ex-1.2.1-solutions

December 28, 2018

```
In [4]: import pandas as pd
```

Exercise 1.2.1

Question 1

Part 1

The first question asks us to obtain the distribution for the hair color of the 20 students. The hair color data is the following list:

```
In [4]: hair_color=['Red', 'Blond', 'Blond', 'Brown', 'Brown', 'Red',
         'Blond', 'Blond', 'Brown',
        'Black', 'Blond', 'Red', 'Red', 'Brown', 'Black', 'Brown', 'Red',
         'Black',
        'Brown', 'Blond']
```

We wrap the list in a Pandas datframe and use the
function to find the distribution of the hair color data:

```
In [7]: hair_df=pd.DataFrame({'hair_color':hair_color})
```

```
In [17]: hair_dist=hair_df.hair_color.value_counts()
         dist
```

```
Out[17]: Brown    6
         Blond    6
         Red      5
         Black    3
         Name: hair_color, dtype: int64
```

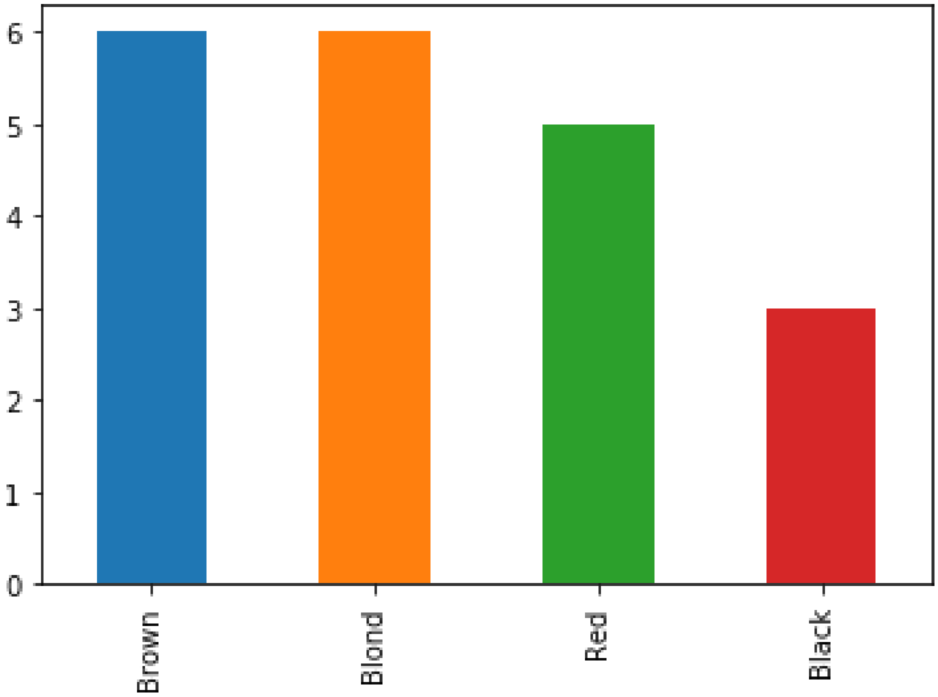
The above is the distribution of the data.

Part 2

This question goes on to ask us to draw a histogram of the hair data. We invoke the Pandas function on the hair_dist object in order to get a plot of the distribution.

```
In [18]: hair_dist.plot(kind='bar')
```

```
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7fde4a71a400>
```



The last part of the question asks us to compute the sample proportion of blonds. We do not see any definition of the sample proportion, so we use Google to find one. Well, nothing jumps out when we search for the term sample proportion, so we go back to the text. For a distrbution consisting of 17 R's and 3 L's, the sample proportion is given as $\frac{3}{16}$ We guess that the numerator is the count of items in the class in the distribution we want the sample proportion for, and the denominator is the sum of the counts of the other classes in the distribution. Therefore, to answer this question, the sample proportion is $\frac{6}{(6+5+3)-1} = \frac{6}{13}$.

Question 4

This is not really I question, it is a task but I am not sure what to call it in this outliine format I'm making up now. Anyway, question 4 is asking for the distribution of left and right handed players in the baseball player data in Appendix A. Well, we had a hard time trying to copy and past that data in order to work with it, so the best thing to do is convert the book pdf to tiff, and then use tesseract to do the OCR.

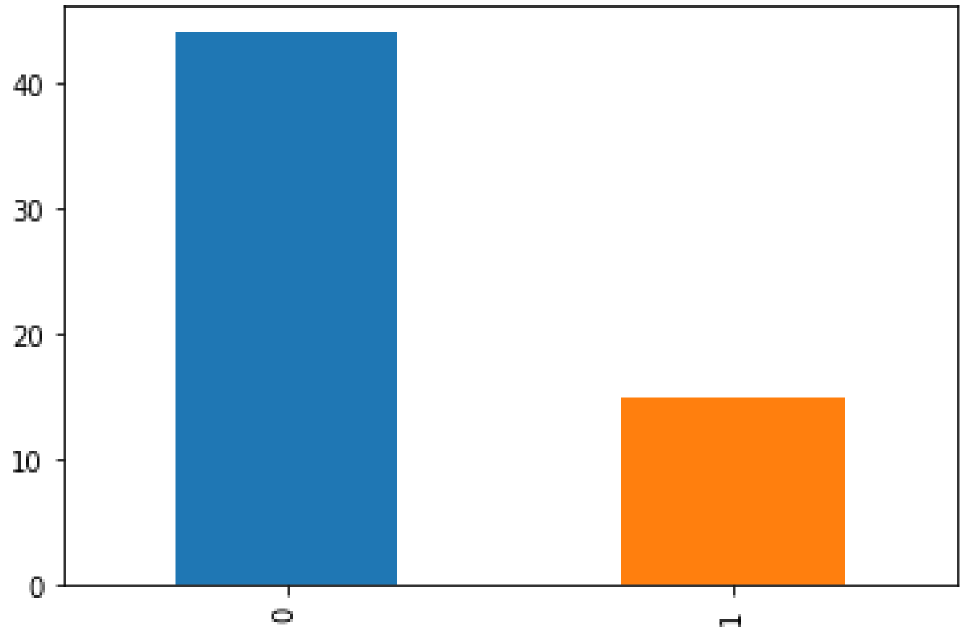
In order to do this task, we read the baseball data into a dataframe, then use
to get the distribution, then
with a bar chart type to draw the histogram.

```
In [21]: baseball_data=pd.read_csv('../baseball-data.csv', sep = ' ', header=None)
         baseball_dist = baseball_data[3].value_counts()
         baseball_dist
```

```
Out[21]: 0    44
         1    15
         Name: 3, dtype: int64
```

```
In [22]: baseball_dist.plot(kind='bar')
```

```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x7fde4a6ea518>
```



Ok, I don't know why the labels on the axes are rotated ninety degrees from what I think they should be, but I'll take it.

Question 5

Question 5 is asking us to obtain the sample proportion of left handed players. The sample proportion is $\frac{15}{43} \approx 0.35$. In this question the authors state that about 11% of males in America are left handed, and they ask us if the sample proportion seems high. Since our sample proportion approximately 35%, it seems that our sample has a high proportion of left-handed players. We think this is because players with a dominant left hand have an advantage over right-handed players in the game of baseball.

Question 6

Question 6 asks us to find the distribution and proportion of ones in the following data:

```
In [5]: q6_data = pd.Series([1, 1, 1, 1, 0, 0, 0, 1, 1 ,0,
        1, 0, 0, 0, 1, 0, 0, 0, 1, 0,
        0, 0, 1, 1, 1, 0, 0, 1, 1, 1,
        0, 0, 0, 1, 0, 1, 0, 0, 1, 0])
```

```
In [6]: q6_data.value_counts()
```

```
Out[6]: 0    22
        1    18
        dtype: int64
```

The first two lines of output in the cell above give the distribution. The sample proportion of ones is $\frac{18}{22-1} \approx 0.8571$.

References

[1] A.Abebe, J. Daniels, and J. W. McKean, Statistics and Data Analysis, 2nd ed. Western Michigan University,Kalamzoo, MI: Statistical Computation Lab (SCL), 2001. [E-book] Available: http://www.stat.wmich.edu/s160/hcopy/book.pdf.

[2] StackOverflow user Alexander, August 2017. Available: StackOverflow, https://stackoverflow.com/posts/31029857/revisions . [Accessed December 27, 2018].