Learning Objectives

Scatter Plot

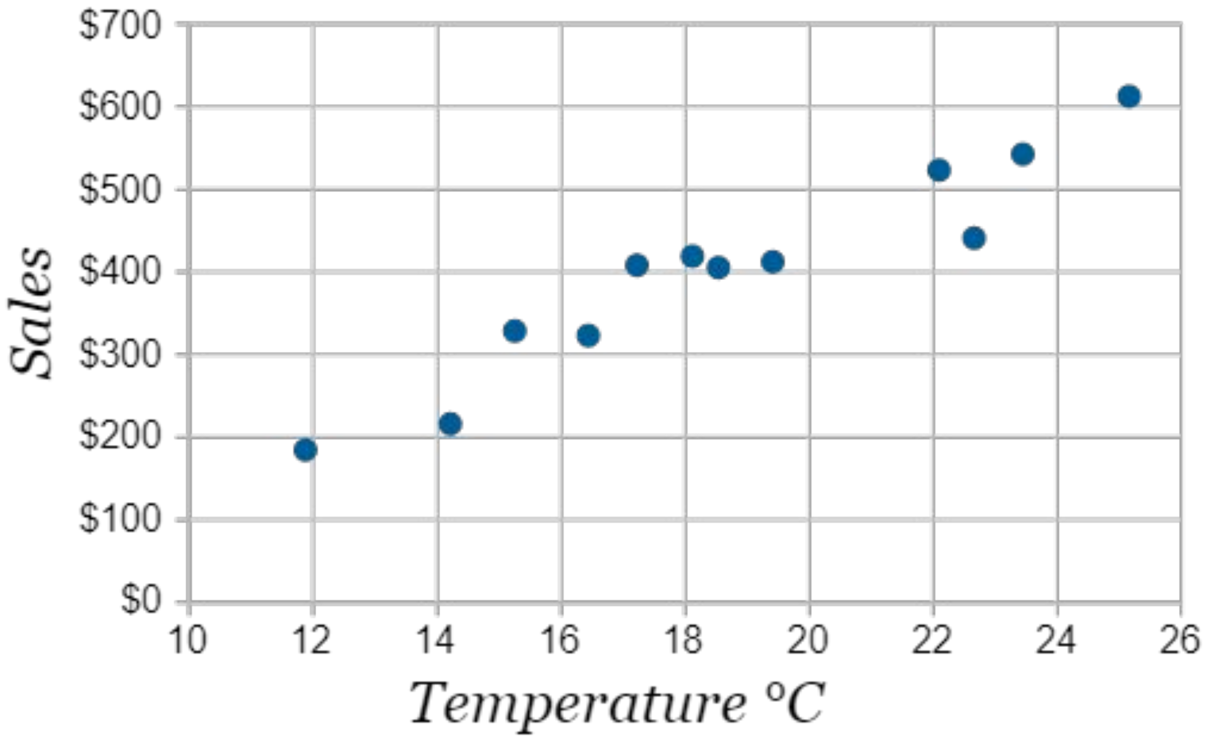
Outliers

Correlation in Data Science

DPhi

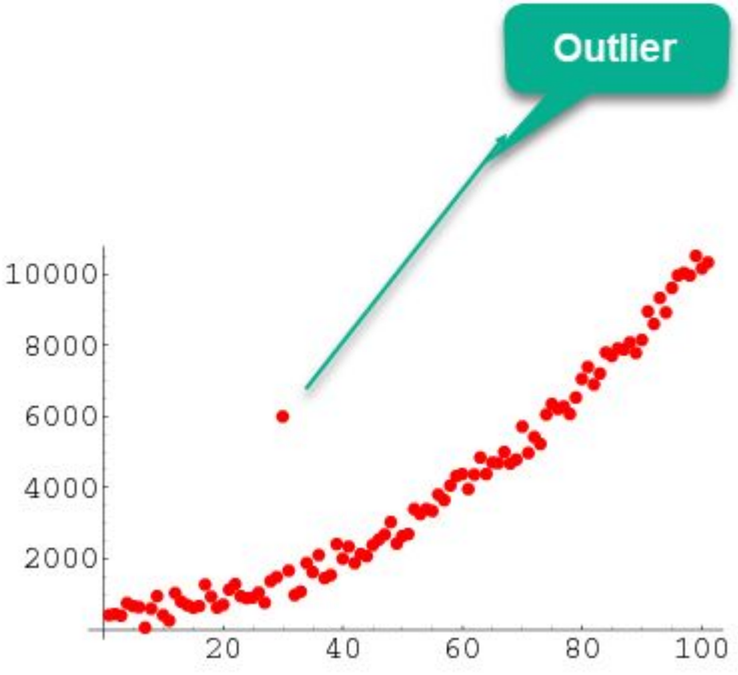
Scatter Plot

Scatter plots are used for interpreting trends in data. Below is an example of a scatter plot between temperature and ice cream sales in dollars. **What is the trend in this scatter plot?** Roughly we can say that as temperature increases ice cream sales increases.



Outlier

In statistics, an outlier is a data point that differs significantly from other observations.



Correlation

Correlation is a statistical measure.

It is a measure of the strength of a linear relationship between two quantitative variables

Now you may ask **what is a variable?** - If we go back to scatter plot example: temperature and ice-cream sales are variables. Variable is often interchangeable used as features too.

Target variable - In data science, The “target variable” is the variable whose values are to be modeled and predicted by other variables in the dataset.

Importance of Correlation

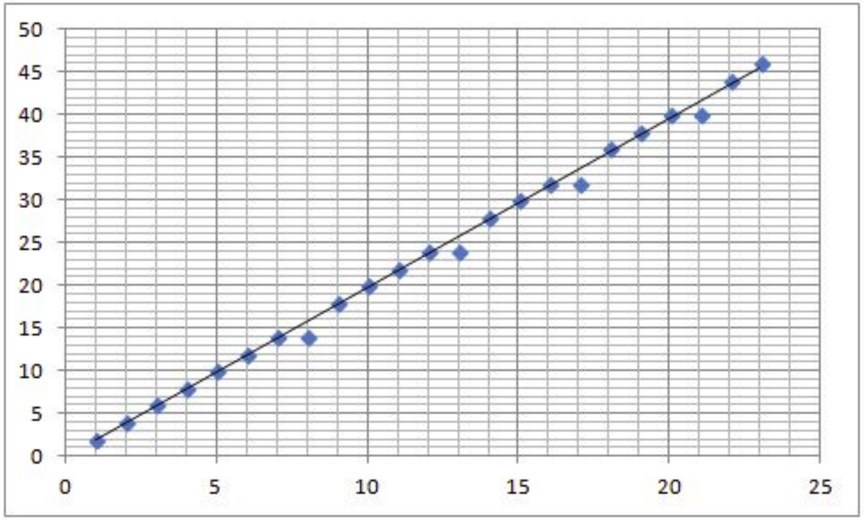
Every single successful data science project revolves around finding accurate correlations between the input and target variables. However more than often, we oversee how crucial correlation analysis is.

It is recommended to perform correlation analysis before and after data gathering and transformation phases of a data science project.

Positive Correlation

Two features (variables) can be positively correlated with each other. It means that when the value of one variable increases then the value of the other variable(s) also increases (also decreases when the other decreases).

Strong Positive Correlation:



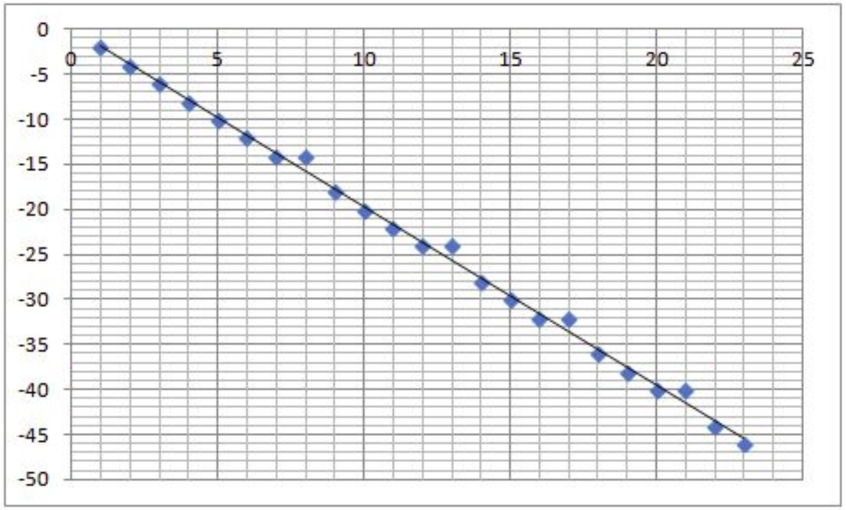
Real life examples:

- The more time you spend running on a treadmill, the more calories you will burn.
- As the temperature goes up, ice cream sales also go up.
- As the level of water lowers in a fish tank, the volume of the habitat for the fish decreases.

Negative Correlation

Two features (variables) can be negatively correlated with each other. This occurs when the value of one variable increases and the value of other variable(s) decreases (inversely proportional).

Strong Negative Correlation:



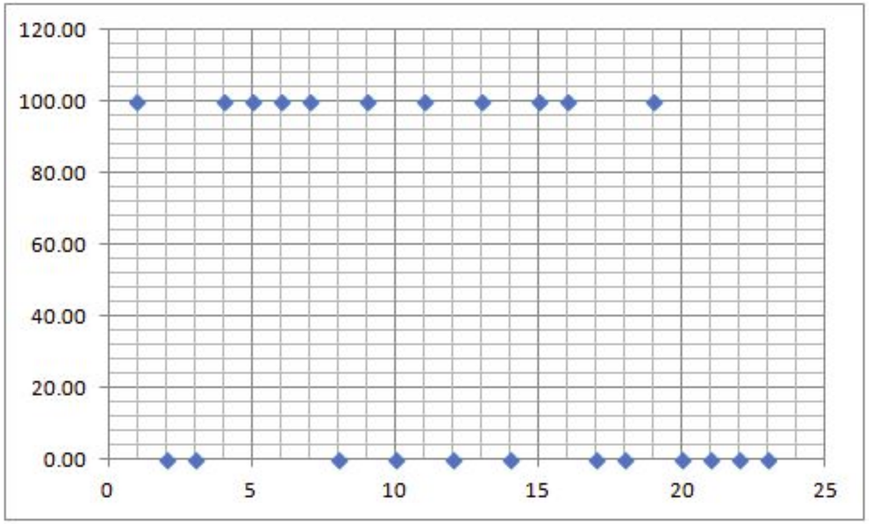
Real life examples:

- As the weather gets colder, air conditioning costs decrease.
- The more vitamins one takes, the less likely one is to have a deficiency.
- The more one works, the less free time one has.

Zero/No Correlation

Two features might not have any relationship with each other. This happens when the value of a variable is changed then the value of the other variable is not impacted.

No Correlation:



Real life examples:

- There is no relationship between the amount of tea drunk and level of intelligence.
- It was raining this morning and the grocery store was out of bananas.
- The temperature in Mars and the stock market have an almost zero correlation because the price of the stock market will not depend on the temperature in Mars.

Notebook and Dataset

- Next, we'll be looking at a pre-recorded session on Data Visualization with Matplotlib and Pandas

- **Link to the Notebook:** **https://dphi.tech/notebooks/853/manish_kc_06/day-4-notebook-data-visualization-in-python**

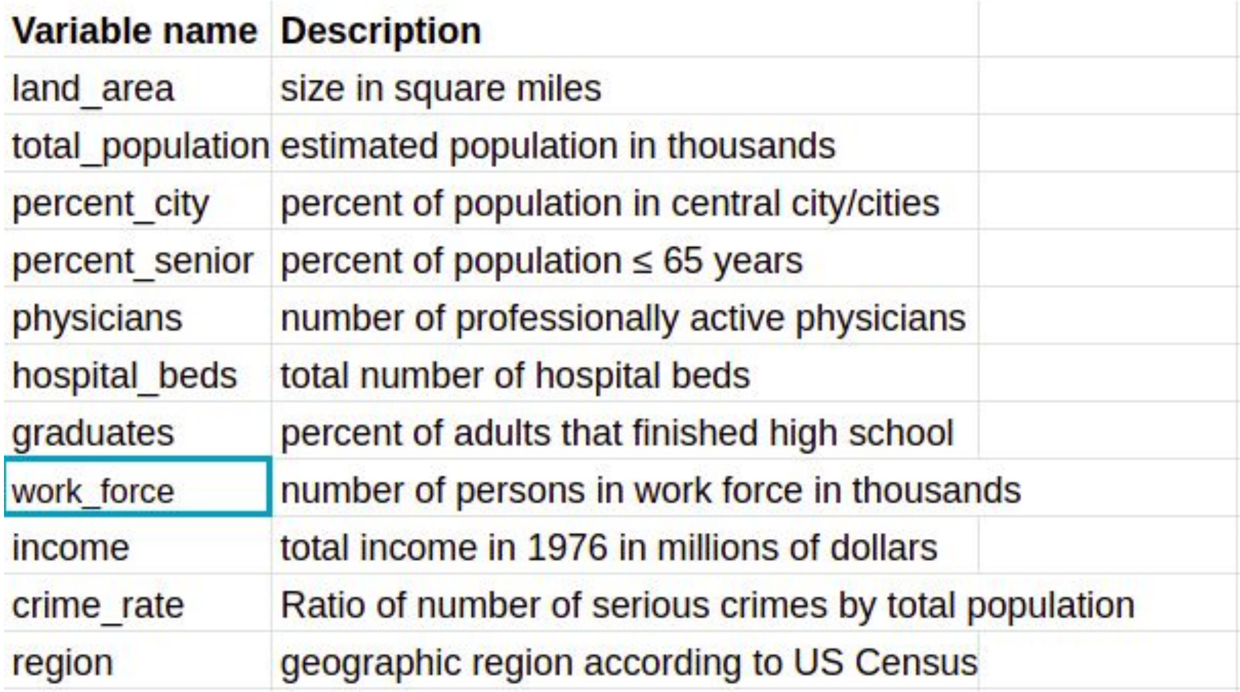
- **Link to the Dataset:** **https://github.com/dphi-official/Datasets/blob/master/Standard_Metropolitan_Areas_Data-data.csv**

- Go through the dataset and try to understand what the columns represent.

DPhi

Dataset Description

It contains data of 99 standard metropolitan areas in the US. The data set provides information on 10 variables for each area for the period 1976-1977. The areas have been divided into 4 geographic regions: 1=North-East, 2=North-Central, 3=South, 4=West. The variables provided are listed in the table below:



Variable name	Description
land_area	size in square miles
total_population	estimated population in thousands
percent_city	percent of population in central city/cities
percent_senior	percent of population ≤ 65 years
physicians	number of professionally active physicians
hospital_beds	total number of hospital beds
graduates	percent of adults that finished high school
work_force	number of persons in work force in thousands
income	total income in 1976 in millions of dollars
crime_rate	Ratio of number of serious crimes by total population
region	geographic region according to US Census

Slide Download Link

You can download these slides from the below link:

https://docs.google.com/presentation/d/1GiYoMN9Zly4r12RR_0Iwd6IkzMwUYwMTeo4L5R9UA1A/edit?usp=sharing

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.