Exploratory Data Analysis

Manas Dasgupta

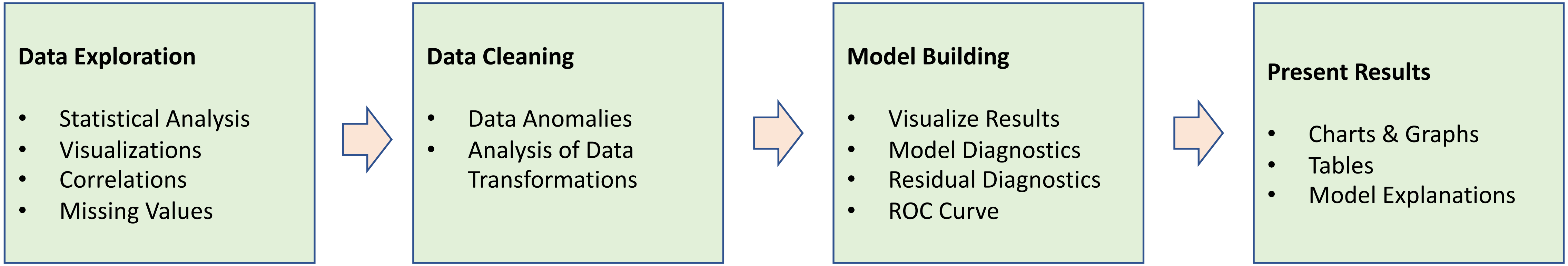
Exploratory Data Analysis (EDA)

Introduction to EDA

Exploratory Data Analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

EDA in various Phases of DS/ML



Data Exploration

- Statistical Analysis
- Visualizations
- Correlations
- Missing Values

Data Cleaning

- Data Anomalies
- Analysis of Data Transformations

Model Building

- Visualize Results
- Model Diagnostics
- Residual Diagnostics
- ROC Curve

Present Results

- Charts & Graphs
- Tables
- Model Explanations

Goals of EDA

For Direct Business Purposes

- Analysis and Finding Patterns and Insights
- Creating Reports and Visualizations

For Machine Learning

- Pre-processing Data for Machine Learning
- Establishing Assumptions about Data

Types of EDA

There are four primary types of EDA:

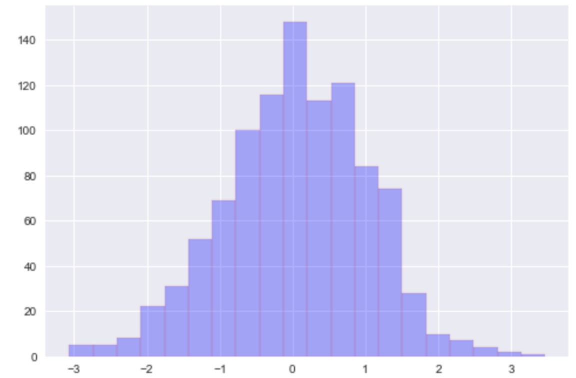
Univariate non-graphical: This is simplest form of data analysis, where the data being analysed consists of just one variable. Since it's a single variable, it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

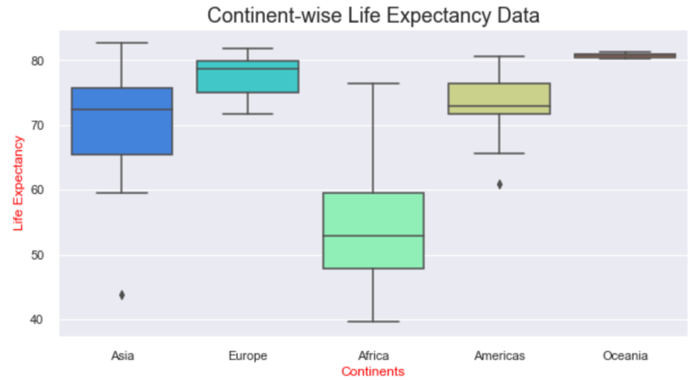
Univariate graphical: Non-graphical methods don't provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include:

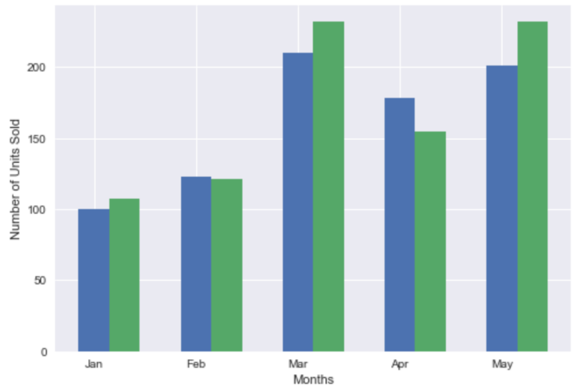
- Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values.
- Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.

Multivariate nongraphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.

Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.



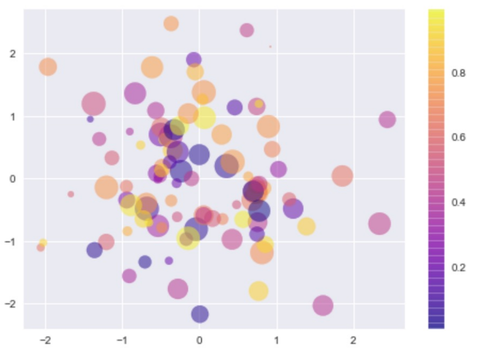


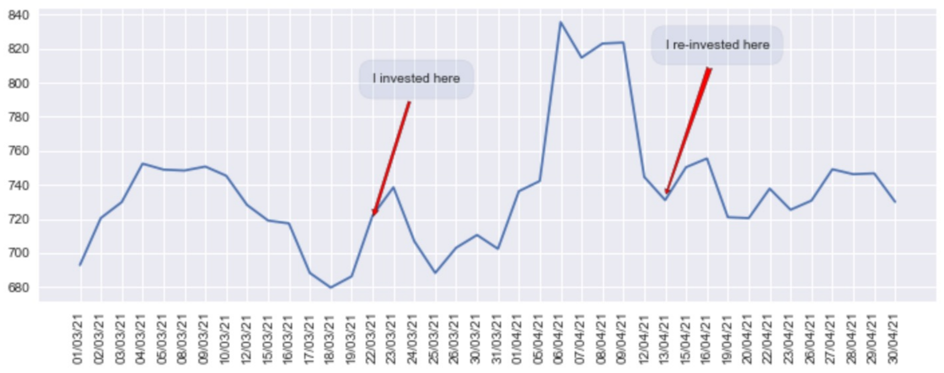


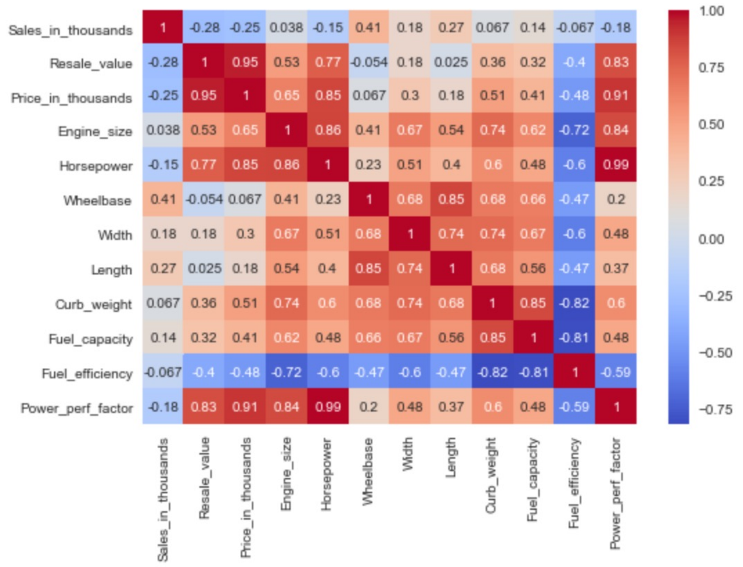
Types of EDA

Other common types of multivariate graphics include:

- **Scatter plot**, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.
- **Multivariate chart**, which is a graphical representation of the relationships between factors and a response.
- **Run chart**, which is a line graph of data plotted over time.
- **Bubble chart**, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot.
- **Heat map**, which is a graphical representation of data where values are depicted by color.







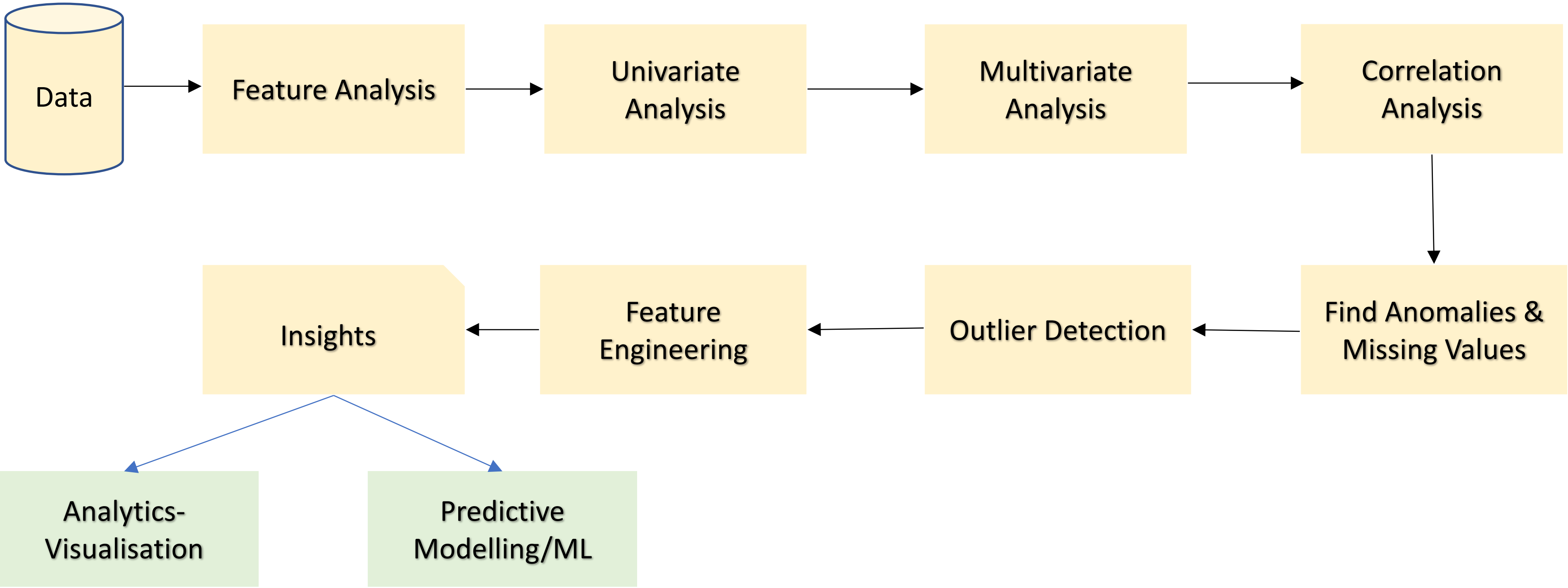
	Sales_in_thousands	Resale_value	Price_in_thousands	Engine_size	Horsepower	Wheelbase	Width	Length	Curb_weight	Fuel_capacity	Fuel_efficiency	Power_perf_factor
Sales_in_thousands	1	-0.28	-0.25	0.038	-0.15	0.41	0.18	0.27	0.067	0.14	-0.067	-0.18
Resale_value	-0.28	1	0.95	0.53	0.77	-0.054	0.18	0.025	0.36	0.32	-0.4	0.83
Price_in_thousands	-0.25	0.95	1	0.65	0.85	0.067	0.3	0.18	0.51	0.41	-0.48	0.91
Engine_size	0.038	0.53	0.65	1	0.86	0.41	0.67	0.54	0.74	0.62	-0.72	0.84
Horsepower	-0.15	0.77	0.85	0.86	1	0.23	0.51	0.4	0.6	0.48	-0.6	0.99
Wheelbase	0.41	-0.054	0.067	0.41	0.23	1	0.68	0.85	0.68	0.66	-0.47	0.2
Width	0.18	0.18	0.3	0.67	0.51	0.68	1	0.74	0.74	0.67	-0.6	0.48
Length	0.27	0.025	0.18	0.54	0.4	0.85	0.74	1	0.68	0.56	-0.47	0.37
Curb_weight	0.067	0.36	0.51	0.74	0.6	0.68	0.74	0.68	1	0.85	-0.82	0.6
Fuel_capacity	0.14	0.32	0.41	0.62	0.48	0.66	0.67	0.56	0.85	1	-0.81	0.48
Fuel_efficiency	-0.067	-0.4	-0.48	-0.72	-0.6	-0.47	-0.6	-0.47	-0.82	-0.81	1	-0.59
Power_perf_factor	-0.18	0.83	0.91	0.84	0.99	0.2	0.48	0.37	0.6	0.48	-0.59	1

Tools of EDA

Various Statistical Techniques you are used to perform EDA:

1. Univariate Analysis and visualizations.
2. Bivariate and Multi-variate visualizations and visualizations for mapping and understanding interactions between different features in the data.
3. Predictive models, such as Linear Regression, use statistics and data to predict outcomes.
4. Classification Modelling for correctly predicting data labels.
5. Unsupervised techniques such as Clustering (e.g. K-Means) that help in finding similar behavioural patterns in Data.
6. Dimensionality Reduction Techniques (e.g. PCA) that help bringing down complex high dimensional data into lower dimensional data without losing (much) information.

High Level EDA Process



Data

Feature Analysis

Univariate Analysis

Multivariate Analysis

Correlation Analysis

Find Anomalies & Missing Values

Outlier Detection

Feature Engineering

Insights

Analytics- Visualisation

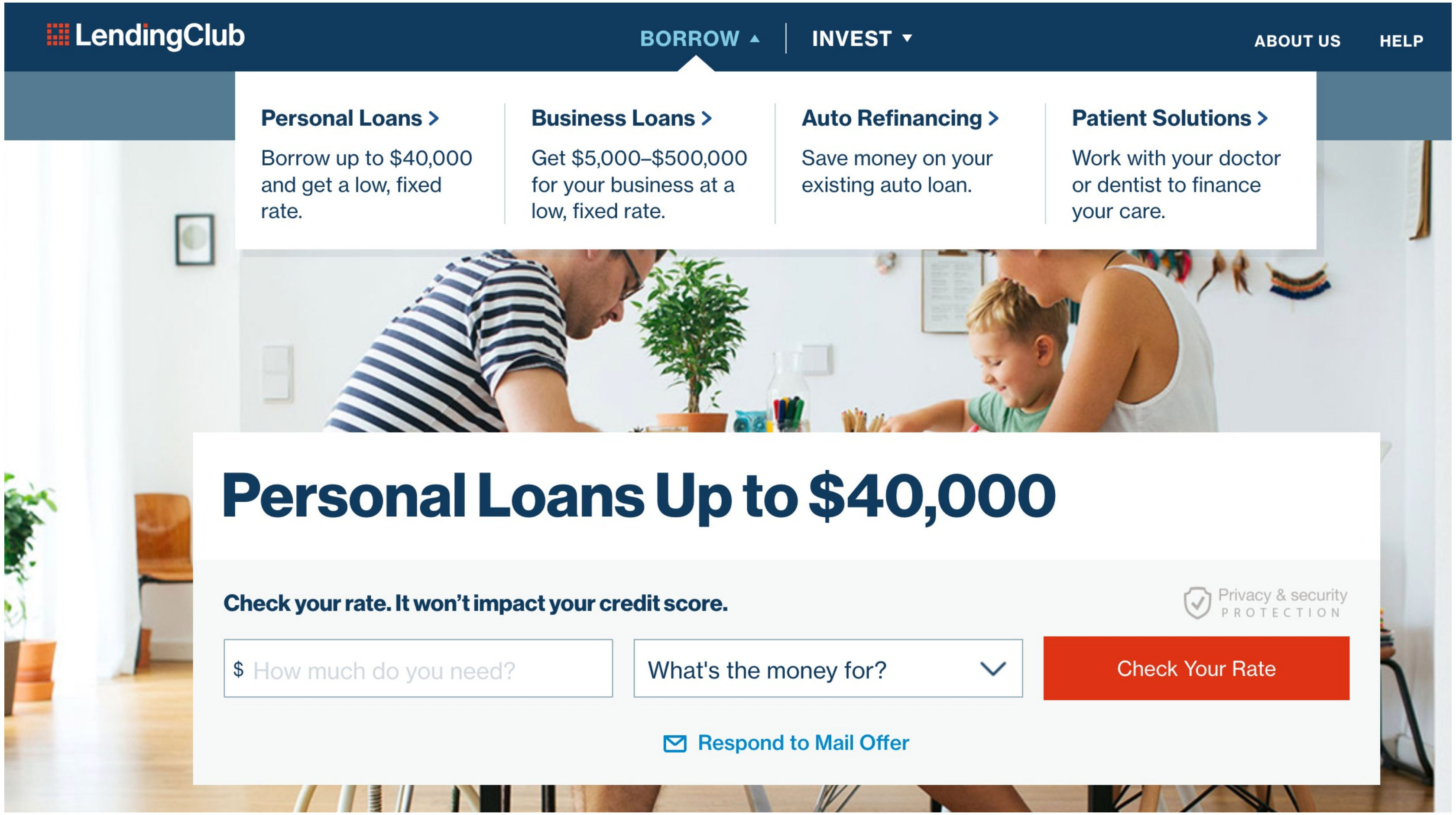
Predictive Modelling/ML

Why EDA in Machine Learning

EDA is a very critical task in Machine Learning for the following purposes

- Find out Missing Data
- Finding Data Anomalies
- Finding Outliers
- Finding class imbalances
- Finding correlations between data
- Finding the right features for inclusion in the model
- Ascertain data types/values and transformation needs
- Ascertain Data Scales for scale transformations
- Ascertain need for derived features which might be more useful and help reduce dimensionality
- Ascertain need for additional data and external data integrations
- Ascertain Noises in Data

Lending Club EDA Case Study



Personal Loans >
Borrow up to $40,000 and get a low, fixed rate.

Business Loans >
Get $5,000–$500,000 for your business at a low, fixed rate.

Auto Refinancing >
Save money on your existing auto loan.

Patient Solutions >
Work with your doctor or dentist to finance your care.

Personal Loans Up to $40,000

Check your rate. It won't impact your credit score.

Lending Club is an US based Consumer Finance Company. They have shared their Loan Transaction data in public for the purpose of Data Science research. The most popular usage of Lending Club Data has become the EDA Case Study.

Lending Club Project Problem Statement

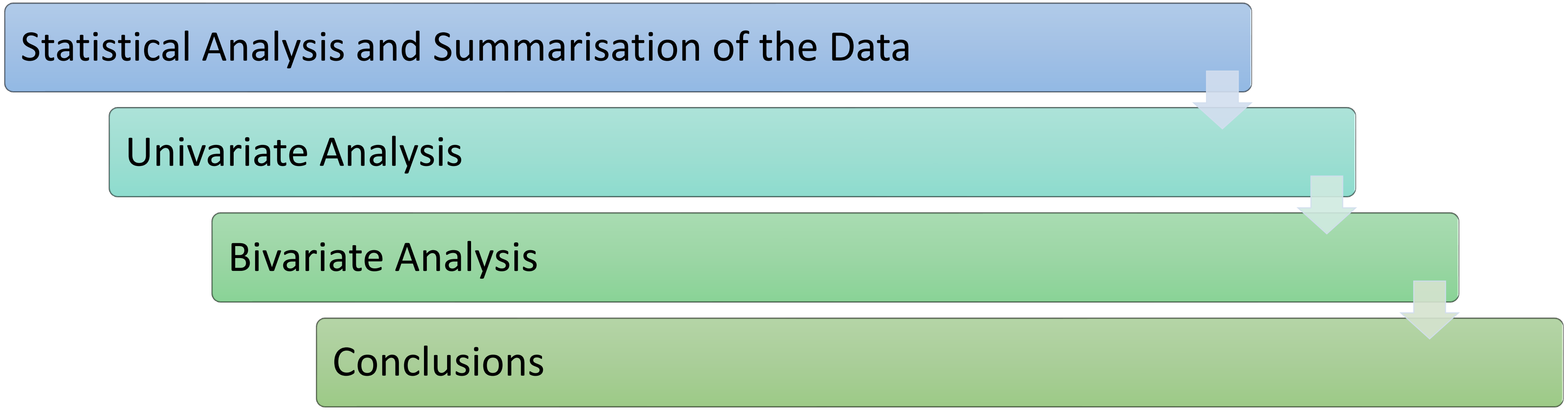
When Lending Club receives a loan application from an individual, the company has to make a decision for loan approval based on the Applicant’s profile and Credit History. This is called Credit Decision. A correct Credit Decision is important to the company as a rejected Credit is loss of business, at the same time, credit given to an un-credit-worthy person leads to risk of default.

To make the Credit Decision process efficient and correct, Lending Club wants to undertake an exercise to analyse past loan data with the incidences of ‘default’s and ‘non-default’s. Through this analysis, Lending Club wants to find what factors in the Applicants Profile must they focus on to ensure Credit is given to the right applicants and reduce chance of Credit Decision errors.

Historically, a manual effort with some programming/tool intervention would have helped in the process with large amount of effort and unreliability in results.

Luckily for us, with Data Science techniques at our disposal, we will look at the automating the entire process and bring out the required insights.

EDA Process



Statistical Analysis and Summarisation of the Data

Univariate Analysis

Bivariate Analysis

Conclusions

As part of the Analysis, we will use a number EDA techniques including abundant amount of Visualisations.

The final goal will be to draw conclusions from the data that would help Lending Club in better profiling of customer for better Credit Decisions.