Adis Mahmic

Richard Price

Data Intensive Fundamentals

2/5/2025

**Car Crashes in New York City**

1. Car crashes are a very common occurrence in everyday life and are especially prominent in big cities such as New York. With the Motor Vehicle Collision Crashes dataset, I aim to provide evidence for my hypothesis that the number one cause of motor vehicle crashes is texting while driving. The dataset used was pulled from Data.Gov from the New York City Police Department. According to the CDC, over 3,100 people were killed and another 424,000 were injured in the United States due to a distracted driver in 2019.

Investigating this dataset can help inform policy decisions, enforcement strategies, and public awareness campaigns aimed at reducing accidents and improving road safety.

**Documentation**

* Created a folder that would house the local repository and cd to it in GitBash

1. Went to GitHub.com, and created a remote repository to store all of my data in.

* Clicked on code at the top right of the repository page, clicked on SSH and copied the link.

1. In GitBash, I cloned the repository to my folder in order to work on the repository locally.

* I did a git Fetch, then pull.
* Installed Git LFS and used it to track .csv files

1. File Used: <https://data.cityofnewyork.us/api/views/h9gi-nx95/rows.csv?accessType=DOWNLOAD>
2. Used curl to transfer datafile into Motor\_Crash\_Collisions.csv

Curl “https://data.cityofnewyork.us/api/views/h9gi-nx95/rows.csv?accessType=DOWNLOAD” >Motor\_Crash\_Collisions.csv

1. Shown the first 10 rows of Motor\_Crash\_Collisions.csv with:

Head -n 10 Motor\_Crash\_Collisions.csv

The data output was formatted weird, The column names did not line up with the actual columns.

A computer screen shot of a black screen

AI-generated content may be incorrect.

* After careful consideration and looking through all of the columns, I have decided to use Contributing Factor Vehicle 1 and 2, number of persons injured, and number of persons killed, number of pedestrians injured, and number of pedestrians killed. This will give a good basis on what the leading reason for car accidents is and how many people are effected, involved or not.

7. I used the awk command “awk -F, 'NR<=10 {print $11 "\t" $12 "\t" $13 "\t" $14 "\t" $17 "\t" $18 "\t" $19 "\t" $20}' Motor\_Crash\_Collisions.csv” to print only the selected rows shown above, and then stored the results into a separate .csv file, Motor\_Crash\_Collisions\_Extracted.csv.

**Upon inspecting the Xcel file, I noticed that column one, Number of Persons Injured, was also showing results from the street name column. A column that I did not include. I am not sure what is causing this error.**

1. I then used the awk command “awk -F, '{print $3 "," $11 "," $12 "," $18 "," $19}' Motor\_Crash\_Collisions.csv > Motor\_Crash\_Collisions\_InitialData.csv” to gather initial data which includes the same data but also includes BOROUGH’S to see which borough has the most car accidents, and how many people where injured or killed.

I then used another command which showed me the top 10 Boroughs with the most number of car accidents.

$ awk -F, '{print $4}' Motor\_Crash\_Collisions\_InitialData.csv | sort | uniq -c | sort -nr | head -10

A screen shot of a computer

AI-generated content may be incorrect.

**I am not sure what the numerical values without a Borough are, it must be some sort of error.**

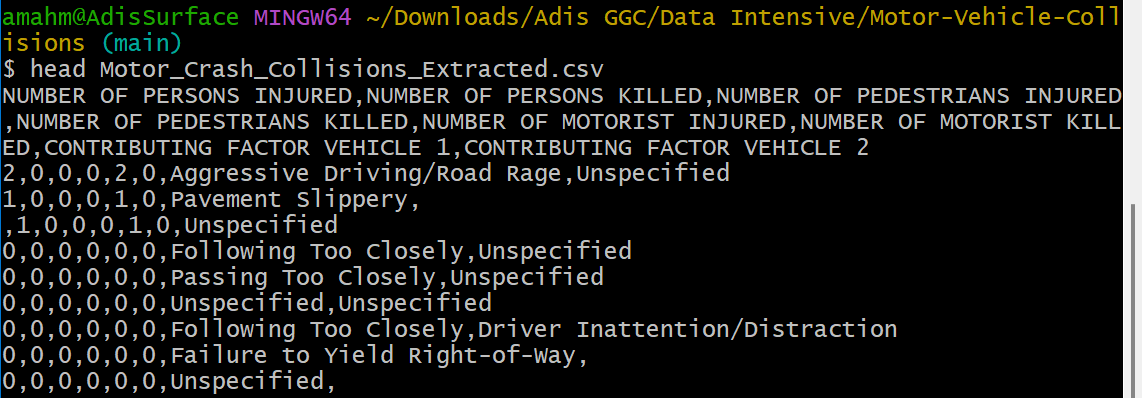
1. Conclusion: The dataset contains a lot of valuable data that if cleaned and manipulated correctly, could yield the results of the most common occurrence of car accidents, and potentially some other information that might be interesting.

**Hadoop Documentation**

4/8/2025

Number 3:

To begin the Hadoop processing job, the head command was run in Git Bash in order to see the columns of the .CSV file and pick which data I wanted to extract for Hadoop to process.



NUMBER OF MOTORIST KILLED, and CONTRIBUTING FACTOR VEHICLE 1 where chosen to extract and use in the Hadoop file Processing.

Created a python program to clean the data even further and narrow down the columns to just NUMBER OF MOTORISTS INJURED, NUMBER OF MOTORISTS KILLED, CONTRIBUTING FACTOR VEHICLE 1, AND CONTRIBUTING FACTOR VEHICLE 2 for easier data processing in the Hadoop job.

After running Hadoop job, the results where different from what I expected. The leading cause of death in driving accidents is Unsafe Speed at 293 killed, and driver Inattention/Distraction in 2nd place 76. I was wrong about texting and driving being the leading cause of death by a long shot. And technically there isn’t a category for texting a driving. I am assuming it is in the Inattention/Distraction column.

My future plan with this data is to write a program to sort the data in order of most deaths to least so that I can see in order the leading cause of deaths while driving.