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?**

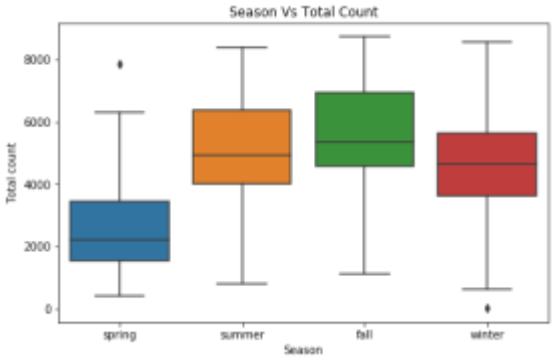


Fig 1

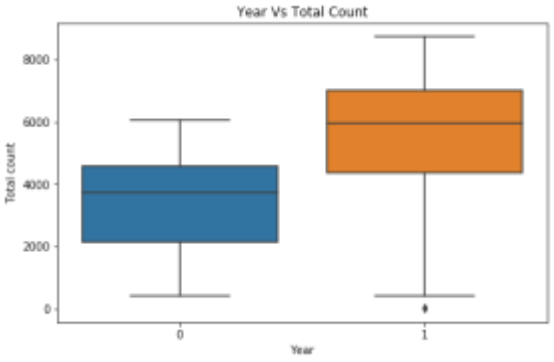


Fig 2

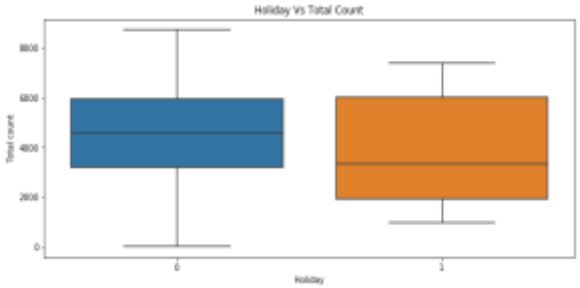


Fig 3

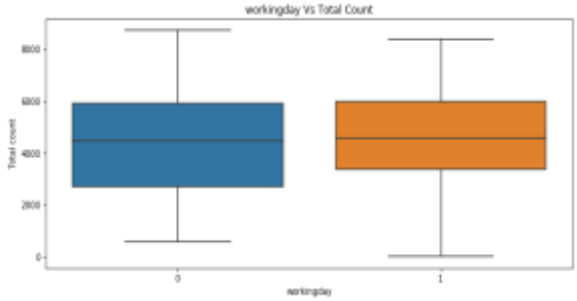


Fig 4

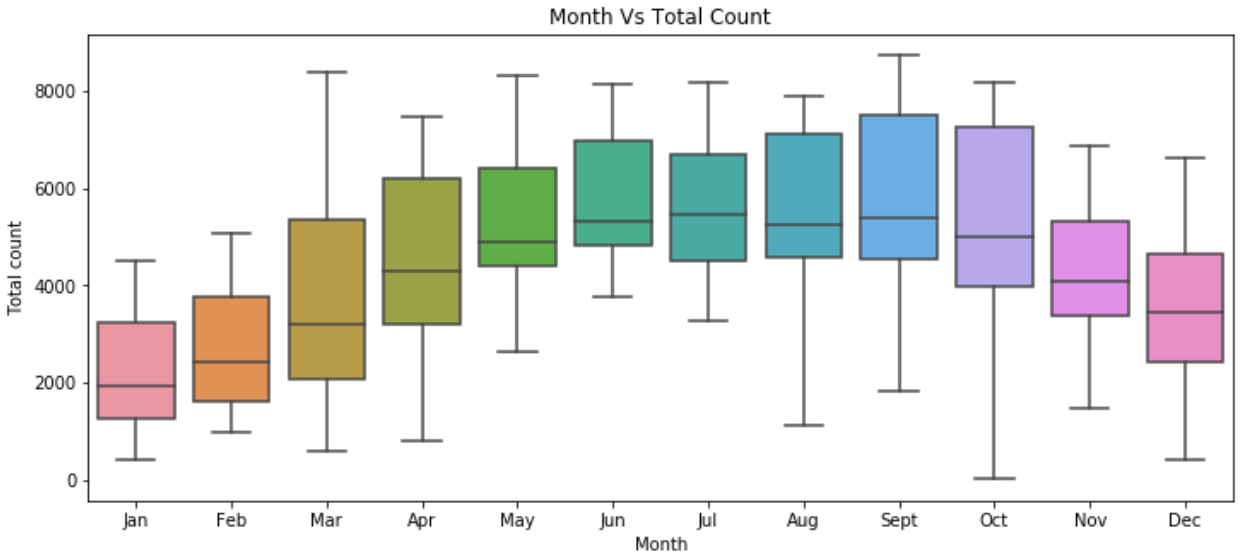


Fig 5

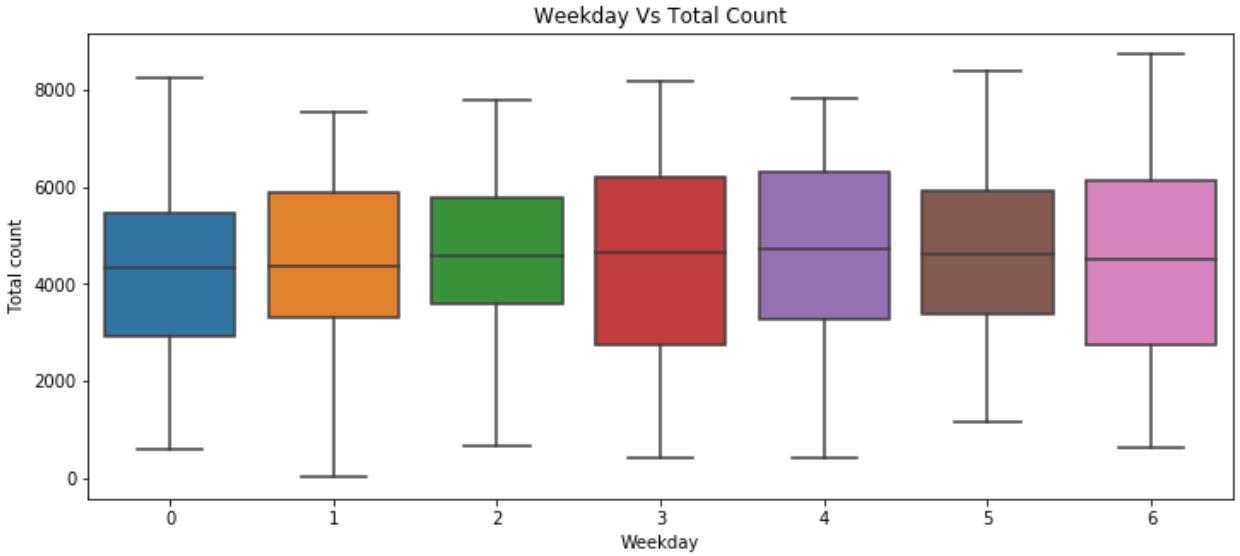


Fig 6

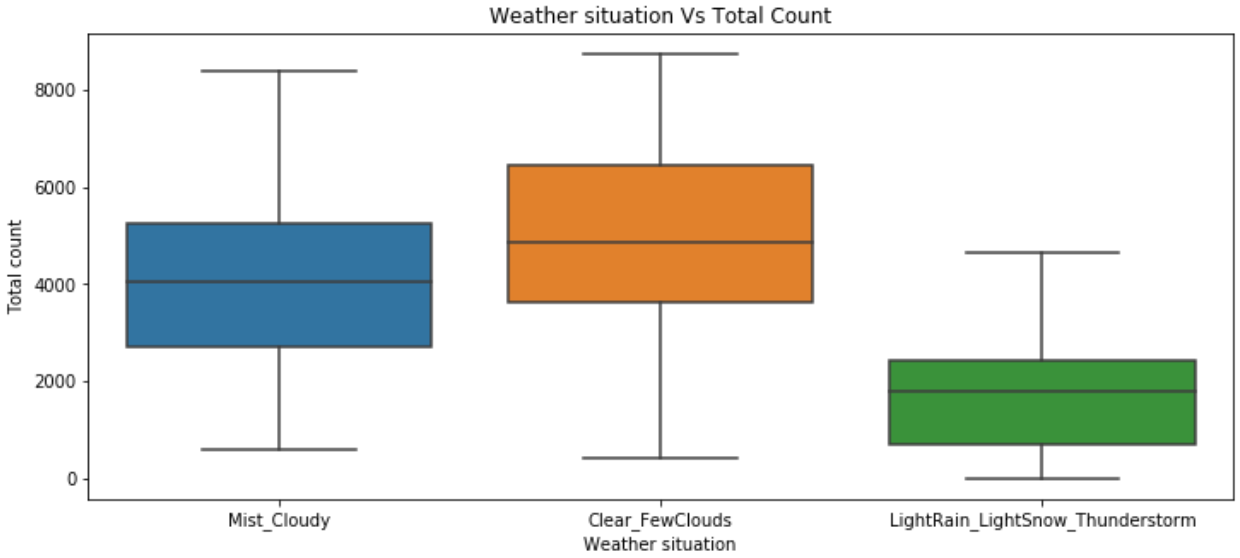


Fig 7

Fig 1 – Season vs Count

a. From the box plot we can observe that the count is low in spring as compared to other season variable.

Fig 2 – Year vs Count

a. From the box plot we can observe that the count of total bikes given on rent increased in number in 2019 as compared to 2018.

Fig 3 – Holiday vs Count

a. From the box plot we can infer that the count of total bikes given on rent is less on a holiday.

Fig 5 – Month vs Count

a. From the box plot we can infer that the count of total bikes given on rent is more the month starting June till Sept – this is in sync with the season with count chart where we see more rentals during summers and fall and a decline in rentals during winter and rainy season.

Fig 6 – Weekday vs Count

a. From the box plot we can infer that the count of total bikes given on rent is similar for all the days of the week.

Fig 7 – Weather Situation vs Count

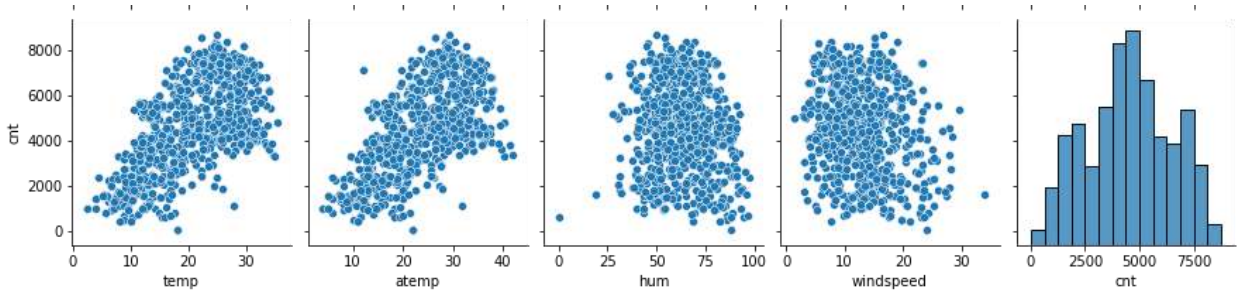
a. From the box plot we can infer that the count of total bikes given on rent reduces drastically during rainy/snowy season.

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

It is important to use drop_first = True during dummy variable creation because:

a. Having "n" dummy variables for a categorical variable with "n" different values will result in perfect multicollinearity among dummy variables created. Dropping the first dummy variable will help in overcoming this issue.
b. Using the coefficient of the remaining variables one can interpret the value for the first dropped variable.

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



As can be seen from the plot above, temp & atemp are the variables which have the highest correlation with the target variable.
These two variables are highly correlated with each other as well.

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

We can validate the assumptions of Linear Regression by residual analysis and draw plot to confirm whether the error is normal distributed or not.

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

Based on the final model, following is the equation of the best fit line:
cnt = yr * 0.2384 - holiday * 0.0754 + temp * 0.4233 - windspeed * 0.1398 - season_spring * 0.1286 + season_winter * 0.0681 - mnth_Dec * 0.0501 - mnth_Jul * 0.0791 - mnth_Nov * 0.0650 - weathersit_LightRain_LightSnow_Thunderstorm * 0.2721 - weathersit_Mist_Cloudy * 0.0738

Looking at the coefficients we can see that the top 3 features contributing significantly are:

1. temp – with coefficient of 0.4233
2. weathersit_LightRain_LightSnow_Thunderstorm – with coefficient of -0.2721
3. yr – with coefficient of 0.2384

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Linear Regression is a statistical technique used to figure out the linear relationship between a dependent variable and one or more independent variables. The objective of Linear Regression is to find the best fit line that minimizes the error coefficient – that is the difference between the observed data points and the regression line.
The linear regression can be represented by the following straight line equation:
Y = b0 + b1x1 + b2x2 + ----- + bnxn + e0

Where y = dependent variable
b0 – y intercept when all the independent variables are zero
b1 to bn – regression coefficients.
e0 – random error term

Regression coefficient signifies the weight of the contribution of independent variables towards the dependent variable.

To fit the regression line – we use the least square method to estimate the regression coefficients that minimizes the sum of all squared residuals (difference between the actual and the predicted values.)

Once the model is fit and we have the best fit line – we can assess the goodness of fit using R-squared or Adjusted R-squared values to check their statistical significance.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet refers to a set of 4 datasets having very different distribution and at the same time having nearly identical simple descriptive statistical properties like mean, variance, correlation and linear regression.
Anscombe's quartet demonstrates how looking only at the summary statistics can be misleading compared to visualizing the data.

The 4 datasets in Anscombe's quartet are:

1. Linear data with no outliers
2. Linear data with one outlier

3. Non-linear data
4. Linear data with high variance

Despite the differences in the data distribution, all these 4 datasets have:

- Same mean x value (9)
- Same mean y value (7.5)
- Same variance (11)
- Same Pearson correlation coeff between x and y (0.816)
- Same linear regression line: y = 3 + 0.5x

But when we plot the datasets and look at the various plots/graphs we can visually see the differences.

Anscombe's quartet highlights the importance of plotting the data to uncover the actual patterns while analyzing the data rather than only reporting the summary statistics.

He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

3. What is Pearson's R?

Pearson's R is also known as the Pearson correlation coefficient, is a statistical measure of linear correlation between two sets of data. It quantifies the strength and direction of the linear relation between two variables.

It is the ratio between the covariance of two variables and the product of their standard deviations, hence it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

Few important properties:

- Pearson's R values range from -1 to +1. A value of 0 indicates no linear correlation. +1 indicates a perfect positive correlation and -1 indicates a perfect negative correlation.
- The sign of correlation represents the direction of association. A positive value means both variables move in same direction and a negative value represents they move in opposite direction.
- It measures the strength of linear association. The values close to 0 denotes a weak correlation and the values close to 1 (absolute value) denotes a strong correlation.

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

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. There are various techniques used to normalize/standardize the features in the dataset.
During analysis, we should note that the feature scaling only affects the coefficients and none of the other parameters like t-stat, p-value, r-square etc.
The main purpose of scaling are:

- To standardize the variables so that they are comparable to each other. It prevents one variable from dominating over other variables. Suppose if the range of values of one variable is very large as compared to other variables, on doing the regression the coefficient of this variable might come very different from other coefficients.

- Scaling also helps in removing the outliers that can negatively impact model performance. It also helps in reducing distortion and skewness in the data.
- After scaling the variables, it is noticed that it helps machine learning optimization algorithms converge faster and hence helps in faster computation.

Normalized scaling (also known as Min Max scaling) rescales the range between 0 and 1.
Standardized scaling rescales the data to have a mean of 0 and standard deviation of 1.

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

An infinite value of VIF indicates the presence of perfect multicollinearity between independent variables. In other words we can say that it happens when one independent variable is a perfect linear function of another independent variable. In the case of perfect correlation, we get R2 as 1, which lead to 1/(1-R2) as infinity.

Another reason could be if there are duplicate variables ie. the two variables measures the same thing (duplicate variables). Removing duplicates here can solve the infinite VIF problem.

To solve this infinite VIF problem we need to identify and drop the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q plot is also known as quantile-quantile plot. It is used to check if a dataset follows a normal distribution or not. It is an important tool in hand during linear regression analysis. In a Q-Q plot, the quantiles of the observed data are plotted against the quantiles of the comparison distribution.

Following are the uses/importance of Q-Q plot:

- A Q-Q plot of the residuals against a normal distribution can detect non-normality. Normality of the residuals is a key assumption in linear regression.
- A Q-Q plot can be used to identify outliers. Points deviating from the trend line indicate potential outliers.
- Patterns in the residual Q-Q plots highlight skewness etc.
- Q-Q plots provide a visual assessment of linear regression assumptions like normality, homoscedasticity etc.