**Group 19**

Michael DeBello 720462656

Zachary Dickson 720469328

Amanda Liu 720042603

**Process Book**

**－ Overview and Motivation：**

We chose to analyze car collisions in one of the most traffic-ridden cities in the world and evaluate trends in the data. We used a data set provided by the New York Police Department that broke down every motor vehicle collision in NYC based on the location of the accident and any injuries that resulted from it. The data includes motor vehicle collisions from July 2012 to November 2017 that is updated manually every month on the NYPD website. Each record represents a crash in NYC by borough, precinct, cross street, and the specific time the crash happened, along with details on the contributing factors to the collision, as well as the type of vehicles involved. From this dataset, we are able to determine how dangerous or safe different intersections are in NYC based on the number of accidents there are. We have also been able to determine the likelihood of an accident occurring based on factors such as the inattentiveness of the driver or the time of day. We also determined the probability that an individual involved in an accident could be injured or killed.

We were initially interested in looking into New York City accidents because one of our group members, who is from New York City, noticed a difference in the driving habits, attitudes, and vehicle safety in New York City compared to Chapel Hill. We wanted to analyze the data of accidents from one of the worst traffic cities in the world and examine where and why they occur. Ultimately, we wanted to determine this so that we could consider potential implementations of more stringent safety regulations and modifications. In particular, we were interested to learn more about different intersections that crashes occurred in, the different types of vehicles that caused them, and at what time they occurred.

**－ Related Work:**

We were inspired by the mpg data set that we initially worked with in this class. We found out that we can sort the vehicles into different categories in order to figure out whether it related to the number of collisions happened. But in this case, we grouped them by the types of the vehicles instead of the manufacturers or models, such as passenger vehicles, taxis, vans, etc. **－ Initial Questions:**

1. What are the most dangerous/safe intersections in NYC?
2. Which type of vehicles (pickup truck, passenger, taxi, etc.) most frequently caused a collision?
3. What are the most popular contributing factors that led to a collision and is there a certain type of vehicle and/or area of NY associated with them?
4. What is the probability that people were injured or killed from the collision?

Originally, our questions did not lead to very insightful conclusions, such as determining the most dangerous and safest intersections in New York City based on the highest number of collisions per street. But, over time our questions developed more into questions like what times more accidents occurred at and why. We also started to realize that New York City is very large and a rather broad geographic area, so we started to ask if there were specific trends within the different boroughs of New York City (differences in collision frequency, number killed) in order to determine if the layout of the streets and the type of driving (concentrated, traffic-based vs. more spread out) had any influence on the accident patterns. We also wanted to see how the average number of collisions per month within each borough compared and differed throughout time. We also became increasingly curious with the distribution of accident-related deaths across the city and thought the latitude and longitudinal coordinates were too valuable and insightful to pass up using. Therefore, some of out new questions included: At what time intervals were accidents more and less frequently occurring? What is the amount of accidents across the different boroughs so that we can look at differences among the borough and see how that plays a vital role in collisions? And what does the distribution of deaths caused by the injuries look like across the different boroughs?

**－ Exploratory Data Analysis:**

In answering most of our questions, we started out by utilizing bar graph visuals in order to portray the most dangerous and the safest streets in NYC. Within each of our intersection-based graphs we broke the streets down into 3 types of categorical variables to quantify “dangerous” in 3 different ways-- number of accidents, number injured, and number killed. In order to get to this we utilized dyplr commands learned in the first portion of the class to mutate a new variable Num\_injured which was the sum of every injured variable (people in the vehicles, pedestrians, cyclists, motorists). After doing the same for number killed by mutating a new num\_killed variable, we group\_byed the cross street name and summarised count, num\_killed and num\_injured. Finding the max of each variable, we were able to look at the top streets in terms of the highest frequency of accidents, highest injury toll, and highest death toll. Using min, we did the same to find the safest streets in the city. We then made 2 bar graphs out of this to plot the most dangerous and safest NYC cross streets, each broken down into 3 bars by number of collisions, number injured, and number killed. From these data mutations, we then combined the number of people injured and the number of people killed to figure out the proportion of crashes that resulted in an injury or death. We got the proportion by taking the mean of the conditional statement of whether someone got injured or killed.

We also wanted to find out what type of vehicle had most frequently caused a collision. The NYPD collision data reported the different types of vehicles by multiple codes, so one vehicle could have multiple different characteristics like being a passenger vehicle and a sports utility vehicle. Using the dplyr function group\_by we grouped together each of the first two vehicle type codes. Once we had grouped those together we were able to take the union of those two to combine them into each of the different types. So, for each vehicle type, they were combined into two parts of whether they were characterized using type 1 or type 2. We then created a bar graph that illustrated the 10 most frequent car types that got into a collision, broken down by the 2 vehicle codes. But performing a coord\_flip to make the vehicles types much easier to read on the y-axis. The next thing we wanted to look at was the highest contributing factor to a crash. Just like the type of vehicles, there can be multiple reasons for the cause of a crash happening. We used the same process for the contributing factors that we used with the vehicle types by grouping together the factors and showing the top ten factors that led to a crash.

Next, we wanted to look at 1 hour intervals to determine if there were certain times during the day in which collisions were far more common, data which could be leveraged in order to make traffic directing more efficient and safer for all during those times. We accomplished this by changing the normal time into seconds, which we did by multiplying the hours by 3600 and the minutes by 60, then adding both of those to the number of seconds. Then we grouped together each of these times by a 1 hour interval using the function inner\_join by every 3600 seconds. Once we had grouped together everything into 24 intervals, we then reverted all of the times back into normal times with hours, minutes, and seconds. We then were able to illustrate with a bar graph what times collisions most frequently happened. We originally had evaluated it with 30 minute intervals and graphed those results, but that did not turn out as planned since the 24 times made the graphs look really jumbled and made the x-axis look really crammed, ultimately making it harder to read. Since we wanted to show the trend of the plot, we decided to make it into 1 hour intervals.

As our data manipulation developed, some of these questions became more detailed and evolved to: which boroughs are more common for collisions and which are more dangerous? Furthermore, we became curious as to the distribution of accidents/people killed because of vehicle collisions. Our questions became less about which variables had the highest count in a category K and more about the associations between the high collision rates and certain locations within NYC to determine if there existed a certain driving correlation.

For this reason, we then created a map of New York City to look at the distribution of those killed and whether there were any stark differences. To do this we used the select dplyr function to make a data frame of the number of persons killed variable and the associated latitude and longitude, then making a tibble with a variable type of number of persons killed and using an int variable to quantify that for each observation (collision). We did this by doing group\_by latitude and longitude and then summarise by the total. Then, getting a map of New York and using the ggmap function and geom\_point, we plotted the number of killed in an accident, against the x-axis latitude and x-axis longitude and breaking the color down by the number of people in order to portray the location and how many people were killed in accidents around the city of New York.

In order to look at how the average number of accidents per month varied over time within each of the 5 boroughs throughout the years of the dataset, we filtered by the borough variable and then used group\_by to organize by the date of the year and borough and summarised an average accidents per month variable that was the mean of the number of accidents at each of those times and boroughs. From this new data frame, we were able to use geom\_line to create a line graph depicting the ebbs and flows of the average number of accidents by borough across time from 2012 to 2018, which was broken down into 5 colors based on the 5 boroughs.

**－ Final Analysis: What did you learn about the data? How did you answer the questions? How can you justify your answers?**

We answered the first question regarding the most dangerous and safest intersections through tables and bar graphs. These showed that the most dangerous intersections were 3rd Avenue, Broadway, 2nd Avenue, 5th Avenue, and 8th Avenue. Broadway is a very renown road and perhaps one of the busiest in the country in terms of both vehicle and pedestrian traffic, so the fact that there are crashes there does not surprise us. The same goes for third avenue. These are both heavy on tourists and there’s a lot going on so the margin for error is every so slight and a slight eye off the road will cause an accident because of the lack of space. What is interesting though is that although 3rd avenue had over 200 more accidents than Broadway, Broadway had almost 20 more people killed, which is significant relative to the low magnitude of deaths. Likewise, 3rd, 2nd, 5th and 8th Avenues make sense as to why they have a high volume of crashes, since they run very close to big landmarks like the Empire State Building and have high tourist volumes. It could stem from the fact that some people just stop in the middle of the road to observe buildings without regards for other people behind them. Since New Yorkers are generally aggressive drivers, a random stop or even just going slower than normal to look around could easily cause a wreck, even for a split second. On the other hand, we found that the safer intersections were generally private roads that were not as traffic heavy. It’s also important to note that the safest intersections we found had to have at least 1 crash on the road for it to make the collision list, so there could be other roads that had zero crashes on them. It is interesting to note how, of the safest roads with one collision, 2 of them involved 2 people getting injured, probably because the relative speed was faster than that of a traffic ridden Broadway or 3rd Avenue and higher speeds equal higher likelihood of injury. We also derived the probability of an injury or death per crash, which was around 19% of people that get injured or killed per crash. This is an interesting percentage since you would think that a higher amount of people get injured or killed in a crash because of how dangerous cars are. We originally predicted it to be higher, around 30-35%. But, we figured that this may be because New York City is a very busy and jam packed city, so cars are prone to be stuck in traffic a lot and aren’t travelling very fast because of this. Therefore, a majority of the collisions are just fender benders or light taps because of the proximity and clustering of vehicles on the street, and thus a slim portion of them actually result in any harm, no less fatality. Also, it may just simply be because cars have evolved to become a lot safer and more readily prepared to prevent injuries and deaths.

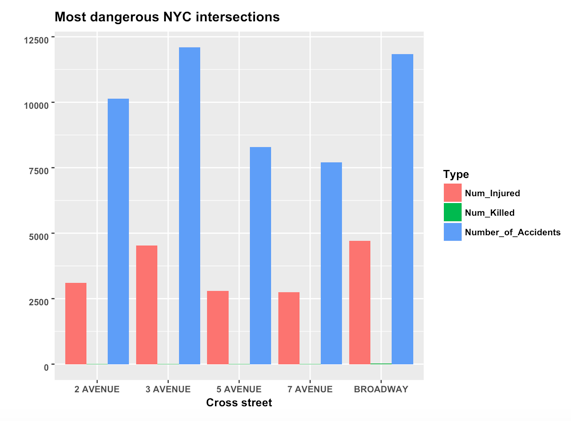
After grouping together different car types and making bar graphs of them, we found that passenger vehicles were the most frequent to cause a collision by over twice the amount of the next most frequent, which was sports utility vehicles. However, this could have been attributed to the large amount and relative proportion of “passenger vehicles” in New York City compared to others, automatically making them more likely to be involved in more accidents because of a much higher volume. Another interesting point to make about this bar plot was that taxis happened to be the third highest, and not much ahead compared to other vehicle types behind it. The point of this result are twofold--you would figure that there are so many taxis in New York and taxi drivers are usually the most aggressive in order to get passengers to their destinations quicker, leading to more accidents. However, it makes some sense and is reassuring when you think about it because they have strangers (customers) in the car and need to exercise due care. What’s most interesting from this plot, and quite striking, is that passenger vehicles and sport/utility vehicles kill the competition with by far with about 1.03 million and 370,000 more collisions, respectively, than the third most common vehicle type in a year. We found that passenger vehicles were involved in twice as many collisions as sport vehicle were and almost 15 times as much as taxis were.

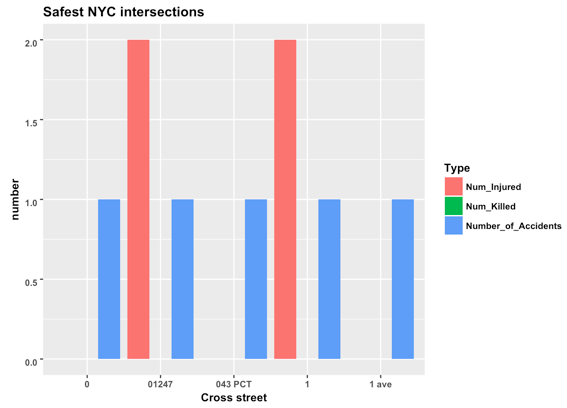
From our contributing factor plot, we found that driver inattention was the leading cause by far with over 3 times as much frequency than any other contributing cause. This can be attributed to the nature of the city, which is generally characterized by a lot of hustle and bustle. If a driver gets distracted even for a split second, then it will almost always result in an accident because of the proximity of vehicles and bumper to bumper traffic, and that split second counts as driver inattention. The rise of commercials and ads regarding texting and driving reaffirm this maximum as a lot of the inattention is probably due to looking at smartphones. Being inattentive on a highway would not result in nearly as many accidents, given the greater time to react and more space to work with. Something that we found very eye-opening was that driver inexperience was the cause with the lowest count of collisions. Driving is hard as it is, but driving in New York City, especially as a new driver, is a whole new level of difficulty. We assumed that maybe driving inexperience was lower since it is generally not what’s reported as the cause for a collision, and something for tangible is normally reported that often results from inexperience, such as a wrong turn, blind spot, or inattentiveness.

Plotting the time intervals of the day against the number of accidents for this year, we found that the number of accidents rises suddenly between 8am and 9am and continues to rise through 3 pm. The maximum hour in terms of accidents was at 5pm, which makes complete sense considering it is rush hour, the time at which most people are commuting home from work. Having a relative spike all of a sudden from the 8am to 9am interval makes sense as well, since most people are commuting to work at this time in the morning and there are, thus, a large number of vehicles on a packed road. After 6pm, the number of collisions consistently decreases, which fits our predictions because people are eating dinner and off the road, leading to a lower chance of an accident. Fitting with this trend is the fact that the lows in terms of accidents occur between 12am and 6am when the roads are pretty much empty. However, because there isn’t a negative correlation between lateness at this time and number of accidents, and the relative maxima for this interval of hours is 1 am and not 12 am, one can infer pretty clearly that this relative peak is largely caused by drunk driving accidents given the time of night.

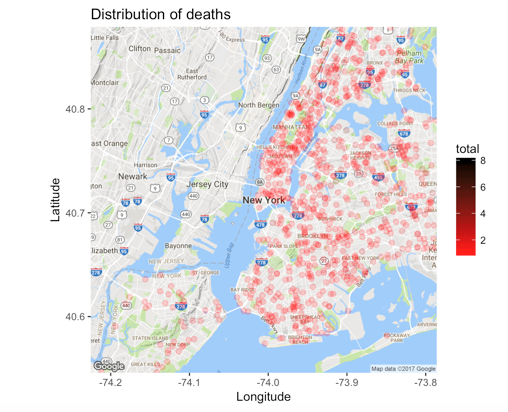
Looking at our map showing the distribution of people killed among the 5 boroughs, it should come as no surprise that the highest concentration of deaths occurred in Manhattan given that it’s not only the busiest of the 5 boroughs, but it is the most dense in terms of streets, traffic, and tightness of lanes. However, the results are more insightful when you compare the map of fatalities with the geom\_line plot of average accidents per month by borough. Within the average accidents per month by borough graph, Manhattan, astonishingly, isn’t even at the top in terms of average number of collisions, though it has the highest clustering of deaths, which we predicted it would given the nature of the traffic there. Brooklyn, surprisingly, is at the top for most points across the timeline for average collisions per month. According to the map, the lowest count of deaths by auto accident and lowest density occurs within Staten Island, which fits with its lowest average number of accidents a month of all the boroughs and the thought process that Staten Island is almost entirely suburban and residential areas that tend to be spread out compared to the more city-like nature of the other 4 boroughs, making it far less dangerous. What is most interesting to note about the average collisions per month by borough is that within each borough, the average collisions a month values peak at the beginning of a new year most likely because of the inclimate weather and high volume of snow and icy road conditions during this time of the year.

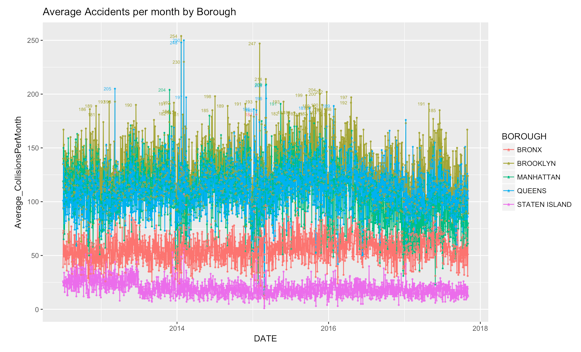
**－ Presentation:**Final Results:



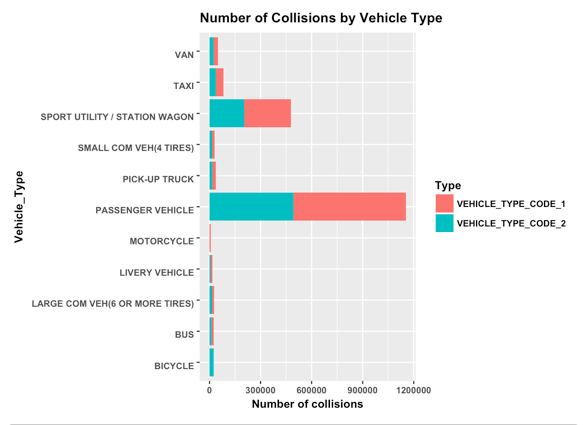
Here we have the 5 most dangerous cross streets in all of NYC according to the number of total accidents on each street, gathered with the number of people injured and number killed on each of those cross streets. Broadway and 3rd Avenue, being two of the busiest streets in the world, are no surprise to be up there in count as traffic is very dense there. Looking closely, we can see number killed in green even though it is such a small quantity compared to number injured and number of accidents. Broadway has 20 more deaths than 3rd avenue even though 3rd has over 200 more accidents. This death count could have to do with the fact that Broadway runs diagonally through another major avenue in New York, which is the recipe for danger. These 5 make sense because they are some of the busiest not just in the city but in the country and have a lot of landmarks.

This plot shows the 5 safest intersections in NYC in terms of lowest number of accidents, but interestly enough, 2 of them have involved 2 injuries.

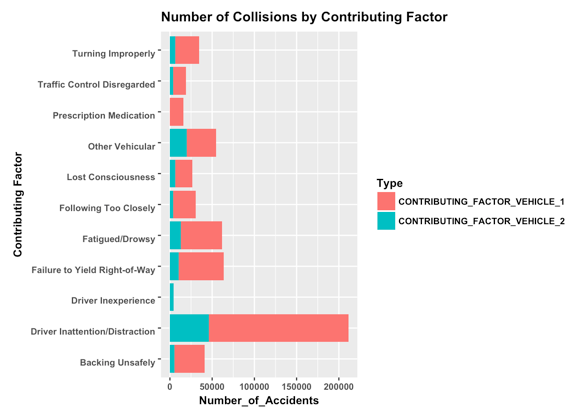
We plotted all the deaths occurring from accidents in this data set onto their location, the gradient of the point being the number of people killed within that accident. Manhattan has the highest concentration of deaths out of all boroughs while Staten Island has by far the lowest, most likely because it’s residential and the others are much more city-like and compressed, leading to more dangerous accidents. In addition, we determined the probability of one being killed or injured in an accident in New York to be about 19%, which is relatively low compared to our previous predictions. We think this is because a large quantity of the accidents that occur in NY are fender benders or light taps simply because traffic and density is so high and vehicles are packed so closely together, making it very easy to bump someone else.



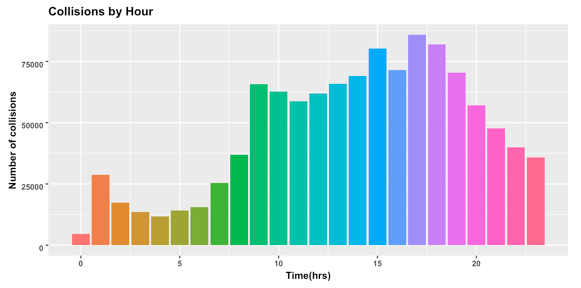
We used geom\_line to show through the whole life of the dataset--from 2012 to 2018--how the average number of collisions per month changes for each borrow. Counter to the results of the ggmap, Manhattan does not even have the highest overall number of collisions per month most of the time. Brooklyn is in first for most of the timeline while Staten Island has the most. The average number of collisions per month peaks for each borrow at the beginning and/or end of each year due to the inclimate weather in New York involving snow and icy roads during this time of the year, causing slippage and lots of accidents.



This bar graph shows the number of total collisions for each type of vehicle. Passenger vehicles were involved in by far the most collisions, with sport utility vehicles not even close in second, but this is probably due to the sheer quantity of passenger vehicles compares to other vehicles, automatically raising the chances of being involved in an accident. To our surprise, no vehicle is even close to that of sport or passenger vehicles, and taxis aren’t really in the running. One would think that taxis would have a higher frequency of accidents due to their ubiquity throughout New York and their contribution to the faced paced mood, wanting to get customers to their destination quickly.



This bar graph shows the number of accidents for each contributing factor from either one of the vehicles involved in it. Driver inattentiveness is by far the leading cause but surprisingly, driver experience is deadlast, although you would think it would be up there because driving in New York City is challenging due to the aggressiveness and fast pace but being a new driver adds a whole nother realm of difficulty to it. This is probably because most accident causes are not reported as driver inexperience, but something more tangible and physical in the moment of the accident.



Lastly, we looked at how often collisions occurred through each hour of the day. It is no surprise that at about 8am, accidents spike and that they are at their max at 5 pm, as these times represent the commute to work in the morning into the city and rush hour when everyone is leaving the city, two of the busiest times on the road in NY.