**ANALYSIS OF WORLD HAPPINESS DATA**

**Data Management Project Report**

**By: Prerit Saxena, M12366923**

**Summary**

An analysis of World Happiness Report has been conducted using an open source data-set. The objective of the project includes studying the various factors which lead up to the calculation of a “Happiness Index” for each country and understanding its distribution throughout the world.

The process involves ingesting the data-set in SQL, cleaning the dataset and then using it in R through RODBC server. Further, a regression model and clustering analysis has been done in R. Finally, the data has been visualized through Tableau in which the data has been imported from Microsoft SQL Server using the SQL server connection.

**Data Overview**

**Source:** The dataset is an open source dataset from a report published on world happiness, first in 2012 and thereafter every year. This report is an outcome of the survey results of Gallup World Poll which takes representative sample from each country and asks them questions in the form of Cantril ladder, which is asking respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale.

The dataset is present here for 2015-2017 : <https://www.kaggle.com/unsdsn/world-happiness/data>

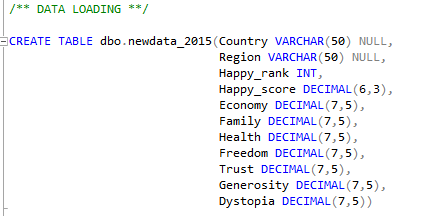
**Data Description:**

The datasets are identical except for the year they contain information of and have the following columns:

* **Country**: Name of the country
* **Region**: Region of the world, the country belongs to
* **Happiness** **Rank**: Rank of the country according to happiness score
* **Happiness Score**: Metric measured as a combination of various factors
* **Economy (GDP per capita):** The extent to which GDP contributes to happiness
* **Family**: The extent to which Family contributes to happiness
* **Health (Life Expectancy):** The extent to which Life Expectancy contributes to happiness
* **Freedom**: The extent to which Freedom contributes to happiness
* **Trust (Government Corruption):** The extent to which trust in government contributes to happiness
* **Generosity**: Generosity of the general public and its contribution to happiness
* **Dystopia Residual**: Contribution to Dystopia residual to happiness. Dystopia is an imaginary country that has the world’s least happy people. The purpose of having this is to have a lower benchmark so that all countries do positively against it. This variable has no physical significance.

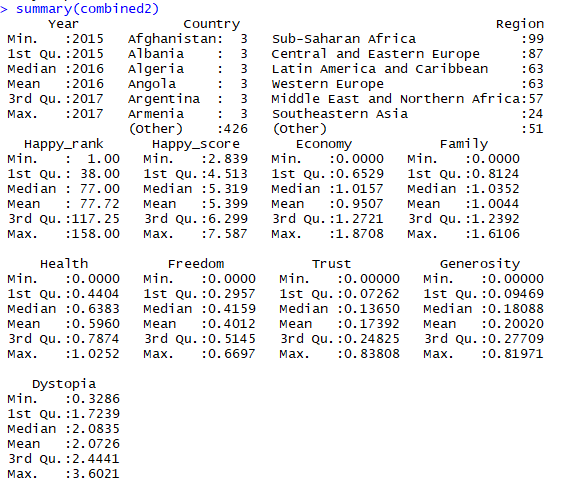
**Data Loading:**

Data is present in the form of 3 csv files, one for each year – 2015,2016 and 2017. The files are imported in a Database “DM\_Project” through the import task. A general schema of the imported file is here:



**Summary:**

Summary() function in R gives a summary of the combined dataset for all three years



**Data Preparation**

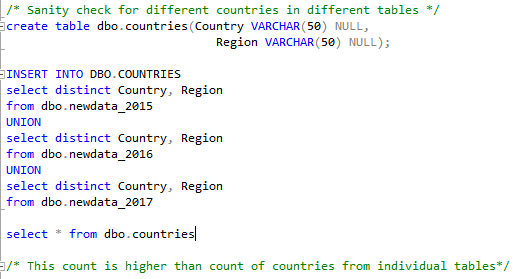
For the purpose of analyzing the data in an analysis tools like R, data needs to be in the correct format. To carry out this, data needs to be checked for normalization first and then data cleaning is required (to remove any inconsistencies present in the dataset).

**Normalizing Data:**

The dataset is already in a normalized form, i.e. present in three separate tables – each one pertaining to a year of data. However, later, for putting a structure to the data for regression analysis in R, we will combine all three datasets into one and include a column of Year in it.

**Data Cleaning:**

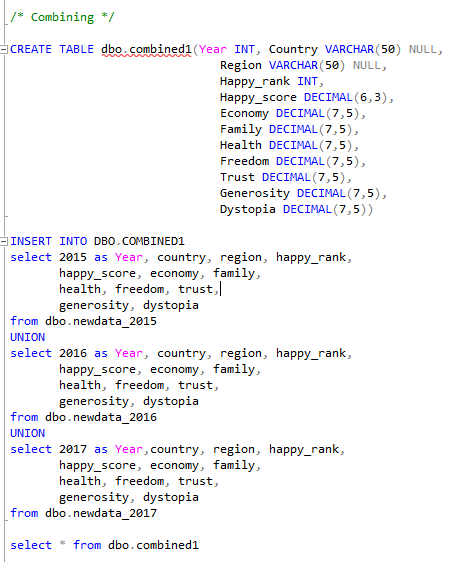
First, the data has been checked for any inconsistencies in country names. This has been checked by creating a master list of countries combining countries and regions from all three datasets and checking overlap between it and individual datasets.



The total count for combined list is 163 which is higher than newdata\_2015(158), newdata\_2016(157) and newdata\_2017(155). This implies that there are few countries which are not a part of all three datasets and hence, including them in final combined dataset will result in null values.

Hence, we will combine the dataset in two steps:

Step-1: Combine the data from all three tables into one dataset



Step-2: The next step is to create a de-duped list of countries and regions following by pulling year for every entry of this list using a FULL JOIN. This will create a table which is unique at Country, Region and Year level.

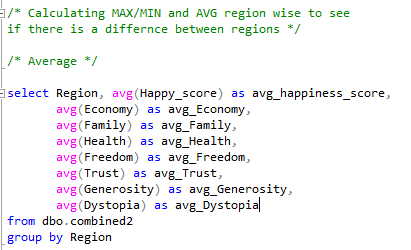
Then the remaining columns have been included in this list using a LEFT JOIN with the combined table. A complex nested SQL query has been used for achieving this.

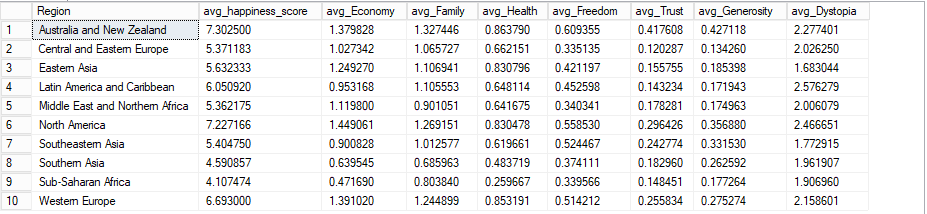


**Data Statistics**

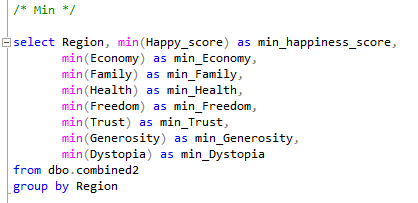
Using SQL to obtain basic statistics from the combined dataset. MIN/MAX/AVG has been calculated at region level for all the variables.

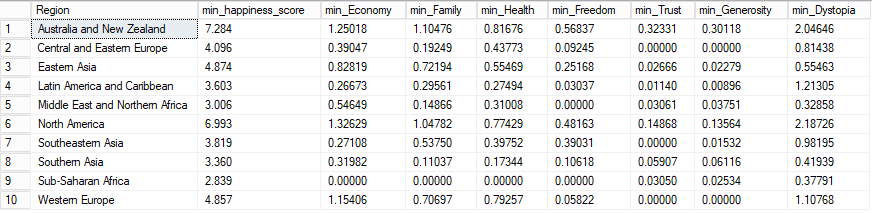
**Average:**



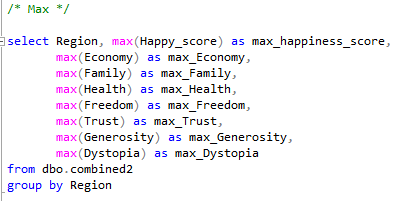


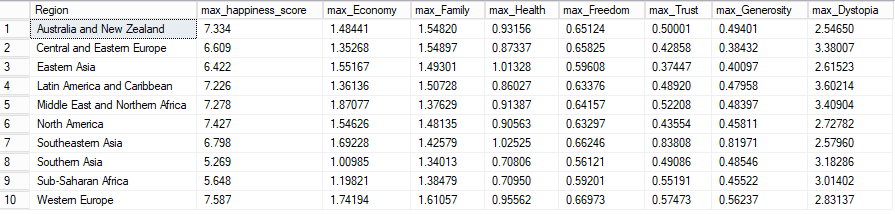
**Minimum:**



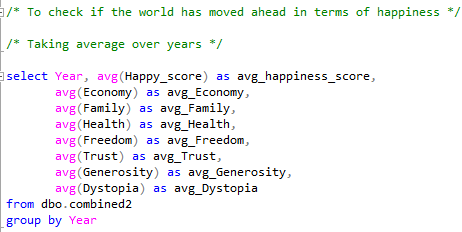


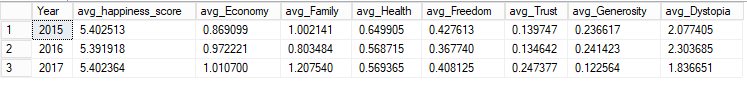
**Maximum:**





Also, we can take the average of variables over the last three years to see if there has been a trend in happiness or other factors over the years.



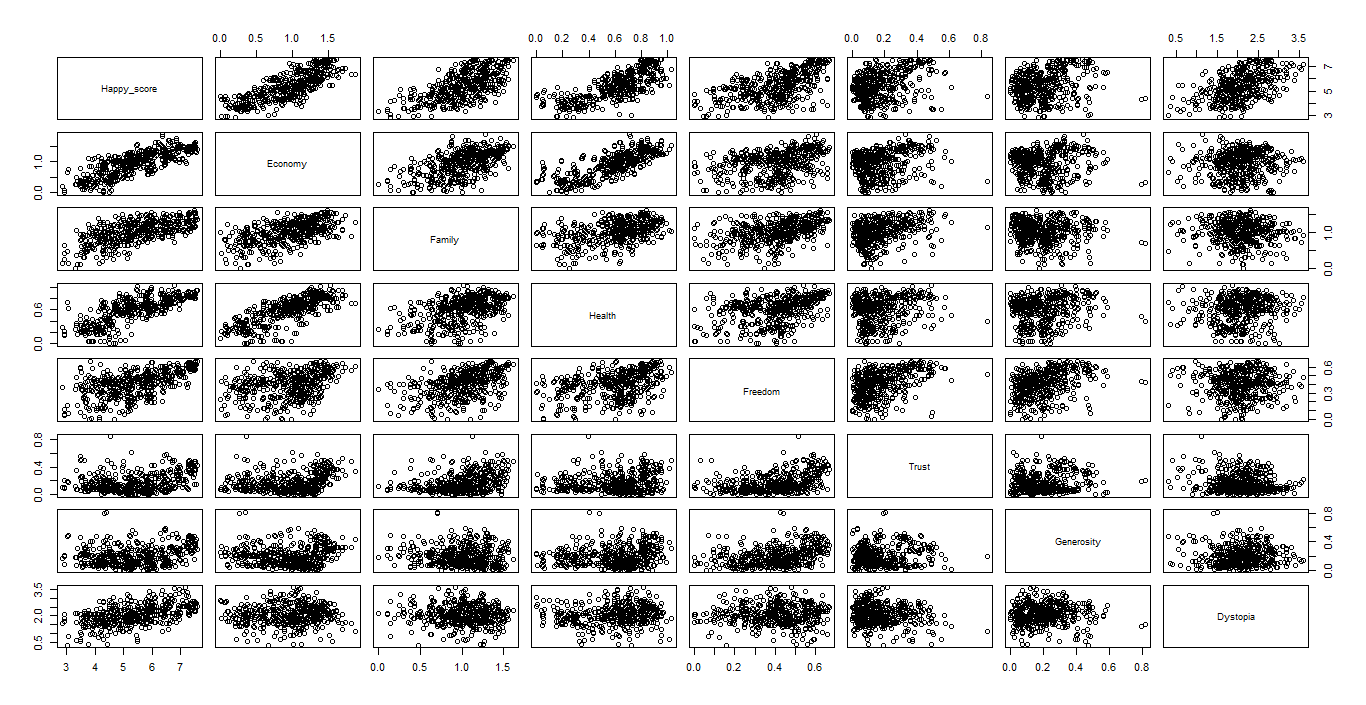


This is indicative of the fact overall world happiness has not changed over the years but there is certainly a trend over the years for factors like Economy.

**Exploratory Analysis and Data Modeling**

**Data Exploration:**

R has been used for the process of exploratory data analysis and data modeling. A basic regression model assuming happiness score as target variable and remaining variables as regressors has been created. This is followed by clustering the countries to see if the clusters are indicative of the regions.

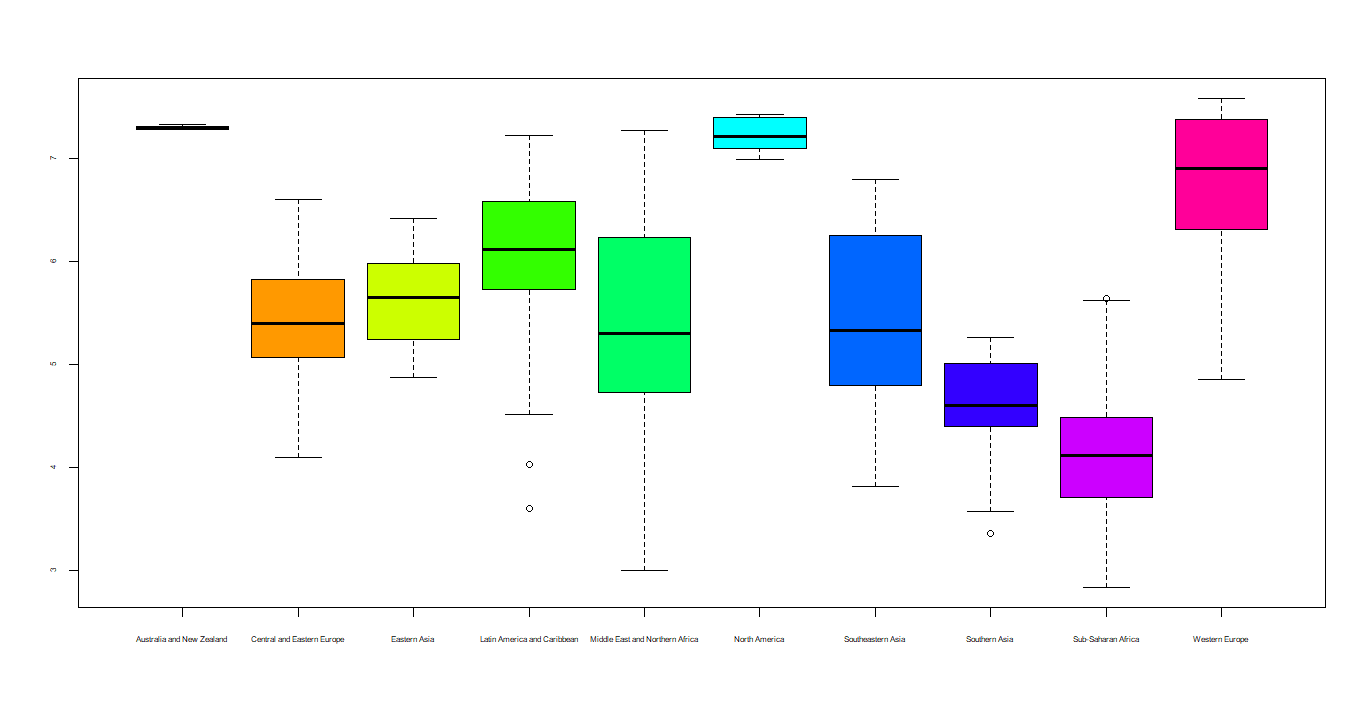


The correlation plot in R shows a strong relation between happiness and factors like Economy but not so strong with factors like Generosity. Few other factors like Economy and Family can also be seen as highly correlated.

**Exploring through other plots:**

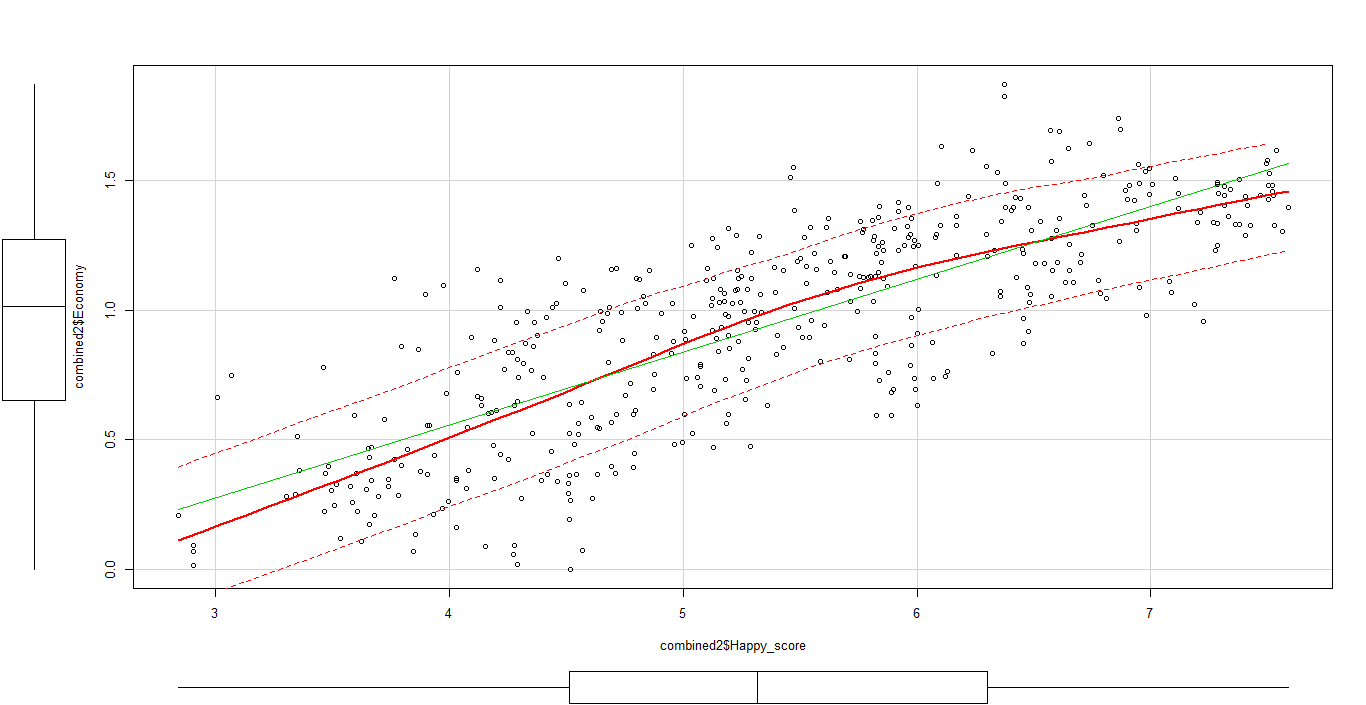
Creating a box-plot of happiness score by region





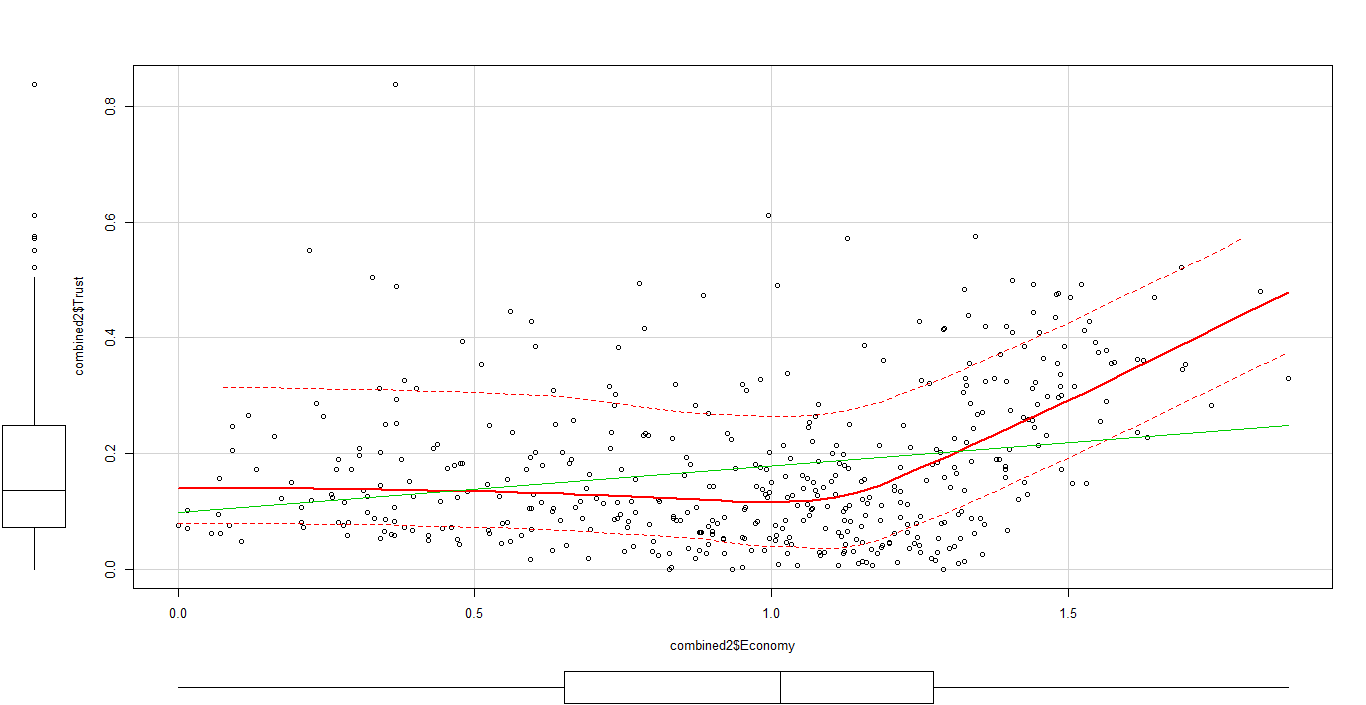
This graph shows North America with highest overall happiness score and Sub Saharan with lowest overall happiness score.

Creating a scatter plot of happiness score vs Economy



The graph shows that happiness increases with the strength of economy in each country. A reason why probably more developed countries have higher happiness scores.

Creating a scatter plot of Economy vs Trust

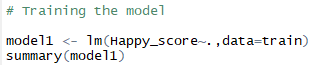


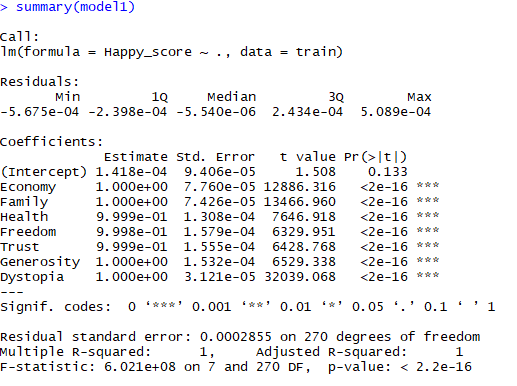
This graph shows a surprising trend that trust score remains fairly constant for economies on the lower side but then rises for strong economies.

**Data Modeling:**

Linear Regression

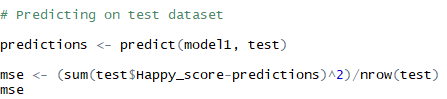
Linear Regression model using lm() in R has been created after dividing the data into training and testing. Here’s a summary of the regression model





The model shows all coefficients to be close to 1 and an R-squared value of 1 which means the model explains the variance in the data very well.

Next, predictions on the test dataset can be made using this model.



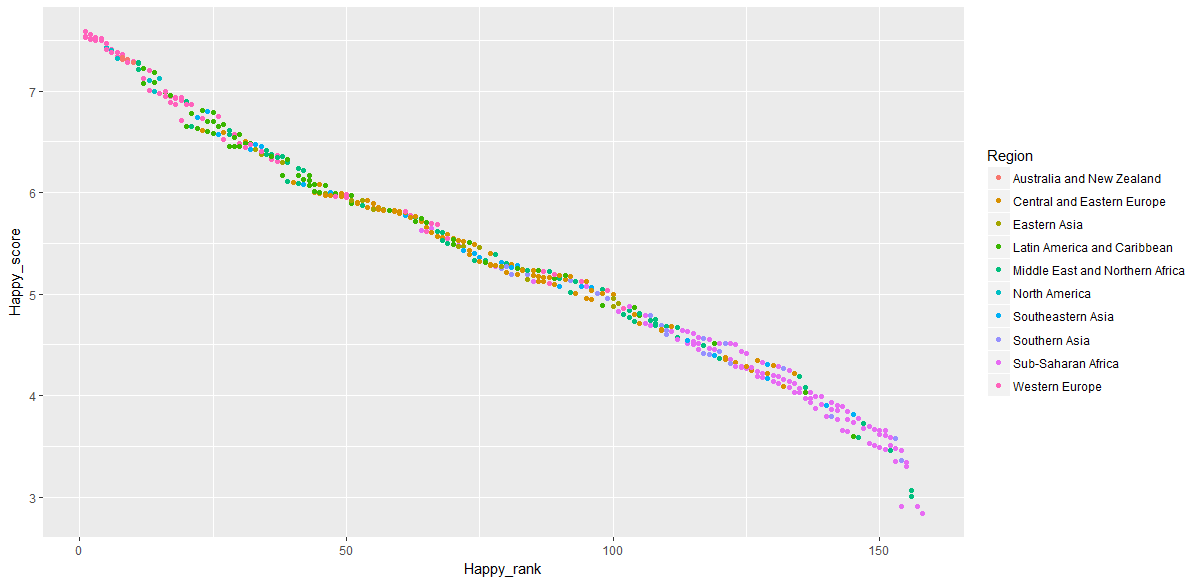


A very low Mean-Squared Error(MSE) indicates good predictions on the test on the test data.

Clustering

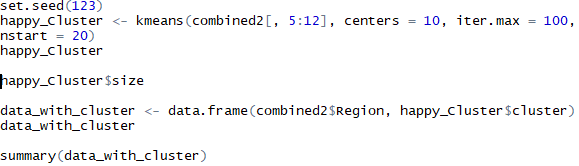
Exploring the possibility of clusters in raw-data using scatter plot of the combined dataset. Our objective in this exercise will be to see if clusters formed through passing the data into a clustering algorithm will give similar results.

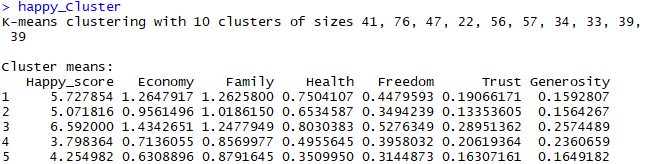




This graph indicates that happiness score of countries in different regions is not randomly distributed, rather there is a grouping which segregates the regions.

Now, its time to pass all the factors into our k-means clustering algorithm.

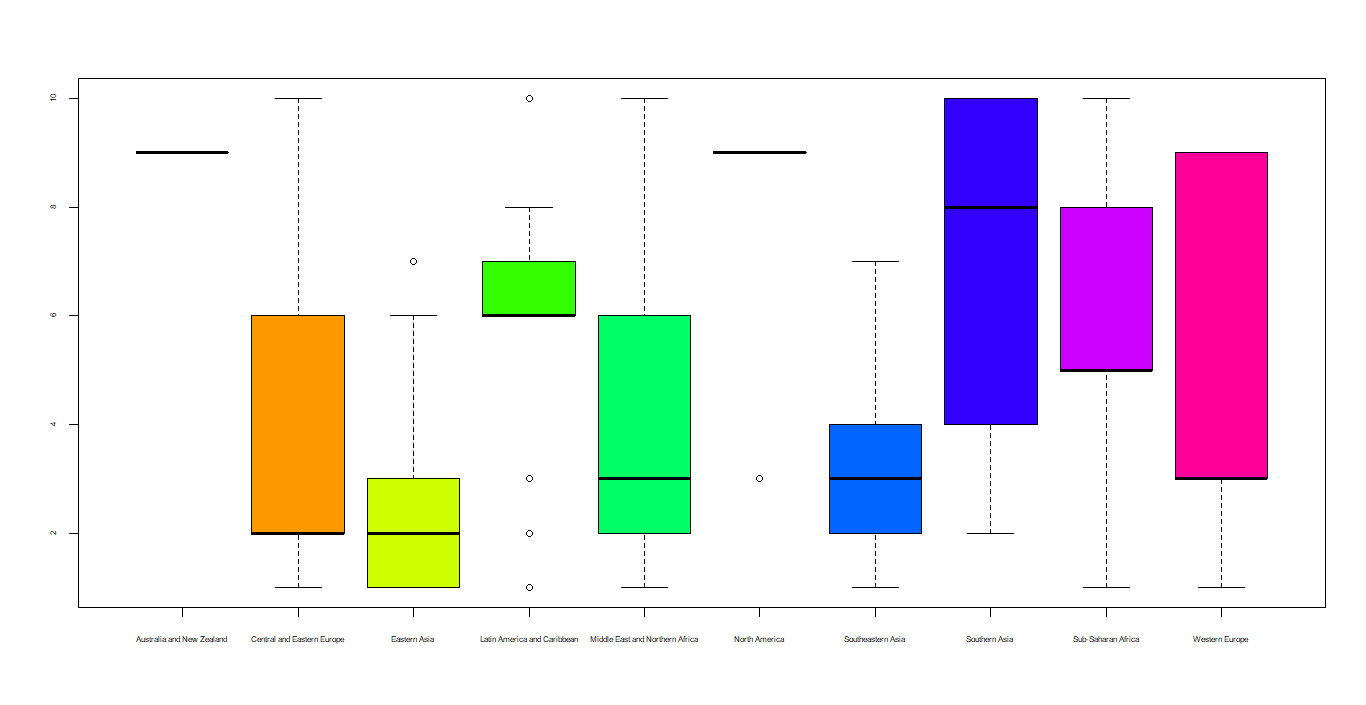




This algorithm creates 10 clusters and assigns a cluster value to every record. Also notice, a country may fall in different clusters in different years but possibility of that happening is very low as in general, countries do not undergo this amount of drastic change in consecutive years.

Now, lets create a box-plot using the cluster values





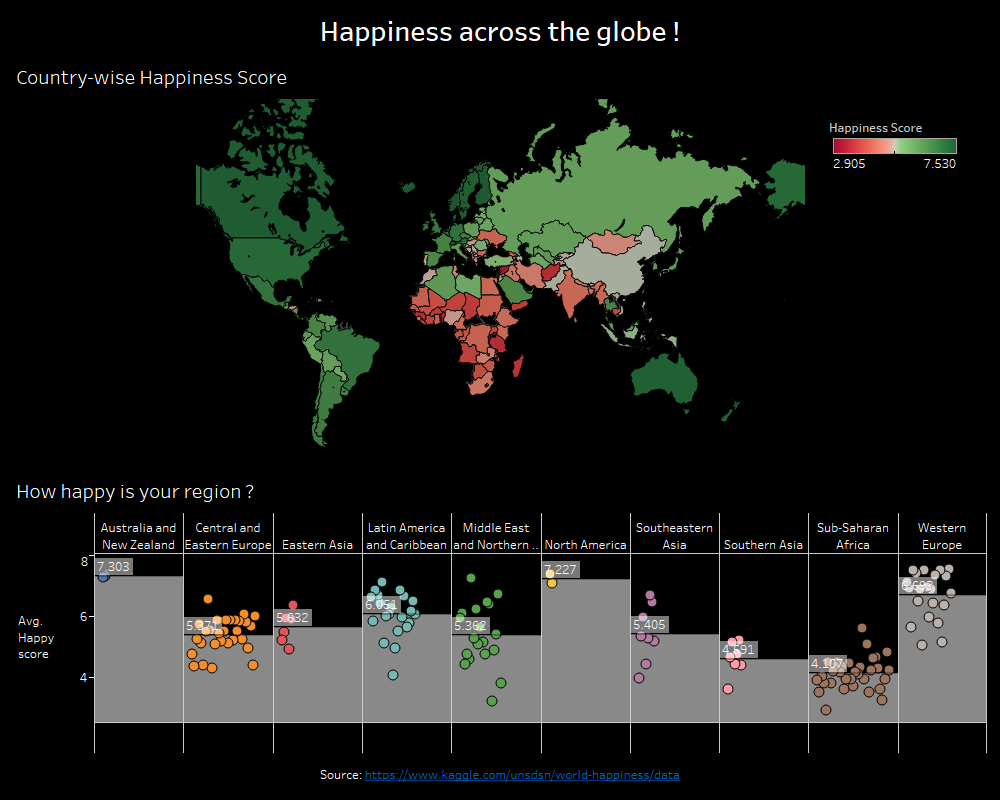
This box-plot is very similar to the box-plot seen initially with happiness score values. Had the linear regression model been poor, this plot might have differed significantly from the earlier one.

**Visualizing data through Tableau**

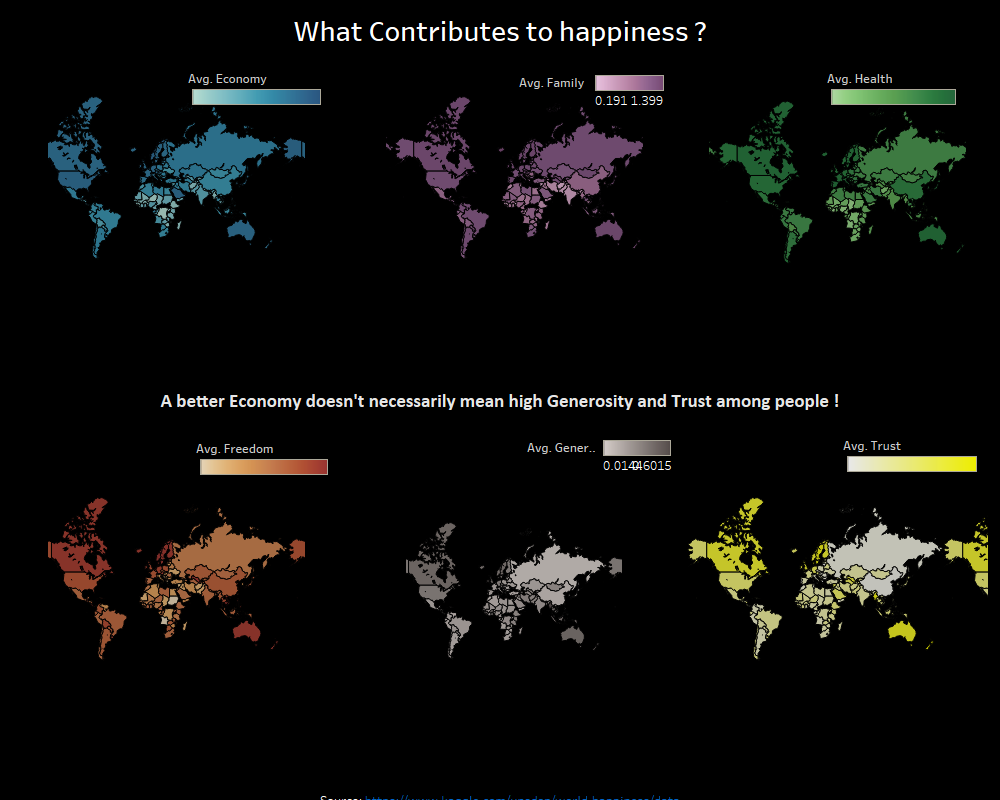
Data Visualization software-Tableau has been used for the visualizing the data. This involves creating a connection from SQL Server to Tableau using the in-build functionality of Tableau and then importing the combined dataset.

The story is summarized using 2 Dashboards:

**Dashboard 1: Overall Happiness Score**



**Dashboard 2: Happiness Contributors**



**Summary**

After carrying out the analysis and plotting the data, we can arrive at the following conclusion

* Happiness across the world is highly dependent of region. Countries in North America are far happier than countries in the Sub-Saharan region of Africa which lag behind in development
* Happiness score is merely an overall estimate. It is a combination of various factors like Economy, Family, Trust etc. which in turn contribute to happiness
* Happiness score is not directly correlated with all other factors. While factors like economy show a high correlation, ones like trust or generosity show a lower correlation
* Factors, within themselves, are also not directly correlated. Economy and Trust show a very small correlation which means a good economy is not definitive of trust of people in their governments
* Clustering on factors also show a similar pattern as plotting happiness score by region

**Challenges**

* Input data was not in its cleanest format. Few formatting changes in raw files had to be done to make the formatting similar for all three files
* Multiple files had to be created in SQL for Data Preparation. Had the dataset been larger, this would have used significant amount of space on the system
* A lot of knowledge of Tableau was required to create the dashboards. The dashboards are interactive and can be seen using Tableau Public