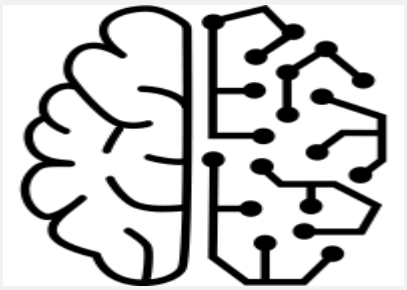


CA Assessment 1: Data Science & Machine Learning

Student name: Sam Cullen

Student number: C00250093

Lecturer name: Greg Doyle

Submission Dates:

Monday 2nd October	(Portfolio Specification)
Friday 27th October	(Advanced Portfolio Draft)
Friday 27th November	(Final Portfolio Submission)
Monday 4th December	(Presentation and demonstration)

Table of Contents

Table of Contents

A general timeline for building the portfolio website will follow:

- Documentation and brain-storming: 23/09/2023 – 02/10/2023
- Development and advanced portfolio draft: 02/10/2023 – 27/10/2023
- Further development and final submission: 27/10/2023 – 17/11/2023

1. Introduction

For my portfolio, I will be using a mix of tools and languages for designing my portfolio website. Using Visual Studio Code as the editor, and the extensions included helped make the general development process easy to track the progress in creating, saving, and committing the work.

For building the website, the technologies used most consist of HTML, CSS, and JavaScript for adding interactivity and functionality where necessary to improve the overall User Experience. To make building this portfolio easier, extensions such as Live Server which hosts a look-a-like website port give a preview of what the website would look like before deploying it. The Prettier extension helps keep the code neat and properly formatted, which makes it easier to understand the code.

Throughout the development of the website, There will be consistent commits to a GitHub repository which will help maintain a well-organised project. Each commit will include a descriptive commit message to document the changes added, or removed.

There will also be a README file in my GitHub repository. This will include information about the project, how to run it locally, and any other relevant details that might be useful to visitors.

2. Portfolio Specification

The general basis of the design and layout of my Portfolio will be:

2.1 Design

The primary color scheme chosen for this portfolio website is a simple and effective black-and-white contrast. This approach to design is not only visually appealing for the viewers but also serves a practical purpose and simplicity.

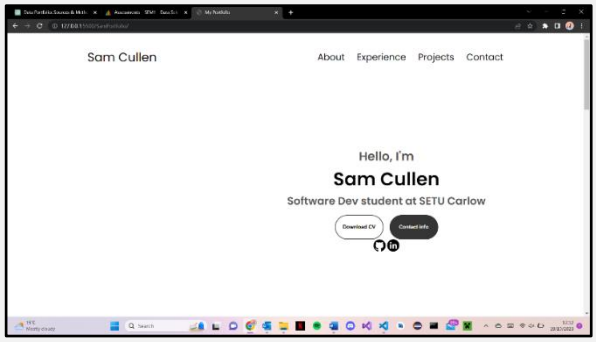
White is the dominant background color of the website, giving the website a clean look, it serves as a backdrop for text, images, and other elements in the website ensuring that the content is easily accessible.

The use of black in the design in contrast to the white backdrop provides depth and contrast, creating a visually appealing hierarchy to the page. It is used to draw attention to key elements such as the navigation bar, headings, and text.

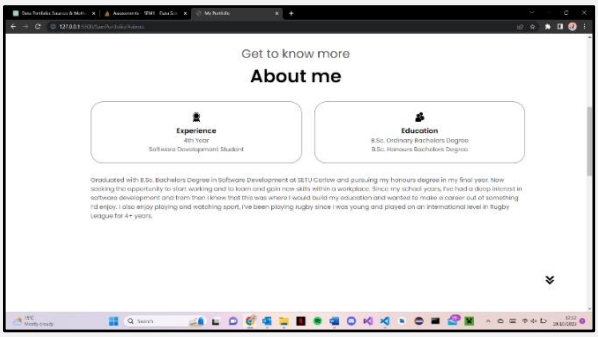
The balance between black and white is carefully used to not have too much of one or the other and to ensure both are appealing to each other throughout the entire webpage.

2.2 Layout

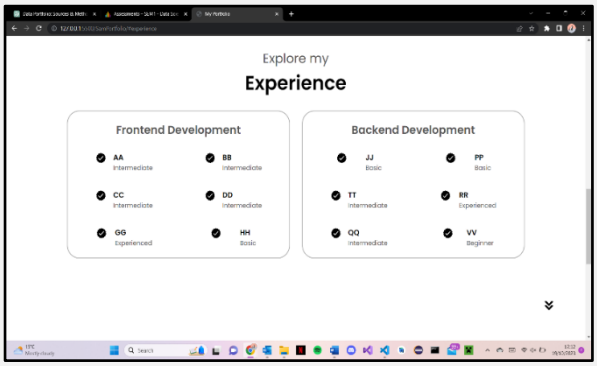
- Front page: Will have my name, short text about me, links to my GitHub, and LinkedIn, and the option to view and download my CV all available.
- Navigation bar: At the top right corner of the first page, there will be options to traverse through my Portfolio easier than scrolling. Options such as 'About' Experience', 'Project', and 'Contact'. Depending on the device the portfolio is being viewed on, a hamburger menu will prompt to keep the page from becoming crowded if viewed on a smaller device.



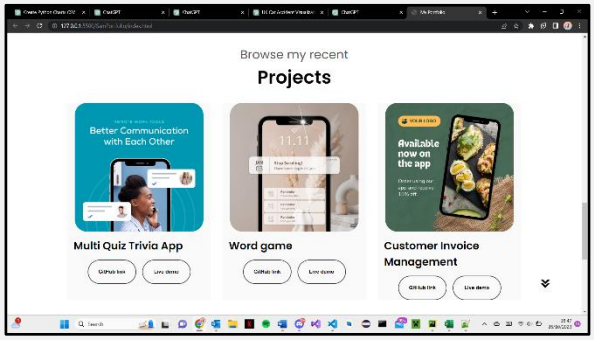
- The About Me section will offer additional insights into my background and achievements, including a brief description of my education.



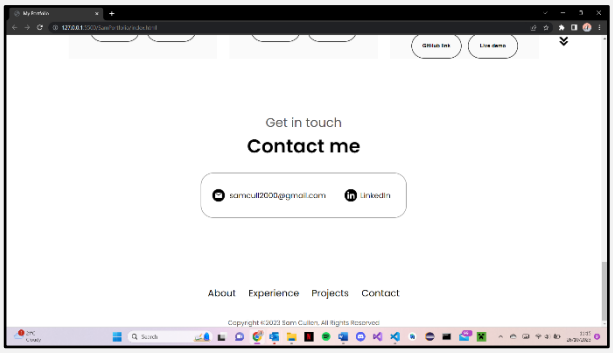
- The Experience section will include two different tables:
  - First table showcasing frontend languages learned throughout college and in personal time,
  - The other table demonstrates the backend languages, tools, IDEs, and other skills acquired in college and in personal time.



- A section on Projects will be available to show off projects during the 4 years in college, each project shown will have:
  - GitHub link attached for exploration
  - A short description of how each project works and the technologies used.



- Finally, a section on Contact Us will include
- My Email, LinkedIn, and GitHub will once more be featured at the bottom of the website.
- Finally, the portfolio website will be deployed on Netlify publicly and the URL will be available in my CV for reference.

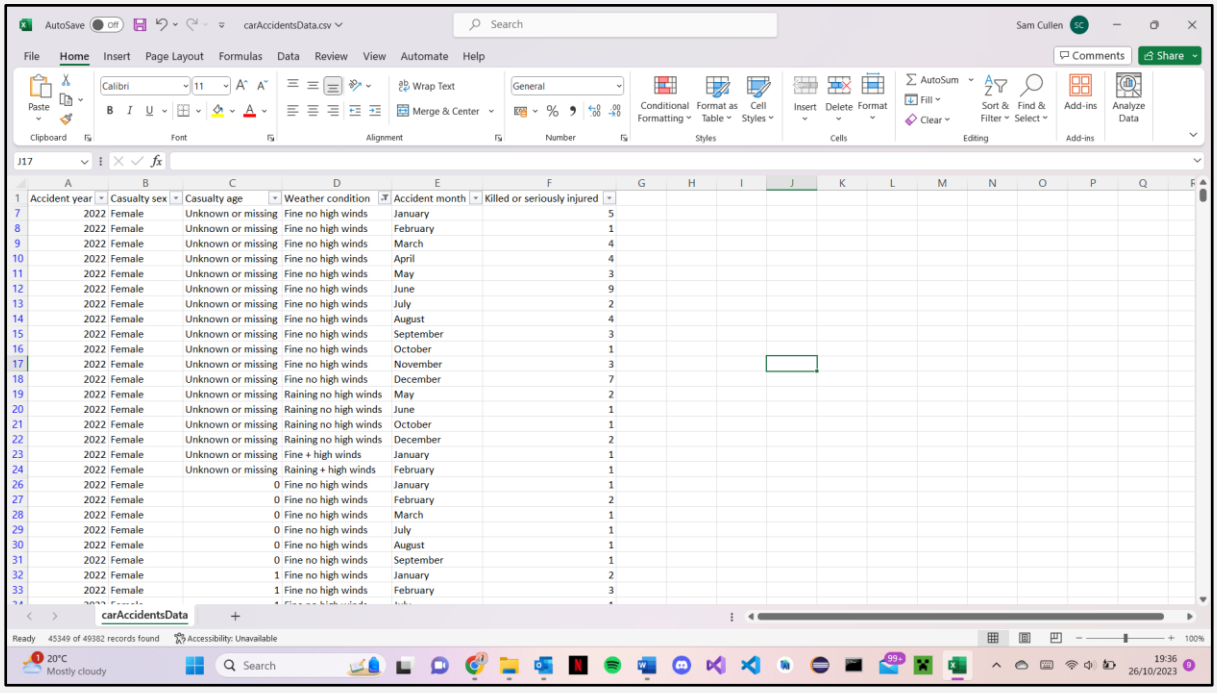


3. Machine Learning for Car Accidents in the UK

Car accidents are a significant public safety concern in the United Kingdom, leading to injuries, property damage, and fatalities. To proactively address this issue and enhance road safety, using Machine Learning, we can make future predictions and estimate roughly how many accidents could occur in the future by studying accidents from 2013-2022 and using linear regression and graphs to give a better sight as to how the statistics increase or decrease over the years. This document outlines the steps involved in using a dataset of historical car accidents in the UK to develop predictive models that can forecast future accidents, ultimately contributing to better road safety measures. While predictions cannot provide an exact count, machine learning can improve the ability to anticipate accidents by providing valuable information based on real-time data

The variables going to be used in this assessment are

- Accident year
- Casualty sex
- Casualty age
- Weather conditions
- Accident month
- Kills or seriously injured



	A	B	C	D	E	F
1	Accident year	Casualty sex	Casualty age	Weather condition	Accident month	Killed or seriously injured
7	2022	Female	Unknown or missing	Fine no high winds	January	5
8	2022	Female	Unknown or missing	Fine no high winds	February	1
9	2022	Female	Unknown or missing	Fine no high winds	March	4
10	2022	Female	Unknown or missing	Fine no high winds	April	4
11	2022	Female	Unknown or missing	Fine no high winds	May	3
12	2022	Female	Unknown or missing	Fine no high winds	June	9
13	2022	Female	Unknown or missing	Fine no high winds	July	2
14	2022	Female	Unknown or missing	Fine no high winds	August	4
15	2022	Female	Unknown or missing	Fine no high winds	September	3
16	2022	Female	Unknown or missing	Fine no high winds	October	1
17	2022	Female	Unknown or missing	Fine no high winds	November	3
18	2022	Female	Unknown or missing	Fine no high winds	December	7
19	2022	Female	Unknown or missing	Raining no high winds	May	2
20	2022	Female	Unknown or missing	Raining no high winds	June	1
21	2022	Female	Unknown or missing	Raining no high winds	October	1
22	2022	Female	Unknown or missing	Raining no high winds	December	2
23	2022	Female	Unknown or missing	Fine + high winds	January	1
24	2022	Female	Unknown or missing	Raining + high winds	February	1
26	2022	Female	0	Fine no high winds	January	1
27	2022	Female	0	Fine no high winds	February	2
28	2022	Female	0	Fine no high winds	March	1
29	2022	Female	0	Fine no high winds	July	1
30	2022	Female	0	Fine no high winds	August	1
31	2022	Female	0	Fine no high winds	September	1
32	2022	Female	1	Fine no high winds	January	2
33	2022	Female	1	Fine no high winds	February	3

4. Data Sources and Collection Methods

A range of data sources and collection methods

4.1 Data Sources

Public Datasets: Publicly available datasets such as Kaggle, data.gov.ie, github.com, datahub.io, and more were very beneficial in providing multiple diverse data available to compare and use.

Government Datasets: Many governments release open datasets on a variety of topics like healthcare, finance, and much more. These datasets proved valuable resources while looking for data to compare with other datasets and get an idea of the accuracy and quality of the data represented

4.2 Collection/Storage Methods

Data Entry: In cases where extracting the data automatically is not possible, Microsoft Excel spreadsheets (CSV files) or databases like MySQL, SQLite, MongoDB, and more can be used to enter the data into a structured format. In this specification, Microsoft Excel will be used to store, clean, and join data from various spreadsheets

PyCharm was the IDE used for creating charts and graphs of the data to provide a visual representation, making it easier to grasp the information more effectively.

4.3 Data Cleaning

Before working with the data, various steps were carried out to ensure data quality and reliability:

4.3.1 Handling Missing Data

The first step carried out before taking any dataset was addressing any missing data, using techniques like statistical analysis to help find missing data and remove any unfinished rows to keep the quality of the data. The removal of missing data was done on Microsoft Excel by selecting * rows and columns < special < blanks.

4.3.2 Duplicate Data Removal

Duplicate values can affect analysis and lead to misleading results, finding duplicates and removing them. Using SQL queries to find and manage duplicates using the statements ‘SELECT’, ‘DISTINCT’, and ‘GROUP BY ’ to keep a clean dataset. Same as before, we removed any duplicate values on Microsoft Excel by selecting * rows and columns < Data < Data tools < Remove duplicates

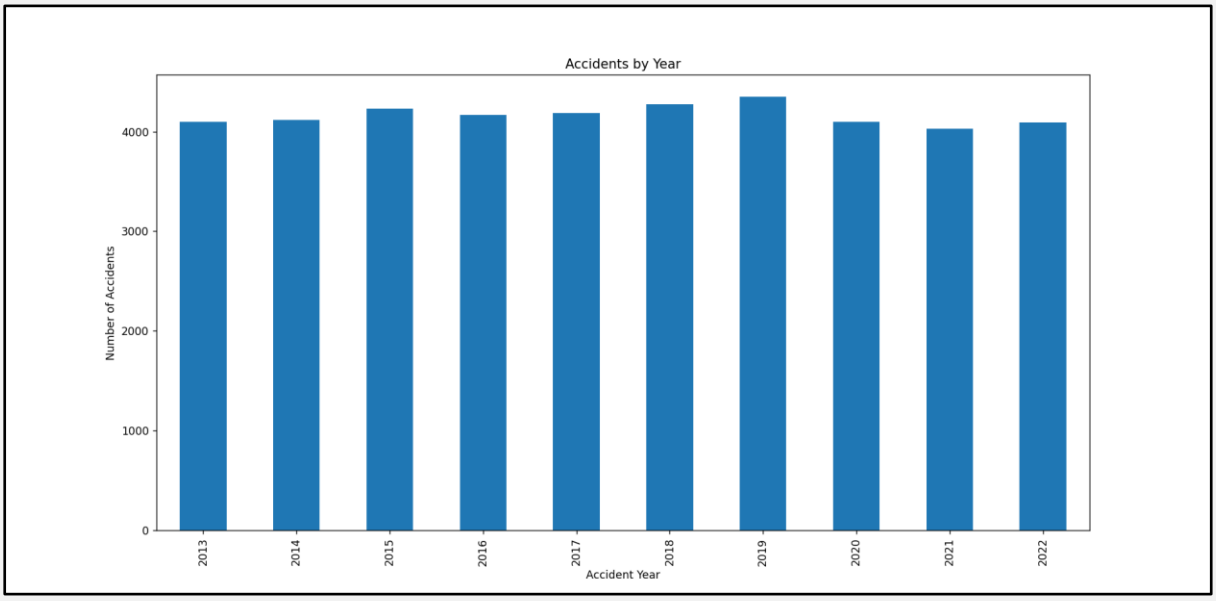
4.3.3 Documentation and Version Control

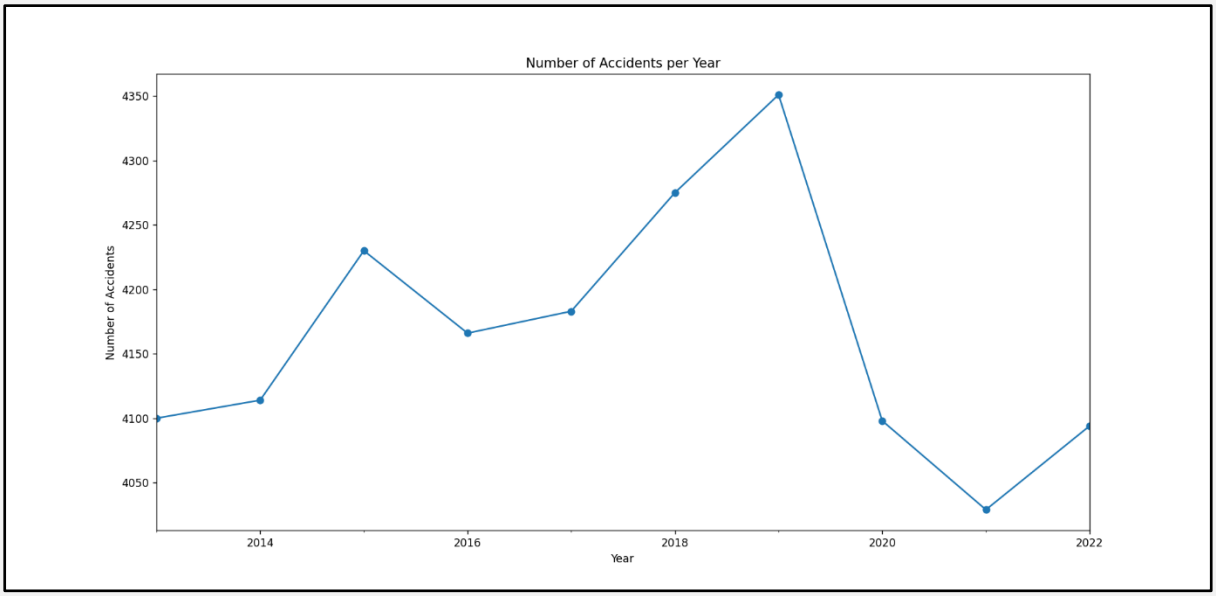
Throughout the data cleaning process, a constant record of the progress is being documented to record the steps taken and changes made, this was crucial in the development of maintaining a productive and high-quality dataset. Version controls like GitHub helped keep track of the progress for saving progress, or going back a step if coming across an issue

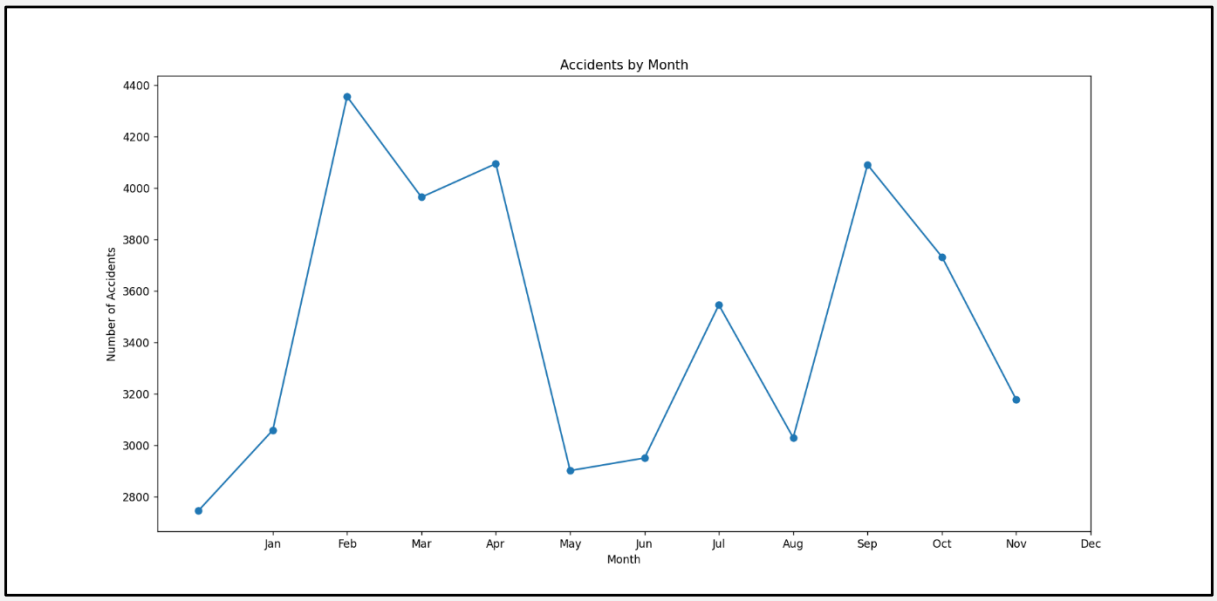
5. Data Graphs

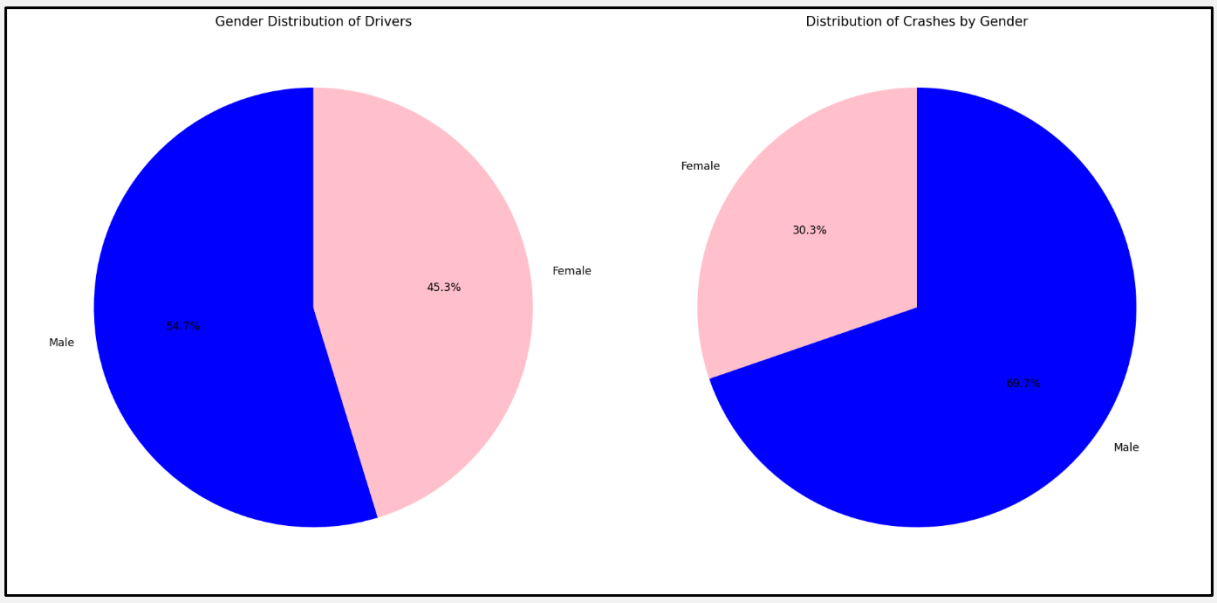
This section represents a comprehensive analysis of car accidents in the UK from 2013 - 2022. To shed light on the dynamics and trends within this dataset, a variety of graphs and charts have been used to illustrate the data, these visual representations not only provide a clear overview of the data but also allow for the identification of possible patterns, insight, and future predictions.

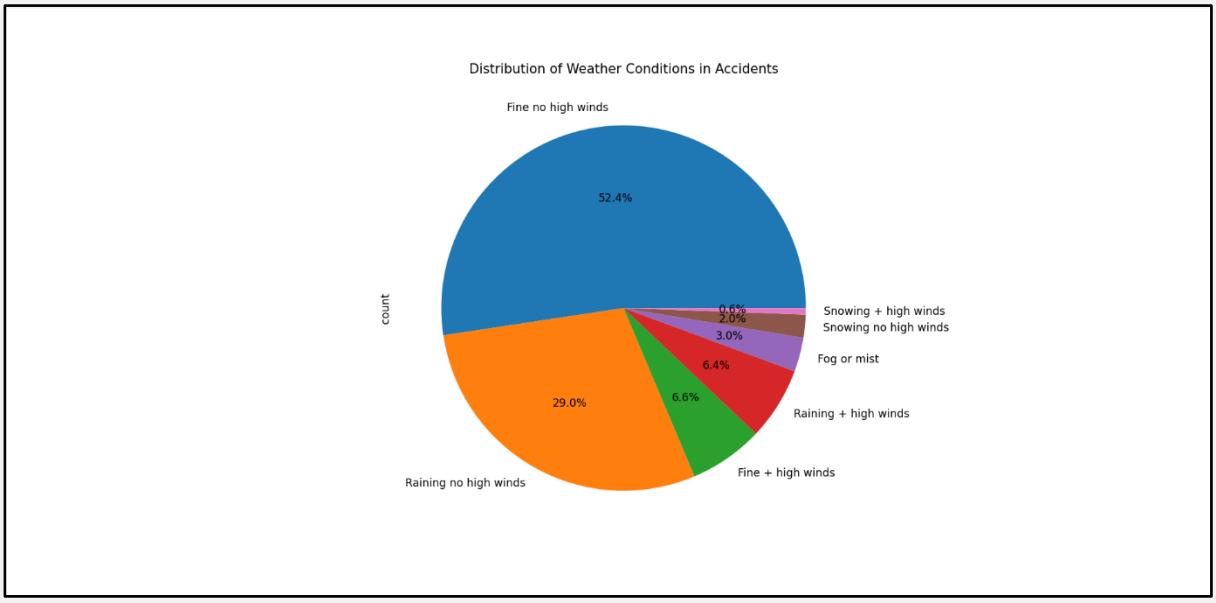
A breakdown of accident types – such as weather conditions, casualty age, casualty sex, accidents by year, and accidents by month is shown through pie charts, bar graphs, and stacked column charts. These representations offer a deeper understanding of the underlying causes and contributing factors to car accidents and allow us to discern any significant shifts in accident occurrences.

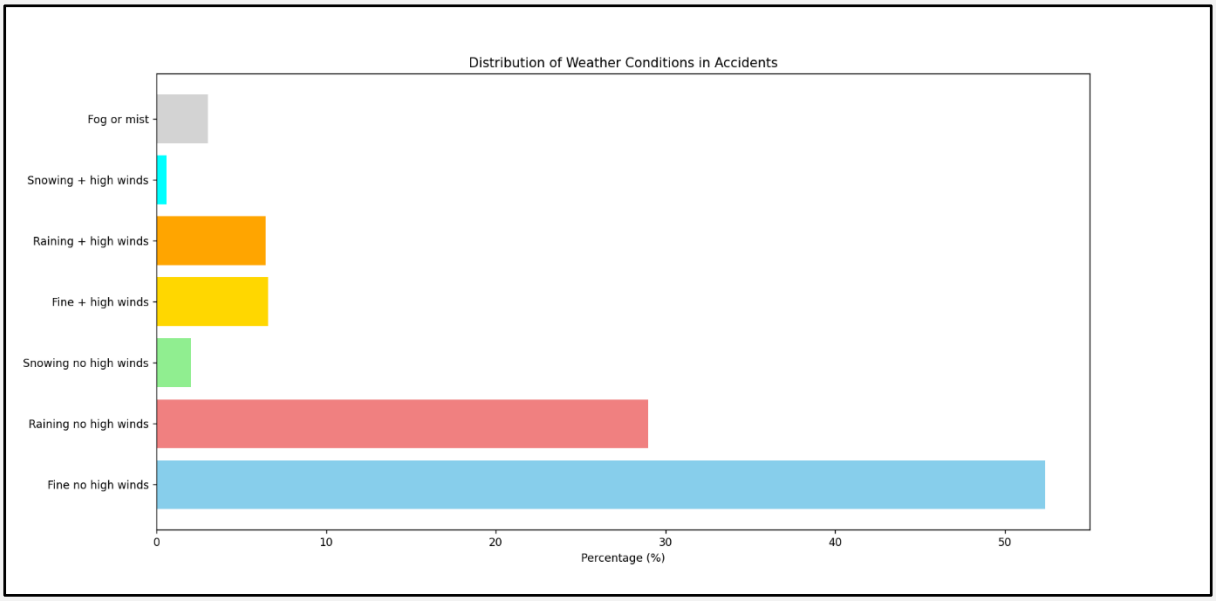


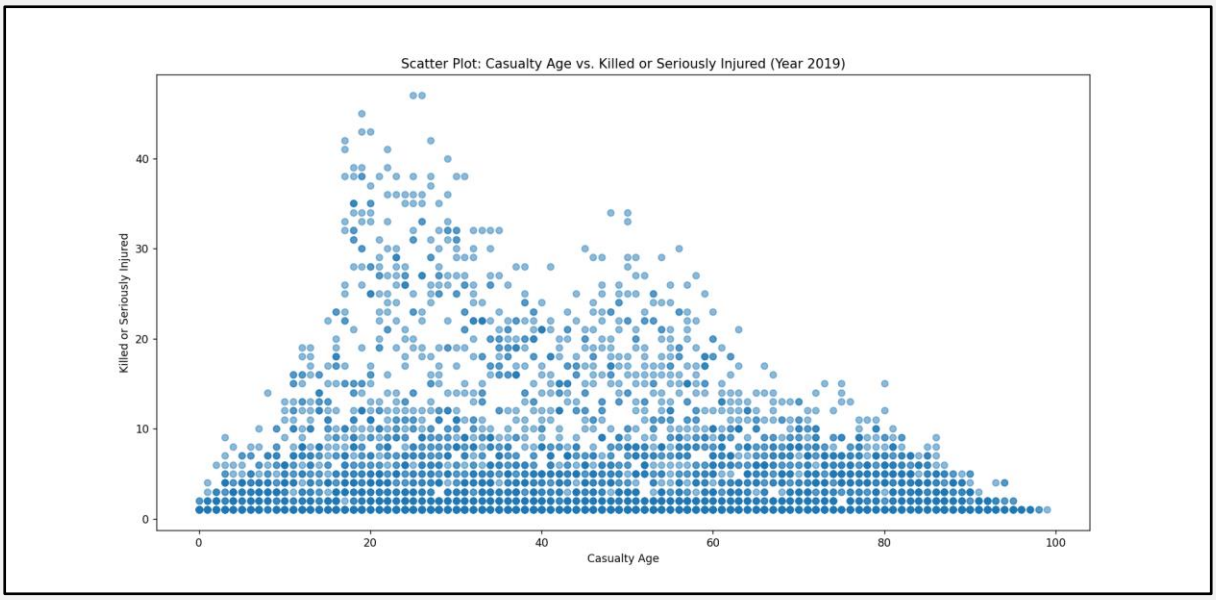


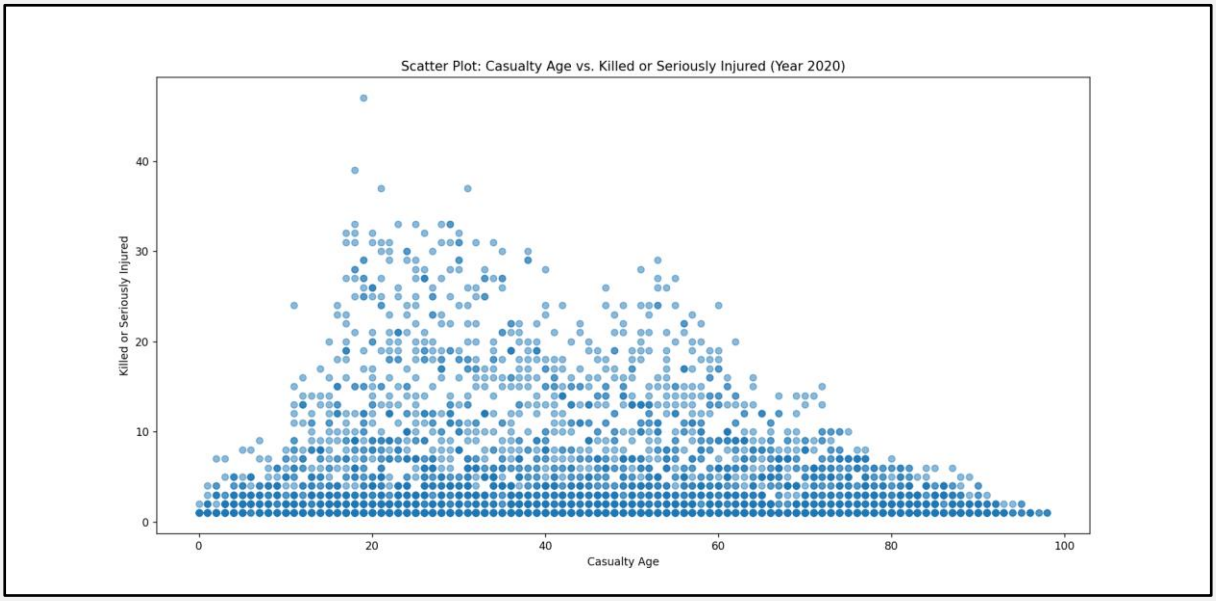

Accidents by Month
Number of Accidents
4400
4200
4000
3800
3600
3400
3200
3000
2800
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

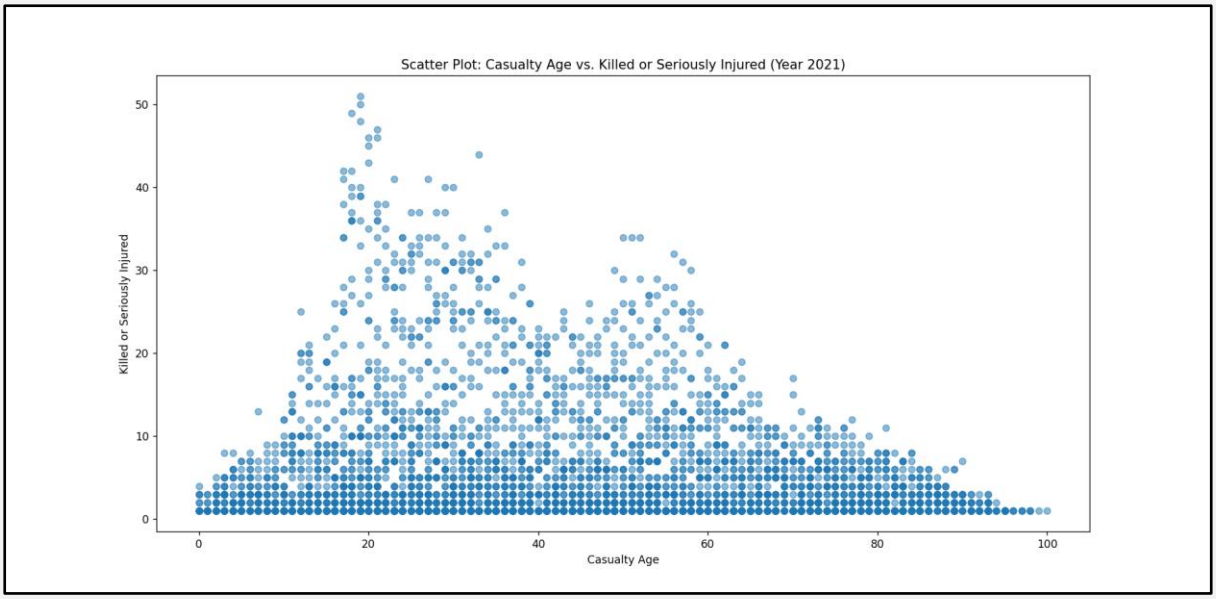

Gender Distribution of Drivers
45.3%
Female
54.7%
Male
Distribution of Crashes by Gender
Female
30.3%
69.7%
Male

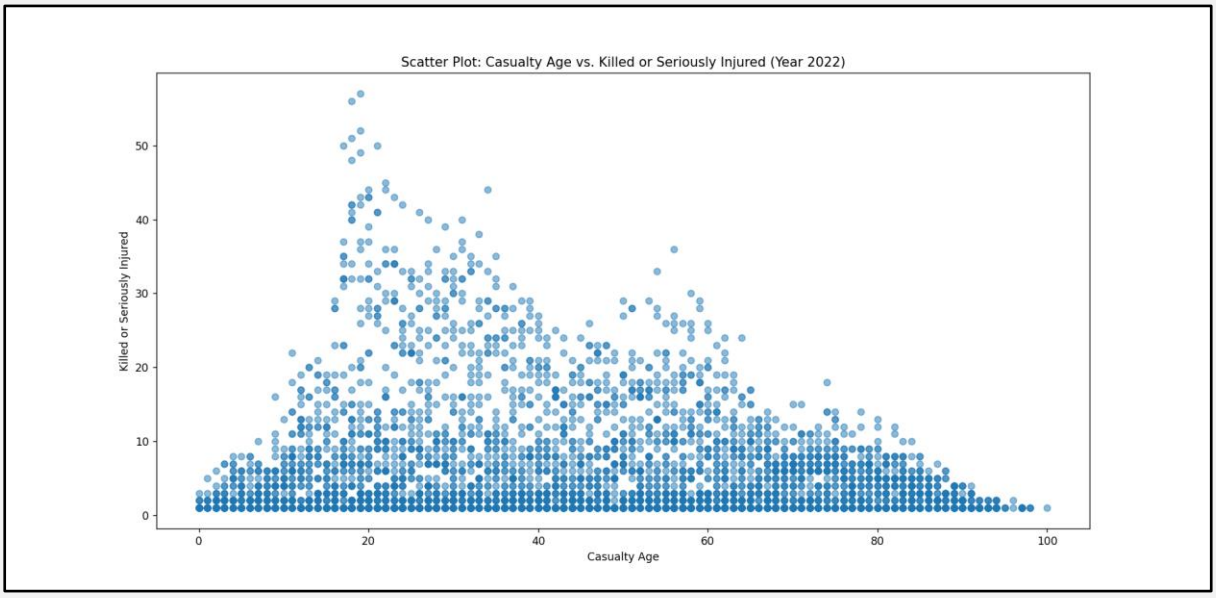

Distribution of Weather Conditions in Accidents
Fine no high winds
52.4%
count
0.6%
2.0%
3.0%
6.4%
6.6%
29.0%
Snowing + high winds
Snowing no high winds
Fog or mist
Raining + high winds
Fine + high winds
Raining no high winds

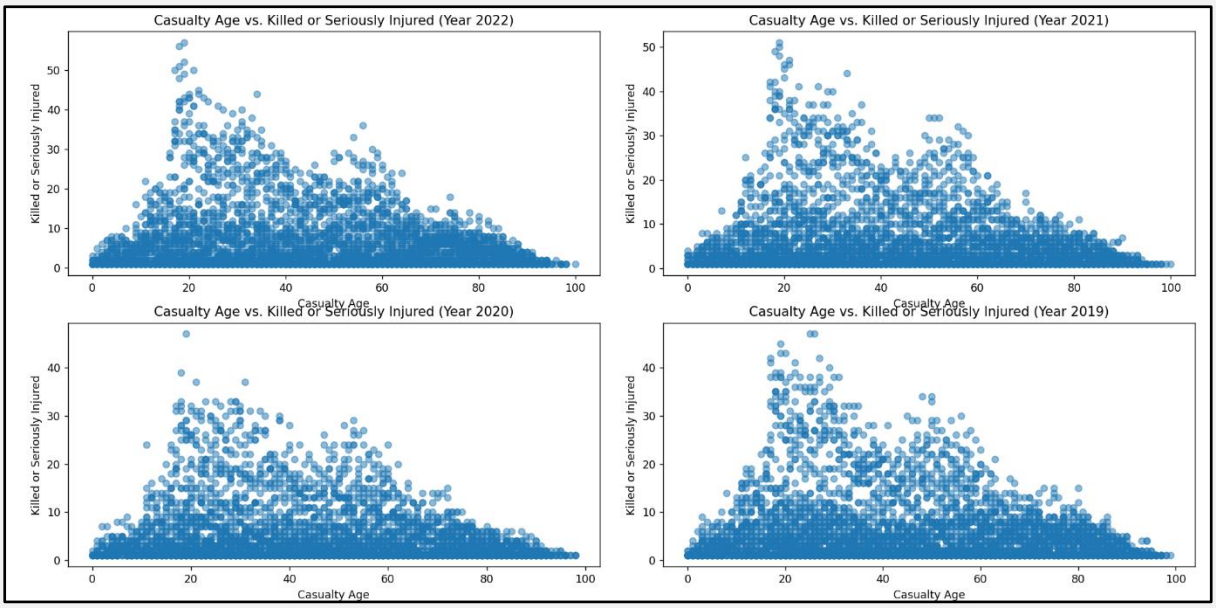

Distribution of Weather Conditions in Accidents
Fog or mist
Snowing + high winds
Raining + high winds
Fine + high winds
Snowing no high winds
Raining no high winds
Fine no high winds
0
10
20
30
40
50
Percentage (%)

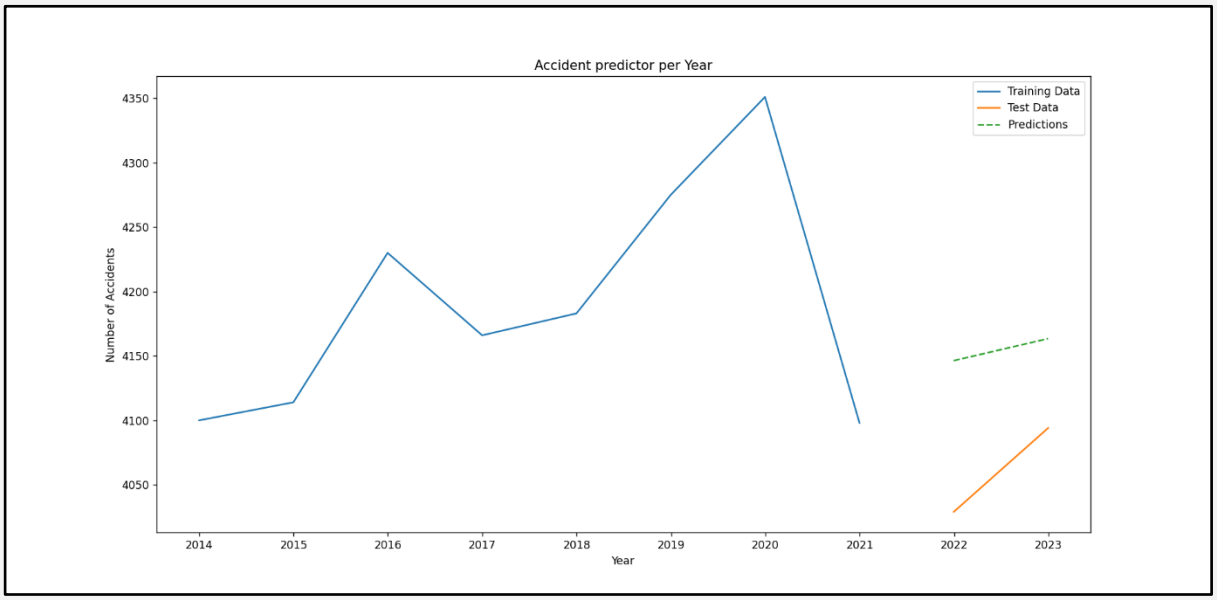

Scatter Plot: Casualty Age vs. Killed or Seriously Injured (Year 2019)
Killed or Seriously Injured
Casualty Age
0
10
20
30
40
0
20
40
60
80
100

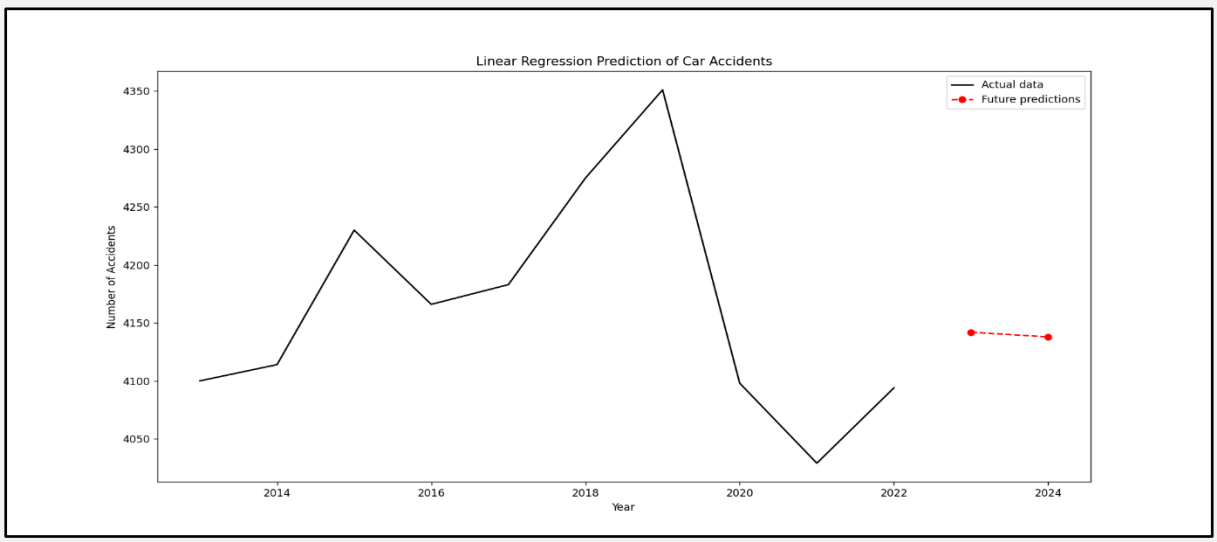

Scatter Plot: Casualty Age vs. Killed or Seriously Injured (Year 2020)
Killed or Seriously Injured
Casualty Age
0
10
20
30
40
0
20
40
60
80
100

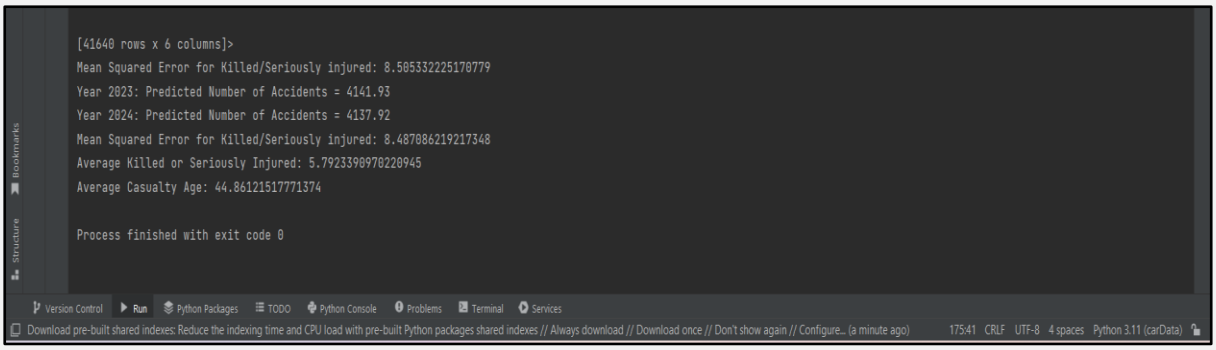

Scatter Plot: Casualty Age vs. Killed or Seriously Injured (Year 2021)
Killed or Seriously Injured
Casualty Age
0
10
20
30
40
50
0
20
40
60
80
100


Scatter Plot: Casualty Age vs. Killed or Seriously Injured (Year 2022)
Killed or Seriously Injured
Casualty Age
0
10
20
30
40
50
0
20
40
60
80
100


Casualty Age vs. Killed or Seriously Injured (Year 2022)
Casualty Age vs. Killed or Seriously Injured (Year 2021)
Casualty Age vs. Killed or Seriously Injured (Year 2020)
Casualty Age vs. Killed or Seriously Injured (Year 2019)
Killed or Seriously Injured
Casualty Age


Accident predictor per Year
Training Data
Test Data
Predictions
Number of Accidents
Year


Linear Regression Prediction of Car Accidents
Actual data
Future predictions
Number of Accidents
Year


[41640 rows x 6 columns]>
Mean Squared Error for Killed/Seriously injured: 8.505332225170779
Year 2023: Predicted Number of Accidents = 4141.93
Year 2024: Predicted Number of Accidents = 4137.92
Mean Squared Error for Killed/Seriously injured: 8.487086219217348
Average Killed or Seriously Injured: 5.7923390970220945
Average Casualty Age: 44.86121517771374
Process finished with exit code 0
Bookmarks
Structure
Version Control
Run
Python Packages
TODO
Python Console
Problems
Terminal
Services
Download pre-built shared indexes: Reduce the indexing time and CPU load with pre-built Python packages shared indexes // Always download // Download once // Don't show again // Configure... (a minute ago)
175:41
CRLF
UTF-8
4 spaces
Python 3.11 (carData)

6. Data Exploring and Analysis

6.1 Exploratory Data Analysis

The first part of the investigation involves using well-known Python tools like Pandas and Matplotlib to perform exploratory data analysis. Several bar graphs are produced as a result of this technique to graphically depict the frequency of accidents over two different temporal dimensions: annually and monthly. Insights into the temporal distribution of accidents are provided by these visualisations, which facilitate the recognition of trends and patterns.

6.2 Descriptive Statistics

To find patterns related to gender in the dataset, statistical techniques are used. Pie charts are a useful tool for clearly displaying the proportion of male and female drivers, providing a visual breakdown of demographics by gender. To provide a more sophisticated picture of the role gender plays in accidents, another pie chart shows the distribution of collisions by gender.

6.3 Weather condition analysis

The code investigates the distribution of accidents under various weather conditions using a combination of pie charts and bar charts. The identification of possible associations between weather patterns and accident incidents is made possible by this analysis, which also offers practical advice for putting safety precautions into place.

6.4 Scatter plots

To visually examine the association between the age of the casualties and the number of fatalities or major injuries, scatter plots are created. These scatter plots provide a thorough analysis of potential age-related risk factors that may affect accident outcomes over time. They span four years, from 2019 to 2022.

6.5 Time Series Forecasting (ARIMA)

The ARIMA (AutoRegressive Integrated Moving Average) model is used to forecast time series in this investigation. The model is used to forecast the number of accidents per year after the data is split into training and testing sets. Mean Squared Error (MSE) is used to evaluate performance, and the results are displayed to show the predicted trends.

6.6 Linear Regression

To forecast the number of accidents in 2023 and 2024, linear regression is used. The model is trained using resampled annual data, and a line plot that displays the predictions next to the real data provides an easy-to-understand illustration of the model's predictive power.

6.7 Calculating Averages

The last steps of the script involve calculating and presenting the average values for two important variables: "Killed or seriously injured" and "Casualty age." several averages offer summary statistics that give a succinct rundown of several important dataset components.

7. References

Responsive Web development tips

https://www.youtube.com/watch?v=vZB1s8J6dhY&ab_channel=Raddy

HTML CSS Nav bar & Hamburger tutorial

https://www.youtube.com/watch?v=Ey-slKj8xA8&ab_channel=Onlinewebustaad

What is Machine Learning?

https://www.youtube.com/watch?v=ukzFI9rgwfU&ab_channel=Simplilearn

Tips to build an ML portfolio

https://www.springboard.com/blog/data-science/machine-learning-portfolio/

Python ML tutorial (helped show how to do it)

https://www.youtube.com/watch?v=7eh4d6sabA0&ab_channel=ProgrammingwithMosh

PyCharm download

https://www.jetbrains.com/pycharm/download/?section=windows

For finding data sets

https://www.kaggle.com/

https://data.gov.ie/

https://www.github.com

https://datahub.io/

https://www.ml4devs.com/articles/datasets-for-machine-learning-and-data-science/

https://www.tableau.com/

https://roadtraffic.dft.gov.uk/#6/55.254/-6.064/basemap-regions-countpoints

Machine Learning algorithms

https://blackboard.itcarlow.ie/bbcswebdav/pid-951930-dt-content-rid-5731704_1/xid-5731704_1

https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies

https://www.clickworker.com/customer-blog/the-6-types-of-machine-learning-algorithms-you-should-know/

https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.08-Random-Forests.ipynb

8. Collaboration Form

Student Name: Sam Cullen

Student No: C00250093

The following is a total list of all students I collaborated with whilst conducting research for this assignment:

Student Name: Sam Cullen

Student Number: C00250093

% of collaboration within submission: 100%