CSE 3020–DATA VISUALIZATION

J Component Report

PATTERN IDENTIFICATION IN METROPOLITAN CITIES

By

Reg. No 19BCE1410	Name Ayyappan K M
Reg.No. 19BCE1491	Name S Naman
Reg.No. 19BCE1211	Name P.Sai Teja

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted to

Dr.Maheshwari S

School of Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

November 2021

ABSTRACT

Understanding the evolution on the social and environmental aspects of a big city, such as Barcelona, is a hard task, given the complex and even subjective answers one might obtain. Driven by the motivation to clarify such questions, this project seeks a better understanding regarding those topics about Barcelona, aiming to comprehend how and if the city has been evolving over the most recent years, by analysing social indicators, such as immigration and emigration, birth rates, life expectancy, unemployment and accidents, and also environmental ones, such as tree plantations.

To achieve this, data visualisations will be used in order to develop clarifications on the subject, by mapping data to visual variables, providing an easier way of understanding Barcelona's people and environment. Thus, this project has the purpose of finding a better and more complete way of presenting answers about Barcelona, extending the textual and statistical information provided by the various data sets.

INTRODUCTION

The motivation behind the choice of this project is to attain a deeper knowledge about the evolution of Barcelona by studying the chosen data, with the help of several different data visualisations. To accomplish this purpose, a main data set is used, comprised of ten different ones where it can be found information about the administration, urban environment, population, territory, economy and business in the city of Barcelona.
The questions this project aims to answer are mainly focused on indicators such as the urban environment and population, like for instance, questions such as "How has the unemployment in relation to the unemployment demand evolved in the last years. Our Project is divided into various sections where it is explained in deeper detail what questions this project aims to answer about Barcelona, the details of the chosen data sets, what other related work has been done and, finally, the project proposal.

To develop the proposed study, it is necessary to determine a guideline defining the main intended goals. This way, the research is conducted. To get an initial overview of Barcelona we pretend to investigate the distribution of the population over districts, gender and age. Life expectancy is a health quality indicator that is quite important when evaluating the quality of life of a population. Because of this, it is pretended to analyse the life expectancy over districts and how it has been evolving in the latest years as well as if the trend that women live more than men also applies to Barcelona. The birth rate of a city is important to know if its population is ageing or not but birth rate isn't the only metric representing the population fluctuation and should be complemented with the immigration rate. It is pretended to understand if the population of Barcelona is not only ageing or not but if it is diversifying and if the entrance of immigrants in the city has an impact on birth rate. Street trees are an important part of a city improving its population's quality of life. According to this, street trees have a bigger influence over people's health than an increase of income.

Unemployment is an important aspect of any city providing rich information about its economical situation. If one is to move to Barcelona then they should investigate this metric in order to understand where they should work and if their gender might influence this choice. Either visiting or living in a city, road safety is always an important part of public health. Because of this it'll be studied what times of the day and which areas of Barcelona are the safest to drive in and which zones should be avoided. Having set our research goals we shall now investigate the dataset we're working with and that will be used to answer all the research questions indicated above.

DATA SET

The full data set is comprised into ten different sections, each one corresponding to the different topics and indicators being studied.

A detailed analysis of each one follows:

Accidents – 2017

A list of accidents handled by the local police in the city of Barcelona, in 2017. This data set includes information about when and where the accident happened including the precise coordinates. It also includes the number of vehicles involved, the number of victims and their injury severity.

Births

Births by district and by neighbourhoods of the city of Barcelona, from 2013 to 2017, for boys and girls.

Population

Population by neighbourhood, by quinquennial ages (recurring every 5 years) and by gender of the city of Barcelona, from 2013 to 2017. Reading registers of inhabitants. The data set also includes unique IDs for the different neighbourhoods and districts.

Life Expectancy

Life Expectancy of the population of Barcelona between 2010 and 2014 containing information about neighbourhood and gender.

Street trees of the city of Barcelona

Name of the species and geographical location of the trees of the city of Barcelona located on public roads. The information contains, among other data, the scientific name, the common name, the height, the direction and the width of the sidewalk, etc. The trees of the parks are not included. The coordinates are expressed in the ETRS89 reference system.

Shapefile of the districts of Barcelona

Shapefile that represents the different districts of the city of Barcelona.

Unemployment

Registered unemployment and unemployment demand by district, neighbourhood and gender in the city of Barcelona. It includes the months from 2013 to 2017 and unique IDs for the districts and neighbourhoods.

Plots and their Observations

In order to study some aspects about the data being used, for example its distribution, the presence of missing values, or their range, the following visualisations were created to provide such results.

Population Distribution

This visualisation allows the verification of what districts are more or less populated and how many man and women constitute the population in each district, with the gender mapped to the colour. Despite of not being present in figure 2, in Tableau it is possible to check these results for every year.

Some conclusions can be derived, such as Example being the most populated district, and only in Ciutat Vella there are more men than women.

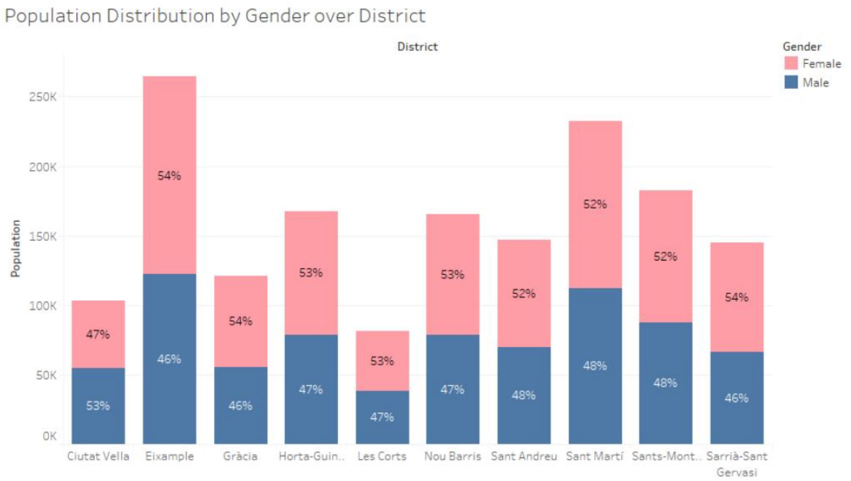


Fig.1. Distribution of Barcelona’s Population by Gender over District

Births Distribution

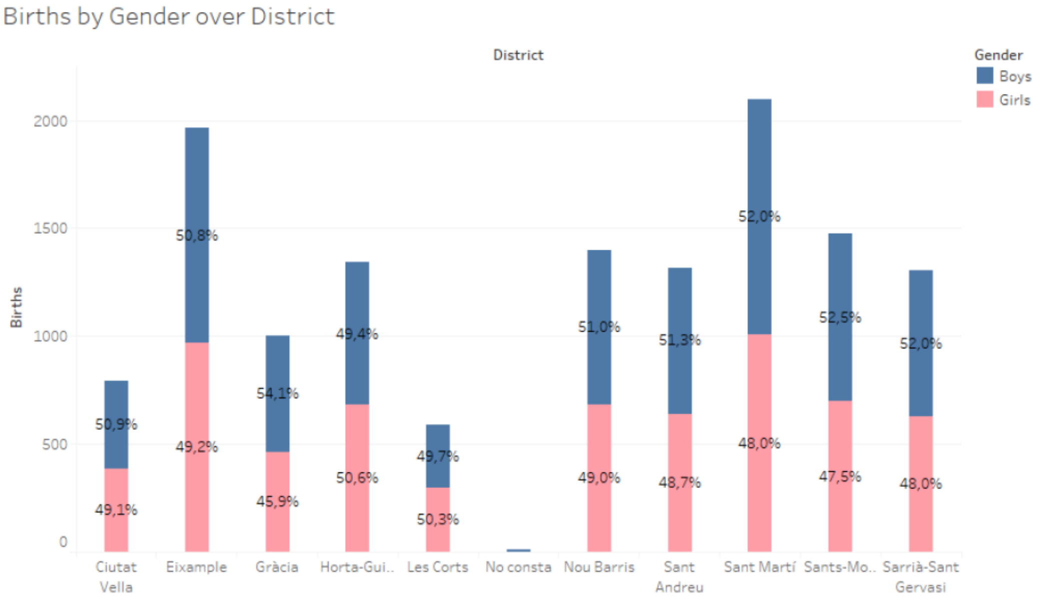


Fig.2. Distribution of Births by Gender over District

INFERENCE FROM ABOVE MODELS

Similar to the visualisation in figure 1, this one presents the number of births by gender (mapped to the colour) for each district in Barcelona, with the possibility to check for every year. It is also possible to see there are some missing values, present in the "No Consta" column: there are a total of seven, only in the year 2013, as it can be seen in figure 3. We can verify there isn’t much difference in the number of births for both genders, since the length of the bar is similar, and that Example, the most populated one, is also the district with more births.

Immigration Distribution

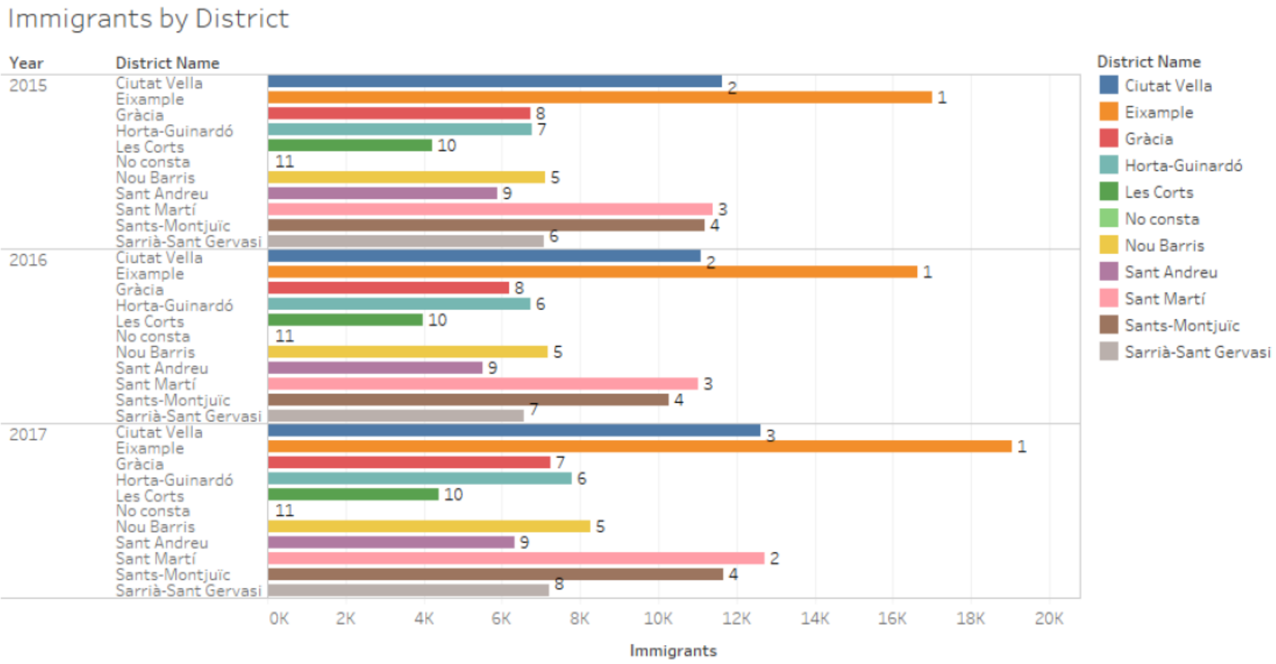


Fig.3. Distribution of Immigrants by District over Year

Inference from Above Model

On figure 3, it is presented a bar chart with the ranked number of immigrants who moved to each district (mapped to the colour), for every year present in the data. Some missing values can be spotted, in the district column "No Consta", totalling 17: 6 in the year 2015, 7 in 2016 and the remaining 2 in 2017.

Along the three years, a tendency can be spotted for most districts: from the first year to the second, there is a slight decrease in the number of immigrants, and from the second to the last year, that number increases to a value which is greater than the first year. Also, Example is for all years the district with most immigrants.

Unemployment Distribution

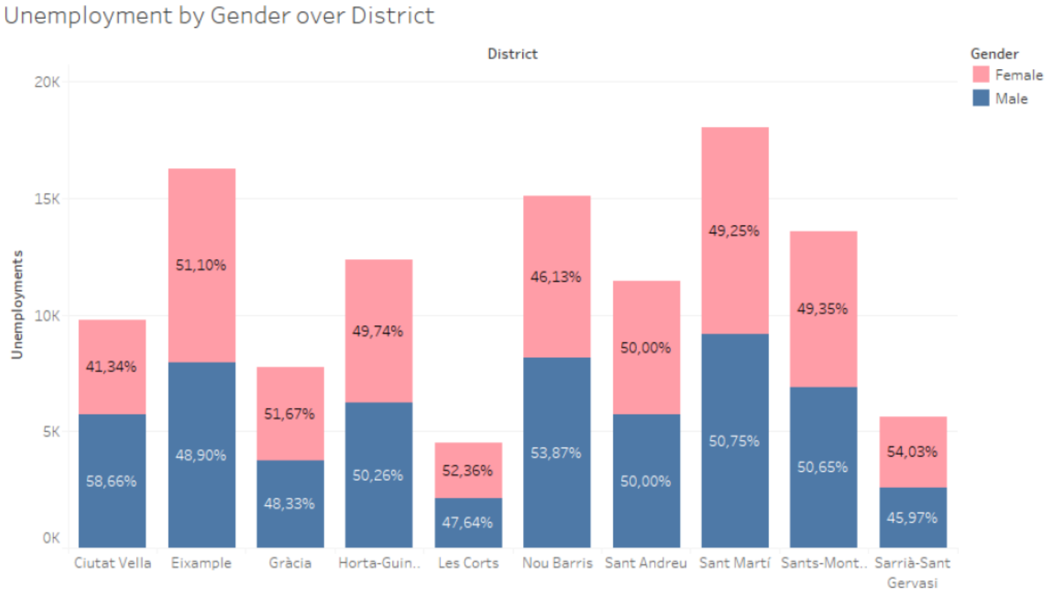


Fig.4. Distribution of Barcelona Unemployment by District

Inference from Above Model

In this visualisation, figure 4, the number of unemployed citizens of Barcelona are presented by gender, which is mapped to the colour, over the districts, being also possible in Tableau to verify these values year by year.

In Sant Marti, the number of unemployment is the greatest, exceeding 18.000 people, while in Les Corts it is the smallest. There isn't much difference between unemployed man and women, however in six districts of the existing ten, there are more unemployed women.

Accidents Distribution

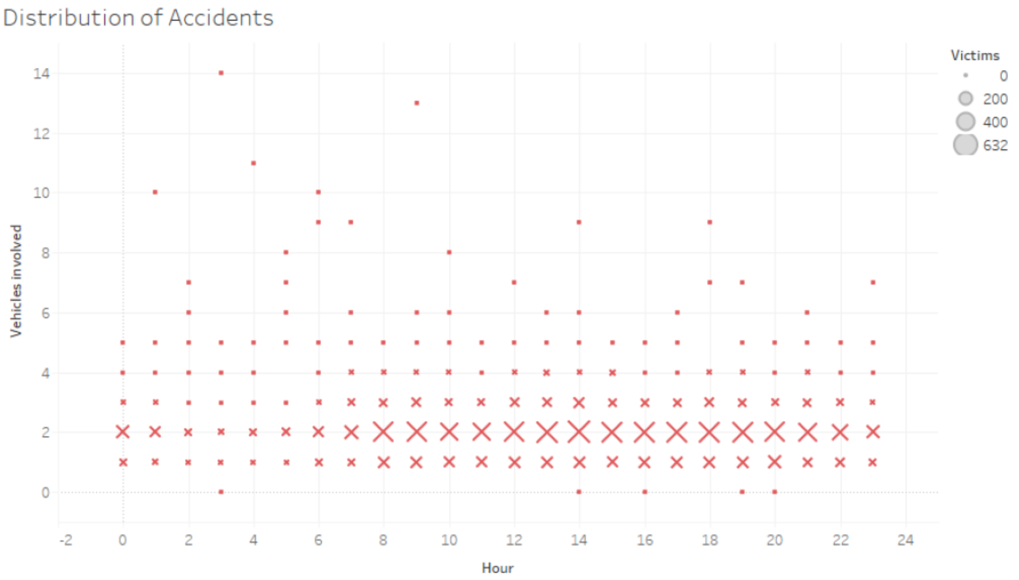


Fig.5. Accident Distribution over a period of time

Inference from Above Model

The final visualisation, displayed on figure 5, shows the distribution of the number of accident victims (mapped to size) by each hour of the day and the number of vehicles. We can see that accidents with most victims involve two vehicles and occur between 1PM and 8PM, with most victims exactly at 2PM.

PROPOSAL AND RESULTS

To analyse the indicators referred on the previous sections we've developed five different dashboards comprised of several visualisations. Both the dashboards and its visualisations are further described in detail.

Population Overview

The dashboard of figure 14 aims to answer several questions about the distribution of population of Barcelona and its life expectancy. This dashboard is interactive so the user can select one or more districts to focus and analyse them in further detail.

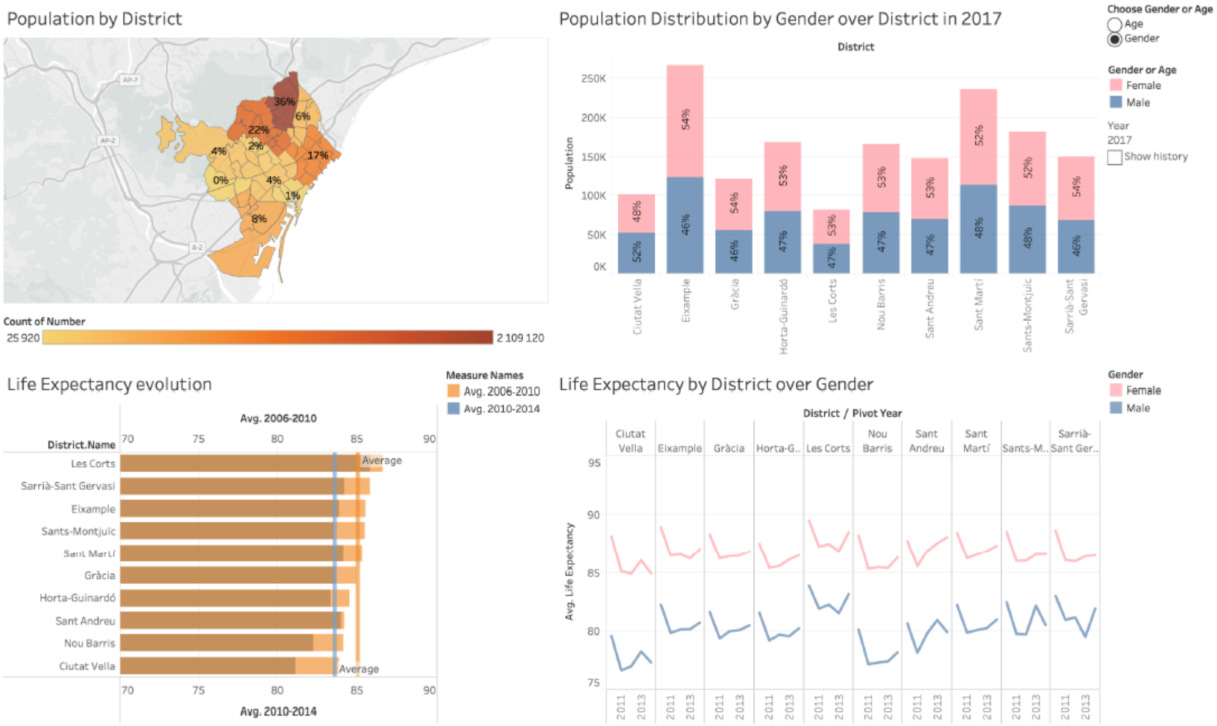
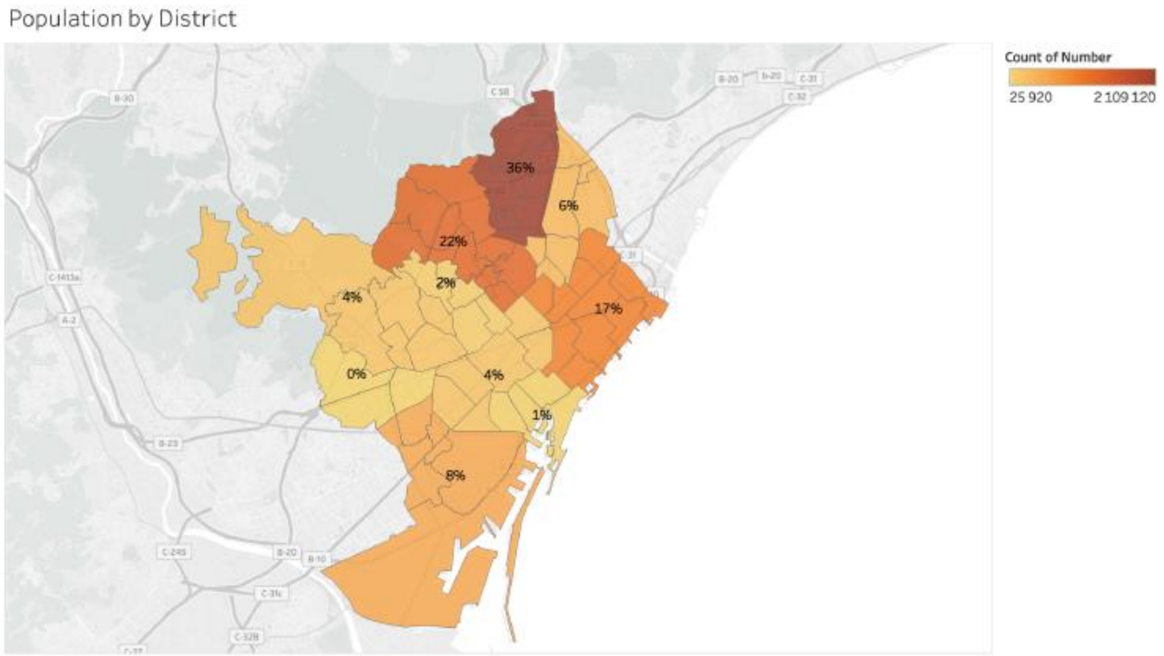


Fig.6. Population Overview Dashboard

The Population Overview dashboard is composed of the following visualisations:

Population by District



Map based on Longitude and Latitude. Color shows count of Number. The marks are labeled by % of Total Count of Number. Details are shown for District Name.

Fig.7. Population By District

Inference from above Model

The visualisation Population by District on figure 7 shows the percentage of the total population that lives in each district. The number of individuals is mapped to the intensity of the colour and the labels show the percentage over the total of Barcelona’s population.

Population Distribution by Gender or Age over District:

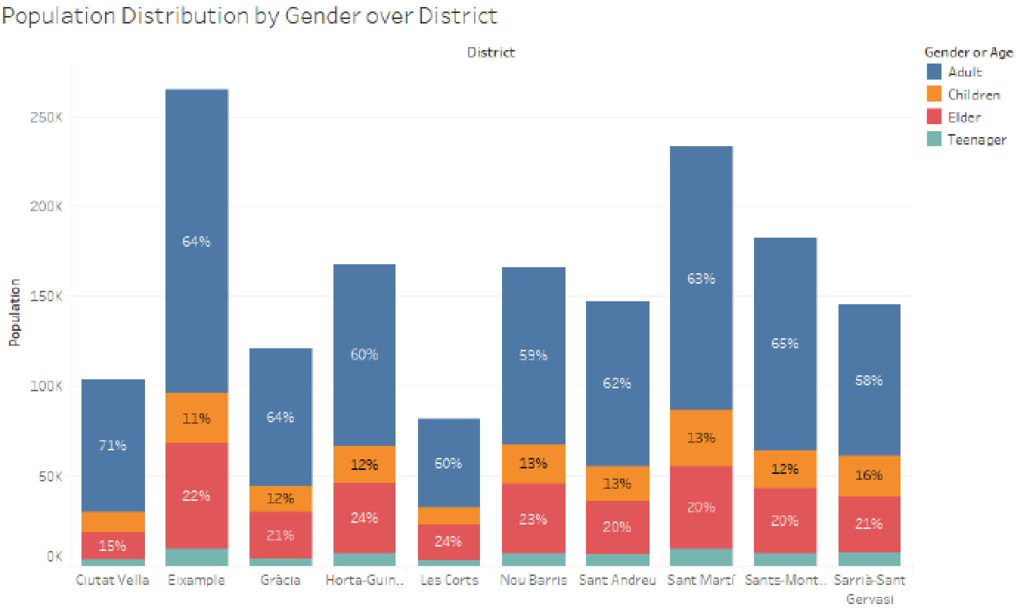


Fig.8. Population Distribution by Gender over District when Age is chosen

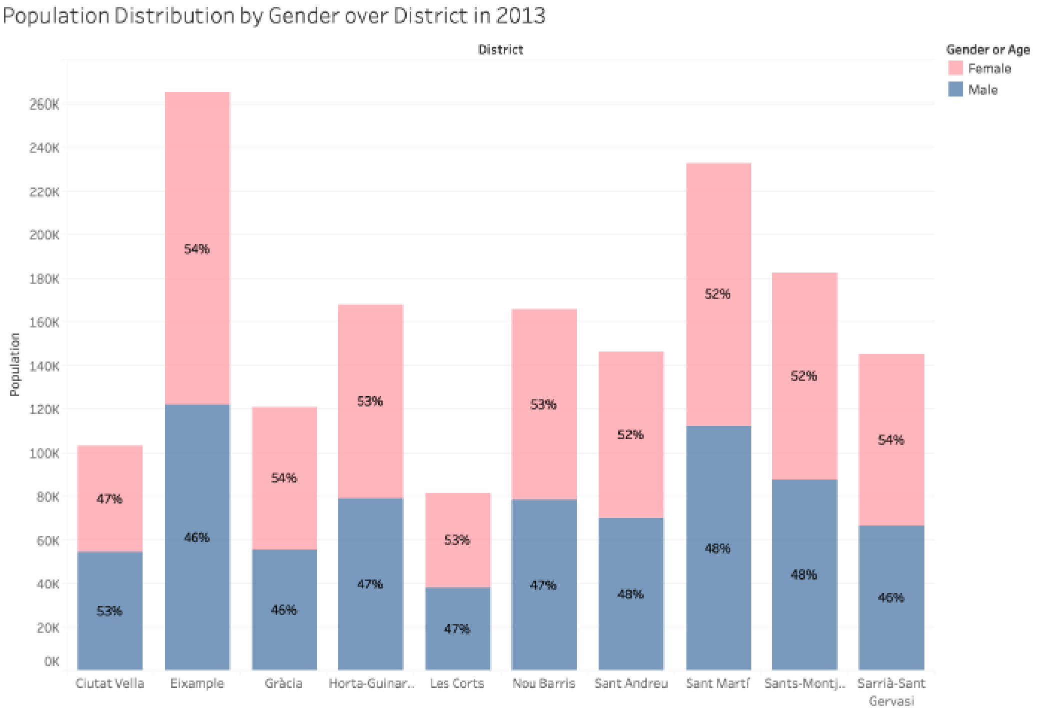


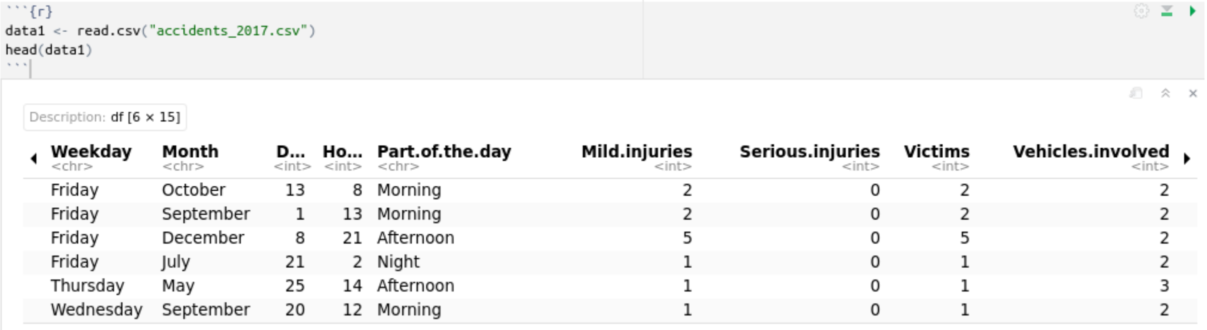
Fig.9. Population Distribution by Gender over District when Gender is chosen

Inference from Above Models

In the visualisation Population Distribution by Age/ Gender over District, shown on figures 8 and 9, the user can choose between visualising the data by age - figure 8 - or gender -figure 9. It shows the distribution of the age groups, or gender, by district between the years of 2013 and 2017. In the case Age is chosen, the different age groups - Children, Teenager, Adult and Elder - are mapped to the colour and the labels show the percentage over the total of the district's population. In case the Gender is chosen, the gender is mapped to the colour and the labels represent, once more, the percentage over the total of the district's population. In both cases, the x-axis contains the districts and the y-axis has the population number. This visualisation answers the question of "How is the population distributed over districts and gender?" by presenting the districts over the columns and population over the rows.

Machine Learning Models

Data Preprocessing:



~~~
```{r}
data1 <- read.csv("accidents_2017.csv")
head(data1)
```
~~~

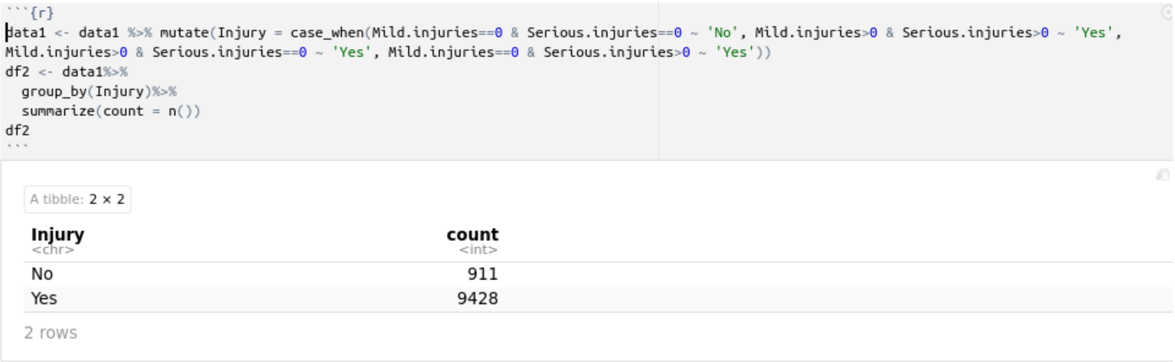
Description: df [6 × 15]

Weekday <chr>	Month <chr>	D... <int>	Ho... <int>	Part.of.the.day <chr>	Mild.injuries <int>	Serious.injuries <int>	Victims <int>	Vehicles.involved <int>
Friday	October	13	8	Morning	2	0	2	2
Friday	September	1	13	Morning	2	0	2	2
Friday	December	8	21	Afternoon	5	0	5	2
Friday	July	21	2	Night	1	0	1	2
Thursday	May	25	14	Afternoon	1	0	1	3
Wednesday	September	20	12	Morning	1	0	1	2

Fig.10. Initial Data View

Count of people with mild injuries According to intensity

Since there is no proper result column in the dataset for injuries we have calculated a new column called injuries such that it has values either yes or no based on the columns mild injuries and serious injuries.



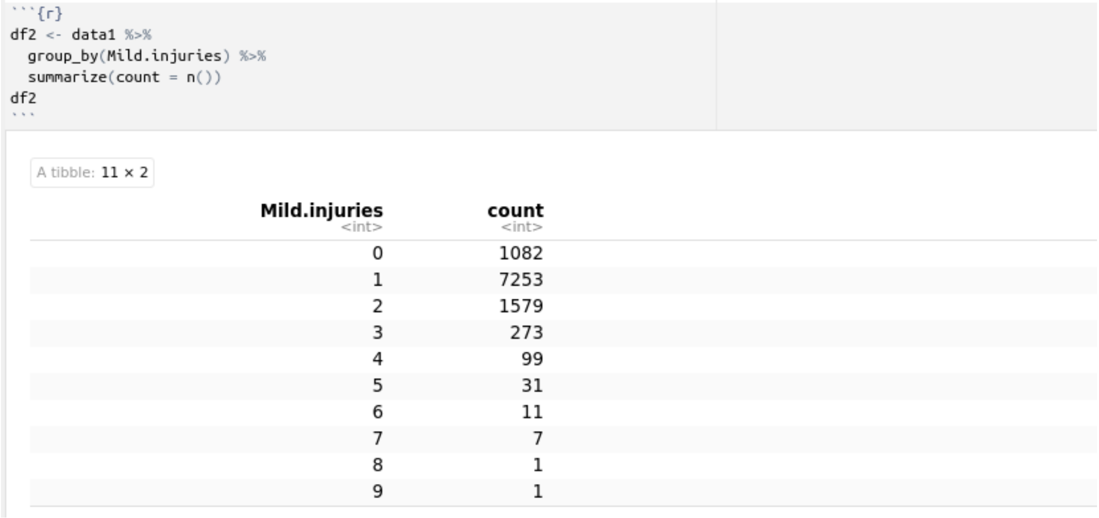
~~~~
```{r}
data1 <- data1 %>% mutate(Injury = case_when(Mild.injuries==0 & Serious.injuries==0 ~ 'No', Mild.injuries>0 & Serious.injuries>0 ~ 'Yes',
Mild.injuries>0 & Serious.injuries==0 ~ 'Yes', Mild.injuries==0 & Serious.injuries>0 ~ 'Yes'))
df2 <- data1%>%
  group_by(Injury)%>%
  summarize(count = n())
df2
```
~~~~

A tibble: 2 × 2

Injury <chr>	count <int>
No	911
Yes	9428

2 rows

Fig.11. Injury count

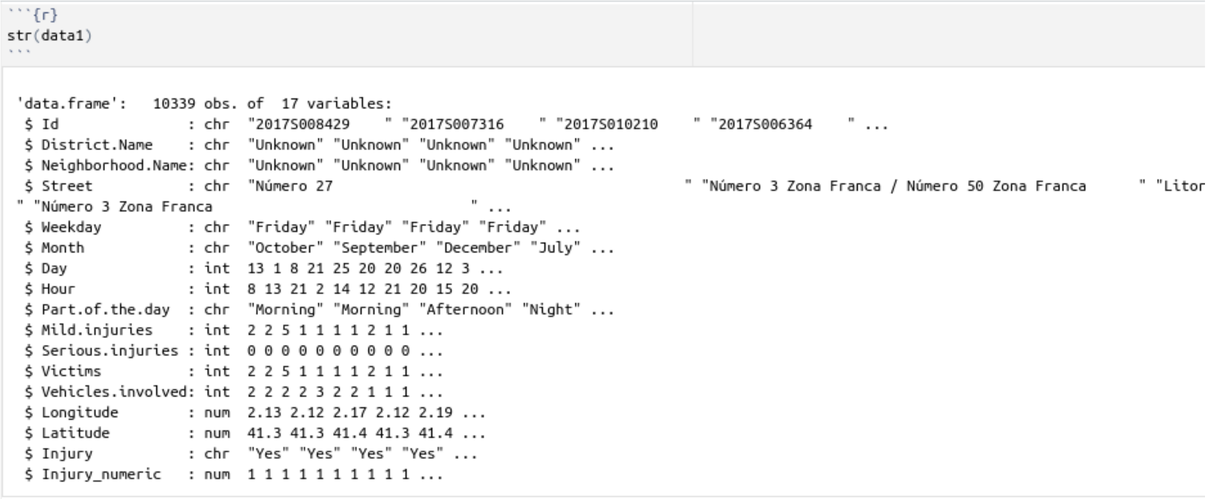


~~~~
```{r}
df2 <- data1 %>%
  group_by(Mild.injuries) %>%
  summarize(count = n())
df2
```
~~~~

A tibble: 11 × 2

Mild.injuries <int>	count <int>
0	1082
1	7253
2	1579
3	273
4	99
5	31
6	11
7	7
8	1
9	1

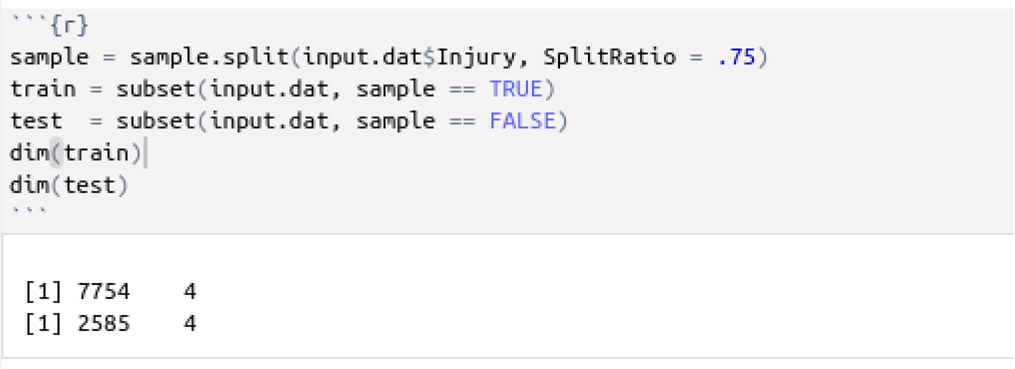
Fig.12.Injury count in new Column



~~~~
```{r}
str(data1)
```

'data.frame':	10339 obs. of  17 variables:
 $ Id               : chr  "2017S008429    " "2017S007316    " "2017S010210    " "2017S006364    " ...
 $ District.Name    : chr  "Unknown" "Unknown" "Unknown" "Unknown" ...
 $ Neighborhood.Name: chr  "Unknown" "Unknown" "Unknown" "Unknown" ...
 $ Street           : chr  "Número 27                                        " "Número 3 Zona Franca / Número 50 Zona Franca      " "Litor
" "Número 3 Zona Franca                              " ...
 $ Weekday          : chr  "Friday" "Friday" "Friday" "Friday" ...
 $ Month            : chr  "October" "September" "December" "July" ...
 $ Day              : int  13 1 8 21 25 20 20 26 12 3 ...
 $ Hour             : int  8 13 21 2 14 12 21 20 15 20 ...
 $ Part.of.the.day  : chr  "Morning" "Morning" "Afternoon" "Night" ...
 $ Mild.injuries    : int  2 2 5 1 1 1 1 2 1 1 ...
 $ Serious.injuries : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Victims          : int  2 2 5 1 1 1 1 2 1 1 ...
 $ Vehicles.involved: int  2 2 2 2 3 2 2 1 1 1 ...
 $ Longitude        : num  2.13 2.12 2.17 2.12 2.19 ...
 $ Latitude         : num  41.3 41.3 41.4 41.3 41.4 ...
 $ Injury           : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ Injury_numeric   : num  1 1 1 1 1 1 1 1 1 1 ...
~~~~

Fig.13. Structure of the Data

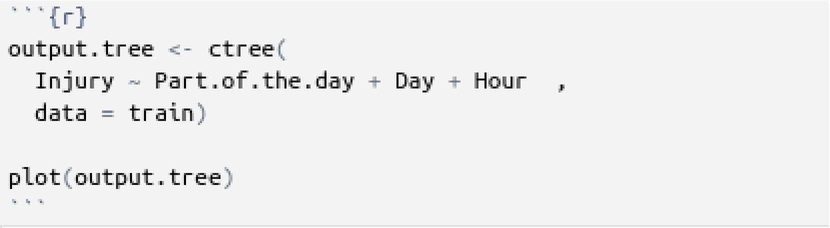


~~~
```{r}
sample = sample.split(input.dat$Injury, SplitRatio = .75)
train = subset(input.dat, sample == TRUE)
test  = subset(input.dat, sample == FALSE)
dim(train)
dim(test)
```
~~~

```
[1] 7754    4
[1] 2585    4
```

Fig.14. Split into Test and Train Set

Decision Tree



~~~
```{r}
output.tree <- ctree(
  Injury ~ Part.of.the.day + Day + Hour   ,
  data = train)

plot(output.tree)
```
~~~

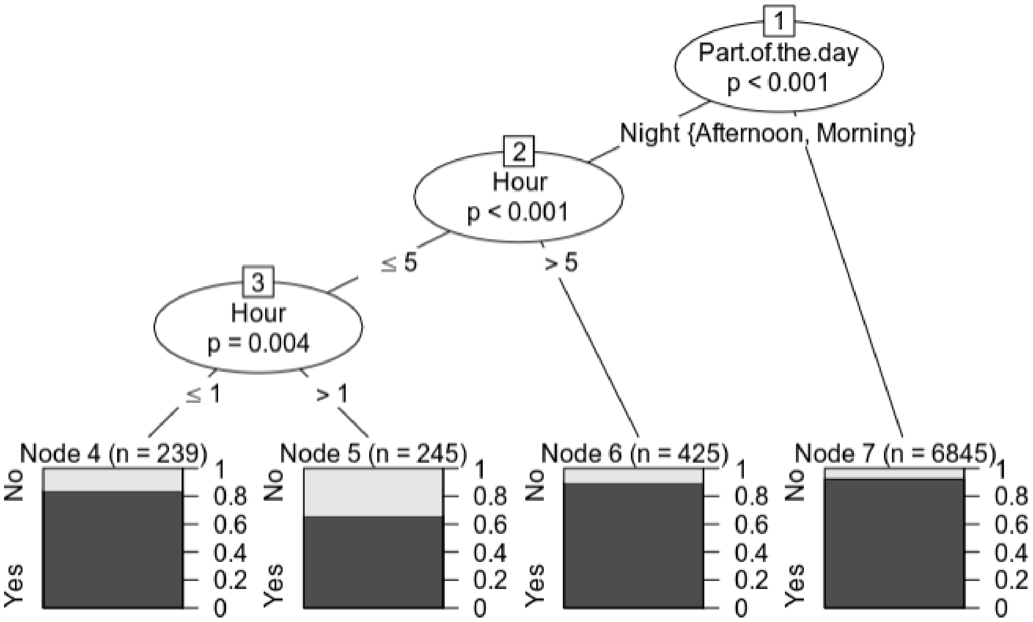
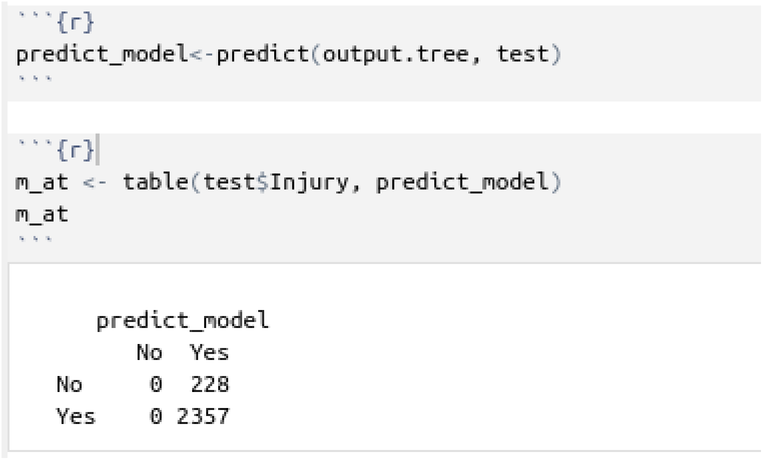


Fig.15. Decision Tree Model

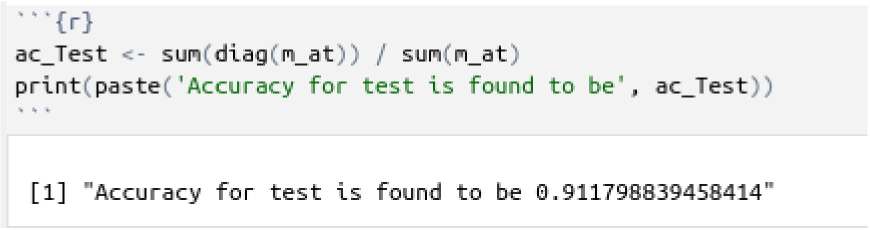


~~~
```{r}
predict_model<-predict(output.tree, test)
```
~~~

~~~
```{r}
m_at <- table(test$Injury, predict_model)
m_at
```
~~~

```
     predict_model
        No  Yes
  No     0  228
  Yes    0 2357
```

Fig.16. Prediction and confusion matrix

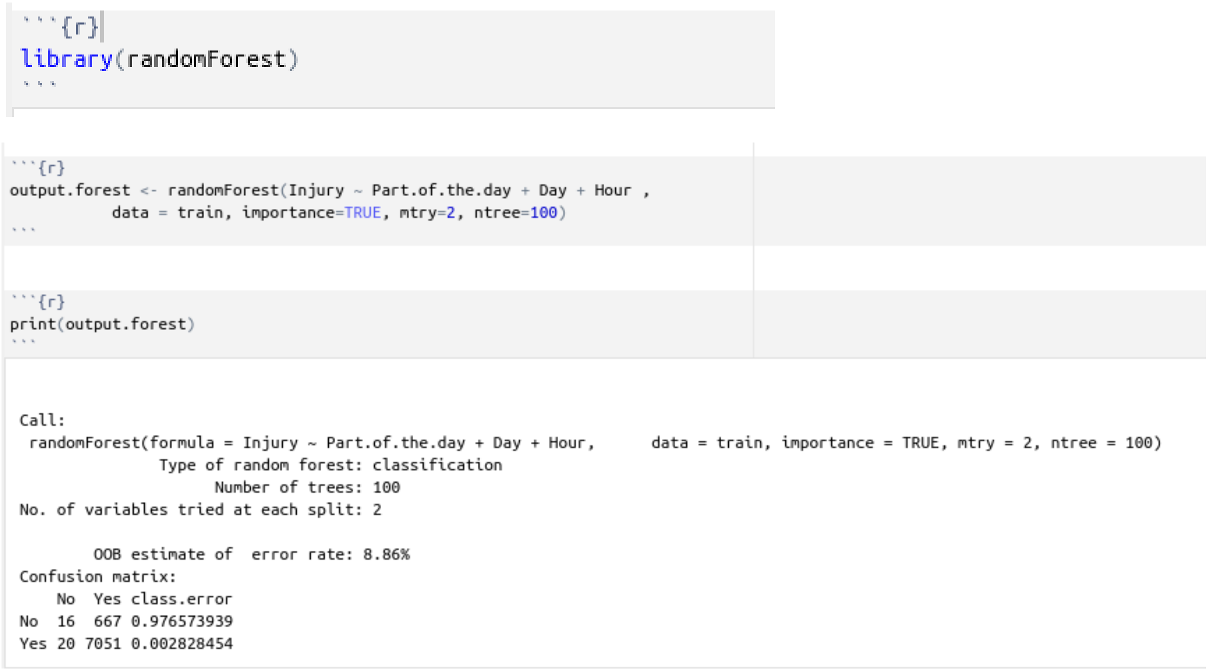


~~~
```{r}
ac_Test <- sum(diag(m_at)) / sum(m_at)
print(paste('Accuracy for test is found to be', ac_Test))
```
~~~

```
[1] "Accuracy for test is found to be 0.911798839458414"
```

Fig.17. Accuracy based on test data

Random Forest



~~~
```{r}
library(randomForest)
```
~~~

~~~
```{r}
output.forest <- randomForest(Injury ~ Part.of.the.day + Day + Hour ,
           data = train, importance=TRUE, mtry=2, ntree=100)
```
~~~

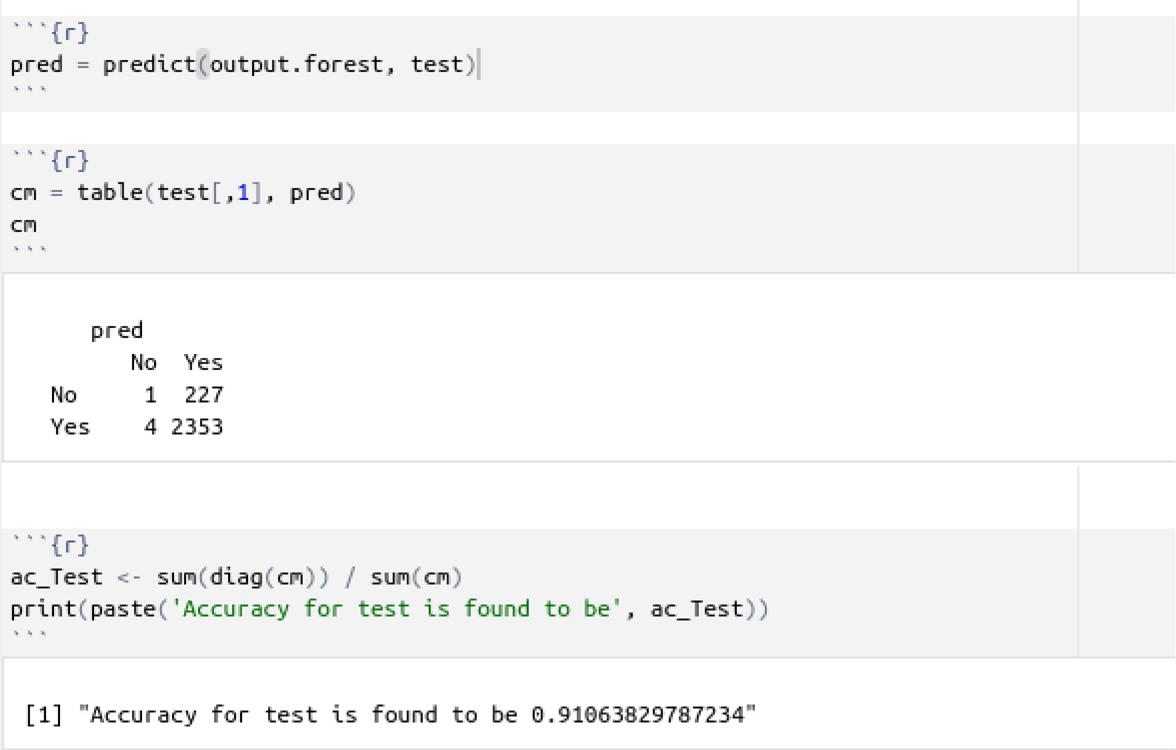
~~~
```{r}
print(output.forest)
```
~~~

```
Call:
 randomForest(formula = Injury ~ Part.of.the.day + Day + Hour,      data = train, importance = TRUE, mtry = 2, ntree = 100)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 2

        OOB estimate of  error rate: 8.86%
Confusion matrix:
    No  Yes class.error
No  16  667 0.976573939
Yes 20 7051 0.002828454
```

Fig.18. Random Forest Model

Inference From above Model

As we can see the random forest model on default spans 500 trees and we have giver mtry as 2 which is number of variables available for splitting at each tree node.
Also the model gives the OOB error rate of **8.86%.**



~~~
```{r}
pred = predict(output.forest, test)
```
~~~

~~~
```{r}
cm = table(test[,1], pred)
cm
```
~~~

```
     pred
        No  Yes
  No     1  227
  Yes    4 2353
```

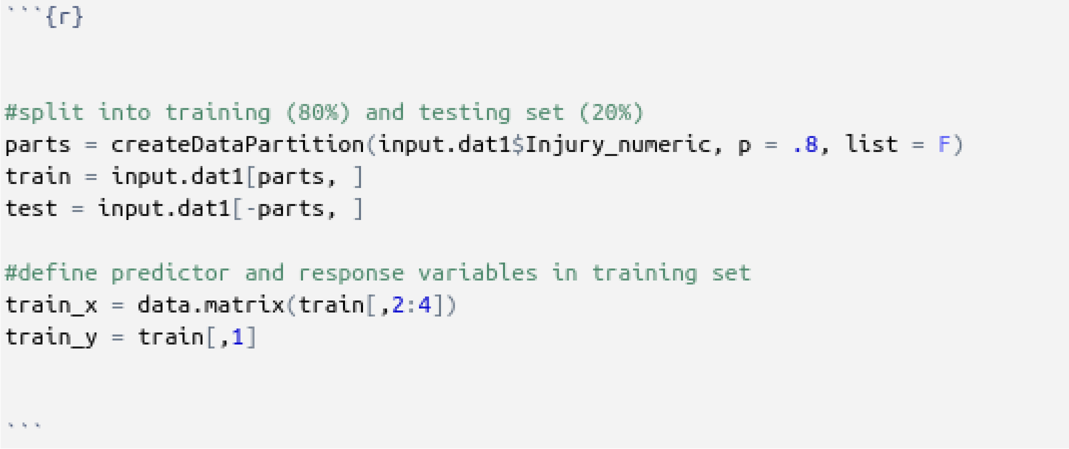
~~~
```{r}
ac_Test <- sum(diag(cm)) / sum(cm)
print(paste('Accuracy for test is found to be', ac_Test))
```
~~~

```
[1] "Accuracy for test is found to be 0.91063829787234"
```

Fig.19.Prediction and test Results

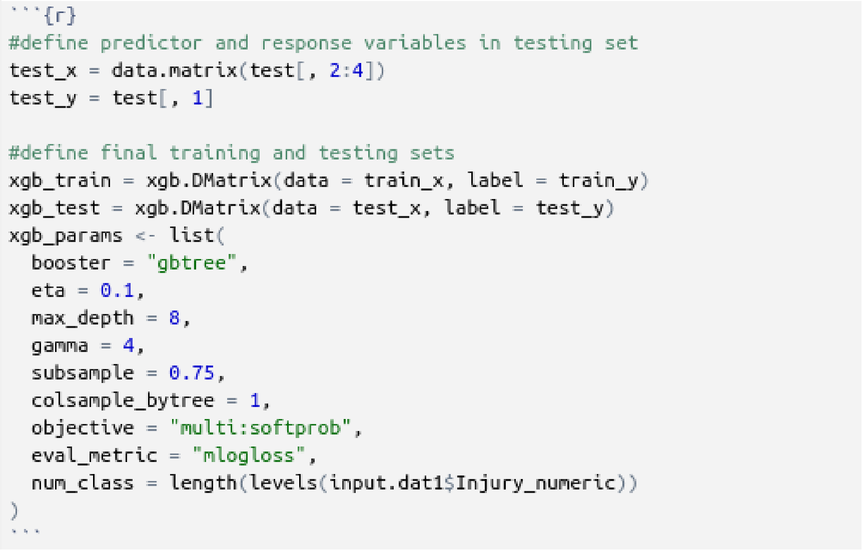
XGBoost Algorithm

Data Preprocessing

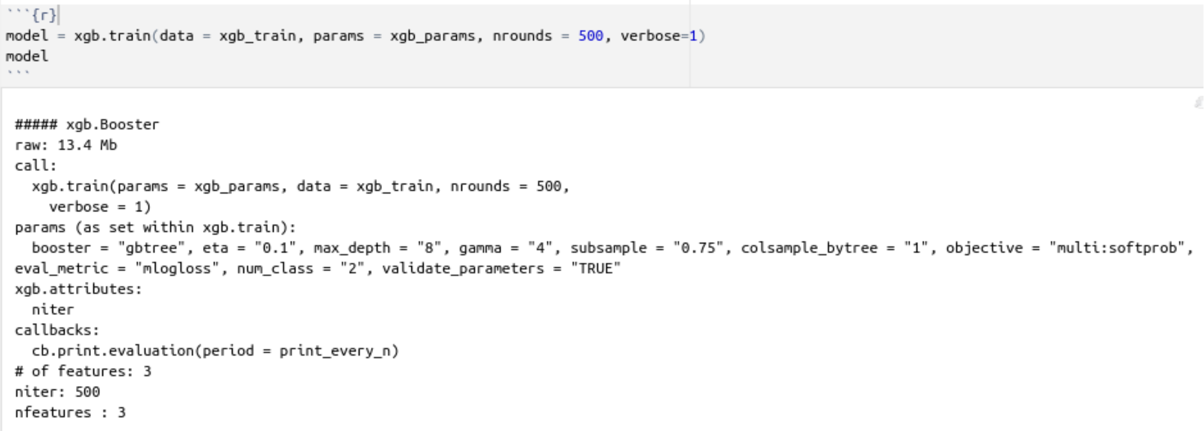


~~~
```{r}


#split into training (80%) and testing set (20%)
parts = createDataPartition(input.dat1$Injury_numeric, p = .8, list = F)
train = input.dat1[parts, ]
test = input.dat1[-parts, ]

#define predictor and response variables in training set
train_x = data.matrix(train[,2:4])
train_y = train[,1]


```
~~~

Fig.20 Data Preprocessing



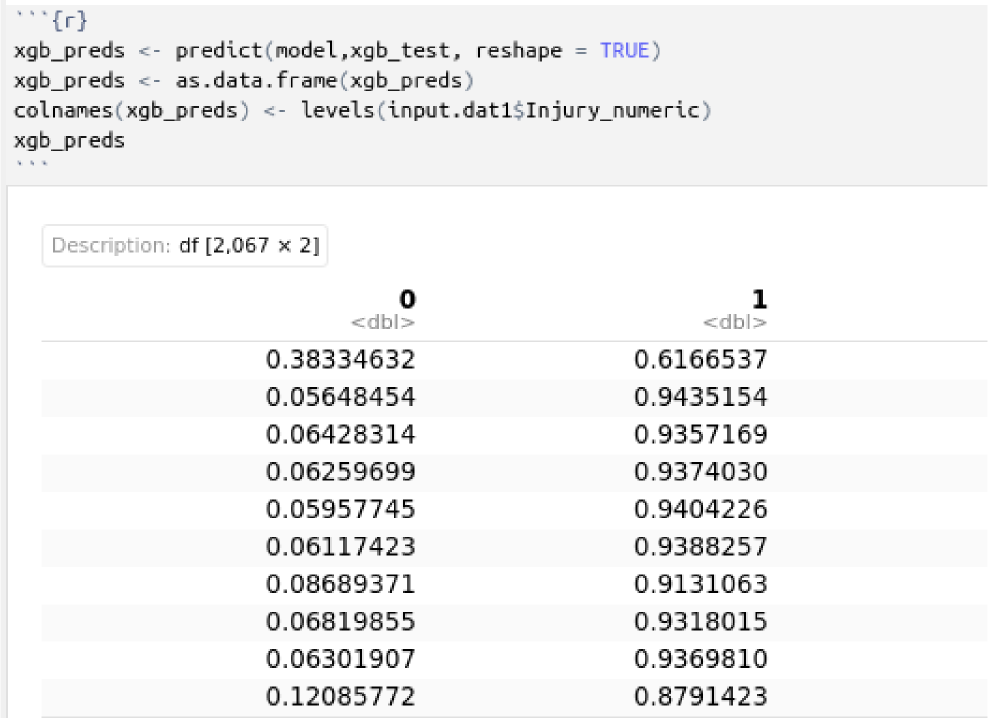
~~~
```{r}
#define predictor and response variables in testing set
test_x = data.matrix(test[, 2:4])
test_y = test[, 1]

#define final training and testing sets
xgb_train = xgb.DMatrix(data = train_x, label = train_y)
xgb_test = xgb.DMatrix(data = test_x, label = test_y)
xgb_params <- list(
  booster = "gbtree",
  eta = 0.1,
  max_depth = 8,
  gamma = 4,
  subsample = 0.75,
  colsample_bytree = 1,
  objective = "multi:softprob",
  eval_metric = "mlogloss",
  num_class = length(levels(input.dat1$Injury_numeric))
)
```
~~~



~~~
```{r}
model = xgb.train(data = xgb_train, params = xgb_params, nrounds = 500, verbose=1)
model
```
~~~

```
##### xgb.Booster
raw: 13.4 Mb
call:
  xgb.train(params = xgb_params, data = xgb_train, nrounds = 500,
    verbose = 1)
params (as set within xgb.train):
  booster = "gbtree", eta = "0.1", max_depth = "8", gamma = "4", subsample = "0.75", colsample_bytree = "1", objective = "multi:softprob",
eval_metric = "mlogloss", num_class = "2", validate_parameters = "TRUE"
xgb.attributes:
  niter
callbacks:
  cb.print.evaluation(period = print_every_n)
# of features: 3
niter: 500
nfeatures : 3
```

Fig.21. Model Training

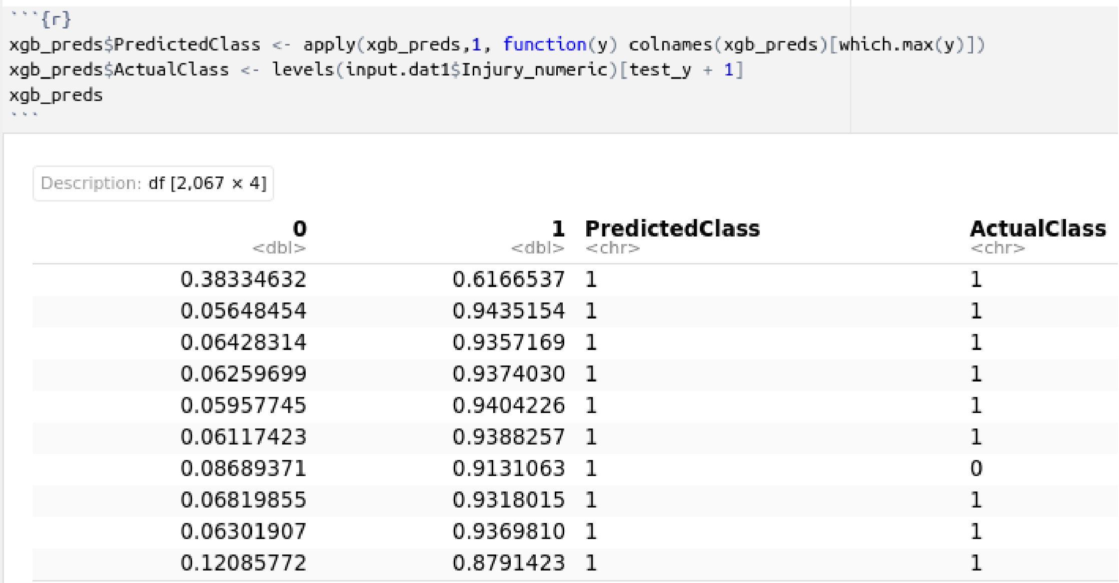


~~~
```{r}
xgb_preds <- predict(model,xgb_test, reshape = TRUE)
xgb_preds <- as.data.frame(xgb_preds)
colnames(xgb_preds) <- levels(input.dat1$Injury_numeric)
xgb_preds
```
~~~

Description: df [2,067 × 2]

0 <dbl>	1 <dbl>
0.38334632	0.6166537
0.05648454	0.9435154
0.06428314	0.9357169
0.06259699	0.9374030
0.05957745	0.9404226
0.06117423	0.9388257
0.08689371	0.9131063
0.06819855	0.9318015
0.06301907	0.9369810
0.12085772	0.8791423

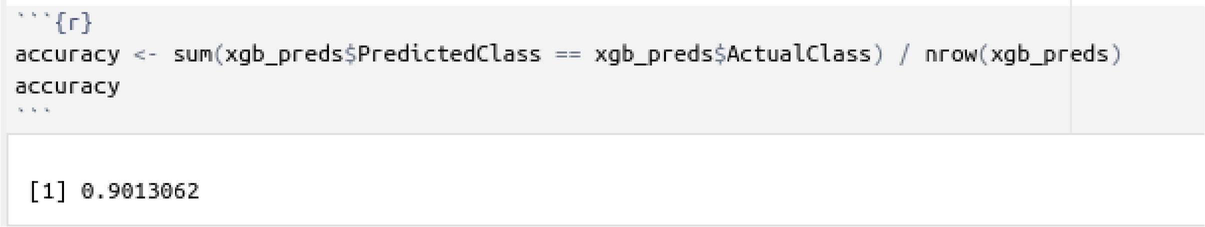
Fig.22.XGBOOST



~~~
```{r}
xgb_preds$PredictedClass <- apply(xgb_preds,1, function(y) colnames(xgb_preds)[which.max(y)])
xgb_preds$ActualClass <- levels(input.dat1$Injury_numeric)[test_y + 1]
xgb_preds
```
~~~

```
Description: df [2,067 × 4]

           0          1  PredictedClass        ActualClass
       <dbl>      <dbl>  <chr>                 <chr>
  0.38334632  0.6166537  1                     1
  0.05648454  0.9435154  1                     1
  0.06428314  0.9357169  1                     1
  0.06259699  0.9374030  1                     1
  0.05957745  0.9404226  1                     1
  0.06117423  0.9388257  1                     1
  0.08689371  0.9131063  1                     0
  0.06819855  0.9318015  1                     1
  0.06301907  0.9369810  1                     1
  0.12085772  0.8791423  1                     1
```

Fig.23.Comparing Prediction data and actual data



~~~
```{r}
accuracy <- sum(xgb_preds$PredictedClass == xgb_preds$ActualClass) / nrow(xgb_preds)
accuracy
```

[1] 0.9013062
~~~

Fig,24 .Accuracy

CONCLUSION

As we can see from the above 3 models decision tree has the highest accuracy score of 91.17% but random forest and xgboost also have good accuracy scores and hence we can conclude that decision tree is the appropriate model to predict if a person might be injured based on the the variables such as day, hour and part of the day.

Also from the visualisations done using Tableau we are able to easily understand the a big city like Barcelona where we can infer that the female population is increasing gradually between the years 2014-17 in various districts of Barcelona.The birth rate of a city is important to know if its population is ageing or not but birth rate isn’t the only metric representing the population fluctuation and should be complemented with the immigration rate. It is pretended to understand if the population of Barcelona is not only ageing or not but if it is diversifying and if the entrance of immigrants in the city has an impact on birth rate.

References

[1] Antoni Plascència, Carme Borrell, Andjosep M. Antis Emergency Department and Hospital admissions and deaths from traffic injuries in Barcelona, Spai. A one-year population-based study.

[2] Map of the street trees in Barcelona, Nathalie Richer

[3] NDVI vegetation index and street trees in Barcelona city - v 1.2, Juanjo Vidal

[4] Xavier Vivancos García, Discovering Barcelona Part 1.

[5] Kardan, O., Gozdyra, P., Misic, B., Moola, F., Palmer, L. J., Paus, T., & Berman, M.G. (2015). Neighbourhood greenspace and health in a large urban center. Scientific Reports, 5, 11610.

[6]W. Tysiak, "Regression Analysis of Intrinsic Linear Models with Automated Transformations of Monotone Predictors.

[7] A. Karama, M. Farouk and A. Atiya, "A Multi Linear Regression Approach for Handling Missing Values with Unknown Dependent Variable (MLRMUD).