193.187 Information Visualisation

Group Assignment 1 - Project Proposal

UK Road Safety

Group 14:

Valentin Tian (12426893)

Kashir Hasnain (12350030)

Doan Chau Tran (11736088)

Ildar Fatkullin (11724445)

December 2025

Preface

Our project name is **UK Road Safety**, an interactive visualisation tool to help analysts, government officials and the public explore *where, when, and to whom* serious road injuries (KSI - killed or seriously injured) occur in the UK. With a focus on what risks are by vehicle type, road type, location, time, and demographics.

We use the UK Department for Transport (DfT) road safety open data (latest 5 years): *collisions*, *vehicles*, *casualties*) and its 2024 code list for data definitions.

Part 1 - Data Description

1.1. Selected Topic & Motivation

Our visual tools will let users explore collisions/casualties patterns by place (map), time (trend), and context (road type, environment, demographics, vehicle characteristics). Government officials can use these views to identify high-risk areas and affected population groups, insurance analysts can examine driver and vehicle segments to understand collision risk. These visualisations should help to support evidence-based decision-making and targeted interventions.

1.2. Dataset Description

- **Collisions** - *100,927 rows*, *44 variables* (location, date/time, severity, road class/type, speed limit, junction, light/weather/surface, etc.). One police-reported personal-injury collision. PK: collision_index (unique).
- **Vehicles** - *183,514 rows*, *32 variables* (vehicle type, first point of impact, driver age/sex, etc.). One vehicle involved in a collision. Multiple vehicles may be associated with the same collision. Composite PK: (collision_index, vehicle_reference). FK: Collisions.
- **Casualties** - *128,272 rows*, *23 variables* (type, age/sex bands, injury severity, etc.). One casualty involved in a collision. Several casualties may be linked to one vehicle and one collision. Composite PK: (collision_index, vehicle_reference, casualty_reference). FKs: Collisions and Vehicles.

The dataset is collected by police officers across the UK and aggregated by the DfT. The full dataset contains almost 100 columns across the three tables. For this project, we want to choose the most relevant 18–20 attributes that support our focus on spatial, temporal, contextual, vehicle-based, and person-based analysis.

A. Collision-level attributes (ca. 10 key attributes)
`collision_index` (nominal ID), `date` (date), `time` (time), `day_of_week` (ordinal), `longitude` / `latitude` (continuous), `first_road_class` (ordinal categorical), `road_type` (categorical), `speed_limit` (discrete numeric), `light_conditions` (categorical), `weather_conditions` (categorical), `collision_severity` (ordinal).

B. Vehicle-level attributes (ca. 5–6 key attributes)
`collision_index`, `vehicle_reference` (composite ID), `vehicle_type` (categorical), `vehicle_manoeuvre` (categorical), `first_point_of_impact` (categorical), `sex_of_driver` (categorical), `age_band_of_driver` (ordinal).

C. Casualty-level attributes (ca. 6–7 key attributes)
`collision_index`, `vehicle_reference`, `casualty_reference` (composite ID), `casualty_class` (categorical), `casualty_type` (categorical), `age_band_of_casualty` (ordinal), `sex_of_casualty` (categorical), `casualty_severity` (ordinal), `pedestrian_location` (categorical).

These attributes should allow us to implement:

- - Time-oriented analysis (collision date and time)
- - Geospatial patterns (lat/long, local context)
- - Infrastructure context (road class, road type, speed limit)
- - Environment (light conditions, surface conditions, weather)
- - Vehicle behaviour (direction, manoeuvre, impact point)
- - Demographics (age, sex, casualty class/type)
- - Severity/outcomes (fatal/serious/slight)

In our analysis we do not fully denormalise all three tables into one mega-table. Instead, we keep three clean base tables and create task-specific views: collision-level views (Collisions only), vehicle-level views (Vehicles + Collisions), and casualty-level views (Casualties + Collisions, maybe joined to Vehicles too if we need the injured person's vehicle information).

1.3. Missingness / Quality / Comparability

This data is collected and published by the UK DfT as official statistics, with well-documented data collection and very good quality assurance. The data is very clean and consistent, so we can treat is as high-quality input for our visual analysis. There are some known limitations, such as under-reporting of some light/non-fatal collisions and over time changes in how injury severity is recorded. None of these known issues are critical for the purposes of our project.

1.4. Transformations (planned)

- Decode categorical codes using the official 2024 code list (data guide).
- Prepare demographic groupings - age bands and gender.
- Pre-aggregate tables where necessary.

Part 2 - Data-Users-Tasks

2.1. Data Characterisation

Our data is multi-table tabular data with a strong focus on time and geospatial components, with many categorical descriptors too. It consists of three related tables from the UK DfT open data (latest 5 years): Collisions, Vehicles, and Casualties.

At a high level, the data can be characterised as:

- Multidimensional: Each collision, vehicle and casualty is described by multiple dimensions:
    - Time: date, time, day_of_week, month, year.
    - Space: longitude, latitude, and administrative areas.
    - Road context: first_road_class, road_type, speed_limit, junction details.
    - Environment: light_conditions, weather_conditions, road_surface_conditions.
    - People and vehicles: casualty_type, casualty_severity, age_band_of_casualty, sex_of_casualty, vehicle_type, vehicle_manoeuvre, age_band_of_driver, etc.
- Time-oriented: Collisions and casualties are time-stamped, good for trend analysis (e.g., monthly accidents) and exploration of daily/seasonal patterns.
- Geospatial: Each collision has coordinates, so we can visualise using maps.

This structure will allow us to slice this data along many dimensions: geography, road class/type, time (day/month/year), environment (light/weather/surface), demographics (age/sex, casualty class/type) and severity (fatal/serious/slight).

2.2. Users & Domain Analysis

Our visualization focuses on two main user groups within the "road safety/risk management" domain. These groups are, in our opinion, realistic enough for the purposes of this assignment.

2.2.1. User Group 1 - government officials/local authorities/...

These are analysts and decision-makers working in local authorities, police forces, etc. They are familiar with high-level statistics and basic visualisation types from official reports but are not necessarily sophisticated data scientists. They often have limited time and need clear overviews and simple interactions, not complex dashboards.

Their needs or what they care about the most:

- Overall scale of the problem (how many casualties, how severe, changes over time)
- Where collisions and casualties occur (regions, hotspots).
- Who is affected (vulnerable road users, age and gender groups, driver/passenger/pedestrian).
- Context and risk factors (road type, speed limits, light and weather conditions).

They want a casualty-focused overview that can be filtered by severity, age group, gender, speed limit, etc.

Our first mock-up (Mock-up 1):

- central bars for "Casualties by casualty severity"
- supporting bars for "Casualties by type of vehicle"

- a monthly accidents trend line, and
- a casualties map of the UK.

This design should allow them to quickly answer questions like "Which age groups or road user types drive most of our casualty numbers?".

2.2.2. User Group 2 - insurance analysts/risk managers/...

These are data-literate analysts working for insurance companies or risk consultancies. They are comfortable with advanced data visualization tools and aggregations, and they are interested in how casualty and collision patterns translate into risk for different categories of drivers, vehicles, and regions. These could be claims analysts, pricing and risk-modelling teams, maybe managers who need compact summaries for decision-making.

Their needs or what they care about the most:

- Segmenting risk by vehicle type, vehicle engine capacity (CC), vehicle age/model, driver age band, and gender.
- Understanding how risk changes with road type, speed limit, and distance banding (how far the vehicle is from the driver's home or usual area).
- Seeing patterns like "motorcycles on rural roads at night" or "young drivers in urban areas" and how these correlate with weather and light conditions.
- Examining point of impact (front/rear/side) and casualty severity as a proxy for likely damage and claim size.

They need similar views to User Group 1 but may need some drilling down into specific segments and comparing them side-by-side. They will need filters for vehicle type, engine CC, driver age and distance banding.

2.3. Tasks & Goals

2.3.1. T1 - quick overview of casualties/severity for a selected year

Users: both groups, especially government officials and executives.

Goal: See total casualties and the split between Fatal/Serious/Slight at a glance, and how this has changed relative to the previous year.

Support: central bar chart "Casualties by casualty severity" plus the "Total casualties vs PY" summary in Mock-up 1.

2.3.2. T2 - Identify high-risk casualty groups by severity, age, gender, vehicle

Users: both groups.

Goal: Find which combinations of age band, sex, casualty class (driver, passenger, pedestrian) and vehicle type contribute most to severe outcomes (KSI). Compare drivers vs passengers vs pedestrians, car vs motorcycle vs bus vs taxi vs other.

Support: filters in the sidebar (age, casualty class, gender) and the “Casualties by type of vehicle” bar chart, all linked to the severity chart.

2.3.3. T3 - Detect and compare spatial hotspots of casualties

Users: mainly government officials.

Goal: Identify areas with unusually high casualty counts or a high proportion of severe casualties (maybe to prioritise interventions).

Support: the “Casualties Map” of the UK, filtered by severity and other attributes.

2.3.4. T4 - Explore temporal patterns in collisions and casualties

Users: both groups.

Goal: Understand how collisions and casualties evolve over months and extend this to day-of-week or time-of-day patterns, so users can see peaks, seasonal effects and unusual periods. Insurance analysts may use this to support forecasting.

Support: the “Monthly Accidents” line chart in Mock-up 1, plus potential extensions to hour-of-day / day-of-week views.

2.3.5. T5 - Compare collision risk for driver and vehicle segments

Users: mainly insurance analysts/risk managers.

Goal: Compare driver age groups in terms of collisions or casualties. Analyse how risk changes with vehicle engine CC, vehicle age, and distance banding (near home vs far from home). Relate weather conditions and vehicle point of impact to collision patterns, as a proxy for likely damage and claim severity.

Support: bar and line charts for collisions by driver age, engine-size bands, and distance bands; circle charts for weather and point of impact; filters for vehicle attributes and conditions.

2.3.6. T6 - (Advanced) Compare aggregated vs stratified views

Users: analysts in both groups.

Goal: Compare overall casualty trends with stratified trends (by age group or road type) to see cases where aggregation hides or even reverses patterns (Simpson’s Paradox).

Part 3 - Conceptual Design

3.1. Mock-up 1 - UK Road Safety 2019 – 2024

The main goal of the “Mock-up 1” design is to quickly grasp the scale of the problem, observe how the situation has changed compared to different previous years and break down the analysis across different factors. In addition, the visualizations make it possible to explore how incident severity and incident counts relate to vehicle type, gender, age, speed limit, participant class and geographic location.

Users: policymakers, analysts from ministries involved in transport security.

The dashboard sketch is presented in Appendix B. We also generated two additional digital versions of the dashboard (Appendix C) by uploading the photo of the hand-drawn sketch to ChatGPT and asking it to generate a digital representation.

3.1.1. The rationale for the design decisions

A dashboard was chosen as the primary visual communication tool allowing us to design a coherent interface for addressing users' questions. This approach allowed us to condense the required information into a limited space while preserving both flexibility and interactivity.

The dashboard uses a classic grid-based visual representation, which allows for an even distribution of visual load. There are filters on the left side, at the top is a selection of key metrics for analysis, and below are visualizations and basic information.

3.1.2. The visual encodings

There are several types of plots and tools for information presentation are used in the dashboard:

1. Horizontal Bar Chart is used twice: in the “Casualties by casualty severity” and “Casualties by casualty severity”. It allows to compare categories by number of victims. The length of the bar is a good visual channel to compare numbers, also convenient for ranking and quickly assessing differences between categories.
2. KPI Card is in the centre of the dashboard as its main element. It is used to display the total number of casualties and to compare it with the previous period in %.
3. Pie Chart is used in the “Casualties by gender” block. It is suitable for the task of comparison casualties by gender as we have only 3 categories and want to see a portion of the whole rather than an exact comparison.
4. Line Chart is used in the “Monthly Accidents” graph to show the trend during the year, additionally there is the second trend line of the previous year to show the changing.
5. Dot Density Map is used in the “Casualties Map (GB)” block. It helps to see the distribution of the casualties by regions and allows to see clusters and “hot” regions.

3.1.3. The interaction methods

We have several interaction tools on our dashboard:

1. Filters include: “Current year” and “Prior year” for selecting the analysis periods, “Speed limit” for choosing the road speed limit at the accident location, “Age groups” for filtering casualties by age, “Casualty class” for filtering by the type of participants involved in the accident and “Gender” for filtering by the gender of the individuals in the casualty records.
2. Severity selectors: “Select all”, “Fatal”, “Serious”, “Slight”, “Clear all” allow to create different views by choosing specific accident severity levels and focusing on relevant target groups.
3. Charts navigation: “Hover highlighting” to highlight the selected bar and see the additional information in the context menu. “Zooming” on the map to focus on the regions, “Brushing” on the “Monthly chart” for selecting different periods within a year.

3.1.4. Strengths, limitations, and possible improvements

The main strength of our first sketch is the ability to quickly answer typical analytical questions on the fly: who was injured, how severely, what type of vehicle, as well as the ability to focus on different data points by age, gender, participant type, etc. The dashboard also allows for comparative analysis not only for the current year but also for a selected year in historical perspective (e.g., 2024 vs. 2023, 2024 vs. 2022, 2022 vs. 2019).

Despite the strengths, the specificity of the layout being an early version, there are a number of limitations.

First, the colour scheme used in the layout is random, and the choice of colours should be consistent with the stated goals. For example, using a semantic colour palette based on the severity of the incident.

Second, the map block also requires some refinement. A legend could be added, and the map points could also be colour coded based on the severity of the incident or the density of incidents in a particular region.

Also, the nature of a single-page dashboard does not allow for an even distribution of visual elements, which can create visual noise what is a limitation of the chosen information presentation tool. To avoid visual overload, we will try to distribute visual elements more balanced.

3.2. Mock-up 2 – Vehicle Collisions

3.2.1. The rationale for the design decisions

The dashboard must be able to help risk analysts and managers to quickly identify the accident patterns and risks. Bar, Circle and line are chosen for simplicity and user-friendly interpretation.

The dashboard sketch is presented in Appendix D. We also generated two additional digital versions of the dashboard (Appendix E) by uploading the photo of the hand-drawn sketch to ChatGPT and asking it to generate a digital representation.

3.2.2. The visual encodings

There are several types of plots and tools for information presentation are used in the dashboard:

1. Line Chart is used in the "Driver Age vs Collision Frequency" to show the number of collisions across different age groups of the drivers.
2. Radial Diagram is used in the "Vehicle Point of Impact" to display impact zones (Front, Nearside, Back, Offside) around a circular layout. This helps users to easily spot which parts of vehicles are most frequently hit.
3. The Horizontal Bar Chart is used in the "Distance Banding (Risk Factor)" to illustrate the number of collisions across different distance bands (0KM, 5KM, 10KM).
4. Circular Diagram is used in the "Weather Conditions" block to present the distribution of collisions under different weather types (Fine, Raining or Snowing).

3.2.3. The interaction methods

Filtering: Analysts can filter by Vehicle Type, Vehicle Model/Age, and Vehicle Engine CC and view the data in more detail. Managers can view reports for the year and filter them by month.

Highlighting: Hovering over a data point shows tooltips with percentages, number of collisions in the line and bar chart.

3.2.4. Strengths, limitations, and possible improvements

The dashboard is designed in a straightforward way for users to explore the data. There are multiple sheets in the dashboard to provide both overall and detailed insights. With different colours for percentages to make comparisons clearly visual. Filters and drill-down give flexibility to analysts, managers, and executives to find needed insights.

While it has certain limitations such as it does not include the financial or claim-related datasets. Therefore, the dashboard focuses more on the analysis of accident risk rather than the economic impact of collisions. This limit insurance companies to connect accident frequency with the cost implications or predicting the claim payouts.

We can further improve the dashboard by integrating external datasets providing claim costs, financial exposure. It can therefore extend the dashboard scope beyond the risk assessment. Additionally, predictive analysis could be considered to forecast the future claim volumes based on seasonal trends or weather conditions, enabling insurers to plan resources and manage risk portfolios effectively.

Appendix A

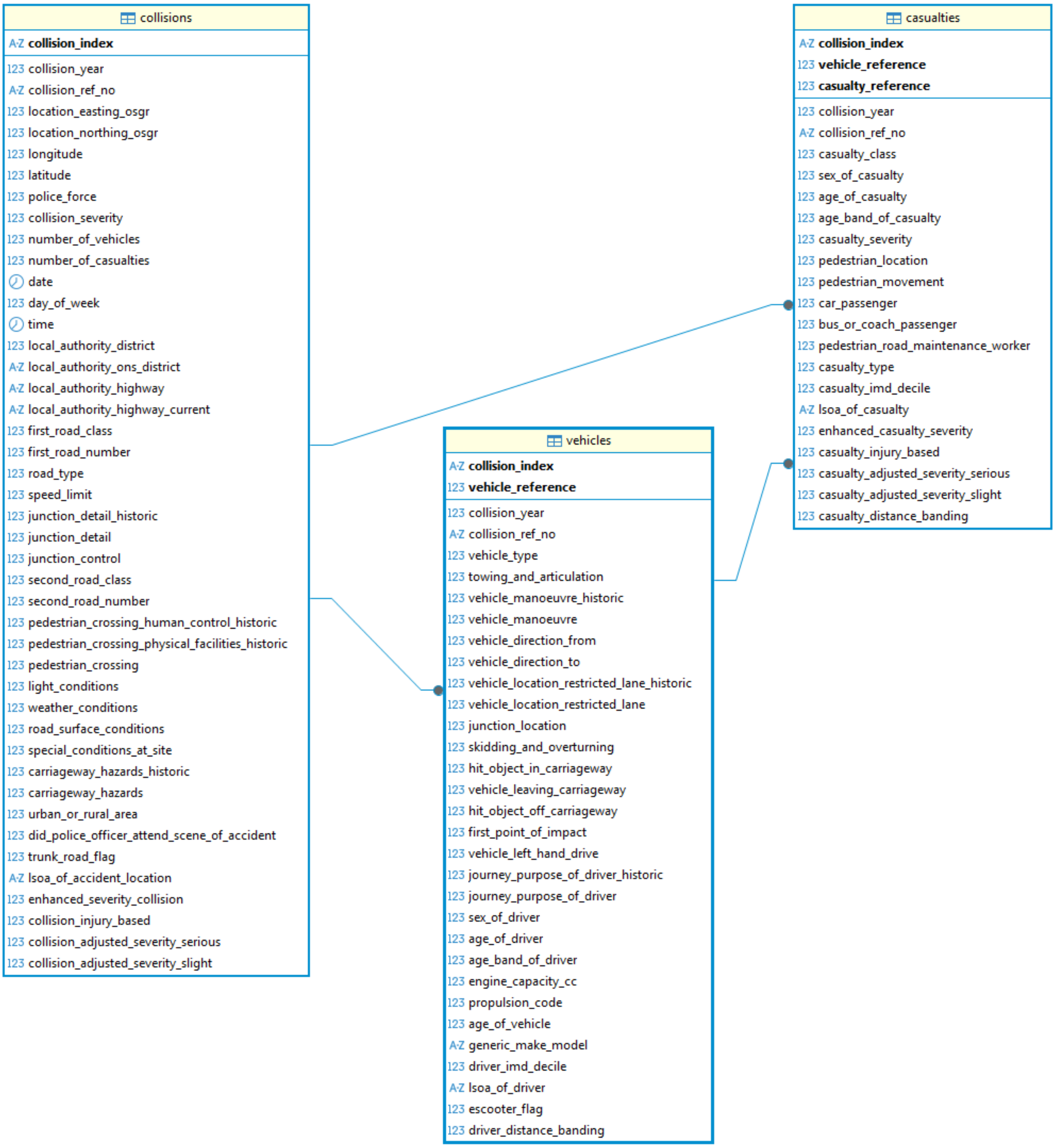
collisions
A-Z collision_index
123 collision_year
A-Z collision_ref_no
123 location_easting_osgr
123 location_northing_osgr
123 longitude
123 latitude
123 police_force
123 collision_severity
123 number_of_vehicles
123 number_of_casualties
date
123 day_of_week
time
123 local_authority_district
A-Z local_authority_ons_district
A-Z local_authority_highway
A-Z local_authority_highway_current
123 first_road_class
123 first_road_number
123 road_type
123 speed_limit
123 junction_detail_historic
123 junction_detail
123 junction_control
123 second_road_class
123 second_road_number
123 pedestrian_crossing_human_control_historic
123 pedestrian_crossing_physical_facilities_historic
123 pedestrian_crossing
123 light_conditions
123 weather_conditions
123 road_surface_conditions
123 special_conditions_at_site
123 carriageway_hazards_historic
123 carriageway_hazards
123 urban_or_rural_area
123 did_police_officer_attend_scene_of_accident
123 trunk_road_flag
A-Z lsoa_of_accident_location
123 enhanced_severity_collision
123 collision_injury_based
123 collision_adjusted_severity_serious
123 collision_adjusted_severity_slight
vehicles
A-Z collision_index
123 vehicle_reference
123 collision_year
A-Z collision_ref_no
123 vehicle_type
123 towing_and_articulation
123 vehicle_manoeuvre_historic
123 vehicle_manoeuvre
123 vehicle_direction_from
123 vehicle_direction_to
123 vehicle_location_restricted_lane_historic
123 vehicle_location_restricted_lane
123 junction_location
123 skidding_and_overturning
123 hit_object_in_carriageway
123 vehicle_leaving_carriageway
123 hit_object_off_carriageway
123 first_point_of_impact
123 vehicle_left_hand_drive
123 journey_purpose_of_driver_historic
123 journey_purpose_of_driver
123 sex_of_driver
123 age_of_driver
123 age_band_of_driver
123 engine_capacity_cc
123 propulsion_code
123 age_of_vehicle
A-Z generic_make_model
123 driver_imd_decile
A-Z lsoa_of_driver
123 escooter_flag
123 driver_distance_banding
casualties
A-Z collision_index
123 vehicle_reference
123 casualty_reference
123 collision_year
A-Z collision_ref_no
123 casualty_class
123 sex_of_casualty
123 age_of_casualty
123 age_band_of_casualty
123 casualty_severity
123 pedestrian_location
123 pedestrian_movement
123 car_passenger
123 bus_or_coach_passenger
123 pedestrian_road_maintenance_worker
123 casualty_type
123 casualty_imd_decile
A-Z lsoa_of_casualty
123 enhanced_casualty_severity
123 casualty_injury_based
123 casualty_adjusted_severity_serious
123 casualty_adjusted_severity_slight
123 casualty_distance_banding

Figure 1. ERD diagram of UK Road Safety open data

Appendix B

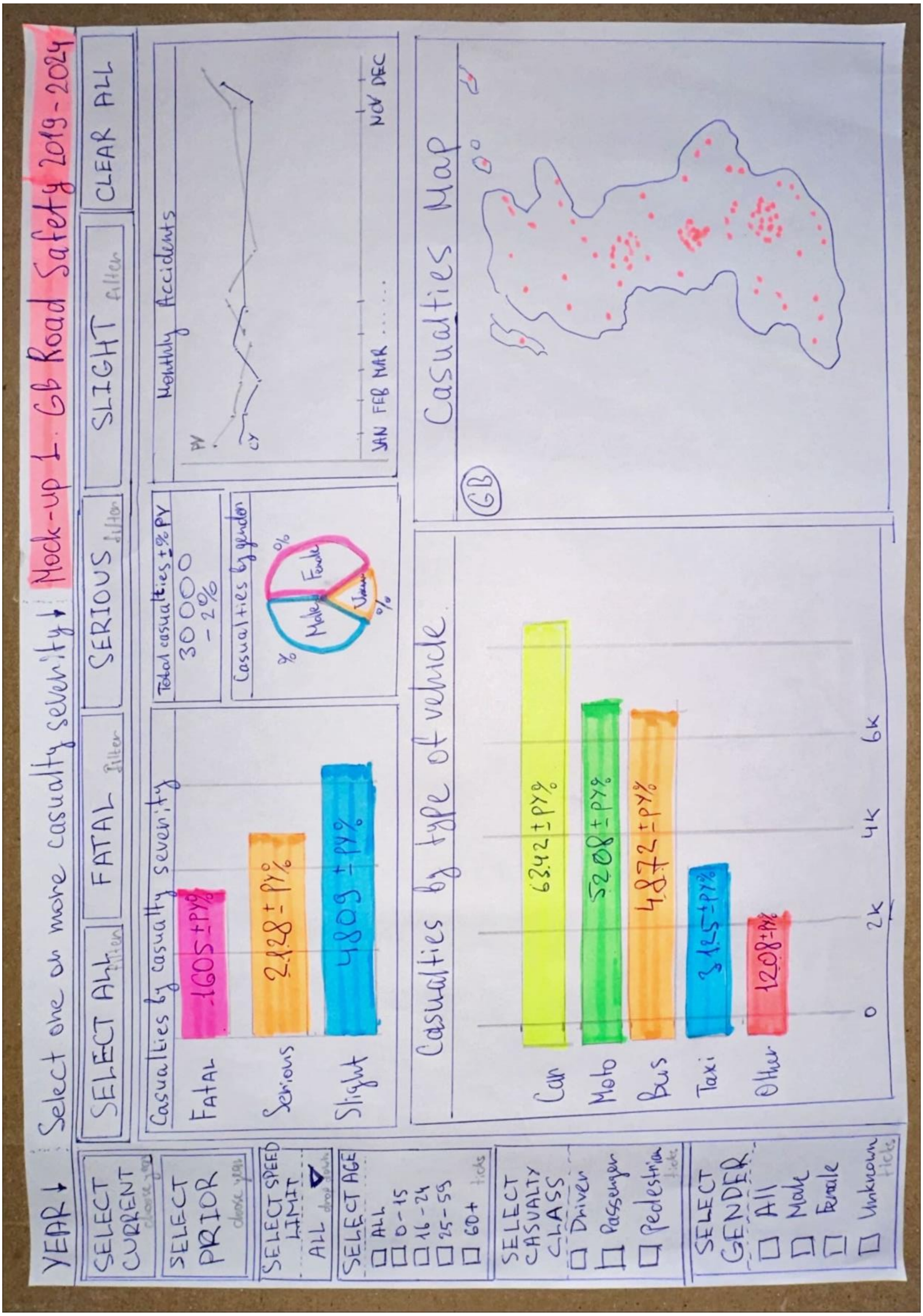

Figure 2. Mock-up 1.

Appendix C

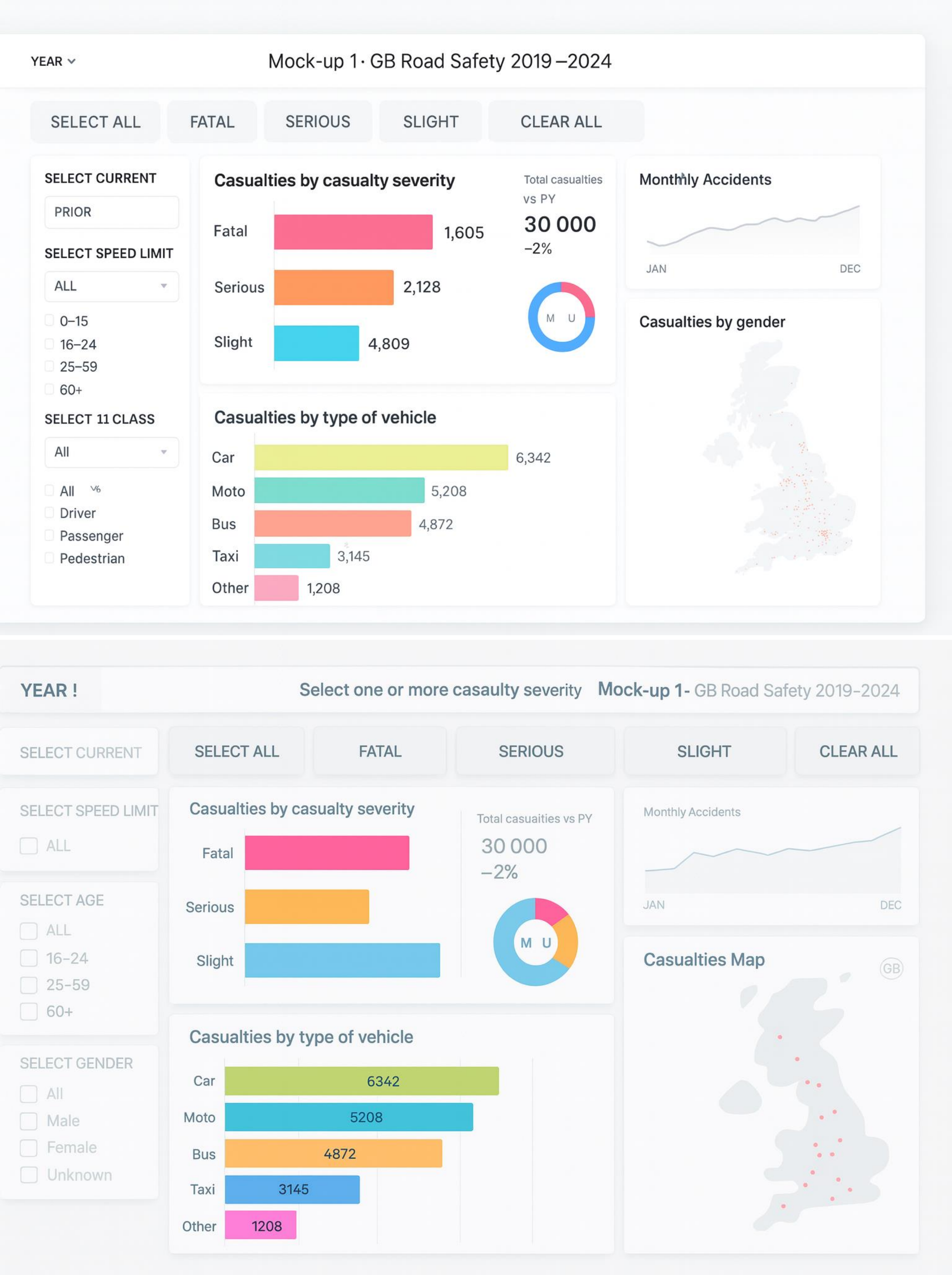

Figure 3. Digital sketches generated by ChatGPT based on “Mock-up 1”

Appendix D

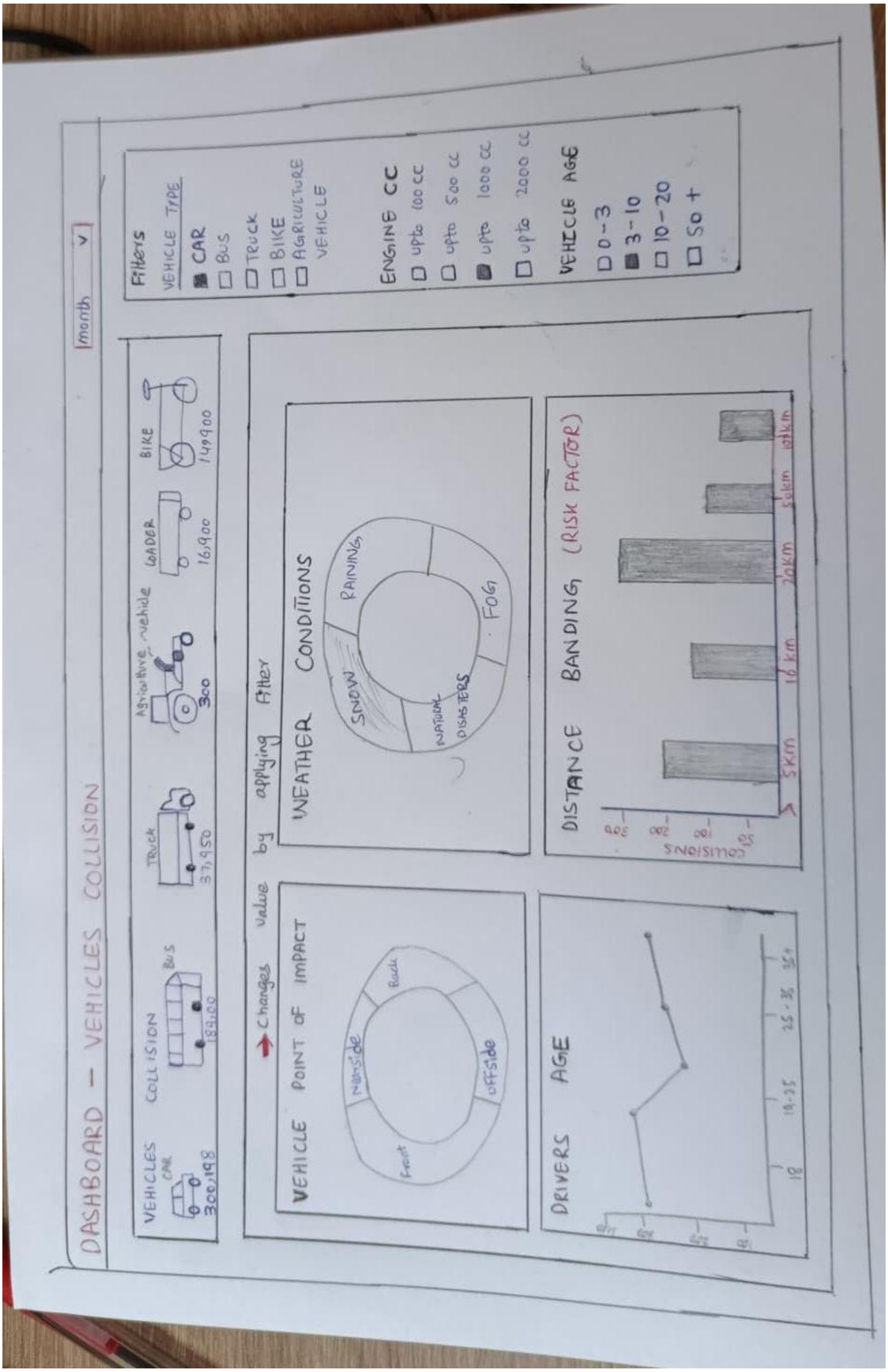

Figure 4. Mock-up 2.

Appendix E

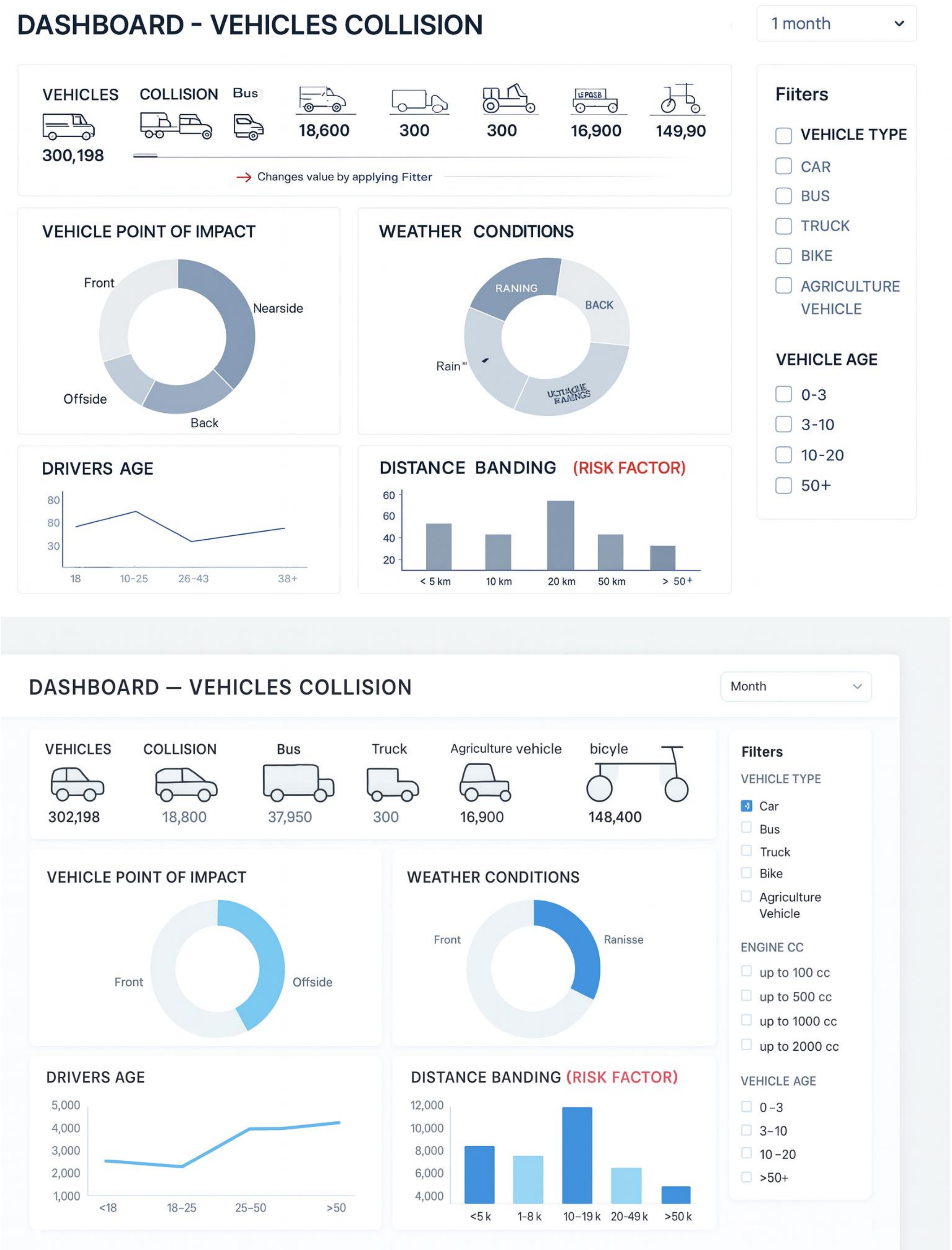

Figure 5. Digital sketches generated by ChatGPT based on “Mock-up 2”