Assignment-based Subjective Questions

- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
    - Based on my analysis Fall season has more customer booking. The lowest booking is in the season spring.
    - Year 2019 (Year:1) has more bookings than year 0 (2018)
    - Average number of users increases till July. September shows the maximum number of users. After september, the average and maximum number of users goes down
    - When there is a holiday, demand has increased
    - More users avail the services on a Good weather day (Clear weather)
    - Booking is almost same on a working day or on a non-working day
- Why is it important to use drop_first=True during dummy variable creation? (2 mark)
    - Number of dummy variables needed to represent 'n' states are n-1. Dropping a variable helps to reduce the correlation between created variables.
- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - 'temp' variable has the highest correlation with the target variable.
- How did you validate the assumptions of Linear Regression after building the model on the training set?
    - Residual analysis helps to confirm if the assumptions of error terms are correct or not.
    - Plotting distplot shows, a normal distribution where mean is at zero which confirms the assumptions
    - Check for multi collinearity to confirm if independent variables are correlated or not
- Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - Temp
    - weathersit
    - year

General Subjective Questions

- Explain the linear regression algorithm in detail

  Linear regression is a statistical model which analyzes the linear relationship between dependednt and a given set of dependant variables. With a linear relationship, when the value of independant variable changes, the value of dependant variable also changes accordingly (either increase or decrese). A linear relation ship may be positve od negative in nature. Linear relation ship can be defined by the equation y = mX + c where m is the slope and C is a constant, y is the dependant variable and X is the independant variable.

  Linear regression can be single linear regression or multi linear regression.
  Assumptions in linear regressions are
    - There exist a linear relationship exist between independant variables and dependant variables
    - Error terms are normally distributed
    - Error terms are idnependant of each other

- Error terms have constant variance

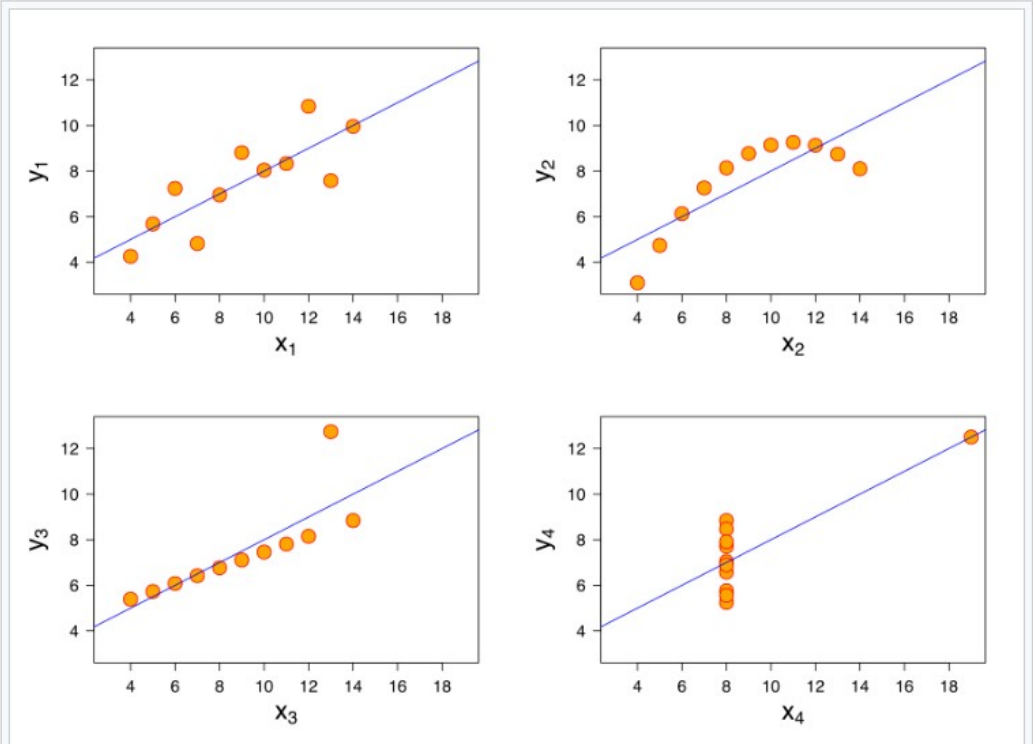
There is no assumption of individual vales, instead the assumptions are on the relationships.

Single linear assumptions are applicable for Multi linear regression also. There are few more considerations for for MLR

- Overfitting
- Multicollinearity
- Feature selection is an important aspect

- Explain the Anscombe's quartet in detail.
  - Anscombe's Quartet was developed by statistician Francis Anscombe. It tells us the importance of visualizing the data before applying various algorithms to buld the model. It consists of four data sets each containaing eleven (x,y) pairs. These data sets seems to have idneitfical statistics but they have very different distribution while plotted.

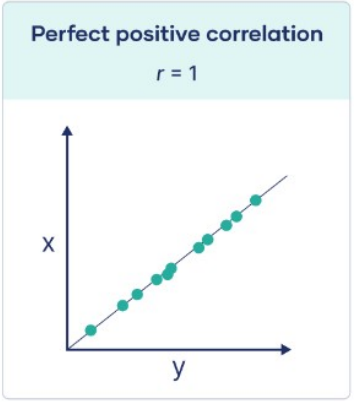
    The below graphs shows the graph plotted for all 4 sets and their data sets are given below. Even though all these sets have simlar statistical description, they shows a different pattern while plotted.

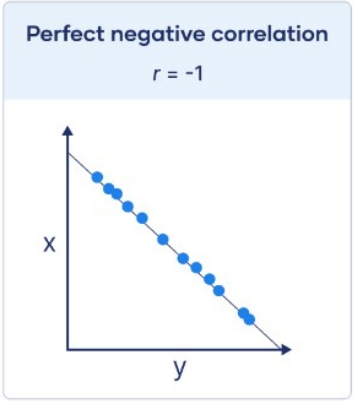


Anscombe's quartet

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

- What is Pearson's R?
  - Pearson Correlation Coefficient (r) is the most common way of measuring a linear correlation. Its a number between -1 and 1 which shows the strength of relationship between variables. A value of zero indicates no relation between the variables and value less than zero indicates a negative correlation and greater than zero indicates a positive correlation.





- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
  - Feature scaling is a technique to standardise the independant variables. Scaling makes the variables at a comparable level and there by their coefficients also would be comparable. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
  - Two common techniques used for scaling are normalised scaling (Min-Max Scaling) and standardised scaling. Normalization tries to keep the variables in a range b/w 0 and 1 but standardisation is not bound to any particular range. Normalization is done by (x-xmin)/((xmax – xmin) wheres as standardisation is done by (x-u)/sigma

- You might have observed that sometimes the value of VIF is infinite. Why does this happen?
  - Infinite value of VIF shows a perfect correlation (multicollinearity) between two independant variables. In this case we get R-squared = 1 which leards to 1/(1-R2) as infinity. Dropping one of the variable helps to address this

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

  The quantile-quantile (Q-Q) plot is a graphical technique to determine if a data set came from some theoretical distribution such as normal or exponential.

  Its used to plot quantiles of the first data set against the quantiles of the second data set. If both sets of quantiles came from same distribution we should see the points forming a line that’s roughly straight.

  Importance of Q-Q plot: Q-Q plot helps to know if the assumptions of a common distribution is justified. If they are from a common distribution, then scale estimators can pool both data sets to obtain estimates of the common scale. If they differe, it is useful to get some understanding of the differences.