Capstone Project 1: Final Report

Distribution of doctors in France

Problem Statement

In France, the distribution of healthcare professional on the whole territory is problematic. Indeed, people living in villages far from big cities complain about the lack of doctors and the very long delay to have a medical appointment.

The purpose of this project is to construct a model with machine learning to predict the distribution of healthcare professional in France considering the municipality, the region, the population of municipalities, the specialty of doctors….

Potential clients of this project could be the French State which manages the public hospitals, private groups that manage private clinics or healthcare professionals who want to settle down and open a practice. This would allow them to learn about areas that lack healthcare professional to know where to locate potential future medical infrastructure.

For that, I will use two sets of data found on the website https://public.opendatasoft.com et the website https://data.opendatasoft.com.

First Dataset: This dataset (149 477 x 19 cells) displays healthcare professional, their location and coordinates, the nature of their activity, the technical acts that they perform in France.
https://public.opendatasoft.com/explore/dataset/annuaire-des-professionnels-de-sante/export/

Second Dataset: This dataset (39724 x 25 cells) displays the population and the area for each municipality in France in 2015.
https://data.opendatasoft.com/explore/dataset/code-postal-code-insee-2015%40public/export/

Both datasets will be used in the csv format and the first step of this project will be the merge of the datasets using zip codes of different municipalities with python.

The project will also be presented as an interactive map where we will be able to observe on a map of the France the distribution of health professionals according to their specialty and the population of each municipality.

Description of the dataset

- After importing both data frames and looking of their structures, I noticed in particular one column of each data frame (‘healthcares’ and ‘municipalities’) that had a lot of null data (>100,000 and > 35,000 respectively). I decided to remove it (the values were not necessary for the project).

- The ‘zip code’ columns, common in both data frames are very important because they could be useful to merge them later. Then, I decided to remove all the null values in ‘zip code’ columns in the two data frames.

- The France is composed of a continental territory and five overseas departments. The data frames include data for all these territories. In this study, we will focus only on metropolitan France. Thus, I deleted the data concerning the five overseas departments in both data frames.

- I noticed that the type of ‘zip code’ column of ‘municipalities’ dataset is still object despite the elimination of the null cells. To understand why, I sorted the data frame by the ‘zip code’ column in descending order. And I found ten rows that not contain a zip code but a string value. I deleted these ten rows.

- I noticed that some name of region in ‘municipalities’ dataset are false so I corrected them.

- I created 4 new columns that could be useful later in ‘municipalities’ dataset.
  - The first one is composed of the object values: ‘north’ and ‘south’ according to the location of the regions. There are 5 regions in the south and 8 regions in the north of the France.
  - Sometime, one zip code can be the same for several municipalities. In prediction to the merge of the two datasets, the population numbers were grouped by each zip code and

then this new dataset was merged with the Municipalities dataset. This second new column is the population by zip code.

- The third and fourth column are the latitude and longitude for each municipality present in the same column in the dataset.

- Finally, to improve our dataset, we added deux new features: poverty_rate and unemployment_rate. They come from:

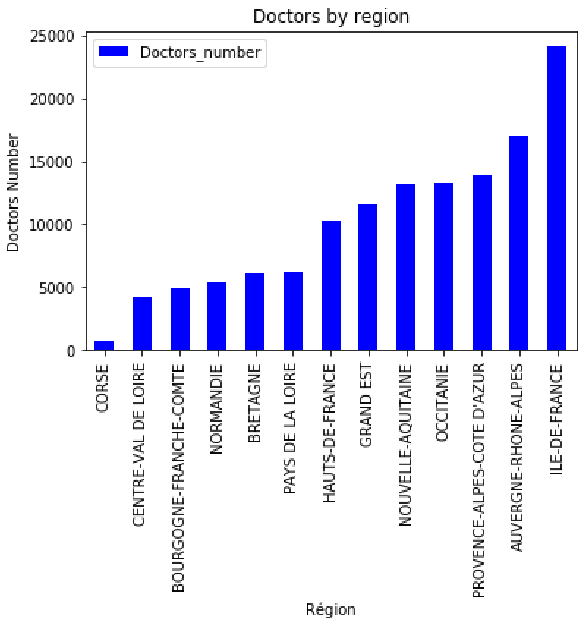
  https://public.opendatasoft.com/explore/dataset/revenus-disponibles-pauvrete-2012-iris/export/
  https://public.opendatasoft.com/explore/dataset/taux-de-chomage-par-zone-demploi-france-2003-2015/export/
  We first group the data by region because there were some missing data into the dataset. And then, we merged the two new features to the municipalities dataset.

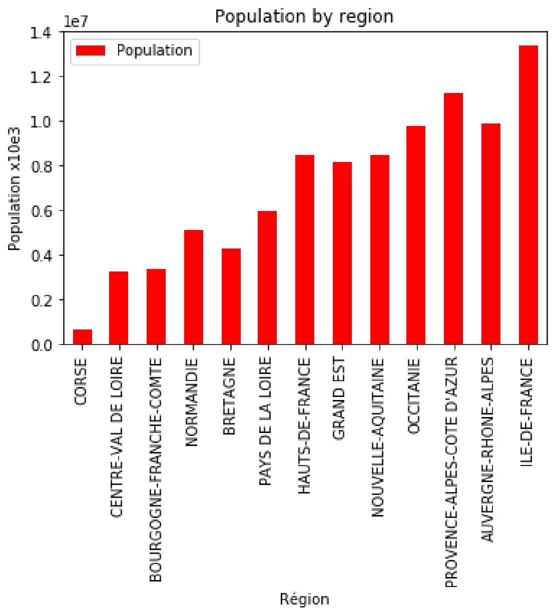
- Then, the important step here was to merge both data frames. We have the 'zip code' column common in each dataset. The problem with the 'zip code' column is that each zip code could match with several different municipality and after the merge with a left join, we obtain a repetition of several rows for this type of zip code. The population was calculated by zip code, so the repeated rows were deleted (the first one was kept) using the column with the name of each doctors.

Visual exploratory data analysis

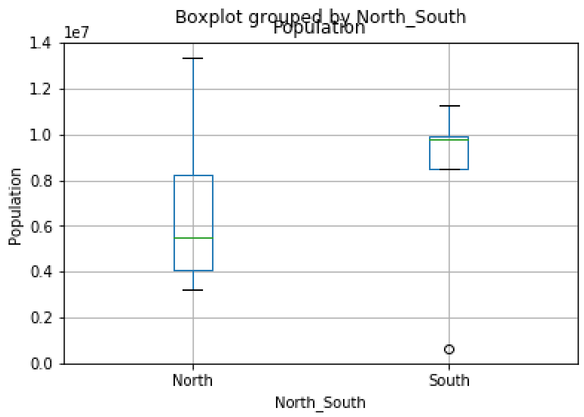
The first plot shows the number of doctors for each region of the France. The values are presented in ascending number. We saw a very little number of doctors in Corse which can be explain by the fact that the Corse a little island near to the France. The biggest number of doctors is in Ile-de-France which the capital of the France.

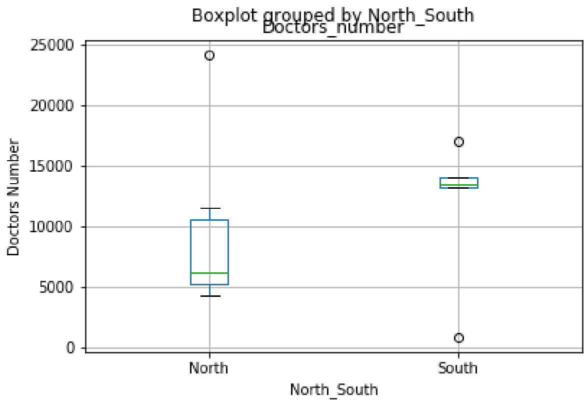
The second plot represents the population for each region of the France. Each region is presented in the same order than the previous plot to compare them. It seems that the population almost follows a similar distribution than the number of doctors.



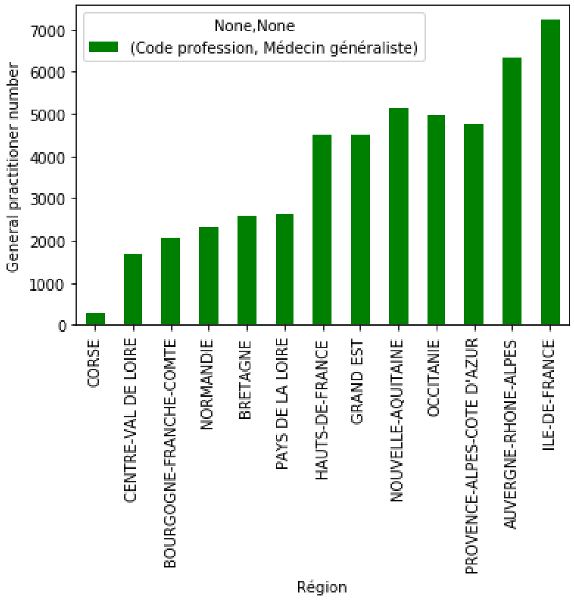


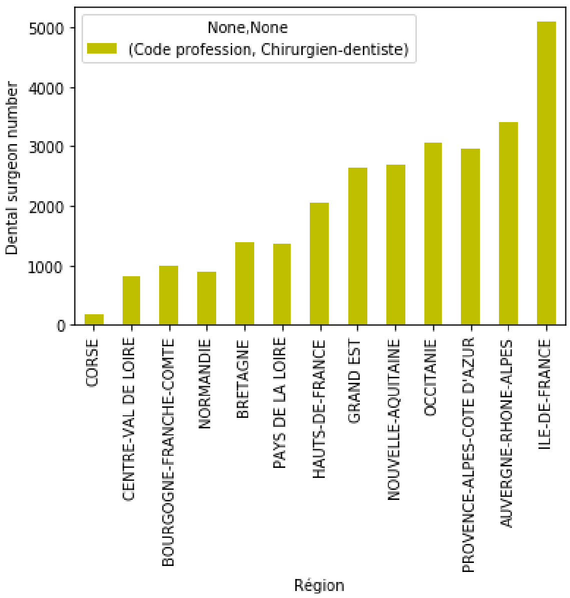
The following plots show the population (first plot) and the number of doctors (second plot) according to the location of the regions (in the north or in the south of the France). It seems there is a lot of more doctors in the regions located in the south of the France but there is also more population on the south of the country.





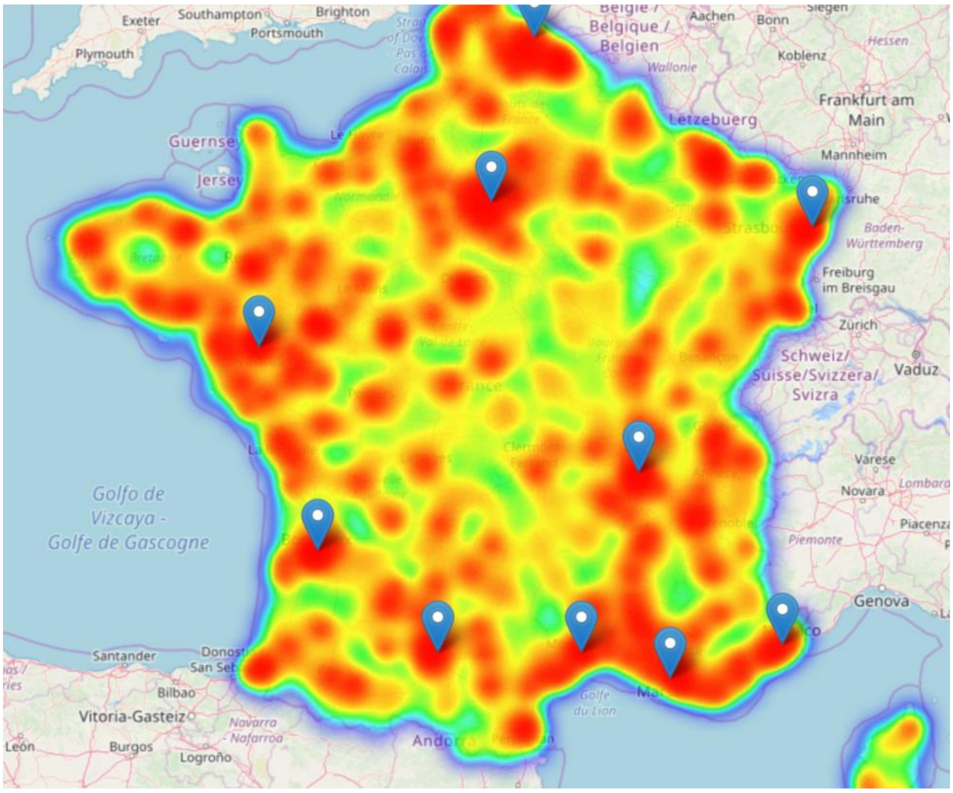
The following plots show the repartition of each general practitioners (first plot) and each dental surgeon (second plot), the two most numerous doctor specialties, for each region (presented in the same order than the previous plots). We can see they seem following the same repartition as the number of all doctors.





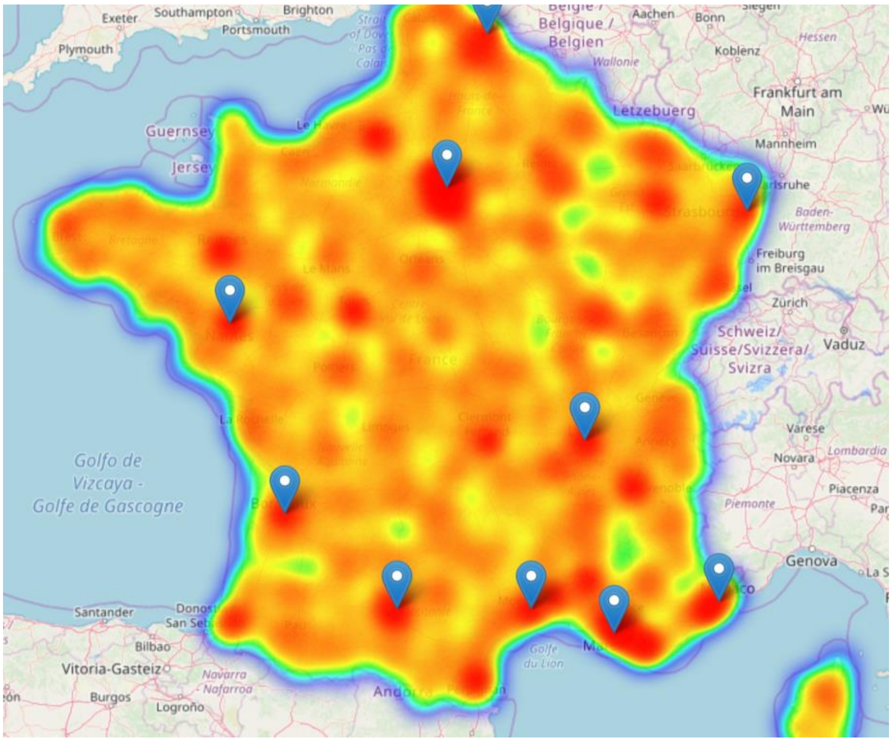
The next plot is a heatmap presenting the density of the population in France. The marker plots show the ten most populated cities in France (Paris, Lyon, Marseille, Toulouse, Bordeaux, Nantes, Strasbourg, Nice, Montpellier and Lille).

Density of the population of the France

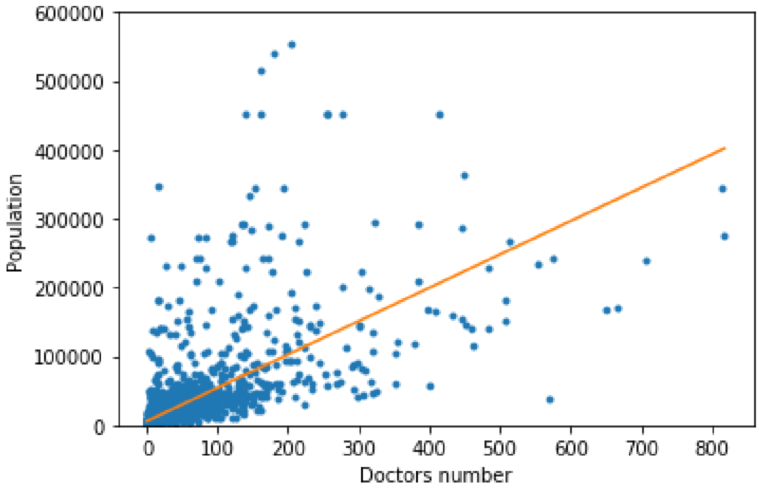


The following plot is a heat map showing the density of doctors in France. We can see there is a lot of doctors in big cities and around the Mediterranean Sea.

Density of number of doctors in the France



The last plot is a scatter plot showing the number of doctors according to the population in France. The orange straight line is the linear regression line. It seems to have a positive correlation between the population and the number of doctors in France. The number of doctors is increasing in the most populated areas of France.



- ⇨ Despite differences between the regions of France, there seems to be a correlation between the number of doctors and the population.

Statistical exploratory data analysis

During the data story part, we saw that it seems there is a correlation between the number of doctors and the population in France. According to the scatter plot showing the population depending on the doctor number and the heatmaps showing variations in population and number of doctors throughout France, we saw there is a higher number of doctors in location with a large population. Let us now take a closer look at the relationship between these two variables from a statistical point of view and regarding the location. To do this, we will perform first a correlation test between these two variables. Then, we will do a one-way ANOVA between the ratio of the number of doctors according to the population in northern France and southern France. Finally, we will do the same thing depending on the regions of France. To do this, we will use the dataset Population_Doctor_by_zipcode, which includes the number of doctors, the population, the ratio between both variables, the regions and the north/south class grouped by zipcode. We took a threshold α=0.05.

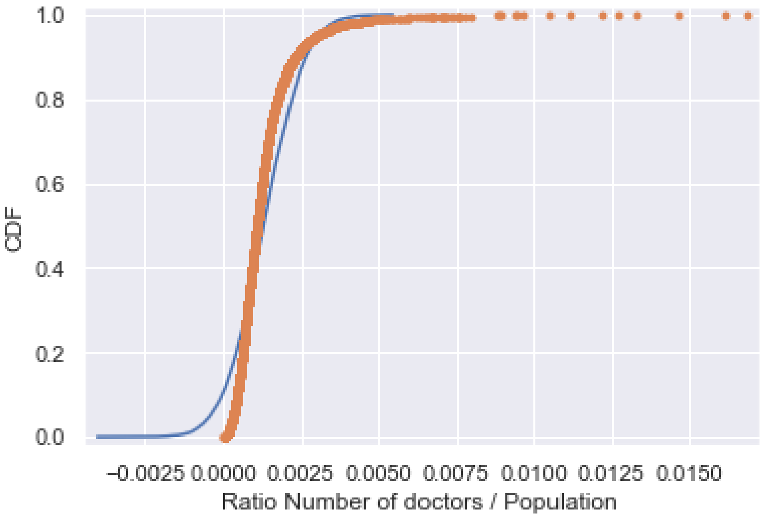
Hypothesis test correlation between the number of doctors and the population

Null Hypothesis: the two variables Number of Doctors and Population are completely uncorrelated.

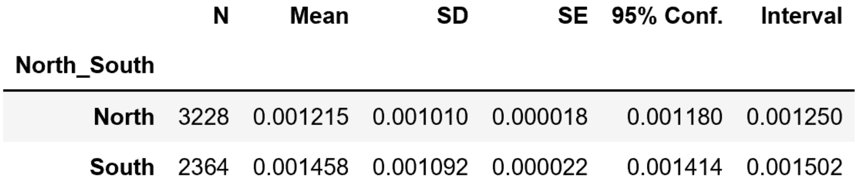
We have simulated the data assuming the null hypothesis is true. We had using Pearson correlation, p as the statistic test. Then, using bootstrapping resampling, we had computed p-value as fraction of replicates that have p at least as large as observed. The procedure done, and in all 10000 of the replicates under the null hypothesis not one had a Pearson correlation coefficient as high as the observed value of 0.65. The p-value is very small, so we have rejected the null hypothesis. There is a significant correlation between the number of Doctors and the Population. There is a higher Number of Doctors for the more populated territories.

One-Way ANOVA between the ratio number of doctors / population on the North and the ratio on the South

First, let’s check the normality of the values by plotting the cumulative distribution function. According to the CDF plot, the data seem normally distributed.



Here a summary of the data using rp.summary_cont().



	N	Mean	SD	SE	95% Conf.	Interval
North_South						
North	3228	0.001215	0.001010	0.000018	0.001180	0.001250
South	2364	0.001458	0.001092	0.000022	0.001414	0.001502

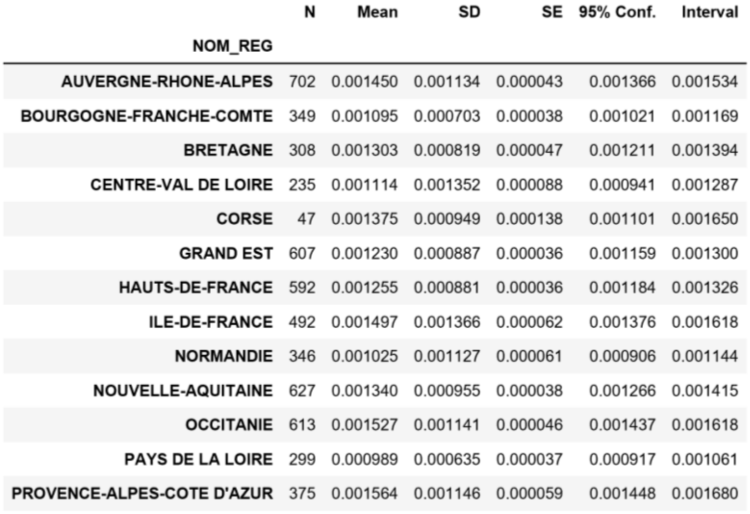
Then, using stats.f_oneway(), we calculated the ANOVA between the ratio Doctors number / Population on the North of the France and the ratio on the South of the France and we obtained:

```
F_onewayResult(statistic=73.51922931198455,
pvalue=1.2747156042454852e-17)
```

There is a significant difference between the ratio on the North and the ration on the South of the France. The ratio Doctors number / Population is higher on the South of the France.

One-Way ANOVA between the ratios number of doctors / population depending on the location (Régions)

Here a summary of the data:



	N	Mean	SD	SE	95% Conf.	Interval
NOM_REG						
AUVERGNE-RHONE-ALPES	702	0.001450	0.001134	0.000043	0.001366	0.001534
BOURGOGNE-FRANCHE-COMTE	349	0.001095	0.000703	0.000038	0.001021	0.001169
BRETAGNE	308	0.001303	0.000819	0.000047	0.001211	0.001394
CENTRE-VAL DE LOIRE	235	0.001114	0.001352	0.000088	0.000941	0.001287
CORSE	47	0.001375	0.000949	0.000138	0.001101	0.001650
GRAND EST	607	0.001230	0.000887	0.000036	0.001159	0.001300
HAUTS-DE-FRANCE	592	0.001255	0.000881	0.000036	0.001184	0.001326
ILE-DE-FRANCE	492	0.001497	0.001366	0.000062	0.001376	0.001618
NORMANDIE	346	0.001025	0.001127	0.000061	0.000906	0.001144
NOUVELLE-AQUITAINE	627	0.001340	0.000955	0.000038	0.001266	0.001415
OCCITANIE	613	0.001527	0.001141	0.000046	0.001437	0.001618
PAYS DE LA LOIRE	299	0.000989	0.000635	0.000037	0.000917	0.001061
PROVENCE-ALPES-COTE D'AZUR	375	0.001564	0.001146	0.000059	0.001448	0.001680

Then we calculated the ANOVA between the ratios Doctors number / Population depending on the region of the France and we obtained:

```
F_onewayResult(statistic=13.474806947107517,
pvalue=6.513761694105628e-28)
```

There is a significant difference between the ratios depending on the region of the France. Then we have made some post hoc comparison using Bonferroni Correction to analyse if there is a significant difference between the ratios of specific regions. For example, let's see if there are significant differences between the ratios in Ile-de-France where the capital of France is located and other regions using stats.ttest_ind(). We obtained:

- `Bonferroni Correction between Ile-de-France and Bretagne: Ttest_indResult(statistic=2.25912195618384, pvalue=0.02414486706798545)`
- `Bonferroni Correction between Ile-de-France and Provence-Alpes-Côte-d'Azur:Ttest_indResult(statistic=-0.7638462263829737, pvalue=0.44516718045069736)`
- `Bonferroni Correction between Ile-de-France and Nouvelle-Aquitaine: Ttest_indResult(statistic=2.257721053343628, pvalue=0.024155482253022397)`

-
  ```
  Bonferroni Correction between Ile-de-France and Occitanie: Ttest_indResult(statistic=-0.3968092790129793, pvalue=0.6915848692456656)
  ```

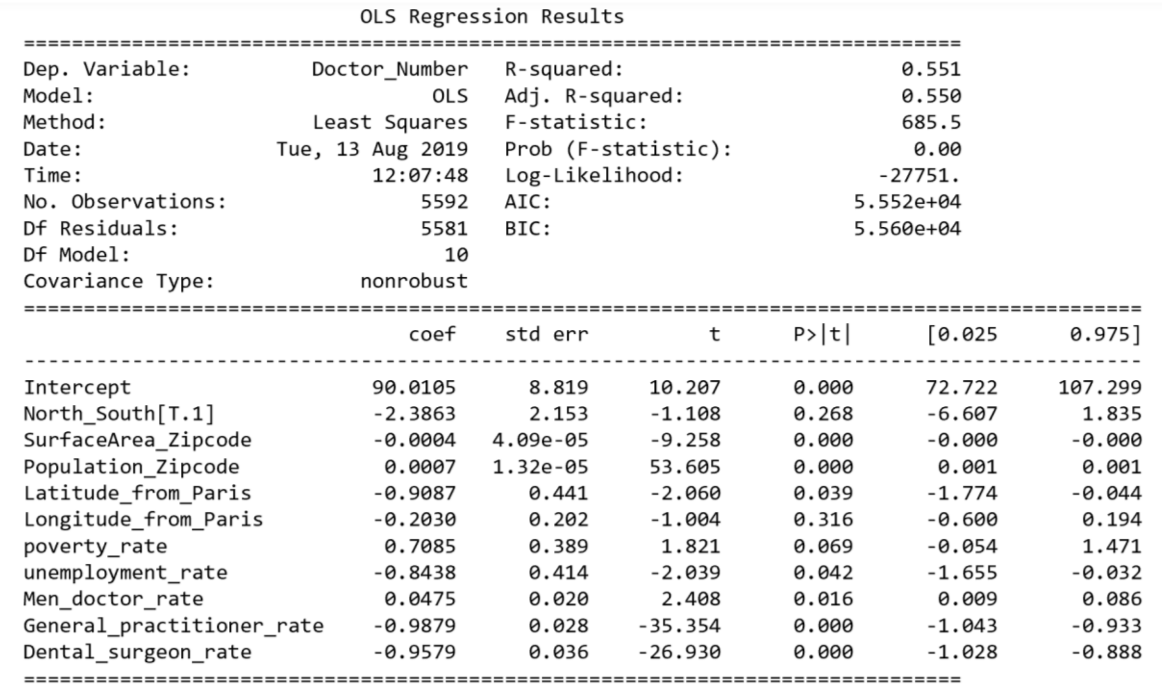
We can see there are significant differences between the ratio of the Ile-de-France and the ratio of the Nouvelle-Aquitaine or the ratio of the Bretagne. And there is no significant difference between the ratio of the Ile-de-France and the ratio of the Occitanie and the ratio of the Provence-Alpes-Côte-d'Azur. Then the ratio Doctors number / Population is very different according to the France regions and it's not depending only of the location of the region (North vs. South). Indeed, Ile-de-France and Bretagne are located on the North and have significantly different ratios and Occitanie or Provence-Alpes-Côte-d'Azur are on the South and do not have significantly different ratios with that of the Ile-de-France. They could be interesting features to predict the number of doctors in the next part.

Machine Learning

For the Machine learning part, we built a new dataset, for each zip-code we had the following column (features) : population, surface area, north/south (0 for the north and 1 for the south), the latitude in relation to Paris, the longitude in relation to Paris, the poverty rate, the unemployment rate, the women doctor's rate, the men doctor's rate, the general practitioner's rate, the dental surgeon's rate, the specialist doctor's rate and the number of doctor (dependant variable).

1. Predict the number of doctors per zip code

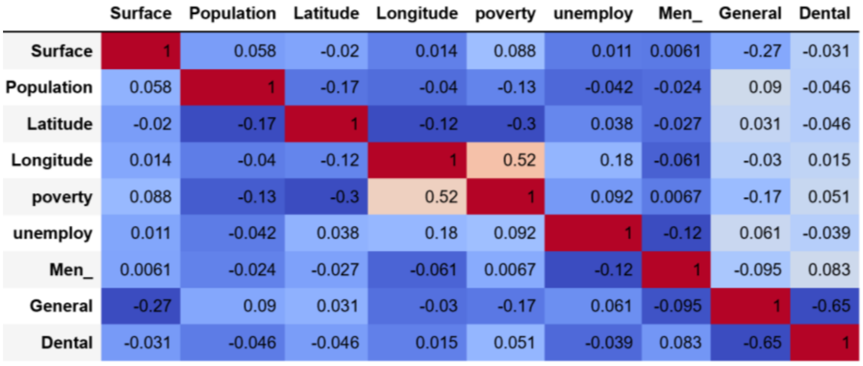
First using Statsmodels we checked the p_value and the coefficients of the features.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Doctor_Number   R-squared:                       0.551
Model:                            OLS   Adj. R-squared:                  0.550
Method:                 Least Squares   F-statistic:                     685.5
Date:                Tue, 13 Aug 2019   Prob (F-statistic):               0.00
Time:                        12:07:48   Log-Likelihood:                -27751.
No. Observations:                5592   AIC:                         5.552e+04
Df Residuals:                    5581   BIC:                         5.560e+04
Df Model:                          10
Covariance Type:            nonrobust
===============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                   90.0105      8.819     10.207      0.000      72.722     107.299
North_South[T.1]            -2.3863      2.153     -1.108      0.268      -6.607       1.835
SurfaceArea_Zipcode         -0.0004   4.09e-05     -9.258      0.000      -0.000      -0.000
Population_Zipcode           0.0007   1.32e-05     53.605      0.000       0.001       0.001
Latitude_from_Paris         -0.9087      0.441     -2.060      0.039      -1.774      -0.044
Longitude_from_Paris        -0.2030      0.202     -1.004      0.316      -0.600       0.194
poverty_rate                 0.7085      0.389      1.821      0.069      -0.054       1.471
unemployment_rate           -0.8438      0.414     -2.039      0.042      -1.655      -0.032
Men_doctor_rate              0.0475      0.020      2.408      0.016       0.009       0.086
General_practitioner_rate   -0.9879      0.028    -35.354      0.000      -1.043      -0.933
Dental_surgeon_rate         -0.9579      0.036    -26.930      0.000      -1.028      -0.888
==============================================================================
```

We can see that the features North_South and Longitude_from_Paris have both a very high p_value. They are not good predictors for the number of doctors by zip code. The other features have a significant p_value or very close to be significant (poverty_rate) (for an α=0.05). The Population_Zipcode, the poverty_rate and the Men_doctor_rate have positive coefficients, the number of doctor increases when the value of these three features also increase. The other features have negative coefficients, the number of doctor increases when the value of these features decrease. The SurfaceArea_Zipcode, the Population_Zipcode, the General_practitioner_rate and the Dental_surgeon_rate have the lower p_values, less than 0.001. If the population increases by one unit (1000 people), the number of doctor increases by 0.0007. If the surface area decreases by one unit (1 square meter), the number of doctor increases by 0.0004. And if the general practitioner rate or the dental surgeon rate decrease by one unit (1%), the number of doctor increases by 0.99 and 0.96 respectively. The coefficient of the SurfaceArea_Zipcode is low, so we checked the relation between the number of doctor and the surface area with a univariate model, but we found the same value as with the multivariate model.

Then, we looked at the relation between the features with a correlation matrix:

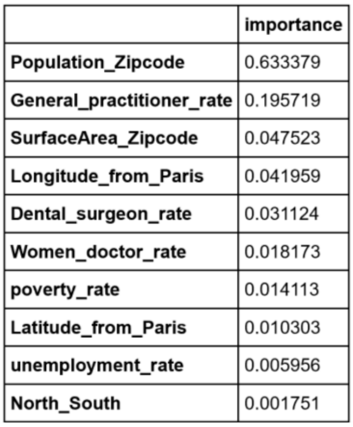


	Surface	Population	Latitude	Longitude	poverty	unemploy	Men_	General	Dental
Surface	1	0.058	-0.02	0.014	0.088	0.011	0.0061	-0.27	-0.031
Population	0.058	1	-0.17	-0.04	-0.13	-0.042	-0.024	0.09	-0.046
Latitude	-0.02	-0.17	1	-0.12	-0.3	0.038	-0.027	0.031	-0.046
Longitude	0.014	-0.04	-0.12	1	0.52	0.18	-0.061	-0.03	0.015
poverty	0.088	-0.13	-0.3	0.52	1	0.092	0.0067	-0.17	0.051
unemploy	0.011	-0.042	0.038	0.18	0.092	1	-0.12	0.061	-0.039
Men_	0.0061	-0.024	-0.027	-0.061	0.0067	-0.12	1	-0.095	0.083
General	-0.27	0.09	0.031	-0.03	-0.17	0.061	-0.095	1	-0.65
Dental	-0.031	-0.046	-0.046	0.015	0.051	-0.039	0.083	-0.65	1

We can observe a positive correlation between the poverty rate and the longitude from Paris (r = 0.52) and a negative correlation between the poverty rate and the latitude from Paris with a lower value (r = -0.30). We can also observe a negative correlation between the general practitioner rate and the dental surgeon rate (r = -0.65).

Then, in order to build a model to predict the number of doctors by zipcode based on our datatset and its features, we used a Random Forest Regressor. First, we split the dataset into two parts: 70% to train the model and 30% to test the model. Then we used GridSearchCV to find the best values for the max_depth parameter for our Random Forest model. Max_depth is the maximum depth of the tree. By using the best values for this parameter and after training the model, we find an accuracy R^2 of 0.93 for the training data and an accuracy R^2 of 0.78 for the test data. The first R^2 score indicates overfitting and the second one is a bit weak. The weakness of this model is certainly due to the fact that we have too few features in our dataset that can have an impact on the number of doctors per postal code. Unfortunately, I was not able to collect more data with public access.

Finally, we looked at the order of importance of the features of our model.



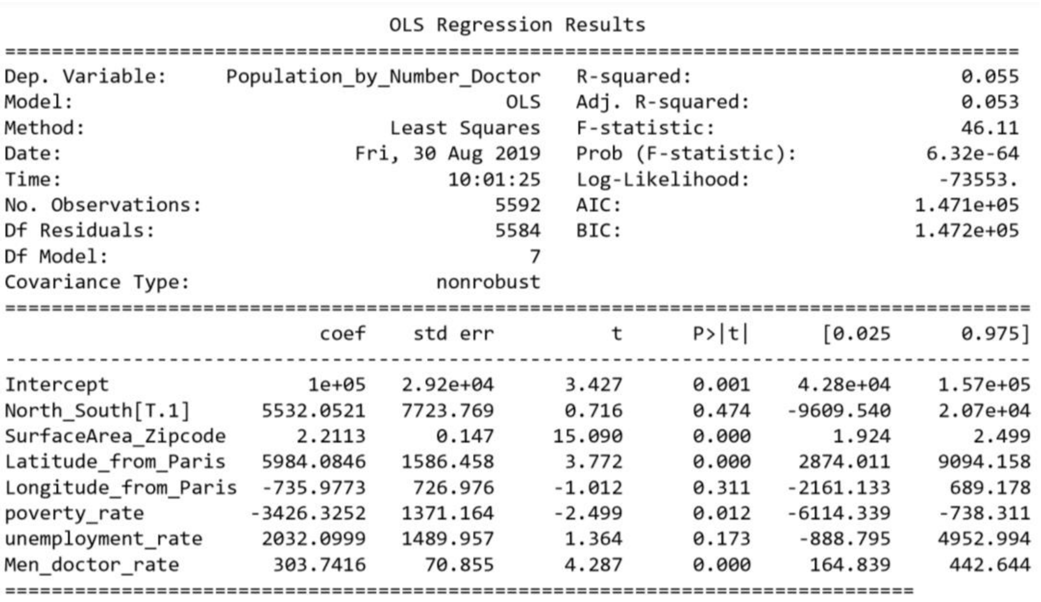
	importance
Population_Zipcode	0.633379
General_practitioner_rate	0.195719
SurfaceArea_Zipcode	0.047523
Longitude_from_Paris	0.041959
Dental_surgeon_rate	0.031124
Women_doctor_rate	0.018173
poverty_rate	0.014113
Latitude_from_Paris	0.010303
unemployment_rate	0.005956
North_South	0.001751

We can see that the two most important features to predict the number of doctors are the Population_Zipcode and the General_practitioner_rate, two features with a very low p_value, with a predominance of Population_Zipcode. Indeed, we had found a correlation between the population and the number of doctors in the statistical and visual representation part.

2. Predict the population by General Practitioner by zip code

For this part, we added a new column: the population divided by the number of General Practitioner by zip code.

Then, using again Statsmodels we checked the p_value and the coefficients of the features.

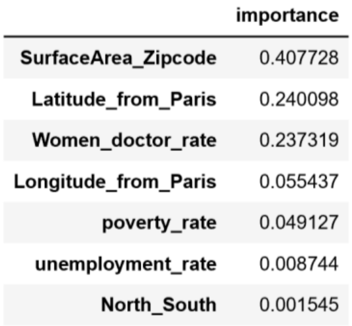


```
                              OLS Regression Results
==============================================================================
Dep. Variable:     Population_by_Number_Doctor   R-squared:                 0.055
Model:                                     OLS   Adj. R-squared:            0.053
Method:                          Least Squares   F-statistic:               46.11
Date:                         Fri, 30 Aug 2019   Prob (F-statistic):     6.32e-64
Time:                                 10:01:25   Log-Likelihood:          -73553.
No. Observations:                         5592   AIC:                   1.471e+05
Df Residuals:                             5584   BIC:                   1.472e+05
Df Model:                                    7
Covariance Type:                     nonrobust
================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                1e+05   2.92e+04      3.427      0.001    4.28e+04    1.57e+05
North_South[T.1]     5532.0521   7723.769      0.716      0.474   -9609.540    2.07e+04
SurfaceArea_Zipcode     2.2113      0.147     15.090      0.000       1.924       2.499
Latitude_from_Paris  5984.0846   1586.458      3.772      0.000    2874.011    9094.158
Longitude_from_Paris -735.9773    726.976     -1.012      0.311   -2161.133     689.178
poverty_rate        -3426.3252   1371.164     -2.499      0.012   -6114.339    -738.311
unemployment_rate    2032.0999   1489.957      1.364      0.173    -888.795    4952.994
Men_doctor_rate       303.7416     70.855      4.287      0.000     164.839     442.644
=======================================================================
```

We can see that the features unemployment_rate, North_South and Longitude_from_Paris have both a very high p_value. They are not good predictors for the population by general practitioner. The other features have a significant p_value (for an α=0.05). The SurfaceArea_Zipcode, the Latitude_from_Paris and the Men_doctor_rate have positive coefficients, the population by general practitioner increases when the value of these three features also increase. If the surface area decreases by one unit (1 square meter), the population by general practitioner increases by 2.21 (*1000). If the men doctor rate decreases by one unit (1%), the population by general practitioner increases by 303.74 (*1000). And if the latitude from Paris increases by one unit (1°), the population by general practitioner increases by

5984.08 (*1000). The poverty_rate feature has a negative coefficient, the population by general practitioner increases by 3426.33 (*1000) when the poverty rate decrease by one unit (1%). Unfortunately, for this linear regression, the R^2 as well as the adjusted R^2 are extremely low ($R^2=0.055$ and adjusted $R^2=0.053$).

Then, we used the same process as before to build a model to predict the population by general practitioner by zip code using a Random Forest Regressor. We find an accuracy R^2 of 0.13 for the training data and an accuracy R^2 of 0.18 for the test data. These accuracies are very weak. To finish, here is the order of importance of the features of our model.



	importance
SurfaceArea_Zipcode	0.407728
Latitude_from_Paris	0.240098
Women_doctor_rate	0.237319
Longitude_from_Paris	0.055437
poverty_rate	0.049127
unemployment_rate	0.008744
North_South	0.001545

To conclude this project, in a first time we built a model to predict the number of doctors by zip code. We saw that the most important feature to predict the number of doctors is the population, there are more doctors in the most populous cities. To go a step further, we wanted to verify that there is the same ratio of general practitioners per number of inhabitants in both large and smaller cities. Thus, in a second time, we built a model to predict to predict the population by general practitioner by zip code. The result of this model could have shown us in which cities (smaller or larger) general practitioners should settle first. Unfortunately, we were not able to obtain a high enough accuracy based on the available features to be able to use the model and draw a conclusion. To improve the quality of this model we would have needed more features that have an impact on our variable to predict.