**Motor Vehicle Accident Predictor**

C964 - Capstone Project

Jerry Sandhu

April 5th, 2024

*\*Note: Each header corresponds to a bullet point in the rubric\**

[**✾A - Letter of transmittal**](#_ow282woezfsd) **3**

[A.1 - Problem summary](#_42zvsabu6ein) 4

[A.2 - Benefits to customer and the decision making process](#_a8cklujzp0vz) 5

[A.3 - Data product outline](#_8gm9jtky351v) 5

[A.4 - Data description](#_46fgpx7zxzgi) 6

[A.5 - Project objectives and hypotheses](#_e28duwsp0dtl) 6

[A.6 - Project methodology outline](#_6imdeiy3fe91) 6

[A.7 - Funding requirements](#_e89zrqqxnup) 7

[A.8 - Solution impact on stakeholders](#_mm4d3wyl0tg4) 7

[A.9 - Ethical and legal considerations](#_58pb3nzhvis7) 8

[A.10 - Personal expertise](#_mqzy838rbksr) 8

[**✾B - Executive summary**](#_vn5ti1bs68sw) **8**

[B.1 - Problem solution and opportunity](#_bc8hlisuw78s) 8

[B.2 - Description of customers](#_6gaxrprx8e62) 9

[B.3 - Gaps in the existing data product](#_odwvpl9atuom) 9

[B.4 - Data to collect and support product lifecycle](#_q8eku5dltfeh) 9

[B.5 - Methodology to support product development](#_out2wckum3u) 9

[B.6 - Deliverables](#_97c6uqy9mxd3) 10

[B.7 - Product implementation plan](#_q92n6q5yuhj0) 10

[B.8 - Methods to validate requirements](#_wy9bp1pygywb) 11

[B.9 - Programming environment](#_ppchdgbg6lpb) 12

[B.10 - Project timeline](#_acky3mhyshg2) 12

[**✾C - Program summary**](#_tmbpcnmlb6p2) **12**

[**✾D - Post implementation report**](#_3rky672j5s3u) **13**

[D.1 - Business vision](#_o1z4kctkw9nn) 13

[D.2 - Dataset](#_164h1pkvigof) 13

[D.3 - Code analysis](#_zbi7vwqsvad1) 14

[D.4 - Hypothesis assessment](#_yb3gpfw5xkum) 15

[D.5 - Data exploration support](#_uua8x9sg1vk7) 15

[D.6 - Product accuracy assessment](#_aweb23qnuzm7) 16

[D.7 - Results](#_j1se7kvxut2t) 17

[D.8 - Source code](#_kcjw6y9gml98) 19

[D.9 - Quick start guide](#_it1sj2mugpiv) 20

[**✾E - Sources**](#_59ndbo7xfbhq) **21**

## ✾A - Letter of transmittal

John Johnson, CEO

Kord Motors

Dear John Johnson,

My name is Jerry Sandhu, a recent graduate in Computer Science from Western Governors University. Throughout my academic journey and internships, I have obtained a large skill set in programming, particularly in the field of machine learning.

As the automobile industry moves towards automated driving, it’s of great importance to use the tools we have to introduce measures to increase safety. As the prominent figure in the leadership of Kord, there exists an expectation for you and your company to spearhead initiatives that prioritize safety for all individuals using your vehicles.

I would like to present a proposal to implement machine learning into Kord's vehicle software, specifically targeting the identification of events that may cause accidents. This proposed program will leverage datasets pertaining to accidents, employing machine learning algorithms to find key factors contributing to such incidents. The insights found from this analysis will be integrated into Kord's autonomous driving systems, therefore enhancing the safety protocols for all road users.

This solution will be implemented within a Jupyter Notebook, ensuring accessibility and ease of use for Kord's software engineers. Moreover, it will be compatible with any standard web browser, facilitating effortless setup and data manipulation.

With my seven years of experience in software development, along with a Bachelor's degree in Computer Science, I am committed to delivering a solution that works with Kord's vision of safer and more advanced automotive technologies.

I look forward to the opportunity of speaking with you to discuss this proposal in further detail.

Sincerely,

Jerry Sandhu

Software Engineer

### A.1 - Problem summary

Increasing road safety and lowering the occurrence of motor vehicle accidents, and so minimizing the injuries and fatalities that come with it, stands as the most significant priority within the autonomous driving industry. The importance of implementing a strong system to increase safety measures for all road users is above all else.

To address this pressing concern, a prediction making program leveraging machine learning and historical accident data is the best solution. While beginning such a program means initial investments in labor, software, and hardware resources for Kord, the long term benefits are significant. Over time, the expenses incurred will be offset by the reduction in accidents and their associated costs with the public.

### A.2 - Benefits to customer and the decision making process

This data product will provide a significant advantage of decreasing accidents involving Kord vehicles, therefore creating a corporate image centered on consumer safety as of the utmost importance. By leveraging insights found as a result of this program, Kord will have the tools to make informed decisions regarding the ability of their autonomous driving systems and vehicle designs. As the dataset of past accident factors continues to increase, Kord's capabilities in accident prevention and safety measures will also improve.

Kord can also capitalize on this program to diversify its market potential by offering accident reduction technology to companies involved in manufacturing electric scooters and boats. This expansion into new markets not only shows Kord's commitment to safety but also shows the public that Kord will be an industry leader in innovative safety problem solving beyond the automotive industry.

### 

### 

### A.3 - Data product outline

The data product will be developed in Python 3.7 with PyCharm, inside a Jupyter Notebook environment, and run as a program online through Google Colab. Its primary function will involve taking a dataset of past vehicle accidents and applying machine learning methods to find underlying patterns and factors contributing to accidents.

Python along with Jupyter Notebook, was selected as the development platform for its accessibility and popularity within the data analysis and machine learning fields. Along with the user-friendly nature of Python and versatile capabilities of Jupyter Notebook, this makes them the most ideal platforms for this project.

### A.4 - Data description

The program will utilize data sourced from Kaggle, containing detailed data containing the date, week, severity, location, light conditions, speed, and number of vehicles involved in accidents across Europe from 2020 to 2021.

After the program is successfully funded, the program's scope will be increased to have data from many different countries and various Kord vehicle types. This increase in datasets will contain an analysis aimed at finding discrepancies in how external factors influence accidents involving different Kord vehicle models, such as cars and trucks.

### A.5 - Project objectives and hypotheses

The primary goal of this project is to develop a user friendly program that uses machine learning to connect incident factors with actual accidents. Software Engineers will have the flexibility to run the program directly from any web browser through Google Colab. Additionally, the program will be open source, allowing for ongoing collaboration for continuous improvement of the product.

Although the current dataset used for the program is limited in scope, there are plans to expand it in the future to include data from all regions and vehicle types. This expansion will enhance the program's effectiveness and its use in accident analysis and prevention efforts.

### A.6 - Project methodology outline

The best methodology to develop this project is the waterfall method. This will ensure that the project is developed in a linear manner with a step by step process, where tasks are completed in a certain order. The order for this program follows the standard waterfall phases:

a) Requirements

b) Design

c) Implementation

d) Verification

e) Deployment

f) Maintenance

The requirements document will be formed with the help of software engineers, with a focus on security and accuracy. The project will be designed based on those requirements. The project will then be implemented into a live environment to test and ensure it’s working to a high standard. Then once the program is fully functioning and deployed, any further issues will have fixes implemented and documented.

### A.7 - Funding requirements

To fund the development and implementation of the data product, the estimated costs are below:

Program: $30,000

Computers: $8,000

Employee labor: $200,000

This brings the total funding requirement to $238,000. Additionally, the annual maintenance cost for the program is projected to be $15,000. It's essential to ensure enough funding to support the launch and ongoing maintenance of the program, which will increase its effectiveness in enhancing safety and reducing accidents.

### 

### A.8 - Solution impact on stakeholders

As a publicly traded company, Kord has investments from large firms like AutoBank and Ace International. The implementation of this product holds a lot of promise for Kord, offering opportunities to increase the value of their assets and look into new industries equipped with enhanced autonomous systems. This move not only has the potential to increase share values, but also increase job security for all stakeholders associated with Kord.

By leveraging advanced safety technologies, Kord can solidify its position as an industry leader towards innovation and high standards. This not only benefits shareholders but also fosters an environment for sustained growth and development within the company.

### A.9 - Ethical and legal considerations

The program will be released as open source, being accessible for distribution and enhancement by others under the Creative Commons license. Data sourcing will be done through contractual agreements with insurance companies, ensuring that only relevant information directly related to vehicle accidents is included. The dataset will not contain any personally identifiable information such as race, age, or sex, which will also get rid of privacy concerns and reduce the need for strong data protection measures.

By adhering to the above principles, the program will promote transparency, collaboration, and innovation for all involved, while also prioritizing privacy and data security. This approach will create an atmosphere of improvement and advancement in accident prevention technologies, which will benefit society as a whole.

### A.10 - Personal expertise

With seven years of programming experience and a Bachelor's degree in Computer Science, along with a Microsoft Azure certificate in Artificial Intelligence, I am well equipped to lead the development of this project. My expertise is within various domains within AI and machine learning, which will allow me to see this project to its finish successfully. I am confident that my experience, education, and certifications qualify me to lead this project, ensuring its success and contributing to the advancement of autonomous driving technologies.

## ✾B - Executive summary

### B.1 - Problem solution and opportunity

The ability for autonomous vehicle systems to safely drive its passengers from one location to another is one of the biggest concerns in the automobile industry. Having a program to accurately predict accidents before they occur is of the utmost importance. There is an opportunity here for Kord to be the leader of this industry and show its stakeholders and the public that it's serious about passenger safety.

### B.2 - Description of customers

The primary customer is Kord. To use the program to predict accidents and use the calculated information in the autonomous systems in vehicles. Software Engineers will collect information such as weather, speed and distance from other vehicles as data fed into the machine learning program to help it learn more about what will cause an accident. As more data is collected, the algorithms used in the program will improve and become more accurate at predicting a future accident.

### B.3 - Gaps in the existing data product

Kords current system takes in past accident data, but does not implement machine learning. This means that the analysis of data isn’t getting more accurate and runs the risk of vehicles making mistakes when they’re in dangerous situations. The existing system also requires dedicated computers and software engineers to be onsite to implement updates to the code. The new solution will be open source and available for engineers to implement fixes from any web browser, provided that they have the correct credentials to access Kords servers. All of this ensures that the data product will provide Kord with more accurate data and faster updates to the program than the existing data product.

### B.4 - Data to collect and support product lifecycle

The data that needs to be collected is information on speed, weather conditions and fatality rates for accidents across the world. An open source database will be created where other automakers, insurance companies and websites such as Kaggle can contribute their data, which will be vetted for accuracy by Kord engineers. This data will be entered into the Jupyter Notebook of the program in python to display Matplotlib visualizations of how the data correlates to accident occurrences.

### B.5 - Methodology to support product development

This project will use the waterfall method of development, where each stage of development will follow a step by step process. Those steps are as follows:

a) **Requirements** - Meet with stakeholders and discuss the requirements for the project to be successful. This includes the costs, functionality of the program and signing of documents to proceed to the next steps.

b) **Design** - Create the design of the program's interface, needed libraries and how the program will be accessed by engineers.

c) **Implementation** - Programming the application itself, collecting and vetting data, training the machine learning model and evaluating the results it showcases.

d) **Verification** - In this step there will be multiple tests done. These include unit and integration testing to ensure that the program is working correctly and free of bugs.

e) **Deployment** - The program will be deployed in a live environment.

f) **Maintenance** - The program will be continued to be monitored for bugs, accurate data and documented for future upgrades.

### 

### B.6 - Deliverables

The deliverables for this project is a Jupyter Notebook that has the application code and collected data that is visualized. Also included is a user guide on how to use the program and a report on the findings from the program.

### 

### 

### B.7 - Product implementation plan

The plan to implement the product is as follows:

1. Download the python libraries, Anaconda Navigator, Google Chrome and PyCharm
2. Collect Data regarding accident factors
   1. Obtain data from Kaggle
   2. Allow other companies to submit vetted data
   3. Organize the data to be coded into the program
3. Processing of data
   1. Use the data to create graphs and charts
4. Create the machine learning model
   1. Code the program to add new data and use it to make new predictions
5. Tune the machine learning model
   1. Tune the accuracy of the model to be more accurate with its predictions
6. Implement the program in a closed environment
   1. Host the Jupyter Notebook on Google Colab
   2. Ensure that the notebook is accessible through different browsers
7. Testing for bugs and performance
   1. Run unit tests
   2. Run integration tests
8. Fully deploy the program in a live environment
9. Maintain the code and documentation
   1. Document upgrades and bug fixes
   2. Develop a user guide

### 

### B.8 - Methods to validate requirements

The method for validating that the product meets the requirements is a post implementation report. This document will allow for checking that each feature of the program from its results to its costs line up with the requirements that were discussed with stakeholders. Testing the results of the program's predictions in a live environment with real vehicles in different situations, with mannequins as passengers will also validate the results.

### B.9 - Programming environment

The programming environment will cost $3,000 and will consist of only a server. The program itself will be hosted on Google Colab and will not have any extra costs attached. The labor costs to complete each phase of development is projected to be $200,000. This cost includes meeting with stakeholders and designing, developing, testing, and updating the program along with its documentation.

### B.10 - Project timeline

The projected timeline for the project is as follows below. The milestones must be completed in descending order and handled by the company's software engineers.

| **Milestone/Deliverable** | **Hours** | **Start date** | **End date** |
| --- | --- | --- | --- |
| Approval of requirements | 8 | 01/01/2024 | 01/03/2024 |
| Program design | 20 | 1/4/2024 | 1/10/2024 |
| Hardware purchases | 2 | 1/11/2024 | 1/12/2024 |
| Program development | 100 | 1/13/2024 | 2/1/2024 |
| Data collection | 5 | 2/2/2024 | 2/5/2024 |
| Live testing | 20 | 2/6/2024 | 2/10/2024 |
| Bug testing | 20 | 2/11/2024 | 2/20/2024 |
| Program deployment | 20 | 2/21/2024 | 3/1/2024 |

## ✾C - Program summary

See attached Python file

## ✾D - Post implementation report

### D.1 - Business vision

Vehicle accidents occur everyday in every country and in every type of vehicle. The push towards autonomous driving takes away the control from drivers and puts it in the hands of the vehicle's computer systems. Leading to an increased need to automatically operate the vehicle in a safe manner to reduce accidents.

The solution to this problem was to implement machine learning in the vehicle's computer systems. Using data sets from previous accidents to predict things that are likely to cause future accidents. The goal was to have an accuracy of 99% when determining that given factors such as weather and road conditions will contribute to an accident. Because the program is open source and can be accessed from any browsers, this means that there is more data for the program to learn from and be easier to access for software engineers.

### 

### 

### 

### 

### 

### 

### D.2 - Dataset

The source of the raw dataset was obtained from Kaggle.

<https://www.kaggle.com/datasets/nextmillionaire/car-accident-dataset/data>

The data was collected by a user on Kaggle named Saher Muhamed who obtained information regarding road accidents in both Kensington and Chelsea that occurred during January 2021 (Muhamed, 2024). The database is hosted on my Github account and accessed remotely by using a CSV reader on the data url in the application. The dataset was further manipulated using Pandas libraries with machine learning.



### 

### D.3 - Code analysis

*Descriptive method:*

The main code has a database graph where all the data regarding dates and accident causes are available to view in an unanalyzed format. This information can be further altered to get datasets that are in line with each other and to remove outliers. This data is read by using a CSV read function on a CSV database I have hosted on my Github page. This method was used because of its ease of use in viewing and modifying data.

*Predictive method:*

The application takes in information from the database using a CSV reader. It also uses Pandas to manipulate data and Seaborn with Matplotlib libraries to display the analyzed data into graphs. Once the data is in the table, all that needs to be done is running each or all of the cells to view visual models of the data. There is also an interactive slider in the program that allows user input of accident factors to determine the accident rate. This method was used because of its simplicity in analyzing data and viewing the results.

### D.4 - Hypothesis assessment

The hypothesis of this project was to collect information about past vehicle accidents and use that information to predict future accidents. The developed program showed promise, but did not have enough data to conclusively predict accidents. As it stands, I assess a rejection of the hypothesis until more data is collected. Once a significant amount of data on worldwide vehicle accidents is collected, the hypothesis can be revisited.

### 

### 

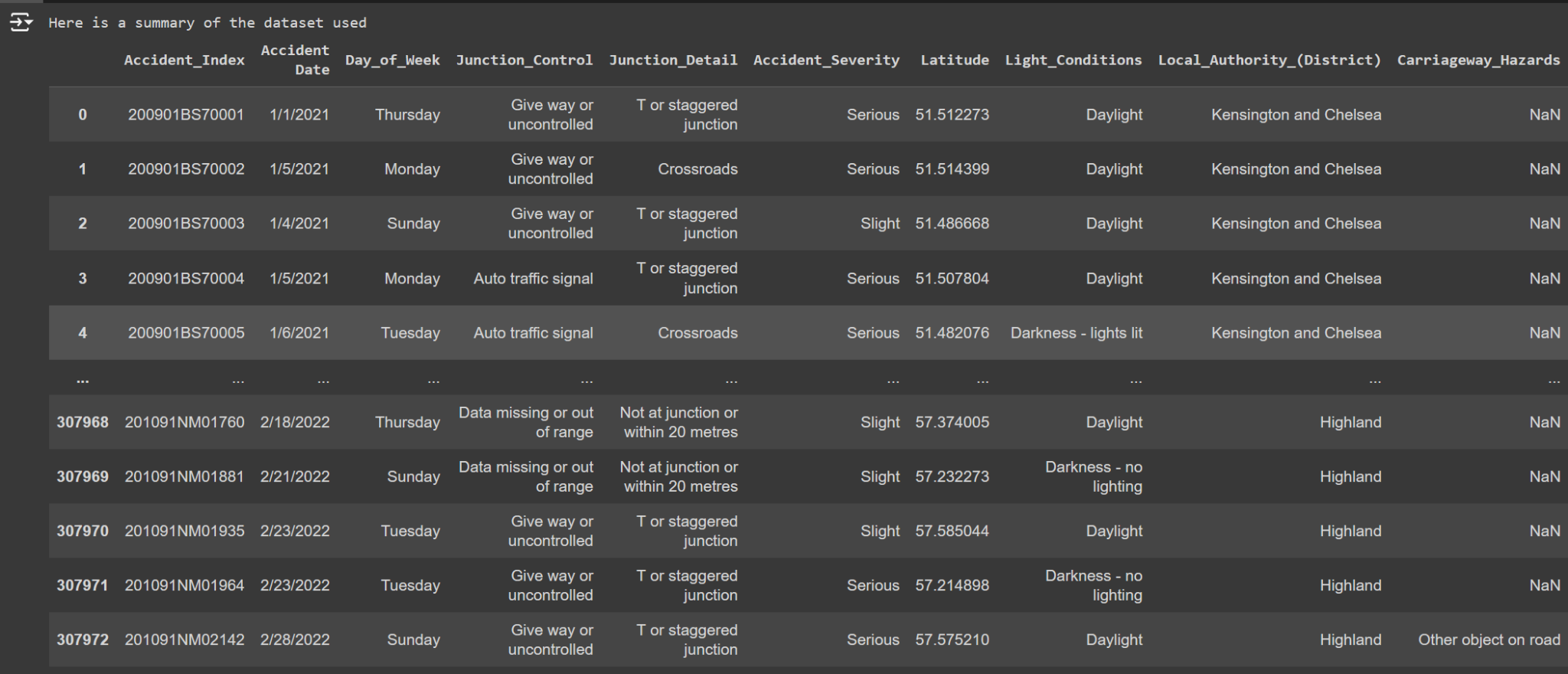
### 

### 

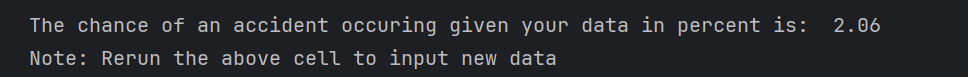
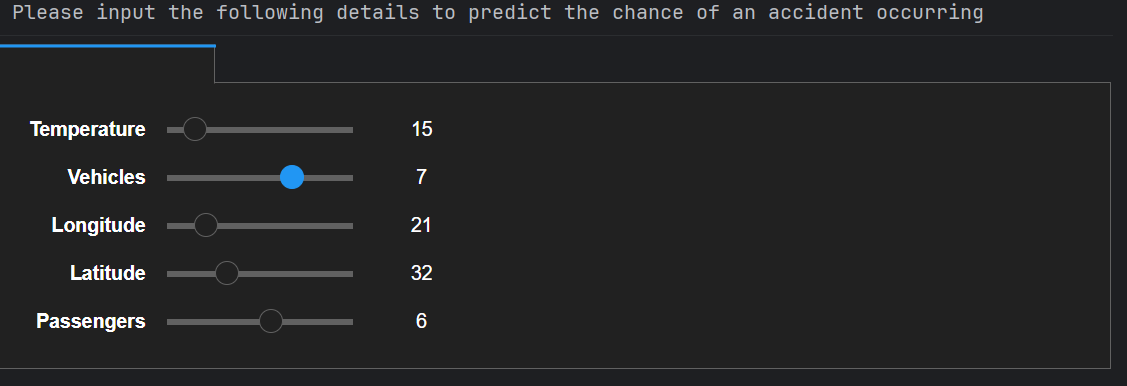
### D.5 - Data exploration support

Visual elements of the data can be seen in section D.7 as the results of the program's output. The storytelling of the data is to show that vehicle accidents don’t just randomly occur, there are factors that lead to them happening. The data analysis in this project shows that to be the case. That factors such as speed, location and weather have a large impact on accidents.

The dataset itself was prepared by downloading a raw dataset off Kaggle, turning it into plain text and then hosting it on Github. The accuracy of the program was detected comparing the program's results from the 2021 dataset with a 2022 raw dataset. An example of the dataset used in the program can be seen below.



An example of the user interactable sliders can be seen below. It detects the values the user inputs and calculates the rate an accident is likely to occur



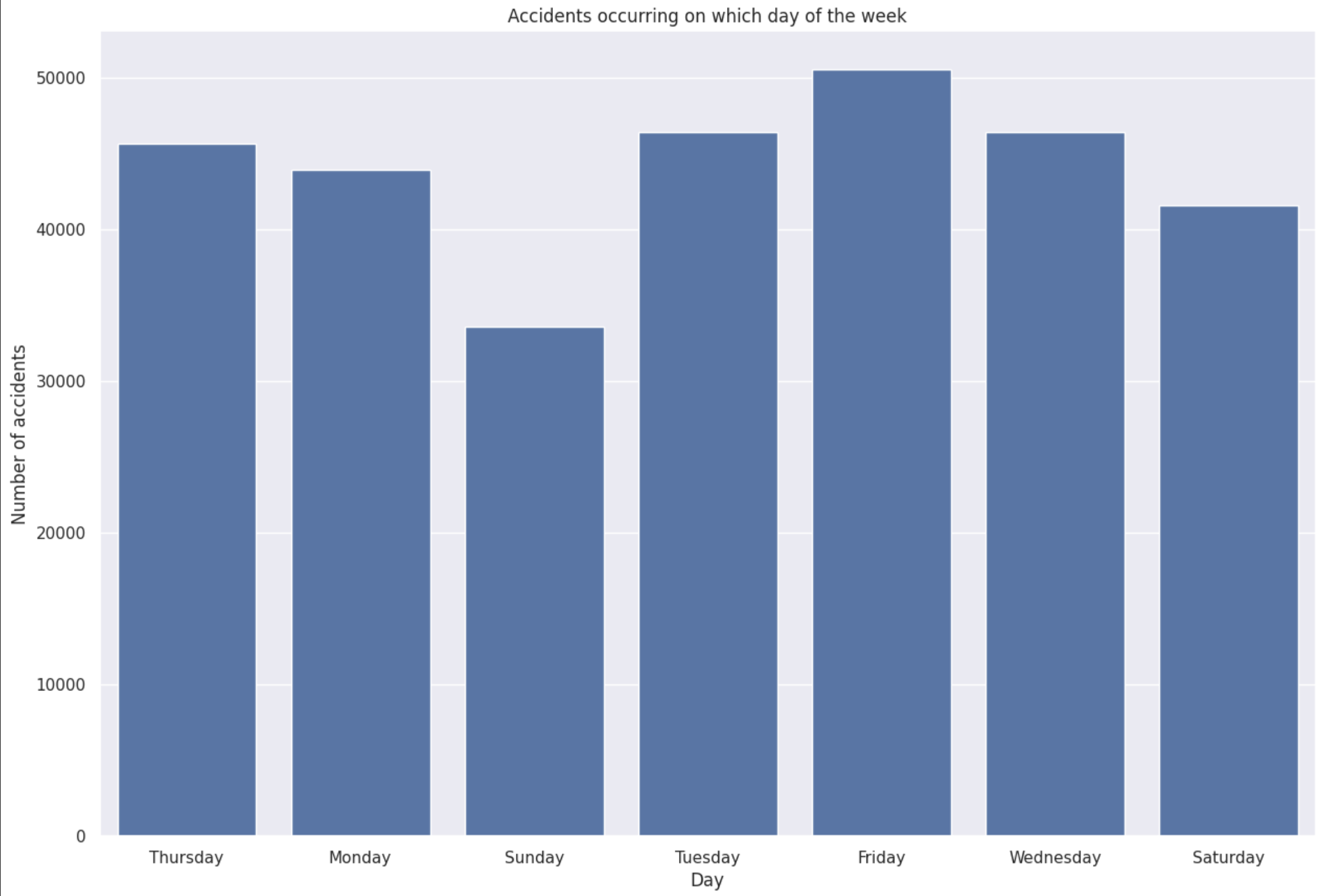
### D.6 - Product accuracy assessment

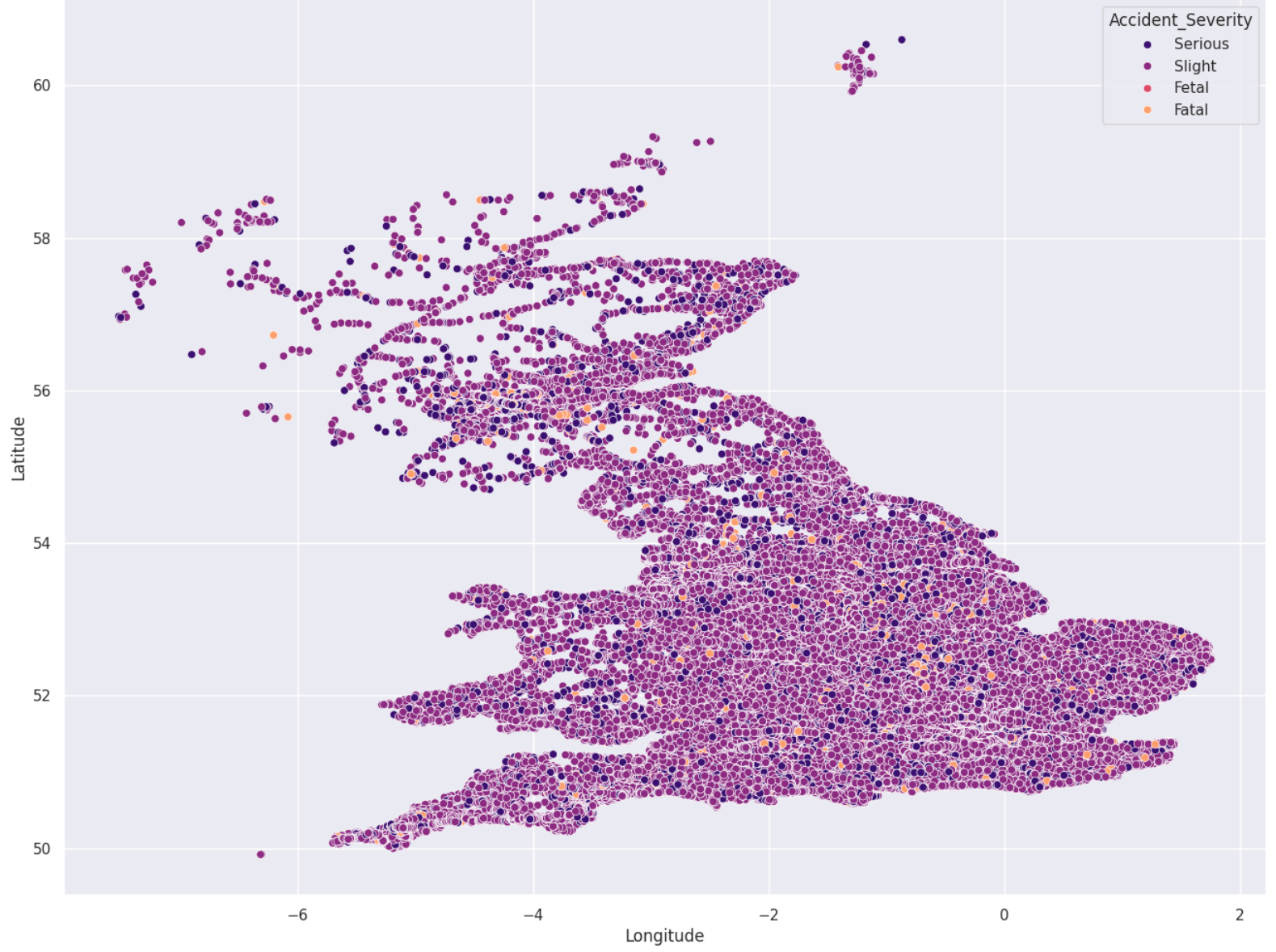
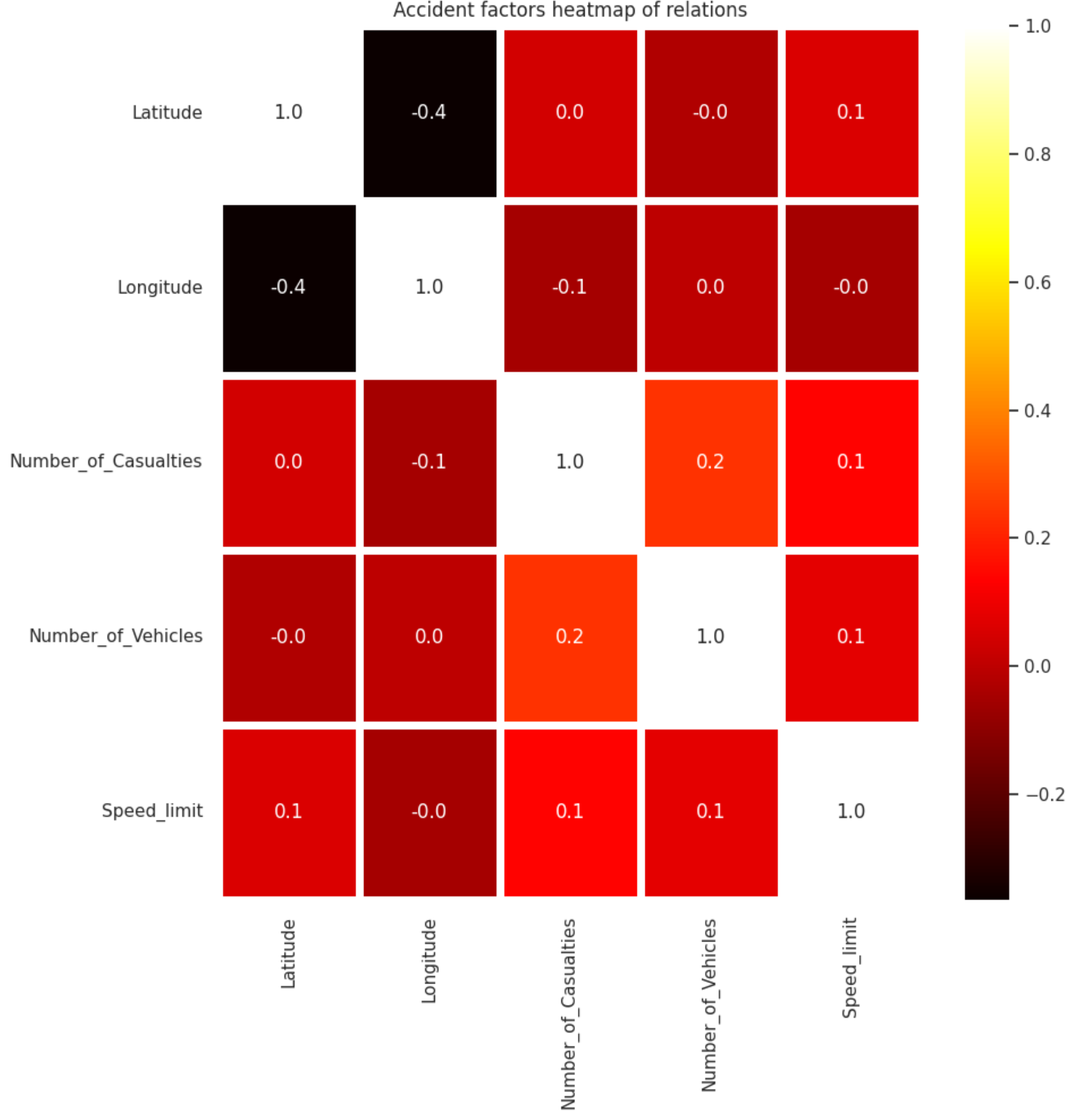
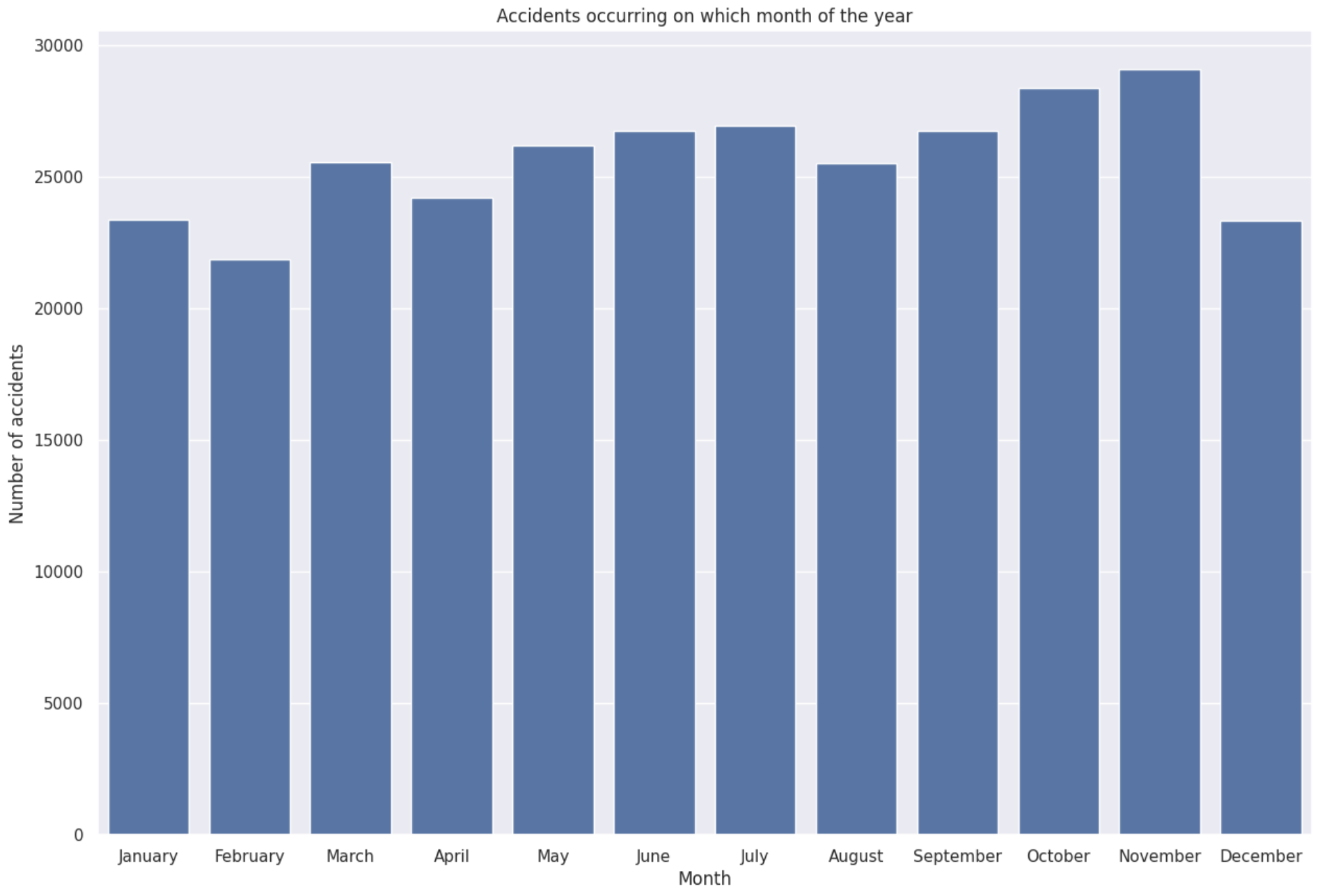
The program was found to have an accuracy of 91.1%, which is within the expected bounds. This was measured by comparing analyzed results from the 2021 dataset of accidents to raw data from 2022. Meaning that the program used the 2021 data to use machine learning to predict accidents in 2021 with 91.1% accuracy.

### D.7 - Results

The 91.1% accuracy was expected, but the lack of datasets means that the results of the machine learning algorithm does not have a high confidence level. The graphed results of the 2021 dataset can be seen below.

The first 2 graphs show the number of accidents on individual days of the week and month. The 3rd graph shows a heatmap of how each accident factor correlated to another factor in regards to accident severity. The 4th graph shows how the location of an accident relates to the severity of an accident.





### 

### D.8 - Source code

See the attached file for the source code.

### 

### 

### 

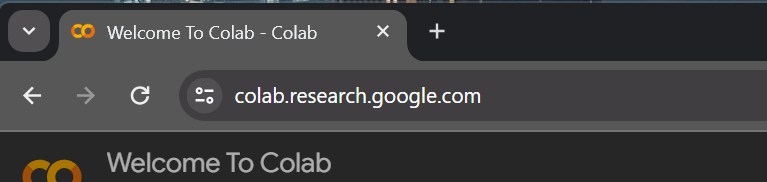
### D.9 - Quick start guide

**Installation**

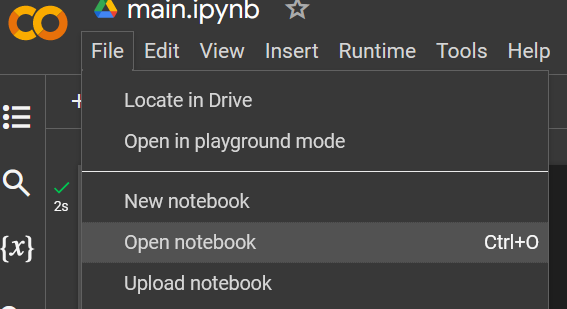
This application is developed using Python and runs within a Jupyter Notebook environment hosted on Google Colab. The prerequisites for using it is a web browser that is compatible with Google Drive as well as the Python file.

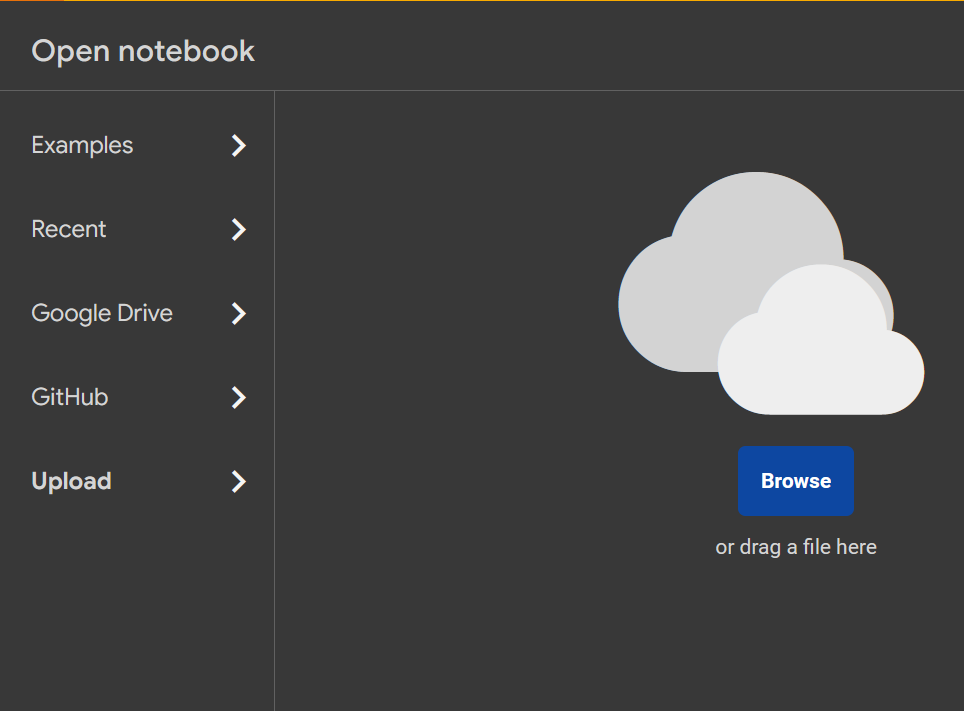
**How to use**

1. Download the Jupyter Notebook file (located within the attached .zip archive)
2. On a web browser, go to [colab.research.google.com](http://colab.research.google.com)

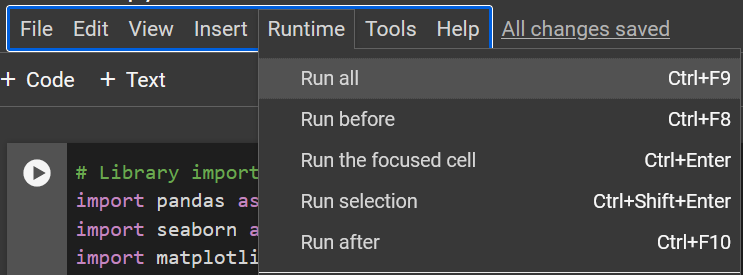


1. Click upload, browse, then select the Jupyter Notebook file

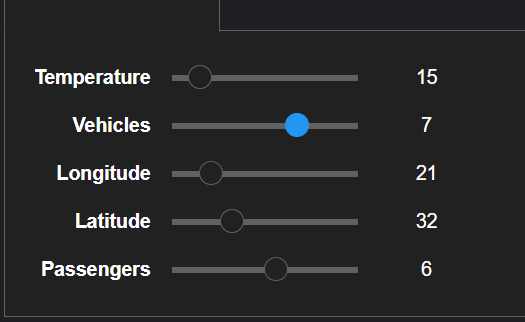




1. On the top bar, click “Runtime”, then on the drop down menu select “Run All”



1. All charts in the application should load within 20 seconds
2. Use the slider at the bottom of the application to input values and get results



## ✾E - Sources

Muhamed, S. (2024, January 10). *Car accident dataset*. Kaggle. https://www.kaggle.com/datasets/nextmillionaire/car-accident-dataset/data