KYPER
DATA TECHNOLOGIES

巨量資料分析

期末專題介紹

2017/05/05

Kyper Data Technologies

資料分析計畫步驟

KYPER

資料分析步驟

總覽

步驟一：總覽全局

步驟二：取得資料

步驟三：發掘與視覺化資料，以取得深入了解

步驟四：修正調整資料，以準備分析

步驟五：分析資料，建立數據模型

步驟六：調整優化模型

步驟七：提出你的見解

資料分析步驟

步驟一：總覽全局

- 了解你的問題領域
- 定義你想要解決的問題
  - 可測量
  - 明確
  - 簡潔
- 範例：是否能在減少員工數量情況下，保持目前公司的產能與品質？

資料分析步驟

步驟二：取得資料

- 決定需要哪些資料？
  - 領域
  - 來源
  - 資料格式
- 從不同來源取得目前現有的資料
- 透過問卷或者量測的方式，取得目前沒有的必要資料

資料分析步驟

步驟三：發掘與視覺化資料

- 了解你手上的數據
  - 資料數量
  - 特徵屬性
    - 類型 (文字、數字、集合...)
    - 範圍
    - 統計量

資料分析步驟

步驟四：修正調整資料

- 如何處理缺失資料?
  - 移除該筆資料
  - 補值
    - 0
    - 最常見值
    - 平均數
- 調整特徵值
  - 正規化
  - 合併特徵值
  - 根據現有特徵值產生新的特徵值

資料分析步驟

步驟五：分析資料

- 決定數據模型
  - Classifier
  - Regression Model
- 準備訓練資料與驗證資料
- 訓練數據模型

資料分析步驟

步驟六: 調整優化模型

- 交叉驗證模型
- 再次調整修正特徵值
- 嘗試不同參數值

KYPER

資料分析步驟

步驟七：提出你的見解

- 數據是否有回答到你原來的問題？ 如何回答？
- 數據是否能夠協助你防衛其他不同的異議？如何防禦？
- 你的結論是否有其限制？是否有未考慮到的死角存在？

資料集介紹

KYPER

Grupo Bimbo

- 墨西哥食品製造公司
- 商業模式:
  食品 -> 銷售點 -> 客戶 -> 一般消費者

KYPER

資料集

- train.csv 主資料
- test.csv 測試資料
- producto_tabla.csv 商品名稱
- cliente_tabla.csv 客戶名稱
- town_state.csv 銷售點位置

KYPER

主資料標籤介紹

- Semana
- Agencia_ID
- Canal_ID
- Ruta_SAK
- Cliente_ID
- Producto_ID
- Venta_uni_hoy
- Venta_hoy
- Dev_uni_proxima
- Dev_proxima
- Demand_uni_equil

	Semana	Agencia_ID	Canal_ID	Ruta_SAK	Cliente_ID	Producto_ID	\
0	3	1110	7	3301	15766	1212	
1	3	1110	7	3301	15766	1216	
2	3	1110	7	3301	15766	1238	
3	3	1110	7	3301	15766	1240	
4	3	1110	7	3301	15766	1242	

	Venta_uni_hoy	Venta_hoy	Dev_uni_proxima	Dev_proxima	Demanda_uni_equil
0	3	25.14	0	0.0	3
1	4	33.52	0	0.0	4
2	4	39.32	0	0.0	4
3	4	33.52	0	0.0	4
4	3	22.92	0	0.0	3

KYPER

檢查資料集

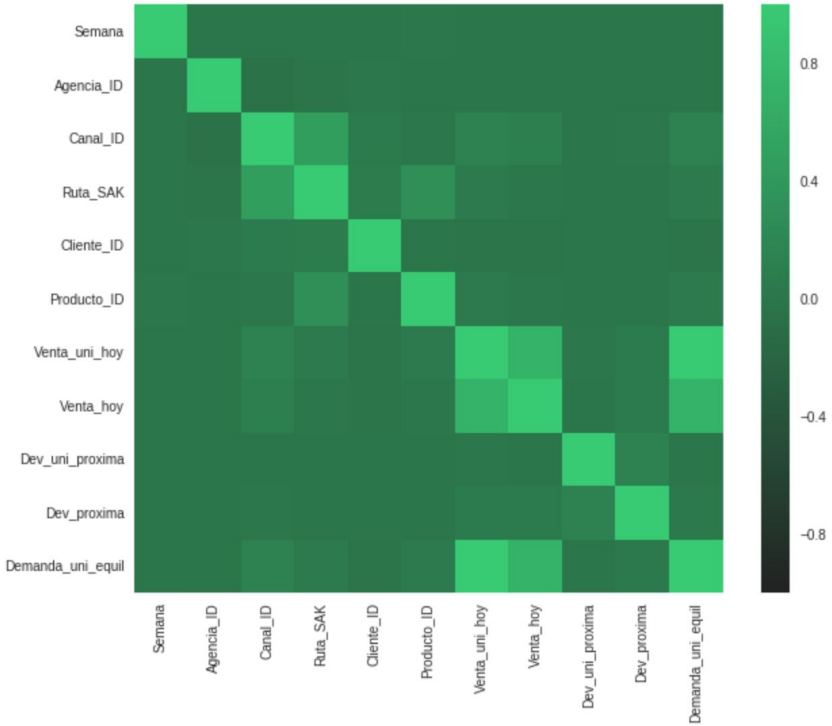
- Check Feature Type
- Duplicate
- Missing value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74180464 entries, 0 to 74180463
Data columns (total 11 columns):
Semana                int64
Agencia_ID            int64
Canal_ID              int64
Ruta_SAK              int64
Cliente_ID            int64
Producto_ID           int64
Venta_uni_hoy         int64
Venta_hoy             float64
Dev_uni_proxima       int64
Dev_proxima           float64
Demanda_uni_equil     int64
dtypes: float64(2), int64(9)
memory usage: 6.1 GB
None
```

KYPER

檢查資料集

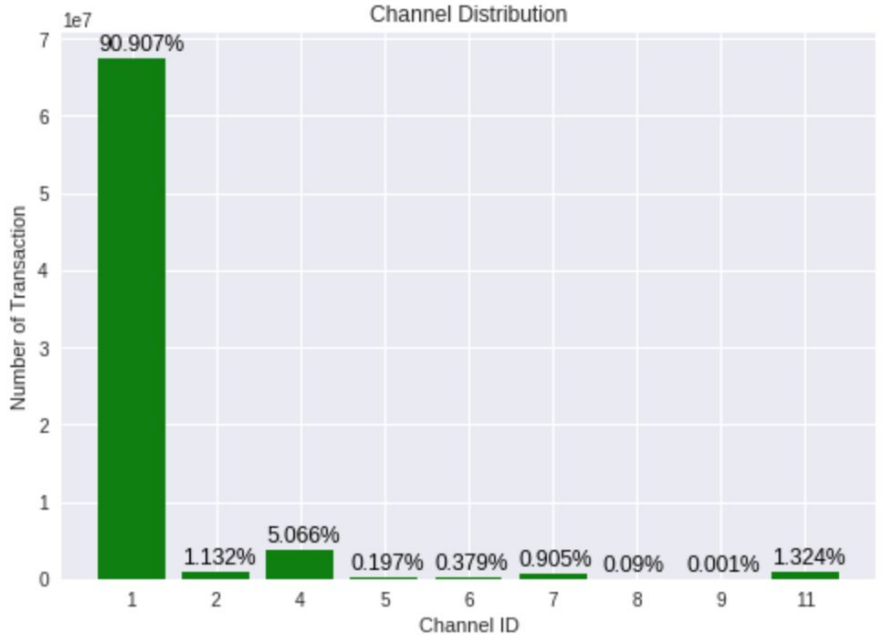
- 各個特徵的相關係數



KYPER

資料標籤分布視覺化 - Category Data

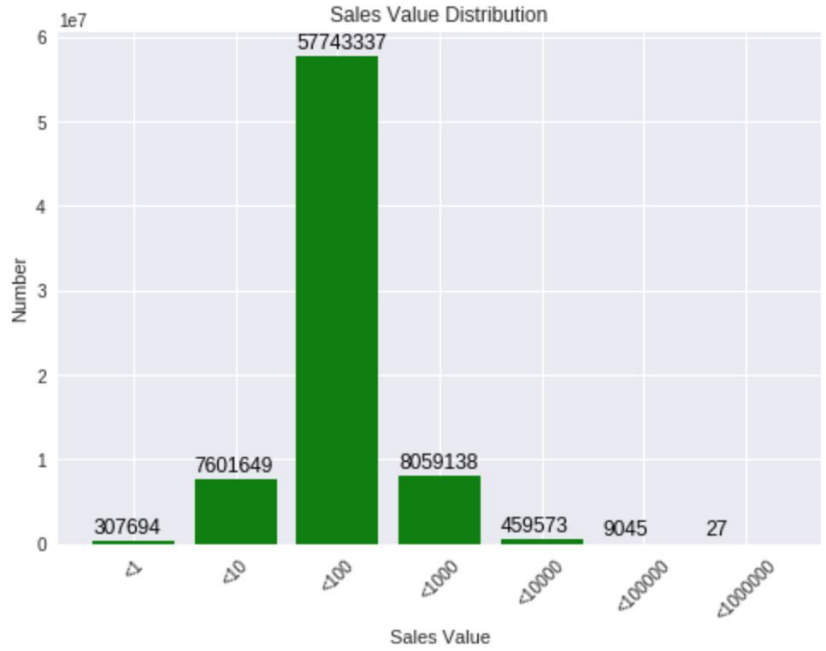
- Ex. Canal_ID



KYPER

資料標籤分布視覺化 - Numerical Data

- Ex. Venta_hoy (Sales)



KYPER

資料集檢索

- Producto_tabla.csv

	Producto_ID	NombreProducto
0	0	NO IDENTIFICADO 0
1	9	Capuccino Moka 750g NES 9
2	41	Bimbollos Ext sAjonjoli 6p 480g BIM 41
3	53	Burritos Sincro 170g CU LON 53
4	72	Div Tira Mini Doradita 4p 45g TR 72

- 拆解標籤內資訊

KYPER

資料集檢索

- Town_state.csv

	Agencia_ID	Town	State
0	1110	2008 AG. LAGO FILT	MÉXICO, D.F.
1	1111	2002 AG. AZCAPOTZALCO	MÉXICO, D.F.
2	1112	2004 AG. CUAUTITLAN	ESTADO DE MÉXICO
3	1113	2008 AG. LAGO FILT	MÉXICO, D.F.
4	1114	2029 AG.IZTAPALAPA 2	MÉXICO, D.F.

- 從質化到量化

KYPER