Factors Affecting Global Tourism

Manay Divatia and Heinrick Gonzales

Abstract

Tourism serves as a vital component of the global economy. Through an analysis of data obtained from Kaggle, derived from the World Bank, we examined 50 variables encompassing 145 countries to investigate the various factors influencing the volume of tourism within nations. Our findings indicate that Gross Domestic Product (GDP) emerges as a significant determinant affecting tourism levels across countries.

Background Information

According to Khan (2020), “transportation, peace, price of different commodities, infrastructure, education, and population affect the tourism industry positively”. The method used for Dr. Khan’s study included qualitative research and an analysis of 18 web articles.

The roots of tourism can be traced back to ancient civilizations, where travelers embarked on journeys for trade, religious pilgrimage, or leisure. Over time, the concept evolved, spurred by advancements in transportation and communication, leading to the emergence of modern tourism as we know it today. In contemporary times, tourism has become a major engine of economic development, contributing significantly to GDP, job creation, and foreign exchange earnings in many countries. Governments recognized the potential of tourism as an economic driver and began investing in infrastructure, marketing, and regulatory frameworks to support the industry's expansion.

Our analysis seeks to investigate different factors that may influence the influx of arrivals into countries from a quantitative perspective. Through an examination of country-level data from GDP to indicators of population happiness, we can understand determinants driving tourism growth. Through our research, we hope that policymakers can leverage these insights to formulate policies to enhance the competitiveness of their respective countries within the global tourism landscape.

Data Collection and Cleaning

To first collect our data, we looked to Kaggle which gave us all the data we needed. Firstly, we found general data on each country like GDP, Population, Land Size, Military power, etc. Next, we found a dataset that provided information from a study done on happiness of a country. This study was done by the World Happiness Report and included information about freedom, happiness ranking, and others. One thing to note with this dataset is that it did not span all countries which limited the data and our findings. Next, we found data on tourism numbers. We decided to define tourism as the number of people coming into a country who aren't immigrating. Once again, this dataset was limited in that it did not provide information for every country. Another big limitation is that the data only spanned up until 2018. To solve this, we looked directly at the data source but we found that the data source stopped their data collection in 2018. We also tried contacting the publisher of the dataset via LinkedIn but got no response. In order for the data to match up, we ensured that the previous datasets we found were also from 2018. Finally, what we found is that some countries were listed with slightly different names for each dataset. For example, the first dataset had the United States listed as "USA" while the second dataset listed it as "United States of America". To fix this issue, we found a fourth dataset

with the standard 3 and 4 letter country codes and used it to merge all the remaining 3 datasets. Ultimately, we were able to retain 145 countries. If we hadn't used the standard country codes, we would only be able to retain 131 countries.

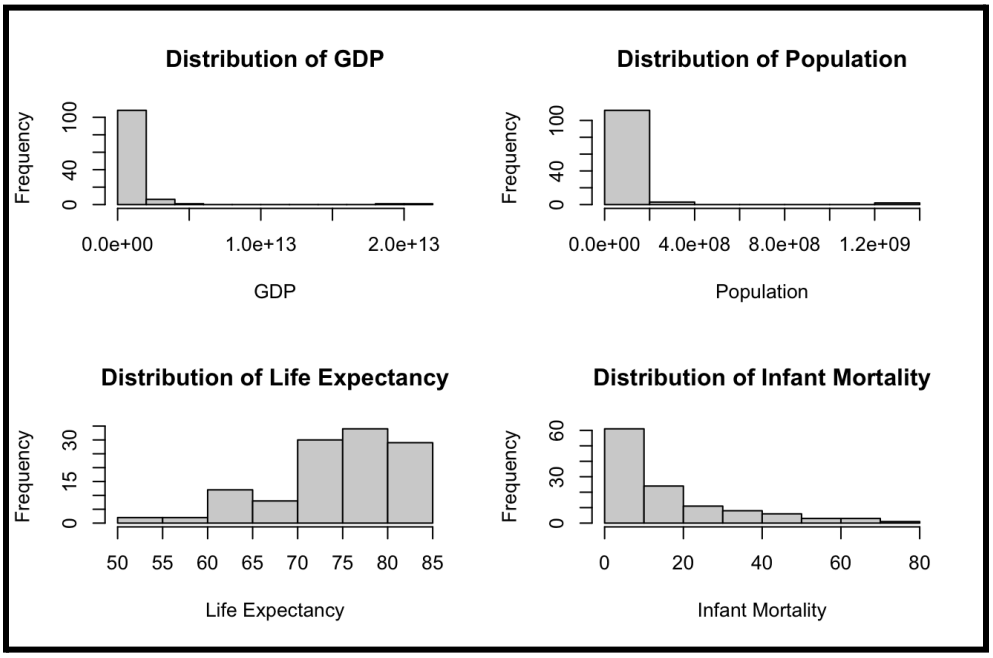
Now that we had our data merged, we had to clean the data. The first thing we did was remove any columns we felt were not relevant. This included columns like country abbreviation, official language, and country capital. What we immediately found was that many columns like population, tourism and gdp were listed as strings instead of numeric values. In fact, we found that over half of the variables had an incorrect type. Additionally, some columns like GDP, minimum wage, and gasoline price had a leading dollar sign character which had to be removed. Similarly, columns like tax revenue, agricultural land percentage, and CPI change had a trailing percentage sign which also had to be removed. After converting types and removing unnecessary characters, we removed all columns with null variables in our data. From here, we had a cleaned dataset that we could perform operations on in order to attain insights.

What we found later in our analysis was that the tourism, GDP, and population variables had significant outliers. In order to combat this, we took the log of those variables and redid our analyses, specifically for the unsupervised learning.

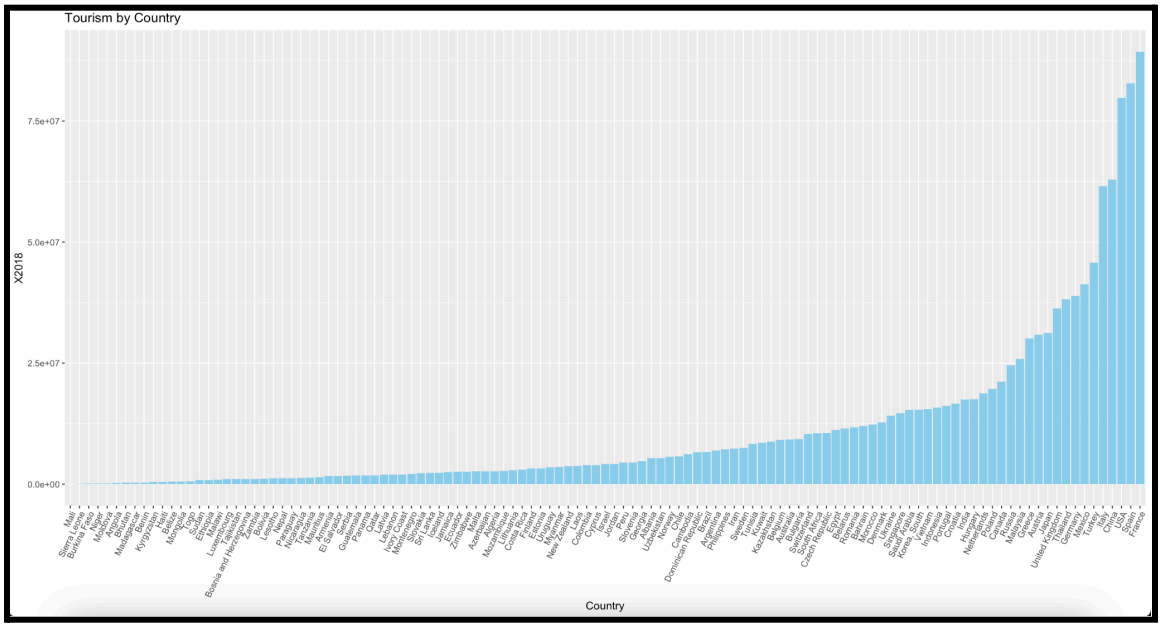
Here is an overview of the columns we decided to keep, their type, and what they describe.

Variable	**Class**	**Description**
Tourism	Integer	Total tourism (defined by WorldBank) in 2018
logTourism	Double	Log (base 10) of the tourism variable
logGDP	Double	Log (base 10) of the GDP variable
GDP	Integer	Gross domestic product in 2018
Forested.Area	Double	The percentage of land area that has vegetation and/or is a forested area
Freedom.to.make.life.choices	Double	Score of freedom based on the World Happiness Report
Latitude	Double	The median latitude of the country
Overall.rank	Integer	The happiness rank based on the World Happiness Report
logPopulation	Double	Log (base 10) of the population variable
Population	Integer	Amount of citizens in a country in 2018
Life.expectancy	Double	Average life expectancy in 2018
Infant.mortality	Double	Infant mortality rate in 2018
Tax.revenue	Double	Percentage of government revenue that comes from taxes
CO2.Emissions	Integer	Amount of CO2 emissions in kilotons
CPI	Double	The Consumer Price Index in 2018
Minimum.wage	Double	Federal minimum wage in 2018
Unemployment.rate	Double	National unemployment rate in 2018
Total.tax.rate	Double	Total tax rate (federal and state) in 2018
TourismPerCapita	Double	Tourism amount per citizen (tourism variable divided by population)
Urban.population	Integer	Amount of population that lives in an urban area

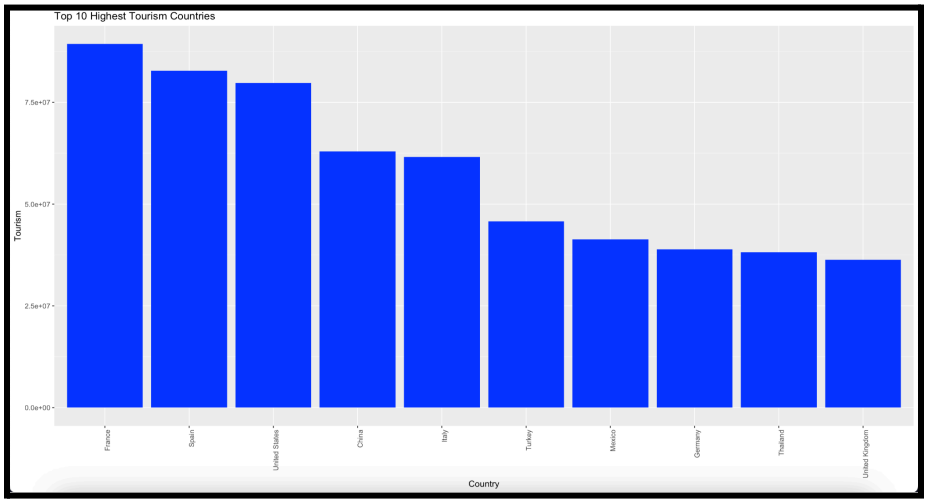
Basic Insights

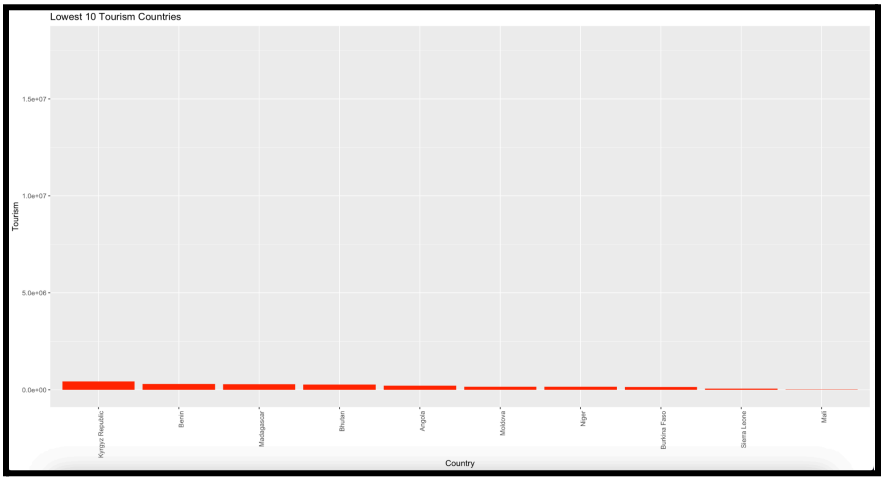


This is a basic histogram depicting the distribution of the variables: GDP, Population, Life Expectancy, and Infant mortality. One big thing we can understand from these graphs is the outliers that exist for GDP and Population. Like mentioned earlier, we chose to make these outliers not so drastic by taking the log of the GDP and Population variables for the future.



This bar graph shows the distribution of tourism for all the countries in our dataset. This graph summarizes the distribution of our tourism variable and shows the long tail that it has. In this context, it shows how there are many countries with very little tourism and a few countries with an incredibly large number of tourists. It also allows us to see that France, Spain, the US, China, and Italy are the leading countries in terms of tourism.





To further show the drastic difference between the top countries and bottom countries in terms of tourism, we plotted 2 bar graphs that show the top 10 and lowest 10 tourism countries. One thing

to note is that the scale of the lowest 10 tourism countries graph is 5 times smaller than the highest 10 graphs because if they were on the same scale, the bar graphs wouldn't even show up. Another interesting insight here is that the countries with the lowest tourism were also relatively small in population and land size. This is part of the reason we decided to look into tourism per capita in our analysis.

Methods & Interpretation of Results

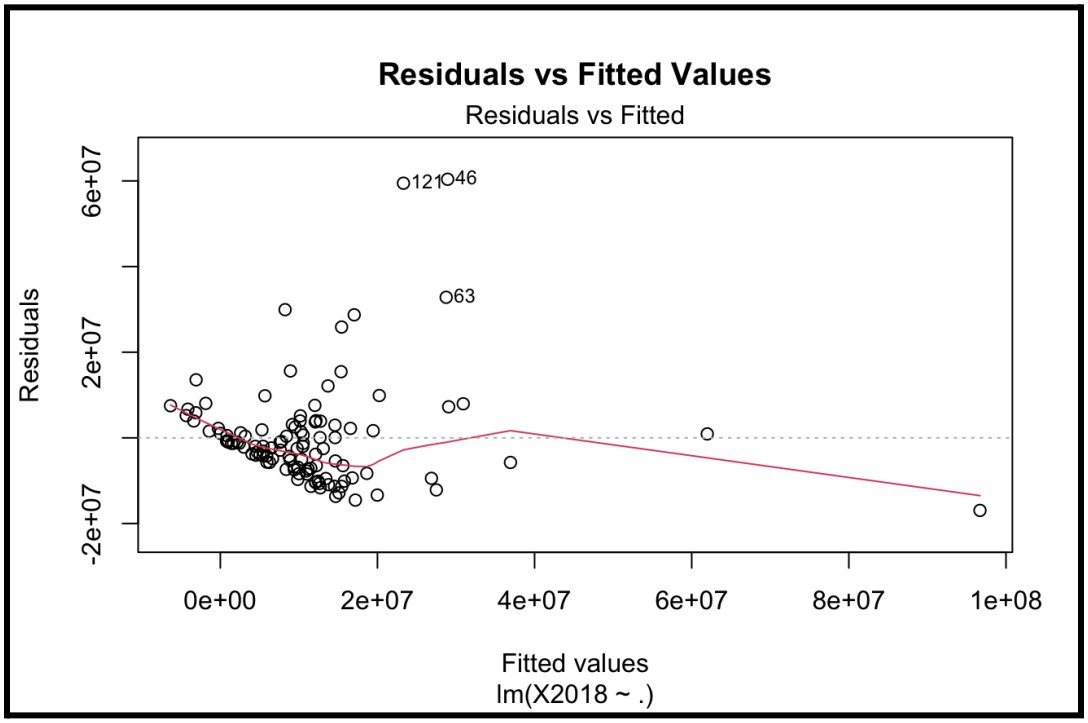
Linear Regression

```
Call:
lm(formula = X2018 ~ . - logX2018, data = relevant_data)

Residuals:
      Min        1Q    Median        3Q       Max
-21619547  -4661980  -2433589   2905138  55241917

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.061e+08  4.206e+07  -2.523   0.0132 *
GDP                           4.228e-06  1.018e-06   4.154 6.99e-05 ***
Forested.Area....            -3.610e+02  5.571e+04  -0.006   0.9948
Freedom.to.make.life.choices -1.053e+07  9.428e+06  -1.116   0.2670
Latitude                      8.632e+04  4.925e+04   1.753   0.0828 .
Overall.rank                 -3.752e+04  4.640e+04  -0.809   0.4207
logPopulation                 4.933e+06  8.953e+05   5.510 2.90e-07 ***
Life.expectancy               5.233e+05  5.045e+05   1.037   0.3022
Infant.mortality              1.837e+04  1.886e+05   0.097   0.9226
Tax.revenue....               1.990e+05  1.913e+05   1.040   0.3007
Co2.Emissions                -5.732e+00  2.690e+00  -2.131   0.0356 *
CPI                          -1.681e+04  2.049e+04  -0.820   0.4142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11360000 on 98 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.6132,    Adjusted R-squared:  0.5698
F-statistic: 14.12 on 11 and 98 DF,  p-value: 7.078e-16
```

We decided to run a linear regression analysis to see which variables were significantly correlated with tourism in 2018. What we found is that GDP, the log of population, and CO2 emissions were significantly correlated with tourism. We decided to use this information in our unsupervised learning portion of the analysis. We also got a decent

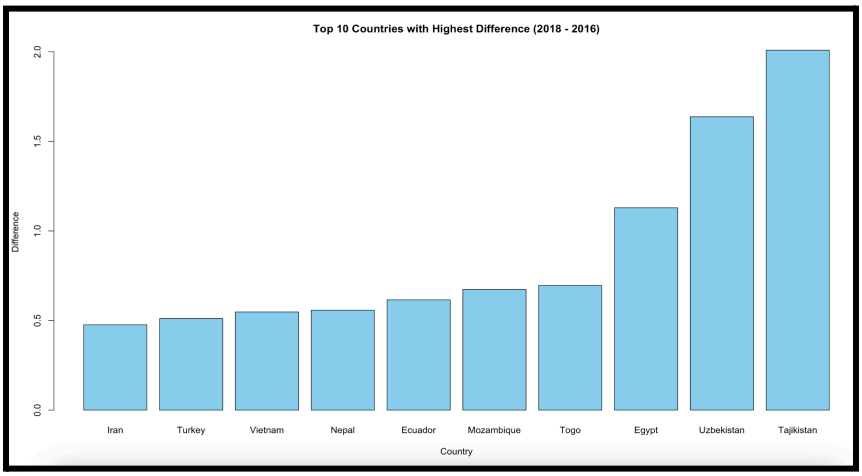
R-squared value of 56.98%. This means that 56.98% of the variation in our data can be explained by this model.

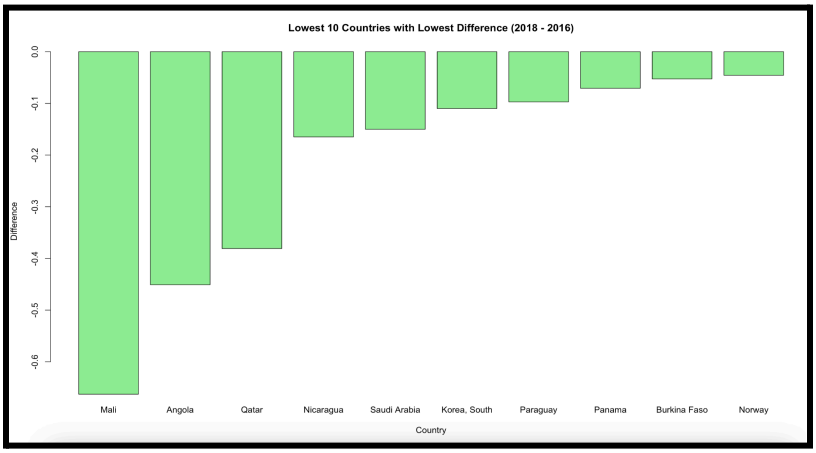
Residuals



We decided to plot the residuals of the linear regression model we had previously. From this graph, we can see that the model underestimates between 0 and 2e+7, overestimates slightly at around 4e+7, and underestimates a lot beyond that. We also see that the residual line is not evenly distributed which could mean that a linear model is not the best choice for this data.

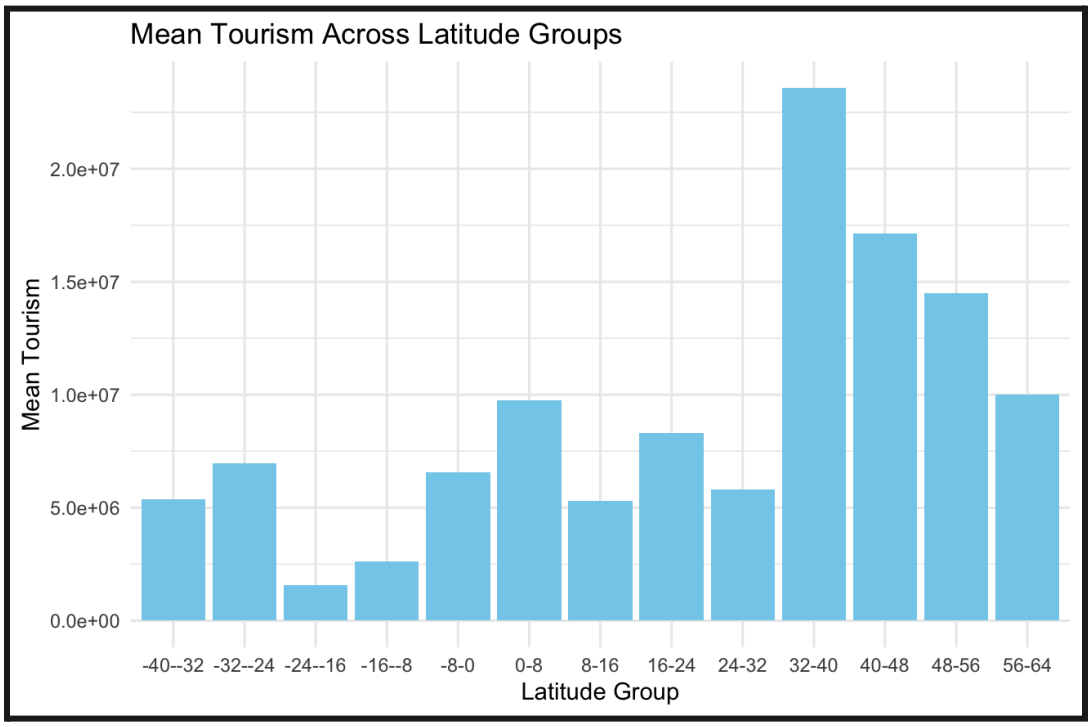
Difference in Differences





From the linear regression model, we decided to create a difference in differences. To begin with, we identified the 10 countries with the highest percentage difference from 2016 to 2018 and the 10 countries with the lowest percentage difference from 2016 to 2018. Tajikistan, Uzbekistan, and Egypt saw the largest increase in tourism percentage with Tajikistan doubling its tourism numbers. On the other hand, Mali, Angola, and Qatar saw the biggest decrease in tourism percentage.

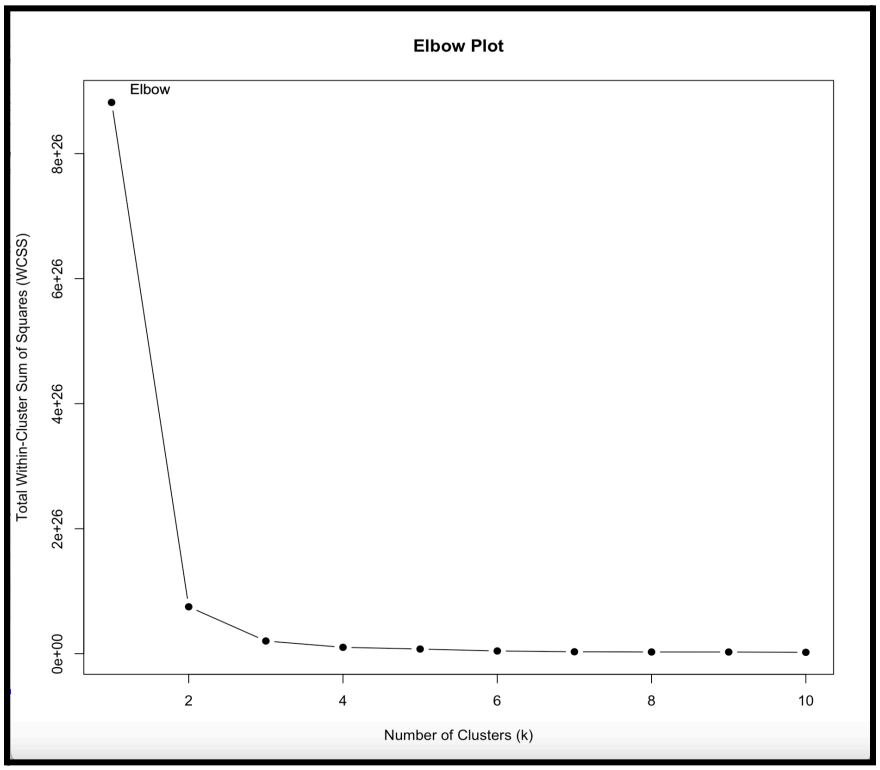
Latitude Analysis



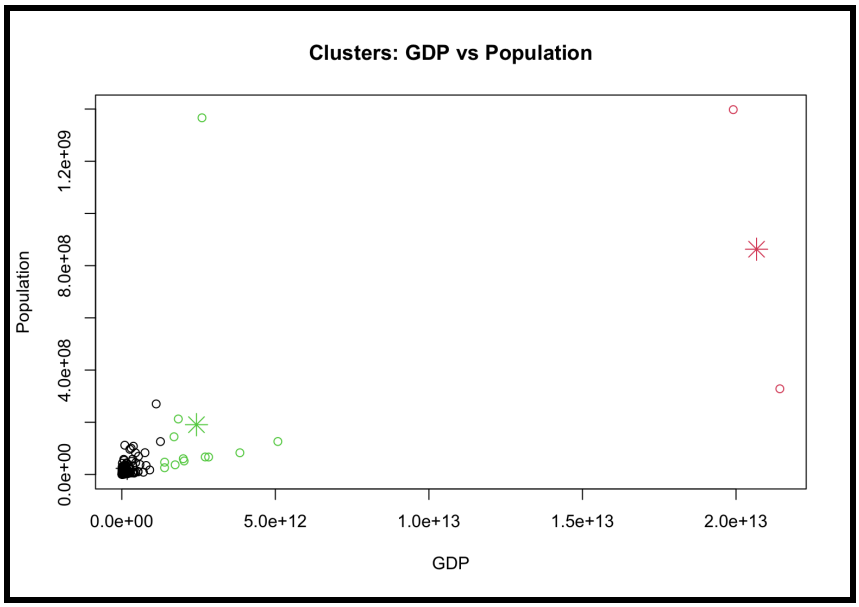
When looking at the results of the linear regression model, we found that latitude was almost a significant model. We decided to create a histogram with bins of size 8 to see if there was some relationship between latitude and mean tourism. What we found is that there was a relationship that we could see from the visualization. The tourism increases drastically from the 24-32 bin to the 32-40 bin. Moreover, we can see that after the 32-40 bin, the mean tourism decreases as the latitude increases and the climate presumably gets colder.

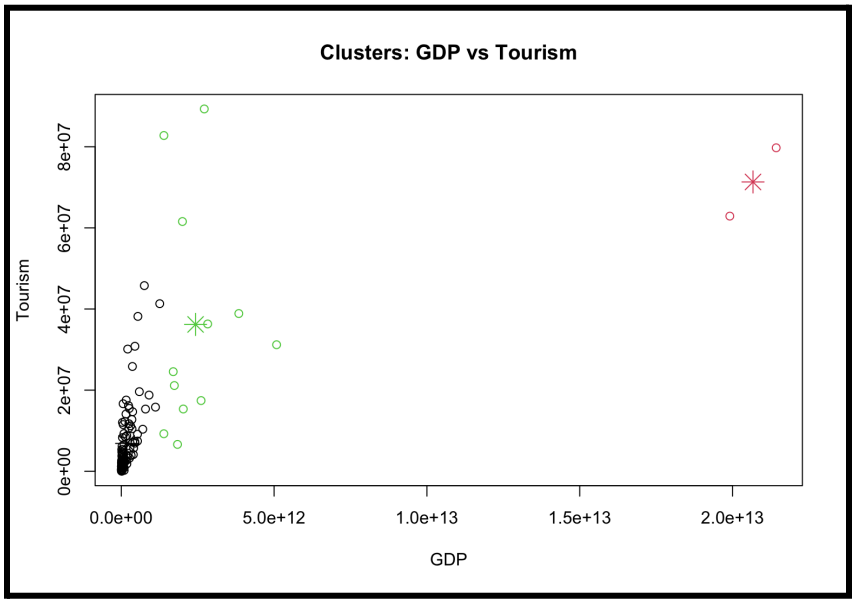
Clustering

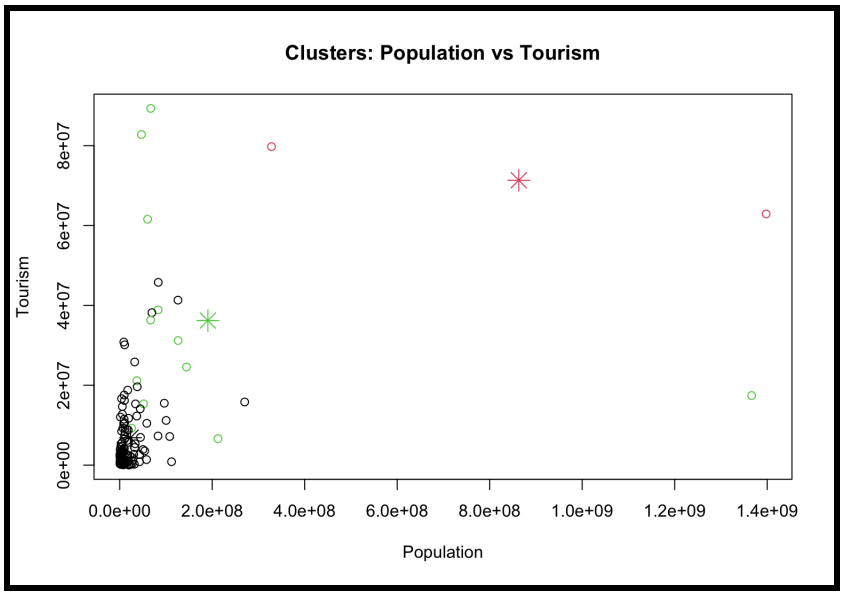
We then decided to do a k-means unsupervised clustering to see if we there were any discernible groups in our data in terms of GDP, population, and tourism.



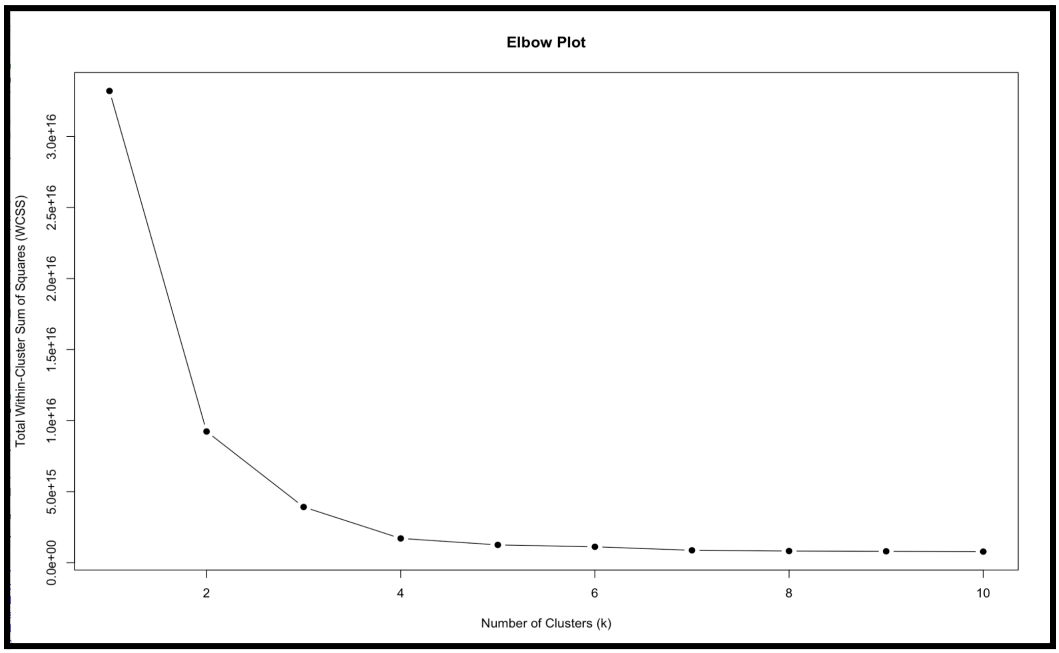
The first thing we did was determine the number of clusters (k) to use. To do this, we built an elbow plot. We saw that after 2 or 3 , the change in total within-cluster sum of squares didn't change much. Because of this, we chose our k to be 3. We could've also chosen are k to be 2 but we went with 3 because we felt that the difference in within-cluster sum of squares between 2 and 3 was a big enough difference.


Clusters: GDP vs Population
Population
0.0e+00
4.0e+08
8.0e+08
1.2e+09
0.0e+00
5.0e+12
1.0e+13
1.5e+13
2.0e+13
GDP

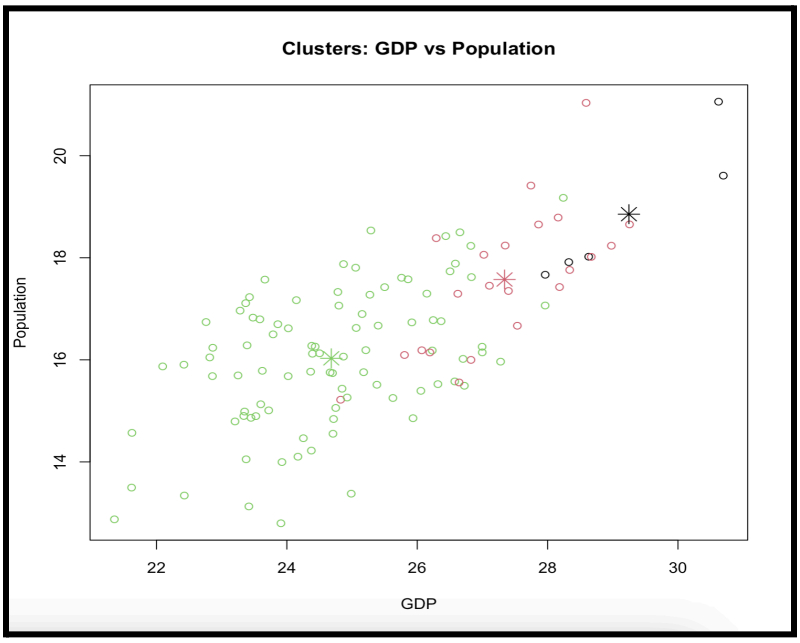

Clusters: GDP vs Tourism
Tourism
0e+00
2e+07
4e+07
6e+07
8e+07
0.0e+00
5.0e+12
1.0e+13
1.5e+13
2.0e+13
GDP

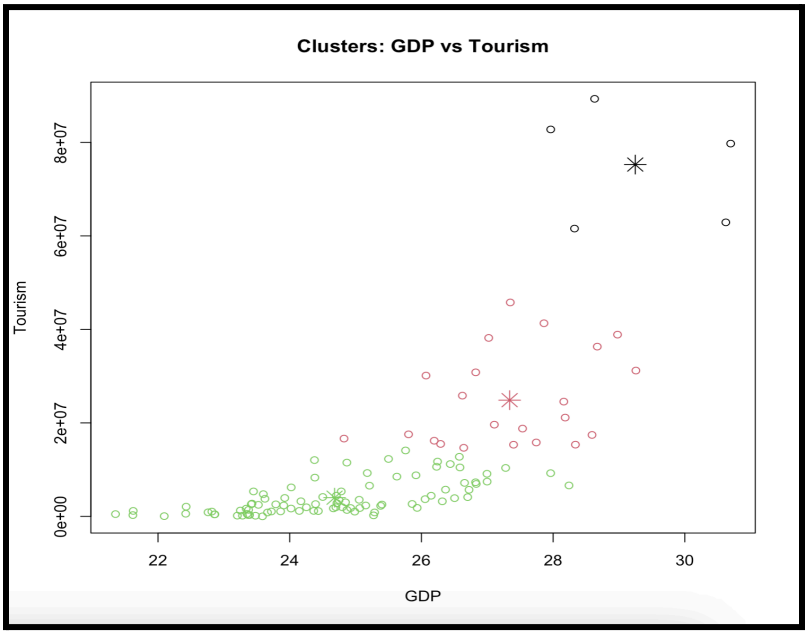


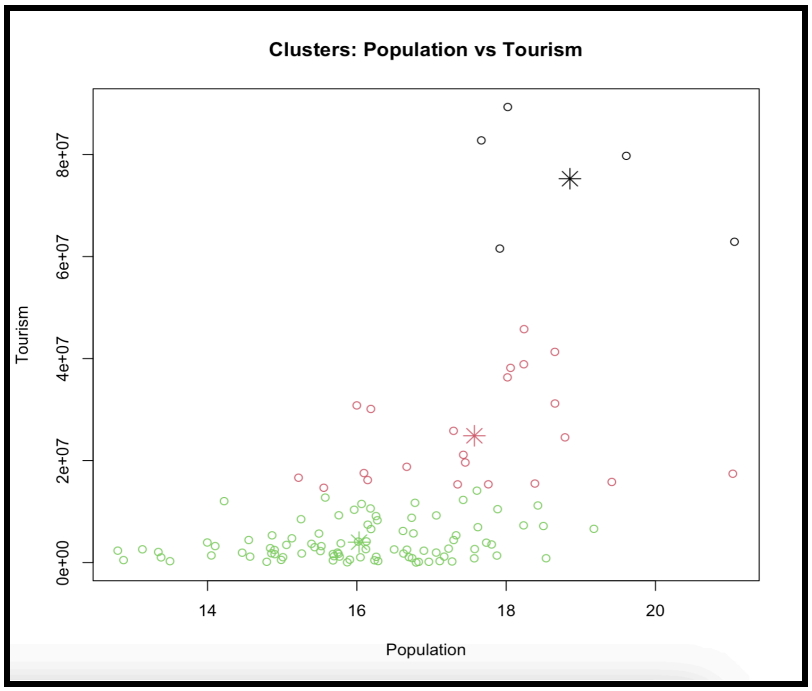
The previous 3 graphs are our findings for the k-means clustering using the GDP, population, and tourism variables. What we can see is that a lot of the data is close to the (0,0) point and a few outliers drag the cluster away. After realizing this, we change our k-means to look at logGDP, logPopulation, and tourism instead.



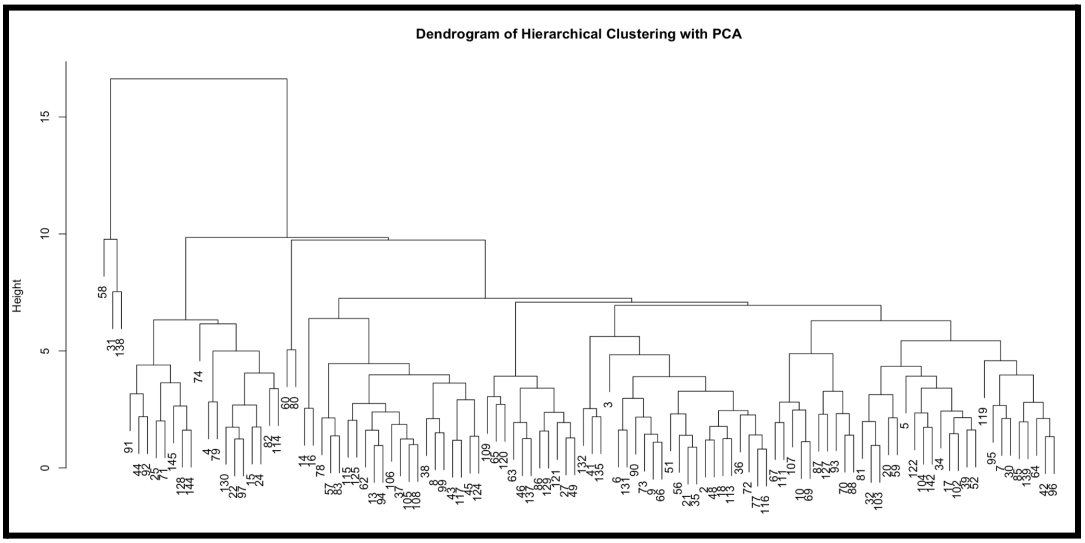
Again, we created an elbow plot for our unsupervised learning for the logGDP, logPopulation, and tourism variables. What we found is that, once again, the number of clusters (k) being 3 was a good enough choice for our clustering. Again, we could've gone with 2 or 4 but chose to stick with 3.







After using the log variables, we can now see 3 discernable cluster where each cluster is a different color and a star depicts the center of that cluster. Generally, it looks like GDP vs Tourism has the most separated groups while GDP vs Population has a bit of overlap. To further gain insights from the k-means clustering, we can look at which countries are in each group. Cluster 1 (green) has a lot of countries with small populations like Egypt, Haiti, Serbia, and Zambia. Cluster 1 also has the most countries in it by far. Cluster 2 (red) has 22 countries in it. What we found is that a lot of these countries are part of the intergovernmental organization G20. They are known to make up many of the world's largest economies and population. Finally, cluster 3 (black) has 5 countries: China, France, Italy, USA, and Spain. These countries have a wide range of populations but all are large tourism destinations. Additionally, just like cluster 2, these countries have some of the largest economies and hold a lot of power in the world.



Finally, we decided to use PCA to get a visualization of a hierarchical clustering for our data as opposed to the non-hierarchical clustering we got from the k-means clustering. What we can see here is a separation between the three countries on the right and the remaining countries in the dendrogram. Specifically, those 3 countries are China, India, and the United States. All those countries differ greatly in tourism. However, all have high populations, significant land area, and large economies.

Conclusion

Our analysis reveals that Gross Domestic Product (GDP) significantly influences the levels of tourism experienced by a country. Notably, countries situated closer to the equator exhibit heightened levels of tourism. However, our analysis indicates that other factors examined do not significantly impact tourism trends.

To improve clustering methods, we incorporated log wage as a variable and compared it to GDP and population size. This produced more clearly delineated clusters, characterized by distinct groups based on population size, land mass, and affiliations with different alliances.

These insights provide valuable guidance for governments seeking to evaluate strategies for boosting tourism levels by benchmarking against peer countries within their respective clusters.

Despite the lack of statistical correlation between many variables and tourism levels, our findings highlight the importance of GDP as a key driver of tourism. This suggests that countries should prioritize efforts to enhance their GDP through investments in infrastructure and government spending to stimulate sustained growth in tourism.

References

Khan, Naushad and Hassan, Absar Ul and Fahad, Shah and Naushad, Mahnoor, Factors Affecting Tourism Industry and Its Impacts on Global Economy of the World (March 23, 2020).