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 : There are about 7 categorical variables **namely season, yr, mnth, holiday, weekday, workingday, weathersit**.

Based on the analysis using thee boxplot on each of these variables, here are the some of the observations:

- Fall and summer has the maximum bookings. It's easier to ride bikes during these seasons and there will be less booking expected whenever it rains.
- The year 2019 has more booking compared to previous year. The number of users have gone up due to the ease of COVID restrictions.
- Months during middle of year has highest booking which corresponds to the seasons of summer and fall.
- There are more bookings when the weather is clear and no rain, thunderstorm, scattered cloud.

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

Ans : This config is used during dummy variable creation to remove the redundancy and It will also help in dealing with multicollinearity scenario which will result in high correlation among multiple independent variables in the model.

It is also important to drop the additional variables which doesn't add value to the model.

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

Ans : 'temp' and 'atemp' has variables has the highest correlation with the target variable 'cnt' looking at the pairplot.

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

Ans : The assumptions are made based on the following.

- **Residual analysis** : The error terms should be normally distributed. This plays an important role in deciding the model can be inferred. The mean value should be 0 in order to decide if the error terms are equally distributed.
- **Error terms are independent of each other :** The predictor variables are independent of each other. Multicollinearity is addressed using VIF for all the variables and it has the values below 5.

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 : The top 3 features contributing significantly towards explaining the demand is -

- Temp with the coeffient of 0.3905
- Year with the coefficient of 0.2364
- September with the coefficient of 0.0647

General Subjective Questions

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

Ans : Linear regression is an algorithm which computes the relationship between a dependent variable and independent variable. It can compute on one or more independent variables.

There are 2 types of Linear regression :

- Simple Linear regression
- Multiple Linear regression.

Simple Linear regression : Simple linear regression is performed where there is single independent and dependent variable. The objective is to figure out residual between the predicted values and actual values. It is represented by the following equation.

Formula : Y = B0 +B1*X + E

Y – Dependent variable
X – Independent variable
B0 - Intercept
B1 - Slope
E – Error

The Mean Squared Error(MSE) is the cost function which will be used to find the difference between the actual and predicted values.

It also uses the optimization techniques like gradient descent.

Multiple Linear regression : Multiple regression is performed on multiple independent variables.

Assumptions of Linear Regression :

- Linear relationship between dependent and independent variables.
- Error terms are normally distributed
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

Steps to achieve the linear regression :

- **Data preparation** : Read the dataset and analyse to perform the Exploratory Data Analysis (EDA) on the dataset. This also includes the data cleaning and derivation.
- **Data Visualization** : Plot the charts to check the linearity between the variables.
- **Dummy variables** : Identify the categorical variables and create the dummy values on each of it.
- **Test and train data** : Segregate the data for training the model and testing the model. General standards would be use 70% of data to train the model and 30% to test data.
- **Rescaling** : Data need to be rescaled to the values between 0 and 1. We can use MinMaxScaler to rescale the feature.

- **Model building** : Building the model holds the significance part in the linear regression. Using the techniques like RFE and manual approach we can eliminate the features which are not relevant to the model building. We need to use VIF and check the P value to eliminate the features as well.
- **Residual analysis** : After the model building phase, the prediction of the model come into the place. In this step we can calculate the predictions for the target variables and verify the linearity.

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

Ans : Anscombe's quartet, consists of four datasets with similar basic statistics but different distributions. Its purpose is to stress the importance of visualizing data rather than relying solely on summary statistics.

The datasets in Anscombe's quartet have similar average, spread, correlation, and regression properties, but they display distinct patterns when graphed. This emphasizes the drawback of depending solely on summary statistics to comprehend a dataset. Anscombe's quartet is frequently employed to underscore the significance of visually exploring and graphically analyzing data.

- In the first one, there will be a linear relationship between x and y.
- In the second one, there will be a non-linear relationship between x and y.
- In the third one, the graph indicates the outliers
- Finally, the fourth one, shows the high corelation with the single data point.

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

Ans : The Pearson's R or the Pearson's coefficient is a coefficient representing the relationship between two variables, i.e. X and Y. It measures the strength

The Pearson's coefficient ranges between -1 to +1, where :

- -1 represents the negative corelation
- +1 represents the positive corelation
- 0 represents no relation at all

The key features of the Pearson's coefficient include :

- Measuring the linear association between variables and assuming that the relationship between them can be represented in a straight line.

- The correlation coefficient is symmetric, meaning that the correlation between varible X and Y is same as variable Y and X.

- Extreme outliers will have a huge impact on the coefficient correlation.

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

Ans : What is scaling? Scaling is a process of converting the values of features of a dataset to a standard range. The aim is to bring all the features in the same range of values, which will help the algorithm to learn the learn the patterns from the data.

Why is scaling performed? The data collected may contain the features that are highly varying in range and units. If scaling is not performed there are chances that algorithm will take only the magnitude into the account and not the units and it will lead to the incorrect modelling.

What is the difference between normalized scaling and standardized scaling?

- Normalized scaling converts all the data into the range of 0 and 1. It is also called as MinMaxScaling and it is imported from sklearn library.
  The below formula is used for normalizing.

  MinMaxScaling : x = x-min(x) / max(x) – min(x)

- Standardized scaling bring all the data into a standard normal distribution which has mean zero and standard deviation 1.

  Standardized : 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)

Ans : The scenario of infite VIF occurs where there is a perfect correlation. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. This is due to perfect R-squared. Value of 1 which will lead to 1/(1-R2) infinity. In order to solve this we need to drop one of the variables from the dataset.

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

Ans : The quantile-quantile plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically the normal distribution. It is also used for checking the assumption of normativity in a dataset.

How to Draw Q-Q plot :
- Collect the data and sort the data.
- Draw a normal distribution curve.
- Find the z-value (cut-off point) for each segment.
- Plot the dataset values against the normalizing cut-off points.

Explain the use and importance of a Q-Q plot in linear regression -
A Q-Q plot used as a tool in regression analysis to validate key assumptions necessary for reliable inference. Specifically, it enables the examination of whether the residuals in a model adhere to a normal distribution. This is crucial because many parametric tests and confidence intervals rely on the assumption of normality in residuals. By employing a Q-Q plot, researchers can visually assess whether the residuals align with the expected normal distribution, helping ensure the robustness of statistical analyses.