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)

Answer:

The categorical columns are 'season','year','month','holiday','weekday','workingday','weathersit' in the dataset.

A function has been created in order to obtain the box plot in visualizing the categorical variables

From the observations the following points can be noted

1) From season column During the summer and fall the total count of booking are comparatively high
2) From Year column it can be noted that the total count of booking increase drastically in the year 2019
3) From Month column it can be noted that there are more bookings from May to October
4) From holiday column we can see that on the holidays the number of bookings are apparently lesser but there are only fewer number of holiday and it would be in appropriate to conclude it.
5) From weekday we can see that the later days of the week such as Thursday, Friday, Saturday, Sunday have higher number of bookings when compared to the Monday, Tuesday and Wednesday
6) From working day, it can be noted that the total bookings are almost equal on the working and non-working days
7) From weather we can note that the total bookings are high whenever the weather is clear.

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

When creating dummy variables from categorical variables, the drop first=True parameter is used to prevent multicollinearity and maintain consistency in the model
Importance: Avoiding the Dummy Variable Trap ensure consistent interpretation of coefficients, and maintain model efficiency.

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

"Temperature", and "feeling temperature" have the highest correlation among the numerical variables

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

Answer:

Validations that are made on the Assumptions of Linear Regression are

Normal distribution of error terms:

1) Error terms should be normally distributed. We validate this assumption about residuals by plotting a distplot of residuals

Multicollinearity check:

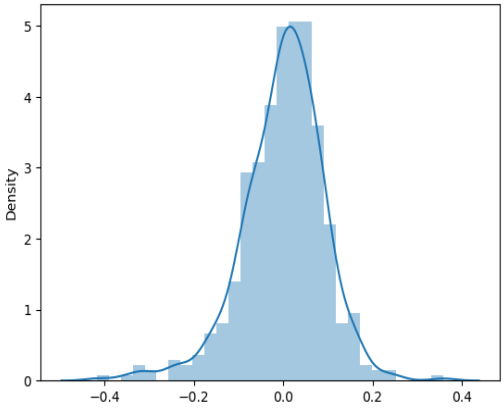
2) There should be insignificant multicollinearity among variables.

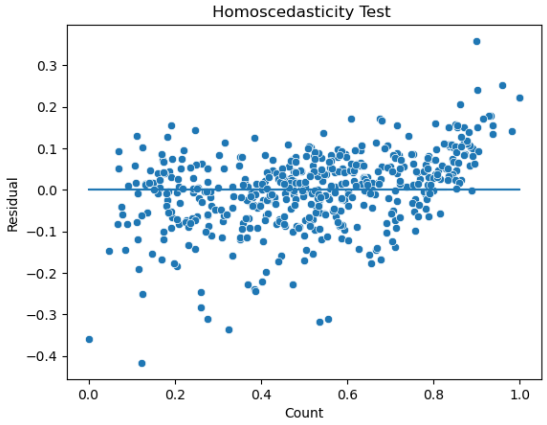
Linear collinear and relationship validation:

3) Linearity should be visible among variables

Homoscedasticity:

4) There should be no visible pattern in residual values. This means they should have similar variance throughout the distribution and this can be validated by making a scatter plot of residuals and a horizontal line passing through 0.





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

Answer:

By looking at the parameter of the model the top 3 features contributing significantly towards explaining the demand of the shared bikes are

1) Temperature 0.4777
2) Year 0.2341
3) weathersit_Light_snow/rain -0.2850

General Subjective Questions

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

Linear regression is a statistical method that uses a line to predict the value of a dependent variable based on the value of one or more independent variables. The line is called the regression line, and it is found by minimizing the sum of the squared errors between the predicted values and the actual values. The linear regression algorithm can be expressed as follows:

$y = mx + b$

where:

- y is the dependent variable
- m is the slope of the line
- b is the y-intercept
- x is the independent variable

Detailed explanation of the linear regression algorithm building:

i. Problem Formulation:

- Identify the problem or question you want to answer.
- Determine the dependent variable (also known as the target variable or output variable) that you want to predict based on the independent variables (also known as features or input variables).

ii. Data Collection:

- Gather a dataset that contains observations of both the dependent and independent variables.
- Ensure the dataset is representative and has enough variability to capture the relationship between the variables.

iii. Data Preprocessing:

- Clean the dataset by handling missing values, outliers, and other data quality issues.
- Split the dataset into training and testing subsets.

iv. Model Representation:

Linear regression assumes a linear relationship between the independent variables and the dependent variable, represented by the equation:

$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$

v. Training the Model:

- Train the model by utilizing the training dataset that is obtained after splitting the complete dataset.

vi. Model Evaluation:

- Checking if the model satisfies the assumptions of linear regression model.

vii. Making predictions utilizing the model.

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but when visualized, they reveal starkly different patterns. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in understanding and interpreting statistical analysis.

The four datasets in Anscombe's quartet have the same statistical properties:

1. Dataset I:
- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
2. Dataset II:
- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26
3. Dataset III:
- x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
- y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
4. Dataset IV:
- x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
- y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91

Despite having the same means, variances, correlations, and regression lines, each dataset has a unique pattern when plotted. This highlights the limitations of relying solely on summary statistics and reinforces the importance of visualizing data.

When visualized, the differences become apparent:

1. Dataset I: Forms a relatively linear relationship between x and y.
2. Dataset II: Shows a non-linear relationship, where the data points resemble a quadratic curve.
3. Dataset III: Contains a single outlier that significantly affects the linear regression line.
4. Dataset IV: Appears to have a clear linear relationship except for an outlier, which has a substantial impact on the regression line.

Anscombe's quartet demonstrates that statistical measures alone cannot capture the complexity of relationships within data. It emphasizes the need for visual exploration and understanding the data's underlying patterns, outliers, and potential pitfalls in drawing conclusions based solely on numerical summaries.

The quartet serves as a cautionary example and encourages data analysts and statisticians to visualize data before performing analyses to gain a more comprehensive understanding of the data and avoid erroneous assumptions.

3. What is Pearson’s R? (3 marks)

Pearson's R or simply as the correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of the association between two continuous variables.

Pearson's R is a value between -1 and 1, where:

- A value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Normalization scales the data to a fixed range, such as [0, 1] or [-1, 1]. This is done by subtracting the minimum value of each feature from the feature and then dividing by the difference between the maximum value and the minimum value
Standardization scales the data to have a standard deviation of 1. This is done by subtracting the mean of each feature from the feature and then dividing by the standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   Variance Inflation Factor (VIF) is a measure used to assess multicollinearity, which is the presence of high correlations between predictor variables in a regression model. VIF quantifies how much the variance of the estimated regression coefficient is increased due to multicollinearity.

In some cases, the VIF value can be infinite or extremely high. This occurs when one or more predictor variables can be perfectly corelated
If there is perfect correlation, then VIF = infinity

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data against the quantiles expected from the theoretical distribution.
Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
The use and importance of a Q-Q plot in linear regression are as follows:

- Checking Normality Assumption
- Detecting Skewness and Outliers
- Assessing Model Fit