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?

There are mainly four categorical variables in the data provided namely season, month, weekday and weather situation.

We can clearly observe that as soon as there is rain (weather situation) there is a huge drop in the demand for shared bikes.

We have no demand or zero demand for shared bikes when there is heavy rain or snow.

The month variable shows the fact that there is a drop in demand from November to February which is when there is heavy snow.

We can also observe from the season variable that there is a relative less demand for shared bikes in the spring season, as it is the season when there is heavy rainfall.

The day of the week have a huge impact on the dependent variable as there visually not much difference in demand for shared bikes on different days.

(All the above observations can be confirmed from plots in the Jupiter notebook submitted)

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

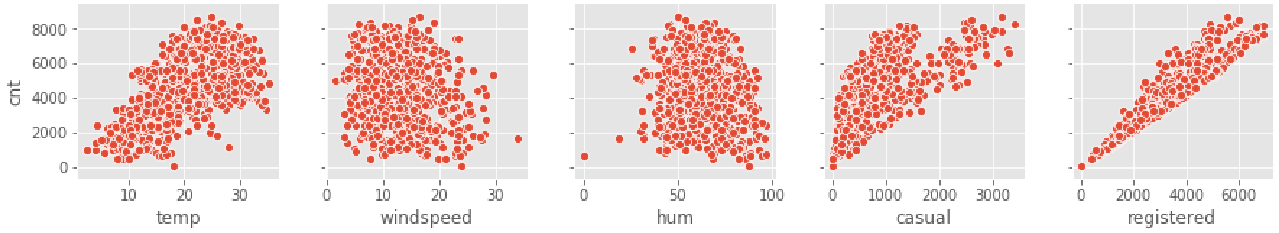
It important to use drop_first=True during dummy variable creation, as rest of the levels that are not dropped can completely explain the level that is dropped, this reduces the multicollinearity present in the data.

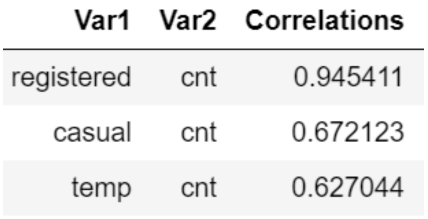
Example:

In the provided data there is a column named season where we have 4 levels namely Spring, Summer, Fall and Winter. If drop_first drops the spring level all the other variables can explain the level Spring as when all the other levels (Summer, Fall and Winter) are zero it would mean that it is Spring season.

This would reduce the multicollinearity in the data.

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





Var1	Var2	Correlations
registered	cnt	0.945411
casual	cnt	0.672123
temp	cnt	0.627044

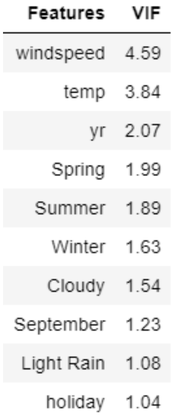
We can clearly see that registered and cnt have the highest correlation, this is because cnt is the sum of the columns registered and casual and therefore can expect it to be highly correlated. Since our target variable must be independent, we drop registered and casual.

So, temp has 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?

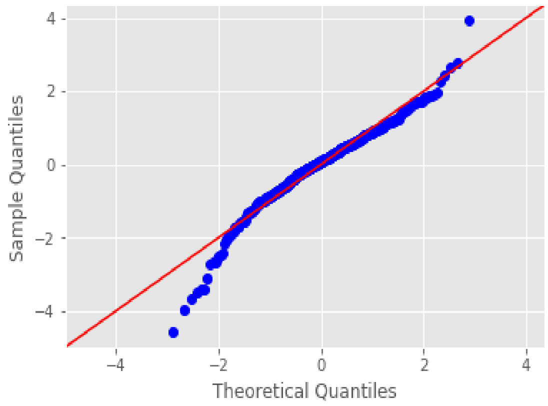
To validate the assumptions of Linear Regression after building the model on the training set:

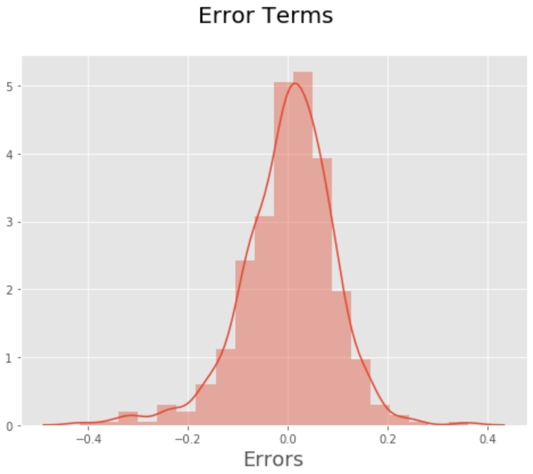
- Checked the VIF or Variance Inflation Factor to insure there is no multicollinearity in the model.



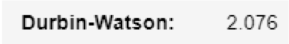
Features	VIF
windspeed	4.59
temp	3.84
yr	2.07
Spring	1.99
Summer	1.89
Winter	1.63
Cloudy	1.54
September	1.23
Light Rain	1.08
holiday	1.04

- Histogram and Q-Q plot to validate the assumption that the residuals are normally distributed.





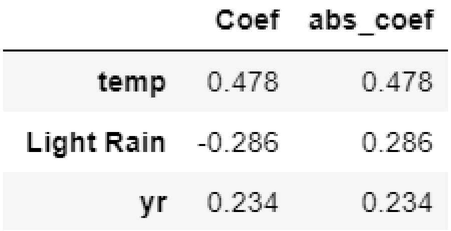
- Autocorrelation is validated with the help of the Durbin-Watson test, which in are case is close to 2.



Durbin-Watson:	2.076

- Linear Relationship is validated by plotting pair plot.

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



	Coef	abs_coef
temp	0.478	0.478
Light Rain	-0.286	0.286
yr	0.234	0.234

- Temp variable represents the temperature on a particular day in Celsius. As the coefficient is positive, we can say that as temperature rises the demand of shared bikes increases. This also tells us that demand for shared bikes will be more in warmer conditions.
- Light Rain variable represents one of the levels in the weather situation variable in the original data. Even initially when performing EDA, we found that Rain and Snow have a large impact on the demand for shared bikes. As the coefficient is negative, we can say that as the amount of rain increases the demand for shared bikes decreases.
- Here yr refers to the year.

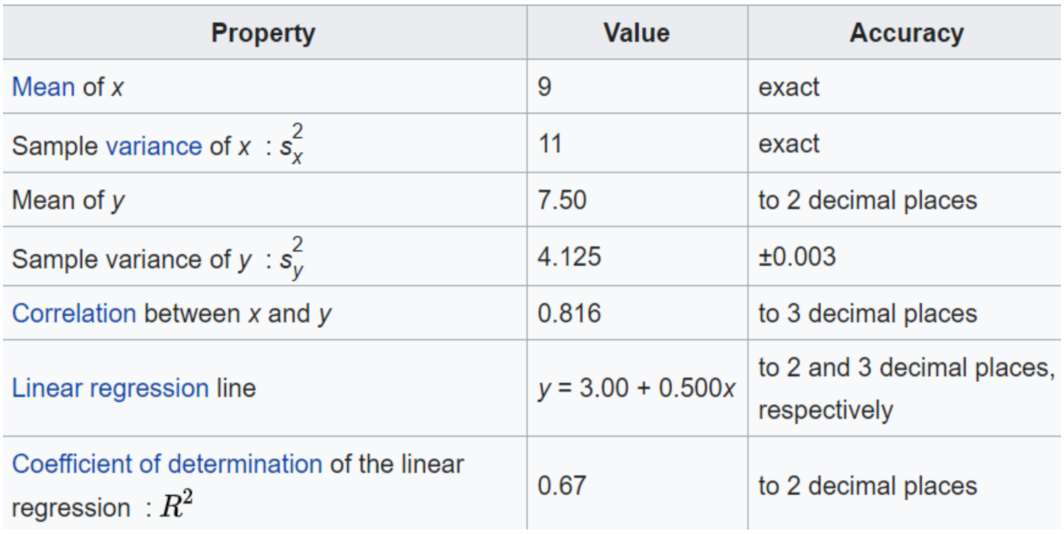
General Subjective Questions

1. Explain the linear regression algorithm in detail.

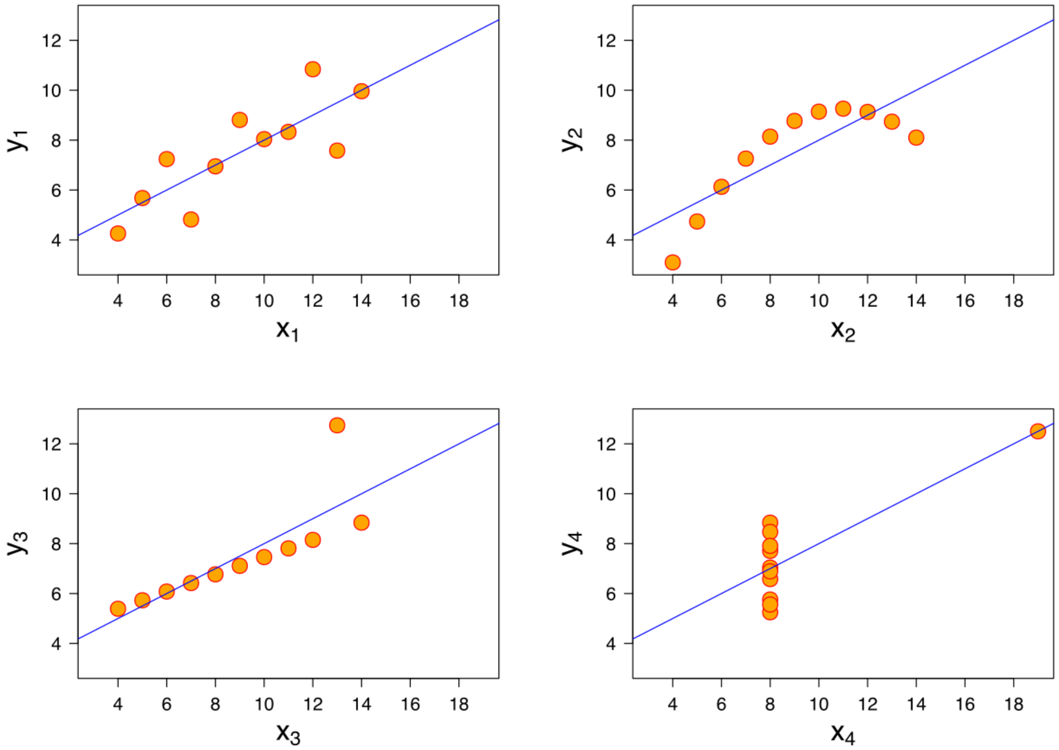
- Linear regression is one of the methods in Supervised learning, we have previous data with the right labels (or output) which can be used to train the model.
- It takes specific set of inputs and gives a single numerical output.
- It is generally used to solve problems related to prediction.
- This method is mainly based on the statical concept of regression.
- Here we follow a linear approach, we make use of the equation of a line, i.e., mx + c.
  Here 'm' is the slope of the line, 'x' is the independent variable and 'c' is the intercept of the line.
- We find the line of best fit to predict the values in the future.
- The line of best fit is found by minimising the RSS or the Residual Sum of Squares, which is by reducing the difference between the actual and the predicted of the dependent variable.
- To minimise the value of RSS we use a method called Ordinary Least Squares, this means that given a regression line in the data we calculate the distance of each point from the regression line and square it, and sum all of these together, this is the quantity that the OLS or Ordinary Least Squares method aims to reduce. This is one of the techniques used to estimate the values of the coefficients.
- There is one more method called the Gradient Descent which is an iterative approach to reduce the error of the model. It initially starts off with a random value for each coefficient and finds the RSS for the given line and then changes the value of the coefficient to reduce the RSS. This process is repeated till the least RSS value is achieved.
- Since we now have the optimal values of the coefficients, we then proceed to make predictions by solving a simple linear equation of multiplying the independent variables with their respective coefficient values to obtain the value of the dependent variable.

2. Explain the Anscombe's quartet in detail.

- As the name suggests 'quartet' which means a set of four people or things, Anscombe's quartet is a group of four data sets that have nearly identical simple descriptive statistics.
- This demonstrates the importance of data visualization before analysis or deriving conclusions only based on the statical description and the effect of outliers in the data.



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places



- We can clearly see that x1 represents a simple linear relationship.
- In x2 we can see that there is no linear relationship as well as the values are not normally distributed.
- In x3 we can see that even though the data is perfectly linear the regression line is slightly pushed towards the one point that is the outlier. This shows how one outlier in the data can affect the correlation coefficient.
- In x4 even though there is no linear relationship shown by the data, one high-leverage point has enough influence to result in a higher correlation coefficient.

3. What is Pearson’s R?

PCC or Pearson’s Correlation Coefficient also referred to as the Pearson’s R, is the correlation coefficient that represents the direction and the strength of the linear relationship between two variables.

It is found by dividing the covariance of the variables by the product of their standard deviations.

$$r_{xy} = \frac{S_{xy}}{S_x \quad S_y} = \frac{\sum (xi - \bar{x})\ (yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right) \left(\sum (yi - \bar{y})^2\right.}}$$

Where, S_x , S_y → Sample Standard Deviation

S_{xy} → Sample Covariance

$\bar{X}, \bar{Y}$ → SampleMean

The value of the PCC ranges between -1 and 1.

1 represents perfectively positive linear relationship.

-1 represents perfectively negative linear relationship.

Values closer to 0 represent that there is neither positive nor negative linear relationship.

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

- Scaling refers to the process of bringing continuous variables to a common scale so that it becomes easer to compare different variables of different ranges.
- Scaling is usually performed on continuous variables that tent to have different ranges and it becomes difficult to compare those variables against each other which in turn leads to selection of variables that may not have influence on the dependent variable.
- There are mainly two ways of scaling data those are normalized scaling and standardized scaling.
- In Standardized scaling the variables are scaled in such a way that their mean is zero and their standard deviation is one.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation }(x)}$$

- In normalized scaling the variables scaled in such a way that their values lie between zero and one. It is also referred to as Min-Max Scaling.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

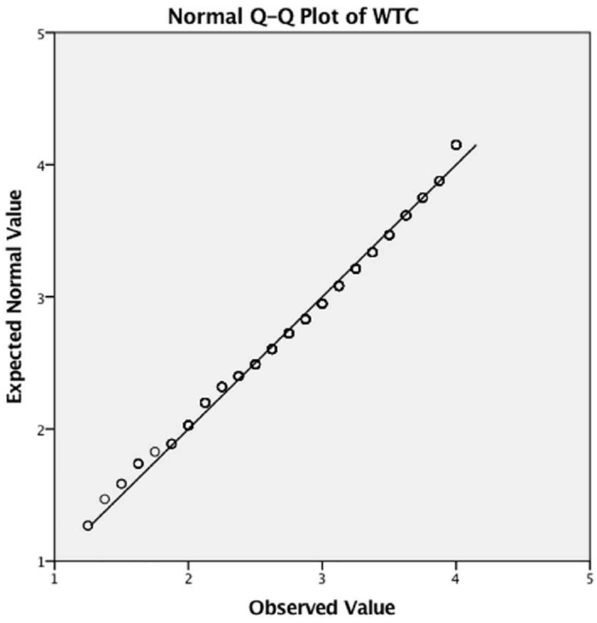
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

When there are variables that are perfectly correlated with each other the VIF value becomes infinite.

In other words when the correlation of two variables turn out to be either 1 or -1 the VIF value becomes infinite.

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

- Q-Q plot or Quantile – Quantile plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- The Q-Q plot forms a roughly straight line if the two distributions being compared are similar.
- It provides a graphical view of how graphical properties like skewness etc are similar or different in the distributions being compared.
- In linear regression the Q-Q plot can be used to check if the residuals are normally distributed by comparing the distribution of the residuals with the normal distribution.
- Following is an example of a Q-Q plot:


Normal Q-Q Plot of WTC
Expected Normal Value
Observed Value
1
2
3
4
5
1
2
3
4
5