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)

Answer: The following observations were made :

- Overall growth in the industry is very evident in all the bar graphs with clear indication of growth from year 2018 to 2019
- May, June and July months on an average are the more popular months as per box plots where not only the rentals are at rise, but also the lower whisker value is noticeably higher than the median values of January and December months. From the bar plot it is evident that Sept of 2019 was highest grosser.
- In the Rentals vs. Weather Situations we see a decrease in the rentals in light snow. Also, notice the absence of the value 'Heavy Rain' from the plot altogether. This could be because of lack of data or we can say people avoid biking in heavy rain due to safety hazards.

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

Answer: We only need m-1 variables to describe m values of a column. For example, if gender has two values, you only need to show M as 1 or 0 it will be sufficient to classify M and F as gender because if the value of M as 1 suggest male and 0 suggest female automatically.

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

Answer: The pair-plot suggest that target variable cnt is having highest correlation with the registered and casual columns. This is because the cnt column is derived from summation of casual and registered.

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

Answer: There are four assumptions associated with a linear regression model:

Linearity: The relationship between X and the mean of Y is linear.

Homoscedasticity: The variance of residual is the same for any value of X.

Independence: Observations are independent of each other.

Normality: For any fixed value of X, Y is normally distributed.

The training set was analysed against the predicted values of the target variable from the finalized model. The residual error analysis of the same resulted in a normalized distributed plot which is what we can say is our validation. This is because in Linear Regression the error distribution is assumed to be normalized at 0.

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

Answer: Based on the final model derived, weather situation, the year and season_spring are my highest explaining factors.

General Subjective Questions

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

Answer: The linear regression algorithm aims to find the best linear fit for a set of data points. We aim to find a linear formula for determining the The best linear fit is determined by minimising the square of the differences between the y-coordinates of the points and the line. If the data is given by {(xi,yi):i=1,...,n} the least squares line will be of the form y = beta_0 + beta_1 x. The regression algorithm aims to minimize the following summation (yi-beta_0-beta_1*xi)^2 over i = 1,...,n. To find this minima, the gradient descent method can be used. The values of beta_0 and beta_1 where the minima is obtained represent the best linear fit.

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

Answer: Anscombe's quartet are a set of four datasets with completely different shape when graphed but certain very similar statistics, for example, mean and variance of both the x and y coordinates, the line of best linear fit and the R^2 coefficient. This demonstrates the advantage/need of graphing the data and visually checking for patterns and outliers about a given dataset before doing any further data analysis.

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

Answer: The Pearson's R, also known as the Pearson's correlation coefficient, is the covariance of two random variables divided by the product of their standard deviations. The value of the Pearson's correlation coefficient lies between -1 and 1. If the value is 0, then there is no linear fit and if it's 1, the two random variables are positively correlated (implying, one increases with the other increasing) and if it's -1, then the two random variables are negatively correlated (one increases with the decrease of the other).

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

Answer: Scaling is a method to bring all the data points to the same observable range. Often times, data can come with different units, to reduce the effect of this, scaling is performed. Standardized scaling converts the entries into draws from a standard normal distribution or the Z-scores. It is done in the following way (xi-mean(x))/sd(x) where sd(x) is the standard deviation of the data and mean(x) is it's mean. Normalized scaling on the other hand, is also known as min-max scaling. It is done by (xi-min(x))/(max(x)-min(x)). It brings all the values to be in the range (0,1) whereas Standardized scaling does no such thing.

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

Answer: The VIF or Variance inflation factor is a measure of the multicollinearity in an ordinary least squares regression analysis. It estimates how much the variance of the regression coefficient is increased because of collinearity. A smaller VIF denotes a bad linear fit and a larger value of VIF indicates a good linear fit. A high VIF can also be indicative of duplicate rows in the data analysis. If the VIF is infinite, it means that the dependent variable is a perfect linear combination of the dependent variables.

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

Answer: A Q-Q plot is a tool to find out whether or not a given sample of values are from a Normal Distribution or not. Given a set of n values, to form a Q-Q plot, the following procedure is used. First, from the standard normal distribution, n samples are taken. Let the data and the samples taken from the normal distribution be arranged in ascending order to get {x1,...,xn} and {y1,...,yn} respectively. Then, plot {(xi,yi):i=1,...,n}. If this plot resembles a straight line, then the data points belong to a normal distribution. Q-Q plots are important in linear regression to evaluate if the error terms come from a normal distribution. For a good linear model, the error terms should come from a normal distribution.