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?**

   categorical variables like winter season was positively correlated, but weathersit categorical variable mist, lightsnow is negatively correlated , and also season spring is negatively correlated, illustrated clearly by below equation

   cnt = 0.236 x yr - 0.086 x holiday + 0.437 x atemp - 0.132 x windspeed - 0.278 x LightSnow - 0.077 x Mist - 0.125 x spring + 0.039 x winter

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

   While dummy variable creation, n categorical levels are created, but n-1 levels are sufficient to explain n levels
   drop_first=True is used to get (n-1) levels out of n categorical levels by removing the first level. Thus it reduces correlations created among dummies and it allows to drop the first variable and identify it through all other columns being 0.

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

   'atemp' has the highest correlation with target variable 'cnt'

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

   Assumptions and their respective verification are as below
   1.Linear relationship :
   Verified by using scatter plot between target and independent variables
   2.No auto-correlation or independence:
   By checking the value for Durbin-Watson in the model built
   3.No Multicollinearity:
   By computing VIF and ensuring model built on VIF < 4
   4.Normal distribution of error terms
   By Residual analysis of error terms and drawing distribution plot by using difference of predicted target variable data from train set and actual target variable data from train set

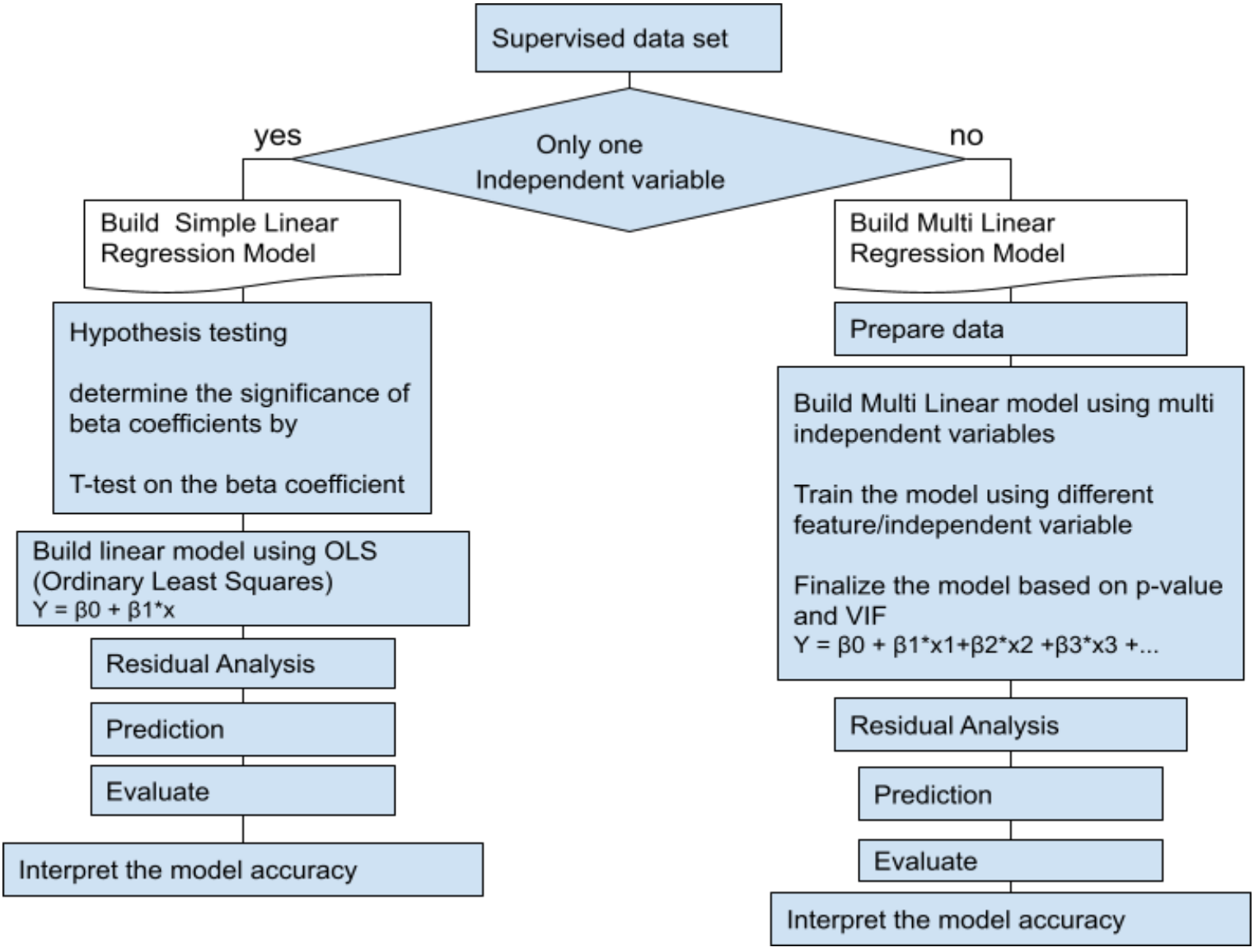
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

   Based on final model top 3 features contributing significantly are
   1. atemp
   2. yr
   3. winter

General Subjective Questions

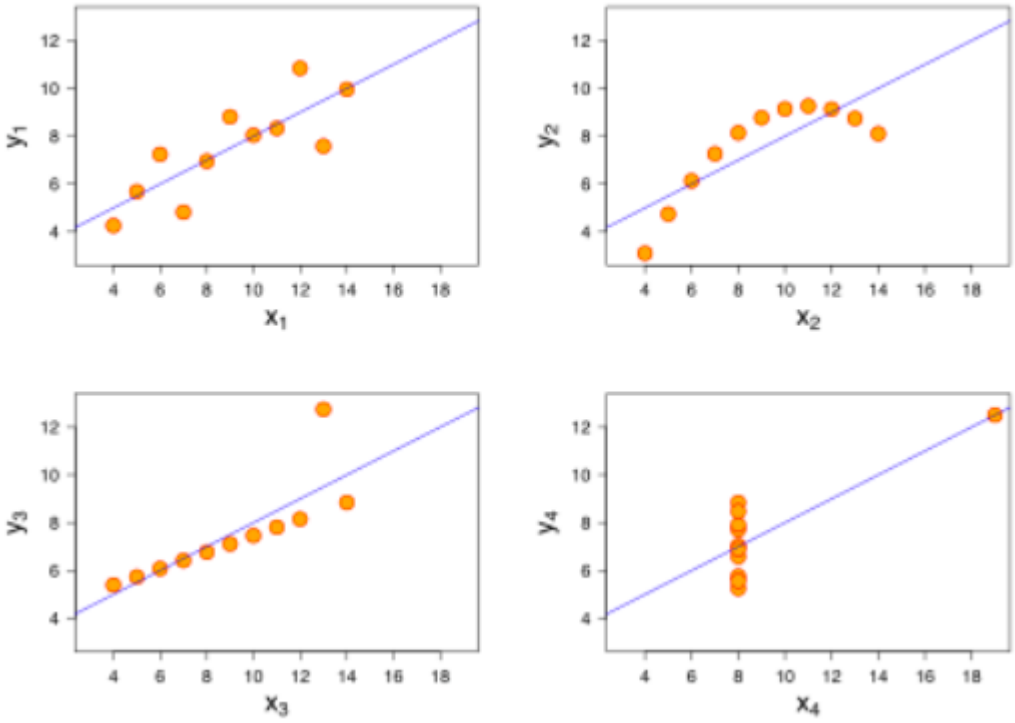
1. **Explain the linear regression algorithm in detail.**

   Linear regression is supervised learning.The output variable to be predicted is continuous in nature.Linear regression performs **the task to predict a dependent variable value (y) based on a given independent variable (x)**. which can be explained in more detail by following algorithm



2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet contains four data sets that have nearly identical statistics, but very different distributions and appear very different when graph is drawn.
This concept uses a dataset with eleven (*x*,*y*) points. Constructed to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. This enabled to counter the assumption among statisticians that "numerical calculations are exact, but graphs are rough.



pic credit:wikipedia

3. **What is Pearson's R?**

Pearson's R is a statistical measure of the strength of a linear relationship between two variables. It is typically represented by the symbol 'r'.
Pearson correlation coefficient can take on values from -1 to +1 a. it indicates whether one variable increases or decreases as the other variable increases or decreases. A Pearson correlation coefficient of 1 indicates a perfect positive (direct) linear relationship, while a Pearson correlation coefficient of -1 indicates a perfect negative (inverse) linear relationship. When Pearson's r is 0 there is no linear relationship between the two variables.

Scatter plot with x and y, can easily show the pearson values tendency

Formula used to calculate:

$r = \sum(x - \overline{x})(y - \overline{y}) / \sqrt{\sum(x - \overline{x})^2 \sum (y - \overline{y})^2}$

where $\overline{x}$ and $\overline{y}$ represent mean values for the respective x and y values.

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

Scaling is method used to normalize the range of independent variables or features of data, it also helps to run the algorithm faster

Scaling should be performed because collected data sets may have varying magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling, we have to do scaling to bring all the variables to the same level of magnitude

Normalized scaling versus Standardized scaling

Normalized scaling It brings all of the data in the range of 0 and 1.

Calculated by formula:

$X = (x - \min(x))/(\max(x) - \min(x))$

Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

Calculated by formula:

$X = (x - \text{mean}(x))/\text{sd}(x)$

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

If there is perfect correlation then VIF tends to be infinite.greater the VIF higher the multicollinearity
This can happen when

- perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity
- perfectly predicted by other variables in the model
- there are too many independent variables and many duplicate data sets

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

(Q-Q) plot is called Quantile-Quantile plot, this helps to compare the sample distribution of the variable at the hand against any other possible distributions visually

Q-Q plot enables visually analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x

It is used to check following scenarios which help in analysis of different distribution
If two data sets

a. come from populations with a common distribution
b. have common location and scale
c. have similar distributional shapes
d. have similar tail behavior