Midterm

Reese Quillian

2022-10-19

```
library(psych)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(ggfortify)
library(here)
```

```
## here() starts at C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggpubr)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(lindia)
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

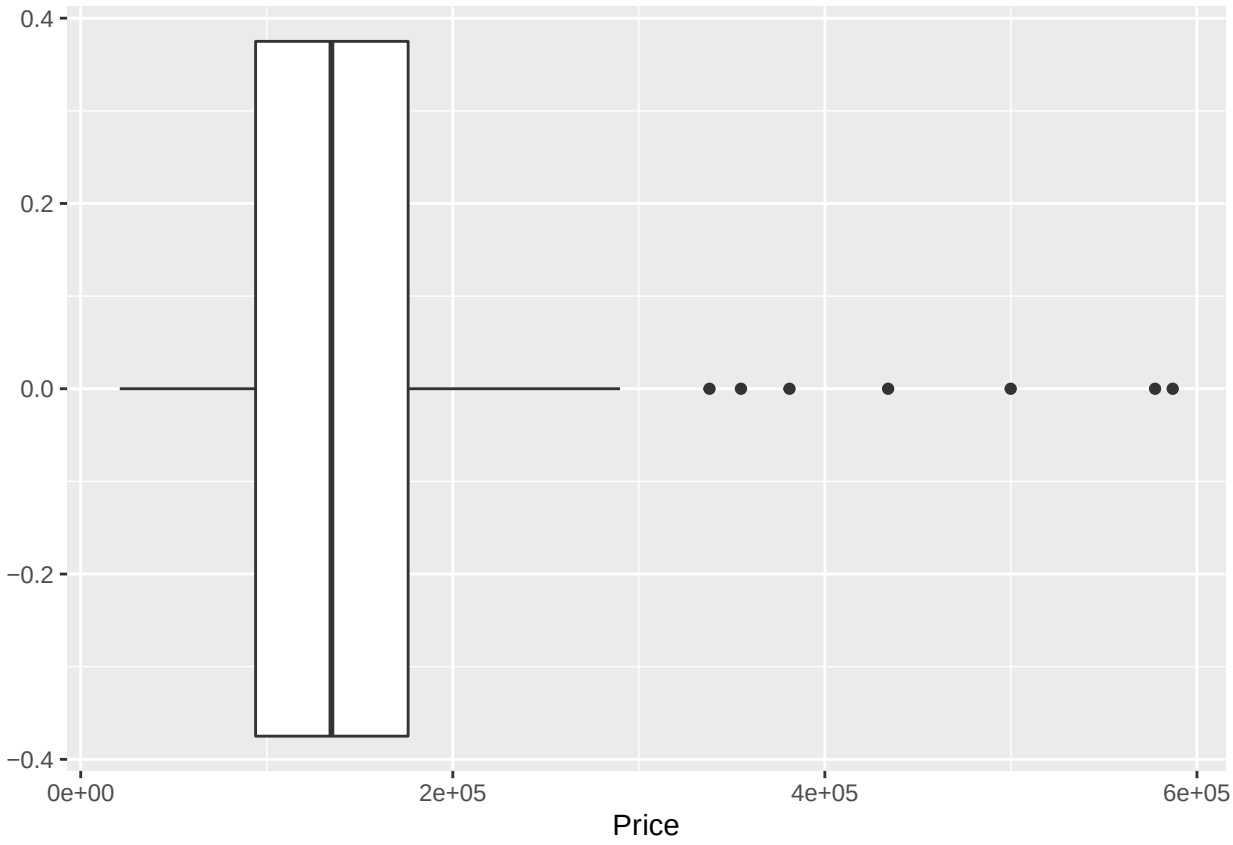
```
setwd("C:/Users/Student/OneDrive - University of Virginia/Documents/SYS4021/")

# read in data
housing_prices <- read.csv("housing-prices.csv")
```

Part 1

Question 1

Use a box plot to determine if the price variable has any outliers. If so how many does it have?

```
ggplot(data=housing_prices, aes(x=Price)) + geom_boxplot()
```

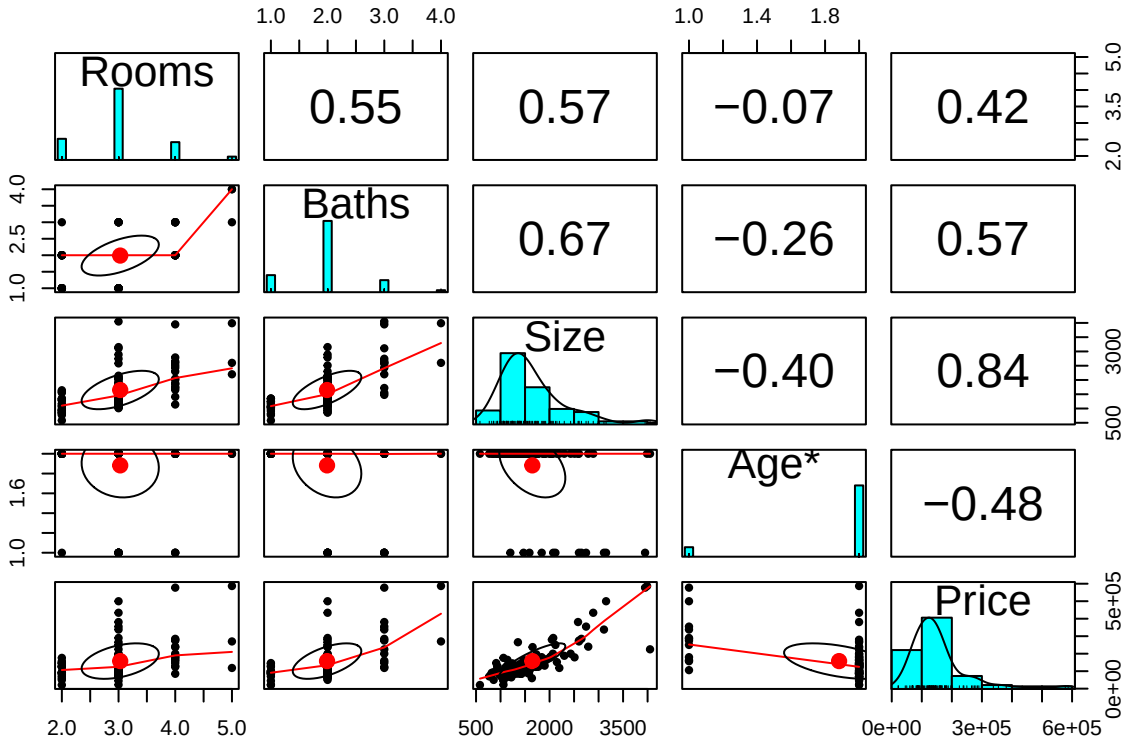



```
# 7 outliers
```

Question 2

Use a scatter plot matrix or individual scatter plots to determine the variable in the data set that has the strongest linear relationship with price. What is it?

```
pairs.panels(housing_prices[,c("Rooms", "Baths", "Size", "Age", "Price")])
```




```
# size
```

Question 3

What is the correlation of the strongest linear relationship to Price?

```
# size and price correlation = 0.84
```

Question 4

Which of the following is a main effects model using Baths as the only predictor variable and y to represent the true value of the observed response, not the modeled prediction? A. log(y)= B0+B1X1 + e where X1 = Baths **B. y= B0+B1X1+e where X1 = Baths** C. y= B0+B1X1 where X1 = Baths D. log(y)= B0+B1X1 where X1 = Baths

```
# B
```

Question 5

Build a linear model, houses.lm1, to predict Price in terms of Baths. Which of the following is true about the relationship between Baths and Price?

```
houses.lm1<-lm(Price~Baths, data=housing_prices)
summary(houses.lm1)
```

```
##
## Call:
## lm(formula = Price ~ Baths, data = housing_prices)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135562  -53562  -15963   27761  341337
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -32035      28729  -1.115    0.267
## Baths          95299      13817   6.897 4.87e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84030 on 100 degrees of freedom
## Multiple R-squared:  0.3224, Adjusted R-squared:  0.3156
## F-statistic: 47.57 on 1 and 100 DF,  p-value: 4.871e-10
```

```
# B1 is positive -> positive relationship
```

Question 6

Use your model in Question 5 to explain the effect on price from having one more bathroom. The predicted selling price of a house decreases/increases by _____ dollars per added bathroom.

```
coef(houses.lm1)
```

```
## (Intercept)       Baths
##   -32035.25    95298.99
```

```
#95298.99
```

Question 7

Which of the following is a main effects model using all predictor variables (Baths, Rooms, Age, and Size) and y to represent the true value of the observed response, not its modeled prediction?

A. y= B0 + B1X1 + B2X2 + B3X3+ B4X4 where X1 = Baths, X2= Rooms, X3 = 1 if Old; 0 otherwise, X4=Size B. y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + e where X1 = Baths, X2 = Rooms, X3 = 1 if Old; 0 otherwise, X4 =Size
C. y= B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X1X2 + B6X1X3 + B7X1X4 + B8X2X3 + B9X2X4 + B10X3X4 where X1 = Baths, X2 = Rooms, X3 = 1 if Old; 0 otherwise, X4 = Size
D. y= B0+B1X1 + B2X2 + B3X3 + B4X4 + B5X1X2 + B6X1X3 + B7X1X4 + B8X2X3 + B9X2X4 + B10X3X4 + e where X1 = Baths, X2= Rooms, X3 = 1 if Old; 0 otherwise, X4 = Size

Question 8

Which of the following is a model that includes the main effect and all interaction terms using ONLY Size and Age as predictors of the response, whose true, observed value is represented by y.

A. y= B0+ B1X1 + B2X2 where X1 = Size, X2 = 1 if Old; 0 otherwise B. y= B0 + B1X1 + B2X2 + e where X1 = Size, X2 = 1 if Old; 0 otherwise
C. y= B0 + B1X1 + B2X2 + B3X1X2 where X1 = Size, X2 = 1 if Old; 0 otherwise D. y= B0 + B1X1 + B2X2 + B3X1X2 + e where X1 = Size X2 = 1 if Old; 0 otherwise

Question 9

How many parameters does your interaction model in Question 8 have? **4**

Coefficients? **3**

Question 10

What is the null hypothesis for the Partial F-test to compare the models you found in Questions 4 and 7?

A. B1=0 B. B1=B2=0 **C. B2=B3=B4=0** D. B4=0

Question 11

Conduct the partial F test to compare the models you found for Questions 4 and 7. What is the exact p-value (not the significance level)?

```
houses.lm2<-lm(Price~Baths+Rooms+Age+Size, data=housing_prices)
summary(houses.lm2)
```

```
##
## Call:
## lm(formula = Price ~ Baths + Rooms + Age + Size, data = housing_prices)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -216746  -31473   -5943   18202  164287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22762.46   31204.42   0.729  0.46748
## Baths         5350.69   12444.17   0.430  0.66817
## Rooms        -8638.90   10163.93  -0.850  0.39744
## AgeOld      -50802.60   18409.25  -2.760  0.00692 **
## Size           118.43      12.08   9.803 3.57e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53850 on 97 degrees of freedom
## Multiple R-squared:  0.7301, Adjusted R-squared:  0.7189
## F-statistic: 65.59 on 4 and 97 DF,  p-value: < 2.2e-16
```

```
anova(houses.lm1, houses.lm2)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Baths
## Model 2: Price ~ Baths + Rooms + Age + Size
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    100 7.0614e+11
## 2     97 2.8128e+11  3 4.2486e+11 48.838 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 12

What do the results of the partial F test comparing the models you found for Questions 4 and 7 suggest with regards to which model you should choose?

A. Reject the null hypothesis and therefore use the larger model with additional terms. B. Fail to reject this hypothesis and therefore use the smaller model without additional terms. C. Reject the null hypothesis and therefore use the smaller model without additional terms. D. Fail to reject this hypothesis and therefore use the larger model with additional terms.

Question 13

Now create a stepwise model from a main effects + interaction model for ALL predictor variables. What is the AIC of the model? Round to the nearest whole number.

```
houses.inter <- lm(Price~(Baths+Rooms+Age+Size)^2,data=housing_prices)
#summary(houses.inter)

#step
houses.step<-step(houses.inter, trace=T)
```

```
## Start:  AIC=2213.5
## Price ~ (Baths + Rooms + Age + Size)^2
##
##               Df  Sum of Sq        RSS    AIC
## - Baths:Size   1 3.3781e+07 2.1859e+11 2211.5
## - Baths:Rooms  1 2.3393e+08 2.1879e+11 2211.6
## - Rooms:Age    1 2.5309e+08 2.1881e+11 2211.6
## <none>                      2.1855e+11 2213.5
## - Baths:Age    1 1.0819e+10 2.2937e+11 2216.4
## - Rooms:Size   1 1.2323e+10 2.3088e+11 2217.1
## - Age:Size     1 3.1939e+10 2.5049e+11 2225.4
##
## Step:  AIC=2211.52
## Price ~ Baths + Rooms + Age + Size + Baths:Rooms + Baths:Age +
##     Rooms:Age + Rooms:Size + Age:Size
##
##               Df  Sum of Sq        RSS    AIC
## - Rooms:Age    1 2.3070e+08 2.1882e+11 2209.6
## - Baths:Rooms  1 3.8711e+08 2.1898e+11 2209.7
```

```
## <none>                      2.1859e+11 2211.5
## - Baths:Age    1 1.2664e+10 2.3125e+11 2215.3
## - Rooms:Size   1 1.3723e+10 2.3231e+11 2215.7
## - Age:Size     1 3.2378e+10 2.5097e+11 2223.6
##
## Step:  AIC=2209.63
## Price ~ Baths + Rooms + Age + Size + Baths:Rooms + Baths:Age +
##     Rooms:Size + Age:Size
##
##               Df  Sum of Sq        RSS    AIC
## - Baths:Rooms  1 3.3832e+08 2.1916e+11 2207.8
## <none>                      2.1882e+11 2209.6
## - Rooms:Size   1 1.3538e+10 2.3236e+11 2213.8
## - Baths:Age    1 1.7792e+10 2.3661e+11 2215.6
## - Age:Size     1 3.2535e+10 2.5135e+11 2221.8
##
## Step:  AIC=2207.78
## Price ~ Baths + Rooms + Age + Size + Baths:Age + Rooms:Size +
##     Age:Size
##
##              Df  Sum of Sq        RSS    AIC
## <none>                     2.1916e+11 2207.8
## - Baths:Age   1 1.7871e+10 2.3703e+11 2213.8
## - Rooms:Size  1 2.6547e+10 2.4570e+11 2217.4
## - Age:Size    1 3.3387e+10 2.5254e+11 2220.2
```

```
summary(houses.step)
```

```
##
## Call:
## lm(formula = Price ~ Baths + Rooms + Age + Size + Baths:Age +
##     Rooms:Size + Age:Size, data = housing_prices)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -145788  -27632   -3353   18162  178857
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.272e+05  9.171e+04   2.478 0.015003 *
## Baths        -1.033e+05  3.993e+04  -2.587 0.011214 *
## Rooms        -5.426e+04  1.741e+04  -3.116 0.002432 **
## AgeOld       -9.262e+04  7.287e+04  -1.271 0.206887
## Size          1.068e+02  3.890e+01   2.746 0.007231 **
## Baths:AgeOld  1.137e+05  4.108e+04   2.769 0.006783 **
## Rooms:Size    2.966e+01  8.789e+00   3.374 0.001076 **
## AgeOld:Size  -1.050e+02  2.775e+01  -3.784 0.000271 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48290 on 94 degrees of freedom
## Multiple R-squared:  0.7897, Adjusted R-squared:  0.774
## F-statistic: 50.42 on 7 and 94 DF,  p-value: < 2.2e-16
```

```
#AIC(houses.step)
```

Question 14

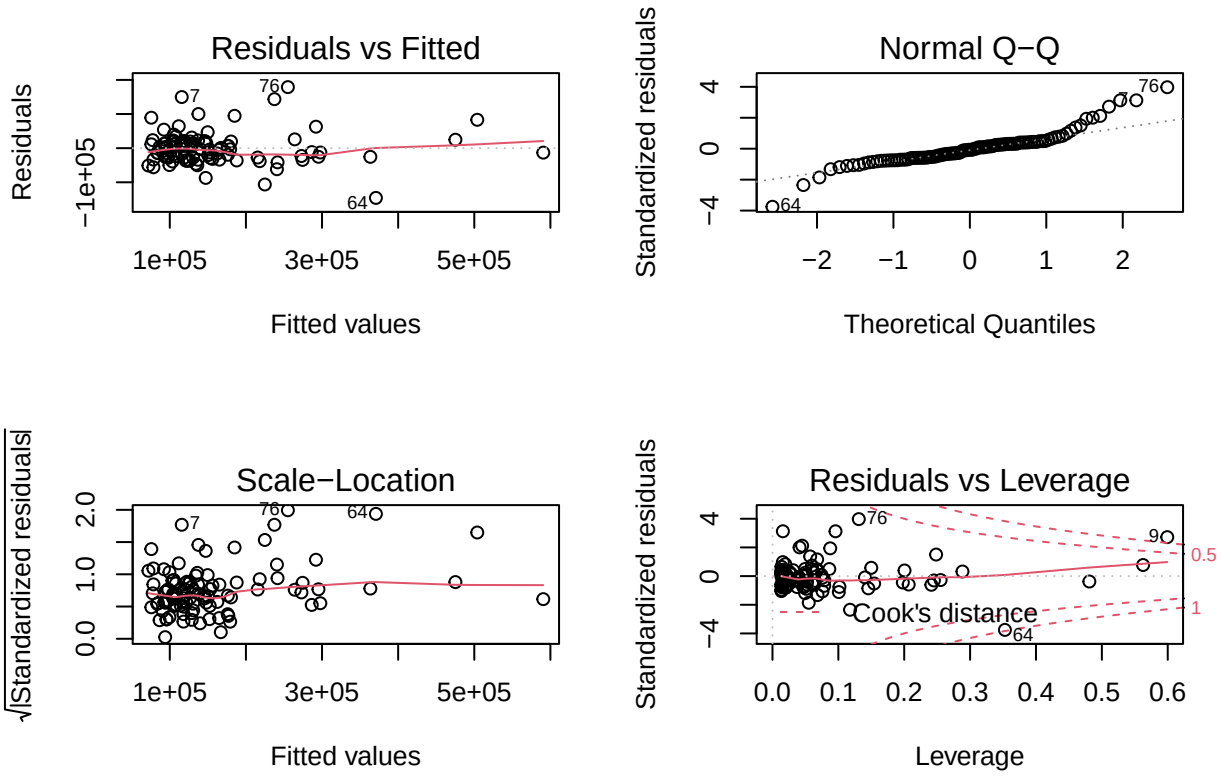
For your stepwise model from Question 13 (a main effects + interaction model for ALL predictor variables). How many parameters are significant at the 0.05 level?

```
# 7 significant paramaeters (includes intercept)
```

Question 15

Model diagnostics

```
par(mfrow = c(2, 2))
plot(houses.step)
```

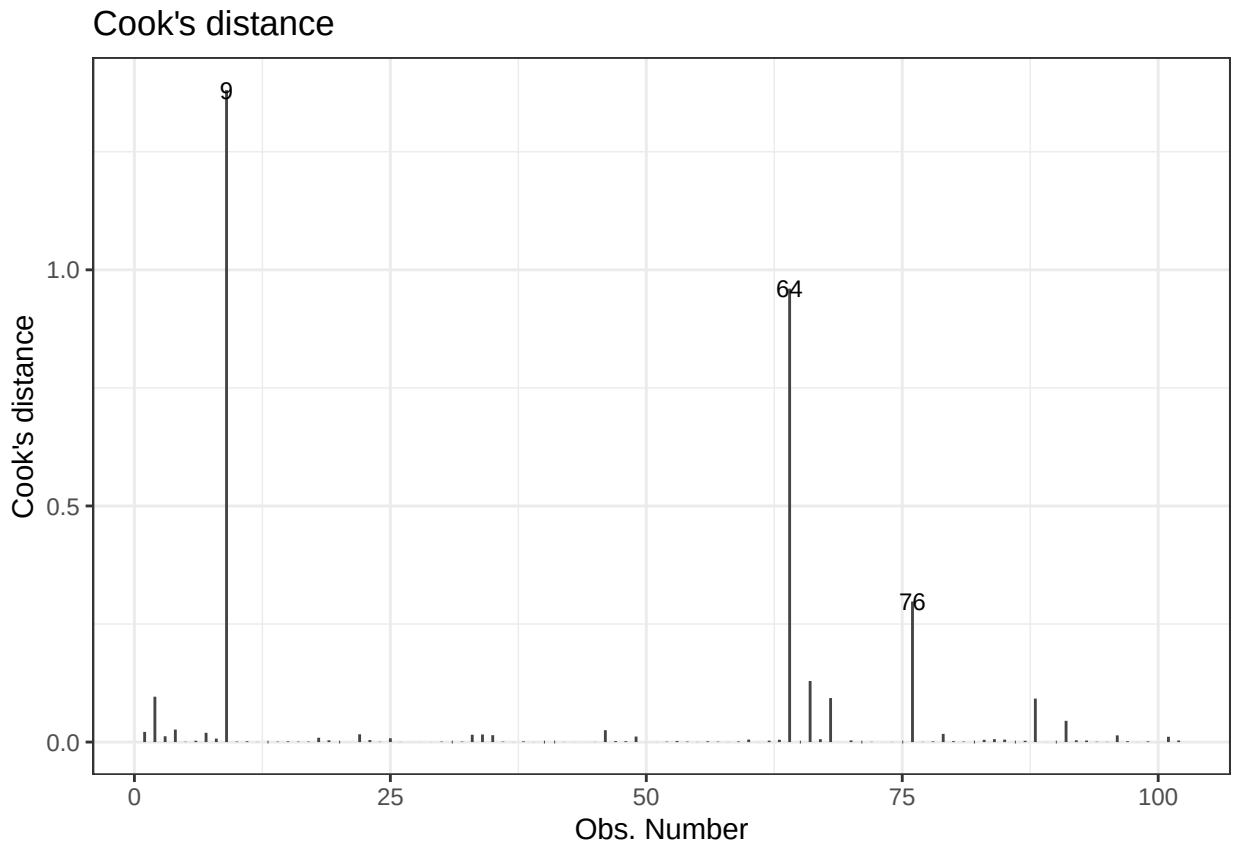



```
# non-constant variance
ols_test_breusch_pagan(houses.step) # p-value <.05 so there is non-constant variance
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
```

```
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##               Data
##  ---------------------------------
##  Response : Price
##  Variables: fitted values of Price
##
##          Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    14.62988
##  Prob > Chi2   =    0.0001308238
```

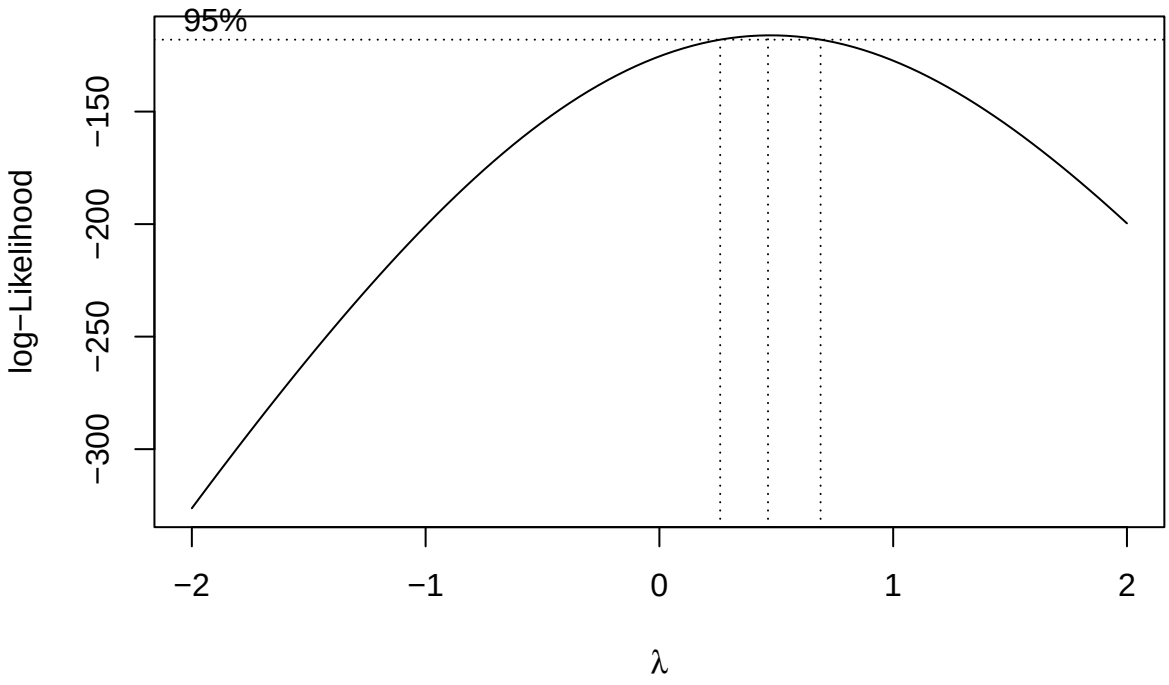
```
autoplot(houses.step, which=4, ncol = 1, label.size = 3) + theme_bw()
```



Question 16

For your stepwise model from Question 13, using Box-Cox plot would you recommend a transformation of the response variable and if so, what type?

```
boxcox(houses.step)
```




```
boxcox(houses.step, plotit = F)$x[which.max(boxcox(houses.step, plotit = F)$y)]
```

```
## [1] 0.5
```

```
# power transformation with L=0.5
```

Question 17

Answer the following questions about the point with the highest Cook's distance.

```
# observation number 9
# other info:
housing_prices[9,]
```

```
##   Rooms Baths Size Age  Price
## 9     5     4 3990 Old 587000
```

Question 18

Which variables could not be included in principal components analysis of this dataset? - anything quantitative (i.e. NOT AGE)

Question 19

Using the correlation matrix, find the principal components for the dataset using all variables that can be included in PCA. Which 2 variables have the largest absolute loadings in the first principal component?

```
houses.corr <- princomp(housing_prices[,c("Price","Rooms","Baths","Size")], cor = T)
houses.corr$loadings
```

```
##
## Loadings:
##       Comp.1 Comp.2 Comp.3 Comp.4
## Price  0.509  0.550  0.252  0.612
## Rooms  0.438 -0.769  0.442  0.145
## Baths  0.494 -0.184 -0.843  0.102
## Size   0.552  0.268  0.172 -0.770
##
##                Comp.1 Comp.2 Comp.3 Comp.4
## SS loadings      1.00   1.00   1.00   1.00
## Proportion Var   0.25   0.25   0.25   0.25
## Cumulative Var   0.25   0.50   0.75   1.00
```

```
# size and price
```

Question 20

In the first PC, Price moves in the same direction as:

rooms, baths, and size because they are all positive

Question 21

Which 2 variables have the largest absolute loadings in the second principal component? rooms and price

Question 22

In the second PC, Price moves in the same direction as: size only

Question 23

Which of the following is true about the scree plot

```
# scree plot
source("PCAplots.R")
```

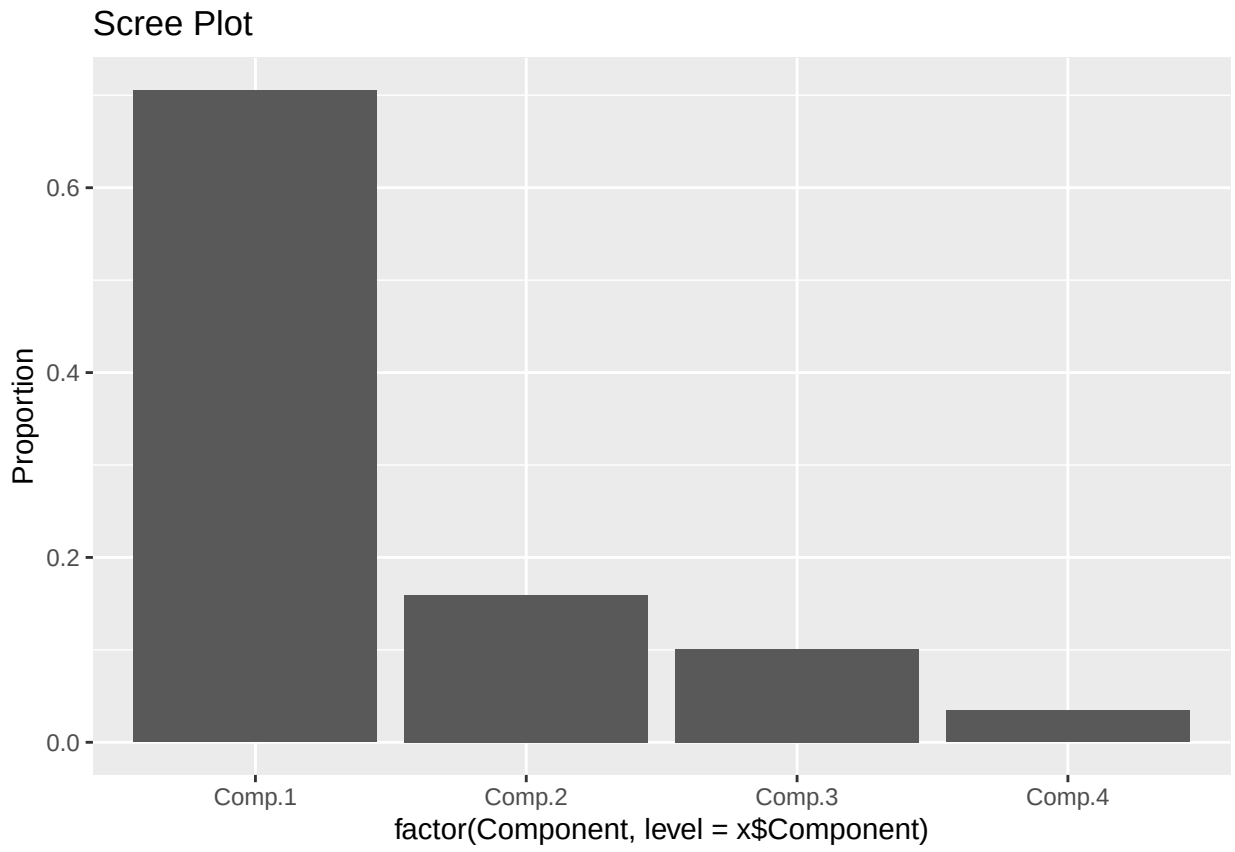
```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following object is masked from 'package:ggpubr':
##
##     mutate

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:here':
##
##     here

##
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':
##
##     alpha, rescale
```

```
ggscreeplot(houses.corr)
```

```
## $var
##    Component Proportion
## 1:    Comp.1 0.70535753
## 2:    Comp.2 0.15909899
## 3:    Comp.3 0.10095153
## 4:    Comp.4 0.03459195
##
## $plot
```

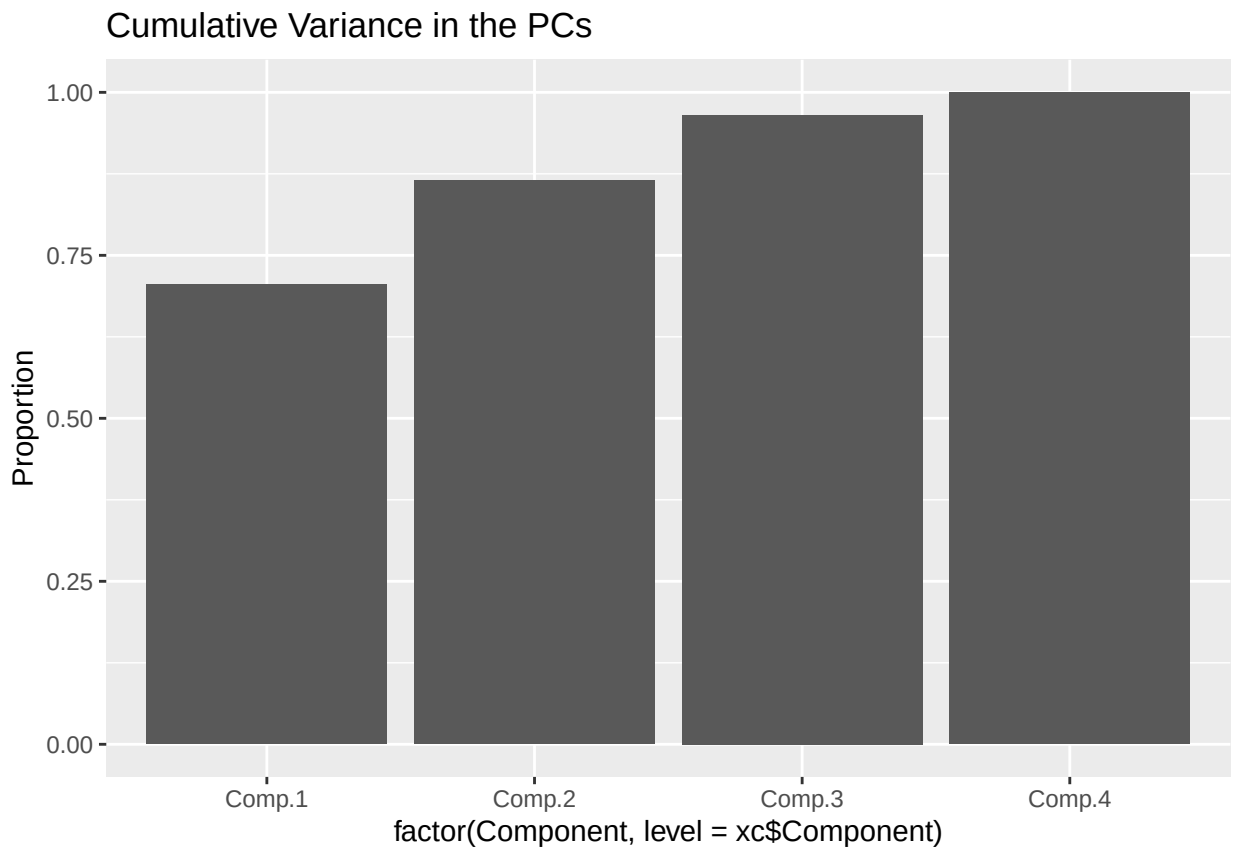


Question 24

How many principal components would it take to account for 90 percent of the variance?

```
cumplot(houses.corr)
```

```
## $cumvar
##    Component Proportion
## 1:    Comp.1  0.7053575
## 2:    Comp.2  0.8644565
## 3:    Comp.3  0.9654081
## 4:    Comp.4  1.0000000
##
## $plot
```

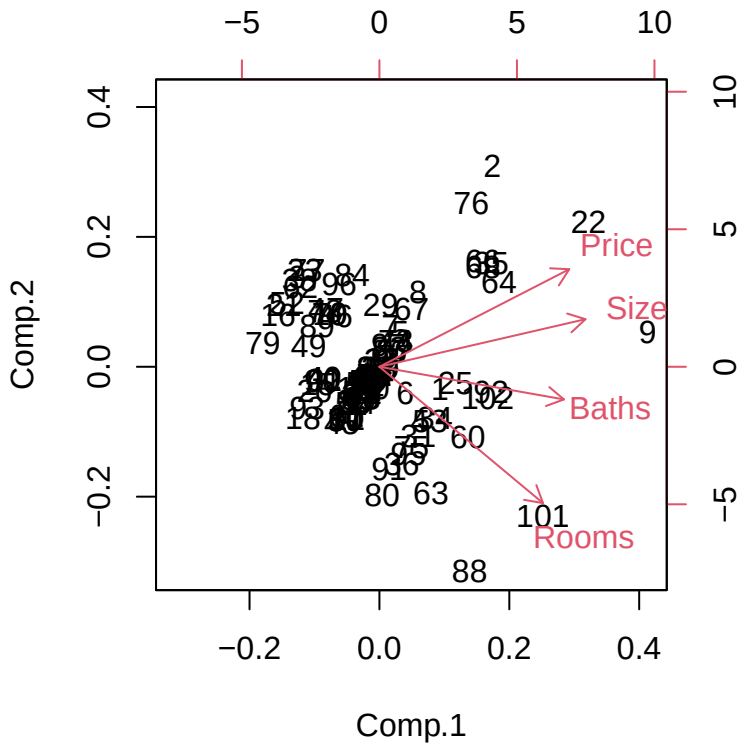



```
# 3 pcs
```

Question 25

Make a biplot of the data in the first two PCs. Which variables appear most correlated in the first two PCs?

```
biplot(houses.corr)
```




```
# price and size
```

Question 26

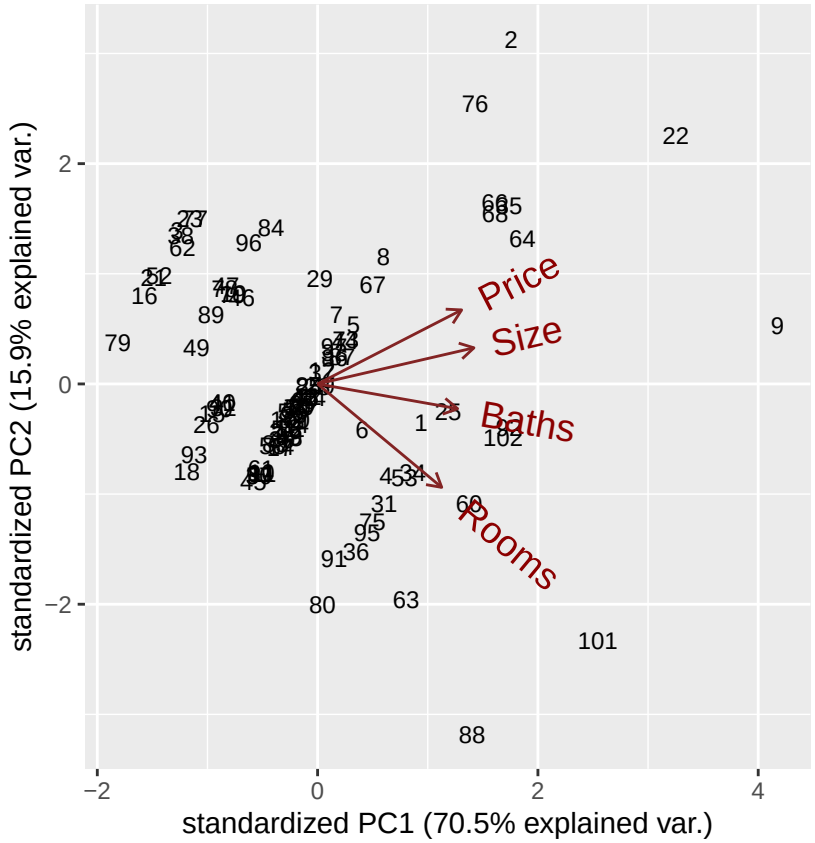
Based on the biplot, which two variables appear most independent in the first two PCs?

closest to perpendicular -> price and rooms

Question 27

Find the observation with the largest absolute value of PC2. What is the size of this house in square feet?

```
ggbiplot(houses.corr, varname.size = 5, labels=row(housing_prices)[,1], plot.obs=TRUE,
         xlim=c(-0.2,0.4), ylim=c(-0.3,0.3))
```




```
# observation 88:

housing_prices[88,]
```

```
##    Rooms Baths Size Age  Price
## 88     5     3 2200 Old 118300
```

```
#housing_prices[2,]
```

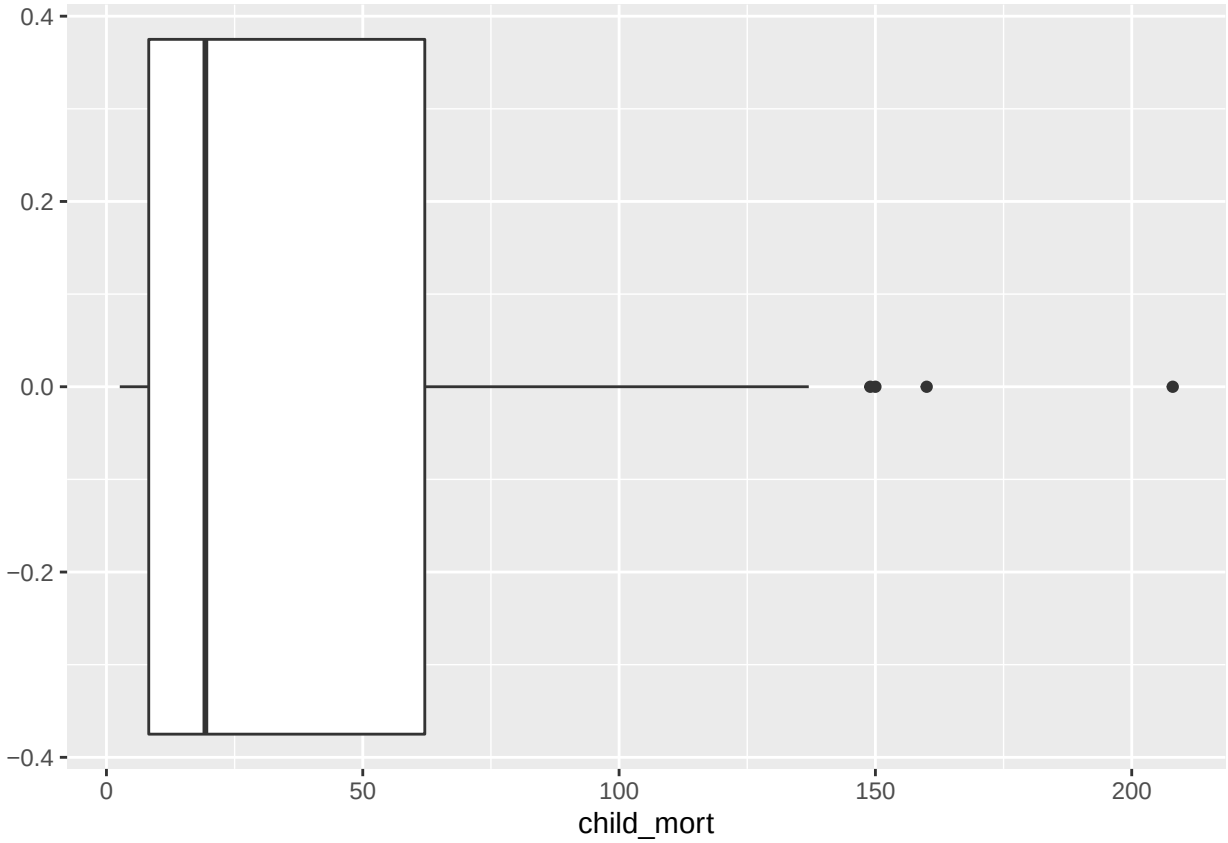
Part 2

Question 28

Make a boxplot of child_mort. How many outliers are there? (Hint: use the statistics of the boxplot if you can't count them in the figure)

```
country <- read.csv("Country-data.csv")
```

```
ggplot(data=country, aes(x=child_mort)) + geom_boxplot()
```




```
boxplot.stats(country$child_mort)
```

```
## $stats
## [1]   2.60   8.25  19.30  62.10 137.00
##
## $n
## [1] 167
##
## $conf
## [1] 12.71608 25.88392
##
## $out
## [1] 149 150 208 160
```

Question 29

What countries do the outliers correspond to?

```
# observations are where child mort = 149, 150, 208, 160
country %>% filter(child_mort %in% c(149,150,208,160))
```

```
##                    country child_mort exports health imports income inflation
## 1 Central African Republic        149    11.8   3.98    26.5    888      2.01
## 2                     Chad        150    36.8   4.53    43.5   1930      6.39
```

```
## 3                   Haiti        208    15.3   6.91    64.7   1500      5.45
## 4            Sierra Leone        160    16.8  13.10    34.5   1220     17.20
##   life_expec total_fer gdpp
## 1       47.5      5.21  446
## 2       56.5      6.59  897
## 3       32.1      3.33  662
## 4       55.0      5.20  399
```

Question 30

Which country has the lowest child mortality rate?

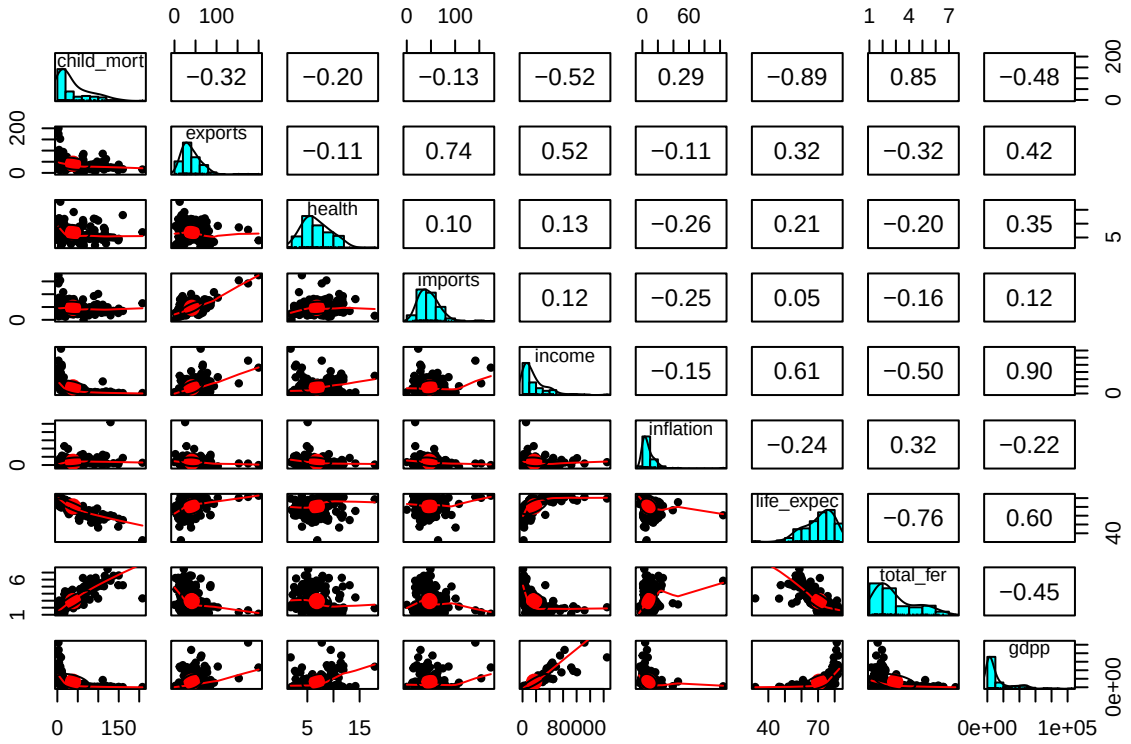
```
country %>% arrange(child_mort) %>% head()
```

```
##      country child_mort exports health imports income inflation life_expec
## 1    Iceland        2.6    53.4   9.40    43.3  38800     5.470       82.0
## 2 Luxembourg        2.8   175.0   7.77   142.0  91700     3.620       81.3
## 3  Singapore        2.8   200.0   3.96   174.0  72100    -0.046       82.7
## 4    Finland        3.0    38.7   8.95    37.4  39800     0.351       80.0
## 5     Sweden        3.0    46.2   9.63    40.7  42900     0.991       81.5
## 6      Japan        3.2    15.0   9.49    13.6  35800    -1.900       82.8
##   total_fer   gdpp
## 1      2.20  41900
## 2      1.63 105000
## 3      1.15  46600
## 4      1.87  46200
## 5      1.98  52100
## 6      1.39  44500
```

Question 31

Make a scatter plot of all variables in the dataset except "Country." Which variable's correlation coefficient with child_mort has the greatest absolute value?

```
pairs.panels(country[,c("child_mort", "exports", "health", "imports", "income","inflation","life_expec"
```



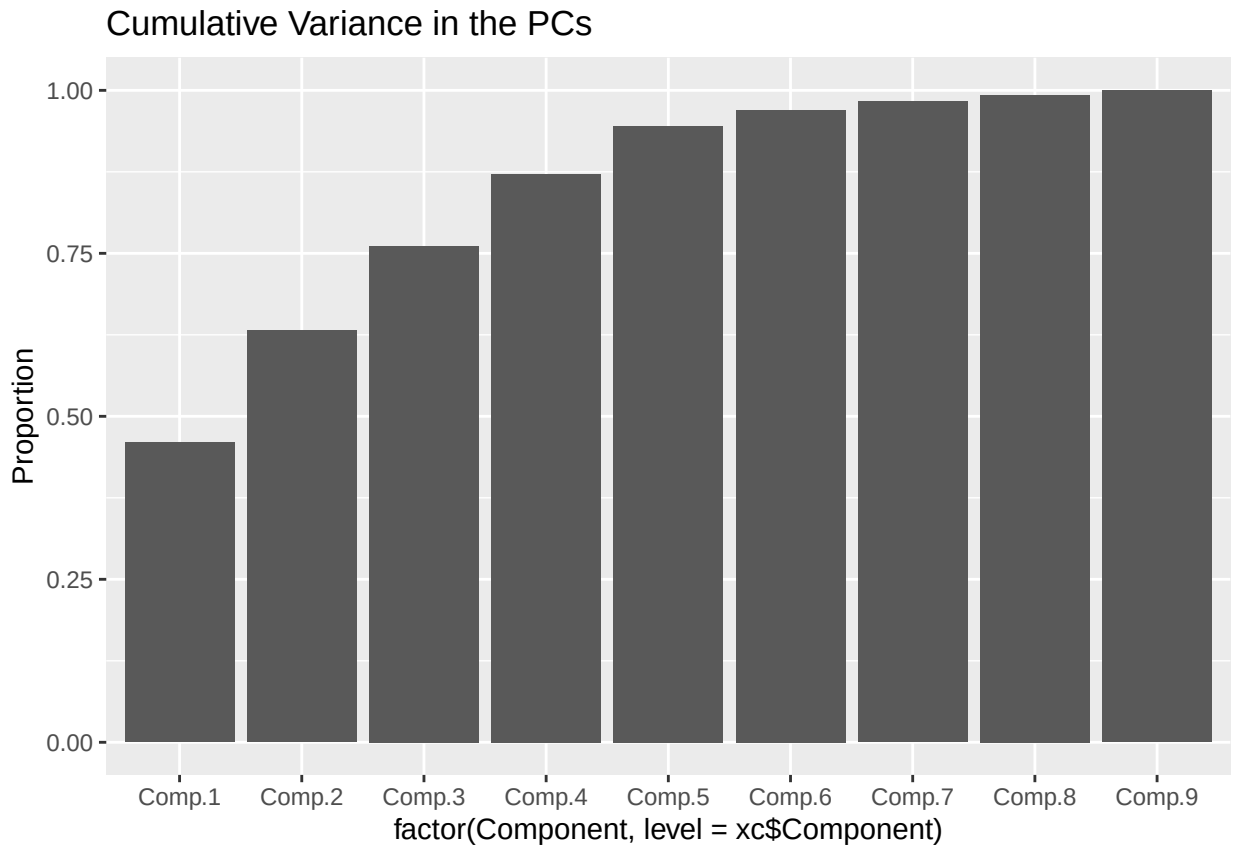
Question 32

Using the correlation matrix, perform PCA on this dataset, excluding the "Country" variable. How many PCs are needed to explain 90% of the variance in the dataset?

```
countries.corr <- princomp(country[,c("child_mort", "exports", "health", "imports", "income","inflation
```

```
cumplot(countries.corr)
```

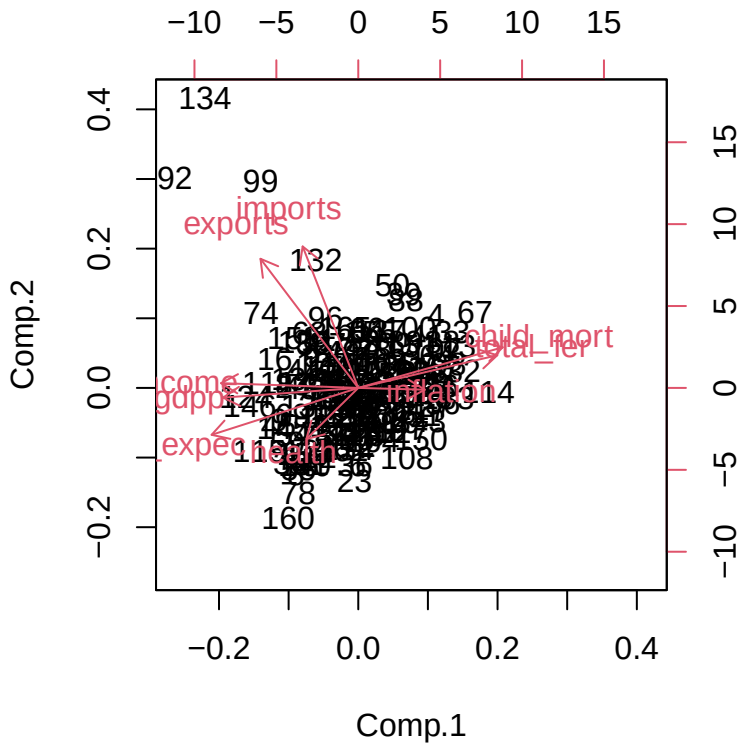
```
## $cumvar
##    Component Proportion
## 1:    Comp.1  0.4595174
## 2:    Comp.2  0.6313337
## 3:    Comp.3  0.7613762
## 4:    Comp.4  0.8719079
## 5:    Comp.5  0.9453100
## 6:    Comp.6  0.9701523
## 7:    Comp.7  0.9827566
## 8:    Comp.8  0.9925694
## 9:    Comp.9  1.0000000
##
## $plot
```



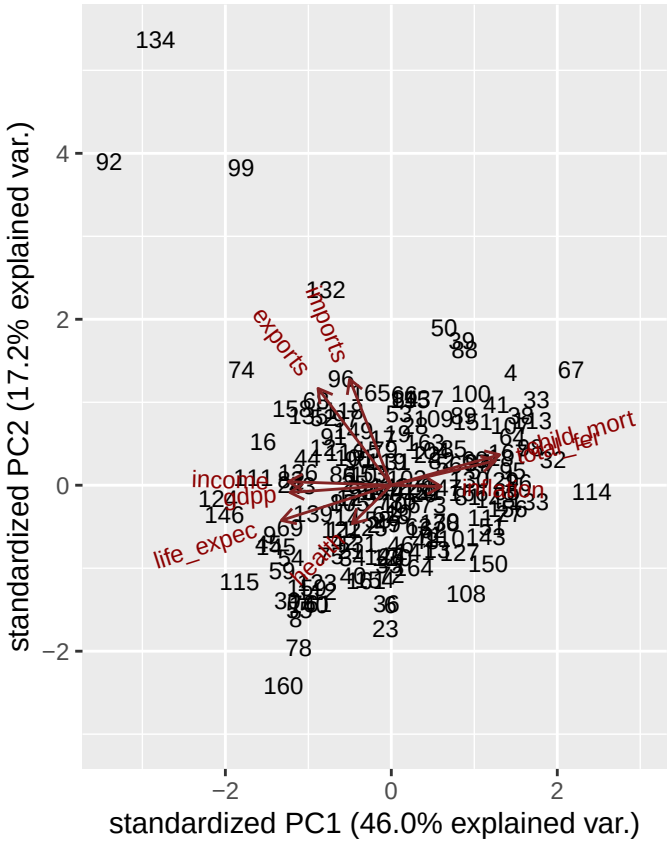
Question 33

Make a biplot of the data in the first two PCs. Which variable is most positively correlated with child mortality in the first two PCs?

```
biplot(countries.corr)
```




```
# without observations
ggbiplot(countries.corr, varname.size = 3, labels=row(country)[,1], plot.obs=TRUE,
         xlim=c(-3,3), ylim=c(-3,3))
```



Question 34

Based on the biplot, which variable is most unrelated to child_mort? imports, almost exactly 90 degrees

Question 35

```
# obs 114
country[114,]
```

```
##     country child_mort exports health imports income inflation life_expec
## 114 Nigeria        130    25.3   5.07    17.4   5150       104       60.5
##     total_fer gdpp
## 114      5.84 2330
```

Question 36

```
# highest in PC 2 is obs 134
country[134,]
```

```
##       country child_mort exports health imports income inflation life_expec
## 134 Singapore        2.8     200   3.96     174  72100    -0.046       82.7
```

```
##     total_fer  gdpp
## 134      1.15 46600
```

Question 37

How many variables vary in the same direction as child_mort in the first PC?

```
countries.corr$loadings
```

```
##
## Loadings:
##            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## child_mort  0.420  0.193         0.371  0.169  0.201         0.683  0.328
## exports    -0.284  0.613 -0.145                       0.707        -0.123
## health     -0.151 -0.243  0.597  0.462 -0.518         0.250         0.113
## imports    -0.161  0.672  0.300        -0.255        -0.592
## income     -0.398        -0.302  0.392  0.247  0.160        -0.353  0.613
## inflation   0.193        -0.643  0.150 -0.715        -0.105
## life_expec -0.426 -0.223 -0.114 -0.204 -0.108 -0.601         0.505  0.294
## total_fer   0.404  0.155         0.378  0.135 -0.751        -0.293
## gdpp       -0.393        -0.123  0.532  0.180        -0.243  0.250 -0.626
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111  0.111
## Cumulative Var  0.111  0.222  0.333  0.444  0.556  0.667  0.778  0.889  1.000
```

Question 38

Which two variables have the largest absolute loading in the second PC?

exports and imports