# Capstone Project - The Battle of the Neighborhoods (week 1)

(english is not my native language, so dont judge my grammatical errors =))

## 1. Problem Background:

Salt Lake City is the capital of Utah. The population is round about 200.000. The city was founded in 1847 by followers of the church. Further more it is one of the headquarters of the „Chruch of Jesus Christ of Latter-day Saints“. It has been developed to a strong outdoor recreation tourist industry based primarily on skiing and outdoor recreation.

Salt Lake City has an area of 110.4 square miles (286 $km^2$) and is historically known as the "Crossroads of the West" for its railroads, when nearby steel, mining and railroad operations provided a strong source of income. All in all Salt Lake City's modern economy is service-oriented.

Due to its raise in population, there are not only positive things, which changed the daily life of people. The crime statistic of Salty Lake City compared to Utah and National shows a lot of differences. The Total crime rate per 100k people is in the City 6907, in Utah 2611 and National 2580. The question is: why are there differences? Are the higher count of venues in the city correlated with a higher crime rate?

Source: https://www.areavibes.com/salt+lake+city-ut/crime/

## 2. Problem Description:

People want to live in a save environment. To fullfil this task, the police is the executive function of the country. But in reality there are some economical concerns. The police men and women have to be paid for their commitment. Of that fact the ressources are limited and have to be splitted over the whole area of the City (286 $km^2$). It would be nice to identify the hotspots of crimes and the gps location where the posibilty of crimes is higher than in other regions of the City. With this knowledge the police could act more efficient and split their limited ressources more accurat. For example: In a specific area, in which the crime rate is very high, there should be a higher presence of police than in a neighborhoods, in which the crime rate is very low.

## 3. Data Description and how to solve the problem

I download the addtitional data file i used to solve the problem via kaggle. This is a platform where scientest upload datasets to very different topics.

The dataset consists of the type of crime, the day, the location data(latidute and longitude), the district and a lot more columns which are not in concern of the problem i described. Round about 60000 cases are puplished in this dataset.

First of all i start with the very important data wrangling and drop the unneccecary columns to get a better overview. In the next step i cluster the loactaion data to get five centers of the different clusters (kmeans).I realised that there are two cluster which only contains 2 and 40 cases (of 60000), so i decided to drop these rows and classified them as outliners.

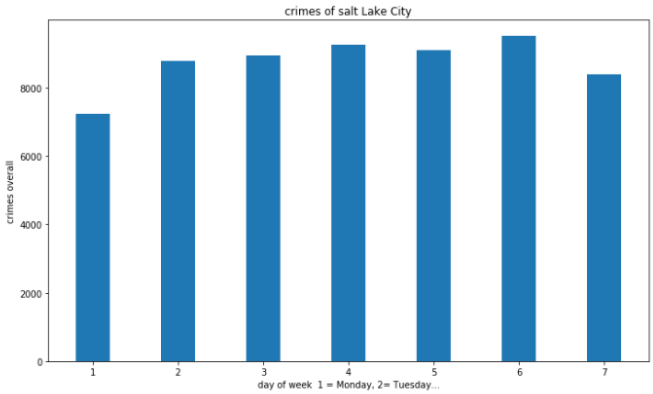
Via Fouresquare i requested the sourounding venues (radius= 400) of the three remaining centers. To compare the results i decided to get the sourounding venues of three other spots in the city. Visualisations show the results and if there are any correlations between the total number venues and the crime rate in a given location.

## 4. Methology and Results

### 4.1 day of week

To get a clearer overview of the dataset i dropped unnessecary rows and columns. In the next step the library " *from colections import Counter"* was used to identify if there are differences of total crimes due to the day of week. The results are shown below.The exploratives Data displayes a smaller account of crimes on Monday and Sunday.

| Wochentag | crime_day |
|---|---|
| 1 | 7238 |
| 2 | 8782 |
| 3 | 8950 |
| 4 | 9259 |
| 5 | 9099 |
| 6 | 9516 |
| 7 | 8396 |




To control the assumptions via a statistical method. I calculated an Annova, which displayed no significant changes due to the excact day.

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| crime_day | 1.0 | 7.578206 | 7.578206 | 1.855421 | 0.231307 |
| Residual | 5.0 | 20.421794 | 4.084359 | NaN | NaN |

### 4.2 Type of crime

in the given dataset there are is an excact description of the crimes involved. With the same procedure as in Cap 4.1. The Type of crime was analysed. The data shows, that the most frequent type of crime in Salt Lake Ciyt is Larcency with an account of 12478.

| | type | value |
|---|---|---|
| 7 | LARCENY | 12478 |
| 5 | PUBLIC ORDER | 8997 |
| 9 | DRUGS | 5282 |
| 6 | ASSAULT | 4383 |
| 4 | PUBLIC PEACE | 4146 |

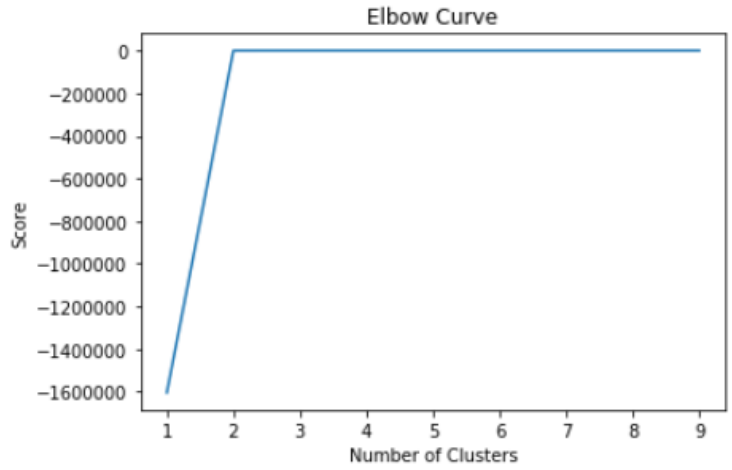
## 4.3 Total count of remaining rows

After the data wrangling there is an total ammout of 55579 rows remaining.
The count method was used to get this information.
To drop the rows, where no information of the gps(latidute and longitude) is given, the following code was used:

```
Y_axis = T[['latitude']]
X_axis = T[['longitude']]
result = Y_axis.isnull()
print(result)
T.dropna(subset=["latitude"], axis=0, inplace=True)
T.dropna(subset=["longitude"], axis=0, inplace=True)
# reset index, because we droped two rows
T.reset_index(drop=True, inplace=True)
T
```

## 4.4 Elbow Curve

The Elbow Curve was conducted to analyse, how many cluster should be used for the data.




## 4.5 Clustering

This remaining dataset was used to cluster the gps data (latitude and longitude). Kmeans was used to analyse the dataset (n_clusters=5, n_init=12). In the following graph the centers and the label of the cluster are shown.

| | latitude | longitude |
|---|---|---|
| count | 55579.000000 | 55579.000000 |
| mean | -111.755781 | 40.672947 |
| std | 5.372417 | 3.058414 |
| min | -118.457241 | -73.285003 |
| 25% | -111.915513 | 40.739895 |
| 50% | -111.899550 | 40.760534 |
| 75% | -111.879530 | 40.767992 |
| max | 88.421094 | 40.840264 |

| cluster_label | latitude | longitude |
|---|---|---|
| 0 | -111.904947 | 40.761073 |
| 1 | 88.421094 | -73.285003 |
| 2 | -118.457241 | 38.119653 |
| 3 | -111.866222 | 40.737924 |
| 4 | -111.966934 | 40.774278 |

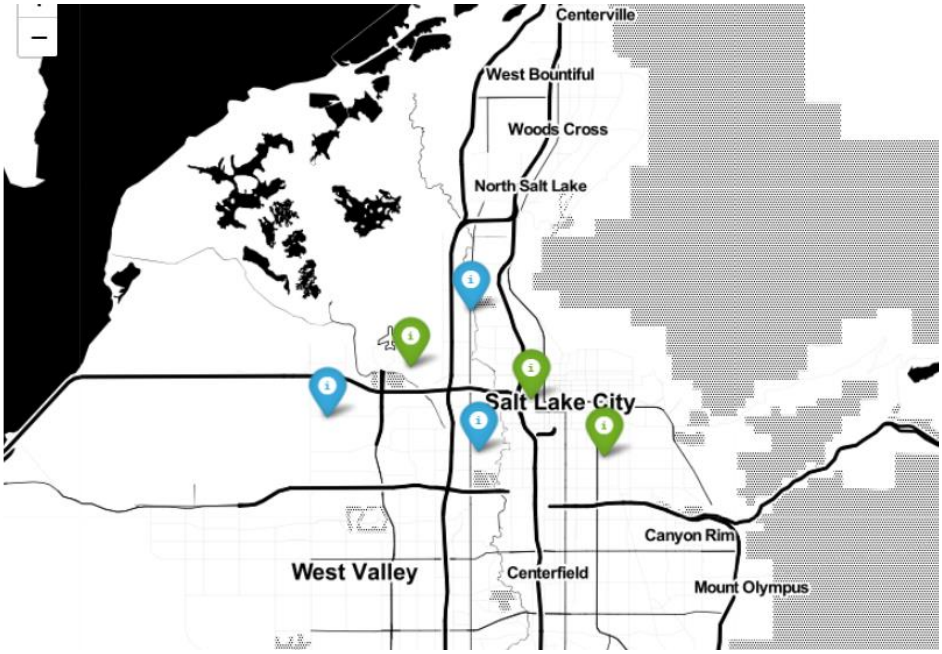
The cluster center with label 1 is very confusing. So i decided to have a closer look. In the next step you can see the count of crimes in each cluster.

```
0    30895
3    18052
4     6590
1       40
2        2
```

Cause oft he lack of information, due to very little crime cases in Cluster 1 and 2 i decided to drop them.

### 4.6 Foliuim

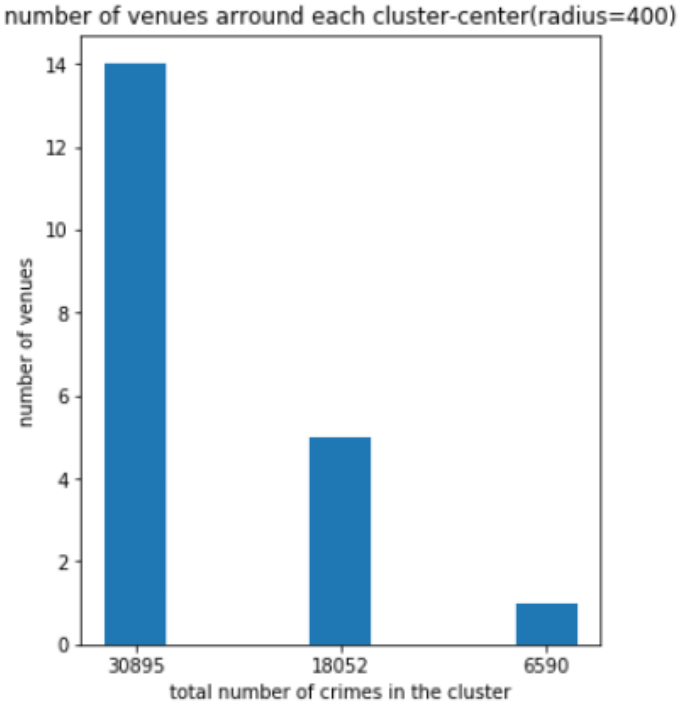
Via Folium the three centers of the cluster a displayed on a map (tiles= stamen Toner) in the color green. Three random created other spots in the city (marked in blue) are shown as well to compare these with the hotspots of crime.



### 4.7 Foursquare

To get informations of the sourounding venues ot the marked centers(radius = 400) in the map, foursqure was used.

| | crime | lon | lat | venues_around |
|---|---|---|---|---|
| 0 | 30895 | -111.906519 | 40.761767 | 14 |
| 1 | 18052 | -111.867350 | 40.738420 | 5 |
| 2 | 6590 | -111.971285 | 40.774869 | 1 |




There is a linear  positive correlation between the venues and the crime in Salt Lake City.

```
array([[1.        , 0.9828037],
       [0.9828037, 1.       ]])
```

In Comparison the three random created locations (marked in blue, folium) showed only a total of 6 sourrounding venuues.

## 5. Discussion and Conclusion

The results of this data analysis shows, that there is a corelation between crimes and venues. On the other hand there is no time dependency of crimes due to the day of week. The results should be treatend carefully, cause there is a limitation of cases and only the data of Salt Lake city was observed.

But this Analyse could be a good stimulation for further testing of different citys in the USA. It is possible that the limited ressources of the police men and women could be better splitted. Location where a lot of venues are should be treated more carefully and parts of a city with fewer

venues could be observed lesser. All in all the results show, that more scietific research of crimes should be done and with the use of mashine learning we can live in a safer environment.