ISOM5610 Project

Team 1

14 December 2018

```
setwd("~/MSBA/ISOM5610/final")
claim <- read.table("Claim.csv", sep = ",", header = TRUE)
summary(claim)
```

```
##     PolicyID          Claim            Exposure            Power
##  Min.   :     1   Min.   :0.00000   Min.   :0.002732   f      :95538
##  1st Qu.:103312   1st Qu.:0.00000   1st Qu.:0.200000   g      :91050
##  Median :206614   Median :0.00000   Median :0.530000   e      :76863
##  Mean   :206596   Mean   :0.03548   Mean   :0.560810   d      :67889
##  3rd Qu.:309867   3rd Qu.:0.00000   3rd Qu.:1.000000   h      :26650
##  Max.   :413169   Max.   :1.00000   Max.   :1.990000   j      :18002
##                                                        (Other):36420
##      CarAge          DriverAge
##  Min.   :  0.000   Min.   :18.00
##  1st Qu.:  3.000   1st Qu.:34.00
##  Median :  7.000   Median :44.00
##  Mean   :  7.533   Mean   :45.32
##  3rd Qu.: 12.000   3rd Qu.:54.00
##  Max.   :100.000   Max.   :99.00
##
##                                  Brand            Gas
##  Fiat                               : 16691   Diesel :205559
##  Japanese (except Nissan) or Korean: 78898   Regular:206853
##  Mercedes, Chrysler or BMW          : 19248
##  Opel, General Motors or Ford       : 37330
##  Renault, Nissan or Citroen         :217822
##  Volkswagen, Audi, Skoda or Seat    : 32575
##  other                              :  9848
##      Region          Density
##  R24    :160392   Min.   :    2
##  R11    : 69603   1st Qu.:   67
##  R53    : 42047   Median :  287
##  R52    : 38675   Mean   : 1983
##  R72    : 31263   3rd Qu.: 1408
##  R31    : 27219   Max.   :27000
##  (Other): 43213
```

```
claim <- claim[-1]
str(claim)
```

```
## 'data.frame':    412412 obs. of  9 variables:
##  $ Claim    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Exposure : num  0.09 0.84 0.52 0.45 0.15 0.75 0.81 0.05 0.76 0.34 ...
##  $ Power    : Factor w/ 12 levels "d","e","f","g",..: 4 4 3 3 4 4 1 1 1 6 ...
##  $ CarAge   : int  0 0 2 2 0 0 1 0 9 0 ...
##  $ DriverAge: int  46 46 38 38 41 41 27 27 23 44 ...
##  $ Brand    : Factor w/ 7 levels "Fiat","Japanese (except Nissan) or Korean",..: 2 2 2 2 2 2 2 2 1 2
##  $ Gas      : Factor w/ 2 levels "Diesel","Regular": 1 1 2 2 1 1 2 2 2 2 ...
```

```
##  $ Region   : Factor w/ 10 levels "R11","R23","R24",..: 9 9 5 5 6 6 9 9 5 1 ...
##  $ Density  : int  76 76 3003 3003 60 60 695 695 7887 27000 ...
```

```
summary(claim$Power)
```

```
##     d     e     f     g     h     i     j     k     l     m     n     o
## 67889 76863 95538 91050 26650 17589 18002  9521  4673  1829  1303  1505
```

```
summary(claim$Gas)
```

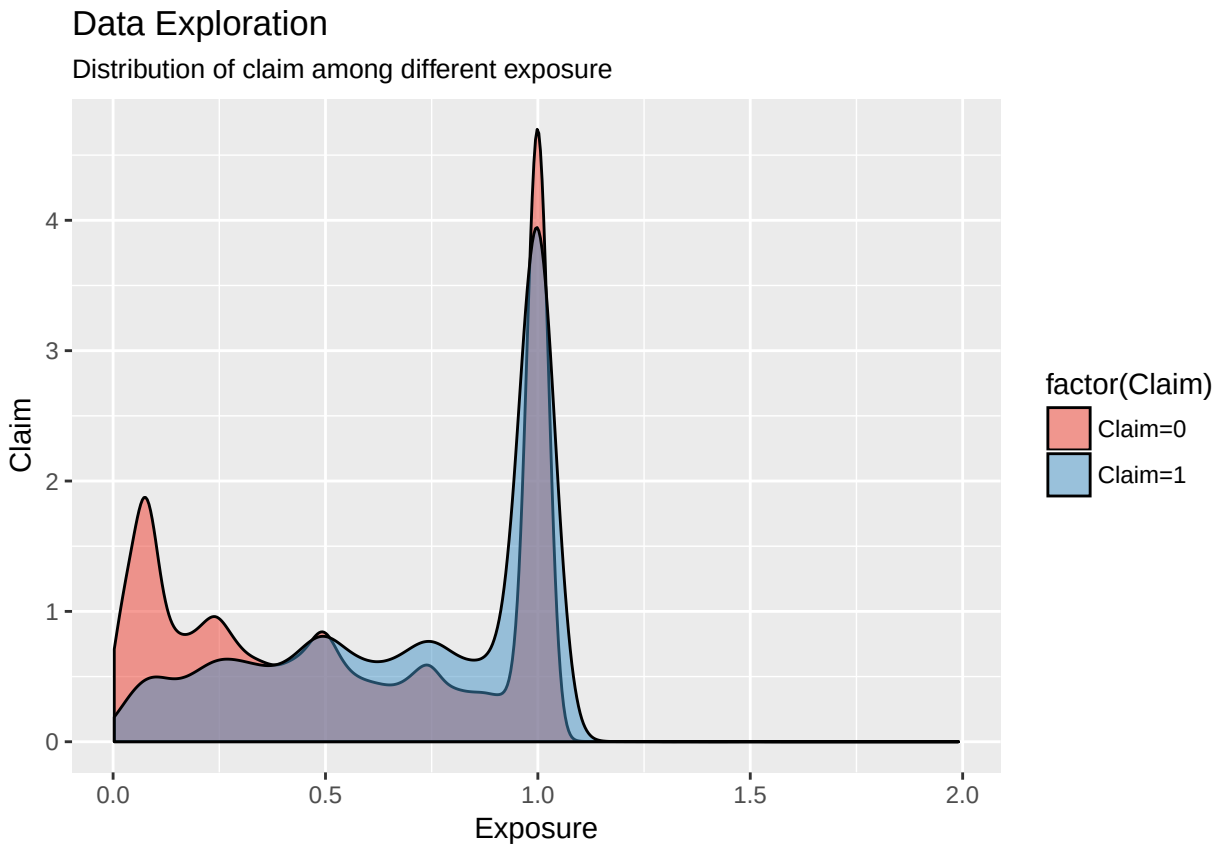
```
##  Diesel Regular
##  205559  206853
```

```
summary(claim$Region)
```

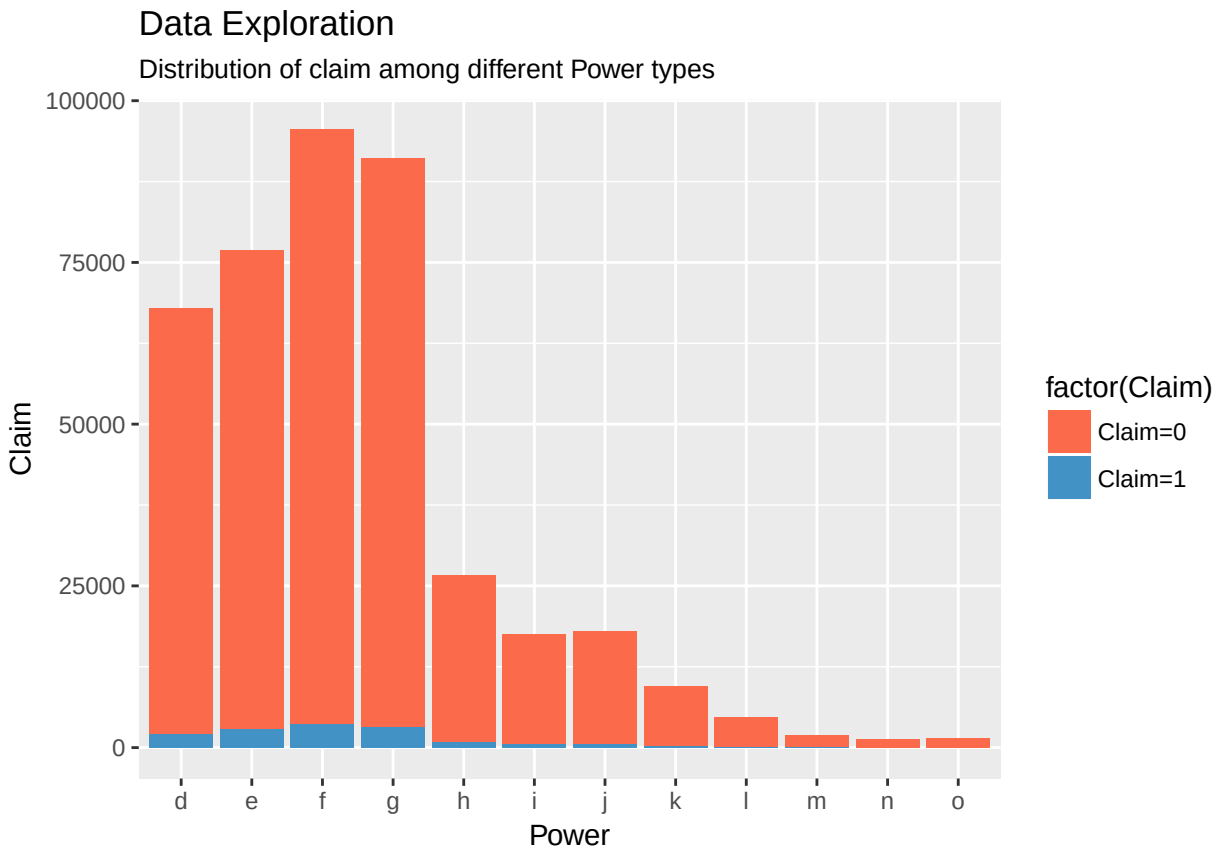
```
##    R11    R23    R24    R25    R31    R52    R53    R54    R72    R74
##  69603   8773 160392  10870  27219  38675  42047  19015  31263   4555
```

There is no missing value. Claim: binary. Power: 12 categories. Brand:7 categories. Gas: binary. Region: 10 regions.

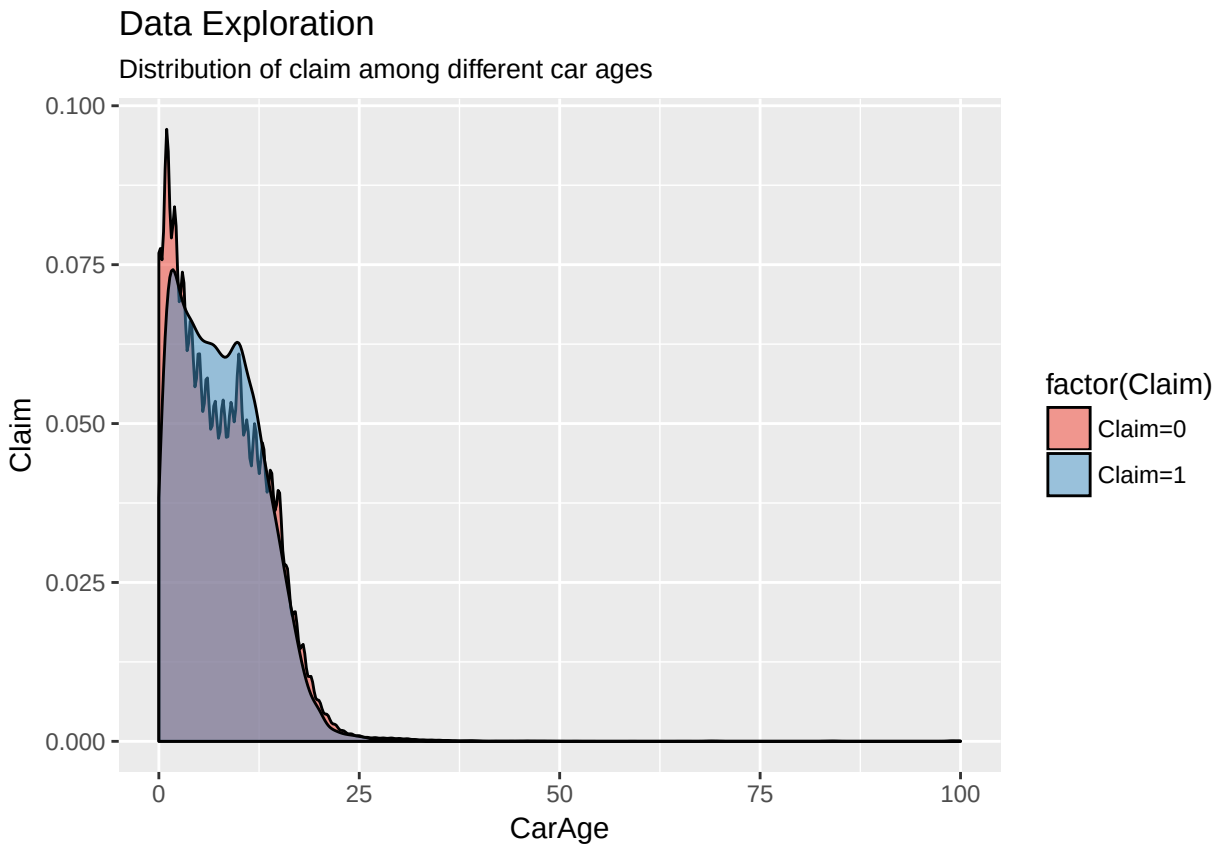
```
library(ggplot2)
library(RColorBrewer)


ds1 <- ggplot(claim, aes(x=Exposure)) +
  geom_density(aes(fill=factor(Claim)),alpha=0.5)+
  labs(title="Data Exploration",
       subtitle="Distribution of claim among different exposure", y="Claim", x="Exposure")+
  scale_fill_manual(values = c(brewer.pal(7, "Reds")[5],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))
plot(ds1)
```



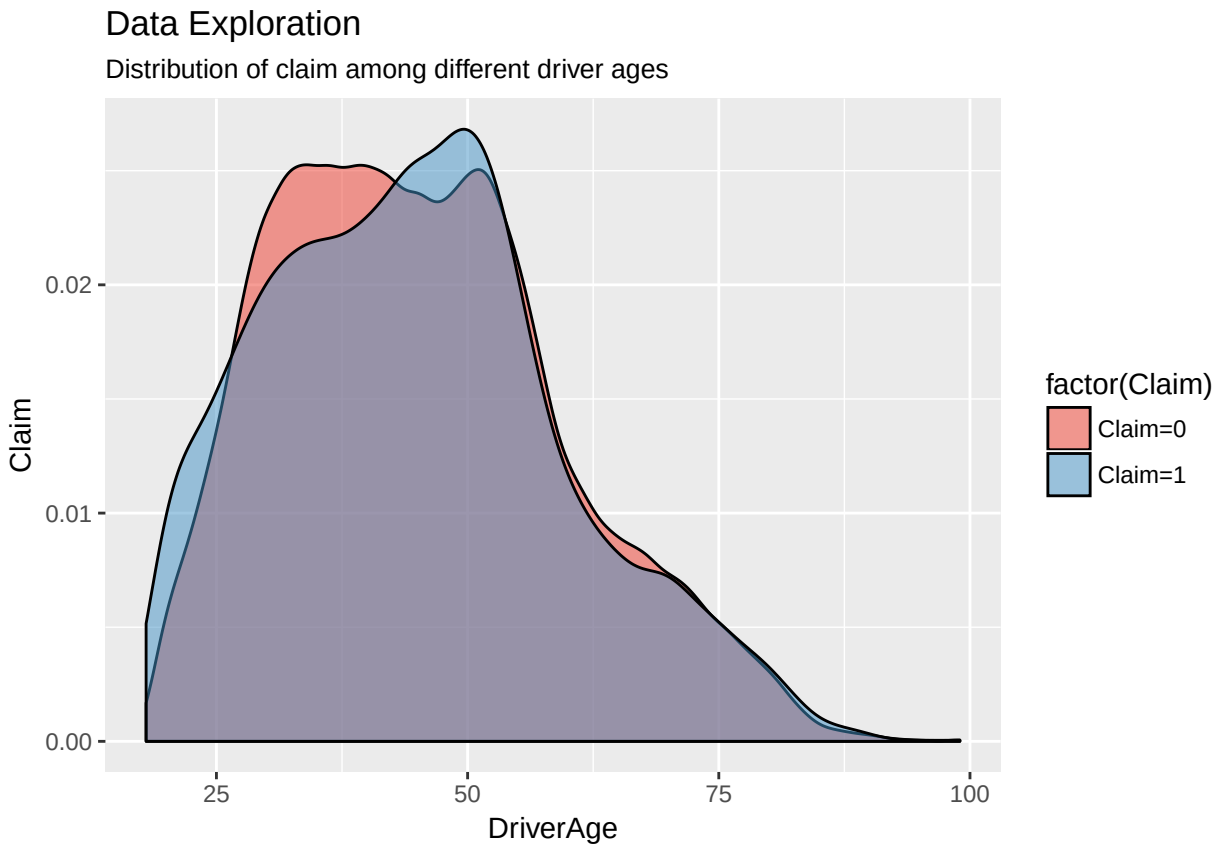

```
bar1 <- ggplot(claim, aes(x = Power)) + geom_bar(aes(fill=factor(Claim)))+
   labs(title="Data Exploration",
        subtitle="Distribution of claim among different Power types", y="Claim", x="Power")+
scale_fill_manual(values = c(brewer.pal(7, "Reds")[4],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))

plot(bar1)
```



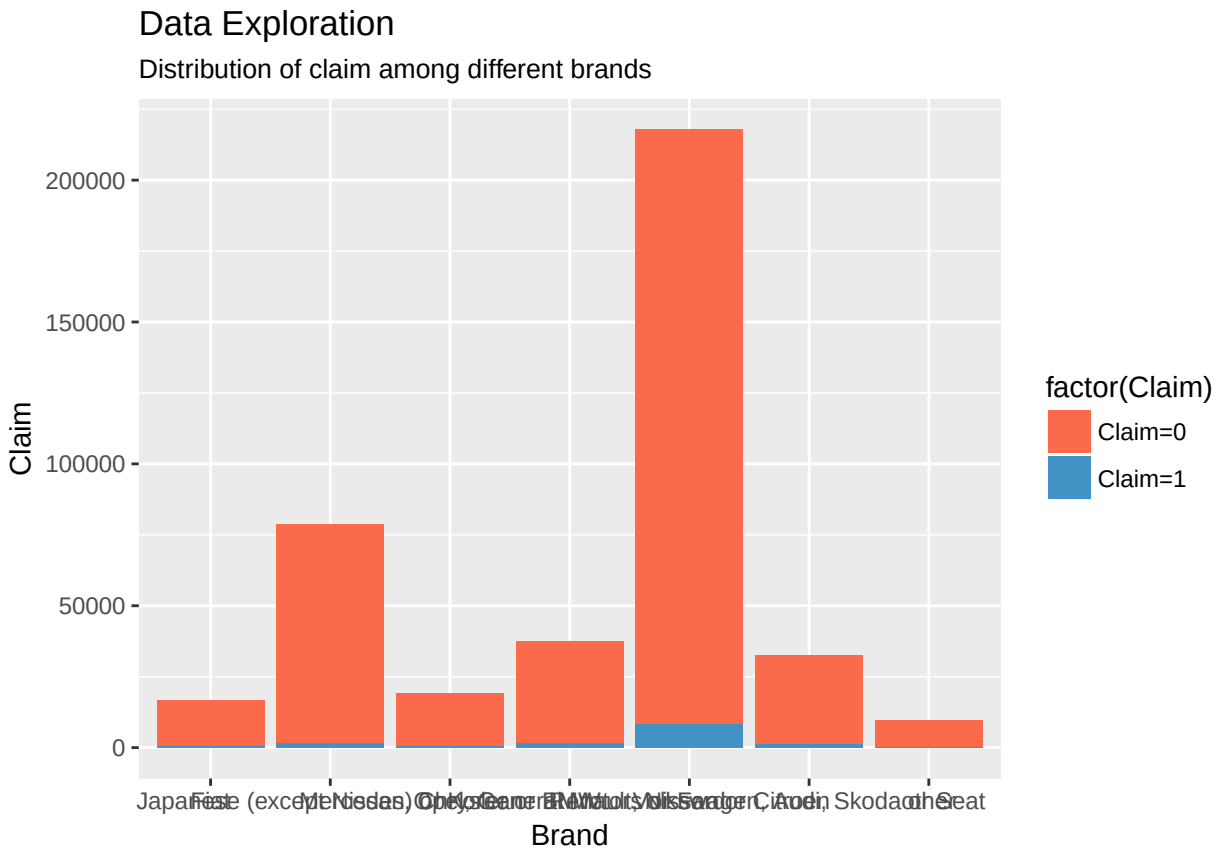

```
ds2 <- ggplot(claim, aes(x=CarAge)) +
  geom_density(aes(fill=factor(Claim)),alpha=0.5)+
  labs(title="Data Exploration",
       subtitle="Distribution of claim among different car ages", y="Claim", x="CarAge")+
  scale_fill_manual(values = c(brewer.pal(7, "Reds")[5],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))
plot(ds2)
```



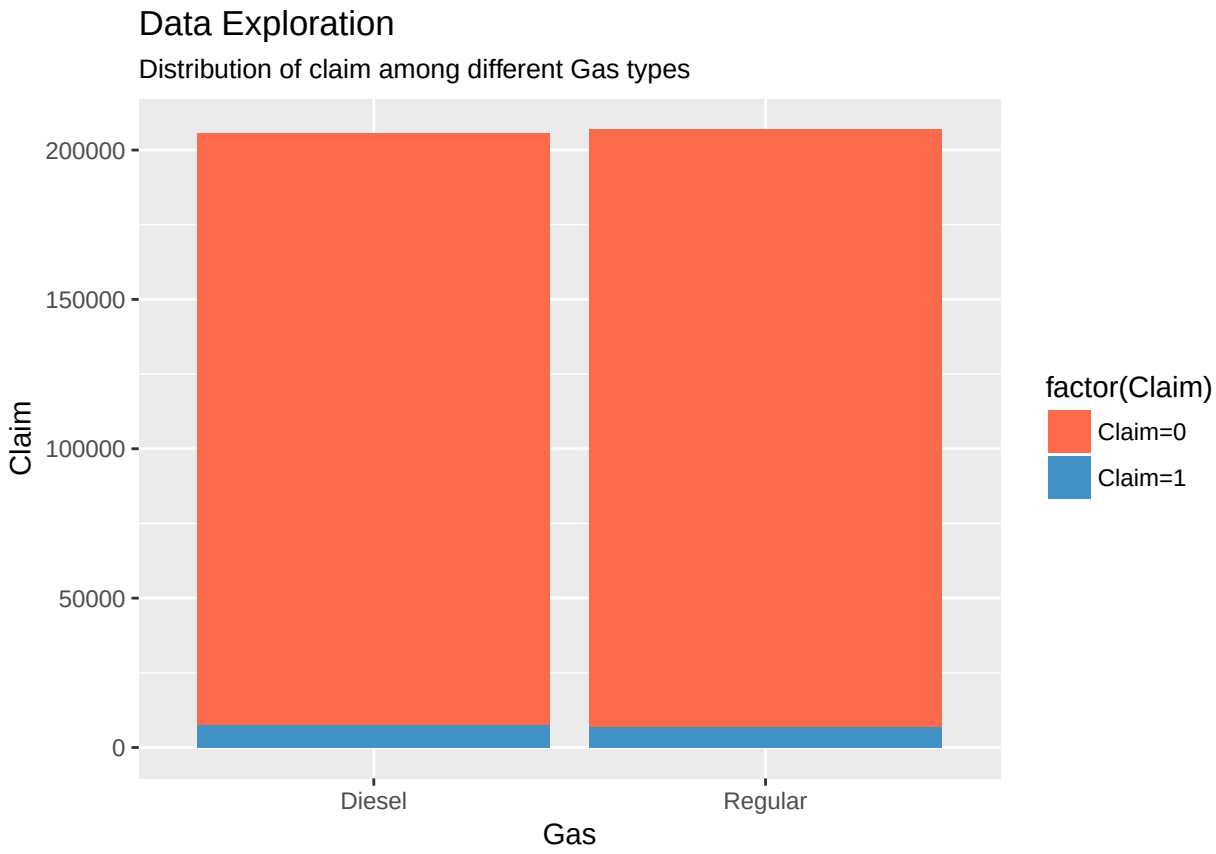

```
ds3 <- ggplot(claim, aes(x=DriverAge)) +
  geom_density(aes(fill=factor(Claim)),alpha=0.5)+
  labs(title="Data Exploration",
       subtitle="Distribution of claim among different driver ages", y="Claim", x="DriverAge")+
  scale_fill_manual(values = c(brewer.pal(7, "Reds")[5],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))
plot(ds3)
```



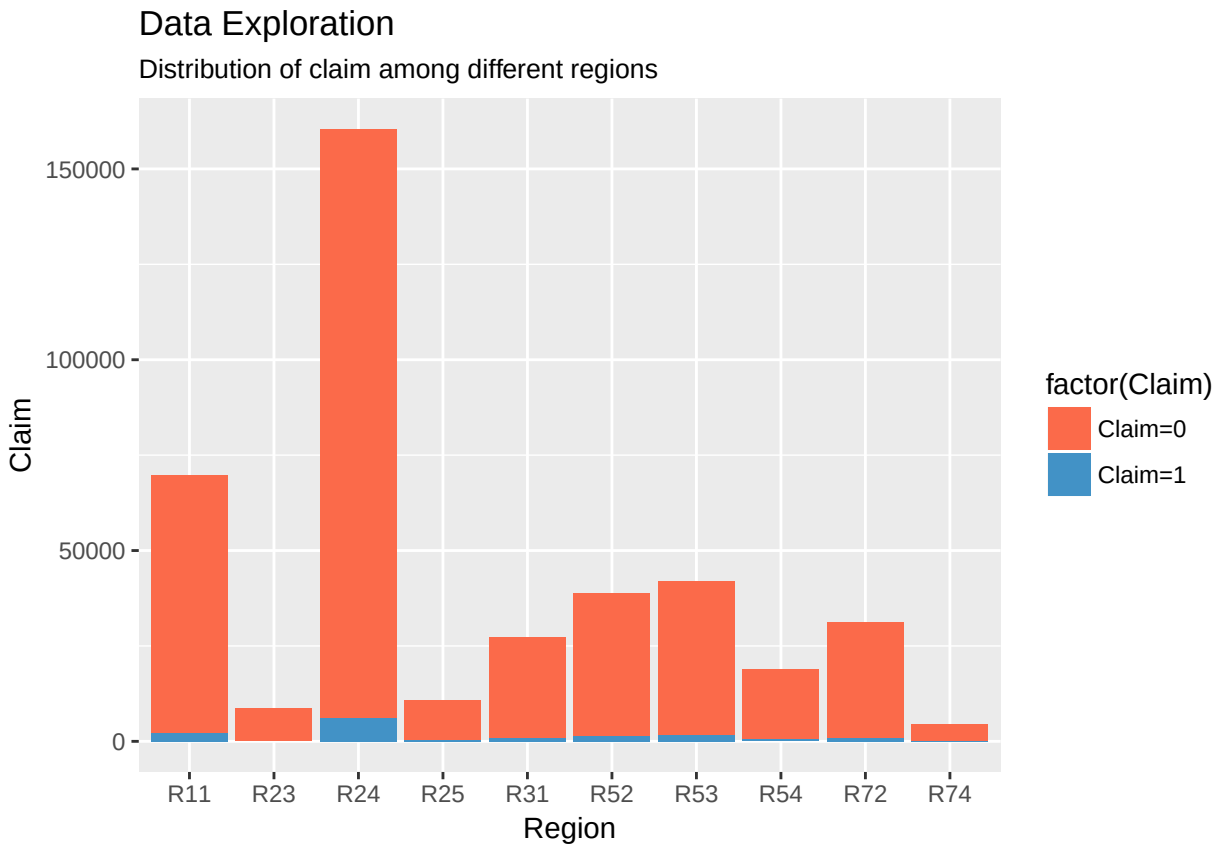

```
bar2 <- ggplot(claim, aes(x = Brand)) + geom_bar(aes(fill=factor(Claim)))+
   labs(title="Data Exploration",
        subtitle="Distribution of claim among different brands", y="Claim", x="Brand")+
scale_fill_manual(values = c(brewer.pal(7, "Reds")[4],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))

plot(bar2)
```



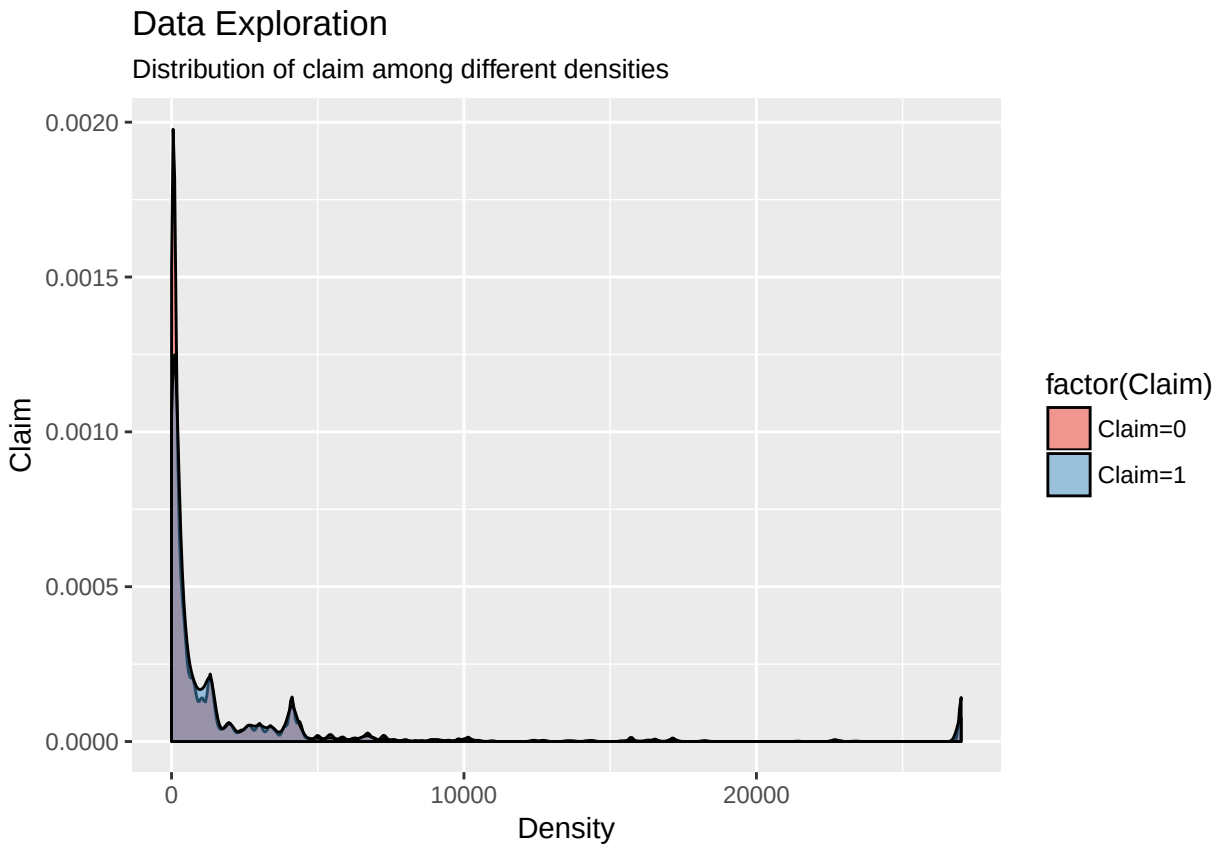

```
bar3 <- ggplot(claim, aes(x = Gas)) + geom_bar(aes(fill=factor(Claim)))+
   labs(title="Data Exploration",
        subtitle="Distribution of claim among different Gas types", y="Claim", x="Gas")+
scale_fill_manual(values = c(brewer.pal(7, "Reds")[4],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))

plot(bar3)
```




```
bar4 <- ggplot(claim, aes(x = Region)) + geom_bar(aes(fill=factor(Claim)))+
   labs(title="Data Exploration",
        subtitle="Distribution of claim among different regions", y="Claim", x="Region")+
scale_fill_manual(values = c(brewer.pal(7, "Reds")[4],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))

plot(bar4)
```




```
ds4 <- ggplot(claim, aes(x=Density)) +
  geom_density(aes(fill=factor(Claim)),alpha=0.5)+
  labs(title="Data Exploration",
       subtitle="Distribution of claim among different densities", y="Claim", x="Density")+
  scale_fill_manual(values = c(brewer.pal(7, "Reds")[5],brewer.pal(7, "Blues")[5]),
                    labels = c("Claim=0", "Claim=1"))
plot(ds4)
```




```
fit <- glm(Claim~., data = claim)
summary(fit)
```

```
##
## Call:
## glm(formula = Claim ~ ., data = claim)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.10920  -0.04872  -0.03397  -0.02103   1.00812
##
## Coefficients:
##                                        Estimate Std. Error t value
## (Intercept)                           2.982e-02  2.085e-03  14.303
## Exposure                              3.907e-02  8.298e-04  47.084
## Powere                                2.462e-03  1.001e-03   2.460
## Powerf                                3.305e-03  9.761e-04   3.386
## Powerg                                2.046e-03  9.536e-04   2.146
## Powerh                                3.086e-03  1.373e-03   2.248
## Poweri                                6.865e-03  1.576e-03   4.356
## Powerj                                5.997e-03  1.576e-03   3.806
## Powerk                                7.821e-03  2.042e-03   3.830
## Powerl                                3.648e-03  2.851e-03   1.280
## Powerm                                5.321e-03  4.466e-03   1.192
## Powern                                5.339e-03  5.200e-03   1.027
## Powero                                6.590e-03  4.825e-03   1.366
## CarAge                               -3.767e-04  5.655e-05  -6.662
```

```
## DriverAge                                -2.368e-04  2.079e-05 -11.391
## BrandJapanese (except Nissan) or Korean -1.366e-02  1.642e-03  -8.320
## BrandMercedes, Chrysler or BMW           3.814e-05  2.024e-03   0.019
## BrandOpel, General Motors or Ford        2.733e-03  1.728e-03   1.582
## BrandRenault, Nissan or Citroen         -2.342e-03  1.491e-03  -1.570
## BrandVolkswagen, Audi, Skoda or Seat     8.929e-04  1.767e-03   0.505
## Brandother                               -2.123e-03  2.356e-03  -0.901
## GasRegular                               -3.028e-03  6.273e-04  -4.827
## RegionR23                                -6.620e-03  2.158e-03  -3.067
## RegionR24                                -2.411e-03  1.097e-03  -2.199
## RegionR25                                -9.430e-04  2.001e-03  -0.471
## RegionR31                                -1.834e-03  1.419e-03  -1.293
## RegionR52                                -1.318e-04  1.319e-03  -0.100
## RegionR53                                -1.562e-04  1.325e-03  -0.118
## RegionR54                                 1.348e-03  1.644e-03   0.820
## RegionR72                                -1.965e-03  1.368e-03  -1.436
## RegionR74                                 4.943e-03  2.883e-03   1.714
## Density                                   4.999e-07  7.425e-08   6.732
##                                         Pr(>|t|)
## (Intercept)                              < 2e-16 ***
## Exposure                                 < 2e-16 ***
## Powere                                  0.013875 *
## Powerf                                  0.000709 ***
## Powerg                                  0.031901 *
## Powerh                                  0.024568 *
## Poweri                                  1.32e-05 ***
## Powerj                                  0.000141 ***
## Powerk                                  0.000128 ***
## Powerl                                  0.200700
## Powerm                                  0.233396
## Powern                                  0.304583
## Powero                                  0.172027
## CarAge                                  2.71e-11 ***
## DriverAge                                < 2e-16 ***
## BrandJapanese (except Nissan) or Korean  < 2e-16 ***
## BrandMercedes, Chrysler or BMW          0.984966
## BrandOpel, General Motors or Ford       0.113720
## BrandRenault, Nissan or Citroen         0.116351
## BrandVolkswagen, Audi, Skoda or Seat    0.613278
## Brandother                              0.367702
## GasRegular                              1.38e-06 ***
## RegionR23                               0.002162 **
## RegionR24                               0.027906 *
## RegionR25                               0.637377
## RegionR31                               0.196095
## RegionR52                               0.920454
## RegionR53                               0.906168
## RegionR54                               0.412398
## RegionR72                               0.150904
## RegionR74                               0.086474 .
## Density                                 1.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.03398005)
##
##     Null deviance: 14114  on 412411  degrees of freedom
## Residual deviance: 14013  on 412380  degrees of freedom
## AIC: -224361
##
## Number of Fisher Scoring iterations: 2
```

```
library(ggplot2)
```