Capital and Flights (Group F)

A Final Report Prepared For S225

Jocelyn, Daniel

4/28/2018

Abstract

Travel, particularly via plane, is dependent upon the capital available to an individual. Using the `Flight Route` dataset from Kaggle [1], we seek to investigate whether or not there is a relationship between a country's change in wealth (using average GDP per capita (PPP) growth from the period 2010-2016) and the number of flights originating in a given country. We also explore how classification of a country as developed, developing, or recently developed affects this relationship. We use parametric and non-parametric regression methods and find that that average GDP growth is a significant predictor ($p = 0.00871$ and 0.006578 respectively) of the log number of originating flights ($n > 30000$), but it is not very explanatory $R^2 < 0.05$ in either case. We also run both parametric ANOVA and the Kruskal-Wallis rank-sum to test for the significance of development status and find similarly significant p-values in both cases $1.551e-12$ and $3.686e-12$ respectively. Finally, we run parametric and non-parametric proportion tests to investigate whether developing countries fly to the most popular developed destinations: Antigua and Barbuda, Iceland, Saint Kitts and Nevis, Bermuda, and Austria. Proportion test results suggest that developing countries don't fly to Austria and Iceland at the same levels, whereas binomial test results suggest that developing nations don't fly to Barbuda, Bermuda, and St. Kitts and Nevis at the same levels. We conclude that both average 2010-2016 per capita GDP (PPP) growth and development status are correlated with the number of flights originating in a country, and that developing nations generally do not fly the same destinations as developed nations.

Introduction

The focus of this project will be on exploring differences in travel, via plane, between countries stratified by development status. One measure to distinguish between developed and developing nations is gross domestic product (GDP) on a per capita basis, i.e. adjusted by the country's population. Additionally, GDP

per capita is often adjusted for the different living costs between countries by applying a purchasing power parity adjustment (PPP). In this project we use GDP per capita (PPP) figures made available from the World Bank. [2]

Our dataset originally consisted of 59,036 flight routes. After wrangling, which involved data cleaning, resolving airport and country codes with country names, and adding GDP and population data, we ended up with a dataset of 33,559 fully-matched flights.

A GDP per capita (PPP) of greater than $12,000 is often considered the unofficial benchmark of a developed nation. [3] As 2010 was the year for which we had the most GDP data, we classified our countries as developing if their GDP per capita (PPP) was greater than $10490.48 (inflation-adjusted from today) in 2010. Countries with GDP figures less than this cutoff were classified as developing countries, and countries that were developing in 2010 but developed in 2016 — the last year for which we have GDP figures — were classified as recently developed.

In this project, we hope to answer the following questions:

1. Does GDP growth positively correlated with flights flown from a country?
2. Have countries that have developed in recent years started to travel more, i.e. is development status a significant explanatory variable?
3. Do developing countries fly to the same destinations as developed

Given that airplane tickets are rarely cheap and often unaffordable for many people, it appears safe to assume that developed countries should have a greater demand for airplane travel than developing countries. By answering our questions we may be able to understand how different travel is for people in developing and developed nations. It may also provide the groundwork for additional questions related to travel.

Initial Look

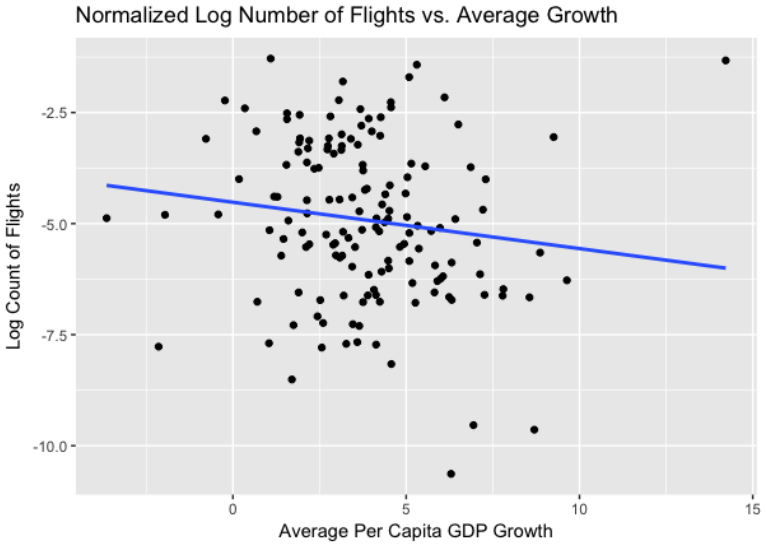
An initial look at the dataset shows that we get different results for source countries with the most flights if we don't adjust for their population:

	Source Country	Development Status	Number of Flights
1	China	0	4961
2	Germany	1	2166
3	India	0	1834
4	Germany	1	1830
5	Spain	1	1701
6	France	1	1511

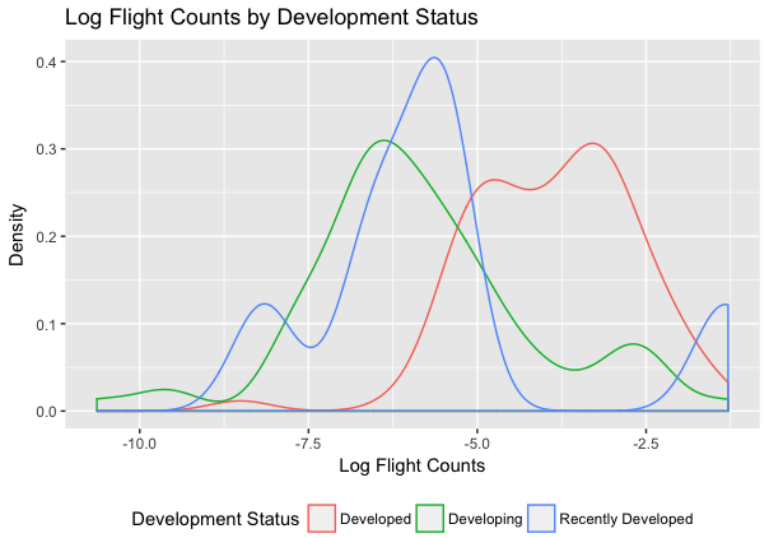
	Source Country	Development Status	Population	Normalized Count
1	Vanuatu	0	28212	0.28
2	Nauru	0	1131	0.27
3	Malta	1	43209	0.24
4	Palau	1	2196	0.18
5	Antigua and Barbuda	1	10305	0.16
6	Seychelles	1	9523	0.12

To adjust for this discrepancy, we choose to use normalized counts in our analysis.

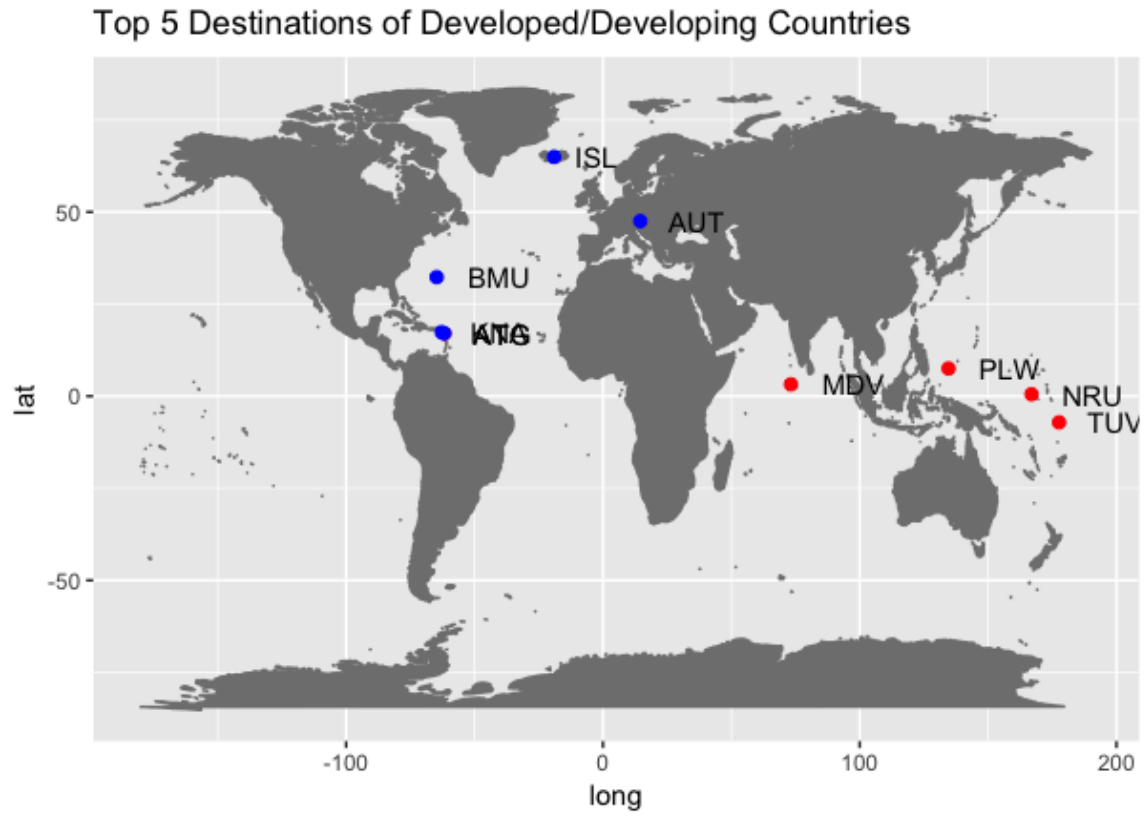
Using the entire dataset, if we look the fit of linear model of the normalized log count of flights originating in a country regressed on average GDP growth, we promising results that motivate further exploration:



If we also look at how our categorical variable, `development_status` affects the normalized (log) number of flights what we find is encouraging. The different slopes are encouraging and suggest further exploration is worthwhile, although it is worth noting that the sample sizes are unequal, i.e. $n(\text{recently developed}) = 7$:



Finally, simple graphical comparisons of the top 5 destinations (normalized) of developed and developing nations are shown below. It's interesting to note how popular developed nation (blue) destinations are clustered in the Carribean, i.e. between North America and Europe, while popular developing nation (red) destinations are clustered in the Pacific, i.e. between Asia and North America.



Results

Regression: Predicting Flights With GDP Growth

Creating `avg_growth`, the average GDP growth of a source country, we fit a parametric, simple linear regression model on the normalized `log_counts` as the response variable and `avg_growth.x` as the predictor variable. The summary output for the parametric model is below, the multiple r-squared for this model is 0.04588 while the adjusted r-squared is 0.03939:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.2952	0.2823	-15.21	5.4910e-32
avg_growth	-0.1723	0.0648	-2.66	8.7123e-03

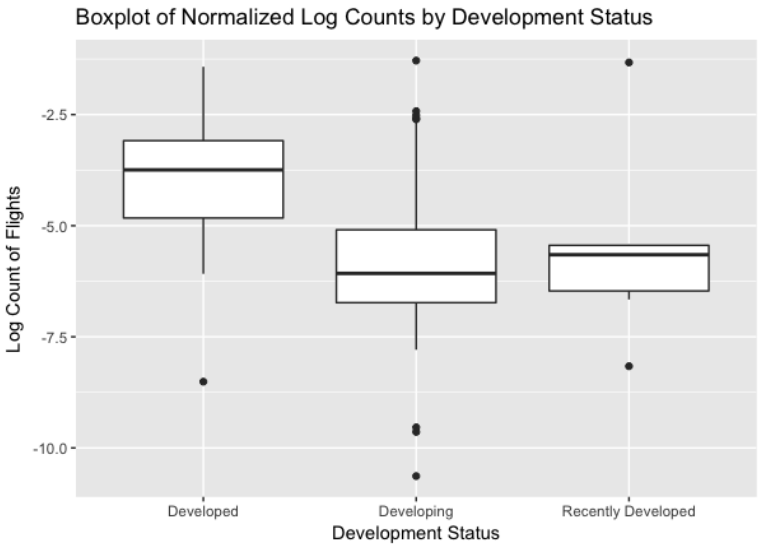
We then ran a non-parametric regression on the same formula, using `summary(mod,overall='drop')` we find that the multiple r-squared is 0.0485:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.1934	0.3124	-13.4246	2.5218e-27
avg_growth	-0.1782	0.06467	-2.757	6.578e-03

In both of these cases we find that `avg_growth` is a significant predictor ($p < 0.01$), but that the model does not explain much variation in `log_count`. That is to say that GDP growth between 2010-2016 does not sufficiently explain why some countries have more originating flights than others. We therefore move onto explore development status as a predictor. As it is a categorical variable, we use ANOVA procedures.

ANOVA: Investigating the Significance of Development Status

Before running ANOVA, we check boxplots:



The boxplot and residual plot (not shown here) suggest that the spreads of developed and developing countries appear to be very similar, but the spread of recently developed countries is slightly smaller. Comparing standard deviations, recently developed countries have the largest, 1.817, and developed countries have the smallest, 1.394. Since the ratio of the largest standard deviation to the smallest is less than 2 (approximately 1.303), the Equal Variance condition appears to be roughly satisfied. The QQ plot shows some divergence from the reference line near both ends, as a result any p-values derived from parametric ANOVA should be interpreted with caution.

Running parametric ANOVA gives us the following results:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
recent_development.x	2	147.34	73.67	32.73	1.551e-12
Residuals	151	339.84	2.25		

At all of the most commonly chosen significance levels, this is a significant p-value. We can reject the null hypothesis that there is no difference in the mean normalized `log_counts` between countries of different development status. We conclude that there is at least one difference in means between development status groups.

From the ANOVA by itself, we cannot determine which pairs have significantly different means. So we conducted Fisher's LSD to determine which pairs are significantly different:

	Developed	Developing
Developing	0.00	
Recently Developed	0.01	0.61

These results suggest that the mean normalized `log_counts` are significantly different between developed

and developing countries (p-value ≈ 0) as well as between developed and recently developed countries (p-value $= 0.01$), at a significance level of 0.05. At any significant α level, however, we cannot reject the null hypothesis that there is no difference in means for developing and recently developed countries.

We next conduct a Krusall-Wallis rank sum test, which provides an alternative to the one-way Parametric ANOVA. Similar to the ANOVA, it tests for differences in center. Unlike the ANOVA, however, which tests for differences in means, Kruskal-Wallis tests for difference in medians. The test gives us a Chi-squared test statistic of 52.653 with two degrees of freedom, and therefore a resulting p-value of $3.686e-12$.

Since the Kruskal-Wallis test assumes a shift model, we consder the `log_count` density plot stratified by development status shown earlier. The distribution of developed appears to be potentially bimodal, but roughly symmetric. The two peaks are close together and of similar height. The distribution of developing is also bimodal, but the second peak is significant smaller than the first. Additionally, it appears roughly symmetric. Finally, the distribution of recently developed is also bimodal - if we ignore that third hump at the far right, which corresponds to the Republic of Nauru (identified as an outlier in analysis). The second peak is smaller and located on the left side of the distribution. Although there is some deviation in terms of shape, the distributions appear similar enough to assume that the shift model assumption of the KW-test is met.

Therefore, we return to our p-value. At all of the most commonly chosen significance levels, this is a significant p-value. We can reject the null hypothesis that there is no difference in the median normalized `log_counts` between countries of different development status. We conclude that there is at least one difference in medians between development status groups.

Although we've concluded that a difference is present, we cannot determine which pairs have significantly different medians Below, we conduct a Dunn test to determine which pairs are significantly different:

	chi2	Z	P	P.adjusted	comparisons
1	52.65	7.13	0.00	0.00	Developed - Developing
2	52.65	2.76	0.00	0.00	Developed - Recently Developed
3	52.65	-0.22	0.41	0.41	Developing - Recently Developed

The results of the Dunn test suggest that the median normalized `log_counts` are significantly different between developed and developing countries (p-value ≈ 0) as well as between developed and recently developed countries (p-value $ = 0.0091$ & $ = 0.0029$), at a significance level of 0.01. At any significant α level, however, we cannot conclude that there is a difference in medians between developing and recently developed countries.

Our results indicate that recently developed countries do not differ from developing countries in terms of

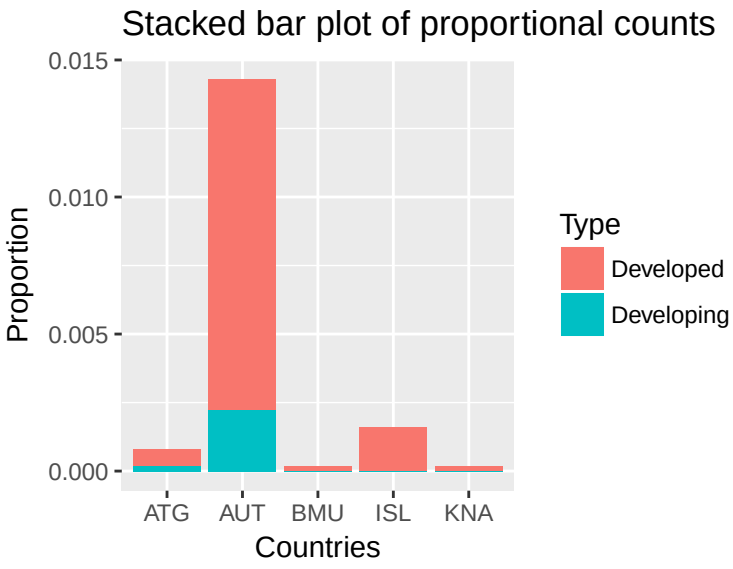
the normalized, natural log number of flights. Additionally, since recently developed countries differ from developed countries, we can conclude that travel has not increased in recently developed countries. In other words, recently developed countries are not distinguishable from developing countries in terms of the mean and median normalized, natural log counts of flights.

Proportion Tests: Exploring Whether Developing Fly to Popular Developed Destinations

After normalization, we find that the top 5 destinations of developed nations are: Antigua and Barbuda, Iceland, Saint Kitts and Nevis, Bermuda, and Austria. For developing nations, the top 5 destinations are: Palau, the Maldives, Nauru, Marshall Islands and Vanuatu. We use proportion tests to compare the proportions of flights to the `top5` developed nation destinations to total nation status flight counts, between developing and developed nations.

So our p is the proporton of developed nation flights are to Antigua and Barbuda, Iceland, Saint Kitts and Nevis, Bermuda, and Austria, while developing nation flight proportions to the same destinations are p_0. Thus, our hypotheses are: $H_0 : p_0 = p$, and $H_A : p_0 \neq p$.

Our question of interest here, therefore, is to test with statistical significance whether developing nations fly to popular developed nation destinations as frequently as developed nations. A stacked bar plot shows some initial issues, namely that for three of the destinations being considered, Bermuda, Iceland, and St. Kitts and Nevis (KNA) there were no flights originating in developing nations whatsoever.



The resulting p-values of the proportion tests are as follows:

We find significant results for Austria (AUT) and Iceland (ISL), p-values of $5.545961e-19$ and $7.775549e-05$,

	pvals	country
1	0.16	ATG
2	0.00	AUT
3	0.40	BMU
4	0.00	ISL
5	0.40	KNA

respectively. However, the other tests are insignificant at a 0.05 significance level. This is likely because conditions parametric proportion test conditions of at least 10 successes and 10 failures are not met for the other three countries. It is also questionable whether we can consider are trials independent. If we ignore these questionable conditions, we'd conclude that developing countries fly to Austria and Iceland at different levels as developed countries.

Because these conditions weren't met for the parametric proportion test, a non-parametric binomial test may prove more useful here. When we conduct this we find the following p-values:

	pvals	country
1	0.00	ATG
2	0.2747	AUT
3	0.00	BMU
4	0.2747	ISL
5	0.00	KNA

Our binomial test results differ from our proportion test results in that they find that the Austria test is now insignificant (p=0.2747). It also finds Iceland insignificant, but not the other 3 countries. So we might conclude that developing countries fly to Barbuda, Bermuda, and St. Kitts and Nevis in differentproportions as developed nations.

Conclusion

References

[1] Kaggle dataset, originally from openflights.org, available at: https://www.kaggle.com/open-flights/flight-route-database

[2] World Bank GDP per capita, (PPP) figures in dollars: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

[3] World Bank definition of high-income economies defines it as $12,236 for the 2018 fiscal year: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519#High_income