**Summary**

Amanda Liu 730042603

Michael Debello 720462656

Zachary Dickson 720469328

* What we have tried

1. Dangerous/safest intersections in the city- we mutated new columns named a number of the injured as well as number killed by combining the columns of persons, pedestrians, cyclists, and motorists injured and killed, respectively.
   * + then group\_by street name and summarised by count of the total incidents, count of injured and of killed
     + then did the same for the safest intersections using min of the count of collisions and setting killed and injured = 0 and find the
   1. type of vehicle that caused the collision- we created another new variable (different from #1) and mutated new columns to show the number of collisions caused by different types of vehicles.
      * We started by grouping the first two types of vehicles and then unified them into one so that we could see the different types of vehicles collectively and summarized the total count per vehicle type using a table.
   2. contributing factors that cause a collision- we used the same process that we used for what type of vehicle except used the contributing factors that caused the collisions.
   3. probability that people were injured or killed from the collision- we mutated two new columns that combined the number of people killed or injured for each collision.
      * Then we took the proportion of the number of people either injured or hurt to the number of collisions.
   4. periods when collisions happened- we first needed to change the time periods so that they were consistent throughout, then we grouped all data points together by every 30 minutes.
      * Once we grouped them together we then were able to summarize the number of collisions that happened during each of the 48 time periods.

* What we found

1. The most dangerous street in terms of number of collisions is 3rd avenue while broadway is the most dangerous both in terms of number killed and number injured. Then by arranging by the top 5 counts, we plotted the streets containing the most intersection collisions and grouped bars by count of collision, injured and killed finding that the top 5 streets were broadway, 2,3,5, and 7 avenue. We found that the “safest” streets are 1 ave, the street between 1st and 2nd ave, but still have to manipulate in order to properly find street names as opposed to things like 0 and 043 PCT, which aren’t streets
2. from the vehicle collision table, we found that passenger vehicles(multiple codes), followed by sport utility/station wagon, and then taxis had the top 3 highest collision counts. To my surprise, taxis were not first despite their high presence/quantity within the city and their aggressive driving nature to get people to their destination quickly. Even after combining all vehicle codes into a singular taxi category on the bar graph, taxi was still well in 3rd for the amount of collisions.
3. within the causes of the collisions, we plotted by combining the 2 vehicle types that were involved in the collision to see overall, how many total incidents by either of the vehicles involved in the collision were caused by a certain factor. What we found, breaking the bars down into vehicle 1 causes and vehicle 2 causes, was that overall, driver intention present by one of the drivers was the most common contributing factor, followed by failure to yield and fatigue.
4. based on the proportion of injured plus killed to the total number of collisions, we found the probability of being injured or killed in an accident in NYC to be almost 20%, astonishingly high and showing that in a jampacked NYC, accidents are harmful due to the nature of the driving
5. After establishing 30 minute intervals starting at midnight, we found the count of collisions within those intervals. After plotting, not surprisingly, the lost stretch of collisions occur between midnight and 5 am. The stretch of the highest occurrence rate of collisions occurs between about noon and 6:30 pm. with the highest overall total coming between 4:30 and 5, right around the time rush hour starts in New York and most people are heading home from work. The spike in collisions occurs all of a sudden between 8:30 and 9 am, fitting because that is when most people are coming into work.

* What we have left to do
  1. we still need to find out the root of the problems and why some streets have more collisions than others. Are their intersections with other streets/avenues more dangerous/confusing to navigate?
  2. We need to try to explain and account for the lowest count of collisions and the highest amount of collisions (e.g. the busiest avenue in the busiest city in the world, broadway, obviously has far more traffic than a small residential street somewhere not downtown)- and thus and far higher accident/collision probability
  3. we need to make more sense out of some of our plots- q1 (safest streets)
* This was data collected since July 1st, 2012