Correlation and Covariance. Visualization.

Nazgul Rakhimzhanova

COURSE SCHEDULE

week	Mid Term (weeks 01-07)	End Term (weeks 08-14)	week
01	Intro: Data Science Area and open source tools for Data Science	Mid Semester Feedback	08
02	NumPy package for data science	Correlation and Covariance	09
03	Pandas package for data science	Sampling and Estimation	10
04	Visualization with matplotlib	Hypothesis testing	11
05	Statistics: Distribution – Normal	Decision Tree	12
06	Exploratory Data Analysis (EDA)	Linear Regression	13
07	**Summary for 6 weeks QA session**	**Summary for 6 weeks QA session**	**14**
15	**Course summary**		

IDEA

- Two variables can have some **relationship**
- Example of relationship,
  - Variable **Experience**, **years**
  - Variable **Salary**, **tenge**

Experience	Salary
0,5	100 000
1	130 000
1,5	150 000
2,5	200 000

IDEA

- So, **Experience** and **Salary** seems to have a relationship – we see that with growth of Experience the Salary grows too.
- **For every 6 month of experience the Salary grows on 20 000 – 30 000 tenge.**

Experience	Salary
0,5	100 000
1 (+0,5)	130 000 (+30 000)
1,5 (+0,5)	150 000 (+20 000)
2,5 (+1)	200 000 (+50 000)

IDEA

- So
- We can somehow **measure** this relationship
- We can calculate a **coefficient** of this relationship.
- We can **use this coefficient** to derive some knowledge (perform analysis)

IDEA

- Correlation – statistical technique which determines how one variables **moves/changes** in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.
- Covariance – it says, two variables are related based on how these variables change in relation with each other. If two variables change in the same direction, like Exp. Grows and Salary grows too.

IDEA

- Correlation, Covariance…
- How can we use it?

CORRELATION, WHY?

- If two variables are closely correlated, then **we can predict one variable from the other.**
- Correlation plays a vital role in identifying the **important variables** on which other variables depend.
- It’s used as the foundation for various modeling techniques.
- Proper correlation **analysis leads to better understanding of data.**
- Correlation contribute towards the understanding of causal relationship(if any).

CORRELATION

- Brain size & IQ
- Chocolate consumption & Weight
- Name & Weight

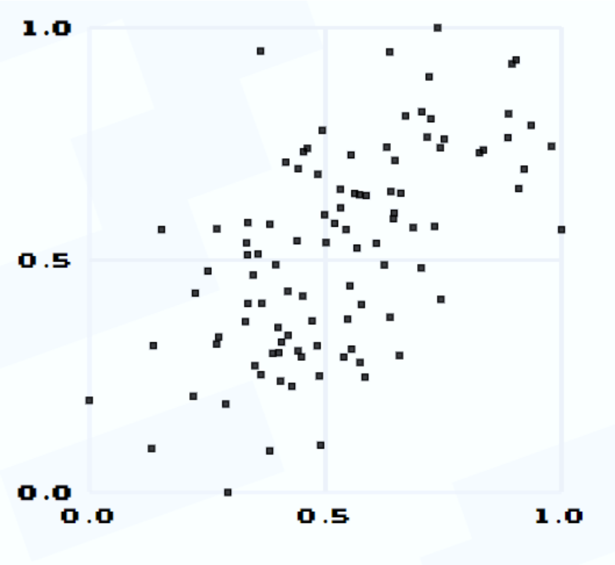
- **Which pair can be correlated?**

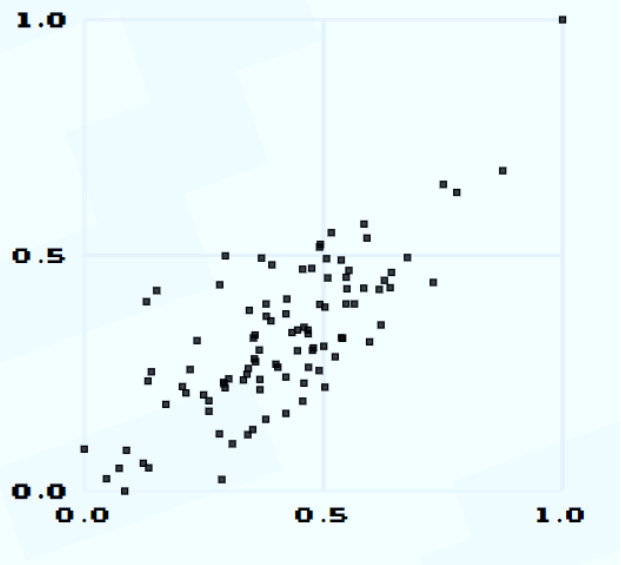
CORRELATION

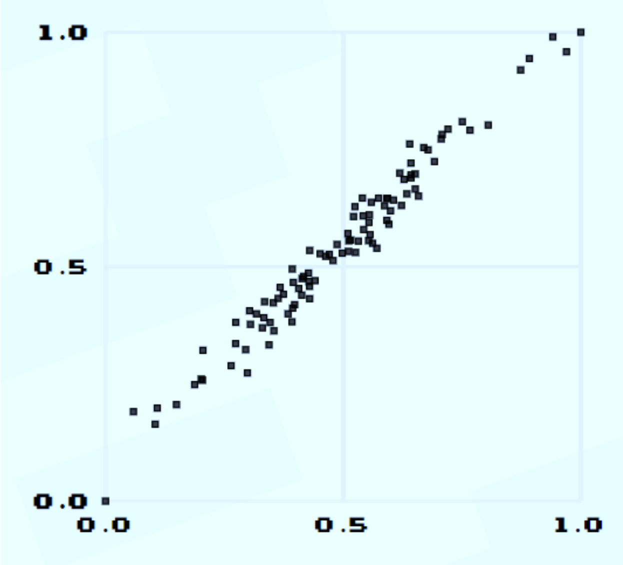
How to visualize?

- **Frequency table** – nominal/ordinal data
- **Scatter plots** – ratio/interval data

CORRELATION

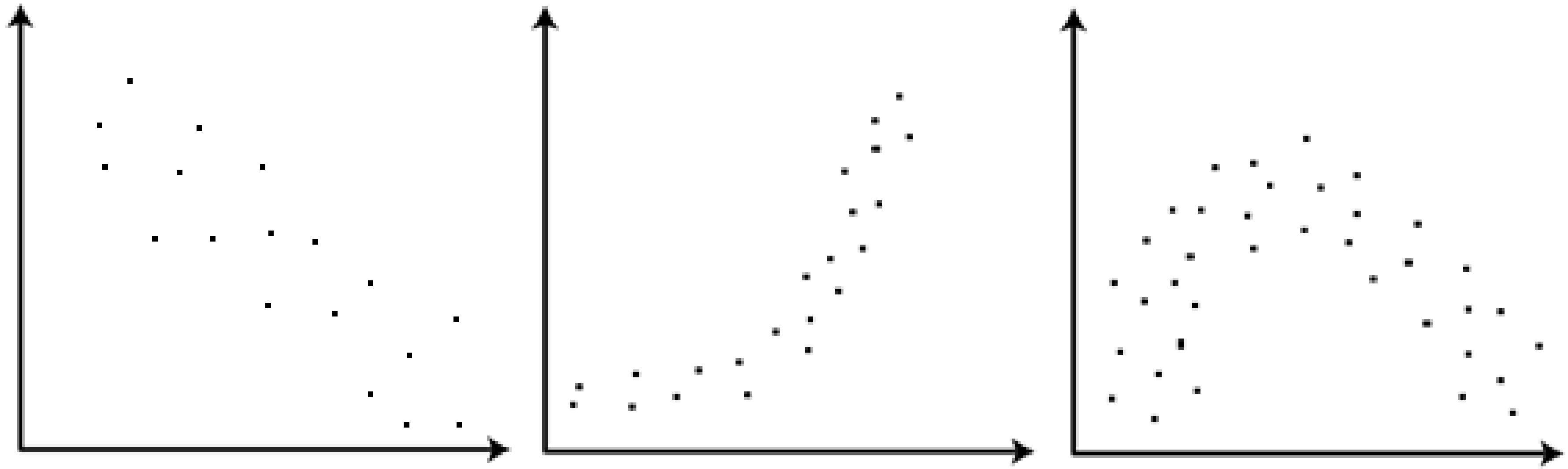






Try Urself: http://guessthecorrelation.com/

CORRELATION



CORRELATION COEFF.

- **Pearson's** - captures the strength and direction of the linear association between two continuous variables. It tries to draw the line of best fit through the data points of two variables. Pearson correlation coefficient indicates how far these data points are away from the line of best fit
- **Spearman** – tries to determine the strength and the direction of the monotonic relationship which exists between two ordinal or continuous variables. In a monotonic relationship two variables tend to change together but not with the constant rate.
- **Kendall**

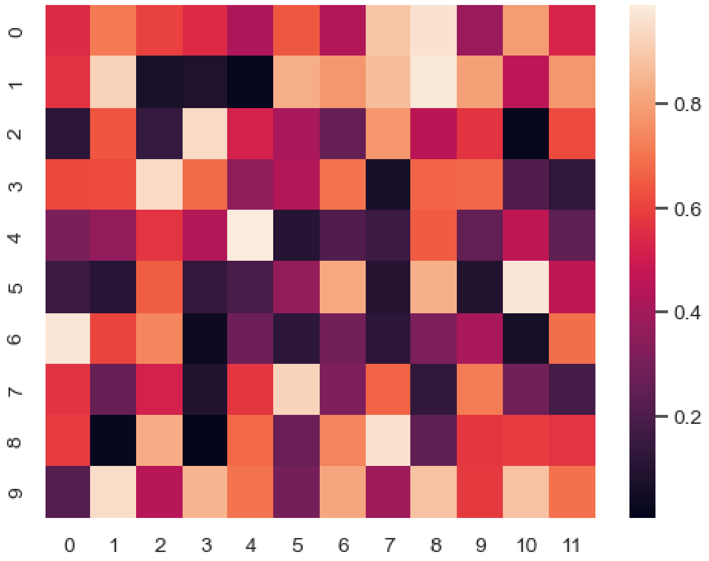
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html

CORRELATION & CAUSATION

- **"Correlation does not imply causation!"**
- Think about if video games cause violence. Let's think:
  - Violent people can play more video games.
  - Video games can cause more violence in people.
  - People are violent, some of them play video games.
  - Which conclusion is correct? We have no idea. Because, Correlation does not imply causation.

CORRELATION VISUALIZATION

- The easiest ways to visualize the correlation are – **scatter** plot and **heatmap**.
- Use seaborn or matplotlib.pyplot scatter plot
- Use seaborn.heatmap()



READINGS

- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://seaborn.pydata.org/tutorial/relational.html
- https://likegeeks.com/seaborn-heatmap-tutorial/