IMPLEMENTING DATABASE SYSTEM AND DATA ANALYSIS OF ACCIDENTS BETWEEN 2016 – 2018

BY

JOSE LUIS DIAZ PULGAR

Table of Contents

Introduction:

Purpose of the project.

The purpose of this project is investigate, create and implement a system to migrate from a spreadsheet to a relational database system with suitable analysis tools in place to evaluate current performance.

To develop this complex task, I am going to divide it into several subtasks following a logical process.

Sub tasks:

- Selecting the cloud solution.
- Defining the database system.
- Obtaining the data and data wrangling.
- Interpreting data. Data visualisation and statistical data.
- Improvements using data science and machine learning.

Scenario.

We are CarMot an insurance company which is specialised in insurance of cars and motorcycles. We have two teams, administrative team who is working with a database focus on our own management system of customers, vehicles and accidents, and a second database managed for our data team.

Selecting a cloud solution.

Why do we have to migrate to the cloud?

- First reason is a cloud solution reduces the operational cost.
- Second the provider of the service helps us with the security of our data, system…
- Third reason is a cloud solution allows us to scale our business when needs change.

There different solutions. The two main solutions are Microsoft Azure and Amazon Web Service. Both of them offer us similar solutions, products and tools and levels of security.

Selecting the solution.

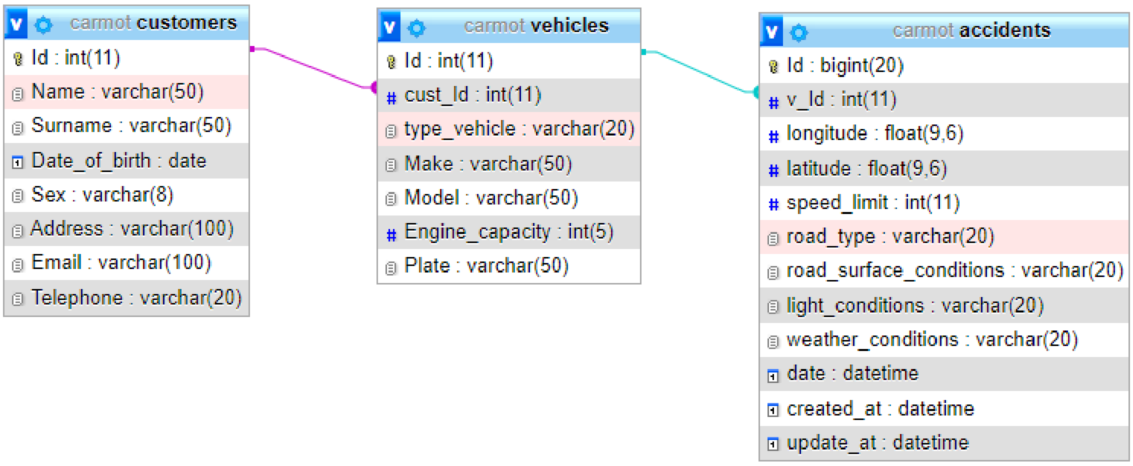
After studying pros and cons of each provider, the company decided Amazon Web Service to develop our system into the cloud.

The main reason was the process of learning this technology for people are not used to it is easy. It saves time for the company and is a less frustrating process of learning for our new employees.

Defining databases.

As it was said in the introduction, the company works with two databases one for customers and another one focused on official data. For both databases the company has chosen as Relational Database Management System ***MySQL***. It is an *open source software*, so it is *free* and according to (Solid IT, 2022) *MySQL is the second most popular database around the world.*

The relationship between different tables in the database is one to many. It means one record in one table can be associated with one or more records in another table. In the following diagram, this relation is showed for one of the databases (The another one has the same relationship).



carmot **customers**
Id : int(11)
Name : varchar(50)
Surname : varchar(50)
Date_of_birth : date
Sex : varchar(8)
Address : varchar(100)
Email : varchar(100)
Telephone : varchar(20)

carmot **vehicles**
Id : int(11)
cust_Id : int(11)
type_vehicle : varchar(20)
Make : varchar(50)
Model : varchar(50)
Engine_capacity : int(5)
Plate : varchar(50)

carmot **accidents**
Id : bigint(20)
v_Id : int(11)
longitude : float(9,6)
latitude : float(9,6)
speed_limit : int(11)
road_type : varchar(20)
road_surface_conditions : varchar(20)
light_conditions : varchar(20)
weather_conditions : varchar(20)
date : datetime
created_at : datetime
update_at : datetime

SQL is the language used to manage the databases. We have different instructions to manage our database:

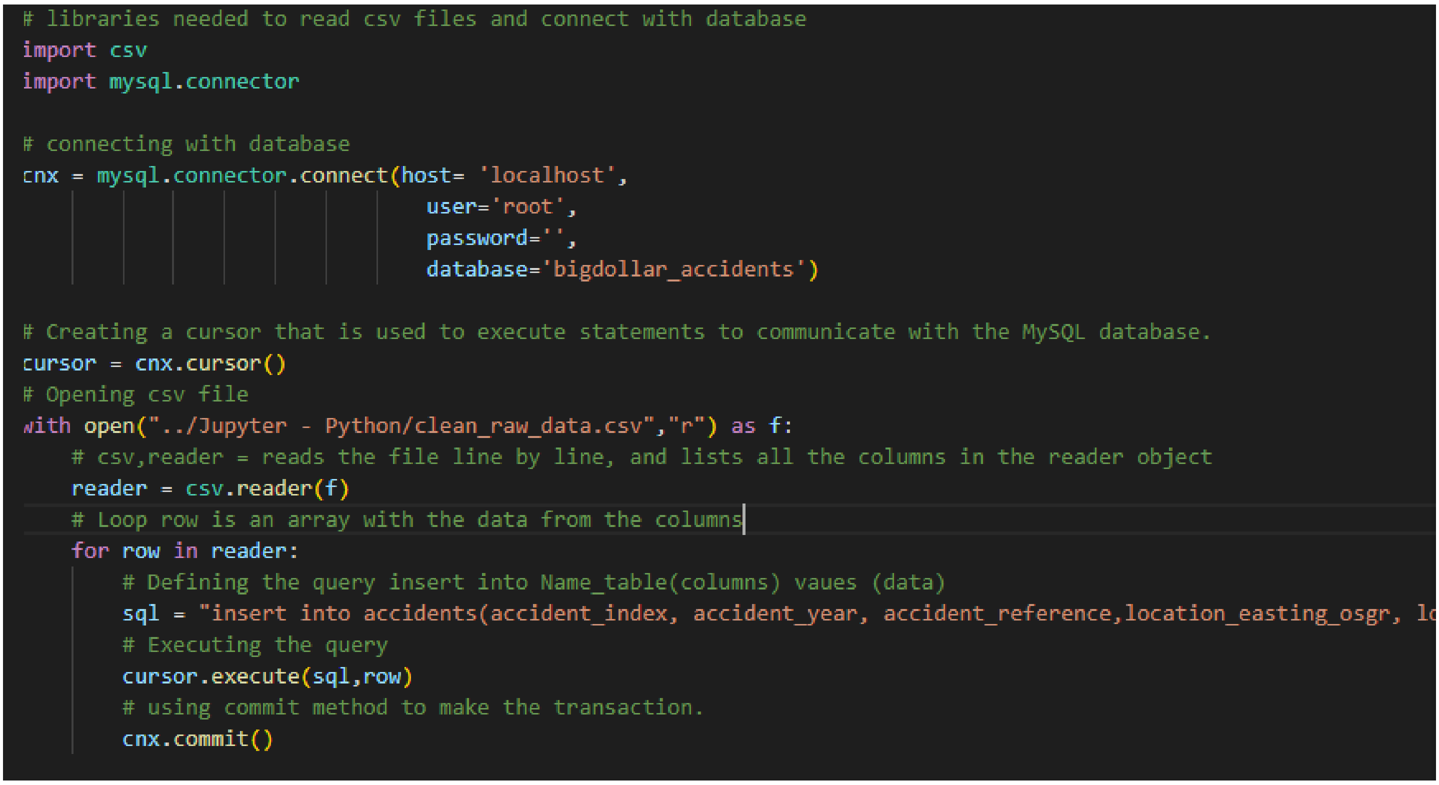
- Create a database: CREATE DATABASE Name_of_database.
- Create a table: CREATE TABLE Name_of_table (
  column1 data-type,
  column2 data-type,
  ………………..…
)

Appendix 1 SQL syntax to create tables to record spreadsheet: "*Accident 5 years.csv*" and "*vehicle 5 years.csv*"

All databases are developed in the cloud.

To migrate all data from local to the cloud we have two options meanly.

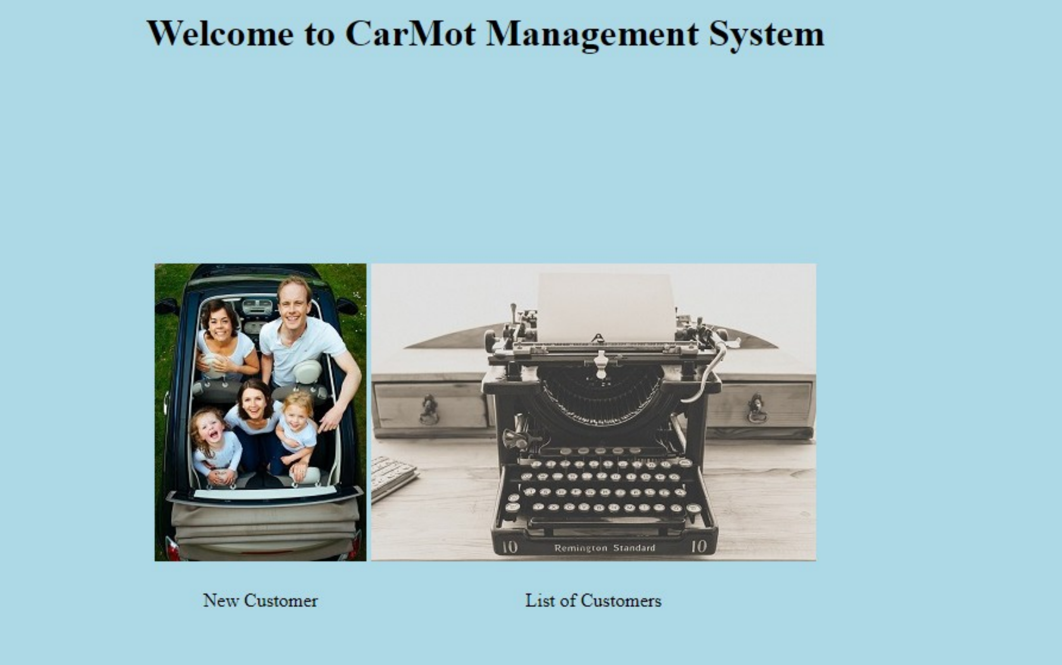
- ***PhpMyAdmin.*** It is a web application to manage MySQL databases. It is easy and intuitive, however the main disadvantage is when the system is importing large files causes us problems. For example we can not import files of more than 128 MB and we have problems when import files with a large number of rows.
- Create a script in ***Python, PHP*** to import data into the database. The process in Python is simple:



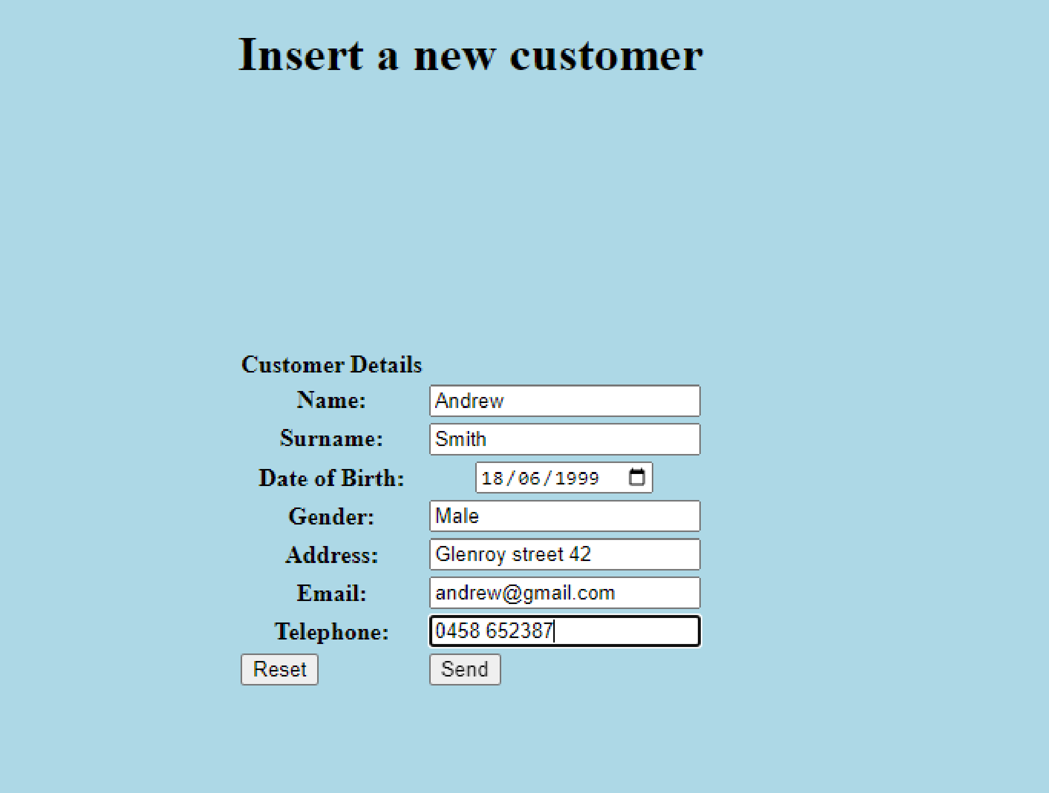
```
# libraries needed to read csv files and connect with database
import csv
import mysql.connector

# connecting with database
cnx = mysql.connector.connect(host= 'localhost',
                                user='root',
                                password='',
                                database='bigdollar_accidents')

# Creating a cursor that is used to execute statements to communicate with the MySQL database.
cursor = cnx.cursor()
# Opening csv file
with open("../Jupyter - Python/clean_raw_data.csv","r") as f:
    # csv,reader = reads the file line by line, and lists all the columns in the reader object
    reader = csv.reader(f)
    # Loop row is an array with the data from the columns
    for row in reader:
        # Defining the query insert into Name_table(columns) vaues (data)
        sql = "insert into accidents(accident_index, accident_year, accident_reference,location_easting_osgr, lo
        # Executing the query
        cursor.execute(sql,row)
        # using commit method to make the transaction.
        cnx.commit()
```

Creating a basic management system.

We create a basic management system that allow us to record new customers, vehicles and accidents.



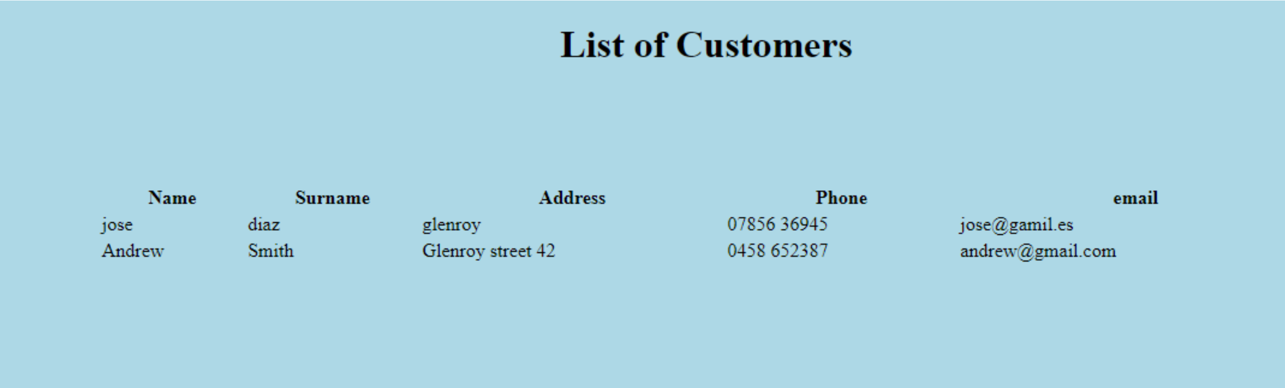
Welcome to CarMot Management System



Insert a new customer

Customer Details

Name:	Andrew
Surname:	Smith
Date of Birth:	18/06/1999
Gender:	Male
Address:	Glenroy street 42
Email:	andrew@gmail.com
Telephone:	0458 652387



List of Customers

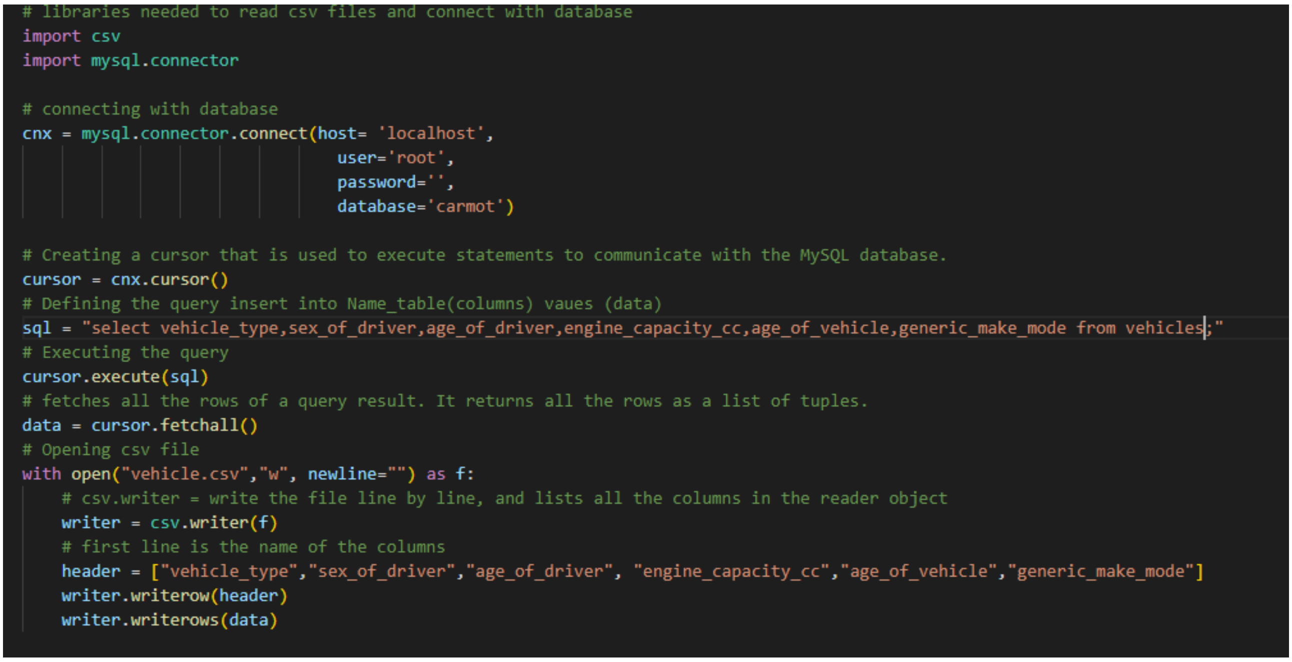
Name	Surname	Address	Phone	email
jose	diaz	glenroy	07856 36945	jose@gamil.es
Andrew	Smith	Glenroy street 42	0458 652387	andrew@gmail.com

Obtaining data and data wrangling.

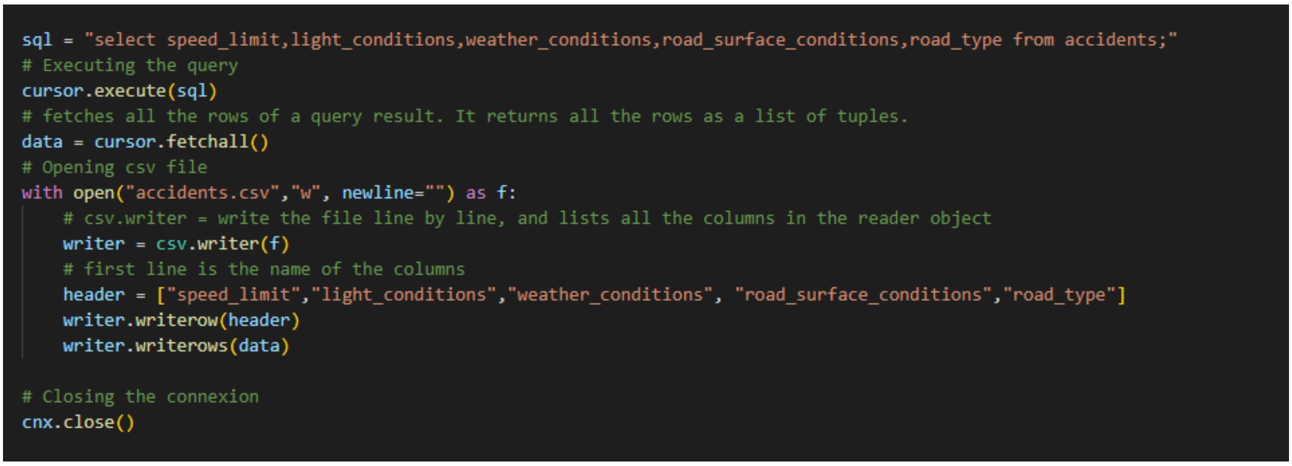
Obtaining data.

Once we have all data recorded in our database in the cloud, we are going to select what data we need.
The SQL syntax is simple: SELECT column_1, column_2,…. FROM Name_of_table

We need to store it into a csv file. The process is showed in the next figure.



```
# libraries needed to read csv files and connect with database
import csv
import mysql.connector

# connecting with database
cnx = mysql.connector.connect(host= 'localhost',
                                user='root',
                                password='',
                                database='carmot')

# Creating a cursor that is used to execute statements to communicate with the MySQL database.
cursor = cnx.cursor()
# Defining the query insert into Name_table(columns) vaues (data)
sql = "select vehicle_type,sex_of_driver,age_of_driver,engine_capacity_cc,age_of_vehicle,generic_make_mode from vehicles;"
# Executing the query
cursor.execute(sql)
# fetches all the rows of a query result. It returns all the rows as a list of tuples.
data = cursor.fetchall()
# Opening csv file
with open("vehicle.csv","w", newline="") as f:
    # csv.writer = write the file line by line, and lists all the columns in the reader object
    writer = csv.writer(f)
    # first line is the name of the columns
    header = ["vehicle_type","sex_of_driver","age_of_driver", "engine_capacity_cc","age_of_vehicle","generic_make_mode"]
    writer.writerow(header)
    writer.writerows(data)
```



```
sql = "select speed_limit,light_conditions,weather_conditions,road_surface_conditions,road_type from accidents;"
# Executing the query
cursor.execute(sql)
# fetches all the rows of a query result. It returns all the rows as a list of tuples.
data = cursor.fetchall()
# Opening csv file
with open("accidents.csv","w", newline="") as f:
    # csv.writer = write the file line by line, and lists all the columns in the reader object
    writer = csv.writer(f)
    # first line is the name of the columns
    header = ["speed_limit","light_conditions","weather_conditions", "road_surface_conditions","road_type"]
    writer.writerow(header)
    writer.writerows(data)

# Closing the connexion
cnx.close()
```

Data wrangling

Data wrangling is a concept that involve the processes of cleaning, structuring and enriching raw data. Before starting with this process, the data team needs set up a strategy to find out what they are looking for.

In our case, we are looking for:

- Getting a “global vision” of what happened between years 2016 – 2020. We are analysing the trends of the number of accidents through the years, the weekday, the time of the day… We need a general idea and we do not need numbers, so the best tool for this purpose is a visual analytic platform as Tableau or Power BI. In this case, we are using tableau, but Power BI would be valid as well.
- We are an insurance company, so we need to understand our customers and their behaviour when they drive. For this purpose, we will use Python and R because we need analyse the data with descriptive statistic.
- The external condition are important as well. Conditions such as speed limit, weather conditions, road surface conditions, road type… help us to know where the accidents occur and why.

The next step is to analyse what type of data we have. In the sources provided, we have two types of data:

- Qualitative data: sex of the driver, make and model of the vehicle, weekday…
- Quantitative data: age of the driver, engine capacity, age of vehicle…

We apply different tools of descriptive statistic for different types of data. This point will be expanded in future sections.

The next step in this process of data wrangling is to deal with lack of information or wrong and nonsense values. In this step, I decided to delete them because we have a large number of values and we maintain enough data to analyse. Other options could be to replace them with the mean, the median or the mode.

Data visualisation.

Global analysis.

As it was said previously, in this type of analysis we use tableau visualizations. **The conclusions** we get from them can be summarise in:

- Trough the years the number of accidents has significant variations from year to year. It is significant the year 2020 with a -20% of variation respect to the year 2019. COVID explains this variation. **Appendix 2 – Figure 1**.

- The trend by the weekday is similar with different values for each year. **Appendix 2 – Figure 2**.

- We have an increment number of accidents between May and July and the last months of the year. **Appendix 2 – Figure 3**.

- The trend of the number accidents is different from Monday to Friday to weekends. From Monday to Friday we have an increment in the number of accidents coinciding with the hours of commuting to work eight o'clock in the morning and 5 o'clock in the afternoon. Some explanations are we have more vehicles in the road, in the morning people drive faster to get to work on time, in the afternoon people are tired and therefore suffer a decrease in their reflexes and distractions are more frequent. **Appendix 2 – Figure 4**.

- If we break down the number of accidents by the weekday through the months in each year, we can find some similar patterns in different years. **Appendix 2 – Figure 5**.

Particular analysis.

We are analysing:

- The characteristics of the people and vehicles that have an accident. We are working with the following columns from the dataset: ***sex_of_driver, age_of_driver, vehicle_type, generic_make_model, engine_capacity_cc, age_of_vehicle***.

- The conditions when an accident happened. We are working with the following columns: **road_type, speed_limit, light_conditions, weather_conditions, road_surface_conditions**.

Methodology.

Quantitative data.

1. Obtain in R studio a histogram, boxplot graph and density function graph. (R)
2. Clean the data deleting outliers and non sense values. (Python and R)
3. Obtain the graphs again as the step 1. (R)
4. Obtain the main statistics and additional graphs.(R and Python)

Qualitative data.

1. Clean the data deleting non sense values. (Python)
2. Obtain table of frequencies and proportions in R studio. (R)
3. Obtain graphs from the results obtained in the previous step. (Python and R)

Analysis of the variables.

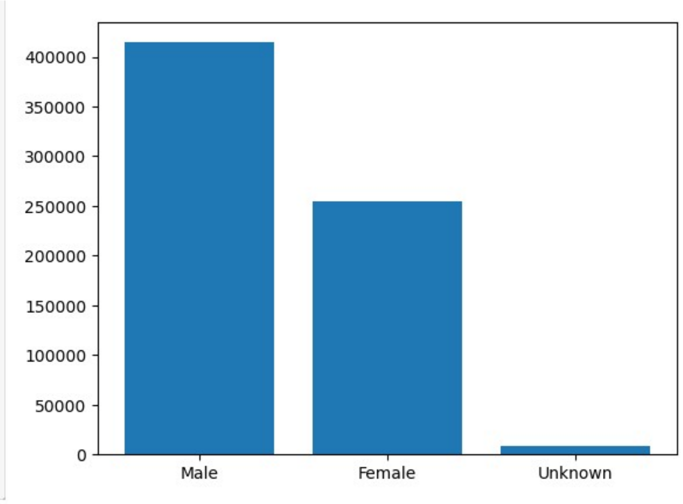
Accidents by sex.

Cars.

Since there is a greater proportion of men than women who drive cars, it is logical to think that they will suffer more accidents. In the period between 2016 – 2020, we can observe that in 56.04 % of accidents men are involved, 33.47% correspond to women and we have around 10% where sex is not defined.

From the data analysed we obtain this table of frequencies for the period 2016 – 2020.

Accidents: 2016 – 2020		
Frequencies		**Proportion %**
Male	446804	56.04
Female	266905	33.47
Unknown	83656	10.49

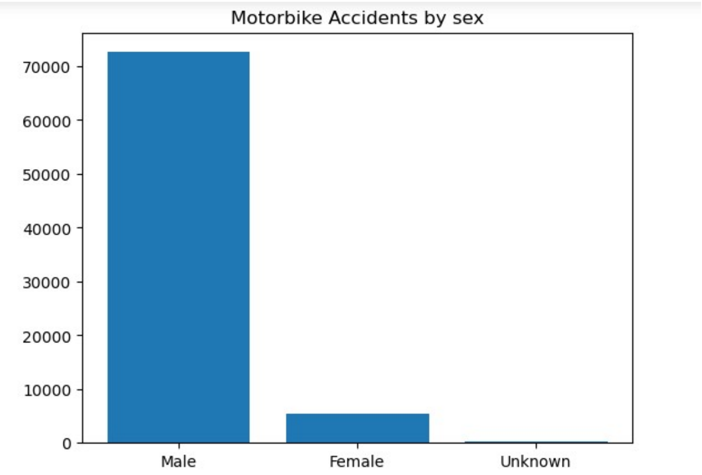


Motorcycles.

In this case we can say the majority of accidents are by men. One explanation to this is the number of male drivers is superior to female drivers.

From the data analysed we obtain this table of frequencies for the period 2016 – 2020.

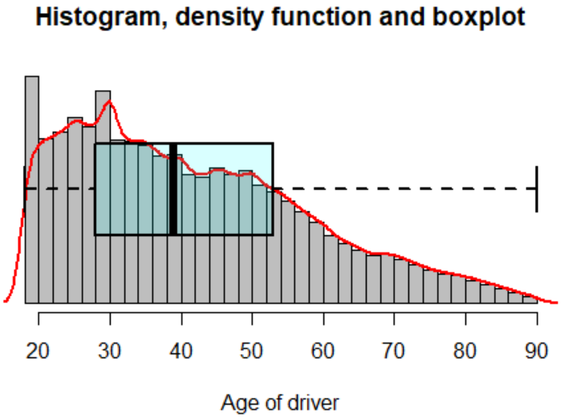
Frequencies		**Proportion %**
Male	**79437**	**91.00**
Female	**5845**	**6.69**
Unknown	**2014**	**2.31**



Accidents by age.

Cars.

After cleaning the data, and set a valid range for the age, we obtain that most of the accidents happened among people who were under 40 years of age.



In the following table we can see the fifteen age ranges that have more accidents.

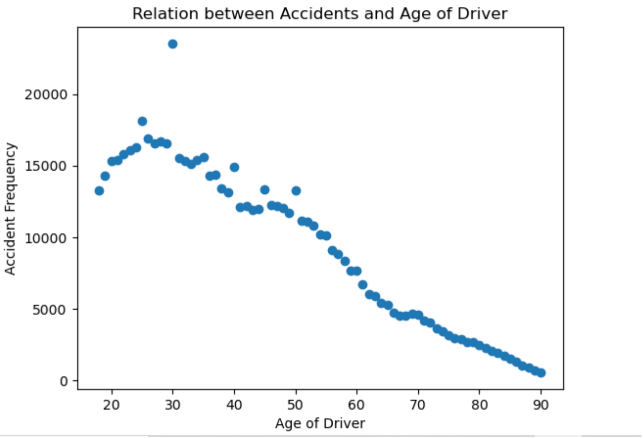
Age of driver	Proportion %
30	3.47
25	2.68
26	2.49
28	2.46
27	2.44
29	2.44
24	2.40
23	2.37
22	2.33
35	2.30
31	2.29
34	2.27
21	2.27
32	2.26
20	2.26

We can say that young people have more probability of having an accident than the rest age ranges and this is because the price for car insurance is greater in young people than mature people.

In the graph on the right side we can see a strong relationship between number of accidents and age.

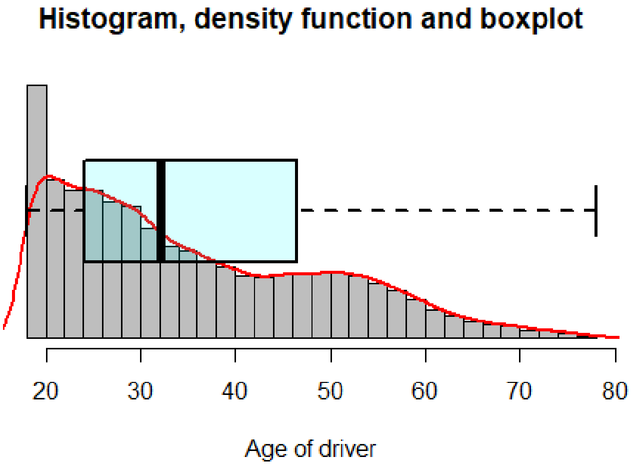
The main statistic for this variable are:

statistics	
mean	41.67
median	39
standard deviation	16.53
mode	30



Motorcycles.

After cleaning the data, and set a valid range for the age, we obtain that most of the accidents happened among people who were under 32 years of age. There is 8 years of difference between car drivers and motorcycle drivers.

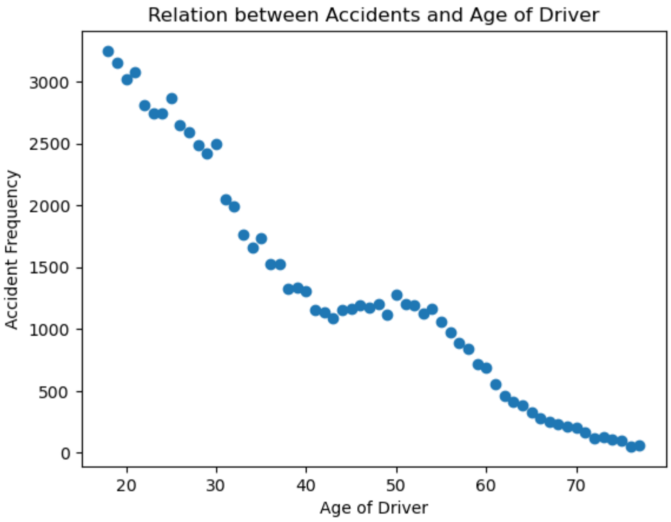


The main statistics are:

statistics	
mean	35.46
median	32
standard deviation	13.72
mode	18

Inexperienced drivers have more accidents than advanced drivers.

Age of driver	Proportion %
18	4.16
19	4.04
21	3.94
20	3.86
25	3.67
22	3.60
24	3.52
23	3.51
26	3.39
27	3.31
30	3.19
28	3.18



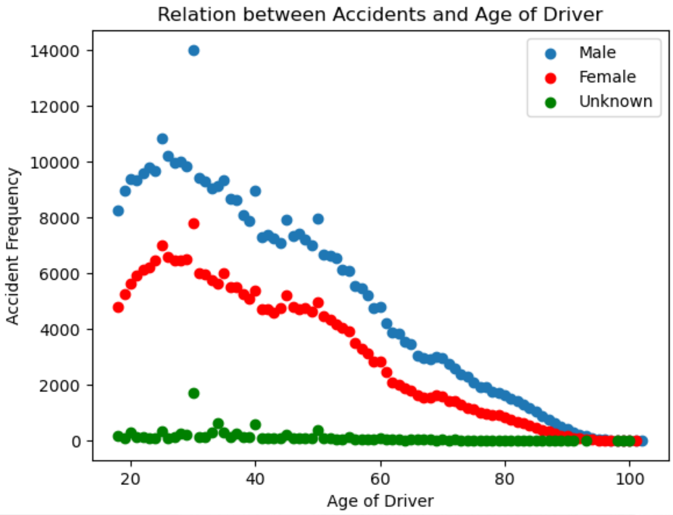
Accidents by sex and age.

Cars.

Combining both variables:

At first glance we observe that there is a difference between the values for both sexes, however the trend is similar in both.

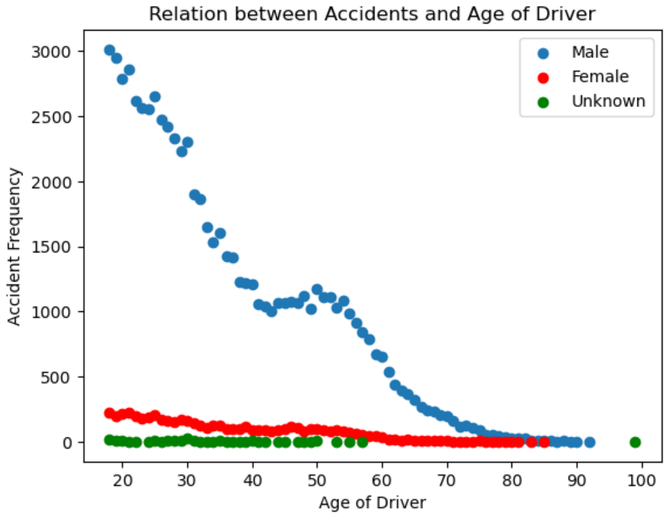
Men have more accidents than women, in both young people have more frequency than mature people. This analysis supports the two previous one done.



The main statistic for both sexes are:

			quantiles				
Sex of driver	**mean**	**standard deviation**	**0%**	**25%**	**50%**	**75%**	**100%**
Male	42.15	16.97	18	28	39	53	102
Female	41.3	16.18	18	28	39	52	101
Unknown	36.9	12.78	18	30	34	43	100

Motorcycles.



It is totally different respect to the graph for cars. The trend for female remains quasi horizontal through the years, while the trend for men starts in high values and drops through the years.

The main statistics are:

			Quantiles				
sex of driver	**mean**	**standard deviation**	**0%**	**25%**	**50%**	**75%**	**100%**
Male	35.65	13.99	18	24	32	46	92
Female	35	12.91	18	24	32	45	85
Unknown	31.23	10.81	18	24	30	36	99

Accidents by Engine capacity and make model.

We have two types of data:

- Engine capacity → Quantitative data
- Make model → Descriptive data

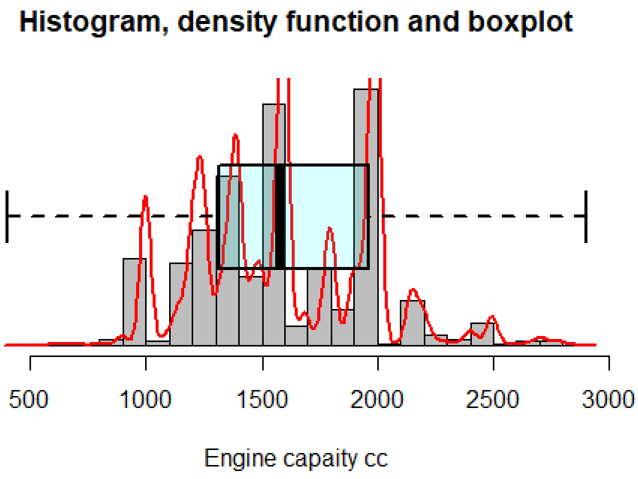
cars.

Engine capacity:

After cleaning the data, and set a valid range for the engine capacity, we obtain that most of the accidents happened among engine capacity between 1100 cc and 2000 cc. With two engine capacity ranges that have more important than the others: ranges 1500-1600 and 1900-2000

In the following table we can see the fifteen engine capacities that have more accidents.

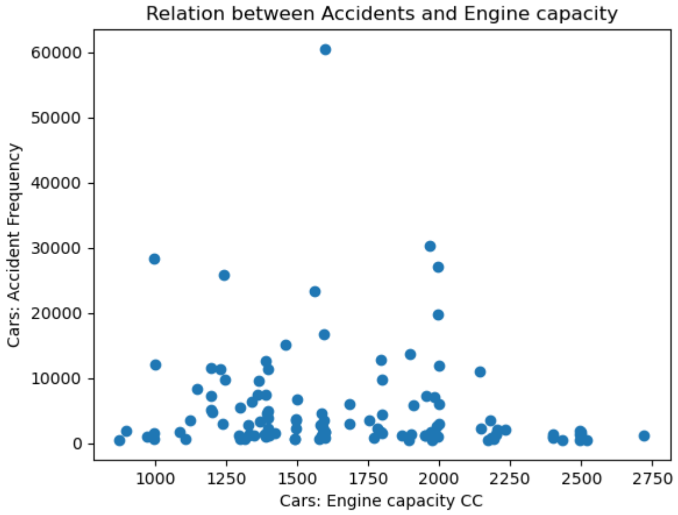
Engine Capacity CC	Proportion %
1598	9.09
1968	4.55
998	4.25
1995	4.06
1242	3.9
1560	3.52
1997	2.96
1596	2.52
1461	2.29
1896	2.06
1796	1.93
1390	1.89
999	1.81
1998	1.8
1198	1.73



The main statistics for this variable are:

statistics	
mean	1605.25
median	1596
standard deviation	368.35
mode	1598

As we can see in the next graph does not exist a relationship between engine capacity and the number of accidents.



Most of the values have are in the same range of number accidents, except the value of 1598 cc that have a big difference with the rest of the values.

Make model:

We are using a descriptive variable so we are calculating its frequency. From the results obtained using R studio, we can say the five make cars which have more accidents are: Ford, Vauxhall, Volkswagen, BMW, and Toyota.

Make	*Frequency*
BMW	7433
FORD	12356
TOYOTA	6391
VAUXHALL	11890
VOLKSWAGEN	9024

And the model with have more accidents for each make is:

Make	*Frequency*
BMW 3 SERIES	2535
FORD FIESTA	4826
TOYOTA YARIS	1642
VAUXHALL CORSA	4051
VOLKSWAGEN GOLF	3695

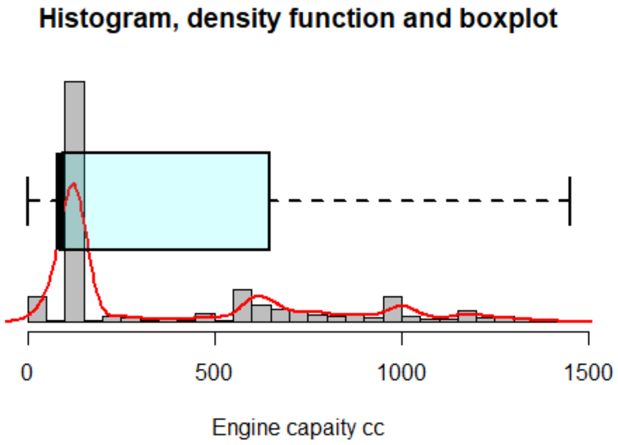
Motorcycles.

Engine capacity.

Once we have cleaned the data, we obtain the following histogram and we can see at first glance the motorcycles of 125 cc have the majority of the number of accidents. The two next ranges of engine capacity are around 600 cc and 1000cc.

If we assume 124 cc as 125 cc, motorcycles with 125 cc of engine capacity have almost 50% of accidents and we see in future graphs, the engine capacity have a huge influence in the number of accidents.

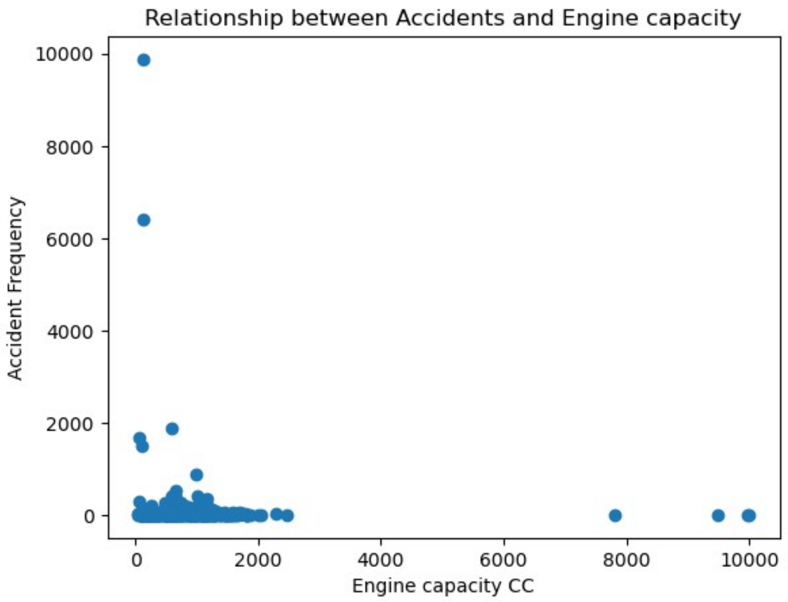
Engine Capacity CC	Proportion %
125	28.92
124	18.88
599	5.50
49	4.77
108	4.36
998	2.59
649	1.53
645	1.24
600	1.24
999	1.22
1170	1.05



The main statistics for this variable are:

statistics	
mean	402.85
median	125
standard deviation	410.09
mode	125

As I have previously mentioned, in this case the engine capacity have a huge influence in the number of accidents and it can be verified with the following graph.



While most of the values are distributed in the same area, the values of 124 cc and 125 cc scape from it.

Make model.

Analysing this qualitative variable we obtain the frequencies for the models of motorcycles.

The five make cars which have more accidents are: Honda, Yamaha, Suzuki, Piaggio and Kawasaki

Make	***Frequency***
HONDA	3901
KAWASAKI	420
PIAGGIO	460
SUZUKI	645
YAMAHA	2334

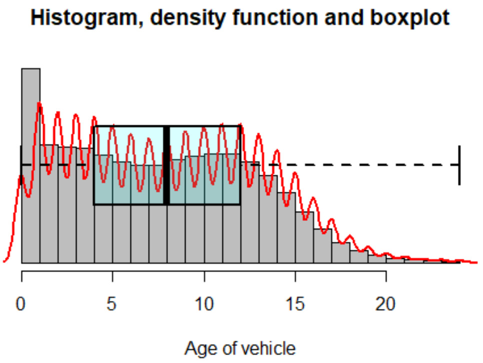
Model with higher engines have less accidents Kawasaki with 1000 cc had just 52 accidents in the period of five years while Yamaha with an engine of 125 cc had 772 accidents. Again we can see engine capacity has a huge influence in the number of accidents.

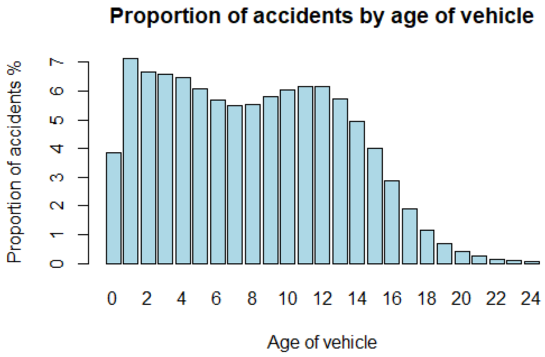
Model	**Frequency**
HONDA WW125	519
KAWASAKI ZX1000	52
PIAGGIO VESPA	170
SUZUKI GSXR 600	68
YAMAHA GPD 125	772

Accidents by age of vehicle.

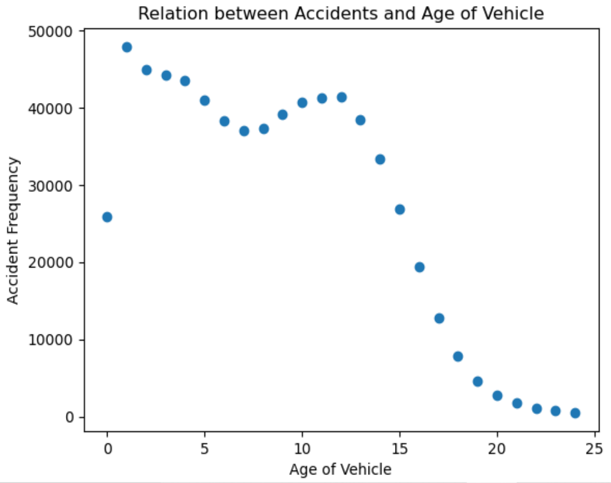
Cars.

After cleaning the data, and set a valid range for the vehicle years old, we obtain that cars with less years old are more prone to accidents. During the second year (1 year – 2 years), we have a 7,12% of the accidents, however the trend is similar between 1 year and 13 years with values between 6% and 7%.





The following table shows the frequency of the number of accidents by the age of vehicle. We can see the proportion are similar in the first years.



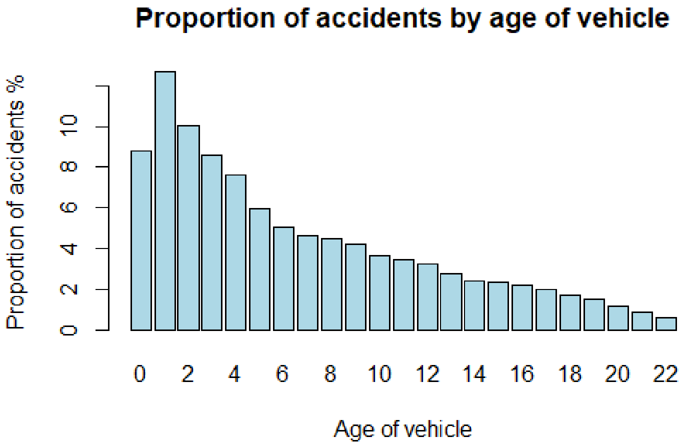
Age of vehicle	Frequency	Proportion %
1	47906	7.12%
2	44867	6.67%
3	44197	6.57%
4	43586	6.48%
12	41370	6.15%
11	41333	6.14%
5	40983	6.09%
10	40726	6.05%
9	39157	5.82%
13	38414	5.71%
6	38279	5.69%
8	37278	5.54%
7	37058	5.51%

The main statistics are:

statistics	
mean	8.05
median	8
standard deviation	5.1
mode	1

Motorcycles.

As we can see in the following table of frequencies, more than 50% of accidents happened in the first five years of motorcycle age, three years less than the car age. If we see the graph, the shape is similar to the right skewed normal distribution.

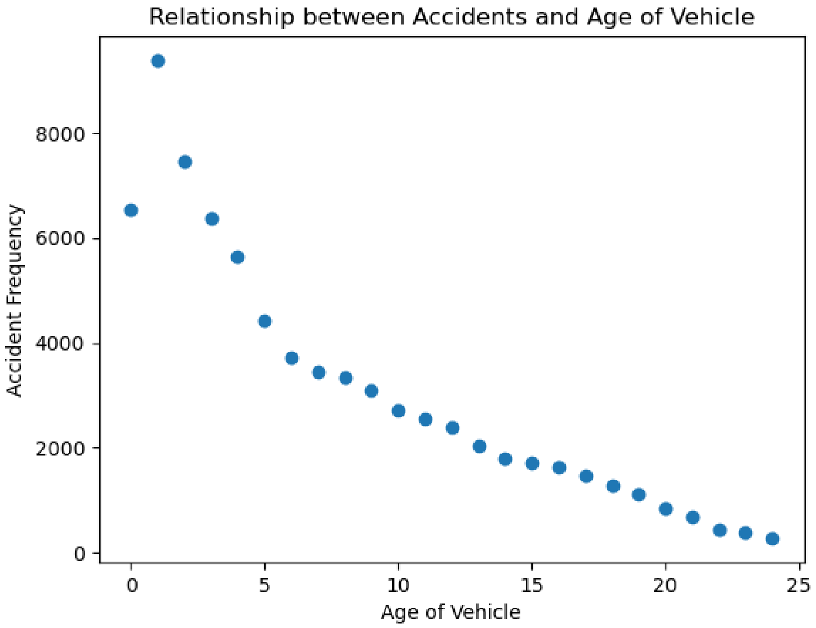


Age of vehicle	Frequency	Proportion %
1	9384	12.66%
2	7449	10.05%
0	6536	8.82%
3	6380	8.61%
4	5635	7.60%
5	4421	5.97%
6	3721	5.02%
7	3447	4.65%
8	3342	4.51%
9	3099	4.18%
10	2716	3.67%

The main statistics for this variable are:

statistics	
mean	6.57
median	5
standard deviation	5.68
mode	1

Checking the relationship between the age of vehicle and motorbike, there is a strong relationship between the two variables.



Summary.

- For both cars and motorcycles men have more accidents than women. This difference is greater in the case of motorcycles.

- In 1 of every 4 accidents people under 28 years old are involved in the case of cars and under 24 years old in the case of motorcycles. It is supported by statistic (Quantile 25%)

- Young people have more accidents with cars, however the age mean is 41.67 while in the case of motorcycles is 35.46

- While the trend of the number of accidents for both sexes through the age is similar in the case of car drivers, in the case of motorcycles drivers is totally different.

- For both cases there is a value of engine capacity which has a huge influence in the number of accidents around 1600 cc for cars and 125 cc for motorcycles. However there is no relationship between the engine power and the number of accidents. In other words, a larger engine capacity does not mean an increase in the number of accidents.

- The previous point is reinforced if we analyse the make and the model. Ford fiesta have the largest number of accidents in cars, while in the case of motorcycles is Yamaha GPD 125.

- In the case of the age of the vehicle the number of accidents is concentrated in the first years of the vehicle's age. The mean of car's age is 8.05 years and motorcycle's age is 6.57. While the trend in the case of cars is similar in the first years, it drops in the case of motorcycles.

To conclude, as insurance company our target with more risk would be a thirty years old female who drives a year old Ford fiesta or an eighteen years old male who drives a year old Yamaha GPD 125.

Road and weather conditions.

The following statements are supported by the different tables of frequencies and proportions that are shown in the **appendix 3**.

- More than 50% of accidents occur in road with a speed limit of 30 mph. It means more of then occur in built up areas, where we have roads with street lighting.

- Related to the previous point we have around 21% of accidents during the periods of darkness but with street lights present, again in built up areas.

- More than 70% accidents occur during daylight.

- 80% of accidents occur with good weather conditions and this variable is related to road surface conditions that says to us that 71% of accidents with a dry road surface.

Improvements using data science and machine learning.

As insurance company our target is customers and their behaviour.

We focus on:

- Price of insurance. How can offer a better price not only for new customers but for old ones.
- Customer behaviour. Analysing and determining variables that define the driving style of our customers.
- Determine the probability of having an accident by our customers.

The company will use several machine techniques to carry out data science task and they can be summarised as follows:

- *Multiple linear regression* to predict the price of new customers. We use several explanatory variables to to predict the outcome of a response variable. In this case the response variable is the price of insurance, and the explanatory variable could be variables such as age of driver, gender of driver, where they live, make model and age of vehicle…

- Related to the previous point we can group people in different groups using techniques of *clustering*.

- *Logistic regression*. This technique is useful for modelling the probability of an event occurring as a function of other factors. For example we can calculate the probability of accident for our customers, or another example is we want to offer a new service for our customers we can analyse the probability that they will acquire the new service or not. We could analyse variables such as age of customer, level of studies, family income, where they live, when the company phone them how receptive are to a telephone conversation…

- *Decision trees*. We try to find the variable that allows dividing the dataset into logical groups that are most different from each other.

Appendix 1

```
CREATE TABLE IF NOT EXISTS `accidents` (
 `accident_index` varchar(50) NOT NULL,
 `accident_year` int(10) DEFAULT NULL,
 `accident_reference` varchar(20) DEFAULT NULL,
 `location_easting_osgr` float DEFAULT NULL,
 `location_northing_osgr` float DEFAULT NULL,
 `longitude` float DEFAULT NULL,
 `latitude` float DEFAULT NULL,
 `police_force` int(11) DEFAULT NULL,
 `accident_severity` int(11) DEFAULT NULL,
 `number_of_vehicles` int(11) DEFAULT NULL,
 `number_of_casualties` int(11) DEFAULT NULL,
 `date` varchar(20) DEFAULT NULL,
 `day_of_week` int(11) DEFAULT NULL,
 `time` varchar(10) DEFAULT NULL,
 `local_authority_district` int(11) DEFAULT NULL,
 `local_authority_ons_district` varchar(20) DEFAULT NULL,
 `local_authority_highway` varchar(20) DEFAULT NULL,
 `first_road_class` int(11) DEFAULT NULL,
 `first_road_number` int(11) DEFAULT NULL,
 `road_type` int(11) DEFAULT NULL,
 `speed_limit` float DEFAULT NULL,
 `junction_detail` int(11) DEFAULT NULL,
 `junction_control` int(11) DEFAULT NULL,
 `second_road_class` int(11) DEFAULT NULL,
 `second_road_number` int(11) DEFAULT NULL,
 `pedestrian_crossing_human_control` int(11) DEFAULT NULL,
 `pedestrian_crossing_physical_facilities` int(11) DEFAULT NULL,
 `light_conditions` int(11) DEFAULT NULL,
 `weather_conditions` int(11) DEFAULT NULL,
 `road_surface_conditions` int(11) DEFAULT NULL,
 `special_conditions_at_site` int(11) DEFAULT NULL,
 `carriageway_hazards` int(11) DEFAULT NULL,
 `urban_or_rural_area` int(11) DEFAULT NULL,
 `did_police_officer_attend_scene_of_accident` int(11) DEFAULT NULL,
 `trunk_road_flag` int(11) DEFAULT NULL,
 `lsoa_of_accident_location` varchar(20) DEFAULT NULL,
 PRIMARY KEY (`accident_index`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

```
CREATE TABLE IF NOT EXISTS `vehicles` (
 `accident_index` varchar(50) DEFAULT NULL,
 `accident_year` int(10) DEFAULT NULL,
 `accident_reference` varchar(20) DEFAULT NULL,
 `vehicle_reference` int(4) DEFAULT NULL,
 `vehicle_type` int(4) DEFAULT NULL,
 `towing_and_articulation` int(4) DEFAULT NULL,
 `vehicle_manoeuvre` int(4) DEFAULT NULL,
 `vehicle_direction_from` int(4) DEFAULT NULL,
 `vehicle_direction_to` int(4) DEFAULT NULL,
 `vehicle_location_restricted_lane` int(4) DEFAULT NULL,
 `junction_location` int(4) DEFAULT NULL,
 `skidding_and_overturning` int(4) DEFAULT NULL,
 `hit_object_in_carriageway` int(4) DEFAULT NULL,
 `vehicle_leaving_carriageway` int(4) DEFAULT NULL,
 `hit_object_off_carriageway` int(4) DEFAULT NULL,
 `first_point_of_impact` int(4) DEFAULT NULL,
 `vehicle_left_hand_drive` int(4) NOT NULL,
 `journey_purpose_of_driver` int(4) NOT NULL,
 `sex_of_driver` int(2) NOT NULL,
 `age_of_driver` int(4) NOT NULL,
 `age_band_of_driver` int(4) NOT NULL,
 `engine_capacity_cc` int(6) NOT NULL,
 `propulsion_code` int(4) NOT NULL,
 `age_of_vehicle` int(4) NOT NULL,
 `generic_make_model` varchar(50) NOT NULL,
 `driver_imd_decile` int(4) NOT NULL,
 `driver_home_area_type` int(4) NOT NULL,
 KEY `FK_index` (`accident_index`),
CONSTRAINT `FK_index` FOREIGN KEY (`accident_index`) REFERENCES `accidents`
(`accident_index`) ON UPDATE CASCADE ON DELETE CASCADE

) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4;
```

Appendix 2

Figure 1

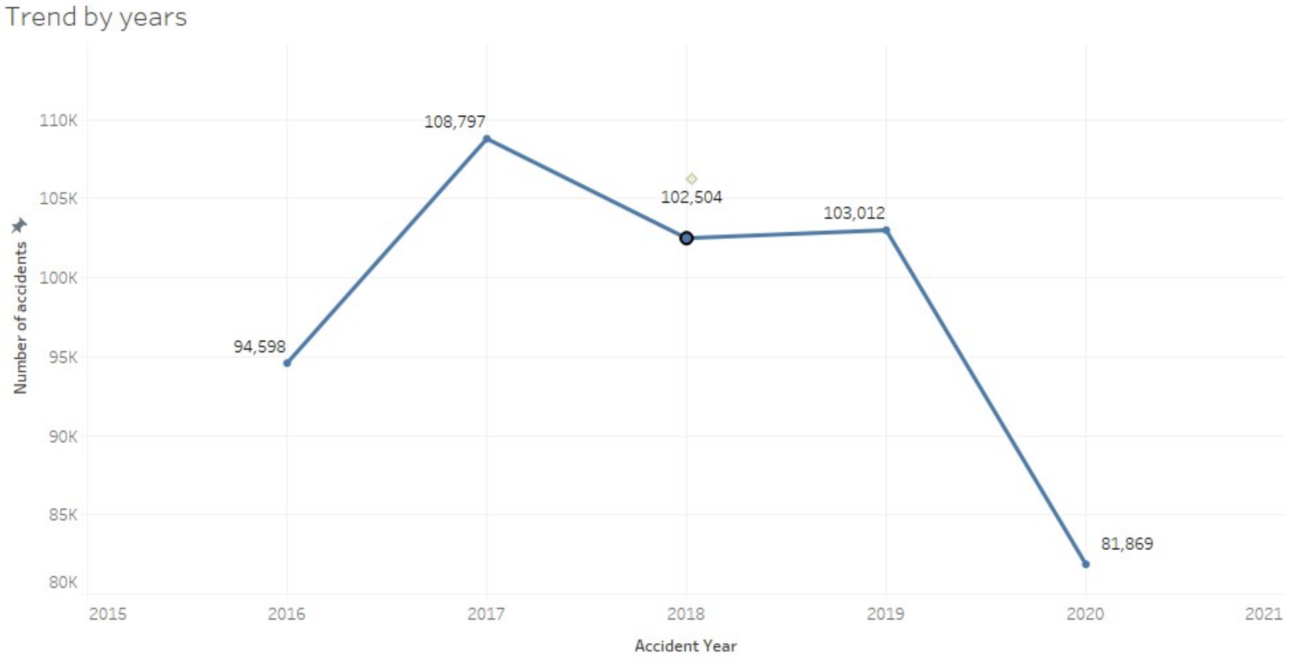


Figure 2

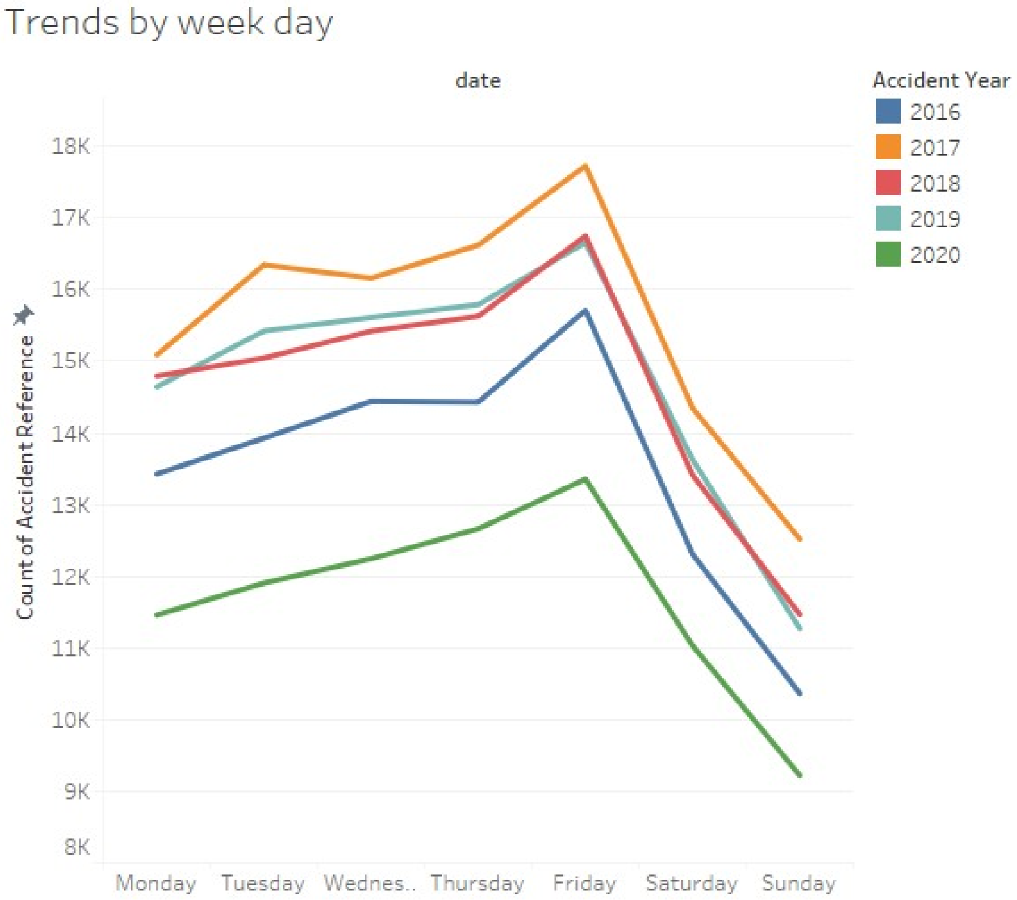


Figure 3

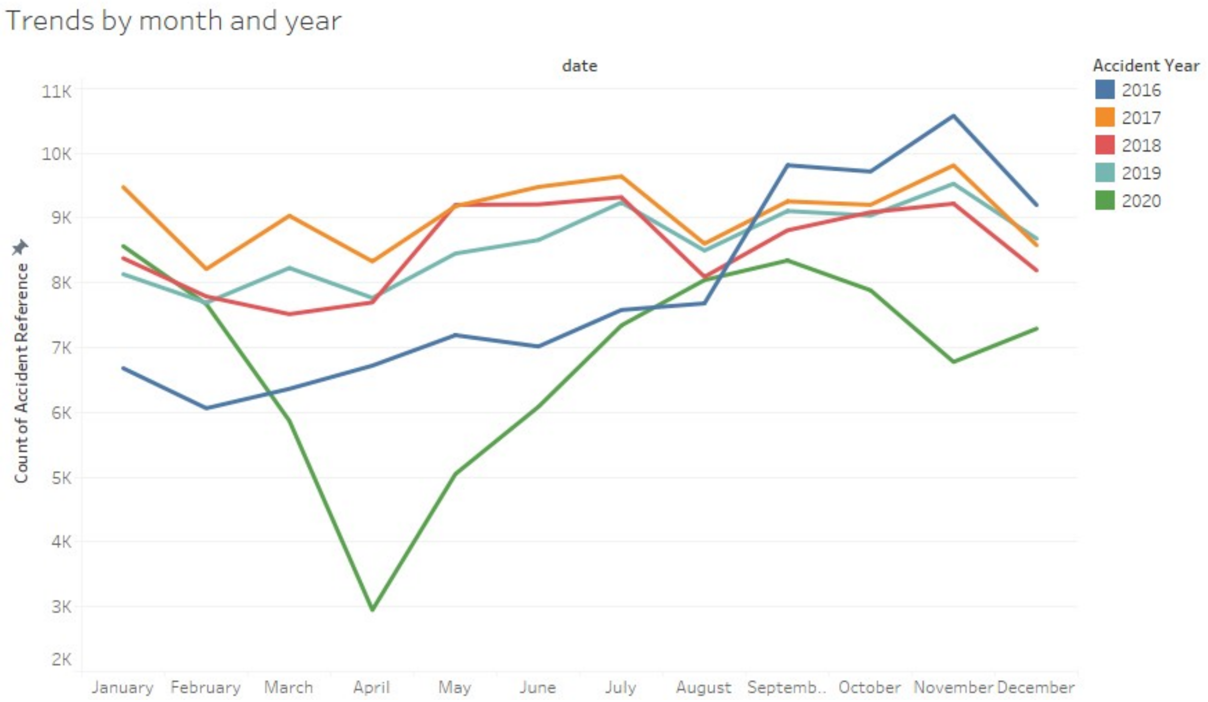


Figure 4

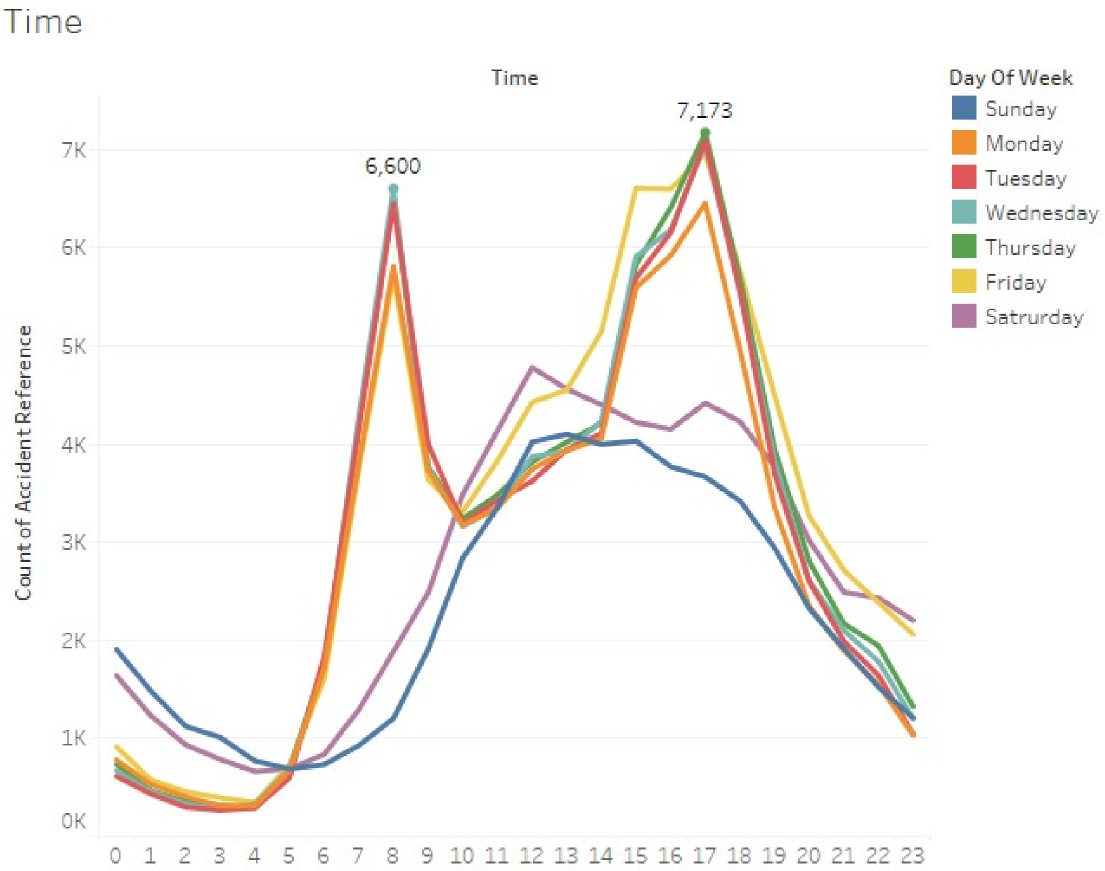
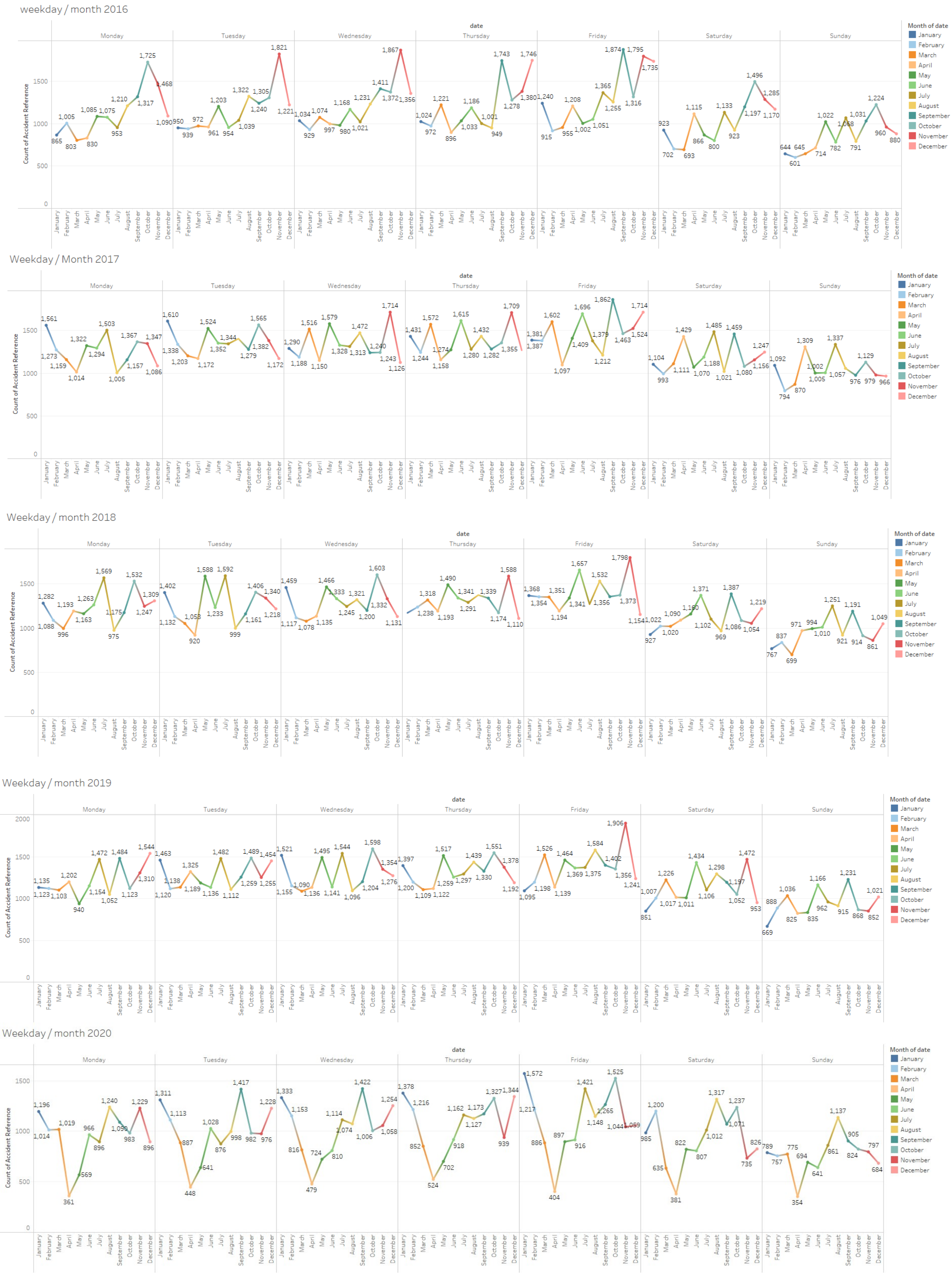
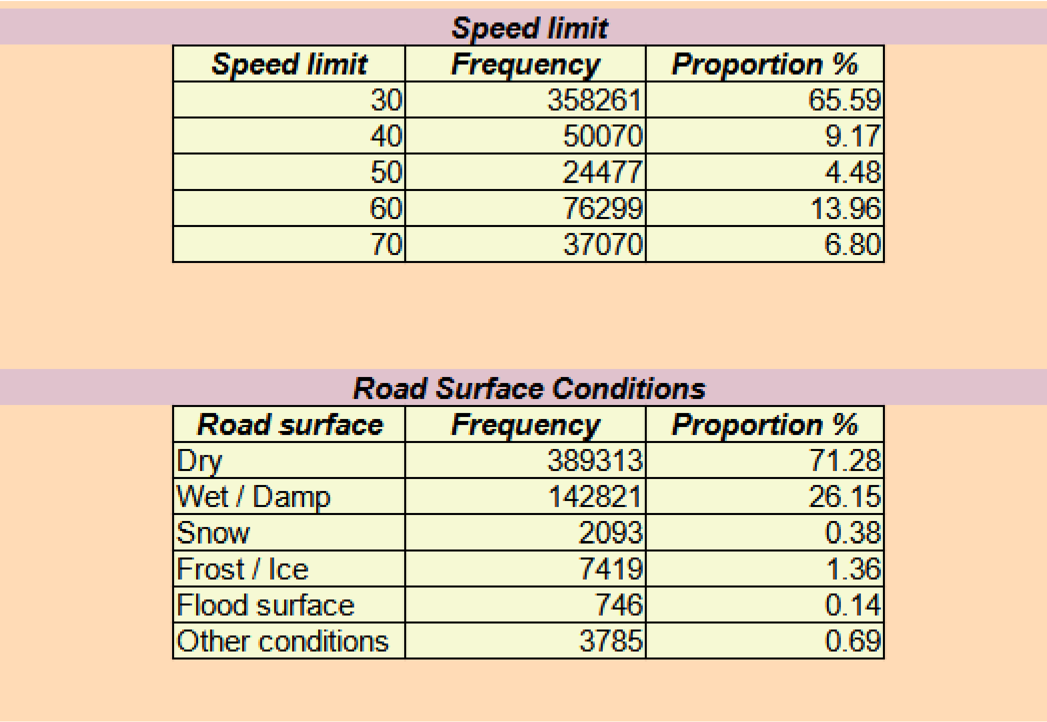


Figure 5a,5b,5c,5d,5e



Appendix 3

Tables of frequencies and proportions of accidents for different conditions.

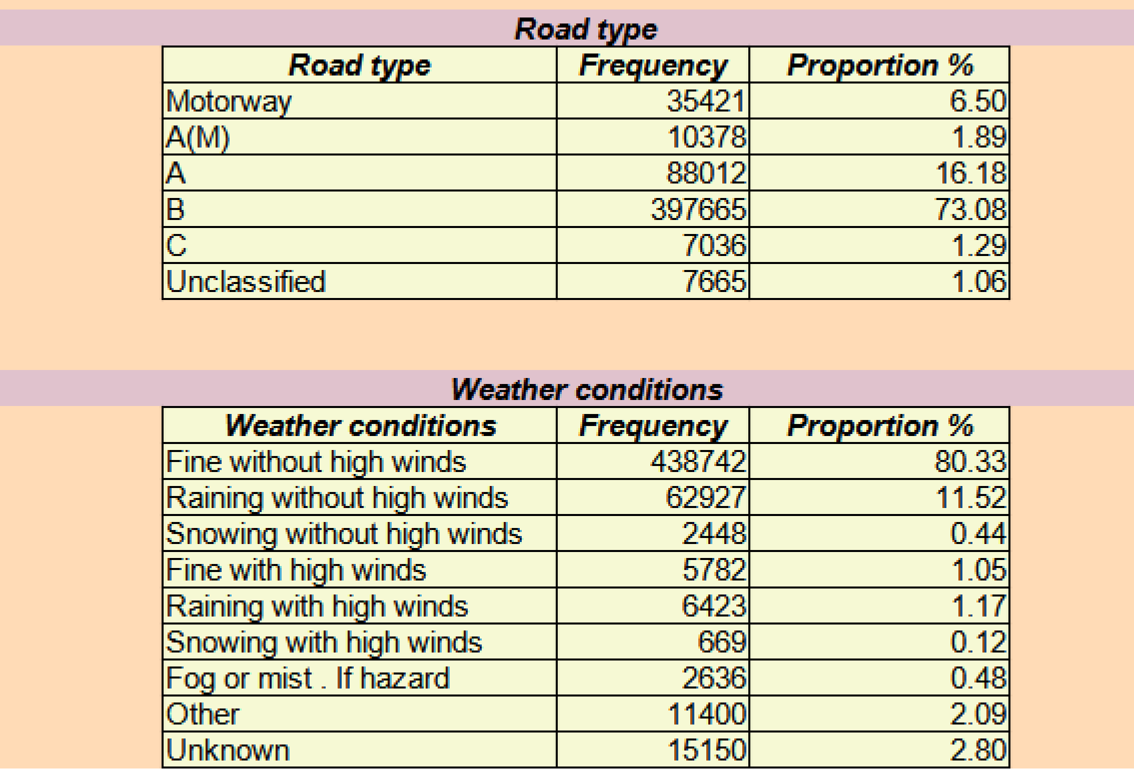


Speed limit

Speed limit	Frequency	Proportion %
30	358261	65.59
40	50070	9.17
50	24477	4.48
60	76299	13.96
70	37070	6.80

Road Surface Conditions

Road surface	Frequency	Proportion %
Dry	389313	71.28
Wet / Damp	142821	26.15
Snow	2093	0.38
Frost / Ice	7419	1.36
Flood surface	746	0.14
Other conditions	3785	0.69

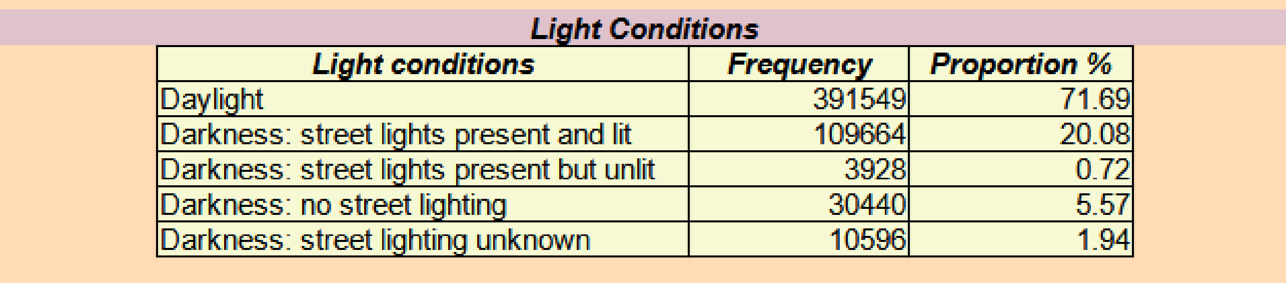


Road type

Road type	Frequency	Proportion %
Motorway	35421	6.50
A(M)	10378	1.89
A	88012	16.18
B	397665	73.08
C	7036	1.29
Unclassified	7665	1.06

Weather conditions

Weather conditions	Frequency	Proportion %
Fine without high winds	438742	80.33
Raining without high winds	62927	11.52
Snowing without high winds	2448	0.44
Fine with high winds	5782	1.05
Raining with high winds	6423	1.17
Snowing with high winds	669	0.12
Fog or mist . If hazard	2636	0.48
Other	11400	2.09
Unknown	15150	2.80



Light Conditions

Light conditions	Frequency	Proportion %
Daylight	391549	71.69
Darkness: street lights present and lit	109664	20.08
Darkness: street lights present but unlit	3928	0.72
Darkness: no street lighting	30440	5.57
Darkness: street lighting unknown	10596	1.94

References.

DB-Engines Ranking (no date). solid IT gmbh. Available at: https://db-engines.com/en/ranking (Accessed: December 8, 2022).