INDE546_HW8

Michael Shieh

3/6/2020

Principle Component Analysis (PCA)

The six variable I chose was “Age” (from question 43: How old are you), “CommuteTime” (from question 2: On average, how many minutes does it take you to get to UW from your home), “UberDollar” (from question 12: In the past 7 days, how much did you spend on Uber, Lyft, or other ride-hailing apps in US Dollars), “QualityEd” (from question 37: Please indicate how much you agree or disagree with… I think the UW provides a quality education), “SortWaste” (from question 40: How often do you properly sort waste into trash, recyclables and compost), and Downtown (from question 5: How often do you go to downtown Seattle)

```
survey <- survey %>%
  rename(Age = How.old.are.you., CommuteTime = On.average..how.many.minutes.does.it.t
ake.you.to.get.to.the.U..Washington.from.your.home.,
         UberDollar = In.the.past.7.days..how.much.did.you.spend.on.Uber.Lyft.or.othe
r.ride.hailing.apps..in.US.dollars..,
         QualityEd = Please.indicate.how.much.you.agree.or.disagree.with.the.followin
g.statements...I.think.the.UW.provides.a.quality.education.,
         SortWaste = How.often.....do.you.properly.sort.waste.into.trash..recyclables
.and.compost.,
         Downtown = How.often.do.you...go.to.downtown.Seattle..) %>%
  mutate(UberDollar = replace_na(UberDollar, 0)) %>%
  filter(Age >= 0, CommuteTime >= 0)
```

First, I try to see the correlation between each variables. Since the range of each variable varies a lot, I standardize the data first before proceeding.

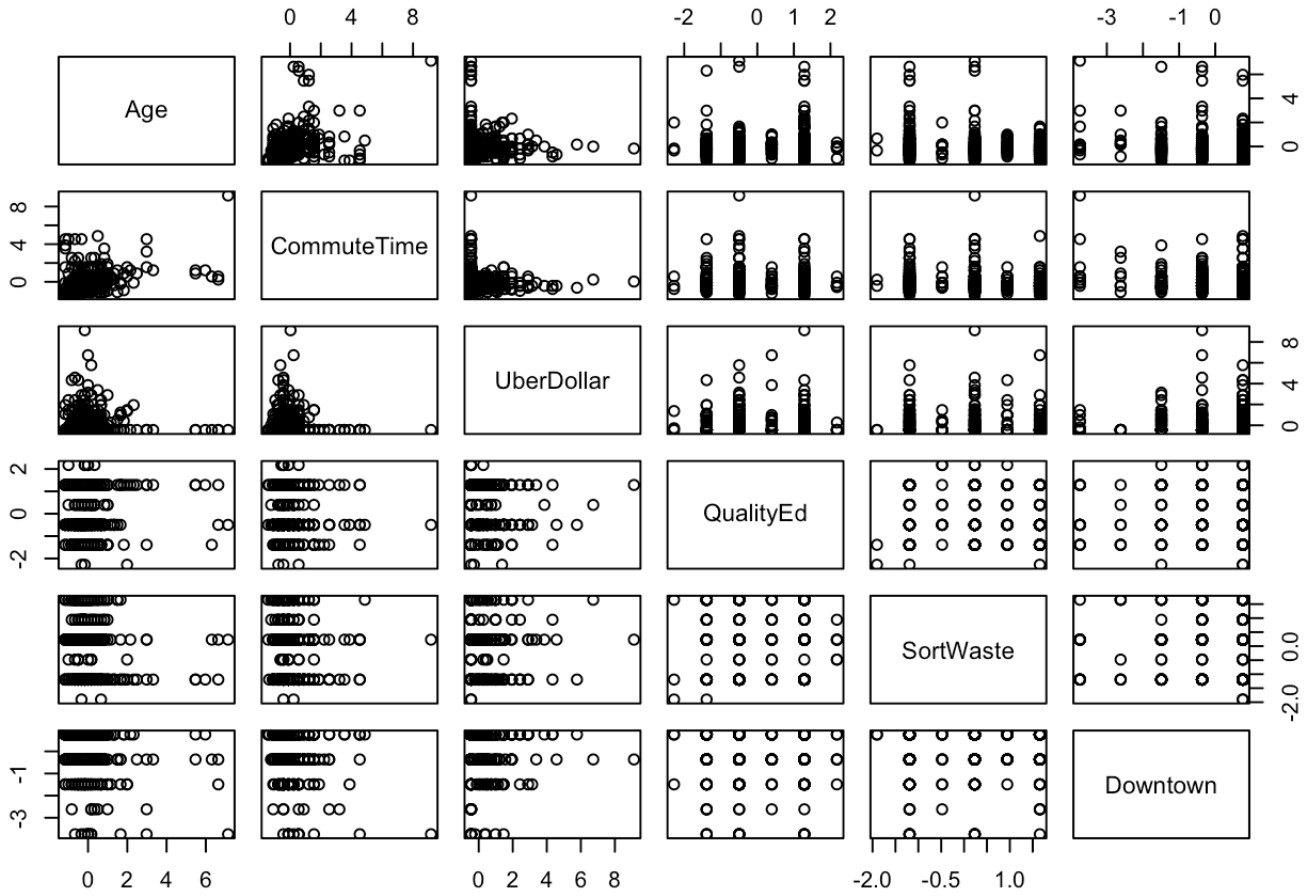
```
survey <- survey %>% select(Age, CommuteTime, UberDollar, QualityEd, SortWaste, Downt
own)
survey$QualityEd <- as.numeric(survey$QualityEd)
survey$SortWaste <- as.numeric(survey$SortWaste)
survey$Downtown <- as.numeric(survey$Downtown)

survey <- survey %>%
  mutate(Age = (Age - mean(Age))/sd(Age), CommuteTime = (CommuteTime - mean(CommuteTi
me))/sd(CommuteTime),
         QualityEd = (QualityEd - mean(QualityEd))/sd(QualityEd), SortWaste = (SortWa
ste - mean(SortWaste))/sd(SortWaste),
         UberDollar = (UberDollar - mean(UberDollar))/sd(UberDollar), Downtown = (Dow
ntown - mean(Downtown))/sd(Downtown))
print(xtable::xtable(cor(survey)), type = "html", html.table.attribute = "border = 2"
)
```

	Age	CommuteTime	UberDollar	QualityEd	SortWaste	Downtown
Age	1.00	0.35	-0.02	0.07	-0.11	-0.17
CommuteTime	0.35	1.00	-0.09	-0.07	-0.09	-0.16
UberDollar	-0.02	-0.09	1.00	0.04	0.01	0.01
QualityEd	0.07	-0.07	0.04	1.00	-0.07	0.06
SortWaste	-0.11	-0.09	0.01	-0.07	1.00	0.06
Downtown	-0.17	-0.16	0.01	0.06	0.06	1.00

As shown aboive, most correlation between variables are small. Next, I plot

```
plot(survey)
```




```
survey_pca <-prcomp(survey, scale = TRUE)
print(survey_pca$sdev)
```

[1] 1.2339704 1.0465939 0.9968165 0.9578558 0.9228632 0.7868615

```
print(xtable::xtable(survey_pca$rotation), type = "html", html.table.attribute = "bor
der = 2")
```

	PC1	PC2	PC3	PC4	PC5	PC6
Age	0.60	-0.16	-0.06	0.34	0.15	-0.69
CommuteTime	0.60	0.14	0.03	0.21	0.34	0.67
UberDollar	-0.13	-0.37	-0.86	-0.03	0.30	0.10
QualityEd	-0.02	-0.77	0.21	0.38	-0.38	0.25
SortWaste	-0.28	0.44	-0.24	0.79	-0.21	0.02
Downtown	-0.42	-0.16	0.38	0.26	0.76	-0.08

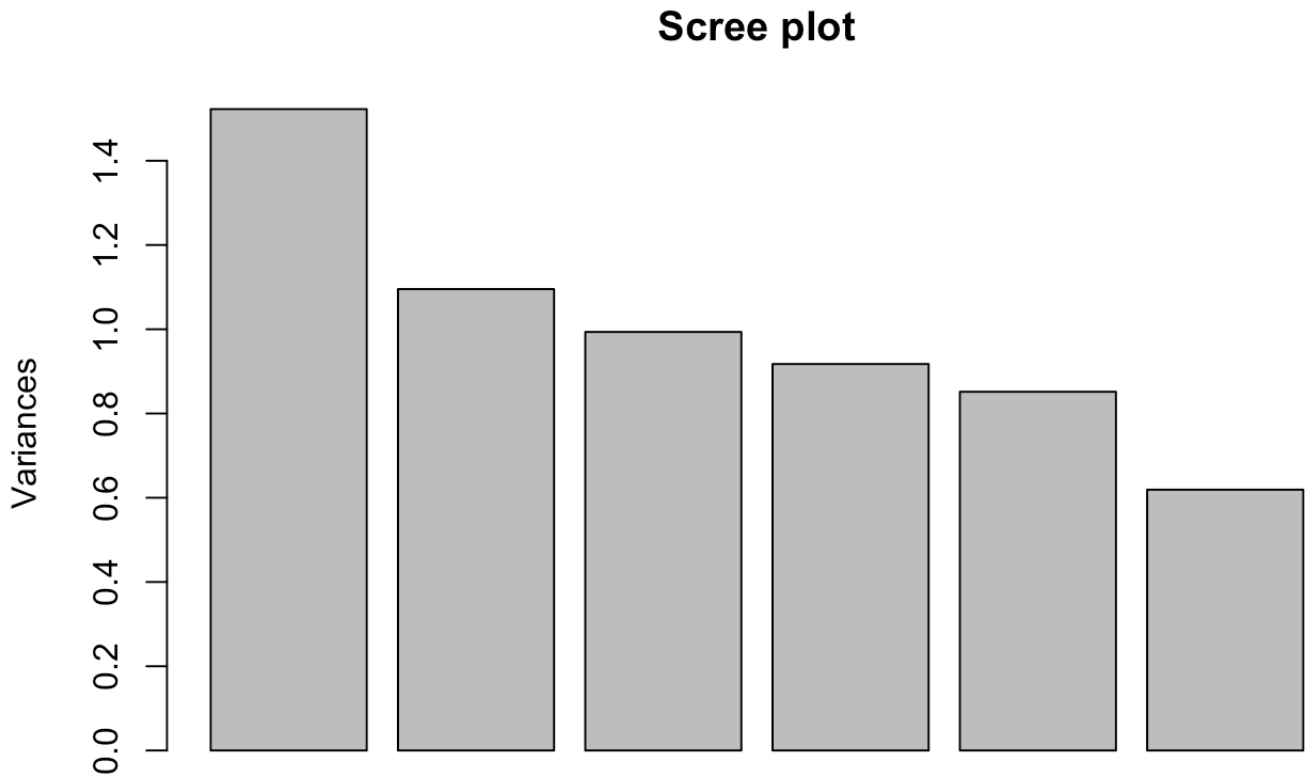
```
print(xtable::xtable(summary(survey_pca)), type = "html", html.table.attribute = "bor
der = 2")
```

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.2340	1.0466	0.9968	0.9579	0.9229	0.7869
Proportion of Variance	0.2538	0.1826	0.1656	0.1529	0.1419	0.1032
Cumulative Proportion	0.2538	0.4363	0.6019	0.7549	0.8968	1.0000

It seems that PC1 is highly correlated with the variables "Age" and "CommuteTime", where the PC1 increases with the increase of age and commute time.Meanwhile, PC2 only highly correlated with the variable "QualityEd". I'd say this PC is the primary measure of the quality of UW education, where the score of PC2 decreases with the increasing of quality education score. Similarly, PC3 can be regarded as the measure of spendings on Uber/Lyft, PC4 as the measure of the frequency of properly sorting waste into appropriate bins, and PC5 as the frequency of going to downtown Seattle. Lastly, PC6 is highly correlated with age (positive) and commute time (negative), but the impact of the two variables are opposite.

I would choose to keep the first 5 principle components because I think that no principle components explains a significant amount of variance. Keeping more of them should be a sensible thing to do considering all of them explained a similar proportion of variance. Keeping the first 5 should explain enough variance. Next, I plot the scree plot to see if my decision would fit the elbow rule.

```
plot(survey_pca, main = "Scree plot")
```



From the scree plot there is no clear sign of an elbow existed visually, unless I only keep the first principle components. However, by doing so I could only explain a small amount of variance, which is unacceptable.

Factor Analysis

Setup for Factor Analysis, and initially select to use 2 factors

```
Age <- cor(survey)[1,]
CommuteTime <- cor(survey)[2,]
UberDollar <- cor(survey)[3,]
QualityEd <- cor(survey)[4,]
SortWaste <- cor(survey)[5,]
Downtown <- cor(survey)[6,]
A_f <- rbind(Age, CommuteTime, UberDollar, QualityEd, SortWaste, Downtown)
```

```
factanal(covmat = A_f, factors = 2, rotation = "none")
```

```
##
## Call:
## factanal(factors = 2, covmat = A_f, rotation = "none")
##
## Uniquenesses:
##         Age CommuteTime  UberDollar   QualityEd   SortWaste    Downtown
##       0.624       0.652       0.992       0.204       0.965       0.917
##
## Loadings:
##             Factor1 Factor2
## Age          0.609
## CommuteTime  0.585
## UberDollar
## QualityEd            0.892
## SortWaste   -0.170
## Downtown    -0.281
##
##                Factor1 Factor2
## SS loadings      0.826   0.819
## Proportion Var   0.138   0.137
## Cumulative Var   0.138   0.274
##
## The degrees of freedom for the model is 4 and the fit was 0.0044
```

Since the variable “UberDollar” did not correlate with any of the factor, I suspected there should be one more factor, so I re-run the setup with 3 factors.

```
factanal(covmat = A_f, factors = 3, rotation = "none")
```

```
##
## Call:
## factanal(factors = 3, covmat = A_f, rotation = "none")
##
## Uniquenesses:
##         Age CommuteTime  UberDollar   QualityEd   SortWaste    Downtown
##       0.600       0.663       0.716       0.005       0.967       0.916
##
## Loadings:
##             Factor1 Factor2 Factor3
## Age                  0.622
## CommuteTime          0.576
## UberDollar          -0.110   0.520
## QualityEd    0.997
## SortWaste           -0.168
## Downtown            -0.279
##
##                Factor1 Factor2 Factor3
## SS loadings      1.014   0.836   0.284
## Proportion Var   0.169   0.139   0.047
## Cumulative Var   0.169   0.308   0.356
##
## The degrees of freedom for the model is 0 and the fit was 5e-04
```

It seems Factor1 is very highly correlated with Quality of Education, and Factor 3 is correlated with Uber Spendings. Also, Factor 2 is highly correlated with Age and Commute Time.

Next, I would use oblique rotation to rotate the data to provide more flexibility. I choose oblique rotation because the data is already uncorrelated, so there's no need to eliminate collinearity.

```
fa3 <- factanal(covmat = A_f, factors = 3, rotation = "promax")
fa3
```

```
##
## Call:
## factanal(factors = 3, covmat = A_f, rotation = "promax")
##
## Uniquenesses:
##         Age CommuteTime  UberDollar   QualityEd   SortWaste    Downtown
##       0.600       0.663       0.716       0.005       0.967       0.916
##
## Loadings:
##             Factor1 Factor2 Factor3
## Age                  0.635
## CommuteTime          0.559
## UberDollar                   0.535
## QualityEd    0.998
## SortWaste           -0.169
## Downtown            -0.286
##
##                Factor1 Factor2 Factor3
## SS loadings      1.013   0.827   0.296
## Proportion Var   0.169   0.138   0.049
## Cumulative Var   0.169   0.307   0.356
##
## Factor Correlations:
##         Factor1 Factor2 Factor3
## Factor1  1.0000  0.0176  -0.100
## Factor2  0.0176  1.0000  -0.153
## Factor3 -0.1005 -0.1531   1.000
##
## The degrees of freedom for the model is 0 and the fit was 5e-04
```

The factors I got from factor analysis are actually very similar to some of the principle components I got earlier. Factor 1 and PC2 both only highly correlated with quality of UW education, Factor 3 and PC3 represents the spendings on Uber/Lyft, and Factor 2 and PC1 are both similarly highly correlated with age and commute time. Even after I allowed factors to be correlated, all three factors end up approximately uncorrelated. I think if the variables are uncorrelated in the beginning, it doesn't matter which type of analysis I use, I would obtain somewhat similar results.