1. What is the purpose of census income survey?

Census Income survey collected the data and it is used to measure an economics and well being of the population. In this case, it is the US Population between year 1994 and 1995.

2. What are the uses of census income survey?

The census income survey is used along with other specifics datas to research about economic cases. For example, a project “Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador” com. The data could be analyzed to **research on trends** and **construct the forecast** of the direction of the countries’ economics.

From the survey, it can be used to be a **basis for decision making** to **set policies** and **planning** for various fields and levels.

3. Identify type for each attribute (**NOMINAL** , **ORDINAL**, **INTERVAL**, **RATIO**) in the data file.

Attribute Name	**Description**	**Attribute Type**
age	Age of the worker	**RATIO**
class_worker	Class of worker	**NOMINAL**
ind_code	Industry code	**NOMINAL**
occ_code	Occupation code	**NOMINAL**
education	Level of education	**ORDINAL**
wage_per_hour	Wage per hour	**RATIO**
hs_college	Enrolled in educational institution last week	**NOMINAL**
marital_stat	Marital status	**NOMINAL**
major_ind_code	Major industry code	**NOMINAL**
major_occ_code	Major occupation code	**NOMINAL**
Wrace	Race	**NOMINAL**
hisp_origin	Hispanic origin	**NOMINAL**
sex	Sex	**NOMINAL**
union_member	Member of a labor union	**NOMINAL**
unemp_reason	Reason for unemployment	**NOMINAL**
full_or_part_emp	Full- or part-time employment status	**NOMINAL**
capital_gains	Capital gains	**RATIO**
capital_losses	Capital losses	**RATIO**
stock_dividends	Dividends from stocks	**RATIO**
tax_filer_stat	Tax filer status	**NOMINAL**
region_prev_res	Region of previous residence	**NOMINAL**
state_prev_res	State of previous residence	**NOMINAL**
det_hh_fam_stat	Detailed household and family	**NOMINAL**

	status	
det_hh_summ	Detailed household summary in household	NOMINAL
mig_chg_msa	Migration code - change in MSA	NOMINAL
mig_chg_reg	Migration code - change in region	NOMINAL
mig_move_reg	Migration code - move within region	NOMINAL
mig_same	Live in this house one year ago	NOMINAL
mig_prev_sunbelt	Migration - previous residence in sunbelt	NOMINAL
num_emp	Number of persons that worked for employer	RATIO
fam_under_18	Family members under 18	NOMINAL
country_father	Country of birth father	NOMINAL
country_mother	Country of birth mother	NOMINAL
country_self	Country of birth	NOMINAL
citizenship	Citizenship	NOMINAL
own_or_self	Own business of self-employed	RATIO
vet_question	Fill included questionnaire for Veterans Administration	NOMINAL
vet_benefits	Veterans benefits	RATIO
weeks_worked	Weeks worked in the year	INTERVAL
year	Year of survey	INTERVAL
income _50k	Income less than or greater than $50,000	NOMINAL

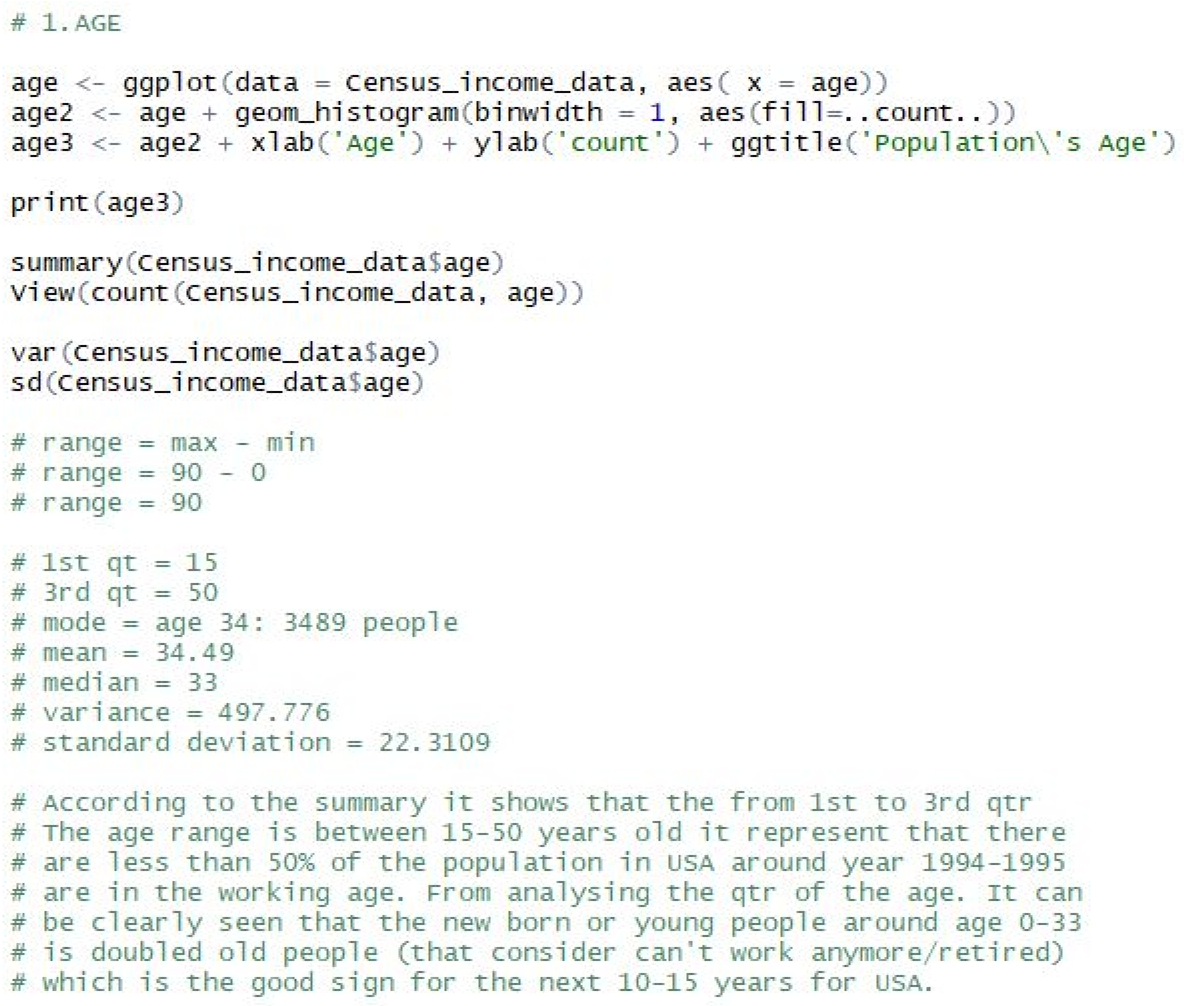
4. Verify data quality:

- implement data cleaning approaches to census-income data in R
- Explain what you do to clean the data.
- Provide the R codes

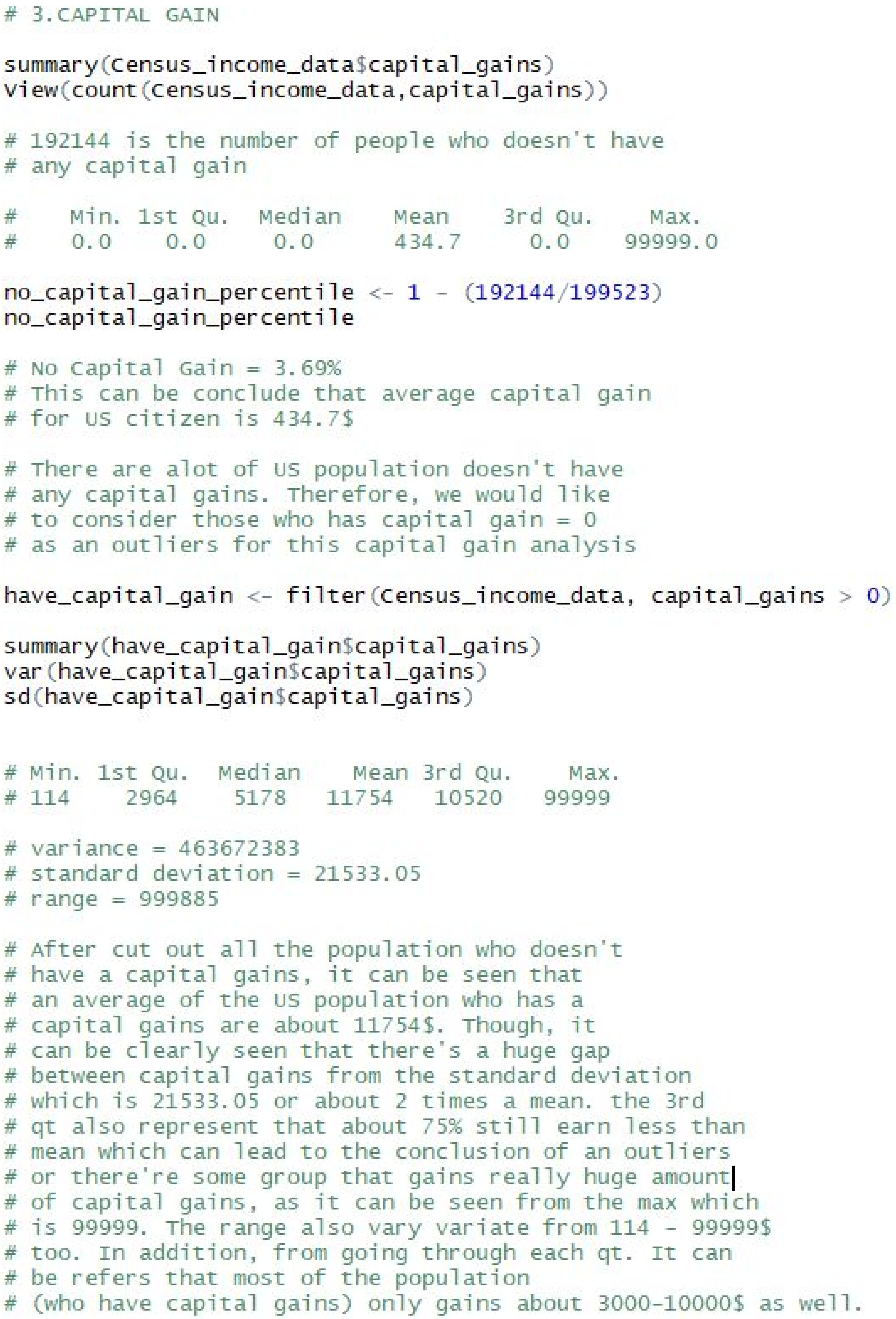
For the data quality, which are outliers, missing values, inconsistent values, and duplicated data. We did checking all if there’s any missing values exist in the given dataset. However, for duplicated data, we don’t see any point of checking it since in the given dataset there’s no unique identifiers for any rows. It also possible that we might have to combine all the attributes in order to create one unique identifier for the data, though we won’t use all of the following attributes anyway. Therefore, checking duplicated data is irrelevant.

Next, for an outliers and inconsistent values, most of them actually depends on an attribute itself and how we are going to use it. Therefore, we have look through all of an attributes as well as did some basic visualizing it in our R code, “DataQuality&Analysis.R”, feel free to read it, but insert all of them into a document wouldn’t be a good idea.

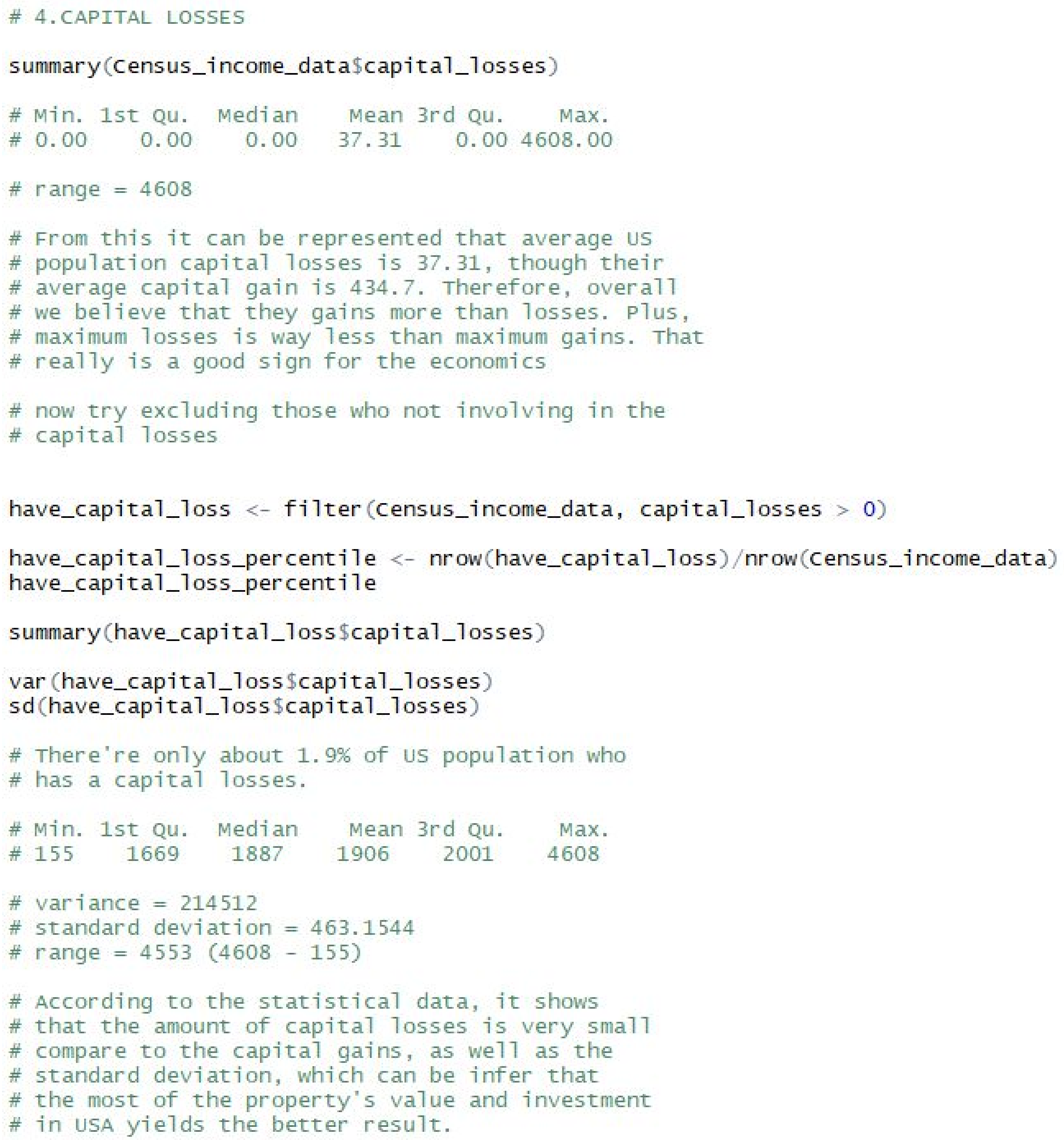
5. Give simple appropriate statistics (range, mode, mean, median, variance, counts, etc.) for 10 most important attributes and describe what they mean or if you found something interesting. (The file name “AppropriateStatistics.R”)



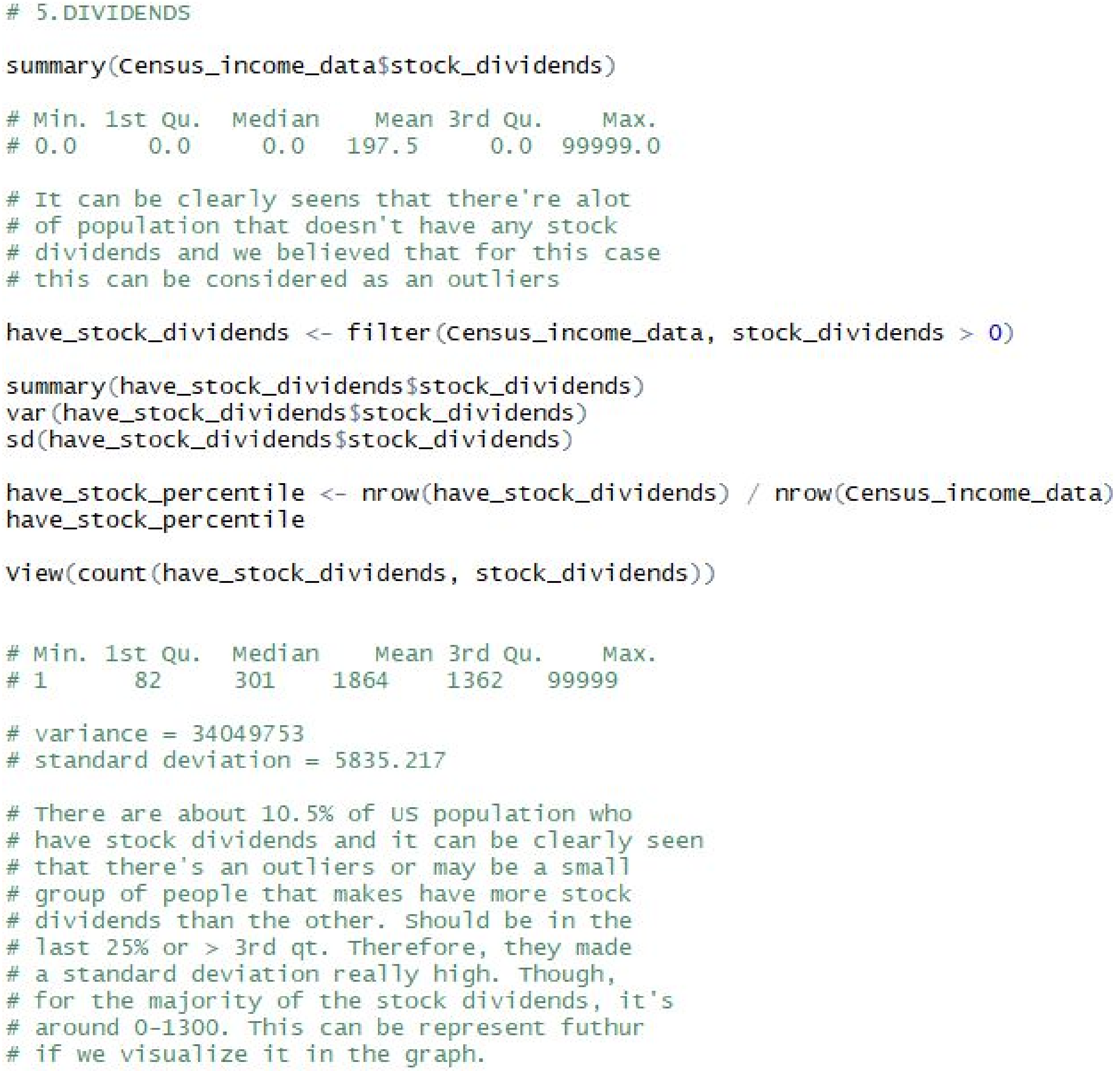
```
# 1.AGE

age <- ggplot(data = Census_income_data, aes( x = age))
age2 <- age + geom_histogram(binwidth = 1, aes(fill=..count..))
age3 <- age2 + xlab('Age') + ylab('count') + ggtitle('Population\'s Age')

print(age3)

summary(Census_income_data$age)
View(count(Census_income_data, age))

var(Census_income_data$age)
sd(Census_income_data$age)

# range = max - min
# range = 90 - 0
# range = 90

# 1st qt = 15
# 3rd qt = 50
# mode = age 34: 3489 people
# mean = 34.49
# median = 33
# variance = 497.776
# standard deviation = 22.3109

# According to the summary it shows that the from 1st to 3rd qtr
# The age range is between 15-50 years old it represent that there
# are less than 50% of the population in USA around year 1994-1995
# are in the working age. From analysing the qtr of the age. It can
# be clearly seen that the new born or young people around age 0-33
# is doubled old people (that consider can't work anymore/retired)
# which is the good sign for the next 10-15 years for USA.
# 2.WAGE/HOUR

summary(Census_income_data$wage_per_hour)
var(Census_income_data$wage_per_hour)
sd(Census_income_data$wage_per_hour)
View(count(Census_income_data, wage_per_hour))

# range = max - min
# range = 9999 - 0
# range = 9999

# 1st qt = 0
# 3rd qt = 0
# mode = 0
# mean = 55.43
# median = 0
# variance = 75568.06
# standard deviation = 274.8965

# From mean, it can represent that people in USA's
# average wage/hr is 55.43$.
# We might use average wage/hr x average (hrsOfWork/day x week
# weeks/year) and we might get estimate income of
# US for that year as well.

# It can be clearly seen that more than 75% of the
# US population doesn't have any wage/hour, which
# can lead to the conclusion that they doesn't really
# work or they might earn their incomes from a different
# way


workers <- filter(Census_income_data, wage_per_hour > 0)

nrow(Census_income_data)
nrow(workers)

worker_percentile <- nrow(workers) / nrow(Census_income_data)
worker_percentile

# total rows = 199523
# worker rows = 11304
# worker percentile = 5% of the population
```



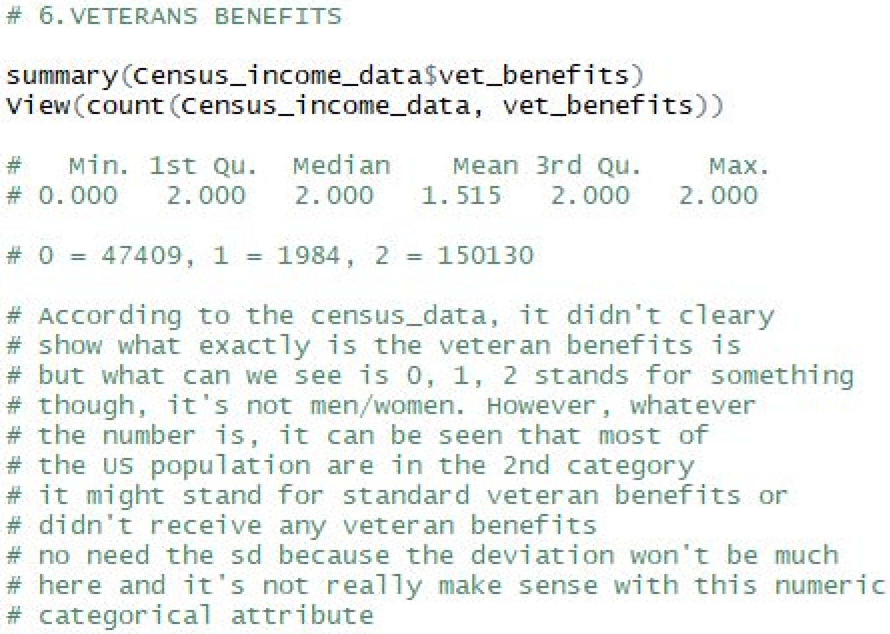
```
# 3.CAPITAL GAIN

summary(Census_income_data$capital_gains)
View(count(Census_income_data,capital_gains))

# 192144 is the number of people who doesn't have
# any capital gain

#    Min. 1st Qu.  Median    Mean    3rd Qu.    Max.
#    0.0    0.0     0.0      434.7     0.0    99999.0

no_capital_gain_percentile <- 1 - (192144/199523)
no_capital_gain_percentile

# No Capital Gain = 3.69%
# This can be conclude that average capital gain
# for US citizen is 434.7$

# There are alot of US population doesn't have
# any capital gains. Therefore, we would like
# to consider those who has capital gain = 0
# as an outliers for this capital gain analysis

have_capital_gain <- filter(Census_income_data, capital_gains > 0)

summary(have_capital_gain$capital_gains)
var(have_capital_gain$capital_gains)
sd(have_capital_gain$capital_gains)


# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 114    2964    5178   11754   10520   99999

# variance = 463672383
# standard deviation = 21533.05
# range = 999885

# After cut out all the population who doesn't
# have a capital gains, it can be seen that
# an average of the US population who has a
# capital gains are about 11754$. Though, it
# can be clearly seen that there's a huge gap
# between capital gains from the standard deviation
# which is 21533.05 or about 2 times a mean. the 3rd
# qt also represent that about 75% still earn less than
# mean which can lead to the conclusion of an outliers
# or there're some group that gains really huge amount
# of capital gains, as it can be seen from the max which
# is 99999. The range also vary variate from 114 - 99999$
# too. In addition, from going through each qt. It can
# be refers that most of the population
# (who have capital gains) only gains about 3000-10000$ as well.
```



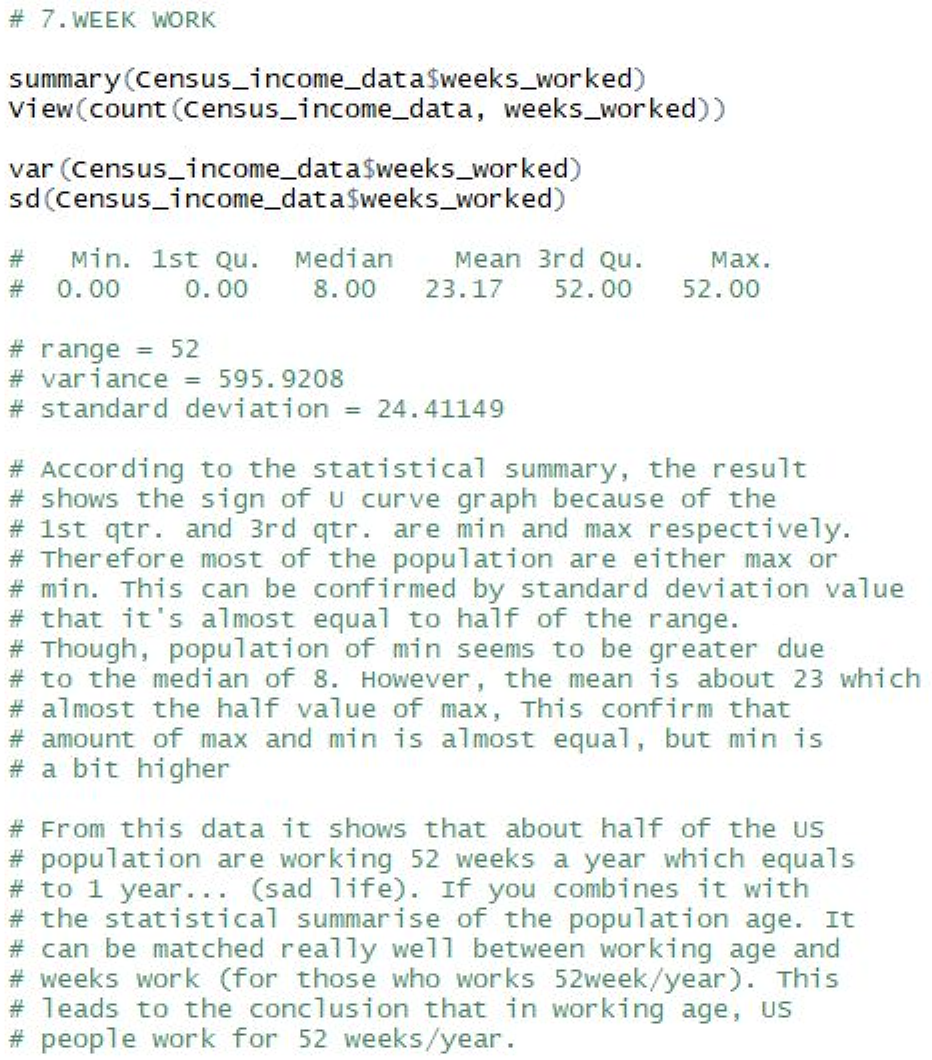
```
# 4.CAPITAL LOSSES

summary(Census_income_data$capital_losses)

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00    0.00    0.00   37.31    0.00 4608.00

# range = 4608

# From this it can be represented that average US
# population capital losses is 37.31, though their
# average capital gain is 434.7. Therefore, overall
# we believe that they gains more than losses. Plus,
# maximum losses is way less than maximum gains. That
# really is a good sign for the economics

# now try excluding those who not involving in the
# capital losses


have_capital_loss <- filter(Census_income_data, capital_losses > 0)

have_capital_loss_percentile <- nrow(have_capital_loss)/nrow(Census_income_data)
have_capital_loss_percentile

summary(have_capital_loss$capital_losses)

var(have_capital_loss$capital_losses)
sd(have_capital_loss$capital_losses)

# There're only about 1.9% of US population who
# has a capital losses.

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 155    1669    1887    1906    2001    4608

# variance = 214512
# standard deviation = 463.1544
# range = 4553 (4608 - 155)

# According to the statistical data, it shows
# that the amount of capital losses is very small
# compare to the capital gains, as well as the
# standard deviation, which can be infer that
# the most of the property's value and investment
# in USA yields the better result.
```



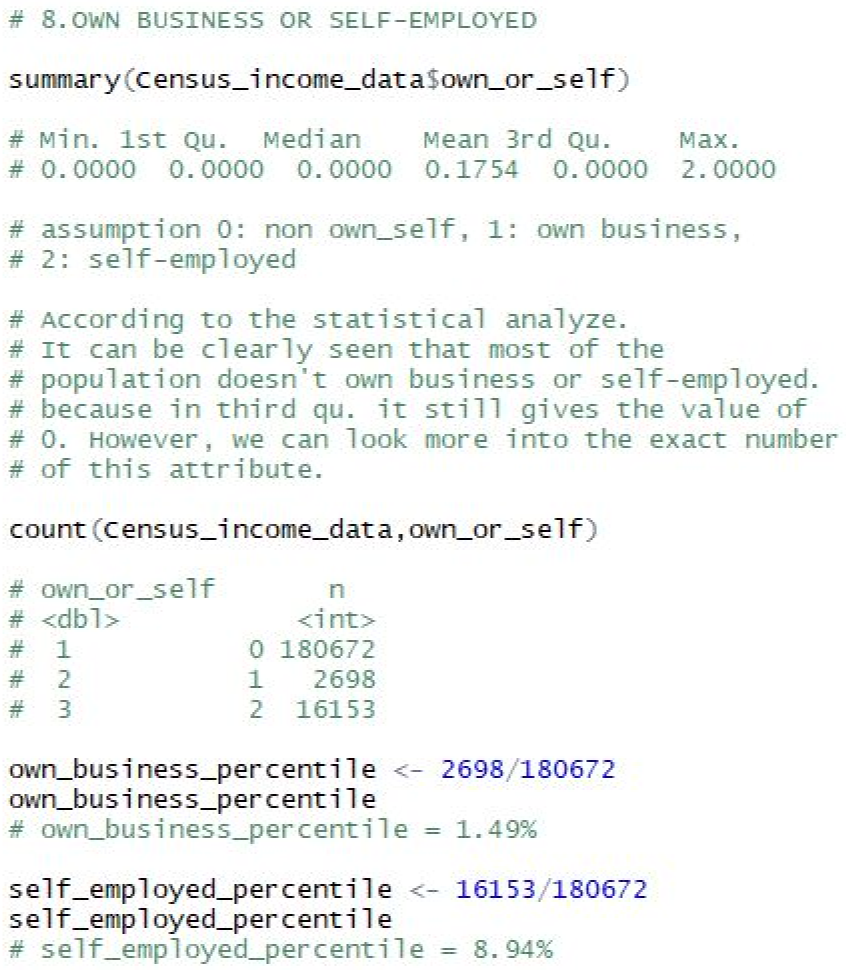
```
# 5.DIVIDENDS

summary(Census_income_data$stock_dividends)

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0     0.0     0.0   197.5     0.0  99999.0

# It can be clearly seens that there're alot
# of population that doesn't have any stock
# dividends and we believed that for this case
# this can be considered as an outliers

have_stock_dividends <- filter(Census_income_data, stock_dividends > 0)

summary(have_stock_dividends$stock_dividends)
var(have_stock_dividends$stock_dividends)
sd(have_stock_dividends$stock_dividends)

have_stock_percentile <- nrow(have_stock_dividends) / nrow(Census_income_data)
have_stock_percentile

View(count(have_stock_dividends, stock_dividends))


# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1      82     301    1864    1362   99999

# variance = 34049753
# standard deviation = 5835.217

# There are about 10.5% of US population who
# have stock dividends and it can be clearly seen
# that there's an outliers or may be a small
# group of people that makes have more stock
# dividends than the other. Should be in the
# last 25% or > 3rd qt. Therefore, they made
# a standard deviation really high. Though,
# for the majority of the stock dividends, it's
# around 0-1300. This can be represent futhur
# if we visualize it in the graph.
```



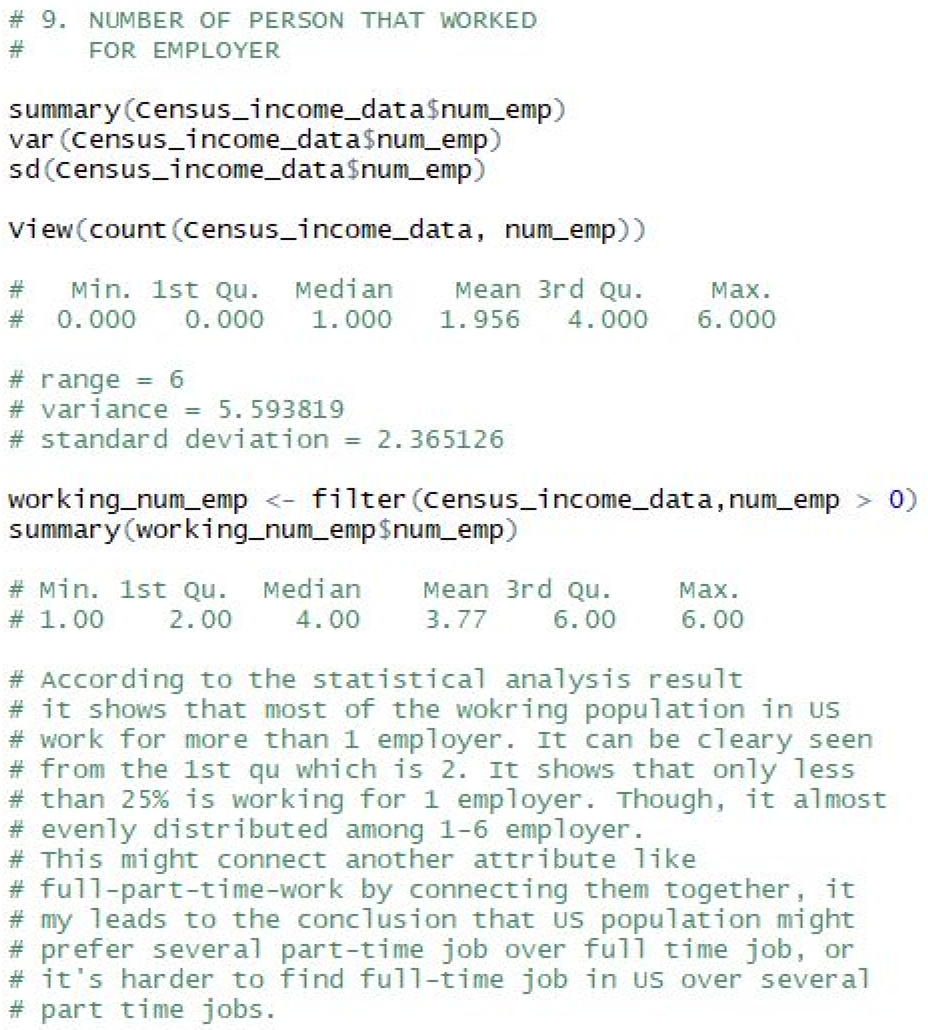
```
# 6.VETERANS BENEFITS

summary(Census_income_data$vet_benefits)
View(count(Census_income_data, vet_benefits))

#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.000   2.000   2.000   1.515   2.000   2.000

# 0 = 47409, 1 = 1984, 2 = 150130

# According to the census_data, it didn't cleary
# show what exactly is the veteran benefits is
# but what can we see is 0, 1, 2 stands for something
# though, it's not men/women. However, whatever
# the number is, it can be seen that most of
# the US population are in the 2nd category
# it might stand for standard veteran benefits or
# didn't receive any veteran benefits
# no need the sd because the deviation won't be much
# here and it's not really make sense with this numeric
# categorical attribute
```



```
# 7.WEEK WORK

summary(Census_income_data$weeks_worked)
View(count(Census_income_data, weeks_worked))

var(Census_income_data$weeks_worked)
sd(Census_income_data$weeks_worked)

#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  0.00    0.00    8.00   23.17   52.00   52.00

# range = 52
# variance = 595.9208
# standard deviation = 24.41149

# According to the statistical summary, the result
# shows the sign of U curve graph because of the
# 1st qtr. and 3rd qtr. are min and max respectively.
# Therefore most of the population are either max or
# min. This can be confirmed by standard deviation value
# that it's almost equal to half of the range.
# Though, population of min seems to be greater due
# to the median of 8. However, the mean is about 23 which
# almost the half value of max, This confirm that
# amount of max and min is almost equal, but min is
# a bit higher

# From this data it shows that about half of the US
# population are working 52 weeks a year which equals
# to 1 year... (sad life). If you combines it with
# the statistical summarise of the population age. It
# can be matched really well between working age and
# weeks work (for those who works 52week/year). This
# leads to the conclusion that in working age, US
# people work for 52 weeks/year.
```



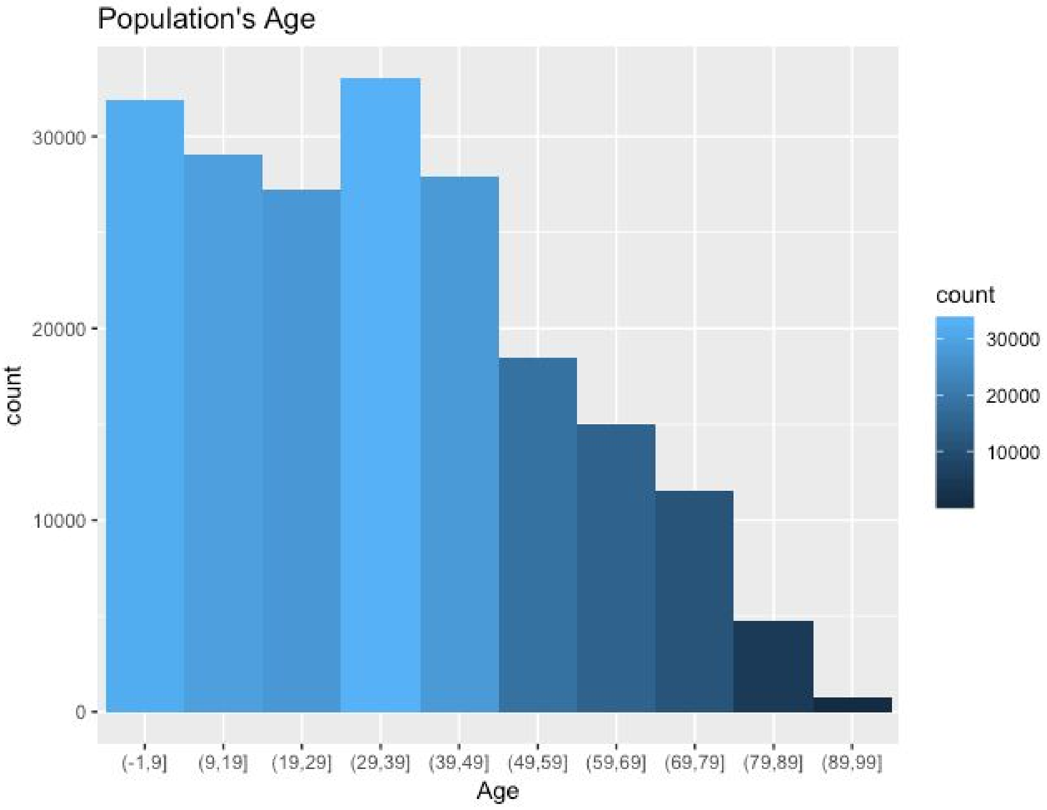
```
# 8.OWN BUSINESS OR SELF-EMPLOYED

summary(Census_income_data$own_or_self)

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0000  0.0000  0.0000  0.1754  0.0000  2.0000

# assumption 0: non own_self, 1: own business,
# 2: self-employed

# According to the statistical analyze.
# It can be clearly seen that most of the
# population doesn't own business or self-employed.
# because in third qu. it still gives the value of
# 0. However, we can look more into the exact number
# of this attribute.

count(Census_income_data,own_or_self)

# own_or_self       n
# <dbl>           <int>
#  1           0 180672
#  2           1   2698
#  3           2  16153

own_business_percentile <- 2698/180672
own_business_percentile
# own_business_percentile = 1.49%

self_employed_percentile <- 16153/180672
self_employed_percentile
# self_employed_percentile = 8.94%
```



```
# 9. NUMBER OF PERSON THAT WORKED
#    FOR EMPLOYER

summary(Census_income_data$num_emp)
var(Census_income_data$num_emp)
sd(Census_income_data$num_emp)

View(count(Census_income_data, num_emp))

#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  0.000   0.000   1.000   1.956   4.000   6.000

# range = 6
# variance = 5.593819
# standard deviation = 2.365126

working_num_emp <- filter(Census_income_data,num_emp > 0)
summary(working_num_emp$num_emp)

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.00    2.00    4.00    3.77    6.00    6.00

# According to the statistical analysis result
# it shows that most of the wokring population in US
# work for more than 1 employer. It can be cleary seen
# from the 1st qu which is 2. It shows that only less
# than 25% is working for 1 employer. Though, it almost
# evenly distributed among 1-6 employer.
# This might connect another attribute like
# full-part-time-work by connecting them together, it
# my leads to the conclusion that US population might
# prefer several part-time job over full time job, or
# it's harder to find full-time job in US over several
# part time jobs.
# 10. INCOME

categoryToNumber <- function(x){
  if(x == "-50000"){
    x <- 0.0
  }
  else{
    x <- 1.0
  }
  return(x)
}

temp_income_50k <- pull(Census_income_data,income_50k)
vector_income_50k <- c(temp_income_50k)

number_income_50k <- sapply(temp_income_50k, categoryToNumber)
summary(number_income_50k)

View(count(Census_income_data, income_50k))

# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00000 0.00000 0.00000 0.06206 0.00000 1.00000

# We transform the data of US population from
# those who get less than 50000 to 0 and greater
# than 50000 to 1. As a result, we found that
# there are only few people in US who earn more
# than 50k. There're less than 25%. Therefore,
# it's no need for sd nor variance here.

more_than_50k <- 12382/199190
more_than_50k

# There are only 6.2% of the US population
# who earns more than 50k. This percentage
# also similar to those who earn huge amount
# of capital gains and business/self-emplyed
# as well. It might possible to be the same
# group of people. Though, it's easy to trace
# them from capital gains, we can just set it
# more than 50k.

percentile_cap <- count(filter(Census_income_data, capital_gains > 50000)) / 199190
percentile_cap

# no... there are only 0.1% who gains more
# than 30k from capital gains... or about
# 390 people... Therefore, might be from
# the other businesses...
# now its harder to trace...
```

6. Visualize 10 most important attributes appropriately (histogram, bar chart, etc.). Provide an interpretation for each chart

6.1 Age

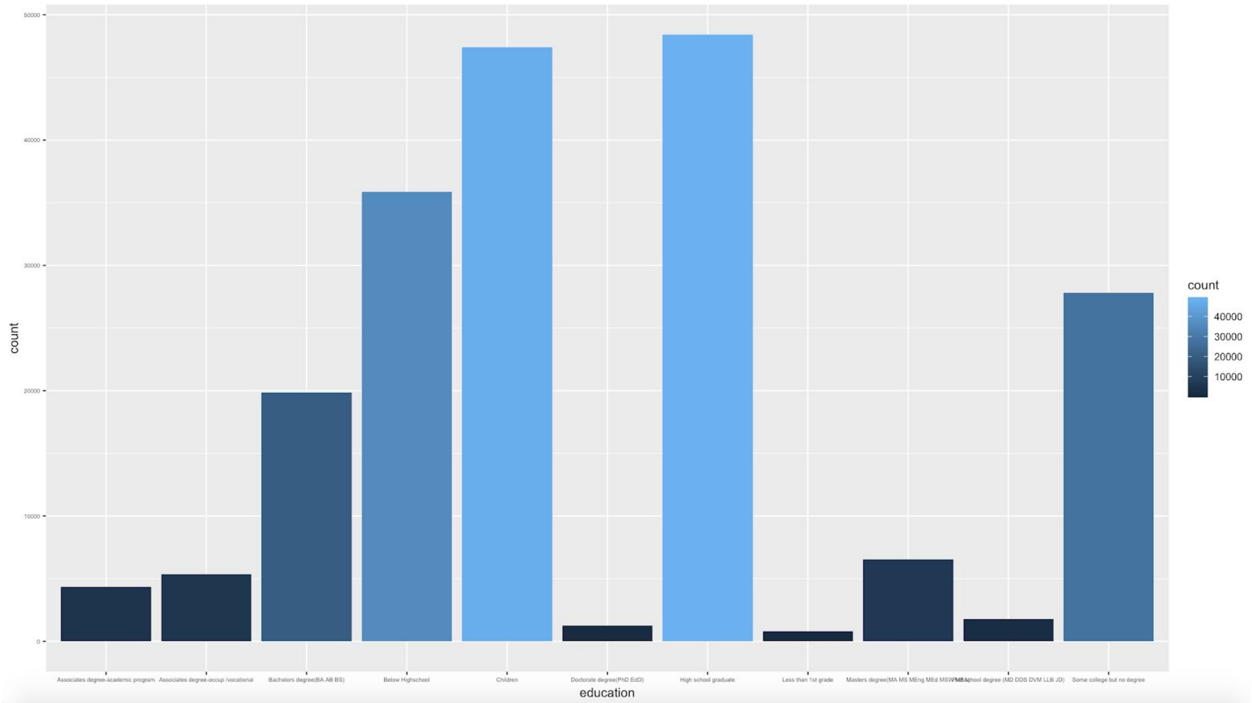


The graph above represents the number of workers in each age range in the US between 1994-1995.

According from the graph, the highest population of worker's age-range was between 29-39 years old. Which could mean that people in this age were one of the important force to drive the business.

However, there was a dramatic decline of workers after the age of 49 years old, it was possible to assume that there's a high amount of people who start their retirement after the age of 49.

6.2 Education

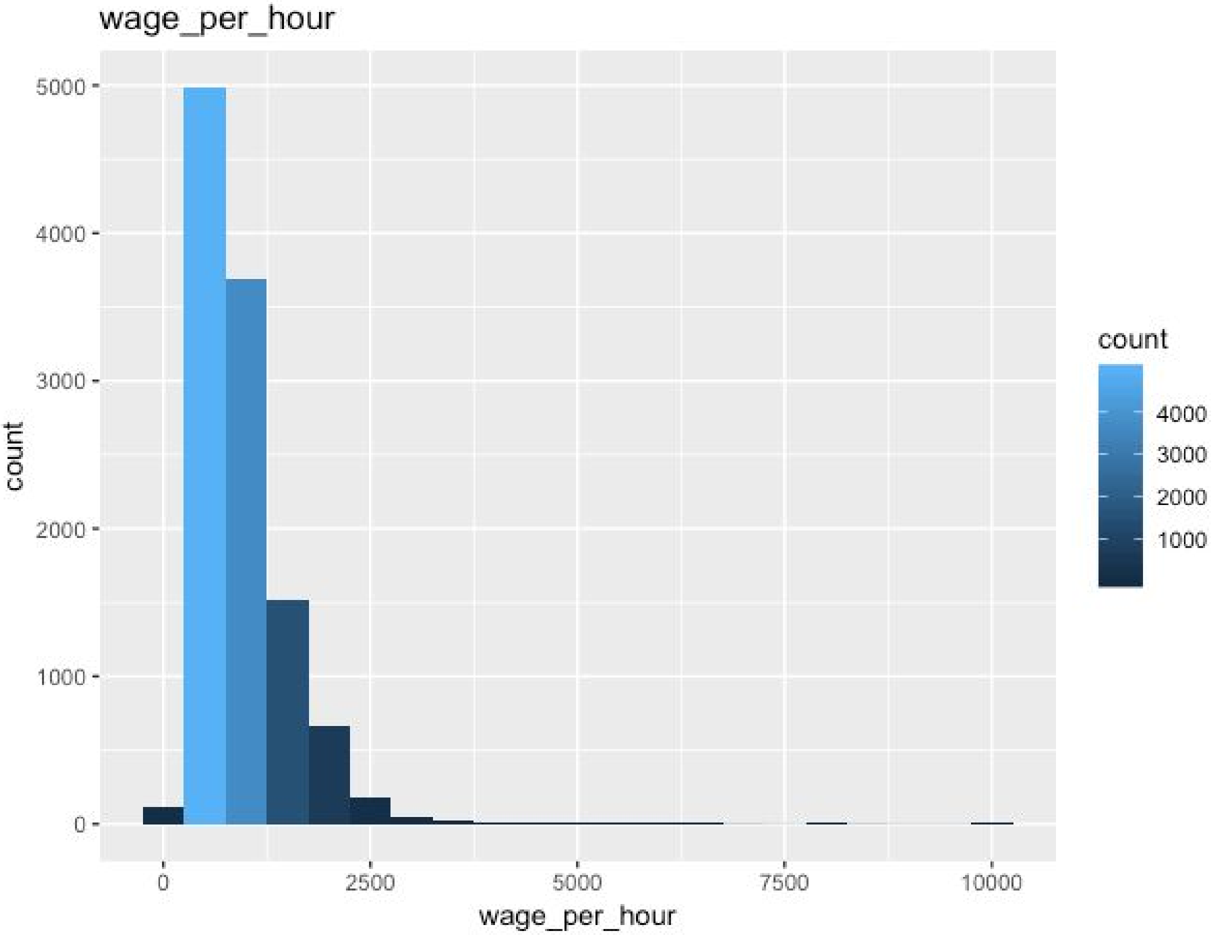


Associate Degree- Academic Program
Associate Degree- Occup/ Vocational
BA
School
Children
PhD
High- School Graduate
Less than 1st grade
MA
Prof School
Some college but no degree

The graph above shows the record of the level of education of the people in the US between 1994-1995.

According from the graph, it has shown that between 1994-1995, only approximately half of the people decided to continue their studying after high school. The majority of people tended to start to work after high school instead of pursuing higher degree.

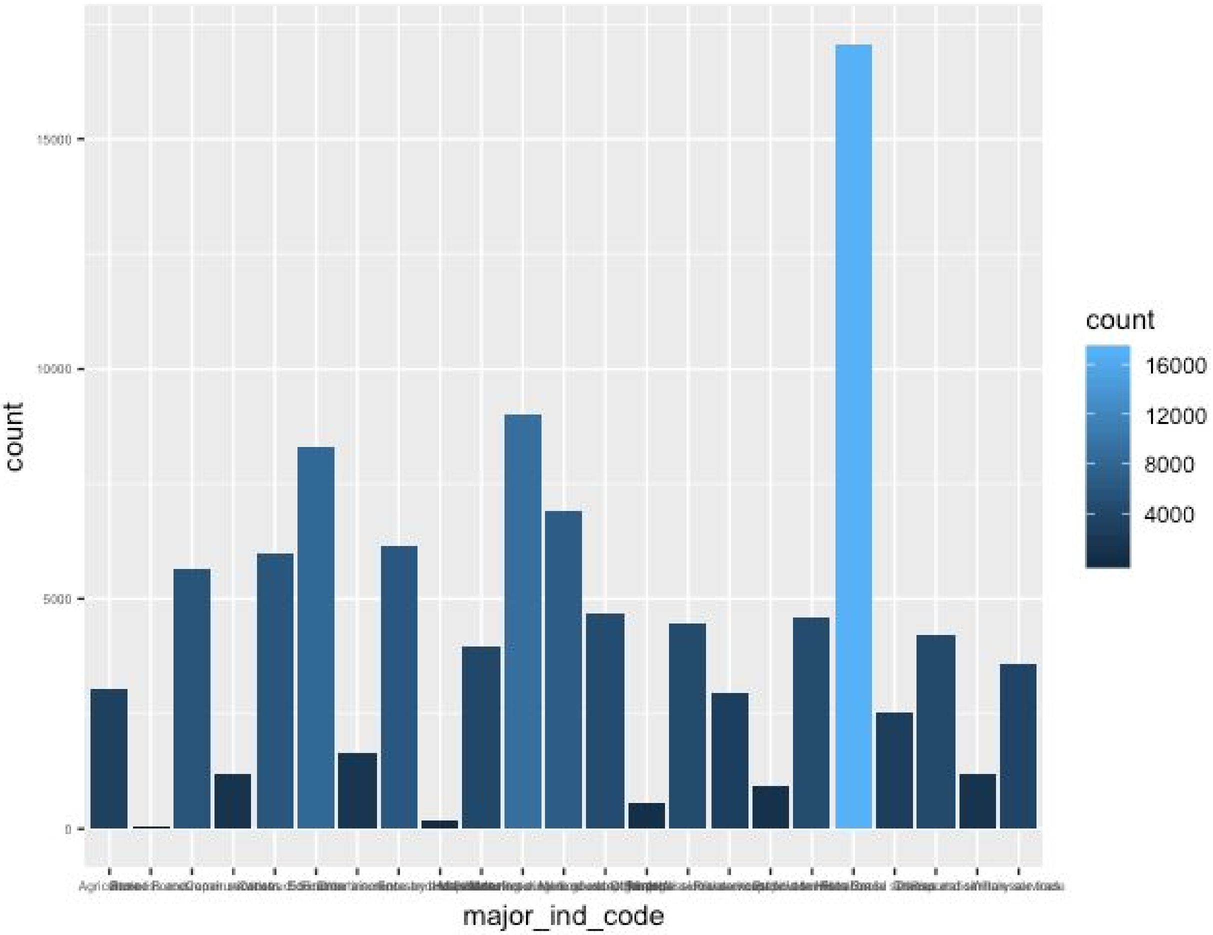
6.3 Wage per hour



The graph indicates number of people in the US with their wage per hour, the data was collected during 1994-1995. The data shown only the data of workers who earn more than 0 USD per hour.

The average wage per hour was approximately 1000 USD, excluding the number of people who earn 0 USD per hour or no data.

6.4 Major Ind Code



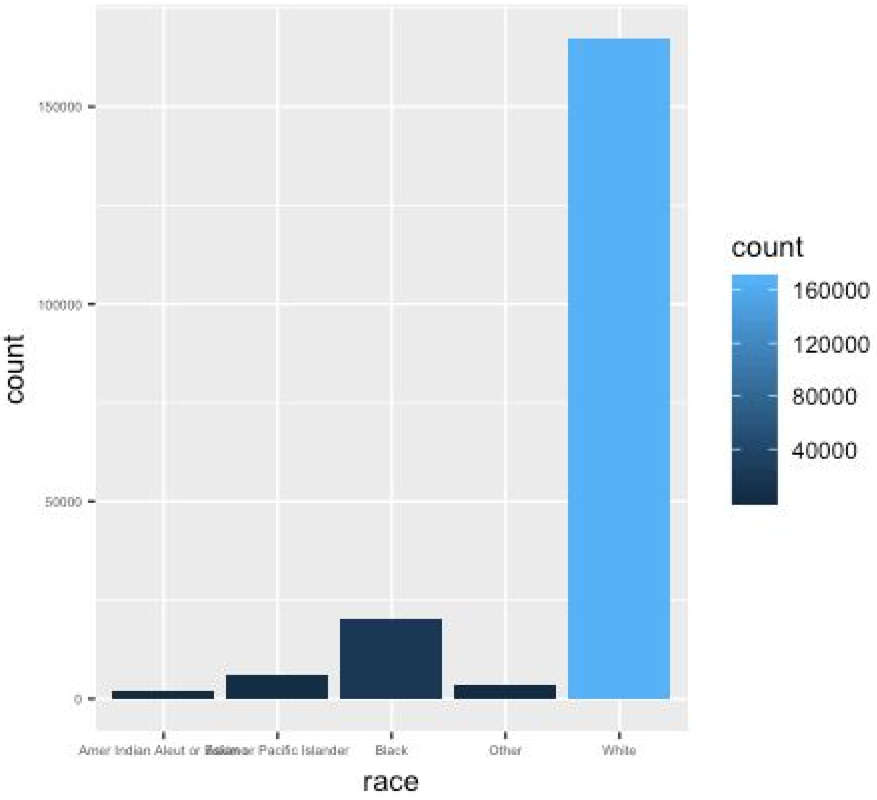
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

(From left to right) 1.Agriculture, 2.Armed Force, 3.Business and repair service, 4.Communications, 5.Construction, 6.Education, 7.Entertainment, 8.Finance Insurance and real estate, 9.Forestry and fisheries, 10.Hospital services, 11.Manufacturing-durable goods, 12. Manufacturing-nondurable goods, 13.Medical except hospital, 14.Mining, 15.Other professional services, 16.Personal services except private HH, 17.Private household services, 18.Public administration, 19.Retail trade, 20.Social services, 21.Transportation, 22.Utilities and sanitary services, 23.Wholesale trade

The graph represents the number of workers in each major industry fields in the US during 1994-1995.

According from the graph, the number of people who worked for the retail trade was obviously the highest among all the jobs, following by manufacturing-durable goods. It could be assumed that the majority of business in the US was mainly on producing and distributing products.

6.5 Race



Amer Indian Aleut Or Eskimo

Asian Or Pacific Islander

Black

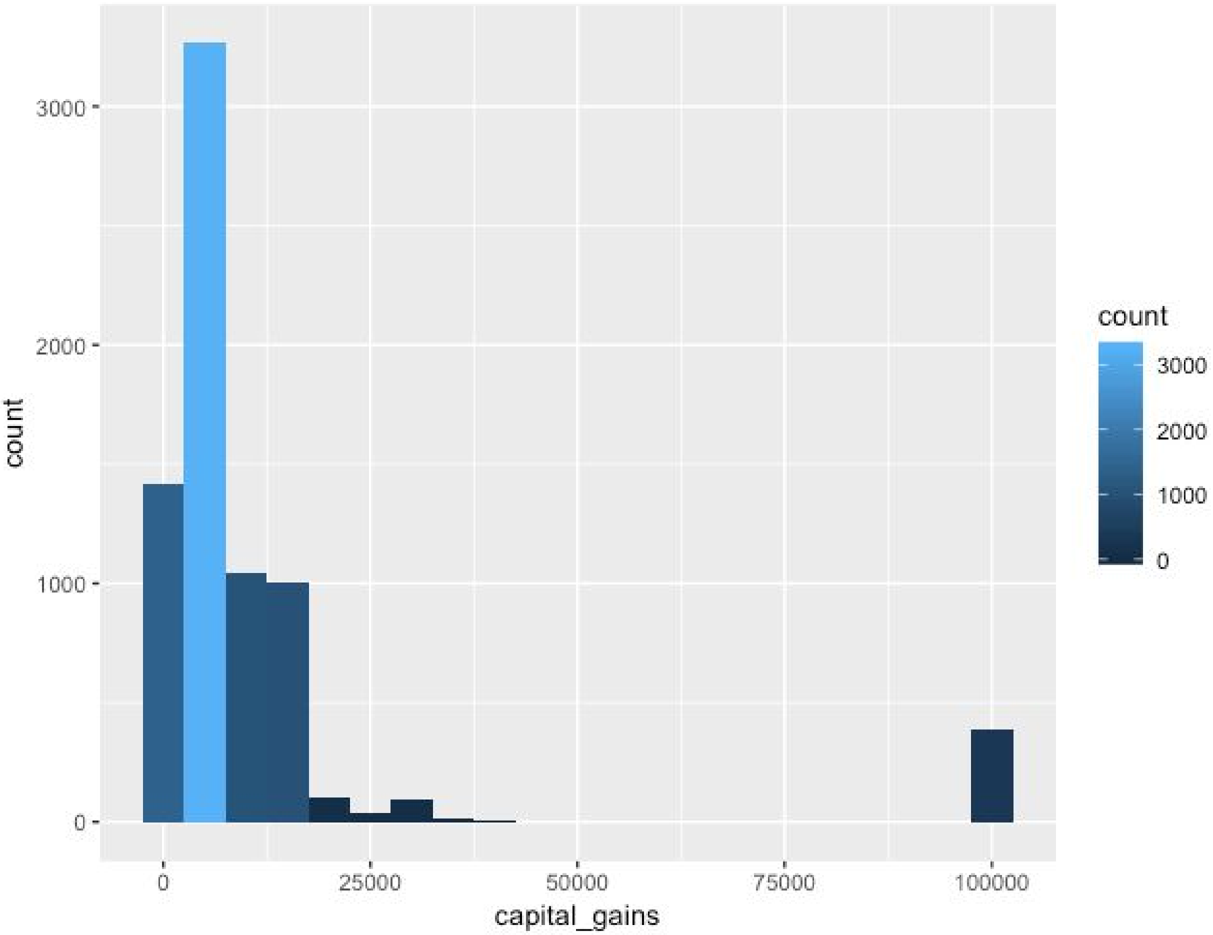
Other

White

The chart indicates the number of people of each races; American Indian or Eskimo, Asian or Pacific Islander, Black, White. The data was collected in the US between 1994-1995.

According from the graph, the population of white people ranked the highest, following by black people. The number of white and black people had quite a explicit difference comparing to other 3 races which were quite similar in number. It could be concluded that white and black people are the majority of population in the US.

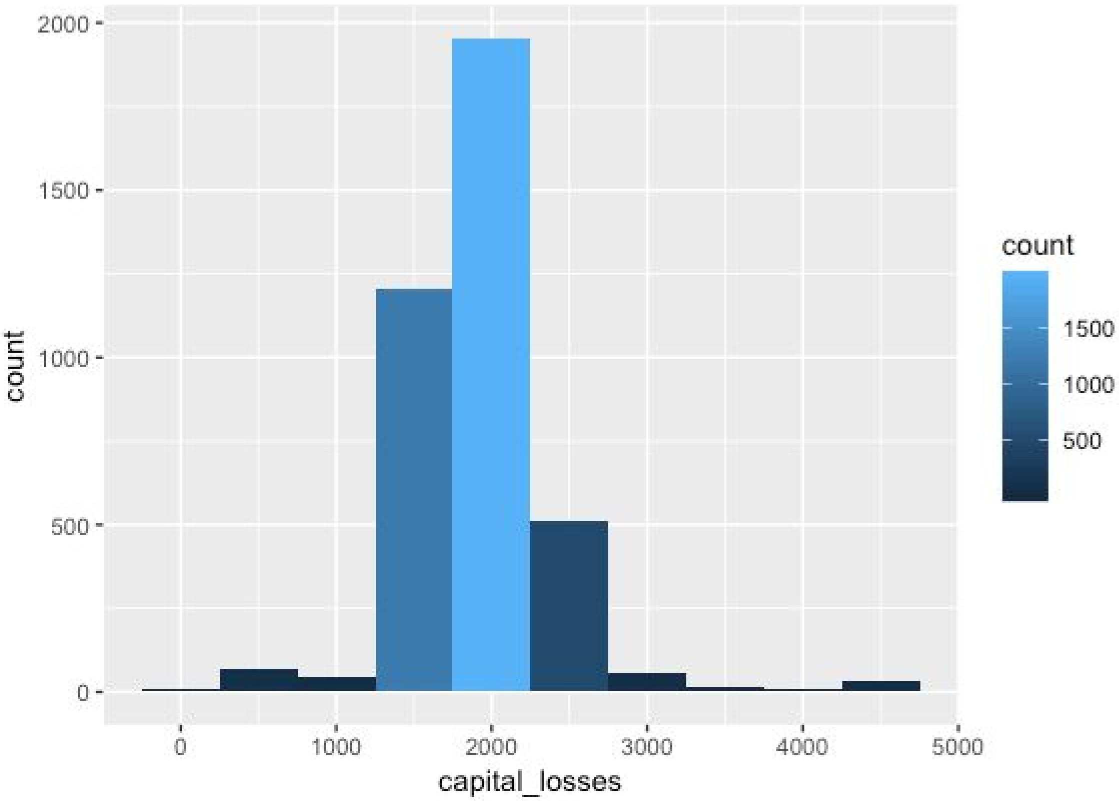
6.6 Capital Gain



The graph represents number of people who has capital gains and their amount of capital gains in the US during 1994-1995.

According the the graph, the majority of workers had capital gain from 0 - 25,000 USD , however, there is a massive gap between the majority of people and a group of people who has capital gains of 100,000 USD. This data could reflects on the economic inequality.

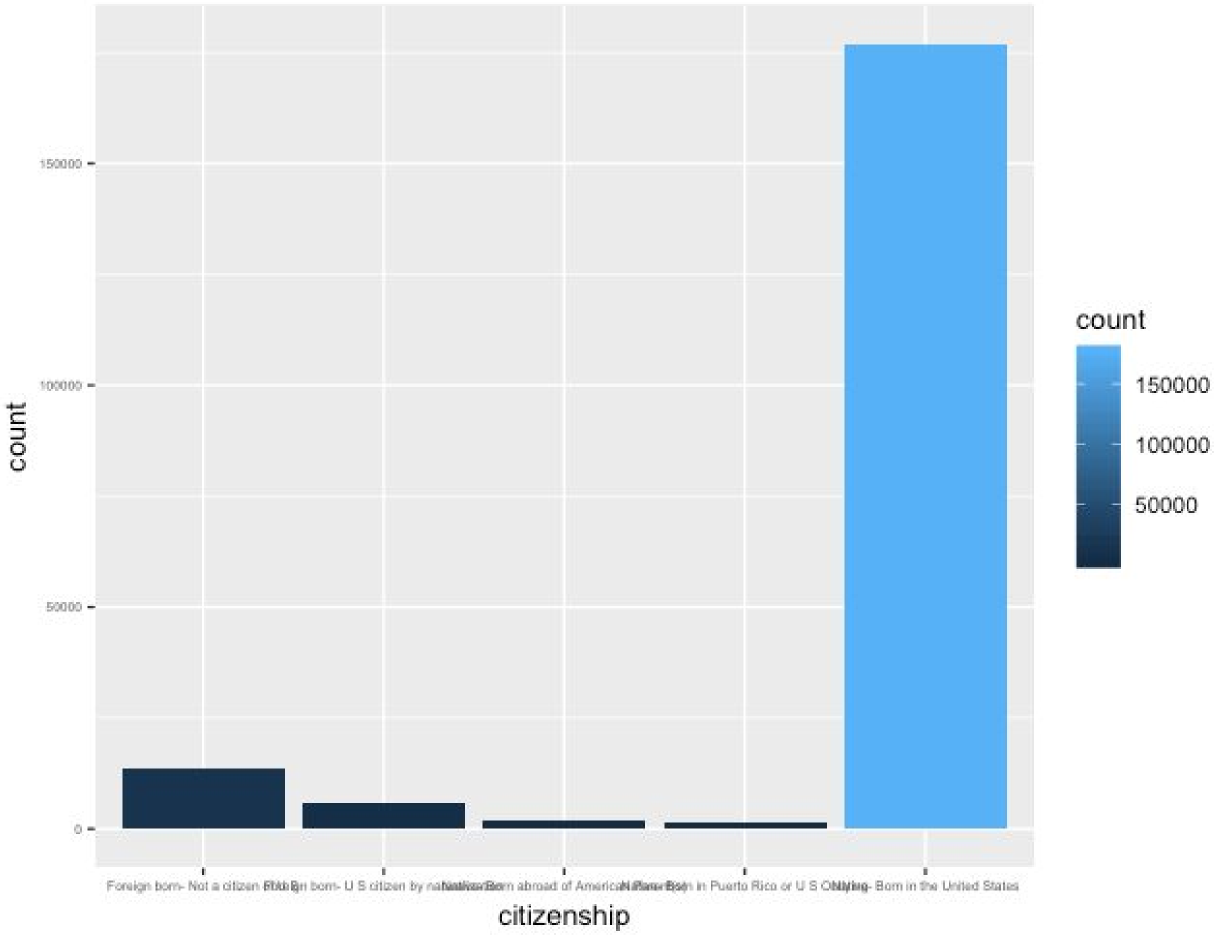
6.7 Capital Loss



The bar graph represents number of people who has capital losses and their amount of capital losses in the US during 1994-1995.

According from the graph, the workers had capital loss of 1500-2000 USD averagely during that year.

6.8 Citizenship



Foreign born- Not a citizen of US

Foreign Born- US citizen by naturalization

Native- Born abroad of American Parent(s)

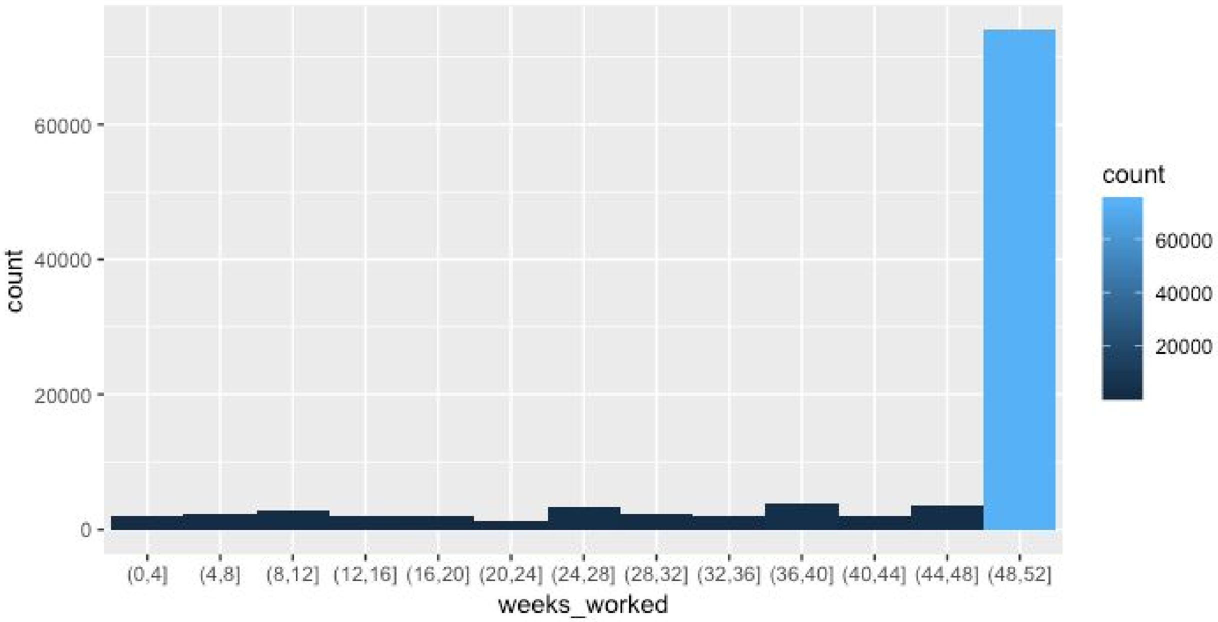
Native Born Of Puerto Rico or US Outlying

Native Born In the US

The bar graph represents number of people and their citizenship in the US during 1994-1995.

According from the graph, the majority of the workers in the US were the native. However, the foreign born and non-citizen was number 2 which was more than the foreign born-US citizen and even more than the native who was born abroad and the native who was born in the US outlying.

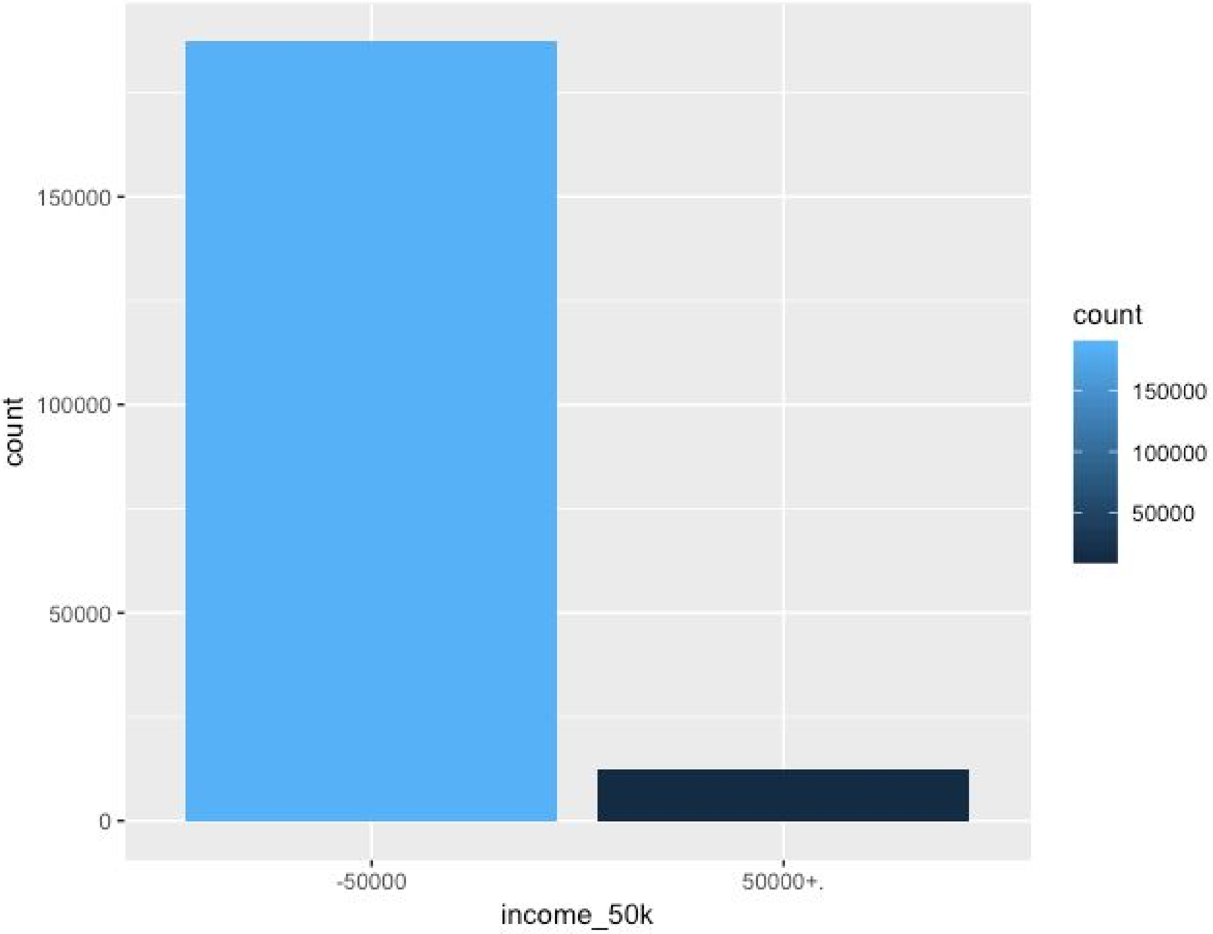
6.9 Week worked



The graph indicates the number of people and their length of work in unit of weeks, collected in the US between 1994-1995.

The majority of the workers in the US worked 48-52 weeks a year which is almost all year round. It could be concluded that the workers in the US rarely have holidays.

6.10 Income 50k

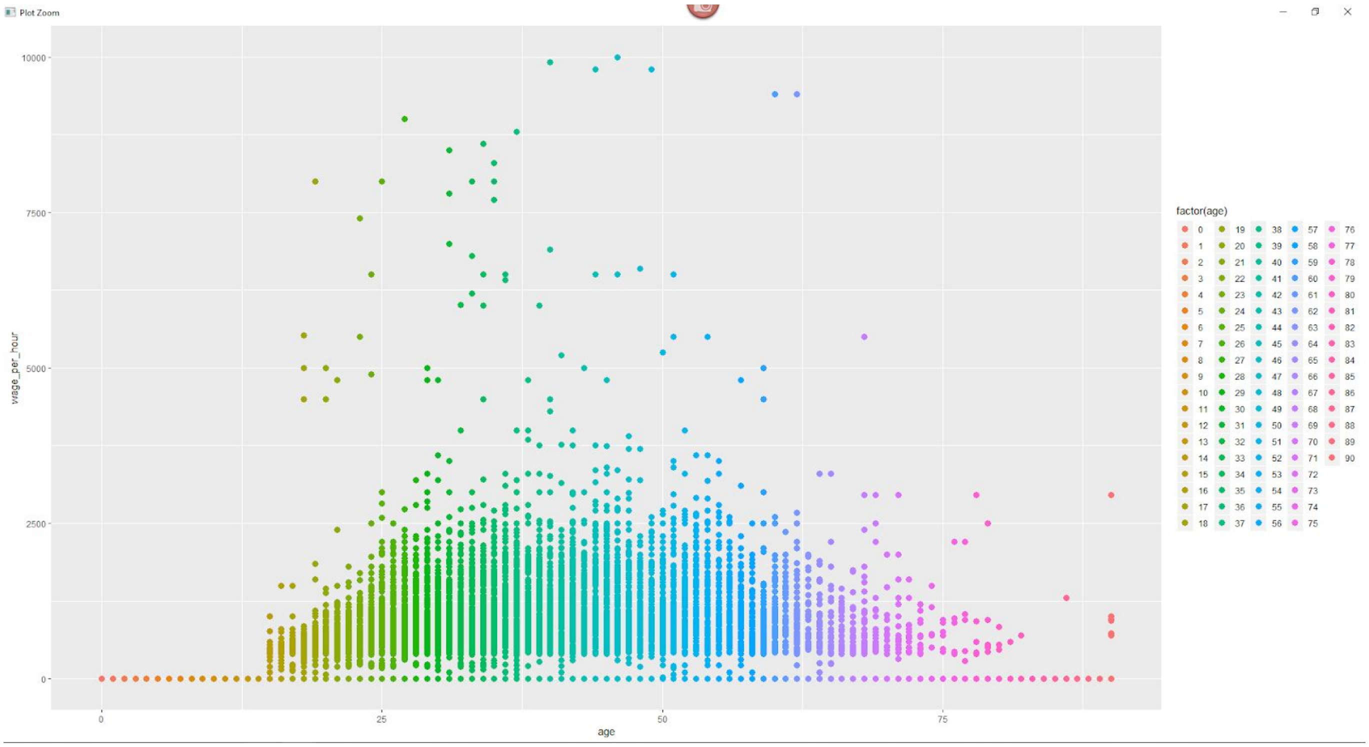


The graph indicates the number of people in the US who earn more or less than 50,000 USD a year in the US during 1994-1995.

The majority of workers in the US had income lower than 50,000 USD a year. Only approximately 7% of the people who earned higher than 50,000 USD.

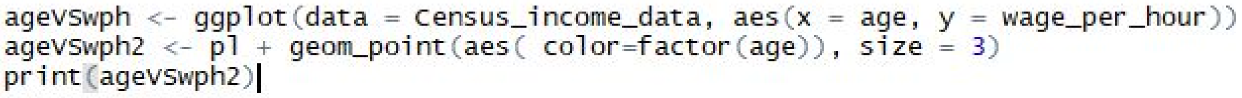
7. Explorer relationships between attributes for 5 relationships. Look at the attributes and then scatter plots, correlation, etc. as appropriate. Explain the results.

7.1. Age vs Wage/Hour



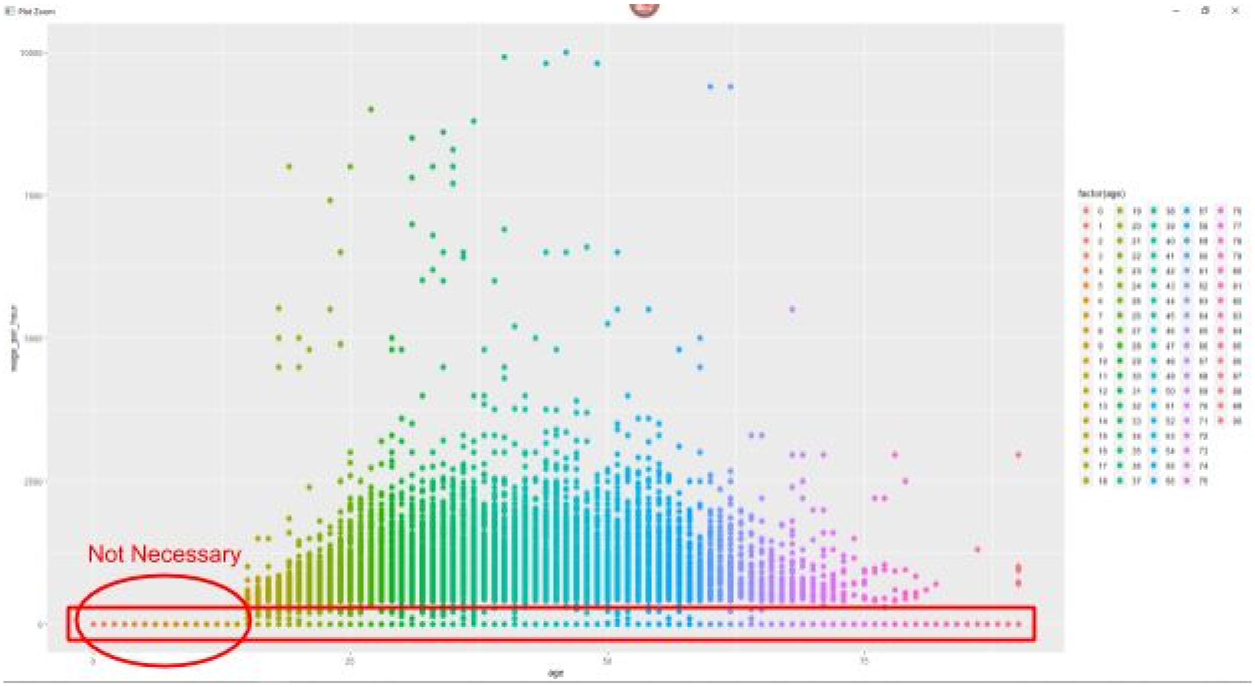
A scatter plot above is presenting an **Age** of the US population between 0 to 90 on an x-axis and **wage_per_hour** of them on the y-axis. At first we tried to filter out some of the values that we believe it would be outliers or doesn't necessary for the result out. However, in the end we believe that show it all is easier to visualize and explain our idea toward these values. In addition, it doesn't really affect the relationship of the other points as well.

Code

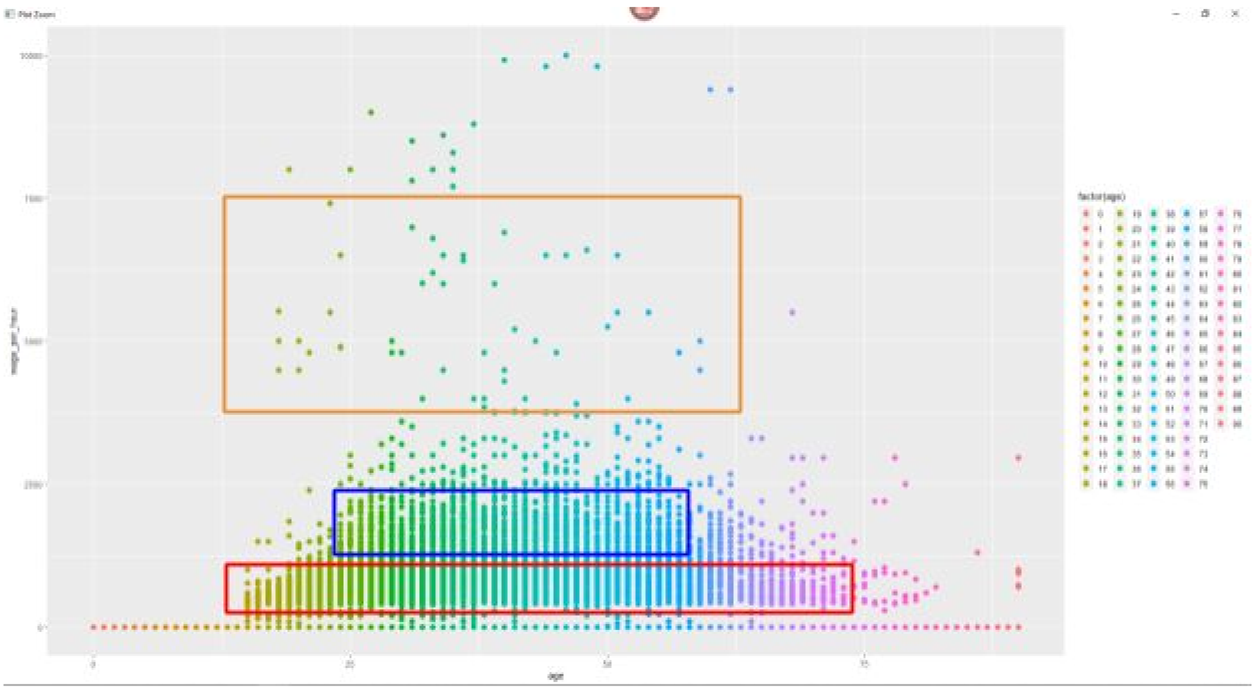


```
agevSwph <- ggplot(data = Census_income_data, aes(x = age, y = wage_per_hour))
agevSwph2 <- pl + geom_point(aes( color=factor(age)), size = 3)
print(agevSwph2)
```

Analyzing



According to the scatter plot, it can be clearly seen that there are 2 parts which are not necessary. First, there are those who earns 0$ / hour, it is the people who doesn't have any wage/hour. It can be interpret as 2 ways, one is they do not have an income and second is they do not work (have passive income). Second, there are those who age under 16, which can be interpreted as the citizen age under 16 are not allowed to work in US. However, these information is not necessary if we are focusing on the relationship between Age and wage_per_hour because they do not earn any wage, though it can represents some other useful facts.

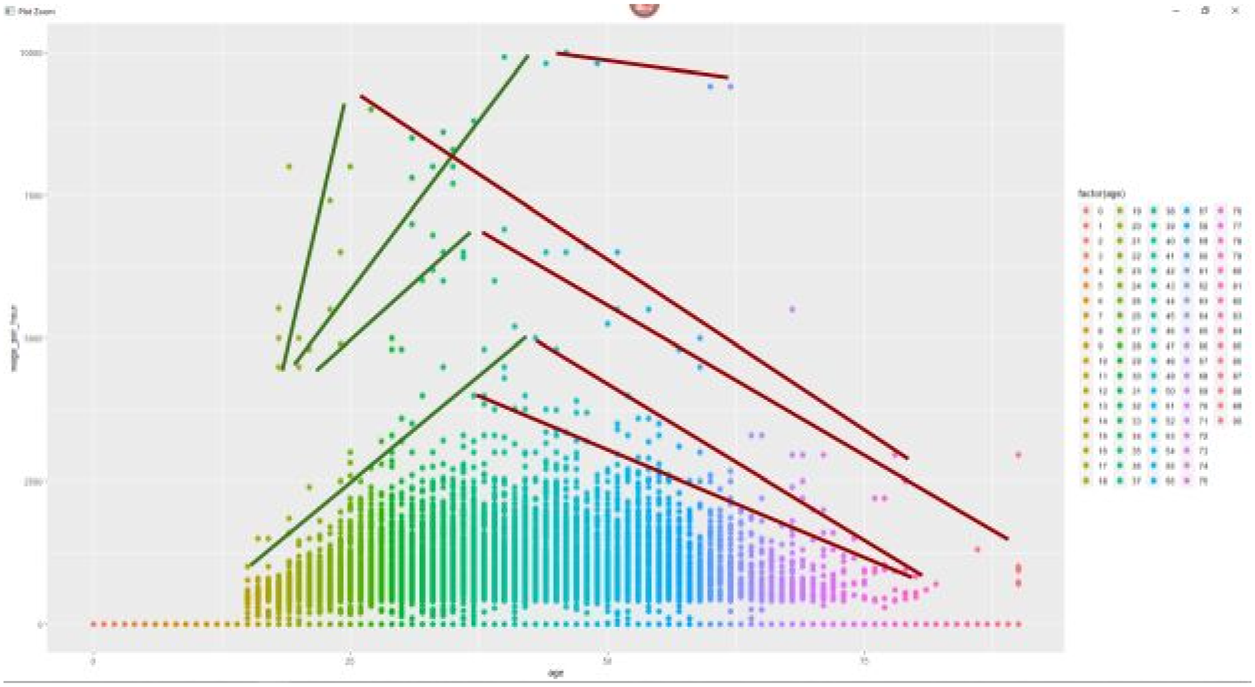


Now, according to the scatter plot above, we classified the points into 3 groups of red, blue, and yellow hence we will discuss it one by one respectively.

First, the red group, it is the highest populated group. It can be seen that most of the points are packed into this red group, the density of the points per block is very high and its consume roughly 5 blocks of the graph. It represents a people who earns around 1-1250$ / hour and the age range is between 16 to 70 which covers all the working age range. Therefore, this leads to the conclusion that 1-1250$ is the basic salary/hour for the US population around 1994 to 1995 because most of the worker in every age range earn at least this much.

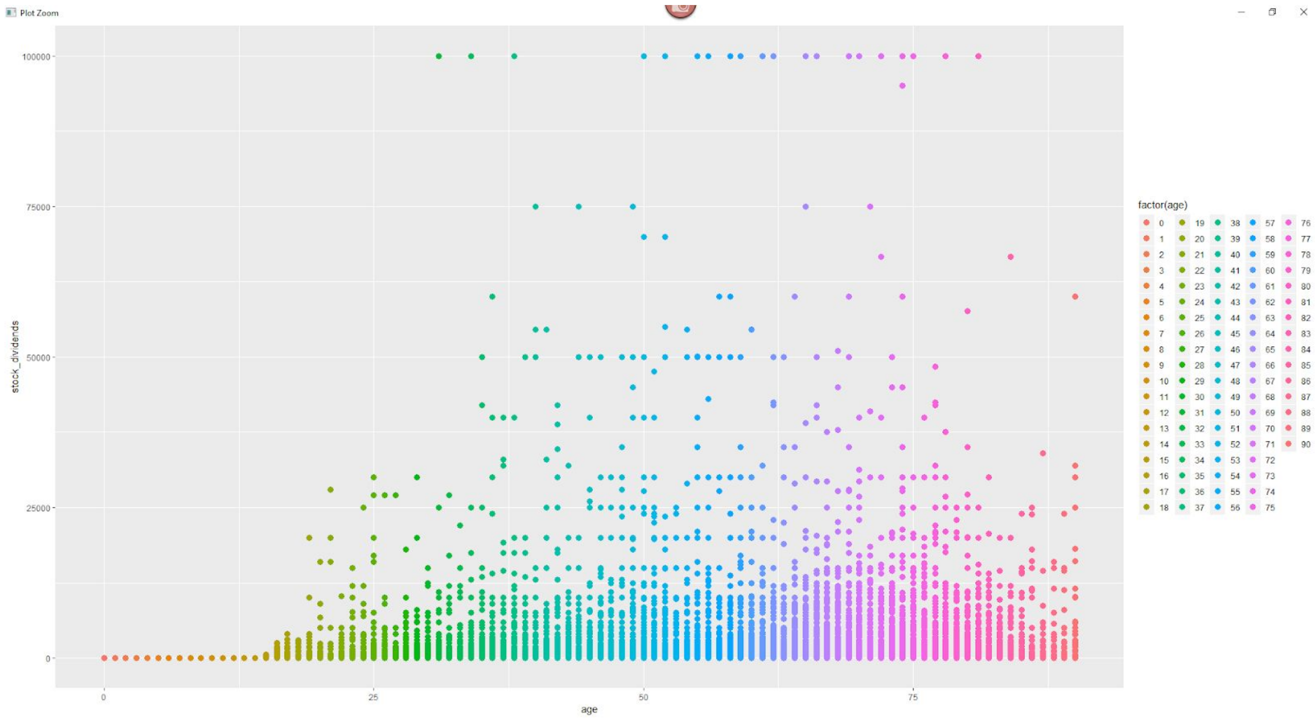
Next, the blue group, it is the second highest populated group. In this group, the density of the points in each box also highly populated as well. However, it only takes about 2.5 boxes. This group's age range is around 25 to 60 and the wage/hour around 1250 to 2500$. An ages between 25 to 60 are considered to be most effectively working period of human life. 25 also an age that most of the college student graduate for a year or 2 years as well. This blue group should be working on the higher position than the prior group. However, apart from the college's graduation, there's another assumption for this blue group. There is a high chance that the blue group is promoted from the red group as you can see on the box between age range of 16 to 25, there is an upward trends. However, due to the high density of people in other age range in both blue and red group it's hard to if there's other upward trends around this working age or not.

Last, the yellow group, it is very low density of points per block that we have to make it into and points per 3 blocks instead. This group represent those who earns around 4000 to 7500$ between age 18 - 60. However, we tried to separate age range into 4 groups according to the block in the graph which are 16 to 25, 25 to 37.5, 37.5 to 50, and 50 to 62.5. There exist the similarities between these 4 groups. Each group have almost the same number of 9 to 12 people to earns wage/hour within the given range. However, between these 4 groups, the people around age 25 to 37.5 wage/hour is the highest. In contrary, 16 to 25 and 50 to 62.5 are the lowest. This can leads to an assumption that 25 to 37.5 is the peak time of making the money while 50 to 62.5 might be a bit too old and 16 to 25 might be too early or lacking of experience. There also exist some upward trends between 16 to 37.5 as well that helps guarantee an assumption of too early or lacking of experience as well. In addition, between age range of 37.5 to 62.5, it is very clear that there is a downward trends exist it is either drastic or steady, but it surely is a downward trends as they grow older.

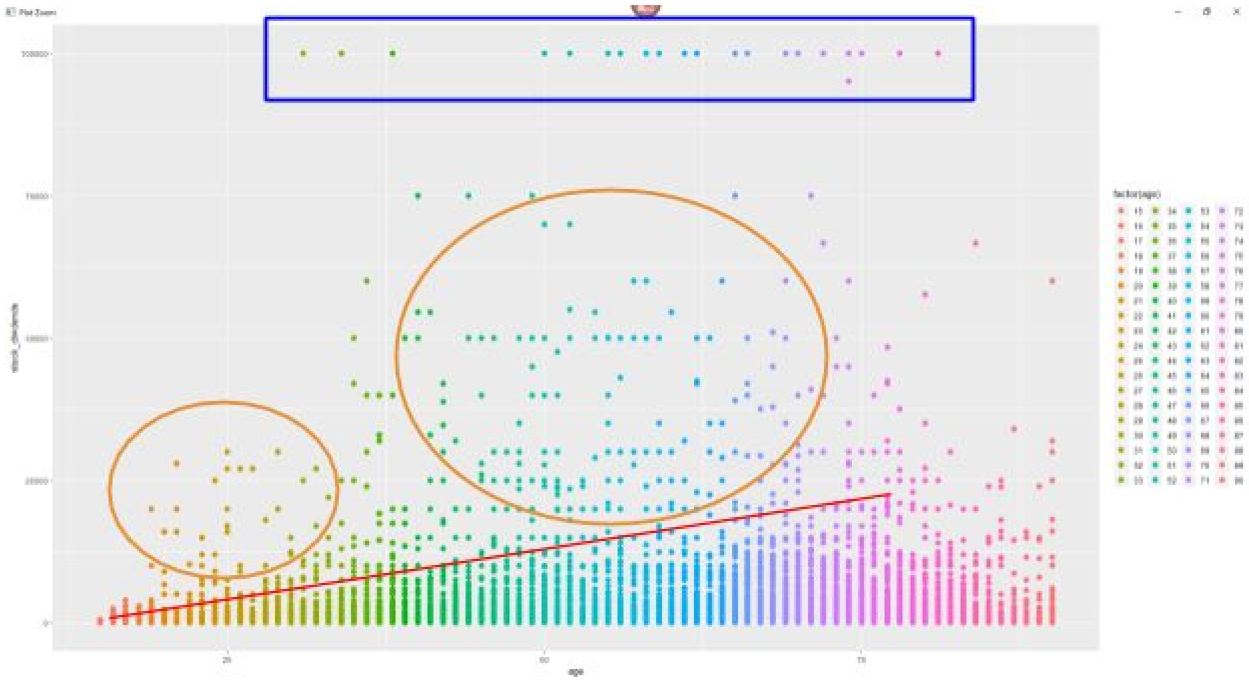


After analyzing several boxes just now, we did see some trends occurring on the graph. It can be clearly seen that between age of 16 to 35 it most of the plot are leading towards the upward trends. However, as the people gets older the trends tend to become a downward in the similar rate. Therefore, this leads to another assumption or suggestion that we should earn as much as possible during our growth time, so that when we get older we still earns a lot just less than our peak time. Trying to be an outlier might be a good idea as well because an outlier’s downward trends seems to be the steadiest one.

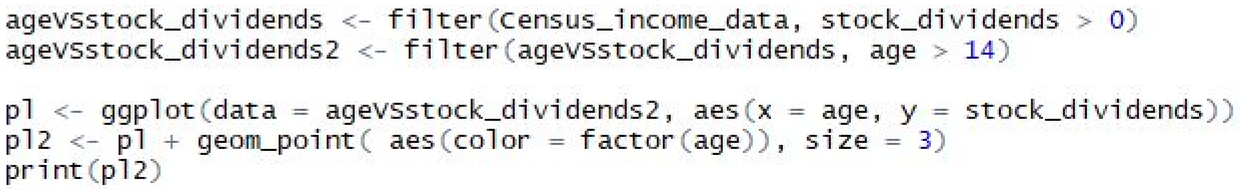
7.2. Age vs Stock Dividends



From the given scatter plot, there is a chance that there's a relationship between these 2 attributes. Though, the outliers such as Age less than 15 affect the visualization too much. Therefore, this will be cleaned out.



Code



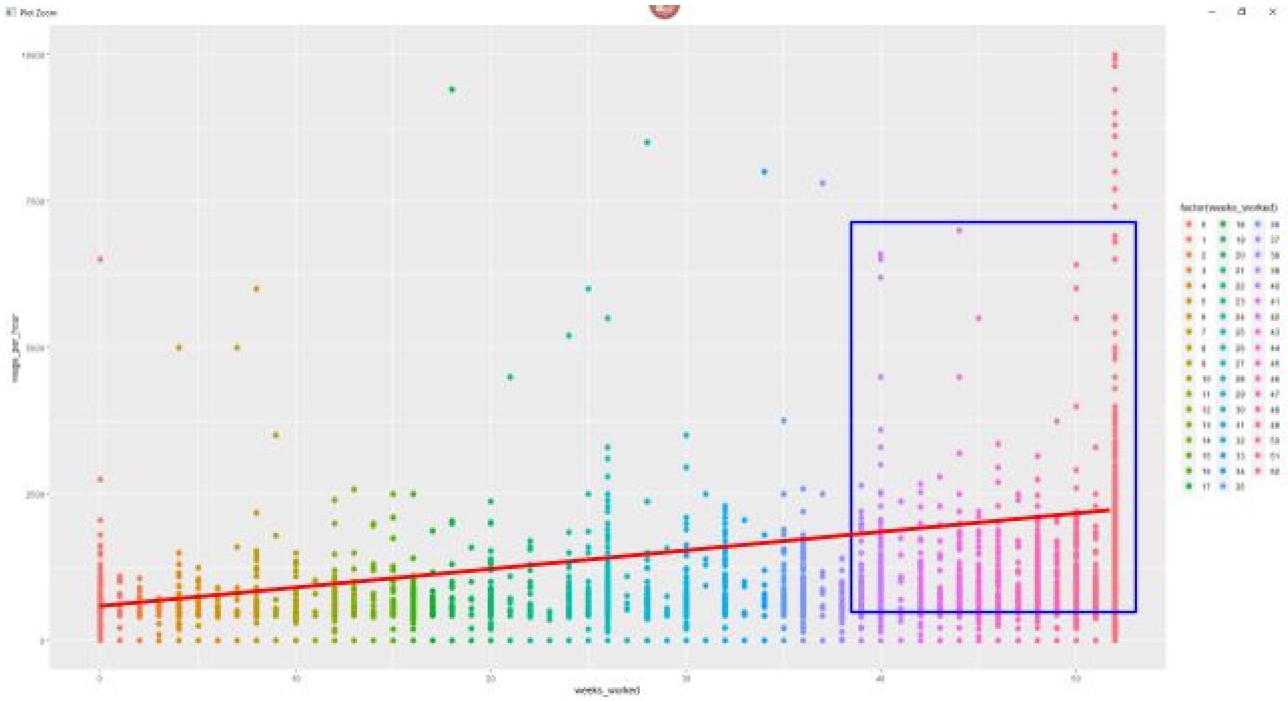
```
ageVSstock_dividends <- filter(Census_income_data, stock_dividends > 0)
ageVSstock_dividends2 <- filter(ageVSstock_dividends, age > 14)

pl <- ggplot(data = ageVSstock_dividends2, aes(x = age, y = stock_dividends))
pl2 <- pl + geom_point( aes(color = factor(age)), size = 3)
print(pl2)
```

Analyzing

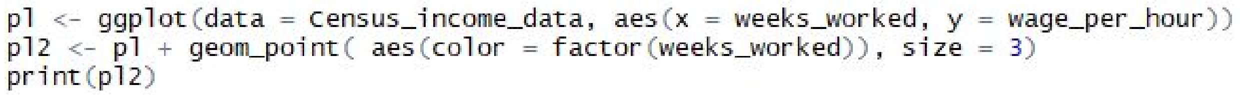
After did some cleaning, the visualization become better. Now, there are 3 things that should be concerned. First, red line,there is an upward trends for the amount of people who received the stock dividends as an age is increased until around age 75. Second, yellow circle, The population of the people who gains huge amount of the stock dividends betweenage 37.5 to 75 is greater than those who gains around age 15 to 37.5. However, between age 37.5 and 75 amount of people who gains huge amount of stock dividends are quite evenly distributed among the 12.5 years of age range. Last, blue rectangle, it is very clear to see that most of the people who gains an outlier amount of the stock dividends are more than 50 years old, usually around 55 to 75 years old.

From this analysis, it leads to similar conclusion as the Age vs Wage/Hour which are as the people get older they tends to earns more stock dividends. This might due to the difference in experience, money, and connection as the people works more and gets older. Though, there are some exceptional young people who earns a lot from stock dividends as well and there are definitely an outliers.

7.3. Weeks Worked VS Wage/Hour



Code



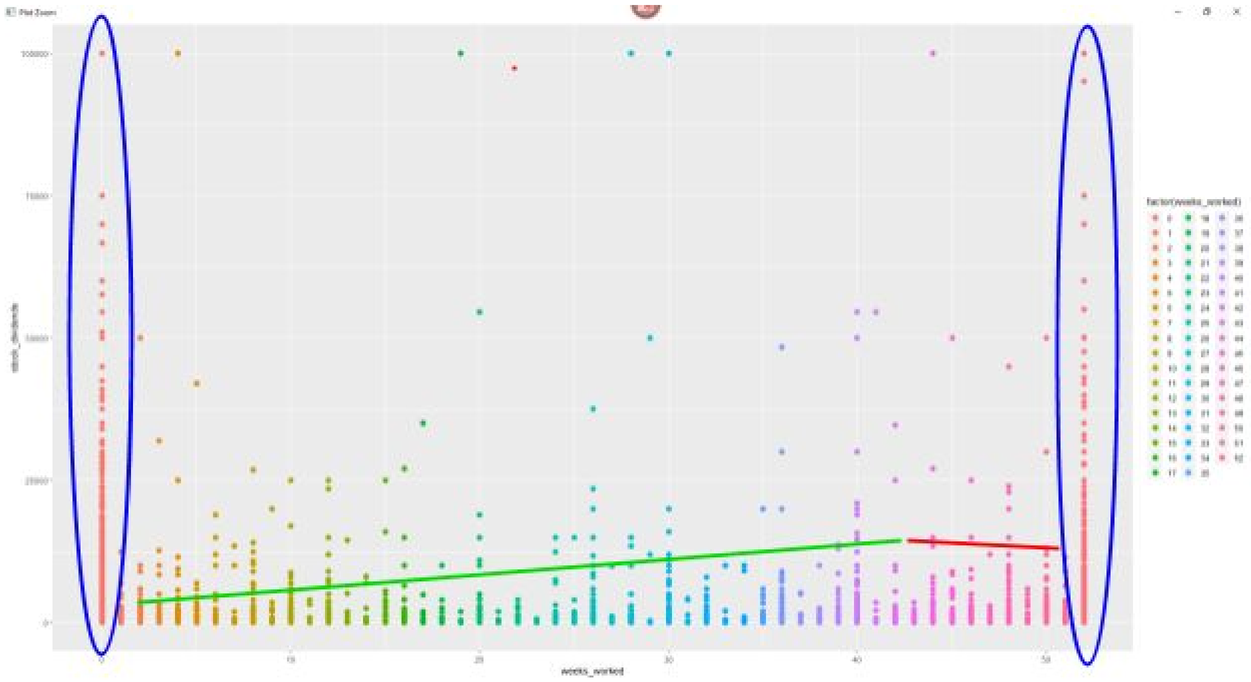
```
pl <- ggplot(data = Census_income_data, aes(x = weeks_worked, y = wage_per_hour))
pl2 <- pl + geom_point( aes(color = factor(weeks_worked)), size = 3)
print(pl2)
```

Analyzing

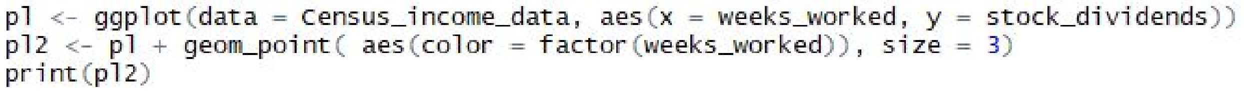
According to the graph representing above, there are 2 facts that it clearly represented. First, the blue rectangle, it shows that between 40 to 52 weeks it has a very high density of the points in the graph compare to the other part. It shows that most of the US population have to work for more than 40 weeks/year. Second, red line, it exist an upward trends, it shows that not only the population(points) that keep increasing as the week_worked is increased, but the wage/hour as well. As the people work more there have a change to get higher wage/hour as well as it is represented in the blue rectangle. Special note for the 52 weeks, it clearly shows that most of the population that earns high or out standing wage/hour are in working 52 weeks/year.

In conclusion, this graph leads to an assumption that the more weeks people works the higher chance for them to earns more wage/hour in US between year 1994 and 1995.

7.4. Weeks Worked VS Stock Dividends



Code



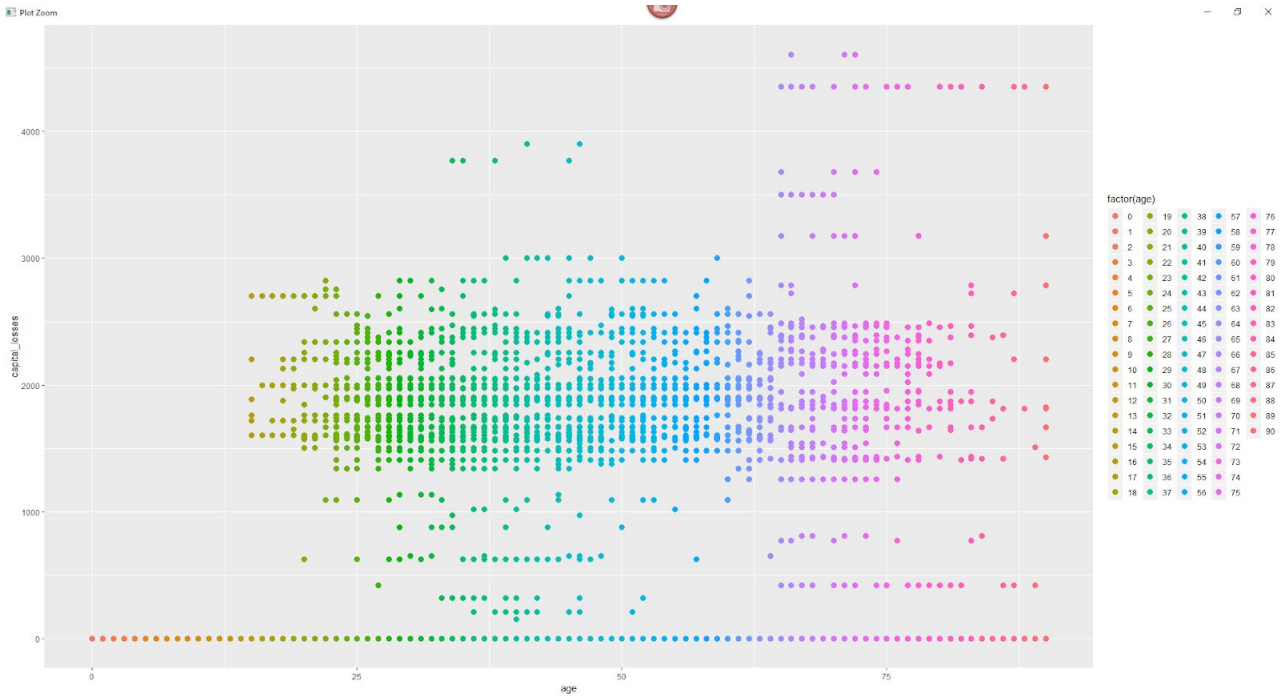
```
pl <- ggplot(data = Census_income_data, aes(x = weeks_worked, y = stock_dividends))
pl2 <- pl + geom_point( aes(color = factor(weeks_worked)), size = 3)
print(pl2)
```

Analyzing

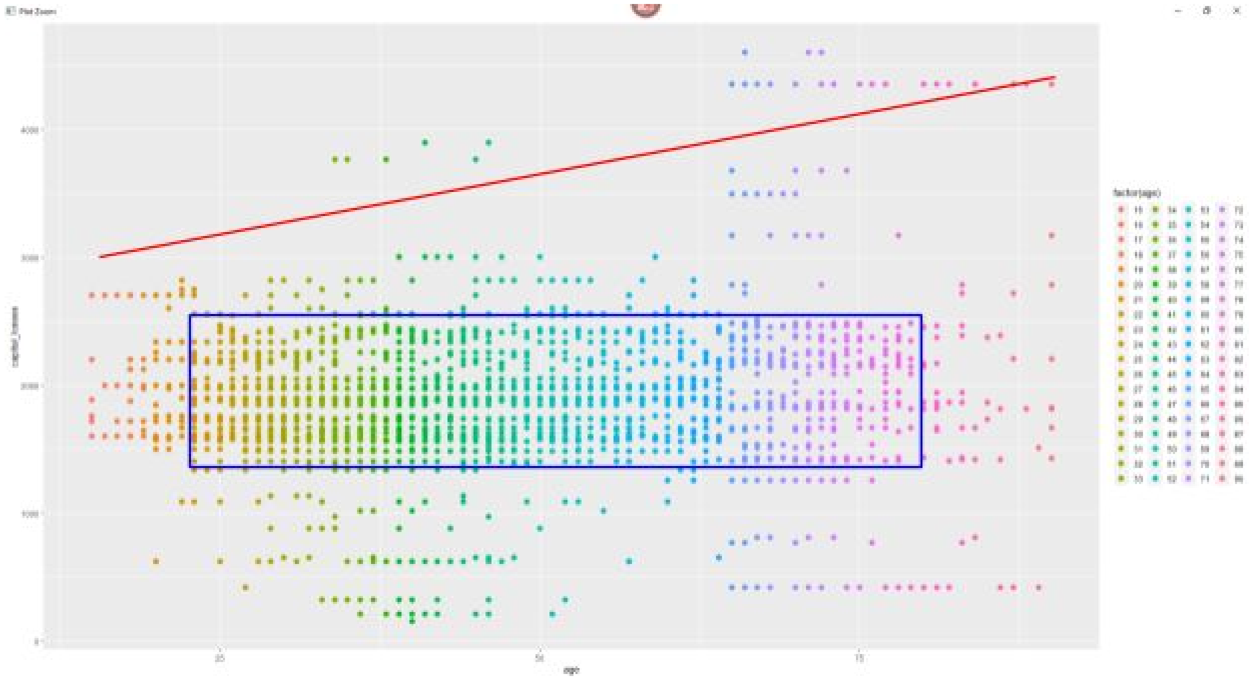
According to the graph above, it represents 3 things. First, green line, an upward trends from the working 1 week/year to 40 weeks/year it shows the slightly increasing the amount of money earns from the stock dividends. However, the second red line presents the a steady downward trends which continue from working 41 weeks/year to 51 weeks/year. Third, the blue oval, the non-working people and worker who works 52 weeks/year. The graph shows that these 2 groups of people have one thing in common, they earns a lot from the stock dividends this can be leads to an assumption that they either own a business itself, very high ranked employees, experienced in working or live by playing stocks.

In conclusion, even though this analysis doesn’t shows much relationship between these 2 attributes such as how the more the worker works the more they earns the stock dividends, but it gives an idea of the 2 types of US people who earns a lot from the stock dividends.

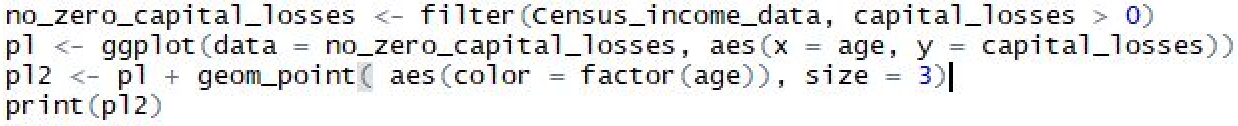
7.5. Age VS Capital Losses



As for the graph presented above, it is clear that capital losses equals to 0 is an outliers and shouldn't be in the graph it affects the visualization. Therefore, it will be cleaned out.



Code



```
no_zero_capital_losses <- filter(Census_income_data, capital_losses > 0)
pl <- ggplot(data = no_zero_capital_losses, aes(x = age, y = capital_losses))
pl2 <- pl + geom_point( aes(color = factor(age)), size = 3)
print(pl2)
```

Analyzing

As points where the capital losses equals to zero has been removed, the age which less than 15 also have been automatically removed as well because it seems like all population that age less than 15 doesn't gain or loss any capital according to this dataset.

Therefore, as the visualization of this scatter plot is getting better, it can be clearly seen that there are 2 major facts have been presented to us. First, the blue rectangle, it presents that almost every range of age, people are highly lost their capital around 1500 to 2500$. Next, the red line, the higher the age, the more maximum capital losses they loss or tendency to loss more capital. However, most of the US population are losing the capital in the similar amounts and not much compared to what they gains which should be alright.

(Note: read about this more on capital gains statistical analysis #5.3-4)