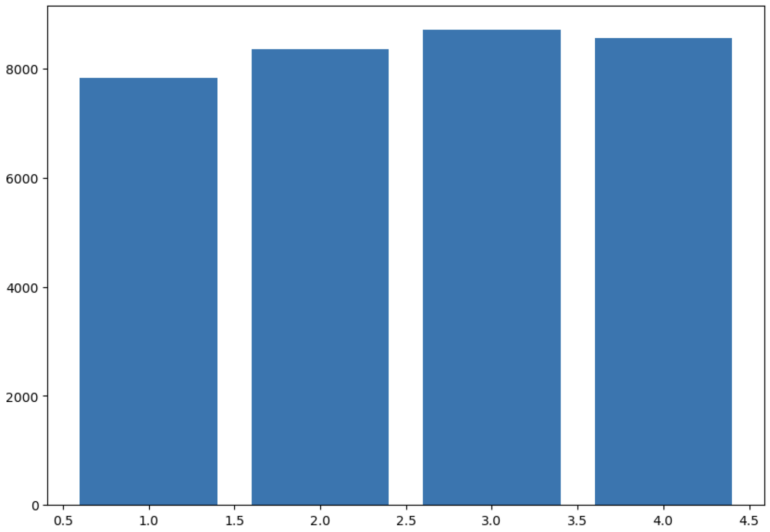
Student: Abhi Prasad

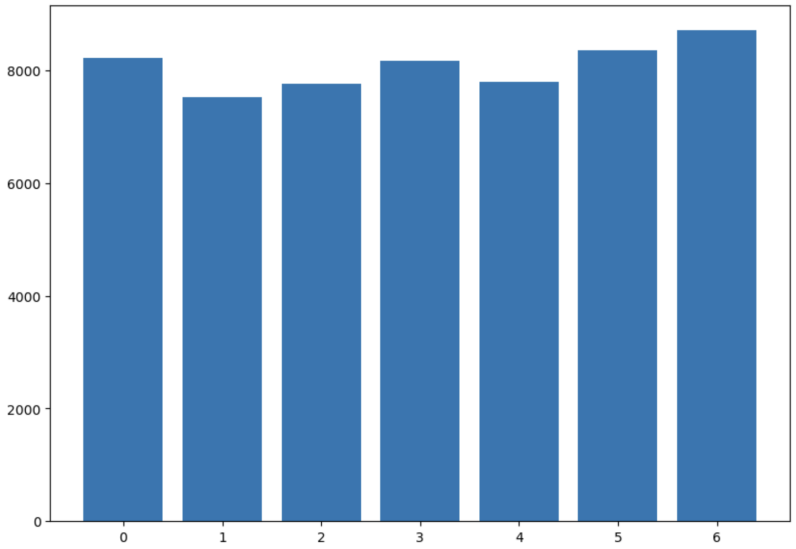
Assignment-based Subjective Questions

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

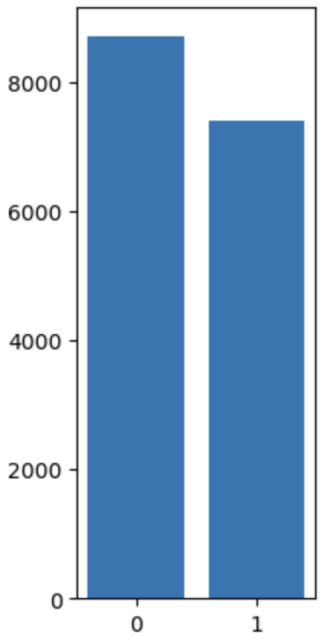
a. Season has no significant impact on the total count. Fall has a better relation but the difference is not as significant



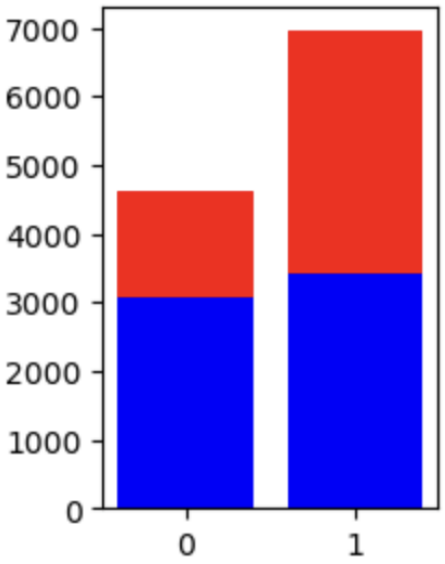
b. Weekday or not doesn't have significant impact. Saturday (6) and Sunday(7) are getting more traction compared to Monday (1)



c. Non-holidays have more bike rentals than holidays. Implying that weekday users are using the bike rental app more, probably for last mile commute.



d. The number of registered users as a percentage of casual users increased from 2018 to 2019.



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

When we create dummy variables, we can consider it as binary value calculation. If there are k variables, then k-1 dummy variables can be created. Reason as follows:

If there are three dummy variables created, Morning, Noon, Night then the breakdown is as follows

Morning	Noon	Night	Interpreted value
1	0	0	Morning
0	1	0	Noon
0	**0**	1	Night

So whether we say 001 or 00, the value is implied to be night. At larger level, additional variables restricts the model from being simpler and leaner.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Atemp and temp have a 0.99 correlation making them redundant. The correlation with other variables also is similar for both variables.
Similarly Registered and Cnt have high correlation of 0.95. Though their correlation with Holiday and Casual is high, making them not entirely similar.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Validated if the day of the week has any statistical significance and it didn't. Weekday or weekend did not add much difference to the overall R-squared and adjusted R-squared values.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Temperature, Clear weather and Summer are helping the bike rentals when looked at overall count.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
Linear regression algorithm takes existing data, tries to derive patterns from it to help us predict future values based on derived coefficients.

A typical linear regression algorithm is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It assumes a linear relationship between the inputs and the target, represented by a linear equation:
y = mx + b

- "y" - target variable
- "x" - input feature,
- "m" - slope
- "b" - intercept.

The algorithm learns these coefficients from the training data to minimize the sum of squared differences between predicted and actual target values. Linear regression is widely used for

tasks like predicting house prices, stock prices, and other numerical outcomes, providing interpretable results.

As values increase, the equation changes to

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta n*Xn$

2. Explain the Anscombe's quartet in detail. (3 marks)
Anscombe's quartet is a set of four datasets with almost identical summary statistics but significantly different data patterns. It shows the importance of data visualization beyond just summary statistics, showing that datasets can have similar statistical properties yet require entirely different analytical approaches due to their unique structures.

a. Linear relationship. In this case usually the data and visual representation are aligned
b. Non-Linear relationship. In this case, the difference can be significant. For example, though mean, median, etc are same for two datasets a and b and a being linear and b being non-linear, only the visual representation shows this difference.
c. Outlier. This type of data helps to be seen visually where how far off the outlier is becomes very evident.
d. Clustered data. This type of data is similar to non-linear in its visual unpredictability.

3. What is Pearson's R? (3 marks)
Pearson's correlation coefficient (Perarson's R) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r = 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases proportionally.
- r = -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases proportionally.
- r = 0 indicates no linear relationship between the variables.

Pearson's R is widely used in data analysis, research, and machine learning to assess the degree of association between two variables. It's particularly valuable when you want to determine whether changes in one variable can be used to predict changes in another. However, it's important to note that Pearson's R assumes a linear relationship and is sensitive to outliers.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Scaling is a data preprocessing technique used to standardize or normalize the range of features (independent variables) in a dataset. It's performed to bring all features to a similar scale or range, which can be crucial for many machine learning algorithms. Here's why scaling is done and the difference between normalized scaling and standardized scaling:

Scaling is Performed for:

- Equal Weight: Scaling ensures that all features contribute equally to model training, preventing features with larger scales from dominating the learning process.
- Gradient Descent: Many optimization algorithms, like gradient descent, converge faster when features are on a similar scale, avoiding convergence issues.
- Regularization: Regularization techniques (e.g., L1 and L2 regularization) assume that all features are on a similar scale, making scaling essential.
- Distance Metrics: Distance-based algorithms, such as k-means clustering or k-nearest neighbors, are sensitive to feature scales. Scaling makes these algorithms work more effectively.

Normalized Scaling (Min-Max Scaling): It transforms features to a specific range (e.g., [0, 1]). This scaling is suitable when you want to preserve the original data distribution.

X_normalized = (X - X_min) / (X_max - X_min)

Standardized Scaling (Z-score Scaling): It standardizes features to have a mean of 0 and a standard deviation of 1. Standardization centers the data around zero and is useful when the data should be normally distributed or when comparing features with different units.

X_standardized = (X - X_mean) / X_std

The choice between normalization and standardization depends on the specific requirements of your data and the machine learning algorithm you're using.

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

It happens because VIF formula is 1/(1-R2) So if R2 (R squared of any value is equal to 1, the value becomes 1/0 = infinity.
This happens due to perfect collinearity between the independent variables. In such cases, the VIF formula, which involves dividing the variance of the estimated coefficient of a variable by its corresponding variance without considering other variables, breaks down. When two or more variables are perfectly correlated, it becomes impossible to separate their individual effects, leading to infinite VIF values.

Perfect multicollinearity can arise due to data errors, duplicated variables, or inappropriate feature engineering. To address this issue, it's crucial to identify and remove one or more of the highly correlated variables, or if necessary, reconsider the data preprocessing steps to avoid perfect multicollinearity.

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 (Quantile-Quantile) plot is a graphical tool used in statistics to check if a dataset follows a normal distribution. In a Q-Q plot, the quantiles of the observed data are plotted

against the quantiles of the expected theoretical distribution (e.g., a normal distribution). If the points on the plot roughly form a straight line, it suggests that the data follows the theoretical distribution.

In linear regression, Q-Q plots are important for several reasons:

1. Assumption Checking: Q-Q plots help check the assumption of normality of residuals, which is crucial for linear regression. Deviations from a straight line in the Q-Q plot may indicate that the residuals are not normally distributed, potentially impacting the validity of regression results.
2. Detecting Outliers: Outliers in the data can affect regression models. Q-Q plots can reveal the presence of outliers as deviations from the expected distribution in the plot.
3. Model Evaluation: Q-Q plots are valuable in evaluating the quality of a linear regression model. A well-behaved Q-Q plot suggests that the model's assumptions are met, increasing confidence in the model's reliability.