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)

Ans: For Season: Count is maximum for fall season.

For weathersit: Count is maximum in weathersit 1 (Clear, Few clouds, Partly cloudy, Partly cloudy) and is minimum in weathersit 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

For months: Counts in middle months of the year is higher than in the beginning and the end.

For year: Significant increase in number of counts in 2019 from 2018.

For weekday: Count median is almost same for all days.

For working day: Count is higher for working day.

For holiday: Count is lower on holidays.

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

Ans: If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. If we don't do that, we would have an extra column that adds no new information that the previous columns don't provide when looked together. It will just increase multicollinearity. Drop_first = True helps us in achieving this. In our assignment we have year as 2018 and 2019, after using drop_first = True we have column of 2019 only. 2018 is default value if not 2019. Similarly, with month and week days one value is less which acts as default if no value is present.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: Looking at pair-plot from numerical variables temp and atemp has highest correlation with target variable cnt with value of 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

1. Linear Relationship: Linear regression assumes that there exists a linear relationship between the dependent variable and the predictors. Pair-wise scatterplots are helpful in validating the linearity assumption.
2. Homoscedasticity: It means that the residuals have constant variance no matter the level of the dependent variable. Look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.

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

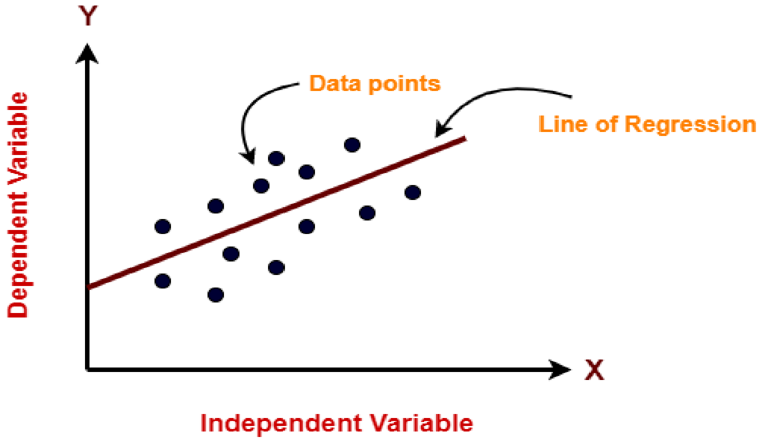
Ans:

1. Temperature
2. Weather category 3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
3. Mid-year months. (May to September)

General Subjective Questions

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

Ans: In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset. So, regression is a machine learning technique where the model predicts the output as a continuous numerical value. Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.



In the figure above, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

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

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. All the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

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

Ans: The Pearson correlation coefficient also known as Pearson's r, is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; Thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

Ans: Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalized Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardized scaling: It 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

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

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) 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).

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.