Homework#5

Jason Seda

2025-10-09

Group Members: Riyesh Nath, Jason Seda, and Bamba Cisse

Summary

In our experiment, we attempt to create a model that allows us to predict potential income_midpoint given certain parameters from household census data. For our model, we will use linear regression model.

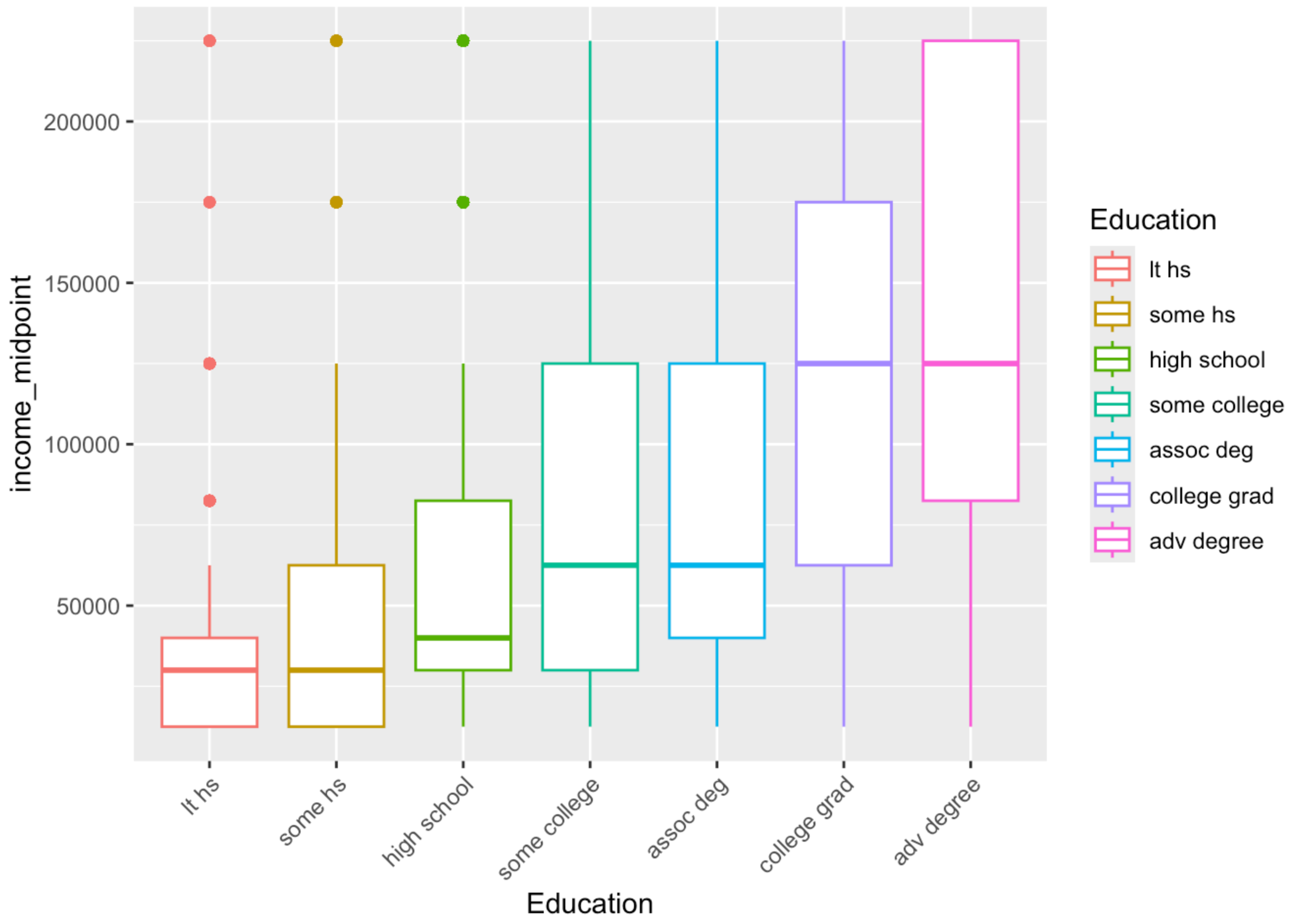
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
## Loading required package: viridisLite
```

Data Exploration

For the first step, we want to examine the parameters that might be useful for our linear regression. Lets examine by doing some data exploration. We will focus on a subcategory, using NorthEast Region

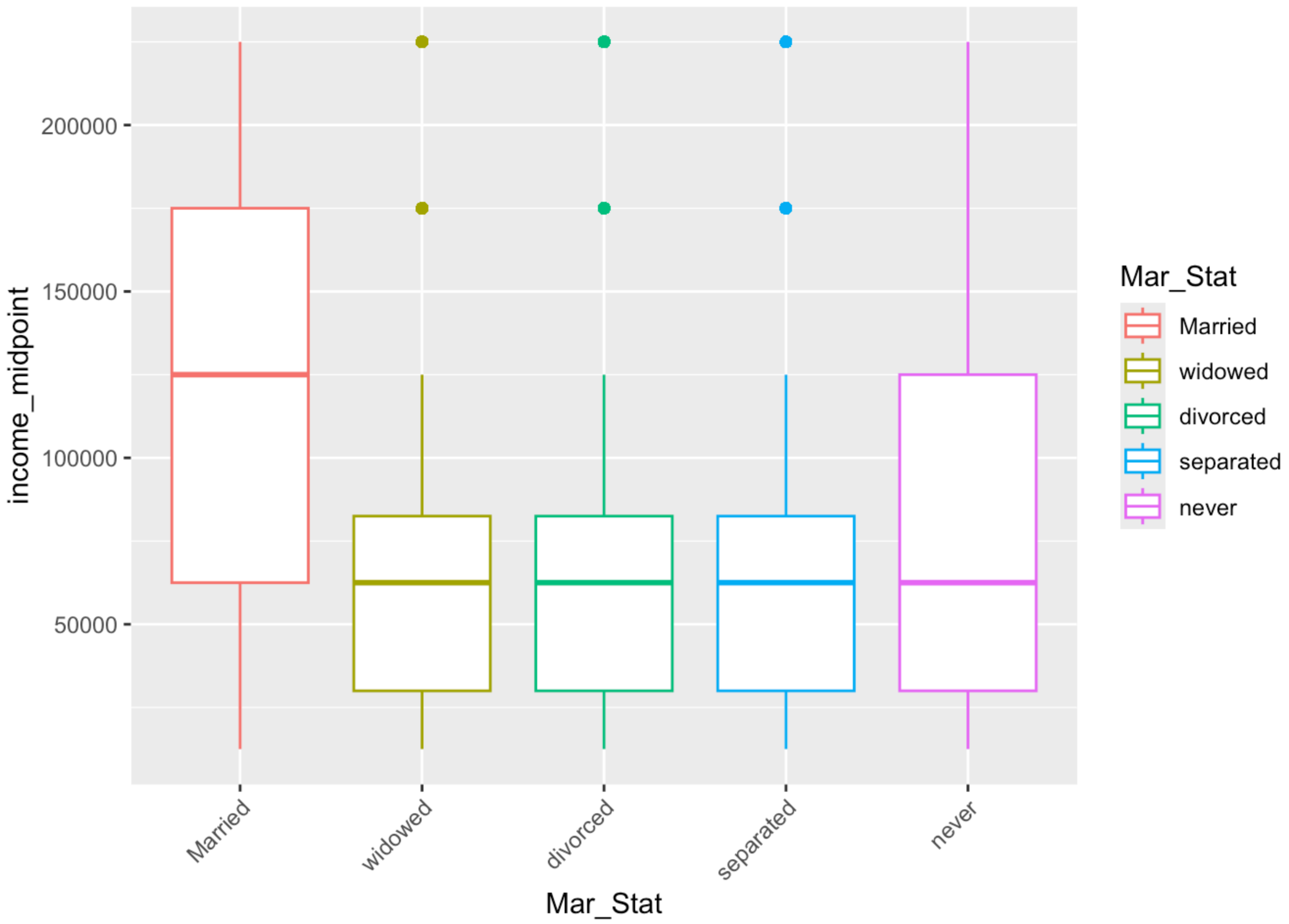
Use ANOVA for categorical data to see which one has a greater effect on income_midpoint

```
## Warning: Removed 30158 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



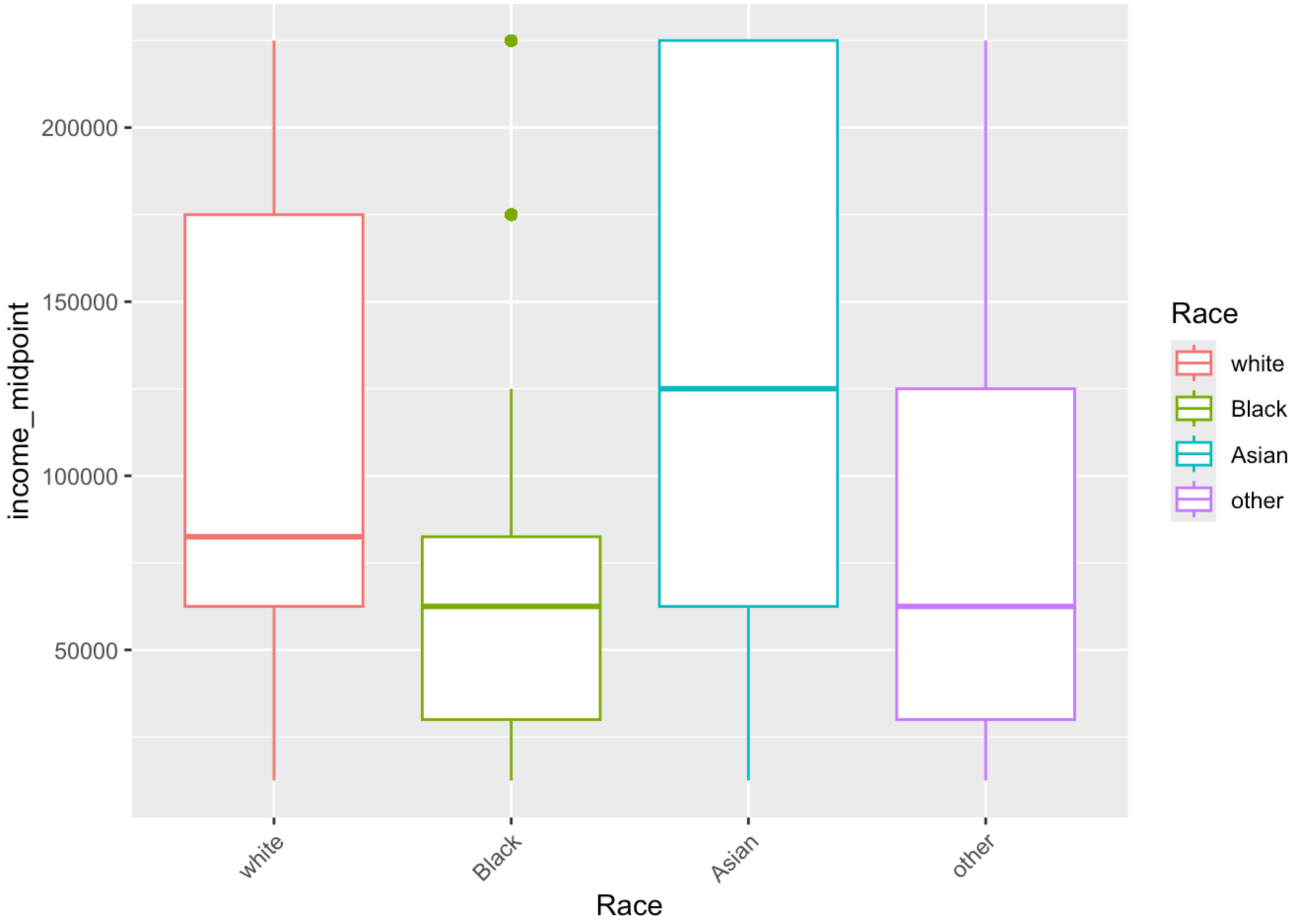
```
##                  Df    Sum Sq   Mean Sq F value Pr(>F)    
## Education         6 1.023e+14 1.706e+13    4502 <2e-16 ***
## Residuals    121389 4.600e+14 3.789e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 30158 observations deleted due to missingness
```

```
## Warning: Removed 29250 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



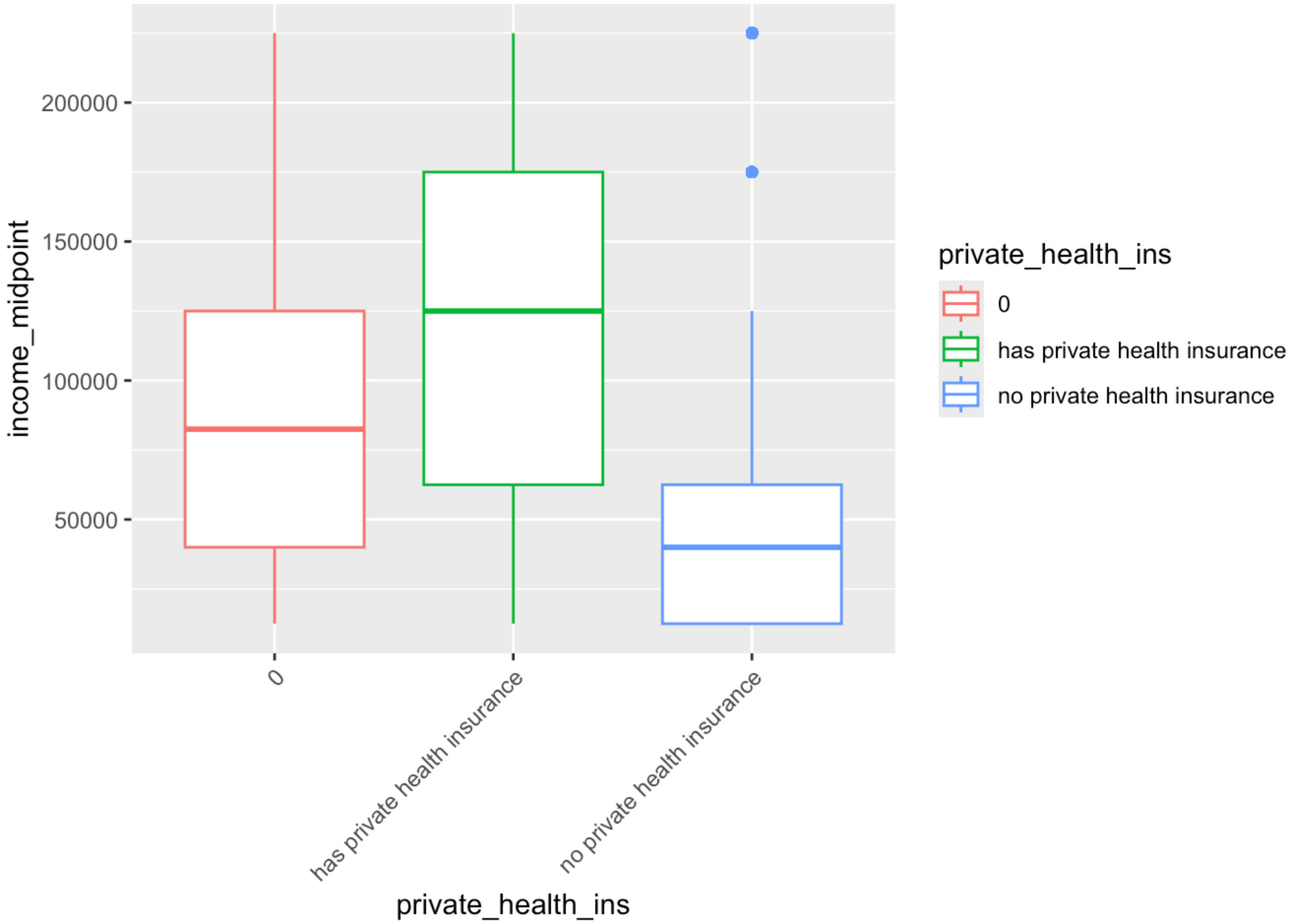
```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)
## Mar_Stat           4 9.006e+13 2.251e+13    5789 <2e-16 ***
## Residuals     120968 4.704e+14 3.889e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 29250 observations deleted due to missingness
```

```
## Warning: Removed 30158 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



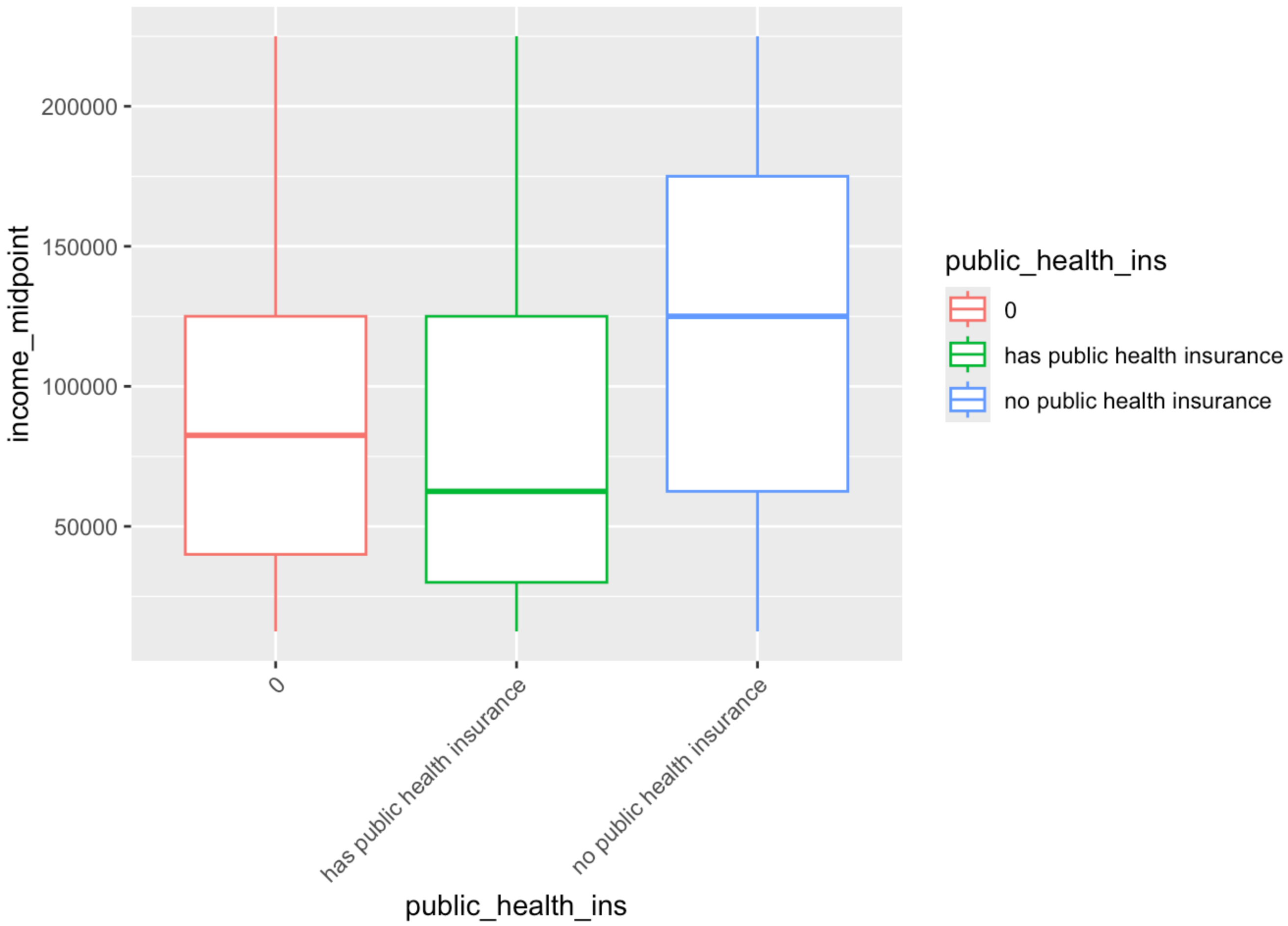
```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## Race             3 1.309e+13 4.362e+12     964 <2e-16 ***
## Residuals   121392 5.492e+14 4.525e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 30158 observations deleted due to missingness
```

```
## Warning: Removed 9611 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



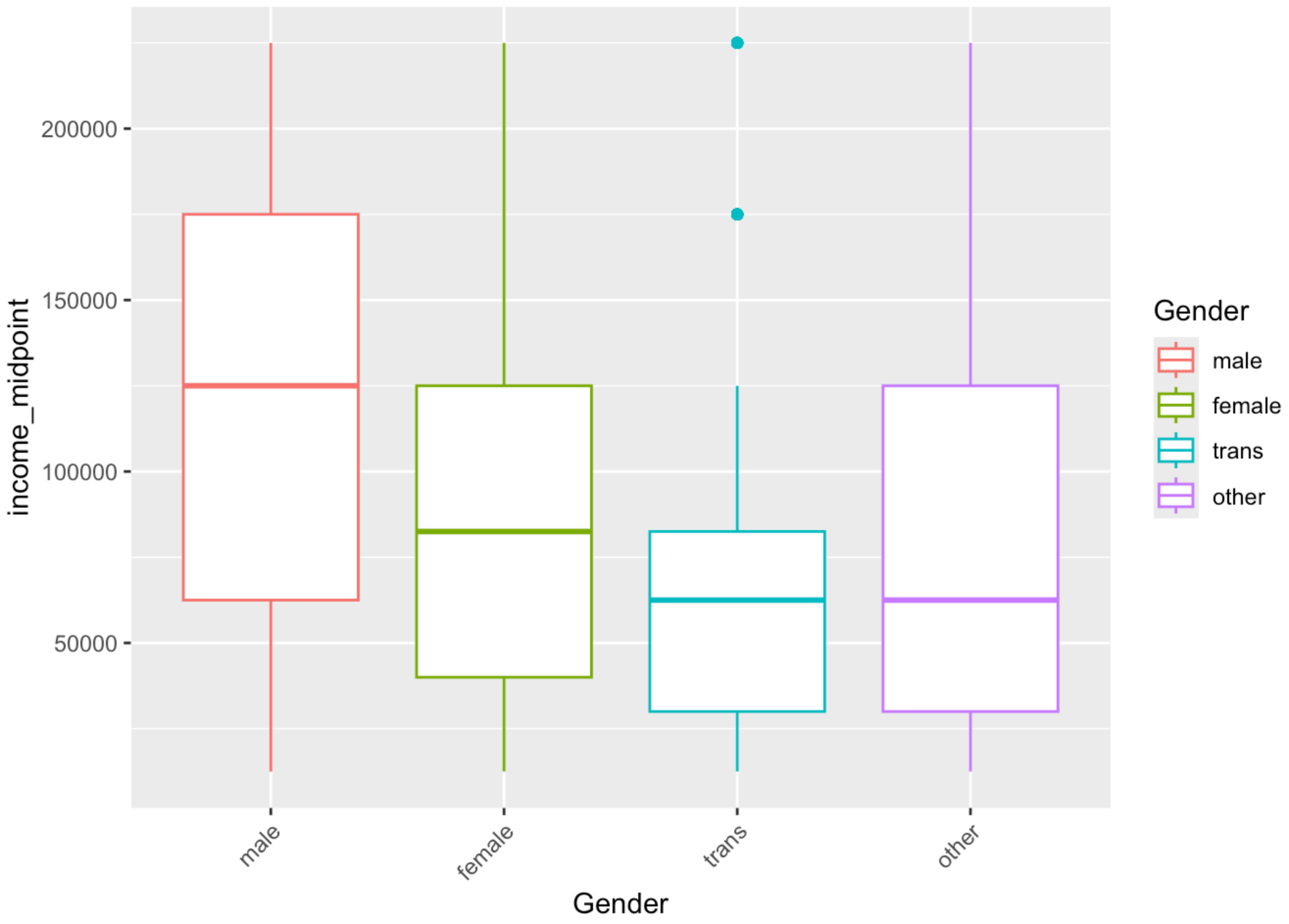
```
##                         Df    Sum Sq   Mean Sq F value Pr(>F)
## private_health_ins       2 6.156e+13 3.078e+13    7503 <2e-16 ***
## Residuals           117395 4.816e+14 4.102e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 9611 observations deleted due to missingness
```

```
## Warning: Removed 9190 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



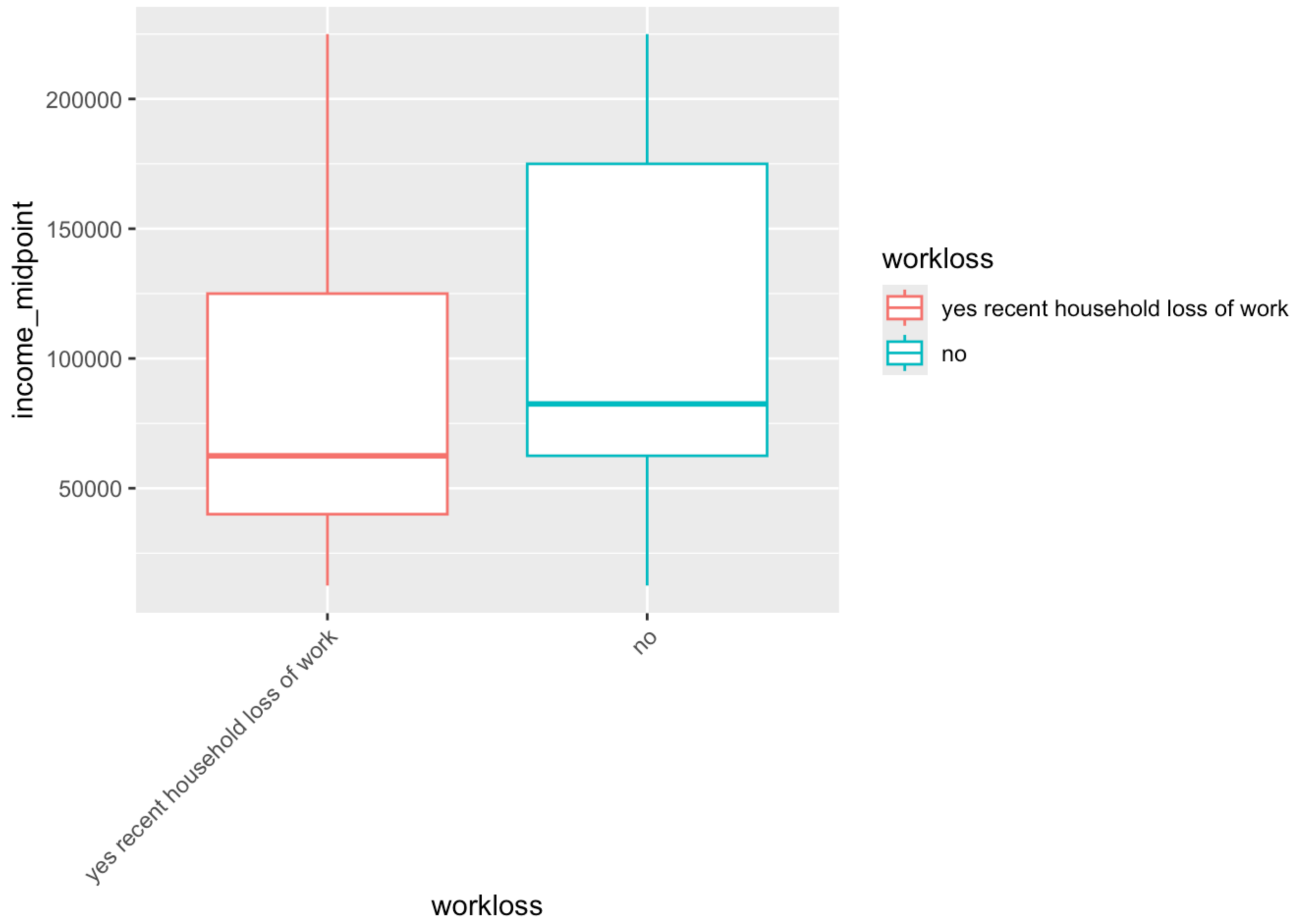
```
##                          Df    Sum Sq   Mean Sq F value Pr(>F)    
## public_health_ins         2 5.805e+13 2.902e+13    7117 <2e-16 ***
## Residuals            112526 4.589e+14 4.078e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 9190 observations deleted due to missingness
```

```
## Warning: Removed 30158 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)    
## Gender           3 1.199e+13 3.998e+12   881.8 <2e-16 ***
## Residuals   121392 5.503e+14 4.534e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 30158 observations deleted due to missingness
```

```
## Warning: Removed 27545 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
##                   Df    Sum Sq   Mean Sq F value Pr(>F)    
## workloss           1 1.016e+13 1.016e+13    2234 <2e-16 ***
## Residuals     121214 5.512e+14 4.548e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 27545 observations deleted due to missingness
```

Income Midpoint comparison vs. education:

There is a clear and consistent positive correlation between education and income. For individuals with schooling up to the associate degree level, income tends to rise gradually, showing a modest but steady upward trend. However, after completing a bachelor’s degree, there is a substantial and more pronounced increase in income levels. This sharp jump highlights a significant differentiation between those with a college education and those with only some postsecondary or associate-level attainment, emphasizing the strong economic advantage associated with higher educational attainment.

Income vs. marital status:

There is no substantial separation in income levels across most categories, except for a clear and notable increase among married individuals. This jump is largely attributed to the presence of multiple earners within the same household, which naturally raises the combined household income. Apart from this distinction, income differences among other groups remain relatively moderate and do not exhibit significant divergence.

Income vs. Race:

While, in principle, income should not differ by race or ethnicity, the data indicate a clear tendency for higher income levels among individuals identifying as Asian or White in the Northeast. This pattern may be influenced by cultural norms, educational attainment, and occupational distribution prevalent within these groups. Indirectly, these factors contribute to reinforcing existing socioeconomic stereotypes, even if the underlying causes are more structural than individual.

Income vs. private Health insurance obtained:

Those with private health insurance consistently show higher overall median income levels.

Income vs. public health insurance:

A similar pattern appears when comparing public health insurance to income levels, suggesting that lower-income individuals are generally less likely to hold any form of health insurance.

Income vs. Gender:

It shows a general income differentiation of male vs all “others”, yet, trans gender shows a much lower differentiation between those of the trans category.

Income vs. Workloss:

As expected, those with a recent loss in work have a lower overall expected income midpoint.

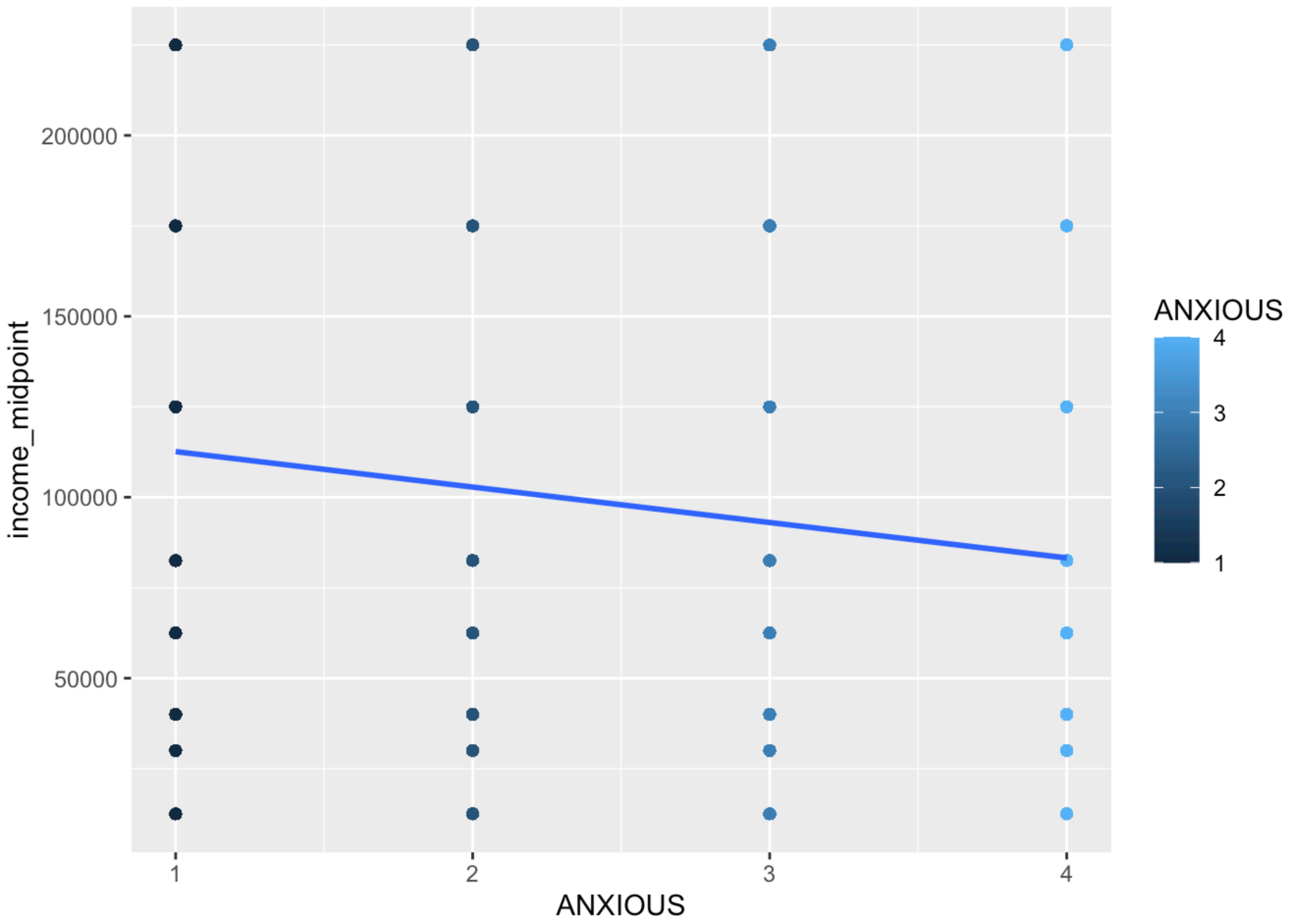
Use geom_point to see which continuous data has a greate effect on income_midpoint

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

```
## Warning: Removed 13253 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

```
## Warning: Removed 13253 rows containing missing values or values outside the scale ran
ge
## (`geom_point()`).
```



Income vs Anxiety:

When plotting general anxiety levels against income midpoints, we see a gradual decline in income as anxiety increases. This could suggest that financial instability and job insecurity contribute to higher anxiety, or conversely, that chronic anxiety limits access to higher-paying roles due to challenges in high-pressure environments. It’s also likely the relationship is cyclical, anxiety both results from and reinforces lower income. This underscores the need to view mental health as a key factor in economic opportunity.

```
##
## Call:
## lm(formula = income_midpoint ~ Education + Mar_Stat + Race +
##     private_health_ins + public_health_ins + Gender + workloss +
##     ANXIOUS, data = filtered_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -167256  -38324   -5122   38066  239139
##
## Coefficients: (1 not defined because of singularities)
##                                                Estimate Std. Error t value
## (Intercept)                                     96094.5     3318.1  28.961
## Educationsome hs                                  654.4     3771.7   0.173
## Educationhigh school                             8267.9     3260.1   2.536
## Educationsome college                           20378.1     3234.4   6.300
## Educationassoc deg                              21384.1     3276.0   6.527
## Educationcollege grad                           48607.5     3219.5  15.098
## Educationadv degree                             64780.8     3225.0  20.087
## Mar_Statwidowed                                -38980.6     2413.7 -16.150
## Mar_Statdivorced                               -38265.7      736.3 -51.973
## Mar_Statseparated                              -34230.6     1567.5 -21.837
## Mar_Statnever                                  -41584.9      515.2 -80.719
## RaceBlack                                      -14804.6      874.6 -16.928
## RaceAsian                                        1237.4      924.5   1.338
## Raceother                                       -7220.1     1109.5  -6.508
## private_health_inshas private health insurance  10216.3      755.4  13.525
## private_health_insno private health insurance  -15990.1     1034.3 -15.459
## public_health_inshas public health insurance   -27122.4      740.6 -36.622
## public_health_insno public health insurance          NA         NA      NA
## Genderfemale                                   -13736.6      463.2 -29.657
## Gendertrans                                    -23282.8     4093.4  -5.688
## Genderother                                    -16355.8     2632.5  -6.213
## worklossno                                      12492.3      576.1  21.685
## ANXIOUS                                         -5064.9      216.9 -23.353
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## Educationsome hs                                 0.8623
## Educationhigh school                             0.0112 *
## Educationsome college                          2.99e-10 ***
## Educationassoc deg                             6.74e-11 ***
## Educationcollege grad                           < 2e-16 ***
## Educationadv degree                             < 2e-16 ***
## Mar_Statwidowed                                 < 2e-16 ***
## Mar_Statdivorced                                < 2e-16 ***
## Mar_Statseparated                               < 2e-16 ***
## Mar_Statnever                                   < 2e-16 ***
## RaceBlack                                       < 2e-16 ***
## RaceAsian                                        0.1807
## Raceother                                      7.70e-11 ***
## private_health_inshas private health insurance  < 2e-16 ***
## private_health_insno private health insurance   < 2e-16 ***
## public_health_inshas public health insurance    < 2e-16 ***
## public_health_insno public health insurance          NA    
## Genderfemale                                    < 2e-16 ***
## Gendertrans                                    1.29e-08 ***
## Genderother                                    5.24e-10 ***
## worklossno                                      < 2e-16 ***
## ANXIOUS                                         < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52740 on 57362 degrees of freedom
##   (3590 observations deleted due to missingness)
## Multiple R-squared:  0.4171, Adjusted R-squared:  0.4169 
## F-statistic:  1955 on 21 and 57362 DF,  p-value: < 2.2e-16
```