**Project Implementation**

**3.1 An introduction to dataset:**

The data set contains demographic information about a set of towns in New York state. The response "MALE\_FEM" is the number of males in the town for every 100 females. The predictors are the percentage under the age of 18, the percentage between 18 and 65, and the percentage over 65 living in the town (all expressed in percents such as "57.0"), along with the town's total population.

The data was recorded in 1990 by US Census for New York city population and Demographics Which is mainly uses:

1. To group customers based on variables including age, gender etc.

By grouping consumers using demographic data, it allows businesses to understand each segment, what they want and how they want it. It will help the business to market each product differently based on the consumer groups they are targeting.

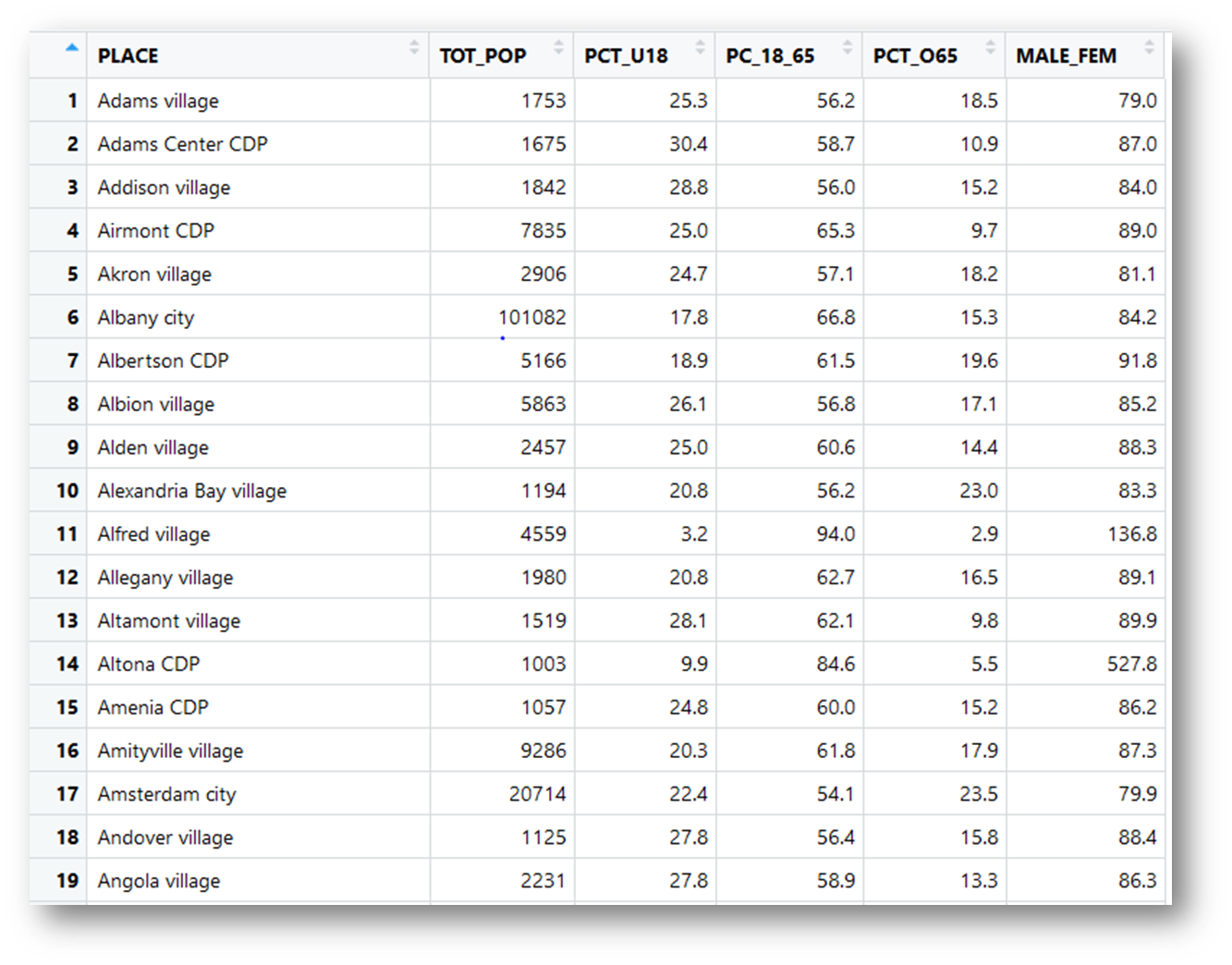
1. Businesses can use demographics to determine the next step of their growing business.

By understanding what their customers are looking for, they would be able to determine things such as; would customers like a new location and if so, how far would they be willing to go out of their way to get to there. They could also help to avoid making costly mistakes and this will help to push the business forward.

1. To customize products to specific consumer groups.

Demographic data can be used to customize products by finding out what is wanted by consumers, and changing the existing product to fit the desired outcome. The whole idea of customizing products would be to save time and money on products that aren’t wanted or that don’t fit your consumer specification.

Snippet of Dataset:



**Introduction to columns of Dataset:**

1. PLACE: This column contains the names of towns of New York city.

2. TOT\_POP: This column contains no. of people living in town.

3. PCT\_U18: this column contains the no. of people who are less than 18 years.

4. PC\_18\_65: this column contains the no. of people who has age between 18 to 65 years.

5.PCT\_O65: this column contain the no. of people who are older than 65 years.

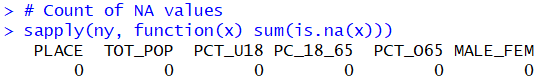
6. MALE\_FEM: This column contains the no. of males for every 100 females.

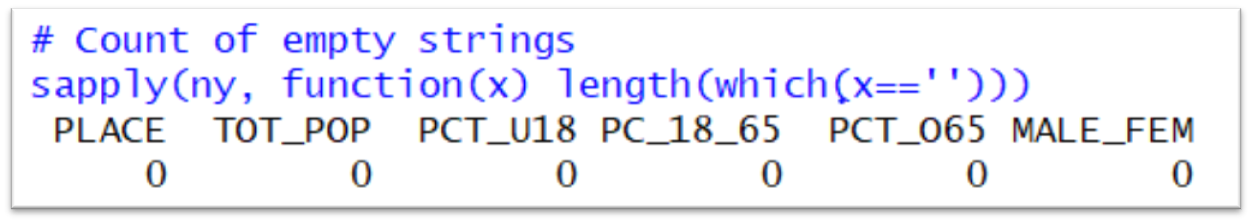
**3.2 Exploratory Data analysis:** Before start building a model let’s have a look towards data.

1) First we have to check for null values.

1. NULL values are missing values from the data which can decrease the performance of model, so it is important to remove NULL values,
2. To remove NULL values, we can

* drop those rows which contains NULL values.
* Replace them with either mean, median or mode, according to data.





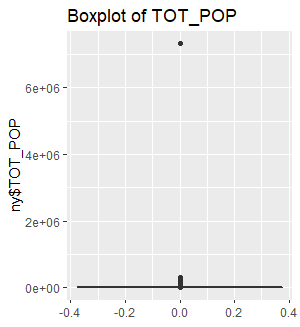
There are no presence of NULL values, so we can move further.

2) Now we have to find outliers in the data:

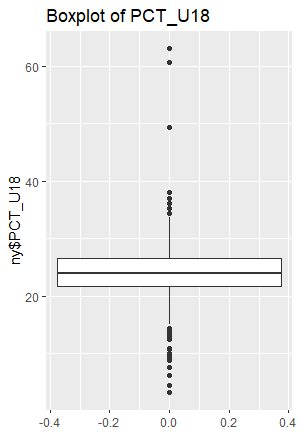
* Outliers are data point that deviates significantly from the normal objects as if it were generated by different mechanism.
* Effect of outliers on model is quite ambiguous, they might be important for model so it is important to make a model with them and another model without them.
* Outliers can be detected using box and whisker method. It is a standardized way of displaying the distribution of data based on the five-number summary: minimum, first quartile, median, third quartile, and maximum.
* The values which are greater than the value of 1.5\*(Q3-Q1)+Q3 and values that are less than 1.5\*(Q3-Q1)-Q1 are considered as outliers.

Box-plots of every column:

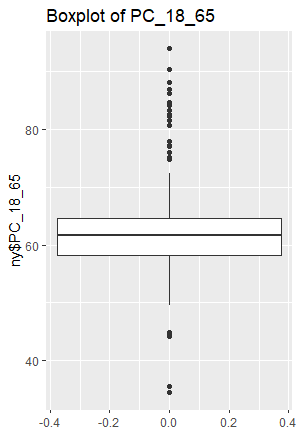
* Boxplot of total population



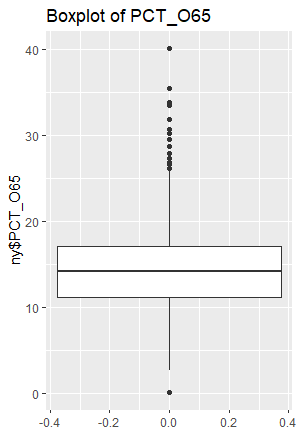
* Boxplot of no. of peoples with age less than 18



* Boxplot of no. of peoples with age between 18 and 65



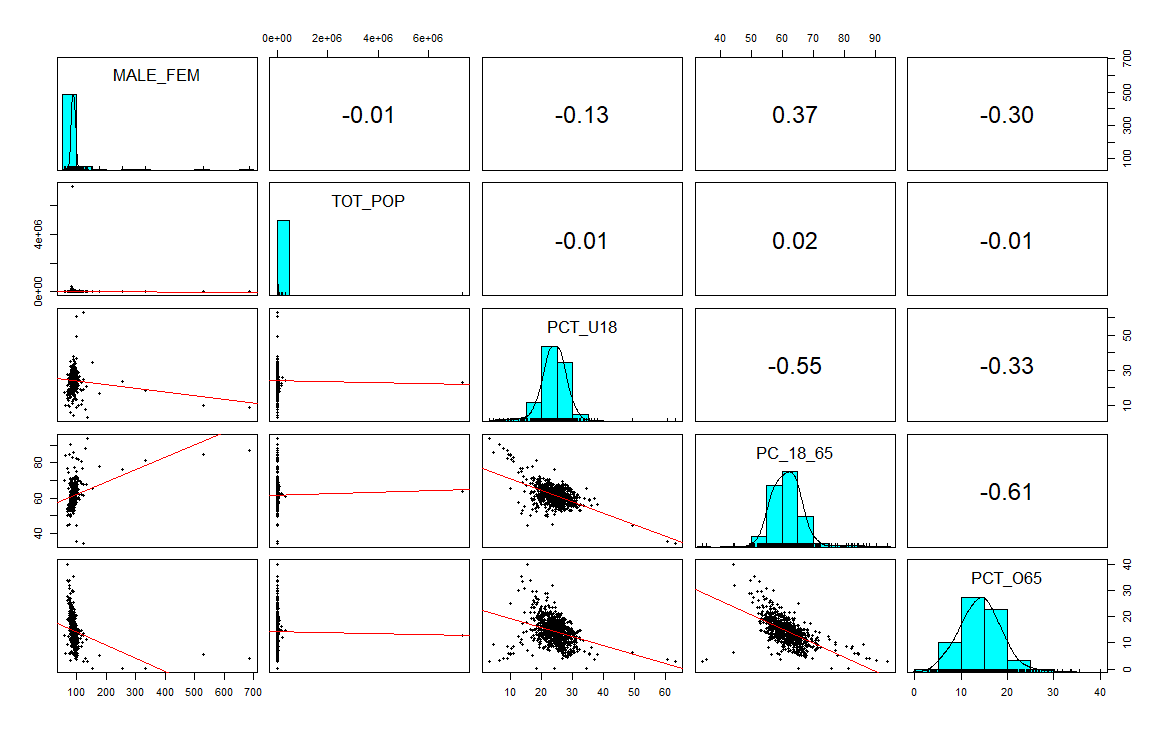
* Boxplot of no. of peoples with age greater than 65



As we can see there are too much outliers in the dataset so we have to make another dataset after removing outliers.

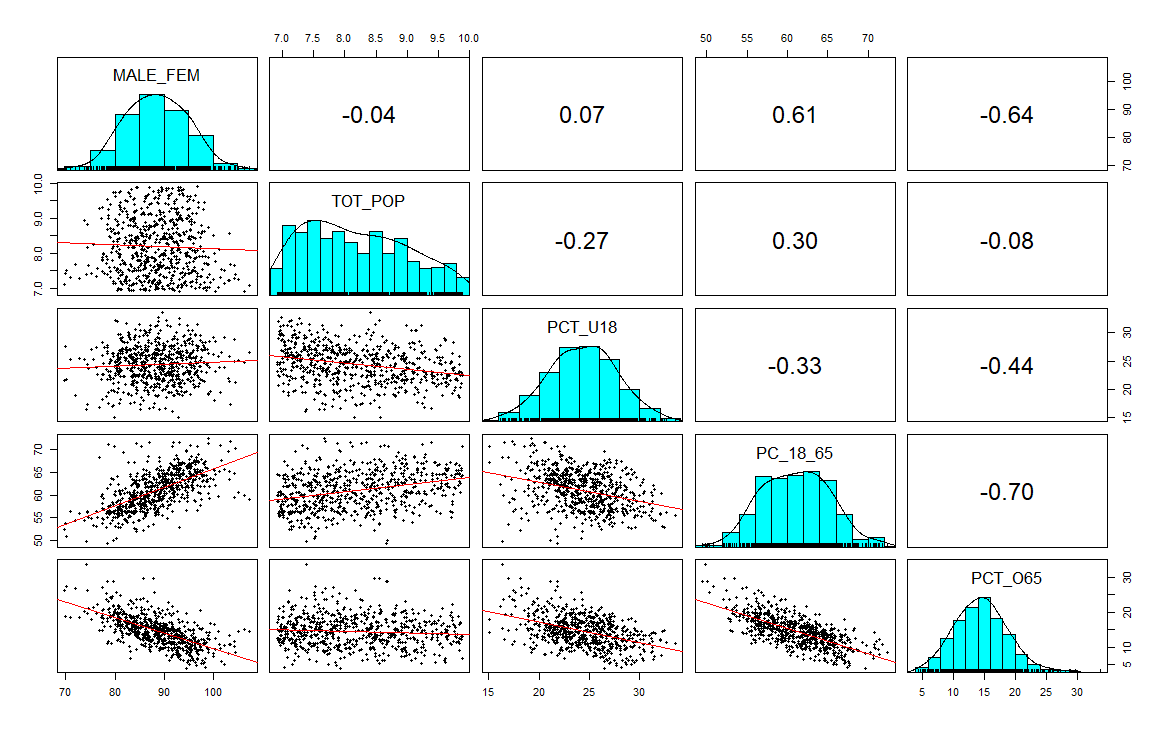
Now, lets visualize both the datasets

* Before removing the outliers



The columns MALE\_FEM and TOT\_POP are right skewed.

* After Removing outliers



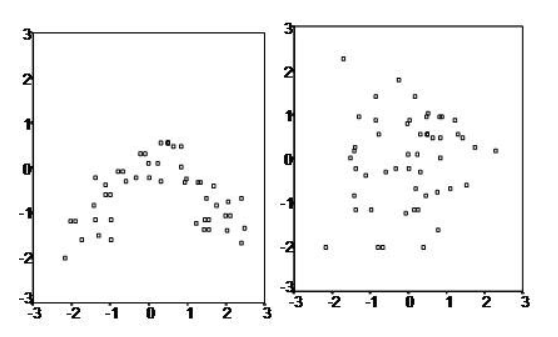
All the columns are normally distributed now but columns PCT\_O65 and PC\_18\_65 show collinearity.

**3.3 Why Multiple Regression?**

In the dataset we can clearly see that the target variable which is MALE\_FEM column is a continuous variable so in order to keep the simplest model regression can be applied here and we have more than one predictor variables that are TOT\_POP, PCT\_U18, PC\_18\_65, PCT\_O65, so multiple regression will be best to start.

Assumptions of Multiple Linear Regression:

1. multiple linear regression requires the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatterplots. The following two examples depict a curvilinear relationship (left) and a linear relationship (right).



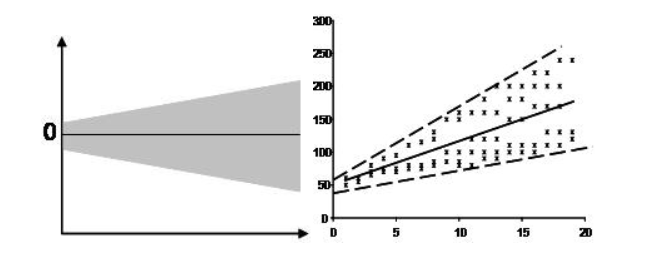
1. the multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.
2. multiple linear regression assumes that there is no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be checked multiple ways:

* + Correlation matrix – When computing a matrix of Pearson’s bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.
  + Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.

If multicollinearity is found in the data, one possible solution is to center the data. To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.

4. The last assumption of multiple linear regression is homoscedasticity. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.



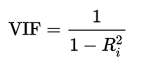
**3.4 Model Building and Evaluation:**

To build a model first we need to know what predictor variables are important, this process is called feature selection. Feature selection is useful on a variety of fronts: it is the best weapon against the Curse of Dimensionality; it can reduce overall training times; and it is a powerful defense against overfitting, increasing generalizability.

Another thing to pay attention is multicollinearity. Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

Variance Inflation factor is a famous method to detect multicollinearity, a Variance inflation factor is a measure of the amount of multicollinearity in a set of multiple regression variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression. When significant multicollinearity issues exist, the variance inflation factor will be very large for the variables involved.

VIFs are usually calculated by software, as part of regression analysis. You’ll see a VIF column as part of the output. VIFs are calculated by taking a predictor, and regressing it against every other predictor in the model. This gives you the R-squared values, which can then be plugged into the VIF formula. “i” is the predictor you’re looking at (e.g. x1 or x2):



Variance inflation factors range from 1 upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

* 1 = not correlated.
* Between 1 and 5 = moderately correlated.
* Greater than 5 = highly correlated.

Exactly how large a VIF has to be before it causes issues is a subject of debate. What is known is that the more your VIF increases, the less reliable your regression results are going to be. In general, a VIF above 10 indicates high correlation and is cause for concern.

In the dataset multicollinearity is present between PC\_18\_65 and PCT\_O65. After calculating VIF for TOT\_POP, PCT\_U18, PC\_18\_65, PCT\_O65 the values are 1.1249, 4422.1200, 7110.809964, 7761.9390 respectively. There are too high values for VIF which is strong evidence of multicollinearity, so we have to drop the column with highest value which is PCT\_O65. After dropping column PCT\_O65 the VIF for TOT\_POP, PCT\_U18, PC\_18\_65 are 1.122564, 1.16511, 1.18760 respectively. Now there is no multicollinearity in these columns and we are ready to build regression model.

In Regression model we include TOT\_POP, PCT\_U18, PC\_18\_65 as predictor variables and MALE\_FEM as target variable. After predicting values, we are able to get accuracy of approx. 50%.

**3.5 Conclusion**