Bike Sharing Assignment

By Shweta Alukuru Trikutam

Assignment-based Subjective Questions

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

Answer: From the analysis of the categorical variables in the dataset, several insights emerge regarding their impact on the dependent variable:

1. Seasonal Demand Variation: Fall season exhibits the highest demand for rental bikes, indicating a seasonal trend where certain seasons may have a higher demand compared to others.
2. Yearly Demand Trend: There is a noticeable increase in demand for the next year, suggesting a positive trend in overall demand growth over time.
3. Monthly Demand Fluctuation: Demand shows a consistent upward trend month on month until June, with September emerging as the month with the highest demand. However, demand tends to decrease post-September.
4. Impact of Holidays: The presence of a holiday correlates with a decrease in demand, implying that holidays may influence the rental bike demand negatively.
5. Weekday Demand Variation: Weekday does not provide a clear indication of demand variation, suggesting that other factors may have a more significant influence on rental bike demand.
6. Weather Conditions Influence: Clear weather situations (weathersit) are associated with the highest demand, indicating that weather conditions play a crucial role in determining demand patterns.
7. Seasonal Peaks and Troughs: Bike sharing is more prevalent during September, while it tends to decrease during the end and beginning of the year, indicating seasonal peaks and troughs in demand.

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

Answer:
drop_first=True is important to use, as it helps in reducing the extra columns that gets created

during dummy variable creation. Hence it reduces the correlations created among dummy variables.
Dropping the first columns as (p-1) dummies can explain p categories . In weathersit, the first column was not dropped so as not to lose the info about severe weather conditions.

Q 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer:
Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation with the target variable (cnt).

Q 4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
Answer:
Residual Analysis:

1. Errors are normally distributed with a mean of 0.
2. Actual and predicted results follow the same pattern.
3. The error terms are independent of each other.

R2 value for test predictions:
R2 value for predictions on test data (0.828) is almost the same as R2 value of train data (0.822). This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data).

Homoscedasticity:
We can observe that variance of the residuals (error terms) is constant across predictions. i.e., error term does not vary much as the value of the predictor variable changes.

Plot Test vs Predicted value test:
The prediction for test data is very close to actuals.

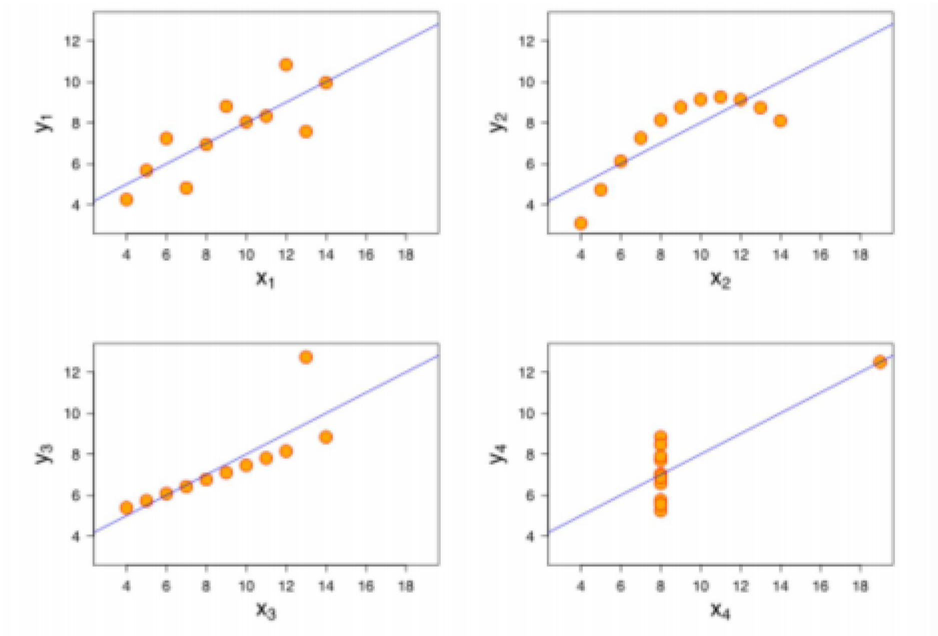
Q 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
The top 3 features are:
1. year (positive corelation)
2. temp (positive corelation)
3. windspeed (negative corelation)

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail. (4 marks)
Answer:
Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.
The linear regression model can be represented by the following equation:
y= a0+a1x+ ε
The linear regression model provides a sloped straight line representing the relationship between the variables.
y= Dependent Variable (Target Variable) x= Independent Variable (predictor Variable) a0= intercept of the line (Gives an additional degree of freedom) a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.
The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.
In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

Q 2. Explain the Anscombe's quartet in detail. (3 marks)
Answer:
Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

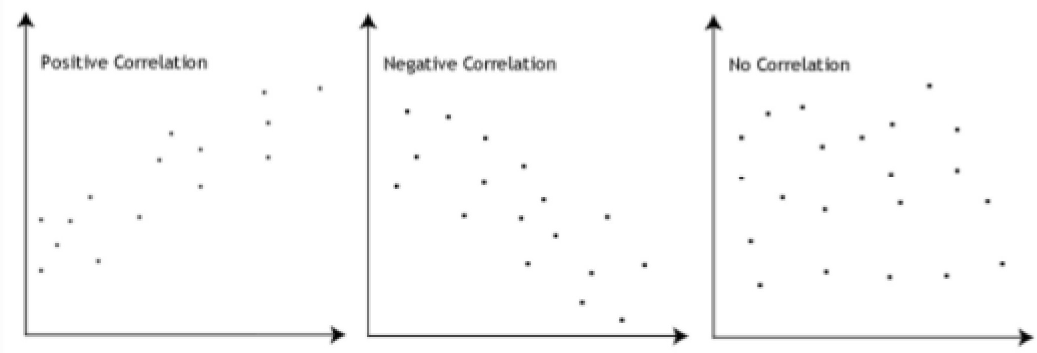


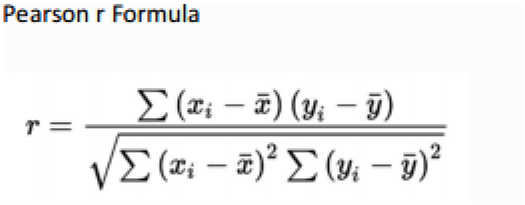
Q 3. What is Pearson’s R? (3 marks)
Answer:

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association





Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- r =correlation coefficient
- xi =values of the x-variable in a sample
- x̄ =mean of the values of the x-variable
- yi =values of the y-variable in a sample
- ȳ =mean of the values of the y-variable

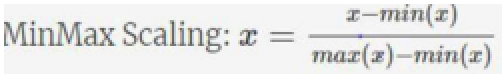
Q 4. What is scaling? Why is scaling performed? What is the difference between normalized Scaling and standardized scaling? (3 marks)

Answer:

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. When we collect a data set it contains features highly varying in magnitudes, units, and range. If scaling is not done, then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

1. Normalization/Min-Max Scaling:

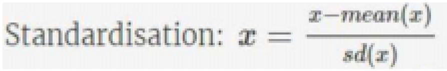
It brings all of the data in the range of 0 and 1.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

2. Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with the same distributions.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.