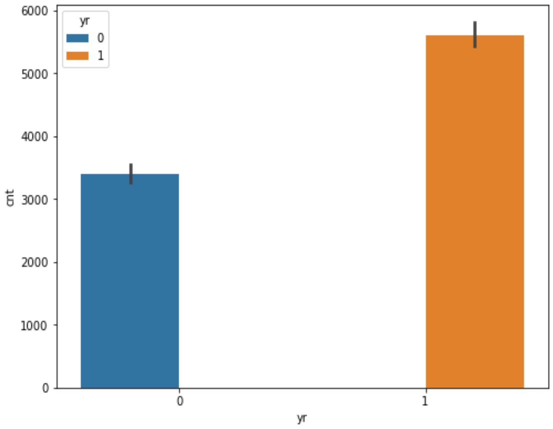
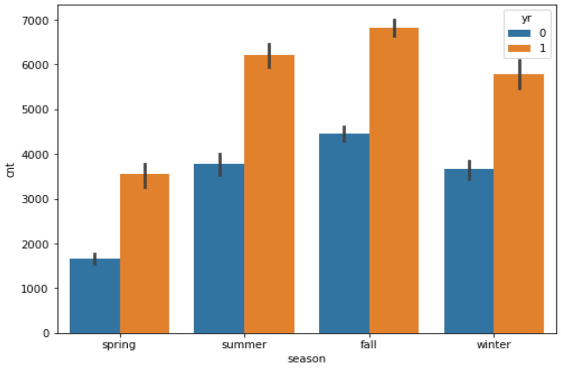
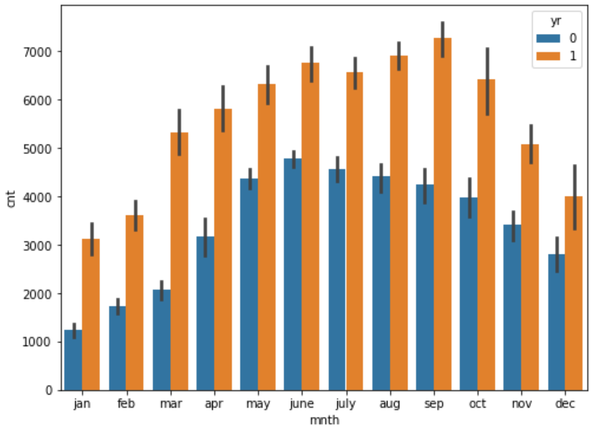
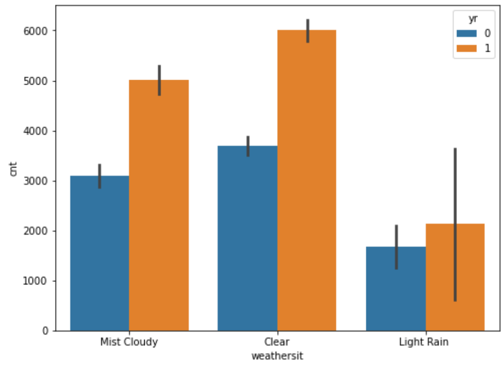
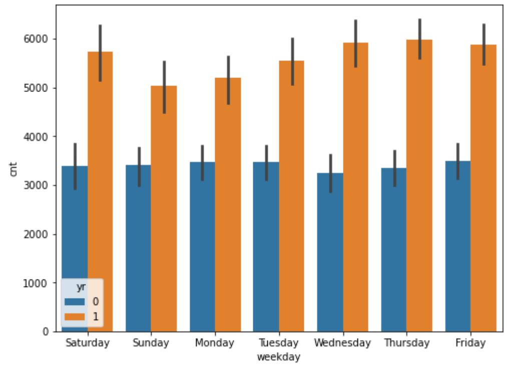
Assignment-based Subjective Questions

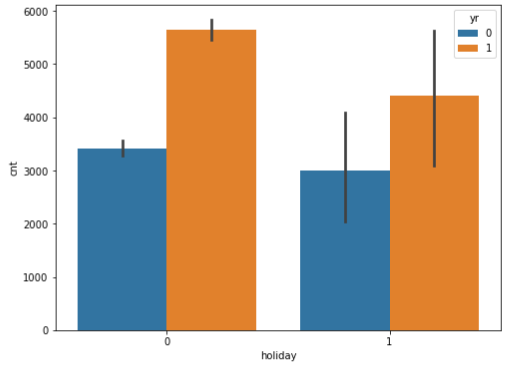
Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

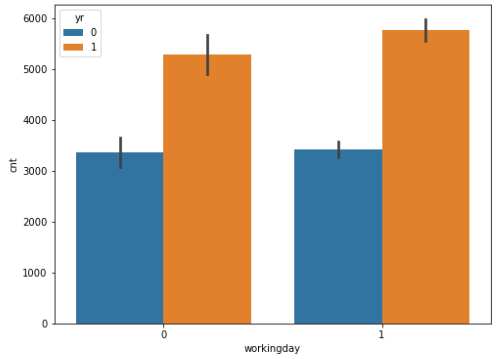
Ans:

 	Feature: Yr (Year) Yr 0: 2018 Yr 1: 2019 As per analysis, it is clean that Year is having a good impact. Booking in year 2019 are high compared to year 2018. This will be a good predictor variable.
 	Feature: season Season Summer and Fall are having more bookings compares to others in both years. This visible influence make season is good predictor variable.
 	Feature: mnth (Month) There is an increase in bookings from Mar to Sep and dec in other months. There are both slopes in this variable’s data which makes it a good predictor variable.

<table>
<tr>
<td>
</td>
<td>Feature: weathersit (Weather Situation)

There is a visible increase in bookings due Clear weather situation compared to others. It makes weather situation a good influencer variable.</td>
</tr>
<tr>
<td>
</td>
<td>Feature: weekday

All days are having similar bookings. It seems no significance impact on target.</td>
</tr>
<tr>
<td>
</td>
<td>Feature: holiday

This is an impact due to holiday on target which make it a potential predictor variable.</td>
</tr>
<tr>
<td>
</td>
<td>Feature: workingday

All days are having similar bookings. It seems no significance impact on target.</td>
</tr>
</table>

Categorical Features year, session, month, weather situation and holiday are having impact on target whereas features working day, weekday are not impacting the target variable.

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

Ans: Dummy variables are the Numeric representation of categorical data that can only take on one of two values: zero or one.

Newly generated Dummy variables count will match with number of different values that the categorical variable has. These variables may have high correlation which can impact the target variable also.

We only need *n*-1 variable to describe *n* different values. This drop_first=True due the same automatically. It reduces the extra column created during dummy variable creation. 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 : Among numerical variables, temp and atemp variables are having highest correlation with target variable with value 0.63. As these two are highly correlated with each other, we can keep one variable only. Later RFE automatically removes atemp. Hence, temp is the highly correlated with target variable.

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

Ans : Below are the Linear regression assumptions:

1. Linear Relationship: To validate this, scatter plot is generated between predicted and outcome values, and best fit line is drawn.
2. Normal distribution of Error term: Validated with histogram plot that Error term is following normal distribution curve.
3. No overfit: Checked R-square values of test and train model which are close.
4. Little or No Multicollinearity: Keep those variables only whose VIF is in acceptable range <5
5. Homoscedasticity: Validated with residual plot

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

Ans: Equation for best fitted line can be drawn as below

cnt = 0.2218 + 0.2343 x yr - 0.0922 x holiday + 0.4502 x temp - 0.1496 x windspeed + 0.0501 x 3 + 0.0443 x 4 + 0.0551 X 5 + 0.0837 X 9 - 0.2891 x Light Rain - 0.0804 x Misty Cloudy - 0.0901 x spring + 0.0785 x winter

From the equation, we can see that top 3 features are

1. temp (temperature),
2. yr (Year),
3. Light Rain.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

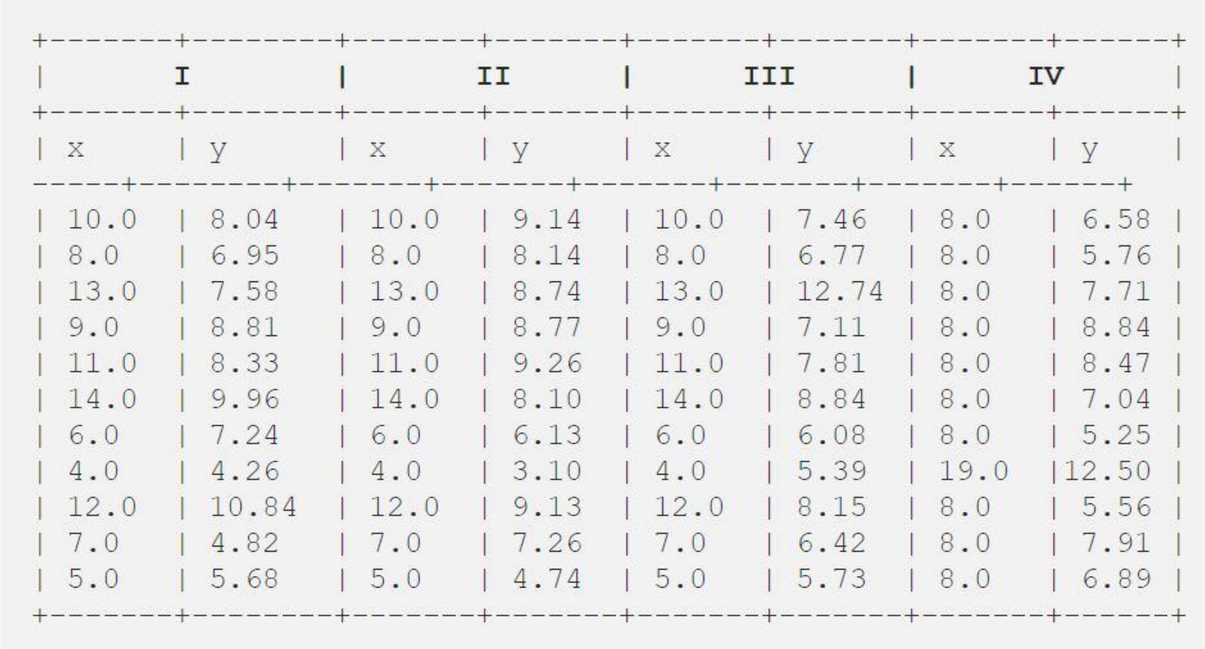
y = Dependent variable from dataset

2. Explain the Anscombe’s quartet in detail.

Ans : 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.

Simple understanding:

Once Francis John “Frank” Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.



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

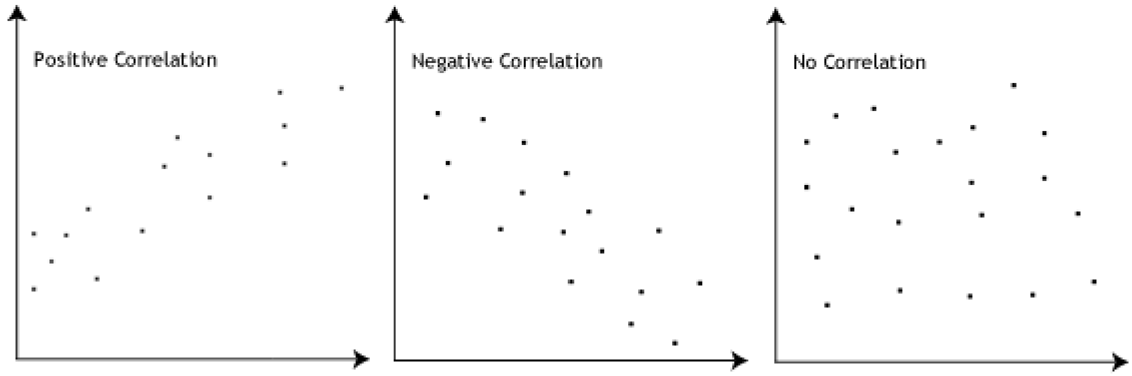
After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

3. What is Pearson’s R?

Ans : Pearson's r or Pearson correlation coefficient (PCC) is a measure of linear correlation between two sets of data. It is the covariance of two variables divided by the product of their standard deviations. It is a normalised measurement of the covariance with values varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope

- r = -1 means the data is perfectly linear with a negative
- r = 0 means there is no linear association



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

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

Ans: Scaling is data Pre-Processing step for normalize the data within a particular range which is applied to independent variables. It helps in acceleration of calculations.

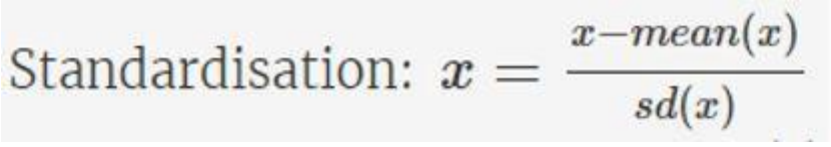
Mostly, the acquired data set comprises features with widespread in magnitudes, units and range. If scaling is not done, then algorithm only consider magnitude and not units hence resulting in incorrect modelling. To get rid from this issue, Scaling is the to set all the variables to the same level of magnitude.It only impact **the coefficients** and no impact on other factors like **t-statistic, F-statistic, p-values, R-squared**, etc.\

Scaling Types:

Normalization/Min-Max Scaling: It maps all of the data between 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization replaces the values by their Z scores. It converts values into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Ans: If VIF is infinite then we can say that there is perfect correlation between two independent variables. Because

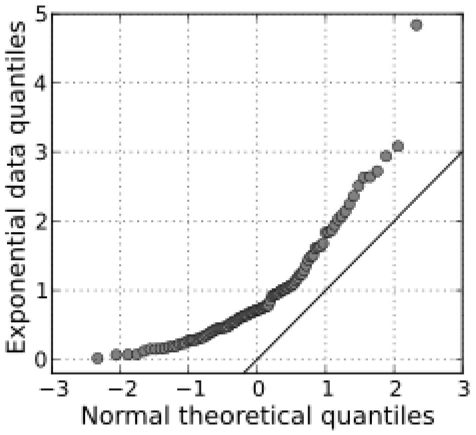
VIF = 1/(1-R2) if VIF is infinite, it means that R2=1

And R2=1 means two variables are in perfect correlation. In this case, corresponding variable may be expressed by linear combination of other variables. To solve this problem, we need to drop one of the variable from data based on business need. Ex. In this Bike data set, target variable ‘cnt’ is also expressed as casual + registered. It means if we keep casual and registered in model, we will get infinite VIF for same.

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

Ans: 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. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



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.