**Introduction and Business 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**

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 within a Radius of 10km, Foursquare API is used to get these data.
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).