Bike Sharing Rental Assignment Questions

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: Many of the categorical variables were collinear and has minimal impact. Hence they were removed to derive a subset of relevant fields.

   Of these fields, It was observed that the holiday, season and weather had a significant impact on the bike rental.

   Holiday, windspeed and weather had a negative coefficient. It indicated that the rentals were negatively impacted due to holidays or in a windy weather or if it not was a clear and sunny weather.

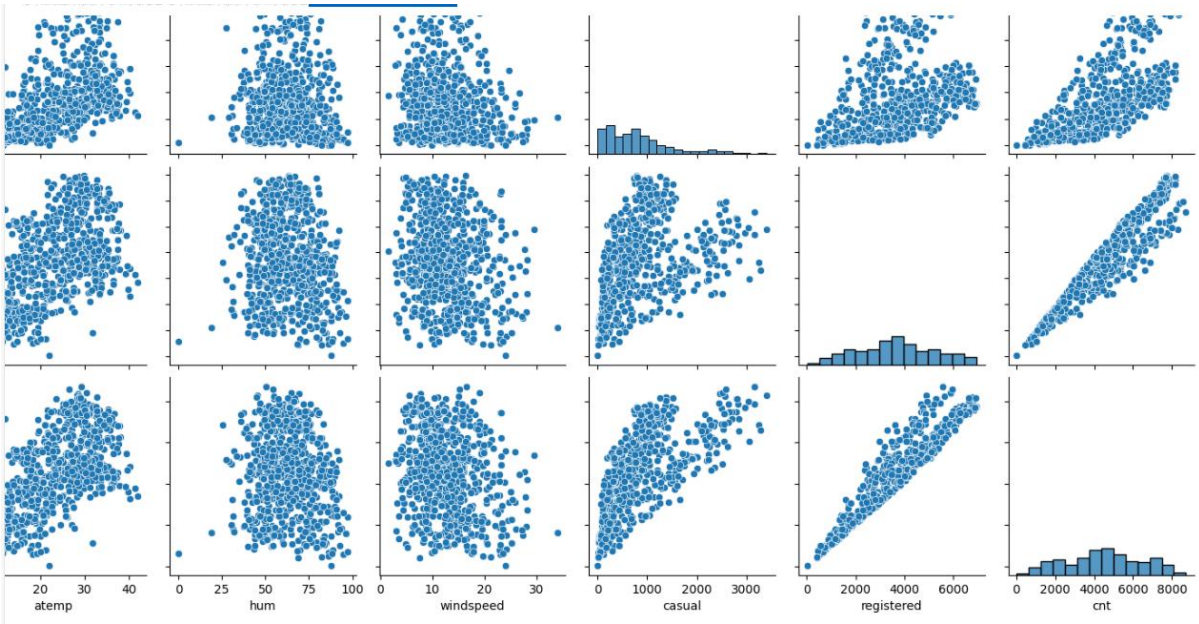
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   Answer:

   Drop_first=true helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   Answer: If you look at the picture below, registered renters has the highest correlation with the cnt variable.



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

   Answer:

   I checked the following:

   a. *The R2 score was verified. It stabilized around 80% which showed that the model was good.*
   b. *The pvalues of all the independent variables was checked. There were near to 0. It indicated that they were significant*
   c. *The Fstat value was high.*
   d. *The VIF of all the variables in the final set was verified. All were below 5 which indicated low collinearity.*

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: The top 3 features that contributed significantly are :

   a. *Holiday played a significant role in rentals.*
   b. *Weather had a direct impact. The clear the skies, the high were the rentals*
   c. *Temp had a positive impact.*

General Subjective Questions

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

Answer:

Overview

Linear regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. The goal is to find the linear

Assumptions of Linear Regression

Linearity: The relationship between the independent and dependent variables is linear.

Independence: Observations are independent of each other.

Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.

Normality: The residuals of the model are normally distributed.

Steps in the Linear Regression Algorithm

1. *Data Collection: Gather the data that includes the dependent variable (target) and the independent variables (features).*
2. *Data Pre-processing:*
    - *Handling Missing Values: Fill or remove missing data points.*
    - *Feature Scaling: Normalize or standardize the data if necessary.*
    - *Splitting Data: Divide the data into training and testing sets.*
3. *Model Initialization: Initialize the parameters (weights) of the linear model. For simple linear regression, this includes the slope ((b1)) and intercept ((b0)).*
4. *Hypothesis Function: Define the hypothesis function, which is the linear equation:*

   h(x)=b0+b1x

   For multiple linear regression, it extends to:

   h(x)=b0+b1x1+b2x2+…+bnxn

5. *Cost Function: Define the cost function to measure the error between the predicted values and the actual values. The most common cost function for linear regression is the Mean Squared Error (MSE):*

   J(b0,b1)=2m1i=1∑m(h(xi)−yi)2

   where (m) is the number of training examples, (h(x_i)) is the predicted value, and (y_i) is the actual value.

6. *Gradient Descent: Use gradient descent to minimize the cost function by iteratively updating the parameters.*

7. *Model Training: Iterate over the training data, updating the parameters using the gradient descent algorithm until the cost function converges to a minimum.*
8. *Model Evaluation: Evaluate the model's performance using metrics such as R-squared, Mean Absolute Error (MAE), or Root Mean Squared Error (RMSE) on the testing set.*
9. *Prediction: Use the trained model to make predictions on new data*

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

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".

Importance of Anscombe's Quartet:

Anscombe's quartet demonstrates several important lessons in data analysis:

Graphical Analysis: Always visualize your data before performing statistical analysis. Graphs can reveal patterns, relationships, and anomalies that summary statistics might miss.

Outliers and Influential Points: Outliers and high-leverage points can significantly impact statistical measures and regression models. Identifying and understanding these points is crucial.

Misleading Statistics: Identical statistical properties do not guarantee similar data distributions. Relying solely on summary statistics can lead to incorrect conclusions.

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

Answer:

The Pearson correlation coefficient, often denoted as Pearson's r, is a measure of the linear relationship between two quantitative variables. It quantifies the strength and direction of this relationship, producing a value between -1 and 1.

Key Points:

Value Range:

1: Perfect positive linear relationship.

0: No linear relationship.

-1: Perfect negative linear relationship.

Interpretation:

Positive values indicate that as one variable increases, the other also increases.

Negative values indicate that as one variable increases, the other decreases.

Values close to 0 suggest a weak linear relationship.

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

Answer:

Scaling in machine learning refers to the process of transforming the features of your data so that they are on a similar scale. This is crucial because many machine learning algorithms perform better or converge faster when the features are on a relatively similar scale and close to normally distributed.

In machine learning, feature scaling is employed for a number of purposes:

- *Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.*
- *Algorithm performance improvement: When the features are scaled, several machine learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly.*
- *Preventing numerical instability: Numerical instability can be prevented by avoiding significant scale disparities between features.*

Difference between Normalization and Standardization:

Normalization, also known as Min-Max scaling, transforms the data to fit within a specific range, typically [0, 1] or [-1, 1]. Scales the data to a fixed range. Highly sensitive to outliers, as they can significantly affect the min and max values.

Standardization, also known as Z-score normalization, transforms the data to have a mean of 0 and a standard deviation of 1. Does not bound the data to a specific range. Less sensitive to outliers compared to normalization, as it does not rely on min and max values.

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

(3 marks)

Answer:
The Variance Inflation Factor (VIF) can sometimes be infinite, and this typically indicates a severe issue with multicollinearity in your regression model.

Formula

The VIF for a predictor (X_i) is calculated as:

$$VIF(X_i) = 1 - R_i^2$$

where (R_i^2) is the coefficient of determination of the regression of (X_i) on all other predictors.

Why VIF Can Be Infinite:

1. ***Perfect Multicollinearity**: If (R_i^2) is 1, it means that the predictor (X_i) is perfectly linearly dependent on the other predictors. In this case, the denominator of the VIF formula becomes zero, leading to an infinite VIF value.*
2. ***Near-Perfect Multicollinearity**: Even if (R_i^2) is very close to 1, the VIF value can become extremely large, indicating severe multicollinearity.*

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

(3 marks)

Answer:
Q-Q plots are a valuable diagnostic tool in linear regression for assessing the normality of residuals. By visually inspecting the Q-Q plot, analysts can ensure that the assumptions of linear regression are met, leading to more reliable and valid results.
In most cases, this type of plot is used to determine whether or not a set of data follows a normal distribution.
*As a rule of thumb, **the more that the points in a Q-Q plot lie on a straight diagonal line, the more** normally distributed the data.*

Importance of Q-Q Plots in Linear Regression

1. ***Assumption Verification**: Ensuring that residuals are normally distributed is crucial for the validity of hypothesis tests and confidence intervals in linear regression.*
2. ***Model Diagnostics**: Identifying deviations from normality can indicate potential issues with the model, such as the presence of outliers, skewness, or other anomalies.*
3. ***Improving Model Fit**: If the residuals are not normally distributed, transformations of the data or alternative modeling approaches may be necessary to improve the model fit.*