Data science applied to the fight against traffic accidents

Anton Barrera Mora (me@antonio-barrera.cyou)

April 2023

Contents

Introduction

In this project, we intend to show how the "life cycle" applies to data mining (DM) projects.The 'project life cycle' refers to the sequential phases that a project goes through from its initiation to its completion. It provides a structured framework for managing projects and encompasses all the activities, processes, and deliverables involved in the project's lifespan. We are talking abour a general framework that establishes the stages or phases a project goes through from initiation to completion. The project lifecycle provides a structure for managing the project as a whole, from defining the objectives to delivering the results.

In the other hand, while the CRISP-DM "What Is CRISP DM? - Data Science Process Alliance" (n.d.) model specifically focuses on the processes and steps involved in data mining, the project lifecycle is a broader framework that can be applied to different types of projects, including data mining projects. Then the CRISP-DM model can be considered as a specific methodology within the project lifecycle of a data mining project. It provides detailed guidelines on the specific stages to be followed when conducting a data mining project, such as understanding the business, understanding the data, preparing the data, modeling, evaluating, and deploying.

We will limit ourselves to the first three phases:

1. Understanding the business.
2. Understanding the data.
3. Data preparation.To achieve our objectives, we selected dataset sourced from the National Highway Traffic Safety Administration (NHTSA). The Fatality Analysis Reporting System (FARS) was developed by the NHTSA in the United States to provide a comprehensive measure of road safety. This dataset specifically pertains to the year 2020 and consists of accident records that capture significant descriptive data. Each accident in the dataset involves at least one fatality.

Phase 1. Business understanding

Problem:

We are requested to analyze the causes of road accidents in the United States, whether they are of human, material, or environmental origin. With this information, they hope to annually review the trends in this matter with the help of the model and adjust the intervention plan, whether through investment, campaigns, or training. They are also very interested in identifying specific states and cities where the intervention in road safety should be increased or modified, as well as any aspects related to infrastructure. Finally, we are asked to review the three-year time series to understand the evolution of road accidents.

Objective:

This data mining project aims to explore the dataset, uncover hidden patterns, and in future if potentially possible, develop predictive models to identify factors contributing to severe accidents. The insights gained from this analysis will provide valuable information for road safety initiatives and support evidence-based decision-making in accident prevention strategies.

The primary analytical objective of this project is to gain insights into the factors that contribute to the severity of an accident and to define what constitutes a severe accident. By applying data mining techniques, we aim to uncover patterns and relationships within the dataset that can help us understand the key factors associated with severe accidents.

Based on the above and in summary, we will undertake the initial phases to design a data mining model that allows us to understand the following in an updated manner:

1. The evolution of the time series of fatal accidents.
2. Major causes of accidents.
3. Incident volume by states and cities, including "black spots" in the road network.

Product:

Documented data mining model tailored to extract relevant information in the defined areas of interest (time series, black spots, and causes).

Tasks:

Definition of the study population: Traffic accidents in the USA from 2018 to 2020.

Data collection:

Primary data (provided by the contracting entity).
For more information on the dataset and the FARS, please refer to the National Highway Traffic Safety Administration's Crash Data Systems website: FARS.

Relevant variables:

In a holistic approach, the variables or factors in this project are established around the number of traffic accidents, human causes, material causes, and environmental causes that can influence their occurrence.

Success criteria:

An informative data mining model that fulfills its function of helping to prevent accidents by anticipating material wear (emergence of new black spots in the road network of states or cities), the appearance of new drugs and areas of incidence, and serves as a primary tool for decision-making.

Phase 2. Understanding the Data

In this phase, we will identify the necessary dataset to achieve the set objective and gain a holistic understanding of its structure. We will also address ethical considerations in data usage.

Objectives:

- Selection and identification of the necessary dataset in relation to the set objectives.
- Initial filtering of the data to eliminate redundant tables and records identified during the initial analysis.
- Attention to data ethics aspects:
  - Ownership of the dataset
  - Transparency in transactions
  - Consent
  - Currency
  - Privacy
  - Openness
  - Compliance with legal aspects of data usage, including storage and security.

Product:

- Descriptive document of the available data, including details about the data itself, its origin, availability, storage, privacy, security, and other legal aspects. Also includes information about variables: their nature, structure, problems, validity, and appropriateness.
- Final datasets suitable for the next phase to achieve the project objectives.

Tasks:

Selection of the dataset: The National Highway Traffic Safety Administration (NHTSA) in the United States provides a public dataset on traffic accidents called the Fatality Analysis Reporting System (FARS), which collects detailed information on fatal traffic accidents nationwide since 1975.

FARS contains information about the characteristics of vehicles involved in accidents, drivers and passengers, road and weather conditions, as well as the cause and consequences of the accidents. The information is collected through reports from law enforcement authorities, accident investigations, and medical and autopsy records.

FARS data is publicly available in CSV format and can be downloaded from the NHTSA website. There are also online tools and software packages that can assist in analyzing FARS data.

It should be noted that FARS only includes information on traffic accidents with fatalities, so it does not provide data on the entirety of traffic accidents in the United States. However, it can still be a valuable source of information for research on road safety and accident prevention.

The analytical objective is to understand the main causes of accidents, black spots in the road network, and the temporal evolution of accidents. We have selected the main dataset for accidents in 2020, called 'FARS2020NationalCSV'. It is a version of the FARS dataset that includes information on fatal traffic accidents that occurred in the year 2020 in the USA. It includes several tables that provide detailed information about vehicles, drivers, and passengers involved in the accidents, as well as the circumstances of the accidents.

The selected tables from the overall dataset that will substantiate the project objectives are:

'accident'

Information about the fatal traffic accident. Some of the variables included in this table are:

"ST_CASE": the case number assigned to the accident.

"STATE": numeric code of the state where the accident occurred.

"STATENAME": name of the state where the accident occurred corresponding to its numeric code.

"COUNTY": numeric code of the county where the accident occurred.

"COUNTYNAME": names of the counties corresponding to their numeric codes.

"CITY": numeric code of cities.

"CITYNAME": names of the cities corresponding to their numeric codes.

"FATALS": number of fatalities.

"DAY": day of the accident.

"MONTH": month in which the accident occurred.

"HOUR": the hour of the day when the accident occurred (in 24-hour format).

"MINUTE": minutes of the hour when the accident occurred.

"NHS": codes indicating if the accident occurred on a National Highway System (NHS) road.

"NHSNAME": confirmation if the accident belongs to an NHS road.

"ROUTE": code indicating the type of route where the accident occurred.

"ROUTENAME": name of the route corresponding to the numeric code.

"TWAY_ID": the identification of the roadway where the accident occurred.

"TWAY_ID2": the identification of a second roadway if necessary.

"MILEPT": the mile at which the accident occurred.

"LATITUDE": the latitude of the accident location.

"LONGITUDE": the longitude of the accident location.

"MAN_COLL": code indicating the type of collision that occurred.

"MAN_COLLNAME": description of the collision type.

"RUR_URB": codes specifying the area of occurrence.

"RUR_URBNAME": confirmation of the area where the accident occurred.

"LGT_COND": code indicating the lighting condition of the vehicle at the time of the accident.

"LG_CONDNAME": description of the lighting condition based on the previous code.

"WEATHER": code indicating the weather conditions.

"WEATHERNAME": name of the adverse weather condition.

"DRUNK_DR": number of drunk drivers.

This table provides valuable information about the characteristics of fatal traffic accidents, including the date, time, location, and number of victims. These data can be used to identify patterns and trends in traffic accidents, which in turn can be utilized to develop more effective accident prevention strategies. Additionally, it serves as the basis for calculating the time series and other statistics related to the total number of victims and forecasting.

'driverrf'

Contains information about the drivers involved in the fatal traffic accident. Some of the variables included in this table are:

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"DRIVERRF": numeric code indicating references to driver circumstances, such as driver's license information, improper vehicle handling, and various aspects related to the driver.

"DRIVERRFNAME": description of the above codes regarding relevant driver and driving aspects at the time of the accident.

This table provides valuable information about the characteristics of drivers involved in fatal traffic accidents. Variables related to the driver's condition and substance use are important for our analytical objectives.

'vehicle'

Contains information about the vehicles involved in the fatal traffic accident. Some of the variables considered for the analytical objective are:

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"MAKE": numeric code for the vehicle's make.

"MAKENAME": make name.

"MODEL": numeric code for the vehicle's model.

"MAK_MOD": numeric code for the make and model.

"MAK_MODNAME": make and model name.

"MOD_YEAR": the vehicle's model year.

"TOW_VEH": whether the vehicle was being towed at the time of the accident.

"MCARR_ID": identification of the commercial vehicle involved in the accident.

"GVWR": gross vehicle weight rating.

"VEH_YEAR": the vehicle's manufacturing year.

This table provides valuable information about the characteristics of vehicles involved in traffic accidents with fatalities. Variables related to the vehicle's manufacturing date and model can help identify and characterize issues with the vehicles in use.

'weather'

In FARS, it contains information about the weather conditions at the time of the fatal traffic accident. Some of the variables included in this table are:

"ST_CASE": the case number assigned to the accident.

"STATENAME": the state where the accident occurred.

"WEATHER": the numeric code describing the weather conditions at the time of the accident.

"WEATHERNAME": description of the weather conditions described in the previous variable at the time of the accident.

It provides information about the weather conditions at the time of the fatal traffic accident and can be used to conduct more detailed analyses of road safety and accident prevention under different weather conditions. In our case, we are interested in understanding the multiple variables involved in accidents and their effects, but this type of information can also enable corrective measures in areas where certain weather phenomena are more frequent.

'distract'

Contains data on the distractions that drivers were experiencing at the time of the fatal traffic accident. Some of the variables included are:

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"DRDISTRACT": the numeric code describing the distraction the driver was experiencing at the time of the accident.

"DRDISTRACTNAME": description of the type of distraction.

It provides information about a wide range of distractions that drivers experienced at the time of the accident. Some of the distractions captured in the table include:

- Talking or interacting with a passenger in the vehicle.
- Eating or drinking while driving.
- Using a mobile phone or electronic device.
- Observing something outside the vehicle, such as an accident or scenery.
- Touching or adjusting the radio or vehicle controls.
- Looking at or reaching for something inside the vehicle.
- Other types of distractions, such as being fatigued, being emotionally disturbed, or being distracted by something outside the vehicle.

'drugs'

Contains information about the presence of drugs in drivers involved in fatal traffic accidents. Some of the variables included in this table are:

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"DRUGSPEC": numeric code for the type of drug test conducted.

"DRUGSPECNAME": descriptive specification of the type of drug test conducted.

"DRUGRES": numeric code for the test results.

"DRUGRESNAME": descriptive specification of the test results, specifying the name of the substance.

Understanding the main causes of accidents is the analytical objective, but the variables included in this table can help us understand trends in drug use and highlight areas where this occurs, enabling targeted awareness and information campaigns among the population.

'vision'

Contains information about vision problems of drivers involved in fatal traffic accidents. Some of the variables included in this table are:

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"VISION": the numeric code describing the vision problem the driver had at the time of the accident.

"VISION_NAME": detailed description of the vision problem.

Understanding the variables in this table can be useful for analyzing the relationship between drivers' vision problems and fatal traffic accidents.

'Factor'

Contains information about different factors contributing to fatal traffic accidents. Vehicle conditions, environmental conditions, driver behavior, and other external factors.

"ST_CASE": the case number assigned to the accident.

"VEH_NO": the vehicle number involved in the accident.

"VEHICLECC": numeric code describing the factor that contributed to the accident. There are several codes available for mechanical failures observed in the vehicle.

"VEHICLECCNAME": provides descriptive information about the mechanical problem referred to by each code.

It can help us understand the main mechanical causes of traffic accidents, relate them to vehicle models and age, and help develop prevention and control measures.

Exploratory Analysis We are going to make a first approach to explore the chosen dataset and get to know its structure as the number of records, variables, type of variables, problems, determine the value of the data and other important aspects before starting to work with the dataset.

```
# Set paths to different accident datasets
  # 2020 Dataset
path_acc20 = "./FARS2020NationalCSV/accident.CSV"
    # 2019 Dataset
path_acc19 = "./FARS2019NationalAuxiliaryCSV/ACC_AUX.CSV"
      # 2018 Dataset
path_acc18 = "./FARS2018NationalAuxiliaryCSV/ACC_AUX.CSV"

# Set paths to additional 2020 files for further analysis
path_veh = "./FARS2020NationalCSV/vehicle.CSV"
path_per = "./FARS2020NationalCSV/person.CSV"
path_driverrf = "./FARS2020NationalCSV/driverrf.CSV"
path_distract = "./FARS2020NationalCSV/Distract.CSV"
path_drimpair = "./FARS2020NationalCSV/Drimpair.CSV"
path_factor = "./FARS2020NationalCSV/Factor.CSV"
path_vision = "./FARS2020NationalCSV/Vision.CSV"
path_wea = "./FARS2020NationalCSV/weather.CSV"



# Load accident data for each year for time series and other analyses
# 2020 Accidents - complete data
acc20 <- read.csv(path_acc20, row.names=NULL)
# 2019 Accidents - auxiliary data
acc19 <- read.csv(path_acc19, row.names=NULL)
# 2018 Accidents - auxiliary data
acc18 <- read.csv(path_acc18, row.names=NULL)

# Load additional 2020 data
# Vehicles
veh <- read.csv(path_veh, row.names=NULL)
```

```r
# Persons
per <- read.csv(path_per, row.names=NULL)

# Additional information about the driver
drivrf <- read.csv(path_driverrf, row.names=NULL)

# Information about driver impairments
impair <- read.csv(path_drimpair, row.names=NULL)

# Information about mechanical factors
fact <- read.csv(path_factor, row.names=NULL)

# Information about visibility problems
visio <- read.csv(path_vision, row.names=NULL)

# Information about driver distractions
distract <- read.csv(path_distract, row.names=NULL)

# Information about weather conditions
wea <- read.csv(path_wea, row.names=NULL)
```

Once the dataset is loaded, we proceed to the exploration of the dataset.

1. Verifying the structure of the accident set from 2018 to 2020 inclusive: ACCIDENTS 2020

```r
# Main file acc20

# Check the first few rows of data for each variable
head(acc20)
```

```
##   STATE STATENAME ST_CASE VE_TOTAL VE_FORMS PVH_INVL PEDS PERSONS PERMVIT
## 1     1   Alabama   10001        1        1        0    0       4       4
## 2     1   Alabama   10002        4        4        0    0       6       6
## 3     1   Alabama   10003        2        2        0    0       2       2
## 4     1   Alabama   10004        1        1        0    0       5       5
## 5     1   Alabama   10005        1        1        0    0       1       1
## 6     1   Alabama   10006        2        2        0    0       3       3
##   PERNOTMVIT COUNTY     COUNTYNAME CITY       CITYNAME DAY DAYNAME MONTH
## 1          0     51    ELMORE (51)    0 NOT APPLICABLE   1       1     1
## 2          0     73 JEFFERSON (73)  350     BIRMINGHAM   2       2     1
## 3          0    117   SHELBY (117)    0 NOT APPLICABLE   2       2     1
## 4          0     15   CALHOUN (15)    0 NOT APPLICABLE   3       3     1
## 5          0     37     COOSA (37)    0 NOT APPLICABLE   4       4     1
## 6          0    103   MORGAN (103)    0 NOT APPLICABLE   4       4     1
##   MONTHNAME YEAR DAY_WEEK DAY_WEEKNAME HOUR      HOURNAME MINUTE MINUTENAME NHS
## 1   January 2020        4    Wednesday    2 2:00am-2:59am     58         58   0
## 2   January 2020        5     Thursday   17 5:00pm-5:59pm     18         18   0
## 3   January 2020        5     Thursday   14 2:00pm-2:59pm     55         55   0
## 4   January 2020        6       Friday   15 3:00pm-3:59pm     20         20   0
## 5   January 2020        7     Saturday    0 0:00am-0:59am     45         45   0
## 6   January 2020        7     Saturday   16 4:00pm-4:59pm     55         55   0
##                         NHSNAME ROUTE                 ROUTENAME
```

```
## 1 This section IS NOT on the NHS     4                 County Road
## 2 This section IS NOT on the NHS     6 Local Street - Municipality
## 3 This section IS NOT on the NHS     3               State Highway
## 4 This section IS NOT on the NHS     4                 County Road
## 5 This section IS NOT on the NHS     4                 County Road
## 6 This section IS NOT on the NHS     3               State Highway
##                       TWAY_ID TWAY_ID2 RUR_URB RUR_URBNAME FUNC_SYS
## 1                        cr-4               1       Rural        5
## 2    martin luther king jr dr               2       Urban        4
## 3                       sr-76   us-280       1       Rural        4
## 4 CR-ALEXANDRIA WELLINGTON RD               1       Rural        7
## 5                       CR-63               1       Rural        5
## 6                       sr-36               1       Rural        4
##      FUNC_SYSNAME RD_OWNER                    RD_OWNERNAME MILEPT MILEPTNAME
## 1 Major Collector        2            County Highway Agency      0       None
## 2  Minor Arterial        4 City or Municipal Highway Agency      0       None
## 3  Minor Arterial        1             State Highway Agency     49         49
## 4           Local        2            County Highway Agency      0       None
## 5 Major Collector        2            County Highway Agency      0       None
## 6  Minor Arterial        1             State Highway Agency    390        390
##   LATITUDE LATITUDENAME  LONGITUD LONGITUDNAME SP_JUR              SP_JURNAME
## 1 32.43313  32.43313333 -86.09485    -86.09485      0 No Special Jurisdiction
## 2 33.48466  33.48465833 -86.83954 -86.83954444      0 No Special Jurisdiction
## 3 33.29994  33.29994167 -86.36964 -86.36964167      0 No Special Jurisdiction
## 4 33.79507  33.79507222 -85.88349 -85.88348611      0 No Special Jurisdiction
## 5 32.84841  32.84841389 -86.08355 -86.08354722      0 No Special Jurisdiction
## 6 34.50894  34.50894167 -86.67486 -86.67485556      0 No Special Jurisdiction
##   HARM_EV               HARM_EVNAME MAN_COLL
## 1      42       Tree (Standing Only)        0
## 2      12 Motor Vehicle In-Transport        6
## 3      34                     Ditch        0
## 4      42       Tree (Standing Only)        0
## 5      42       Tree (Standing Only)        0
## 6      12 Motor Vehicle In-Transport        2
##                                                                    MAN_COLLNAME
## 1 The First Harmful Event was Not a Collision with a Motor Vehicle in Transport
## 2                                                                         Angle
## 3 The First Harmful Event was Not a Collision with a Motor Vehicle in Transport
## 4 The First Harmful Event was Not a Collision with a Motor Vehicle in Transport
## 5 The First Harmful Event was Not a Collision with a Motor Vehicle in Transport
## 6                                                                 Front-to-Front
##   RELJCT1 RELJCT1NAME RELJCT2          RELJCT2NAME TYP_INT         TYP_INTNAME
## 1       0          No       1         Non-Junction       1 Not an Intersection
## 2       0          No       1         Non-Junction       1 Not an Intersection
## 3       0          No       3 Intersection-Related       3      T-Intersection
## 4       0          No       1         Non-Junction       1 Not an Intersection
## 5       0          No       1         Non-Junction       1 Not an Intersection
## 6       0          No       1         Non-Junction       1 Not an Intersection
##   WRK_ZONE WRK_ZONENAME REL_ROAD REL_ROADNAME LGT_COND       LGT_CONDNAME
## 1        0         None        4  On Roadside        2 Dark - Not Lighted
## 2        0         None        1   On Roadway        3     Dark - Lighted
## 3        0         None        4  On Roadside        1           Daylight
## 4        0         None        4  On Roadside        1           Daylight
## 5        0         None        4  On Roadside        2 Dark - Not Lighted
```

```
## 6        0         None        1   On Roadway        2 Dark - Not Lighted
##   WEATHER WEATHERNAME SCH_BUS SCH_BUSNAME    RAIL       RAILNAME NOT_HOUR
## 1       1       Clear       0          No 0000000 Not Applicable       99
## 2       2        Rain       0          No 0000000 Not Applicable       17
## 3       2        Rain       0          No 0000000 Not Applicable       14
## 4      10      Cloudy       0          No 0000000 Not Applicable       99
## 5       2        Rain       0          No 0000000 Not Applicable        0
## 6       1       Clear       0          No 0000000 Not Applicable       17
##    NOT_HOURNAME NOT_MIN NOT_MINNAME ARR_HOUR                   ARR_HOURNAME
## 1       Unknown      99     Unknown        3                  3:00am-3:59am
## 2 5:00pm-5:59pm      18          18       17                  5:00pm-5:59pm
## 3 2:00pm-2:59pm      58          58       15                  3:00pm-3:59pm
## 4       Unknown      99     Unknown       99 Unknown EMS Scene Arrival Hour
## 5 0:00am-0:59am      45          45        0                  0:00am-0:59am
## 6 5:00pm-5:59pm       0           0       17                  5:00pm-5:59pm
##   ARR_MIN                       ARR_MINNAME HOSP_HR
## 1      10                                10      99
## 2      26                                26      99
## 3      15                                15      99
## 4      99 Unknown EMS Scene Arrival Minutes      99
## 5      55                                55      88
## 6      19                                19      18
##                         HOSP_HRNAME HOSP_MN                        HOSP_MNNAME
## 1                           Unknown      99 Unknown EMS Hospital Arrival Time
## 2                           Unknown      99 Unknown EMS Hospital Arrival Time
## 3                           Unknown      99 Unknown EMS Hospital Arrival Time
## 4                           Unknown      99 Unknown EMS Hospital Arrival Time
## 5 Not Applicable (Not Transported)      88  Not Applicable (Not Transported)
## 6                      6:00pm-6:59pm      51                                51
##   FATALS DRUNK_DR
## 1      3        1
## 2      1        0
## 3      1        0
## 4      1        0
## 5      1        0
## 6      1        0
```

```
# Analyze numeric variables
summary(acc20)
```

```
##      STATE        STATENAME            ST_CASE          VE_TOTAL
##  Min.   : 1.00   Length:35766       Min.   : 10001   Min.   : 1.00
##  1st Qu.:12.00   Class :character   1st Qu.:122078   1st Qu.: 1.00
##  Median :26.00   Mode  :character   Median :260917   Median : 1.00
##  Mean   :27.16                      Mean   :272387   Mean   : 1.56
##  3rd Qu.:42.00                      3rd Qu.:420477   3rd Qu.: 2.00
##  Max.   :56.00                      Max.   :560115   Max.   :15.00
##     VE_FORMS         PVH_INVL             PEDS          PERSONS
##  Min.   : 1.000   Min.   : 0.00000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.: 0.00000   1st Qu.:0.0000   1st Qu.: 1.000
##  Median : 1.000   Median : 0.00000   Median :0.0000   Median : 2.000
##  Mean   : 1.517   Mean   : 0.04269   Mean   :0.2285   Mean   : 2.173
##  3rd Qu.: 2.000   3rd Qu.: 0.00000   3rd Qu.:0.0000   3rd Qu.: 3.000
##  Max.   :15.000   Max.   :10.00000   Max.   :8.0000   Max.   :61.000
```

```
##     PERMVIT          PERNOTMVIT          COUNTY        COUNTYNAME
##  Min.   : 0.000   Min.   :0.0000   Min.   :  1.00   Length:35766
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.: 31.00   Class :character
##  Median : 2.000   Median :0.0000   Median : 71.00   Mode  :character
##  Mean   : 2.163   Mean   :0.2387   Mean   : 93.06
##  3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:117.00
##  Max.   :61.000   Max.   :9.0000   Max.   :999.00
##       CITY        CITYNAME             DAY           DAYNAME
##  Min.   :   0   Length:35766       Min.   : 1.00   Min.   : 1.00
##  1st Qu.:   0   Class :character   1st Qu.: 8.00   1st Qu.: 8.00
##  Median : 120   Mode  :character   Median :16.00   Median :16.00
##  Mean   :1436                      Mean   :15.71   Mean   :15.71
##  3rd Qu.:2080                      3rd Qu.:23.00   3rd Qu.:23.00
##  Max.   :9999                      Max.   :31.00   Max.   :31.00
##      MONTH          MONTHNAME             YEAR         DAY_WEEK
##  Min.   : 1.000   Length:35766       Min.   :2020   Min.   :1.000
##  1st Qu.: 4.000   Class :character   1st Qu.:2020   1st Qu.:2.000
##  Median : 7.000   Mode  :character   Median :2020   Median :4.000
##  Mean   : 6.898                      Mean   :2020   Mean   :4.114
##  3rd Qu.:10.000                      3rd Qu.:2020   3rd Qu.:6.000
##  Max.   :12.000                      Max.   :2020   Max.   :7.000
##  DAY_WEEKNAME            HOUR         HOURNAME            MINUTE
##  Length:35766       Min.   : 0.00   Length:35766       Min.   : 0.00
##  Class :character   1st Qu.: 7.00   Class :character   1st Qu.:14.00
##  Mode  :character   Median :15.00   Mode  :character   Median :30.00
##                     Mean   :13.94                      Mean   :29.24
##                     3rd Qu.:19.00                      3rd Qu.:45.00
##                     Max.   :99.00                      Max.   :99.00
##   MINUTENAME             NHS           NHSNAME             ROUTE
##  Length:35766       Min.   :0.0000   Length:35766       Min.   :1.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.:2.000
##  Mode  :character   Median :0.0000   Mode  :character   Median :3.000
##                     Mean   :0.5877                      Mean   :3.901
##                     3rd Qu.:1.0000                      3rd Qu.:6.000
##                     Max.   :9.0000                      Max.   :9.000
##   ROUTENAME           TWAY_ID            TWAY_ID2            RUR_URB
##  Length:35766       Length:35766       Length:35766       Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :1.662
##                                                           3rd Qu.:2.000
##                                                           Max.   :9.000
##  RUR_URBNAME           FUNC_SYS      FUNC_SYSNAME          RD_OWNER
##  Length:35766       Min.   : 1.000   Length:35766       Min.   : 1.00
##  Class :character   1st Qu.: 3.000   Class :character   1st Qu.: 1.00
##  Mode  :character   Median : 4.000   Mode  :character   Median : 1.00
##                     Mean   : 6.038                      Mean   :19.96
##                     3rd Qu.: 5.000                      3rd Qu.: 4.00
##                     Max.   :99.000                      Max.   :99.00
##  RD_OWNERNAME           MILEPT          MILEPTNAME           LATITUDE
##  Length:35766       Min.   :    0.0   Length:35766       Min.   : 19.09
##  Class :character   1st Qu.:    2.0   Class :character   1st Qu.: 32.99
##  Mode  :character   Median :   80.0   Mode  :character   Median : 36.17
##                     Mean   :19990.7                      Mean   : 36.90
```

```
##                    3rd Qu.:  955.5                      3rd Qu.: 40.45
##                    Max.   :99999.0                      Max.   :100.00
##  LATITUDENAME         LONGITUD       LONGITUDNAME          SP_JUR
##  Length:35766       Min.   :-165.30   Length:35766       Min.   :0.00000
##  Class :character   1st Qu.: -97.90   Class :character   1st Qu.:0.00000
##  Mode  :character   Median : -87.81   Mode  :character   Median :0.00000
##                     Mean   : -84.59                      Mean   :0.04029
##                     3rd Qu.: -81.52                      3rd Qu.:0.00000
##                     Max.   :1000.00                      Max.   :9.00000
##   SP_JURNAME           HARM_EV      HARM_EVNAME           MAN_COLL
##  Length:35766       Min.   : 1.00   Length:35766       Min.   : 0.000
##  Class :character   1st Qu.: 8.00   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :12.00   Mode  :character   Median : 0.000
##                     Mean   :18.31                      Mean   : 1.929
##                     3rd Qu.:30.00                      3rd Qu.: 2.000
##                     Max.   :99.00                      Max.   :99.000
##  MAN_COLLNAME         RELJCT1        RELJCT1NAME           RELJCT2
##  Length:35766       Min.   :0.00000   Length:35766       Min.   : 1.000
##  Class :character   1st Qu.:0.00000   Class :character   1st Qu.: 1.000
##  Mode  :character   Median :0.00000   Mode  :character   Median : 1.000
##                     Mean   :0.07283                      Mean   : 2.368
##                     3rd Qu.:0.00000                      3rd Qu.: 2.000
##                     Max.   :9.00000                      Max.   :99.000
##  RELJCT2NAME          TYP_INT       TYP_INTNAME           WRK_ZONE
##  Length:35766       Min.   : 1.000   Length:35766       Min.   :0.00000
##  Class :character   1st Qu.: 1.000   Class :character   1st Qu.:0.00000
##  Mode  :character   Median : 1.000   Mode  :character   Median :0.00000
##                     Mean   : 1.764                      Mean   :0.04748
##                     3rd Qu.: 1.000                      3rd Qu.:0.00000
##                     Max.   :99.000                      Max.   :4.00000
##  WRK_ZONENAME         REL_ROAD      REL_ROADNAME          LGT_COND
##  Length:35766       Min.   : 1.000   Length:35766       Min.   :1.000
##  Class :character   1st Qu.: 1.000   Class :character   1st Qu.:1.000
##  Mode  :character   Median : 1.000   Mode  :character   Median :2.000
##                     Mean   : 2.557                      Mean   :1.961
##                     3rd Qu.: 4.000                      3rd Qu.:3.000
##                     Max.   :99.000                      Max.   :9.000
##  LGT_CONDNAME         WEATHER       WEATHERNAME           SCH_BUS
##  Length:35766       Min.   : 1.000   Length:35766       Min.   :0.000000
##  Class :character   1st Qu.: 1.000   Class :character   1st Qu.:0.000000
##  Mode  :character   Median : 1.000   Mode  :character   Median :0.000000
##                     Mean   : 9.725                      Mean   :0.001426
##                     3rd Qu.: 2.000                      3rd Qu.:0.000000
##                     Max.   :99.000                      Max.   :1.000000
##  SCH_BUSNAME           RAIL             RAILNAME           NOT_HOUR
##  Length:35766       Length:35766       Length:35766       Min.   : 0.00
##  Class :character   Class :character   Class :character   1st Qu.:16.00
##  Mode  :character   Mode  :character   Mode  :character   Median :99.00
##                                                           Mean   :61.39
##                                                           3rd Qu.:99.00
##                                                           Max.   :99.00
##  NOT_HOURNAME         NOT_MIN       NOT_MINNAME           ARR_HOUR
##  Length:35766       Min.   : 0.00   Length:35766       Min.   : 0.00
##  Class :character   1st Qu.:34.00   Class :character   1st Qu.:16.00
```

```
##  Mode  :character   Median :98.00   Mode  :character   Median :99.00
##                     Mean   :68.46                     Mean   :61.88
##                     3rd Qu.:99.00                     3rd Qu.:99.00
##                     Max.   :99.00                     Max.   :99.00
##  ARR_HOURNAME          ARR_MIN      ARR_MINNAME          HOSP_HR
##  Length:35766       Min.   : 0.00   Length:35766       Min.   : 0.00
##  Class :character   1st Qu.:34.00   Class :character   1st Qu.:88.00
##  Mode  :character   Median :98.00   Mode  :character   Median :88.00
##                     Mean   :68.74                      Mean   :77.59
##                     3rd Qu.:99.00                      3rd Qu.:99.00
##                     Max.   :99.00                      Max.   :99.00
##  HOSP_HRNAME           HOSP_MN      HOSP_MNNAME           FATALS
##  Length:35766       Min.   : 0.00   Length:35766       Min.   :1.000
##  Class :character   1st Qu.:88.00   Class :character   1st Qu.:1.000
##  Mode  :character   Median :88.00   Mode  :character   Median :1.000
##                     Mean   :80.76                      Mean   :1.085
##                     3rd Qu.:99.00                      3rd Qu.:1.000
##                     Max.   :99.00                      Max.   :8.000
##     DRUNK_DR
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2664
##  3rd Qu.:1.0000
##  Max.   :4.0000
```

```r
# Get the names of variables in the accident table
names(acc20)
```

```
##  [1] "STATE"         "STATENAME"     "ST_CASE"       "VE_TOTAL"      "VE_FORMS"
##  [6] "PVH_INVL"      "PEDS"          "PERSONS"       "PERMVIT"       "PERNOTMVIT"
## [11] "COUNTY"        "COUNTYNAME"    "CITY"          "CITYNAME"      "DAY"
## [16] "DAYNAME"       "MONTH"         "MONTHNAME"     "YEAR"          "DAY_WEEK"
## [21] "DAY_WEEKNAME"  "HOUR"          "HOURNAME"      "MINUTE"        "MINUTENAME"
## [26] "NHS"           "NHSNAME"       "ROUTE"         "ROUTENAME"     "TWAY_ID"
## [31] "TWAY_ID2"      "RUR_URB"       "RUR_URBNAME"   "FUNC_SYS"      "FUNC_SYSNAME"
## [36] "RD_OWNER"      "RD_OWNERNAME"  "MILEPT"        "MILEPTNAME"    "LATITUDE"
## [41] "LATITUDENAME"  "LONGITUD"      "LONGITUDNAME"  "SP_JUR"        "SP_JURNAME"
## [46] "HARM_EV"       "HARM_EVNAME"   "MAN_COLL"      "MAN_COLLNAME"  "RELJCT1"
## [51] "RELJCT1NAME"   "RELJCT2"       "RELJCT2NAME"   "TYP_INT"       "TYP_INTNAME"
## [56] "WRK_ZONE"      "WRK_ZONENAME"  "REL_ROAD"      "REL_ROADNAME"  "LGT_COND"
## [61] "LGT_CONDNAME"  "WEATHER"       "WEATHERNAME"   "SCH_BUS"       "SCH_BUSNAME"
## [66] "RAIL"          "RAILNAME"      "NOT_HOUR"      "NOT_HOURNAME"  "NOT_MIN"
## [71] "NOT_MINNAME"   "ARR_HOUR"      "ARR_HOURNAME"  "ARR_MIN"       "ARR_MINNAME"
## [76] "HOSP_HR"       "HOSP_HRNAME"   "HOSP_MN"       "HOSP_MNNAME"   "FATALS"
## [81] "DRUNK_DR"
```

```r
# Perform a general analysis of the structure
summario20 <- as.data.frame(str(acc20, give.attr = FALSE, strict.width = "cut"))
```

```
## 'data.frame':    35766 obs. of  81 variables:
##  $ STATE       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ STATENAME   : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
```

```
##  $ ST_CASE     : int  10001 10002 10003 10004 10005 10006 10007 10008 10009 10..
##  $ VE_TOTAL    : int  1 4 2 1 1 2 1 2 2 2 ...
##  $ VE_FORMS    : int  1 4 2 1 1 2 1 2 2 2 ...
##  $ PVH_INVL    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PEDS        : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ PERSONS     : int  4 6 2 5 1 3 1 2 4 3 ...
##  $ PERMVIT     : int  4 6 2 5 1 3 1 2 4 3 ...
##  $ PERNOTMVIT  : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ COUNTY      : int  51 73 117 15 37 103 73 25 45 95 ...
##  $ COUNTYNAME  : chr  "ELMORE (51)" "JEFFERSON (73)" "SHELBY (117)" "CALHOUN "..
##  $ CITY        : int  0 350 0 0 0 0 330 0 0 1500 ...
##  $ CITYNAME    : chr  "NOT APPLICABLE" "BIRMINGHAM" "NOT APPLICABLE" "NOT APP"..
##  $ DAY         : int  1 2 2 3 4 4 7 8 9 10 ...
##  $ DAYNAME     : int  1 2 2 3 4 4 7 8 9 10 ...
##  $ MONTH       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ MONTHNAME   : chr  "January" "January" "January" "January" ...
##  $ YEAR        : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ DAY_WEEK    : int  4 5 5 6 7 7 3 4 5 6 ...
##  $ DAY_WEEKNAME: chr  "Wednesday" "Thursday" "Thursday" "Friday" ...
##  $ HOUR        : int  2 17 14 15 0 16 19 7 20 10 ...
##  $ HOURNAME    : chr  "2:00am-2:59am" "5:00pm-5:59pm" "2:00pm-2:59pm" "3:00pm"..
##  $ MINUTE      : int  58 18 55 20 45 55 23 15 0 2 ...
##  $ MINUTENAME  : chr  "58" "18" "55" "20" ...
##  $ NHS         : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ NHSNAME     : chr  "This section IS NOT on the NHS" "This section IS NOT o"..
##  $ ROUTE       : int  4 6 3 4 4 3 4 4 4 2 ...
##  $ ROUTENAME   : chr  "County Road" "Local Street - Municipality" "State High"..
##  $ TWAY_ID     : chr  "cr-4" "martin luther king jr dr" "sr-76" "CR-ALEXANDRI"..
##  $ TWAY_ID2    : chr  "" "" "us-280" "" ...
##  $ RUR_URB     : int  1 2 1 1 1 1 2 1 1 1 ...
##  $ RUR_URBNAME : chr  "Rural" "Urban" "Rural" "Rural" ...
##  $ FUNC_SYS    : int  5 4 4 7 5 4 4 5 5 3 ...
##  $ FUNC_SYSNAME: chr  "Major Collector" "Minor Arterial" "Minor Arterial" "Lo"..
##  $ RD_OWNER    : int  2 4 1 2 2 1 4 2 2 1 ...
##  $ RD_OWNERNAME: chr  "County Highway Agency" "City or Municipal Highway Agen"..
##  $ MILEPT      : int  0 0 49 0 0 390 0 0 0 3019 ...
##  $ MILEPTNAME  : chr  "None" "None" "49" "None" ...
##  $ LATITUDE    : num  32.4 33.5 33.3 33.8 32.8 ...
##  $ LATITUDENAME: chr  "32.43313333" "33.48465833" "33.29994167" "33.79507222" ..
##  $ LONGITUD    : num  -86.1 -86.8 -86.4 -85.9 -86.1 ...
##  $ LONGITUDNAME: chr  "-86.09485" "-86.83954444" "-86.36964167" "-85.88348611"..
##  $ SP_JUR      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SP_JURNAME  : chr  "No Special Jurisdiction" "No Special Jurisdiction" "No"..
##  $ HARM_EV     : int  42 12 34 42 42 12 8 12 12 12 ...
##  $ HARM_EVNAME : chr  "Tree (Standing Only)" "Motor Vehicle In-Transport" "Di"..
##  $ MAN_COLL    : int  0 6 0 0 0 2 0 1 1 2 ...
##  $ MAN_COLLNAME: chr  "The First Harmful Event was Not a Collision with a Mot"..
##  $ RELJCT1     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RELJCT1NAME : chr  "No" "No" "No" "No" ...
##  $ RELJCT2     : int  1 1 3 1 1 1 3 1 8 1 ...
##  $ RELJCT2NAME : chr  "Non-Junction" "Non-Junction" "Intersection-Related" "N"..
##  $ TYP_INT     : int  1 1 3 1 1 1 2 1 1 1 ...
##  $ TYP_INTNAME : chr  "Not an Intersection" "Not an Intersection" "T-Intersec"..
##  $ WRK_ZONE    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ WRK_ZONENAME: chr  "None" "None" "None" "None" ...
##  $ REL_ROAD    : int  4 1 4 4 4 1 1 1 1 1 ...
##  $ REL_ROADNAME: chr  "On Roadside" "On Roadway" "On Roadside" "On Roadside" ...
##  $ LGT_COND    : int  2 3 1 1 2 2 3 1 2 1 ...
##  $ LGT_CONDNAME: chr  "Dark - Not Lighted" "Dark - Lighted" "Daylight" "Dayli"..
##  $ WEATHER     : int  1 2 2 10 2 1 1 1 10 10 ...
##  $ WEATHERNAME : chr  "Clear" "Rain" "Rain" "Cloudy" ...
##  $ SCH_BUS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SCH_BUSNAME : chr  "No" "No" "No" "No" ...
##  $ RAIL        : chr  "0000000" "0000000" "0000000" "0000000" ...
##  $ RAILNAME    : chr  "Not Applicable" "Not Applicable" "Not Applicable" "Not"..
##  $ NOT_HOUR    : int  99 17 14 99 0 17 19 7 20 10 ...
##  $ NOT_HOURNAME: chr  "Unknown" "5:00pm-5:59pm" "2:00pm-2:59pm" "Unknown" ...
##  $ NOT_MIN     : int  99 18 58 99 45 0 23 21 0 3 ...
##  $ NOT_MINNAME : chr  "Unknown" "18" "58" "Unknown" ...
##  $ ARR_HOUR    : int  3 17 15 99 0 17 19 7 20 10 ...
##  $ ARR_HOURNAME: chr  "3:00am-3:59am" "5:00pm-5:59pm" "3:00pm-3:59pm" "Unknow"..
##  $ ARR_MIN     : int  10 26 15 99 55 19 29 28 10 7 ...
##  $ ARR_MINNAME : chr  "10" "26" "15" "Unknown EMS Scene Arrival Minutes" ...
##  $ HOSP_HR     : int  99 99 99 99 88 18 88 88 99 10 ...
##  $ HOSP_HRNAME : chr  "Unknown" "Unknown" "Unknown" "Unknown" ...
##  $ HOSP_MN     : int  99 99 99 99 88 51 88 88 99 29 ...
##  $ HOSP_MNNAME : chr  "Unknown EMS Hospital Arrival Time" "Unknown EMS Hospit"..
##  $ FATALS      : int  3 1 1 1 1 1 1 1 1 1 ...
##  $ DRUNK_DR    : int  1 0 0 0 0 0 0 0 0 0 ...
```

```
knitr::kable(summario20, caption = "Summary of the acc20 table")
```

Table: Summary of the acc20 table

We can observe that the 2020 accident table has 3576 records or instances of accidents with 81 variables. At first glance, we can already see that some variables are not necessary for the project as they provide unnecessary data. All numeric variables are of integral type, there are no date records, and the remaining variables can be categorized as text strings.

ACCIDENTS 2019

```
# Auxiliary file acc19

# Let's check the first rows of data for each variable
head(acc19)
```

```
##   ST_CASE YEAR STATE COUNTY FATALS A_CRAINJ A_REGION A_RU A_INTER A_RELRD
## 1   10001 2019     1     81      1        1        4    2       1       1
## 2   10002 2019     1     55      1        1        4    2       1       1
## 3   10003 2019     1     29      1        1        4    1       1       1
## 4   10004 2019     1     55      1        1        4    1       1       1
## 5   10005 2019     1      3      1        1        4    2       1       4
## 6   10006 2019     1     85      1        1        4    1       1       1
##   A_INTSEC A_ROADFC A_JUNC A_MANCOL A_TOD A_DOW A_CT A_WEATHER A_LT A_MC
## 1        2        1      2        2     1     1    2         1    1    2
## 2        2        1      2        2     2     1    2         2    2    2
## 3        2        1      2        2     2     1    3        13    1    2
## 4        2        1      2        1     2     1    1         1    1    2
```

```
## 5        2        1      2        1     2     1    1         6    2    2
## 6        2        1      2        5     1     1    2         1    1    2
##   A_SPCRA A_PED A_PED_F A_PEDAL A_PEDAL_F A_ROLL A_POLPUR A_POSBAC A_D15_19
## 1       1     2       2       2         2      2        2        3        2
## 2       1     2       2       2         2      1        2        3        2
## 3       1     2       2       2         2      2        2        3        2
## 4       2     1       1       2         2      1        2        3        2
## 5       1     2       2       2         2      1        2        1        2
## 6       2     2       2       2         2      2        2        3        2
##   A_D16_19 A_D15_20 A_D16_20 A_D65PLS A_D21_24 A_D16_24 A_RD A_HR A_DIST
## 1        2        2        2        2        2        2    2    2      2
## 2        2        2        2        2        2        2    2    2      2
## 3        2        2        2        2        1        1    2    2      2
## 4        2        2        2        2        2        2    2    2      2
## 5        2        2        2        2        2        2    1    2      2
## 6        2        2        2        1        2        2    2    2      2
##   A_DROWSY BIA SPJ_INDIAN INDIAN_RES
## 1        2   0          0          0
## 2        2   0          0          0
## 3        2   0          0          0
## 4        2   0          0          0
## 5        2   0          0          0
## 6        2   0          0          0
```

```r
# An analysis of numeric variables
summary(acc19)
```

```
##     ST_CASE            YEAR          STATE           COUNTY
##  Min.   : 10001   Min.   :2019   Min.   : 1.00   Min.   :  0.00
##  1st Qu.:121862   1st Qu.:2019   1st Qu.:12.00   1st Qu.: 31.00
##  Median :260792   Median :2019   Median :26.00   Median : 71.00
##  Mean   :272182   Mean   :2019   Mean   :27.14   Mean   : 92.17
##  3rd Qu.:420474   3rd Qu.:2019   3rd Qu.:42.00   3rd Qu.:115.00
##  Max.   :560121   Max.   :2019   Max.   :56.00   Max.   :997.00
##      FATALS         A_CRAINJ    A_REGION           A_RU          A_INTER
##  Min.   :1.000   Min.   :1   Min.   : 1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:1   1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:2.00
##  Median :1.000   Median :1   Median : 5.000   Median :2.000   Median :2.00
##  Mean   :1.086   Mean   :1   Mean   : 5.407   Mean   :1.566   Mean   :1.88
##  3rd Qu.:1.000   3rd Qu.:1   3rd Qu.: 7.000   3rd Qu.:2.000   3rd Qu.:2.00
##  Max.   :8.000   Max.   :1   Max.   :10.000   Max.   :3.000   Max.   :3.00
##     A_RELRD         A_INTSEC        A_ROADFC         A_JUNC         A_MANCOL
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1.00
##  Median :1.000   Median :2.000   Median :4.000   Median :2.000   Median :1.00
##  Mean   :2.048   Mean   :1.751   Mean   :3.682   Mean   :1.829   Mean   :1.98
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:3.00
##  Max.   :6.000   Max.   :3.000   Max.   :7.000   Max.   :4.000   Max.   :7.00
##      A_TOD           A_DOW            A_CT         A_WEATHER
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 1.00
##  Median :2.000   Median :1.000   Median :1.000   Median : 1.00
##  Mean   :1.526   Mean   :1.403   Mean   :1.506   Mean   :10.25
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 2.00
```

```
##  Max.   :3.000   Max.   :3.000   Max.   :3.000   Max.   :99.00
##       A_LT            A_MC          A_SPCRA          A_PED
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.866   Mean   :1.852   Mean   :1.742   Mean   :1.813
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_PED_F         A_PEDAL        A_PEDAL_F         A_ROLL
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.815   Mean   :1.974   Mean   :1.974   Mean   :1.785
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_POLPUR        A_POSBAC        A_D15_19        A_D16_19
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :2.000   Median :2.000
##  Mean   :1.989   Mean   :2.352   Mean   :1.913   Mean   :1.915
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :3.000   Max.   :2.000   Max.   :2.000
##     A_D15_20        A_D16_20        A_D65PLS        A_D21_24
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.885   Mean   :1.887   Mean   :1.788   Mean   :1.867
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_D16_24          A_RD            A_HR            A_DIST
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.761   Mean   :1.512   Mean   :1.941   Mean   :1.914
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_DROWSY          BIA             SPJ_INDIAN         INDIAN_RES
##  Min.   :1.000   Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:2.000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :2.000   Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :1.981   Mean   :0.007077   Mean   :0.004688   Mean   :0.008063
##  3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :2.000   Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
```

```
# The names of the variables in the accident table
names(acc19)
```

```
##  [1] "ST_CASE"    "YEAR"       "STATE"      "COUNTY"     "FATALS"
##  [6] "A_CRAINJ"   "A_REGION"   "A_RU"       "A_INTER"    "A_RELRD"
## [11] "A_INTSEC"   "A_ROADFC"   "A_JUNC"     "A_MANCOL"   "A_TOD"
## [16] "A_DOW"      "A_CT"       "A_WEATHER"  "A_LT"       "A_MC"
## [21] "A_SPCRA"    "A_PED"      "A_PED_F"    "A_PEDAL"    "A_PEDAL_F"
## [26] "A_ROLL"     "A_POLPUR"   "A_POSBAC"   "A_D15_19"   "A_D16_19"
## [31] "A_D15_20"   "A_D16_20"   "A_D65PLS"   "A_D21_24"   "A_D16_24"
```

```
## [36] "A_RD"       "A_HR"       "A_DIST"     "A_DROWSY"   "BIA"
## [41] "SPJ_INDIAN" "INDIAN_RES"
```

```
# and perform a general analysis of the structure
summario19 <- as.data.frame(str(acc19, give.attr = FALSE, strict.width = "cut"))
```

```
## 'data.frame':    33487 obs. of  42 variables:
##  $ ST_CASE   : int  10001 10002 10003 10004 10005 10006 10007 10008 10009 1001..
##  $ YEAR      : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
##  $ STATE     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY    : int  81 55 29 55 3 85 55 69 3 101 ...
##  $ FATALS    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ A_CRAINJ  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ A_REGION  : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ A_RU      : int  2 2 1 1 2 1 1 1 2 2 ...
##  $ A_INTER   : int  1 1 1 1 1 1 1 2 2 2 ...
##  $ A_RELRD   : int  1 1 1 1 4 1 3 1 4 1 ...
##  $ A_INTSEC  : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ A_ROADFC  : int  1 1 1 1 1 1 1 6 4 5 ...
##  $ A_JUNC    : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ A_MANCOL  : int  2 2 2 1 1 5 1 1 1 1 ...
##  $ A_TOD     : int  1 2 2 2 2 1 1 2 1 2 ...
##  $ A_DOW     : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ A_CT      : int  2 2 3 1 1 2 1 1 1 1 ...
##  $ A_WEATHER : int  1 2 13 1 6 1 13 1 1 1 ...
##  $ A_LT      : int  1 2 1 1 2 1 2 2 2 2 ...
##  $ A_MC      : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_SPCRA   : int  1 1 1 2 1 2 2 1 2 2 ...
##  $ A_PED     : int  2 2 2 1 2 2 2 2 2 1 ...
##  $ A_PED_F   : int  2 2 2 1 2 2 2 2 2 1 ...
##  $ A_PEDAL   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_PEDAL_F : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_ROLL    : int  2 1 2 1 1 2 1 1 2 2 ...
##  $ A_POLPUR  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_POSBAC  : int  3 3 3 3 1 3 2 1 2 3 ...
##  $ A_D15_19  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D16_19  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D15_20  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D16_20  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D65PLS  : int  2 2 2 2 2 1 1 2 1 2 ...
##  $ A_D21_24  : int  2 2 1 2 2 2 2 2 2 2 ...
##  $ A_D16_24  : int  2 2 1 2 2 2 2 2 2 2 ...
##  $ A_RD      : int  2 2 2 2 1 2 1 1 1 2 ...
##  $ A_HR      : int  2 2 2 2 2 2 2 2 2 1 ...
##  $ A_DIST    : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_DROWSY  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ BIA       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SPJ_INDIAN: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INDIAN_RES: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
knitr::kable(summario19, caption = "Summary of the acc20 table")
```

Table: Summary of the acc20 table

We can observe that in this case, the data in the 2019 'Auxiliary' file are of integral type, and we have 33,487 instances or accidents with 42 variables. The file lacks descriptions in text strings, but the variable names and primary key match the previous one, so we can unify them using the table as a child table and using the primary key as a foreign key.

ACCIDENTS 2018

```
# Auxiliary file acc18

# Let's check the first rows of data for each variable
head(acc18)
```

```
##   ST_CASE YEAR STATE COUNTY FATALS A_CRAINJ A_REGION A_RU A_INTER A_RELRD
## 1   10001 2018     1    121      1        1        4    1       1       2
## 2   10002 2018     1    127      2        1        4    2       1       3
## 3   10003 2018     1     21      1        1        4    1       1       4
## 4   10004 2018     1      3      1        1        4    1       1       4
## 5   10005 2018     1     73      1        1        4    2       1       1
## 6   10006 2018     1     49      1        1        4    1       1       2
##   A_INTSEC A_ROADFC A_JUNC A_MANCOL A_TOD A_DOW A_CT A_WEATHER A_LT A_MC
## 1        2        1      2        1     1     1    1         1    1    2
## 2        2        1      3        1     2     2    1         2    2    2
## 3        2        1      2        1     2     1    2        13    1    2
## 4        2        1      2        1     1     1    1        13    2    2
## 5        2        1      2        2     1     1    2         1    1    2
## 6        2        1      2        1     2     2    1         1    2    2
##   A_SPCRA A_PED A_PED_F A_PEDAL A_PEDAL_F A_ROLL A_POLPUR A_POSBAC A_D15_19
## 1       2     2       2       2         2      1        2        2        2
## 2       1     2       2       2         2      2        1        3        2
## 3       2     2       2       2         2      2        2        3        2
## 4       2     2       2       2         2      2        2        2        2
## 5       2     2       2       2         2      2        2        3        2
## 6       1     1       1       2         2      2        2        2        2
##   A_D16_19 A_D15_20 A_D16_20 A_D65PLS A_D21_24 A_D16_24 A_RD A_HR A_DIST
## 1        2        2        2        2        2        2    1    2      2
## 2        2        2        2        2        1        1    1    2      2
## 3        2        2        2        2        2        2    1    2      2
## 4        2        2        2        2        2        2    1    2      2
## 5        2        2        2        2        2        2    2    2      2
## 6        2        2        2        2        1        1    1    2      2
##   A_DROWSY BIA SPJ_INDIAN INDIAN_RES
## 1        2   0          0          0
## 2        2   0          0          0
## 3        2   0          0          0
## 4        2   0          0          0
## 5        2   0          0          0
## 6        2   0          0          0
```

```
# An analysis of numeric variables
summary(acc18)
```

```
##     ST_CASE            YEAR          STATE           COUNTY
##  Min.   : 10001   Min.   :2018   Min.   : 1.00   Min.   :  1.00
##  1st Qu.:121777   1st Qu.:2018   1st Qu.:12.00   1st Qu.: 31.00
```

```
##  Median :260782   Median :2018   Median :26.00   Median : 71.00
##  Mean   :271496   Mean   :2018   Mean   :27.07   Mean   : 91.53
##  3rd Qu.:420504   3rd Qu.:2018   3rd Qu.:42.00   3rd Qu.:114.00
##  Max.   :560101   Max.   :2018   Max.   :56.00   Max.   :999.00
##      FATALS          A_CRAINJ    A_REGION          A_RU          A_INTER
##  Min.   : 1.000   Min.   :1   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.:1   1st Qu.: 4.000   1st Qu.:1.000   1st Qu.:2.000
##  Median : 1.000   Median :1   Median : 5.000   Median :2.000   Median :2.000
##  Mean   : 1.086   Mean   :1   Mean   : 5.398   Mean   :1.577   Mean   :1.877
##  3rd Qu.: 1.000   3rd Qu.:1   3rd Qu.: 7.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :20.000   Max.   :1   Max.   :10.000   Max.   :3.000   Max.   :3.000
##     A_RELRD        A_INTSEC        A_ROADFC         A_JUNC         A_MANCOL
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.00   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.00   Median :2.000   Median :4.000   Median :2.000   Median :1.000
##  Mean   :2.06   Mean   :1.756   Mean   :3.684   Mean   :1.836   Mean   :1.969
##  3rd Qu.:4.00   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :6.00   Max.   :3.000   Max.   :7.000   Max.   :4.000   Max.   :7.000
##      A_TOD          A_DOW           A_CT         A_WEATHER          A_LT
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   : 1.00   Min.   :1.000
##  1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 1.00   1st Qu.:2.000
##  Median :2.00   Median :1.000   Median :1.000   Median : 1.00   Median :2.000
##  Mean   :1.53   Mean   :1.405   Mean   :1.507   Mean   :10.56   Mean   :1.868
##  3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 4.00   3rd Qu.:2.000
##  Max.   :3.00   Max.   :3.000   Max.   :3.000   Max.   :99.00   Max.   :2.000
##       A_MC          A_SPCRA          A_PED          A_PED_F
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.854   Mean   :1.746   Mean   :1.813   Mean   :1.815
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_PEDAL        A_PEDAL_F         A_ROLL         A_POLPUR
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.974   Mean   :1.974   Mean   :1.782   Mean   :1.989
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_POSBAC        A_D15_19        A_D16_19        A_D15_20
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :2.341   Mean   :1.911   Mean   :1.913   Mean   :1.881
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :3.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##     A_D16_20        A_D65PLS        A_D21_24        A_D16_24
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.884   Mean   :1.799   Mean   :1.864   Mean   :1.754
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##       A_RD            A_HR           A_DIST         A_DROWSY
```

```
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.502   Mean   :1.941   Mean   :1.922   Mean   :1.979
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##       BIA             SPJ_INDIAN         INDIAN_RES
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.000000   Median :0.000000
##  Mean   :0.007665   Mean   :0.006457   Mean   :0.009316
##  3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.000000
```

```
# The names of the variables in the accidents table
names(acc18)
```

```
##  [1] "ST_CASE"    "YEAR"       "STATE"      "COUNTY"     "FATALS"
##  [6] "A_CRAINJ"   "A_REGION"   "A_RU"       "A_INTER"    "A_RELRD"
## [11] "A_INTSEC"   "A_ROADFC"   "A_JUNC"     "A_MANCOL"   "A_TOD"
## [16] "A_DOW"      "A_CT"       "A_WEATHER"  "A_LT"       "A_MC"
## [21] "A_SPCRA"    "A_PED"      "A_PED_F"    "A_PEDAL"    "A_PEDAL_F"
## [26] "A_ROLL"     "A_POLPUR"   "A_POSBAC"   "A_D15_19"   "A_D16_19"
## [31] "A_D15_20"   "A_D16_20"   "A_D65PLS"   "A_D21_24"   "A_D16_24"
## [36] "A_RD"       "A_HR"       "A_DIST"     "A_DROWSY"   "BIA"
## [41] "SPJ_INDIAN" "INDIAN_RES"
```

```
# And let's perform a general analysis of the structure
summary18 <- as.data.frame(str(acc18, give.attr = FALSE, strict.width = "cut"))
```

```
## 'data.frame':    33919 obs. of  42 variables:
##  $ ST_CASE   : int  10001 10002 10003 10004 10005 10006 10007 10008 10009 1001..
##  $ YEAR      : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
##  $ STATE     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY    : int  121 127 21 3 73 49 97 73 97 117 ...
##  $ FATALS    : int  1 2 1 1 1 1 1 1 1 1 ...
##  $ A_CRAINJ  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ A_REGION  : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ A_RU      : int  1 2 1 1 2 1 2 2 2 2 ...
##  $ A_INTER   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ A_RELRD   : int  2 3 4 4 1 2 1 1 4 1 ...
##  $ A_INTSEC  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_ROADFC  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ A_JUNC    : int  2 3 2 2 2 2 2 2 3 2 ...
##  $ A_MANCOL  : int  1 1 1 1 2 1 1 2 1 1 ...
##  $ A_TOD     : int  1 2 2 1 1 2 1 2 2 2 ...
##  $ A_DOW     : int  1 2 1 1 1 2 2 1 2 2 ...
##  $ A_CT      : int  1 1 2 1 2 1 1 2 1 1 ...
##  $ A_WEATHER : int  1 2 13 13 1 1 1 13 13 1 ...
##  $ A_LT      : int  1 2 1 2 1 2 2 2 2 2 ...
##  $ A_MC      : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ A_SPCRA   : int  2 1 2 2 2 1 2 2 1 2 ...
##  $ A_PED     : int  2 2 2 2 2 1 1 1 2 1 ...
```

```
##  $ A_PED_F   : int  2 2 2 2 2 1 1 1 2 1 ...
##  $ A_PEDAL   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_PEDAL_F : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_ROLL    : int  1 2 2 2 2 2 2 2 2 2 ...
##  $ A_POLPUR  : int  2 1 2 2 2 2 2 2 2 2 ...
##  $ A_POSBAC  : int  2 3 3 2 3 2 3 3 1 3 ...
##  $ A_D15_19  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D16_19  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D15_20  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D16_20  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_D65PLS  : int  2 2 2 2 2 2 1 2 2 2 ...
##  $ A_D21_24  : int  2 1 2 2 2 1 2 2 2 1 ...
##  $ A_D16_24  : int  2 1 2 2 2 1 2 2 2 1 ...
##  $ A_RD      : int  1 1 1 1 2 1 2 2 1 2 ...
##  $ A_HR      : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_DIST    : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ A_DROWSY  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ BIA       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SPJ_INDIAN: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ INDIAN_RES: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
knitr::kable(summary18, caption = "Summary of the acc18 table")
```

Table: Summary of the acc18 table

We observe that the accidents table for 2018 contains 33,919 instances and the same 42 variables as in the previous case, with analogous structures.

Furthermore, the primary/foreign key is the case number:

ST_CASE - Accident identifier

We will study the following aspects of the data:

TEMPORAL TREND STUDY

FATAL - Annual fatalities (2018 to 2020)

CAUSE STUDY

DRUNK_DR - Number of drunk drivers

DAY_WEEKNAME - Interest in weekends

HOUR - Hour

HOURNAME - Time slot

MINUTE - Minute

DRDISTRAC - Code for driver distractions

DRDISTRACNAME - Distraction specification

MOD_YEAR - Year of vehicle model

L_TYPE - License type, but we are interested only in those without a license

WEATHER - Code for weather conditions at the time

WEATHERNAME - Weather condition specification

DRUNK_DR - Number of positive alcohol tests for drivers involved

DRIVERRF - Type of driving infractions or aspects. This record comes from a table that allows us to select the involved or responsible drivers

AGE - Age, filtering the age of officially responsible drivers

DRIVERRFNAME - Textual specification of the responsible driver

AIR_BAG - Code to specify airbag behavior

AIR_BAGNAME - Textual specification of airbag behavior

DRUGS - Codes for the presence of drugs

DRUGSNAME - Textual specification

VISION - Codes for visibility elements (mirrors, windows, etc.)

VISIONNAME - Specification of the type of anomaly in vision-related elements

VEHICLECC - Code for factors that may contribute to the accident

VEHICLECCNAME - Specification of the factors

DRIMPAIR - Codes for detected physical impairments

DRIMPAIRNAME - Specification of detected psychophysical impairment aspects

LOCATION OF BLACK SPOTS

COUNTY - County code

COUNTYNAME - County name

CITY - City code

CITYNAME - City name

ROUTE - Route code or road type

ROUTE - Specification of the road type

RUR_URB - Code 1-2 to specify if it is rural or urban, respectively

MILEPT - Mile point number

LATITUDE

LONGITUDE

2. Feature processing and management I: Cleaning

The next step is to ensure that there are no empty or null values.

```
print('NA')
```

```
## [1] "NA"
```

```
# Checking the accident files
print ('Accidents 2020')
```

```
## [1] "Accidents 2020"
```

```
colSums(is.na(acc20))
```

```
##        STATE    STATENAME      ST_CASE     VE_TOTAL     VE_FORMS     PVH_INVL
##            0            0            0            0            0            0
##         PEDS      PERSONS      PERMVIT   PERNOTMVIT       COUNTY   COUNTYNAME
##            0            0            0            0            0            0
##         CITY     CITYNAME          DAY      DAYNAME        MONTH    MONTHNAME
##            0            0            0            0            0            0
##         YEAR     DAY_WEEK DAY_WEEKNAME         HOUR     HOURNAME       MINUTE
##            0            0            0            0            0            0
##   MINUTENAME          NHS      NHSNAME        ROUTE    ROUTENAME      TWAY_ID
##            0            0            0            0            0            0
##     TWAY_ID2      RUR_URB  RUR_URBNAME     FUNC_SYS FUNC_SYSNAME     RD_OWNER
##            0            0            0            0            0            0
## RD_OWNERNAME       MILEPT   MILEPTNAME     LATITUDE LATITUDENAME     LONGITUD
##            0            0            0            0            0            0
## LONGITUDNAME       SP_JUR   SP_JURNAME      HARM_EV  HARM_EVNAME     MAN_COLL
##            0            0            0            0            0            0
## MAN_COLLNAME      RELJCT1  RELJCT1NAME      RELJCT2  RELJCT2NAME      TYP_INT
##            0            0            0            0            0            0
##  TYP_INTNAME     WRK_ZONE WRK_ZONENAME     REL_ROAD REL_ROADNAME     LGT_COND
##            0            0            0            0            0            0
## LGT_CONDNAME      WEATHER  WEATHERNAME      SCH_BUS  SCH_BUSNAME         RAIL
##            0            0            0            0            0            0
##     RAILNAME     NOT_HOUR NOT_HOURNAME      NOT_MIN  NOT_MINNAME     ARR_HOUR
##            0            0            0            0            0            0
## ARR_HOURNAME      ARR_MIN  ARR_MINNAME      HOSP_HR  HOSP_HRNAME      HOSP_MN
##            0            0            0            0            0            0
##  HOSP_MNNAME       FATALS     DRUNK_DR
##            0            0            0
```

```
print('Accidents 2019')
```

```
## [1] "Accidents 2019"
```

```
colSums(is.na(acc19))
```

```
##    ST_CASE       YEAR      STATE     COUNTY     FATALS   A_CRAINJ   A_REGION
##          0          0          0          0          0          0          0
##       A_RU    A_INTER    A_RELRD   A_INTSEC   A_ROADFC     A_JUNC   A_MANCOL
##          0          0          0          0          0          0          0
##      A_TOD      A_DOW       A_CT  A_WEATHER       A_LT       A_MC    A_SPCRA
##          0          0          0          0          0          0          0
##      A_PED    A_PED_F    A_PEDAL  A_PEDAL_F     A_ROLL   A_POLPUR   A_POSBAC
##          0          0          0          0          0          0          0
##   A_D15_19   A_D16_19   A_D15_20   A_D16_20   A_D65PLS   A_D21_24   A_D16_24
##          0          0          0          0          0          0          0
##       A_RD       A_HR     A_DIST   A_DROWSY        BIA SPJ_INDIAN INDIAN_RES
##          0          0          0          0          0          0          0
```

```
print('Accidents 2018')
```

```
## [1] "Accidents 2018"
```

```
colSums(is.na(acc18))
```

```
##    ST_CASE       YEAR      STATE     COUNTY     FATALS   A_CRAINJ   A_REGION
##          0          0          0          0          0          0          0
##       A_RU    A_INTER    A_RELRD   A_INTSEC   A_ROADFC     A_JUNC   A_MANCOL
##          0          0          0          0          0          0          0
##      A_TOD      A_DOW       A_CT  A_WEATHER       A_LT       A_MC    A_SPCRA
##          0          0          0          0          0          0          0
##      A_PED    A_PED_F    A_PEDAL  A_PEDAL_F     A_ROLL   A_POLPUR   A_POSBAC
##          0          0          0          0          0          0          0
##   A_D15_19   A_D16_19   A_D15_20   A_D16_20   A_D65PLS   A_D21_24   A_D16_24
##          0          0          0          0          0          0          0
##       A_RD       A_HR     A_DIST   A_DROWSY        BIA SPJ_INDIAN INDIAN_RES
##          0          0          0          0          0          0          0
```

```
# Checking the auxiliary files
print('driverrf')
```

```
## [1] "driverrf"
```

```
colSums(is.na(drivrf))
```

```
##        STATE    STATENAME      ST_CASE       VEH_NO     DRIVERRF DRIVERRFNAME
##            0            0            0            0            0            0
```

```
print('Factor')
```

```
## [1] "Factor"
```

```
colSums(is.na(fact))
```

```
##         STATE     STATENAME       ST_CASE        VEH_NO     VEHICLECC
##             0             0             0             0             0
## VEHICLECCNAME
##             0
```

```
print('impair')
```

```
## [1] "impair"
```

```
colSums(is.na(impair))
```

```
##        STATE    STATENAME      ST_CASE       VEH_NO     DRIMPAIR DRIMPAIRNAME
##            0            0            0            0            0            0
```

```
print('Person')
```

```
## [1] "Person"
```

```
colSums(is.na(per))
```

```
##             STATE         STATENAME           ST_CASE          VE_FORMS
##                 0                 0                 0                 0
##            VEH_NO            PER_NO           STR_VEH            COUNTY
##                 0                 0                 0                 0
##               DAY           DAYNAME             MONTH         MONTHNAME
##                 0                 0                 0                 0
##              HOUR          HOURNAME            MINUTE        MINUTENAME
##                 0                 0                 0                 0
##           RUR_URB       RUR_URBNAME          FUNC_SYS      FUNC_SYSNAME
##                 0                 0                 0                 0
##           HARM_EV       HARM_EVNAME          MAN_COLL      MAN_COLLNAME
##                 0                 0                 0                 0
##           SCH_BUS       SCH_BUSNAME              MAKE          MAKENAME
##                 0                 0              8172                 0
##           MAK_MOD          BODY_TYP     BODY_TYPNAME          MOD_YEAR
##                 0              8172                 0              8172
##      MOD_YEARNAME           TOW_VEH      TOW_VEHNAME          SPEC_USE
##                 0              8172                 0              8172
##      SPEC_USENAME          EMER_USE     EMER_USENAME          ROLLOVER
##                 0              8172                 0              8172
##      ROLLOVERNAME           IMPACT1      IMPACT1NAME          FIRE_EXP
##                 0              8172                 0              8172
##      FIRE_EXPNAME               AGE          AGENAME               SEX
##                 0                 0                 0                 0
##           SEXNAME           PER_TYP      PER_TYPNAME           INJ_SEV
##                 0                 0                 0                 0
##       INJ_SEVNAME          SEAT_POS     SEAT_POSNAME          REST_USE
##                 0                 0                 0                 0
##      REST_USENAME          REST_MIS     REST_MISNAME           AIR_BAG
##                 0                 0                 0                 0
##       AIR_BAGNAME          EJECTION     EJECTIONNAME           EJ_PATH
##                 0                 0                 0                 0
##       EJ_PATHNAME          EXTRICAT     EXTRICATNAME          DRINKING
##                 0                 0                 0                 0
##      DRINKINGNAME           ALC_DET      ALC_DETNAME         ALC_STATUS
##                 0                 0                 0                 0
##    ALC_STATUSNAME          ATST_TYP     ATST_TYPNAME           ALC_RES
##                 0                 0                 0                 0
##       ALC_RESNAME             DRUGS        DRUGSNAME          DRUG_DET
##                 0                 0                 0                 0
##      DRUG_DETNAME           DSTATUS      DSTATUSNAME          HOSPITAL
##                 0                 0                 0                 0
##      HOSPITALNAME               DOA          DOANAME          DEATH_DA
##                 0                 0                 0                 0
##      DEATH_DANAME          DEATH_MO     DEATH_MONAME          DEATH_YR
##                 0                 0                 0                 0
```

```
##      DEATH_YRNAME          DEATH_HR      DEATH_HRNAME          DEATH_MN
##                 0                 0                 0                 0
##      DEATH_MNNAME          DEATH_TM      DEATH_TMNAME           LAG_HRS
##                 0                 0                 0                 0
##       LAG_HRSNAME          LAG_MINS      LAG_MINSNAME          WORK_INJ
##                 0                 0                 0                 0
##      WORK_INJNAME          HISPANIC      HISPANICNAME          LOCATION
##                 0                 0                 0                 0
##      LOCATIONNAME          HELM_USE      HELM_USENAME          HELM_MIS
##                 0                 0                 0                 0
##      HELM_MISNAME          VPICMAKE      VPICMAKENAME         VPICMODEL
##                 0              8172                 0              8172
##     VPICMODELNAME     VPICBODYCLASS VPICBODYCLASSNAME       ICFINALBODY
##                 0              8172                 0              8172
##   ICFINALBODYNAME
##                 0
```

```
print('Vehicles')
```

```
## [1] "Vehicles"
```

```
colSums(is.na(veh))
```

```
##             STATE         STATENAME           ST_CASE           VEH_NO
##                 0                 0                 0                 0
##          VE_FORMS           NUMOCCS       NUMOCCSNAME               DAY
##                 0                 0                 0                 0
##           DAYNAME             MONTH         MONTHNAME              HOUR
##                 0                 0                 0                 0
##          HOURNAME            MINUTE        MINUTENAME           HARM_EV
##                 0                 0                 0                 0
##       HARM_EVNAME          MAN_COLL      MAN_COLLNAME          UNITTYPE
##                 0                 0                 0                 0
##      UNITTYPENAME           HIT_RUN       HIT_RUNNAME          REG_STAT
##                 0                 0                 0                 0
##      REG_STATNAME             OWNER         OWNERNAME              MAKE
##                 0                 0                 0                 0
##          MAKENAME             MODEL           MAK_MOD       MAK_MODNAME
##                 0                 0                 0                 0
##          BODY_TYP      BODY_TYPNAME          MOD_YEAR      MOD_YEARNAME
##                 0                 0                 0                 0
##               VIN           VINNAME             VIN_1             VIN_2
##                 0                 0                 0                 0
##             VIN_3             VIN_4             VIN_5             VIN_6
##                 0                 0                 0                 0
##             VIN_7             VIN_8             VIN_9            VIN_10
##                 0                 0                 0                 0
##            VIN_11            VIN_12           TOW_VEH       TOW_VEHNAME
##                 0                 0                 0                 0
##           J_KNIFE       J_KNIFENAME          MCARR_I1      MCARR_I1NAME
##                 0                 0                 0                 0
##          MCARR_I2      MCARR_I2NAME          MCARR_ID      MCARR_IDNAME
##                 0                 0                 0                 0
```

```
##          V_CONFIG      V_CONFIGNAME          CARGO_BT      CARGO_BTNAME
##                 0                 0                 0                 0
##           HAZ_INV       HAZ_INVNAME          HAZ_PLAC      HAZ_PLACNAME
##                 0                 0                 0                 0
##            HAZ_ID        HAZ_IDNAME           HAZ_CNO       HAZ_CNONAME
##                 0                 0                 0                 0
##           HAZ_REL       HAZ_RELNAME           BUS_USE       BUS_USENAME
##                 0                 0                 0                 0
##          SPEC_USE      SPEC_USENAME          EMER_USE      EMER_USENAME
##                 0                 0                 0                 0
##           TRAV_SP       TRAV_SPNAME          UNDERIDE      UNDERIDENAME
##                 0                 0                 0                 0
##          ROLLOVER      ROLLOVERNAME          ROLINLOC      ROLINLOCNAME
##                 0                 0                 0                 0
##           IMPACT1       IMPACT1NAME          DEFORMED      DEFORMEDNAME
##                 0                 0                 0                 0
##             TOWED         TOWEDNAME            M_HARM        M_HARMNAME
##                 0                 0                 0                 0
##          FIRE_EXP      FIRE_EXPNAME           DR_PRES       DR_PRESNAME
##                 0                 0                 0                 0
##           L_STATE       L_STATENAME            DR_ZIP        DR_ZIPNAME
##                 0                 0                 0                 0
##          L_STATUS      L_STATUSNAME            L_TYPE        L_TYPENAME
##                 0                 0                 0                 0
##          CDL_STAT      CDL_STATNAME          L_ENDORS      L_ENDORSNAME
##                 0                 0                 0                 0
##           L_COMPL       L_COMPLNAME          L_RESTRI      L_RESTRINAME
##                 0                 0                 0                 0
##            DR_HGT        DR_HGTNAME            DR_WGT        DR_WGTNAME
##                 0                 0                 0                 0
##          PREV_ACC      PREV_ACCNAME         PREV_SUS1     PREV_SUS1NAME
##                 0                 0                 0                 0
##         PREV_SUS2     PREV_SUS2NAME         PREV_SUS3     PREV_SUS3NAME
##                 0                 0                 0                 0
##          PREV_DWI      PREV_DWINAME          PREV_SPD      PREV_SPDNAME
##                 0                 0                 0                 0
##          PREV_OTH      PREV_OTHNAME          FIRST_MO      FIRST_MONAME
##                 0                 0                 0                 0
##          FIRST_YR      FIRST_YRNAME           LAST_MO       LAST_MONAME
##                 0                 0                 0                 0
##           LAST_YR       LAST_YRNAME          SPEEDREL      SPEEDRELNAME
##                 0                 0                 0                 0
##          VTRAFWAY      VTRAFWAYNAME          VNUM_LAN      VNUM_LANNAME
##                 0                 0                 0                 0
##          VSPD_LIM      VSPD_LIMNAME            VALIGN        VALIGNNAME
##                 0                 0                 0                 0
##          VPROFILE      VPROFILENAME          VPAVETYP      VPAVETYPNAME
##                 0                 0                 0                 0
##          VSURCOND      VSURCONDNAME          VTRAFCON      VTRAFCONNAME
##                 0                 0                 0                 0
##          VTCONT_F      VTCONT_FNAME          P_CRASH1      P_CRASH1NAME
##                 0                 0                 0                 0
##          P_CRASH2      P_CRASH2NAME          P_CRASH3      P_CRASH3NAME
##                 0                 0                 0                 0
```

```
##           PCRASH4       PCRASH4NAME           PCRASH5       PCRASH5NAME
##                 0                 0                 0                 0
##          ACC_TYPE      ACC_TYPENAME            DEATHS          DR_DRINK
##                 0                 0                 0                 0
##      DR_DRINKNAME          TRLR1VIN      TRLR1VINNAME          TRLR2VIN
##                 0                 0                 0                 0
##      TRLR2VINNAME          TRLR3VIN      TRLR3VINNAME          VPICMAKE
##                 0                 0                 0                 0
##      VPICMAKENAME         VPICMODEL     VPICMODELNAME     VPICBODYCLASS
##                 0                 0                 0                 0
## VPICBODYCLASSNAME       ICFINALBODY   ICFINALBODYNAME         GVWR_FROM
##                 0                 0                 0                 0
##     GVWR_FROMNAME           GVWR_TO       GVWR_TONAME         TRLR1GVWR
##                 0                 0                 0                 0
##     TRLR1GVWRNAME         TRLR2GVWR     TRLR2GVWRNAME         TRLR3GVWR
##                 0                 0                 0                 0
##     TRLR3GVWRNAME
##                 0
```

```
print('vision')
```

```
## [1] "vision"
```

```
colSums(is.na(visio))
```

```
##      STATE  STATENAME    ST_CASE     VEH_NO     VISION VISIONNAME
##          0          0          0          0          0          0
```

```
print('Blanks')
```

```
## [1] "Blanks"
```

```
# Checking the accident files
print ('Accidents 2020')
```

```
## [1] "Accidents 2020"
```

```
colSums(acc20=="")
```

```
##        STATE    STATENAME      ST_CASE     VE_TOTAL     VE_FORMS     PVH_INVL
##            0            0            0            0            0            0
##         PEDS      PERSONS      PERMVIT   PERNOTMVIT       COUNTY   COUNTYNAME
##            0            0            0            0            0            0
##         CITY     CITYNAME          DAY      DAYNAME        MONTH    MONTHNAME
##            0            0            0            0            0            0
##         YEAR     DAY_WEEK DAY_WEEKNAME         HOUR     HOURNAME       MINUTE
##            0            0            0            0            0            0
##   MINUTENAME          NHS      NHSNAME        ROUTE    ROUTENAME      TWAY_ID
##            0            0            0            0            0            0
##     TWAY_ID2      RUR_URB  RUR_URBNAME     FUNC_SYS FUNC_SYSNAME     RD_OWNER
```

```
##        26997           0           0           0           0           0
## RD_OWNERNAME       MILEPT   MILEPTNAME     LATITUDE LATITUDENAME     LONGITUD
##           0           0           0           0           0           0
## LONGITUDNAME       SP_JUR   SP_JURNAME      HARM_EV  HARM_EVNAME     MAN_COLL
##           0           0           0           0           0           0
## MAN_COLLNAME      RELJCT1  RELJCT1NAME      RELJCT2  RELJCT2NAME      TYP_INT
##           0           0           0           0           0           0
##  TYP_INTNAME     WRK_ZONE WRK_ZONENAME     REL_ROAD REL_ROADNAME     LGT_COND
##           0           0           0           0           0           0
## LGT_CONDNAME      WEATHER  WEATHERNAME      SCH_BUS  SCH_BUSNAME         RAIL
##           0           0           0           0           0           0
##     RAILNAME     NOT_HOUR NOT_HOURNAME      NOT_MIN  NOT_MINNAME     ARR_HOUR
##           0           0           0           0           0           0
## ARR_HOURNAME      ARR_MIN  ARR_MINNAME      HOSP_HR  HOSP_HRNAME      HOSP_MN
##           0           0           0           0           0           0
##  HOSP_MNNAME       FATALS     DRUNK_DR
##           0           0           0
```

```
print('Accidents 2019')
```

```
## [1] "Accidents 2019"
```

```
colSums(acc19=="")
```

```
##    ST_CASE       YEAR      STATE     COUNTY     FATALS   A_CRAINJ   A_REGION
##          0          0          0          0          0          0          0
##       A_RU    A_INTER    A_RELRD   A_INTSEC   A_ROADFC     A_JUNC   A_MANCOL
##          0          0          0          0          0          0          0
##      A_TOD      A_DOW       A_CT  A_WEATHER       A_LT       A_MC    A_SPCRA
##          0          0          0          0          0          0          0
##      A_PED    A_PED_F    A_PEDAL  A_PEDAL_F     A_ROLL   A_POLPUR   A_POSBAC
##          0          0          0          0          0          0          0
##   A_D15_19   A_D16_19   A_D15_20   A_D16_20   A_D65PLS   A_D21_24   A_D16_24
##          0          0          0          0          0          0          0
##       A_RD       A_HR     A_DIST   A_DROWSY        BIA SPJ_INDIAN INDIAN_RES
##          0          0          0          0          0          0          0
```

```
print('Accidents 2018')
```

```
## [1] "Accidents 2018"
```

```
colSums(acc18=="")
```

```
##    ST_CASE       YEAR      STATE     COUNTY     FATALS   A_CRAINJ   A_REGION
##          0          0          0          0          0          0          0
##       A_RU    A_INTER    A_RELRD   A_INTSEC   A_ROADFC     A_JUNC   A_MANCOL
##          0          0          0          0          0          0          0
##      A_TOD      A_DOW       A_CT  A_WEATHER       A_LT       A_MC    A_SPCRA
##          0          0          0          0          0          0          0
##      A_PED    A_PED_F    A_PEDAL  A_PEDAL_F     A_ROLL   A_POLPUR   A_POSBAC
##          0          0          0          0          0          0          0
```

```
##   A_D15_19   A_D16_19   A_D15_20   A_D16_20   A_D65PLS   A_D21_24   A_D16_24
##          0          0          0          0          0          0          0
##       A_RD       A_HR     A_DIST   A_DROWSY        BIA SPJ_INDIAN INDIAN_RES
##          0          0          0          0          0          0          0
```

```
# Checking the auxiliary files
print('driverrf')
```

```
## [1] "driverrf"
```

```
colSums(drivrf=="")
```

```
##        STATE    STATENAME      ST_CASE       VEH_NO     DRIVERRF DRIVERRFNAME
##            0            0            0            0            0            0
```

```
print('Factor')
```

```
## [1] "Factor"
```

```
colSums(fact=="")
```

```
##         STATE     STATENAME       ST_CASE        VEH_NO     VEHICLECC
##             0             0             0             0             0
## VEHICLECCNAME
##             0
```

```
print('impair')
```

```
## [1] "impair"
```

```
colSums(impair=="")
```

```
##        STATE    STATENAME      ST_CASE       VEH_NO     DRIMPAIR DRIMPAIRNAME
##            0            0            0            0            0            0
```

```
print('Person')
```

```
## [1] "Person"
```

```
colSums(per=="")
```

```
##             STATE         STATENAME           ST_CASE          VE_FORMS
##                 0                 0                 0                 0
##            VEH_NO           PER_NO           STR_VEH            COUNTY
##                 0                 0                 0                 0
##               DAY           DAYNAME             MONTH         MONTHNAME
##                 0                 0                 0                 0
##              HOUR          HOURNAME            MINUTE        MINUTENAME
```

```
##               0               0               0               0
##         RUR_URB      RUR_URBNAME         FUNC_SYS     FUNC_SYSNAME
##               0               0               0               0
##         HARM_EV      HARM_EVNAME         MAN_COLL     MAN_COLLNAME
##               0               0               0               0
##         SCH_BUS      SCH_BUSNAME            MAKE         MAKENAME
##               0               0              NA            8172
##         MAK_MOD         BODY_TYP     BODY_TYPNAME         MOD_YEAR
##            8172              NA            8172              NA
##     MOD_YEARNAME         TOW_VEH      TOW_VEHNAME         SPEC_USE
##            8172              NA            8172              NA
##     SPEC_USENAME         EMER_USE     EMER_USENAME         ROLLOVER
##            8172              NA            8172              NA
##     ROLLOVERNAME         IMPACT1      IMPACT1NAME         FIRE_EXP
##            8172              NA            8172              NA
##     FIRE_EXPNAME             AGE         AGENAME              SEX
##            8172               0               0               0
##         SEXNAME         PER_TYP      PER_TYPNAME         INJ_SEV
##               0               0               0               0
##      INJ_SEVNAME         SEAT_POS     SEAT_POSNAME         REST_USE
##               0               0               0               0
##     REST_USENAME         REST_MIS     REST_MISNAME          AIR_BAG
##               0               0               0               0
##      AIR_BAGNAME         EJECTION     EJECTIONNAME         EJ_PATH
##               0               0               0               0
##      EJ_PATHNAME         EXTRICAT     EXTRICATNAME         DRINKING
##               0               0               0               0
##     DRINKINGNAME          ALC_DET      ALC_DETNAME        ALC_STATUS
##               0               0               0               0
##    ALC_STATUSNAME         ATST_TYP     ATST_TYPNAME          ALC_RES
##               0               0               0               0
##      ALC_RESNAME           DRUGS        DRUGSNAME         DRUG_DET
##               0               0               0               0
##     DRUG_DETNAME         DSTATUS      DSTATUSNAME         HOSPITAL
##               0               0               0               0
##     HOSPITALNAME             DOA         DOANAME         DEATH_DA
##               0               0               0               0
##     DEATH_DANAME         DEATH_MO     DEATH_MONAME         DEATH_YR
##               0               0               0               0
##     DEATH_YRNAME         DEATH_HR     DEATH_HRNAME         DEATH_MN
##               0               0               0               0
##     DEATH_MNNAME         DEATH_TM     DEATH_TMNAME          LAG_HRS
##               0               0               0               0
##      LAG_HRSNAME         LAG_MINS     LAG_MINSNAME         WORK_INJ
##               0               0               0               0
##     WORK_INJNAME         HISPANIC     HISPANICNAME         LOCATION
##               0               0               0               0
##     LOCATIONNAME         HELM_USE     HELM_USENAME         HELM_MIS
##               0               0               0               0
##     HELM_MISNAME         VPICMAKE     VPICMAKENAME        VPICMODEL
##               0              NA            8172              NA
##    VPICMODELNAME    VPICBODYCLASS VPICBODYCLASSNAME      ICFINALBODY
##            8172              NA            8172              NA
##  ICFINALBODYNAME
```

```
##             8172
```

```
print('Vehicles')
```

```
## [1] "Vehicles"
```

```
colSums(veh=="")
```

```
##            STATE        STATENAME          ST_CASE           VEH_NO
##                0                0                0                0
##         VE_FORMS          NUMOCCS      NUMOCCSNAME              DAY
##                0                0                0                0
##          DAYNAME            MONTH        MONTHNAME             HOUR
##                0                0                0                0
##         HOURNAME           MINUTE       MINUTENAME          HARM_EV
##                0                0                0                0
##       HARM_EVNAME         MAN_COLL      MAN_COLLNAME         UNITTYPE
##                0                0                0                0
##      UNITTYPENAME          HIT_RUN      HIT_RUNNAME         REG_STAT
##                0                0                0                0
##      REG_STATNAME            OWNER        OWNERNAME             MAKE
##                0                0                0                0
##          MAKENAME            MODEL          MAK_MOD       MAK_MODNAME
##                0                0                0                0
##          BODY_TYP      BODY_TYPNAME         MOD_YEAR      MOD_YEARNAME
##                0                0                0                0
##               VIN          VINNAME            VIN_1            VIN_2
##                8                8                8                8
##             VIN_3            VIN_4            VIN_5            VIN_6
##                8                9                9               12
##             VIN_7            VIN_8            VIN_9           VIN_10
##                16               35               44               68
##            VIN_11           VIN_12          TOW_VEH       TOW_VEHNAME
##                96              131                0                0
##           J_KNIFE       J_KNIFENAME          MCARR_I1      MCARR_I1NAME
##                0                0                0                0
##          MCARR_I2      MCARR_I2NAME          MCARR_ID      MCARR_IDNAME
##                0                0                0                0
##          V_CONFIG      V_CONFIGNAME         CARGO_BT      CARGO_BTNAME
##                0                0                0                0
##           HAZ_INV       HAZ_INVNAME         HAZ_PLAC      HAZ_PLACNAME
##                0                0                0                0
##            HAZ_ID        HAZ_IDNAME          HAZ_CNO       HAZ_CNONAME
##                0                0                0                0
##           HAZ_REL       HAZ_RELNAME          BUS_USE       BUS_USENAME
##                0                0                0                0
##          SPEC_USE      SPEC_USENAME         EMER_USE      EMER_USENAME
##                0                0                0                0
##           TRAV_SP       TRAV_SPNAME         UNDERIDE      UNDERIDENAME
##                0                0                0                0
##          ROLLOVER      ROLLOVERNAME         ROLINLOC      ROLINLOCNAME
##                0                0                0                0
##           IMPACT1       IMPACT1NAME         DEFORMED      DEFORMEDNAME
```

```
##               0               0               0               0
##           TOWED       TOWEDNAME          M_HARM      M_HARMNAME
##               0               0               0               0
##        FIRE_EXP    FIRE_EXPNAME        DR_PRES     DR_PRESNAME
##               0               0               0               0
##         L_STATE     L_STATENAME         DR_ZIP      DR_ZIPNAME
##               0               0               0               0
##        L_STATUS    L_STATUSNAME         L_TYPE      L_TYPENAME
##               0               0               0               0
##        CDL_STAT    CDL_STATNAME       L_ENDORS    L_ENDORSNAME
##               0               0               0               0
##         L_COMPL     L_COMPLNAME       L_RESTRI    L_RESTRINAME
##               0               0               0               0
##          DR_HGT      DR_HGTNAME         DR_WGT      DR_WGTNAME
##               0               0               0               0
##        PREV_ACC    PREV_ACCNAME      PREV_SUS1   PREV_SUS1NAME
##               0               0               0               0
##       PREV_SUS2   PREV_SUS2NAME      PREV_SUS3   PREV_SUS3NAME
##               0               0               0               0
##        PREV_DWI    PREV_DWINAME       PREV_SPD    PREV_SPDNAME
##               0               0               0               0
##        PREV_OTH    PREV_OTHNAME       FIRST_MO    FIRST_MONAME
##               0               0               0               0
##        FIRST_YR    FIRST_YRNAME        LAST_MO     LAST_MONAME
##               0               0               0               0
##         LAST_YR     LAST_YRNAME       SPEEDREL    SPEEDRELNAME
##               0               0               0               0
##        VTRAFWAY    VTRAFWAYNAME       VNUM_LAN    VNUM_LANNAME
##               0               0               0               0
##        VSPD_LIM    VSPD_LIMNAME         VALIGN      VALIGNNAME
##               0               0               0               0
##        VPROFILE    VPROFILENAME       VPAVETYP    VPAVETYPNAME
##               0               0               0               0
##        VSURCOND    VSURCONDNAME       VTRAFCON    VTRAFCONNAME
##               0               0               0               0
##        VTCONT_F    VTCONT_FNAME       P_CRASH1    P_CRASH1NAME
##               0               0               0               0
##        P_CRASH2    P_CRASH2NAME       P_CRASH3    P_CRASH3NAME
##               0               0               0               0
##         PCRASH4     PCRASH4NAME        PCRASH5     PCRASH5NAME
##               0               0               0               0
##        ACC_TYPE    ACC_TYPENAME         DEATHS        DR_DRINK
##               0               0               0               0
##     DR_DRINKNAME        TRLR1VIN    TRLR1VINNAME        TRLR2VIN
##               0               0               0               0
##     TRLR2VINNAME        TRLR3VIN    TRLR3VINNAME         VPICMAKE
##               0               0               0               0
##     VPICMAKENAME        VPICMODEL    VPICMODELNAME    VPICBODYCLASS
##               0               0               0               0
## VPICBODYCLASSNAME      ICFINALBODY  ICFINALBODYNAME        GVWR_FROM
##               0               0               0               0
##     GVWR_FROMNAME         GVWR_TO      GVWR_TONAME        TRLR1GVWR
##               0               0               0               0
##     TRLR1GVWRNAME        TRLR2GVWR    TRLR2GVWRNAME        TRLR3GVWR
```

```
##                 0                 0                 0                 0
##     TRLR3GVWRNAME
##                 0
```

```
print('vision')
```

```
## [1] "vision"
```

```
colSums(visio=="")
```

```
##      STATE  STATENAME    ST_CASE     VEH_NO     VISION VISIONNAME
##          0          0          0          0          0          0
```

Certain variables in the 'per' (person) table contain blank values, such as MAKENAME or MAK_MOD, which refer to the vehicle models and other specifications in the 'veh' (vehicles) table. However, these variables are not relevant for our project's objective, so we will discard those records.

At this point, working with the large number of tables and variables becomes challenging. Therefore, we will need to create the tables on which we will work. By doing so, we will implicitly remove the variables that are not necessary for the project's objective.

In this initial feature engineering phase (second phase), we will select variables based on the descriptive work conducted earlier. In the next phase, we will delve deeper into feature engineering based on the specific requirements of each objective.

TIME SERIES ANALYSIS 2018-2020

We need to create a table that includes the number of victims (FATALS), the year (YEAR), and the accident number or case code (ST_CASE), which serves as the key.

```
#Install various packages as needed
if(!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if(!require('Rmisc')) install.packages('Rmisc'); library('Rmisc')
if(!require('dplyr')) install.packages('dplyr'); library('dplyr')
if(!require('xfun')) install.packages('xfun'); library('xfun')
if(!require('magrittr')) install.packages('magrittr'); library('magrittr')
if (!require('factoextra')) install.packages('factoextra'); library('factoextra')
if(!require('pracma')) install.packages('pracma'); library('pracma')
```

```
print('Creating table for the 2018-2020 FATALS time series')
```

```
## [1] "Creating table for the 2018-2020 FATALS time series"
```

```
# Select the necessary columns
accidentData18_20 <- rbind(
  acc18 %>% select(ST_CASE, YEAR, FATALS),
  acc19 %>% select(ST_CASE, YEAR, FATALS),
  acc20 %>% select(ST_CASE, YEAR, FATALS)
)

# Sort the table by year
analysis18_20 <- accidentData18_20 %>% arrange(YEAR, ST_CASE)

# Show the content of the new table
print('New analysis18_20 table')
```

```
## [1] "New analysis18_20 table"
```

```
# Remove the accidentData18_20 table
rm(accidentData18_20)

# Analyze the headers of the new table
head(analysis18_20)
```

```
##   ST_CASE YEAR FATALS
## 1   10001 2018      1
## 2   10002 2018      2
## 3   10003 2018      1
## 4   10004 2018      1
## 5   10005 2018      1
## 6   10006 2018      1
```

FACTORS PRESENT IN ACCIDENTS (2020)

Previously, we have described the variables to consider or study in order to understand their influence on the outcome. Now, we will create a new table that only includes the variables of interest, discarding others.

```
print('Creating tables with factors related to accidents and excluding the rest')
```

```
## [1] "Creating tables with factors related to accidents and excluding the rest"
```

```
# Selecting the necessary columns from each table
# Unique primary key records
acc20_select <- acc20 %>% select(ST_CASE, DAY_WEEKNAME, HOUR, MINUTE, LGT_COND, LGT_CONDNAME)
wea20_select <- wea %>% select(ST_CASE, WEATHER, WEATHERNAME)

# Merging both tables using the merge function
accidentWeather20 <- merge(acc20_select, wea20_select, by = "ST_CASE")


# Composite primary key records
veh_select <- veh %>% select(ST_CASE, VEH_NO, MOD_YEAR)
per_select <- per %>% select(ST_CASE, VEH_NO, AGE, AIR_BAG, AIR_BAGNAME, DRINKING, DRINKINGNAME, DRUGS,
drivrf_select <- drivrf %>% select(ST_CASE, VEH_NO, DRIVERRF, DRIVERRFNAME)
fact_select <- fact %>% select(ST_CASE, VEH_NO, VEHICLECC, VEHICLECCNAME)
impair_select <- impair %>% select(ST_CASE, VEH_NO, DRIMPAIR, DRIMPAIRNAME)
visio_select <- visio %>% select(ST_CASE, VEH_NO, VISION, VISIONNAME)
distract_select <- distract %>% select(ST_CASE,VEH_NO, DRDISTRACT, DRDISTRACTNAME)

# Combining each table using the merge function
accidentFacts20 <- merge(veh_select, per_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)
accidentFacts20 <- merge(accidentFacts20, drivrf_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)
accidentFacts20 <- merge(accidentFacts20, fact_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)
accidentFacts20 <- merge(accidentFacts20, impair_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)
accidentFacts20 <- merge(accidentFacts20, visio_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)
accidentFacts20 <- merge(accidentFacts20, distract_select, by = c("ST_CASE", "VEH_NO"), all = TRUE)

# Removing datasets that have been merged
rm(acc20_select)
```

```
rm(distract_select)
rm(drivrf_select)
rm(fact_select)
rm(impair_select)
rm(per_select)
rm(veh_select)
rm(visio_select)
rm(wea20_select)

# Analyzing the result using the head() function
print("Accident Facts Table")
```

```
## [1] "Accident Facts Table"
```

```
head(accidentFacts20)
```

```
##   ST_CASE VEH_NO MOD_YEAR AGE AIR_BAG     AIR_BAGNAME DRINKING
## 1   10001      1     1997  22      20    Not Deployed        8
## 2   10001      1     1997  22      20    Not Deployed        8
## 3   10001      1     1997  24       1 Deployed- Front        9
## 4   10001      1     1997  21       1 Deployed- Front        8
## 5   10002      1     1993  51      20    Not Deployed        8
## 6   10002      1     1993  40      20    Not Deployed        9
##          DRINKINGNAME DRUGS           DRUGSNAME DRIVERRF DRIVERRFNAME VEHICLECC
## 1        Not Reported     8        Not Reported        0         None          0
## 2        Not Reported     8        Not Reported        0         None          0
## 3 Reported as Unknown     9 Reported as Unknown        0         None          0
## 4        Not Reported     8        Not Reported        0         None          0
## 5        Not Reported     8        Not Reported        0         None          0
## 6 Reported as Unknown     9 Reported as Unknown        0         None          0
##   VEHICLECCNAME DRIMPAIR                    DRIMPAIRNAME VISION
## 1    None Noted       99 Reported as Unknown if Impaired      0
## 2    None Noted       99 Reported as Unknown if Impaired      0
## 3    None Noted       99 Reported as Unknown if Impaired      0
## 4    None Noted       99 Reported as Unknown if Impaired      0
## 5    None Noted       99 Reported as Unknown if Impaired      0
## 6    None Noted       99 Reported as Unknown if Impaired      0
##             VISIONNAME DRDISTRACT DRDISTRACTNAME
## 1 No Obstruction Noted         96   Not Reported
## 2 No Obstruction Noted         96   Not Reported
## 3 No Obstruction Noted         96   Not Reported
## 4 No Obstruction Noted         96   Not Reported
## 5 No Obstruction Noted         96   Not Reported
## 6 No Obstruction Noted         96   Not Reported
```

```
print("Accident Weather Table")
```

```
## [1] "Accident Weather Table"
```

```
head(accidentWeather20)
```

```
##   ST_CASE DAY_WEEKNAME HOUR MINUTE LGT_COND       LGT_CONDNAME WEATHER
## 1   10001    Wednesday    2     58        2 Dark - Not Lighted       1
## 2   10002     Thursday   17     18        3     Dark - Lighted       2
## 3   10003     Thursday   14     55        1           Daylight       2
## 4   10004       Friday   15     20        1           Daylight      10
## 5   10005     Saturday    0     45        2 Dark - Not Lighted       2
## 6   10006     Saturday   16     55        2 Dark - Not Lighted       1
##   WEATHERNAME
## 1       Clear
## 2        Rain
## 3        Rain
## 4      Cloudy
## 5        Rain
## 6       Clear
```

We have now two tables that collect different variables that may be related to the occurrence of accidents. We decided to unify them into a single table:

```
print("Merging tables with different keys")
```

```
## [1] "Merging tables with different keys"
```

```
accFacts20 <- merge(accidentFacts20, accidentWeather20, by = "ST_CASE", all = TRUE)


# Removing the tables that have been merged
rm(accidentFacts20)
rm(accidentWeather20)

# Viewing the result using the head() function
print("Applying head() and names() to the new table accFacts20")
```

```
## [1] "Applying head() and names() to the new table accFacts20"
```

```
head(accFacts20)
```

```
##   ST_CASE VEH_NO MOD_YEAR AGE AIR_BAG     AIR_BAGNAME DRINKING
## 1   10001      1     1997  22      20    Not Deployed        8
## 2   10001      1     1997  22      20    Not Deployed        8
## 3   10001      1     1997  24       1 Deployed- Front        9
## 4   10001      1     1997  21       1 Deployed- Front        8
## 5   10002      1     1993  51      20    Not Deployed        8
## 6   10002      1     1993  40      20    Not Deployed        9
##          DRINKINGNAME DRUGS           DRUGSNAME DRIVERRF DRIVERRFNAME VEHICLECC
## 1        Not Reported     8        Not Reported        0         None         0
## 2        Not Reported     8        Not Reported        0         None         0
## 3 Reported as Unknown     9 Reported as Unknown        0         None         0
## 4        Not Reported     8        Not Reported        0         None         0
## 5        Not Reported     8        Not Reported        0         None         0
## 6 Reported as Unknown     9 Reported as Unknown        0         None         0
##   VEHICLECCNAME DRIMPAIR                    DRIMPAIRNAME VISION
## 1    None Noted       99 Reported as Unknown if Impaired      0
```

```
## 2    None Noted       99 Reported as Unknown if Impaired      0
## 3    None Noted       99 Reported as Unknown if Impaired      0
## 4    None Noted       99 Reported as Unknown if Impaired      0
## 5    None Noted       99 Reported as Unknown if Impaired      0
## 6    None Noted       99 Reported as Unknown if Impaired      0
##             VISIONNAME DRDISTRACT DRDISTRACTNAME DAY_WEEKNAME HOUR MINUTE
## 1 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 2 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 3 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 4 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 5 No Obstruction Noted         96   Not Reported     Thursday   17     18
## 6 No Obstruction Noted         96   Not Reported     Thursday   17     18
##   LGT_COND       LGT_CONDNAME WEATHER WEATHERNAME
## 1        2 Dark - Not Lighted       1       Clear
## 2        2 Dark - Not Lighted       1       Clear
## 3        2 Dark - Not Lighted       1       Clear
## 4        2 Dark - Not Lighted       1       Clear
## 5        3     Dark - Lighted       2        Rain
## 6        3     Dark - Lighted       2        Rain
```

```
names(accFacts20)
```

```
##  [1] "ST_CASE"        "VEH_NO"         "MOD_YEAR"       "AGE"
##  [5] "AIR_BAG"        "AIR_BAGNAME"    "DRINKING"       "DRINKINGNAME"
##  [9] "DRUGS"          "DRUGSNAME"      "DRIVERRF"       "DRIVERRFNAME"
## [13] "VEHICLECC"      "VEHICLECCNAME"  "DRIMPAIR"       "DRIMPAIRNAME"
## [17] "VISION"         "VISIONNAME"     "DRDISTRACT"     "DRDISTRACTNAME"
## [21] "DAY_WEEKNAME"   "HOUR"           "MINUTE"         "LGT_COND"
## [25] "LGT_CONDNAME"   "WEATHER"        "WEATHERNAME"
```

ROAD BLACK SPOTS

In line with the two previous tables that collect the different aspects we intend to analyze, we proceed to create a third table with geographic data to later determine where the road black spots are located. We start with the base of the 2020 accident table, which contains all the information we need:

```
print("Creating the table with the location of accidents")
```

```
## [1] "Creating the table with the location of accidents"
```

```
accBpoint20 <- acc20 %>% select(ST_CASE, STATE, STATENAME, COUNTY, COUNTYNAME, CITY, CITYNAME, ROUTE, R


# Analyzing the content with head() and the variables with names()
print("Table head")
```

```
## [1] "Table head"
```

```
head(accBpoint20)
```

```
##   ST_CASE STATE STATENAME COUNTY     COUNTYNAME CITY       CITYNAME ROUTE
```

```
## 1   10001     1   Alabama     51    ELMORE (51)    0 NOT APPLICABLE     4
## 2   10002     1   Alabama     73 JEFFERSON (73)  350     BIRMINGHAM     6
## 3   10003     1   Alabama    117   SHELBY (117)    0 NOT APPLICABLE     3
## 4   10004     1   Alabama     15   CALHOUN (15)    0 NOT APPLICABLE     4
## 5   10005     1   Alabama     37     COOSA (37)    0 NOT APPLICABLE     4
## 6   10006     1   Alabama    103   MORGAN (103)    0 NOT APPLICABLE     3
##                      ROUTENAME RUR_URB RUR_URBNAME MILEPT LATITUDE  LONGITUD
## 1                  County Road       1       Rural      0 32.43313 -86.09485
## 2 Local Street - Municipality       2       Urban      0 33.48466 -86.83954
## 3                State Highway       1       Rural     49 33.29994 -86.36964
## 4                  County Road       1       Rural      0 33.79507 -85.88349
## 5                  County Road       1       Rural      0 32.84841 -86.08355
## 6                State Highway       1       Rural    390 34.50894 -86.67486
```

```
print("Variable Names")
```

```
## [1] "Variable Names"
```

```
names(accBpoint20)
```

```
##  [1] "ST_CASE"     "STATE"       "STATENAME"   "COUNTY"      "COUNTYNAME"
##  [6] "CITY"        "CITYNAME"    "ROUTE"       "ROUTENAME"   "RUR_URB"
## [11] "RUR_URBNAME" "MILEPT"      "LATITUDE"    "LONGITUD"
```

At this point, we have achieved the objectives for this phase, as we have: 1. Description of the dataset and the variables represented in it. 2. Clean dataset with an initial phase of feature management.

Finally, we proceed to clean the workspace before moving on to the third phase.

```
# Removing unnecessary elements except for the 3 tables we will work with
rm(acc18, acc19, acc20, distract, drivrf, fact, impair, per, veh, visio, wea)

# Removing 'paths' and 'summaries'
rm(list = ls(pattern = "^pat"))
rm(list = ls(pattern = "^summa"))
```

Phase 3. Data Preparation

Objectives:

In this phase, we will continue with feature engineering, which involves selecting and transforming variables or features of the data to improve the performance of the machine learning model. In this case, we will adapt the different tables and variables to the project's needs.

Deliverable:

Obtain a dataset from 3 tables that contain the relevant variables for the 3 aspects of the project:

1. Analyze the evolution of accidents in the 2018 to 2020 series.
2. Identify blackspots in the road network.
3. Explore factors that may influence the occurrence of accidents.

These tables will have undergone data transformation and dimensionality reduction methods.

Tasks:

1. Feature processing and management II TABLE FOR THE 2018-2020 TIME SERIES

The table for the analysis of the time series, 'analisis18_20', contains only 3 variables as we saw earlier.

```
# Table analysis
head(analysis18_20)
```

```
##   ST_CASE YEAR FATALS
## 1   10001 2018      1
## 2   10002 2018      2
## 3   10003 2018      1
## 4   10004 2018      1
## 5   10005 2018      1
## 6   10006 2018      1
```

```
names(analysis18_20)
```

```
## [1] "ST_CASE" "YEAR"    "FATALS"
```

Feature processing (or engineering) is typically applied when there is a large number of variables and the goal is to reduce their dimensionality by eliminating irrelevant or highly correlated variables. However, in this case, there are only four variables in the table and all of them appear to be relevant for the analysis. Therefore, we will not perform any feature engineering on this table.

TABLE OF FACTORS IN ACCIDENTS

The resulting table from Phase 2, 'accFacts20', contains, as we will see below, 96966 records and 27 variables.

```
#Table factors
head(accFacts20)
```

```
##   ST_CASE VEH_NO MOD_YEAR AGE AIR_BAG     AIR_BAGNAME DRINKING
## 1   10001      1     1997  22      20    Not Deployed        8
## 2   10001      1     1997  22      20    Not Deployed        8
## 3   10001      1     1997  24       1 Deployed- Front        9
## 4   10001      1     1997  21       1 Deployed- Front        8
## 5   10002      1     1993  51      20    Not Deployed        8
## 6   10002      1     1993  40      20    Not Deployed        9
##          DRINKINGNAME DRUGS           DRUGSNAME DRIVERRF DRIVERRFNAME VEHICLECC
## 1        Not Reported     8        Not Reported        0         None         0
## 2        Not Reported     8        Not Reported        0         None         0
## 3 Reported as Unknown     9 Reported as Unknown        0         None         0
## 4        Not Reported     8        Not Reported        0         None         0
## 5        Not Reported     8        Not Reported        0         None         0
## 6 Reported as Unknown     9 Reported as Unknown        0         None         0
##   VEHICLECCNAME DRIMPAIR                    DRIMPAIRNAME VISION
## 1    None Noted       99 Reported as Unknown if Impaired      0
## 2    None Noted       99 Reported as Unknown if Impaired      0
## 3    None Noted       99 Reported as Unknown if Impaired      0
## 4    None Noted       99 Reported as Unknown if Impaired      0
## 5    None Noted       99 Reported as Unknown if Impaired      0
```

```
## 6    None Noted       99 Reported as Unknown if Impaired      0
##             VISIONNAME DRDISTRACT DRDISTRACTNAME DAY_WEEKNAME HOUR MINUTE
## 1 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 2 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 3 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 4 No Obstruction Noted         96   Not Reported    Wednesday    2     58
## 5 No Obstruction Noted         96   Not Reported     Thursday   17     18
## 6 No Obstruction Noted         96   Not Reported     Thursday   17     18
##   LGT_COND       LGT_CONDNAME WEATHER WEATHERNAME
## 1        2 Dark - Not Lighted       1       Clear
## 2        2 Dark - Not Lighted       1       Clear
## 3        2 Dark - Not Lighted       1       Clear
## 4        2 Dark - Not Lighted       1       Clear
## 5        3     Dark - Lighted       2        Rain
## 6        3     Dark - Lighted       2        Rain
```

```
names(accFacts20)
```

```
##  [1] "ST_CASE"        "VEH_NO"         "MOD_YEAR"       "AGE"
##  [5] "AIR_BAG"        "AIR_BAGNAME"    "DRINKING"       "DRINKINGNAME"
##  [9] "DRUGS"          "DRUGSNAME"      "DRIVERRF"       "DRIVERRFNAME"
## [13] "VEHICLECC"      "VEHICLECCNAME"  "DRIMPAIR"       "DRIMPAIRNAME"
## [17] "VISION"         "VISIONNAME"     "DRDISTRACT"     "DRDISTRACTNAME"
## [21] "DAY_WEEKNAME"   "HOUR"           "MINUTE"         "LGT_COND"
## [25] "LGT_CONDNAME"   "WEATHER"        "WEATHERNAME"
```

The large number of variables makes it challenging to work with the data. At this point, we will perform a Principal Component Analysis (PCA). The goal is to find linear combinations of the original variables that explain the most variability in the data. By doing so, we can reduce the number of variables to those that truly contribute relevant information:

```
# To perform a PCA we need to work only with the numerical variables.
print("Filtering variables to exclude numerical variables")
```

```
## [1] "Filtering variables to exclude numerical variables"
```

```
accFacts20_num <- accFacts20[, !grepl("NAME", names(accFacts20))]


# Looking for missing values or NA
sum(is.na(accFacts20_num))
```

```
## [1] 52762
```

We observe that there are indeed missing values in the data. Additionally, upon visual inspection and as a result of merging different tables, we can see values that are completely out of range (9999, 99, 0.99), which require further handling.

```
library("pracma")
print("Revisamos cada variable para conocer como se distribuyen los valores ausentes")
```

```
## [1] "Revisamos cada variable para conocer como se distribuyen los valores ausentes"
```

```r
# establecemos un limite razonable de impresiones para no saturar el documento
options(max.print=20)


print("MOD_YEAR")
```

```
## [1] "MOD_YEAR"
```

```r
which(is.na(accFacts20_num$MOD_YEAR))
```

```
##  [1]  22  38  41  48  78  88  95 101 109 117 171 173 188 234 266 268 292 301 311
## [20] 320
##  [ reached getOption("max.print") -- omitted 8579 entries ]
```

```r
print("AGE")
```

```
## [1] "AGE"
```

```r
which(is.na(accFacts20_num$AGE))
```

```
##  [1]  294  360  434  603 1202 1826 2213 2751 3093 3589 4102 4428 4429 4518 5564
## [16] 5948 6788 6789 7008 7022
##  [ reached getOption("max.print") -- omitted 272 entries ]
```

```r
print("DRINKING")
```

```
## [1] "DRINKING"
```

```r
which(is.na(accFacts20_num$DRINKING))
```

```
##  [1]  294  360  434  603 1202 1826 2213 2751 3093 3589 4102 4428 4429 4518 5564
## [16] 5948 6788 6789 7008 7022
##  [ reached getOption("max.print") -- omitted 272 entries ]
```

```r
print("DRUGS")
```

```
## [1] "DRUGS"
```

```r
which(is.na(accFacts20_num$DRUGS))
```

```
##  [1]  294  360  434  603 1202 1826 2213 2751 3093 3589 4102 4428 4429 4518 5564
## [16] 5948 6788 6789 7008 7022
##  [ reached getOption("max.print") -- omitted 272 entries ]
```

```r
print("DRIVERRF")
```

```
## [1] "DRIVERRF"
```

```r
which(is.na(accFacts20_num$DRIVERRF))
```

```
##  [1]  22  38  41  48  78  88  95 101 109 117 171 173 188 234 266 268 292 301 311
## [20] 320
##  [ reached getOption("max.print") -- omitted 8579 entries ]
```

```r
print("HOUR")
```

```
## [1] "HOUR"
```

```r
which(is.na(accFacts20_num$MINUTE))
```

```
## integer(0)
```

```r
print("MINUTE")
```

```
## [1] "MINUTE"
```

```r
which(is.na(accFacts20_num$MINUTE))
```

```
## integer(0)
```

```r
print("WEATHER")
```

```
## [1] "WEATHER"
```

```r
which(is.na(accFacts20_num$WEATHER))
```

```
## integer(0)
```

In general, when using the 'which' function without limits, we observe that there are approximately 8579 NA values. In the case of the vehicle's age, it is clear that not all records capture this information, as there are accidents involving non-motorized vehicles. Therefore, we know that accidents with 'VEH_NO' = 0 correspond to accidents involving non-motorized vehicles. Additionally, we know that among the vehicles involved, 'VEH_NO' represents the vehicle considered to have caused the accident. Hence, during the statistical analysis in the upcoming phases, we need to take this into account.

We also observe that there are a large number of discrete variables, especially related to technical aspects of the vehicle at the time of the accident, driver-related details, substance consumption, etc., which need to be transformed into continuous variables.

Another aspect to consider is the presence of attributes that have taken on values like "type 9999" due to the merging of tables. We will address each case individually.

```r
print("Filtramos valores = 0 en VEH_NO que se referen a otros involucrados o vehiculos sin motor ")
```

```
## [1] "Filtramos valores = 0 en VEH_NO que se referen a otros involucrados o vehiculos sin motor "
```

```r
accFacts20_num_filtrado <- accFacts20_num[accFacts20_num$VEH_NO != 0, ]

print("imputamos a los valores 'MOD_YEAR'= NA el valor del vehiculo principal en el mismo caso")
```

```
## [1] "imputamos a los valores 'MOD_YEAR'= NA el valor del vehiculo principal en el mismo caso"
```

```r
for (i in unique(accFacts20_num_filtrado$ST_CASE)) {
  vehiculo_principal_mod_year <- na.omit(accFacts20_num_filtrado$MOD_YEAR[accFacts20_num_filtrado$ST_CA
  if (length(vehiculo_principal_mod_year) > 0) {
    accFacts20_num_filtrado$MOD_YEAR[accFacts20_num_filtrado$ST_CASE == i & is.na(accFacts20_num_filtra
  }
}

print("Comprobamos los valores NA en 'MOD_YEAR")
```

```
## [1] "Comprobamos los valores NA en 'MOD_YEAR"
```

```r
print("MOD_YEAR")
```

```
## [1] "MOD_YEAR"
```

```r
which(is.na(accFacts20_num$MOD_YEAR))
```

```
##  [1]  22  38  41  48  78  88  95 101 109 117 171 173 188 234 266 268 292 301 311
## [20] 320
##  [ reached getOption("max.print") -- omitted 8579 entries ]
```

We observed different problems, such as "9999" values in the dates of the models and others in the variable AGE.

```r
print("Eliminamos valores imposibles en MOD_YEAR y AGE")
```

```
## [1] "Eliminamos valores imposibles en MOD_YEAR y AGE"
```

```r
accFacts20_num_filtrado <- subset(accFacts20_num_filtrado, !(MOD_YEAR > 2020 & AGE > 90))
print("eliminamos edades imposibles o casi, para conducir")
```

```
## [1] "eliminamos edades imposibles o casi, para conducir"
```

```r
accFacts20_num_filtrado <- subset(accFacts20_num_filtrado, !(AGE > 90))
```

We realise that for the purposes of the analysis it is of little relevance to analyse those involved other than the main vehicle, so we rectify to correct and obtain a table containing the causal cars and drivers, including all the factors we want to study in the following phases:

```r
print("Estableciendo el numero de caso = VEH_NO=1")
```

```
## [1] "Estableciendo el numero de caso = VEH_NO=1"
```

```r
accFacts20_num_filtrado_veh1 <- accFacts20_num_filtrado %>%
  filter(VEH_NO == 1) %>%
  distinct(ST_CASE, .keep_all = TRUE)

print("cambiamos el nombre a la tabla y limpiamos el entorno de trabajo")
```

```
## [1] "cambiamos el nombre a la tabla y limpiamos el entorno de trabajo"
```

```r
accFacts20v2 <- accFacts20_num_filtrado_veh1
rm(accFacts20_num, accFacts20_num_filtrado, accFacts20_num_filtrado_veh1)

print("Eliminamos la variable AIR_BAG porque cometimos un error seleccionandola y no tiene utilidad para
```

```
## [1] "Eliminamos la variable AIR_BAG porque cometimos un error seleccionandola y no tiene utilidad pa
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-AIR_BAG)
```

```r
#Tenemos que convertir en binaria la variable "Drinking"
print("Creamos una nueva variable binaria 0/1 (1= para conductor bebido (probado) y 0= no bebido o no p
```

```
## [1] "Creamos una nueva variable binaria 0/1 (1= para conductor bebido (probado) y 0= no bebido o no
```

```r
# Codificar la variable DRINKING como binaria (0/1)
accFacts20v2$DRINKING <- ifelse(accFacts20v2$DRINKING == 1, 1, 0)
accFacts20v2$DRINKING <- ifelse(accFacts20v2$DRINKING %in% c(8, 9), 0, accFacts20v2$DRINKING)

print('En la nueva variable, hemos respetado la presuncion de inocencia, es decir, los valores 8 y 9 que
```

```
## [1] "En la nueva variable, hemos respetado la presuncion de inocencia, es decir, los valores 8 y 9 q
```

We have coded the variable 'Drinking' as binary: YES or No alcohol 1/0. We repeat the process with the variable 'Drugs':

```r
#Tenemos que convertir en binaria la variable "Drugs"
print("Creamos una nueva variable binaria 0/1 (1= para conductor drogado (probado) y 0= no drogado o no
```

```
## [1] "Creamos una nueva variable binaria 0/1 (1= para conductor drogado (probado) y 0= no drogado o n
```

```r
# Codificar la variable DRINKING como binaria (0/1)
accFacts20v2$DRUGS <- ifelse(accFacts20v2$DRUGS == 1, 1, 0)
accFacts20v2$DRUGS <- ifelse(accFacts20v2$DRUGS %in% c(8, 9), 0, accFacts20v2$DRUGS)

print('En la nueva variable, hemos respetado la presuncion de inocencia, es decir, los valores 8 y 9 que
```

```
## [1] "En la nueva variable, hemos respetado la presuncion de inocencia, es decir, los valores 8 y 9 q
```

On the other hand, we initially selected the variable 'DRIVERRF' because it contains valuable information. It is important to exclude accidents involving fire or police vehicles, as well as accidents where vehicles are being towed. These correspond to the codes 97, 96, 95, 94, 86, and 16. Afterward, we will remove the 'DRIVERRF' variable as it is no longer needed for our analysis.

```
print("creamos un vector con los valores que pretendemos excluir de la variable DRIVERRF")
```

```
## [1] "creamos un vector con los valores que pretendemos excluir de la variable DRIVERRF"
```

```
valores_no_deseados <- c(97, 96, 95, 94, 86, 16)
accFacts20v2_filtrado <- accFacts20v2[!accFacts20v2$DRIVERRF %in% valores_no_deseados, ]


print("Eliminamos la variable DRIVERRF")
```

```
## [1] "Eliminamos la variable DRIVERRF"
```

```
accFacts20v2 <- accFacts20v2_filtrado %>%
  select(-DRIVERRF)
rm(accFacts20v2_filtrado)
```

The variable VEHICLECC refers to mechanical problems in the vehicle and specifies the type, we convert it to binary, 0= no problems and 1= mechanical problems.

```
#Tenemos que convertir en binaria la variable "VEHICLECC"
print("Creamos una nueva variable binaria 0/1 (1= para problemas mecanicos y 0= no problemas mecanicos)
```

```
## [1] "Creamos una nueva variable binaria 0/1 (1= para problemas mecanicos y 0= no problemas mecanicos
```

```
# Codificar la variable DRINKING como binaria (0/1)
accFacts20v2$VEHICLECC <- ifelse(accFacts20v2$VEHICLECC == 1, 1, 0)
accFacts20v2$VEHICLECC <- ifelse(accFacts20v2$VEHICLECC %in% c(2,3,4,5,6,7,8,9,10,12,13,14,15,16,17,97)
accFacts20v2$VEHICLECC <- ifelse(accFacts20v2$VEHICLECC %in% c(98, 99), 0, accFacts20v2$VEHICLECC)


print('En la nueva variable o recodificacion, hemos respetado la "presuncion de inocencia mecanica", es
```

```
## [1] "En la nueva variable o recodificacion, hemos respetado la \"presuncion de inocencia mecanica\",
```

The variable 'DRIMPAIR' is very interesting as it contains codes to describe different situations where the drivers' psychophysical abilities are affected. These situations may include blindness, deafness, physical injuries, and more. We will exclude the effects of alcohol and drugs by setting their values to 0, as we have encoded them in other specific variables. The 'DRIMPAIR' variable will encompass all the psychological and physical aspects that may have influenced the accident.

```
#Tenemos que convertir en binaria la variable "DRIMPAIR"
print("Creamos una nueva variable binaria 0/1 (1= para problemas fisicos y psicologicos (no por consumo
```

```
## [1] "Creamos una nueva variable binaria 0/1 (1= para problemas fisicos y psicologicos (no por consume
```

```
# Codificar la variable DRIMPAIR como binaria (0/1)
accFacts20v2$DRIMPAIR <- ifelse(accFacts20v2$DRIMPAIR == 1, 1, 0)
accFacts20v2$DRIMPAIR <- ifelse(accFacts20v2$DRIMPAIR %in% c(4,5,6,7,8,10,96), 1, accFacts20v2$DRIMPAIR
accFacts20v2$VEHICLECC <- ifelse(accFacts20v2$VEHICLECC %in% c(98, 99, 9, 95), 0, accFacts20v2$VEHICLECC


print('En la nueva variable o recodificacion, los valores codificados como 99 "desconocido" los hemos c
```

```
## [1] "En la nueva variable o recodificacion, los valores codificados como 99 \"desconocido\" los hemos
```

```
print("La nueva variable es por tanto representantiva de aquellos problemas fisicos o psicologicos no di
```

```
## [1] "La nueva variable es por tanto representantiva de aquellos problemas fisicos o psicologicos no d
```

```
print("una simple revision visual nos permite ver que aun tenemos algunos valores NA o infinitos, por lo
```

```
## [1] "una simple revision visual nos permite ver que aun tenemos algunos valores NA o infinitos, por l
```

```
print("Eliminamos valores NA comunes a VEHICLECC y DRIMPAIR")
```

```
## [1] "Eliminamos valores NA comunes a VEHICLECC y DRIMPAIR"
```

```
accFacts20v2 <- subset(accFacts20v2, !is.na(MOD_YEAR) & !is.na(AGE))
```

We will now proceed with the variable 'VISION', which refers to difficulties in visibility such as fog, smoke, or defective reflective elements, among others. We will convert this variable into a binary variable, indicating whether or not visibility problems were present at the time of the accident. Visibility problems may include issues related to the design of the road (structural).

```
#Tenemos que convertir en binaria la variable "VISION"
print("Creamos una nueva variable binaria 0/1 (1= para problemas de visibildad y 0= no presentes)")
```

```
## [1] "Creamos una nueva variable binaria 0/1 (1= para problemas de visibildad y 0= no presentes)"
```

```
# Codificar la variable DRIMPAIR como binaria (0/1)
accFacts20v2$VISION <- ifelse(accFacts20v2$VISION == 1, 1, 0)
accFacts20v2$VISION <- ifelse(accFacts20v2$VISION %in% c(2,3,4,5,6,7,8,9,10,11,12,13,14,97,98), 1, accFa
accFacts20v2$VISION <- ifelse(accFacts20v2$VISION %in% c(95, 99), 0, accFacts20v2$VISION)


print('En la nueva variable o recodificacion, los valores codificados como 99 "desconocido" los hemos co
```

```
## [1] "En la nueva variable o recodificacion, los valores codificados como 99 \"desconocido\" los hemos
```

```
print("La nueva variable es por tanto representantiva de aquellos problemas fisicos o psicologicos no di
```

```
## [1] "La nueva variable es por tanto representantiva de aquellos problemas fisicos o psicologicos no d
```

We need to express the years of manufacture of the vehicles in years of age. So let's create a new variable OLD and delete MOD_YEAR

```
print("vamos a crear una variable nueva desde el punto de partida de MOD_YEAR")
```

```
## [1] "vamos a crear una variable nueva desde el punto de partida de MOD_YEAR"
```

```r
# Obtener el año actual
anyo_actual <- as.numeric(format(Sys.Date(), "%Y"))


# Calcular la antigüedad en años
accFacts20v2$ANTIGUEDAD <- anyo_actual - accFacts20v2$MOD_YEAR


# Convertir la antigüedad a valor absoluto como precaución, aunque no espero valores negativos
accFacts20v2$ANTIGUEDAD <- abs(accFacts20v2$ANTIGUEDAD)


print("Una vez creada la variable 'ANTIGUEDAD', nos deshacemos de la variable 'MOD_YEAR'")
```

```
## [1] "Una vez creada la variable 'ANTIGUEDAD', nos deshacemos de la variable 'MOD_YEAR'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-MOD_YEAR)
```

We consider making the new variable binary:

```r
print("consideraremos un coche antiguo cuando tenga mas de 10 años, por lo que convertiremos en binaria
```

```
## [1] "consideraremos un coche antiguo cuando tenga mas de 10 años, por lo que convertiremos en binaria
```

```r
# Creamos una nueva variable binaria que indique si el vehículo tiene 10 años o más
accFacts20v2$OLD <- ifelse(accFacts20v2$ANTIGUEDAD >= 10, 1, 0)
```

```r
print("Una vez creada la variable 'OLD', nos deshacemos de la variable 'ANTIGUEDAD'")
```

```
## [1] "Una vez creada la variable 'OLD', nos deshacemos de la variable 'ANTIGUEDAD'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-ANTIGUEDAD)
```

Another factor to consider is distractions. In a world where technology makes us increasingly hyperconnected, these devices have become a double-edged sword. Additionally, apart from technology-related distractions, other "classic" distractions have always been present. We want to evaluate the impact of distractions on accidents, so we will convert the variable 'DRDISTRAC' into another binary variable, considering 0 for no distractions or unspecified distractions, and 1 for reported distractions.

```r
print("Modificamos la variable a binaria 0/1 (1= para problemas de distracciones al volante y 0= no pres
```

```
## [1] "Modificamos la variable a binaria 0/1 (1= para problemas de distracciones al volante y 0= no pre
```

```r
# Codificamos la variable DRDISTRAC como binaria (0/1)
accFacts20v2$DRDISTRACT <- ifelse(accFacts20v2$DRDISTRACT == 99, 0, 1)
```

The variable 'HOUR' is relevant as it signifies a possible cause that can influence the psychophysical conditions of those involved in accidents. We will consider the hours between 6:00 PM (18:00) and 6:00 AM (6:00) as nighttime. We will create a new variable called 'NIGHT_HOUR' to represent this.

```r
# Codificar la variable HOUR como binaria (0/1)
print("creamos la nueva variable binaria NIGHT_HOUR")
```

```
## [1] "creamos la nueva variable binaria NIGHT_HOUR"
```

```r
accFacts20v2$NIGHT_HOUR <- ifelse(accFacts20v2$HOUR >= 18 | accFacts20v2$HOUR < 6, 1, 0)

#Eliminamos la variable 'HOUR' y 'MINUTE'
accFacts20v2 <- accFacts20v2 %>%
  select(-HOUR)
accFacts20v2 <- accFacts20v2 %>%
  select(-MINUTE)
```

The variable 'LGT_COND' refers to the road conditions, but we will discard it because we can study the nighttime factor using the 'NIGHT_HOUR' variable. Additionally, we will address the 'AGE' column, which pertains to the age of the primary individuals involved, in intervals.

```r
#Eliminamos la variable 'LGT_COND'
print("eliminamos la variable 'LGT_COND'")
```

```
## [1] "eliminamos la variable 'LGT_COND'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-LGT_COND)

print("Dividimos en intervalos la variable 'AGE'")
```

```
## [1] "Dividimos en intervalos la variable 'AGE'"
```

```r
# Dividir la columna 'AGE' en intervalos usando la función cut()
accFacts20v2 <- accFacts20v2 %>%
  mutate(AGE_GRUP = cut(AGE, breaks = c(-Inf, 16, 44, 72, 100, Inf), labels = c("<=16", "17-44", "45-72
```

As a precaution before removing the 'AGE' variable, we will create another binary variable called 'AGE_BIN' to categorize individuals as young (0) or senior (1). We will consider individuals with an age less than or equal to 25 as young, and those with an age greater than 25 as senior:

```r
# Crear la variable binaria 'AGE_BIN'
print("creamos la variable 'AGE_BIN'")
```

```
## [1] "creamos la variable 'AGE_BIN'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  mutate(AGE_BIN = ifelse(AGE <= 25, 0, 1))
```

Finally, we will address the 'WEATHER' variable, which we will also convert into a binary variable called 'WEA_BIN'. We will assign a value of 0 for clear weather conditions and a value of 1 for inclement weather conditions.

```r
print("Creamos la variable convirtiendo los registros segun se naturaleza 0= despejado y 1= No despejad
```

```
## [1] "Creamos la variable convirtiendo los registros segun se naturaleza 0= despejado y 1= No despeja
```

```r
# Crear la variable binaria 'WEA_BIN'
accFacts20v2 <- accFacts20v2 %>%
  mutate(WEA_BIN = ifelse(WEATHER %in% c(1, 98, 99), 0, 1))
```

We eliminate the variables no longer needed:

```r
#Eliminamos la variable 'AGE'
print("eliminamos la variable 'AGE'")
```

```
## [1] "eliminamos la variable 'AGE'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-AGE)

print("eliminamos la variable 'VEH_NO'")
```

```
## [1] "eliminamos la variable 'VEH_NO'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-VEH_NO)

print("eliminamos la variable 'WEATHER'")
```

```
## [1] "eliminamos la variable 'WEATHER'"
```

```r
accFacts20v2 <- accFacts20v2 %>%
  select(-WEATHER)

#copia de seguridad
accFacts20vBACK <- accFacts20v2

#retomamos la tabla principal
accFacts20 <- accFacts20v2
```

We continue with the PCA once all the variables of interest have been reviewed, cleaned and transformed:

```r
# Cargar librería para PCA
library(stats)
options(max.print = 1000)

# Creamos una nueva tabla excluyendo las variables 'ST_CASE' y 'AGE_GRUP'
accFacts_pca <- accFacts20[, !(colnames(accFacts20) %in% c("ST_CASE", "AGE_GRUP"))]

# Realizar el PCA
pca_result <- prcomp(accFacts_pca, scale. = TRUE)
```

```r
# Obtener los resultados del PCA
pca_variances <- pca_result$sdev^2
pca_proportions <- pca_variances / sum(pca_variances)
pca_loadings <- pca_result$rotation

# Imprimir los resultados
cat("Varianzas explicadas por cada componente principal:\n")
```

```
## Varianzas explicadas por cada componente principal:
```

```r
cat(pca_variances, "\n")
```

```
## 1.394765 1.169018 1.110145 1.046596 0.9920903 0.959853 0.9282325 0.8881005 0.8453004 0.6658995
```

```r
cat("\nProporciones de varianzas explicadas por cada componente principal:\n")
```

```
##
## Proporciones de varianzas explicadas por cada componente principal:
```

```r
cat(pca_proportions, "\n")
```

```
## 0.1394765 0.1169018 0.1110145 0.1046596 0.09920903 0.0959853 0.09282325 0.08881005 0.08453004 0.06658
```

```r
cat("\nLoadings de cada variable en cada componente principal:\n")
```

```
##
## Loadings de cada variable en cada componente principal:
```

```r
print(pca_loadings)
```

```
##                    PC1         PC2          PC3          PC4         PC5
## DRINKING    0.64869055 -0.06991905  0.174599508 -0.104334481  0.13061016
## DRUGS       0.53269818 -0.12116224  0.369702935  0.068065304  0.08119061
## VEHICLECC  -0.03233221 -0.34893545  0.030547657  0.498779313  0.29830129
## DRIMPAIR   -0.21078773  0.11194831  0.443602143 -0.001755419  0.21964628
## VISION     -0.09133966 -0.61394664 -0.037592423 -0.331131657  0.04321455
## DRDISTRACT -0.17286349  0.16135948 -0.028547002 -0.118186887  0.88972091
## OLD         0.10731112 -0.15967538 -0.008917767  0.671185771  0.03470705
## NIGHT_HOUR  0.42390063  0.07765486 -0.420723777 -0.272406301  0.19316963
## AGE_BIN    -0.06259350  0.06419617  0.674845331 -0.210064505 -0.07627966
## WEA_BIN    -0.12886539 -0.63837893  0.019656164 -0.207987028  0.03622782
##                    PC6         PC7         PC8         PC9         PC10
## DRINKING   -0.009238251 -0.05248954 -0.03584235  0.06018361 -0.712835867
## DRUGS       0.062024653  0.06202043  0.45369036 -0.20276839  0.546812347
## VEHICLECC   0.550323938 -0.41256682 -0.25143611  0.03383604  0.032684080
## DRIMPAIR   -0.540384498 -0.63552022  0.04426017  0.01201013 -0.001055829
## VISION     -0.164012871  0.02281430 -0.23241704 -0.64809264 -0.018309995
## DRDISTRACT  0.032597014  0.35223515  0.09976727 -0.03778457 -0.034374611
## OLD        -0.546210635  0.38813255 -0.24648399  0.03595308  0.013180842
## NIGHT_HOUR -0.181612000 -0.22718899 -0.46393220  0.23583738  0.412673801
## AGE_BIN     0.173672804  0.29336529 -0.56577741  0.17647194  0.138300916
## WEA_BIN    -0.108359347  0.07991234  0.26007421  0.66668419  0.025971277
```

In the principal component analysis conducted in our study, it was found that the first three principal components explain a significant portion of the total variance, with the first principal component explaining approximately 14% of the variance, the second principal component explaining around 12%, and the third principal component explaining around 11%. This suggests that these three principal components capture the majority of the variability in the original data. Furthermore, loading patterns of the variables on each principal component were identified, indicating the direction and magnitude of the influence of each variable on the principal components. For example, it was found that the variable 'DRINKING' has a strong positive influence on the first principal component, while the variable 'DRUGS' has a strong negative influence on the second principal component. These results help us understand how the original variables contribute to the structure of the principal components and how they relate to each other in our analysis Jolliffe (2002). We have decided to keep all the variables.

Finally, the table 'accBpoint20', we look for NA or infinite values:

```
print("Comprobando NA")
```

```
## [1] "Comprobando NA"
```

```
sum(is.na(accBpoint20))
```

```
## [1] 0
```

```
print("Comprobando encabezados")
```

```
## [1] "Comprobando encabezados"
```

```
head(accBpoint20)
```

```
##   ST_CASE STATE STATENAME COUNTY     COUNTYNAME CITY       CITYNAME ROUTE
## 1   10001     1   Alabama     51    ELMORE (51)    0 NOT APPLICABLE     4
## 2   10002     1   Alabama     73 JEFFERSON (73)  350     BIRMINGHAM     6
## 3   10003     1   Alabama    117   SHELBY (117)    0 NOT APPLICABLE     3
## 4   10004     1   Alabama     15   CALHOUN (15)    0 NOT APPLICABLE     4
## 5   10005     1   Alabama     37     COOSA (37)    0 NOT APPLICABLE     4
## 6   10006     1   Alabama    103   MORGAN (103)    0 NOT APPLICABLE     3
##                     ROUTENAME RUR_URB RUR_URBNAME MILEPT LATITUDE  LONGITUD
## 1                  County Road       1       Rural      0 32.43313 -86.09485
## 2 Local Street - Municipality       2       Urban      0 33.48466 -86.83954
## 3               State Highway       1       Rural     49 33.29994 -86.36964
## 4                  County Road       1       Rural      0 33.79507 -85.88349
## 5                  County Road       1       Rural      0 32.84841 -86.08355
## 6               State Highway       1       Rural    390 34.50894 -86.67486
```

```
print("nombres de variables")
```

```
## [1] "nombres de variables"
```

```
names(accBpoint20)
```

```
##  [1] "ST_CASE"     "STATE"       "STATENAME"   "COUNTY"      "COUNTYNAME"
##  [6] "CITY"        "CITYNAME"    "ROUTE"       "ROUTENAME"   "RUR_URB"
## [11] "RUR_URBNAME" "MILEPT"      "LATITUDE"    "LONGITUD"
```

We observe that there are no outliers or special issues with this final table. Additionally, all the variables are relevant for creating a map of black spots, so we have decided to keep all the variables, concluding Phase 3. In this project, we are not going to model, evaluate, or deploy, as the objective was to put into practice in a comprehensible way the feature engineering process, which is considered one of the most complex and time-consuming parts of a data mining project.

BIBLIOGRAPHY

Jolliffe, Ian T. 2002. *Principal Component Analysis for Special Types of Data*. Springer.

"What Is CRISP DM? - Data Science Process Alliance." n.d. https://www.datascience-pm.com/crisp-dm-2/.