1A. Explore the In Class Survey

January 16, 2019

1 Explore the In Class Survey

During the first class, you filled out the first-day survey. In this lab, you will explore the responses.

Run the following code to read the data into a `pandas` `DataFrame` whose columns are the survey questions. Each row represents one student's response to the questions.

```
In [1]: import pandas as pd
        import requests
        %matplotlib inline

        API_KEY = "AIzaSyAu1_itQekOyIXrIKIfn9sJrLVVGCL3Unc"
        SPREADSHEET_ID = "1NzcMTL7INHHee8CDdcSPPgyKuiEg0YTkqoOkJbI9JmQ"

        url = "https://sheets.googleapis.com/v4/spreadsheets/%s/values/A1:J100?key=%s" % (
            SPREADSHEET_ID,
            API_KEY
        )
        req = requests.get(url)
        df = pd.DataFrame(req.json()["values"])
        df = df.rename(columns=df.iloc[0]).drop(0)

        df.head()
```

Out[1]:

	Timestamp	What is your major?	How many older siblings do you have?	How many younger siblings do you have?	Which would you rather be?
1	1/8/2019 8:42:05	STAT	0	2	the worst player on a great team
2	1/8/2019 8:42:10	STAT	1	0	the best player on a horrible team
3	1/8/2019 8:42:24	CSC	1	0	the best player on a horrible team
4	1/8/2019 8:42:27	CSC	1	0	the worst player on a great team
5	1/8/2019 8:42:29	political science	0	0	the worst player on a great team

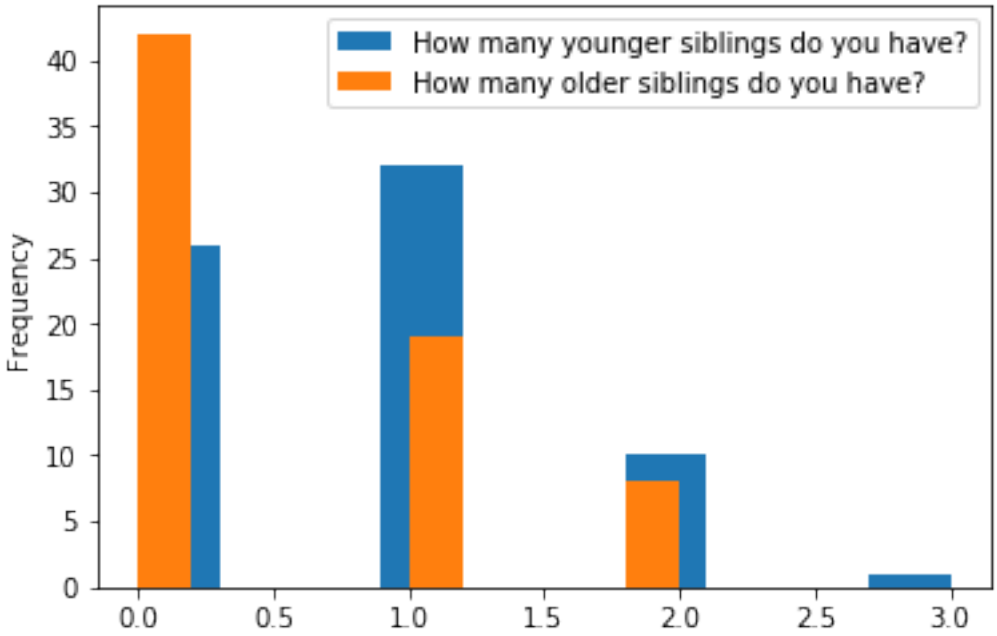
```
  What is your favorite color? What day of the month were you born?  \
1                       orange                                    10
2                 Forest Green                                    29
3                         Blue                                     4
4                         blue                                    25
5                          red                                    18

  Do you think that we would need more or less than 100000 basketballs?  \
1                                                 more
2                                                 less
3                                                 more
4                                                 more
5                                                 more

  How many basketballs do you think we would need? Prompt
1                                             1500   1000
2                                              700   1000
3                                        100000000   1000
4                                             3000   1000
5                                             2750   1000
```

1.1 Question 1

Calculate the number of siblings (total, both older and younger) each student has. Make a graphic that visualizes this information. Explain what you see.

```
In [2]: older_siblings_count = pd.to_numeric(df["How many older siblings do you have?"])
        younger_siblings_count = pd.to_numeric(df["How many younger siblings do you have?"])

        younger_siblings_count.plot.hist(legend=True),
        older_siblings_count.plot.hist(legend=True)

        (older_siblings_count.sum(), older_siblings_count.mean(),
        younger_siblings_count.sum(), younger_siblings_count.mean(),
        older_siblings_count.sum() + younger_siblings_count.sum())
```

```
Out[2]: (35, 0.50724637681159424, 55, 0.79710144927536231, 90)
```



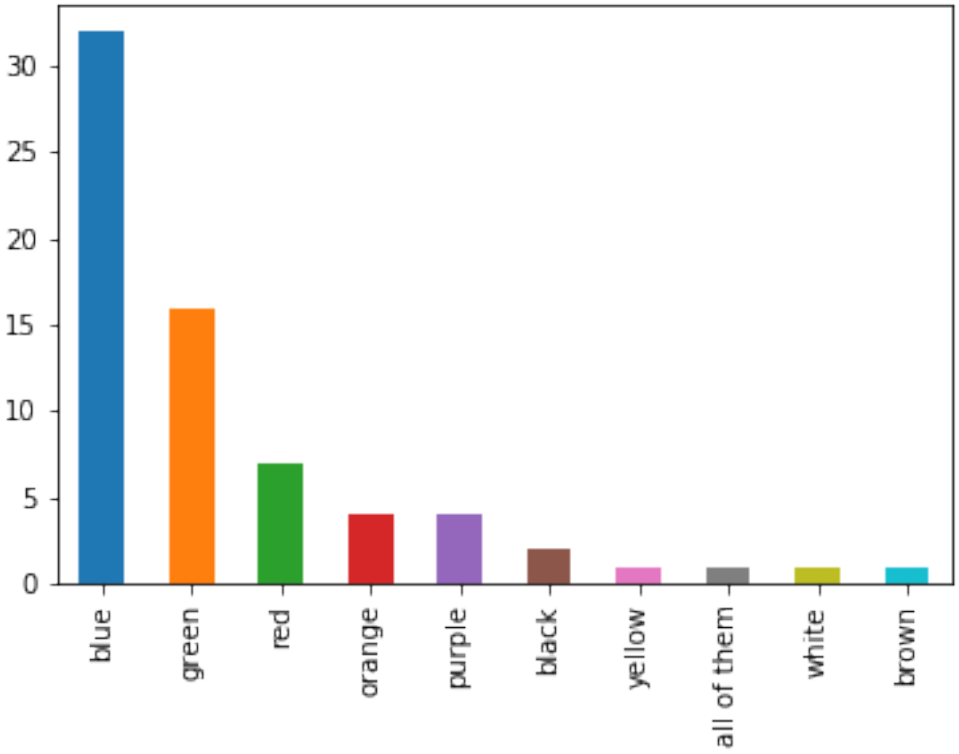
TYPE YOUR WRITTEN EXPLANATION HERE. In this survey it appears that all of the students combined have 55 younger siblings and 35 older siblings. On average one of every 2 students from the survey has an older sibling whereas students from the survery have a younger sibling on average about 80% of the time. However, looking at the distributions of siblings may tell a different story. Observe that the older sibling distribution is skewed right meaning that the mean is greater than what it should be. Much of the data is closer to 0, indicating that perhaps less than one of every two students has a sibling. While observing the younger sibling distribution, we see that the distribution peaks at 1 and the distribution looks somewhat normally shaped. This indicates that perhaps 4 out every 5 students do have a younger sibling.

1.2 Question 2

Make a graphic that visualizes the favorite colors of students in DATA 301. Explain what you see.
(*Hint:* You might have to clean the data a bit first.)

```
In [3]: df["New Color"] = df["What is your favorite color?"].str.lower()

        for color in ["blue", "red", "green"]:
            df.loc[df["New Color"].str.contains(color), "New Color"] = color

        df["New Color"] = df["New Color"].str.replace("navy" , "blue")
        df["New Color"] = df["New Color"].str.replace("burgundy" , "red")
        df["New Color"] = df["New Color"].str.replace("spaghetti" , "red")
        df["New Color"] = df["New Color"].str.replace("magenta" , "purple")
        df["New Color"] = df["New Color"].str.replace("sage" , "green")
        df["New Color"] = df["New Color"].str.replace("mint" , "green")
        df["New Color"].value_counts().plot.bar()
```

```
Out[3]: <matplotlib.axes._subplots.AxesSubplot at 0x7f0f463c0828>
```



TYPE YOUR WRITTEN EXPLANATION HERE. From the visualization, blue is the most common favorite color for students in the survery followed by green and red. Respondents said blue nearly twice as many times as the next highest color, green. In order to reach this conclusion, I first cleaned the data by changing the responses to all lower case - facilitating the process of grouping the colors into families of colors. Next, because there were still repsonses such as navy, burgundy, and mint, I manually grouped those colors into their respective families.

1.3 Question 3

Remember that wacky question about how many basketballs would fit in the classroom? Unbeknownst to you, I actually presented the question differently to the two sections.

- The morning section was first asked, "Do you think that we would need more or less than 1,000 basketballs?"
- The afternoon section was first asked "Do you think that we would need more or less than 100,000 basketballs?"

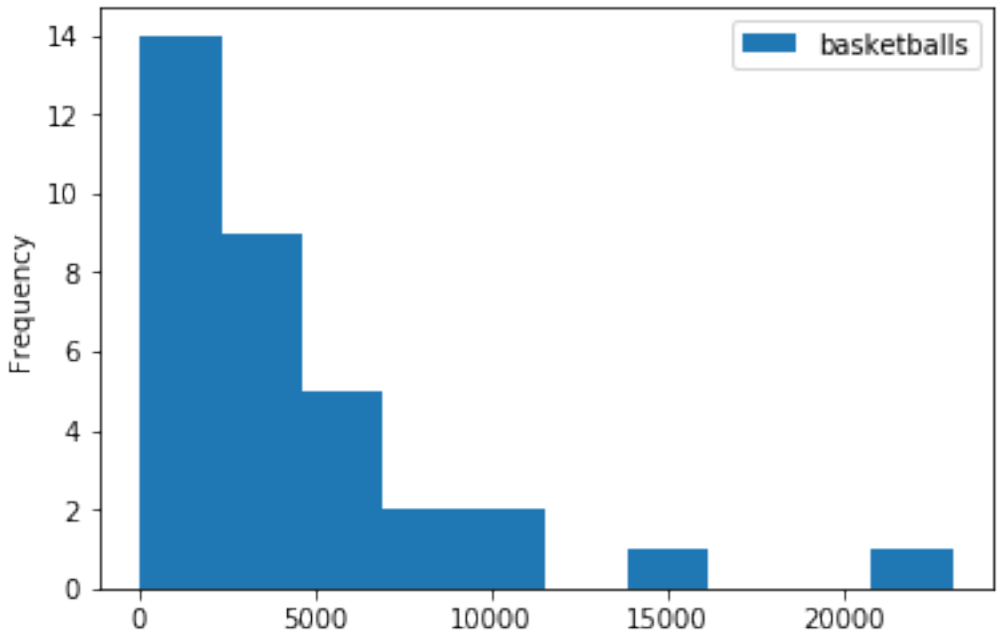
The exact number that each student was given in the prompt is stored in the "Prompt" column of the `DataFrame`.

The purpose of this exercise was to test a famous effect in psychology called the "anchoring effect". The hypothesis is that the afternoon section, which was presented with the higher "anchor", would guess larger numbers than the morning section.

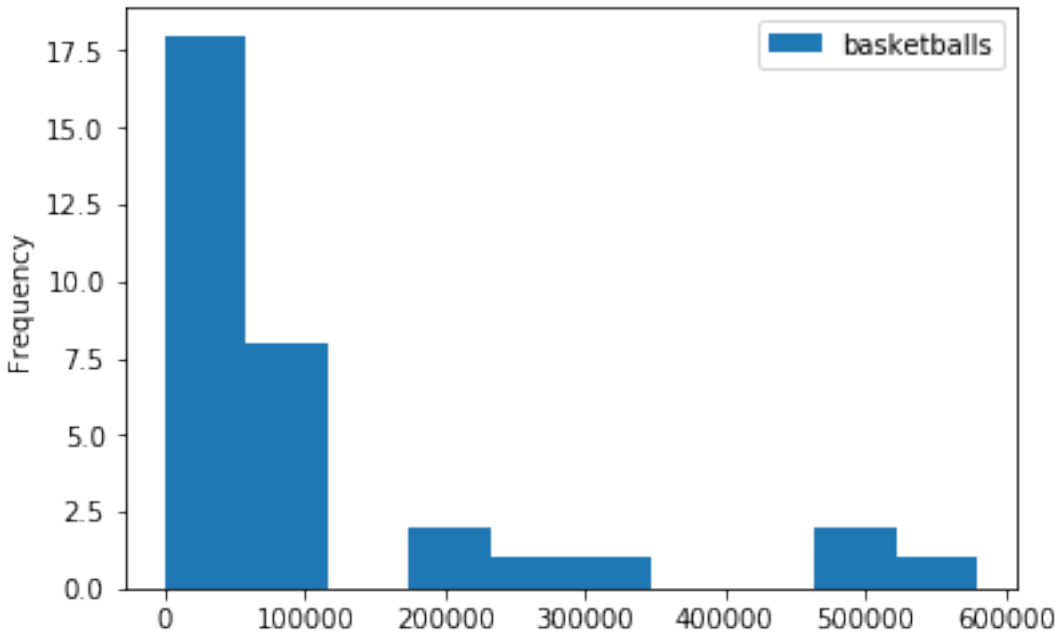
Does the data provide evidence of an anchoring effect? Explain your approach and state your conclusions.

(*Hint:* There are many reasonable approaches to this problem. You will get full credit for any reasonable approach, as long as you carefully justify it.)

```
In [4]: df["basketballs"] = pd.to_numeric(
            df["How many basketballs do you think we would need?"])

        inds = list(range(1,37))
        del inds[16]
        del inds[2]
        section1 = df.loc[inds]
        section2 = df[36:69]

In [5]: section1.plot.hist()
        section1["basketballs"].describe()

Out[5]: count       34.000000
        mean      4371.794118
        std       4666.828952
        min         15.000000
        25%       1500.000000
        50%       2875.000000
        75%       5000.000000
        max      23076.000000
        Name: basketballs, dtype: float64
```



```
In [6]: section2.plot.hist()
        section2["basketballs"].describe()

Out[6]: count        33.000000
        mean     106606.151515
        std      154047.439848
        min           1.000000
        25%       10000.000000
        50%       50000.000000
        75%      100001.000000
        max      580000.000000
        Name: basketballs, dtype: float64
```



The data does provide evidence of the anchoring effect. I approached this problem by first by making a quick histogram of responses in hope there would have been a clear bimodal shape. This would have hopefully shown 2 distinct responses due to how the quesiton was framed - with 1000 balls in the prompt and 100000 balls in the prompt. However, the data appears did not have this shape. I then split the data by how the question was asked in the survery - 1000 balls, 100000 balls. Due to 2 extreme responses in the 1000 ball survery, I removed these responses in order to create effective histograms. Otherwise, it would have been nearly impossible to determine if there was an anchoring effect in the data. After making the histograms and reviewing the summary statistics, there is evidence of anchoring. If we look at the upper 25% of the 1000 balls distribution we find that the 75th percentile is 5000 balls, so this means 75% of the data is less than 5000, making it clear that people are choosing numbers closer to 1000. Next, if we look at the 25th percentile of the 100000 balls distribution we see that the 25th percentile is 1000, so this means 75% of the data is larger than 1000 balls and is approaching 100000 balls since the maximum is 580000. Also,

the mean of the of the 100000 balls distribution is 97000, practically the same as phrased in the quesiton. Thus, the data provides evidence of anchoring.

1.4 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 PolyLearn.