The Learning Hub

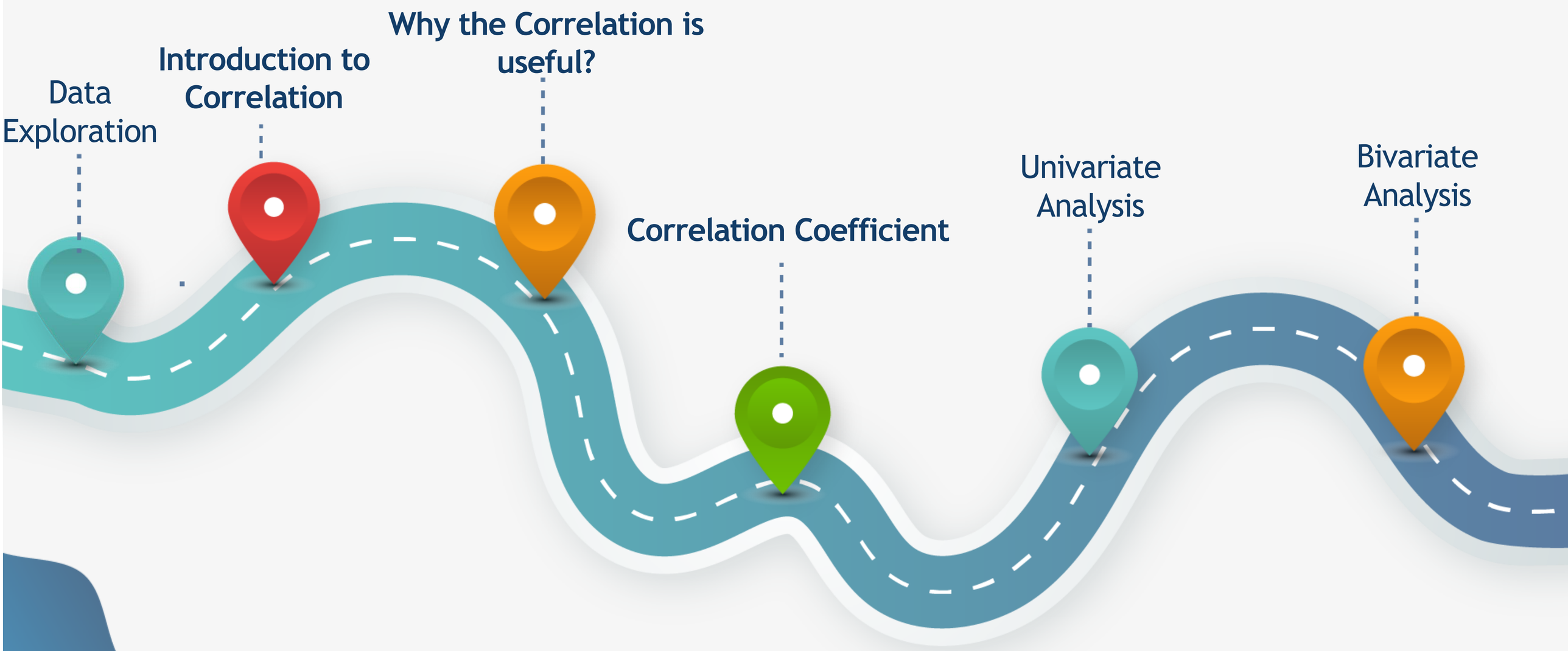
Data Science Advanced

Data Exploration

Education and Training Solutions

2023

The Learning Hub



Data Exploration

Data Exploration

Data Exploration (Exploratory Data Analysis EDA), is the first step in data analysis where users look at and understand their data with statistical techniques and visualization tools. This step helps identify patterns and problems in the dataset.

Why is data exploration important?

- Humans are visual learners, suitable to reuse visual data much more easy than numerical data. Accordingly, it's challenging for data scientists to review thousands of rows of data points and infer meaning without backing.
- Data visualization tools and rudiments like colors, shapes, lines, graphs and angles aid in effective data disquisition of metadata, enabling relations or anomalies to be detected.

How machine learning is applied to data exploration

- Machine learning can significantly assist in data exploration when big quantities of data are involved.
- for a machine learning model to be precise, data analysts must take the following steps before performing the analysis.

- Identify and define all variables and their role in the data set.
- apply univariate analysis for continuous variables, using a histogram, box plot, or scatter plot. For categorical variables(those that can be grouped by category), bar maps can be used.

How machine learning is applied to data exploration

- applying bivariate analysis, to determine the relationship strength between variables. This can be completed by applying data visualization tools, like Tableau.
  - Continuous and Continuous: scatter plots
  - Categorical and Categorical: stacked column chart
  - Categorical and Continuous: boxplots combined with swarm plots.

- Account for any missing values and outliers.

How machine learning is applied to data exploration

- Homogeneity in variance
- Normally distributed data: Various statistical techniques assume normality, such as linear regressions and t-tests. Histograms can be used to show data distributions.
- Collinearity in covariates, Interaction between variables, and Independence in the dataset.

Introduction to Correlation

Introduction to Correlation

Correlation: describes the relation between variables to know how the change of one variable will affect the other.

- **Variables**: the input features that will be used to predict the target variable (label).

Examples:

1. Temperature goes up, ice cream sales go up.
2. Hair grows, more shampoo you will need.
3. Running on a treadmill, burns more calories.

Why the Correlation is useful?

Why the Correlation is useful?

- Correlation between variables leads to a better understanding of the data.
- Have a different number of modelling techniques.
- To know which important variables depend on each other.
- If two variables are correlated to each other, then we can predict one variable from the other.

Correlation Coefficient

Correlation Coefficient

A correlation coefficient : is a number between -1 and 1 that describes the relationship's strength between variables.

The rationale behind utilizing correlation for feature selection is that a good variable is closely related to the target class .

Correlation Coefficient

Correlation coefficient value	Correlation type	Description
1	Perfect positive correlation	When one variable increases, the other variables increase, and vice versa.
0	No correlation	There is no relationship between the variables
-1	Perfect negative correlation	When one variable increases, the other variables decrease, and vice versa.

Correlation Coefficient

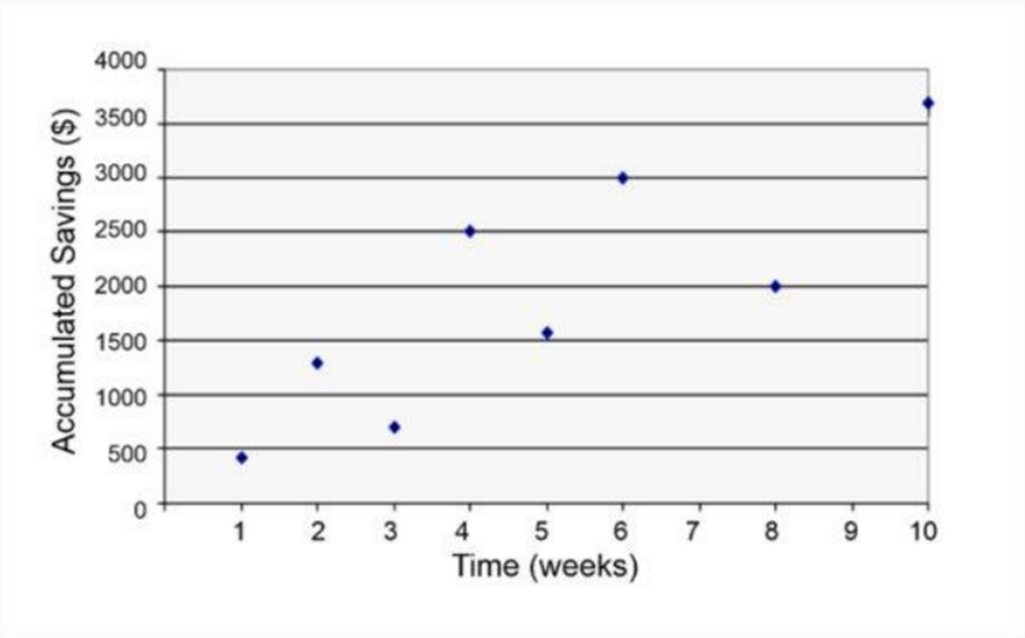
➢ **The correlation coefficient** describes how closely the data fit on a line.

To explore the correlation between two measurement variables, we can begin by creating a scatter plot. In the case of a linear relationship, a straight line of best fit will be drawn taking all data points into account on a scatter plot.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

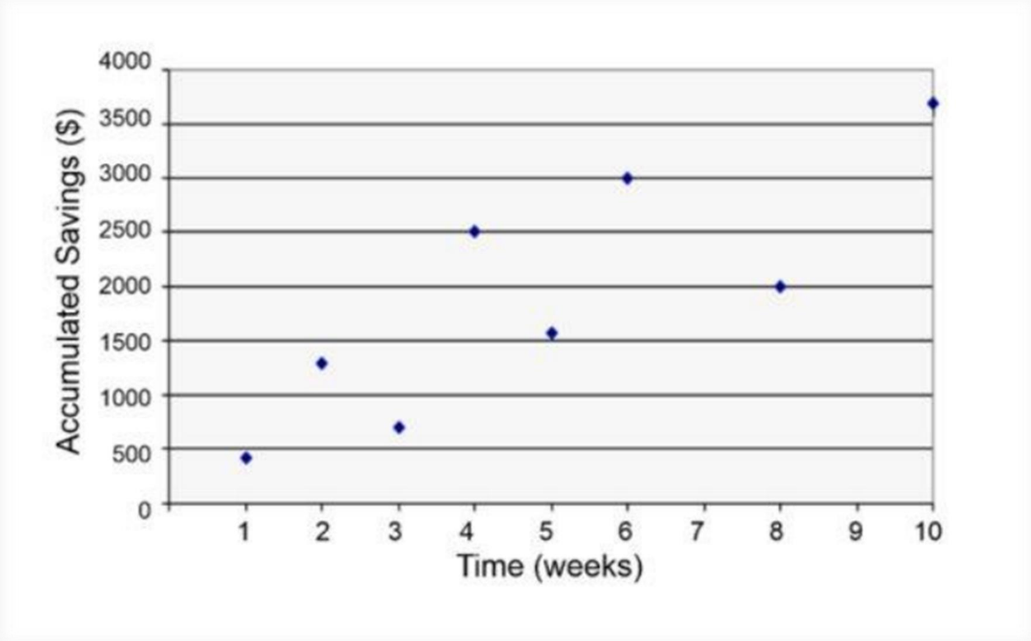
Correlation Coefficient

- The possible independent variable values are shown on the horizontal axis (the X-axis)
- The possible output values is shown on the vertical axis (the Y-axis).



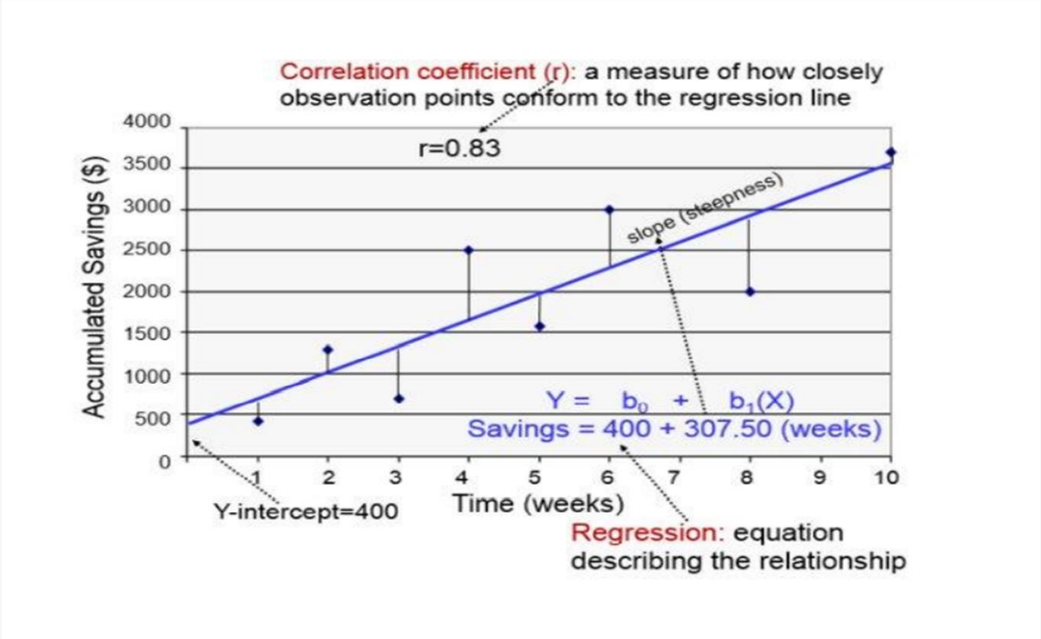
Correlation Coefficient

- That is obvious that the overall trend is a linear increase over time.



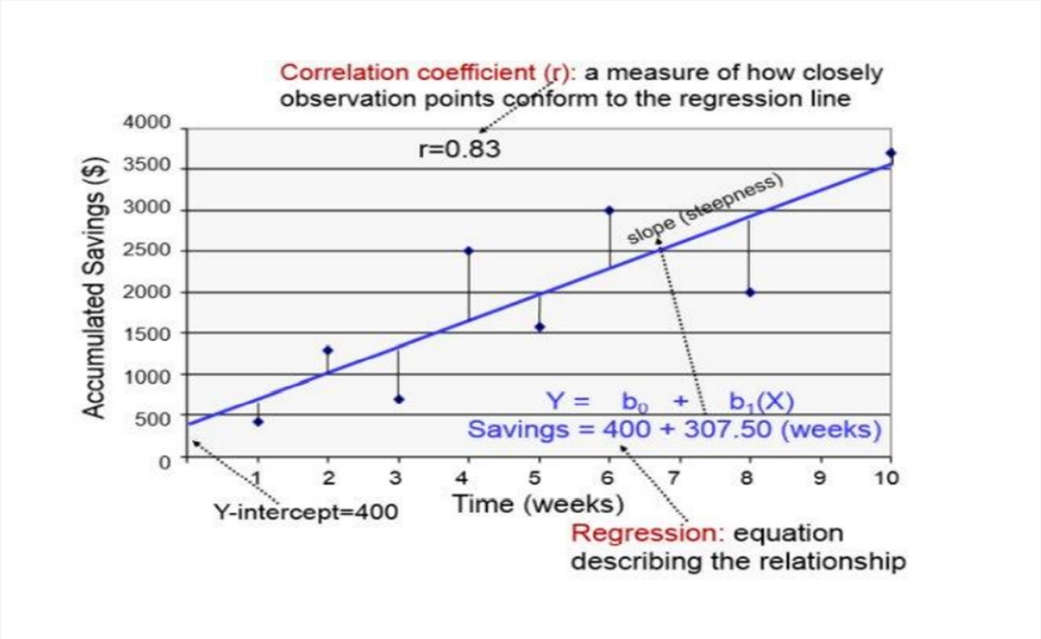
Correlation Coefficient

- In fact, the vision can be clarified
- using correlation analysis and simple linear regression analysis



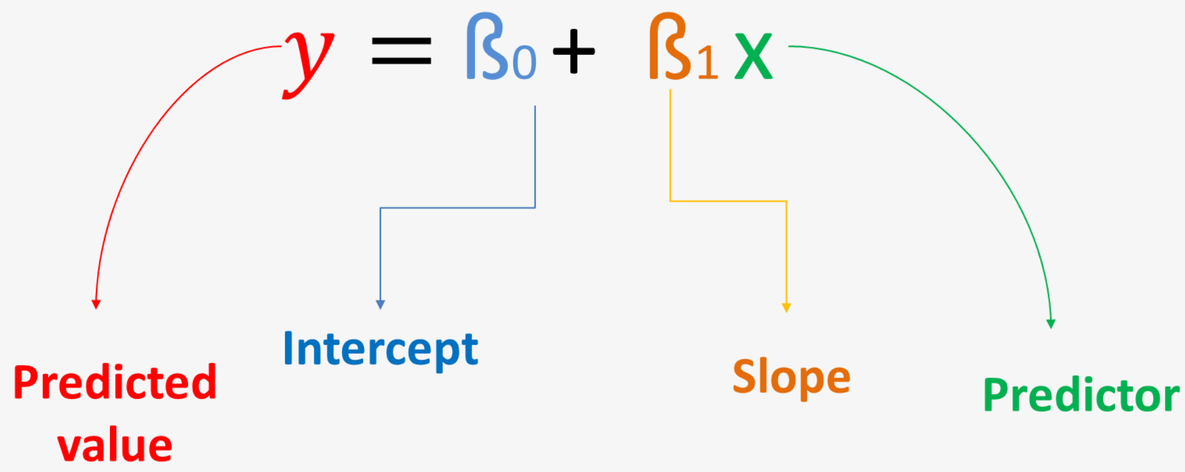
Correlation Coefficient

- The regression line is determined from a mathematical model that minimizes the errors between the scatter points and the line.



Correlation Coefficient

- In this kind of analysis the relationship between two variables can be described with a simple linear regression equation, the general form of which is:



Univariate Analysis

Univariate Analysis

- Univariate analysis is the term refers to the analysis of one variable and is the simplest form of analysing data. “Uni” means “one”, so in other words when data has only one variable.

- The major purpose of the univariate analysis is to understand the distribution of values for a single variable and describe it

- It takes data, summarizes and finds patterns in the data.

23

Why Univariate Statistics?

- The univariate analysis describes each variable in a data set on its own.
- It looks at the range of values, such as the central tendency of the values.
- It explores the pattern of the variables' response

24

How to Analyse One Variable

➢ Raw Data

Raw data typically refers to a matrix, where each row contains measurements and each column represents a variable that describes some property of each measurement.

How to Analyse One Variable

❖ **Example**: Raw data for a study of injuries among county workers (first 10 cases)

Injury Report No	County Name	Cause of Injury	Severity of Injury
1	County A	Fall	3
2	County B	Auto	4
3	County C	Fall	6
4	County C	Fall	4
5	County B	Fall	5
6	County A	Violence	9
7	County A	Auto	3
8	County A	Violence	2
9	County A	Violence	9
10	County B	Auto	3

26

How to Analyse One Variable

❖Raw Data

Gathering data for each column separately.

Severity of Injury
3
4
6
4
5
9
3
2
9
3

How to Analyse One Variable

❖ Frequency Distribution

- frequency distribution is a table that represents the frequency of various outcomes in a sample of the data for the variable,

- it can be obtained by identifying the lowest and highest values of the variable, and then putting all the values of the variable in order from lowest to highest.

- Next, count the number of the appearance of each value of the variable. This is a count of the frequency with which each value occurs in the data set.

How to Analyse One Variable

❖ Frequency Distribution

For example, for the variable "Severity of Injury," the values range from 2 to 9.

Severity of Injury	Injuries No with this severity
2	1
3	3
4	2
5	1
6	1
9	2
Total	**10**

29

How to Analyse One Variable

❖ Grouped Data

- Decide on whether the data should be grouped into classes
- The severity of injury ratings can be abbreviated into just a few categories or groups.
- Grouped data usually has from 3 to 7 groups.
- There should be no groups with a frequency of zero (for example, there are no injuries with a severity rating of 7 or 8).

30

How to Analyse One Variable

❖Grouped Data

Severity of Injury	Injuries No with this severity
Mild (1-3)	4
Moderate(4-6)	4
Severe(7-9)	2
Total	**10**

How to Analyse One Variable

❖Cumulative Distributions

Cumulative frequency distributions include a third column in the table (this can be done with either simple frequency distributions or with grouped data).

How to Analyse One Variable

❖ Cumulative Distributions

Severity of Injury	Injuries No with this severity	Cumulative Frequency
2	1	1
3	3	4
4	2	6
5	1	7
6	1	8
9	2	10

How to Analyse One Variable

❖Percentage Distributions

Frequencies can also be presented in the form of percentage distributions and cumulative percentages.

Severity of Injury	Injuries No with this severity	Cumulative Frequency
2	10	10
3	30	40
4	20	60
5	10	70
6	10	80
9	20	100

34

Bivariate Analysis

Bivariate Analysis

- Continuous-Continuous Analysis
- Continuous Categorical Analysis
- Categorical Categorical data analysis

Bivariate Analysis

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

Bivariate Analysis

Caloric intake X	Weight Y
3500	250
2000	225
4500	380
1500	110
2250	145

Types of Bivariate Analysis

- ✓ Scatter Plot
- ✓ Regression
- ✓ Correlation Coefficients

39

Continuous-Continuous Analysis

- One effective way to explore the relationship between two continuous variables is with a scatter plot.

- A scatter plot represents the observed values of a pair of variables as points on a coordinate grid.

  The values of one of the variables are aligned to the values of the horizontal axis and the other variable values to the vertical axis. A relationship between these two variables is seen as a pattern in the plotted points.

40

Continuous-Continuous Analysis

There are 2 ways to measure the correlation in the case of continuous and categorical variables:

- **point biserial correlation**

- **logistic regression**

41

The Point Biserial Correlation

The point biserial correlation ranges from -1 to +1.

The assumption of point biserial calculation is that the continuous variable is normally distributed.

Logistic Regression

- in the case of the existence of a relationship between the categorical and continuous variable, we should be able to construct an accurate predictor of the categorical variable from the continuous variable.
- If the resulting classifier has a high degree of fit, is accurate, sensitive, and specific we can conclude the two variables share a relationship and are actually correlated.

43

Categorical-Categorical data analysis

- There are two different ways to find associations between categorical variables.
  - Distance metrics such as Euclidean distance or Manhattan distance
  - Statistical metrics such as the chi-square test or Goodman Kruskal's lambda.

- A high value for chi-square means there is a low correlation between two sets of data.

Relative strengths and weaknesses

- Distance metrics are more axiom and simpler to understand. It describes if one variable can be perfectly predicted by another variable when plotted in a high dimensional space, the two variables will comp or be very close to each other.

45

Relative strengths and weaknesses

- The disadvantage of approaches relying on distance metrics is that they are sensitive to scale. For example, small changes in scale cause significant changes in distance metrics. This behaviour is not desirable to understand the goodness of fit between different features
- distance metrics are not comparable between variable pairs containing different number of categories

46

References

Deviations, O.T.S. (2018) An overview of correlation measures between categorical and continuous variables, Medium. Medium. Available at: https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365 (Accessed: January 20, 2023).

Liu, M. (2020) *2 - data exploration*, *Machine Learning Blog | ML@CMU | Carnegie Mellon University*. Available at: https://blog.ml.cmu.edu/2020/08/31/2-data-exploration/ (Accessed: January 20, 2023).

THANK YOU