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. The categorical variables evidence the below:

- weathersit: The bike demand is highest during clear weather and drops as the weather moves to misty and then to snow.
- weekday: the medians of the weekdays are very similar and distributed, therefore carry little significance to the bike demands and it can be dropped
- season: carries significance to the bike demand and highest demand is in fall, followed by summers.
- month: the bike demand peaks during the mid of the year and is flatter near year end and beginning, it can be directly related to the existing weather during these months.
- holiday: bike demand is lower on a holiday
- workingday: does no have a significant impact on the bike demand and hence can be dropped
- yr: year on year bike demand is increasing

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

Ans. Drop first drops the first variable during the dummy set creation and hence doesn't widen the dataframe, the number of variables created from the a categorical variable with x values are therefore x-1.

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

Ans. Temperature has the highest correlation with target variable.

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

Ans.

- Assumption 1: There is a linear relationship between the dependent variables and the independent variables- verified by a scatterplot of the prediction set vs test set.
- Assumption 2:The independent variables are not too highly correlated with each other- verified by plotting a heatmap correlation matrix after selecting the features of the model.
- Assumption 3: Residuals should be normally distributed with a mean of 0 and variance σ. – verified by residual analysis histogram.

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.

1. atemp – highest significance(positive correlation)
2. light_snow_rain: negative correlation
3. yr : positive correlation

General Subjective Questions

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

Ans. Linear regression algorithm is a machine learning algorithm based on supervised learning performing a regression task. Regression builds a model based on independent variables and uses it to predict a target variable. It statically calculates the relationship between the independent variables and target variable and uses it to predict values of target variable based on linear relationship using the equation of a straight-line y =mx+c.

In case of multiple variables, the equation takes form of $y=m_1 x_1+m_2x_2....\ m_nx_n+\ c\ +e$

where m_n is the coefficient of the variable x_n and c is the intercept constant, e is the error resulting due to error terms arising in the equation.

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 known as the cost function calculated as RMSE of the difference between predicted value of y and true value of y.

This is done by assigning random values to m & c initially and then iterating their values till the cost function reduces at each iteration reaching a minimum. This is known as gradient descent.

An example of a linear algorithm regression is a prediction model for stock prices basis various variables like price trends, similar business stocks etc. Another example is a weather prediction model basis humidity, pressure, temperature etc.

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

Ans. Anscombe's quartet is a compiled dataset to emphasise the importance of the visual illustrations by using graphs. It uses four datasets of values for x & y.

Studying the values in the dataset do not have much of resemblance to any patterns in the y values against x, however they appear relevant on a scatterplot.

It was designed to clear the common understanding that graphs are rough and mathematical calculations are exact.

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

Ans. Pearson's R is the measure of the linear correlation between two dataset. It is the ratio between the covariance of two variables(bivariate correlation) and the product of their standard deviations; hence it is a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

For example, the relation between height of kids in a primary school and their age is positively linearly correlated and shall have values above zero and below 1. For negative correlation, the value shall be between 0 & -1.

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

Ans. Scaling is a technique used for normalising the data within a particular range during the pre-processing phase.

Scaling is performed to process the data within a particular range as data may be in different units (like time expressed in minutes or hours in different columns) and may be largely varying in magnitude. If scaling is not done then algorithm shall only take the magnitude into account and may process incorrect results, moreover it also speeds up the algorithm. Scaling only affects the coefficients of the variables and no other analytical values.

Normalized scaling is also known as min max scaling and calculates values between 0 & 1's. Whereas Standardization replaces the values by their Z scores. It brings the data into a standard normal distribution which has mean zero and standard deviation one.

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

Ans. Variance Inflation factor is indicator for the correlation of a variable to all other variables in a dataset. It is calculated by VIF = $1/(1-r^2)$, here if the r^2 value is taken as 1 for a perfectly linear model(or for a overfitted model), then denominator reduces to zero and hence VIF value as infinity.

The infinite values of VIF arise when the model is perfectly linear by observations or due to overfitting of the variables, hence a careful analysis of the data is essential prior fitting into the model and the highly correlated variables to be dropped, else the model shall follow the observations instead of the fitted line.

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

Ans. Q-Q plot, Quantile-Quantile is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, exponential or Uniform. It also helps to determine if two data sets come from populations with a common distribution.

In linear regression if the training and test data sets were received separately and then we can confirm that both the data sets are from populations with same distributions using a Q-Q plot.