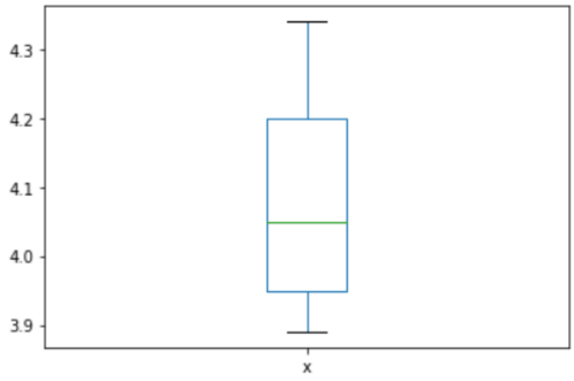
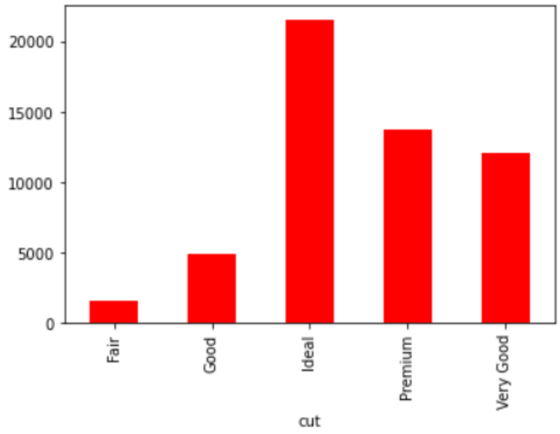
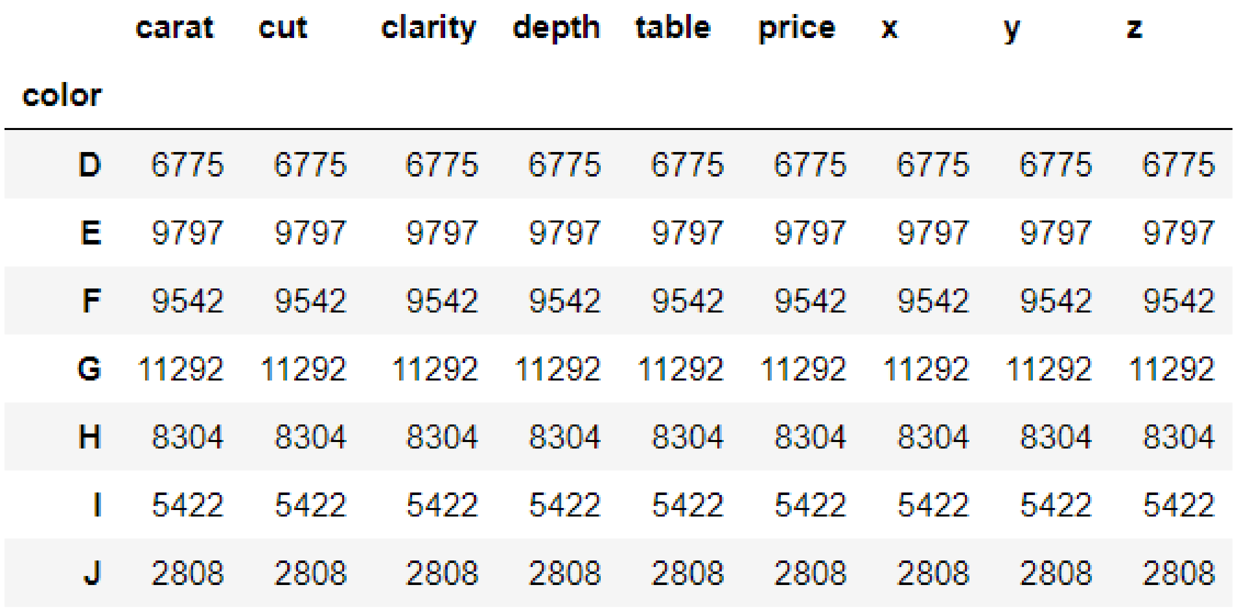
Project Name: Data-Wrangling Fundamentals

Table of Contents

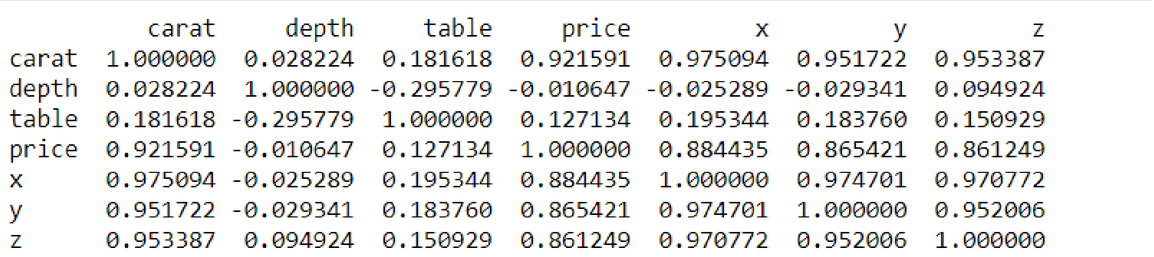
Demo



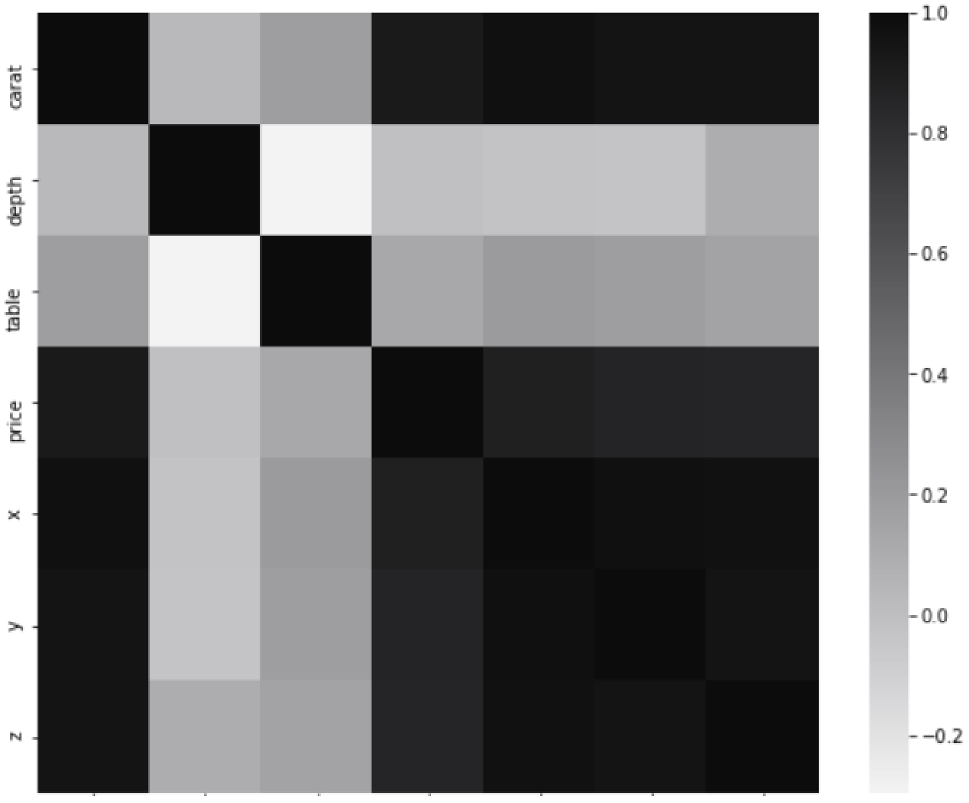




	carat	cut	clarity	depth	table	price	x	y	z
color									
D	6775	6775	6775	6775	6775	6775	6775	6775	6775
E	9797	9797	9797	9797	9797	9797	9797	9797	9797
F	9542	9542	9542	9542	9542	9542	9542	9542	9542
G	11292	11292	11292	11292	11292	11292	11292	11292	11292
H	8304	8304	8304	8304	8304	8304	8304	8304	8304
I	5422	5422	5422	5422	5422	5422	5422	5422	5422
J	2808	2808	2808	2808	2808	2808	2808	2808	2808



```
          carat     depth     table     price         x         y         z
carat  1.000000  0.028224  0.181618  0.921591  0.975094  0.951722  0.953387
depth  0.028224  1.000000 -0.295779 -0.010647 -0.025289 -0.029341  0.094924
table  0.181618 -0.295779  1.000000  0.127134  0.195344  0.183760  0.150929
price  0.921591 -0.010647  0.127134  1.000000  0.884435  0.865421  0.861249
x      0.975094 -0.025289  0.195344  0.884435  1.000000  0.974701  0.970772
y      0.951722 -0.029341  0.183760  0.865421  0.974701  1.000000  0.952006
z      0.953387  0.094924  0.150929  0.861249  0.970772  0.952006  1.000000
```



Overview

This is a Data Wrangling Basics.

Data wrangling involves processing the data in various formats like - merging, grouping, concatenating etc. for the purpose of analysing or getting them ready to be used with another set of data. Python has built-in features to apply these wrangling methods to various data sets to achieve the analytical goal.

This repository contains the code for Data Wrangling using python's various libraries.

It used numpy, pandas and matplotlib libraries.

These libraries help to perform individually one particular functionality.

Numpy is used for working with arrays. It stands for Numerical Python.

Pandas objects rely heavily on Numpy objects.

Matplotlib is a plotting library.

The purpose of creating this repository is to gain insights into how to wrangle or transform data.

These python libraries raised knowledge in discovering these libraries with practical use of it.

It leads to growth in my ML repository.

These above screenshots will help you to understand flow of output.

Motivation

The reason behind building this is, I know that wrangling costs analytics professionals as much as 80% of their time therefore, to become ML Professional this is one of the skillsets necessary in bucket-list. One very common truth of life is that, when we hear proverbs or saying, we do not understand it much or completely rather we can absorb it fully when we go through it. Right. That means, here information from advice format to practical self-experience format got transferred and then we could understand it completely. Similarly, in Data Wrangling is the process of converting and mapping data from its raw form to another format with the purpose of making it more valuable and appropriate for advance tasks such as Data Analytics and Machine Learning. For companies, their clients are important and so for employee because its all about business and decisions taken for business. Data Wrangling exactly here comes for rescue as Data wrangling is the art of providing the right information to business analysts to make the right decision on time. Data wrangling also provides organisations with the right information in a short span of time to access the right information thereby helping make strategic decisions for the business. One of the most common steps taken in data science work is data wrangling to solve complex business problems.

One of the important goal of Data Wrangling is to drive better decisions based on data in short time span and that is what I am trying to achieve by doing this project and that is also one of the important qualities that employee should have. Performing data wrangling in right direction will rescue an organization 6 on 10 times from drowning. When analyst can reach to precise decision through this process then it is good deal. Hence, I continue to gain knowledge while practicing the same and spread literary wings in tech-heaven.

Technical Aspect

Numpy contains a multi-dimensional array and matrix data structures. It works with the numerical data. Numpy is faster because is densely packed in memory due to its homogeneous type. It also frees the memory faster.

Pandas module mainly works with the tabular data. It contains Data Frame and Series. Pandas is 18 to 20 times slower than Numpy. Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data.

Matplotlib is used for EDA. Visualization of graphs helps to understand data in better way than numbers in table format. Matplotlib is mainly deployed for basic plotting. It consists of bars, pies, lines, scatter plots and so on. Inline command display visualization inline within frontends like in Jupyter Notebook, directly below the code cell that produced it.

Installation

Using intel core i5 9th generation with NVIDIA GFORECE GTX1650.

Windows 10 Environment Used.

Already Installed Anaconda Navigator for Python 3.x

The Code is written in Python 3.8.

If you don't have Python installed then please install Anaconda Navigator from its official site. If you are using a lower version of Python you can upgrade using the pip package, ensuring you have the latest version of pip, *python* -m pip install --*upgrade pip and press Enter.*

Run/How to Use/Steps

Keep your internet connection on while running or accessing files and throughout too.
Follow this when you want to perform from scratch.

Open Anaconda Prompt, Perform the following steps:

```
cd <PATH>
pip install numpy
pip install pandas
pip install matplotlib
```

You can also create requirement.txt file as, pip freeze > requirements.txt
run files.

Follow this when you want to just perform on local machine.
Download ZIP File.
Right-Click on ZIP file in download section and select Extract file option, which will unzip file.
Move unzip folder to desired folder/location be it D drive or desktop etc.
Open Anaconda Prompt, write cd <PATH> and press Enter.
eg: cd C:\Users\Monica\Desktop\Projects\Python Projects 1\13)Data_Wrangling\Project_1_Data_Wrangling_Fundamentals

In Anconda Prompt, pip install -r requirements.txt to install all packages.
Open in Jupyter Notebook, <filename>.ipynb
That is,

Open in Jupyter Notebook, 1)Fundamentals_Data_Wrangling.ipynb
This takes diamonds.csv file as input dataset.
Please be careful with spellings or numbers while typing filename and easier is just copy filename and then run it to avoid any silly errors.
Note: cd <PATH>
[Go to Folder where file is. Select the path from top and right-click and select copy option and paste it next to cd one space <path> and press enter, then you can access all files of that folder] [cd means change directory]

Directory Tree/Structure of Project

Folder: 13)Data_Wrangling > Project_1_Data_Wrangling_Fundamentals

1)Fundamentals_Data_Wrangling.ipynb

To Do/Future Scope

Can try with another complex dataset.

Technologies Used/System Requirements/Tech Stack

NumPy

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

matplotlib

Credits

PyLadiesRemote Webcasts Channel