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)

   Ans - As we can see from the jupyter notebook, box plots plotted against the categorical variables, following conclusions can be drawn:

   - mnth – Jun. Jul has the highest demand for rental bikes as compared to other months
   - yr: The demand for the rental bikes has increased in the succeeding year
   - holiday: The demand on non-holiday is more as compared to the holiday
   - weekday: There is not much conclusiveness in between the weekdays as far as demand of rental bikes are considered.
   - weathersit: Bad weather contributes greatly in the downfall for demand of rental bikes

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

   Ans – It is important because it helps in eliminating the use of extra variable. Therefore, it helps in reducing the correlations among different independent variables. Generally, if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. In essence, it avoids redundant features.

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

   Ans - The numerical variable 'temp' has the highest correlation with the target variable 'cnt' As we dropped 'atemp' to avoid multicollinearity in the data.

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

   Ans- This is done using residual error analysis:

   - As analysed in the jupyter notebook, residual error follows the normal distribution
   - Plotting distplot, regplot and qq plot to understand error terms.
   - Which conclude there is a linear relationship between the predicted values (y_pred) and the actual values (y_test)

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

   Ans- Based on the model, following 3 featured contribute significantly towards the such explanation:

   - `temp` - A coefficient value of `4826.1702` indicated that a unit increase in temp variable increases the bike hire numbers by 4826.1702 units.

   - `Light snow/rain` - A coefficient value of `-2273.2060` indicated that, a unit increase in Light snow/rain variable decreases the bike hire numbers by -2273.2060 units.

- yr - A coefficient value of 2007.8235 indicated that a unit increase in yr variable increases the bike hire numbers by 2007.8235 units.

General Subjective Questions

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

   Ans: Linear Regression is a Machine Learning algo used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variables. So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.

   Hypothesis function for Linear Regression:
   Y = a + bx
   where y = dependent variable
   a = intercept b = coefficient of x
   x = independent variable

   By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the a and b values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y). Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y). To update a and b values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. It starts with random a and b values and then iteratively updating the values, reaching minimum cost.

   A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is.

   The R-squared values range between 0 and 1. A value of 0.835 means that the explanatory variable can explain 83.5 percent of the variation in the observed values of the dependent variable.

   A value of 1 means that a perfect prediction can be made. A value of 0 means the explanatory variable does not help at all in predicting the dependent variable.
   Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

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

   Ans- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is generally used to

illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties. They have very different distributions and appear differently when plotted on scatter plots. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order 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, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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

   Ans- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. For example, a child's height increases with his increasing age (different factors affect this biological change). So, we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r.
   There are certain requirements for Pearson's Correlation Coefficient:
   - Scale of measurement should be interval or ratio
   - Variables should be approximately normally distributed
   - The association should be linear
   - There should be no outliers in the data

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

   Ans- It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
   Scaling just affects the coefficients and none of the other parameters like t-statistic, F□statistic, p-values, R-squared, etc.

   Normalization Scaling:
   - It brings all the data in the range of 0 and 1.
   - sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores.
- It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

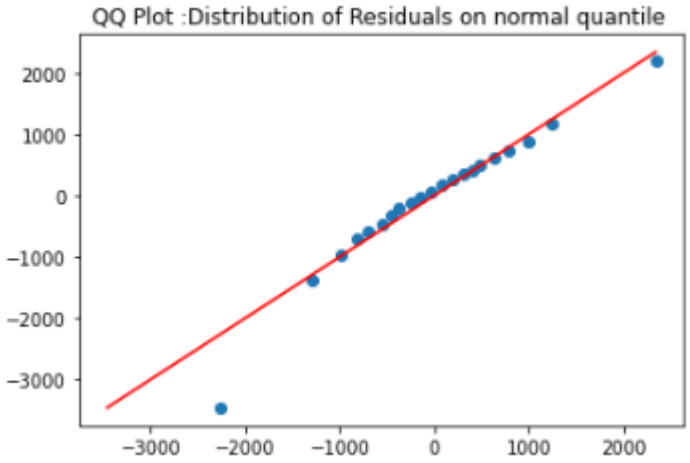
5. . You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

   Ans: If VIF is infinity, then it means it has a perfect relation. A large value of VIF indicates that there is a correlation between the variables.
   In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

   Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
   These are the plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

   A Q Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.