Final_Project_Proposal

Lance Ding, Grayson Stark, Kristy Tran

2023-04

```
knitr::opts_chunk$set(echo = TRUE)
pacman::p_load(tidyverse)
```

For our final project, we would like to do an analysis on this dataset of salary information. It is a dataset generated from user responses to the 2021 version of this survey on Ask a Manager. Each row records one respondee's answers to the survey, and each column records the respondees' answers to one question.

The dataset is saved locally in a subdirectory at `./Datasets/salary.csv`*.*

Basic Exploration

```
salary <- read.csv("./Datasets/salary.csv")
summary(salary)
```

```
##   Timestamp          How.old.are.you.   What.industry.do.you.work.in.
##  Length:28043       Length:28043       Length:28043
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Job.title
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  If.your.job.title.needs.additional.context..please.clarify.here.
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  What.is.your.annual.salary...You.ll.indicate.the.currency.in.a.later.question..If.you.are.part.time
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  How.much.additional.monetary.compensation.do.you.get..if.any..for.example..bonuses.or.overtime.in.an
##  Min.   :        0
##  1st Qu.:        0
##  Median :     2000
##  Mean   :    18208
##  3rd Qu.:    10000
##  Max.   :120000000
##  NA's   :7350
##  Please.indicate.the.currency If..Other...please.indicate.the.currency.here..
##  Length:28043                 Length:28043
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
##
##  If.your.income.needs.additional.context..please.provide.it.here.
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  What.country.do.you.work.in. If.you.re.in.the.U.S...what.state.do.you.work.in.
##  Length:28043                 Length:28043
##  Class :character             Class :character
##  Mode  :character             Mode  :character
##
##
##
##
##  What.city.do.you.work.in.
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  How.many.years.of.professional.work.experience.do.you.have.overall.
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
```

```
##
##  How.many.years.of.professional.work.experience.do.you.have.in.your.field.
##  Length:28043
##  Class :character
##  Mode  :character
##
##
##
##
##  What.is.your.highest.level.of.education.completed. What.is.your.gender.
##  Length:28043                                       Length:28043
##  Class :character                                   Class :character
##  Mode  :character                                   Mode  :character
##
##
##
##
##  What.is.your.race...Choose.all.that.apply..    X             X.1
##  Length:28043                                Mode:logical   Mode:logical
##  Class :character                            NA’s:28043     NA’s:28043
##  Mode  :character
##
##
##
##
##    X.2            X.3            X.4            X.5
##  Mode:logical   Mode:logical   Mode:logical   Mode:logical
##  NA’s:28043     NA’s:28043     NA’s:28043     NA’s:28043
##
##
##
##
##
```

```
# skimr::skim(salary)
```

skim() outputs a unicode character that cannot be rendered. I have commented it out to allow for the compilation of the pdf

Running `skim()` and `summary()` on the data tells us that there are **28043** rows and **24** columns. Of these, we have 17 `character` columns, 1 `numeric` column and 6 `logical` columns. A quick look at the actual data reveals that we actually have 6 empty columns, which correspond to the 6 `logical` columns, 1 question column that corresponds to the `numeric` column, and 17 other question columns that correspond to the 17 `character` columns. For more details on the dataset and its columns, consult the codebook.

Research Questions and Methods

There are a few potential research questions that we are interested in answering:

1. Is there an association between the **highest level of education attainment** and **annual salary**?
2. Is there an association between **age**, **race** and **annual salary**?
3. What industries have employees with the largest **work experience** to **age** ratio?
4. Is there a significant difference in **income/education** between **men/women** or **white/minority**?

To help answer these questions, we will likely need to employ these cleaning techniques:

- *Separating/Uniting*
  - *We may manually dirty up the dataset and to showcase an additional cleaning technique if we end up needing more than what materializes from the following*
- Cleaning Variable Names
  - Many of the variable names are crude, long, and contain spaces
- Cleaning Missing Values
  - Some individuals answered "prefer not to answer" to some questions, and many left irrelevant questions blank
- Re-coding Variables Values
  - Some entries in the column corresponding to the work country can be cleaned up and united e.g. US, United States and United States of America to US
- Cleaning Strings
  - Some entries in the column corresponding to the work city also have state information, which we do not want and can use `stringr` to clean
- Cleaning Factors
  - Certain columns can be reordered for easier visualization. Also, many columns actually contain categorical data - respondees had to pick responses from radio buttons or check boxes - meaning we can factor them and do categorical computations on them

Visualizations

- To answer our first research question, we could employ a **box-and-whisker** plot.
- To answer our second/third questions, we could use a **scatterplot**.
- To answer our fourth question, we could utilize a **facetted bar graph/histogram**.