Finding Correlations between datasets using Dynamic Time Warping

Jeet Kamdar
NYU Tandon School of Engineering
New York, New York
jeet.kamdar@nyu.edu

Khushnaseeb Ali
NYU Tandon School of Engineering
New York, New York
kk3521@nyu.edu

Muhammad Hassan Farooqui
NYU Tandon School of Engineering
New York, New York
mhf303@nyu.edu

ACM Reference Format:
Jeet Kamdar, Khushnaseeb Ali, and Muhammad Hassan Farooqui. 2018. Finding Correlations between datasets using Dynamic Time Warping. In *Proceedings of Project Report (Big Data Analysis).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Cities are growing at a phenomenal rate, bustling with people and economic activity. This brings with it new sets of challenges in terms of providing basic services, public transportation, affordable housing, schools, infrastructures such as roads, bridges and highways. With an increasing amount of data that is being gathered, it is possible to leverage this to have a better understanding of cities and help in management and planning by city administrators.

It is possible to find patterns, both expected and unexpected, from the vast datasets available and experts can gain greater insights into how different aspects interact. They can generate hypothesis and theories from this, create policies and test them. While some relationships exist across all cities, some are unique to every city. Discovering these relationships can be difficult as these datasets have components that vary over space and time.

One challenge that exists with identifying the relationships across datasets is that certain relations exist only under certain extreme conditions, while no such relation exists under normal conditions. For example, the total number of daily trips falls when the temperatures are very low on some days of January but normal variations in temperature do not affect it. This pattern is expected as such extreme variations in weather affects human activity and can be seen through the urban datasets. We use Dynamic Time Warping to test and investigate these relationships that exist in the data over space and time at different resolutions.

2 PROBLEM FORMULATION

Given a collection of urban datasets, can we find out the correlations that exist among different pairs of datasets. If we are given two or more datasets, we have to find out if they have some attributes that are correlated which would imply that the datasets themselves are correlated. By *correlated attributes* we mean the following: Assume we have two datasets where dataset A has a set of attributes $\{a0, a1...am\}$ and dataset B has attributes $\{b0, b1...bn\}$. We will conclude that A and B are correlated if $\exists$ a $\in$ A which is correlated with $\exists$ b $\in$ B. In other words, if at least one attribute from each of the two datasets is correlated then the two datasets are correlated. This correlation metric is defined in more detail below.

3 METHOD

We propose a framework that uses Dynamic Time Warping (DTW) to calculate the similarity measure between two datasets. DTW is an algorithm that computes the similarity between two temporal sequences (which may vary in speed) using dynamic programming. We can use this algorithm to compute correlations between temporal datasets by ordering the columns in chronological order and then feeding pairs of columns to the dynamic time warping algorithm. The problem here is that DTW is not naturally normalized for inter dataset comparison, nor is it directly extensible to spatial-temporal data, which is why we use normalizations to make sense of the distance measure that DTW produces and make inferences about the similarity accordingly. The normalization works as follows:

$$\beta_{DTW}(X, Y) = 1 - \frac{DTW(X, Y)}{DTW(X, 0) + DTW(0, Y)} \tag{1}$$

So the data we feed to the DTW algorithm is *aligned* through a preprocessing phase. By that we mean that we pick a pair of datasets, we select the relevant columns we want to use in order to find the similarity measure, filter data from the dataset according to the same time range and then order the columns in chronological order and then feed the data to our DTW algorithm.

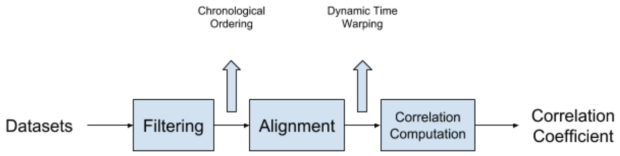
4 ARCHITECTURE AND DESIGN

Our framework has three main directories, src and Aggregated Data. The Aggregated Data directory has the processed versions of the raw files stored on hdfs. The Aggregated Data directory has a subdirectory for each dataset we have analyzed. For example it has a directory named Weather which has subdirectories that have different aggregated versions of the weather data for different granularities of time. Those files are processed using pyspark scripts stored in the src/Aggregation Scripts folder. This folder has scripts that use map reduce and spark SQL to aggregate the data. The DTW directory hosts the scripts that read the aligned files from the Aggregated Data directory and runs the similarity measure algorithm on them and prints values that range from 0 to 1 for all pairs of columns that belong to the two datasets. The DTW directory also has a plot.py script. This script plots graphs for each

Big Data Analysis, Spring 2018, New York, NY, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

pair of columns inside the specified dataset using the matplotlib library. This is useful to visualize the two datasets together and visually see if the correlation metric makes sense or not. It also helps us gain more insight into the trends hidden inside the data.



The DTWTest directory hosts the scripts that read the aligned files from the aligned-files directory and runs the similarity measure algorithm on them and prints a value that ranges from 0 to 1 which represents the similarity between the two columns that belong to different datasets.

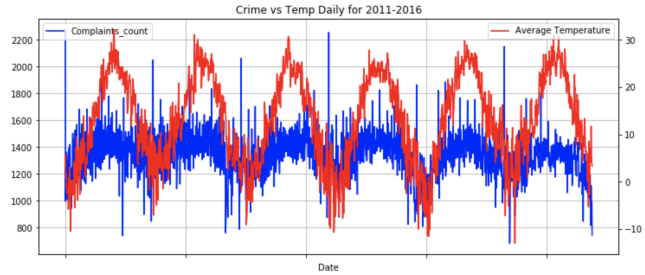
5 TECHNICAL DEPTH AND INNOVATION

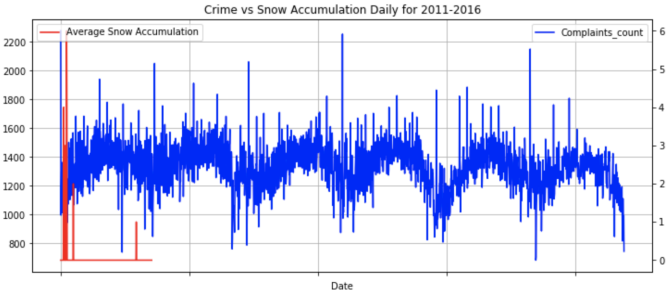
As indicated earlier, DTW is a dynamic programming solution. It has a $O(n^2)$ running time complexity which means that any reasonably large dataset can break it. So currently we are using the fastDTW implementation of the DTW algorithm which is an approximation of the standard DTW algorithm and it has a O(n) running time complexity. However this still limits the scalability of the solution to some extent so we have come up with a viable solution for this. We have decided to aggregate the data. For example for weather if we have temperature entries for every minute we can calculate the average temperature for each hour. Another example would be citibike trips. If we have an entry for each trip that happens, we can calculate the total number of trips for each hour and use that for our analysis. This would reduce the number of rows we would have to feed to our fastDTW algorithm and it would never break since we have control over the granularity of the data.

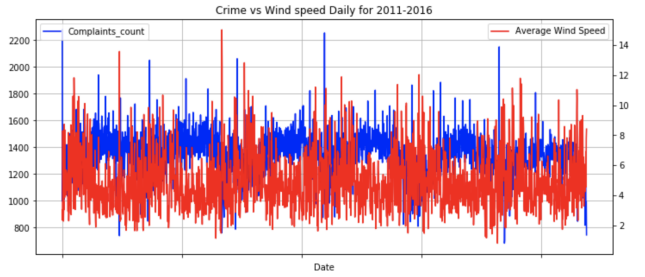
6 RESULTS

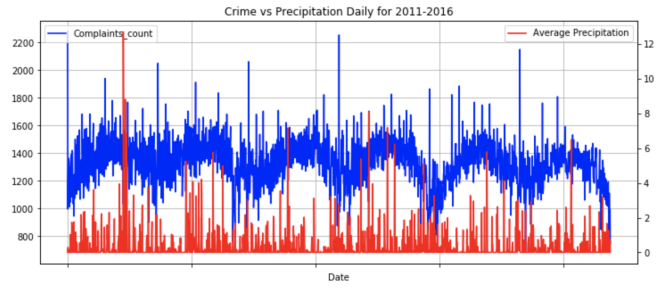
We picked seven datasets namely NYC Crime, taxi, weather, demographics, poverty, 311 and vehicle collisions datasets. Then we picked pairs of two from the datasets mentioned above. For each pair, we aggregated the two datasets according to various temporal resolutions(yearly,monthly, daily and specific to a month/year) and spatial resolution(borough wise).Post cleaning, these aggregated datasets corresponding to a particular temporal/spatial resolution are then fed to the fastDTW algorithm. If the correlation outputted by the algorithm was around or greater than 0.5, the datasets were declared to be correlated. We also plotted the graphs for each pair of datasets on various spatial and temporal resolution to confirm the same. Below are some of the pairs of datasets that we analyzed along with the summarized results of application of fastDTW algorithm to different pairs of datasets and corresponding graphs.

1. Crime and Weather Correlation Analysis: We tried to investigate if there is a correlation between crime and various weather attributes like daily temperature temperature, snow accumulation, precipitation and wind speed. Here we have used data from 2011-2016 only. We can see from the following graphs that crime is highly correlated with temperature and wind speed as compared to snow and precipitation.





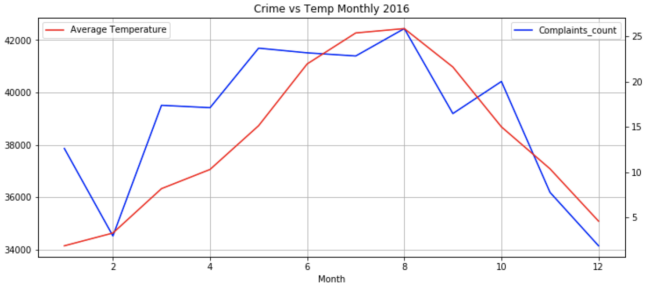




We derived some interesting observation from the graphs above. First, many crimes decrease as the temperatures get colder. Assault and murder complaints are highest in summer as compared to winter. Warmer temperatures bring out more aggression leading to more assaults and also there are more people outdoors thus increasing the numbers of potential victims. Second, burglary complaints

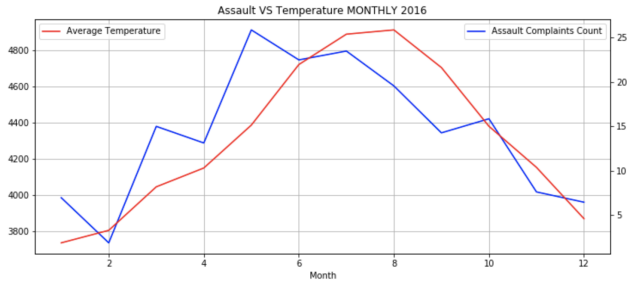
Crime and Weather From 2011-2016 Daily.

Weather Parameters	Total Complaints Registered
Avg. Temperature	0.62
Avg. Wind Speed	0.47
Avg. Snow Accumulation	0.14
Avg. Precipitation	0.39

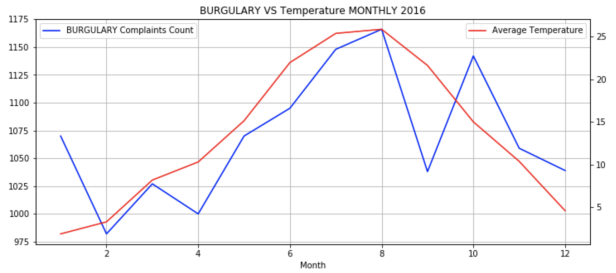


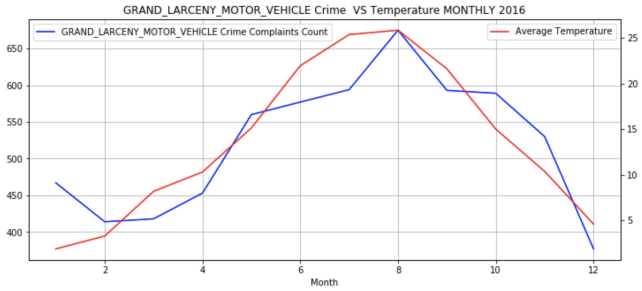
Crime Categories and Weather - 2016 Monthly

Crime Category	Average Temperature
Assault	0.62
Murder	0.75
Burglary	0.67
Vehicle and Traffic Crimes	0.64
Grand Larceny and Motor theft	0.79



are highest in summer as many people go on vacations leaving there houses attended while, also some people start to leave their windows open to try to cool off their houses. Additionally, they tend to leave the house more often to attend different summer activities. Both of these activities give burglars a perfect opportunity to enter the home. Third, while most crime rates seem to drop with the temperatures, there is one exception: grand larceny or motor theft. Car theft complaints actually jump when temperatures goes below zero. This is likely because car owners leave their vehicles running with the heat on to warm up their cars or defrost their windows. An unattended car that's running serves as the perfect opportunity for a car thief. Also while crime seems to spike when



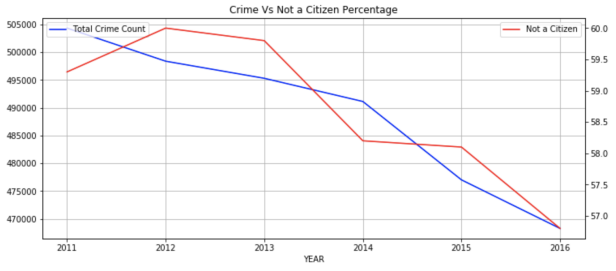


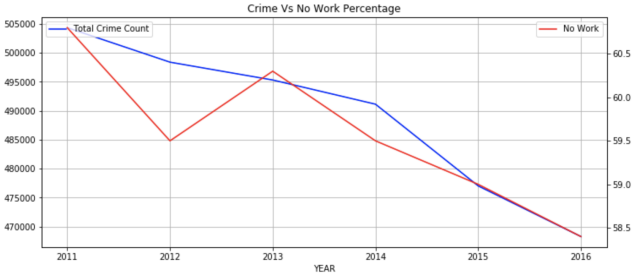
it's warm outside, if temperatures rise above 32 degrees Celsius, it starts to decline again because as temperatures soar to extremes, people seek shelter inside and spend less time outside.

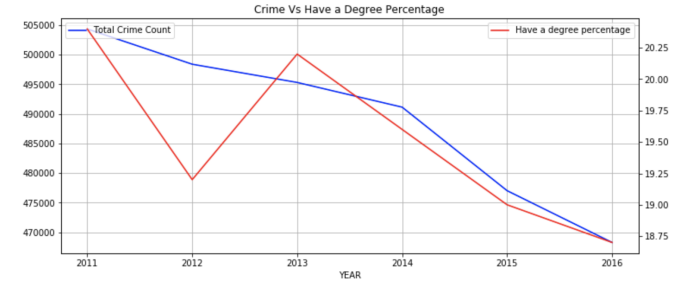
2. Demographics and Crime Analysis: Here we took the census data from the year 2011-2016. Since the census data is available only yearly, all the aggregations were done on yearly basis. We then tried to explore a relationship between crime and it's various categories with various demographic factors like Per Capita Income, unemployment, education immigrants' percentage using the fastDTW algorithm. The results are described below along with the graphs that model them.

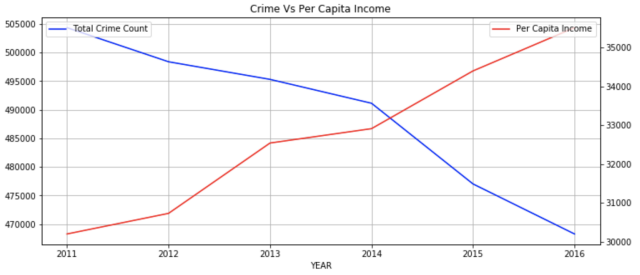
Crime and Demographics - 2011-2016

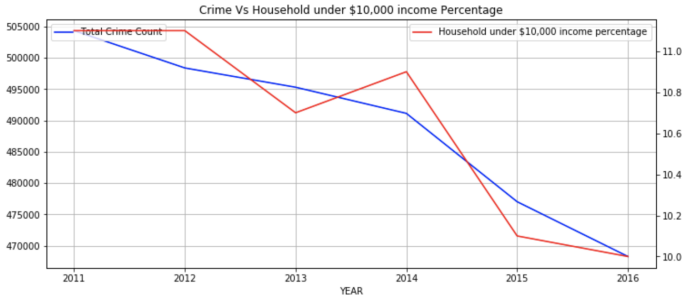
Demographic Factors	Correlation with Crime
Per CAPITA Income	0.02
No work Percentage	0.78
Not a Citizen Percentage	0.76
Full Time Work Percentage	0.65
Bachelor's Degree or higher Percentage	0.76
Household under $10,000 income percentage	0.84









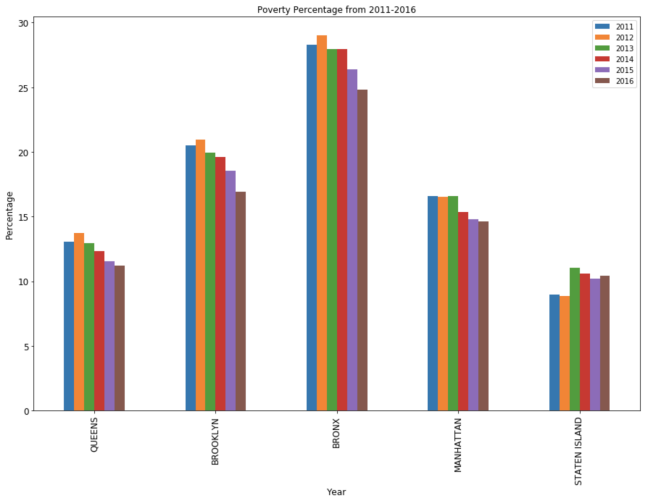


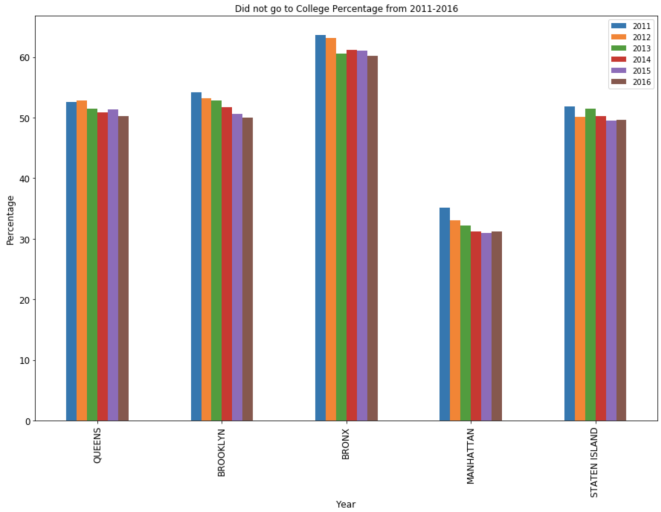
We can see from the graphs above that crime rate decreases with increase in per capita income, percentage of population with a degree and percentage of people having a full time work option. We can also see that crime rate decreases with a decrease in percentage of households under 10000 dollars income, percentage of immigrants and percentage of unemployed people.

3. Poverty and Crime: We analyzed crime and poverty data in both spatial and temporal resolutions to find some correlations. We can see from the following graphs that as the poverty rate decreases crime also decreases. Also there is a high correlation between burglary and poverty. We have calculated the correlation of various types of crimes with poverty and percentage of people who went to college.

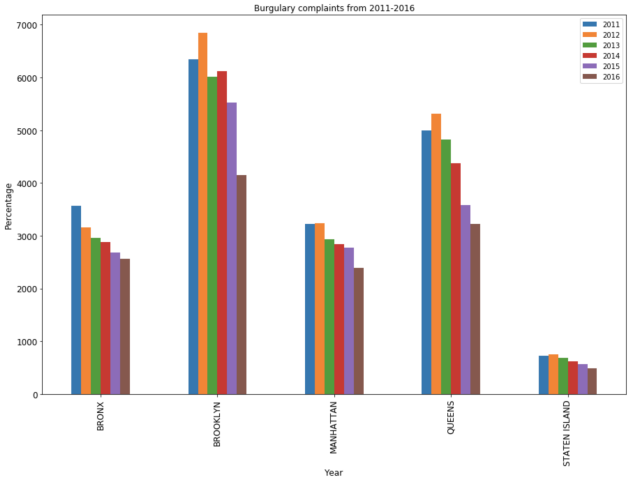
Crime Categories and Poverty - 2011-2016

	Crime Categories		
Borough	Assault	Burglary	Murder
Brooklyn	0.74	0.89	0.54
Bronx	0.24	0.64	0.56
Manhattan	0.30	0.77	0.76
Staten Island	0.37	0.22	0.63
Queens	0.79	0.94	0.78



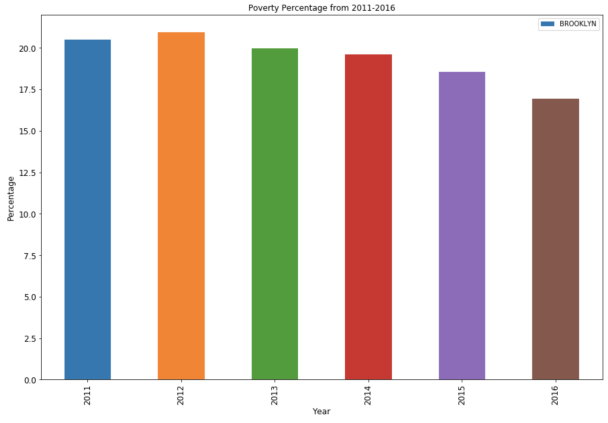


To dig deeper, we analyzed poverty and crime trends in a specific Borough - Brooklyn. We found the same relation here as well which is crime decreases with decrease in poverty rate. Also crime decreases as the percentage of people who went to college increases.

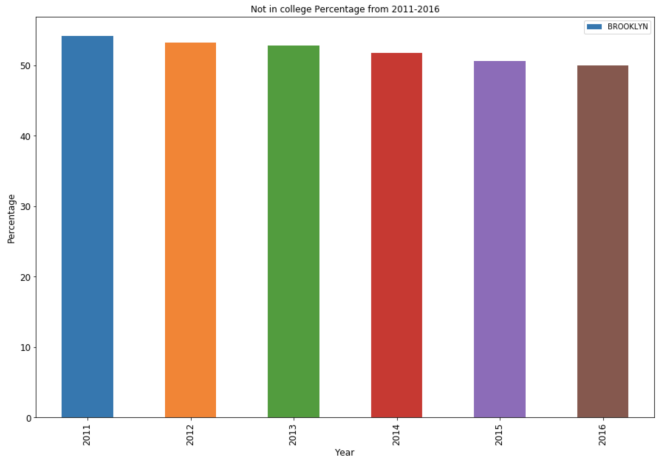


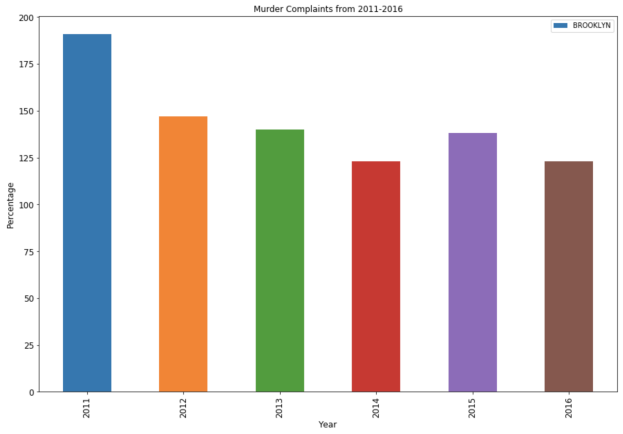
Crime Categories and Percentage of Population that did not go to College - 2011-2016

	Crime Categories		
Borough	Assault	Burglary	Murder
Brooklyn	0.73	0.70	0.69
Bronx	0.13	0.78	0.84
Manhattan	0.24	0.63	0.85
Staten Island	0.59	0.68	0.243
Queens	0.44	0.76	0.90

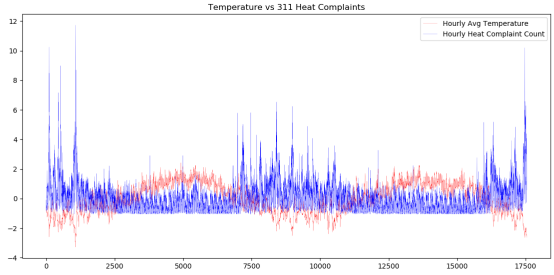


4. Taxi and weather: We noticed a very strong correlation (0.60) between the average number of trips every hour with the average hourly temperature. We hypothesize that because people would prefer to stay indoors and travel less when the weather is cold. While examining the relationship for extreme conditions, we found a drastic fall between the number of trips with very low temperatures. There was also a relationship between the average fare and temperature (0.59). We found no such relation between either weather and average tips (0.35) or with weather and trip distance (0.09). The average tips seem to be fairly constant and seems like the change in weather does not affect the way people tip.



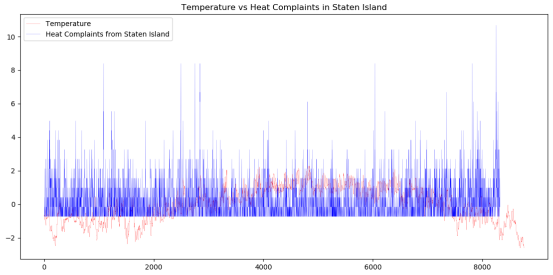


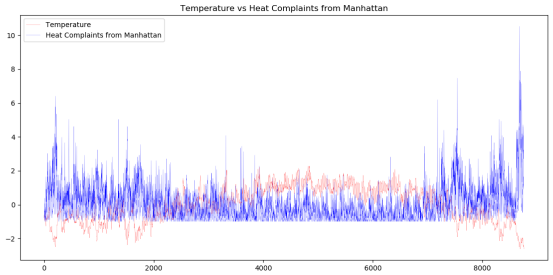
5. 311 and weather: We observed a lot of interesting correlations between weather and 311 complaints. We found a very strong correlation of temperature with heat and hot water complaints(0.62). The relationship is inverse in nature, the number of complaints



increases with a decrease in temperature. This makes sense as they are only required in cold weather. Also, this correlation is strongest in Manhattan(0.60) whereas no such relation is observed in Staten Island(0.44). Manhattan has a very high proportion of renters whereas Staten Island has a very high proportion of home-owners. This explains the correlation as a high number of renters means a higher number of such complaints. This relationship also

exists in Brooklyn(0.58), Queens(0.57), and Bronx(0.56) which also have a very high proportion of renters. We also saw a correlation between complaints about traffic signal and street sign condition with weather(0.52). Also, there was a stronger correlation in the outer boroughs than in Manhattan(0.40). The correlation is strongest in Brooklyn(0.54) and Staten Island(0.56).

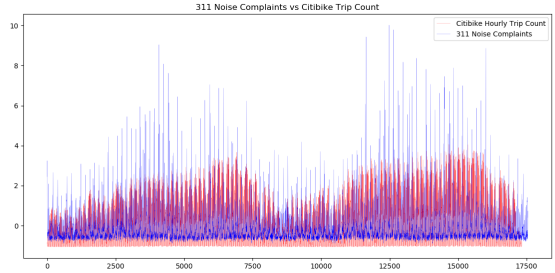




We also came across a relationship between the number of Lead complaints and weather(0.55), but no such relationship existed specifically for any of the boroughs. This is probably a false positive, likely because of the relatively small number of such complaints. Another relationship that was noticed was missed garbage collection and overflowing litter baskets complaints for Manhattan(0.50) but no such relationship existed for the entire city. This is likely due to the high population density in Manhattan.

6. 311 and Citi: We found a correlation between the heat and hot water complaints with the number of Citi bike trips(0.54). The relationship was inverse in nature, the number of Citi bike trips fell with the increase in the number of these complaints and increased with the decrease in the number of such complaints. The number of these complaints have a strong correlation with the temperature(0.62) as discussed in the previous section. This fact helps explain this correlation between the number of Citi bike trips and the heat and hot water complaints, that the number of trips fall with fall in temperature and increase back when the temperatures rise. This correlation was the stronger in Manhattan(0.52) and Brooklyn(0.54) than in Staten Island. This could probably be explained by the fact that we had not found a correlation between temperature and these complaints in Staten Island.

There was a very interesting relationship between the number of noise complaints and the number of Citi bike trips(0.50). There



seems to be a strong correlation of noise complaints with the number of Citi bike trips than with weather, and the relationship is direct, an increase in the number of bike trips corresponds with an increase in the number of noise complaints. We hypothesize that these are the same days when people stay out late in the night which leads to increased number of noise complaints from the streets and that people probably sleep in early during winter leading to lesser disturbances and noise complaints.

We also found a correlation between the complaints about the traffic signal and street sign condition(0.56) and the number of Citi bike trips. There is a curious relationship between these two. The number of these complaints are the highest right around the time when the number of Citi bike trips have increased and before they reach their peak for the year. We hypothesize that this is probably because that more Citi bikers might also mean more pedestrians and these are two are more likely to notice and make a 311 call than someone traveling in a car. We observed another relationship between the average city bike trip duration and the number of snow complaints(0.56) but this seems to another false positive because of the relatively very low number of snow complaints being made.

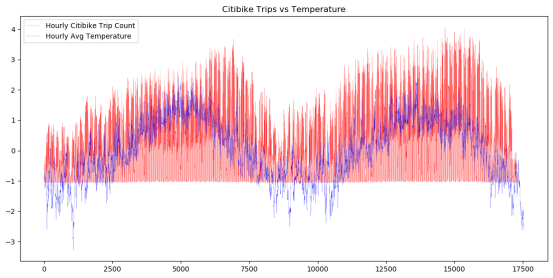
7. Taxi and 311: We found no relation between the number of taxi trips with 311 calls. But there was a relationship between the average fare and the animal abuse complaints (0.51). This is possibly because such complaints are highest during the day, and this is when the average taxi fares are the highest. There was also a relationship that was noticed between average fares and the noise complaints(0.50). On further examining the data, we found that an inverse relationship exists between the two, the complaints are made mostly during the late night hours which is when the average fare is relatively lower.

8. Taxi and Citi: There was no such relation that we observed between the two datasets. We had expected to see a direct relationship between the number of Citi bike trips and the number of taxi trips(0.40). This could be because they have different peak hours, Citi bikes plot has dual peaks with one around 9 am and another one around 6 pm, this is when people commute to work, and then falls during the other hours. The taxi trip count on the other hand generally has a single peak at around 3 pm. Also, Citi bike riders are more likely to be affected by changes in weather which further disrupts the relationship.

9. Taxi and collisions: No significant relationship was observed between the two datasets. We had expected to find a relation between the collision count and the taxi trip count, but this could probably be explained by the fact that increased taxi activity during peak hours also corresponds with increased traffic. With the traffic moving slowly, there are probably fewer chances of an accident.

10. 311 and collisions: We came across an interesting correlation between complaints about the street sign and traffic signal condition and the number of accidents(0.54). A further analysis of the spatial data is required to ascertain if any of these were the cause of the accidents and to find out if a relationship actually exists between the two or that these are completely unrelated events merely happening during the same hour of the day. There was also a strong correlation between the number of lead complaints and the number of deaths due to accidents(0.66). We believe this is a false positive because of the low number of lead complaints.

11. Citi Bikes and weather: We observed a very strong correlation between the number of Citi bike trips and the average temperature(0.63). Higher temperatures are more favorable to ride a bike and this can be observed from the direct relationship between the increase in the number of Citi bike trips with an increase in temperature and correspondingly decreasing during unfavorable weather conditions.



12. Citi Bikes and collisions: We found a relationship between the average duration of Citi bike trips and the number of accidents(0.54). We believe this is because during the daytime, with higher traffic volumes and work, the Citi bikers would keep their rides short compared to later in the evening when they are more leisurely. Higher flowing traffic would also mean a greater number of accidents.

7 CHALLENGES AND ISSUES

1. Alignment issue The issue with ordering on the basis of time was that every dataset had different date and timestamps formats. This would mean that if we tried to order the datasets on the basis of their own date formats, the alignment would be wrong. Because the ordering is done on the basis of string comparisons. So during aggregation we used the python dateutils library to parse the dates and convert them into a single format. This would allow us to align datasets without errors.

2. Null and Junk Values: A lot of columns had null values and junk values like 9999. We had to carefully examine and deal with those values.

3. Different values for the same column name: Certain columns which refer to the same variable had different representations in different datasets. For example borough column in Poverty data had numeric values (1,2,3,4,5) where as the same column in crime data had string values (“Brooklyn”, “Bronx”). We had to transform the datasets to have a common representation.

4. Different encodings for different years for the same dataset: The encoding for some of the years for the taxi dataset uses utf-8 while for others it uses utf-16. This brings its own challenges like the different ways in which NULL value is represented.

5. Some columns exist only in a few years while they do not for the other years for that dataset. Working with these columns, analyzing and finding correlations can be a challenge when they do not exist over an extended period of time.

8 CONCLUSION

The objective of this project was to use dynamic time warping to find correlations between datasets. We were able to verify that weather has correlations with almost all other datasets. This however was a bit obvious from the beginning so we dug deeper and found some interesting relationships for different boroughs like that in Citi bikes and 311 calls about the difference in Heat and Hot Water complaints between Manhattan and Staten Island or that of grand larceny/motor theft with temperature. Along the way we got to understand the strengths and weaknesses of DTW. DTW was able to identify relationships when there were wave like trends in the datasets. On the other hand, if the the columns had “almost linear” trends and if the rate of change of one trend was negative and the it was positive for the other trend, DTW could not find that correlation even though a negative correlation does exist. An example is of “per capita income” and “crime”. Plotting graphs show that a clear negative correlation exists among the two. Another problem with DTW is that if the data fluctuates a lot and no clear trend exists, DTW tends to give very high correlation values for such data. For example in the case of “vehicle collisions” where the values for hourly injuries and hourly deaths varies a lot, DTW gives very high correlations of this dataset with every other dataset. To summarize, even though DTW gave very good results in most cases, a different correlation measure should be used for cases where it falls short like the ones mentioned above.

9 FUTURE WORKS

We have solved the issue of scalability using aggregation of the huge datasets since the original DTW algorithm is not scalable. Another way to achieve scalability would be to write a scalable version of DTW using map reduce. That would allow any number of input rows to be processed without having to aggregate it. This would be beneficial if aggregation leads to loss in information. For example if instead of using wind speed for each minute we aggregate it to get the average wind speed for each hour or day, we would lose the variations that happen inside of that hour or day. In other cases aggregation might be necessary. An example of that is of taxi trips. The taxi dataset has an entry for every trip that takes place. If we

want to find the correlation between the number of taxi trips and temperature we would have to aggregate those rows to get the total number of trips within a time frame. So in conclusion both aggregation and a scalable version of DTW is necessary.

Another nice addition to our framework would be to add the functionality to create a graph if the correlation between two columns would be 0.5. We could give a deviation d from 0.5 such that if the correlation c between two columns is 0.5-d < c < 0.5+d then a graph is created for the user to look at and analyze.

Another thing we want to do is to make the framework such that if it is given two datasets, it computes the correlations directly without having to write the alignment files. This would be possible if we can write a generic aligner script which can figure out which of the columns are numerical. The problem with that is that all the numerical columns inside a data are not always relevant. For example a lot of the files have location coordinates which do not make sense to be used for computing correlations. This makes this task rather difficult since dataset cleaning also has to be automated in this case. But this is something that would be nice in the future.

10 CODE REPOSITORY

The code can be found at following the following github repository `github.com/muhammadhassaanfarooqui/BigData`. The repository has three section - src, plots and aggregated data and Json-To-DataFrame-Converter. src contains is again composed two folders - aggregation scripts and correlation analysis python notebooks/scripts. Aggregation scripts folder contains various scripts to aggregate the original datasets into various spatial and temporal resolutions which are then stored in the folder aggregated data. Json-To-DataFrame-Converter which contains the code to read a json file into spark dataframe. Correlation analysis notebooks contain the code for determining the correlation between two datasets using FastDTW and plotting graphs.Plots directory contains various plots providing a visual representation of the relationship between features of two datasets.

REFERENCES

[1] S. Salvador and P. Chan. *Toward Accurate Dynamic Time Warping in Linear Time and Space.* Intell. Data Anal., 11(5):561âĂŞ580, 2007.
[2] E. Keogh and A. Ratanamahatana. *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.* Information, communication and society, 15(5). 662-679.
[3] Boyd, D., & Crawford, K. (2012) *Everything You Know About Dynamic Time Warping is Wrong. Workshop on Mining Temporal and Sequential Data.* 2004.
[4] `https://opendata.cityofnewyork.us/`