</talentlabs>

Assignment 6 - Exploratory Data Analysis and Plotting Systems in Python (21 points)

Instructions

1. Answer the below question in the boxes if needed.
2. For coding exercises, code in a single google colab notebook and zip all your code before submission.
3. Please submit the assignment through TalentLabs Learning System.

Part 1: Concept Questions

Question 1 (1 point)
Which of these are graph plotting systems in Python? Select all that are correct.

1. Scikit - Learn
2. Pandas
3. Numpy
4. ggplot
5. Matplotlib
6. Tidyverse
7. Seaborn
8. Tableau

- Pandas
- Ggplot
- Matplotlib
- Seaborn

Question 2 (2 points)
Mark the steps that are part of a Exploratory Data Analysis project.

1. Build a model
2. Plot a histogram and boxplot to answer a question
3. Fetch data from a website
4. Make a dashboard for your stakeholders.
5. Removing Missing Values
6. Look at outliers.
7. Create tables and write data into a database

- Plot a histogram and boxplot to answer a question
- Removing Missing Values
- Look at outliers
- Fetch data from a website

Question 3 (2 points)

What inconsistencies do you spot in the `pakistan_intellectual_capital.csv` dataset?
We are looking for inconsistencies of the type:

- data entry errors (could be related to different ways of looking at value or data type related)
- missing values
- duplicates

Tell us what inconsistencies do you spot in:

1) Department
2) Designation
3) Year
4) Country

Example:
Field: Terminal Degree
Inconsistencies: Duplicates
Explanation: This column has duplicates such as phd and PhD, MS and MSCS (can be seen as data entry errors).

Tips
You can load the data and use pandas and numpy in python to see if the columns have any of the inconsistencies mentioned *(optional but recommended, code not required for submission)*, if they do mention them in the format "column_name:inconsistency type", if they don't - then write - "column_name:no inconsistency".

See and give an example of the inconsistency in the column. (2 points)

Note: You can upload the dataset in a google collab notebook using:

```
dataframe =
pd.read_csv("/content/pakistan_intellectual_capital.csv",index_col=0)
```

Ignore column 'S#'

Your Answer:

1. Department:

Inconsistency: There are variations in the naming convention for the same department.
Example: 'Computer Science & IT' and 'Computer Science' appear to refer to the same department but are labeled differently.

2. Designation:

Inconsistency: There are different designations like 'Assistant Professor' and 'Adjunct Professor', which might be the same role.
Example: 'Assistant Professor' and 'Adjunct Professor'.

3. Year:

Inconsistency: The Year column contains different formats, such as missing values (NaN) and decimal points (2005.0).
Example: Some rows have a year as 2005.0, while others are missing the year entirely.

4. Country:

Inconsistency: There might be different formats for the same country.
Example: South Korea and SouthKorea

Question 4 (2 points)
Match the examples below to where these types of analytics are (Descriptive or Predictive)?

Data Analytics Example	**Descriptive / Predictive**
Early Detection of Allergic Reactions	Predictive
What genres and TV shows interest their subscribers most	Descriptive
Change in Year over Year customer behavior	Descriptive
Forecasting Future Cash Flow for a company	Predictive

Question 5 (2 points)

Identify and label at least 5 elements of this graph. Annotate by editing the image.

Hint: Look at “elements of a graph” slides.

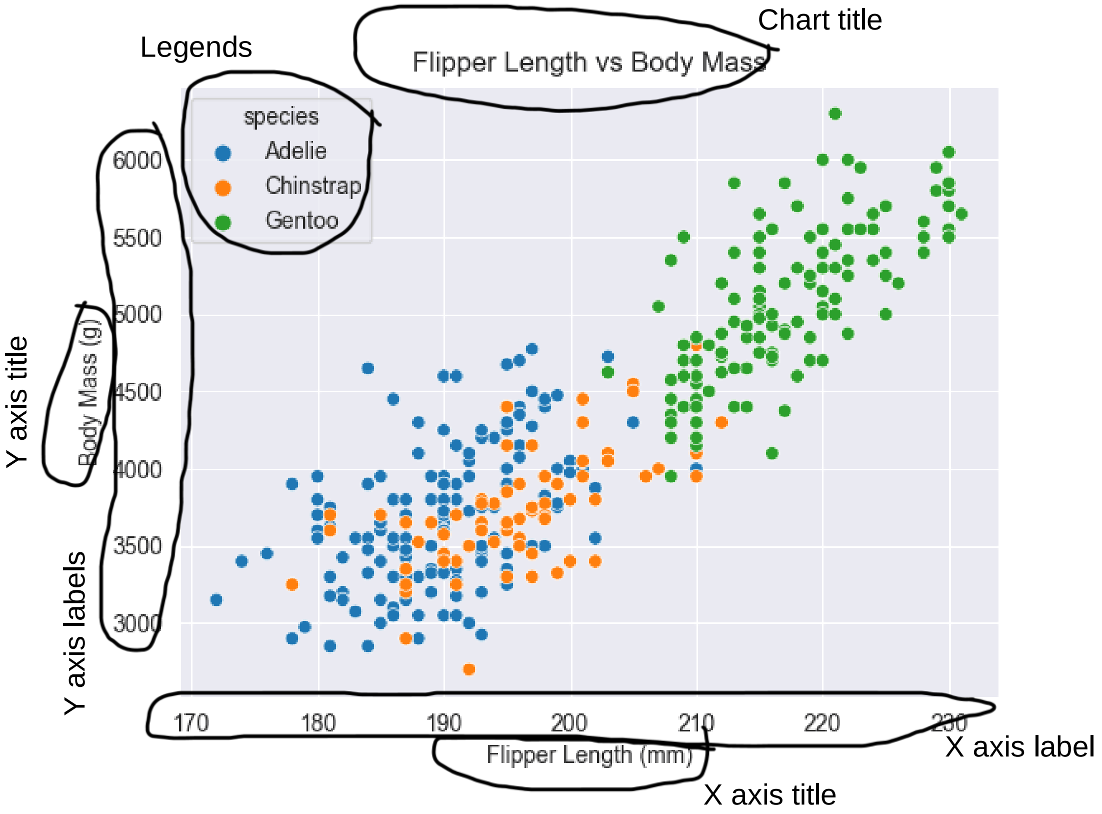


Chart title

Legends

Y axis title

Y axis labels

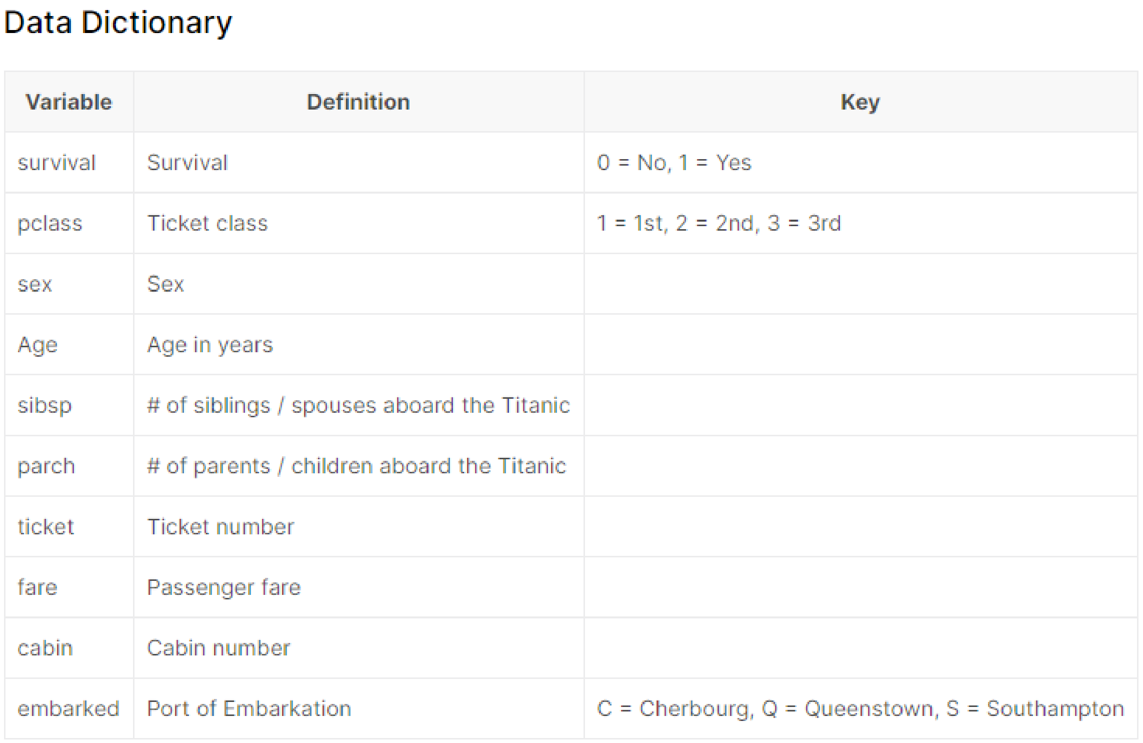
X axis label

X axis title

Question 6 (6 points)

Load the titanic dataset using seaborn using:

```
import seaborn as sns
df = sns.load_dataset('titanic');
```



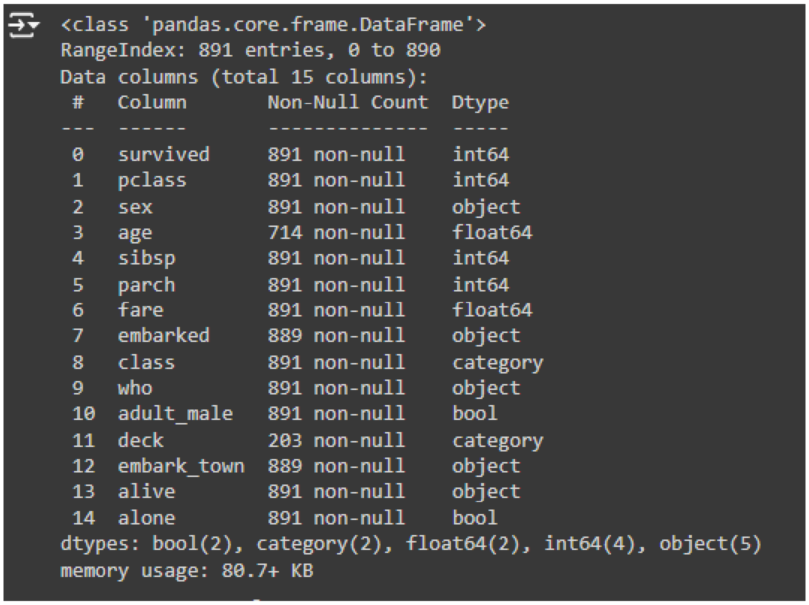
Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

1. How many columns and rows does the dataset have? (½ point)

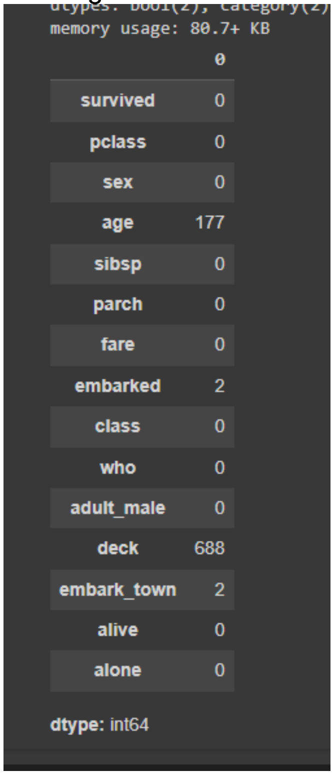
Rows: 891
Column: 15

2. Print the column data types and number of missing values (½ point)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

Missing values:



```
dtypes: bool(2), category(2)
memory usage: 80.7+ KB
                    0
   survived         0
     pclass         0
        sex         0
        age       177
      sibsp         0
      parch         0
       fare         0
   embarked         2
      class         0
        who         0
 adult_male         0
       deck       688
embark_town         2
      alive         0
      alone         0

dtype: int64
```

3. Run descriptive statistics on the dataset and report the mean and standard deviation for
    - age
    - fare, and

   And the most frequent value for
    - sex
    - embark_town.

   (2 point)

Age:
Mean - 29.69911764705882
Standard Deviation - 14.526497332334042

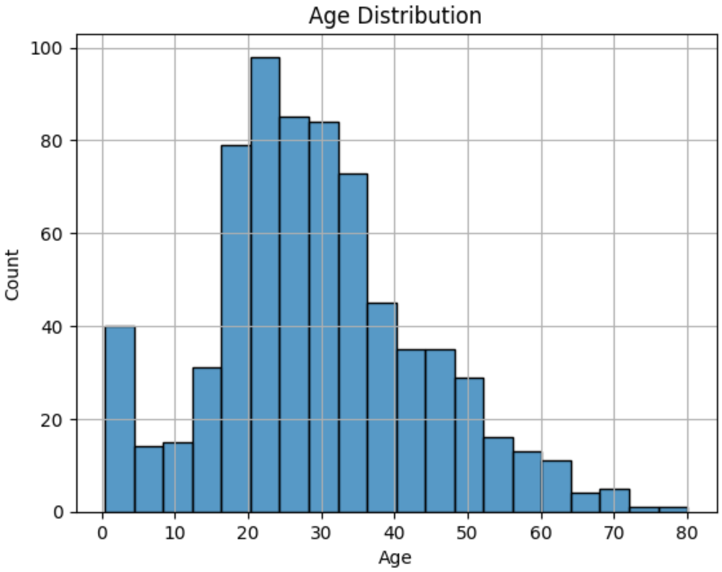
Fare:
Mean - 32.204207968574636
Standard Deviation - 49.6934285971809

Sex:
Most frequent value - male

Embark_town:
Most frequent value - Southampton

4. The most convenient way to take a quick look at a univariate distribution in seaborn is the distplot() function. By default, this will draw a histogram. Plot the histogram of age and add a title, x label, y label, gridlines. Count all the infants on board (age less than 3) and all the children ages 5-10. (3 points)

Screenshot of the chart:



Number of infants onboard (age less than 3): 24
Number of children ages 5-10: 24

Part 2: Coding Exercises

My link: https://colab.research.google.com/drive/1qW9n6StK4V4B5gdx1uXJHZ-dih14fg82?usp=sharing

For each of the exercises below, please write code in the same Google Colaboratory notebook or Jupyter Notebook, and create visualizations according to the instructions. You should also include the Google Colaboratory notebook or Jupyter Notebook in your submission.

Question 1 (2 points)

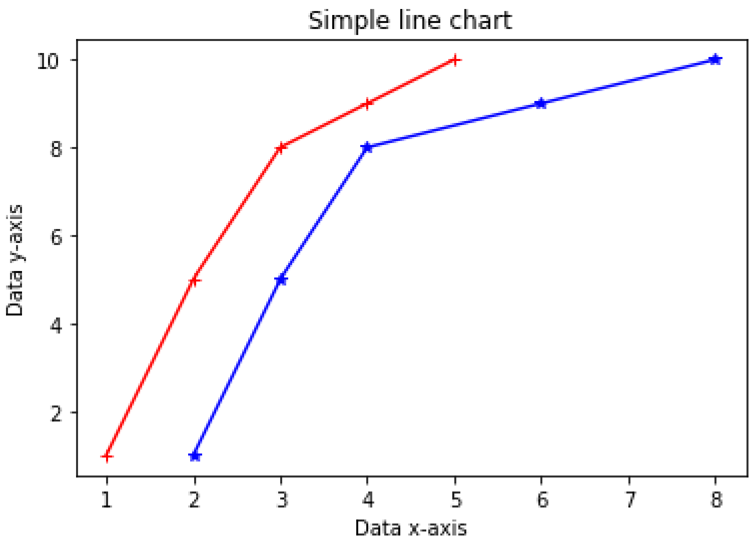
Write a Python program to plot several lines with different format styles in one command using arrays. Also give it a title, x label and y label shown below for the chart. You could use any color for each.

The arrays are below:

```
a = np.array([1,2,3,4,5])
b = np.array([2,3,4,6,8])
c = np.array([1,5,8,9,10])
```

Use a and b on the x axis, c on the y axis.

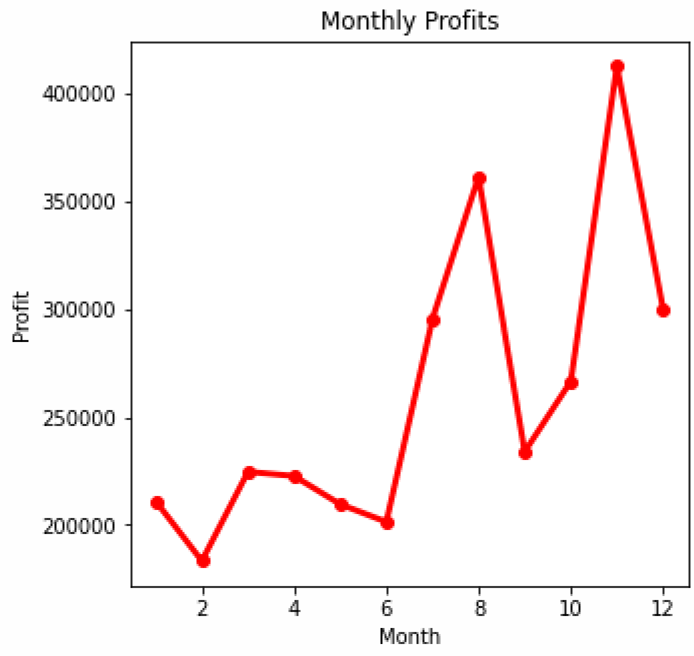
A sample output is included below:



Question 2 (2 points)
Ingest the company_sales_data.csv (attached in the assignment materials) and work to get total profit of all months and show line plot with the following style properties. Generated line plot must include following Style properties:

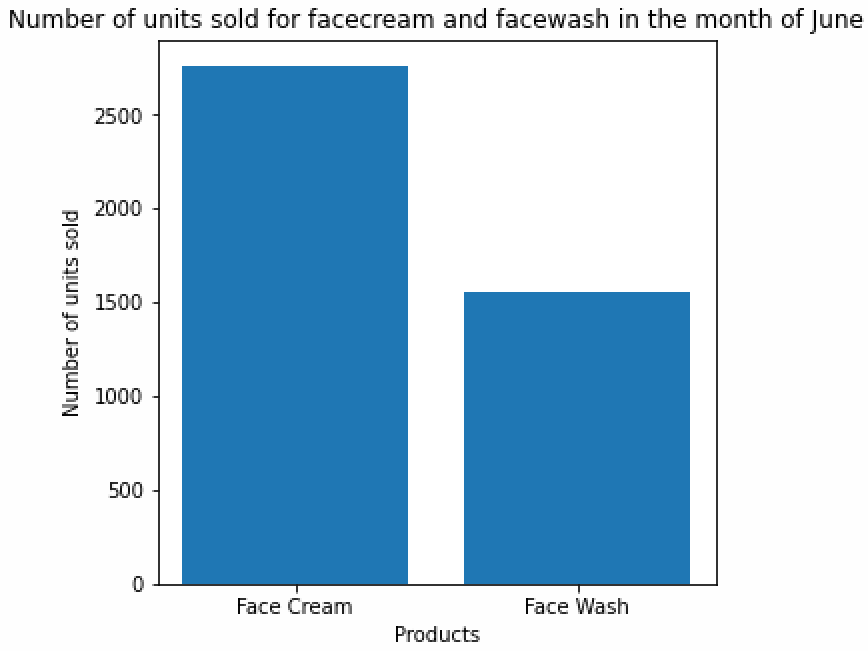
- X label name = Month
- Y label name = Profit
- Title Monthly Profits
- Add a circle marker.
- Make a line plot
- Line marker color as red
- Line width should be 3

Sample output:



Question 3 (2 points)
Ingest the company_sales_data.csv (attached in the assignment materials) and for each product column we see the number of units sold for various months, Read face cream and face wash product sales data and show it using the bar chart. The bar chart should display the number of units sold for facecream and face wash in the month of June. Add a separate bar for face cream and face wash in the same chart. Add title, labels mentioned below in sample. (2 points).

Sample output below:


Number of units sold for facecream and facewash in the month of June
Number of units sold
Face Cream
Face Wash
Products