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?

Answer: Here are some categorical variable from dataset and respective dependent variable ('cnt')

- Median bike rents are increasing year on as year 2019 has a higher median then 2018
- Overall spread in the months plot is reflection of season plot as fall months have higher median
- Week days and weekend have almost the same median.
- Clear weather is most optimal for bike renting

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

Answer: A variable with n level can be represented by n-1 dummy variable. So if we remove the first column then we can represent the data. If the value of variable from 2 to n is 0, it means that the value of 1st variable is 1.

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

Answer: By looking at the pair plot TEMP variable has the highest (0.63) correlation with target variable 'cnt'

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

Answer:

Validate the assumptions:

- Best be tested with scatter plots [For examples depict two cases, where no and little linearity is present.
- Linear regression analysis requires all variables to be multivariate normal.
- Best be checked with a histogram or a Q-Q-Plot.

- 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.

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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes

1. weathersit_Light_Snow(negative correlation)
2. Year_2019 (Positive correlation)
3. temp(Positive correlation)

General Subjective Questions

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

   Answer:

   - Linear regression algorithm is a supervised learning method where predicted output is continues and constant slope.
   - It is explain the relationship between a dependent and an independent variable using a straight line (slope), intercept and linear coefficient
   - Independent variable also called as “Predictor variable” and dependent variables as “Output variables”
   - Simple linear regression and multiple linear regression
2. Explain the Anscombe’s quartet in detail. (3 marks)

   Answer:

   - Anscombe's quartet having the four data sets that have nearly identical simple descriptive statistics (mean, variance, standard deviation etc, it is very different distributions and appear very different in graphed.), each dataset contain eleven(x,y) points.
   - They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
   - It has 4 sets with 11 points to get the standard deviation, mean and correlation between x & y
   - It will visualizing our data is important as summary statistics.

3. What is Pearson’s R? (3 marks)

   Answer:

- Pearson's r (also called as Pearson correlation coefficient or bivariate correlation) is a numerical summary of the strength of the linear association between the variables. It depends on the variables trends go up and down together, the correlation coefficient will be positive.
- It 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. It also has a numerical value that lies between -1.0 and +1.0.
- Pearson's correlation coefficient is the covariance between two variables divided by the product of their standard deviations

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

Answer:

Scaling – scaling is a technique to make them closer to each other or in simpler words. It is standardized the independent feature in the data in a fixed ranges. It will change the range of data.

Scaling performed – It will help to data normalization and removed the complexity of data in future use.

Difference between Normalized Scaling and Standardized Scaling

Normalized Scaling	Standardized Scaling
Dataset values that range between 0 and 1	Data values that range mean of 0 and standard deviation of 1 (Z-Score)
It is really affected by outliers.	It is much less affected by outliers.
Features are of different scales When used.	When we want to ensure zero mean and unit standard deviation.

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

Answer:

- It will show perfect correlation between two independent variables.

- Infinite VIF value indicates the corresponding variable may be expressed exactly by linear combination of other variables.

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

Answer:

- Quantile-Quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
- Q-Q plots, It is theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.
- It helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let's discuss some of the probability distributions.