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?

- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The weekday box plots indicates that more bikes are rent during saturday.
- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The weather sit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather

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

- In machine learning, the parameter **drop_first** typically refers to a setting used in one-hot encoding or dummy variable encoding.
- drop_first=True is used during the dummy variable creation to avoid the dummy variable trap and multicollinearity issues.
- By dropping the first category using drop_first=True, we create a reference category. This means that if all the dummy variables are 0, it implies the reference category is present. Having a reference category helps to avoid the dummy variable trap because we remove the redundant information that can be inferred from the other dummy variables. It also ensures that the model has a full rank and the variables are linearly independent
- we can avoid the dummy variable trap, reduce multicollinearity, and ensure the model's stability and interpretability.

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

- “Temp” had the highest correlation coefficient of 0.63

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

- We can plot the residual distribution. After plotting it comes out to be normal distribution with a mean value of 0. This is how we can validate the assumptions of Linear Regression after building the model on training set.

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

- yr(0.2478)
- holiday(-0.0703)
- Mist_cloudy(-0.0905)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables only.

•The dataset is divided into test and training data

•Train data is divided into features(independent) and target (dependent) datasets

- A linear model is fitted using the training dataset. Internally the api’s from python uses gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares.

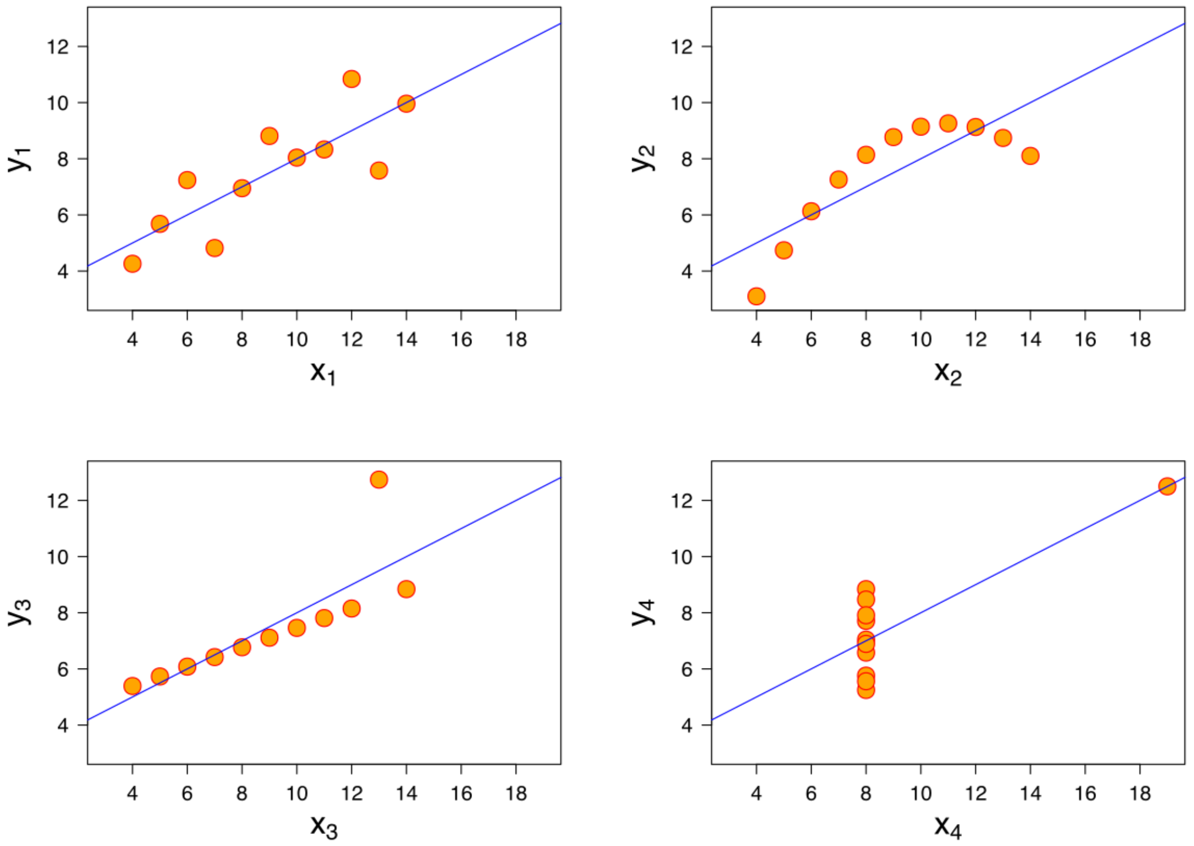
- In case of multiple feature, the predicted variable is a hyperplane instead of line. The predicted variable takes the following form:

$Y = \beta0+ \beta1x1+ \beta2x2+ \beta3x3+.....+ \beta nxn$

- The predicted variable is than compared with test data and assumptions are checked.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quarter comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically.The simple statistics consist of mean,sample variance of x and y , correlation coefficient ,linear regression line and R-square value.Anscombe’s Quarter shows that multiple data sets with many similar properities can still be vasty different from one another when graphed.The graphs are shown below.



1-First plot (top left) appears to be simple linear relationship

2.The second plot (top right) is not distributed normally and correlation coefficient is irrelevant as it shows a nonlinear relationship

3.The third plot (bottom left) is linear but has different regression line. This is happening because of the outliers present in the data

4. The fourth plot (bottom right) does not show linear relationship however due to outliers the statistic

3. What is Pearson’s R?

Pearson’s R measures the strength of association of two variables. It is the covariance of two variables divided by the product of their standard deviation. It has a value from +1 to -1.

- ❖ A value of 1 means a total positive linear correlation. It means that if one variable increase then will also increase
- ❖ A value of 0 means no correlation
- ❖ A value of -1 means a total negative correlation. It means that if one variable increase then will decrease.

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

Scaling of a variable is performed to keep a variable in certain range. Scaling is a pre-processing step in linear regression analysis. The reason we scale a variable to make the computation of gradient descent faster. The step size of gradient descent are generally low of accuracy, if the data has some small variables(values in the range of 0-1) and some big variables(values in the range of 0-1000) than the time taken by gradient descent algorithm will be huge.

Normalization/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.

Standardization Scaling:

 Standardization 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). 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.

Formula of Normalized scaling:

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

Formula of Standardized scaling:

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

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

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

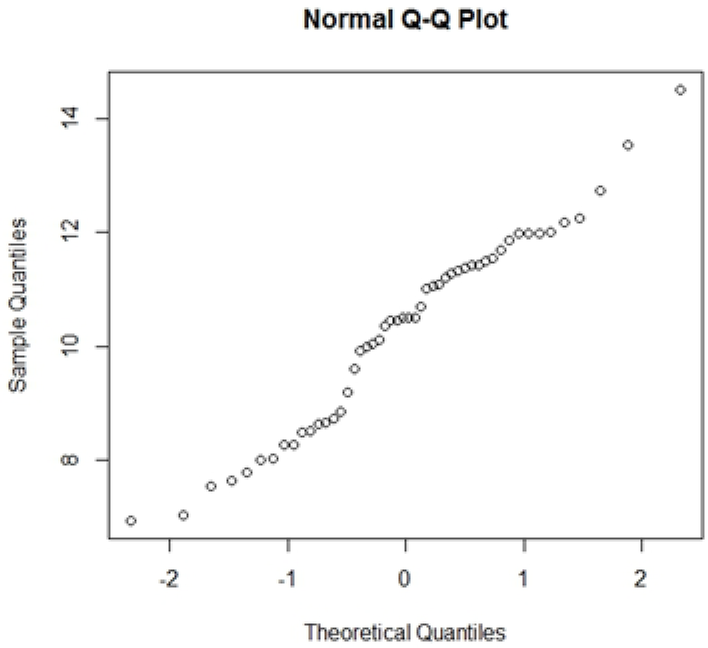
Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



<u>Use of Q-Q plot in Linear Regression</u>: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

<u>Importance of Q-Q plot: Below are the points:</u>

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.