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.)

1. 10% for the first digit being a 1, and 10% for the first digit being a 9.
2. 10% for the last digit beina a 1, and 10% for the last digit being a 9.

Question 1

The S&P 500 (https://en.wikipedia.org/wiki/S%26P_500_Index) 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 [5]:
```

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

```
Out[5]:
```

	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.

From the data set, the Name column would be the unit of observation and a natural variable to use as the index.

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 (https://pandas.pydata.org/pandas-docs/stable/text.html) 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?

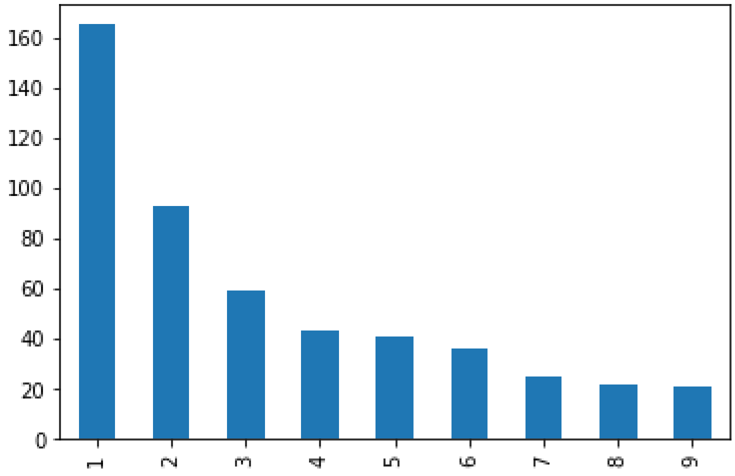
In [6]:

```
# ENTER YOUR CODE HERE.
import matplotlib # python library used to create 2D graphs from python script
%matplotlib inline

df.volume = df.volume.apply(str) # turns the numbers into strings
first_digits = df.volume.str[0] # extracts the first character of each string

first_value = first_digits.value_counts() # returns a series of the first value
percentage = first_value / first_value.sum() # calculate the percentages
first_value.plot.bar() # print bar graph
print(percentage * 100) # print data set
```

```
1    32.673267
2    18.415842
3    11.683168
4     8.514851
5     8.118812
6     7.128713
7     4.950495
8     4.356436
9     4.158416
Name: volume, dtype: float64
```



ENTER YOUR WRITTEN EXPLANATION HERE.

By comparing question 2 from question 0, there's a 33% value of 1's and 4% value of 9's being the first digit.

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.

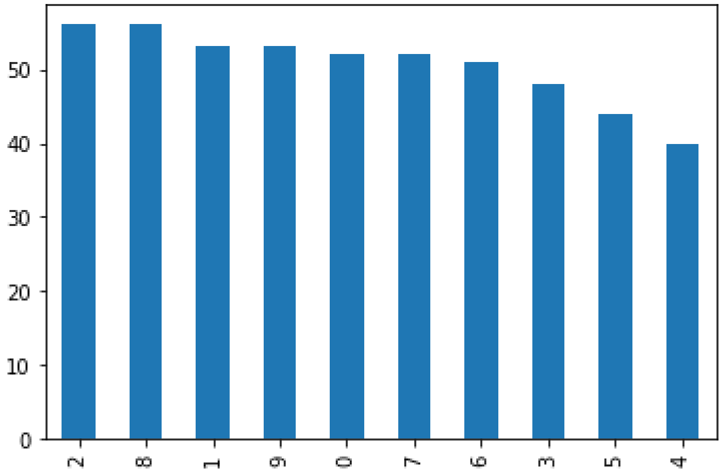
In [7]:

```
# ENTER YOUR CODE HERE.
import matplotlib # python library used to create 2D graphs from python script
%matplotlib inline

df.volume = df.volume.apply(str) # turns the numbers into strings
first_digits = df.volume.str[-1] # extracts the last character of each string

first_value = first_digits.value_counts() # returns a series of the first value
percentage = first_value / first_value.sum() # calculate the percentages
first_value.plot.bar() # print bar graph
print(percentage * 100) # print data set
```

```
2    11.089109
8    11.089109
1    10.495050
9    10.495050
0    10.297030
7    10.297030
6    10.099010
3     9.504950
5     8.712871
4     7.920792
Name: volume, dtype: float64
```



ENTER YOUR WRITTEN EXPLANATION HERE.

By comparing question 3 from question 0, there's a 10% value of 1s and 10% value of 9s being the last digit.

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 (https://pandas.pydata.org/pandas-docs/stable/text.html)](https://pandas.pydata.org/pandas-docs/stable/text.html) of `pandas` and then convert this variable to a quantitative variable.)

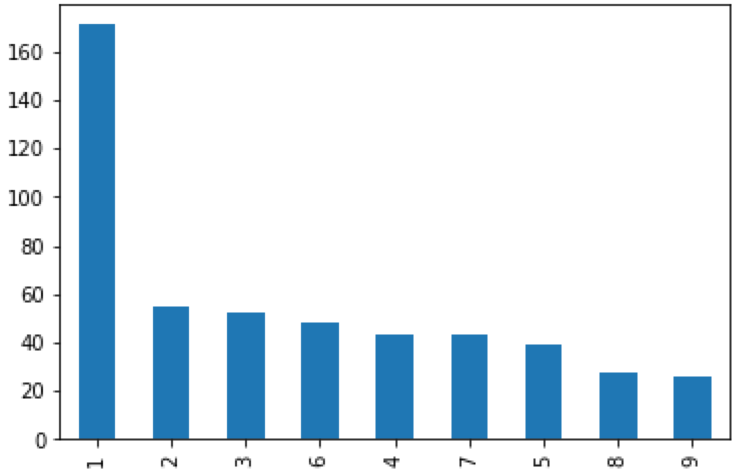
In [8]:

```
# ENTER YOUR CODE HERE.
import matplotlib # python library used to create 2D graphs from python script
%matplotlib inline

df.volume = df.close.apply(str) # turns the numbers into strings
first_digits = df.close.str[1] # extracts the first character of each string

first_value = first_digits.value_counts() # returns a series of the first value
percentage = first_value / first_value.sum() # calculate the percentages
first_value.plot.bar() # print bar graph
print(percentage * 100) # print data set
```

```
1    33.861386
2    10.891089
3    10.297030
6     9.504950
4     8.514851
7     8.514851
5     7.722772
8     5.544554
9     5.148515
Name: close, dtype: float64
```



ENTER YOUR WRITTEN EXPLANATION HERE.

From the data set. close is the unit of observation. There's a 34% value of 1's and 5% value of 9's.

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. Go to `File > Export Notebook As > PDF`.
2. Double check that the entire notebook, from beginning to end, is in this PDF file. (If the notebook is cut off, try first exporting the notebook to HTML and printing to PDF.)
3. Upload the PDF to iLearn.
4. Have the TA check your lab to obtain credit.