Relationship between Number of Starbucks store and Prosperity

```
## NULL
```

Introduction

Starbucks was first established in Seattle at 1971. Fifty years later, there are about 25,000 stores around the world, and Starbucks has become an important culture beyond just a cup of coffee. In urban area like Chicago and New York, we can easily see the Starbucks in every block but it is hard to see in the rural area. By observing this kind of situation, our group came up with a simple question: "what makes the difference?" Our group think that the number of Starbucks stores could be one of the indicators that represent the prosperity of city or country. To check our claim, we made a hypothesis: null hypothesis with "there is no relationship between the number of Starbucks stores and the indicators that represents the prosperity of city or country" versus the alternative hypothesis with "there is some relationship between the number of Starbucks stores and the indicators that represents the prosperity of city or country". In other words, we set our hypothesis $H_0 : B_j = 0$ and $H_a : B_j \neq 0$. Note that we used 7 variables including "GDP", "Happiness Score", "Cost of Living", "Rent Price", "Quality of Life", "Safety index" and "Population".

Our group strongly believe that if we can answer out question and find such indicator, we can inversely estimate that indicator by counting number of Starbucks. Furthermore , our group strongly believe that inverse estimation can be used in various fields including economics, business administration, and social studies. For example, economists can predict the estimated GDP of a country by just counting the number of Starbucks, and social scientists can make a further research of how Starbucks and coffee impact people's wellness and happiness. We expect Starbucks could be one of the important and indescribable index if our research has a meaningful result. In order to conduct our research, we first cleaned the data collected from variety sources and tested our hypothesis.

Data explanation

1) Data Source and Data Description

We used the total of six datasets for this project.

The first data we used is directory.csv from https://www.kaggle.com/starbucks/store-locations. This dataset includes a record for every Starbucks or subsidiary store location in operations worldwide. This data was scraped from the Starbucks store locator webpage by https://github.com/chrismeller/. There are a number of including Brand, Store Number, Store Name, Owenership Type, Street Address, City, and Country.

The second data we used is cost_of_living.csv from https://www.kaggle.com/dumbgeek/countries-dataset-2020?select=Cost+of+living+index+by+country+2020.csv updated by Varun Yadav. This dataset includes various variables such as cost of living and rent of each countires worldwide counted as the index number.

The third dataset we used is quality_of_live.csv from https://www.kaggle.com/dumbgeek/countries-dataset-2020?select=Quality+of+life+index+by+countries+2020.csv updated by Varun Yadav. This data includes variables realted to the indicators of the overall quality of live such as safety index, health care index and quality of life index.

The fourth dataset we used is happiness.csv from https://www.kaggle.com/unsdsn/world-happiness?select=2017.csv. There are two significant indicators in this dataset, which are happiness ranking of the countries worldwide and the happiness score of the those countries. These happiness scores and ranking use data from the Gallup World Poll. The happiness scores are entirely based on answers to the main life evalution question answered by respondents asked in the poll and the happiness ranking are assgined to each countries based on the scores.

The fifth dataset we used is economic.csv from https://www.kaggle.com/nottisani/worldwide-economics-gdp updated by Gabriela McDavid. The dataset includes a number of variables that help us to get a sense of the economic status of countires worldwide such as the net GDP and the size of the population.

The last dataset we used is gdpstate.csv from https://www.bea.gov/data/gdp/gdp-state.

2) Data Cleaning

Firstly, we cleaned up six dataset separately in order to select variables that we needed to address the topic of our research and curiousity. In addition, since the original dataset we used have their own distinct form of a data frame, we tidyed up each dataset having similar form of a data frame in order for us to combine the dataset and make the single data frame for the analysis. We made the two data frames, one for the analysis of the number of Starbucks in countries worldwide and its relation to the indicators of the overall quality of life, which are the net GDP, happiness score, cost of living, rent, quality of life index, safety index and the size of population and the other for the analysis of the number of Starbucks in 50 states of United States and its association with the net GDP of each states (We conducted this analysis because the United States is a huge outlier when it comes to the number of Starbucks stores and the net GDP).

Most importantly, for the orginal directory.csv dataset, the country names were recorded as an abbreviation such as "KR" for "South Korea" and "JP" for "Japan", so we revised the name of the countries into their full name. Also, there were some errors in the name of the countries so that we couldn't count the number of Starbucks stores in each countries properly. In order to address this type of error, we looked into the variables of "Street address," "City," "Province," and "Postcode" and figured out that some country names like South Korea, China and Japan were in these variables, not in the column of "Country". After fixing this error, we counted the number of Starbucks based on the countries and made a new data frame named "starbucks". Without huge difficulties, we cleaned up the rest of the dataset and combined them with the newly made starbucks dataframe (the number of Starbucks in each countries). In order to analyze the number of Starbucks in each states in the United States and its association with GDP of each states, we only filtered out the country of United States in the data frame of "starbucks" and made a new column of "states". Lastly, we recounted the number of Starbucks stores based on the states and combine this data frame with the gdpstate.csv.

```
# # A tibble: 54 x 9
#    Country starbucks   GDP Happiness.Score Cost_of_living  Rent Quality.of.Life~
#    <chr>       <int> <int>           <dbl>          <dbl> <dbl>            <dbl>
#  1 SLOVAK~         3    87            6.10           44.5 16.1              153.
#  2 SOUTH ~         3   315            4.83           42.9 16.6              132.
#  3 BULGAR~         5    50            4.71           36.7  9.64             130.
#  4 PANAMA          5    52            6.45           54.2 24.8              108.
#  5 FINLAND         8   232            7.47           70.3 26.2              190.
#  6 MOROCCO         9   101            5.24           34.3  8.94             105.
#  7 CYPRUS         10    19            5.62           57.9 20.5              148.
#  8 COLOMB~        11   292            6.36           30.7  9.58             106.
#  9 PORTUG~        11   199            5.20           49.5 21.8              163.
# 10 HUNGARY        16   122            5.32           40.8 14.0              128.
# # ... with 44 more rows, and 2 more variables: Safety.Index <dbl>,
# #   Population <dbl>
```

The above dataframe is our final combined data for the analysis of the relationship between the number of starbucks in countries worldwide and significant indicators related to the overall quality of life.

```
#             State  code       GDP store
# 1         alabama  1000  221030.7    85
# 2          alaska  2000   54292.9    49
# 3         arizona  4000  350718.3   488
# 4        arkansas  5000  127761.3    55
# 5      california  6000 2975083.0  2821
# 6        colorado  8000  372452.9   481
# 7     connecticut  9000  279782.3   123
# 8        delaware 10000   74186.7    25
# 9         florida 12000 1050298.4   694
# 10        georgia 13000  602023.9   326
# 11         hawaii 15000   93100.5    99
# 12          idaho 16000   79090.8    67
# 13       illinois 17000  863039.5   575
# 14        indiana 18000  368424.5   221
# 15           iowa 19000  190147.0    89
# 16         kansas 20000  171718.8    94
# 17       kentucky 21000  207849.4   116
# 18      louisiana 22000  253236.1    84
# 19          maine 23000   64557.0    30
# 20       maryland 24000  411619.1   257
# 21  massachusetts 25000  570464.2   273
# 22       michigan 26000  521803.4   283
# 23      minnesota 27000  371929.7   184
# 24    mississippi 28000  113578.5    32
# 25       missouri 29000  317949.1   188
# 26        montana 30000   50692.3    36
# 27       nebraska 31000  124705.4    58
# 28         nevada 32000  169179.6   253
# 29  new hampshire 33000   84584.1    29
# 30     new jersey 34000  612979.3   261
# 31     new mexico 35000  100079.9    76
# 32       new york 36000 1705010.2   645
# 33 north carolina 37000  567451.7   338
# 34   north dakota 38000   56286.8    13
# 35           ohio 39000  675029.7   378
# 36       oklahoma 40000  198595.8    79
# 37         oregon 41000  241978.1   359
# 38   pennsylvania 42000  778374.6   357
# 39   rhode island 44000   59924.6    27
# 40 south carolina 45000  235286.9   131
# 41   south dakota 46000   53239.0    25
# 42      tennessee 47000  362737.1   180
# 43          texas 48000 1795635.1  1042
# 44           utah 49000  181622.7   101
# 45        vermont 50000   32981.0     8
# 46       virginia 51000  533510.4   432
# 47     washington 53000  575416.7   757
# 48  west virginia 54000   77632.5    25
# 49      wisconsin 55000  337553.1   145
```

```
# 50        wyoming 56000   39703.2    23
```

The above dataframe is our final combined data for the second analysis of the relationship between the number of Starbucks in each of 50 states in America and its association with the net GDP.

Method & Analysis

After cleaning the dataset, our group mainly used "linear regression" for the analysis so that we are able to observe the dependency between the number of Starbucks and other variables. Our group checked whether there is a relationship between specific variables and number of Starbucks, we checked the associated p-value and found that only "GDP" and "Population" seemed meaningful for initial hypothesis. Hence, we made deeper analyzations on both "GDP" and "Population". We then used R-squared to check the fitness of our model. Moreover, our group detected that the number of Starbucks in USA seems to be too large compared to other datasets and therefore, we decided to subdivide the "USA" into "each states" to prevent USA being an outlier in terms of the number of Starbucks. Furthermore, to help the audience look our data at ease, our group used Choropleth map to compare the distribution of Starbucks in USA and GDP at glances. As non-linearity of the response-predictor relationship is one of the main concern in linear regression model, our group also checked whether there is a non-linear association with each variables to check which model is suitable by checking the associated p-value.

Results

1) Country: Relationship between Number of Starbucks and overall Quality of Life

(a) Linear regression

```
#
# Call:
# lm(formula = starbucks ~ . - Country, data = Final_data)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2116.79   -90.84    36.74   185.20  2125.34
#
# Coefficients:
#                          Estimate Std. Error t value Pr(>|t|)
# (Intercept)           1234.30033  949.14952   1.300    0.200
# GDP                      0.71098    0.04183  16.999  < 2e-16 ***
# Happiness.Score       -119.15772  175.00133  -0.681    0.499
# Cost_of_living         -10.20238    9.63865  -1.058    0.295
# Rent                    16.17272   13.93950   1.160    0.252
# Quality.of.Life.Index   -0.90751    5.46128  -0.166    0.869
# Safety.Index            -4.73674    9.18879  -0.515    0.609
# Population              -2.96726    0.49730  -5.967 3.25e-07 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 675.8 on 46 degrees of freedom
# Multiple R-squared:  0.8856, Adjusted R-squared:  0.8682
# F-statistic: 50.87 on 7 and 46 DF,  p-value: < 2.2e-16
```

From multiple linear regression, we found out that "GDP" and "Population" were highly statistically significant variabels with extremely low p-values. Hence we may reject the null hypothesis for "GDP" and "Population".

```
# [1] "R squared Value"
```

```
# [1] 0.8855986
# [1] "Adjust R squared Value"
# [1] 0.8681897
```

According to summary of our model, both r-squared(0.8804866) and adjusted r-squared(0.8622998) had values close to 1, so we can say that our model is good and reliable.

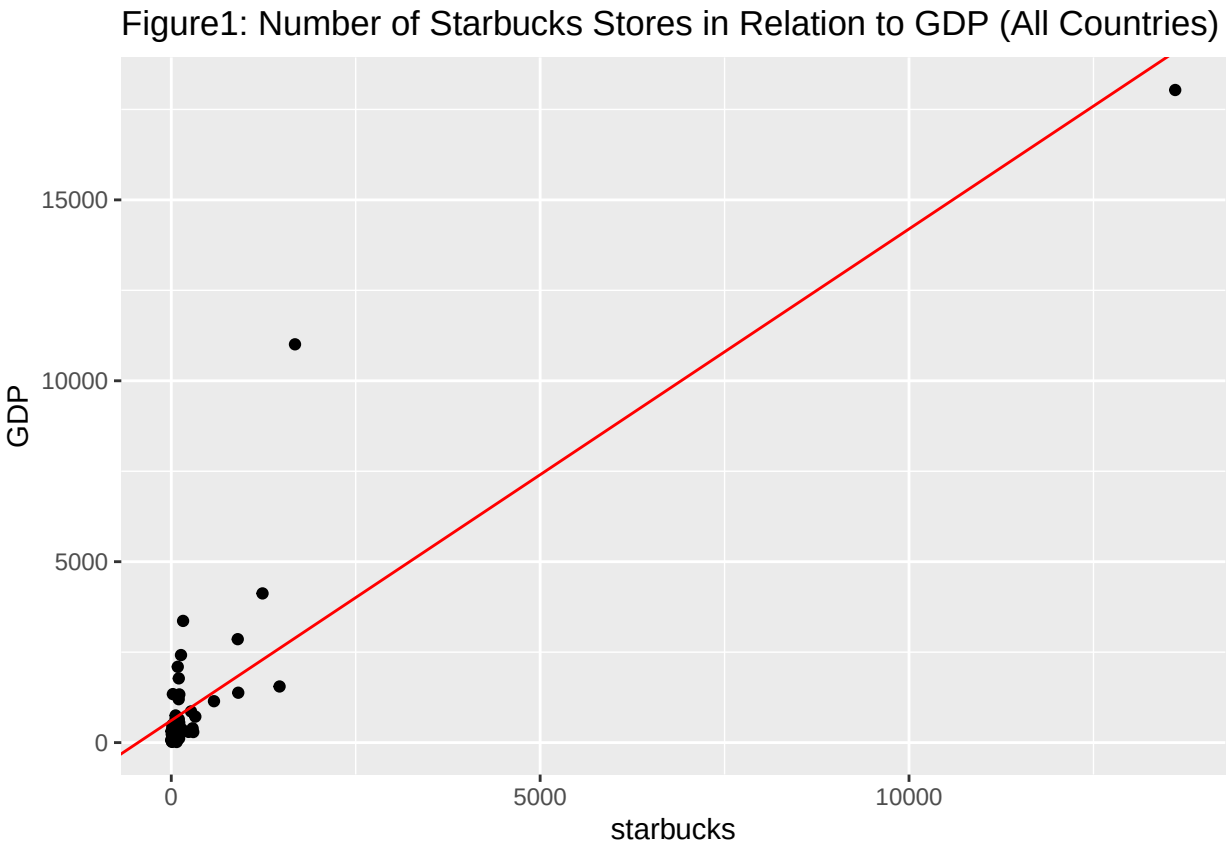
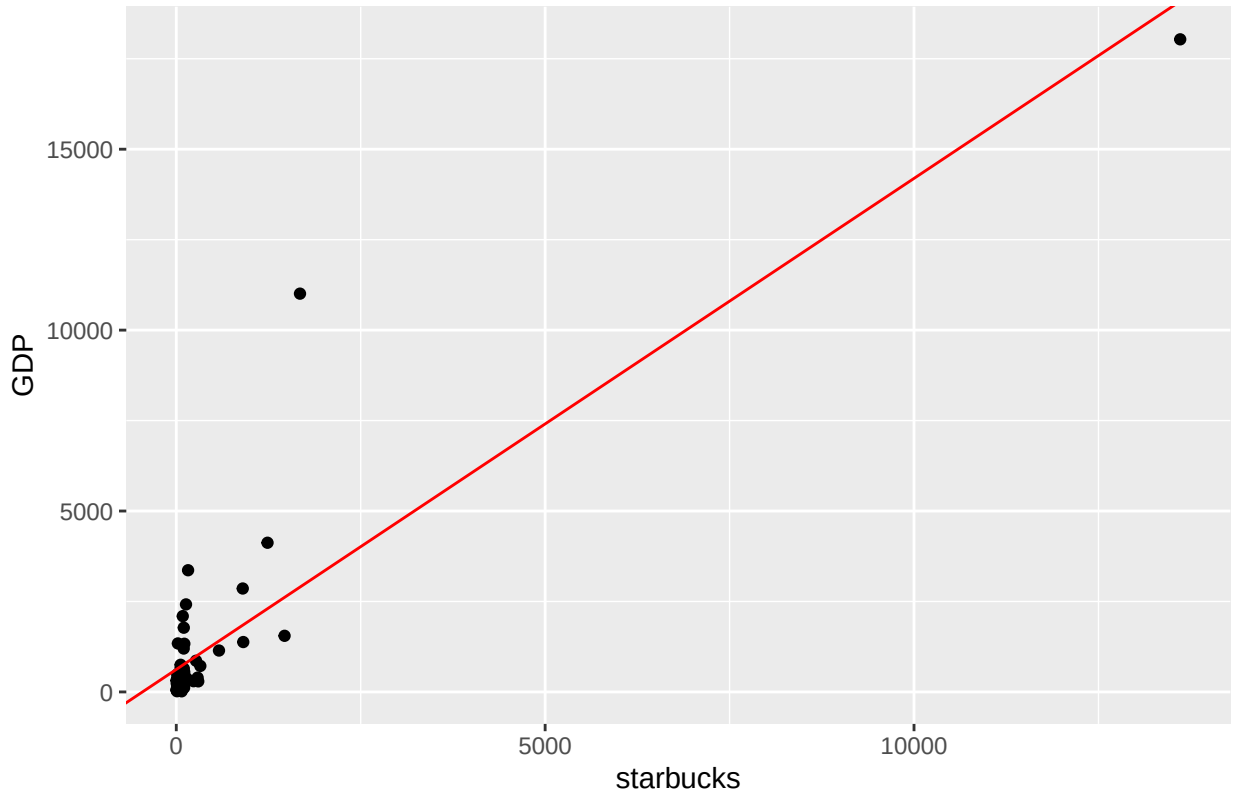
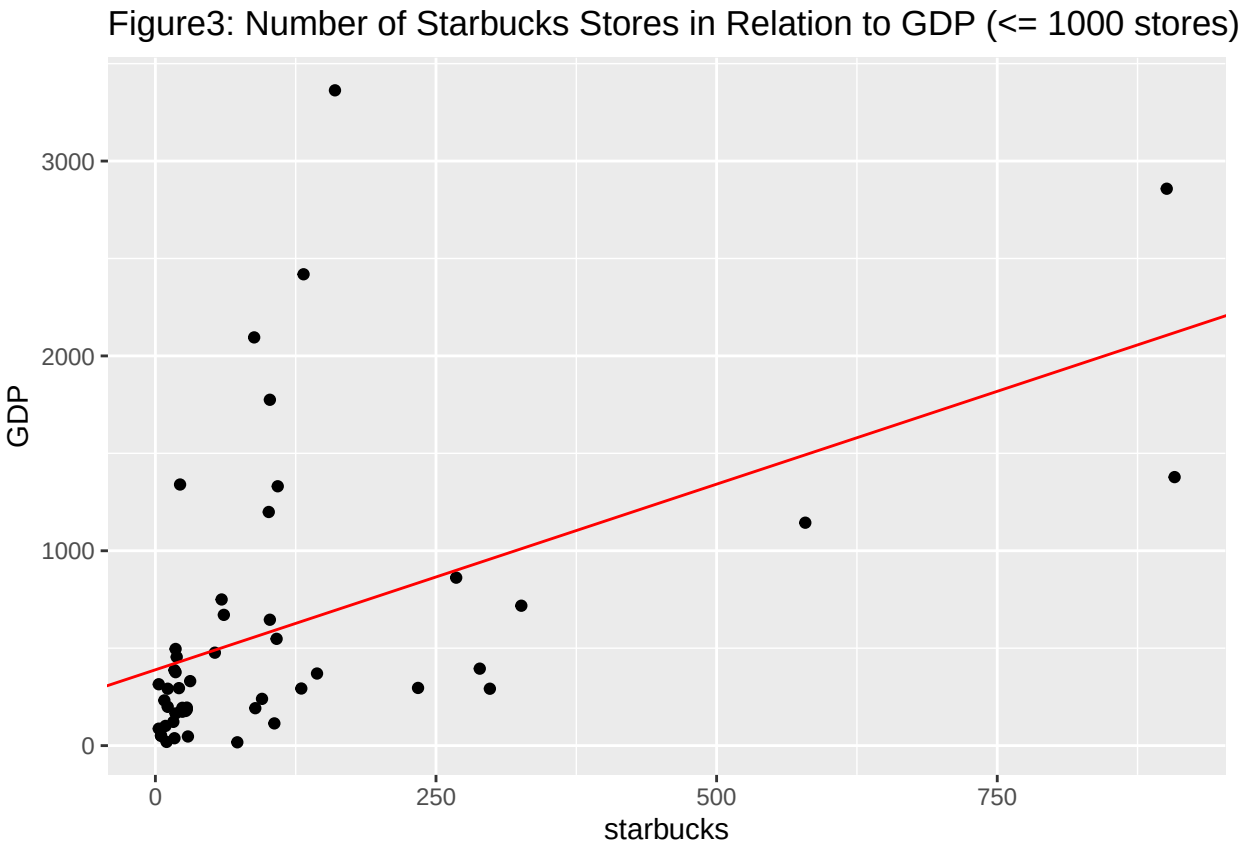


Figure2: Number of Starbucks Stores in Relation to GDP (exceptU.S.)





Above *Figure1 Figure2 Figure3* show that linearity of starbucks stores with countries' GDP and all of them show good linearity. The *Figure1*, as U.S. overwhelmingly has too many stores, on the *Figure2*, we singled out U.S. and represented the rest of countries. Also, on the *Figure3* in order to check possible distortions of our data, we checked countries with stores less than 1,000. Regardelss of the number of stores of each country, there was a good linearity between GDP and number of stores. However there were some points deviating from the regression line, which leaves the room for further alternative analysis.

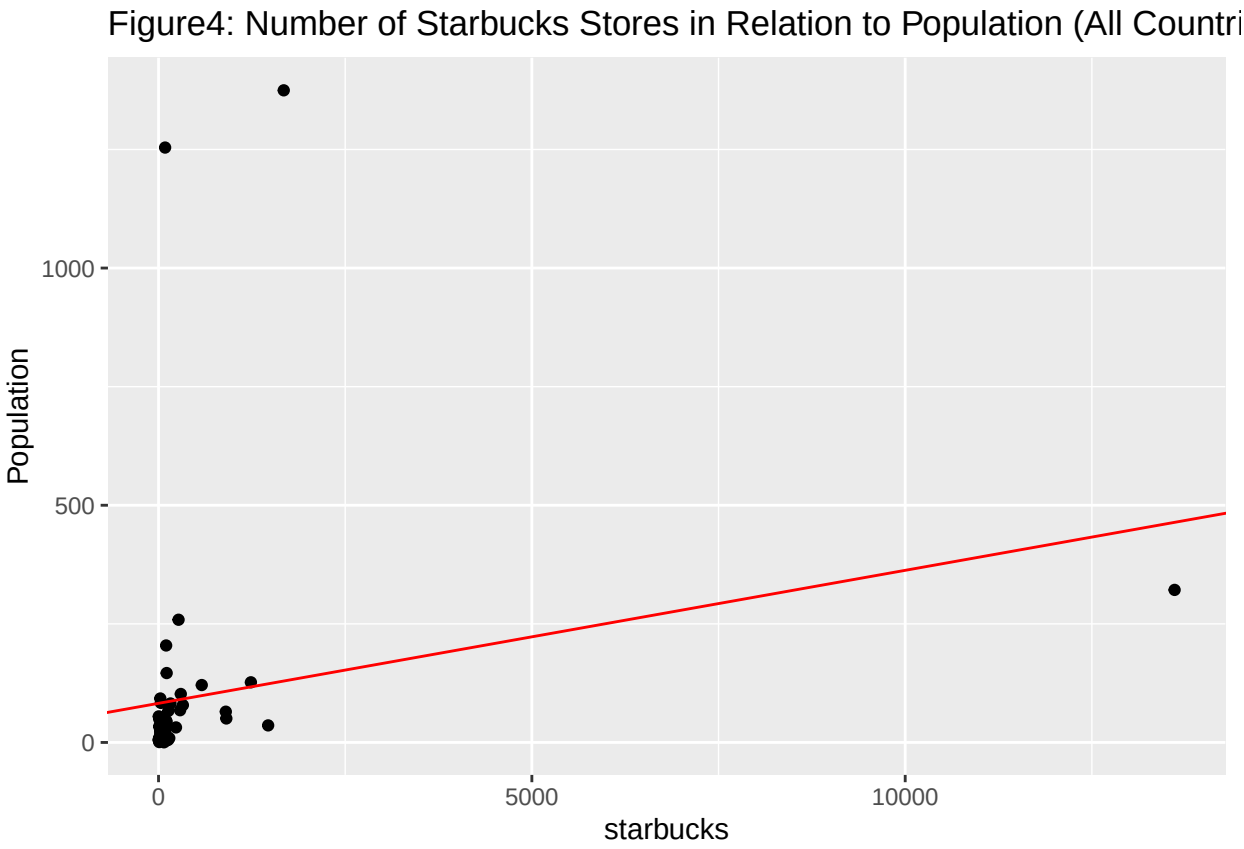


Figure4: Number of Starbucks Stores in Relation to Population (All Countri
Population
1000
500
0
0
5000
10000
starbucks

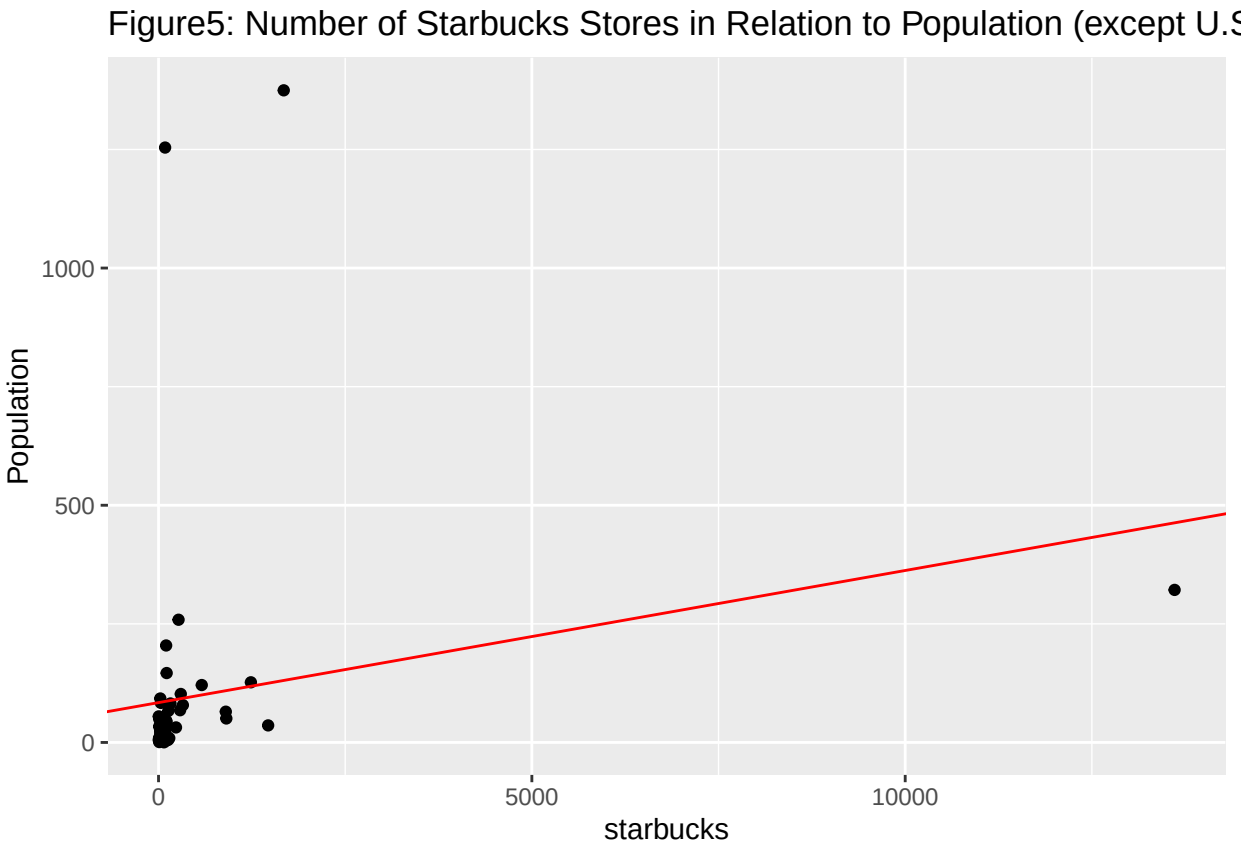
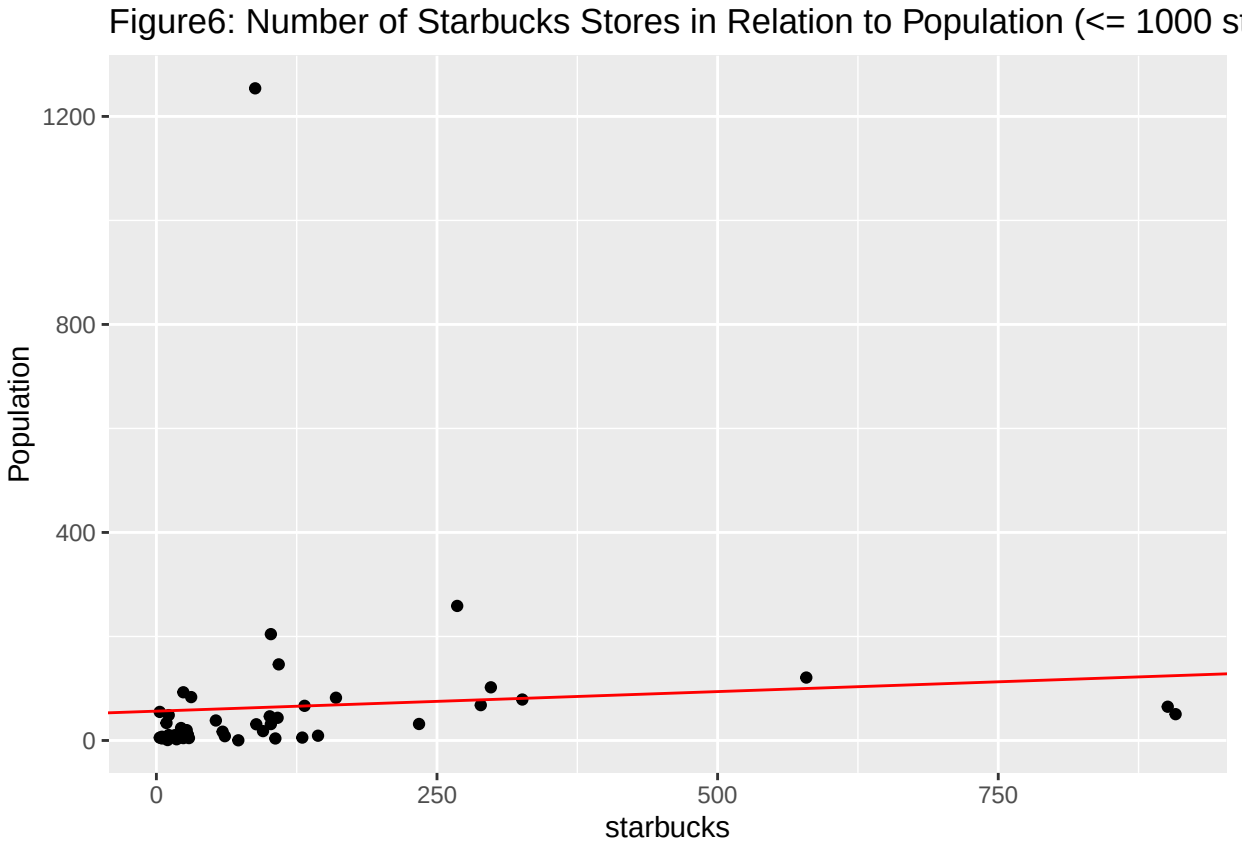


Figure5: Number of Starbucks Stores in Relation to Population (except U.S
Population
1000
500
0
0
5000
10000
starbucks



Using the separated datsets, we found above *Figure4 Figure5 Figure6* showed good linearity of starbucks stores with countries.' But there were still some few points far off from the regression line, so we had to take additional approach.

(b) Non-linear Association

```
#
# Call:
# lm(formula = starbucks ~ poly(GDP, 4), data = Final_data)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -550.18  -76.37  -25.32   44.95 1101.95
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)     443.09      33.88  13.079  < 2e-16 ***
# poly(GDP, 4)1 11990.42     248.95  48.164  < 2e-16 ***
# poly(GDP, 4)2  5285.20     248.95  21.230  < 2e-16 ***
# poly(GDP, 4)3  2964.03     248.95  11.906 4.51e-16 ***
# poly(GDP, 4)4   345.75     248.95   1.389    0.171
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 248.9 on 49 degrees of freedom
# Multiple R-squared:  0.9835, Adjusted R-squared:  0.9821
# F-statistic: 728.5 on 4 and 49 DF,  p-value: < 2.2e-16
```

Additionally, we went on to check whether there were non-lineaer association in this datset. Using poly() function, in terms of GDP, (except 4th polynomial) they had siginificantly low p-values. At this point, we may say that there is a non-linear association between the number of Starbucks stores and the level of GDP of each countries.

```
#
# Call:
# lm(formula = starbucks ~ poly(Population, 4), data = Final_data)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -5264.4  -267.2   -48.3   295.2  5022.3
#
# Coefficients:
#                       Estimate Std. Error t value Pr(>|t|)
# (Intercept)              443.1      166.0   2.670   0.0103 *
# poly(Population, 4)1    2828.1     1219.5   2.319   0.0246 *
# poly(Population, 4)2   -6875.1     1219.5  -5.638 8.42e-07 ***
# poly(Population, 4)3    -775.0     1219.5  -0.636   0.5281
# poly(Population, 4)4    7410.1     1219.5   6.076 1.78e-07 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1219 on 49 degrees of freedom
# Multiple R-squared:  0.6032, Adjusted R-squared:  0.5708
# F-statistic: 18.62 on 4 and 49 DF,  p-value: 2.307e-09
```

Also, we wanted to check non-linear assocaition win terms of Population. Except third polynomial, the rest of them were statistically significant.

2) US: Relationship between Number of Starbucks and State GDP

(a) Linear regression

```
#
# Call:
# lm(formula = store ~ GDP, data = State_GDP_SB)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -611.52  -56.00   -3.00   29.77  599.44
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept) -3.901e+01  2.948e+01  -1.323    0.192
# GDP          7.598e-04  4.460e-05  17.035   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 164.2 on 48 degrees of freedom
# Multiple R-squared:  0.8581, Adjusted R-squared:  0.8551
# F-statistic: 290.2 on 1 and 48 DF,  p-value: < 2.2e-16
```

According to Starbucks data, there was a total of 24,459 Starbucks stores in the world. However, we found that more than half of the Starbucks stores were found in the United States, which 13,608 stores. So we decided to separate US data and analyzed the correlation between the number

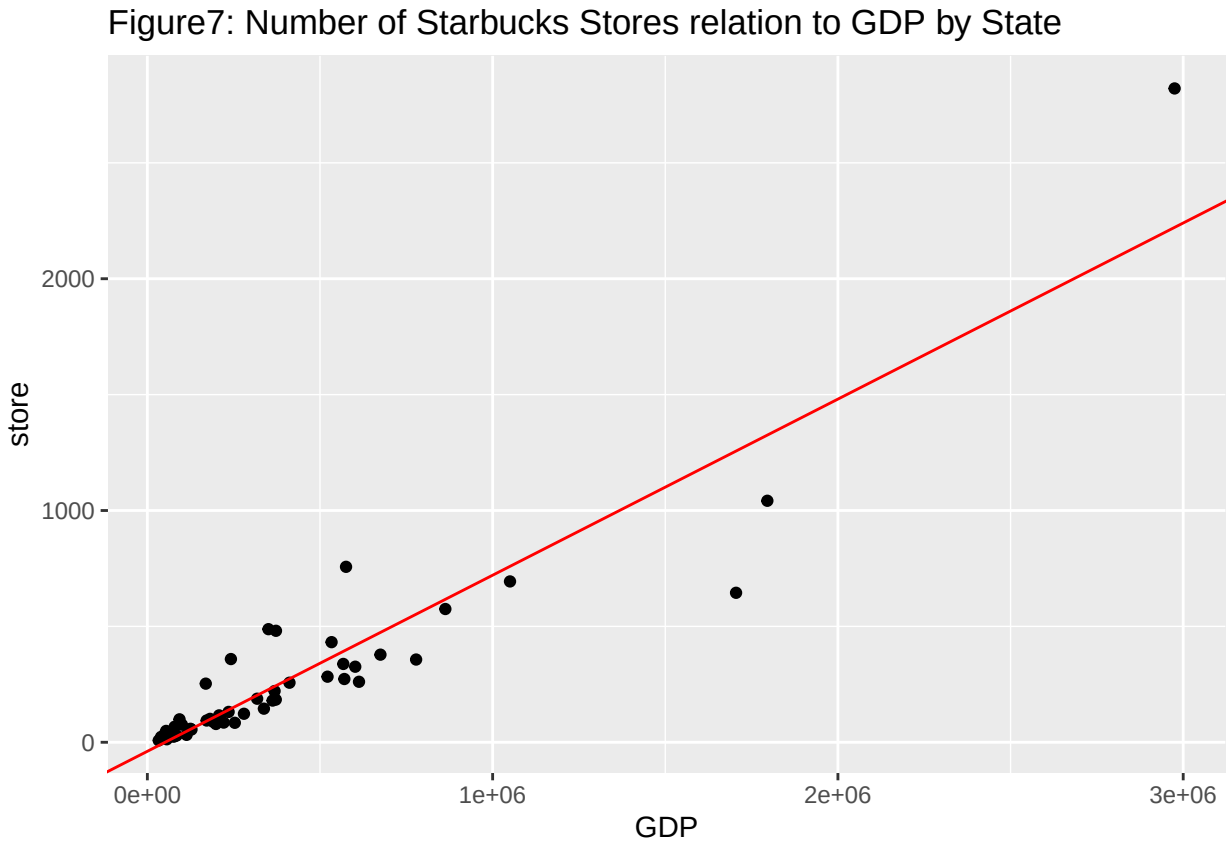
of Starbucks stores in each state and each state GDP.

Figure7 and *Figure8* are the choropleth map that shows the number of GDP and stores in the US. Both figures show similar trends that California and Texas have both high GDP and number of the store. However, to find a more accurate correlation, we implemented a linear model of the number of stores and GDP.

Figure9 shows the linear regression graph our linear model. The plot is normally distributed and each point are lined well on the straight dashed line.Moreover, the p-value of the linear model was 2e-16 which is smaller than 0.05, and R squared and adjusted R squared is 0.85, which close to 1. Therefore, since we got a significant p-value, R squared, and adjust R squared, we can conclude it is statistically significant between the number of Starbucks stores and GDP.

```
# [1] "R squared Value"
# [1] 0.8580715
# [1] "Adjust R squared Value"
# [1] 0.8551146
```

According to summary of our model, both r-squared(0.8804866) and adjusted r-squared(0.8622998) had values close to 1, so we can say that our model is good and reliable.



(b) Choropleth map

Figure8: Choropleth map of GDP by State

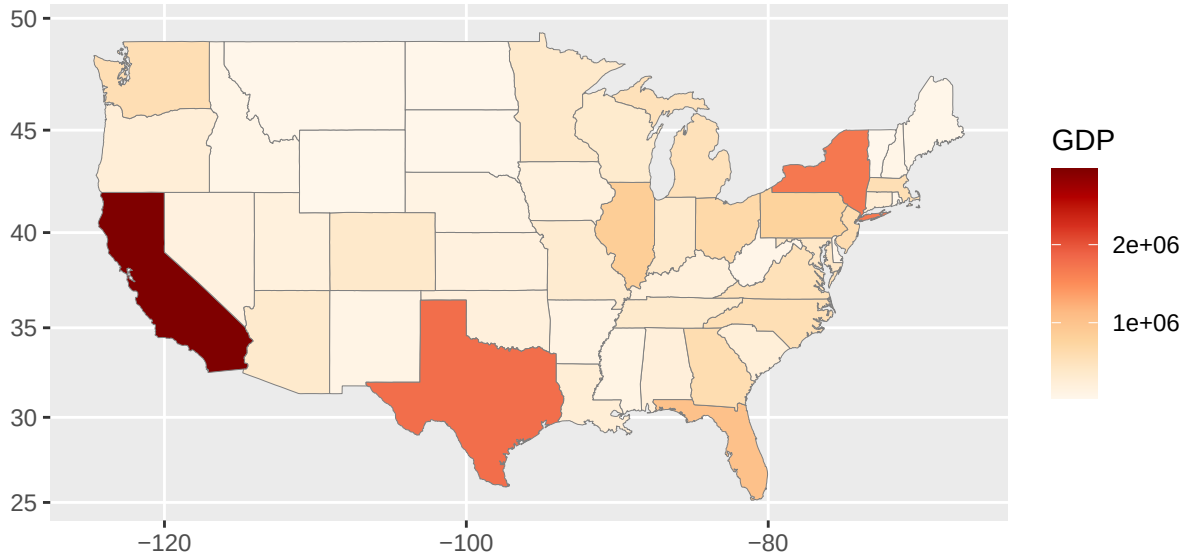
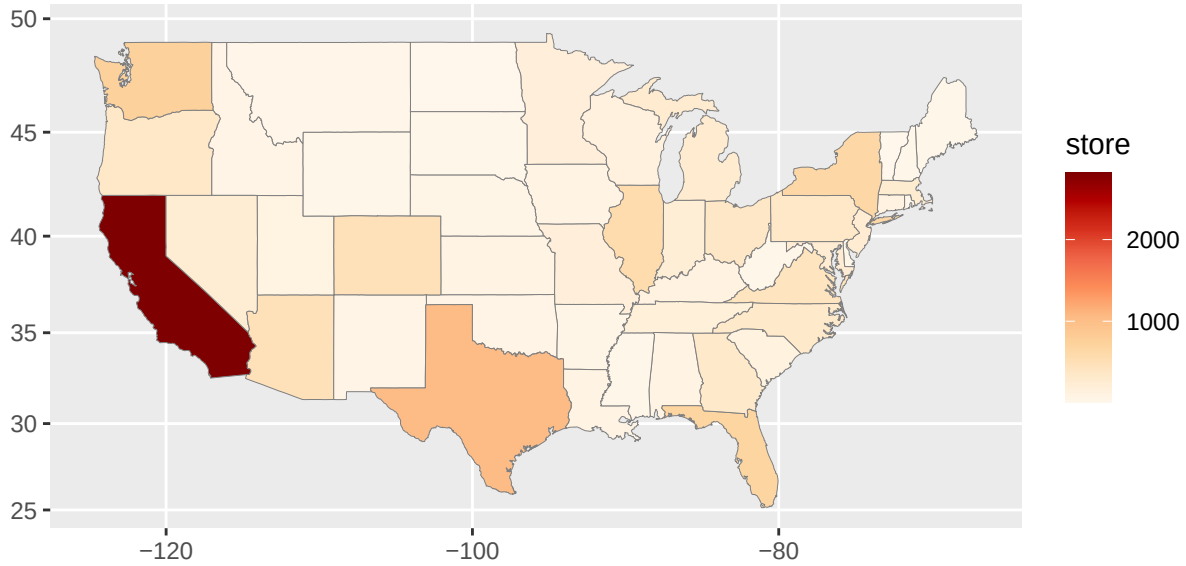


Figure9: Choropleth map of Number Starbucks Store by State



(c) Non-linear Association

```
#
# Call:
# lm(formula = store ~ poly(GDP, 4), data = State_GDP_SB)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -193.25  -56.47  -21.68   15.90  376.65
#
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)     270.34      14.95  18.083  < 2e-16 ***
# poly(GDP, 4)1  2797.52     105.71  26.463  < 2e-16 ***
# poly(GDP, 4)2   662.83     105.71   6.270 1.24e-07 ***
# poly(GDP, 4)3   589.67     105.71   5.578 1.32e-06 ***
# poly(GDP, 4)4    67.28     105.71   0.636    0.528
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 105.7 on 45 degrees of freedom
# Multiple R-squared:  0.9449, Adjusted R-squared:   0.94
# F-statistic: 192.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

We had check non-linear association terms between number of Starbucks store and GDP using Polynomial Regression. From the result,except quadratic term, the rest of them were statistically significant.

Conclusion

We were interested in the relationship between the number of starbucks stores and other various countries' data. Then, we found out that United States had overwhelmingly many starbucks stores compared to other countries, so as a country scale, including United States in the data had the possibility of distorting data. Accordingly, we separated United States and delved into it in a more domestic and microscopic way. According to subsequent linear regression analyzing the correlation between the number of starbucks stores with other predictors on a U.S. state level, we discovered that 'GDP' was the most explanatory predictor. We found out the relation with GDP using linear regression model and geographic visualizations. Back to a country scale excluding United States, 'GDP' and 'Pouplation' of countries had higher correlation with the number of starbucks stores according to linear regression model. To quench our curiosity, we also went on to check non-linear association and re-confirmed that 'GDP' and 'Population were still statistically significant. We have done by dissecting this project into two parts: United State's domestic level, and international level. By looking at U.S. domestic level and international level with United States, GDP was assumed to be the strong predictor correlated with the number of stores. Of course, there existed limitation on these results. The number of stores in starbucks dataset was not accurate, due to missing countries (Italy) and wrong data inputs. Also, the correlation we found is not equivalent to causality. Therefore, more explanations and further analysis has to be done. Also, off the top of our head, we intuitively can predict that the more prosperous and richer the region or country is, the more starbucks stores located at. Though there was not a novel discovery in our project, it is an undeniable fact that 'GDP' is the powerful predictor correlated with the number of starbucks stores. To sum up, in terms of the indicators of 'GDP' and 'population' we were able to reject the null hypotheses.