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.

1. I believe about more than 20% of the values would have a first digit of 1. I believe 1 is a pretty common number since it is the first cardinal number so it would make sense that a significant percentage of the values would have a first digit of 1. I believe around 5% of the values would have a first digit of 9. Since 9 is the last digit when counting up from 1, a smaller percentage of values would have 9 as the last digit.
2. I believe around 5% of the values would have the last digit of 1. Since the last digit is one that comes at the end, it would make more sense to choose a number that when counting, also comes at the end. I believe around 20% of the values would have the last digit of 9 for the same reason. It would naturally make sense to choose a number that comes last when counting for the last digit of a value.

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 [1]:

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

df_comp = df.set_index("Name")
print(df_comp)
```

```
            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 in this data set is the company.

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.)

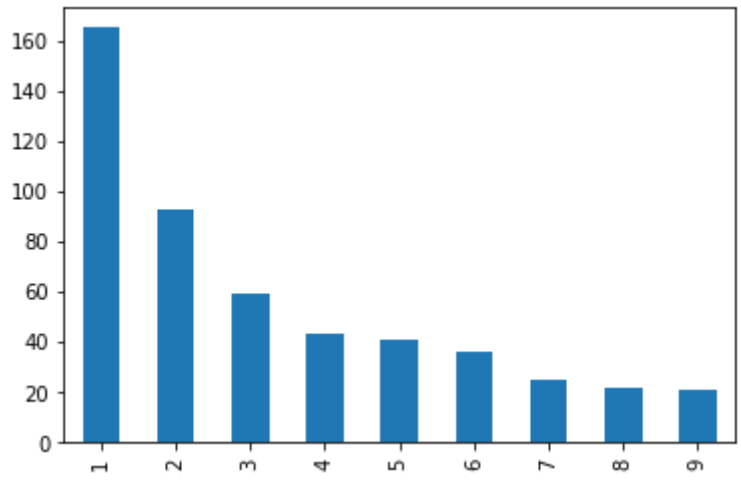
How does this compare with what you predicted in Question 0?

In [2]:

```
# ENTER YOUR CODE HERE.
df.volume = df.volume.apply(str)
first_digits = df.volume.str[0]
first_digits.value_counts()

import matplotlib
%matplotlib inline

fn = first_digits.value_counts()
fn.plot.bar()
```

Out[2]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbf1da4d50>
```



ENTER YOUR WRITTEN EXPLANATION HERE. This supports what I predicted in Question 0. The most common first digit is 1.

Question 3

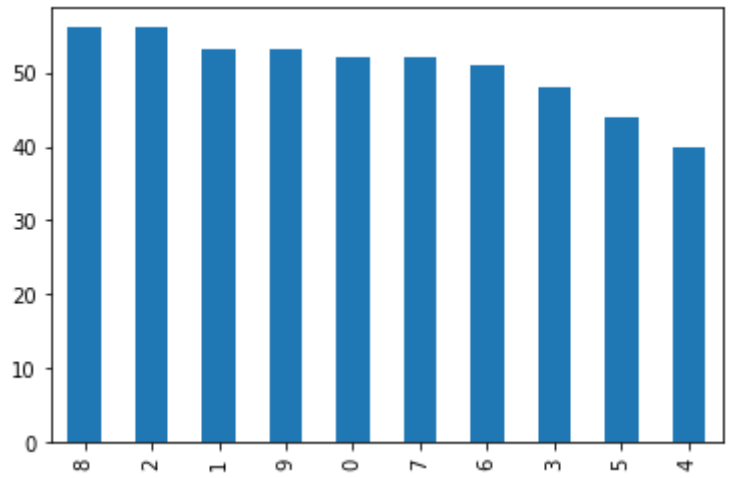
Now, repeat Question 2, but for the distribution of *last* digits. Again, make an appropriate visualization and compare with your prediction in Question 0.

In [3]:

```
# ENTER YOUR CODE HERE.
df.volume = df.volume.apply(str)
last_digit = df.volume.str[-1]
last_digit.value_counts()

import matplotlib
%matplotlib inline

ln = last_digit.value_counts()
ln.plot.bar()
```

Out[3]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbf29f35d0>
```



ENTER YOUR WRITTEN EXPLANATION HERE. This does not support what I predicted in Question 0. From the bar chart, we can see that there is almost an equal amount of last digit occurences for all numbers. No number is significantly higher than the other but 4 could be considered the least common.

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) of `pandas` and then convert this variable to a quantitative variable.)

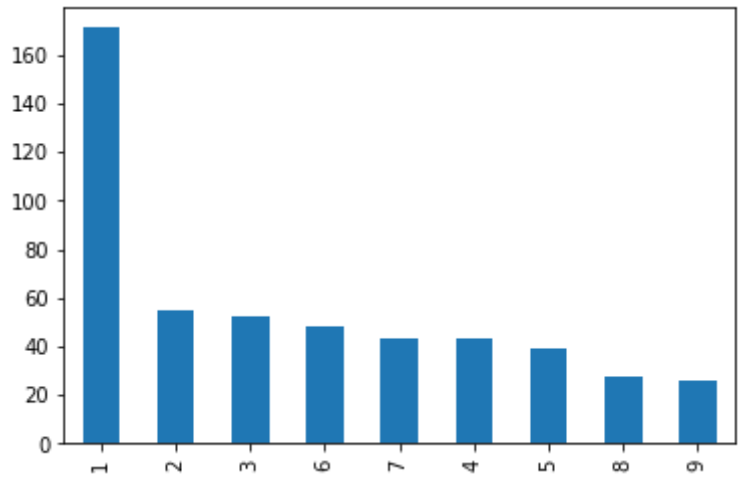
In [4]:

```
# ENTER YOUR CODE HERE.
df.close = df.close.apply(str)
first_digit = df.close.str[1]
first_digit.value_counts()
print(first_digit)

import matplotlib
%matplotlib inline

fn = first_digit.value_counts()
fn.plot.bar()
```

```
0      5
1      1
2      1
3      1
4      9
      ..
500    7
501    8
502    1
503    5
504    7
Name: close, Length: 505, dtype: object
```

Out[4]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbf2c03090>
```



ENTER YOUR WRITTEN EXPLANATION HERE. We can see that once again 1 is the most common first digit. The more higher digits (8 and 9) are the least common digits.

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.

In [ ]: