SMML Class 1 Lab Take-Home

Sunghee Lee

8/30/2022

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.

```
HLSK<-readRDS(
  "/Users/ceciliayao/Desktop/Graduate Study/1st sem 2022-2023/SurvMeth 685/SurvMeth-685-
#You need to change the path in the line above with yours as this
#is unique to my computer
```

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

```
#Check the codebook from ICPSR website: 37635-0001-Codebook-ICPSR.pdf
names(HLSK)
```

```
##   [1] "UNIQUE_NUM"       "STARTLANG"        "MAIN_COMP"
##   [4] "CPN_NUM"          "DEVICEMOBILE"     "INTVLANG"
##   [7] "SITE"             "AQ1_PUB"          "AQ1_1"
##  [10] "AQ2_KOREAN"       "AQ3"              "AQ4"
##  [13] "AQ5_ENGLISH"      "AQ5_KOREAN"       "AQ6"
##  [16] "AQ6_1"            "AQ7"              "AQ8"
##  [19] "AQ8_INDICATOR"    "BQ1_MALE"         "BQ2"
##  [22] "BQ2A"             "BQ3"              "BQ4"
##  [25] "BQ5_1"            "BQ5_2_PUB"        "BQ5_3_PUB"
##  [28] "BQ5_4"            "CQ1"              "CQ1_EXP_LOCATION"
##  [31] "CQ1_EXP_SCALE"    "CQ2"              "CQ3"
##  [34] "CQ4"               "CQ5"               "CQ6"
##  [37] "CQ7"               "CQ8"               "CQ9"
##  [40] "CQ10"              "CQ11"              "CQ2_11"
##  [43] "CQ12"              "EXP_AGREE"         "CQ13"
##  [46] "CQ14"              "CQ14_EXP"          "CQ15"
##  [49] "CQ16"              "CQ17"              "CQ17_EXP"
##  [52] "DQ1"               "DQ2"               "DQ3"
##  [55] "DQ3_1"             "DQ4"               "DQ5"
##  [58] "DQ6"               "DQ7"               "DQ7_1"
##  [61] "DQ8"               "DQ9"               "DQ10_PUB"
##  [64] "DQ11"              "DQ11_EXP_SCALE"    "DQ12_UNIT"
##  [67] "DQ12_PUB"          "HEIGHT_CM_PUB"     "DQ13_UNIT"
##  [70] "DQ13_PUB"          "WEIGHT_KG_PUB"     "BMI_PUB"
##  [73] "BMI_CAT_PUB"       "BMI_OBESE_PUB"     "DQ14"
##  [76] "DQ15"              "DQ16"              "DQ17"
##  [79] "DQ18"              "DQ19"              "DQ20"
##  [82] "DQ21"              "DQ22"              "DQ23"
##  [85] "DQ24"              "EQ1"               "EQ2"
##  [88] "EQ3"               "EQ4"               "EQ1_4_GRID"
##  [91] "FQ1"               "FQ2"               "FQ3"
##  [94] "FQ4"               "FQ5"               "FQ6"
##  [97] "FQ1_6_GRID"        "FQ7"               "FQ8"
## [100] "FQ9"               "FQ10"              "FQ11"
## [103] "FQ12"              "FQ13"              "FQ14"
## [106] "FQ15"              "GQ1"               "GQ2"
## [109] "GQ3"               "GQ5"               "GQ6"
## [112] "GQ7"               "GQ8"               "GQ9"
## [115] "GQ10"              "GQ12"              "GQ13"
## [118] "GQ14"              "GQ15_PUB"          "GQ16"
## [121] "GQ17"              "GQ18"              "GQ19"
## [124] "GQ20"              "GQ21"              "GQ22"
## [127] "GQ23"              "GQ24_HOBBY"        "GQ24_NONE"
## [130] "GQ24_OTHER"        "GQ24_POLITICAL"    "GQ24_PROFESSION"
## [133] "GQ24_REL"          "GQ24_SCHOOL"       "GQ24_VOLUNTEER"
## [136] "GQ25"              "GQ26"              "GQ27"
## [139] "GQ28"              "GQ29"              "GQ30"
## [142] "GQ31"              "GQ32"              "HQ1"
## [145] "HQ2"               "HQ3"               "HQ4"
## [148] "HQ4_EXP"           "HQ5"               "HQ6"
## [151] "HQ6_EXP"           "HQ7"               "HQ8"
## [154] "HQ9"               "HQ10"              "HQ10_EXP"
## [157] "HQ11"              "HQ12"              "HQ12_EXP"
## [160] "HQ13"              "HQ14"              "HQ15"
## [163] "HQ15_EXP"          "HQ16"              "HQ16_EXP"
## [166] "HQ17"              "HQ18"              "HQ18_EXP"
## [169] "HQ19"             "HQ20"             "HQ20_EXP"
## [172] "HQ21"             "JQ1"              "JQ1_EXP_LOCATION"
## [175] "JQ2"              "JQ2_EXP"          "JQ3"
## [178] "JQ3_EXP"          "JQ4"              "JQ4_EXP"
## [181] "JQ5"              "KQ1_EMPLOYER"     "KQ1_GOVASSIST"
## [184] "KQ1_KOREAN"       "KQ1_OTHER"        "KQ1_PURCHASE"
## [187] "KQ1_INSURED"      "KQ2"              "KQ3"
## [190] "KQ4"              "KQ5"              "KQ6"
## [193] "LQ1"              "LQ2"              "LQ3_PUB"
## [196] "LQ5"              "LQ6"              "LQ7"
## [199] "LQ8"              "LQ11"             "POV_LT200"
## [202] "LQ12"             "LQ13"             "LQ14"
## [205] "LQ15"             "LQ16"             "LQ17"
## [208] "LQ18"             "LQ19"             "NUMKR"
## [211] "FU_PARTICIPATE"
```

```
summary(HLSK$LQ3_PUB)
```

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

```
mean(HLSK$LQ3_PUB, na.rm=T)
```

```
## [1] 63905.6
```

```
mn<-function(x){
  mean(x, na.rm=T)
  }
mn(HLSK$LQ3_PUB)
```

```
## [1] 63905.6
```

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

```
summary(HLSK$LQ3_PUB)
```

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

```r
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
  }

Mode(HLSK$LQ3_PUB)
```

```
## [1] 5000
```

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

```r
var(HLSK$LQ3_PUB, na.rm=T)
```

```
## [1] 3726068140
```

```r
sd(HLSK$LQ3_PUB, na.rm=T)
```

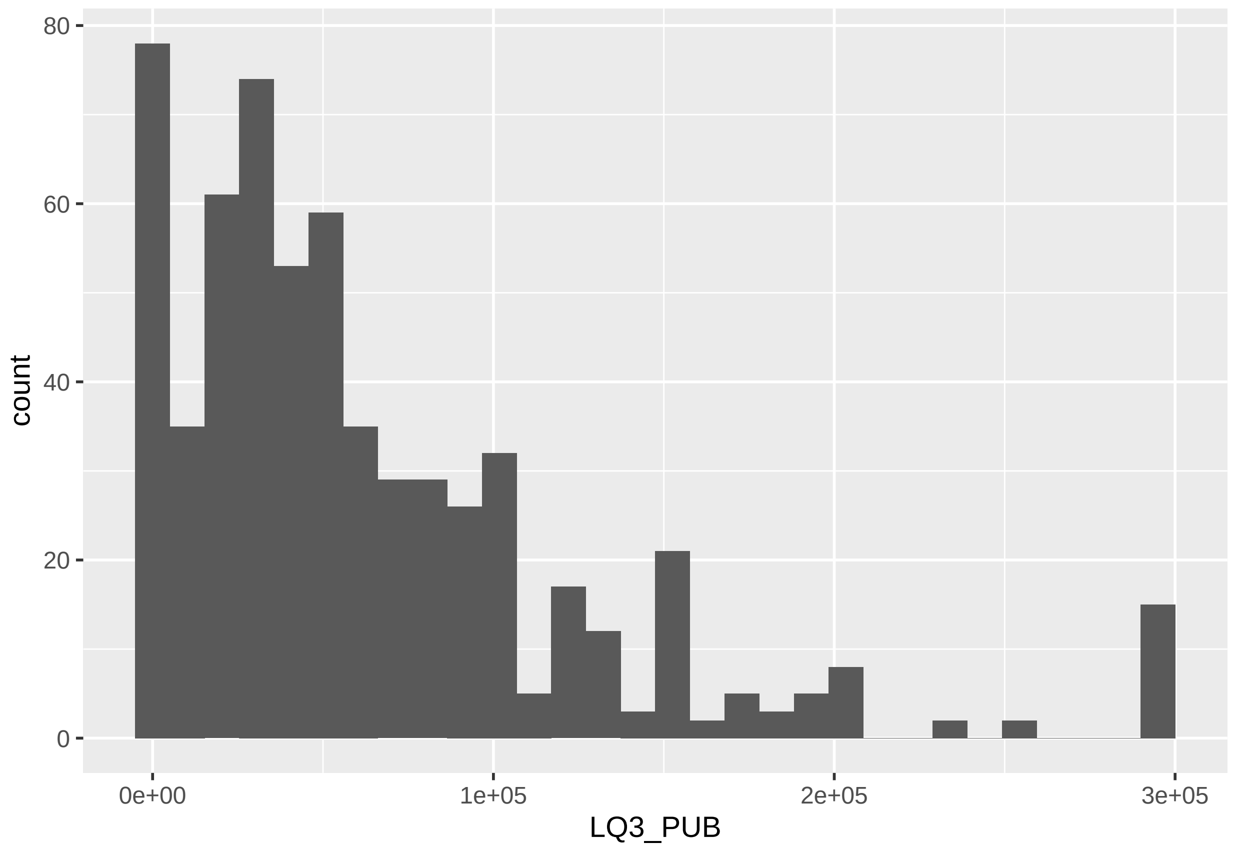
```
## [1] 61041.53
```

```r
sqrt(var(HLSK$LQ3_PUB, na.rm=T))
```

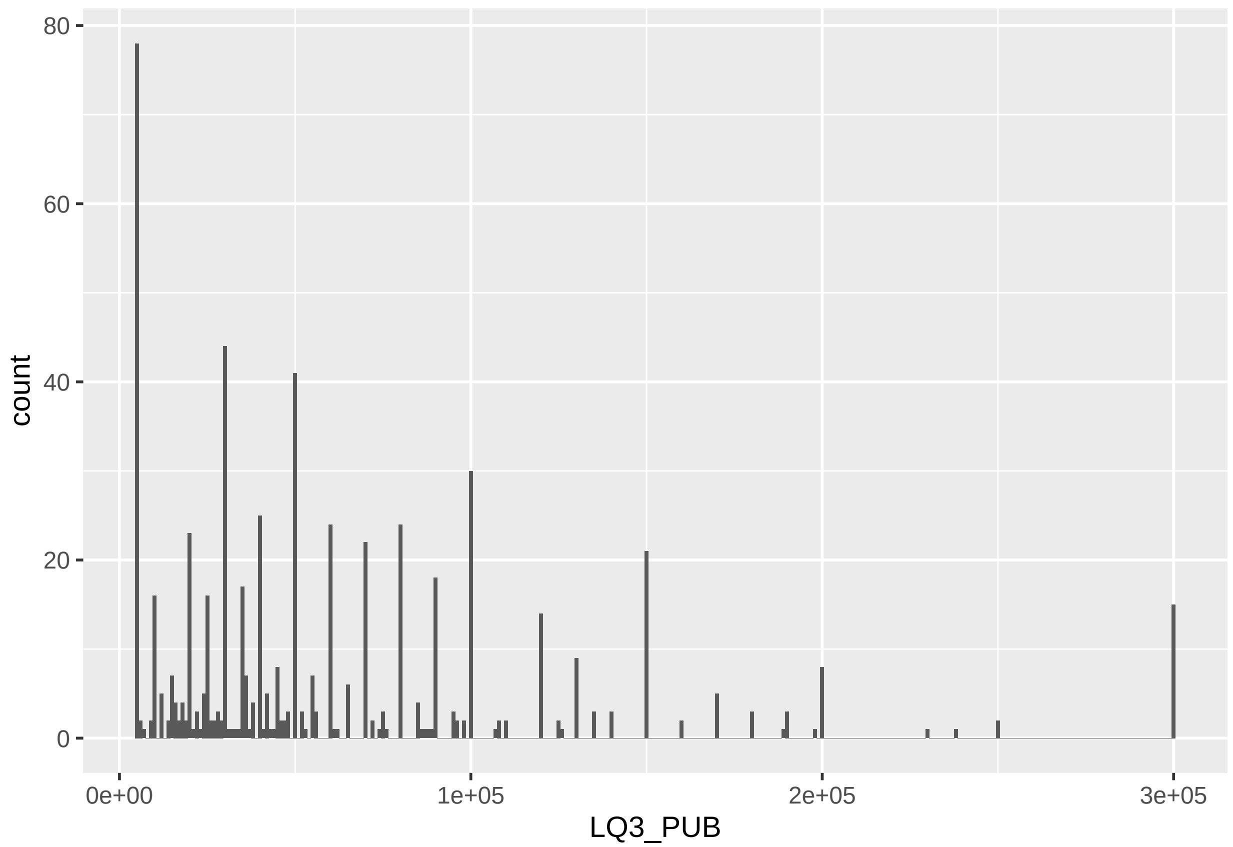
```
## [1] 61041.53
```

5. Visualize the income using a histogram and a box plot. Do you see the patterns in #3 and #4 in these? What are the benefits of each visualization method? How about drawbacks?

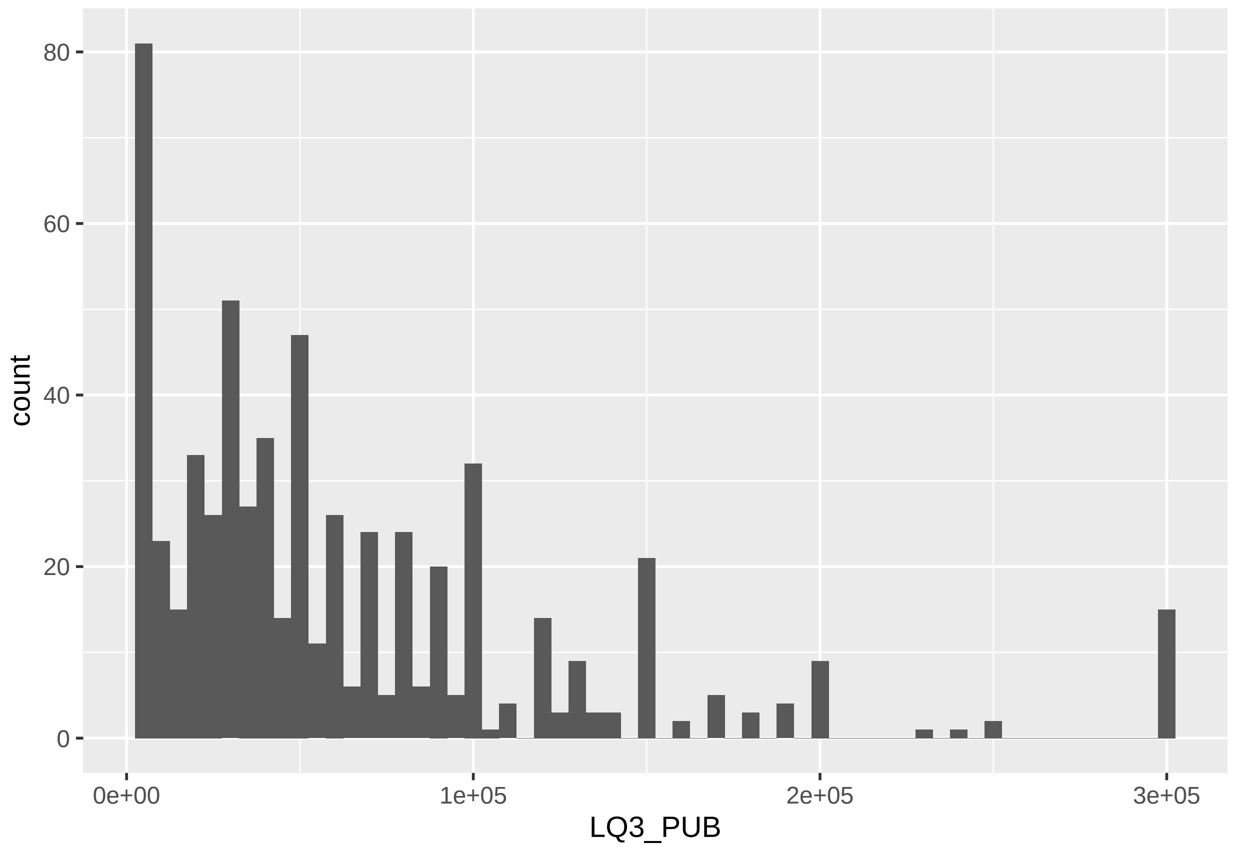
```r
library(ggplot2)
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_histogram()
```



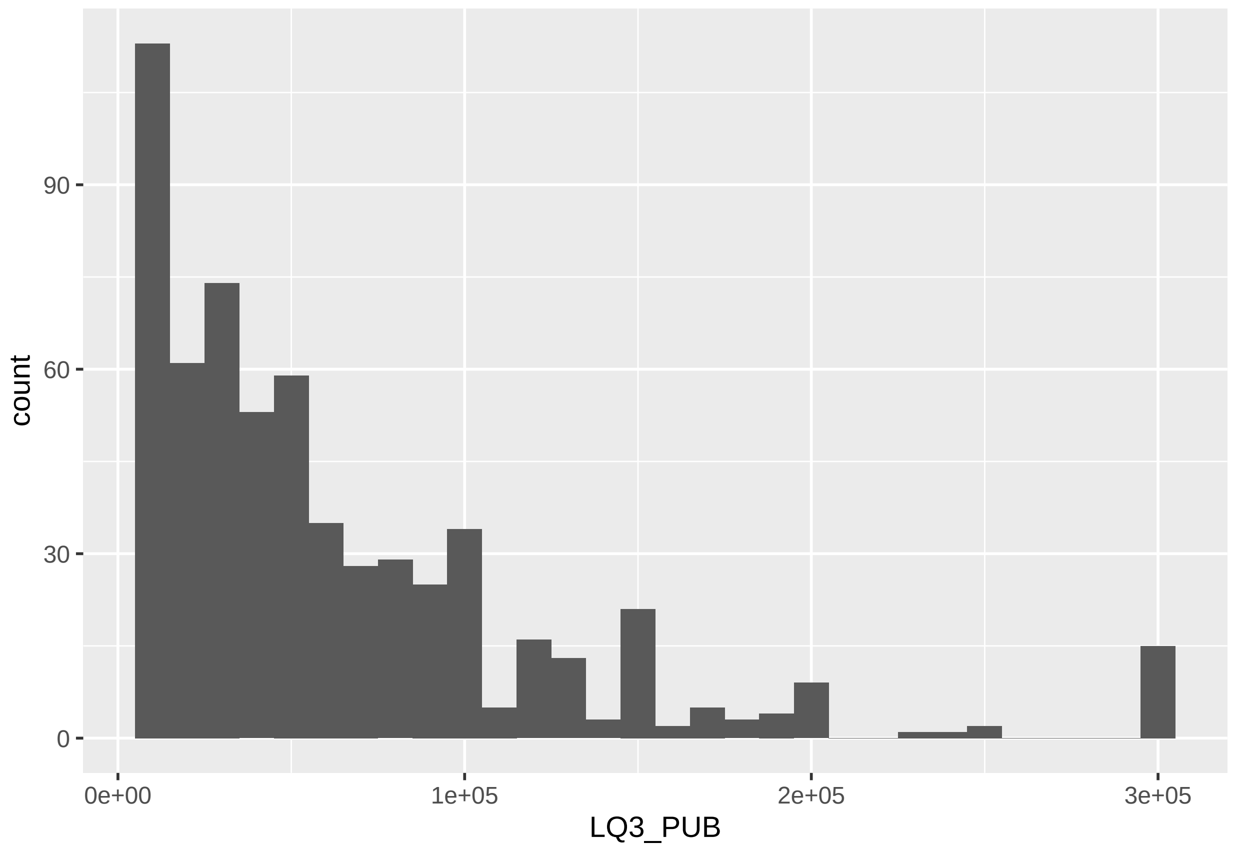
```r
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_histogram(binwidth=1000)
```



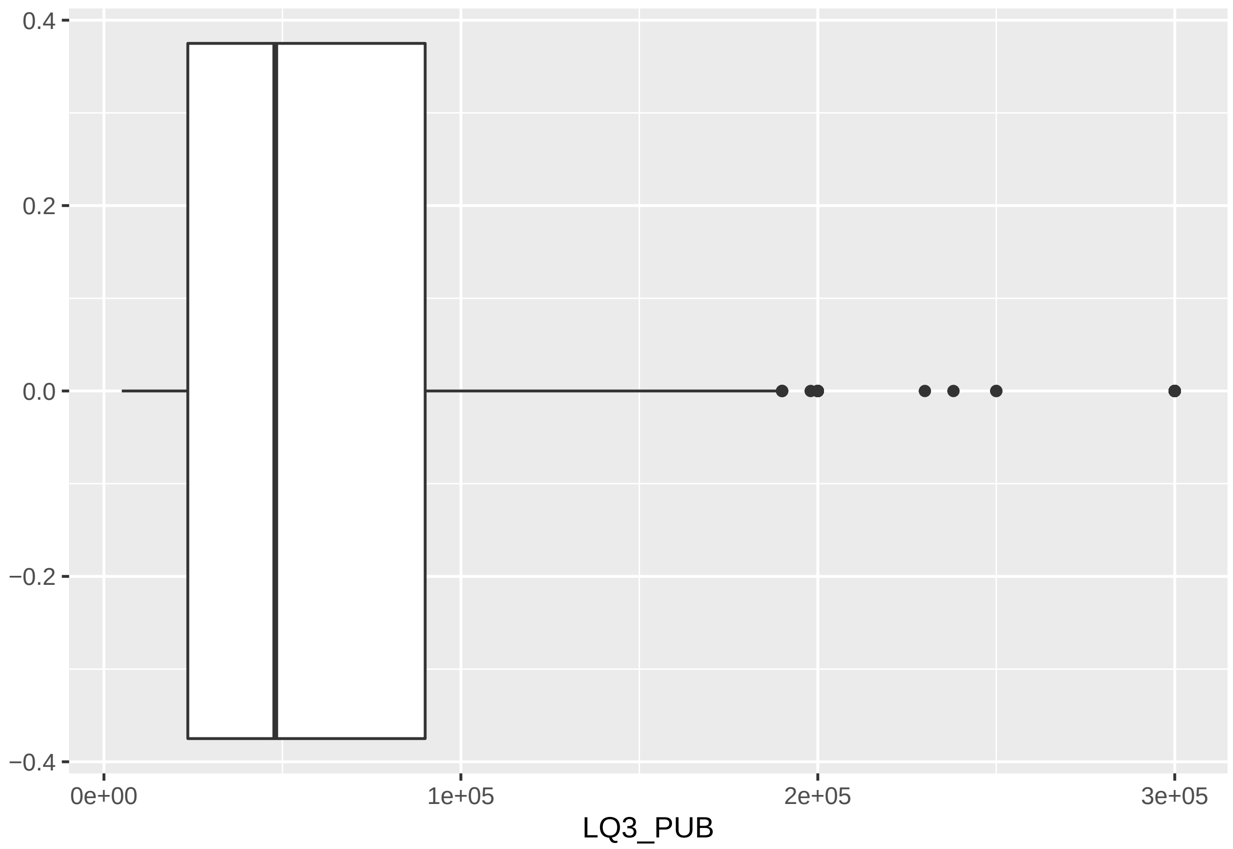
```
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_histogram(binwidth=5000)
```



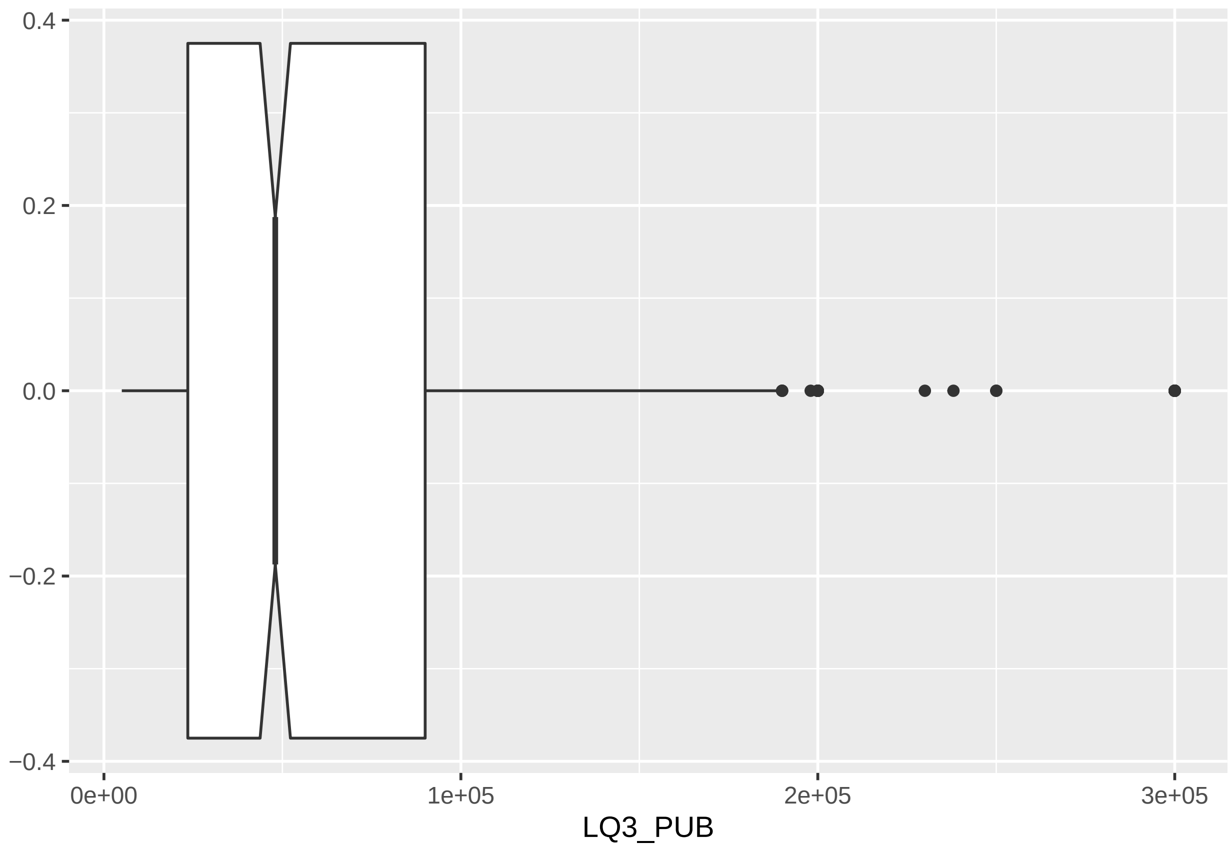
```
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_histogram(binwidth=10000)
```



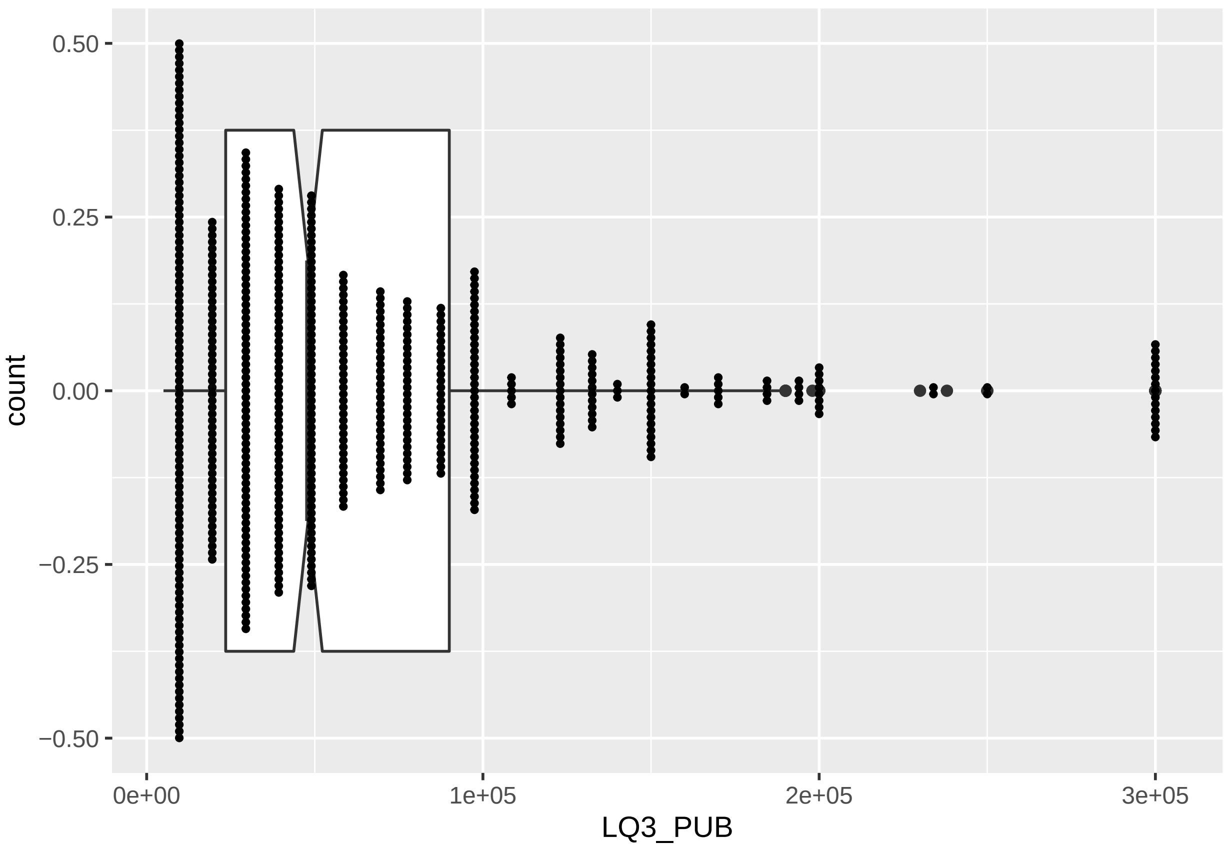
```
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_boxplot()
```



```
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_boxplot(notch=TRUE)
```



```
ggplot(HLSK, aes(x=LQ3_PUB)) + geom_boxplot(notch=TRUE) +
  geom_dotplot(binaxis='x', stackdir='center', dotsize=0.2)
```



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?

- What are potential factors that may influence immigrants' income (y_i) given the HLSK data?
  - Examples
    - Time of immirgration (e.g., before or after age 18)
    - Country where the final degree was obtained (e.g., US. vs. non-US)
    - Years in the U.S.
- Is having a final degree from the US associated with higher income for Korean immigrants than a degree from elsewhere?
  - Formula: $\mu_{US} > \mu_{Non-US}$
- What is the relationship between years in the U.S. and immigrants' income?
  - Formula: $\rho_{income,USyrs}$
- What is the effect of one more year in the U.S. on immigrants' income?
  - Formula: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$