**CHICAGO TRAFFIC CRASHES ANALYSIS**

By DATA CHAMPS

Team Members

Shivani Konyala

Yeshwanth Nellikanti

**Table of Contents**

[Introduction](#_heading=h.gjdgxs) 2

[Data Qualit](#_heading=h.30j0zll)y 2

[Research Questions](#_heading=h.1fob9te) 3

[Rejected Research Questions:](#_heading=h.3znysh7) 4

[Analysis](#_heading=h.2et92p0) 4

[Presenting the findings](#_heading=h.3dy6vkm) 7

a. What is the hour-wise number of crashes in Chicago from 2018 to 2020? 7

b. Biannually, how many crashes have been reported in Chicago city from 2018 to 2020? 9

c. Biannually, what are the top 5 contributory causes of fatalities reported in traffic crashes from 2018 to 2020? 10

[Conclusion](#_heading=h.1t3h5sf) 13

Works Cited 14

# 

# **Introduction**

As part of the Big Data project, our team Data Champs is tasked with analyzing a large dataset. The team is responsible for selecting a large data dataset, developing research questions to break down the information, performing data analysis, and detailing the results of sightings.

As a team we went through several datasets from the suggested sites, we investigated some of the datasets like a dual credit card, cybersecurity, and a few others. Eventually, the team selected Chicago Traffic Crashes, as it helps to analyze the traffic crashes of Chicago. This dataset has data on the Traffic Crashes in Chicago state from 2013 but the data of all districts are not available until 2017. So, we intended to analyze the data from 2018 as it has information of all the districts of Chicago.

Each record of this dataset represents a traffic crash that occurred in Chicago with necessary columns like Date, time, and primary cause, posted speed, weather, and others. This dataset was selected as the most recent data was available. The dataset selected has complete details of crashes that occurred from 2013 up to March 2021. As it has complete data from 2018 to 2020, we have selected this dataset. Based on this dataset we have developed some of the research questions for analyzing the traffic crashes dataset. We intended to use Chicago Traffic Crashes Dataset with a total record of 488K and the dataset is open source. It can be downloaded from the below link.

(<https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if>)

# **Data Quality**

The traffic crashes dataset contains all the granular details required for the research questions the team has picked. For two of the questions, all the required data is available without any blanks from 2018 to 2020. However, for one of the research questions, we need the primary contributory cause of the crash. The Details in the primary contributory cause are not blank but for most of the records, the cause was unable to determine. Knowing about the details of the primary cause would aid the Chicago traffic department in analyzing the crashes and take necessary measures to avoid crashes in the future. Despite this discrepancy, we have selected to analyze the primary cause for the fatalities occurring in Chicago. The dataset contains 49 columns in total, out of these columns some of the records have blank values but that does not affect the analysis as they are not required for the research questions.

# **Research Questions**

Initially, we came up with three research questions which we have discussed together and submitted for review. We had received a conveying response from the professor on how we should investigate a few of our questions to provide a value that would be useful and placing our questions in an inverted pyramid fashion. Next, we rephrased a couple of our questions and sent for a second review through email to which we got a response to go-ahead for the questions below.

The research questions selected are:

1. What is the hour-wise number of crashes in Chicago from 2018 to 2020?

2. Biannually, how many crashes have been reported in Chicago city from 2018 to

2020?

3. Biannually, what are the top 5 contributory causes of fatalities reported in traffic crashes from 2018 to 2020?

While planning the research questions we also evaluated whether these questions provide any value. We presumed that each of our questions provides answers that would be helpful and important for the Chicago Traffic department.

# **Rejected Research Questions:**

The team has jotted down several research questions and of which below are few research questions that were rejected due to reasons mentioned below.

1. What was the Lighting Condition when most of the crashes happened?

This question was rejected because 3.5% of the lighting conditions data is missing in the dataset and this is a missing data discrepancy.

1. During which Weather Condition the crashes happened?

With a missing data of 4.2% records, the weather condition at which the crash occurred is rejected as the data analysis will not be appropriate.

1. What are the top streets that have the greatest number of crashes in Chicago from 2018 to 2021? (Street = Street No + Street Name)

With the possibility of a street being too long, it would be difficult to find the exact location of a crash based on the street details provided.

# **Analysis**

Below are the SPARK SQL queries to extract the required traffic crashes data.

1. **What is the hour-wise number of crashes in Chicago from 2018 to 2020?**

|  |
| --- |
| HourlyDetails = spark.sql(“  SELECT crash\_hour AS Crash\_hour,  COUNT(\*) AS NumberOfCrashes  FROM global\_temp.TrafficCrashDetails  WHERE YEAR(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2018,2019,2020)  GROUP BY crash\_hour  ORDER BY CAST(crash\_hour AS INT)  ").show(24); |

1. **Biannually, how many crashes have been reported in Chicago city from 2018 to**

**2020?**

|  |
| --- |
| BianuualCrashes = spark.sql("  SELECT  A.Crash\_Year,  A.Biannual,  COUNT(A.Crash\_Record\_ID) AS NumberOfCrashes  FROM (SELECT  YEAR(TO\_DATE(crash\_date,'mm/dd/yyyy')) AS Crash\_Year,  CASE  WHEN crash\_month <7 THEN 'Jan-Jun'  ELSE 'July-Dec'  END AS Biannual,  Crash\_Record\_ID  FROM global\_temp.TrafficCrashDetails  WHERE YEAR(To\_Date(crash\_date,'mm/dd/yyyy')) IN (2018,2019,2020))A  GROUP BY A.Crash\_Year, A.Biannual  ORDER BY A.Crash\_Year, A.Biannual  ").show(); |

1. **Biannually, what are the top 5 contributory causes of fatalities reported in traffic crashes from 2018 to 2020?**

|  |
| --- |
| **i. Jan-June 2018**  FatalityCauses = spark.sql("  SELECT PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2018)  AND Crash\_Month <7  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

|  |
| --- |
| **ii. July-Dec 2018**  FatalityCauses = spark.sql("  SELECT PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2018)  AND Crash\_Month >6  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

|  |
| --- |
| **iii. Jan-June 2019**  FatalityCauses = spark.sql("  SELECT PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2019)  AND Crash\_Month <7  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

|  |
| --- |
| **iv. July-Dec 2019**  FatalityCauses = spark.sql("  SELECT PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2019)  AND Crash\_Month >6  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

|  |
| --- |
| **v. Jan-June 2020**  FatalityCauses = spark.sql("  SELECT PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2020)  AND Crash\_Month <7  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

|  |
| --- |
| **vi. July-Dec 2020**  FatalityCauses = spark.sql("  SELECT  PRIM\_CONTRIBUTORY\_CAUSE,  SUM(CAST(INJURIES\_FATAL AS INT)) AS NoOfFatals  FROM global\_temp.TrafficCrashDetails  WHERE year(TO\_DATE(crash\_date,'mm/dd/yyyy')) IN (2020)  AND Crash\_Month >6  GROUP BY PRIM\_CONTRIBUTORY\_CAUSE  ORDER BY NoOfFatals DESC LIMIT 5  ").show() |

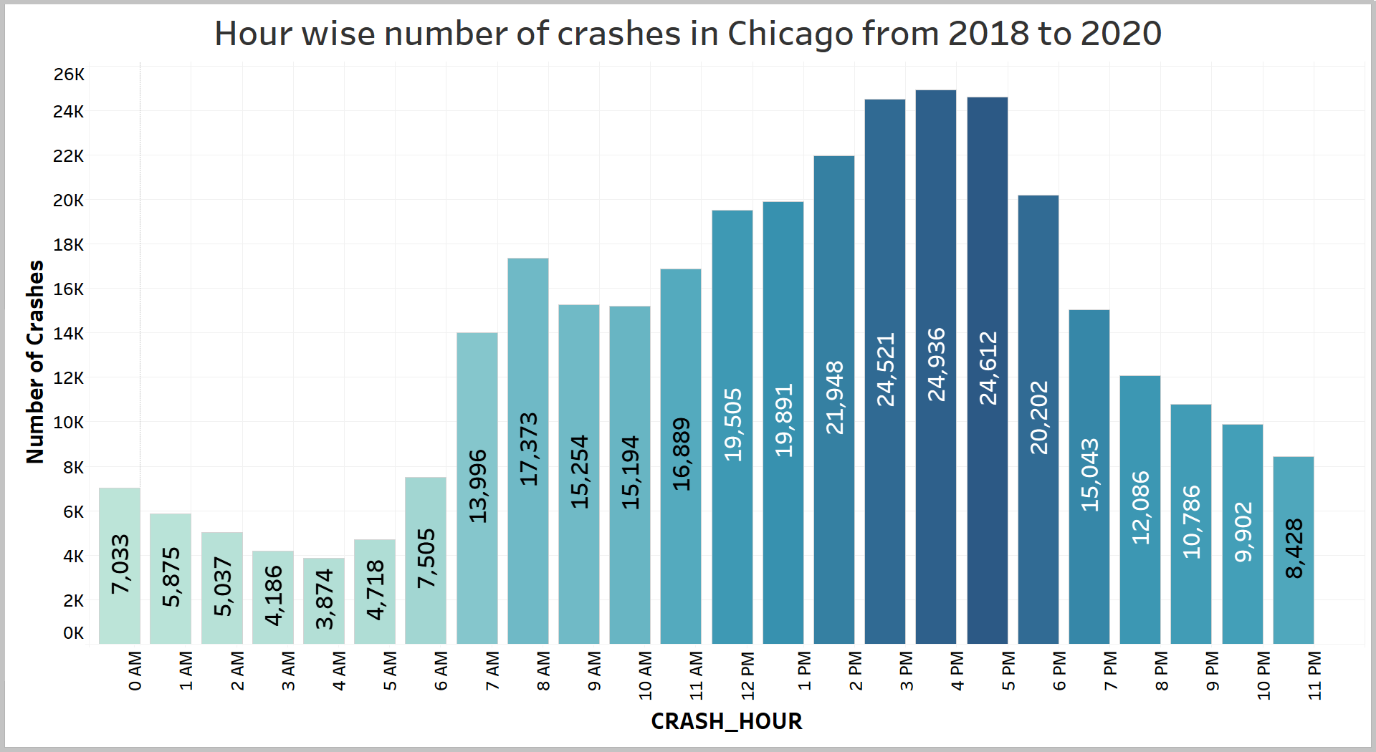
# **Visualizing the Findings**

To visualize the findings of the research questions, we have used tableau software, and excel.

# **Presenting the Findings**

1. **What is the hour-wise number of crashes in Chicago from 2018 to 2020?**

To this research question, we have analyzed the data that covers over the period from January 2018 through December 2020. This was the first research that we thought about and the reason to choose this question is to know at what time the crashes are most likely to occur in Chicago. The results could be useful for the Chicago Police Department (CPD), road safety authorities, and the public. The results are useful for the Chicago Police Department to know at what time they need more officers to control the traffic and to avoid crashes. It helps the Chicago Traffic Department to take appropriate measures for avoiding traffic crashes, and for the public, the hour-wise data would alert them to be more cautious while driving. The findings are visualized in figure1.

**Figure 1. Hour Wise number of crashes in Chicago**

We have used a simple bar graph to represent the findings of the hour-wise number of crashes that occurred in Chicago from 2018 to 2020. The reason to choose this bar graph is to represent the findings clearly with the x-axis representing the hour and the y-axis representing the number of crashes that occurred at that hour.

From the graph, we can observe that the highest traffic crashes have occurred at 4 PM which is around 3.24 percent higher than the adjacent hours. From 3 PM to 5 PM, we can observe that the number of crashes reported is more compared to the other hours. This is due to people rushing back to their homes from offices. From evening around 6 PM to early morning 4 AM, we can observe a steady decline in the number of crashes as it is obvious that the number of people driving at the night hours will be less compared to other hours.

Besides that, we can observe from the bar graph that the number of crashes has suddenly increased at 5 AM. This is because most of the pubs and restaurants are closed at 4 AM in Chicago and some people usually start their work in the early morning’s around 5 AM. There is a steady increase in the number of crashes from 5 AM till 8 AM and there is a decrease in crashes at 9 AM which is due to the fact that people usually reach their office by 8 AM.

As mentioned earlier, these data can be used by the Chicago Police Department (CPD), Chicago Traffic Department, and the public. The Chicago government can plan for high visibility enforcement strategies as a standard practice to create awareness about traffic enforcement. Such kind of actions can reduce traffic crashes with increased execution. Also, they can plan for more officers to be operated when the crashes are high. Public and visitors of Chicago can make use of this data to know when traffic crashes are more likely to occur based on which they can be cautious when driving at that time and schedule their plan accordingly.

1. **Biannually, how many crashes have been reported in Chicago city from 2018 to 2020?**

Chart, bar chart

Description automatically generatedTo this research question, we have analyzed the data that covers over the period from January 2018 through December 2020. This data question was chosen because it would be helpful to the Department of Transportation to find out during which half of the year the traffic crash rate is increasing or decreasing over the period in Chicago city based on which they can find out the reason and take measurable actions to control the crashes. The biannual number of crashes each year is displayed in Figure2.

**Figure2. Biannual, number of crashes from 2018 to 2020.**

We have used a bar chart to visualize the findings of biannually, the number of crashes recorded in Chicago. The label represents the total number of crashes. The x coordinate shows the time frames from January to June and July to December. A bar chart is used because it shows the number of crashes well by dividing three consecutive years into half.

The visual display shows that crashes are most common in the second half of every year. Though in the first two years of the first half (January-June) crashes are close to the second half (July-December). The most notable measure in the result is a significant difference that can be seen in the third year where the first half crashes are 35 percent less, while the second half is 11 percent less than the previous two years.

The data leads us to believe that, though the total number of crashes has gone down over the period, they are common in the second half of every year. This could be due to reasons like the weather is warm enough in summer for the locals to go to the beach and do outdoor activities. In addition, most of the visitors visit Chicago city during the second half of the year as it is the best time for concerts, holiday events and many of the festivals take place.

The Department of Transportation can make use of this data. They can compare the results of both biannual periods, support the development of useful tools, implementation, evaluation of safety programs that tend to reduce road traffic crashes.

1. **Biannually, what are the top 5 contributory causes of fatalities reported in traffic crashes from 2018 to 2020?**

This research question is picked to analyze the top 5 contributory causes of fatalities in Chicago from January 2018 to Dec 2020 on a biannual basis. There are 41 different primary contributory causes in the dataset. Out of all the causes, we are interested to investigate the most contributory causes which are leading to death. The data analysis helps the Chicago traffic department to invest their money and time on the measures which can reduce traffic crashes. The traffic department can create necessary public awareness and impose new rules to get control over the traffic crashes.

In figure 3, the timeline is of 6 intervals from 2018 to 2020 divided biannually. Ranks are allocated from 1 to 5. Rank 1 describes the top primary contributory cause which leads to the highest fatalities in the biannual period. Similarly, the next ranks listed give the information about the immediate primary contributory causes in descending order. For each cycle, the top 5 contributory causes of fatalities as per the dataset are given the ranks from 1-5. The topmost primary contributory causes are unable to determine, the physical condition of the driver, failing to yield, disregarding traffic signals, failing to reduce speed, exceeding the speed limit, and operating vehicle issues. These factors are leading to the death of people in traffic crashes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | **Jan-Jun (2018)** | **Jul-Dec (2018)** | **Jan-Jun (2019)** | **Jul-Dec (2019)** | **Jan-Jun (2020)** | **Jul-Dec (2020)** |
|  |
| **1** | **Unable to determine 16** | **Unable to determine 18** | **Unable to determine 16** | **Unable to determine 12** | **Unable to determine 23** | **Unable to determine 34** |  |
| **2** | **The physical condition of Driver 10** | **Failing to Yield 9** | **Disregarding Traffic Signals 8** | **Failing to Reduce speed 8** | **Failing to Yield 7** | **The physical condition of Driver 13** |  |
| **3** | **Exceeding Speed Limit 9** | **Disregarding Traffic Signals 5** | **The physical condition of Driver 7** | **Failing to Yield 8** | **The physical condition of Driver 6** | **Disregarding Traffic Signals 11** |  |
| **4** | **Failing to Reduce speed 4** | **Exceeding Speed Limit 5** | **Operating Vehicle Issue 4** | **Disregarding Traffic Signals 4** | **Disregarding Traffic Signals 5** | **Failing to Yield 9** |  |
| **5** | **Failing to Yield 4** | **Failing to Reduce speed 5** | **Failing to Reduce speed 2** | **The physical condition of Driver 4** | **Failing to Reduce speed 4** | **Failing to Reduce speed 8** |  |

**Figure 3: Biannual, top 5 prime contributory causes for fatalities from 2018 to 2020.**

From the data in the grid, we can observe that in each cycle the Rank 1 contributory cause is recorded as “Unable to determine”. While briefing the incident happened, if the responsible officer could not find or recognize the factor which led to the crash then the officer updates the contributory cause as “Unable to determine”.

In almost every cycle we can see that the physical condition of the driver, failing to yield, failing to reduce speed, and disregarding traffic signals have caused severe traffic crashes. If the driver is intoxicated with alcohol or did not have proper knowledge of traffic rules, speed limits, school zones, then the driver becomes a prime cause for the traffic crash. Failing to yield and disregarding traffic signals causes severe collisions where the cars with high speed may collide and people involved in such crashes are most likely to be severely injured.

This data can be used by the Chicago traffic department to find out the most severe contributory causes of traffic crashes. With further camera installations and speedometers, it would help the officers to find the exact reason for the cause. The Traffic department should create awareness to the public about the driving rules, speed limits, and signals.

# **Conclusion**

With this project, we have learned how to analyze the large dataset with the Hadoop platform. Hortonworks sandbox helped to manage Big data and with advanced options provided to use spark in Ambari made it easier to work with Spark SQL. Working with spark helped us to extract the data quickly. We have chosen Python as a domain-specific language to manipulate the views created. With this project, we got an opportunity to explore the Hadoop platform.

Initially, we have downloaded our large dataset and uploaded it to HDFS. Once the file was uploaded, we started to explore the hive and observed that each query was taking a lot of time to execute due to the settings. Later, we moved to explore the spark and made use of a web shell client available in advanced links of Hortonworks. In the web shell client, we imported pyspark and spark session, read the CSV file, and stored the CSV file data as a temporary view. In the next steps, we have used this temp view to query the data using Spark SQL. Working with spark SQL felt exciting as it was easy and like SQL with some changes. We were able to find out the data quality issues, and further picking the research questions which added value to the project made us think differently. Visualizations helped us to present our findings more understandably.

We cannot consider anything we would do differently if we must do this project again. However, if there is not any missing data for weather and lighting conditions, we would have considered it as one of our research questions to determine which weather condition was most prevalent for each type of crash.

In conclusion, this project helped us to learn different big data concepts and we were able to apply the concepts learned in class. The Project provided us with a better hands-on experience to explore big data concepts.

**WORKS CITED**

*Chicago Data Portal: Chicago Traffic Crashes* [**https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if**](https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if)