Data Cleaning, Data Exploration, Data Modelling and Model Deployment using R

Famous Salami | salamifamous@gmail.com

15-10-2023

Dataset: Records from an insurance company's recently insured drivers.

.

Importing Libraries

```r
#clears the R environment
rm(list=ls())
```

```r
#sets working directory to current directory
setwd(getwd())
```

.

Loading the dataset

```r
#Reading the data file
insuredb <- read.csv("insurance.csv", stringsAsFactors = T)
```

.

Inspecting and understanding the data

```r
#structure of the data
str(insuredb)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ insurance      : num  232 310 506 174 527 ...
##  $ driver_age     : int  67 54 47 80 54 32 44 66 25 61 ...
##  $ car_value      : Factor w/ 534 levels "£0.7K","£1,400",..: 265 36 359 436 520 432 508 73 319 73 .
##  $ num_accident   : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ annual_mileage : int  9000 11950 12900 3730 9130 8200 6900 12820 6730 6380 ...
##  $ car_age        : Factor w/ 167 levels "0Y 10M","0Y 11M",..: 100 139 91 155 8 68 103 134 91 131 ..
##  $ excess         : int  50 100 200 150 0 100 50 100 100 0 ...
```

```
##  $ car_reg        : Factor w/ 1000 levels "AA14NVN","AA15LSH",..: 567 96 705 165 215 522 651 886 426
##  $ gender         : Factor w/ 4 levels "F","female","M",..: 3 4 3 1 2 4 3 4 2 4 ...
##  $ alarm          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 2 1 ...
##  $ insurance_group: Factor w/ 3 levels "Group 1","Group 2",..: 2 2 3 2 2 2 1 2 1 1 ...
##  $ crime_area     : Factor w/ 4 levels "","high","low",..: 3 4 4 2 2 4 3 3 3 4 ...
##  $ country        : Factor w/ 1 level "UK": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
#in-depth view of the data
describe(insuredb)
```

```
## insuredb
##
##  13  Variables      1000  Observations
## ---------------------------------------------------------------------------------
## insurance
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0      989        1    310.8    122.6    171.8    185.3
##      .25      .50      .75      .90      .95
##    224.8    290.6    374.9    466.9    527.0
##
## lowest : 150.32 150.5  151.08 152    152.51, highest: 668.88 676.06 680.02 726.74 780.42
## ---------------------------------------------------------------------------------
## driver_age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       70        1    47.52    19.36       21       25
##      .25      .50      .75      .90      .95
##       34       47       60       71       77
##
## lowest : 17 18 19 20 21, highest: 82 83 84 85 86
## ---------------------------------------------------------------------------------
## car_value
##        n  missing distinct
##     1000        0      534
##
## lowest : £0.7K  £1,400 £1,900 £1.5K  £1.6K , highest: £9.5K  £9.6K  £9.7K  £9.8K  £9.9K
## ---------------------------------------------------------------------------------
## num_accident
##        n  missing distinct     Info     Mean      Gmd
##     1000        0        6    0.857    0.799   0.9457
##
## Value          0     1     2     3     4     5
## Frequency    466   339   137    49     6     3
## Proportion 0.466 0.339 0.137 0.049 0.006 0.003
##
## For the frequency table, variable is rounded to the nearest 0
## ---------------------------------------------------------------------------------
## annual_mileage
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      968       32      659        1    10055     3471     4994     6070
##      .25      .50      .75      .90      .95
##     8055    10075    12150    14010    15146
##
## lowest :  1530  1640  1650  1800  2070, highest: 18540 18610 18630 18790 19300
## ---------------------------------------------------------------------------------
```

```
## car_age
##        n  missing distinct
##     1000        0      167
##
## lowest : 0Y 10M 0Y 11M 0Y 2M  0Y 3M  0Y 8M , highest: 9Y 5M  9Y 6M  9Y 7M  9Y 8M  9Y 9M
## --------------------------------------------------------------------------------
## excess
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       11    0.972    152.5    104.1        0       50
##      .25      .50      .75      .90      .95
##      100      150      200      300      350
##
## Value          0    50   100   150   200   250   300   350   400   450   500
## Frequency     72   136   211   229   139   106    53    36    14     3     1
## Proportion 0.072 0.136 0.211 0.229 0.139 0.106 0.053 0.036 0.014 0.003 0.001
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## car_reg
##        n  missing distinct
##     1000        0     1000
##
## lowest : AA14NVN AA15LSH AB18WSH AC66EKU AD18VHJ
## highest: ZW69EWH ZW69ZJP ZX19RPG ZZ20BVA ZZ65GGQ
## --------------------------------------------------------------------------------
## gender
##        n  missing distinct
##     1000        0        4
##
## Value           F female      M   male
## Frequency     297    198    303    202
## Proportion  0.297  0.198  0.303  0.202
## --------------------------------------------------------------------------------
## alarm
##        n  missing distinct
##     1000        0        2
##
## Value         No   Yes
## Frequency    393   607
## Proportion 0.393 0.607
## --------------------------------------------------------------------------------
## insurance_group
##        n  missing distinct
##     1000        0        3
##
## Value      Group 1 Group 2 Group 3
## Frequency      394     486     120
## Proportion   0.394   0.486   0.120
## --------------------------------------------------------------------------------
## crime_area
##        n  missing distinct
##     1000        0        4
##
## Value                high    low normal
```

```
## Frequency      22    302    297    379
## Proportion  0.022  0.302  0.297  0.379
## ---------------------------------------------------------------------------------
## country
##        n  missing distinct    value
##     1000        0        1       UK
##
## Value        UK
## Frequency  1000
## Proportion    1
## ---------------------------------------------------------------------------------
```

```
#summary of the data, for insight in slightly different form
summary(insuredb)
```

```
##    insurance       driver_age      car_value    num_accident   annual_mileage
##  Min.   :150.3   Min.   :17.00   £15,200:  7   Min.   :0.000   Min.   : 1530
##  1st Qu.:224.8   1st Qu.:34.00   £15,700:  7   1st Qu.:0.000   1st Qu.: 8055
##  Median :290.6   Median :47.00   £12,300:  6   Median :1.000   Median :10075
##  Mean   :310.8   Mean   :47.52   £14,000:  6   Mean   :0.799   Mean   :10055
##  3rd Qu.:374.9   3rd Qu.:60.00   £14,700:  6   3rd Qu.:1.000   3rd Qu.:12150
##  Max.   :780.4   Max.   :86.00   £17,400:  6   Max.   :5.000   Max.   :19300
##                                  (Other):962                   NA's   :32
##     car_age         excess          car_reg       gender    alarm
##  7Y 0M   : 17   Min.   :  0.0   AA14NVN:  1   F     :297   No :393
##  4Y 0M   : 16   1st Qu.:100.0   AA15LSH:  1   female:198   Yes:607
##  4Y 7M   : 15   Median :150.0   AB18WSH:  1   M     :303
##  3Y 0M   : 14   Mean   :152.5   AC66EKU:  1   male  :202
##  4Y 11M  : 14   3rd Qu.:200.0   AD18VHJ:  1
##  4Y 4M   : 14   Max.   :500.0   AD19GNG:  1
##  (Other):910                    (Other):994
##  insurance_group  crime_area  country
##  Group 1:394            : 22   UK:1000
##  Group 2:486      high  :302
##  Group 3:120      low   :297
##                   normal:379
##
##
##
```

.

```
# Observing the first 6 rows
head(insuredb)
```

Observing the first 6 rows

```
##   insurance driver_age car_value num_accident annual_mileage car_age excess
## 1    231.69         67   £26,500            0           9000   4Y 2M     50
## 2    310.47         54   £12,000            1          11950   7Y 5M    100
```

```
## 3    506.26         47    £34.5K            1          12900   3Y 5M    200
## 4    174.29         80    £5,100            0           3730   8Y 9M    150
## 5    526.97         54    £9,300            1           9130 10Y 10M     0
## 6    502.09         32    £48.3K            0           8200   1Y 6M    100
##   car_reg gender alarm insurance_group crime_area country
## 1 OO19TCA      M   Yes         Group 2        low      UK
## 2 BZ65SRB   male   Yes         Group 2     normal      UK
## 3 SG69XEX      M   Yes         Group 3     normal      UK
## 4 DZ14WGG      F   Yes         Group 2       high      UK
## 5 FE12JKB female    No         Group 2       high      UK
## 6 NJ71VPX   male   Yes         Group 2     normal      UK
```

```r
min(insuredb$driver_age)
```

```
## [1] 17
```

The dataset comprises 1,000 observations across 13 variables (columns). It comprises of information on insurance premiums, driver demographics, accident rates and car-related details within the United Kingdom.

The insurance premiums range from £150.32 to £780.42, showcasing variability in policy costs. Driver ages span within the range of 17 to 86 years, with the average age being around 47.

'crime_area' contains 22 missing values.

'annual_mileage' also contains 32 missing values identified as NAs

The 'gender' variable is meant to comprise categorical data but is messed up with some inconsistent values.

More insights would be shared on this data with visualization tool, Power BI.

.

Data Cleaning and Preparation

The features 'car_reg' and 'country' would not be suitable as predictor because the columns contain 1,000 unique values which is same as total observations and cannot guarantee suitable scientific prediction.

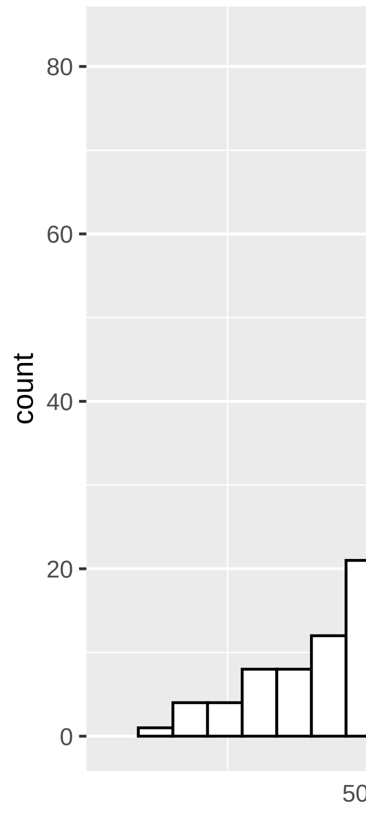
```r
insuredb$car_reg <- NULL #to remove the car_reg column

insuredb$country <- NULL #to remove the country column

summary(insuredb)
```

```
##    insurance       driver_age      car_value    num_accident   annual_mileage
##  Min.   :150.3   Min.   :17.00   £15,200:  7   Min.   :0.000   Min.   : 1530
##  1st Qu.:224.8   1st Qu.:34.00   £15,700:  7   1st Qu.:0.000   1st Qu.: 8055
##  Median :290.6   Median :47.00   £12,300:  6   Median :1.000   Median :10075
##  Mean   :310.8   Mean   :47.52   £14,000:  6   Mean   :0.799   Mean   :10055
##  3rd Qu.:374.9   3rd Qu.:60.00   £14,700:  6   3rd Qu.:1.000   3rd Qu.:12150
##  Max.   :780.4   Max.   :86.00   £17,400:  6   Max.   :5.000   Max.   :19300
##                                  (Other):962                   NA's   :32
##     car_age         excess         gender     alarm     insurance_group
##  7Y 0M   : 17   Min.   :  0.0   F     :297   No :393   Group 1:394
##  4Y 0M   : 16   1st Qu.:100.0   female:198   Yes:607   Group 2:486
```

```
##  4Y 7M  : 15   Median :150.0   M     :303            Group 3:120
##  3Y 0M  : 14   Mean   :152.5   male  :202
##  4Y 11M : 14   3rd Qu.:200.0
##  4Y 4M  : 14   Max.   :500.0
##  (Other):910
##   crime_area
##        : 22
##  high  :302
##  low   :297
##  normal:379
##
##
##
```

.

```r
#To determine best value to fix for the missing values of 'annual_mileage'
ggplot(data = insuredb, aes(x = annual_mileage)) +
  geom_histogram(bins = 30, na.rm = TRUE, color="black", fill="white", position = "stack")
```

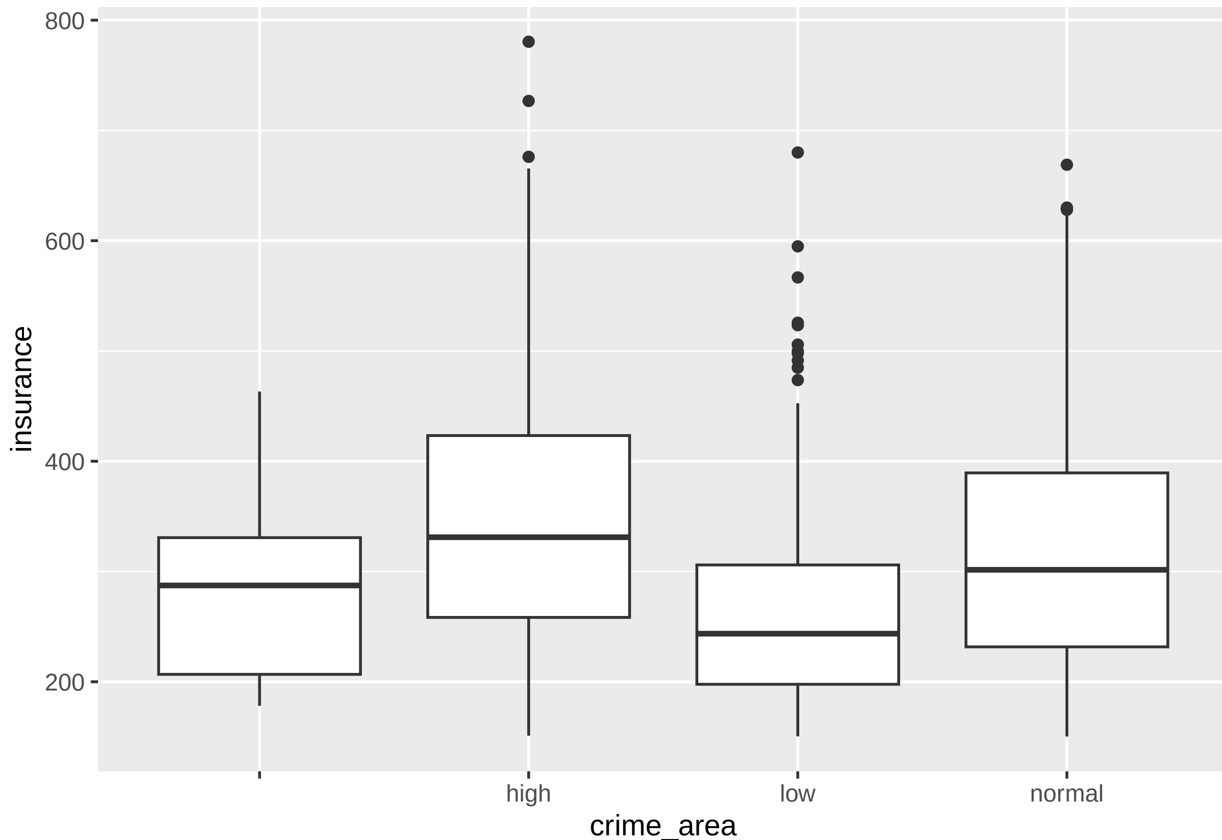


Features 'annual_mileage' and 'crime_area' has NULL values, needs to be fixed

The histogram above is normally symmetric, therefore the mean of the column values is suitable as a replacement for the missing values.

```r
#get the mean of the values
getMean = mean(insuredb$annual_mileage, na.rm = TRUE)

insuredb$annual_mileage[is.na(insuredb$annual_mileage)] <- getMean
#summary(insuredb)
```

```r
#To determine best value to fix for the missing values of 'crime_area'

ggplot(data = insuredb, aes(x = crime_area, y = insurance)) + geom_boxplot()
```




```r
#to replace missing values with 'low' being the median
insuredb$crime_area[(insuredb$crime_area == '')] <- 'low'

insuredb$crime_area = droplevels(insuredb$crime_area)
```

.

```r
#gender column transformation
prepGender <- function(genderStr){
  cSgenderStrtr <- as.character(genderStr)
  if(str_length(genderStr) == 1){
```

```r
    if(genderStr == 'M') genderStr = as.factor('Male')
    if(genderStr == 'F') genderStr = as.factor('Female')
  }else{
    genderStr = str_to_sentence(genderStr)
  }
  genderStr = as.factor(genderStr)
}

insuredb$gender = sapply(insuredb$gender, prepGender)
#summary(insuredb$gender)
```

Transforming the 'gender' column to have consistent values, 'Male' and 'Female'

.

```r
#car_value column transformation to get rid of £ and K from the values
carValue <- function(cStr){
  cStr <- as.character(cStr)
  if(str_detect(cStr, 'K')){
    newVal = str_remove_all(cStr, "[£K]")
    newVal = as.numeric(newVal)
    cStr = newVal * 1000
  }
  if(str_detect(cStr, ",")){
    newVal = str_remove_all(cStr, "[£,]")
    cStr = newVal
  }
  cStr = as.numeric(cStr)
}

insuredb$car_value = sapply(insuredb$car_value, carValue)
#summary(insuredb)
```

Transforming the 'car_value' column to have consistent numerical data

.

```r
#car_age column transformation to get rid of Y(ear) and M(onth)

carAge <- function(cStr){
  cStr <- as.character(cStr)
  if(str_detect(cStr, ' ')){
    spVal = unlist(str_split(cStr, " "))
    yr = str_remove_all(spVal[1], 'Y')
    mth = str_remove_all(spVal[2], 'M')
    nyr = as.numeric(yr)
```

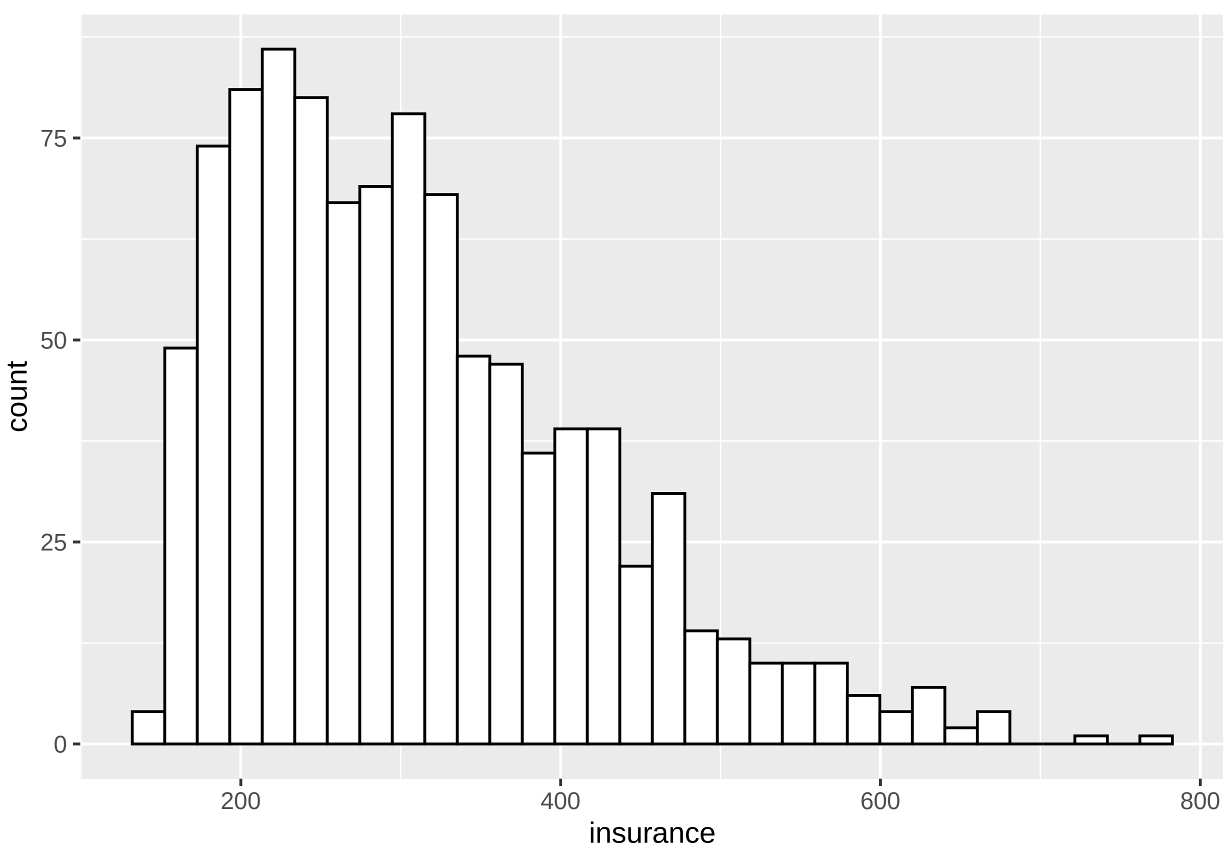
```
    nmth = as.numeric(mth)
    yr = 12*nyr
    cStr = yr + nmth
  }
  cStr = as.numeric(cStr)
}

insuredb$car_age = sapply(insuredb$car_age, carAge)
```

Transforming the 'car_age' column to have consistent numerical data

.

Exploratory Data Analysis

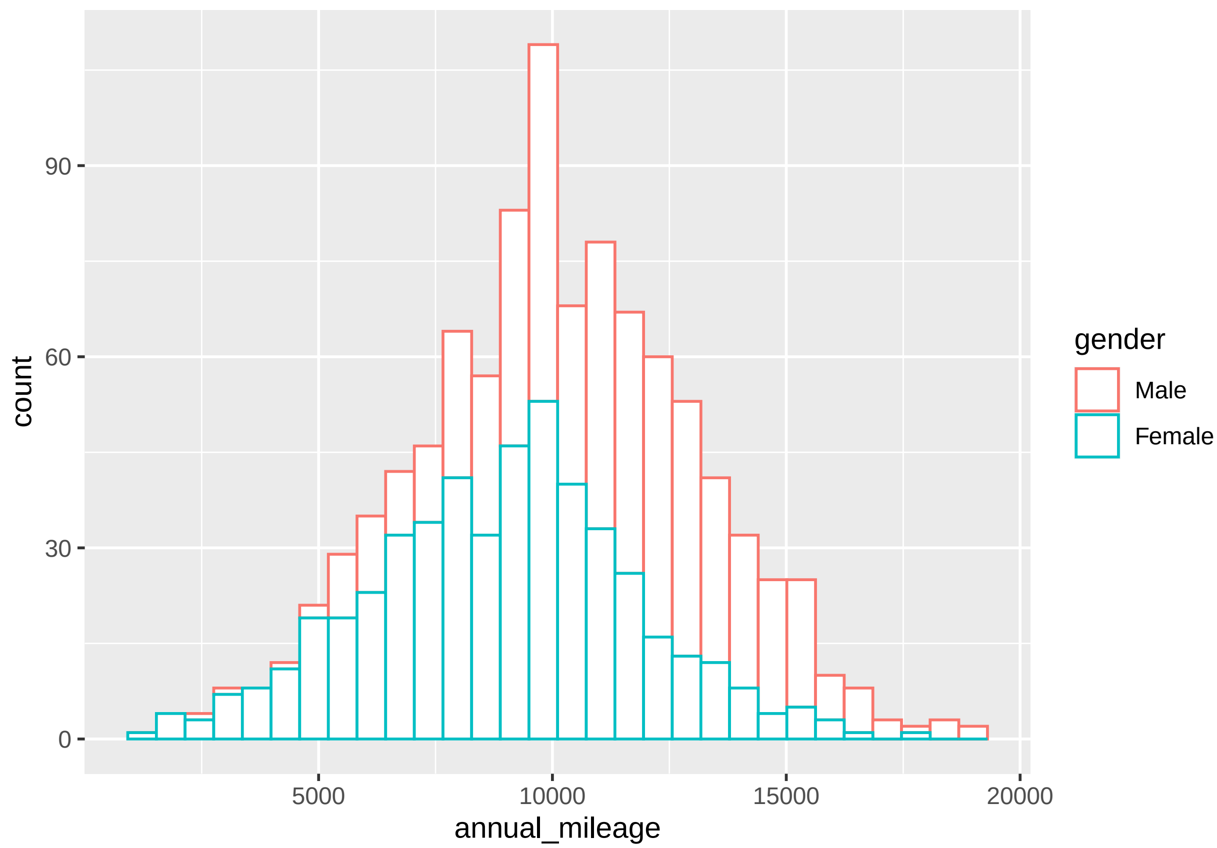
```
ggplot(data = insuredb, aes(x = insurance)) +
  geom_histogram(bins = 32, na.rm = TRUE,color="black", fill="white")
```



The histogram of insurance is skewed(positive) to the right.

Most of the drivers spent less than the average insurance premium of (£310.8) while a few of the drivers spent more, up to (£700+).

```r
ggplot(data = insuredb, aes(x = annual_mileage, color=gender)) + geom_histogram(bins = 30, na.rm = TRUE
```



The histogram of annual_mileage is symmetrical. Shows a perfect balance in the miles traveled by drivers annually.

- Key takeaway: for every person who drives more miles than the average (10,055 miles) annually, there's another person who drives fewer miles than the average.

.

```r
# Calculate the gender with the highest number of accidents
result <- aggregate(annual_mileage ~ gender, data = insuredb, FUN = sum)
#result
male_miles <- result[result$gender == "Male", "annual_mileage"]
female_miles <- result[result$gender == "Female", "annual_mileage"]

cat("Total number miles traveled by males annually is:",
    sprintf("%s", prettyNum(male_miles, big.mark = ",", decimal.mark = ".")), ' miles\n')
```

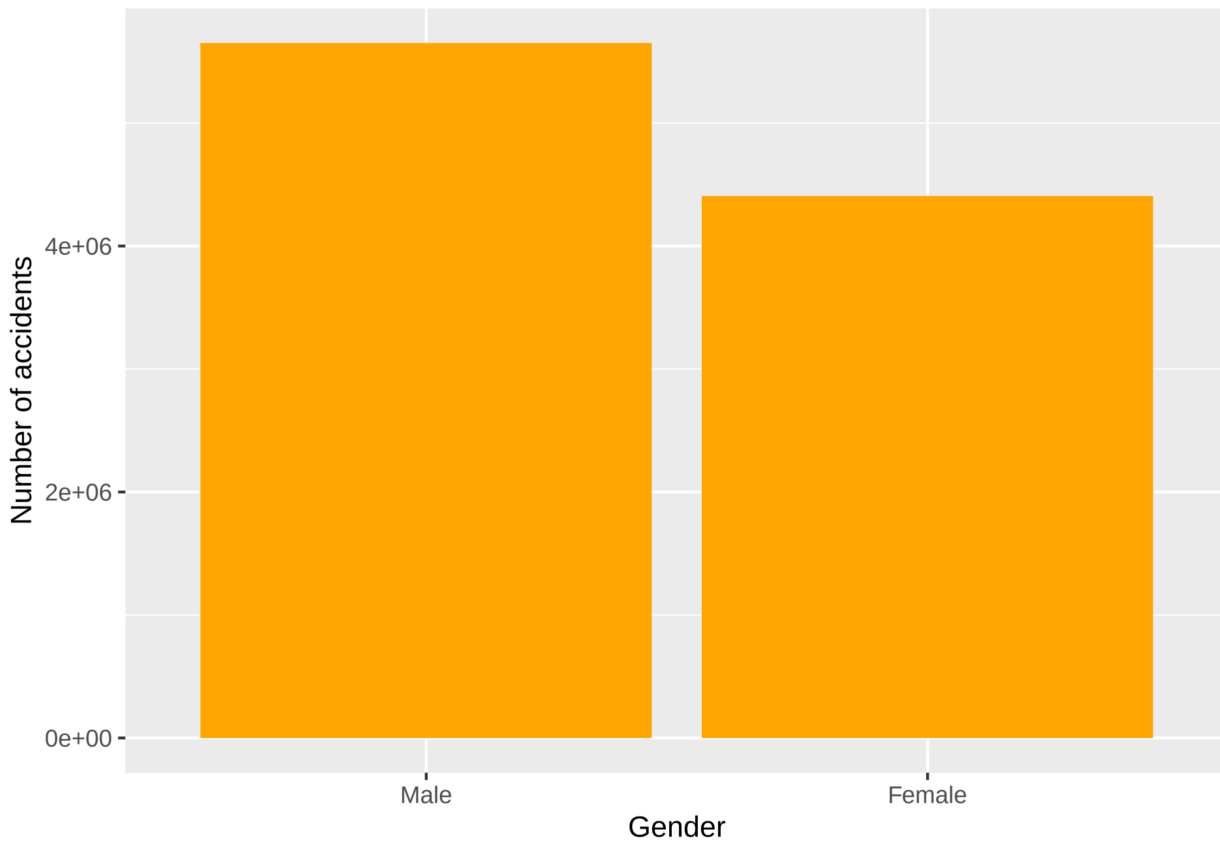
Determining gender with highest annual mileage

```
## Total number miles traveled by males annually is: 5,649,643  miles
```

```r
cat("Total number miles traveled by females annually is:",
    sprintf("%s", prettyNum(female_miles, big.mark = ",", decimal.mark = ".")),'miles')
```

```
## Total number miles traveled by females annually is: 4,405,822 miles
```

```r
ggplot(result, aes(x = gender, y = annual_mileage)) + geom_bar(stat = "identity", fill = "orange") +
  labs(x = "Gender", y = "Number of accidents")
```



.

```r
# Calculate the gender with the highest number of accidents
result <- aggregate(num_accident ~ gender, data = insuredb, FUN = sum)
#result
male_accidents <- result[result$gender == "Male", "num_accident"]
female_accidents <- result[result$gender == "Female", "num_accident"]

cat("Total number accidents by males is:", male_accidents, '\n')
```

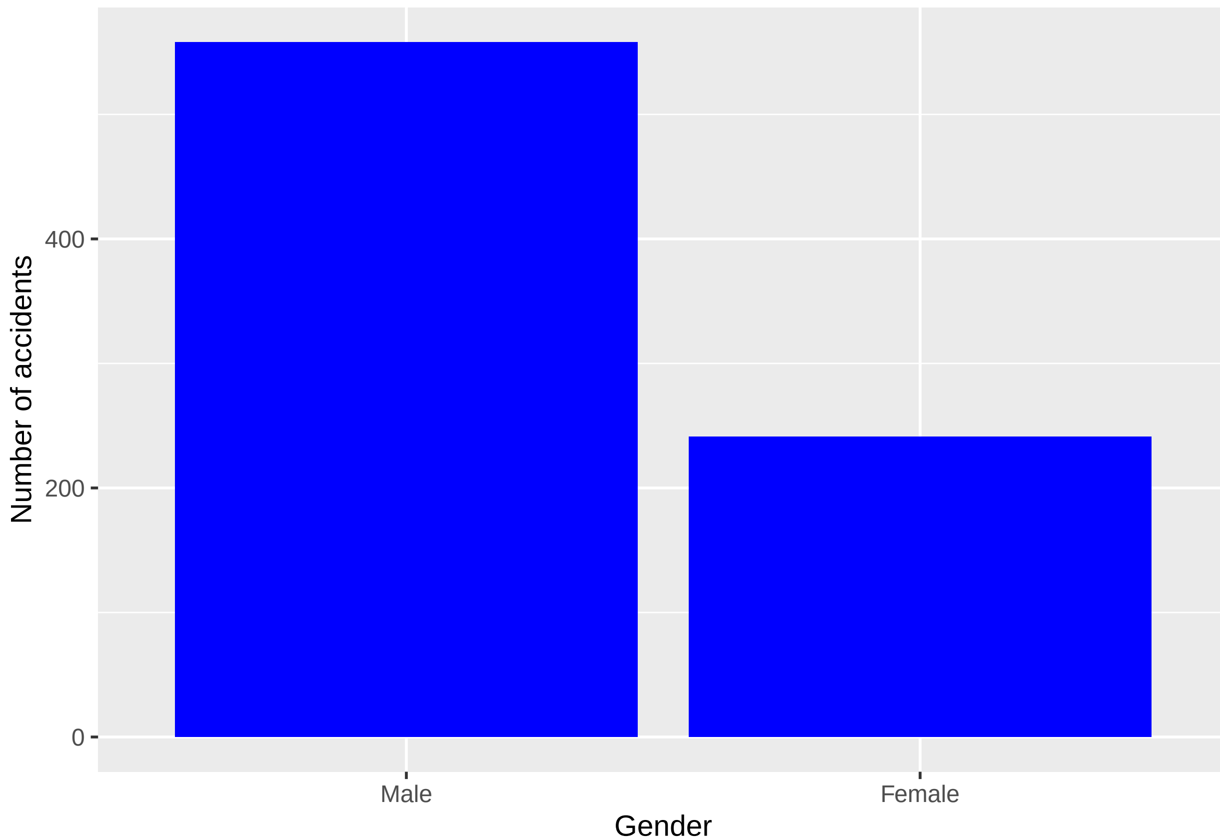
Determining gender with highest number of accidents

```
## Total number accidents by males is: 558
```

```r
cat("Total number accidents by females is:", female_accidents)
```

```
## Total number accidents by females is: 241
```

```r
ggplot(result, aes(x = gender, y = num_accident)) + geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Gender", y = "Number of accidents")
```



.

```r
# Calculate the gender with the highest number of accidents
result <- aggregate(car_value ~ gender, data = insuredb, FUN = sum)
#result
male_cars <- result[result$gender == "Male", "car_value"]
female_cars <- result[result$gender == "Female", "car_value"]

cat("Total males cars value is:", sprintf("£%s", prettyNum(male_cars, big.mark = ",")), '\n')
```

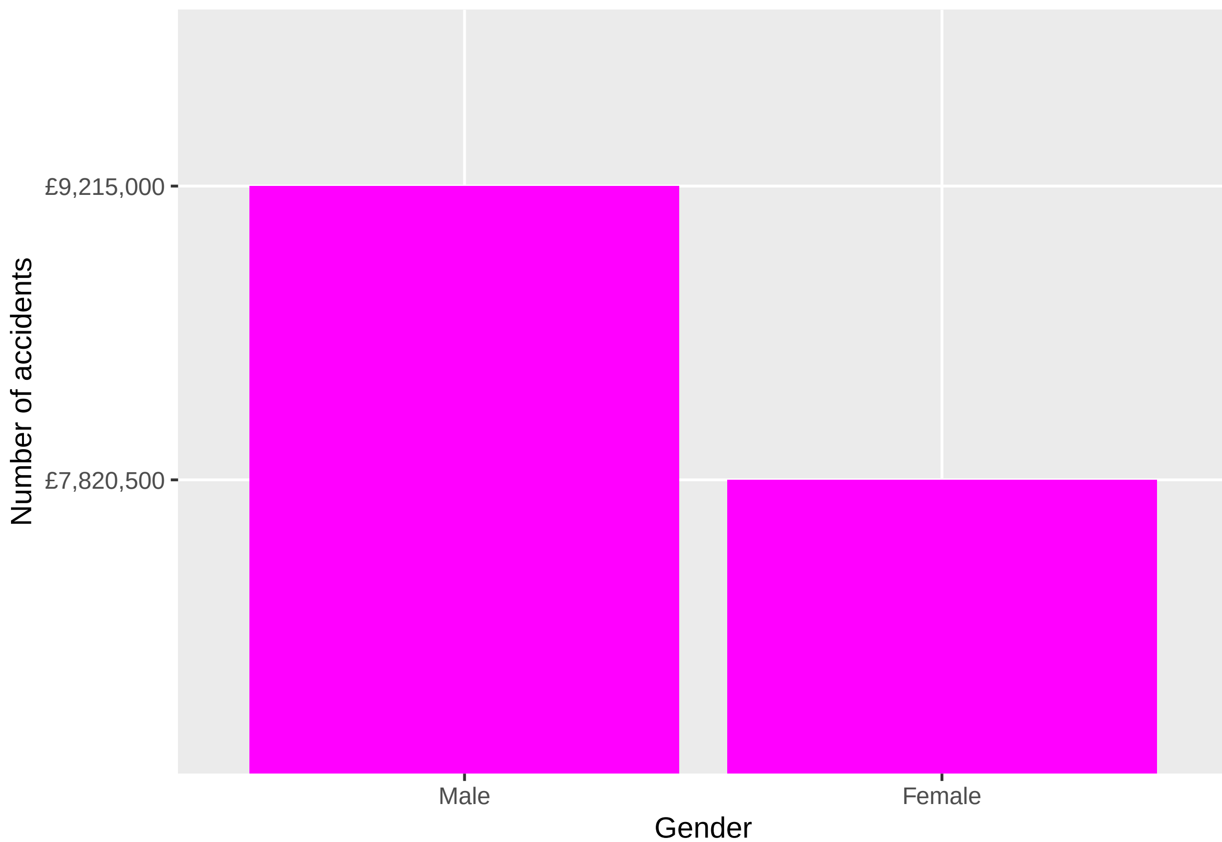
Determining gender with highest cars value (most expensive cars)

```
## Total males cars value is: £9,215,000
```

```r
cat("Total females cars value is:", sprintf("£%s", prettyNum(female_cars, big.mark = ",")))
```

```
## Total females cars value is: £7,820,500
```

```r
ggplot(result, aes(x = gender, y = sprintf("£%s", prettyNum(car_value, big.mark = ",")))) +
  geom_bar(stat = "identity", fill = "magenta") +
  labs(x = "Gender", y = "Number of accidents")
```



.

```r
#result <- insuredb[which.max(insuredb$num_accident), "gender"]
result <- aggregate(insurance ~ insurance_group, data = insuredb, FUN = sum)
#result
group1_insurance <- result[result$insurance_group == "Group 1", "insurance"]
group2_insurance <- result[result$insurance_group == "Group 2", "insurance"]
group3_insurance <- result[result$insurance_group == "Group 3", "insurance"]

g1_2Currency <- sprintf("£%s", prettyNum(group1_insurance, big.mark = ",", decimal.mark = "."))
g2_2Currency <- sprintf("£%s", prettyNum(group2_insurance, big.mark = ",", decimal.mark = "."))
g3_2Currency <- sprintf("£%s", prettyNum(group3_insurance, big.mark = ",", decimal.mark = "."))

cat("Total insurance premium paid by Group 1 is:", g1_2Currency, '\n')
```

Determining insurance group with highest aggregated insurance premiums

```
## Total insurance premium paid by Group 1 is: £111,268.6
```

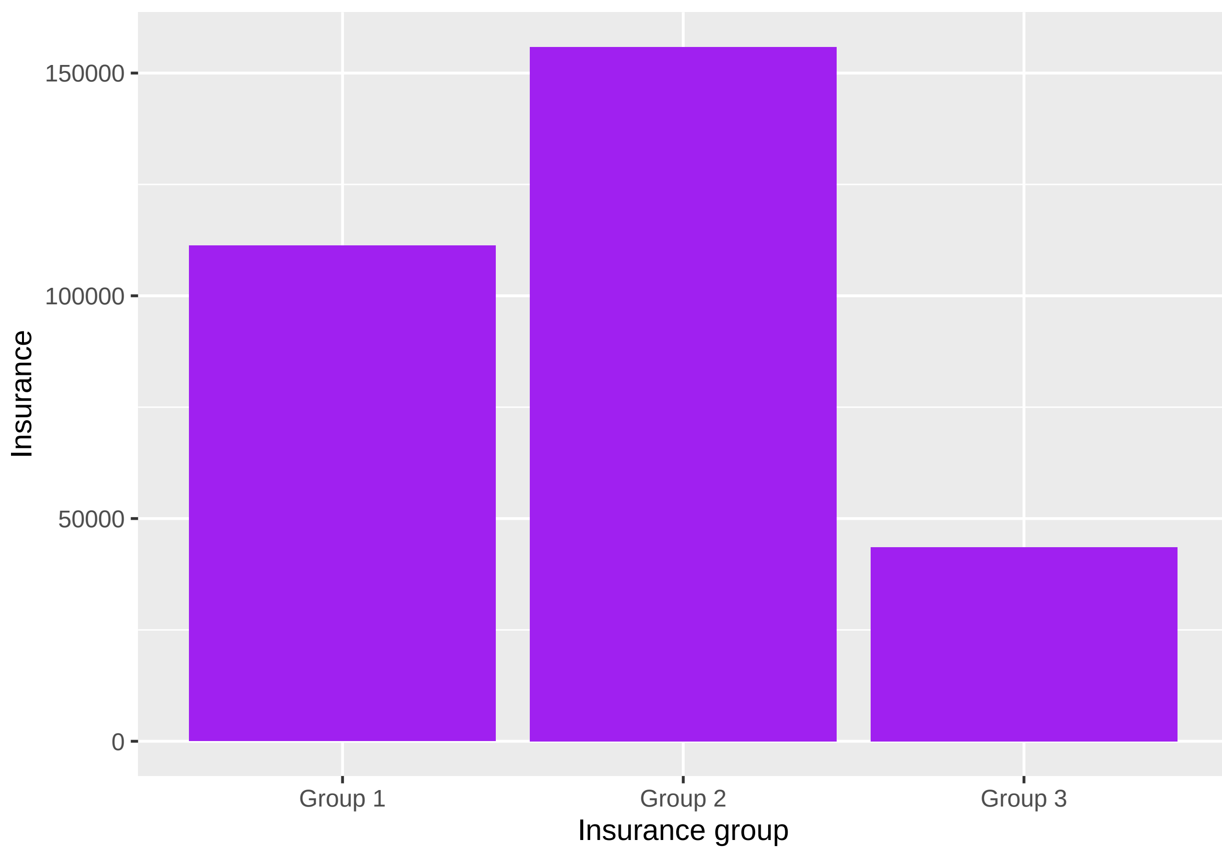
```r
cat("Total insurance premium paid by Group 2 is:", g2_2Currency, '\n')
```

```
## Total insurance premium paid by Group 2 is: £155,881.1
```

```r
cat("Total insurance premium paid by Group 3 is:", g3_2Currency)
```

```
## Total insurance premium paid by Group 3 is: £43,613.35
```

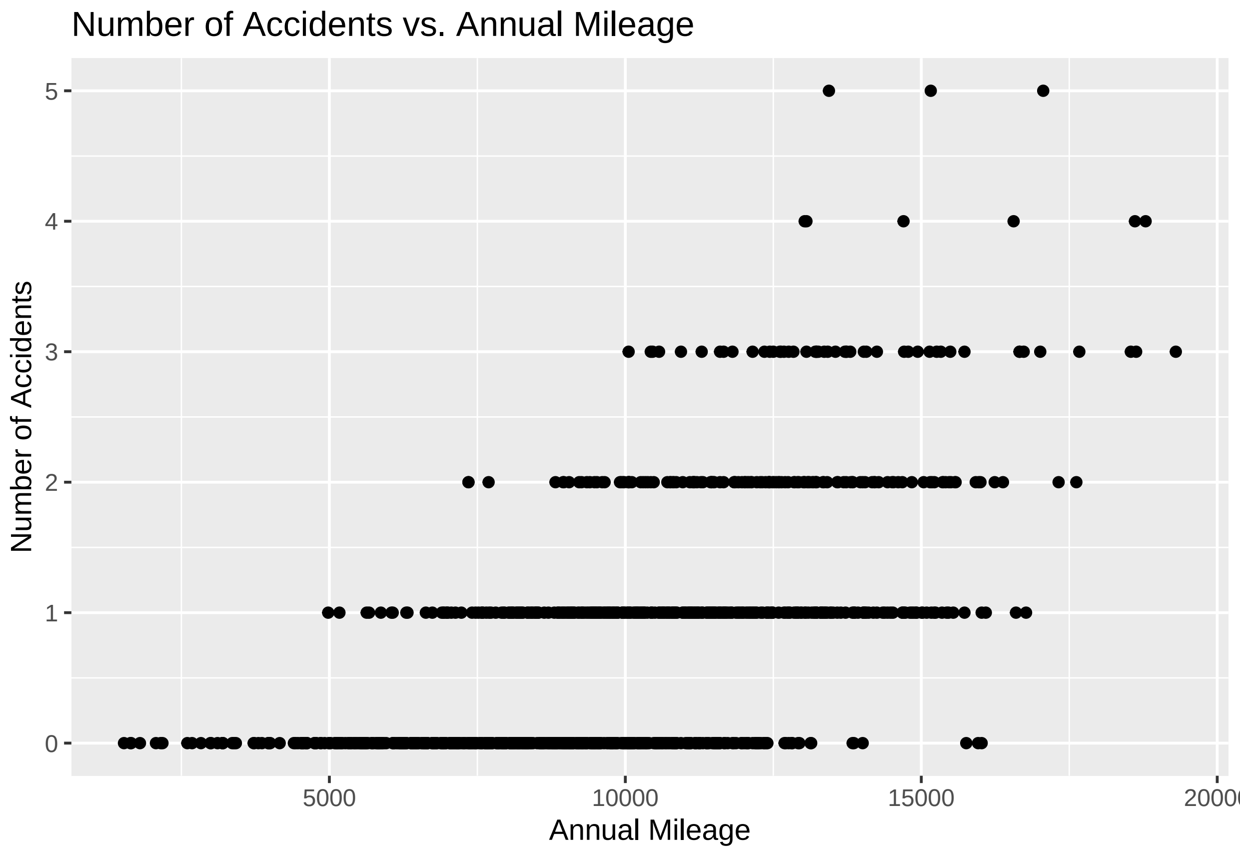
```r
ggplot(result, aes(x = insurance_group, y = insurance)) + geom_bar(stat = "identity", fill = "purple") +
labs(x = "Insurance group", y = "Insurance")
```



.

```r
ggplot(insuredb, aes(x = annual_mileage, y = num_accident)) +
  geom_point() +
  labs(x = "Annual Mileage", y = "Number of Accidents") +
  ggtitle("Number of Accidents vs. Annual Mileage")
```

Determining relationship between number of accident and driver's annual mileage



The plot reveals that as the number of miles driven annually increases, the number of accidents also increases. Therefore, if a driver drives a lot of miles each year, there's a higher chance he/she may be involved in more accidents and vice versa.

- Application: An important insight for insurance or safety considerations.
- Decision support: It is important to consider annual mileage when assessing the risk of accidents in the context of insurance or safety planning.

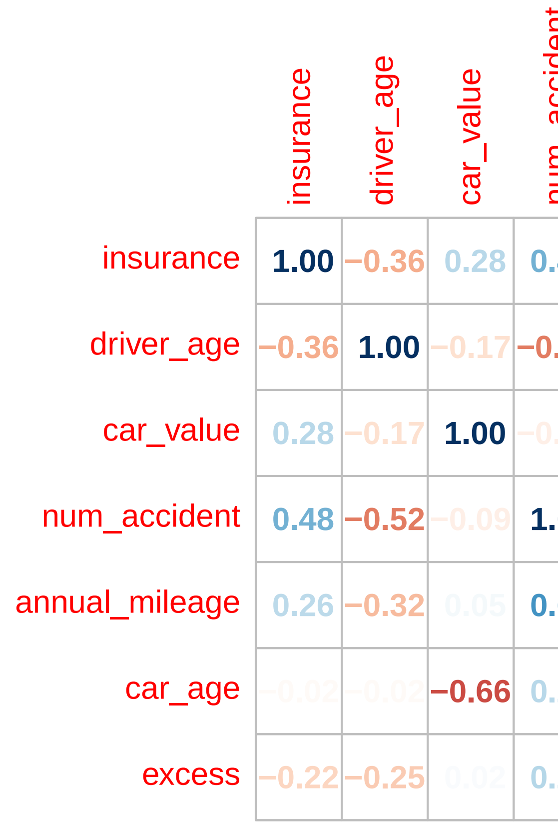
.

```
write.csv(insuredb, file.path(getwd(), "new_insurancedb.csv"), row.names = FALSE)
```

Write the final transformed dataset to current directory for visualization in Power BI

.

```
#corrplot(cor(insuredb[(1:7)]))

corrplot(cor(insuredb[(1:7)]), method = "number")
```



Correlation between insurance and other key variables

.

Conclusion

- Insurance vs. Driver Age: Negative correlation between insurance costs and driver age. As a driver's age increases, insurance costs tend to decrease.
- Insurance vs. Car Value: Positive correlation between insurance costs and the value of the car. Cars with higher value have higher insurance costs.
- Insurance vs. Number of Accidents: Positive correlation between insurance costs and the number of accidents. For drivers involved in more accidents, their insurance costs seem to be higher.
- Insurance vs. Annual Mileage: Positive correlation between insurance costs and the number of miles driven annually. The more miles driven, the higher the insurance costs.
- Insurance vs. Car Age: Weak negative correlation between insurance costs and the age of the car. As the car gets older, insurance costs may slightly decrease.
- The male gender are significantly involved in more accidents than the females, this might be because the males travel more miles than the females annually.
- The male gender drives more expensive cars than the females but with a moderate margin.

.

Future Works

My next work on this would be to:

- build a predictive model
- build a risk assessment model
- create a Power BI visualization