Data Cleaning using Python with Pandas Library.

According to this article, data cleaning and organizing constitutes 57% of the total weight when it comes to the part of the data science.

Tanu N Prabhu Follow
Jun 21, 2019 · 4 min read ★

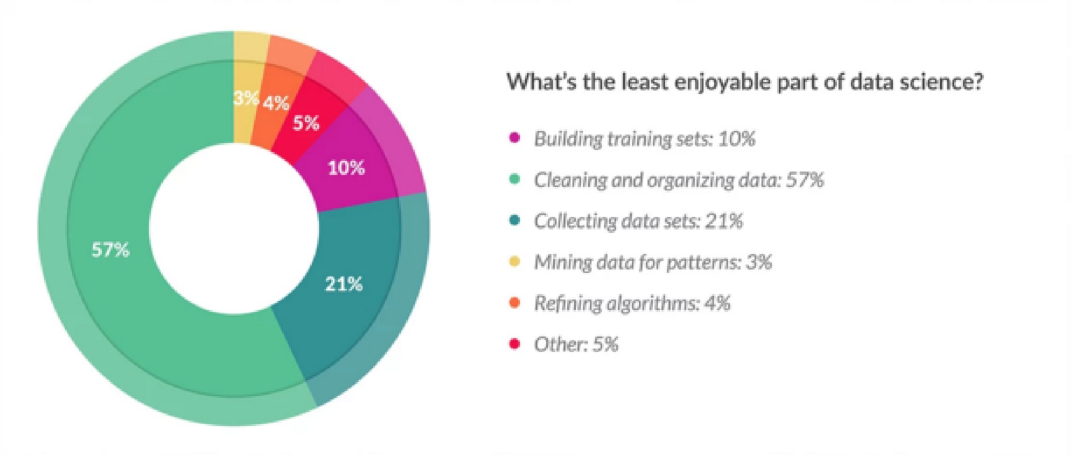


Image Credits: whatsthebigdata.com

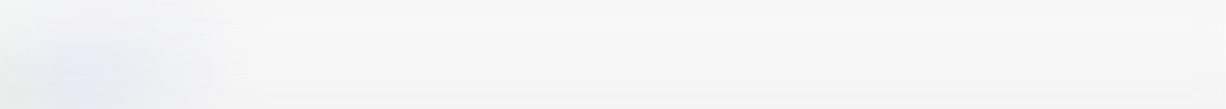
The entire data cleaning process is divided into sub-tasks as shown below.

1. *Importing the required libraries.*
2. *Getting the data-set from a different source (Kaggle) and displaying the dataset.*
3. *Removing the unused or irrelevant columns.*
4. *Renaming the column names as per our convenience.*
5. *Replacing the value of the rows and make it more meaningful.*

Even though this tutorial is small, but it's a good way to start on small things and get our hands dirty later on. I will make sure that everyone with no prior experience in python programming or don't know what is data science or data cleaning can easily understand this tutorial. I did not know python in the first place, so even for me, this was a good place to start. One thing with python is that the code is self-explanatory, your focus should not be what the code does, because the code pretty much says what it does, rather you should tell why did you choose to do this, the "**why**" factor is important than the "**what**" factor. Moreover, the entire source code can be found in my **GitHub**.

Step 1: Importing the required libraries.

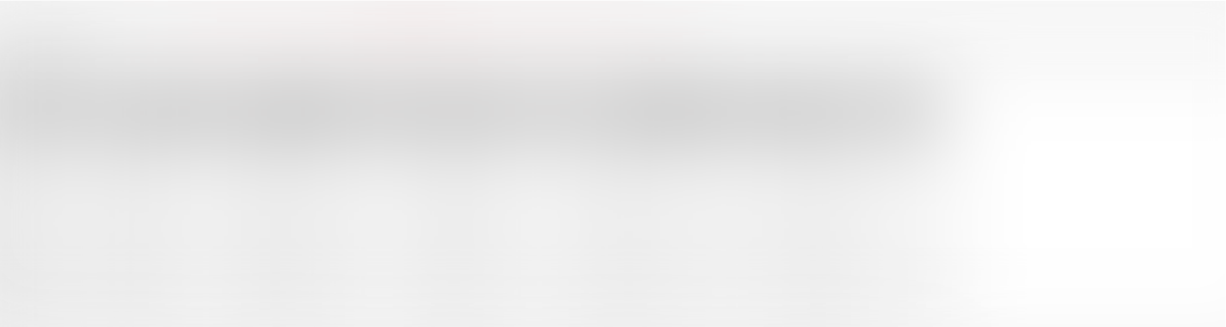
This step involves just importing the required libraries which are pandas, numpy and csv. These are the necessary libraries when it comes to data science.



Step 2: Getting the data-set from a different source and displaying the data-set.

This step involves getting the data-set from a different source, and the link for the data-set is provided below.

Data-set Download



Note: If you are using **Jupyter Notebook** to practice this tutorial then there should be no problem to read the CSV file. But if you are a Google fan like me, then you ought to use **Google Colab** which is the best according to me, for practicing data science, then you must follow some steps in order to load or read the CSV file. So this article helps you to solve this issue. I personally recommend everybody to go through this article. I followed the second step to read the CSV file in that article. Choose the best one and start working.

. . .

Step 3: Removing the unused or irrelevant columns

This step involves removing irrelevant columns such as cp, fbs, thalach, and many more, and the code is pretty much self-explanatory.

. . .

Step 4: Renaming the column names as per our convenience.

This step involves renaming the column names because many column names are kinda confusing and hard to understand.

. . .

Step 5: Replacing the value of the rows if

necessary.

This step involves replacing the incomplete values or making the values more readable, such as in here the **Sex** field consists of the values **1** and **0** being **1** as Male and **0** as Female, but it often seems ambiguous for the third person, so changing the value to an understandable one is a good idea.

So the above is the overall simple data cleaning process obviously this is not the actual cleaning process at an industry level, but this is a good start, so let’s start from small and then go for huge data-sets which then involves more cleaning process. This was just to give an idea as to how the process of data cleaning looks like in a beginners perspective. Thank you guys for spending your time reading my article, stay tuned for more updates. Let me know what is your opinion about this tutorial in the comment section below. Also if you have any doubts regarding the code, comment section is all yours. Have a nice day.

216 claps

WRITTEN BY

Tanu N Prabhu

Follow

MSc Student | University of Regina, Canada | Loves Data Science | Discord Group:https://discord.gg/Kgkske | GitHub: https://github.com/Tanu-N-Prabhu/Python

towards data science

Towards Data Science

Follow

A Medium publication sharing concepts, ideas, and codes.

See responses (1)

More From Medium

3 Ways to Get Real-Life Data Science Experience Before Your First Job

Terence S in Towards Data Science

How You Should Read Research Papers According To Andrew Ng (Stanford Deep Learning Lectures)

Richmond Alake in Towards Data Science

Features You Likely Don’t Use in Python 3 — But You Should

Amritansh Sagar in Towards Data Science

Ten SQL Concepts You Should Know for Data Science Interviews

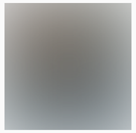
Terence S in Towards Data Science

Sktime: a Unified Python Library for Time Series Machine Learning

Alexandra Amidon in Towards Data Science

Will AutoML Be the End of Data Scientists?

Frederik Bussler in Towards Data Science



Top 9 Data Science certifications to know about in 2020

Rashi Desai in Towards Data Science

The Most Elegant Python Object-Oriented Programming

Christopher Tao in Towards Data Science