Final project

Leo

11/06/22

First include basic settings and related packages.

```
> library(tidyverse)
```

```
## Warning:  'tidyverse' R 4.2.1
```

```
## Warning:  'tibble' R 4.2.1
```

```
> library(skimr)
```

```
## Warning:  'skimr' R 4.2.1
```

```
> library(MASS)
```

```
## Warning:  'MASS' R 4.2.1
```

```
> library(openxlsx)
```

```
## Warning:  'openxlsx' R 4.2.1
```

```
> library(pander)
```

```
## Warning:  'pander' R 4.2.1
```

Read and roughly view data 'JAAD RAW DATA' in project's working directory

```
> rawdata <- read.xlsx("JAAD RAW DATA.xlsx")
> head(rawdata)
```

```
##   FEMALE AGE.YEARS AGA NHS LS
## 1      1        71   0  NA  0
## 2      1        30   0  NA  0
## 3      1        68   1  NA  1
## 4      1        72   0  NA  0
## 5      1        64   1  NA  3
## 6      1        51   0  NA  0
```

```
> skim(rawdata)
```

Table 1: Data summary

Name	rawdata
Number of rows	175
Number of columns	5
Column type frequency:	
numeric	5
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
FEMALE	0	1.0	0.30	0.46	0	0	0	1	1	
AGE.YEARS	0	1.0	64.99	14.48	23	54	66	75	93	
AGA	0	1.0	0.67	0.47	0	0	1	1	1	
NHS	53	0.7	3.42	1.84	1	2	3	5	7	
LS	122	0.3	0.85	1.15	0	0	0	2	3	

Effect of age on AGA level in male or female

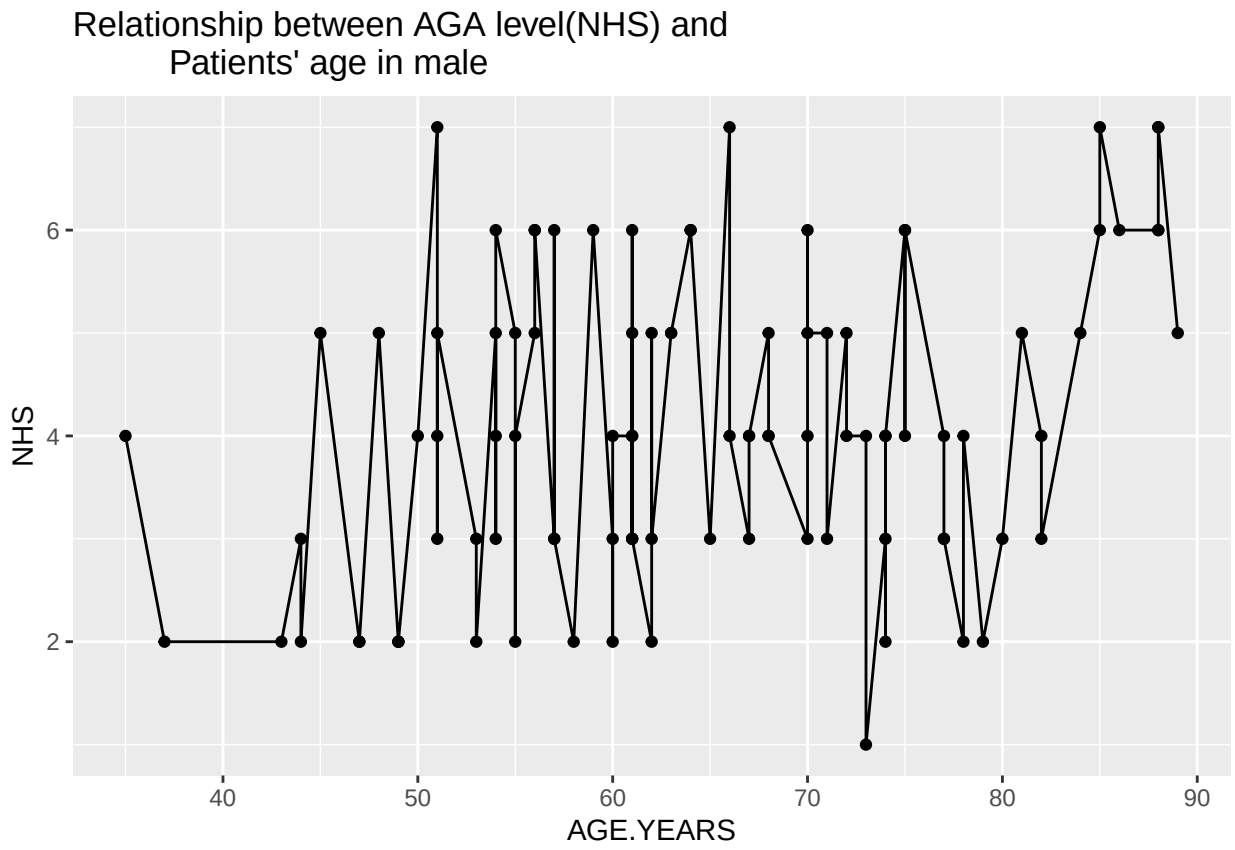
1st. Divide data into two groups via gender and delete col with NA

```
> male_data <- rawdata[grep(pattern="0",rawdata$FEMALE),]
> male_data <- male_data[grep(pattern="1",male_data$AGA),]
> male_data <- male_data[,-5]
> female_data <- rawdata[grep(pattern="1",rawdata$FEMALE),]
> female_data <- female_data[grep(pattern="1",female_data$AGA),]
> female_data <- female_data[,-4]
```

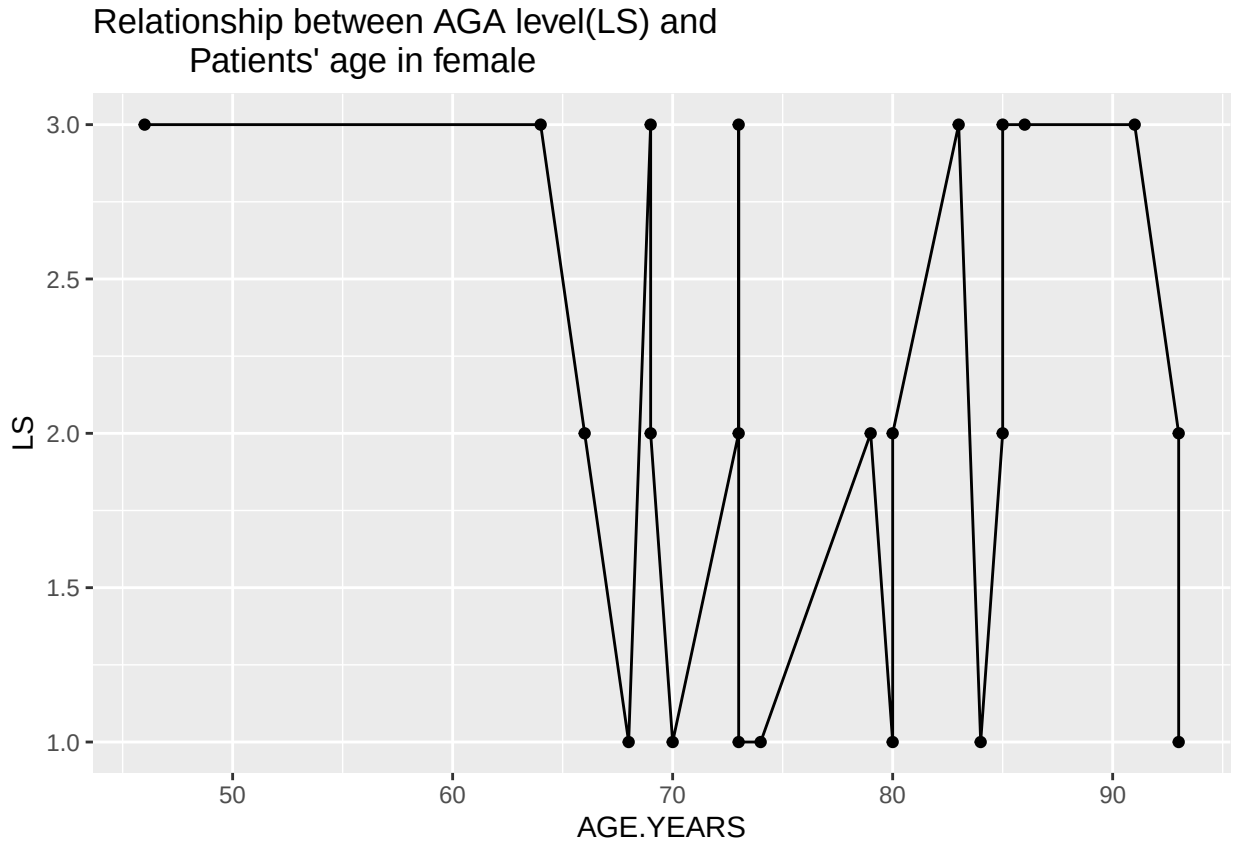
2nd. Conduct some EDA

Point and line plot of AGA level against AGE.YEARS

```
> male_data%>%
+   ggplot(mapping = aes(x = AGE.YEARS,y = NHS)) +
+   geom_point() +
+   geom_line() +
+   ggtitle("Relationship between AGA level(NHS) and
+           Patients' age in male")
```




```
> female_data%>%
+   ggplot(mapping = aes(x = AGE.YEARS,y = LS)) +
+   geom_point() +
+   geom_line() +
+   ggtitle("Relationship between AGA level(LS) and
+           Patients' age in female")
```



3rd. Fit a linear regression model to explain the AGA level

```
> full.model1 <- lm(NHS ~ .,data = male_data)
> summary(full.model1)
```

```
##
## Call:
## lm(formula = NHS ~ ., data = male_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3964 -1.1124 -0.1881  1.0713  3.4368
##
## Coefficients: (2 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.63146    0.75908   2.149  0.03418 *
## FEMALE            NA         NA      NA       NA
## AGE.YEARS    0.03788    0.01156   3.278  0.00147 **
## AGA               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.431 on 94 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.09301
## F-statistic: 10.74 on 1 and 94 DF,  p-value: 0.001468
```

```
> full.model2 <- lm(LS ~ .,data = female_data)
> summary(full.model2)
```

```
##
## Call:
## lm(formula = LS ~ ., data = female_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11667 -0.96458 -0.02083  0.88125  1.07500
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.683333   1.306809   2.053   0.0533 .
## FEMALE             NA         NA      NA       NA
## AGE.YEARS   -0.008333   0.016904  -0.493   0.6274
## AGA                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8595 on 20 degrees of freedom
## Multiple R-squared:  0.01201,    Adjusted R-squared:  -0.03739
## F-statistic: 0.243 on 1 and 20 DF,  p-value: 0.6274
```

4th. Variable Selection(only one variable maybe meaningless)

```
> stepwiseSelection1 <- stepAIC(full.model1,direction = "both",
+                               trace = FALSE,k = log(NROW(male_data)))
> summary(stepwiseSelection1)
```

```
##
## Call:
## lm(formula = NHS ~ AGE.YEARS, data = male_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3964 -1.1124 -0.1881  1.0713  3.4368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.63146    0.75908   2.149  0.03418 *
## AGE.YEARS    0.03788    0.01156   3.278  0.00147 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.431 on 94 degrees of freedom
## Multiple R-squared:  0.1026, Adjusted R-squared:  0.09301
## F-statistic: 10.74 on 1 and 94 DF,  p-value: 0.001468
```

```
> pander::pander(stepwiseSelection1)
```

Table 3: Fitting linear model: NHS ~ AGE.YEARS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.631	0.7591	2.149	0.03418
AGE.YEARS	0.03788	0.01156	3.278	0.001468

```
> stepwiseSelection2 <- stepAIC(full.model2,direction = "both",
+                               trace = FALSE,k = log(NROW(female_data)))
> summary(stepwiseSelection2)
```

```
##
## Call:
## lm(formula = LS ~ 1, data = female_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04545 -1.04545 -0.04545  0.95455  0.95455
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0455     0.1799   11.37 1.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8439 on 21 degrees of freedom
```

```
> pander::pander(stepwiseSelection2)
```

Table 4: Fitting linear model: LS ~ 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.045	0.1799	11.37	1.958e-10

5th. Interpretation

$$\hat{NHS(male)} = 1.631 + 0.0378 * AGE.YEARS$$

Keep other covariates unchanged, the NHS of male is expected to increase by 0.0378 with every unit increase of AGE.YEARS.

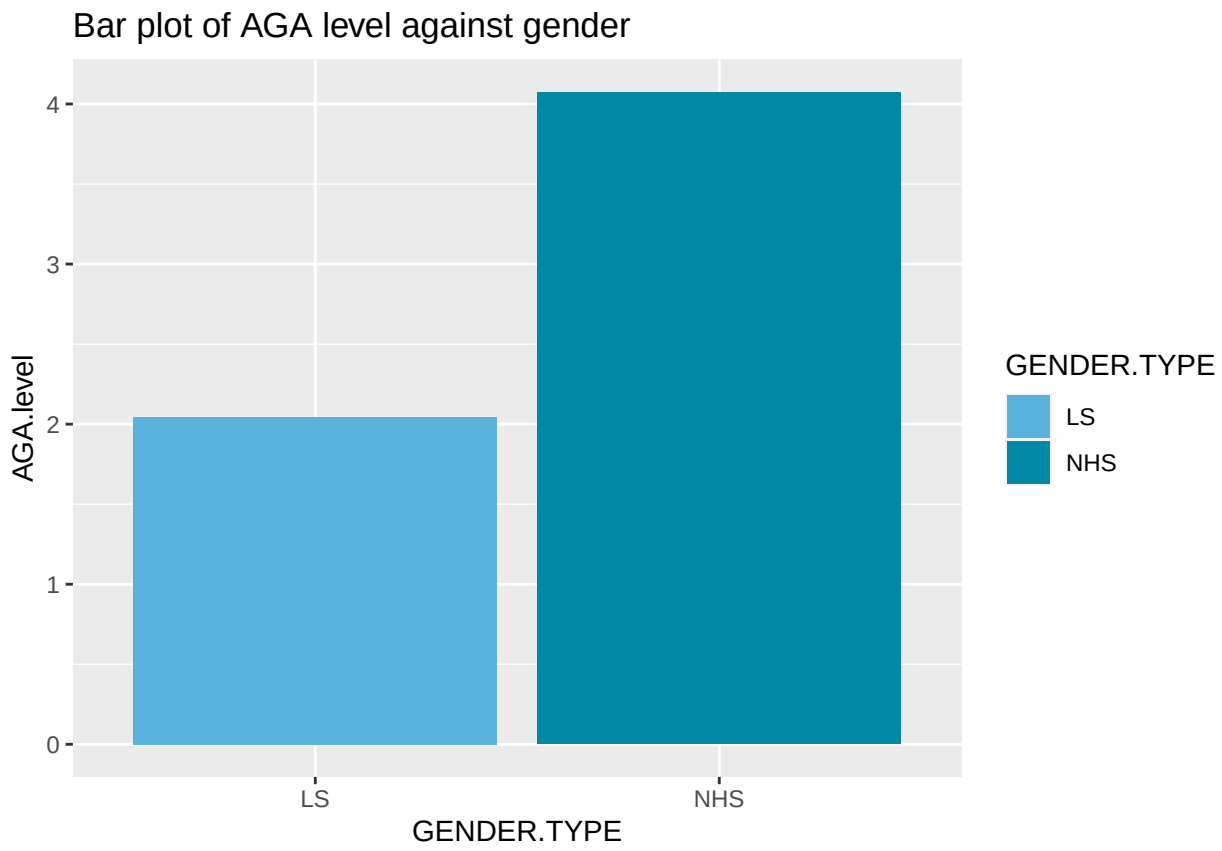
By fitting models, we discovered that there is little correlation between level of AGA(LS) and AGE.YEARS in female.(Maybe it's because the female sample is too small)

Effect of gender on AGA(T-test)

Conduct some EDA

Bar plot of AGA level against gender

```
> AGA_data <- rawdata[grep(pattern="1",rawdata$AGA),]
> combine_data <- AGA_data%>%
+   pivot_longer(c(NHS,LS),names_to = "GENDER.TYPE",values_to = "AGA.level")
> combine_data <- na.omit(combine_data)
> colorset <- c("#58B2DC","#0089A7")
> combine_data%>%
+   group_by(GENDER.TYPE) %>%
+   summarise(n = n(), AGA.level = mean(AGA.level)) %>%
+   ggplot(aes(x = GENDER.TYPE, y = AGA.level, fill = GENDER.TYPE)) + geom_bar(stat = "identity") +
+   ggtitle("Bar plot of AGA level against gender") +
+   scale_fill_manual(values = colorset)
```



Delete cols including NA and conduct T-test

```
> gender_data <- rawdata[,-c(4,5)]
> ttest<- t.test(gender_data$AGA~gender_data$FEMALE)
> pander::pander(ttest)
```

Table 5: Welch Two Sample t-test: `gender_data$AGA` by `gender_data$FEMALE` (continued below)

Test statistic	df	P value	Alternative hypothesis
4.778	84.26	7.41e-06 * * *	two.sided

mean in group 0	mean in group 1
0.7869	0.4151

Discovery

T-test shows that p is less than 0.01, which means there is true differences between male and female in AGA. So the effect of gender on AGA is considerable.