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)

From the analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualisation

1. The bike rentals have increased in 2019 compared to 2018, YoY growth.
2. More bikes are rented in the months of August to October.
3. In Summer, Fall and Clear days more bikes are rented.
4. Bike rentals are more in moderate temperature (26 - 30)
5. Bike rentals are more in thursday, friday and saturday
6. More bikes are rented on holidays.

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

Because when we use encoding techniques to convert categorical variables into another form an extra column is created. 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)

'temp' variable has the highest 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)

Validated the assumption of Linear Regression Model based on

- Normality of error terms: Error terms are normally distributed. Plotted a histogram and check for distribution
- Multicollinearity check: There is insignificant multicollinearity among variables. Using a heatmap checked for the variables.
- Linear relationship validation: Verified using scatter plots between dependent and independent variables.
- Homoscedasticity: There should be no visible pattern in residual values. Checked by plotting of residual vs y_pred plot
- Independence of residuals: No auto-correlation. Durbin-Watson Statistics will always assume a value between 0 and 4. Using durbin_watson function our model showed 2.088 which is almost equal to 2, so we can say that there is little or no auto-correlation

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

- Year
- Temp
- Light Snow

General Subjective Questions

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

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with a given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Linear regression is of the following two types:

- Simple Linear regression : Where only one independent variable is present and the relationship between this and the dependent variable has to be found.

- Multiple Linear regression : Where multiple independent variables are present and the model has to find the relationship with the dependent variable.

Assumptions - The following are some assumptions about the dataset that is made by Linear Regression model:

- Multicollinearity – in Linear regression model assumes that there is very little or no multicollinearity in the data.
- Auto-correlation – Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data.
- Relationship between variables – Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms – Error terms should be normally distributed
- Homoscedasticity – There should be no visible pattern in residual values.

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed, each dataset consists of eleven coordinates.

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

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be

positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

Feature Scaling is a technique to standardise the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Mainly two Scaling approaches exist:

Normalization/Min-Max scaling : In this approach all of the data is brought into range 0 and 1. Sklearn.preprocessing.MinMaxScalar helps to implement normalisation.

Standardization scaling: replaces the values by their z scores. It brings all of the data into a standard normal distribution which has mean zero and std deviation one. Sklearn.preprocessing.scale helps to implement Standardisation in python.

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. Large value of VIF indicates that there is a correlation between the variables.

VIFi = 1 / (1 − Ri2). If the R-squared value is equal to 1 then the denominator of the above formula becomes 0 and the overall value becomes infinite.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction of points below the given value. That is, the 0.3quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.