Week1 DSC630

Moshe Burnstein

2023-03-16

Summary of data and questions to explore

The Adult dataset originated in the US Census Bureau and displays multiple demographics with a target binary variable of making more than $50k a year, from the 1994 Census (Kohavi, 1996).The continuous features include age of a subject, final weight which is the base weight that the CPS assigns to a subject to represent the number of subjects that the sample represents in the population, capital gains and losses, and hours per week worked. Categorical features include type of work, level of education, years of education, marital status, type of job, role in household, race, sex, and native country. This dataframe contains 32,561 observations and 15 columns and is housed in the UCI Machine Learning Repository. It is a font of information about our population. I want to compare males to females relative to years of education. Do males spend more years educating themselves? Which race suffers the most capital gains losses? Is the median age of females greater than males? Females have traditionally shown longer lifespans than males (World Data, n.d.). Is one race in America heartier than any other? Or do lifespans in this country congregate around similar numbers? Does any race work more hours than others? How does age affect this?

Load Adult dataset

```
adult_df <- read.csv('adult.data', header = FALSE)
```

Check that loaded properly

```
head(adult_df)
```

```
##   V1                V2     V3         V4 V5                  V6
## 1 39         State-gov  77516  Bachelors 13       Never-married
## 2 50  Self-emp-not-inc  83311  Bachelors 13  Married-civ-spouse
## 3 38           Private 215646    HS-grad  9            Divorced
## 4 53           Private 234721       11th  7  Married-civ-spouse
## 5 28           Private 338409  Bachelors 13  Married-civ-spouse
## 6 37           Private 284582    Masters 14  Married-civ-spouse
##                   V7             V8     V9     V10  V11 V12 V13            V14
## 1       Adm-clerical  Not-in-family  White    Male 2174   0  40  United-States
## 2    Exec-managerial        Husband  White    Male    0   0  13  United-States
## 3  Handlers-cleaners  Not-in-family  White    Male    0   0  40  United-States
## 4  Handlers-cleaners        Husband  Black    Male    0   0  40  United-States
## 5     Prof-specialty           Wife  Black  Female    0   0  40           Cuba
## 6    Exec-managerial           Wife  White  Female    0   0  40  United-States
##      V15
## 1  <=50K
## 2  <=50K
## 3  <=50K
## 4  <=50K
## 5  <=50K
## 6  <=50K
```

Check for null values

```
which(is.na(adult_df))
```

```
## integer(0)
```

```
sum(is.na(adult_df))
```

```
## [1] 0
```

Add column names

```
colnames(adult_df) <- c('Age', 'WorkClass', 'Fnlwgt', 'Education', 'Education_Num', 'Marital_Status',
                        'Occupation', 'Relationship', 'Race', 'Gender', 'Capital_Gain',
                        'Capital_Loss', 'Hours_per_Week',
                        'Native_Country', 'Earned_more_than_50k')

head(adult_df)
```

```
##   Age         WorkClass Fnlwgt  Education Education_Num      Marital_Status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
## 4  53           Private 234721       11th             7  Married-civ-spouse
## 5  28           Private 338409  Bachelors            13  Married-civ-spouse
## 6  37           Private 284582    Masters            14  Married-civ-spouse
##           Occupation   Relationship   Race  Gender Capital_Gain Capital_Loss
## 1       Adm-clerical  Not-in-family  White    Male         2174            0
## 2    Exec-managerial        Husband  White    Male            0            0
## 3  Handlers-cleaners  Not-in-family  White    Male            0            0
## 4  Handlers-cleaners        Husband  Black    Male            0            0
## 5     Prof-specialty           Wife  Black  Female            0            0
## 6    Exec-managerial           Wife  White  Female            0            0
##   Hours_per_Week Native_Country Earned_more_than_50k
## 1             40  United-States                <=50K
## 2             13  United-States                <=50K
## 3             40  United-States                <=50K
## 4             40  United-States                <=50K
## 5             40           Cuba                <=50K
## 6             40  United-States                <=50K
```

Check number of unique values in education columns

```
unique(adult_df$Education_Num)
```

```
##  [1] 13  9  7 14  5 10 12 11  4 16 15  3  6  2  1  8
```

```
unique(adult_df$Education)
```

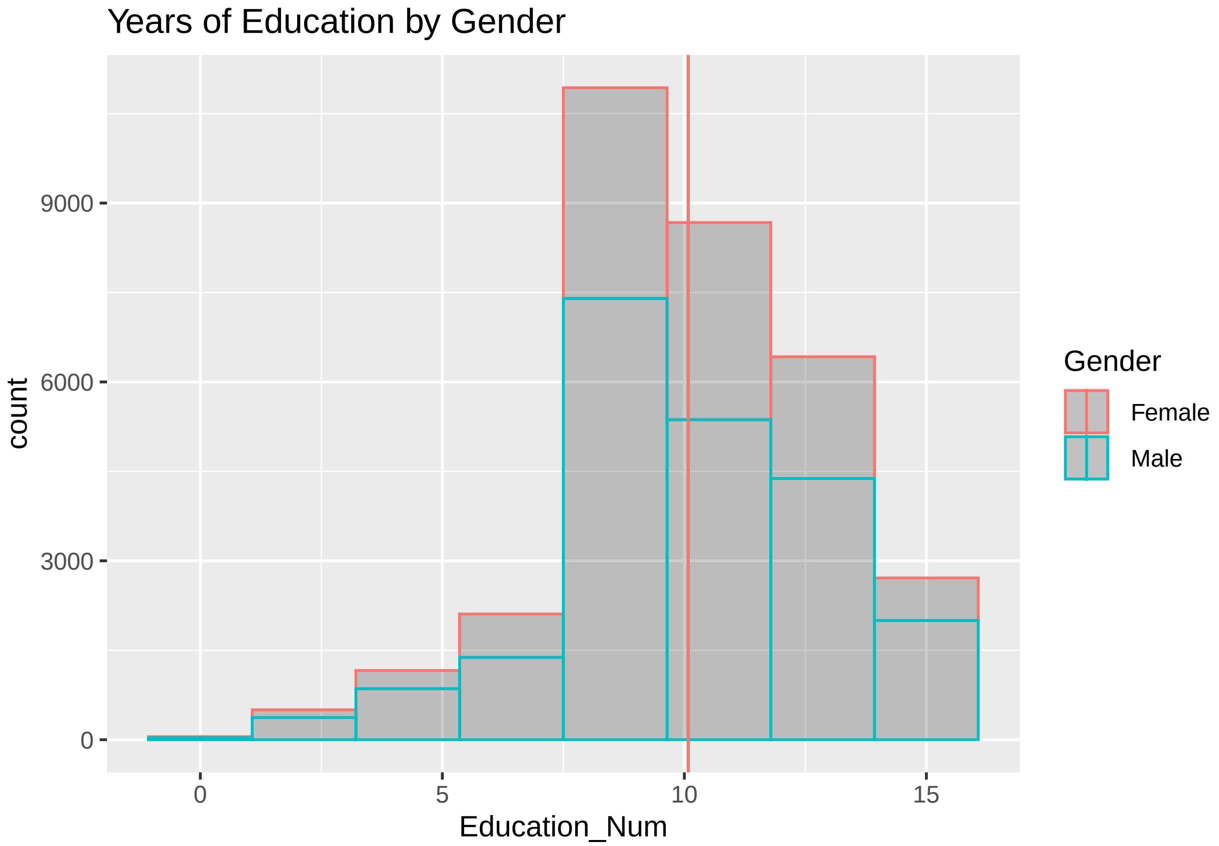
```
##  [1] " Bachelors"    " HS-grad"      " 11th"         " Masters"
##  [5] " 9th"          " Some-college" " Assoc-acdm"   " Assoc-voc"
##  [9] " 7th-8th"      " Doctorate"    " Prof-school"  " 5th-6th"
## [13] " 10th"         " 1st-4th"      " Preschool"    " 12th"
```

Create histogram to display distribution of education years, by gender

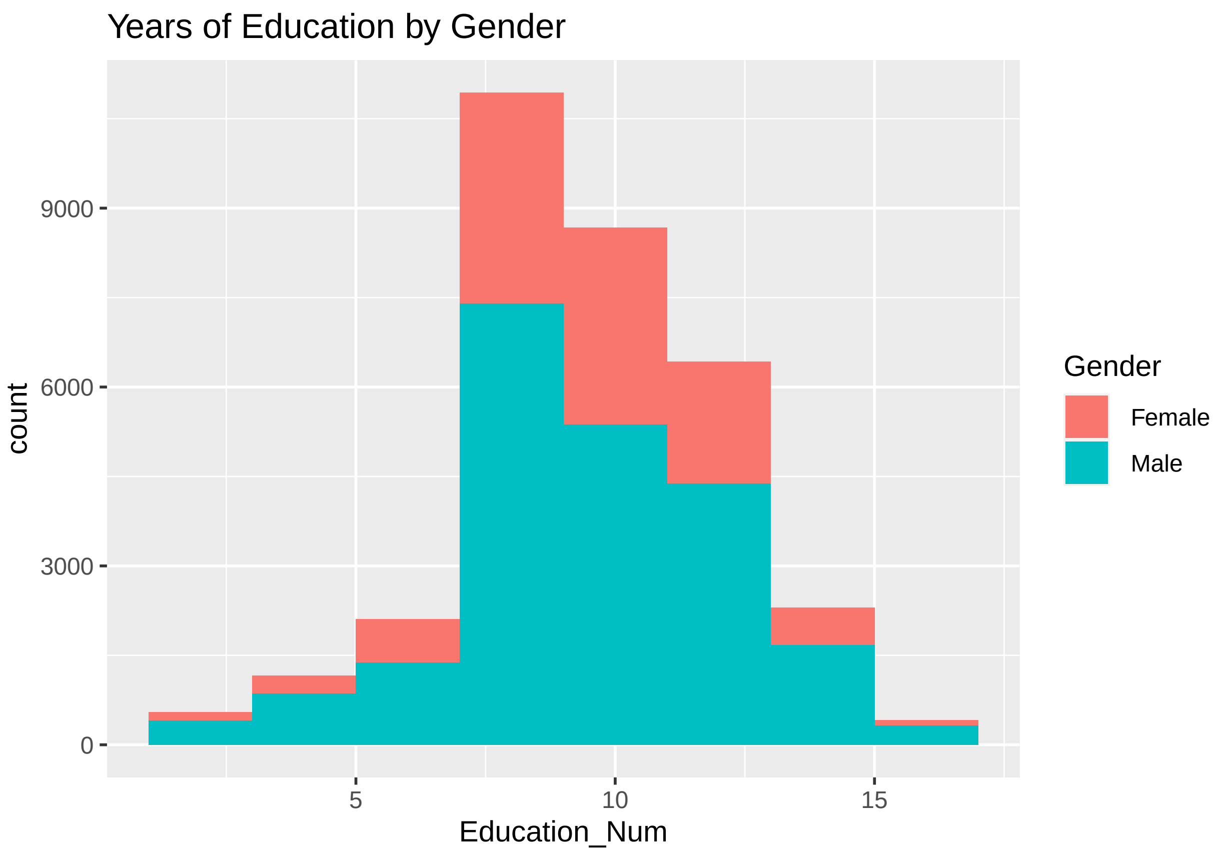
```
library(ggplot2)
library(plyr)
```

```
## Warning: package ’plyr’ was built under R version 4.2.1
```

```
hist_education <- ggplot(adult_df, aes(x=Education_Num, color=Gender)) +
  geom_histogram(fill='black', alpha=0.2, bins = 8) + ggtitle('Years of Education by Gender') +
  geom_vline(data = adult_df, aes(xintercept=mean(Education_Num), color=Gender), linetype=1)
hist_education
```

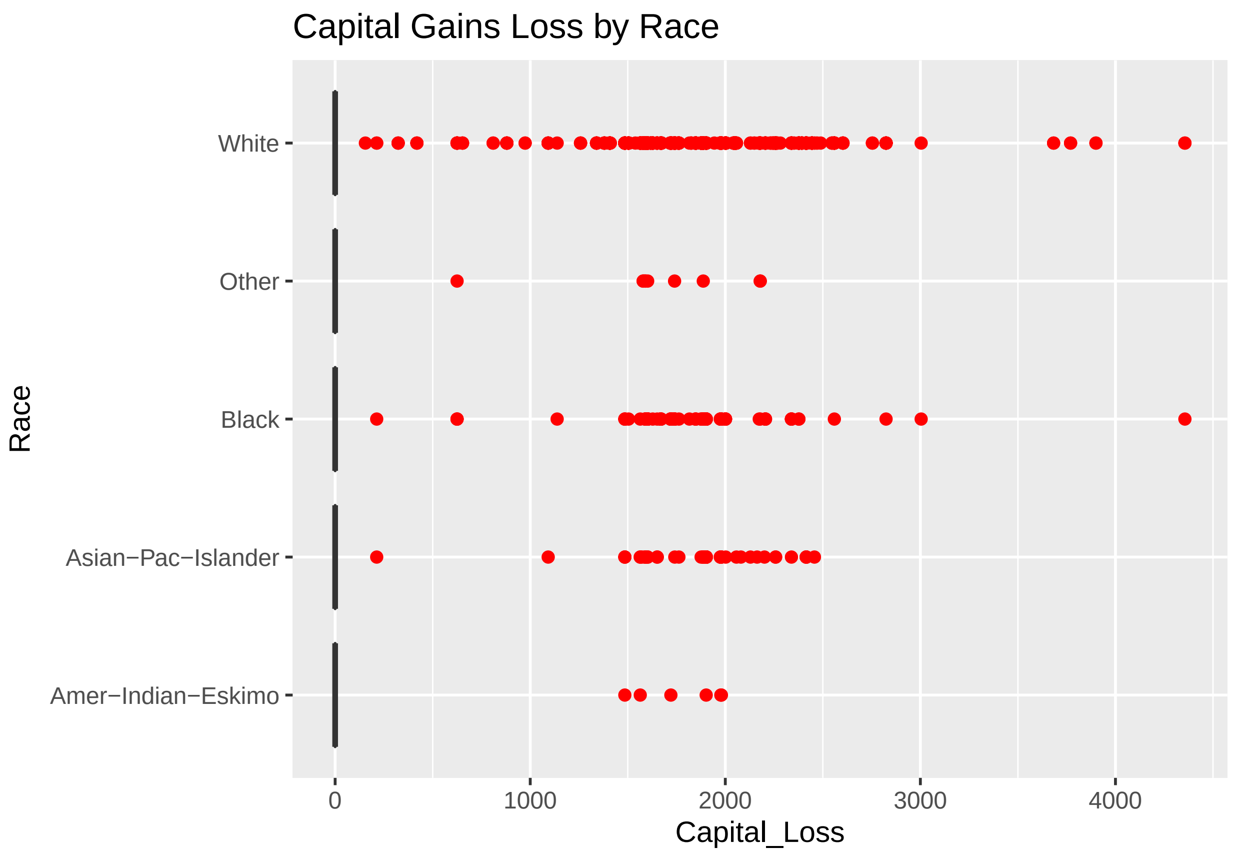


```
histo_education <- ggplot(adult_df, aes(Education_Num, fill=Gender)) + geom_histogram(binwidth = 2)+ gg
histo_education
```



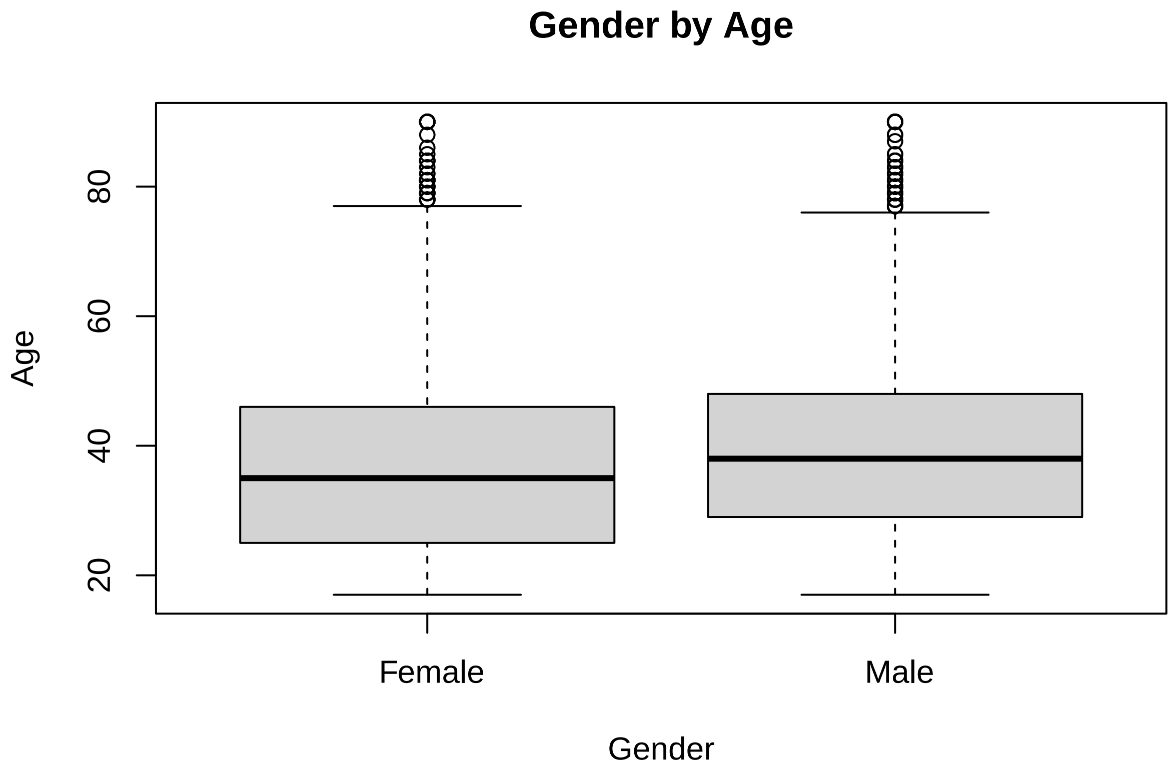
Create Boxplots

```
capital_loss_by_race <- ggplot(adult_df, mapping=aes(x=Race, y=Capital_Loss, colors(distinct = FALSE)))
  geom_boxplot(outlier.colour = 'red', outlier.shape = 16, outlier.size = 2) +
  coord_flip() + ggtitle('Capital Gains Loss by Race')
capital_loss_by_race
```

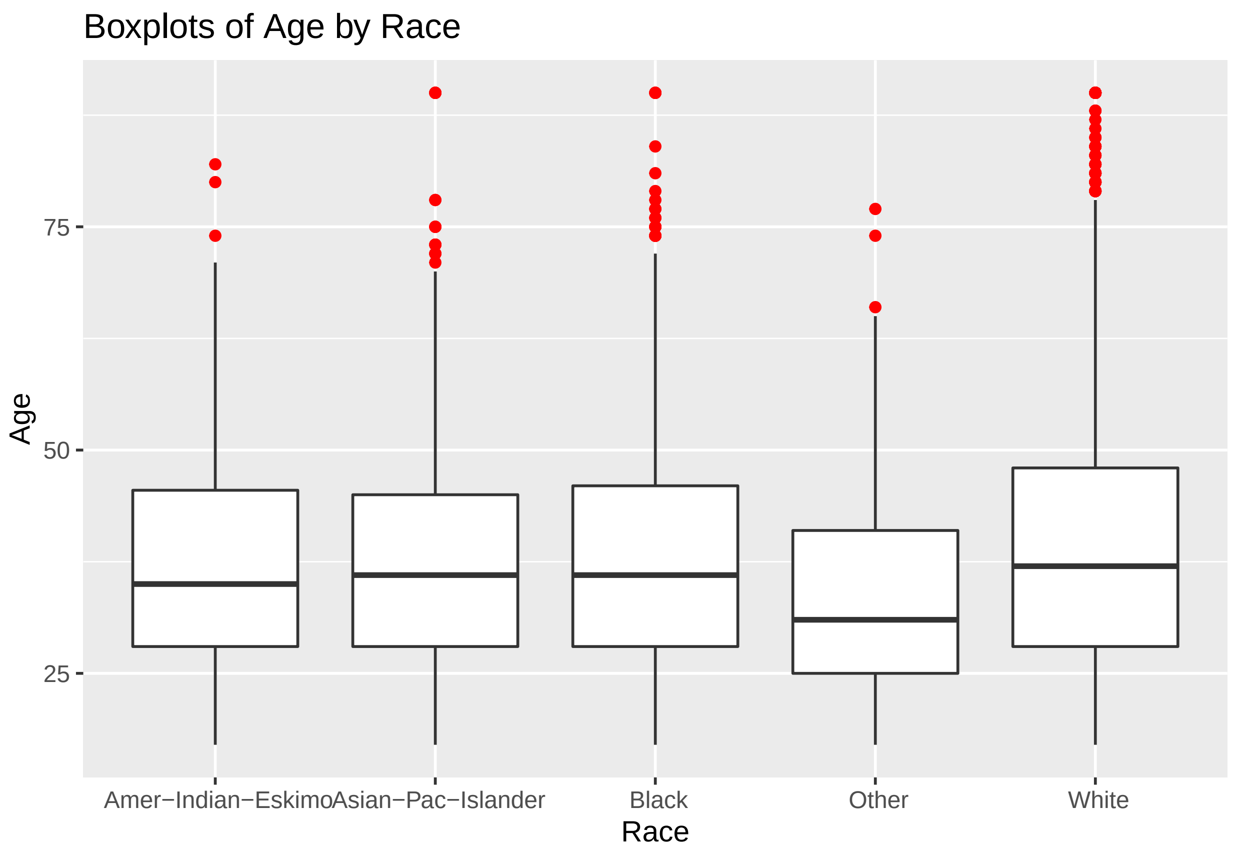


Note how no box appears. This is because the middle 50% of the data is at 0.

```
boxplot(Age ~ Gender, data = adult_df, main='Gender by Age')
```



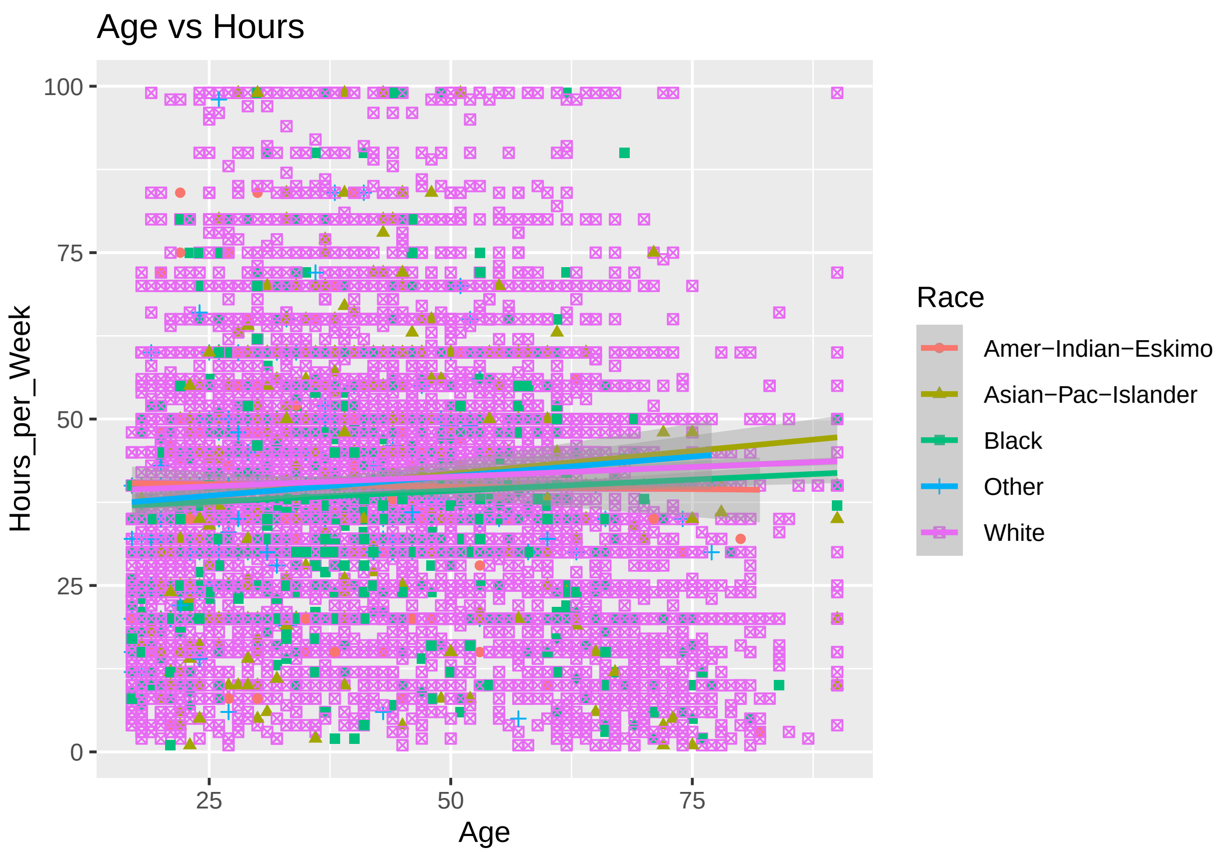
```
age_by_race <- ggplot(adult_df, aes(Race, Age)) + geom_boxplot(outlier.colour = 'red') + labs(title = 'I
age_by_race
```



Plot age vs hours per week worked, by race

```
scatter_age_hours <- ggplot(adult_df, aes(x=Age, y=Hours_per_Week, shape=Race,
                                          color=Race)) + geom_point() + stat_smooth(method = lm) + labs
scatter_age_hours
```

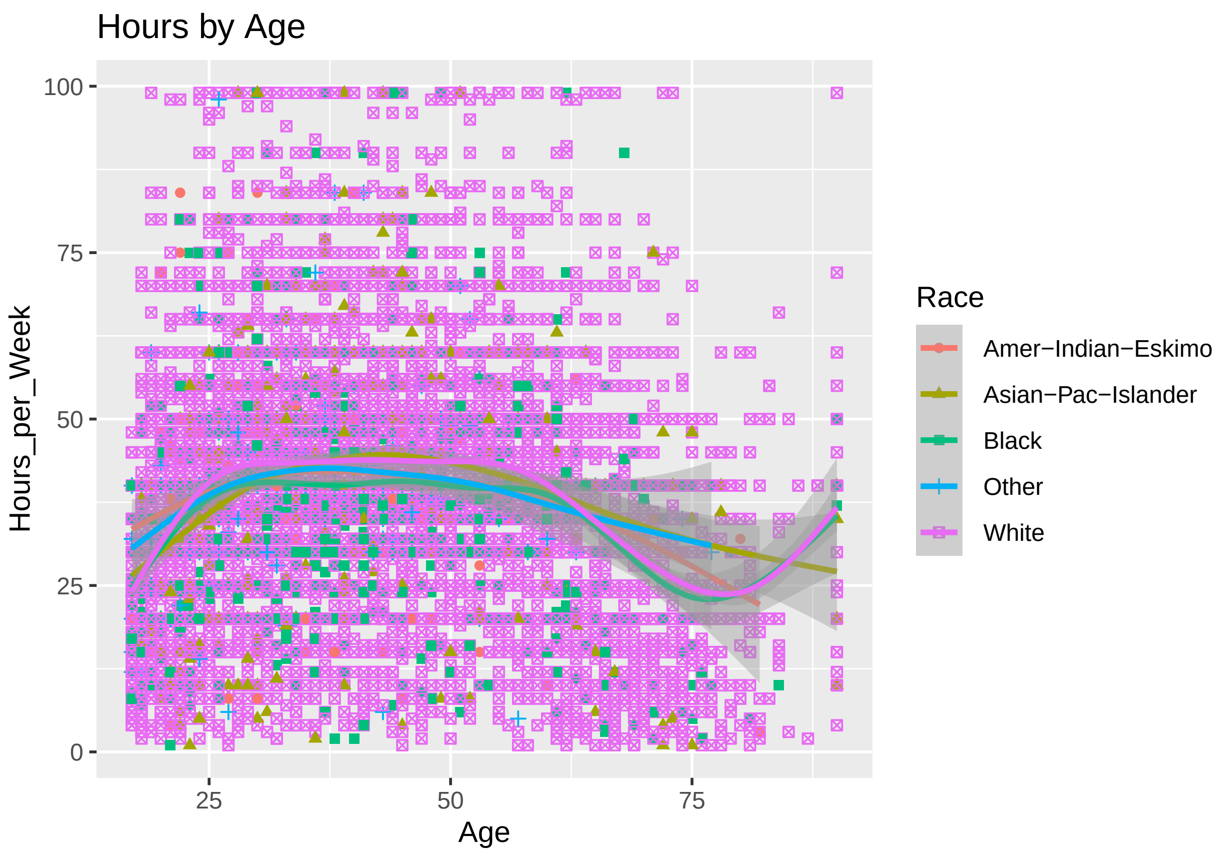
```
## `geom_smooth()` using formula 'y ~ x'
```



Compare smoothing methods

```
scatter_age_hours_smooth <- ggplot(adult_df, aes(x=Age, y=Hours_per_Week, shape=Race,
                                            color=Race)) + geom_point() + stat_smooth() + labs(title = 'H
scatter_age_hours_smooth
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Summary of results and questions answered

Females top males in years of education. Capital gains losses are all outliers because the majority of the population across all races had 0 capital gains losses. The median age of females is surprisingly less than males. The 75th percentile of age is slightly higher for females than for males, however. There is no significant difference in median lifespan or outlier longevity across race in America. Hours per week worked plateaued between 30 and 60 years of age with a significant drop at around 75. Surprisingly, whites who worked older than 75 worked longer hours.

References: Kohavi, R. (1996) Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In: KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Pages 202–207 Retrieved March 14, 2023, from https://dl.acm.org/doi/10.5555/3001460.3001502 Kohavi, R., Becker, B.(1996). UCI Machine Learning Repository [http://archive.ecs.uci.edu/ml/datasets/adult]. Irvine, CA: University of California, School of Information and Computer Science "Life expectancy for men and women" (n.d.) World Data. Retrieved March 14, 2023 from https://www.worlddata.info/life-expectancy.php