Project

Radhika Agarwal, Maia Payne, Victor Tolulope Akangbe

2023-04-30

Abstract:

This regression analysis examines the relationship between various socioeconomic factors and health insurance coverage in the United States. The model utilizes eight explanatory variables to determine which factors are significantly associated with health insurance coverage and the type of coverage obtained. By analyzing the results, policymakers can make informed decisions on how to increase health insurance coverage throughout the country. The study's findings may contribute to the development of policies aimed at improving access to health care services for all Americans, particularly those who are socioeconomically disadvantaged. Overall, the analysis provides valuable insights into the factors that influence health insurance coverage in the United States and may inform future efforts to increase access to health care services.

Introduction:

There have been several studies conducted to determine the relationship between a person's wage and their health coverage status. The purpose of this research is to summarize the current state of health coverage within the respondent level. Many other studies have shown that there is a strong correlation between a person's wages and their likelihood of having health coverage. Dickman et al. (2017) this study found that the rising insurance premiums for employer -sponsored private coverage have broken down wage gains for middle-class Americans. As well as Kuroki (2022) article investigated the effects of minimum wage hikes on the proportion of uninsured people between 2008 and 2018.

The last study we looked at was by Stinson (2003), which investigated a combination model of salaries, job termination risk, and the likelihood of having employer-provided health insurance.

After doing the literature review, we started with the following hypothesis:

- Gender is unlikely to be a factor in determining whether a person has health insurance, as insurance providers typically do not discriminate based on gender.
- Those with more annual incomes are more likely to have health
- People from households with four or fewer members are more likely to have health insurance than those from larger households, possibly because of the higher cost of covering a bigger family.

There have been recent efforts have been made to address many different issues/ opportunities in health insurance within many of the different states. Many policy changes and initiatives have aimed to increase people with insurance (Smith, Horneffer, & O'Connell, 2022). Analyzing the data on health insurance can provide insight into understanding the socioeconomic factors that influence health care coverage in the United States.

The response variables will be health insurance coverage and its type being private or public. They will be measured as a (yes/no) variable based on whether the individual has health insurance.

Initially we started with the following regression model, and will look upon the necessary variables and come up with a final regression.

Health Coverage = $\beta0$ + $\beta1$ (Race) + $\beta2$ (sex) + $\beta3$ (wages) + $\beta4$ (age) + $\beta5$ (Persons Per Family) + $\beta6$ (Citizen) + $\beta7$ (Disability) + $\beta8$ (State)

Loading necessary libraries

```
library(tidyverse)
library(DescTools)
library(dplyr)
library(ggplot2)
library(vcd)
library(readxl)
library(psych)
library(writexl)
library(survey)
library(GGally)
```

Reading the CSV file

```
data = read_csv("final_health_ins_data.csv")
```

Data:

The data that we looked at we gather from the respondent-level Census data from the United States Census Bureau website.

(https://data.census.gov/mdat/#/search?ds=ACSPUMS1Y2021&vv=NPF,AGEP,WAGP&cv=PUBCOV,PRIVCOV,HICOV,SEX,RAC1P&rv=POWSP,CIT,DIS,ucgid&g=0400000US01,02,04,05

,06,08,09,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56)

This is the link to the micro data table from the Census Bureau using ACS 1-Year Estimates Public Use Microdata Sample (2021).

With this data we are able to examine the affects of a person having a health insurance in the United States.

Our data looks like this:

Disability	**Disability**	Yes	1
		No	0
Race	**Race**	White	1
		Others	0
Sex	**Sex**	Male	1
		Female	0
Wages	**Wages or salary income past 12 months**	$4 to 999999 (Rounded and top-coded)	4 - 999999
		Not Employed	0
Age	**Age**	1 to 99	1 to 99
		under 1 year	0
PersonsPerFamily	**Number of persons in family**	Number of persons in family	2 to 20
		N/A (GQ/vacant/non-	1

		family household)	
Citizen	**US Citizenship status**	Yes	1
		No	0
PublicHealthIns	**Public Health Insurance**	Yes	1
		No	0
PrivateHealthIns	**Private Health Insurance**	Yes	1
		No	0
HealthIns	**Health Coverage**	Yes	1
		No	0
State	**State**	Alabama	01
		Alaska	02
		Arizona	04
		Arkansas	05

		California	06
		Colorado	08
		Connecticut	09
		Delaware	10
		District of Columbia	11
		Florida	12
		Georgia	13
		Hawaii	15
		Idaho	16
		Illinois	17
		Indiana	18
		Iowa	19
		Kansas	20
		Kentucky	21

		Louisiana	22
		Maine	23
		Maryland	24
		Massachusetts	25
		Michigan	26
		Minnesota	27
		Mississippi	28
		Missouri	29
		Montana	30
		Nebraska	31
		Nevada	32
		New Hampshire	33
		New Jersey	34
		New Mexico	35

		New York	36
		North Carolina	37
		North Dakota	38
		Ohio	39
		Oklahoma	40
		Oregon	41
		Pennsylvania	42
		Rhode Island	44
		South Carolina	45
		South Dakota	46
		Tennessee	47
		Texas	48
		Utah	49
		Vermont	50

		Virginia	51
		Washington	53
		West Virginia	54
		Wisconsin	55
		Wyoming	56

Cleaning the Data

Before we dive into analysis, we had to clean the data.

- Converted the race attribute to people who are white as 1 and 0 represents all other races combined
- The dataset consists of -1 and NA values in Wages, which we have removed from out data.
- Converted the citizen attribute to people who are citizen of US and not
- Coded all values in attributes from 2 to 0 for better understanding.
- Removed Non-US states from the States attribute.

```
data$Race[data$Race == 2 |data$Race == 3 | data$Race == 4 |data$Race == 5 | data$Race == 6 |data$Race == 7 | data$Race == 8 |data$Race == 9] = 0

data$Wages[data$Wages == -1] = NA
data = data %>% drop_na(Wages)

data$Citizen[data$Citizen == 1 | data$Citizen == 2 | data$Citizen == 3 | data$Citizen == 4 ] = 1

data$Citizen[data$Citizen == 5 ] = 0
```

```r
data$Sex[data$Sex == 2 ] = 0
data$Sex[data$Sex == 1 ] = 1


data$HealthIns[data$HealthIns == 2 ] = 0
data$PublicHealthIns[data$PublicHealthIns == 2 ] = 0
data$PrivateHealthIns[data$PrivateHealthIns == 2 ] = 0

data$Disability[data$Disability == 2 ] = 0
data$Disability[data$Disability == 1 ] = 1

data <- subset(data, !(State %in% c('N',72,166,251,254,301,303,399,555)))
```

Logging Wages

Log transformed Wages to help with unequal variances.

```r
data = data %>% mutate(Wages_log = log10(Wages))
data = data[is.finite(data$Wages_log),]
```

Summarizing the data

```r
summary(data)

##  PersonsPerFamily      Age            Wages        PublicHealthIns
##  Min.   : 1.000   Min.   :16.00   Min.   :     4   Min.   :0.0000
##  1st Qu.: 2.000   1st Qu.:30.00   1st Qu.: 20000   1st Qu.:0.0000
##  Median : 2.000   Median :42.00   Median : 41000   Median :0.0000
##  Mean   : 2.762   Mean   :42.53   Mean   : 59763   Mean   :0.1623
##  3rd Qu.: 4.000   3rd Qu.:55.00   3rd Qu.: 74000   3rd Qu.:0.0000
##  Max.   :20.000   Max.   :95.00   Max.   :787000   Max.   :1.0000
##  PrivateHealthIns   HealthIns           Sex              Race
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.8295   Mean   :0.9234   Mean   :0.5215   Mean   :0.6708
```

```
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
## Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    State              Citizen          Disability         Wages_log
## Length:412650      Min.   :0.0000   Min.   :0.00000   Min.   :0.6021
## Class :character   1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:4.3010
## Mode  :character   Median :1.0000   Median :0.00000   Median :4.6128
##                    Mean   :0.9351   Mean   :0.06814   Mean   :4.5358
##                    3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:4.8692
##                    Max.   :1.0000   Max.   :1.00000   Max.   :5.8960
```

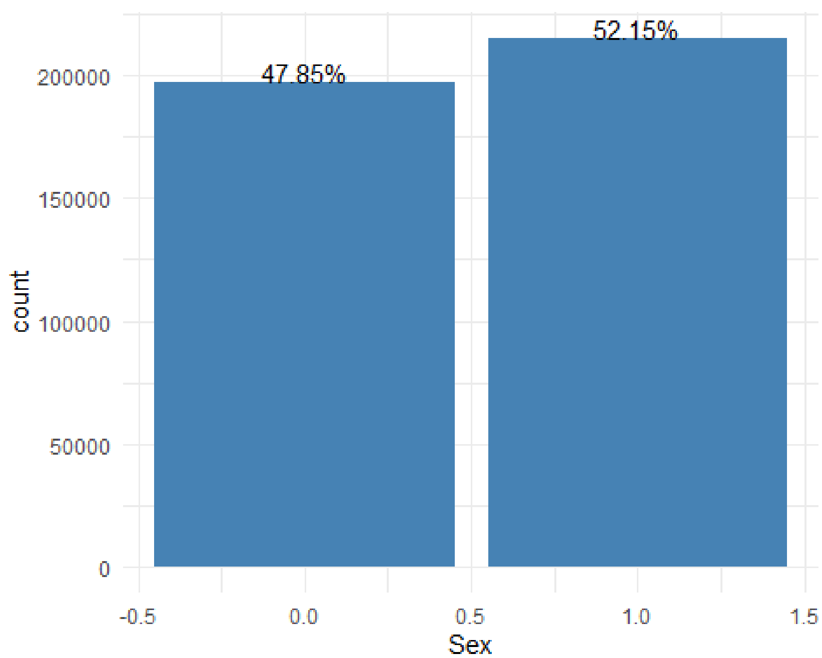
The summary shows the descriptive statistics for the variables in the data set, which includes the measures of minimum, maximum, mean, and median as well as the quartiles. This information is relevant because it provides a comprehensive overview of the data set.

- The Summary of the data is showing that the PersonsPerFamily: Minimum number of persons per family is 1.0 and the maximum number is 20.0. This variable is relevant because it represents the number of people in each family unit in the sample, which could be used to analyze family demographics when researching.
- The data also shows that the min age is 16 and the max is 95 which is relevant because it represents the age of each of the person in the sample.
- The minimum wage is $4, maximum wage is $787,000. This variable is relevant because it represents the income earned by each person in the sample.
- PublicHealthIns, PrivateHealthIns, HealthIns, all are represented in as either 0 being no insurance and 1 being they have health insurance. These variables are relevant because it represents whether each person in the sample has any form of health insurance, which could be used to analyze health insurance coverage and access to healthcare.
- Sex, Race, Citizen, and Disability have either value 0 or 1. These are all important because they could be used to analyze gender, race, disability, and citizenship demographics based on a response of yes or no.
- Lastly, with the wages_log: Minimum value is 0.6021, maximum value is 5.8960. This variable is relevant because it represents the logarithm of wages earned by each person in the sample and could be used to analyze income distribution.

Univariate Analysis

Percentage of people having sex as a Male or a Female

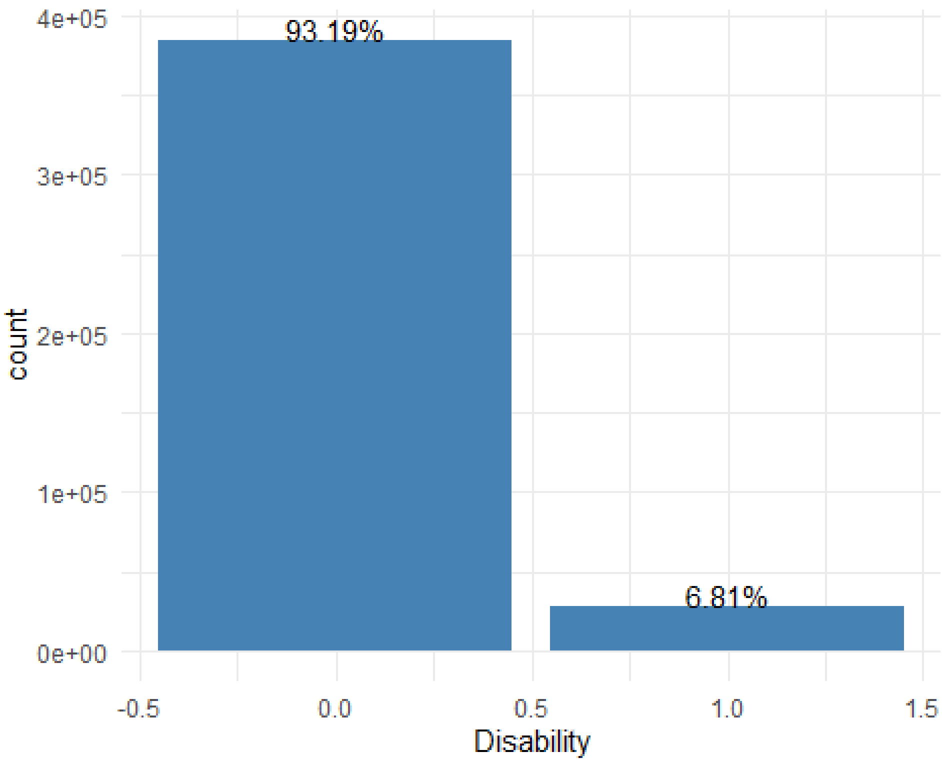
```
ggplot(data, aes(x=Sex)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0)+
  theme_minimal()
```



As we can see at the histogram, there are 47.85% female and 52.15% male in our data set. Although these percentages are unequal, it does not significantly impact on the analysis as the difference is minimal.

Percentage of people with a disability

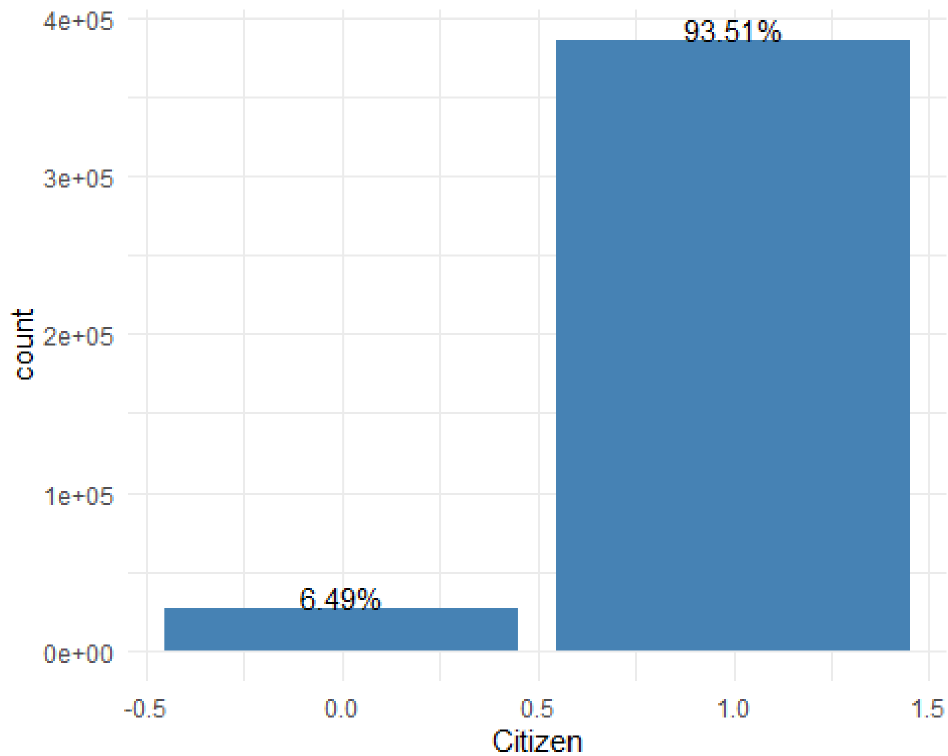
```
ggplot(data, aes(x=Disability)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  theme_minimal()
```



The histogram illustrates that 6.81% of the population have a disability, whereas 93.19% individuals do not have a disability. The graph displays a significant difference between the two categories.

Percentage of people who are US citizen

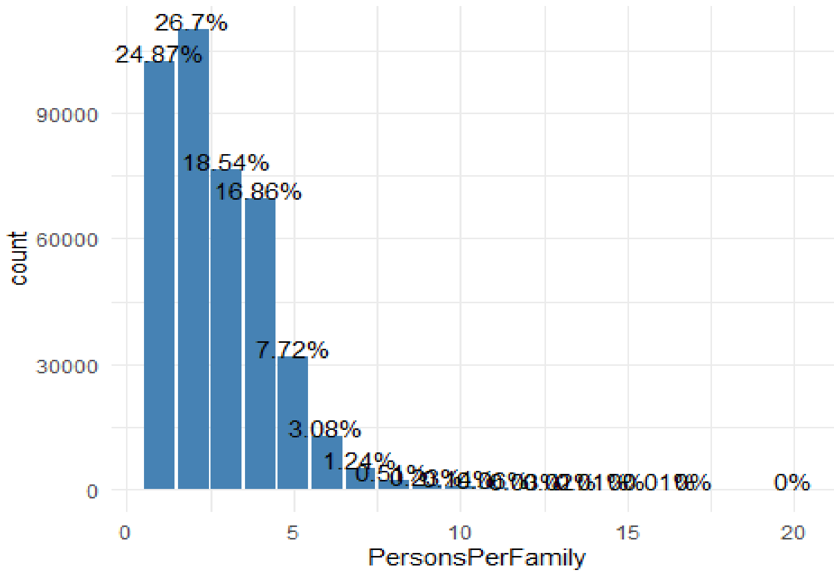
```
ggplot(data, aes(x=Citizen)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  theme_minimal()
```



The majority of the individuals involved identified themselves as American citizens (93.51%), and the remaining individuals (6.49%) identified themselves as foreigners.

Percentage of people per family

```
ggplot(data, aes(x=PersonsPerFamily)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  theme_minimal()
```

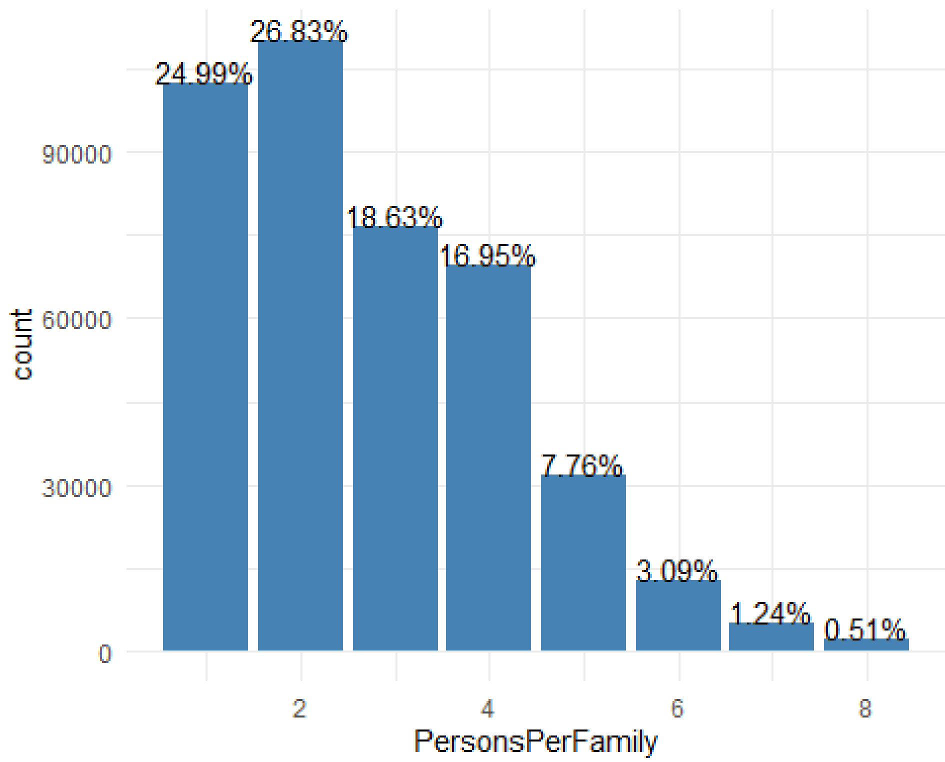


Removing outliers from PersonsPerFamily

```
data <- subset(data, !(PersonsPerFamily %in% c(9,10,11,12,13,14,15,16,17,18,19,20)))
```

Percentage of people per family after removing outliers

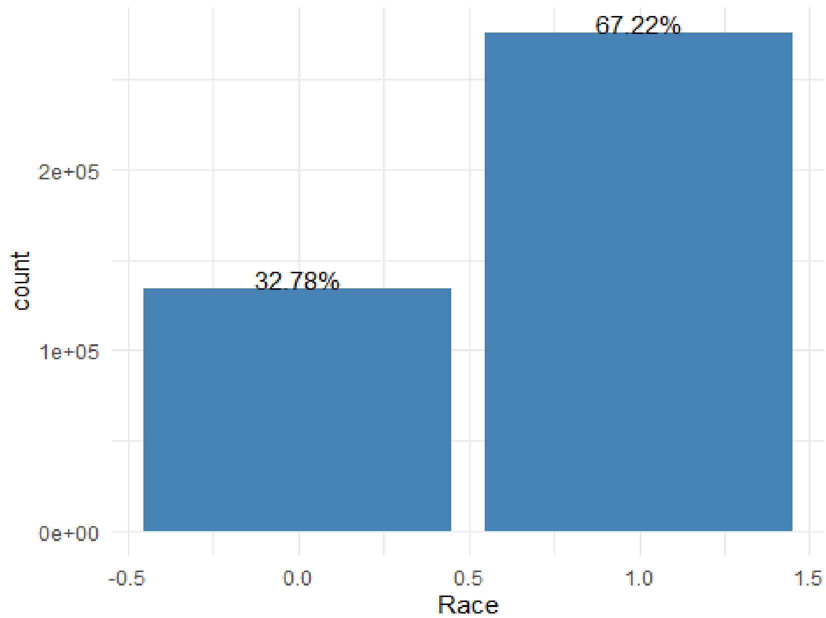
```
ggplot(data, aes(x=PersonsPerFamily)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  theme_minimal()
```



The graph indicates that the family size of 2 people is the most common within our data set.

Percentage of White People

```
ggplot(data, aes(x=Race)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  theme_minimal()
```

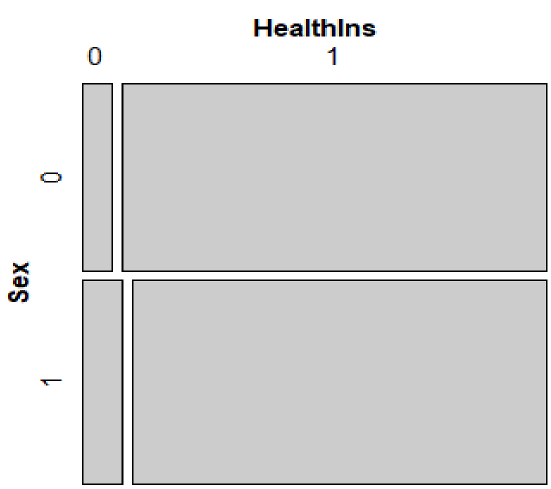


It can be seen that there are more individuals that identify as white 67.22% compared to those who identify as a person of color 32.78%.

Bivariate Analysis On Categorical Variables

Mosaic plot between Sex and Health Insurance

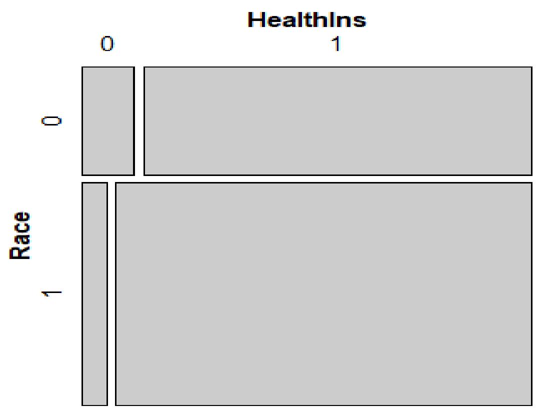
```
mosaic(~Sex+HealthIns, data = data)
```



The mosaic plot above shows the correlation between an individual's biological sex and health insurance to test whether sex is a factor that determines if an individual has health insurance. Based on the results of the mosaic plot, sex does not play a role in determining if an individual has health insurance. The tiles are roughly the same size and shape.

Mosaic plot between Race and Health Insurance

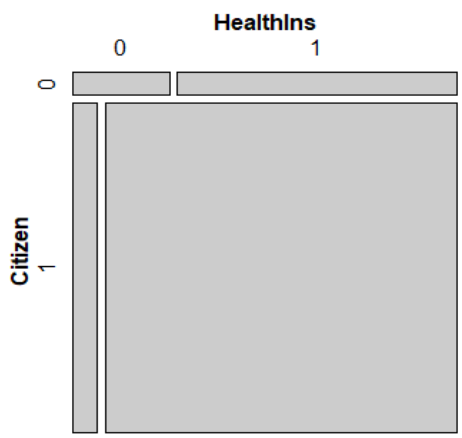
```
mosaic(~Race+HealthIns, data = data)
```



The plot shows that the tiles are roughly the same size and shape Indicating that there is not much effect of race on a person having a Health Insurance.

Mosaic plot between Citizen and Health Insurance

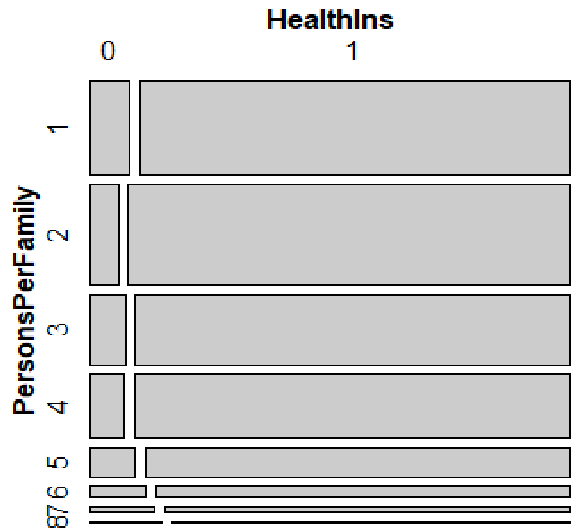
```
mosaic(~Citizen+HealthIns, data = data)
```



The graph shown above indicates that the tiles are not the same size or shape. It tells us that if a person is a U.S. citizen, they are more likely to have health insurance than non-U.S. citizens.

Mosaic plot between Persons per Family and Health Insurance

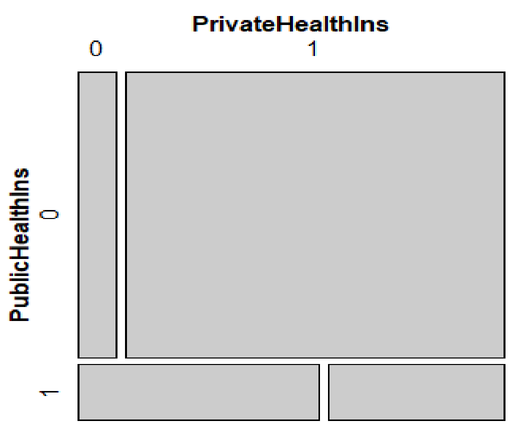
```
mosaic(~PersonsPerFamily+HealthIns, data = data)
```



The plot is between Persons per family and Health Insurance, we see that mostly tiles align to each other, telling us that persons per family does not have much effect on health insurance. If we look closely we can say that if the number of persons in a family is more than 4, chances of them not having health insurance increases.

Mosaic plot between Public Health Insurance and Private Health Insurance

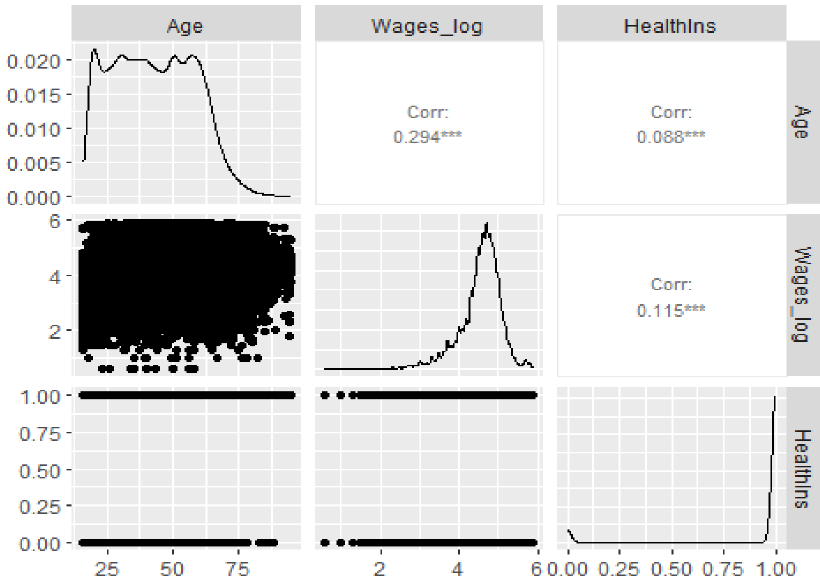
```
mosaic(~PublicHealthIns+PrivateHealthIns, data = data)
```



Bivariate Analysis on Continuous Variables

Correlation plot between Log of Wages, Health Insurance and Age

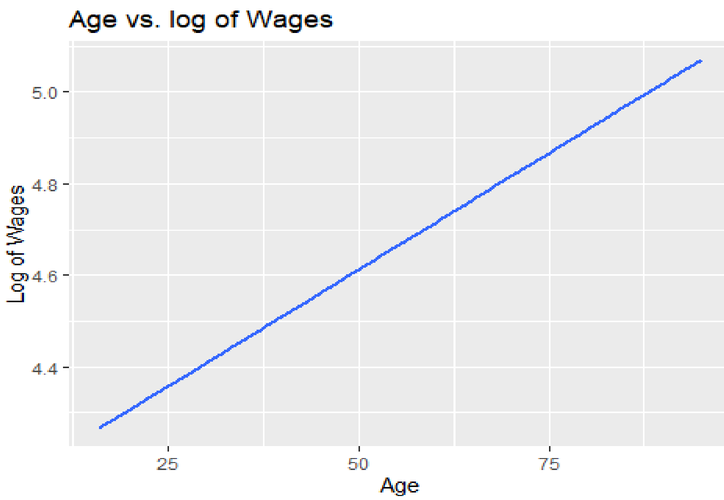
```
ggpairs(data[, c("Age", "Wages_log", "HealthIns")], columns = 1:3,
      upper = list(continuous = wrap("cor", size = 3)))
```



Using correlation plot, we can plot different scatter plot to see the relationship between each variable.

Scatter plot with linear regression line fit between Age and Health Insurance

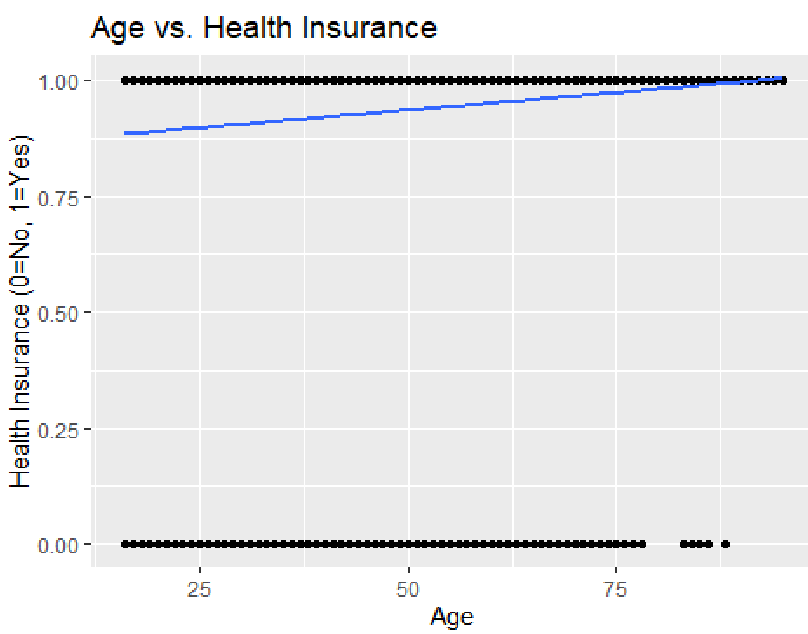
```
ggplot(data, aes(x=Age, y=Wages_log)) +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="Age vs. log of Wages", x="Age", y="Log of Wages")
```



This plot shows that as the age increases a person wages also increases linearly

Scatter plot with linear regression line fit between Age and Health Insurance

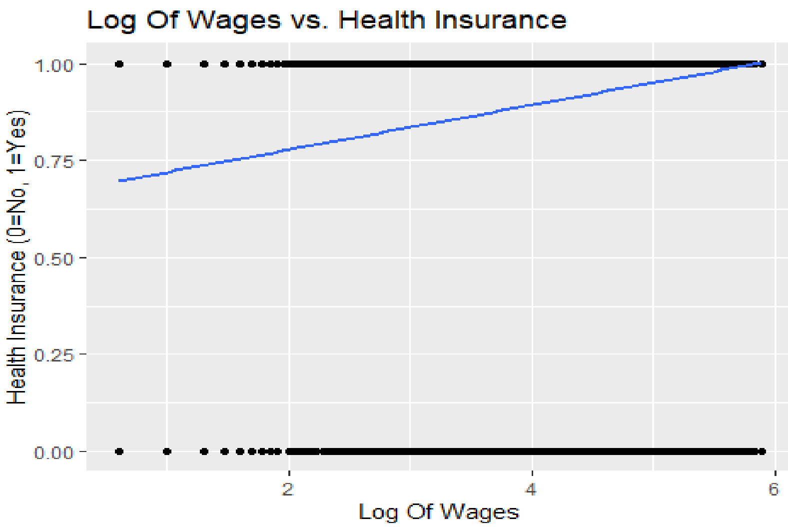
```
ggplot(data, aes(x=Age, y=HealthIns)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="Age vs. Health Insurance", x="Age", y="Health Insurance (0=No, 1=Yes)")
```



This plot shows that with increase in Age, the probability of having a health insurance increases slightly.

Scatter plot with linear regression line fit between log of Wages and Health Insurance

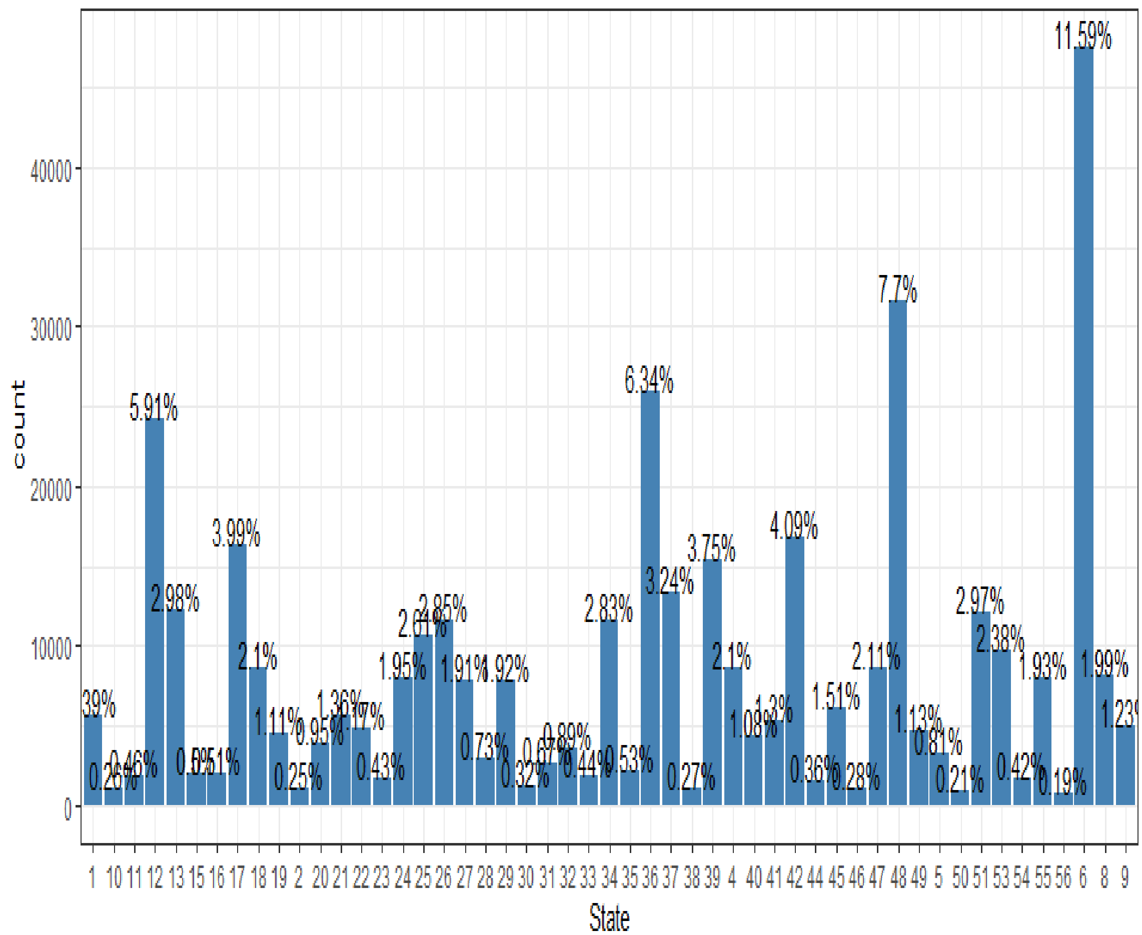
```
ggplot(data, aes(x=Wages_log, y=HealthIns)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  labs(title="Log Of Wages vs. Health Insurance", x="Log Of Wages", y="Health Insurance (0=No, 1=Yes)")
```



This plot shows that there is a linear increase trend between log of Wages and Health Insurance.

Percentage of people per State

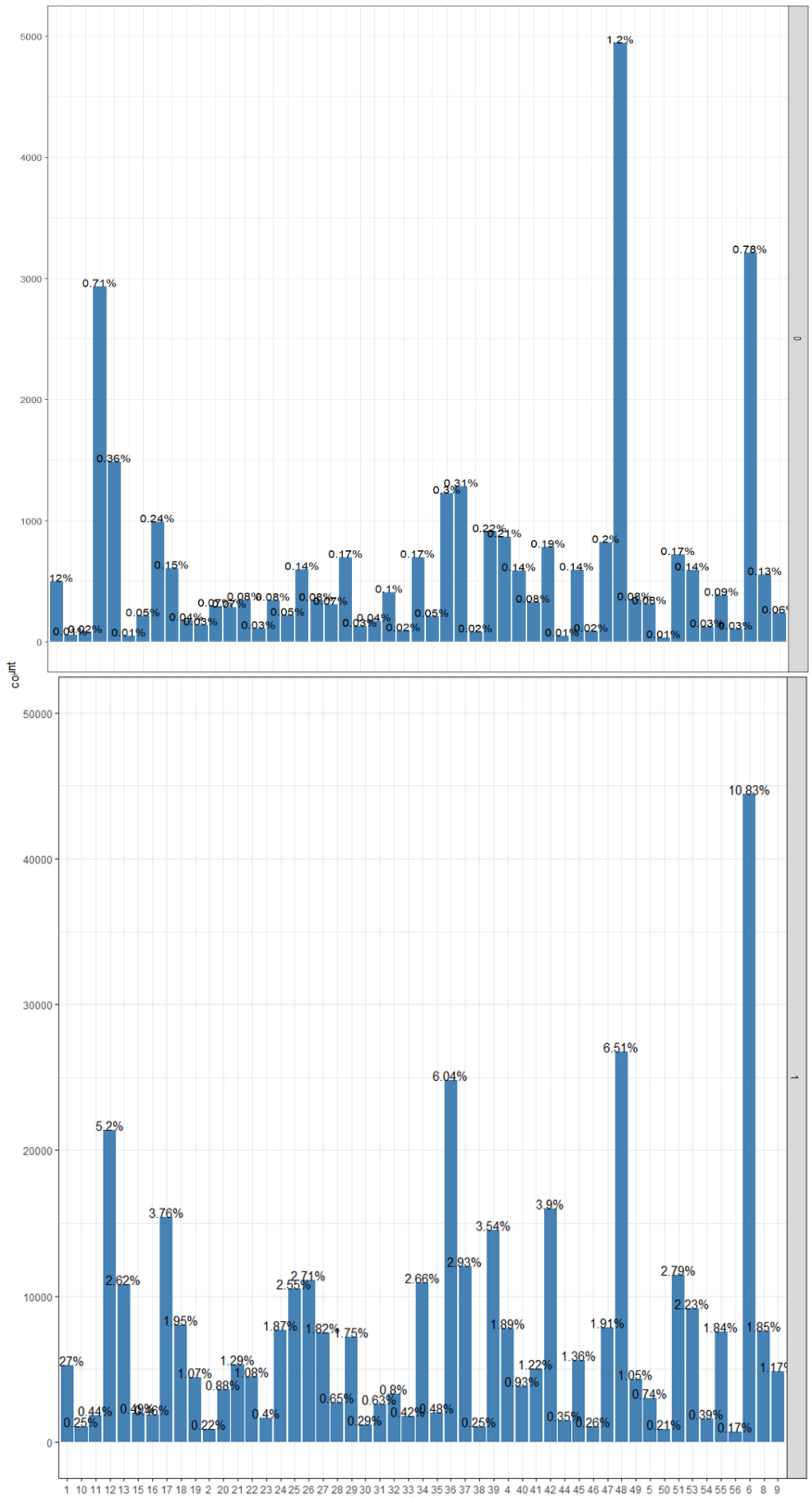
```
ggplot(data, aes(x=State)) +
  geom_bar(fill="steelblue")+
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2), "%")),vjust=0) +
  theme_bw()
```



This plot shows the population percentage of each state.

Percentage of people per State with and without health insurance

```
ggplot(data, aes(x=State, fill=HealthIns)) +
  geom_bar(fill="steelblue") +
  geom_text(stat='count',aes(label = paste0(round(..count../sum(..count..) * 100,2),
"%")),vjust=0) +
  facet_grid(data$HealthIns) +
  scale_y_continuous(limits = c(0, 50000)) +
  theme_bw()
```



count
State
0
1
0
1000
2000
3000
4000
5000
10000
20000
30000
40000
50000
1 10 11 12 13 15 16 17 18 19 2 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 4 40 41 42 44 45 46 47 48 49 5 50 51 53 54 55 56 6 8 9

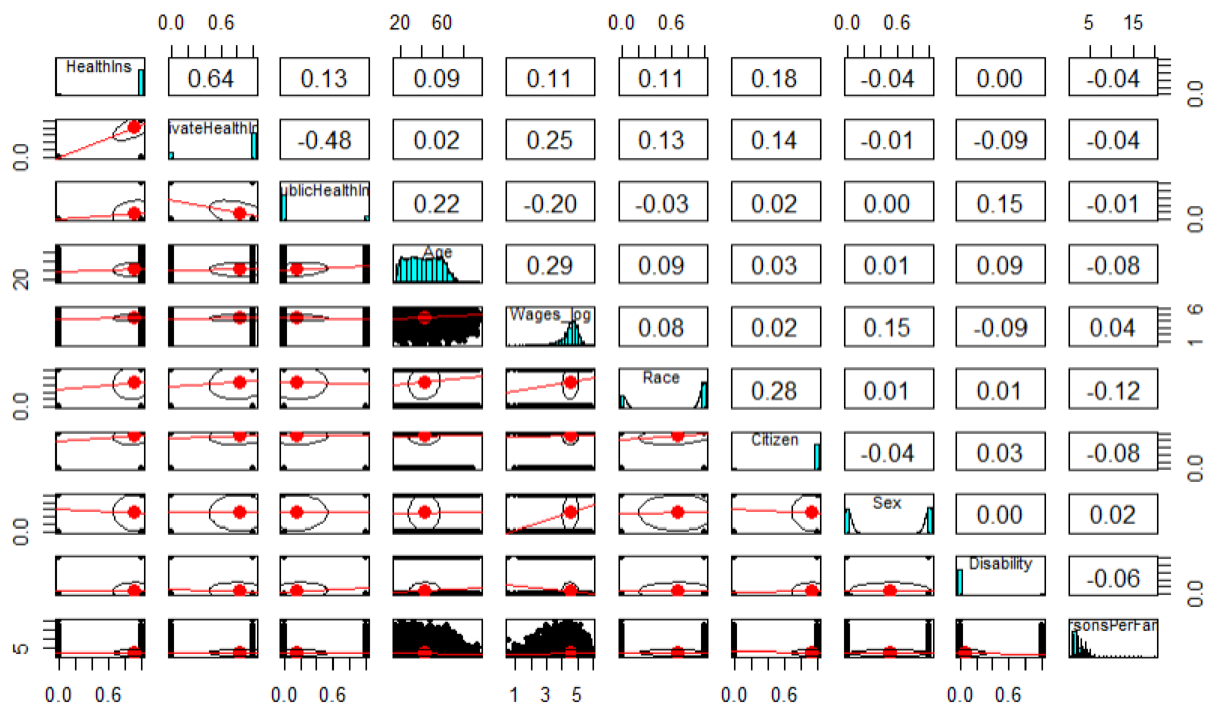
As we can look at this plot:

- Approximately out of 11.59% of the people living in California, 10.83% people have an Health Insurance.
- If we look at Texas, out of 7.7% population, 6.51% have an health Insurance.
- New york has a population of 6.34% of which 6.04% have a Health Insurance
- Florida has a population of 5.91% of which 5.23% have a Health Insurance
- The lowest population is of Wyoming State of 0.19% overall.

Correlation plot between all the variables

```
corrdata = with(data,data.frame(HealthIns, PrivateHealthIns, PublicHealthIns, Age, Wages_log, Race, Citizen, Sex, Disability, PersonsPerFamily))

pairs.panels(corrdata,lm=T)
```



- We can see that health insurance and private health insurance are highly correlated with a value of 0.64.
- The race and citizen correlation is 0.28
- The age and wages_log correlation is 0.29

- Health insurance and citizen correlation is 0.18.

So now we will try to build a linear regression model based on the all analysis we have done so far.

Dividing Age

The code re-codes the Age variable into three categories based on age ranges. This can be useful for simplifying the data and creating categories that are more meaningful or easier to interpret.

```
data$Age[data$Age <= 30 ] = 1
data$Age[data$Age > 30 & data$Age <=60 ] = 2
data$Age[data$Age > 60 ] = 3
```

Linear Regression on the Base Model

```
basemodel = lm(HealthIns ~
PersonsPerFamily+as.factor(Age)+Wages_log+Sex+Race+Citizen+Disability+as.factor(State),
data = data)
summary(basemodel)

##
## Call:
## lm(formula = HealthIns ~ PersonsPerFamily + as.factor(Age) +
##     Wages_log + Sex + Race + Citizen + Disability + as.factor(State),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07984  0.01946  0.05479  0.09626  0.46588
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.4947074  0.0051467  96.121  < 2e-16 ***
## PersonsPerFamily   -0.0012198  0.0002734  -4.462 8.11e-06 ***
## as.factor(Age)2     0.0030042  0.0010247   2.932 0.003370 **
## as.factor(Age)3     0.0452625  0.0013640  33.185  < 2e-16 ***
## Wages_log           0.0541306  0.0008460  63.986  < 2e-16 ***
```

```
## Sex                -0.0293113  0.0008121 -36.093  < 2e-16 ***
## Race                0.0314650  0.0009308  33.803  < 2e-16 ***
## Citizen             0.1682142  0.0017084  98.463  < 2e-16 ***
## Disability          0.0024175  0.0016023   1.509 0.131367
## as.factor(State)10  0.0355219  0.0084885   4.185 2.86e-05 ***
## as.factor(State)11  0.0525873  0.0068008   7.733 1.06e-14 ***
## as.factor(State)12 -0.0237353  0.0037665  -6.302 2.95e-10 ***
## as.factor(State)13 -0.0270402  0.0041016  -6.593 4.33e-11 ***
## as.factor(State)15  0.0897584  0.0065924  13.615  < 2e-16 ***
## as.factor(State)16 -0.0114667  0.0065540  -1.750 0.080192 .
## as.factor(State)17  0.0293170  0.0039333   7.454 9.10e-14 ***
## as.factor(State)18  0.0145309  0.0043677   3.327 0.000878 ***
## as.factor(State)19  0.0423402  0.0050801   8.335  < 2e-16 ***
## as.factor(State)2  -0.0372208  0.0086929  -4.282 1.85e-05 ***
## as.factor(State)20  0.0095238  0.0053181   1.791 0.073323 .
## as.factor(State)21  0.0341651  0.0048162   7.094 1.31e-12 ***
## as.factor(State)22  0.0135820  0.0050109   2.710 0.006719 **
## as.factor(State)23  0.0165342  0.0069989   2.362 0.018157 *
## as.factor(State)24  0.0445208  0.0044357  10.037  < 2e-16 ***
## as.factor(State)25  0.0666386  0.0041969  15.878  < 2e-16 ***
## as.factor(State)26  0.0343067  0.0041320   8.303  < 2e-16 ***
## as.factor(State)27  0.0352788  0.0044566   7.916 2.46e-15 ***
## as.factor(State)28 -0.0123255  0.0057799  -2.132 0.032969 *
## as.factor(State)29 -0.0061183  0.0044502  -1.375 0.169183
## as.factor(State)30 -0.0127161  0.0078771  -1.614 0.106458
## as.factor(State)31  0.0192108  0.0059515   3.228 0.001247 **
## as.factor(State)32 -0.0056206  0.0054175  -1.037 0.299511
## as.factor(State)33  0.0265126  0.0068905   3.848 0.000119 ***
## as.factor(State)34  0.0340733  0.0041405   8.229  < 2e-16 ***
## as.factor(State)35  0.0024374  0.0064432   0.378 0.705213
## as.factor(State)36  0.0467203  0.0037425  12.484  < 2e-16 ***
## as.factor(State)37 -0.0047124  0.0040483  -1.164 0.244406
## as.factor(State)38  0.0082772  0.0083486   0.991 0.321469
```

```
## as.factor(State)39  0.0236218  0.0039663   5.956 2.59e-09 ***
## as.factor(State)4  -0.0045493  0.0043676  -1.042 0.297605
## as.factor(State)40 -0.0404822  0.0051259  -7.898 2.85e-15 ***
## as.factor(State)41  0.0250740  0.0048713   5.147 2.64e-07 ***
## as.factor(State)42  0.0355812  0.0039224   9.071  < 2e-16 ***
## as.factor(State)44  0.0541207  0.0074626   7.252 4.11e-13 ***
## as.factor(State)45 -0.0054460  0.0046952  -1.160 0.246085
## as.factor(State)46  0.0065778  0.0082603   0.796 0.425851
## as.factor(State)47 -0.0078053  0.0043622  -1.789 0.073567 .
## as.factor(State)48 -0.0525751  0.0036837 -14.272  < 2e-16 ***
## as.factor(State)49  0.0189378  0.0050585   3.744 0.000181 ***
## as.factor(State)5  -0.0047725  0.0055766  -0.856 0.392104
## as.factor(State)50  0.0389673  0.0092601   4.208 2.58e-05 ***
## as.factor(State)51  0.0308459  0.0041054   7.513 5.77e-14 ***
## as.factor(State)53  0.0298859  0.0042666   7.005 2.48e-12 ***
## as.factor(State)54  0.0030005  0.0070313   0.427 0.669580
## as.factor(State)55  0.0316447  0.0044428   7.123 1.06e-12 ***
## as.factor(State)56 -0.0482908  0.0097582  -4.949 7.47e-07 ***
## as.factor(State)6   0.0432675  0.0035983  12.024  < 2e-16 ***
## as.factor(State)8   0.0164892  0.0044168   3.733 0.000189 ***
## as.factor(State)9   0.0380445  0.0049435   7.696 1.41e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2559 on 410528 degrees of freedom
## Multiple R-squared:  0.06811,    Adjusted R-squared:  0.06797
## F-statistic: 517.3 on 58 and 410528 DF,  p-value: < 2.2e-16
```

This is the first model that we ran for linear regression model on all are independent variables:

- The intercept represents the predicted value of the dependent variable when all independent variables are equal to zero. In this case, the intercept is not meaningful since it is highly unlikely that all the independent variables would be equal to zero.
- The coefficient estimate for “PersonsPerFamily” is -0.0023813, which means that, holding all other predictors constant, a one-unit increase in the number of people in

the family is associated with a decrease of 0.0023813 in the likelihood of having health insurance.

- R square (R^2) indicates the proportion of the variation in the HealthIns variable that is explained by the independent variables included in the model. Here the R^2 value is 0.06855, which means 6% variation in health insurance is explained by all the independent variables.
- F-Statistic: It indicates whether the model as a whole is significant in explaining the variation in the dependent variable. Here F-statistic is 524.6, which is very large, and the p-value is very small (less than 2.2e-16), providing strong evidence that the model is significant.

Modifications to the Linear Model

After running the full linear regression model, we can see that:

- Disability variable is insignificant and the coefficient is very low of 0.0022907, so we can remove this variable from our analysis.
- PersonPerFamily variable also has very low coefficient of -0.0023813, so we can remove it as well.

```
updatedModel_1 = lm(HealthIns ~
as.factor(Age)+Wages_log+Sex+Race+Citizen+as.factor(State), data = data)
summary(updatedModel_1)

##
## Call:
## lm(formula = HealthIns ~ as.factor(Age) + Wages_log + Sex + Race +
##     Citizen + as.factor(State), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07748  0.01959  0.05477  0.09634  0.46721
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.4919073  0.0050720  96.985  < 2e-16 ***
## as.factor(Age)2     0.0025082  0.0010156   2.470 0.013520 *
```

```
## as.factor(Age)3     0.0458952  0.0013537  33.902  < 2e-16 ***
## Wages_log           0.0539970  0.0008420  64.132  < 2e-16 ***
## Sex                -0.0293665  0.0008118 -36.177  < 2e-16 ***
## Race                0.0317839  0.0009280  34.250  < 2e-16 ***
## Citizen             0.1685340  0.0017069  98.735  < 2e-16 ***
## as.factor(State)10  0.0355969  0.0084887   4.193 2.75e-05 ***
## as.factor(State)11  0.0531422  0.0067998   7.815 5.50e-15 ***
## as.factor(State)12 -0.0236879  0.0037665  -6.289 3.20e-10 ***
## as.factor(State)13 -0.0270169  0.0041017  -6.587 4.50e-11 ***
## as.factor(State)15  0.0894655  0.0065923  13.571  < 2e-16 ***
## as.factor(State)16 -0.0117177  0.0065538  -1.788 0.073792 .
## as.factor(State)17  0.0292264  0.0039333   7.431 1.08e-13 ***
## as.factor(State)18  0.0145059  0.0043678   3.321 0.000897 ***
## as.factor(State)19  0.0424078  0.0050802   8.348  < 2e-16 ***
## as.factor(State)2  -0.0369874  0.0086930  -4.255 2.09e-05 ***
## as.factor(State)20  0.0095632  0.0053182   1.798 0.072146 .
## as.factor(State)21  0.0341700  0.0048164   7.095 1.30e-12 ***
## as.factor(State)22  0.0137499  0.0050109   2.744 0.006070 **
## as.factor(State)23  0.0167849  0.0069988   2.398 0.016474 *
## as.factor(State)24  0.0444817  0.0044358  10.028  < 2e-16 ***
## as.factor(State)25  0.0666974  0.0041969  15.892  < 2e-16 ***
## as.factor(State)26  0.0343320  0.0041321   8.309  < 2e-16 ***
## as.factor(State)27  0.0352707  0.0044567   7.914 2.50e-15 ***
## as.factor(State)28 -0.0122388  0.0057800  -2.117 0.034224 *
## as.factor(State)29 -0.0060928  0.0044503  -1.369 0.170975
## as.factor(State)30 -0.0126091  0.0078772  -1.601 0.109442
## as.factor(State)31  0.0192159  0.0059516   3.229 0.001244 **
## as.factor(State)32 -0.0056760  0.0054176  -1.048 0.294784
## as.factor(State)33  0.0265966  0.0068906   3.860 0.000113 ***
## as.factor(State)34  0.0337966  0.0041402   8.163 3.27e-16 ***
## as.factor(State)35  0.0025710  0.0064433   0.399 0.689878
## as.factor(State)36  0.0466675  0.0037425  12.470  < 2e-16 ***
## as.factor(State)37 -0.0045422  0.0040482  -1.122 0.261849
```

```
## as.factor(State)38  0.0082978  0.0083488   0.994 0.320273
## as.factor(State)39  0.0236524  0.0039664   5.963 2.47e-09 ***
## as.factor(State)4  -0.0045598  0.0043677  -1.044 0.296502
## as.factor(State)40 -0.0404058  0.0051259  -7.883 3.21e-15 ***
## as.factor(State)41  0.0251638  0.0048714   5.166 2.40e-07 ***
## as.factor(State)42  0.0355226  0.0039225   9.056  < 2e-16 ***
## as.factor(State)44  0.0541539  0.0074627   7.257 3.98e-13 ***
## as.factor(State)45 -0.0052769  0.0046951  -1.124 0.261052
## as.factor(State)46  0.0066098  0.0082605   0.800 0.423611
## as.factor(State)47 -0.0077352  0.0043623  -1.773 0.076199 .
## as.factor(State)48 -0.0526944  0.0036837 -14.305  < 2e-16 ***
## as.factor(State)49  0.0183977  0.0050572   3.638 0.000275 ***
## as.factor(State)5  -0.0048198  0.0055767  -0.864 0.387439
## as.factor(State)50  0.0391966  0.0092601   4.233 2.31e-05 ***
## as.factor(State)51  0.0309232  0.0041055   7.532 5.00e-14 ***
## as.factor(State)53  0.0299995  0.0042666   7.031 2.05e-12 ***
## as.factor(State)54  0.0029798  0.0070315   0.424 0.671730
## as.factor(State)55  0.0316359  0.0044429   7.121 1.08e-12 ***
## as.factor(State)56 -0.0482483  0.0097584  -4.944 7.65e-07 ***
## as.factor(State)6   0.0430033  0.0035980  11.952  < 2e-16 ***
## as.factor(State)8   0.0166084  0.0044168   3.760 0.000170 ***
## as.factor(State)9   0.0380196  0.0049436   7.691 1.47e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2559 on 410530 degrees of freedom
## Multiple R-squared:  0.06805,    Adjusted R-squared:  0.06793
## F-statistic: 535.3 on 56 and 410530 DF,  p-value: < 2.2e-16
```

We can see after removing those variables, the adjusted r-square value has increased to 0.06835, which we can say is a better fit.

Further Modifications to the Linear Model

After running the updated linear regression model, we can do the following to increase accuracy:

- As Age increase, we saw that the wage increases. So it would be better to add an interaction between the variables.
- Similarly, for Race and Citizenship there is a strong relationship, so we added an interaction.

```
updatedModel_2 = lm(HealthIns ~
(as.factor(Age)*Wages_log)+Sex+(Race*Citizen)+as.factor(State), data = data)
summary(updatedModel_1)

##
## Call:
## lm(formula = HealthIns ~ as.factor(Age) + Wages_log + Sex + Race +
##     Citizen + as.factor(State), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07748  0.01959  0.05477  0.09634  0.46721
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.4919073  0.0050720  96.985  < 2e-16 ***
## as.factor(Age)2     0.0025082  0.0010156   2.470 0.013520 *
## as.factor(Age)3     0.0458952  0.0013537  33.902  < 2e-16 ***
## Wages_log           0.0539970  0.0008420  64.132  < 2e-16 ***
## Sex                -0.0293665  0.0008118 -36.177  < 2e-16 ***
## Race                0.0317839  0.0009280  34.250  < 2e-16 ***
## Citizen             0.1685340  0.0017069  98.735  < 2e-16 ***
## as.factor(State)10  0.0355969  0.0084887   4.193 2.75e-05 ***
## as.factor(State)11  0.0531422  0.0067998   7.815 5.50e-15 ***
## as.factor(State)12 -0.0236879  0.0037665  -6.289 3.20e-10 ***
## as.factor(State)13 -0.0270169  0.0041017  -6.587 4.50e-11 ***
## as.factor(State)15  0.0894655  0.0065923  13.571  < 2e-16 ***
## as.factor(State)16 -0.0117177  0.0065538  -1.788 0.073792 .
## as.factor(State)17  0.0292264  0.0039333   7.431 1.08e-13 ***
## as.factor(State)18  0.0145059  0.0043678   3.321 0.000897 ***
```

```
## as.factor(State)19  0.0424078  0.0050802   8.348  < 2e-16 ***
## as.factor(State)2  -0.0369874  0.0086930  -4.255 2.09e-05 ***
## as.factor(State)20  0.0095632  0.0053182   1.798 0.072146 .
## as.factor(State)21  0.0341700  0.0048164   7.095 1.30e-12 ***
## as.factor(State)22  0.0137499  0.0050109   2.744 0.006070 **
## as.factor(State)23  0.0167849  0.0069988   2.398 0.016474 *
## as.factor(State)24  0.0444817  0.0044358  10.028  < 2e-16 ***
## as.factor(State)25  0.0666974  0.0041969  15.892  < 2e-16 ***
## as.factor(State)26  0.0343320  0.0041321   8.309  < 2e-16 ***
## as.factor(State)27  0.0352707  0.0044567   7.914 2.50e-15 ***
## as.factor(State)28 -0.0122388  0.0057800  -2.117 0.034224 *
## as.factor(State)29 -0.0060928  0.0044503  -1.369 0.170975
## as.factor(State)30 -0.0126091  0.0078772  -1.601 0.109442
## as.factor(State)31  0.0192159  0.0059516   3.229 0.001244 **
## as.factor(State)32 -0.0056760  0.0054176  -1.048 0.294784
## as.factor(State)33  0.0265966  0.0068906   3.860 0.000113 ***
## as.factor(State)34  0.0337966  0.0041402   8.163 3.27e-16 ***
## as.factor(State)35  0.0025710  0.0064433   0.399 0.689878
## as.factor(State)36  0.0466675  0.0037425  12.470  < 2e-16 ***
## as.factor(State)37 -0.0045422  0.0040482  -1.122 0.261849
## as.factor(State)38  0.0082978  0.0083488   0.994 0.320273
## as.factor(State)39  0.0236524  0.0039664   5.963 2.47e-09 ***
## as.factor(State)4  -0.0045598  0.0043677  -1.044 0.296502
## as.factor(State)40 -0.0404058  0.0051259  -7.883 3.21e-15 ***
## as.factor(State)41  0.0251638  0.0048714   5.166 2.40e-07 ***
## as.factor(State)42  0.0355226  0.0039225   9.056  < 2e-16 ***
## as.factor(State)44  0.0541539  0.0074627   7.257 3.98e-13 ***
## as.factor(State)45 -0.0052769  0.0046951  -1.124 0.261052
## as.factor(State)46  0.0066098  0.0082605   0.800 0.423611
## as.factor(State)47 -0.0077352  0.0043623  -1.773 0.076199 .
## as.factor(State)48 -0.0526944  0.0036837 -14.305  < 2e-16 ***
## as.factor(State)49  0.0183977  0.0050572   3.638 0.000275 ***
## as.factor(State)5  -0.0048198  0.0055767  -0.864 0.387439
```

```
## as.factor(State)50  0.0391966  0.0092601   4.233 2.31e-05 ***
## as.factor(State)51  0.0309232  0.0041055   7.532 5.00e-14 ***
## as.factor(State)53  0.0299995  0.0042666   7.031 2.05e-12 ***
## as.factor(State)54  0.0029798  0.0070315   0.424 0.671730
## as.factor(State)55  0.0316359  0.0044429   7.121 1.08e-12 ***
## as.factor(State)56 -0.0482483  0.0097584  -4.944 7.65e-07 ***
## as.factor(State)6   0.0430033  0.0035980  11.952  < 2e-16 ***
## as.factor(State)8   0.0166084  0.0044168   3.760 0.000170 ***
## as.factor(State)9   0.0380196  0.0049436   7.691 1.47e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2559 on 410530 degrees of freedom
## Multiple R-squared:  0.06805,	Adjusted R-squared:  0.06793
## F-statistic: 535.3 on 56 and 410530 DF,  p-value: < 2.2e-16
```

We can see after adding interactions between those variables, the adjusted r-square value has increased to 0.07619, which we can say is a better fit.

Interaction model with PrivateHealthIns as the response variable

```
updatedModel_3 = lm(PrivateHealthIns ~
(as.factor(Age)*Wages_log)+Sex+(Race*Citizen)+as.factor(State), data = data)
summary(updatedModel_3)

##
## Call:
## lm(formula = PrivateHealthIns ~ (as.factor(Age) * Wages_log) +
##     Sex + (Race * Citizen) + as.factor(State), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23965  0.01536  0.11974  0.19692  1.21765
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.854e-01  9.524e-03  29.966  < 2e-16 ***
```

```
## as.factor(Age)2          -7.981e-01  1.120e-02 -71.293  < 2e-16 ***
## as.factor(Age)3          -5.168e-01  1.524e-02 -33.918  < 2e-16 ***
## Wages_log                 8.637e-02  1.879e-03  45.958  < 2e-16 ***
## Sex                      -3.883e-02  1.124e-03 -34.547  < 2e-16 ***
## Race                      1.241e-01  5.721e-03  21.687  < 2e-16 ***
## Citizen                   1.686e-01  2.639e-03  63.863  < 2e-16 ***
## as.factor(State)10       -8.824e-03  1.174e-02  -0.752  0.45230
## as.factor(State)11        9.499e-03  9.406e-03   1.010  0.31255
## as.factor(State)12       -5.189e-02  5.209e-03  -9.960  < 2e-16 ***
## as.factor(State)13       -3.170e-02  5.673e-03  -5.588 2.30e-08 ***
## as.factor(State)15        8.811e-02  9.118e-03   9.664  < 2e-16 ***
## as.factor(State)16       -5.414e-02  9.064e-03  -5.973 2.34e-09 ***
## as.factor(State)17       -1.530e-02  5.440e-03  -2.812  0.00493 **
## as.factor(State)18       -3.391e-02  6.041e-03  -5.614 1.98e-08 ***
## as.factor(State)19       -1.835e-03  7.027e-03  -0.261  0.79397
## as.factor(State)2        -1.078e-01  1.202e-02  -8.962  < 2e-16 ***
## as.factor(State)20        2.249e-05  7.356e-03   0.003  0.99756
## as.factor(State)21       -3.461e-02  6.661e-03  -5.196 2.04e-07 ***
## as.factor(State)22       -7.375e-02  6.930e-03 -10.642  < 2e-16 ***
## as.factor(State)23       -4.412e-02  9.680e-03  -4.558 5.16e-06 ***
## as.factor(State)24       -3.082e-03  6.136e-03  -0.502  0.61543
## as.factor(State)25        4.215e-03  5.805e-03   0.726  0.46776
## as.factor(State)26       -2.087e-02  5.715e-03  -3.652  0.00026 ***
## as.factor(State)27       -2.833e-02  6.164e-03  -4.596 4.30e-06 ***
## as.factor(State)28       -4.344e-03  7.994e-03  -0.543  0.58686
## as.factor(State)29       -2.910e-02  6.155e-03  -4.727 2.28e-06 ***
## as.factor(State)30       -8.531e-02  1.089e-02  -7.830 4.87e-15 ***
## as.factor(State)31       -4.703e-03  8.232e-03  -0.571  0.56775
## as.factor(State)32       -4.437e-02  7.493e-03  -5.922 3.18e-09 ***
## as.factor(State)33       -1.424e-02  9.530e-03  -1.494  0.13516
## as.factor(State)34       -7.832e-03  5.727e-03  -1.368  0.17144
## as.factor(State)35       -1.164e-01  8.912e-03 -13.058  < 2e-16 ***
## as.factor(State)36       -3.835e-02  5.176e-03  -7.408 1.29e-13 ***
```

```
## as.factor(State)37        -2.548e-02  5.599e-03  -4.551 5.33e-06 ***
## as.factor(State)38        -2.504e-03  1.155e-02  -0.217  0.82831
## as.factor(State)39        -3.347e-02  5.486e-03  -6.101 1.05e-09 ***
## as.factor(State)4         -6.896e-02  6.041e-03 -11.415  < 2e-16 ***
## as.factor(State)40        -5.587e-02  7.089e-03  -7.880 3.28e-15 ***
## as.factor(State)41        -5.036e-02  6.738e-03  -7.475 7.72e-14 ***
## as.factor(State)42        -4.782e-03  5.425e-03  -0.881  0.37805
## as.factor(State)44        -1.937e-02  1.032e-02  -1.877  0.06057 .
## as.factor(State)45        -2.993e-02  6.494e-03  -4.610 4.03e-06 ***
## as.factor(State)46        -2.438e-03  1.142e-02  -0.213  0.83100
## as.factor(State)47        -3.554e-02  6.033e-03  -5.890 3.85e-09 ***
## as.factor(State)48        -6.266e-02  5.095e-03 -12.300  < 2e-16 ***
## as.factor(State)49         1.292e-02  6.995e-03   1.847  0.06475 .
## as.factor(State)5         -7.578e-02  7.713e-03  -9.825  < 2e-16 ***
## as.factor(State)50        -3.351e-02  1.281e-02  -2.617  0.00888 **
## as.factor(State)51         3.134e-03  5.678e-03   0.552  0.58102
## as.factor(State)53        -1.526e-02  5.901e-03  -2.587  0.00969 **
## as.factor(State)54        -7.405e-02  9.725e-03  -7.615 2.65e-14 ***
## as.factor(State)55        -1.049e-02  6.145e-03  -1.706  0.08795 .
## as.factor(State)56        -7.989e-02  1.350e-02  -5.919 3.24e-09 ***
## as.factor(State)6         -4.033e-02  4.977e-03  -8.103 5.37e-16 ***
## as.factor(State)8         -4.238e-02  6.109e-03  -6.937 4.02e-12 ***
## as.factor(State)9         -4.584e-02  6.838e-03  -6.703 2.04e-11 ***
## as.factor(Age)2:Wages_log  1.758e-01  2.513e-03  69.960  < 2e-16 ***
## as.factor(Age)3:Wages_log  1.003e-01  3.401e-03  29.480  < 2e-16 ***
## Race:Citizen              -6.468e-02  5.859e-03 -11.039  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3539 on 410527 degrees of freedom
## Multiple R-squared:  0.1094, Adjusted R-squared:  0.1092
## F-statistic: 854.4 on 59 and 410527 DF,  p-value: < 2.2e-16
```

In this we can see the r-squared increase to 0.1092.

Interaction model with PublicHealthIns as the response variable

```
updatedModel_4 = lm(PublicHealthIns ~
(as.factor(Age)*Wages_log)+Sex+(Race*Citizen)+as.factor(State), data = data)
summary(updatedModel_4)

##
## Call:
## lm(formula = PublicHealthIns ~ (as.factor(Age) * Wages_log) +
##     Sex + (Race * Citizen) + as.factor(State), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30950 -0.14534 -0.08661 -0.02278  1.14115
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.4298253  0.0089168  48.204  < 2e-16 ***
## as.factor(Age)2        0.4438008  0.0104814  42.342  < 2e-16 ***
## as.factor(Age)3        1.1955821  0.0142653  83.811  < 2e-16 ***
## Wages_log             -0.0891610  0.0017594 -50.676  < 2e-16 ***
## Sex                    0.0243804  0.0010523  23.169  < 2e-16 ***
## Race                  -0.0047735  0.0053563  -0.891  0.37282
## Citizen                0.0458443  0.0024711  18.552  < 2e-16 ***
## as.factor(State)10     0.0455645  0.0109918   4.145 3.39e-05 ***
## as.factor(State)11     0.0457399  0.0088064   5.194 2.06e-07 ***
## as.factor(State)12     0.0181986  0.0048772   3.731  0.00019 ***
## as.factor(State)13    -0.0006763  0.0053112  -0.127  0.89868
## as.factor(State)15     0.0218702  0.0085365   2.562  0.01041 *
## as.factor(State)16     0.0430004  0.0084864   5.067 4.04e-07 ***
## as.factor(State)17     0.0340958  0.0050933   6.694 2.17e-11 ***
## as.factor(State)18     0.0360619  0.0056558   6.376 1.82e-10 ***
## as.factor(State)19     0.0453907  0.0065785   6.900 5.21e-12 ***
## as.factor(State)2      0.0722656  0.0112565   6.420 1.36e-10 ***
## as.factor(State)20     0.0012058  0.0068866   0.175  0.86100
```

```
## as.factor(State)21         0.0754725  0.0062367  12.101  < 2e-16 ***
## as.factor(State)22         0.0978001  0.0064885  15.073  < 2e-16 ***
## as.factor(State)23         0.0543351  0.0090628   5.995 2.03e-09 ***
## as.factor(State)24         0.0568487  0.0057444   9.896  < 2e-16 ***
## as.factor(State)25         0.0634553  0.0054351  11.675  < 2e-16 ***
## as.factor(State)26         0.0501362  0.0053506   9.370  < 2e-16 ***
## as.factor(State)27         0.0603101  0.0057712  10.450  < 2e-16 ***
## as.factor(State)28         0.0011644  0.0074846   0.156  0.87637
## as.factor(State)29         0.0102955  0.0057626   1.787  0.07400 .
## as.factor(State)30         0.0633757  0.0102000   6.213 5.19e-10 ***
## as.factor(State)31         0.0157555  0.0077068   2.044  0.04092 *
## as.factor(State)32         0.0423349  0.0070152   6.035 1.59e-09 ***
## as.factor(State)33         0.0354804  0.0089226   3.976 7.00e-05 ***
## as.factor(State)34         0.0300960  0.0053615   5.613 1.99e-08 ***
## as.factor(State)35         0.1259792  0.0083435  15.099  < 2e-16 ***
## as.factor(State)36         0.0803970  0.0048463  16.589  < 2e-16 ***
## as.factor(State)37         0.0213434  0.0052419   4.072 4.67e-05 ***
## as.factor(State)38         0.0142689  0.0108107   1.320  0.18687
## as.factor(State)39         0.0469128  0.0051361   9.134  < 2e-16 ***
## as.factor(State)4          0.0601224  0.0056557  10.630  < 2e-16 ***
## as.factor(State)40         0.0153249  0.0066374   2.309  0.02095 *
## as.factor(State)41         0.0769671  0.0063079  12.202  < 2e-16 ***
## as.factor(State)42         0.0366509  0.0050794   7.216 5.38e-13 ***
## as.factor(State)44         0.0649712  0.0096633   6.723 1.78e-11 ***
## as.factor(State)45         0.0293345  0.0060795   4.825 1.40e-06 ***
## as.factor(State)46         0.0070870  0.0106963   0.663  0.50761
## as.factor(State)47         0.0252151  0.0056487   4.464 8.05e-06 ***
## as.factor(State)48         0.0012979  0.0047700   0.272  0.78555
## as.factor(State)49        -0.0079316  0.0065485  -1.211  0.22582
## as.factor(State)5          0.0728183  0.0072212  10.084  < 2e-16 ***
## as.factor(State)50         0.0703260  0.0119907   5.865 4.49e-09 ***
## as.factor(State)51         0.0322256  0.0053164   6.062 1.35e-09 ***
## as.factor(State)53         0.0472099  0.0055249   8.545  < 2e-16 ***
```

```
## as.factor(State)54        0.0901061  0.0091050   9.896  < 2e-16 ***
## as.factor(State)55        0.0307029  0.0057532   5.337 9.47e-08 ***
## as.factor(State)56        0.0327588  0.0126359   2.593  0.00953 **
## as.factor(State)6         0.0764596  0.0046592  16.410  < 2e-16 ***
## as.factor(State)8         0.0498156  0.0057192   8.710  < 2e-16 ***
## as.factor(State)9         0.0680460  0.0064022  10.629  < 2e-16 ***
## as.factor(Age)2:Wages_log -0.0927885  0.0023531 -39.433  < 2e-16 ***
## as.factor(Age)3:Wages_log -0.1761929  0.0031837 -55.342  < 2e-16 ***
## Race:Citizen             -0.0282414  0.0054858  -5.148 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3314 on 410527 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:   0.19
## F-statistic:  1633 on 59 and 410527 DF,  p-value: < 2.2e-16
```

In this we can see the r-squared increase to 0.19

We can say that this model best fits with Public Health insurance.

Results:

After running the regression analysis, we came across the best fitted model as:

lm(formula = HealthIns ~ (as.factor(Age) * Wages_log) + Sex + (Race * Citizen) + as.factor(State), data = data)

As per the hypothesis we started with, we can conclude from the regression output that Wages are positively correlated with Health Insurance as the coefficient is 0.0539970, which means that, holding all other predictors constant, a one-unit increase in the Wages is associated with an increase of 0. 0539970 in the likelihood of having health insurance.

By looking at the best fit model, we can conclude that persons per family doesn't play an efficient role in determining whether a person will have health insurance coverage, as it is not included in the final model.

The coefficient of -0.0293665 for Sex indicates that women are less likely to have health insurance than men, holding all other variables in the model constant. More specifically, for a one-unit increase in the Sex variable from 0 (female) to 1 (male), the predicted log-odds of having health insurance decrease by 0.0293665 units, or about 2.93%. This means that after controlling for age, wages, race, citizenship status, and state of residence, women have lower odds of having health insurance compared to men.

If we see the public health insurance model, it has a coefficient of 0.0243804 for sex, which means it is more likely that women will have a public health insurance compared to private health insurance.

Looking at the coefficients of the age variable in the three models, we can see that age has a significant effect on all three types of insurance.

- In the model for HealthIns, we see that the coefficient estimates for as.factor(Age)2 and as.factor(Age)3 are both positive and significant, indicating that individuals in the age groups 31-60 and over 60 are more likely to have health insurance than those in the youngest age group (under 30).
- In the model for PrivateHealthIns, we see that the coefficient estimates for as.factor(Age)2 and as.factor(Age)3 are both negative and significant, indicating that individuals in the age groups 31-60 and over 60 are less likely to have private health insurance than those in the youngest age group (under 30).
- In the model for PublicHealthIns, we see that the coefficient estimates for as.factor(Age)2 and as.factor(Age)3 are both positive and significant, indicating that individuals in the age groups 31-60 and over 60 are more likely to have public health insurance than those in the youngest age group (under 30).

Therefore, age seems to have a complex relationship with different types of health insurance, with older individuals having a higher likelihood of having health insurance in general, but a lower likelihood of having private health insurance specifically.

Conclusion:

Although the variables used in the analysis are informative in predicting health insurance coverage, the results suggest that they may not be sufficient for accurate predictions.

The analysis aimed to investigate the impact of income on health insurance coverage and how different socioeconomic factors affects the likelihood of having health insurance. While the data provided some insights, there were limitations to the accuracy of the models.

To improve accuracy, future analyses may need to incorporate survey weights and explore alternative modeling techniques such as logistic regression or decision trees.