BumbleC

Report

By

Mr.	**Kulawut**	**Makkamoltham**	**6388030**
Miss	**Ariya**	**Phengphon**	**6388040**
Mr.	**Khunathip**	**Suravanit**	**6388064**
Mr.	**Peerawat**	**Sorosthunyapong**	**6388104**
Miss	**Pitchaya**	**Teerawongpairoj**	**6388133**
Miss	**Sasima**	**Srijanya**	**6388196**

A Report Submitted in Partial Fulfillment of
the Requirements for

ITCS495 Special Topics in Database and Intelligent Systems

Faculty of Information and Communication Technology
Mahidol University

2023

Table of Contents

Introduction

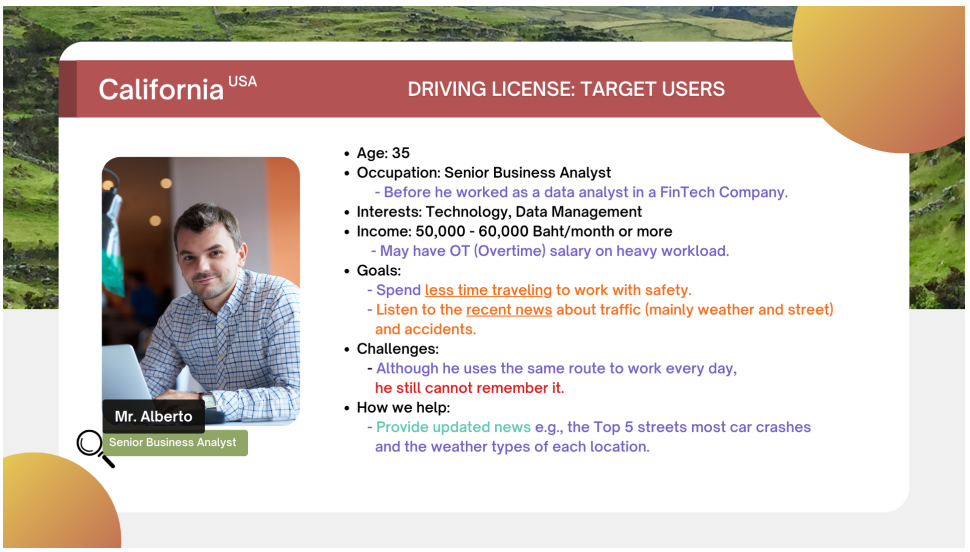
Business Domain Overview

Our dataset, centered around car crash incidents, is extensive and multifaceted. This means that its comprehensive details can be interpreted and applied across a range of business domains, particularly those related to routes and vehicle incidents involving cars. It allows researchers to tailor their focus and problem-solving approach within these domains. After the discussion and gaining a deeper understanding of these attributes, it becomes obvious that several common conditions can contribute to car incidents. These include several common conditions that may cause the car incidents such as weather conditions (e.g., clear, rain) and road details (e.g., roadway surface and road types).

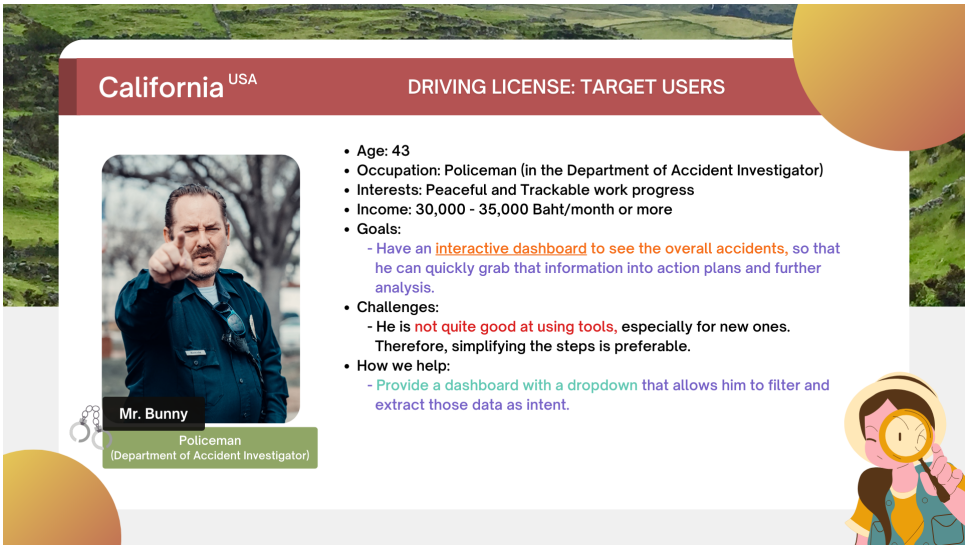
When these factors are integrated, in the short term, it can serve as a valuable resource allowing users to access this relevant data in their daily lives. This information can be readily incorporated into navigation systems or news updates to provide timely warnings and alerts, ensuring that individuals are informed and prepared when they are in hazardous situations on the road. In the long term, this dataset can be leveraged for predictive analysis to forecast and allocate resources to specific locations. This may involve implementing warning signs, adjusting speed limits, or installing traffic lights where necessary. Through analysis of the dataset, the aim is to minimize accidents and mitigate their impact.

Target Users

There are three primary target user groups, each selected based on their potential involvement and the benefits they can derive from the car crash dataset. These user groups possess distinct characteristics and responsibilities, which consequently influence their specific perspectives and requirements when interacting with the dashboards, discussed in the **Objectives and Scope of the project.** To be more specific on their objectives, the below figure describes the user persona of each group. (However, when creating a dashboard, we select User 1 Commuters and Drivers and User 2 Law Enforcement Agencies to work on.)



1) **Commuters and Drivers:** Individuals can access accident data to plan their routes more safely, avoid or increase awareness in high-accident areas, and make informed decisions about their commutes.



2) **Law Enforcement Agencies (Police Departments, Department Of Transportation (DOT)):** Police departments or related organizations can utilize this dataset to effectively allocate resources such as barricades, warning plates, and maximum speed limit signs, and enforce traffic laws in high-risk areas to minimize accidents and provide immediate assistance when needed.

3) **Traffic Safety Apps and Navigation Services:** Mobile apps and navigation services (e.g., Google Maps, Waze) can integrate the data to provide real-time alerts, offer alternate routes, and enhance the overall driving experience.

Objectives and Scope of the Project

<u>User 1:</u> Mr. Alberto (Commuters and Drivers)

Background: Alberto, a senior business analyst, faces a daily challenge during his commute to work. The issue is the unpredictable traffic jams and incidents that could happen at any time. Therefore, knowing which road areas or other factors (e.g., weather types) should be avoided would help him to be safer and arrive at his work quicker.

- **Question 1:** What is the top 5 most street that has car crashes?
    - **Solution 1:** Create a *Bar Chart* to display the top 5 streets with the highest number of car accidents, providing Mr. Alberto with a clear visual image of which streets to avoid.
- **Question 2:** What are the top 5 primary factors of car crashes with the most severe type of injuries that occur on the streets?
    - **Solution 2:** Visualize the top 5 primary factors of car crashes in Chicago through a *Stacked Bar Chart* with the proportion of the most severe injuries (the maximum type of injuries) in each factor. This will assist Mr. Alberto in determining which primary factors of crashes and seeing how much they cause injuries are more common in the city.
- **Question 3:** What is the number of total cases (in percentage and number), categorized by the crash types?
    - **Solution 3:** Use a *Card* to display the number of all cases that happen. This will assist Mr. Alberto in distinguishing the difference between injury and no-injury types.
- **Question 4:** What is the total number of injuries (people) involved in each case?
    - **Solution 4:** Use a *Multi-row Card* to show the number of injuries of Fatal, Incapacitating (Severe hurt, Unable to move), and Non-incapacitating (Moderate to Small hurt). This will assist Mr. Alberto to see the difference between the three groups of injuries.
- **Question 5:** What day of the week has the most car crashes occurred?
    - **Solution 5:** Use a *Line Chart* to display crash frequency by day of the week to allow Mr. Alberto to decide which day he should go outside.
- **Question 6:** What times of the day have the most car crashes occur?
    - **Solution 6:** Use a *Line Chart* to display crash frequency by hour to allow Mr. Alberto to decide whether to leave for work sooner or later to avoid peak accident times.
- **Question 7:** Analyze the relationship between the different types of weather and the total injuries that happened during car crashes.
    - **Solution 7:** Employ a *Donut Chart* to visualize the distribution of car crash injuries across different weather conditions in order to avoid when the weather (e.g., clear, rain) can cause a high possibility of car crashes.
- **Question 8:** Analyze the relationship between the different types of lightning conditions and the total injuries that happened during car crashes.

    - **Solution 8:** Employ a *Pie Chart* to visualize the distribution of car crash injuries across different lighting conditions in order to avoid when the lightning (e.g., day, night) can cause a high possibility of car crashes.
- **Question 9:** How can we effectively visualize the distribution of car crashes on a map to identify hotspots and patterns?
    - **Solution 9:** Generate a *Heat Map (Power BI Visuals)* to illustrate the distribution of car crashes in each location in Chicago. This will assist Mr.Alberto in seeing which zones of the country he should avoid and be more cautious in specific areas where these devices are located.

<u>User 2:</u> Mr. Bunny (Law Enforcement Agencies)

Background: Mr. Bunny can use the analyzed data to reduce accidents by enforcing the law. For example, some streets are under construction, and some of them have a higher accident rate than others. Therefore, the Law Enforcement Agencies need to allocate barriers or signs to prevent possible accidents. In addition, there are many cases in which people violate traffic laws, and Mr. Bunny can enforce more punishments for appropriate incidents.

- **Question 1:** Which areas have the highest frequency of accidents, and which territory police officer is assigned to patrol (beat id)?
    - **Solution 1:** Use an iConMap *V3* to visualize the territory of which zones have many crashes so that they can provide more police/law enforcement agencies to the required area to facilitate the users and prevent the workload.
- **Question 2:** Show the car crash frequency between years and months.
    - **Solution 2:** Use a *Matrix* to visualize the history of accidents to see the overall image separated by months and years. This tabulated format allows for a comprehensive view of the data, enabling the relevant person to gain insights into accident patterns and related information.
- **Question 3:** Which Chicago Police Department Beat ID needs to employ more or less staff based on the Beat presence at the accident scene and report type (Not On Scene/ On Scene)?
    - **Solution 3:** Create a *Stacked Column Chart* to visually represent the correlation between Beat ID occurrences and report types. This chart will help determine if the Chicago Police Department should consider increasing police officers in areas with an increase in both accident rates and desk reports. (Use with Beat ID Filter)
- **Question 4:** What is the relationship between weather and road surface conditions to see the number of car accident rates in each type of weather?
    - **Solution 4:** Use a *Stacked Bar Chart* to analyze the total amount of car accidents to see which type of weather has the highest rate of accidents, and in each weather, split each of them into road surface conditions to see the relationship between weather and road surface conditions. Therefore, the relevant department to address road surface issues and plan repairs for safer road conditions.
- **Question 5:** What is the average duration in seconds for each Police Beat to notify a crash based on the type of report?

    - **Solution 5:** Use a *Bar Chart* to show the time taken for police to notice the crash for each beat and categorize by the type of report. This information can help each Beat assess their response speed and use the data for training and improving their efficiency in future operations.

- **Question 6:** What is the number of crashes by the trafficway type?
    - **Solution 6:** Use a *Bar Chart* to show the highest number of crashes according to traffic way types (e.g., The undivided road lane has the highest number of crashes.) This helps the police to notice which types of roads possibly cause an accident the most so that they can allocate more resources to reduce the number of crashes.
- **Question 7:** What is the number of crashes in each traffic control by traffic control device condition?
    - **Solution 7:** Use a *Stacked Bar Chart* to show which types of traffic control are related to each crash. For instance, no control causes the most car crashes and splits to see whether the devices are working fine or not, if not then we can report them to the nearest station for fixing.
- **Question 8:** What streets that have a road defect contribute to the crash 5% or more?
    - **Solution 8:** Use a *Bar Chart* to show the defective street names that cause crashes for more than and equal to 5 percent of crashes. This helps the staff to pay more attention to those streets that have a high number of crashes, and allocate staff to those areas.
- **Question 9:** What are the primary causes of Fatal or Incapacitating injuries based on percentage?
    - **Solution 9:** Use a *TreeMap* to show the percentage of people in the Fatal or Incapacitating injuries group (severe), so that they can present cause people to get more damage and find ways to alleviate the number in these groups.
- **Question 10:** What are the trends of crashes in the week of day colored by a group of damage caused?
    - **Solution 10:** Use a *Ribbon Chart* to show the number of crashes each day (from Sunday to Saturday) and color separated by the amount of damage caused. This can help the police to see how severe the crashes occur and estimate the damage in money.
- **Question 11:** Show the increase/decrease of all cases.
    - **Solution 11:** Use a *Waterfall Chart* to show the increase and decrease by year. This can help the staff to see the number changes in cases, to see how many cases it causes to be solved in each year.

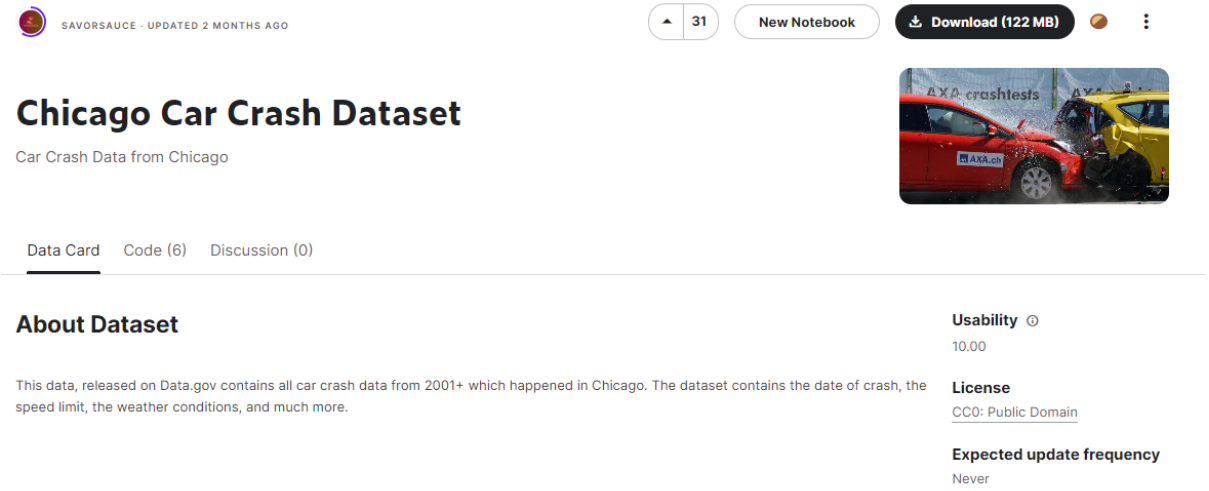
<u>User 3:</u> Ms. Camelon (Representative of Traffic Safety Apps and Navigation Services)
Background: (Didn't include this user in the dashboard)

The goal of increasing road safety and improving the driving experience for users is aligned with Ms. Camelon's mission to improve traffic safety applications and navigation services through the integration of data-driven insights.

- **Question 1:** Which route streets are the safest (low accident rates) for users considering the current weather conditions?
    - **Solution 1:** Utilize the *Stacked Bar Chart* and a Card to illustrate the streets with the lowest accident rates, along with the weather conditions. This way Ms. Camelon can provide the users with safe street routes with the customized weather filter by using Weather Filter.
- **Question 2:** What are the most common primary and secondary contributory causes of car crashes in Chicago, and is there a correlation between these causes and the severity of injuries?
    - **Solution 2:** Create a *Grouped Bar Chart* showing the top primary and secondary contributory causes. Additionally, use a scatter plot to illustrate whether there is a correlation between these causes and the severity of injuries. Ms. Camelon can use this information to address not only the primary but also secondary causes of accidents and assess their impact on injury severity.
- **Question 3:** Where are the high-density accident areas in Chicago, and are those areas with more accidents during low-light conditions?
    - **Solution 3:** Create a *Heat Map* to display high-density accident locations in Chicago based on lighting conditions. Ms. Camelon can use this information to identify areas with more accidents during low-light conditions and plan for user alerts about lighting in those areas.
- **Question 4:** What are the most common types of trafficway, and do they have any correlation with the severity of accidents in Chicago?
    - **Solution 4:** Use matrix visualization with 'TRAFFICWAY_TYPE' in the row and 'MOST_SEVERE_INJURY' in the column. This can provide a tabular view of the data with counting. and help identify patterns between traffic types and injury severity. Ms. Camelon can then use this information to assess whether certain types of trafficways are more prone to severe accidents.
- **Question 5:** How do different types of crashes vary based on the number of units involved and the day of the week?
    - **Question 5:** Create a clustered column chart with 'NUM_UNITS' on the x-axis and the number of accidents on the y-axis. You can use different colors for the columns to show different days of the week ('CRASH_DAY_OF_WEEK'). This chart will show different numbers of related units. according to the day of the week. This analysis can help Ms. Camelon understand patterns in accidents and tailor traffic safety strategies according to specific crash types and days of the week.

Data Description

Dataset Name: Chicago Car Crash Dataset [2]
Usability: 10.0 (scored by Kaggle)
Tags/Categories: Tabular, Automobiles and Vehicles, Beginner, United States, North America



Data Overview

This dataset, released on Data.gov, encompasses all car crash data between 2013 and 2023 which happened in Chicago, United States of America (USA). The dataset includes information such as the crash date, speed limit, weather conditions (e.g., clear, rain), the traffic control device or nearby crash location (e.g., pedestrian crossing sign, school zone), and road details such as roadway surface and road types. With this comprehensive dataset, various aspects can be analyzed and leveraged to gain insights into traffic accidents, particularly those involving cars. It serves as a valuable resource for understanding the patterns and underlying causes of these incidents.

The dataset consists of 746,498 records. Each record has CRASH_RECORD_ID with a fixed length of 128 characters as a key. For more explanation of each attribute, see the **Field Description section [1]** below.

Data Characteristics

1. **Reliability**
   The data is reliable because the data comes from An official website of the United States government (data.gov)
2. **Originality**
   The publisher provided the source of this dataset, which can be found in Traffic crashes - crashes[3]. The data hasn't been modified, so it is considered original.
3. **Comprehensiveness**
   The records are reported by the administrative department, which is reliable. Moreover, this dataset can be used to analyze to find causes of the crash and improve road quality. Therefore, this dataset is comprehensive.

4. **Current**
   The data is still current since it was released in July 2023 on the Kaggle platform. In addition, The latest data recorded in this dataset is on 30 July 2023 at 8:50:00 PM
5. **Cited**
   Since this data is created by the government department, the maintainer, Jonathan Levy is mentioned on the website. Thus, this dataset is cited

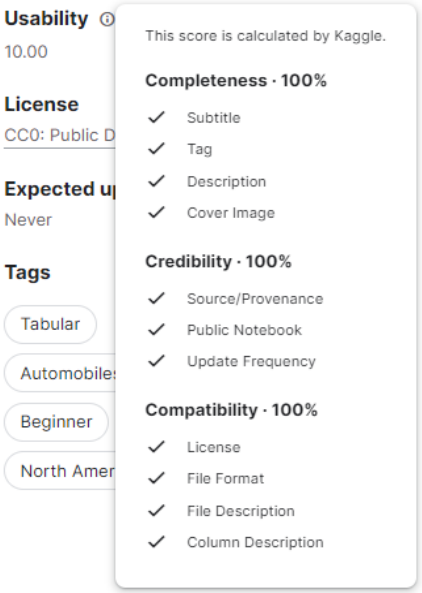
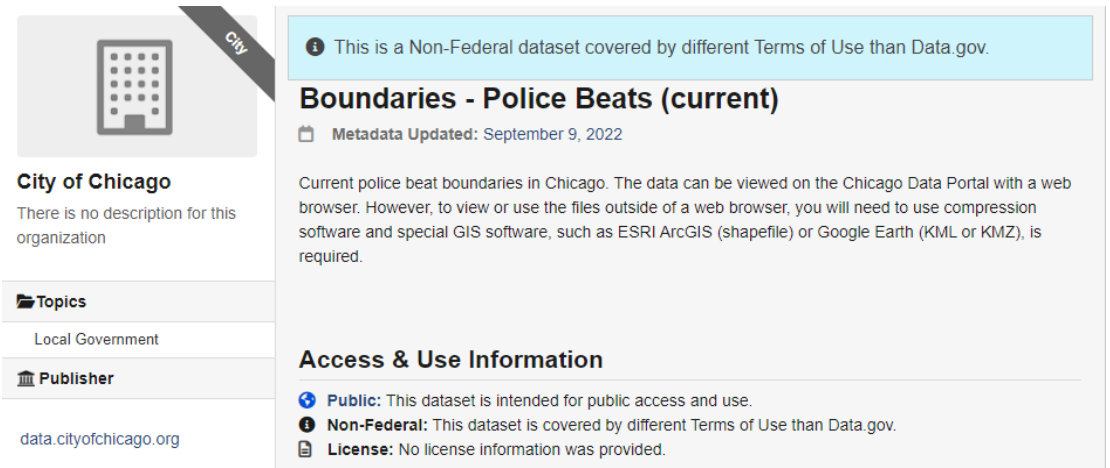


Figure 1: Data Scoring by Kaggle

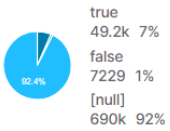
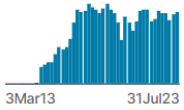
Additional Dataset

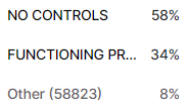
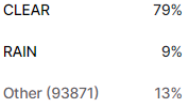
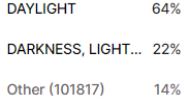
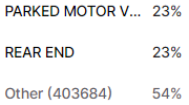
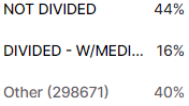


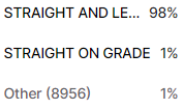
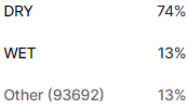
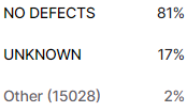
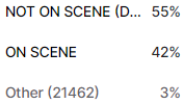
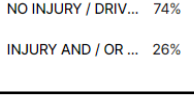
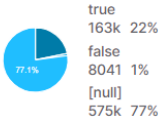
Dataset Name: Boundaries - Police Beats (current) [4]
Description: This dataset contains 277 records with 4 fields including

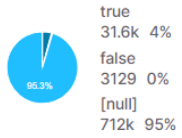
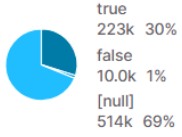
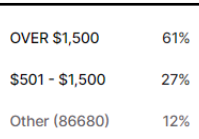
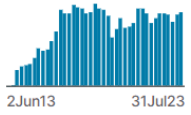
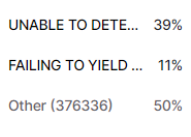
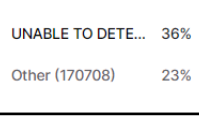
- the_geom - Text field containing Well Known Text for Police Beat Boundaries.
- District - Integer field containing District number (area)
- SECTOR or BEAT - Integer field containing Beat or Sector number
- BEAT_NUM - Integer field containing Beat number which is the smallest unit.

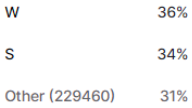
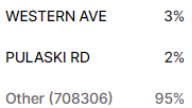
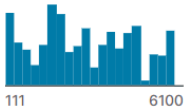
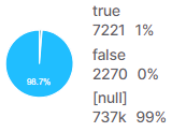
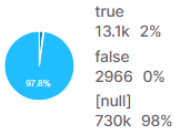
Field Description [1]

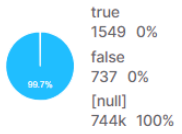
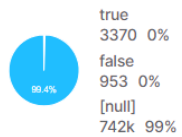
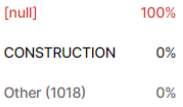
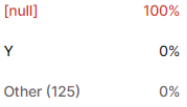
Column Name	Distribution	Data type	Description	Null records	Unique records
CRASH_RECORD_ID		Text	Unique Record ID serves as a primary key in this dataset	0	746498
RD_NO		Text	Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.	4307	742192
CRASH_DATE_EST_I		Text	Crash date estimated by desk officer or reporting party (only used in cases where a crash is reported at a police station days after the crash). (Y, N)	690109	2
CRASH_DATE		DateTime	Date and time of the crash	0	> 10000
POSTED_SPEED_LIMIT		Number	Posted Speed limit	0	45

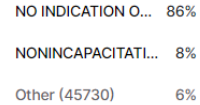
Column Name	Distribution	Data type	Description	Null records	Unique records
TRAFFIC_CONTROL_DEVICE		Text	Traffic control device present at the crash location	0	19
DEVICE_CONDITION		Text	The condition of the traffic control device	0	8
WEATHER_CONDITION		Text	Weather condition at the time of crash	0	12
LIGHTING_CONDITION		Text	Lighting condition at the time of crash	0	6
FIRST_CRASH_TYPE		Text	Type of the first collision in crash	0	18
TRAFFICWAY_TYPE		Text	Trafficway type https://masscrashreportmanual.com/crash/trafficway-description/	0	20

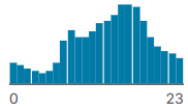
Column Name	Distribution	Data type	Description	Null records	Unique records
LANE_CNT		Number	Total number of through lanes in either direction, excluding turn lanes	547494	41
ALIGNMENT		Text	Street alignment at crash location	0	6
ROADWAY_SURFACE_COND		Text	Road surface condition	0	7
ROAD_DEFECT		Text	Road defects	0	7
REPORT_TYPE		Text	Administrative Report Type	21222	3
CRASH_TYPE		Text	Crash Type that classify the general severity	0	2
INTERSECTION_RELATED_I		Text	An observation by the police officer whether an intersection played a role in the crash. (Y, N)	575368	2

Column Name	Distribution	Data type	Description	Null records	Unique records
NOT_RIGHT_OF_WAY_I		Text	Whether the crash began or first contact was made outside of the public right-of-way. (Y, N)	711724	2
HIT_AND_RUN_I		Text	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid (Y, N)	513706	2
DAMAGE		Text	Estimated damage	0	3
DATE_POLIC_NOTIFIED		DateTime	Date when the police notify the crash	0	> 10000
PRIM_CONTRIBUTORY_CAUSE		Text	The most significant cause of the crash judged by the officer	0	40
SEC_CONTRIBUTORY_CAUSE		Text	The second most significant cause of the crash judged by the officer	0	40

Column Name	Distribution	Data type	Description	Null records	Unique records
STREET_NO		Number	Street Number	0	> 10000
STREET_DIRECTION		Text	Street Direction	4	4
STREET_NAME		Text	Street Name	1	1627
BEAT_OF_OCCURRENCE		Text	Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74	5	276
PHOTOS_TAKEN_I		Text	Whether a photo of the crash taken by an officer (Y, N)	737007	2
STATEMENTS_TAKEN_I		Text	Whether statements were taken from the unit(s) involved in crash (Y, N)	730402	2

Column Name	Distribution	Data type	Description	Null records	Unique records
DOORING_I		Text	Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash (Y, N)	744212	2
WORK_ZONE_I		Text	Whether the crash occurred in an active work zone (Y, N)	742175	2
WORK_ZONE_TYPE		Text	Work zone type	743128	4
WORKER_PRESENT_I		Text	Whether construction workers were present in an active work zone at crash location (Y, N)	745383	2
NUM_UNITS		Number	Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.	0	17

Column Name	Distribution	Data type	Description	Null records	Unique records
MOST_SEVERE_INJURY		Text	Most severe injury sustained by any person involved in the crash	1630	5
INJURIES_TOTAL		Number	Total persons sustaining fatal, incapacitating, non-incapacitating.	1619	20
INJURIES_FATAL		Number	Total persons sustaining fatal injuries in the crash	1619	5
INJURIES_INCAPACITATING		Number	Total people who suffered incapacitating injuries which make them unable to do normal activities.	1619	10
INJURIES_NON_INCAPACITATING		Number	Total people who suffered non incapacitating injuries such as bruises, abrasion, or small cuts that don't prevent them from engaging in normal activities.	1619	19
INJURIES_REPORTED_NOT_EVIDENT		Number	Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness,	1619	13

Column Name	Distribution	Data type	Description	Null records	Unique records
			claims of injuries not evident, limping, complaints of pain, nausea, and hysteria.		
INJURIES_NO_INDICATION		Number	Total persons sustaining no injuries in the crash	1619	46
INJURIES_UNKNOWN		Number	Total persons for whom injuries sustained, if any, are unknown. This field only contains 0	1619	1
CRASH_HOUR		Number	Hour of the crash	0	24
CRASH_DAY_OF_WEEK		Number	Day of the week of the crash	0	7
CRASH_MONTH		Number	Month of the crash	0	12

Column Name	Distribution	Data type	Description	Null records	Unique records
LATITUDE		Number	Latitude of the crash location	4908	> 10000
LONGITUDE		Number	Longitude of the crash location	4908	> 10000
LOCATION		Point	Crash location	4908	> 10000

Data Manipulation

Workflow

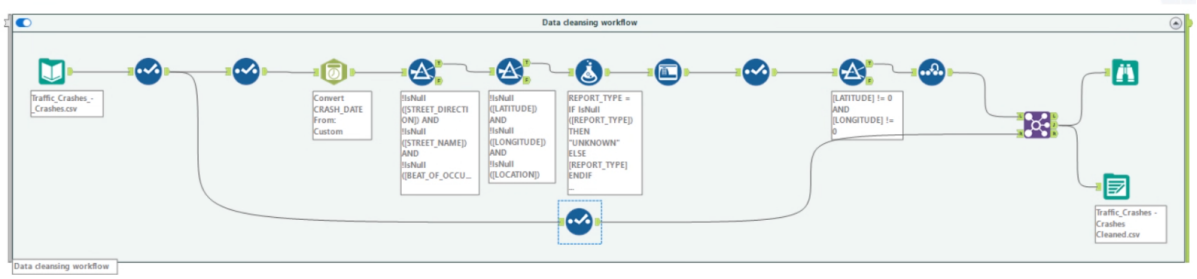


Figure 2: Data cleansing workflow

Description: This workflow is used to clean the given dataset, see Modifications section for changes in the dataset. The first selection is used to remove columns containing too much null such as DATE_EST_I, HIT_AND_RUN_I, LANE_CNT, etc. Then the second Selection tool is used to select columns to be cleaned while the Selection at the bottom will pick the columns that don't need to be processed. In the Formula tool, there are 3 formulas, converting null values of REPORT_TYPE and RD_NO to null and changing the datetime format to day of the week (CRASH_DAY_OF_WEEK_TEXT). Then, the next 2 filters are used to delete rows with null values on several columns. Next, Autofield is used with the Selection tool to automatically and convert the datatype and the Selection tool allows the inspection of the converted datatype. After that, The last filter filters latitude and longitude containing 0. After that, the imputation is used on the field with the INJURIES prefix to replace the null value with the median. The reason why the median is used is that the dataset mostly contains 0, using the mean will just cause the complication since it has decimal points when averaging. On the other hand, using mode should work the same, but in the long run, the median might be more effective. Finally, combine the columns using the Join tool, which is slightly faster than using one lane without filtering used and unused columns.

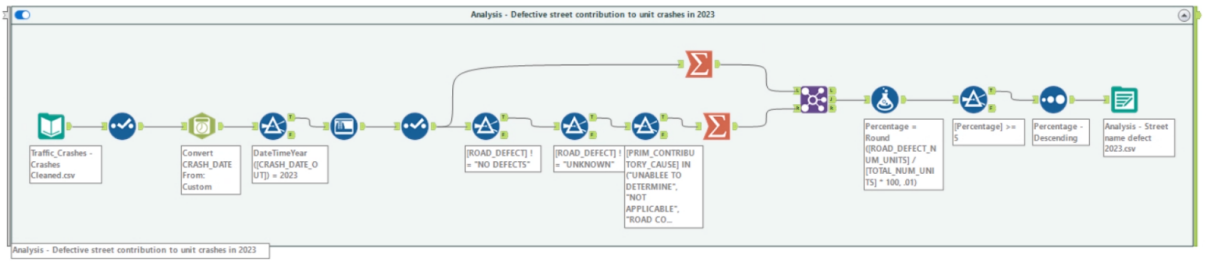


Figure 3: Workflow of street names that have road defect contribution to the number of units involved in the crash 5 percent or more in 2023

Description: This workflow was used to analyze complex data, listing the street names that have road defect contribution to the number of units involved in the crash 5 percent or more in 2023. First, use the filter to get only the needed columns namely, CRASH_DATE, ROAD_DEFECT, CRASH_TYPE, PRIM_CONTRIBUTORY_CAUSE, STREET_NAME, NUM_UNITS. Next, convert the datetime to Alteryx datetime and filter only the year 2023. Then, use autofield and select to automatically change the datatypes. After that, filter out NO

DEFECTS street and UNKNOWN. Moreover, filter PRIM_CONTRIBUTORY_CAUSE to get only the target Primary cause of the crashes. Next, use the Summarization tool to calculate the Summation of NUM_UNITS grouped by STREET_NAME, ROAD_DEFECT_ AND PRIM_CONTRIBUTORY_CAUSE and use inner join with the Summarization tool with the same setting but without the filter. Lastly, calculate the percentage by dividing the number of units caused by road defect by the total of NUM_UNITS in the category, choose only rows more than 5 percent, and sort by Percentage in descending. The result of the analysis is shown in Figure 4.

Record	STREET_NAME	TOTAL_NUM_UNITS	ROAD_DEFECT	PRIM_CONTRIBUTORY_CAUSE	ROAD_DEFECT_NUM_UNITS	Percentage
1	RICHARDS DR	5	OTHER	NOT APPLICABLE	3	60
2	BURNHAM AVE	7	SHOULDER DEFECT	ROAD CONSTRUCTION/MAINTENANCE	2	28.57
3	86TH PL	7	OTHER	NOT APPLICABLE	2	28.57
4	54TH PL	8	RUT, HOLES	NOT APPLICABLE	2	25
5	LILL AVE	9	WORN SURFACE	VISION OBSCURED (SIGNS, TREE LIMBS, BUILDIN...	2	22.22
6	CONCORD PL	20	DEBRIS ON ROADWAY	WEATHER	4	20
7	TILDEN ST	12	OTHER	ROAD CONSTRUCTION/MAINTENANCE	2	16.67
8	101ST ST	15	RUT, HOLES	ROAD ENGINEERING/SURFACE/MARKING DEFEC...	2	13.33
9	NORTH WATER ST	10	RUT, HOLES	NOT APPLICABLE	1	10
10	SHORE DR	21	RUT, HOLES	NOT APPLICABLE	2	9.52
11	HAUSSEN CT	21	RUT, HOLES	ROAD ENGINEERING/SURFACE/MARKING DEFEC...	2	9.52
12	LA CROSSE AVE	11	RUT, HOLES	VISION OBSCURED (SIGNS, TREE LIMBS, BUILDIN...	1	9.09
13	NEWLAND AVE	11	OTHER	ROAD CONSTRUCTION/MAINTENANCE	1	9.09
14	OKETO AVE	12	RUT, HOLES	NOT APPLICABLE	1	8.33
15	ORIOLE AVE	28	OTHER	ROAD CONSTRUCTION/MAINTENANCE	2	7.14
16	LAWRENCE DR	16	RUT, HOLES	ROAD ENGINEERING/SURFACE/MARKING DEFEC...	1	6.25
17	BYRON ST	54	RUT, HOLES	WEATHER	3	5.56
18	MIDWAY PLAISANCE	18	RUT, HOLES	ROAD ENGINEERING/SURFACE/MARKING DEFEC...	1	5.56
19	FLETCHER ST	39	SHOULDER DEFECT	VISION OBSCURED (SIGNS, TREE LIMBS, BUILDIN...	2	5.13

Figure 4: Result of Workflow of street names that have road defect contribution to the number of units involved in the crash 5 percent or more in 2023

The result contains 19 records with 6 fields namely STREET_NAME, TOTAL_NUM_UNITS, ROAD_DEFECT, PRIM_CONTRIBUTORY_CAUSE, ROAD_DEFECT_NUM_UNITS (Number of units involved in each Road Defect and Primary Contributory Cause), and Percentage. The Primary Contributory Cause is used to ensure that the crash might happen because of the road defect. In this case, VISION OBSCURED is counted too because it might be the signs from ROAD_DEFECT.

Analysis with Alteryx

1. List the Street name and calculate the percentage of car crashes that are probably caused by Road Defects in 2023 that have a percentage of 5 percent or more, and filter by Primary contributory cause in this list: ("UNABLE TO DETERMINE", "NOT APPLICABLE", "ROAD CONSTRUCTION/MAINTENANCE", "VISION OBSCURED (SIGNS, TREE LIMBS, BUILDINGS, ETC.)", "WEATHER", "ROAD ENGINEERING/SURFACE/MARKING DEFECTS"). (Grouped by Street name, Road defect, and Primary contributory cause)
2. [Changed to Power Query] Count the number of crashes by record ID grouped by Street name and year and add a total column.
3. [Changed to Power Query] Create a summary of Total crashes in each year and create a total column for counting the total number of crashes of each street name.
4. [Changed to Power Query] Time taken for police to be notified about the crashes of each beat. Show the result in seconds and in hh:mm:ss format.

5. [Changed to Power Query] Total fatal or incapacitating injuries based on Primary and Secondary Contributory cause.

Modification

1. Convert CRASH_DAY_OF_WEEK from 1 - 8 to Sunday - Saturday
2. Removed the records that contain null values of STREET_DIRECTION, STREET_NAME, BEAT_OF_OCCURENCE, LATITUDE, LONGITUDE, and LOCATION
3. Removed records with Latitude and Longitude equal to 0.
4. Using imputation and replacing null values with median in INJURIES_TOTAL, INJURIES_FATAL, INJURIES_INCAPACITATING, INJURIES_NON_INCAPACITATING, INJURIES_REPORTED_NOT_EVIDENT, INJURIES_NO_INDICATION, AND INJURIES_UNKNOWN
5. The fields that contain null values of more than 100,000 records are removed from the table.
6. Replace null values of REPORT_TYPE and RD_NO with ‘UNKNOWN’
7. Added CRASH_DAY_OF_WEEK_TEXT (Sunday - Saturday) from date, the one in the dataset is in Integer format (1-7).

Power BI Visualization

Draft Dashboard

To streamline the process and minimize potential issues when working with Power BI, we have chosen to plan the dashboard's layout. This involves determining which components should be included and assigning importance to each (chart size and placing location), ensuring that the final dashboard is user-friendly and easy to comprehend. However, since in user 2, we are getting more familiar with the data and tools, we decided not to do the draft dashboard, and cite some components from user 1 instead to reduce time.

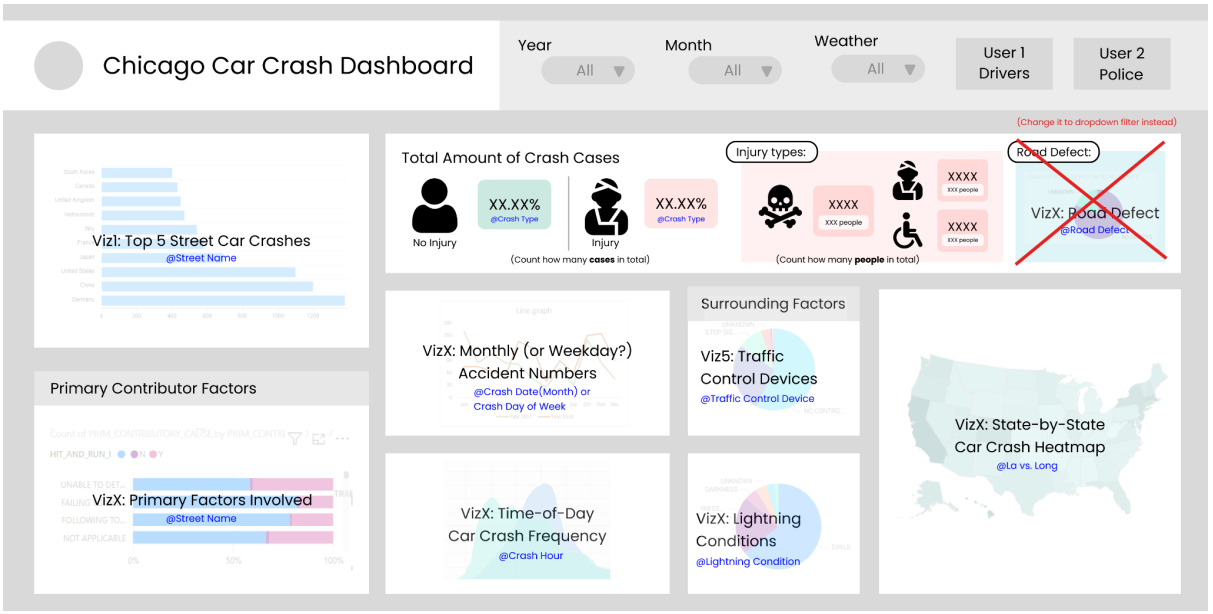
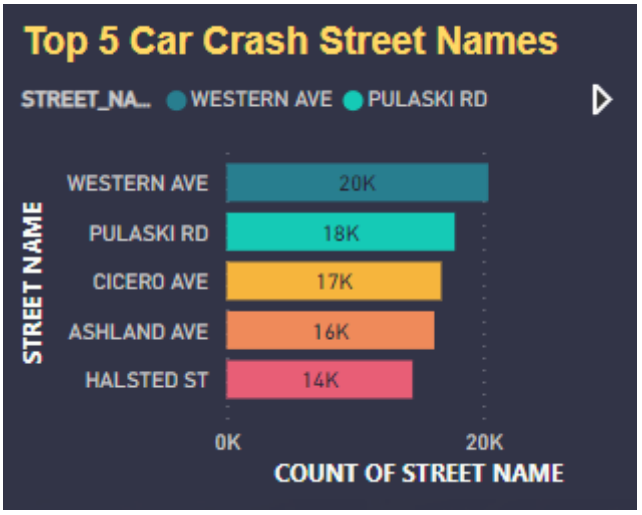


Figure 5: User 1's Draft Dashboard

Visualization for Commuters and Drivers

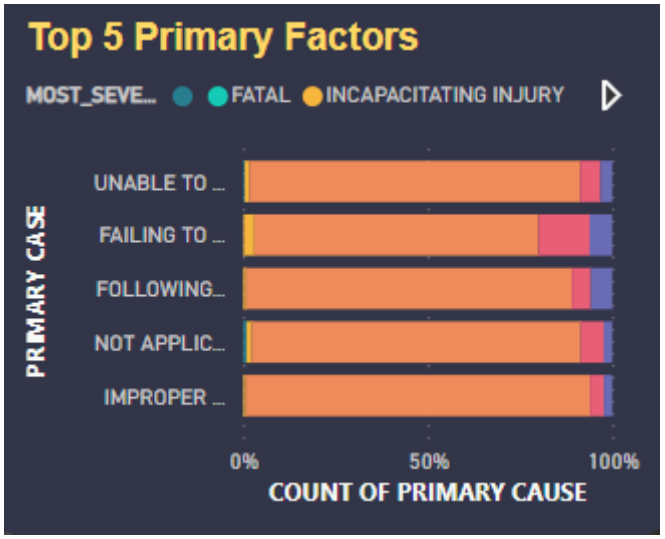
Name: Top 5 Car Crash Street Names



Visualization 1: Top 5 Car Crash Street Names

Description: This visualization is a Clustered Bar Chart showing the top 5 streets with the most car crashes in Chicago, USA. The chart illustrates the distribution of the total number of car accidents by different street names in Chicago indicated by the colors. On top of that, the displayed data is also sorted and ranked with the highest number of car accidents for the top 5 street names. This visual can provide the users with helpful information used to make a decision about route avoidance.

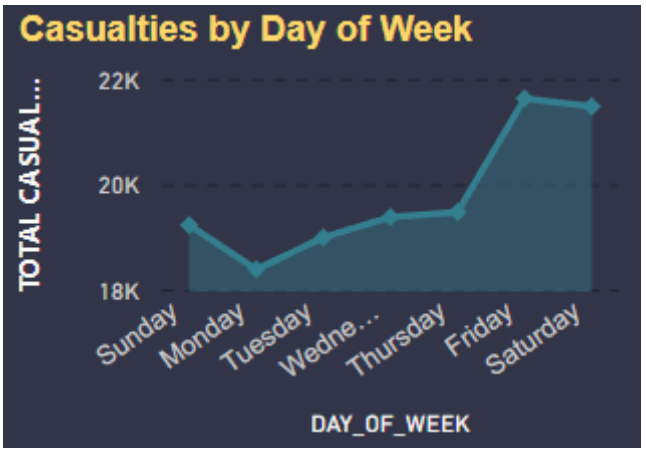
Name: Top 5 Primary Factors



Visualization 2: Top 5 Primary Factors

Description: This visualization displays a 100% Stacked Bar Chart that visually presents the primary causes of car crash accidents in Chicago, USA, and the corresponding percentages of most severe injury found (the maximum number among injury types) incidents associated with each cause. Users can utilize this data to practice caution and understand the consequences of injuries associated with these key factors.

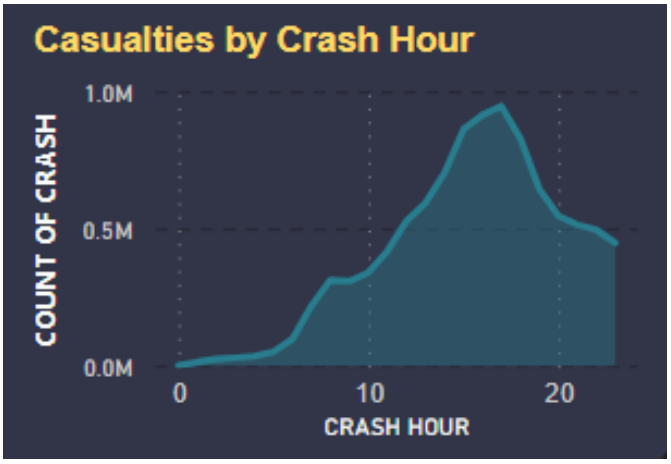
Name: Casualties by Day of Week



Visualization 3: Casualties by Day of Week

Description: This visualization presents a Stacked Area Chart, which illustrates casualties by the day of the week, indicating the frequency of car crash accidents occurring on each day. This allows the users to get an overall view and encourages them to be cautious on days with a higher accident rate.

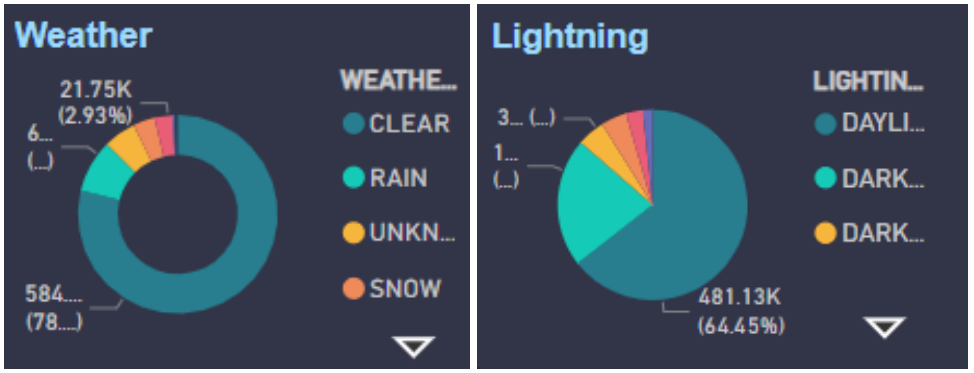
Name: Casualties by Crash Hour



Visualization 4: Casualties by Crash Hour

Description: This visualization shows a Stacked Area Chart that depicts casualties based on the time of day, showing the frequency of car crash accidents at each hour. Provides the users an overall perspective on the hours of the day when car accidents are most likely to occur and enables them to be cautious during those times that have a higher accident rate.

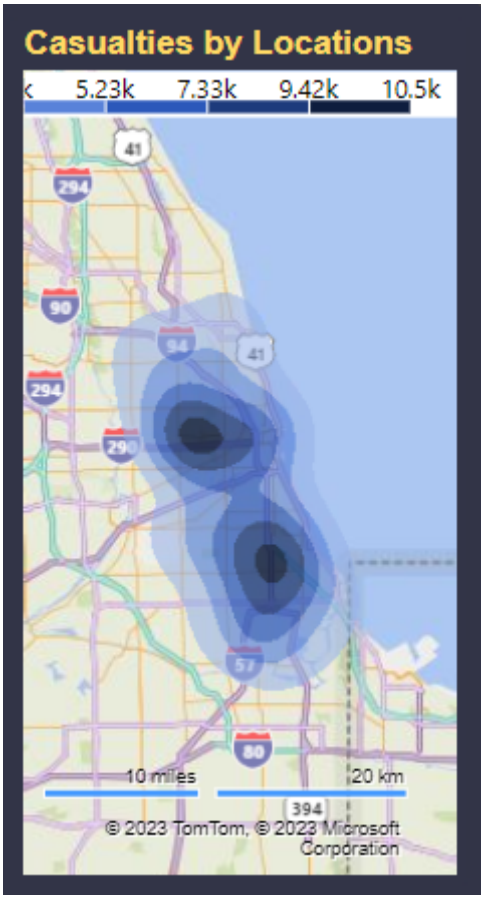
Name: Environmental Related



Visualization 5: Environmental Related

Description: This visualization employs a Donut and Pie Chart respectively to display the percentage of environmental factors contributing to car crash incidents, encompassing both weather and lighting conditions. Users can utilize this visual representation to assess the likelihood, in terms of percentages, of each specific weather and lighting condition leading to an accident. The weather uses a Donut chart since it allows the user to see the individual values of weather types.

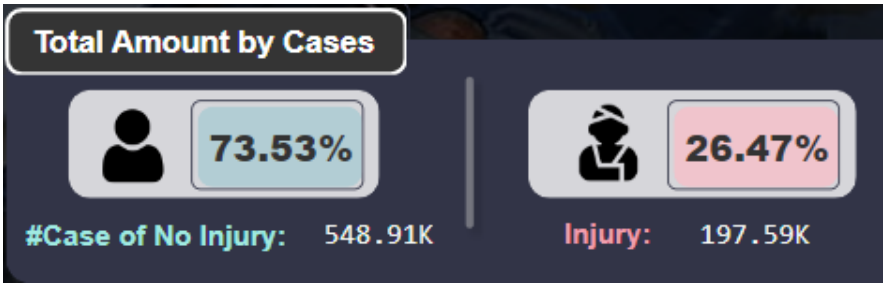
Name: Casualties by Locations



Visualization 6: Casualties by Locations

Description: This visualization utilizes Google Maps for Power BI to present a heat map that pinpoints the count of car crash casualties categorized by various locations within Chicago, USA. This visual aids users in obtaining a comprehensive overview of Chicago, showing areas or streets with the number of car accidents.

Name: Total Amount by Cases



Visualization 7: Total Amount by Cases

Description: This visualization employs cards to show the total count of car accident cases in Chicago, USA, and classification into two groups, including those with no injury and those with injuries. This way, users will visually see both the total amount and percentage of accident cases involving drivers who either sustained an injury or did not sustain an injury after being in an accident.

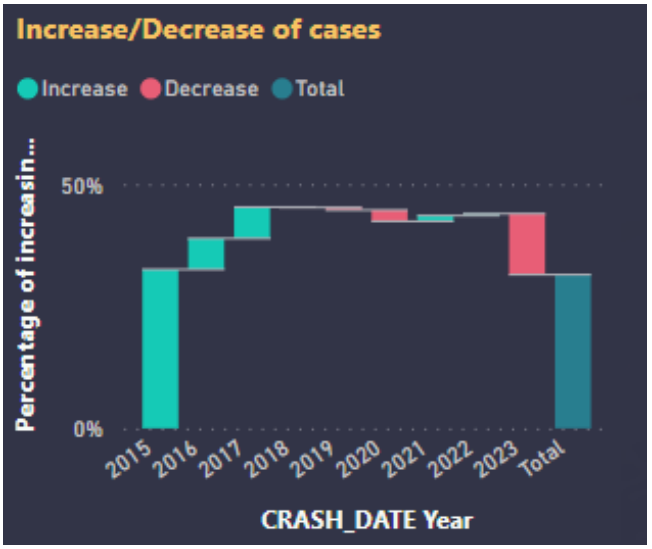
Name: Total Number of Injuries

Visualization 8: Total Number of Injuries

Description: This visualization shows the number of injuries from accidents, including fatal, incapacitating, and non-incapacitating. It is useful to prevent future accidents, improve safety precautions in various areas, such as transportation, workplaces, or public spaces, and reduce the incidence of accidents.

Visualization for Law Enforcement Agencies

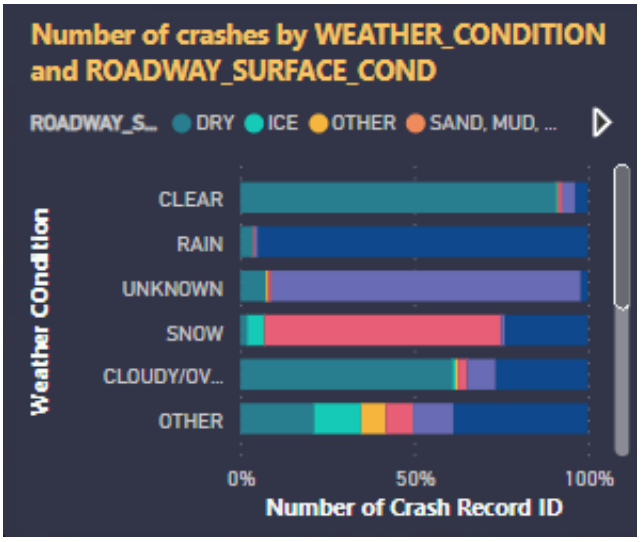
Name: Increase/Decrease of cases



Visualization 9: Increase/Decrease of cases

Description: This visualization employs a Waterfall Chart to illustrate the percentage of accident cases that have increased or decreased each year. Users also have the option to drill down for more detailed insights by quarters and months. This visual benefits users in monitoring the trends of increase and decrease of car crashes in Chicago, USA, facilitating continued improvement of the agencies' performance.

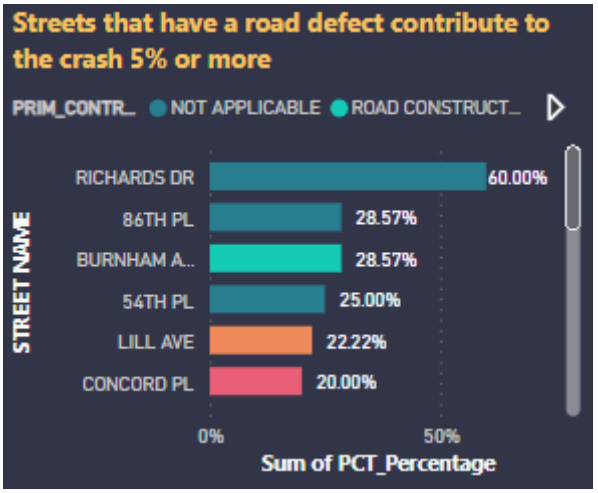
Name: Number of Crashes by Weather Condition and Road Surface Condition



Visualization 10: Number of Crashes by Weather Condition and Road Surface Condition

Description: This visualization shows the number of crashes that are separated by weather conditions and road surface conditions, which use a 100% Stacked Bar Chart. This visualization provides information to help the related agencies see which weather or road surface leads to high accident rates.

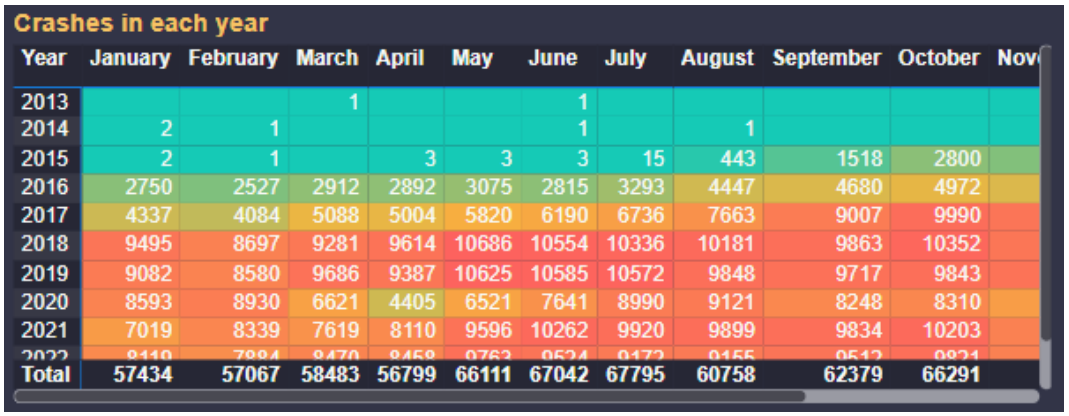
Name: Streets that have a Road Defect Contribute to the Crash 5% or More



Visualization 11: Streets that have a Road Defect Contribute to the Crash 5% or More

Description: This visualization uses a Stacked Bar Chart to illustrate the streets in Chicago, USA, where road defects have contributed to 5% or more of the crashes. Users can refer to this visual to gain insights into which streets are most affected by road defects and assess additional relevant information.

Name: Crashes in Each Year

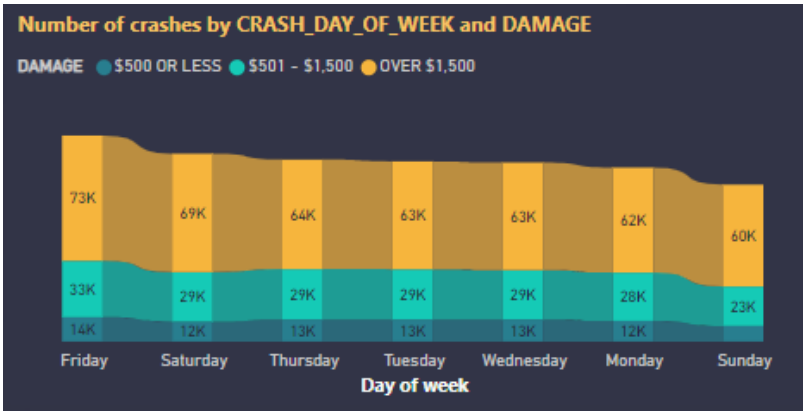


Year	January	February	March	April	May	June	July	August	September	October	Nov
2013			1			1					
2014	2	1				1		1			
2015	2	1		3	3	3	15	443	1518	2800	
2016	2750	2527	2912	2892	3075	2815	3293	4447	4680	4972	
2017	4337	4084	5088	5004	5820	6190	6736	7663	9007	9990	
2018	9495	8697	9281	9614	10686	10554	10336	10181	9863	10352	
2019	9082	8580	9686	9387	10625	10585	10572	9848	9717	9843	
2020	8593	8930	6621	4405	6521	7641	8990	9121	8248	8310	
2021	7019	8339	7619	8110	9596	10262	9920	9899	9834	10203	
2022	8119	7884	8470	8458	9763	9524	9172	9155	9512	9821	
Total	57434	57067	58483	56799	66111	67042	67795	60758	62379	66291	

Visualization 12: Crashes in Each Year

Description: This visualization utilizes a Matrix to provide a visual representation of the count of crashes in each year, segmented by months. This format allows users to observe the historical data and patterns of car crash accidents in Chicago, USA. Red cells indicate a high frequency of crash occurrences, while green cells represent a lower rate of crash incidents.

Name: Number of Crashes by Crash Day of Week and Damage



Visualization 13: Number of Crashes by Crash Day of Week and Damage

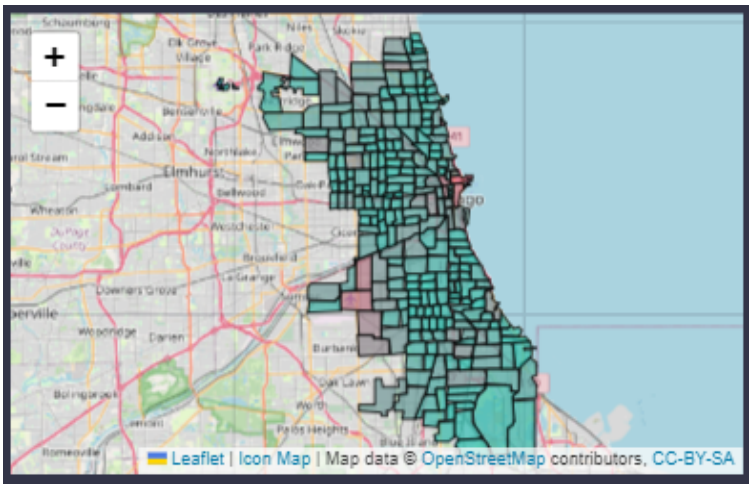
Description: This visualization uses a Ribbon Chart to illustrate the count of crashes by the day of the week and their associated damage values in Chicago, USA. Users can utilize this visual representation to observe the distribution of damage values for car crashes each day of the week and make further assessments or comparisons as needed.

Name: Total Number of Crashes, Injuries, and Beats

Visualization 14: Total Number of Crashes, Injuries, and Beats

Description: This visualization employs a Card to provide a summary of the total count of Crashes, Injuries, and Beats. Users can use this visual to understand an overall perspective of the car crash situation in Chicago, USA.

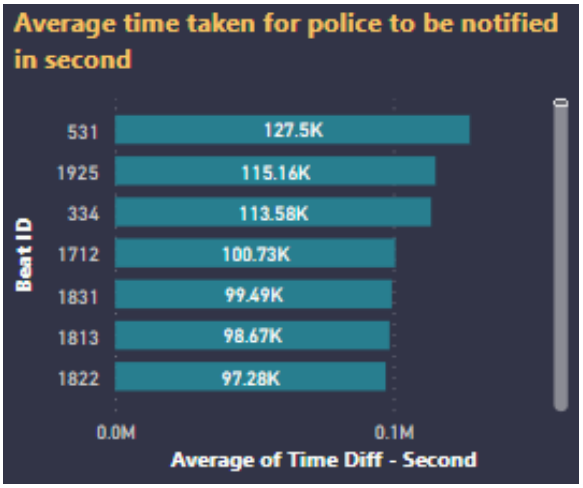
Name: Map Represent Beat Num Information



Visualization 15: Map Represent Beat Num Information

Description: This visualization utilizes an iconMapV3 to create a map that displays Beat information, incorporating details including Beat boundaries [4], Beat Number, the count of crashes, District, and Sector. It distinguishes different numbers of crashes on the map with various colors. Users can take advantage of this visual to understand the associations between Beat IDs, sectors, and districts, and observe the crash count for each district.

Name: Average Time Taken for Police to be Notified in Second

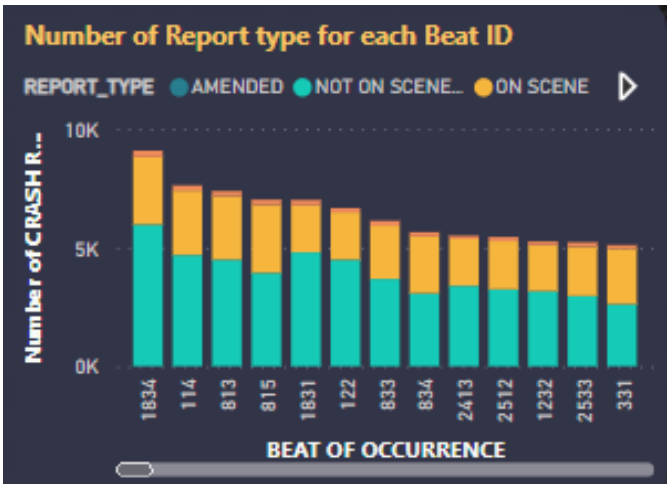


Visualization 16: Average Time Taken for Police to be Notified in Second

Description: This visualization employs a Stacked Bar Chart to illustrate the average time it takes for the police to notice a crash for each Beat, measured in seconds. This data allows

each Beat to evaluate their response speed and use this information for training and enhancing their operational efficiency in the future.

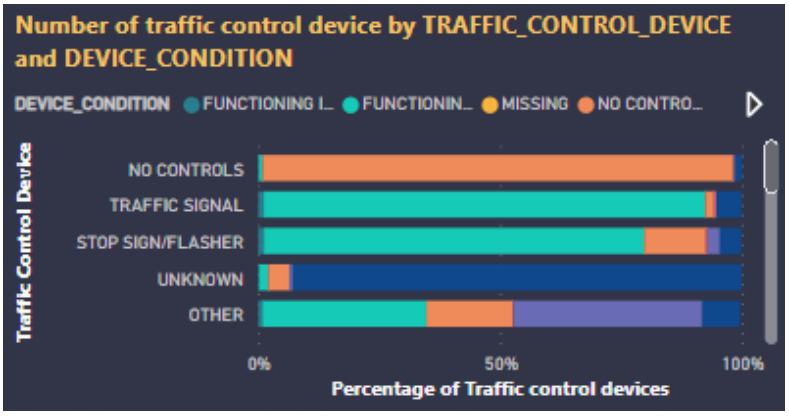
Name: Number of Report Types for Each Beat ID



Visualization 17: Number of Report Types for Each Beat ID

Description: This visualization employs a Stacked Column Chart to illustrate the count of report types for each Beat ID in the Chicago Police Department, USA. Users can utilize this insight data to collaborate with relevant departments and plan improvements and training for future enhancements.

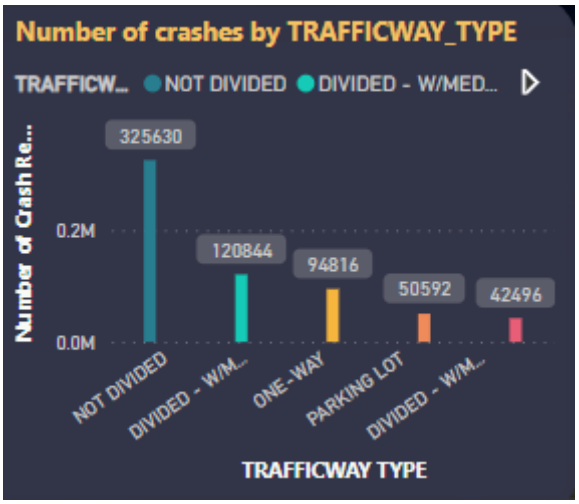
Name: Number of Traffic Control Device by Traffic Control Device and Device Control



Visualization 18: Number of Traffic Control Devices by Traffic Control Device and Device Control

Description: This visualization utilizes a 100% Stacked Bar Chart to display the percentage of traffic control devices categorized by the type of traffic control device and its device control status. Users can refer to this visual to gain a comprehensive overview of the current status of each traffic control device and determine whether each traffic control device is in which beat number area of device control.

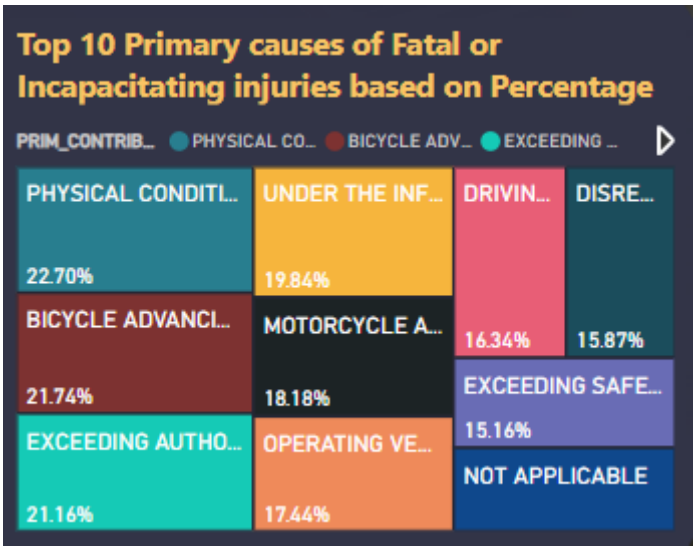
Name: Number of Crashes by Trafficway Type



Visualization 19: Number of Crashes by Trafficway Type

Description: This visualization employs a Clustered Column Chart to represent the count of crashes separated by trafficway type in Chicago, USA. Users can use this visual to identify which types of trafficways have the highest number of car crash accidents, allowing the related department to be more cautious in those specific areas.

Name: Top 5 Primary Causes of Fatal or Incapacitating Injuries



Visualization 20: Primary Causes of Fatal or Incapacitating Injuries Based on Percentage

Description: This visualization employs a Treemap of the top 10 primary causes of fatal or incapacitating injuries in car accidents. The percentage is calculated from the Sum of fatal and incapacitating injuries divided by the total injuries. Law enforcement agencies can gain insights into the reasons behind accidents and use this information to devise strategies for preventing these specific causes.

Dashboard

Figure 6: Dashboard 1 (User 1)

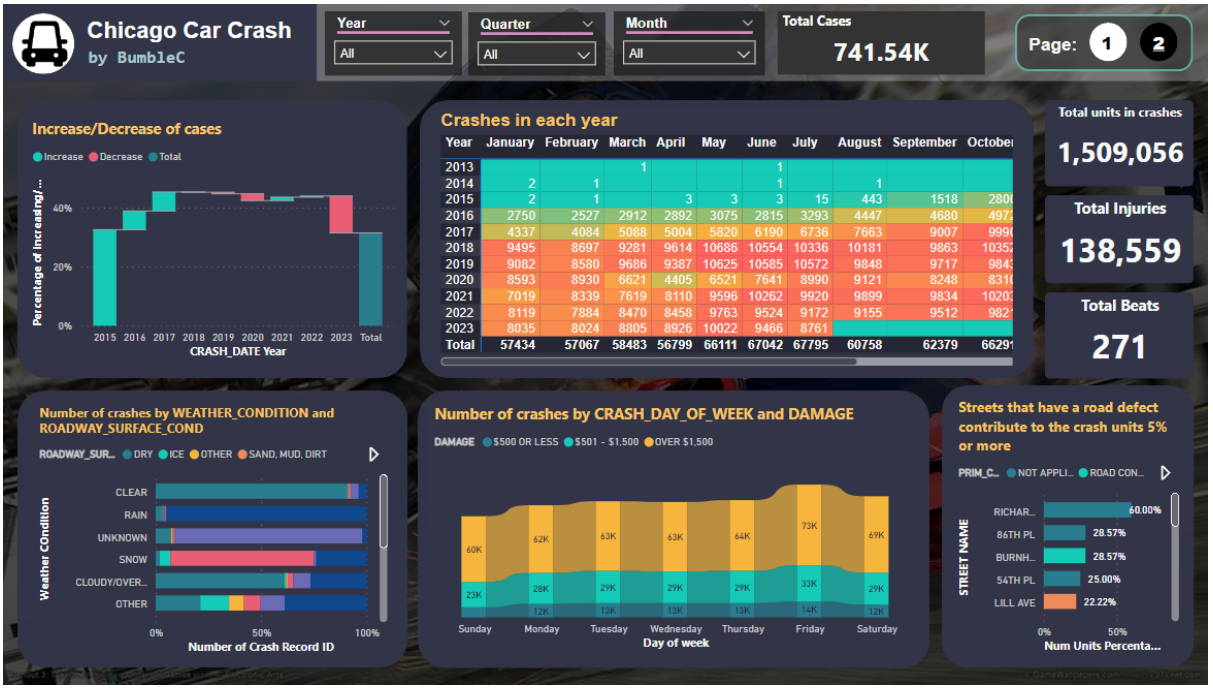


Figure 7: Dashboard 2 (User 2)

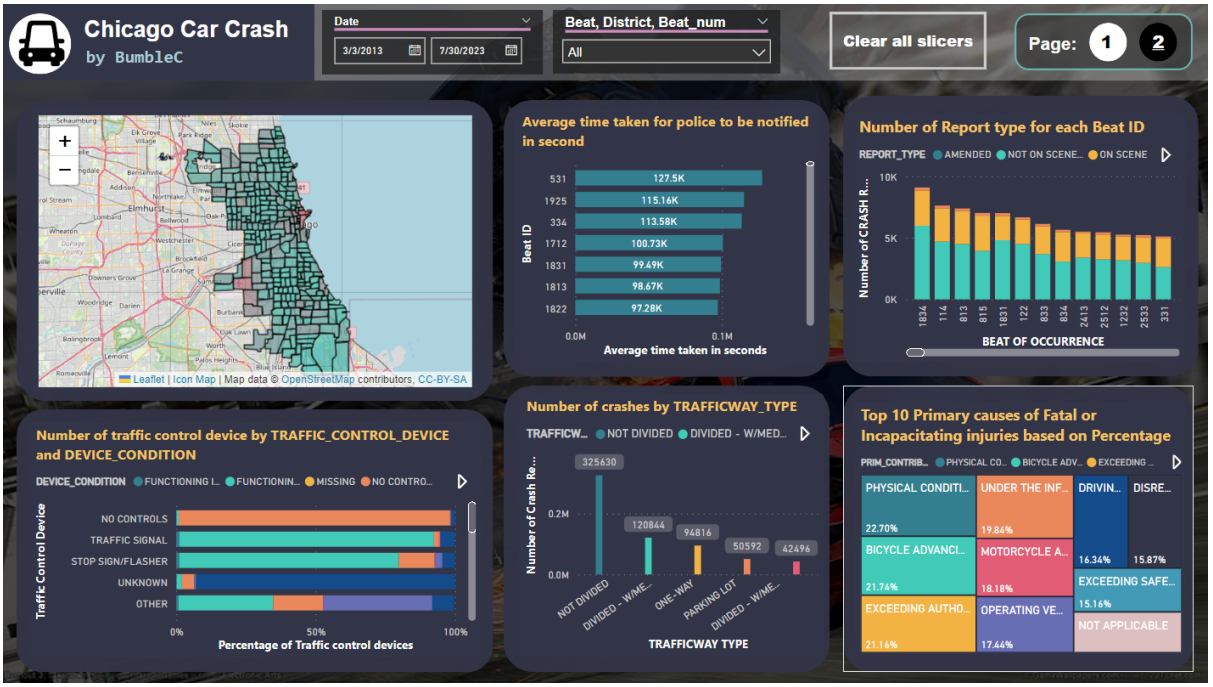


Figure 8: Dashboard 2 (User 2)

Discussion and Conclusion

Discussion

Alteryx and Power Query in Power BI can do almost the same tasks. However, Alteryx is easier and simpler with complex analysis than Power Query. In Power BI, several limitations make some analyses unusable in visualization. First, the map has a limit of 30,000 data points which is not enough for the Chicago Car Crash dataset. In addition, plotting 30,000 data points also caused the dashboard to be very laggy and occasionally crash the program. The solution are using a heatmap to show the area instead, but this visualization needs an interaction on the first time loading. The other solution is IconMapV3 which is used in Dashboard 2. It can show the Beat boundaries using Well-Known Text (WKT) from the boundary dataset. However, it cannot dynamically show the data points instead of the boundary when zooming in, which prevents pointing out the traffic device locations. Consequently, another map is needed and it will trade with a performance from the first problem.

Conclusion

There are many business domains involved with car crashes such as insurance, car manufacturers, auto repairs, etc. However, this project focuses on Public Safety and Legal Services. The target users are commuters or drivers, Law Enforcement Agencies (Police Departments, Department Of Transportation (DOT)), and Traffic Safety Apps and Navigation Services. The usage of drivers and Traffic Safety Apps and Navigation Services have a similar goal, preventing routes with high chances of crashes. On the other hand, the Police Department can use the data to allocate appropriate resources to prevent crashes, find the cause of the crashes to find a solution (proposing the adjustment to the authoritative parties), or improve the operations. The main dataset in this project is the **Chicago Car Crash Dataset.** Another dataset is **Boundaries - Police Beats (current)** or Chicago Police Beats Boundary, which can be used with the main dataset perfectly since they are in the same domain and organization. The main dataset needs to be processed in Alteryx while the additional dataset is cleaned. In Power BI, The first visualization has 12 visualizations with 4 sliders for commuters or drivers to gain general insights into car crashes and areas or streets to prevent. Moreover, there are the other 2 dashboards are for Police with different purposes. The second visualization consists of 9 visualizations with 3 sliders for displaying data overview and facts while The third visualization contains 6 visualizations with 2 sliders for analyzing Beats, resources, and operations.

References

[1] C. of Chicago, “Traffic crashes - crashes: City of Chicago: Data Portal,” Chicago Data Portal, https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if (accessed Sep. 23, 2023).

[2] SavorSauce, “Chicago Car Crash Dataset,” Kaggle, https://www.kaggle.com/datasets/nathaniellybrand/chicago-car-crash-dataset (accessed Oct. 10, 2023).

[3] “Traffic crashes - crashes,” Catalog, https://catalog.data.gov/dataset/traffic-crashes-crashes (accessed Oct. 12, 2023).

[4] Boundaries - Police Beats (current), https://catalog.data.gov/dataset/boundaries-police-beats-current (accessed Oct. 15, 2023).