CADT IDT វិទ្យាស្ថានបច្ចេកវិទ្យាឌីជីថល Institute of Digital Technology

# Power BI Final Project Report
# UK Accident 10 Years History

Member: Hoy Seiha, Mony Sovann, Ngang Puthsabbos, Pov Phearum
Subject: Information Visualization using Power BI
Lecturer: Mr Chan Sophal
Date: 21/11/2023

# I. Introduction

Road accidents are one of the major concerns in the United Kingdom. To understand more about the causes of road accidents or the variables that can change the severity of road accidents, we need data to analyze the trends in road accidents. With the analysis we get from the data we can formulate decisions to decrease road accidents and save people's lives.

In this project, we will study the dataset from UK road accidents in the ten-year period from 2005 to 2014. From the dataset, we will apply the data preprocessing techniques on the data then, use PowerBI to visualize the insights in the dataset.

# II. Dataset

## 1. Description

The dataset detailed the road accidents in the UK from 2005 to 2014 which we got from Kaggle. In this dataset, there are three main tables which are Accident, Vehicles, and Casualties. Also, there is a Lookup file that allows you to see all the variable codes in the main tables. Each table has 1640597 unique values.

The Accident table describes each accident which includes accident severity, weather, road conditions, and other details about the surroundings around the accidents with a total of 32 columns. From this table, we can clearly see how external variables can affect accidents. Most importantly we can see the severity of the accidents which are classified into three types which are fatal, serious, and slight.

Secondly, the Vehicles tables describe the condition and properties of the vehicles in the accident. This includes the vehicle type, vehicle model, and vehicle engine capacity with a total of 22 columns. In addition, we can also get information about the driver's and passenger's which includes age and gender. From this, we can see how vehicles can affect the severity and the causes of accidents.

Finally, the Casualties table describes the casualties in the accident. The table includes information about casualties which are casualty severity, age, gender, and casualty class (pedestrian, driver, or passenger) with a total of 15 columns. Even though the Vehicles table also includes information about the casualties, this table provides more information about the casualties.

## 2. Data Source

In this project, we connect our data to PowerBI through Restful API. In order to connect PowerBI through Restful API, there are two steps which are:

1. Connect to MySQL database
2. Use FastAPI framework as backend
3. Connect with PowerBI

### 2.1. MySQL Database

We chose MySQL to store our table because of its simplicity and ease of access. In order to convert our table into MySQL, first we must convert some data types. One data type we must convert is the date column. The date column in our table from Kaggle was in the format “dd-mm-yyyy”, but in MySQL, the default Date format is “yyyy-mm-dd.” We made these changes so we can easily input the data into MySQL. To do this we use Jupyter Notebook to assist us with changing the Date format. For larger tables with 1640597 unique values, we would manually input them into the MySQL data source folder.

```python
def convert_date_format(date):
    try:
        # Try to convert the date from '%d/%m/%Y' to '%Y-%m-%d'
        return pd.to_datetime(date, format="%d/%m/%Y").strftime("%Y-%m-%d")
    except ValueError:
        # If the date is already in the '%Y-%m-%d' format, leave it as is
        return date


df["Date"] = df["Date"].apply(convert_date_format)
```

**Figure 1:** A code snapshot of converting Date format

## 2.2 FastAPI

FastAPI is a modern and fast framework for building RESTful APIs in Python. With FastAPI we would be able to share our dataset with other people and they can retrieve the data in PowerBI for visualization. In this framework, there are three parts which are:

1. Database configuration file
2. Models for our tables
3. The routing file for APIs

The database configuration file database.py will connect the framework to our MySQL database. In this file, we will configure the database username and password to connect with MySQL.

```python
from sqlalchemy import create_engine
from sqlalchemy.orm import sessionmaker
from sqlalchemy.ext.declarative import declarative_base


URL_DATABASE = 'mysql+pymysql://mysql_user_password/schema_name'
engine = create_engine(URL_DATABASE)

SessionLocal = sessionmaker(autocommit=False, autoflush=False, bind=engine)

Base = declarative_base()
```

**Figure 2:** A code snapshot of connecting to MySQL from FastAPI

In order to create different tables in MySQL we need to create different models for each table. The models can be seen in models.py. In this file, we will create different classes and their properties. Their properties will represent the columns of each table. Most importantly, we must set 'Accident_Index' as the primary in the Accidents, Vehicles, and Casualties table, since it is the only primary key we get from Kaggle.

```python
from sqlalchemy import Boolean, Column, ForeignKey, Integer, String
from database import Base


class Casualties(Base):
    __tablename__ = "casualties0514"
    Accident_Index = Column(String(15), primary_key=True, index=True)
    Vehicle_Reference = Column(Integer)
    Casulty_Reference = Column(Integer)
    Casulty_Class = Column(Integer)
    Sex_of_Casulty = Column(Integer)
    Age_of_Casulty = Column(Integer)
    Age_Band_of_Casulty = Column(Integer)
    Casulty_Severity = Column(Integer)
    Pedestrian_Location = Column(Integer)
    Pedestrian_Movement = Column(Integer)
    Car_Passenger = Column(Integer)
    Bus_or_Coach_Passenger = Column(Integer)
    Pedestrian_Road_Maintenance_Worker = Column(Integer)
    Casulty_Type = Column(Integer)
    Casulty_Home_Area_Type = Column(Integer)
```

**Figure 3:** A code snapshot of creating the table Casualties

Finally, in order to route the database to an API we need a routing file that we can see in main.py. In this file, we will have different routing functions which are mostly POST and GET. For the POST function, it is used to insert the data into the database. On the other hand, the GET function will be used to get the data from the database. This will be the routing function that we will use in PowerBI to retrieve all the data.

Furthermore, some tables have a lot of rows which can be time and memory-consuming whenever we want to retrieve the dataset. For easier access, we decided to use pagination to get the data by some amount each time. With pagination, we can get the data faster and easier.

```python
@app.post("/acc_severity/", status_code=status.HTTP_201_CREATED)
async def create_severity(accident_severity: AccidentSeverity, db: db_dependency):
    db_severity = models.Accident_Serverity(**accident_severity.dict())
    db.add(db_severity)
    db.commit()


@app.get("/accidents", status_code=status.HTTP_200_OK)
async def get_all_accidents(db: db_dependency, skip: int = 0, limit: int = 100):
    db_accidents = db.query(models.Accidents).offset(skip).limit(limit).all()
    if db_accidents is None:
        raise HTTPException(status_code=404, detail="Accidents not found")
    return db_accidents
```

**Figure 4:** A code snapshot of the POST function for ‘Accident Severity’ and the GET function for ‘Accidents’

## 2.3 Connect with PowerBI

At this stage, we will have the API from FastAPI to retrieve the data from MySQL through the GET function. Still, the API we have is a local host on the hosting computer. In order to share the dataset with the team, we must forward the route to the public network, then the team will be able to retrieve the data. To solve this problem we use NGROK to forward the localhost IP Address to the public. With the NGROK link provided, the whole team can then access the dataset.

Secondly, in order for PowerBI to connect to the database we use blank query and Power Query M in advance editor to get the data from the API. The next step to get the whole table from the dataset there are a few steps in advance.

1. Transpose the table: When we get the data from the API, the ‘column headers’ will be in one column and the designated values will be in the second column.
2. Delete repeating column header: After transposing the table we will get duplicates of the table header. We have to delete every other row from the table. Then we can use the first row as the column header.
3. Data types: The data types of each column will be any, therefore we must clearly define the data type of each column. We can refer to the data types based on the model in FastAPI.

```
let
    Source = Web.Contents("https://1e3d-103-16-62-134.ngrok-free.app/vehicles?skip=0&limit=10000"),
    JsonContent = Json.Document(Source),
    TableFromList = Table.FromList(JsonContent, Splitter.SplitByNothing(), null, null, ExtraValues.Error),

    // Function to convert each record in the list to a table and transpose it
    recordToTable = (record) => Table.Transpose(Record.ToTable(record)),

    // Use List.Transform to apply the function to each record in the list
    tables = List.Transform(Table.Column(TableFromList, "Column1"), recordToTable),

    // Combine all the transposed tables into a single table horizontally
    combinedTable = Table.Combine(tables),
    #"Promoted Headers" = Table.PromoteHeaders(combinedTable, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"Accident_Index", type any},
    {"Vehicle_Reference", Int64.Type}, {"Vehicle_Type", Int64.Type}, {"Towing_and_Articulation", Int64.Type},
    {"Vehicle_Manoeuvre", Int64.Type}, {"Vehicle_Location_Restricted_Lane", Int64.Type}, {"Junction_Location", Int64.Type},
    {"Skidding_and_Overturning", Int64.Type}, {"Hit_Object_in_Carriageway", Int64.Type}, {"Vehicle_Leaving_Carriageway", Int64.Type},
    {"Hit_Object_off_Carriageway", Int64.Type}, {"First_Point_of_Impact", Int64.Type}, {"Left_Hand_Drive", Int64.Type},
    {"Journey_Purpose_of_Driver", Int64.Type}, {"Sex_of_Driver", Int64.Type}, {"Age_of_Driver", Int64.Type},
    {"Age_Band_of_Driver", Int64.Type}, {"Engine_Capacity_CC", Int64.Type}, {"Propulsion_Code", Int64.Type},
    {"Age_of_Vehicle", Int64.Type}, {"Driver_IMD_Decile", Int64.Type}, {"Driver_Home_Area_Type", Int64.Type}}),

    // Add an index column
    #"Added Index" = Table.AddIndexColumn(#"Changed Type", "Index", 1, 1, Int64.Type),

    // Remove even rows
    #"Filtered Rows" = Table.SelectRows(#"Added Index", each Number.Mod([Index], 2) <> 0),
```

**Figure 5:** A code snapshot in Power Query to get the ‘Vehicle’ table from API

## 3. Data Preprocessing

After loading the data onto PowerBI we decided to clean our data on PowerBI since we only have to take care with missing values. Since our dataset is divided into categories for example the severity of each accident, we did not look into outliers.

From Kaggle the dataset does not have any missing values but with a closer look into the Lookup files we can see that in each column, when there is a missing value the data will be represented as '-1.' With this we decided to convert all '-1' into null in PowerBI using Power Query.

```
#"Replaced Value" = Table.ReplaceValue(#"Reordered Columns",-1,null,Replacer.ReplaceValue,{"Engine_Capacity_CC"}),
```

**Figure 6:** A code snapshot in Power Query to replace '-1' with 'null' in the 'Engine_Capacity_CC' column

After replacing '-1' with null, we decided to fill it up from the next value. We decided to fill up the missing values because each column is a distinct category which are ordinal data type. Therefore it would not be best to find the average or use imputation to fill in missing values.

```
#"Filled Up" = Table.FillUp(#"Replaced Value",{"Engine_Capacity_CC"}),
```

**Figure 7:** A code snapshot in Power Query to replace 'null' with the next value in the 'Engine_Capacity_CC' column

# III. Exploratory Data Analysis

- Accidents based on surrounding (Road, Weather, Area, Ligh Condition, Journey Purpose):

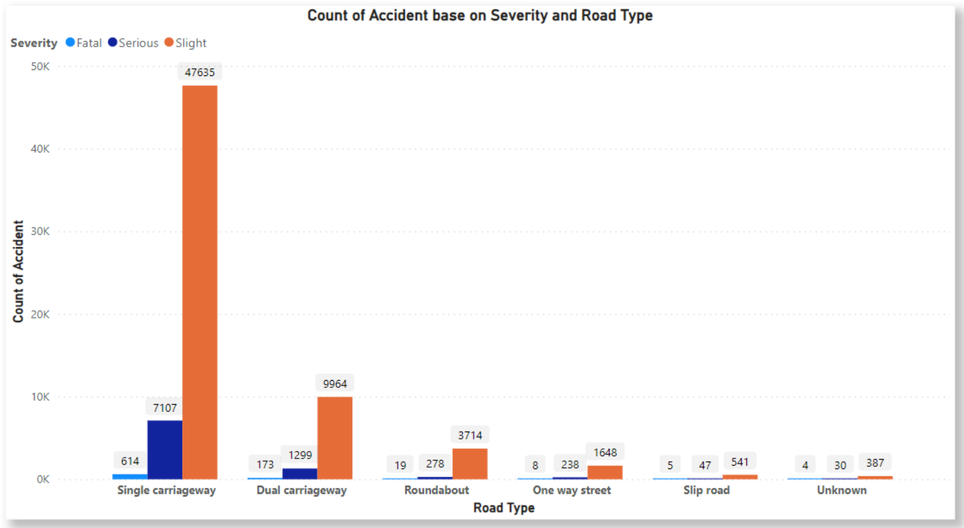


**Figure 8:** A bar graph showing the number of accidents based on road types

**Graph Count of Accident based on Severity and Road type** shows which type of road has the most accidents in different accident serverity, so based on this graph we found that the majority of accidents occur on single carriageways because on this type of road any type of vehicle travel in the same land. So it is easy to have an accident.

- Severity Distribution:
  - Fatal accidents account for approximately 5% of all accidents.
  - Serious accidents make up about 20% of the total accidents.
  - Slight accidents constitute the remaining 75% of accidents.
- Road Type Distribution:
  - Single carriageways account for about 50,000 accidents.
  - Dual carriageways contribute to about 11,000 accidents.
  - Roundabouts, one-way streets, slip roads, and unknown road types account for the remaining accidents.

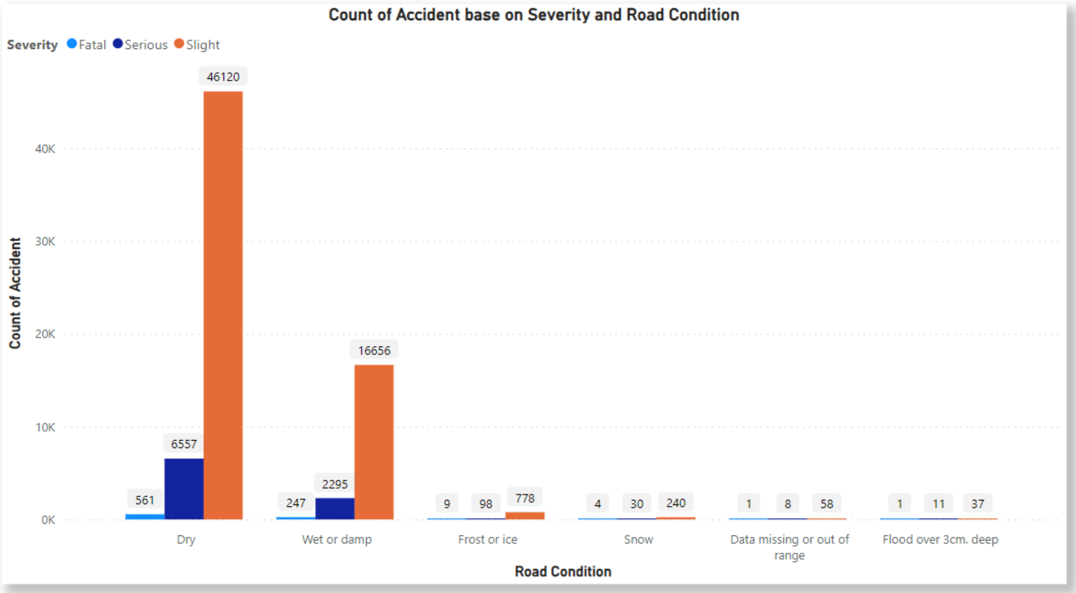


**Figure 9:** A bar graph showing the number of accidents based on road condition

**Graph Count of Accident based on Severity and Road Condition** shows which condition of the road has the most accidents in different accident severity, so based on this graph we found that the majority of accidents occur on roads in good condition (Dry), because most people travel on this condition of the road more than another.

- Severity Distribution:
  - Fatal accidents account for approximately 0.4% of all accidents.
  - Serious accidents make up about 7% of the total accidents.
  - Slight accidents constitute the remaining 92.6% of accidents.
- Road Condition Distribution:
  - Roads in good condition (Dry) account for about 54,046 accidents.
  - Roads with wet or damp conditions contribute to about 24,895 accidents.
  - Frost or ice, snow, data missing or out of range, and flood over 3cm. deep account for the remaining accidents.

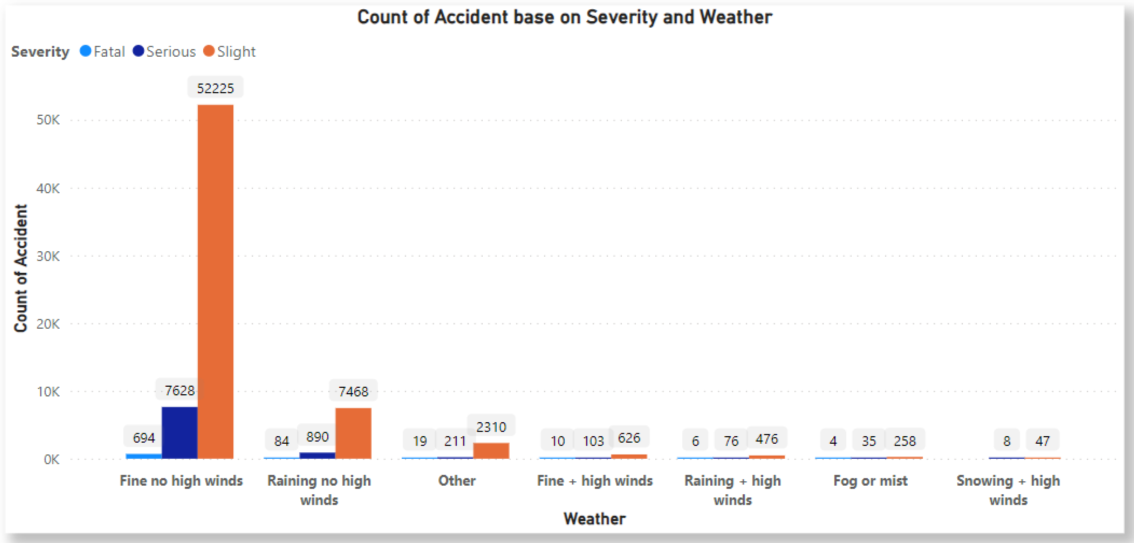


**Figure 10:** A bar graph showing the number of accidents based on weather

**Graph Count of Accidents based on Severity and Weather** shows the weather when the most accidents occur, so based on this graph we found that the majority of accidents occur in fine weather without high winds, because most people travel when the weather is fine.

- Severity Distribution:
  - Fatal accidents account for approximately 0.3% of all accidents.
  - Serious accidents make up about 6 of the total accidents.
  - Slight accidents constitute the remaining 93.7% of accidents.
- Weather Distribution:
  - Fine weather without high winds accounts for about 60000 accidents.
  - Raining weather with high winds contributes to about 8,000 accidents.
  - "Other" weather conditions, including fine weather with high winds, rainy weather with high winds, fog or mist, and snowing weather with high winds, account for the remaining accidents.

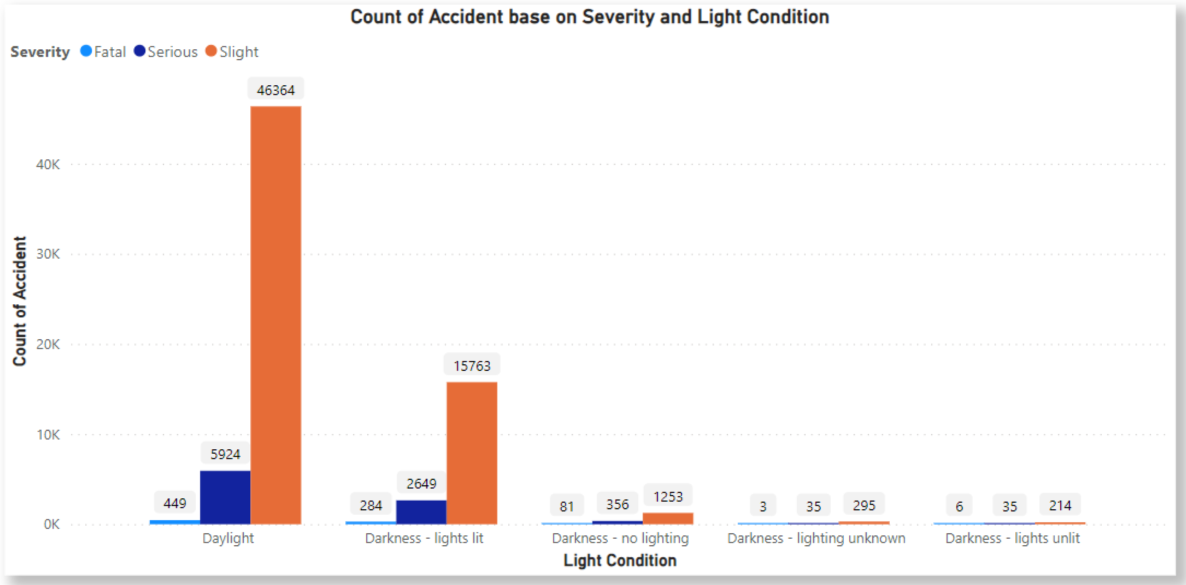


**Figure 11:** A bar graph showing the number of accidents based on light conditions

**Graph Count of Accidents based on Severity and Light Condition** shows the light condition that has the most accidents occur, so based on this graph we found that the majority of accidents occur under daylight conditions because most people travel in the morning or in the afternoon.

- Severity Distribution:
  - Fatal accidents account for approximately 1.7% of all accidents.
  - Serious accidents make up about 4.7% of the total accidents.
  - Slight accidents constitute the remaining 93.6% of accidents.
- Light Condition Distribution:
  - Daylight accounts for about 50,000 accidents.
  - Darkness with lights lit contributes to about 18,000 accidents.

- Darkness with no lighting, darkness-lighting unknown, and darkness lights unlit account for the remaining accidents

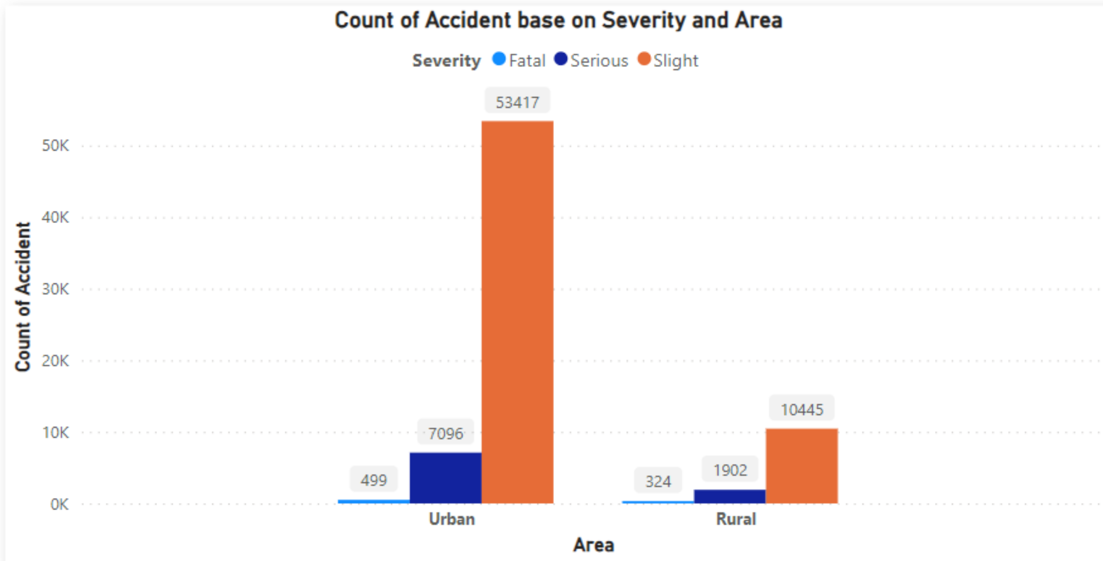


**Figure 12:** A bar graph showing the number of accidents based on areas of accidents

**Graph Count of Accidents based on Severity and Area** shows the area where the most accidents occur, so based on this graph we found that the majority of accidents occur in urban areas because most people live in this area.

- Severity Distribution:
  - Fatal accidents account for approximately 0.8% of all accidents.
  - Serious accidents make up about 2.7% of the total accidents.
  - Slight accidents constitute the remaining 96.5% of accidents.
- Area Distribution:
  - Urban areas account for about 67,000 accidents.
  - Rural areas contribute to about 12,000 accidents.

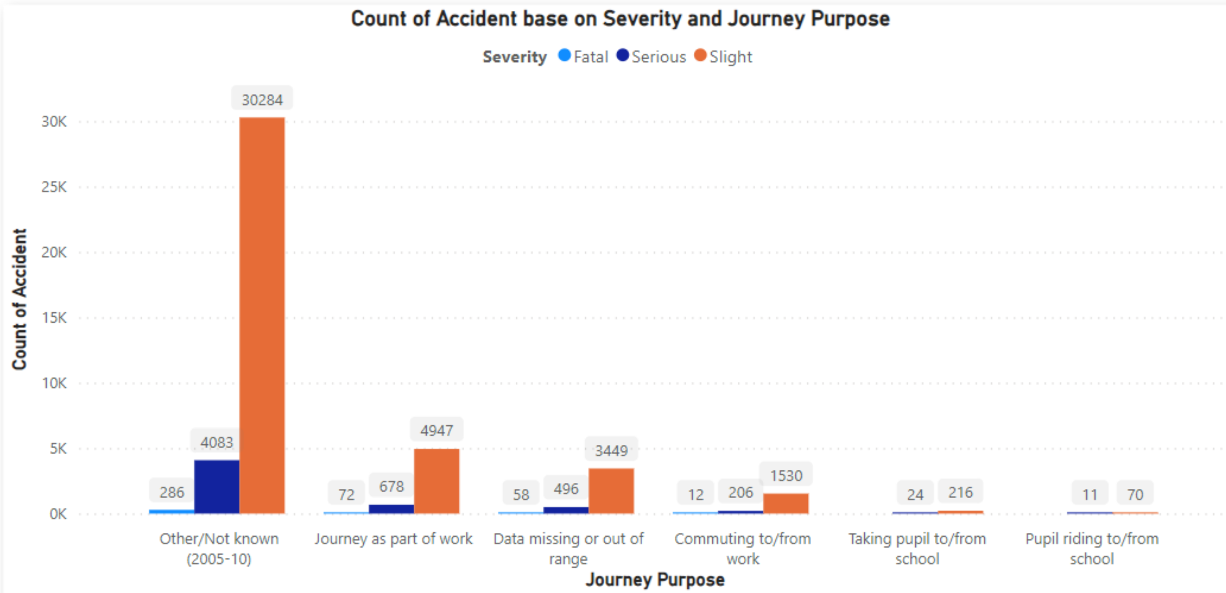


**Figure 13:** A bar graph showing the number of accidents based on journey purpose

**Graph Count of Accidents base on Severity and Journey Purpose** shows the purpose of the casualty's journey that has the most accidents occur, so based on this graph we do not know about the purpose of their journey that the majority of accidents occur , because mostly people travel form far the scene.

- Severity Distribution:
  - Fatal accidents account for approximately 0.5% of all accidents.
  - Serious accidents make up about 1.9% of the total accidents.
  - Slight accidents constitute the remaining 97.6% of accidents.
- Journey Purpose Distribution:
  -other accounts for about 35,000 accidents
  - Journey as part of work accounts for about 5,500 accidents.
  - Commuting to/from work, Taking pupils to/from school, Pupil riding to/from school and Other/Not known account for the remaining accidents

**Insight:**

- Single-carriageways are the most prevalent road type for accidents, suggesting a need for enhanced safety measures on these roads.
- Comparing severity across road types reveals that single-carriageways are generally more prone to accidents, while roundabouts and one-way streets appear to be safer options.
- Roads in good condition (Dry) are the most prevalent road type for accidents, suggesting a need for proactive safety measures even under favorable conditions.
- Fine weather without high winds is the most prevalent weather condition for accidents, suggesting that overall driving behavior and road conditions play a significant role in accident occurrences.
- Daylight is the most prevalent light condition for accidents, suggesting that visibility and lighting conditions play a significant role in accident occurrences.
- Focusing on visibility-enhancing measures, such as proper lighting and improved road infrastructure, could help mitigate accidents during low-visibility conditions.
- Urban areas are the most prevalent area for accidents, suggesting that driving behaviors and road conditions in urban environments play a significant role in accident occurrences.
- Fatal accidents are the least common type, but they have a significant impact, emphasizing the importance of proactive safety measures.
- Lighter accidents, such as slight accidents, are more frequent, indicating a broader focus on overall road safety is necessary.
- Focusing on traffic management, road infrastructure improvements, and public education campaigns in urban areas could help mitigate accidents.
- Emphasizing safe driving practices, such as adhering to speed limits and maintaining a safe distance from other vehicles, is important for both urban and rural environments.

- Accidents based on Vehicles:

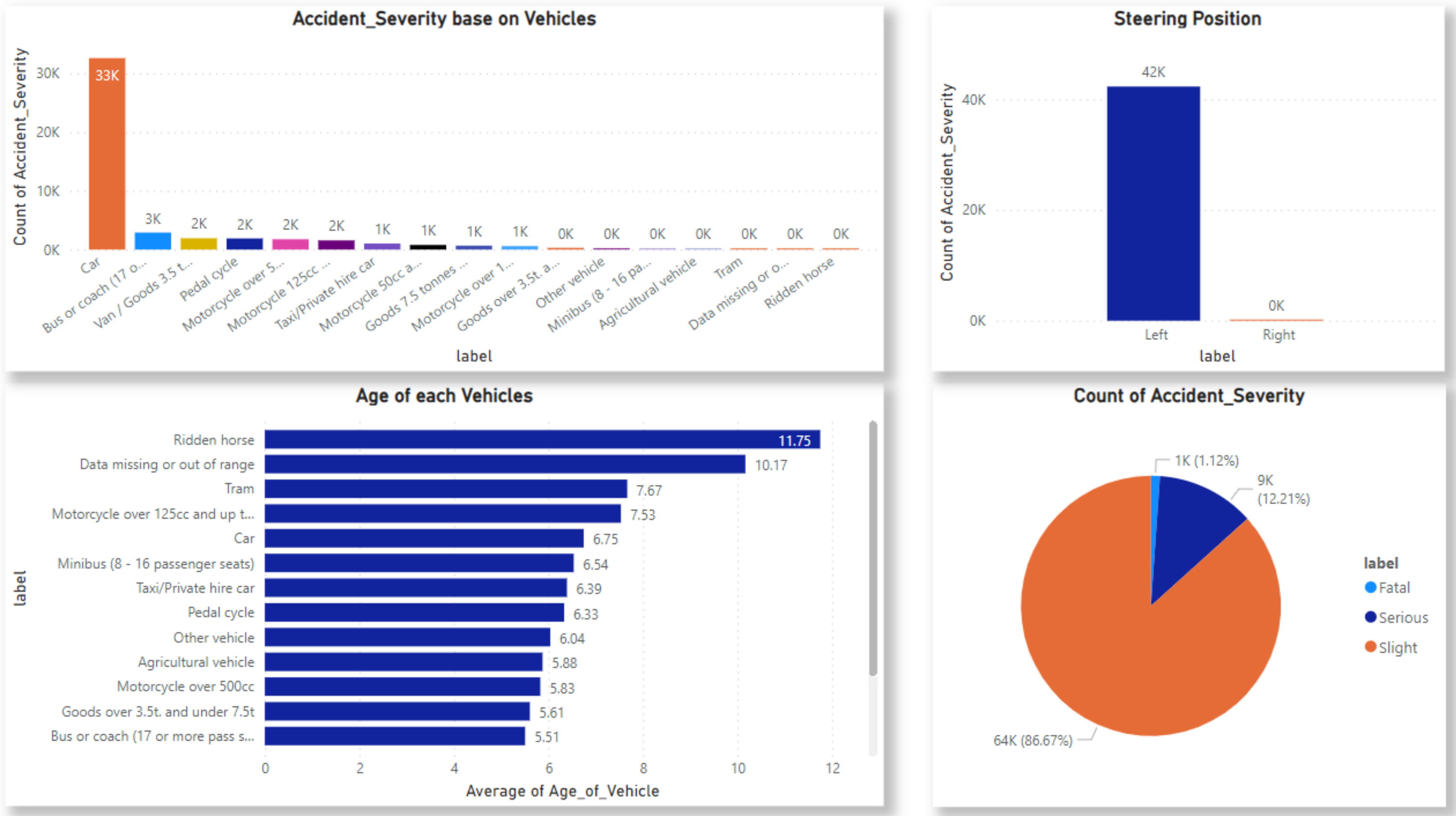


**Figure 14:** Bar graphs and a pie chart showing the effects of vehicles on road accidents

**Graph Accident_Severity based on Vehicles** shows which vehicle has the most accidents, so based on this graph we found the most accidents are cars because in the UK at least one car in each home is why cars mostly cause accidents. **Graph Steering Position** shows which between left steer and right steer has the most accidents, based on this graph we found the left steer caused more accidents because in the UK most cars have left steer, but according to law drivers also have to drive on the left lane. **Graph Age of each vehicle** shows the age of each vehicle. **Graph Count of Accident_Severity** shows the types of accidents: Fatal is 823 accidents(1.12%), Serious is 8999 accidents(12.21%), and Slight is 63889 accidents(86.67%). So based on this graph most accidents are Slight.

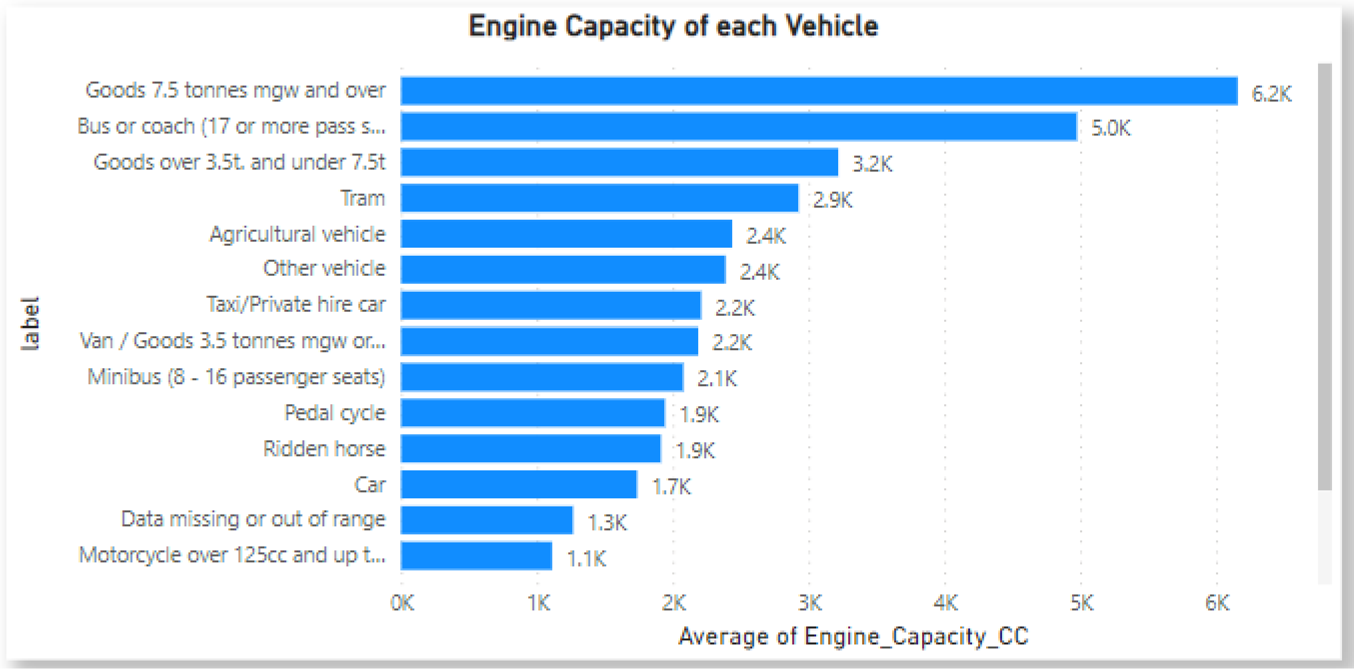


**Figure 15:** A bar graph showing the average engine capacity of vehicle type

**Graph Engine Capacity of each Vehicle** shows which vehicle has the highest engine capacity, so based on this graph the highest engine capacity is Good 7.5 tonnes mgw and over. If we compare a car and a Good 7.5 tonnes mgw, cars have more accidents than a Good 7.5 tonnes mgw but if we look at the probability of Fatal, Good 7.5 tonnes mgw has more than a car.

**Insight:**

- Accidents with larger vehicles tend to increase the fatality rate.
- More accidents impact cars because of most people driving cars in the UK.
- More accidents from Left steering because in the UK people drive on the left lane.

- Accidents Based on Human Factors

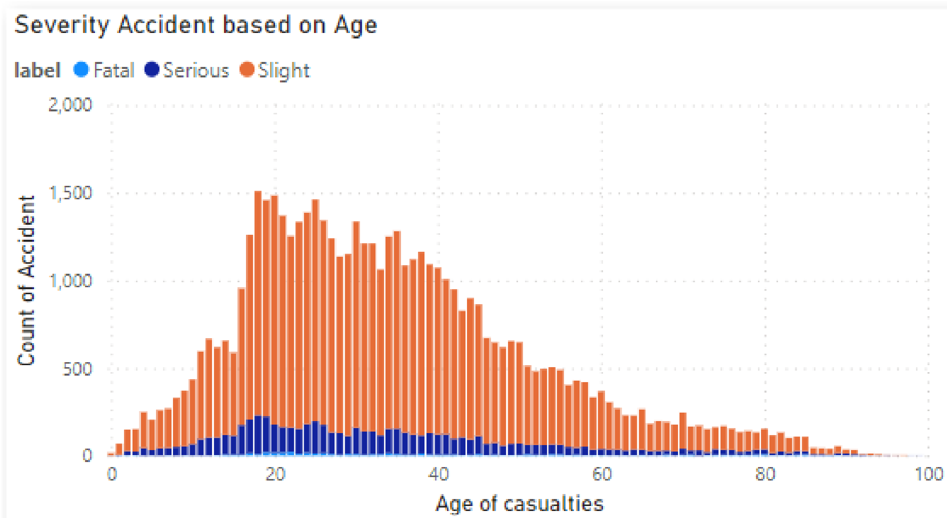


**Figure 16:** A bar graph showing the number of accidents based on age

**Graph Road Accident based on Age** show the severity of accident based on age we can see that at the age of 18 there is the most slight accident and also a serious accident. On the other hand, for fatality accidents, we can see that the age of 22 is the most fatal accident.

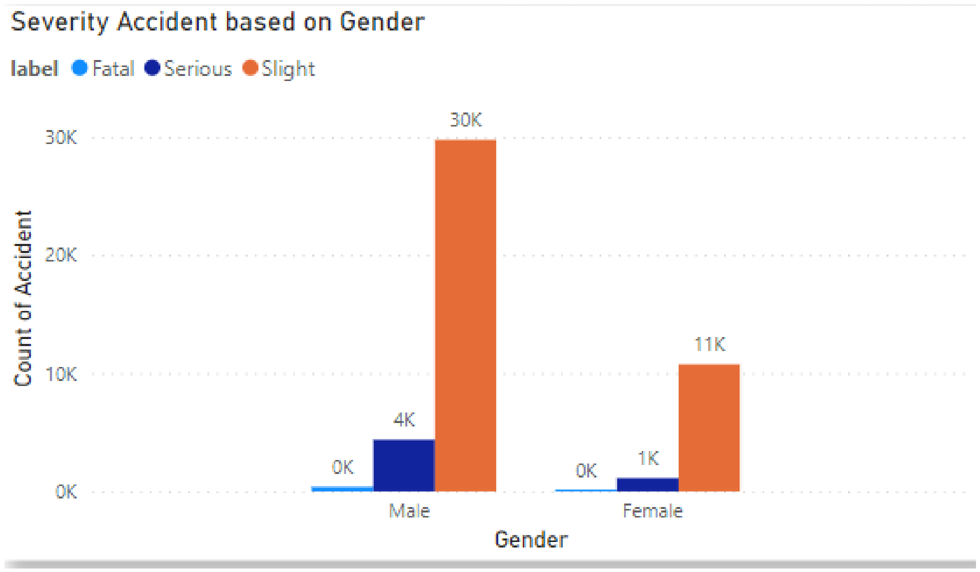


**Figure 17:** A bar graph showing the number of accidents based on gender

**Graph Road Accident based on Gender** shows the severity of accidents based on gender so we can see that more men are having more road accidents than women. From this, we can foresee that more men are on the road because the population of the UK is made up of 51% women and girls.

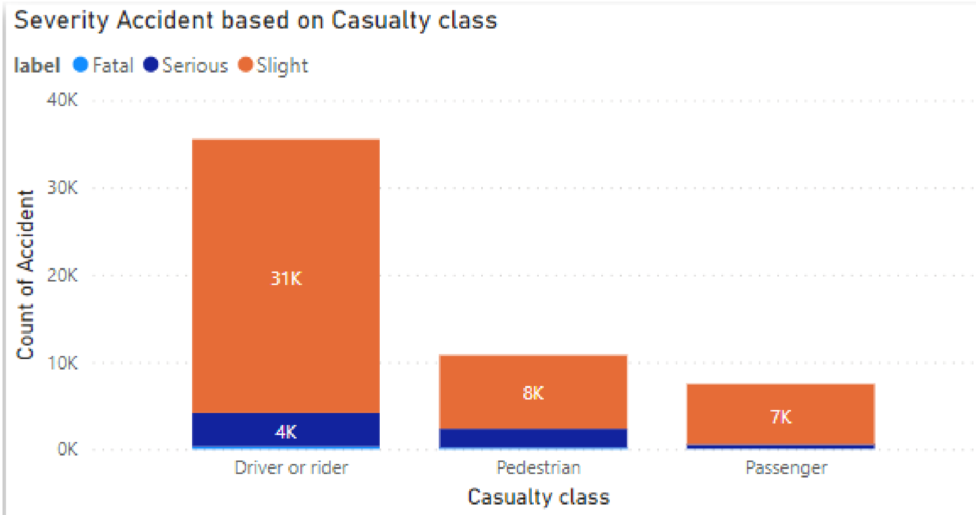


**Figure 18:** A bar graph showing the number of accidents based on casualty class

**Graph Road Accident based on Casualty class** shows the classes of casualties in a vehicle. From this, we can see that if you are a driver or a rider, then you will have a high chance of getting into an accident, but we can not neglect the fact that most people drive or ride on their own. As for severity, we can see that as a pedestrian you will have a high probability of a fatal accident. In addition, if you are a passenger then you are the safest.

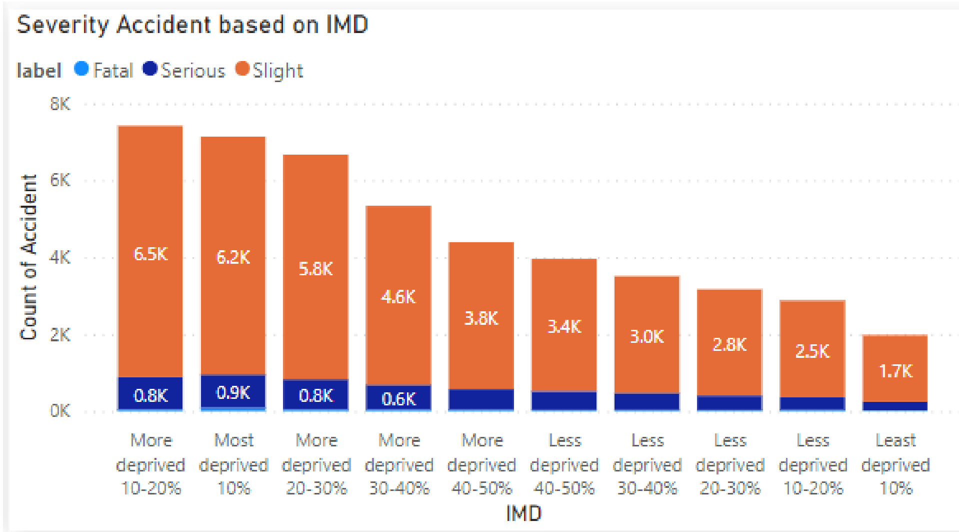


**Figure 19:** A bar graph showing the number of accidents based on IMD

**Graph Severity Accident based on IMD** shows that people with a higher deprived percentage of IMD will have a higher chance of getting into a road accident. This is because people with a higher percentage of IMD means they have a harder life than lower IMD. This can relate to their education or their income. Poor education will lead to risky driving.

# IV. Conclusion

From the overall analysis of the dataset, we can identify the relationships between the variables (surrounding, vehicles, and human factors) on the severity of road accidents. In the surrounding factor, we can see that the road types are the main impact on the severity of road accidents. Next looking onto the Vehicles factor, we can see that the vehicles with the left steering wheel tends to face more road accidents. Also we can also see that the severity of the accident is based on the size of the vehicles. Finally, looking at the human factor we ca see a correlation with younger people, where they are likely to have more road accidents than older people.

## V. Appendix

1. Github: https://github.com/SeihaHoy/powerbiapi.git
2. PowerBI for Surrounding visualization: https://app.powerbi.com/links/2U9YyWzytt?ctid=1e9461ec-5362-4329-ae46-61fa3e91c6d2&pbi_source=linkShare
3. PowerBI for Vehicles visualization: https://app.powerbi.com/links/qRDlsPA011?ctid=1e9461ec-5362-4329-ae46-61fa3e91c6d2&pbi_source=linkShare
4. PowerBI for Human factor visualization: https://app.powerbi.com/links/hiDPt7dxg5?ctid=1e9461ec-5362-4329-ae46-61fa3e91c6d2&pbi_source=linkShare