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

- Most number of bookings are made in Fall season.
- September has highest number of bookings while median for bike rental count is highest for July. Trend of booking bikes increases from July and by October the booking trend decreases as winter approaches.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

Answer: It is crucial to employ drop_first=True when generating dummy variables to mitigate the problem of multicollinearity in regression models. Multicollinearity arises when there is a high correlation among two or more independent variables in a regression model, resulting in redundant information and complicating the model's interpretation. When dummy variables are being constructed, opting for drop_first=True eliminates one dummy variable for each categorical variable. This action establishes a reference category by dropping the initial dummy variable, with the remaining ones indicating the presence or absence of other categories. This strategy is instrumental in preventing perfect correlation among the dummy variables.

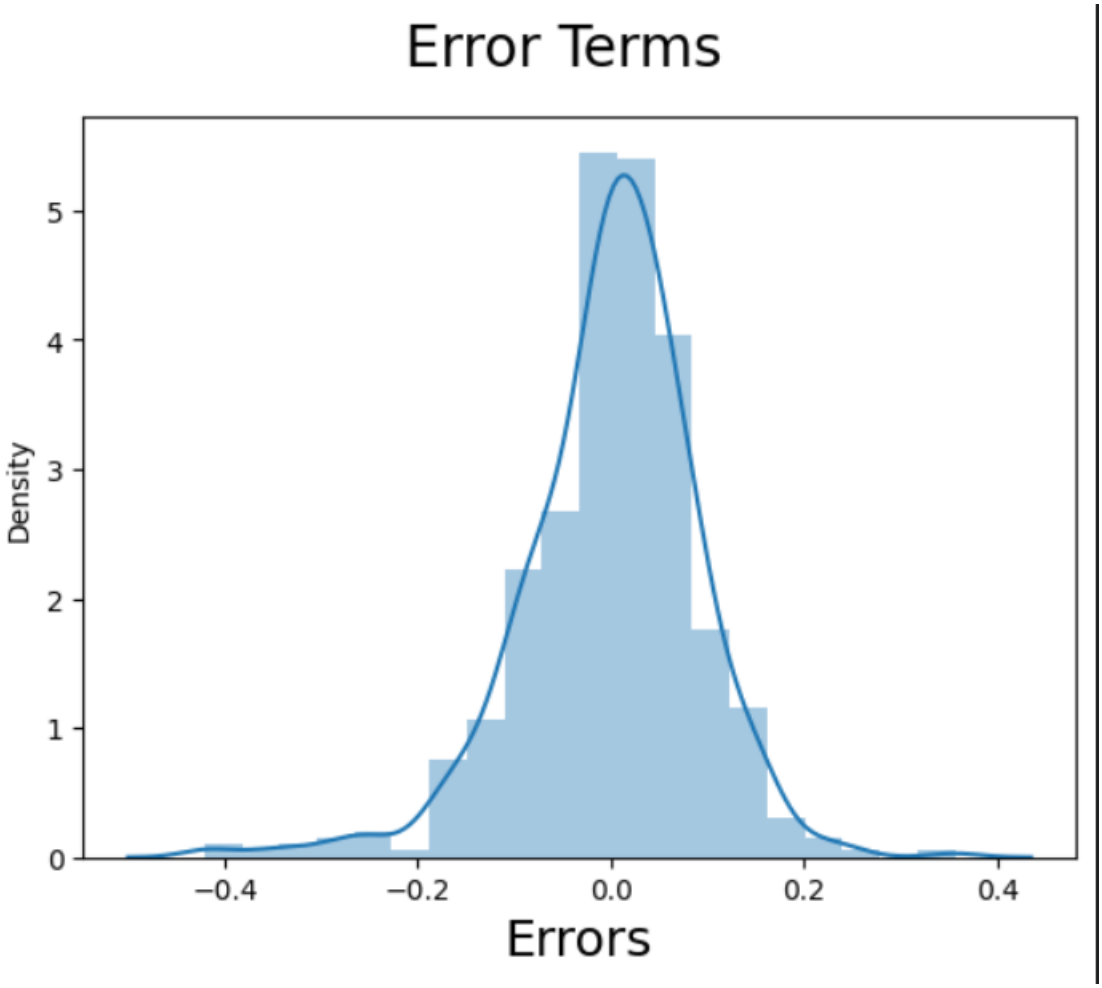
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Looking at the pairplot before building the model, 'temp' and 'atemp' variable has the highest correlation with the target variable.

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

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

1. **Error/Residual analysis**: For Linear regression problems, error terms should be normally distributed (which is in fact, one of the major assumptions of linear regression). My Linear regression error graph is also normally distributed as shown below:



2. **Error terms are independent of each other**: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are 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)

Answer: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp (Coef: 0.4923)
- year (Coef: 0.2338)
- sep (Coef: 0.0721)

General Subjective Questions

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

Linear regression is a popular and widely used algorithm for predicting a continuous numeric output based on one or more input variables. It assumes a linear relationship between the input variables and the output variable. Here's a detailed explanation of the linear regression algorithm:

- Data Preparation:

    - Gather a dataset that contains the input variables (also called features or independent variables) and the corresponding output variable (also called the target or dependent variable).
    - Split the dataset into a training set and a test set. The training set is used to train the linear regression model, while the test set is used to evaluate its performance.
- Model Representation: o Linear regression aims to find the best-fitting linear equation that represents the relationship between the input variables and the output variable.
    - In its simplest form, linear regression is represented by the equation: Y = β0 + β1*X1 + β2*X2 + ... + βn*Xn, where Y is the predicted output variable, β0 is the intercept, and β1 to βn are the coefficients corresponding to X1 to Xn (input variables).

- Model Training:
    - During training, the linear regression model adjusts the values of the coefficients (β0, β1, β2, ..., βn) to minimize the difference between the predicted values and the actual values in the training data.
    - The most common approach to training is called Ordinary Least Squares (OLS), which minimizes the sum of squared differences between the predicted values and the actual values.
- Model Evaluation:
    - Once the model is trained, it is evaluated using the test set to assess its performance on unseen data.
    - Common evaluation metrics for linear regression include mean squared error (MSE), root mean squared error (RMSE), and R-squared, which measures the proportion of variance in the target variable that is explained by the model.
- Making Predictions:
    - After the model is trained and evaluated, it can be used to make predictions on new or unseen data.
    - Given new input values, the model applies the learned coefficients to calculate the predicted output variable.
- Assumptions of Linear Regression:
    - The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables.
- Assumptions about the residuals:
    - Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
    - Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
    - Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.
    - Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.
- Assumptions about the estimators:
    - The independent variables are measured without error.
    - The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.
- Variations and Extensions:

- There are variations and extensions to linear regression, such as multiple linear regression (with more than one input variable), polynomial regression (using polynomial terms), and regularization techniques like Ridge regression and Lasso regression to handle overfitting and improve generalization.

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

Answer: Anscombe's quartet refers to a set of four datasets that have nearly identical simple descriptive statistics but exhibit significantly different patterns when visualized and analyzed. These datasets were created by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization and the limitations of relying solely on summary statistics. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Pearson's correlation coefficient, commonly referred to as Pearson's R or simply as the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who developed this measure in the late 19th century.

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

- A value of +1 indicates a perfect positive linear relationship between the variables, meaning that as one variable increases, the other variable increases proportionally.
- A value of -1 indicates a perfect negative linear relationship between the variables, meaning that as one variable increases, the other variable decreases proportionally.
- A value of 0 indicates no linear relationship between the variables, suggesting that there is no systematic change in one variable when the other variable changes.

Key features of Pearson's R:

- Linearity: Pearson's R measures the linear association between variables. It assumes that the relationship between the variables can be represented by a straight line.
- Scale-invariant: Pearson's R is unaffected by changes in the scale or units of measurement of the variables.

- Symmetry: The correlation coefficient is symmetric, meaning that the correlation between variable A and variable B is the same as the correlation between variable B and variable A.
- Affected by outliers: Extreme outliers can have a strong influence on the correlation coefficient, as it is sensitive to extreme values.

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

Answer: Scaling is the process of transforming the features of a dataset to a standard range. The goal is to bring all features to a similar magnitude, making it easier for machine learning algorithms to learn patterns from the data. Scaling is particularly important for algorithms that are sensitive to the scale of the input features, such as distance-based methods (e.g., k-nearest neighbors) and optimization-based methods (e.g., gradient descent).

Why Scaling is Performed:

- Equal Weightage: Scaling ensures that all features contribute equally to the computation of distances, coefficients, or other measures in machine learning algorithms. Without scaling, features with larger magnitudes may dominate the learning process.
- Convergence: For optimization algorithms like gradient descent, scaling can help the algorithm converge faster by providing a more balanced landscape for the minimization process.
- Performance: Scaling can improve the performance of models that use distance-based metrics, such as k-nearest neighbors or support vector machines.
- Regularization: Regularization methods, such as L1 and L2 regularization, are sensitive to the scale of the input features. Scaling helps to ensure that regularization is applied uniformly across all features.

Normalized Scaling:

Normalized scaling, also known as Min-Max scaling, transforms the features to a specific range, usually between 0 and 1. The formula for Min-Max scaling is:

$X_{\text{normalized}}=(X-X_{\text{min}})/(X_{\text{max}}-X_{\text{min}})$

Standardized Scaling:

Standardized scaling, also known as z-score normalization, transforms the features to have a mean of 0 and a standard deviation of 1. The formula for standardization is:

$X_{\text{standardized}}=(X-\mu)/\sigma$

Difference between Normalized Scaling and Standardized Scaling:

- Range:
    - Normalized scaling brings the values within a specified range (typically 0 to 1).
    - Standardized scaling transforms the values to have a mean of 0 and a standard deviation of 1.

- Sensitivity to Outliers:
    - Normalized scaling may be sensitive to outliers as it depends on the minimum and maximum values.
    - Standardized scaling is less sensitive to outliers because it uses the mean and standard deviation.
- Interpretability:
    - Normalized scaling retains the original distribution and interpretability of the data within the specified range.
    - Standardized scaling centers the data around the mean, making it more interpretable in terms of standard deviations from the mean.

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

Answer: 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. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

A Quantile-Quantile (Q-Q) 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. Q-Q plots are particularly useful for checking the assumption of normality in a dataset.

How a Q-Q Plot Works:

- Ordered Data: The observed data points are first sorted in ascending order.
- Theoretical Quantiles: Corresponding quantiles from a theoretical distribution (e.g., the normal distribution) are calculated.
- Plotting: The ordered data quantiles are plotted against the theoretical quantiles.

If the data follows the theoretical distribution (e.g., normal distribution), the points on the Q-Q plot should fall along a straight line. Deviations from a straight line suggest departures from the assumed distribution.

Use and Importance in Linear Regression:

Q-Q plots are essential in linear regression for several reasons:

- Normality Assumption: Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed. Checking the normality of residuals is crucial for the validity of statistical inferences.

- Model Assumption Checking: The Q-Q plot helps assess whether the residuals follow a normal distribution. If the points on the Q-Q plot deviate from a straight line, it suggests that the residuals may not be normally distributed.
- Identifying Outliers: Q-Q plots can help identify outliers or skewness in the data. Outliers may appear as points that deviate significantly from the expected straight line.
- Data Transformation: If the Q-Q plot indicates non-normality, it might be necessary to consider data transformations to make the residuals more normally distributed. Common transformations include logarithmic or square root transformations.
- Decision Making: The results of the Q-Q plot influence decisions about the appropriateness of the linear regression model. If the residuals are approximately normally distributed, it adds credibility to the regression analysis. Conversely, departures from normality may warrant further investigation or alternative modelling approaches.