Project1-1

Zhihao Chen zc4284

0. Introduction

The first data was called US voter turnout, which includes number of age-eligible voters versus total votes counted by state and year. The second data was called US average tuition, which includes the avrage tuition by state and year. Both of these data were found on the github rfordatascience website, and they are interesting becasue I think there might be a potential correlation between the voter turnout and college tuition in some area of the US.

1. Tidying: Rearrange Wide/Long

The tuition data was first pivot longer to create new rows for each state with each year, and then the year was seperate into two parts to make the year more tidyer. Unnecessary columns are removed. For the turnout data, columns that contain unnecessary information were removed.

```
tuition_2 <- tuition%>%pivot_longer(c(2,3,4,5,6,7,8,9,10,11,12,13))%>%separate(name, into=c("yea
r","unknown"), convert=T)%>%rename(state = State)%>%select(-unknown)
glimpse(tuition_2)
```

```
## Observations: 600
## Variables: 3
## $ state <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Al…
## $ year  <int> 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013…
## $ value <dbl> 5682.838, 5840.550, 5753.496, 6008.169, 6475.092, 7188.954…
```

```
turnout_2 <- turnout%>%select(-X, -icpsr_state_code, -alphanumeric_state_code)
glimpse(turnout_2)
```

```
## Observations: 936
## Variables: 4
## $ year            <int> 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, …
## $ state           <fct> United States, Alabama, Alaska, Arizona, Arkansa…
## $ votes           <int> 83262122, 1191274, 285431, 1537671, 852642, 7513…
## $ eligible_voters <int> 227157964, 3588783, 520562, 4510186, 2117881, 24…
```

2. Joining/Merging

Data tuition_2 was joined with data turnout_2 using left_join, and the joined data was piped into na.omit to remove any row with NA. These two data were joined by two columns, year and states, so there is no data being lost during the joining. Column value was renamed to avg_tuition.

```
temp <- tuition_2%>%left_join(turnout_2)%>%na.omit()%>%rename(tuition = value)
```

```
## Joining, by = c("state", "year")
```

```
## Warning: Column `state` joining character vector and factor, coercing into
## character vector
```

```
glimpse(temp)
```

```
## Observations: 273
## Variables: 5
## $ state           <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Ala…
## $ year            <int> 2004, 2008, 2010, 2014, 2004, 2006, 2008, 2010, …
## $ tuition         <dbl> 5682.838, 6475.092, 8071.134, 9496.084, 4328.281…
## $ votes           <int> 1890317, 2105622, 1503232, 1191274, 314502, 2383…
## $ eligible_voters <int> 3292608, 3454510, 3472582, 3588783, 452124, 4651…
```

3. Wrangling

A new column called rate was calculated with votes and eligible_voters, which represents the actual turnout rate for a given year and state. To understand the center and spread of the tution, the mean and standard deviation of tuition was calculated, and we can see there is a great difference in tuition across states. In order to better understand the variance of tuition across the US, a robust statistic is required, since the range of variable tuition is large and may contain outliers. Thus, the median absolute deviation (MAD) of avg_tuition was calculated. This statistic measures the dispersion of the tuition across states, and a value of 1602.384 of MAD indicates a great variance in the tuition. Next, the data was arranged by rate to see which state has the highest voting turnout, and interestingly Minnesota has a relatively high voting turnout from 2004 to 2012. The min and max of number of eligible voters base on state were found, the min and max of number of votes base on year were found.

By grouping by state and year, we can measure the mean and see the 1 over rate. And then the quantile of tuition of each state was found, we can see a rough distribution of tuition can be observed. Next, a correlation was found between turnout rate and tuition of California, and there is no correlation between them.

```
# mutate()
temp <- temp%>%mutate(rate = votes/eligible_voters)
glimpse(temp)
```

```
## Observations: 273
## Variables: 6
## $ state           <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Ala…
## $ year            <int> 2004, 2008, 2010, 2014, 2004, 2006, 2008, 2010, …
## $ tuition         <dbl> 5682.838, 6475.092, 8071.134, 9496.084, 4328.281…
## $ votes           <int> 1890317, 2105622, 1503232, 1191274, 314502, 2383…
## $ eligible_voters <int> 3292608, 3454510, 3472582, 3588783, 452124, 4651…
## $ rate            <dbl> 0.5741093, 0.6095284, 0.4328860, 0.3319437, 0.69…
```

```
# group_by(), summarize(), select()
temp%>%group_by(state)%>%summarize(mean(tuition), sd(tuition))
```

```
## # A tibble: 49 x 3
##    state       `mean(tuition)` `sd(tuition)`
##    <chr>                 <dbl>         <dbl>
##  1 Alabama               7431.         1697.
##  2 Alaska                5376.          716.
##  3 Arizona               7678.         2398.
##  4 Arkansas              6702.          688.
##  5 California            7210.         1920.
##  6 Colorado              7071.         1832.
##  7 Connecticut           9376.         1146.
##  8 Delaware              9616.         1351.
##  9 Florida               5039.         1231.
## 10 Georgia               6010.         1682.
## # … with 39 more rows
```

```
temp%>%select(tuition)%>%
  mutate(median = median(tuition), dev = tuition-median, absdev = abs(dev), MAD=median(absdev))
```

```
## # A tibble: 273 x 5
##    tuition median    dev absdev   MAD
##      <dbl>  <dbl>  <dbl>  <dbl> <dbl>
##  1   5683.  7476. -1793.  1793. 1602.
##  2   6475.  7476. -1001.  1001. 1602.
##  3   8071.  7476.   595.   595. 1602.
##  4   9496.  7476.  2020.  2020. 1602.
##  5   4328.  7476. -3148.  3148. 1602.
##  6   4919.  7476. -2557.  2557. 1602.
##  7   5075.  7476. -2400.  2400. 1602.
##  8   5759.  7476. -1717.  1717. 1602.
##  9   6026.  7476. -1450.  1450. 1602.
## 10   6149.  7476. -1327.  1327. 1602.
## # … with 263 more rows
```

```
# arrange()
temp%>%arrange(desc(rate))
```

```
## # A tibble: 273 x 6
##    state          year tuition   votes eligible_voters  rate
##    <chr>         <int>   <dbl>   <int>           <int> <dbl>
##  1 Minnesota      2004   8144. 2842912         3609185 0.788
##  2 Minnesota      2008   9024. 2921147         3740142 0.781
##  3 Minnesota      2012  10793. 2950780         3861598 0.764
##  4 Wisconsin      2004   6575. 3016288         4006948 0.753
##  5 Maine          2004   7058.  751519         1003792 0.749
##  6 Wisconsin      2008   7373. 2997086         4120694 0.727
##  7 Oregon         2004   6579. 1851671         2550887 0.726
##  8 New Hampshire  2008  11168.  719643          992226 0.725
##  9 Maine          2008   8764.  744456         1036242 0.718
## 10 Colorado       2008   6284. 2422236         3382959 0.716
## # … with 263 more rows
```

```
temp%>%group_by(state)%>%summarize(min(eligible_voters), max(eligible_voters))
```

```
## # A tibble: 49 x 3
##    state       `min(eligible_voters)` `max(eligible_voters)`
##    <chr>                        <int>                  <int>
##  1 Alabama                    3292608                3588783
##  2 Alaska                      452124                 520562
##  3 Arizona                    3717055                4510186
##  4 Arkansas                   1969208                2117881
##  5 California                21132533               24440416
##  6 Colorado                   3192647                3800664
##  7 Connecticut                2429634                2577311
##  8 Delaware                    584817                 681526
##  9 Florida                   11811921               13914216
## 10 Georgia                    5878186                6725041
## # … with 39 more rows
```

```
temp%>%group_by(year)%>%summarize(min(votes), max(votes))
```

```
## # A tibble: 6 x 3
##    year `min(votes)` `max(votes)`
##   <int>        <int>        <int>
## 1  2004       245789     12589367
## 2  2006       196217      8899059
## 3  2008       256035     13743177
## 4  2010       190822     10529134
## 5  2012       250701     13202158
## 6  2014       171153      7513972
```

```
temp%>%group_by(state, year)%>%summarize(1/rate)
```

```
## # A tibble: 273 x 3
## # Groups:   state [49]
##    state    year `1/rate`
##    <chr>   <int>    <dbl>
##  1 Alabama  2004     1.74
##  2 Alabama  2008     1.64
##  3 Alabama  2010     2.31
##  4 Alabama  2014     3.01
##  5 Alaska   2004     1.44
##  6 Alaska   2006     1.95
##  7 Alaska   2008     1.46
##  8 Alaska   2010     1.89
##  9 Alaska   2012     1.70
## 10 Alaska   2014     1.82
## # … with 263 more rows
```

```
temp%>%group_by(state)%>%do(data.frame(t(quantile(.$tuition))))
```

```
## # A tibble: 49 x 6
## # Groups:   state [49]
##    state          X0.  X25.  X50.   X75.  X100.
##    <chr>       <dbl> <dbl> <dbl>  <dbl>  <dbl>
##  1 Alabama     5683. 6277. 7273.  8427.  9496.
##  2 Alaska      4328. 4958. 5417.  5959.  6149.
##  3 Arizona     5138. 5626. 7449.  9810. 10414.
##  4 Arkansas    5772. 6278. 6659.  7190.  7606.
##  5 California  5286. 5476. 7046.  8939.  9361.
##  6 Colorado    4704. 5768. 7016.  8532.  9299.
##  7 Connecticut 7984. 8368. 9827. 10037. 10664.
##  8 Delaware    8353. 8682. 8995. 10534. 11515.
##  9 Florida     3848. 3953. 4830.  6136.  6495.
## 10 Georgia     4298. 4646. 5630.  7497.  8063.
## # … with 39 more rows
```

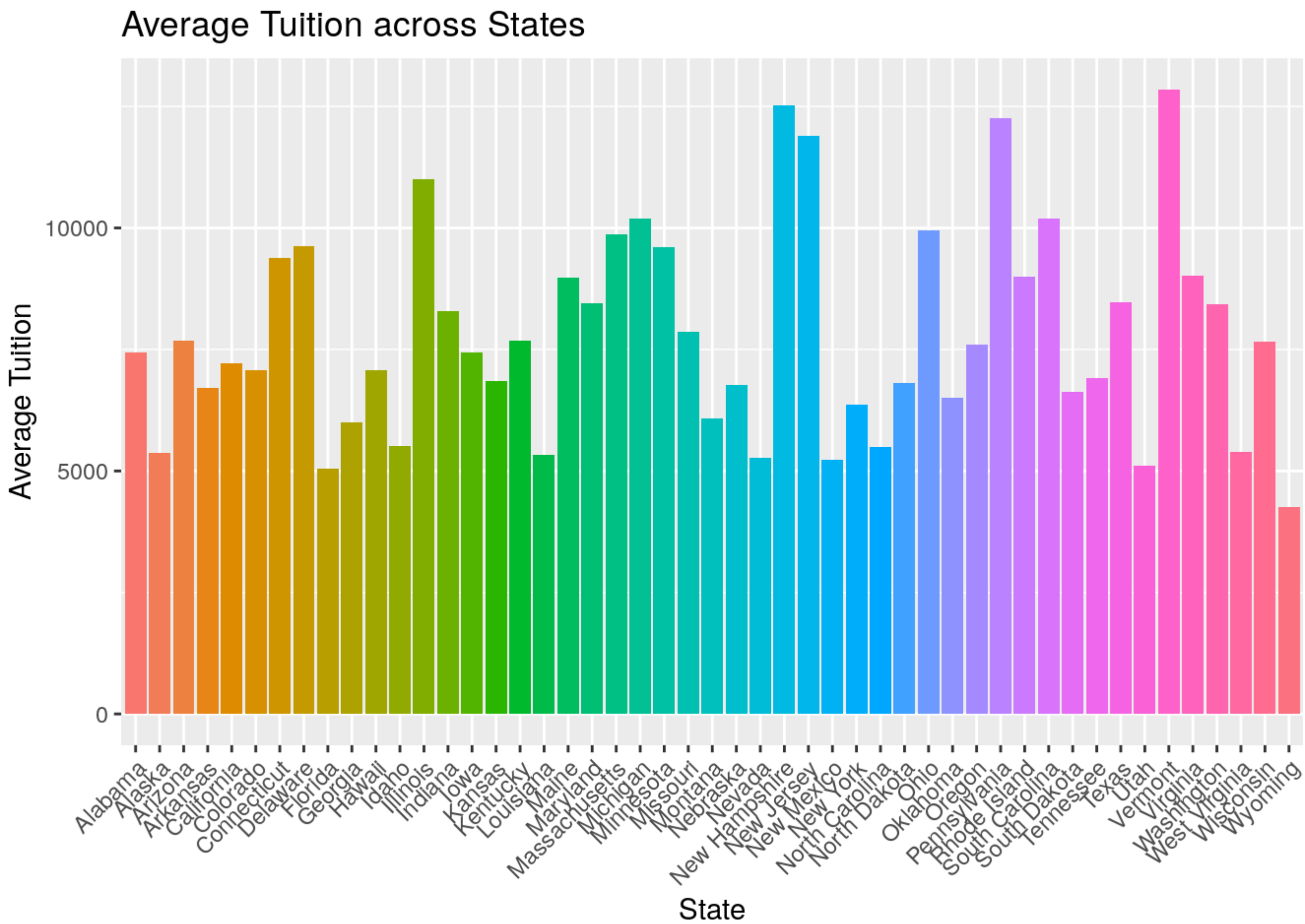
```
# filter()
temp%>%filter(state=="California")%>%select(rate, tuition)%>%cor()
```

```
##               rate    tuition
## rate     1.0000000 -0.4082598
## tuition -0.4082598  1.0000000
```

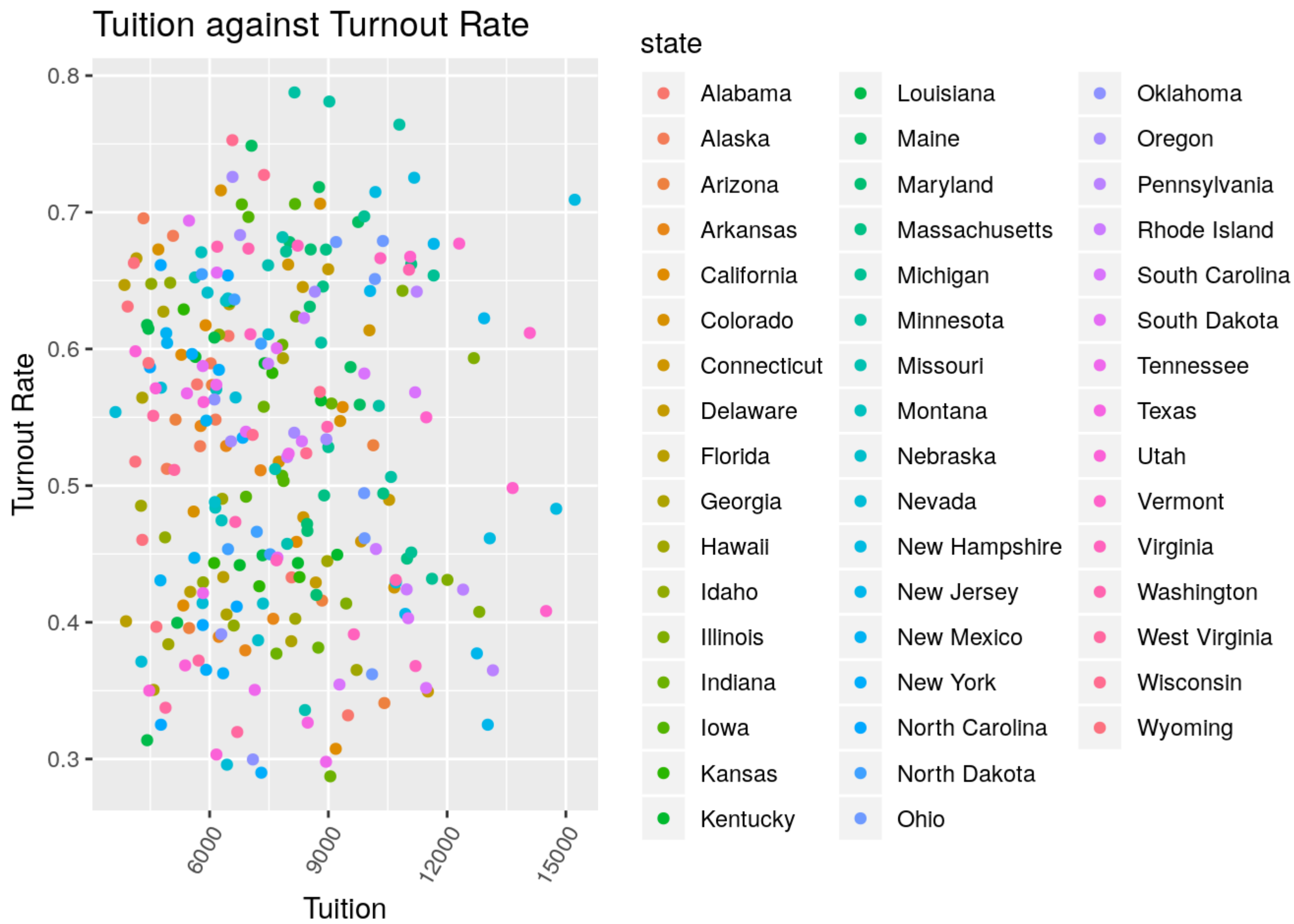
4. Visualizing

The first plot demonstrates the average tuition across different states, and the mean of tuition from differnt year was calculated within the fun.y function. The second plot demonstrate the relationship between tuition and turnout rate across differnt state, but there is no clear linear correlation among them. The potential effect was measured during the dimentionality reduction section.

```
ggplot(temp, aes(state))+
  geom_bar(aes(y=tuition,fill=state), stat="summary", fun.y="mean")+
  theme(axis.text.x = element_text(angle=45, hjust=1), legend.position="none")+
  labs(title = "Average Tuition across States", x = "State", y = "Average Tuition")
```



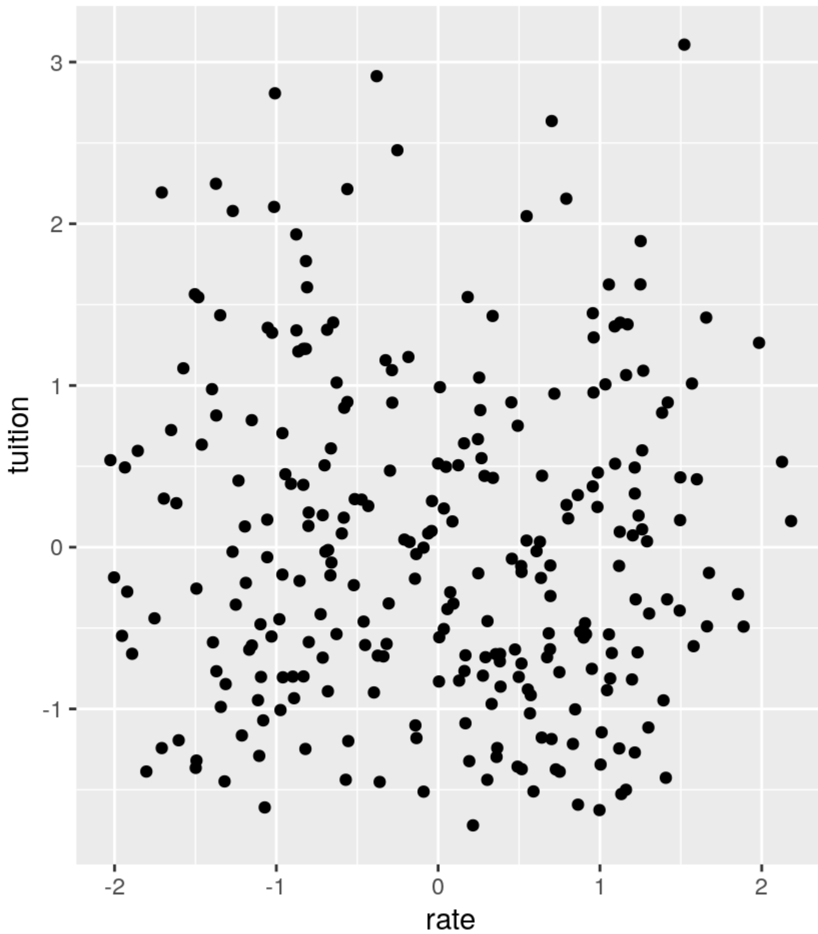

```
ggplot(temp, aes(tuition, rate, color=state))+
  geom_point()+
  theme(axis.text.x=element_text(angle=60, hjust=1))+
  labs(title = "Tuition against Turnout Rate", x = "Tuition", y = "Turnout Rate")
```



5. Dimensionality Reduction

From the result of PCA and correpounding plot, we can see that the votes variable and the eligible voters variable are strongly correlated, but the other two variables do not address much variation on other variables.

```
temp2 <- temp%>%select(-year, -state)
temp2_scaled = data.frame(scale(temp2))
ggplot(temp2_scaled, aes(x = rate, y = tuition))+geom_point()+coord_fixed()
```

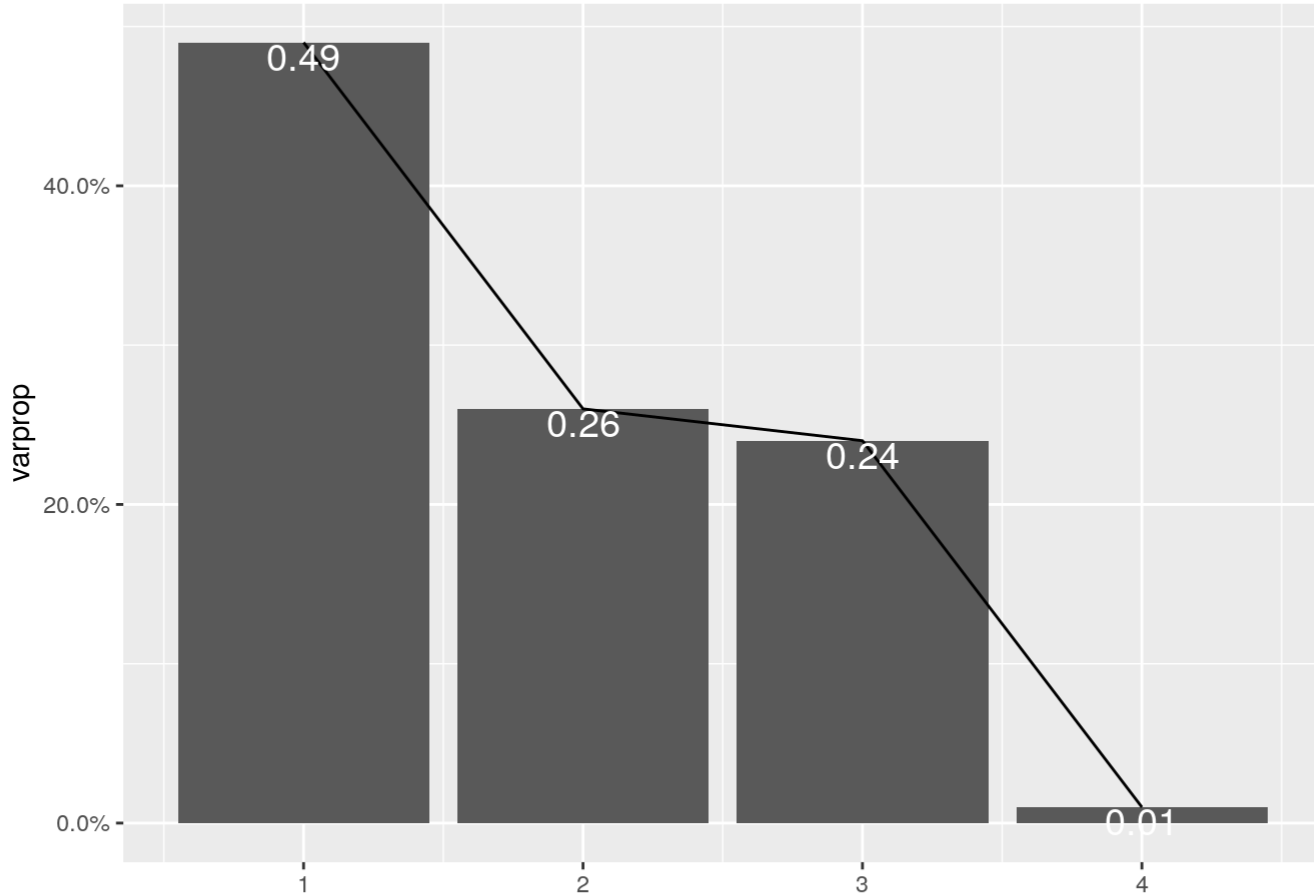



```
temp_pca<-princomp(temp2_scaled)
summary(temp_pca, loadings = T)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3      Comp.4
## Standard deviation     1.3954926 1.0267524 0.9786094 0.161405002
## Proportion of Variance 0.4886398 0.2645241 0.2402993 0.006536838
## Cumulative Proportion  0.4886398 0.7531639 0.9934632 1.000000000
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4
## tuition         -0.106 -0.535  0.838
## votes           -0.703  0.141        -0.697
## eligible_voters -0.703        -0.129  0.696
## rate                   0.830  0.530  0.173
```

```
eigval<-temp_pca$sdev^2
varprop=round(eigval/sum(eigval),2)

ggplot()+geom_bar(aes(y=varprop,x=1:4),stat="identity")+xlab("")+geom_path(aes(y=varprop,x=1:4))
+
 geom_text(aes(x=1:4,y=varprop,label=round(varprop,2)),vjust=1,col="white",size=5)+
 scale_y_continuous(breaks=seq(0,.6,.2),labels = scales::percent)+
 scale_x_continuous(breaks=1:10)
```



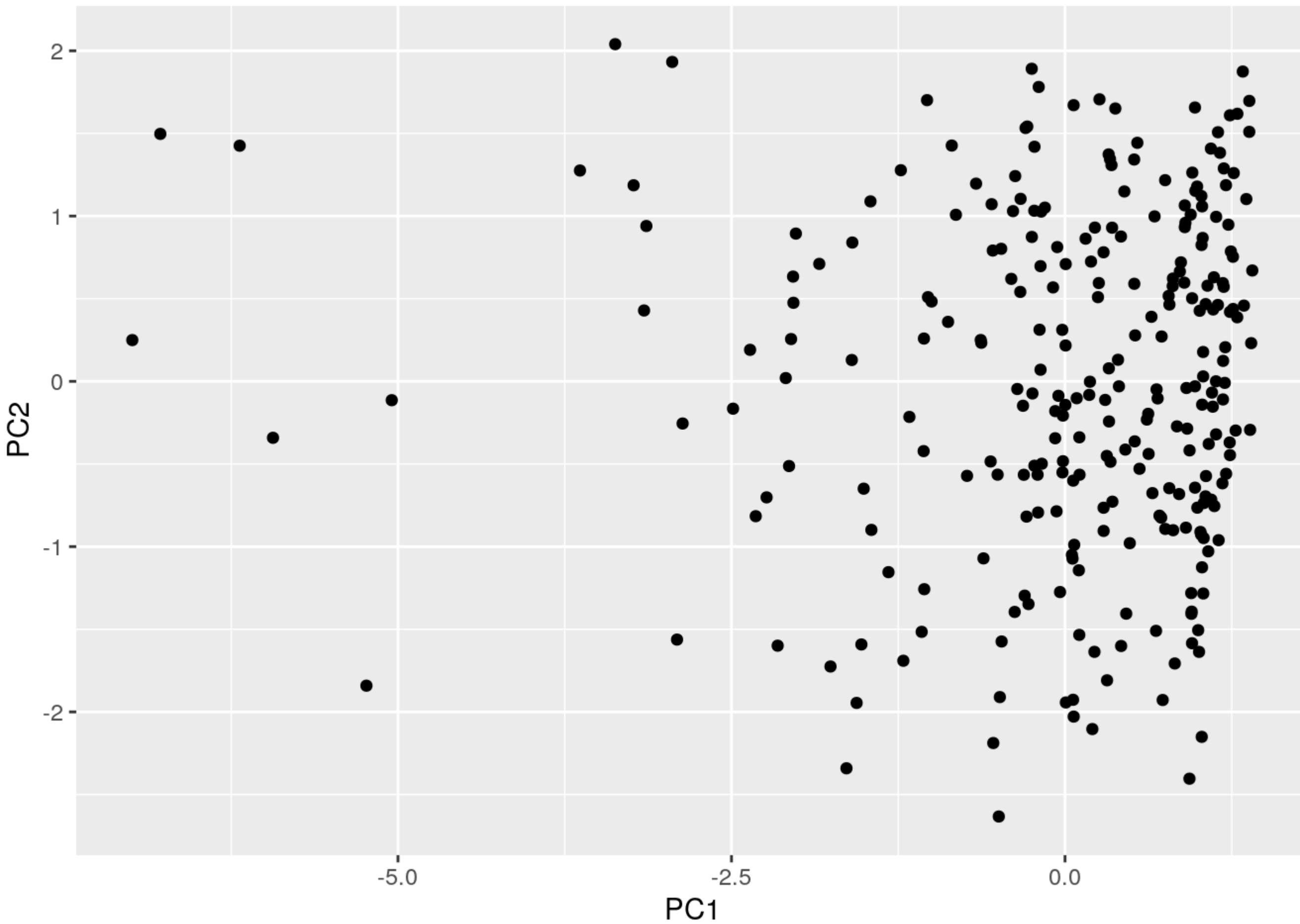

```
round(cumsum(eigval)/sum(eigval),2)
```

```
## Comp.1 Comp.2 Comp.3 Comp.4
##   0.49   0.75   0.99   1.00
```

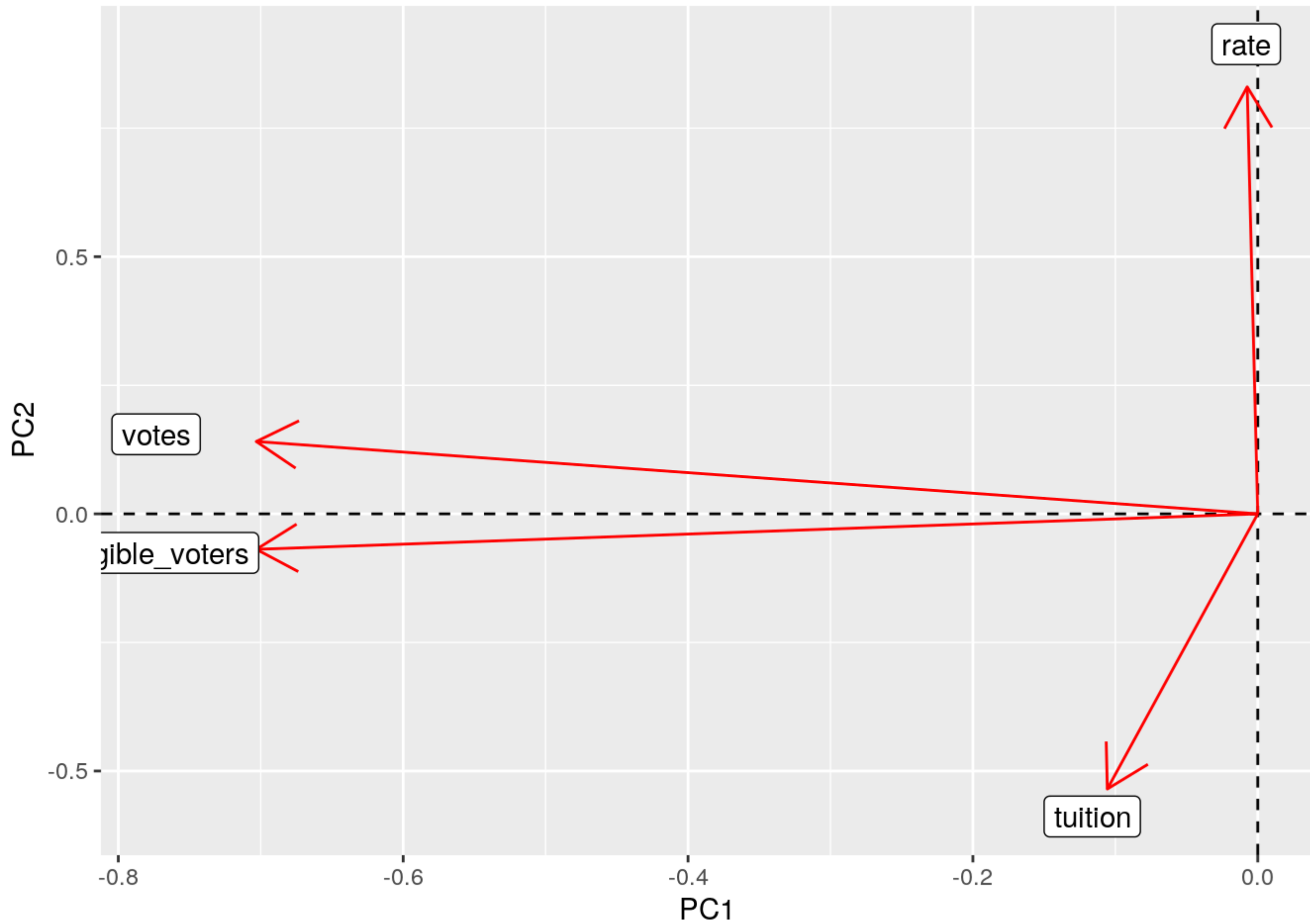
```
eigval
```

```
##     Comp.1     Comp.2     Comp.3     Comp.4
## 1.94739969 1.05422040 0.95767633 0.02605157
```

```
ggplot()+geom_point(aes(temp_pca$scores[,1], temp_pca$scores[,2]))+xlab("PC1")+ylab("PC2")
```




```
temp_pca$loadings[1:4,1:2]%>%as.data.frame%>%rownames_to_column%>%
ggplot()+geom_hline(aes(yintercept=0),lty=2)+
 geom_vline(aes(xintercept=0),lty=2)+ylab("PC2")+xlab("PC1")+
 geom_segment(aes(x=0,y=0,xend=Comp.1,yend=Comp.2),arrow=arrow(),col="red")+
 geom_label(aes(x=Comp.1*1.1,y=Comp.2*1.1,label=rowname))
```


rate
votes
gible_voters
tuition
0.5
0.0
-0.5
PC2
-0.8
-0.6
-0.4
-0.2
0.0
PC1