# 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?

   There are some categorical affecting dependent variables, which are mnth, weathersit and season. On the other hand, holiday, weekday, workingday don’t have any affect on dependent variable.

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

   It took me sometimes to investigate why my VIF returned inf. It turned out I didn’t use drop_first = True. VIF return inf (infinity) because strong multicollinearity was occurred.

   Multicollinearity is a term used in data analytics that describes the occurrence of two exploratory variables in a linear regression model that is found to be correlated through adequate analysis and a predetermined degree of accuracy.

   In another words, two independent varibles are strongly correlated. If all two variables are used to predict the dependent variable, it causes multicollinearity.

   In this assignment, for example, variable weathersit has only 3 cardinalities. If all 3 cardinalities are transformed to variables, either at least one or all of them are correlated with each other. Therefore, using drop_first = True are essential to eliminate multicollinearity.

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

   Following variables have the highest correlation:

   - dteday (after adjusting from text to numeric)
   - atemp
   - yr

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

   Following steps are done to validate the model:

   - Predicting the model on test set, which created prior the training.
   - Performing residual analysis on the test dataset to determine if the residual is normal distribution with peak at 0.
   - Calculating R-squared on test data to make sure the R-squared is similar.

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

   Because my final model has only 4 features, their contribution are in following order:

- atemp with slop 0.4
- yr with slop 0.2
- Sep with slop 0.05
- spring with slop -0.1

## General Subjective Questions

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

In mathematics, the short form linear regression can be simply wroten in this equation:

$$y = \theta + \alpha x$$

Where:

$y\ is\ outcome\ variables\ or\ depedent\ variable.$
$\theta\ is\ the\ intercept, the\ default\ value\ when\ x\ equals\ 0.$
$\alpha\ is\ the\ slope\ of\ the\ line.$
$x\ is\ the\ depedent\ variable.$

The general form of linear regression is

$$y = \theta + \sum_{1}^{n} \alpha_i x_i$$

In this equation, a change of $y$ is defined by the any changes of $xi$ given the slope of alpha i.

Linear regression is the most basic machine learning algorithm where we train a model to predict the outcome of continuous variable (depedent variable) of our data based on some other independent variables.

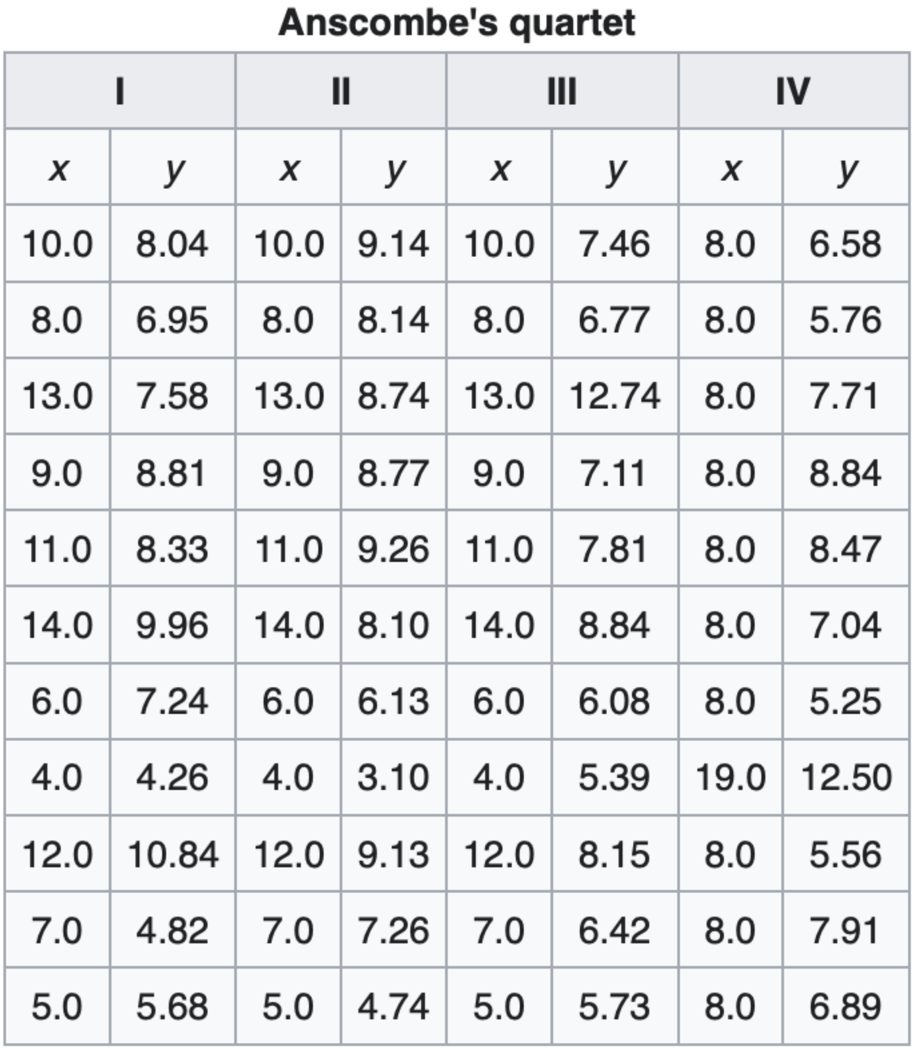
In the linear regression model assignment, we need to predict the demand of bike shares (unknown data) using data such as date, season, number of registered customers, other weather conditions (known data). Some changes in known data doesn’t reflect in the changes of other known data. That’s why we call them independent variables. For example, changes in season doesn’t result in changes in working day or holiday. In contrast, changes in independent variables (such as season, weather condition) may lead to a change in number of bike sharing demands.

### 2. Explain the Anscombe’s quartet in detail

Anscombe’s quartet are four datasets that almost identically equal in statisticals summary (such as mean, standard deviation, correlation, etc), yet differently visualise on graphs.

In another words, if we visualise them on charts, we will see completely different 4 charts, but when we examine statistical summary (mean, standard deviation, correlation), we have identical result.

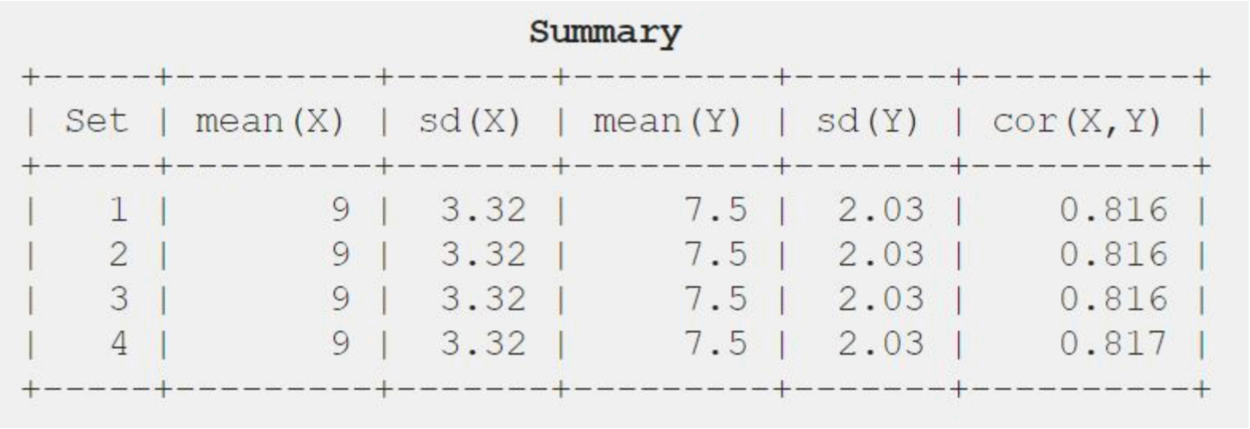
Following example is copied from wikipedia. The table below shows the example of Anscombe's quartet.



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 |

Here is the statistical summary.



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

As we can see, all datasets have idential mean, standard deviation, and correlation.

## 3. What is Pearson’s R

Pearson's R represents the linear correlation betweens variables. There are three possible scenario of correlation. Possitive correlation is when variables tend to go up and down together. Whereas if variables go in different direction it is negative correlation. Zero correlation means there is no linear correlate at all.

Pearson’s R correlation is an useful indicator to determine if a change in one variable correlates (but may or may not cause) to another variables. In another word, Pearson’s R correlation (or other correlation algorithm) doesn’t imply causation.

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

During steps of data processng, sometimes we need to scale independent variables. It normalised the data within a certain range. Besides, it’s usefl to boost calculations in a algorithm.

Scaling data is essentially important, because data collection sometimes contains features highly varying in weights, units and range. Scaling data helps to get all the variables to the same level of scale. If this is not done properly, algorithm only considers magnitude in account and not units so it leads to inapproriate result.

Normalisation transform all of the data in the range of 0 and 1, in which 0 equals the minimum values and 1 are the maximum values in the original data. On the other hand, standardisation replaces original values with their Z-score. The new dataset has mean zero and standard deviation one.

Imagine, the original dataset may varry a lot in range. The standardisation compresses it a bit, but it’s not bounded to a certain range. While the normalisation compresses it way better in range of 0 to 1. However, the normalisation shouldn’t be applied if there is outliers as a matter of fact, the extrem outliers are always 1.

Normalisation is usually used when we don’t know about the distribution while standardisation is useful for normal distribution.

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

If there is a perfect correlation amongs variables, the $R^2$ is equal 1. The VIF = $\frac{1}{1-R^2}$ so if $R^2$ = 1 the VIF returns infinity. In my asssignment, it happens when I didn’t drop first columns of months, wheathersit during dummy variables creation.

VIF is useful to detect multicollinearity, and without dropping first columns, strong multicollinearity happens. After dropping first columns during dummy variables creation, infinity VIF vanished.

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

Quantile-Quantile (Q-Q) plot, is useful to detect if a dataset is normal distribution. Besides, it also helps to determine if two dataset comes from same population with a common distribution.

In linear regression model, Q-Q plots is useful for model evaluation in term of residual analysis. Residual is the difference between actual values and predicted values. Given train and test dataset, if we can conclude that both set of residuals come from the same population, we are pretty sure that our model perform on test dataset as well as on train dataset.