**Step By Step Guide To Automated Exploratory Data Analysis Using Sweetviz in Python**



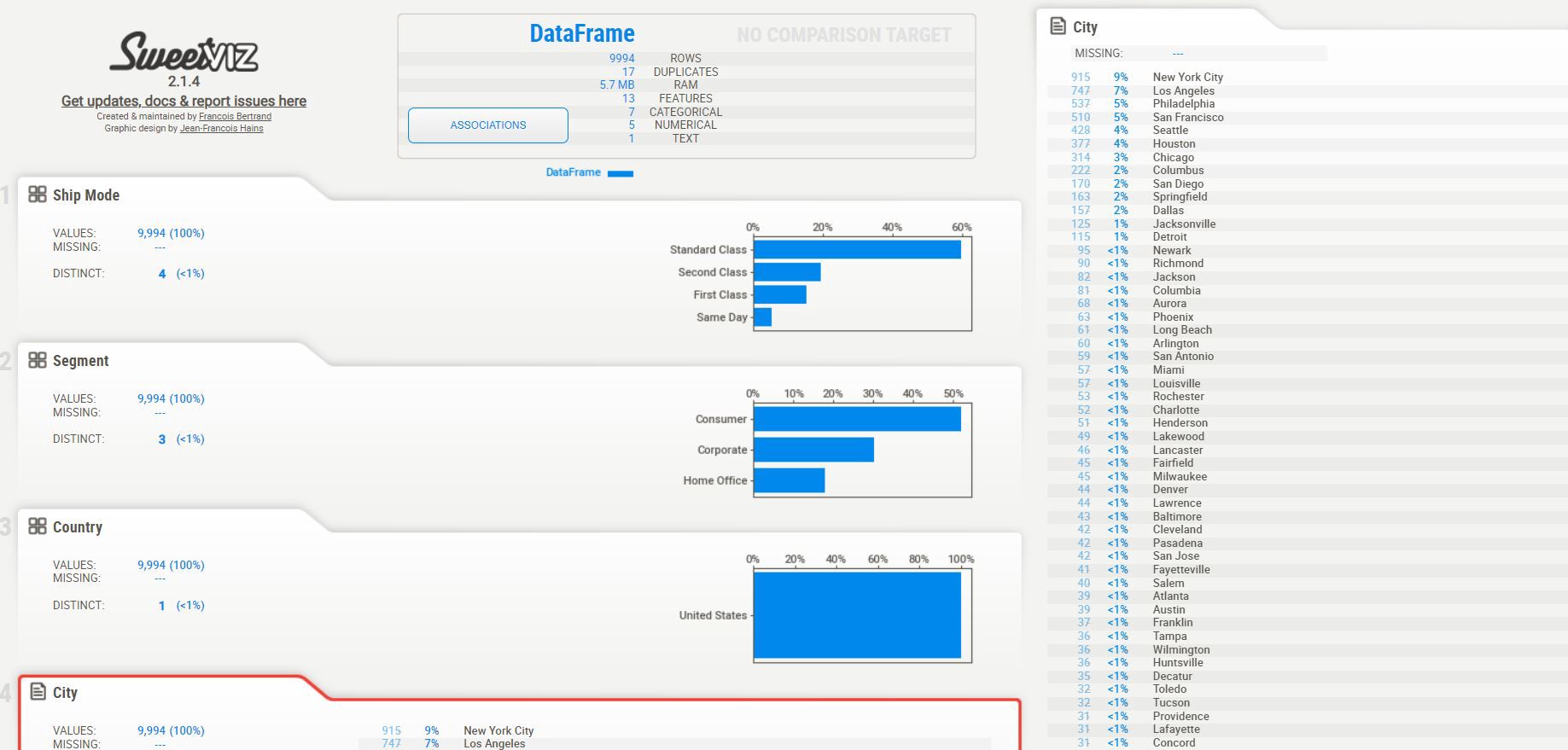
Photo by [Luke Chesser](https://unsplash.com/@lukechesser?utm_source=medium&utm_medium=referral) on [Unsplash](https://unsplash.com?utm_source=medium&utm_medium=referral)

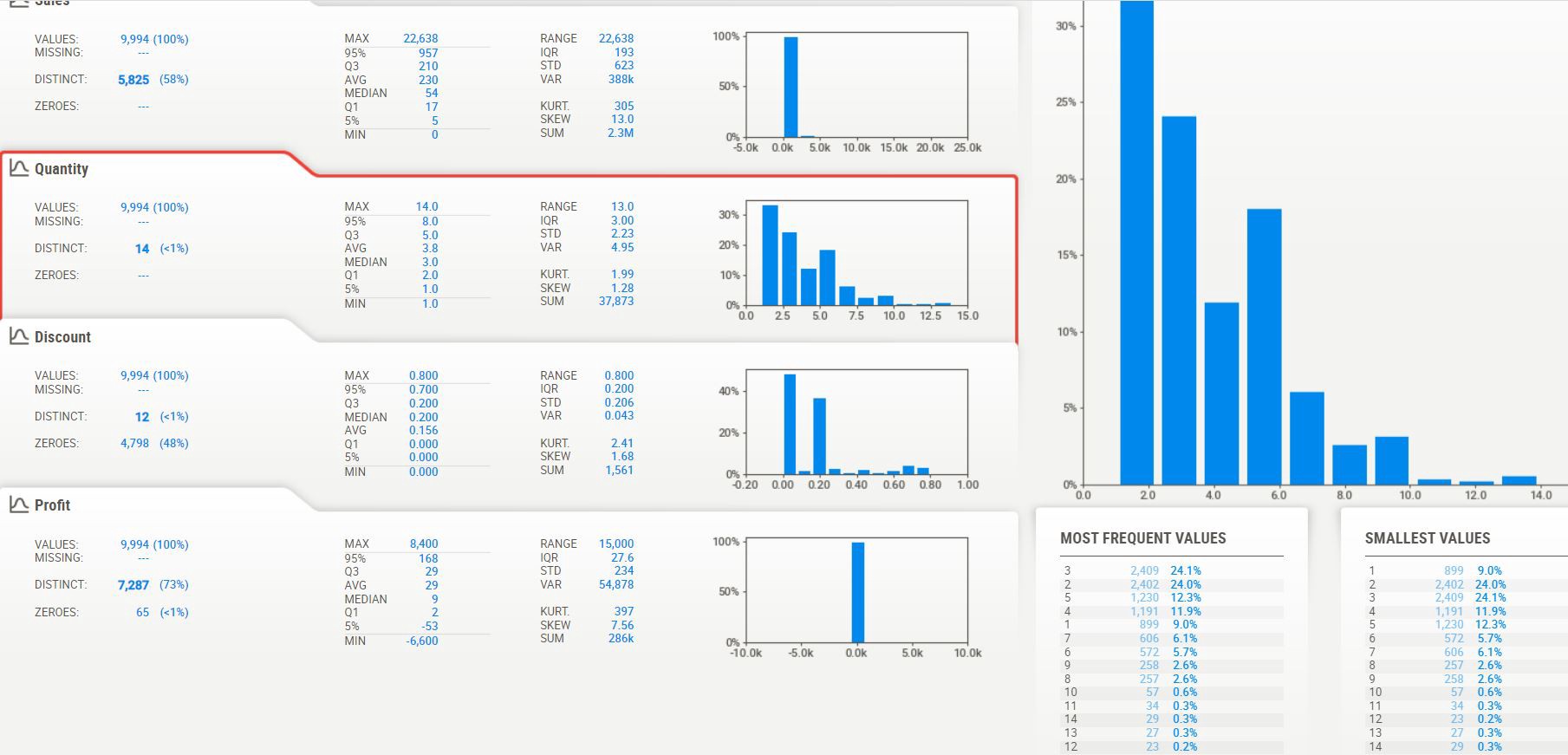
**Table of Contents**

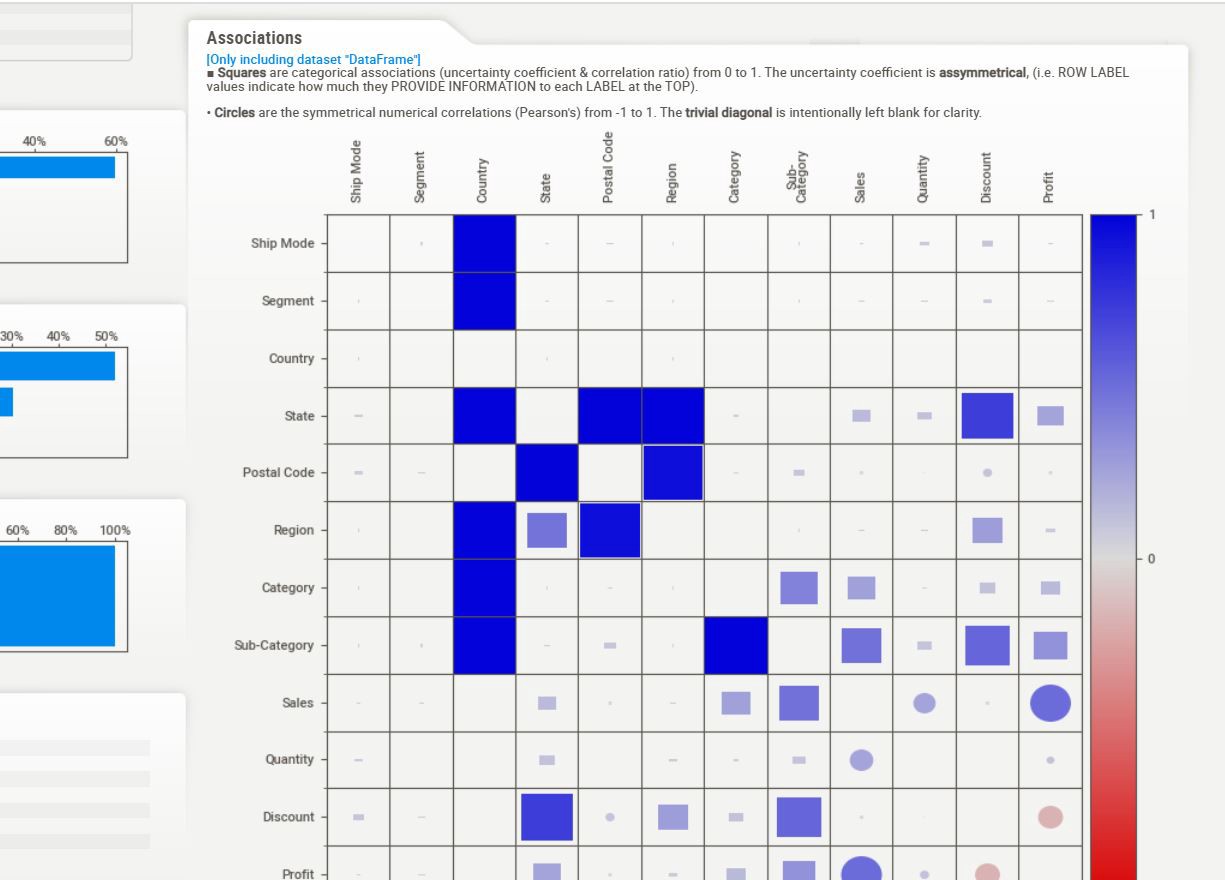
· [Abstract](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#711c)  
· [Casual Introduction to Data Science & Analytics](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#6dc2)  
∘ [Skills of a Data Scientist](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#cbb5)  
∘ [Key lessons from the Industry](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#b93d)  
· [What are the different analytics levels?](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#0c9d)  
· [What is Sweetviz?](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#e165)  
· [Analyzing Data with Sweetviz](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#976c)  
∘ [Step 1: Install Sweetviz and Read Raw Data](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#5c92)  
∘ [Step 2: Create the Report](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#ab82)  
∘ [Step 3: Display the Report](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#e9da)  
∘ [Generating EDA summary by Category](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#2e8b)  
· [Conclusion](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#0b56)  
· [Reference](https://towardsdatascience.com/step-by-step-guide-to-automated-exploratory-data-analysis-using-sweetviz-in-python-1061ed3f91da#14c9)

**Abstract**

With just one line of code, the open-source Python module Sweetviz can create stunning visualizations for exploratory data analysis (EDA). Additionally, it produces a summarized report and can be used to build interactive dashboards. An entirely self-contained HTML application is the output produced.







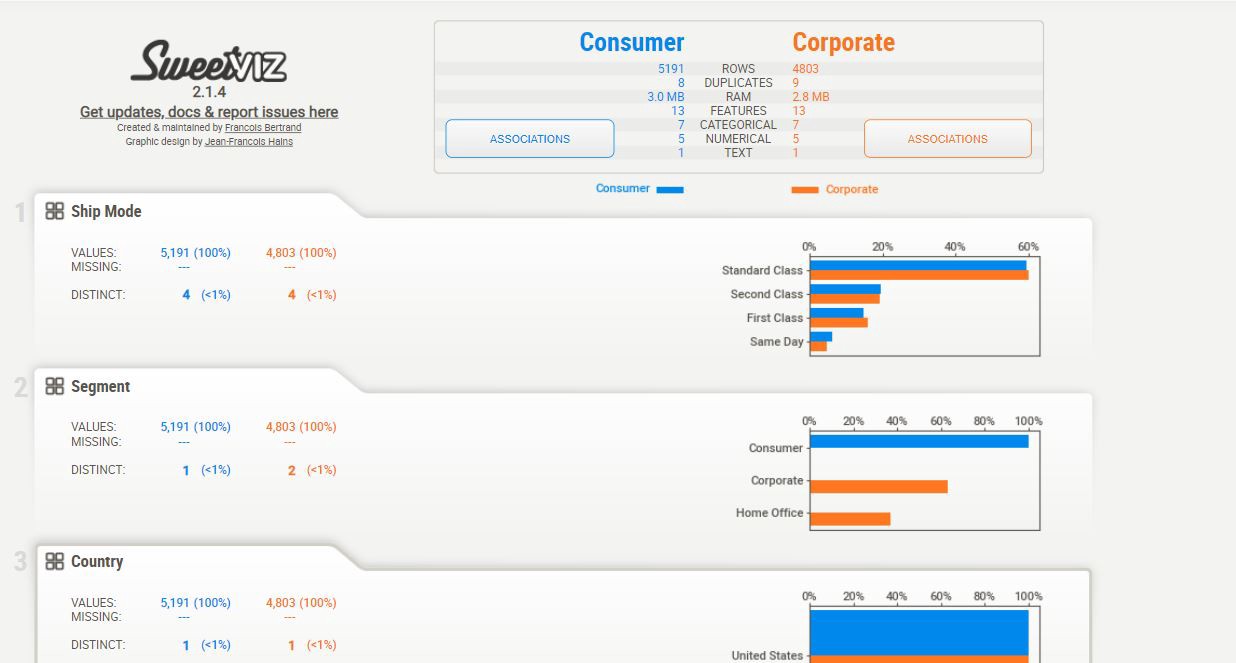


Image — Sample output from Sweetviz. Source — Author.

**Casual Introduction to Data Science & Analytics**

Data science is the technology of handling and extracting value from data. Data science is not just big data. It involves working with data to achieve value and even generate market trends that allow a business to benchmark performances. Data Science is often defined as the process of using statistical and machine learning techniques on big multi-structured data to identify:

* Correlation
* Causation
* Classify and predict events
* Identify patterns and anomalies
* Probabilities

Data science facilitates data-driven decision-making.

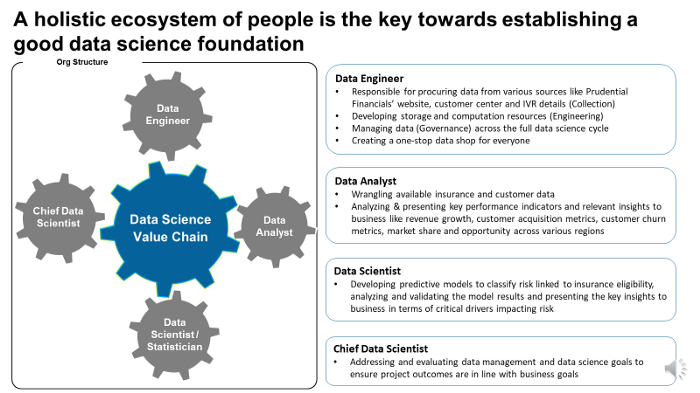


Figure 1. Different roles driving the data science value chain in an Organization. Source: Author.

**Skills of a Data Scientist**

As the data science industry evolves, a data scientist should know mathematics, business, and technology.

* Business — Data Scientists should have domain knowledge of business to understand the problem space and propose a solution that best suits the problem space.
* Mathematics (Mathematics + Statistics + Machine Learning) — Should be able to understand statistics, machine learning models, and algorithms statistics and apply them to structured and unstructured data.
* Technology (Programming) — Should be comfortable working with R, Python, SAS, SQL, Hadoop, etc., and handle backend and frontend programming.

**Key lessons from the Industry**

* Everyone should be curious to work with data (a critical quality).
* Communication skills are underrated.
* Data Collection and wrangling are the biggest challenges.
* A data scientist is better at statistics than a software engineer and better at software than a statistician.
* The data science industry is nascent, and roles need to be better defined; hence one gets to work with different business stakeholders.

As discussed earlier, analyzing data becomes a key component before an algorithm can be developed and used. Let’s understand the basic level of analytics before jumping into how we can optimize data EDA using SweetViz.

**What are the different analytics levels?**

Analytics levels can be classified using a DIPP framework. The definition and examples are listed below:

* **Descriptive** — gain insights from data. E.g., sales by regions and territories for a pharmaceutical company for a specific brand of drug.
* **Inquisitive** — Understand specific trends or findings from data using descriptive statistics. E.g., if sales of the region for a drug are low, why did sales go down
* **Predictive** — making a prediction using statistical and machine learning models. E.g., forecasting sales of a drug using historical data so that goals can be set for sales representatives.
* **Prescriptive** — Recommendation of decisions to business stakeholders using optimization or solution. E.g., to improve sales, prescribers in a region should be targeted by sales representatives.

**What is Sweetviz?**

With just one line of code, the open-source Python module Sweetviz enables the creation of attractive, minutely accurate visuals for exploratory data analysis. Additionally, it produces a report summary and can aid in developing interactive dashboards. An entirely self-contained HTML application is the output produced. The technology quickly creates reports by comparing datasets and showing the target values. With the help of SweetViz, you can quickly characterize target properties, training vs. testing data, and other data types (Dey, 2021).



Image Source — [*https://github.com/fbdesignpro/sweetviz*](https://github.com/fbdesignpro/sweetviz), License Free to Use and Distribute— <https://github.com/fbdesignpro/sweetviz/blob/master/LICENSE>

There are the following features in Sweetviz (Bertrand, 2022):

* **Creating Target analysis (applicable when training models are required):** How boolean or numerical goal values connect to other characteristics.
* **Comparing and contrasting several dataset types:** Distinct datasets (such as those used for training versus test data or even training versus validation) or intra-set features (e.g., male versus female, category 1 versus category 2).
* **Identifying relationships of diverse types:** To provide comprehensive information for all data kinds, Sweetviz can easily find numerical associations like Pearson’s correlation, categorical associations like uncertainty coefficient, and categorical-numerical data types.
* **Inference for Data Type:** With an optional manual override, SweetViz identifies numerical, category, and text features automatically.
* **Learn about summary data**, including Type, unique values, missing values, duplicate rows, most common values, and numerical analysis, including min/max/range, quartiles, mean, mode, standard deviation, total, median absolute deviation, coefficient of variation, kurtosis, and skewness.

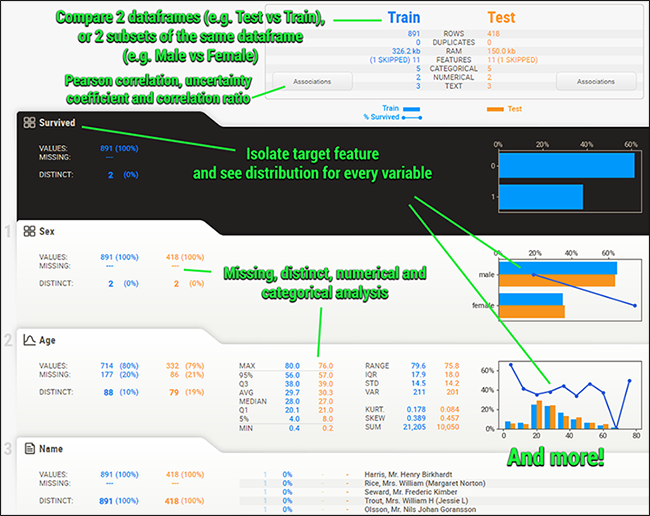


Figure 2. Sample output from SweetViz. I*mage source:* [*https://github.com/fbdesignpro/sweetviz*](https://github.com/fbdesignpro/sweetviz)*.* License Free to Use and Distribute — <https://github.com/fbdesignpro/sweetviz/blob/master/LICENSE>

**Analyzing Data with Sweetviz**

The data used for this analysis can be found [here](https://www.kaggle.com/datasets/bravehart101/sample-supermarket-dataset?resource=download) (license — public domain hence free to use and distribute). This is an example of a superstore dataset, a sort of simulation where you conduct in-depth data analysis to provide ideas on how the business can boost earnings while reducing losses.

To use Sweetviz we need to install it. Note this is an external package and doesn’t come with Anaconda Navigator installation.

**Step 1: Install Sweetviz and Read Raw Data**

! pip install --user sweetviz  
  
import pandas as pd  
  
data = pd.read\_csv('SampleSuperstore.csv')  
  
data.head()

The first five records from the data are illustrated below. Since the data pertains to a store, we have information about the different shipments, including Shipping Mode (Ship Mode), Customer Segment, Country, City, State, Postal Code, and Region to which the product is shipped. Category and Subcategory of item/product shipping are also available along with Sales KPIs (Key Performance Indicators), including Sales, Quantity, Discount, and Profit.



Figure 3. Sample output — First 5 rows of the data. Source — Author.

**Step 2: Create the Report**

There are 3 main functions for creating reports:

* **analyze ()** — Allows the user to perform fundamental Exploratory Data Analysis (EDA) on the data. This is synonymous with creating descriptive statistics and produces summaries, numerical distribution, 5-point statistics (mean, median, mode, standard deviation, etc.), and correlations on all attributes from the data. Note it won’t produce any other multivariate analysis except correlation plots.
* **compare()** — Used for comparing Training and Test data and is particularly useful when building Machine Learning Models.
* **compare\_intra()** — This function allows users to compare two categorical attributes in a data. E.g., from the above data, if we want to create a comparison summary for consumer vs. other customer segments across all other attributes, we can achieve the same using compare\_intra. Note that only binary comparisons are performed using this function. It currently doesn’t support any multi-category comparisons.

import sweetviz as sv  
analyze\_report = sv.analyze(data, pairwise\_analysis="on")  
analyze\_report.show\_html('Data EDA.html', open\_browser=True)  
analyze\_report.show\_notebook(w=None,   
 h=None,   
 scale=0.8,  
 layout='vertical',  
 filepath=None)

There are two components to running a SweetViz. One, we generate the report using analyze(), and two, we display the report either within the notebook or using an HTML file using show\_html() or show\_notebook().

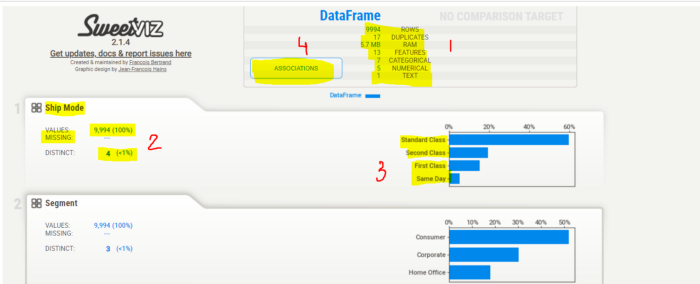


Figure 4. Sample Output of analyze(). Source — Author.

Figure 4 illustrates a quick snap shot of the sample out. Few key areas to highlight here:

1. **Section 1** — This section provides information about the data, including the number of rows, duplicates, if any, total features or columns, and the data type of each column.
2. **Sections 2 & 3** — This is where we get a univariate analysis of each variable. We get the number of records, missing value count, number of distinct categories, and a count plot of the feature.
3. **Section 4** — This provides a correlation analysis of all variables. You need to click on the association tab to get the correlation matrix.

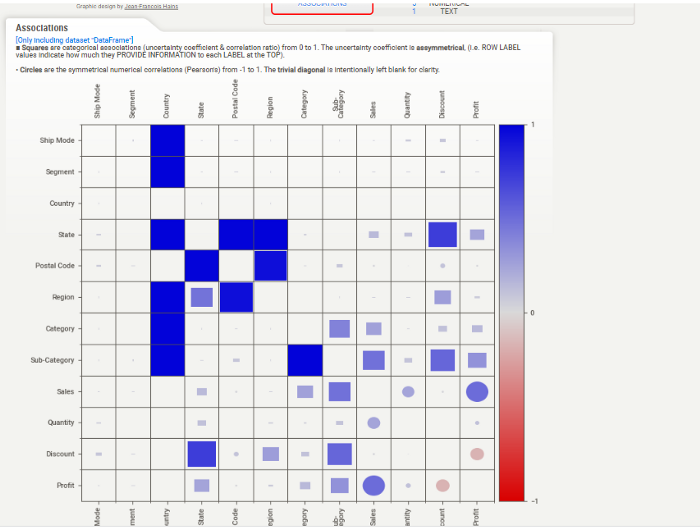


Figure 5. Illustrates the correlation matrix of all attributes from the data. The time required to complete correlations and other relationships can be quadratic. Until a data collection contains “association auto threshold” features, the default value (“auto”) will operate without notice. Beyond that point, you must explicitly pass the pairwise analysis=” on” (or “off”) argument because processing that many characteristics would be time-consuming. Additionally covered by this parameter is the creation of association graphs (based on [Drazen Zaric’s theory](https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec)). Source — Author.

Similarly, for numerical variables, Sweeviz produces basic statistics of the data (section 1 below), distribution plots with the flexibility to decide the intervals or buckets, and correlation analysis with numerical and categorical values. Note — The distribution plot is a histogram with the Y axis representing the proportion of data for each numerical bucket.

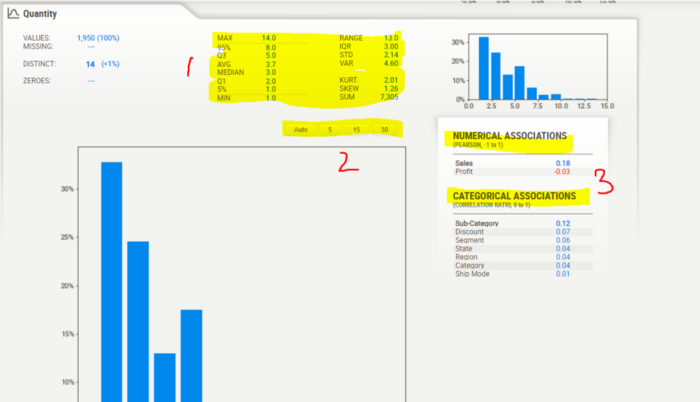


Figure 6. Illustrates the output of numerical variables. Source — Author.

**Step 3: Display the Report**

Although we have covered this previously, there are two ways in which the Sweetviz report can be displayed. 1. show\_html() 2. show\_notebook()

An HTML report will be generated by show\_html() and saved to the specified file directory. Options are available for:

1. **layout:** either a “widescreen” or a “vertical” arrangement. When the mouse is over each feature, the widescreen layout shows details on the right side of the screen. The updated vertical layout (as of version 2.0) is more condensed horizontally and allows clicking to expand each detailed area.
2. **scale** the entire report with a floating-point number (scale= 0.8 or None). Adapt reports to any output, and this is helpful.
3. **open\_browser**: Allows a web browser to open automatically and display the report. You can disable it here if this is something that you don’t want to happen sometimes (or if it interferes with particular IDEs).

As of version 2.0, the new function show\_notebook() will embed an IFRAME element showing the report inside a notebook (e.g., Jupyter, Google Colab, etc.).

1. **w (width):** Sets the width of the report’s output window (the complete report may not fit; for the report itself, use layout and/or scale). w can be expressed as a string percentage (w=”100%”) or as a pixel count (w=900).
2. **h (height):** Sets the height of the report’s output window with the h (height) command. The window can be stretched to be as tall as all the features (h=”Full”) or as many pixels (h=700).
3. **scale/layout:** similar functionality as show\_html()

**Generating EDA summary by Category**

The compare\_intra() function allows users to compare two categorical attributes in a data. The output remains broadly in line with the analyze function, except that we have the summaries by categories. Please note that this function can only produce binary comparisons. Hence to create this analysis, we have considered “Consumer” as category one and the rest as Others (Corporate/Head Office). Still, we have defined or tagged this category in the report as “Corporate.” Quick example below.

my\_report = sv.compare\_intra(data, data["Segment"] == "Consumer", ["Consumer", "Corporate"])  
  
my\_report.show\_notebook(w=None,   
 h=None,   
 scale=0.8,  
 layout='vertical',  
 filepath=None)

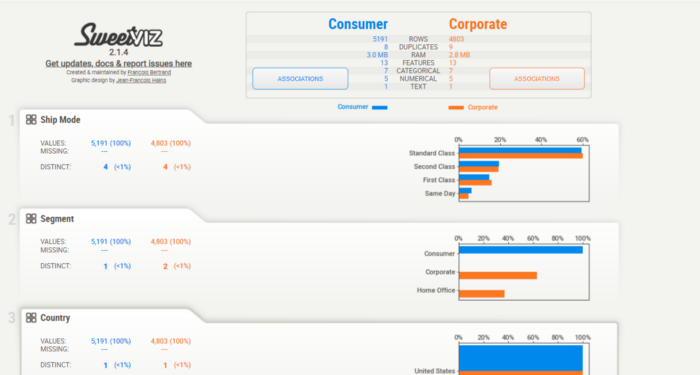


Figure 7. Illustrates the output of compare\_intra(). Source — Author.

**Conclusion**

Python currently supports multiple other packages that allow quick EDA on the data, including pandas profiling. Although Sweetviz doesn’t necessarily produce a comprehensive output of the data, it can be used to understand the data structure and underlying trends purely from the perspective of univariate analysis. Especially the summary statistics, in general, allows user to strategize on data cleaning processes which becomes a key to model accuracies later on. Complex summaries involving multi-variate analysis still need to be performed manually or using business intelligence tools.

Some useful links for your reference.

**[Introduction to Data Processing using Descriptive Statistics and Statistical Charts in Python](https://towardsdatascience.com/introduction-to-data-processing-using-descriptive-statistics-and-statistical-charts-in-python-9857a60c481b" \t "_blank)**

**[A complete hands-on guide to test Data Assumptions (MCAR, MAR, MNAR, Central Tendency, Skewness, and Outliers) in…](https://towardsdatascience.com/introduction-to-data-processing-using-descriptive-statistics-and-statistical-charts-in-python-9857a60c481b" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/introduction-to-data-processing-using-descriptive-statistics-and-statistical-charts-in-python-9857a60c481b" \t "_blank)

**[Statistics for Data Science — A beginners guide to Descriptive Statistics in Python](https://towardsdatascience.com/statistics-for-data-science-a-beginners-guide-to-descriptive-statistics-in-python-9e0daa30809a" \t "_blank)**

**[A working example of Central Tendency, Dispersion, Standard Deviation, and Correlation using Excel, Python, and…](https://towardsdatascience.com/statistics-for-data-science-a-beginners-guide-to-descriptive-statistics-in-python-9e0daa30809a" \t "_blank)**

[towardsdatascience.com](https://towardsdatascience.com/statistics-for-data-science-a-beginners-guide-to-descriptive-statistics-in-python-9e0daa30809a" \t "_blank)

**Reference**

1. Bertrand, F. (2022, June 3). Features. Retrieved from GitHub website: <https://github.com/fbdesignpro/sweetviz>
2. Dey, V. (2021, July 24). Step By Step Guide To Data Analysis Using SweetViz. Retrieved December 3, 2022, from Analytics India Magazine website: <https://analyticsindiamag.com/step-by-step-guide-to-data-analysis-using-sweetviz/#:~:text=What%20is%20SweetViz%3F>
3. Sample Superstore Dataset. (n.d.). Retrieved December 3, 2022, from [www.kaggle.com](http://www.kaggle.com) website: <https://www.kaggle.com/datasets/bravehart101/sample-supermarket-dataset?resource=download>

*About the Author: Advanced analytics professional and management consultant helping companies find solutions for diverse problems through a mix of business, technology, and math on organizational data. A Data Science enthusiast, here to share, learn and contribute; You can connect with me on* [*Linked*](https://www.linkedin.com/in/angel-das-9532bb12a/) *and* [*Twitter*](https://twitter.com/dasangel07_andy)*;*