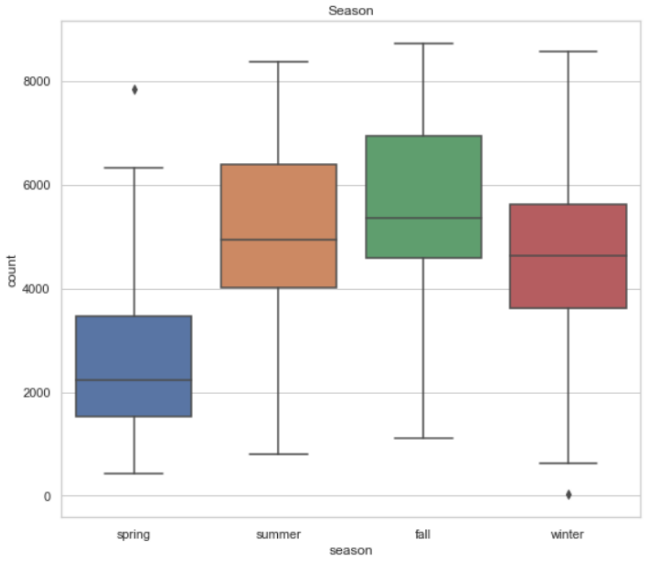
Name- Vibhu Dutt Raturi

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)

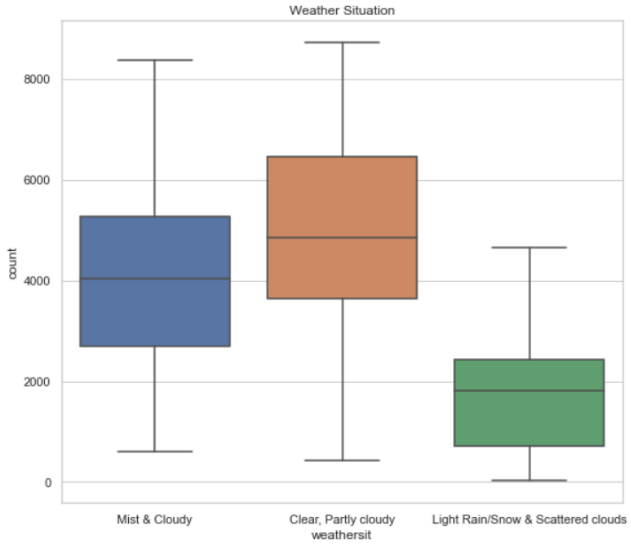
Answer – There were 7 categorical variables in the dataset provided.

'season' – There were 4 categories in this variable. Spring, Summer, Fall and Winter.



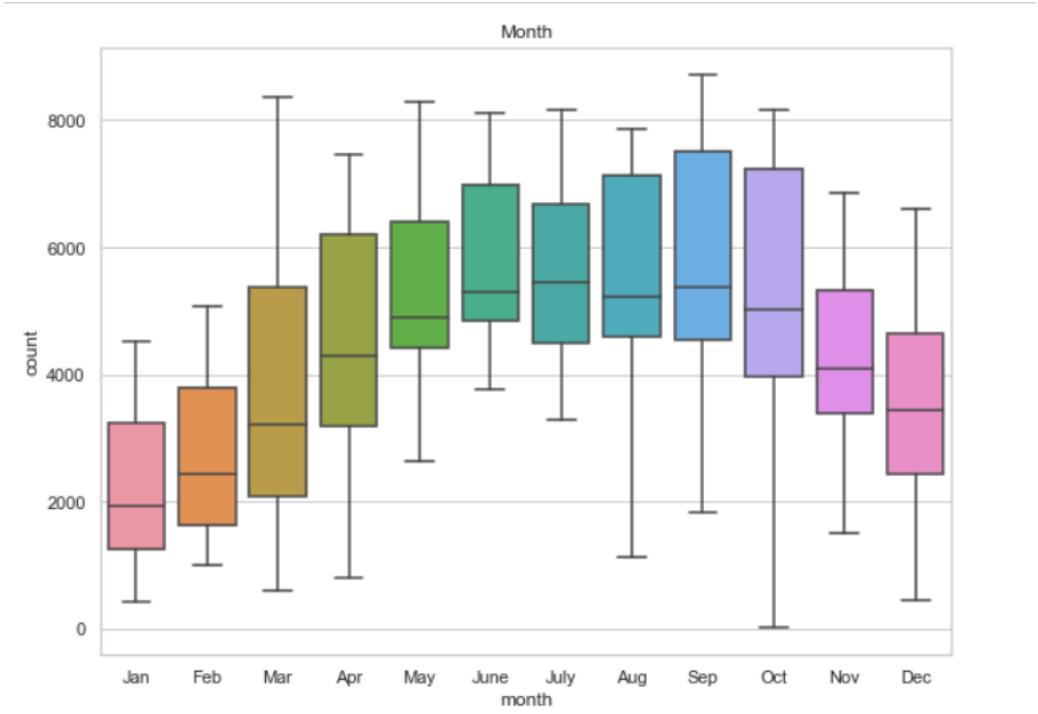
This shows the distribution of season with respect to total count of bike sharing. Fall season has most number of counts followed by summer, winter and spring.

'weathersit' – There were 4 categories in this variable. Mist & Cloudy, Clear, Partly Cloudy, Light Rain/Snow & Scattered Clouds and Heavy Rain, Thunderstorm, Mist & Snow.



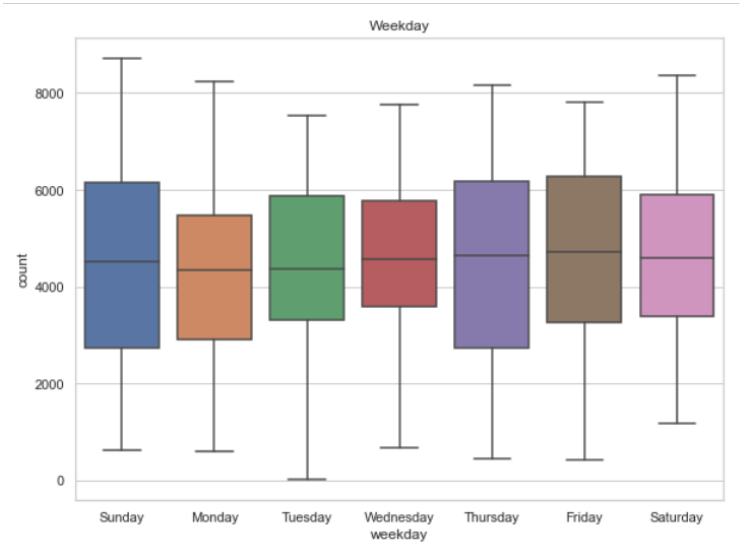
This shows the distribution of Weather Situation with respect to total count of bike sharing and we can see when the weather is bad ie Rain/Snow there is a significant drop in bike sharing count. Usually, people prefer bike sharing when the weather is Clear/Partly Cloudy.

‘month’ – This variable has months from Jan to Dec.



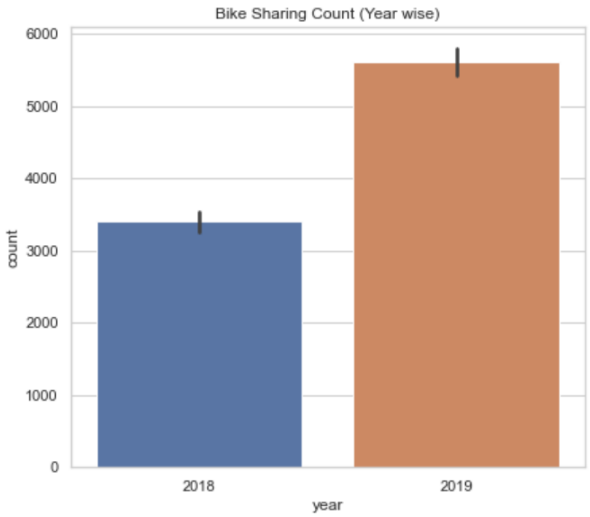
This shows the distribution of month with respect to total count of bike sharing and as we can see that the business starts picking from the very first month of January and from June to October, bike sharing is the most with September month being most popular, as we have already seen fall season having the most bike sharing counts. After October there is a significant dip towards the end of a year.

‘weekday’ – This variable has 7 days of a week.



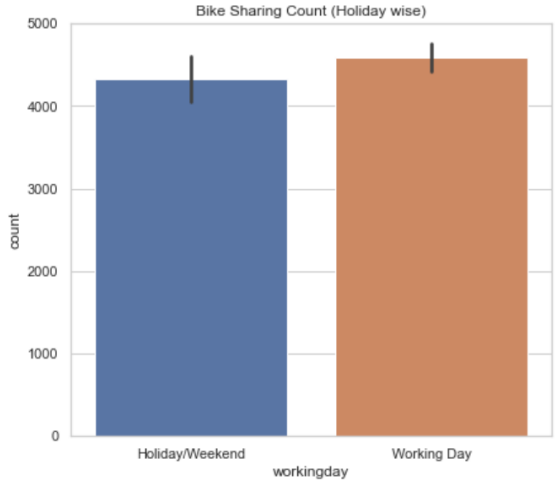
The distribution of Weekday with respect to total count and mostly the visual seems almost same for all the days with Friday & Sunday having more counts.

‘year’ – This variable has two categories (0: 2018, 1:2019)

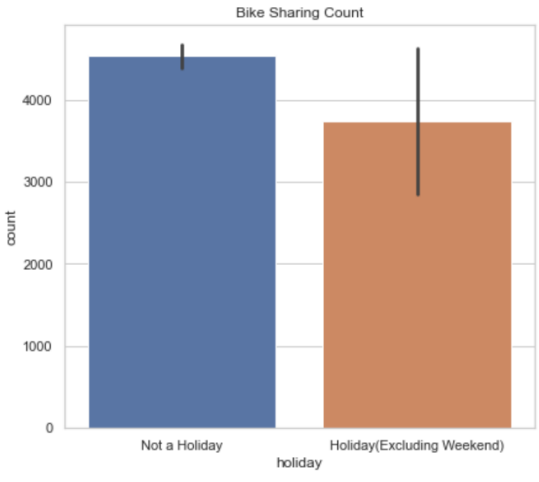


We can clearly see that how the business jumps up from 2018 to 2019 in terms of bike sharing total counts.

‘workingday’ - if day is neither weekend nor holiday is 1, otherwise is 0.



Bike Sharing Count on a working day is more than the count if it’s a Holiday or if it’s a weekend.



This Visual above gives us an insight that bike sharing count increases if it is a non-holiday i.e. a weekday or weekend but bike sharing count drops if it is Holiday.

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

   Answer - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   Dropping your first categorical variable is possible because if every other dummy column is 0, then this means your first value would have been 1.

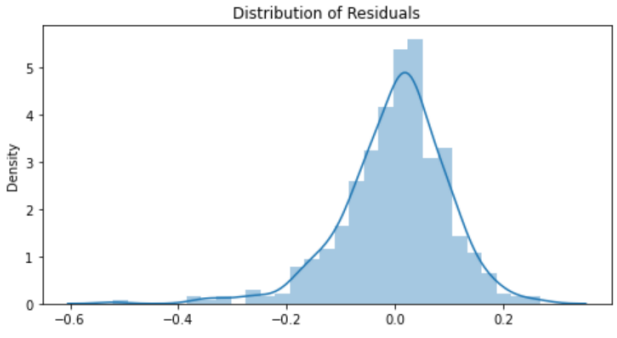
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Answer – Looking at the pair plot 'registered' variable seems to have the highest correlation with the target variable 'cnt'.

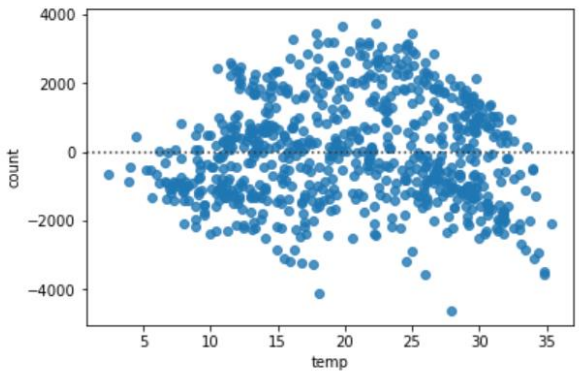
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Answer – We can validate the assumptions of Linear Regression
   Multicollinearity – Independent variables should not be multicollinear with each other that's one assumption in linear regression so while we are building a model, we need to check for variance inflation factor (VIF) which gives excludes the constant and target variable and gives us a value for each independent variable, if VIF >5 for a specific variable it means it is highly dependent on other independent variables.

   Residual Analysis – When a model is ready, we need to validate assumptions related to error terms as well. We check distribution of residuals which is the difference in actual y_train value and the y_predicted value by our model and plot these errors. It should follow a normal distribution. Which we can check using distplot or a q-q plot.



   Residual Analysis – One more assumption related to residuals is that they should have constant variance throughout and should not have a pattern for eg variance increasing with increase in value of y. This is called Homoscedasticity. Which we can check using a scatter plot or a residplot.



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

   Answer – Top three features in our model
   1- 'temp' – The coefficient value is 0.5499
   2- 'Light Rain/Snow & Scattered Cloud' – The coefficient value is -0.2871. (negative sign implies negative correlation)
   3- 'year' – The coefficient value is 0.2331

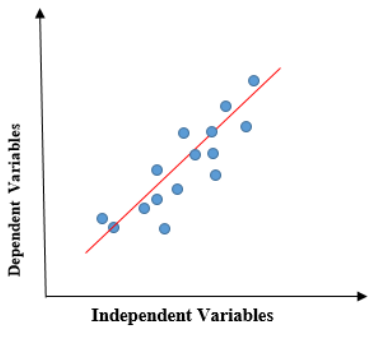
General Subjective Questions

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

   Answer - Linear regression is a statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression.

   If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

   The linear regression model gives a sloped straight line describing the relationship within the variables



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate the line of best fit we use the equation y = mx +c

Where y = Dependent variable, x = Independent variable, m= Slope of the line(linear regression coefficient which is called Beta1), c=intercept of the line(Beta0)

The cost function helps to figure out the best possible values for b0 and b1, which provides the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

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

Answer - Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The below image has 4 datasets with 11 data points.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary of these 4 datasets

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

3. What is Pearson's R? (3 marks)

Answer - Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a

population and the letter “r” for a sample.
Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association
r > 8 means there is a strong association

Pearson’s R formula -

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

   Answer - It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not 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.
   It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

   Normalization/Min-Max Scalar - It brings all of the data in the range of 0 and 1.
   sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
   It is useful when we don’t know about the distribution.
   Affected by outliers
   Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
   It is useful when the feature distribution is Normal or Gaussian.
   Less affected by outliers

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

   Answer - If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve it we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

   Answer - Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like

normal, uniform, exponential.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.