Twins among the four Asian tigers?
A comparison of Hong Kong and Singapore from economic and location data.

Fabien Nugier
IBM Applied Data Science Capstone Project on Coursera

a. ***Abstract:*** Hong Kong and Singapore are two major financial hubs in the world economy. Not only these two "tigers", as they are often called, are similar geographically, they are also similar in terms of economic indicators like the growth domestic product and standards of living. We propose here to revisit some of their similarities and try to point out differences between the two economies using data science methods. We especially use location data in order to characterize and compare their different neighborhoods with machine learning techniques. In other words, we try to examine the two Asian tigers from data and see whether they are twin tigers or not.

b. ***Context of study:*** This project is written in the context of the final project of the IBM Data Science certification on Coursera. The data science presented in this paper is done with Jupyter notebook and is available on my Git Repository at the following address:

`https://github.com/FabienNugier/Coursera_Capstone`

c. ***Period:*** The project was started on Oct. 24, 2019 and submitted on November 1, 2019, spending about 50% of my work time on it.

November 1, 2019

Contents

I. INTRODUCTION: PRESENTATION OF THE TWO TIGERS

Hong Kong (HK) and Singapore (SG) are two of the four Asian tigers – along with Taiwan and South Korea – and they share multiple common features which make them suitable for a comparison without too much calibration, as we would need for example when comparing very large and small economies.

According to Wikipedia [1, 2], the populations of Hong Kong and Singapore in 2018 were respectively 7,451,000 and 5,638,676, hence Hong Kong's population being roughly 32% higher than Singapore. Still in 2018, the labour force in Hong Kong was 3,955,349 while being 3,377,908 in Singapore. This makes Hong Kong labor force just 17% higher than Singapore's, bringing the two economies to a relatively equal footing in terms of working populations.

In terms of geography, Singapore's area is 725.1 km^2 while Hong Kong reaches 1,108 km^2. Both territories host major world container ports, going hand in hand with their free trade policies, Singapore being ranked 2nd while Hong Kong ranking 7th. Economic sectors are slightly different between the two tigers, as summarized in the Table I taken from Wikipedia, Singapore being much more present in the industry sector than Hong Kong.

Asian Tiger	agriculture	industry	services
Hong Kong	0.1 %	7.6 %	92.3 %
Singapore	0.7 %	25.6 %	73.7 %

TABLE I: GDP by sector (2017 estimates) data.

In terms of Growth Domestic Product (GDP), which corresponds to the total monetary value of all the finished goods and services produced within a country border [3], the two tigers are doing very well considering the size of their population, and this is certainly in part to be attributed to their low taxation rates for companies, their important presence on international financial markets and their high economic freedom policies. In 2018, Singapore was ranked 34th in the world with a GDP of 364,157 million USD, and Hong Kong 35th with 362,993 million USD (US dollar). In 2017, the GDP growth of the two tigers was respectively 3.8% (HK) and 3.7% (SG). In 2018, the GDP growth was also quite similar between them with 3.0% (HK) and 3.1% (SG). Since the two economies have very close GDPs but slightly smaller population for Singapore, this makes Singapore's GDP per capita (i.e. GDP / population) higher. In 2018, Hong Kong's GDP per capita was 48,717 USD, ranking 14th in the world, while Singapore ranked 7th with 64,582 USD.

According to Table II, Hong Kong and Singapore were respectively ranked 2nd and 3rd in trade-to-GDP ratio among all the countries in the world (data from 2017). Although this indicator tends to

favor small economies (others to mention being Luxembourg and Ireland), it is still a clear indication that these two Asian tigers are centers of trade and major actors in the world economy. Considering this, we can easily wonder how people's life is affected by the economy on a daily basis. One very important indicator of the standard of living is the Consumer Price Index (CPI) which measures the average price of a basket of consumer goods and services [4].

Asian Tiger	Exports of goods and services (% of GDP)	Imports of goods and services (% of GDP)	Imports and Exports (% of GDP)
Hong Kong	188.0 %	187.1 %	375.1 %
Singapore	173.3 %	149.1 %	322.4 %

TABLE II: Trade-to-GDP ratio according to the world bank's 2017 data.

Last, but not least, we should remind ourselves that Hong Kong and Singapore are megacities hosting millions of people, and we can wonder how these cities (or group of connected cities) compare in terms of shops, restaurants, facilities, etc. According to an online article on *Culture Trip* [5], Hong Kong prevails in terms of attractions and shops, but both places are world-class top locations when it comes to restaurants and multicultural cuisines. Housing in Singapore is more affordable than Hong Kong. Both cities have very low crime rates and high cost of living, but Singapore does better in education and pollution rankings. As we can see, the two tigers have similarities and variations that need to be further explored and quantified.

We will thus consider here the GDP and CPI in the first place in order to understand Singapore and Hong Kong on a global scale. We will then focus more on location data in order to get a more detailed description of their economic activities. This work can benefit different kinds of stakeholders, for example a company or a bank hesitating between Singapore and Hong Kong for a forthcoming implantation. It can also be useful to a travel agency wishing to compare some cities in the world with these two locations, and better serve requirement of clients. We describe the data collected for this work in Sec. II, followed by the methodology applied to it in Sec. III. We then present the results of our analysis in Sec. III, discuss further directions that could be explored in Sec. V and conclude in Sec. VI.

II. DATA SOURCES

In this section we briefly explain how the data is collected and which operations are done on it to make it usable for data exploration and modeling.

A. Growth Domestic Product data

We obtain GDP data as well as other economic indicators from the UNdata website [6]. These economic indicators concern both Singapore and Hong Kong in the period 1981-2017, but are given in local currencies. In order to make them comparable, we need currency exchange rates with another currency, e.g. the USD. We use exchange rates data given by *Investing.com* which provides monthly

averaged values for HKD/USD and SGD/USD (see e.g. [7]). Since we have month averages, we take their average over each year to match with our year estimates of the GDPs. Cleaning the data, we get values for HKD/USD going from 1977 to 2017, while for SGD/USD the values are only available from 1981 to 2017. This data allows us to convert GDP values for each year into USD currency.

In addition to GDP values, the csv (comma-separated values) file offered by UNdata also contains expenditures of "final consumption", "household consumption" and "general government final consumption". These data columns are also converted in USD for comparison.

B. Consumer Price Index data

CPI data for Singapore is obtained from the *Department of Statistics Singapore* [8]. It provides yearly averaged prices for about 30 types of goods and also averaging a basket of goods according to their category of prices (all items lowest 20%, Middle 60%, Highest 20%). The data spans the years range 1993-2018.

CPI data for Hong Kong is obtained from the *Census and Statistics Department* [9] and contains about 25 prices by month and by year averages. We only exploit year averages of 5 indexes: "Food", "Housing", "Clothing and footwear", "Durable goods" and "Transport". After cleaning the data, the years range goes from 1982 to 2018.

C. Localisation and Venues data

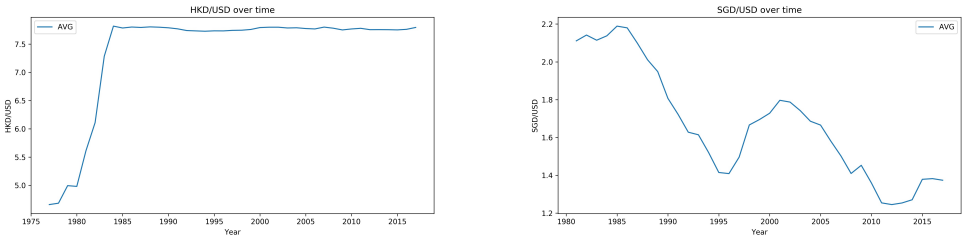
We obtain the list of several location names in each of the 28 districts of Singapore from *iProperty.com* [10] and use *Latitude.to* [11] to find their latitude and longitude that we add by hand to a csv file. When different names appear in *Latitude.to*'s search engine, we take the location that appear the most relevant (usually variations are very small). We also do an average of locations by district in order to estimate the center of each of the 28 districts.

We obtain the districts names of Hong Kong taking the names from *Wikipedia* [12] and using *Latitude.to* to build a csv file by hand. Additionally, we scrape a page on *Geodatos.net* [13] to get the cities names and locations in Hong Kong. Both sets of locations are different as cities in Hong Kong are located in the valleys while districts can encompass large areas of mountains.

Finally, all the data concerning venues in Singapore and Hong Kong, such as shops, restaurants, facilities, etc., are generated directly from commands inside the *Jupyter notebook* [14], using a *FOURSQUARE developers* account [15] to gather all the venues at multiple locations. This data accounts for a large portion of all the data used in this work.

III. METHODOLOGY

Different Python libraries are used in the `Jupyter` notebook in order to go through data collection, data understanding, visualization, modeling and evaluation. We use `NumPy` to deal with arrays, `Pandas` for dataframes, `Matplotlib` and `Seaborn` for visualization, `Scikit-learn` for machine learning, `Folium` for displaying interactive maps, `BeautifulSoup` for HTML scraping, and some other libraries are used punctually.



(a) HKD/USD for the period 1977-2017 (b) SGD/USD for the period 1981-2017

FIG. 1: Currency exchange rates

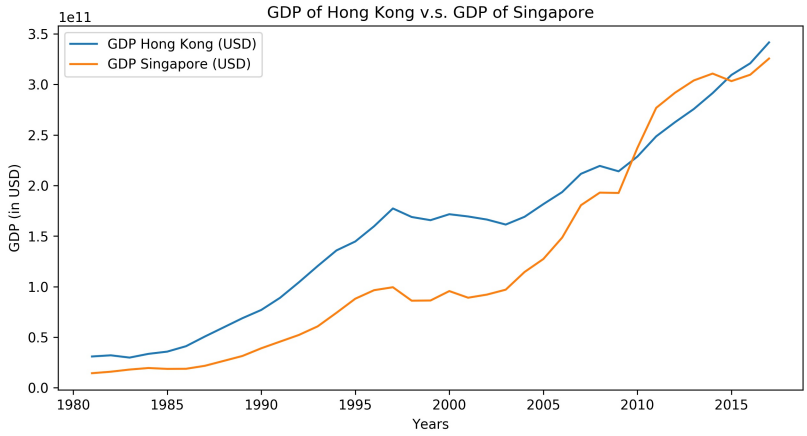


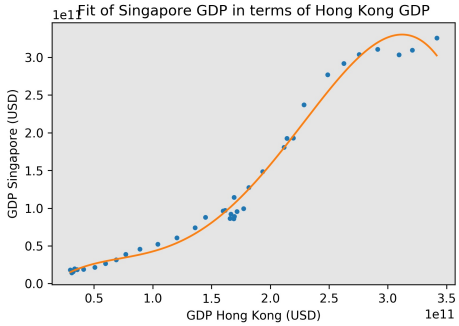
FIG. 2: GDPs evolution from 1981 to 2017 (in USD).

A. GDP and CPI

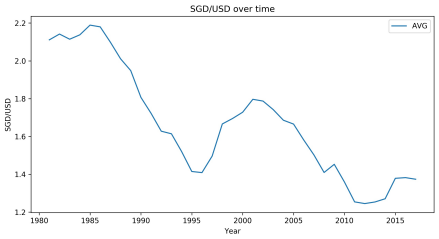
The dataframe obtained from UNdata contains rows for both Singapore and Hong Kong. After reshaping the dataframe, we form two of them called `df_HK` and `df_SG`. From the currency exchange rates dataframes called `price_HKDUSD` and `price_SGDUSD` we can plot the time evolution of HKD/USD and SGD/USD, as shown in Fig. 1a and 1b.

These currency rates allow us to convert the dataframes `df_HK` and `df_SG` from local currencies to USD. We select the GDP columns of these two dataframes and form a new dataframe called `df_GDP`. Plotting the columns against the years, we obtain the evolution of Singapore and Hong Kong GDPs over time in USD, as shown in Fig. 2.

We can also decide to fit the GDP of Hong Kong to see what type of fit best describe its evolution. We find that a polynomial fit works pretty well, with a best fit given by $f(t) = 4.927 \cdot 10^{7}\ t^2 - 1.888 \cdot 10^{11}\ t + 1.808 \cdot 10^{14}$ and t is in years.The corresponding plot is shown in Fig. 3a. As another interesting exercise, we also plot the GDP of Singapore in terms of Hong Kong's GDP. We find that a 4th-order polynomial of the form $GDP_{SG}(x) = -2.27 \cdot 10^{-34}\ x^4 + 1.364 \cdot 10^{-22}\ x^3 - 2.24 \cdot 10^{-11}\ x + 1.728\ x - 1.988 \cdot 10^{10}$, where x is the GDP of Hong Kong, best describes the data. The fit is shown in Fig. 3b, the R^2 value is 0.986, which shows a very good fit, and the MSE is $2.03 \cdot 10^{21}$ (USD^2).

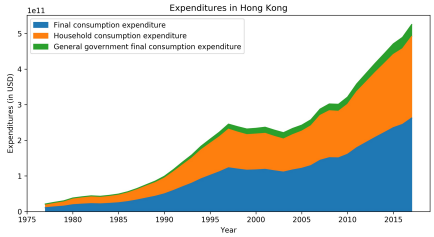


(a) Fit of Hong Kong GDP with a polynomial

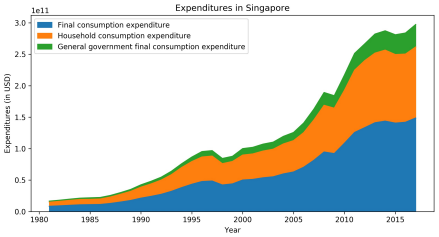


(b) Fit of Singapore GDP in terms of Hong Kong GDP

FIG. 3: Fits of GDPs



(a) Hong Kong expenditures for the period 1977-2017



(b) Singapore expenditures for the period 1981-2017

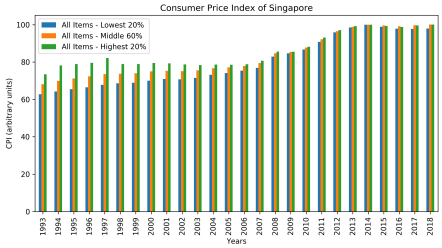
FIG. 4: Expenditures

Though not very useful in practice, this fit is still interesting as a curiosity.

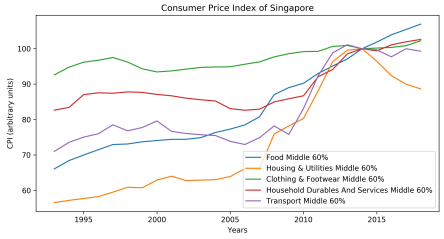
We also use the exchange rates to convert the expenditure values in USD for the two Asian tigers. The area plots are shown in Fig. 4a and 4b and clearly show that expenditures follow the GDP trend with pretty fair consistency.

The CPI values of the two tigers are obtained from different sources and as such put in two separated dataframes respectively called `CPI_SG` and `CPI_HK`. For Singapore, we first plot the CPI computed for all surveyed items, but separated in terms of price range, as shown in Fig. 6a. We then plot the CPI for 5 types of goods, focusing on the goods of intermediary price, as shown in Fig. 6b.

For Hong Kong, the CPI data is not separated in price ranges, so we select 5 categories of goods which are similar to those plotted for Singapore: food, housing, clothing, durable goods, transport. We first show a line plot of the corresponding CPIs, as displayed in Fig. **??**, and for a better representation of the variations during the period 1982-2017 we also plot a box chart in Fig. **??**.

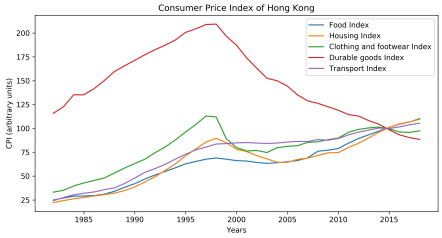


(a) Singapore's CPI for all items grouped by price range.

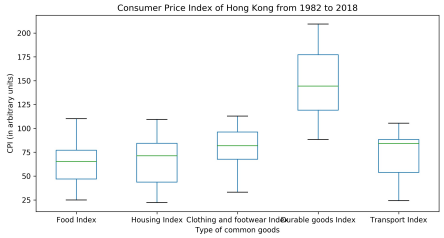


(b) Singapore's CPI for 5 types of goods (taking middle range price).

FIG. 5: CPIs of Singapore



(a) Hong Kong's CPI for 5 types of goods.



(b) Same but in a box plot.

FIG. 6: CPIs of Hong Kong

B. Venues from localisation data and k-means clustering

The location data for the two tigers are collected in two different phases. In the first one, we just collect few data for predetermined locations. In the second one, more locations are generated through random values of latitude and longitude around these first locations.

For Singapore, we first build a dataframe called `Dist_SG` that contains the names of sub-districts such as: Raffles Place, Marina, Cecil, Tanjong Pagar, etc. One location named "Bugis" is not included as its location on *Latitude.to* cannot be found. Using Folium, we plot a first map of Singapore presented in Fig. 7. Since we want a description in terms of districts, simpler for discussion, we take a location average of latitude and longitude over the different sub-districts and build a dataframe called `Dist_SG_concat`. Importing venues data from FOURSQUARE, with a limit of 100 venues per location, we build a dataframe called `Venues_SG` that contains several venues per district. We proceed to a one-hot encoding (`SD_onehot`) of it and group the venues (`SG_grouped`) which allow us to create a map with the top 5 venues in each district, as illustrated by the map of Fig. fig:SG-2. The circles displayed have a radius of 1 km and corresponds to the area in which the venues were searched for. Finally, we use a k-means clustering with $k = 3$ in order to cluster the districts into categories. The map of the districts according to these 3 categories is displayed in Fig. 9.

The logic is the same for Hong Kong, but as often in data science it is hard to start with a uniform source of data. First, the districts of Hong Kong do not seem to have a number like those in Singapore, second there is no table of them available, so we first scrape a Wiki page to get the

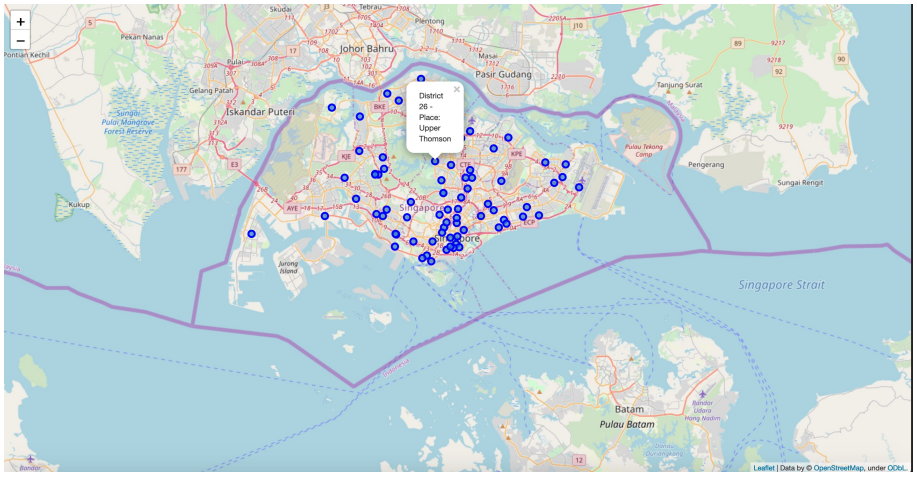


FIG. 7: Singapore sub-districts.

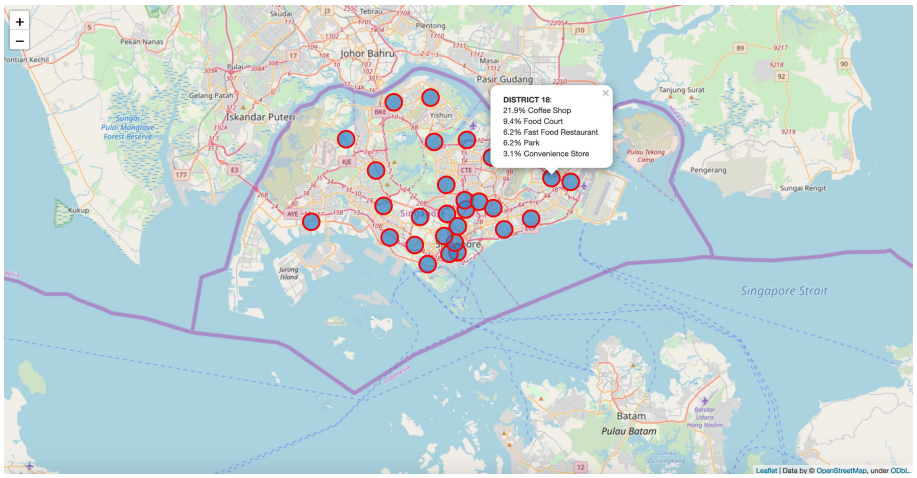


FIG. 8: Singapore districts with top 5 venues.

names of the cities of Hong Kong and their locations, building a dataframe called `Cities_HK`. The corresponding map is shown in Fig. 10. We then build a dataframe of Hong Kong districts from their names on Wikipedia and using *Latitude.to* to get their position. Using a radius of 1 km and a limit of 100 on the number of venues, we then collect the venues from FOURSQUARE and create similar dataframes as Singapore (`HK_onehot` and `HK_grouped`). We obtain a map with the top 5 venues per district, as presented in Fig. 11. The k-means clustering is similar to the previous one and gives a map which is not very satisfying considering the lack of a real separationg of districts, as it can be seen in Fig. 12.

As explained above, the first stage of map generations is based on well-defined data locations. In order to improve the k-means clustering, we need better statistics. For that reason, we decide to generate location data by creating random locations around the previously presented location spots. For simplicity of notations, the names of dataframes are kept the same as before, but they contain much more locations. For Singapore, we use the districts centers to generate the new data, and we

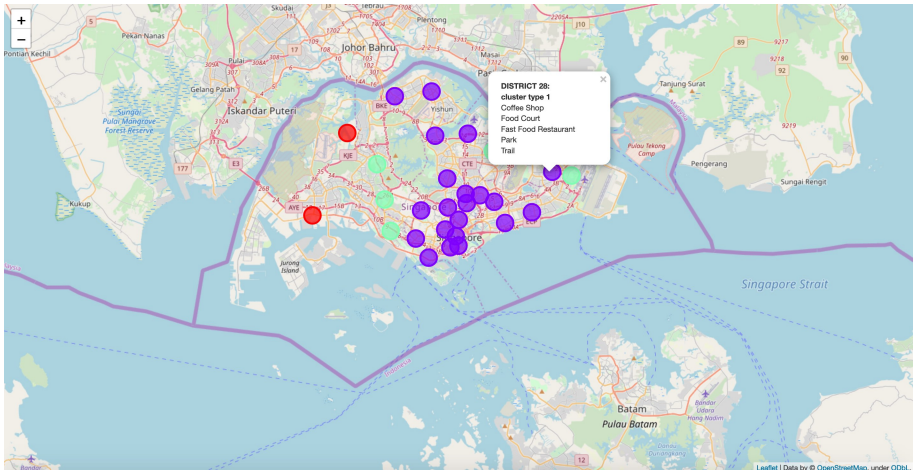


FIG. 9: Singapore districts clustered by k-means.

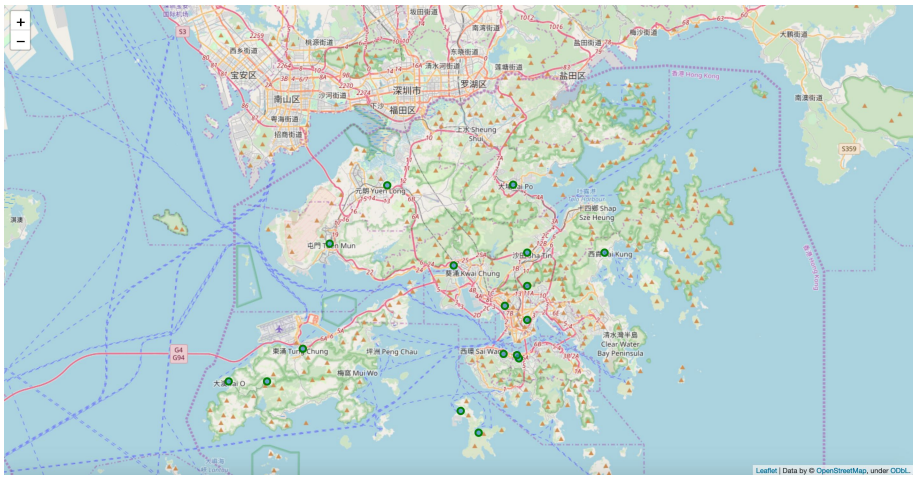


FIG. 10: Hong Kong cities.

get venues within a radius of 500 meters. Instead of taking an arbitrary number for the number of clusters in the k-means method, we do the clustering for k between 2 and 15 and count the number of locations in each cluster. The bar plot corresponding to these values is shown in Fig. 13. From that table we retain the value $k = 7$ as a good value and obtain the corresponding map of clustered locations in Fig. 14. It is clear from this map that the random generation of locations and the better evaluation of k's value lead to a more meaningful clustering of data.

We follow the same procedure for Hong Kong, but instead of taking the district centers to generate random locations this time, we instead chose to start with the cities locations as they allow a better centering on populated areas. Indeed, Hong Kong having mountain areas, choosing a location inside one of these areas can easily lead to no venues returned by FOURSQUARE (which actually happens for 43 locations out of 221, compared with only 9 out of 228 for Singapore). Testing the different values of k for the k-means clustering, we obtain the bar chart of Fig. 15. We then chose $k = 15$ as our best k and obtain the map of clustered locations shown in Fig. 16. Here also we notice that the

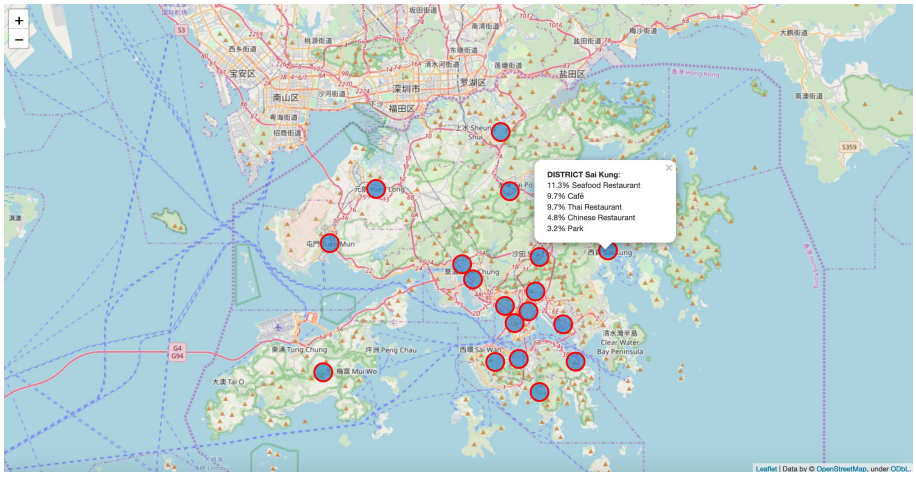


FIG. 11: Hong Kong districts with top 5 venues.

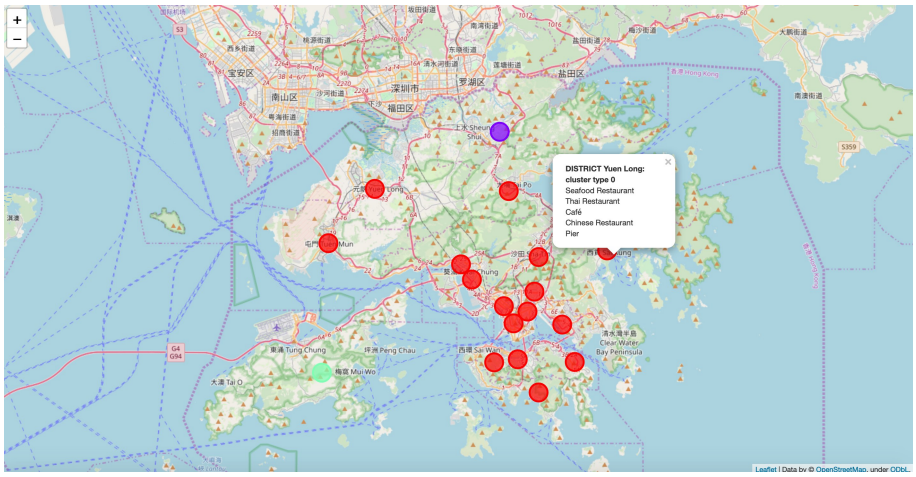


FIG. 12: Hong Kong districts clustered by k-means.

clustering is much better with the random generation of locations.

From the k-means clustering made over Singapore and Hong Kong, we find some clusters who appear to have an intrinsic description, as requested from clustering. We find the following most common venues among the most important clusters of Singapore:

- cluster 1: Café, Coffee Shop, Food Court, Chinese Restaurant, Japanese Restaurant.
- cluster 2: Bus Station, Food, Park, Zoo Exhibit, Café.
- cluster 5: Chinese Restaurant, Asian Restaurant, Food Court, Seafood Restaurant, Hotel.
- cluster 6: Coffee Shop, Food Court, Asian Restaurant, Chinese Restaurant, Fast Food Restaurant.

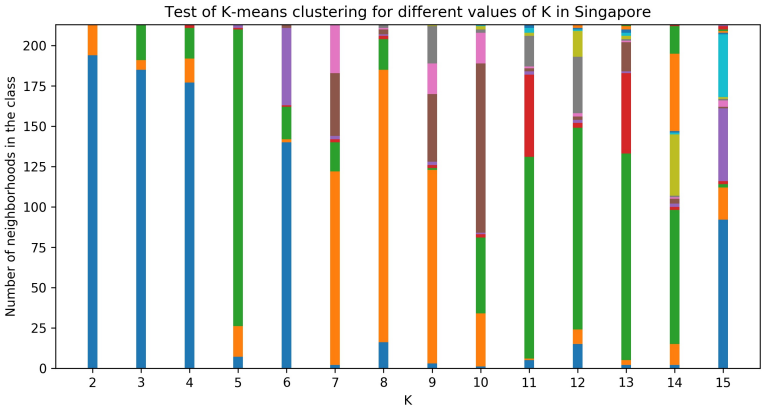


FIG. 13: K-means cluster populations for Singapore

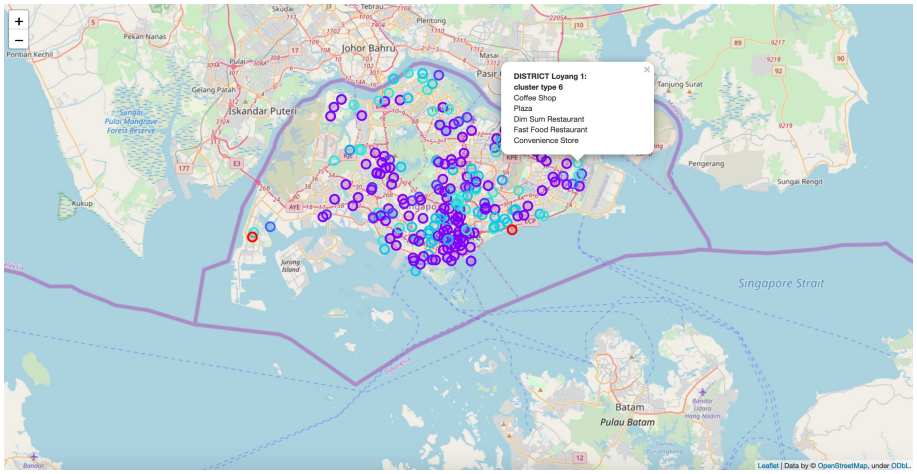


FIG. 14: Map of Singapore with clustered locations.

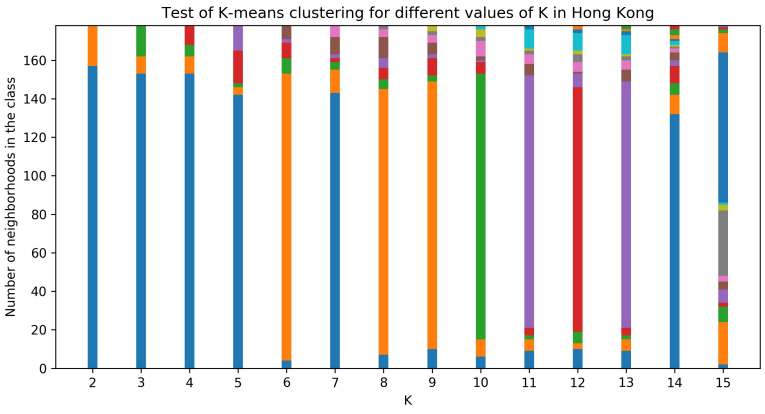


FIG. 15: K-means cluster populations for Hong Kong

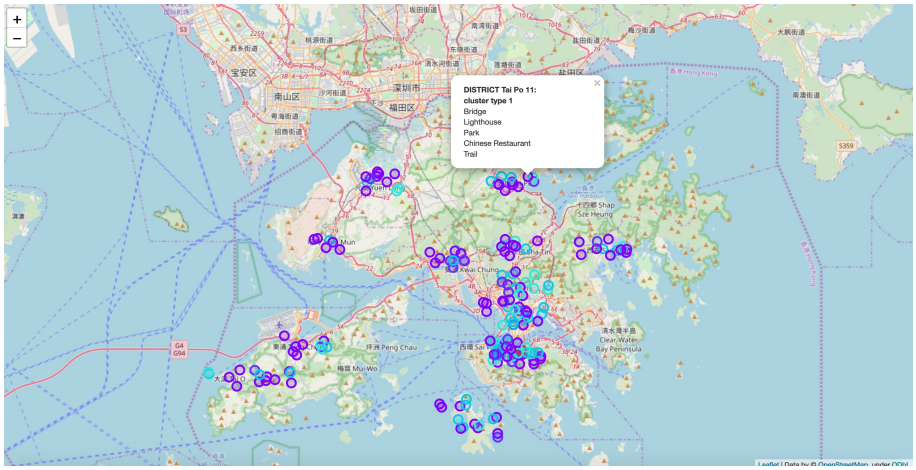


FIG. 16: Map of Hong Kong with clustered locations.

Doing the same for Hong Kong, we get the most important clusters

- cluster 1: Café, Seafood Restaurant, Chinese Restaurant, Dessert Shop, Coffee Shop.
- cluster 7: Fast Food Restaurant, Chinese Restaurant, Shopping Mall, Coffee Shop, Cha Chaan Teng.
- cluster 10: Café, Coffee Shop, Hotel, Chinese Restaurant, Zoo.
- cluster 11: Scenic Lookout, Zoo, Trail, Mountain, Fish Market.

C. Decision tree and KNN classifications

We go deeper into the comparison of Singapore and Hong Kong by creating a decision tree of the combined locations from the two places. We do a train/test split of 70%/30% and go to a depth up to 10, which is enough to reduce the initial sample of 273 elements to 87 non-classified elements, as shown in Fig. 17. Such a tree can be used to take any location and, as long as it can be written inside the same set of venues as the tree is built from, the tree can provide a classification as SG or HK for that location. The accuracy of the decision tree on the test set is 0.762.

Since we have a classification method, we would like to compare different cities in the world with Hong Kong and Singapore. For that we create a table of 10 cities from the world (other than the two tigers) and get their locations on *Locate.to*. We also get their venues from FOURSQUARE, taking a radius of 5 km around their city center and a limit of 200 venues in order to have a good representation of the cities as a whole. However, like mentioned above, we need the venues to be the same as the ones used to build the decision tree if we want to use our decision tree on them. For that reason, we merge the 10 cities with the dataframe of Hong Kong and Singapore random locations, and we train a new model on them. We choose to test K Nearest Neighbor (KNN) algorithm on this data. Testing different values of K, we obtain the accuracy plot of Fig. 18.

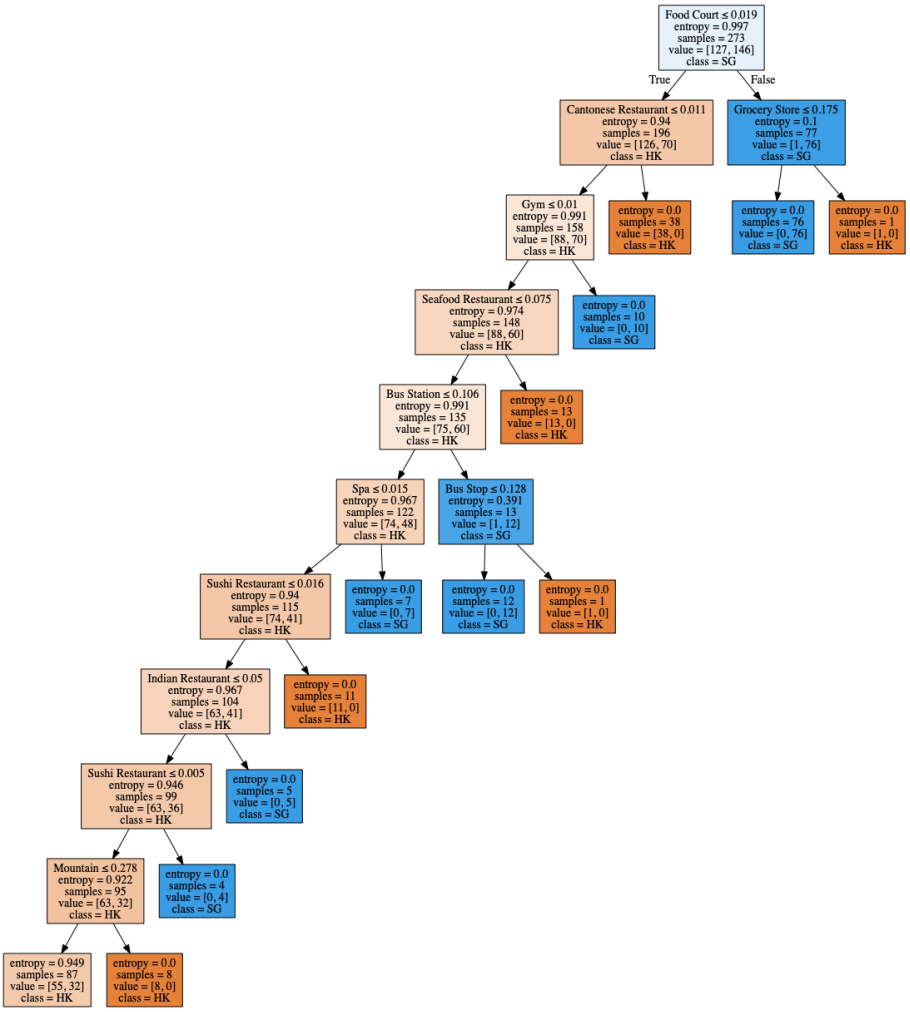


FIG. 17: Decision tree to classify locations in terms of their venues.

The training/test split is 80%/20% and the best accuracy is given at $K = 16$. For that value of K we obtain a training set accuracy of 0.852 and a test set accuracy of 0.898. the F1-score is 0.90 and the Jaccard similarity score is 0.90 also. Finally, the world cities that get classified as HK are Paris, London, Taipei, Jakarta, Kuala Lumpur and Guangzhou. The world cities that get classified as SG are Tokyo, New York, Shanghai and Seoul.

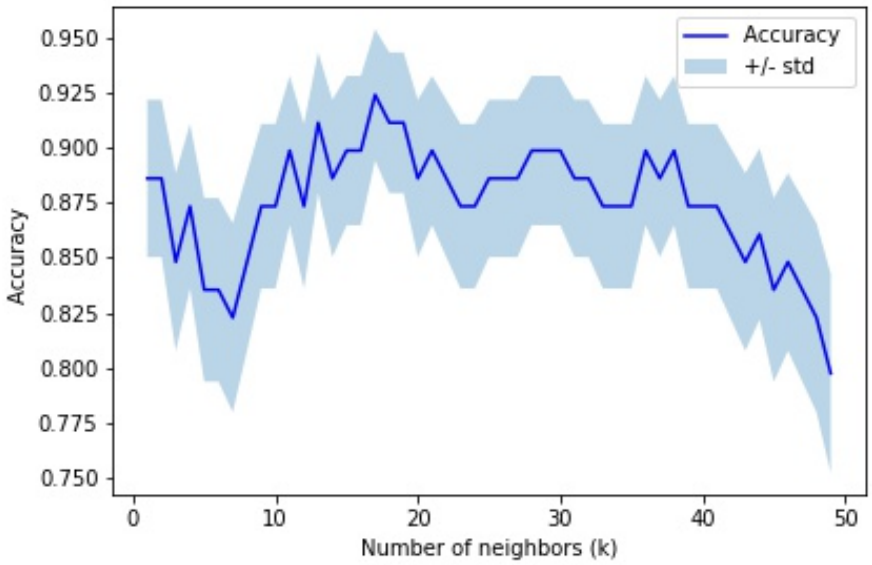


FIG. 18: Accuracy of the KNN model for different values of K.

IV. RESULTS: WHERE THE TIGERS LOOK ALIKE AND WHERE THEY DON'T

A. GDP and CPI

B. Venues from localisation data

V. DISCUSSION: WHAT COULD BE IMPROVED

VI. CONCLUSIONS

Let us conclude by quoting a song that many kids around the world know. In French, this song is called *Frère Jacques*, in English *Brother John*, but in Chinese this song is translated as "two tigers" (兩隻老虎), and the lyrics say:

兩隻老虎，兩隻老虎，跑得快，跑得快。
一隻沒有耳朵，一隻沒有尾巴，眞奇怪！眞奇怪！

Two tigers running, two tigers running, running fast, running fast.
One does not have ears, the other one has no tail, strange! Strange!

As we have seen, the two Asian tigers that are Singapore and Hong Kong are very similar, so similar on GDP that we could wonder in they are not twin tigers. Nevertheless some differences clearly exist between them, like the two tigers in the song. In the end, the important element from the song may not be that both tigers are unique, but that they are both running very fast!

Acknowledgement

This short paper is the final assignment of the **IBM Data Science** series of courses that I followed on Coursera between September 25, 2019 and November 1, 2019. This series contains 9 courses, most of them at intermediate level: *What is Data Science?*, *Open Source tools for Data Science*, *Data Science Methodology*, *Python for Data Science and AI*, *Databases and SQL for Data Science*, *Data Analysis with Python*, *Data Visualization with Python*, *Machine Learning with Python*, *Applied Data Science Capstone*.

This last course being the initial motivation of this work, I wish to thank the organizers of the course and previous courses with Coursera for giving me the opportunity to work on this topic that I chose according to my personal interests.

References

[1] *Economy of Hong Kong*, Wikipedia, https://en.wikipedia.org/wiki/Economy_of_Hong_Kong

[2] *Economy of Singapore*, Wikipedia, https://en.wikipedia.org/wiki/Economy_of_Singapore

[3] *Gross Domestic Product - GDP*, Investopedia, https://www.investopedia.com/terms/g/gdp.asp

[4] *Consumer Price Index - CPI*, Investopedia, https://www.investopedia.com/terms/c/consumerpriceindex.asp

[5] *Hong Kong vs Singapore - Which City Does it Better?*, Culture Trip, https://theculturetrip.com/asia/china/hong-kong/articles/hong-kong-vs-singapore-which-city-does-it-better/

[6] *GDP by Type of Expenditure at current prices - National currency*, UNdata, http://data.un.org/Data.aspx?q=GDP+by+Type+of+Expenditure+at+current+prices+&d=SNAAMA&f=grID%3a101%3bcurrID%3aNCU%3bpcFlag%3a0

[7] *USD/HKD - US Dollar Hong Kong Dollar*, Investing.com, https://www.investing.com/currencies/usd-hkd-historical-data

[8] *M212431 - Consumer Price Index (CPI) For Households In Different Income Groups, Base Year 2014 = 100, Annual*, Department of Statistics Singapore, https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=6771

[9] *Table E501 : Consumer Price Index at Commodity/Service Section/Group level (October 2014 - September 2015 = 100)*, Census and Statistics Department, https://www.censtatd.gov.hk/hkstat/sub/sp270.jsp?productCode=D5600001

[10] *Singapore District Guides*, iProperty.com, https://www.iproperty.com.sg/resources/District_Guide.aspx#

[11] *Find GPS coordinates for any address or location.*, Latitude.to, https://latitude.to/

[12] *Districts of Hong Kong*, Wikipedia, https://en.wikipedia.org/wiki/Districts_of_Hong_Kong

[13] *Geographic coordinates of Hong Kong*, Geodatos.net, https://www.geodatos.net/en/coordinates/hong-kong

[14] Jupyter, https://jupyter.org/

[15] FOURSQUARE Developers, https://developer.foursquare.com/ See also the search engine at https://foursquare.com/city-guide