Health_assignment_2

Mathijs Lenderink

2022-11-06

Assignment 2: Risk Adjustment

Reading in libraries and data

```
library(tidyverse)
library(ggplot2); theme_set(theme_bw())
library(patchwork)
library(mlogit)
```

```
data <- read.csv('data_assignment2.csv', sep = ',')
```

Exploratory data analysis

```
head(data)
```

```
##       ID Gender Age_category   Insurer Order_age Income_source Limited_coverage
## 1  20824   Male        [0,5] Insurer A         1         Child                0
## 2  49573   Male        [0,5] Insurer A         1         Child                0
## 3  71451   Male        [0,5] Insurer B         1         Child                0
## 4  76844   Male        [0,5] Insurer A         1         Child                0
## 5 179479   Male        [0,5] Insurer D         1         Child                0
## 6 304970   Male        [0,5] Insurer A         1         Child                0
##   Unhealthy_region Healthcare_cost Population_density
## 1                0               0                  3
## 2                1               0                  4
## 3                0               0                  4
## 4                0               0                  1
## 5                1               0                  3
## 6                0               0                  2
```

Before I summarize the data I first set the categorical variables to categorical data type.

```
categorical_cols <- c("Gender", "Age_category", "Insurer", "Income_source")
data[categorical_cols] <- lapply(data[categorical_cols], factor)
summary(data)
```

```
##        ID             Gender        Age_category         Insurer
##  Min.   :      1   Female:519359   (35,40]:105892   Insurer A:298515
##  1st Qu.: 250007   Male  :476949   (40,45]:104025   Insurer B:249245
##  Median : 500016                   (30,35]: 99326   Insurer C:229069
##  Mean   : 500009                   (45,50]: 95418   Insurer D:169819
##  3rd Qu.: 750011                   (25,30]: 86222   Insurer E: 49660
##  Max.   :1000000                   (50,55]: 82777
##                                    (Other):422648
##    Order_age                    Income_source    Limited_coverage
##  Min.   : 1.000   Child                : 63984   Min.   :0.00000
##  1st Qu.: 7.000   Pension              :139614   1st Qu.:0.00000
##  Median : 9.000   Student              : 34282   Median :0.00000
##  Mean   : 9.451   Unemployment Benefits: 36790   Mean   :0.07059
##  3rd Qu.:12.000   Working              :721638   3rd Qu.:0.00000
##  Max.   :24.000                                  Max.   :1.00000
##
##  Unhealthy_region Healthcare_cost Population_density
##  Min.   :0.0000   Min.   :    0   Min.   :1
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:2
##  Median :0.0000   Median : 9696   Median :3
##  Mean   :0.1495   Mean   : 8145   Mean   :3
##  3rd Qu.:0.0000   3rd Qu.:12451   3rd Qu.:4
##  Max.   :1.0000   Max.   :27030   Max.   :5
##
```

```
dim(data)
```

```
## [1] 996308     10
```

The data summary shows that we have data of 996.308 people of which we know:

- ID: id of person
- Gender: gender of person (male/female)
- Age_category: in which age category the person falls, see below a summary of age categories and distribution of age.
- Order_age: the age_category ordered from low to high
- Insurer: Which insurer the person has
- Income_source: the source of income
- Limited_coverage: whether the person has limited coverage (yes/no)
- Unhealthy_region: whether the person lives in an unhealthy region (yes/no)
- Healthcare_cost: The healthcare cost
- Population_density: how densely populated the area where the person lives is measured on a scale of 1 to 5

To check whether there are missing values:

```
sapply(data, function(x) sum(is.na(x)))
```

```
##                 ID             Gender       Age_category            Insurer
##                  0                  0                  0                  0
##          Order_age      Income_source   Limited_coverage   Unhealthy_region
##                  0                  0                  0                  0
##    Healthcare_cost Population_density
##                  0                  0
```

There are no missing values.

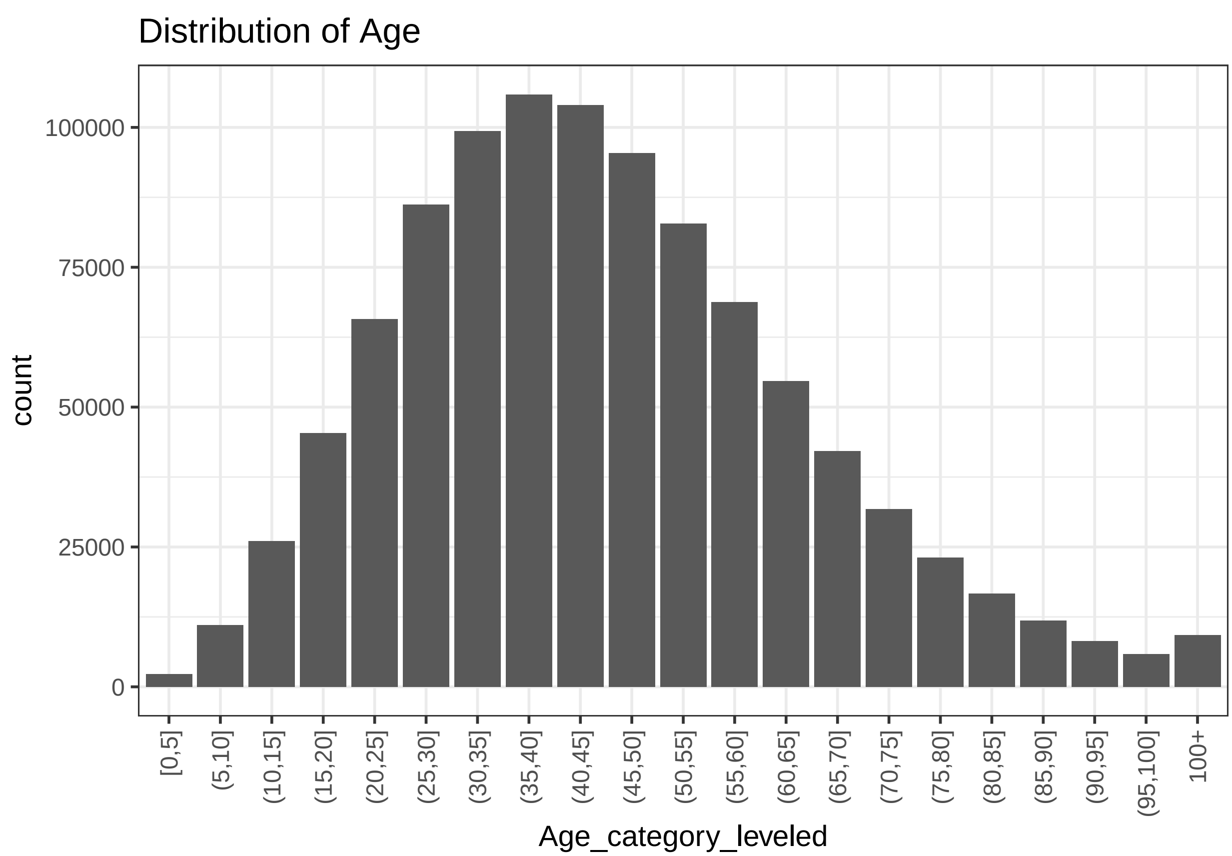
For visualization and clarity purposes i set the age_category levels to increasing categories starting from category [0,5]

```
age_levels <- c( "[0,5]", "(5,10]","(10,15]", "(15,20]", "(20,25]", "(25,30]", "(30,35]", "(35,40]",
"(40,45]", "(45,50]",  "(50,55]", "(55,60]", "(60,65]",
"(65,70]", "(70,75]", "(75,80]", "(80,85]", "(85,90]", "(90,95]",
"(95,100]", "100+")
data$Age_category_leveled <- factor(data$Age_category, levels = age_levels)
```

Data visualisation

Basic graphs

```
ggplot(data = data, aes( x = Age_category_leveled))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Age")
```



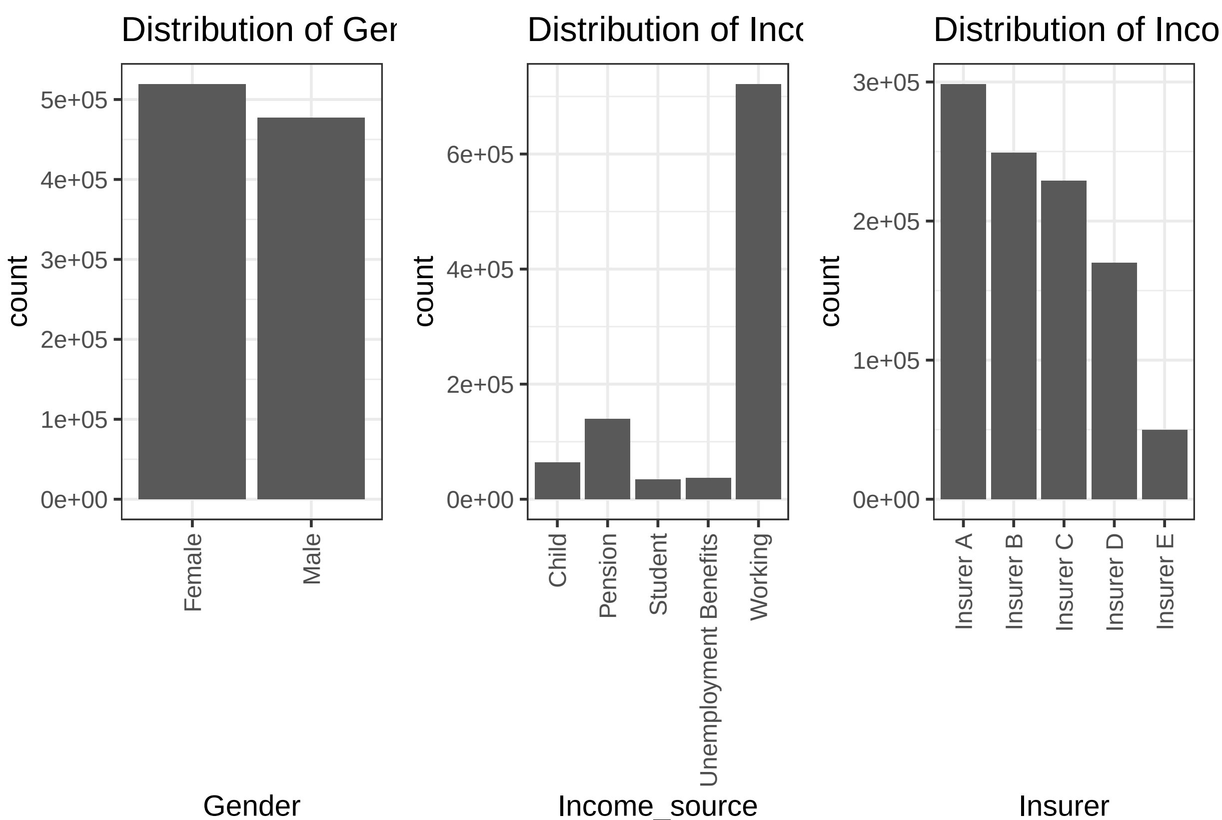

```
Gender_dist <- ggplot(data = data, aes( x = Gender))+
  geom_bar()+
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Gender")
```

```
Income_dist <- ggplot(data = data, aes( x = Income_source))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Income_source")
```

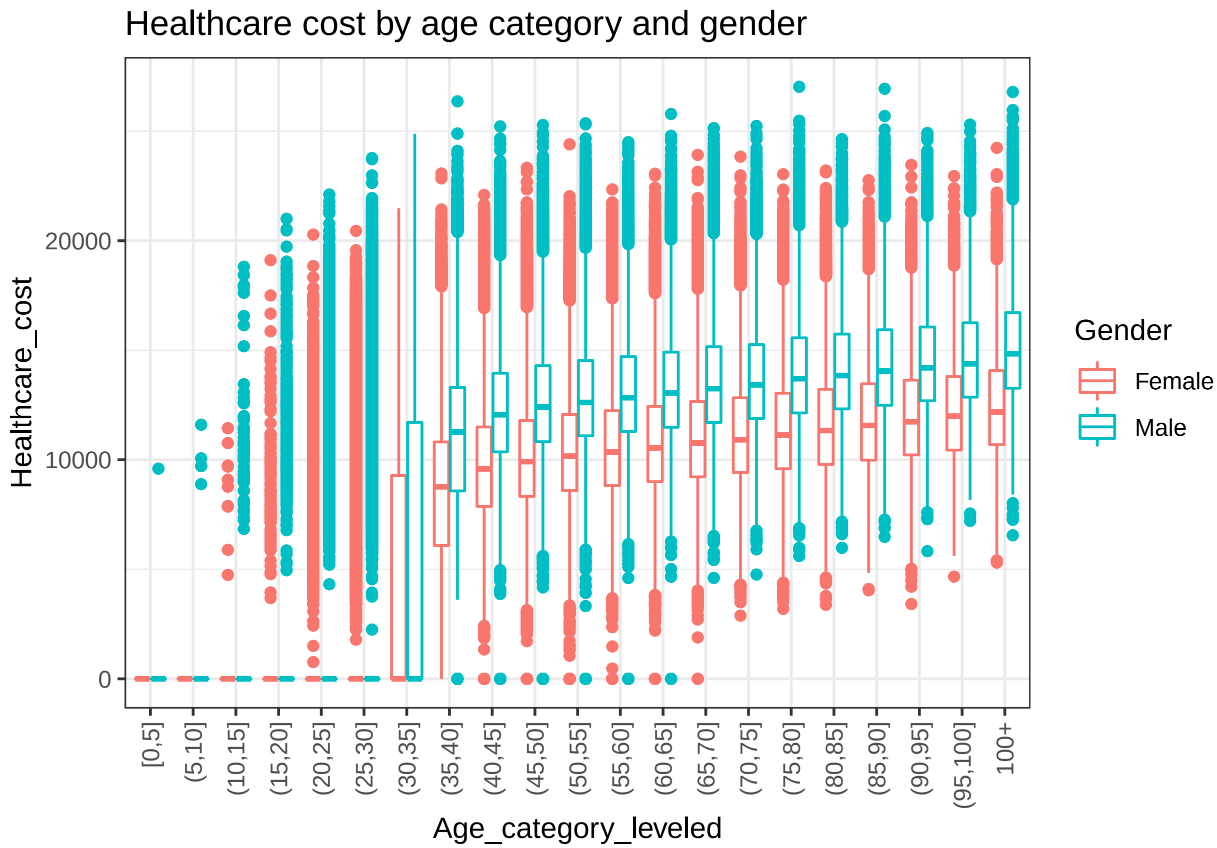
```
Insurer_dist <- ggplot(data = data, aes( x = Insurer))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Income_source")
```

```
Gender_dist + Income_dist + Insurer_dist
```



Exploratory Graphs

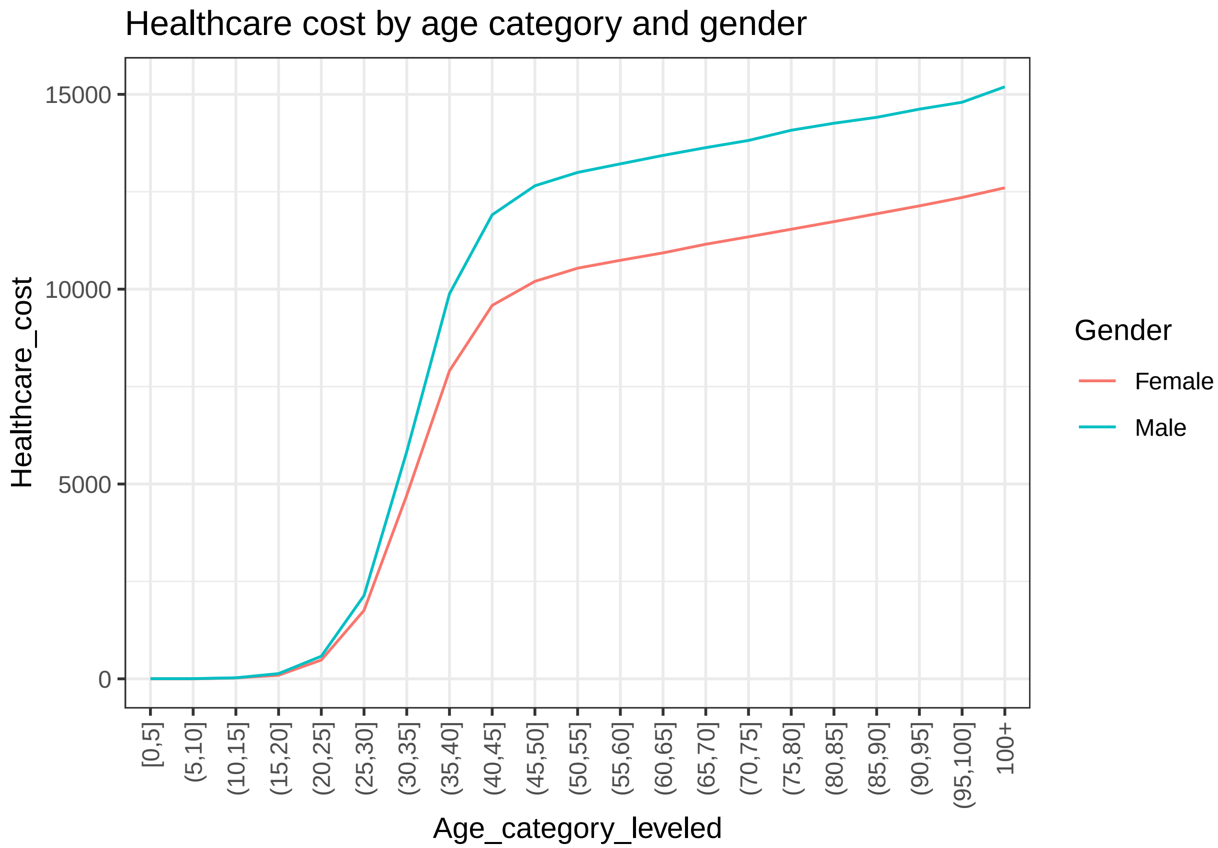
```
ggplot(data = data, aes( x = Age_category_leveled, y = Healthcare_cost, color = Gender))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Healthcare cost by age category and gender")
```



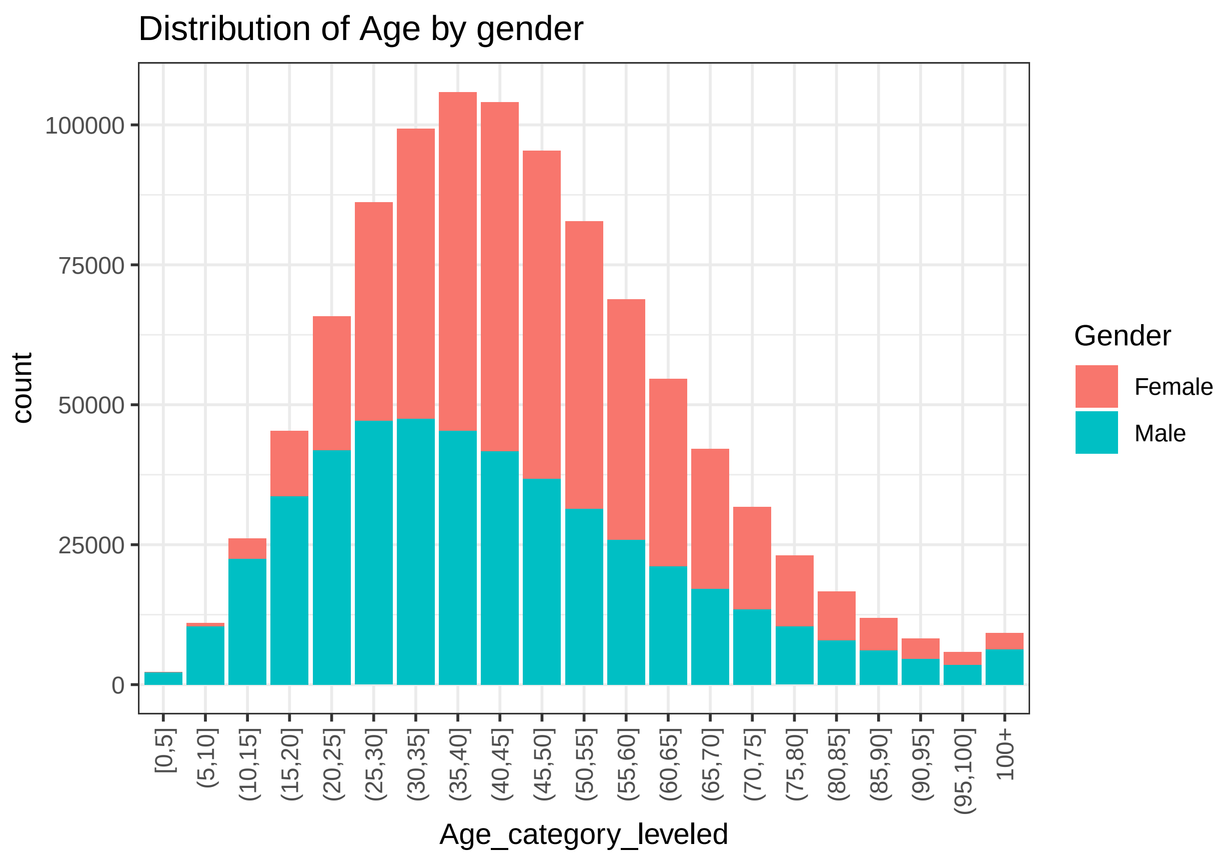
This boxplot shows that healthcare costs are increasing over age category and that overall the female healthcare costs are lower than the male healthcare costs. The below line graph shows the difference by gender.

```
ggplot(data = data, aes( x = Age_category_leveled, y = Healthcare_cost, colour = Gender))+
  stat_summary(aes(y = Healthcare_cost, group = Gender), fun.y = mean, geom = "line")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Healthcare cost by age category and gender")
```

```
## Warning: 'fun.y' is deprecated. Use 'fun' instead.
```




```
ggplot(data = data, aes( x = Age_category_leveled, fill = Gender))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Age by gender")
```



In the graph above we can see that the share of males in the youngest categories is very large, also in the oldest categories this difference can be observed.

Estimating model based on Age and Gender

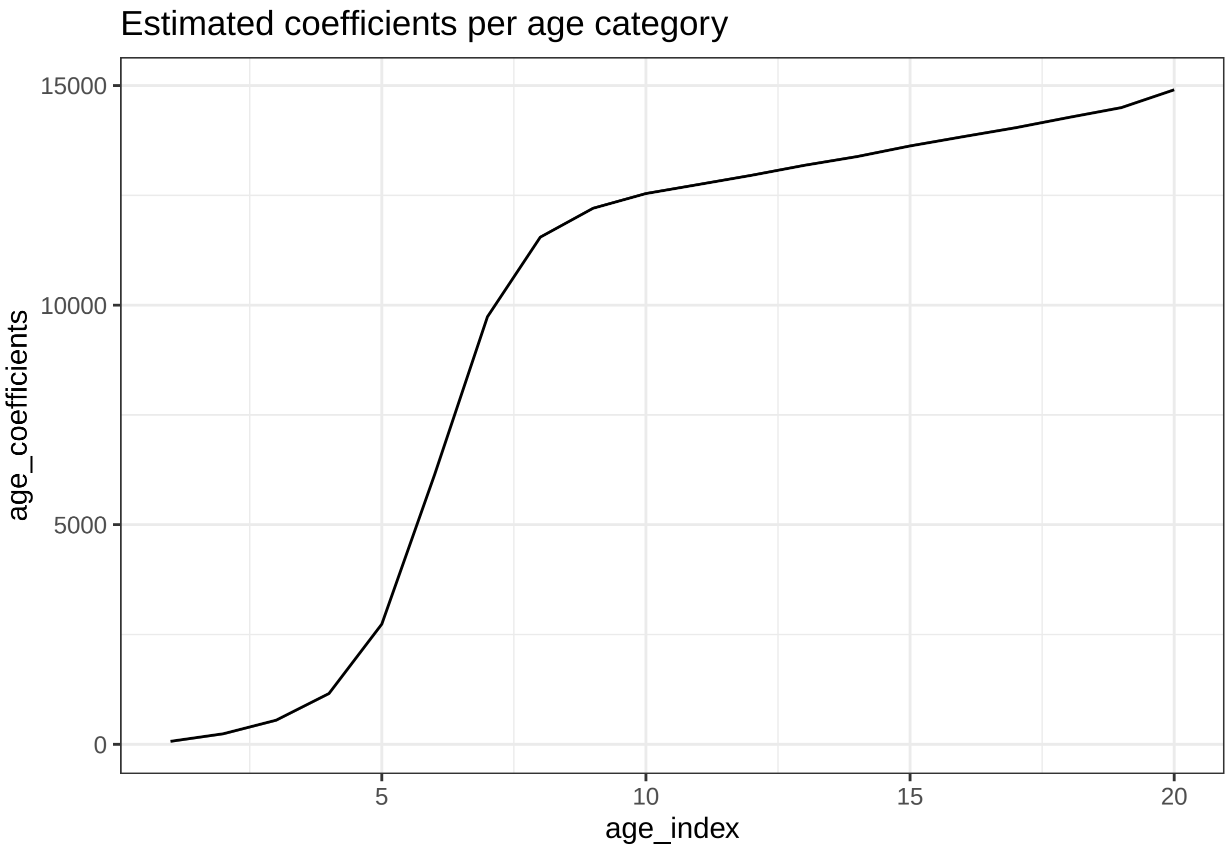
```
model1 <- lm(Healthcare_cost ~ Age_category_leveled + Gender, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12991.0  -1960.4   -388.5   1644.3  21003.6
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1745.620     79.094 -22.070  < 2e-16 ***
## Age_category_leveled(5,10]       67.637     86.357   0.783  0.43350
## Age_category_leveled(10,15]     240.471     82.041   2.931  0.00338 **
## Age_category_leveled(15,20]     548.903     80.669   6.804 1.02e-11 ***
## Age_category_leveled(20,25]    1156.457     80.102  14.437  < 2e-16 ***
```

```
## Age_category_leveled(25,30]   2737.504     79.815  34.298  < 2e-16 ***
## Age_category_leveled(30,35]   6141.315     79.706  77.050  < 2e-16 ***
## Age_category_leveled(35,40]   9735.011     79.672 122.189  < 2e-16 ***
## Age_category_leveled(40,45]  11547.458     79.698 144.890  < 2e-16 ***
## Age_category_leveled(45,50]  12205.643     79.780 152.991  < 2e-16 ***
## Age_category_leveled(50,55]  12541.077     79.922 156.917  < 2e-16 ***
## Age_category_leveled(55,60]  12747.848     80.137 159.077  < 2e-16 ***
## Age_category_leveled(60,65]  12956.523     80.457 161.037  < 2e-16 ***
## Age_category_leveled(65,70]  13183.380     80.917 162.925  < 2e-16 ***
## Age_category_leveled(70,75]  13382.899     81.568 164.070  < 2e-16 ***
## Age_category_leveled(75,80]  13624.450     82.554 165.036  < 2e-16 ***
## Age_category_leveled(80,85]  13834.795     83.951 164.795  < 2e-16 ***
## Age_category_leveled(85,90]  14039.950     85.920 163.407  < 2e-16 ***
## Age_category_leveled(90,95]  14272.745     88.863 160.615  < 2e-16 ***
## Age_category_leveled(95,100] 14497.314     92.662 156.453  < 2e-16 ***
## Age_category_leveled100+     14903.376     87.791 169.759  < 2e-16 ***
## GenderMale                    1780.134      7.731 230.257  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3724 on 996286 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.6076
## F-statistic: 7.346e+04 on 21 and 996286 DF,  p-value: < 2.2e-16
```

The simple model with age and gender above shows that, just like in the graphs, the costly individuals are older individuals and males in general. Males have on average 1780 more healthcost. The older the individual the more healthcosts you will on average have, this can be observed from the increasing coefficient of the age_categories. The older the category the higher the coefficient estimate of the age_category, meaning that on average an individual will have higher health costs when they fall in a higher age category.

```
age_coefficients <- model1$coefficients[2:21]
age_index <- seq(20)
data_age_coeff <- data.frame(age_index, age_coefficients)
ggplot(data_age_coeff, aes(x = age_index, y = age_coefficients))+
  geom_line()+
  ggtitle("Estimated coefficients per age category")
```



A second model based on the ordered age:

Order_age is a variable ranging from 1 to 24 depending on the age category, the higher the number the higher the age category. a one increase in the Order_age means one higher age category. For msot of the data (except above age 100) this means that a person is in an age class of 5 years higher. See below for a table of Order_age values per Age category.

```
age_table <- data %>%
  group_by(Order_age)%>%
  distinct(Age_category)
colnames(age_table) <- c("Age_category", "age_index")
age_table$age_index = age_table$age_index - 1
age_table <- age_table[0:21,]
age_table
```

```
## # A tibble: 21 x 2
## # Groups:   age_index [21]
##    Age_category age_index
##    <fct>            <dbl>
##  1 [0,5]                0
##  2 (5,10]               1
##  3 (10,15]              2
##  4 (15,20]              3
##  5 (20,25]              4
```

```
##   6 (25,30]              5
##   7 (30,35]              6
##   8 (35,40]              7
##   9 (40,45]              8
## 10 (45,50]              9
## # ... with 11 more rows
```

```
model2 <- lm(Healthcare_cost ~ Order_age + Gender, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Order_age + Gender, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16385  -3750   -326   3139  19239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2058.339     12.895  -159.6   <2e-16 ***
## Order_age    1032.626      1.139   906.9   <2e-16 ***
## GenderMale    926.755      8.912   104.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4400 on 996305 degrees of freedom
## Multiple R-squared:  0.4524, Adjusted R-squared:  0.4524
## F-statistic: 4.115e+05 on 2 and 996305 DF,  p-value: < 2.2e-16
```

The outcome of this regression shows that on average when you increase Order_age by 1, so fall in an age category higher, you will have 1032.6 increased health costs. The outcome from this regression also shows that you will on average have 926.8 increased health costs when you are male instead of female.

To conclude from both regressions the groups that are profitable and the groups that are loss-making:

- Profitable:
  - Females on average have lower healthcare costs.
  - Young people under 30 are more profitable (see the coefficients per age category graph above), there is a large jump between age category [25-30] with a value 2737 and age category [30-35] with a value 6141. See the graph for the exact change in coefficient. But it depends on the premium of the indivduals at what age the indivuals become loss-mkaing on average.
- Loss-making:
  - Males on average have higher healthcare costs.
  - Older individuals have increased healthcare costs (see the coefficients per age category graph above). Older people are more likely to be loss making.

Model Extension

I will now extend the model using other available data and analyze whether this extra data increases the accuracy of the model.

```
summary(model1)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12991.0  -1960.4   -388.5   1644.3  21003.6
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1745.620     79.094 -22.070  < 2e-16 ***
## Age_category_leveled(5,10]     67.637     86.357   0.783  0.43350
## Age_category_leveled(10,15]   240.471     82.041   2.931  0.00338 **
## Age_category_leveled(15,20]   548.903     80.669   6.804 1.02e-11 ***
## Age_category_leveled(20,25]  1156.457     80.102  14.437  < 2e-16 ***
## Age_category_leveled(25,30]  2737.504     79.815  34.298  < 2e-16 ***
## Age_category_leveled(30,35]  6141.315     79.706  77.050  < 2e-16 ***
## Age_category_leveled(35,40]  9735.011     79.672 122.189  < 2e-16 ***
## Age_category_leveled(40,45] 11547.458     79.698 144.890  < 2e-16 ***
## Age_category_leveled(45,50] 12205.643     79.780 152.991  < 2e-16 ***
## Age_category_leveled(50,55] 12541.077     79.922 156.917  < 2e-16 ***
## Age_category_leveled(55,60] 12747.848     80.137 159.077  < 2e-16 ***
## Age_category_leveled(60,65] 12956.523     80.457 161.037  < 2e-16 ***
## Age_category_leveled(65,70] 13183.380     80.917 162.925  < 2e-16 ***
## Age_category_leveled(70,75] 13382.899     81.568 164.070  < 2e-16 ***
## Age_category_leveled(75,80] 13624.450     82.554 165.036  < 2e-16 ***
## Age_category_leveled(80,85] 13834.795     83.951 164.795  < 2e-16 ***
## Age_category_leveled(85,90] 14039.950     85.920 163.407  < 2e-16 ***
## Age_category_leveled(90,95] 14272.745     88.863 160.615  < 2e-16 ***
## Age_category_leveled(95,100] 14497.314     92.662 156.453  < 2e-16 ***
## Age_category_leveled100+    14903.376     87.791 169.759  < 2e-16 ***
## GenderMale                   1780.134      7.731 230.257  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3724 on 996286 degrees of freedom
## Multiple R-squared:  0.6076, Adjusted R-squared:  0.6076
## F-statistic: 7.346e+04 on 21 and 996286 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(Healthcare_cost ~ Age_category_leveled + Gender + Income_source + Limited_coverage + Unheal
summary(model3)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Income_source + Limited_coverage + Unhealthy_region + Population_density,
##     data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -16078.5  -1727.1     80.4   1934.7  17612.1
##
## Coefficients:
##                                     Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                        -2345.061     71.214  -32.930  < 2e-16 ***
## Age_category_leveled(5,10]            90.585     77.356    1.171 0.241594
## Age_category_leveled(10,15]          266.038     73.489    3.620 0.000294 ***
## Age_category_leveled(15,20]          487.520     73.670    6.618 3.65e-11 ***
## Age_category_leveled(20,25]          932.276     78.329   11.902  < 2e-16 ***
## Age_category_leveled(25,30]         2353.064     78.637   29.923  < 2e-16 ***
## Age_category_leveled(30,35]         5408.188     78.567   68.835  < 2e-16 ***
## Age_category_leveled(35,40]         8633.811     78.583  109.869  < 2e-16 ***
## Age_category_leveled(40,45]        10274.548     78.633  130.665  < 2e-16 ***
## Age_category_leveled(45,50]        10879.236     78.708  138.222  < 2e-16 ***
## Age_category_leveled(50,55]        11197.633     78.826  142.055  < 2e-16 ***
## Age_category_leveled(55,60]        11413.486     79.001  144.473  < 2e-16 ***
## Age_category_leveled(60,65]        11609.820     79.262  146.474  < 2e-16 ***
## Age_category_leveled(65,70]        11769.217     85.422  137.778  < 2e-16 ***
## Age_category_leveled(70,75]        11967.433     89.353  133.934  < 2e-16 ***
## Age_category_leveled(75,80]        12196.587     90.077  135.401  < 2e-16 ***
## Age_category_leveled(80,85]        12421.364     91.108  136.336  < 2e-16 ***
## Age_category_leveled(85,90]        12619.808     92.570  136.327  < 2e-16 ***
## Age_category_leveled(90,95]        12879.083     94.775  135.892  < 2e-16 ***
## Age_category_leveled(95,100]       13077.553     97.651  133.921  < 2e-16 ***
## Age_category_leveled100+           13501.194     93.973  143.670  < 2e-16 ***
## GenderMale                          1784.937      6.926  257.721  < 2e-16 ***
## Income_sourcePension                1445.198     51.461   28.083  < 2e-16 ***
## Income_sourceStudent                1353.378     34.663   39.043  < 2e-16 ***
## Income_sourceUnemployment Benefits  1430.432     37.271   38.379  < 2e-16 ***
## Income_sourceWorking                1365.035     33.049   41.303  < 2e-16 ***
## Limited_coverage                   -3909.255     14.214 -275.026  < 2e-16 ***
## Unhealthy_region                    3862.047      9.374  411.977  < 2e-16 ***
## Population_density                    -1.951      2.363   -0.826 0.408921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3336 on 996279 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6851
## F-statistic: 7.743e+04 on 28 and 996279 DF,  p-value: < 2.2e-16
```

Population density will be left out of the model as the estimated effect size is very small and the coefficient is not significant. The other variables I will leave in the model, in the appendix summaries can be found on the models with and without the other variables. From those I conclude that model fit (R-squared) and statistical significance are optimal when I leave in all the variables except population density. The other added variables do seem to have a significant effect on the health costs based on the estimated coefficients and the statistical significance of these coefficients. Another important measure to check whether the model has become more accurate with the added variables is the value of the R-squared of the model. Compared to the model without the added variables we see an increase in the R-squared. R-squared value of the basic age and gender model: 0.6076 R-squared value of the model with added variables: 0.6852

When we remove the population density variable from the model we are left with the following model:

```
model4 <- lm(Healthcare_cost ~ Age_category_leveled + Gender + Income_source + Limited_coverage + Unheal
summary(model4)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Income_source + Limited_coverage + Unhealthy_region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16074.6  -1727.2     78.5   1935.0  17614.0
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       -2350.873     70.865  -33.174  < 2e-16 ***
## Age_category_leveled(5,10]           90.534     77.356    1.170 0.241855
## Age_category_leveled(10,15]         266.000     73.489    3.620 0.000295 ***
## Age_category_leveled(15,20]         487.475     73.670    6.617 3.67e-11 ***
## Age_category_leveled(20,25]         932.276     78.329   11.902  < 2e-16 ***
## Age_category_leveled(25,30]        2353.053     78.637   29.923  < 2e-16 ***
## Age_category_leveled(30,35]        5408.182     78.567   68.835  < 2e-16 ***
## Age_category_leveled(35,40]        8633.808     78.583  109.869  < 2e-16 ***
## Age_category_leveled(40,45]       10274.535     78.633  130.665  < 2e-16 ***
## Age_category_leveled(45,50]       10879.234     78.708  138.222  < 2e-16 ***
## Age_category_leveled(50,55]       11197.634     78.826  142.055  < 2e-16 ***
## Age_category_leveled(55,60]       11413.468     79.001  144.473  < 2e-16 ***
## Age_category_leveled(60,65]       11609.804     79.262  146.474  < 2e-16 ***
## Age_category_leveled(65,70]       11769.253     85.422  137.778  < 2e-16 ***
## Age_category_leveled(70,75]       11967.492     89.353  133.935  < 2e-16 ***
## Age_category_leveled(75,80]       12196.640     90.077  135.402  < 2e-16 ***
## Age_category_leveled(80,85]       12421.407     91.108  136.337  < 2e-16 ***
## Age_category_leveled(85,90]       12619.849     92.570  136.327  < 2e-16 ***
## Age_category_leveled(90,95]       12879.102     94.775  135.892  < 2e-16 ***
## Age_category_leveled(95,100]      13077.584     97.651  133.921  < 2e-16 ***
## Age_category_leveled100+          13501.266     93.973  143.671  < 2e-16 ***
## GenderMale                         1784.935      6.926  257.720  < 2e-16 ***
## Income_sourcePension               1445.114     51.461   28.082  < 2e-16 ***
## Income_sourceStudent               1353.337     34.663   39.042  < 2e-16 ***
## Income_sourceUnemployment Benefits  1430.392     37.271   38.378  < 2e-16 ***
## Income_sourceWorking               1365.002     33.049   41.302  < 2e-16 ***
## Limited_coverage                  -3909.250     14.214 -275.025  < 2e-16 ***
## Unhealthy_region                   3862.047      9.374  411.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3336 on 996280 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6851
## F-statistic: 8.03e+04 on 27 and 996280 DF,  p-value: < 2.2e-16
```

Compared to the previous model the R-squared has not changed, which also indicates that the population density did not add accuracy to the model.

Model analysis

From the estimated coefficients of the final model we can observe the following:

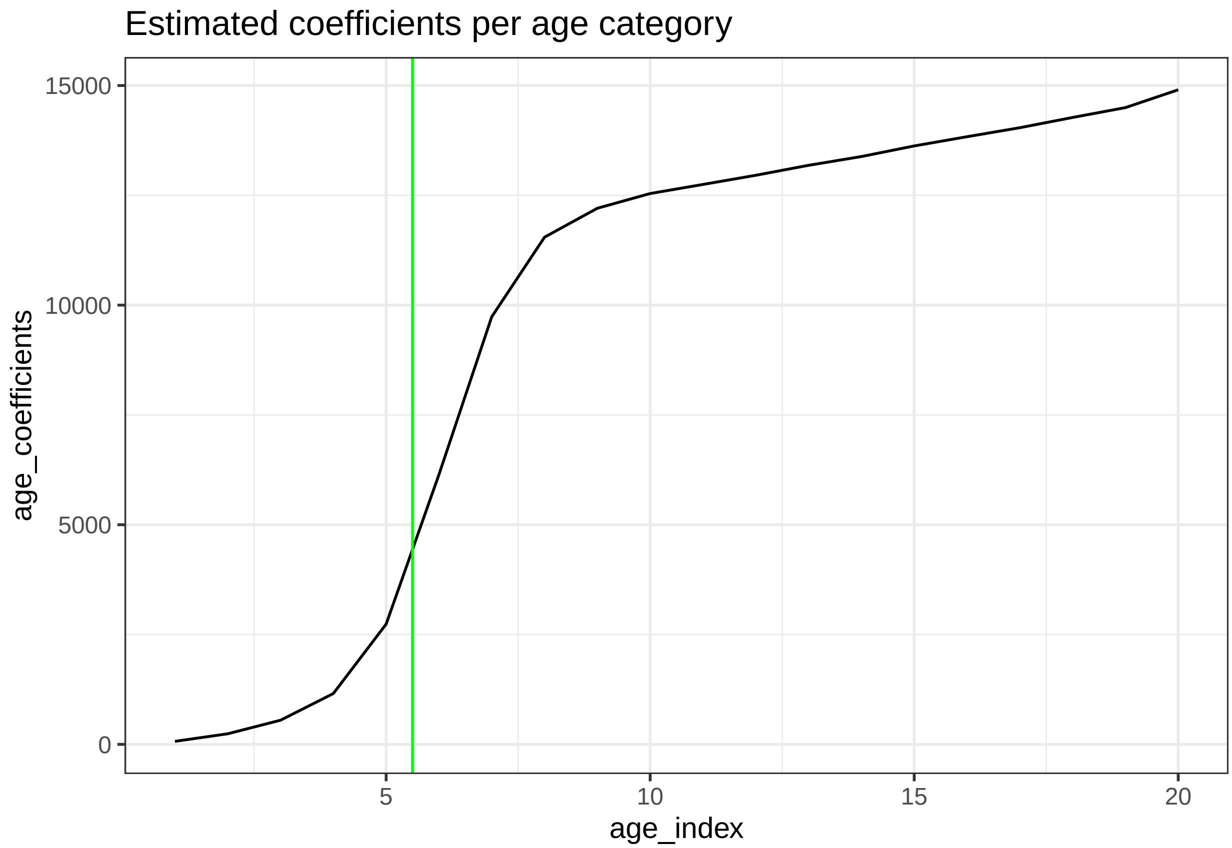
- Increased age category has on average the result that you have more health costs.
- When you are male you will on average have 1785 more health costs.
- The different income sources have different sized effects on the health costs. It can be expected that someone who has unemployment benefits have on average a higher health cost than someone who works and it can be expected that students (often young and healthy) will have lower health costs than people who live on a pension. This reasoning can be seen in the estimated coefficients: *Income_sourcePension: 1445.114* Income_sourceStudent: 1353.337 *Income_sourceUnemployment Benefits: 1430.392* Income_sourceWorking: 1365.002
- Whether or not you have limited coverage on your insurance has a large effect. When you have limited coverage on your insurance you have on average 3909.3 less health costs compared to someone who has full coverage.
- Whether or not you live in an unhealthy region also has a large effect on your health costs. When you live in an unhealthy region you have on average 3862.0 more health costs than someone who does not live in an unhealthy region.

Further analysis using graphs

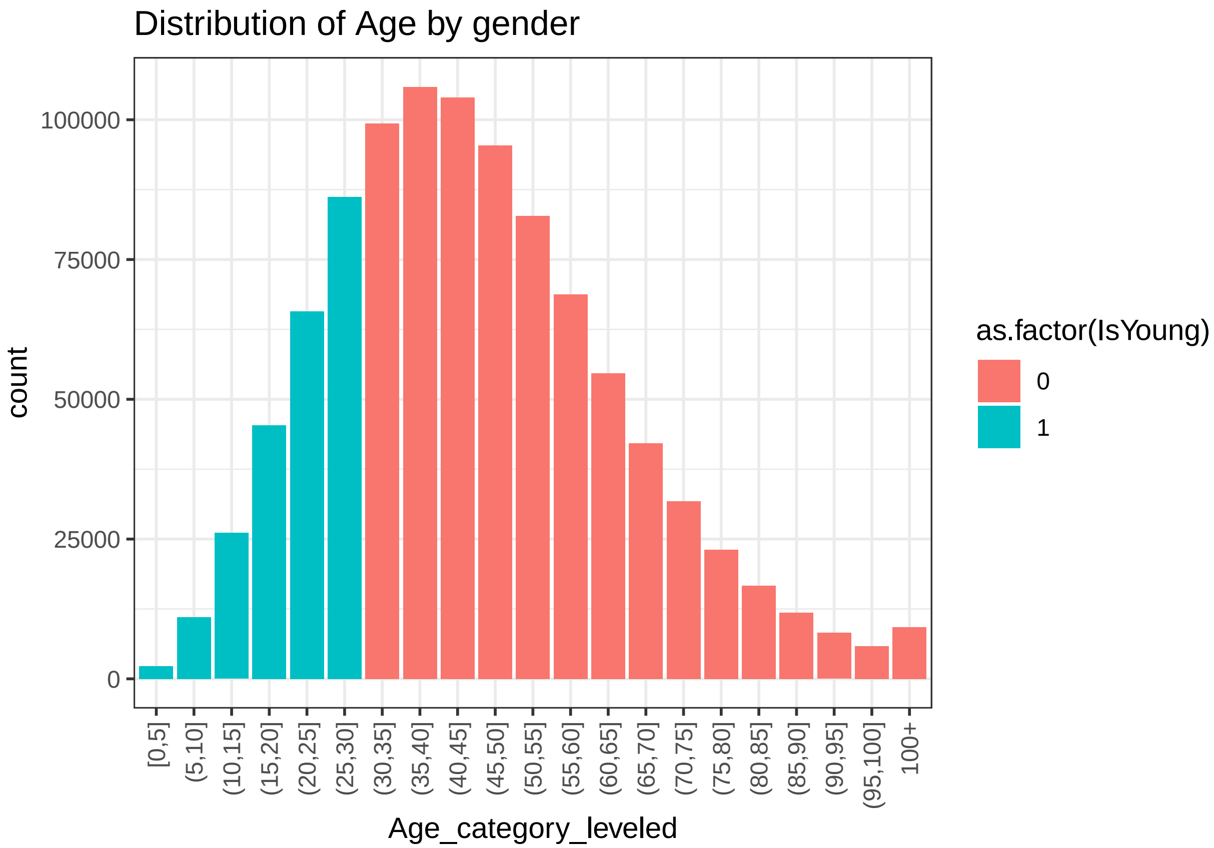
As we concluded above the individuals who are in the category [30-35] and above have a much higher health cost than those younger than them. So for analysis I split this group in two where one group is everyone under 30 and one group is everyone above 30. In the grpah below this shows that I will include everyone up to the green line.

```
data$IsYoung <- ifelse(data$Order_age <= 6, 1, 0)
```

```
ggplot(data_age_coeff, aes(x = age_index, y = age_coefficients))+
  geom_line()+
  ggtitle("Estimated coefficients per age category")+
  geom_vline(xintercept = 5.5, colour = 'green')
```

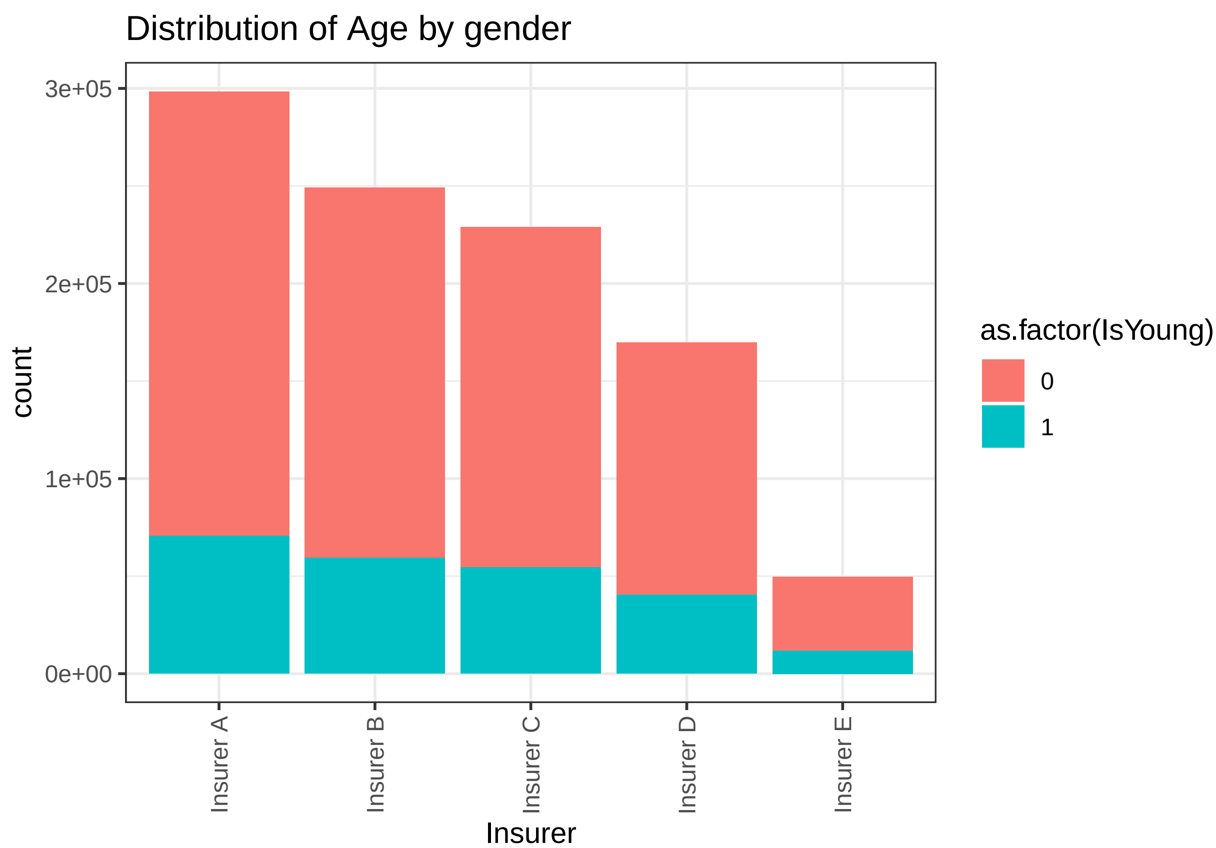



```
ggplot(data = data, aes( x = Age_category_leveled, fill = as.factor(IsYoung)))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Age by gender")
```



Using this split I can now visualize which insurer has the most profitable individuals.

```
ggplot(data = data, aes( x = Insurer, fill = as.factor(IsYoung)))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))+
  ggtitle("Distribution of Age by gender")
```




```
data%>%
  group_by(Insurer)%>%
  summarise_at(vars(IsYoung), funs(mean(.)))
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
```

```
## # A tibble: 5 x 2
##   Insurer   IsYoung
##   <fct>       <dbl>
## 1 Insurer A   0.237
## 2 Insurer B   0.238
## 3 Insurer C   0.238
## 4 Insurer D   0.238
## 5 Insurer E   0.239
```

From this analysis we can observe that Insurer E has the highest share of people under 30, namely 23.9%. However the percentages between insurers do not differ much. The lowest percentage is 23.67%, while the highest (from insurer E) is 23.91%

Appendix

```
model_A1 <- lm(Healthcare_cost ~ Age_category_leveled + Gender  + Limited_coverage + Unhealthy_region +
summary(model_A1)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Limited_coverage + Unhealthy_region + Population_density,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16078.6  -1703.5     77.4   1936.6  17626.5
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 -2341.500     71.280  -32.849  < 2e-16 ***
## Age_category_leveled(5,10]     90.426     77.428    1.168 0.242856
## Age_category_leveled(10,15]   265.540     73.557    3.610 0.000306 ***
## Age_category_leveled(15,20]  1097.774     72.353   15.173  < 2e-16 ***
## Age_category_leveled(20,25]  2269.503     71.933   31.550  < 2e-16 ***
## Age_category_leveled(25,30]  3699.819     71.646   51.640  < 2e-16 ***
## Age_category_leveled(30,35]  6761.831     71.498   94.574  < 2e-16 ***
## Age_category_leveled(35,40]  9994.904     71.438  139.909  < 2e-16 ***
## Age_category_leveled(40,45] 11639.102     71.458  162.881  < 2e-16 ***
## Age_category_leveled(45,50] 12244.864     71.531  171.183  < 2e-16 ***
## Age_category_leveled(50,55] 12563.459     71.658  175.326  < 2e-16 ***
## Age_category_leveled(55,60] 12779.320     71.850  177.861  < 2e-16 ***
## Age_category_leveled(60,65] 12975.723     72.137  179.876  < 2e-16 ***
## Age_category_leveled(65,70] 13195.423     72.550  181.881  < 2e-16 ***
## Age_category_leveled(70,75] 13410.381     73.134  183.368  < 2e-16 ***
## Age_category_leveled(75,80] 13639.634     74.018  184.275  < 2e-16 ***
## Age_category_leveled(80,85] 13864.525     75.271  184.196  < 2e-16 ***
## Age_category_leveled(85,90] 14063.128     77.036  182.553  < 2e-16 ***
## Age_category_leveled(90,95] 14322.552     79.675  179.763  < 2e-16 ***
## Age_category_leveled(95,100] 14521.206     83.081  174.784  < 2e-16 ***
## Age_category_leveled100+    14945.190     78.714  189.868  < 2e-16 ***
## GenderMale                   1780.921      6.932  256.926  < 2e-16 ***
## Limited_coverage            -3828.769     14.103 -271.488  < 2e-16 ***
## Unhealthy_region             3861.618      9.383  411.549  < 2e-16 ***
## Population_density             -1.799      2.365   -0.761 0.446856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3339 on 996283 degrees of freedom
## Multiple R-squared:  0.6846, Adjusted R-squared:  0.6846
```

```
## F-statistic: 9.009e+04 on 24 and 996283 DF,  p-value: < 2.2e-16
```

```
model_A2 <- lm(Healthcare_cost ~ Age_category_leveled + Gender + Income_source  + Unhealthy_region + Po
summary(model_A2)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Income_source + Unhealthy_region + Population_density, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16074  -1802     41   1735  18570
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2341.199     73.867 -31.695  < 2e-16 ***
## Age_category_leveled(5,10]           90.403     80.239   1.127 0.259880
## Age_category_leveled(10,15]         265.471     76.228   3.483 0.000497 ***
## Age_category_leveled(15,20]         486.538     76.416   6.367 1.93e-10 ***
## Age_category_leveled(20,25]         961.099     81.248  11.829  < 2e-16 ***
## Age_category_leveled(25,30]        2549.378     81.564  31.256  < 2e-16 ***
## Age_category_leveled(30,35]        5956.546     81.469  73.114  < 2e-16 ***
## Age_category_leveled(35,40]        9551.568     81.438 117.287  < 2e-16 ***
## Age_category_leveled(40,45]       11363.316     81.459 139.497  < 2e-16 ***
## Age_category_leveled(45,50]       12018.729     81.528 147.418  < 2e-16 ***
## Age_category_leveled(50,55]       12350.828     81.648 151.270  < 2e-16 ***
## Age_category_leveled(55,60]       12568.537     81.829 153.595  < 2e-16 ***
## Age_category_leveled(60,65]       12765.476     82.100 155.487  < 2e-16 ***
## Age_category_leveled(65,70]       12924.858     88.498 146.047  < 2e-16 ***
## Age_category_leveled(70,75]       13123.143     92.580 141.749  < 2e-16 ***
## Age_category_leveled(75,80]       13352.410     93.333 143.063  < 2e-16 ***
## Age_category_leveled(80,85]       13577.315     94.403 143.823  < 2e-16 ***
## Age_category_leveled(85,90]       13775.940     95.921 143.617  < 2e-16 ***
## Age_category_leveled(90,95]       14035.382     98.210 142.912  < 2e-16 ***
## Age_category_leveled(95,100]      14234.062    101.197 140.658  < 2e-16 ***
## Age_category_leveled100+          14658.093     97.378 150.528  < 2e-16 ***
## GenderMale                         1780.384      7.184 247.829  < 2e-16 ***
## Income_sourcePension                286.935     53.199   5.394 6.91e-08 ***
## Income_sourceStudent                212.817     35.697   5.962 2.50e-09 ***
## Income_sourceUnemployment Benefits  275.577     38.414   7.174 7.30e-13 ***
## Income_sourceWorking                206.096     34.001   6.061 1.35e-09 ***
## Unhealthy_region                   3861.457      9.724 397.115  < 2e-16 ***
## Population_density                   -1.715      2.451  -0.700 0.484182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3461 on 996280 degrees of freedom
## Multiple R-squared:  0.6613, Adjusted R-squared:  0.6612
## F-statistic: 7.203e+04 on 27 and 996280 DF,  p-value: < 2.2e-16
```

```
model_A3 <- lm(Healthcare_cost ~ Age_category_leveled + Gender + Income_source + Limited_coverage  + Po
summary(model_A3)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Income_source + Limited_coverage + Population_density, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12993.9  -2093.8   -277.6   1884.7  20508.9
##
## Coefficients:
##                                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                      -1744.926     77.025  -22.654  < 2e-16 ***
## Age_category_leveled(5,10]          67.879     83.686    0.811  0.41730
## Age_category_leveled(10,15]        241.134     79.503    3.033  0.00242 **
## Age_category_leveled(15,20]        459.375     79.699    5.764 8.22e-09 ***
## Age_category_leveled(20,25]        929.384     84.739   10.968  < 2e-16 ***
## Age_category_leveled(25,30]       2347.782     85.072   27.598  < 2e-16 ***
## Age_category_leveled(30,35]       5399.785     84.997   63.529  < 2e-16 ***
## Age_category_leveled(35,40]       8624.198     85.013  101.445  < 2e-16 ***
## Age_category_leveled(40,45]      10265.673     85.067  120.677  < 2e-16 ***
## Age_category_leveled(45,50]      10873.078     85.149  127.695  < 2e-16 ***
## Age_category_leveled(50,55]      11194.866     85.276  131.278  < 2e-16 ***
## Age_category_leveled(55,60]      11399.811     85.465  133.385  < 2e-16 ***
## Age_category_leveled(60,65]      11607.865     85.748  135.372  < 2e-16 ***
## Age_category_leveled(65,70]      11767.360     92.412  127.336  < 2e-16 ***
## Age_category_leveled(70,75]      11948.138     96.665  123.604  < 2e-16 ***
## Age_category_leveled(75,80]      12189.570     97.449  125.087  < 2e-16 ***
## Age_category_leveled(80,85]      12399.776     98.564  125.804  < 2e-16 ***
## Age_category_leveled(85,90]      12604.733    100.146  125.864  < 2e-16 ***
## Age_category_leveled(90,95]      12837.355    102.530  125.206  < 2e-16 ***
## Age_category_leveled(95,100]     13061.682    105.642  123.641  < 2e-16 ***
## Age_category_leveled100+         13467.268    101.663  132.469  < 2e-16 ***
## GenderMale                        1785.240      7.493  238.267  < 2e-16 ***
## Income_sourcePension              1437.640     55.672   25.823  < 2e-16 ***
## Income_sourceStudent              1347.327     37.500   35.929  < 2e-16 ***
## Income_sourceUnemployment Benefits  1401.408     40.321   34.756  < 2e-16 ***
## Income_sourceWorking              1349.507     35.754   37.745  < 2e-16 ***
## Limited_coverage                 -3907.915     15.377 -254.135  < 2e-16 ***
## Population_density                  -1.917      2.556   -0.750  0.45317
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3609 on 996280 degrees of freedom
## Multiple R-squared:  0.6315, Adjusted R-squared:  0.6315
## F-statistic: 6.324e+04 on 27 and 996280 DF,  p-value: < 2.2e-16
```

```r
model_A4 <- lm(Healthcare_cost ~ Age_category_leveled + Gender + Income_source + Limited_coverage + Unhe
summary(model_A4)
```

```
##
## Call:
## lm(formula = Healthcare_cost ~ Age_category_leveled + Gender +
##     Income_source + Limited_coverage + Unhealthy_region, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16074.6  -1727.2     78.5   1935.0  17614.0
##
## Coefficients:
##                                   Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                      -2350.873     70.865  -33.174  < 2e-16 ***
## Age_category_leveled(5,10]          90.534     77.356    1.170 0.241855
## Age_category_leveled(10,15]        266.000     73.489    3.620 0.000295 ***
## Age_category_leveled(15,20]        487.475     73.670    6.617 3.67e-11 ***
## Age_category_leveled(20,25]        932.276     78.329   11.902  < 2e-16 ***
## Age_category_leveled(25,30]       2353.053     78.637   29.923  < 2e-16 ***
## Age_category_leveled(30,35]       5408.182     78.567   68.835  < 2e-16 ***
## Age_category_leveled(35,40]       8633.808     78.583  109.869  < 2e-16 ***
## Age_category_leveled(40,45]      10274.535     78.633  130.665  < 2e-16 ***
## Age_category_leveled(45,50]      10879.234     78.708  138.222  < 2e-16 ***
## Age_category_leveled(50,55]      11197.634     78.826  142.055  < 2e-16 ***
## Age_category_leveled(55,60]      11413.468     79.001  144.473  < 2e-16 ***
## Age_category_leveled(60,65]      11609.804     79.262  146.474  < 2e-16 ***
## Age_category_leveled(65,70]      11769.253     85.422  137.778  < 2e-16 ***
## Age_category_leveled(70,75]      11967.492     89.353  133.935  < 2e-16 ***
## Age_category_leveled(75,80]      12196.640     90.077  135.402  < 2e-16 ***
## Age_category_leveled(80,85]      12421.407     91.108  136.337  < 2e-16 ***
## Age_category_leveled(85,90]      12619.849     92.570  136.327  < 2e-16 ***
## Age_category_leveled(90,95]      12879.102     94.775  135.892  < 2e-16 ***
## Age_category_leveled(95,100]     13077.584     97.651  133.921  < 2e-16 ***
## Age_category_leveled100+         13501.266     93.973  143.671  < 2e-16 ***
## GenderMale                        1784.935      6.926  257.720  < 2e-16 ***
## Income_sourcePension              1445.114     51.461   28.082  < 2e-16 ***
## Income_sourceStudent              1353.337     34.663   39.042  < 2e-16 ***
## Income_sourceUnemployment Benefits  1430.392     37.271   38.378  < 2e-16 ***
## Income_sourceWorking              1365.002     33.049   41.302  < 2e-16 ***
## Limited_coverage                 -3909.250     14.214 -275.025  < 2e-16 ***
## Unhealthy_region                  3862.047      9.374  411.977  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3336 on 996280 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6851
## F-statistic: 8.03e+04 on 27 and 996280 DF,  p-value: < 2.2e-16
```