Final report: Correlation analysis on spatiotemporal data

Heejong Kim (hk2451)
Cool name pending team
heejong.kim@nyu.edu

Jun Yuan (jy50)
Cool name pending team
junyuan@nyu.edu

Jiin Nam (jn1664)
Cool name pending team
jn1664@nyu.edu

KEYWORDS

Data cleaning, Correlation analysis, Mutual information, Spatiotemporal data

1 INTRODUCTION

In this report, we introduce the pipeline of correlation analysis based on the spatiotemporal dataset. Our project follows the steps of data cleaning, Data aggregation, correlation calculation, visualization and analysis (Fig. 1).

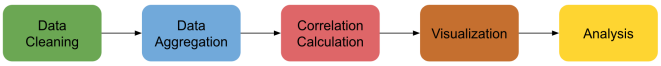

Figure 1: Pipeline.

Problem. As pointed by Erwig et al.[11], spatiotemporal data is the data which labels have been added showing where and when they were collected. And the research on the spatiotemporal data can even trace back to nomadic tribes who use the "data" to schedule their seasonal migration and grazing activities. In the modern world, we have more and more complex spatiotemporal datasets of which the urban data take a large proportion. Scientists and analysts keep studying on that for better urban planning, resource distribution, and environmental monitoring. For example, Becker et al used cellar network data to analyze data flow and help urban planners to under city dynamics[2]; Castro et al. used taxi GPS traces build urban traffic models and predict road capacity[3].

In this project, we focus our research on the urban datasets related to New York City so that we can better understand the development of this city and aid the planners to make better decisions. However, with the development of data storage and recording, we are facing challenges of dealing with more and more complex data. In the following parts, we will demonstrate what challenges we face and how we solve them, as well as the method we use to analyze spatiotemporal data. In the end, we illustrate some case studies to make sense of the results we get and evaluate the work we have done.

Contribution. In terms of big data, we often face a lot of data issues. So we target the main data quality issues in the 7 datasets we process. And then we demonstrate the rules to do the data cleaning and aggregation which can be a guide for people who deal with big urban data. In this project, we calculate mutual information both within a dataset and across different datasets. The correlation analysis based on the data processing and facts really help to make insights on how the city functions and make better decisions on urban planning.

2 RELATED WORK

2.1 Big Data Mining

To deal with the big data, Wu et al.[20] summarize the challenges in big data mining and proposed a big data processing model. In our project, we use **Spark** which provides us many powerful and expressive functions to process all our cleaning and calculation works. We also use **Dumbo**, the Hadoop Distributed File System, to store all our big data and codes.

We need to process really large datasets in this project. Let us use *311 Complaints* data as an example, it has over 17.6 million rows and 41 columns where there are different data types and ranges. Since we have so many columns and rows, we have to well organize the records and pick up useful information we want. MacAfee et al.[12] highlight the importance of data cleaning and organizing skills. Furthermore, they point out the increasing values of visualization and related techniques. In the paper from Rahm et al.[16], they summarize the main problems for data cleaning, including single-source and multi-source problems, as well as current approaches. For the single-source problems, we target some data quality issues in the following data cleaning section and introduce the corresponding solutions. And for the multi-source problem which mainly related to the naming and structural conflict, we include it in the data aggregation section.

2.2 Correlation Analysis

Correlation analysis plays an important role in big data analysis. The information gathered from big data engine provides insights that how we can make our lives better by studying the correlations between the features. For example, based on the correlations between spatio information and the search keywords, Google Flu Trends[5] helps to prevent the spread of flu. Also, you can make sense of the way to the career success by researching on the correlations between the objective/subjective success and a wide range group of working factors[13].

To understand the correlation, we can use correlation coefficients, such as Pearson correlation coefficient and its variants which summarize the association. There are still some discussion and analysis on correlation measures in recent years. Song et al.[18] compare several correlation measures including mutual information, correlation coefficient, and model-based indices. Steuer et al.[19] present that mutual information and the Pearson correlation have an almost one-to-one correspondence when measuring gene pairwise relationships within their investigated datasets. In our project, since most of the datasets we process are categorical data, it is more convenient to calculate mutual information. Therefore, we analyze the correlation based on the pairwise mutual information.

In addition, there are some helpful visualization tools to detect the correlations and outliers among data, like Table Lense[17]. We

	311 Complaints	Weather	Collision	Property	Taxi	Bike	Crime
Data Incompleteness	9	0	6	3	1	0	0
Data Incorrectness	1	0	1	1	1	1	1
Data Inconsistency	2	1	2	2	2	2	2
Data Redundancy	13	0	5	4	5	5	2

Table 1: Number of columns for each data quality issue in the datasets we process.

generate a lot of visualization results which can be checked in the supplementary materials and include 3 case studies in this report.

3 DATA CLEANING

We focus our analysis on New York City (NYC) related data in this project. The datasets we use include *311 Service Requests from 2010 to Present*[7], *NYPD Motor Vehicle Collisions*[10], *NYC taxi*[15], *NYC Weather*[14], *Citi bike trip*[8], *Crime*[9], and *Property price*.
To analyze these large dateset, we have to clean some noises and keep the data in a well defined form. So we target several data quality issues for all the datasets at first. We use 311 dataset as an example, the data quality issues are below:

- *Incompleteness.* In the dataset, some columns have mostly incomplete properties, such as the column 'Landmark', 'Vehicle Type', and other 7 columns have over 90% cells are null. There are also some cells with values like "None" or "Unspecific". In some cases, like the 'Vehicle Type' we have mentioned, we cannot get the information from other columns. Since the information in these columns is so incomplete that cannot contribute useful knowledge to the final correlation result, we could just drop it.
  However, in some cases, we can restore the value according to the values in other columns. For example, if the record has a missing ZIP code, we can still know the location by the latitude and longitude (GPS), and then map the ZIP code to the location. So for the latter case, we use ZIP code mapping functions to restore the information which is demonstrated in the following aggregation section.
- *Incorrectness.* With the increasing of the data volume, there can be some incorrect information in the table. For example, there can be some weird content like '999' for an address. And for the NYC data we used, the ZIP code listed in the column "Incident Zip" includes some numbers which are not valid ZIP code in New York City.
- *Inconsistency.* For the same content, there can be some different formats of expression, like "YYYY-MM-DD" and "MM-DD-YYYY" for a date, or "xxxxx" and "xxxxx-xxxx" format for ZIP code. The inconsistency of format could cause wrong results in the end.
- *Redundancy.* Attributes like *zip code*, *latitude*, *longitude*, *address*, *street name1*, etc. are describing the same information which is "location". They just provide the information at a different granularity or from a different angle.

As shown in both the Table1 and Fig.2, the issues exist in most of the datasets.

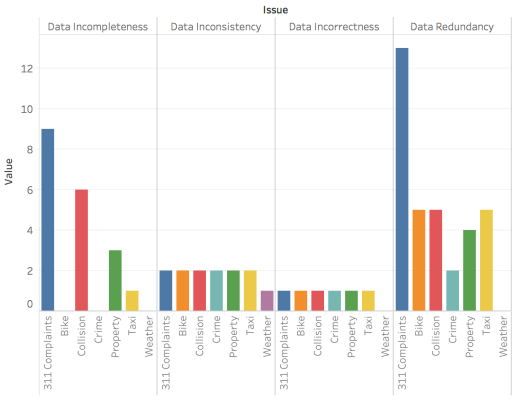

Figure 2: Data quality issues exist in most of the datasets.

3.1 Incompleteness

To clean the incomplete data, we list possible expressions which indicate a missing value, including "None", "Unspecified", "N/A", or just a null value. Some values like null are just empty values. And the value like "Unspecified" does not represent empty, it is a valid value recorded by officers when it is hard to define a specific issue type. So we also consider "Unspecific" as a symbol of incompleteness.
For each missing value we list above, we count the times they appear in each column, if this column has over 90% of missing data, we could say this column is just not useful for the correlation because no matter how other features change, it stays just a missing value. And to figure out the incomplete data we can restore later, we check whether all the related information is empty. If the record has empty values in all the ZIP code, latitude, and longitude fields, it is useless; but either it has a ZIP code or the coordinate information, we can still process it.
As demonstrated in the Fig. 2, we have most incomplete columns in the *311 Complaints* dataset, while *Crime* and *Weather* have relatively complete records.

3.2 Incorrectness

Incorrect information can be classified into 2 categories. The first category is obviously incorrect data, like the '999' address and "99999" ZIP code; the second is hidden incorrect data which mainly comes from location information outside New York City.
To clean the first type of incorrectness, we check the distinct values for the attribute and spot the meaningless values as outliers. As for the second type, because in this project we only analyze the NYC data, the record with non-NYC location does not need to be considered. So we use a list of zip codes in New York City[6] for

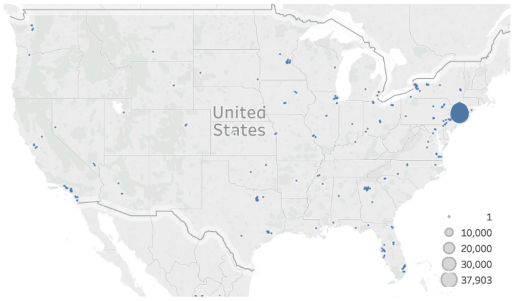

Figure 3: Zip code distribution in the U.S. For 311 dataset used for this report, most of the ZIP codes fall in the New York City. After explicitly deleting some obvious incorrectness like '99999', we still get several outlier ZIP codes which are valid but cannot contribute more to the correlation analysis as the NYC ZIP codes do.

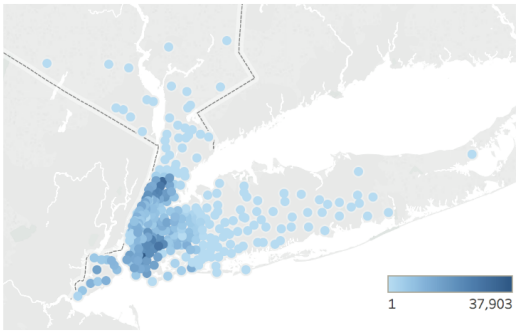

Figure 4: Zip code distribution in New York City. The ZIP codes in New York City have more complete properties in the datasets.

validation. After figuring out anomalous values identified by the 2 categories, we delete the records with those anomalous zip codes. The Fig. 3 shows the ZIP code distribution across the country before cleaning, and the Fig. 4 shows the ZIP code distribution only in New Tork city after cleaning the incorrect records.

3.3 Inconsistency & Redundancy

And when it comes to the inconsistency, we can use some format functions to regulate the format. For example, the date can be processed by the corresponding "DateTime formatting function" in spark. We used the function to change all the date in the form of "MM/DD/YYYY". In addition, because we want to compare the attributes in different datasets by year or by month, we generate new attributes *year*, *month*, *day* for correlation calculation based on different date granularity. And for the inconsistency for time, we have time in the form 24-hour clock and 12-hour clock. After data processing, we change all the time into the 24-hour clock form and keep only hour information.

Among those attributes describing similar information, we keep only one of them. For example, we only use *zipcode* for location. As in the Fig. 2 shows, in each dataset, we should deal with the inconsistency related to zip code and data, and redundancy related to the location information.

4 DATA AGGREGATION

After data cleaning, we can filter some noises. But we still need to do data aggregation for better comparison.

4.1 ZIP Code Mapping

As demonstrated in the previous section, we face the challenge related to the location granularity inconsistency. In *311 complaints* data, we have both ZIP code and GPS information. However, we only have GPS information for the taxi data, crime data, and citibike data. And we only have zip code information in the left datasets. So we apply the crossing number algorithm to convert GPS points to ZIP codes to support diverse space resolution for that cases such as *NYC Taxi* data. The algorithm checks whether a point is in a polygon by counting the number of edges of the polygon meet with a line connecting the point and the other point places out of the polygon. Chirigati et al. also use this algorithm to deal with different location resolution in the data polygamy framework[4].

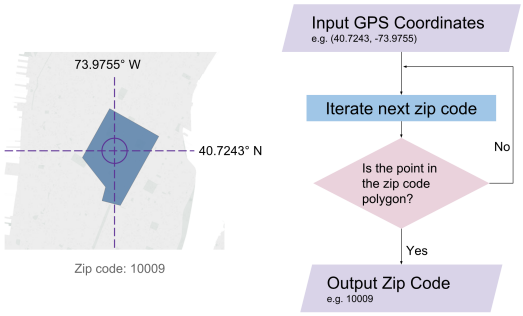

Figure 5: Zip code mapping algorithm. We check iterate all the zip code polygon and check whether the GPS point is in the polygon.

As demonstrated in the Fig. 5, with the given polygon set for ZIP code, the algorithm can allow checking if a GPS coordinate(point) is in any of polygons representing ZIP codes. If it is, our algorithm converts the point to the mapped ZIP code. Table. 2 shows mapping instances from the NYC taxi data. In our project, we apply our mapping algorithm to *NYC taxi*, *Crime*, and *Citibike* data which does not have the zip code attribute.

4.2 Transformation & Categorization

For the attributes like date and time, they are structural and contain more detailed information inside. So we extract new attributes like *Year*, *Month*, *Day*, *Hour* for later filtering operations. And we ignore attributes like *Minute* and *Second* because they are too detailed and

GPS to ZIP code mapping		
Longitude	Latitude	ZIP code
-73.916661142	40.828848333	10456
-73.784556739	40.697338138	11433
-73.740890627	40.714806463	11429
-73.975517666	40.724369992	10009
-73.873141669	40.753770398	11372
...	...	...

Table 2: Results converting GPS to Zip code using Crossing number algorithm

it is hard to find similarities on them among the data. Moreover, we transform *pick up time* and *drop off time* into *trip duration*, and *year* into *age* since we are more interested in these new attributes. We also meet some name conflict cross different datasets so that we need to transform the attributes to avoid errors in future processing. For example, the *Year* in some dataset only represents the year information; while the *Year* in the *Weather* dataset is in the form of "yyyy/mm" which includes both year and month information. In this case, we need to change the attribute name from "Year" to "Date" in the *Weather* data.

In addition, due to the correlation measurement we choose, we need categorical data for the mutual information calculation. However, a few attributes have continuous values such as *trip distance* in *NYC taxi* data, and *temperature* in *Weather* data. So we plot the distinct values and linearly map them into segmented and floored integers each represents a range of values. For example, a passenger whose *age* is *20* can be 20, 29 years old or some age in between.

5 CORRELATION CALCULATION

Our team's designated correlation measure is "Mutual information". The formal definition of the mutual information is:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}), \quad (1)$$

where $p(x, y)$ is the joint probability function of X and Y and $p(x)$ and $p(y)$ are the marginal probability from X and Y each. X and Y represent each attribute in our analysis. The marginal probability is count number of rows having the same category divided by the number of rows in the attributes. The joint probability is obtained by counting the number of rows having two attribute's categories together.

We join the different tables based on date and generate the mutual information matrices by year. In our project, we consider both the mutual information among attributes in each dataset and attributes across different datasets. The higher the mutual information is, the more correlated the two attributes are.

6 RESULTS

For each pair of datasets, we generate mutual information matrix by year. We also make the line charts to show the mutual information change by year. To evaluate the results we generate, we want to explain the results by facts. Here we look into some cases and find the interesting stories behind the data.

In the matrix charts as shown from Fig. 6 to Fig. 10, x-position and y-position represent different attributes. And the color at the coordinate (a, b) shows the mutual information value of the two attributes a and b. Here we use color intensity to demonstrate different values sequentially as shown in the legend in each matrix chart. As for the line chart shown in Fig. 11, different color hue represents different attribute pair. And we use x-position to encode the year change and y-position the mutual information values.

6.1 Case Study 1: Complaint

Complaint & Collision . The correlation results of Complaint attributes themselves show recognizable patterns as shown in Fig. 6 *street name* and *zipcode* attributes have high correlation to *complaint type* which implies different locations may have different complaint types. For example, "COLONIAL AVENUE" has more complaint type called "Street Condition" signifying that there could be recurring problems to cause bad street conditions or people near the street tent to be more sensitive about the street condition than others.

Complaint attributes related to locations tend to be highly correlated to most of the collision attributes. It can be induced based on common sense regarding traffic accidents. Traffic accidents, in most cases, cause noise, bad street conditions, or blocked driveways. People around the areas occurred the accidents are more likely exposed to inconvenient situations leading to more complaints than a normal situation.

The overall correlations between two datasets have been decreased year by year from 2014 to 2018. There could be two possible explanations. First, the way of handling the accidents has been getting professional and fast so that people around the areas suffer less inconvenient than before since there have been better insurances or services dealing with the accidents. Second, understanding of the unpleasant situation caused by traffic accidents has been getting raised in general and occur fewer complaints about it.

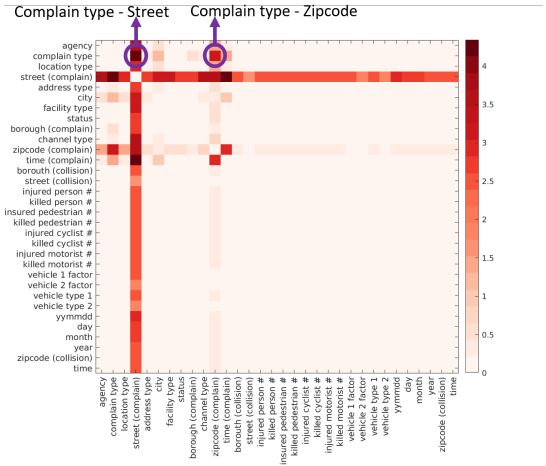

Figure 6: Mutual information matrix in 2018 between *Complaint* and *Collision*

Complaint & Crime. In the results for complaints and crime, we can find *street* and *zipcode* correlate to most other attributes. This

makes sense because in different locations there are different major complaints and crimes. We use the mutual information matrix in 2013 as an example as shown in Fig. 7.
The high correlation between time and street is also interesting. One possible reason for this can be that during the daytime, there are more complaints and crimes in the business area; while there are more complaints and crimes during the night in the residential area.

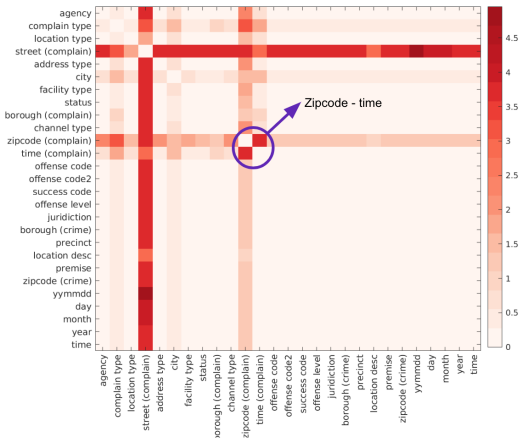

Figure 7: Mutual information matrix in 2013 between *Complaint* and *Crime*

6.2 Case Study 2: Weather

We can see many high correlation pairs from weather attributes. In this case study section, we show "Weather-Collision", "Weather-Complain", and "Weather-Bike".

Weather Collision. Collision data has attributes related to how many people and vehicles involved in the collision and weather data has attributes such as wind speed, temperature, and precipitation. Correlation matrix Fig. 8 is an example correlation matrix from collision-weather of 2013. From 2013 to 2017, the correlation results have same features. First, the weather has a high correlation with overall collisions. If you look at the Fig. 8, we can see the warm colored lines. The lines are time attribute and temperature attribute from weather data. The lines explain that all attribute from collision and weather has a relationship with the attribute. From the fact that we found from the relationship, we could find an interesting post which explains the relationship between temperature and accidents [1]. The study found that accidents increase in winter, but days with temperature above 80^oF increase fatal accidents in comparison with days of 50^o-60^oF.

Weather Complain. We find same lines from the correlation matrix of "Weather-Complain". The correlation results between weather and complain have similar results overall from 2011 to 2017. Fig. 9 explains that the weather attributes are related to overall attributes from complaint and weather attributes itself. Mostly they are related to time and temperature. We can intuitively understand the result since cold weather will result in complaints about heat and hot weather will make A/C complaint.

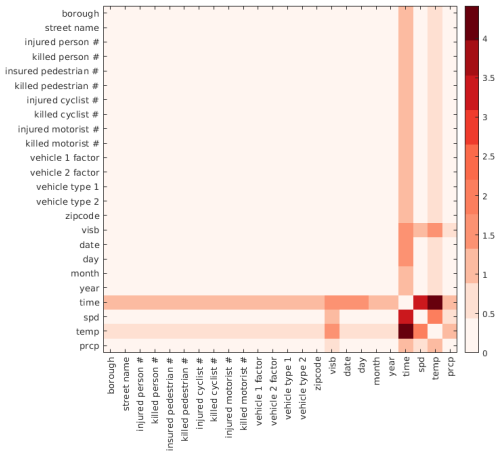

Figure 8: Correlation result between attributes from collision and weather in 2013

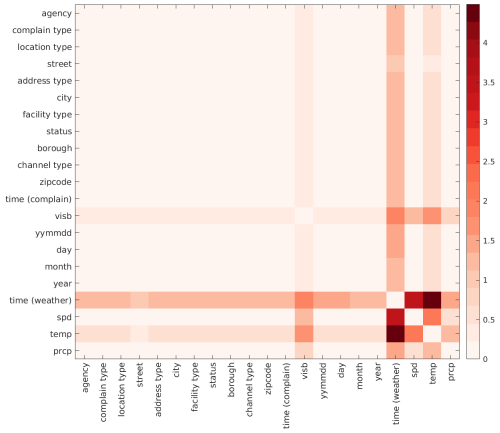

Figure 9: Correlation result between attributes from weather and complain in 2011

Weather Bike. The figure 10 shows that almost every attributes are related to the attribute *temperature* in *Weather*. The attribute *yymmdd* in *Bike* can represent the number of bike usage on that day. The high correlation between *yymmdd* and *temperature* can be interpreted that temperature changes affect bike usage rate. Hot or cold weather possibly discourage bike users from choosing it as means of transportation while reasonable temperature encourages them to bike.
An interesting correlation found here is between *gender* and *temperature*. There should be some different tendencies of bike usage between female and male users depending on temperature. For instance, female users tend to tolerate hot temperature so they can bike even the weather is really hot. While male users tend to bike more than female users when it is cold since cold weather is bearable for them.

Weather Taxi. In general, weather attributes such as *visibility*, *wind speed rate*, *temperature* and *the depth of precipitation* have

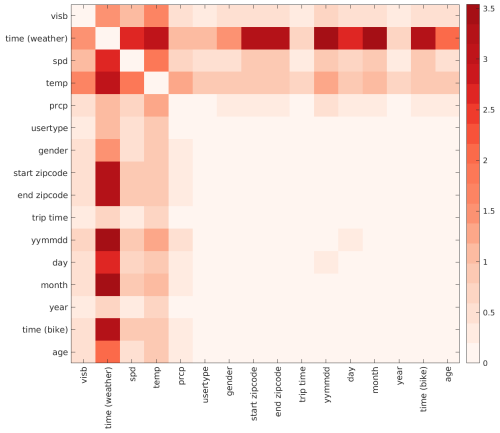

Figure 10: Correlation result between attributes from weather and bike in 2013

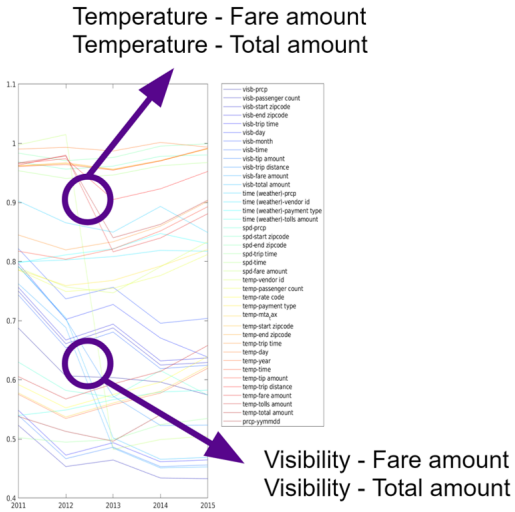

Figure 11: Mutual information year by year change from 2011 to 2015 for all attribute pairs from Weather and Taxi

higher correlation across most of Taxi attributes. This is because people prefer to take a taxi in bad weather such as raining, windy, hot, or cold.

One of the interesting correlation results is that the correlations between the weather and some taxi trip attribute decrease year by year, dramatically between 2012 and 2013 as shown in Figure 10. One possible interpretation for this phenomenon is as follows. The yellow taxi utilization rate got decreased in general since better taxi services such as Uber, Lyft, and Via have become available and taxi users have moved to the services year by year. These taxi services are fast and easy to reserve especially under the bad weather conditions. As a result, yellow taxi usage has not significantly increased even though the weather is not good. The situation change led to less correlation between weather condition and the taxi usage.

7 GIT REPOSITORY & SUPPLEMENTARY

Git repository. **Cool name pending** or go to the next url: https://github.com/DS-GA1004/Cool_name_pending.git

Slide. Click for the slide: **Correlation analysis on spatiotemporal data** or visit the url: https://goo.gl/cAuLQP

Additional figures. Click for additional figures on **Google drive** or visit the url: https://goo.gl/y1R4Ej

REFERENCES

[1] [n. d.]. WARM WEATHER AND INCREASED ACCIDENT RATES. https://www.hughesandcoleman.com/warm-weather-increased-accident-rates/. ([n. d.]). Accessed: 2018-05-14.

[2] Richard A Becker, Ramon Caceres, Karrie Hanson, Ji Meng Loh, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2011. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing* 10, 4 (2011), 18–26.

[3] Pablo Samuel Castro, Daqing Zhang, and Shijian Li. 2012. Urban traffic modelling and prediction using large scale taxi GPS traces. In *International Conference on Pervasive Computing*. Springer, 57–72.

[4] Fernando Chirigati, Harish Doraiswamy, Theodoros Damoulas, and Juliana Freire. 2016. Data polygamy: the many-many relationships among urban spatio-temporal data sets. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, 1011–1025.

[5] Samantha Cook, Corrie Conrad, Ashley L Fowlkes, and Matthew H Mohebbi. 2011. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS one* 6, 8 (2011), e23610.

[6] City Data. 2017. New York, New York (NY) Zip Code Map - Locations, Demographics - list of zip codes. (2017). http://www.city-data.com/zipmaps/New-York-New-York.html

[7] NYC Open Data. 2017. 311 Service Requests from 2010 to Present | NYC Open Data. (2017). https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

[8] NYC Open Data. 2017. Citi Bike NYC. (2017). https://www.citibikenyc.com/system-data

[9] :NYC Open Data. 2017. NYPD Complaint Data Historic | NYC Open Data. (2017). https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[10] NYC Open Data. 2017. NYPD Motor Vehicle Collisions | NYC Open Data. (2017). https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

[11] Martin Erwig, Ralf Hartmut Gu, Markus Schneider, Michalis Vazirgiannis, et al. 1999. Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica* 3, 3 (1999), 269–296.

[12] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. 2012. Big data: the management revolution. *Harvard business review* 90, 10 (2012), 60–68.

[13] Thomas WH Ng, Lillian T Eby, Kelly L Sorensen, and Daniel C Feldman. 2005. Predictors of objective and subjective career success: A meta-analysis. *Personnel psychology* 58, 2 (2005), 367–408.

[14] National Oceanic and Atmospheric Administration. 2017. Climate Data Online. (2017). http://www7.ncdc.noaa.gov/CDO/dataproduct

[15] The City of New York. 2017. NYC Taxi Limousine Commission - Trip Record Data. (2017). http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

[16] Erhard Rahm and Hong Hai Do. 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23, 4 (2000), 3–13.

[17] Ramana Rao and Stuart K Card. 1994. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 318–322.

[18] Lin Song, Peter Langfelder, and Steve Horvath. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* 13, 1 (2012), 328.

[19] Ralf Steuer, Jürgen Kurths, Carsten O Daub, Janko Weise, and Joachim Selbig. 2002. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* 18, suppl_2 (2002), S231–S240.

[20] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. 2014. Data mining with big data. *IEEE transactions on knowledge and data engineering* 26, 1 (2014), 97–107.