Report 1

Business and Data understanding

Project Title:

“UK Road Safety – Accident Severity Analysis”

Team:
Melesse Perschewski, Kateryna Kantsyr (Group 6)
Github

Business Understanding

Background

Road accidents in the United Kingdom pose social and economic challenges. Understanding where, when and why accidents happen could improve public safety, reduce the severity and support as a base for decision-making. The Kaggle datasets UK Road Safety: Traffic Accidents and Vehicles provides insight on these accidents and the involved vehicles. However, for this project, only the Accident_Information.csv dataset will be used.

Business Goals

The project aims to provide insight into factors influencing the severity of road accidents. Additionally, the project tries to develop a model, predicting the severity of an accident based on its factors, potentially giving people an even clearer understanding of what contributes to severe accidents.

In detail the goals are as stated below:

1. Analyse where, when and why accidents happen
2. Identify key factors influencing the severity of an accident.
3. Develop a predictive machine learning model that predicts accidents as either slight or severe based on the given factors.

Assessing the situation

- Dataset: Kaggle dataset UK Road Safety: Traffic Accidents and Vehicles Accident_Information.csv (705.45 MB)
- Tools: GitHub, Kaggle, Jupyter Notebook
- Hardware: Personal

Requirements, assumptions and constraints

The project must be finished within the course deadline, including checkpoint submissions. A finished project is expected to provide an analysis of the data as specified before and a functioning predictive model with documented performance. Both team members are required to be able to access the dataset.
The dataset is assumed to be complete for the specified years, imbalances are assumed to be manageable using the learned techniques and it is assumed to be stable throughout the years.
Since the dataset is rather large it may be slow to work with due to limited hardware.
Variables may be noisy, incomplete or inconsistent, limiting their usefulness for prediction tasks.

Risks and contingencies

Risk: Dataset too large for smooth local processing
Solution: Using Google Collab for faster processing

Risk: Data inconsistent or missing
Solution: Make use of cleaning techniques

Risk: Poor model performance caused by imbalance
Solution: Try alternative models

Terminology

- *Accident Severity*: Binary column in the dataset, describing whether an accident was serious or slight
- *Slight accident*: Accident resulting in slight injury
- *Serious accident:* Accident resulting in serious injury
- *Environmental features*: eg. weather, light conditions, road surface condition

Costs and Benefits

The project leads to no financial costs. The only costs consist of time and computational resources.
Benefits of the project are insights in the factors causing road accidents, a working predictive model and experience for the team.

Data-mining goals

1. Explore correlations related to accident severity and report them
2. Build a classification predicting the outcome of accidents as serious or slight

Data-mining success criteria

1. A predictive model, with acceptable performance
2. Interpretable results for which factors influence the severity
3. A report on the question when, where and why accidents happen

Data Understanding

Outline data requirements

To meet the goals of the project - understanding when, where, and under what conditions road accidents occur in the United Kingdom and predicting their severity - the data should include variables that describe the main aspects of each accident.

First, **time-related** information such as *the date, year, day of the week*, and *time of day* is needed to identify temporal patterns and trends.
Second, **environmental data** (for example, *weather*, *lighting*, and *road surface conditions*) help to understand how surroundings influence accident severity.
Third, **spatial data**, including *latitude*, *longitude*, and whether *the area is urban or rural*, are necessary to determine where accidents are most common and to detect high-risk locations.
Finally, **outcome variables**, such as *the number of vehicles*, *number of casualties*, and the *severity of the accident*, are essential for both descriptive analysis and the predictive model.

Verify data availability

The dataset *UK Road Safety - Accidents and Vehicles* was obtained from **Kaggle** and published by the *UK Department for Transport, stored in CSV format*. From this collection,

only the file **Accident_Information.csv** was used. It originally contains about **2 million accident records and 34 variables** covering the years **2005–2017**, with each row representing a single reported accident.

For the purpose of this project, data from **2012–2017** were selected, resulting in approximately **830,000 records** after filtering and initial cleaning.

The table below lists the **12 key variables** most relevant for analysis and modeling.

Variable	Description	Type	Measurement Level
Accident_Index	Unique identifier for each accident	object	ID
Date	Date of the accident	datetime	Temporal
Time	Time of day (HH:MM)	object	Temporal
Day_of_Week	Day when the accident occurred	object	Ordinal
Year	Year of the accident	int	Temporal
Weather_Conditions	Weather during the accident	object	Nominal
Light_Conditions	Lighting at the time of the accident	object	Nominal
Road_Surface_Conditions	Condition of the road surface	object	Nominal
Urban_or_Rural_Area	Whether the area is urban or rural	object	Nominal
Speed_limit	Legal speed limit at the location (mph)	float	Ratio
Number_of_Vehicles	Number of vehicles involved	float	Ratio
Accident_Severity	Severity level (Slight / Serious / Fatal)	object	Ordinal (Target)

A preliminary inspection in Python confirmed that all essential variables - such as date, time, location, weather, lighting, road surface, and accident severity - **are present and correctly formatted**. The dataset was successfully imported into *Jupyter Notebook* using *Pandas* without structural issues.

The **analysis of missing values** showed that most key variables are complete. Some attributes, such as 2nd_Road_Number(≈0.5%) and Carriageway_Hazards (≈98%), contain missing or undefined values, but these are not critical for the main objectives and can be handled during the data preparation stage.

Checks on numeric and spatial fields confirmed that all latitude and longitude values (100%) fall within the valid UK geographic range, and numeric variables such as Speed_limit, Number_of_Vehicles, and Number_of_Casualtiescontain realistic values. **No duplicate** *Accident_Index records* were found.

The distribution of the *target variable* Accident_Severity indicates a class imbalance, with Slight (83.8%), Serious (15.0%), and Fatal (1.2%) accidents. This imbalance will be addressed later during modeling.

A more detailed initial data analysis and its results are available in the Jupyter Notebook file stored in the project's GitHub repository

Define selection criteria

To ensure that the analysis focuses on recent and comparable data, only records from **2012 to 2017** were included in the study.
This **five-year period** provides a balanced and reliable timeframe with sufficient data coverage, representing the **most recent complete data** available in the dataset. It also

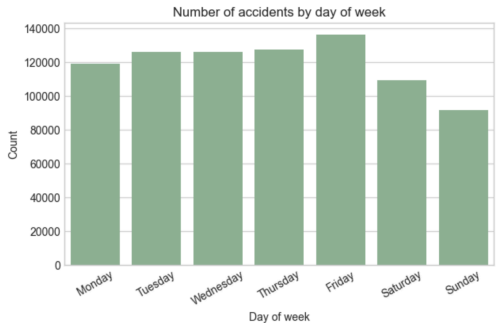
aligns with the project's objective of analyzing current accident patterns and trends in the United Kingdom.

Furthermore, records were retained only if they contained a **valid accident date**, a defined **severity level**, and information on at least **one involved vehicle**.
Spatial boundaries were applied to include only accidents that occurred **within the United Kingdom** (Latitude 49-60, Longitude -10 - 2).
Numerical attributes such as Speed_limit, Number_of_Vehicles, and Number_of_Casualties were checked to ensure that they fall within **reasonable and legally possible ranges**.

After applying these selection criteria, the dataset was reduced to approximately **830,000 records**, forming a representative and consistent subset for further analysis.

Exploring data

An initial **Exploring data stage** was conducted to better understand the main patterns in the dataset after filtering and cleaning.
Several **visualizations** were created to illustrate accident distributions by time, environmental conditions, and location.
Some of these results are demonstrated in the figures below.



Most accidents occurred on **weekdays**, with **Friday** showing the highest frequency.
The majority of cases took place under **daylight** and *"Fine, no high winds"* weather conditions, meaning that accidents often occur in normal driving circumstances rather than in extreme weather.
The most common speed limit in accident locations is **30 mph**, typical for **urban areas**, which suggests that dense traffic environments contribute to the higher number of incidents.

Temporal analysis indicated that accidents are most frequent during **daytime hours (8:00-18:00)**, aligning with daily traffic peaks.
A spatial overview also confirmed a higher concentration of accidents around **large cities and major highways**, while rural regions showed lower densities.

Overall, this preliminary exploration helped identify key behavioral patterns in the data and supported the selection of variables for further descriptive and predictive analysis.

Verifying Data Quality

The final stage of data understanding assessed the overall **quality, accuracy, and consistency** of the dataset. Checks confirmed that variable formats are consistent, relationships between attributes are logical, and no critical errors are present.

Minor missing values in non-essential columns (e.g., 2nd_Road_Number, Carriageway_Hazards) do not affect the main analytical goals and will be handled later if necessary. Categorical variables use standardized codes, and numerical values fall within realistic and legally valid ranges.

The dataset shows strong internal coherence, with each accident linked to at least one vehicle and a valid severity level.

Overall, the data are **complete, reliable, and representative**, providing a solid foundation for further data preparation and modeling.

Planning part

Task 1: Initial Data cleaning (5 hours each member)

- Includes downloading and the first look at the dataset
- Performing basic data cleaning such as handling invalid values

Task 2: Exploratory Data Analysis (8 hours each)

- Identifying relevant features for modeling
- acquiring relevant data for Business Goal 1
- visualizing found patterns
- preparing summary of the findings for the report

Task 3 Feature Engineering (5 hours each):

- selecting variables to be used for the model
- encoding categorical data
- producing a clean dataset for modeling

Task 4: Model Development, Fine tuning and Evaluation (9 hours each):

- building and training models such as RandomForest or Decision Trees
- performing train/test splits
- hyperparameter tuning
- evaluating the models based on metrics such as accuracy and precision
- includes documenting the process and results

Task 5: Interpretation and reporting (3 hours each):

- interpreting findings
- linking findings to business goals
- prepare presentation for Poster Session