Where is the French Silicon Valley?

Matt Sonnati[1]

March 30, 2019

1. Introduction

1.1. Background

In 2018, California GDP reached about $ 2.9 trillions, ranking California as the 5th biggest GDP in the world, ahead of UK (6th) and France (7th).1 Taking a deeper look, we see that most of California's economic activity is centered around its four major cities: San Francisco, San Jose, Los Angeles and San Diego.2 The area located between San Francisco and San Jose, is famously known as the Silicon Valley. This area is the hometown of some of the biggest Tech companies in the world: Google, Apple and Facebook. But what does make this place so special? Several researchers have tried to answer this trillion-dollar question.

In 1994, UC Berkeley Professor AnnaLee Saxenian compared the Silicon Valley to the Boston area known as Route 128, trying to understand what are the key features explaining Silicon Valley success.[3] One of her findings was that the clustering of certain types of companies and institutions such as university in a given area does not by itself explain the success of Silicon Valley.[4] The comparison with Route 128 area around Boston was interesting because, like Silicon Valley, it was the home of prestigious university (Harvard, MIT) and had a lot of technology advanced companies (Digital Equipment Corp., Apollo Computer) but as it turned out, it wasn't as successful in term of startup creation and growth.

Subsequent studies pointed out others Silicon Valley competitive advantages such as: spirit of cooperation, risk-taking culture, great talent attraction, cultural diversity, nice living environment, top-notch universities and research centers, etc.[5,6] Based on these findings, policymakers around the world have been trying to replicate Silicon Valley model of success. Consequently, the concept of Startup Ecosystem has emerged and is defined as people, startups in their various stages and various types of organizations in a location (physical or virtual), interacting as a system to create and scale new startup companies.[7] Some reports compare on a regular basis the various startup ecosystems in the world but, by far, Silicon Valley is always on the top spot.[8]

---

[1] LinkedIn, Github

In this study, we will focus on the comparison of the major French Startup Ecosystem with Silicon Valley and Boston area.

In June 2018, during the inaugural speech of Station F in Paris, French President Emmanuel Macron stated that France was to become a "startup Nation".[9] In addition to the recent opening of the new start-up incubator Station F in Paris, France has been updating its ecosystem for the last 10 years to make it more startup-friendly both for entrepreneurs and investors.[10] Among the key initiatives in favor of startups, we can highlight the creation of the FrenchTech label,[11] the Public Investment Bank[12] (Bpifrance) and the research tax-credit[13] (CIR).

All of these initiatives have deeply impacted the French startup scene and today, most of the major French cities have promising local startup ecosystems. It is a great time to take a deeper look at those startup ecosystems and try to understand how they compare to Silicon Valley.

1.2. Problem

The question that this study tries to answer is: "Where is the French Silicon Valley?", that-is-to-say which French city has the closest startup ecosystem to the one of the Silicon Valley.

Among the several possible approaches, we choose to explore the geographical clustering of specific type of companies and organizations such as universities in specific areas centered around 13 major French cities and two US areas: Silicon Valley on the west coast and Boston area (Route 128) on the east coast.

Using Foursquare Places API geo dataset, we will try to characterize each of these startup ecosystems and compare them through modelling (K-means clustering, linear regression).

1.3. Interest

The result of this data analysis will be of great interest for those who want to better understand the startup landscape in France. Several uses case for this analysis could be considered:

- Local development policymakers will be interested to benchmark how their local startup ecosystem compare to other in France and in the US,

- Startup investors will be interested to identify which startup ecosystem offer the best investing opportunities,

- Startup founder will be interested to know where the best startup ecosystems are to help them potentially chose where to create their startup.

2. Data acquisition and preparation

2.1. Data sources

2.1.1. Foursquare Places API

Foursquare is a location technology company that publish apps enabling their users to check in and leaves comment about places they visit.[14] As a result, they developed over the years one of the biggest places database with more than 105 millions places labeled in more than 300 categories with corresponding address, tips, photos, etc.

In April 2016, Foursquare CEO Jeff Glueck predicted that Chipotle's sales will be down nearly 30% by the next quarter based on the data analysis of their place database.[15] Developers around the world can also access part of these data through their Foursquare Places API.[16] Several tiers are available.

Data used in this study were accessed through a Personal Free account with the following restrictions:

- 105M places
- 2 Photos & 2 Tips per Venue
- 2 Queries per Second (QPS)
- 1 App per Account
- 99,500 Regular Calls / Day
- 500 Premium Calls / Day

Based on these restrictions, we chose to focus our analysis on the number of places in selected categories for each of the 15 startup ecosystems.

2.1.2. List of startup ecosystems

The list of French cities selected for this study is as follows: Bordeaux, Clermont-Ferrand, Grenoble, Lille, Lyon, Marseille Aix-en-Provence, Montpellier, Nantes, Nice Sophia Antipolis, Paris, Rennes, Strasbourg and Toulouse.

These 13 major French cities were selected according to two articles which listed the major startup ecosystem in France in 2018.[17,18] Noteworthy, most of these cities are also accredited as part of FrenchTech label, which is a good indicator of the level of maturity of the corresponding local startup ecosystems.[19]

Obviously, Silicon Valley was added as main the reference point for the best Startup Ecosystem in the world. Boston area (Route 128) was also added as a reference point, based on the original article from UC Berkeley Professor AnnaLee Saxenian.[3,4]

2.2. Data acquisition

Data were acquired and prepared using one jupyter notebook for each of the 15 startup ecosystems studied here. These notebooks are available on github.[20]

Silicon Valley (SV) is not centered in a specific city. In fact, SV extends from San Francisco in the North to San Jose in the South. As a result, we chose Palo Alto, CA as the approximate center of the area and asked Foursquare API for the list of up to 50 cities within a 50 km radius from Palo Alto, CA. The list of places in each of the selected categories was then obtained within a 5 km radius around each of the city in the list for the specific startup ecosystem. Duplicates were dropped.

The process described here-above was repeated for each of the 14 others startup ecosystems and are detailed in the notebooks listed in Table 1 below.

Table 1. List of the 15 startup ecosystems studied and the corresponding number of places extracted for each area before and after feature selection.

Country	Area	Number of places (before feature selection)	Number of places (after feature selection)
USA	Silicon Valley	11277	8192
USA	Boston area	9301	5906
France	Paris area	8276	5886
France	Lille area	4120	3031
France	Lyon area	2961	2052
France	Marseille Aix-en-Provence area	2182	1357
France	Bordeaux area	1755	1269
France	Toulouse area	1729	1207
France	Nice Sophia Antipolis area	1824	1148
France	Nantes area	1548	1082
France	Montpellier area	1456	1006
France	Strasbourg area	1393	896
France	Rennes area	1111	770
France	Clermont-Ferrand area	832	565
France	Grenoble area	799	548
	Total	50564	34915

2.3. Features selection

Foursquare has more than 300 categories.[21] It goes from Restaurants to Stores and some of these categories can widely impact any analysis just by their sheer number of places. As a result, after some preliminary data exploration, we had to focus on a close selection of 46 categories that are directly related to the life of startups, that-is-to-say places that constitutes a startup ecosystem.[7]

After data preparation, we got 34915 places combined in a 46 x 15 dataframe corresponding to the 46 categories and the 15 startup ecosystems. The corresponding jupyter notebook is available here on github.[20]

The first category selected was Tech startup as this is the most direct indicator of the activity of a startup ecosystem. We then selected all the services that are required to the common operations of a startup (Financial or Legal Service, Coworking Space, Business Center, etc.), categories related to potential academic or research partners of startup (College & University, College Lab, Laboratory, Research Station, etc.), categories related to potential industrial partners of startup (Factory, Office, Industrial Estate), categories related to potential governmental partners of startup (Government Building, Capitol Building, City Hall) and categories related to various events and cultural centers that are required for startup to network and meet potential users, customers and investors (Event Space, Conference, Museum, etc.).

All other categories were dropped either because the number of place in the corresponding category was null or negligible, or because, to the opposite, the sheer number was hindering any insightful conclusion (e.g. the number of French restaurant in France is not really helpful when trying to compare French startup ecosystem to US ones).

One of the most problematic part of the feature selection was the need to drop two key places categories which were Bank and Lawyer. One may think that these two categories are closely related to a startup ecosystem and thus must be include in the analysis. However, the preliminary data analysis carried out with these two categories yielded to inconclusive results. These two categories both correspond to large number of places in each of the startup ecosystem studied and were almost always in the top 10 categories with the most places. While counterintuitive, we ended up deciding to drop these two categories and after this we were able to compare more easily the various ecosystems. This point will be discussed in more details in the Results and Discussion part of the present report.

3. Methodology

3.1. Exploratory Data Analysis

As a framework for our Exploratory Data Analysis (EDA), we followed the Data Science Methodology CRISP-DM.[22] The objective of this EDA was to get a satisfactory data understanding before pursuing to modelling and evaluation. As stated previously, we used geo data extracted from Foursquare Places API and we focused on the types (categories) and quantities of companies and organizations that are related to startup ecosystems.

3.1.1. Raw data

We started by taking a look at the raw data without any normalization. Data were sorted by descending number of places in the Tech startup category as this is the main feature of interest. In term of categories, the list of categories was sorted by descending number of places for Silicon Valley area as this is our reference. The top 5 area and top 5 categories are listed in Table 2.

Table 2: Number of places for the top 5 areas and top 5 categories.

	Financial or Legal Service	Tech Startup	Business Service	Coworking Space	Factory	Total
Silicon Valley	1002	924	890	851	595	8192
Paris area	493	615	470	603	448	5886
Boston area	903	459	851	604	460	5906
Lille area	192	304	124	365	463	3031
Lyon area	211	204	166	232	254	2052

Silicon Valley has by far the most places in the whole range of the 46 selected categories with a total of 8192 places. After Silicon Valley comes Boston and Paris which have respectively 5906 and 5886 places. While Boston comes 2nd in term of overall places, Paris is far ahead of Boston in term of Tech startup. For these three areas, the top 5 categories of places are: Financial or Legal Service, Tech Startup, Business Service, Coworking Space, Factory. Lille comes 4th with Factory as the category with the largest number of places. Lyon is the 5th city in term of Tech startup. Figure 1 shows the number of places in each of the top 30 categories for the 15 startup ecosystems.

The results for the 10 other startup ecosystems are much closer, with a total number of places comprised between 548 and 1357. As seen on Figure 2, the top 5 and the bottom 10 startup ecosystems are widely different in term of absolute number of places. This may be related to the difference of population densities in those areas. Normalization will be helpful to carry out a relative analysis of the various areas on a more similar basis.

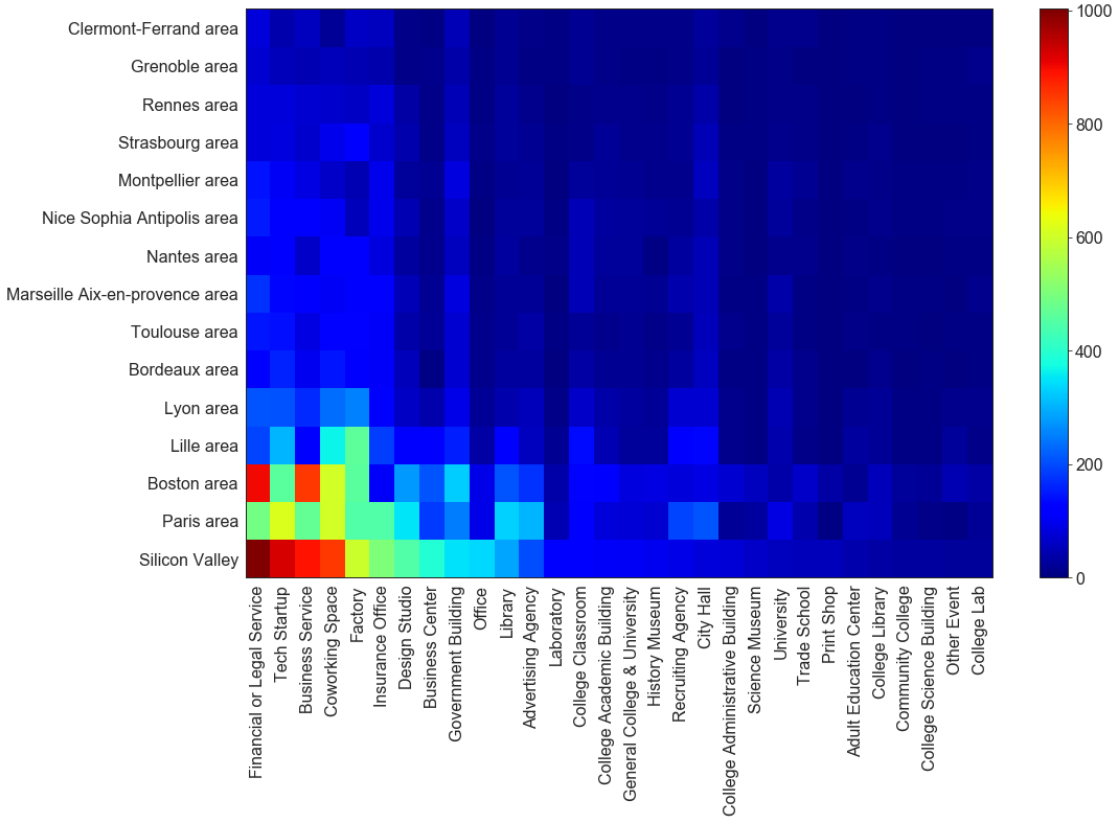

Figure 1: Heatmap of the numbers of places in each of the top 30 categories for each startup ecosystem.

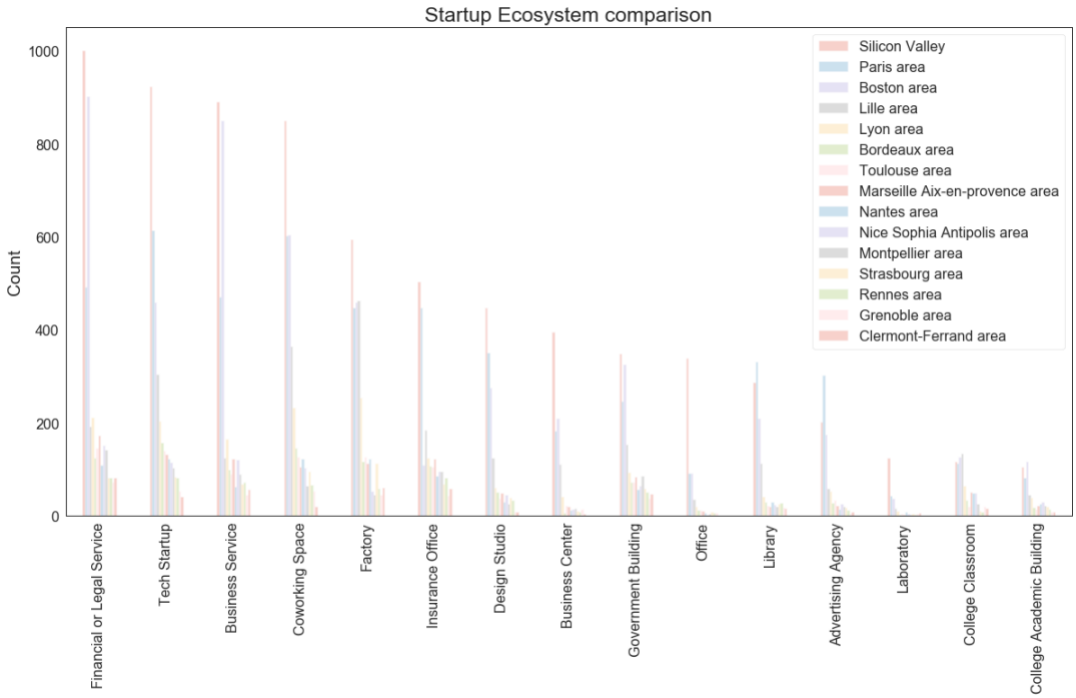

Figure 2: Bar plot of the numbers of places in each of the top 15 categories for each startup ecosystem.

3.1.2. Normalized data

As described previously, there is a wide difference between the numbers of places of the top 5 areas with the most places and the bottom 10. In order to better compare the various startup ecosystems, we thus decided to normalize the data based on the number of Tech startup in each startup ecosystem. The corresponding ratio of the top 5 categories are shown in Table 3.

Table 3: Normalized ratio of places for the top 5 categories.

	Financial or Legal Service	Tech Startup	Business Service	Coworking Space	Factory
Silicon Valley	1.0844	1	0.9632	0.9210	0.6439
Paris area	0.8016	1	0.7642	0.9805	0.7285
Boston area	1.9673	1	1.8540	1.3159	1.0022
Lille area	0.6316	1	0.4079	1.2007	1.5230
Lyon area	1.0343	1	0.8137	1.1373	1.2451
Bordeaux area	0.7848	1	0.6329	0.9241	0.7342
Toulouse area	1.0357	1	0.6429	0.9071	0.9000
Marseille Aix-en-Provence area	1.3008	1	0.9173	0.7895	0.8496
Nantes area	0.8852	1	0.5164	1.0082	1.0000
Nice Sophia Antipolis area	1.3217	1	1.0435	0.8957	0.4522
Montpellier area	1.3786	1	0.8641	0.6311	0.4369
Strasbourg area	0.9643	1	0.8214	1.1429	1.3452
Rennes area	1.0000	1	0.9012	0.8272	0.7284
Grenoble area	1.3846	1	0.8654	1.0192	0.8846
Clermont-Ferrand area	1.9286	1	1.3333	0.4762	1.4286

- As seen in Paris and Boston areas have the largest ratio of Advertising agencies.

- Clermont-Ferrand, Silicon Valley and Grenoble areas have the largest ratio of laboratories. Bordeaux and Montpellier have the lowest ratio of Laboratories.

, the descending order of categories is more or less the same for any of the 15 startup ecosystems. However, when looking more closely, a few observations can be highlighted:

- Boston and Clermont-Ferrand areas have the largest ratio of Financial or Legal services, followed by Grenoble, Montpellier, Nice-Sophia Antipolis and Marseille areas. Silicon Valley has almost a 1:1 ratio of Financial or Legal services. Lille area has the lowest ratio of Financial or Legal services.

- Boston and Clermont-Ferrand areas have both the largest ratio of Business services, while Lille, Nantes, Toulouse and Bordeaux areas have the lowest ratio of Business services.

- Clermont-Ferrand and Montpellier areas both seem to have a relatively lower ratio of Coworking spaces compared to other startup ecosystems. Boston, Lille, Lyon and Strasbourg areas have relatively larger ratio of Coworking spaces.

- Lille, Clermont-Ferrand, Strasbourg and Lyon areas have relatively larger ratio of Factories. Montpelier and Nice-Sophia Antipolis areas have the lowest ratio of Factories.

- Clermont-Ferrand area has the largest ratio of Insurance offices, while Boston area has the lowest.
- Boston, Paris, Silicon Valley and Strasbourg areas have the largest ratio of Design Studio. Grenoble and Clermont-Ferrand have the lowest ratio of Design Studio.
- Boston area and Silicon Valley have the largest ratio of Business centers, followed by Lille, Paris and Grenoble areas. All other areas have relatively low ratio of Business centers.
- Clermont-Ferrand and Montpellier areas have the largest ratio of Government buildings compared to other startup ecosystems.
- Silicon Valley by far as the largest ratio of Office.
- Paris and Boston areas have the largest ratio of Libraries.
- Paris and Boston areas have the largest ratio of Advertising agencies.
- Clermont-Ferrand, Silicon Valley and Grenoble areas have the largest ratio of laboratories. Bordeaux and Montpellier have the lowest ratio of Laboratories.

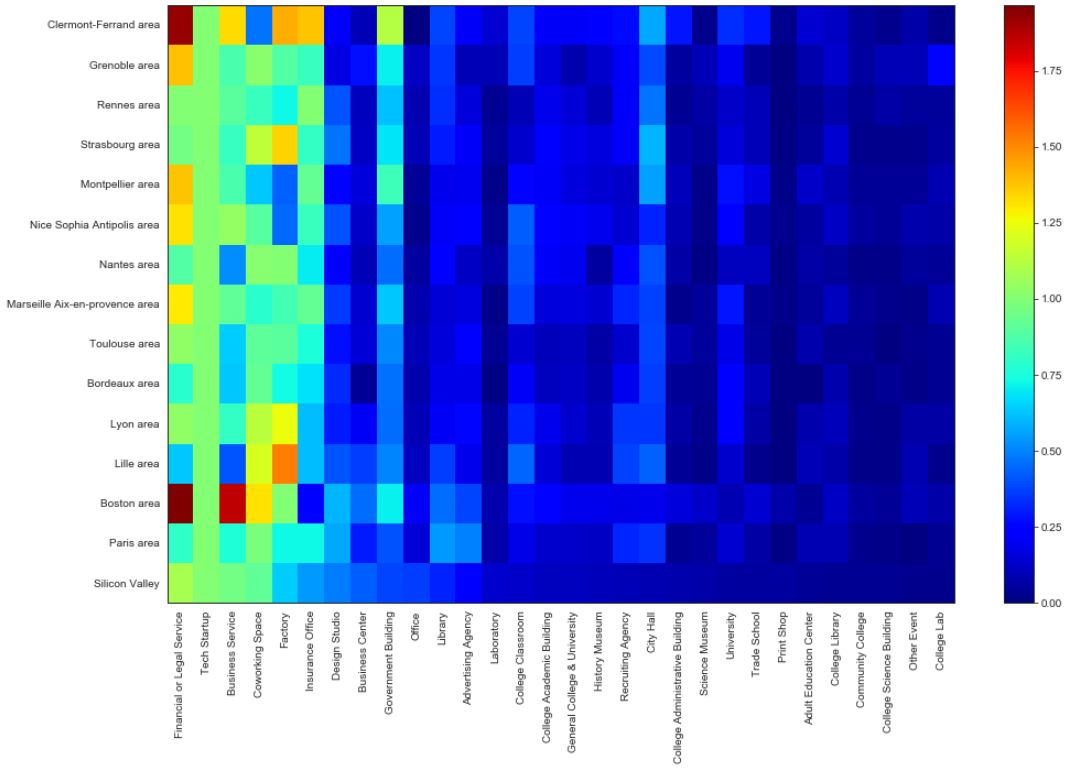

Figure 3: Heatmap of the normalized ratio of places in each of the top 30 categories for each startup ecosystem.

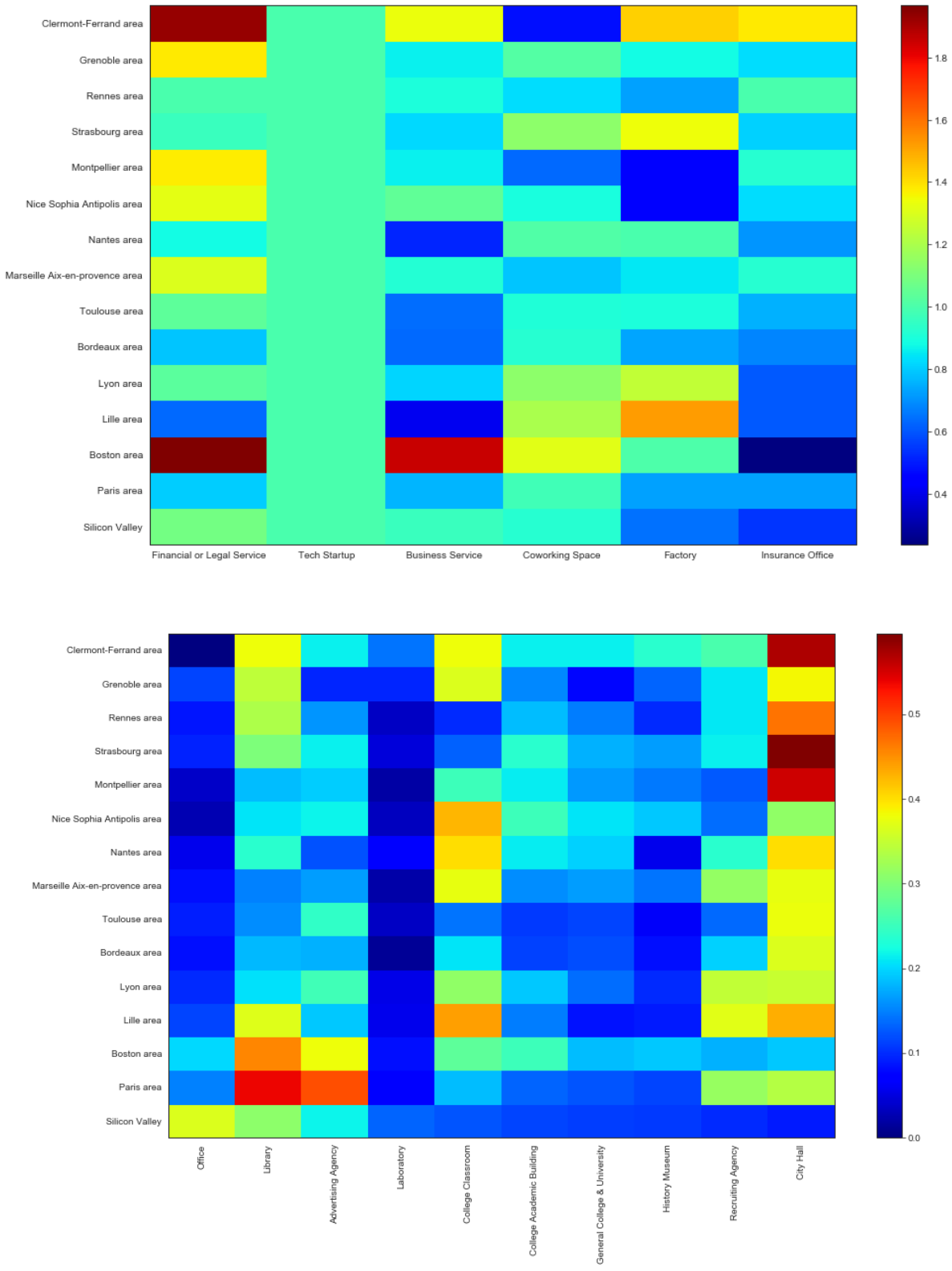

Figure 4: Heatmaps of the normalized ratio of places for selected categories.

3.2. Machine Learning

Our objective is to model an ideal startup ecosystem based on Silicon Valley example. Among the many approaches, we choose two machine learning techniques: K-means clustering and linear regression.

3.2.1. K-Means clustering

First, we applied K-Means clustering to the raw number of places in each category and for each startup ecosystem. As seen in Figure 5, K-means clustering of the 15 startup ecosystems into 6 clusters based on the raw data leads to one cluster for the bottom 10 areas and one cluster for each of the top 5 areas: Silicon Valley, Boston, Paris, Lille and Lyon.

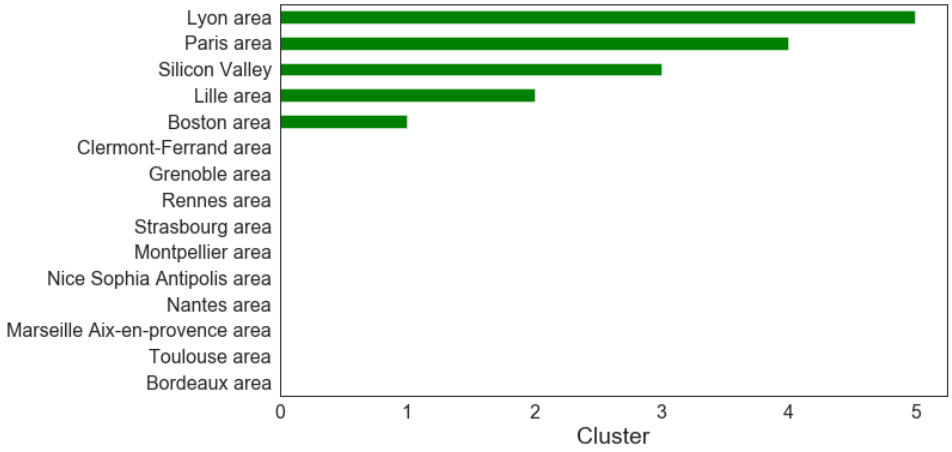

Figure 5: Raw data K-means clustering of the 15 startup ecosystems into 6 clusters (k = 6).

Model evaluation of K-means clustering using the Elbow method indicates that the best value for k is at the spot where sum of squared distances (SSD) starts to flatten out and forming an elbow. As seen on Figure 6, there is no sudden flattening meaning that the more cluster the smaller the SSD.

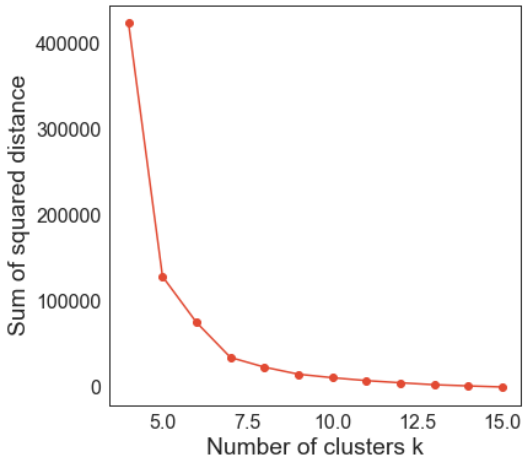

Figure 6: Sum of squared distances for K-means clustering with k comprised between 4 and 15.

After normalizing the number of places in each category based on the number of Tech startup in each startup ecosystems, we ran K-means clustering. As seen Figure 7, K-means clustering of the 15 startup ecosystems into 6 clusters based on the normalized data leads to one cluster with Silicon Valley and Paris, one cluster with Boston, one cluster with Clermont-Ferrand, one cluster with Lille area, one cluster with Lyon, Strasbourg, Nantes, Toulouse and Bordeaux areas and one cluster with Grenoble, Rennes, Montpellier, Nice Sophia Antipolis and Marseille-Aix-en-Provence areas.

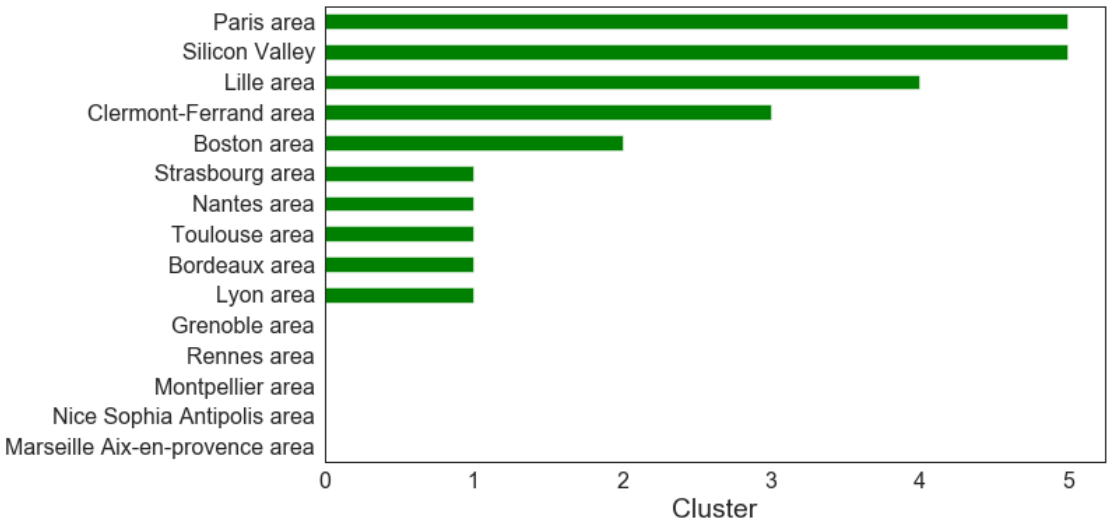

Figure 7: Normalized data K-means clustering of the 15 startup ecosystems into 6 clusters (k = 6).

When increasing the number of clusters to 7, we see that Paris and Silicon Valley are still in the same cluster while the bigger cluster starts to divide into smaller ones (Figure 8).

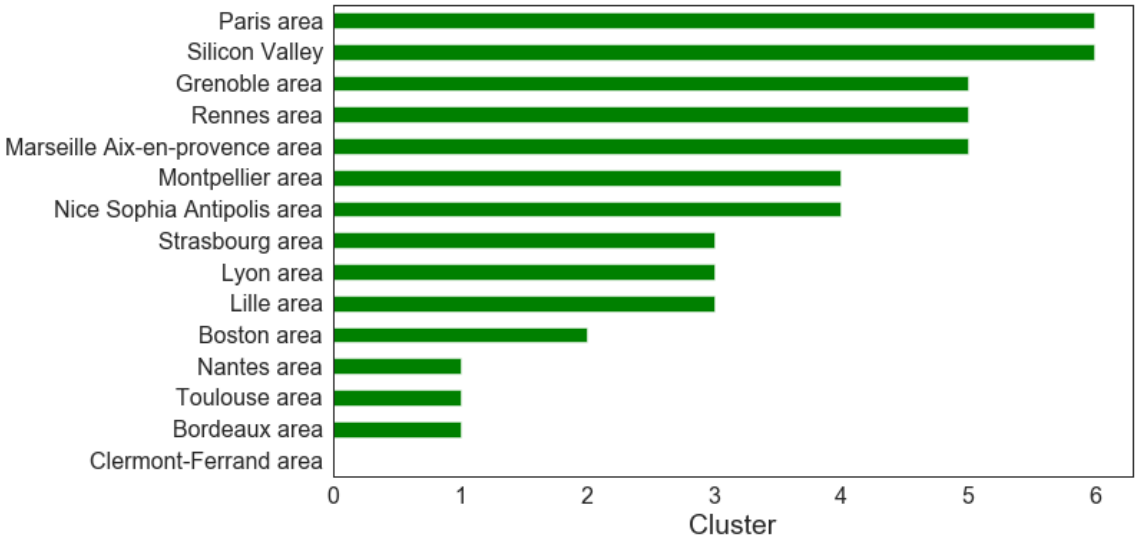

Figure 8: Normalized data K-means clustering of the 15 startup ecosystems into 7 clusters (k = 7).

When increasing the number of clusters to 8, we see that Paris and Silicon Valley are no longer in the same cluster (Figure 9).

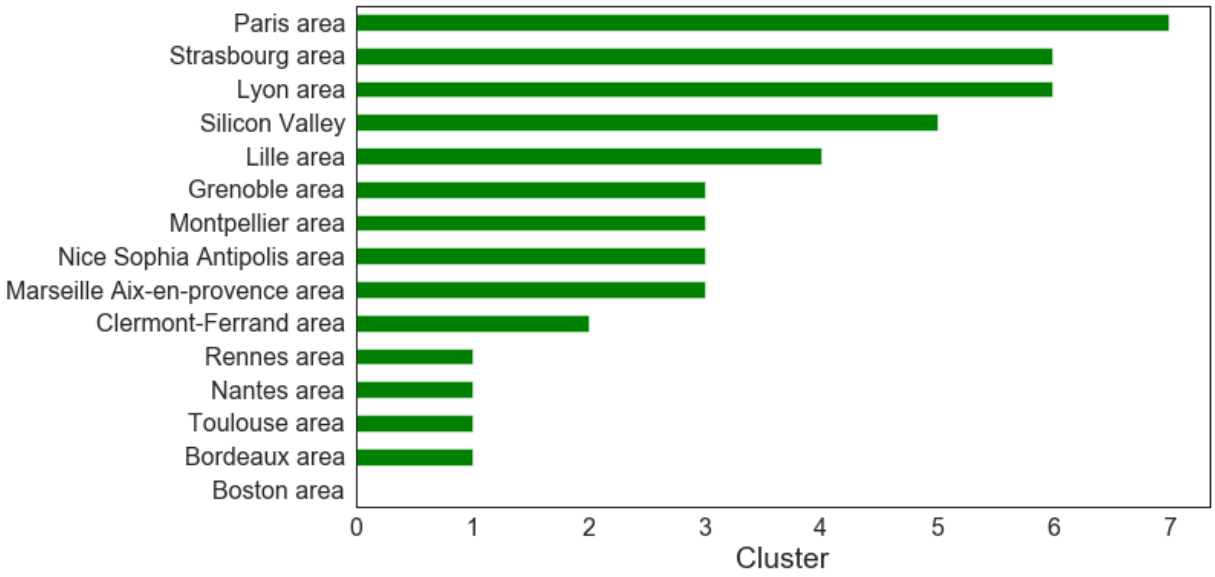

Figure 9: Normalized data K-means clustering of the 15 startup ecosystems into 8 clusters (k = 8).

Model evaluation of K-means clustering using the Elbow method indicates that the best value for k is at the spot where SSD starts to flatten out and forming an elbow. As we can see on the Figure 10, there is no sudden flattening meaning that the more cluster the smaller the SSD.

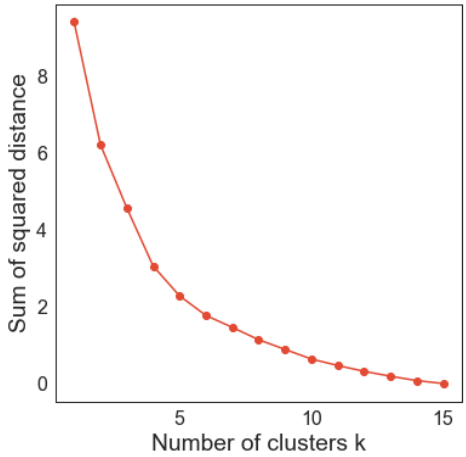

Figure 10: SSD for K-means clustering with k comprised between 4 and 15.

3.2.2. Linear regression

Using raw data, we decided to evaluate how the number of places in each of the top 5 categories correlates with the number of Tech startup in each startup ecosystem. As seen on Figure 11, we clearly see the distinction between the data points corresponding to the bottom 10 areas (around 100 places) and the data points corresponding to the top 5

areas. Noteworthy, Silicon Valley has the largest number of places in each of the top 5 categories and is easily identified as the upper right corner data point on the graphs. The number of coworking spaces seems to be the best indicator of the number of Tech startups.

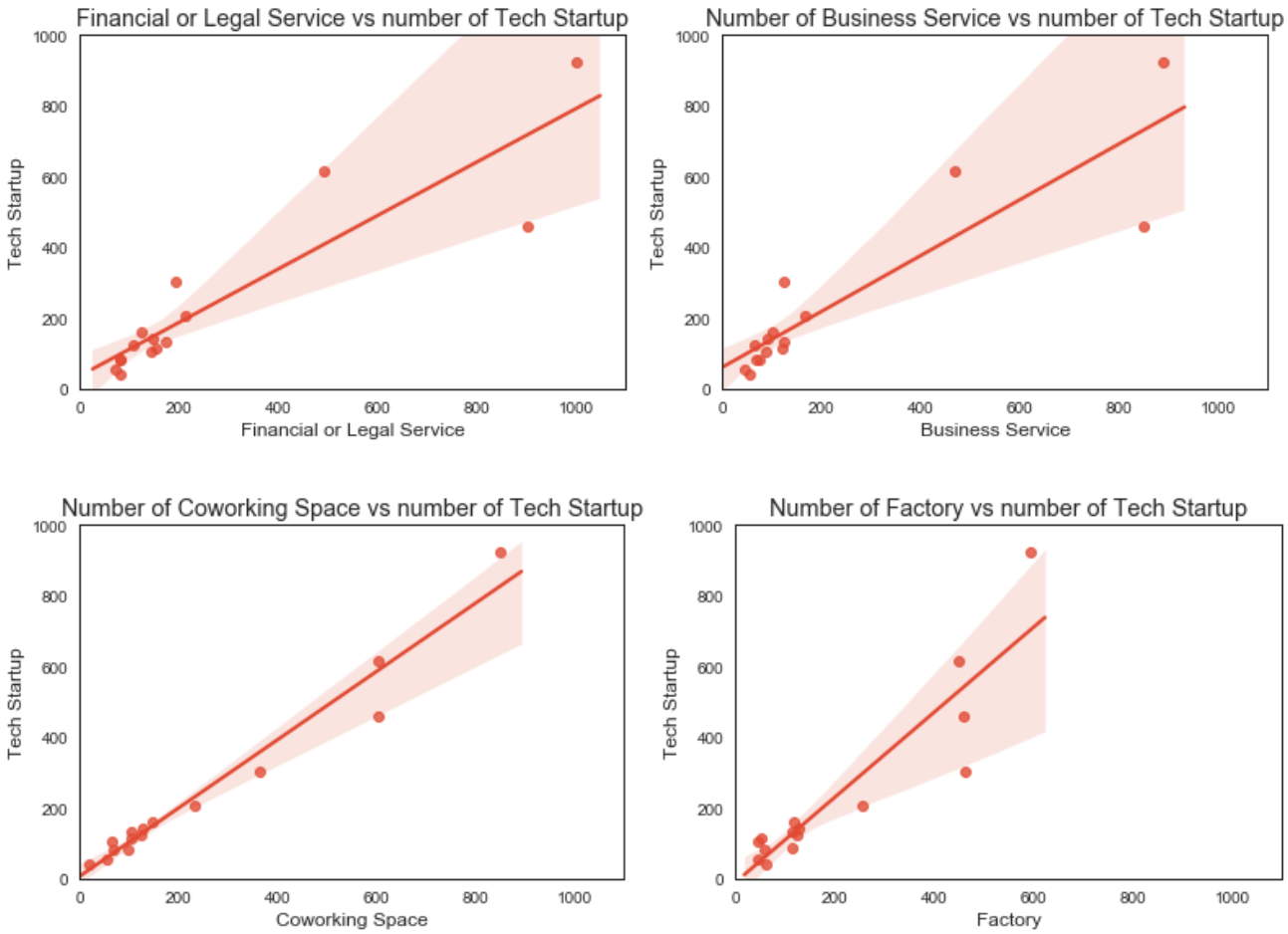

Figure 11: Linear regression of the number of Financial or Legal Service, Business service, Coworking space and Factory with the number of Tech startup in each of the 15 startup ecosystems.

Another way to investigate the similarities between the various startup ecosystems is to compare them to Silicon Valley using linear regression. As seen on Figure 12, we plotted the normalized ratio of places in each category for the various startup ecosystems as a function of the normalized ratio of places in each category for Silicon Valley.

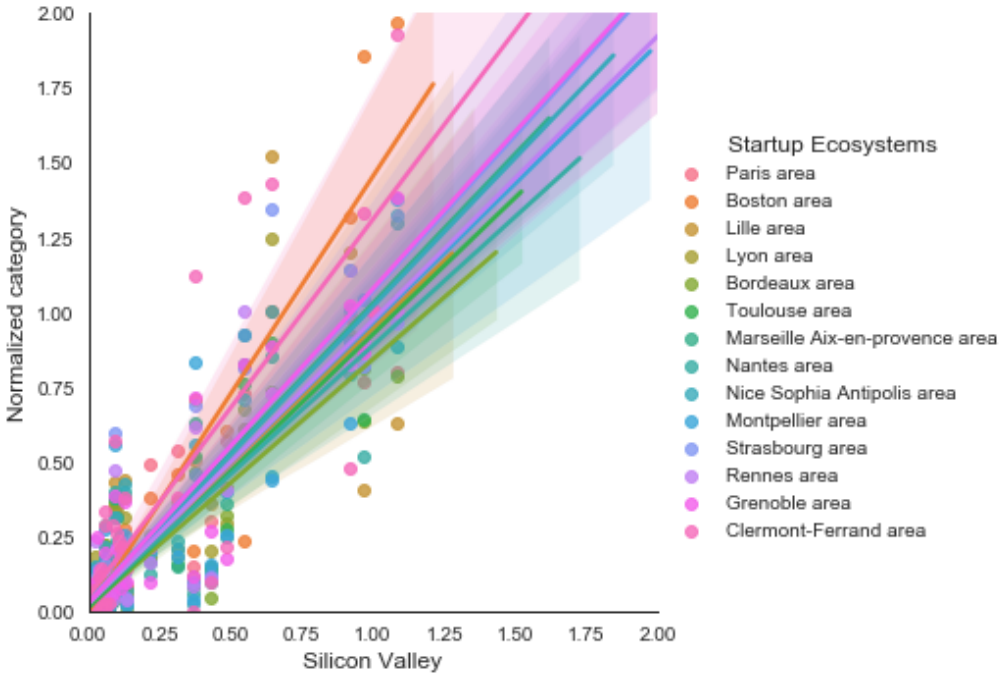

Figure 12: Linear regression of the normalized ratio of places for each startup ecosystem compared to Silicon Valley.

Ideally, the closer to 1 the slope and the correlation coefficient are, the more similar the startup ecosystem is to the Silicon Valley. Table 4 shows the Pearson correlation coefficient, intercept, p-value and slope of the 15 startup ecosystems.

Table 4: Linear regression parameters of the 15 startup ecosystems.

	Pearson correlation coef.	intercept	p-value	slope
Bordeaux area	0.9128	0.0166	1.01E-18	1.0089
Silicon Valley	1	0	0.00E+00	1
Paris area	0.9360	-0.0064	1.40E-21	0.9570
Toulouse area	0.9238	0.0176	5.81E-20	0.9344
Nantes area	0.8797	0.0193	8.44E-16	0.8996
Rennes area	0.9134	0.0109	8.82E-19	0.8799
Nice Sophia Antipolis area	0.9238	0.0060	5.76E-20	0.8603
Marseille Aix-en-Provence area	0.9126	0.0077	1.06E-18	0.8341
Lyon area	0.9076	0.0124	3.45E-18	0.8247
Grenoble area	0.9223	0.0071	8.70E-20	0.8104
Montpellier area	0.8645	0.0221	9.80E-15	0.8035
Strasbourg area	0.8841	0.0181	3.93E-16	0.7532
Lille area	0.8033	0.0374	1.83E-11	0.7166
Boston area	0.9391	0.0230	4.95E-22	0.6069
Clermont-Ferrand area	0.8174	0.0376	4.21E-12	0.5305

Paris area has both the closest correlation coefficient to 1 and the closest slope to 1. This seems to indicate that Paris startup ecosystem would be one of the closest to Silicon Valley.

Bordeaux also has a slope very close to 1 and a good correlation coefficient. Toulouse, Nantes, Rennes and Nice-Sophia Antipolis also have both slope and correlation coefficient close enough to 1. This could indicate that their startup ecosystems would be quite close to Silicon Valley one.

Boston has a higher correlation coefficient than Paris but the slope inferior to 1. Lille, Boston and Clermont-Ferrand have the lowest slopes (0.53 - 0.72). This indicate that their startup ecosystems are significantly different from Silicon Valley one.

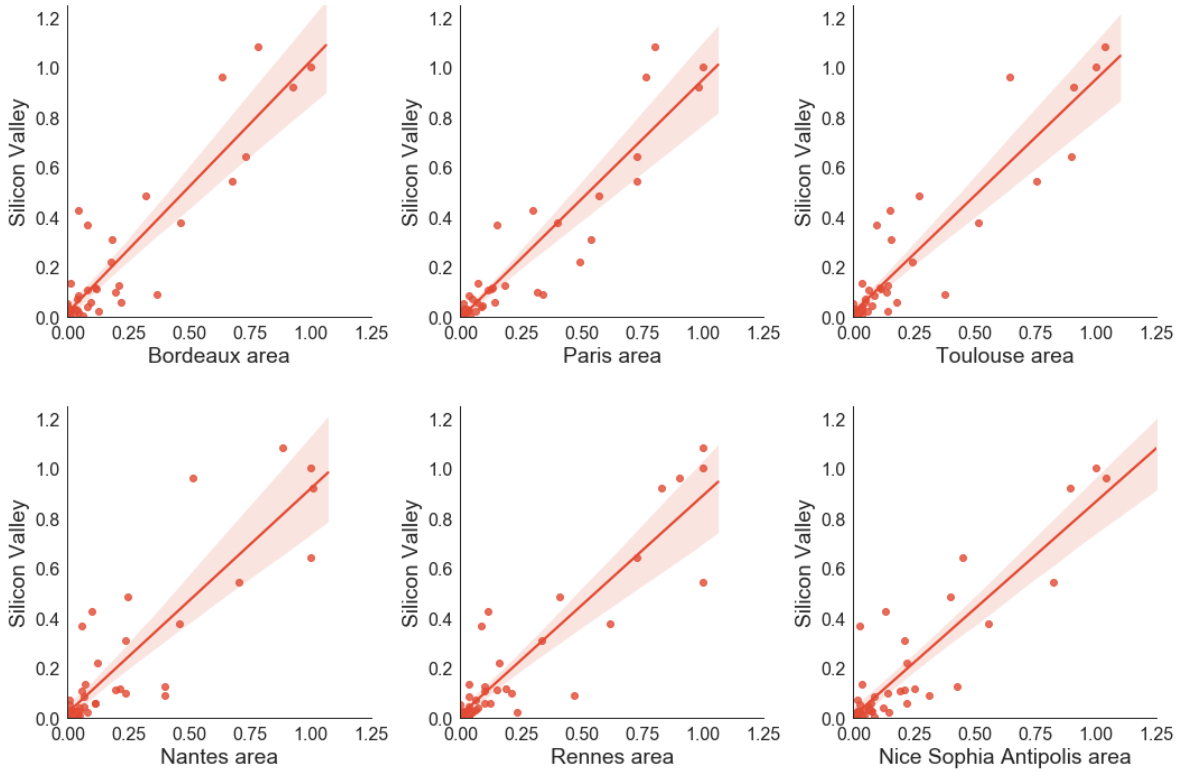

Figure 13: Top 6 startup ecosystems closest to Silicon Valley according to linear regression.

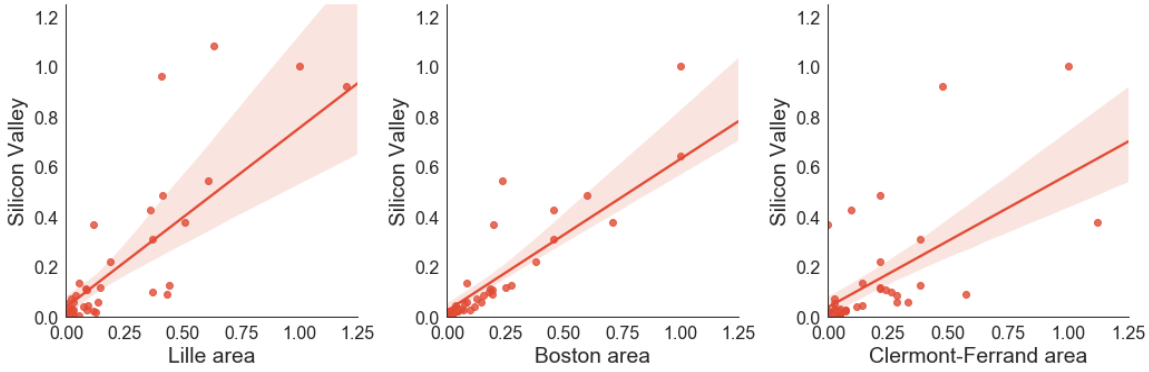

Figure 14: The 3 startup ecosystems that are the most different to Silicon Valley according to linear regression.

4. Results and discussions

4.1. Startup ecosystem comparisons

4.1.1. Silicon Valley is unique

According to our analysis, Silicon Valley sets itself apart from all the other 14 startup ecosystems. By far, Silicon Valley has the largest number of Tech startups (924) according to Foursquare Places API. It is the same in term of overall number of places with 8192 places related to the 46 selected categories, which represents 23% of the total number of places analyzed. These large numbers may be linked to both the high density of Tech startup in this area but also to the fact that people in this area could be more prone to be active user of applications such as those published by Foursquare. It is worth noting that this analysis only used Foursquare Places API and thus that the data are, by definition, biased by the number of local users of the corresponding apps.

Modelling using K-means clustering, both on raw and normalized data, tends to set Silicon Valley on its own cluster alone starting at respectively k = 3 and k = 8 clusters (Figure 5 and Figure 9). This means that Silicon Valley has a specific distribution of number of places in each of the 46 selected categories that is significantly different from the other startup ecosystems.

Linear regression model evaluation leads to similar result as no regression had Pearson correlation coefficient superior to 0.94 (Table 4).

As a result, we would state that according to our analysis, Silicon Valley startup ecosystem remains quite unique in the world and no place is exactly similar, even at a smaller scale.

4.1.2. Paris area as the closest French contender to Silicon Valley

The main objective of this study was to answer the title-question “Where is the French Silicon Valley?” using data from Foursquare Places API.

Based on our findings, the most probable answer to this question would be Paris area. This may seem obvious, but we have to acknowledge that data tends to support this fact in several ways:

- Based on absolute number of Tech startups, Paris with 615 places listed in the Tech startup category is the closest to Silicon Valley and its 924 Tech startup places (Table 2). Even if Paris area comes 2nd with 33% less Tech startups than Silicon Valley, it is still far superior to any other French startup ecosystems.

- Interestingly, K-means clustering on normalized data tends to associate Paris area and Silicon Valley in the same cluster up to k = 7 clusters (Figure 8). This means that Paris would be the closest to Silicon Valley in term of distribution of number of places in each of the 46 selected categories. However, above k = 8 cluster, K-means clustering sets Paris area and Silicon Valley in two separate clusters alone (Figure 9). This means that, while these two areas are close, when increasing the level of clustering, the similarity between Paris and Silicon Valley are not close enough to be kept in the same cluster. This support the fact that Paris startup ecosystem is unique in its own way, like Silicon Valley is.

- According to linear regression, Paris has the best compromise in term of slope and Pearson correlation coefficient (Figure 13). This means that Paris distribution of number of places over the 46 selected categories is the closest to Silicon Valley one. Bordeaux area is the only startup ecosystem with a slope closer to 1 than Paris area but its Pearson correlation coefficient is worse.

These findings lead to our conclusion that, while still significantly different from Silicon Valley, Paris would be the closest French startup ecosystem to Silicon Valley if we had to select any. Of course, this is only based on the analysis of data from Foursquare Places API only. Other features and additional data sources should be considered before awarding "officially" the title of "French Silicon Valley" to any place in the world, especially in France.

4.1.3. Boston (Route 128) vs Silicon Valley

In our analysis, we also looked at Boston area to have another US reference with which to compare the 13 French startup ecosystems we selected. Boston area hosts prestigious universities (MIT, Harvard) and large Tech companies, yet it is not regarded as a startup ecosystem with the same level of successful than its rival on the west coast.

According to our analysis, we can only confirm that Boston area differs significantly from Silicon Valley:

- Boston comes 3rd in term of absolute number of Tech startups, with 50% less places than Silicon Valley and 25% less places than Paris (Table 2). The comparison is even more striking when looking at the normalized ratio of places in the top 4 categories (Financial or Legal Service, Business Service, Coworking Space and Factory) which are very high compared to other startup ecosystems. Actually, if we look at these ratios the other way around, we could argue that it is in fact the number of Tech startups in this area which is unusually low compared to other startup ecosystems.

- K-means clustering modelling both on raw and normalized data, tends to set Boston area on its own cluster alone very early, starting at respectively k = 3 and k = 3 clusters (Figure 5). This means that Boston has a very distinct distribution of number of places compared to other startup ecosystems.

- Linear regression modelling indicates similar results since Boston area has one of the lowest slopes (Figure 14). Noteworthy, its Pearson correlation coefficient is surprisingly high considering how bad the slope is. This may indicate that while Boston data exhibit a more linear behavior than most other startup ecosystems, this behavior is yet significantly different from the one exhibited by Silicon Valley data.

Boston seems different from Silicon Valley, with a surprisingly low number of Tech Startup in comparison to its number of Financial or Legal Services, Business Services, Coworking Spaces and Factories.

4.1.4. Lille, Clermont Ferrand, Strasbourg and Lyon: the industrial heritage

Among the French startup ecosystems, four areas distinguish themselves: Lille, Clermont Ferrand, Strasbourg and Lyon. These areas have surprisingly high Factory/Tech startup ratio, either indicating an unusually high number of Factory or an unusually low number of Tech startup. This observation is supported by the major industrial history of these areas, which were, and still are in some way, the manufacturing arms of France.

While K-means clustering didn't cluster all these 4 areas into one unique cluster, they ended up rapidly either in their own cluster or in the same cluster. This means that these areas really set themselves apart from other startup ecosystems. To the exception of Lyon, all three other areas have the worse slope and worse Pearson correlation coefficient according to linear regression (Figure 14). This means that these areas have strongly different distribution of places compared to Silicon Valley.

4.2. Category selection

4.2.1. Impact of top categories

Among the 46 selected categories of places, the top 10 categories account for 24 786 places which represents 71% of the 34 915 places analyzed. If we look only at the top 5 categories, they account for about 50% of the places. Because of this disproportionate distribution, we have to keep in mind that the top categories may have a strong impact on any modelling we could do.

Category-based normalizing using the number of Tech startup in each startup ecosystem was helpful to compare the various startup ecosystems to Silicon Valley. However, we could also consider in a future work to try another normalizing approach. Area-based normalizing using the number of places in Silicon Valley for each category could be a complementary approach and could help mitigate the potential disproportionate impact of top categories on the remaining categories.

4.2.2. Why Bank and Lawyer categories were not selected

Bank and Lawyer categories were not selected among the 46 categories used for this data analysis. While it may seem counterintuitive to drop these categories as one would obviously argue that these types of places are definitely part of the life of any startup, we had to drop them based on our preliminary data analysis. Indeed, during the first rounds of analysis, we were unable to draw any conclusion. Bank and Lawyer categories have both very large number of places and were in the top 10 categories for many of the 15 startup ecosystems studied here. Actually, for Paris area, Bank was the number 1 category with the most places. Because of the impact of these two categories, K-means clustering wasn't helpful to identify any similarity between the various areas and lead to either cluster of one or cluster of too many.

After dropping these two categories, results were significantly improved. To advance a possible explanation, we could argue that Bank and Lawyer categories are too vast and too common to describe startup ecosystem. Places listed in the Bank category often refers more to local office for individual customer (personal finance) than Finance service specially adapted to startups. The same is true for places listed in the Lawyer category which may very well refer to divorce lawyers, personal injury lawyers, immigration lawyers, etc. In addition, we have to mention that all the finance and legal needs a startup can have are still taking into account within our analysis thanks to the Finance and Legal service category.

4.3. Area selection

4.3.1. Paris-Saclay inclusion

Since Paris area seems to be one of the French contenders to Silicon Valley, we took a closer look at the area included in our analysis. As mentioned in part 2.2. Data acquisition of this present report, we selected up to 50 cities within a 50 km radius from Paris as what we called the "Paris area" or "Paris startup ecosystem". A map of these 50 cities around Paris is shown in Figure 15.

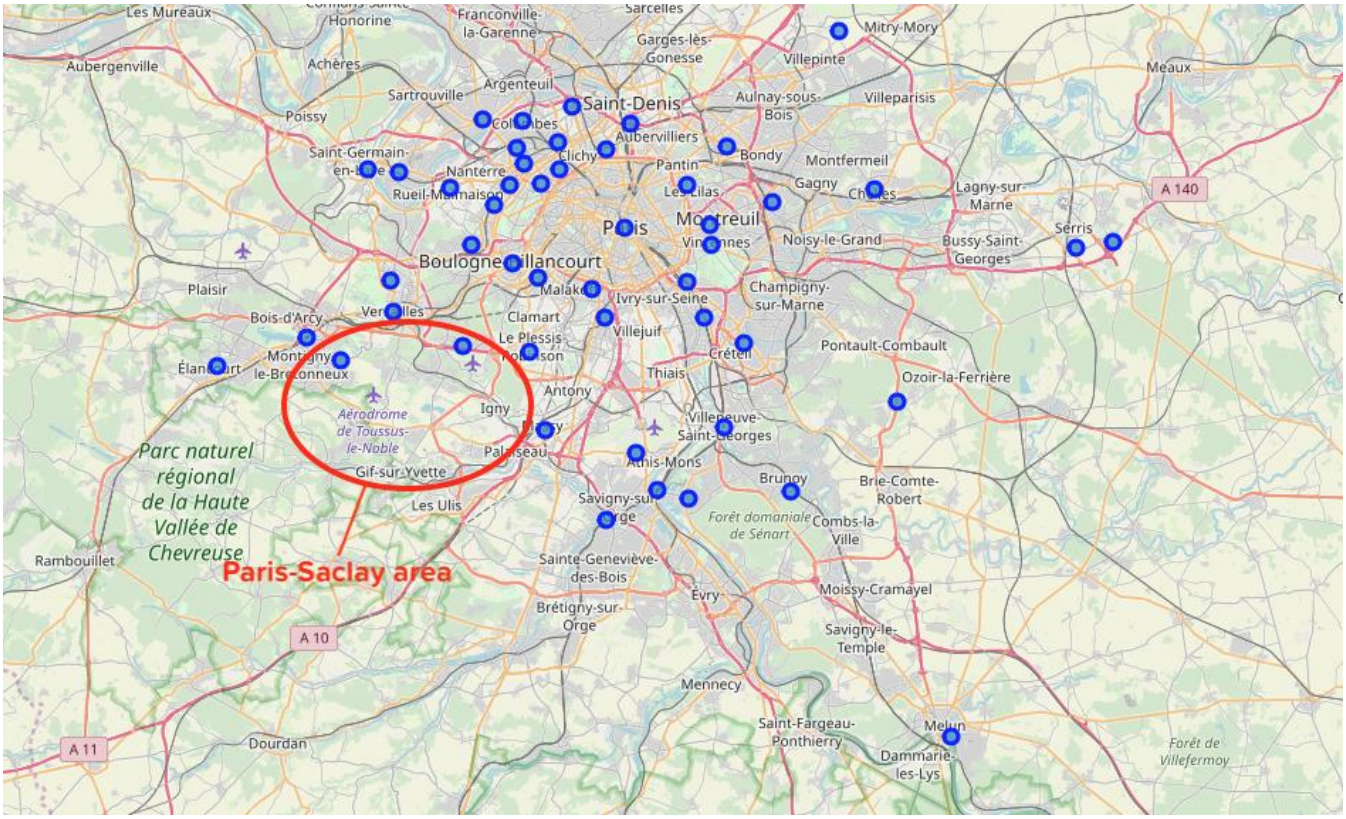

Figure 15: Map of the 50 cities included within "Paris area" (Etampes is missing in the south west). Paris-Saclay area is indicated in red.

French policymaker actively supports the development of the French startup landscape. The opening of Station F in Paris was one of the most iconic projects of 2018.[9,10] Station F ambition is to host up to 1000 startups within the same place and become one of center of the French Tech.

Another project in the south-west of Paris is under development: Paris-Saclay campus.[23] This area already hosts some of the most prestigious engineering schools in France: École polytechnique, CentraleSupélec, Orsay Mathematical Institute, ENSAE, etc.[24] In addition to these impressive schools, several large companies have opened R&D center in this area, accounting to up to 15% of the public and private R&D in France.

As seen in Figure 15, we have included some peripheral cities to Paris-Saclay area but no cities directly within. This may be due to the fact that this area is relatively new and Foursquare Places API doesn't have much data about it

yet. After manual control of Paris area dataset, we made sure that most of the prestigious places listed above were in the dataset.

4.3.2. Lille and Belgium border

Lille is located in the North of France and is very close to Belgium border. As a result, when selecting the top 50 cities within 50 km of Lille, we obviously ended up with Belgian cities. We choose to leave these cities and the corresponding places as they would definitely be part of Lille startup ecosystem even if from another country. Belgium and France are both in the EU and share French as a common language so we could easily see people and startups collaborate within this area, whatever their nationality. As seen in Figure 16, three Belgian cities were included within Lille area: Tournai, Mouscron, Kortrijk

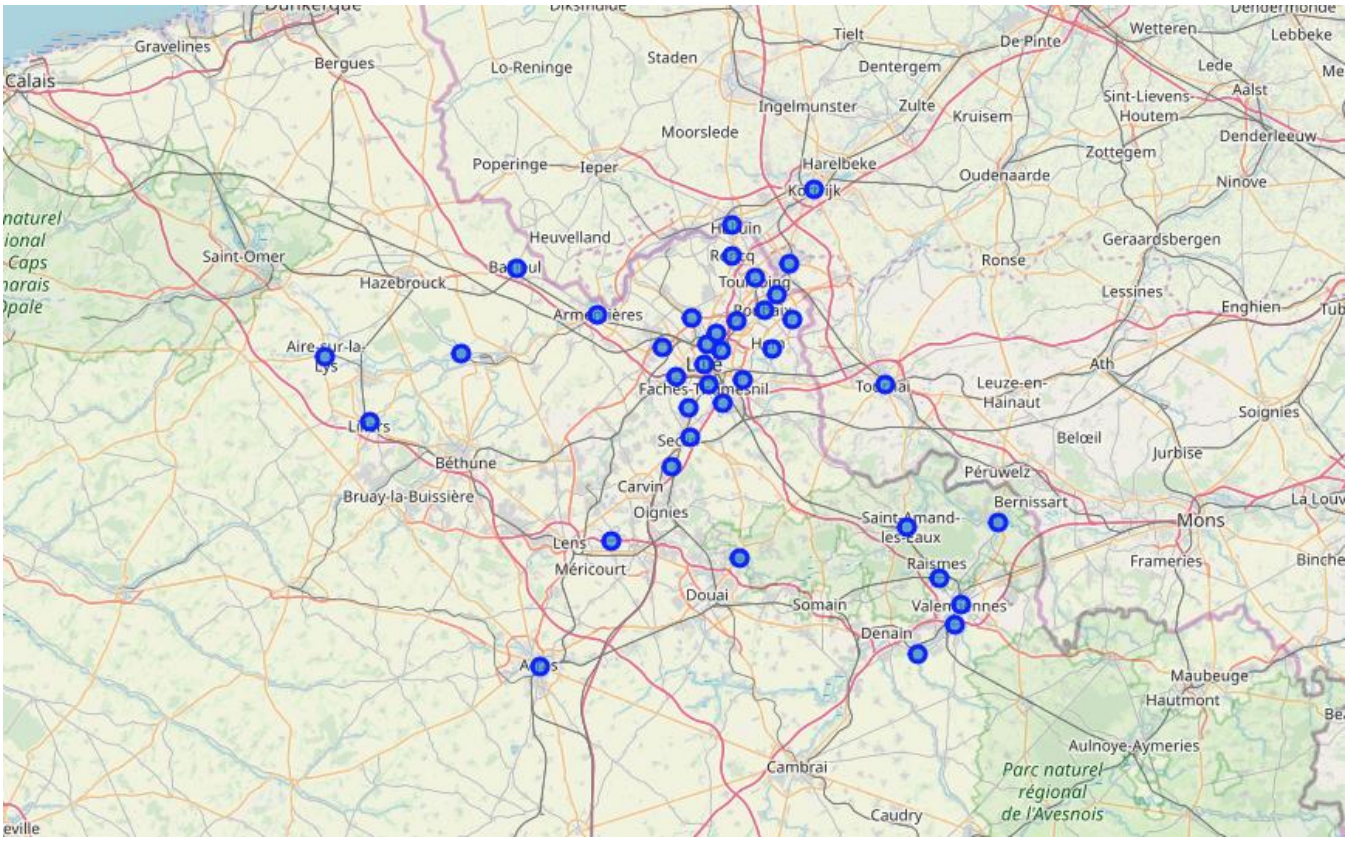

Figure 16: Map of the cities included within "Lille area". Border with Belgium on the east is indicated by the bold purple line

4.4. Foursquare Places dataset

4.4.1. Account limitations

As mentioned in part 2.1 Data source, a free personal Foursquare developer account was used to extract the data from Foursquare Places API. While free, this account limited the scope of our study. For example, the number of

results per search request is limited to 50. This means that in some case such as Silicon Valley and Paris area, the number of cities included in each area was limited to 50 not because there are no more than 50 cities within this area but because of Foursquare API limitations.

We were also limited by the nature of the data we could get from Foursquare Places API. For example, it would have been interesting to have access to places' stats over time, users' comments (called tips), places' description, ratings or like counts.

4.4.2. Early-adopter bias

As mentioned previously, there is an obvious bias in the dataset used in this study. This bias is due to the fact that most of these data were added by users of apps published by Foursquare. As a result, both the quantity and the quality of data in each area is directly linked to the number of active Foursquare users in said area.

For example, it is not surprising that Silicon Valley has the largest number of places in our dataset since this area is both densely populated and also has a larger than usual proportion of early-adopter users compared to other area. We could even argue that this early-adopter bias is an indicator in itself of a lively and efficient startup ecosystem. If true, this could mean that just by evaluating the amount of Foursquare data on a specific place and its completeness compared to other data sources, we would be able to predict how effective a local startup ecosystem is. As noted by other researchers, local early-adopters are key to the success of early stage startup as they represent their first potential customers/users.[3–6] People with such mindset are more likely to try new products or services offered by these startups and thus contribute to their local success, helping them find their market fit and adjust their business model before scaling. Thus, the more users and the more data in Foursquare database about a specific area, the more likely this area would have a large number of users/customers with an early-adopter mindset.

Conclusions

The main objective of this study was to answer the title-question “Where is the French Silicon Valley?” using data from Foursquare Places API. We used machine learning techniques such as K-means clustering and linear regression to answer this question.

Our answer to this question would be that, while still significantly different from Silicon Valley, Paris would be the closest French startup ecosystem to Silicon Valley if we had to select any. Of course, this is only based on the analysis of data from Foursquare Places API only and should be confirmed by further analysis.

The on-going Paris-Saclay development project in Paris south-west could be a push forward that may very well help the French capital close the gap with Silicon Valley.

Boston seems different from Silicon Valley, with a surprisingly low number of Tech Startup in comparison to its number of Financial or Legal Services, Business Services, Coworking Spaces and Factories.

Among the French startup ecosystems, we could see that four areas distinguished themselves: Lille, Clermont Ferrand, Strasbourg and Lyon. These areas have surprisingly high Factory/Tech Startup ratio, either indicating an unusually high number of Factory or an unusually low number of Tech startup. This observation is supported by the major industrial history of these areas, which were, and still are in some way, the manufacturing arms of France.

Limitations based on the type and quantity of data that we were able to extract from Foursquare Places API prevent any definitive conclusions to be drawn from this analysis. Further analysis and cross-validation with additional features and other data sources would be required before awarding “officially” the title of “French Silicon Valley” to any place in the world, especially in France.

With a larger access to Foursquare data, it could be very interesting to look for example at the time series data related to the evolution of the various startup ecosystems over the last 10 years.

Bibliography

[1] "Comparison between U.S. states and sovereign states by GDP," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Comparison_between_U.S._states_and_sovereign_states_by_GDP. [Accessed 29 03 2019].

[2] L. M. SEGARRA, "California's Economy Is Now Bigger Than All of the U.K.," Fortune, 5 May 2018. [Online]. Available: http://fortune.com/2018/05/05/california-fifth-biggest-economy-passes-united-kingdom/. [Accessed 29 03 2019].

[3] J. Fox, "What Still Makes Silicon Valley So Special," 05 12 2014. [Online]. Available: https://hbr.org/2014/12/what-still-makes-silicon-valley-so-special. [Accessed 29 03 2019].

[4] A. Saxenian, «Silicon Valley Versus Route 128,» 1 02 1994. [En ligne]. Available: https://www.inc.com/magazine/19940201/2758.html . [Accès le 29 03 2019].

[5] K. AMADEO, «Silicon Valley, America's Innovative Advantage,» 20 11 2018. [En ligne]. Available: https://www.thebalance.com/what-is-silicon-valley-3305808. [Accès le 30 03 2019].

[6] A. Crompton, "What's Special About Silicon Valley?," 22 02 2018. [Online]. Available: https://medium.com/swlh/silicon-valley-e2a3cb55f1ee. [Accessed 30 03 2019].

[7] "Startup ecosystem," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Startup_ecosystem.

[8] R. Empson, " Startup Genome Ranks The World's Top Startup Ecosystems: Silicon Valley, Tel Aviv & L.A. Lead The Way," Techcrunch, 2012. [Online]. Available: https://techcrunch.com/2012/11/20/startup-genome-ranks-the-worlds-top-startup-ecosystems-silicon-valley-tel-aviv-l-a-lead-the-way/. [Accessed 30 03 2019].

[9] L. Alderman, "Macron Vowed to Make France a 'Start-Up Nation.' Is It Getting There?," The New York Times, 23 05 2018. [Online]. Available: https://www.nytimes.com/2018/05/23/business/emmanuel-macron-france-technology.html. [Accessed 30 03 2019].

[10] L. Alderman, B. Morenne, E. Peltier "Why France Is Taking a Lesson in Culture From Silicon Valley," The New York Times, 29 06 2017. [Online]. Available: https://www.nytimes.com/2017/06/29/business/station-f-tech-incubator-france.html. [Accessed 30 03 2019].

[11] FrenchTech, [Online]. Available: https://www.lafrenchtech.com/en. [Accessed 30 03 2019].

[12] "Bpifrance, la banque des entrepreneurs," Bpifrance, [Online]. Available: https://www.bpifrance.fr. [Accessed 30 03 2019].

[13] "Crédit d'impôt recherche (CIR)," Service Public Pro, 21 01 2019. [Online]. Available: https://www.service-public.fr/professionnels-entreprises/vosdroits/F23533. [Accessed 30 03 2019].

[14] "Foursquare (company)," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Foursquare_(company). [Accessed 30 03 2019].

[15] J. Glueck, "Foursquare Predicts Chipotle's Q1 Sales Down Nearly 30%; Foot Traffic Reveals the Start of a Mixed Recovery," Medium, 12 04 2016. [Online]. Available: https://medium.com/foursquare-direct/foursquare-predicts-chipotle-s-q1-sales-down-nearly-30-foot-traffic-reveals-the-start-of-a-mixed-78515b2389af. [Accessed 30 03 2019].

[16] "Places API," Foursquare, [Online]. Available: https://developer.foursquare.com/places-api. [Accessed 30 03 2019].

[17] N. Leon, "Où est la Silicon Valley française ? Cap au Nord," Welcome to the Jungle, 28 06 2018. [Online]. Available: https://www.welcometothejungle.co/articles/silicon-valley-francaise-nord. [Accessed 30 03 2019].

[18] N. Leon, "Où est la Silicon Valley française ? Cap au Sud," Welcome to the Jungle, 4 07 2018. [Online]. Available: https://www.welcometothejungle.co/articles/tech-silicon-valley-francaise-province-sud. [Accessed 30 03 2019].

[19] "French Tech," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/French_Tech. [Accessed 30 03 2019].

[20] M. Sonnati, "where_is_the_french_silicon_valley," Github, [Online]. Available: https://github.com/MattSonnati/where_is_the_french_silicon_valley. [Accessed 30 03 2019].

[21] "Venue Categories," Foursquare, [Online]. Available: https://developer.foursquare.com/docs/resources/categories. [Accessed 30 03 2019].

[22] "Cross-industry standard process for data mining," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining. [Accessed 30 03 2019].

[23] Paris Saclay Startup, [Online]. Available: https://www.paris-saclay-startup.com/#/. [Accessed 31 03 2019].

[24] M. L. TUTOUR, Capital, 15 11 2018. [Online]. Available: https://www.capital.fr/votre-carriere/paris-saclay-la-silicon-valley-a-la-francaise-1315895. [Accessed 30 03 2019].