# General findings

## Cities.csv

- 6 variables: store_id, storetype_id, store_size, city_id_old, country_id, city_code
- 63 observations
- Variables: 1 text, 1 numeric, and 4 categorical
- No missing cells
- No duplicated rows

## Variables

### Store_id

- The column has the stores identifier.
- All IDs are unique - all distinct.
- There are no missing values.
- It is a text column.
- All the observations start with a “S” letter followed by three numbers.
- Zero is the character more represented.

### Storetype_id

- Identifies the type of store
- There are no missing values.
- It is a categorical variable.
- 4 levels: ST01, ST02, ST03 and ST04
- ST04 most representative - 63.5%, followed by ST03 - 31.7%
- ST01 - 2 observations and ST02 only 1.

### Store_size

- The columns represents the size of each store.
- There are no missing values.
- It is a numeric variable - integer.
- Integer values.
- 50.8% of observations are distinct values
- There are no negative, infinite numbers or zeros.
- The maximum size is 63, and the minimum is 8
- The mean is 24.8, and the median is 20
- The standard deviation is 12.6, and the variance is 159
- Higher frequency for 14, 17, and 19 (7.3%)
- Smaller store sizes are more common than bigger store sizes - data is skewed to lower values
- Highest representation between 12 and 20
- Presence of outliers (60 and 63)

### City_id_old

- The column has an old city identifier
- There are no missing values.
- It is a categorical variable.

- All the observations start with a “C” letter followed by three numbers.
- 19 distinct city_id (30.2%)
- C014 is the city_id most represented (50.8%)

**Country_id**

- Represents the country identifier
- There are no missing values.
- It is a categorical variable.
- Only 1 level (Country) - Turkey - in all observations.

**City_code**

- The column represents a new city identifier
- There are no missing values.
- It is a categorical variable.
- 19 distinct cities_code (30.2%)
- Istanbul is the most represented city_code (50.8%)
- Encoding issues - special characters like Ş, İ, and ı in city names are not properly encoded or decoded, resulting in characters like ? (Sanl?urfa, ?zmir，Diyarbak?r, Adapazar? and Eski?ehir)

**Correlations**

- Citiy_code is high correlated with city_id_old (1.00)
- Store_size is high correlated with storetype_id (0.695)
- The remaining variables are not correlated or have very low correlation

**Corrected city names**

Istanbul
Antalya
Sanliurfa
Konya
Izmir
Samsun
Kahramanmaras
Van
Denizli
Erzurum
Adana
Gaziantep
Diyarbakir
Kayseri
Bursa
Mersin
Ankara
Adapazari
Eskisehi

## Product.csv

- 10 variables: product_id, product_length, product_depth, product_width, cluster_id, hierarchy1_id, hierarchy2_id, hierarchy3_id, hierarchy4_id, hierarchy5_id
- 699 observations
- Variables: 4 text, 3 numeric, and 3 categorical
- 100 missing cells (1.4%)
- No duplicated rows

## Variables

### Product_id

- All IDs are unique - product identifier.
- There are no missing values.
- It is a text column.
- All the observations start with a “p” letter followed by three numbers.
- 699 different products
- Zero is the character more represented.

### Product_length

- There are 18 missing values (2.6%).
- It is a numeric variable - float
- 18.1% of observations are distinct values
- There are no negative or infinite numbers, and 1 zero.
- The maximum product_lengh is 100, and the minimum is 0.
- The mean is 7.23, and the median is 5.
- The standard deviation is 8.51, and the variance is 72.46.
- Higher frequency for 5 (12.3%).
- Smaller product lengths are more common than bigger product lengths -data is skewed to lower values
- Highest representation between 0 and 10.
- Presence of outliers (59, 62, 70, and 100).

### Product_depth

- There are 16 missing values (2.3%).
- It is a numeric variable.
- 23.9% of observations are distinct values
- There are no negative or infinite numbers, and 1 zero.
- The maximum product_depth is 165, and the minimum is 0.
- The mean is 18.46, and the median is 17.
- The standard deviation is 14.27, and the variance is 203.63.
- Higher frequency for 28 (4.9%).
- Smaller product depths are more common than bigger product depths - data is skewed to lower values
- Highest representation between 0 and 25.

- Presence of outliers (77, 80, 88, 89, 100, 160, 165).

**Product_width**

- There are 16 missing values (2.3%).
- It is a numeric variable.
- 21.1% of observations are distinct values
- There are no negative or infinite numbers, and 1 zero.
- The maximum product_widht is 100, and the minimum is 0.
- The mean is 13.45, and the median is 10.8.
- The standard deviation is 10.14, and the variance is 102.76.
- Higher frequency for 10 (5.7%).
- Smaller product widths are more common than bigger product widths - data is skewed to lower values
- Highest representation between 6 and 18.
- Presence of outliers (100).

**Cluster_id**

- Cluster identifier for grouping similar products maybe
- There are 50 missing values (7.2%).
- It is a categorical variable.
- 10 different levels (from cluster_0 to cluster_9)
- cluster_0 is the most represented cluster_id (64.4%)

**Hierarchy1_id**

- First level of product hierarchy
- There are no missing values.
- It is a categorical variable.
- 4 different hierarchies (from H00 to H03).
- H03 is the most represented hierarchy1_id (41.8%)
- H02 has only 11 observations (1.6%)

**Hierarchy2_id**

- Second level of product hierarchy
- There are no missing values.
- It is a categorical variable.
- Seems related to hierarchy1_id - 2 first numbers - code of hierarchy1_id followed by 2 numbers identifying hierarchy2_id
- 18 different hierarchies (from H0000 to H0004; H0105 to H0108; H0209 to H0210; H0311 to H0317)
- H03 products (hierarchy1_id) are the most represented
- H0313 is the most represented hierarchy2_id (14.4%)

**Hierarchy3_id**

- Third level of product hierarchy
- There are no missing values.
- It is a categorical variable (**Should be converted to categorical)**.
- Seems related to the previous hierarchies - 2 first numbers - code of hierarchy_1 followed by 2 numbers identifying hierarchy2_id and 2 last numbers related to hierarchy3_id
- 79 different hierarchies.

- H031302 is the most represented hierarchy1_id (5.6%)
- However, the products distribution seems more balanced

**Hierarchy4_id**

- Fourth level of product hierarchy
- There are no missing values.
- It is a categorical variable (**Should be converted to categorical)**.
- Seems related to the previous hierarchies (see hierarchy_3 example)
- 168 different hierarchies.
- H03130200 is the most represented hierarchy1_id (2.6%)

**Hierarchy5_id**

- Fifht level of product hierarchy
- There are no missing values.
- It is a categorical variable (**Should be converted to categorical)**.
- Seems related to hierarchy4_id (see hierarchy_3 example)
- 373 different hierarchies.
- H031302501 is the most represented hierarchy1_id (1.1%)

**Correlations**

- **hierarchy1_id** is highly overall correlated with **hierarchy2_id**
- **product_width** is correlated with **hierarchy1_id** and **product_depth**

## Sales.csv

- 14 variables (6 numeric, 5 categorical, 2 text, 1 dateTime)
- 8886058 observations
- 28.4% missing cells
- Last entry refers to 31/10/2019

### Correlations

- Promo_bin_1 is highly overall correlated with promo_bin_2 and liscount_type_2
- Promo_bin_2 is highly overall correlated with promo_bin_1 and 3 other fields (promo_discount_type_2, promo_discount_2 and promo_type_2)
- Promo_discount_2 is highly overall correlated with promo_bin_2 and 2 other field (promo_discount_type_2 and promo_type_2)
- Promo_discount_type_2 is highly overall correlated with promo_bin_1 and 3 other fields (promo_bin_2, promo_discount_2 and promo_type_2)
- Promo_type_2 is highly overall correlated with promo_bin_2 and 2 other fields (promo_discount_2 and promo_discount_type_2)
- Revenue is highly overall correlated with sales
- Sales is highly overall correlated with revenue

### Imbalanced data

- Promo_type_1 is highly imbalanced (77.3%)
- Promo_type_2 is highly imbalanced (99.1%)

### Missing values

- Sales has 302296 (3.4%) missing values
- Revenue has 302296 (3.4%) missing values
- Stock has 302296 (3.4%) missing values
- Price has 91381 (1.0%) missing values
- Promo_bin_1 has 7653515 (86.1%) missing values
- Promo_bin_2 has 8873337 (99.9%) missing values
- Promo_discount_2 has 8873337 (99.9%) missing values
- Promo_discount_type_2 has 8873337 (99.9%) missing values

**Skewed data**

- Sales is highly skewed (γ1 = 1557.844936) to the left
- Revenue is highly skewed (γ1 = 815.4548181) to the left
- Stock is highly skewed (γ1 = 24.21927272) to the left

**Unique values**

- Unnamed: 0 is uniformly distributed
- Unnamed: 0 has unique values
- “Unnamed: 0 “ column seems to represent sales number, starting at 1

**Zeros**

- Sales has 7048907 (79.3%) zeros
- Revenue has 7049979 (79.3%) zeros

**Variables**

**Unnamed: 0**

- All distinct - index column
- There are no missing values.
- Real number column.

**Store_id**

- Identifier for the store
- Distinct 63 Stores
- No Missing Values
- Text Column
- All Store_id’s have 5 characters (start With “S” Followed By 4 Numbers)
- Most represented store_id: S0038 - 3.8% followed by S0085 - 3.7%

**Product_id**

- Identifier for the product
- 615 Distinct Products
- No Missing Values
- Text Column
- All product_id’s have 5 characters (start With “P” Followed By 4 Numbers)
- All products have a similar frequency

**Date**

- Date of the record
- 1033 Distinct Dates
- No Missing Values
- yyyy-mm-dd format
- Minimum Date: 2017-01-02 00:00:00

- Maximum Date: 2019-10-31 00:00:00
- Higher frequency on the most recent dates
- Datetime Column

**Sales**

- Column represents number of unit sales for each product
- 5435 Distinct Sales
- 3.4% Missing Values
- Float Number Column (decimal numbers present)
- 7048907 Zeros (79.3%)
- Maximum: 43301
- Minimum: 0
- Sum: 4063622
- Mean: 0.47
- Median: 0
- Standard deviation: 21.29
- Variance: 453.29
- Most Common Values: 0 (79.3%), 1 (9.5%), 2 (3.4%)
- number of outliers: 1534855
- max outlier value: 43301.0
- min outlier value: 0.018

**Revenue**

- Column represents the resultant revenue for each product sale
- 12155 Distinct Values
- Real Number Column
- 302296 (3.4% Missing Values)
- Mean: 2.29
- Median: 0
- Standard deviation: 54.07
- Variance: 2923.35
- Maximum: 84197.961
- Minimum: 0
- Sum: 84197.961
- 7049979 Zeros (79.3%)
- Number of outliers: 1533783
- Max outlier value: 84197.961
- Min outlier value: 0.01
- No currency available

**Stock**

- The column represents the number of products available to sell
- 9039 Distinct Values
- Real Number
- 302296 (3.4%) Missing Values
- Mean: 16.01
- Median: 8
- Standard deviation: 37.52
- Variance: 1407.52
- Maximum: 4655
- Minimum: 0
- Sum: $1.3738952 \times 10^8$
- Common Values: 4 (7.0%), 3 (6.9%), 6 (6.8%), 2 (6.6%), 5 (6.4%)
- number of outliers: 674701
- max outlier value: 4655.0
- min outlier value: 36.525

**Price**

- The column represents the price of each product
- 606 Distinct Values
- Real Number Column
- 91381 Missing Values (1.0%)
- Mean: 15.75
- Median: 8
- Standard deviation: 32.77
- Variance: 1074.44
- Minimum: 0.01
- Maximum: 1599
- Number of outliers: 747066
- Max outlier value: 1599.0
- Min outlier value: 37.25
- No currency available

**Promo_type_1**

- Column is associated with the type of first promotion - different promotional codes
- 17 Distinct Values
- 0 Missing Values
- Categorical Column
- Common Values: Pr14 (86.1%), Pr05 (6.2%), Pr10 (2.4%)
- Promo_type_1 is represented by 4 characters (2 letters and 2 numbers)

**Promo_bin_1**

- Binned promotion rate for applied promo_type_1
- 5 Distinct Values
- 7653515 (86.1%) Missing Values
- Categorical Column - 5 levels (verylow, low, moderate, high and veryhigh)
- Common Values: Verylow (5.8%), Low (2.9%), Moderate (2.2%), High (1.6%), Veryhigh(1.3%)

**Promo_type_2**

- Column is associated with the type of second promotion
- 4 Distinct Values
- 0 Missing Values
- Categorical Column
- Common Value: PR03 (99.9%)
- Promo_type_2 is represented by 4 characters (2 letters and 2 numbers)

**Promo_bin_2**

- Binned promotion rate for applied promo_type_2
- Categorical Column
- 3 Distinct categories: verylow, high and veryhigh
- 8873337 (99.9%) Missing Values
- Verylow is the most represented category (50.6%)

**Promo_discount_2**

- Column represents the discount rate for applied promo_type_2
- 6 Distinct Values
- 8873337 (99.9%) Missing Values
- Mean: 30.11
- Median: 20
- Standard deviation: 11.85
- Variance: 140.44
- Minimum: 16
- Maximum: 50
- Real Number Column
- Common Values: 20, 33, 50, 35, 40
- number of outliers: 1534855
- max outlier value: 43301.0
- min outlier value: 0.018

**Promo_discount_type_2**

- Column represents the type of discount applied
- 4 Distinct Values
- 8873337 (99.9%) Missing Values
- Categorical Column
- Promo_discount_type_2 is represented by 4 characters (2 letters and 2 numbers)

## Forecast Revenue.csv

- 4 variables (1 text, 2 categorical, 1 numeric)
- 1943 observations
- 0 missing values
- There are different stores with forecast revenue, different from the store of the sales dataset
- There is no reference to data between 25/09/2019 and 30/09/2019.

## Correlations

- Date is highly overall correlated with WEEk

## Variables

### Store

- The column has the store identifier
- 63 distinct values
- No missing values
- Text variable
- Seems to be the store_id variable of cities.csv

### Date

- 33 distinct values
- No missing values
- Categorical variable
- dd/mm/yyyy format
- first date: 10/1/2019
- last date: 10/22/2019

### Fcst_revenue

- 795 distinct values
- No missing values
- Mean: 457.18
- Median: 295
- Minimum: 11
- Maximum: 3443
- Sum: 888310.3
- Standard deviation: 490.11
- Variance: 240212.68
- Data is skewed to lower values
- No currency available

### WEEk

- 5 distinct values (representing weeks 41 to 44)
- No missing values
- Categorical variable

# Exploratory analysis

- Months with the **most sales** in each year:
    - 2017 - February and April
    - 2018 - May and October
    - 2019 - February and August

- Months with **fewer sales**:
    - 2017- June and July
    - 2018 - February and November
    - 2019 - May and July
- No records of **stock**, **sales** and **revenue** in October 2019
- There are sales with **decimal numbers**, probably associated with different selling of products which are sold by meter, instead of units.
- 57/63 stores have decimal numbers in sales: ['S0002' 'S0003' 'S0005' 'S0010' 'S0012' 'S0014' 'S0015' 'S0016' 'S0020' 'S0022' 'S0023' 'S0026' 'S0032' 'S0036' 'S0038' 'S0039' 'S0040' 'S0045' 'S0046' 'S0050' 'S0052' 'S0055' 'S0056' 'S0058' 'S0059' 'S0061' 'S0062' 'S0067' 'S0068' 'S0071' 'S0072' 'S0073' 'S0080' 'S0083' 'S0085' 'S0086' 'S0088' 'S0089' 'S0091' 'S0092' 'S0094' 'S0095' 'S0097' 'S0099' 'S0102' 'S0104' 'S0107' 'S0108' 'S0109' 'S0120' 'S0122' 'S0126' 'S0131' 'S0132' 'S0136' 'S0142' 'S0143']
- 11/564 products have decimal numbers in sales: ['P0413' 'P0561' 'P0316' 'P0176' 'P0610' 'P0630' 'P0550' 'P0725' 'P0031' 'P0155' 'P0484']
- All store types have decimal numbers in sales
- **Stores** with the **most sales**:
    - S0085 - 90070
    - S0062 - 68952
    - S0026 - 67361
    - S0020 - 62658

- Some stores **do not** show data for 2017 sales, possibly they only opened after 2017 (S0005, S0036, S0046, S0061, S0071, S0076, S0092, S0109).
- **S0007** only shows values for **2019**, which means it possibly only opened in 2019.

- When **sales = 0,** there were **no sales**, and the **stock** value remains **unchanged**. Meaning that each row with sales = 0 represents the stock of the respective product, for that day.
- When a sale occurs, the stock value decreases accordingly.
- Every day, each product has **only** a row showing if there were or not sales.

- **Stores by city (top 4)**
    - Instanbul - 32
    - Antalya - 5

    - Sanliurfa - 3
    - Konya - 3
- **Istanbul** has the majority of stores (**6x more** than Antalya)
- S0085 has the **higher revenue** (2.07M, 3-year cumulative), followed by S0097, S0026 and S0062
- In general, **2018** was the year which had more revenue.
- Istanbul is the city with more revenue (12.28 M, 3-year cumulative), followed by Bursa and Konya.

- **Top sellers by city**

| City | Top 3 products |
|---|---|
| **Istanbul** | **P0103**<br>**P0664**<br>**P0694** |
| **Ankara** | **P0103**<br>**P0503**<br>**P0183** |
| **Bursa** | **P0233**<br>**P0237**<br>**P0263** |
| **Adapazari** | **P0543**<br>**P0336**<br>**P0437** |
| **Konya** | **P0453**<br>**P0125**<br>**P0536** |
| **Eskisehir** | **P0090**<br>**P0506**<br>**P0491** |
| **Kayseri** | **P0456**<br>**P0125**<br>**P0131** |
| **Mersin** | **P0608**<br>**P0652**<br>**P0103** |
| **Samsun** | **P0663**<br>**P0664**<br>**P0499** |
| **Kahramanmaras** | **P0212**<br>**P0129**<br>**P0277** |

| | |
|---|---|
| **Denizli** | **P0325**<br>**P0286**<br>**P0406** |
| **Gaziantep** | **P0177**<br>**P0067**<br>**P0212** |
| **Diyarbakir** | **P0663**<br>**P0695**<br>**P0712** |
| **Adana** | **P0005**<br>**P0051**<br>**P0131** |
| **Sanliurfa** | **P0131**<br>**P0297**<br>**P0712** |
| **Van** | **P0125**<br>**P0131**<br>**P0015** |
| **Antalya** | **P0325**<br>**P0664**<br>**P0261** |
| **Izmir** | **P0261**<br>**P0125**<br>**P0348** |
| **Erzurum** | **P0005**<br>**P0054**<br>**P0067** |

- The number of products sold (top 3) in each city is similar, except for Istanbul where the top 3 products sold could be almost 30x higher.
- Product P0103 seems to have more sells, followed by P0364, P033 and P0569
- Some products do not show data for 2017 or 2018 sales, probably are new collections

# Konya findings

## Overview

**Dataset statistics**
532614 observations
34 variables
11.4% missing values

**Variable types**
10 numeric
21 categorical
1 text
1 boolean
1 DateTime

# Variables

## Unnamed: 0
Numeric
All distinct
No missing or infinite values
No zeros

## Store_id
Categorical
No missing values
3 stores: S0094, S0142 and S0030
Most represented:
- S0094 - 50.2%
- S0142 - 38.2%
- S0030 - 11.6%

## Product_id
Text
No missing values
480 products
Most represented:
- p0453 - 0.6%
- p0125 - 0.6%
- p0536 - 0.6%
- p0015 - 0.6%

## Date
Date
No missing values
Min - 2017-01-02
Max  - 2019-09-30

## Sales
Numeric
No missing or infinite values
86.6% zeros

## Revenue
Numeric
No missing or infinite values
86.6% zeros

## Stock
Numeric
No missing or infinite values
0.4% zeros

## Price
Numeric
0.2% missing values
No infinite values
No zeros

## Promo_type_1
Categorical
No missing values
16 promo_type_1
Most represented:
- PR14 - 85.8%
- PR05 - 6.7%
- PR10 - 2.4%

## Promo_bin_1

Categorical
85.8% missing values
5 promo_bin_1
Most represented:
verylow - 5.7%
low - 3.0%
moderate - 2.3%

## Promo_type_2

Categorical
No missing values
4 promo_type_2
Most represented:
P03 - 99.9%

## Promo_bin_2

Categorical
99.9% missing values
3 promo_bin_2
Most represented:
verylow - 0.1%

## Promo_discount_2

Categorical
99.9% missing values
5 promo_discount_2
Most represented:
20.0 - 0.1%

## Promo_discount_type_2

Categorical
99.9% missing values
5 promo_discount_2
Most represented:
PR02 - <0.1%

## Product_length

Numeric
0.6 % missing values
No infinite values
0.1% zeros

## Product_depth

Numeric
0.6 % missing values
No infinite values
0.1% zeros

## Product_width

Numeric
0.6 % missing values
No infinite values
0.1% zeros

## Cluster_id

Categorical
No missing values
10 cluster_id
Most represented:
cluster_0 - 58.5%
cluster_9 - 8.4%
cluster_4 - 6.9%
cluster_3 - 6.7%

## Hierarchy1_id

Categorical
No missing values
4 hierarchy1_id
Most represented:
H00 - 42.2%
H01 - 31.0%
H03 - 26.5%

## Hierarchy2_id

Categorical
No missing values
18 hierarchy2_id
Most represented:
H0108 - 15.9.0%
H0003 - 14.6%
H0002 - 10.5%
H0313 - 10.1%

## Hierarchy3_id

Categorical
No missing values
77 hierarchy3_id
Most represented:
H000312 - 6.0%
H010601 - 5.0%
H010807 - 4.5%
H000004 - 4.2%

## Hierarchy4_id

Categorical
No missing values
151 hierarchy4_id
Most represented:
H00031200 - 4.7%
H01080500 - 2.8%
H00010210 - 2.7%
H00000405 - 2.6%

## Hierarchy5_id

Categorical
No missing values
292 hierarchy5_id
Most represented:
H0001021012 - 2.1%
H0000040501 - 2.1%

## Storetype_id

Categorical
No missing values
2 storetype_id
Most represented:
ST04 - 88.4%
ST03 - 11.6%

## Store_size

Categorical
No missing values
3 store_size
Most represented:
45 - 50.2%
31 - 38.2%
13 - 11.6%

## City_old_id

Categorical
No missing values
1 city_old_id

## Country_id

Categorical
No missing values
1 country_id

## City_code

Categorical
No missing values
1 city_code

## Day

Numeric
No missing or infinite values
No zeros

## Weekday

Categorical
No missing values
7 weekday

## Season

Categorical
No missing values
4 season
Most represented:
3 - 28.2%
2 - 27.4%
1 - 26.0%

## Week

Numeric
No missing or infinite values
No zeros

## Holiday

Boolean
No missing values

**Month_name**

Categorical
No missing values
12 month_name

Most represented:
Jul - 9.5%
Aug - 9.5%
May - 9.3%

## Missing values

- price (0.2% missing values)
- promo_bin_1 (85.8% missing values)
- promo_bin_2 (99.9% missing values)
- promo_discount_2 (99.9% missing values)
- promo_discount_type_2 (99.9% missing values)
- product_length (0.6% missing values)
- product_depth (0.6% missing values)
- product_width (0.6% missing values)

## Outliers (numeric variables)

- sales (13.40%)
- revenue (13.39%)
- stock (8.62%)
- price (9.35%)
- product_length (8.02%)
- product_depth (2.84%)
- product_width (7.03%)

## Correlations(>0.600 or <-0.600)

- season with week.

## Distributions (numeric variables)

- sales - skewed to the right
- revenue - skewed to the right
- stock - skewed to the right
- price - skewed to the right
- product_length - skewed to the right
- product_depth - skewed to the right
- product_width - skewed to the right

# Possible variable transformations

- Variables to drop
    - prom_bin_1 (85.8% missing values)
    - promo_bin_2 (99.9% missing values)
    - promo_discount_2 (99.9% missing values)
    - promo_discount_type_2 (99.9% missing values)
    - hierarchy2_id (is a subcategory of hierarchy1_id)
    - hierarchy3_id (is a subcategory of hierarchy2_id)
    - hierarchy4_id (is a subcategory of hierarchy3_id)
    - hierarchy5_id (is a subcategory of hierarchy4_id)
    - city_old_id (only 1 value)
    - country_id (only 1 value)
    - city_code (only 1 value)

- Variables to keep
    - store_id
    - product_id
    - storetype_id
    - store_size
    - date
    - sales
    - stock
    - price
    - promo_type_1
    - promo_type_2
    - cluster_id
    - hierarchy1_id
    - weekday
    - season
    - holiday,
    - month_name

- Variables for one hot encoding
    - store_id
    - storetype_id
    - store_size
    - product_id
    - promo_type_1
    - promo_type_2
    - cluster_id
    - hierarchy1_id
    - weekday

    - season

**General findings for Konya**

- There was a high increase in sales at the end of 2017
- increase in sales for months: 5-6 and 9
- store S0094 has the highest sales for both years
- all 3 stores were opened since 2017 and have sales registered
- S0094 store has the highest revenue, across the years.
- Some products only have sales from 2018. Probably new products inserted in 2018.
- Some products started selling only in 2019 (new products).
- Products selling from 2017 registered more sales in 2019, compared with new products.
- S0094 store has the highest number of products.
- 109 products are present in the three stores.
- The amount of products in each cluster is unevenly distributed: cluster_0 has 300, cluster_1 has 4, cluster_2 has 9, and the remaining 6 clusters have between 10 and 34 products.
- There are no repeated products across clusters.
- Four clusters have products from every hierarchy1_id and cluster_1 only has products from one hierarchy1_id
- All clusters are present in every store.
- Sales, stock, price, and revenue from 2017 to 2019 have different trends for products within each cluster.
- There is no clear relationship among products within each cluster.