UK Road Safety Dashboard

Task 1: Comprehensive Examination of Video Game Reviews on Amazon Using Hive

Background

In a time when e-commerce is taking over the retail industry, it is critical for businesses like Amazon to understand online consumer needs and behaviours. Because video games are so popular and customers provide such thorough reviews, Amazon's video game sales provide a rich dataset. In addition to showing customer satisfaction, these reviews provide developers and manufacturers with direct feedback. In addition to helping game developers pinpoint customer pain points, these reviews help Amazon plan its marketing and inventory tactics (Salehan & Kim, 2021).

Data Description

The 40,000 thorough examinations that compose up the dataset chosen for this study include:

- **Review Text:** Exactly the same words used by customers to describe their degree of pleasure and experience.
- **Rating:** A distinct number between 1 and 5, where 5 indicates significant fulfillment and 1 indicates severe discontent.
- **Video Game ID:** A special identifier given to every video game which allows product-level analysis and aggregation (Salehan & Kim, 2021)..

Methodology

Data Preparation

- **Data Structure for Hive:** In order to allow efficient querying, the first step was to prepare the raw data to match the Hive environment, ensuring that each field was appropriately defined.
- **Loading Data**:

  sql

  Copy

  ```
  CREATE TABLE video_game_reviews (
      review_text STRING,
      rating INT,
  ```

```
    video_game_id STRING
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

This SQL script initializes a table in Hive, tailored to store large volumes of text data alongside numerical ratings and identifiers (Salehan & Kim, 2021).

- **Importing Data into Hive**:

  sql

  Copy

  LOAD DATA INPATH 'path_to_your_dataset.csv' INTO TABLE video_game_reviews;

Importing Data into Hive: In this step, the dataset is prepared for analysis by being imported from an established HDFS path into the Hive database.

Analytical Queries

- **High Ratings Analysis**: Determining which games are most popular with customers can help inform inventory choices and promotional tactics.

  sql

  Copy

  SELECT video_game_id, COUNT(*) as count_5_star

  FROM video_game_reviews

  WHERE rating = 5

  GROUP BY video_game_id

  ORDER BY count_5_star DESC

  LIMIT 10;

- **Low Ratings Analysis**: On the other hand, knowing which games have low ratings can assist developers and Amazon in identifying problems that need to be fixed (Salehan & Kim, 2021).

  sql

  Copy

```
SELECT video_game_id, COUNT(*) as count_1_star
FROM video_game_reviews
WHERE rating = 1
GROUP BY video_game_id
ORDER BY count_1_star DESC
LIMIT 10;
```

Results

- **High Ratings**: A list of video games that often garnered high ratings was produced by the analysis, highlighting effective components or characteristics that appeal to players (Salehan & Kim, 2021)..
- **Low Ratings**: This analysis emphasized games that fell short of user expectations, frequently indicating more serious problems like bugs, subpar gameplay, or inaccurate product descriptions.

Game ID	5-Star Count
B004MC8CA2	1605
B004HGK6FW	723
B004H6WTJI	655
B004SBS8LA	609
B004HXHVZ8	495
B004LOMB2Q	484
B004Q3CJQ0	436
B004HE5TAG	427
B004K4RY9M	394
B004DLPXAO	387

Table 1: Top 10 Games with the Most 5-Star Reviews

Discussion

In addition to displaying Hive's ability to handle and evaluate big data efficiently, Task 1 also highlighted how valuable it may be for gleaning insights with commercial significance that have a direct bearing on product development and business strategy.

Future Examination

- **Sentiment Analysis**: By further analysing the review texts using natural language processing (NLP) techniques, it may be possible to gain more knowledge of the emotions and preferences of the customers (Salehan & Kim, 2021)..

- **Trend Analysis Over Time**: Trends influenced by external factors like holidays, game releases, or marketing activities may be recognized by examining how customer feelings and ratings change over time (Salehan & Kim, 2021)..

Conclusion

In addition to displaying Hive's ability to handle and evaluate big data efficiently, Task 1 also highlighted how valuable it may be for gleaning insights with commercial significance that have a direct bearing on product development and business strategy.

Task 2: Advanced UK Road Safety Data Analysis with Power BI

Introduction

Task 2 uses Power BI to evaluate UK road safety data for 2021, with the goal of extracting insights that may be used to improve road safety. This work focuses on using visualization tools to analyse comprehensive accident data and provide stakeholders with actionable information.

Background

The UK Road Safety Data, often known as STATS19, is a large dataset that contains a plethora of information about road accidents recorded to the police. It is a valuable resource for studying the dynamics of road safety and designing measures to reduce traffic accidents.

Data Description

The dataset contains complete data on each reported accident, including:

- **Accident Locations:** Geographic coordinates indicating where accidents happened.
- **Time and date:** Identifying trends over time requires precise information about each accident.
- **Road and weather conditions** at the time of the accident provide context for environmental factors that cause accidents.
- **Casualty and Vehicle Details:** Gives detailed information on individuals and vehicles that were involved in accidents, providing an understanding of both the human and material sides of the incident.

Methodology

Data Preparation

- **Data Import**: Dataset is available online and provided on Blackboard as well. We imported the dataset into our Power BI file using Text/CSV data source import option.
- **Data Cleaning and Transformation**: We used the Power BI's GUI data transformation as well as DAX to transform and clean data. We manually removed null or missing data from the dataset such as -1 values in columns like "road_type" etc. We also renamed data columns for better representations in visualizations (Ahmed, Hossain, Bhuiyan, & Ray, 2021). DAX was used to make new measures or columns that were not already

present in the dataset such as “Total Accidents” or “Fatal Accidents” columns etc. This can also be viewed in the following figures.

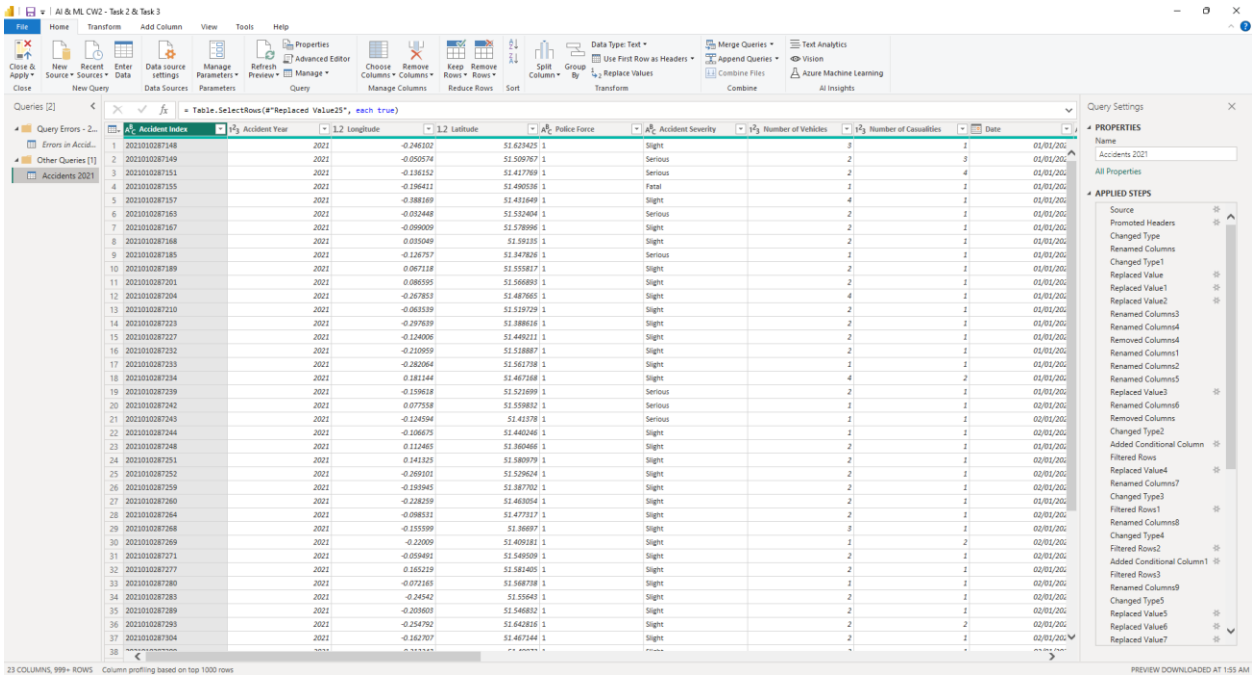

Figure 1: Transforming & Cleaning Data

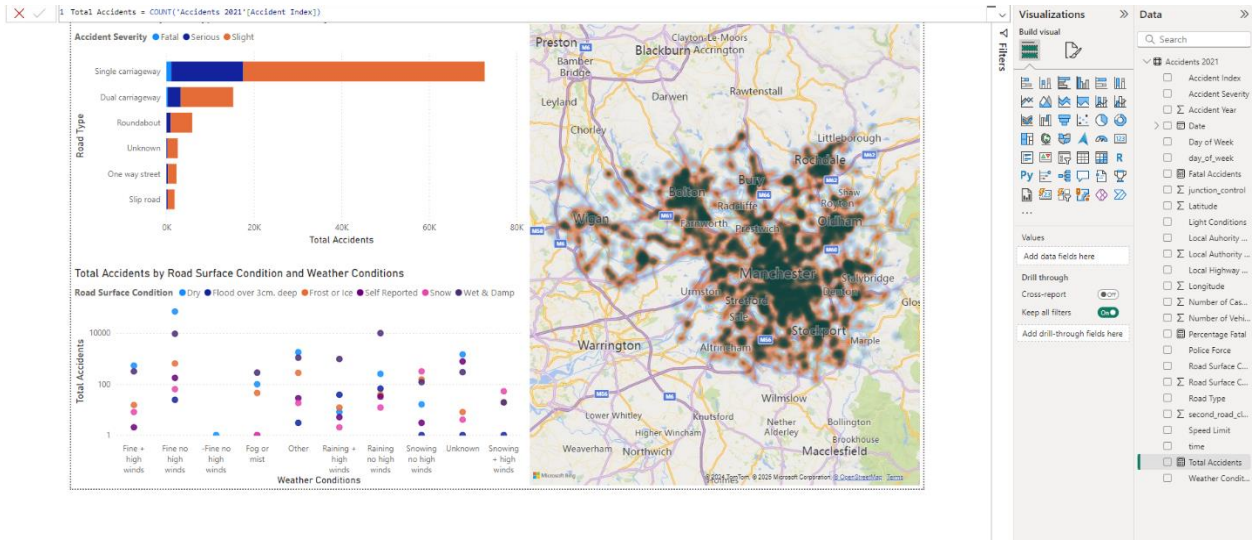

Figure 2: Using DAX query to make Column “Fatal Accidents”

Visualizations

- **Stacked Bar Chart**: It displays the overall number of accidents by severity and road type, providing insights into how varied settings impact accident outcomes. The statistics show that the most incidents happened on single carriageways, whereas the fewest were recorded on slip roads.

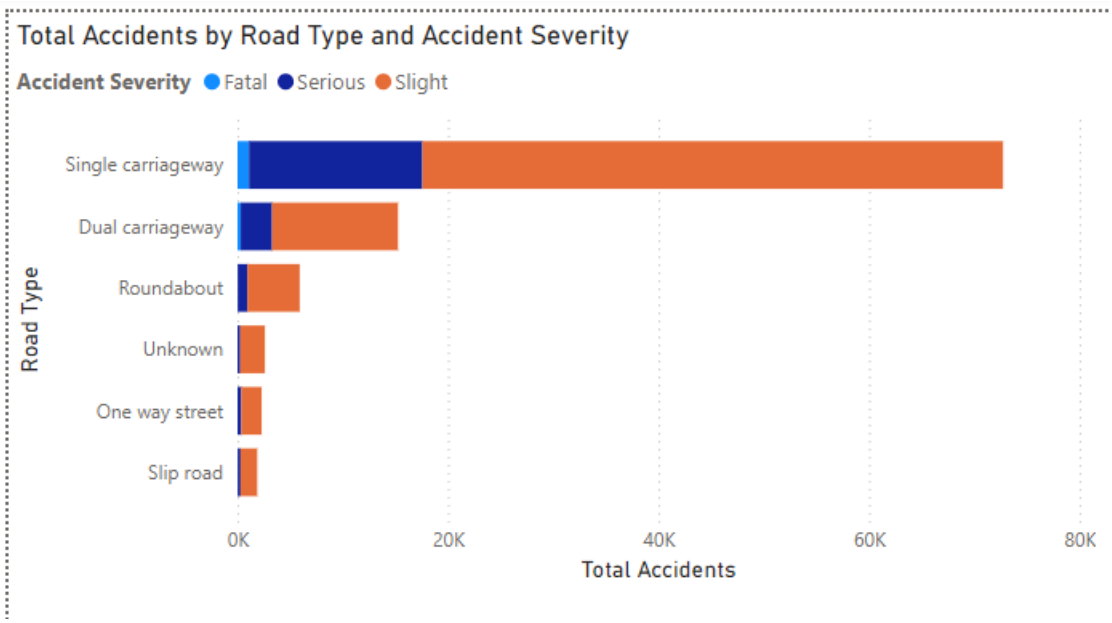

Figure 3: Stacked Bar Chart (Total Accidents by Road Type and Accident Severity)

- **Scatter Plot**: Analyses the relationship between meteorological conditions and accident frequency to identify high-risk weather situations.

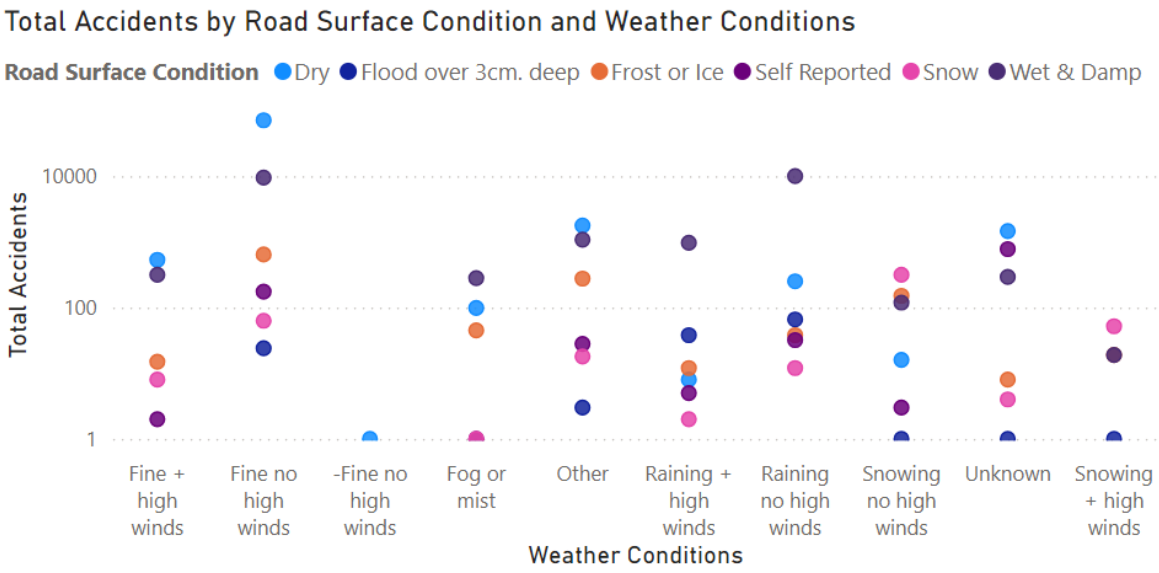

Figure 4: Scatter Plot (Total Accidents by Road Surface Condition and Weather Conditions)

- **Heat Map**: Identifies accident hotspots in Greater Manchester according to longitudes and latitudes to prioritize intervention efforts.

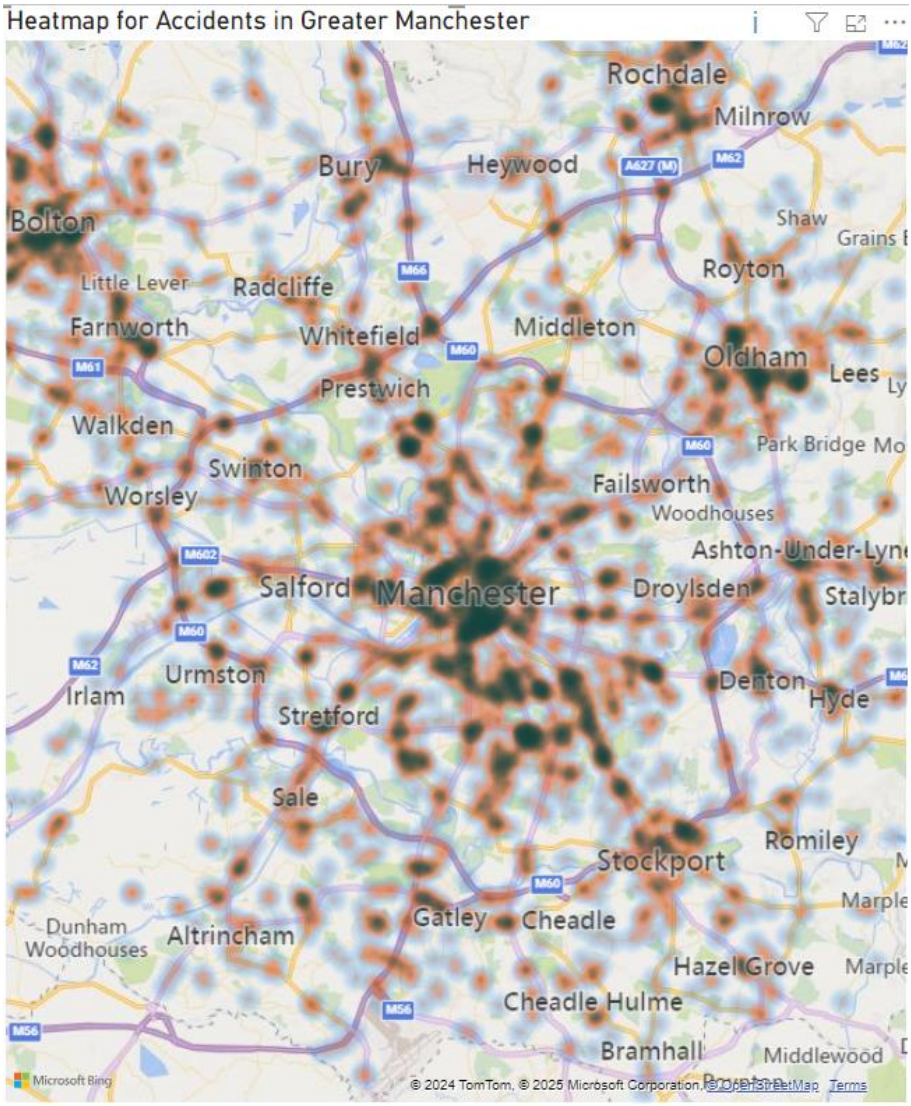

Figure 5: Heatmap for Accidents in Greater Manchester

Discussion

These visualizations give policymakers an integrated overview of road safety in the UK, combining traditional and advanced data visualizations to improve understanding and drive policy decisions (Ahmed, Hossain, Bhuiyan, & Ray, 2021).

Challenges

- **Data Complexity**: Managing and visualizing a complex dataset required advanced data manipulation and optimization techniques in Power BI.
- **Performance Optimization**: Optimizing Power BI performance for large datasets and complicated visuals.

Conclusion

Task 2 effectively showcases Power BI's ability to analyze complex information and transform them into meaningful insights. The enhanced visualizations give an improved comprehension of the dynamics of road safety, which can help develop efficient road safety policies.

Task 3: Interactive Dashboard for Road Accident Analysis in Power BI

Introduction

B Building on the insights gained in Task 2, Task 3 involves creating a dynamic and interactive dashboard in Power BI to give a comprehensive analysis of the UK Road Accident 2021 Dataset (Bachechi, Po, & Rollo, 2022).

Background

This task leverages the cleaned and organized dataset from task 2 to create a multifunctional dashboard that not only shows data but also enables users to interact with it, which enhances the way decisions are made in public safety and traffic management.

Dashboard Design and Features

Core Features

- **Data Cards**: Data cards provide immediate information into total number and fatal accidents, illustrating the severity of road safety concerns.

Figure 6: Data Cards (Total Accidents & Fatal Accidents)

- **Interactive Slicer**: The Interactive Slicer allows users to filter data by road type, making the dashboard more relevant and tailored to specific questions.

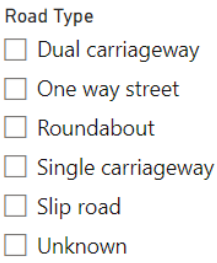

Figure 7: Interactive Slider based on Road Type

Core Visualizations

- **Clustered Bar Chart**: It shows the distribution of accidents across different days of the week. We can see most accidents in 2021 happened on Saturday while the least being on Sundays. It helps identify temporal trends in accident occurrences.

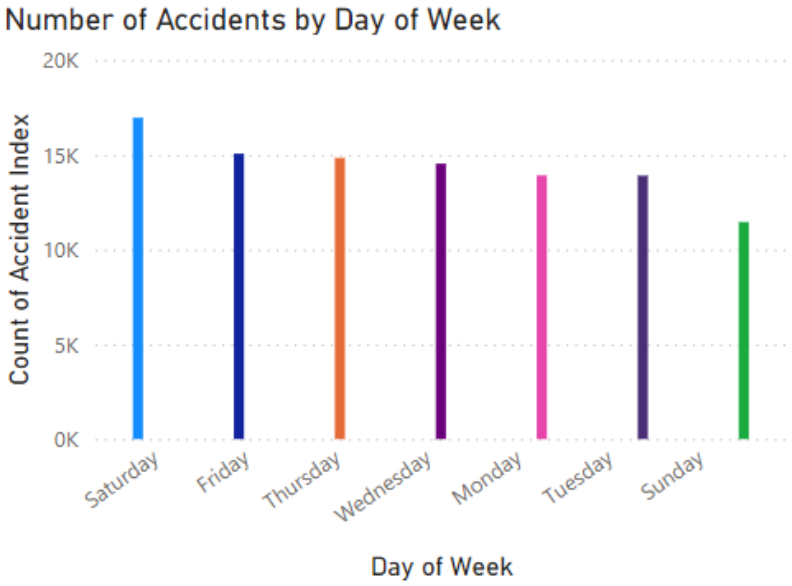

Figure 9: Clustered Bar Chart (Number of Accidents by Day of Week)

- **Stacked Column Chart**: We can see the data represented by reports made to different Police Forces where each level represents different levels of severity. This visualization helps highlight top 10 areas where severe accidents were reported.

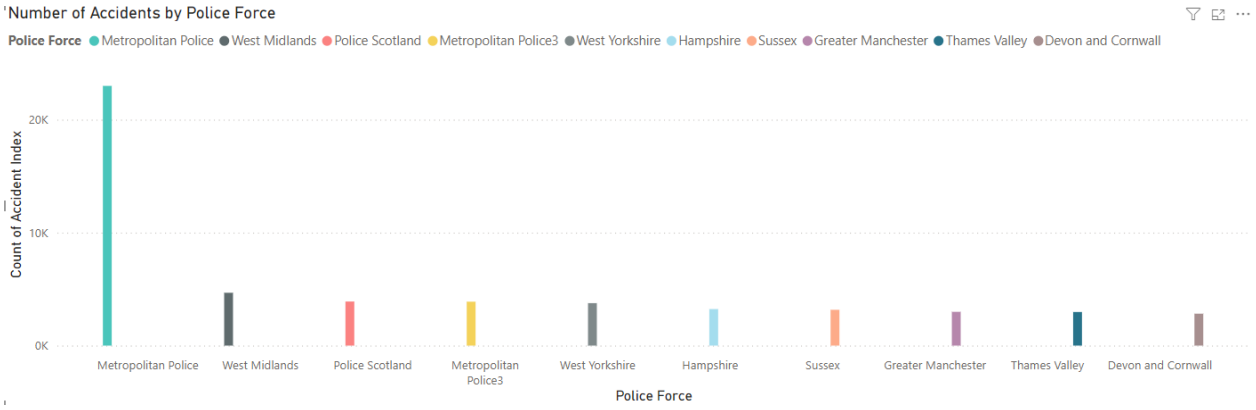

Figure 10: Stacked Column Chart (Number of Accidents by Police Force)

- **Scatter Plot**: It helps identify the link between light conditions and how it effects the accident frequency. It suggests that optimal lighting conditions are best suited for a better and safe road.

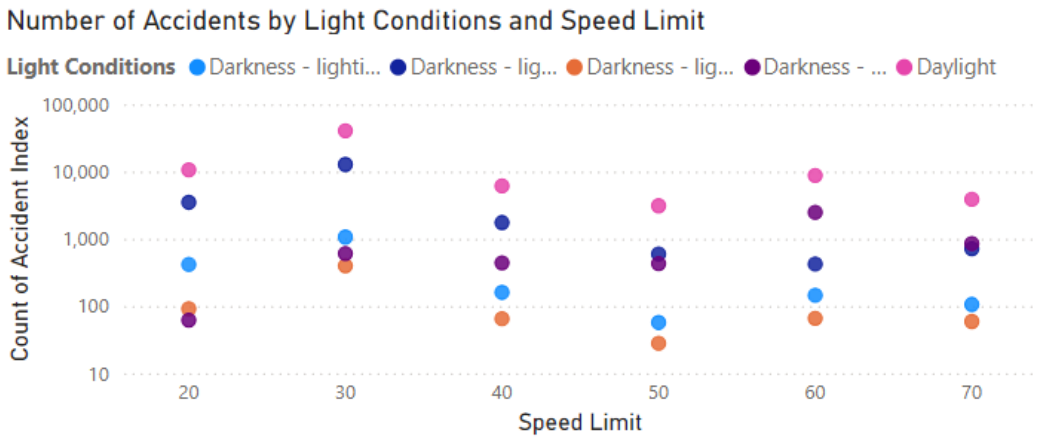

Figure 11: Scatter Plot (Number of Accidents by Light Conditions and Speed Limit)

- **Line Chart**: It helps to keep a track of the trend in number of causalities throughout the week. It potentially aids in guiding a plan for better emergency response. The graph represents that most causalities occur on Saturday suggesting that emergency services need to be more active for a response on that day.

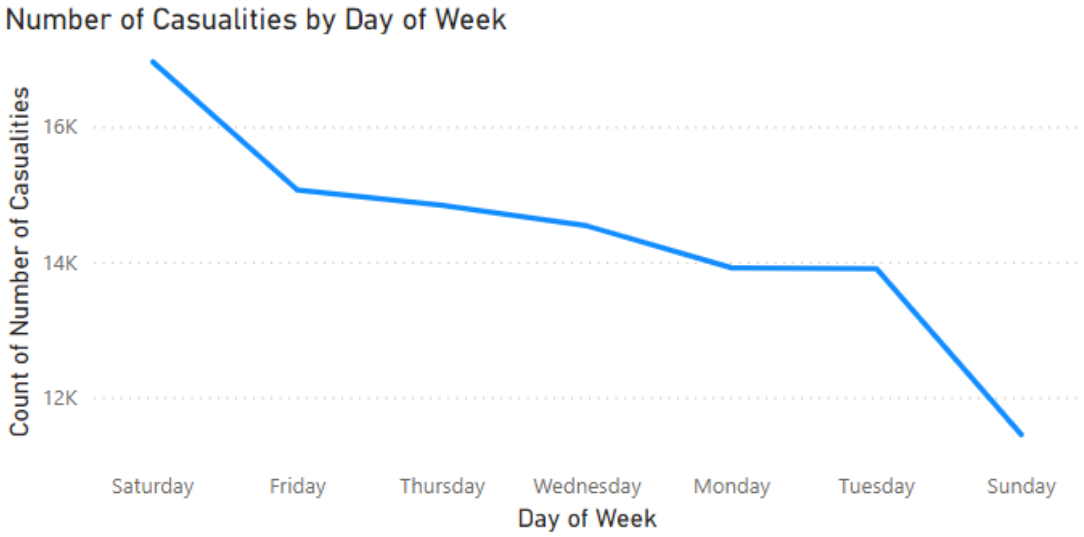

Figure 12: Line Chart (Number of Causalities by Day of Week)

Discussion

The dashboard's interactive features of Power BI offer an extensive structure for deep analysis, allowing stakeholders to examine data in a user-friendly manner. This study showcases Power BI's outstanding features for generating interactive elements that may dramatically improve public safety measures (Bachechi, Po, & Rollo, 2022; Salehan & Kim, 2016).

Challenges

- **Data Integration**: Integrating real-time data sources while making sure that interactive visualisation features keep performing optimally.
- **User Interface Design**: It is very important to create a simple and clean dashboard user interface to make sure that the visualization is both informative and intuitive. It should accommodate a wide range of users from decision or policy makers to traffic planners.

Conclusion

Task 3 demonstrates Power BI's extensive features for generating engaging and informative dashboards. The extensive analysis and strategic planning enabled by this dashboard highlight how modern data visualization techniques may have a significant influence on road safety management and public safety measures.

Reference list

Ahmed, R., Shaheen, S. and Philbin, S.P. (2022). The Role of Big Data Analytics and decision-making in Achieving Project Success. *Journal of Engineering and Technology Management*, [online] 65, p.101697. doi:https://doi.org/10.1016/j.jengtecman.2022.101697.

Ahmed, S., Hossain, M.A., Bhuiyan, M.M.I. and Ray, S.K. (2021). *A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity*. [online] IEEE Xplore. doi:https://doi.org/10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069.

Bachechi, C., Po, L. and Rollo, F. (2022). Big Data Analytics and Visualization in Traffic Monitoring. *Big Data Research*, 27, p.100292. doi:https://doi.org/10.1016/j.bdr.2021.100292.

Salehan, M. and Kim, D.J. (2016). Predicting the Performance of Online Consumer reviews: a Sentiment Mining Approach to Big Data Analytics. *Decision Support Systems*, 81, pp.30–40. doi:https://doi.org/10.1016/j.dss.2015.10.006.