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: I have done analysis on categorical columns using the boxplot and scatter plot. Below are the few points we can infer from the visualisation –

   - Fall season have highest bookings
   - Booking increase towards mid of the year, specifically from june to sept
   - Bookings are more if weather is clear
   - Bookings are more in 2019 year as compare to 2018
   - Bookings are less on holidays as people spend time with family.
   - Booking seems to be almost equal on working and non working days
   - Booking seems to be similar on all days of week with slightly more on Thursday to Sunday.

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

   Answer: drop_first=true is important to avoid extra unwanted columns during dummy variable creations. For example, in a given data set, season is a categorical variable with 4 possible values. Now you can create just 3 dummy variables and get the same information instead of 4 dummy variables. As having 0 in all 3 dummy variables means 1 for the 4th possible variable.
   Syntax for adding dummies for season column is below :
   dummies = pd.get_dummies(bike_rental_data.season, drop_first = True)
   bike_rental_data = pd.concat([bike_rental_data, dummies], axis = 1)

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

   Answer: Looking at the pair plot, temperature has the highest correlation with the target variable which is 0.63. Same can also be seen in heatmap as well.

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

   Answer: I have validated the assumption of Linear Regression Model based on below assumptions –
   - Plotted the distPlot for residual analysis and confirmed that residuals are normally distributed and centred around zero.
   - Calculate VIF for all independent variables and ensure that no variable has more than 5 VIF and hence verify the Multicollinearity of the variables with other variables.
   - Plotted scatter plot for residuals which shows horizontal straight line which proves homoscedasticity

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

   Answer: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes
   - Temperature with coefficient 0.48
   - September with coefficient 0.06
   - Year with a coefficient 0.23 tells us the popularity of bikes is increasing every year.
   - Winter with coefficient 0.05

General Subjective Questions

1. Explain the linear regression algorithm in detail

   Answer: Linear regression is a statistical method used for modelling the relationship between a target variable and one or more independent variables by fitting a linear equation to the observed data. Intent is to find the best-fit linear equation with minimum difference between predicted and actual target variable.

   - Linear regression is of two types :
     - Simple Linear Regression
     - Multiple Linear Regression

   - Sample linear equation for Single linear regression looks like below -
     - $Y = \beta_0 + X \beta_1$

   - Following are the assumption we consider while doing linear regression -
     - Residuals are normal distributed and centred around zero
     - Independent Variables does not have strong correlated with each other

   - Main steps involved in performing linear regression
     - Preparing/Cleaning data set and visualise relations using pairplots
     - Splitting the data set into train and test data sets.
     - Scale the columns if required.
     - Building the linear model and identify key variables using technic like RFE
     - Residual Analysis to verify assumptions
     - Test prediction on the model build.

2. Explain the Anscombe’s quartet in detail.

   Answer: Anscombe's quartet emphasised the importance of EDA in understanding data sets. He explained such data sets which have similar mean, variance, correlation and linear regression but on plotting graphs, differ significantly when graphically visualised. One of the reasons for this can be the presence of outliers in the data set. Hence, relying solely on numerical summaries can be misleading, and graphical exploration is essential for uncovering nuances in data patterns

3. What is Pearson's R?

Answer: Pearson's correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1

- r=1 indicates a perfect positive linear relationship.
- r=−1 indicates a perfect negative linear relationship
- r=0 indicates no linear relationship.

R = 1 or -1 indicates a strong relationship while R = 0 indicates weak relationship.
Example - In finance, Pearson's r might be used to measure the correlation between the return of two stocks, helping investors understand how changes in one stock's value relate to changes in another.

4. What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Answer: Scaling is a preprocessing step in data analysis and machine learning where the values of variables are transformed to be within a specific range. The goal is to ensure that no variable dominates the others in terms of scale. Scaling is performed for below reasons:

- Scaling ensures that coefficients or feature importance scores in linear models are comparable across variables. For example, if we have a distance in thousands, then the coefficient might be small and it gives an interpretation that distance has less impact as compared to other variables whereas if we scale it down in 0 to 1 range, then it can give more appropriate coefficient values.
- Few machine learning algo are sensitive to scale and hence, we should always check for scaling as pre request if applicable before model building.

Difference between normalised scaling and standardised scaling:

- Normalised scaling is also known as min-max scaling, helps to transform the values of a given column in range from 0 to 1. On the other hand, Standardised scaling helps to transform values of column with mean 0 and standard deviation 1
- Formula for normalised scaling is
  X(normalised) = X - X(Min) / X(Max)- X(Min)
  where as formula for standardised scaling is
  X(standardised) = X - mean / standard deviation

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

Answer: VIF is a measure used to quantify the extent of multicollinearity in a regression analysis. Formula of VIF is as below:
VIF(j) = 1 divided by (1- R(j) square) where R is the regression of jth variable, assuming the jth variable as dependent variable and other variable has independent variables.
VIF is infinite only when we are getting R(j) as 1 which signifies perfect multicollinearity among variables which means that this variable can be easily perfectly predicted with the help of other independent variables. One hypothetical An example can be having two columns with exactly the same values. In the real world scenario, it is very rare to get R(j) as 1 and there will be differences in the values even at decimal level.

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

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess the normality of a distribution by comparing observed samples to the theoretical normal distribution.

Generally this is used in residuals analysis in linear regression where we can check how close the normal distribution of residuals are from the ideal normal distributions. Ideal normal distribution is with mean zero and standard deviation 1. Having closure to 45 degree in Q-Q plot shows that distribution of residuals are closer to normal distributions. If there are a significant number of points having variations from 45 degree diagonal, then it is an indicator of residuals being not normally distributed and hence rejecting our assumption in linear regression.