R Module 4 Rubric

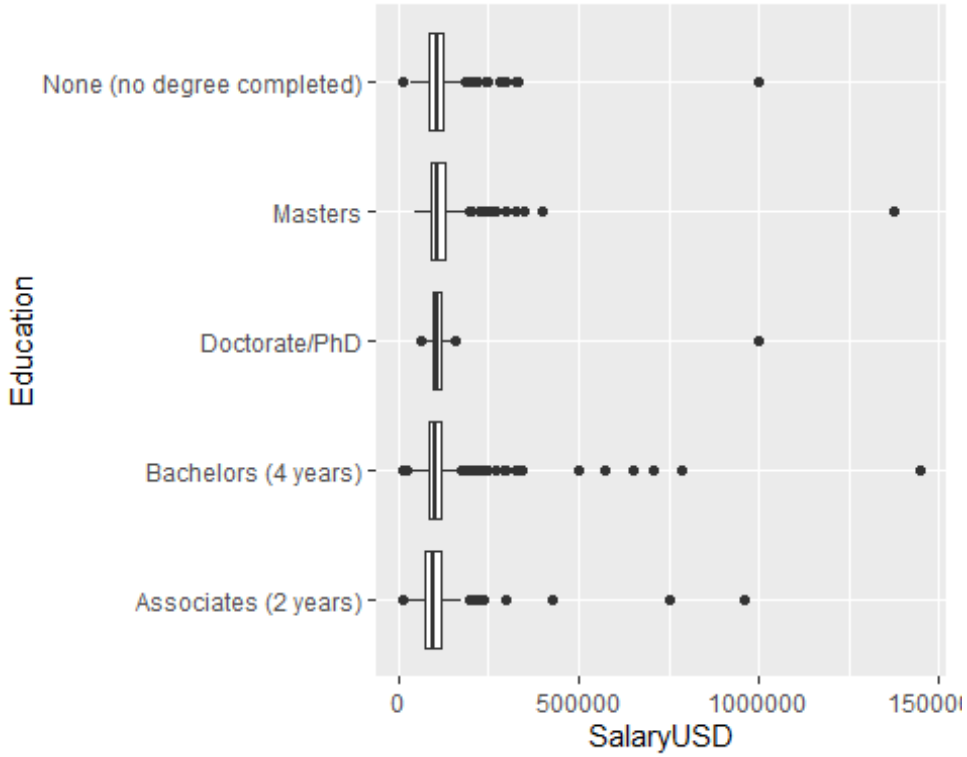
This R Module introduces R Markdown, and is designed to get students comfortable with using the format to produce reports and generate documents. Rather than taking screenshots and pasting into a .docx, students should be able to run code directly in their .Rmd, and have code and figures embedded in their document.

From this point on, students should be using R Markdown. I definitely don't want to dictate any sort of major change to the syllabus, so that decision is up to you. However, I think using R Markdown, while there is a learning curve, will lead to better code and an easier time trying to debug students' code and projects.

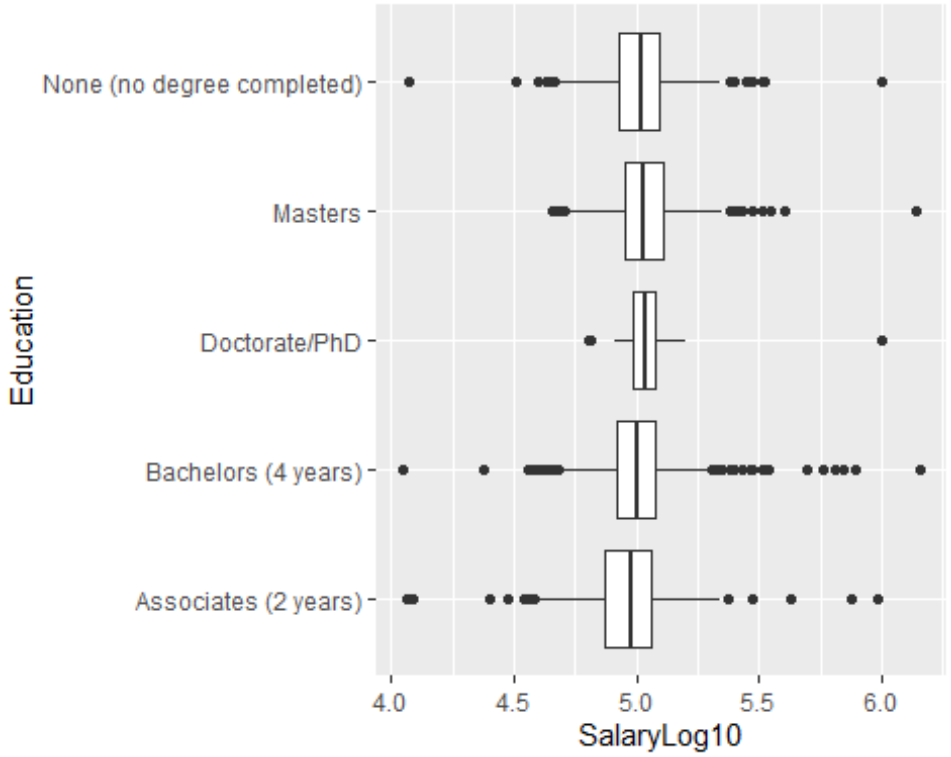
Assignment

```
library(tidyverse)
library(readxl)

survey <- read_xlsx(path = "data/salary_survey.xlsx")

survey_sub <- survey %>%
  dplyr::select(`Survey Year`, Country, PrimaryDatabase, SalaryUSD,
YearsWithThisDatabase, YearsWithThisTypeOfJob, Education)

survey_sub <- survey_sub %>%
  dplyr::filter(
    YearsWithThisDatabase <= 50,

    # Same for the following:
    YearsWithThisTypeOfJob <= 50,

    # We're only interested in the U.S.
    Country == "United States",

    # We want to filter out "missing values"
    Education != "Not Asked",

    # Some respondents put in their hourly wage rather than their yearly
salary;
    # it's doubtful that anyone only makes $13 USD per year working in this
kind
    # of job!

    SalaryUSD > 1000
  )
```

```
survey_sub %>%
  ggplot(aes(y = Education, x = SalaryUSD)) +
  geom_boxplot()
```




```
survey_sub <- survey_sub %>%
  mutate(
    SalaryLog10 = log10(SalaryUSD)
  )


# We'll pop this back into our boxplots...
survey_sub %>%
  ggplot(aes(y = Education, x = SalaryLog10)) +
  geom_boxplot()
```




```
survey_sub$Education <-
  factor(
    survey_sub$Education,

    # We set this argument to TRUE when the order of our factor matters, or
if
    # we intend to compare the "amount" of education (a PhD is a greater
    # "amount" of education than a Bachelors, for example.)
    ordered = TRUE,

    # The `levels` argument requests a character vector of the different
factor
    # levels in the dataset, and the order we want them to be in.
    levels = c(
      "None (no degree completed)",
      "Associates (2 years)",
      "Bachelors (4 years)",
      "Masters",
      "Doctorate/PhD"
    )
  )


summary(survey_sub)
```

```
##   Survey Year     Country          PrimaryDatabase      SalaryUSD
YearsWithThisDatabase
##  Min.   :2017   Length:4494        Length:4494        Min.   :  11100
Min.   : 0.00
##  1st Qu.:2017   Class :character   Class :character   1st Qu.:  85000
1st Qu.: 6.00
##  Median :2018   Mode  :character   Mode  :character   Median : 102000
Median :10.00
##  Mean   :2018                                         Mean   : 107496
Mean   :11.31
##  3rd Qu.:2018                                         3rd Qu.: 122000
3rd Qu.:16.00
##  Max.   :2019                                         Max.   :1450000
Max.   :38.00
##  YearsWithThisTypeOfJob                      Education     SalaryLog10
##  Min.   : 0.000         None (no degree completed): 671   Min.   :4.045
##  1st Qu.: 3.000         Associates (2 years)      : 500   1st Qu.:4.929
##  Median : 5.000         Bachelors (4 years)       :2540   Median :5.009
##  Mean   : 7.386         Masters                   : 759   Mean   :5.004
##  3rd Qu.:10.000         Doctorate/PhD             :  24   3rd Qu.:5.086
##  Max.   :40.000                                         Max.   :6.161
```

```
factors <-
  # The `cut()` function takes a character vector and "cuts" it into a
factor,
  # by looking at values that fall within each break. Think of it as plotting
a
  # histogram with a certain number of bins, and assigning each of those bins
as
  # a factor level.
  cut(
    survey_sub$SalaryLog10,
    breaks = c(4, 5, 6, 7),
    labels = c("5 Figures", "6 Figures", "7 Figures")
  )

survey_sub$SalaryFigs <- factors

summary(survey_sub)
```

```
##   Survey Year     Country          PrimaryDatabase      SalaryUSD
YearsWithThisDatabase
##  Min.   :2017   Length:4494        Length:4494        Min.   :  11100
Min.   : 0.00
##  1st Qu.:2017   Class :character   Class :character   1st Qu.:  85000
1st Qu.: 6.00
##  Median :2018   Mode  :character   Mode  :character   Median : 102000
Median :10.00
##  Mean   :2018                                         Mean   : 107496
Mean   :11.31
```

```
##  3rd Qu.:2018                                          3rd Qu.: 122000
3rd Qu.:16.00
##  Max.   :2019                                          Max.   :1450000
Max.   :38.00
##  YearsWithThisTypeOfJob                      Education     SalaryLog10
SalaryFigs
##  Min.   : 0.000         None (no degree completed): 671   Min.   :4.045
5 Figures:2182
##  1st Qu.: 3.000         Associates (2 years)      : 500   1st Qu.:4.929
6 Figures:2309
##  Median : 5.000         Bachelors (4 years)       :2540   Median :5.009
7 Figures:   3
##  Mean   : 7.386         Masters                   : 759   Mean   :5.004
##  3rd Qu.:10.000         Doctorate/PhD             :  24   3rd Qu.:5.086
##  Max.   :40.000                                           Max.   :6.161
```

```
survey_clean <- survey_sub %>%
  # We're not interested in these columns, so we can exclude them with a "-"
  dplyr::select(-PrimaryDatabase, -Country, SalaryLog10) %>%
  # Likewise, we want to ignore the cases where the salary is 7 figures, so
we set our filter criterion to the observations where SalaryFigs is not (!=
represents "is not") "7 Figures".
  dplyr::filter(SalaryFigs != "7 Figures") %>%
  # Even though we filtered out the cases of 7-figure salaries, the "7
Figures"
  # level still exists within our data. We use the `droplevels()` function to
  # remove unused factor levels. Nothing actually changes about our data
itself,
  # but it helps keep our results tidy.
  droplevels()
```

```
summary(survey_clean)
```

```
##   Survey Year     SalaryUSD       YearsWithThisDatabase
YearsWithThisTypeOfJob                      Education
##  Min.   :2017   Min.   :  11100   Min.   : 0.00         Min.   : 0.000
None (no degree completed): 671
##  1st Qu.:2017   1st Qu.:  85000   1st Qu.: 6.00         1st Qu.: 3.000
Associates (2 years)      : 500
##  Median :2018   Median : 102000   Median :10.00         Median : 5.000
Bachelors (4 years)       :2538
##  Mean   :2018   Mean   : 106616   Mean   :11.31         Mean   : 7.382
Masters                   : 758
##  3rd Qu.:2018   3rd Qu.: 122000   3rd Qu.:16.00         3rd Qu.:10.000
Doctorate/PhD             :  24
##  Max.   :2019   Max.   :1000000   Max.   :38.00         Max.   :40.000
##   SalaryLog10        SalaryFigs
##  Min.   :4.045   5 Figures:2182
##  1st Qu.:4.929   6 Figures:2309
```

```
##  Median :5.009
##  Mean   :5.003
##  3rd Qu.:5.086
##  Max.   :6.000

survey_clean %>%
  ggplot(aes(x = YearsWithThisTypeOfJob, y = SalaryFigs, fill = SalaryFigs))
+
  geom_boxplot() +
  facet_grid(rows = vars(Education)) +
  labs(
    x = "Years with this type of job",
    y = "Salary Figures",
    title = "Years Experience vs. Salary",
    fill = "Figures"
  )
```