Assignment-based Subjective Questions and Answers

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

Answer :

- Bike rentals are generally higher during the Fall season and lower in the Spring.
- The months of May to October see a significant increase in bike rentals.
- Bike rentals in the year 2019 surpassed those in 2018.
- Clear weather conditions lead to higher bike rental counts.
- Among weekdays, Friday records the highest bike rental counts, followed by Saturday and Sunday.
- The demand of bike is almost similar either on working day or non-working day.
- Bike demand doesn't show a much difference between working day and Holiday.

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

Answer :

Using "drop_first=True" during dummy variable creation is important to avoid multicollinearity, eliminate redundancy, enhance model interpretability, and prevent the dummy variable trap.

It ensures that one level of the categorical variable is excluded from the dummy variables, resulting in n-1 dummy variables, where n represents the number of levels/categories in the original variable. This approach improves model performance and provides reliable coefficient estimates when dealing with categorical data.

Example :

```
dummy = pd.get_dummies(dataset["column_name"], drop_first=True)
```

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

Answer :

The variable 'temp' exhibits 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?

Answer :

After building a Linear Regression model on the training set, it is essential to validate its assumptions to ensure the reliability of predictions and statistical inferences. Here are the key steps for validation:

- **Residual Analysis**: Check for patterns or trends in the differences between actual and predicted values (residuals). Randomly distributed residuals around zero are preferred.
- **Normality of Residuals**: Ensure that residuals follow a normal distribution using statistical tests and visual inspection.
- **Homoscedasticity**: The variance of residuals should be consistent across all levels of predictor variables, as seen in scatter plots.
- **Independence of Residuals**: Residuals should be independent and not show any autocorrelation or patterns.
- **Linearity**: Verify that the relationship between predictors and the response variable is linear, observed through plots.
- **Multicollinearity**: Check for high correlation between predictor variables, which can impact the model's stability and interpretation.
- **Outliers and Influential Points**: Identify and examine outliers and influential points that can distort the model and its results.
- **Error Terms**: Ensure that the error terms are normally distributed with mean zero and constant variance.

Use statistical tests and visualisations for validation. Addressing and validating these assumptions will lead to a more robust and accurate Linear Regression model.

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

Answer :

Top 3 features that influence the demand of the shared bikes are :

“Temperature”, with a positive coefficient of 0.4712.

"Year" with a positive coefficient value of 0.2330.

"September" (transition from summer to winter) month shows a positive impact on Bike rentals.

General Subjective Questions and Answers

1. Explain the linear regression algorithm in detail.

Answer :

Linear regression is a supervised learning algorithm used for predicting the relationship between a dependent variable (target) and one or more independent variables (features) in a continuous numeric context.

The linear regression algorithm can be expressed mathematically as follows:

y = mx + b

where :

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

There are two types of linear regression.

1.Simple Linear Regression (SLR) : When dependent variable is predicted using only one independent variable.

2. Multiple Linear Regression (MLR) : When dependent variable is predicted using multiple independent variable.

MLR equation :

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

where :

- y is dependent variable (target).
- x1, x2, ..., xn are multiple independent variables (features).
- b0 is Intercept.
- b1, b2, ..., bn are the coefficients corresponding to the independent variables.

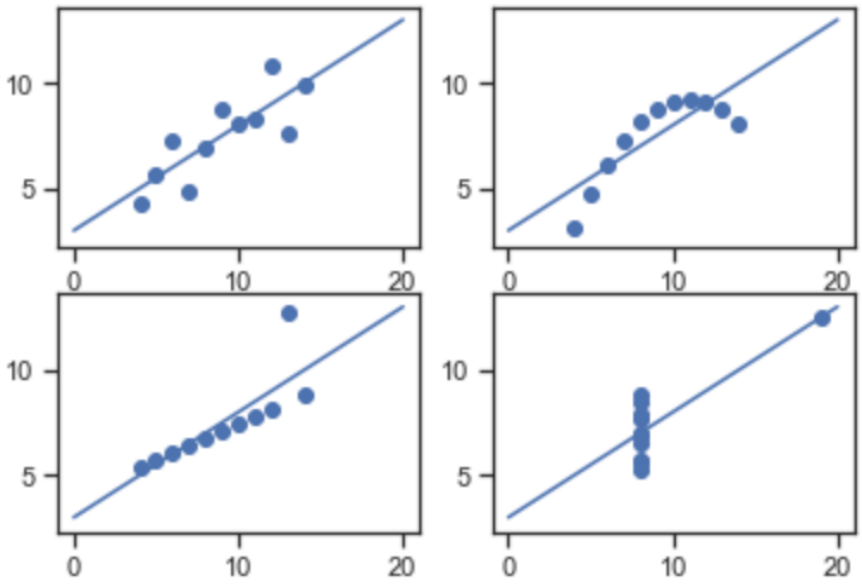
2. Explain the Anscombe’s quartet in detail.

Answer :

Anscombe's quartet is a set of four data sets that have nearly identical summary statistics when analysed using various basic statistical measures, such as mean, variance, correlation, and linear regression.

They have very different distributions and appear differently when plotted on graph.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.



Graphical Representation of Anscombe’s Quartet

The first scatter plot appears to be a simple linear relationship.

The second graph is not distributed normally, while there is a relation between them, but doesn’t show a linear relationship.

Third graph shows the distribution is linear but should have a different regression line where the calculated regression is offset by one outlier which exerts enough influence to lower the correlation coefficient.

The fourth plot shows a data-point which 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?

Answer :

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

It is denoted by the symbol “r”.

The value of Pearson's R ranges from -1 to +1 :

Pearson's R	Interpretation
-1	Perfect negative correlation
-0.7 to -0.9	Strong negative correlation
-0.3 to -0.6	Moderate negative correlation
0	No correlation
0.3 to 0.6	Moderate positive correlation
0.7 to 0.9	Strong positive correlation
1	Perfect positive correlation

Pearson's R is a very common statistical measure, and it is used in a wide variety of fields like :

- Psychology to study the relationship between different personality traits.
- Business to study the relationship between sales and marketing spending.
- Education to study the relationship between student test scores and hours spent studying.

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

Answer :

Scaling is a pre-processing step to normalize or standardise the range of independent variables or features of data, or transforming the features (variables) of a dataset to a similar range. When features have varying numerical ranges, certain machine learning algorithms may be biased towards features with larger values, leading to unfair results and slow convergence during training. Scaling prevents such biases, ensuring all features contribute equally to the learning process and enabling effective and fair analysis by these algorithms.

Normalization and standardization are two techniques used to scale data in machine learning. They both involve transforming the data so that it has a mean of 0 and a standard deviation of 1.

However, there are some key differences between the two techniques.

Normalized Scaling (Min-Max Scaling):

- Also known as Min-Max scaling.
- Transforms data to a specific range (usually between 0 and 1).
- Formula: x_normalized = (x - min(x)) / (max(x) - min(x)).
- Sensitive to outliers due to the influence of extreme values on the range.
- Suitable when data distribution is bounded within a specific range and preserving relationships between data points is important.

Standardized Scaling (Z-Score Normalization):

- Also known as z-score scaling.
- Transforms data to have zero mean and unit variance.
- Formula: x_standardized = (x - mean(x)) / std(x).
- Less sensitive to outliers as it doesn't rely on minimum and maximum values.
- Useful when data has varied scales, follows a Gaussian distribution, or when minimizing the impact of outliers is necessary.
- Commonly used for distance-based algorithms like k-nearest neighbours and clustering.

In summary, when deciding between normalized scaling and standardized scaling, it's essential to consider the data distribution, its characteristics, and the requirements of the machine learning model. Understanding these factors will aid in making the right choice for data pre-processing.

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

Answer :

Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression analysis. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated with each other.

VIF becomes “infinite” when perfect multicollinearity exists, where one predictor can be perfectly predicted from others.

The formula for VIF is :

$VIF_i = 1 / (1 - R^2_i)$

Where:

- VIF_i is the Variance Inflation Factor for the i-th predictor variable.
- R^2_i is the coefficient of determination (R-squared) obtained from regressing the i-th predictor variable on all other predictor variables.

If an independent variable can be completely described by other independent variables, that means it has perfect correlation and has an R-squared value of 1. So, the result will be VIF = 1/(1-1) provides VIF=1/0, calculates to “infinity”.

VIF values close to or equal to infinity signify perfect multicollinearity, which makes it impossible to estimate unique coefficients for the involved predictor variables.

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

Answer :

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a sample dataset with a theoretical distribution such as a normal, exponential, or uniform distribution.

In linear regression, it is used to check the assumption of normality of the residuals (differences between observed and predicted values). If the residuals are normally distributed, then the Q-Q plot will be a straight line. However, if the residuals are not normally distributed, then the Q-Q plot will deviate from a straight line.

The importance of a Q-Q plot in linear regression are :

- Assess Plausibility of Distribution: A Q-Q plot helps assess if a dataset plausibly follows a theoretical distribution like normal, exponential, or uniform.
- Compare Distributions: It can determine if two datasets come from populations with a common distribution.
- Assess Normality of Residuals: In linear regression, it is used to check if the residuals are normally distributed.
- Identify Model Problems: Deviations from a straight line in the Q-Q plot indicate non-normality, highlighting potential problems with the model.
- Accuracy of Inferences: Normality of residuals is crucial for accurate standard errors and coefficient significance in linear regression.
- Gain Distribution Insights: Q-Q plots provide insights into skewness, kurtosis, and presence of outliers in the data.

In conclusion, a Q-Q plot is a powerful tool to validate assumptions, detect issues, and gain distributional insights in linear regression.