STAT 410 Project Report

Countries' Cost of Living

The purpose of this analysis is to investigate the factors that influence the cost of living in various countries around the world. I chose this topic because it aligned with my personal interests, and because cost of living is an important consideration for virtually every person or organization that needs to organize finances in some way or form. The everyday person who has a salary; pays taxes; and has rent, utilities, and groceries, among other expenses may be able to use the findings from this project to gain insight into the cost of living of their respective country, and what it means in relation to their personal finances. Multinational businesses can use this analysis to inform their decisions in building connections and expanding internationally, as cost of living directly affects wages, overhead costs, transportation costs, and many other key factors. Combined with knowledge of the socioeconomic, cultural, and political status of the countries in question, we may be able to draw conclusions about the reasoning behind a country's cost of living, and what future trends may be likely.

Two datasets were used in the conduction of this analysis: one that indicated the cost of living of each country, and the other that had vital statistics and other information of each country. The cost of living dataset originates from Numbeo.com, a website created in 2009 by former software engineer Mladen Adamovic. Numbeo's Cost of Living Index aims to provide an accurate metric of cost of living by combining information from local user surveys, along with "authoritative sources" such as "websites of supermarkets, taxi company websites, governmental institutions, newspaper articles, other surveys, etc." Price estimates for a wide variety of commodities - which fall under the categories clothing and shoes, markets, rent per month,

restaurants, sports and leisure, transportation, and utilities - are collected, aggregated, and processed to generate a formula that indicates the country's "consumer basket plus rent." The other dataset came from World Bank's World Development Indicators data manipulation tool, which allows users to generate datasets and visualizations from a selection of 266 countries/regions; 1445 economic, social, and political features; and years ranging from 1974 to the present. These data would serve as the predictors which I would use to predict cost of living.

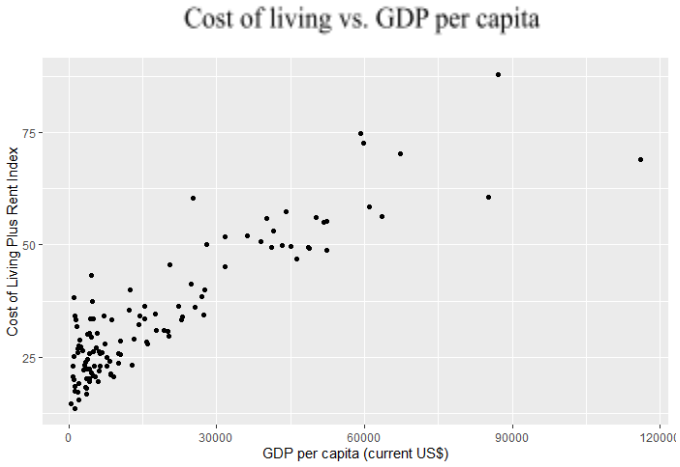
Substantial cleaning of the World Bank dataset was necessary in order to convert it into a format eligible for data analysis. Immediately apparent upon preparation of the dataset from the World Bank website was that the "shape" or structure was undesirable and confusing: the years were on the columns, while the features and countries were both on the rows. As the Numbeo dataset came from the year 2020, the data were filtered such that only results from that year appeared. The data were then pivoted such that the countries (the observations of the analysis) were on each row, and the predictors were on the columns. Next came the tedious process of handling NA values, which comprised the vast majority of the values in the entire dataset. Beginning with the rows, I removed the countries above the 90% quantile for number of missing values, as the majority of these included regions that weren't countries at all (e.g. American Samoa, British Virgin Islands, Isle of Man), or very small countries (e.g. Liechtenstein, Nauru, Tuvalu). I then examined the columns, and noted that a significant portion of them had NA values for all 195 rows; I removed these. I repeated this process until I was left with a dataset with 165 countries and 143 predictors, taking care to not determine a cutoff for the number of NA values that would remove important countries/columns, or as I found a couple times, everything from the dataset. I considered imputing missing values so that there would be more

columns to use as predictors, however, I decided against this as I wasn't particularly comfortable with the packages necessary for imputation, as it was not covered in class.

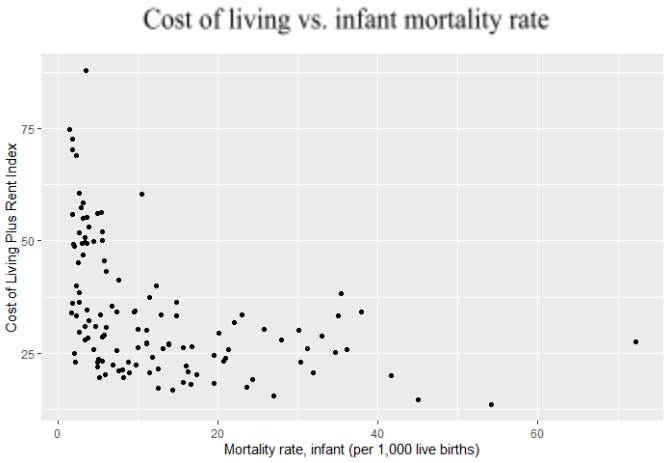
Next, upon joining the two datasets, I noted that a few of the countries from the World Bank were not matched with their respective observations in the cost of living. To ensure as many countries were included in the analysis as possible, I examined names of countries that were not joined and identified patterns in their reporting that differed between the two datasets. For instance, the country whose capital is Cairo is listed as "Egypt, Arab Rep." in World Bank, while it is simply listed as "Egypt" in COL. Several other instances of superfluous commas and titles after the country name were identified, such as "Korea, Rep." instead of "South Korea." I mutated the country names to match those in the COL data, using regular expressions to change common patterns such as the comma situation, and handling others on a case by case basis, for example changing "Kyrgyz Republic" to its more colloquial name "Kyrgyzstan." I then examined the columns and noticed that many of them were either almost the exact same thing as another column (e.g. GDP per capita, constant 2017 international $ and GDP per capita, current international $), strangely specific and thus not particularly useful to the analysis (e,g, Secondary education, duration (referring to the length of secondary school in the country, not the length of secondary education the average student receives)), or a function of other columns (e.g. Age Dependency Ratio, a ratio of working age population to total population). I removed these, along with predictors that would inherently favor countries with a larger population, such as GDP and any population figure that was a raw number rather than a rate or percentage. This left me with 24 predictors, with which I could successfully conduct my analysis.

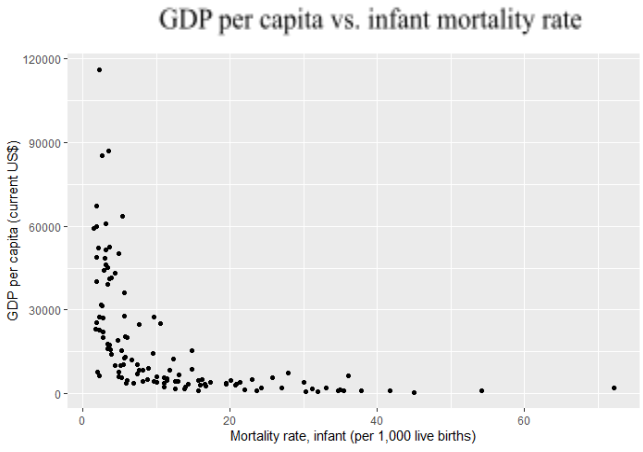
My initial speculation was that GDP per capita, and by extension factors associated with GDP per capita, would correlate with a higher cost of living, as countries with a richer

population would be able to pay greater prices for goods and services. This was affirmed by my exploratory data analysis, which showed that the two variables had a fairly strong linear relationship.



As an example of one predictor associated with GDP per capita, I looked at infant mortality rate, and how it related with both cost of living and GDP per capita itself.

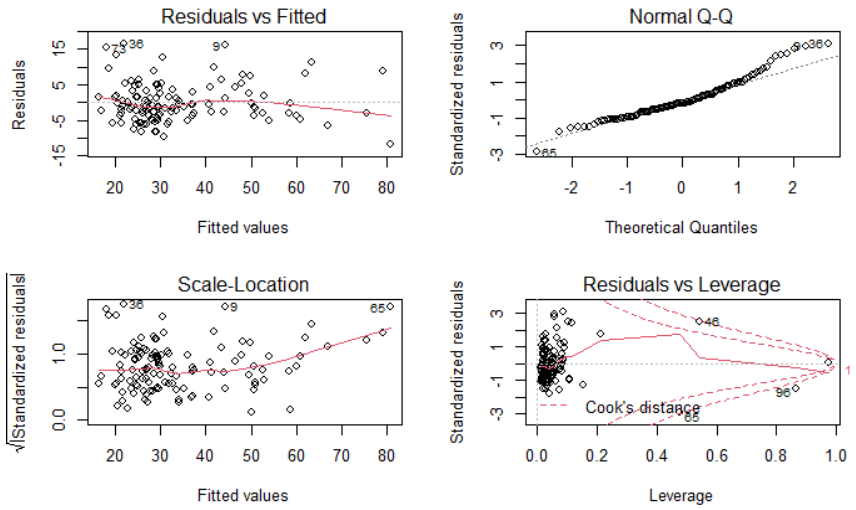




Indeed, there is an association between mortality rate and GDP per capita; this is something to note with caution during the model building process. In addition, the relationship does not seem perfectly linear, interestingly.

I decided to perform variable selection by AIC in order to obtain a model predicting cost of living based upon the predictors in the data. Initially, as an exploratory exercise, I fit an MLR model containing each of the predictors from the data, and found that while the p-value was sufficiently low and the adjusted R squared was a respectable 0.8387, only four of the 24 predictors were significant. Therefore, I determined that a much simpler model would be more appropriate. I decided to utilize variable selection with forward search, as I felt that beginning with no variables and adding predictors individually would likely be the most effective method of computation, given that only a few of the predictors may be significant. I decided against using a similar approach that minimized purely RSS as opposed to AIC, as this may have been prone to adding extraneous variables. This resulted in a model with seven predictors: GDP per capita, GDP per capita % growth, labor force participation rate, population density, foreign direct investment (% of GDP), conversion factor (i.e. exchange rate against the US dollar), and % urban population. Of these parameters, five were significant at a $p < .05$ (and .01) level, and the adjusted R squared improved slightly to 0.8514. Out of curiosity, I performed a backward AIC search on the data to see if inclusion of any other variables would be relevant, however, this yielded the exact same model as the forward search. I then performed a partial F test with this model and a reduced model with the insignificant parameters removed, to see if the model could be simplified without sacrificing accuracy. Each of the five parameters in the reduced model remained significant with the removal of the other two, and the ANOVA test did not show statistical significance; I could not conclude that the coefficients of the insignificant variables were significantly different from zero. However, the AIC for the reduced model was higher by about two points, increasing from 764.48 to 766.49. In spite of the partial F test, I decided to proceed with the full model albeit with caution, keeping the reduced model stored as a variable.

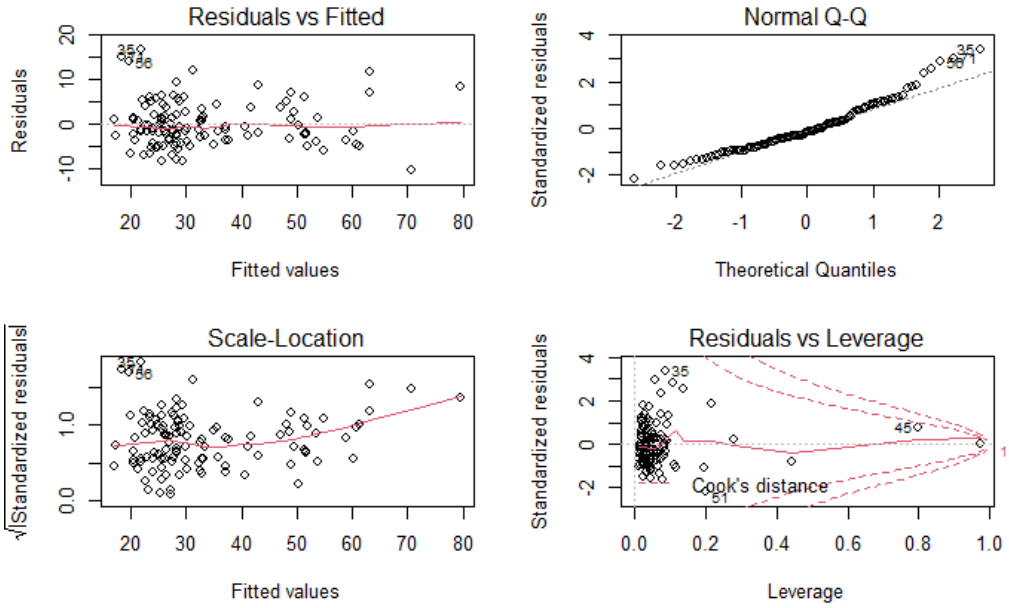
I then examined the diagnostic plots to test the validity of the model assumptions and to see if there were any significant outliers and/or leverage points.



The right tail of the Normal Q-Q plot was somewhat concerning as it curved upward above the line a bit, potentially violating linearity, which was also reflected in the Scale-Location plot. I noticed that there were a few observations whose predicted values were rather far off from their actual values, notably 9, 36, and 65, which corresponded to the Bahamas, Ethiopia, and Luxembourg, respectively. I speculated that due to the Bahamas' large tourism industry and presence, prices would be driven above their predicted value. Luxembourg has a very high GDP per capita due to its small population (just over 600,000) and business friendly governance, potentially leading the model to estimate its cost of living as higher than it actually is. I could not reason an explanation for Ethiopia's unusually high cost of living from my background knowledge and brief research. Next, I moved onto the points that the Residuals vs. Leverage plot indicated as being bad leverage points. Along with the aforementioned Luxembourg, Singapore and Hungary were shown to be beyond Cook's Distance. I reasoned that because Singapore is an exclusively urban country consisting of one continuous metropolitan area, its

urban population value of 100 may have caused the model to overestimate its cost of living. I could not identify an anomaly about Hungary that would have caused its outlier/leverage point status. Thus, I removed from the dataset the countries for which I could ascertain a socioeconomic anomaly that may have caused the model to inaccurately predict its cost of living.

Upon fitting the same model on the dataset minus the Bahamas, Luxembourg, and Singapore, I found that only four predictors showed significance; foreign direct investment lost its significance under the new data. In addition, adjusted R squared showed a modest increase. Upon examination of the diagnostic plots, I saw that this model suffered many of the same flaws as the previous model, namely an unsatisfactorily high right tail on the Normal Q-Q plot. I identified the outliers on the plot as Jordan, Myanmar, and Ethiopia, for none of which I could identify a valid explanation for their high residual values.



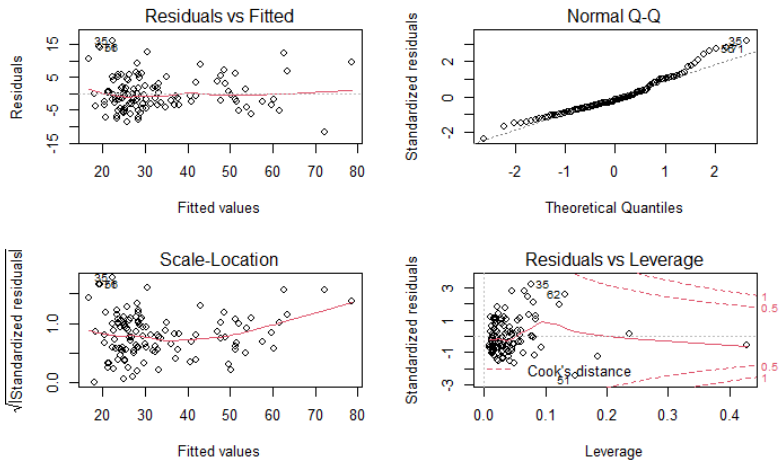
For this reason, along with not wanting to enter an endless loop of removing outliers, and also seeing that there were no longer any bad leverage points, I decided to stop at this model. I then ran another partial F test, this time comparing the full model against a reduced model with the

now three insignificant parameters removed. This yielded an even less significant p-value of 0.1158, and a comparison of the AIC values of the two models showed that the reduced model was only greater by a mere 0.3 (726.79 vs. 726.47). This was also a significant improvement on the model that included outliers, which had a greater AIC value by about 38 points. For these reasons, I decided to keep the reduced model as my final model, as it greatly simplified the model while only increasing error by a relatively trivial amount. This reduced model had GDP per capita, GDP per capita % growth, labor force participation rate, and population density as predictors, each of which were significant at a .01 level; the intercept was also significant at a .05 level. The adjusted R squared was 0.8534, and the p value was sufficiently low.

```
Coefficients:
                                                                                  Estimate
(Intercept)                                                                     8.242e+00
`GDP per capita (current US$)`                                                  6.350e-04
`GDP per capita growth (annual %)`                                             -4.022e-01
`Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)` 1.868e-01
`Population density (people per sq. km of land area)`                           4.660e-03
```

Final model R outputs

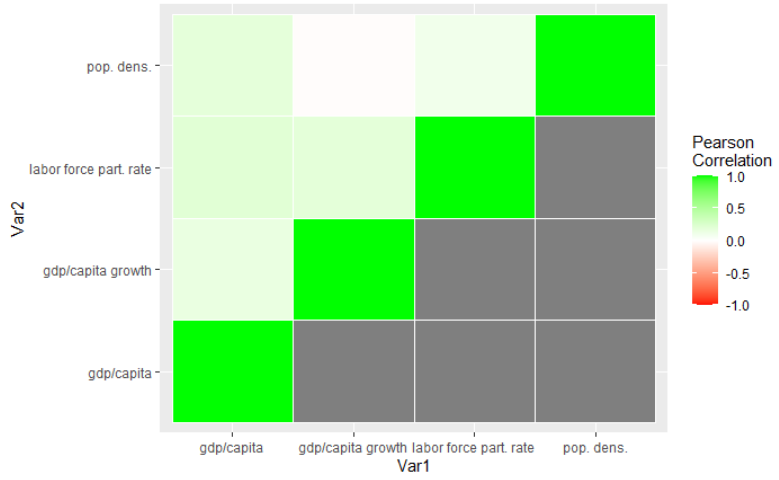
```
                                                                                  Pr(>|t|)
(Intercept)                                                                      0.013528
`GDP per capita (current US$)`                                                    < 2e-16
`GDP per capita growth (annual %)`                                               0.000234
`Labor force participation rate, total (% of total population ages 15+) (modeled ILO estimate)` 0.000539
`Population density (people per sq. km of land area)`                            0.006887
```




```
Residual standard error: 5.247 on 112 degrees of freedom
Multiple R-squared:  0.8584,    Adjusted R-squared:  0.8534
F-statistic: 169.8 on 4 and 112 DF,  p-value: < 2.2e-16
```

Finally, I wanted to test the correlation between the parameters in my model to ensure that they didn't have significant correlations with each other, as the infant mortality rate/GDP per capita plot brought to my attention. I generated a heatmap plotting all four variables against each other

by Pearson correlation, which showed that the highest correlation coefficient was 0.209, between GDP per capita and labor force participation rate. Thus, I concluded that the predictors did not have much significant correlation between each other.



From the final model, we can conclude that GDP per capita, labor force participation rate, and population density positively correlate with a country's cost of living, while GDP per capita growth has a negative correlation. The GDP per capita finding more or less affirms my initial hypothesis, and the other predictors may have interesting implications. Because developing countries such as China and India tend to have higher percentage rates of GDP per capita growth, it follows that countries with higher GDP per capita growth rates have a lower cost of living. Alternatively, one could make the opposite conclusion: countries that are growing in GDP per capita are doing so as a result of their lower cost of living, as businesses may be attracted to lower operating costs. It is no surprise, then, that outsourcing has become a commonplace part of the post-war world. Population density can be explained by the fact that countries with a higher population density are more dominated by urban areas, which tend to have a higher cost of living than rural areas. Countries with a high labor force participation rate, perhaps, have less people reliant on familial networks or social welfare programs for their economic survival,

increasing cost of living. In order to make firm conclusions, however, a greater understanding of the macroeconomic principles of the world than my own is necessary. It is important to note that the coefficient estimates in the R output do not scale directly to the size of the effect of the given predictor. GDP per capita ranges from a few hundred to upwards of six figures, while GDP per capita growth and labor force participation rate are percentages and thus limited to a 0-100 scale. Population density varies from the single digits to the thousands. One must take these varying scales into account when making conclusions about the magnitude of the effects of each predictor.

There are a few limitations of my analysis that may be considered in future research. One such limitation is imputation: I did not use this tool in this project, however, further investigation may find that it allows for the use of predictors I eliminated due to excessive NA values. Another is the fact that I only considered linear regression; other methods of fit may be worth discussing, including possibly an exponential or polynomial regression, as these methods may be able to address the upward trending Q-Q plot line. In addition, in spite of what was suggested to me, I did not include interaction variables in my regression, as none that I experimented with showed significance. Further investigation into this topic may show interesting correlations. Moreover, one could attempt to fit the model to countries which did not have a cost of living index present in the Numbeo dataset, in order to predict their cost of living. These factors, while not present in my analysis, may result in interesting and useful findings through further exploration of this topic.

References

Numbeo

- Data source:

  https://www.numbeo.com/cost-of-living/rankings_by_country.jsp?title=2020

- Methodology: https://www.numbeo.com/common/motivation_and_methodology.jsp

World Bank

- https://databank.worldbank.org/source/world-development-indicators