Feng Jiayi_3035772028

Problem Set

Due: 2023-12-3 23:59 (HKT)

General Introduction

In this Problem Set, you will apply data science skills to wrangle and visualize the replication data of the following research article:

Cantú, F. (2019). The fingerprints of fraud: Evidence from Mexico's 1988 presidential election. *American Political Science Review*, *113*(3), 710-726.

Requirements and Reminders

- You are required to use **RMarkdown** to compile your answer to this Problem Set.
- Two submissions are required (via Moodle)
  - A `.pdf` file rendered by `Rmarkdown` that contains all your answer.
  - A compressed (in `.zip` format) R project repo. The expectation is that the instructor can unzip, open the project file, knitr your `.Rmd` file, and obtain the exact same output as the submitted `.pdf` document.
- The Problem Set is worth 30 points in total, allocated across 7 tasks. The point distribution across tasks is specified in the title line of each task. Within each task, the points are evenly distributed across sub-tasks. Bonus points (+5% max.) will be awarded to recognize exceptional performance.
- Grading rubrics: Overall, your answer will be evaluated based on its quality in three dimensions
  - Correctness and beauty of your outputs
  - Style of your code
  - Insightfulness of your interpretation or discussion
- Unless otherwise specified, you are required to use functions from the `tidyverse` package to complete this assignments.
- Fo some tasks, they may be multiple ways to achieve the same desired outcomes. You are encouraged to explore multiple methods. If you perform a task using multiple methods, do show it in your submission. You may earn bonus points for it.
- You are encouraged to use Generative AI such as ChatGPT to assist with your work. However, you will need to acknowledge it properly and validate AI's outputs. You may attach selected chat history with the AI you use and describe how it helps you get the work done. Extra credit may be rewarded to recognize creative use of Generative AI.
- This Problem Set is an individual assignment. You are expected to complete it independently. Clarification questions are welcome. Discussions on concepts and techniques related to the Problem Set among peers is encouraged. However, without the instructor's consent, sharing (sending and requesting) code and text that complete the entirety of a task is prohibited. You are strongly encouraged to use *CampusWire* for clarification questions and discussions.

Background

In 1998, Mexico had a close presidential election. Irregularities were detected around the country during the voting process. For example, when 2% of the vote tallies had been counted, the preliminary results showed the PRI's imminent defeat in Mexico City metropolitan area and a very narrow vote margin between PRI and FDN. A few minutes later, the screens at the Ministry of Interior went blank, an event that electoral authorities justified as a technical problem caused by an overload on telephone lines. The vote count was therefore suspended for three days, despite the fact that opposition representatives found a computer in the basement that continued to receive electoral results. Three days later, the vote count resumed, and soon the official announced PRI's winning with 50.4% of the vote.

What happened on that night and the following days? Were there electoral fraud during the election? A political scientist, Francisco Cantú, unearths a promising dataset that could provide some clues. At the National Archive in Mexico City, Cantú discovered about 53,000 vote tally sheets. Using machine learning methods, he detected that a significant number of tally sheets were *altered*! In addition, he found evidence that the altered tally sheets were biased in favor of the incumbent party. In this Problem Set, you will use Cantú's replication dossier to replicate and extend his data work.

Please read Cantú (2019) for the full story. And see Figure 1 for a few examples of altered (fraudulent) tallies.

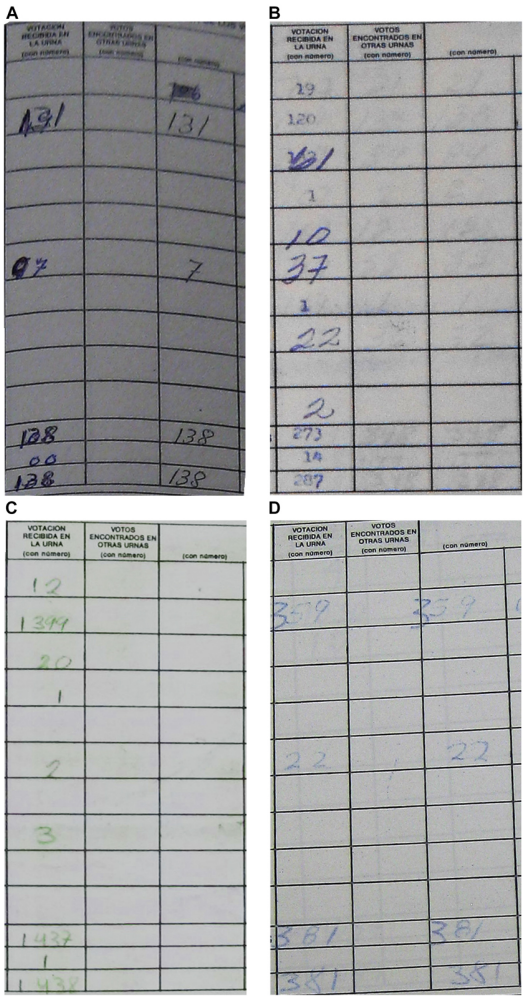


Figure 1: Examples of altered tally sheets (reproducing Figure 1 of Cantú 2018)

Task 0. Loading required packages (3pt)

For Better organization, it is a good habit to load all required packages up front at the start of your document. Please load the all packages you use throughout the whole Problem Set here.

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(patchwork)
library(sf)
library(ggthemes)
library(stringi)
```

Task 1. Clean machine classification results (3pt)

Cantú applys machine learning models to 55,334 images of tally sheets to detect signs of fraud (i.e., alteration). The machine learning model returns results recorded in a table. The information in this table is messy and requires data wrangling before we can use them.

Task 1.1. Load classified images of tally sheets

The path of the classified images of tally sheets is `data/classification.txt`. Your first task is loading these data onto R using a `tidyverse` function. Name it `d_tally`.

Note:

- Although the file extension of this dataset is `.txt`, you are recommended to use the `tidyverse` function we use for `.csv` files to read it.
- Unlike the data files we have read in class, this table has *no column names*. Look up the documentation and find a way to handle it.
- There will be three columns in this dataset, name them `name_image`, `label`, and `probability`.

Print your table to show your output.

```
d_tally <- read.csv("data/classification.txt", header=FALSE)
names(d_tally) <- c("name_image","label", "probability")
#print(d_tally)
```

Note 1. What are in this dataset?

Before you proceed, let me explain the meaning of the three variables.

- `name_image` contains the names of of the tallies' image files (as you may infer from the `.jpg` file extensions. They contain information about the locations where each of the tally sheets are produced.
- `label` is a machine-predicted label indicating whether a tally is fraudulent or not. `label = 1` means the machine learning model has detected signs of fraud in the tally sheet. `label = 0` means the machine detects no sign of fraud in the tally sheet. In short, `label = 1` means fraud; `label = 0` means no fraud.
- `probability` indicates the machine's certainty about its predicted `label` (explained above). It ranges from 0 to 1, where higher values mean higher level of certainty.

Interpret `label` and `probability` carefully. Two examples can hopefully give you clues about their correct interpretation. In the first row, `label = 0` and `probability = 0.9991`. That means the machine thinks this tally sheet is NOT FRAUDULENT with a probability of 0.9991. Then, the probability that this tally sheet is fraudulent is `1 - 0.9991 = 0.0009`. Take another example, in the 11th row, `label = 1` and `probability = 0.935`. This means the machine thinks this tally sheet IS FRAUDULENT with a probability of 0.935. Then, the probability that it is NOT FRAUDULENT is `1 - 0.9354 = 0.0646`.

Task 1.2. Clean columns `label` and `probability`

As you have seen in the printed outputs, columns `label` and `probability` are read as `chr` variables when they are actually numbers. A close look at the data may tell you why — they are "wrapped" by some non-numeric characters. In this task, you will clean these two variables and make them valid numeric variables. You are required to use `tidyverse` operations to for this task. Show appropriate summary statistics of `label` and `probability` respectively after you have transformed them into numeric variables.

```
d_tally <- d_tally |>
  mutate(label = str_remove_all(label, "\\[\\[|\\]\\]"),
         probability = str_remove_all(probability, "\\[\\[|\\]\\]"),
         label = as.numeric(label),
         probability = as.numeric(probability))
summary(d_tally)
```

```
##   name_image            label         probability
##  Length:55334       Min.   :0.0000   Min.   :0.5000
##  Class :character   1st Qu.:0.0000   1st Qu.:0.8185
##  Mode  :character   Median :0.0000   Median :0.9710
##                     Mean   :0.3623   Mean   :0.8926
##                     3rd Qu.:1.0000   3rd Qu.:0.9996
##                     Max.   :1.0000   Max.   :1.0000
```

Task 1.3. Extract state and district information from `name_image`

As explained in the note, the column `name_image`, which has the names of tally sheets' images, contains information about locations where the tally sheets are produced. Specifically, the first two elements of these file names indicates the **states'** and districts' identifiers respectively, for example, `name_image = "Aguascalientes_I_2014-05-26 00.00.10.jpg"`. It means this tally sheet is produced in state **`Aguascalientes`**, district `I`. In this task, you are required to obtain this information. Specifically, create two columns named `state` and `district` as state and district identifiers respectively. You are required to use `tidyverse` functions to perform the task.

```
#method--separate
d_tally <- d_tally |>
  separate(name_image, into = c("state", "district"), sep = "_", remove = FALSE) |>
  mutate(district = str_remove(district, ".jpg$"))

# method--stringr
d_tally <- d_tally |>
  mutate(split_names = str_split_fixed(name_image, "_", n = 3),
         state = split_names[, 1],
         district = split_names[, 2])|>
   select(-split_names)
summary(d_tally)
```

```
##   name_image           state              district             label
##  Length:55334       Length:55334       Length:55334       Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000
##                                                           Mean   :0.3623
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##   probability
##  Min.   :0.5000
##  1st Qu.:0.8185
##  Median :0.9710
##  Mean   :0.8926
##  3rd Qu.:0.9996
##  Max.   :1.0000
```

Task 1.4. Re-code a state's name

One of the states (in the newly created column `state`) is coded as "`Estado de Mexico`." The researchers decide that it should instead re-coded as "**`Edomex`**." Please use a `tidyverse` function to perform this task.

Hint: Look up functions `ifelse` and `case_match`.

```
#method1--ifelse
d_tally <- d_tally |>
  mutate(state = ifelse(state == "Estado de Mexico", "Edomex", state))
#method2--case_match
d_tally <- d_tally |>
  mutate(state = case_when(state == "Estado de Mexico" ~ "Edomex",TRUE ~ state))
summary(d_tally)
```

```
##   name_image           state              district            label
##  Length:55334       Length:55334       Length:55334       Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000
##                                                           Mean   :0.3623
##                                                           3rd Qu.:1.0000
##                                                           Max.   :1.0000
##   probability
##  Min.   :0.5000
##  1st Qu.:0.8185
##  Median :0.9710
##  Mean   :0.8926
##  3rd Qu.:0.9996
##  Max.   :1.0000
```

Task 1.5. Create a *probability of fraud* indicator

As explained in Note 1, we need to interpret `label` and `probability` with caution, as the meaning of `probability` is conditional on the value of `label`. To avoid confusion in the analysis, your next task is to create a column named `fraud_proba` which indicates the probability that a tally sheet is is fraudulent. After you have created the column, drop the `label` and `probability` columns.

Hint: Look up the `ifelse` function and the `case_when` function (but you just need either one of them).

```
d_tally <- d_tally |>
  mutate(fraud_proba = ifelse(label == 0, 1 - probability, probability)) |>
  select(-label, -probability)
```

Task 1.6. Create a binary *fraud* indicator

In this task, you will create a binary indicator called `fraud_bin` in indicating whether a tally sheet is fraudulent. Following the researcher's rule, we consider a tally sheet fraudulent only when the machine thinks it is at least 2/3 likely to be fraudulent. That is, `fraud_bin` is set to TRUE when `fraud_proba` is greater to `2/3` and is FALSE otherwise.

```
d_tally <- d_tally |>
  mutate(fraud_bin = if_else(fraud_proba >= 2/3, TRUE, FALSE))
```

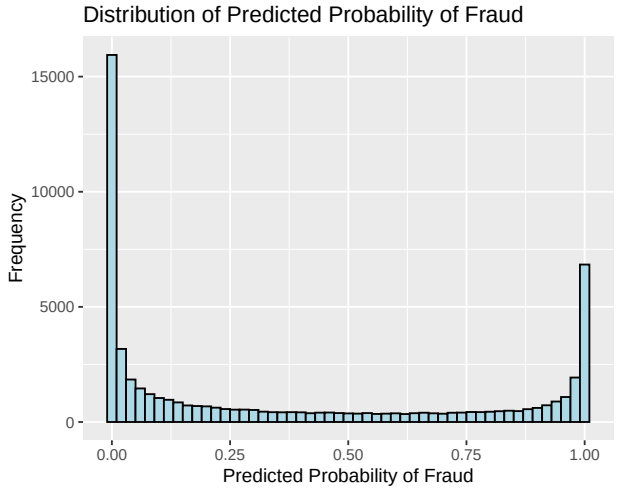
Task 2. Visualize machine classification results (3pt)

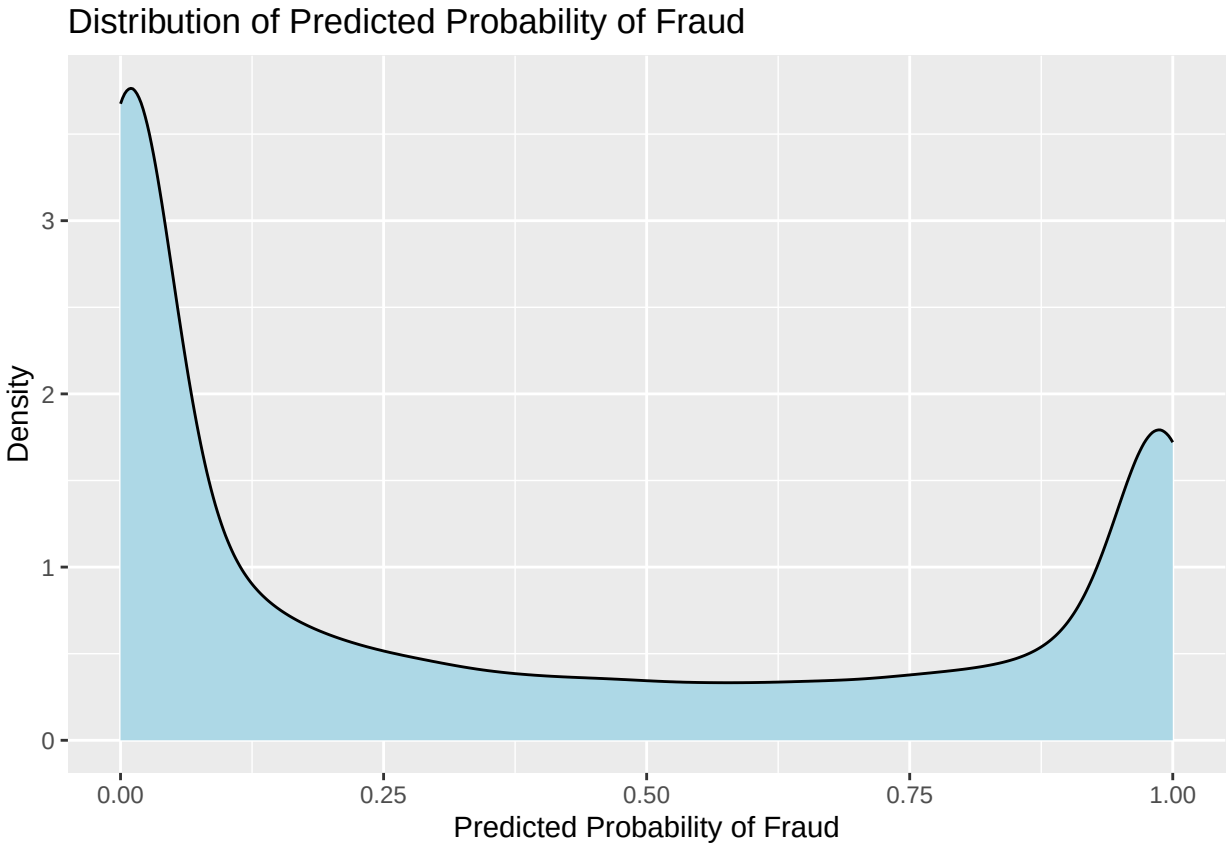
In this section, you will visualize the `tally` dataset that you have cleaned in Task 1. Unless otherwise specified, you are required to use the `ggplot` packages to perform all the tasks.

Task 2.1. Visualize distribution of `fraud_proba`

How is the predicted probability of fraud (`fraud_proba`) distributed? Use two methods to visualize the distribution. Remember to add informative labels to the figure. Describe the plot with a few sentences.

```
ggplot(d_tally, aes(x = fraud_proba)) +
  geom_histogram(binwidth = 0.02, fill = "lightblue", color = "black") +
  labs(title = "Distribution of Predicted Probability of Fraud",
       x = "Predicted Probability of Fraud",
       y = "Frequency")
```





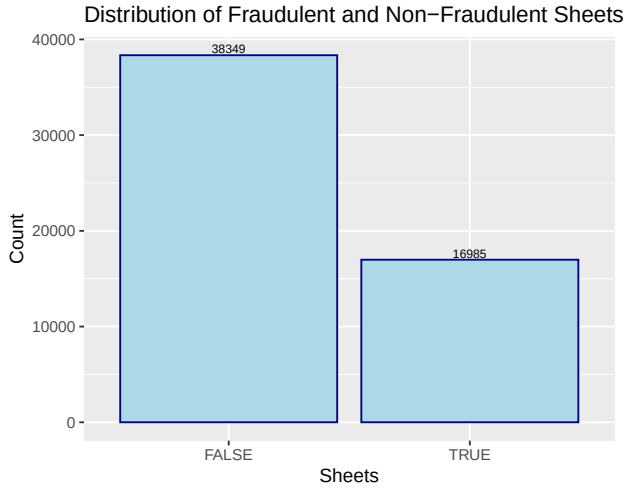
From the first plot, we know that the possibility distribution in the middle is relatively even, and it becomes larger at the ends of true and fraudulent. The distribution of probabilities is highest for extremely likely to be true, which may exceed 15,000 ballots, and second for extremely likely to be fraudulent, which may exceed 5,000 ballots.

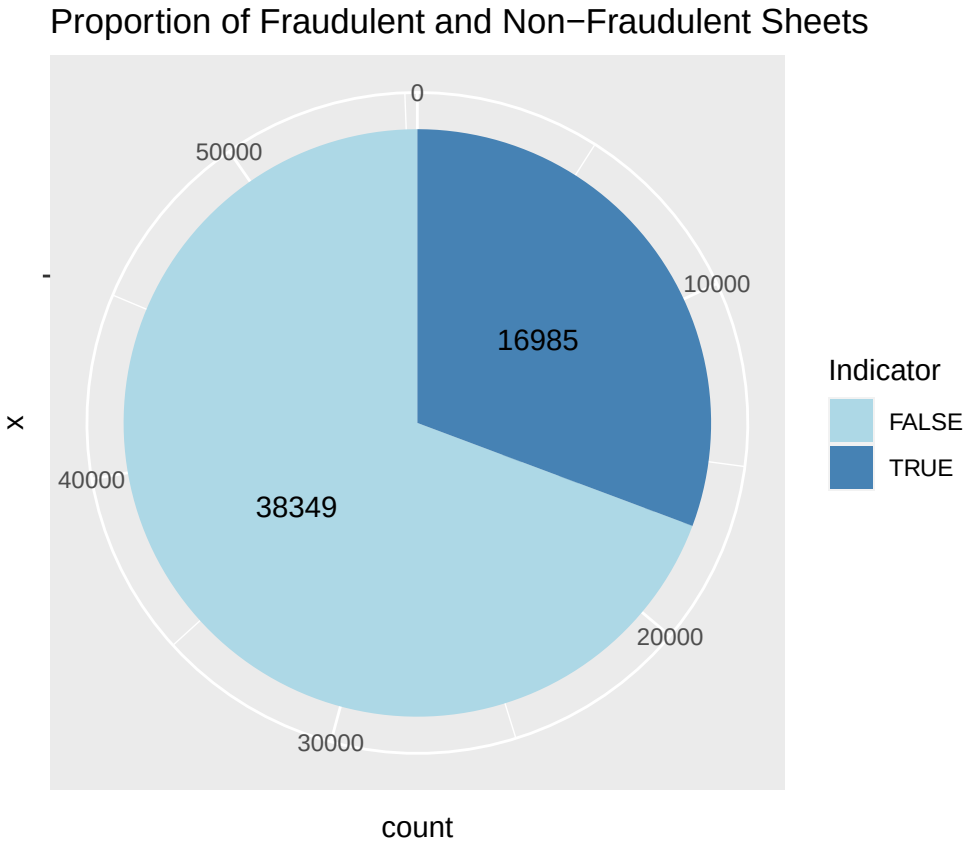
From the second plot, we know that there are two high peaks which indicate a higher concentration of probabilities in true tally sheets and fraudulent tally sheets. Also, the density of probabilities of true sheets is higher than Fraudulent ones.

Task 2.2. Visualize the distribution of `fraud_bin`

How many tally sheets are fraudulent and how many are not? We may answer this question by visualizing the binary indicator of tally-level states of fraud. Use at least two methods to visualize the distribution of `fraud_bin`. Remember to add informative labels to the figure. Describe your plots with a few sentences.

```
d_fraud <- d_tally |>
  group_by(fraud_bin)|>
  summarise(count = n())
ggplot(d_fraud, aes(x = fraud_bin, y = count)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "navy") +
  geom_text(aes(label = count), vjust = -0.2, color = "black", size = 2.5) +
  labs(title = "Distribution of Fraudulent and Non-Fraudulent Sheets",
       x="Sheets",
       y = "Count")
```





From the plots, we know that the fraudulent sheets are 38349 and the non-fraudulent sheets are 16985. Also, fraudulent sheets are much more than non-fraudulent sheets.

Task 2.3. Summarize prevalence of fraud by state

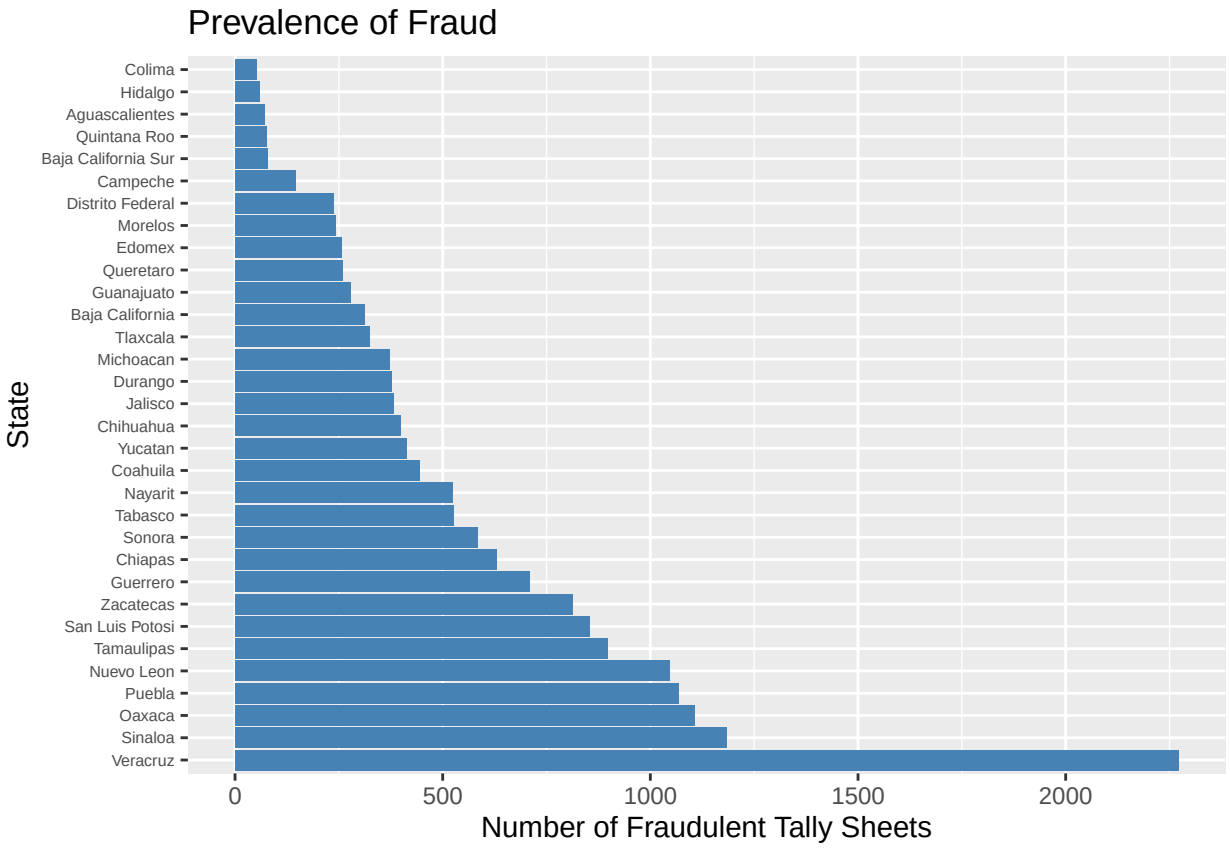
Next, we will examine the between-state variation with regards to the prevalence of election fraud. In this task, you will create a new object that contains two state-level indicators regarding the prevalence of election fraud: The count of fraudulent tallies and the proportion of fraudulent tallies.

```
d_state <- d_tally |>
  group_by(state) |>
  summarise(n_fraud = sum(fraud_bin == 1),
            prop_fraud = mean(fraud_bin == 1) * 100)
```

Task 2.4. Visualize frequencies of fraud by state

Using the new data frame created in Task 2.3, please visualize the *frequencies* of fraudulent tallies of every state. Describe the key takeaway from the visualization with a few sentences.

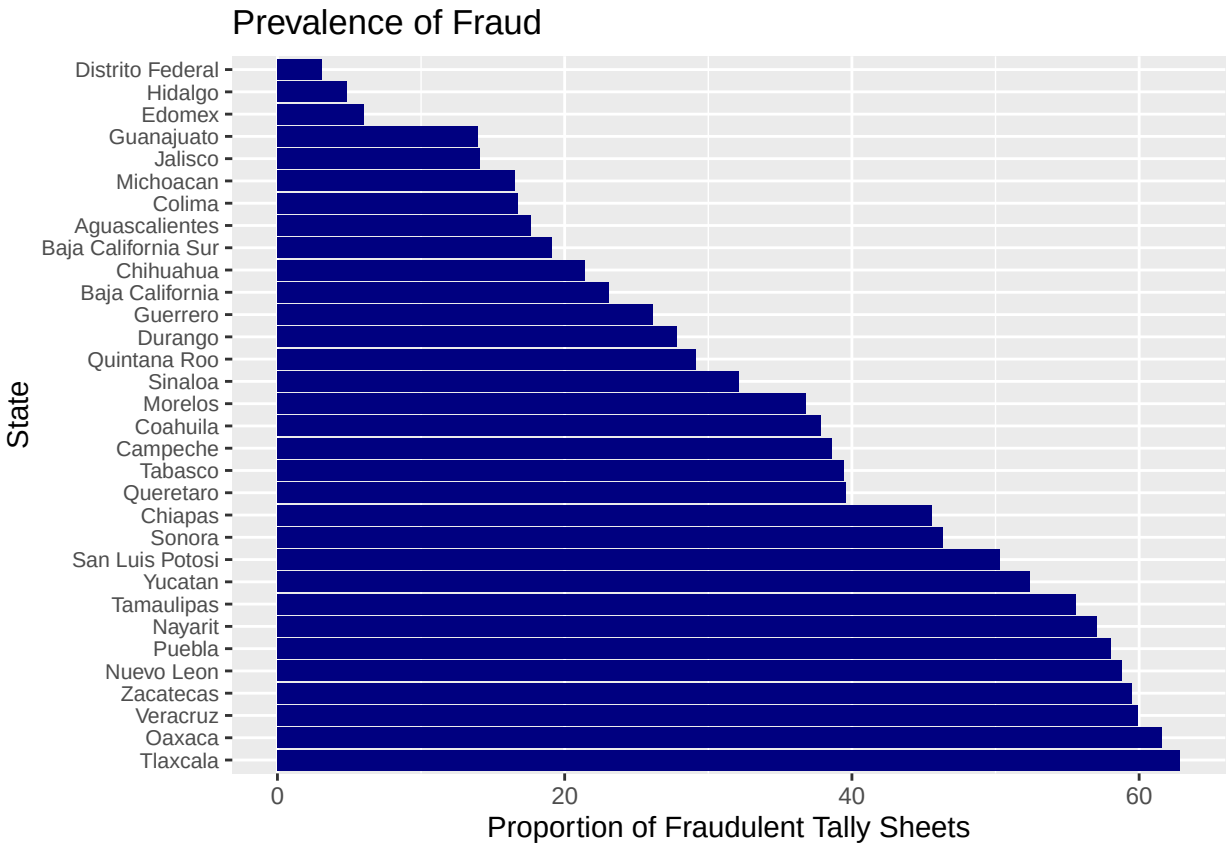
Feel free to try alternative approach(es) to make your visualization nicer and more informative.



Task 2.5. Visualize proportions of fraud by state

Using the new data frame created in Task 2.3, please visualize the *proportion of* of fraudulent tallies of every state. Describe the key takeaway from the visualization with a few sentences.

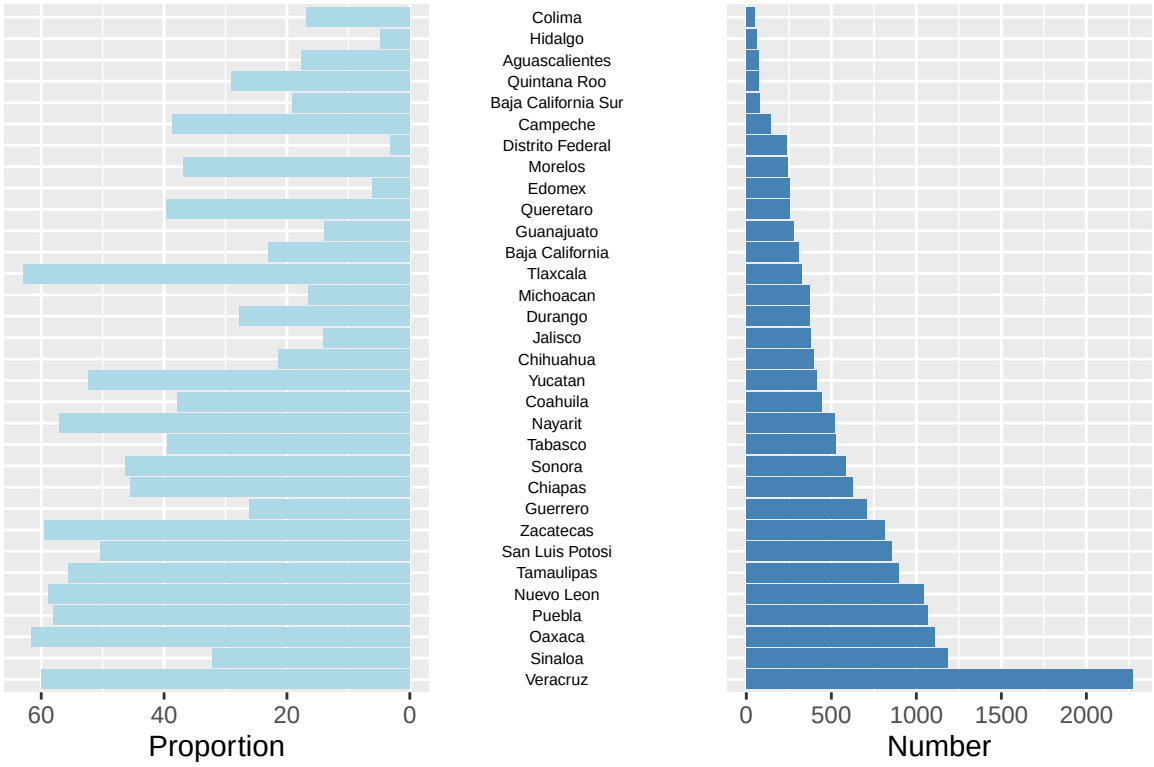
Feel free to try alternative approach(es) to make your visualization nicer and more informative.



Task 2.6. Visualize both proportions & frequencies of fraud by state

Create data visualization to show BOTH the *proportions* and *frequencies* of fraudulent tally sheets by state in one figure. Include annotations to highlight states with the highest level of fraud. Add informative labels to the figure. Describe the takeaways from the figure with a few sentences.

```
plot1 <- ggplot(d_state, aes(x = n_fraud, y = reorder(state, -n_fraud))) +
  geom_bar(aes(fill = "Frequency"), stat = "identity", position = "identity") +
  scale_fill_manual(values = "steelblue") +
  labs(x = "Number", y = "",
       title = "") +
  theme(axis.text.y = element_blank(), axis.ticks.y = element_blank(),
        legend.title = element_blank(), legend.position = "none")

plot2 <- ggplot(d_state, aes(x = prop_fraud, y = reorder(state, -n_fraud))) +
  geom_bar(aes(fill = "Proportion"), stat = "identity", position = "identity") +
  scale_fill_manual(values = "lightblue") +
  labs(x = "Proportion", y = "",
       title = "") +
  theme(axis.text.y = element_blank(), axis.ticks.y = element_blank(),
        legend.title = element_blank(), legend.position = "none")

plot2 <- plot2 + scale_x_reverse()

state_names <- ggplot(d_state, aes(x = 0, y = reorder(state, -n_fraud), label = state)) +
  geom_text(hjust = 0.5, size = 2) +
  theme_void()

combined_plot <- plot2 + state_names + plot1 +
  plot_layout(ncol = 3, widths = c(1.5, 0.8, 1.5))

combined_plot
```



One can see that Veracruz has the highest number of frauds and Tlaxcala has the highest level of proportion of frauds.

Task 3. Clean vote return data (3pt)

Your next task is to clean a different dataset from the researchers' replication dossier. Its path is `data/Mexican_Election_Fraud/dataverse/VoteReturns.csv`. This dataset contains information about vote returns recorded in every tally sheet. This dataset is essential for the replication of Figure 4 in the research article.

Task 3.1. Load vote return data

Load the dataset onto your R environment. Name this dataset `d_return`. Show summary statistics of this dataset and describe the takeaways using a few sentences.

```
d_return <- read.csv("data/VoteReturns.csv")
summary(d_return)
```

```
##      foto              seccion            casilla              dtto
##  Length:53499       Length:53499       Length:53499       Length:53499
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       dto            municipio             edo              entidad
##  Min.   :  1.000   Length:53499       Length:53499       Length:53499
##  1st Qu.:  3.000   Class :character   Class :character   Class :character
##  Median :  6.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :  8.704
##  3rd Qu.: 10.000
##  Max.   :341.000
##  NA's   :4
##     pagina                p1                 p2                p3
##  Length:53499       Min.   :     0.0   Min.   :    0.0   Min.   :   0.0
##  Class :character   1st Qu.:   250.0   1st Qu.:   67.0   1st Qu.:  98.0
##  Mode  :character   Median :   530.0   Median :  245.0   Median : 233.0
##                     Mean   :   671.9   Mean   :  343.3   Mean   : 319.3
##                     3rd Qu.:   941.5   3rd Qu.:  482.0   3rd Qu.: 442.0
##                     Max.   :364105.0   Max.   :48225.0   Max.   :9127.0
##                                                          NA's   :1
##        p4                p5                pan               pri
##  Min.   :    0.0   Min.   :   0.00   Min.   :   0.00   Min.   :   0.0
##  1st Qu.:   73.0   1st Qu.:   0.00   1st Qu.:   2.00   1st Qu.:  52.0
##  Median :  222.0   Median :  13.00   Median :  18.00   Median : 107.0
##  Mean   :  369.7   Mean   :  29.36   Mean   :  56.88   Mean   : 162.7
##  3rd Qu.:  464.0   3rd Qu.:  36.00   3rd Qu.:  72.00   3rd Qu.: 195.0
##  Max.   :21265.0   Max.   :6650.00   Max.   :4436.00   Max.   :6080.0
##
##       pps               psm               pms              pfcrn
##  Min.   :   0.00   Min.   :   0.000   Min.   :   0.00   Min.   :   0.00
##  1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:   0.00   1st Qu.:   0.00
##  Median :   9.00   Median :   1.000   Median :   2.00   Median :  11.00
##  Mean   :  35.04   Mean   :   3.637   Mean   :  12.19   Mean   :  34.17
##  3rd Qu.:  47.00   3rd Qu.:   3.000   3rd Qu.:  13.00   3rd Qu.:  45.00
```

```
##  Max.   :1056.00   Max.   :1802.000   Max.   :5511.00   Max.   :1011.00
##
##       prt              parm            noregis           nombrenore
##  Min.   :  0.000   Min.   :   0.00   Min.   :   0.0000   Length:53499
##  1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:   0.0000   Class :character
##  Median :  0.000   Median :   5.00   Median :   0.0000   Mode  :character
##  Mean   :  1.912   Mean   :  20.44   Mean   :   0.8175
##  3rd Qu.:  1.000   3rd Qu.:  23.00   3rd Qu.:   0.0000
##  Max.   :592.000   Max.   :1170.00   Max.   :1604.0000
##                                      NA's   :1
##      otros            otroscan             pan2              pri2
##  Min.   :   0.000   Length:53499       Min.   :   0.000   Min.   :   0.000
##  1st Qu.:   0.000   Class :character   1st Qu.:   0.000   1st Qu.:   0.000
##  Median :   0.000   Mode  :character   Median :   0.000   Median :   0.000
##  Mean   :   3.171                      Mean   :   1.475   Mean   :   3.941
##  3rd Qu.:   0.000                      3rd Qu.:   0.000   3rd Qu.:   0.000
##  Max.   :1734.000                      Max.   :1239.000   Max.   :2651.000
##  NA's   :4
##       pps2              psm2              pms2              pfcrn2
##  Min.   :  0.0000   Min.   :  0.000   Min.   :  0.0000   Min.   :   0.0000
##  1st Qu.:  0.0000   1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:   0.0000
##  Median :  0.0000   Median :  0.000   Median :  0.0000   Median :   0.0000
##  Mean   :  0.7557   Mean   :  0.116   Mean   :  0.3039   Mean   :   0.7968
##  3rd Qu.:  0.0000   3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:   0.0000
##  Max.   :680.0000   Max.   :429.000   Max.   :427.0000   Max.   :1319.0000
##
##       prt2             parm2             noregis2            otro2
##  Min.   :  0.000   Min.   :  0.0000   Min.   :  0.00000   Min.   : 0.000000
##  1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.:  0.00000   1st Qu.: 0.000000
##  Median :  0.000   Median :  0.0000   Median :  0.00000   Median : 0.000000
##  Mean   :  0.073   Mean   :  0.5122   Mean   :  0.01837   Mean   : 0.002935
##  3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.:  0.00000   3rd Qu.: 0.000000
##  Max.   :429.000   Max.   :429.0000   Max.   :259.00000   Max.   :26.000000
##
##       pan3              pri3             pps3              psm3
##  Min.   :   0.00   Min.   :   0.0   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:   0.00   1st Qu.:   0.0   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :   0.00   Median :  32.0   Median :  0.00   Median :  0.000
##  Mean   :  39.36   Mean   :  93.5   Mean   : 22.08   Mean   :  2.094
##  3rd Qu.:  45.00   3rd Qu.: 127.0   3rd Qu.: 21.00   3rd Qu.:  1.000
##  Max.   :2194.00   Max.   :6080.0   Max.   :921.00   Max.   :856.000
##                    NA's   :1                         NA's   :2
##       pms3              pfcrn3            prt3              parm3
##  Min.   :   0.000   Min.   :  0.00   Min.   :  0.000   Min.   :   0.00
##  1st Qu.:   0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:   0.00
##  Median :   0.000   Median :  0.00   Median :  0.000   Median :   0.00
##  Mean   :   7.803   Mean   : 21.63   Mean   :  1.077   Mean   :  12.68
##  3rd Qu.:   5.000   3rd Qu.: 23.00   3rd Qu.:  1.000   3rd Qu.:  11.00
##  Max.   :8932.000   Max.   :992.00   Max.   :413.000   Max.   :1170.00
##  NA's   :1          NA's   :1
##     noregis3           otro3               suma            nulos
##  Min.   :  0.0000   Min.   :   0.0000   Min.   :   0.0   Min.   :   0.00
##  1st Qu.:  0.0000   1st Qu.:   0.0000   1st Qu.:  82.0   1st Qu.:   0.00
##  Median :  0.0000   Median :   0.0000   Median : 217.0   Median :   3.00
```

```
##  Mean   :  0.3498   Mean   :   0.3016   Mean   : 296.4   Mean   :  21.93
##  3rd Qu.:  0.0000   3rd Qu.:   0.0000   3rd Qu.: 420.0   3rd Qu.:  11.00
##  Max.   :747.0000   Max.   :1353.0000   Max.   :9962.0   Max.   :8770.00
##                     NA's   :1           NA's   :1        NA's   :1
##      total             suma1              nulos1            total1
##  Min.   :    0.0   Min.   :   0.000   Min.   :   0.000   Min.   :   0.000
##  1st Qu.:   90.0   1st Qu.:   0.000   1st Qu.:   0.000   1st Qu.:   0.000
##  Median :  229.0   Median :   0.000   Median :   0.000   Median :   0.000
##  Mean   :  315.7   Mean   :   4.865   Mean   :   0.635   Mean   :   7.175
##  3rd Qu.:  440.0   3rd Qu.:   0.000   3rd Qu.:   0.000   3rd Qu.:   0.000
##  Max.   :16811.0   Max.   :3333.000   Max.   :1600.000   Max.   :2787.000
##  NA's   :1         NA's   :2          NA's   :2          NA's   :2
##      suma2             nulos2            total2         inciden
##  Min.   :   0.0   Min.   :   0.00   Min.   :   0.0   Length:53499
##  1st Qu.:   0.0   1st Qu.:   0.00   1st Qu.:   0.0   Class :character
##  Median :   0.0   Median :   0.00   Median :   0.0   Mode  :character
##  Mean   : 176.9   Mean   :  11.38   Mean   : 192.6
##  3rd Qu.: 280.0   3rd Qu.:   5.00   3rd Qu.: 299.0
##  Max.   :7633.0   Max.   :7734.00   Max.   :9855.0
##  NA's   :2        NA's   :2         NA's   :2
##  representante_pan  representante_pri  representante_pps  representante_pms
##  Length:53499       Length:53499       Length:53499       Length:53499
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  representante_psm  representante_pfcrn representante_prt  representante_parm
##  Length:53499       Length:53499        Length:53499       Length:53499
##  Class :character   Class :character    Class :character   Class :character
##  Mode  :character   Mode  :character    Mode  :character   Mode  :character
##
##
##
##
##  protesta_pan       protesta_pri       protesta_pps       protesta_pms
##  Length:53499       Length:53499       Length:53499       Length:53499
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  protesta_psm       protesta_pfcrn     protesta_prt       protesta_parm
##  Length:53499       Length:53499       Length:53499       Length:53499
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  protesta_otro       presidente         secretario           primer
##  Length:53499       Length:53499       Length:53499       Length:53499
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    segundo            observa              var79           salinas
##  Length:53499       Length:53499       Min.   :   1.0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:   1.0   1st Qu.:  63.0
##  Mode  :character   Mode  :character   Median :   1.0   Median : 115.0
##                                        Mean   : 131.2   Mean   : 174.4
##                                        3rd Qu.:   2.0   3rd Qu.: 206.0
##                                        Max.   :9999.0   Max.   :6080.0
##                                        NA's   :53422
##    clouthier           ibarra           castillo        ppsccs
##  Min.   :   0.00   Min.   :  0.000   Min.   :   0   Min.   :   0.00
##  1st Qu.:   3.00   1st Qu.:  0.000   1st Qu.:   0   1st Qu.:   1.00
##  Median :  23.00   Median :  0.000   Median :   1   Median :  12.00
##  Mean   :  61.37   Mean   :  2.185   Mean   :   4   Mean   :  37.67
##  3rd Qu.:  78.00   3rd Qu.:  2.000   3rd Qu.:   3   3rd Qu.:  51.00
##  Max.   :4436.00   Max.   :592.000   Max.   :1802   Max.   :1056.00
##
##     pfcrnccs          parmccs            nrccs             noregccs
##  Min.   :   0.00   Min.   :   0.00   Min.   :0.000000   Min.   :   0.0000
##  1st Qu.:   1.00   1st Qu.:   0.00   1st Qu.:0.000000   1st Qu.:   0.0000
##  Median :  14.00   Median :   6.00   Median :0.000000   Median :   0.0000
##  Mean   :  36.85   Mean   :  21.98   Mean   :0.006654   Mean   :   0.1439
##  3rd Qu.:  48.00   3rd Qu.:  25.00   3rd Qu.:0.000000   3rd Qu.:   0.0000
##  Max.   :1319.00   Max.   :1170.00   Max.   :1.000000   Max.   :1125.0000
##
##       occs           otrosccs           cardenas
##  Min.   :0.0000   Min.   :   0.000   Min.   :   0.00
##  1st Qu.:1.0000   1st Qu.:   0.000   1st Qu.:  10.00
##  Median :1.0000   Median :   0.000   Median :  53.00
##  Mean   :0.9942   Mean   :   3.106   Mean   :  99.75
##  3rd Qu.:1.0000   3rd Qu.:   0.000   3rd Qu.: 141.00
##  Max.   :1.0000   Max.   :1734.000   Max.   :2280.00
##
```

The dataset contains information about various variables related to elections in different states and districts.

Note 2. What are in this dataset?

This table contains a lot of different variables. The researcher offers no comprehensive documentation to tell us what every column means. For the sake of this problem set, you only need to know the meanings of the following columns:

- `foto` is an identifier of the images of tally sheets in this dataset. We will need it to merge this dataset with the `d_tally` data.
- `edo` contains the names of states.
- `dto` contains the names of districts (in Arabic numbers).
- `salinas`, `clouthier`, and `ibarra` contain the counts of votes (as recorded in the tally sheets) for presidential candidates Salinas (PRI), Cardenas (FDN), and Clouthier (PAN). In addition, the summation of all three makes the total number of **presidential votes**.
- `total` contains the total number of **legislative votes**.

Task 3.2. Recode names of states

A state whose name is `Chihuahua` is mislabelled as `Chihuhua`. A state whose name is currently `Edomex` needs to be recoded to `Estado de Mexico`. Please re-code the names of these two states accordingly.

```
d_return$edo <- ifelse(d_return$edo == "Chihuhua", "Chihuahua", d_return$edo)
d_return$edo <- ifelse(d_return$edo == "Edomex", "Estado de Mexico", d_return$edo)
```

Task 3.3. Recode districts' identifiers

Compare how districts' identifiers are recorded differently in the tally (`d_tally`) from vote return (`d_return`) datasets. Specifically, in the `d_tally` dataset, `district` contains Roman numbers while in the `d_return` dataset, `dto` contains Arabic numbers. Recode districts' identifiers in the `d_return` dataset to match those in the `d_tally` dataset. To complete this task, first summarize the values of the two district identifier columns in the two datasets respectively to verify the above claim. Then do the requested conversion.

```
d_tally|>
  group_by(district)|>
  count()
```

```
## # A tibble: 40 x 2
## # Groups:   district [40]
##    district     n
##    <chr>    <int>
##  1 I         6218
##  2 II        6251
##  3 III       5065
##  4 IV        4513
##  5 IX        2490
##  6 V         5101
##  7 VI        4246
##  8 VII       3262
##  9 VIII      2956
## 10 X         1904
## # i 30 more rows
```

```
d_return|>
  group_by(dto)|>
  count()
```

```
## # A tibble: 42 x 2
## # Groups:   dto [42]
##      dto     n
##    <int> <int>
##  1     1  5976
##  2     2  6095
##  3     3  4865
##  4     4  4217
##  5     5  4942
##  6     6  4127
##  7     7  3008
##  8     8  2782
##  9     9  2524
## 10    10  1875
## # i 32 more rows
```

```
d_return$dto <- as.roman(d_return$dto)
```

Task 3.4. Create a `name_image` identifier for the `d_return` dataset

In the `d_return` dataset, create a column named `name_image` as the first column. The column concatenate values in the three columns: `edo`, `dto`, and `foto` with an underscore `_` as separators.

```
d_return <- d_return |>
  mutate(name_image = paste(edo, dto, foto, sep = "_"))|>
  select(name_image, edo, dto, foto, everything())
```

Task 3.5. Wrangle the `name_image` column in two datasets

As a final step before merging `d_return` and `d_tally`, you are required to perform the following data wrangling. For the `name_image` column in BOTH `d_return` and `d_tally`:

- Convert all characters to lower case.
- Remove ending substring `.jpg`.

```
d_return <- d_return |>
  mutate(name_image = tolower(name_image)) |>
  mutate(name_image = str_remove_all(name_image, "\\.jpg$"))

d_tally <- d_tally |>
  mutate(name_image = tolower(name_image)) |>
  mutate(name_image = str_remove_all(name_image, "\\.jpg$"))
```

Task 3.6 Join classification results and vote returns

After you have successfully completed all the previous steps, join `d_return` and `d_tally` by column `name_image`. This task contains two part. First, use appropriate `tidyverse` functions to answer the following questions:

- How many rows are in `d_return` but not in `d_tally`? Which states and districts are they from?
- How many rows are in `d_tally` but not in `d_return`? Which states and districts are they from?

```
#Revise "Edomex"
d_return <- d_return |>
  mutate(edo = ifelse(edo == "Estado de Mexico", "Edomex", edo))

#rows in d_return but not in d_tally
rows_only_in_return <- anti_join(d_return, d_tally, by = "name_image")
num_rows_only_in_return <- nrow(rows_only_in_return)
states_districts_only_in_return <- rows_only_in_return |>
  select(edo, dto) |>
  distinct()

#rows in d_tally but not in d_return
rows_only_in_tally <- anti_join(d_tally, d_return, by = "name_image")
num_rows_only_in_tally <- nrow(rows_only_in_tally)
states_districts_only_in_tally <- rows_only_in_tally |>
  select(state, district) |>
  distinct()

#which states and districts are they from?
```

We can see that there are 210 rows in in d_return but not in d_tally and 2368 rows in d_tally but not in d_return.

Second, create a dataset call `d` by joining `d_return` and `d_tally` by column `name_image`. `d` contains rows whose identifiers appear in *both* datasets and columns from *both* datasets.

```
d_return <- d_return |>
  mutate(dto = as.character(dto))
d <- inner_join(d_tally, d_return,
                by = c("name_image", "state"="edo", "district"= "dto"))

# Print the first few rows of the merged dataset
head(d)
```

```
##                            name_image          state district fraud_proba
## 1 aguascalientes_i_2014-05-26 00.00.10 Aguascalientes        I  0.00080401
## 2 aguascalientes_i_2014-05-26 00.00.17 Aguascalientes        I  0.04277194
## 3 aguascalientes_i_2014-05-26 00.00.25 Aguascalientes        I  0.42309284
## 4 aguascalientes_i_2014-05-26 00.00.31 Aguascalientes        I  0.03494918
## 5 aguascalientes_i_2014-05-26 00.00.38 Aguascalientes        I  0.13024312
## 6 aguascalientes_i_2014-05-26 00.00.45 Aguascalientes        I  0.21174937
##   fraud_bin                foto seccion casilla dtto      municipio    entidad
## 1     FALSE 2014-05-26 00.00.10       1      84      AGUASCALIENTES AGUASCALIE
## 2     FALSE 2014-05-26 00.00.17      85      85    1 AGUASCALIENTES AGUASCALIE
```

```
## 3     FALSE 2014-05-26 00.00.25      45    45-A    1 AGUASCALIENTES       AGUA
## 4     FALSE 2014-05-26 00.00.31      86      86    1 AGUASCALIENTES      AGUAS
## 5     FALSE 2014-05-26 00.00.38      87      87    1                        1
## 6     FALSE 2014-05-26 00.00.45       1    87-A    7 AGUASCALIENTES      AGUAS
##   pagina   p1  p2  p3  p4 p5 pan pri pps psm pms pfcrn prt parm noregis
## 1    128  919 453 497 497 45 263 167  28   5   5    13   0    7     488
## 2    129  795 264 545 483 61 306 165  23  11  12     8   2    5       0
## 3    130  767 450 316 316  0 192  88  10   1   1     8   1   10       0
## 4    131 1243 578 666 614 60 432 173  19   2   4    10   1   14       0
## 5    132  718 333 384 349 35 181 145  15   6   4    12   1    7       0
## 6    133  710 299 411 411 31   0   0   0   0   0     0   0    0       0
##   nombrenore otros otroscan pan2 pri2 pps2 psm2 pms2 pfcrn2 prt2 parm2 noregis2
## 1                0            0    0    0    0    0      0    0     0        0
## 2                0            0    0    0    0    0      0    0     0        0
## 3                0            0    0    0    0    0      0    0     0        0
## 4                0            0    0    0    0    0      0    0     0        0
## 5                0            0    0    0    0    0      0    0     0        0
## 6                0            0    0    0    0    0      0    0     0        0
##   otro2 pan3 pri3 pps3 psm3 pms3 pfcrn3 prt3 parm3 noregis3 otro3 suma nulos
## 1     0  263  167   28    5    5     13    0     0      488     0  488     9
## 2     0  306  165   23   11   12      8    2     5        0     0  532    13
## 3     0  192   88   10    1    1      8    1    10        0     0  311     5
## 4     0    0    0    0    0    0      0    0     0        0     0  655    11
## 5     0  181  145   15    6    4     12    1     7        0     0  371    13
## 6     0  170  170   21    4   15     14    1     7        0     0    0     0
##   total suma1 nulos1 total1 suma2 nulos2 total2 inciden representante_pan
## 1   497     0      0      0   488      9    497 NINGUNA               Si
## 2   545     0      0      0   532     13    545 NINGUNO               Si
## 3   316     0      0      0   311      5    316                       Si
## 4   666     0      0      0     0      0      0                       Si
## 5   184     0      0      0   371     13    184 NINGUNO               Si
## 6     0     0      0      0   402      9    411                       Si
##   representante_pri representante_pps representante_pms representante_psm
## 1                Si                No                No                Si
## 2                No                No                No                No
## 3                Si                Si                No                Si
## 4                Si                No                No                No
## 5                Si                Si                No                Si
## 6                Si                No                No                Si
##   representante_pfcrn representante_prt representante_parm protesta_pan
## 1                  No                No                 No          Si
## 2                  No                No                 No          No
## 3                  No                No                 No          No
## 4                  Si                No                 No          No
## 5                  Si                No                 No          No
## 6                  Si                No                 No          No
##   protesta_pri protesta_pps protesta_pms protesta_psm protesta_pfcrn
## 1           Si           No           No           Si             No
## 2           No           No           No           No             No
## 3           No           No           No           No             No
## 4           No           No           No           No             No
## 5           No           No           No           No             No
## 6           No           No           No           No             No
##   protesta_prt protesta_parm protesta_otro presidente secretario primer segundo
```

```
## 1           No            No            No         Si         Si     Si      Si
## 2           No            No            No         Si         Si     Si      No
## 3           No            No            No         Si         Si     Si      Si
## 4           No            No            No         Si         Si     Si      Si
## 5           No            No            No         Si         Si     Si      Si
## 6           No            No            No         Si         Si     Si      Si
##                            observa var79 salinas clouthier ibarra castillo
## 1 EL DISTRITO FEDERAL NO ES LEGIBLE    NA     167       263      0        5
## 2                                   1    NA     165       306      2       11
## 3                                        NA      88       192      1        1
## 4                                        NA     173       432      1        2
## 5                                        NA     145       181      1        6
## 6                                        NA     170       170      1        4
##   ppsccs pfcrnccs parmccs nrccs noregccs occs otrosccs cardenas
## 1     28       13       7     0        0    1        0       48
## 2     23        8       5     0        0    1        0       36
## 3     10        8      10     0        0    1        0       28
## 4     19       10      14     0        0    1        0       43
## 5     15       12       7     0        0    1        0       34
## 6     21       14       7     0        0    1        0       42
```

Task 4. Visualize distributions of fraudulent tallies across candidates (6pt)

In this task, you will visualize the distributions of fraudulent tally sheets across three presidential candidates: **Sarinas (PRI)**, **Cardenas (FDN)**, and **Clouthier (PAN)**. The desired output of is reproducing and extending Figure 4 in the research article (Cantu 2019, pp. 720).

Task 4.1. Calculate vote proportions of Salinas, Clouthier, and Cardenas

Before getting to the visualization, you should first calculate the proportion of votes (among all) received by the three candidates of interest. As additional background information, there are two more presidential candidates in this election, whose votes received are recorded in `ibarra` and `castillo` respectively. Please perform the tasks in the following two steps on the `d` dataset:

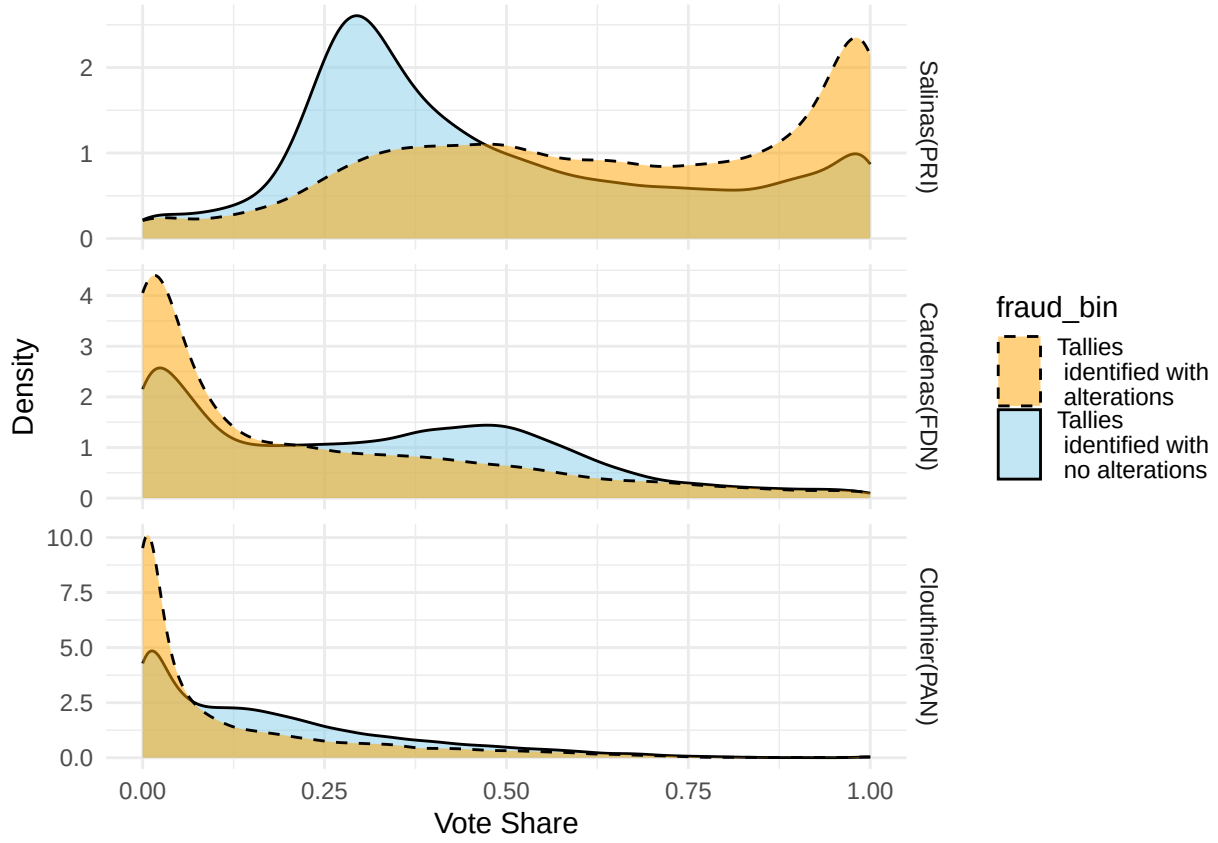
- Create a new column named `total_president` as an indicator of the total number of votes of the 5 presidential candidates.
- Create three columns `salinas_prop`, `cardenas_prop`, and `clouthier_prop` that indicate the proportions of the votes these three candidates receive respectively.

```
d <- d |>
  mutate(total_president = salinas + cardenas + clouthier + ibarra + castillo)
d <- d |>
  mutate(salinas_prop = salinas / total_president,
         cardenas_prop = cardenas / total_president,
         clouthier_prop = clouthier / total_president)
```

Task 4.2. Replicate Figure 4

Based on all the previous step, reproduce Figure 4 in Cantu (2019, pp. 720).

```
v <- d |> select(salinas_prop, cardenas_prop, clouthier_prop, fraud_bin)
v <- v |>
  rename("Salinas(PRI)" = salinas_prop,
         "Cardenas(FDN)" = cardenas_prop,
         "Clouthier(PAN)" = clouthier_prop)
v <- v |>
  pivot_longer(cols = c("Salinas(PRI)", "Cardenas(FDN)", "Clouthier(PAN)"),
               names_to = "candidate")
v$candidate <- factor(v$candidate, levels = c("Salinas(PRI)", "Cardenas(FDN)", "Clouthier(PAN)"))

plot_faceted <- ggplot(v, aes(x = value, fill = fraud_bin, linetype = fraud_bin)) +
geom_density(alpha = 0.5) +
facet_grid(candidate ~ ., scales = "free_y") +
labs(x = "Vote Share", y="Density")+
scale_fill_manual(values = c("orange", "skyblue"),
labels = c("Tallies \n identified with \n alterations", "Tallies \n identified with \n no alterations")
breaks = c(TRUE, FALSE)) +
scale_linetype_manual(values = c("dashed", "solid"),
labels = c("Tallies \n identified with \n alterations", "Tallies \n identified with \n no alterations")
breaks = c(TRUE, FALSE)) +
theme_minimal()

plot_faceted
```



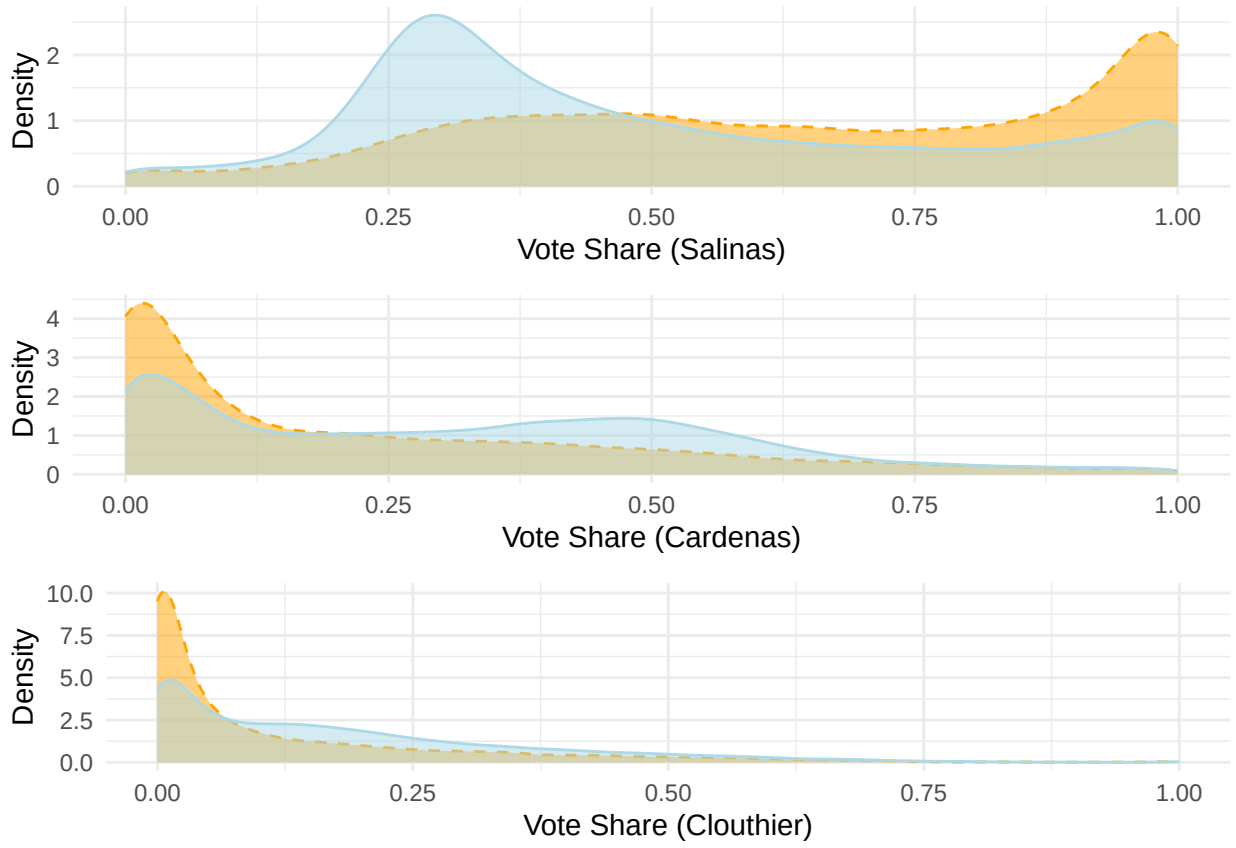
Note: Your performance in this task will be mainly evaluated based on your output's similarity with the original figure. Pay attention to the details. For your reference, below is a version created by the instructor.

Task 4.3. Discuss and extend the reproduced figure

Referring to your reproduced figures and the research articles, in what way is the researcher's argument supported by this figure? Make an alternative visualization design that can substantiate and even augment the current argument. After you have shown your alternative design, in a few sentences, describe how your design provides visual aid as effectively as or more effectively than the original figure.

Note: Feel free to make *multiple* alternative designs to earn bonus credits. However, please be selective. Only a design with major differences from the existing ones can be counted as an alternative design.

```
library(ggplot2)
library(gridExtra)

# Filter TRUE and FALSE tallies for Salinas
salinas_true <- d[d$fraud_bin == TRUE, ]
salinas_false <- d[d$fraud_bin == FALSE, ]

# Create the density plot for Salinas
plot_salinas <- ggplot() +
  geom_density(data = salinas_true, aes(x = salinas_prop), linetype = "dashed", color = "orange", fill =
  geom_density(data = salinas_false, aes(x = salinas_prop), fill = "lightblue", color = "lightblue", alp
  labs(x = "Vote Share (Salinas)", y = "Density") +
  theme_minimal()

# Filter TRUE and FALSE tallies for Cardenas
cardenas_true <- d[d$fraud_bin == TRUE, ]
cardenas_false <- d[d$fraud_bin == FALSE, ]

# Create the density plot for Cardenas
plot_cardenas <- ggplot() +
  geom_density(data = cardenas_true, aes(x = cardenas_prop), linetype = "dashed", color = "orange", fill
  geom_density(data = cardenas_false, aes(x = cardenas_prop), fill = "lightblue", color = "lightblue", a
  labs(x = "Vote Share (Cardenas)", y = "Density") +
  theme_minimal()

# Filter TRUE and FALSE tallies for Clouthier
clouthier_true <- d[d$fraud_bin == TRUE, ]
clouthier_false <- d[d$fraud_bin == FALSE, ]

# Create the density plot for Clouthier
plot_clouthier <- ggplot() +
  geom_density(data = clouthier_true, aes(x = clouthier_prop), linetype = "dashed", color = "orange", fi
  geom_density(data = clouthier_false, aes(x = clouthier_prop), fill = "lightblue", color = "lightblue"
  labs(x = "Vote Share (Clouthier)", y = "Density") +
  theme_minimal()

# Combine the density plots using grid.arrange
combined_plots <- grid.arrange(plot_salinas, plot_cardenas, plot_clouthier, nrow = 3)
```




```
combined_plots
```

```
## TableGrob (3 x 1) "arrange": 3 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (2-2,1-1) arrange gtable[layout]
## 3 3 (3-3,1-1) arrange gtable[layout]
```

The reproduced figure in 4.2 shows the resultant vote share distributions for the three main candidates, with the solid lines representing the densities of the true tallies and the dashed lines representing the fraud tallies.

It shows the occurrence of unaltered tallies with vote shares exceeding 90% for Salinas, and the research argues that argue that the plot indicates two possible scenarios: the official candidate enjoyed exceptional popularity; there is an irregularity in the vote distribution commonly associated with electoral fraud.

Note: Feel free to suggest *multiple* alternative designs to earn bonus credits. However, please be selective. Only a design with major differences from the existing ones can be counted as an alternative design.

Task 5. Visualize the discrepancies between presidential and legislative Votes (6pt)

In this task, you will visualize the differences between the number of presidential votes across tallies. The desired output of is reproducing and extending Figure 5 in the research article (Cantu 2019, pp. 720).

Task 5.1. Get district-level discrepancies and fraud data

As you might have noticed in the caption of Figure 5 in Cantu (2019, pp. 720), the visualized data are aggregated to the *district* level. In contrast, the unit of analysis in the dataset we are working with, `d`, is *tally*. As a result, the first step of this task is to aggregate the data. Specifically, please aggregate `d` into a new data frame named `sum_fraud_by_district`, which contains the following columns:

- `state`: Names of states
- `district`: Names of districts
- `vote_president`: Total numbers of presidential votes
- `vote_legislature`: Total numbers of legislative votes
- `vote_diff`: Total number of presidential votes minus total number of legislative votes
- `prop_fraud`: Proportions of fraudulent tallies (hint: using `fraud_bin`)

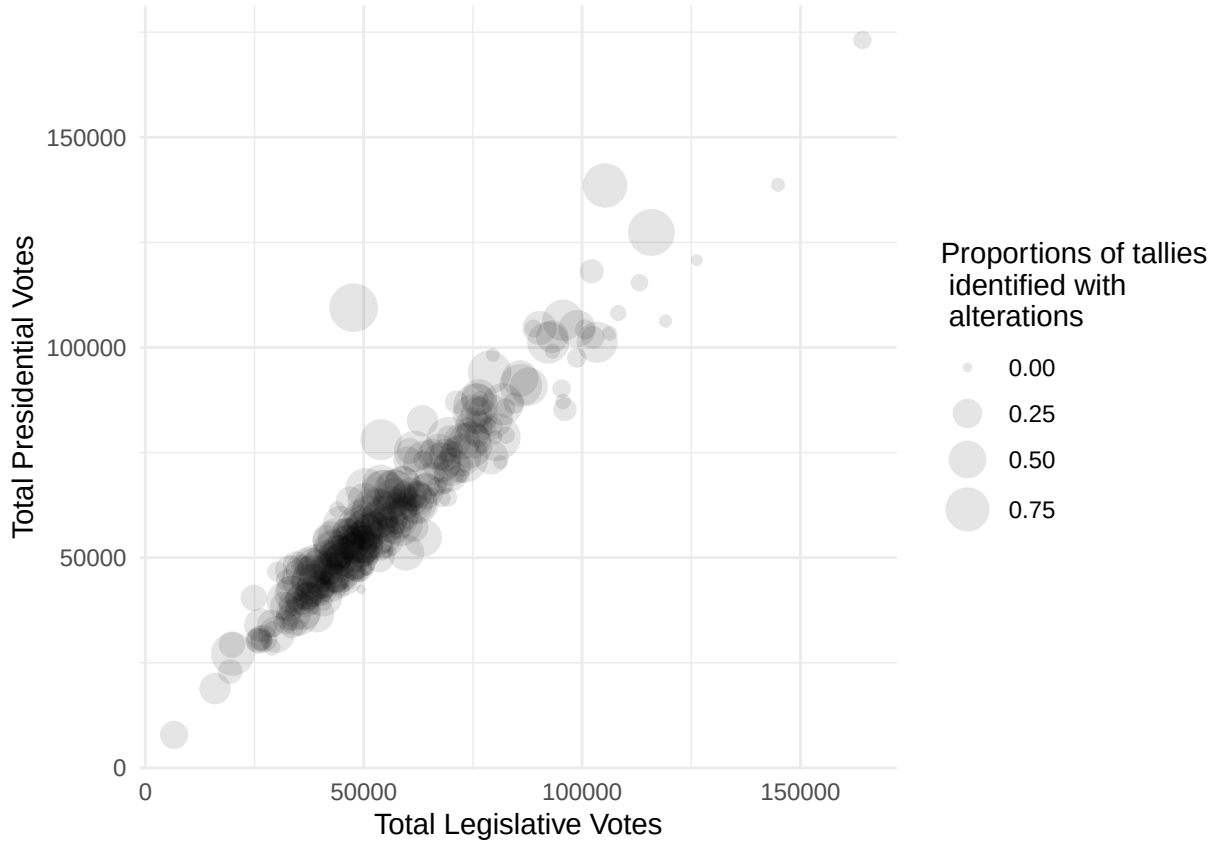
```
sum_fraud_by_district <- d |>
  group_by(state, district) |>
  summarize(
    vote_president = sum(total_president),
    vote_legislature = sum(total),
    vote_diff = sum(total_president) - sum(total),
    prop_fraud= sum(fraud_bin)/n())
  print(sum_fraud_by_district)
```

```
## # A tibble: 300 x 6
## # Groups:   state [32]
##    state           district vote_president vote_legislature vote_diff prop_fraud
##    <chr>           <chr>             <int>            <int>     <int>      <dbl>
##  1 Aguascalientes  I                118139           102213     15926     0.135
##  2 Aguascalientes  II                58722            55271      3451     0.215
##  3 Baja California I                 75385            60550     14835     0.171
##  4 Baja California II                44630            32429     12201     0.0960
##  5 Baja California III               79072            75940      3132     0.132
##  6 Baja California IV               104627            90270     14357     0.375
##  7 Baja California V                 55792            48971      6821     0.152
##  8 Baja California VI                64986            60596      4390     0.368
##  9 Baja Californi~ I                 52226            47569      4657     0.259
## 10 Baja Californi~ II                30405            26641      3764     0.0933
## # i 290 more rows
```

Task 5.2. Replicate Figure 5

Based on all the previous step, reproduce Figure 5 in Cantu (2019, pp. 720).

```
ggplot(sum_fraud_by_district, aes(x = vote_legislature, y = vote_president, size = prop_fraud)) +
  geom_point(alpha = 0.1) +
  scale_size_continuous(range = c(1, 8)) +
  labs(x = "Total Legislative Votes", y = "Total Presidential Votes", size = "Proportions of tallies \n
  theme_minimal()
```



Note 1: Your performance in this task will be mainly evaluated based on your output's similarity with the original figure. Pay attention to the details.

Note 2: The instructor has detected some differences between the above figure with Figure 5 on the published article. Please use the instructor's version as your main benchmark.

Task 5.3. Discuss and extend the reproduced figure

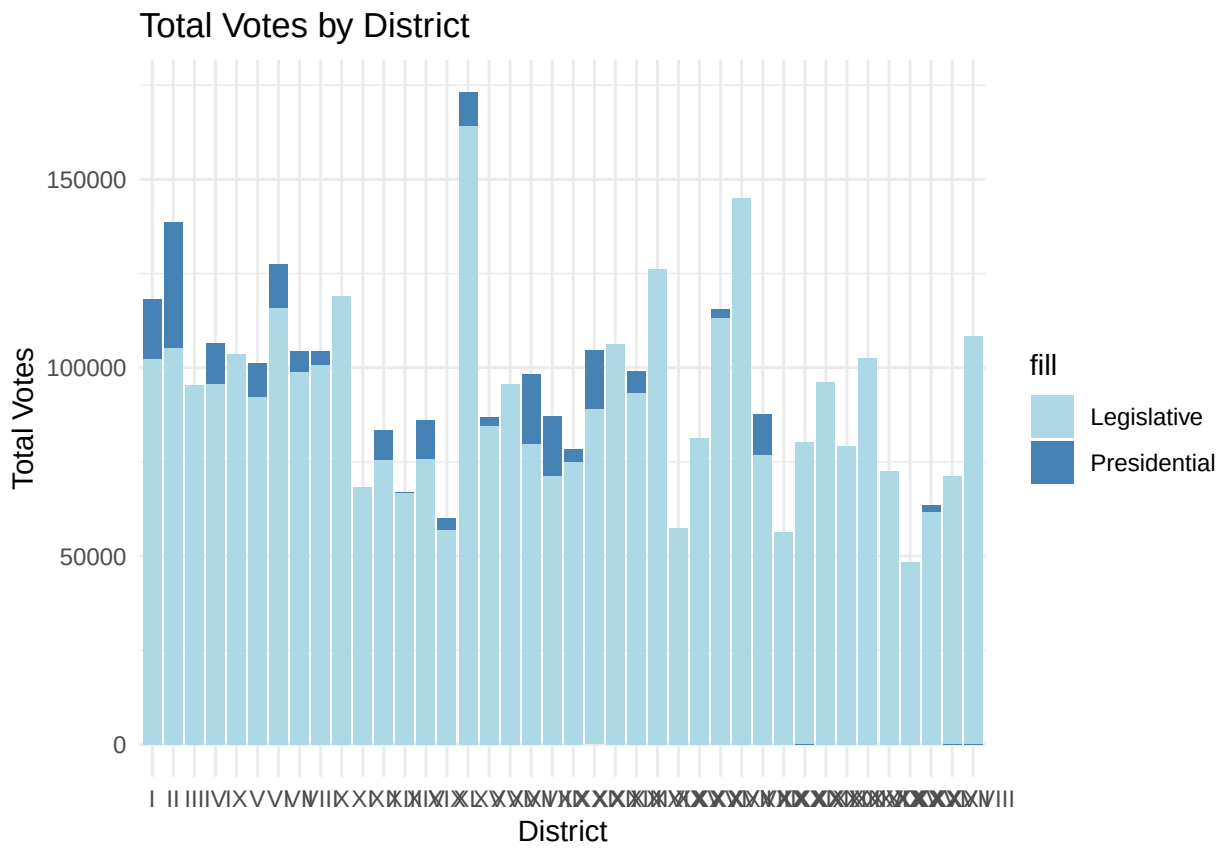
Referring to your reproduced figures and the research articles, in what way is the researcher's argument supported by this figure? Make an alternative visualization design that can substantiate and even augment the current argument. After you have shown your alternative design, in a few sentences, describe how your design provides visual aid as effectively as or more effectively than the original figure.

Note: Feel free to make *multiple* alternative designs to earn bonus credits. However, please be selective. Only a design with major differences from the existing ones can be counted as an alternative design.

```
summary(sum_fraud_by_district)
```

```
##     state              district         vote_president   vote_legislature
##  Length:300         Length:300         Min.   :  7842   Min.   :  6577
##  Class :character   Class :character   1st Qu.: 47140   1st Qu.: 42652
##  Mode  :character   Mode  :character   Median : 56201   Median : 50854
##                                        Mean   : 60692   Mean   : 56106
##                                        3rd Qu.: 71729   3rd Qu.: 66224
##                                        Max.   :173120   Max.   :164206
##                                                         NA's   :1
##    vote_diff        prop_fraud
##  Min.   :-12829   Min.   :0.00000
##  1st Qu.:  1074   1st Qu.:0.06441
##  Median :  3853   Median :0.21671
##  Mean   :  4605   Mean   :0.29073
##  3rd Qu.:  7550   3rd Qu.:0.46661
##  Max.   : 61767   Max.   :0.99306
##  NA's   :1
```

```
ggplot(sum_fraud_by_district, aes(x = district)) +
  geom_col(aes(y = vote_president, fill = "Presidential"), position = "dodge") +
  geom_col(aes(y = vote_legislature, fill = "Legislative"), position = "dodge") +
  labs(x = "District", y = "Total Votes", title = "Total Votes by District") +
  scale_fill_manual(values = c("Presidential" = "steelblue", "Legislative" = "lightblue", alpha=0.5)) +
  theme_minimal()
```



The plot shows that there are a number of points departs from the line X=Y, which aligns with the research's argument–there are significant inconsistencies in the results announced by electoral authorities.

Task 6. Visualize the spatial distribution of fraud (6pt)

In this final task, you will visualize the spatial distribution of electoral fraud in Mexico. The desired output of is reproducing and extending Figure 3 in the research article (Cantu 2019, pp. 720).

Note 3. Load map data

As you may recall, map data can be stored and shared in **two** ways. The simpler format is a table where each row has information of a point that "carves" the boundary of a geographic unit (a Mexican state in our case). In this type of map data, a geographic unit is is represented by multiple rows. Alternatively, a map can be represented by a more complicated and more powerful format, where each geographic unit (a Mexican state in our case) is represented by an element of a `geometry` column. For this task, I provide you with a state-level map of Mexico represented by both formats respectively.

Below the instructor provide you with the code to load the maps stored under the two formats respectively. Please run them before starting to work on your task.

```
# IMPORTANT: Remove eval=FALSE above when you start this part!

# Load map (simple)
map_mex <- read_csv("data/map_mexico/map_mexico.csv")
# Load map (sf): You need to install and load library "sf" in advance
map_mex_sf <- st_read("data/map_mexico/shapefile/gadm36_MEX_1.shp")
map_mex_sf <- st_simplify(map_mex_sf, dTolerance = 100)

#?st_simplify
```

Bonus question: Explain the operations on `map_mex_sf` in the instructor's code above.

In the step on map_mex_sf, st_simplify() is used to simplify lines by removing vertices, reducing the size of the spatial data. In this step, a tolerance level of 100 is specified.

Note: The map (sf) data we use are from https://gadm.org/download_country_v3.html.

Task 6.1. Reproduce Figure 3 with `map_mex`

In this task, you are required to reproduce Figure 3 with the `map_mex` data.

Note:

- Your performance in this task will be mainly evaluated based on your output's similarity with the original figure. Pay attention to the details. For your reference, below is a version created by the instructor.
- Hint: Check the states' names in the map data and the electoral fraud data. Recode them if necessary.

```
#m <- d |> select(state, fraud_proba, fraud_bin)
#map_mex <- map_mex |> rename(state = state_name_official)
# Remove accents from state names in map_mex_sf (from ChatGPT)
#map_mex$state <- stri_trans_general(map_mex$state, "Latin-ASCII")
#merged_d1 <- merge(map_mex, m, by = "state")

#map_1 |>
  #ggplot(aes(x = long, y = lat)) +
  #geom_map(
   # map = map_mex,
    #aes(map_id = state, fill = prop_fraud),
    #color = "black", size = 0.1) +
  #scale_fill_gradient(low = "white", high = "black") +
  #labs(fill = "Proportion \n of altered \n tallies") +
#oord_map() +
 # theme_void()

#map
```

Task 6.2. Reproduce Figure 3 with `map_mex_sf`

In this task, you are required to reproduce Figure 3 with the `map_mex` data.

Note:

- Your performance in this task will be mainly evaluated based on your output's similarity with the original figure. Pay attention to the details. For your reference, below is a version created by the instructor.
- Hint: Check the states' names in the map data and the electoral fraud data. Recode them if necessary.

```
map_mex_sf <- map_mex_sf |> rename(state = NAME_1)
# Remove accents from state names in map_mex_sf (from ChatGPT)
map_mex_sf$state <- stri_trans_general(map_mex_sf$state, "Latin-ASCII")
merged_d2 <- merge(map_mex_sf, d_state, by = "state")

map <- ggplot() +
  geom_sf(data = merged_d2, aes(fill = prop_fraud), color = "black") +
  scale_fill_gradient(low = "white", high = "black", na.value = "gray", name = "Proportion \n of altered
  labs(caption = "Figure 5: Rates of Tallies Classified as Altered by State\n") +
  theme_void() +
  theme(plot.caption = element_text(hjust = 0.5, margin = margin(t = 10, b = 10)))

map
```

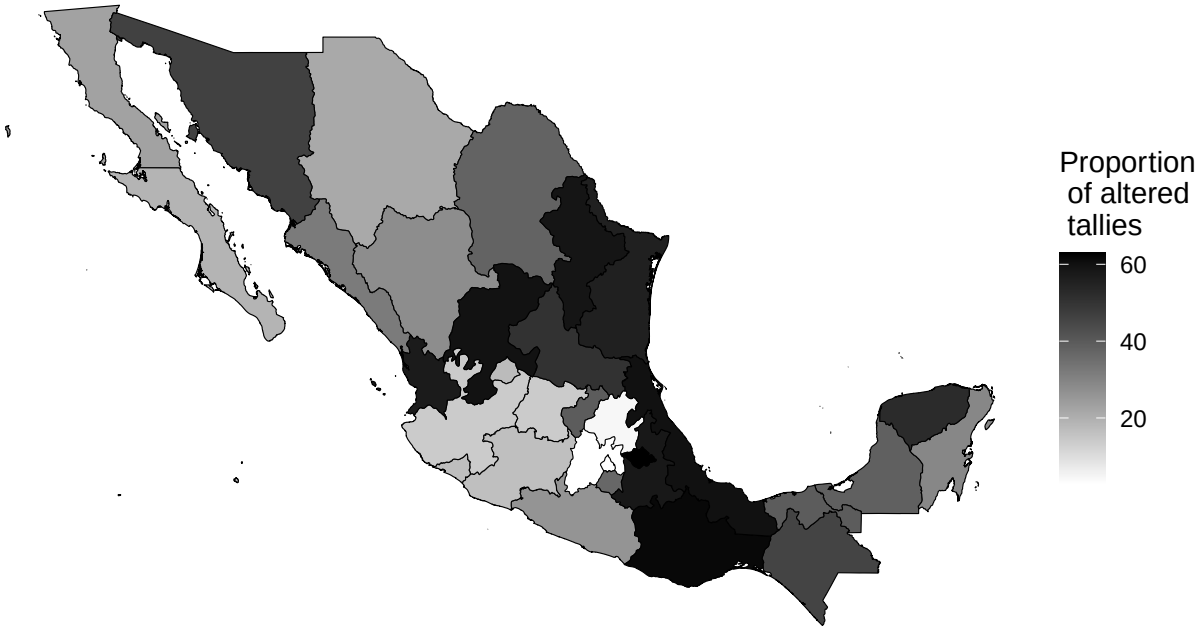


Figure 5: Rates of Tallies Classified as Altered by State

Task 6.3. Discuss and extend the reproduced figures

Referring to your reproduced figures and the research articles, in what way is the researcher's argument supported by this figure? Make an alternative visualization design that can substantiate and even augment the current argument. After you have shown your alternative design, in a few sentences, describe how your design provides visual aid as effectively as or more effectively than the original figure.

Note: Feel free to make *multiple* alternative designs to earn bonus credits. However, please be selective. Only a design with major differences from the existing ones can be counted as an alternative design.

```
# YOUR CODE HERE
```

This plot provides a visual representation of the spatial distribution of predicted fraud tallies. The researcher argues that at the state level in Mexico, the rates of altered tallies (presumably referring to manipulated or fraudulent voting records) vary significantly, which one can see from the grayscale change of the map. According to the figure, there are relatively darker parts in the south, which indicates most of the tallies with alterations are concentrated in the southern region of the country, aligning with the researcher's writing.