Subjective Questions

Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Below are the effect of Categorical variables on the target variable “cnt”.

From season vs cnt:

- It is observed that the count of total rental bikes is more in fall season and least in spring season.
- In fall a median of over 5000 total rental bikes bookings is observed.

From mnth vs cnt:

- It is observed that the booking are more in the month of September which is approximately over 7000 and has a median count of around 5000. January has the least amount of bookings somewhere over 3000.

From weathersit vs cnt:

- It is observed that the bookings are high during Clear, Few clouds, Partly cloudy, Partly cloudy weather with over 6000 bookings with a median of around 5000. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds weather has the least number of bookings.

From holiday vs cnt:

- It is observed that the maximum amount of bookings on holidays and non-holidays seems to be same, which is around 6000. But the median is more on non-holidays which is over 4000. So comparatively, bookings are more on non-holidays.

From weekday vs cnt:

- It is observed that the bookings are more on weekdays. More bookings happened on Wednesdays and Thursday and least being on Sunday.

From workingday vs cnt:

- It is observed that the median is almost the same for both working and non-working day which is over 4000 bookings. But the probability of customers booking is observed more on working days.

Effect of yr on cnt:

- It is observed that year 2019 has more number of bookings comparatively to 2018.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Variable “temp” has the highest and a positive correlation with the target variable “cnt”.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

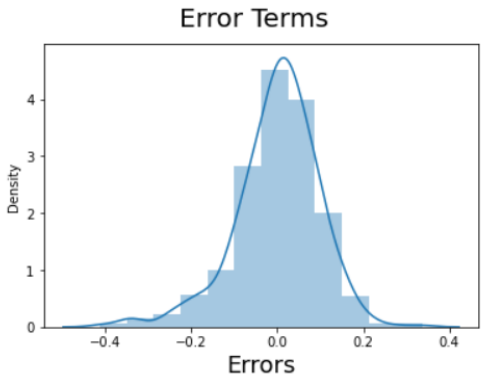
The assumptions of Linear Regression could be validated from:

- The equation of best fitted surface based on the final model built i.e., lr7 in our case

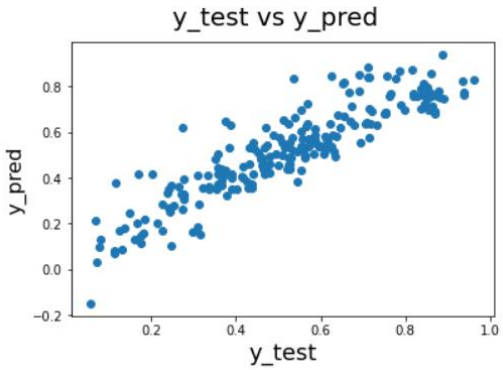
Equation of best fitted surface based on model lr7:
cnt = 0.052028 + (yr_1 × 0.228547) + (temp × 0.585685) + (season_summer × 0.078760) + (season_winter × 0.137323) + (mnth_September × 0.100220) + (weekday_Saturday × 0.019739) − (weathersit_Mist_Cloudy × 0.070643) − (weathersit_LightSnow_LightRain × 0.318299)

- Collinearity is observed between the independent and dependant variables.
- **Residual Analysis:**

When a histogram plot is plotted for all the error terms, we could see it is normally distributed.



* **Homoscedasticity**: Variance of residuals (error terms) is constant across predictions.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Temp, yr, weathersit and season are the top features contributing significantly towards explain the demand of shared bikes.

yr : yr represents year. From the analysis, year 2019 has increase in demand for shared bikes comparatively to year 2018.

Season: The season analysis indicates that more bikes are rent during fall season.

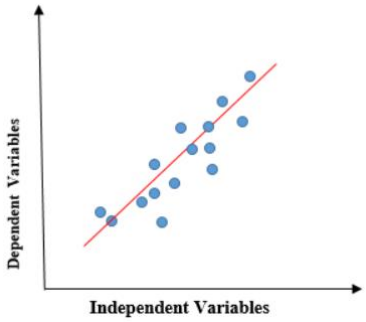
Temp: It has a positive correlation with the variable “cnt” stating the demand for shared bikes.

weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds A coefficient value of ‘-0.318299’ indicated that, w.r.t weathersit_Clear_FewClouds, a unit increase in weathersit_LightSnow_LightRain, decreases the bike hire numbers by 0.318299 units.

General Subjective Questions:

1. Explain the linear regression algorithm in detail. (4 marks)

- Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.
- If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there is more than one input variable, such linear regression is called **multiple linear regression**.
- The linear regression model gives a sloped straight line describing the relationship within the variables.

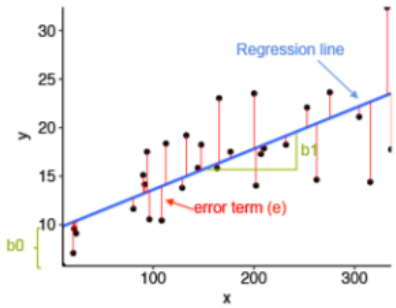


- The above graph presents the linear relationship between the dependent variable and independent variables.
- When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.
The mathematical equation of linear regression can be written as y = b0 + b1*x + e, where – • y is the predicted (dependent variable)

$$y = b0 + b1*x + e$$

y= Dependent Variable.
x= Independent Variable.
b0= intercept of the line.
b1 = Linear regression coefficient.
e = error term

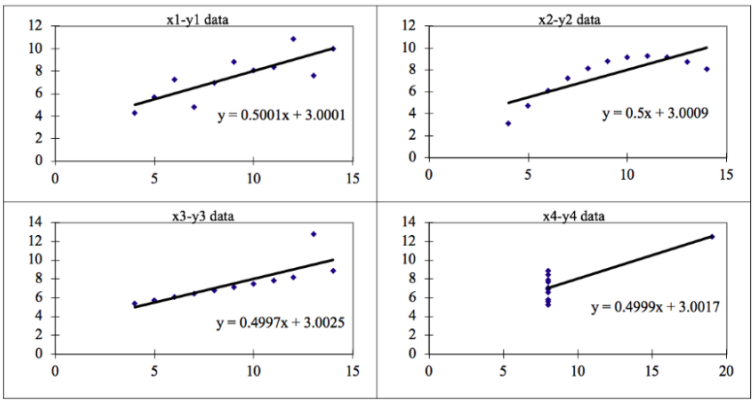


The graph of linear regression above shows –
• the best-fit regression line is in blue
• the intercept (b0) and the slope (b1) are shown in green
• the error terms (e) are represented by vertical red lines

Need of a Linear regression:
As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable.

2. Explain the Anscombe's quartet in detail. (3 marks)

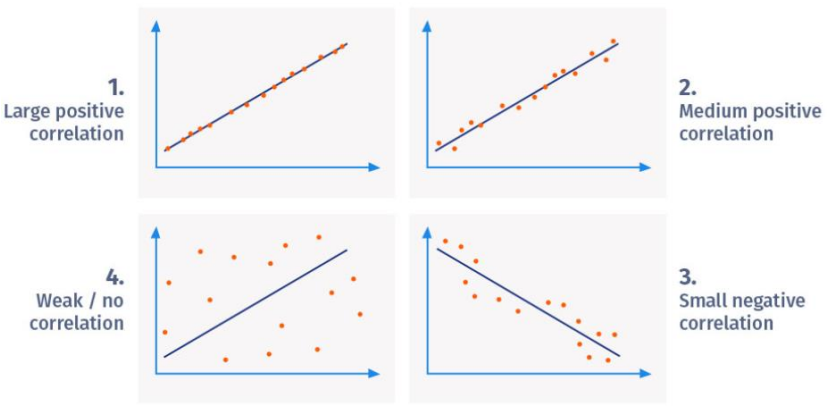
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. These four plots can be defined as follows:



The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model
The four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

3. What is Pearson's R? (3 marks)

- Pearson correlation coefficient or Pearson's correlation coefficient or Pearson's r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.
- In simple words, Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.
- This approach is based on covariance and thus is the best method to measure the relationship between two variables.
- The Pearson coefficient correlation has a high statistical significance. It looks at the relationship between two variables.
- It seeks to draw a line through the data of two variables to show their relationship.
- The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, scaling is done to bring all the variables to the same level of magnitude.
- Normalized/Min-Max Scaling:
  It brings all of the data in the range of 0 and 1.
  sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

- Standardization:
  It replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

  sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

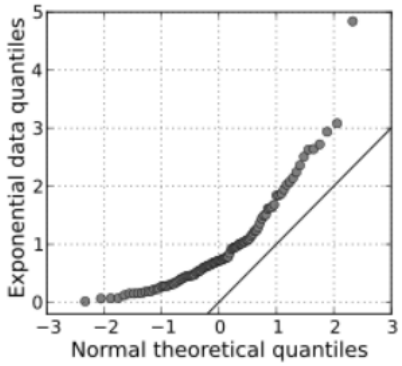
- If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
  [VIF = 1/1-R2]

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile.
- For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45 degree reference line:



- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.
- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Submitted by:
Suhas S