Assignment-based Subjective Questions

QUESTION -1

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

ANSWER -

- The inference that We could derive were:
- Season: Almost 32% of the bike booking were happening in season3(fall) with a median of over 5000 booking (for the period of 2 years). This was followed by season2(summer) & season4(winter) with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- Mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- Weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

  weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

  workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

QUESTION – 2

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

ANSWER –

- Use of drop_first=True is very important because by using drop_first=True helps prevent multicollinearity, ensures model identifiability, and simplifies the interpretation of regression coefficients. This practice is especially important in linear regression models and models that are sensitive to multicollinearity issues.

QUESTION – 3

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

ANSWER –

variables like atemp , temp etc. show a positive correlation with the target variable cnt.

QUESTION - 4

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

ANSWER -

- After building the model on the training set we can assumptions of Linear Regression as follows:
- **Linearity Assumption:**
- Check scatterplots of actual vs. predicted values: Plot the predicted values against the actual values to visually inspect if the relationship is linear.
- **Constant Variance Assumption:**
- Residuals vs. Fitted Values Plot: Examine the spread of residuals across the range of predicted values. Ideally, the spread should be constant.
- **Normality of Residuals:**
    - Q-Q Plot: Check the quantiles of the residuals against the quantiles of a normal distribution. Deviations from a straight line suggest departures from normality.

- Shapiro-Wilk test or Anderson-Darling test: These are statistical tests for normality. A low p-value indicates non-normality.

- **Multicollinearity:**
  - Variance Inflation Factor (VIF): Calculate the VIF for each predictor variable. High VIF values (typically above 10) may indicate multicollinearity.

Normality of Predictors :

- Check the distribution of predictor variables. If they are highly skewed or not normal, consider transformations or robust regression techniques.

QUESTION – 5

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

ANSWER -

- Temperature (temp) - A coefficient value of '0.375922' indicated that a unit increase in temp variable increases the bike hire numbers by 0.375922 units.

- Weather Situation 3 (weathersit_3)(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered) - A coefficient value of '-0.333164' indicated that, w.r.t Weathersit_3, a unit increase in Weathersit_3 variable decreases the bike hire numbers by 0.333164 units.

- Year (yr) - A coefficient value of '0.232965' indicated that a unit increase in yr variable increases the bike hire numbers by 0.232965 units.

General Subjective Questions

QUESTION – 1

Explain the linear regression algorithm in detail.

ANSWERE –

Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The goal is to minimize the sum of squared differences between observed and predicted values. It assumes a linear relationship and estimates coefficients through methods like least squares or gradient descent. The model's performance is evaluated, and it can be used for making predictions on new data.
It include various steps like:

- **Assumptions:** Linearity: Assumes a linear relationship between the independent and dependent variables.
- **The Linear Equation:** The linear regression model can be expressed as a linear equation:
- **Cost Function:** The objective is to minimize the sum of squared differences between the observed and predicted values.
- **Parameter Estimation:** The coefficients are estimated using methods like the least squares method. The goal is to find the values that minimize the cost function.
- **Gradient Descent :** Alternatively, iterative optimization algorithms like gradient descent can be used to minimize the cost function. The algorithm adjusts the coefficients step by step in the direction that reduces the cost.
- **Model Evaluation:** Once the model is trained, it's important to evaluate its performance using metrics or Mean Squared Error (MSE) on a separate test dataset.
- **Prediction:** After training and evaluation, the model can be used to make predictions on new, unseen data by plugging in the values of the independent variables into the linear equation.
- Linear regression is a simple yet powerful algorithm used in various fields for predicting numerical values. While it has its assumptions and limitations, it serves as a foundational concept for more complex regression techniques.

QUESTION – 2

Explain the Anscombe’s quartet in detail.

ANSWER –

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data

1. Dataset I:
    - Mean of x: 9.0
    - Mean of y: 7.5
    - Correlation coefficient: 0.816
    - Linear regression equation: $\hat{y}=3.0+0.5x$
    - Residual standard deviation: 1.25
2. Dataset II:
    - Mean of x: 9.0
    - Mean of y: 7.5
    - Correlation coefficient: 0.816
    - Linear regression equation: $\hat{y}=3.0+0.5x$
    - Residual standard deviation: 1.25
3. Dataset III:
    - Mean of x: 9.0
    - Mean of y: 7.5
    - Correlation coefficient: 0.816
    - Linear regression equation: $\hat{y}=3.0+0.5x$
    - Residual standard deviation: 1.25
4. Dataset IV:
    - Mean of x: 9.0
    - Mean of y: 7.5
    - Correlation coefficient: 0.816
    - Linear regression equation: $\hat{y}=3.0+0.5x$
    - Residual standard deviation: 1.25

Anscombe's quartet underscores the importance of data visualization in understanding the underlying patterns and relationships within a dataset. Relying solely on summary statistics can be misleading, as datasets with different structures may produce the same statistical summaries. Visualization helps identify outliers, patterns, and potential issues that may be overlooked when only looking at summary statistics.

QUESTION – 3

What is Pearson’s R?

ANSWER –

Pearson's correlation coefficient, often denoted as r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables.

- The closer r is to 1 or -1, the stronger the linear relationship.

- A positive r indicates a positive linear relationship, while a negative r indicates a negative linear relationship.

Pearson's correlation coefficient is widely used in various fields, including statistics, biology, economics, and social sciences, to assess the degree of association between two variables. However, it does not imply causation; a strong correlation does not necessarily mean one variable causes the other.

QUESTION – 4

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling ?

ANSWER –

Define : Scaling is a preprocessing step in data analysis and machine learning where the values of the features (variables) in a dataset are transformed to bring them to a similar scale. The goal is to ensure that no particular feature dominates the others, especially when using algorithms that are sensitive to the scale of the input features.

Why is scaling performed :

- **Algorithm Sensitivity:** Many machine learning algorithms are sensitive to the scale of the input features. Algorithms like k-nearest neighbors, support vector machines, and neural networks often perform better when features are on a similar scale.
- **Convergence Speed:** Gradient-based optimization algorithms, such as those used in linear regression or neural networks, may converge faster when features are scaled. It helps prevent certain features from dominating the updates during the optimization process.
- **Distance-Based Algorithms:** Algorithms that rely on distances between data points, like k-nearest neighbors, can be affected by the scale of the features. Scaling helps ensure that distances are computed more accurately.

Differences between Normalized Scaling and Standardized Scaling**:**

1. Scale Range:
    - Normalized Scaling: Scales the features to a specific range (e.g., [0, 1]).
    - Standardized Scaling: Centers the data around 0 and scales it to have a standard deviation of 1.
2. Distribution Shape:

- Normalized Scaling: Maintains the relative differences between data points and does not assume a specific distribution.
- Standardized Scaling: Assumes a Gaussian distribution and preserves the shape of the original distribution.

3. Sensitivity to Outliers:
    - Normalized Scaling: Can be sensitive to outliers, especially if the range is influenced by extreme values.
    - Standardized Scaling: Less sensitive to outliers due to the use of the mean and standard deviation.

QUERSTION – 5

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

ANSWER –

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the severity of multicollinearity in a set of predictor variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it challenging to assess the individual effect of each variable on the dependent variable.

If the VIF for a variable is infinite, it indicates a perfect linear relationship between that variable and the other predictors in the model. This perfect correlation leads to an R_i2 of 1, and when you plug this value into the VIF formula, you get an infinite result.

QUESTION – 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 whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is particularly useful in linear regression and other statistical analyses to check the assumption of normality and identify departures from normality.

Assumption of Normality:

- One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) are normally distributed.

Identifying Outliers:

- Outliers in the data or residuals can be detected by examining points that deviate significantly from the expected straight line in the Q-Q plot.

Model Adequacy:

- Checking the normality assumption is crucial for assessing the adequacy of the linear regression model.

Inference and Hypothesis Testing:

- Normality assumptions are often important for making statistical inferences and conducting hypothesis tests.

Residual Diagnostics:

- Q-Q plots are part of a broader set of residual diagnostic tools used to evaluate the assumptions of a linear regression model.

Q-Q plots play a crucial role in assessing the normality assumption of residuals in linear regression. They provide a visual way to identify deviations from normality, outliers, and potential issues with the regression model. Addressing such deviations is important for building reliable models and making valid statistical inferences.