Exploratory Analysis of Crash Data of Philadelphia

Shu, Xingxing

Prof. Erick Guerra

1. Introduction

Traffic crashes are a severe problem in Philadelphia which takes nearly 100 people's lives and injures around 300 in over 10,0000 incidents every year. On November 7, 2016, Mayor of Philadelphia, James F. Kenny created a plan called Vision Zero Task Force set a target of zero traffic deaths by 2030. Though only one year after this plan launched, Philadelphia had 19% decrease in traffic deaths to 78 people (the lowest since 2012), this number bounced back to 91 in 2018. During last five years from 2013 to 2017, the average traffic deaths per 100,000 residents is 6.06 which is more than twice as New York's. Also according to report from PennDOT, more than 50% of traffic deaths occurred in just 12% of Philadelphia's streets.

In this project, I will use traffic crashes data from OpenDataPhilly and gather some other relevant data to answer following questions:

- Where do most crashes happened in city?
- What are the characteristics of crashes in Philly?
- Which streets are dangerous and which intersections have the most crashes?
- Where are these dangerous streets or intersections located? Do traffic crashes affect all neighborhoods in Philly equally?
- What environmental factors have impact on traffic crashes and to what extend?
- What are the temporal patterns of crashes in Philly?
- How can we improve circumstances of current roads to save more people's lives?

Figure 1.1 Philadelphia Crashes from 2011 to 2017

The figure 1.1 shows spatial distribution of crashes from 2011 to 2017 in city of Philadelphia. And from the Year Two Update 2019 report of Vision Zero, it shows though crashes occur in all areas of Philadelphia, there still a concentration along expressways and some other specific part of city. Figure 1.2 shows two areas in the city (Harrowgate Park and Schuylkill Expressway near 30th station) where have the highest concentration of crashes in past 5 years.

Figure 1.2 High Collision Frequency Locations

Though I will not create a model to predict where traffic crashes are likely to occur, analysis of past records is helpful enough to policymakers and urban planners for their decision making and future designs.

2. Methods

2.1 Data Collection

For the analysis, I gathered 4 types of data from open sources which includes:

Vehicular Crashes Data: This dataset contains crash data in city of Philadelphia from 2011 to 2017 with 77,463 rows of data. Then I selected useful and information-complete columns for latter analysis which shows in Table 1:

Table 1. Crash Data Glossary

Variable Name	Description	Type
crash_year	Year when the crash occurred	TEXT
crash_month	Month when the crash occurred	TEXT
day_of_week	Day of the week code when crash occurred	TEXT
time_of_day	Time of day when crash occurred	TEXT
illumination	Code that defines lightling at crash scene	TEXT
weather	Code for the weather type at time of crash	TEXT
road_condition	Road surface condition code	TEXT
collision_type	Collision category that defines the crash	TEXT
intersect_type	Code that defines the intersection type	TEXT
tcd_type	Code that defines the traffic control device	TEXT
location_type	Code that define the crash location	TEXT
fatal_count	Total amount of fatalities involved	NUMBER
injury_count	Total count of all injuries sustained	NUMBER
person_count	Total people involved	NUMBER
total_unit	Total count of all pedestrians and vehicles	NUMBER
latitude	GPS latitude	NUMBER
longitude	GPS longitude	NUMBER
bicycle_death_count	Total amount of bicyclist fatalities	NUMBER
ped_death_count	Total amount of pedestrian fatalities	NUMBER

But there is a fatal flaw of the data I collected from website that all crash data of recent years do not have a complete timestamp, namely, they all do not have information of which day of the month the crash occurred. Though there's a column describe the day of the week when crash occurred, there's no way to transform it into "ymd-hms" form. Let alone further temporal analysis.

Lucky enough, I found crash data with information of which day the crash occurred. I combined it with time columns in Table 1 to get a complete timestamp in form like "201208300610" which means the crash occurred at 6:10 A.M. on August 30 in 2012. This step was done in Excel. So as for temporal analysis, I will use this dataset instead of crash data of recent years.

Census Data: Neighborhoods of Philly was downloaded using *tidycensus* package in R which contains boundaries of census tracts with detailed information of census tracts of Philadelphia like percentage of white, average commute time to work, median household income etc. I selected 10 characteristics of tracts and the full variables are show in Table 2:

Table 2: Census Data Glossary

Variable Name	Description	Type
Total_Pop	Total population	Int
Med_Inc	Median household income	Int
White_Pop	Population of white	Int
Travel_Time	Total time on traveling	Int
Means_of_Transport	Average number of people	Int
Total_Public_Trans	Number of people taking public transport	Int
Med_Age	Median age	Double
Percent_White	Percentage of white	Double
Mean_Commute_Time	Average commute time for workers	Double
Percent_Taking_Public_Trans	Percentage of people taking public tranport	Double

Weather Data: Weather data is downloaded for the reason that it might plays a very important role in traffic conditions. I gathered weather data of Philadelphia airport weather station using *riem* package in R between May 1, 2012 and August 1, 2012 for these 3 months have the most crashes. Why 2012? In order to find out if there's a correlation between weather conditions and crashes, I need to make scatterplots with x-axis of weather variables and y-

axis of count of crashes. The bridge link these two datasets together is time, that is I'll summarize the average counts of crashes and the average like temperature in every 1 hour. This required complete timestamps which is only available to 2012. And weather data contains three types of conditions: precipitation, wind speed and temperature. (shows in Figure 2.1)

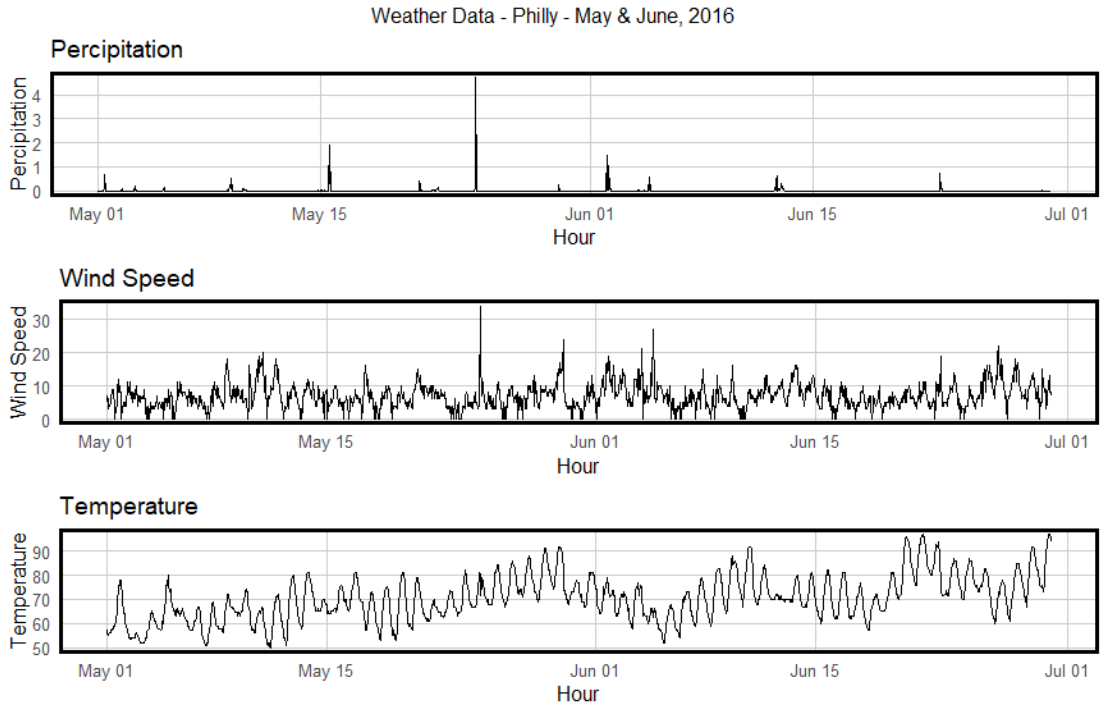


Figure 2.1 Weather Data of Philadelphia in May & June, 2012

Geographic Data: Centerlines of streets of Philadelphia and boundary of Philadelphia was downloaded from Opendataphilly as shapefiles.

2.2 Data Wrangling

Step 1: My analysis begins with street centerlines of Philadelphia In ArcGIS. First, I need to break each road into separate lines at their intersections and extract all intersection points by topology in ArcGIS. Then I create buffers for these two objects: 15 meters for road centerlines and radius of 10 meters around intersection points. Then I clip the road centerlines file by the intersection buffers to get road lines without roads in intersection buffers to avoid overlapping. To locate which roads or intersections have the most crashes, I spatial joined the crash points to these two shapefiles in ArcGIS.

Also, I used spatial join tool to calculate number of crashes of each census tract in the past 5 years and add a field calculating crash density by dividing total number of crashes by area of each tract (unit: count / km^2).

Step 2: Divide the crash raw data from 2011 to 2017 into multiple files in Python. First group is grouped by crash years and months and calculate the total number of injured and fatal people; Second group is grouped and calculate the total number of victims by months and the time of a day to see how it changes through a day. Others are grouped by different characteristics of collisions.

Step 3: Crash data with complete timestamp of 2012 was read in R and I use lubridate package to parse it into the time form I expect: "2012-08-30 06:10:00" and extract which week of the year the crash was in. Then I need to combine these three data panels together to do further analysis. First is to join crash data with weather data based on the timestamp; Then is to join this new data panel with census data use spatial join.

2.3 Exploratory Analysis

Characteristics of collisions: I want to explore the relationship between some external environmental features, like road conditions, illumination or internal features like crash locations and collision types with number of collisions. To have a better view of the impact of weather on occurrence of crashes, I calculated the total number of crashes in rainy and non-rainy days in May and June by week, which is from week 18 to week 26. And I plotted temperature against average count of crashes by week in the same period.

Temporal Analysis: I want to explore how crashes changes in different day of a week; How does it change at different hour of the day; How does it change at different month of the year; Is there a difference between weekday and weekend. What time of day is more dangerous? Do crashes is correlated with traffic rush hours? This part of analysis was done in both Python and R.

Spatial analysis: First I tried to find out which part of city has more crashes and foe cyclists and pedestrian, which part of city is dangerous to them? Next, we know that crash shows some temporal

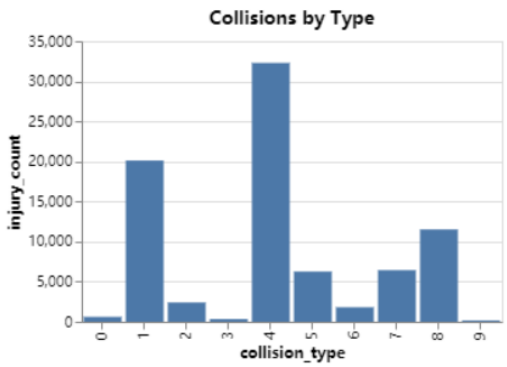
correlation, but what about spatial autocorrelation? How do collisions change in different tracts in different time, do they occurred repeatedly at some specific areas or they actually shows different spatial patterns at different time? So I plotted the sum of crash counts of each tract by week (still week 18 to week 26 for less workload and more data in this interval), by day of the week and by hour of the day. Also I spatial joined the total crash points to segmented street buffers and intersection buffers to find out spatial distributions of collisions more specific. I did this part of analysis in R and ArcGIS.

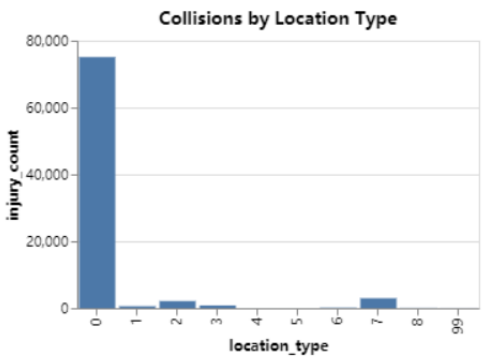
Socioeconomic Analysis: To see whether crashes occur spatially equal to all tracts, I calculated the Pearson correlation coefficient for each of these 10 Census variables. Finally, small multiple scatterplots are created in R.

3. Results

3.1 Characteristic Analysis

Figure 3.1.1 shows the interior characteristics of collisions from 2011 to 2017. The top three collision types are: Angle, Rear-end and hit pedestrian; as for where these collisions occurred, most of them are not applicable in this record. It seems that environmental factors have more impact on collisions like over 75,000 crashes occurred in dark conditions without street light or at dusk. But weather or road conditions do not have much impact as we expected that collisions in rainy days or on wet roads do not necessarily means more crashes. Though nearly 11,000 collisions happened in rainy days, over 80% of crashes occurred under no adverse conditions.





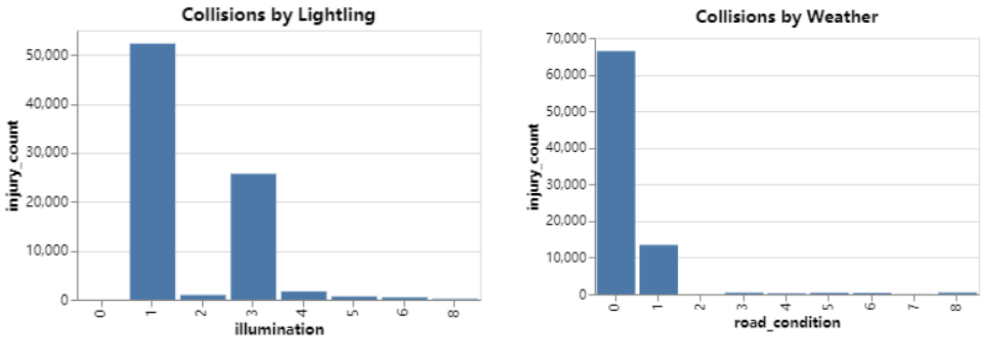


Figure 3.1.1 Injury Count by Different Types of Collisions

Then I calculate the average count of crashes from May 1st to June 30th. In Figure 3.1.2, in all 9 weeks counts, we observe number of crashes in non-rainy days are significantly higher than in rainy days. That is precipitation have very little impact on crashes than we expected.

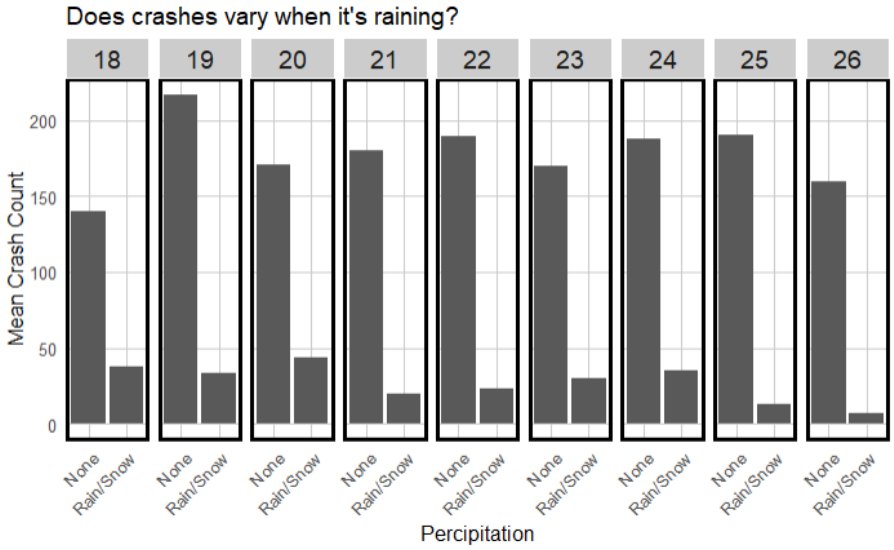


Figure 3.1.2 Average Crash Count by Rain / No-Rain

Also I checked the correlations between crash counts and temperature by week. In several weeks, we observe some positive correlation as temperature climbs from 50 ℉ to 80℉ in week 19 and from 60 to 90 ℉ in week 25. But calculated the correlation coefficients of temperature and crashes, the coefficients are nearly

the same whether we considered weather conditions in or not. Thus weather factors do not have much impact on crashes in Philly.

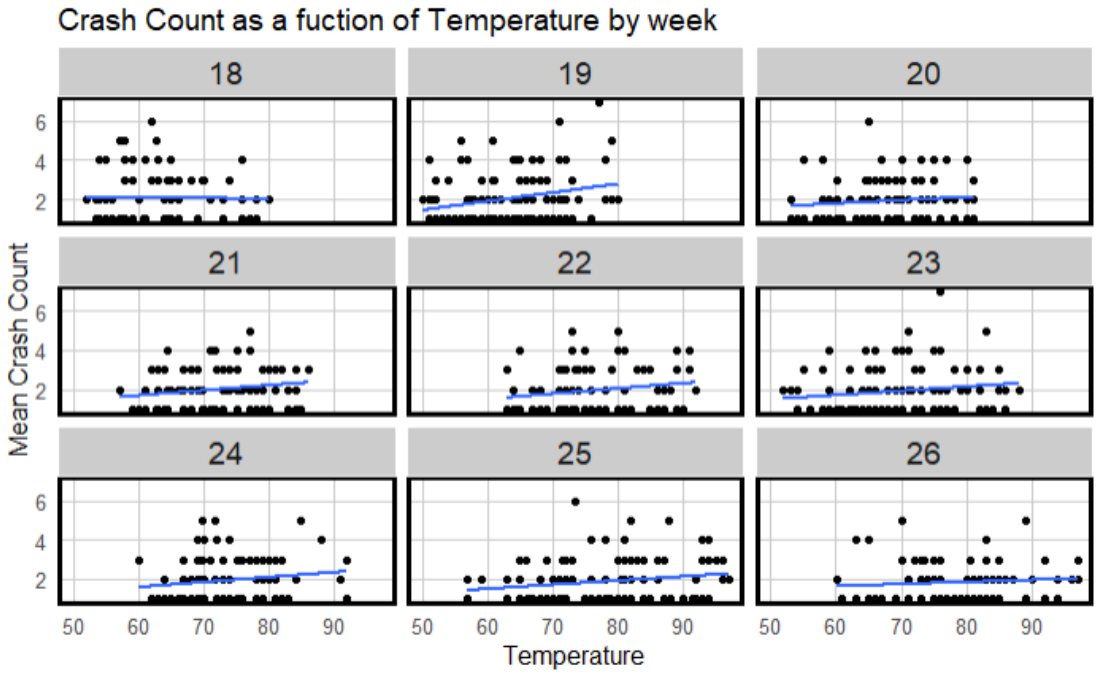


Figure 3.1.3 Crash Count as a function of temperature by week

3.2 Temporal Analysis

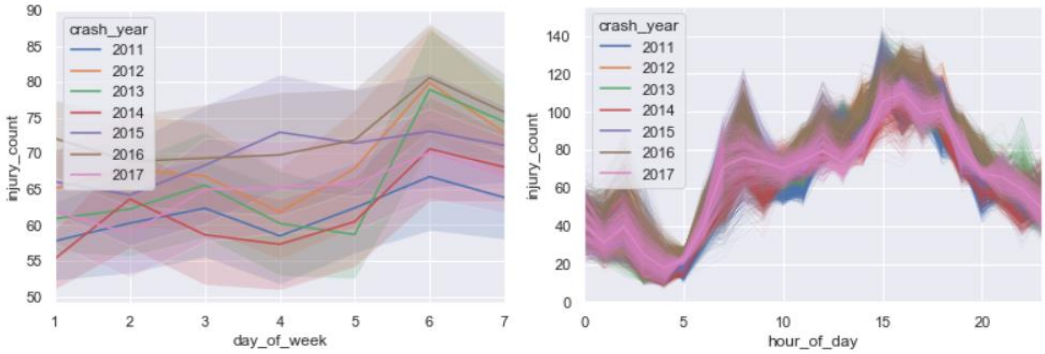


Figure 3.2.1 Sum of Yearly Crashes by day of a week / hour of a day

Here we examine the temporal patterns of collisions. I calculate the count of injuries of each year by day of a week and by hour of a day to see how crashes differ. In figure 3.2.1, though polylines are not concentrated together, the peak of each year show on day 6 which is Friday and the second peak of collision shows on Monday. And there is a “U”-shape between Monday and Friday where

Wednesday lies at the bottom of this “U” shape. And yearly count of collisions are increasing as the lines represent each year tend to be higher than previous year.

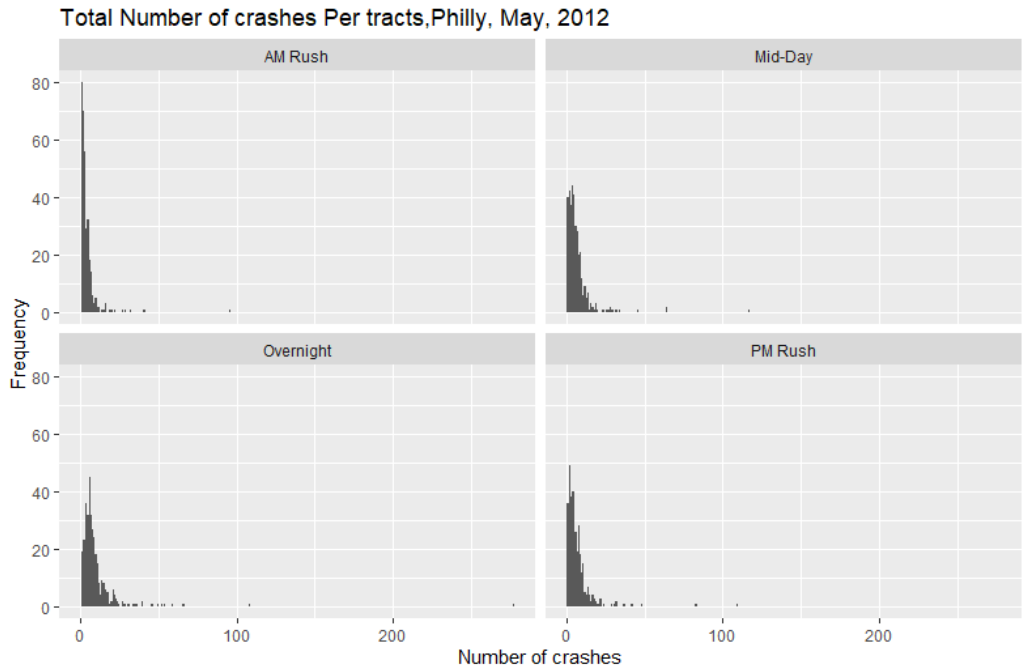


Figure 3.2.2 Sum of Crashes by tracts in different period of a day

Figure on the right in Figure 3.2.1 shows these polylines seem to be more concentrated than lines in figure on the left, namely it has more clear patterns than day of a week: the trough of a day occurs at 4 a.m. and the number climbs way up to nearly 5 times at 7 to 8 a.m., though there is a little decrease after morning rush, the number of collisions still keep at a very high level and achieves another peak of the day at afternoon rush hour, then the counts take a nose dive in evening.

It seems collisions follow the pattern of traffic volumes of the day, especially from the figure on the right. So I divided hour of the day into four periods: morning rush, mid-day, afternoon rush and overnight. Made histograms of crashes in each of these four periods and result shows in Figure 3.2.2. We do observe in the morning rush, collisions counts are highest but for other 3 periods, no obvious differences are observed.

To get a better view of how injuries or fatality changes over time, I created two bubble plots as show in figure 3.2.3. The x-axis shows

time of the day and the y-axis shows day of the week. Injuries have more clear temporal patterns than fatalities. Generally speaking, weekdays and weekends are slightly different for more injuries occurs at midnight on weekends but on weekdays, more injuries happened between 7:00 to 18:00. As for crashes with fatalities, they tend to occur after 18:00, especially in midnight on weekends. Shortly speaking, injuries concentrate in the working hours of the day and fatalities have more in the rest hours.

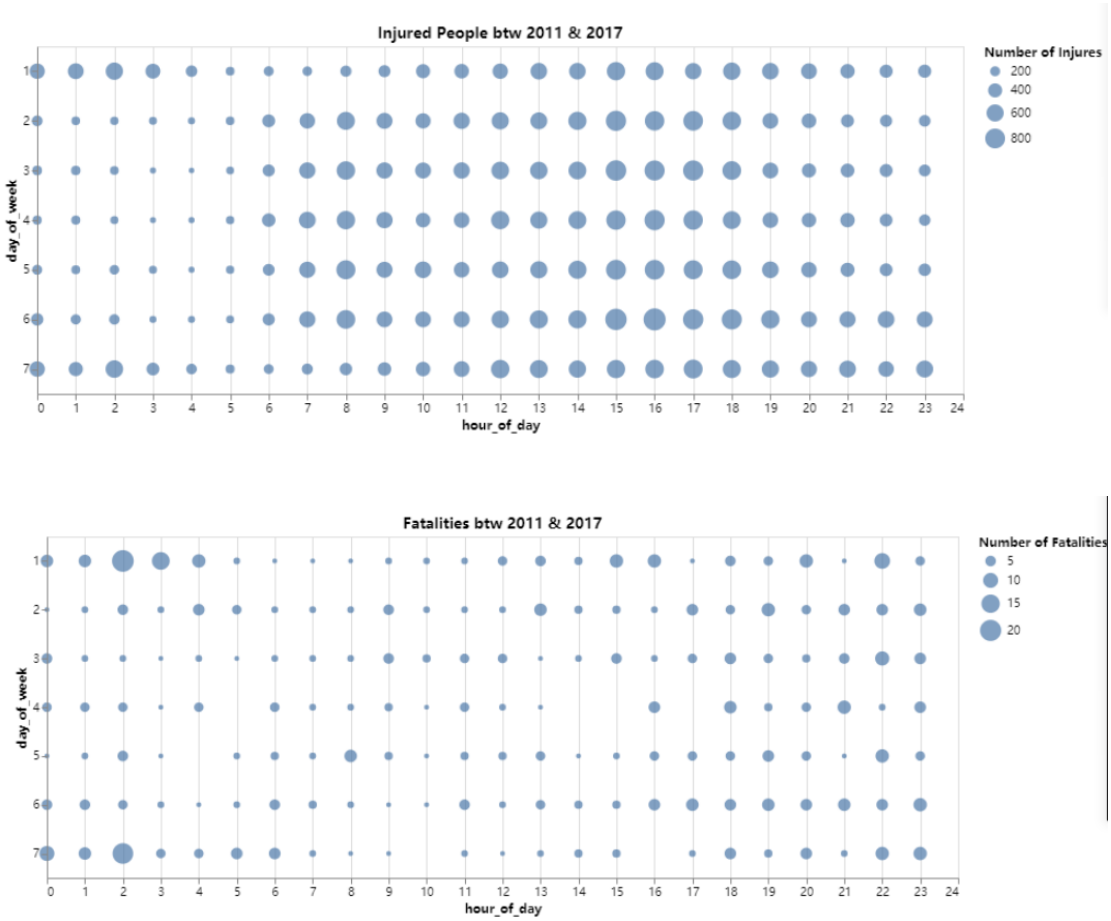


Figure 3.2.3 Sum of Crashes by day of a week & hour of a day

What about months or do different month has different level of collisions? Heatmap in figure 3.2.4 shows the injuries and fatalities with x-axis of crash months and y-axis of crash years. We observed that in general, December, January and February have the lowest count of collisions with injuries while the peak of a year usually happened in April, May and June.

Fatalities do not seem to have patterns compare with injuries, but in July, August and September, there are little bit more

fatalities than the rest months. And concluded from rows, from 2011 to 2017, both numbers of injuries and fatalities are increasing which is frustrating and head towards to another direction of what Vision Zero expects.

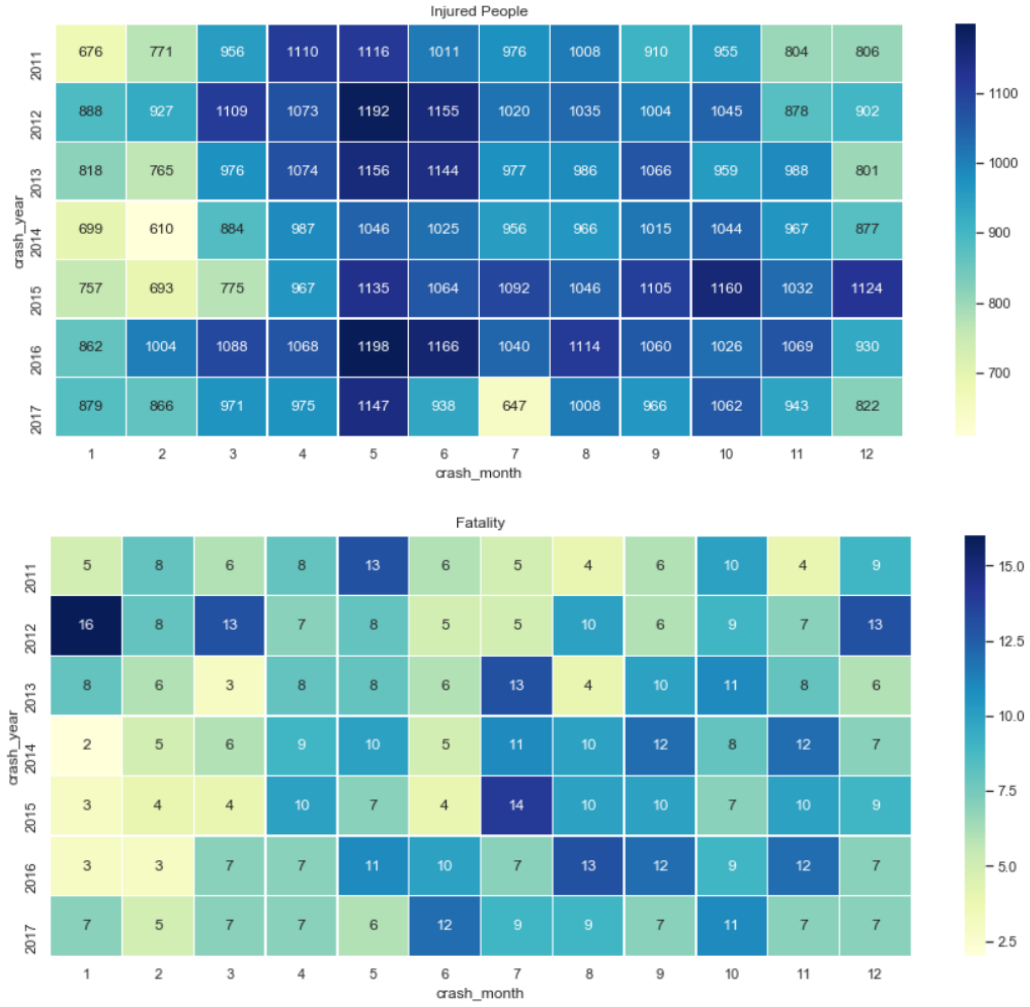


Figure 3.2.4 Sum of Crashes by year & month

3.3 Spatial Analysis

First, I plot all these crash points on map as show in figure 1.1, but it's not very useful and can not tell us more information about crashes in city of Philadelphia. Then I created hexagons with side length of 500 meters as my research unit and calculate the number of crashes within. The result is in figure 3.3.1.

Obviously, distributions of collisions are not random. On some roads in cities, there are more collisions on average, and some road types (such as expressways) have more collisions, regardless of their location in the city. In addition, most road segments in Philadelphia were free of accidents. We observe clear concentration along expressways like Schuylkill, Roosevelt, Delaware and Van Street expressway. And on North Broad street and areas around City Hall also have more than average crashes. All these places are high in traffic volumes and are arteries of Philly.

Figure 3.3.1 Spatial Heatmap of Crashes

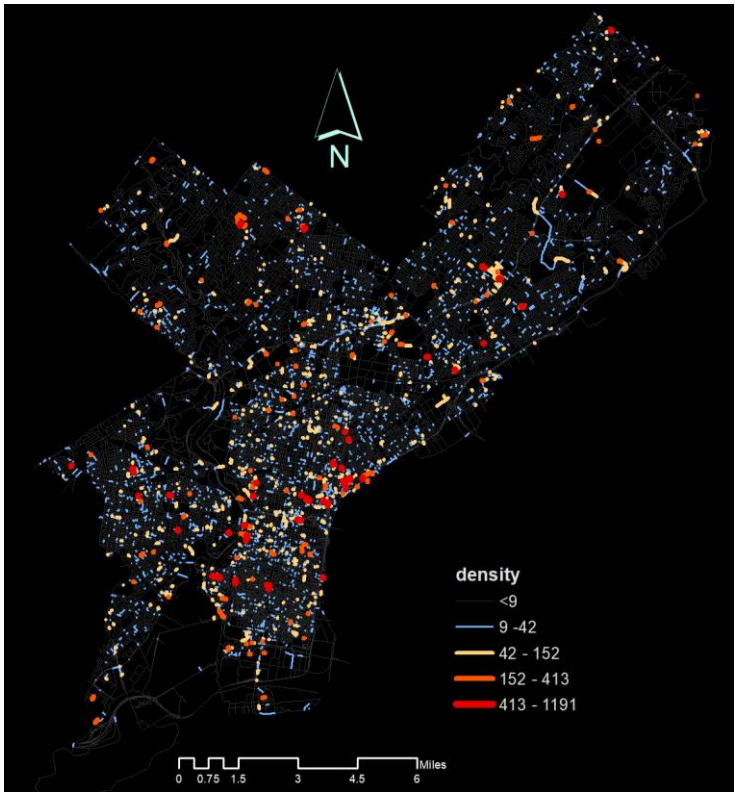


Figure 3.3.2 Crash Density of street segments (per km)

Figure 3.3.3 Sum of Crashes within 10m buffers of Intersections

To get a detailed view of distribution of collisions in city, I calculate the density of collisions in each segmented street and intersections. I standardize the count of crashes by calculating the density with the unit of crash per kilometer and for intersections. Since intersections are all the same in sizes, so there is no need to calculate the density but simply use the number within each of them.

As I what I expected, most streets of Philly do not have collisions and streets with crash points more than 400 are mostly expressways of city or streets connect city street to these arteries. In Figure 3.3.3, we observe more points in map indicates that more collisions occurred at places near intersections. Also in this map, crash count more than 30 are all intersections of expressways with urban streets. Like intersection of Van St Expy and Schuylkill Expy; Entry and exit ramp of Vine St Expy at 16th street.

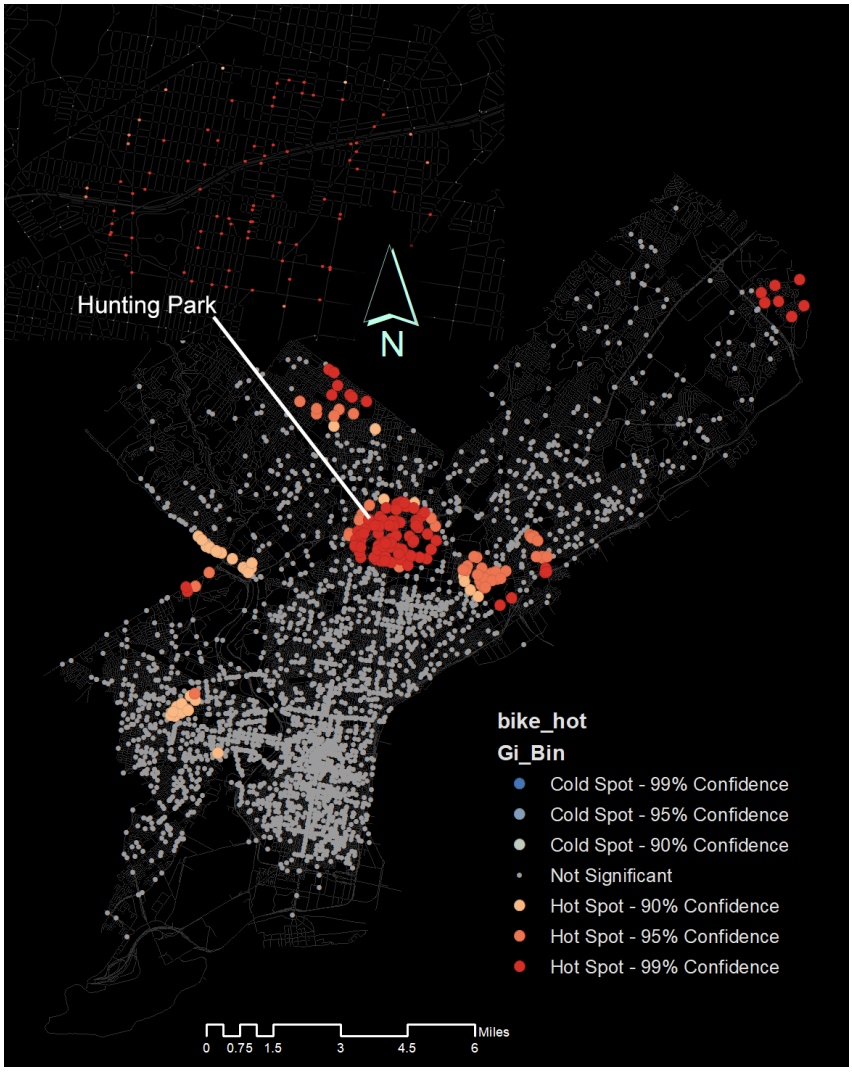


Figure 3.3.4 Cyclist Crash Hotspots

Then I select collisions with fatalities of cyclists and pedestrians to find out which places in city are dangerous to these two kinds of people, or did hot spots analysis of these two kind of collisions by volume of people involved in ArcGIS. Hotspots of cyclists are show in figure 3.3.4 and hotspots of pedestrian are show in figure 3.3.5.

As show in figure 3.3.4, no significant cold spots are observed and hotspots of 99% significance concentrate at the north part of Philly around Hunting Park and Mechanicsville where at the far northeast part of Philly; Hotspots of 95% significance locate at Harrowgate and Frankford on the east side of Hunting park; Hotspots of 90% significance locate at 52nd street and along the north bank of Schuylkill River at Manayunk.

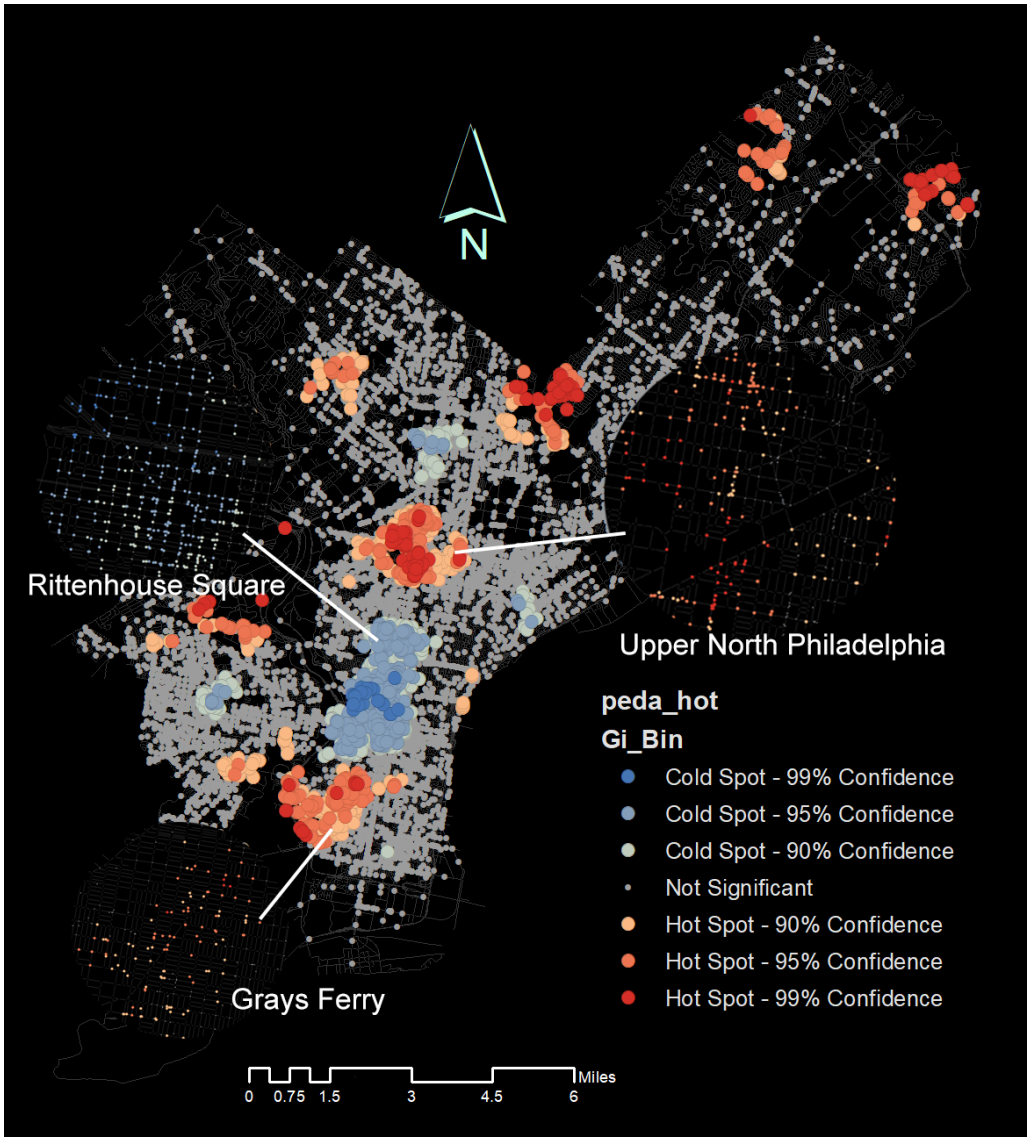


Figure 3.3.5 Pedestrian Crash Hotspots

Concluded from figure 3.3.5, multiple clusters of hotspots and coldspots are exposed on map. All coldspots concentrate in the region which south border is Lombard Street and north side is Spring Garden St, west to Schulykill River and east to Broad street. This is interesting because this area is of high volume of both traffic and population, but it is where cold spots of pedestrian collisions concentrated. As for hotspots of pedestrian, points of 99% significance locate in Upper North Philadelphia around Temple University Hospital, and north boundary of Philly at Lawncrest and Oak Lane/ East Oak Lane. Also in far northeast part near city boundary in Mechanicsville.

This result is not what I expect that I hope to see concentration of hotspots at center city or some other places with high volume of population; This may be the consequence of considering number of pedestrian involved of each collision points, or to say that though collisions with cyclists or pedestrian may have more incidents at center city or some other places where we think they should based on our common sense, these incidents are not severe or not many people were got involved. Most collisions in center city have 1-2 people involved, but places in Upper North Philadelphia, Grays Ferry have more injuries and more fatalities which makes them hotspots after we add the weight of number of people involved.

Figure 3.3.6 Sum of Crashes by tract and week

Next, let's look at the how spatial pattern of collisions changes by time throughout census tracts of Philadelphia. Figure 3.3.6 shows the distribution of crashes of tracts by week from May 1st to June 30th. During these 8 weeks in 2012, the spatial pattern remains stable and do not change vey much. In each case, crash appears to be concentrated in southern, western tracts at the north bank of Schuylkill River, South Philadelphia East, north part of Philly. Though count of crash increase from week 18 to week 25 as a whole in the city ,especially in the north part. Patterns of crash do repeatedly every week.

Then take a look at the sum of collisions of tracts by day of the week. In the result of my temporal analysis previous, crashes do rise and down in 7 days of the week, peaks on Friday then follows Monday, the low point of the week is on Wednesday. In Figure 3.3.7, patterns of crashes are nearly the same and similar to patterns by week. Besides tracts with expressways, other tracts with high count of crash are in Northeastern Philly and in Southern part.

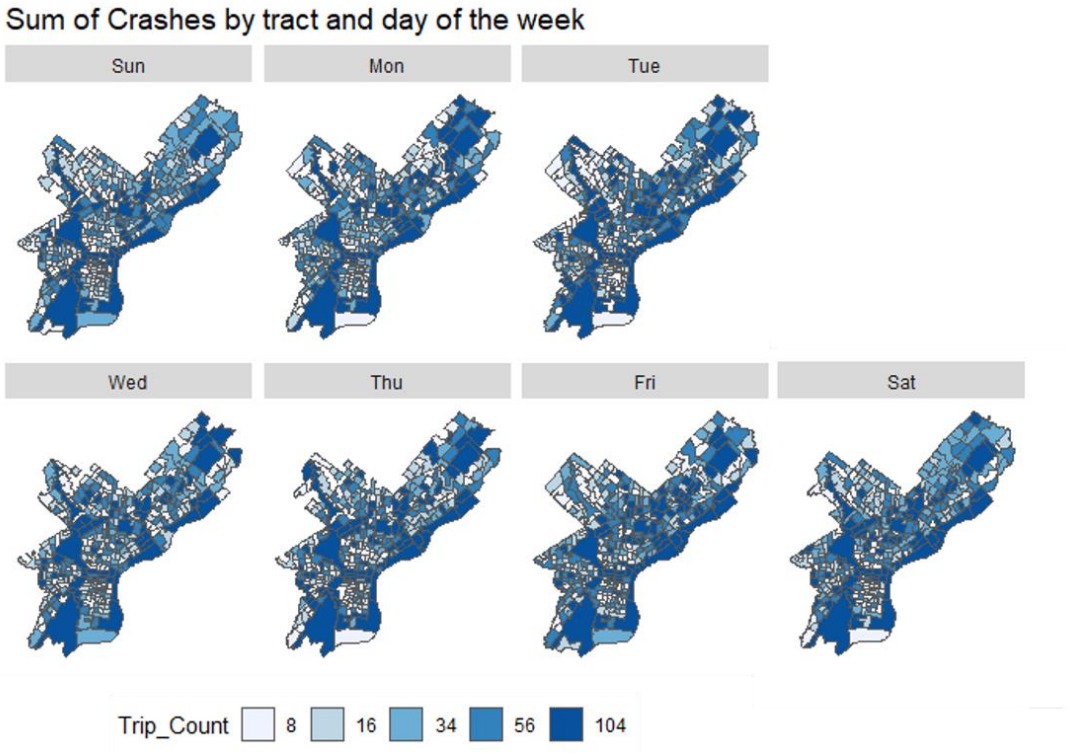


Figure 3.3.7 Sum of Crashes by tract and day of the week

Finally, Figure 3.3.8 visualizes the spatial distribution of crash by hours. Similarly, crashes are concentrated in several large tracts with expressways of high traffic volume, but as time changes, the crash count seems to gradually increase outward from the central part of city and made these tracts ‘hollowed’. But north eastern Philly like tracts in Frankford, Lawncrest, Olney and Feltonville are always have high crash count across all day. This may be traffic conditions there are complex with multiple arteries pass through and different types of roads intersect there.

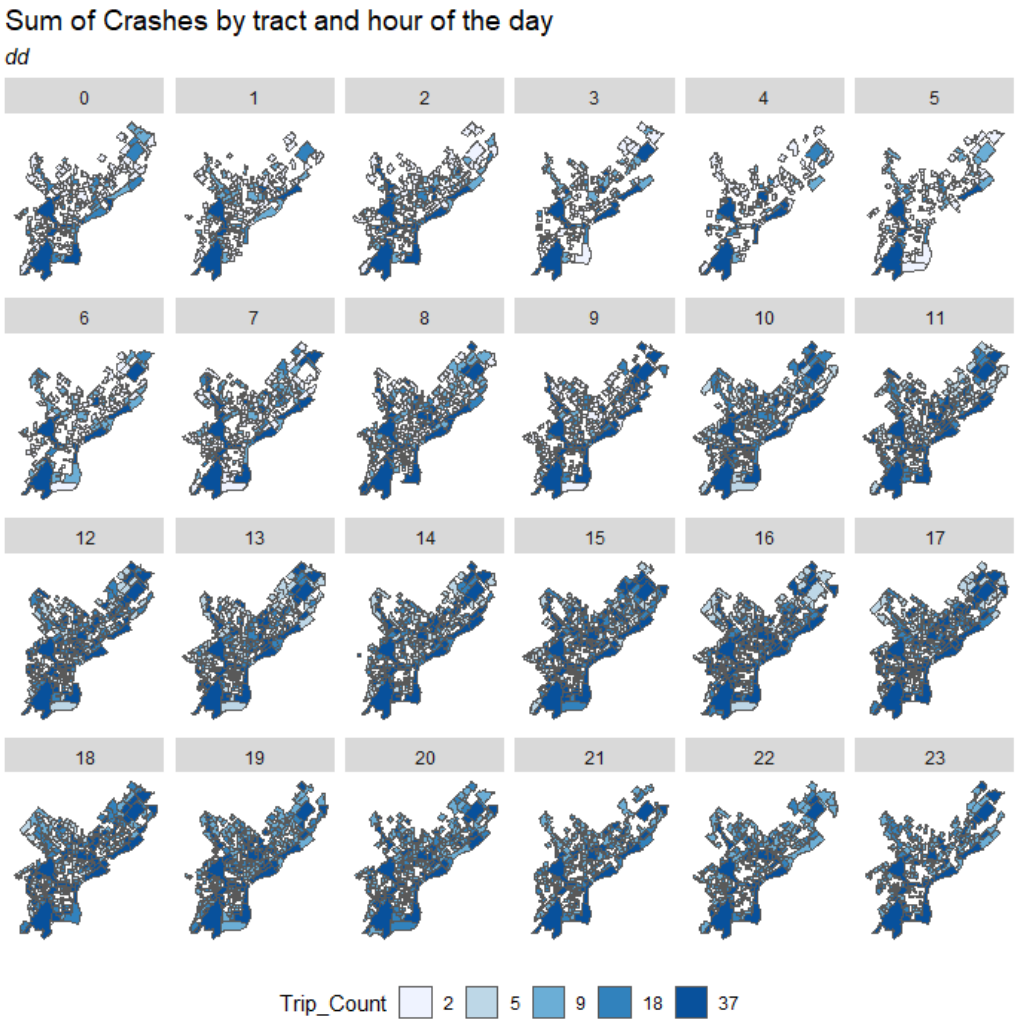


Figure 3.3.8 Sum of Crashes by tract and hour of the day

3.4 Does crashes differs in census tracts?

This section briefly analyzes correlation between crash count by tract for year 2012 as a function of 8 Census variables: aggregate travel time to work, mean commute time for workers, means of

transportation to work, median household income, percent of taking public transportation, total population, total number of peopple taking public transportation exclude taxi and tota white population.

As results of analysis in previous, clear patterns of spatial and temporal have revealed. Since crashes concentrate at specific part of city, this makes me thinking if it's correlated with socioeconomic characteristics of census tracts in Philly. Like in Figure 3.4.1, besides center city, tracts near to city border have less crash density of inner tracts.

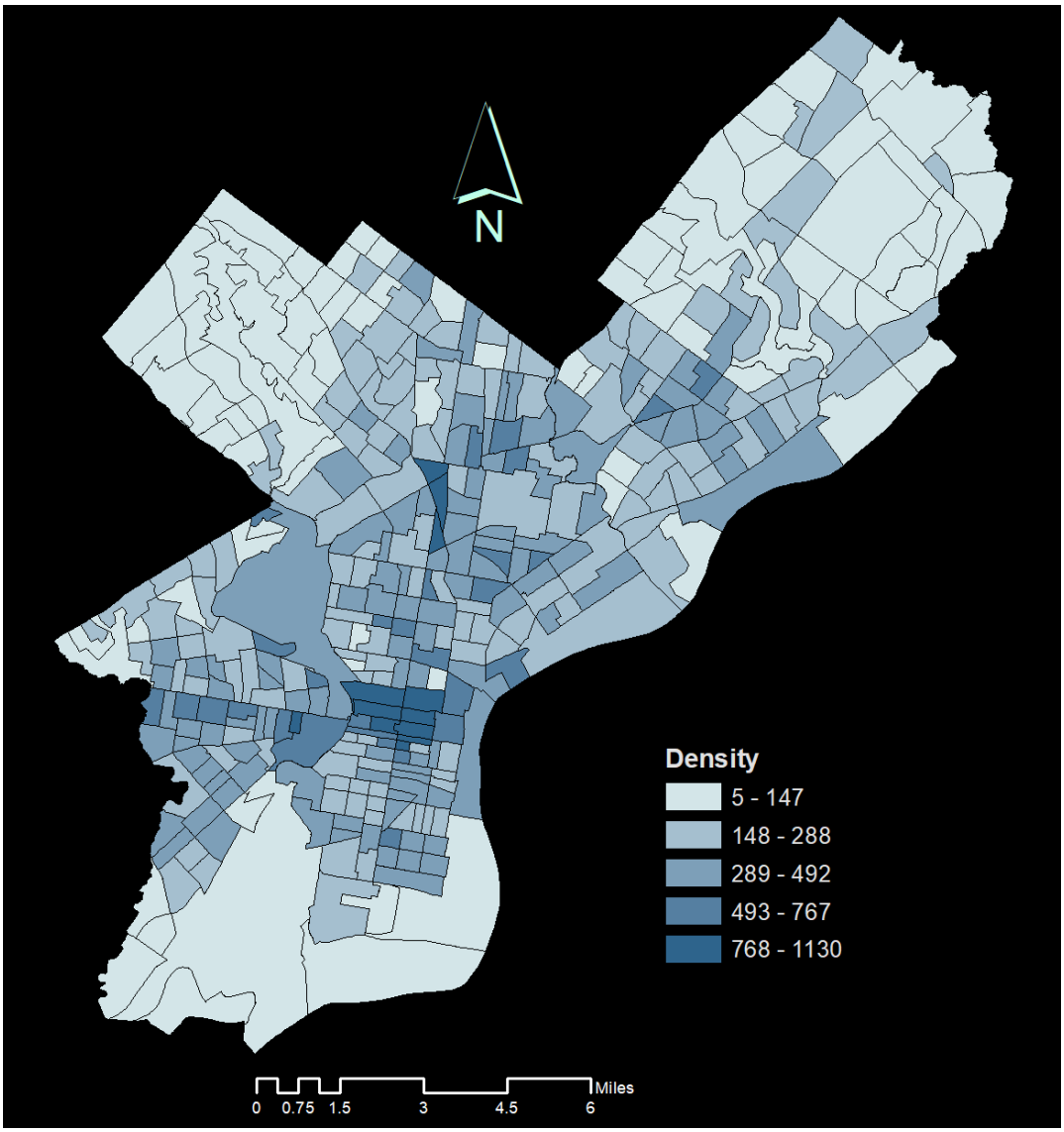


Figure 3.4.1 Density of Crashes by tract

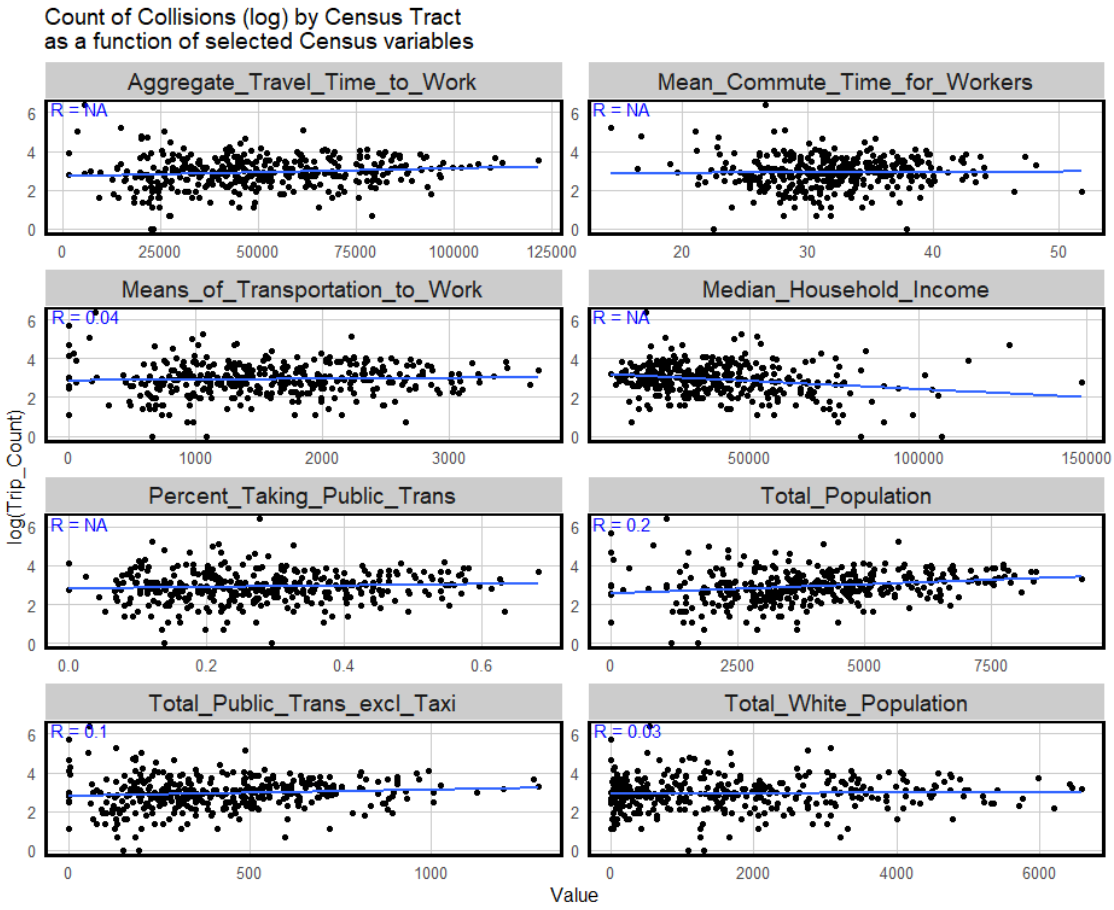


Figure 3.4.2 Do Crashes correlated with socioeconomic variables?

Scatterplots with R squared show the relationship between crash count and each of these 8 variables. Surprisingly, half of them do not have any relationship with crash count. The rest 4 variables though their R square is not NA, coefficients are still very small and none of them are significantly correlated with crash count. So concentration of crashes in specific regions or census tracts in city is not due to difference of socioeconomic characteristics of tracts.

4. Discussion

The results of exploratory analysis reveal the fact that crash of Philadelphia is not randomly distributed in city but have patterns spatially and temporally. Crash accompanied with traffic volume of roads and it has morning rush too which crash peaks in the morning time from 5:00 to 8:00, and the in the rest of the day, crash count climbs to its second peak from 15:00 to 17:00.

In weekdays, crashes with high injuries are concentrate in worktime (7:00 to 18:00) while in weekends, crashes with high injuries are concentrate in midnight. Crashes with fatalities do not have clear patterns like injury does, but more crashes with fatalities are occurred in midnight too. On Friday, Philly has the highest crash count then follows Monday while the day with the lowest crash count is on Wednesday.

In a year, December, January and February have the lowest count of collisions with injuries while the peak of a year usually happened in April, May and June. Fatalities do not seem to have patterns compare with injuries, but in July, August and September, there are little bit more fatalities than the rest months. And from 2011 to 2017, both numbers of injuries and fatalities are increasing.

Spatially speaking, most streets of Philly do not have collisions or have very few collisions occurred. Specific area of Philly has the most number of crashes and most of them locate along Schuylkill, Van Street and Delaware expressways and arteries of city like Broad Street, Roosevelt Blv, chestnut street and walnut street on west philly with high traffic volumes. Compare crash density of street and intersections, I find out the fact that intersections have more crashes than streets do and intersections with high crash counts are where expressways meet boulevards of city.

Dangerous places for cyclists are in northern Philly like Hunting Park and Mechanicsville where at the far northeast part of Philly. Other places like Harrowgate and Frankford on the east side of Hunting park and area near 52nd street and along the north bank of Schuylkill River at Manayunk, for these places have more than average cyclist-involved collisions; For pedestrians, Upper North Philadelphia around Temple University Hospital, north boundary of Philly at Lawncrest, Oak Lane/ East Oak Lane and in far northeast part near city boundary in Mechanicsville are dangerous for the same reason that collisions occurred at these places have more than average pedestrians involved.

Crash patterns of tracts by week, day of the week and hour of the day are very similar or repeated. Tracts with high crash count are around the border of Philly or where expressways pass through. No obvious difference is observed from these multiple maps, but despite the rise and fall of crash count within tracts, some tracts in

the northern part of Philly are constantly high in crash count. I do not know exactly why this happen, but given to the complex road conditions and traffic volume brought by several arteries passed through, these may be part of the answer.

Last, I calculate the correlation coefficients between crash count and multiple weather and socioeconomic characteristics. Found out that none of them are significantly correlated with crashes in Philly. This indicates that crashes in Philly are more of spatial rather than environmental or socioeconomic.

5. Conclusion

This project uses Python, R and ArcGIS to conduct an exploratory analysis of crash data from 2011 to 2017 of Philadelphia and for crash data of 2012 of Philly for more detailed timestamps. The analysis contains four perspectives: spatial, temporal, socioeconomic and environmental. Results can serve as a guide for locate dangerous roads or tracts in Philly or give some advice for planners which part of city need improvement of road design, or a reference for citizens to know where is dangerous and keep that in mind just in case. Moreover, more specific and further research can be applied in future with more recent and detailed dataset.

References

Analyzing Philadelphia Crash Data, By Daniel McGlone on April 2nd, 2014: https://www.azavea.com/blog/2014/04/02/analyzing-philadelphia-crash-data/

The Philadelphia Pedestrian's Guide to the City's Most Dangerous Streets, by Jonathan Geeting, Feb. 21, 2014: https://nextcity.org/daily/entry/the-philadelphia-pedestrians-guide-to-the-citys-most-dangerous-streets

Geography of Crashes in Philadelphia, By Carlos Bonilla on September 19th, 2016: https://www.azavea.com/blog/2016/09/19/geography-crashes-philadelphia/

Three year Action Plan, Vision Zero, Sep, 2017: http://visionzerophl.com/uploads/attachments/cj8a9vbdj074ojnd66ah3mxxi-2017-vz-action-plan-final.pdf

Year Two Update, 2019: http://visionzerophl.com/uploads/attachments/ck181ipfv1isp9pd66ww0iw0f-file-print-pages-hq-visionzero-y2-update.pdf

Traffic crash analysis with point-of-interest spatial clustering, Ruo Jia, Anish Khadka, Inhi Kim, Accident Analysis & Prevention, Volume 121, Dec.2018, Pages 223-230. https://www.sciencedirect.com/science/article/pii/S0001457518306481#!

Crash Analysis Report- the City of Cambridge: https://www.cambridgema.gov/~/media/Files/policedepartment/SpecialReports/CrashAnalysisReport_Final_05252017.pdf

Philadelphia Crash Analysis Standards & Recommendations: https://www.dvrpc.org/Reports/17068.pdf