Assignment Subjective questions

Assignment-based Subjective Questions

Ques1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: I have used Box plot on the categorical variables to carry out the analysis, below are the observation:

- Bike sharing is high for Fall season, when compared with the other season
- Number of bike sharing is less for holidays.
- Bike sharing is more for Clear, Few clouds, Partly cloudy, Partly cloudy, which seems logical for doing cycling.
- For 2019 the bike sharing has increased when compared to 2018 which is good in terms of business perspective.
- Sharing is almost same whether there is a working day or not.
- Midyear months have high number of bike sharing number.

Ques2: Why is it important to use drop_first=True during dummy variable creation?

Answer: drop_first=True is important to use as it reduces one extra column created, which reduces the correlation among the dummy variables. Like for example for coin flip there are two possibility head or tail, we could create one dummy variable "Head" which will hold value 1 if it is head or 0 when it is not head implying it is tail. So it is better to use n-1 dummy column's when having n categorical columns as it avoids Multicollinearity among dummy variables.

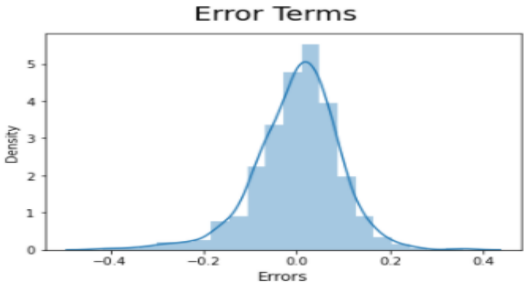
Ques3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: temp and atemp have highest correlation with the target variable i.e. cnt.

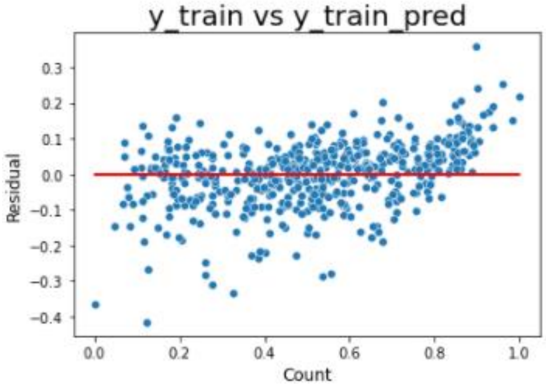
Ques4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: To validate the assumptions of LR, I used multiple checks:

1. **Normality of error terms**: Error terms should be normally distributed



2. **Multicollinearity checks:** There should be insignificant multicollinearity among variables
3. **Linear Relation:** Variable should have linear relationship
4. **Homoscedasticity:** No visible pattern among residual.



5. **Independence of residual:** No autocorrelation among residual points

Ques5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer: The top 3 Features are:

- Temp with coefficient 0.4915
- Yr with coefficient 0.2335
- Light Snow/Light Rain/ Thunderstorm/ Scattered clouds/ Light Rain/ Scattered clouds with coefficient 0.2852

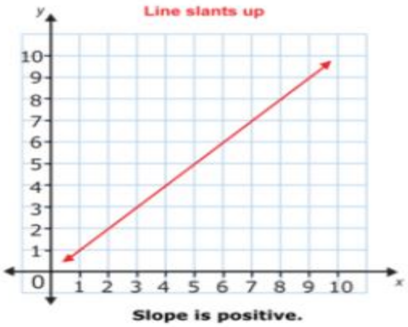
General Subjective Questions

Ques1: Explain the linear regression algorithm in detail.

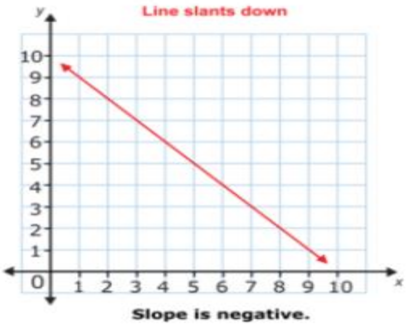
Answer: Linear regression is a method to find the best straight line on the given data, which means it establishes linear relationship between dependent and independent variables. The basic formula to represent linear relationship is “Y=mX+C” ,

where Y is dependent variable ,X is independent variable, m is the slope and C is Y intercept for X=0.

Linear relationship could be negative or positive in nature, i.e one variable increases other could decrease or increase.



Slope is positive.



Slope is negative.

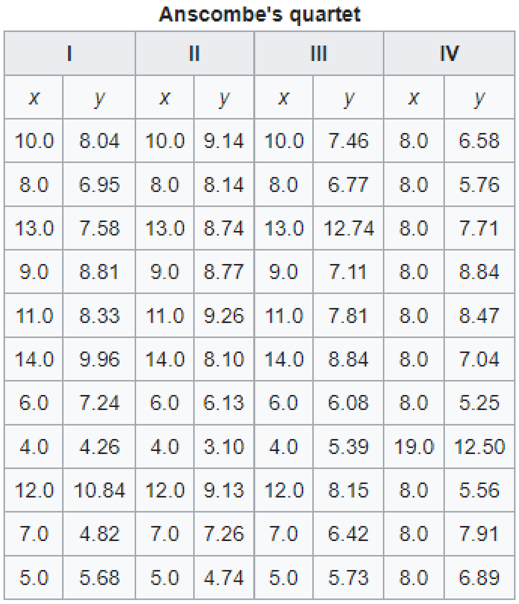
Regression is divided into 2 categories:

1. Simple linear regression: When dependent variable is predicted using only one independent variable. Equation for this is Y=m X+c

2. Multi linear regression: when dependent variable is predicted using more than one independent variable. Equation for this is Y=m1 X1+m2 X2+…….+mn Xn+c.

Ques2: Explain the Anscombe's quartet in detail.

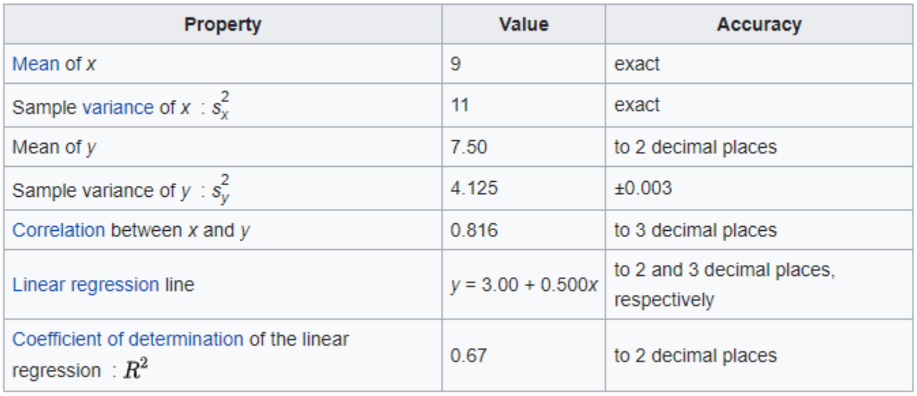
Answer: Anscombe's quartet was constructed by statistician Francis Anscombe in 1973 to demonstrate importance of graphing data when analyzing it and the effect of outliners and other observations on stats properties. It consist of 4 data set that have identical simple descriptive stats but still have different distributions so they appear very different when they are plotted on the graph.



Anscombe's quartet

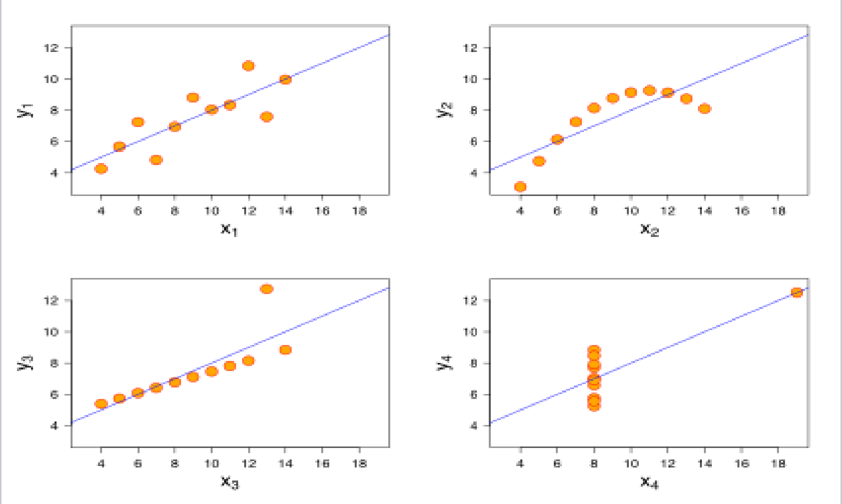
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	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

Data Property:



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

When the points are plotted on the graph. Below graphs are created.



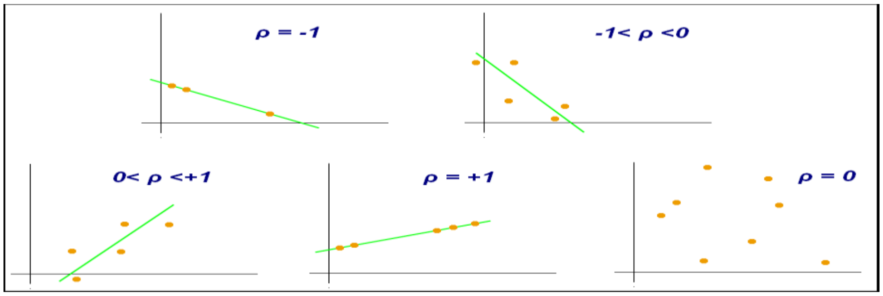
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

- The first scatter plot in top left appears to be in a simple linear relationship, corresponding to two variables correlated where y is modelled as gaussian with mean linearly dependent on x.
- The second graph in top right while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph in bottom left is a modelled relationship and is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph in the bottom right, shows an e.g. when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Ques3: What is Pearson's R?

Answer: Pearsons's correlation coefficient is covariance of two variables divided by the product of their standard deviation. If variables tend to go up and down together, the correlation coefficient will be positive. If variables tend to go up and down in opposite direction i.e if one increases and other decreases or vice versa ,then the correlation will be negative.

The Pearson's R could take value from -1 to 1. Value of 0 will imply there is no relationship between two variables. 1 if correlation is positive and -1 if the correlation is negative. Below is graph for p value.



Ques4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling is the method to pre process the data, which is used on the variables to bring it into a normalize range. Also it helps in speeding up calculation.

Most of the times the data collected have features that are highly varying in magnitudes, range and units. If the scaling is not done then the algo would take only magnitude in account and not units which would result in incorrect modeling. In order to avoid this issue, we have to perform scaling to bring all variable in same range. Scaling only affects the coefficient and not any other parameters like t-statistic, F-statistics, p- value etc.

Sr no.	Normalized scaling	Standardized scaling
1	Min and max value of features are used for scaling	Mean and standard deviation is used for scaling
2	Scikit-learn provides transformer called MinMaxScaler for normalization	Scikit-learn provides transformer called StandardScaler for standardization
3	It is used when features are of different scales.	It is used when we want Zero mean and unit standard deviation.
4	The values lies between 0 to 1 or -1 to 1.	It is not bound by any range.
5	It is affected by outliers.	It is less affected by outliers.

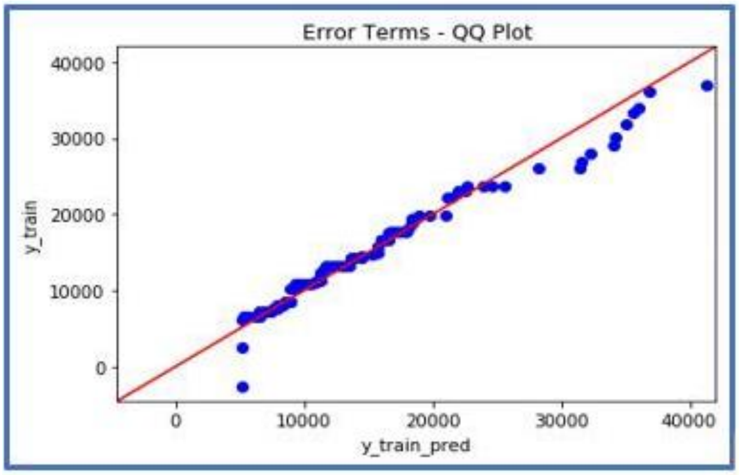
Ques5 : You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: VIF=infinite means that there is perfect correlation between two independent variable. When there is perfect correlation we get R2 as 1 which lead to 1/(1-R2) as infinite. In order to solve this we would have to drop the variable which is causing this perfect multicollinearity. An infinite VIF indicates that the corresponding variables may be expressed exactly by linear combination of other variable, which show infinite VIF as well.

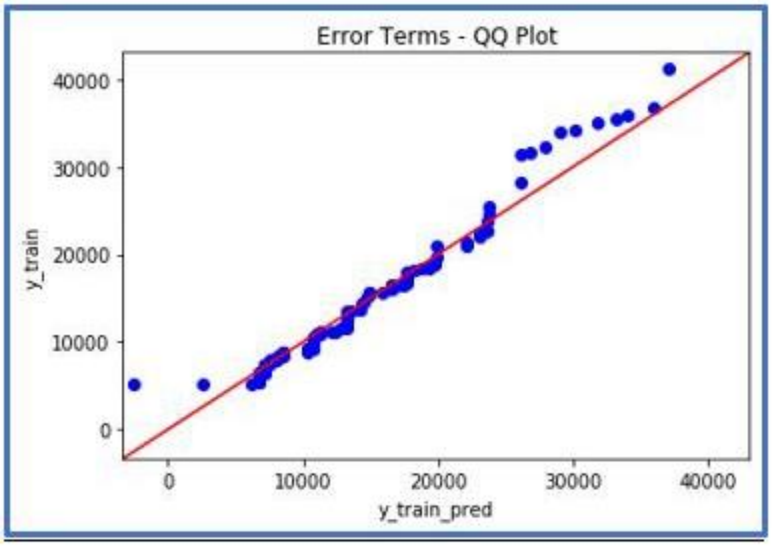
Ques6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: The Quantile-Quantile (Q-Q) plot, is graphical technique which helps us assess if a data set plausibly came from some theoretical distribution such as normal, uniform or exponential distribution. Also, it helps to determine if two data set from population with normal distribution. This helps in scenario of linear regression when we have received train and test data separately, we could use Q-Q plot to confirm the data sets are from populations with same distributions.

Advantages: Many distributional aspects like common location and scale, same distribution shape, population with common distribution, have same tail behavior or detecting outliners can all be detected using the plot. It could be used with sample size also.



- Same distribution: If all points lie’s on or close to line at angle of 45 degree from x- axis.
- Y- values < X-values: If y quantiles are lower than x.
- Y- values > X-values: If x quantiles are lower than y.
- Different distribution: if all points lie away from straight line at angle of 45 degree from x- axis.


Error Terms - QQ Plot
y_train
y_train_pred