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)**
   - The categorical variables where Season, Weather sit, Month and weekday
   - Created dummies for these variables, and dropped the unnecessary
   - After final model and VIF, what I could infer is:
     1. Count of users is highly related to Winter season
     2. September month is the best month, followed by July for the users
     3. Sunday is the best day
     4. The weathers Spring, mist and light rain are the best

2. **Why is it important to use drop first=True during dummy variable creation? (2 mark)**
   - Drop first = true drops the sorts the variable and drops the first column after creating dummies.
   - As we don't need the first column after creating dummies because we can define the rest of the columns using following columns.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   - Temperature

4. **How did you validate the assumptions of Linear Regression after building the model on the Training set? (3 marks)**
   - Plotted histogram to check residuals
   - Used VIF to check multicollinearity
   - Dropped variables with high p-values and VIF
   - Used scatter plot to check to understand the spread of y_test and y_pred
   - Used qq plot

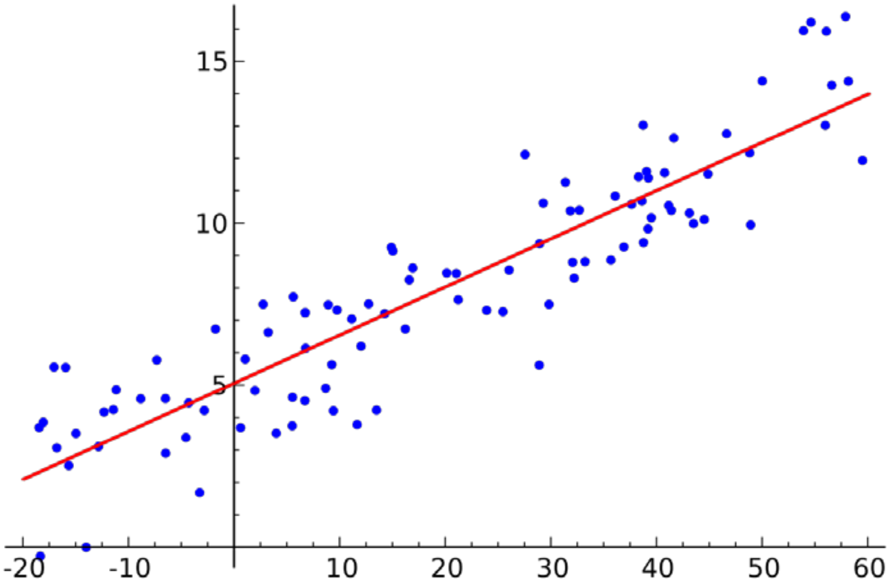
5. **Based on the final model, which are the top 3 features contributing significantly towards Explaining the demand of the shared bikes? (2 marks)**

   - Temperature(The lower the temperature the better)
   - Year(Demand increased from 2018)
   - Winter(As the temperature is low)

General Subjective Questions

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

- Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

  The line can be modelled based on the linear equation shown below**.**

  y = a_0 + a_1 * x

- The motive of the linear regression algorithm is to find the best values for a_0 and a_1.

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

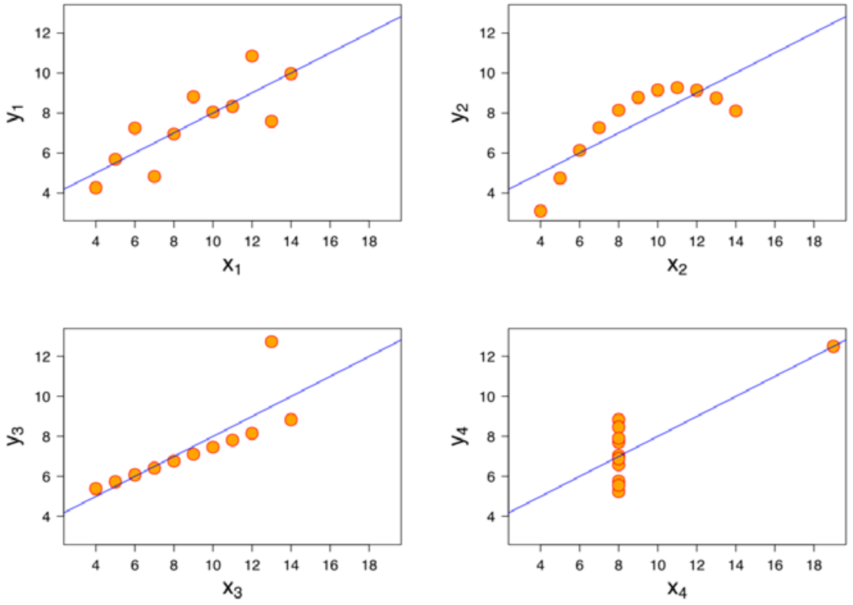
- Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely,**,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- The summary statistics show that the means and the variances were identical for x and y across the groups :

  1. Mean of x is 9 and mean of y is 7.50 for each dataset.
  2. Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
  3. The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

- When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



1. Dataset I appears to have clean and well-fitting linear models.
2. Dataset II is not distributed normally.
3. In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
4. Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

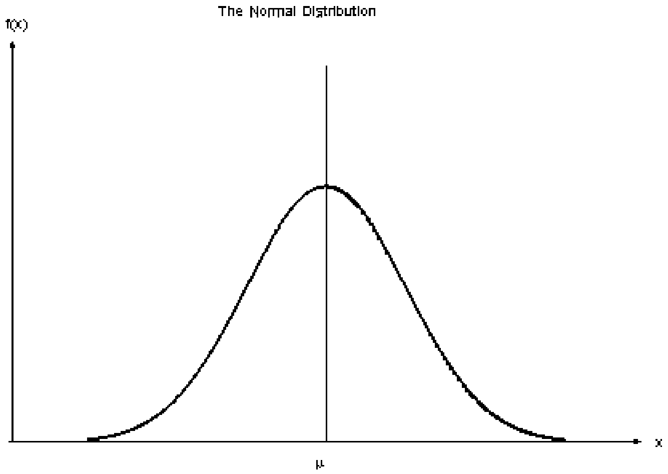
- This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

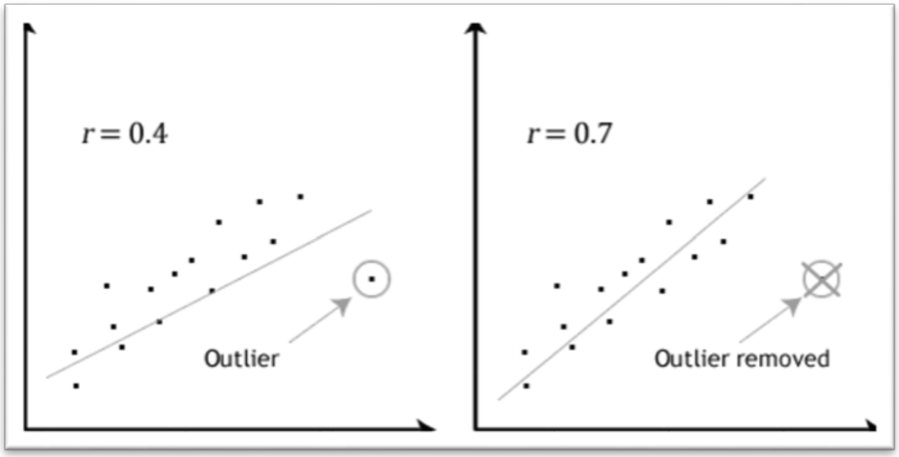
- Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

Assumptions

1. For the Pearson r correlation, both variables should be **normally distributed**. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'. A simple way to do this is to determine the normality of each variable separately using the Shapiro-Wilk Test.

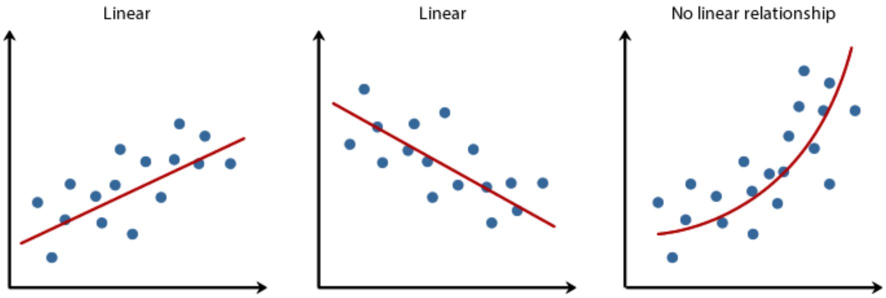


2. There should be **no significant outliers**. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.



3. Each variable should be **continuous** i.e. interval or ratios for example weight, time, height, age etc. If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

4. The two variables have a **linear relationship**. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric.



Copyright 2014. Laerd Statistics.

5. The observations are **paired observations.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable. For example if you're calculating the correlation between age and weight. If there are 12 observations of weight, you should have 12 observations of age. **I.e. no blanks.**

6. **Homoscedascity**. I've saved best for last. The hard is hard to pronounce but the concept is simple. Homoscedascity simply refers to '**equal variances**'. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the

circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.

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

- This means that you're transforming your data so that it fits within a specific scale, like 0–100 or 0–1. You want to scale data when you're using methods based on measures of how far apart data points, like support vector machines or KNN With these algorithms, a change of "1" in any numeric feature is given the same importance.
- For example, you might be looking at the prices of some products in both Yen and US Dollars. One US Dollar is worth about 100 Yen, but if you don't scale your prices methods like SVM or KNN will consider a difference in price of 1 Yen as important as a difference of 1 US Dollar! This clearly doesn't fit with our intuitions of the world. With currency, you can convert between currencies. But what about if you're looking at something like height and weight? It's not entirely clear how many pounds should equal one inch (or how many kilograms should equal one meter).
- By scaling your variables, you can help compare different variables on equal footing.

What is Normalization?

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

What is Standardization?

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

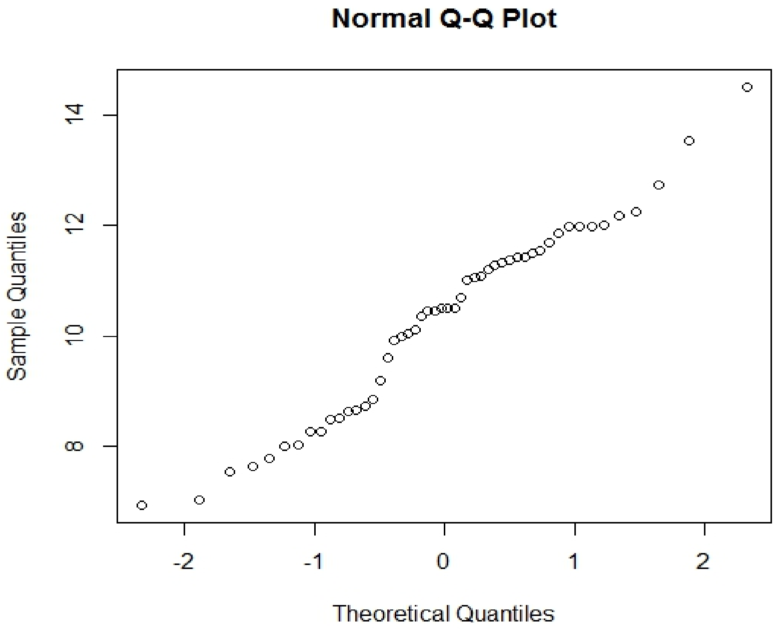
Here’s the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)**
   - An infinite VIF value indicates that the corresponding variable may be expressed exactly by a liner combination of other variables.
   - Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictor’s approaches unity, the corresponding VIF becomes infinite.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**(3 marks)
   - The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.
   - A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2