Capstone Project Coursera

BRALIC MADRID Goran

Introduction

Last February, Forbes published their annual NBA franchises valuations. And like the past few years, the Knicks, the New York franchise owned by James Dolan, were at the top with an estimated value of 4.6 billion dollars. As an NBA fan, it was weird to see that the franchise who can be considered as one of the worst, with the lowest winning percentage for several years, was in fact the richest one. In the other hand, the Toronto Raptors (the actual NBA champion) and the Milwaukee Bucks (the team with the most victories in the 2018-2019 NBA season) were ranked 10th and 19th with respectively estimated values of “only” 2.1 billion and 1.58 billion dollars. Of course, the value of a franchise is not only linked to the sports results, and of course in the past the Knicks were a good team, but to have such a gap between the bests actuals teams and the worst one got me curious.

I live in Europe and I’ve never been to the USA, but when I think about New York, the first to come to my mind is the cultural part. NY is a huge beautiful and rich city with a lot of things to do. Monuments, museums, parks, restaurants and all other venues. But when we talk about Milwaukee, nothing comes to my mind. Ranked by its estimated population, Milwaukee is 31st largest city in the USA, which is not bad, but it isn’t New York. These two cities don’t have the same recognition from the USA and from the world. And this may be why the values are so different.

In comparison with European soccer, best teams are the one with the most estimated valuate. But in most cases these teams are all in big and rich cities as capitals and are big institutions who are historically always been good teams. Moreover, in European soccer teams aren’t franchises like in the NBA, and all the rules and leagues operations are different.

The main purpose of this assignment is to find a correlation between franchise values and the socio-economic data of a city. If there is a correlation maybe I will try to determine which cities have a good potential to be included in the NBA in case of expansion.

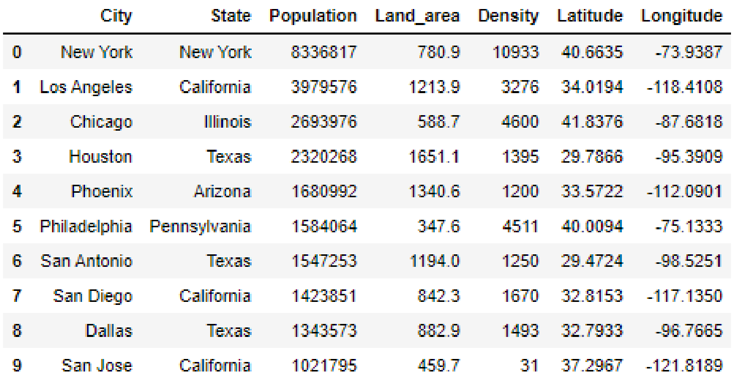
Data needed

First, for this project, I need the NBA team’s names, and the estimated value of each franchise. I will use Foursquare to have information about the venues of a city (I will use the number of venues in each city). For this I need to have the coordinates of each NBA city. The other information will be geographical, social and economic information of cities, like the area, the estimated population, the mean income, and other. Of course, I will also use sport information like the results or the number of championships.

All this data can be found on the internet. The most part will come from web scrapping Wikipedia pages, but other information’s can be directly download from some sites.

Data loading and data processing

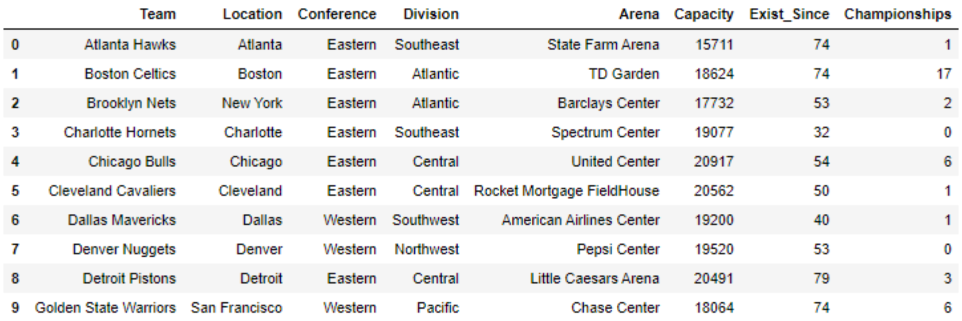
First, we're going to get information about the 150 biggest USA cities. I web scrapped a Wikipedia page to get the fallowing information about cities: Name of the city, State, population, land area, density of population and the coordinates.
https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population



	City	State	Population	Land_area	Density	Latitude	Longitude
0	New York	New York	8336817	780.9	10933	40.6635	-73.9387
1	Los Angeles	California	3979576	1213.9	3276	34.0194	-118.4108
2	Chicago	Illinois	2693976	588.7	4600	41.8376	-87.6818
3	Houston	Texas	2320268	1651.1	1395	29.7866	-95.3909
4	Phoenix	Arizona	1680992	1340.6	1200	33.5722	-112.0901
5	Philadelphia	Pennsylvania	1584064	347.6	4511	40.0094	-75.1333
6	San Antonio	Texas	1547253	1194.0	1250	29.4724	-98.5251
7	San Diego	California	1423851	842.3	1670	32.8153	-117.1350
8	Dallas	Texas	1343573	882.9	1493	32.7933	-96.7665
9	San Jose	California	1021795	459.7	31	37.2967	-121.8189

Now that we have the data about USA cities, were going to get data about NBA cities. Like with the first dataset, I web scrapped a Wikipedia page to get the information. The important attributes that I to collect are: Team name, location, conference, division, name of the arena, arena capacity and the creation date. I manually add a column with the championships.

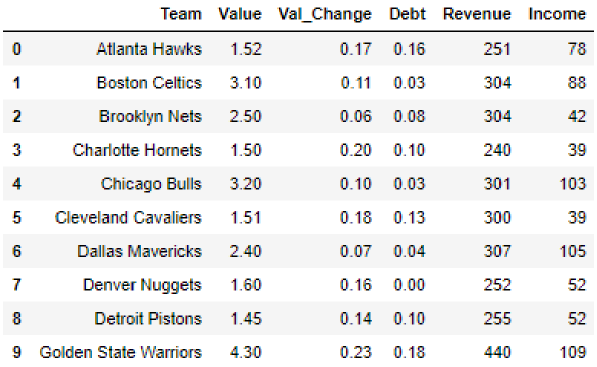
https://en.wikipedia.org/wiki/National_Basketball_Association



	Team	Location	Conference	Division	Arena	Capacity	Exist_Since	Championships
0	Atlanta Hawks	Atlanta	Eastern	Southeast	State Farm Arena	15711	74	1
1	Boston Celtics	Boston	Eastern	Atlantic	TD Garden	18624	74	17
2	Brooklyn Nets	New York	Eastern	Atlantic	Barclays Center	17732	53	2
3	Charlotte Hornets	Charlotte	Eastern	Southeast	Spectrum Center	19077	32	0
4	Chicago Bulls	Chicago	Eastern	Central	United Center	20917	54	6
5	Cleveland Cavaliers	Cleveland	Eastern	Central	Rocket Mortgage FieldHouse	20562	50	1
6	Dallas Mavericks	Dallas	Western	Southwest	American Airlines Center	19200	40	1
7	Denver Nuggets	Denver	Western	Northwest	Pepsi Center	19520	53	0
8	Detroit Pistons	Detroit	Eastern	Central	Little Caesars Arena	20491	79	3
9	Golden State Warriors	San Francisco	Western	Pacific	Chase Center	18064	74	6

For the data about franchise value, I manually scrapped a Forbes page, and I obtained the fallowing information: Franchise Value, Value Change from precedent year, debt value, revenue and the operating incomes (Earnings before interest, taxes, depreciation and amortization).

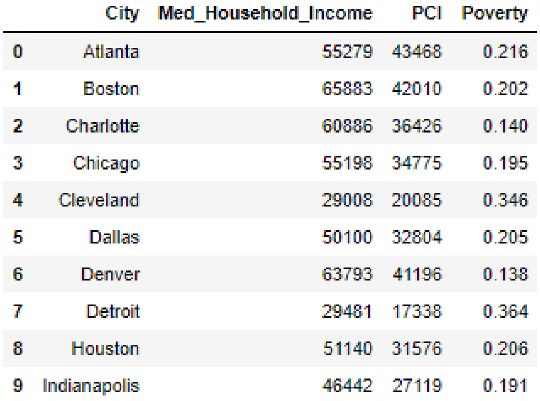
https://www.forbes.com/nba-valuations/list



	Team	Value	Val_Change	Debt	Revenue	Income
0	Atlanta Hawks	1.52	0.17	0.16	251	78
1	Boston Celtics	3.10	0.11	0.03	304	88
2	Brooklyn Nets	2.50	0.06	0.08	304	42
3	Charlotte Hornets	1.50	0.20	0.10	240	39
4	Chicago Bulls	3.20	0.10	0.03	301	103
5	Cleveland Cavaliers	1.51	0.18	0.13	300	39
6	Dallas Mavericks	2.40	0.07	0.04	307	105
7	Denver Nuggets	1.60	0.16	0.00	252	52
8	Detroit Pistons	1.45	0.14	0.10	255	52
9	Golden State Warriors	4.30	0.23	0.18	440	109

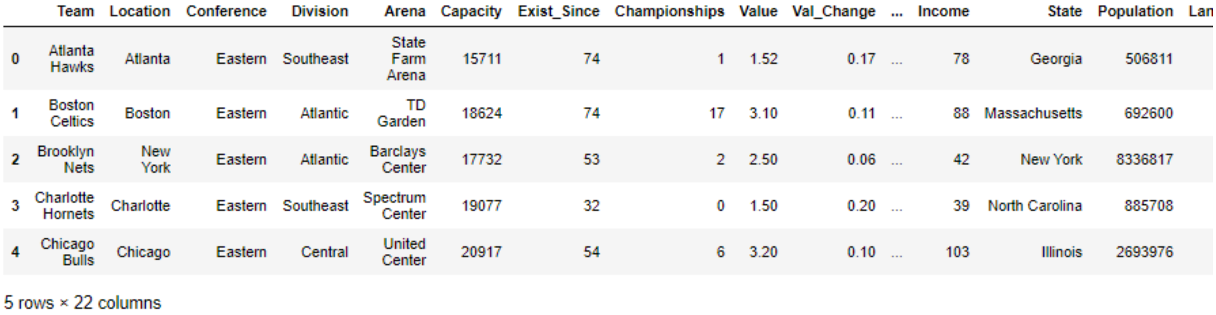
The last data was retrieved in US Census site. The important features are Per capita income in past 12 months (in 2018 dollars), Median household income (in 2018 dollars) and the percentage of persons in poverty. I manually add Toronto because it's not an USA city (Canada) and so was not in the site.

https://www.census.gov/quickfacts/fact/table/US/PST045219



	City	Med_Household_Income	PCI	Poverty
0	Atlanta	55279	43468	0.216
1	Boston	65883	42010	0.202
2	Charlotte	60886	36426	0.140
3	Chicago	55198	34775	0.195
4	Cleveland	29008	20085	0.346
5	Dallas	50100	32804	0.205
6	Denver	63793	41196	0.138
7	Detroit	29481	17338	0.364
8	Houston	51140	31576	0.206
9	Indianapolis	46442	27119	0.191

Now that I have all the data from the web, I can merge all the data frames to have a good and complete one.



	Team	Location	Conference	Division	Arena	Capacity	Exist_Since	Championships	Value	Val_Change	...	Income	State	Population	Lan
0	Atlanta Hawks	Atlanta	Eastern	Southeast	State Farm Arena	15711	74	1	1.52	0.17	...	78	Georgia	506811	
1	Boston Celtics	Boston	Eastern	Atlantic	TD Garden	18624	74	17	3.10	0.11	...	88	Massachusetts	692600	
2	Brooklyn Nets	New York	Eastern	Atlantic	Barclays Center	17732	53	2	2.50	0.06	...	42	New York	8336817	
3	Charlotte Hornets	Charlotte	Eastern	Southeast	Spectrum Center	19077	32	0	1.50	0.20	...	39	North Carolina	885708	
4	Chicago Bulls	Chicago	Eastern	Central	United Center	20917	54	6	3.20	0.10	...	103	Illinois	2693976	

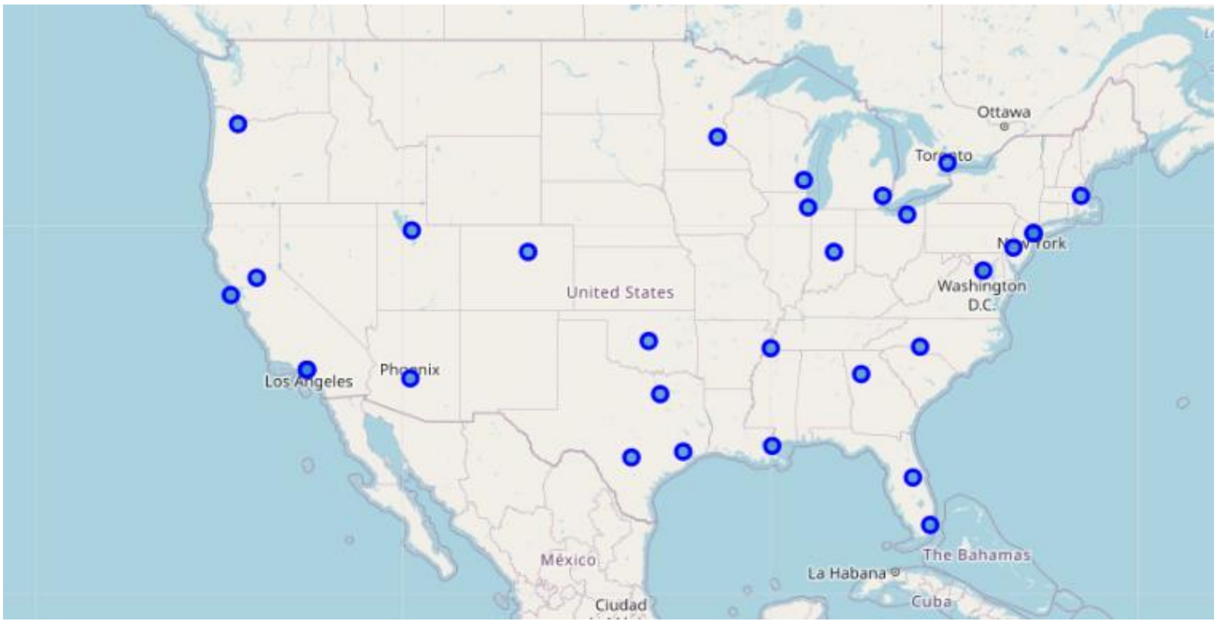
5 rows × 22 columns

Now that the data frame is complete, we can make our Foursquare request in order to get more information.

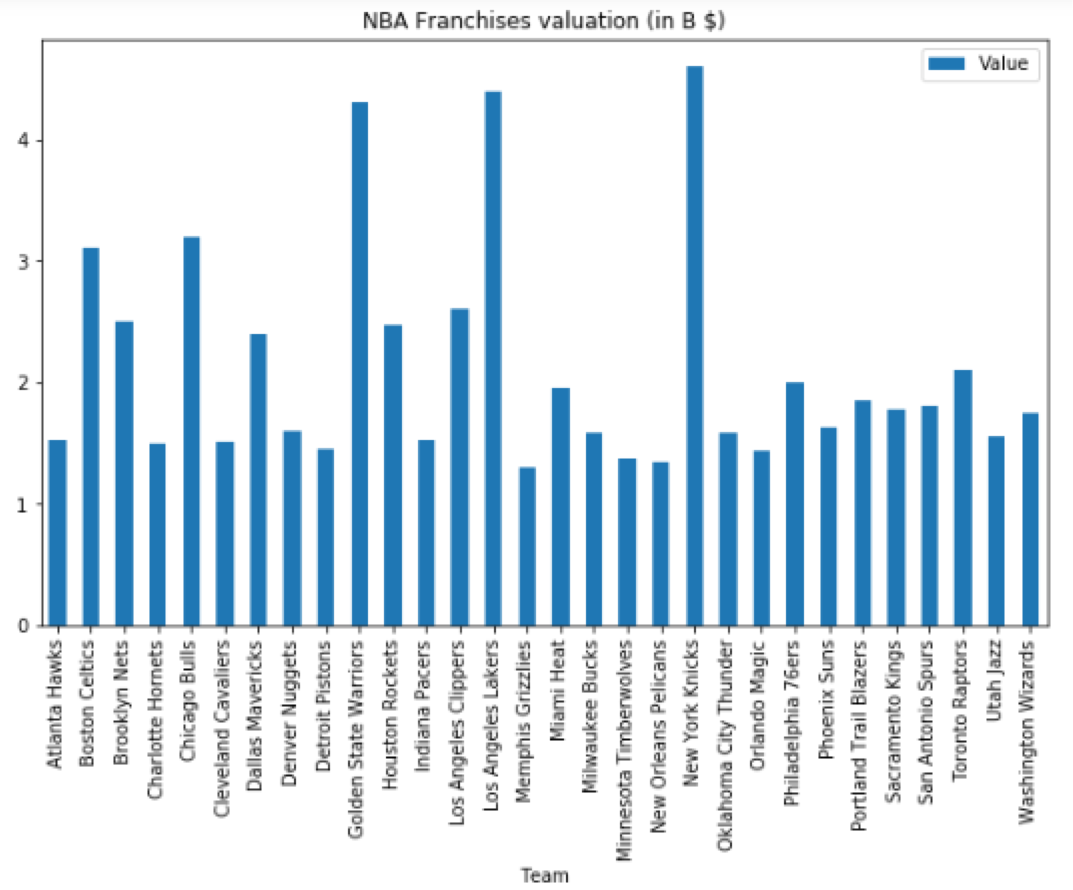
I just realize that Foursquare is not adapted to my project. Indeed, Foursquare only return 50 results per query. This project is based on big cities and Foursquare is more adapted to smaller places like neighborhoods or streets. I decide to not use Foursquare.

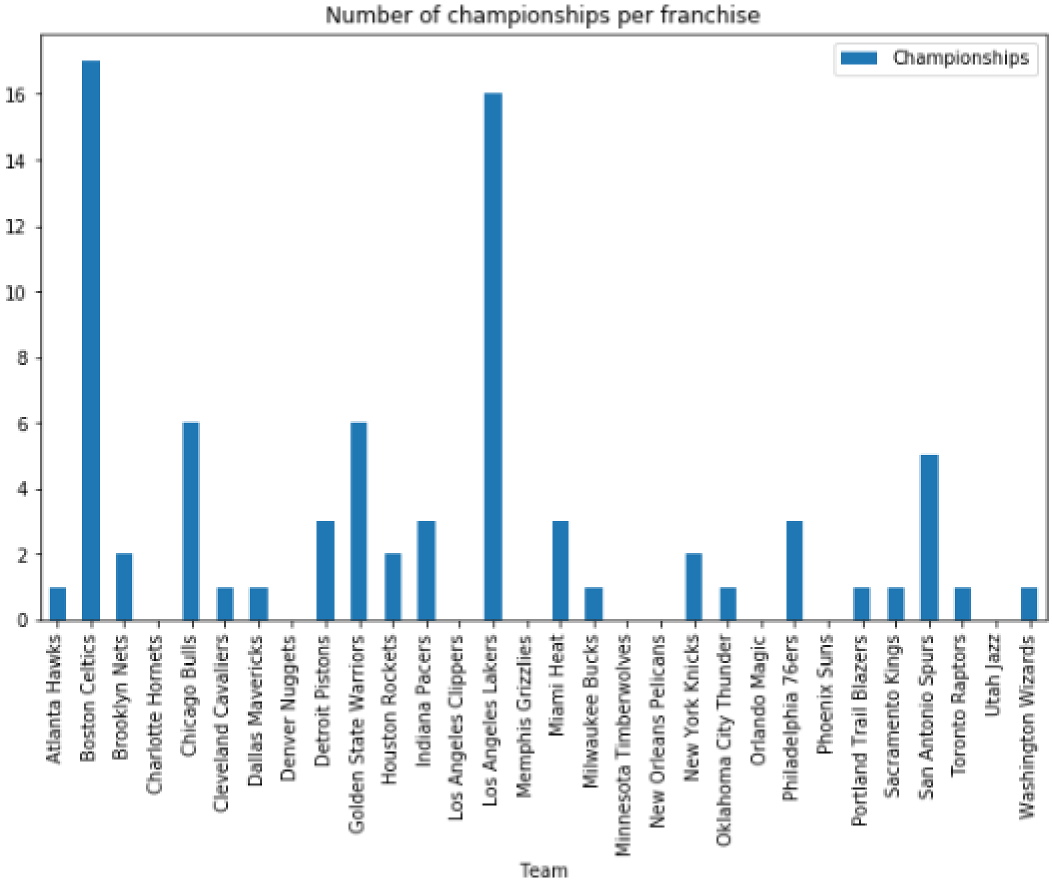
Data Visualization

First let's have a geographical representation of NBA Cities using Folium.

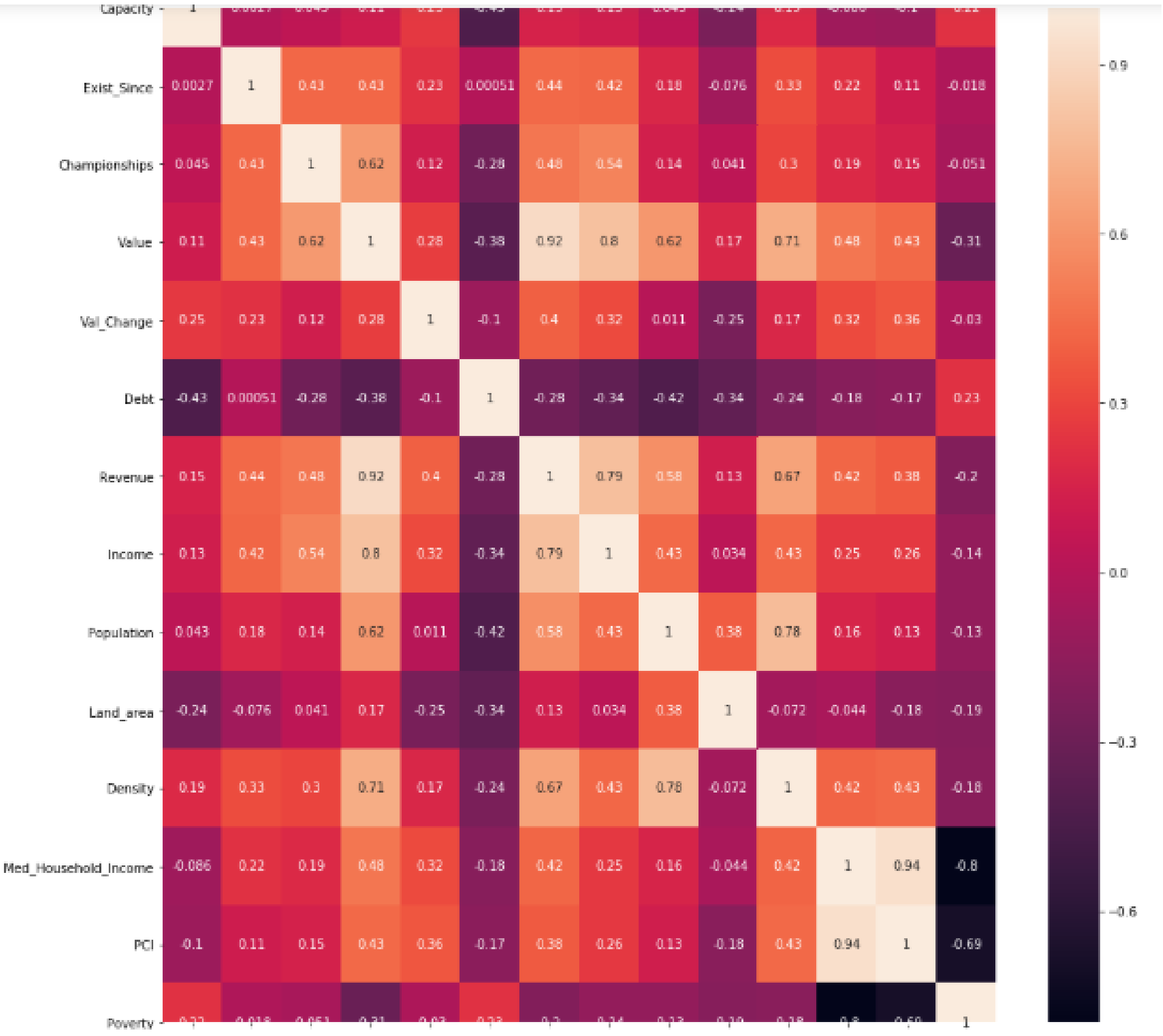


I did two bar plots to represent franchise valuation and franchises championship. These two plots didn't really seem to be strongly correlated. They might be a small correlation, because teams as Los Angeles Lakers, Boston Celtics, Chicago Bulls and Golden State Warriors are the most successful teams, and they are also between the most valuated ones.

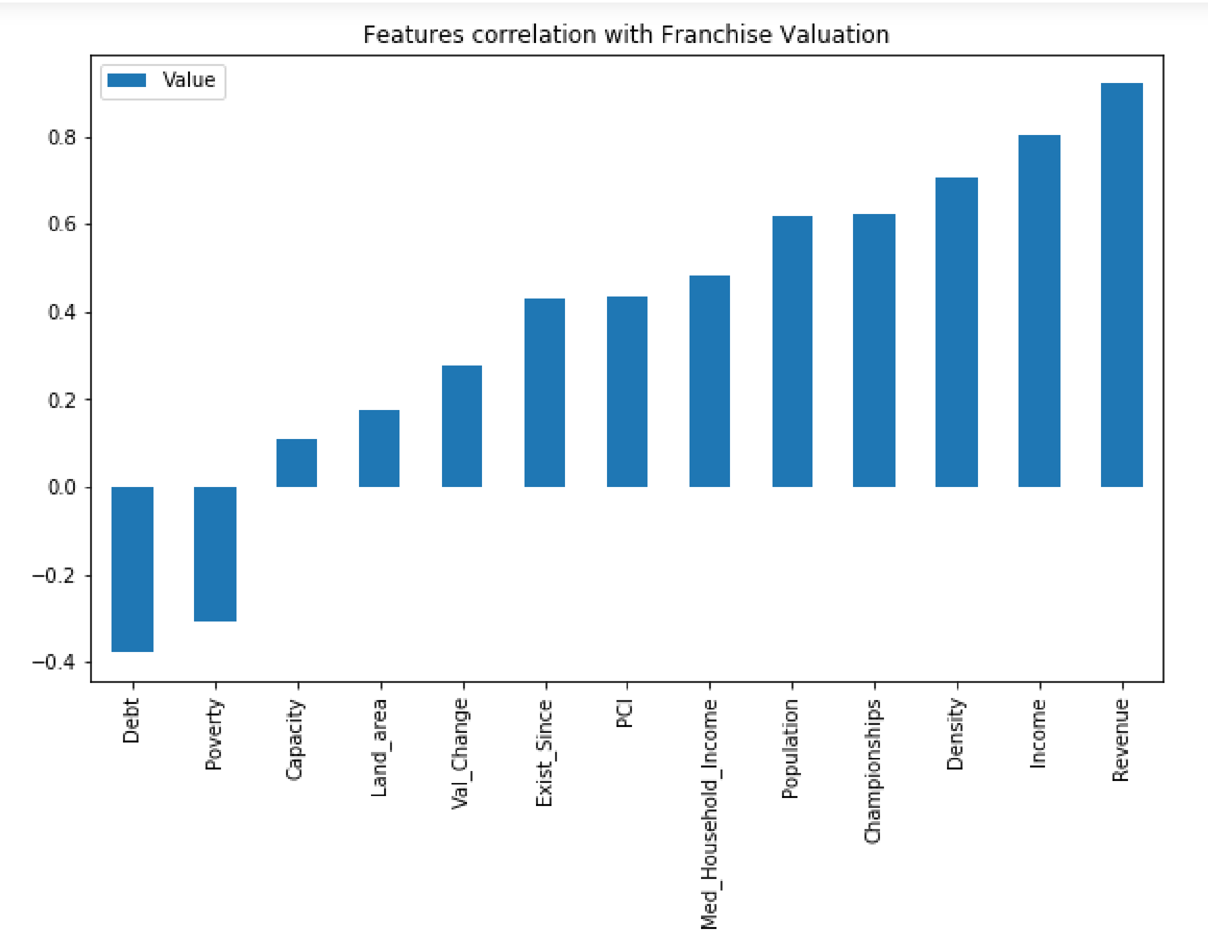




Let's try to find which features are the most correlated to Value. To do so, I made a correlation matrix, and a bar plot to show the most correlated features.



Capacity	1	0.0027	0.045	0.11	0.25	-0.43	0.15	0.13	0.043	-0.24	0.19	-0.086	-0.1	0.22
Exist_Since	0.0027	1	0.43	0.43	0.23	0.00051	0.44	0.42	0.18	-0.076	0.33	0.22	0.11	-0.018
Championships	0.045	0.43	1	0.62	0.12	-0.28	0.48	0.54	0.14	0.041	0.3	0.19	0.15	-0.051
Value	0.11	0.43	0.62	1	0.28	-0.38	0.92	0.8	0.62	0.17	0.71	0.48	0.43	-0.31
Val_Change	0.25	0.23	0.12	0.28	1	-0.1	0.4	0.32	0.011	-0.25	0.17	0.32	0.36	-0.03
Debt	-0.43	0.00051	-0.28	-0.38	-0.1	1	-0.28	-0.34	-0.42	-0.34	-0.24	-0.18	-0.17	0.23
Revenue	0.15	0.44	0.48	0.92	0.4	-0.28	1	0.79	0.58	0.13	0.67	0.42	0.38	-0.2
Income	0.13	0.42	0.54	0.8	0.32	-0.34	0.79	1	0.43	0.034	0.43	0.25	0.26	-0.14
Population	0.043	0.18	0.14	0.62	0.011	-0.42	0.58	0.43	1	0.38	0.78	0.16	0.13	-0.13
Land_area	-0.24	-0.076	0.041	0.17	-0.25	-0.34	0.13	0.034	0.38	1	-0.072	-0.044	-0.18	-0.19
Density	0.19	0.33	0.3	0.71	0.17	-0.24	0.67	0.43	0.78	-0.072	1	0.42	0.43	-0.18
Med_Household_Income	-0.086	0.22	0.19	0.48	0.32	-0.18	0.42	0.25	0.16	-0.044	0.42	1	0.94	-0.8
PCI	-0.1	0.11	0.15	0.43	0.36	-0.17	0.38	0.26	0.13	-0.18	0.43	0.94	1	-0.69
Poverty	0.22	-0.018	-0.051	-0.31	-0.03	0.23	-0.2	-0.14	-0.13	-0.19	-0.18	-0.8	-0.69	1



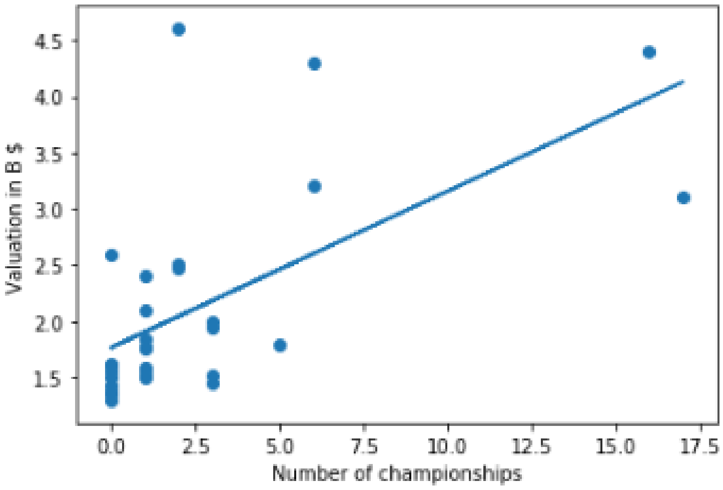
With this plot we can see that the most important features are franchise revenue, the income, the city density (linked to population) and championships. The other features cannot be considered as correlated.

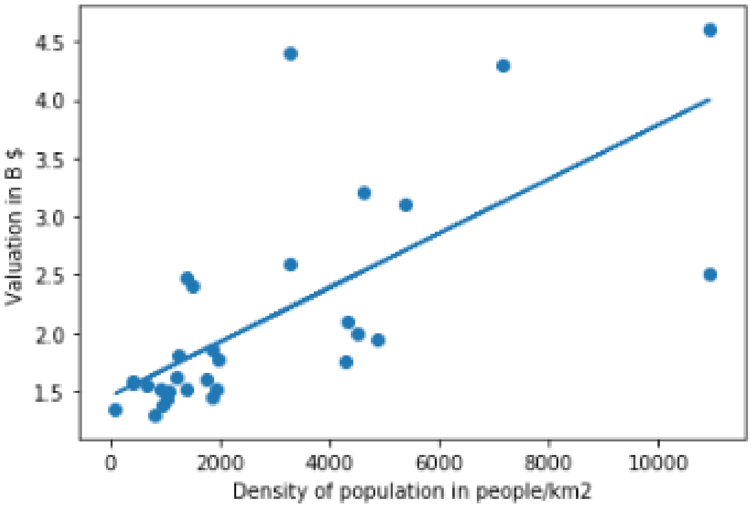
This tell us that franchise value not only depends on the economic part, and franchises can't be fully compared to enterprises. The sport and the geo-social side are also very important in the valuation.

Data analysis

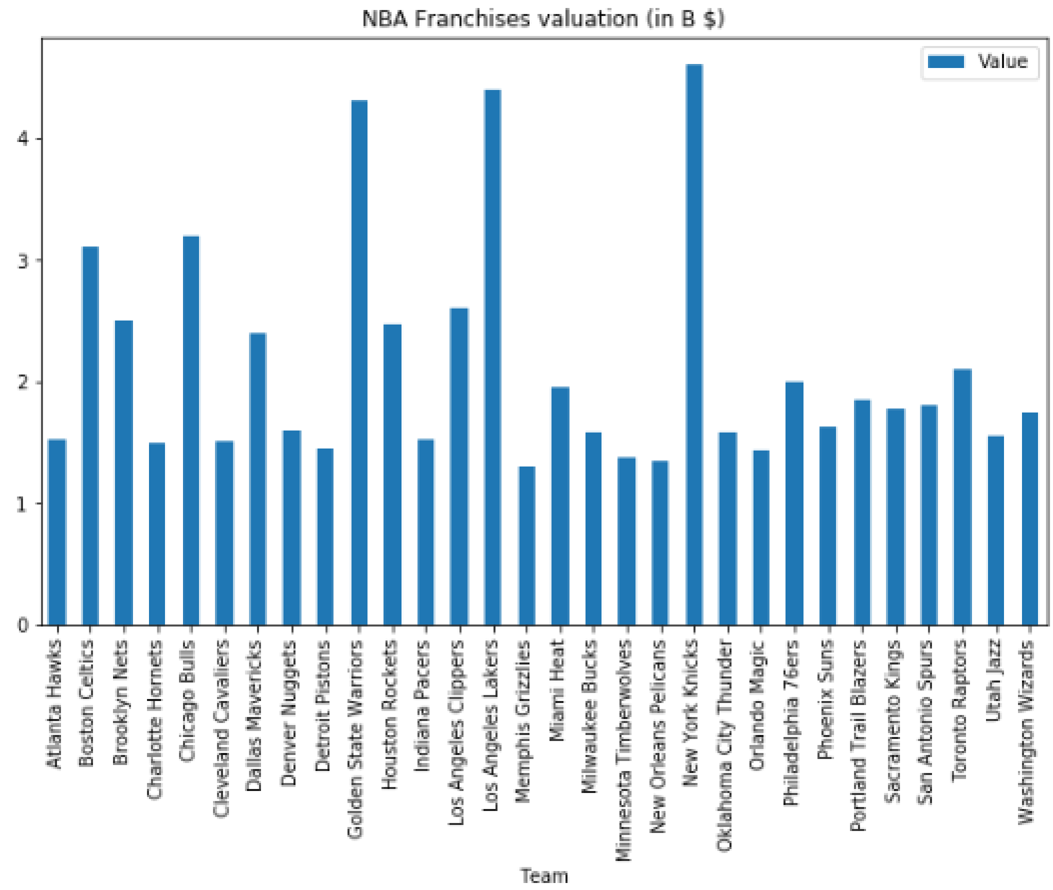
Now that we have all the considered relevant data, we can explore it more deeply.

Because the amount of data is too low it is difficult to create some model. Let's start by some regression with the features founded before.



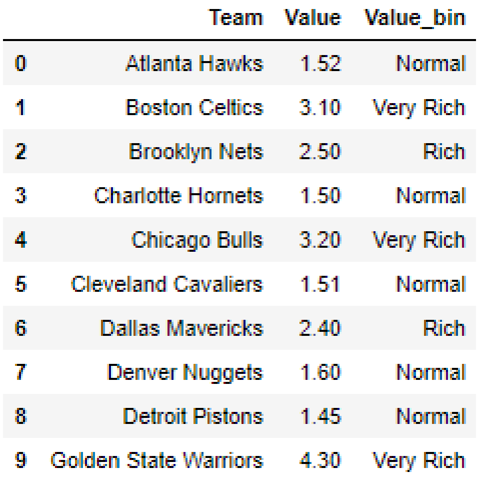


These plots show the variation and relationship between franchise valuation to Championships and Density of population. Regressions are difficult in here because we don't have a lot of data. But we clearly can see that the most successful a team is the greater the valuation. Also, if the city density of population grows, the valuation of that franchise will also grow.



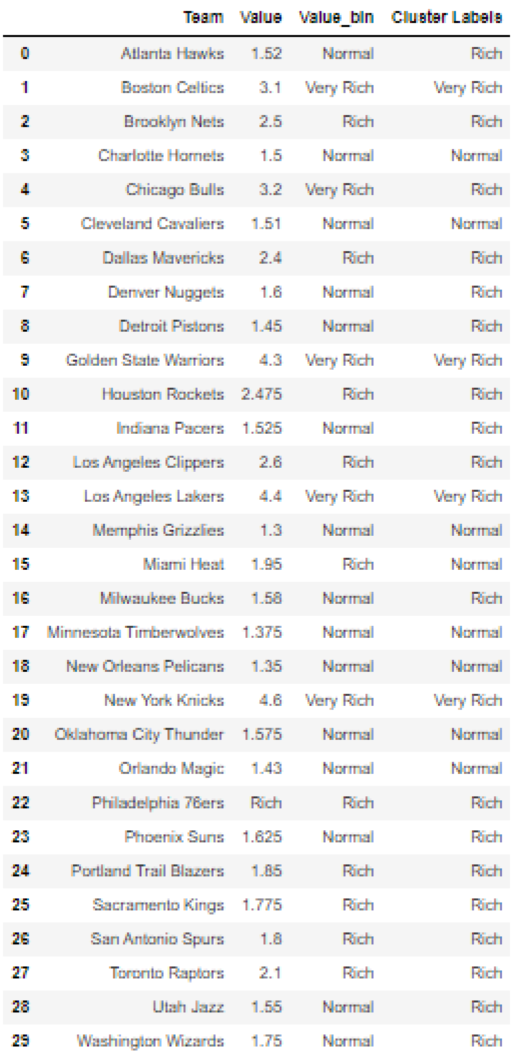
In this bar plot, we can see that there are 3 categories of team: Very Rich, Rich, and Normal. The very rich franchises can be considered as the franchises with a value higher than 3B dollar, the rich with a value between 1.75B and 3B and the normal with a value lower than 1.75B dollar.

We obtain the fallowing separation doing a binning.



	Team	Value	Value_bin
0	Atlanta Hawks	1.52	Normal
1	Boston Celtics	3.10	Very Rich
2	Brooklyn Nets	2.50	Rich
3	Charlotte Hornets	1.50	Normal
4	Chicago Bulls	3.20	Very Rich
5	Cleveland Cavaliers	1.51	Normal
6	Dallas Mavericks	2.40	Rich
7	Denver Nuggets	1.60	Normal
8	Detroit Pistons	1.45	Normal
9	Golden State Warriors	4.30	Very Rich

Now let's try to make some clustering and with the important features and see if the groups that we just create are clustered together. The features considered in this clustering are: Championships, Longevity of the team, Revenue of the franchise, Income of the franchise, the city population, the city density of population and the median household income in the city.



	Team	Value	Value_bin	Cluster Labels
0	Atlanta Hawks	1.52	Normal	Rich
1	Boston Celtics	3.1	Very Rich	Very Rich
2	Brooklyn Nets	2.5	Rich	Rich
3	Charlotte Hornets	1.5	Normal	Normal
4	Chicago Bulls	3.2	Very Rich	Rich
5	Cleveland Cavaliers	1.51	Normal	Normal
6	Dallas Mavericks	2.4	Rich	Rich
7	Denver Nuggets	1.6	Normal	Rich
8	Detroit Pistons	1.45	Normal	Rich
9	Golden State Warriors	4.3	Very Rich	Very Rich
10	Houston Rockets	2.475	Rich	Rich
11	Indiana Pacers	1.525	Normal	Rich
12	Los Angeles Clippers	2.6	Rich	Rich
13	Los Angeles Lakers	4.4	Very Rich	Very Rich
14	Memphis Grizzlies	1.3	Normal	Normal
15	Miami Heat	1.95	Rich	Normal
16	Milwaukee Bucks	1.58	Normal	Rich
17	Minnesota Timberwolves	1.375	Normal	Normal
18	New Orleans Pelicans	1.35	Normal	Normal
19	New York Knicks	4.6	Very Rich	Very Rich
20	Oklahoma City Thunder	1.575	Normal	Normal
21	Orlando Magic	1.43	Normal	Normal
22	Philadelphia 76ers	Rich	Rich	Rich
23	Phoenix Suns	1.625	Normal	Rich
24	Portland Trail Blazers	1.85	Rich	Rich
25	Sacramento Kings	1.775	Rich	Rich
26	San Antonio Spurs	1.8	Rich	Rich
27	Toronto Raptors	2.1	Rich	Rich
28	Utah Jazz	1.55	Normal	Rich
29	Washington Wizards	1.75	Normal	Rich

This clustering is effective, the clusters seems to respect the binning made before.