Data Journalism

Data Analysis

Part 1

Angelica Lo Duca
angelica.loduca@iit.cnr.it

No one ever made a decision because of a number. They need a story.

Daniel Kahneman

Insight is the discovery of non-trivial, complex, deep, unexpected, or relevant truths about the information

Three types of analysis

DESCRIPTIVE ANALYTICS

Analyse the past

What happened?

DIAGNOSTIC ANALYTICS

Analyse the present

What is happening right now?

PREDICTIVE ANALYTICS

Predict the future

What will happen?

Descriptive Analysis

Si basa sul calcolo di alcune **metriche** o **indici**

- Indici di frequenza
- Indici di tendenza centrale
- Indici di variabilità

Indici di Frequenza

Descrivere una singola variabile nel dataset

1

COUNT

Data una variabile, contare quante volte appare una certa categoria

2

PERCENTUALE

percentuale relativa al conteggio precedente

Indici di Tendenza Centrale

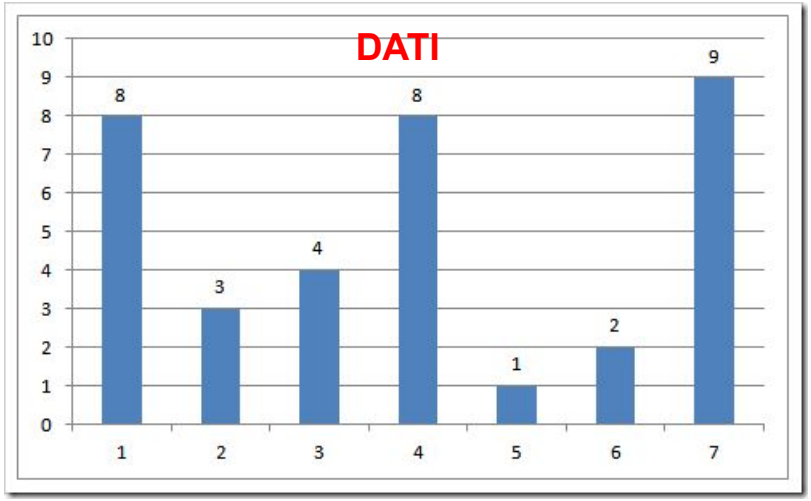
Descrivere i dati con un solo valore

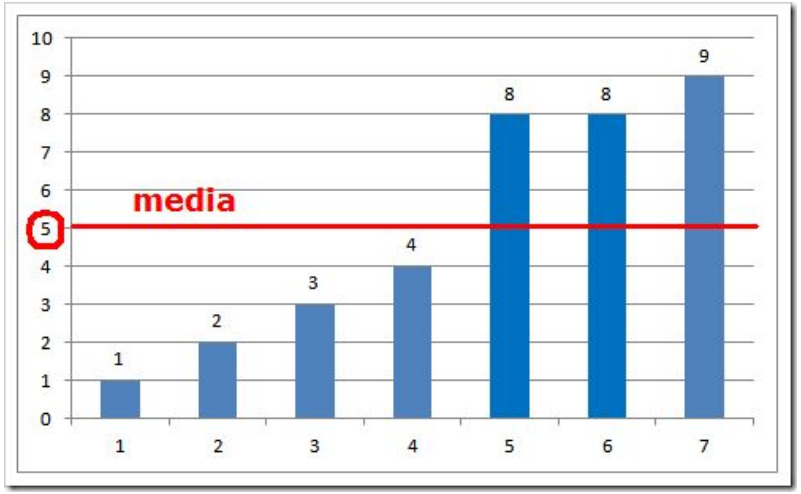
1. **MEDIA ARITMETICA**
   Somma dei dati

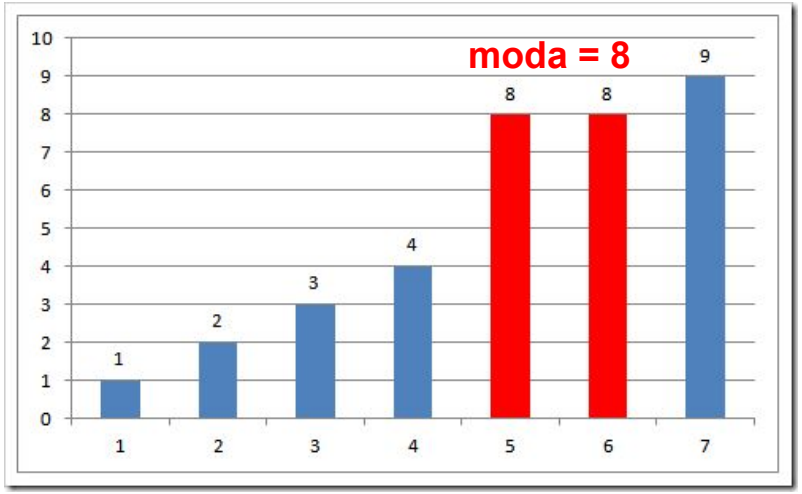
   Numero dei dati

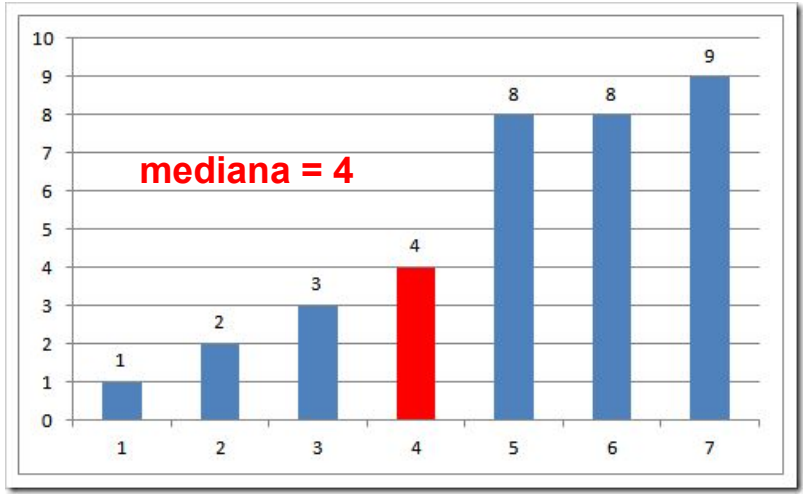
2. **MEDIANA (o 50° percentile)**
   valore al di sotto del quale cade la metà dei dati (valore centrale)

3. **MODA**
   valore che ricorre con maggiore frequenza


DATI
10
9
8
7
6
5
4
3
2
1
0
8
3
4
8
1
2
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
media
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
moda = 8
1
2
3
4
8
8
9
1
2
3
4
5
6
7


10
9
8
7
6
5
4
3
2
1
0
mediana = 4
1
2
3
4
8
8
9
1
2
3
4
5
6
7

When to use the MEAN

Both the following conditions must be satisfied:

- Data are scaled, i.e. data with equal intervals, such as time, temperature, speed
- Data distribution is quite normal, i.e. there are not outliers

When to use the MEDIAN

One of the following condition is satisfied:

- data are ordinal (first, second, third, …)
- distribution is skewed or non normal

When to use the MODE

When you want to know the most frequent value.

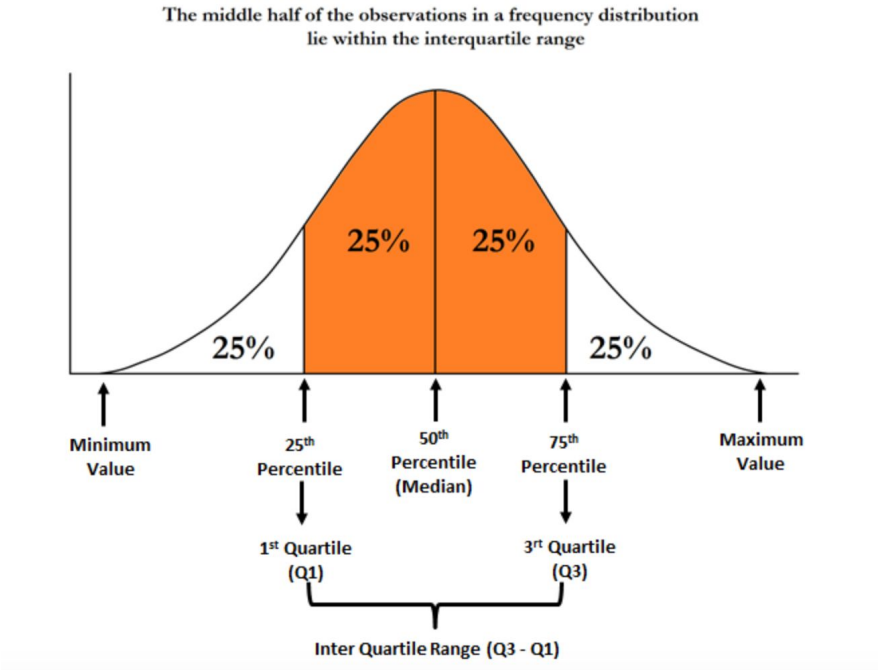
Indici di Variabilità

Descrivere la variabilità dei dati

1. **MAXIMUM** valore massimo
   MINIMUM valore minimo
   RANGE Differenza tra il valore massimo e il valore minimo
2. **QUARTILE**
3. **VARIANZA**
   dispersione dei valori del dataset attorno al valor medio. La deviazione standard è la radice quadrata della varianza

Quartile

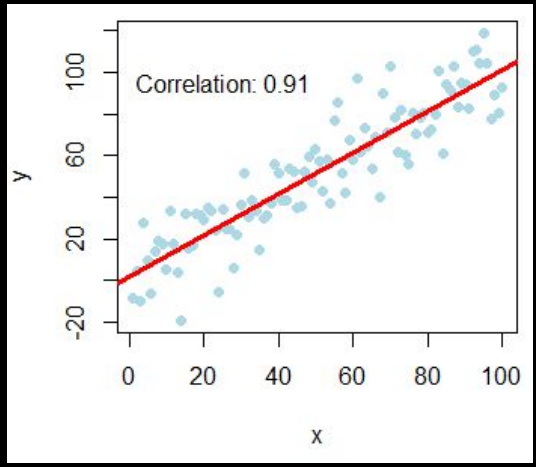
I quartili dividono un set di dati in 4 parti uguali e si riferiscono ai valori del punto tra i quarti. Il Quartile inferiore (Q1) è il punto tra il 25% più basso di valori e il 75% più alto di valori. È anche chiamato il 25 ° percentile. Il secondo quartile (Q2) è il centro del set di dati. È anche chiamato 50 ° percentile, o mediana. Il quartile superiore (Q3) è il punto tra il 75% più basso e il 25% più alto di valori. È anche chiamato il 75 ° percentile.

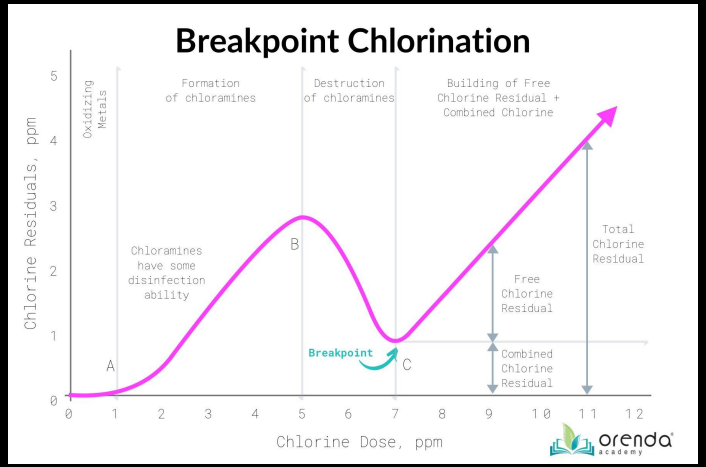


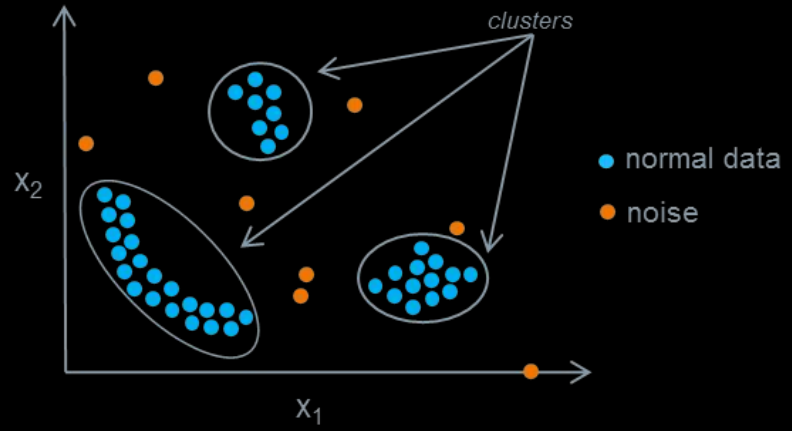
What can we do with our collected data?

Breakpoints/Anomalies/Surprise/Novelty Discovery

Correlation Discovery







Breakpoints/Anomalies/Surprises/Novelties Discovery

Introduction

Discover structural changes, novelties, surprises, anomalies in data

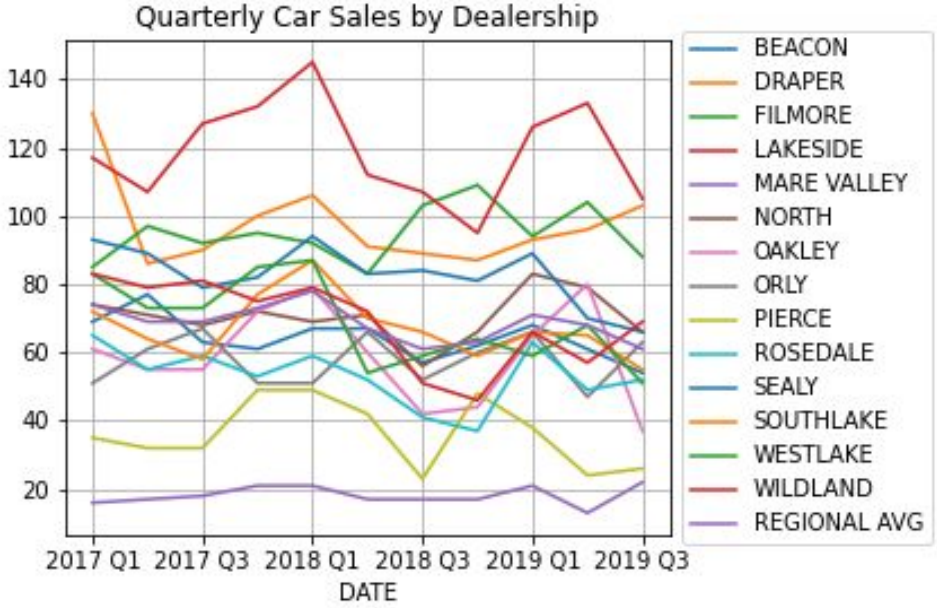
Identify conflicts in data

Conflict — a fact, which happens and changes the current situation. In the business and financial sector, the conflict is also known as breakpoint event. You should ask what is causing this change.

Example Starting situation - after data collection

Quarterly Car Sales by Dealership

This plot is unreadable!



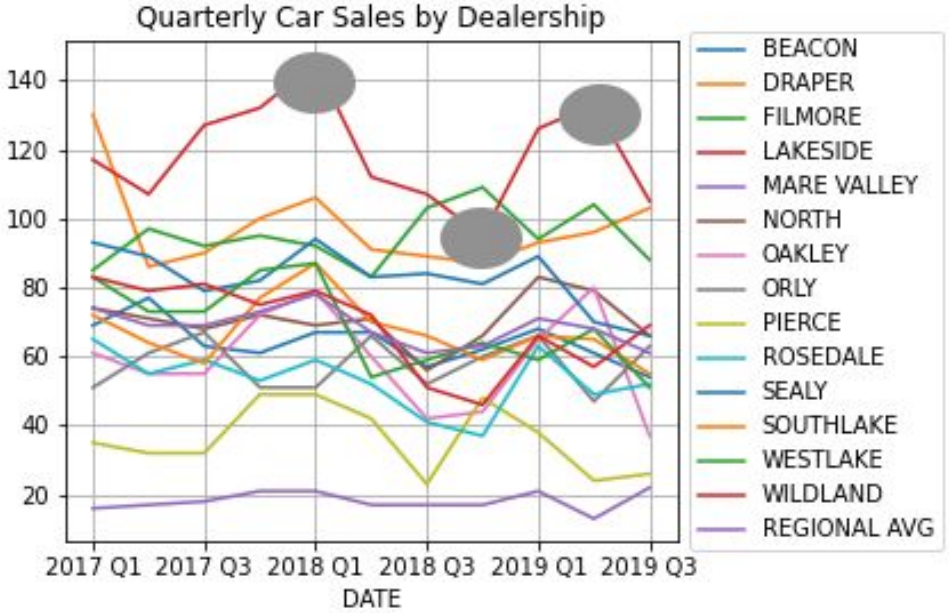
If you can’t explain it simply, you don’t understand it well enough.
ALBERT EINSTEIN

Diachronic view

In order to identify a conflict, we can look at the graph horizontally (diachronic view) or vertically (synchronic view).

By looking at the graph horizontally, we search for changes over the time.

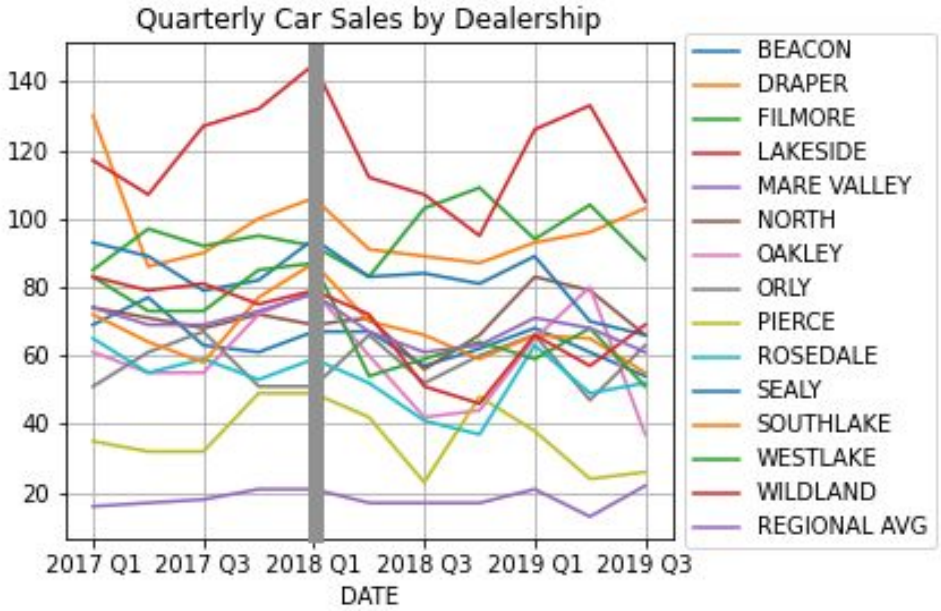
For example, with respect to Lakeside, we can identify different conflicts, which correspond to peaks.



Synchronic view

Analysing the plot vertically, we search for the behaviour of the different dealerships.

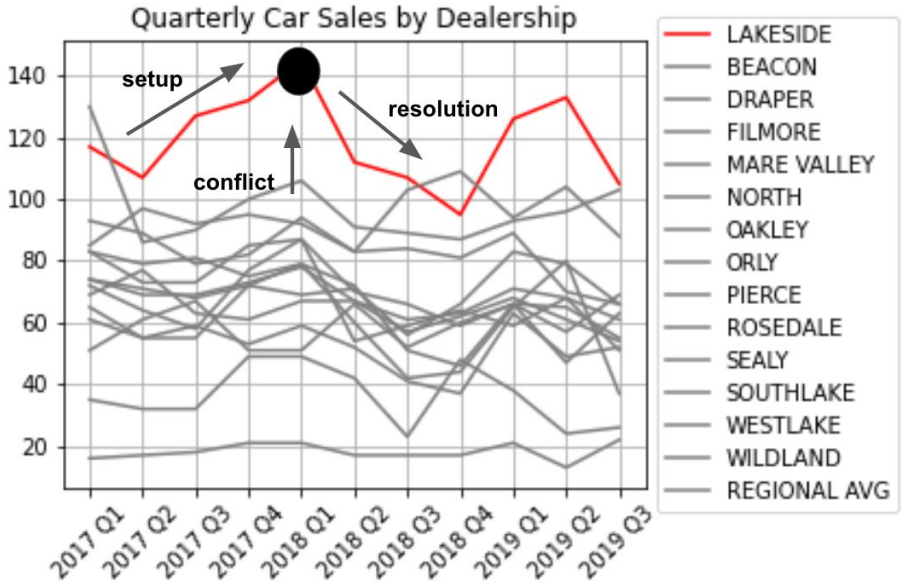
For example, in the first quarter of 2018 almost all cars have a peak, followed by a drop. Why?



Every conflict should be analysed separately, i.e. it should be represented by a different narrative and series of plots.

Setup - Conflict - Resolution

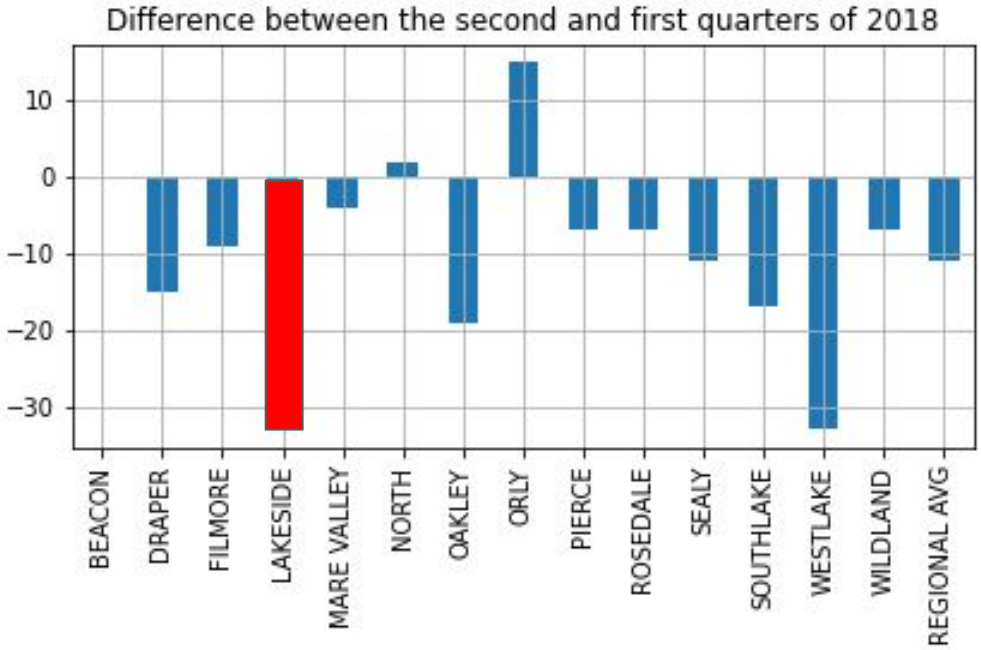
We note that the Lakeside sales increase up to the first quarter of 2018, when something happens and then they decrease until the fourth quarter of 2018, when something else happens and Lakeside sales begin again to increase.



And now?

Calculate the difference between after and before the conflict.

Almost all the dealerships have a negative value, except for Orly and North. The Lakeside and Westlake sales experience the worst situation.



And now? (cont.)

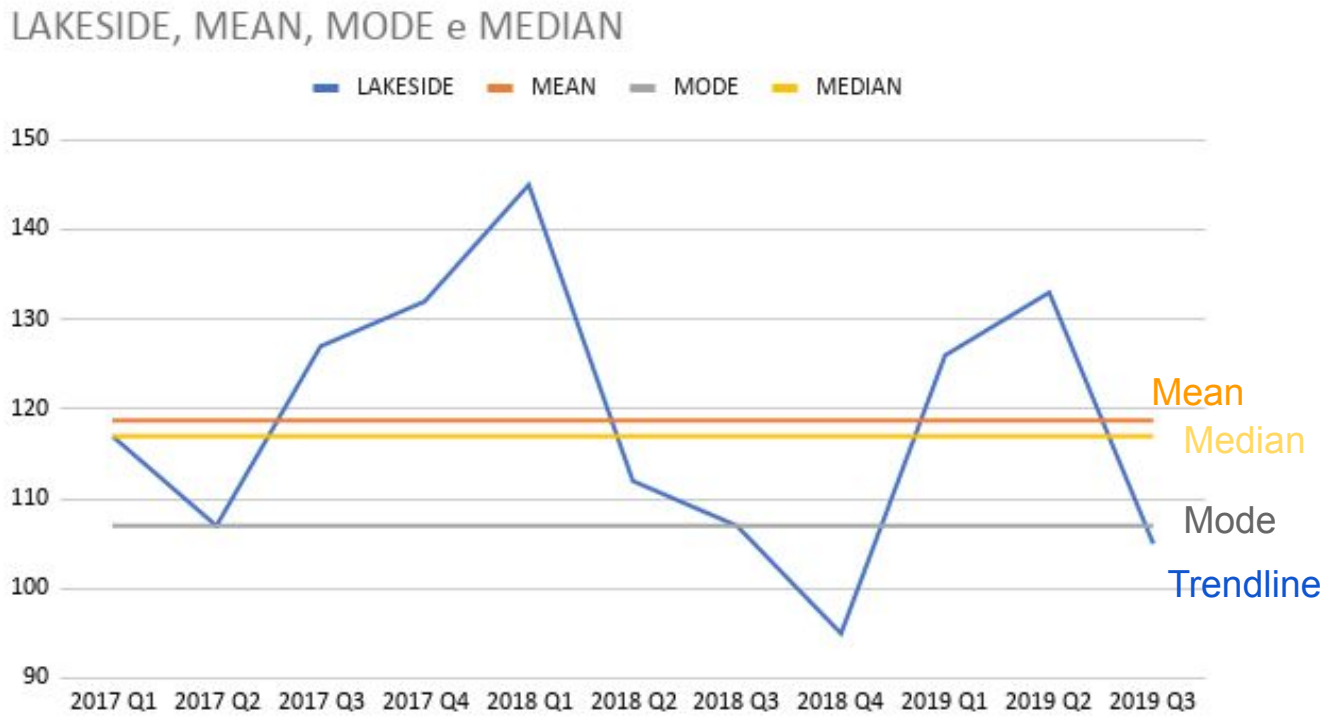
What happened in the first quarter of 2018, which negatively influenced car sales? Let us try to google and search for the answer.

I found this interesting article and this report, which explain that in 2018 there was an incredibly increase of light trucks sales (about 70%), which produced a decrease of car sales.

Orly (O'Reilly Automotive), instead, did not experience this decrease because of a different policy.

And now? (cont.)

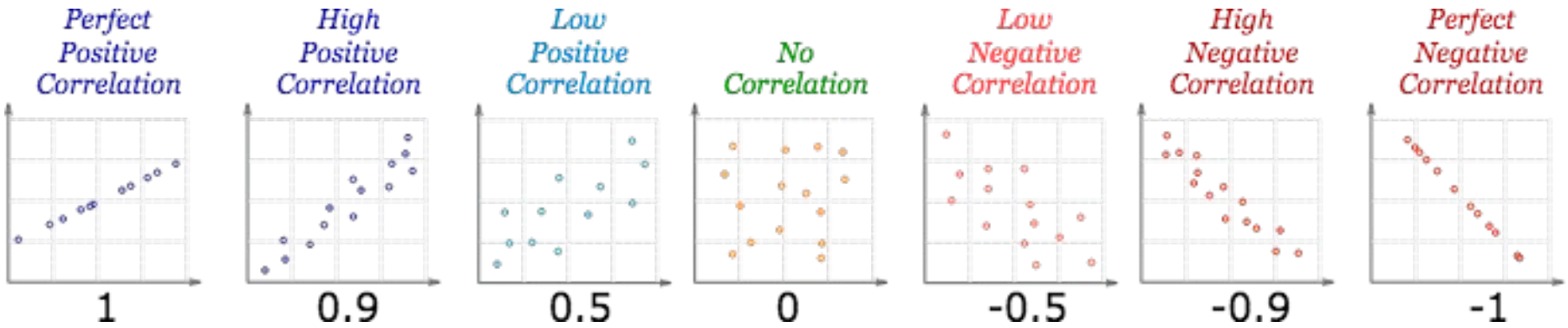
Calculate the difference between the conflict and the mean / mode /median value



Correlation Discovery

Correlation

Discover if two observations are correlated.



Pearson Coefficient

It shows the linear relationship between two sets of data.

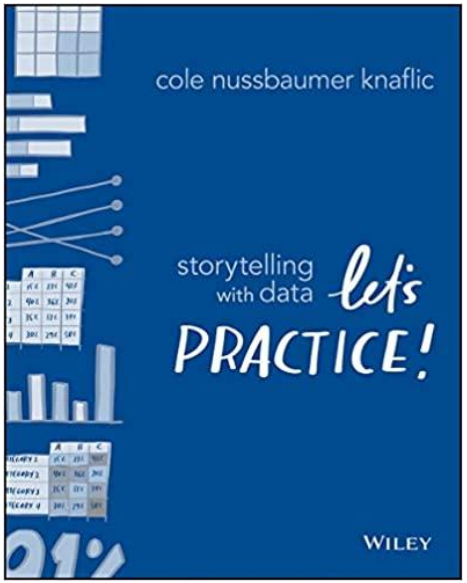
```
import numpy as np

x_simple = np.array([-2, -1, 0, 1, 2])

y_simple = np.array([4, 1, 3, 2, 0])

my_rho = np.corrcoef(x_simple, y_simple)
```

Examples



Extracted from [Storytelling With Data. Let's practice]

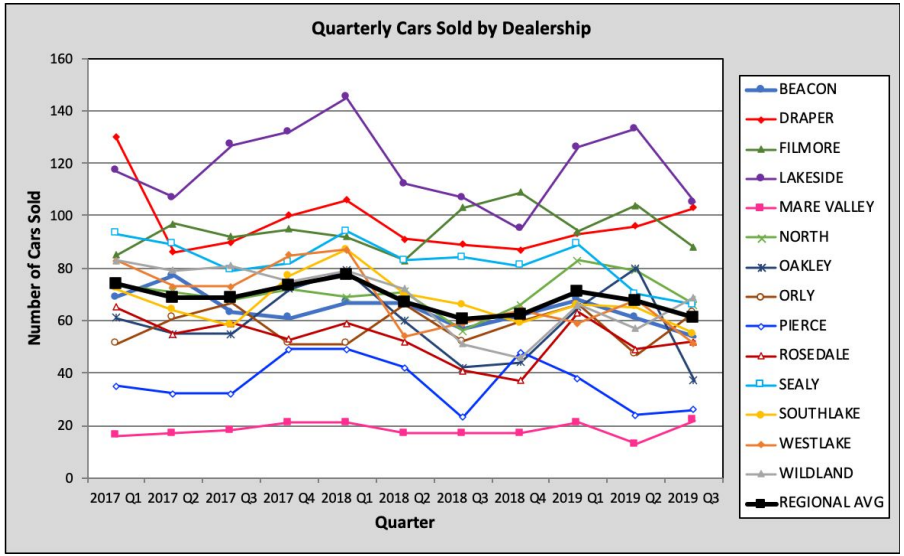
Example 1

https://community.storytellingwithdata.com/exercises/one-little-changeand-a-redesign

Solution

https://docs.google.com/spreadsheets/d/1UTUYLlrPO188ftgB3VZ1lAWhAeQ-hnnb/edit#gid=1631368711

https://docs.google.com/presentation/d/17lKG9Mp3bRr8X_9IjKr9xdp7FCacUzOm/edit#slide=id.p1

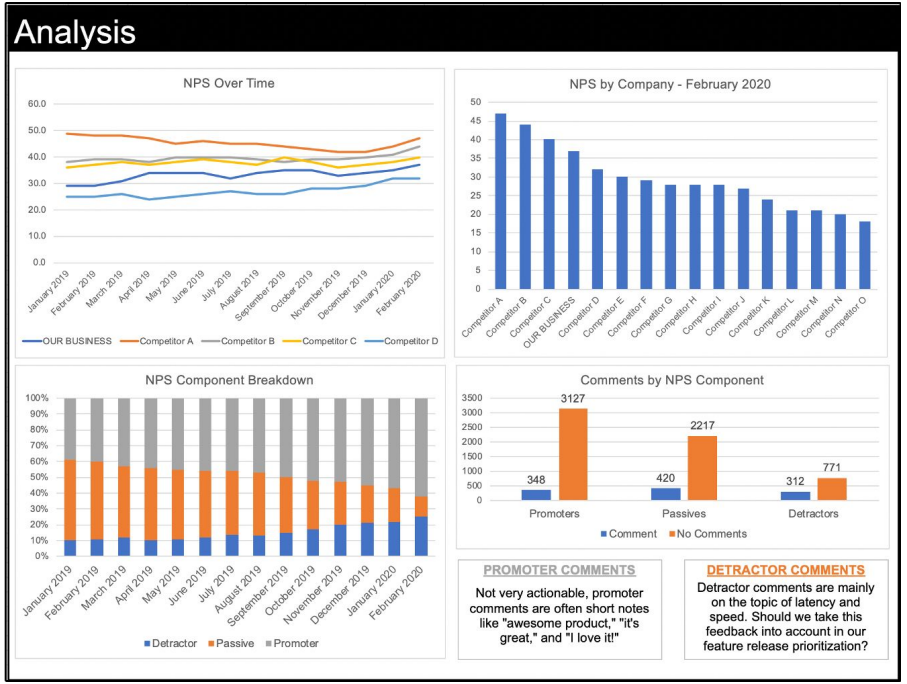


Exercise 2

https://community.storytellingwithdata.com/exercises/lets-give-this-slide-a-makeover

Solution

https://drive.google.com/file/d/1UYfZRpqLveC2n1FU57_nFZnBHFVz342A/view



Example 3

https://community.storytellingwithdata.com/exercises/table-takeaways

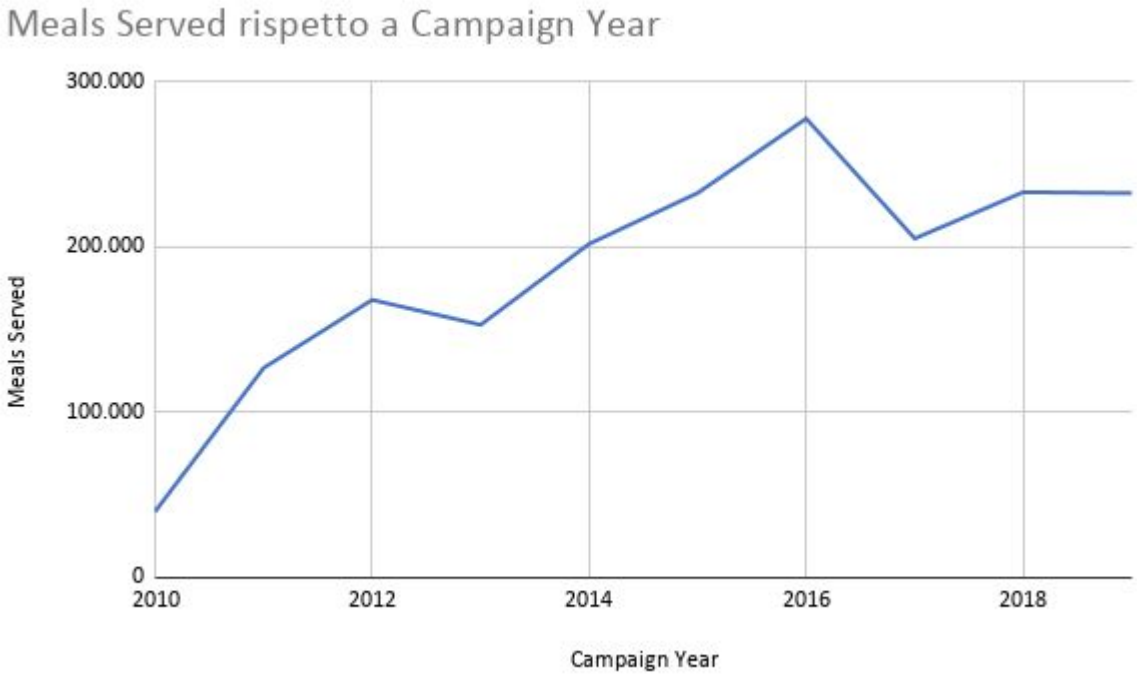
Meals served over time

Campaign Year	Meals Served
2010	40,139
2011	127,020
2012	168,193
2013	153,115
2014	202,102
2015	232,897
2016	277,912
2017	205,350
2018	233,389
2019	232,797
2020	154,830

Example 3 (cont.)

Solution

https://community.storytellingwithdata.com/videos/become-a-data-viz-superstar-part-1



More resources

Storytelling with Data

Exercises with solutions

Exercises from community (requires registration)

Learning Videos