# Desafío 4 - Trabajo Final: Tienda Pippo

Integrantes:

* Bardauil, Joaquín
* Marquez, Hector
* Neustadt, Alejandro
* Pero, Felipe

Nuestro trabajo se genera a través de conseguir una base de datos de e-comerce de un comercio de blanquería, al que llamamos ficticiamente Pippo.

La base es un Excel que tiene 2 hojas, una con el detalle de las compras y la otra con la facturación de estas, las fechas de la base son desde abril’18 a mediados de junio’19.

Concertamos con el negocio enmascarar datos sensibles de la base de datos y por eso la primera parte del trabajo fue justamente que uno de nosotros realizara esa tarea. Luego la base quedo disponible para el equipo y empezamos a explorarla en conjunto.

Basándonos en el trabajo que habíamos realizado en los desafíos previos arrancamos con una limpieza y ordenamiento de datos. Y para coordinar en principio dividimos algunas tareas entre nosotros y luego en función de nuestras preferencias individuales, temas propiamente dicho.

Tuvimos en cuenta los consejos de usar el GitHub y avanzamos en el uso de éste a medida que avanzaba nuestro trabajo, también con el uso del Google Colab, mail, Wapp y Drive. Comunicación no nos faltó.

Luego como cada notebook de limpieza o preparación de datos iba modificando las salidas de archivos que podían ser usados en otros notebooks que estaban en proceso, generamos una tabla de detalle de archivos con entradas y salidas que permitía controlar avances y visualizar todo el conjunto. También un notebook general que va corriendo en orden cada trabajo individual.

Realizamos una Regresión Lineal Múltiple utilizando como variable target la cantidad de productos demandada (Quantity) y cuatro variables explicativas: mes, día, código postal y costo de envío. El resultado fue un R^2 bajísimo, el cual se mantenía así sin importar que variable explicativa incorporemos. Luego chequeamos los supuestos de Gauss Markov y vimos que eran varios los supuestos que no se cumplían, lo que podría explicar la mala performance de la regresión.

Posteriormente, intentamos predecir la variable categórica “product line” mediante modelos de clasificación. Tanto en la Regresión Logística como en Naive Bayes obtuvimos un nivel de accuracy relativamente alto, pero viendo la matriz de confusión pudimos observar que ambos modelos predecían siempre el mismo valor. Esto se debe a que las clases de la variable product line se encuentran desbalanceadas: la clase dormitorio representa el 60% de la muestra.

Como los resultados de estas predicciones no fueron satisfactorios, decidimos no incluirlos en la presentación.

Dicho esto, vamos al objetivo, los datos y notebooks:

* Conocer los patrones de consumo,
* Identificar clientes típicos
* Identificar productos típicos
* Días donde más se compra
* Tipos de artículos con más ventas
* Desarrollar **modelos predictivos** que contribuyan a captar nuevos clientes y aumentar las ventas futuras.
  + Recomendaciones
  + Pronósticos de ventas

Datos

* Limpieza de Datos
  + Base dwh y trabajo en Local de ventas (solo en PDF)
  + Limpieza\_Pippo
* Análisis Exploratorio
  + Líneas\_y\_tipos\_de\_productos\_Pippo
  + Análisis\_Pippo
* Modelos de Grafos
  + Preparacion\_grafo\_Pippo
  + Análisis\_grafo\_Pippo
  + Análisis\_grafo\_Pippo\_Tipo Producto
* Modelos de Predicción
  + Crearemos **Predicciones Temporales** con Redes Neuronales
    - **Simples**
    - **Multivariables**
    - **Embeddings**
  + Para predecir el comportamiento de:
    - Ventas diarias en $
    - Ventas diarias en Q
    - Productos de mayor venta
    - Tipos de Productos
  + Finalmente compararemos los diferentes métodos de predicción.
    - En total 12 notebooks detalladas abajo con las series, mas una de preparación datos.
    - Se observa en cada predicción con el tipo 3 Embeddings es superadora respeto **a la Red Neuronal Simple y Multivariables**



Los pronósticos para primer semana de Junio’19 y cada caso son:



* Modelos de Recomendaciones
  + Usamos 2 modelos de Recomendaciones, mas uno exploratorio didáctico a modo de ejemplo que pudo trabajar con pocos datos.
    - **Surprise**
    - **Turicreate** (modelos de filtrado colaborativos para recomendar productos a los clientes utilizando datos de compra)
  + Para recomendar según ID o Tipo de producto:
    - Productos de mayor venta
    - Tipos de Productos
  + Finalmente compararemos los diferentes modelos usados.
    - En total 6 notebooks detalladas abajo, mas una exploratoria didáctica



Para el modelo usado con Turicreate internamente realiza una selección del mejor método entre 3 modelos (Popularidad, Coseno y Person) y 3 vistas de datos (ventas, Dummies y un normalizado de ventas), seleccionando la mejor respuesta en el tipo Coseno y Dummies, a modo de ejemplo se muestra una de las tablas resumen, el total está en los notebooks o el resumen de comparaciones de métricas que realizamos.

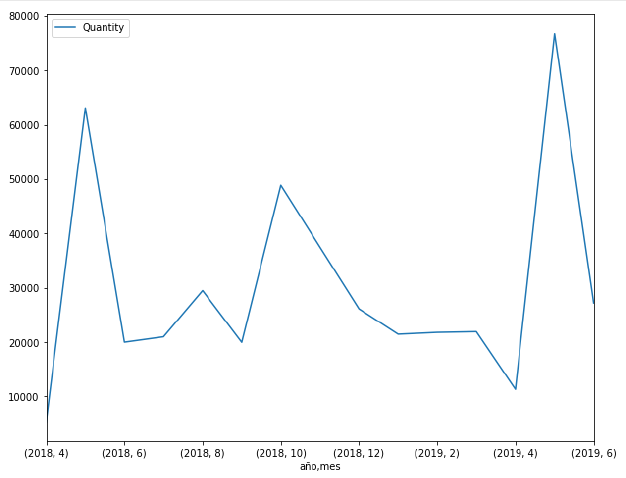


Detalle de notebooks

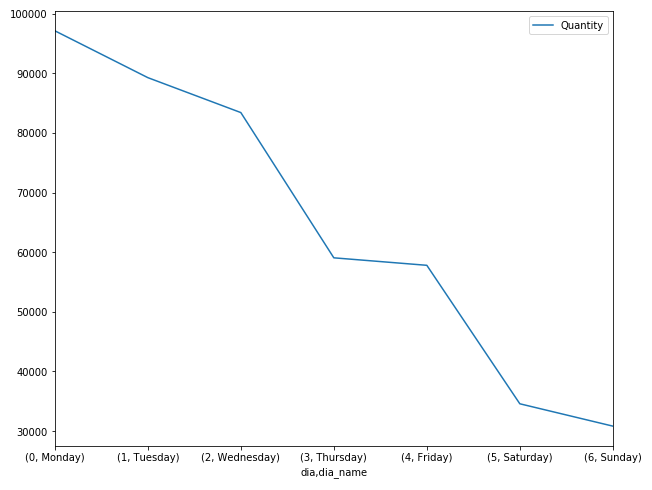


Detalle de visualizaciones y agrupamientos

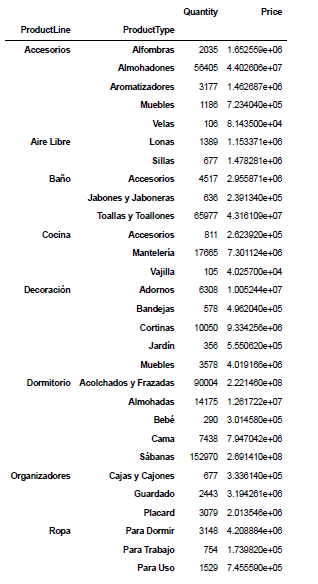
Ventas en el periodo de análisis



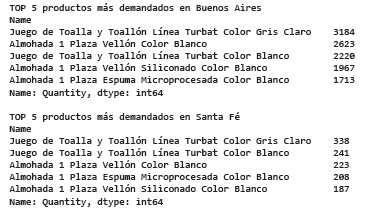
Ventas por días de la semana (lunes es cero)



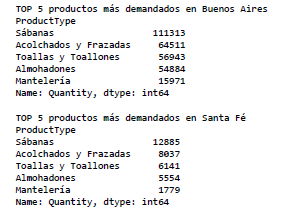
Grupby por Línea y Tipos de productos



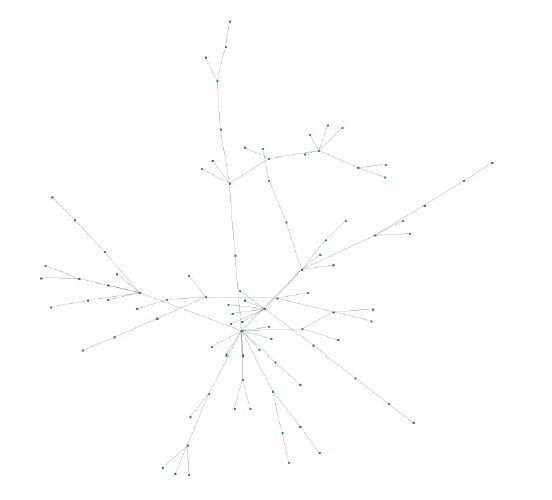
Productos mas demandados por zonas o lugares



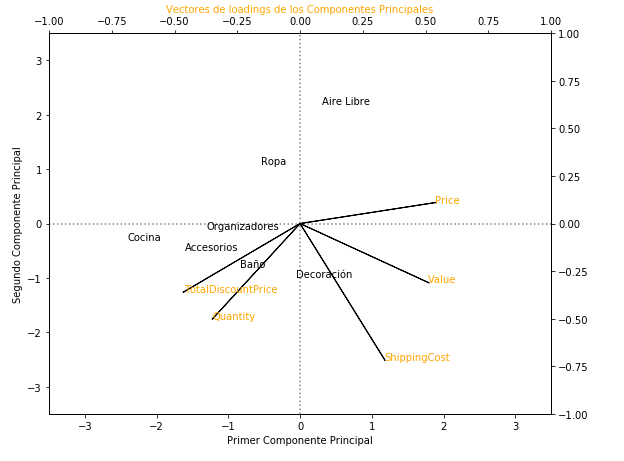
Tipos de productos mas demandados por zonas y lugares

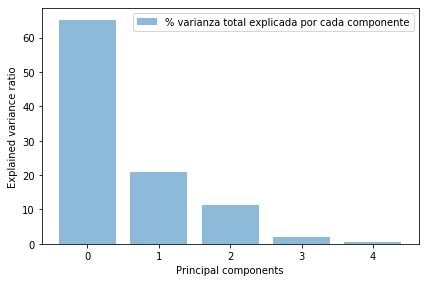
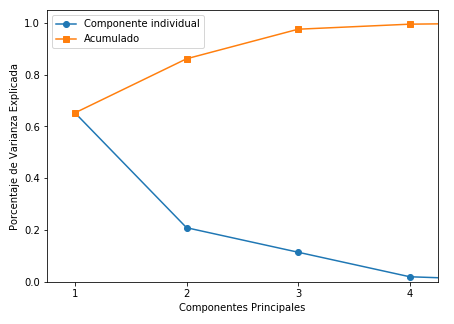


Grafos

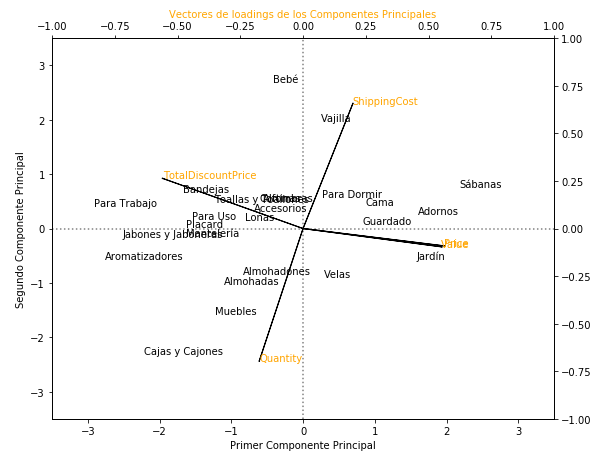


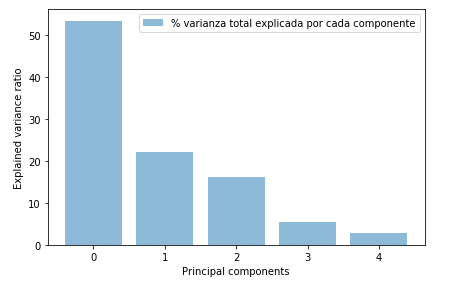
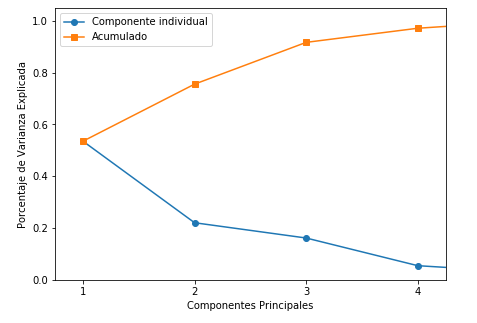
Armamos un PCA exploratorio por línea de producto





Por tipo de producto





Gráficos a lo analizado por zonas y productos

