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?

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.)

ENTER YOUR WRITTEN EXPLANATION HERE.
We think that about 30% the values will start with 1 and 5% will start with 9. This is because usually first digits in a string are 1s and then random numbers follow and so on. Therefore, we believe that there will be a high percentage of 1s rather than 9s.
We think that about 10% will end with 1 and 10% will end with 9. This is because since it is the last digit in a string, any number can have a possibilty of being at the end. There isnt a usual way of ending the strings like there is with beginning them so it can be any random number from 1-9 as the last digit of the string so that means that all numbers 1-9 will have about 10% possibility of being the last digit.

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`.

In [1]:

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

Out[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

ENTER YOUR WRITTEN EXPLANATION HERE.
The unit of observation is the companies in the S&P 500. We chose this because we are trying to see the values of the cost of the open and close data for each company so we are observing the data for each company.
The name is the variable that could be the index and can be sorted alphabetically. It is very simple to identify the company by its name and easily look it up the data for the company that way and by making it alphabetized it is easier to scroll through and find the necessary data.

Question 2

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

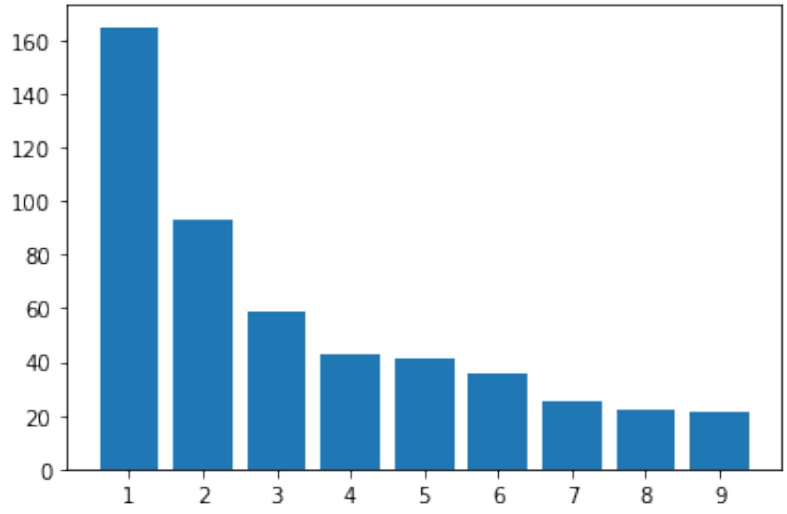
Extract the first digit of every value in this column. (*Hint:* First, turn the numbers into strings. Then, use the text processing functionalities of `pandas` to extract the first character of each string.) Make an appropriate visualization to display the distribution of the first digits. (*Hint:* Think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

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

In [2]:

```
# ENTER YOUR CODE HERE.
df.volume = df.volume.apply(str)
# print(df.dtypes)
# Get the frequencey of each value of the series "df.quantity.str[1]"
digit_frequency = df.volume.str[0].value_counts()
print(digit_frequency)
import matplotlib.pyplot as plt
%matplotlib inline
plt.bar(digit_frequency.index,digit_frequency)
plt.show()
```

```
1    165
2     93
3     59
4     43
5     41
6     36
7     25
8     22
9     21
Name: volume, dtype: int64
```



ENTER YOUR WRITTEN EXPLANATION HERE.
In our prediction we said that 1 would appear the most as the first digit, which was correct, but 1 appears way more often then we predicted; however, we predicited that 9 would appear the least and at a very small percentage, which was correct as seen in the bar graph. As we see in the bar graph, the frequency of the numbers 1-9 appearing as the first digit gets lower and lower as we go through the numbers 1-9 respectively. This shows that 1 has the highest count for being the first digit and 9 has the lowest count (and 2 has the second highest and so forth). AS we go doen the list of numbers 1-9 respectively, there is a decrease in frequency when counting the amount of times that specific number appears as the first digit. This is because usually first digits in a string, when it comes to identifying something like the voulume #/string, are 1s and then random numbers follow. Then, it moves onto 2s then 3s and so on. That is why 9 being the first digit happens the least often (has the smallest frequecy). Therefore, there that is why there is a high percentage of 1s rather than 9s.

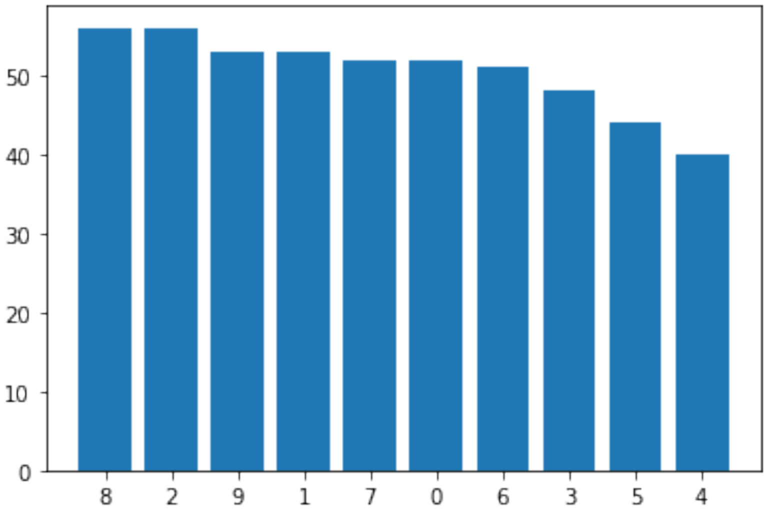
Question 3

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

In [3]:

```
# ENTER YOUR CODE HERE.
df.volume = df.volume.apply(str)
# print(df.dtypes)
# Get the frequencey of each value of the series "df.quantity.str[1]"
digit_frequency = df.volume.str[-1].value_counts()
print(digit_frequency)
import matplotlib.pyplot as plt
%matplotlib inline
plt.bar(digit_frequency.index,digit_frequency)
plt.show()
```

```
8    56
2    56
9    53
1    53
7    52
0    52
6    51
3    48
5    44
4    40
Name: volume, dtype: int64
```



ENTER YOUR WRITTEN EXPLANATION HERE.

In our prediction, we guessed that 1 and 9 will appear the same amount of times as the last digit, which is correct. This is clearly shown in the bar graph, but also seen when looking at the number frequencies counted. As we look at the bar graph, we can see that each number between 1-9 appears almost the same amount of times for being the last digit. The numbers 1-9 all have a similiar and almost same frequency when counting the amount of times that specific number appears as the last digit. This is because since it is the last digit in a string, when it comes to identifying something like the voulume #/string, any number can have a possibilty of being at the end. It is very random and the numbers are placed as the last digit randomly. There isn't a usual or specific way of ending the strings like there is with beginning them so it can be any random number from 1-9 as the last digit of the string so that means that all numbers 1-9 will have about 10% possibility or an equal possibility of being the last digit, which is seen here that they had similar and almost equal frequencies of appearing as the lst digit in the string for volume.

Question 4

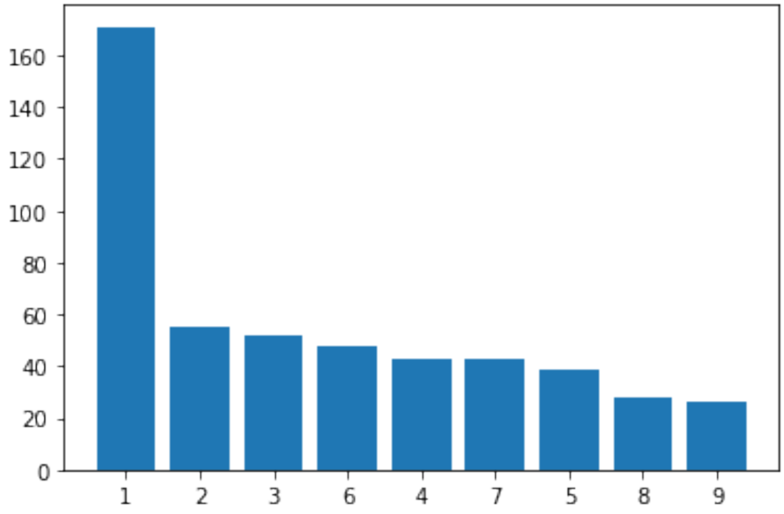
Maybe the `volume` column was just a fluke. Let's see if the first digit distribution holds up when we look at a very different variable: the closing price of the stock. Make a visualization of the first digit distribution of the closing price (the `close` column of the `DataFrame`). Comment on what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

In [4]:

```
# ENTER YOUR CODE HERE.
df.close = df.close.apply(str)
# print(df.dtypes)
# Get the frequencey of each value of the series "df.quantity.str[1]"
digit_frequency = df.close.str[1].value_counts()
print(digit_frequency)
import matplotlib.pyplot as plt
%matplotlib inline
plt.bar(digit_frequency.index,digit_frequency)
plt.show()
```

```
1    171
2     55
3     52
6     48
4     43
7     43
5     39
8     28
9     26
Name: close, dtype: int64
```



ENTER YOUR WRITTEN EXPLANATION HERE.
When looking at the graph, we can see that 1 as the starting digit appears way more than any other number in the closing price and 9 as the starting digit apprears the least again. There is a big drop from the frequency of 1 appearing as the first digit of the closing price to numbers 2-9. For numbers 2-9, there isn't that drastic of a difference in frequency between them as there is with them and the frequency of the number 1. There is only a slight difference and decrease as we go through and look at the frequency of numbers 2-9 respectively. We are looking at the cost column and each string in it. We are now looking for the first number in the string and not the first symbol. Since the first part/character of the string is the dollar sign symbol and we want to look at the first number in the string, we would put 1 in our code instead of 0 because the 0 in each string in the cost column would be the dollar symbol. So instead instead of putting 0 we put 1 in our code to account for the $ symbol in the cost column

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.