1A. Distribution of First Digits

October 1, 2021

1 The Distribution of First Digits

In this lab, you will explore the distribution of first digits in real data. For example, the first digits of the numbers 52, 30.8, and 0.07 are 5, 3, and 7 respectively. In this lab, you will investigate the question: how frequently does each digit 1-9 appear as the first digit of the number?

1.1 Question 0

Make a prediction.

1. Approximately what percentage of the values do you think will have a *first* digit of 1? What percentage of the values do you think will have a first digit of 9?
2. Approximately what percentage of the values do you think will have a *last* digit of 1? What percentage of the values do you think will have a last digit of 9?

(Don't worry about being wrong. You will earn full credit for any justified answer.)

EXPLANATION: 1. I think it will be a 10% of the first digit being 1. Also 10% of the first digit being 9. We have 1 digit out of 10 possible digits which .1 -> 10% 2. For both I think it will be 10%. Same for 1 digit out of 10 possible digits which is .1 -> 10%

1.2 Question 1

The S&P 500 is a stock index based on the market capitalizations of large companies that are publicly traded on the NYSE or NASDAQ. The CSV file `sp500.csv` contains data from February 1, 2018 about the stocks that comprise the S&P 500. We will investigate the first digit distributions of the variables in this data set.

Read in the S&P 500 data. What is the unit of observation in this data set? Is there a variable that is natural to use as the index? If so, set that variable to be the index. Once you are done, display the `DataFrame`.

[1]:

```python
# ENTER YOUR CODE HERE.
import pandas as pd
df = pd.read_csv("sp500.csv")
df.head()
df.set_index('Name', inplace = True)
df
```

```
[1]:             date     open    close    volume
     Name
     AAL   2018-02-01   $54.00   $53.88   3623078
     AAPL  2018-02-01  $167.16  $167.78  47230787
     AAP   2018-02-01  $116.24  $117.29    760629
     ABBV  2018-02-01  $112.24  $116.34   9943452
     ABC   2018-02-01   $97.74   $99.29   2786798
     ...          ...      ...      ...       ...
     XYL   2018-02-01   $72.50   $74.84   1817612
     YUM   2018-02-01   $84.24   $83.98   1685275
     ZBH   2018-02-01  $126.35  $128.19   1756300
     ZION  2018-02-01   $53.79   $54.98   3542047
     ZTS   2018-02-01   $76.84   $77.82   2982259

     [505 rows x 4 columns]
```

ENTER YOUR WRITTEN EXPLANATION HERE. - The unit of observation is Name because the names of the companies are unique and easily identifiable. You can search with the name to find the first digit distribution of that company.

1.3 Question 2

We will start by looking at the `volume` column. This variable tells us how many shares were traded on that date.

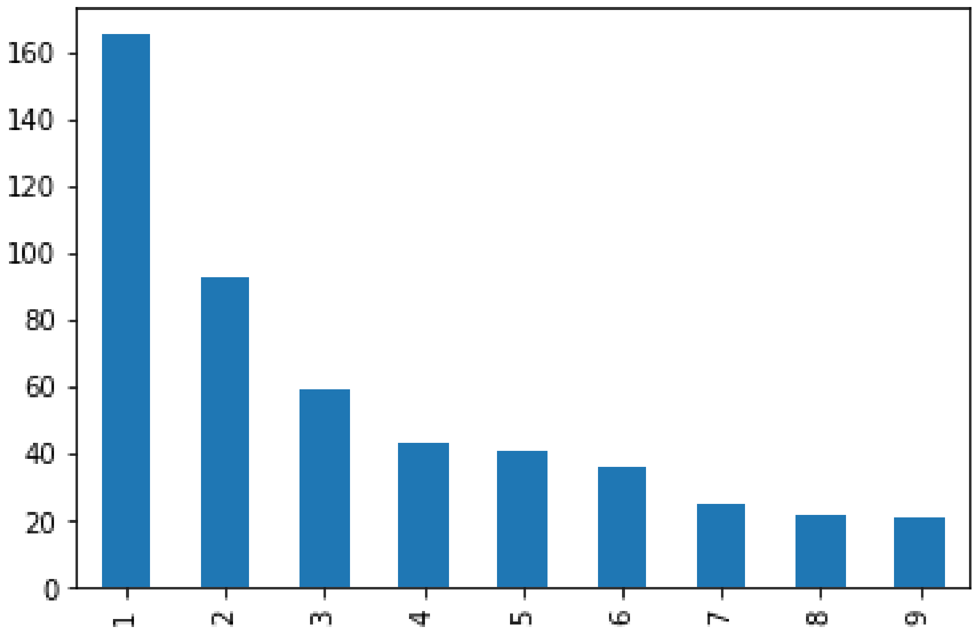
Extract the first digit of every value in this column. (*Hint:* First, turn the numbers into strings. Then, use the text processing functionalities of `pandas` to extract the first character of each string.) Make an appropriate visualization to display the distribution of the first digits. (*Hint:* Think carefully about whether the variable you are plotting is quantitative or categorical.)

How does this compare with what you predicted in Question 0?

```
[2]: # ENTER YOUR CODE HERE.
     df.volume = df.volume.apply(str)
     first_digits = df.volume.str[0]
     first = first_digits.value_counts() #Returns a series containing counts of unique values

     import matplotlib
     %matplotlib inline
     #use bar
     first.plot.bar()
```

```
[2]: <AxesSubplot:>
```



ENTER YOUR WRITTEN EXPLANATION HERE.

- The bar graph visualization shows us that our hypothesis is incorrect. We hypothesized that there was a 10% chance that 1 and 9 would show up as the first digit. However, the graph shows that first digit usually is 1 and is most unlikely 9.

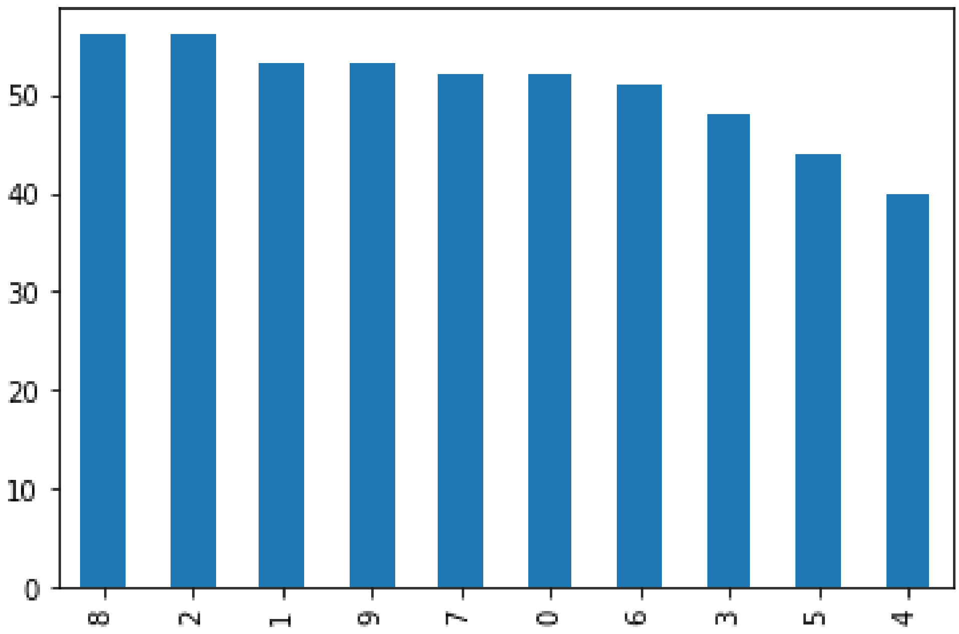
1.4 Question 3

Now, repeat Question 2, but for the distribution of *last* digits. Again, make an appropriate visualization and compare with your prediction in Question 0.

```
[3]: # ENTER YOUR CODE HERE.
     df.volume = df.volume.apply(str)
     last_digits = df.volume.str[-1]
     last = last_digits.value_counts() #Returns a series containing counts of unique␣
      ↪values

     import matplotlib
     %matplotlib inline
     #use bar
     last.plot.bar()
```

```
[3]: <AxesSubplot:>
```



ENTER YOUR WRITTEN EXPLANATION HERE.

- The bar graph visualization shows us that our hypothesis is incorrect. We hypothesized that there was a 10% chance that 1 and 9 would show up as the last digit. However, the graph shows that 1 and 9 have close to the same likeliness of being the last digit.

1.5 Question 4

Maybe the `volume` column was just a fluke. Let's see if the first digit distribution holds up when we look at a very different variable: the closing price of the stock. Make a visualization of the first digit distribution of the closing price (the `close` column of the `DataFrame`). Comment on what you see.

(*Hint:* What type did `pandas` infer this variable as and why? You will have to first clean the values using the text processing functionalities of `pandas` and then convert this variable to a quantitative variable.)

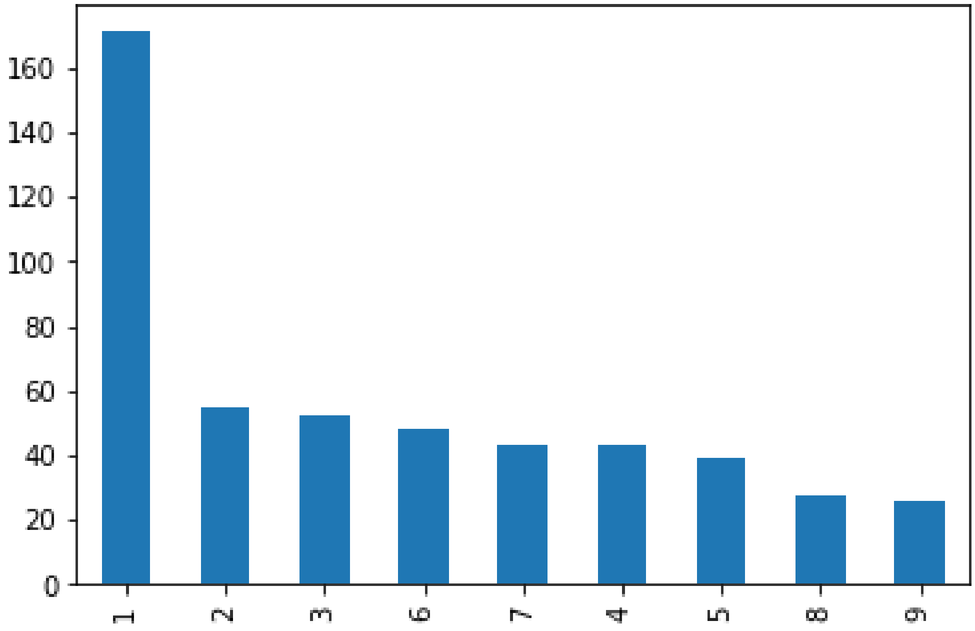
[4]:
```
# ENTER YOUR CODE HERE.
df.close = df.close.apply(str)
first_digits = df.close.str[1]
print(first_digits)
first = first_digits.value_counts()

import matplotlib
%matplotlib inline
first.plot.bar()
```

```
Name
```

```
AAL     5
AAPL    1
AAP     1
ABBV    1
ABC     9
       ..
XYL     7
YUM     8
ZBH     1
ZION    5
ZTS     7
Name: close, Length: 505, dtype: object
```

```
[4]: <AxesSubplot:>
```



ENTER YOUR WRITTEN EXPLANATION HERE. - Similarily to question 2, with the volume, 1 seems to be the most likely fist digit while 9 is the least likely first digit. After that, digits 2 to 8 are about the same but descend in numerical order. - first_digits is of type object because we made it into a string.

1.6 Submission Instructions

Once you are finished, follow these steps:

1. Restart the kernel and re-run this notebook from beginning to end by going to `Kernel > Restart Kernel and Run All Cells`.

2. If this process stops halfway through, that means there was an error. Correct the error and repeat Step 1 until the notebook runs from beginning to end.
3. Double check that there is a number next to each code cell and that these numbers are in order.

Then, submit your lab as follows:

1. Demo your lab to obtain credit.
2. Upload your .ipyn Notebook to iLearn and pdf to Gradescope.