Crime Analytics: Visualization of Incident Reports

Introduction

In this entry, I'm analyzing crime incident reports data from summer of 2014 in the cities of Seattle and San Francisco. We got the data from the first assignment, Coursera MOOC "Communicating Data Science Results" by the University of Washington.

The data was provided as comma separated value files for each city where each entry in the file is an incident report in the corresponding city. The data over these datasets doesn't complain in the schema, so preprocessing and data manipulation was required to obtain the data we require to perform this analysis. The scripts used to format the data according to our needs as provided to allow the reader reproduce our results. Note that some data manipulation operations were required to be performed manually, such as generate 2 "Comma Separate Value (CVS)" files to extract the "incident report date" columns from the files.

The visualizations generated in this entry was generated by using Excel. The spreadsheet used for this purpose is also provided.

Questions to be answered in this report

In this report, we want to answer the following questions:

- How many incidents are reported daily in each city?
- How many incidents are reported each weekday during the 2014's summer's period?
    - Can we compare each day?
- How many incidents are reported each month?

Reproducibility

This exercise is made to be reproducible. Feel free to get the scripts to reproduce the experiments described here from https://github.com/ulisesmx/data_visualization. Note that you might require a system with a Python interpreter installed as well as all the files from that github repository.

All commands must run from the directory you downloaded the repository.

First steps

As a first step, we generated two simpler CVS files, one for each city, with the dates of each report. We don't care about other data different from the report date. The name of these files will be "seattle_dates.csv" and "sanfrancisco_dates.csv" and they will only include the columns of the date of the incident. With this information, we will be able to apply our required process. To obtain those files, just run the following command:

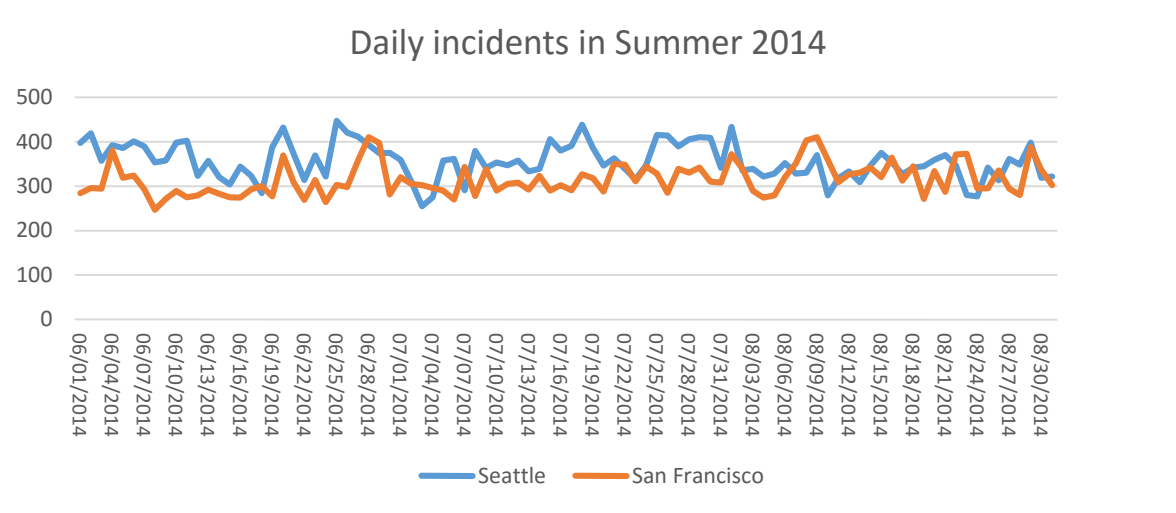
```
$ python get_dates.py
```

Daily incidents

Now, we will obtain the daily incident report for each city. We can combine the data to create a visualization that compares both cities in this period of time. Run the following commands to obtain the required “cvs” files to generate following visualization:

```
$ python count_day_incidents.py sanfrancisco_dates.csv > sfc.csv
$ python count_day_incidents.py seattle_dates.csv > seattle.csv
```

You can create a visualization with the information of files “sfc.csv” and “seattle.csv”.



In this first graph, we can see how the reports are generated daily. We can see that in general, Seattle has more crimes per day than San Francisco. In August, it seems like the crime incidents tend to be similar in both cities.

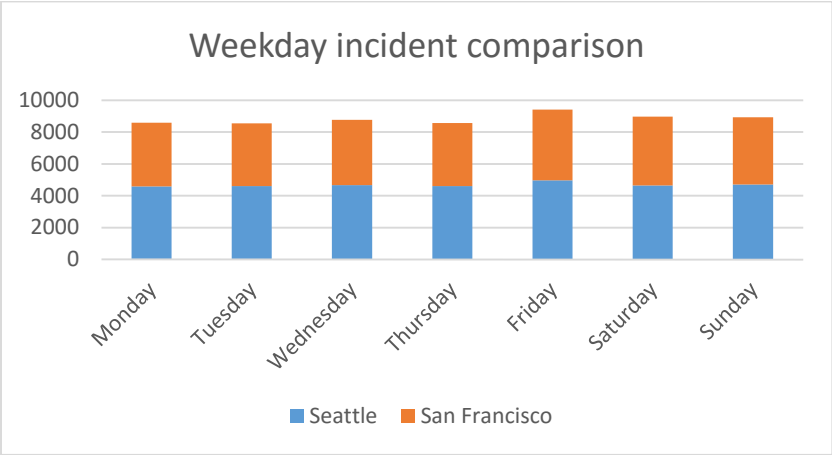
There is not too much information we can grab from this visualization, just a general comparison between both cities, and we can infer in which seasons the incidents are more common in each city.

Weekday incidents

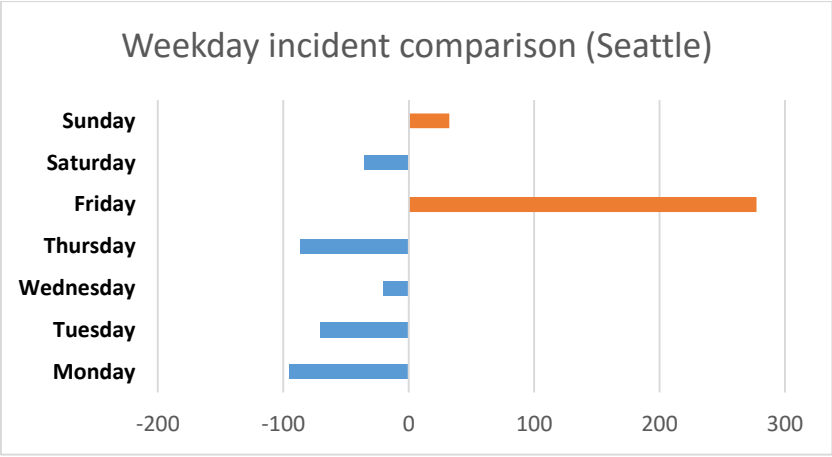
Now, we will obtain the daily incident report for each city. We can combine the data to create a visualization that compares both cities in this period of time. Run the following commands to obtain the required “cvs” files to generate following visualizations:

```
$ python count_weekday_incidents.py sanfrancisco_dates.csv > sfc.csv
$ python count_weekday_incidents.py seattle_dates.csv > seattle.csv
```

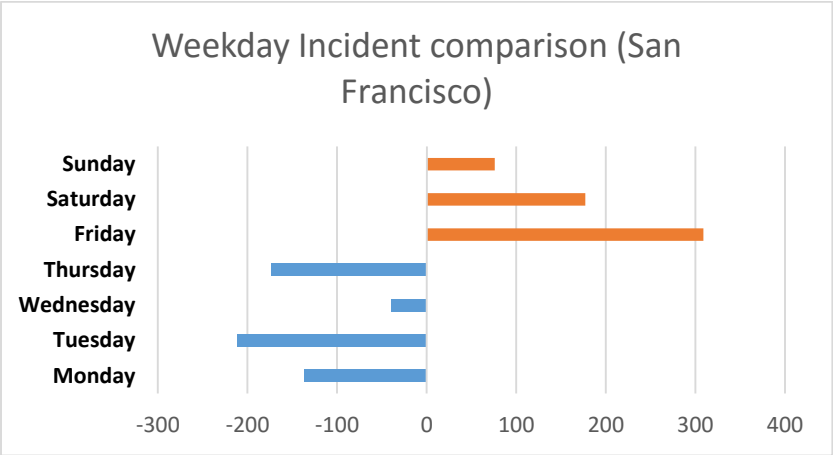
You can create the visualizations with the information of files “sfc.csv” and “seattle.csv”.



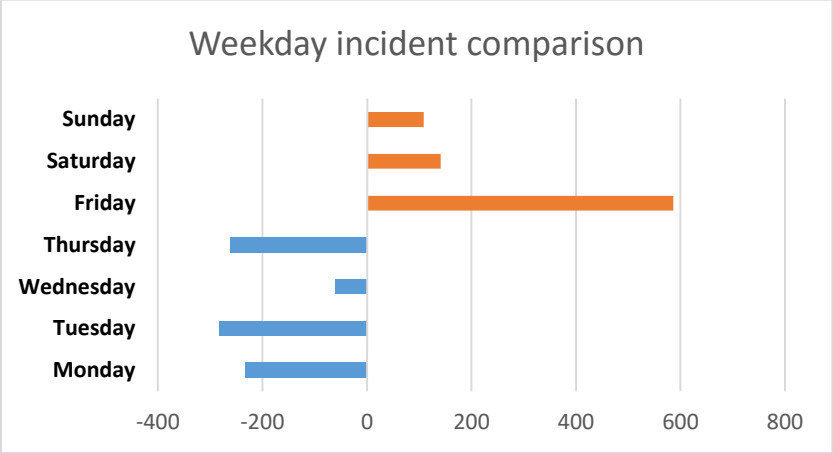
In this graph, we can compare the incidents at each week day. It seems like the most dangerous day is Friday, followed by the other weekend days (Saturday and Sunday). We can take a deeper look at this information by comparing the incidents in each day. In this case, we will compare the incidents of each city against the mean of each one in to generate visualizations of comparisons of each day.



In the case of Seattle, Sunday and Friday are the days that are above the weekday average, thus, these are the most dangerous days. Also, it seems like Wednesday it’s also more dangerous than other days.



In the case of San Francisco, the weekend days are the most dangerous days. Also, we can see that the other days are behaving similarly to their pairs at the Seattle analysis.



Finally, we compare the two cities average and it looks like the San Francisco graph, Having the weekend days as the most dangerous. This agrees with our first-weekday graph where the trend is that weekends are the most dangerous days, particularly the Fridays.

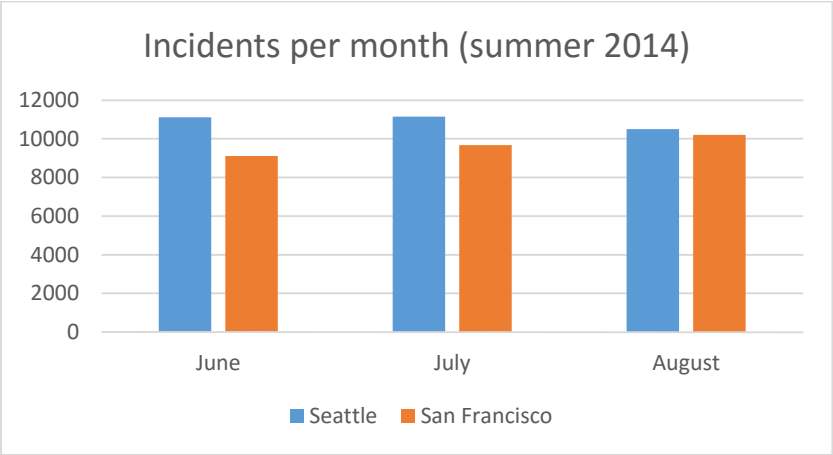
We can conclude from this that most of the reports are done during the weekend.

Month incidents

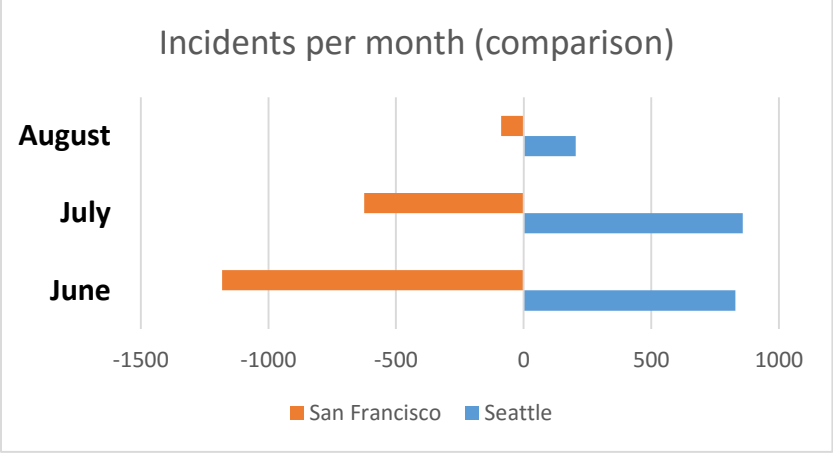
Finally, we will obtain the daily incident report for each city. We can combine the data to create a visualization that compares both cities in this period of time. Run the following commands to obtain the required “cvs” files to generate following visualizations:

```
$ python count_month_incidents.py sanfrancisco_dates.csv > sfc.csv
$ python count_month_incidents.py seattle_dates.csv > seattle.csv
```

You can create the visualizations with the information of files “sfc.csv” and “seattle.csv”.



In this visualization, we can compare each city incidents per month, where we can clearly see that Seattle has more incidents than San Francisco, as seen in the first visualization comparing the days. Also, we can see that August is a month where both cities have a similar number of reports. Let’s take a look on the month average per city to see how far and how different are each city:



In this last visualization, we can see clearly the difference between the report number in each city at each month. With this visualization, we can conclude that in august, it seems to have a more similar behavior in the incidents in both cities.

Final conclusions

The visualizations provided helped us to understand better the data, much more than the raw tables that we can get from automated tools. With visualizations, we can explore and understand the data in a better way than just provided numbers. This is because we are able to abstract data very well by using visual means.

Finally, generate visualizations is easy if our data is compliant with a schema that we know, but this is not usually the case. After we get the data, we have to understand how it is provided and how we can fit our purposes. This is usually the harder step. For this purpose, we provide the used scripts to generate the formatted data that we used to generate these graphs.