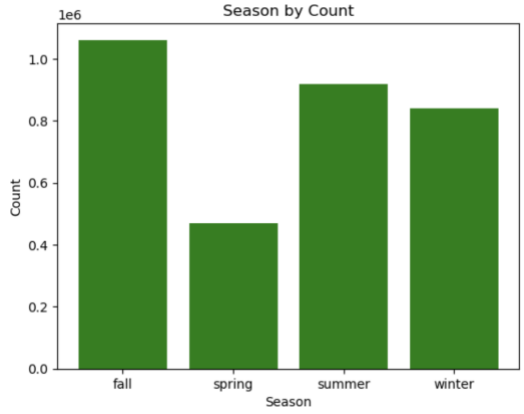
Assignment-based Subjective Questions

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

→ Season, Weathersit, Holiday, Year, Month, Weekday, Workingday are the categorical variables from dataset.

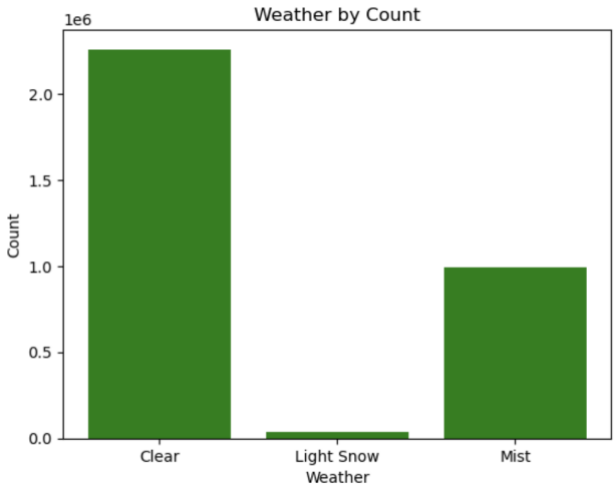
1. **Season:** Season varies spring, summer, fall to winter.
   This indicates the dataset covers all four seasons, which can impact the dependent variable due to seasonal effects.

   Fall season is the highest count of bike sharing
   Summer and Winter constitute **52%** of Bike sharing and Fall alone constitute **33%** of the Bike sharing counts.



2. **Weather:** Clear, Mist, Light Snow, Heavy Rain types of weather where Heavy rain don't have any data points of sharing bike count and it is expected due to weather condition.

   68% of the Bike sharing happens when the weather is Clear.
   Where most people avoid Light snow weather which constitute to only less than **1%**.



3.Holiday and Weekdays:

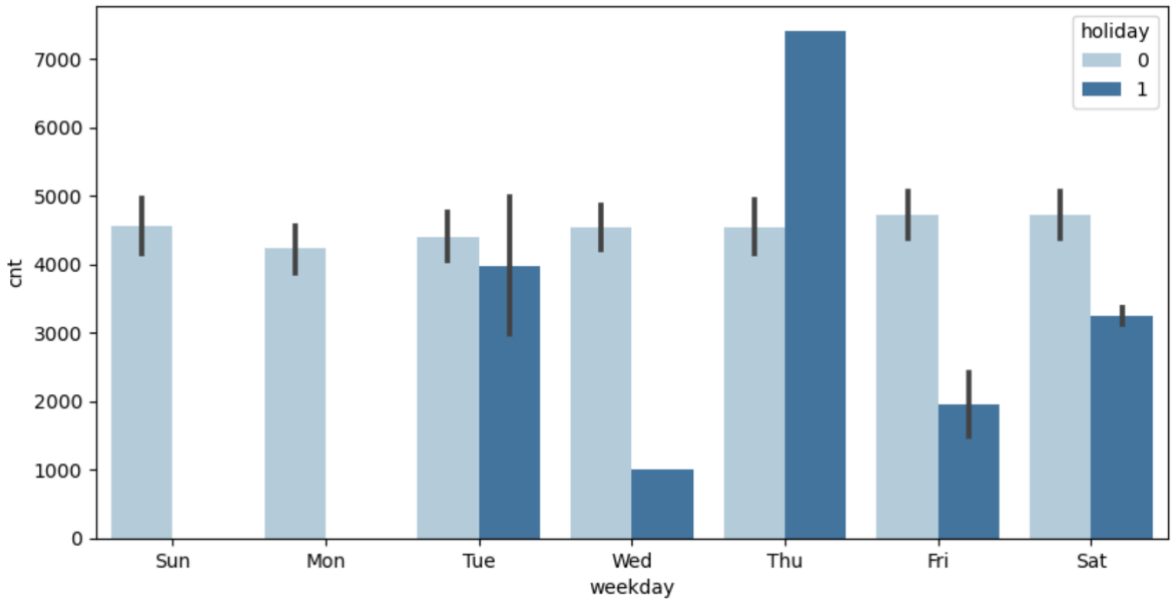
96% of the non-holidays have a bike sharing and people tends to rest over the day of holiday.

Irrespective of whether it's a holiday or not, bike sharing is happening.

There is no comparison for a holiday bike share with assuming highest usage through segment corporate employee which needs additional data to be explored.

Only on **Thursday** when it is holiday have BoomBike sharing have **160%** more shares than non-holiday Thursday.

No Bike share happened on Sunday and Monday when it is holiday.



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

By using drop_first=True during dummy variable creation, we ensure that we avoid **multicollinearity** in our regression models, which leads to good results. This step removes one category from each set of dummy variables, which prevents redundancy or duplicate and makes the model more efficient.

The coefficients of the remaining variables become easier to analyze, and overall helps in the model improvement.

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

The Pair-Plot tells us that there is a Linear Relation between 'temp','atemp' and 'cnt'.

'**atemp'** variable has the **highest** correlation with the target variable 'cnt' - 0.63
Followed by 'temp' 0.627

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

1. Linear relationship should be there between the dependent and the independent variables

2. Each observation should be unique. So need to cross validate the data.

3. Checking multicollinearity among variables, which can increase standard errors and affect the interpretation of coefficients. Examining correlations among predictors

4. Identify outliers and data points that may significantly affect the regression model

5. Checking the residuals (the differences between observed and predicted values) for normality and randomness

6. Checking the presence of correlation in the error terms which reduces the accuracy of the model

7. Conducting a Durbin-Watson (DW) statistic test.
   If DW=2, no auto-correlation

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

Based on my OLS regression results, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. **yr (Year)**: Coefficient = 2020.9785, t-value = 30.128, p-value < 0.001
   - This indicates that the year variable has a strong positive effect on bike demand.
2. **temp (Temperature)**: Coefficient = 2626.4103, t-value = 2.369, p-value = 0.018
   - Despite a lower t-value compared to 'yr', temperature still significantly contributes to bike demand.
3. **Light Snow**: Coefficient = -1688.1887, t-value = -7.430, p-value < 0.001
   - This negative coefficient suggests that the presence of light snow has a significant negative impact on bike demand.

These features are identified based on their coefficients, corresponding t-values and low p-values. They are crucial for understanding how variations in these factors affect the demand for shared bikes according to the model.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression tends to establish a relationship between dependant variable and Independent variable by finding best fit of the straight line.

Formula :- Y = mx+c where m = slope & c=Intercept

Linear regression at each x finds the best estimate for y.
Model predicts a single value, therefore there is distribution error.

Assumption of Linear regression :
1. Linear relationship between x and y should be there.
2. Error terms should be normally distributed and
3. Error terms should be independent to each other
4. Error term should have constant variance i.e., Homoscedasticity

Linear regression model involves below steps:
Data Understanding
Data Cleaning
Performing EDA (exploratory data analysis)
Model Building
Model Evaluation

1. Data Understanding :
   This stage involves gaining insights into the dataset, understanding the variables (x and y), their data types (whether it is continuous or categorical) and their relationships.

   Which involves getting the summary statistics using Univariate, Bivariate or multivariate technique.

   Checking correlation for all the variables

   Visualization also helps in understanding the data

2. Data Cleaning: mainly focuses on pre-processing raw data to ensure quality and consistency, making it suitable for modelling.

   Which involves handling missing data, outliers and transforming the data using standard scaler or standardization

3. EDA: involves exploring data visually and statistically to uncover patterns, anomalies, and relationships that helps model development.

4. Model Building : Building the Linear model for the above data by splitting the data into train and test.

   We can use OLS method which is one of the method.
   Developing the model to predict y in train data and run that prediction in test data.

5. Model Evaluation: Evaluating the performance and validity of the regression model to ensure it meets requirements and makes reliable predictions.
   Mean Squared Error (MSE),
   Root Mean Squared Error (RMSE),
   R-squared (R2)
   to measure model accuracy and goodness of fit.

Checking residuals for assumptions (normality, homoscedasticity, independence).

These stages collectively ensure that the linear regression model is built on clean, understood data, and thoroughly evaluated to provide reliable predictions and insights.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet is a series of four datasets that are used as a demonstration of the limitations of using summary statistics alone to analyse a dataset and the benefits of plotting and visualising data. The four datasets have the same or similar Mean values, Variances, Correlation, Coefficient of determination and line of best fit however when the data is displayed in a plot it can be seen that the individual data points are very different.

   Descriptive statistics alone are not enough to accurately analyse a dataset. It could be decided by looking at the summary statistics that these data sets are very similar.

   They have exact or close to exactly the same mean values, variances, correlations, linear regression lines and coefficients of determination. However when we plot these datasets as scatter plots it can be seen that they are very different.

   So we have to choose dataset that shows a linear regression line that correctly represents the data trend. This showcases the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3. **What is Pearson's R?**

   Pearson's R is the common way of measuring a linear correlation. Also be called as bivariate correlation, It measures between –1 and 1 that provides direction of the relationship between two variables.

   Three types of Pearson's R :
   Positive correlation: ranges between 0 and 1, when one variable changes, the other variable changes in the **same direction**.

   No correlation: ranges 0, there is **no relationship** between the variables.

   Negative correlation: ranges between 0 and -1, when one variable changes, the other variable changes in the **opposite direction**.

   The Pearson correlation assumption are as follows:

   - Both variables are quantitative: We need quantifiable variable.
   - The variables are normally distributed**:** We have to create a histogram of each variable to verify whether the distributions are approximately normal.
   - The data have no outliers **:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers. Looking for points that are outside.

- The relationship is linear**:** The relationship between the two variables can be described reasonably well by a straight line. We can use a scatterplot to check whether the relationship between two variables is linear.

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

Scaling is a pre-processing technique that transforms feature values to a similar scale, ensuring all features contribute equally to the model. It's essential for datasets with features of varying ranges, units, or magnitudes. Common techniques include standardization, normalization, and min-max scaling.

Data set contains features highly varying in ranges. Scaling should be done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Min-Max Scaling:
It brings all of the data in the range of 0 and 1. Min-Max Scaler helps to implement normalization in python.

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

Normalization	**Standardization**
Rescales values to a range between 0 and 1	Centers the data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/sd

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

The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

The VIF quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity. Specifically, the VIF for a predictor variable is calculated as the ratio of the variance of the coefficient estimate when that variable is included in the model to the variance of the coefficient estimate when that variable is not included in the model.

Which calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated

When choosing a VIF threshold, you should take into account that multicollinearity is a lesser problem when dealing with a large sample size compared to a smaller one.

If VIF = Infinity then an infinite value of VIF for a independent variable indicates that it can be perfectly predicted by other variables in the model.

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 is a graphical tool used in statistics to assess whether a given data set follows a normal distribution. It is a type of probability plot that compares the quantiles of the data set to the quantiles of a specified theoretical distribution.

In a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the theoretical distribution. If the points on the plot fall approximately along a straight line, then it suggests that the data is well-modelled by the theoretical distribution. Deviations from a straight line indicate deviations from the assumed distribution.

Q-Q plots are particularly useful for visually inspecting the fit of a data set to a theoretical distribution, identifying outliers, and assessing the presence of skewness or other departures from the assumed distribution. They are commonly used in exploratory data analysis and model diagnostics in statistics.

Q-Q plots summarize any distribution visually and very useful to determine :
If two populations are of the same distribution
If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
Skewness of distribution