Day 10 of #100daysofmathandstats: Exploration of binary and categorical data (Contd...)

By Harsh Kathiriya

Outline

- Correlation introduction
- Correlation coefficient
- Correlation matrix
- Scatterplot

Correlation Introduction

- Exploratory data analysis in many modeling projects involves examining correlation among predictors, and between **predictors and a target variable.**
- Variables X and Y (each with measured data) are said to be **positively correlated** if high values of X go with high values of Y, and low values of X go with low values of Y
- If **high values of X** go with **low values of Y,** and vice versa, the variables are negatively correlated.

Correlation coefficient

- A metric that measures the extent to which numeric variables are associated with one another **(ranges from –1 to +1).**
- To compute **Pearson’s correlation coefficient**, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y})}{(n-1)s_x s_y}$$

Real examples of correlation coefficient

- Positive correlations
  - The more money you save, the more financially secure you feel.
  - As the temperature goes up, ice cream sales also go up.
  - The more gasoline you put in your car, the farther it can go.
- Negative correlations
  - Time Spent Watching TV vs. Exam Scores
  - Time Spent Running vs. Body Fat
  - Less study time vs chance of getting high scores

Correlation matrix

- A **table** where the variables are **shown on both rows and columns,** and the cell values are the correlations between the variables.
- Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.

Example of correlation matrix

T	CTL	FTR	VZ	LVLT	
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

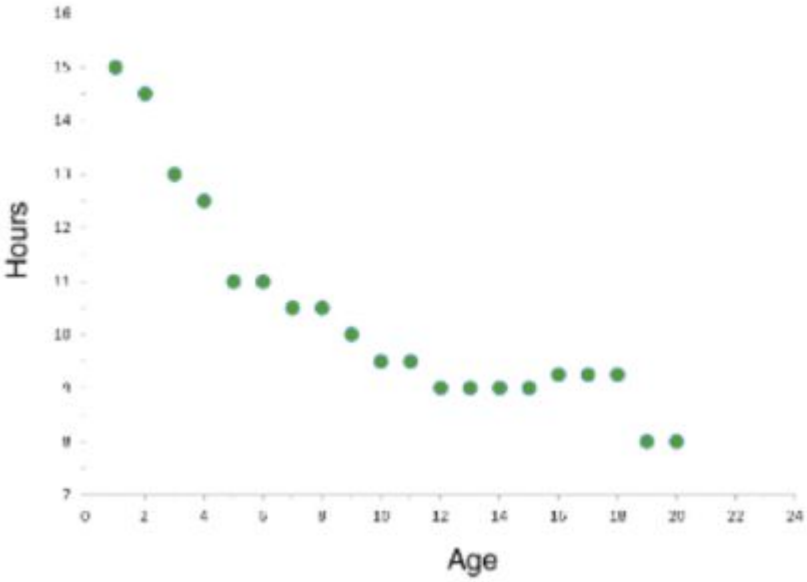
- Table shows the correlation between the daily returns for telecommunication stocks from July 2012 through June 2015.

Scatter Plot

- A plot in which the x-axis is the value of one variable, and the y-axis the value of another.
- The **x-axis represents one variable** and the **y-axis another**, and each **point** on the graph is a **record**
- This is very useful when we wants to see some relations between features.

Real example of scatter plot

This is a scatter plot showing the amount of sleep needed per day by age.



Careers that use scatter plot a lot

- Economist
- Operations research analyst
- Market research analyst
- Management analyst
- Data Scientist

Thank you

Github Link: https://github.com/harsh9898/100daysofstatandmath

Don’t forget to post your queries or feedbacks on the post.

Share or like for the benefit of others.