Iowa Vehicle Crash Data Analysis by team DataVerse

INDEX:

1. Deliverable 1
    - 1.1. Problem Statement
        - 1.1.1. Challenges to Address
        - 1.1.2. Who the Consumer Is
        - 1.1.3. Potential Benefit
        - 1.1.4. Business Overview – New Business
        - 1.1.5. Project Description
    - 1.2. Business Impact: Cost Reduction, Revenue Generation, and New Business Opportunities
        - 1.2.1. Cost Reduction
        - 1.2.2. Revenue Generation
        - 1.2.3. New Business Line
    - 1.3. Methodology
        - 1.3.1. Vehicle Crash Data Processing and Analysis Workflow
        - 1.3.2. Azure Architecture Diagram
        - 1.3.3. Tools
        - 1.3.4. Dataset Description
    - 1.4. Summary

2. Deliverable 2
    - 2.1. Data Acquisition
    - 2.2. Data Sources
        - 2.2.1. Vehicle Crash in Iowa
        - 2.2.2. People Injured in Iowa Crashes
        - 2.2.3. Drivers and Vehicles involved in Iowa crash
    - 2.3. Data Volume
    - 2.4. Focus Attributes
    - 2.5. Filtering
    - 2.6. Data Cleaning

3. Deliverable 3
    - 3.1. Data Process
        - 3.1.1. Software and System Used
        - 3.1.2. Steps Taken
    - 3.2. Assumptions
    - 3.3. Visualization
        - 3.3.1. Bar Chart of Crash Severity vs Environmental Conditions
        - 3.3.2. Heat Map of Crashes by Time and Day of the Week
        - 3.3.3. Tree map of Vehicle Make vs Number of Fatalities
        - 3.3.4. Clustered Column Chart of Drug or Alcohol Test Results vs Number of Fatalities
    - 3.4. Recommendations

1. Deliverable 1

1.1 Problem Statement:

Iowa faces a significant public safety issue due to the high number of vehicle crashes, leading to fatalities, injuries, and substantial economic losses. The primary challenge is to reduce both the frequency and severity of these crashes. This can be achieved by analysing crash data to uncover key patterns and risk factors that contribute to these incidents and identifying high-risk locations, understanding contributing factors, and implementing targeted interventions. This project aims to utilize Big Data analytics and machine learning tools to provide insights and help make informed strategic decisions, ultimately making Iowa's roads safer.

1.1.1 Challenges to Address:

- **High-Risk Locations:** Identifying specific crash hotspots, such as dangerous intersections, rural roads, or highways with high traffic volumes. The goal is to analyse why these areas are more prone to crashes and address factors like weather conditions, road surfaces, traffic patterns, and driver behaviours like speeding or distracted driving or lack of safety barriers.
- **Severity:** Investigating what makes certain crashes more severe than others. This includes exploring factors such as impaired driving (alcohol or drugs), distracted driving, weather conditions (fog, ice), and road surface quality. Understanding these elements will help in prioritizing interventions that could save lives, such as infrastructure improvements, improved traffic signs or traffic control measures.
- **Temporal Patterns:** Crashes often occur more frequently at specific times, such as during rush hours, late at night, or under adverse weather conditions. By analysing these time-based patterns, we can recommend adjusting traffic light timings during high-risk periods or increasing patrols during peak accident hours.

1.1.2 Who the Customer Is:

- **Iowa Department of Transportation (DOT):** The Iowa DOT is a primary customer, relying on these insights to make informed decisions about infrastructure improvements and safety interventions like where to add traffic signals, improve road markings, or implement other safety measures
- **Local Law Enforcement Agencies:** City and county officials will benefit from data-driven insights for targeted enforcement, such as focusing on DUI checks or speed enforcement in high-risk areas.
- **Public Health and Safety Organizations:** These organizations can tailor public safety campaigns based on the insights enabling them to implement preventative measures such as promoting seat belt use or warning against texting while driving during high-risk times.
- **Insurance Companies:** The analysis could help insurance companies refine risk assessments for drivers, leading to more personalized insurance plans and potentially reducing premiums for safer driving behaviours or regions with lower crash probabilities.
- **Iowa Drivers and Pedestrians:** The ultimate beneficiaries of this project will be the residents and travellers within Iowa. By proactively identifying and addressing hazardous road conditions and alcohol-related factors, the project aims to reduce the frequency of crashes, potentially saving lives and minimizing injuries.

1.1.3 Potential Benefits:

- **Enhanced Public Safety:** This project aims to reduce the frequency and severity of crashes, ultimately saving lives and decrease the number of injuries on Iowa's roads.
- **Economic Savings:** By preventing crashes, the state can save on costs related to emergency responses, medical care, legal expenses, and property damage etc and this project helps people to lower insurance costs, fewer emergency response expenditures, and a general decrease in the economic burden associated with vehicle accident.
- **Efficient Use of Resources:** Data-driven insights allow for smarter allocation of resources, such as targeting road improvements in the most dangerous areas or deploying law enforcement more effectively.

1.1.4 Business Overview – New Business:

Safe Routes Analytics, a newly established consulting firm, focuses on improving road safety and optimizing transportation systems for municipal governments, transportation departments, and private infrastructure companies. By leveraging data-driven insights, the company aims to reduce vehicle accidents and enhance transportation planning. Working with the Iowa Department of Transportation (Iowa DOT) open data, one of its first initiatives involves analysing the "Vehicle Crashes in Iowa" dataset, which includes crash location, time, cause, and conditions. Safe Routes plans to develop predictive models and actionable recommendations to reduce crash frequency and severity across Iowa's roadways.

In addition to crash data, the analysis will incorporate an alcohol usage dataset, allowing the team to examine the role that alcohol consumption plays in traffic incidents. By correlating alcohol usage trends with crash data, Safe Routes aims to uncover insights that will help local authorities and businesses, such as bars and restaurants, better understand the connection between alcohol consumption and vehicle crashes. This could inform targeted interventions like stricter enforcement of DUI laws in high-risk areas or the implementation of alcohol awareness programs.

1.1.5 Project Description:

The project will focus on conducting a comprehensive analysis of the dataset, looking for correlations between crash severity and factors like road conditions, weather patterns, alcohol involvement, and driver behaviour. The addition of the alcohol usage dataset will allow for deeper analysis into whether certain regions, times, or events (e.g., holidays or sporting events) are associated with increased alcohol-related crashes. Machine learning models will be employed to predict crash likelihood under varying conditions, providing a tool for decision-makers to implement preventive measures.

This project has the potential to create a ripple effect of benefits, from improving safety on the roads to influencing how transportation infrastructure and alcohol-related policies are developed across the state. Through data analysis and predictive modelling, Safe Routes Analytics (new business) is poised to contribute to Iowa's future as a safer place to live, work, and travel.

1.2 Business Impact: Cost Reduction, Revenue Generation, and New Business Opportunities

1.2.1 Cost Reduction:

- **Accident-Related Costs**: By predicting high-risk crash zones and periods, public and private sectors can proactively prevent accidents, resulting in fewer insurance claims, reduced legal liabilities, and lower healthcare costs. Additionally, focusing infrastructure maintenance and improvements on crash-prone areas helps avoid unnecessary spending.

- **Emergency Services Optimization**: Predictive insights will enable more efficient deployment of emergency resources. Knowing high-risk times and locations allows ambulances, police, and fire services to strategically position themselves, cutting down response times and lowering overall operational costs.
- **Targeted Traffic Management**: Government bodies and city planners can use predictions to install traffic control systems (e.g., speed cameras, traffic lights, signs) in areas of the highest need, ensuring a cost-effective approach to traffic management.

1.2.2 Revenue Generation:

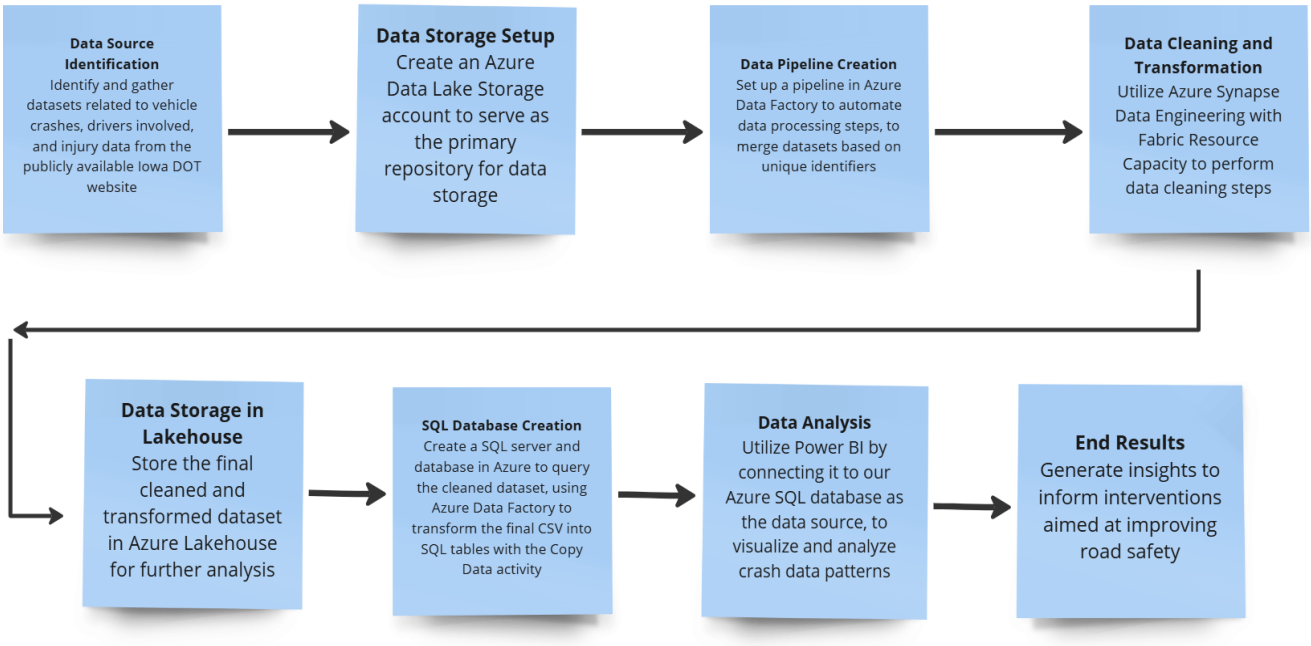
- **Customized Insurance Premiums:** Insurance companies can develop more personalized premium plans by leveraging predictive models. They can identify high-risk drivers and charge appropriate premiums while rewarding safer drivers with lower rates. This data-driven approach can lead to an increase in revenue as premiums more accurately reflect risk profiles.
- **Fines and Citations:** Authorities can increase traffic monitoring and enforce fines in high-risk areas using predictive insights. This allows for more efficient placement of speed cameras and patrols, boosting revenue through traffic citations while promoting safer driving behaviours.
- **Dynamic Pricing for Roads:** Predictive analytics enables transportation agencies to implement dynamic pricing systems. For example, higher fees can be charged during peak risk times, while encouraging road use during safer periods. This can generate additional revenue while balancing traffic flow and reducing accidents.
- **Partnerships with Automotive and Safety Tech Companies:** Automotive companies could utilize predictive crash data to enhance vehicle safety systems (e.g., Advanced Driver Assistance Systems), improving crash avoidance. Partnerships between government agencies and road safety companies for smart road technologies (such as intelligent traffic lights or road sensors) could also open new revenue streams.

1.2.3 New Business Lines

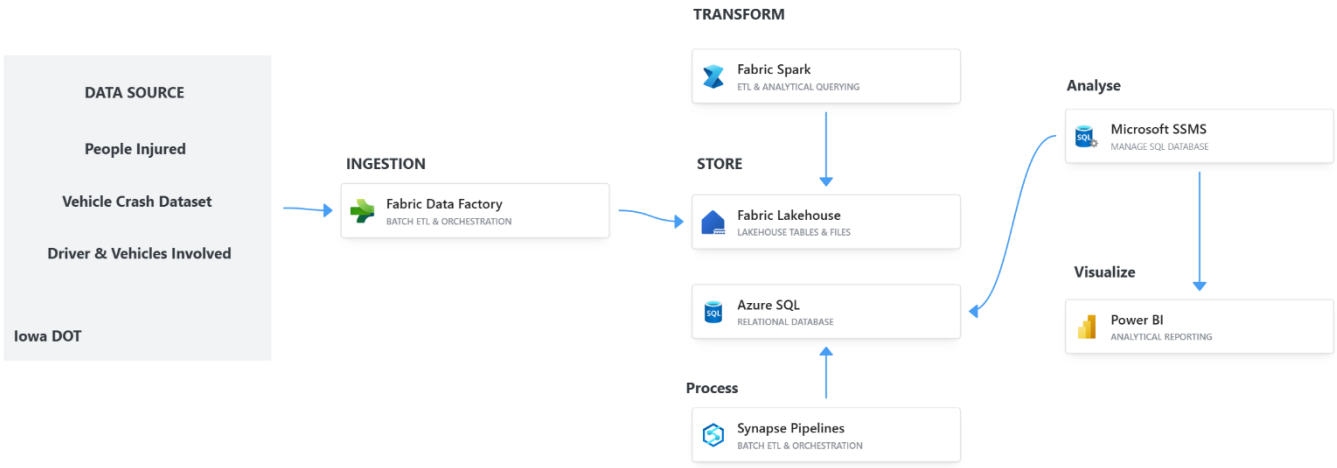
- **Road Safety Analytics Services:** Using predictive insights, a new business could emerge around selling crash data analytics as a service to transportation agencies, municipalities, and private companies. This could involve real-time crash forecasting, which can be offered as a subscription or consulting service to help with urban planning, traffic management, and safety initiatives.
- **Smart City Infrastructure:** Smart city solutions, enabled by predictive crash data, could be designed to improve road safety. This includes real-time adaptive traffic lights that adjust based on predicted crash risks. Companies involved in these technologies can offer these solutions to governments and businesses as part of a broader innovative city framework, creating a new market for safety-focused traffic management solutions.
- **Vehicle Telematics and Driver Monitoring**: Another potential business line involves real-time driver monitoring services based on predictive models. Vehicle fleets, insurance companies, or even individual drivers could adopt these solutions to track driving behaviour and reduce risk, leading to lower insurance premiums and safer roads.

1.3 Methodology:

1.3.1 Vehicle Crash Data Processing and Analysis Workflow



1.3.2 Azure Architecture Diagram:



1.3.3 Tools:

Several tools will be used to handle the preparation, cleaning, and analysis tasks, **Software and Systems Used:**

- **Azure Data Lake Storage:** Primary repository for storing all datasets. This enables scalable, secure storage for large volumes of data, making it easily accessible for data processing and analysis. Datasets are uploaded to Azure Data Lake Storage and accessed via Azure Data Factory and other tools for further processing.
- **Microsoft Fabric:** Used for setting up a workspace for centralized data processing, management, and storage, ensuring efficient data handling at scale.

- **Azure SQL:** Employed for structured data storage, querying, and management, enabling efficient handling of relational data within the project.
- **Power BI:** Utilized for visualizing crash data to identify patterns related to road conditions, driver behaviour, and crash severity, providing actionable insights through interactive dashboards.
- **Azure Data Factory:** Used to build and manage data pipelines for merging, transforming, and transferring data between storage solutions. This enables streamlined data integration from multiple datasets using unique identifiers.
- **Lakehouse:** Acts as a staging area for data processing, allowing for cleaning, transforming, and storing merged datasets, making them ready for downstream analysis.
- **Azure Lakehouse Notebook:** Used for data wrangling, attribute transformation, and other processing tasks, supporting data cleaning and transformation in a collaborative, reproducible environment.
- **PySpark:** Employed for large-scale data processing and distributed computing. PySpark allows efficient data wrangling and transformation tasks, making it well-suited for handling large datasets.
- **Synapse Data Engineering:** Used for advanced data engineering tasks, including data integration, preparation, and transformation at scale, enhancing data flow and processing capabilities across the project.
- **Microsoft SQL Server Management Studio (SSMS):** SSMS is used to manage and query SQL Server databases within the project, allowing for efficient database administration, querying, and data manipulation.

1.3.4 Dataset Description:

The dataset used for this project is titled "Vehicle Crashes in Iowa", which provides comprehensive information about vehicle crashes across Iowa. The dataset includes the following key attributes:

- Crash Date and Time: Timestamp of when the crash occurred.
- Crash Location: Latitude and longitude coordinates, providing specific geographical locations of crashes.
- Road Conditions: Details about road conditions at the time of the crash, such as whether it was dry, wet, icy, etc.
- Weather Conditions: Weather data at the time of the crash, which can help correlate crash frequency with adverse weather.
- Vehicle Information: Number and types of vehicles involved in the crash (e.g., passenger vehicles, trucks, motorcycles).
- Crash Severity: Information about the severity of crashes, including injury types and fatality counts.
- Driver Details: Driver age, gender, and other demographic information.

The dataset is a public source provided by the **Iowa Department of Transportation**, making it reliable for in-depth crash analysis. It is available via the following link:
Vehicle Crashes in Iowa Dataset

1.4 Summary:

The report highlights Iowa's public safety issue due to frequent vehicle crashes, aiming to reduce crash frequency and severity through data analysis and predictive modelling. SafeRoutes Analytics will analyse crash data alongside factors like weather, road conditions, and alcohol involvement to identify high-risk locations and patterns. By using machine learning models, the project will provide actionable insights for decision-makers, improving road safety and optimizing resource allocation. The initiative promises

economic savings through reduced accidents and new business opportunities, such as personalized insurance plans and smart city infrastructure solutions.

2. Deliverable 2

2.1 Data Acquisition:

The dataset used for this analysis was obtained from the official Iowa Data Portal, available at https://data.iowa.gov/. The portal provides open access to a wide range of data curated by the State of Iowa for public use.

For this project, we specifically sourced all four links related to vehicle crash data from the Iowa Department of Transportation (DOT) portal, ensuring comprehensive coverage of relevant incidents. These datasets contain detailed information such as crash locations, vehicle types, contributing factors, and crash outcomes. The four links were retrieved directly from the Iowa DOT's dataset repository, ensuring accuracy and consistency in the data.

The datasets are publicly available and can be downloaded in multiple formats, including CSV and JSON making them suitable for various types of analysis. For this analysis, we chose the CSV format due to its compatibility with various data analysis tools and ease of manipulation. Prior to download, we reviewed the metadata and supporting documentation to confirm that these datasets met the specific requirements for our analysis.

2.2 Data Sources: (Description, Data Type, Attributes)

2.2.1 Vehicle Crash in Iowa:

https://data.iowa.gov/Crashes/Vehicle-Crashes-in-Iowa/tw78-ziwj/data

The Vehicle Crashes in Iowa dataset contains comprehensive information on vehicle crashes across Iowa, beginning from January 2009. It includes key details such as the date and time of each crash, the precise location (latitude and longitude), and contributing factors like road conditions, weather, and driver behaviour. The dataset also documents the severity of each crash, including fatalities, injuries, and property damage. Additionally, it provides information about the vehicles involved, number of occupants, and contributing causes, offering valuable insights for analysing traffic safety trends and accident prevention measures.

Attribute Group	Attributes	Data Types
Case Information	Iowa DOT Case Number, Law Enforcement Case Number, Date of Crash, Month of Crash, Day of Week, Time of Crash, Hour	Number, Date, Text, Time
Location Details	DOT District, City Name, County Name, Route with System, Location Description, Location	Text, Point
Crash Event Details	First Harmful Event, Location of First Harmful Event, Manner of Crash/Collision, Major Cause	Text
Driver Influence	Drug or Alcohol	Text
Environmental Factors	Environmental Conditions, Light Conditions, Surface Conditions, Weather Conditions	Text

Roadway Details	Roadway Contribution, Roadway Type, Roadway Surface, Work Zone	Text
Crash Severity	Crash Severity, Number of Fatalities, Number of Injuries, Number of Major Injuries, Number of Minor Injuries, Number of Possible Injuries, Number of Unknown Injuries	Text, Number
Damage and Impact	Amount of Property Damage, Number of Vehicles Involved	Number
Occupant Details	Total Number of Occupants, Travel Direction	Number, Text

2.2.2 People Injured in Iowa Crashes:

https://data.iowa.gov/Crashes/People-Injured-in-Iowa-Crashes/66tm-a8uq/about_data

This dataset provides deidentified information on individuals injured in vehicle crashes in Iowa. It includes data on various injury levels such as fatalities, suspected serious incapacitating injuries, suspected minor injuries, and complaints of pain or possible injury. The dataset contains details about the injured person's seating position, use of occupant protection, airbag deployment, ejection status, and whether they were trapped. Additionally, the date and location of each crash are included, offering insights into crash severity and contributing circumstances related to occupant safety.

Attribute Group	**Attributes**	**Data Types**
Case Information	Iowa DOT Case Number, Person Key, Year of Crash, Date of Crash	Number, Text, Date
Demographics	Gender, Age Category, Age	Text, Number
Injury Details	Injury Status, Major Cause	Text
Driver Influence	Drug or Alcohol Related?	Text
Occupant Details	Seating Position, Occupant Projection Used?, Occupant Protection	Text
Ejection Details	Ejection, Ejection Path	Text
Safety Features	Airbag Deployment, Occupant Trapped?	Text
Crash Location	Crash Location (Latitude and Longitude)	Point

2.2.3 Drivers and Vehicles involved in Iowa crash:

https://data.iowa.gov/Crashes/Drivers-and-Vehicles-Involved-in-Iowa-Crashes/vjnp-m8t4/about_data

This dataset provides detailed information on both drivers and vehicles involved in crashes within the Vehicle Crashes in Iowa dataset. Deidentified driver information includes their age, gender, license state, any charges filed, results of alcohol and drug tests (if applicable), whether their vision was obscured, and other contributing circumstances to the crash. Vehicle details encompass make, model, style, license plate state, most damaged area, and the extent of damage. Additionally, data on the number of occupants, crash sequence, crash severity, surface conditions, and the date of the crash are provided, enabling comprehensive analysis of factors leading to vehicle accidents.

Attribute Group	**Attributes**	**Data Types**
Case Information	Unit Key, Iowa DOT Case Number, Date of Crash	Integer, Date
Driver Details	Driver Age, Driver Gender, Driver License State, Driver Charged	Integer, String, Boolean
Test Results	Alcohol Test Results, Drug Test, Drug Test Results	String, Boolean

Driver Condition	Vision Obscured, Driver Contributing Circumstances 1 & 2	String
Vehicle Information	Vehicle Configuration, Cargo Body, Vehicle Year, Vehicle Make, Vehicle Model, Vehicle Style	Integer, String
Crash Sequence	Vehicle Action, Sequence of Events - 1st to 4th Event, Most Harmful Event	String
Speed and Control	Speed Limit (MPH), Traffic Controls	Integer, String
Damage Details	Fixed Object Struck, Most Damaged Area, Vehicle Damage Extent	String
Crash Severity	Crash Severity, Major Cause	Integer, String
Surface and Environment	Surface Conditions, Drug or Alcohol Related, Road Type, Work Zone, Location	String, Boolean

2.3 Data Volume:

Dataset	Number of Entries	Number of Fields
Vehicle Crash in Iowa	543,337	37
Iowa Dot Winter Road Conditions	1,011	50
People Injured in Iowa Crashes	181,000	18
Drivers and Vehicles involved in Iowa crash	1,120,000	40

2.4 Focus Attributes:

Given the project's aim of analysing crashes, several key attributes will be the focus of the analysis

- **Crash Severity**: The severity level of each crash (minor, moderate, severe) will be a key focus, especially when analysing factors like weather conditions and driver behaviour.
- **Road Condition**: The condition of the road at the time of the crash (icy, wet, snow-covered, etc.) will be an essential focus, especially when integrated with weather data from the Winter Road Conditions dataset.
- **Demographic Details of Injured People**: From the People Injured in Iowa Crashes dataset, demographic attributes such as **age**, **gender**, and **injury severity** will be examined to identify high-risk groups in specific types of crashes.
- **Driver Risk Behaviour**: As driver behaviour often plays a key role in crashes, attributes like **distraction**, **alcohol involvement**, and **speeding** will be critical for identifying the causes of severe crashes.

2.5 Filtering:

Effective filtering will ensure that the analysis is relevant and focused on the key patterns:

- **Time Period**: The dataset will be filtered to focus on the winter months, especially to correlate with the Iowa DOT Winter Road Conditions dataset. Only crashes from October to March will be included, as this is when winter weather significantly affects road conditions in Iowa.
- **Crash Severity**: Data will be filtered to prioritize crashes with significant injury or fatality outcomes, as these are the most critical for understanding how road conditions or driver behaviour contribute to severe outcomes.

- **Driver Behaviour**: Specific focus will be placed on crashes involving risky driver behaviour, such as **distracted driving**, **alcohol involvement**, or **speeding**, to analyze their correlation with severe accidents.

2.6 Data Cleaning:

To ensure the data is in a suitable format for analysis, the following cleaning steps will be applied:

- **Handling Missing Data:** Critical fields such as crash severity, injury details, and road conditions will be closely examined. Minor missing values will be addressed through imputation techniques, while records with substantial gaps in essential fields will be excluded if they undermine the integrity of the analysis.
- **Outlier Management:** Outliers that significantly deviate from standard values, such as unusually high injury counts or crash severity scores, will be analysed. If these outliers are found to skew overall trends, they will be adjusted or removed to maintain the robustness of the results.
- **Ensuring Data Type Consistency:** We will standardize data types across all datasets, ensuring that fields like dates are uniformly formatted, numerical values are accurate, and categorical data (e.g., road conditions) have consistent labels. This step aims to eliminate discrepancies and enhance data interoperability.
- **Data Integration:** The datasets will be merged based on shared keys like Crash ID or Crash Date, linking crash details, road conditions, injury records, and driver behaviour. This approach will create a comprehensive dataset that accurately represents each incident and its associated factors.
- **Excluding Irrelevant Data:** Non-essential columns that do not directly contribute to the crash analysis, such as internal identifiers, agency names, or extraneous geospatial details, will be removed to streamline the dataset.
- **Eliminating Duplicates:** Duplicate entries, resulting from the same crash being reported multiple times or across different datasets, will be identified and removed to prevent double-counting and ensure data integrity.
- **Focus on Seasonal Relevance:** To align with the study's emphasis on winter conditions, we will filter out data from non-winter months, focusing exclusively on crash incidents influenced by adverse weather conditions during the winter season

3. <u>Deliverable 3</u>

Comprehensive Data Analysis and Insights:

3.1 Data Process

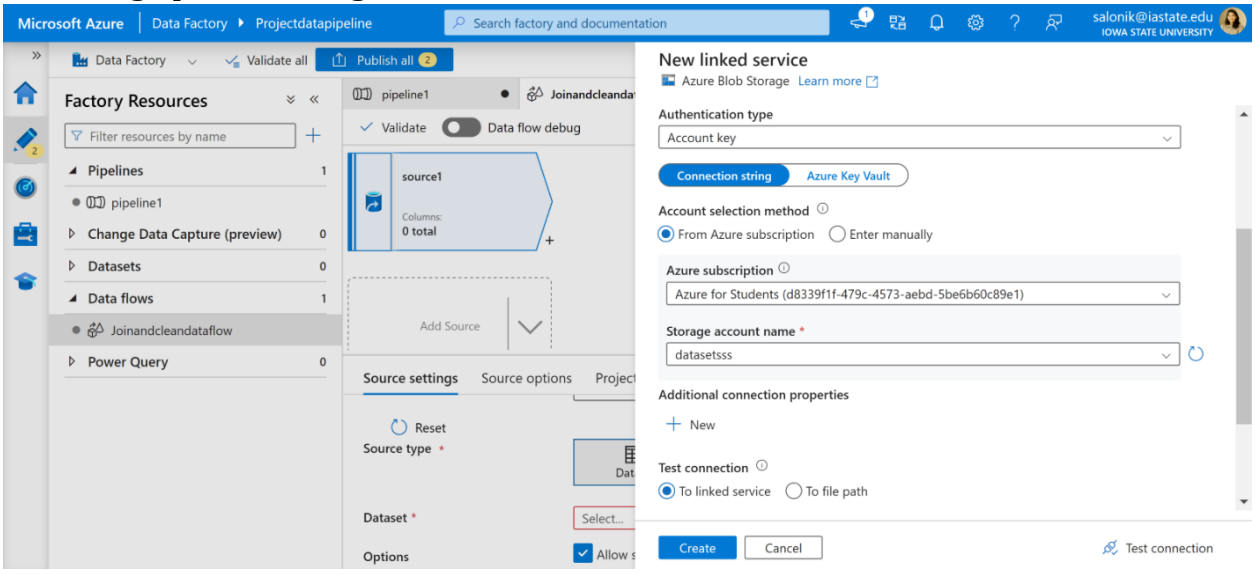
3.1.1 Software and Systems Used

- **Azure Data Lake Storage:** Primary cloud repository for scalable, secure storage and easy access to datasets.
- **Azure Fabric:** Centralized workspace setup for efficient data processing, management, and storage.
- **Azure SQL:** Database for structured data storage, querying, and management of relational data.

- **Power BI/Tableau:** Visualization tools for identifying crash data patterns via interactive dashboards.
- **Azure Data Factory:** Pipeline builder for streamlined data merging, transformation, and transfer.
- **Lakehouse:** Staging area for cleaning, transforming, and preparing datasets for analysis.
- **Azure Lakehouse Notebook:** Collaborative tool for data wrangling, transformation, and reproducible processing.
- **PySpark:** Distributed computing tool for large-scale data processing and efficient transformation.
- **Synapse Data Engineering:** Platform for scalable data integration, preparation, and advanced engineering tasks.
- **Microsoft SQL Server Management Studio (SSMS):** Tool for managing and querying SQL Server databases.

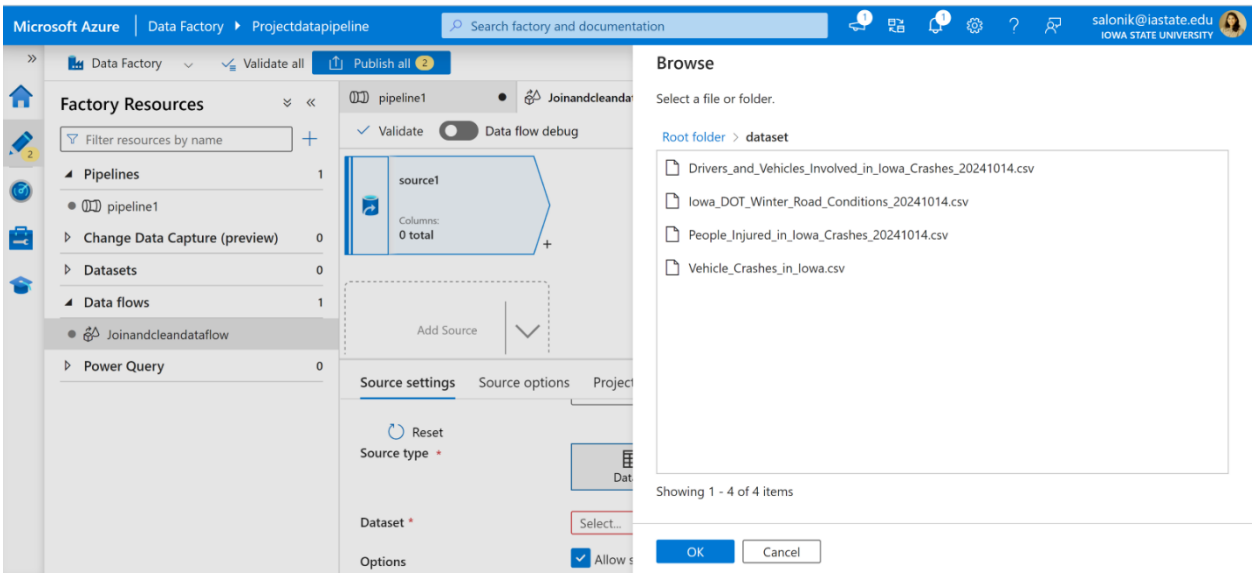
3.1.2 Steps Taken

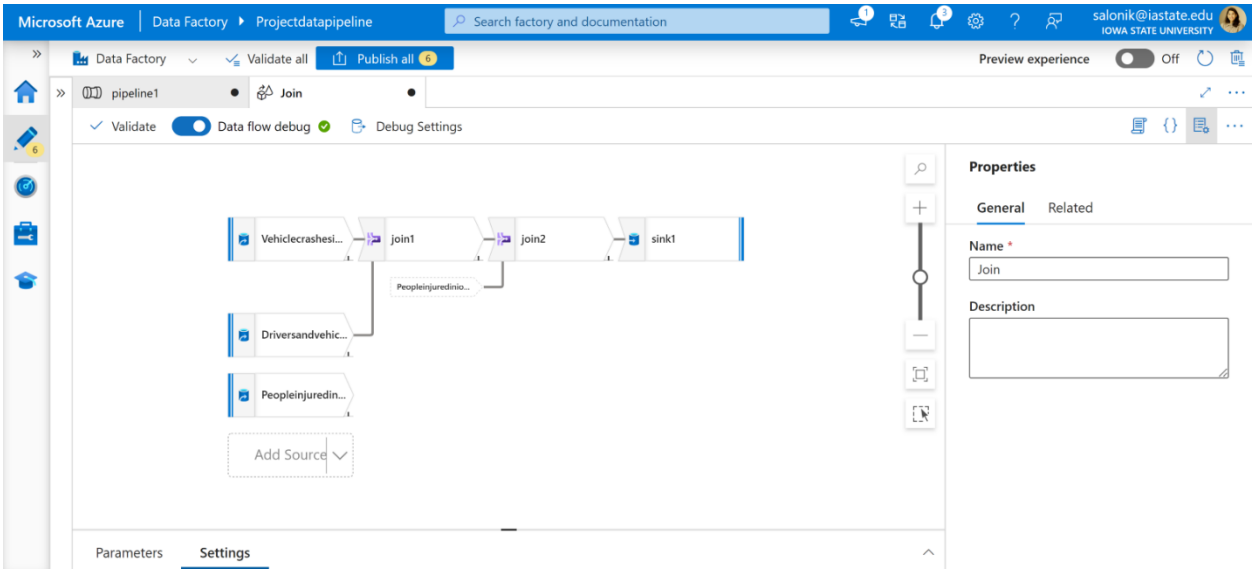
- **Setting Up Cloud Storage:**

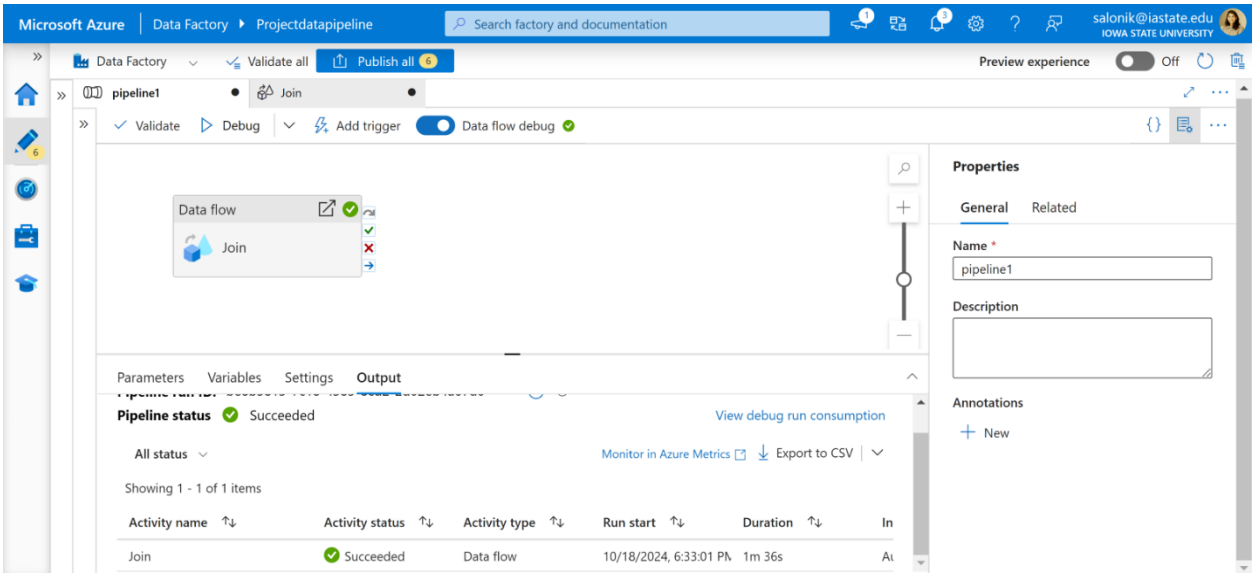


  - We created an Azure Data Lake Storage account to serve as our primary repository for data storage.
  - We set up containers within the data lake to organize our datasets, which included the vehicle crash data, road conditions data, and injury data.
  - We then uploaded each dataset into the respective containers in Azure Data Lake Storage for easy access and management

- **Creating a Data Pipeline in Azure:**


Microsoft Azure
Data Factory
Projectdatapipeline
Search factory and documentation
salonik@iastate.edu
IOWA STATE UNIVERSITY
Data Factory
Validate all
Publish all 2
Factory Resources
Filter resources by name
Pipelines 1
pipeline1
Change Data Capture (preview) 0
Datasets 0
Data flows 1
Joinandcleandataflow
Power Query 0
pipeline1
Validate
Data flow debug
source1
Columns:
0 total
Add Source
Source settings
Source options
Reset
Source type *
Dataset *
Select...
Options
Browse
Select a file or folder.
Root folder > dataset
Drivers_and_Vehicles_Involved_in_Iowa_Crashes_20241014.csv
Iowa_DOT_Winter_Road_Conditions_20241014.csv
People_Injured_in_Iowa_Crashes_20241014.csv
Vehicle_Crashes_in_Iowa.csv
Showing 1 - 4 of 4 items
OK
Cancel

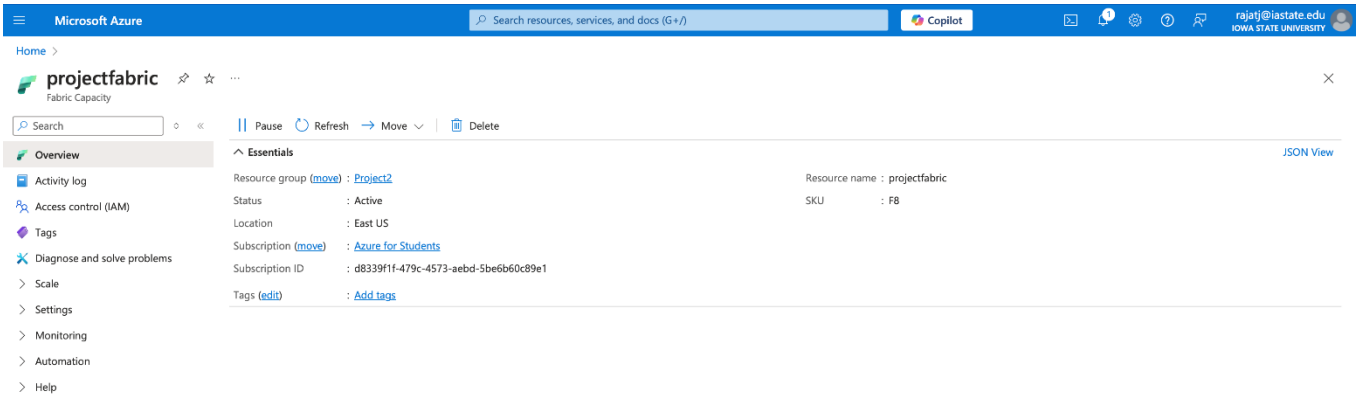

Microsoft Azure
Data Factory
Projectdatapipeline
Search factory and documentation
salonik@iastate.edu
IOWA STATE UNIVERSITY
Data Factory
Validate all
Publish all 6
Preview experience
Off
pipeline1
Join
Validate
Data flow debug
Debug Settings
Vehiclecrashesi...
join1
join2
sink1
Peopleinjuredinio...
Driversandvehic...
Peopleinjuredin...
Add Source
Properties
General
Related
Name *
Join
Description
Parameters
Settings



- We set up a pipeline in Azure Data Factory to automate the data processing steps. We configured the pipeline to identify and merge the three major datasets (vehicle crashes, road conditions, and injury data) using the unique identifier, Iowa DOT Case Number.
- We applied the Join function in the pipeline to merge records across datasets based on the Iowa DOT Case Number, ensuring each row represents a unique incident with relevant details from all sources.
- Finally, we used the Sink function to store the final merged dataset within Azure Data Lake Storage, preparing it for further analysis.

- **Fabric Setup and Workspace Configuration**

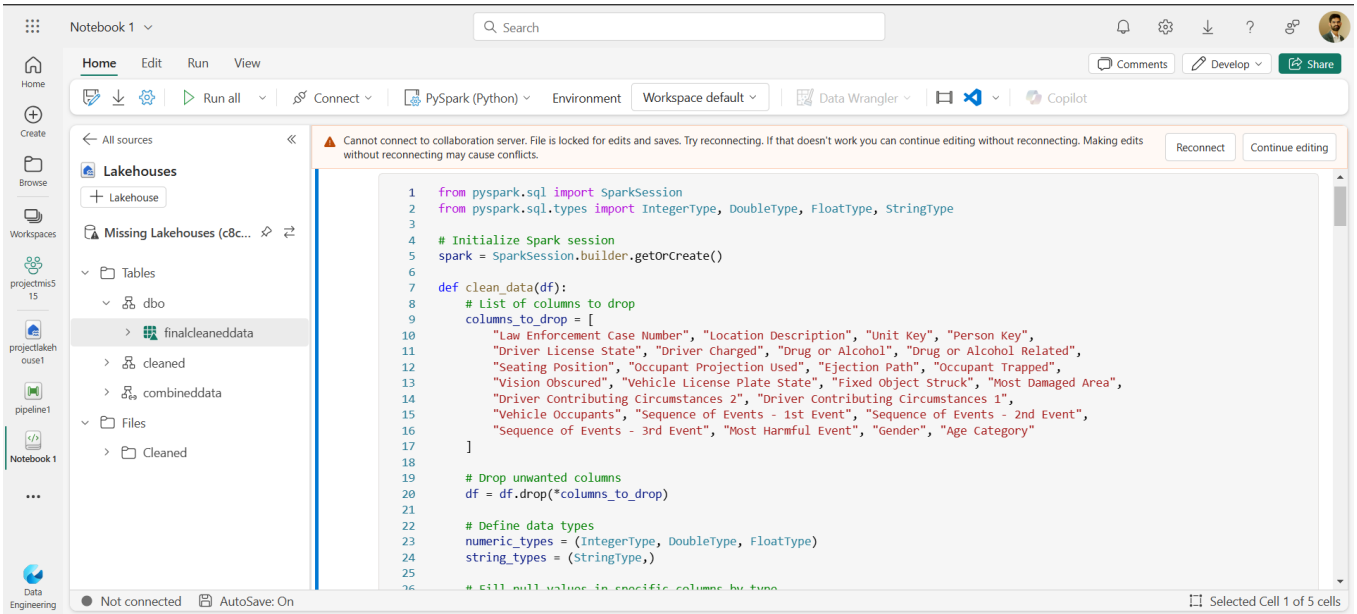
  - **Fabric Capacity Creation:** We set up Fabric capacity within Azure to handle our data processing and storage needs for large datasets, ensuring scalable and efficient data operations.
  - **Fabric Workspace Creation:** We created a dedicated Fabric workspace to centralize our project resources and manage data flow.
  - **Lakehouse Setup:** We established a Lakehouse within the Fabric workspace to store, clean, and manage the merged dataset. This Lakehouse served as our staging area for data preparation and further transformations.



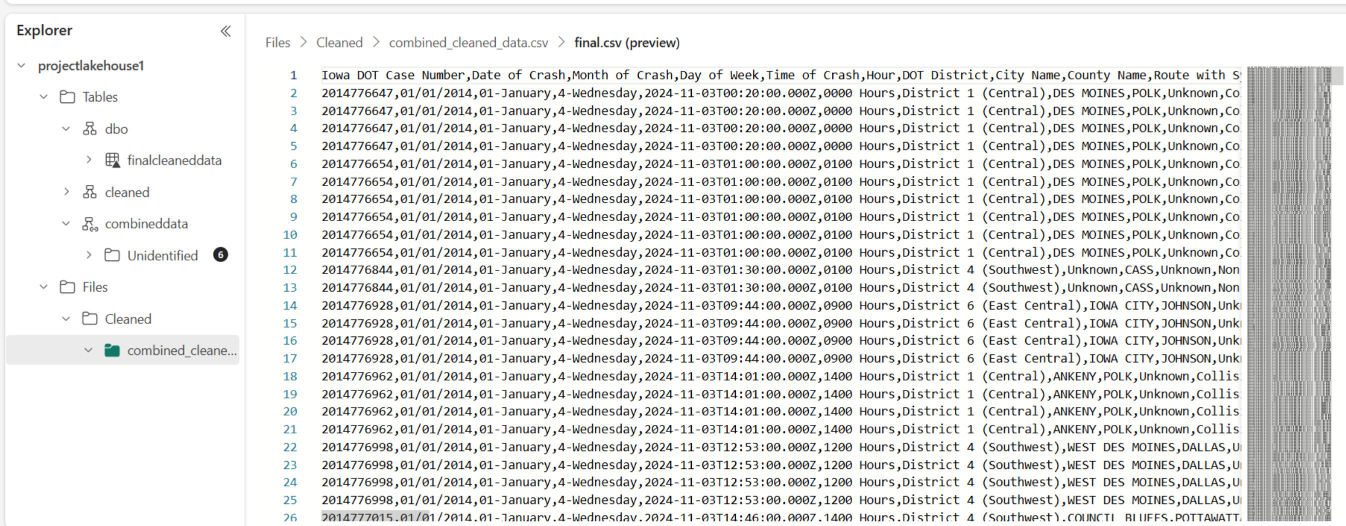
- **Data Aggregation, Cleaning & Pre-processing:**

  - We have created a Py spark code in synapse data engineering, that will iterate through all the 6 datasets we have and drops the unwanted columns; defines the correct datatypes to all the columns and replaces all the null values for columns with specific datatypes & then combines all the datasets into a single file and stores it in azure Lakehouse.

    Py spark code used for Data processing & Data Cleaning:

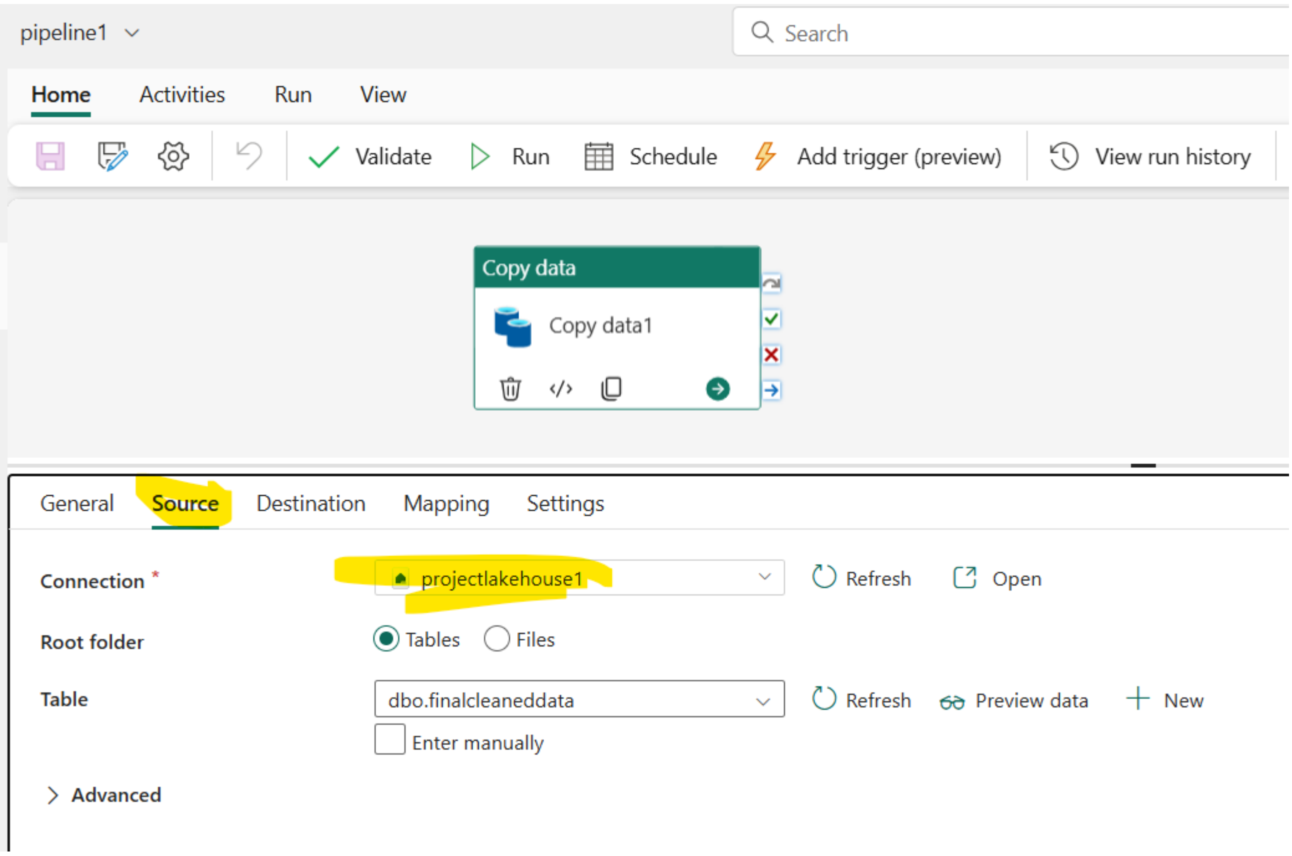


Output file stored in Lakehouse:

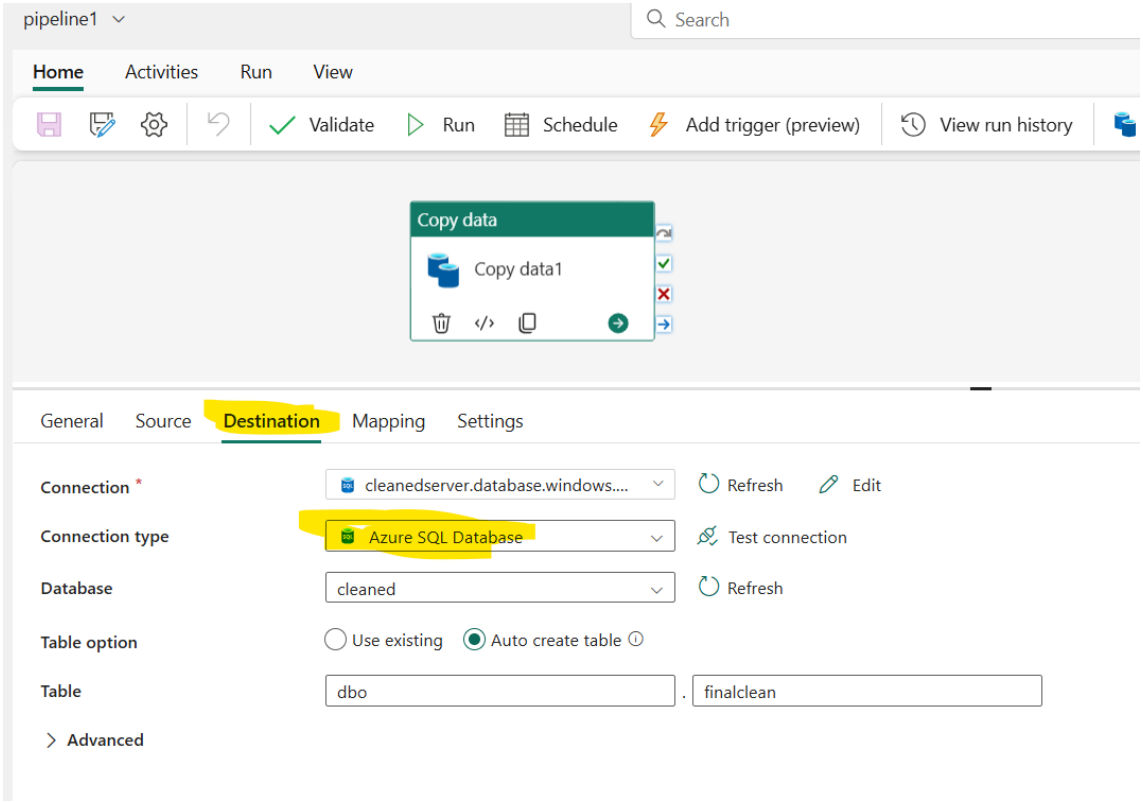


And then, for further analysis, we have created a SQL server & SQL database in Azure, to query on this cleaned dataset. To transform this final.csv file into SQL tables, we have used Azure Data Factory and ran a pipeline to load this into Azure SQL database. We have used, Copy Data activity to get them to load the csv file into SQL database in tables form.

Source - Lakehouse:

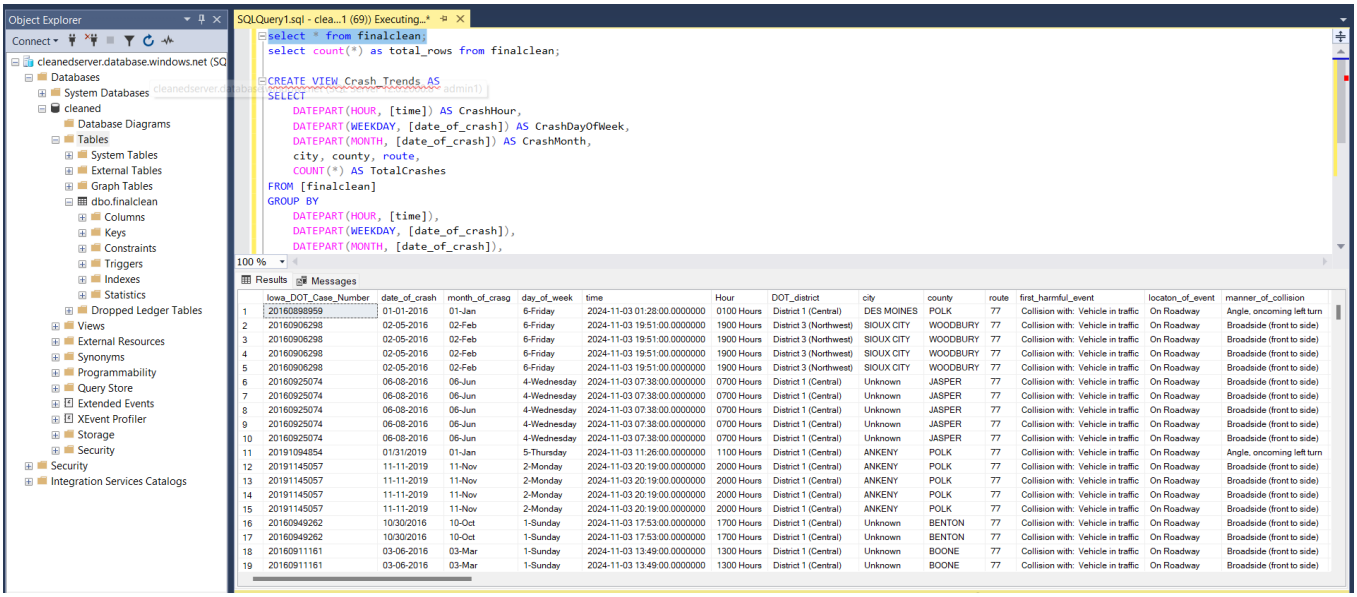


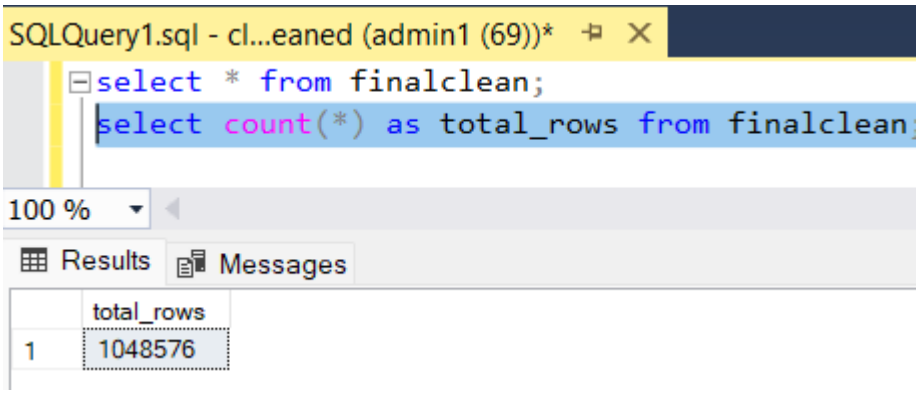
- **Destination – Azure SQL database:**



- **Data:**

After the tables are loaded, we have used Microsoft SQL server Management Studio to connect and interact with the database we have created. (screenshot of table final_clean loaded from lakehouse)





- **Descriptive Analysis:**

After that for descriptive analytics, we ran a SQL query that aggregates several columns and creates a final view named crash_trends which provides insights into crash trends including time, location, date & hour.

```sql
CREATE VIEW Crash_Trends2 AS
SELECT
    DATEPART(HOUR, [time]) AS CrashHour,
    DATEPART(WEEKDAY, [date_of_crash]) AS CrashDayOfWeek,
    DATEPART(MONTH, [date_of_crash]) AS CrashMonth,
    city, county, route,
    COUNT(*) AS TotalCrashes
FROM [finalclean]
GROUP BY
    DATEPART(HOUR, [time]),
    DATEPART(WEEKDAY, [date_of_crash]),
    DATEPART(MONTH, [date_of_crash]),
    city, county, route;
```

00 %

Messages

```
Commands completed successfully.

Completion time: 2024-11-04T16:45:01.5859565-06:00
```

```sql
    city, county, route;

select * from Crash_Trends;
```

100 %

Results | Messages

	CrashHour	CrashDayOfWeek	CrashMonth	city	county	route	TotalCrashes
1	0	1	1	AMES	STORY	Unknown	6
2	0	1	1	ANKENY	POLK	77	3
3	0	1	1	ARNOLDS PARK	DICKINSON	77	3
4	0	1	1	BETTENDORF	SCOTT	77	3
5	0	1	1	BETTENDORF	SCOTT	US 67	6
6	0	1	1	BURLINGTON	DES MOINES	77	8
7	0	1	1	BURLINGTON	DES MOINES	Unknown	4
8	0	1	1	CARROLL	CARROLL	77	2
9	0	1	1	CEDAR FALLS	BLACK HAWK	77	4
10	0	1	1	CEDAR RAPIDS	LINN	77	3
11	0	1	1	CLEAR LAKE	CERRO GORDO	77	1
12	0	1	1	CLINTON	CLINTON	77	2
13	0	1	1	CLINTON	CLINTON	Unknown	2
14	0	1	1	CLIVE	DALLAS	77	1
15	0	1	1	COLFAX	JASPER	77	2
16	0	1	1	COLO	STORY	US 30	4
17	0	1	1	CORALVILLE	JOHNSON	I-80	1
18	0	1	1	COUNCIL BLUFFS	POTTAWATTAMIE	77	13
19	0	1	1	DAVENPORT	SCOTT	77	24
20	0	1	1	DAVENPORT	SCOTT	I-280	1
21	0	1	1	DAVENPORT	SCOTT	I-80	2
22	0	1	1	DAVENPORT	SCOTT	Unknown	6
23	0	1	1	DES MOINES	POLK	77	14
24	0	1	1	DES MOINES	POLK	I-235	2
25	0	1	1	DES MOINES	POLK	Unknown	20

3.2 Assumptions

- **Uniformity of Data Collection Methods Across Datasets:** We assumed that each dataset (drivers and vehicles involved, people injured, and vehicle crashes) followed a similar methodology for data collection. This assumption allows us to merge these datasets under the same schema without adjusting for potential variations in data collection procedures across different sources. Consistent data collection methods ensure compatibility for columns like Iowa_DOT_Case_Number, which was used as a unique identifier for merging.

- **Temporal Consistency of Date and Time Formats:** We assumed that all date and time fields, such as date_of_crash, month_of_crash, and time, were formatted consistently across the merged datasets. This assumption allowed us to accurately convert and manipulate these fields for analysis without needing extensive reformatting. Ensuring that these temporal fields are uniform is crucial for accurate time-based visualizations and analyses, such as understanding crash trends over different days and times.

- **Driver and Vehicle Information Completeness:** We assumed that the information regarding driver demographics (like driver_age and gender_driver) and vehicle details (such as vehicle_make and vehicle_year) was sufficiently complete to draw meaningful conclusions. This means that while some records might have missing data, we presumed that the majority of records contained the essential information needed for our analysis. This assumption helped streamline the cleaning process and supported the use of these variables in visualizations and predictive modelling without extensive imputations.

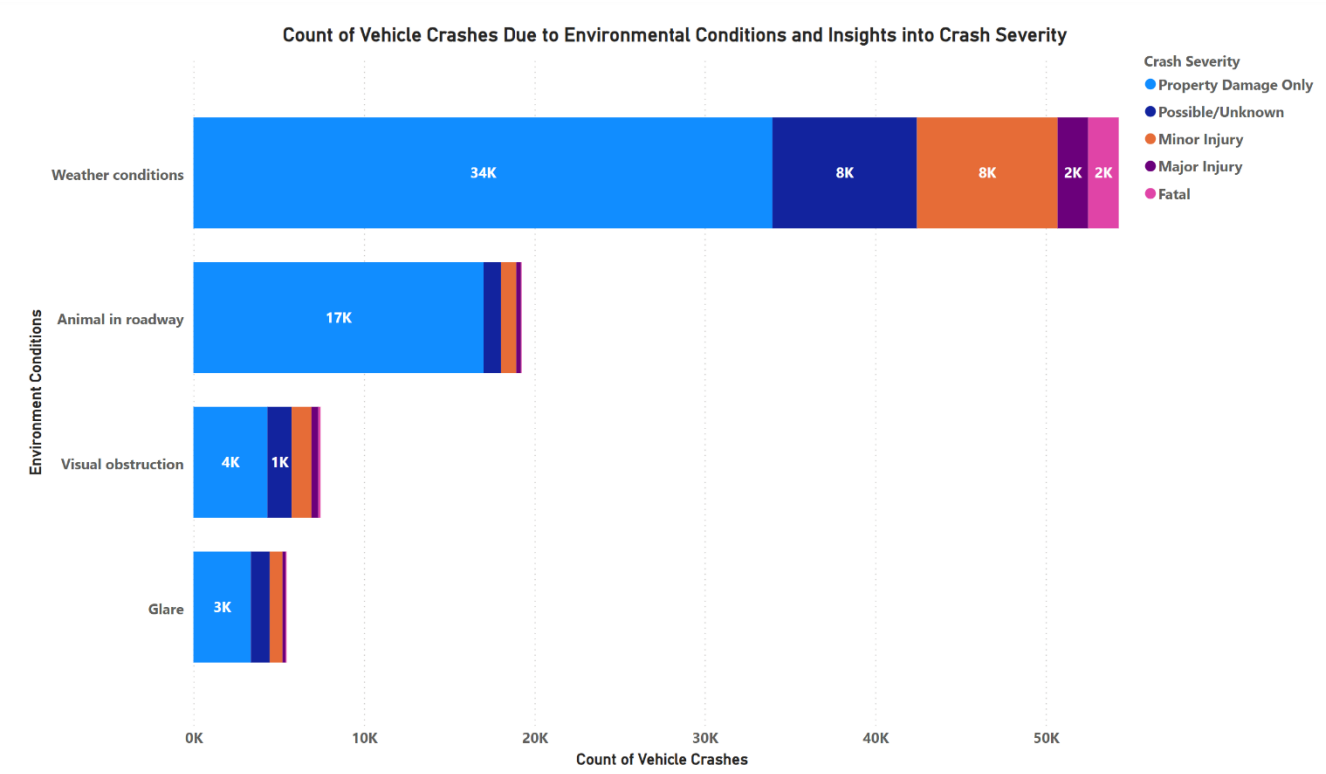
- **Data Integrity and Referential Consistency:** We assumed that the relationships between the datasets were maintained correctly during the merging process, specifically regarding foreign key references such as Iowa_DOT_Case_Number. This assumption implies that every case number in the merged dataset accurately corresponds to its respective records in the original datasets for drivers, vehicles, and injuries. Ensuring referential integrity is crucial for accurate analysis and prevents issues such as orphaned records or duplicated entries, which could lead to misleading insights in the final combined dataset.

3.3 Visualizations

We utilized Power BI to visualize the data by connecting it to our Azure SQL database as the data source.

3.3.1 Bar Chart of Crash Severity vs. Environmental Conditions

This bar chart visualizes the relationship between crash severity and environmental conditions during the incidents. By comparing the frequency of different crash severity levels across various environmental conditions stakeholders can identify patterns that may indicate higher risks under certain conditions. Insights derived from this visualization can inform policymakers and safety organizations about the need for enhanced road safety measures, such as better signage or road maintenance during adverse weather conditions.



3.3.2 Heat Map of Crashes by Time and Day of the Week

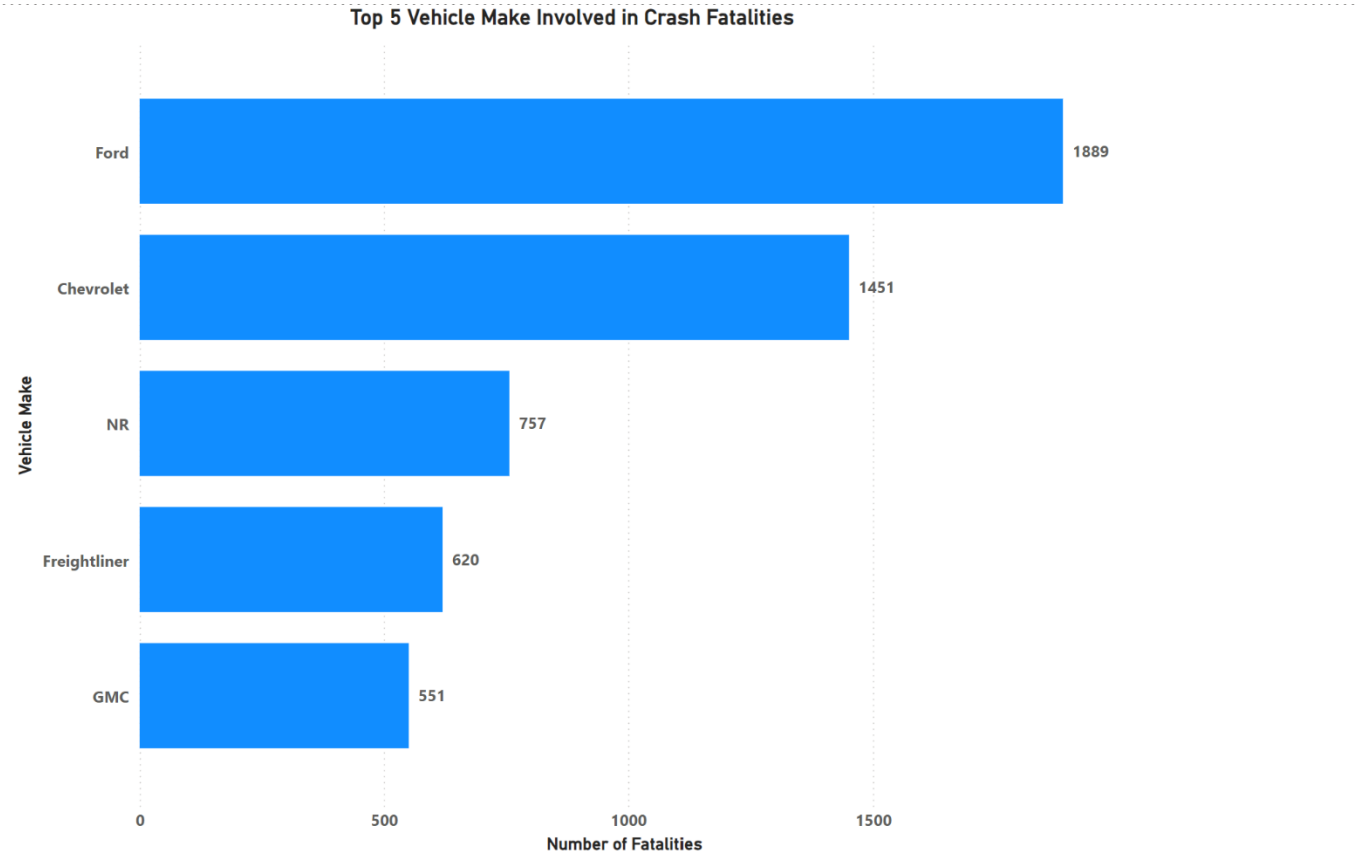
The heat map displays the distribution of crashes across different times of the day and days of the week, providing a visual representation of when crashes are most likely to occur. By analyzing this data, safety officials can identify peak periods for crashes, allowing them to allocate resources more effectively, such as increasing law enforcement presence or implementing public awareness campaigns during high-risk times. This visualization helps in understanding temporal patterns in crash occurrences, which can lead to improved traffic safety strategies.

day_of_week	0600 Hours	0700 Hours	0800 Hours	1200 Hours	1300 Hours	1400 Hours	1500 Hours	1600 Hours	1700 Hours	1800 Hours	Total
1-Sunday	1776	1839	2392	7012	6686	6575	7043	9030	6734	6417	**55504**
2-Monday	4980	10951	8851	11468	9405	10298	14448	13978	14576	7904	**106859**
3-Tuesday	5407	12814	9817	9499	8570	9743	14373	14040	15353	8101	**107717**
4-Wednesday	5117	11447	9385	9570	9079	10416	13785	14074	14910	9030	**106813**
5-Thursday	4587	10527	8904	9805	8884	10049	15001	15321	15725	8627	**107430**
6-Friday	4530	9971	8708	11774	10841	12078	17292	18420	16713	10640	**120967**
7-Saturday	2459	2935	4092	9558	8714	8785	9007	8360	8666	7639	**70215**
Total	**28856**	**60484**	**52149**	**68686**	**62179**	**67944**	**90949**	**93223**	**92677**	**58358**	**675505**

3.3.3 Tree map of Vehicle Make vs. Number of Fatalities

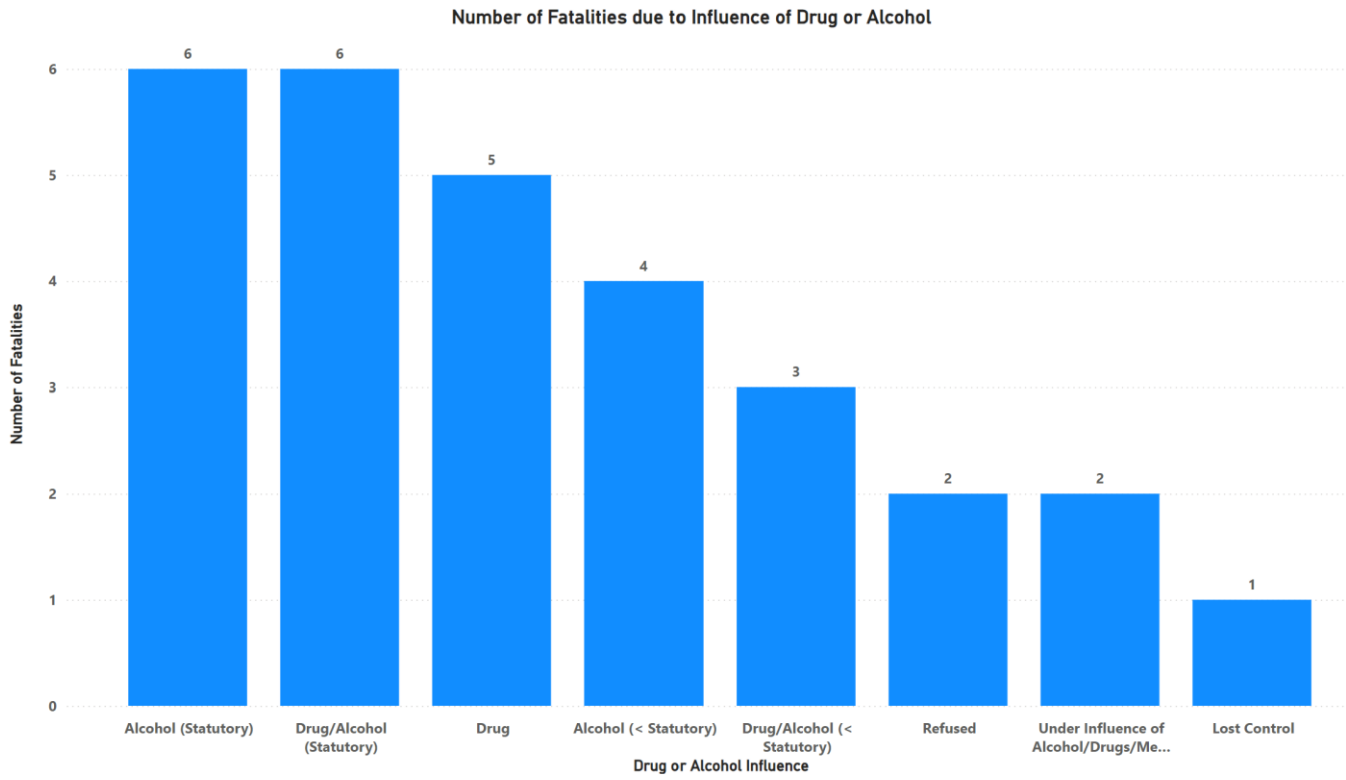
This tree map illustrates the relationship between different vehicle makes and the number of fatalities resulting from crashes involving those vehicles. Each rectangle represents a vehicle make, sized according

to the number of fatalities associated with it. This visualization helps stakeholders, such as manufacturers and regulatory bodies, identify which vehicle makes are involved in more fatal crashes. Understanding these trends can drive safety improvements, encourage manufacturers to enhance vehicle safety features, and inform consumer awareness regarding vehicle safety performance.



3.3.4 Clustered Column Chart of Drug or Alcohol Test Results vs. Number of Fatalities

The clustered column chart visualizes the distribution of fatalities in crashes based on the results of drug or alcohol tests conducted on drivers involved in these incidents. Each slice represents a category of drug or alcohol involvement. By examining the proportions of fatalities linked to different test results, policymakers and public health officials can assess the impact of substance use on road safety. This visualization can inform targeted interventions, such as increased sobriety checkpoints or educational campaigns focused on the dangers of impaired driving.



3.4 Recommendations

- **Enhance Road Safety Measures Based on Environmental Conditions**

Given the bar chart findings that link crash severity to specific environmental conditions, local governments and road safety organizations should prioritize road maintenance and signage improvements in areas prone to adverse weather conditions. Implementing measures such as improved lighting, clearer signage, and increased visibility of road markings during rainy or foggy weather can reduce the likelihood of severe crashes, ultimately saving lives and reducing injuries on the road.

- **Targeted Traffic Enforcement During High-Risk Times**

The heat map indicates specific times and days of the week when crashes are more prevalent. Traffic enforcement agencies should allocate resources and increase police presence during these peak periods, focusing on educating drivers about safe driving practices and enforcing speed limits and DUI laws. This proactive approach can help reduce the frequency of crashes, improve overall road safety, and foster a culture of responsible driving among the community.

- **Collaborate with Vehicle Manufacturers for Improved Safety Features**

Insights from the tree map show a correlation between certain vehicle makes and higher fatality rates in crashes. Collaboration between regulatory bodies and vehicle manufacturers should be encouraged to promote the development and implementation of advanced safety features, such as automatic braking systems and lane-keeping assistance, in high-risk vehicles. This partnership can lead to a reduction in fatalities and improve the overall safety of the vehicles on the road.

- **Implement Educational Campaigns Focused on Impaired Driving**

The clustered column chart highlights the significant impact of drug and alcohol involvement in fatal crashes. Public health agencies and community organizations should develop and launch educational campaigns aimed at raising awareness about the dangers of impaired driving. These campaigns can include community events, social media outreach, and partnerships with local businesses to promote responsible drinking and alternative transportation options, thereby reducing the incidence of impaired driving and enhancing overall community safety.