Analysis of World Happiness Report

Group Members	Neptun code
ZHAO WEIYI	FVG8Z9
GU YIJIA	L6AMMN
WANG KE	ESQEM4
XIONG ZIJIAN	H5WEKI

1. Introduction

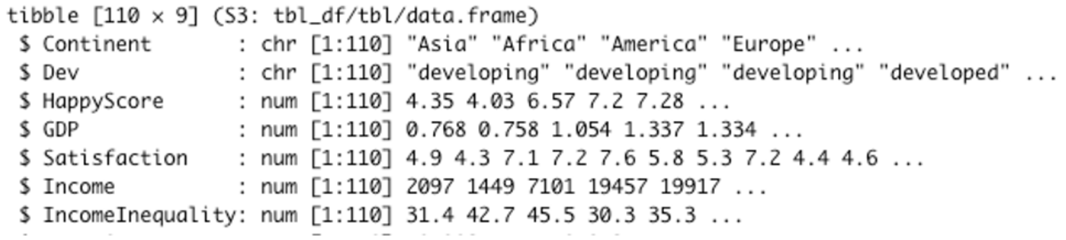
This database was found from Kaggle and it includes 110 observations. For the variables, it has Continent, Dev, HappyScore, GDP, Satisfaction, Income, and IncomeInequality. What is more, to find what kind of factors can affect the final happy score of people, the necessary statistical tests such as basic descriptive statistics for numeric variables, frequencies and figures for qualitative variables, and so on will be shown. After that, we try to find the relationship between different variables, so the confidence interval estimation for variables and analysis of the relationship between variables will be introduced as well. Finally, three rational and possible hypotheses will be given to test the reliability. To prove the hypothesis, we will use R to check and after that, the conclusion will be done in this essay.

2. Dataset Description

2.1 Short description about variables, type of variables, scales of measurement

This dataset includes Continent, Dev, HappyScore, GDP, Satisfaction, Income, and IncomeInequality. It is worth mentioning that for the unit of measurement, the variable so-called Dev simply means developing or developed, HappyScore is “the score of Happiness” measured in points. Satisfaction represents “the score of satisfaction towards the whole society” measured in points. Income means “annual income per capita” measured in dollars. GDP represents “GDP growth rate” measured in percent and Income inequality measured in“ GINI Index”.

Table 1 Data type of variables



```
tibble [110 × 9] (S3: tbl_df/tbl/data.frame)
 $ Continent       : chr [1:110] "Asia" "Africa" "America" "Europe" ...
 $ Dev             : chr [1:110] "developing" "developing" "developing" "developed" ...
 $ HappyScore      : num [1:110] 4.35 4.03 6.57 7.2 7.28 ...
 $ GDP             : num [1:110] 0.768 0.758 1.054 1.337 1.334 ...
 $ Satisfaction    : num [1:110] 4.9 4.3 7.1 7.2 7.6 5.8 5.3 7.2 4.4 4.6 ...
 $ Income          : num [1:110] 2097 1449 7101 19457 19917 ...
 $ IncomeInequality: num [1:110] 31.4 42.7 45.5 30.3 35.3 ...
```

(Source: author's own work based on R Studio)

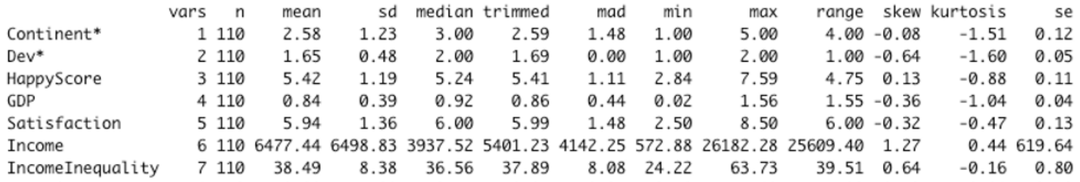
From table 1 we can check the data type of these 7 variables. For "Continent" and "Dev", it is obviously qualitative variables. Thus, R shows character. For the rest 5 variables, all of them are numeric variables. Also, it is worth introducing the scales of measurement. " Continent" and "Dev" are names used to identify an attribute of the element. Thus, "Continent" and "Dev" are nominal. However, for the variable "HappyScore", "GDP", "Satisfaction", "Income", and "IncomeInequality", the zero point of all these variables represent that nothing exists. In other words, the minimum value of all of these variables is the smallest, it cannot be negative. Therefore, variable "HappyScore", "GDP", "Satisfaction", "Income", and "IncomeInequality" are ratio scale of measurement.

2.2 Issue of missing data, imputation, outlier, etc.

In the process of data sorting, we did not find some values like zero values, NA values, and so on. Therefore, we do not need to make a kind of imputation in this case. However, for the outliers, the basic test for it should be done. In other words, first of all, it is necessary to check the outliers are exist or not. If it exists, we need to remove them to guarantee the accuracy of the analysis of the data, then we are able to use box plot to check twice if it is necessary. If not, we only need to focus on the rest part of testing.

The method we find the outlier is compare the difference between the mean and trimmed. Basically, if there is a huge difference between these two numbers, then it represents there is an outlier. After comparing these values, we found there is a big difference mean and trimmed in variable Income in Table 2. In more detail, the mean of Income is 6477.44 and the trimmed of it is 5401.23. Thus, there must be one or more outliers in the variable "Income".

Table 2 Find the outliers

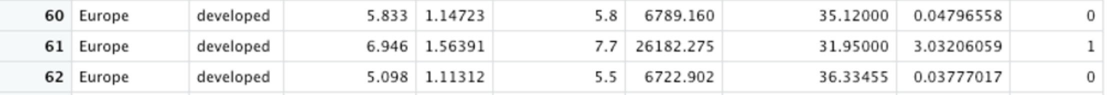


```
                  vars   n    mean      sd  median trimmed     mad    min      max    range  skew kurtosis     se
Continent*           1 110    2.58    1.23    3.00    2.59    1.48   1.00     5.00     4.00 -0.08    -1.51   0.12
Dev*                 2 110    1.65    0.48    2.00    1.69    0.00   1.00     2.00     1.00 -0.64    -1.60   0.05
HappyScore           3 110    5.42    1.19    5.24    5.41    1.11   2.84     7.59     4.75  0.13    -0.88   0.11
GDP                  4 110    0.84    0.39    0.92    0.86    0.44   0.02     1.56     1.55 -0.36    -1.04   0.04
Satisfaction         5 110    5.94    1.36    6.00    5.99    1.48   2.50     8.50     6.00 -0.32    -0.47   0.13
Income               6 110 6477.44 6498.83 3937.52 5401.23 4142.25 572.88 26182.28 25609.40  1.27     0.44 619.64
IncomeInequality     7 110   38.49    8.38   36.56   37.89    8.08  24.22    63.73    39.51  0.64    -0.16   0.80
```

(Source: author's own work based on R Studio)

After that, it is necessary to detect and remove the outlier. First of all, we need to calculated the z-score, if it is less than 3, then it belongs to normal value category. However, if z-score is greater than 3, then it can be considered as an outlier. In other words, we try to give a brand a new meaning to "PW$outlier", which means the outlier in the dataset PW. And then use "if else" function to define the outlier equal to 1 and normal value equal to 0. If we run these functions and check the dataset immediately, we can find that there is a new column so called "Inocme_z" had been added into the dataset. And then we can view the dataset PW, table 4 indicates that there is an outlier in the row 61. Here, row 61 represents the country so-called Luxembourg.

Table 3 Detecting the outliers



60	Europe	developed	5.833	1.14723	5.8	6789.160	35.12000	0.04796558	0
61	Europe	developed	6.946	1.56391	7.7	26182.275	31.95000	3.03206059	1
62	Europe	developed	5.098	1.11312	5.5	6722.902	36.33455	0.03777017	0

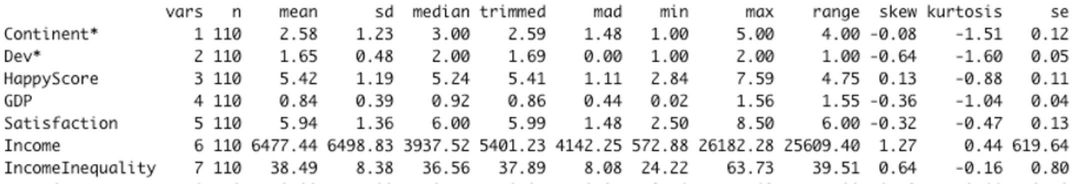
(Source: author's own work based on R Studio)

As we mentioned before, to keep the accuracy of the data, we must remove the outlier from the original dataset. However, it is only show which one is the outlier until now. In order to get a clear and clean dataset, we collect all the non-outlier values to a new dataset, which named "PW1". Until now, the dataset which include all non-outlier values is made successfully.

2.3 Descriptive statistics for numeric variables

In order to get the more meaningful information, we need to check the mean, median and other important descriptive statistics for numeric variables. We are able to use R to check all of these information by function “psych::describe()”, but before that the package should be installed into R.

Table 4 Descriptive statistics



```
                 vars   n    mean      sd  median trimmed     mad    min      max    range  skew kurtosis     se
Continent*          1 110    2.58    1.23    3.00    2.59    1.48   1.00     5.00     4.00 -0.08    -1.51   0.12
Dev*                2 110    1.65    0.48    2.00    1.69    0.00   1.00     2.00     1.00 -0.64    -1.60   0.05
HappyScore          3 110    5.42    1.19    5.24    5.41    1.11   2.84     7.59     4.75  0.13    -0.88   0.11
GDP                 4 110    0.84    0.39    0.92    0.86    0.44   0.02     1.56     1.55 -0.36    -1.04   0.04
Satisfaction        5 110    5.94    1.36    6.00    5.99    1.48   2.50     8.50     6.00 -0.32    -0.47   0.13
Income              6 110 6477.44 6498.83 3937.52 5401.23 4142.25 572.88 26182.28 25609.40  1.27     0.44 619.64
IncomeInequality    7 110   38.49    8.38   36.56   37.89    8.08  24.22    63.73    39.51  0.64    -0.16   0.80
```

(Source: author’s own work based on R Studio)

Table 4 indicates the descriptive statistics for numeric variables in very detail. Here we think it is worth introducing the meaning of the skewness and kurtosis. The skewness of variable HappyScore is 0.13, which is between -0.5 and 0.5, the data is fairly symmetrical. The kurtosis of variable HappyScore is -0.88, which is less than 3, so we are able to consider that the frequency distribution of the variable is relatively scattered on both sides of the mode. For variable GDP, the skewness is-0.36, which is between -0.5 and 0.5 as well. Therefore, we can say the data is fairly symmetrical. The kurtosis value is -1.04, which is less than 3. This means the distribution is shorter, tails are thinner than the normal distribution. The skewness of variable Satisfaction is -0.32, which is between -0.5 and 0.5. Thus, the situation is similar as before. The data is almost fairly symmetrical. The kurtosis of variable Satisfaction is -0.47, which is less than 3. Actually, for the variables rest, all kurtosis of them is less than 3. For the variable Income, the situation of skewness is different. The value is 1.27, which is greater than 0.5. Thus, it represents this data is positively skewed, the data are moderately skewed. Similarly, the skewness of “IncomeInequality” is 0.64, which is greater than 0.5 as well. So it can be considered as a kind of positively skewed.

3. Analysis of results

3.1 Ratios, frequencies and figures for qualitative variables

After showing the descriptive statistics of all the numeric variables, we calculated the ratios and frequencies of qualitative variables and made the bar graph to show the level of development proportions within the ‘continent’ qualitative variable clearly.

The tables below show the frequencies and ratios of the qualitative variable ‘continent’, ‘dev’. We combine these two together to make it more easily to find the differences of these different categories in the variables.

Table 5 Frequency table for qualitative variables

Frequencies of “Continent” variable					
Group	*Africa*	*America*	*Asia*	*Europe*	*Oceania*
Frequency	*32*	*18*	*25*	*33*	*1*

Frequencies of “Dev” variable		
Group	*Developed countries*	*Developing countries*
Frequency	*38*	*72*

(Source: author’s own work based on R Studio)

Table 6 Proportion for qualitative variables

Proportion of each continent	
Africa	*0.294*
America	*0.165*
Asia	*0.229*
Europe	*0.303*
Oceania	*0.009*
Proportion of “Developed” or “Developing” groups	
Developed	*0.339*
Developing	*0.661*

(Source: author’s own work based on R Studio)

For the level of development countries, we can find the number of developing countries is nearly twice than the developed countries. If we focus on the number of countries in the continent, we can find the most biggest amount is Europe. Africa has two lower countries than Europe. And in this dataset, we find there is only one country (Australia) contained in the Oceania.

In the below graph we made, we can obviously find that all of the European countries and Oceania in the ‘Happy’ dataset are developed countries. However, for America and

Asia continents, there are only several developed countries. And there is even no one developed country in Africa.

Looking at table 7, we calculated the mean of every continent, and we can find that: mean (happy score of Africa) is 4.23, mean (happy score of America) is 6.34, mean (happy score of Asia) is 5.15, mean (happy score of Europe) is 6.17, mean (happy score of Oceania) is 7.284.

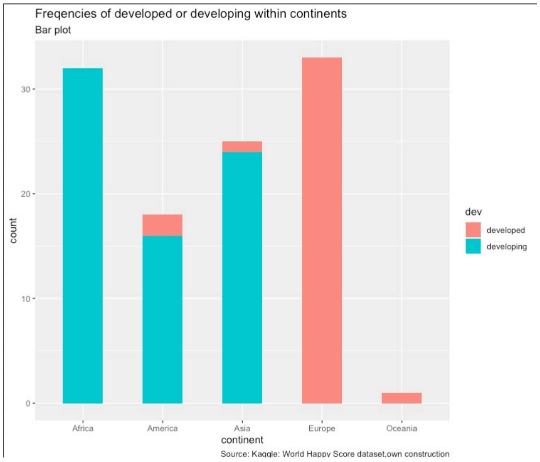
For these two developed continents, Oceania has the highest happy score, and Europe has the relatively higher happy score though it is lower than America. However, for Africa (has no developed countries), it has the lowest happy score. So there is the possibility that it exists the positive relationship between the level of development of continents and their happy scores. The higher the development, the happier they are. Later we will find if it is true through other tests.

Table 7 Mean of every continent

Mean value of happiness score for each continent	
Africa	*4.239*
America	*6.341*
Asia	*5.15*
Europe	*6.166*
Oceania	*7.284*

(Source: author's own work based on R Studio)

Figure 1 Frequencies of "Dev" within "Continents"



(Source: author's own work based on R Studio)

3.2 Confidence interval estimation for variables

We cannot get the data of all countries in the world, so we find the representative countries as more as possible and use the statistical results of the sample to predict the overall result. However, because the sample is not the whole population and we get different results from different samples, so there is the error interval between the sample and the whole population, and we want to test if there is 95% probability of confidence interval containing the sample of population. In other words, we want to know if we can be 95% about the true mean of population in the confidence interval.

3.2.1 For happy scores

Table 8 Confidence interval estimation for happy scores

Confidence interval estimation for happy scores	
Lower limit bound	*5.184*
Upper limit bound	*5.628*

(Source: author's own work based on R Studio)

Table 9 Confidence interval stratified by continents

Confidence interval estimation for happy scores in case of each continent		
	Lower limit bound	Upper limit bound
Africa	4.005	4.473
America	5.964	6.717
Asia	4.839	5.462
Europe	5.816	6.515
Oceania	NA	NA

(Source: author's own work based on R Studio)

Table 8 helps us calculate the happy scores confidence interval [5.184, 5.628], which concludes the mean of 5.42. Therefore, we could say that there is a 95% probability the interval contains the sample of the population. We are 95% confident that the average happy score is between 5.184 and 5.628. The random interval has a 95% probability of containing the true value μ, so the results we get later have higher reliability.

After that, we calculate the more specific confidence intervals stratified by the 'continent' variable. We can find in table 9, the continent Africa, America, Asia, and Europe all have 95% confidence intervals. But there is no confidence interval value for

Oceania because there is the point estimate for the only country Australia of continent Oceania in this dataset. We can conclude that we are 95% confident that the mean value of happy score of each continent is in its own confidence interval.

3.2.2 For GDP

Table 10 Confidence interval estimation for GDP

Confidence interval estimation for GDP	
Lower limit bound	*0.763*
Upper limit bound	*0.908*

(Source: author's own work based on R Studio)

The confidence interval of GDP is [0.763 ,0.908], which contains the GDP mean 0.84. Therefore, there is the 95% confidence interval that contains the sample of population. We are 95% confident that the average GDP growth rate is between 0.763% and 0.908%. The random interval has a 95% probability of containing the true value μ, so the results we get later has the higher reliability.

3.2.3 For income

Table 11 Confidence interval estimation for income

Confidence interval estimation for income	
Lower limit bound	*5124.321*
Upper limit bound	*7469.003*

(Source: author's own work based on R Studio)

There is one outlier which the z-value is higher than 3, so we screen it out before analyzing. The confidence interval of income is [5124.321, 7469.003], which contains the income mean 6477.44. Therefore, there is the 95% confidence interval that contains the sample of population. We are 95% confident that the average annual income per capita is between 5124.321 dollars and 7469.003 dollars. The random interval has a 95% probability of containing the true value μ, so the results we get later has the higher reliability.

3.2.4 For Satisfaction

Table 12 Confidence interval estimation for satisfaction scores

Confidence interval estimation for satisfaction scores	
Lower limit bound	*5.668*
Upper limit bound	*6.178*

(Source: author's own work based on R Studio)

The confidence interval for satisfaction scores is [5.668, 6.178], which contains the mean value 5.94. Therefore, there is the 95% confidence interval that contains the sample of population. We are 95% confident that the average satisfaction score is between 5.668 and 6.178. The random interval has a 95% probability of containing the true value μ, so the results we get later has the higher reliability.

3.2.5 For Income Inequality

Table 13 Confidence interval estimation for income inequality

Confidence interval estimation for income inequality	
Lower limit bound	*36.979*
Upper limit bound	*40.129*

(Source: author's own work based on R Studio)

The confidence interval of Income Inequality is [36.979, 40.129], which contains the mean of income inequality 38.49. Therefore, there is the 95% confidence interval that contains the sample of population. We are 95% confident that the income inequality is between 36.979 and 40.129. The random interval has a 95% probability of containing the true value μ, so the results we get later has the higher reliability.

In short, we are 95% confident that the true mean of these variables' population is in the confidence interval. And the results we obtain later will be reliable.

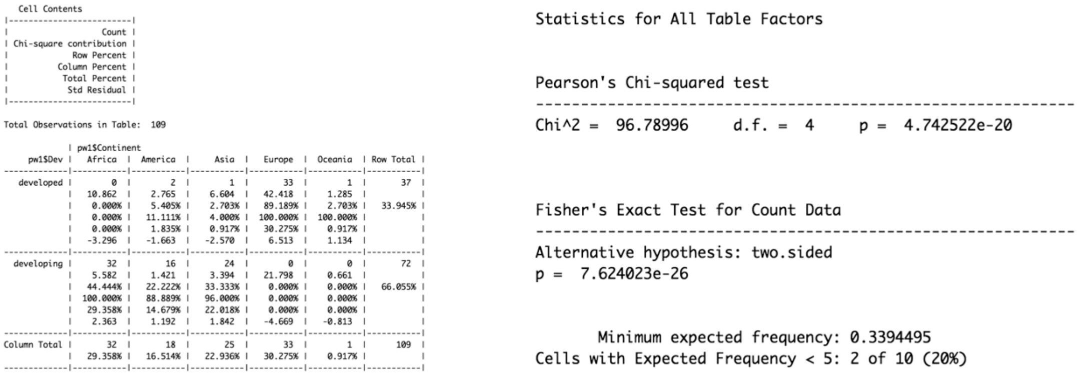
3.3 Analysis of relation between variables

3.3.1 Analysis of Cross Table

From this part, we will do several different types of analysis, and find the relation between variables. Firstly, we can do Chi-squared test from a cross table, which is an analysis of the correlation between two qualitative variables. In the case of the score of happiness, we have two qualitative variables, one is the type of continent, and the other is the degree of development of a country. In this project, the value of the Chi-square statistic is approximately 96.79. The p-value appears in the same row approaches to zero. The result is significant if this value is equal to or less than the designated alpha level, which is 5%. And in this project, the p-value is smaller than the standard alpha value, so we would reject the null hypothesis that the two qualitative variables are independent of each other. To put it simply, the result is significant, and the data suggests that the variables “Continent” and “Degree of development of a country” are associated with each other.

Fisher’s exact test is a way to determine if there are nonrandom associations between two qualitative variables as well. From the table above, the p-value of Fisher’s exact test is less than 5%. Therefore, in this case, there would be a statistically significant association between the two qualitative variables. Practically speaking, this situation is common in real lives, for instance, a European country is more likely to be a developed country, however, an African country tends to be less developed and is more likely to be a developing country.

Table 14 Cross table between two qualitative variables



```
   Cell Contents
|-------------------------|
|                   Count |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  109

             | pw1$Continent
     pw1$Dev |    Africa |   America |      Asia |    Europe |   Oceania | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
   developed |         0 |         2 |         1 |        33 |         1 |        37 |
             |    10.862 |     2.765 |     6.604 |    42.418 |     1.285 |           |
             |    0.000% |    5.405% |    2.703% |   89.189% |    2.703% |   33.945% |
             |    0.000% |   11.111% |    4.000% |  100.000% |  100.000% |           |
             |    0.000% |    1.835% |    0.917% |   30.275% |    0.917% |           |
             |    -3.296 |    -1.663 |    -2.570 |     6.513 |     1.134 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
  developing |        32 |        16 |        24 |         0 |         0 |        72 |
             |     5.582 |     1.421 |     3.394 |    21.798 |     0.661 |           |
             |   44.444% |   22.222% |   33.333% |    0.000% |    0.000% |   66.055% |
             |  100.000% |   88.889% |   96.000% |    0.000% |    0.000% |           |
             |   29.358% |   14.679% |   22.018% |    0.000% |    0.000% |           |
             |     2.363 |     1.192 |     1.842 |    -4.669 |    -0.813 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |        32 |        18 |        25 |        33 |         1 |       109 |
             |   29.358% |   16.514% |   22.936% |   30.275% |    0.917% |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
```

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  96.78996     d.f. =  4     p =  4.742522e-20



Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  7.624023e-26



        Minimum expected frequency: 0.3394495
Cells with Expected Frequency < 5: 2 of 10 (20%)
```

(Source: author’s own work based on R Studio)

Although we can get the association between continent variable and “degree of development of a country” variable from the cross table, Chi-square value is hard to interpret, mainly because the larger the table and the number of cases, the larger Chi-square becomes. Cramer’s V is a more interpretable measure of association, and its value is between 0 and 1. If Cramer’s V equals 0, then it indicates there is no association between two categorical variables. If the value is between 0 and 0.3, it indicates a weak association, if the value is between 0.3 and 0.7, then it means a moderate association, if the value is between 0.7 and 1, it indicates a strong association, and if the value is 1, then it means deterministic relationship. The Cramer's V formula is captured as follows

$$V = \sqrt{\frac{\chi^2}{n \min\left(p-1,\ q-1\right)}}$$

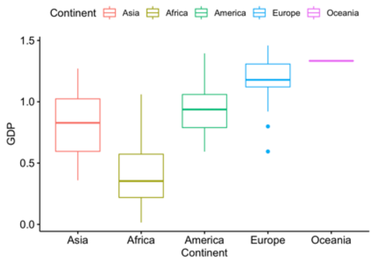
where "n" means the number of observations, "p" means the number of row, and "q" means the number of column. And after calculating the value of Cramer's V, we can get it is approximately 0.94. This value indicates there is a strong association between continent and the degree of development of a country.

3.3.2 Analysis of ANOVA table

Let us discuss why we use ANOVA. When doing some experiments, we usually divide the samples into different groups and give different treatments. After the experiment, we may be asked that are there any significant differences in the performance of these groups? At this point, we can use ANOVA to answer these questions and can be considered as a way to find the relation between a qualitative and a quantitative variable. If we want to obtain the variability of a distribution, for a more intuitive approach, we can read the range or interquartile range of a boxplot. If box plots for different groups seem identical, then they are likely to have very similar variances, or we can say there are no significant differences between groups.

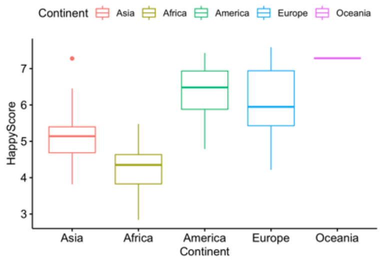
When having a look at figure 2, it is a boxplot comparing the five continents' groups and GDP growth rate. For all groups, the range and interquartile range are different from each other, which means it is likely that variances are different from each other, and the average GDP growth rate is significantly different between developing and developed countries. For table 15, the output "Pr(>F)" corresponds to the p-value of the test. As the p-value approaches zero and is less than the significance level of 5%, so we can reject the null hypothesis stating that the average values of GDP growth rate for all groups are the same. Now we can conclude that there are significant differences among the five continents.

Figure 2 Boxplot comparing continents and GDP growth rate



(Source: author's own work based on R Studio)

Figure 3 Boxplot comparing continents and happiness



(Source: author's own work based on R Studio)

Table 15 ANOVA table between continents and GDP growth rate

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Pw$Continent	4	9.998	2.4995	40.22	<2e-16
Residuals	105	6.525	0.0621	-	-

(Source: author's own work based on R Studio)

Let us consider the relationship between qualitative variable "Continent" with another quantitative variable "the score of happiness". Considering figure 3, for all the five groups, the range, and interquartile range are different from each other, which indicates the average values of happiness score is significantly different within the five continents. From the ANOVA table 16, the p-value is very small and is less than 5% significance level. Therefore, we can reject the null hypothesis that average values of happiness score within five continents are the same, and conclude that there are significant differences between five groups of continents.

Table 16 ANOVA table between continents and happiness

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Pw$Continent	4	83.52	20.88	32.18	<2e-16
Residuals	104	67.48	0.649	-	-

(Source: author's own work based on R Studio)

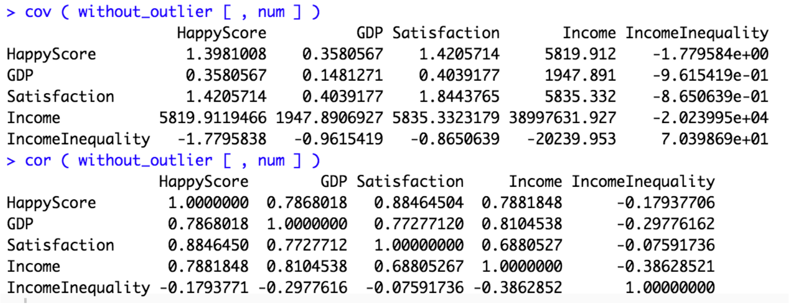
In this project, we have two qualitative variables and five quantitative variables in total. After conducting ANOVA tests of each quantitative with each qualitative respectively, we can get for each combination of quantitative and qualitative variable, the p-values of ANOVA tests are less than the significance level 5%, which means there are significant differences and among five continents.

3.3.3 Analysis of covariance and correlation

Analysis of correlation between quantitative variables contributes to the understanding of mass behavior. In this project, with the help of correlation, it is possible to have a correct idea of the determinants of happiness score.

The covariance is a measure of the linear association between two quantitative variables. Covariance is in the original units of the variables, and variables on scales with bigger numbers and with wider distributions will necessarily have bigger covariances. For example, "income" is measured in dollars, and "GDP growth rate" is measured in percent, thus the covariance value of "income" will be larger than the covariance value of "GDP growth rate". With positive values, covariances indicate a positive relationship, and on contrary, with negative values, covariances indicate a negative relationship. Covariance of a variable with itself is simply the variance. The table 17 is about the covariance matrix, and if we take a look at the first column, we can obtain happiness score's covariance with income inequality (-1.86) is negative. And other covariance values in the first column are positive, which indicates a positive relationship between happiness score with GDP growth rate, satisfaction score towards the society, and income. Practically speaking, the higher GDP per capita growth rate is, the higher the happiness score people will get. Similarly, the higher the average annual income per capita is, and the higher people's satisfaction with society is, then the higher the happiness score people will get. If income inequality in a country is serious, then the happiness score in this country will be low, which proves the negative relation.

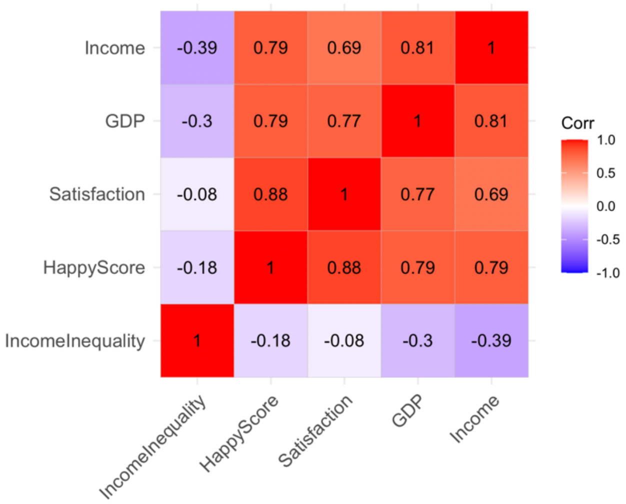
Table 17 Covariance matrix and correlation matrix



```
> cov ( without_outlier [ , num ] )
                   HappyScore          GDP Satisfaction       Income IncomeInequality
HappyScore          1.3981008    0.3580567    1.4205714     5819.912    -1.779584e+00
GDP                 0.3580567    0.1481271    0.4039177     1947.891    -9.615419e-01
Satisfaction        1.4205714    0.4039177    1.8443765     5835.332    -8.650639e-01
Income           5819.9119466 1947.8906927 5835.3323179 38997631.927    -2.023995e+04
IncomeInequality   -1.7795838   -0.9615419   -0.8650639   -20239.953     7.039869e+01
> cor ( without_outlier [ , num ] )
                 HappyScore        GDP Satisfaction     Income IncomeInequality
HappyScore        1.0000000  0.7868018   0.88464504  0.7881848      -0.17937706
GDP               0.7868018  1.0000000   0.77277120  0.8104538      -0.29776162
Satisfaction      0.8846450  0.7727712   1.00000000  0.6880527      -0.07591736
Income            0.7881848  0.8104538   0.68805267  1.0000000      -0.38628521
IncomeInequality -0.1793771 -0.2977616  -0.07591736 -0.3862852       1.00000000
```

(Source: author's own work based on R Studio)

Figure 4 Correlation heatmap



(Source: author's own work based on R Studio)

The correlation matrix is shown in table 17 as well, and to compute correlation, we divide the covariance by the standard deviation of both variables to remove units of measurement. Correlation is a measure of linear association and does not necessarily indicate causation. Correlation, unlike covariance, is constrained to being between -1 and 1. Values near -1 indicate a strong negative linear relationship, and values near 1 indicate a strong positive linear relationship. The figure above is the so-called correlation heatmap, which is a transformation of the correlation matrix and is a good figure to depict relation. Comparing the correlation matrix with the covariance matrix, the correlation matrix has the same relationship with the covariance matrix. The numbers on the heatmap are correlation coefficients. The correlation coefficient between “GDP growth rate” and “happiness score” is relatively high. Therefore, the positive relationship between them is relatively strong. Besides, we can get the relationship between GDP growth rate and income as well. The relationship between them is relatively high and approaches 1, which means there is a strong positive relationship between GDP growth rate and income. This relationship is common in practice, because if a person has more disposable income, then the private consumption will increase, and we know private consumption is part of GDP, thus the GDP growth rate will increase in an indirect way.

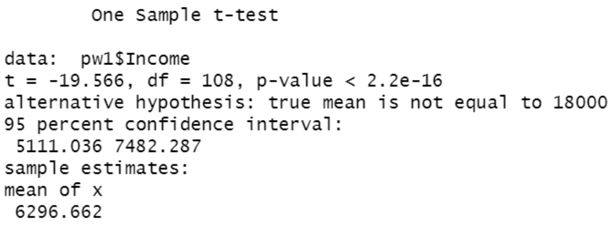
3.4 Hypothesis tests for variables

After analyzing different attributions of statistics, we could hold some hypothesis tests to check the quality, accuracy, proportion of variances and samples, some estimations made out of our mind, or simply just come from common senses. It is really important to test the hypothesis, with elements including null hypothesis (which is our pre-settled statement), alternative hypothesis, significant value (in this project work we always assume the significant level is suggested by R which is 5%), and p-value.

3.4.1 Mean test

Quoted from BBC news “Let's put the world's average salary - in dollars - of $1,480 a month, or almost $18,000 a year” (Ruth Alexander, 2012). Thus, we decide to test if we could get the same conclusion from our sample. The null statement should be “mean income for samples is 18000”.

Table 18 T-test for mean of income



```
        One Sample t-test

data:  pw1$Income
t = -19.566, df = 108, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 18000
95 percent confidence interval:
 5111.036 7482.287
sample estimates:
mean of x 
 6296.662 
```

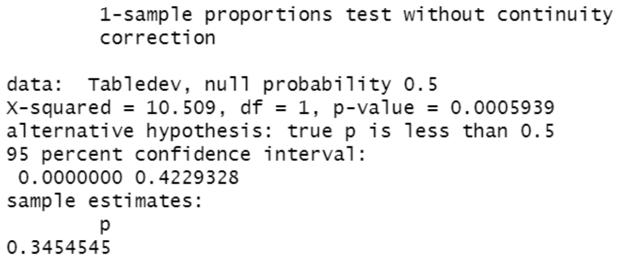
(Source: author’s own work based on R Studio)

It turns out we derive result as p-value is less than 2.2e-16, thus we reject this null statement. Moreover, the actual sample estimation on annual income is about 6296.662, which proves the decision we made.

3.4.2 Proportion test

Assume in a theoretical and ideal case, the expected proportion of developing countries' group is about 50%. Therefore, we have the proportion of developed countries should be 50%. Here we create a dummy first for conveniently holding group-by test later, which means if the country is a developing one, then it is recorded as 1, otherwise 0.

Table 19 Proportion test for developing countries



```
        1-sample proportions test without continuity
        correction

data:  Tabledev, null probability 0.5
X-squared = 10.509, df = 1, p-value = 0.0005939
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.4229328
sample estimates:
        p
0.3454545
```

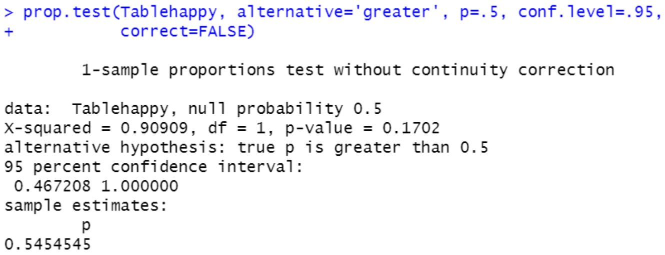
(Source: author's own work based on R Studio)

It comes out the p-value is very small and less than 5%, which means we need to reject the null hypothesis. The estimated proportion for developed countries is 34.5% according to the output of R.

Creating a dummy variable is an essential part in ratio and proportion hypothesis test. By defining "1" and "0" for some values. We could get counting tools to calculate samples that have different levels on certain things. For example, we divide the "Happy score" variable into two groups, one is for the happy score value lower than the average score (5.42), and the other is for the happy score is higher than average.

To test if the samples of low scores and high scores hold balanced volumes, here we use the proportion test as well. The null hypothesis, in this case, should be the true proportion that the happiness score above 5.42 is less or equal to 50%. The alternative hypothesis should be: the true proportion is more than 50%. It turns out the p-value is about 0.17, which is greater than 5%, thus we fail to reject the null hypothesis. The following output gives us an estimation that the proportion should be around 54.54%.

Table 20 Proportion test on happy scores



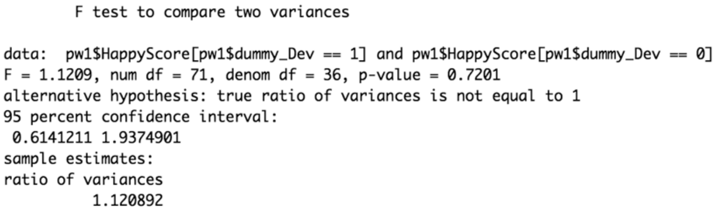
```
> prop.test(Tablehappy, alternative='greater', p=.5, conf.level=.95,
+           correct=FALSE)

	1-sample proportions test without continuity correction

data:  Tablehappy, null probability 0.5
X-squared = 0.90909, df = 1, p-value = 0.1702
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.467208 1.000000
sample estimates:
        p 
0.5454545 
```

(Source: author's own work based on R Studio)

3.4.3 Two independent samples T test

In this part, we will do two independent samples T-test. First, we set a statement that the average happiness score of developed and developing countries should be the same, and then we will test if this statement is true or not. Before conducting the T-test, we will check whether the two variances from the two samples are equal or not. If they are equal, then we can use an equal variance version of the T-test, otherwise, we will use the not equal variance version of the T-test. From table 21, we can see that the p-value is greater than 5%, which means we can't reject the null hypothesis that the two samples' variances are equal. Therefore, in the following T-test, we will use the equal variance version of the T-test.

Table 21 Two samples variance test

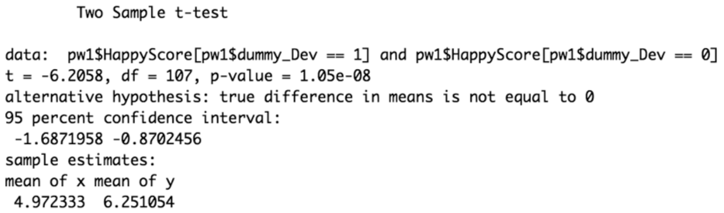


```
	F test to compare two variances

data:  pw1$HappyScore[pw1$dummy_Dev == 1] and pw1$HappyScore[pw1$dummy_Dev == 0]
F = 1.1209, num df = 71, denom df = 36, p-value = 0.7201
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.6141211 1.9374901
sample estimates:
ratio of variances 
          1.120892 
```

(Source: author's own work based on R Studio)

In the previous part, we know we need to use the equal variance version of the T-test. When doing such a test, we set the null hypothesis that the difference between the average happiness score of the developing countries' group and the average happiness score of the developed countries' group is 0. From table 22, the p-value is less than 5%, which means we need to reject the null hypothesis we just set. And we can get the conclusion that the average happiness score in developed countries' group and developing countries' group is different. The last two rows in table 22 prove the decision we made. This phenomenon is common in our daily life because people in developed countries tend to have a high income, which means the happiness score in developed countries tends to be high.

Table 22 Two independent samples T- test

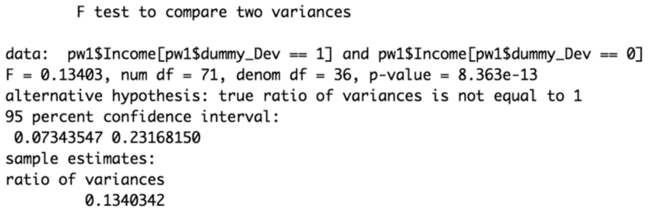


```
          Two Sample t-test

data:  pw1$HappyScore[pw1$dummy_Dev == 1] and pw1$HappyScore[pw1$dummy_Dev == 0]
t = -6.2058, df = 107, p-value = 1.05e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6871958 -0.8702456
sample estimates:
mean of x mean of y 
 4.972333  6.251054 
```

(Source: author's own work based on R Studio)

Similarly, we can do the same test for average income in developing countries' group and in developed countries' group. First, we will test the two samples' variances are equal or not. From table 23, we can get the p-value is less than 5% and approaches 0, which means we need to reject the null hypothesis that the two samples' variance are equal, thus, we will use the not equal variance version of the T-test.

Table 23 Two samples variance test

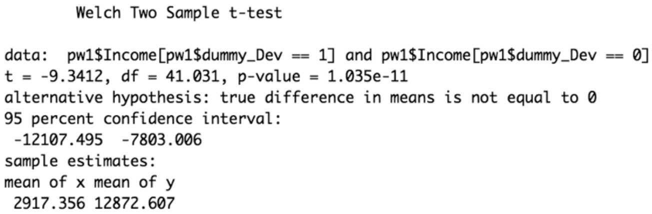


```
          F test to compare two variances

data:  pw1$Income[pw1$dummy_Dev == 1] and pw1$Income[pw1$dummy_Dev == 0]
F = 0.13403, num df = 71, denom df = 36, p-value = 8.363e-13
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.07343547 0.23168150
sample estimates:
ratio of variances 
         0.1340342 
```

(Source: author's own work based on R Studio)

Here we do the second two independent samples variance test of expected value. The null hypothesis is that the difference in average income between developed countries' group and developing countries' group is the same. From table 24, we can get the p-value is less than 5%, which means we need to reject the null hypothesis we set before. And the conclusion is that the average income in developing countries' group and developed countries' group is not the same. And when have look at the last two rows of table 24, we can get the income estimation, which proves the decision we made. And this phenomenon is common as well because we know most developed countries tend to have higher income levels than developing countries.

Table 24 Two independent samples T- test



```
	Welch Two Sample t-test

data:  pw1$Income[pw1$dummy_Dev == 1] and pw1$Income[pw1$dummy_Dev == 0]
t = -9.3412, df = 41.031, p-value = 1.035e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12107.495  -7803.006
sample estimates:
mean of x mean of y 
 2917.356 12872.607 
```

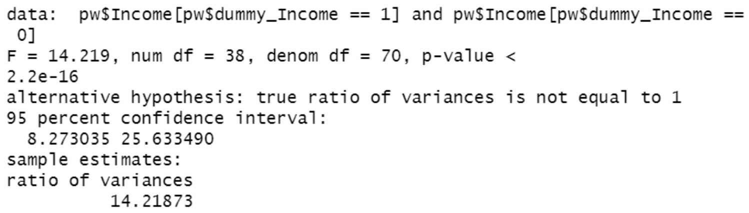
(Source: author's own work based on R Studio)

As we divide the "Happy score" variable into two categories, one is for the happy score value lower than the average score (5.42), and the other is for the happy score higher than average. We set the null hypothesis as the ratio between the number of happiness score that is upper and lower than the mean should be equal to 1. The variance ratio test turns out that the probability is about 48.8 %, which means we fail to reject the null hypothesis that the ratio of variance is equal to 1. And the actual ratio is 1.21. It is quite near 50%, which means we have relatively equivalent volumes on happy score scales. The conclusion we derive later on the endings is not biased, but relatively fair

In past experience, the whole world has a very extremely greatly variance in GDP growth. But in this dataset, we still need to test how equivalent the countries we have. Here it has the null hypothesis that the ratio between the number of lower-GDP-developing-speed countries and higher-GDP-developing-speed countries should be 1.

As a result, its p-value is relatively large, so we fail to reject the null hypothesis, which reflects our guessing before: the whole world has a very extreme greatly variance in GDP growth so that our sample is not well- distributed on GDP growth scale. But overall it has no influence on how fair our conclusion would be.

Table 25 Proportion test on higher-income countries test



```
data:  pw$Income[pw$dummy_Income == 1] and pw$Income[pw$dummy_Income ==
 0]
F = 14.219, num df = 38, denom df = 70, p-value <
2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  8.273035 25.633490
sample estimates:
ratio of variances
          14.21873
```

(Source: author's own work based on R Studio)

As we have tested the GDP ratio, we could also generate more insights on the income test. As we know, income could be one of the top elements that affect GDP. We still assume that the ratio of income variances batten lower income and higher income is equal to 1. Back in 1800, global inequality between countries was much lower than it is today (Max Roser, 2013). However, from table 25, it turns out the value is even smaller than 2.2e-16 so that we reject the null hypothesis. It means two samples of income variable have a big variance. The smaller value could be extremely smaller than the bigger values. It can be seen from the first part of the essay, which is the scale of measurement and quartile.

4. Summary of the conclusions

In this project work, we use 2 qualitative variables and 5 quantitative variables with 109 observations to analyze the degree of happiness of 109 countries (without outliers). First, we describe the type of variables and the scale of measurement of each variable. And then we find there is no missing data, but an outlier exists in the dataset. After detecting outlier, we check the descriptive statistics including mean, median, quartile, skewness, kurtosis, and so on. Then we check the relationship between two qualitative variables including ratios, frequencies, figures, and cross table. In the cross table, we can get calculate Cramer's V by using Chi-square and get there is a strong association between two qualitative variables. ANOVA table is analyzed and we get happiness score and GDP growth rate in five continent groups are totally different. After that, the covariance matrix is used to show a rough positive or negative relationship between two variables. A correlation heatmap, as a graphical representation of a correlation matrix, shows the positive or negative relationship in a detailed way. For instance, the correlation coefficient between income and GDP growth rate is closes to 1, which indicates a strong positive relationship between them. Practically speaking, if people live in a country with a high GDP growth rate, have a high income, then their happiness will be high. And if someone lives in a country with high-income inequality, then his happiness will be low.

References

Max Roser (2013) - "Global Economic Inequality". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/global-economic-inequality' [Online Resource].

Ruth Alexander, “Where are you on the global pay scale?”, 2012, Retrieved from: “https://www.bbc.com/news/magazine-17512040”