SMML Class 1 Lab Take-Home

John Kubale

8/27/2024

Review. Take home exercise

Data from the Health and Life Study of Koreans (HLSK) is available on Canvas, "HLSK.RDS". The codebook and other associated materials are available from https://www.icpsr.umich.edu/web/RCMD/studies/37635

1. Download the data from Canvas. You can read the data into R using the "readRDS" function. 1a. How would you import the data using a relative filepath (hint: "..")?

```
hlsk <- #readRDS(here( "UMD/classes/stat_mod_ML_1_SURV615/class_1/", "HLSK.RDS")) |>
  readRDS("~/UMD/classes/stat_mod_ML_1_SURV615/class_1/HLSK.RDS") |>
  as_tibble() |>
  # change the name of income variable
  rename(income = LQ3_PUB)
```

2. Find the household annual income variable. What difference do you see in this, compared to income in Wage and psid?

- The range of income is very wide, however, it is right skewed with a few top earners of $300,000 at the top end. In this case the mean is not useful due to it's skewness, so here I report the median of $63,906 with 50% of income falling in between $23,500 and $90,000.

```
#Check the codebook from ICPSR website: 37635-0001-Codebook-ICPSR.pdf
hlsk |> select(income) |> summary()
```

```
##      income
##  Min.   :  5000
##  1st Qu.: 23500
##  Median : 48000
##  Mean   : 63906
##  3rd Qu.: 90000
##  Max.   :300000
##  NA's   :28
```

3. What is the minimum, mean, mode, median and maximum of the income? Write your

own function so that na.rm = TRUE by default.

```
hlsk |> select(income) |> drop_na() |>
  summarise(across(income, list(min, mean, Mode, median, max))) |>
  rename(minimum=1, mean=2, mode=3, median=4, maximum=5)
```

```
## # A tibble: 1 x 5
##   minimum   mean  mode median maximum
##     <dbl>  <dbl> <dbl>  <dbl>   <dbl>
## 1    5000 63906.  5000  48000  300000
```

4. What is the variance and standard deviation of the income?

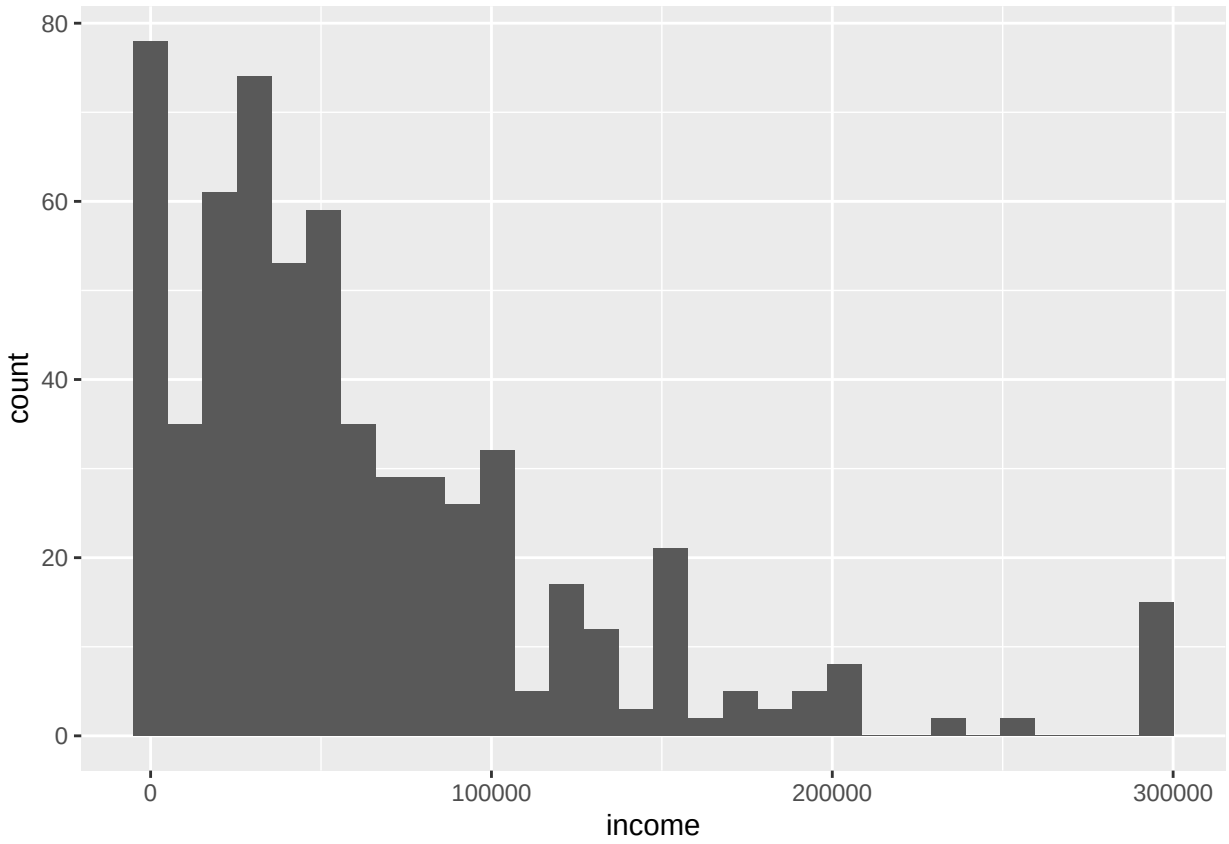
```
hlsk |> select(income) |> drop_na() |>
  summarise(across(income, list(sd, var))) |>
  rename(stand_deviation=1, variance=2)
```

```
## # A tibble: 1 x 2
##   stand_deviation    variance
##             <dbl>       <dbl>
## 1          61042. 3726068140.
```

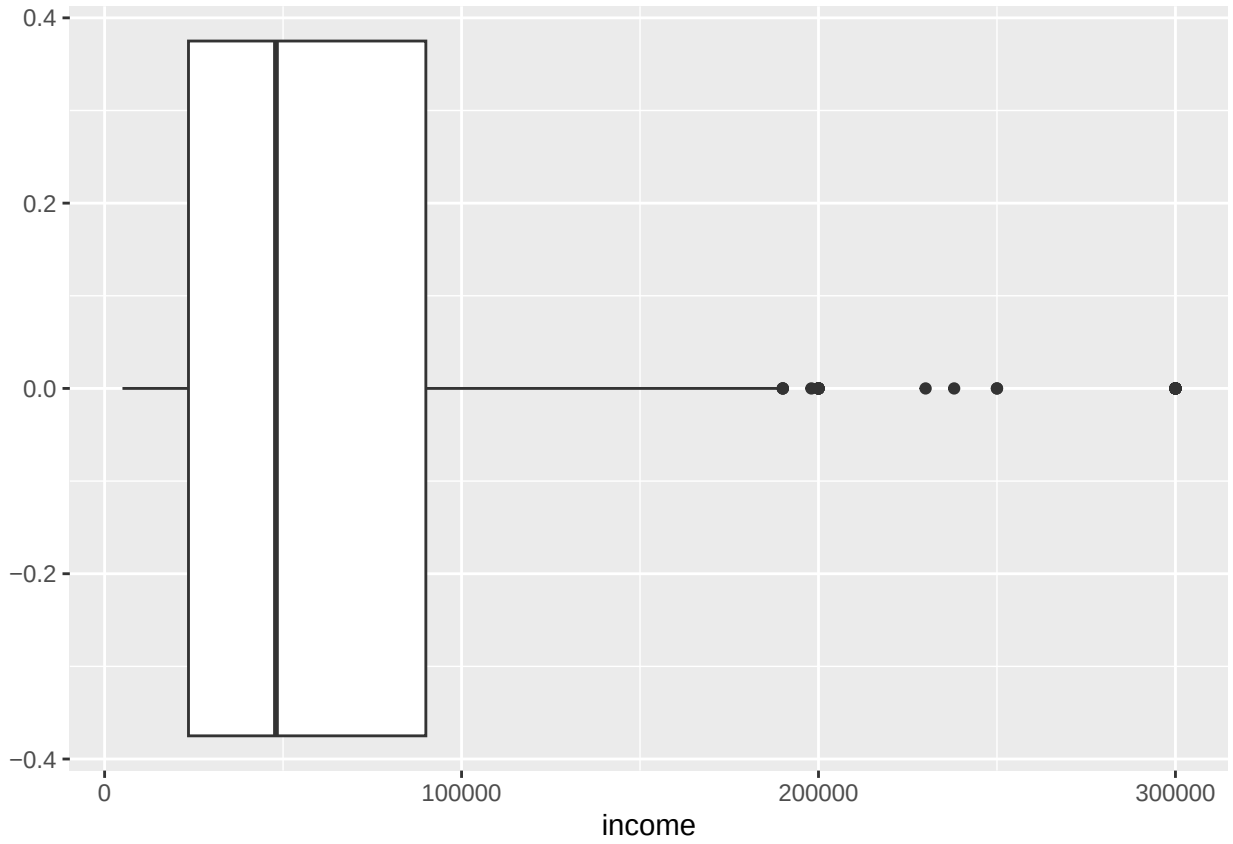
5. Visualize the income using a histogram and a box plot. What does notch=TRUE mean and when might this be useful (hint: ?geom_boxplot)? What are the benefits of each visualization method? How about drawbacks?

- Notch = true separates out the values within the first and third IQR. This helps us assess whether medians in the distribution differ. This is more useful with additional boxplots in the visual.
- Histograms allow us to examine the kernal desities of plausable values within a distribution of a continous random variable. However, when there are few values, it is difficult to find the proper bin size yet there are formulas for this. The boxplot helps us visualize summary values such as min, max, mean, and interquartile range in additional to outliers. However, we cannot determine from this plot how disperse values are from the mean. Adding the dotplot gives us a better visual or idea of where most values are in this distribution, yet it seems that this type of plot becomes more valuable as the sample size increases.

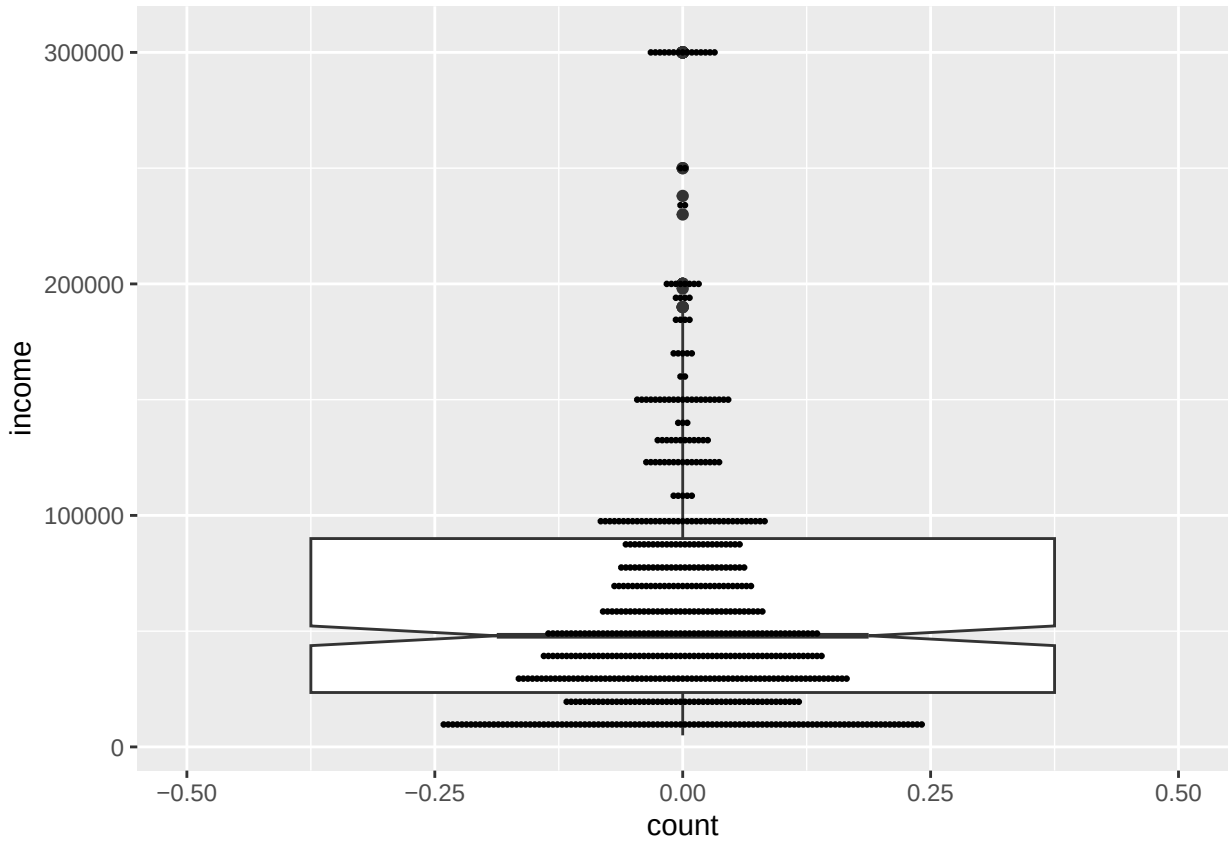
```
hlsk |> ggplot(aes(x=income)) + geom_histogram()
```




```
hlsk |> ggplot(aes(x=income)) +  geom_boxplot()
```




```
hlsk |> ggplot(aes(x=income)) + geom_boxplot(notch=TRUE) +
   # we can also overlay a dot plot onto the boxplot using geom_dotplot and adding a "
  geom_dotplot(binaxis='x', stackdir='center', dotsize=0.2) +
  coord_flip()
```



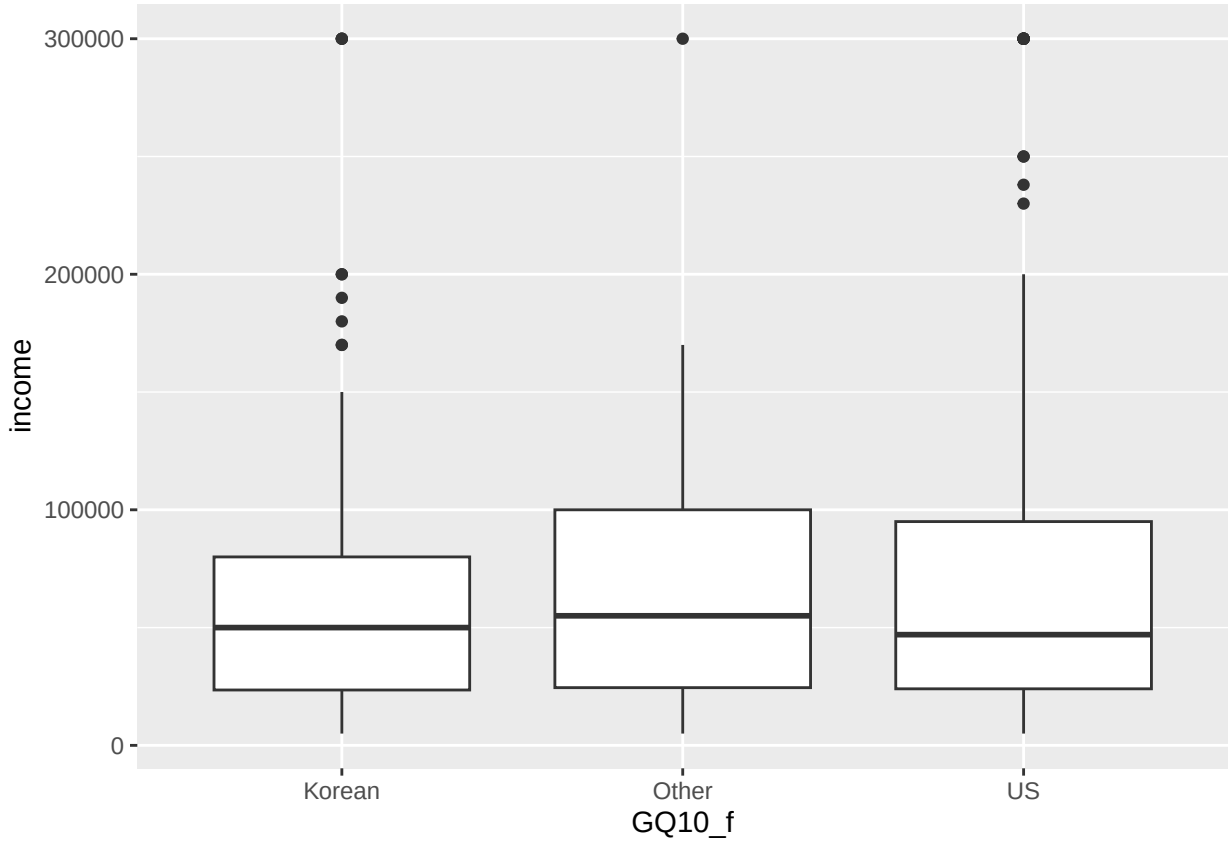
6. Going over the codebook and think about what kind of stories you want to learn about the income. How would you express those stories with formulas? Hint: mathematical formulas can be written in R using LaTeX by enclosing in '$ $' (e.g., $\hat{\beta_0}+$). A useful cheatsheet for this notation can be found at: https://kapeli.com/cheat_sheets/LaTeX_Math_Symbols.docset/Contents/Resources/Documents/index.

- An important story to untangle would be regional and demographic factors. For instance, we could use geography since these data were collected in Louisiana and Michigan in addition to age, although we may want to consider binning age somehow. We can express this linear relationship as:

$$\hat{E[Y_{income_i}]} = \beta_0 + \beta_1 \times region_i + \beta_2 \times age_i + \epsilon$$

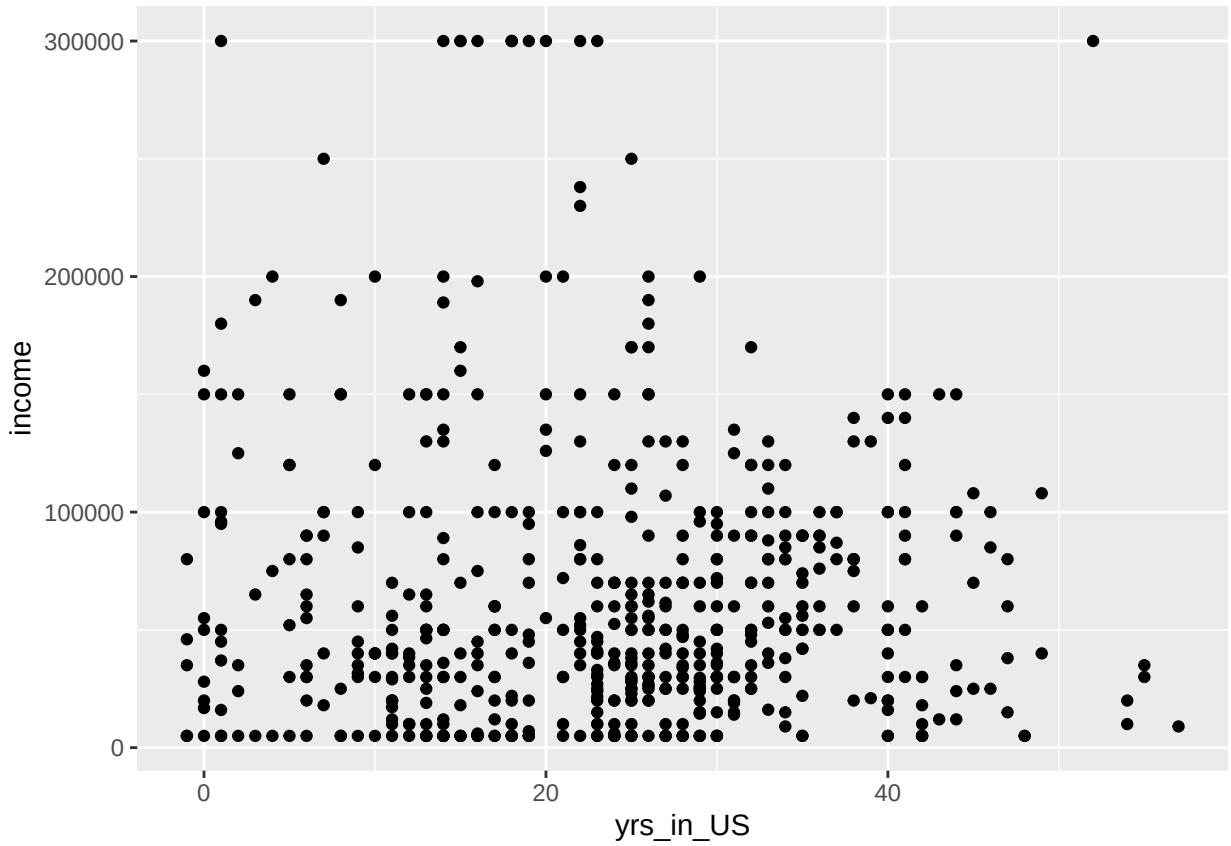
- What are potential factors that may influence immigrants' income (y_i) given the HLSK data?
- Perhaps salaries are not completely reported by immigrants because a lot of their labor is paid in cash and not accounted for through employer's taxes.
- Is having a final degree from the US associated with higher income for Korean immigrants than a degree from elsewhere?
- Based on a simple boxplot, it does not appear that there is a clear difference in income for Koreans based on where they received their last degree from.

```r
hlsk |> # respondents are all identified as Korean
  select(income,  GQ10) |>
  drop_na() |>
  mutate(GQ10_f = case_when(
    GQ10 == 1 ~ "US",
    GQ10 == 2 ~ "Korean",
    GQ10 == 3 ~ "Other")) |>
  ggplot(aes(y=income, x=GQ10_f)) +
  geom_boxplot()
```



- What is the relationship between years in the U.S. and immigrants' income?
- Using a scatterplot to visualize time in the US and income, there is no clear relationship between these two variables.

```r
hlsk |> # respondents are all identified as Korean
  select(income, GQ5, AQ1_PUB) |>
  drop_na() |>
  mutate(birth_year = 2017 - AQ1_PUB,
         yrs_in_US = GQ5 - birth_year) |>
  ggplot(aes(x=yrs_in_US, y=income)) +
  geom_point()
```



- What is the effect of one more year in the U.S. on immigrants' income?
- For every 1 year in the US, we would expect Koreans to make $224 less dollars.

```
dat <- hlsk |> # respondents are all identified as Korean
  select(income, GQ5, AQ1_PUB) |>
  drop_na() |>
  mutate(birth_year = 2017 - AQ1_PUB,
         yrs_in_US = GQ5 - birth_year)

mod <- lm(income ~ yrs_in_US, data=dat)

summary(mod)
```

```
##
## Call:
## lm(formula = income ~ yrs_in_US, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63526 -40366 -17057  26384 243357
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
```

```
## (Intercept)  68301.8     5533.1  12.344 <0.0000000000000002 ***
## yrs_in_US     -224.2      210.6  -1.065                0.287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60170 on 576 degrees of freedom
## Multiple R-squared:  0.001964,   Adjusted R-squared:  0.0002312
## F-statistic: 1.133 on 1 and 576 DF,  p-value: 0.2875
```