Research on Diabetes and Income in US

Group F

January 29, 2017

Introduction

In this paper, we research the relaship between diabetes rates and income number in US overtime and seek to explore the pattern between the two. We will use R to analyze the data patterns in this data set. We will use a few build-in functions in R library to demonstrate the attribute of the data set and the relation between different variables. We will also give charts to better present the relationship.The timeframe in scope of this research is from 2009 to 2013.

1 Chapter 1: Data Pre-Processing

We will need to clean up the data and take only what ever we need from the data set now. For the diabete data set, since the popluation can vary in different states and counties, we will use the percentage as the parameter. For the income data set, we will use the "median income in past 3 years" to represent the income values in different states in different years. Since those data comes in with different data sets, we will need to combine data and create new tables with only the data we needed.

First of all, we are going to shrink the data set of diabetes rate in every state. This task itself can be perform in excel itself. We are going to combine all the counties in one state into one data line and take the average diabete value to represent the whole state.

Then, we will need to process the data on the income level. We have income data separated in 5 different data sets and now we are going to consolidate the data in every state across 2009-2013.

Both pre-processed dataset can be found at the end of this paper.

2 Chapter 2: High level property of the data

Average diabetes rate across different states: By running the below R command against the new data set, we can see that the average diabete rate across this 5 year period in different state can differ by more than 9 percent and ranges from 6.34 percent to 15.73 percent. The variance of the data is about 4, which looks relatively small compared to data itself and suggest that the data is trend to cluster in similar area.

```
updatedDM <- read.table("updatedDM.csv",sep=",",header=T)

updatedDM$mean=rowMeans(updatedDM[,c("Year2009", "Year2010","Year2011","Year2012","Year2013")], na.rm=TRUE)
updatedDM$mean
```

```
##  [1] 14.598  7.102 10.438 12.632 10.718  6.344  8.322 10.874  8.140 11.756
## [11] 12.344  8.796  9.346 10.044 11.338  9.686 10.666 12.616 12.768 10.100
## [21] 10.540  8.668 11.162  8.616 14.268 11.180  8.818  9.562  8.896  9.306
## [31]  9.210  8.810  9.632 12.018  9.890 11.650 12.424  9.786 10.852  8.332
## [41] 13.598 10.140 12.774 10.566  8.044  7.898 11.212  9.750 13.406  9.244
## [51] 15.730 15.724
```

```
min(updatedDM$mean,na.rm = TRUE)
```

```
## [1] 6.344
```

```
max(updatedDM$mean,na.rm = TRUE)
```

```
## [1] 15.73
```

```
mean(updatedDM$mean,na.rm = TRUE)
```

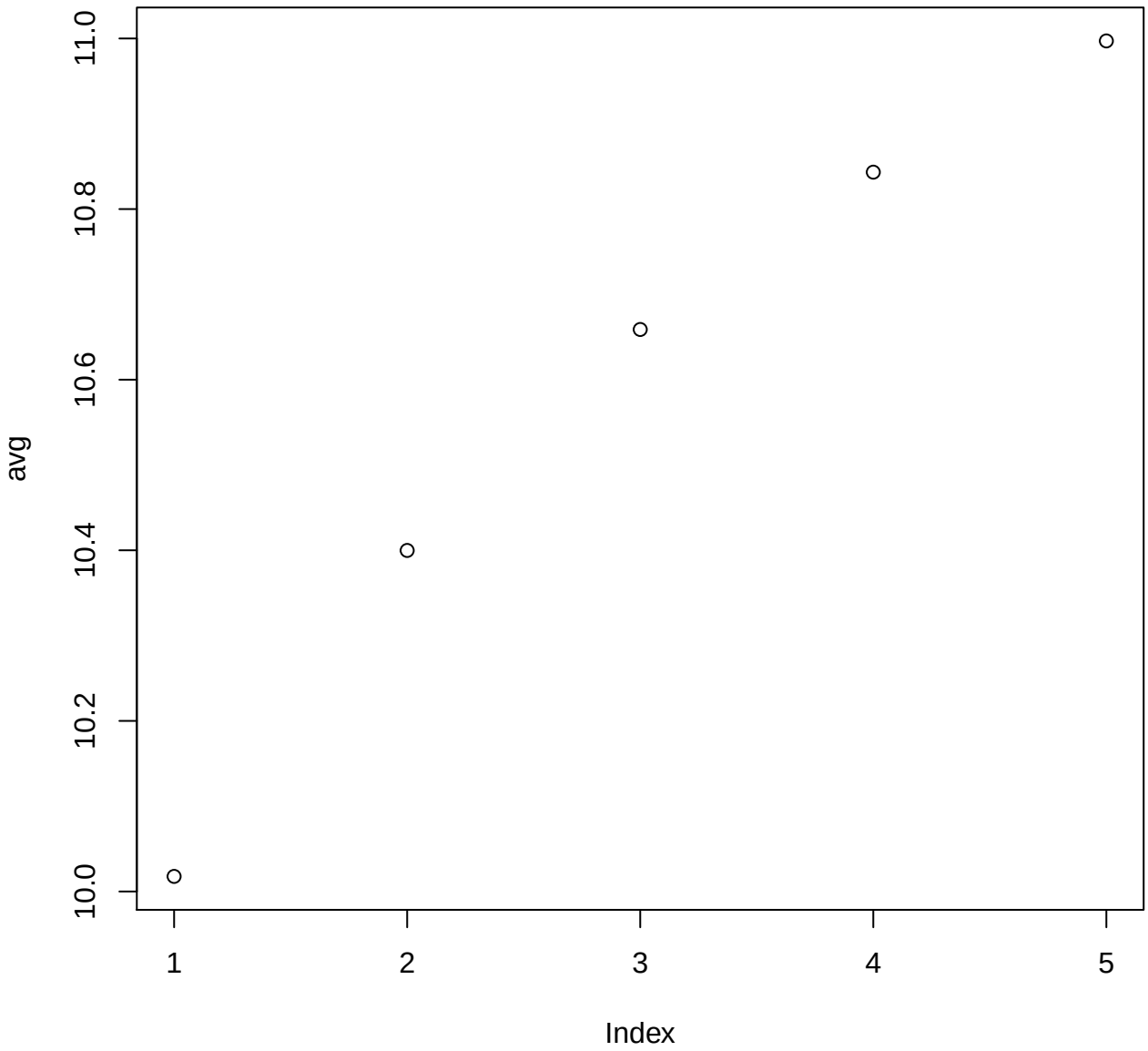
```
## [1] 10.58335
```

```
var(updatedDM$mean,na.rm = TRUE)
```

```
## [1] 4.359945
```

Average diabetes rate across different states through 5 years: In the below R code, we are calculating the average diabetes rate in the whole US across the time series from 2009 to 2013. We can easily see that the tread is upward through these years.

```
avg<-c(mean(updatedDM$Year2009,na.rm = TRUE)
       ,mean(updatedDM$Year2010,na.rm = TRUE)
       ,mean(updatedDM$Year2011,na.rm = TRUE)
       ,mean(updatedDM$Year2012,na.rm = TRUE)
       ,mean(updatedDM$Year2013,na.rm = TRUE))
plot(avg)
```



3 Chapter 3: Relationship between Diabetes rate and income

Sum of the mpg number in of all Cars

```
sum(mtcars$mpg,na.rm = TRUE)
```

```
## [1] 642.9
```

Median mpg

```
median(mtcars$mpg,na.rm = TRUE)
```

```
## [1] 19.2
```

Variance,variance is used to measure how much the data are divided

```
var(mtcars$mpg,na.rm = TRUE)
```

```
## [1] 36.3241
```

Standard deviation,standard deviation is used to measure how much the data are divided

```
sd(mtcars$mpg,na.rm = TRUE)
```

```
## [1] 6.026948
```

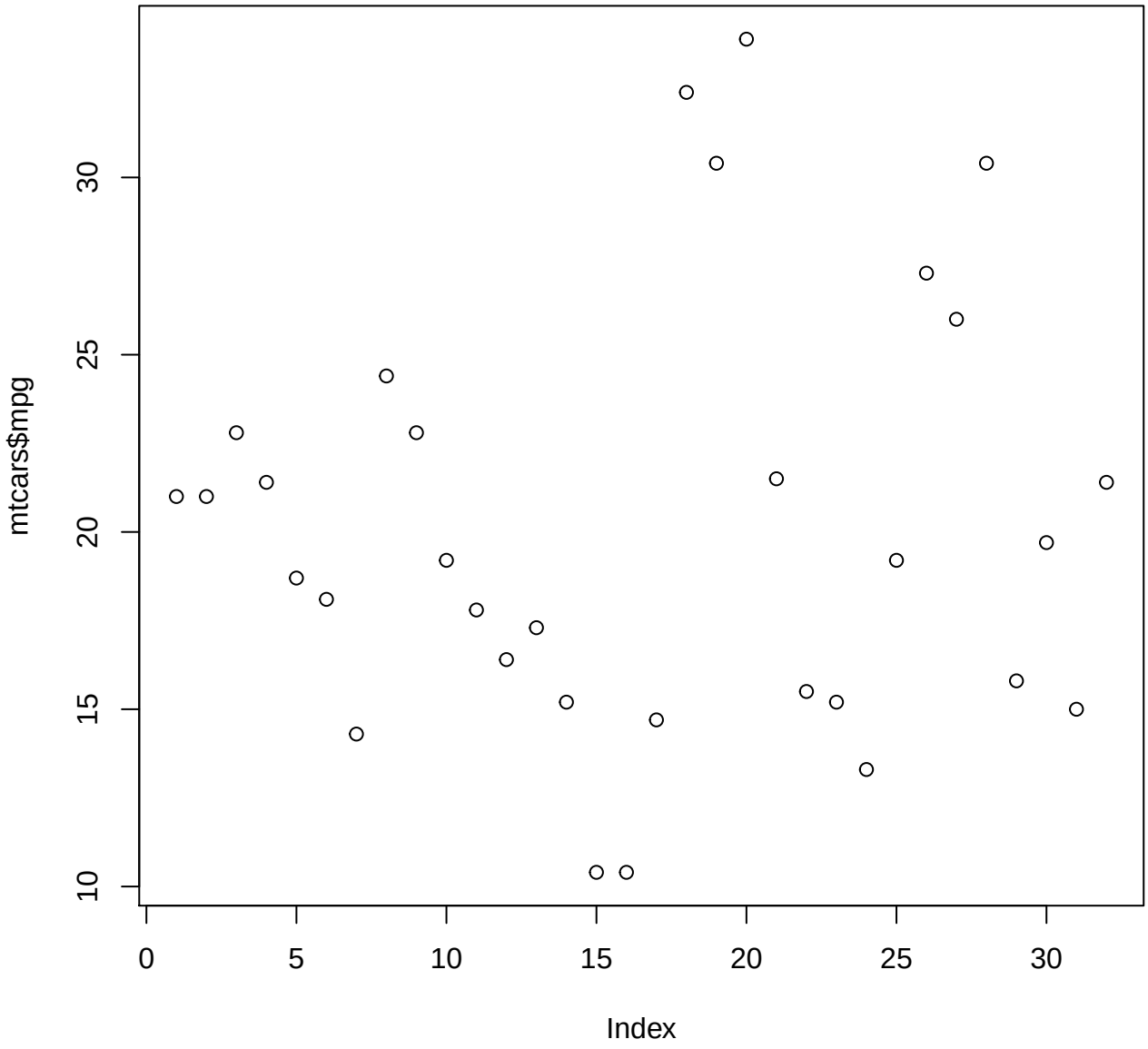
MAD, mad is mean absolute deviation, which is also used to distribution of the data values

```
mad(mtcars$mpg,na.rm = TRUE)
```

```
## [1] 5.41149
```

Plot sample

```
plot(mtcars$mpg)
```



4 Chapter 2: Data Frame and Relation

In the last chapter, we mainly looked at the properties of the data set by checking 1 variable each time. Now we are going to explore how different variables in the set are related to each other.

create a data frame called cars and get 4 varialbes fromt mtcars data set then put them into the data frame. use gear as orderred factor in this cars data frame for the am variable, we are replacing value 0 using "auto" and value 1 using "manual". Also transfer it into a factor.

```
cars<-transform(mtcars[c(1,2,9,10)],gear=ordered(gear),am=factor(am,label=c("auto","manual")))

str(cars,vec.len=2)
```

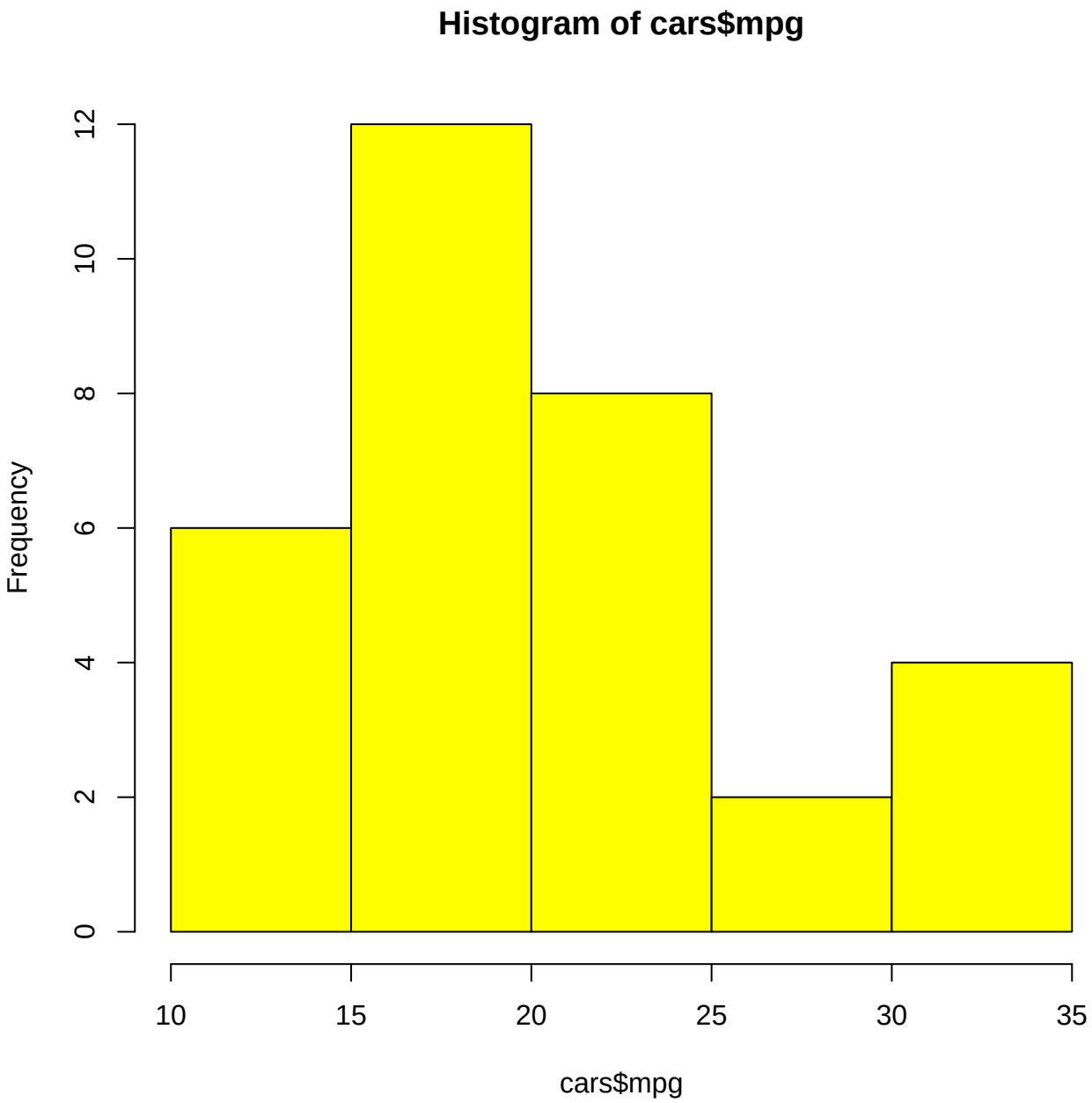
```
## 'data.frame': 32 obs. of  4 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 ...
##  $ cyl : num  6 6 4 6 8 ...
##  $ am  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 ...
##  $ gear: Ord.factor w/ 3 levels "3"<"4"<"5": 2 2 2 1 1 ...
```

We are going to put the am field into a table and display it As shown in the below result, there are 19 auto and 13 manual cars in the data set. The percentage are 0.59 and 0.41 respectively

```
amtable<-table(cars$am)
amtable

##
##   auto manual
##     19     13

amtable/sum(amtable)

##
##    auto  manual
## 0.59375 0.40625
```

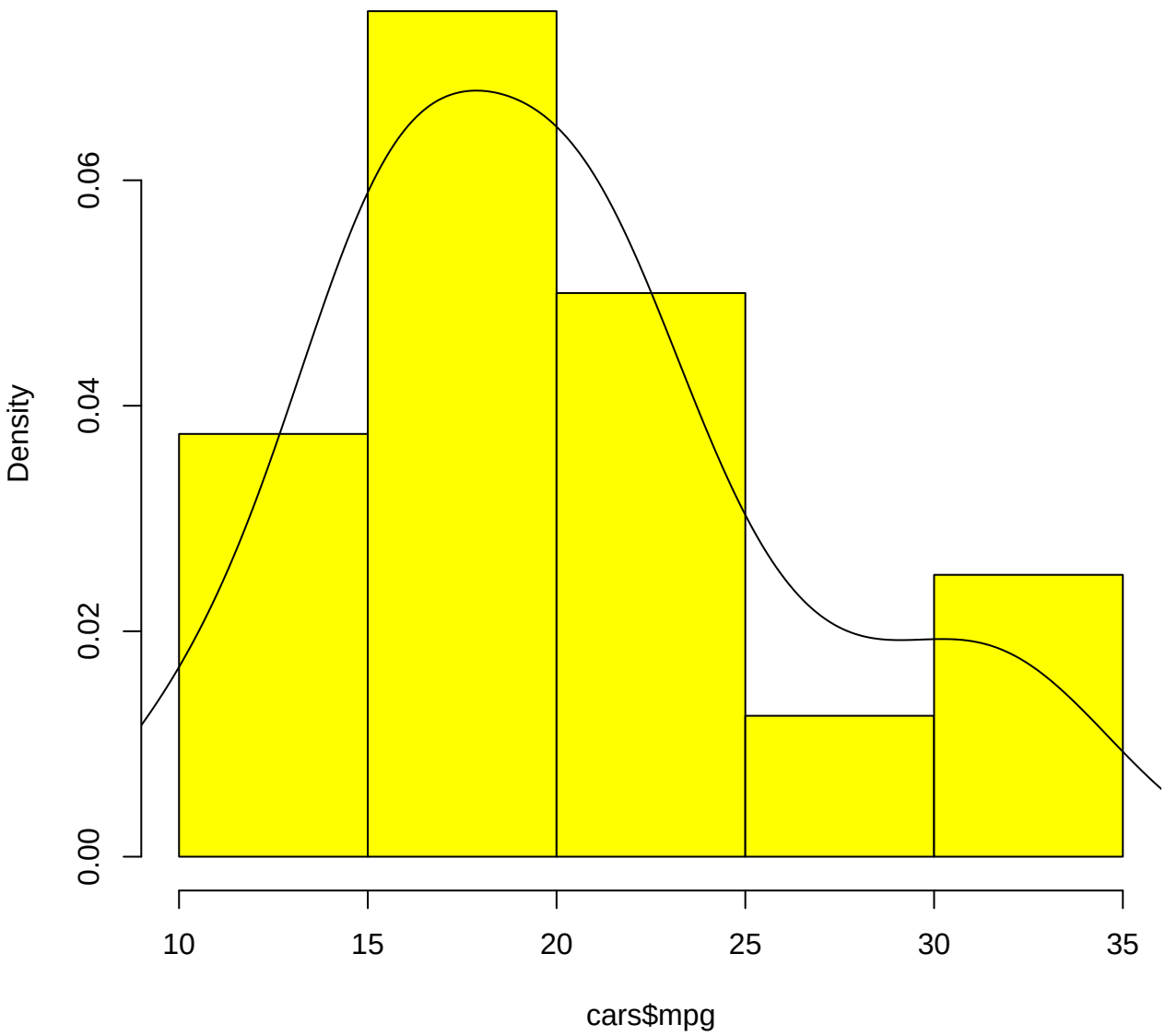
Here we are creating a histogram graph showing the frequency of each of the mpg ranges in the cars data frame. We see 15-20 mpg is showing the highest frequency and 25-30 is showing the lowest

```
hist(cars$mpg,col="yellow")
```



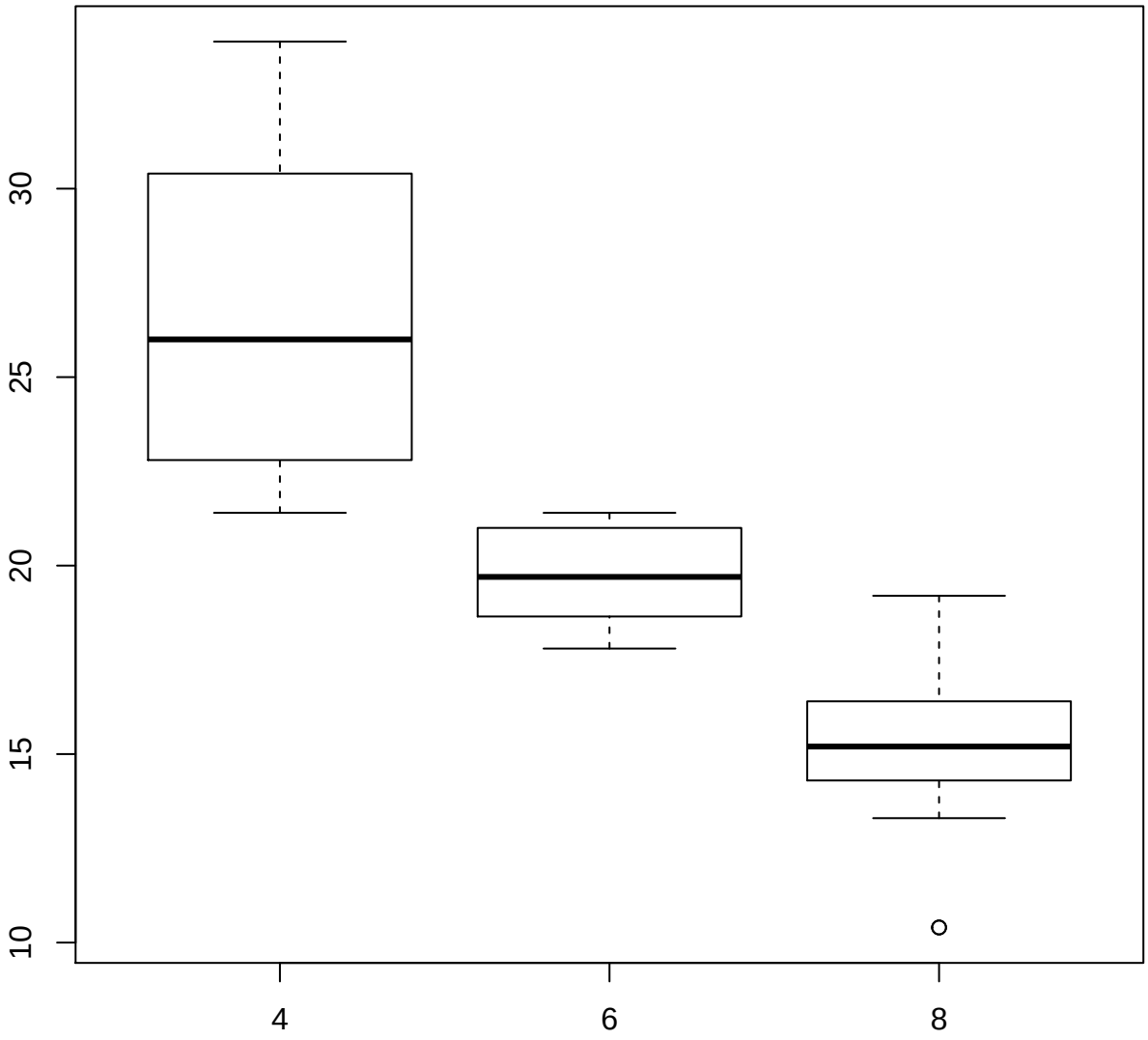
Here we are creating a density plot for the car-mpg field and also add it into the hitogram we had in the last step. We see the shape of the density plot is matching the shape of histogram

```
mpgdens<-density(cars$mpg)
hist(cars$mpg,col="yellow",freq=FALSE)
lines(mpgdens)
```



Here we first define cyl as a factor in the cars data frame, then we use the newly defined factor as x variable and mpg as y variable to create boxplot

```
cars$cyl<-as.factor(cars$cyl)
boxplot(mpg~cyl,data=cars)
```



correlation: correlation measure the relationship of 2 variables. It can range from -1 to 1, where negative value means negatively related, positive means postively related and 0 means not related.

We see below that the mpg and weight is having very high negative relationship. This is also true in real world as the car with higher weight consumes more gas and have more powerful engine, which will in turn lower the mpg

```
with(mtcars,cor(mtcars$mpg,mtcars$wt))

## [1] -0.8676594
```

mpg and displacement also shows significant negative relationship. This is because better engine usually have higher displacement. Since more displacement means more gas consumption in one time unit. Mpg of those cars will be lower

```
with(mtcars,cor(mtcars$mpg,mtcars$disp))

## [1] -0.8475514
```

We see below that the mpg and house power is having very high negative relationship. This is similar to the relationship between mpg and weight, as more powerful engine tends to consume more gas , which will in turn lower the mpg

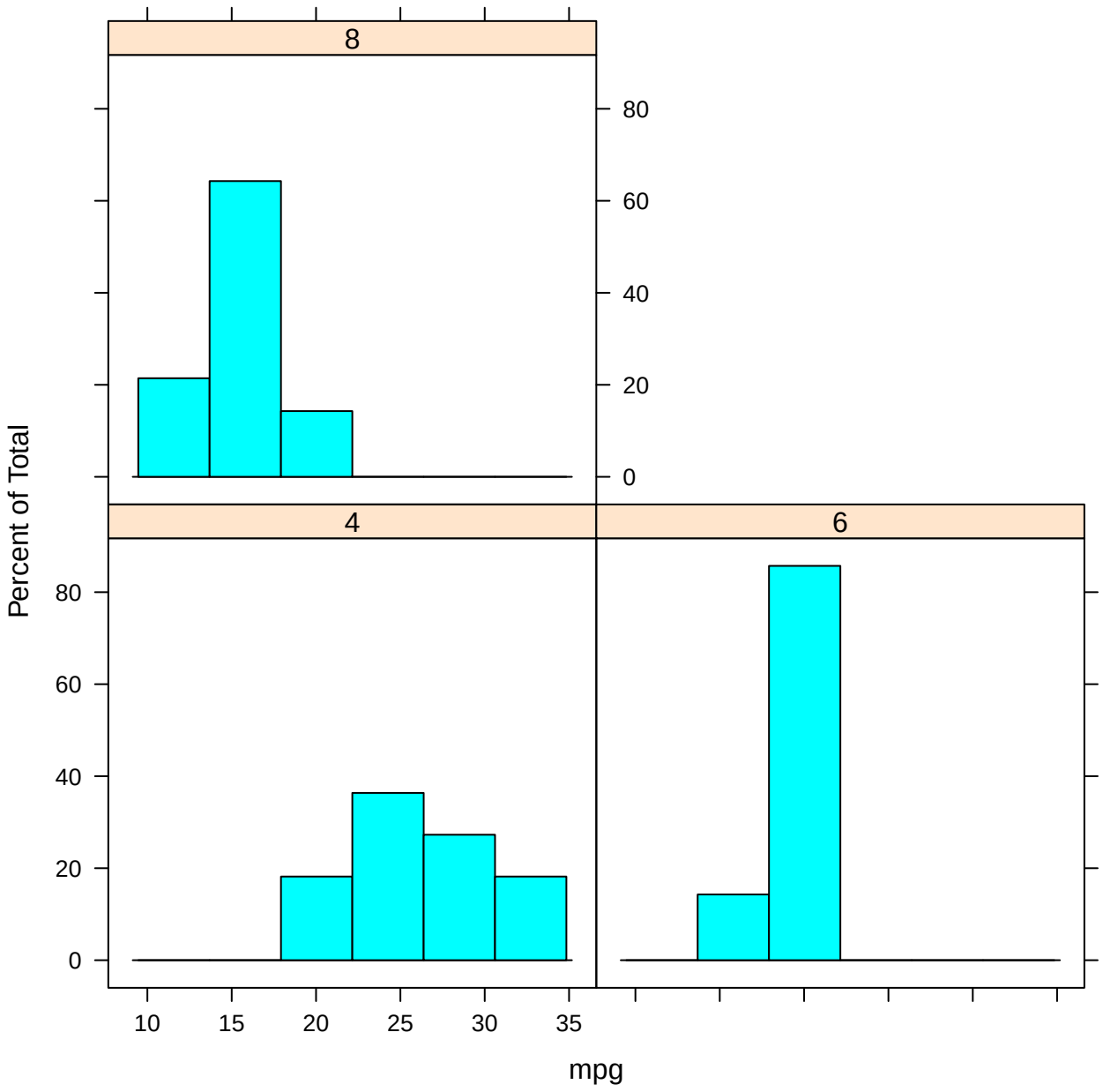
```r
with(mtcars,cor(mtcars$mpg,mtcars$hp))

## [1] -0.7761684
```

5 Chapter 3: Modelling

In this chapter we will build different models to further explore the relationship between different variables we found inlast chapter.

Here we are going to build a histogram of mpg of each different types of engine based on the cyl number on each engine.

```r
library("lattice")
histogram(~mpg|cyl,data=cars)
```



Observation: We can clearly see that the mpg range of each engine tends to cluster together. The shows the cyl of each engine does impact the overall mpg performance and since the engines are structurely similar to each other, the mpg trends to be very close to each other.

We will now build a table of mean using a AOVModel we build on cars data set and mpg/cyl field.

```
AOVModel<-aov(mpg ~cyl,data=cars)
AOVModel

## Call:
##    aov(formula = mpg ~ cyl, data = cars)
##
## Terms:
##                       cyl Residuals
## Sum of Squares  824.7846  301.2626
## Deg. of Freedom        2        29
##
## Residual standard error: 3.223099
## Estimated effects may be unbalanced

model.tables(AOVModel,type="mean")

## Tables of means
## Grand mean
##
## 20.09062
##
##  cyl
##         4     6    8
##     26.66 19.74 15.1
## rep 11.00  7.00 14.0
```

Observation: In the above result, we have the data for the overall average value of mpg and the average of each of the cyl engines. This confirms our idea before that the lower cyl engine tends to have higher mpg since they tend to consume less gas in a time unit.

We will now build a table of mean using a AOVModel we build on cars data set and mpg/am field.

```
AOVModel<-aov(mpg ~am,data=cars)
AOVModel

## Call:
##    aov(formula = mpg ~ am, data = cars)
##
## Terms:
##                        am Residuals
## Sum of Squares  405.1506  720.8966
## Deg. of Freedom        1        30
##
## Residual standard error: 4.902029
## Estimated effects may be unbalanced

model.tables(AOVModel,type="mean")

## Tables of means
## Grand mean
##
## 20.09062
##
##  am
##      auto manual
##     17.15  24.39
## rep 19.00  13.00
```

Observation: In the result, we see the auto cars have lower mpg than manual cars. This is due to that auto transmission tend to have a delay when switching transmission and lack the ability to forecast the road condition. In contrast, since human driver can such ability using manual, they can save gas and achieve higher mpg

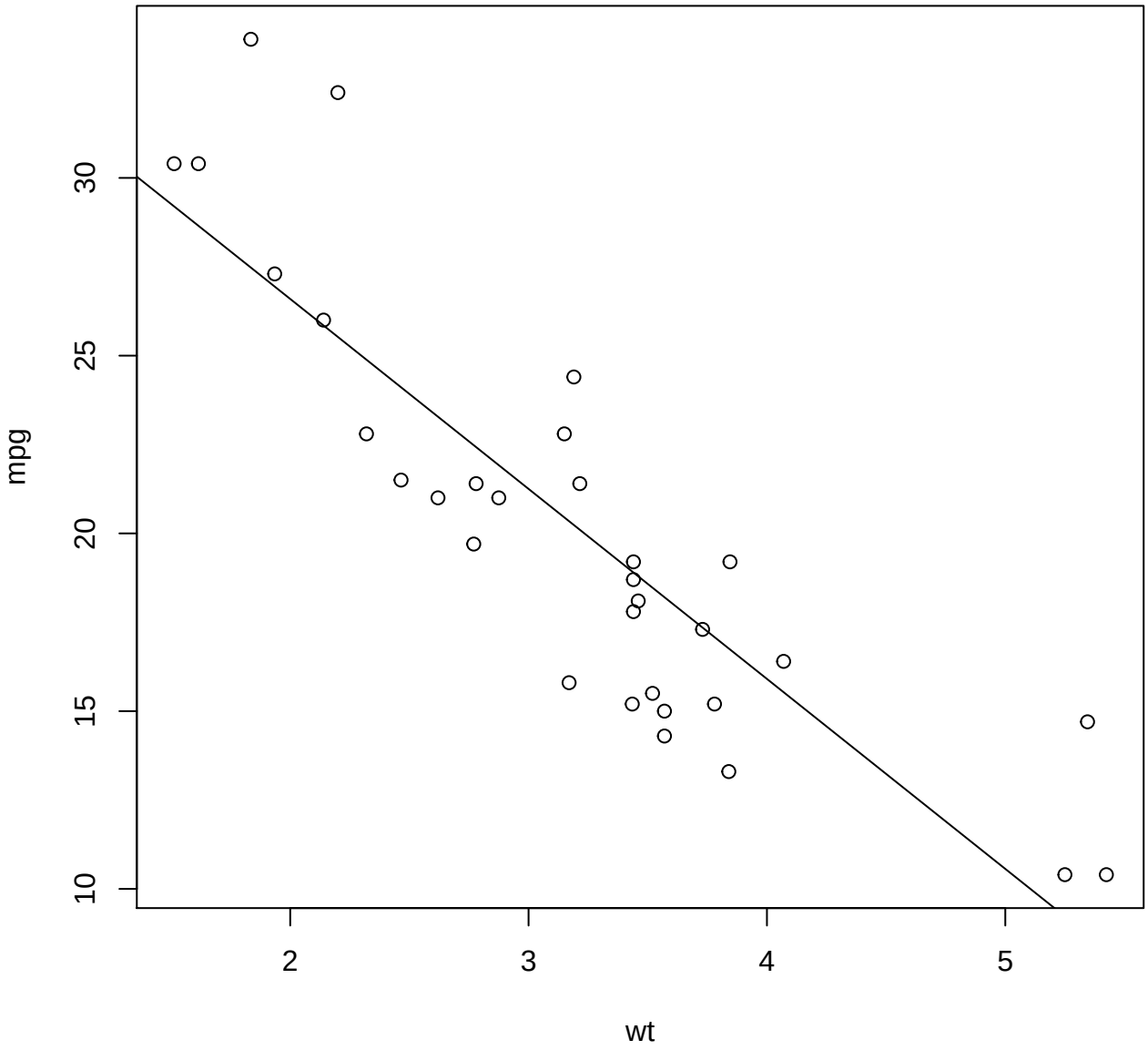
We will now build a table of mean using a AOVModel we build on cars data set and mpg/gear field.

```
AOVModel<-aov(mpg ~gear,data=cars)
AOVModel

## Call:
##    aov(formula = mpg ~ gear, data = cars)
##
## Terms:
##                      gear Residuals
## Sum of Squares  483.2432  642.8040
## Deg. of Freedom        2        29
##
## Residual standard error: 4.708042
## Estimated effects may be unbalanced

model.tables(AOVModel,type="mean")

## Tables of means
## Grand mean
##
## 20.09062
##
##  gear
##         3     4     5
##     16.11 24.53 21.38
## rep 15.00 12.00  5.00
```

Observation: In the result, we see the auto cars have 4 gear is having the most mpg efficiency, with 5 comes after it and 3 in the last position.

At last we will build a liner model with mpg and wt variable in mtcars data set.

```
Model<-lm(mpg ~wt,data=mtcars)
coef.Model<-coef(Model)
coef.Model

## (Intercept)          wt
##   37.285126   -5.344472

plot(mpg ~wt,data=mtcars)
abline(coef=coef.Model)
```



In the plot chart, we clear see that in the liner model, there is a negative relation between wt and mpg. Also, the plots are fitting pretty closely with the liner model we build.

6 Summary

In this paper, we explored different functions in R by analzing the mtcars data set. We have analyzed different attributes of the data set and worked on the relation between different vairalbes using different measures. We also build statistic models to further simply the patterns in the variables.

7 DataSet

Diagnosed Diabetes Percentage	2009	2010	2011	2012	2013
Alabama	13.9	14.08	14.48	15.04	15.49
Alaska	6.78	7.37	7.2	7.14	7.02
Arizona	9.96	10.08	10.46	10.6	11.09
Arkansas	11.98	12.49	12.68	12.67	13.34
California	10.14	10.48	10.8	10.88	11.29
Colorado	6	6.15	6.49	6.59	6.49
Connecticut	7.75	8.29	8.53	8.56	8.48
Delaware	10.33	10.6	10.57	11.2	11.67
District of Columbia	8.2	8.1	8.2	8.1	8.1
Florida	11.26	11.6	11.88	12.01	12.03
Georgia	11.8	12.35	12.43	12.45	12.69
Hawaii	8.24	9.16	8.96	8.86	8.76
Idaho	8.87	9.43	9.48	9.71	9.24
Illinois	9.05	9.67	10.29	10.7	10.51
Indiana	10.59	10.95	11.35	11.76	12.04
Iowa	8.75	9.19	9.86	10.35	10.28
Kansas	10.24	10.57	10.8	10.84	10.88
Kentucky	12.27	12.55	12.46	12.59	13.21
Louisiana	12.15	12.67	12.96	13.17	12.89
Maine	9.66	10.16	10.23	10.26	10.19
Maryland	10.24	10.27	10.48	10.65	11.06
Massachusetts	8.51	8.63	8.51	8.63	9.06
Michigan	10.65	11.07	11.32	11.4	11.37
Minnesota	8.55	8.47	8.65	8.55	8.86
Mississippi	13.4	14.07	14.53	14.68	14.66
Missouri	10.36	10.91	11.53	11.62	11.48
Montana	8.39	8.59	8.86	8.91	9.34
Nebraska	9.29	9.5	9.47	9.72	9.83
Nevada	8.82	8.95	8.87	8.75	9.09
New Hampshire	8.68	8.95	9.4	9.62	9.88
New Jersey	8.89	9.05	9.29	9.32	9.5
New Mexico	7.87	8.66	9.08	9.19	9.25
New York	9.08	9.59	9.95	9.91	9.63
North Carolina	11.36	11.94	12.1	12.36	12.33
North Dakota	9.58	9.53	10	10.27	10.07
Ohio	10.99	11.16	11.69	11.98	12.43
Oklahoma	11.48	12.27	12.65	12.69	13.03
Oregon	8.93	9.56	9.72	10.23	10.49
Pennsylvania	10.45	10.64	10.99	10.9	11.28
Rhode Island	7.58	7.7	8.3	8.84	9.24
South Carolina	12.59	13.22	13.63	14.18	14.37
South Dakota	9.55	10.13	10.1	10.59	10.33
Tennessee	11.68	12.15	12.8	13.44	13.8
Texas	10.3	10.63	10.8	10.7	10.4
Utah	7.68	7.82	8.12	8.29	8.31
Vermont	7.38	7.85	8.03	8.14	8.09
Virginia	10.86	11.15	11.34	11.5	11.21
Washington	9.29	9.68	9.76	9.88	10.14
West Virginia	13.1	13.24	13.25	13.5	13.94
Wisconsin	8.83	9.14	9.11	9.43	9.71
Wyoming	14.33	15.19	15.91	16.25	16.97
Puerto Rico	14.31	15.14	15.91	16.25	17.01

State	2009Income	2010Income	2011Income	2012Income	2013Income
Alabama	42651.60	42218.35	42245.00	43350.04	43195.94
Alaska	63505.11	61872.44	60566.33	61065.77	61730.67
Arizona	47106.02	47093.48	48319.00	48688.58	49562.34
Arkansas	39391.83	38599.90	39806.33	40605.99	40760.32
California	56862.50	56418.36	56074.00	56221.86	56883.25
Colorado	59963.68	59669.49	59803.00	60179.85	60727.26
Connecticut	65213.43	65958.13	67165.33	66843.79	66904.62
Delaware	53031.65	53195.58	55420.67	54306.54	52838.62
District of Columbia	55280.00	55280.00	56565.67	60533.90	61364.70
Florida	45896.96	45350.48	46136.00	46174.85	47114.24
Georgia	46569.95	44991.88	45642.33	47171.30	47958.30
Hawaii	61055.39	59124.64	59605.33	59747.83	59881.77
Idaho	48299.35	47528.18	48348.33	48640.14	49846.59
Illinois	53412.82	52810.66	52801.00	52284.45	54043.92
Indiana	46579.25	46155.52	46165.67	46706.89	47804.95
Iowa	50422.40	50503.78	51321.67	52109.68	53695.85
Kansas	47527.20	46722.26	46847.33	48537.57	50003.33
Kentucky	41489.42	42090.70	42331.00	41687.14	41706.96
Louisiana	42527.81	41896.48	42946.00	40659.77	40461.86
Maine	48032.09	48081.36	49648.00	50121.31	50487.09
Maryland	65182.77	64596.03	67468.67	69920.01	69826.10
Massachusetts	59981.02	60923.19	62808.67	64153.34	64372.93
Michigan	48888.49	47870.76	48281.00	49549.26	50055.74
Minnesota	56955.52	55063.15	56868.67	58640.63	61161.85
Mississippi	36650.06	36849.63	39078.33	39591.79	40193.82
Missouri	47407.89	47459.74	48058.00	48247.78	49402.53
Montana	42778.18	42004.62	41753.33	43226.06	43863.83
Nebraska	50333.38	51503.91	53926.67	54755.01	54777.12
Nevada	53964.10	53081.83	51263.33	49759.04	47371.12
New Hampshire	66654.02	66303.30	67287.00	68415.23	69453.28
New Jersey	64143.21	65172.93	65071.67	65548.09	64669.52
New Mexico	43789.91	43997.69	44732.33	44604.75	43221.39
New York	50372.41	50656.39	51547.33	50599.54	51553.93
North Carolina	43228.84	43274.56	44786.67	44620.36	43395.41
North Dakota	49450.14	50846.89	53827.33	55672.69	55946.21
Ohio	47808.98	46752.25	46696.00	46092.60	45887.16
Oklahoma	45506.74	45576.77	47008.33	47755.16	47691.13
Oregon	50865.66	50938.28	51735.33	52555.39	54066.99
Pennsylvania	49828.72	49826.47	50086.67	51245.22	52767.61
Rhode Island	53584.09	52771.44	52142.33	53495.47	55158.84
South Carolina	42944.54	42058.83	42065.00	43078.04	43437.04
South Dakota	48416.27	48168.04	47352.67	48461.47	51165.12
Tennessee	40894.62	40025.65	41524.33	42265.84	43302.79
Texas	47143.20	47600.59	49195.00	50591.03	52169.07
Utah	58721.83	59857.27	58438.00	58234.87	59877.30
Vermont	50618.57	53490.45	54804.67	55808.08	54982.06
Virginia	61150.60	61543.78	62775.67	64042.53	66014.52
Washington	58963.61	58329.70	59369.67	59789.57	60692.29
West Virginia	40627.41	40824.22	42801.33	43765.43	42580.62
Wisconsin	51762.62	51484.31	52574.33	53082.90	54342.08
Wyoming	52009.80	53235.52	54458.33	56043.93	56834.52