**Seattle: Data Analysis of the city**

Carlos Espejo Peña

October 18, 2020

**1. Introduction: Description and purpose of this project**

Seattle is a seaport city on the West Coast of the United States. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States.

In this data analysis of the city of Seattle, the objective will be to find patterns within it using data. Firstly, this research starts analyzing which are the most affected areas by Fire Alarms and if this is related with the population rate in these areas. Secondly, it is also possible to connect the employment ratio with the crime one. Usually, neighborhoods with more business and nearear to the city centre tend to have a higher number of incidents. Let's see if Seattle follows this trend.

Gathering all these concerns, it would be possible to create maps and information charts for the city of Seattle.

Finally, this project will end up in exploring each of the neighborhoods so as to find clusters in which they can be placed. To this aim, this clusterization will be based on the top venues hold in each of the districts. Therefore, different types of venues will be found, spotting the trending ones in each borough.

**1.1. Tools that will help in the development of this project**

1. python libraries:
   * **Data Wrangling and Manipulation**: pandas
   * **Modeling**: scikit-learn, numpy
   * **Web scraping**: requests, geopy
   * **Visualization**: matplotlib, seaborn, folium, pywaffle, wordcloud, PIL
2. **Foursquare API**: Tool to obtain the location data of the main venues in each neighborhood.

**2. Data Description**

These are the data sources we will use for working on this project:

* **seattle-neighborhoods.geojson** (1) -> geojson data needed to conform the choropleth maps.
* **Seattle\_Census\_Data.csv** (2) -> Data of the population of Seattle per neighborhood.
* **Seattle\_Employement\_Data.csv** (3) -> Data of the employment in Seattle per neighborhood.
* **Seattle\_Real\_Time\_Fire\_911\_Calls\_Last500.csv** (3) -> Last 500 Fire 911 Calls in the city of Seattle.
* **SPD\_Crime\_Data\_2020\_Last500.csv** (3) -> Last 500 crimes committed reported by the Seattle Police Department in the city of Seattle.
* **Forsquare API** (4) -> To get the most common venues of the Boroughs of Seattle.

Please Note: This data has been cleaned from the root sources since they were holding unnecessary information.

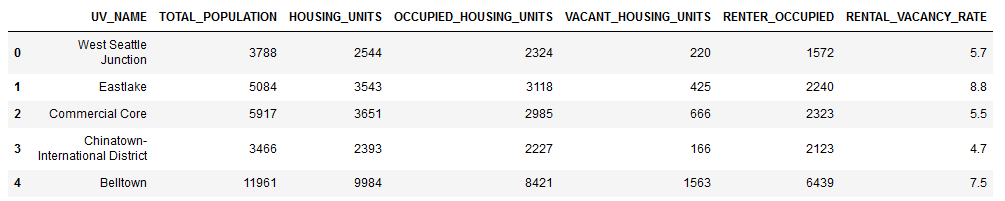
**3. Methodology**

During this section it will be discussed and described all the exploratory data analysis that were performed in this project. Statistical approaches will be used to test read data, and what machine learnings were used and why.

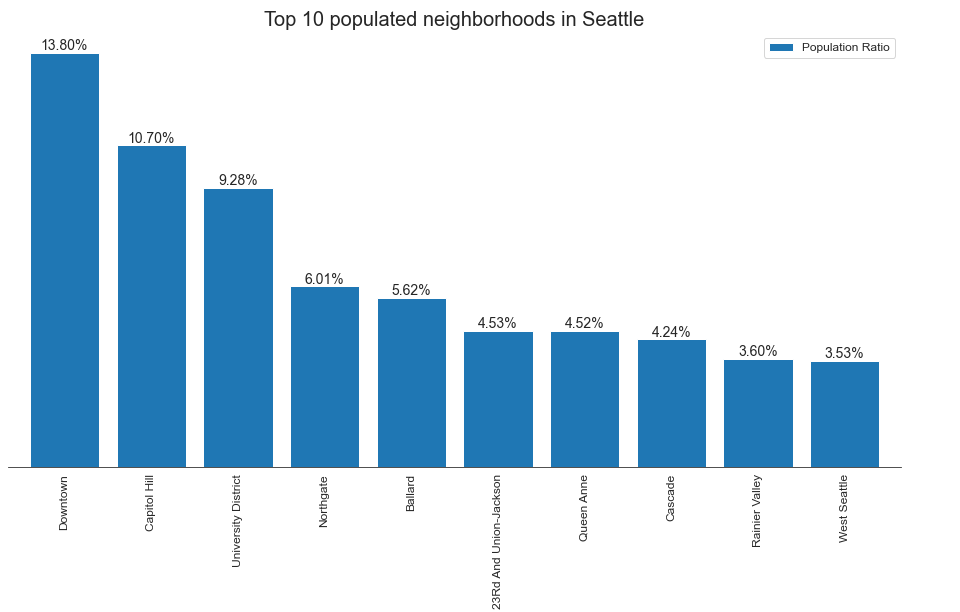
### 3.1. Importing & Exploring all the data for the project

Importing all the main packages for the project:

**3.1.1. Reading Seattle Census Data**

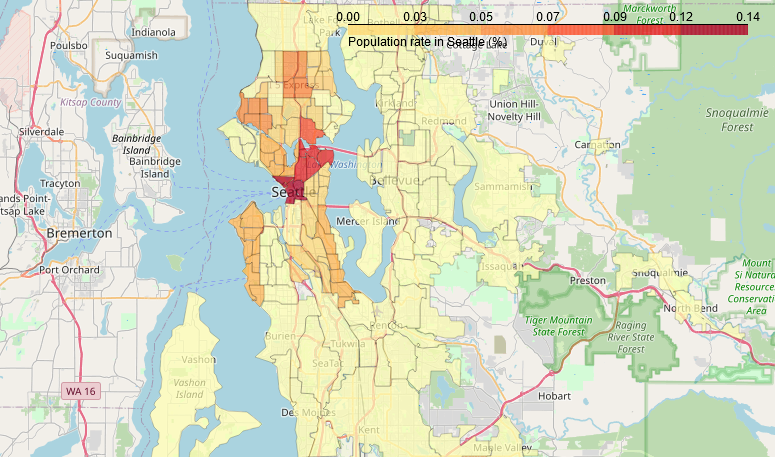


After reading the data with pandas, I calculated the “Population\_Ratio” by standardizing the values. To this aim, the ratio was conformed using the percentage of contribution of each neighborhood to the city. This resulted in the bar chart showed below:

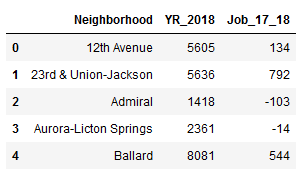


From the above bar chart, it is easy to spot that most of the population lives nearby the city centre, specifically in Downtown Seattle and Capitol Hill.

Let’s see how this data is displayed on Seattle’s map:

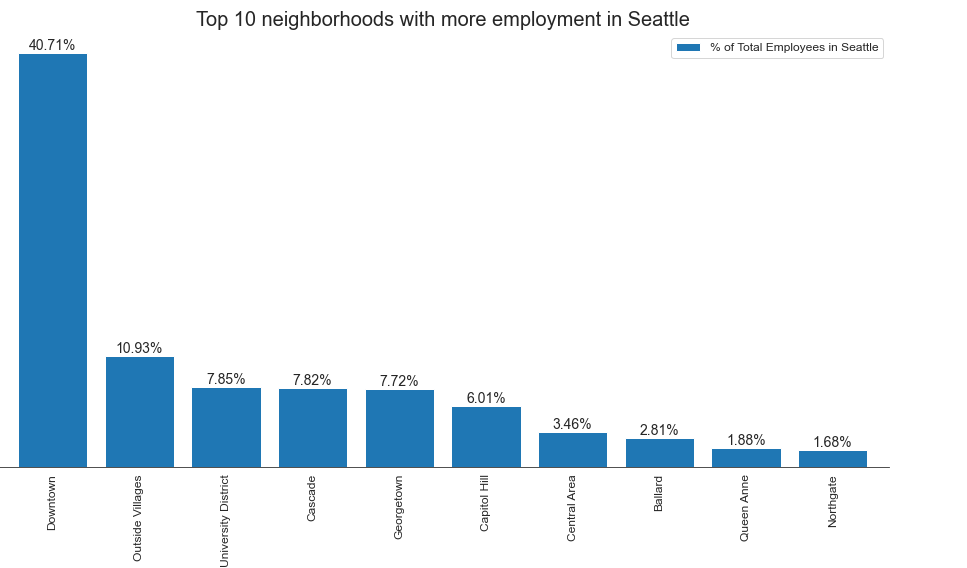


**3.1.2 Reading Seattle Employment Data**



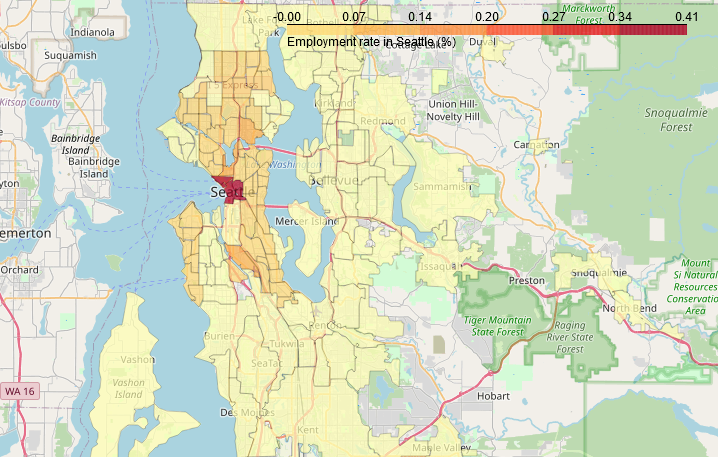
In this dataset we can find “YR\_2018”, which is the number of employees register per district in 2018.

After reading the data with pandas, I calculated the “Employment\_Ratio” by standardizing the values. To this aim, the ratio was conformed using the percentage of contribution of each neighborhood to the city. This resulted in the bar chart showed below:



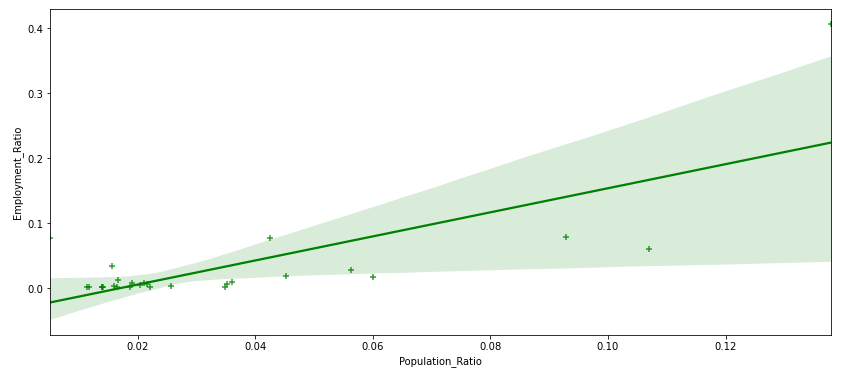
This employment ratio chart gives us an expected insight. It is due to the fact that the most employed neighborhoods are those located nearer to the city centre, so that is why Downtown holds more than the 40% of the active population of Seattle.

Let’s see how this data is displayed on Seattle’s map:

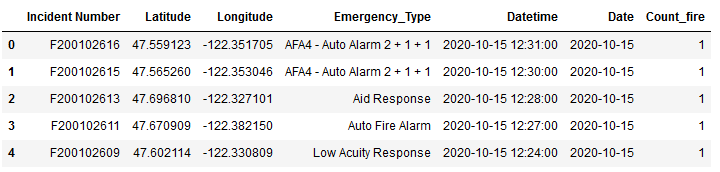


##### ***Correlation between the Population\_Ratio and the Employment\_Ratio in the city of Seattle***

As we can see in this regression plot, population and employment have a strong positive correlation, which makes common sense:



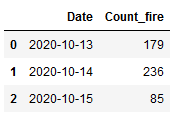
**3.1.3. Reading Seattle Fire 911 Calls Data**



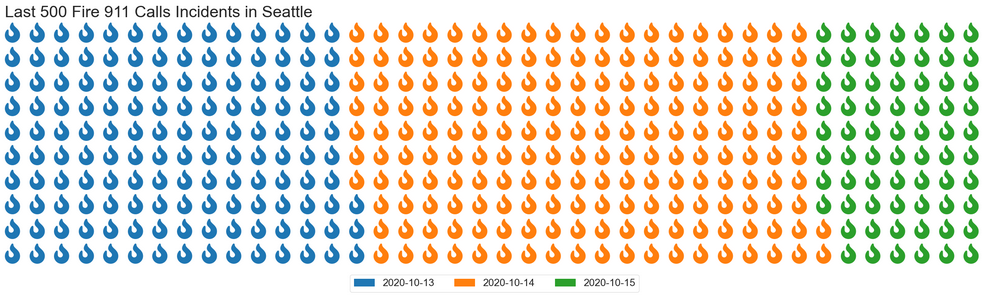
When reading this data, “Data” and “Count\_fire” columns were missing. Using pandas library, we added these last two columns which will be essential for our analysis in the methodology section.

***Creating a Waffle to show the number of incident calls***

We grouped the data to be easily used by pywaffle:

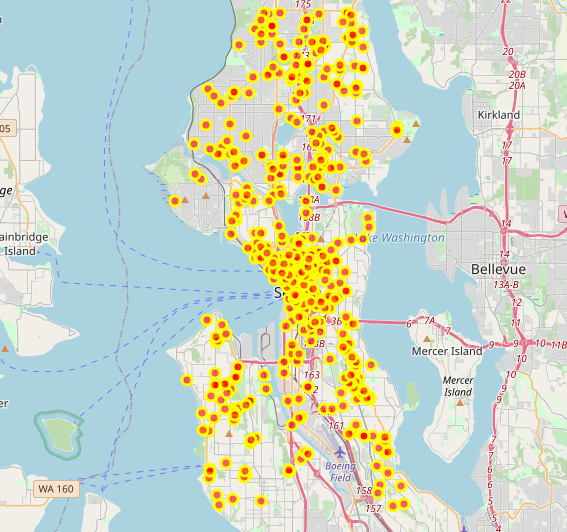


In the data frame, it is detailed that the last 500 fire 911 calls were made in 3 days.



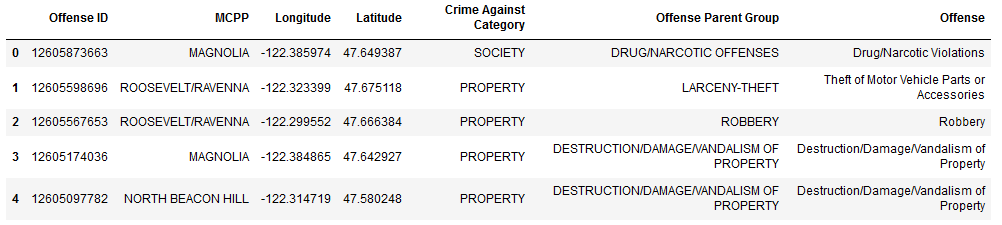
From the Waffle Chart above, it is easy to spot the day in which the 911 service received more calls from Seattle citizens. Therefore, as we can also affirm from the grouped data frame, the day with more incidents was the 14th of October of 2020.

Let’s also visualize this data on a map, so we can have a look where most of the incidents took place:



As we can see in the map, most of the fire incidents are located in those areas with the highest population ratio, which makes sense since there are more housing units and local businesses.

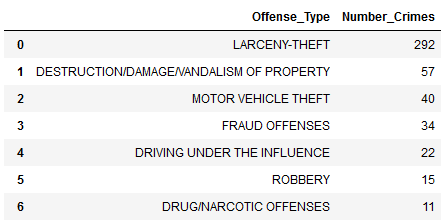
**3.1.4. Reading SPD Crime Data**



In this case, we will use a different method to have a quick view of the data. To this aim, a word cloud will be plotted to see which crimes are the most common in a quick glance.

***Creating a Word Cloud to show the most common crimes in Seattle***

Firstly, we will group the data using pandas in order to have the adequate format to be used by the word cloud:

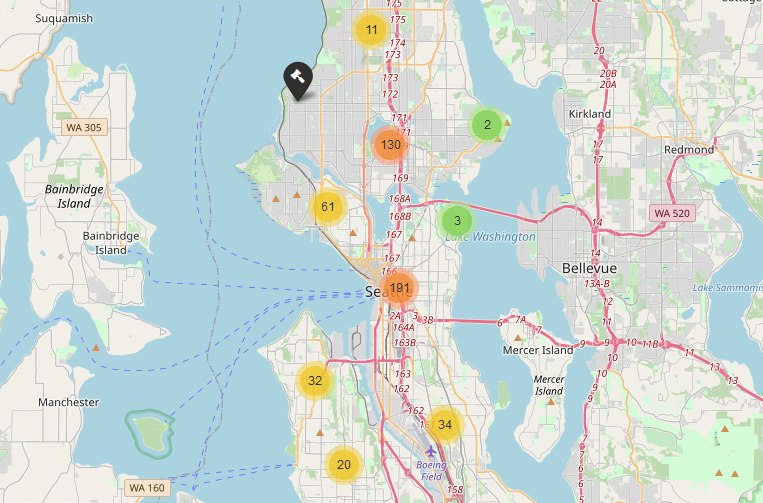
 

To plot the word cloud, I used a template of the map of Seattle so as to place this analysis in context. From the word cloud it is possible to deduce the most common reported crimes in the city of Seattle by its Police Department. Therefore, among the most common incidents that could be highlighted, we can find:

* Offenses
* Theft / Larceny
* Robbery
* Vandalism
* Destruction

**Creating a cluster map plot for the Crimes committed in Seattle**

To build this map, the folium library was used and all the incidents were grouped using a clustering visualization based on proximity of nodes:



In this map, we find a similar figure as with the Fire incidents map. Those neighborhoods with the highest population and employment ratio are the ones which holds most of the criminality of Seattle. What is more, the top offenders in this city are Downtown Seattle and First Hill, which are located in the city centre.

It is also important to remark the impact of crimes in the north area of the city, especially in the whereabouts of Green Lake and Woodland Park. However, Madison Park presumes to be the most secure district, holding only 3 incidents out of the last 500 reported by Seattle Police Department.

### 3.2. Process to cluster neighborhoods in Seattle based on top venues

After cleaning the dataset and defining a new df (df\_seattle), we will start exploring and analysing the df venues following this process:

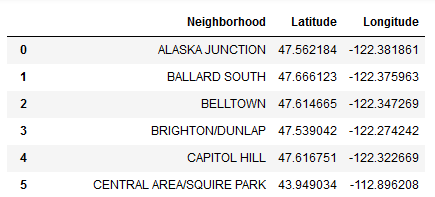
* Exploring each neighborhood in Seattle
  + Cleaning the dataset.
  + Defining a function to obtain all the data from venues.
  + Define the credentials to Foursquare.
* Getting the venues per neighborhood.
  + Analyse each neighborhood: Top 10 venues.

Finally, moving to the results section:

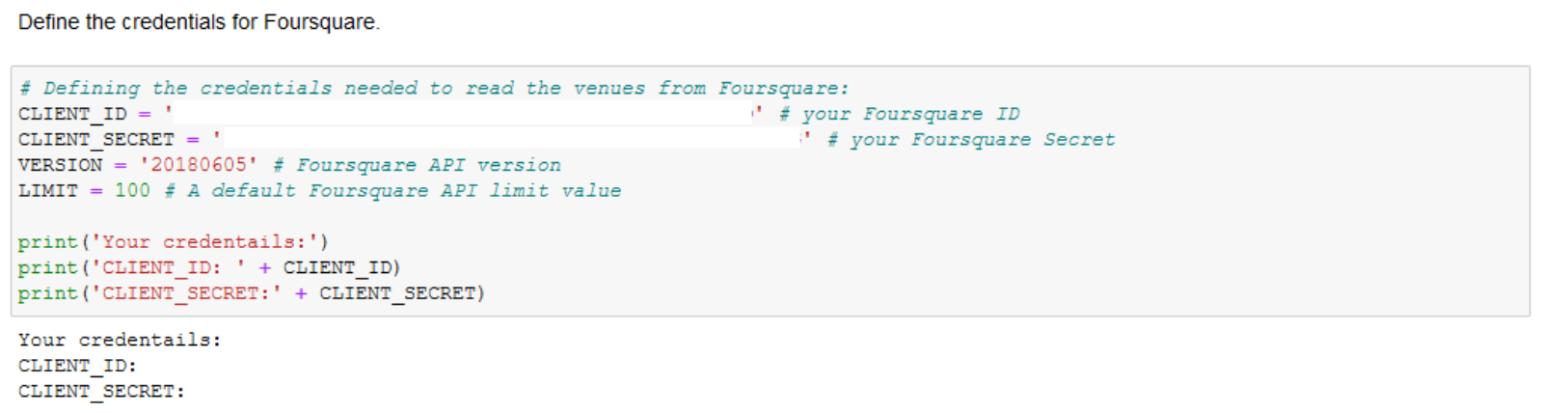
* Clustering Seattle's neighborhoods.
  + Examine Seattle's clusters.

**3.2.1. Exploring each neighborhood in Seattle**

This process starts by cleaning the dataset and defining the “df\_seattle” before working on it. It contains the coordinates (latitude, longitude) of each of the 41 neighborhoods in Seattle that we will use to build the model:

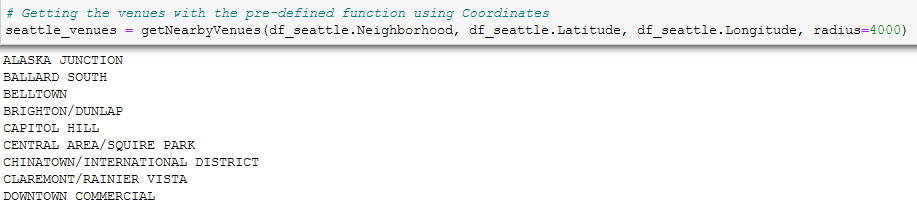


It is time to initialize the connection to the Foursquare API, providing our credentials:



**3.2.2. Getting the venues per Neighborhood**

A function was built (“getNearbyVenues”) to extract all the venues around each borough:

After manipulating the data, we obtain a data frame with 242 unique venues categories. The table below shows the top 10 most common venues per borough:



**3.2.4. Clustering Seattle's neighborhoods**

The analysis of each neighborhood will be carried out using the clustering technique that consists of exploring and reviewing the neighborhoods, segmenting them and finally grouping them into clusters to obtain the results.

To this aim, a K-Means algorithm will be used. This is a specialized unsupervised Machine Learning model for clustering analysis for handling clustered data. It takes a bunch of nearest points and uses them to learn how to label other cases, so it classifies cases based on their similarity.

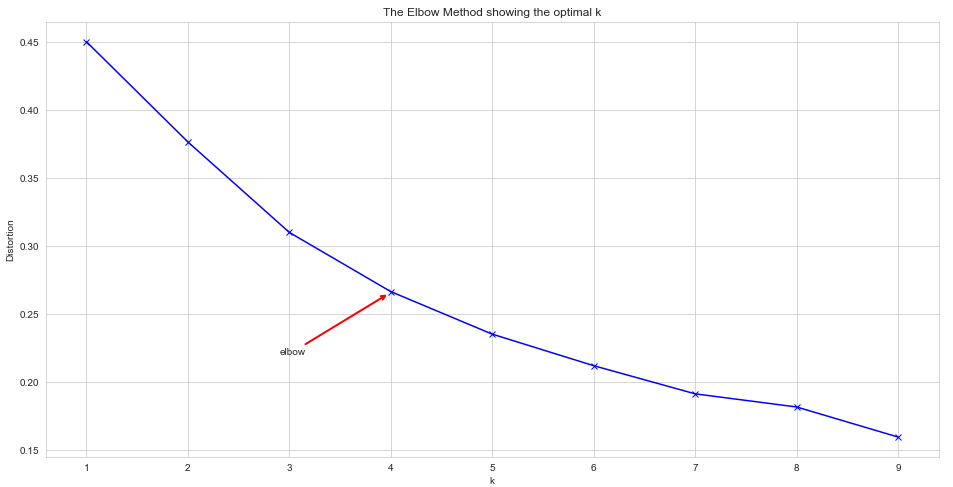
##### ***K-Means Elbow Method: Deciding the best value for k (num\_clusters)***

K-Means is an unsupervised machine learning algorithm that groups data into k number of clusters. The number of clusters is user-defined and the algorithm will try to group the data even if this number is not optimal for the specific case.

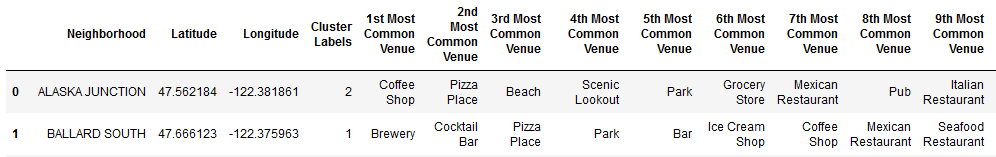
Therefore, we have to come up with a technique that somehow will help us decide how many clusters we should use for the K-Means model.

The Elbow method is a very popular technique and the idea is to run k-means clustering for a range of clusters k (let’s say from 1 to 10) and for each value, we are calculating the sum of squared distances from each point to its assigned center(distortions).

When the distortions are plotted and the plot looks like an arm then the “elbow” (the point of inflection on the curve) is the best value of k.

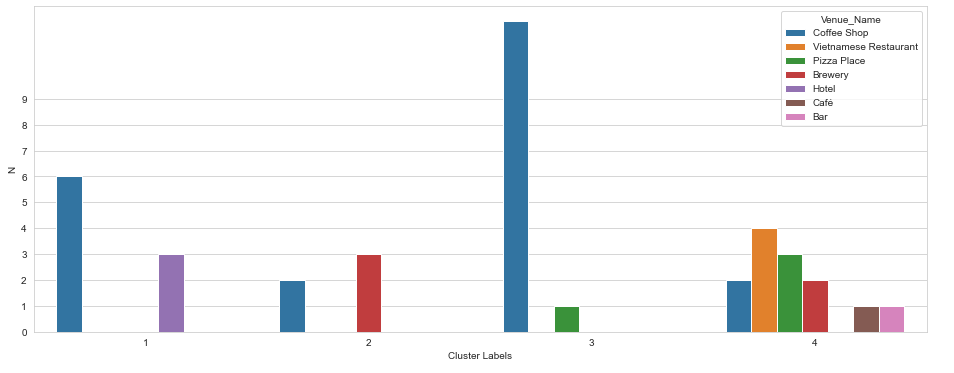


Therefore, the “elbow” is the number 4 which is optimal for this case. Now we can run a K-Means using 4 as the number of clusters.



***Defining the Clusters' Names***

We create a bar chart to see the number of times a venue appears in one cluster as “1st Most Common Venue”.



Labels for each cluster can be determined once the graph above is examined:

* **Cluster 1**: "Hotels Venues"
* **Cluster 2**: "Brewery Venues"
* **Cluster 3**: "Coffee Shop Venues"
* **Cluster 4**: "Restaurants & Bars Venues"

**4. Results**

In this section, the previous information will be merge into final plots. This will consist of elaborating a quick view of the details from the dataset used in this project, after the exploratory data analysis it was done in the methodology section. This is the index for the results section:

* Plotting Relation Maps in Seattle:
  + Fire 911 Calls reported: Plotted on the population choropleth map.
  + SPD Crimes reported: Plotted on the employment choropleth map.
* Mapping the venues' clusters on the city of Seattle:
  + Creating the map visualization.
  + Examine Seattle's neighborhoods.

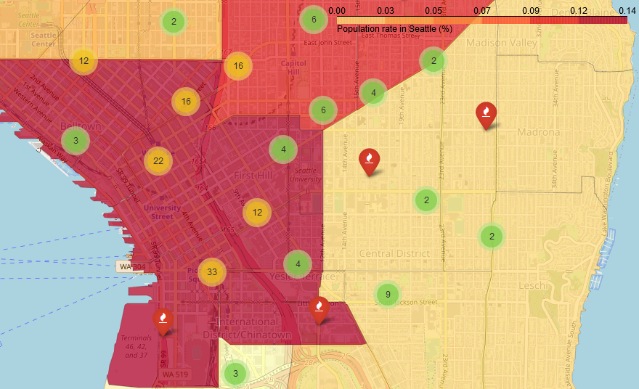
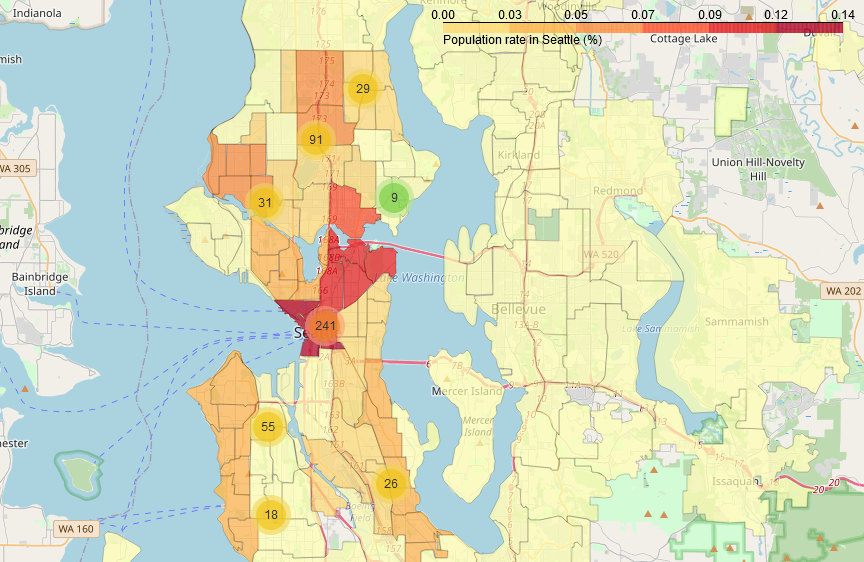
This analysis will provide a better understanding of the city of Seattle.

### 4.1. Plotting Relational Maps in Seattle

In this section, the previous maps exposed will be merged in order to find more insightful conclusions. Therefore, two maps will be plotted according to the estimated relation that some factors can have:

* **Population map** (Choropleth map) + **Fire map** (Map with Markers).
* **Employment map** (Choropleth map) + **Crime map** (Cluster Map).

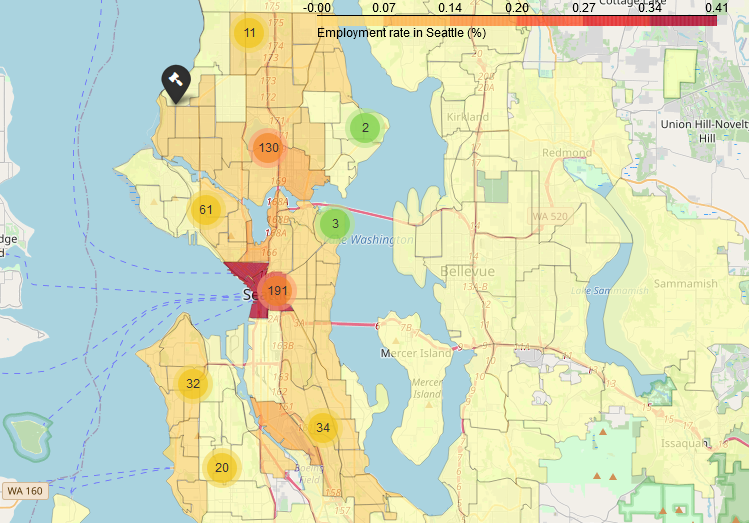
**4.1.1. Fire 911 Calls reported in Seattle: Plotted on the population choropleth map**

**

##### ***Comments of the resulting map***

As it was predicted in the introduction to this project, the major number of Fires reported in the 911 Calls are produced in those neighborhoods in the city with the most population.

**4.1.2. SPD Crimes reported in Seattle. Plotted on the employment choropleth map**



**Comments of the resulting map**

Once again, the initial prediction was right! Most of the Crimes reported by Seattle Police Department (SPD) are produced in those boroughs in the city with the most employment. This is due to the fact that they are the most crowded areas in the city.

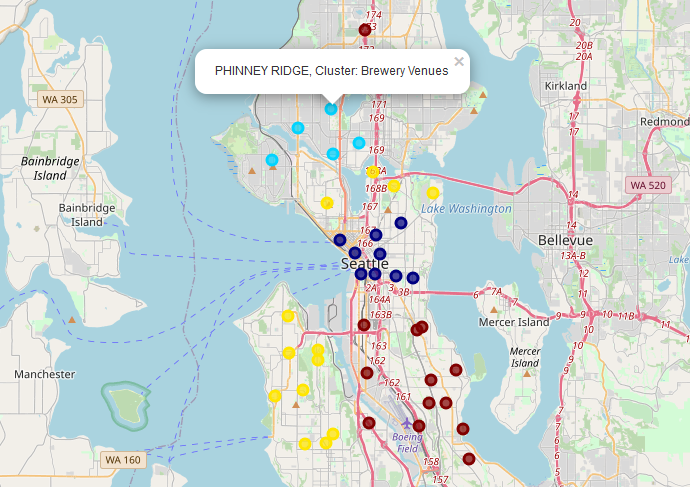
As it was shown in the regression plot between population and employment, they are positively correlated. This means that areas with more population have the most of employees, which makes natural sense. Therefore, if the major number of crimes are produced in areas with the highest employment ratios, they are occurring in those neighborhoods with the highest population ratios.

### 4.2. Mapping the venues' clusters on the city of Seattle

### In this section, we will visualize the clusters dividing each of the boroughs of the city into clusters. This is the result of applying the K-Means algorithm with a k=4 to our initial dataset.

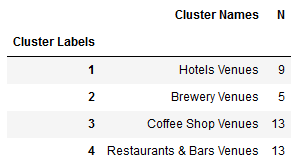
**4.2.1. Creating the map visualization**

In this map we can visualize the details for each neighborhood being categorized within a cluster, depending on its 1st most common venues:



**4.2.2. Examine Seattle's neighborhoods**

We can have a general view of the number of elements (N) contained in each of the conformed clusters:



**5. Discussion**

**As it was discussed in the introduction, Seattle is the largest populated city in the state of Washington. The amount of daily data that its metropolitan area produces in each of the districts is more than remarkable. Thereupon, in the middle of this complex metropolis, it is possible to find countless insights and analysis based on data.**

**I started by exploring the datasets which were used in this project. Using pandas for data wrangling and manipulation clarified the way in which the data frames could be analysed and represented. This lead to create some insightful graphs for data visualization. Population and employment ratios were disposed using bar charts and choropleth maps. Fire 911 calls were presented using a waffle chart to see the day with the most incidents. Seattle Police Department’s reported crime data were ordered in a word cloud created inside the map of Seattle. These visualizations enhance the way in which data insights can be obtained.**

**The correlation between “population - fires” or “employment - crimes” helps to understand how the city is organized. What is more, it can be a powerful decision-making tool to use so as to determine either where to open the next fire or police station in the city.**

**Moving to the algorithm selected to classify the venues, K-Means clustering method was the chosen one. I used the Elbow method to determine the optimum value for k (number of clusters), which resulted to be 4. To create this model, 41 neighborhoods of the city of Seattle were used. Nevertheless, if we want to improve the accuracy of this solution, more location segments in the city should be trained.**

**The objective of this project was to find the main most common venues that compose each of the boroughs in Seattle, so they could be divided into groups based on this criterion. The problem to tackle was to meet a model that could work on its own to discover invisible data trends for human eye. Therefore, unsupervised machine learning seemed to be the adequate approach to this case. Although other classification models could have been used in order to conform the different groups, clustering segmentation is the most suitable one for partitioning data into segments with similar characteristics. Moreover, this solution is the most visual as it can be easily represented in a map.**

**6. Conclusion**

In conclusion, most of the incidents occur in the downtown since it is the area with more activity and population. In order to reduce these numbers, the town hall of Seattle is able to use this data sources to direct their decisions. As I mentioned in the previous section, a data-driven decision-making system used by city managers could contribute on how to optimally organize the police and fire stations in Seattle.

Apart from the benefits data can have on city management, it could also be used to inform and guide tourism. Clustering helps to determine the most common venues in each area. Therefore, it could act as a recommendation system for visitors. Clusters can show the best locations for booking hotels and restaurants, or looking for a place to have either a beer or a coffee.

As far as we have data, it would be possible to even implement this model and analysis to any city in the world.