Bike Sharing Assignment – Subjective Questions

Assignment Based Subjective Questions :

Que 01 : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans :

- Summer and Fall seasons have high demand of bikes
- Demand for Rental bikes is higher from Month 5 to Month 10.
- In 2019, demand was higher than 2018. So we can say that, people are accepting rental bikes and in future, demand might increase
- Usually on holidays, demand for bike is higher
- People favours renting a bike when weather is Light Snow, Light Rain, Scattered clouds.

Que 02 : Why is it important to use drop_first=True during dummy variable creation?

Ans : drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Dropping this redundant variable helps in reducing **multicollinearity.**

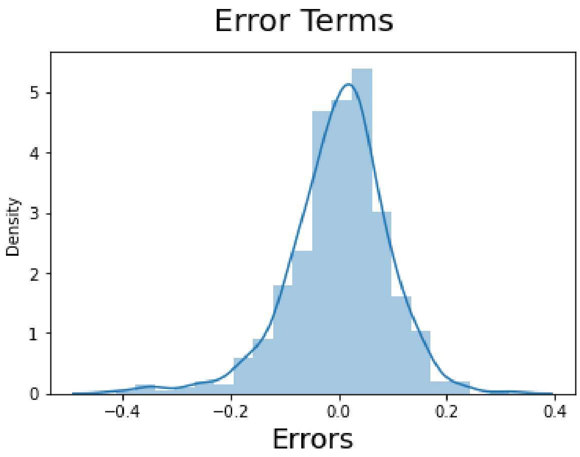
Que 03 : Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans : From Pairplot and Heatmap, we can say that 'atemp' column has high correlation with target variable which is around 0.65, followed by 'temp' column (0.64)

Que 04 : How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans :

- Error terms which is difference between target variable and predicted target variable is plotted using histogram. It is observed that, error terms follow Normal Distribution with mean close to zero. This validates the assumption of linear regression.



Que 5 : Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans : From final model, parameters obtained are as follows

```
const           0.075325
yr              0.233056
workingday      0.056323
temp            0.549936
windspeed      -0.155238
season_2        0.087429
season_4        0.131826
mnth_9          0.097236
weekday_6       0.067688
weathersit_2   -0.081347
weathersit_3   -0.288021
```

So top 3 features contributing towards explaining demand of shared bikes are:

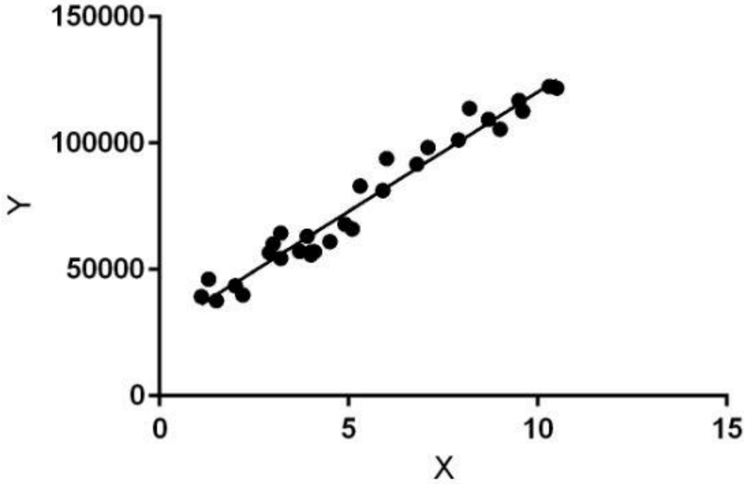
1) Temperature (temp)
2) Weather situation 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
3) Year (2019)

General Subjective Questions

Que 01 : Explain the linear regression algorithm in detail.

Ans :

- Linear Regression Algorithm is a machine learning algorithm based on supervised learning
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).
- 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.** Based on the given data points, we try to plot a line that models the points the best.



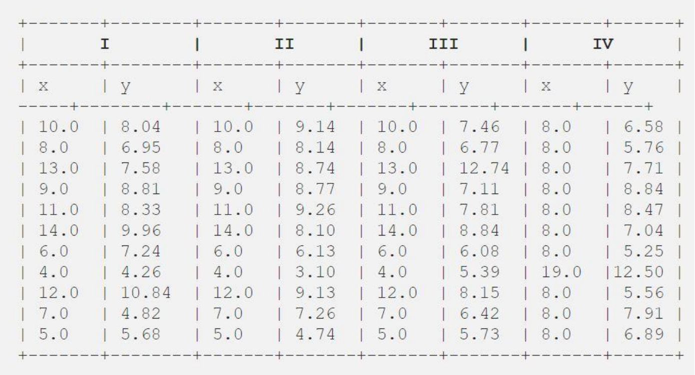
- In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.
- Examples : Prediction of trends and Sales targets, Price Prediction – Using regression to predict the change in price of stock or product.
- To calculate coefficients, we will use the least square criterion, which means we will find a line that will decrease the sum of squared errors

Algorithm :

Que 02 : Explain the Anscombe's quartet in detail

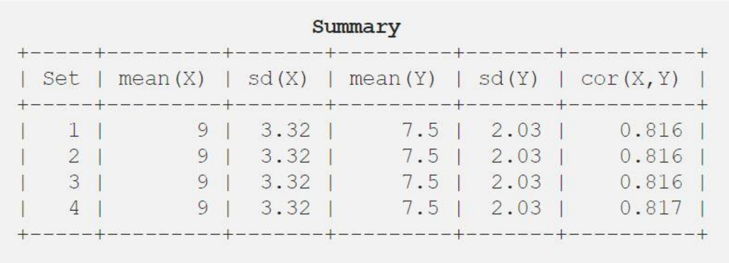
Ans : Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Anscombe's found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. This data set was as follows.



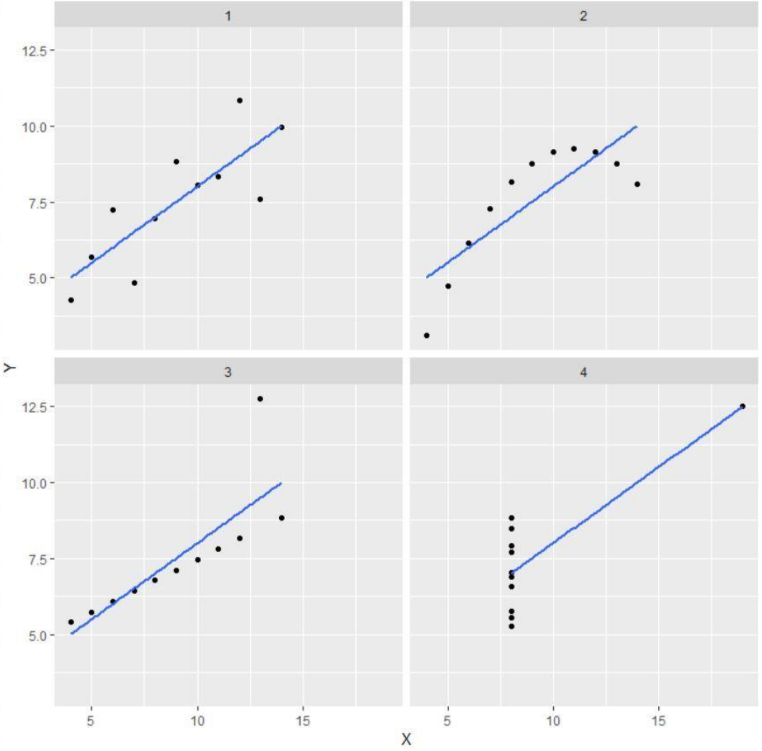
```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+-------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Council then found the mean, standard deviation, and correlation between x and y. Summary of this is as follows :



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

When this data sets were plotted,



It was observed that, Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application : The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Que 03 : What is Pearson’s R?

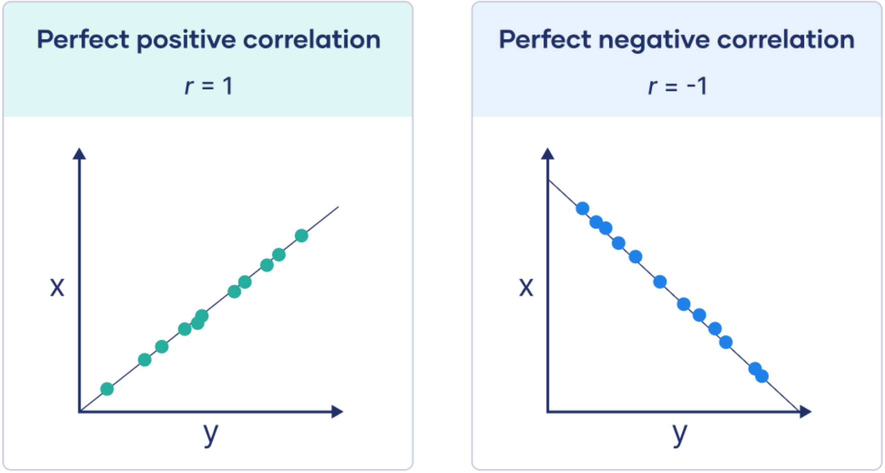
Ans : The Pearson correlation coefficient (***r***) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is also known by many names:

- Pearson’s *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, *r* is negative. When the slope is positive, *r* is positive.

When *r* is 1 or –1, all the points fall exactly on the line of best fit:


Perfect positive correlation

r = 1

Perfect negative correlation

r = -1

Scribbr

Below is a formula for calculating the Pearson correlation coefficient (*r*):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Que 04 : What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans :

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
1. Ease of interpretation

2. Faster convergence for gradient descent methods You can scale the features using two very popular method:

- Standardized Scaling **:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

- MinMax Scaling (Normalized Scaling) : The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

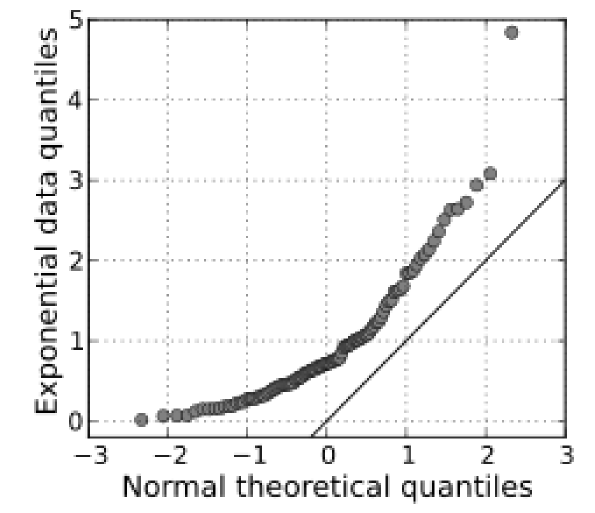
It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

Que 05 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans :

- 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) as infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- 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).

Que 06 : What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans :

- 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.
- 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.