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: For season- Fall, Summer, and Winter there is high demand in rental bike sharing compared to spring season- which shows very low demand. There is sharp increase in amount of rental bike sharing from 2018 to 2019, nearly 100% hike. Month playing very interesting role, from January to June there is steady rise in demand of rental bike sharing, which remains steady till September and after that it declines in stable way. Holiday, weekday and working day predictor variables do not have much impact on rental bike sharing-- all levels of categorical variable have same effect. Clear and Mist weather have high sharing of rental bike while light_snow weather has very low demand.

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

   Answer: In dummy variable creation if we will not use drop_first then it will create all separate columns for all levels of categorical variables but when we set it as True then N-1 columns will create where N is total level of categorical variable. Thus, we can encode categorical variables with less dimensions.

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

   Answer: temperature and feeling temperature

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

   Answer: 1) check linearity between dependent and independent variables using scatter plot, 2) calculate residuals (y_train-y_train_pred), plot distplot of residual- plot should be normally distributed with mean zero 3) check variance of residual by plotting scatter plot of residual vs y_train (actual data points) – there should not be any visible pattern means all residual points must be independent with each other.

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

   Answer: Weathersit, year and season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer: Linear regression is a supervised machine learning algorithm which explain the relationship between dependent variable (continuous) and single or more independent variables. It will try to find best fit line or best fit plane between dependent and independent variables. There are few assumptions on which algorithm work:1) there is linear relationship between dependent and independent variables, 2) residuals (RSS) should be normally distributed for each data point, 3) All residuals should be independent with each other 4) Variance in dependent variable should be constant.

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a set of four datasets that have identical statistical properties, but are vastly different when graphed. It emphasizes on the importance of graphical data analysis and caution against to rely solely on statistical summary. The quartet demonstrates that summary statistics can be misleading when dealing with complex data. Graphical analysis is essential to understanding the nature of the data and any underlying relationships. Anscombe's quartet serves as a powerful reminder to statisticians and data analysts to not rely solely on summary statistics, but to explore and visualize the data before drawing conclusions or making predictions.

3. What is Pearson's R?

   Answer: Pearson's R is also known as pearson's correlation coefficient. It measures the linear relationship between the two continuous variables, it ranges from -1 to +1, where +1 shows exact positive correlation, -1 shows exact negative correlation while 0 shows no relation between continuous variables.

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

   Answer: Scaling is a pre-processing technique in which features of a dataset are transformed to specific range without changing distribution. If scaling is not performed then features will be on different scales, so those features which have high scale then they will have high coefficients and model would end up with bias. It will be easy to compare the features since they will have same scale after scaling. Normalized scaling is also called min-max scaling which ranges between 0 & 1. Standardized scaling is z-score scaling in which data transformed with mean 0 and standard deviation of 1.

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

   Answer: VIF-Variance Inflation factor is a measure of multi-collinearity between predictor variables. Normally, when VIF value >=5 , then we can say that feature is highly correlated with other predictor variables and we can drop that feature. But sometimes when VIF becomes infinite then we can say that more than 1 features are perfectly correlated with each other. So, there would be very high multi-collinearity among predictor variables.

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

   Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a dataset to a theoretical distribution. The plot displays the quantiles of the data against the quantiles of a specified theoretical distribution. In linear regression, Q-Q plots are used to assess the normality assumption of the residuals. By examining the Q-Q plot of the residuals, we can visually inspect whether they follow a normal distribution. If the plot shows a straight line, this indicates that the residuals follow a normal distribution, and the normality assumption is satisfied. However, if the plot shows a curved or nonlinear pattern, this indicates that the residuals are not normally distributed and may violate the normality assumption.