ANA 600 Midterm

##SETTING UP - In class Monday!

STEP 1: Modify the code chunk below to set your working directory and load the ACS.csv provided in Slack STEP 2: Load favorite packages

```
myDataLocation <- "C:\\Program Files\\R\\ANA 600\\R Code\\Midterm Questions"
setwd(myDataLocation)
MyData <- read.csv(file = "ACS.csv", header = TRUE)

library(knitr)
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(ggplot2)
library(ggformula)
```

```
## Loading required package: scales
```

```
## Loading required package: ggridges
```

```
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
```

```
library(supernova)
```

```
##
## Attaching package: 'supernova'
```

```
## The following object is masked from 'package:scales':
##
##     number
```

```
library(lsr)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by thi
s.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:scales':
##
##     rescale
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

##INSTRUCTIONS

For this midterm, I would like you to perform a basic exploratory data analysis (EDA). An EDA consists of the procedures and concepts we have been practicing in class to date. You will be exploring your data set, reviewing the variables, and answering questions about specific aspects of the dataset.

Many questions will contain two parts. - Part A will test your coding knowledge. - Part B will ask you to enter written answers into the Markdown (white space) - Please note that some questions only have Part A.

##QUESTION 1 *10pts* Part A: Enter code to produce the structure of your dataframe Part B: 1. Indicate the number of rows and columns: Rows = Columns =
2. Which variables would be appropriate for examination by histogram? 3. Why are these variables appropriate for histograms?

```
str(MyData)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ Sex            : int  0 1 1 0 1 1 1 1 0 0 ...
##  $ Age            : int  31 31 75 80 64 14 78 35 70 18 ...
##  $ Married        : int  0 0 0 0 1 0 1 0 1 0 ...
##  $ Income         : num  60 0.36 0 0 0 ...
##  $ HoursWk        : num  40 12 40 13.2 32.7 ...
##  $ Race           : chr  "white" "black" "white" "white" ...
##  $ USCitizen      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HealthInsurance: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Language       : int  1 0 0 0 0 0 0 1 0 0 ...
```

1. Rows=1000 Columns=9
2. Age, Income, HoursWk
3. These are quantitative variables that possess discrete data that can be repeated.

##QUESTION 2 *10pts* Part A: Generate descriptive statistics (fav_stats) for one of the variables you listed in Question 1:PartB2, and save them to a new object. Name your object using the format 'VariableName.stats'

Part B: Provide the min, mean, median, max, and IQR for this variable.

```
#Your code goes here!
#1) Generate descriptive stats for your categorical variable from question 1.

HoursWk.stats <- favstats(~ HoursWk, data = MyData)

#2) Print these new objects to view the output
print(HoursWk.stats)
```

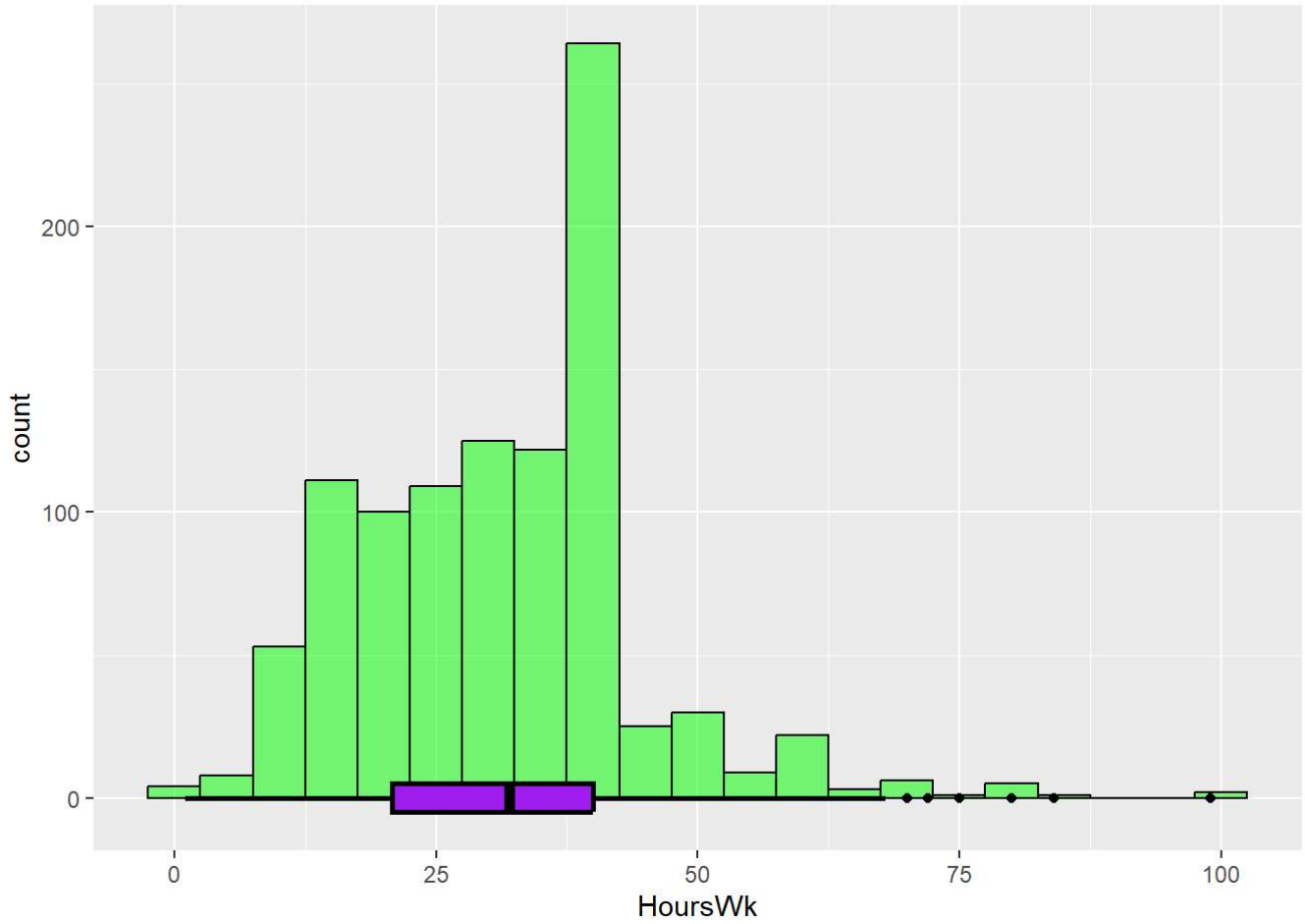
```
##  min       Q1 median Q3 max     mean       sd    n missing
##    1 20.83396     32 40  99 31.41504 13.21737 1000       0
```

min = 1, mean = 31.41504, median = 32, max = 99, IQR = Q3-Q1 = 40-20.83396 = 19.16604

##QUESTION 3 *10pts* Part A: Write code to produce a histogram of the variable you indicated in the last question.

Part B: Describe the shape of the histogram

```
gf_histogram(~ HoursWk, data = MyData, fill = "green", color = "black", linewidth = 0.5, binwidt
h = 5) %>%
  gf_boxplot(fill = "purple", width = 10, color = "black", linewidth = 1)
```



This is a histogram of the HoursWk variable. The mode of this data is understandably 40, and the median and the mean are around 31 to 32. The distribution is Left-skewed, and looks to be uni-modal.

##QUESTION 4 *10pts* Part A: Recode Sex into new variable called “Gender” where 1 = Male and 0 = Female.

```
MyData$Gender = recode(MyData$Sex,"0" = "Female","1" = "Male")
```

##QUESTION 5 *5pts* Part A: Check that your code worked by viewing the first and last six rows of your new variable

```
head(MyData$Gender)
```

```
## [1] "Female" "Male"   "Male"   "Female" "Male"   "Male"
```

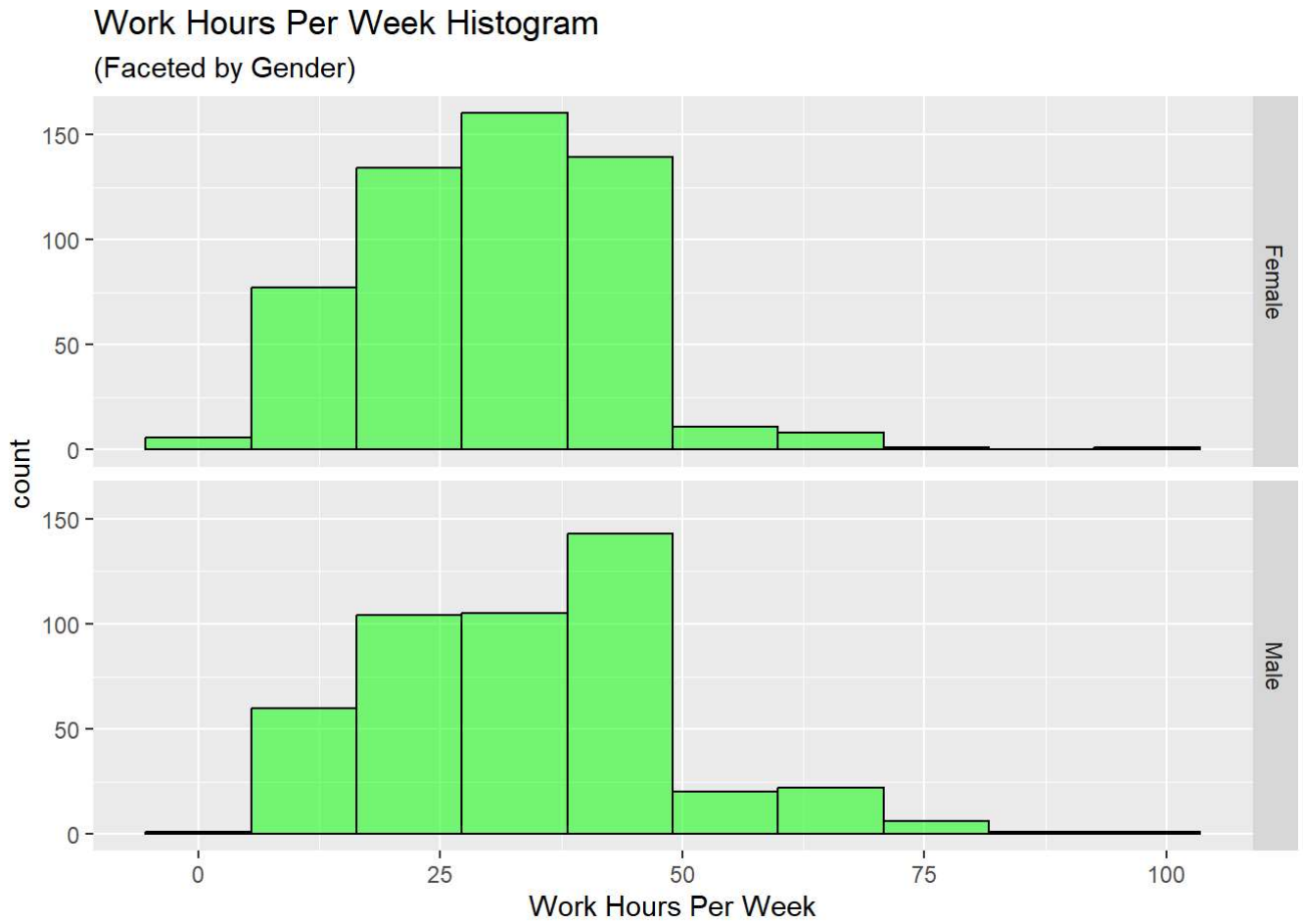
```
tail(MyData$Gender)
```

```
## [1] "Female" "Female" "Female" "Male"   "Female" "Female"
```

##QUESTION 6: *20pts* Part A: 1. Rewrite your histogram from Question 3 and this time fill it in using the 'Gender' variable from question 4. (Note: This will allow you to differentiate between male and female instances.) 2. Add 10 bins 3. Label the title and x-axis of the histogram using gf_labs=title"", x="" 4. separate the histogram by gender using gf_facet_grid(variable ~.)

Part B: What can you determine from the new histogram output?

```
gf_histogram(~ HoursWk, data = MyData, fill = "green", color = "black", linewidth = 0.5, bins =
10, title = "Work Hours Per Week Histogram", subtitle = "(Faceted by Gender)", xlab = "Work Hour
s Per Week") %>%
  gf_facet_grid(Gender ~ .)
```



Males work over 40 hours a little more often than females. Females work under 40 hours a little more often than males.

##QUESTION 7: *15pts* Part A: Create a new categorical variable called Age3Cat from the Age variable. Part B: 1.How many instances of each category are there in your new variable? (*hint: tally) 2. What is the proportion of each category?

```
favstats(MyData$Age)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0	20	41	58	94	40.071	22.99562	1000	0

1 row

```
MyData$Age3Cat <- cut(MyData$Age, breaks = c(-Inf, 12, 20, 35, 50, 65, Inf), labels = c("Childho
od", "Adolescence", "Early Adult", "Middle Adult", "Late Adult", "Eldership"))


#Check your work
tally(~ Age3Cat, data = MyData)
```

```
## Age3Cat
##    Childhood  Adolescence  Early Adult Middle Adult   Late Adult    Eldership
##          154          100          175          221          196          154
```

```
tally(~ Age3Cat, data = MyData, format = "proportion")
```

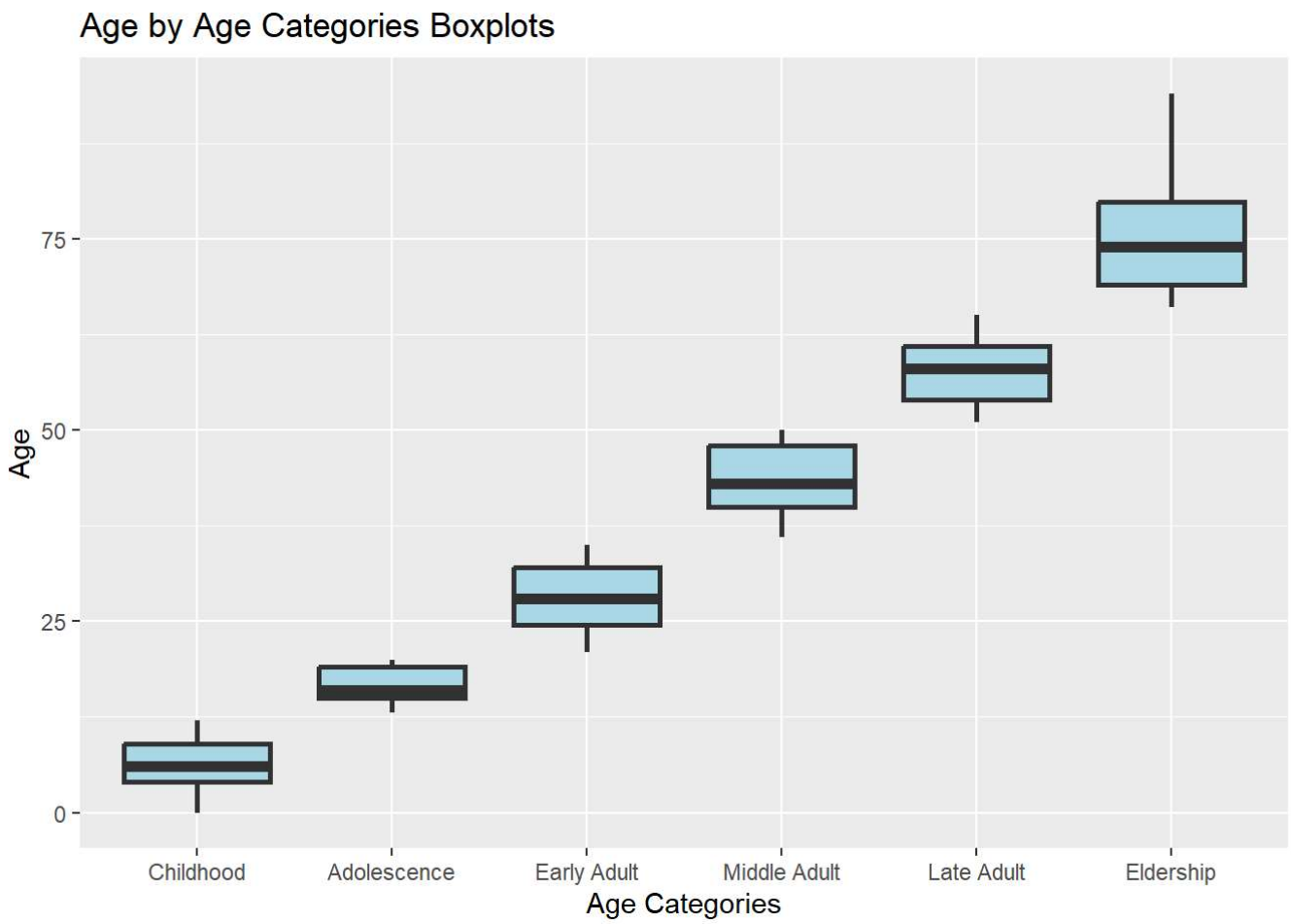
```
## Age3Cat
##    Childhood  Adolescence  Early Adult Middle Adult   Late Adult    Eldership
##        0.154        0.100        0.175        0.221        0.196        0.154
```

Childhood 154, Adolescence 100, Early Adult 175, Middle Adult 221, Late Adult 196, Eldership 154

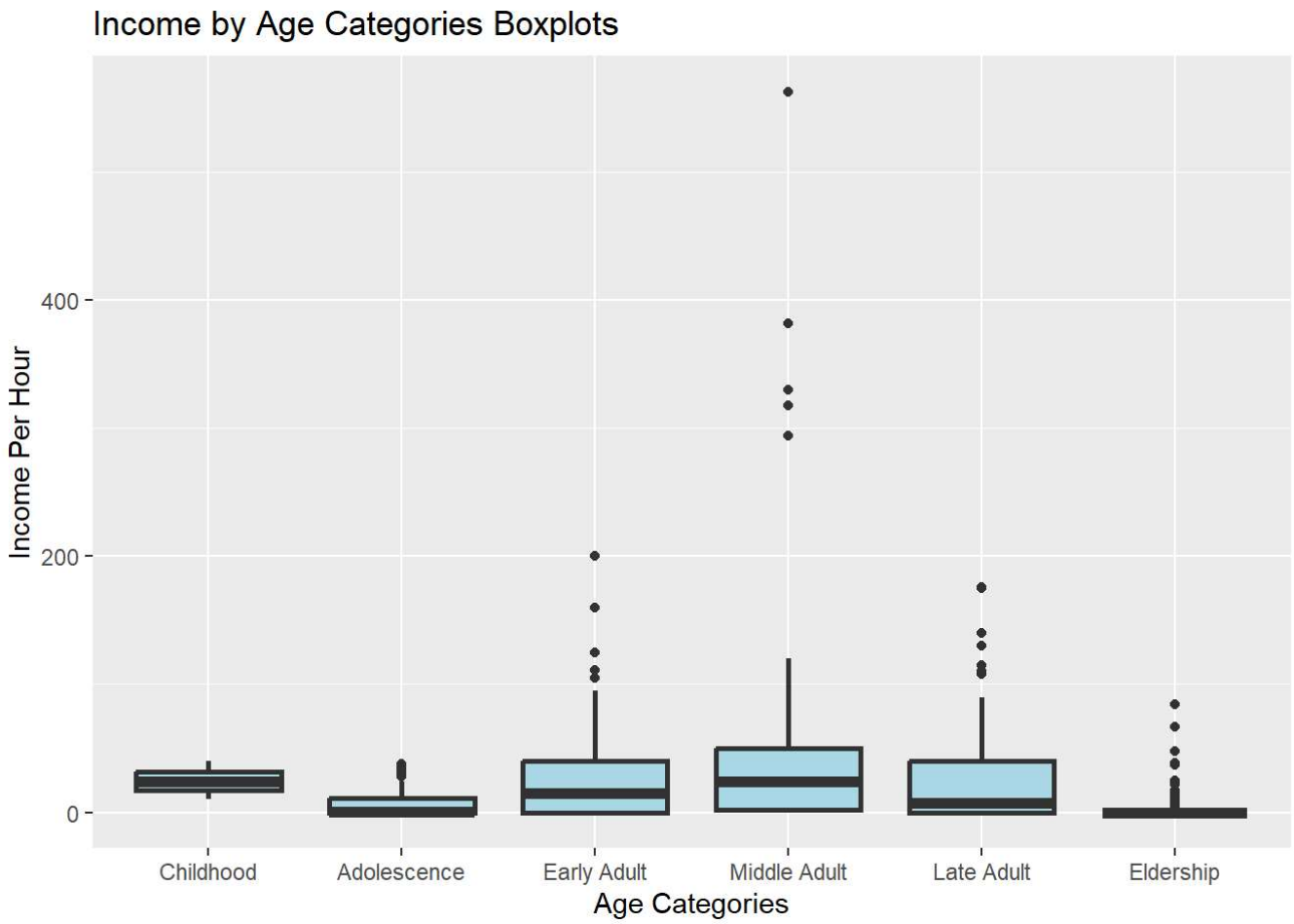
##QUESTION 8: *20pts* Part A: Evaluate Income and Age by your Age3Cat variable. 1. Create a box plot of Age by Age3Cat. 2. Create a box plot of Income by Age3Cat.

Part B: 1. Which group appears to have the greatest spread for age? 2. Which group appears to have the most outliers for income? (Estimate the max income based on the chart.) _____

```
gf_boxplot(Age ~ Age3Cat, data = MyData, fill = "lightblue", border = "darkblue", outpch = 21, o
utbg = "darkblue", linewidth = 1, title = "Age by Age Categories Boxplots", xlab = "Age Categori
es", ylab = "Age")
```




```
gf_boxplot(Income ~ Age3Cat, data = MyData, fill = "lightblue", border = "darkblue", outpch = 2
1, outbg = "darkblue", linewidth = 1, title = "Income by Age Categories Boxplots", xlab = "Age C
ategories", ylab = "Income Per Hour")
```

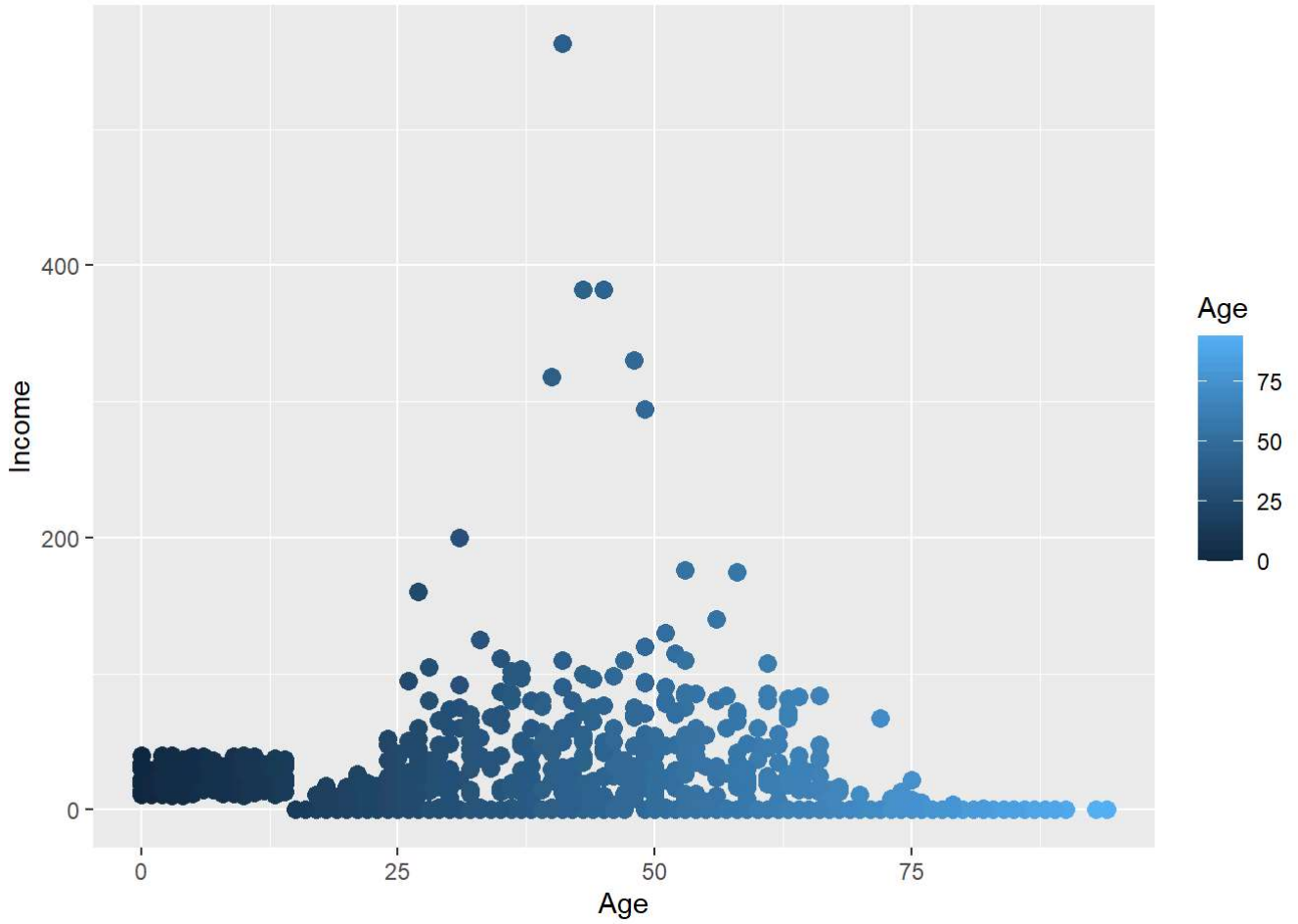


Eldership appears to have the greatest spread in their age category. Eldership has the most outliers for their income category. The maximum income per hour is approximately $575.

##QUESTION 9: *15pts* Part A: Your boss has asked you to investigate whether there is an association between age and income among employees. Make a chart that displays how much employees make (Income) based on their age. Visualize this using a scatterplot - Include these arguments: size = 3, color=~Age

Part B: Do you notice anything?

```
gf_point(Income ~ Age , data = MyData, size = 3, color = ~ Age)
```



There is cube-like cluster in bottom left corner of the scatterplot.

##QUESTION 10:*20pts* There's some data that doesn't seem to make sense. It looks like a cube on the left side of your scatterplot! We want to remove that so we can visualize Income by Age the correct way.

Part A: Filtering exercise. Step 1. Create a new data frame object called MyData_Filtered AND Step 2: Using the filter function, capture into your new dataframe, only the rows in which Income > 10 AND Income < 200 AND Age > 17

Part B: How many rows remain in your filtered dataframe?

Extra Credit (5pts) If you can perform complete this using one filter function!

```
MyData_Filtered <- filter(MyData, Income > 10, Income < 200, Age > 17)
tally(MyData_Filtered)
```

n
<int>
352

1 row

352

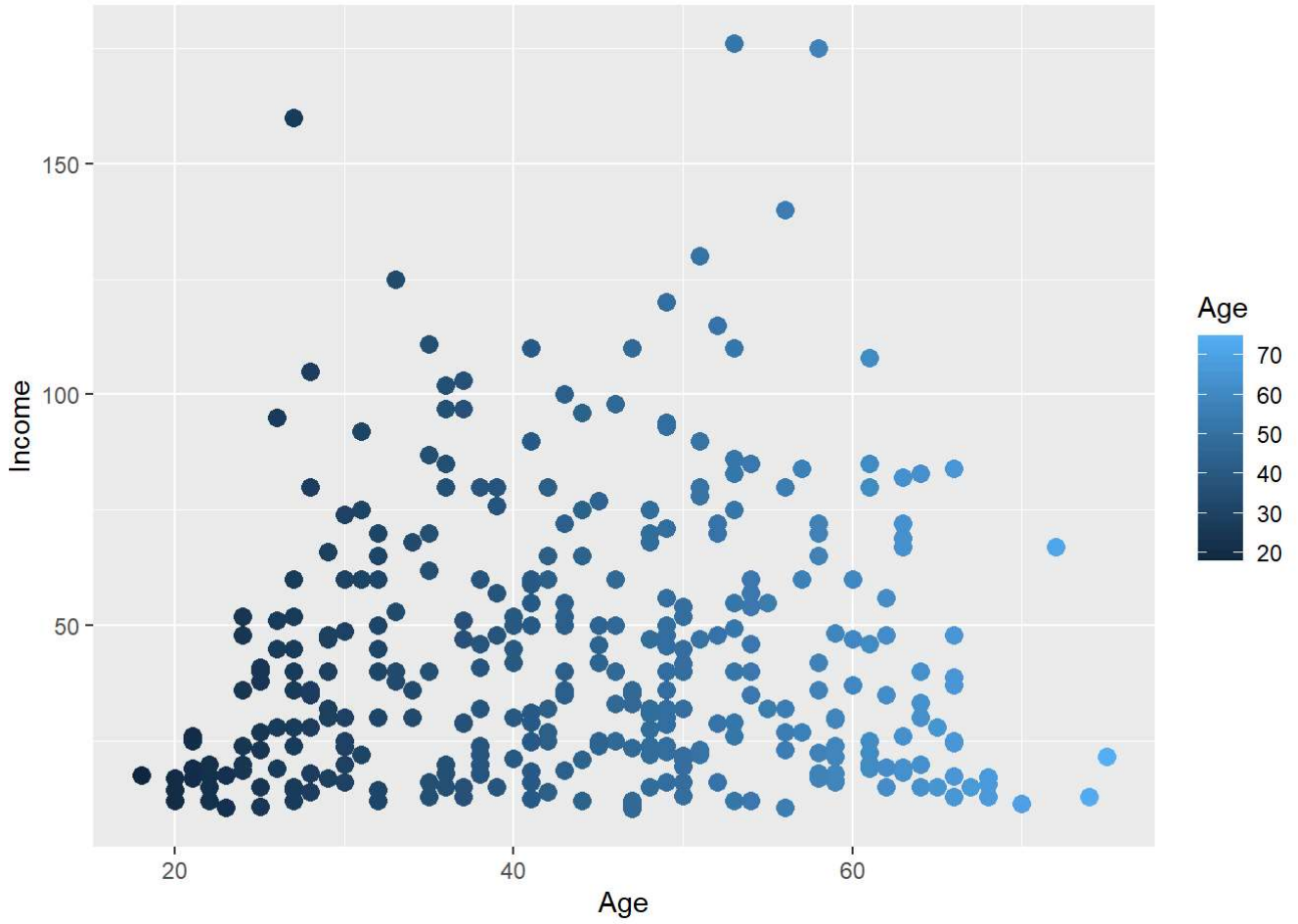
##QUESTION 11: *10pts*

Part A: 1. Let's try our scatterplot visualization from Question 9 again. - Remember to include these arguments: size = 3, color=~Age

Part B: What do you notice from our new scatterplot?

Extra Credit: See if you can add a regression line through the scatter plot

```
gf_point(Income ~ Age , data = MyData_Filtered, size = 3, color = ~ Age)
```



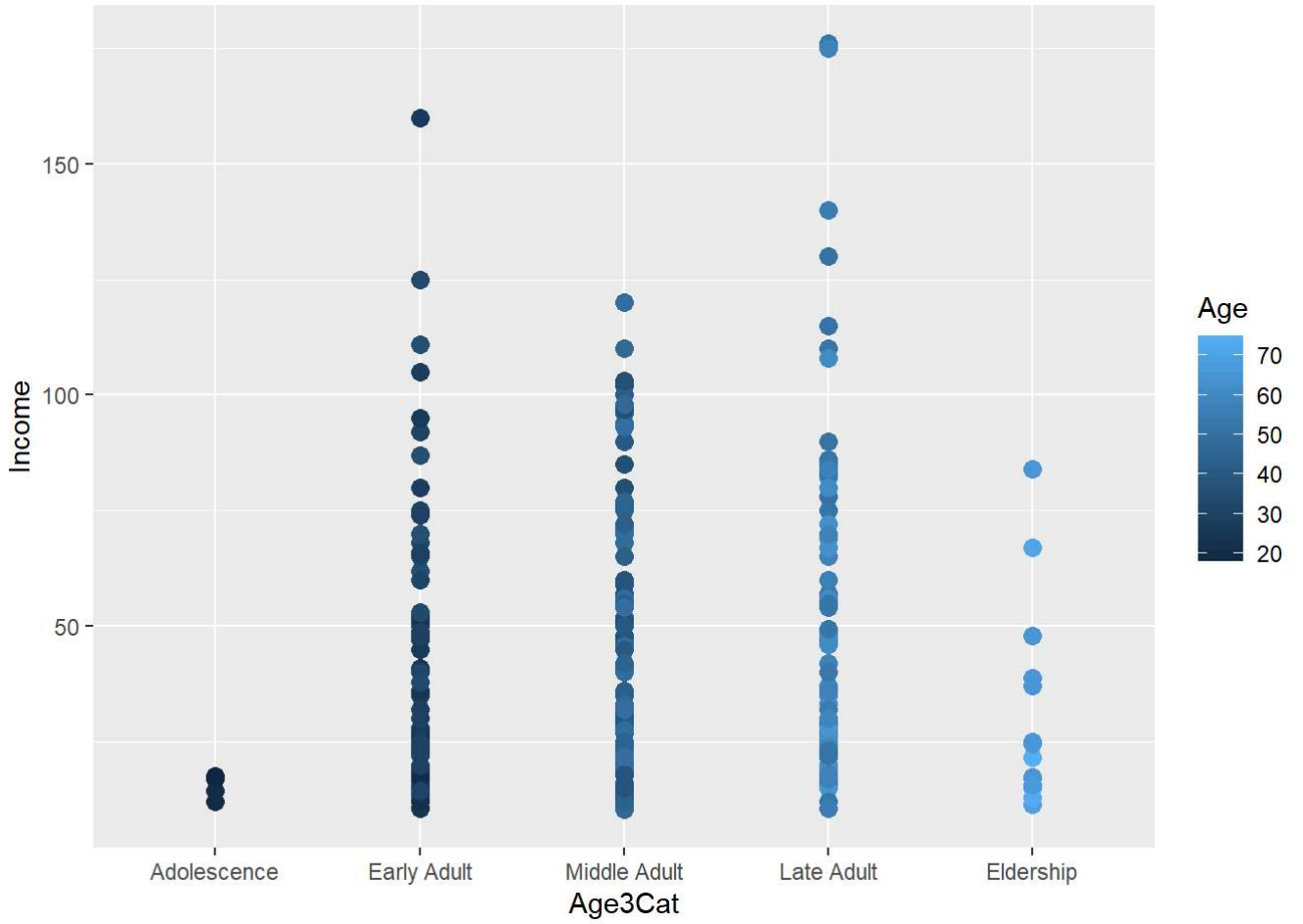
The cube-like cluster in bottom left corner of the scatterplot was removed. Under 17, top, and bottom income earners were removed as well.

##QUESTION 12: *15pts* Part A: Create another scatter plot, but this time using Income by Age3Cat. - Remember to include these arguments: size = 3, color=~Age

Part B: Based on your investigation, does there appear to be an association between age and income? Why or why not?

Part C: Because there is an association, does this mean that Age is the cause of higher income? Why or why not?

```
gf_point(Income ~ Age3Cat , data = MyData_Filtered, size = 3, color = ~ Age)
```



There appears to be a relationship between age and income. The scatter plot clearly shows the Early, Middle, and Late Adults categories will have higher incomes than Adolescence and Eldership categories. Did not apply a linear regression model, because this would probably merit a curvelinear model.

Age with the knowledge and experience that comes with it would likely benefit an individual's income up to a point, but the difficulties often associated with Eldership would eventually deprive an individual of earning potential.

End of Midterm.