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?

A) To get detailed observation from the categorical variables, we can see the boxplot graphs how exactly data is been distributed.

   a. From the 'Season' box plot we can say that more bikes are given rent during the Fall season.
   b. The 'Year' box plot represents that most of the bikes are rented during 2019 where the x variable has values '0' and '1's which indicated yr : year (0: 2018, 1:2019) as per Data Dictionary
   c. 'Month' box plot represents Most of the bikes given rent in September month and in January very few bikes are given comparatively.
   d. From the 'Holiday' and 'Working day' box plots, we can see that most of the customers are opting for bike rentals on working days than on weekend days may be due to high floating rates.
   e. The 'Weekday' box plot indicated that more bikes are taken for rent from Saturday to Monday.
   f. The 'Weather' box plot indicates that bikes are taken for rent more in 'A' climate conditions which represents "Clear, Few Clouds, Partly Cloudy Weather" as per Data Dictionary.

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

A) When we use Dummy variable creation for categorical variables it is important to use **'Drop_first='True'** because it creates a new column and we need to remove the extra column to reduce the correlation among the dummy variables.

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

A) By looking into the Pair plot, we can state the target variable 'cnt' has high correlation with the **"Temp"** variable.

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

A) By plotting the Residual error using the normal distribution plot we can validate the assumption of the Linear regression. If the plot maintains the linear relation between Dependant variables (Test and predicted) we can see that by plotting the 'y_test vs y_pred' values using scatter plot.

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

A) The Top 3 features contributing significantly towards explaining the demand for shared bikes are:

   a. Temp
   b. Year
   c. Weekday_Mon

General Subjective Questions

1. Explain the linear regression algorithm in detail

A) In linear regression, we have two types which are Simple Linear Regression (SLR) and Multiple Linear Regression (MLR). In SLR we can see the relationship between a dependent variable and one independent variable using an as straight line. We can visualize the straight line using a scatter plot for better understanding.

The standard equation for SLR is as follows: $Y = \beta_o + \beta_1 X$. Where β_0 is the Intercept and β_1 is a slope. For getting the best-fit line we need to have a low RSS value. For knowing the strength of the Linear Regression model, we can verify through R^2 and RSE (Residual Standard Error)

If the R^2 value is closer to or equal to 1 means we have the best fit line or if the value is lower than 0.5 then we need to improve the value because at that stage the deviation is higher with the data points.

In MLR we can see the relationship between one dependent variable and multiple independent variables. The standard equation for MLR is as follows: $Y = \beta_0 + \beta_1 X_1 + \beta 2 X_2 +---+ \beta_p X_p + €$. Almost we have the same procedure for both SLR and MLR. But there are a few more things we need to take into consideration for MLR such as:

1. Adding more variables for training that doesn't mean we get the best model sometimes more variables can cause an 'Overfit' problem which may result in high training accuracy and low testing accuracy.
2. Multicollinearity between the columns/variables.
3. Among all the features selecting the important feature is an important aspect.
4. Using VIF (Variance Inflation Factor) we can drop the unwanted features if the value is greater than 10
5. Feature Scaling is another most important aspect. When we have more independent variables the values of those variables may vary on a large scale which leads to a model with different coefficients. To overcome this issue, we need to perform scaling technique methods either Standardizing or Minmax scaling.

2. Explain the Anscombe's quartet in detail.

A) Anscombe's quartet comprises a dataset that has nearly identical same but if we see them, they are different when visualized. Each dataset consists of two columns (x,y) with 11 rows. This dataset is to analyze the data know the effect of the outlier on statistical properties and then visualize the data.

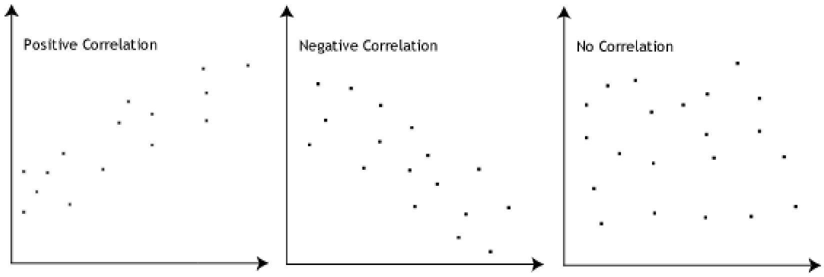
3. What is Pearson's R?

A) In Statistics the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (OOMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations, so it is important to normalize the covariance between -1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r=correlation coefficient
- xi=values of the x-variable in a sample
- $\bar{x}$=mean of the values of the x-variable
- y_i=values of the y-variable in a sample
- $\bar{y}$=mean of the values of the y-variable

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A) Normalized Scaling:

    1. The minimum and maximum values of features are scaled.
    2. Used when features are of different scales.
    3. Scales value between [0,1] or [-1,1].
    4. It is affected by outliers.
    5. Scikit learn provides a transformer called ‘MinMax’ scaler for normalization.
    6. This transformation squeezes the end dimensional data into IN dimensional unit hypercube.
    7. It is useful when we don’t know about the distribution.
    8. It is often called scaling normalization

Standardized scaling:

1. Mean and the standard deviation is used for scaling.
2. It is used when we want to ensure zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. It is much less affected by outliers.
5. Scikit learn provides a transformer cold ‘StandardScaler ‘for standardization.
6. It translates the data to the mean vector of original data to the origin and squeezes or expands.
7. It is useful when the feature distribution is normal or gaussian.
8. It is often called a Z-score. Nomination

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

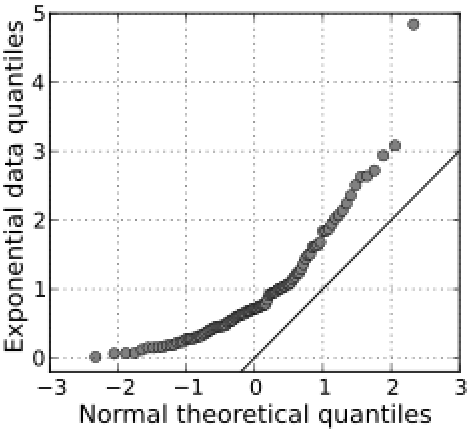
A) If there is a perfect correlation then **VIF** is cold Infinity this shows a perfect correlation between 2 independent variables. In the case of perfect correlation, we get R2 = 1, which leads to 1/(1-R2) Infinity. To solve this problem, we need to drop one of the variables from the data set which is causing this perfect multicollinearity problem an infinite with value indicates that the corresponding value variable may be expressed exactly by the linear combination of other variables (which shows 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.

A) Q-Q plots (Quantile Quantile) plots are plots of 2 quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

For example, the medium is a quantile with 50% of the data falling below that point and the other 50% lying above it. The purpose of Q-Q plots is to find out if 2 sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot if the 2 datasets come from a common distribution, the points will fall on the reference line

Below is the image of a Q-Q plot showing the 45-degree reference line:



If the 2 distributions being compared a similar, the points in the QQ plot will approximately lie on the line y=x. If the distributors are linearly related, the points in the Q-Q plot will approximately lie on a plane, but not necessarily on the line y=x. Q-Q plots can also be used as a graphical means of estimating parameters in a locational 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 a similar or different in 2 distributions.