FinalProjectEDA

Eddy D. Varela

2/22/2020

For my final project, I will be analyzing a dataset on car accidents in the United States and risk assesment from an insurance company based on several features. The accident dataset contains all the reported car accidents from Feb 2016 to Dec 2019 and contains many useful features. Features like severity, location (city, county,state), weather information (temperature, humidity, visibility, and precipitation), and civil twilight will be useful in modeling car accidents. Furthermore I will be exploring different factors that contribute differences in car insurance premiums.

Some challenges that I foresee are working with this massive dataset of 1.5M entries and 49 features but I believe it will still be a very useful dataset. I may consider working with smaller samples at first and removig extrenous features. Conversely, I may focus on a particular time period (say the year 2016) and and compare with other years.

```
car_accident_data = read.csv("./data/US_Accidents_Dec19.csv",header = TRUE, nrows = 50000)
head(car_accident_data)
```

```
##    ID   Source TMC Severity          Start_Time            End_Time
## 1 A-1 MapQuest 201        3 2016-02-08 05:46:00 2016-02-08 11:00:00
## 2 A-2 MapQuest 201        2 2016-02-08 06:07:59 2016-02-08 06:37:59
## 3 A-3 MapQuest 201        2 2016-02-08 06:49:27 2016-02-08 07:19:27
## 4 A-4 MapQuest 201        3 2016-02-08 07:23:34 2016-02-08 07:53:34
## 5 A-5 MapQuest 201        2 2016-02-08 07:39:07 2016-02-08 08:09:07
## 6 A-6 MapQuest 201        3 2016-02-08 07:44:26 2016-02-08 08:14:26
##   Start_Lat Start_Lng End_Lat End_Lng Distance.mi.
## 1  39.86515 -84.05872      NA      NA         0.01
## 2  39.92806 -82.83118      NA      NA         0.01
## 3  39.06315 -84.03261      NA      NA         0.01
## 4  39.74775 -84.20558      NA      NA         0.01
## 5  39.62778 -84.18835      NA      NA         0.01
## 6  40.10059 -82.92519      NA      NA         0.01
##                                                                                Description
## 1 Right lane blocked due to accident on I-70 Eastbound at Exit 41 OH-235 State Route 4.
## 2                                  Accident on Brice Rd at Tussing Rd. Expect delays.
## 3           Accident on OH-32 State Route 32 Westbound at Dela Palma Rd. Expect delays.
## 4                    Accident on I-75 Southbound at Exits 52 52B US-35. Expect delays.
## 5             Accident on McEwen Rd at OH-725 Miamisburg Centerville Rd. Expect delays.
## 6    Accident on I-270 Outerbelt Northbound near Exit 29 OH-3 State St. Expect delays.
##   Number                   Street Side         City     County State
## 1     NA                   I-70 E    R       Dayton Montgomery    OH
## 2   2584                 Brice Rd    L Reynoldsburg   Franklin    OH
## 3     NA            State Route 32    R Williamsburg   Clermont    OH
## 4     NA                   I-75 S    R       Dayton Montgomery    OH
## 5     NA Miamisburg Centerville Rd    R       Dayton Montgomery    OH
## 6     NA            Westerville Rd    R  Westerville   Franklin    OH
##      Zipcode Country   Timezone Airport_Code   Weather_Timestamp
## 1      45424      US US/Eastern         KFFO 2016-02-08 05:58:00
## 2 43068-3402      US US/Eastern         KCMH 2016-02-08 05:51:00
## 3      45176      US US/Eastern         KI69 2016-02-08 06:56:00
## 4      45417      US US/Eastern         KDAY 2016-02-08 07:38:00
```

```
## 5      45459      US US/Eastern         KMGY 2016-02-08 07:53:00
## 6      43081      US US/Eastern         KCMH 2016-02-08 07:51:00
##   Temperature.F. Wind_Chill.F. Humidity... Pressure.in. Visibility.mi.
## 1           36.9            NA          91        29.68             10
## 2           37.9            NA         100        29.65             10
## 3           36.0          33.3         100        29.67             10
## 4           35.1          31.0          96        29.64              9
## 5           36.0          33.3          89        29.65              6
## 6           37.9          35.5          97        29.63              7
##   Wind_Direction Wind_Speed.mph. Precipitation.in. Weather_Condition
## 1           Calm              NA              0.02        Light Rain
## 2           Calm              NA              0.00        Light Rain
## 3             SW             3.5                NA          Overcast
## 4             SW             4.6                NA     Mostly Cloudy
## 5             SW             3.5                NA     Mostly Cloudy
## 6            SSW             3.5              0.03        Light Rain
##   Amenity  Bump Crossing Give_Way Junction No_Exit Railway Roundabout
## 1   False False    False    False    False   False   False      False
## 2   False False    False    False    False   False   False      False
## 3   False False    False    False    False   False   False      False
## 4   False False    False    False    False   False   False      False
## 5   False False    False    False    False   False   False      False
## 6   False False    False    False    False   False   False      False
##   Station  Stop Traffic_Calming Traffic_Signal Turning_Loop Sunrise_Sunset
## 1   False False           False          False        False          Night
## 2   False False           False          False        False          Night
## 3   False False           False           True        False          Night
## 4   False False           False          False        False          Night
## 5   False False           False           True        False            Day
## 6   False False           False          False        False            Day
##   Civil_Twilight Nautical_Twilight Astronomical_Twilight
## 1          Night             Night                 Night
## 2          Night             Night                   Day
## 3          Night               Day                   Day
## 4            Day               Day                   Day
## 5            Day               Day                   Day
## 6            Day               Day                   Day
```

```r
# Cleaning data to capture the numerical and binary values
numNameList = c("Precipitation.in.", "Wind_Speed.mph.","Visibility.mi.",
                "Pressure.in.","Humidity...","Temperature.F.","Severity")
locNameList = c("Start_Time", "End_Time", "Start_Lat", "Start_Lang",
                "Street", "City", "County", "Zipcode", "State" )

numIdx= sort(match(numNameList, names(car_accident_data)))
locIdx= sort(match(locNameList, names(car_accident_data)))

# Traffic calming, turning loop, all false, so removing them
tfIdx = c(44, 46:47)

numericalData = car_accident_data[,numIdx]
locData = car_accident_data[,locIdx]
tfData = car_accident_data[,tfIdx]
```

Summary statistic

```
averages = apply(numericalData, 2, mean, na.rm = TRUE)
variances = apply(numericalData,2, var, na.rm = TRUE)
iqrs= apply(numericalData,2,IQR, na.rm = TRUE)
medians = apply(numericalData,2,median, na.rm = TRUE)

# Build a table
dataTable <-data.frame(
  Mean = c(averages["Severity"], averages["Temperature.F."],
           averages["Humidity..."], averages["Pressure.in."],
           averages["Visibility.mi."], averages["Wind_Speed.mph."],
           averages["Precipitation.in."]),

  Variance = c(variances["Severity"], variances["Temperature.F."],
           variances["Humidity..."], variances["Pressure.in."],
           variances["Visibility.mi."], variances["Wind_Speed.mph."],
           variances["Precipitation.in."]),

  Median = c(medians["Severity"], medians["Temperature.F."],
           medians["Humidity..."], medians["Pressure.in."],
           medians["Visibility.mi."], medians["Wind_Speed.mph."],
           medians["Precipitation.in."]),

  IQR = c(iqrs["Severity"], iqrs["Temperature.F."],
           iqrs["Humidity..."], iqrs["Pressure.in."],
           iqrs["Visibility.mi."], iqrs["Wind_Speed.mph."],
           iqrs["Precipitation.in."]))

rownames(dataTable) = c("Severity", "Temp", "Humidity",
                        "Pressure", "Visibility", "Wind Speed", "Precip.")
library(xtable)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")
print(xtable(dataTable), comment=FALSE)
```

	Mean	Variance	Median	IQR
Severity	2.40	0.24	2.00	1.00
Temp	64.98	174.26	64.40	17.50
Humidity	61.72	471.18	63.00	32.00
Pressure	29.98	0.04	29.97	0.17
Visibility	9.38	3.84	10.00	0.00
Wind Speed	8.74	20.03	8.10	5.70
Precip.	0.03	0.00	0.01	0.03

Histogram of TF values

```
mainTitle = c("Traffic Signal", "Sunrise/Sunset", "Civil Twilight")
xlabels = c("Traffic Signal present", "Time of day", "Day/Night")
color = c("red", "blue", "green")

  "Traffic Signal Present?"
```

```
## [1] "Traffic Signal Present?"
```

```
  countTrue = length(which(tfData$Traffic_Signal == "True"))
  countFalse = length(which(tfData$Traffic_Signal == "False"))
```

```
  "True: "
```

```
## [1] "True: "
```

```
  (countTrue / (countFalse+countTrue)) *100
```

```
## [1] 8.216
```

```
  "False: "
```

```
## [1] "False: "
```

```
  (countFalse / (countFalse+countTrue)) *100
```

```
## [1] 91.784
```

```
  "Sunrise or Sunset"
```

```
## [1] "Sunrise or Sunset"
```

```
  dayCount = length(which(tfData$Sunrise_Sunset == "Day"))
  nightCount = length(which(tfData$Sunrise_Sunset == "Night"))
  "Sunrise: "
```

```
## [1] "Sunrise: "
```

```
  (dayCount / (dayCount+nightCount)) *100
```

```
## [1] 65.702
```

```
  "Sunset: "
```

```
## [1] "Sunset: "
```

```
  (nightCount/(dayCount + nightCount)) * 100
```

```
## [1] 34.298
```

```
  "Civil Twilight"
```

```
## [1] "Civil Twilight"
```

```
  day = length(which(tfData$Civil_Twilight == "Day"))
  night = length(which(tfData$Civil == "Night"))
  "Daytime:"
```

```
## [1] "Daytime:"
```

```
  (day / (day+night)) *100
```

```
## [1] 69.562
```

```
  "NightTime:"
```

```
## [1] "NightTime:"
```

```
  (night / (day + night)) * 100
```

```
## [1] 30.438
```

```
# tf_hist = hist(tfData)
```

The car insurance data (from Australia) provides 67k observations with 11 features about a client's 'exposure' to risk. This value goes from 0-1 and we may want to explore if this value is a function of some combination of variables. This dataset contains features like vehicular value, number of claims, the cost of each claim, vehicle information(vehicle age, vehicle body), client information (age and gender). It was suprisingly difficult

to find a US dataset around insurance claims cost so I will use this dataset as a representative proxy to model an insurance pricing strategy.

```
car_insurance_data = read.csv("./data/car.csv", header= TRUE)
head(car_insurance_data)
```

```
##   veh_value  exposure clm numclaims claimcst0 veh_body veh_age gender area
## 1      1.06 0.3039014   0         0         0    HBACK       3      F    C
## 2      1.03 0.6488706   0         0         0    HBACK       2      F    A
## 3      3.26 0.5694730   0         0         0      UTE       2      F    E
## 4      4.14 0.3175907   0         0         0    STNWG       2      F    D
## 5      0.72 0.6488706   0         0         0    HBACK       4      F    C
## 6      2.01 0.8542094   0         0         0    HDTOP       3      M    C
##   agecat            X_OBSTAT_
## 1      2 01101    0    0    0
## 2      4 01101    0    0    0
## 3      2 01101    0    0    0
## 4      2 01101    0    0    0
## 5      2 01101    0    0    0
## 6      4 01101    0    0    0
```

After exploring these two datasets, I would have more information about vehicular accidents in the United States paired with information and formulas to find the risk that some person may posess based on their heuristics.

I found this extra dataset from AllState's insurance claim challenge and I feel like deriving some information from it would be useful for my final project. It contains information about the driver like education level, employment status, how long they have been insured, coverage, marital status, income, location code, months since last claim, claim ammount, reason for claim, vehicle class and size. Therefore these rich features will allow me to get very granular with my analysis.

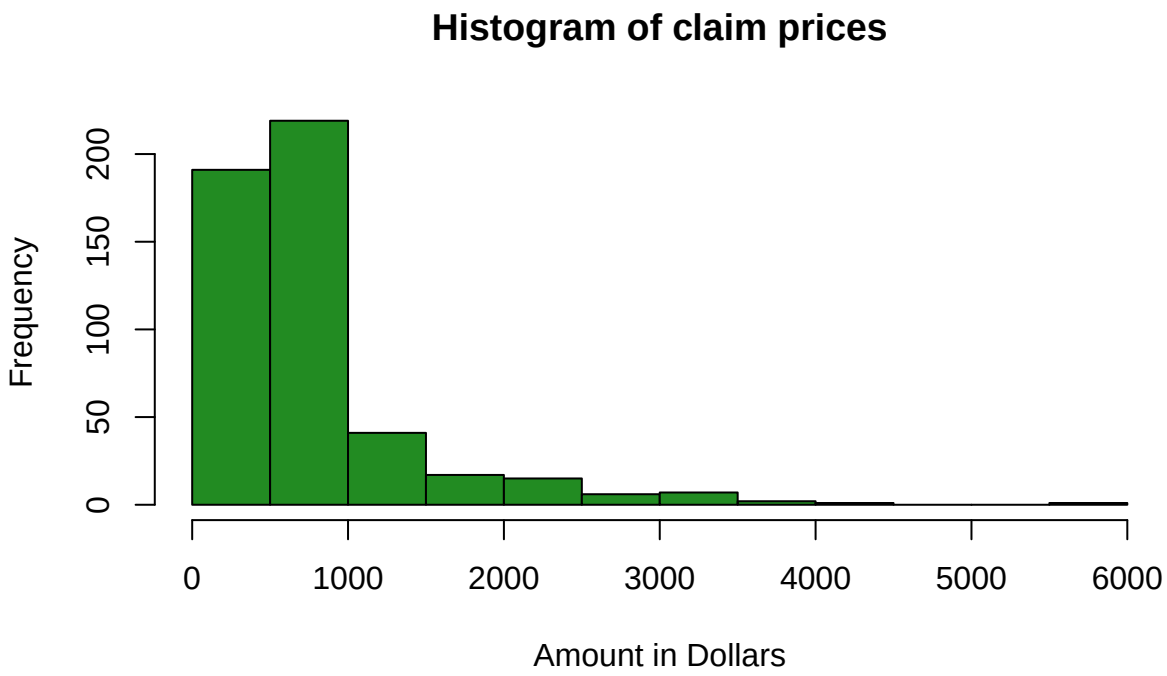
```
claims_data = read.csv('./data/claims.csv',nrows = 500,header = TRUE)
head(claims_data)
```

```
##   Customer Country State.Code    State Claim.Amount Response Coverage
## 1  BU79786      US         KS   Kansas     276.3519       No    Basic
## 2  QZ44356      US         NE Nebraska     697.9536       No Extended
## 3  AI49188      US         OK Oklahoma    1288.7432       No  Premium
## 4  WW63253      US         MO Missouri     764.5862       No    Basic
## 5  HB64268      US         KS   Kansas     281.3693       No    Basic
## 6  OC83172      US         IA     Iowa     825.6298      Yes    Basic
##   Education Effective.To.Date EmploymentStatus Gender Income Location.Code
## 1  Bachelor           2/24/11         Employed      F  56274      Suburban
## 2  Bachelor           1/31/11       Unemployed      F      0      Suburban
## 3  Bachelor           2/19/11         Employed      F  48767      Suburban
## 4  Bachelor           1/20/11       Unemployed      M      0      Suburban
## 5  Bachelor            2/3/11         Employed      M  43836         Rural
## 6  Bachelor           1/25/11         Employed      F  62902         Rural
##   Marital.Status Monthly.Premium.Auto Months.Since.Last.Claim
## 1        Married                   69                      32
## 2         Single                   94                      13
## 3        Married                  108                      18
## 4        Married                  106                      18
## 5         Single                   73                      12
## 6        Married                   69                      14
##   Months.Since.Policy.Inception Number.of.Open.Complaints
```

```
## 1                            5                        0
## 2                           42                        0
## 3                           38                        0
## 4                           65                        0
## 5                           44                        0
## 6                           94                        0
##   Number.of.Policies    Policy.Type       Policy Claim.Reason
## 1                  1 Corporate Auto Corporate L3    Collision
## 2                  8  Personal Auto  Personal L3 Scratch/Dent
## 3                  2  Personal Auto  Personal L3    Collision
## 4                  7 Corporate Auto Corporate L2    Collision
## 5                  1  Personal Auto  Personal L1    Collision
## 6                  2  Personal Auto  Personal L3         Hail
##   Sales.Channel Total.Claim.Amount Vehicle.Class Vehicle.Size
## 1         Agent           384.8111  Two-Door Car      Medsize
## 2         Agent          1131.4649 Four-Door Car      Medsize
## 3         Agent           566.4722  Two-Door Car      Medsize
## 4   Call Center           529.8813           SUV      Medsize
## 5         Agent           138.1309 Four-Door Car      Medsize
## 6           Web           159.3830  Two-Door Car      Medsize
```

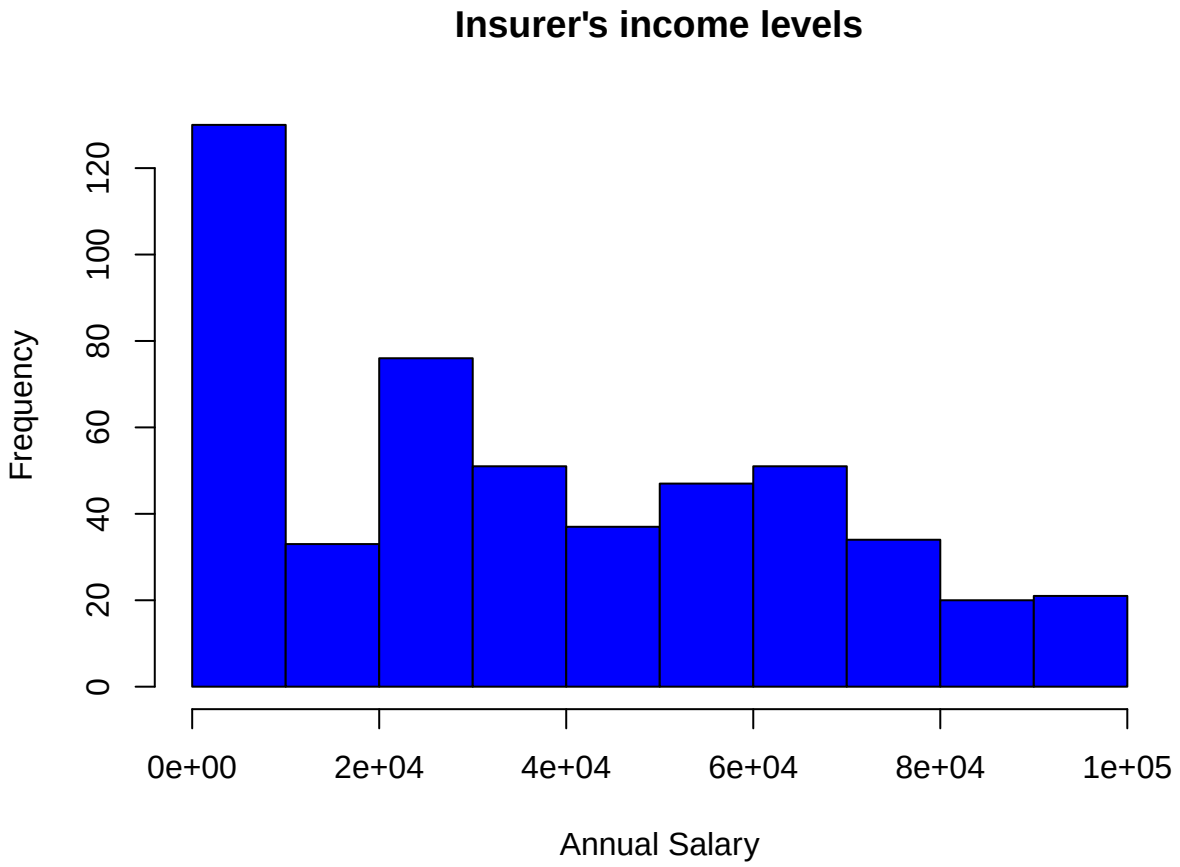
Histogram of claim amounts

```
hist(claims_data$Claim.Amount, xlab = "Amount in Dollars", main="Histogram of claim prices", col = "for
```

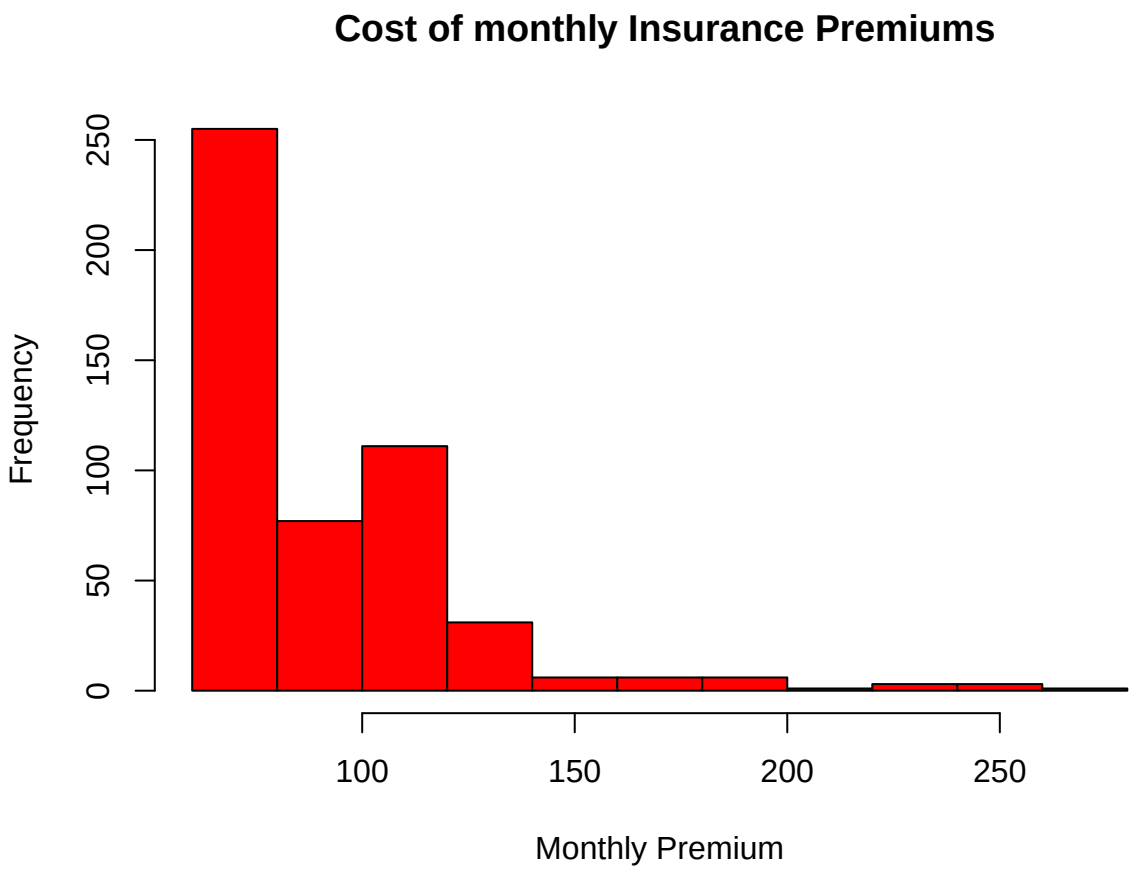


Histogram of insurer's income

```
hist(claims_data$Income, main = "Insurer's income levels", xlab = "Annual Salary" ,col ="blue")
```




```
hist(claims_data$Monthly.Premium.Auto, col = "red", main = " Cost of monthly Insurance Premiums", xlab =
```


Cost of monthly Insurance Premiums
Frequency
0
50
100
150
200
250
100
150
200
250
Monthly Premium