Linear Regression Assignment

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)

- The demand for bikes has seen an increase in 2019 as compared to 2018
- There is most demand during the fall season and the least demand in the spring season.
- From the months May through October we see the trend for maximum bike demand.
- Bike demand is also high when the weather situation is - Clear, Few clouds, Partly cloudy, Partly cloudy
- Weekday or holiday does not cause a major difference in the demand for bikes.

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

- While creating dummy variables, we use k-1 variables for a variable with k values. The first variable becomes redundant and cause dummy variables to be highly correlated. Collinearity between dummy variables is also reduced by deleting the first variable using drop_first=True.

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

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
- Residual analysis - the distribution of residuals should be normal with a mean around zero. We use a distplot to visualise this

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 (Positive correlation)
- Light rain light snow (negative correlation)
- Spring (negative correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear Regression is a form of regression used when the target variable is a continuous variable. It is an interpolation technique to predict the influence of an independent variable on the dependent variable.
   - The equation used is: y=m1*x1 + m2*x2 + m3*x3….+mn *xn +c where y is the target variable and x1…xn are predictor or independent variables.
   - Post exploratory analysis, data is split into test and train. The train dataset is then checked for collinear variables, high VIF, high p-value and R-squared values with iteratively dropping variables or adding variables to come to a final model.
   - Checks to perform - residual analysis - the distplot of distribution of residuals should be normal with a mean around zero.
   - The test data is then tested on the final model and R-squared is checked.

2. Explain the Anscombe's quartet in detail. (3 marks)
   - Anscombe's quartet is a group of four datasets which seem very identical descriptive statistics and appear extremely different when graphed.
   - It highlights the importance of visualising data to confirm the validity of any data relationships.

3. What is Pearson's R? (3 marks)
   - Pearson's R is a measure of of how strong the correlation is between 2 variables.
   - The value lies between +1 and -1, with values closer to +1 indicating a positive linear correlation while values closer to -1 indicate a negative linear correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - During data preprocessing we have a step to normalise data within an appropriate range. This is also used to speed up calculations. If scaling is not done, a dataset varying highly in magnitude will result in incorrect modelling as the algorithm will primarily take magnitude into account.
   - Normalisation or min-max scaling brings all data within the range of 0 to 1.
     - X = x-min(x) / max(x) - min(x)
   - Standardisation Scaling replaces values with their Z scores.
     - X = x - mean(x) / sd(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   - VIF = 1/(1-R^2)
   - If R^2 is high and close to 1, the denominator will become 0 making VIF infinite, denoting perfect correlation in variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
   - Q-Q or Quantile-Quantile plots, plot quintiles of a sample distribution against quantiles of a theoretical distribution. This helps in understanding how the variables are distributed with respect to another. This is useful specially to determine if the samples are from the same population.