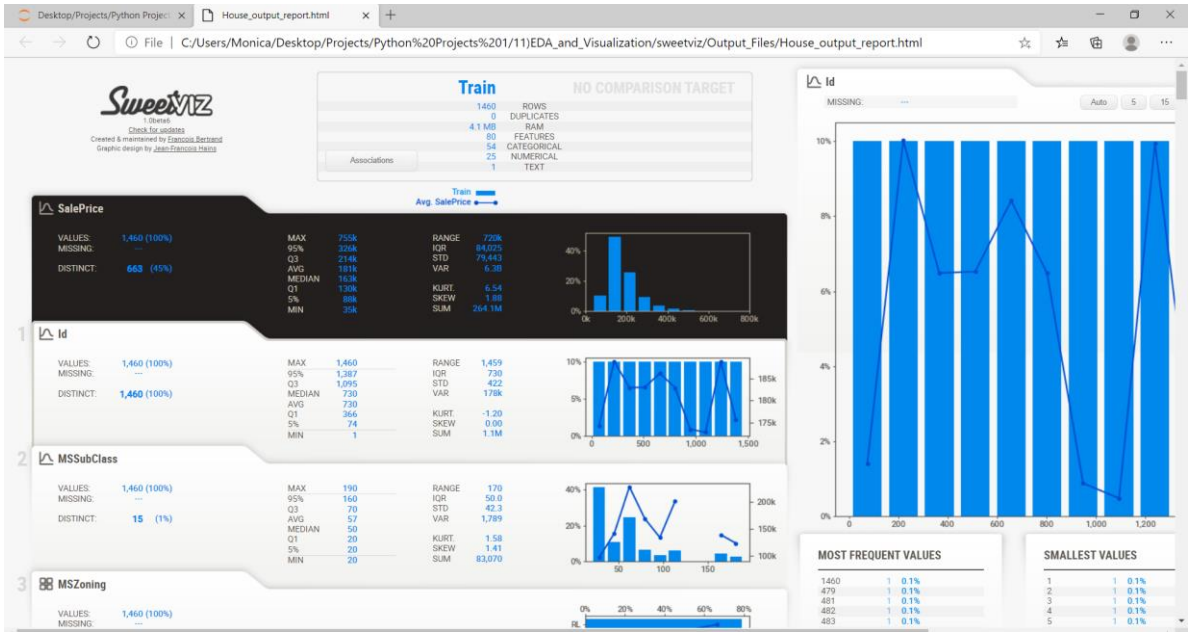
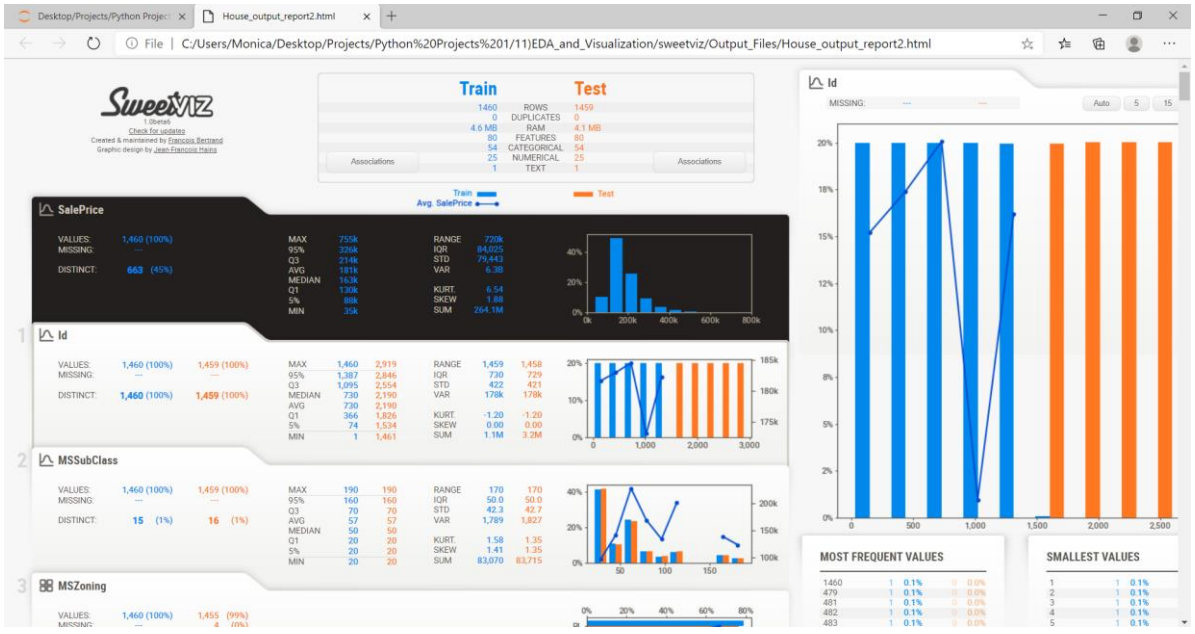
Project Name: EDA With Visualization Using Sweetviz in 1 line - 2

Table of Contents

Demo





Overview

This is an Exploratory Data Analysis with Visualization using Titanic Dataset.
EDA is essentially a type of storytelling for statisticians.
It allows us to uncover patterns and insights, often with visual methods, within data.
EDA is often the first step of the data modelling process.
This repository contains the code for EDA along with Visualization using python's various libraries.
It used pandas and sweetviz libraries.
These libraries help to perform individually one particular functionality.
Pandas objects rely heavily on Numpy objects.
Sweetviz is an open source Python library that generates beautiful, high-density visualizations to kickstart EDA with a single line of code. Output is a fully self-contained HTML application.
pandas-profiling will not work properly when you have many features or records in dataset because it ran out of memory.
pandas-profiling is integrated in sweetviz, so sweetviz is built on top of pandas-profiling.
The purpose of creating this repository is to gain insights into EDA and Visualization.
These python libraries raised knowledge in discovering these libraries with practical use of it.
It leads to growth in my ML repository.
This above screenshot will help you to understand flow of output.

Motivation

The reason behind building this is, to maximize I as analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract. It is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format. EDA is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important for me especially when I arrive at modelling the data in order to apply Machine learning. Another motive is, as a master's student I have learnt Data Mining Subject which has somewhere led me to also learn about EDA. Though, EDA and Data Mining has difference. EDA distinguishes itself from data mining, even though the two are closely related, as many EDA techniques have been adopted into data mining. Also, the goals of the two are very similar: EDA indeed makes sure that you explore the data in such a way that interesting features and relationships between features will become clearer. In EDA, you typically explore and compare many different variables with a variety of techniques to search and find systematic patterns. Data mining, on the other hand, is concerned with extracting patterns from the data. Those patterns provide insights into relationships between variables that can be used to improve business decisions. Also, in both cases, you have no a priori expectations or expectations that are not complete about the relations between the variables. However, in general, Data Mining can be said to be more application-oriented, while EDA is concerned with the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. As a result, Data Mining accepts a "black box" approach to data exploration and doesn't only use techniques that are also used in EDA but also

techniques such as Neural Networks to generate valid predictions but don't identify the specific nature of the relationships between the variables on which the predictions are based. Exploratory Data Analysis (EDA) is used on the one hand to answer questions, test business assumptions, generate hypotheses for further analysis. On the other hand, you can also use it to prepare the data for modelling. The thing that these two probably have in common is a good knowledge of your data to either get the answers that you need or to develop an intuition for interpreting the results of future modelling. There are a lot of ways to reach these goals: you can get a basic description of the data, visualize it, identify patterns in it, identify challenges of using the data, etc. Hence, I continue to gain knowledge while practicing the same and spread intellectual wings in tech-heaven. Reason for doing EDA with sweetviz is, generates a report with all the information easily available, without writing each command separately. For client, if we can do work faster it will be better so it saves lot of time and win-win for all. I obtained grasp on this abstraction. It is a reward for me after practicing it manually. I will always recommend to practice manually first then once we get hang on it then can try with other libraries.

Mostly 60% of the time spent by a data scientist is on Exploratory Data Analysis (EDA).

The reasons to perform EDA is as follows:-

1. It helps us to visualize the data and can identify which feature in the data is affecting the desired output.
2. It helps in finding the missing values in the given dataset.
3. It helps in finding the data types of the columns and complete analysis of the dataset.

To reduce the time to perform EDA there is a powerful library known as **SweetViz**.

Its goal is to help quick analysis of target characteristics, training vs testing data, and other such data characterization tasks. It is an inspiration taken from Pandas-Profiling which I have done in Project22.

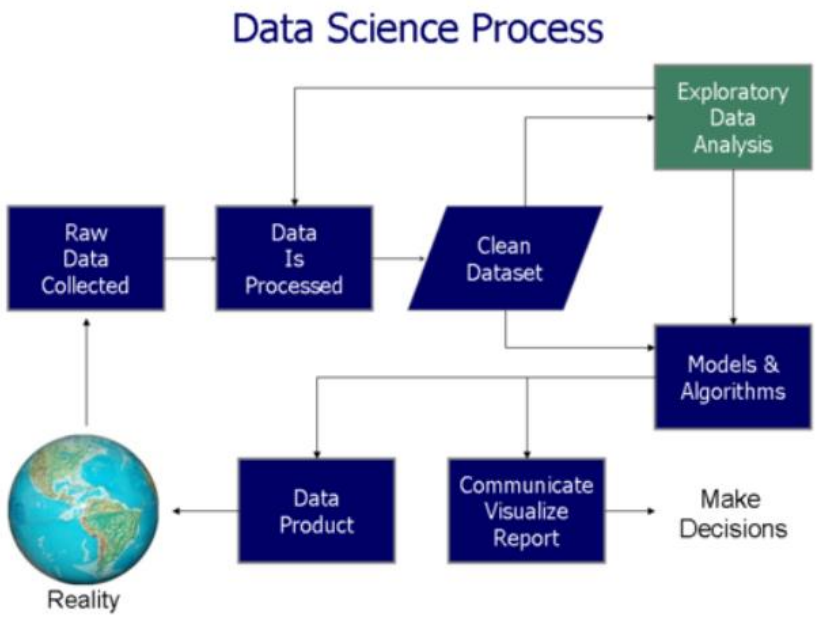
Need for Report: There are 3 main functions for creating reports:

1. analyze(…)
2. compare(…)
3. compare_intra(…)

Technical Aspect

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.

Using Sweetviz, EDA report is ready and contains a lot of information for all the attributes. It's easy to understand and is prepared in just 3 lines of code. Sweetviz can also be used to visualize the comparison of test and train data. Compare() function of Sweetviz is used for comparison of the dataset. It is called automating EDA.



Data Science Process Illustration

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 pandas
pip install sweetviz

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\11)EDA_and_Visualization\sweetviz

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)EDA_With_Visualization_using_sweetviz_in_1line_2.ipynb
This takes train.csv and test.csv file as input dataset.
It creates House_output_report.html and House_output_report2.html files as output in same working folder. Then if you wish you can transfer it by creating it empty to Output_Files Folder as I did.
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]

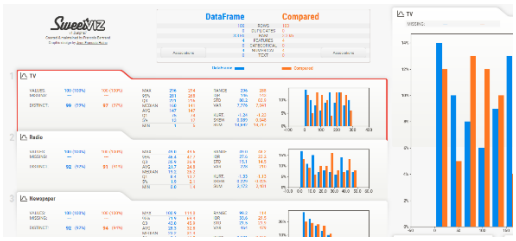
Directory Tree/Structure of Project

Folder: 11)EDA_and_Visualization > sweetviz
1)EDA_With_Visualization_using_sweetviz_in_1line_2.ipynb

To Do/Future Scope

Insights from the Generated Report OR What information does this report give, we can add.

Technologies Used/System Requirements/Tech Stack

Credits

Krish Naik Channel, towardsdatascience.com