Assignment Based Subjective Question Answers-

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3marks)

   Solution-

   Train R-Squared=0.824

   Train R-Squared adjusted=0.821

   Test R-Squared=0.820

   Test R-Squared adjusted=0.812

   Top three predictor variables

   temp (Temperature):

   -Per Unit Increase in temp yields bike bookings raised by 0.563615 times.

   weathersit_3 (Weather Situation 3):

   -Per Unit Increase in weathersit_3 yields bike bookings decreased by -0.306992 times.

   yr (Year):

   -Per Unit Increase in yr yields bike bookings raised by 0.230846 times.

Inferences-

A. Temperature increases (correlated with summers, a pleasant time in the US), leads more people outdoor for adventure. Therefore more rentals.

B. Weather Situation 3, Meaning light snow and showers, makes more people avoid 2 wheeler travelling, therefore a strong negative correlation with rise in rains and light snow.

C. Logical, as each year business grows.

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

   Solution-

   We do NOT need another column for "Uknown".

   It can be necessary for some situations, while not applicable for others. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.

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

Solution-

Temp has the highest correlation with cnt.

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

   Solution-

   1. Normality of error terms - Error terms should be normally distributed

   2. Multicollinearity check o There should be insignificant multicollinearity among variables.

   - Linear relationship validation o Linearity should be visible among variables

   3. Homoscedasticity - There should be no visible pattern in residual values.

   4. Independence of residuals- No auto-correlation

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

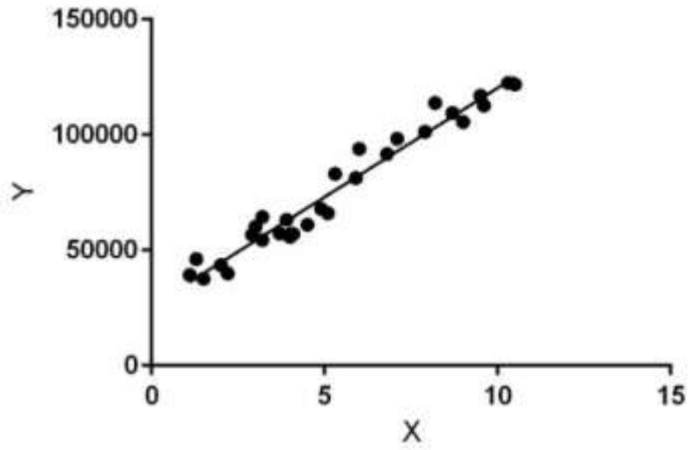
   Solution-

   1. Temp
   2. Weathersituation3
   3. Yr(Year)

General Subjective Questions

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

   Solution Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on - the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model
Hypothesis Testing for linear regression

$Y=B_0+B_1*X$

While training the model we are given : x: input training data (univariate - one input variable(parameter))

y: labels to data (supervised learning)

When training the model - it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best B_0 and B_1 values.

B_0: intercept

B_1: coefficient of x

Once we find the best B_0 and B_1 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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

Anscombe's quartet is a famous dataset in statistics that consists of four sets of data that have nearly identical statistical properties but look very different when plotted. This dataset was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.
Each set in Anscombe's quartet contains 11 data points with two independent variables (x and y). Despite having the same means, variances, correlations, and linear regression coefficients, the plots of the data look quite distinct, showcasing the limitations of relying solely on summary statistics for understanding data.

The quartet is often used to highlight the following key points:

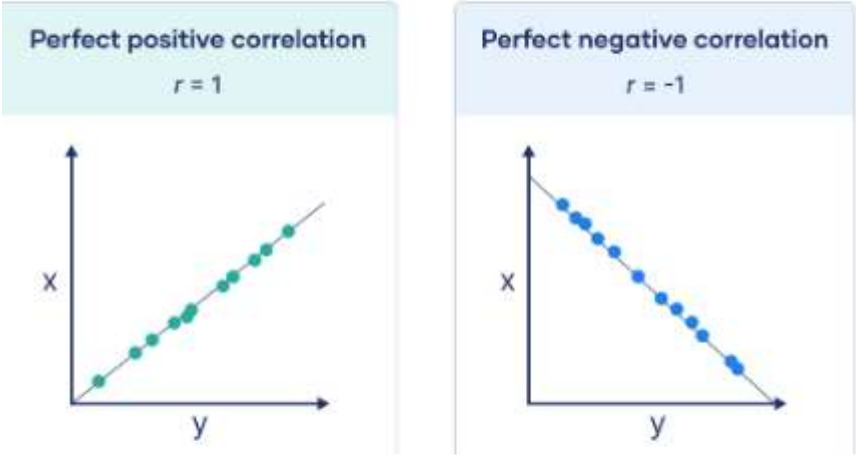
1. Graphical Visualization: It demonstrates the importance of graphical visualization in understanding data patterns and relationships.
2. Diverse Relationships: The quartet shows that different datasets can exhibit various relationships (linear, non-linear, outliers, etc.) even if their summary statistics are similar.
3. Outliers and Influential Points: One dataset in the quartet includes an influential outlier that significantly affects the regression line, demonstrating the importance of detecting outliers.
4. Regression and Correlation: It illustrates that while summary statistics like means, variances, and correlations provide useful information, they might not capture the complete picture.

   Anscombe's quartet serves as a reminder to explore and visualize data comprehensively before drawing conclusions, making predictions, or fitting models. It also highlights the limitations of relying solely on numerical measures without considering the underlying data distribution and relationships.

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

Solution

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It measures how well the relationship between two variables can be described by a straight line.



The Pearson correlation coefficient has values between -1 and 1:

- -1: A perfect negative linear correlation. As one variable increases, the other decreases in a completely predictable manner.
- 0: No linear correlation. The variables do not have a linear relationship.
- 1: A perfect positive linear correlation. As one variable increases, the other also increases in a completely predictable manner.

Pearson's correlation coefficient is widely used to assess the linear relationship between two variables in various fields, including statistics, data analysis, social sciences, and natural sciences. It's important to note that Pearson's correlation coefficient specifically measures linear relationships; it might not capture non-linear relationships between variables.

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

Solution

Scaling is a pre processing technique used in data analysis and machine learning to transform numerical variables so that they have a similar scale or distribution. Scaling is important because many machine learning algorithms, particularly those based on distance metrics or gradient descent optimization, are sensitive to the scale of the input features. Scaling helps ensure that all variables contribute equally to the analysis and prevents variables with larger scales from dominating the results.

There are two common scaling methods: normalized scaling and standardized scaling.

1. Normalized Scaling (Min-Max Scaling):
   - Normalized scaling transforms the features to a specific range, typically between 0 and 1.

     The formula for normalized scaling is:

     X normalized = X-Xmin/Xmax-Xmin

     Where X is the original value, X_{min} is the minimum value of the feature, and X_{max} is the maximum value of the feature.
   - Normalized scaling is appropriate when you have a bounded dataset and you want to maintain the original relationships between data points. However, it can be sensitive to outliers.
2. Standardized Scaling (Z-Score Scaling):
   - Standardized scaling transforms the features to have a mean of 0 and a standard deviation of 1.
   - The formula for standardized scaling is: standardized= –

     Xstandardized=X−μ/σ
   - Where X is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature.
   - Standardized scaling is robust to outliers and is suitable when the dataset has a more Gaussian-like distribution.
   - It's commonly used when the distribution of the data is not known in advance and when you want to ensure that variables are comparable across different models.

In summary, scaling is performed to ensure that variables have a consistent scale or distribution, which can improve the performance and convergence of various machine learning algorithms. Normalized scaling transforms variables to a specific range, while standardized scaling transforms them to have a mean of 0 and a standard deviation of 1. The choice between the two depends on the nature of the data and the requirements of the analysis.

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

Solution

A VIF (Variance Inflation Factor) value becoming infinite usually indicates a perfect linear relationship between one or more predictor variables in your regression model. This situation is also known as "perfect multicollinearity."

Perfect multicollinearity occurs when one predictor variable can be perfectly predicted from a linear combination of other predictor variables in the model. This leads to numerical instability in the calculations of VIF and other regression-related statistics. Here's why this can happen:

1. Redundant Information: When two or more predictor variables are perfectly correlated, they essentially contain the same information. In such cases, one variable can be perfectly predicted from the others.

2. Matrix Inversion: In the context of calculating VIF, one step involves the inversion of a matrix that contains the correlations between the predictor variables. If the matrix is singular (not invertible) due to perfect multicollinearity, it can lead to infinite VIF values.
3. Indeterminate Coefficients: Perfect multicollinearity leads to indeterminate coefficients in the regression model. The algorithm can't distinguish which variable contributes what, leading to numerical issues.
4. Highly Correlated Variables: Even if not perfect, very high correlations between variables can lead to extremely large VIF values, indicating strong multicollinearity.

To address this issue:

- Examine your data to identify highly correlated variables and consider removing one of them from the model.
- Focus on the substantive meaning of the variables and domain knowledge. Sometimes, perfect multicollinearity might be a result of data measurement or coding issues.
- If the variables are conceptually distinct, consider transformations or aggregations that preserve their meaningful differences.

In practice, encountering infinite VIF values due to perfect multicollinearity is rare, but identifying and resolving multicollinearity issues is crucial for accurate and interpretable regression modeling.

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

Solution

A Q-Q plot, which stands for Quantile-Quantile plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution (often the normal distribution). It's particularly useful for checking if a dataset follows a specific distribution and for identifying deviations from that distribution.

Here's how a Q-Q plot works:

1. Theoretical Quantiles: The Q-Q plot first calculates the quantiles of the theoretical distribution (e.g., normal distribution) that correspond to the quantiles of the dataset being analyzed.
2. Data Quantiles: The quantiles of the actual dataset are then plotted against the corresponding quantiles from the theoretical distribution.

If the data points in the Q-Q plot closely follow a straight line, it indicates that the data follows the theoretical distribution. Deviations from the straight line suggest differences in distribution.

In the context of linear regression, Q-Q plots are valuable for:

3. Checking Normality: Linear regression models often assume that the residuals (errors) follow a normal distribution. By creating a Q-Q plot of the residuals, you can assess if this assumption holds. If the residuals closely align with the straight line in the Q-Q plot, it suggests that the normality assumption is reasonable.
4. Identifying Outliers and Skewness: Deviations from the straight line in the Q-Q plot might indicate the presence of outliers or skewness in the data. Outliers could cause the residuals to deviate from a normal distribution.
5. Assessing Assumption Validity: Linear regression relies on several assumptions, including normality of residuals. A Q-Q plot can provide visual evidence of how well the assumptions are met, helping you decide if the model's results are valid.

In summary, a Q-Q plot is a powerful graphical tool that aids in understanding the distribution of data and how it compares to a theoretical distribution. In linear regression, it's crucial for validating assumptions and identifying potential issues like non-normality, outliers, and skewness in residuals. It provides a visual way to assess the validity of the model and determine whether its assumptions are met.