

Anàlisi i interpretació d'interaccions d'una botiga amb RFID

Madrid, David

Curs 2014-2015

Directors:

Dra. Anna Carreras Coch

Dr. David F. Nettleton

GRAU EN ENGINYERIA EN INFORMÀTICA



Universitat
Pompeu Fabra
Barcelona

Escola
Superior Politècnica

Treball de Fi de Grau

Als meus avis, Ana, Salazar, Julia i Serafín.

Als meus pares, Marina i Emiliano.

Per estar sempre quan se us necessita.

Agraïments

Voldria donar les gràcies als meus tutors de projecte, Anna Carreras i David Nettleton, per tot l'assessorament rebut i tota l'atenció que m'ha sigut prestada, i sobretot, per l'amabilitat, el bon tracte amb el que sempre m'han tractat i la paciència que han tingut amb mi.

A en Gabriel Martins, pels tots els seus consells i la seva ajuda en aspectes tècnics de programació, que es va prestar en ajudar-me en els moments en que la situació ho requeria.

A tot l'equip de Keonn, i en especial a en Rafael Pous, per donar-me la possibilitat de participar en un projecte d'RFID ambiciós i de gran importància actualment.

Al personal de la botiga, per tota informació prestada amb els catàlegs de productes a la fase inicial.

A tots els amics que m'han donat ànims i tot el seu suport, per a motivar-me encara més amb la realització del projecte, i han dedicat part del seu temps a que jo no perdés el meu.

I finalment, però no menys important, a la meva família, per fer de mi el que sóc actualment, per donar-me i inculcar-me uns valors i sobretot per ensenyar-me a que tot i que a vegades no es parteix de la millor situació, retirar-se no és una opció.

A tots i cadascú de vosaltres, moltíssimes gràcies, són únics.

Resum

Avui dia a les botigues continuen tenint problemes amb l'inventari i no hi ha una bona coordinació entre el que es té emmagatzemat, el que es disposa a les tendes físiques, o inclús el que apareix a les tendes *online*.

RFID apareix com una alternativa que sembla resoldre aquest problema. En aquest projecte, s'ha realitzat una aproximació del que seria la implementació d'un sistema que treballa amb les dades de l'inventari de cada producte, emmagatzemats gràcies als *tags*, que al seu torn són llegits pel lector RFID a cada interacció dels clients de la tenda als emproadors.

Es pretén realitzar un anàlisi de la informació que proporciona l'inventari, descobrint així característiques de les quals se'n podria extreure una valuosa informació que ajudaria a la tenda de roba a enfocar millor les seves campanyes de moda, a millorar la seva distribució interna o a tenir una sèrie de mètriques útils per al futur.

Abstract

Nowadays stores continue having problems with their inventory and there is not a good coordination between what they have stored, what they dispose in physical stores, or even what appears in online stores.

RFID appears as an alternative that seems to solve this problem. In this project, it has been performed what would be an implementation of a system that works with the data from the inventory of each product, stored by tags, which in turn are read by the RFID reader on each customer interaction in the changing rooms.

It's intended to perform an analysis of the information provided by the inventory, discovering then features from which we could extract valuable information that would help the store to focus on their fashion campaigns, improve their internal distribution or have a series of metrics useful for the future.

Índex

Agraïments	v
Resum	vii
Abstract	ix
Llista de figures	xiii
Llista de taules	xiv
Capítol 1: Introducció	1
1.1 Context i motivacions del projecte	1
1.2 Planificació del projecte	5
Capítol 2: Estat de l'art	7
2.1 Treballs relacionats	7
2.2 Projectes relacionats	10
Capítol 3: Tecnologia i programari	17
3.1 RFID	17
3.2 R	19
Capítol 4: Metodologia	21
4.1 Plantejament inicial	21
4.2 Anàlisi de la informació	22
4.3 Diagrama relacional	26
4.4 Tractament de la informació	27
Capítol 5: Resultats de l'anàlisi	31
5.1 Anàlisi segons els productes vistos	31
5.2 Anàlisi segons les famílies de productes vistos	34
5.3 Anàlisi segons les composicions vistes	37
5.4 Anàlisi comparatiu mensual	38
Capítol 6: Conclusions	41
6.1 Resolució i conclusions	41
6.2 Treballs futurs	42

Capítol 7: Problemàtica	43
7.1 Principals problemes i requeriments derivats	43
Bibliografia	45
Annexos	47
Annex 1: Project charter	47
Annex 2: Codi del projecte	51

Llista de figures

Capítol 1: Introducció

Figura 1.1 Implantació de la tecnologia RFID en funció de la mida d'empresa	2
Figura 1.2 Objectius i usos de la tecnologia RFID en funció de la mida d'empresa	3
Figura 1.3 Work Breakdown Structure del projecte	5

Capítol 2: Estat de l'Art

Figura 2.1 Tag en l'etiqueta del producte	11
Figura 2.2 Sistema de lectura de tags de Zara	12
Figura 2.3 Burberry magic mirrors	14
Figura 2.4 Personalització al magic mirror	14
Figura 2.5 Tag postvenda Burberry actiu	15

Capítol 3: Tecnologia i programari

Figura 3.1 AdvanReader-150	17
Figura 3.2 Advantenna p-33	18
Figura 3.3 AdvanCloud	18

Capítol 4: Metodologia

Figura 4.1 Diagrama relacional de les dades	26
Figura 4.2 Llista enllaçada aplicant un anàlisi a les dades en JSON	28
Figura 4.3 Mastertable de les dades	30

Capítol 5: Resultats de l'anàlisi

Figura 5.1 Nom i preu dels productes més vistos	32
Figura 5.2 Nom i preu dels productes menys vistos	33
Figura 5.3 Llista dels productes menys vistos	33
Figura 5.4 Famílies i composició de productes més vistos	35
Figura 5.5 Famílies i composició de productes menys vistos	36
Figura 5.6 Composició i preu dels productes més vistos	37

Figura 5.7 Composició i preu dels productes menys vistos	38
Figura 5.8 Comparativa mensual dels productes més vistos	39

Annex 1: Project Charter

Figura A1.1 Diagrama de l'estimació temporal del projecte	49
Figura A1.2 Taula de rols i responsabilitats del projecte	49
Figura A1.3 Taula de stakeholders del projecte	50

Llista de taules

Capítol 1: Introducció

Taula 1.1 Objectius i usos de la tecnologia RFID a petites empreses al 2014	3
Taula 1.2 Objectius i usos de la tecnologia RFID a mitjanes i grans empreses al 2014	4

Capítol 4: Metodologia

Taula 4.1 Fragment del fitxer seasonData	27
Taula 4.2 Dataframe resultant de sessiondata	29

Capítol 1: Introducció

1.1 Context i motivacions del projecte

Un dels objectius, si no el primer, que comparteixen la majoria d'empreses dedicades a la venda de productes tèxtils és obtenir el màxim possible d'ingressos minimitzant els costos. A tot això l'hem de sumar que el client és sempre la base del negoci. Per tant, s'ha de procurar facilitar qualsevol interacció que tingui a veure amb el client i la tenda, ja sigui amb el propi personal o amb el producte.

Per aquest motiu, són moltes les empreses que cada cop inverteixen més temps i recursos en la bona organització interna de la cadena de subministrament, o en la correcta elaboració de l'inventari dels productes.

Aquest últim punt no ens ha de sonar massa estrany: quants cops hem vist a una pàgina web que una tenda té un producte disponible i al anar-hi ens diuen que no ho està? O que hi és però a una altra tenda de l'empresa? També no podem oblidar els casos de productes on no s'ha aplicat la rebaixa de preu que a les etiquetes apareix. I com aquests, molts més casos que poden donar-se, i tots provenen del mateix lloc: l'inventari.

És per això que actualment les empreses intenten cuidar al màxim tot el catàleg de productes i informació del que disposen. Abans això no era una tasca fàcil, no hi havia res automatitzat i a poc a poc es va començar a automatitzar. Amb l'aparició de noves tecnologies s'ha pogut millorar el rendiment donat, com és el cas de la tecnologia de la identificació a través de ràdio freqüència (RFID).

Com podem observar a la *Figura 1.1*, l'any 2009 un 20% de les grans empreses (250 o més treballadors) ja adoptaven RFID en els seus inventaris i productes. És una xifra no massa alta però hem de considerar que en aquell moment aquesta tecnologia estava en fase d'introducció a Espanya i necessitava créixer i augmentar el número d'adopció. Pel que fa a les Pimes, un 8.9% de mitjanes empreses (de 50 a 249 treballadors) disposaven de lectors i *tags* a les seves tendes; i un 3.1% a les petites empreses (de 10 a 49 treballadors). En les microempreses el percentatge d'adopció era d'un 0.8%.

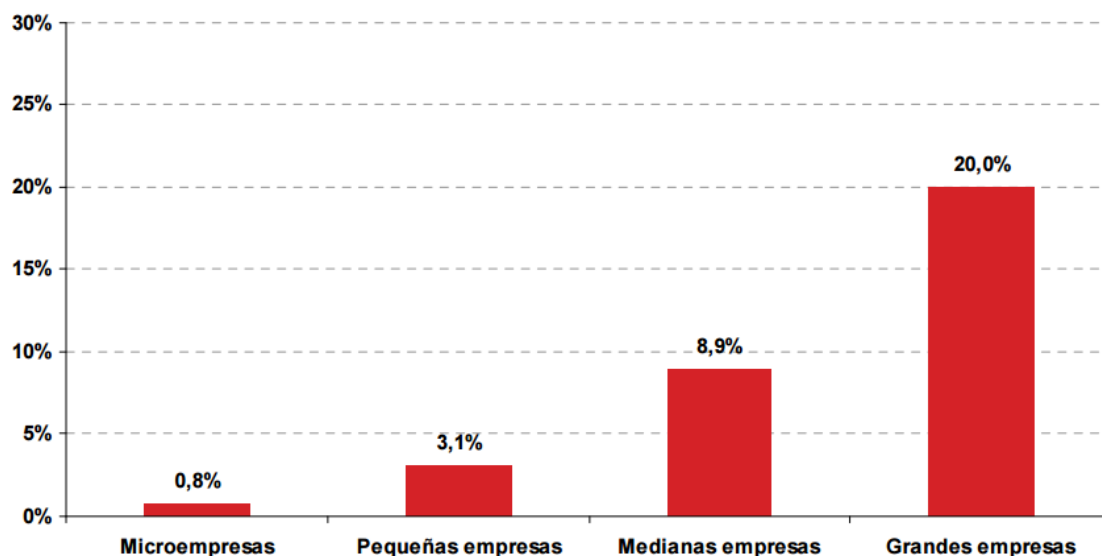


Figura 1.1 Implantació de la tecnologia RFID en funció de la mida d'empresa. Font: ONTSI a partir de les dades del INE 2009

Si ho mirem per sectors, les primeres empreses a adoptar aquesta tecnologia van ser aquelles relacionades amb activitats d'emmagatzematge i transport (12.2%). Seguides pel sector financer (8.2%), la informàtica, les telecomunicacions i els audiovisuals (7.5%), el comerç de caràcter majorista (6.2%) i finalment el minorista (5%).

També podem enfocar més la visió cap a un punt de vista com el de l'ús o aplicacions donat a l'RFID. Així doncs, segons la *Figura 1.2*, l'any 2009 es podia observar que les grans empreses i Pimes es decantaven per un major seguiment i control de les seves cadenes de subministrament i dels seus inventaris, com reflecteix el seu 44.4% seguit de prop per la identificació de persones i control d'accessos, amb un 40.5%

En canvi les empreses de menys de 10 treballadors, les microempreses, donaven més ús a l'RFID amb la seva implantació en sistemes de pagament (peatges, transport de passatgers, etc.) amb un 48.5%. La segona opció era la identificació de productes, amb un 38.7%.

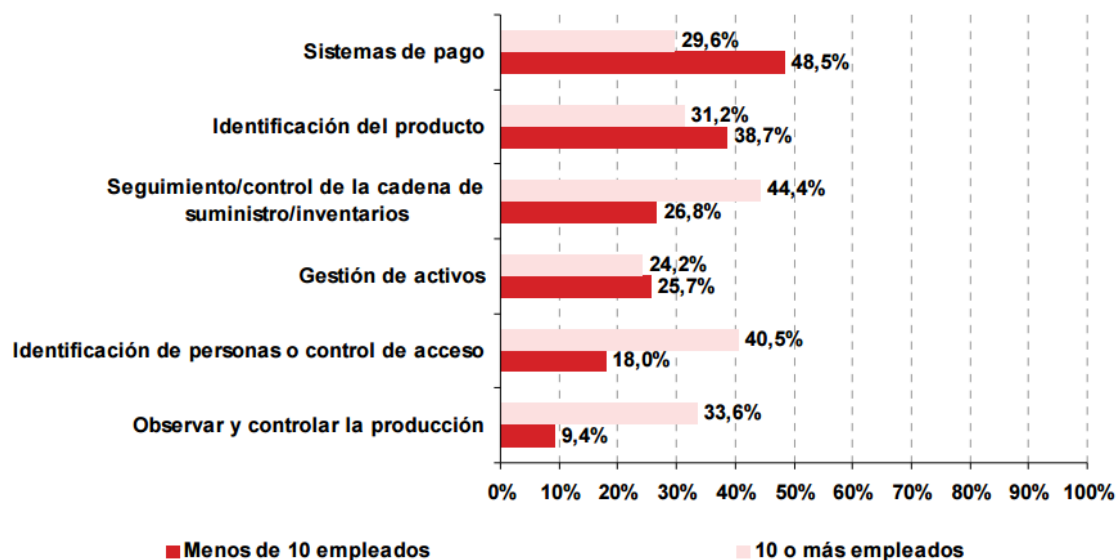


Figura 1.2 Objectius i usos de la tecnologia RFID en funció de la mida d'empresa.
Font: ONTSI a partir de les dades del INE 2009

Si mirem informes recents d'implantació de la tecnologia RFID, tal com la Taula 1.1 i la Taula 1.2 mostren, podem comprovar que ara mateix és una tecnologia incipient que està en continua expansió.

Podem observar un creixement en tots els tipus d'empreses: un 1.3% en les microempreses, un 10% més a les petites empreses, un 26.6% a les mitjanes empreses i un 41.6% a les grans empreses, contrastant amb les dades del 2009. Per tant estem davant d'una nova forma d'emmagatzematge d'informació i gestió d'inventaris i persones.

Enquesta de l'ús de les TIC i Comerç Electrònic (CE) en les empreses 2013-2014	
Resultats nacionals	
Variables de us de TIC (primer trimestre de 2014 i Comerç Electrònic (CE) 2013). Resultats per empreses amb menys de 10 empleats per principals variables.	
Unitats percentatges	
32.1 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID per a la identificació de persones o control d'accés	0.6
32.2 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID com part del procés de producció o del servei d'entrega del producte	0.8
32.3 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID per a la identificació del producte després del procés de producció	0.7

Taula 1.1 Objectius i usos de la tecnologia RFID a petites empreses al 2014. Font: INE 2014

Enquesta de l'ús de les TIC i Comerç Electrònic (CE) en les empreses 2013-2014			
Resultats nacionals			
Variabls de us de TIC (primer trimestre de 2014) per agrupació d'activitat (excepte CNAE 56, 64-66 i 95.1), principals variables i mida de l'empresa.			
Unitats percentatges			
	De 10 a 49	De 50 a 249	De 250 i més
Total d'Empreses			
32.1 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID per a la identificació de persones o control d'accés	5.6	17.2	29.8
32.2 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID com part del procés de producció o del servei d'entrega del producte	4.7	11.4	20.3
32.1 Ús de tecnologies d'identificació per radiofreqüència (RFID): % d'empreses que fan ús d'instruments de RFID per a la identificació del producte després del procés de producció	2.8	6.9	11.5

*Taula 1.2 Objectius i usos de la tecnologia RFID a mitjanes i grans empreses al 2014.
Font: INE 2014*

Moltes tendes encara no disposen encara de RFID o estan començant a implantar-lo, per tant sembla una bona idea poder realitzar un anàlisi de tota la informació que s'obté de les antenes de tendes per extreure conclusions que poden ajudar a enfocar les noves campanyes de roba en l'actual panorama.

1.2 Planificació del projecte

Per portar un control de com estaria planificat el projecte va semblar adient l'elaboració d'un *Work Breakdown Structure* tal com es pot veure a la *Figura 1.3*

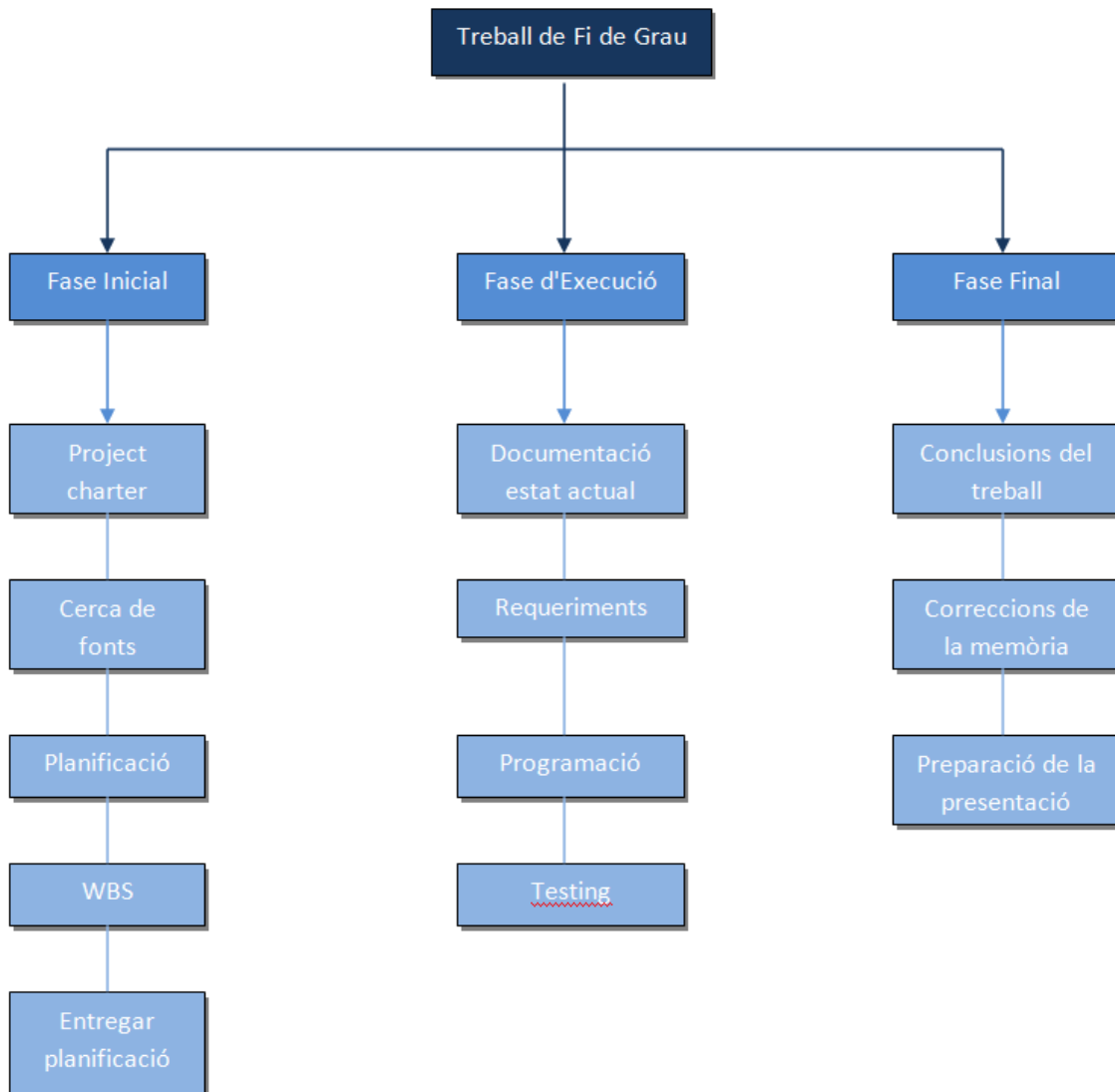


Figura 1.3 Work Breakdown Structure del projecte

Capítol 2: Estat de l'Art

2.1 Treballs relacionats

Com hem pogut veure, l'ús de RFID no és gens nou i és adient cercar informació de treballs ja elaborats relacionats amb el nostre objectiu de controlar l'*stock* de l'inventari.

Hi ha certs articles que aborden una sèrie de qüestions interessants a l'hora d'implementar l'RFID a la tenda, una petita mostra serien els següents:

En l'article de George Roussos [1] s'incorpora el sistema de *Vendor Managed Inventory (VMI)* on és el dependent el que especifica tots els valors i quantitats a enviar. Es basa en la unió de dues tecnologies: el suport de l'intercanvi electrònic de dades entre socis i distribuïdors, i els identificadors del producte, on sovint vénen a ser codis de barres, que ofereixen una identificació estàndard i regulada. A través de l'ús del RFID a través del nivell d'*Stock Keeping Unit (SKU)* es millora l'eficiència a través d'una sèrie d'escaneigs de l'*stock* de la tenda que fan augmentar la precisió de la informació emmagatzemada als inventaris.

Molts treballs es centren en el tema de la seguretat i la privacitat de la incorporació de l'RFID a les tendes, com és el cas de l'article de George Roussos i Theano Moussouri [2] on es va dur a terme un anàlisi quantitatiu i qualitatiu amb un lector RFID. Per unanimitat de les respostes, van poder afirmar que la informació obtinguda i guardada per les tendes era una violació directa a la seva privacitat. Per tant és un punt a tenir en compte, la gent pensa que el tema de controlar l'entorn és molt més important que les oportunitats comercials que el RFID pot tenir.

A l'article de Kristina Huber, Lisa Houck e Igor Vinogradov [3] tracten també el tema de l'RFID des de la seguretat i la privacitat des de la part del consumidor. És interessant veure com donen informació per a desmentir certs mites, rumors o pors que la gent ha creat envers la tecnologia RFID.

Però sobretot s'endinsen en un punt vital: el màrqueting. On s'afirma que al contrari del que la gent pensa, la tecnologia RFID no permet a les tendes recollir cap informació addicional de la qual s'agafaria també amb una targeta de fidelitat, targetes de crèdit específiques de botigues o simplement amb targetes de crèdit normals. Es parla del fet que aquesta presumpta informació extra guardada amb el RFID no té cap sentit, ja que aquesta tecnologia actua només com a coordinadora dels mètodes de màrqueting directe o enfocat. Però també es diu que encara que hi puguin haver noves oportunitats de màrqueting directe, el seu cost per a la implementació serà bastant elevat, i afirmen que amb l'ús de l'RFID s'augmentarà la rebaixa de preus i promocions de productes.

Parlant exclusivament de privacitat, l'enfocament més estès i estudiat és la problemàtica de la informació emmagatzemada pel *tag*. Hi hauria dos mètodes més estesos:

1. La primera seria la del *tag killing*, exposada a l'article de Barbara Keifenheim [4], on es té implementat un mètode que deshabilita la funcionalitat que el *tag* conté un cop que el consumidor adquireix.

El principal al·licient per al seu ús és que permet un alt grau de privacitat al consumidor a un cost que es pot negligir. Però com a gran inconvenient és que les tendes configurarien *killer kiosks*, on seria responsabilitat del consumidor "optar" a realitzar aquesta pràctica o no.

En conseqüència, pot ocasionar un augment de l'error humà [5], ja que estarien deshabilitant un procés manualment per milions de clients. Altre inconvenient a tenir en compte és que potser només hi ha un cert nombre de *killer kiosks* a les tendes, donant lloc a la prevenció dels clients d'eliminar *tags* a causa de les molèsties.

També es podria eliminar un *tag* de manera temporal gràcies a un software de bloqueig. Però en principi hi hauria un comportament anomenat *waking up* amb els *tags* que permetrien tornar a fer servir el mateix *tag* per analitzar consumidors.

2. L'altra alternativa es coneix com a *blocker tag* [6], on es fa referència a un mètode d'engany dels lectors RFID mentre escanegen. La manera que tenen d'aconseguir això és simulant que el consumidor està intentant llegir molts *tags* de cop al lector simultàniament. També es pot bloquejar selectivament un *tag* en concret, per exemple si és notificat pel proveïdor a través del seu codi identificador.

Com a avantatges, aquest sistema ofereix un cost d'implementació baix, ja que no cal modificar cap paràmetre del *tag*, però com a inconvenient s'ha de remarcar que les mesures de protecció que ofereix són limitades:

- Els consumidors no poden confirmar que els *blocker tags* estiguin actualment funcionant i, per aquest motiu, no es pot afirmar que la seva privacitat estigui sent protegida.
- L'usuari és en tot moment qui realitza les accions amb els *tags*. Això pot generar inquietuds sobre el control de la privacitat de la informació.

2.2 Projectes relacionats

És important veure també com empreses reals han integrat l'RFID, i com resolen problemes quotidians que es pot trobar una persona a l'hora d'utilitzar aquesta tecnologia. A continuació s'exposen una sèrie d'exemples per veure les solucions donades a aquesta tecnologia:

1. M&S

L'important grup Mark & Spencer va implantar a la seva cadena d'establiments un sistema basat en el RFID. Al seu departament de moda, concretament va començar a provar amb vestits, tal com es pot veure a la *Figura 2.1*.



Font: <http://retaildesignblog.net/>

El seu objectiu era augmentar l'eficiència del control de l'*stock*. Tal com va dir Stuart Senior, director del departament de Tecnologies de la Informació [7]:

"Amb un control automàtic de l'stock podem fer-nos una imatge perfectament del gènere que tenim a les tendes i al magatzem. Si podem millorar la precisió amb les dades, podem reabastir l'stock d'una manera més precisa."



Van observar que l'actual pràctica de fer els abastiments a través d'estimacions segons les dades dels punts de venda de les tendes suposava en un 8-12% el percentatge d'errors que hi havia entre les dades que disposaven.

Amb l'RFID van augmentar la precisió d'aquestes estimacions, cosa que va derivar en un augment significatiu en la disposició dels productes i, per contrapartida, de les seves vendes.

Figura 2.1: Tag en l'etiqueta del producte Font: <http://www.wired.com/>

2. Inditex

El grup Inditex fa un temps que va incorporar ja l'RFID a les seves tendes, i esperen tenir cap a 2016 un cobriment del 100% a tots els establiments de Zara. Per a tal fi, l'empresa ha adquirit 500 milions de xips RFID.

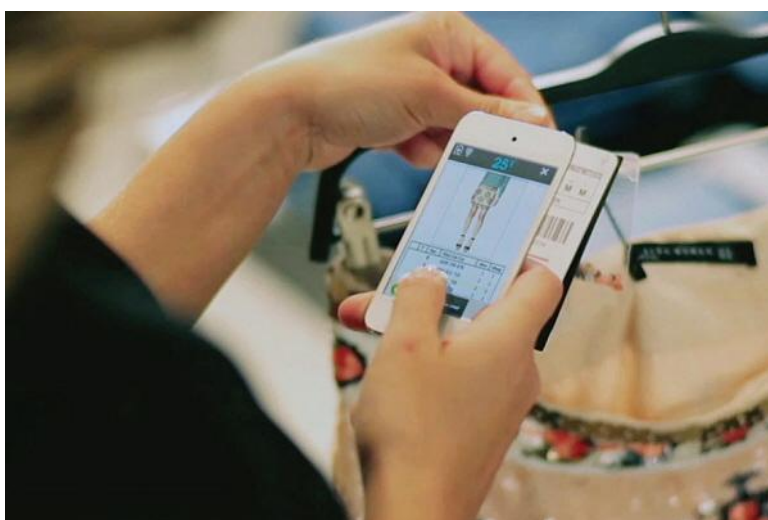


Font: <http://interiordesignshop.net/>

L'objectiu d'Inditex és realitzar un millor seguiment dels seus *stocks* i poder reabastir les seves tendes més ràpidament amb nous productes.

Tal com el president de Tyco Retail Solutions, empresa encarregada de la incorporació de l'RFID a Inditex, Ricardo Arroyo diu [8]:

“La seguretat en la informació de l'inventari fa que no es perdin vendes. Trobem que, en general, existeix un desfasament d'entre el 20% i el 35% en l'inventari que compara el magatzem i la tenda.”



Amb els nous xips RFID integrats a cada producte del l'inventari, tal com es pot veure a la *Figura 2.2*, ara quan es faci una venda de qualsevol producte, s'emet una ordre al magatzem per tal que el producte torni a estar a la tenda.

Figura 2.2 Sistema de lectura de tags de Zara. Font: <http://www.industrysolutions.co.kr/>

En aplicar a les seves tendes la tecnologia de l'RFID, es va poder comprovar que l'atenció al client és de millor qualitat, es va incrementar el control de la seguretat, els inventaris van passar a ser més eficients, i també hi va haver una millor agilitat i precisió respecte als processos de recepció de les peces de roba.

A Inditex solen emprar dos tipus de *tags*: uns plans, col·locats habitualment a sobre de les etiquetes dels preus dels productes, i uns altres; aplicats a les etiquetes de seguretat de les peces. Tot i que aquests últims són els que més utilitza la companyia.

Un altre punt a favor de l'RFID que va fer decantar-se més per aquest tipus de *tags* és la seva fàcil reutilització, ja que qualsevol persona de la tenda les pot desactivar fàcilment en cobrar el producte.

3. Burberry

L'empresa creada per Thomas Burberry va començar a adoptar el RFID a les seves botigues sobre l'any 2012. El seu objectiu era millorar el control de l'inventari però també potenciar l'experiència del client.



Font: <http://www.luxe-en-france.com/>

Per millorar l'estada dels clients dins del provador o a l'interior de la tenda, van incorporar *magic mirrors*, com es pot veure a les *Figures 2.3 i 2.4*, capaços de transformar contingut multimèdia de cada producte i rangs en una pantalla digital. Analitzant el nivell de satisfacció de l'usuari van adonar-se que ajuda al client a sentir-se especial i per un altre banda serveix d'eina per a la integració amb el comerç social. Aquests miralls es troben fixos dins la tenda, per tant, faciliten l'optimització i el *testing*, sense necessitat de provar amb molts dispositius dels clients.



Figura 2.3 Burberry magic mirrors. Font: <http://fashionbi.com/>

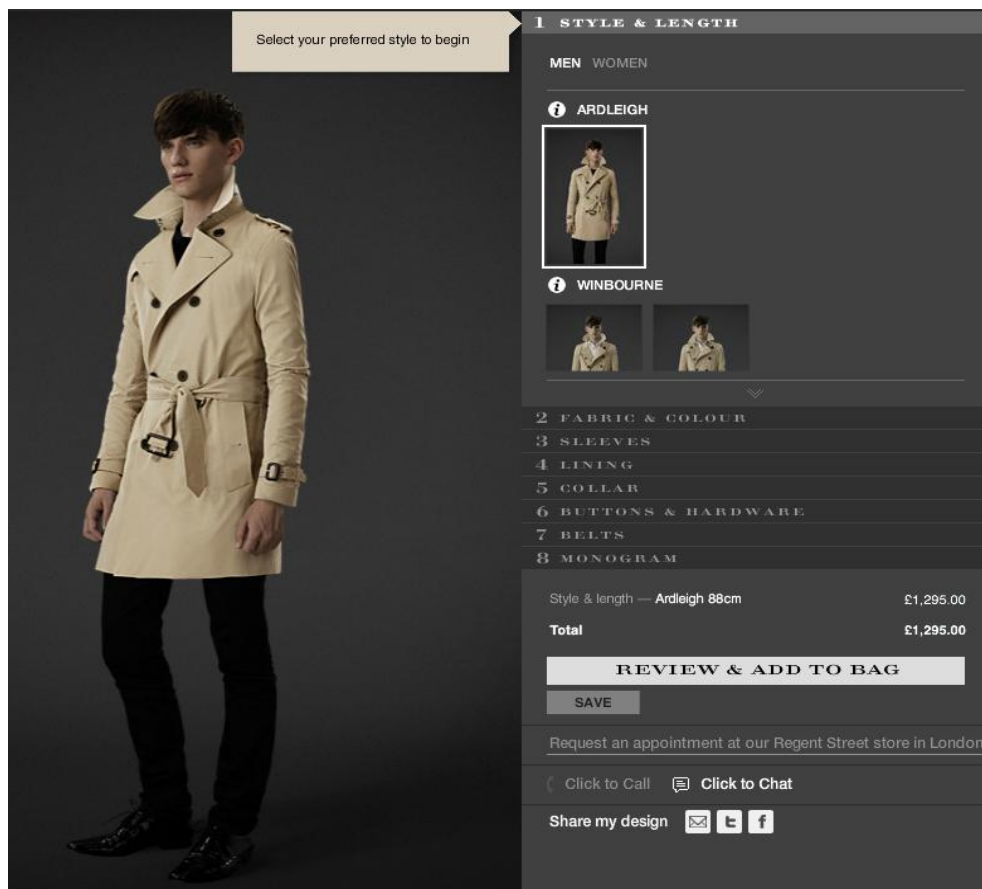


Figura 2.4 Personalització al magic mirror. Font: <https://vmcfashion.wordpress.com/>

Com sol passar a les altres empreses de moda que incorporen l'RFID, el personal de la tenda, desactivarà els tags quan una persona realitzi una compra. Però, almenys a Burberry hi cap la possibilitat, sempre amb el consentiment del client, que els *tags* no es desactivin. L'objectiu seria realitzar un seguiment de la vida de la peça, sempre i quan estigui dins el rang d'un lector RFID. Un exemple a la *Figura 2.5*.

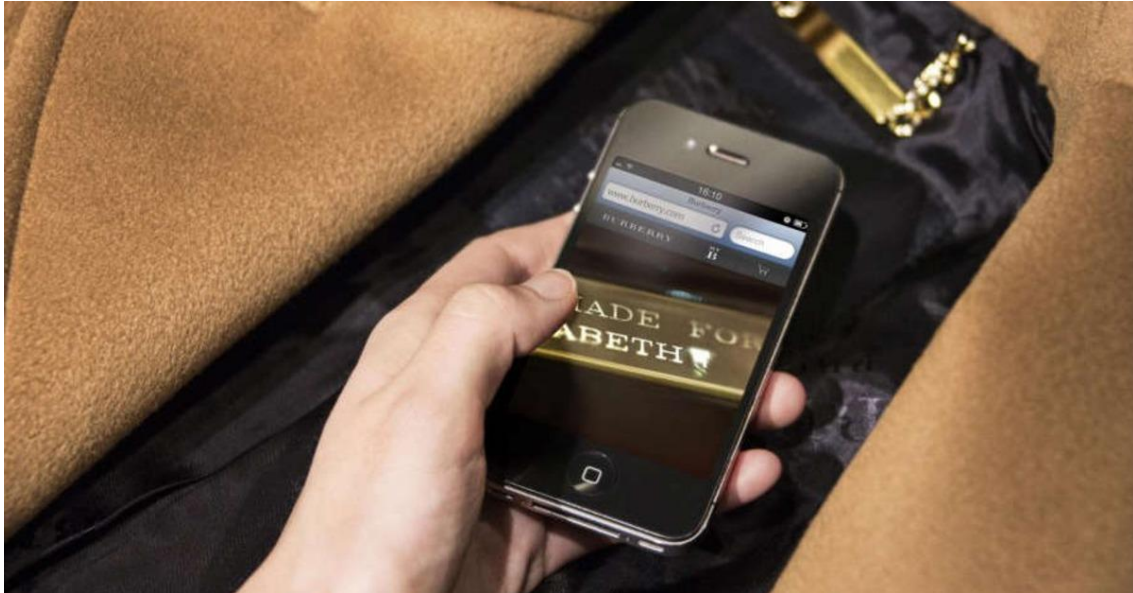


Figura 2.5 Tag postvenda Burberry actiu. Font: <http://mashable.com/>

Capítol 3: Tecnologia i programari

3.1 RFID

La base sobre la que gira el projecte és l'anàlisi d'informació captada pel sistema d'RFID per a l'elaboració de l'inventari. Per tant, entenem que la incorporació i l'ús de la tecnologia RFID juguen un paper fonamental.

El sistema sobre el qual es basa la captació en si és simple: la tenda té etiquetades totes les peces de roba mitjançant *tags*. Qui s'encarrega d'obtenir la informació és el lector, en aquest cas es tracta de l'AdvanReader-150 de l'empresa Keonn, tal com es pot veure a la *Figura 3.1*. Hi ha un col·locat en cadascun dels tres proveïdors que disposa la tenda que, juntament amb una pantalla que fa la funció de *magic mirror* mostren a l'usuari la peça de roba que ha estat escanejada a dins del proveïdor i que es disposa a provar-se. Aquesta funció permet a l'usuari poder accedir a certes opcions sobre l'article que vol provar-se. Algunes poden ser evidents com el color, el gènere del teixit o la talla. D'altres, serveixen per a facilitar l'experiència, com per exemple la visualització de productes recomanats segons el producte que apareix a la pantalla, o la possibilitat de demanar al personal de la tenda que et portin una altra peça de roba sense que l'usuari s'hagi de moure.



Figura 3.1 AdvanReader-150. Font: <http://www.keonn.com/>

Totes les accions que un usuari pot realitzar dins de l'emprovador són captades a través de les antenes (veure *Figura 3.2*), i informàticament es puja tota la informació al *cloud*. Amb aquesta informació és sobre la qual es treballa i es fan les proves de funcionament del sistema, ja que son dades reals a temps real de totes les interaccions que han succeït dins de la tenda en un període concret, tal com mostra la *Figura 3.3*.

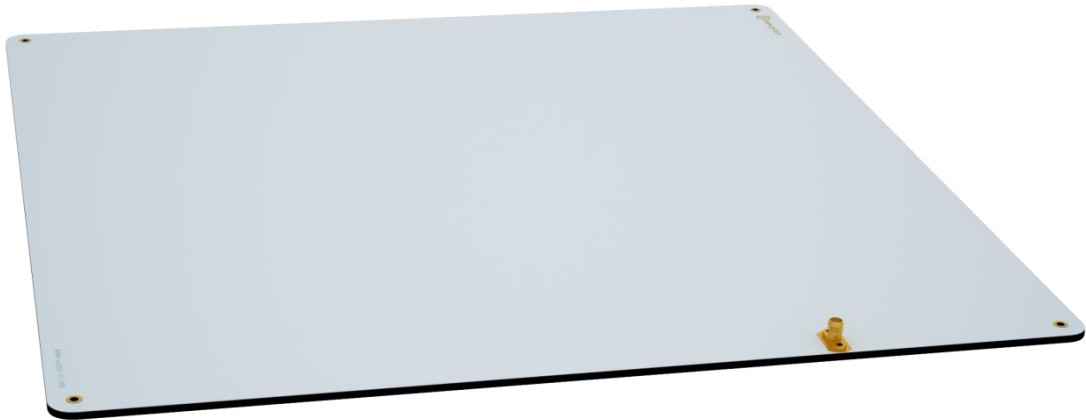


Figura 3.2 Advantenna p-33. Font: <http://www.keonn.com/>

The screenshot shows the AdvanCloud web application interface. The page title is "Welcome to Advancloud" and the URL is "https://advancloud.keonn.com/advancloud/appreports/generate". The page features a navigation menu with options like "App", "App Template", "App Item", "App User", "Permission", "App Item Category", and "Attribute Type". Below the navigation menu, there are search filters for "App", "App Template", "App Item", "App User", "Permission", "App Item Category", and "Attribute Type". The "App" filter is set to "start" and the "Report" filter is set to "ReportByProducts". The "Start Date" and "End Date" fields are empty. The "Report Format" is set to "Screen". A "GENERATE" button is visible below the filters.

code	views	requests	retailer	pct	name
1000132328439	3		office		CANDIE'S SNOW BR...
1000132328439	5	2	desigual	40	CANDIE'S SNOW BR...
1000132328439	7		gumption		CANDIE'S SNOW BR...
1000766065069	4		office		ELLE FEMME TROPI...
1000766065069	4		gumption		ELLE FEMME TROPI...
1001571386738	1		office		DANA BUCHMAN SA...
1001571386738	2		hk		DANA BUCHMAN SA...
1001571386738	8		desigual		DANA BUCHMAN SA...
1001571386738	36	9	gumption	25	DANA BUCHMAN SA...
1001795309889	2	1	gumption	50	ELLE RED ALERT D...
1001795309889	2		desigual		ELLE RED ALERT D...
1002563839607	2		gumption		ELLE FEMME TROPI...
1002563839607	4		office		ELLE FEMME TROPI...
1002649589280	2		gumption		ELLE LADY LUXE
1003143320225	2		office		JENNIFER LOPEZ N...
1004075075528	2		office		LAUREN CONRAD L...
1005481946358	1		gumption		JENNIFER LOPEZ W...
1006476991612	1		office		LAUREN CONRAD T...
1007037765437	1		gumption		LAUREN CONRAD S...
1007037765437	1		hk		LAUREN CONRAD S...

Figura 3.3 AdvanCloud. Font: <http://www.keonn.com/>

3.2 R

Cada projecte té una estructura marcada, i un objectiu definit. Aquest és un projecte adreçat al *data mining* on l'estructura de la informació pren el valor màxim.

Es treballa amb un volum important d'informació, on no tot ha de ser correcte i coherent, però tot estarà estructurat en taules d'on s'han d'extreure les dades.

Segons *Wikipedia*, R és un llenguatge de programació emprat per un gran conjunt d'estadístics i *data miners* per desenvolupar software estadístic i d'anàlisi de dades. Aquesta és una de les grans raons per escollir R com a llenguatge i no un altre, però encara hi ha més motius [9]:

1. La comunitat: és un servei *open source*, amb una gran comunitat darrere, amb multitud de fòrums i recursos *online*, com poden ser cursos interactius.
2. L'entorn de programació: RStudio és un *IDE* que facilita moltes accions quotidianes, a la vegada que es té ple control sobre totes les coses que succeeixen a l'elaboració dels codis, i permet la descàrrega de llibreries i paquets externs d'una manera molt intuïtiva. També incorpora l'ús de manuals per ajudar a l'usuari a fer servir els recursos de programació dels que disposem amb R.
3. La manipulació de les dades: és realment fàcil entendre i programar en R un software adreçat a *data mining*. Té unes funcions preestablertes que faciliten a l'usuari la tasca d'haver de crear el seu propi codi per a l'extracció d'informació, com podria passar amb l'ús d'un altre llenguatge (*e.g.* Java).
4. La visualització de les dades: tal com es parla en el punt 3, si és un llenguatge adreçat a manipular dades, també ho és per visualitzar el resultat d'aquestes manipulacions. Tot d'una manera molt senzilla i intuïtiva, mantenint una complexitat que permet a l'usuari poder crear imatges complexes que reproduïxin amb precisió els resultats de la informació.
5. Prediccions: permet la funcionalitat de l'aprenentatge automàtic. On, mitjançant un conjunt de dades de *testing* i la resta de *training*, permet arribar a entendre la informació que es necessita per a què sense repetir el procés, tinguem a l'abast aquesta informació proporcionada pel sistema.

Capítol 4: Metodologia

4.1 Plantejament inicial

Trobem dues parts ben diferenciades dins del procés a realitzar amb les dades:

1. El tractament de la informació
2. L'anàlisi qualitatiu i quantitatiu

Al tenir un volum important d'informació és adient aplicar un tractament a les dades extretes del lector RFID. Netejar, canviar el format o fins i tot, transformar solen ser els processos més emprats per a obtenir l'*input* de la informació.

Un cop les dades estan correctament estructurades i són perfectament accessibles, es procedeix a analitzar els resultats: es realitzen anàlisis que ens serveixen per a obtenir certes conclusions que són l'objectiu del projecte. Ens podem ajudar d'eines de suport, com taules, gràfiques, per a fer que la visualització dels resultats sigui més intuïtiva i aportï més informació a l'hora d'extreure les conclusions que se'n deriven.

4.2 Anàlisi de la informació

L'objectiu d'aquesta secció és poder agrupar tota la informació per a que passi a estar disponible en una sola taula. Per a tal fi, és necessari tractar cada fitxer per separat aplicant-hi les transformacions que calgui:

1. *SeasonData.csv*:

Representa les dades de les diferents sessions en què es proven els productes.

Conté la següent informació útil:

- *Retailer*: localització de la botiga.
- *Shop*: codi identificador de la botiga.
- *Screen*: codi identificador de l'emprovador on s'efectua la interacció.
- *Datetime*: data i hora de l'inici de la sessió.
- *Sessionnum*: codi identificador de la sessió.
- *Sessiondata*: Totes les dades relacionades amb la sessió, en format JSON.

Sessiondata és una llista amb la mateixa estructura, formada per:

- *Scanned*: llista dels productes que s'han escanejat dins de l'emprovador.
 - *Timestamp*: codificació numèrica de la data i l'hora.
 - *Epc*: codi identificatiu del *tag* llegit per l'antena RFID.
 - *Code*: codi de barres del producte.

- *Requested*: llista dels productes que s'han demanat dins de l'emprovador.
 - *Timestamp*: ídem a l'atribut de la llista *scanned*.
 - *Assistant*: personal de la botiga que ha atès al client.
 - *Status*: estat actual del producte demanat (pendent, sense *stock*, servit, eliminat, etc.).
 - *Code*: codi del producte demanat (veure *productID*).
 - *RequestID*: número identificador de la petició.
 - *Type*: tipus de petició realitzada (producte, assessorament, etc.).
 - *Resptime*: duració en ms de la petició fins que ha estat resolta.

- *Socialcodes*: codi identificador generat quan l'usuari realitza alguna interacció que fa referència amb alguna xarxa social.

- *Viewed*: llista dels productes que l'usuari ha visualitzat a través del *magic mirror*. Poden ser eleccions pròpies, o a través de les recomanacions del *software*.
 - *Timestamp*: ídem a l'atribut de la llista *scanned*.
 - *Name*: nom del producte seleccionat.
 - *Code*: codi del producte seleccionat. (veure *productID*).

- *Sessionstart*: *timestamp* que indica l'inici de la sessió.
- *Sessionend*: *timestamp* que indica el final de la sessió.

2. *Skucodes.csv*

Dóna informació més específica sobre cada producte. En concret, amb els següents camps:

- *ProductID*: codi identificador intern del producte.
- *SkuID*: número de referència o *stock-keeping* unit del producte.
- *Productdata*: llista amb informació concreta sobre els productes, en format JSON.

Productdata és una llista amb la mateixa estructura, formada per:

- *Colour*: llista amb la informació de color del producte.
 - *Name*: nom del color.
 - *Code*: codi hexadecimal del color.
- *RVcolor*: codi decimal del color.
- *Family*: família del producte (bossa, camisa, cinturó, etc.).
- *Name_es*: nom del producte en espanyol.
- *Description_es*: descripció detallada del producte, en espanyol.
- *Size*: talla del producte.
- *Recommended*: llista de productes recomanats pel sistema recomanador del *magic mirror*.

3. *Maestros.txt*

Dóna més informació específica sobre cada producte. En concret amb els següents camps:

- *SkuIDsize*: codi identificador del producte format per la unió del *skuID+size*.
- *Price*: preu en euros del producte que s'està emprovant a la sessió.
- *Percentage*: percentatge entre 0 i 100 del teixit del producte.
- *Composition*: composició o tipus de teixit amb el qual està fet el producte.
- *Ean13*: codi de barres del producte, ídem al valor.
- *Size*: ídem a l'atribut de *skucodes*.

4.3 Diagrama relacional

Es treballa amb la informació emmagatzemada al *cloud*. En concret necessitem tres arxius: *seasonData.csv*, *skucodes.csv* i *maestros.txt*. Que és on obtindrem la majoria d'informació útil de les dades dels productes. La informació ve relacionada d'una manera en concret que es pot veure a la *Figura 4.1*:

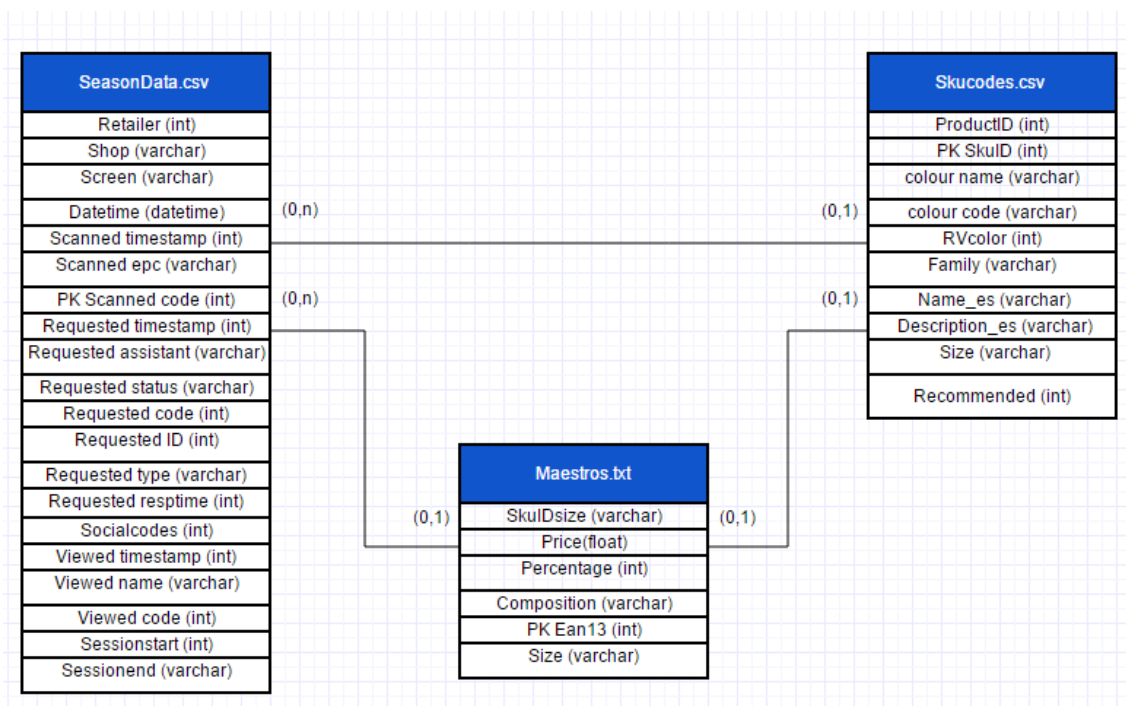


Figura 4.1 Diagrama relacional de les dades

4.4 Tractament de la informació

Un cop es té definida la informació continguda dins dels nostres fitxers de l'*AdvanCloud*, és necessària una transformació del contingut:

Tal com es mostra a la *Taula 4.1*, els camps *sessiondata* i *productdata* es troben en format JSON, i es necessita recollir la seva informació i concatenar-la amb la resta de camps de cada document.

	retailer	store	screen	datetime	sessionnum	sessiondata
1	bcn	shop01	f02.fittingroom	2015/03/18 10:25:41	234047	{"scanned":{"timestamp":"1426673998497","epc":"303602c25c0cb24000000001","code":"8433815130013"},{
2	bcn	shop01	001.fittingroom	2015/03/18 10:52:40	234079	{"scanned":{"timestamp":"1426675652045","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
3	bcn	shop01	001.fittingroom	2015/03/18 11:06:53	234103	{"scanned":{"timestamp":"1426676607407","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
4	bcn	shop01	001.fittingroom	2015/03/18 11:30:20	234109	{"scanned":{"timestamp":"1426677904812","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
5	bcn	shop01	001.fittingroom	2015/03/18 11:41:40	234114	{"scanned":{"timestamp":"1426678543478","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
6	bcn	shop01	f03.fittingroom	2015/03/18 12:18:15	234117	{"scanned":{"timestamp":"1426673705284","epc":"303602c25c0cb24000000001","code":"8433815130013"},{
7	bcn	shop01	001.fittingroom	2015/03/18 12:19:11	234118	{"scanned":{"timestamp":"1426680037737","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
8	bcn	shop01	001.fittingroom	2015/03/18 12:52:01	234126	{"scanned":{"timestamp":"1426682811416","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
9	bcn	shop01	f01.fittingroom	2015/03/18 13:23:49	234137	{"scanned":{"timestamp":"1426684252582","epc":"303602c25c0cb18000000002","code":"8433815129987"},{
10	bcn	shop01	f01.fittingroom	2015/03/18 13:29:49	234140	{"scanned":{"timestamp":"1426685063975","epc":"303602c25c0c780000000002","code":"8433815127709"},{
11	bcn	shop01	001.fittingroom	2015/03/18 13:14:51	234132	{"scanned":{"timestamp":"1426684183490","epc":"303602c25c0c0c0000000001","code":"8433815129956"},{
12	bcn	shop01	f01.fittingroom	2015/03/18 13:43:39	234150	{"scanned":{"timestamp":"1426685423494","epc":"303602c25c0c648000000002","code":"8433815137142"},{
13	bcn	shop01	f01.fittingroom	2015/03/18 13:53:19	234159	{"scanned":{"timestamp":"1426686229760","epc":"303602c25c0c0c0000000001","code":"8433815130334"},{
14	bcn	shop01	f01.fittingroom	2015/03/18 14:02:39	234163	{"scanned":{"timestamp":"1426687050742","epc":"303602c25c0c0c0000000001","code":"8433815130334"},{
15	bcn	shop01	001.fittingroom	2015/03/18 14:19:21	234167	{"scanned":{"timestamp":"1426687865894","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
16	bcn	shop01	001.fittingroom	2015/03/18 14:58:31	234174	{"scanned":{"timestamp":"1426690401847","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
17	bcn	shop01	001.fittingroom	2015/03/18 15:07:51	234176	{"scanned":{"timestamp":"1426690774835","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
18	bcn	shop01	001.fittingroom	2015/03/18 15:40:52	234184	{"scanned":{"timestamp":"1426692944294","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
19	bcn	shop01	001.fittingroom	2015/03/18 16:34:42	234193	{"scanned":{"timestamp":"1426696029516","epc":"303602c25c00c0c000000002","code":"8433815133632"},{
20	bcn	shop01	f01.fittingroom	2015/03/18 17:10:46	234203	{"scanned":{"timestamp":"1426698309308","epc":"303602c25c0cb20000000001","code":"8433815130006"},{
21	bcn	shop01	001.fittingroom	2015/03/18 17:15:12	234204	{"scanned":{"timestamp":"1426698604883","epc":"303602c25c00c0c000000002","code":"8433815133632"},{

Taula 4.1 Fragment del fitxer *seasonData*

Amb la funció *fromJSON* de R, es pot fer fàcilment. I ens permet obtenir una llista enllaçada dels objectes com la de la *Figura 4.2* que els conté, és una forma de jerarquitzar que ajuda bastant quan les nostres llistes contenen un volum important de variables.

També ens dóna informació del tipus de dades que contenen: cada fila del dataframe correspon a una observació, i cada observació pot tenir més o menys columnes, indicades amb el nom de "variables".

En aquest cas, hem utilitzat *nums* per enters, *logi* per dades lògiques (booleans), *chr* per cadenes de caràcters i *factors* per vectors de dades específiques. Aquests vectors venen dividits en nivells (*levels*), si algú d'ells conté més d'una variable, aquesta variable formaria part del següent nivell del vector.

Per indicar que a un camp no hi ha informació s'ha emprat NA, de tipus lògic.

values	
JSONproductdata	Large list (5895 elements, 15.8 Mb)
JSONsessiondata	Large list (2363 elements, 6.3 Mb)
{"scanned":[{"timestamp":1426673998497,"epc":"303602c25c0cb24000000001","code":"8433815130013"}]	
..\$ scanned.timestamp	: num 1.43e+12
..\$ scanned.epc	: Factor w/ 1 level "303602c25c0cb24000000001": 1
..\$ scanned.code	: Factor w/ 1 level "8433815130013": 1
..\$ requested.timestamp	: logi NA
..\$ requested.status	: logi NA
..\$ requested.code	: logi NA
..\$ requested.requestID	: logi NA
..\$ requested.type	: logi NA
..\$ sessiontime	: num 342701
..\$ viewed.timestamp	: num 1.43e+12
..\$ viewed.name	: Factor w/ 1 level "Short quilted windbreaker": 1
..\$ viewed.code	: Factor w/ 1 level "1538120307": 1
..\$ sessionstart	: num 1.43e+12
..\$ sessionend	: num 1.43e+12
{"scanned":[{"timestamp":1426675652845,"epc":"303602c25c0d0cc0000000002","code":"8433815133632"}]	
..\$ scanned.timestamp	: num 1.43e+12
..\$ scanned.epc	: Factor w/ 1 level "303602c25c0d0cc0000000002": 1
..\$ scanned.code	: Factor w/ 1 level "8433815133632": 1
..\$ requested.timestamp	: logi NA
..\$ requested.status	: logi NA
..\$ requested.code	: logi NA
..\$ requested.requestID	: logi NA
..\$ requested.type	: logi NA
..\$ sessiontime	: num 308115

Figura 4.2 Llista enllaçada aplicant un anàlisi a les dades en JSON

Un cop es *parseja* la informació continguda en format JSON, es procedeix a generar un *dataframe* com el de la Taula 4.2 amb tota la informació extreta dels documents a partir de la funció *data.frame()*. D'aquesta manera tenim la nostra informació estructurada de manera que podem accedir-hi ràpidament per ordre de posició dins la llista inicial. Sabent que sempre seguirem una estructura *list.item*, podem aplicar una funció que ens permeti accedir a les dades i guardar-les al nostre *dataframe*.

Simplement hem de considerar que hi hagi camps en les dades on aquestes no estiguin informades o contenint informació, i aplicar-hi una condició per tal que a l'hora de crear el nostre *dataframe* el procés sigui el correcte i els resultats obtinguts, els esperats.

Al document de Maestros.txt no és necessari aquest pas, la informació ve separada per ";" i només amb la lectura del fitxer R sap distingir les variables i les emmagatzema a un *dataframe* correctament.

	scanned.timestamp	scanned.epc	scanned.code	requested.timestamp	requested.status	requested.code	requested.requestID	requested.type	sessiontime
1	1.426674e+12	303602c25c0cb24000000001	8433815130013	NA	NA	NA	NA	NA	342701
2	1.426676e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	308115
3	1.426677e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	205843
4	1.426678e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	316310
5	1.426679e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	357695
6	1.426674e+12	303602c25c0cb24000000001	8433815130013	NA	NA	NA	NA	NA	7389962
7	1.426681e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	313643
8	1.426683e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	310110
9	1.426684e+12	303602c25c0cb18000000002	8433815129987	NA	NA	NA	NA	NA	776751
10	1.426685e+12	303602c25c0c788000000002	8433815127709	NA	NA	NA	NA	NA	325389
11	1.426681e+12	303602c25c0cb0c0000000001	8433815129956	NA	NA	NA	NA	NA	308133
12	1.426685e+12	303602c25c0d648000000002	8433815137142	NA	NA	NA	NA	NA	795940
13	1.426686e+12	303602c25c0cba40000000001	8433815130334	NA	NA	NA	NA	NA	569719
14	1.426687e+12	303602c25c0cba40000000001	8433815130334	NA	NA	NA	NA	NA	308783
15	1.426688e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	495977
16	1.426690e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	310364
17	1.426691e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	497419
18	1.426693e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	308106
19	1.426696e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	453094
20	1.426698e+12	303602c25c0cb200000000001	8433815130006	NA	NA	NA	NA	NA	337729
21	1.426699e+12	303602c25c0d0c0000000002	8433815133632	NA	NA	NA	NA	NA	307889

Taula 4.2 Dataframe resultant de sessiondata

El següent pas és la unió dels tres *dataframes* resultants a cada un dels tres arxius i poder obtenir així un *dataframe* global, tal com es pot apreciar a la *Figura 4.3*.

Per a poder realitzar la unió es fa ús de la funció *merge()*. Aquesta funció ens permet unir dos *dataframes* a partir d'un element comú. S'ha procedit a unir el *dataframe* de *SeasonData* amb el de *Maestros*, a partir de l'element comú entre els dos: el codi de barres.

Operem de la mateixa manera que a la primera unió per concatenar el resultat obtingut amb el *dataframe* de *Skucodes*. Únicament s'ha de tenir en compte que la variable del codi de barres en aquesta taula es diu *skuid* i l'hem de canviar tal com li hem dit en les altres dues taules i obtenir finalment tota la informació en una sola taula.

La taula resultant, anomenada *mastertable* no deixa de ser res més que la concatenació de tots els camps, però sabem que totes les dades estan informades (veure *Capítol 7*).

mastertable		4725 obs. of 28 variables
scanned.code	: Factor w/ 897 levels	"8433815130013",...: 1 1 2 2 2 2 2 2 2 ...
scanned.timestamp	: num	1.43e+12 1.43e+12 1.43e+12 1.43e+12 1.43e+12 ...
scanned.epc	: Factor w/ 1138 levels	"303602c25c0cb24000000001",...: 1 1 2 2 2 2 2 2 2 ...
requested.timestamp	: num	NA NA NA NA NA NA NA NA NA ...
requested.status	: chr	NA NA NA NA ...
requested.code	: chr	NA NA NA NA ...
requested.requestID	: num	NA NA NA NA NA NA NA NA NA ...
requested.type	: chr	NA NA NA NA ...
sessiontime	: num	7389962 342701 306708 357695 312879 ...
viewed.timestamp	: num	1.43e+12 1.43e+12 1.43e+12 NA NA ...
viewed.name	: Factor w/ 325 levels	"Short quilted windbreaker",...: 1 1 5 NA NA NA NA NA NA ...
viewed.code	: Factor w/ 302 levels	"1538120307","1538100310",...: 1 1 5 NA NA NA NA NA NA ...
sessionstart	: num	1.43e+12 1.43e+12 1.43e+12 1.43e+12 1.43e+12 ...
sessionend	: num	1.43e+12 1.43e+12 1.43e+12 1.43e+12 1.43e+12 ...
skuidsize	: Factor w/ 86982 levels	"000000000001TU",...: 38493 38493 80708 80708 80708 80708 80...
price	: Factor w/ 563 levels	"0,01","1,00",...: 170 170 170 170 170 170 170 170 170 ...
percentage	: int	100 100 8 42 13 8 8 37 8 37 ...
composition	: Factor w/ 205 levels	"Acero Inox","Acetato",...: 151 151 136 66 151 136 136 9 136...
size.x	: Factor w/ 68 levels	"00","1","10",...: 59 59 63 63 63 63 63 63 63 ...
colour.name	: Factor w/ 76 levels	"ROSA PASTEL",...: 8 8 67 67 67 67 67 67 67 ...
colour.code	: Factor w/ 49 levels	"#FF00FF","#E1D8D1",...: 8 8 1 1 1 1 1 1 1 ...
rvcolor	: Factor w/ 77 levels	"30","06","01",...: 8 8 68 68 68 68 68 68 68 ...
productcolor	: num	1.54e+11 1.54e+11 6.10e+11 6.10e+11 6.10e+11 ...
family	: chr	"CHAQUETAS" "CHAQUETAS" "BOLSOS" "BOLSOS" ...
name_es	: chr	"Plumas corto" "Plumas corto" "Tote tejido canvas estampado pitÁ*n" "Tote tejido...
description_es	: chr	"Plumas de mujer corto, cuello tira y bolsillos delantero" "Plumas de muj...
size.y	: chr	"L" "L" "TU" "TU" ...
productid	: num	1.54e+09 1.54e+09 6.10e+09 6.10e+09 6.10e+09 ...

Figura 4.3 Mastertable de les dades

Capítol 5: Resultats de l'anàlisi

L'objectiu del projecte és extreure conclusions sobre la informació obtinguda. Necessitem, doncs, un mitjà per a entendre com s'organitzen les dades, què està succeint i com afecta això a la botiga per a la seva campanya de moda. És adient l'ús del suport gràfic per a visualitzar més clarament els resultats de l'anàlisi.

L'anàlisi consta de quatre seccions, on ens centrem a entendre el comportament de les persones en entrar als emprovadors i com això podria ajudar de cara a potenciar les vendes de la companyia.

5.1 Anàlisi segons els productes vistos

Una primera aproximació cap a entendre el que està passant dins els emprovadors és fer un anàlisi de quins són els productes que més i que menys es proven. Això ajuda a entendre quins són els productes estrella de la botiga i quins són els que més s'ha de potenciar. Aquestes dades podrien ajudar en la distribució dels productes dins de la tenda.

Segons les dades obtingudes, els productes més vistos (provats a dins dels emprovadors) es poden veure a la *Figura 5.1*:

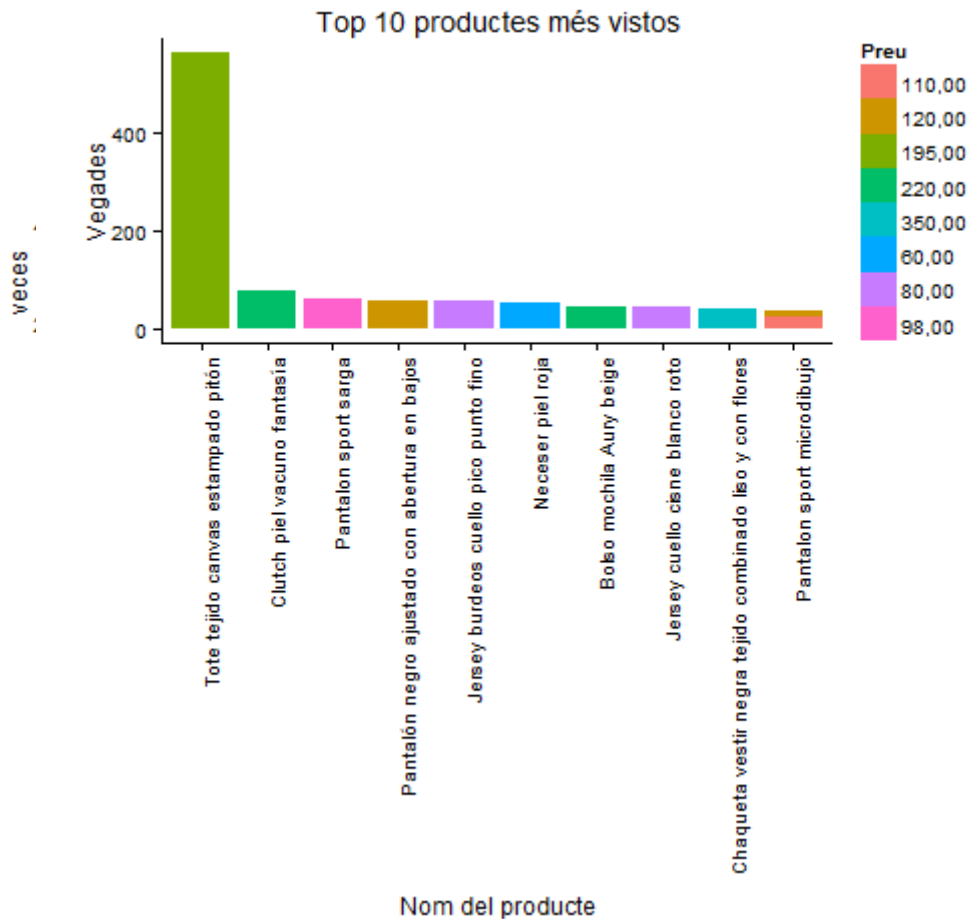


Figura 5.1 Nom i preu dels productes més vistos

D'aquestes dues taules podem apreciar un aspecte molt interessant: hi ha un producte que destaca per sobre de tots. Aquest producte correspon a una bossa i és una dada curiosa: en un estudi realitzat a la mateixa botiga fa un any on s'arribava a la conclusió que les bosses eren el producte menys venut de la botiga.

Podria ser que el resultat del primer producte fos fruit del *testing* de la tenda, i que a la majoria de persones que entressin a l'emprovador els donessin bosses per si volien provar-se-les. Ja que aquest primer producte ha tingut 564 interaccions i el segon 78. Així que si descartem aquest primer element tindríem el segon com a nou producte més vist. Però dona la casualitat que el segon també és una bossa. Per tant sembla que han fet un bon treball de màrqueting i han trobat la manera de fer que la gent es comenci a interessar per les bosses, fent que un 4 dels 10 productes estrella de la botiga siguin bosses.

Si realitzem l'operació inversa i filtrem els productes que menys es prova la gent, obtenim els següents resultats vistos a la *Figura 5.2*.

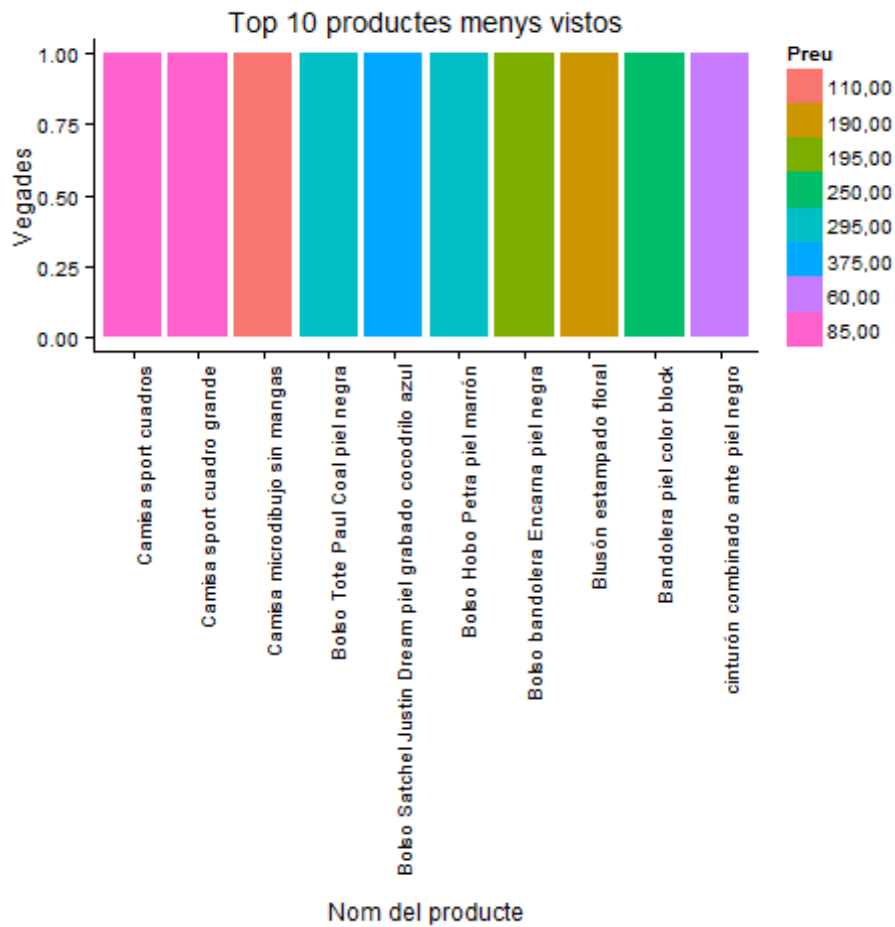


Figura 5.2 Nom i preu de productes menys vistos

Podem veure que aquesta gràfica ens retorna un aspecte prou regular, pel que podria ser una bona idea fer-ne una llista amb els 10 primers productes, tal com a la *Figura 5.3*:

1	Bolso Hobo Petra piel marrón
2	Bolso Tote Paul Coal piel negra
3	Camisa sport cuadros
4	Blusón estampado floral
5	Bolso Satchel Justin Dream piel grabado cocodrilo azul
6	Camisa sport cuadro grande
7	Camisa microdibujo sin mangas
8	cinturón combinado ante piel negro
9	Bandolera piel color block
10	Bolso bandolera Encarna piel negra

Figura 5.3 Llista dels productes menys vistos

La primera dada que es pot extreure és que tots els productes de la tenda són escanejats als emproadors. Pot ser curiós, però donar-se'n compte no és trivial.

Ha semblat adient informar del preu amb una nova variable, representada a través del color a la gràfica, així es pot saber ràpidament quins són els preus dels productes que s'està provant la gent d'una manera molt fàcil i intuïtiva.

Tornem a observar que succeeix una cosa interessant amb les dades: 5 dels 10 productes menys vistos són bosses. Sembla doncs que, observant aquestes dades i contrastant amb les anteriors sobre productes més vistos, podem confirmar que hi ha hagut una certa millora en el tema de les bosses, tot i que encara sembla que es resisteix una mica.

Com a tret rellevant, dir que 6 dels 10 productes que apareixen, són complements, siguin bosses o cinturons. És evident que hi ha una probabilitat més elevada que la gent no s'emprovi complements i es decanti cap a peces de roba per a parts del cos més grans. Aquest tipus de peces no necessiten la intimitat que un emproador proporciona, i podria ser el cas que molta gent se les prova als miralls de la tenda.

5.2 Anàlisi segons les famílies de productes vistos

Els resultats visualitzats ens donen una idea precisa de l'estat de les interaccions dins la botiga, però sovint no interessa quedar-se amb productes en concret, sinó que es necessita saber l'estat de tota una família a l'inventari. Amb aquestes dades, es té una visió més global del que succeeix i es poden pensar estratègies de màrqueting diferents si s'observa que el resultat no és el desitjat.

Podem veure per exemple, les dades de les famílies més vistes a la *Figura 5.4*:

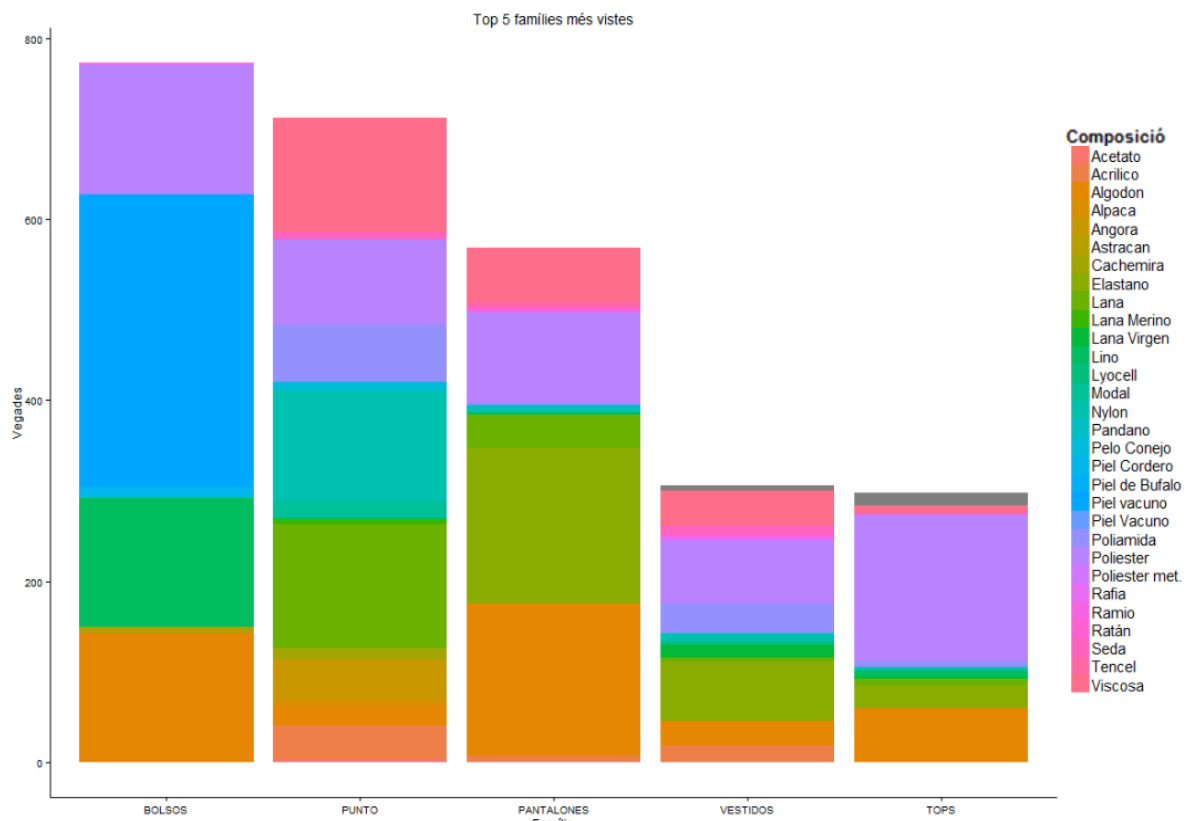


Figura 5.4 Famílies i composició de productes més vistos

Ara podem estar segurs dels resultats obtinguts als apartats previs: les bosses són el producte més provat, probablement degut al més que possible *testing* que s'hagi pogut produir.

La següent família serien els productes fets de punt. Això té certa lògica, ja que a la temporada de gener i febrer, les vendes de jerseis de punt van augmentar i es podia veure com cada botiga en feia un màrqueting bastant important. Podem fixar-nos també en què el material més emprat seria la llana, per tant estem davant d'aquest cas, jerseis gruixuts d'hivern.

La resta de peces de roba poden encaixar en el perfil del primer trimestre: pantalons i vestits. Roba que cobreix una superfície gran. Podem veure que els materials serien el cotó, l'elastà i el polièster. Sembla que els resultats tenen sentit.

L'última família, encara que posi *tops*, en aquest cas no es refereix a cap peça de vestir femenina, sinó a camises. Ja que segons les dades proporcionades aquesta família engloba a algunes altres més (veure Capítol 7).

Si fem el mateix però amb les famílies amb menys visites als emproadors, obtenim la *Figura 5.5*:

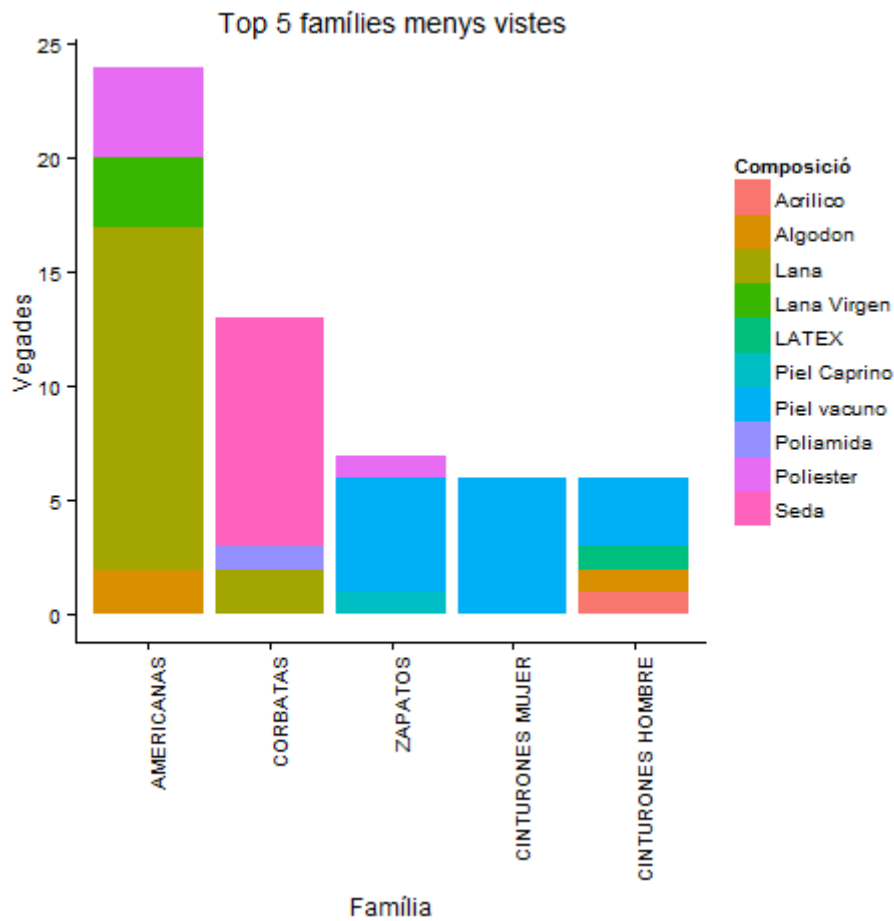


Figura 5.5 Famílies i composició de productes menys vistos

Amb aquestes dades, si pensem en el cas de les bosses, observem com no hi apareixen en les 5 famílies menys vistes, i com la millora és real.

Com a tret significatiu es pot veure que 4 de les 5 famílies amb menys visites són accessoris, i tal com succeïa amb els productes menys vistos, la família menys vista són els cinturons (tant de dona com d'home).

Per tant sembla que seria necessària una acció per a impulsar les vendes d'accessoris, tal com es va fer amb les bosses. Es podria portar a dependents als emproadors amb certs productes en funció del que el client estigui visualitzant i veure si encaixaria.

Les americanes, si observem el material, són de llana. Encara que sigui hivern, no sol ser el teixit principal respecte a la compra d'aquest tipus de producte. Tot i que dels 24 cops que s'han provat americanes, 15 eren de llana.

5.3 Anàlisi segons les composicions vistes

Una bona idea és visualitzar quina composició de productes s'està provant més l'usuari que entra a la botiga, aquesta informació es pot observar a partir dels resultats de la *Figura 5.6*:

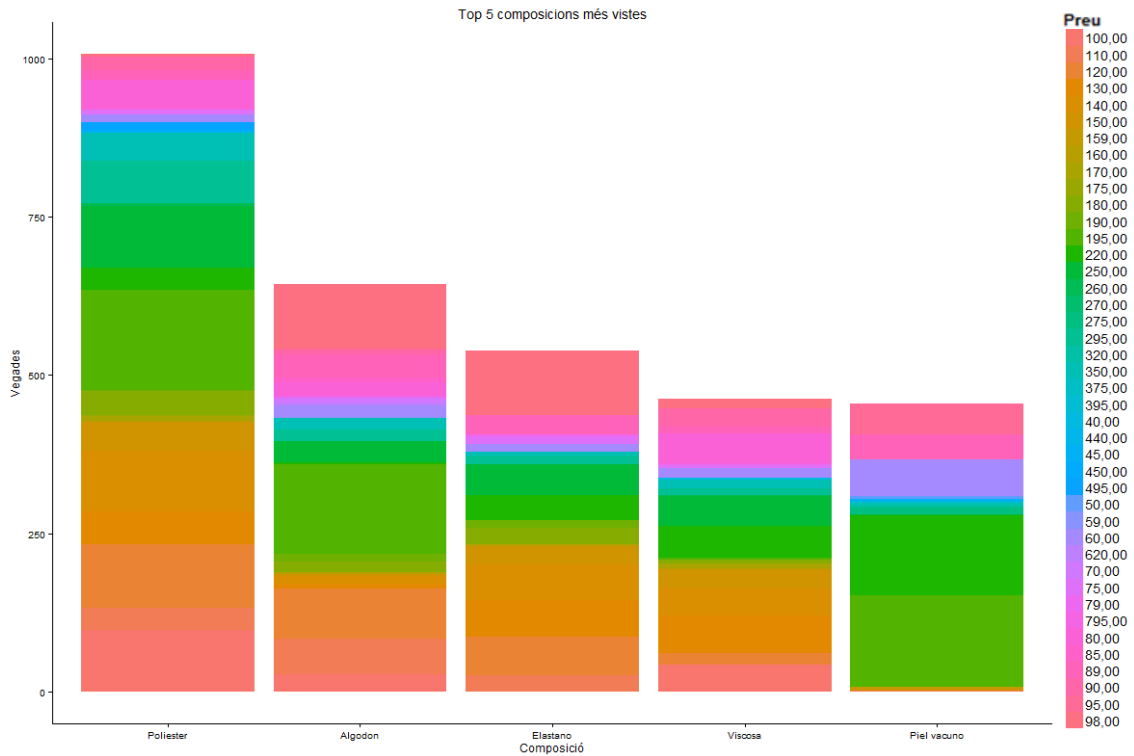


Figura 5.6 Composició i preu de productes més vistos

Podem observar com el polièster és la composició més provada pels clients, seguida del cotó i l'elastà. Amb la *Figura 5.4* podríem fer-nos una idea, però ara podem estar segurs. Aquest gràfic pot servir d'indicador per veure la tendència de teixits que està més de moda entre els clients i redistribuir les peces a la tenda de manera adient.

D'altra banda podem pensar a la inversa: visualitzar quines són les composicions menys vistes pels clients per tal de retirar *stock* o simplement tenir una visió més exacta de la situació i actuar conseqüentment. Això es pot veure a la *Figura 5.7*:

```
> mintop5comp
[1] "Acero Inox" "acrilico" "ACRILICO" "Acrílico" "ACRÍLICO"
```

Figura 5.7 Llista de la composició dels productes menys vistos

El motiu pel qual s'ha considerat fer una llista és que a les composicions menys vistes, hi ha moltes amb índex zero, cosa que ens indica que cap peça de roba ha estat provada si estava feta amb un teixit concret. Si es volgués representar la informació en un gràfic no es veuria res, i no queda del tot adient.

El primer que es pot observar de la llista és que tot i que retorna cinc composicions, quatre d'elles són acrílic. Això ens dona a entendre que les dades proporcionades no són correctes o que necessiten un filtratge especial. (*veure Capítol 7*).

5.4 Anàlisi comparatiu mensual

La millor manera d'obtenir una idea global de la situació, i del que està succeint, és analitzar per intervals de temps. En aquest cas, al tenir dades referents als tres primers mesos, s'ha cregut convenient crear tres gràfics. Cadascú per mostrar els productes que més es prova la gent a cada més, i obtenir una idea del que més li agrada a la gent, del que continuen emprovant-se, o del que perd més importància.

Per tant, podem afirmar que és la *Figura 5.8* dóna molta informació rellevant sobre el cicle de vida dels productes o les tendències de la gent, entre d'altres.

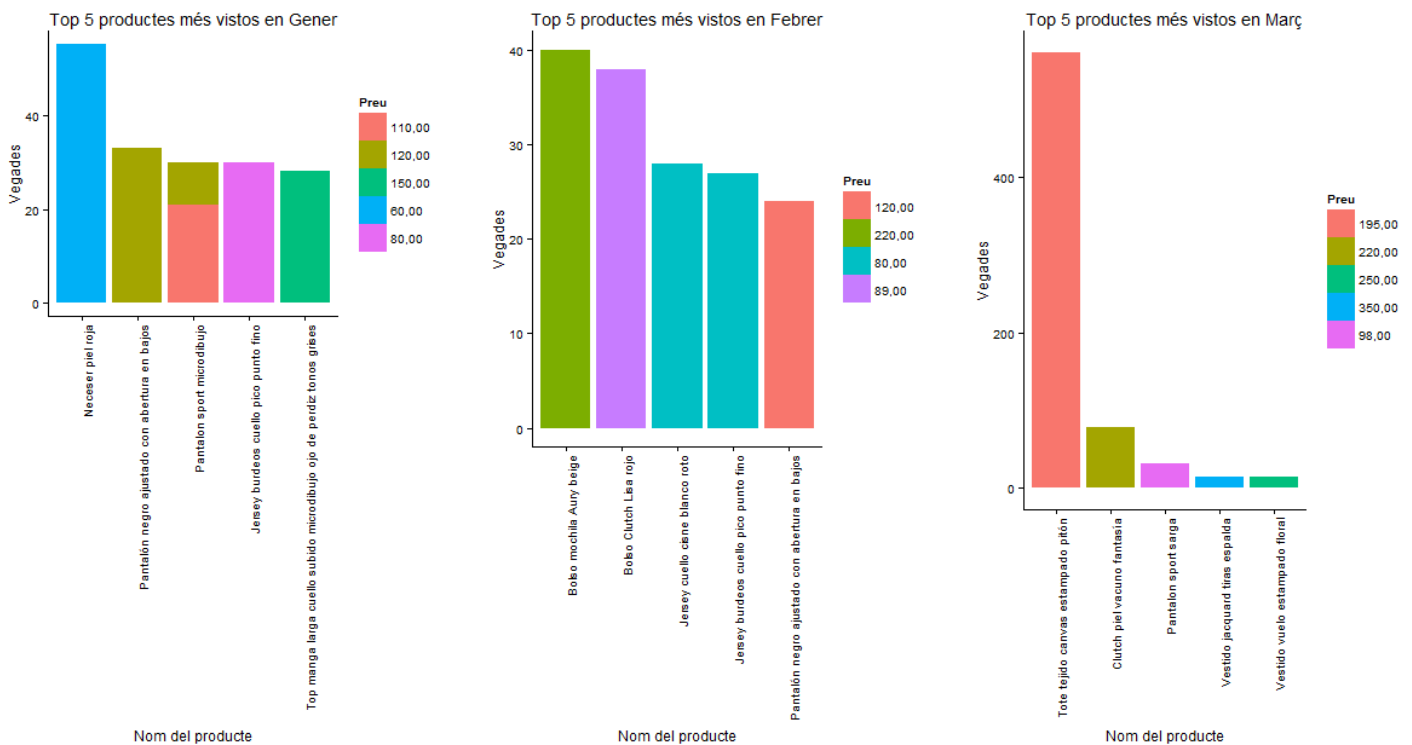


Figura 5.8 Comparativa mensual dels productes més vistos

Tal com podem observar, quasi cap producte dins del top es repeteix per a l'altre més: els pantalons que al gener ocupen la segona posició, al febrer són a la cinquena; i el jersei *burdeos* es manté a la quarta posició del gener al febrer.

Una altra cosa interessant és que es pot visualitzar un canvi de tendència: al gener i al febrer, la gent es prova més pantalons o jerséis. En canvi cap al març, si no tenim en compte aquesta bossa que probablement sigui resultat del *testing* intern, veiem com els vestits comencen a obtenir més visites cap als emproadors.

Com apunt específic, el top 1 a cada més és un complement, i en concret una bossa. Inclús es pot observar com també és el top 2 tant al febrer com al març. Per tant, ens serveix com a indicador del que ja hem comentat de la tendència en augment dels clients emprovant-se bosses.

Capítol 6: Conclusions

6.1 Resolució i conclusions

Durant l'elaboració del projecte, hem pogut observar que les botigues de roba continuen tenint molts problemes a l'hora de saber amb certesa els productes dels quals disposen.

Aquest fet és degut a un inventari ineficient: s'ha pogut veure com tenen diferents fitxers per a categoritzar els productes, on en alguns hi ha redundància en la informació que hi contenen, i com a vegades no estan preparats per a unir-se en un de sol.

S'ha pogut comprovar que, mitjançant la tecnologia RFID és molt fàcil portar un control ordenat i fiable de tota la informació de l'inventari, i a temps real.

Segons les dades extretes a l'anàlisi hem pogut observar com a la campanya de primavera-hivern de 2015:

1. Els jerséis de llana han començat a emergir com a producte a tenir en compte.
2. Les bosses han remuntat respecte a l'any passat i ara es proven moltes més.
3. Es necessita emprendre un pla d'acció amb els complements, són el producte menys provat, sobretot en homes. Una campanya de màrqueting enfocada o promocions d'aquest tipus d'articles podria ser una bona alternativa per augmentar la demanda.
4. A partir de març, és quan més es nota que la gent comença a fer el canvi entre la roba d'hivern i de primavera.

6.2 Treballs futurs

Sempre que un acaba un projecte, sorgeixen idees per a millorar-lo i fer-lo més atractiu. Una idea podria ser crear un sistema d'aprenentatge automàtic, on una persona hagués introduït una mostra de sessions i interaccions de campanyes de moda passades, i com a mostra de *testing* les últimes campanyes més actuals. Així es podria fer un software intel·ligent, capaç de detectar tendències de moda i patrons de conducta de les persones, observant que solen emprovar-se i que és el que menys vegades entra a l'emprorador.

Es disposaria d'una eina molt potent capaç d'elaborar prediccions i comparar-les sobre més prediccions, per a obtenir un refinament de les dades.

També es podria canviar la manera en la qual està implementada la solució. Es podria elaborar un sistema autònom, que realitzés connexions cap al *cloud* on s'emmagatzema la informació en un temps definit, i que descarregués tot per a extreure'n un anàlisi detallat. I finalment, fer una migració cap a plataformes mòbils. Seria interessant tenir una aplicació que pogués controlar *reports* proveïts pel nostre sistema i els envies cap als nostres *smartphones*, on l'usuari pogués controlar la situació d'una manera ubiqua.

I sense abandonar el tema dels *smartphones*, es podria potenciar la integració de l'ús de les xarxes socials als emproadors, oferir bonificacions, punts o descomptes exclusius si estant dins de l'emproador realitzen alguna acció social, com compartir una fotografia de la roba que s'estan provant, donar *like* a *Facebook* o *follow* a *Twitter*, etc.

Veient la situació actual, la integració de les xarxes socials seria una característica interessant a millorar en el futur, ja que aporta informació molt útil, potent i de manera gratuïta, sobre el client, que si s'explota degudament, genera informació de precisió, molt acurada.

També cabria la possibilitat d'integrar la nostra aplicació amb Shinnyapps, una plataforma que ens proveeix d'una sèrie de serveis (*PaaS*), i que ens permet fer un *deploy* de qualsevol aplicació que sigui creada amb qualsevol *IDE* de R, com és el cas d'RStudio. També permetria poder veure desenvolupaments d'altres persones ja que es pot compartir el disseny, d'una manera fàcil i intuïtiva.

Capítol 7: Problemàtica

7.1 Principals problemes i requeriments derivats

Al llarg del projecte, s'hi han trobat certs problemes que s'han hagut de solucionar, altres s'han pogut tolerar. Tot en funció d'una sèrie de requeriments en que s'ha hagut de treballar. Els principals problemes han estat:

1. La informació de l'inventari de la que es disposa no és prou correcta o no està correctament tractada i és possible trobar certs errors, com per exemple:
 - No notificar valors buits amb algun caràcter concret (s'ha utilitzat NA per aquest fi).
 - Tenir famílies de productes incloses dins d'una família, com és el cas de la família TOPS que fa referència a samarretes, camises, tops de dona o bruses entre d'altres. O la família PUNTO, ja que ho associem a teixits i no a roba. Això pot portar a tenir productes duplicats: qualsevol família de productes feta de punt: jerseis, accessoris, jaquetes, americanes, etc.
 - Composicions de teixits duplicats, com per exemple: Algodón, algodón, ALGODON o Algodon. Al tractar-les diferent, pot comportar problemes i el millor seria usar una composició identificadora per teixit.
2. En realitzar la unió de dos *dataframes* amb un valor que actuï de clau de referència entre els dos, es necessita que tots els camps estiguin informats, per tant els camps buits necessiten contenir alguna informació. Si no es realitza aquesta transformació al fer la unió obtindrem un *dataframe* format per la unió del primer i el segon, però totes les dades del segon *dataframe* apareixeran buides.
3. Les dades del *cloud* no són les dades efectives al 100%. Com és el cas de la informació que apareix al fitxer de *maestros.txt*. No hi ha cap manera d'accedir a

informació com el preu, la composició o el percentatge del teixit en les dades del *cloud* i es depèn de la informació que la tenda ens proporciona.

4. La integració dels tres fitxers no és trivial. Com es comenta al problema 3, el fitxer de *maestros.txt* no s'extreu de l'*AdvanCloud*, per tant no té capçaleres. S'ha de crear una fila amb aquesta informació on la variable del codi de barres sigui la clau de referència.
5. La idea de barrejar dades en un XLS amb dades en format JSON. Observant els problemes que podia donar, o la realització de tractaments d'informació específics, es creu convenient tenir els documents en format XLS o JSON únicament. De l'altra manera, serà necessari un pre-processament d'aquesta informació per a ser adaptada amb la resta.

Bibliografía

- [1] Roussos, G. "Enabling RFID in retail". (March 2006). University of London
- [2] Roussos, G., Moussouri, T. "Privacy, security and trust in ubiquitous commerce". (2004).
- [3] Huber, K., Houck, L., Vinogradov, I. "RFID within the Retail Environment".
- [4] Keifenheim, B. "An Ethical Evaluation of Radio Frequency Identification". (March.8, 2007). University of Minnesota
- [5] Electronic Privacy Information Center. "Radio Frequency Identification (RFID) Systems". (February.27, 2007-March.3, 2007).
- [6] Privacy Rights Clearinghouse. " RFID Position Statement of Consumer Privacy and Civil Liberties Organizations " (November.20, 2003-March.3, 2007).
- [7] Thomas, D. " Marks & Spencer begins the UK's largest in-store trial of RFID tagging technology". <http://www.computerweekly.com/feature/Marks-amp-Spencer-begins-the-UKs-largest-in-store-trial-of-RFID-tagging-technology>. *Últim accés: 12/06/2014*
- [8] Riera, S. "Quién controla el RFID de Inditex" <http://www.modaes.es/equipamiento/equipamiento/20141001/quien-controla-el-rfid-de-inditex.html>. *Últim accés: 12/06/2015*
- [9] Martins, G. "6 reasons why you should start using R" <http://gmdias.com/2015/06/24/6-reasons-why-you-should-start-using-r/> *Últim accés: 15/06/2015*

Annexos

Annex 1. Project Charter

1. Introducció i motivacions

Aquest document descriu d'una manera breu i concisa els diferents aspectes a tenir en compte perquè l'anàlisi i interpretació d'interaccions d'una botiga amb RFID sigui satisfactori.

D' ençà que a la universitat vaig començar a conèixer el món de l'*Internet of Things* que em va agradar. El control que es pot tenir sobre objectes connectats a la xarxa, i sobretot, l'abast que pot oferir aquest tipus de tecnologia van esdevenir fonamentals perquè em plantejés la realització d'aquest treball. En particular, amb l'ús de la tecnologia RFID, a causa del potencial que té i que ofereix, semblava molt atractiu poder integrar-la en el projecte.

2. Projecte i descripció del producte

Es procedirà a través d'un sistema d'antenes RFID a la captació de dades rellevants de les interaccions persona-producte més comuns i habituals que succeeixen dins d'una botiga (ja sigui entrar-hi, emprovar-se un producte, pagar o marxar). Posteriorment es realitzarà l'anàlisi dels casos ja esmentats elaborant un estudi amb qui la botiga podria extreure conclusions útils a través de dades numèriques, demogràfiques o relacionades amb els productes. En l'estudi es preveu endinsar-se en màrqueting, localització i posicionament de productes dins la botiga, campanyes de vacances, relacions entre els productes emprovats i comprats, maximitzar les compres, etc.

3. Abast

a. Objectius

La primera fase consistirà en l'anàlisi de la informació: en aquest primer objectiu serà necessari la neteja d'errors (establint un llindar de tolerància) i després el posterior emmagatzematge de la informació, per a dotar-la d'un control mínim (BB.DD, full de càlcul Excel, etc.).

La segona fase donarà lloc a l'estudi i interpretació: extreure relacions, patrons entre les dades, amb la fi de generar un conjunt d'hipòtesis o fets rellevants que facin que el projecte tingui èxit.

Les tecnologies usades seran: RFID per a la part de l'extracció d'informació, Java per a la part de l'anàlisi (per la simplicitat, nivell de domini per la part de la implementació, i la flexibilitat que dona) i Weka o similars a l'hora de fer una interpretació analítica dels resultats (per la bona usabilitat que ofereix, la ràpida adaptació i el nivell de domini).

b. Entregues

Aplicació d'anàlisi de la informació: capaç de, a partir d'un volum d'informació correctament estructurada extreure'n d'altra visualment més agradable e intel·ligible per a una persona.

Anàlisi quantitatiu de l'ús de l'aplicació: un anàlisi de com important és el factor de la tolerància en el sistema i com depenen directament d'ell les conclusions extretes.

c. Limitacions

El projecte està pensat per extreure informació i conclusions a partir d'un conjunt de dades recollides donat. És a dir, no té "coneixement propi" ni aprenentatge automàtic. No pot fer "refeed" i poder recomanar a partir de resultats obtinguts.

No es guardarà en principi un històric de la informació per motius de privacitat.

El sistema serà de tipus *web-app* a causa de motius de programari, però podria ser interessant tenir accés a través d'*smartphones* i *tablets* que facilitessin la interacció de l'usuari.

4. Duració

a. Estimació temporal del projecte

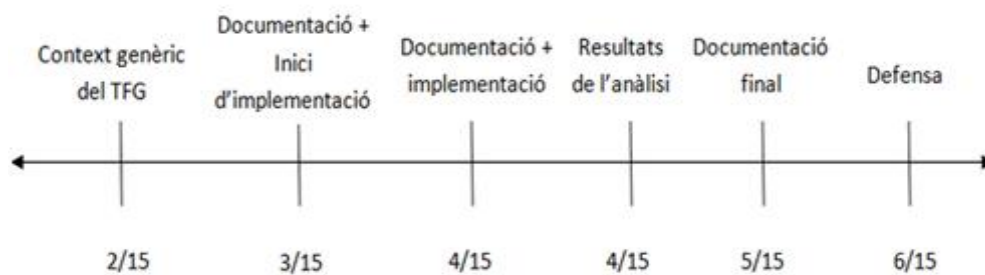


Figura A1.1 Diagrama de l'estimació temporal del projecte

5. Requeriments imposats

El projecte assumeix que la divulgació dels resultats queda taxativament prohibida i la informació serà d'ús propi i amb caràcter privat.

6. Riscos

El principal risc del projecte és la neteja de la informació. Les dades generades pel sistema RFID i la configuració del propi ve donat i no es té accés a modificacions, cosa que pot induir a l'error de certes dades. Això no pot esdevenir cap desviació en l'objectiu del projecte.

7. Organització del projecte

a. Rols i responsabilitats

Nom	Rol en el projecte
Anna Carreras	Tutor principal
David Nettleton	Tutor de suport
David Madrid	Implementador

Figura A1.2 Taula de rols i responsabilitats del projecte

b. *Stakeholders* del projecte

Stakeholder	Rol en el projecte
Anna Carreras	Tutor principal
David Nettleton	Tutor de suport
David Madrid	Autor i implementador
La botiga de roba	Extern i generador d'informació
Rafael Pous (Keonn Technologies)	Extern

Figura A1.3 Taula de stakeholders del projecte

Annex 2. Codi del projecte

```
#cargamos las librerias y paquetes necesarios
library(rjson)
library(RJSONIO)
library(ggplot2)
library(directlabels)
install.packages("devtools")
library(devtools)
install_github("kassambara", "easyGgplot2")
library(easyGgplot2)

seasondata = read.csv("seasonData.csv") #cargar datos del csv de sesiones de Keonn
sessiondata = seasondata[["sessiondata"]] #guardar los datos del dataframe de sessiondata
JSONsessiondata <- mapply(fromJSON, as.character(seasondata$sessiondata)) #crear una variable lista

#creamos los que sera la estructura de la tabla, insertando los valores que tendra y convertimos a NA los campos vacios
JSONsessiondata <- lapply(JSONsessiondata, function(list.item){
  data.frame(scanned.timestamp = ifelse(is.null(list.item$scanned[[1]]$timestamp), NA, list.item$scanned[[1]]$timestamp),
    scanned.epc = ifelse(is.null(list.item$scanned[[1]]$epc), NA, list.item$scanned[[1]]$epc),
    scanned.code = ifelse(is.null(list.item$scanned[[1]]$code), NA, list.item$scanned[[1]]$code),
    requested.timestamp = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$timestamp, NA),
    #requested.assistant = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$assistant, NA),
    requested.status = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$status, NA),
    requested.code = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$code, NA),
    requested.requestID = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$requestid, NA),
    requested.type = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$type, NA),
    #requested.resptime = ifelse(length(list.item$requested) > 0, list.item$requested[[1]]$resptime, NA),
    sessiontime = ifelse(is.null(list.item$sessiontime), NA, list.item$sessiontime),
    viewed.timestamp = ifelse(length(list.item$viewed) > 0, list.item$viewed[[1]]$timestamp, NA),
    viewed.name = ifelse(length(list.item$viewed) > 0, list.item$viewed[[1]]$name, NA),
    viewed.code = ifelse(length(list.item$viewed) > 0, list.item$viewed[[1]]$code, NA),
    sessionstart = ifelse(is.null(list.item$sessionstart), NA, list.item$sessionstart),
    sessionend = ifelse(is.null(list.item$sessionend), NA, list.item$sessionend))
})

df.sessiondata <- do.call(rbind, JSONsessiondata) #los valores numericos se convertiran
#df.sessiondata <- format(do.call(rbind, JSONsessiondata), scientific = FALSE) #respetamos la notacion
rownames(df.sessiondata) <- seq(1, nrow(df.sessiondata))

#####

skudata = read.csv("skuCodes.csv")
productdata = skudata[["productdata"]]

JSONproductdata <- mapply(fromJSON, as.character(skudata$productdata))
JSONproductdata <- lapply(JSONproductdata, function(list.item){
  data.frame(#colour.name = ifelse(is.null(list.item$colour$name), NA, list.item$colour$name),
    #colour.code = ifelse(is.null(list.item$colour$code), NA, list.item$colour$code),
    colour.name = ifelse(is.null(list.item$colour[[1]]), NA, list.item$colour[[1]]),
    colour.code = ifelse(is.null(list.item$colour[[2]]), NA, list.item$colour[[2]]),
    rvcolor = ifelse(is.null(list.item$rvcolor), NA, list.item$rvcolor),
    productcolor = ifelse(is.null(list.item$productcolor), NA, list.item$productcolor),
    family = ifelse(is.null(list.item$family), NA, list.item$family),
    name_es = ifelse(is.null(list.item$name_es), NA, list.item$name_es),
    description_es = ifelse(is.null(list.item$description_es), NA, list.item$description_es),
    size = ifelse(is.null(list.item$size), NA, list.item$size)
  )
})

df.skucodes <- do.call(rbind, JSONproductdata)
#df.skucodes <- format(do.call(rbind, JSONproductdata), scientific = FALSE)

# add a new column
df.skucodes$productid <- skudata$productid
df.skucodes$skuid <- skudata$skuid
rownames(df.skucodes) <- seq(1, nrow(df.skucodes))
#df.skucodes <- format(df.skucodes,scientific = FALSE)

#####

maestros <- read.table("MAESTROS_KEONN.2015.04.17.txt", header = TRUE, sep=";", na.strings = c("", "NA")) #rellena lo vacio a NA

#####

#unimos los dos dataframes de sessiondata y maestros a traves de la columna scanned.code que comparten
#al usar all.x = true hacemos que cojamos la informacion de maestros y la cargemos a sessiondata
mastertable <- merge(df.sessiondata, maestros, by = "scanned.code", all.x = TRUE)

df.skuAuxiliar <- df.skucodes #creamos una tabla auxiliar
colnames(df.skuAuxiliar)[10] <- "scanned.code" #cambiamos el nombre de skuid por scanned.code

#idem al primer merge, pero ahora con el dataframe de skuAuxiliar
mastertable <- merge(mastertable, df.skuAuxiliar, by = "scanned.code", all.x = TRUE)

#####
```

```

#TOP10 PRODUCTOS MAS VISTOS

# calcular frecuencias
tab <- table(mastertable$scanned.code)
#ordenar (de menor a mayor por frecuencia)
tab_s <- sort(tab)
#cojer las 10 primeras
#se usa tail ya que las ultimas son las que contienen las frecuencias maximas, con head para las minimas
maxtop10code <- tail(names(tab_s), 10)
#subconjunto del dataframe, filtrado por el top 10 de frecuencias maximas
master_s <- subset(mastertable, scanned.code %in% maxtop10code)
#ordenar los niveles de los factor
master_s$scanned.code <- factor(master_s$scanned.code, levels = rev(maxtop10code))
#plot
#cargamos un archivo de salida de tipo png
png('img/CodisMesVistos.png')
#creamos una imagen de tipo ggplot, escogiendo los datos de los ejes, labels, leyenda o tipo de grafico
ggplot(master_s, aes(x = scanned.code, fill = as.factor(price))) + geom_bar() + theme_classic() +
labs(title = "Top 10 productos más vistos") + labs(x = "Codi de producte") + labs(y = "Vegades") +
theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
dev.off()

#cambiamos la codificación de los datos para verlos correctamente
Encoding(Encoding(mastertable$name_es) <- "UTF-8")
# calcular frecuencias
tab1 <- table(mastertable$name_es)
#ordenar
tab1_s <- sort(tab1)
#cojer las 10 primeras
maxtop10name <- tail(names(tab1_s), 10)
#subconjunto del dataframe
master1_s <- subset(mastertable, name_es %in% maxtop10name)
#ordenar los niveles de los factor
master1_s$name_es <- factor(master1_s$name_es, levels = rev(maxtop10name))
#plot
png('img/ProductesMesVistos.png')
ggplot(master1_s, aes(x = name_es, fill = as.factor(price))) + geom_bar() + theme_classic() +
labs(title = "Top 10 productes més vistos") + labs(x = "Nom del producte") + labs(y = "vegades") +
theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
dev.off()

#TOP10 PRODUCTOS MENOS VISTOS

# calcular frecuencias
tab2 <- table(mastertable$scanned.code)
#ordenar
tab2_s <- sort(tab2)
#cojer las 10 primeras
mintop10code <- head(names(tab2_s), 10)
#subconjunto del dataframe
master2_s <- subset(mastertable, scanned.code %in% mintop10code)
#ordenar los niveles de los factor
master2_s$scanned.code <- factor(master2_s$scanned.code, levels = rev(mintop10code))
#plot
png('img/CodisMenysVistos.png')
ggplot(master2_s, aes(x = scanned.code, fill = as.factor(price))) + geom_bar() + theme_classic() +
labs(title = "Top 10 productes menys vistos") + labs(x = "Codi del producte") + labs(y = "vegades") +
theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
dev.off()

#cambiamos la codificación de los datos para verlos correctamente
Encoding(Encoding(mastertable$name_es) <- "UTF-8")
# calcular frecuencias
tab3 <- table(mastertable$name_es)
#ordenar
tab3_s <- sort(tab3)
#cojer las 10 primeras
mintop10name <- head(names(tab3_s), 10)
#subconjunto del dataframe
master3_s <- subset(mastertable, name_es %in% mintop10name)
#ordenar los niveles de los factor
master3_s$name_es <- factor(master3_s$name_es, levels = rev(mintop10name))
#plot
png('img/ProductesMenysVistos.png')
ggplot(master3_s, aes(x = name_es, fill = as.factor(price))) + geom_bar() + theme_classic() +
labs(title = "Top 10 productes menys vistos") + labs(x = "Nom del producte") + labs(y = "Vegades") +
theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
dev.off()

```

```

#TOP5 FAMILIAS MAS VISTAS

# calcular frecuencias
tab4 <- table(mastertable$family)
#ordenar
tab4_s <- sort(tab4)
#cojer las 5 primeras
maxtop5fam <- tail(names(tab4_s), 5)
#subconjunto del dataframe
master4_s <- subset(mastertable, family %in% maxtop5fam)
#ordenar los niveles de los factor
master4_s$family <- factor(master4_s$family, levels = rev(maxtop5fam))
#plot
#lo adaptamos ya que la leyenda no cabe
png('img/FamiliesMesVistes.png', width = 1200, height = 800)
ggplot(master4_s, aes(x = family, fill = as.factor(composition))) + geom_bar() + theme_classic() +
labs(title = "Top 5 families més vistes") + labs(x = "Familia") + labs(y = "Vegades") +
labs(fill = "Composició") + theme(legend.text=element_text(size=14))
dev.off()

# calcular frecuencias
tab5 <- table(mastertable$family)
#ordenar
tab5_s <- sort(tab5)
#cojer las 5 primeras
maxtop5fam <- head(names(tab5_s), 5)
#subconjunto del dataframe
master5_s <- subset(mastertable, family %in% maxtop5fam)
#ordenar los niveles de los factor
master5_s$family <- factor(master5_s$family, levels = rev(maxtop5fam))
#plot
png('img/FamiliesMenysVistes.png')
ggplot(master5_s, aes(x = family, fill = as.factor(composition))) + geom_bar() + theme_classic() +
labs(title = "Top 5 families menys vistes") + labs(x = "Familia") + labs(y = "vegades") +
theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Composició")
dev.off()

#TOP5 COMPOSICIONES MAS VISTAS

# calcular frecuencias
tab6 <- table(mastertable$composition)
#ordenar
tab6_s <- sort(tab6)
#cojer las 5 primeras
maxtop5comp <- tail(names(tab6_s), 5)
#subconjunto del dataframe
master6_s <- subset(mastertable, composition %in% maxtop5comp)
#ordenar los niveles de los factor
master6_s$composition <- factor(master6_s$composition, levels = rev(maxtop5comp))
#plot
#lo adaptamos ya que la leyenda no cabe
png('img/ComposicionsMesVistes.png', width = 1200, height = 800)
ggplot(master6_s, aes(x = composition, fill = as.factor(price))) + geom_bar() + theme_classic() +
labs(title = "Top 5 composicions més vistes") + labs(x = "Composició") + labs(y = "Vegades") +
labs(fill = "Preu") + theme(legend.text=element_text(size=14))
dev.off()

#TOP5 COMPOSICIONES MENOS VISTAS

# calcular frecuencias
tab7 <- table(mastertable$composition)
#ordenar
tab7_s <- sort(tab7)
#cojer las 5 primeras
mintop5comp <- head(names(tab7_s), 5)

```

```

#COMPARATIVA PRODUCTOS MÁS VISTOS

#cogemos timestamp y cortamos la string hasta 10, obteniendo el epoch
scanned.month <- substr(format(mastertable$scanned.timestamp, scientific = FALSE),1,10)
#inicializamos nuestro vector
month = NULL
#extraemos de cada epoch su mes, y lo guardamos en el vector
for(i in 1:length(scanned.month)){
  month[i] <- format(as.POSIXct(as.double(scanned.month[i]), origin="1970-01-01"), "%B")
}
#duplicamos nuestra mastertable
MT <- mastertable
#añadimos la columna months
MT$months <- month

#extraemos toda la información de la mastertable del mes de enero
MTenero <- subset(MT, months == "enero")
#cambiamos la codificación de los datos para verlos correctamente
Encoding(Encoding(MTenero$name_es) <- "UTF-8") |
# calcular frecuencias
tab8 <- table(MTenero$name_es)
#ordenar
tab8_s <- sort(tab8)
#cojer los 5 primeros
maxtop5jan <- tail(names(tab8_s), 5)
#subconjunto del dataframe
master8_s <- subset(MTenero, name_es %in% maxtop5jan)
#ordenar los niveles de los factor
master8_s$name_es <- factor(master8_s$name_es, levels = rev(maxtop5jan))

MTfebrero <- subset(MT, months == "febrero")
#cambiamos la codificación de los datos para verlos correctamente
Encoding(Encoding(MTfebrero$name_es) <- "UTF-8")
# calcular frecuencias
tab9 <- table(MTfebrero$name_es)
#ordenar
tab9_s <- sort(tab9)
#cojer los 5 primeros
maxtop5feb <- tail(names(tab9_s), 5)
#subconjunto del dataframe
master9_s <- subset(MTfebrero, name_es %in% maxtop5feb)
#ordenar los niveles de los factor
master9_s$name_es <- factor(master9_s$name_es, levels = rev(maxtop5feb))

MTmarzo <- subset(MT, months == "marzo")
#cambiamos la codificación de los datos para verlos correctamente
Encoding(Encoding(MTmarzo$name_es) <- "UTF-8")
# calcular frecuencias
tab10 <- table(MTmarzo$name_es)
#ordenar
tab10_s <- sort(tab10)
#cojer los 5 primeros
maxtop5mar <- tail(names(tab10_s), 5)
#subconjunto del dataframe
master10_s <- subset(MTmarzo, name_es %in% maxtop5mar)
#ordenar los niveles de los factor
master10_s$name_es <- factor(master10_s$name_es, levels = rev(maxtop5mar))

#plots
enero <- ggplot(master8_s, aes(x = name_es, fill = as.factor(price))) + geom_bar() + theme_classic() +
  labs(title = "Top 5 productos más vistos en Enero") + labs(x = "Nom del producto") + labs(y = "Vegades") +
  theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
febrero <- ggplot(master9_s, aes(x = name_es, fill = as.factor(price))) + geom_bar() + theme_classic() +
  labs(title = "Top 5 productos más vistos en Febrero") + labs(x = "Nom del producto") + labs(y = "Vegades") +
  theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")
marzo <- ggplot(master10_s, aes(x = name_es, fill = as.factor(price))) + geom_bar() + theme_classic() +
  labs(title = "Top 5 productos más vistos en Marzo") + labs(x = "Nom del producto") + labs(y = "Vegades") +
  theme(axis.text.x = element_text(hjust=1, angle=90)) + labs(fill = "Preu")

png('img/ComparativaMensual.png')
#usamos un multiplot para visualizar los tres meses en una matriz 1x3
ggplot2::multiplot(enero, febrero, marzo, cols = 3)
dev.off()

```

