BUS240F Final Report

World Happiness Report

Team 2

Ailin Zhang, Chan Khine, Lihao Xu, Lurui Yu, Mengzhou Wu

Table of Content

1. Context / Domain / Market of Consultancy 3
2. Focus of the Consultancy 3
3. Data Quality 4
    3.1. Data Quality Assurance 4
    3.2 Data Preparation and Wrangling 5
4. Data Visualizations (Insights and Interpretations) 7
    4.1. Heat Map 7
    4.2. Boxplot by Continent 8
    4.3. Histogram & Facet Chart 10
    4.4. Line Chart 11
    4.5. Violin Chart 12
    4.6. Cluster Analysis 14
    4.7. Outlier Analysis 16
    4.8. Regression with Outliers 17
    4.9. Regression without Outliers 19
5. Business Reflections 22
6. Reflection About the Process 22

1. Context / Domain / Market of Consultancy

We are a company specializing in providing immigration consulting services for high net worth clients all over the world. Our clients are mainly individuals who want to migrate to other countries to live a better life. Under this circumstance, our clients may chase after a business-friendly environment (investment immigration), high wages (skilled workforce), or a country with good social welfare. Our client's main concern is the factors of different countries, including how happy they would be in a certain country, economic situation, government trust, etc., and we are a unicorn in this field who controls first-hand information and resources. We aim to offer immigration services that focus on popular and targeted immigration destinations instead of all countries in the world. To better give our clients a thorough understanding of the pros and cons of each country, we mainly formed our report using the data from The World Happiness Report and also the International Migration Database from 2016 to 2020, 5 consecutive years for our clients' reference. We utilize the data to make visualizations from broad views to narrative views regarding the concerns of our clients.

2. Focus of the Consultancy

Moving to a new place to live is an exciting, and life-changing decision, not to mention immigrating to another country. This decision should be based on a lot of evaluations, thoughts, or personal preferences. Various factors can be helpful in deciding on moving to another country, such as concerns about safety, economics, etc. Therefore, according to the customer's preference of different factors, we will search for the factors and analyze the suitability of different countries for our clients. For example, if our customers are particularly concerned about the economic development level and medical level of the country they plan to immigrate to, we will search and

analyze the level of GDP and life expectancy of each country, and provide them with the most suitable countries for immigration. After all, we will enumerate a group of lists that are most suitable and proper for them to immigrate to after analyzing the advantages and disadvantages of each country, and give our customers the options.

3. Data Quality

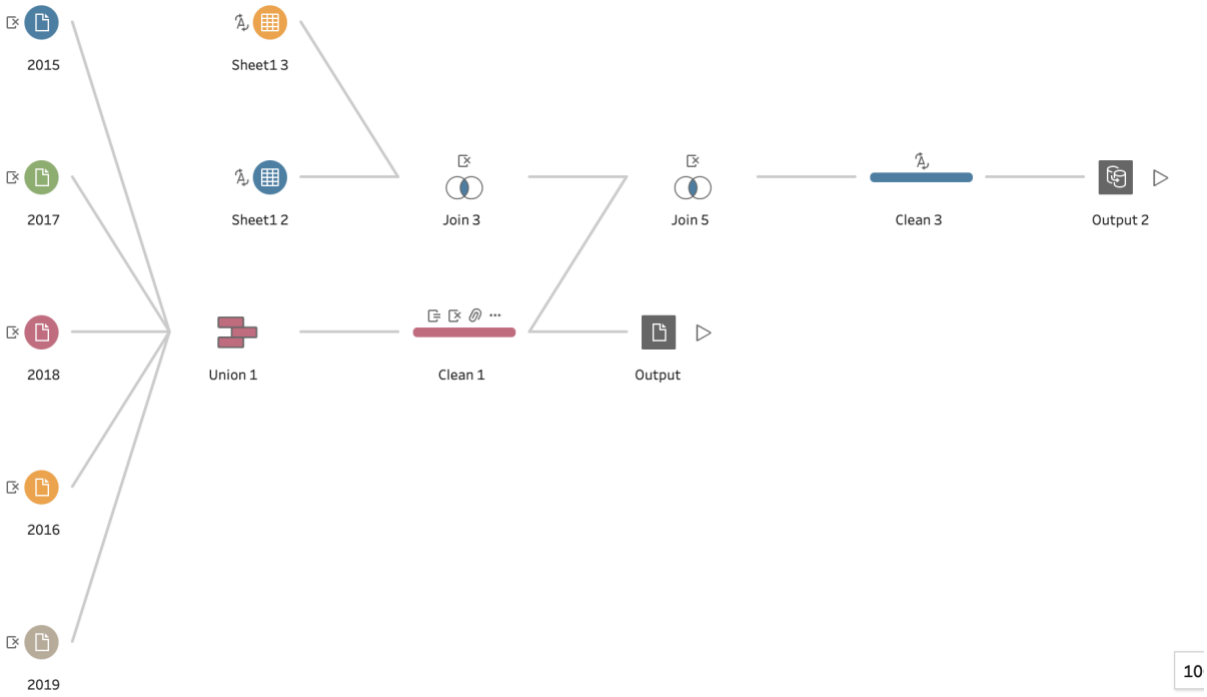
3.1. Data Quality Assurance

The world happiness data is sourced from World Happiness Report 2020, and the data is downloaded from Kaggle. The data downloaded from Kaggle is a compilation of 6 years from 2015 to 2020. We downloaded the original data and compared it with the data from Kaggle to confirm the data's quality. There are in total six data files listed from the year 2015 to the year 2020. These data files are credible in that they are not created but sourced and sorted. The variables in each file are collected from different surveys, research, and organizations. For example, the Happiness score is collected from a SWB survey released on Gallup World Poll (GWP), data of GDP per capita is collected from the World Development Indicators (WDI), Healthy Life Expectancy (HLE) is based on data extracted from the World Health Organization's (WHO) data repository, Freedom is the national average of responses from GWP survey, and Corruption recorded the national average of the survey in GWP. We also cross-checked the public data with organizations such as WHO and the data correctness and cleanliness are high. We can conclude that for public information, the data quality is high.

As a company that provides immigration services, we realize that world happiness data is not sufficient for conducting our business. Therefore we add the international migration data from the OECD (Organization for Economic Co-operation and Development) database. The new dataset

contains two variables that would interest our clients the most: total inflows of migrants and foreign employment by gender. These two factors reflect how hard it is to immigrate and get a job.

3.2 Data Preparation and Wrangling



To ensure that our current data is easy to analyze and create visualizations, we used Tableau Prep to clean and merge our data. As shown in the figure above, we merged all the years from the World Happiness Report to create a single file with all years combined. We also decided to change the *Year* variable to a string type from numeric since changing it to a date type will add month and day, but we only wanted to use year. During the clean step, we removed duplicates, renamed and grouped misspelled countries or outdated country names, and dealt with null values by either removing them from the dataset or adding some median values if we have data for that country for four years.

After ensuring the data quality from the merged World Happiness dataset, we merged it with the two sheets from the international migration dataset. We used the inner join feature from Tableau Prep to ensure that we have all the data we need for each country after joining. Then, we removed duplicate rows and dealt with null values. Hence, we obtained clean data on 32 countries for each year. One concern after merging these two datasets is that we eliminated too many countries from our list. However, we concluded that this is reasonable since the 32 countries on the list are the popular immigration destinations. We decided to focus on the countries that welcome immigrants and are commonly chosen as immigration destinations instead of distributing our attention to every country in the world.

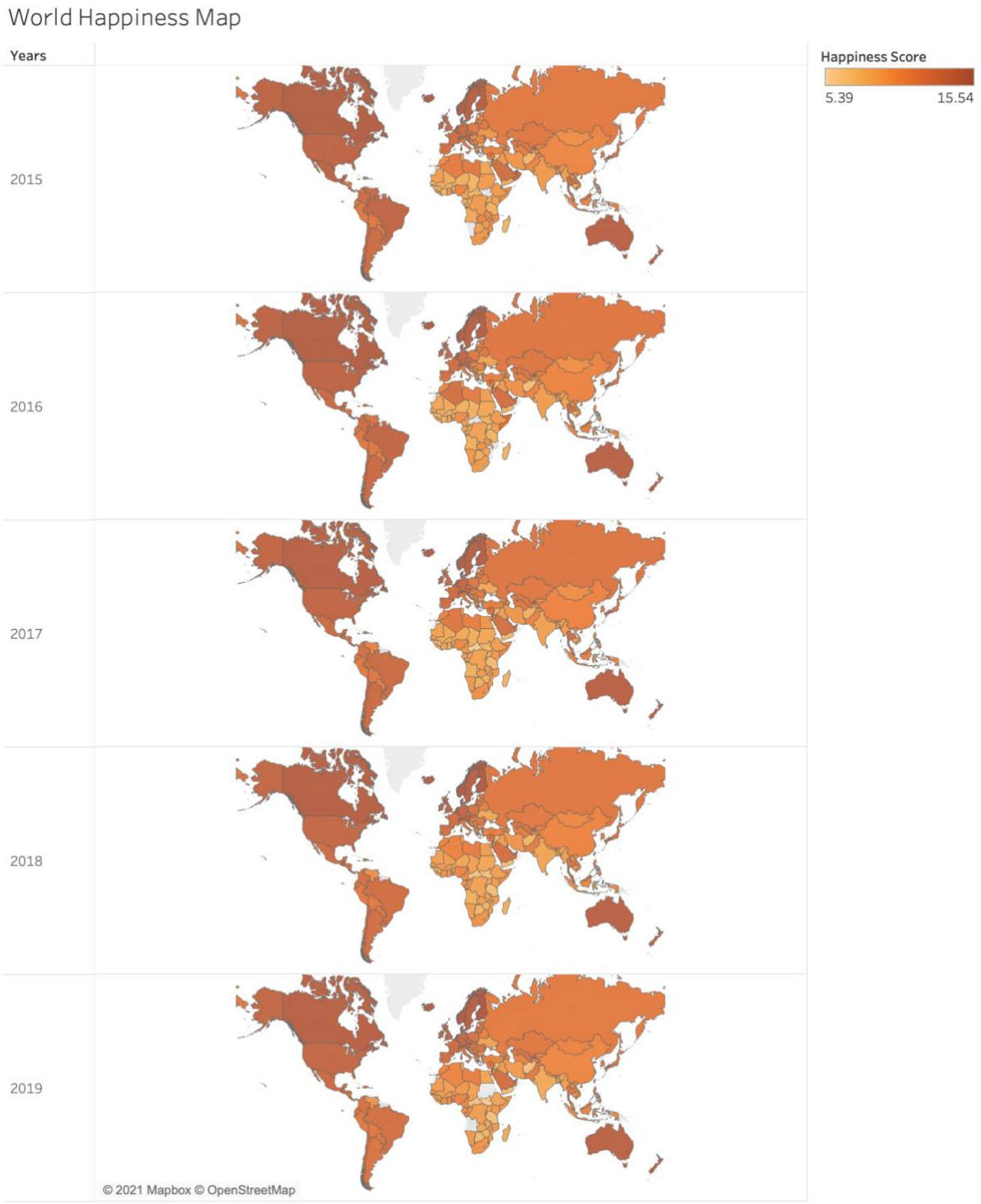
```
$ Years                              : chr
$ Generosity                         : num
$ Happiness_Rank                     : chr
$ Country                            : chr
$ Happiness_Score                    : num
$ GDP_per_Capita                     : num
$ Life_Expectancy                    : num
$ Freedom                            : num
$ Government_Corruption              : num
$ Total_Inflows_of_Migration         : num
$ Foreign_born_Men_Employment_Rate   : num
$ Foreign_born_Women_Employment_Rate: num
```

```
$ Years                              : chr
$ Generosity                         : num
$ Happiness_Rank                     : chr
$ Country                            : chr
$ Happiness_Score                    : num
$ GDP_per_Capita                     : num
$ Life_Expectancy                    : num
$ Freedom                            : num
$ Government_Corruption              : num
$ Total_Inflows_of_Migration         : int
$ Foreign_born_Men_Employment_Rate   : num
$ Foreign_born_Women_Employment_Rate: num
```

As shown in the picture above, we also used R to re-check the data types after we got our final clean dataset. We first renamed almost all columns because the names have spaces in them. Thus, we replaced spaces with underscores. We used *str* to check the data types in the dataset. As shown in the left figure, we recognized that we can change the total inflow of the migration column to integer data type since it is the total number of people. We also used *na.omit* in R to guarantee that we did not miss any null values when cleaning in Tableau Prep. Hence, we are confident that our dataset is complete and clean to start data analysis for our clients.

4. Data Visualizations (Insights and Interpretations)

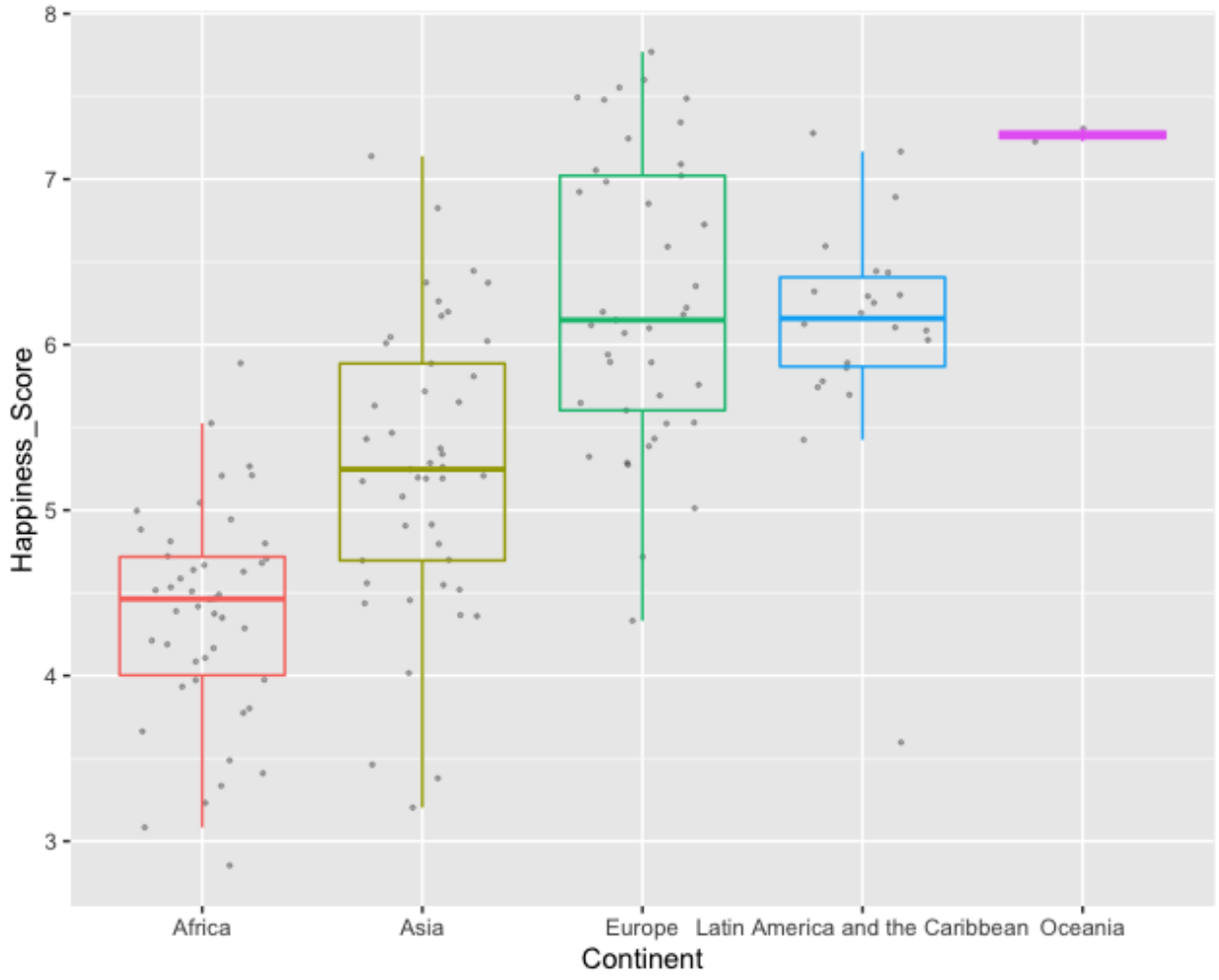
4.1. Heat Map



Map based on Longitude (generated) and Latitude (generated) broken down by Years. Color shows sum of Happiness Score. Details are shown for Country.

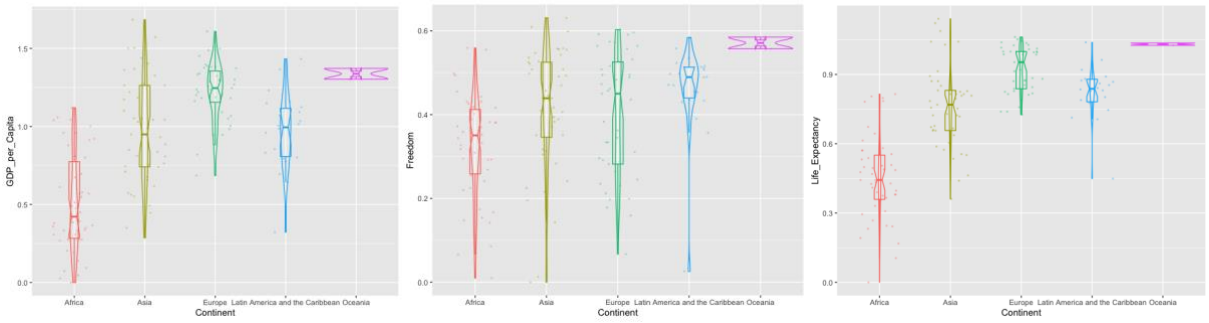
We created a heat map to see the overview of countries and their standings on happiness scores from 2015 to 2019. The darker the color is, the higher the happiness score. In South America, the majority of countries cluster around a score of six on the happiness scale. In Europe, Finland comes out on top of the world for a second consecutive year, and it is not difficult to see why. The country is well-known for its stable work-life balance, supported by a comprehensive welfare state.

4.2. Boxplot by Continent



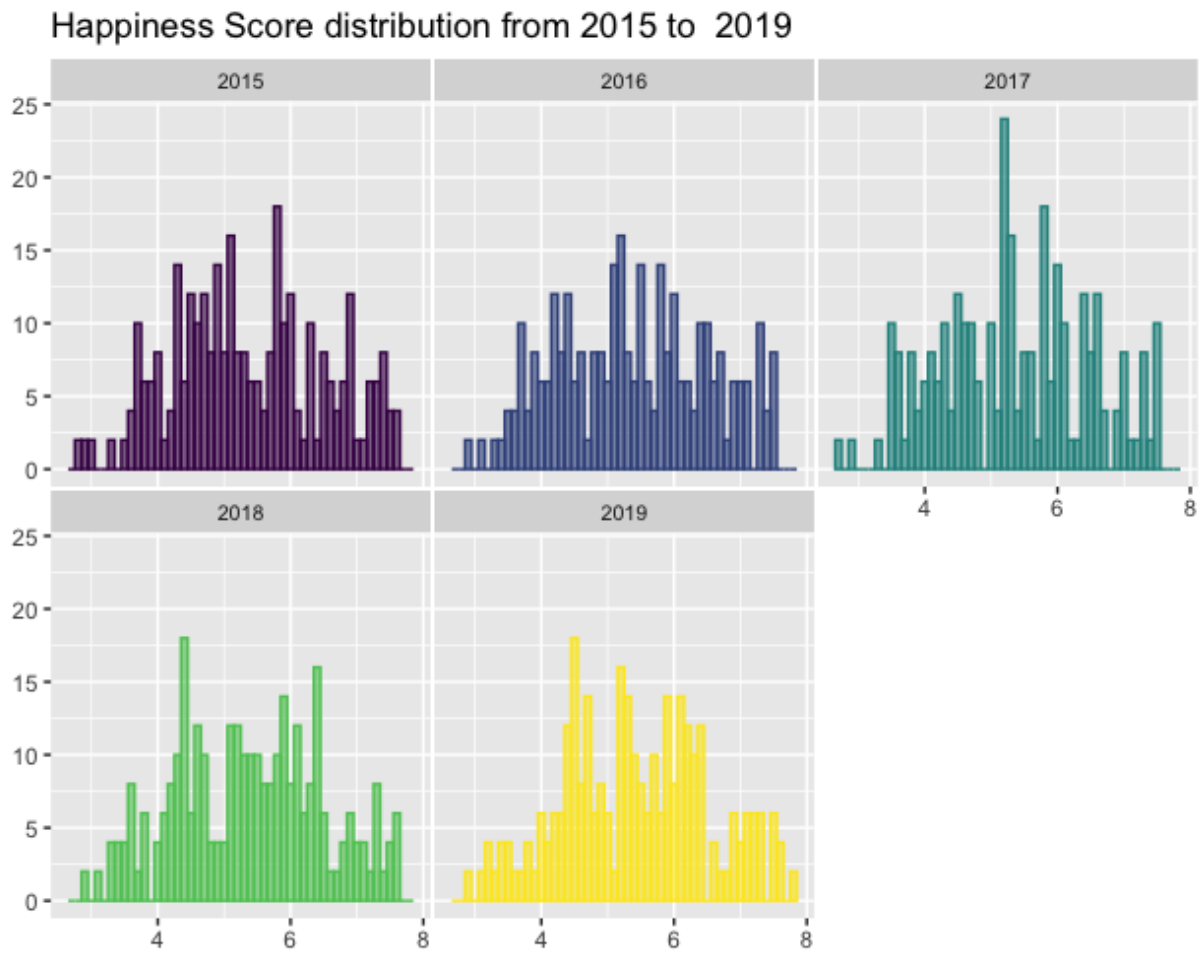
We grouped the data according to the continent to which it belongs and drew a boxplot to see how it distributes. It can be seen from the figure that the European continent has the highest

happiness score, and Africa has the lowest. Since our data for the Oceania continent only contains 2 countries - New Zealand and Australia - we are not going to further discuss it due to insufficient data. Next, we wanted to explore the possible factors that might contribute to the relationship of happiness scores on each continent.



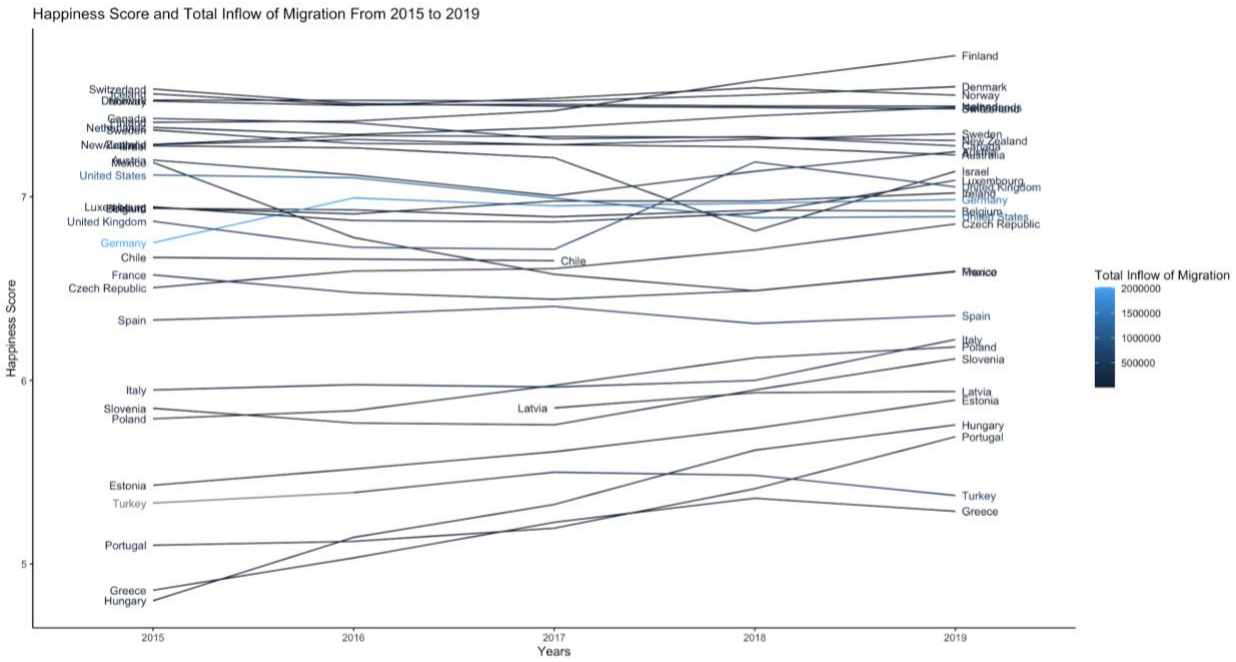
To figure out the potential factors that might influence happiness score, we combined the boxplot and violin plot to show each continent's life expectancy, GDP per capita, and freedom by continent. The middle figure shows each continent's freedom. It seems that the freedom indicator does not differ from each other; each continent has a close median freedom score, thus we concluded that the freedom indicator will not affect each country's happiness score too much. The other two graphs, GDP per Capita and Life Expectancy follow the same path as happiness score with European courtiers being highest and Africa relatively lower than other continents.

4.3. Histogram & Facet Chart



The above visualization is a combination of Histogram Plot and Facet Plot. The Chart shows the relationship between the score of happiness and the frequency of the score. It is clearly shown that each year the distribution of the score is like a bell-shape, which looks normally distributed since most countries are in the 4-6 range while some countries are larger than 6 or smaller than 4. Also, we use the facet plot to showcase the time difference of each year. We can see from the graph that the scores of happiness are going down from time to time, and the numbers of countries whose scores exceed 6 are decreasing, moving to the 4-6 zone. We believe the reason for this is the economic downturns and the instability of some countries and regions.

4.4. Line Chart

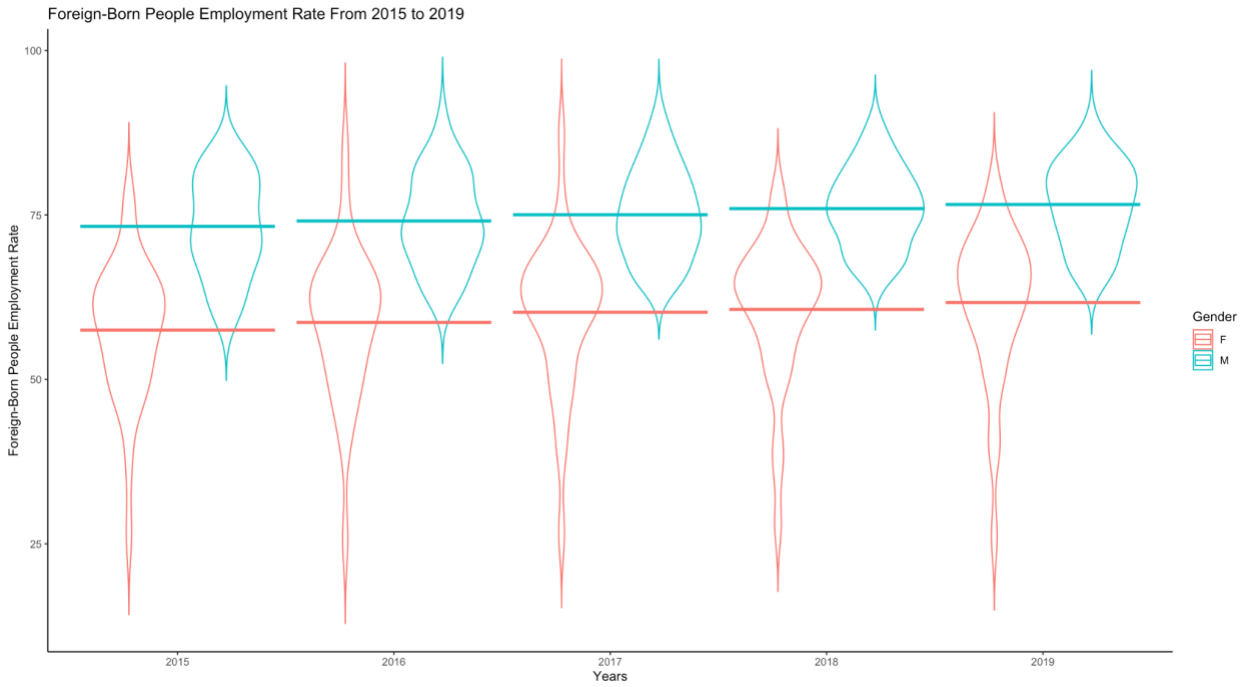


The above visualization is the line chart that shows the total inflow of migration via happiness score for 37 countries from 2015 to 2019. Since there are 37 countries, the labels are overlapping for the countries that have happiness scores that are more than 6.5. It seems that we can distinguish the countries into 2 categories that are lower or higher than a 6.2 happiness score. These countries that are lower or higher than 6.2 happiness scores seem to fluctuate within that range of 4.7 to 6.2 and 6.2 to 7.7 happiness scores. These trend lines for each country raise an interesting question to search for the factors that change the happiness score. However, line charts do not show that detail.

Using the total inflow of migration for each country, we can see that most of the countries have less than 500,000 for the total inflow of migration. It is interesting to observe this occurrence since one might think people will migrate to countries with higher happiness scores. We can only see the light blue colors (1,500,000-2,000,000) for Germany and the United States. The United States might have more migration as people from several countries might be seeking asylum as

well as people migrating for work. It is intriguing to see that Germany seems to have the most total inflow of migration although its happiness score is not in the top 10 countries. This shows that there might be other factors that people consider when they decide to migrate to other countries.

4.5. Violin Chart

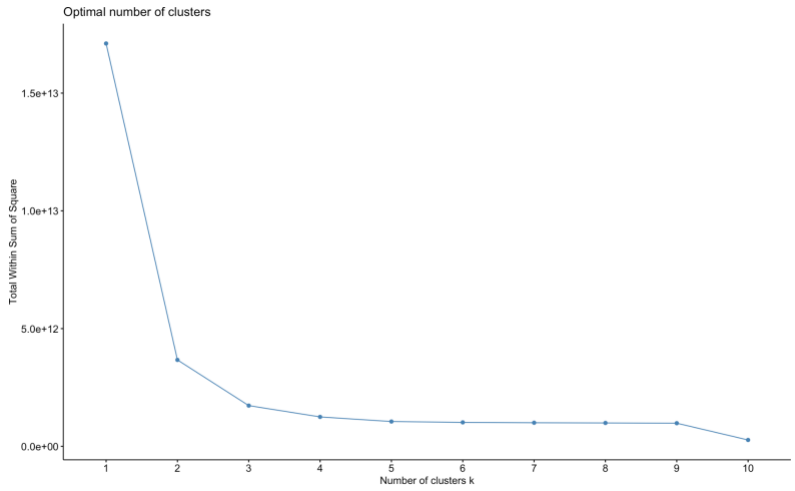


After analyzing the happiness score distribution for both countries, continents, and the inflow of migrations, we decided to explore the data regarding the overall foreign-born population employment rate. The above visualization contains the violin plots that show the employment rate of foreign-born men and women from 2015 to 2019. Red represents women and green represents men. It seems that the mean crossbars in each violin plot are larger as the years increase from 2015 to 2019. However, if we compare the two plots for men and women, the mean for the foreign-born men employment rate is between 70% to 80% while the mean for the foreign-born women employment rate is between 50% to 75%.

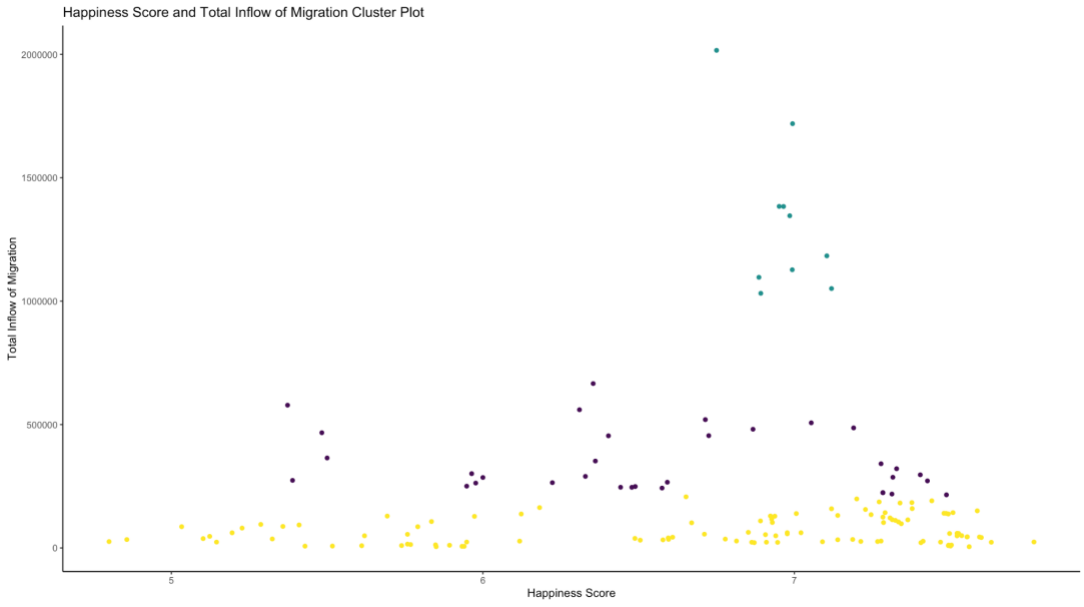
Although the distributions within each violin plot are quite similar for the plots of the same gender, we can see the difference between the distributions for the violin plots for men's and women's employment rates. For instance, while the violin plots for the male employment rate have wider distribution overall and have shorter narrow tails, the plots for the female employment rate have narrower and longer tails. This is interesting because most of the male employment rate is between 65% to 85% while most of the female employment rate is between 55% to 70%. There might be some form of gender inequality for employment within these countries. However, violin plots do not tell us which countries have the largest or the smallest employment rates.

By looking at the visualizations, we can see that the employment rate for foreign-born men and women is getting higher overall. However, the tails for the violin plots for 2016 and 2017 show that the employment rate reaches as high as 100%. Furthermore, the lower tails for the violin plots for the foreign-born men employment rate seems to be increasing from 50% (2015) to about 57% (2019). Nevertheless, the lower tails for the women seem to be around the same employment rate of 20%.

4.6. Cluster Analysis

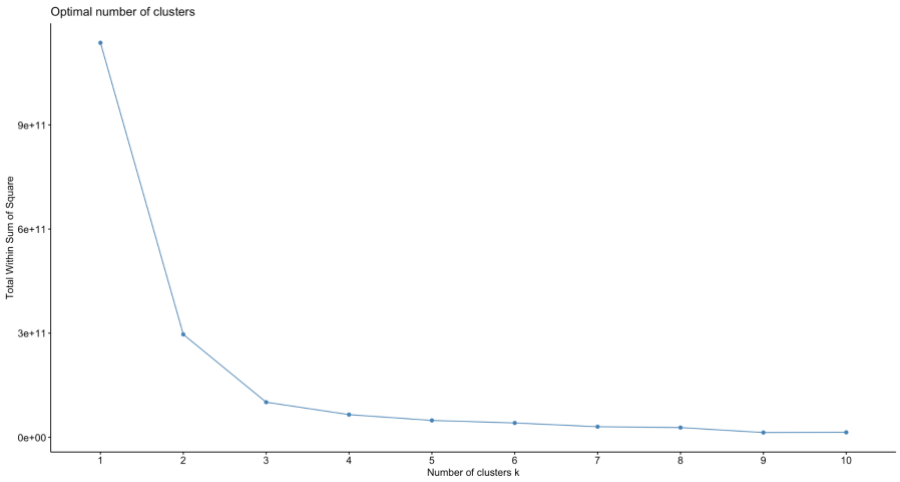


The above visualization represents the variance within the clusters for each number of clusters. The variance decreases as the number of clusters increases. Since there is a bend at 3 for the number of clusters, we choose 3 as the optimal number of clusters for K-means.

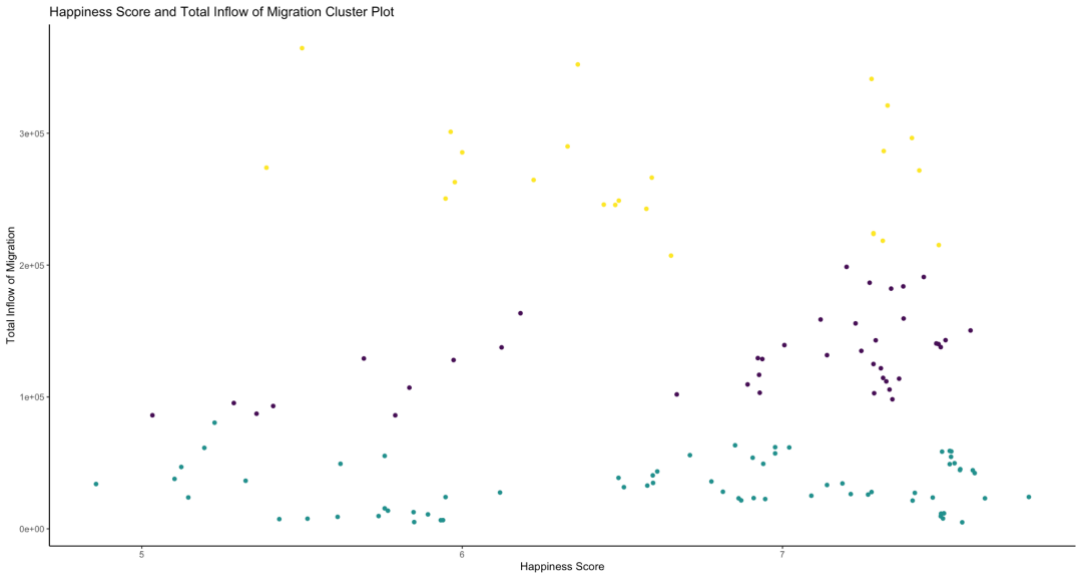


The above visualization is plotted using K-means clustering. We can see there are three clusters: (1) low total inflow of migration on the full range of happiness scores, (2) 250,000 to 750,000 total inflow of migration on the range of happiness score from 5.4 to 7.6, and (3) high

total inflow of migration on happiness score around 7. By looking at this visualization, we suspected that outliers might be influencing the K-means clustering since there are a few points where the total inflow of migration is greater than 1,500,000. Thus, we search the outliers using the mvoutlier function and remove them in r. The following are the plots we created after removing the outliers.



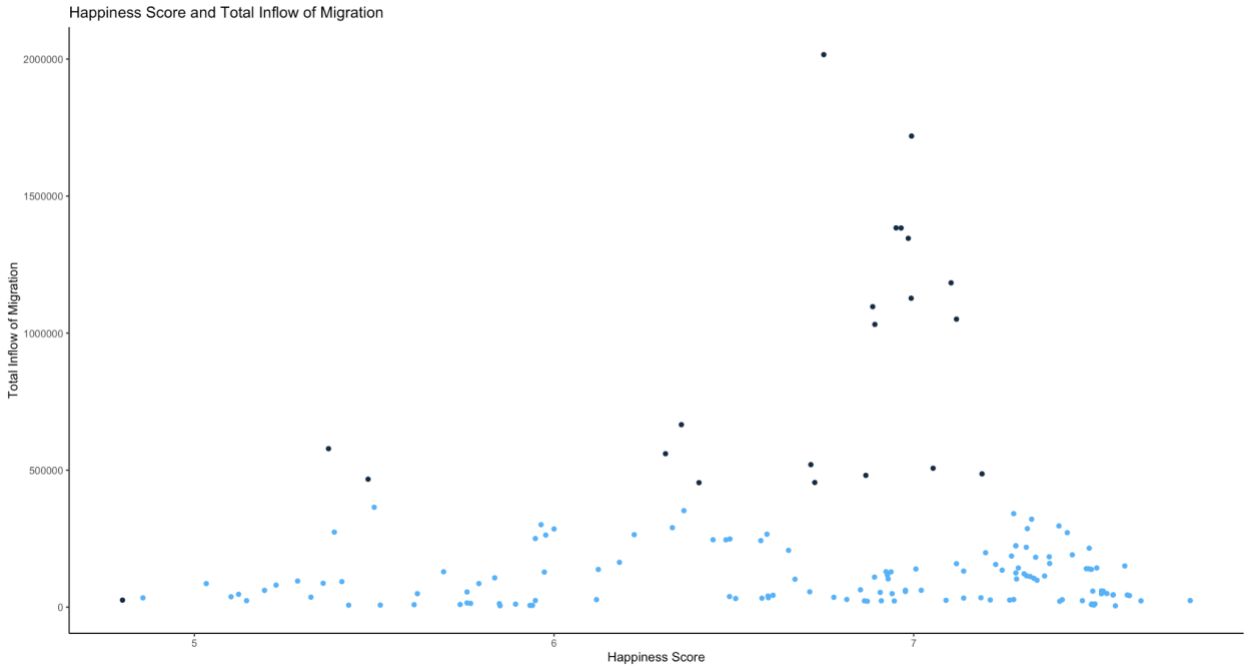
Interestingly, the optimal number of clusters is the same as the one with outliers.



The above visualization is plotted using K-means clustering. We can see that the clusters do not seem to be spread out as much as the first cluster analysis graph since we removed the outliers.

These two cluster plots show that cluster plots are sensitive to outliers. For instance, after removing outliers for the total inflow of migration greater than 350,000, the third cluster is less spread out from the other two. Although it can be interesting to analyze the variables' relationships within each cluster group, it will be harmful even if we use the wrong number of clusters or include outliers during the cluster analysis.

4.7. Outlier Analysis



The above visualization shows the outlier analysis of the Happiness Score and Total Inflow of Migration. The outliers are shown in dark blue colors. We can see that most of the outliers are around 500,000 total inflow of migration. However, it is interesting to see the one outlier at the lowest happiness score and total inflow of migration. Even though outlier analysis might not give much valuable information, knowing that outliers exist can help us to decide whether or not to remove them. Removing outliers can be helpful to our visualizations since outliers can give us biased information if we try to find regression analysis.

4.8. Regression with Outliers

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.04127 -0.22926  0.01064  0.18144  0.90395

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.1949     0.4519   7.071 5.68e-11 ***
Generosity            1.7190     0.3218   5.342 3.39e-07 ***
GDP.per.Capita        0.4666     0.3016   1.547    0.124
Life.Expectancy       1.6666     0.3373   4.941 2.07e-06 ***
Freedom               1.3491     0.3098   4.355 2.47e-05 ***
Government.Corruption 1.7016     0.3415   4.983 1.73e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3734 on 148 degrees of freedom
Multiple R-squared:  0.768,     Adjusted R-squared:  0.7602
F-statistic:    98 on 5 and 148 DF,  p-value: < 2.2e-16
```

```
t test of coefficients:

                    Estimate Std. Error t value  Pr(>|t|)
(Intercept)          3.19493    0.51802  6.1676 6.292e-09 ***
Generosity           1.71897    0.35882  4.7906 4.001e-06 ***
GDP_per_Capita       0.46659    0.35039  1.3316 0.1850315
Life_Expectancy      1.66665    0.28981  5.7508 4.917e-08 ***
Freedom              1.34906    0.34586  3.9006 0.0001452 ***
Government_Corruption 1.70162   0.40665  4.1845 4.881e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

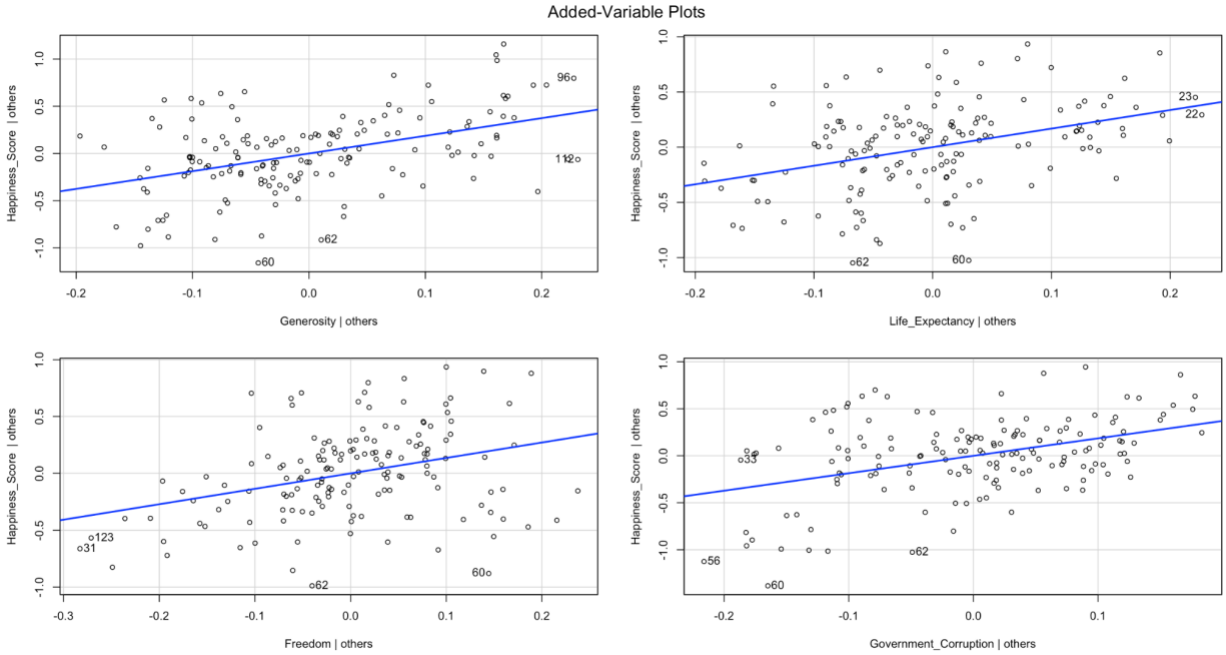
The result of regression analysis with outliers shows that all the explanatory variables, except for GDP per capita, are statistically significant. The figure on the right shows the heteroskedastic robust standard errors. Generosity, life expectancy, freedom, and government corruption score all have a positive effect on the happiness score. With every unit increase in generosity, life expectancy, freedom, and government corruption score, the happiness score will increase 1.72 units, 1.67 units, 1.35 units, and 1.70 units respectively. R-squared is 0.77 which means that 77% of the variations for Happiness Score can be explained by the regressors. Since GDP per capita is not significant, we remove GDP per capita from the regression and obtain the following result.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.07716 -0.22462  0.01484  0.20174  0.84568

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.7308     0.2916  12.796  < 2e-16 ***
Generosity            1.8745     0.3071   6.104 8.55e-09 ***
Life_Expectancy       1.6821     0.3387   4.966 1.85e-06 ***
Freedom               1.3581     0.3111   4.365 2.37e-05 ***
Government_Corruption 1.8571     0.3279   5.663 7.41e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3751 on 149 degrees of freedom
Multiple R-squared:  0.7643,    Adjusted R-squared:  0.7579
F-statistic: 120.8 on 4 and 149 DF,  p-value: < 2.2e-16
```

```
t test of coefficients:

                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)           3.73085    0.26391 14.1367 < 2.2e-16 ***
Generosity            1.87449    0.34931  5.3662 3.014e-07 ***
Life_Expectancy       1.68208    0.28875  5.8254 3.387e-08 ***
Freedom               1.35809    0.35171  3.8614 0.0001676 ***
Government_Corruption 1.85709    0.38845  4.7807 4.153e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

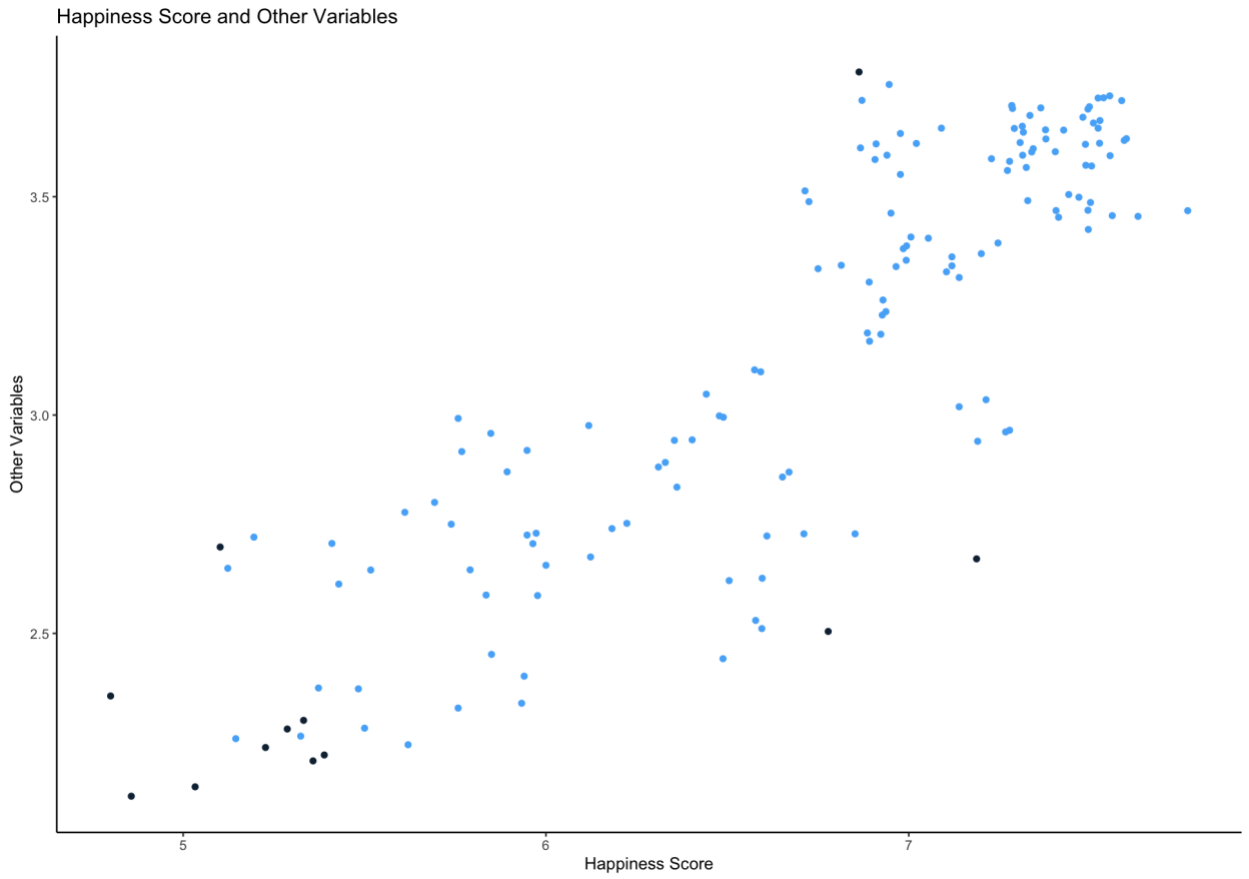
All the regression coefficients are statistically significant after we removed GDP per capita. With every unit increase in generosity, life expectancy, freedom, and government corruption score, the happiness score will increase 1.87 units, 1.68 units, 1.36 units, and 1.86 units

respectively. R-squared decreases to 0.76 from 0.77 after we removed GDP per capita. We plotted the multiple regression using avplots since we have 5 regressors.



The x-axis represents a single predictor variable and the y-axis represents a response variable. The blue line is the relationship between the predictor and the response variable while holding other variables constant. The points labels in each plot are representing two observations with the largest residuals and two observations with the largest partial leverage. If our clients want us to predict the happiness score, we will use the second regression without the GDP per capita variable.

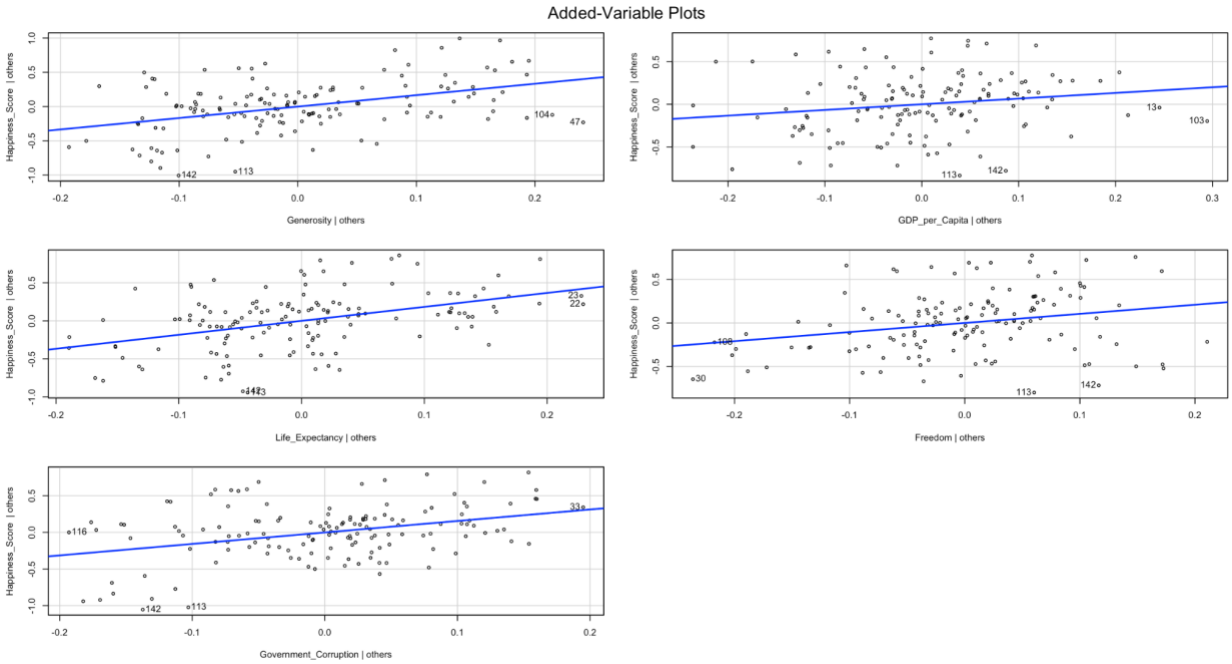
4.9. Regression without Outliers



Using mvoutlier in R, we plotted the outliers in the multiple regression model from section 4.8. The outliers are in dark blue. Since we suspect that these outliers might be affecting the regression calculation, we recalculated the regressions after removing outliers.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.86273 -0.20103  0.00269  0.19050  0.76778

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.9901     0.4839   6.179 7.01e-09 ***
Generosity           1.6704     0.3026   5.520 1.66e-07 ***
GDP_per_Capita       0.6666     0.3148   2.117  0.03606 *
Life_Expectancy      1.8363     0.3288   5.585 1.23e-07 ***
Freedom              1.0475     0.3484   3.006  0.00315 **
Government_Corruption   1.5717     0.3421   4.595 9.77e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3457 on 136 degrees of freedom
Multiple R-squared:  0.7534,    Adjusted R-squared:  0.7444
F-statistic: 83.11 on 5 and 136 DF,  p-value: < 2.2e-16
```

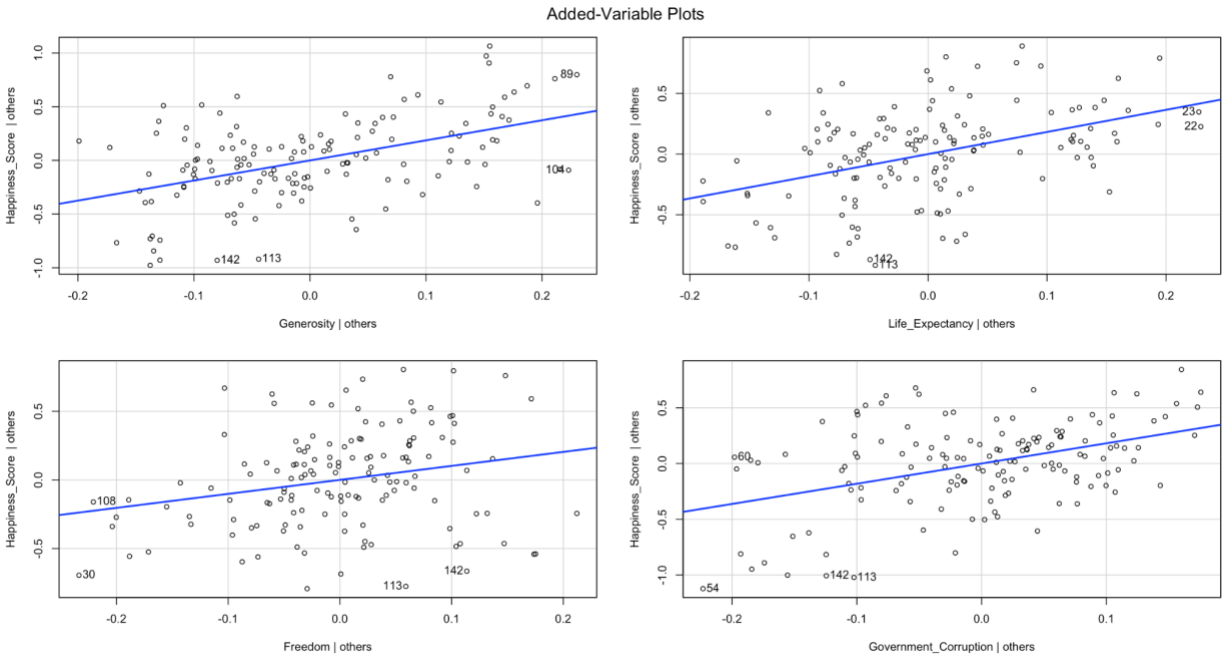
```
t test of coefficients:

                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)           2.99006    0.49431  6.0489 1.330e-08 ***
Generosity            1.67042    0.35396  4.7192 5.809e-06 ***
GDP_per_Capita        0.66659    0.32742  2.0359 0.0437042 *
Life_Expectancy       1.83632    0.29126  6.3048 3.755e-09 ***
Freedom               1.04748    0.39010  2.6852 0.0081511 **
Government_Corruption  1.57171    0.40585  3.8727 0.0001666 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



GDP per capita's coefficient becomes statistically significant after outliers are removed, but freedom's coefficient becomes less significant. All regressors still have a positive effect on the happiness score. With every unit increase in generosity, GDP per capita, life expectancy, freedom, and government corruption score, the happiness score will increase 1.67 units, 0.66 units, 1.84 units, 1.05 units, and 1.57 units respectively. R-squared is 0.75 which is a 0.02 decrease from the previous model with outliers.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.83668 -0.19145  0.01907  0.21209  0.77411

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.8057     0.2965  12.836  < 2e-16 ***
Generosity            1.8720     0.2909   6.436 1.91e-09 ***
Life_Expectancy       1.8285     0.3329   5.492 1.87e-07 ***
Freedom               1.0202     0.3526   2.894  0.00443 **
Government_Corruption 1.8136     0.3265   5.555 1.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.35 on 137 degrees of freedom
Multiple R-squared:  0.7453,    Adjusted R-squared:  0.7379
F-statistic: 100.2 on 4 and 137 DF,  p-value: < 2.2e-16
```

```
t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)           3.80575    0.27422 13.8786 < 2.2e-16 ***
Generosity            1.87201    0.34392  5.4432 2.344e-07 ***
Life_Expectancy       1.82852    0.29032  6.2982 3.820e-09 ***
Freedom               1.02022    0.39643  2.5736   0.01113 *
Government_Corruption 1.81364    0.39017  4.6484 7.778e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



To compare with the multiple regression model with GDP per capita removed in section 4.8, we calculated the new regression line. After GDP per capita is removed, freedom's coefficient is only significant on the 5% significance level. Thus, removing GDP per capita is adding the omitted variable bias in the multiple regression model. R-squared is 0.75. We need to be careful to check the significance of the coefficients after we remove the outliers. As we can see from the regression, focusing on the significance of the coefficients reduces the R-squared of the model. However, we concluded that we should use the regression with GDP per capita after we remove the outliers since the changes in R-square are minimal.

5. Business Reflections

The visualizations in section 4 offer meaningful insights that could help our clients when they are deciding to migrate. Using these visualizations, we could recommend some countries that have the most happiness scores. However, after we analyze the different variables in the dataset, we cannot guarantee that even the countries with the highest happiness score will satisfy our clients' desires. For example, even though Northern European Countries rank the highest among all countries, we still offer our clients different destinations for them to choose. Every individual is different so we would offer personal tailored service for our clients to match their needs for them to choose the best migration destinations. For us to tailor our recommendations to each client, we will need to find and incorporate more data that could explain the inherent similarities or differences between countries such as religion, cuisine, languages, and so on. By incorporating these data and using advanced analytics, we can provide analysis tailored to our clients. We might also be able to inform our clients if they might face potential discrimination due to our analysis of the data. Nevertheless, it would be the clients' decision to choose the country that they think would best fit their needs.

6. Reflection About the Process

The journey we took to reach this final report felt long and short at the same time since we were overwhelmed by the amount of learning to be able to create these visualizations and interpret them meaningfully. Tableau Prep was easy to use as it has a user interface that we can easily manipulate. We found it easy to integrate and reshape the data since the dataset are clearly displayed. However, we discovered that some of the recommendations suggested by Tableau Prep made our data worse than it was before we click it. For instance, it would group Austria and

Australia together even though they are different countries. Furthermore, we have to manually look through the data in Tableau Prep so the data might not be as clean as we think. Thus, being able to use R to further confirm the cleanliness of our dataset was helpful even though learning R was challenging. It was also helpful to have others look at the visualizations to check if their interpretation is the same as the information we want to present in the visualizations.

After working with the data from World Happiness Report and International Migration, we seldom felt perplexed by the results of our analysis. For instance, from the histogram and facet plots, we learned that the number of countries with higher happiness scores is decreasing. However, we are not sure what is causing this decrease in happiness. We also learned that the difference between men and women employment rate is large even though the data is only on the foreign-born population employment rate. These results somehow make us wonder if the world is indeed reaching to a point where equality is difficult to achieve. We should have further looked into the employment rates for countries with high happiness scores. Unfortunately, we did not have enough time to do the analysis.

Overall, although the journey was challenging, it was worthwhile to learn new skills and have new insights. We started this journey with the World Happiness Report so that we can analyze countries with high happiness scores and the factors that contribute to them. However, we end our journey by recognizing the flaws in this world. It was quite a humble experience to learn and analyze these datasets. As a team, we hope we at least call attention to some of these problems even though we were mainly focused on finding ways to help our clients.