**REPORT**

**Introduction**

This project intends to analyse and cluster global financial cities based on their economic and social effectiveness.

**Problem**

ABC LLC is a professional asset manager investing in multitude of asset classes ranging from real estate, public equities, private equity and fixed income securities globally. The firm has recently raised an amount of USD500mn with the mandate to invest globally and is expected to allocate predetermined proportions of the portfolio to clusters of global financial cities that corresponds with their risk profiles. This identification of clusters of global financial cities would help determine the risk premiums to be taken into account when deciding upon an adequate expected return per each cluster. The determination of such expected returns are crucial in assessing the performance of ABC LLC as an asset manager and would eventually cascade to remain/exit decisions from certain geographies altogether.

**Data Sources**

1. Firstly, global financial cities are identified based on Global Financial Centres Index which is published Z/Yen Group (consultancy and venture firm based in London) and the China Development Instituted (based out of Shenzhen) on a semi-annual basis. The index is comprehensive in terms of considering by considering dozens of indices compiled by World Bank, OECD and the Economist Intelligence Unit. The data as of March 26, 2020 is sourced directly from: <https://en.wikipedia.org/wiki/Global_Financial_Centres_Index>
2. To further explore the surroundings of these cities at a later stage (stage 3 and stage 5), latitude and longitude data is sourced using the Nominatim function of the geopy.geocoders.
3. The country corresponding to each Centre is required in a later stage (stage 4 data matching) to allocate features that can only be matched based on the country in which the Centre is located in; for this Foursquare API is used based on the coordinate data (obtained in stage 2) for each Centre and the corresponding country could be identified.
4. A feature set that best enables to identify the quality of each Centre is identified after careful research and is as follows:
   * GDP per capita in US$ (PPP, constant basis) – The data is obtained from three sources where each source is matched sequentially till all values are obtained:
     1. <https://en.wikipedia.org/wiki/List_of_cities_by_GDP_(PPP)_per_capita> – Based on OECD statistics
     2. <http://www.worldcitiescultureforum.com/assets/city_data/Average_income_per_capita_per_year_%28ppp%29_5112018.csv>
     3. World Bank data by country
   * Financial inclusivity:
     1. Measured using ‘Account Ownership at a financial institution % of population’ – World Bank data
   * Quality of Governance and Ease of doing Business:
     1. Business extent of disclosure index – World Bank data
     2. Ease of doing business index – World Bank data
     3. Government effectiveness index – World Bank data
     4. Rule of law estimate – World Bank data
   * Infrastructure quality:
     1. Logistics performance index – World Bank data
   * Quality of financial markets:
     1. Market capitalization % of GDP – World Bank data
   * Social quality:
     1. Gender Inequality index – UNDP.org
5. Further to identify similarity in the most common venues in each city/Centre, venues within a Radius of 10km was obtained and value counts were filtered out. The venue data was obtained using Foursquare API.
6. Finally, all the above data is adequately preprocessed using standardization for numerical data and dummy variables for categorical data such as; most common venue categories (as identified in stage 5).

**Data Cleaning and Preprocessing**

The abovementioned data shall be combined to form a single dataframe before applying a machine learning algorithm to derive the analysis. Before doing so, following methods were used in cleansing the data.

**GDP per Capita by city (PPP, constant terms)**

The data for city-wise GDP per Capita (in USD PPP constant terms) was obtained from two primary sources (refer Data Sources) which provided fairly reasonable estimates of most commonly known financial cities around the world. However, city wise GDP breakdowns are scarce for lesser known cities in which case it is assumed that the GDP per capita (USD, PPP constant) for the corresponding country for the city under consideration is a relevant proxy.

Furthermore, in exhausting all the publicly available data considered to be reliable, data for several cities/centres was not available. However, upon deeper observation it was found that these cities/centres are part of countries which are unlikely to be investable at present given the seclusion of these countries either through global economic sanctions or under-development. Therefore, it was decided that these centres be removed from the analysis, although this amounted only to a few in the dataset.

Rest of the features identified in the Data Sources section is only compiled on a country basis, and was considered an adequate proxy given the mutual exclusivity of the centres on a country basis.

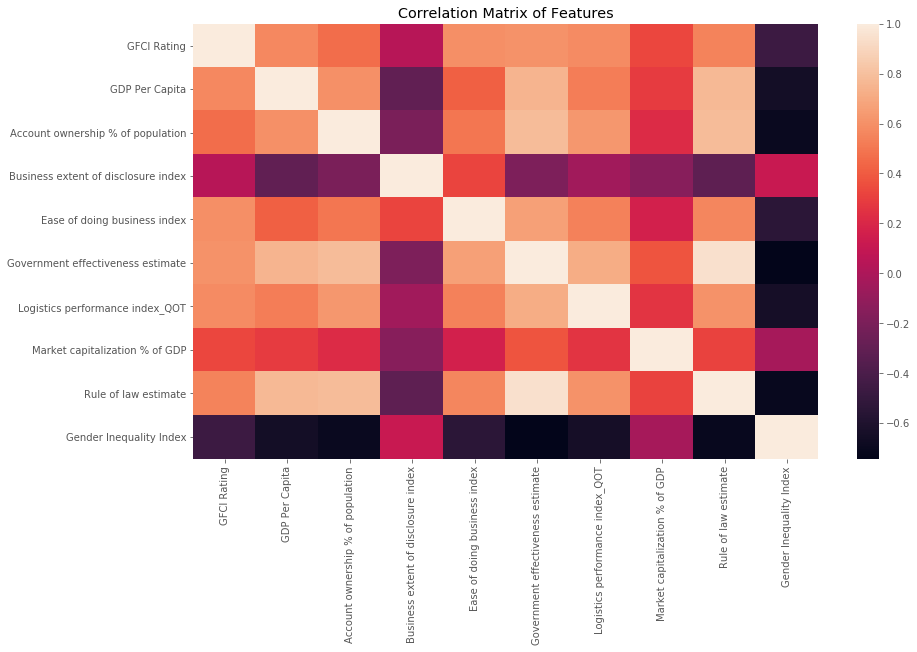
**Data Preprocessing**

All the numerical data was standardized using **sklearn’s StandardScaler** method and the categorical variables (10 most common venues per each city) was converted to dummy variables before feeding into the machine learning algorithm.

**Exploratory data Analysis and Feature selection**

Basic exploratory data analysis was carried out to identify whether meaningful relationships existed between the Global Financial Centres Index and that of the main features considered.

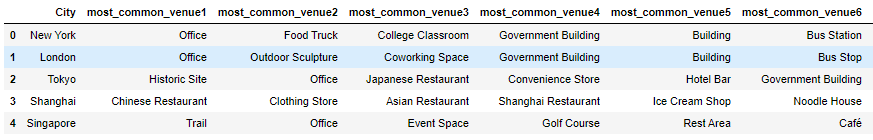
The following correlation matrix in the form of a heatmap was obtained using **Seaborn**:



As noted above most features had reasonably fair correlation with that of the Global Financial Centres Index. However, given the GFCI itself is of weighted nature and constituent of different factors, the correlation is far from perfect. Nevertheless, these features were not made redundant by the less than perfect correlation with that of GFCI index as these features are strongly weighed in deciding the quality of the financial markets in global context by most investors before consideration of investing.

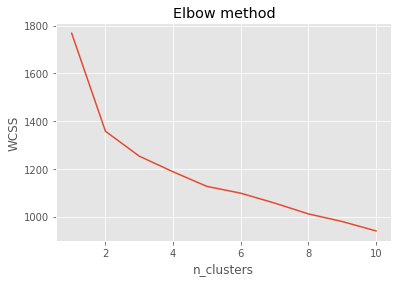
**Venue Data**

In addition to the above numerical features, it was considered rational to take into account the most common venues around the Centre of the city as a key categorical variable in the analysis. It is believed that significant similarities would exist amongst cities in terms of its venue layup often showcasing the level of sophistication in lifestyles around the financial cities. The following is an excerpt of the most common venues of some cities analysed in the dataset.



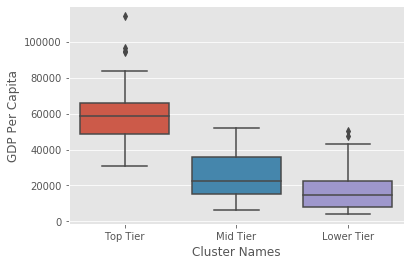
**Clustering model**

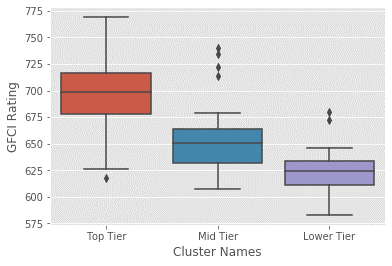
**KMeans** clustering algorithm was used to carry out the above analysis with the number of clusters to be used for the analysis determined through the Elbow method of analyzing the **‘within cluster sum of squares’** **(WCSS);**

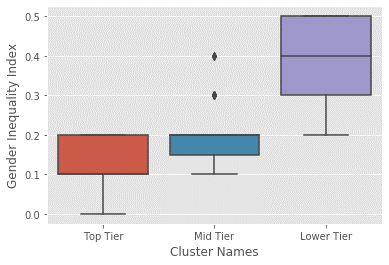


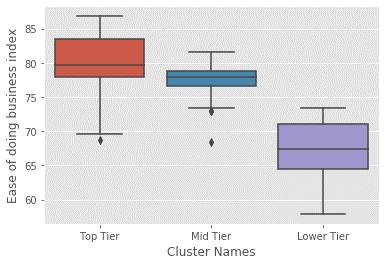
Based on the elbow method above, it was can be observed that the adequate amount of clusters to be used was in the range of 2-3 clusters. Therefore, the analysis was based on 3 clusters of which the basis of naming is as follows;

The three identified clusters were named as **Top Tier, Mid Tier and Lower Tier** representing the quality of the city’s financial standing in terms of economic prowess as well as social development.



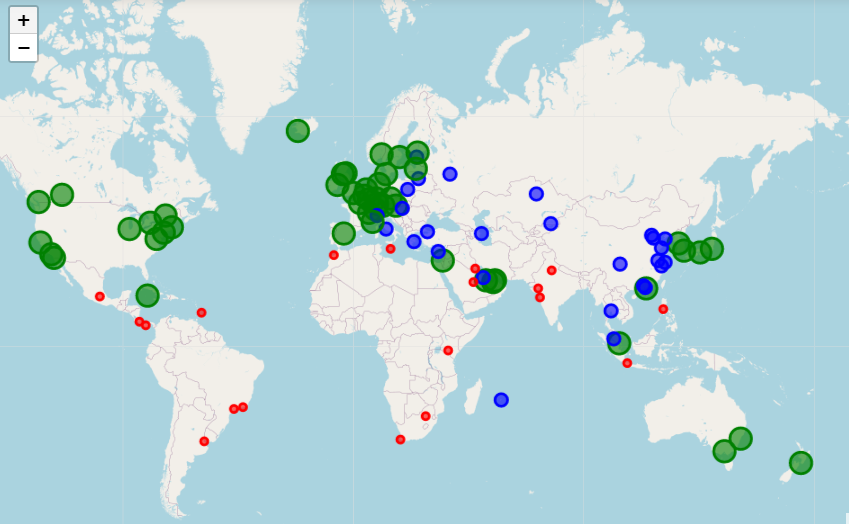






As noticed from the above boxplots against some of the features, the Top Tier cities tend to be the most successful in terms of economic power alongside business governance and social equity. The most notable difference in the boxplots between the Top Tier and Lower Tier cities is where the Top Tier seems to have significantly lower gender inequality indexes. Meanwhile, ease of doing business is rated much higher in Top Tier cities. Furthermore, GDP per capita is much higher in Top Tier cities. Whereas, the Mid-Tier cities falls somewhat in between the contrasting clusters of Top Tier and Lower Tier, although it can be noted that the Mid-Tier cities are much closer to Top Tier cities in terms of most features than to the Lower Tier.

**A Geographical representation of the clustered cities is as follows;**

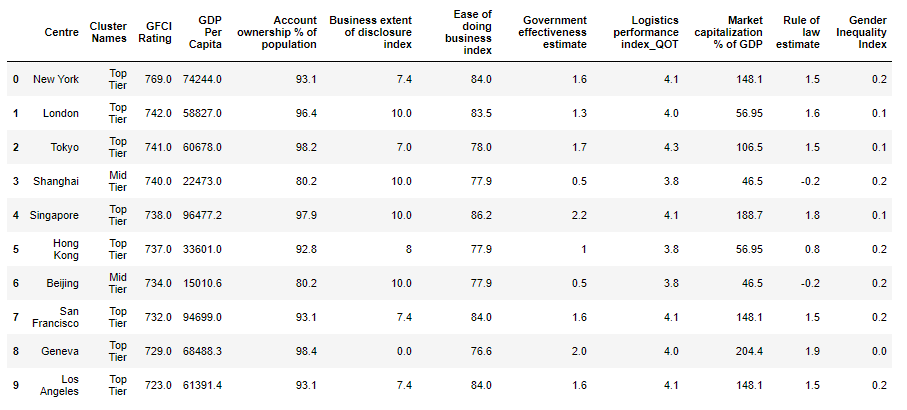


**Top Tier**

**Mid Tier**

**Lower Tier**

**A summary of the results are as follows;**



**Conclusion**

From the above analysis it can be observed that the Top tier cities showcases strong economic success, robust business governance standards and better infrastructure and social conditions which all in summary would intuitively translate to a less risky investment. Although, this analysis does not delve into determining the exact amount of risk premiums to be considered, it serves a crucial purpose of acting as a strong sanity check before estimating risk premiums. Following the analysis, the asset manager would be made aware that the risk premiums (to be used in estimated expected/required rate of returns) shall be the lowest for the Top Tier cities whereas relatively higher risk premiums should be used for the Lower Tier cities in representing the higher risk inherent in such investments. Furthermore, careful consideration needs to be taken in estimating the risk premiums for the Mid Tier cities which are likely to transform to Top Tier cities in the near future, therefore attributing significantly higher risk premiums relative to that of Top Tier cities would be an unfair and an inaccurate assessment.

**Future directions**

In the course of conducting the above analyses, it was noted that the city specific data which are publicly accessible was scarce although premium data was available from selective paid APIs. Therefore, the availability of such data would undoubtedly improve the accuracy of the above clustering model.