ESTIMATING ANNUAL TRAFFIC FOR AIRPORT SIZING

Applied Data Science Capstone

Karan Vaish | June, 2020

INTRODUCTION

Key Stakeholders

- City Planning Committee of *hypothetical country*
- Airport Authority of *hypothetical country*
- Potential Airlines that are interested in operating turboprop aircraft

Context

- A certain hypothetical country in South America is planning to open a new airport in one of its newly established metropolitan cities.
- However, a key component of such an endeavour is estimating the annual traffic – particularly the estimated number of annual flights to/from the city
- This is necessary to *plan* the capacity of the airport and the runway design. Furthermore, partnering airlines are interested to know if there is a business opportunity here, and has also indicated their interests in estimating yearly traffic

Business Problem

- The Airport Authority of the country reached out seeking help in estimating the yearly traffic that can be expected
- The city planning committee has indicated that the number and type of local businesses might be good indicators of flight traffic
- **It is my responsibility to**

  a)Verify if the local businesses are a good determinant of overall flight traffic, and if yes -

  b) build a model that correlates air travel with the number and type of local businesses in the region

DATA SOURCES, DESCRIPTION AND USAGE

#	DATA	SOURCE	DESCRIPTION	USE/FEATURES THAT CAN BE EXTRACTED
1	2019 Flight Schedule Data	Cirium.org	A complete list of all flights and routes as scheduled on daily basis in South America along with the operating airlines: 1. Date of scheduled flight 2. Origin Airport IATA code 3. Destination Airport IATA code 4. Operating flight name and code	Will be used to identify key routes and airports across the region, and avail yearly annual traffic for airports. For exploratory analysis – to identify key hubs in South America
2	Airport Info	Wikipedia	Lat and Lon information for airports against IATA codes	To match airport codes with City Names, airport names, and lat lon info for plotting
3	City Location info	Geocoder	Lat Lon info for each city corresponding to the airports.	Will be used to feed into foursquare API to extract information about businesses in the city
4	City Details	Foursquare API	Aggregated number of venues for each category in each city. A radius of 1 km from city centre was used. The categories are foursquare's primary categories. See slide 6 for examples	Will be used as the independent variable in modelling the air traffic vs city characteristics

METHODOLOGY – DATA WRANGLING

- The flight schedule data is downloaded from Cirium database as a tsv file. The raw data contains the following informations:
- The tsv file is read and stored into a pandas dataframe. Because of its size (2M rows), the DataFrame was stored as a table on Google BigQuery. Upload and download is much faster that way.
- From the flight schedule dataframe, origin and destination airport codes were concatenated to build ‘sectors’ and simple count() function was used to identify total number of flights on the route.
- Using pivot tables and groupby function each airlines market share on each route was identified. Marketshare was calculated by the ratio of the number of flights a particular airline operates on the given route vs total number of flights on that route
- All the airport codes were extracted into a separate dataframe and Wikipedia pages were for complete airport name and city name and latitude and longitude information for each airport.
- A separate table was used to store all the city names and geocoder was used to extract latitude and longitude information for all the cities

METHODOLOGY –EXPLORATION & VISUALIZATION

- Using Folium and matplotlib libraries I created this map to visualize the dominant routes in South America
- In the first layer, I used while loops to add lines (polygons) for each sector.
- The weight of the line was scaled depending on the number of flights on that route.
- I also added another set of polygon lines with low opacity and black colours to indicate that a certain route exists but the number of flights on that route are very very low
- On the second layer of the map, I added green markers to clearly highlight where each airport exists
- Added tool tips to show route, airport and airline information

METHODOLOGY – DATA MODELLING

- I used statsmodel API as it provides statistical outputs like shown on the right.
- OLS statistical method was used to run a simple multiple linear-regression model and find the fits.
- For the independent variables, I used foursquare API to get the total number of venues in a given city for each category. This category list was also extracted from foursquare's documentation.
- The dependant variables were total inbound and outbound traffic at the city's airport

OLS Regression Results

Dep. Variable:	Total Traffic	R-squared:	0.318
Model:	OLS	Adj. R-squared:	0.297
Method:	Least Squares	F-statistic:	14.67
Date:	Sat, 20 Jun 2020	Prob (F-statistic):	2.16e-21
Time:	16:27:46	Log-Likelihood:	-3633.5
No. Observations:	325	AIC:	7289.
Df Residuals:	314	BIC:	7331.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1692.8091	1826.719	0.927	0.355	-1901.348	5286.967
Arts & Entertainment	-20.1033	221.130	-0.091	0.928	-455.188	414.981
College & University	110.5195	149.640	0.739	0.461	-183.904	404.943
Event	4309.1871	461.325	9.341	0.000	3401.508	5216.866
Food	-198.9760	202.728	-0.981	0.327	-597.853	199.901
Nightlife Spot	-213.6644	190.901	-1.119	0.264	-589.272	161.943
Outdoors & Recreation	-363.3970	197.724	-1.838	0.067	-752.428	25.634
Professional & Other Places	440.6717	254.381	1.732	0.084	-59.835	941.178
Residence	338.4825	121.587	2.784	0.006	99.254	577.711
Shop & Service	-347.8972	230.289	-1.511	0.132	-801.001	105.207
Travel & Transport	571.3730	132.147	4.324	0.000	311.367	831.379

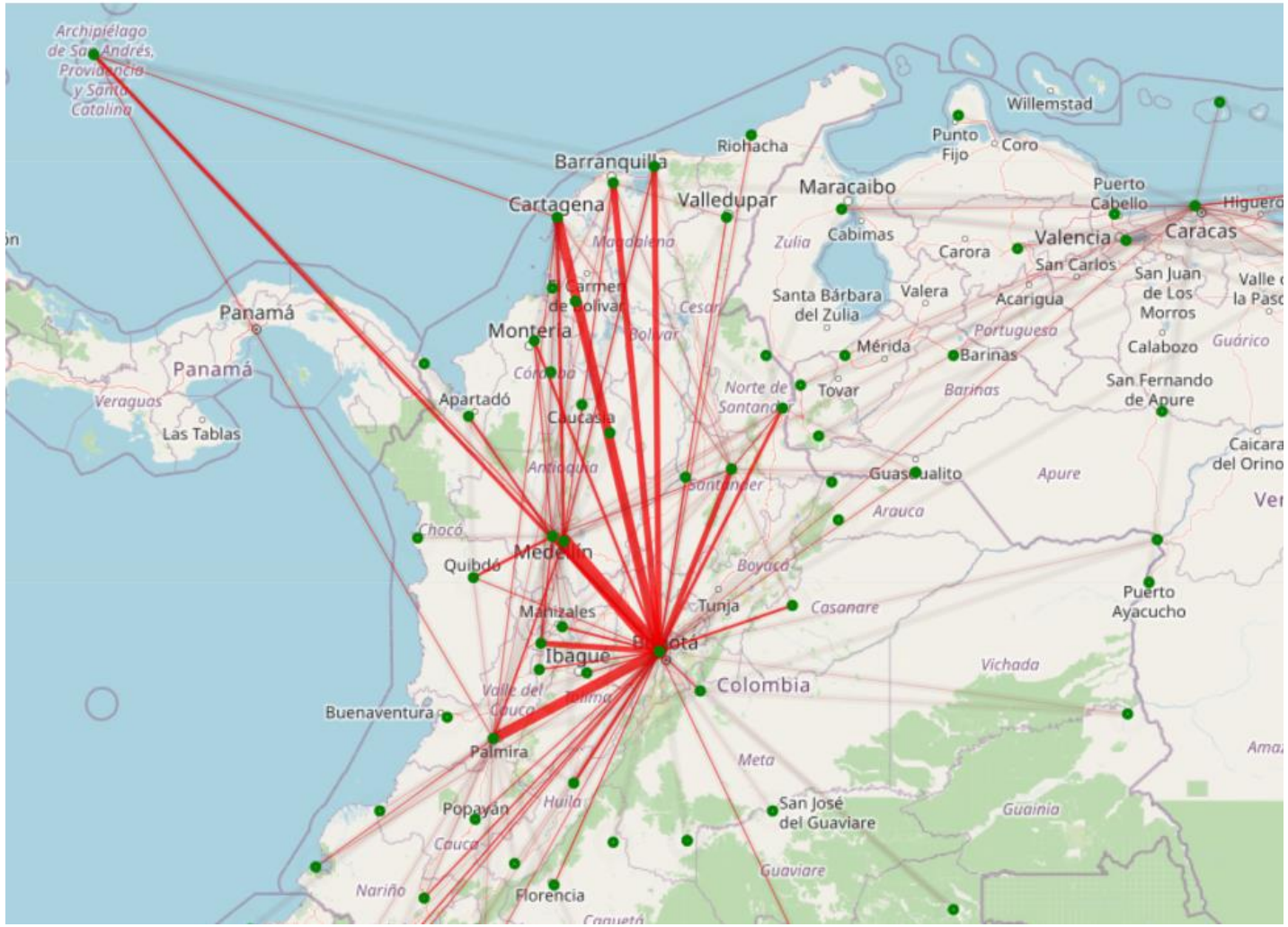
Omnibus:	436.629	Durbin-Watson:	2.056
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49704.791
Skew:	6.361	Prob(JB):	0.00
Kurtosis:	62.234	Cond. No.	177.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

RESULTS AND DISCUSSION

General Visualization Inferences

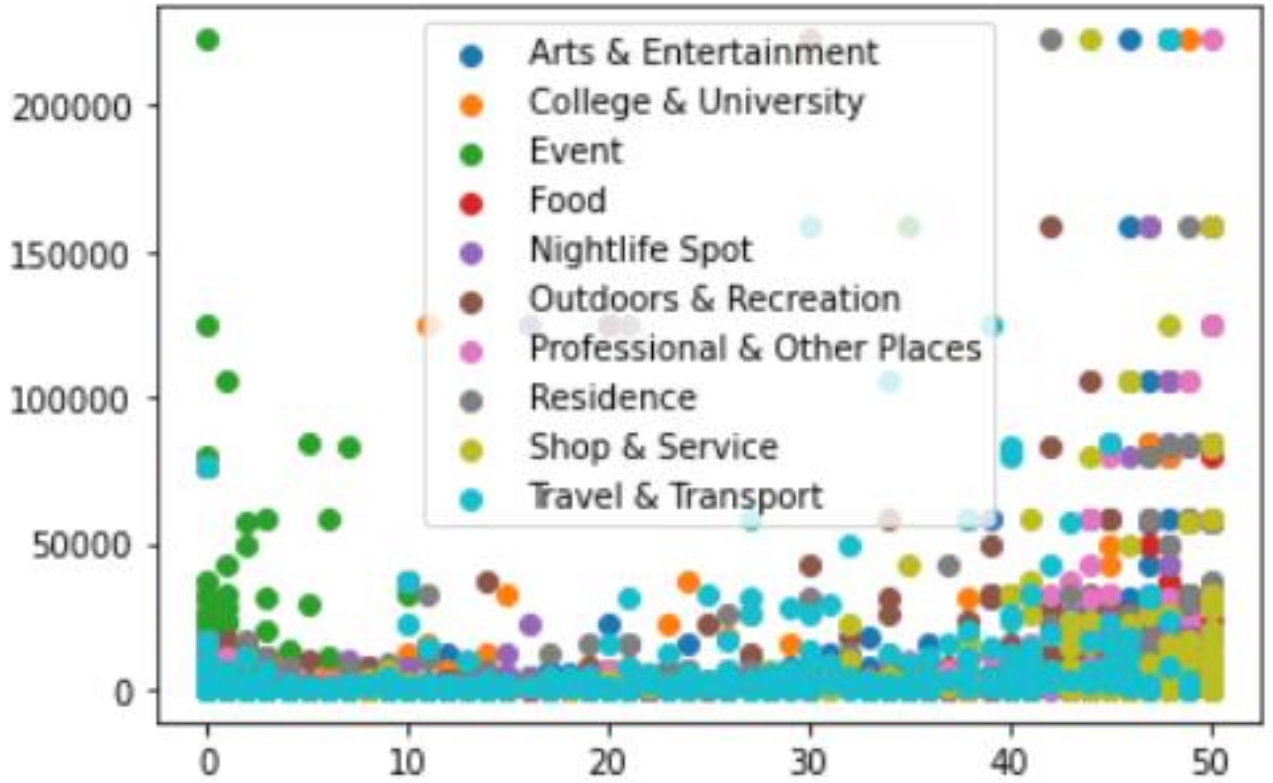
- Most of the dominant routes connect the biggest three metropolitans in each country. And there are several metropolitans in each country
- However each country usually has one single hub in the biggest metropolitan city.
- The hub city connects to several other metropolitan cities however those other cities do not connect with other metropolitans. For example in the image shown – Bogota is the central hub in Colombia. Bogota has flights to Medellin, Palmira, Cartagena etc. However, Medellin's only dominant route is to Bogota.
- This implies that city's size is not a very good indicator of the total number of flight traffic. It is rather a question of whether it is a designated 'hub' or not. E.g. Medellin vs Bogota.
- This might counter City Planning Committee's advise to leverage city size for estimating Air-traffic



RESULTS AND DISCUSSION

General Visualization Inferences

- Another quick visualization was to build quick scatter plots to see if there is any direct correlation that exists between any of the categories and the total number of flights
- From this graph we find that there is very little correlation between any of these categories and the overall traffic
- Further analysis and modelling is performed to verify this



RESULTS AND DISCUSSION

OLS Modelling -

R-squared

- An R-Squared of 0.318 was found. Which means that only 31.8% of the variation in flight traffic can be explained by the businesses in the city.
- Consequently there are other factors that need to be accounted in the model to explain the remaining 70% of behaviour.
- This can also be corroborated by our inference from the exploratory visualization step. In fact, what will determine the yearly traffic is if airlines designate it to be a ‘hub’ or not.

Prob (F-Static)

- A Prob (F-Statistic) of 2.16 e-21 indicates that although the city ‘size’ explains only 31.8% of the variation in air traffic, however the probability of the null hypothesis being true is very minimal.

R-squared:	0.318
Adj. R-squared:	0.297
F-statistic:	14.67
Prob (F-statistic):	2.16e-21
Log-Likelihood:	-3633.5
AIC:	7289.
BIC:	7331.

RESULTS AND DISCUSSION

OLS Modelling -

Coefficients

- As evident, the strongest and largely influencing category is ‘Event’ which includes all entertainment venues ranging from sporting events to performing arts etc. This draws the largest amount of air traffic into the city
- The second most important ones are professional places and residences which are reflective of the economic strength of the city. Needless to say these would be the key driver.
- Counterintuitively Nightlight Spots has a negative influence on the total air traffic in the city. It can be argued that a poor nightlife despite the existence of prominent businesses and other places would indicate that travellers prefer to ‘fly back’ at the end of the work day. This would inadvertently increase air traffic.
- College and University also affect air traffic but not exceedingly. This is an intuitive result as typically in academic institutions students fly in to stay for long periods of time.

	coef
const	1692.8091
Arts & Entertainment	-20.1033
College & University	110.5195
Event	4309.1871
Food	-198.9760
Nightlife Spot	-213.6644
Outdoors & Recreation	-363.3970
Professional & Other Places	440.6717
Residence	338.4825
Shop & Service	-347.8972
Travel & Transport	571.3730

RESULTS AND DISCUSSION

OLS Modelling -

pvalue

- A smaller pvalue implies a stronger accuracy of estimation between the category and the result
- From the output it can be observed that the top five categories have a significant impact on the behaviour of air-traffic.
- However, while trying to refine the model it was observed that Foursquare API allows a maximum of 50 results for each category. For large cities like LIMA, Salvador etc, the total number of venues for many categories might exceed 50. This could drastically, or even fatally, effect the accuracy of the result. To get uncapped results, we require the services of other places-API service providers e.g. google API etc. which are paid services and are out of scope for this assignment

Categories	pvalue
Event	1.5890490458671623e-18
Travel & Transport	1.8395127339423568e-05
Residence	0.005474767858685008
Outdoors & Recreation	0.04303743953484459
Professional & Other Places	0.08041437067492377
Shop & Service	0.13169196777241302
Nightlife Spot	0.20748975494611482
Food	0.3265754153075531
const	0.3518601695986314
College & University	0.4486045871862543

RESULTS AND DISCUSSION

Omnibus:	436.629	Durbin-Watson:	2.056
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49704.791
Skew:	6.361	Prob(JB):	0.00
Kurtosis:	62.234	Cond. No.	177.

OLS Modelling -

Overall result

- A low value of Prob(omnibus) indicates that the errors are not normally distributed. This means that estimating the right value along with its errors would be very difficult. In fact, the skewness is also very high. Ideally skew should be zero.
- The durbin-Watson is a test for homoscedasticity or the propagation of error. We hope to have a value between 1 and 2. In this case, the data is close, but within limits
- The Cond. No. measures the sensitivity of a function's output as compared to its input. When we have multicollinearity, we can expect much higher fluctuations to small changes in the data, hence, we hope to see a relatively small number, something below 30. In this case we see significant multicollinearity between few categories, which is quite reasonable.

CONCLUSIONS

- Overall the data is not a good model for predicting airline traffic.
- The business venue information explain only to a limited extent the expected annual air-traffic for a given city.
- From visual exploratory analysis it was verified that the larger cities tend to have greater air-traffic.
- However, whether the airport is a designated hub for an airline or not *significantly* impacts the air-traffic. There are many more economic considerations behind the utilization of an airport as a hub.
- This reflects in the regression analysis performed using statsmodel API. However, the accuracy *or the inaccuracy* of the model remains in question because of the limitations posed by the Foursquare API.

RECOMMENDATIONS

- For a more accurate modelling of the relationship between air traffic and characteristics of the city, it will be required to get a sense whether the airport in discussion will be utilized as a hub for an airline or not
- Hubs are purely discretionary, and each airline would have a separate set of criteria to posit the suitability of the given airport as a hub
- However, if the city wants to profit from the commercial aviation the city planning committee should make space for building greater ‘Event’ and ‘Professional Services’ venues within the city.
- Before this approach is completely disqualified, other places API need to be referenced which do not cap the overall results to 50 venues

*****