ISSN:2790-1688 DOI: 10.56028/aetr.4.1.420.2023

A Brief Introduction to Exploratory Data Analysis

Xuanzhe Song

School of Zhejiang University of Technology, Hangzhou 310000, China.

1569194409@qq.com

Abstract. EDA is a process of tidying, processing and analyzing acquired data sets. Data science is an interdisciplinary field that combines scientific methods, systems, and processes from statistics, information science, and computer science to provide insight into phenomena through structured or unstructured data. This article will describe this process in detail in terms of purpose and method, including upload, tidy and visualization. In addition, the dataset of world Internet users was processed to conclude that the number of Internet users in Africa, the Americas and Europe accounted for more than 50% of the total population of each continent, Asia and Oceania for more than 30%, and finally the Middle East for 10 percent. In the example above, the basic profile of Internet users on each continent and the differences between each other can be seen very clearly and intuitively. This process of using data visualization to make it usable can be seen in all walks of life.

Keywords: EDA; dataset; Colab; visualization.

1. Introduction

This is an era of data, but also an era of competition based on data. More than 90 percent of Fortune 500 companies have established data analysis departments. IBM, Microsoft, Google and other well-known companies are actively investing in data business, establishing data departments and training data analysis teams. [1,2]Governments and more and more enterprises realize that data and information have become the intellectual assets and resources of enterprises, and data analysis and processing capabilities are becoming increasingly dependent on technical means.

The Internet itself has the characteristics of digitization and interactivity, which brings the possibility of data collection, collation and research.

The ultimate purpose of data analysis is to make better use of the data through this process, more effective analysis, so as to draw conclusions. For the individual, it helps to understand something or a phenomenon; For enterprises, it is used to guide the direction of improvement of products or services; For the government, it can show the macro phenomenon and help the country to make better policies.

Usually, raw data is collected from reality. After that, it will be processed to be a dataset with some cleaning. The next step is EDA, the theme of this paper. The primary purpose of this process is to transform the sorted data into models that are convenient and intuitive to understand, and to use these models to form visual reports for people to make decisions.

All data science projects (should) start with an exploratory data analysis (EDA), which Wikipedia defines as “an approach of analyzing data sets to summarize their main characteristics”, which often use statistical graphics and other data visualization methods.

2. How to accomplish EDA

Next I will show how to use Google Colab[3] to realize the exploration of datasets.

2.1 Find a suitable dataset

The first thing we need to know is how data sets come from. It can come from every aspect of life you can imagine, from transaction data（E-commerce data, Internet click data, company production data）from mobile data, from human data（emails, documents, pictures, audio, video）

Kaggle(https://www.kaggle.com/) supports a variety of dataset publication formats .Not only are open, accessible data formats better supported on the platform, they are also easier to work with for more people regardless of their tools.

Select ‘Datasets’, which is on the left side of the webpage, then we can see the datasets uploaded by other people. Click on it, and you can find the download

method in the upper right corner. At this point, the datasets are downloaded from the Internet to the computer, and step one is completed.

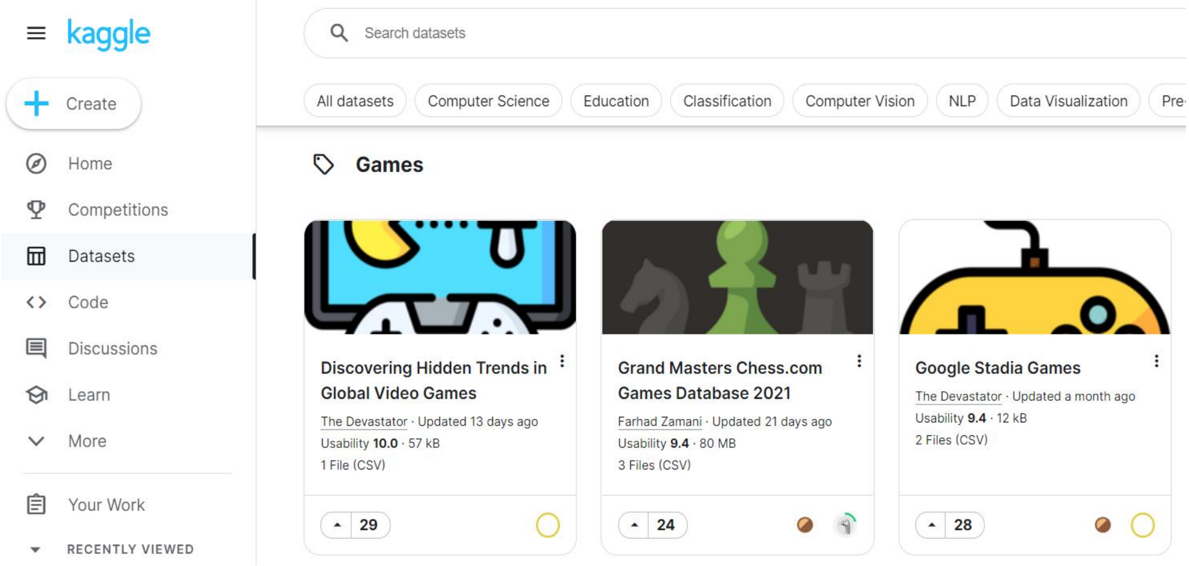


Figure 1. Kaggle Website Page

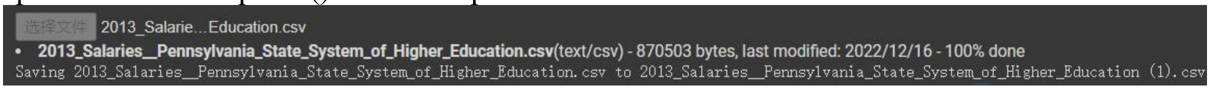
2.2 Upload the datasets

The file format we just downloaded is “.csv”, and the default software to open it on the PC is Excel. We need a way to upload this file to Colab.pandas is a numpy-based tool created for solving data analysis tasks. Here I would like to introduce pandas. pandas incorporates a large library and standard data models, providing the tools needed to efficiently manipulate large data sets.

Pandas is not part of Python's standard library but is by far the most common DF implementation for data science. Therefore, you need to import Pandas before using it. The standard alias is pd

The following is demonstration:

```
from google. colab import files
uploaded = files. upload()          // import datasets into colab's cloud document
```



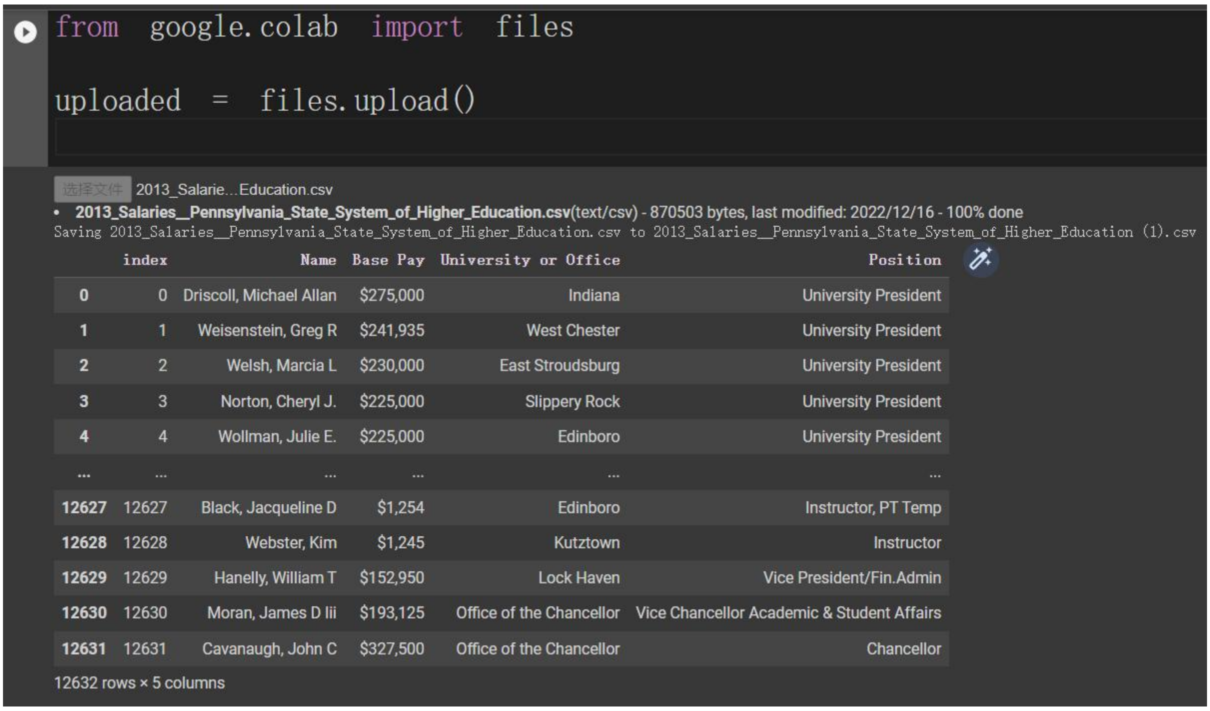
```
2013_Salarie...Education.csv
• 2013_Salaries__Pennsylvania_State_System_of_Higher_Education.csv(text/csv) - 870503 bytes, last modified: 2022/12/16 - 100% done
Saving 2013_Salaries__Pennsylvania_State_System_of_Higher_Education.csv to 2013_Salaries__Pennsylvania_State_System_of_Higher_Education (1).csv
```

Figure 2. Result of Successfully Uploading File

Figure 2 shows the result of a successful run of the import.The next step is to import pandas, use the csv file uploaded into the pandas framework, and name the file.

```
import pandas as pd             // Introduce panda database
wage_df = pd.read_csv()          // Give the uploaded file a name
wage_df          // Output this file
```

These two lines of code run as follows：



```
from  google.colab  import  files

uploaded  =  files.upload()
```

选择文件 2013_Salarie...Education.csv

- **2013_Salaries__Pennsylvania_State_System_of_Higher_Education.csv**(text/csv) - 870503 bytes, last modified: 2022/12/16 - 100% done

```
Saving 2013_Salaries__Pennsylvania_State_System_of_Higher_Education.csv to 2013_Salaries__Pennsylvania_State_System_of_Higher_Education (1).csv
```

	index	Name	Base Pay	University or Office	Position
0	0	Driscoll, Michael Allan	$275,000	Indiana	University President
1	1	Weisenstein, Greg R	$241,935	West Chester	University President
2	2	Welsh, Marcia L	$230,000	East Stroudsburg	University President
3	3	Norton, Cheryl J.	$225,000	Slippery Rock	University President
4	4	Wollman, Julie E.	$225,000	Edinboro	University President
...	...	...	...	...	...
12627	12627	Black, Jacqueline D	$1,254	Edinboro	Instructor, PT Temp
12628	12628	Webster, Kim	$1,245	Kutztown	Instructor
12629	12629	Hanelly, William T	$152,950	Lock Haven	Vice President/Fin.Admin
12630	12630	Moran, James D Iii	$193,125	Office of the Chancellor	Vice Chancellor Academic & Student Affairs
12631	12631	Cavanaugh, John C	$327,500	Office of the Chancellor	Chancellor

12632 rows × 5 columns

Figure 3. Running Result of Dataset

As you can see in Figure 3, part of the datasets has been shown.

2.3 Tidy the datasets

Actually, we get datasets that are too cluttered, or don't qualify for 'Tidy Datasets'[4], so we need to tidy them up.

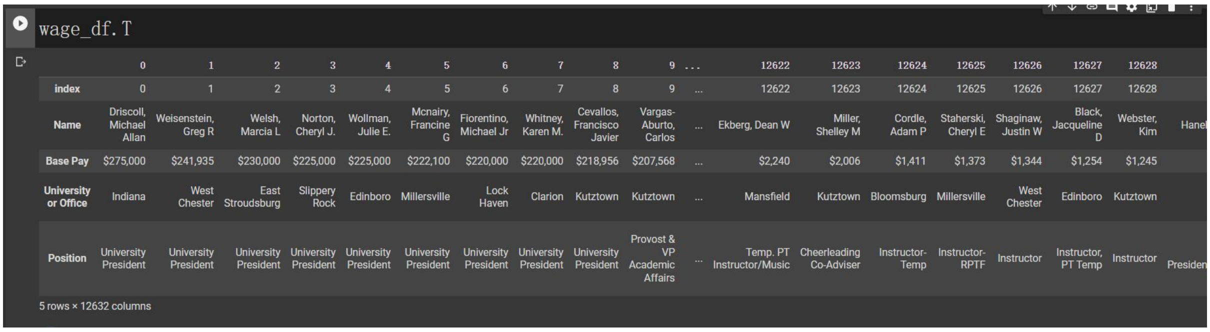
In tidy data:

Each variable must have its own column.

Each observation must have its own row.

Each type of observational unit forms a table

There is a lot of code in python that can do this, such as using T to transpose datasets and using strip to remove bits of a variable.

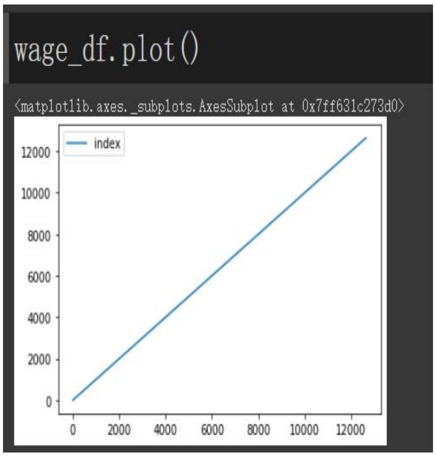


```
wage_df.T
```

	0	1	2	3	4	5	6	7	8	9	...	12622	12623	12624	12625	12626	12627	12628	
index	0	1	2	3	4	5	6	7	8	9	...	12622	12623	12624	12625	12626	12627	12628	
Name	Driscoll, Michael Allan	Weisenstein, Greg R	Welsh, Marcia L	Norton, Cheryl J.	Wollman, Julie E.	Mcnairy, Francine G	Fiorentino, Michael Jr	Whitney, Karen M.	Cevallos, Francisco Javier	Vargas-Aburto, Carlos	...	Ekberg, Dean W	Miller, Shelley M	Cordle, Adam P	Staherski, Cheryl E	Shaginaw, Justin W	Black, Jacqueline D	Webster, Kim	Hanel
Base Pay	$275,000	$241,935	$230,000	$225,000	$225,000	$222,100	$220,000	$220,000	$218,956	$207,568	...	$2,240	$2,006	$1,411	$1,373	$1,344	$1,254	$1,245	
University or Office	Indiana	West Chester	East Stroudsburg	Slippery Rock	Edinboro	Millersville	Lock Haven	Clarion	Kutztown	Kutztown	...	Mansfield	Kutztown	Bloomsburg	Millersville	West Chester	Edinboro	Kutztown	
Position	University President	University President	University President	University President	University President	University President	University President	University President	University President	Provost & VP Academic Affairs	...	Temp. PT Instructor/Music	Cheerleading Co-Adviser	Instructor-Temp	Instructor-RPTF	Instructor	Instructor, PT Temp	Instructor	Presiden

5 rows × 12632 columns

Figure 4. Transform the Rows and Columns of the Dataset



```
wage_df.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff631c273d0>
```

Figure 5. Visualization of Data

You can also graph datasets using x.plot (), but that's obviously not the way to do it. There is no way to select specific variables or add additional elements to the diagram to make it more intuitive. So next step is how to solve this problem.

2.4 Visualize the Data

Now comes the final and most critical step in EDA, visualizing datasets. The common library used is Seaborn[5]. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.It is possible for most Seaborn plotting functions to work with data that has been constructed or loaded using the Pandas or Numpy libraries (e.g. data frames and arrays), as well as built-in Python data structures (e.g. lists and dictionaries).

There are many kinds of plots in Seaborn, such as relplot to show the relationship between variables, displot to show the state of data distribution and catplot to show classified data.Here I use relplot and I choose another datasets because two columns of data is needed to form a plot. This dataset indicates how many Internet users each country has and the percentage of the total population in this country.

Table 1. Partial Display of Dataset

	Country	Region	Population	Internet Users	% of Population
0	World	NaN	7920539977	5424080321	68.48
1	Afganistan	Asia	40403518	9237489	22.86
2	Albania	Europe	2872758	2191467	76.28

```
mport seaborn as sns                              // Import seaborn database
import numpy as np                                   // Import numpy database
import pandas as pd                               // Import pandas database
import matplotlib.pyplot as plt                    // Import database
x0.Region.value_counts().plot(kind='bar')        // Name the file
plt.show()                                           // Output chart
```

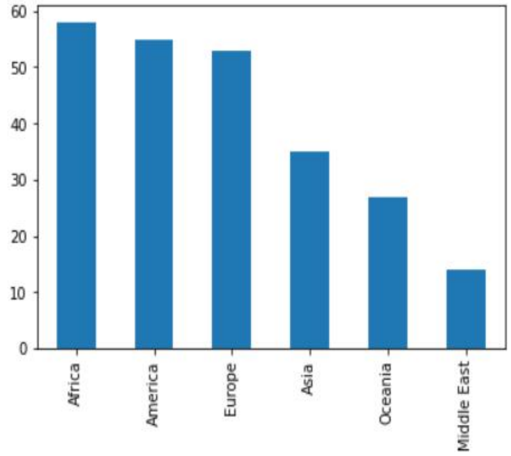


Figure 6. The proportion of people using the Internet in each state

From the chart above we can see the percentage of Internet users in the total population of each continent.

sns.pairplot(d, kind='scatter', hue=None) // Output the dot plot

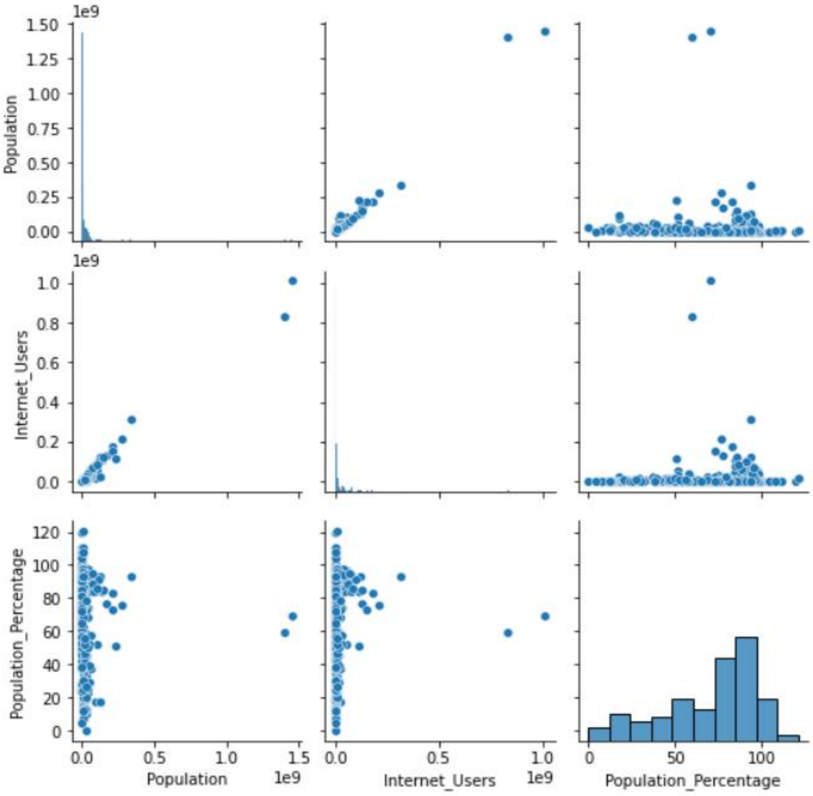


Figure 7. Graph of Each Two Variables

The plot uses two different variables as the horizontal and vertical coordinates to show the relationship between them. From Figure 7, we can see that most plots are not easy to draw conclusions, perhaps we can distinguish them by different continents

3. Conclusion

In conclusion, pandas is very helpful for compiling and displaying datasets. It is a very powerful library for handling data. Seaborn, meanwhile, is an indispensable library in the process of visualizing datasets. It plays a key role in all the steps of EDA, in my opinion. The processing of datasets in the final analysis is to express the result as a plot, so as to draw a clear conclusion intuitively and make a decision from a human perspective.

References

[1] Morgenthaler S. Exploratory data analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2009, 1(1): 33-44.

[2] Chatfield C. Exploratory data analysis[J]. European journal of operational research, 1986, 23(1): 5-13.

[3] Hoyos-Rivera G J, Gomes R L, Willrich R, et al. Colab: A new paradigm and tool for collaboratively browsing the web[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2006, 36(6): 1074-1085.

[4] García S, Luengo J, Herrera F. Data preprocessing in data mining[M]. 2015.

[5] [Bisong E, Bisong E. Matplotlib and seaborn[J]. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 2019: 151-165.