Descriptive_Stats

May 19, 2022

1 Descriptive Statistics

Estimated time needed: **30** minutes

In this lab, you'll go over some hands-on exercises using Python.

1.1 Objectives

- Import Libraries
- Read in Data
- Lab exercises and questions

1.2 Import Libraries

All Libraries required for this lab are listed below. The libraries pre-installed on Skills Network Labs are commented. If you run this notebook in a different environment, e.g. your desktop, you may need to uncomment and install certain libraries.

[ ]:
```
#! mamba install pandas==1.3.3
#! mamba install numpy=1.21.2
#!  mamba install matplotlib=3.4.3-y
```

Import the libraries we need for the lab

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as pyplot
```

Read in the csv file from the URL using the request library

[2]:
```
ratings_url = 'https://cf-courses-data.s3.us.cloud-object-storage.appdomain.
 ↪cloud/IBMDeveloperSkillsNetwork-ST0151EN-SkillsNetwork/labs/teachingratings.
 ↪csv'
ratings_df=pd.read_csv(ratings_url)
```

1.3 Data Description

Variable	Description
minority	Does the instructor belong to a minority (non-Caucasian) group?
age	The professor’s age
gender	Indicating whether the instructor was male or female.
credits	Is the course a single-credit elective?
beauty	Rating of the instructor’s physical appearance by a panel of six students averaged across the six panelists and standardized to have a mean of zero.
eval	Course overall teaching evaluation score, on a scale of 1 (very unsatisfactory) to 5 (excellent).
division	Is the course an upper or lower division course?
native	Is the instructor a native English speaker?
tenure	Is the instructor on a tenure track?
students	Number of students that participated in the evaluation.
allstudents	Number of students enrolled in the course.
prof	Indicating instructor identifier.

1.4 Display information about the dataset

1. Structure of the dataframe
2. Describe the dataset
3. Number of rows and columns

print out the first five rows of the data

```
[3]: ratings_df.head()
```

```
[3]:  minority  age  gender credits    beauty  eval division native tenure  \
  0      yes   36  female    more  0.289916   4.3    upper    yes    yes
  1      yes   36  female    more  0.289916   3.7    upper    yes    yes
  2      yes   36  female    more  0.289916   3.6    upper    yes    yes
  3      yes   36  female    more  0.289916   4.4    upper    yes    yes
  4       no   59    male    more -0.737732   4.5    upper    yes    yes

     students  allstudents  prof  PrimaryLast  vismin  female  single_credit  \
  0        24           43     1            0       1       1              0
  1        86          125     1            0       1       1              0
  2        76          125     1            0       1       1              0
  3        77          123     1            1       1       1              0
  4        17           20     2            0       0       0              0

     upper_division  English_speaker  tenured_prof
  0               1                1             1
  1               1                1             1
  2               1                1             1
  3               1                1             1
  4               1                1             1
```

get information about each variable

[4]:

```
ratings_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 19 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   minority        463 non-null    object 
 1   age             463 non-null    int64  
 2   gender          463 non-null    object 
 3   credits         463 non-null    object 
 4   beauty          463 non-null    float64
 5   eval            463 non-null    float64
 6   division        463 non-null    object 
 7   native          463 non-null    object 
 8   tenure          463 non-null    object 
 9   students        463 non-null    int64  
 10  allstudents     463 non-null    int64  
 11  prof            463 non-null    int64  
 12  PrimaryLast     463 non-null    int64  
 13  vismin          463 non-null    int64  
 14  female          463 non-null    int64  
 15  single_credit   463 non-null    int64  
 16  upper_division  463 non-null    int64  
 17  English_speaker 463 non-null    int64  
 18  tenured_prof    463 non-null    int64  
dtypes: float64(2), int64(11), object(6)
memory usage: 68.9+ KB
```

get the number of rows and columns - prints as (number of rows, number of columns)

[5]:

```
ratings_df.shape
```

[5]:

```
(463, 19)
```

1.5 Lab Exercises

1.5.1 Can you identify whether the teachers' Rating data is a time series or cross-sectional?

Print out the first ten rows of the data

1. Does it have a date or time variable? - No - it is not a time series dataset
2. Does it observe more than one teacher being rated? - Yes - it is cross-sectional dataset

   The dataset is a Cross-sectional

[6]:

```
ratings_df.head(10)
```

```
[6]:   minority  age  gender credits    beauty  eval division native tenure  \
     0      yes   36  female    more  0.289916   4.3    upper    yes    yes
     1      yes   36  female    more  0.289916   3.7    upper    yes    yes
     2      yes   36  female    more  0.289916   3.6    upper    yes    yes
     3      yes   36  female    more  0.289916   4.4    upper    yes    yes
     4       no   59    male    more -0.737732   4.5    upper    yes    yes
     5       no   59    male    more -0.737732   4.0    upper    yes    yes
     6       no   59    male    more -0.737732   2.1    upper    yes    yes
     7       no   51    male    more -0.571984   3.7    upper    yes    yes
     8       no   51    male    more -0.571984   3.2    upper    yes    yes
     9       no   40  female    more -0.677963   4.3    upper    yes    yes

        students  allstudents  prof  PrimaryLast  vismin  female  single_credit  \
     0        24           43     1            0       1       1              0
     1        86          125     1            0       1       1              0
     2        76          125     1            0       1       1              0
     3        77          123     1            1       1       1              0
     4        17           20     2            0       0       0              0
     5        35           40     2            0       0       0              0
     6        39           44     2            1       0       0              0
     7        55           55     3            0       0       0              0
     8       111          195     3            1       0       0              0
     9        40           46     4            0       0       1              0

        upper_division  English_speaker  tenured_prof
     0               1                1             1
     1               1                1             1
     2               1                1             1
     3               1                1             1
     4               1                1             1
     5               1                1             1
     6               1                1             1
     7               1                1             1
     8               1                1             1
     9               1                1             1
```

1.5.2 Find the mean, median, minimum, and maximum values for students

Find Mean value for students

```
[7]: ratings_df['students'].mean()
```

```
[7]: 36.62419006479482
```

Find the Median value for students

```
[8]: ratings_df['students'].median()
```

```
[8]: 23.0
```

Find the Minimum value for students

```
[9]: ratings_df['students'].min()
```

```
[9]: 5
```

Find the Maximum value for students

```
[10]: ratings_df['students'].max()
```

```
[10]: 380
```

1.5.3 Produce a descriptive statistics table

```
[11]: ratings_df.describe()
```

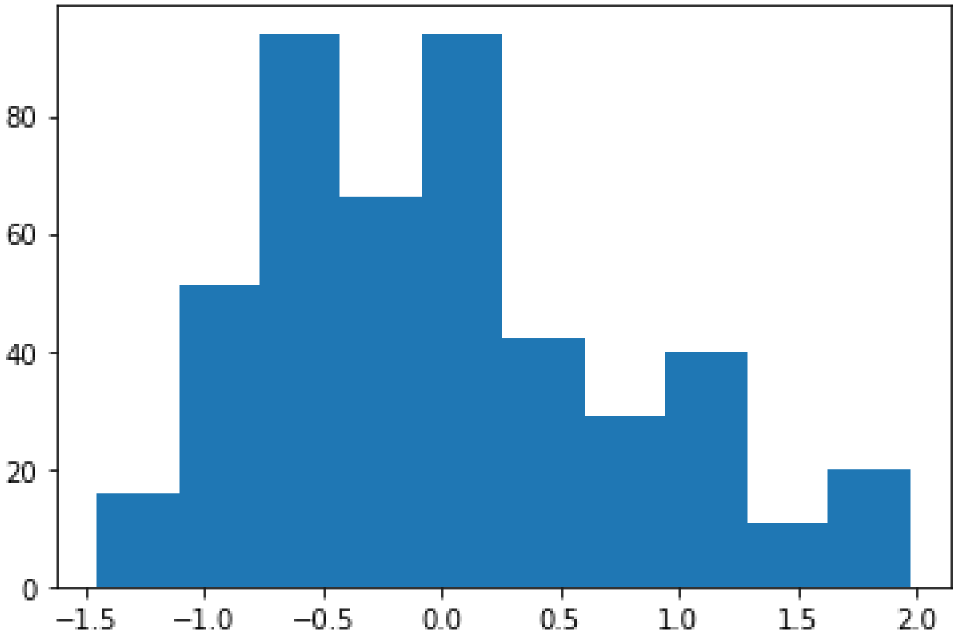
```
[11]:
                age        beauty        eval    students  allstudents  \
count  463.000000  4.630000e+02  463.000000  463.000000   463.000000
mean    48.365011  6.271140e-08    3.998272   36.624190    55.177106
std      9.802742  7.886477e-01    0.554866   45.018481    75.072800
min     29.000000 -1.450494e+00    2.100000    5.000000     8.000000
25%     42.000000 -6.562689e-01    3.600000   15.000000    19.000000
50%     48.000000 -6.801430e-02    4.000000   23.000000    29.000000
75%     57.000000  5.456024e-01    4.400000   40.000000    60.000000
max     73.000000  1.970023e+00    5.000000  380.000000   581.000000

             prof  PrimaryLast      vismin      female  single_credit  \
count  463.000000   463.000000  463.000000  463.000000     463.000000
mean    45.434125     0.203024    0.138229    0.421166       0.058315
std     27.508902     0.402685    0.345513    0.494280       0.234592
min      1.000000     0.000000    0.000000    0.000000       0.000000
25%     20.000000     0.000000    0.000000    0.000000       0.000000
50%     44.000000     0.000000    0.000000    0.000000       0.000000
75%     70.500000     0.000000    0.000000    1.000000       0.000000
max     94.000000     1.000000    1.000000    1.000000       1.000000

       upper_division  English_speaker  tenured_prof
count      463.000000       463.000000    463.000000
mean         0.660907         0.939525      0.779698
std          0.473913         0.238623      0.414899
min          0.000000         0.000000      0.000000
25%          0.000000         1.000000      1.000000
50%          1.000000         1.000000      1.000000
75%          1.000000         1.000000      1.000000
max          1.000000         1.000000      1.000000
```

1.5.4 Create a histogram of the beauty variable and briefly comment on the distribution of data

using the matplotlib library, create a histogram

```
[12]: pyplot.hist(ratings_df['beauty'])
```

```
[12]: (array([16., 51., 94., 66., 94., 42., 29., 40., 11., 20.]),
       array([-1.45049405, -1.10844234, -0.76639063, -0.42433892, -0.08228722,
               0.25976449,  0.6018162 ,  0.94386791,  1.28591962,  1.62797133,
               1.97002304]),
       <BarContainer object of 10 artists>)
```



here are few conclusions from the histogram most of the data for beauty is around the -0.5 and 0 the distribution is skewed to the right therefore looking at the data we can say the mean is close to 0

1.5.5 Does average beauty score differ by gender? Produce the means and standard deviations for both male and female instructors.

Use a group by gender to view the mean scores of the beauty we can say that beauty scores differ by gender as the mean beauty score for women is higher than men

```
[16]: ratings_df.groupby('gender').agg({'beauty':['mean', 'std', 'var']}).
       ↪reset_index()
```

```
[16]:    gender    beauty
                    mean      std       var
      0  female  0.116109  0.81781  0.668813
      1    male -0.084482  0.75713  0.573246
```

1.5.6 Calculate the percentage of males and females that are tenured professors. Will you say that tenure status differ by gender?

First groupby to get the total sum

```
[18]: tenure_count = ratings_df[ratings_df.tenure == 'yes'].groupby('gender').
        ↪agg({'tenure': 'count'}).reset_index()
```

Find the percentage

```
[19]: tenure_count['percentage'] = 100 * tenure_count.tenure/tenure_count.tenure.sum()
      tenure_count
```

```
[19]:    gender  tenure  percentage
      0  female     145   40.166205
      1    male     216   59.833795
```

1.6 Practice Questions

1.6.1 Question 1: Calculate the percentage of visible minorities are tenure professors. Will you say that tenure status differed if teacher was a visible minority?

```
[20]: ## insert code here
      tenure_count = ratings_df.groupby('minority').agg({'tenure': 'count'}).
        ↪reset_index()
      # Find the percentage
      tenure_count['percentage'] = 100 * tenure_count.tenure/tenure_count.tenure.sum()
      ##print to see
      tenure_count
```

```
[20]:   minority  tenure  percentage
      0       no     399   86.177106
      1      yes      64   13.822894
```

Double-click **here** for the solution.

1.6.2 Question 2: Does average age differ by tenure? Produce the means and standard deviations for both tenured and untenured professors.

```
[21]: ## insert code here
      ratings_df.groupby('tenure').agg({'age':['mean', 'std']}).reset_index()
```

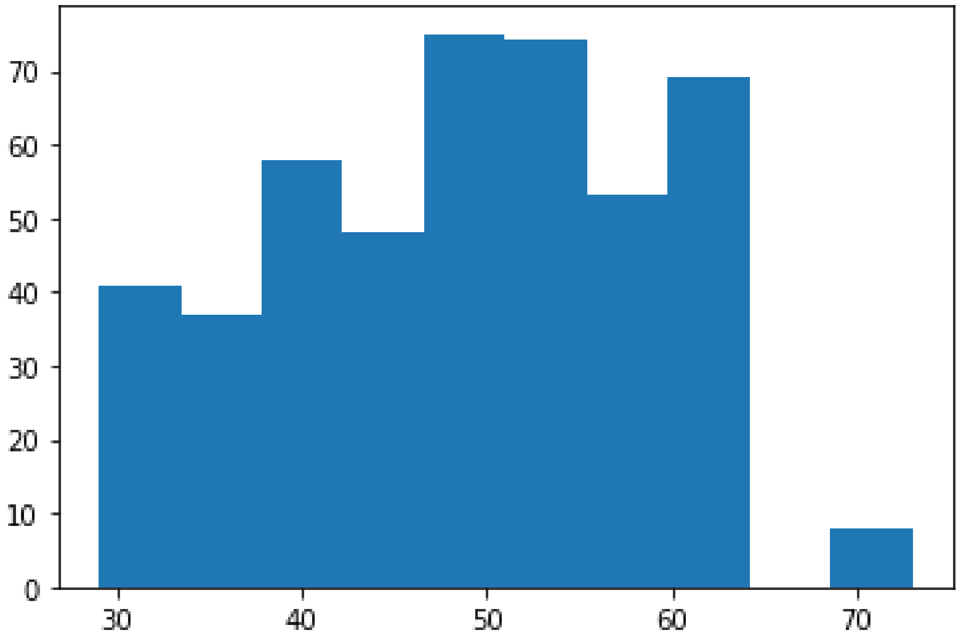
```
[21]:   tenure        age
                     mean        std
       0     no  50.186275   6.946372
       1    yes  47.850416  10.420056
```

Double-click **here** for the solution.

1.6.3 Question 3: Create a histogram for the age variable.

```
[22]: ## insert code here
      pyplot.hist(ratings_df['age'])
```

```
[22]: (array([41., 37., 58., 48., 75., 74., 53., 69.,  0.,  8.]),
       array([29. , 33.4, 37.8, 42.2, 46.6, 51. , 55.4, 59.8, 64.2, 68.6, 73. ]),
       <BarContainer object of 10 artists>)
```

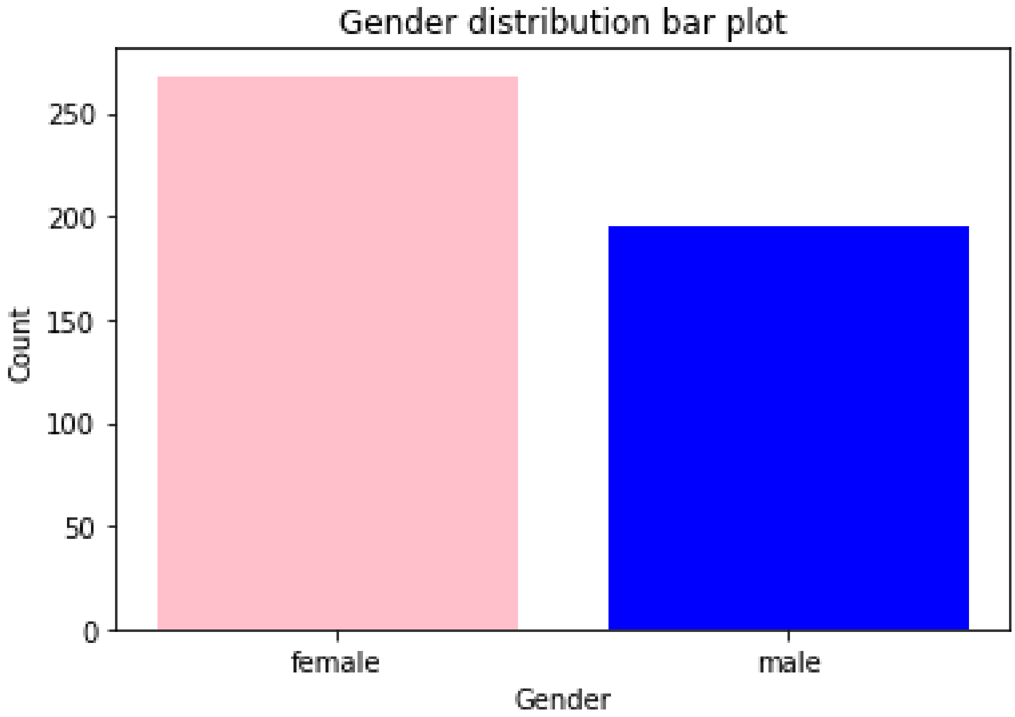


Double-click **here** for the solution.

1.6.4 Question 4: Create a bar plot for the gender variable.

```
[23]: ## insert code here
      pyplot.bar(ratings_df.gender.unique(),ratings_df.gender.
       ↪value_counts(),color=['pink','blue'])
      pyplot.xlabel('Gender')
      pyplot.ylabel('Count')
      pyplot.title('Gender distribution bar plot')
```

```
[23]: Text(0.5, 1.0, 'Gender distribution bar plot')
```



Double-click **here** for the solution.

> Note:Bar plot can be rendered vertically or horizontally. Try to replace **pyplot.bar** with **pyplot.barh** in the above cell and see the difference.

1.6.5 Question 5: What is the Median evaluation score for tenured Professors?

```
[ ]: ## insert code here
```

Double-click **here** for the solution.

1.7 Authors

Aije Egwaikhide is a Data Scientist at IBM who holds a degree in Economics and Statistics from the University of Manitoba and a Post-grad in Business Analytics from St. Lawrence College, Kingston. She is a current employee of IBM where she started as a Junior Data Scientist at the Global Business Services (GBS) in 2018. Her main role was making meaning out of data for their Oil and Gas clients through basic statistics and advanced Machine Learning algorithms. The highlight of her time in GBS was creating a customized end-to-end Machine learning and Statistics solution on optimizing operations in the Oil and Gas wells. She moved to the Cognitive Systems Group as a Senior Data Scientist where she will be providing the team with actionable insights using Data Science techniques and further improve processes through building machine learning solutions. She recently joined the IBM Developer Skills Network group where she brings her real-world experience to the courses she creates.

1.8 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-08-14	0.1	Aije Egwaikhide	Created the initial version of the lab
2022-05-10	0.2	Lakshmi Holla	Added exercise for Bar plot