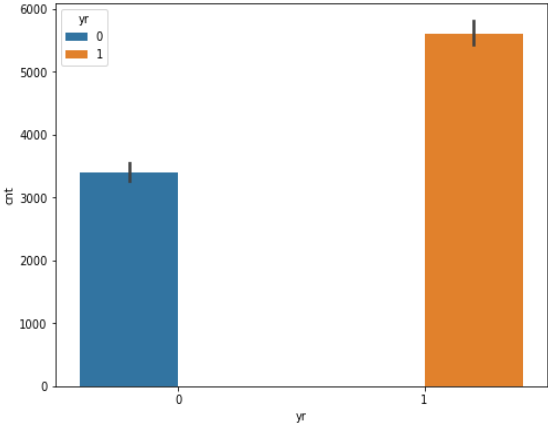
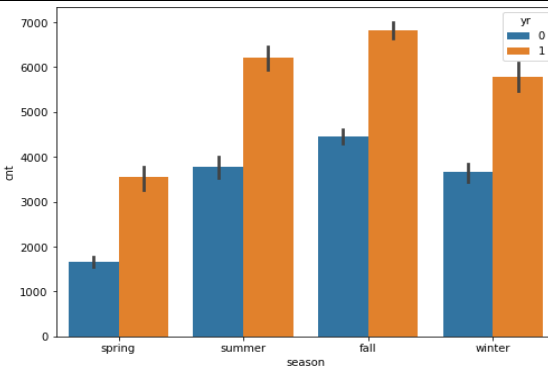
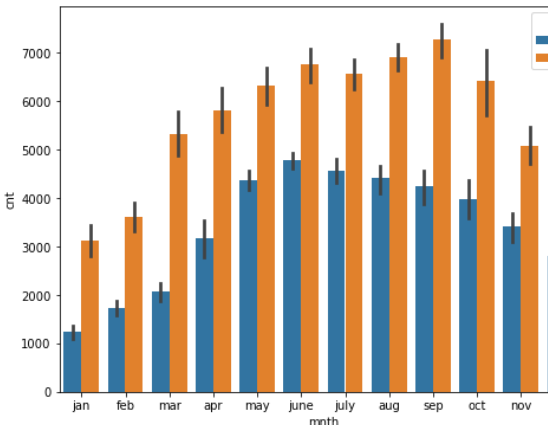
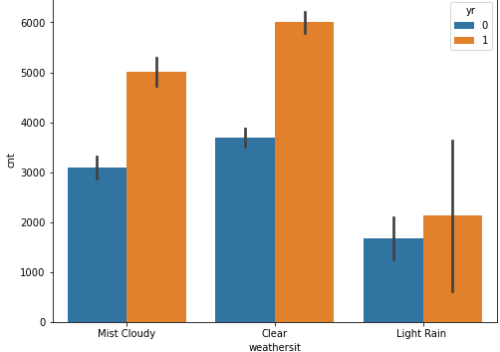
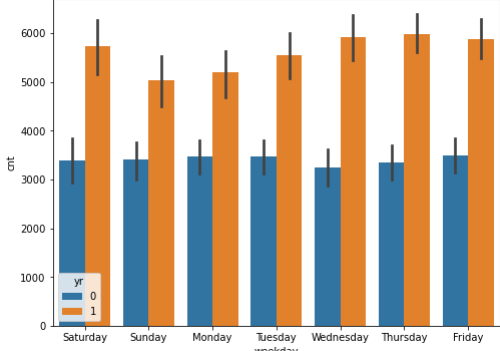
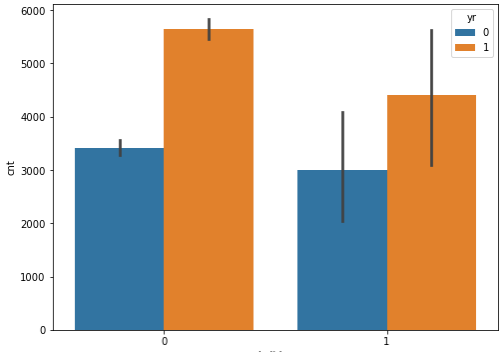
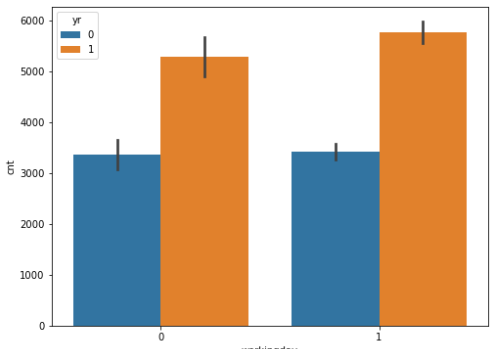


Assignment-based Subjective Questions

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

Ans:

 <p>A bar chart comparing the count of bookings for two years. The x-axis is labeled 'yr' with values 0 and 1. The y-axis is labeled 'cnt' and ranges from 0 to 6000. For yr 0 (blue bar), the count is approximately 3400. For yr 1 (orange bar), the count is approximately 5600. Error bars are present on both bars.</p>	<p>Feature: Yr (Year) Yr 0: 2018 Yr 1: 2019</p> <p>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.</p>
 <p>A grouped bar chart showing the count of bookings by season for two years. The x-axis is labeled 'season' with categories spring, summer, fall, and winter. The y-axis is labeled 'cnt' and ranges from 0 to 7000. For each season, there are two bars: blue for yr 0 and orange for yr 1. In all seasons, the count for yr 1 is higher than for yr 0. Summer and fall have the highest counts in both years. Error bars are present on all bars.</p>	<p>Feature: season</p> <p>Season Summer and Fall are having more bookings compares to others in both years. This visible influence make season is good predictor variable.</p>
 <p>A grouped bar chart showing the count of bookings by month for two years. The x-axis is labeled 'month' with categories from jan to dec. The y-axis is labeled 'cnt' and ranges from 0 to 7000. For each month, there are two bars: blue for yr 0 and orange for yr 1. The data shows a seasonal trend where bookings are lower in winter months and higher in summer months. Year 1 consistently has higher counts than Year 0 across all months. Error bars are present on all bars.</p>	<p>Feature: mnth (Month)</p> <p>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.</p>

	<p>Feature: weathersit (Weather Situation)</p> <p>There is a visible increase in bookings due Clear weather situation compared to others. It makes weather situation a good influencer variable.</p>
	<p>Feature: weekday</p> <p>All days are having similar bookings. It seems no significance impact on target.</p>
	<p>Feature: holiday</p> <p>This is an impact due to holiday on target which make it a potential predictor variable.</p>
	<p>Feature: workingday</p> <p>All days are having similar bookings. It seems no significance impact on target.</p>

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

$$\text{cnt} = 0.2218 + 0.2343 \times \text{yr} - 0.0922 \times \text{holiday} + 0.4502 \times \text{temp} - 0.1496 \times \text{windspeed} + 0.0501 \times 3 + 0.0443 \times 4 + 0.0551 \times 5 + 0.0837 \times 9 - 0.2891 \times \text{Light Rain} - 0.0804 \times \text{Misty Cloudy} - 0.0901 \times \text{spring} + 0.0785 \times \text{winter}$$

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

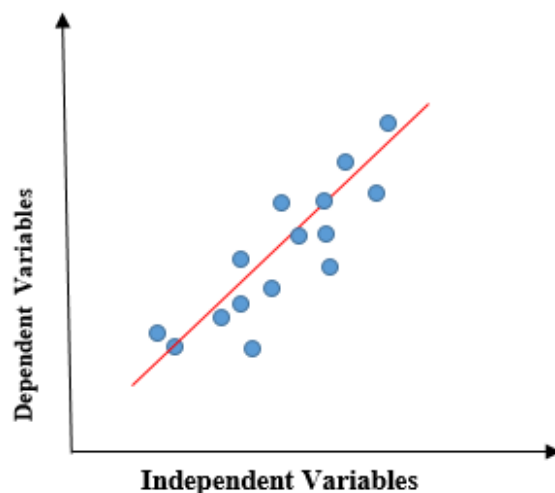
1. temp (temperature),
2. yr (Year),
3. Light Rain.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is one type of machine learning form in which model is trained to predict the outcome/behaviour based on some independent variables. There is a linear correlation between two or more variables where x-axis represent independent variable and y-axis represent target variable.

If there is only 1 input variable, then it is simple linear regression and if there are more than 1 input variable then it is multiple linear regression. LR model shows a sloped straight line describing the relation within the variables.



linear regression equation is below:

$$y = a + bx$$

Where a and b given by the formulas:

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

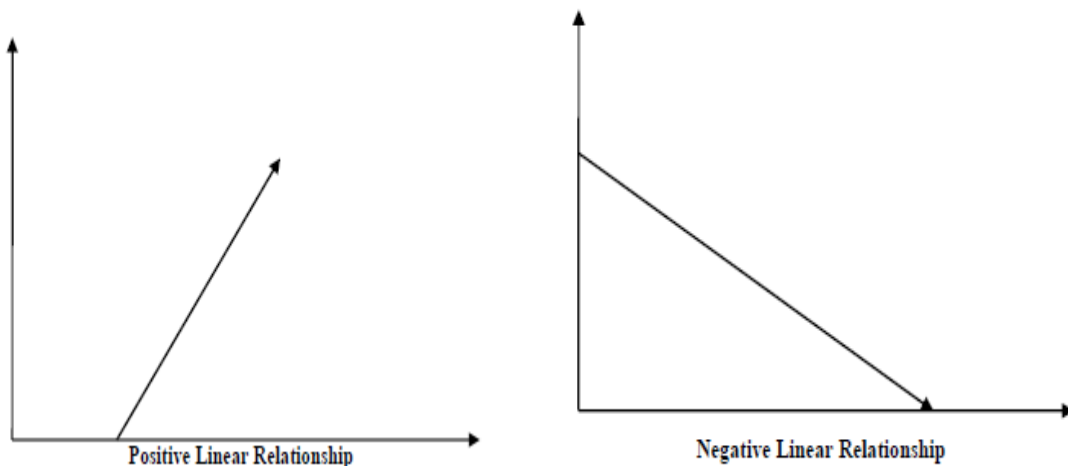
$$a(\text{intercept}) = \frac{n \sum y - b(\sum x)}{n}$$

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

A linear relationship is positive if both independent and dependent variable increases. And it is negative positive if independent increases and dependent variable decreases.



Below are some assumptions related to Linear Regression model and their validation steps –

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: Validate with histogram plot that Error term is following normal distribution curve.
3. No overfit: Check 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: Validate with residual plot

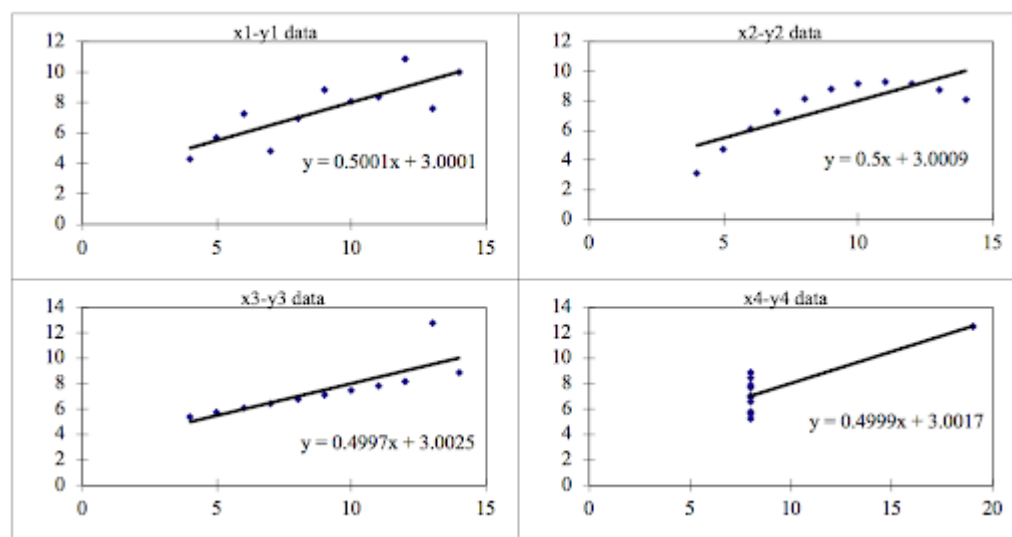
2. Explain the Anscombe's quartet in detail.

Ans: Statistician Francis Anscombe constructed this quartet in 1973 to show the importance of plot a graph data before analysing and the effect of outlier in stats properties.

Anscombe's quartet consists of 4 datasets which are almost identical and simple stats properties but looks different when plotted on graph. There are 11 points in each dataset.

Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

These datasets summary is almost matching but when we plot these it looks like below



We can describe these datasets graphs as:

Data Set 1: fits in linear regression.

Data Set 2: cannot fit in linear regression because the data points are non-linearly plotted.

Data Set 3: outliers exist in the data set, which can't be handled.

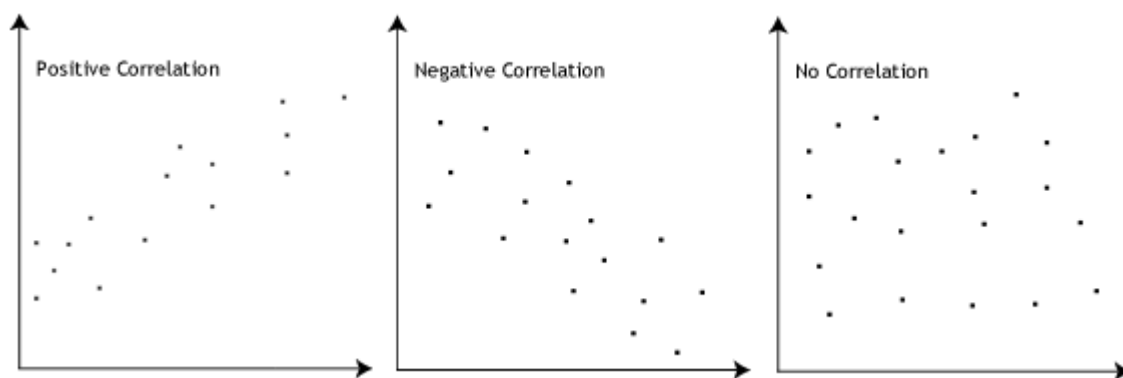
Data Set 4: more outliers in the data set, which also can't be handled.

As we can see, this quartet gives us understanding related to data visualize importance and how easily regression algo can be erroneous. So, it tells us we need to visualize data to create a well fit model before interpreting and model the data.

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).

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

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

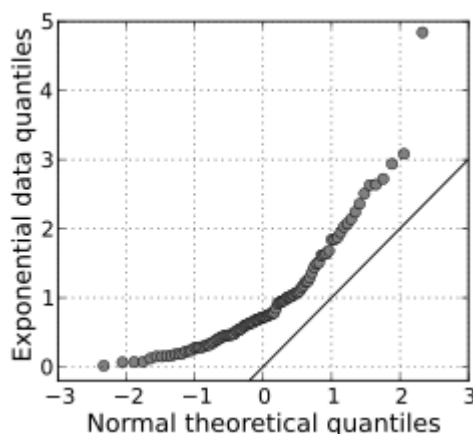
$$\text{VIF} = 1/(1-R^2) \text{ if VIF is infinite, it means that } R^2=1$$

And $R^2=1$ 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 variables 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.