Submitted by : Priyanka Asthana

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? (3 marks)
   Categorical variables like month, season, weather and weekday have different outcome on count of bike riders. During months of May, Jun, July, Aug, Sep demand of bike goes high. Similarly demand is high during clear weather followed by Misty weather. On weekdays demand is bike riders is high. Similarly demand is higher in fall and summer season

2. Why is it important to use **drop first=True** during dummy variable creation? (2 mark)
   It is important to use drop first=True as it drops the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. In case of Housing dataset, if one variable is furnished and semi_furnished, then the third one is obviously unfurnished. So we do not need 3rd variable to identify the unfurnished. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   temp(0.63) and yr(0.57) have the highest co-relation with target variable cnt respectively.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   Assumptions of Linear Regression are validated using the following criteria using the model created. So based on lr15
   1. Linear Relationship was shown for yr variable
   2. Homoscedasticity
   3. Absence of Multicollinearity
   4. Independence of residuals (absence of auto-correlation)
   5. Normality of Errors

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   Temperature and summer season

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

The goals of Linear regression is to find out if there is any correlation between the variables and find the best fit line for the dataset. To implement linear regression we need to follow some steps

Step 1: Reading and Understanding the data

Step 2: Data Pre-processing and data preparation - Check for multicollinearity and visualize the data using boxplot

Step 3: Data Visualization

Step 4: Create dummy variables

Step 5: Rescaling the Features

Step 6: Splitting the Data into Training and Testing Sets

Step 7: Build the model using iterative approach. Watch for R-squared value, p-value, VIF. Drop the variables with high p-value and high VIF. Such variables are insignificant. Repeat till VIFs and p-values both are within an acceptable range. So that it can be used for prediction

Step 8: Residual Analysis of the train data

Step 9: Making Predictions Using the Final Model

Step 10: Model Evaluation

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**.There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x, y points in all four datasets.

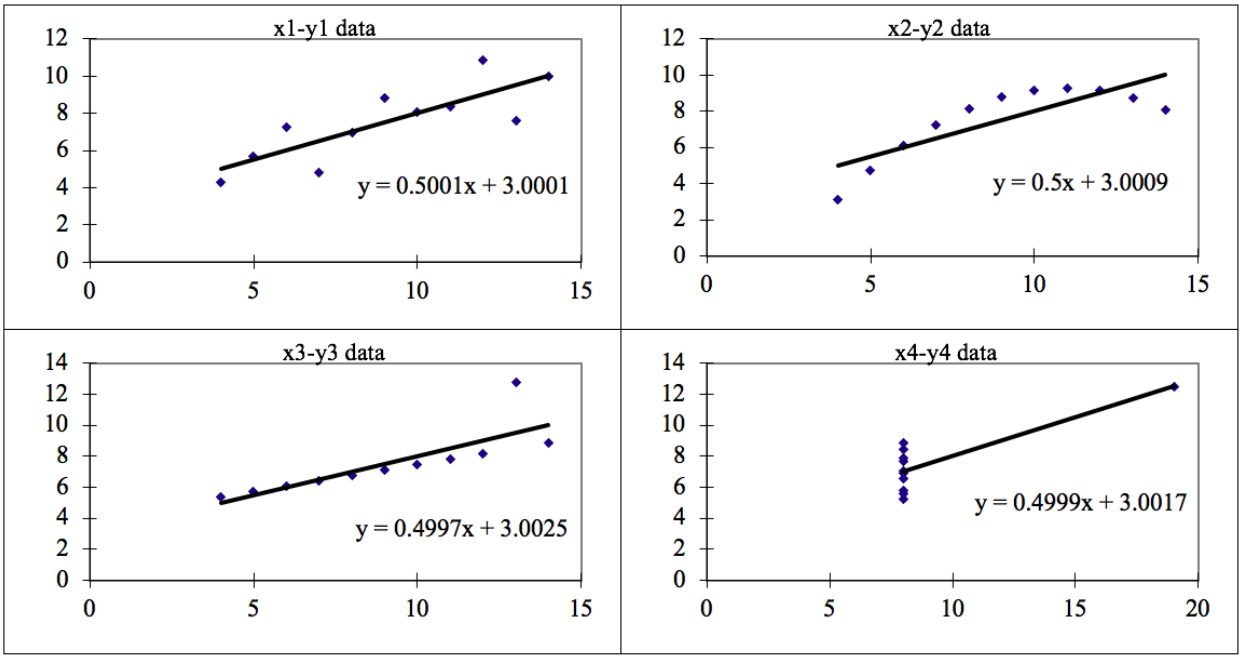
This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

The statistical information for all these four datasets are approximately similar and can be computed as follows:

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	

However when plotted



The four datasets can be described as:

Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)

   Pearson's correlation coefficient is a statistical measure of the strength of a linear relationship between paired data. In a sample it is denoted by r and is by design constrained as follows -1 <=r <= 1

   Furthermore:

   - Positive values denote positive linear correlation;
   - Negative values denote negative linear correlation;
   - A value of 0 denotes no linear correlation;
   - The closer the value is to 1 or –1, the stronger the linear correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Scaling is a way to reduce variance in dataset as it contain features highly varying in magnitudes, units and range. In scaling data is transformed to fit within a range.

   Normalization fits the data such that the resulting distribution lies within a shape.

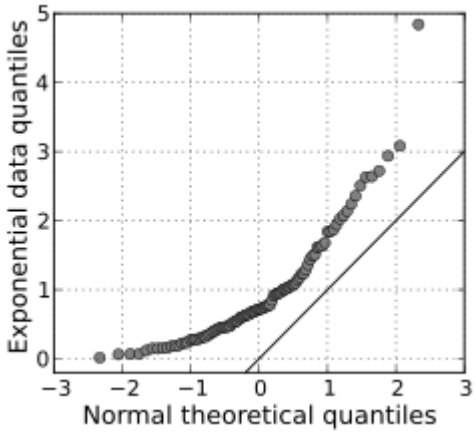
Standardization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where σ is the variance and $\bar{x}$ is the mean.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   High VIF shows a perfect correlation between two independent variables.
   In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
   One of the variables should be dropped from the dataset which is causing perfect multicollinearity.

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



(3 marks)