528HW1

Coco_Luo

2023-01-09

Introduction

- The population of the CHS study consists of adults aged 65 years and older recruited in 1989–1990 from four communities: Forsyth County, North Carolina; Sacramento County, California; Washington County, Maryland; and Pittsburgh, Pennsylvania
- The outcome variable is mortality
- Our aim is to investigate on the association of mortality with exercise variables and a number of other potential risk factors

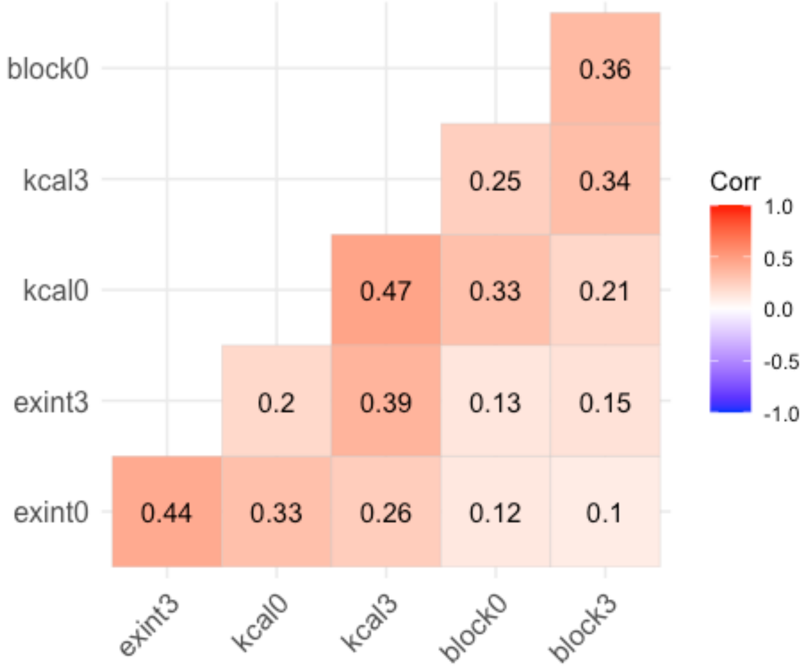
EDA

I first deleted all the missing values in the variable columns we are interested in analysis, I did this for the ease of visualization and further analysis. We do not want missing variables to mess our analysis. After deleting the missing values in the exercise variable, we still have 1668 observations left, which is still large enough to provide convincing results.

I would like to take a first look at the summary statistics of baseline measure of exercise intensity(exint0), Baseline measure of blocks walked in last 2 weeks(block0), and baseline measure of estimated kilocalories expended (kcal0); I will also look at the same variables measured in 3 year recruitment. I considered all the six variables as exercise variables.

I found that the mean of baseline is 1.534, but the mean of exercise after 3 years is lower: 1.5. The mean for block0 is 48.66, and the mean for block3 is 54.93; The mean for kcal0 is 1420 with median 810, and the mean for kcal3 is 1174.6 with median 694.4. It seems that at 1998, people walked less blocks and spend more calories on average than those recorded in 3 years, this might indicates that people are engaging in more intense activities at that time.

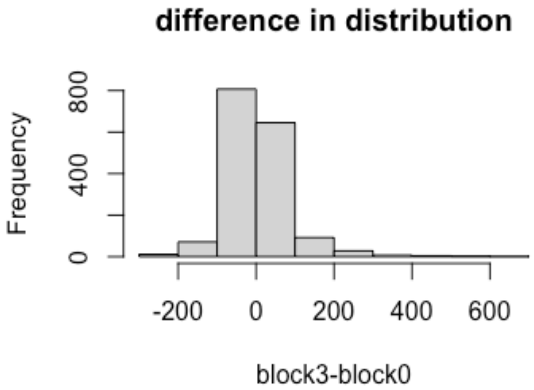
To better visualize their relationships and changes over time. I used a correlation plot which provides a good measure on how those variables are related, kcal0 and kcal3 has a correlation score 0.47, exint0 and exint3 has a correlation score 0.44, exint3 and kcal3 has a correlation 0.39, block0 and kcal0 is 0.36. It makes sense that all those variables are related to each other in some extend since walking is also a kind of activity and all kinds of activities or exercise will spend some energy. None of the correlation scores are higher than 0.6.

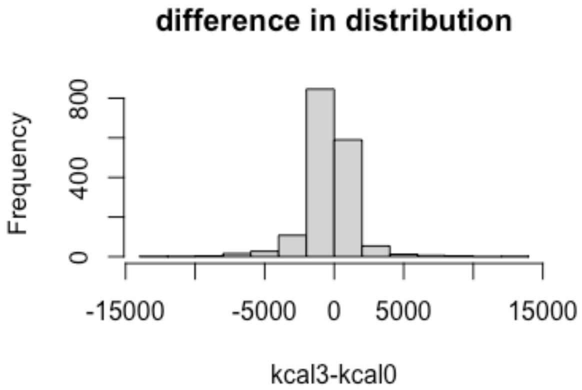


Exercise variables and their proportions over time

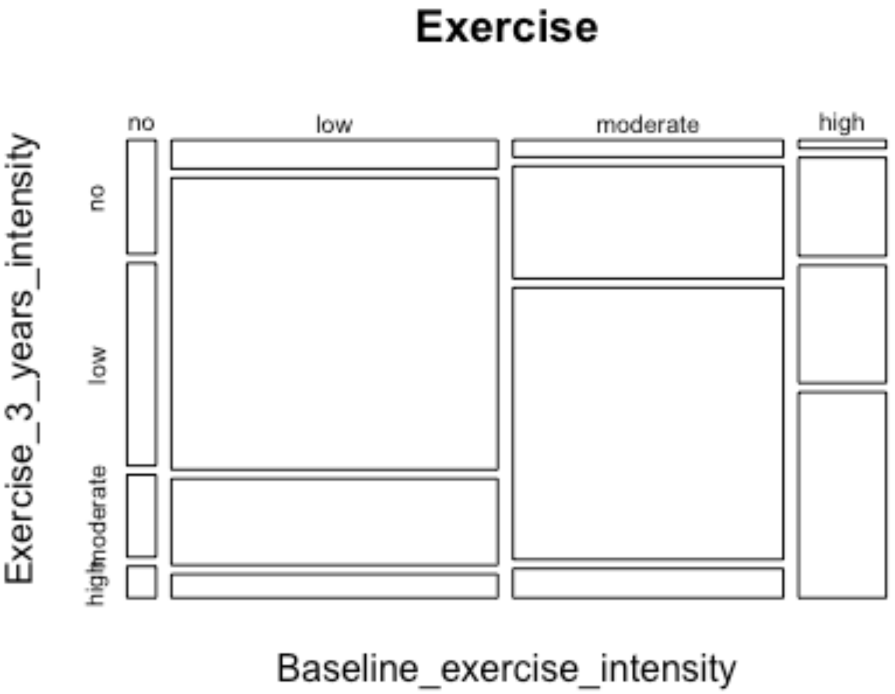
Then, I used table combined with the Mosaic plots to visualize how the exercise variables change over time. From all four categories, it seems that for each category, the majority of people stays, lots of no exercise person starts to exercise a little, many people changed between low to medium intensity. Mosaic plot is a good way to provide a summary of the data and allows for the identification of correlations between distinct variables. In other words, independence is demonstrated when all of the boxes in the same category have the same areas. There are a few disadvantages to using mosaic plots. For one, they can be difficult to read because of the large number of small squares, and the sizes of the squares do not always accurately reflect the relative proportions of the variables. It can also be difficult to determine the exact proportions of each variable. Further, the plot can be cluttered and difficult to interpret if there are a large number of categorical variables. Because of those problems, I decided to add a table above the plots for users to better review the exact numbers for each category and compare the differences. Finally, I made the color of the mosaic plot to be all white so that there would not be too much inky.

To better view how block and kcal changes, I used the historgam to visualize the distribution of the difference in the pairs of variables. Both histograms show me an approximate normal distirbution centered at 0. Since most of the values are around zero, there are not much differences in the block variable and kcal variable 3 years later.



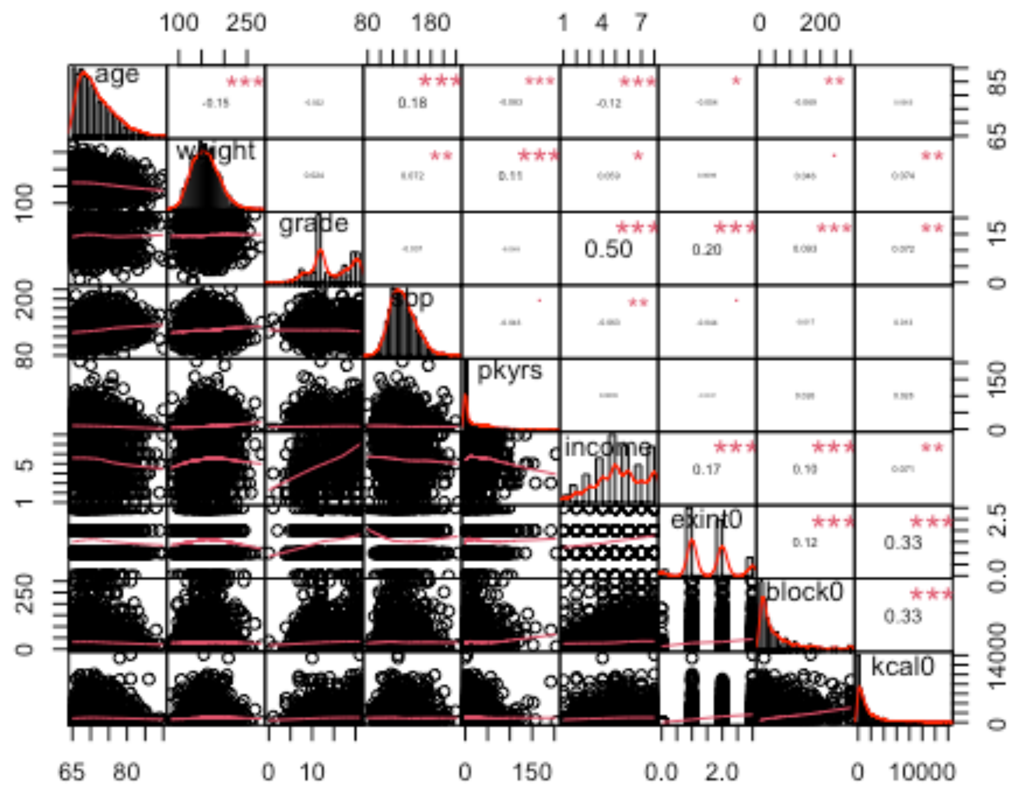


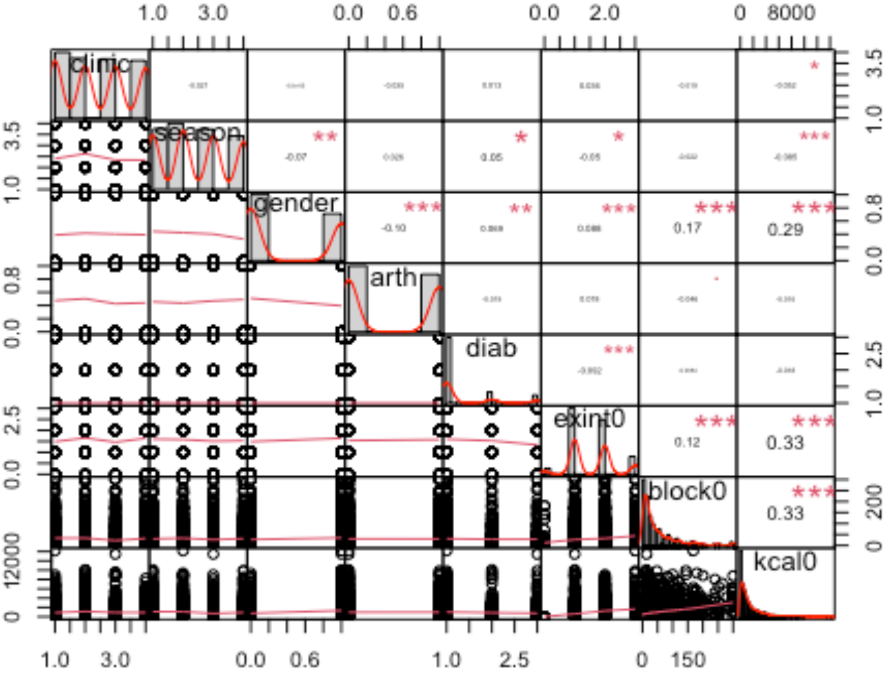
	no	low	moderate	high
no	18	32	13	5
low	51	517	153	42
moderate	25	165	399	43
high	4	47	56	98

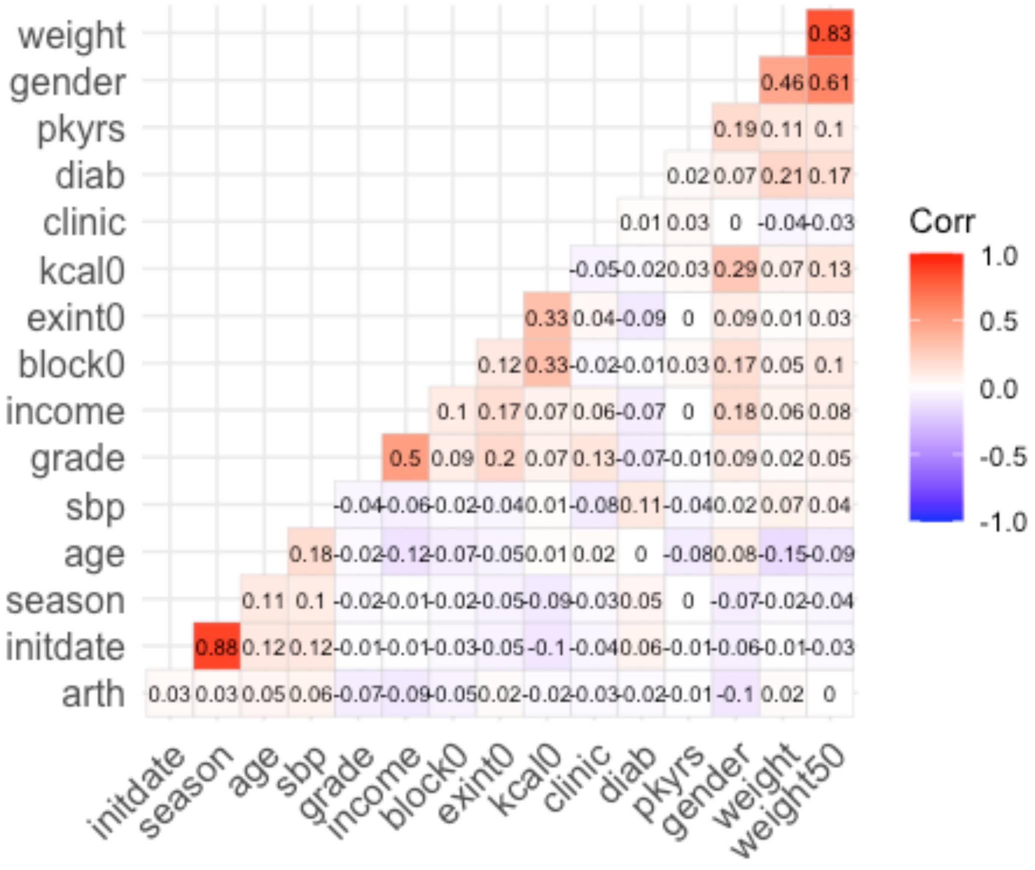


Now, if we do not think about mortality and myocardial infarction status, and only consider about the relationship between baseline exercise variable with all the other baseline variables, again, I used the correlation plots to visualize relationships for numerical baseline variables and another barplot for factor variables ones. At the end, I visualize all the correlation scores where I did not found any significate large correlations between exercise variables and other baseline variables. However, I only found season and initdate as a correlation as high as 0.88, weight and weight50 has a correlation as high as 0.83. Other than, initdate, age, sbp, season and diab are negatively related to exercise, where the other baseline variables are positively related; initdate, age, sbp, clinic, season, arth, and diab are negatively related to block walked , where the other baseline variables are positively related; inidate, clinic, seasonm arth, and diab are negatively related to kcal, where the other baseline variables are positively related. Most of the correlation score lower below 0.1 and there are no relationships with scores higher than 0.33.

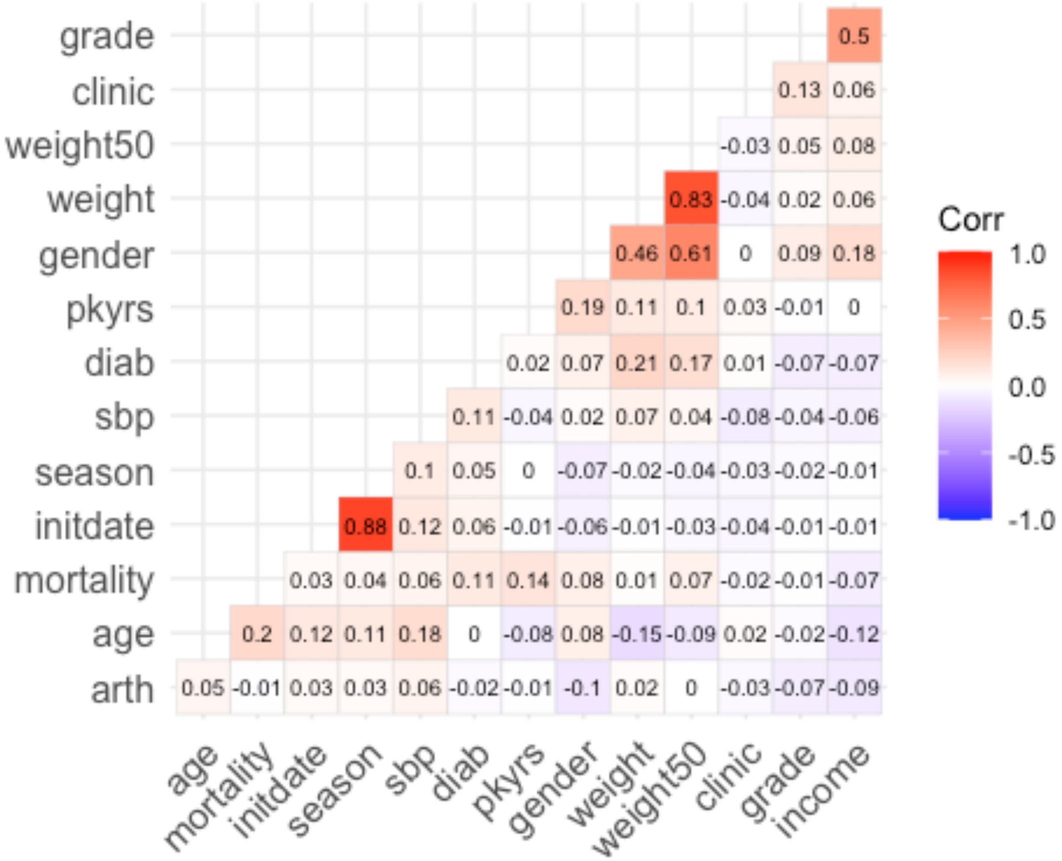
Although the relationships are weak, but we can still see that males spend more calories, older people get less exercise, heavier weighted individuals spend more energy during exercise but they do not exercise as frequent as lighter individuals, better educated people exercise more often, people have arthritis issue tend do less exercise, higher blood pressure person do less exercises, diabetes affect time spend on exercise, and the higher people earn the more exercise they will do.



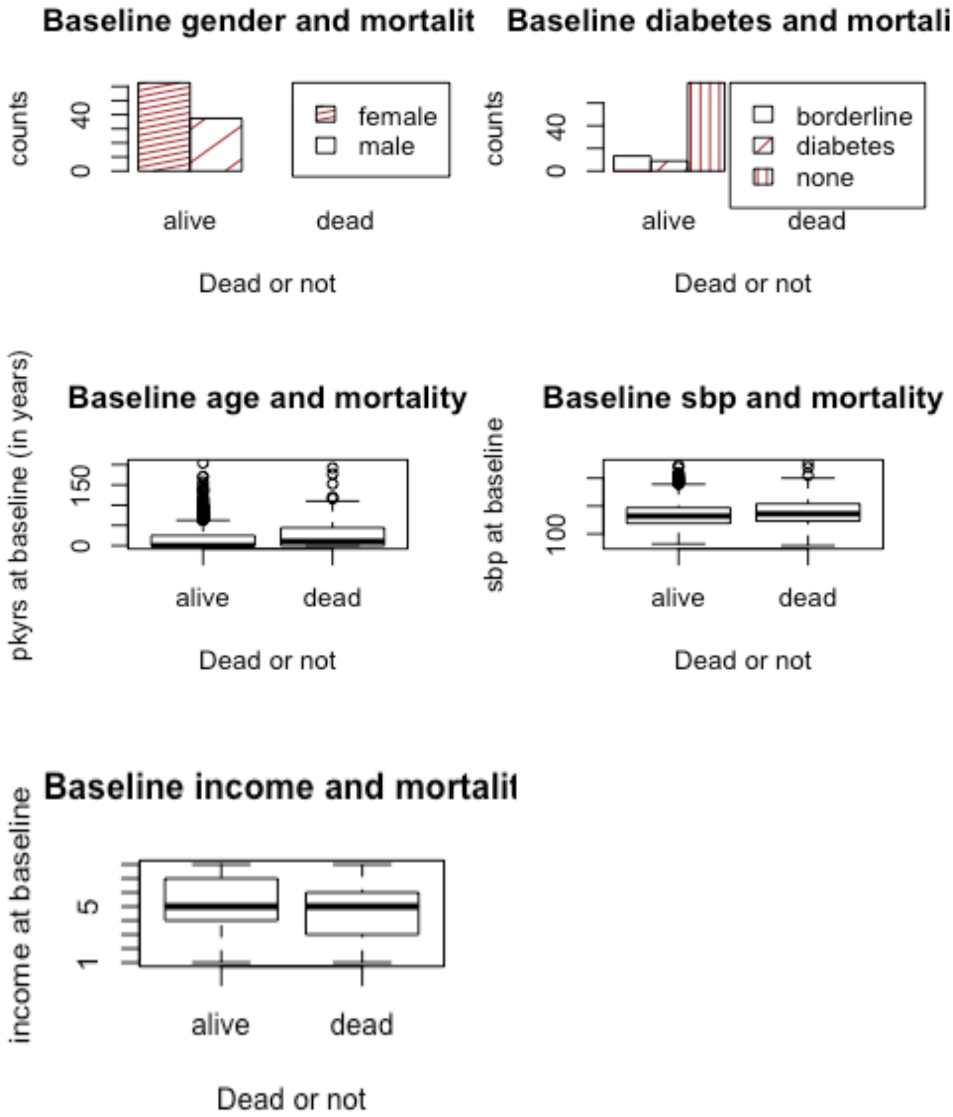

clinic
season
gender
arth
diab
exint0
block0
kcal0
1.0
3.0
0.0
0.6
0.0
2.0
0
8000
3.5
1.0
0.8
0.0
2.5
1.0
200
0
12000
0
1.0
3.0
0.0
0.6
1.0
2.5
0
150
*
**
-0.07
*
0.05
*
-0.05
***
***
-0.10
**
***
***
0.17
***
0.29
***
***
0.12
***
0.33
***
0.33


weight
gender
pkyrs
diab
clinic
kcal0
exint0
block0
income
grade
sbp
age
season
initdate
arth
initdate
season
age
sbp
grade
income
block0
exint0
kcal0
clinic
diab
pkyrs
gender
weight
weight50
Corr
1.0
0.5
0.0
-0.5
-1.0
0.83
0.46 0.61
0.19 0.11 0.1
0.02 0.07 0.21 0.17
0.01 0.03 0 -0.04 -0.03
-0.05 -0.02 0.03 0.29 0.07 0.13
0.33 0.04 -0.09 0 0.09 0.01 0.03
0.12 0.33 -0.02 -0.01 0.03 0.17 0.05 0.1
0.1 0.17 0.07 0.06 -0.07 0 0.18 0.06 0.08
0.5 0.09 0.2 0.07 0.13 -0.07 -0.01 0.09 0.02 0.05
-0.04 -0.06 -0.02 -0.04 0.01 -0.08 0.11 -0.04 0.02 0.07 0.04
0.18 -0.02 -0.12 -0.07 -0.05 0.01 0.02 0 -0.08 0.08 -0.15 -0.09
0.11 0.1 -0.02 -0.01 -0.02 -0.05 -0.09 -0.03 0.05 0 -0.07 -0.02 -0.04
0.88 0.12 0.12 -0.01 -0.01 -0.03 -0.05 -0.1 -0.04 0.06 -0.01 -0.06 -0.01 -0.03
0.03 0.03 0.05 0.06 -0.07 -0.09 -0.05 0.02 -0.02 -0.03 -0.02 -0.01 -0.1 0.02 0

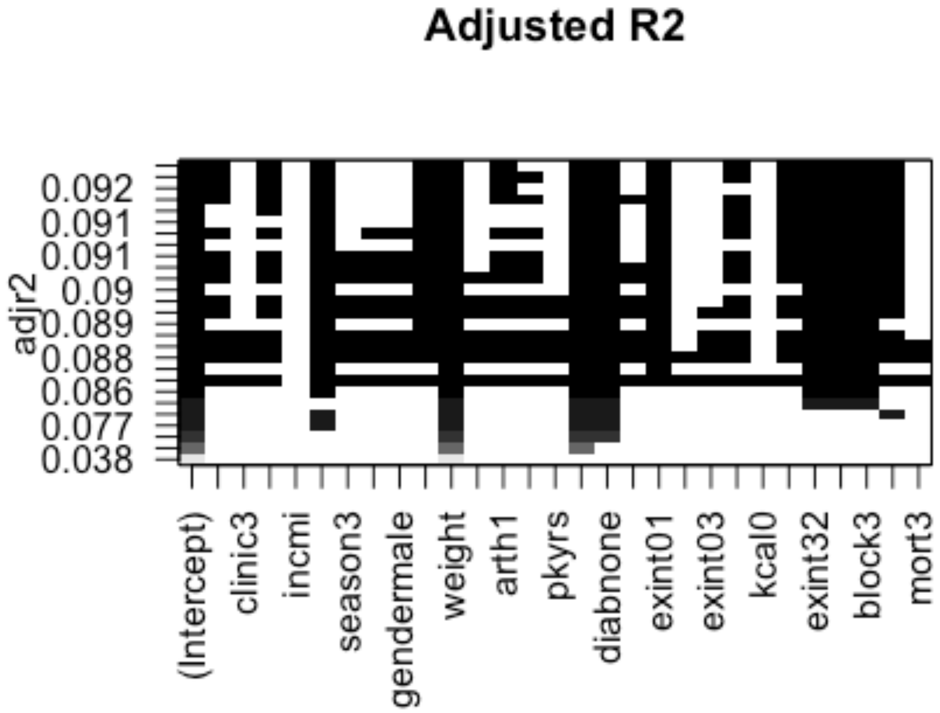
Finally, I looked for the association between mortality at the end of the study 1998 and baseline variables, not including the exercise variables. From the correlation plots, it seems that pack years of smoking history and diabetes have a higher correlation score with mortality. Systolic blood pressure, gender and income are slight related to mortality. The other baseline variables have correlation score less than 0.05.



To look into more details, as 0 means alive and 1 means dead (mortality), for the first two plots I used the percentage bar plot to make sure that all the levels are measured on the same scale. For gender, I see that more females alive than dead, but more males dead than alive; for diabetes, more people diead tha alive; people have a longer smoking history are more likely to die; people have blood pressure have higher chance to die, and people have higher income is more likely to live. Rather than using the color to distinguish different categories, I used different textures, in this way, not only there would not be too much ink, but is also friendly for color blindness.


Baseline gender and mortalit
counts
0
40
alive
dead
female
male
Dead or not
Baseline diabetes and mortali
counts
0
40
borderline
diabetes
none
alive
dead
Dead or not
Baseline age and mortality
pkyrs at baseline (in years)
0
150
alive
dead
Dead or not
Baseline sbp and mortality
sbp at baseline
100
alive
dead
Dead or not
Baseline income and mortalit
income at baseline
1
5
alive
dead
Dead or not

Conclusion



From the EDA, it seems that exercise is beneficial with respect to mortality in healthy individuals over 65 years of age. This is because we see more people alive with at least some degree of exercise than people with no exercise. People who exercise more also walked more blocked and spend more calories.

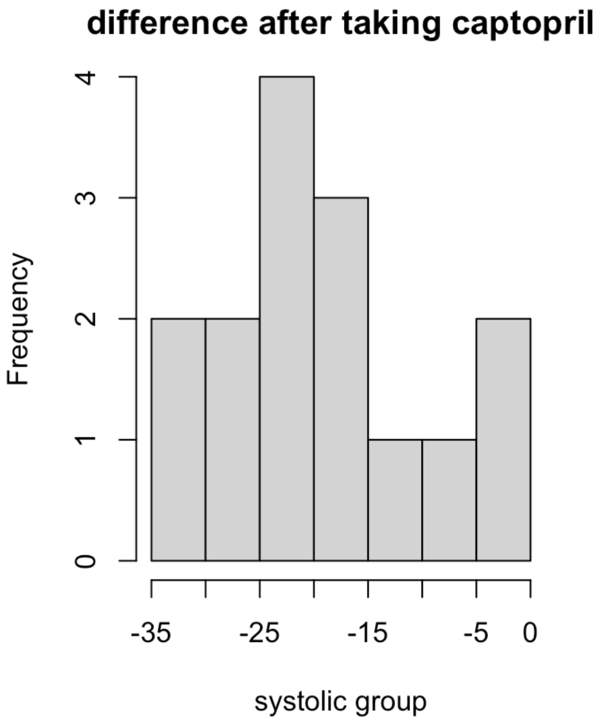
Since mortality is a binary outcome variable, we can fit the multiple logistic regression. To find the best model, I ran the best subset selection by fitting the full model first and than let the algorithm to choose the model that provides the highest adjusted R^2. The data I used included all variables in the current dataset excluding inidate and weight50 since they are highly associated with season and weight, and those variables does not impact much on the aim of this analysis. From the graphical output, we can see that the model with fall season, age, weight, borderline diabetes, income, exint0 with low intensity, exint3, block0, block3, kcal3 gives us the best fitted model amomg other models with $R_{adj}^2 = 0.91$ and least number of chosen variables. I will go ahead and fit my final model.

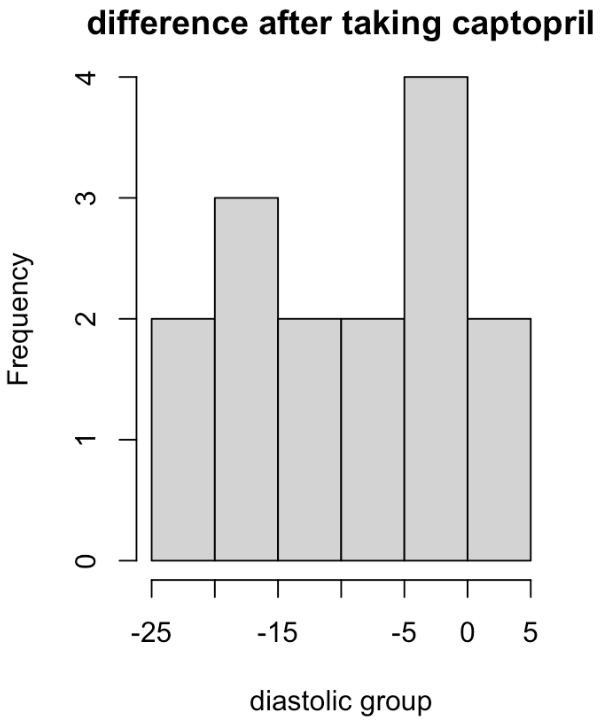
Both the exploratory data analysis and model selection suggest that exercise is goo with respecct to mortality in healthy individuals over 65 years of age. I made the conclusion and end the study.

Question 2

Our scientific question is to evaluate the effectiveness of captopril for decreasing blood pressure (outcome variable: systolic and dialostic response to the drug) where the population is 15 patients with moderate essential hypertension, supine systolic and diastolic blood pressures immediately before and two hours after taking 25 mg of the drug captopril. The table provided by the researchers, however, did not provide us a clear illustration on whether the captopril leads to an increase or decrease of blood pressure. Although we can see the differences, it is extremely not clear on the trends. It is also not clear on the comparison between the systolic and the diastolic group.

To improve this table, I think we can use two tables to separate the parameters we are interested to look at. We can also use different colors to mark the postive differences and negative differences so that viewers can see whether there is a decrease or increase. I also think about having a summary statistic table to show the average, median and IQR range of the differences in our population group.





To better view changes, I used the historgam to visualize the distribution of the difference in blood pressure (mm Hg) before and after taking the captopril. In the systolic group, we can see that captopril have a moderate effect on decreasing the blood pressure as most of the people have a strong decrease from -20 to -35. There are a few people have week decrease from 0 to -10. On the other side, looking at the diastolic group, the effect is not as strong, there are even a few positive numbers indicating that the blodd pressure increases.

Appendix

```
# code for the model
df = na.omit(CHSdataEx1)
df <- df[ -c(3, 10) ]
df$exint3 = as.factor(df$exint3)
df$exint0 = as.factor(df$exint0)
df$mortality = as.factor(df$mortality)
df$season = as.factor(df$season)
df$clinic = as.factor(df$clinic)
df$gender = as.factor(df$gender)
df$arth = as.factor(df$arth)
df$diab = as.factor(df$diab)
library(leaps)
# Fit the model
suppressWarnings(regsubsets.out <- regsubsets(
  mortality ~., data = df,
               nbest = 1,       # 1 best model for each number of predictors
               nvmax = NULL,    # NULL for no limit on number of variables
               force.in = NULL, force.out = NULL,
               method = "exhaustive"))
plot(regsubsets.out, scale = "adjr2", main = "Adjusted R2")
```