# Best city to study in

Coursera - Applied Data Science Capstone Project

Joseph M Camilleri – 6-Aug-2020

# Agenda

- Helping students decide where best to study
- Data Acquisition
- Data Cleansing & Preparation
- Exploratory Data Analysis
- Machine Learning: K-means clustering
- The results
- Conclusion and Future directions

# Helping students decide where best to study

- Many students from around the world study in a city other than their home town.
- Such students have a choice to make, as to where to relocate for one or more years. This may be far from an easy decision and the answer can have lifelong implications for the student's future.
- In 2017 there were 1.7 million students from abroad who were undertaking tertiary level studies across the EU-28. In 2017-18, 342,000 US students studied abroad for credits. Student mobility between 2011 and 2016 has seen an average year-on-year growth rate of 4%.
- The objective was to build a generic model which can be used by different students to help them make this critical decision in their lives.

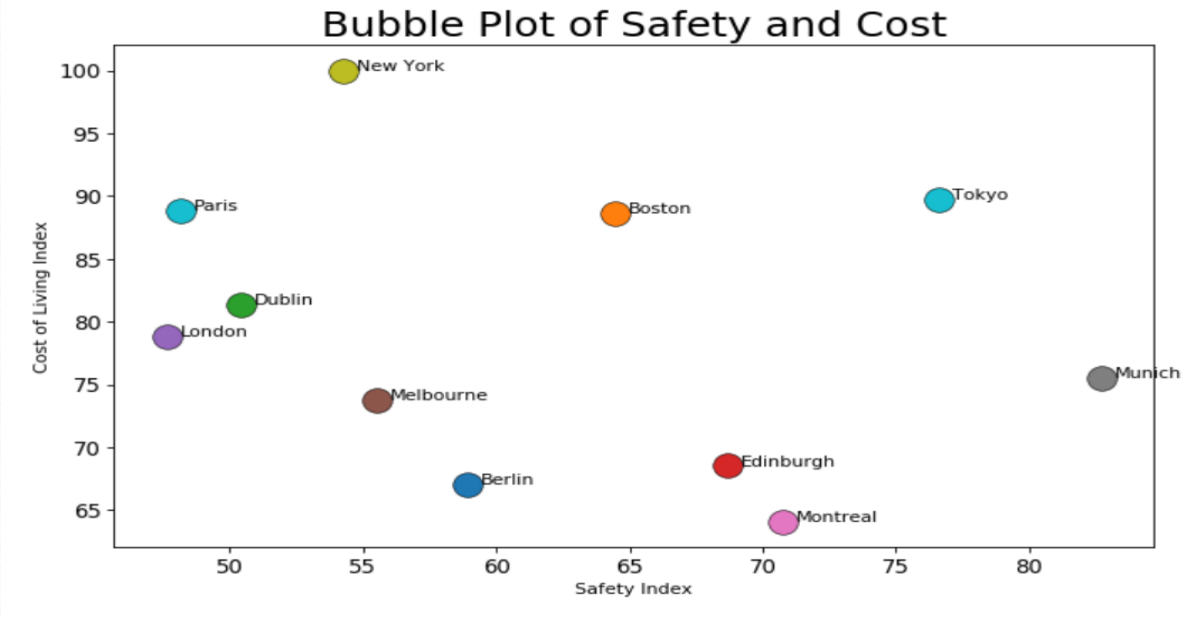
# Data Acquisition

- The student inputs the cities being considered as well as their personal interests directly into a Jupyter notebook cell. Naturally, these will be different for each student and the number of cities and interest will vary.
- City safety index and cost of living index scores are obtained by scrapping 2 separate pages from https://www.numbeo.com/.
- Geographic coordinates for each city are obtained from geolocator of Geopy python client for geocoding web services.
- Data about the venues at each city was obtained using the Foursquare API.

# Data Cleansing & Preparation

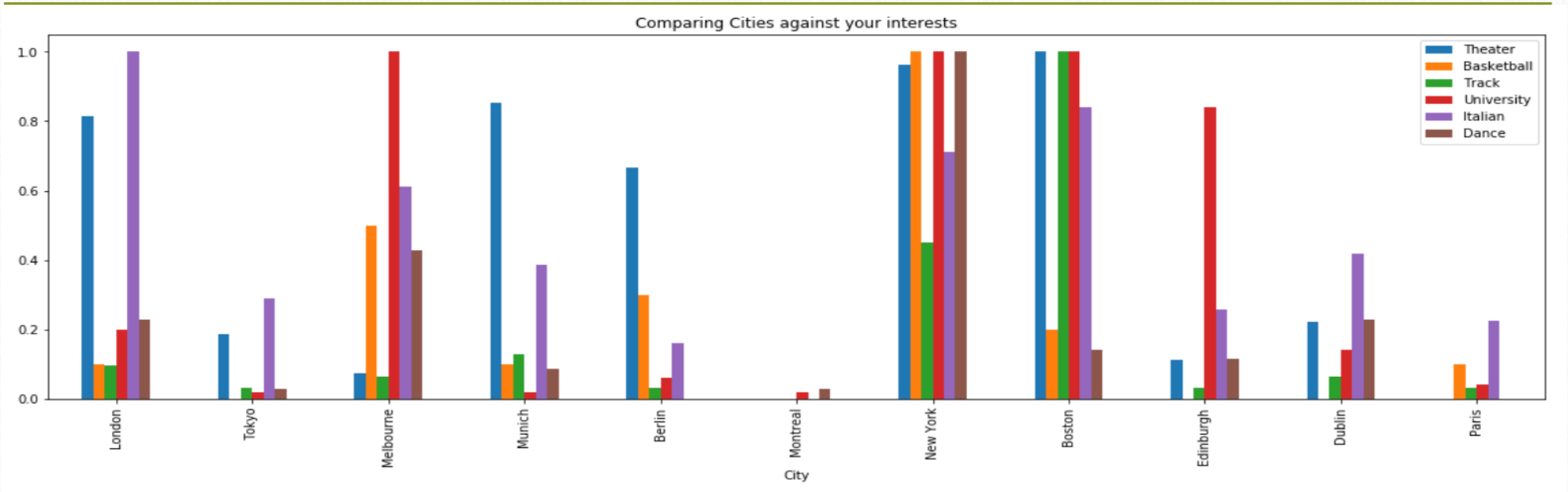
- City names from the web site needed cleaning to get them in a consistent format that matches the student's input;
- City data was augmented with safety and cost of living indices; latitude and longitude and also a column with the number of venues in each city corresponding to each of the student's interest;
- Venues data for each city from the Foursquare API explore function was also transformed into a one hot encoding dataframe, converted into a frequency dataframe for each city by venue category and sorted to extract the top 10 venues in each city of interest.

# Exploratory Data Analysis: Safety and cost of living



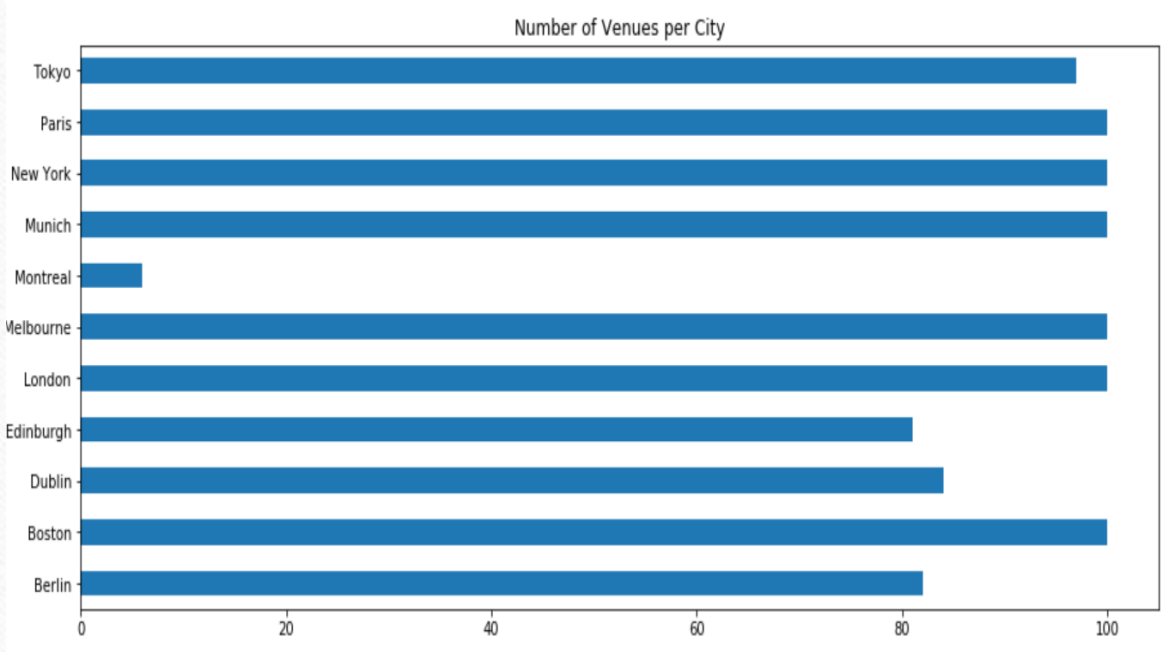

- From this graph it is very easy to see that Munich is the safest city while Montreal enjoys the lowest Cost of Living from amongst the selected cities.

# Exploratory Data Analysis: Matching cities to the student's interests



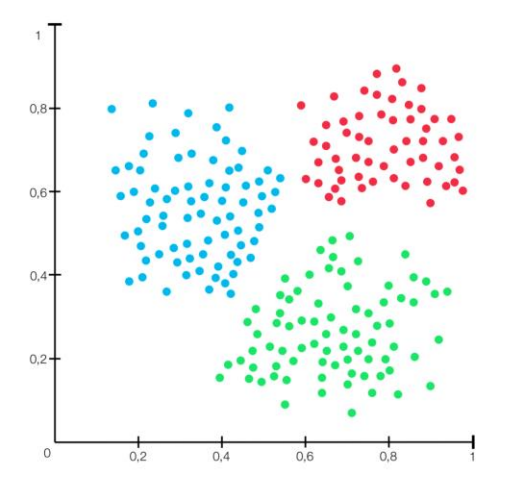

- Here one can observe that New York and Boston most closely match the interests of this student. The student may next consider the relative importance of each interest. For example, Boston offer more venues for "Track" activities, whilst New York offers more in terms of "Basketball" and "Dance" venues.

# Exploratory Data Analysis: Number of venues per city




- In order to explore and compare the venues in each city, I used the "explore" function of the Foursquare API, to obtain the venues near each city centre.
- The maximum number of venues returned was 100, which seems to be the limit for the free API licence.

# Machine Learning: K-means clustering




- The next objective was to compare the cities to determine which are more alike than others. This comparison was done using the K-means clustering technique based on the most frequent venue categories in each city.
- The venues data for each city were collated for K-means clustering using the Foursquare API "explore" function and the data prepared into a dataframe with the 10 most common venues for each city ordered by frequency.
- The number of clusters was designed to very depending on the number of cities entered, by dividing the number of cities by 3, while ensuring the number of clusters is not less than 2.

# Machine Learning: K-means clustering

**Examine Cluster 0 of similar cities by most common venues**

| | City | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | London | 0 | Theater | Hotel | Cocktail Bar | Plaza | Pub | Japanese Restaurant | Ice Cream Shop | Steakhouse | Monument / Landmark | Bakery |
| 2 | Melbourne | 0 | Coffee Shop | Café | Bar | Dessert Shop | Cocktail Bar | Shopping Mall | Clothing Store | Sushi Restaurant | Burger Joint | Juice Bar |
| 3 | Munich | 0 | Café | Bavarian Restaurant | Plaza | Clothing Store | Hotel | Boutique | German Restaurant | Coffee Shop | Italian Restaurant | Seafood Restaurant |
| 4 | Berlin | 0 | Hotel | German Restaurant | Coffee Shop | Italian Restaurant | Wine Bar | Café | Clothing Store | Boutique | Bookstore | Exhibit |
| 6 | New York | 0 | Coffee Shop | Hotel | Café | Sandwich Place | Clothing Store | Cocktail Bar | Park | Burger Joint | Salad Place | Pizza Place |
| 8 | Edinburgh | 0 | Hotel | Bar | Restaurant | Pub | Café | Indian Restaurant | Whisky Bar | Comedy Club | Coffee Shop | Cocktail Bar |
| 9 | Dublin | 0 | Pub | Coffee Shop | Hotel | Restaurant | Bed & Breakfast | Clothing Store | Café | Theater | Department Store | Italian Restaurant |
| 10 | Paris | 0 | French Restaurant | Ice Cream Shop | Plaza | Art Gallery | Gay Bar | Cocktail Bar | Wine Bar | Pub | Coffee Shop | Tea Room |

**Examine Cluster 1 of similar cities by most common venues**

| | City | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | Montreal | 1 | Bus Station | Convenience Store | Residential Building (Apartment / Condo) | Furniture / Home Store | Business Service | Yoga Studio | Event Space | Food Court | Food & Drink Shop | Flower Shop |

**Examine Cluster 2 of similar cities by most common venues**

| | City | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Tokyo | 2 | Historic Site | Café | Convenience Store | Italian Restaurant | French Restaurant | Japanese Restaurant | Park | Chinese Restaurant | Lounge | Sake Bar |
| 7 | Boston | 2 | Italian Restaurant | Historic Site | Coffee Shop | Seafood Restaurant | Bakery | Sandwich Place | Pub | Park | Hotel | Salad Place |

- The results of the K-means clustering where displayed by listing similar cities in each cluster iteratively, depending on the number of clusters as shown above.

# The results:
# safety and cost of living

- In order to discuss possible results I will assume that the student is considering the 10 cities below and also that the interests of the student are the 5 listed here:

**Cities being considered**

| | City | Country |
|---|---|---|
| 0 | London | United Kingdom |
| 1 | Tokyo | Japan |
| 2 | Melbourne | Australia |
| 3 | Munich | Germany |
| 4 | Berlin | Germany |
| 5 | Montreal | Canada |
| 6 | New York | United States |
| 7 | Boston | United States |
| 8 | Edinburgh | United Kingdom |
| 9 | Dublin | Ireland |
| 10 | Paris | France |

**Student's Interests**

| | interest |
|---|---|
| 0 | Theater |
| 1 | Basketball |
| 2 | Track |
| 3 | University |
| 4 | Italian |
| 5 | Dance |

1. The least costly city to live in is Montreal, which also enjoys a good level of safety;
2. the safest city to live in is Munich, with a cost of living below the average of the cities being considered;
3. New York is the most expensive and amongst the less safe cities; while
4. London is the least safe city enjoying average cost of living.

# The results: Matching cities to the student's interests

1. Montreal, and to a lesser degree Tokyo, are the least cities likely to meet the expectations of this student;
2. New York, and Boston as a close second, are the two cities most likely to satisfy the student's interests;
3. The student may next consider the relative importance of his/her interests. For example Boston offer more venues for "Track" activities, whilst New York offers more in terms of "Basketball" and "Dance".
4. Munich, which is the safest city, addresses all of the student's interest albeit to a lesser extent (compared to say Boston and New York) as there are less venues of each category of interest, however "theatre" is quite well represented in Munich.

# The results:
# Comparing cities on venues

1. Most of the cities being considered are similar with a couple of notable exceptions.
2. Montreal Canada is one such exception and stands out by itself, as having the least frequent venues dedicated to leisure time. While all the other cities have leisure venues as their most frequent venues. For Montreal it is only the 6th most frequent type of venue which is of a leisurely nature. Montreal seems to be more of a relatively quiet residential city when compared to the other 9 cities being considered.
3. Tokyo and Boston also standout as having a predominance of historic venues, so students of history may appreciate these two cities for this reason.

# Conclusion and Future directions

- Given the nature of the analysis, the results will be also extracted during the data exploration, and this also helps the student consider his/her options in terms of choosing a city to study in.
- The model developed here does not provide a simple definitive answer of one 'best city', but provides food for thought and leaves the final decision to the student.
- The model can be developed further by incorporating weights representing relative importance to the student of safety, cost of living and each of the interests in order to provide an answer to the question: "Which is the best city to study in?"
- The model can also be improved further by including other factors such as University rankings within each city and cost of tuition.
- Finally, one could develop this model into a fully operational commercial application with an HTML front end deployed to the web, where a student would enter the list of cities and his/her interests to obtain the results of the analysis with a single mouse click.

# Thank you