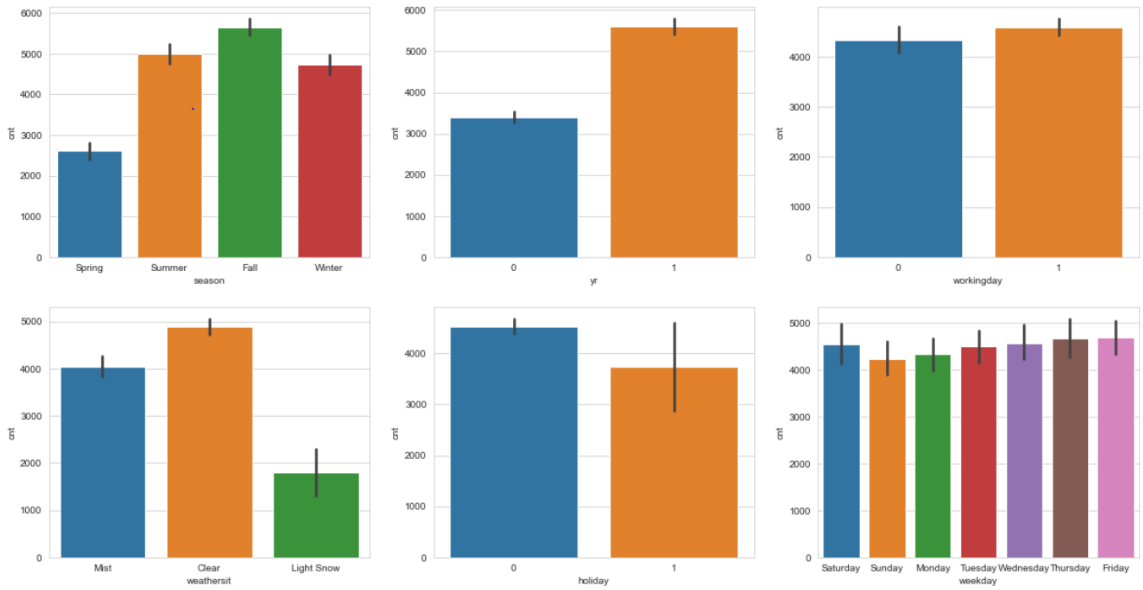
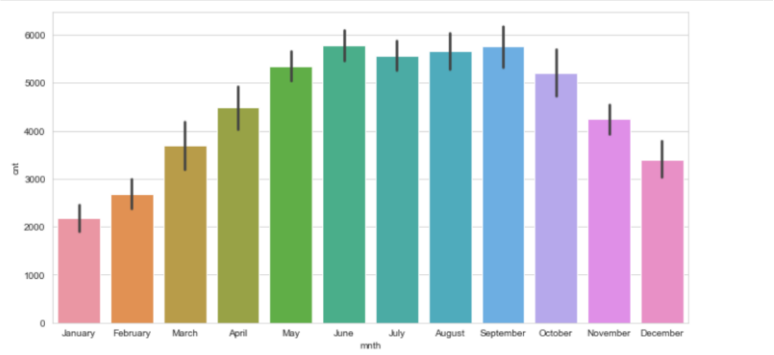
Assignment-based Subjective Questions

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 Vs Count** – Most number of bikes was taken in the 'Fall' season (June to September) followed by 'Summer'
- **Year Vs Count –** Most number of bikes was taken during the '2019' (The impact of count may be of gaining popularity)
- **Working day Vs Count –** Working & Non-Working Day has not have much significant
- **Weather Situation Vs Count –**
    - Most number of bikes was rented during the (Clear, Few clouds, partly cloudy) conditions
    - Very less number of bike was rented on (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)
    - Also, it is inferred, there were no (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) situation for the past 2 years
- **Holiday Vs Count –** Holiday has the negative impact on the number of bikes rented.
- **Weekday Vs Count –** Count wise the number of bike rented on Saturday is high but not very high compared to other day.
- **Month Vs Count –** It is almost known from the season most of the bike was taken between May and October (summer and Fall season). Most number of bikes rented on September.



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

To avoid the Dummy Variable Trap which means the dummy variable is created for the categorical variable, if we use all the values of categorical column which does not explain the relationship well. The extra column that is redundant can be removed. By dropping this dummy variable, we can reduce the co-relation among the dummy variables. The main reason is to avoid adding multicollinearity added to the model

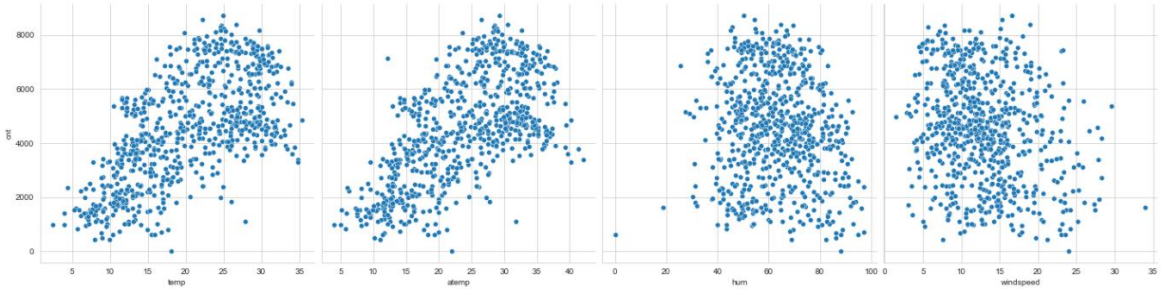
If 3 types of values present in the categorical value (Example - car, Bike, Train)
The dummy variables be

Value	Car	Bike	Train
Car	1	0	0
Bike	0	1	0
Train	0	0	1

This can also be explained as Car- 10., Bike-01 and train -00, so dropping the train dummy variable.
So basically, if n level of categorical value we need to create n-1 dummy variables.

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

From the pair plot it is seen that the Highest co-relation is between '**Temp**' and target variable when compared with other numerical variable. It almost forms the linear plot between the variables, with increase in temp which cause increase in the count.



The Co-relation between temp and 'Cnt' is positive and about 63% inferred from the heatmap.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    - **Linearity -** The linear relationship between the independent and target variable is studied, In the data set the 'Temp' and 'cnt' has the linear plot find using the pair plot
    - **Homoscedasticity –** The residuals have constant variance no matter the level of the dependent variable. The variance of the error terms is constant across the values of the dependent variable from the residual plot. In the data set all were residuals were plotted and the values are scattered.
    - **Absence of Multicollinearity –** Multicollinearity refers to the two or more independent variables co-related. This can be identified using the Variance Influence Factor method. In the data set all the VIF values of feature are less than 5.
    - **Normality of Errors –** From the histogram, it is identified the error terms are normally distributed and the most of value is centred towards zero. The mean of residuals should be zero

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

From the Final model the top features are:

- **Temperature** has the maximum impact on the target variables with the positive coefficient of 0.422. Increase in temp, count increases.
- **(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) Weather situation -** this is the 2nd highest demand with the negative coefficient of 0.289, Increase in this weather situation, count decreases.
- **Year –** This has the positive coefficient of 0.235, Increase in year, count increases.

General Subjective Questions

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

Linear Regression is used in the supervised model. It is used to analyse the linear relationship between the dependent and the independent variables. It is of two types simple and multiple linear regression. The value of the one of independent variables change (increase or decrease) the dependent variable also change (increase or decrease)

It is represented by Y= MX + C

Y – Dependent variable, X- independent variable, M – slope, C – Constant

- The Data set needs to be understood and visualize to find out the linear relationship of the independent and dependent variable.

- The dummy variable column should be created for the categorical value of a column and the column should be treated with outlier method
- The test set and train set is created to train and test the model
- The RIF used to find the best suitable feature of the model
- The OLS (Ordinary least square) method is used to learn the co-efficient and intercept of the selected feature
- The selected feature should be removed one by one which has P value >0.05 and Variable Influence Factor(VIF) > 5
- The residual analysis is done by checking the error terms are normally distributed and the value centred to zero.
- The model should be predicted with the test set based on the R-squared and Adjusted R-squared value

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

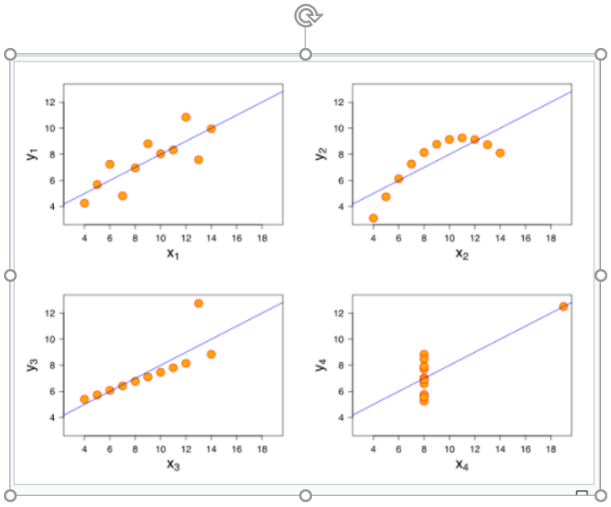
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

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 and other influential observations on statistical properties.

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

Anscombe's quartet

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



For all the set the Mean of X is 9 , Y is 7.50 . Co-relation is 0.816

Regression line is Y= 3+ 0.5X

R-square – 0.67

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

Pearson's r is the measurement of the strength of the relationship between two variables and their association with each other. Pearson's correlation coefficient calculates the effect of change in one variable when the other variable changes.

The Value of the Pearson's R will be range between 1 and -1

- **Positive linear relationship:** In most cases, the income of a person increases as his/her age increases. Its R value is 1
- **Negative linear relationship**: If the vehicle increases its speed, the time taken to travel decreases, and vice versa. . Its R value is -1
- **No relationship** between 2 variables if the R value is Zero

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

- Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.
- Most of the times, collected data set contains features highly varying in magnitudes, units and range. In this case there may be incorrect modelling because of the varying range of values in the independent variables. This can be rectified by scaling to bring all the variables to the same level of magnitude. Also, it helps in speeding up the calculations in an algorithm.
- 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

Normalized scaling

It brings all of the value of the variable in the range of 0 and 1.

Formula: Minmax Scaling X = X – Min(x)/Max(x)- Min(x)

Standardized scaling

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)

Formula: X = (X – Mean(x))/Sd(x)

Normalized scaling	**Standardized scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
It is really affected by outliers.	It is much less affected by outliers.

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

VIF = infinity, only for the perfect correlation. If there is any perfect correlation between two independent variables then R2 =1, which lead to 1/(1-R2) infinity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity with other.

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

- Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- 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.
- 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.

Importance of Q-Q plot in Linear regression

- Two datasets/sample can be of different size.
- Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.
- One of the important assumptions of Linear Regression is that the residual of the model is normally distributed. This can be assessed using Q-Q plot.