BIKE SHARING CASE STUDY

minhngc4795@gmail.com

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)

- **Seasons:** Summer and fall rents are higher, especially from June to September. This might assist guide decisions on how to distribute resources (such as bikes and employees) during busy months.
- **Weather conditions:** Rentals are greater on sunny, good-condition days, which may influence when maintenance or repairs are scheduled.

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

- Using drop_first=True is important because it helps to **avoid multicollinearity in the dataset.**
- We produce n binary (0/1) variables, one for each level, when we create dummy variables for a category variable with n levels. Nevertheless, having n binary variables might cause issues in a regression model because they are all perfectly correlated. This means that one variable may be predicted perfectly from the others, resulting in overfitting and model instability.
- By setting drop first=True, we eliminate one of the binary variables, generally the one corresponding to the category variable's first level. Because the remaining n-1 binary variables are no longer fully correlated with each other, the multicollinearity problem is resolved.

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

- If the **registered column** is not removed before plotting the pairplot, it will have the strongest correlation with the cnt column.
- Otherwise, the **temp / atemp columns** will be the most highly connected with the cnt.

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

I used the following ways to validate the assumption:

- **Summary of the Linear Model:** Verify the model using the Linear Regression model summary (P-value, F-stat, etc.)
- I calculated the **variance inflation factor (VIF)** for each independent variable to test for multicollinearity. A variable with a high VIF is significantly correlated with other variables. We will ensure that all features have an adequate VIF.
- **Residual normality:** It demonstrates that a key assumption of linear regression is the normal distribution of residuals.
- **Plot showing residuals** (differences between real and expected values) vs predicted values. A good model should contain residuals that are spread randomly about zero, with no discernible patterns or trends.

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

In my model, the top 3 features that have the highest correlation with the target columns:

- Temp: 0.4756 coef
- Yr: 0.2339 coef
- Weathersit_3: -0.2257 coef

General Subjective Questions

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

Linear regression is a supervised learning algorithm used to predict a continuous target variable based on one or more independent variables. The goal of linear regression is to find the best-fit line through the data, which represents the relationship between the independent and dependent variables.

- The general form of a linear regression equation is:
  - $y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta n x n + \varepsilon$
- Where y is the dependent variable we want to predict, x1, x2, ..., xn are the independent variables, β0 is the intercept term, β1, β2, ..., βn are the coefficients (also known as weights or parameters), and ε is the error term (or noise).
- The linear regression algorithm works by finding the values of β0, β1, β2, ..., βn that minimizes the sum of the squared errors between the predicted and actual values of y, across all the data points.
  - The most common method used to find the optimal values of β is called Ordinary Least Squares (OLS) regression.
- Once we have the values of β, we can use the linear regression equation to predict the value of y for new data points, by substituting the values of the independent variables into the equation.

In summary, linear regression is a simple and efficient approach for predicting a continuous target variable based on one or more independent variables. It finds the best-fit line across the data that minimizes the sum of squared errors between the predicted and actual values of the target variable.

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

The Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but very different appearances when graphed. It serves as a reminder of the importance of graphing data to fully understand its properties, and not relying solely on summary statistics.

- Dataset I: This dataset shows a positive slope linear connection between x and y. There is significant distribution around the line, with a few outliers.
- Dataset II: This dataset features a non-linear connection between x and y with a distinct curvature pattern. There is significant distribution around the curve, with a few outliers.
- Dataset III: This dataset exhibits a linear connection between x and y, but it has one outlier that is far away from the other data points. The outlier has a big impact on the regression line, and the correlation would be considerably weaker without it.
- Dataset IV: Except for one outlier that is faraway from the other data points, this dataset shows a perfect linear connection between x and y. The outlier has a significant impact on the regression line, and the connection would be perfect if it did not exist.

Anscombe's quartet is intended to demonstrate the need of visualizing data before drawing statistical assumptions. While statistical measurements might give important information about a dataset, they are not always enough for capturing the data's complexity.
Visualizing the data may frequently reveal patterns and correlations that are not obvious from the summary statistics alone, and it can aid in the identification of outliers or other odd aspects.

3. What is Pearson’s R? (3 marks)

- Pearson's R is a measure of the linear connection between two variables. It is also known as Pearson's correlation coefficient. It has a value between -1 and 1, with -1 indicating a fully negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfectly positive linear relationship.
- The Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations:
    - r = cov(X,Y) / (std(X) * std(Y))

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

- The process of transforming variables to a common scale in order to compare them is referred to as scaling. Scaling is frequently performed on features or input variables in the context of machine learning and data analysis to guarantee that they are on a comparable scale and to boost the performance of particular algorithms.
- There are several reasons why scaling is performed:
  - To improve the performance of machine learning algorithms: Many machine learning algorithms use distances as a measure of similarity. If the features have different scales, then one feature might dominate the distance calculation, making the other features relatively insignificant.
  - To reduce the impact of outliers: Outliers can have a large impact on the mean and variance of a feature, and can skew the analysis. Scaling can help to reduce the impact of outliers and make the data more robust to extreme values.
  - To facilitate the interpretation of coefficients: When performing regression analysis, the interpretation of the coefficients becomes easier when the features are on a similar scale.
- The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the shape of the distribution and compresses it to a smaller range, while standardized scaling preserves the shape of the distribution and transforms it to have a mean of 0 and a standard deviation of 1.

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

- In some cases, the VIF value may be infinite. This happens when one or more predictor variables in the model are perfectly collinear with the other predictor variables.
- Perfect collinearity means that one variable is a linear combination of the other variables, so there is no unique solution for the regression coefficients.

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

- A Q-Q plot, also known as a quantile-quantile plot, is a graphical technique for determining the normality of a distribution by comparing it to a theoretical normal distribution (t-distribution). **Q-Q plots are often used in linear regression to test the assumption of residual normality.**
- The following are the applications and significance of Q-Q graphs in linear regression:
  - Determine if the residuals are normally distributed
  - Detect probable outliers or heavy-tailed distributions
  - Improve the model accuray