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?

Ans: Season, Weather Situation, holiday, month, working day and weekday were the categorical variables in the dataset. A boxplot was used to visualise these. These variables influenced our dependent variable in the following ways:

Season: The boxplot revealed that the spring season had the lowest value of cnt, while the fall season had the highest value of cnt. Summer and winter had cnt values that were in the middle.

Weather Situation: When there is heavy rain/snow, there are no users, indicating that the weather is extremely unfavourable. The highest count was observed when the weather forecast was 'Clear, Partly Cloudy.'

Holiday: Rentals were found to be lower during the holidays.

Month: September had the most rentals, while December had the fewest. This observation is comparable to the one made in weathersit. The weather in December is typically cold and snowy.

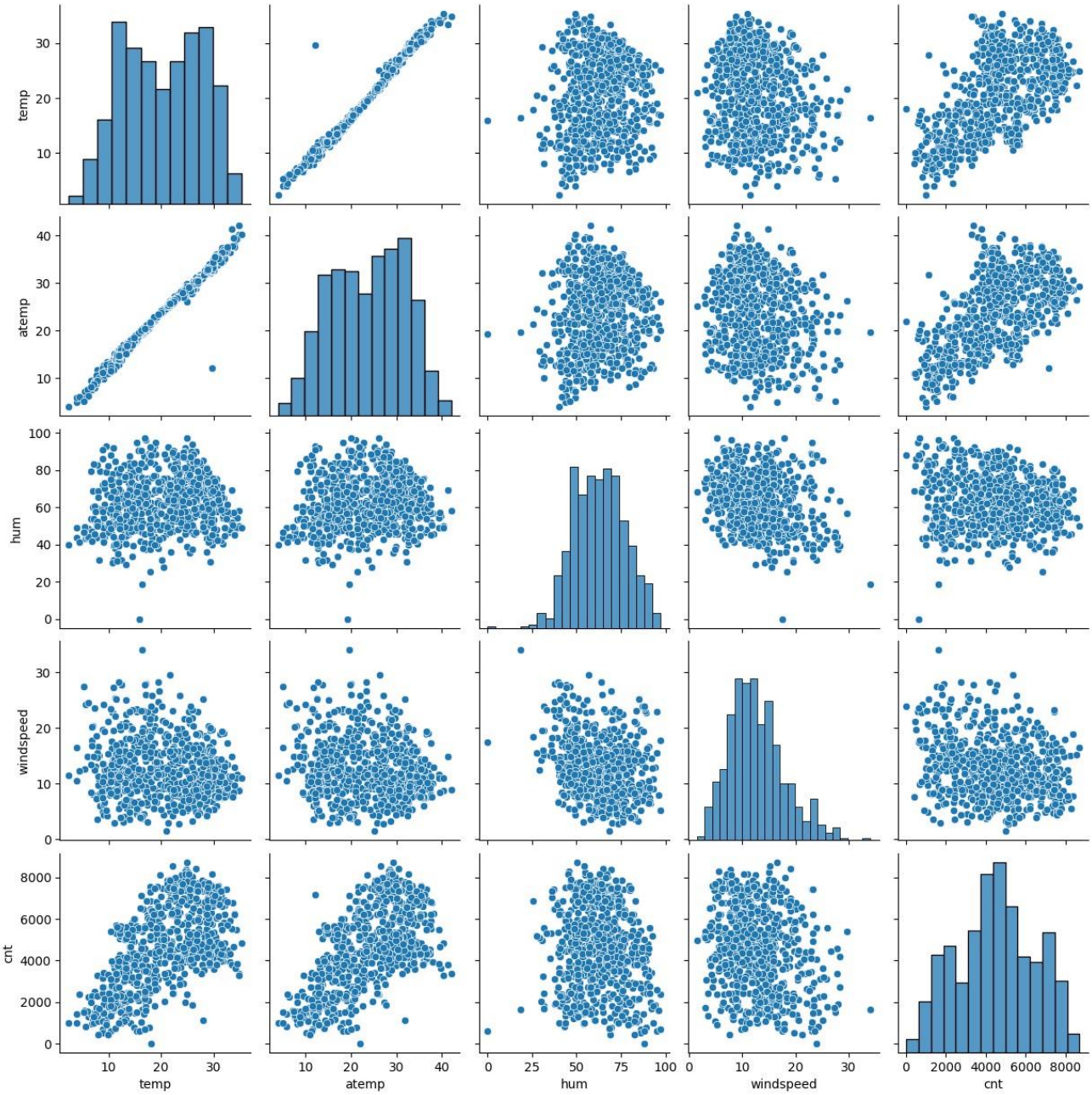
Weekday: Weekends saw a significant increase in book hiring compared to weekdays.
Working day: It had little effect on the dependent variable.

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

Ans: drop_first=True is used during the dummy variable creation as it helps in reducing the extra column created. 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?

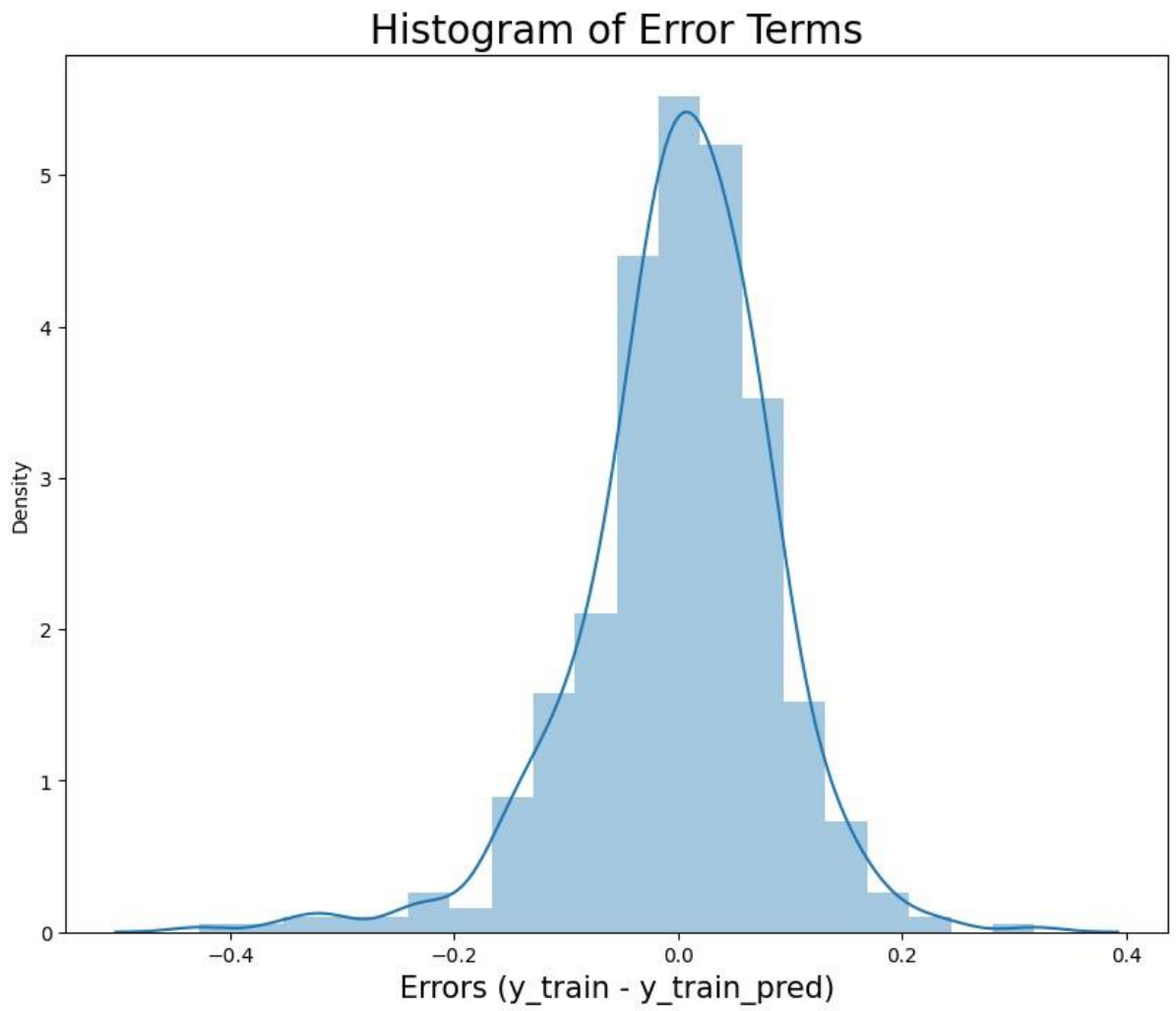
Ans: “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).



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

Ans: The distribution of residuals should be normal and centred around 0. (The mean is 0).

We test this residuals assumption by producing a distplot of residuals to see if they follow a normal distribution or not. The residuals are scattered around mean=0 as seen in the diagram above.



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

Ans: The top 3 predictor variables the influence the bike booking according to our final model are:

Temperature(temp): With a coefficient of 0.4694, a unit increase in the temp variable increases the number of bike rentals by 0.4694 units.

Weather Situation light(weathersit_light): With a coefficient of -0.2539, a unit increase in the weathersit_light variable reduces the number of bike hires by 0.2539 units as compared to weathersit_mist.

Year(yr): With a coefficient of 0.2320, a unit increase in the yr variable increases the number of bike rentals by 0.2320.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets. There are 2 types of Linear Regression:

Simple Linear Regression -
This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression -
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.
In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

3. What is Pearson's R?

Ans: The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

Below is a formula for calculating the Pearson correlation coefficient (r):

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Ans: Feature Scaling also known as data normalization is a data preprocessing step. It is a method used to normalize the range of independent variables or features of data. Independent variables are of different types. The numerical ones can be of type age (0-100 years), salary (in the range of 1000s), dimensions (decimal points), and many more. We don't want our machine learning model to confuse a feature with a larger magnitude as a better one. Feature scaling in Machine Learning would help all the independent variables to

be in the same range, for example- centered around a particular number (0) or in the range (0,1), depending on the scaling technique.

The machine learning models assign weights to the independent variables according to their data points and conclusions for output. In that case, if the difference between the data points is high, the model will need to provide more significant weight to the farther points, and in the final results, the model with a large weight value assigned to undeserving features is often unstable. This means the model can produce poor results or can perform poorly during learning.

Difference between Normalization & Standardization scaling is as follows -

Normalization	Standardization
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

Ans: If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Ans: The quantile-quantile (q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.