Caribbean Economic Analysis
IBM Data Science Capstone Project

Sergio R. Smith

February 2021

1 Introduction

Largely comprised of small island nations, the Caribbean lies in the tropical region spanning the distance between North and South America, to the east of Central America. Caribbean countries share many geographical and political features, but some historical and cultural definitions broaden the scope of the region to include some Central and South American countries with borders on the Caribbean sea. Two features which are common throughout Caribbean countries are the relatively small population compared to larger countries, and consequently the smaller economies.

Many Caribbean countries rely on tourism and hospitality industries, while others rely on the export of goods. A unifying trait of these economies is the high level of sensitivity to both global economic trends, which affect how many visitors come and how much they spend, and global climate, which impacts sea levels and temperature, in turn affecting beach and reef quality and fish availability.

As the world looks keenly at economic balance and climate change, it is important to evaluate the health of small economies such as those in the Caribbean. Of course, it is also crucial for the countries themselves to recognise what factors contribute to or indicate economic health. The goal of this project is to answer the question, **"How are the prices of establishments such as restaurants, hotels and recreational facilities related to the gross domestic product (GDP) of Caribbean countries?"** While not seeking to imply a causal relationship between features, I will use available

data to investigate the extent to which the costs of goods and services describe or predict two measures of GDP.

2 Data needs

In order to meaningfully approach my research question, several pieces of data will be necessary. Fortunately, the data I need are available on the web to those who search well.

First, I will identify those countries which I think are relevant to the study. Multiple agencies and individuals have different ideas of what constitutes a Caribbean country. In the interest of collecting enough data, my list will not be too conservative, but not so liberal that I include countries outside the Caribbean with different population and economic profiles. I also need to make sure that the countries I select have economic data available. Territories such as the British and U.S. Virgin islands may also be excluded if they are counted as part of different economies.

I will need to source data on the GDP of those countries I choose to study. For this project I am studying two measures of GDP: nominal and purchasing power parity (PPP). I need to identify a source which reliably records and publishes these records, which are part or normal economic reporting for a country. The availability of these data may affect the scope of this study; Cuba, for example, does not have readily available documentation of economic measures, and will likely be omitted.

I will use the Foursquare ReST API to collect the data on the costs of goods and services in various countries. I will use data on the venues within a fixed radius of the coordinates of each country's capital. Determining this radius will pose some challenge as there is significant variability in the sizes of capitals within this group. Part of this determination may rely on the precision to which capital coordinates are available. To that note, capital city coordinates are available to be scraped from the web, but will likely have to be cleaned depending on coordinate notation (e.g. 45.30W vs. -45.30).

3 Methodology

Carrying out this project included two major steps: data collection/cleaning and machine learning analysis. In the first phase, I scraped data from the

internet, converted the data to usable formats, and integrated the data into a unified data object for analysis. In the second phase, I used machine learning techniques to build a model of the data and evaluate its effectiveness.

I used a notebook in the JupyterLab service to do all of my work. Jupyterlab provides a Python environment with many data manipulation tools and packages included, and the cell-based notebook structure facilitates real-time debugging and tuning. I used the Anaconda suite to access JupyterLab.

3.1 Data Selection, Collection and Cleaning

For this project, I needed to define what countries to study before collecting and analysing data. After looking into multiple definitions and lists of Caribbean countries, I settled on those countries which were members of the regional body CARICOM. These 15 member states comprise many of the dominant Caribbean economies and governments, and were for the most part independent states whose GDP would not be counted as part of a larger economy. I would use the coordinates of the capital cities, which I thought would return the most venues - the major assumption here being that the capital cities are the largest cities, which seems to be true for Caribbean countries.

In order to collect price data on venues in the capital cities, I needed the coordinates of these cities to pass to the Foursquare API. Rather than use Google's geolocator service, which has given me trouble before, I opted to scrape the coordinates from a table I found online. This table includes capitals for a huge list of countries, which would eventually be pared down to only the countries I decided to study. I imported the table using the *requests* package, and processed it using the *BeautifulSoup* package. The data then needed to be cleaned of web table formatting, and the coordinate data was processed into a uniform integer format. Two Countries of interest were not on this list and were researched independently and added.

Next, I collected data on the GDP of the countries of interest. Wikipedia provided lists of two measures of GDP, nominal and purchasing power parity (PPP), in two tables on separate web pages. These tables contained data on Latin American and Caribbean countries, which don not match the list of countries I had in mind. I created a general function for importing tables from webpages, which I used to scrape this data. I processed the data in a similar manner to the coordinate data and dropped unnecessary columns (gross GDP, for example, is much less useful than per capita GDP for this

analysis). One of my countries of interest was not listed in these tables, presumably it was considered part of another nation's economy. Rather than combine GDP data from multiple (possibly inconsistent) sources, I elected to drop this country from my final list.

Once the coordinate and GDP data were collected and formatted, I performed a *merge* to combine them, keeping only the (now 14) countries of interest to this study. As a useful illustration, I used *folium* to create a map showing relative nominal GDP of cities, seen in Figure 1.

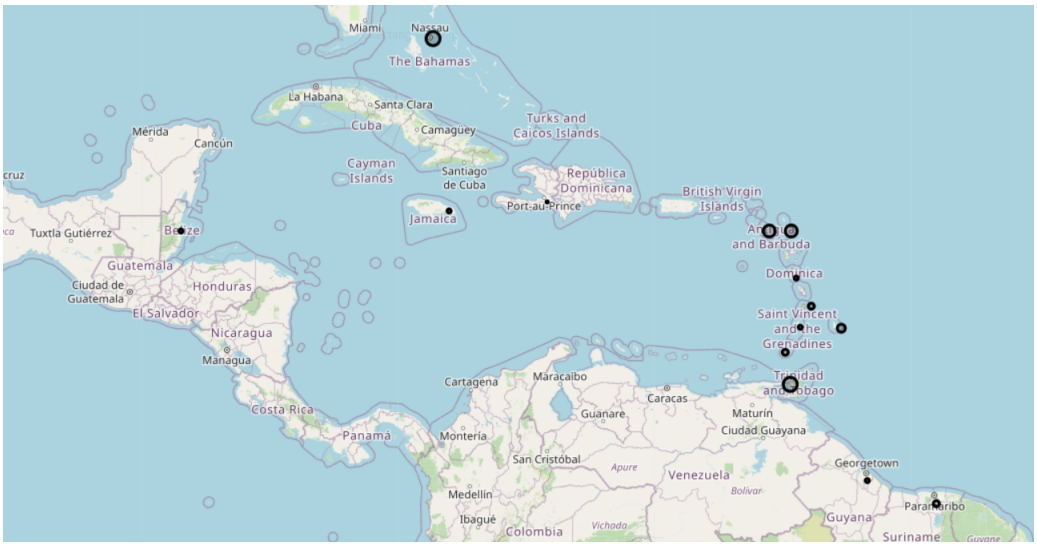


Figure 1: The Caribbean, with relative PPP per capita of selected countries

Next I set about collecting the data on venue prices. Foursquare organises venues into seven general categories, including Arts and Entertainment, Shopping, and Outdoor and Recreation. My hope was to collect data on venues of each of these categories, which would each serve as independent variables. Significant testing revealed that certain limitations (addressed in the Discussion) severely limited the available data. In the end, I only used the Food and Nightlife Spot categories.

I iterated over the list of countries, categories and venues in turn, averaging venue prices by category and recording each category for each country. The program involved a basic API call to collect venue IDs for matching venues in each country, and a "premium" call to get details of each venue, including price. Price was recorded in tiers from 1 to 4. Care was taken to use error-catching structures in case no venues or prices were returned.

Finally, the venue price data was appended to Data were only available for 13 of the 14 countries examined, and thus a second entry was dropped from the final data set as the new data was combined with the coordinate and GDP data. Another map was produced, comparing the prices of the two categories for each capital (Figure 2).

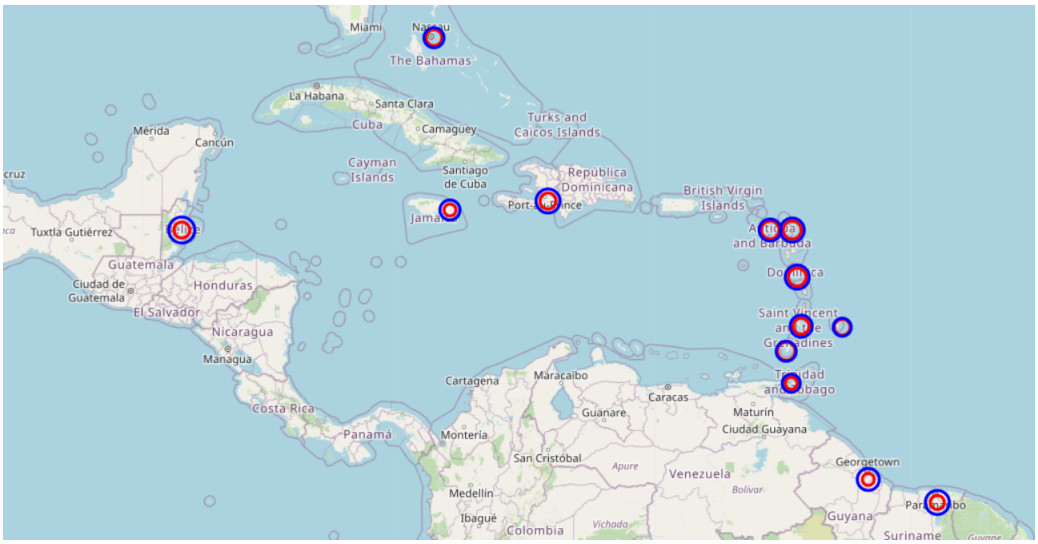


Figure 2: The Caribbean, with relative Food (red) and Nightlife (blue) price

3.2 Data analysis

3.2.1 Preliminary investigations

Before addressing the primary problem, I wanted to investigate a couple of simpler questions to get a feel for the data's properties.

First, I wanted to check whether the two types of GDP, nominal and PPP, are correlated. Both are measures of the economic output of countries, but PPP takes into consideration the buying power within the economy. Intuitively, I would guess that the two were at least weakly correlated, and could test this using simple linear regression. Using the *sklearn* package, I built a regression model and trained it with one GDP as the independent variable and the other as the dependent variable.

Next, I wondered if Food prices and Nightlife prices were correlated for given countries. One might suspect that countries with expensive products in

one category might have high prices in the other. Again, I built a regression model, using one category as the independent variable and the other as a dependent variable.

3.2.2 Primary analysis

Th goal of this project has been to build a model which could use venue prices to explain or predict the GDP, by two different measures, of Caribbean countries. In this section I addressed this.

I trained a multiple linear regression model with two explanatory variables (average Food and Nightlife venue prices for each country) and one dependent variable (either nominal or PPP GDP, one at a time). For each type of GDP, I created a model of the form

$$GDP = \beta_0 + \beta_1 * (Food\ price) + \beta_2 * (Nightlife\ price)$$

where the intercept β_0 and the coefficients β_1, β_2 are the outputs of the machine learning algorithm. In addition to fitting the model, I also created and inspected 3D plots of the GDP as a function of the two prices for all countries, and based on those plots I planned follow-up analysis.

Details about decisions made regarding model evaluation are addressed in the Discussion section.

3.2.3 Follow-up analysis

Based on observations made of the scatter plots showing GDP vs prices, I suspected that the Nightlife values might be more predictive of GDP than Food venue values, and that the Food values might be skewing the model and making it less effective. I trained a simple linear regression model to examine the relationship between these two values, and plotted the results with the regression line.

4 Results

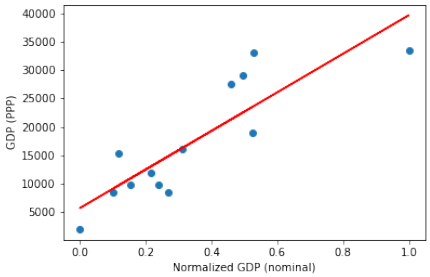
The final data used for analysis is given in Figure 3. It should be noticed that the units for the GDP (nominal) and GDP (PPP) are not shown, but for the primary analysis these values are both normalised anyway. Their original units are United States dollars and International dollars, respectively.

Country	GDP (PPP) per capita	GDP (nominal) per capita	Capital	Latitude	Longitude	Food average price tier	Nightlife average price tier
Antigua and Barbuda	27542	14159	Saint John's	17.20	-61.48	1.894737	2.400000
Barbados	18866	16082	Bridgetown	13.05	-59.30	1.526316	1.875000
Belize	8467	3734	Belmopan	17.18	-88.30	1.500000	2.500000
Dominica	9726	7709	Roseau	15.20	-61.24	1.736842	2.357143
Grenada	16033	9824	Saint George's	12.05	-61.75	1.722222	1.947368
Guyana	8524	8649	Georgetown	6.50	-58.12	1.125000	2.200000
Haiti	1940	732	Port-au-Prince	18.40	-72.20	1.600000	2.312500
Jamaica	9726	5221	Kingston	18.00	-76.50	1.277778	2.066667
Saint Kitts and Nevis	29098	15246	Basseterre	17.17	-62.43	1.894737	2.125000
Saint Vincent and the Grenadines	11965	7033	Kingstown	13.10	-61.10	1.533333	2.250000
Suriname	15362	4199	Paramaribo	5.50	-55.10	1.315789	2.333333
The Bahamas	33516	30027	Nassau	25.05	-77.20	1.650000	2.000000
Trinidad and Tobago	33026	16197	Port of Spain	10.67	-61.52	1.157895	1.700000

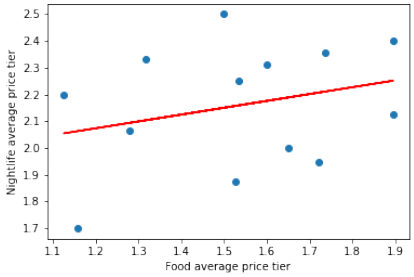
Figure 3: Dataframe showing all data used

4.1 Preliminary investigation

First I present the results of the preliminary investigation which attempted to correlate the two measures of GDP together and then to correlate the two category prices. The regression results are in Table 1, while the scatter plots with regression lines are shown in 4. Note that in both cases, the data has NOT been normalised.



(a) Nominal vs PPP GDP regression



(b) Food vs. Nightlife price regression

Figure 4: Preliminary analysis graphs

Property	Value
Intercept	4822
Slope	1.16
Regression Score	0.741

(a) Nominal vs PPP regression

Property	Value
Intercept	1.77
Slope	0.255
Regression Score	0.078

(b) Food vs Nightlife price regression

Table 1: Preliminary analysis results

4.2 Primary analysis

The planes produced by the multiple regression have three parameters: the slopes of the plane with respect to each independent variable, and the intercept. The results for the two different GDP measures are reported in Table 2. Scatter plots were also produced for the nominal GDP as a function of the two categories' values. Multiple perspectives serve as effective 2D graphs, showing how GDP varies with each of the two categories' prices, and are shown in Figure 5. GDP is normalised for the scatter plot. All variables are normalised for the regression results table.

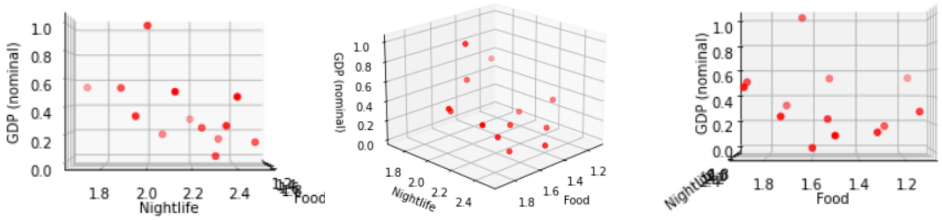


Figure 5: GDP vs Food and Nightlife prices

4.3 Follow-up analysis

Finally, the results of the follow-up analysis are given here in Figure 6. This includes scatter plots of both types of GDP against the Nightlife values for each country, shown with the corresponding regression lines. I also include Table 3 showing the regression statistics. All values are normalised for these regressions and plots.

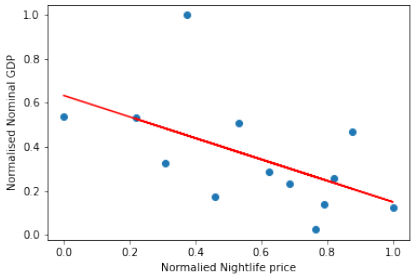
Property	Value
Intercept	0.520
Food Slope	0.326
Nightlife Slope	-0.589
Regression Score	0.464

(a) Nominal GDP multiple regression results

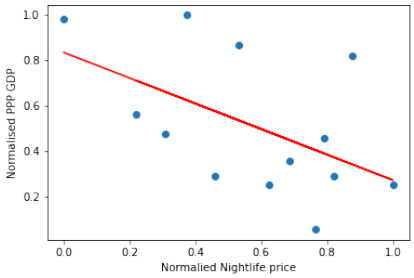
Property	Value
Intercept	0.702
Food Slope	0.386
Nightlife Slope	-0.684
Regression Score	0.433

(b) PPP GDP multiple regression results

Table 2: Multiple regression results



(a) GDP (nominal) vs. Nightlife price with regression



(b) GDP (PPP) vs. Nightlife price with regression

Figure 6: Preliminary analysis graphs

Property	Value
Intercept	0.634
Slope	-0.485
Regression Score	0.301

(a) GDP (nominal) vs. Nightlife price regression

Property	Value
Intercept	0.836
Slope	-0.562
Regression Score	0.277

(b) GDP (PPP) vs. Nightlife price regression

Table 3: Follow-up analysis results

5 Discussion

5.1 Data availability and model evaluation

Before getting into discussing the results of the analysis, I should discuss the availability of data, which played a large part in directing the progress

of this project. Coordinate and GDP data were readily available, with the exception of one country whose economy was a subset of another; I would likely have excluded that country anyway because its economy would be difficult to compare with those of independent states.

On the other hand, data on venues in these Caribbean countries was very sparse. I had hoped to use multiple regression with far more than two independent variables. However, in my testing, only two out of seven venue categories returned enough venues (and particularly venues with price data available) to perform my analysis. Even for these two categories, one country had no price data available and had to be dropped from the analysis altogether. I believe this is a result of the relatively small geographic and population sizes of the countries and cities investigated. I had hoped that, given the reliance of many countries on tourism, a steady flow of tourists would have reviewed more venues using Foursquare's review service. This was not the case.

In addition to the small number of venues available in many cases, those venues which were reported often did not have price information available. In these cases, I could not check more venues arbitrarily for price because I was limited in the number of "premium calls" for the Foursquare API. Each time I requested data on a given venue (to even check if price was available), that call counted against a small allotment for my personal account. I regret that the lowest available paid tier is far outside the typical non-commercial user's budget, and I suspect that this project could have benefited significantly from a larger API call allotment.

Ordinarily, constructing a model would have included more testing to measure effectiveness. This would include performing a test/train split on the data set and building the model iteratively using training subsets. I made a decision not to use a test/train split because my data contained only 13 records (countries). I felt that even a 25% test set would result in too small a training set to build a proper model. Instead, I opted to train the model on all available data and rely on the R^2 regression score to evaluate the model.

The discussion that follows is provided given these caveats about the size of the data set and means of evaluation.

5.2 Interpreting the data

As mentioned above, the regression models produced in this analysis use the data provided to create a straight line (or a plane, with multiple inde-

pendent variables) and return values for the coefficients (slopes) associated with each independent variables as well as an intercept or offset.

The slope of the line indicates how much the the dependent variable increases when the independent (predictive) variable increases. A slope of 0 indicates that the dependent variable is constant with respect to the independent variable. A large value for the slope suggests a high degree of dependence, and a negative slope indicates that one variable decreases as the other increases.

The regression score, or R^2 score, tells us how well the data fits the model, and thus how confident we can be in the model. It is a measure of error, or how far each data point is from the regression line. A score of 1 indicates perfect agreement of data and model, and a score of 0 indicates that the data points are essentially randomly distributed. A negative score is also possible if the model fits the data particularly poorly.

5.3 Preliminary investigation

The first question I asked of the data was whether or not there was a correlation between the two measures of GDP studied. The results showed a strong relationship between the two, with a regression score of 0.741 (1 being the best possible score). This is perhaps unsurprising, as both are metrics used to assess the health of an economy and are similar in derivation. The slope was close to 1, suggesting a nearly 1:1 relationship. However, I take that with caution because this data was not normalised and the two quantities had different dollar units.

The second question was whether the prices of the two venue categories were correlated. This appears not to be the case, with a very poor regression score of 0.078. The two quantities appear to be virtually unrelated. This did surprise me, as I expected that as restaurant prices increased for a country, so might bar, nightclub, and party prices. I am not convinced my hypothesis is incorrect, and I suspect that the small amount of data (only a few prices per category in some countries) contributed to this low value. As it stands, we cannot infer a correlation from this data.

5.4 Primary analysis

The multiple regression model for each GDP type returned two coefficients or slopes, one for each venue category, as seen in Table 2. The slopes

for both independent variables roughly matched between the two types of GDP, which is not surprising given that we have already established a correlation between the two types of GDP. Similarly, the regression scores do not vary significantly. With regression scores $\tilde{0}.45$, we can infer that there is some meaningful relationship between the predictive and predicted variables. The slopes do not indicate that GDP is hugely dependent on either category price in this model, be we are at least reassured that the data does fit the model reasonably well (as it should, given that the model was trained on this data). The coefficients suggest that GDP (both types) increases slightly with restaurant prices and decreases moderately with Nightlife venue prices.

5.5 Follow-up analysis

Based on the distribution of points in the 3D scatter plot, I suspected that there could be a stronger relationship between Nightlife venues and GDP, and that the multiple regression might disguise this. I followed up by training a simple regression model using only the Nightlife independent variable. The results were not overwhelming, with a regression score of only $\tilde{0}.3$ in both cases. It appears that the multiple regression model fit the data better than using a single independent variable.

6 Conclusion

Despite issues regarding data availability, I was able to carry out not only the main analysis for this project, but also ask related questions of the data. I have shown by use of a multiple regression model that the prices of certain types of venues can be used with moderate reliability to predict GDP, and hence are useful as predictors of economic health in Caribbean countries. I was also able to confirm a relationship between two measures of GDP for these 13 countries.

There is a lot of room for follow-up with this project. Certainly it is worth trying to get more data on the countries studied. Given the daily API call limits, the data would have to be collected over several days, or collected using a commercial account on the Foursquare API, which charges a monthly fee and a fee per API call.

Building the model using more Caribbean countries might also be interesting. Having more data records would make splitting the data into testing

and training sets more feasible, so a better measure of model effectiveness could be achieved. Similarly, testing how this model performs on countries outside of the Caribbean would be interesting. It could be worthwhile to train similar models on countries outside the Caribbean, and testing it on the countries I studied.

The results of this analysis could be useful to economists interested in alternative measures of economic health and growth. The results at the very least hint that further research is indicated, and further avenues have been addressed above. I consider the project to have been worthwhile and a good demonstration of some strategies and techniques essential to the data science process, and I am pleased to submit it for consideration as an IBM Data Science Capstone project.