Lab 4

AUTHOR
Dina A Al JIbori

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.2     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.3     ✔ tibble    3.2.1
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stat20data)
glimpse(iran)
```

```
Rows: 366
Columns: 9
$ province         <chr> "East Azerbaijan", "East Azerbaijan", "East Azerbaija…
$ city             <chr> "Azar Shahr", "Asko", "Ahar", "Bostan Abad", "Bonab",…
$ ahmadinejad      <int> 37203, 32510, 47938, 38610, 36395, 435728, 20520, 121…
$ rezai            <int> 453, 481, 568, 281, 485, 9830, 166, 55, 442, 391, 238…
$ karrubi          <int> 138, 468, 173, 53, 190, 3513, 74, 46, 211, 126, 173, …
$ mousavi          <int> 18312, 18799, 26220, 12603, 33695, 419983, 14340, 397…
$ total_votes_cast <int> 56712, 52643, 75500, 51911, 71389, 876919, 35295, 163…
$ voided_votes     <int> 606, 385, 601, 364, 624, 7865, 195, 102, 634, 661, 39…
$ legitimate_votes <int> 56106, 52258, 74899, 51547, 70765, 869054, 35100, 162…
```

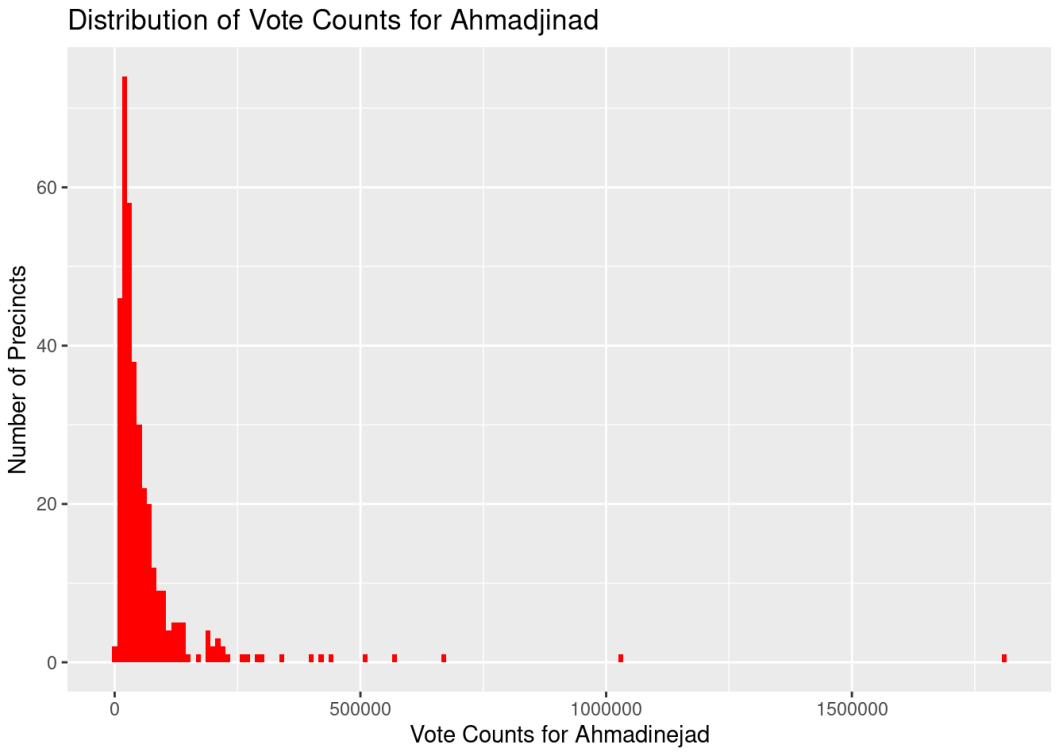
```
library(patchwork)
library(readr)
```

Question 1

Votes per city

Question 2

```
iran %>%
  ggplot(aes(x = ahmadinejad)) +
  geom_histogram(binwidth = 10000, fill = 'red') +
  labs(title = "Distribution of Vote Counts for Ahmadjinad", x=
```




```
iran %>% summarise(mean_vote = mean(ahmadinejad))
```

```
# A tibble: 1 × 1
  mean_vote
      <dbl>
1    66981.
```

```
iran %>% summarise(median_vote = median(ahmadinejad))
```

```
# A tibble: 1 × 1
```

```
  median_vote
        <dbl>
1       35582.
```

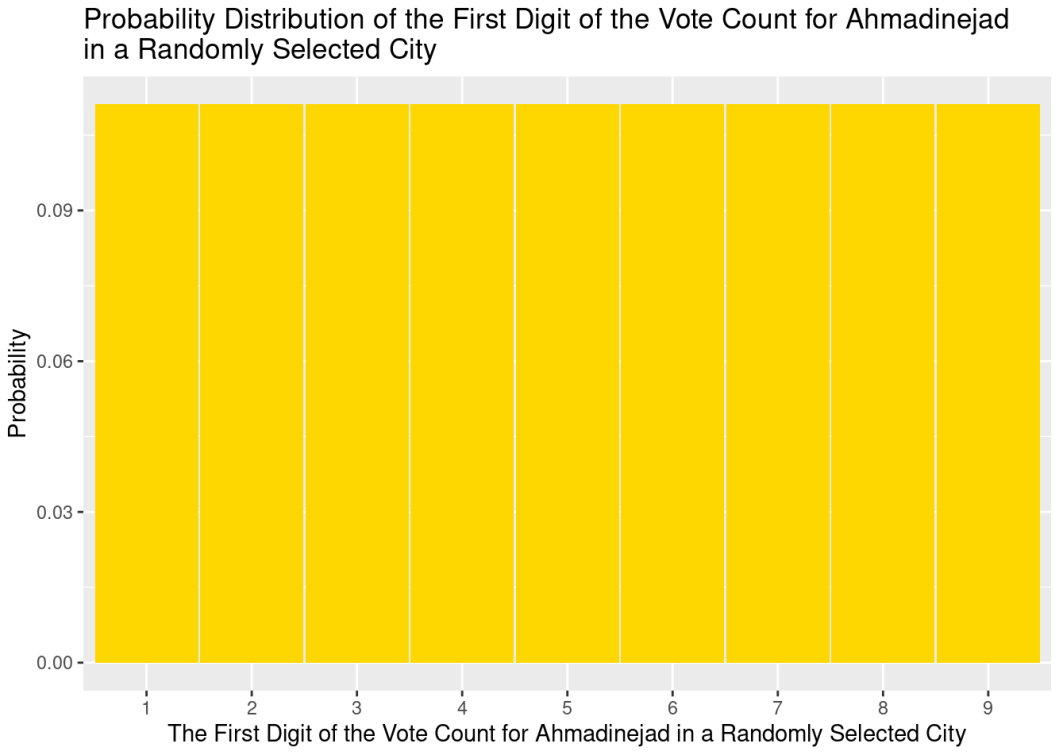
```
iran %>% summarise(iqr_vote = IQR(ahmadinejad))
```

```
# A tibble: 1 × 1
  iqr_vote
     <dbl>
1   46132.
```

The distribution of vote counts for Ahmadinejad is very right skewed with high variability. This means that certain precincts have overwhelming votes for Ahmadinejad. While most other precincts have much lower votes for Ahmadinejad. The average vote is 66981 for Ahmadinejad, although this is a bad measurement for spread for this data since the data is so rightskewed. A better measurement of spread is the median which is 35581 votes. The interquartile range is 46132 votes which is the range where 50% of our votes fall.

Question 3

```
fd_unif <- data.frame(first_digit = seq(1, 9))
fd_unif <- fd_unif %>% mutate(prob = 1/9)
fd_unif %>%
  ggplot(aes(x = factor(first_digit) , y = prob)) +
  geom_col(width = 0.98, fill = "gold") +
  labs(title = "Probability Distribution of the First Digit of
```



Question 4

```
fd_unif <- fd_unif %>% mutate(Exp_val = sum(first_digit*prob))
print(fd_unif)
```

```
  first_digit      prob Exp_val
1           1 0.1111111       5
2           2 0.1111111       5
3           3 0.1111111       5
4           4 0.1111111       5
5           5 0.1111111       5
6           6 0.1111111       5
7           7 0.1111111       5
8           8 0.1111111       5
9           9 0.1111111       5
```

The expected value of X is 5

Question 5

```
fd_unif <- fd_unif %>% mutate(var_prob = (first_digit- Exp_val)
print(fd_unif)
```

```
  first_digit      prob Exp_val  var_prob
1           1 0.1111111       5 1.7777778
```

```
2           2 0.1111111       5 1.0000000
3           3 0.1111111       5 0.4444444
4           4 0.1111111       5 0.1111111
5           5 0.1111111       5 0.0000000
6           6 0.1111111       5 0.1111111
7           7 0.1111111       5 0.4444444
8           8 0.1111111       5 1.0000000
9           9 0.1111111       5 1.7777778
```

```r
fd_unif %>% summarise(sum_prob = sum(var_prob))
```

```
  sum_prob
1 6.666667
```

The variance of X is shown above. It's sum is 6.6667

Question 6

```r
fd_benford <- data.frame(first_digit = seq(1, 9))
fd_benford <- fd_benford %>%
  mutate(prob = log10(1 + 1/first_digit))
print(fd_benford)
```

```
  first_digit       prob
1           1 0.30103000
2           2 0.17609126
3           3 0.12493874
4           4 0.09691001
5           5 0.07918125
6           6 0.06694679
7           7 0.05799195
8           8 0.05115252
9           9 0.04575749
```

```r
fd_benford %>% summarise(sum_prob = sum(prob))
```

```
  sum_prob
1        1
```

This distribution will have an expected value of X that is lower than the Uniform distribution's expected value. This is because the probability is more skewed with the majority of the probability distribution being under 4 while the Uniform distribution has an equally likely chance which causes the expected value to be in the middle of the probability of the set.

Question 7

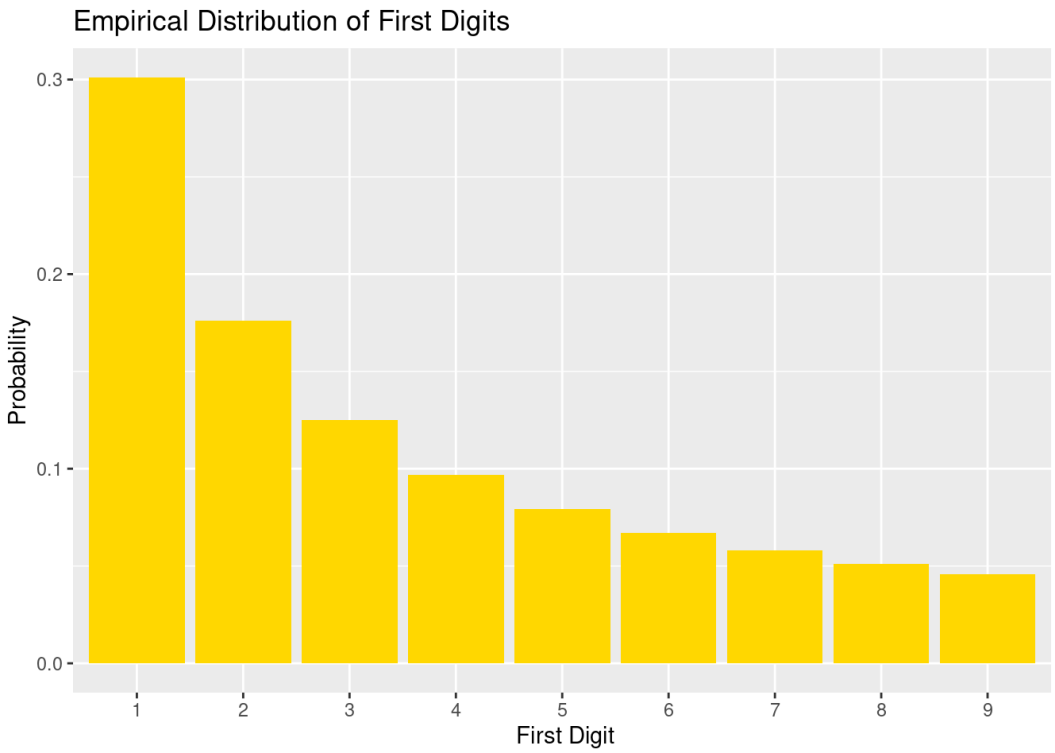
```
fd_benford <- fd_benford %>% mutate(Exp_val = sum(first_digit*p
fd_benford <- fd_benford %>% mutate(var_prob = (first_digit- E>
print(fd_benford)
```

```
  first_digit       prob  Exp_val   var_prob
1           1 0.30103000 3.440237 1.79256031
2           2 0.17609126 3.440237 0.36526302
3           3 0.12493874 3.440237 0.02421420
4           4 0.09691001 3.440237 0.03036527
5           5 0.07918125 3.440237 0.19263694
6           6 0.06694679 3.440237 0.43866126
7           7 0.05799195 3.440237 0.73486890
8           8 0.05115252 3.440237 1.06353455
9           9 0.04575749 3.440237 1.41440819
```

The expected value of X is 3.44 and the variance is shown above.

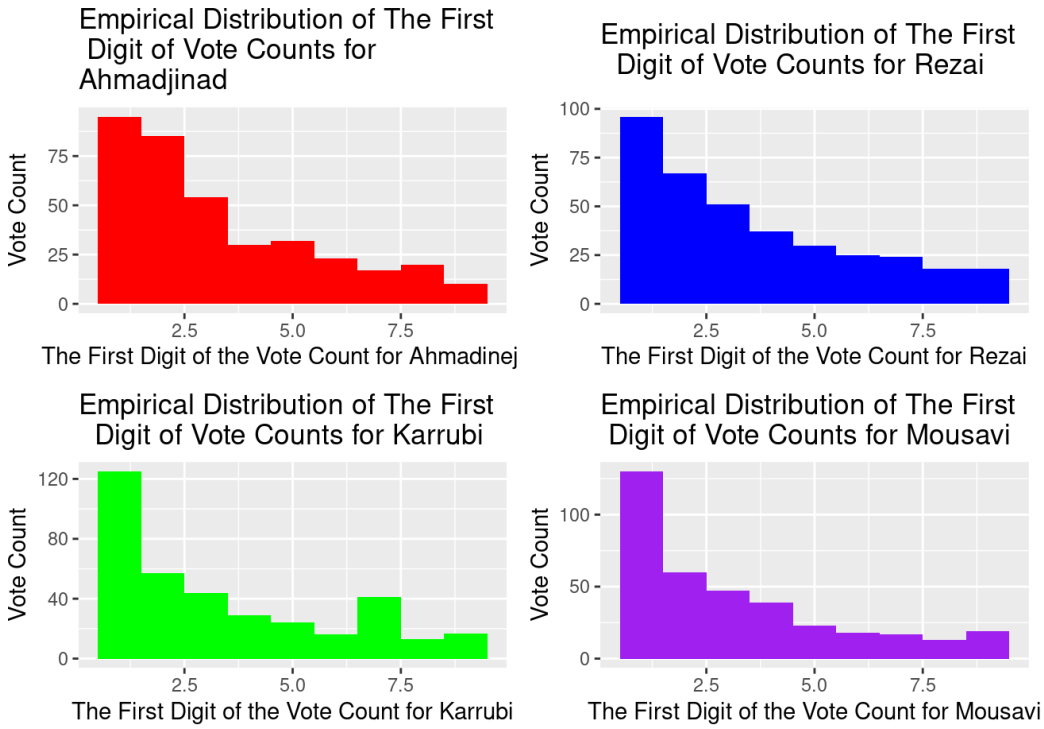
Question 8

```
fd_benford %>%
  slice_sample(n = 366,
               replace = FALSE,
               weight_by = prob) %>%
  ggplot(aes(x=factor(first_digit),y = prob))+
  geom_col(fill = 'gold') + labs(title = "Empirical Distributic
```



Question 9

```
ahmadinejad_graph <- iran %>%
    ggplot(aes(x = get_first(ahmadinejad))) +
  geom_histogram(binwidth = 1, fill = 'red') +
  labs(title = "Empirical Distribution of The First \n Digit of

rezai_graph <- iran %>%
    ggplot(aes(x = get_first(rezai))) +
  geom_histogram(binwidth = 1, fill = 'blue') +
  labs(title = "Empirical Distribution of The First\n  Digit of
karrubi_graph <- iran %>%
    ggplot(aes(x = get_first(karrubi))) +
  geom_histogram(binwidth = 1, fill = 'green') +
  labs(title = "Empirical Distribution of The First \n  Digit o

mousavi_graph <- iran %>%
    ggplot(aes(x = get_first(mousavi))) +
  geom_histogram(binwidth = 1, fill = 'purple') +
  labs(title ="Empirical Distribution of The First \n Digit of
(ahmadinejad_graph + rezai_graph) / (karrubi_graph + mousavi_gr
```



Question 10

The observed first digit distributions are very similar to those simulated from Benford's Law. Although certain candidates have certain outliers that aren't reflected in the Benford Law Probabilities. Rezai has a first-digit distribution most similar to the simulated distribution. Karrubi has a first-digit distribution most different from the simulated ones.

```
florida_votes <- read_csv("https://raw.githubusercontent.com/op
```

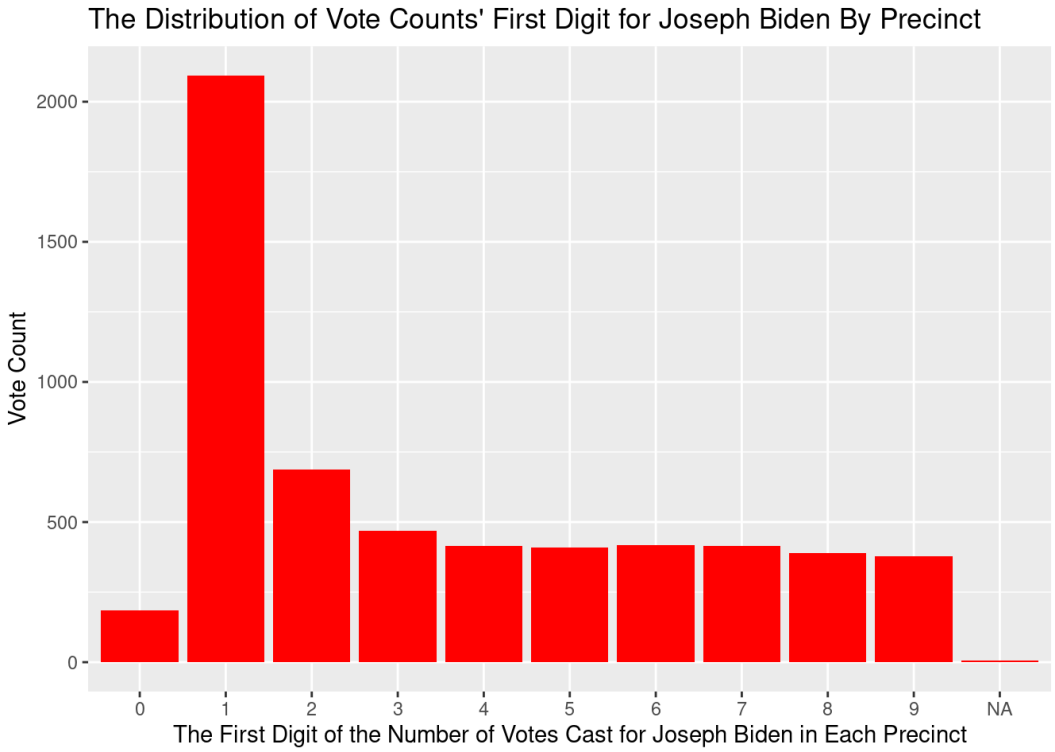
```
Rows: 88666 Columns: 13
── Column specification ────────────────────────────────────────
Delimiter: ","
chr (10): county, precinct, office, candidate, party,
absentee, election_day...
dbl  (2): district, other
num  (1): votes

ℹ Use `spec()` to retrieve the full column specification for
this data.
ℹ Specify the column types or set `show_col_types = FALSE` to
quiet this message.
```

Question 11

I chose to study the state of Florida. The unit of observation is votes per county.The dimensions are 13 by 88666.

Question 12

```
florida_votes %>%
  filter(candidate ==  "Joseph R. Biden") %>%
  ggplot(aes(x = factor(get_first(votes)))) +
  geom_bar(fill = 'red') +
  labs(title = "The Distribution of Vote Counts' First Digit fo
```



Question 13

The election I chose appears to fit Benford's distribution worse than the Iran election. The first digit one has a much higher probability and doesn't decrease in the same way in Benford's distribution.