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?**
   In the dataset, the categorical variables are namely season, yr, mnth, holiday, weekday, workingday, weathersit.
   Their effect on the target variables is visualized using a box plot.
   Inference:
   a. Season: the count of bike sharing is least in spring and high in fall
   b. Yr: there was an increase in bike sharing in the year 2019 when compared to 2018
   c. Mnth: there is a surge in bike sharing from May to September
   d. Holiday: there is surge in bike sharing from May to September
   e. Weekday: not much effected
   f. Workingday: not much difference
   g. Weathersit: most rentals in clear weather but no rentals in heavy rains and snow

2. **Why is it important to use drop_first=True during dummy variable creation?**
   It helps in reducing the extra column created during dummy variable creation. It also reduces correlation among the dummy variables.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Residual Analysis: By plotting histogram of the error terms. Residual distribution should follow a normal distribution.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Atemp: coeff is 0.4409. positively correlated
   Yr: coeff is 0.2327. positively correlated
   Light snow: coeff is -0.2577. negatively correlated

Next page for General Subjective questions

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   It is a type of supervised machine learning algorithm which is used for the prediction of numerical variables. It is based on the equation of a straight line “y=mx+c”. it tries to find the best fit line which can explain the relationsip between the target variable and independent variables. It is broadly classified in into:

   a. Simple Linear Regression(SLR): when the target variable is predicted using one dependent variable.

   b. Multiple Linear Regression(MLR): when the target variable is predicted using multiple predictors.

   The formula for MLR besomes:

   Y=m1x1 + m2x2….mnxn + c

2. **2. Explain the Anscombe’s quartet in detail.**

   Anscombe’s Quartet can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It emphasizes the importance of visualizing the data before applying various algorithms to build models. It suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data etc.

3. **What is Pearson’s R?**

   Pearson’s R also known as Pearson’s correlation coefficient is the test statistics that measure the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

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

   Scaling is bringing down all the independent features within a common scale for ease of interpretability and to avoid weird coefficients

   Scaling should be performed for ease of interpretation and faster convergence of gradient descent.

   Normalized Scaling: All the values are scaled between 0 and 1 by using minimum and maximum values of the data

   Standardized Scaling: the variables are scaled in such a way that the mean is zero and the standard deviation is one.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   VIF is infinite when there is a perfect correlation between two predictors and one variable can be expressed exactly by the linear combination of another.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Q-Q(quantile-quantile) plots are used to compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.
   It answers whether the two datasets comes from population with the common distribution.