**Argentinian Bus Market Report**

Baptiste Cumin

**Summary**

In this project I studied the Argentinian bus market using data on website visits, bus routes and city characteristics. Based on this analysis, I recommend to:

1. Integrate a recommendation engine to increase retention, as website visits to leave from medium-sized cities drops off dramatically.
2. Propose combinations of buses to replace unavailable routes, to capture the 40-70% of searches (by segment) that are for non-existing routes.
3. Target marketing by geographic cluster, as they show different value potential and have different needs from the product.

Methodology

This project takes a customer-centric approach, enriching data on website visitors with information on the routes (number of operators, number of departures), and the cities (distance between cities, population) to understand the unique characteristics of these customers.

I created 4 customer groups based on the geographic location of web visits: Argentina (25% of website visitors), Brazil (34%), other South/Central American countries (SCA, 13%), and the rest of the world (28%). I chose these clusters for 3 reasons: I predicted they would generate different user behavior, they would allow us to target all users as they search our website, and they generated balanced classes. I then looked at the top departure and arrival locations they were searching for, digging into the distance between these places, whether they were travelling outside of the country and how many operators were on these routes. We faced numerous data challenges – see the jupyter notebook for full implementation details.

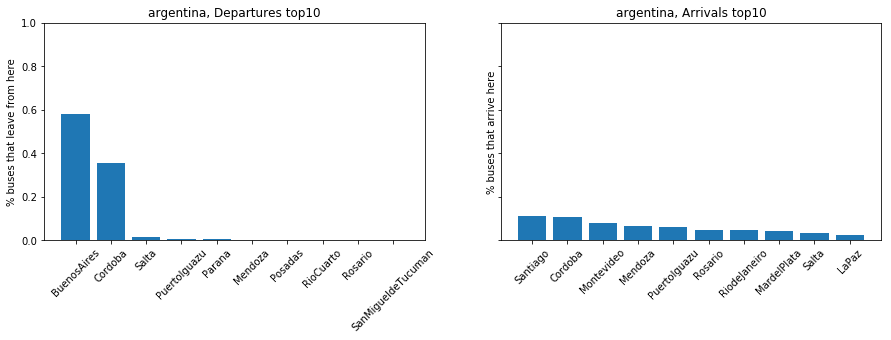
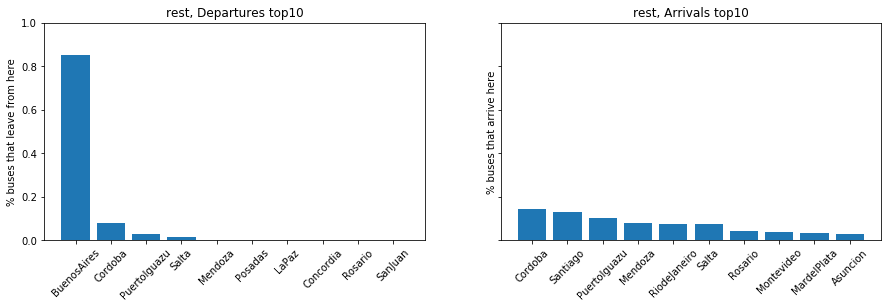
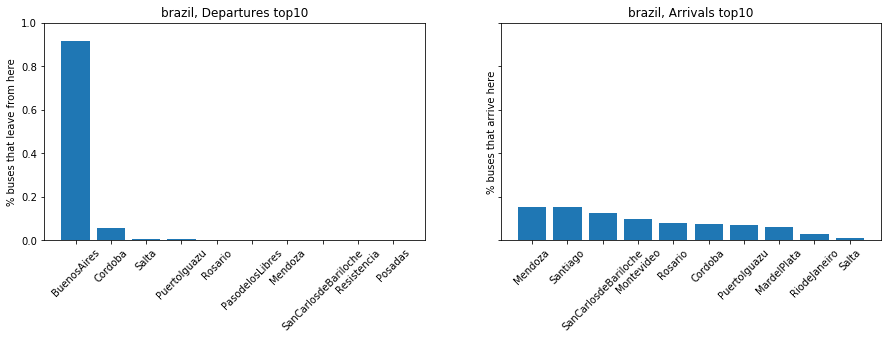
My data visualization focuses on bar graphs, since I had little reliable continuous or time-dependent data[[1]](#footnote-1). This was a day of work to effectively simulate my work in an internship, but I would expand this project by building a presentation with greater story-telling and visualizations.

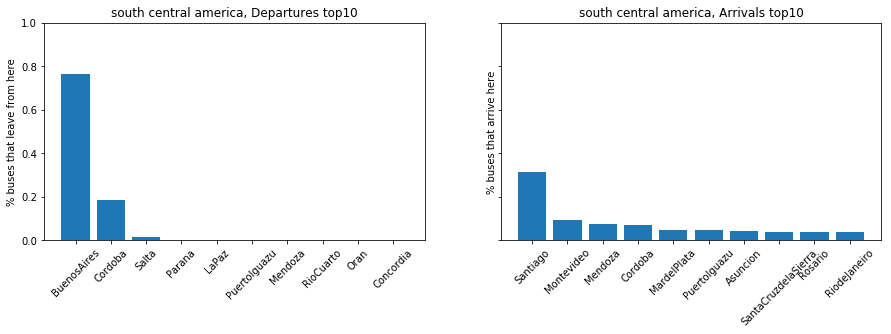
Findings

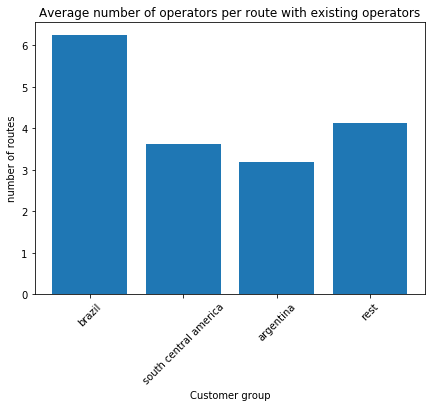
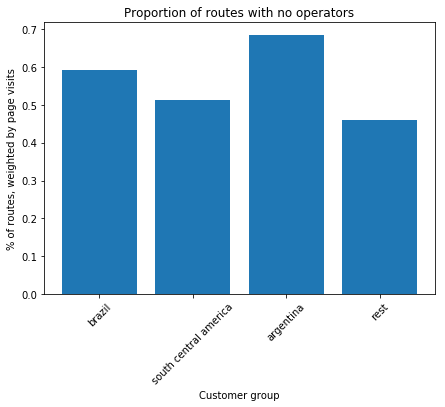
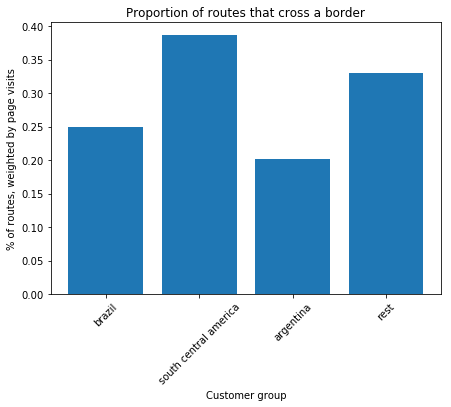
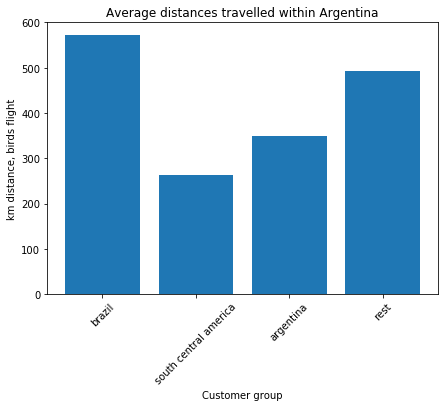
Analysis of arrival and destination searches [appendix 1] shows customers often use our service when they arrive in Argentina, or as they plan the beginning of a trip to Argentina. Most searches are for routes leaving from Buenos Aires or Cordoba (both have large international airports). There seems to be lost user retention: many smaller cities have far more arrivals than departures[[2]](#footnote-2) [appendix 3]. I built a simple route recommendation engine [appendix 4] to incite follow up purchases from these locations.The function takes as input as user’s country, and their arrival city. It returns the top 3 destinations that customers of this geographic segment go on to from this location. It pools all clusters if not enough observations are available. These recommendations could be integrated at checkout, as a side recommendation during the search, or with follow up emails, and we could decide between the first two using A/B testing. The clustering criteria could be made more specific with more user data, and a more complex model including common routes from nearby cities could be developed as an extension.

Clustering analysis delivers actionable insights on valuable product improvements. Argentinian customers are seeking shorter, less popular routes, where less operators are available [appendix 2]. 65% of the searches are non-available routes: capturing even a fraction of this market could have significant business value. The current features to deal with this [appendix 5] can be improved: a product feature could be added to propose a route in 2 bus rides instead of 1: to go from Asuncion to Buenos Aires, and Buenos Aires to Asuncion. This feature would be particularly effective for customers already in Argentina, who may have less flexibility on their destination and could accept this inconvenience.

The customer analysis also shows interesting insights which should allow us to improve follow up marketing to each cluster. Follow-up marketing for the SCA cluster should focus on cross-border destinations: these are generally long-distance, high-value bus tickets with a large profit potential, and this segment searches for them at a disproportionate rate of 38% [appendix 2]. Marketing effort should also be increased for Brazilian customers: they travel long distances within Argentina, on large, established routes with an average of over 6 operators [appendix 2]. This makes our value proposition attractive to this segment. Due to their close geography, they may have higher closing ratios than other segments. These customers could be valuable, and worthy of extensive follow-up marketing.

**Appendix 1**: searched arrivals/destinations, by website visitor segment. Graphs are on the same axis for simple comparison. Note the differences in arrival sparsity, and in departure concentration from BA.



******Appendix 2**: further information on website visitor geographic segment. The first 2 graphs are different metrics for the distribution of route operators.

**Appendix 3**: top 10 departures/arrivals ratios, for cities with over 500 page visits for arrivals (this cuts off BA). Measured on website visits.

|  |  |  |  |
| --- | --- | --- | --- |
| **city** | **searches for departure** | **searches for arrival** | **da ratio** |
| SantaFe | 6 | 703 | 0,009 |
| PuertoMadryn | 6 | 603 | 0,010 |
| MardelPlata | 38 | 2625 | 0,014 |
| SanCarlosdeBariloche | 54 | 3045 | 0,018 |
| Mendoza | 144 | 5551 | 0,026 |
| Rosario | 95 | 3112 | 0,031 |
| PuertoIguazu | 621 | 4074 | 0,152 |
| Posadas | 105 | 507 | 0,207 |
| Salta | 729 | 1965 | 0,371 |
| Cordoba | 8588 | 5592 | 1,536 |

**Appendix 4**: Python code for a simple route recommendation engine, taken from jupyter notebook. The function takes as input a pandas dataframe on website searches (pre-processed such that visitors\_df.city1 is the departure city, and visitors\_df.city2 is the arrival city), the customer’s current country, and the customer’s searched destination. It returns a recommendation of 3 cities to visit next.

#define clusters

cluster1=['Brazil']

cluster2=['Argentina']

cluster3=['Venezuela', 'Peru', 'Chile', 'Columbia','Cuba', 'Costa Rica', 'Ecuador','Mexico',

'Bolivia', 'Uruguay', 'Paraguay', 'Panama', 'Dominican Republic']

cluster4=[c for c in visitors.user\_country.unique() if c not in (cluster3+cluster2+cluster1)]

def recommend\_destination(visitors\_df, destination, country):

cluster=[]

for c in [cluster1, cluster2, cluster3, cluster4]:

cluster=c if country in c else cluster

dep\_destination = visitors\_df.loc[(visitors\_df.city1 == destination) & (visitors\_df.user\_country.isin(cluster))]

#if not enough people in this cluster leaving from this destination, take all clusters

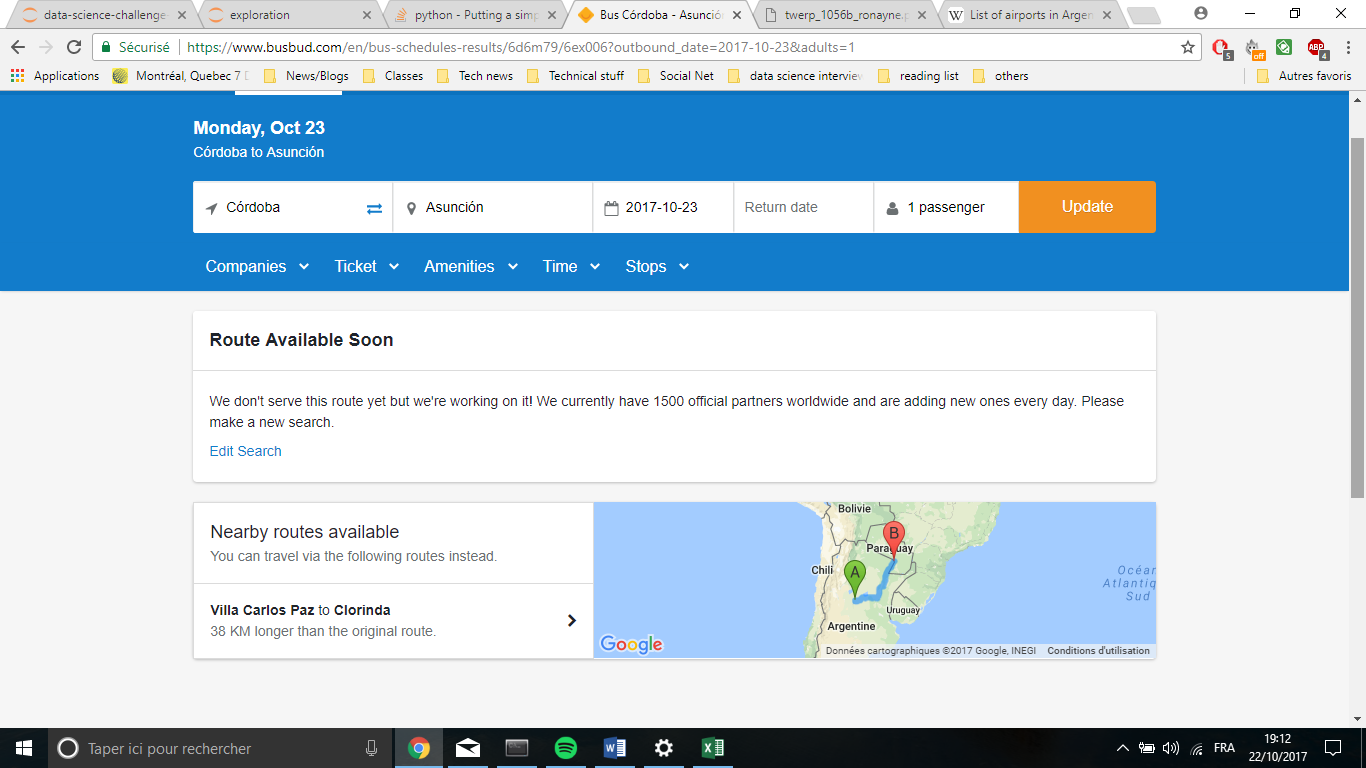
if(sum(dep\_destination.num\_visitors.values)<20):

dep\_destination = visitors\_df.loc[(visitors\_df.city1 == destination)]

dep\_destination\_no = dep\_destination.groupby('city2').agg('sum')[['num\_visitors']]

return dep\_destination\_no.sort\_values(by='num\_visitors', ascending=False) [:3]

**Appendix 5:** current feature to deal with unavailable routes. This could be improved with an additional screen below displaying compound routes.



1. The population metric in the cities data was very sparse – further work could be done to look into this. [↑](#footnote-ref-1)
2. Note this could be due to non-conversion of website visits. I would need further data to dig into this trend. [↑](#footnote-ref-2)