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)
- There is a more usage in months Apr-Oct
- There is an increase with year
- Summer and Fall sees more use
- Clear weather sees more use

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
When we use drop_first = True, there is one less dummy variable created. Encoding for a categorical variable with n values can effectively be done with n-1 dummy variables. This will help to reduce the number of columns in our dataset and its effect will be apparent when there are more categorical variables. If we choose not to use drop_first, we can find the number of occurrences of the categories and we can choose to drop the least used one.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Both temp and atemp has the highest correlation with cnt. Temp and atemp is also highly correlated with each other.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
1. Linear relationship between variables -> pair plots and correlation matrix/ heatmap
2. Error terms are normally distributed -> Used distplot (normal curve) and QQ Plot (linear for normal distribution)
3. Error Terms mean is zero -> used numpy to calculate
4. Error Terms are independent of each other -> Correlation among variables with correlation matrix/ heatmap
5. Error Terms have constant variance (homoscedasticity) -> No visible patterns in the scatter plot created for residuals
6. Multicollinearity -> In summary statistics of OLS

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

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
- Get descriptive statistics metrics and visualise data and do an exploratory data analysis to understand the data
    - Impute missing values, convert data types as needed
    - Correlation among variables
    - Is there linearity
    - Mean, spread of data
- Prepare the data for modelling

- Split data into train and test
  - Add dummy variables for categorical variables
  - Scale the numeric data
- Train the model
  - Choose features via RFE
  - Use OLS or gradient descent to arrive at the optimal model
  - Repeatedly check the p-value (not more than 0.05) and VIF (not more than 5) for the features chosen by RFE and do a manual elimination of these.
- Residual Analysis
  - One of the assumptions of linear regression is residuals, when plotted, follows normal distribution.
  - Find residuals and plot with a distribution plot or a QQ plot
- Predict for test set
  - Use transform (not fit) on the test data set and predict using the optimal model got from training dataset
  - Find R-Squared score – this should be similar to the R-squared for the training dataset. If similar, it means that the model is generalising well.
- Evaluate the model
  - check VIF for the final coefficients (no multicollinearity)
  - f-statistic is high
  - probability of f-statistic is low
  - Durbin-Watson value near 2 (no autocorrelation (relationship of actual and previous value of data) in residuals)

2. Explain the Anscombe's quartet in detail. (3 marks)
Anscombe's quartet underlines the importance of visualizing the data. The Anscombe dataset and other sets like datasaurus dozen datasets has very similar descriptive statistics measures – mean, variance and correlation between the 2 variables. But their visualization is significantly different. Anscombe's quartet, when visualized, has different inferences and different ways of building a regression model or calculating a correlation coefficient.

3. What is Pearson's R? (3 marks)
This is the Pearson's correlation coefficient and it gives the measure of linear correlation between 2 datasets. The range of this is between -1 and + 1 where negative values denote a negative correlation (dependent variable decreases with increase of independent variable) and a positive value indicates a positive correlation (vice-versa). The higher the positive/ negative value, stronger is the correlation. This is only for data with linear correlation. For other kinds of data, some other coefficients like Spearman's is used.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Data ranges can be varied in a dataset. When used directly, this un-scaled data can produce beta coefficients that seem to influence the regression model more but this will not be the case. We need to scale the values to a common range so that our model can be accurate. There are 2 types of scaling – normalized and standardized. Normalization transforms the data between 0 and 1 while standardized scaling is the distance between the point with dataset mean and divided by the standard deviation and this will transform the data into a

range of negative to positive values. Normalization is done for normally distributed data while standardization is done for data with other distributions.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
When VIF is infinite, it means 1-R^2 is 0. Which implies R^2 is 1. R-Squared explains the variance of a dependent variable that is explained by the independent variable. If all of the variance is explained by one variable, the supposed independent variables are highly correlated with each other (they could be redundant) and there is a severe multicollinearity problem.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Quantile-Quantile plots usually compares 2 probability distributions by plotting their quantiles against each other. If the distributions are similar, then the plot should be near the 45degree line we plot.
In linear regression, when we plot the residuals, it should be as close to the 45 degree line showing that the residuals have a normal distribution. (Normal distribution translates to the 45 degree line in Q-Q plot)