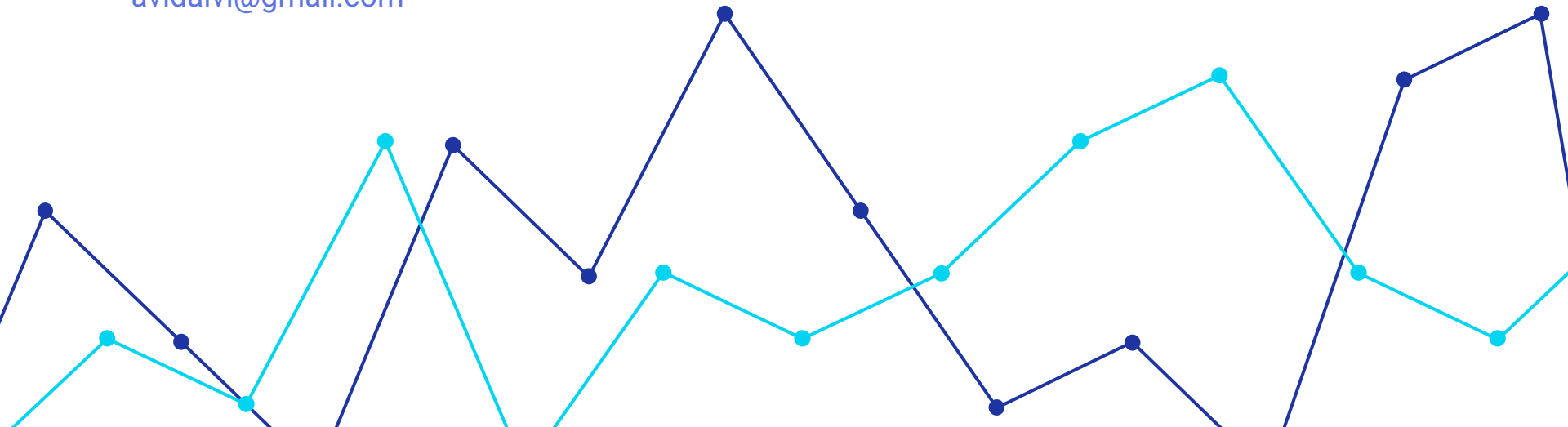


Exemplos de uso de R na empresa

Antonio Vidal Vidal
Senior Data Scientist at Nextail
avidalvi@gmail.com



Sobre mín

Enxeñeiro de Telecomunicacións
Executive MBA

Universidade de Vigo

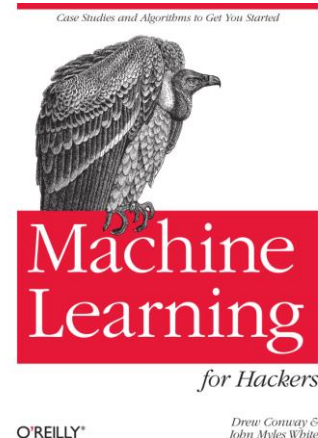
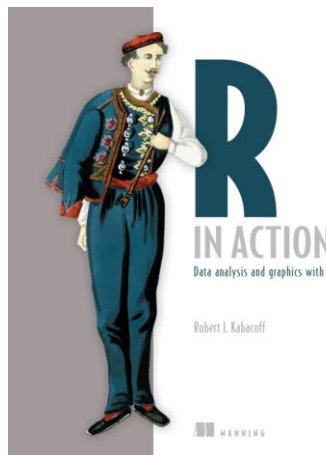
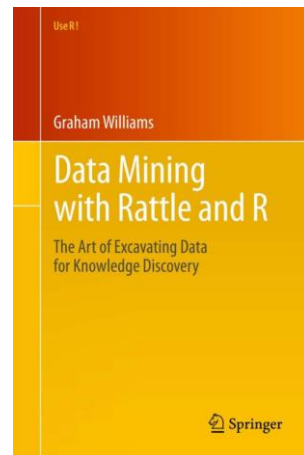
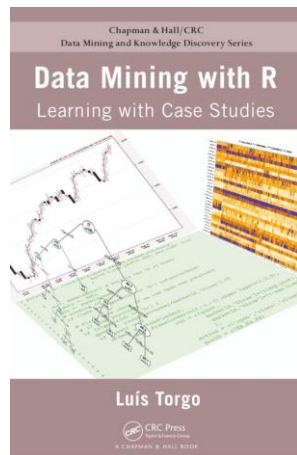
IESIDE
BUSINESS INSTITUTE

INSTITUTO DE
EDUCACIÓN SUPERIOR
INTERCONTINENTAL
DE LA EMPRESA

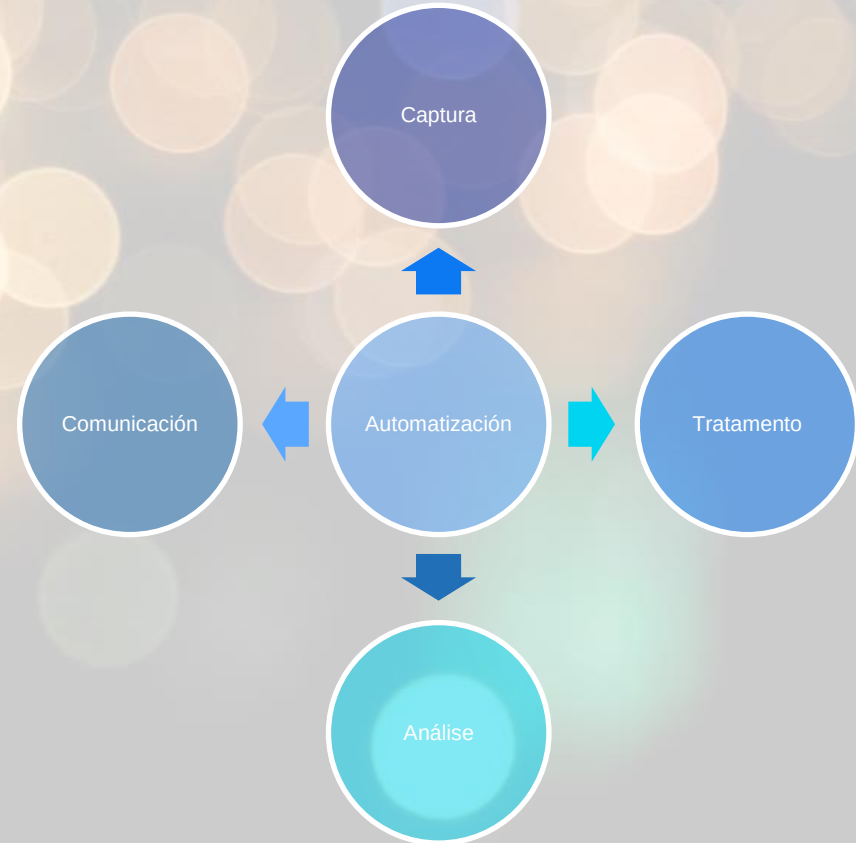
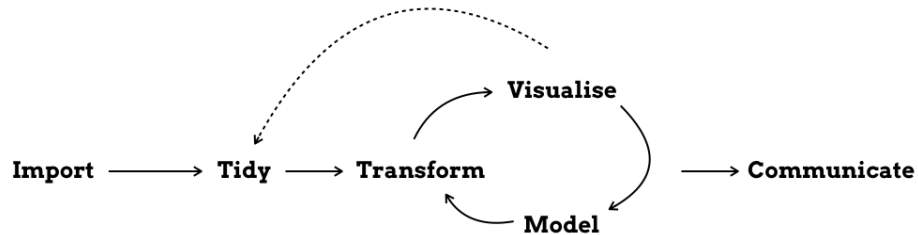


Evolución de R

2.11	2010-04-22	Support for Windows 64-bit systems.
2.12.2	2011-02-25	Last version to support Windows 2000 ^[59]
2.13	2011-04-14	Adding a new compiler function that allows speeding up functions by converting them to byte-code.
2.14	2011-10-31	Added mandatory namespaces for packages. Added a new parallel package.
2.15	2012-03-30	New load balancing functions. Improved serialisation speed for long vectors.
3.0.0	2013-04-03	Support for numeric index values 2^{31} and larger on 64-bit systems.
3.3.3	2017-03-06	Last version to support Microsoft Windows XP.
3.4.0	2017-04-21	Just-in-time compilation (JIT) of functions and loops to byte-code enabled by default.
3.5.0	2018-04-23	Packages byte-compiled on installation by default. Compact internal representation of integer sequences. Added a new serialisation format to support compact internal representations.
3.6.0	2019-04-26	Improved sampling from a discrete uniform distribution, which was noticeably non-uniform on large populations. ^[60] New serialisation format supported since 3.5.0 becomes the default.
4.0.0	2020-04-24	R now uses a <code>stringsAsFactors = FALSE</code> default, and hence by default no longer converts strings to factors in calls to <code>data.frame()</code> and <code>read.table()</code> . Reference counting is used for tracking object sharing, which reduces the need for copying objects. New syntax for raw string constants.
4.1.0	2021-05-18	Introduced <code> ></code> as the pipe operator for base R syntax (similar to the <code>%>%</code> operator of the <code>magrittr</code> package) and the anonymous function shortcut syntax <code>\(x) x+1</code> .



Exemplos de uso de R na empresa



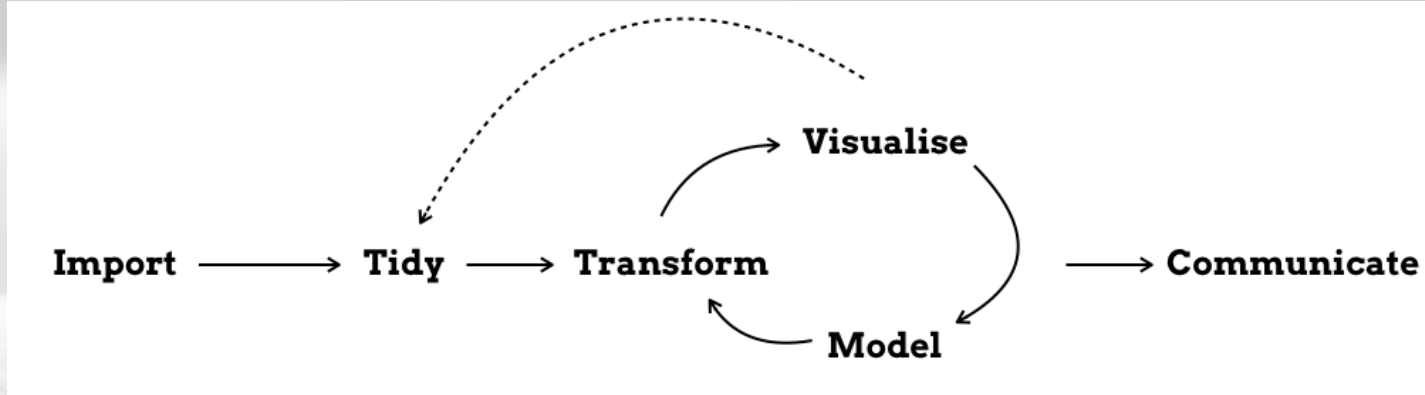
Captura – recollida de datos

Captura de datos	Paquetes
BBDD tradicionais	DBI, RMySQL, RMariaDB, RPostgreSQL, ROracle, odbc, dbplyr
Outras BBDDs	influxdb
Google Cloud	bigrquery
Salesforce	salesforcer

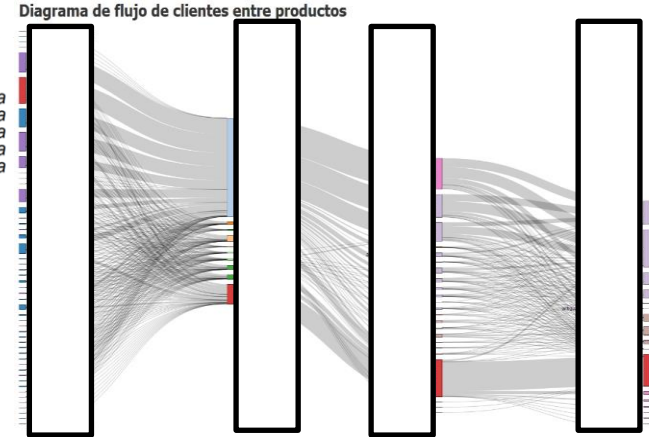
Captura – recollida de datos

Captura de datos	Paquetes
Captura de datos web - scraping	rvest, rSelenium, splashr
Navegación automática	rSelenium
APIs	httr, jsonlite
PDFs	pdftools, tabulizer

Tratamento

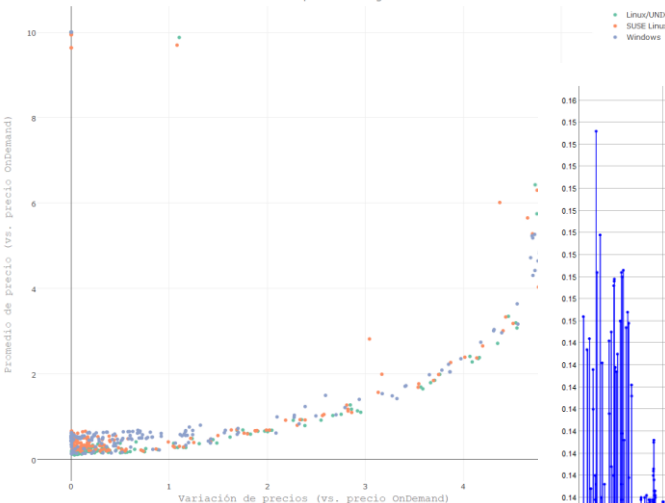


Captura de datos	Paquetes
Importación de datos	readr, readxl, openxlsx, data.table, jsonlite
Limpeza e estandarización de datos	tidyverse, tidyselect, janitor, lubridate, clock, zoo, tsibble, stringr, forcats
EDA	Desc, skimr, Hmisc, ggplot2, plotly, dygraphs

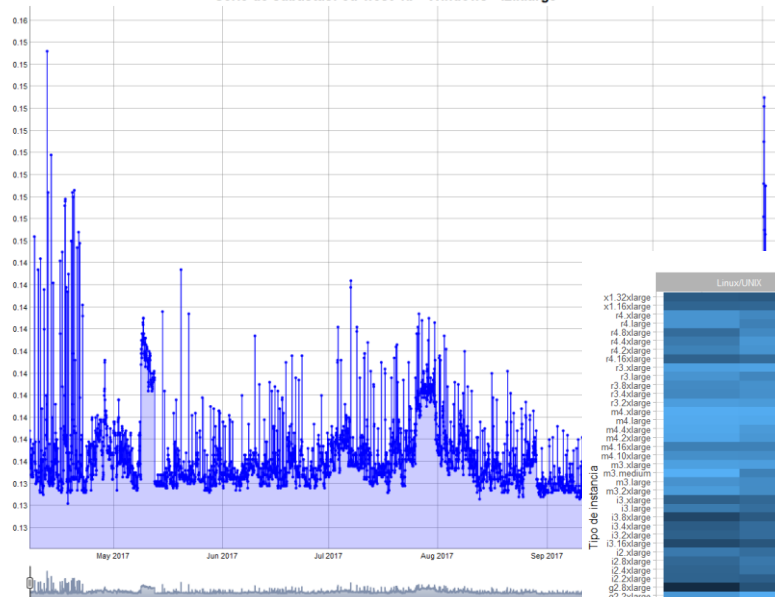


EDA – Precios das instancias spot de AWS

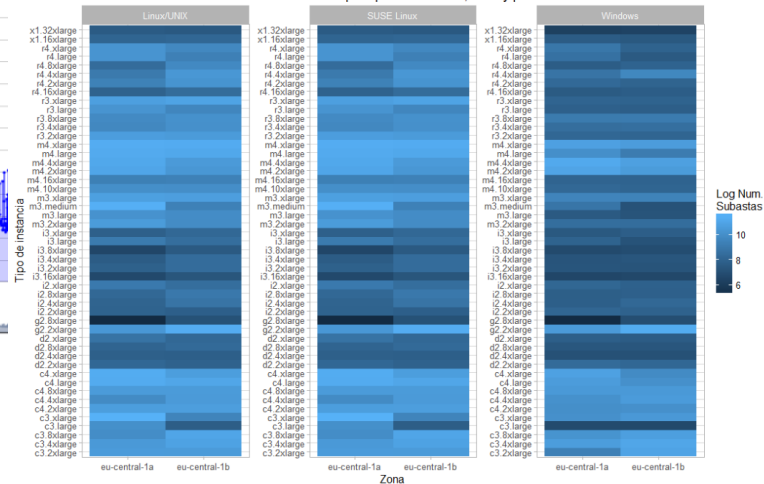
Varlaciones de precio de la region eu-west-1



Serie de subastas: eu-west-1b - Windows - i2.xlarge

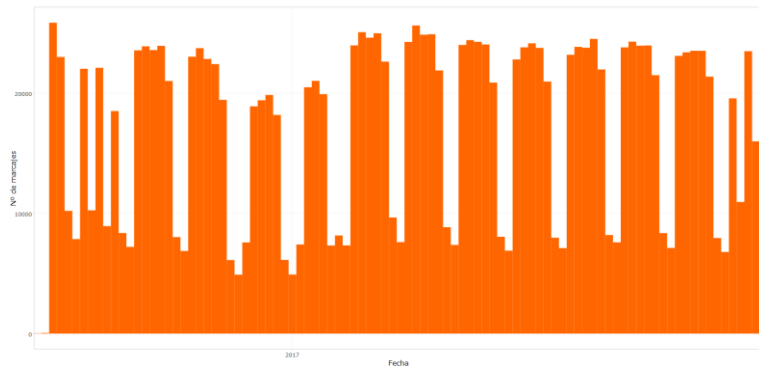


Número de subastas por tipo de instancia, zona y producto

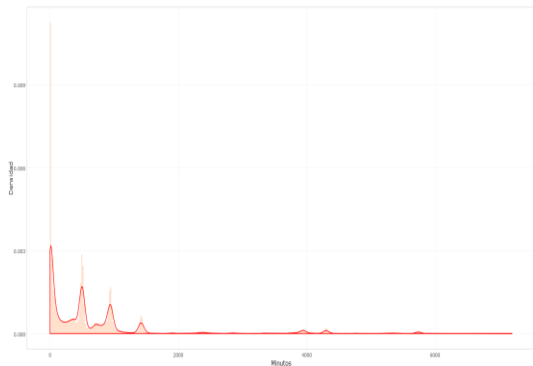


EDA – Identificación de patróns en marcaxes

Distribución marcaxes



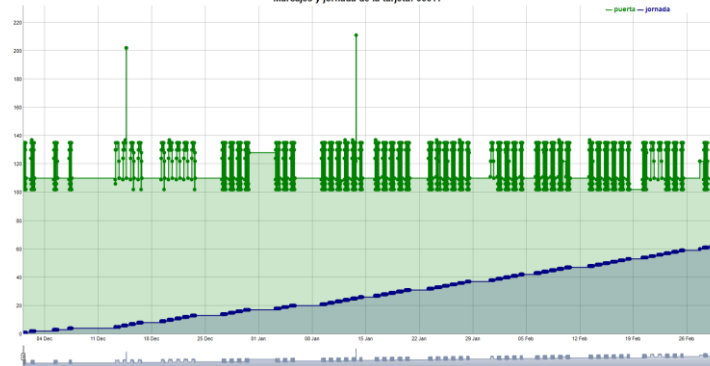
Minutos entre pases de puerta por marcaje



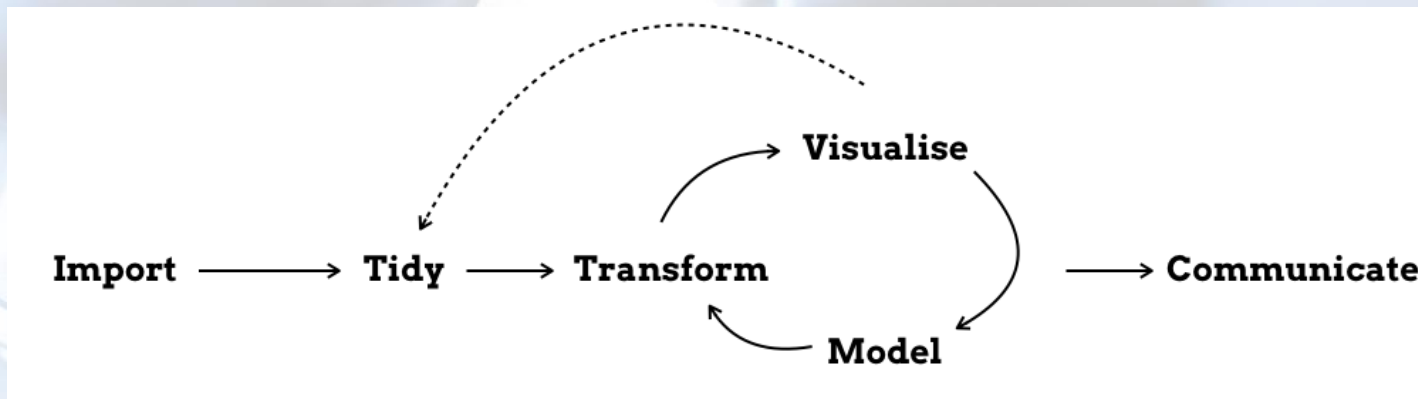
Transiciones entre puertas



Marcaxes y jornada de la tarjeta: 66611

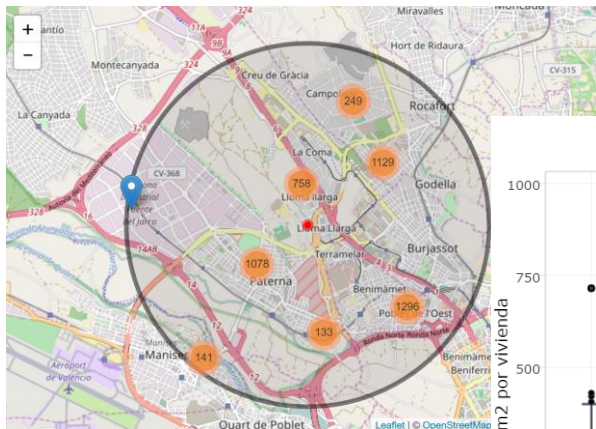


Análise

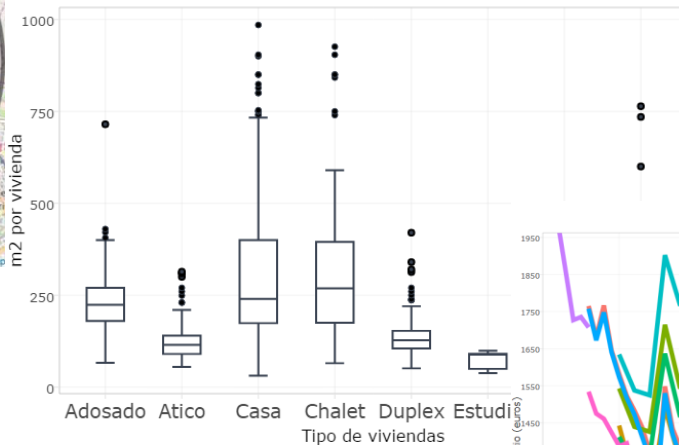


Análise	Paquetes
Creación de modelos	Frameworks: caret, mlr, tidymodels, h2o Texto: tidytext, text2vec, topicmodels Series temporais: forecast, prophet, TSstudio, timetk Deep learning: keras, tensorflow

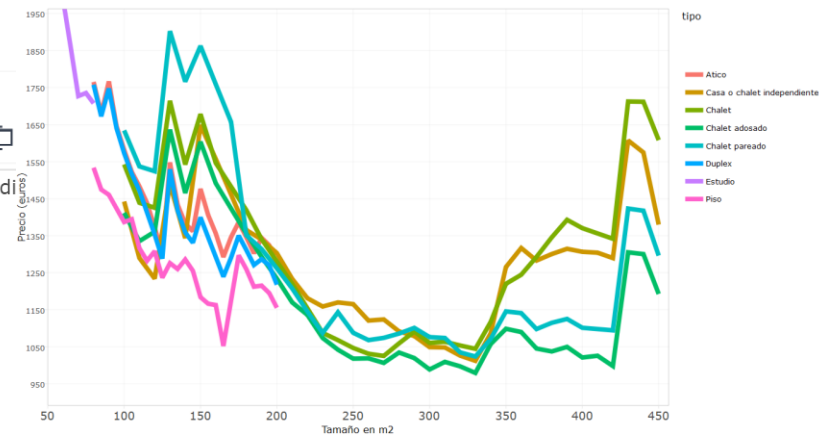
Análise – Precio das vivendas dunha zona



Tamaño de las viviendas



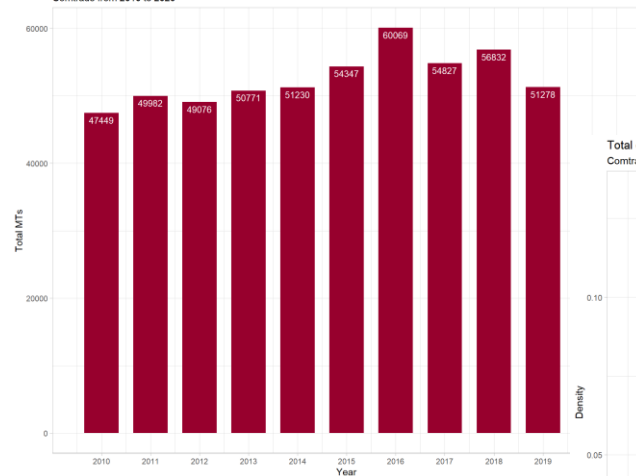
Precio euros/m2



Análise – Estimación do volume de mercado

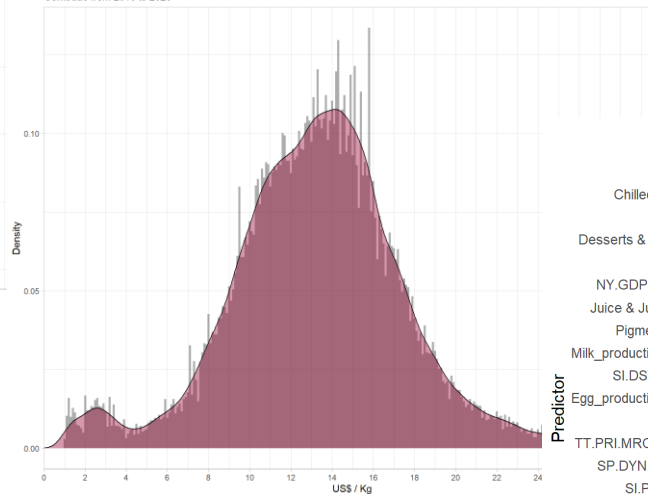
Pectin

Comtrade from 2010 to 2020



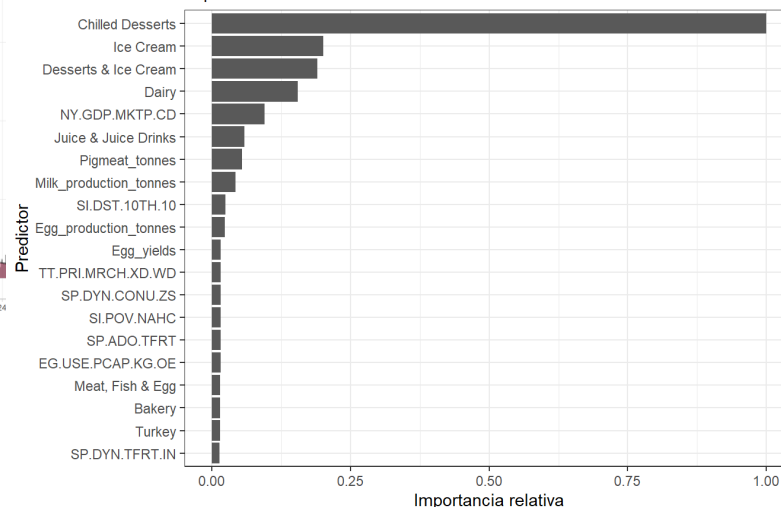
Total distribution of prices

Comtrade from 2010 to 2020

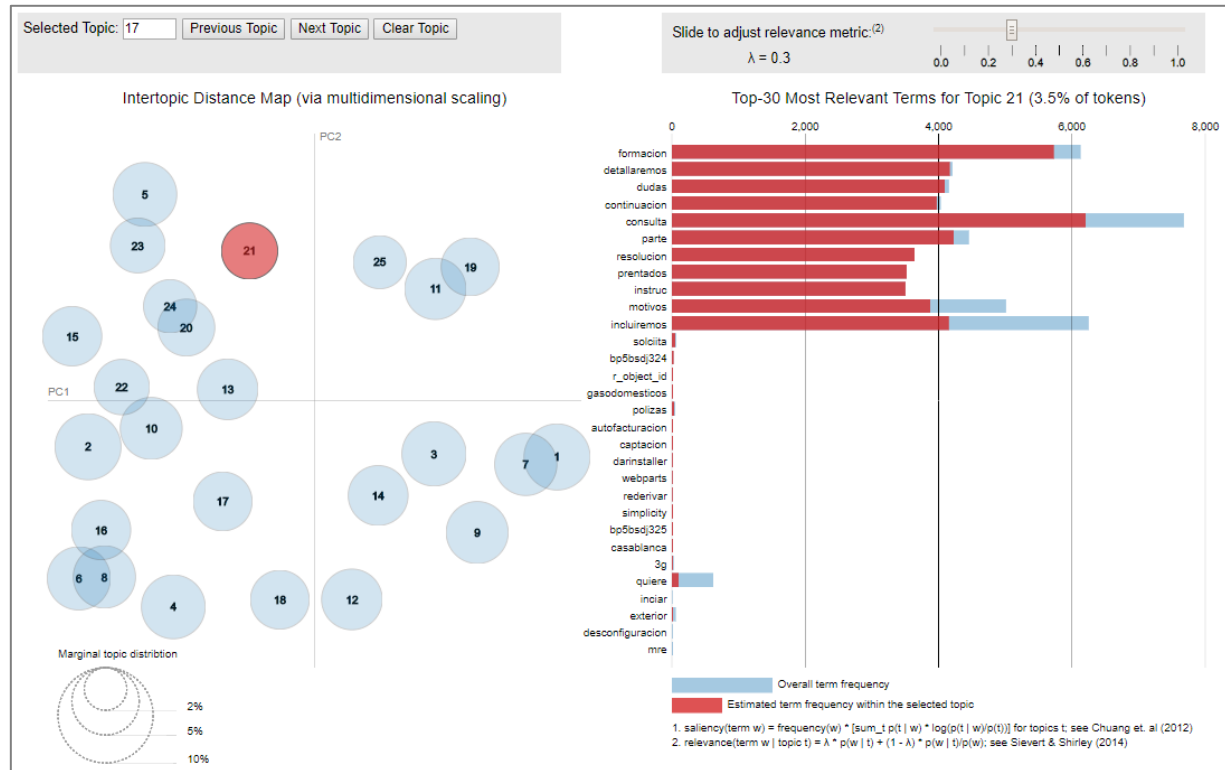
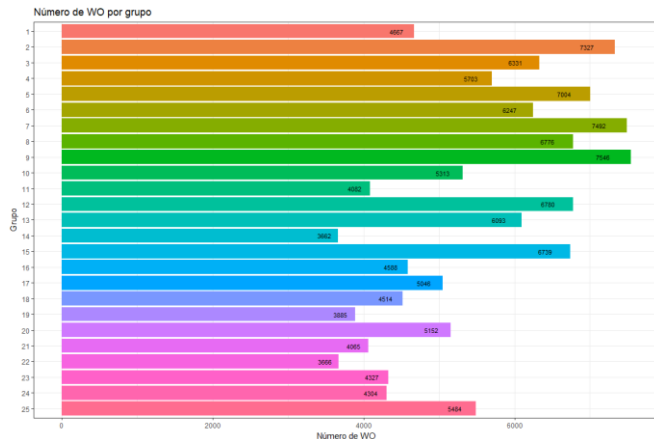


Importancia de los predictores en el modelo GBM

Importancia en base a la reducción del error cuadrático medio

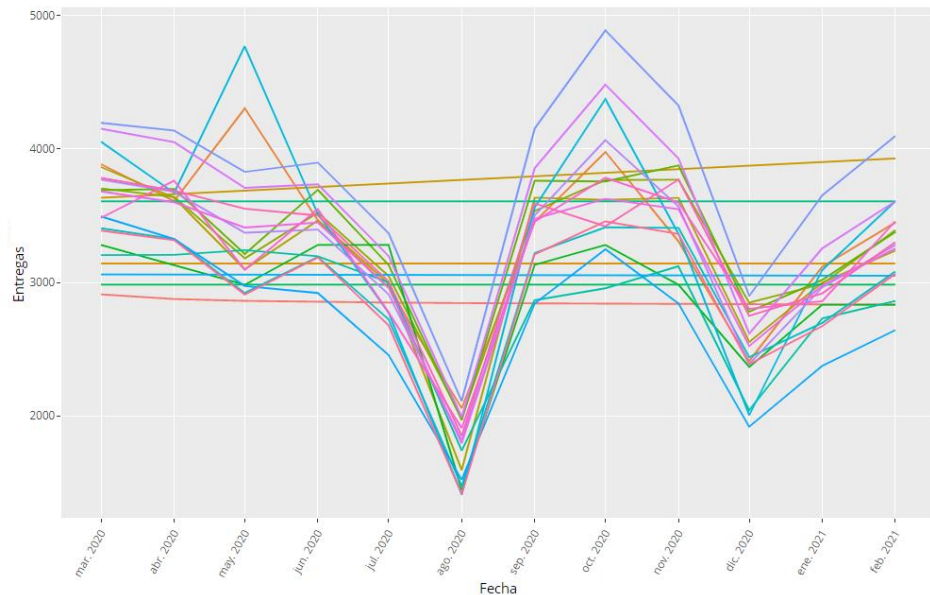


Análise – Ordes de traballo

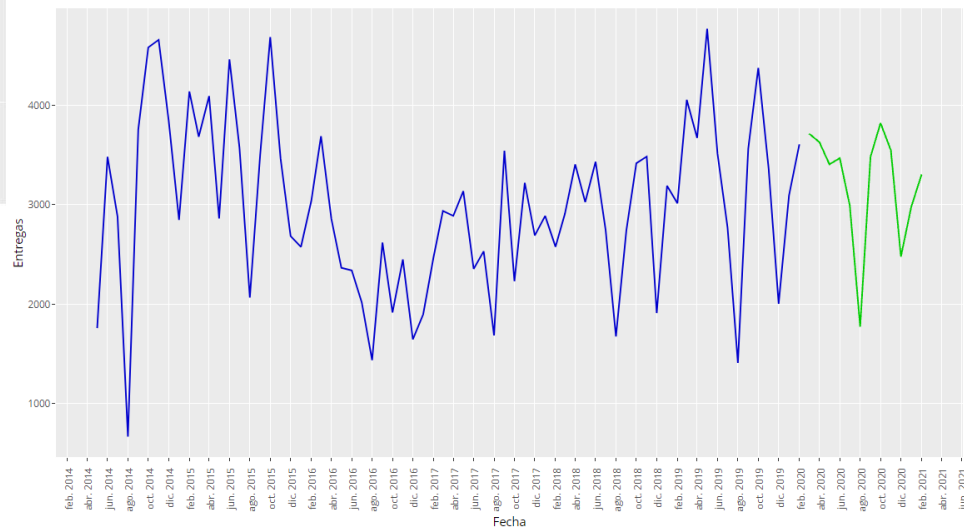


Análise – Predicción de series temporais

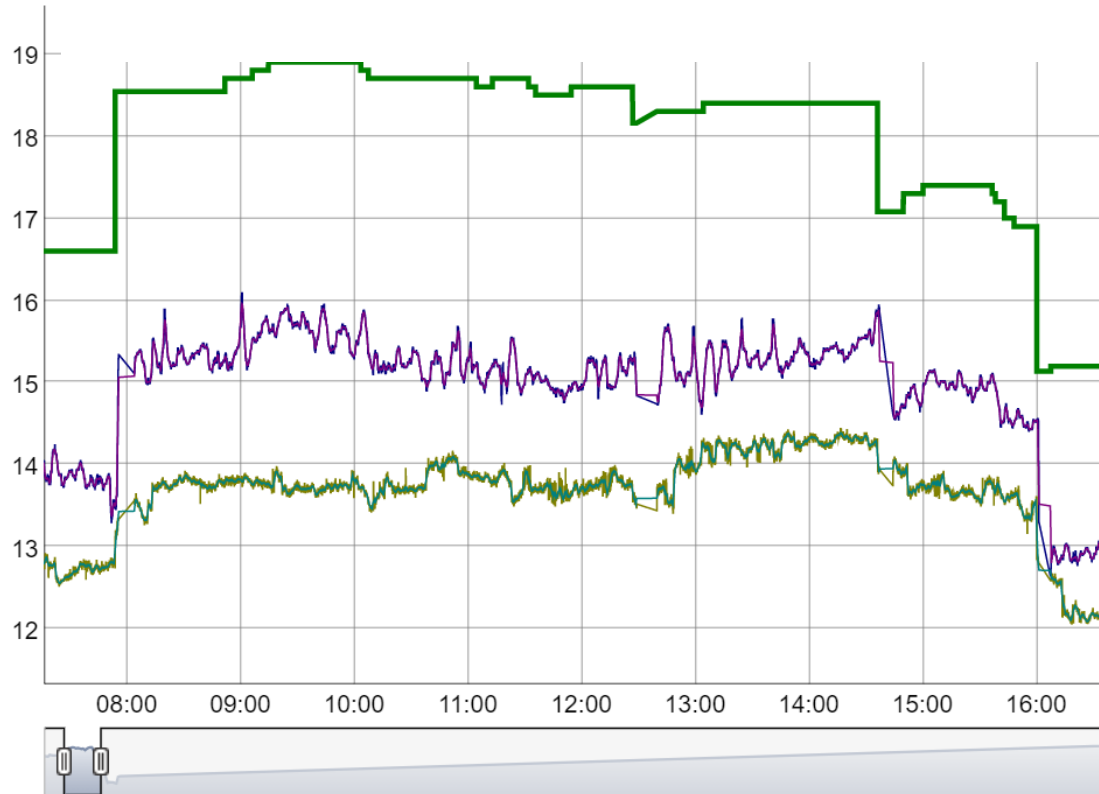
Predicciones de todos lo modelos



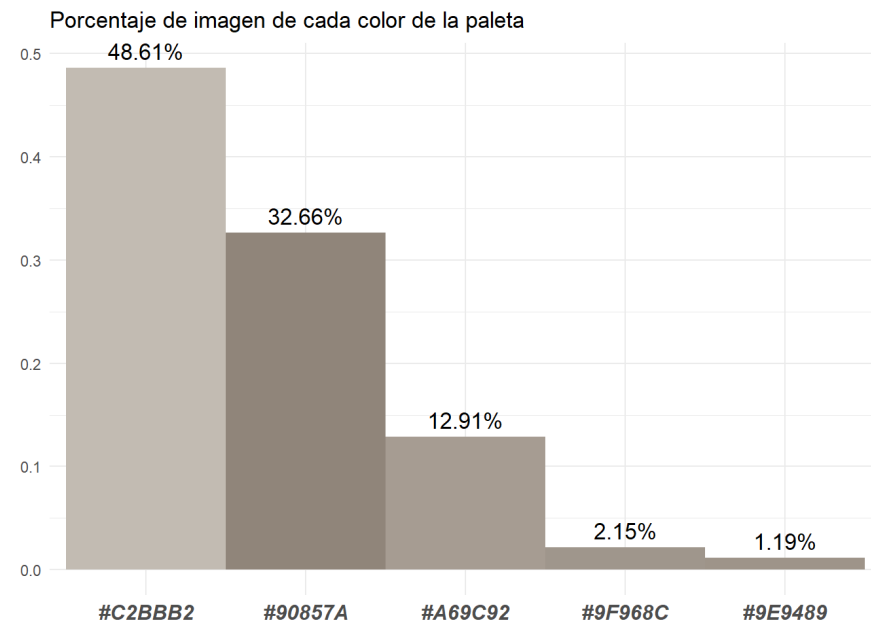
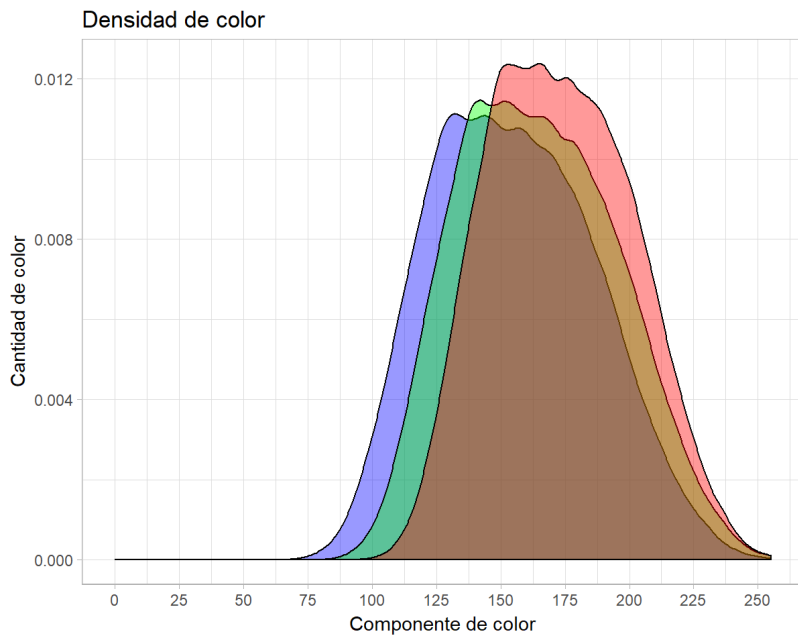
Serie y predicciones



Análise – Predicción industrial. Despliegue tempo real

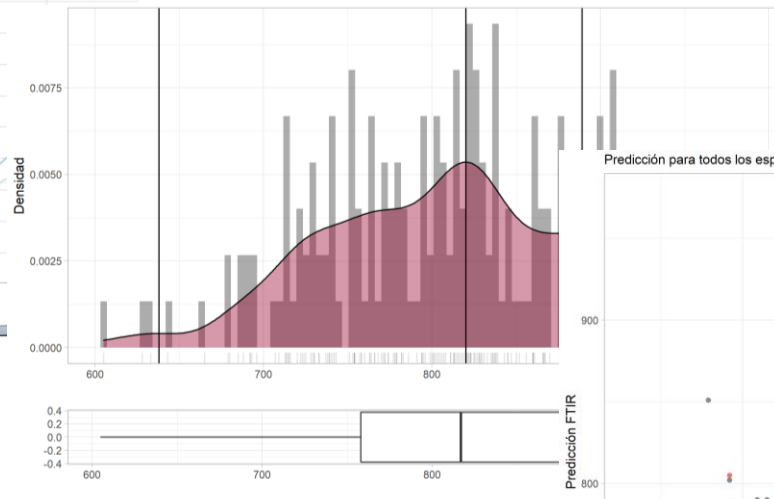


Análise – Paletas de cores

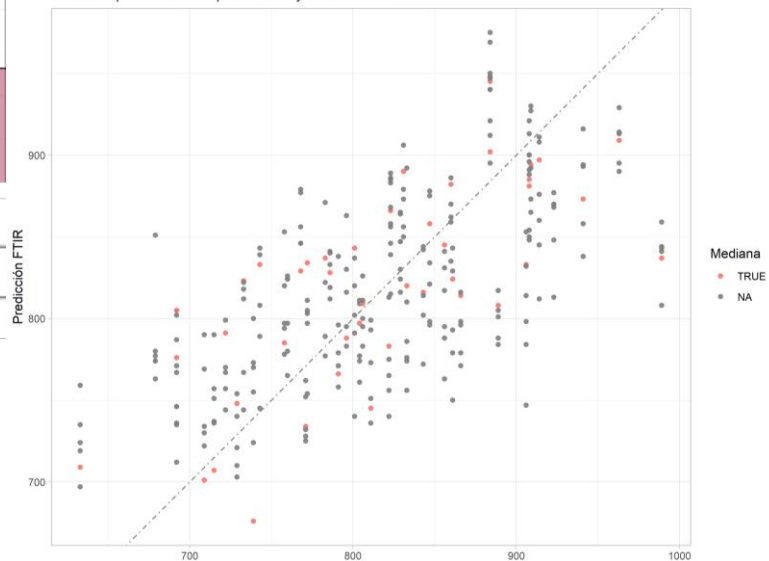


Análise – Predicción con espectros FTIR

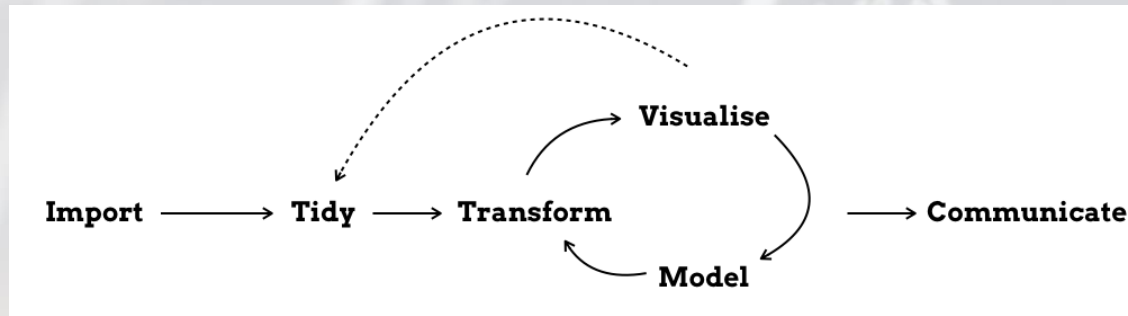
FTIR para diferentes



Predicción para todos los espectros FTIR y las medianas



Comunicación



Comunicación	Paquetes
Creación de informes automatizados	openxlsx, ggplot2, plotly, dygraphs, knitr, rmarkdown, shiny, shinydashboards, leaflet, DT, googlesheets4, googledrive, gmailR
Creación de follas de Excel formateadas	
Integración de datos en Google Spreadsheets	
Creación de aplicaciones en Shiny para facilitar o análise	
Distribución automática por correo	

Aplicacións Shiny – Selección de umbral para campañas

Análisis de selección del umbral

Seleccionar fichero

Seleccionar archivo 141201 an...fijo.csv

[Upload complete](#)

Selección del umbral:

0 0.5 1

Número de bajas estimadas:

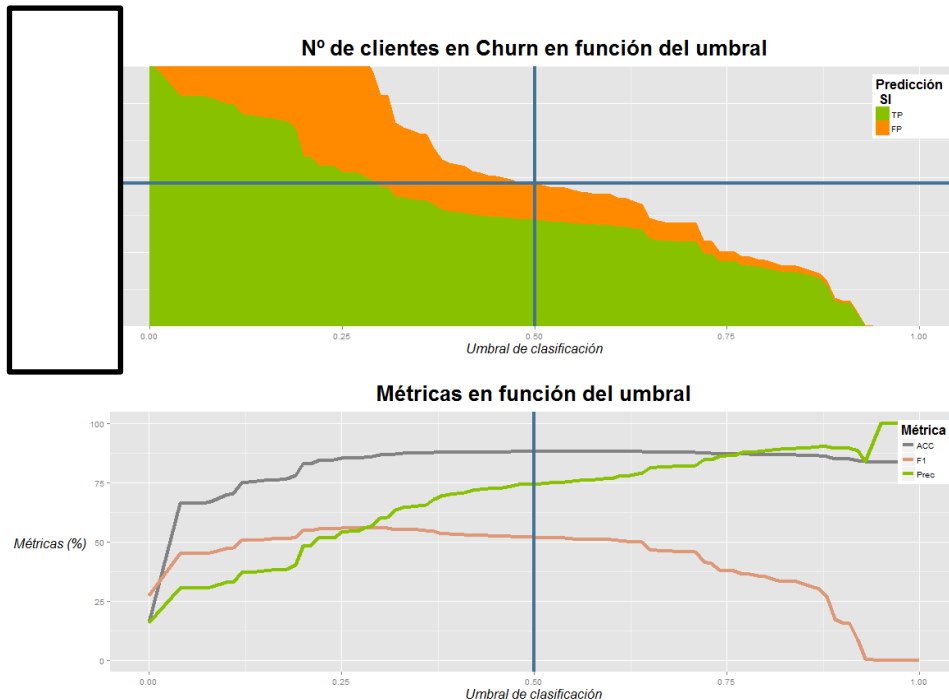
Métricas para el umbral seleccionado

Accuracy = 88.21 %

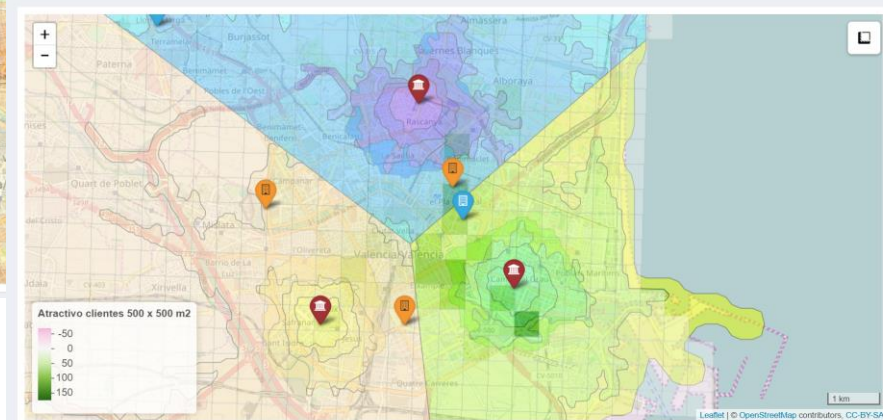
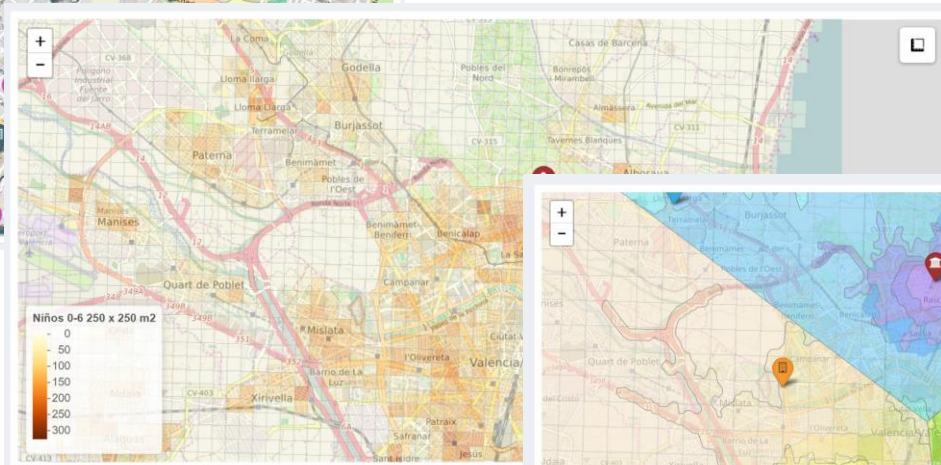
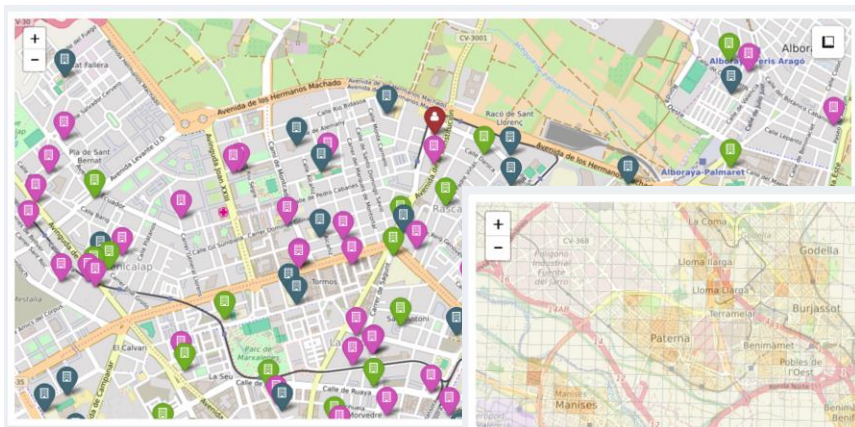
F1 Score = 52.14 %

Precisión = 74.54 %

Matriz de Clasificación



Aplicacións Shiny – Location Intelligence



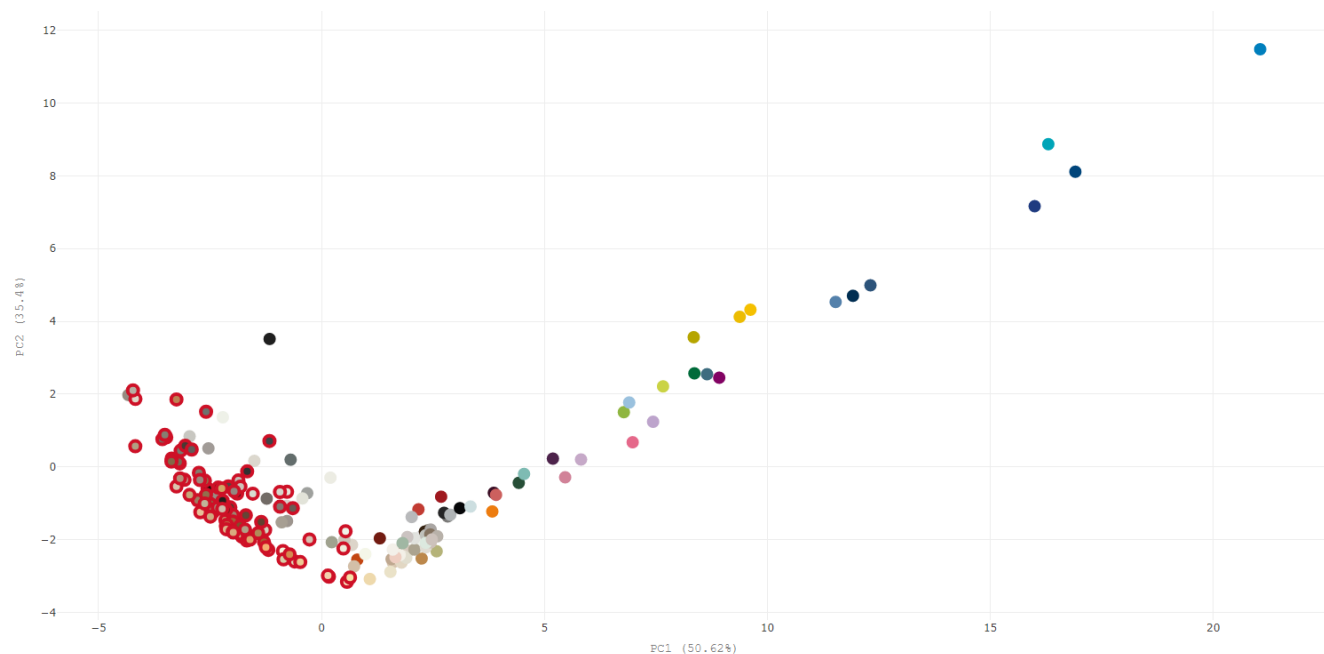
	NIÑOS TOTALES		CL. TOTALES		CL. AJUSTADOS		ATRATIVO
113587		13465		2309		1014	

Aplicacións Shiny – Paletas de cores

Materiales



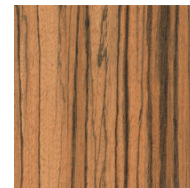
Gama de materiais



Proveedor



Zebrano Essence
Flaeche



Paleta

#CE9261	#543E29
#AA744A	#725439
#BD8254	#8F6643

Aplicacións Shiny – FTIR análise

FTIR analysis tool

Upload FTIR data

Choose or drop xlsx file containing FTIR data ...

Browse...

xlsx

Upload complete

File uploaded

Data	Value
File	
Lote	
Name	
Number of analysis	5

Analysis	Pred
1	856.00
2	823.00
3	866.00
4	818.00
5	841.00

Mean value

840.8

Mean prediction

20.63

Deviation

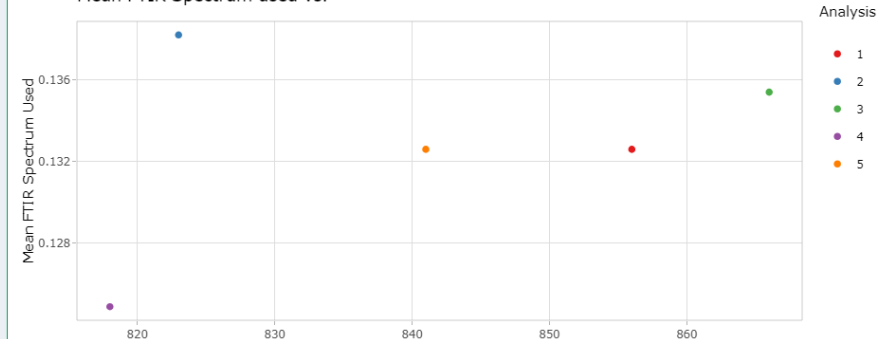
FTIR Spectrum

FTIR spectrum: 1, 2, 3, 4, 5



Mean FTIR Spectrum vs. Prediction

Mean FTIR Spectrum used vs.



Automatización

Capacidade de automatizar as tarefas anteriores asegurando a súa reproducibilidade.

Os modelos e a comunicación de resultados deben ter a capacidade de poñerse en producción

- Personalización de scripts con parámetros
- Programación de execución
- Creación de APIs para modelos: plumber
- Usar versións definidas dos paquetes: checkpoint, renv
- Creación de docker's para asegurar a persistencia no tempo e facilitar o despliegue

Recomendacións para o desenvolvemento con R

- É desenvolvemento de software – hai que ter unha metodoloxía:
 - Definir unha estrutura de proxecto
 - Dividir o proxecto en pequenos pasos secuenciais: definir entradas e saídas, funcionalidade, documentación para cada paso e organízalos mediante un fluxo de traballo
 - Usar unha guía de estilo para facilitar a comprensión
 - Usar un sistema de control de versión para asegurar trazabilidade temporal
 - Documentar o código

Desenrolo con R nas empresas

Para as empresas é importante:

- Proporcionar solucións aos problemas propostos: regra de Pareto
- Capacidade de productización
- Reproducibilidade dos resultados
- Explicabilidade dos resultados
- Iterar o máis rápido posible sobre os resultados para aprender



Exemplos de uso de R na empresa

Antonio Vidal Vidal
Senior Data Scientist at Nextail
avidalvi@gmail.com

