# CHAPTER 2: Exploratory Data Analysis

## Introduction

This chapter is intended to provide a detailed explanation of Exploratory Data Analysis. After completing this chapter, you will be able to perform univariate, bivariate, and multivariate data analysis and draw many meaningful insights from your data.

Exploratory data analysis is a set of techniques that were developed by Tukey, John Wilder in 1970. The philosophy behind this approach was to examine the data before building a model. John Tukey encouraged statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. today data scientists and machine learning analysts spend most of the time in data wrangling and Exploratory data analysis are also known as EDA. But what is this EDA and why it is so important? This chapter explains what EDA is and how to apply EDA techniques to a data set.

Before we get into EDA, First let us understand what data is, different types of data, and how to understand data using descriptive statistics.

### Data

Data is classified into two types Numerical and Categorical

Categorical Data (Qualitative) represent types of data that may be divided into groups. Eg: Car brands – Audi, BMW, and Mercedes. Gender – Male, Female.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Eg: birthdate, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Numerical Data is also known as quantitative data which represents the numerical value(i.e, how much, how often, how many). Numerical data gives information about the quantities of a specific thing. It is again divided into two groups: Discrete and Continuous.

* Discrete data can be usually counted finitely. Those values cannot be subdivided meaningfully
  + Example: Number of students in a class: 10, 20, 30.
* Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.
  + Examples: Hight, Area, Distance, Time

## Levels of Measurement Scales

In statistics, there are four data measurement scales: nominal, ordinal, interval, and ratio. These are simple ways to sub-categorize different types of data.

* Nominal scale: The nominal scale could simply be called “labels”. Note: all these scales are mutually exclusive and none of them have any numerical significance. Here are some examples, below

A screenshot of a computer

Description automatically generated with low confidence

* Ordinal scale: The order of the values is what’s important and significant, but the difference between each one is not really known. Here are some examples, below

Table

Description automatically generated

* Interval Scale: Interval scales are numeric scales in which we know both the order and the exact differences between the values. The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees. The problem with interval scales: they don’t have a “true zero”. For example, there is no such thing as “no temperature”, at least not with celsius. In the case of interval scales, zero doesn’t mean the absence of value but is actually another number used on the scale, like 0 degrees celsius. Negative numbers also have meaning. Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide.
* Ratio scale: Ratio scales are the ultimate nirvana when it comes to data measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an absolute zero which allows for a wide range of both descriptive and inferential statistics to be applied. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios).

## Descriptive Statistics

Descriptive statistics involves organizing, summarizing, and presenting data in an informative way. Descriptive statistics, unlike inferential statistics, seeks to describe the data but does not attempt to make inferences from the sample to the whole population.

Example: You have all the data on how the business is going on, how much inventory you keep, how many customers come to your store, In which month it has been more, at what day of the week it occurs more. Which product is being sold more at what point of time, on what hours is your product sold more. What kind of customers come, do male customers come more at a certain point in time, or do female customers come then. People with children come more, cigarette buyers come more, or beer buyers come more, or grocery item buyers come more.

Descriptive statistics answers all these questions based on data.

### Different types of Descriptive Statistics

Descriptive statistics are broken down into two categories

**The Measure of Central Tendency**

**The Measure of Variability (Spread)**

### The Measure of Central Tendency: Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way “central” to the data set.

* Mean: Mean is a central tendency of the data i.e. a number around which a whole data is spread out. In a way, it is a single number that can estimate the value of the whole data set.
  + Most stable measure
  + Affected by extreme values
  + It May not exist as a data point in the set

The formula for a sample mean:

A picture containing schematic

Description automatically generated

Here n is the size of the data set, x̄ is the sample mean, and x¡ the numbers in sequence, ∑ is the summation of the entire data set.

Similarly, for a population data of size N, the population mean is:

Diagram, schematic

Description automatically generated with medium confidence

Example: The systolic blood pressure of seven middle aged men in:

150, 123, 134, 170, 146, 124, and 113

The mean is = (150+123+134+170+146+124+113)/7 = 137.14

Example: The systolic blood pressure of seven middle-aged men with extreme value in:

150, 123, 134, 170, 146, 124, and 1113

The mean is = (150+123+134+170+146+124+1113)/7 = 280

We can see how to mean is affected by one extreme value

* Median: Median is the value that divides the data into 2 equal parts i.e. the number of values on the right side of it is the same as the number of values on the left side. When data is arranged in either ascending or descending order.
  + It may not exist as a data point in the set
  + Influenced by the position of items, but not their values
  + Median is not influenced by extreme values

Median will be a middle term if the number of terms is odd, and median will be the average of middle 2 terms if the number of terms is even.

Example: if n is odd

The re-ordered systolic blood pressure data:

113, 124, 125, 132, 146, 151, and 170

The median here is 132

For example: if n is even

The re-ordered systolic blood pressure data:

113, 124, 125, 132, 146, 151, 161, and 170

The median here is (132+146)/2 = 139

* Mode: Mode is the most commonly occurring value
  + Mode exists as a data point.
  + Useful for qualitative data.

Example: Six men with high cholesterol participated in the study to investigate the effects of diet on cholesterol levels. At the beginning of the study, their cholesterol levels (mg/dl) were as follows:

366, 327, 274, 292, 274, and 230

Rearrange the data in ascending order as follows:

230, 274, 274, 292, 327, and 366

The mode between the two men having the same cholesterol level = 274

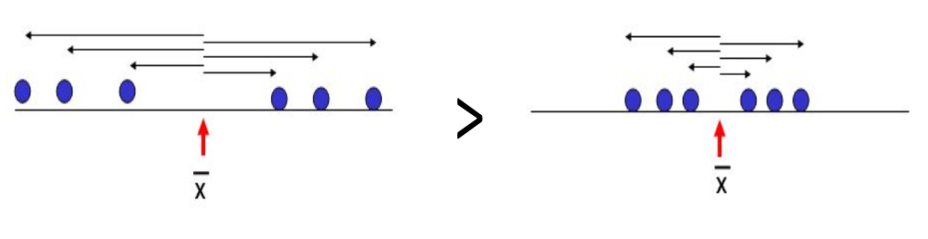
But there could be a data set where there is no mode at all as all values appear the same number of times. If two values appeared at the same time and more than the rest of the values, then the data set is bi-modal. If three values appeared at the same time and more than the rest of the values then the data set is tri-modal and for n modes, that data set is multimodal.

### The Measure of Variability (Spread/Dispersion)

Measures that help us to know about the spread of a data set are called measures of dispersion. The mean, median or mode is usually not by itself a sufficient measure to reveal the shape of a distribution of a data set. We also need a measure that can provide some information about the variation among data set values.

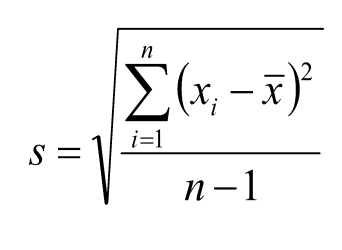
The measures of central tendency and dispersion took together to give a better picture of a data set.

* Standard deviation: Standard deviation is the measurement of the average distance between each quantity and mean. That is, how data is spread out from the mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.



Here (left side) standard deviation is high because, from the mean (x̄), the points are distributed at a longer distance as compared to the right side, where the distance is a bit smaller.

Sample Standard Deviation is denoted by “S”



* + The sample standard deviation has the advantage of being in the same units as the original variables (x).
  + If the standard deviation is small, the data has little spread (i.e., most points fall very near the mean).
  + If standard deviation = 0, there is no spread. This only happens when all data items are the same value.
  + The standard deviation is significantly affected by outliers and skewed distributions.

Population Standard Deviation is denoted by “σ” (sigma)

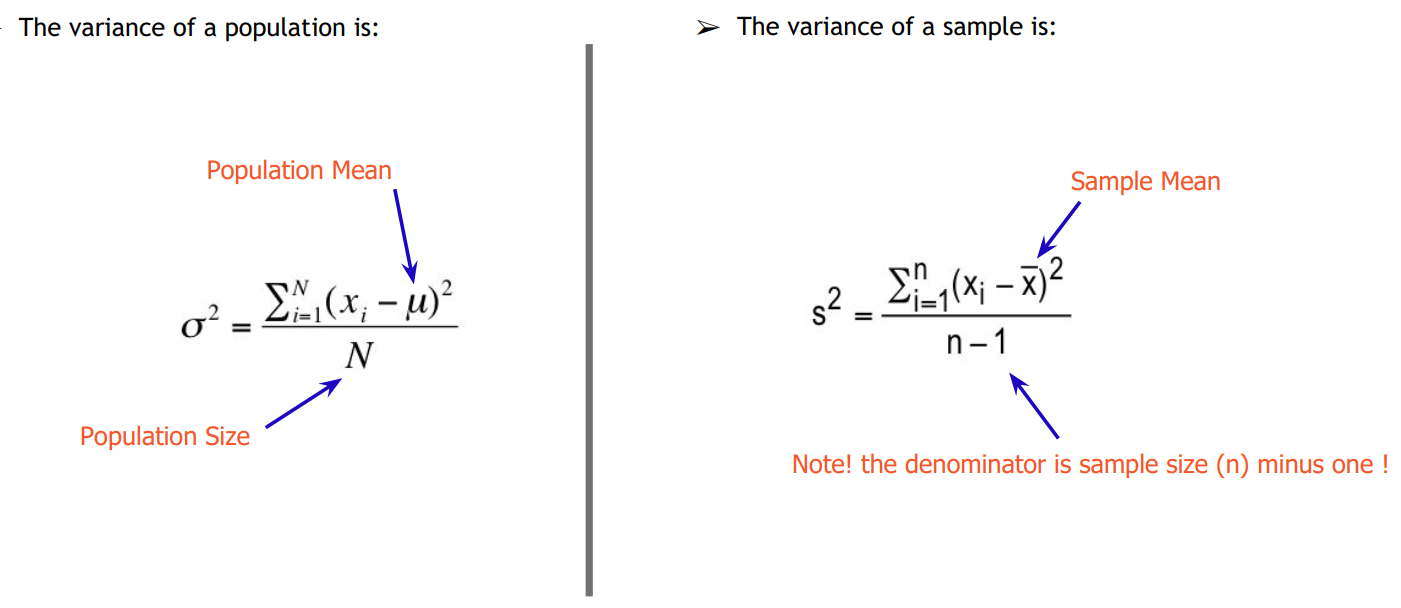
Text, letter

Description automatically generated

* Variance: Variance is a square of the average distance between each quantity and mean. That is, it is a square of standard deviation.



The variance of Population and Sample



* Range: Range is one of the simplest techniques of descriptive statistics. It is the difference between the lowest and highest value.
  + It is easy to calculate.
  + It is implemented for both “best” and “worst” case scenarios.
  + Too sensitive for extreme values.

For example, the minimum and maximum blood pressure are 113 and 170, respectively. Hence the range is 57.

* Percentile: Percentile is a way to represent the position of a value in a data set. To calculate percentile, values in the data set should always be in ascending order.

Example:

12, 24, 41, 51, 67, 67, 85, 99

The median 59 has 4 values less than itself out of 8. It can also be said as: In the data set, 59 is 50th percentile because 50% of the total terms are less than 59. In general, if **k** is **an nth** percentile, it implies that **n%** of the total terms are less than “k”.

* Quartile: In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in ascending order.
  + There are three quartile values. The first quartile value is at the 25th percentile, the second quartile is at the 50th percentile and the third quartile is at the 75th percentile.
  + The second quartile (Q2) is the median of the whole data.
  + The first quartile (Q1) is the median of the upper half of the data.
  + The third quartile (Q3) is the median of the lower half of the data.
  + Quartiles exclude extreme values scientifically

Example: Points scored per game are

12, 24, 41, 51, 67, 67, 85, 99, 115

So here, by analogy,

The middle quartile (50th percentile, Q2) = 67 is the 50th percentile of the whole data.

The first quartile (25th percentile, Q1) = 41 is the 25th percentile and median of the upper half of the data.

Third quartile (75th percentile, Q3) = 85 is 75th percentile and median of the lower half of the data.

Interquartile range, IQR = Q3 – Q1 (Central 50% of data) = 85 – 41 = 44

* Skewness: Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative or undefined.

In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other, which is also called symmetrical distribution, where mean=median=mode.

When a distribution is skewed to the left, the tail on the curve’s left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode and the median. This is also called negative skewness.

When a distribution is skewed to the right, the tail on the curve’s right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode and the median. This is also called positive skewness.

Diagram

Description automatically generated

**Covariance and Correlation:** Covariance studies the direction between two continuous variables and Correlation studies the direction and strength between two continuous variables and helps in understanding how strongly those two variables are associated with each other.

Covariance and correlation are two significant concepts used in statistics and machine learning. One of the most asked questions is the difference between these two terms and how to decide when to use them. Here are some definitions and mathematical formulas used that will help you fully understand covariance and correlation.

What is Covariance Matrix?

A covariance matrix is used to study the direction of the linear relationship between variables. Suppose we have two variables X and Y, then the covariance between these two variables is represented as Cov (X, Y). If ∑(X) and ∑(Y) are the expected values of the variables, the covariance formula can be represented as:

A picture containing text, clock, watch

Description automatically generated

Here are some plots that highlight how the covariance between two variables could look in different directions.

Diagram

Description automatically generated

The covariance values of the variables can lie anywhere between -∞ and + ∞. A negative value indicates a negative relationship whereas a positive value indicates a positive relationship between the variables.

When the unit of observation is changed for one or both two variables, the covariance value changes. However, there is no change in the strength of the relationship.

To better understand the difference between covariance and correlation, let us understand what a correlation matrix is.

What is a Correlation Matrix?

A correlation matrix is used to study the strength of a relationship between two variables. It not only shows the direction of the relationship but also shows how strong the relationship is. The correlation formula can be represented as:

Text

Description automatically generated

Where:

* + Var(X) = standard deviation of X.
  + Var(Y) = standard deviation of Y.

When the two variables move in the same direction, they are positively correlated. On the contrary, when the variables move in the opposite direction, they are negatively correlated.

Chart, scatter chart

Description automatically generated

Fig: Positive relationship

Chart, scatter chart

Description automatically generated

Fig: Negative relationship

The correlation value of the two variables ranges from -1 to +1. A value close to +1 indicates a strong positive relation and a value close to -1 indicates a strong negative correlation.

Now we have got enough understanding of the data and descriptive statistics. Let’s start learning EDA.

## Exploratory Data Analysis

**What is exploratory data analysis?**

Exploratory Data Analysis or EDA is used to take insights from the data. Data scientists and machine learning analysts try to find data patterns, relations, and anomalies in the data using statistical graphs and other visualization techniques.

Following things are part of EDA:

1. Get maximum insights from a data set
2. Uncover underlying structure
3. Extract important variables from the dataset
4. Detect outliers and the anomalies (if any)
5. Test underlying assumptions
6. determine the optimal factor settings

**Why EDA is important?**

the main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. it allows analysts to understand the data better before making any assumptions. the outcome of media helps businesses to know their customers, expand their business and take decisions accordingly.

## Hands-on

**How to perform EDA?**

To understand EDA better let us take an example. We will be working on Automobile Dataset for analysis.

Dataset: <https://www.kaggle.com/toramky/automobile-dataset>

**Step1: Import libraries and load dataset**

A picture containing chart

Description automatically generated

**Step2: Load the dataset**

A picture containing chart

Description automatically generated

Graphical user interface, application

Description automatically generated

We can see that the data set has 26 attributes and column names are missing. you can also observe that there are ‘?’ at some places which means our data has missing values also. he will fill in column names first.

**Step3: Basic data exploration**

Text

Description automatically generated

A picture containing diagram

Description automatically generated

We got our column names. the “price” column is our target variable.

**Check for missing values**

Table

Description automatically generated with low confidence

It is showing that we don't have any null values in our data set, but we have observed earlier that there were ‘?’ Symbols in the data set, which means that these symbols are in the form of an object. let us now check the data types of each attribute.



Table

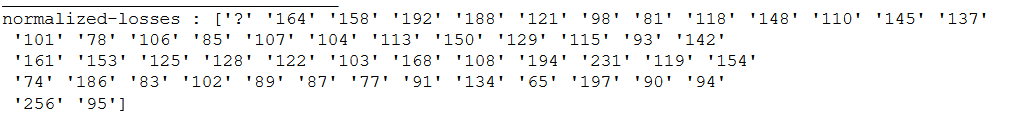
Description automatically generated

We can observe that those columns that have symbols are in the object form as well as some columns should be of an integer type but are of an object type. Now let us detect which columns have symbols and if there are any other symbols too.

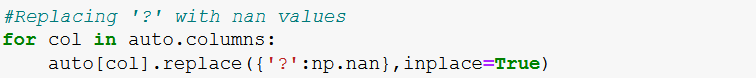
**Step4: Data Cleaning**

A picture containing logo

Description automatically generated



There are null values in our data set in form of ‘?’ only but pandas are not reading them so we will replace them into “np.nan” form.



A screenshot of a computer

Description automatically generated with low confidence

Now we can observe that the ‘?’ symbols have been converted into NaN form. let us check for missing values again.

Table

Description automatically generated

We can observe that how there are missing values in some columns.

**Visualizing the missing values**

With the help of heatmap, we can see the amount of data that is missing from the attribute. with this, we can make decisions whether to drop these missing values or to replace them. Usually dropping the missing value is not advisable but sometimes it may be helpful too.

Chart

Description automatically generated

Now observe that there are many missing values in “normalized\_losses” while other columns have fewer missing values. we can't drop the “normalized\_losses” column as it may be important for our prediction.

**Relacing the Missing Values**

Will be replacing these missing values with median because the number of missing values is this (you can also use mean too).

A picture containing text

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated



We have imputed the numeric variables missing values with median and for categorical variables, we have used the most frequent value to impute.

Now let’s check whether are there any more missing values.

Graphical user interface, table

Description automatically generated

**Step4: Descriptive Statistics**

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generated

Describe function gives the descriptive statistics like count, mean, standard deviation, min, median, and max for all the continuous variables.

We can get a rough idea about the distributions of the variables.

**Step5: Asking analytical questions and visualizations**

This is the most important step in EDA. This trip will decide how much can you think like an analyst. this step varies from person to person in terms of their questioning ability. Try to ask questions related to independent variables and the target variable. For example – how “fule\_type” will affect the price of the car?

Before this let us check the correlation between the variables, this will give us a roadmap on how to proceed further.

**Heat map**



A computer screen capture

Description automatically generated with medium confidence

**Positive Correlation**

* Price – Wheel\_base, length, width, curb\_weight, engine\_size, bore, horsepower
* Wheelbase – length, width, height, curb\_weight, engine\_size, price
* Horsepower – length, width, curb\_weight, engine\_size, bore, price
* Highway mpg – city mpg

**Negative Correlation**

* Price – highway\_mpg, city\_mpg
* Highway\_mpg – wheel-base, length, width, curb\_weight, engine\_size, bore, horsepower, price
* City – wheel\_base, length, width, curb\_weight, engine\_size, bore, horsepower, price

This heatmap has given us great insights into the data

Now let us apply domain knowledge and ask the questions which will affect the price of the automobile.

1. **How does the horsepower affect the price?**

A picture containing diagram

Description automatically generated

Chart, scatter chart

Description automatically generated

We can see that most of the Horsepower value lies between 50-150 has price mostly between 5000-25000, There are outliers also (between 200-300)

Let's see a count between 50-100 i.e., univariate analysis of horsepower

Chart, histogram

Description automatically generated

The average count between 50-100 is 5 is 50 and it is positively skewed.

1. **What is the relationship between engine\_size and price?**

Text

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

We can observe that the pattern is like the Horsepower and price. i.e., The size of the engine is big the price is increasing. Positive correlation between Engine-size and price.

1. **How do highway\_mpg effects price?**

**A picture containing chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

We can see price decreases with an increase in higway\_mpg.

**Let us check the number of doors.**

Text

Description automatically generated with medium confidence

A picture containing logo

Description automatically generated

1. **Relation between no. of doors and price**

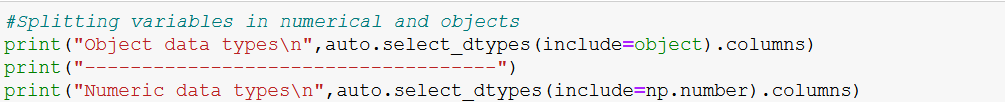


Chart, box and whisker chart

Description automatically generated

With his boxplot, we can visually conclude that there is no significant variation in the price of the vehicle with respect to no. of doors. There are few outliers in the data.

Let’s split the data into numerical and objects



Graphical user interface, text, application, Word

Description automatically generated

**Step6: Outlier Detection and Treatment**

We can have an idea of outlets in the distribution with the command pd.describe() but if we want to have a visual idea of their presence the first thing you have to learn is the use of Box Plots.

Let's have an example of a boxplot for our dataset. We will produce a plot for the quantitative variable. A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartile. the box plot gives us information about the five important numbers of the distribution (min, 1st quartile, median, 3rd quartile, and max).

The values inside the box are between the 1st and the 3rd quartile. The point out of the whiskers that we can see in the graphs represents the outliers. outliers may be plotted individual points.

Here are the types of observations one can make from viewing a box plot:

* What the key values are, such as the average, median, 25th percentile, etc.
* If there are any outliers and what their values are.
* Is the data semantical?
* How tightly is the data grouped?
* If the data is skewed and if so, in what direction.

When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric. when the media is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (skewed right).

From about the upper quartile, a distance of 1.5 times the IQR is measured out and a whisker is drawn up to the largest observed point from the data set that falls within this distance. Similarly, a distance of 1.5 times the IQR is measured out below the lower quartile and a Whiskers drawn up to the lower observed point from the data set that falls within this distance. all other observed points are plotted as outliers.

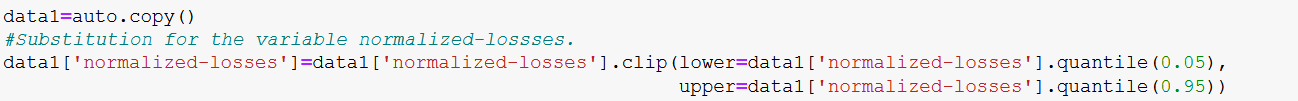
A picture containing crossword puzzle, shoji, text

Description automatically generated

You can remove the outliers using the clip() function to substitute the lower and the upper data with some specific quantiles (for example 0.05 & 0.95).

we will just do it in a copy of the data set apart.

we have outliers in every quantitative variable except for symbolling, bore, curb-weight, and height.



We can compare the difference in the distributions before and after the removal of outliers.

Graphical user interface

Description automatically generated with low confidence

Chart, box and whisker chart

Description automatically generated

Note that the scale is changed (now is 80 -180).

We can do the same with all the other variables changing the upper and lower quantile.

**Step7: Univariate Distributions**

You can use the Count Plot for the categorical features to have an idea of how they are distributed.

**Categorical Features**

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

We can see the highest number of cars sales are from Toyota company, followed by Nissan and Mazda.

Chart

Description automatically generated

Most of the market was captured by gas-fuel-type cars.

Chart

Description automatically generated

Most of the cars are of the std aspirations.

A picture containing logo

Description automatically generated

Most of the cars are having 4 doors.

Chart, bar chart

Description automatically generated

Sedans are the highest number of cars sold followed by hatchback and wagon.

Chart, bar chart

Description automatically generated

The highest number of cars are having front word drive wheels.

Shape, square

Description automatically generated

The majority of the cars are having engines located at the front and every minute no. of cars are having a rear engine.

Chart, bar chart

Description automatically generated

Most of the cars are having the ohc engine type.

Chart, histogram

Description automatically generated

Most of the cars are having 4 cylinders.

Chart, bar chart

Description automatically generated

Most of the cars are having Multi-Point Fuel Injection (MPFI) fuel system, next are 2bbl fuel system cars.

**Numeric Variables**

Let’s plot histograms to see the distributions of the numeric variables.



A close-up of a window

Description automatically generated with low confidence

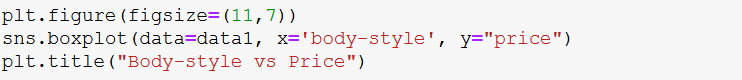
Most of these variables are not normally distributed. **Highway-mpg, length, width, and height** have a bell shape and are almost skewed.

**Engine size and price** are positively skewed.

**Symbolling** is a discrete variable and has its maximum for the value ‘0’.

**Step8: Bivariate**

We have seen the univariate analysis above now let’s see how those variables affect the price in bivariate analysis.



Chart, box and whisker chart

Description automatically generated

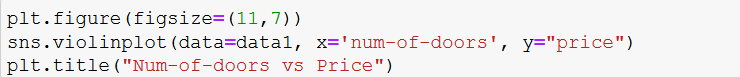
When we analyzed “body-style” individually we came to know, sedans are the most used cars. Now, if we compare the prices of the cars with different body-styles. We got a clear picture of why sedans and hatchbacks are the most used cars. i.e., the price of the hatchback and sedan are low compared to the price of other body-style variant cars. The costliest cars are from hardtop body-style.



Chart, line chart

Description automatically generated

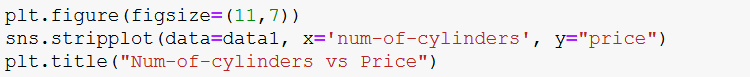
The prices of the “fwd” drive wheel cars are falling less than $30000 and most of the cars are within the range of $10000 while coming to “4wd” prices are less than $20000 and “rwd” drive-wheel cars are costlier compared to other drive wheels.



Chart, radar chart

Description automatically generated

There is not much variation in the price of cars based on the number of doors.



Chart, scatter chart

Description automatically generated

We can clearly see the price of 4-cylinder cars is less compared to all the other cars with a greater number of cylinders.

A picture containing text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

We have plotted “make” against the “price” to see the median cost of the cars. I have drawn a horizontal line at $20000 price to get an idea that most of the car prices are not more than $20000. Only makers like bmw, jaguar, Mercedes-Benz, and porsche are manufacturing the expensive cars.

**Step9: Multivariate Analysis**

In bivariate analysis, we have analyzed the relationship between two variables. Now we are going to see how the price of a car gets affected based on two independent variables.



Chart, scatter chart

Description automatically generated

Fuel type doesn’t seem any impact on the price and the highway mpg of the cars.

Next, we will be using Facet Grid to analyze the different combinations of variables for analysis.



Chart, scatter chart

Description automatically generated

* The price of the cars whose aspiration is “std”, drive-wheels are “rwd” are expensive and highway-mpg is very less compared to other cars.
* The cars which are having aspiration as “std”, drive-wheels as “fwd” are having a price below $20000 and highway-mpg is very good.
* The cars with aspiration as “turbo” are very less in volume and can get clear analysis with limited data.



Chart, scatter chart

Description automatically generated

* Sedan cars are having a wide range of prices, we can see there are low price cars and expensive cars in sedan
* While coming to hatchback we can only see cars in the low range i.e., less than $20000.

Similarly, we can drill down to see a depth analysis of each variable affecting the price of the cars.

**SUMMARY**

In this chapter, you discovered several ways that you can explore your data. We have seen univariate analysis for understanding the distribution of the single variable, bivariate analysis for analyzing the relationship between two variables. You learned how to plot your data using:

* Histograms
* Density Plots
* Boxplots
* Heat Map
* Scatter Plot Matrix

## Assessment

**Choose the appropriate option**

1. Which plots are suitable to perform univariate analysis for numerical variables?
   1. Box Plot
   2. Bar Plot
   3. KDE
   4. Both A & C
2. Data visualization is useful because
   1. Easy to spot patterns and trends
   2. Memorizing important insights
   3. Finding outliers
   4. None of the above.
   5. All the above
3. A Scatterplot is used to understand the relationship between
   1. Two categorical variables
   2. Two continuous variables
   3. One categorical and one continuous variable
   4. None of the above
4. The IQR (Interquartile range) is calculated using \_\_\_\_\_\_\_\_\_\_\_ formula
   1. Q3 – Q1
   2. Q1 - Median
   3. Mean - Median
   4. Lower whisker – Upper whisker
5. Which plot is used to find the outliers robustly?
   1. Scatter Plot
   2. Bar Plot
   3. Box Plot
   4. Pie Plot

**Fill in the spaces with appropriate answers**

1. When we analyze one variable at a time, it is called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_data analysis.
2. An \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is an observation point that is distant from others.
3. In descriptive statistics, a \_\_\_\_\_\_\_\_\_\_ plot is a method for graphically depicting groups of numerical data through their quartiles.
4. The Bivariate for categorical and categorical variables can easily be done with the help of \_\_\_\_\_\_\_\_\_\_\_\_\_\_ or \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ charts.
5. The plot which tells the correlation between two variables that may not be directly related is \_\_\_\_\_\_\_\_\_\_\_ graph/chart.

**True or False**

1. EDA is an approach to analyzing datasets to summarize their main characteristics, often with visual methods.
   1. True
   2. False
2. EDA is used for seeing what the data can tell us after the modeling task
   1. True
   2. False
3. Box plot displays the five-number summary of a set of data
   1. Ture
   2. False
4. A histogram is a plot that lets you discover, and show, the underlying frequency distribution of a set of categorical data.
   1. True
   2. False
5. A scatter plot is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables.
   1. True
   2. False

## Program Assessment

Using the dataset from the below link. Perform the following tasks

https://github.com/fenago/MLBook/blob/main/Chapter%202/Code/Dataset/StudentsPerformance.csv

* Import the dataset
* Validate the missing values
* Perform Univariate Analysis
* Perform Bivariate Analysis
* Perform Multivariate Analysis

## Assessment Solutions

**Choose the appropriate option**

1. D
2. E
3. B
4. C
5. A

**Fill in the spaces with appropriate answers**

1. Univariate
2. Outlier
3. Box plot
4. Value, Size
5. Scatter

**True or False**

1. True
2. False
3. True
4. False
5. True