Analysis 2

Tina Hajinejad

2023-03-21

Required Packages

```
knitr::opts_chunk$set(echo = TRUE)
library(opendatatoronto)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(leaps)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

Question 1

First we get all data sets from Open Data Toronto (opt).

```
#First we get all data sets from Open Data Toronto (opt)


pack_odt <- list_packages()

#Looking for the index Apartment Building Evaluations data set
ind <- which(str_detect(pack_odt$title, 'Apartment Building'))

#Grabbing the id
ABD_id = pack_odt$id[ind]

#OR using search_packages!

ABD_2 <- search_packages("Apartment Building Evaluation")

#Checking if the two ways are correct

ABD_id
```

```
## [1] "4ef82789-e038-44ef-a478-a8f3590c3eb1"
```

```
ABD_2$id
```

```
## [1] "4ef82789-e038-44ef-a478-a8f3590c3eb1"
```

Yes, they are.

Now, there are packages with different types and different dates. We will save the tibbles including these datasets with their id's in ABE_resources.

```
ABE_resources <- ABD_2 %>% list_package_resources()
ABE_resources
```

```
## # A tibble: 4 x 4
##   name                                id                          format last_mod~1
##   <chr>                               <chr>                       <chr>  <date>
## 1 Apartment Building Evaluation       b987be09-0c62-4d7d-928c-~ CSV    2023-03-27
## 2 Apartment Building Evaluation.csv   979fb513-5186-41e9-bb23-~ CSV    2023-03-27
## 3 Apartment Building Evaluation.xml   c86721c4-35e3-44a5-9d64-~ XML    2023-03-27
## 4 Apartment Building Evaluation.json  e5b035b7-91aa-4040-a544-~ JSON   2023-03-27
## # ... with abbreviated variable name 1: last_modified
```

We want the latest csv file, which is the second one in the list:

```
ABE_statistics <- ABE_resources[2,]  %>% get_resource()

write.csv(ABE_statistics, "~/Desktop/UW 2/Data Analysis - Stat 874/A2 Analysis/Apartment Building Evalua

as_tibble(ABE_statistics)
```

```
## # A tibble: 11,753 x 40
##       X_id     RSN YEAR_~1 YEAR_~2 YEAR_~3 PROPE~4  WARD WARDN~5 SITE_~6 CONFI~7
##      <int>   <int>   <dbl>   <dbl>   <dbl> <chr>   <int> <chr>   <chr>     <int>
##  1 2968408 5157421    2023      NA    1973 TCHC       17 Don Va~ 6 TREE~       4
##  2 2968409 5156815    2023      NA    1973 TCHC       17 Don Va~ 15 FIE~       4
```

```
##  3 2968410 5156814    2023      NA    1973 TCHC       17 Don Va~ 13 FIE~       4
##  4 2968411 5157387    2023      NA    1973 TCHC       17 Don Va~ 4 TREE~       4
##  5 2968412 5156871    2023      NA    1973 TCHC       17 Don Va~ 2 TREE~       4
##  6 2968413 5157423      NA      NA    1973 TCHC       17 Don Va~ 8 TREE~       4
##  7 2968414 5186997      NA      NA    2019 PRIVATE    12 Toront~ 200 MA~       6
##  8 2968415 5156142    2023      NA    1889 PRIVATE    13 Toront~ 109 PE~       4
##  9 2968416 5118732    2022      NA    2021 PRIVATE    13 Toront~ 25 NIC~      29
## 10 2968417 5156008    2022      NA    1885 PRIVATE     9 Davenp~ 267 BR~       3
## # ... with 11,743 more rows, 30 more variables: CONFIRMED_UNITS <int>,
## #   EVALUATION_COMPLETED_ON <chr>, SCORE <int>, RESULTS_OF_SCORE <chr>,
## #   NO_OF_AREAS_EVALUATED <int>, ENTRANCE_LOBBY <dbl>,
## #   ENTRANCE_DOORS_WINDOWS <dbl>, SECURITY <dbl>, STAIRWELLS <dbl>,
## #   LAUNDRY_ROOMS <dbl>, INTERNAL_GUARDS_HANDRAILS <dbl>,
## #   GARBAGE_CHUTE_ROOMS <dbl>, GARBAGE_BIN_STORAGE_AREA <dbl>, ELEVATORS <dbl>,
## #   STORAGE_AREAS_LOCKERS <dbl>, INTERIOR_WALL_CEILING_FLOOR <dbl>, ...
```

```
#skim(ABE_statistics)
```

Checking with prof's:

```
#his <- read.csv("Apartment Evaluations 2023.csv")

#skim(his)
```

Question 2

By inspection, we find out that the relevant variables are in columns 13 to 33, but there are extra variables that need to be removed.

```
cleaned <- ABE_statistics[,13:33]

ABE_clean <- cleaned %>% subset(select =-c(RESULTS_OF_SCORE,
                                           NO_OF_AREAS_EVALUATED,
                                           LAUNDRY_ROOMS,GARBAGE_CHUTE_ROOMS
                                           ,ELEVATORS,BALCONY_GUARDS
                                           ,STORAGE_AREAS_LOCKERS))
skim(ABE_clean)
```

Table 1: Data summary

Name	ABE_clean
Number of rows	11753
Number of columns	14
Column type frequency:	
numeric	14
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
SCORE	0	1	73.82	10.57	0	66	74	81	100	
ENTRANCE_LOBBY	2	1	3.71	0.77	1	3	4	4	5	
ENTRANCE_DOORS_WINDOWS	5	1	3.68	0.77	1	3	4	4	5	
SECURITY	6	1	4.13	0.88	1	3	4	5	5	
STAIRWELLS	3	1	3.45	0.79	1	3	3	4	5	
INTERNAL_GUARDS_HANDRAILS	3	1	3.60	0.83	1	3	4	4	5	
GARBAGE_BIN_STORAGE_AREA	1	1	3.61	0.78	1	3	4	4	5	
INTERIOR_WALL_CEILING_FLOOR	2	1	3.49	0.77	1	3	3	4	5	
INTERIOR_LIGHTING_LEVELS	2	1	3.67	0.88	1	3	4	4	5	
GRAFFITI	39	1	4.61	0.76	1	4	5	5	5	
EXTERIOR_CLADDING	9	1	3.55	0.72	1	3	4	4	5	
EXTERIOR_GROUNDS	15	1	3.65	0.75	1	3	4	4	5	
EXTERIOR_WALKWAYS	6	1	3.64	0.74	1	3	4	4	5	
WATER_PEN_EXT_BLDG_ELEMENTS	6	1	3.67	0.74	1	3	4	4	5	

```
names(ABE_clean)
```

```
##  [1] "SCORE"                       "ENTRANCE_LOBBY"
##  [3] "ENTRANCE_DOORS_WINDOWS"      "SECURITY"
##  [5] "STAIRWELLS"                  "INTERNAL_GUARDS_HANDRAILS"
##  [7] "GARBAGE_BIN_STORAGE_AREA"    "INTERIOR_WALL_CEILING_FLOOR"
##  [9] "INTERIOR_LIGHTING_LEVELS"    "GRAFFITI"
## [11] "EXTERIOR_CLADDING"           "EXTERIOR_GROUNDS"
## [13] "EXTERIOR_WALKWAYS"           "WATER_PEN_EXT_BLDG_ELEMENTS"
```

So Now we have everything, and by taking skim(ABE_clean) we see that they are already numeric. Now we have to remove NA's.

```
ABE_no_NA <- na.omit(ABE_clean)

dim(ABE_no_NA)
```

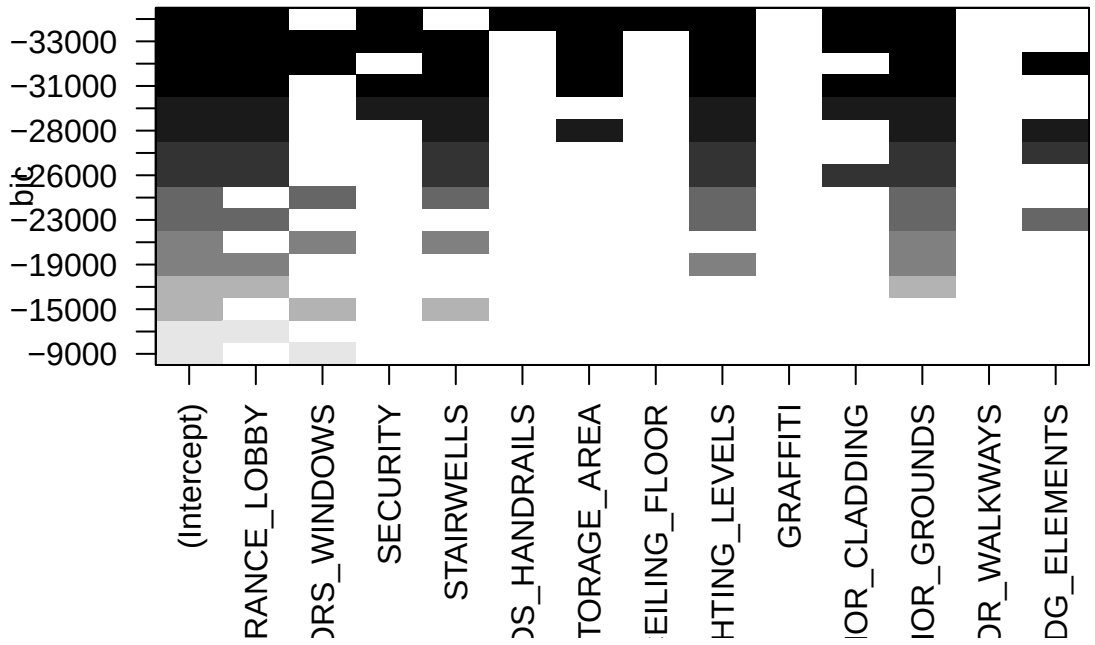
```
## [1] 11676    14
```

```
summary(ABE_no_NA)
```

```
##      SCORE       ENTRANCE_LOBBY  ENTRANCE_DOORS_WINDOWS    SECURITY
##  Min.   : 20.0   Min.   :1.000   Min.   :1.000          Min.   :1.000
##  1st Qu.: 66.0   1st Qu.:3.000   1st Qu.:3.000          1st Qu.:3.000
##  Median : 74.0   Median :4.000   Median :4.000          Median :4.000
##  Mean   : 73.8   Mean   :3.713   Mean   :3.674          Mean   :4.126
##  3rd Qu.: 81.0   3rd Qu.:4.000   3rd Qu.:4.000          3rd Qu.:5.000
##  Max.   :100.0   Max.   :5.000   Max.   :5.000          Max.   :5.000
##    STAIRWELLS    INTERNAL_GUARDS_HANDRAILS GARBAGE_BIN_STORAGE_AREA
##  Min.   :1.000   Min.   :1.000             Min.   :1.000
##  1st Qu.:3.000   1st Qu.:3.000             1st Qu.:3.000
##  Median :3.000   Median :4.000             Median :4.000
##  Mean   :3.451   Mean   :3.602             Mean   :3.604
##  3rd Qu.:4.000   3rd Qu.:4.000             3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000             Max.   :5.000
##  INTERIOR_WALL_CEILING_FLOOR INTERIOR_LIGHTING_LEVELS    GRAFFITI
##  Min.   :1.00                Min.   :1.00             Min.   :1.00
##  1st Qu.:3.00                1st Qu.:3.00             1st Qu.:4.00
##  Median :3.00                Median :4.00             Median :5.00
##  Mean   :3.49                Mean   :3.67             Mean   :4.61
##  3rd Qu.:4.00                3rd Qu.:4.00             3rd Qu.:5.00
##  Max.   :5.00                Max.   :5.00             Max.   :5.00
##  EXTERIOR_CLADDING EXTERIOR_GROUNDS EXTERIOR_WALKWAYS
##  Min.   :1.000     Min.   :1.000    Min.   :1.000
##  1st Qu.:3.000     1st Qu.:3.000    1st Qu.:3.000
##  Median :4.000     Median :4.000    Median :4.000
##  Mean   :3.546     Mean   :3.648    Mean   :3.642
##  3rd Qu.:4.000     3rd Qu.:4.000    3rd Qu.:4.000
##  Max.   :5.000     Max.   :5.000    Max.   :5.000
##  WATER_PEN_EXT_BLDG_ELEMENTS
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :3.668
##  3rd Qu.:4.000
##  Max.   :5.000
```

Question 3

```
regsub_ABE <- regsubsets(SCORE ~ ., data=ABE_no_NA, nbest = 2, really.big = T)
plot(regsub_ABE, scale = "bic")
```




```
# Display the summary of the best models
summary(regsub_ABE)
```

```
## Subset selection object
## Call: regsubsets.formula(SCORE ~ ., data = ABE_no_NA, nbest = 2, really.big = T)
## 13 Variables  (and intercept)
##                               Forced in Forced out
## ENTRANCE_LOBBY                    FALSE      FALSE
## ENTRANCE_DOORS_WINDOWS            FALSE      FALSE
## SECURITY                          FALSE      FALSE
## STAIRWELLS                        FALSE      FALSE
## INTERNAL_GUARDS_HANDRAILS         FALSE      FALSE
## GARBAGE_BIN_STORAGE_AREA          FALSE      FALSE
## INTERIOR_WALL_CEILING_FLOOR       FALSE      FALSE
## INTERIOR_LIGHTING_LEVELS          FALSE      FALSE
## GRAFFITI                          FALSE      FALSE
## EXTERIOR_CLADDING                 FALSE      FALSE
## EXTERIOR_GROUNDS                  FALSE      FALSE
## EXTERIOR_WALKWAYS                 FALSE      FALSE
## WATER_PEN_EXT_BLDG_ELEMENTS       FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          ENTRANCE_LOBBY ENTRANCE_DOORS_WINDOWS SECURITY STAIRWELLS
## 1  ( 1 ) "*"            " "                    " "      " "
## 1  ( 2 ) " "            "*"                    " "      " "
## 2  ( 1 ) "*"            " "                    " "      " "
## 2  ( 2 ) " "            "*"                    " "      "*"
## 3  ( 1 ) " "            "*"                    " "      "*"
## 3  ( 2 ) "*"            " "                    " "      " "
```

```
## 4  ( 1 ) " "            "*"                    " "      "*"
## 4  ( 2 ) "*"            " "                    " "      " "
## 5  ( 1 ) "*"            " "                    " "      "*"
## 5  ( 2 ) "*"            " "                    " "      "*"
## 6  ( 1 ) "*"            " "                    "*"      "*"
## 6  ( 2 ) "*"            " "                    " "      "*"
## 7  ( 1 ) "*"            "*"                    " "      "*"
## 7  ( 2 ) "*"            " "                    "*"      "*"
## 8  ( 1 ) "*"            " "                    "*"      " "
## 8  ( 2 ) "*"            "*"                    "*"      "*"
##          INTERNAL_GUARDS_HANDRAILS GARBAGE_BIN_STORAGE_AREA
## 1  ( 1 ) " "                       " "
## 1  ( 2 ) " "                       " "
## 2  ( 1 ) " "                       " "
## 2  ( 2 ) " "                       " "
## 3  ( 1 ) " "                       " "
## 3  ( 2 ) " "                       " "
## 4  ( 1 ) " "                       " "
## 4  ( 2 ) " "                       " "
## 5  ( 1 ) " "                       " "
## 5  ( 2 ) " "                       " "
## 6  ( 1 ) " "                       " "
## 6  ( 2 ) " "                       "*"
## 7  ( 1 ) " "                       "*"
## 7  ( 2 ) " "                       "*"
## 8  ( 1 ) "*"                       "*"
## 8  ( 2 ) " "                       "*"
##          INTERIOR_WALL_CEILING_FLOOR INTERIOR_LIGHTING_LEVELS GRAFFITI
## 1  ( 1 ) " "                         " "                      " "
## 1  ( 2 ) " "                         " "                      " "
## 2  ( 1 ) " "                         " "                      " "
## 2  ( 2 ) " "                         " "                      " "
## 3  ( 1 ) " "                         " "                      " "
## 3  ( 2 ) " "                         "*"                      " "
## 4  ( 1 ) " "                         "*"                      " "
## 4  ( 2 ) " "                         "*"                      " "
## 5  ( 1 ) " "                         "*"                      " "
## 5  ( 2 ) " "                         "*"                      " "
## 6  ( 1 ) " "                         "*"                      " "
## 6  ( 2 ) " "                         "*"                      " "
## 7  ( 1 ) " "                         "*"                      " "
## 7  ( 2 ) " "                         "*"                      " "
## 8  ( 1 ) "*"                         "*"                      " "
## 8  ( 2 ) " "                         "*"                      " "
##          EXTERIOR_CLADDING EXTERIOR_GROUNDS EXTERIOR_WALKWAYS
## 1  ( 1 ) " "               " "              " "
## 1  ( 2 ) " "               " "              " "
## 2  ( 1 ) " "               "*"              " "
## 2  ( 2 ) " "               " "              " "
## 3  ( 1 ) " "               "*"              " "
## 3  ( 2 ) " "               "*"              " "
## 4  ( 1 ) " "               "*"              " "
## 4  ( 2 ) " "               "*"              " "
## 5  ( 1 ) " "               "*"              " "
```

```
## 5  ( 2 ) "*"               "*"               " "
## 6  ( 1 ) "*"               "*"               " "
## 6  ( 2 ) " "               "*"               " "
## 7  ( 1 ) " "               "*"               " "
## 7  ( 2 ) "*"               "*"               " "
## 8  ( 1 ) "*"               "*"               " "
## 8  ( 2 ) "*"               "*"               " "
##          WATER_PEN_EXT_BLDG_ELEMENTS
## 1  ( 1 ) " "
## 1  ( 2 ) " "
## 2  ( 1 ) " "
## 2  ( 2 ) " "
## 3  ( 1 ) " "
## 3  ( 2 ) " "
## 4  ( 1 ) " "
## 4  ( 2 ) "*"
## 5  ( 1 ) "*"
## 5  ( 2 ) " "
## 6  ( 1 ) " "
## 6  ( 2 ) "*"
## 7  ( 1 ) "*"
## 7  ( 2 ) " "
## 8  ( 1 ) " "
## 8  ( 2 ) " "
```

```
# Select the model with the lowest BIC value
best_model <- which.min(summary(regsub_ABE)$bic)

# Extract the coefficients of the best model
coef(regsub_ABE, id=best_model)
```

```
##                 (Intercept)              ENTRANCE_LOBBY
##                    7.547872                    2.514600
##                    SECURITY   INTERNAL_GUARDS_HANDRAILS
##                    1.795178                    1.951260
##    GARBAGE_BIN_STORAGE_AREA INTERIOR_WALL_CEILING_FLOOR
##                    2.005538                    2.573681
##    INTERIOR_LIGHTING_LEVELS           EXTERIOR_CLADDING
##                    1.850479                    2.704674
##            EXTERIOR_GROUNDS
##                    2.709792
```

```
# Create a linear regression model
model <- lm(SCORE ~ ENTRANCE_LOBBY + SECURITY+ INTERNAL_GUARDS_HANDRAILS+
              GARBAGE_BIN_STORAGE_AREA+INTERIOR_WALL_CEILING_FLOOR+
              INTERIOR_LIGHTING_LEVELS+EXTERIOR_CLADDING
              +EXTERIOR_GROUNDS, data = ABE_no_NA)

# Display the model summary
summary(model)
```

```
##
## Call:
## lm(formula = SCORE ~ ENTRANCE_LOBBY + SECURITY + INTERNAL_GUARDS_HANDRAILS +
##     GARBAGE_BIN_STORAGE_AREA + INTERIOR_WALL_CEILING_FLOOR +
```

```
##     INTERIOR_LIGHTING_LEVELS + EXTERIOR_CLADDING + EXTERIOR_GROUNDS,
##     data = ABE_no_NA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.205  -1.633   0.097   1.719  10.657
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.54787    0.16015   47.13   <2e-16 ***
## ENTRANCE_LOBBY                2.51460    0.04363   57.64   <2e-16 ***
## SECURITY                      1.79518    0.03439   52.20   <2e-16 ***
## INTERNAL_GUARDS_HANDRAILS     1.95126    0.03384   57.66   <2e-16 ***
## GARBAGE_BIN_STORAGE_AREA      2.00554    0.03770   53.20   <2e-16 ***
## INTERIOR_WALL_CEILING_FLOOR   2.57368    0.04041   63.70   <2e-16 ***
## INTERIOR_LIGHTING_LEVELS      1.85048    0.03607   51.31   <2e-16 ***
## EXTERIOR_CLADDING             2.70467    0.04038   66.97   <2e-16 ***
## EXTERIOR_GROUNDS              2.70979    0.04208   64.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.582 on 11667 degrees of freedom
## Multiple R-squared:  0.9398, Adjusted R-squared:  0.9398
## F-statistic: 2.277e+04 on 8 and 11667 DF,  p-value: < 2.2e-16
```

The coefficients derived using lm and using coef(best_model), show the same thing as they should have.

Question 4

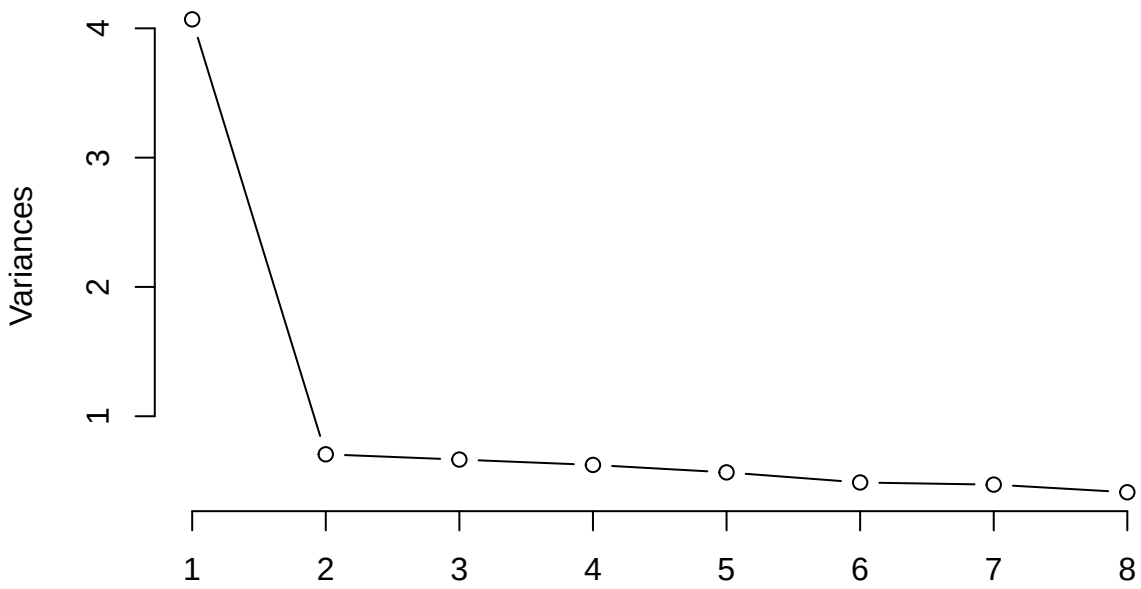
```
PCA = prcomp( ~ ENTRANCE_LOBBY + SECURITY+ INTERNAL_GUARDS_HANDRAILS+
                GARBAGE_BIN_STORAGE_AREA+INTERIOR_WALL_CEILING_FLOOR
              +INTERIOR_LIGHTING_LEVELS+EXTERIOR_CLADDING+
                EXTERIOR_GROUNDS, data = ABE_no_NA,

    scale = TRUE)

PCA
```

```
## Standard deviations (1, .., p=8):
## [1] 2.0173026 0.8399426 0.8153648 0.7896082 0.7522200 0.6982972 0.6864134
## [8] 0.6419254
##
## Rotation (n x k) = (8 x 8):
##                                    PC1         PC2          PC3         PC4
## ENTRANCE_LOBBY              -0.3935988  0.02086506 -0.066371858  0.1871242
## SECURITY                    -0.3485675  0.30167691 -0.447872216 -0.1575014
## INTERNAL_GUARDS_HANDRAILS   -0.3107513  0.62569263  0.644402373 -0.2229822
## GARBAGE_BIN_STORAGE_AREA    -0.3395966 -0.27674311 -0.178405839 -0.7048660
## INTERIOR_WALL_CEILING_FLOOR -0.3596564 -0.04327516 -0.137040753  0.5617006
## INTERIOR_LIGHTING_LEVELS    -0.3704244  0.25619203 -0.217769154  0.2077356
## EXTERIOR_CLADDING           -0.3282915 -0.50457113  0.530751619  0.1536299
## EXTERIOR_GROUNDS            -0.3706025 -0.34406184 -0.004619404 -0.1065691
##                                    PC5         PC6         PC7         PC8
## ENTRANCE_LOBBY              -0.14468185  0.25479736  0.27758297 -0.80144691
## SECURITY                     0.57379357  0.12697988  0.39159238  0.25176969
## INTERNAL_GUARDS_HANDRAILS   -0.16015339  0.07579275  0.01954773  0.12325222
## GARBAGE_BIN_STORAGE_AREA    -0.35047872 -0.38735354  0.07748046 -0.02326703
## INTERIOR_WALL_CEILING_FLOOR -0.49261300 -0.14220870  0.25987913  0.45172867
## INTERIOR_LIGHTING_LEVELS     0.16046562 -0.41225323 -0.69943999 -0.14715885
## EXTERIOR_CLADDING            0.48104482 -0.28296898  0.14383365  0.01302105
## EXTERIOR_GROUNDS            -0.03902251  0.70208235 -0.43101762  0.22951763
```

```
summary(PCA)
```

```
## Importance of components:
##                           PC1     PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     2.0173 0.83994 0.8154 0.78961 0.75222 0.69830 0.6864
## Proportion of Variance 0.5087 0.08819 0.0831 0.07794 0.07073 0.06095 0.0589
## Cumulative Proportion  0.5087 0.59688 0.6800 0.75791 0.82864 0.88960 0.9485
##                           PC8
## Standard deviation     0.64193
## Proportion of Variance 0.05151
## Cumulative Proportion  1.00000
```

```
plot(PCA, type="lines")
```




```
round(head(PCA$x),3)
```

```
##      PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
## 1 -2.396 -0.253  0.180 -0.361  0.740 -2.333  0.416 -0.123
## 2 -1.010  1.212 -1.123 -1.523  0.040 -1.358 -0.325 -0.749
## 3 -1.961  0.051 -0.388 -1.450  0.660 -0.819 -0.697 -0.426
## 4 -2.436 -0.169  0.273 -0.852  1.196 -1.818  0.435 -1.749
## 5 -0.626 -0.130  1.276 -0.746  1.820 -1.611 -0.808 -1.560
## 6 -4.734 -0.442  0.680  0.062 -0.387 -0.232 -0.008  0.189
```

Judging by the plot, PC1 and PC2 will explain the model well. (Based on were the arm is).

Question 5

```
predictors = PCA$x[,1:3]
response = ABE_no_NA$SCORE

model2 <- lm(response ~ PCA$x[,1]+PCA$x[,2]+PCA$x[,3])
summary(model2)
```

```
##
## Call:
## lm(formula = response ~ PCA$x[, 1] + PCA$x[, 2] + PCA$x[, 3])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.7174  -1.6266   0.1067   1.7158  10.0957
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 73.80156    0.02397 3079.027   <2e-16 ***
## PCA$x[, 1]  -5.05217    0.01188 -425.185   <2e-16 ***
## PCA$x[, 2]  -0.25510    0.02854   -8.939   <2e-16 ***
## PCA$x[, 3]   0.32353    0.02940   11.005   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 11672 degrees of freedom
## Multiple R-squared:  0.9394, Adjusted R-squared:  0.9394
## F-statistic: 6.033e+04 on 3 and 11672 DF,  p-value: < 2.2e-16
```

Question 6

Describe briefly two advantages and disadvantages of this PCA-based model over the best subsets model from earlier in this assignment. Advantages:

1. The PCA-based model reduces dimensionality, in the sense that it makes new variables by grouping variables in the same family in a special way, which makes the model simpler and easier to interpret. However, the best subsets model still includes most of the variables, keeping a high dimension which is complex to analyze.
2. As mentioned in the previous point, the PCA-based model groups variables with a similar category, making it easier to understand which component is contributing to the outcome. This issue of multi-colinearity is not dealt with in the best subsets model.

Disadvantages:

1. Although the PCA-base model has its advantages, if we specifically want to know what variables are contributing most to the outcome and in what way, it is generally impossible to find that information from this kind of model. On the other hand, this kind of information is accessible with the best subsets model.
2. The PCA-based model assumes a linear relationship between the variables it's grouping, which may not be true for every case and can sometimes cause inaccuracy, where as the best subsets model can handle non-linear relationship between the variables.

Question 7

```
vif(model)
```

```
##              ENTRANCE_LOBBY                    SECURITY
##                    1.993908                    1.594581
##   INTERNAL_GUARDS_HANDRAILS   GARBAGE_BIN_STORAGE_AREA
##                    1.377786                    1.520492
## INTERIOR_WALL_CEILING_FLOOR   INTERIOR_LIGHTING_LEVELS
##                    1.677618                    1.752209
##           EXTERIOR_CLADDING           EXTERIOR_GROUNDS
##                    1.466414                    1.755484
```

```
vif(model2)
```

```
## PCA$x[, 1] PCA$x[, 2] PCA$x[, 3]
##          1          1          1
```

As explained in the previous question, the PCA-based model has such low inflation factors because what PCA does is create new components out of variables it thinks are similar to each other or fall into the same category to reduce the issue of multicolinearity. This is the reason why we are seeing "1" for all the PCA components. For the best subsets model, we see values larges than one because some of these variables are inherently related to each other.