Date: 03/14/2023

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?***

   After analyzing the categorical columns using boxplots and bar plots, the following conclusions can be made:

   1- Fall and Summer seasons has the highest number of bookings, and there has been a significant increase in bookings from 2018 to 2019 for each season.

   2-The majority of bookings were made in the months of May to October, with a gradual increase from the beginning of the year until the middle, followed by a decrease towards the end of the year.

   3-Clear weather conditions tend to attract more bookings, which is not surprising.

   4-Thursdays, Fridays, Saturdays, and Sundays have more bookings compared to the start of the week.

   5-Bookings tend to be less when it is not a holiday, as people may want to spend time at home with family.

   6-The number of bookings seems to be almost equal on working days and non-working days.

   7-There was an increase in the number of bookings in 2019 compared to the previous year, indicating good progress in the business.

   Based on these insights, it is recommended that BoomBike focus on promoting bookings during the fall season and the months of May to October. They should also consider advertising their services during periods of clear weather. Furthermore, they can focus on providing holiday packages and deals to increase the number of bookings during holidays. Finally, they can continue to analyze trends and adjust their business strategy to optimize bookings and profits.

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

   It is important to use drop_first=True during dummy variable creation to avoid the dummy variable trap. The dummy variable trap occurs when there is perfect multicollinearity among the dummy variables, which happens when one of the dummy variables can be predicted perfectly from the others. This can cause issues with the model fitting process, such as producing unreliable or incorrect coefficient estimates, increasing the standard errors, and reducing the model's predictive power. By dropping the first dummy variable, we remove one of the redundant variables, which helps to avoid the dummy variable trap and improves the accuracy and efficiency of the model.

The syntax for using drop_first is drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.
For example, if we have a categorical column with 3 types of values and we want to create dummy variables for that column, using drop_first=True would result in only 2 dummy variables being created, rather than the original 3. This can help to simplify the dataset and reduce the risk of overfitting.

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

According to the pair-plot analysis among the numerical variables, the variable 'temp' 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?

I have evaluated the assumptions of the Linear Regression Model based on the following 5 criteria:

1-Normality of error terms: The error terms should follow a normal distribution.

2-Multicollinearity check: There should be no significant multicollinearity among the predictor variables.

3-Linear relationship validation: The relationship between the predictor variables and the target variable should be linear.

4-Homoscedasticity: The variance of the error terms should be constant across all values of the predictor variables.

5-Independence of residuals: The error terms should be independent of each other and should not exhibit any pattern of autocorrelation.

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

Based on the coefficients of the final model, the top 3 features contributing significantly towards explaining the demand of shared bikes are:
temperature (temp) with a coefficient of 0.5471
year with a coefficient of 0.2328
Rain/Snowing (Winter)with a coefficient of -0.2883.
But From Understanding the data set, Weather, working days and holidays.

General Subjective Questions

1- <u>Explain the linear regression algorithm in detail.</u>

Linear regression is a statistical modeling technique used to establish a relationship between a dependent variable and one or more independent variables. The basic premise of linear regression is to fit a straight line through a set of data points such that the distance between the line and each data point is minimized. The fitted line can then be used to predict the value of the dependent variable for a given value of the independent variable.

In the case of simple linear regression, there is only one independent variable and one dependent variable, and the relationship between the two is assumed to be linear. The goal is to find the equation of a straight line that best represents the relationship between the two variables. The equation of the line is given by:

$y = \beta 0 + \beta 1x + \varepsilon$

where y is the dependent variable, x is the independent variable, β0 is the intercept, β1 is the slope or regression coefficient, and ε is the error term or random error.

The process of fitting a linear regression model involves estimating the values of the intercept and slope coefficients that best fit the data. This is typically done using a method called least squares, which involves minimizing the sum of squared errors between the actual values of the dependent variable and the predicted values based on the fitted line.

Multiple linear regression extends the simple linear regression model to include more than one independent variable. The equation for multiple linear regression is given by:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta pxp + \varepsilon$

where p is the number of independent variables.

The process of fitting a multiple linear regression model involves estimating the values of the intercept and regression coefficients that best fit the data, again using the method of least squares. The coefficients can be interpreted as the change in the dependent variable associated with a one-unit change in the corresponding independent variable, holding all other independent variables constant.

Linear regression is widely used in various fields, including finance, economics, engineering, and social sciences, for predictive modeling and data analysis.

Positive linear Vs Negative linear regression

In a positive linear relationship, as the value of one variable increases, the value of the other variable also increases. This means that if we plot the relationship on a graph, the

points will form a diagonal line that slopes upward from left to right. A positive coefficient in a linear regression equation indicates that there is a positive relationship between the two variables being analyzed. For example, if we are studying the relationship between temperature and ice cream sales, a positive coefficient would indicate that as temperature increases, so do ice cream sales.

In a negative linear relationship, as the value of one variable increases, the value of the other variable decreases. This means that if we plot the relationship on a graph, the points will form a diagonal line that slopes downward from left to right. A negative coefficient in a linear regression equation indicates that there is a negative relationship between the two variables being analyzed. For example, if we are studying the relationship between age and reaction time, a negative coefficient would indicate that as age increases, reaction time decreases.

In summary, the sign of the coefficient in a linear regression equation tells us whether there is a positive or negative relationship between the two variables being analyzed.

2- <u>Explain the Anscombe's quartet in detail.</u>

Anscombe's quartet is a set of four datasets that have identical statistical properties, yet appear very different when visualized. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data before analyzing it and to demonstrate the potential limitations of relying solely on summary statistics.

Each dataset in the quartet consists of 11 (x, y) pairs. The four datasets have different distributions of x and y values and different levels of correlation between the two variables, yet they all have the same mean, variance, correlation coefficient, and regression line.

The first dataset is a simple linear relationship between x and y. The second dataset is a non-linear relationship between x and y. The third dataset has an outlier that strongly influences the regression line. The fourth dataset has a strong relationship between x and y but only when one outlier is included.

By examining the quartet, it becomes clear that summary statistics alone cannot fully describe a dataset. Visualizing the data is important for understanding the underlying patterns and relationships. Additionally, the quartet demonstrates the importance of data exploration and the potential pitfalls of relying solely on mathematical models to analyze data.

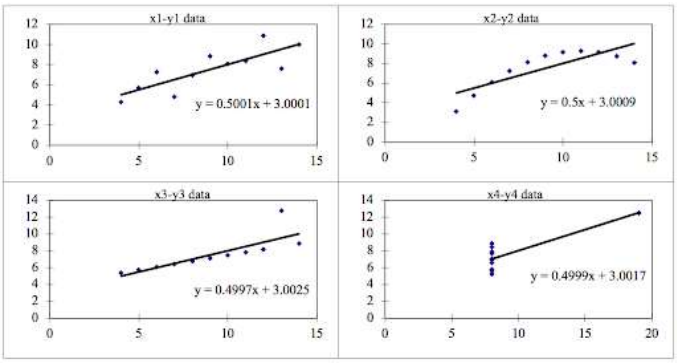

Fig: 1.0

Dataset I: This dataset has a linear relationship between x and y, with a correlation coefficient of 0.816. It appears to be a good fit for linear regression analysis. Fig: 1.0

Dataset II: This dataset is characterized by a non-linear relationship between x and y, with a clear outlier that is not well explained by the linear regression line. It highlights the danger of relying solely on the correlation coefficient to determine the strength of a relationship. Fig: 1.0

Dataset III: This dataset has a strong relationship between x and y that is not well explained by a linear regression line. It has a clear quadratic relationship that is better explained by a higher order regression model. Fig: 1.0

Dataset IV: This dataset has an extreme outlier that is influential in determining the regression line. The outlier heavily influences the correlation coefficient and demonstrates the importance of identifying and handling outliers in statistical analysis. Fig: 1.0

Overall, Anscombe's quartet highlights the importance of visualizing data and the limitations of relying solely on summary statistics like the mean, variance, and correlation coefficient. It emphasizes the need to examine data more closely and to use multiple methods of analysis to fully understand relationships and patterns in data.

3- ***What is Pearson's R?***

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that represents the strength and direction of the linear relationship between two variables. It is denoted by the symbol "r". Pearson's R ranges from -1 to +1, where -1

indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation.

Pearson's R is commonly used in statistics and data analysis to measure the relationship between two continuous variables. It assumes that the variables are normally distributed and have a linear relationship. Pearson's R is sensitive to outliers and can be affected by non-linear relationships between the variables.

Pearson's R can be calculated using the formula:

r = (nΣxy - ΣxΣy) / sqrt[(nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2)]

where:

n is the number of observations

Σxy is the sum of the products of the corresponding values of x and y

Σx and Σy are the sums of the values of x and y, respectively.

Σx^2 and Σy^2 are the sums of the squared values of x and y, respectively

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

Scaling is a preprocessing step in machine learning that transforms the feature values of a dataset to a specific range. It is done to ensure that each feature contributes equally to the analysis and to prevent features with large values from dominating the analysis.

Scaling is performed to bring all features onto the same scale, which makes it easier to compare them and draw meaningful conclusions from the analysis. It also helps to improve the accuracy and performance of machine learning models, especially those that rely on distance-based measures, such as K-Nearest Neighbors and clustering algorithms.

There are two common types of scaling: normalized scaling and standardized scaling. Normalized scaling (also known as Min-Max scaling) rescales the feature values to lie between 0 and 1. Standardized scaling (also known as Z-score scaling) rescales the feature values to have a mean of 0 and a standard deviation of 1.

Normalized scaling is used when the distribution of the data is not known or when the data is not normally distributed. It is a good choice when there are outliers in the dataset. Standardized scaling is used when the distribution of the data is normal or close to normal. It is a good choice when we want to compare the relative importance of different features.

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

The value of VIF (Variance Inflation Factor) can become infinite when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity means that one or more of the predictor variables can be linearly predicted from the other predictor variables. In such cases, the VIF of the variable that can be predicted perfectly becomes infinite because its variance cannot be estimated.

For example, if we have three predictor variables, A, B, and C, and A can be predicted perfectly from B and C, then the VIF of A will become infinite because its variance cannot be estimated independently from B and C.

It is important to note that infinite VIF values only occur in theoretical or simulated data, and rarely occur in real-world datasets. In practice, VIF values greater than 10 are usually considered as an indication of significant multicollinearity among the predictor variables.

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

A Q-Q plot, short for quantile-quantile plot, is a graphical technique used to determine if a dataset theoretical normal distribution.

In linear regression, Q-Q plots are useful for validating the assumption of normality of the error terms. Residuals should be normally distributed around a mean of zero, and the Q-Q plot can help to identify any departures from this assumption. If the residuals are normally distributed, the points on the Q-Q plot will fall approximately along a straight line. However, if there are deviations from a straight line, this may suggest that the errors are not normally distributed, and this may indicate a violation of the assumptions of linear regression.

The importance of the Q-Q plot lies in its ability to visually check the normality assumption of the residuals. If the assumption is violated, it can lead to biased or inefficient estimates of the model parameters. Therefore, the Q-Q plot provides a useful tool to check if the residuals are normally distributed, and if not, to consider possible remedies such as data transformation or the use of non-parametric methods.