Published on July 24, 2021

In Mystery Vault (https://analyticsindiamag.com/category/mystery-vault/)

Step By Step Guide To Data Analysis Using SweetViz

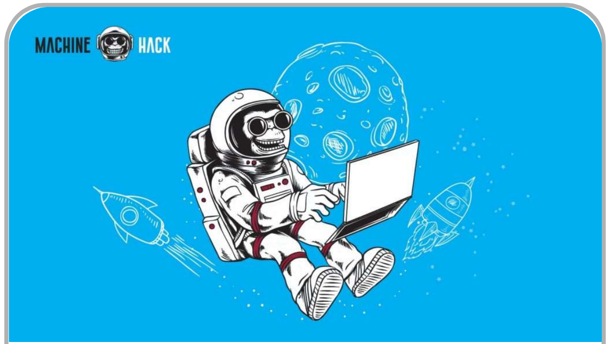
Sweetviz is an open-source Python library that helps generate beautiful, highly detailed visualizations to Exploratory Data Analysis with a single line of code. It also generates a summarised report and can help create interactive dashboards as well. The output generated is a fully self-contained HTML application.

SweetVIZ
Product Trends by Month
Cosmetics Sales Volume
Sales Volume /million pcs
January
February
March
April
May
June
Shampoo
Perfume
Lipstick

Data is an entity that is omnipresent in every aspect of life. Whether it be in spreadsheets, your sales pipeline, social media platforms and more, it can be termed a collection of facts, numbers, words, measurements, observations that have been translated into a form that computers and systems can process. Every piece of information collected or recorded is treated as data that is stored and recorded by systems or humans, individually or combined. Data can be used to measure and understand a wide range of internal and external activities in every ecosystem. With the growth of the world wide web and smartphones becoming technologically superior over the past decade, it has led to a surge in digital data creation. Data churned these days now also include text, audio and video information, and web log and web activity records, which can be categorized as unstructured data. The analysis of such data to gain knowledge of what is happening around and to generate insights that might be useful for decision making is known as Data Analysis. The sole purpose of Data Analysis is to extract useful information from data which can become the pivotal factor for

decision making based upon the data analysis. It is being said that "data is the new oil", the more data one has, the more it will aid the analysis, which might aid in discovering groundbreaking facts.

If a business is not growing, one can look back and acknowledge the mistakes and make a plan again without repeating those mistakes. Viceversa, if the business is growing, then one can further plan on making the business grow even more. All you need to do is analyze your business data and business process. Data Analysis includes the following processes: cleaning, analyzing, interpreting, and visualizing. Data analysis tools make it even easier for users to process and manipulate data, analyze the relationships and correlations between the datasets, and it also helps to identify trends for interpretation. Analysis techniques such as Exploratory analysis or In-depth analysis of data become useful tools for data interpretation. Identifying the data you need for analysis can be another challenge with the vast (https://www.lifewire.com/how-big-is-the-web-4065573) amount of data available. One may choose a data source that isn't reliable or might miss crucial data sources that should be part of the research being conducted. Reliable and complete data is necessary for accurate data analysis.



Descriptive data analysis (https://analyticsindiamag.com/exploring-simfin-api-using-exploratory-data-analysis/) tells us the "What happened?" when analyzing the data. Inferential analysis generalizes or hypothesizes about the question of "What happened?" by comparing statistics from groups within an entire population present in the dataset. Other types of analysis, such as Diagnostic analysis, also known as root cause analysis, aim to answer "Why did the following happen?". Predictive analysis (https://analyticsindiamag.com/comprehensive-guide-to-time-series-analysis-using-arima/) uses data to postulate about future events. It is concerned with "What is likely to happen. The prescriptive analysis is the most advanced form of analysis, as it combines all of your data and analytics, then outputs a model prescription: What action to take.

Using Exploratory Data Analysis, also known as EDA, data scientists can analyze and investigate data sets and summarise their main characteristics, often applying many data visualization methods. One cannot draw reliable conclusions from a massive quantity of data by just glancing over it; data must be taken care of and looked carefully through an analytical perspective. Exploratory Data Analysis helps us determine how to manipulate data sources best to get the answers needed, making it easier for data scientists to discover patterns, spot anomalies, if any, test a hypothesis, or check assumptions. EDA is

aim
GET YOUR ORGANIZATION CERTIFIED
Best Firm for Data Scientists Certification
→

primarily used to see what the present data can reveal beyond the formal hypothesis testing task. Hence, it provides a better understanding of the variables in the dataset being worked upon and describes theirrelationships. It can also help determine if the statistical techniques one is considering for data analysis are appropriate or not. Exploratory Data Analysis is a highly crucial step before you jump to machine learning or modeling of your data. It provides a clear context needed to develop an appropriate model that helps interpret the results correctly.

Exploratory Data Analysis provides a lot of critical information that is easily neglected at times. Therefore, it helps the analysis in the long run, right from framing questions to displaying results. The term was coined by an American mathematician named John Tukey in the 1970s. EDA techniques continue to be a widely used method in the data discovery process today. When data analysis is done properly, several things seem to fall into place, even from a business perspective within an organization. First, data scientists will always know if they have produced results within the required business context; stakeholders will therefore be assured if they are asking the right questions and even discover interesting trends that even they did not know existed.

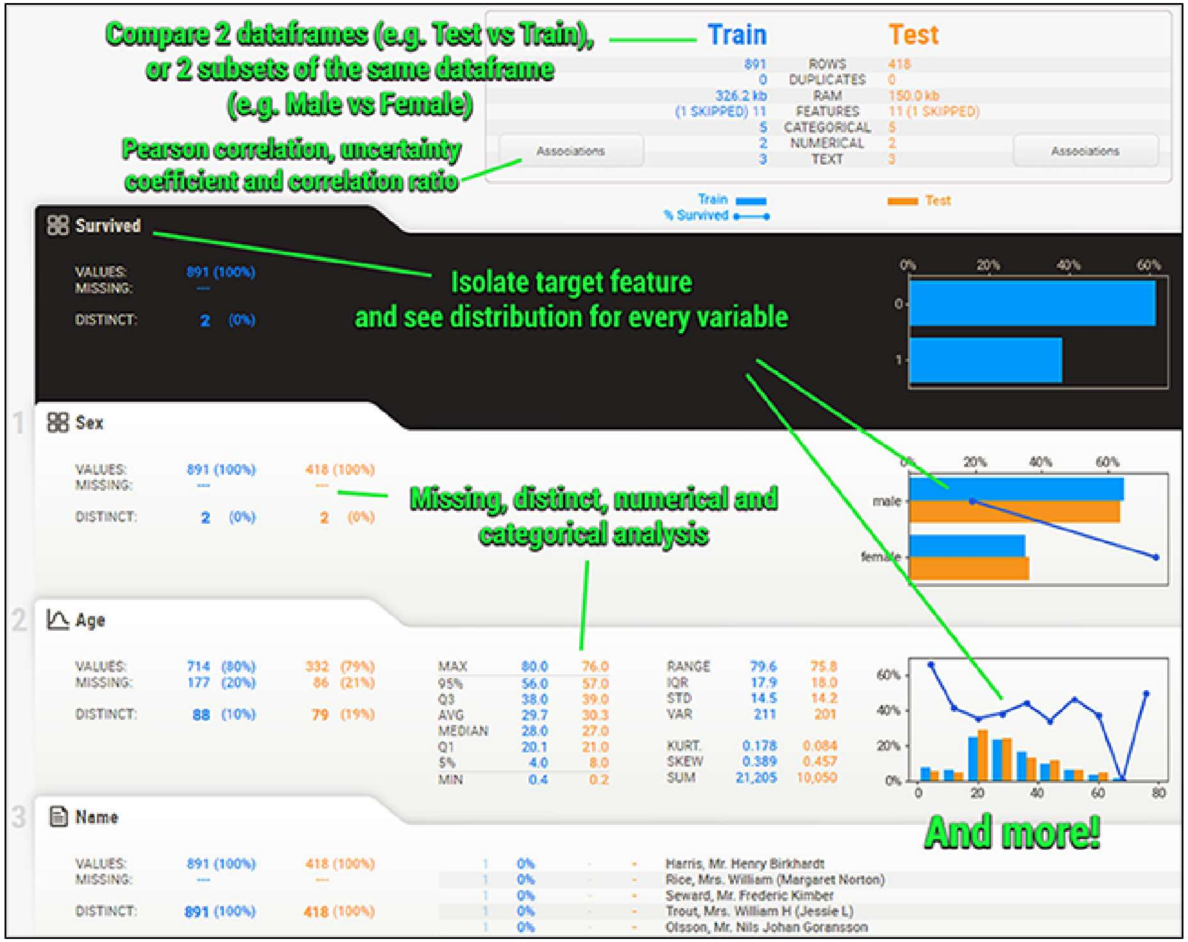
Put simply, exploratory data analysis ensures that business executives are getting the results they are looking for and heading in the right direction just by using data analytics. In addition, the EDA can be easily incorporated into BI software and for Dashboarding purposes, which is crucial when a layman wants to know what's going on in a couple of minutes. Creating a Dashboard (https://analyticsindiamag.com/streamlit-vs-plotlydash-comparison-with-python-examples/)also helps get a breakdown of matters through beautiful visualizations and understanding what further decisions need to be taken.

What is SweetViz?

Sweetviz is an open-source Python library that helps generate beautiful, highly detailed visualizations to Exploratory Data Analysis with a single line of code. It also generates a summarised report and can help create interactive dashboards as well. The output generated is a fully self-contained HTML application (https://analyticsindiamag.com/creating-a-language-translator-app-using-gradio/). The system builds reports around quickly, visualizing the target values and even comparing datasets. SweetViz aims to generate a quick analysis of target characteristics, training vs testing data, and other such data characterization tasks.

Sweetviz comprises of the following features :

- Create Target analysis: How target values boolean or numerical relate to other features
- Visualize and compare different types of datasets: Distinct datasets (e.g. training vs test data) or Intra-set characteristics (e.g. male versus female)
- To Find Mixed-type associations: Sweetviz can discover numerical associations such as Pearson's correlation, categorical associations like uncertainty coefficient, and categorical-numerical data types seamlessly to help provide maximum information for all data types.
- Type inference: SweetViz automatically detects numerical, categorical and text features, with optional manual overrides
- Discover Summary information such as Type, unique values, missing values, duplicate rows, most frequent values and Numerical analysis: min/max/range, quartiles, mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness.



Compare 2 dataframes (e.g. Test vs Train),
or 2 subsets of the same dataframe
(e.g. Male vs Female)
Pearson correlation, uncertainty
coefficient and correlation ratio
Isolate target feature
and see distribution for every variable
Missing, distinct, numerical and
categorical analysis
And more!
Associations
Associations
Train
% Survived
Test
0%
20%
40%
60%
0
1
male
female
0
20
40
60
80

Train		Test
891	ROWS	418
0	DUPLICATES	0
326.2 kb	RAM	150.0 kb
(1 SKIPPED) 11	FEATURES	11 (1 SKIPPED)
5	CATEGORICAL	5
2	NUMERICAL	2
3	TEXT	3

Survived

VALUES:	891 (100%)
MISSING:	---
DISTINCT:	2 (0%)

1 Sex

VALUES:	891 (100%)	418 (100%)
MISSING:	---	---
DISTINCT:	2 (0%)	2 (0%)

2 Age

VALUES:	714 (80%)	332 (79%)
MISSING:	177 (20%)	86 (21%)
DISTINCT:	88 (10%)	79 (19%)

MAX	80.0	76.0
95%	56.0	57.0
Q3	38.0	39.0
AVG	29.7	30.3
MEDIAN	28.0	27.0
Q1	20.1	21.0
5%	4.0	8.0
MIN	0.4	0.2

RANGE	79.6	75.8
IQR	17.9	18.0
STD	14.5	14.2
VAR	211	201
KURT.	0.178	0.084
SKEW	0.389	0.457
SUM	21,205	10,050

3 Name

VALUES:	891 (100%)	418 (100%)
MISSING:	---	---
DISTINCT:	891 (100%)	418 (100%)

1	0%	-	-	Harris, Mr. Henry Birkhardt
1	0%	-	-	Rice, Mrs. William (Margaret Norton)
1	0%	-	-	Seward, Mr. Frederic Kimber
1	0%	-	-	Trout, Mrs. William H (Jessie L)
1	0%	-	-	Olsson, Mr. Nils Johan Goransson

image source: https://github.com/fbdesignpro/sweetviz (https://github.com/fbdesignpro/sweetviz)

Getting Started

In this article, we will try to create an In-Depth Exploratory Data Analysis using SweetViz. We will try to create a dashboard around our data using the SweetViz library and discover some of the features of

SweetViz. The following implementation is inspired by a video tutorial, whose link can be found here. (https://www.youtube.com/watch?v=UR_OK8vBpeY)

Installing The SweetViz Library

First, we will start by installing the SweetViz Library; for this, you can run the following code,

```
#installing the library
! pip install sweetviz
```

Setting up Dependencies

Then we will install the required dependencies and also load our dataset. Here we are using the penguins dataset to analyse, containing all the necessary details for penguins such as their species, island found, body mass, sex and more. We will try to create a Dashboard around the following dataset. You can download the following dataset using the link here. (https://github.com/dataprofessor/data/blob/master/penguins_cleaned.csv)

```
#importing dependencies
import pandas as pd
#loading the dataset
penguins =
pd.read_csv('https://raw.githubusercontent.com/dataprofesso
r/data/master/penguins_cleaned.csv')
```

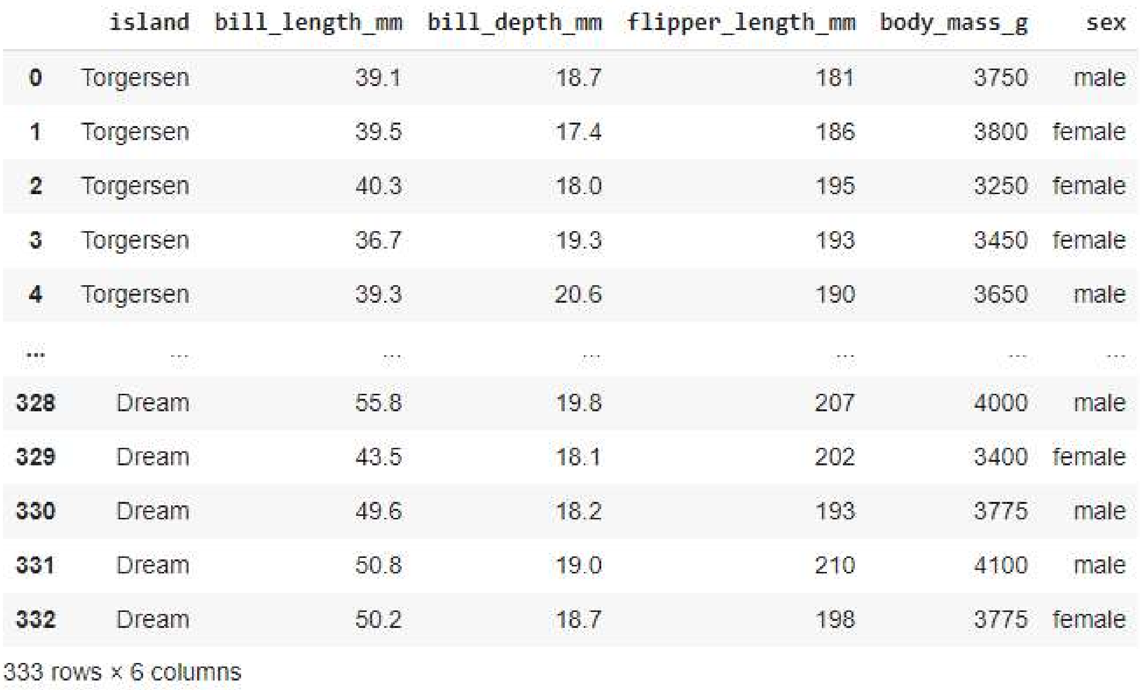
Here we will be trying to discover facts around the species of penguins from our dataset. So we will be splitting the dataset into X and Y, where Y will contain all the information related to the penguin species and X will contain the other information.

```
# Separating X and y
X = penguins.drop('species', axis=1)
y = penguins['species']
```

Taking a look at the data present in X,

```
#displaying data from X
X
```

Output :



	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Torgersen	39.1	18.7	181	3750	male
1	Torgersen	39.5	17.4	186	3800	female
2	Torgersen	40.3	18.0	195	3250	female
3	Torgersen	36.7	19.3	193	3450	female
4	Torgersen	39.3	20.6	190	3650	male
...	...	...	...	...	...	...
328	Dream	55.8	19.8	207	4000	male
329	Dream	43.5	18.1	202	3400	female
330	Dream	49.6	18.2	193	3775	male
331	Dream	50.8	19.0	210	4100	male
332	Dream	50.2	18.7	198	3775	female

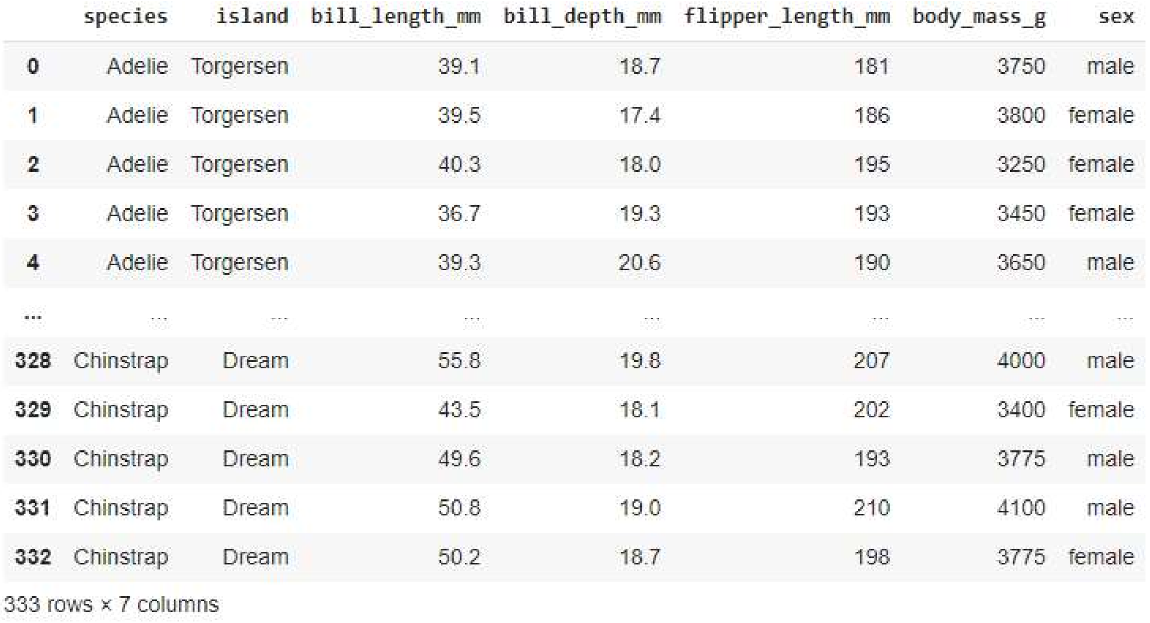
333 rows × 6 columns

As we can see it comprises both quantitative and qualitative data.

Performing Analysis using Sweetviz

Looking at the entire data later performs EDA on the data split with species to analyze the difference.

```
#data before split
penguins
```



	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181	3750	male
1	Adelie	Torgersen	39.5	17.4	186	3800	female
2	Adelie	Torgersen	40.3	18.0	195	3250	female
3	Adelie	Torgersen	36.7	19.3	193	3450	female
4	Adelie	Torgersen	39.3	20.6	190	3650	male
...	...	...	...	...	...	...	...
328	Chinstrap	Dream	55.8	19.8	207	4000	male
329	Chinstrap	Dream	43.5	18.1	202	3400	female
330	Chinstrap	Dream	49.6	18.2	193	3775	male
331	Chinstrap	Dream	50.8	19.0	210	4100	male
332	Chinstrap	Dream	50.2	18.7	198	3775	female

333 rows × 7 columns

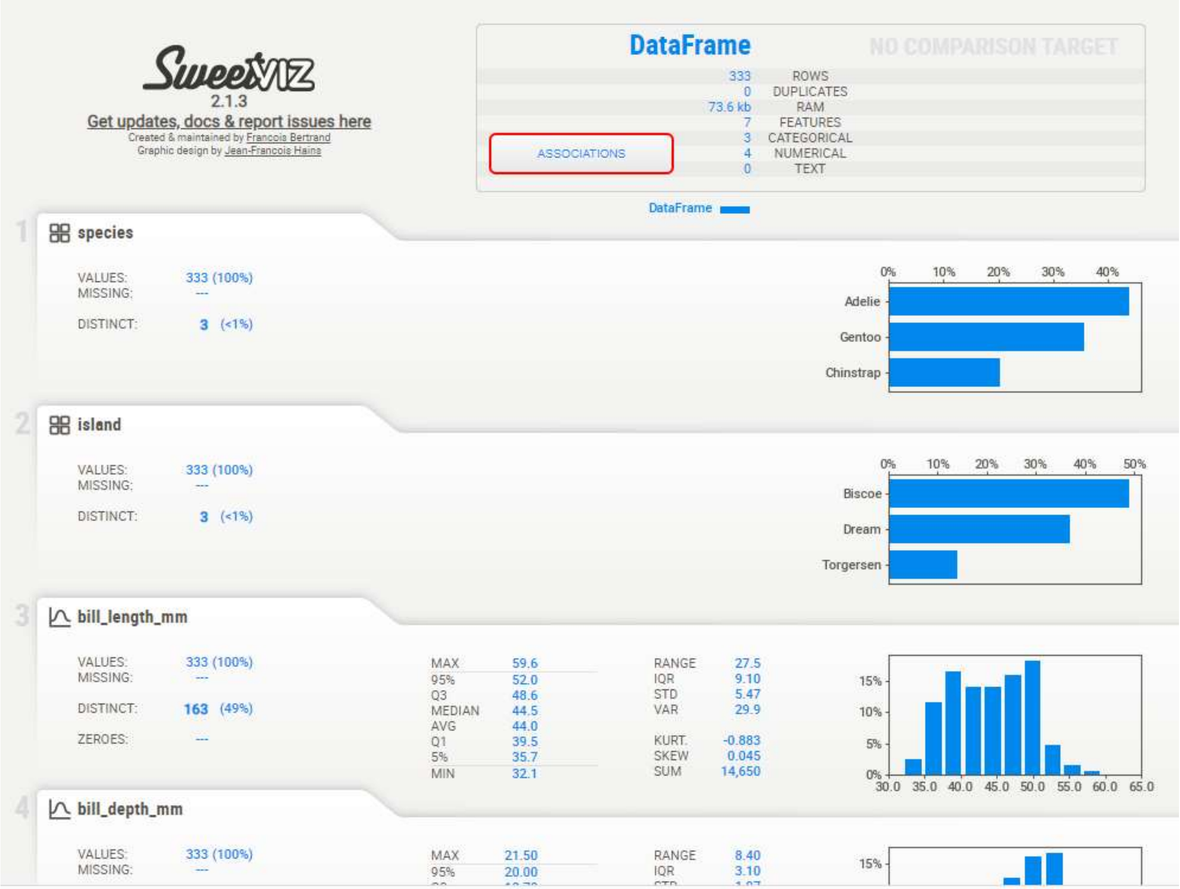
Creating a quick and interactive EDA dashboard on the data,

```
#creating a EDA report
import sweetviz as sv
analyze_report = sv.analyze(penguins)
analyze_report.show_html('analyze.html',
open_browser=False)
```

Displaying Results,

```
import IPython
IPython.display.HTML('analyze.html')
```

This will give us the following output,



SweetVIZ 2.1.3

Get updates, docs & report issues here

Created & maintained by Francois Bertrand

Graphic design by Jean-Francois Hains

DataFrame NO COMPARISON TARGET

333	ROWS
0	DUPLICATES
73.6 kb	RAM
7	FEATURES
3	CATEGORICAL
4	NUMERICAL
0	TEXT

ASSOCIATIONS

DataFrame

1 species

VALUES: 333 (100%)
MISSING: ---

DISTINCT: 3 (<1%)

2 island

VALUES: 333 (100%)
MISSING: ---

DISTINCT: 3 (<1%)

3 bill_length_mm

VALUES: 333 (100%)
MISSING: ---

DISTINCT: 163 (49%)

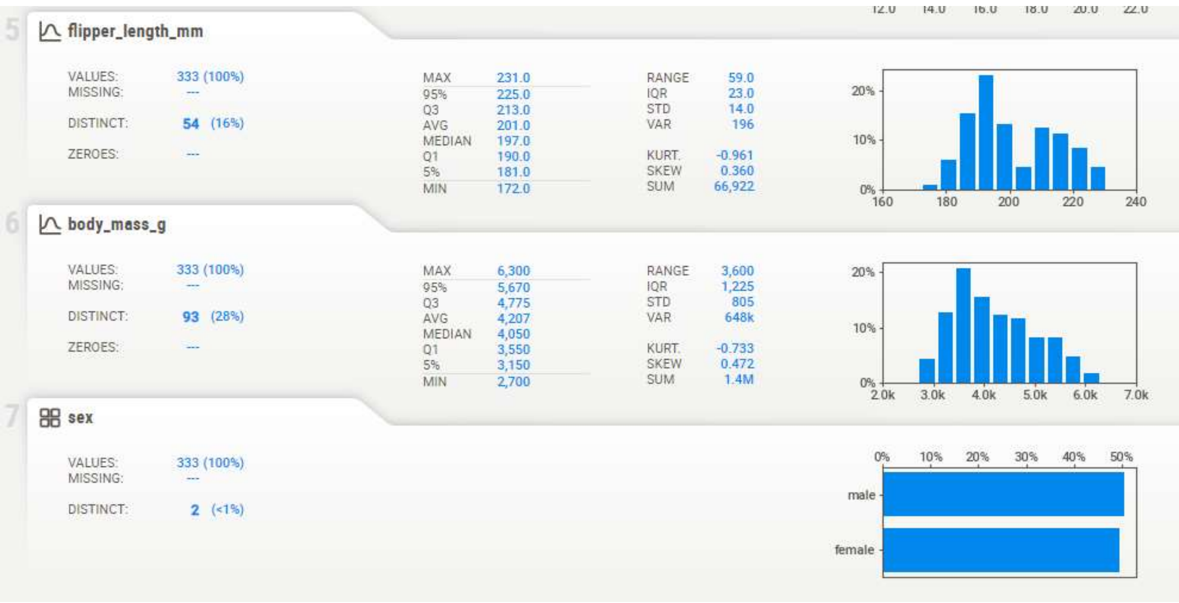
ZEROES: ---

MAX	59.6	RANGE	27.5
95%	52.0	IQR	9.10
Q3	48.6	STD	5.47
MEDIAN	44.5	VAR	29.9
AVG	44.0		
Q1	39.5	KURT.	-0.883
5%	35.7	SKEW	0.045
MIN	32.1	SUM	14,650

4 bill_depth_mm

VALUES: 333 (100%)
MISSING: ---

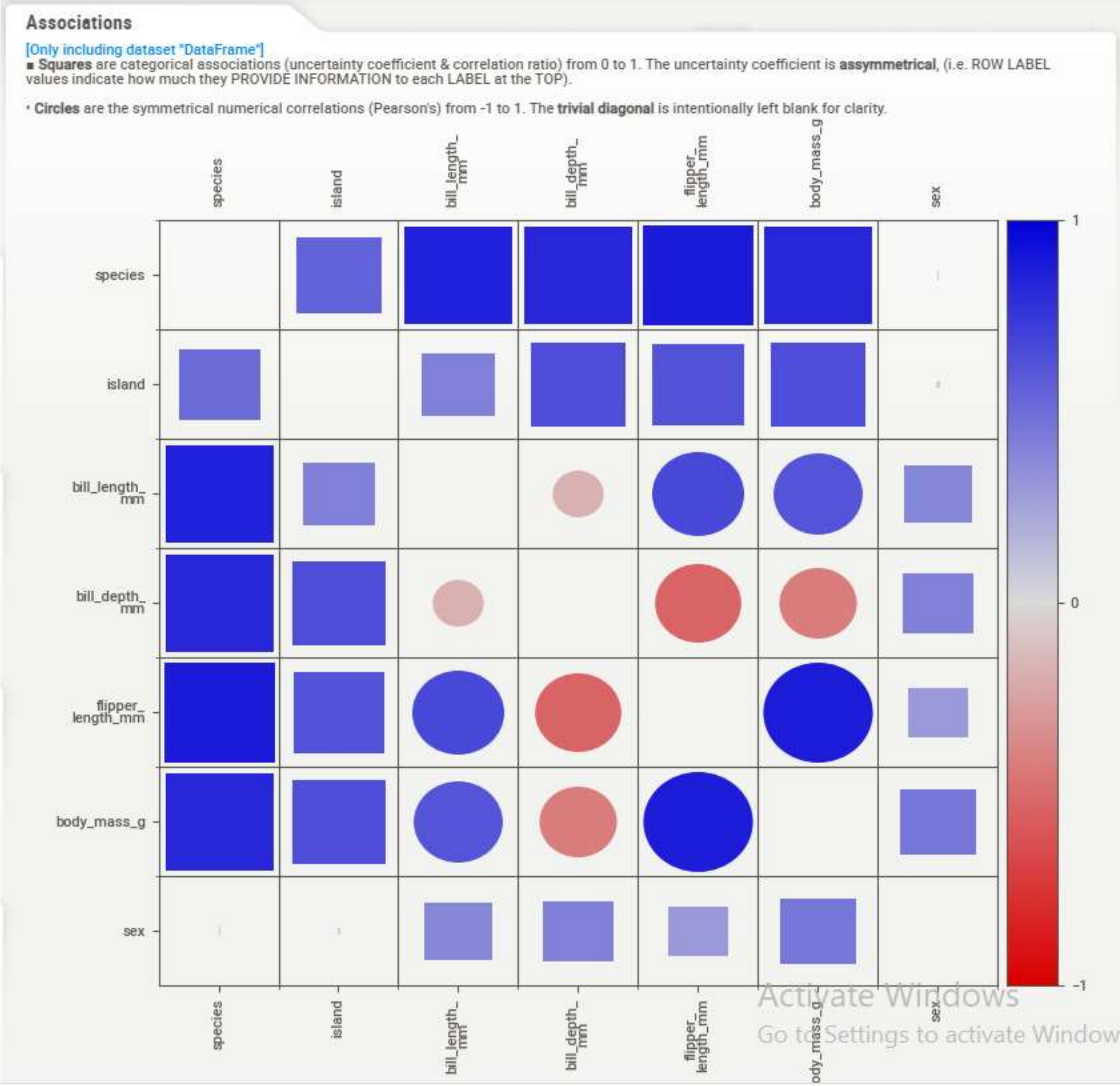
MAX	21.50	RANGE	8.40
95%	20.00	IQR	3.10



```
5 flipper_length_mm

VALUES:     333 (100%)      MAX      231.0      RANGE     59.0
MISSING:    ---             95%      225.0      IQR       23.0
                            Q3       213.0      STD       14.0
DISTINCT:   54  (16%)       AVG      201.0      VAR        196
                            MEDIAN   197.0
ZEROES:     ---             Q1       190.0      KURT.   -0.961
                            5%       181.0      SKEW     0.360
                            MIN      172.0      SUM     66,922

6 body_mass_g

VALUES:     333 (100%)      MAX      6,300      RANGE    3,600
MISSING:    ---             95%      5,670      IQR      1,225
                            Q3       4,775      STD        805
DISTINCT:   93  (28%)       AVG      4,207      VAR       648k
                            MEDIAN   4,050
ZEROES:     ---             Q1       3,550      KURT.   -0.733
                            5%       3,150      SKEW     0.472
                            MIN      2,700      SUM       1.4M

7 sex

VALUES:     333 (100%)
MISSING:    ---

DISTINCT:   2  (<1%)
```

As we can see that using just two lines of code, SweetViz creates an EDA report in a matter of minutes.

You can also create a correlation matrix graph by clicking the Associations Tab,



Associations

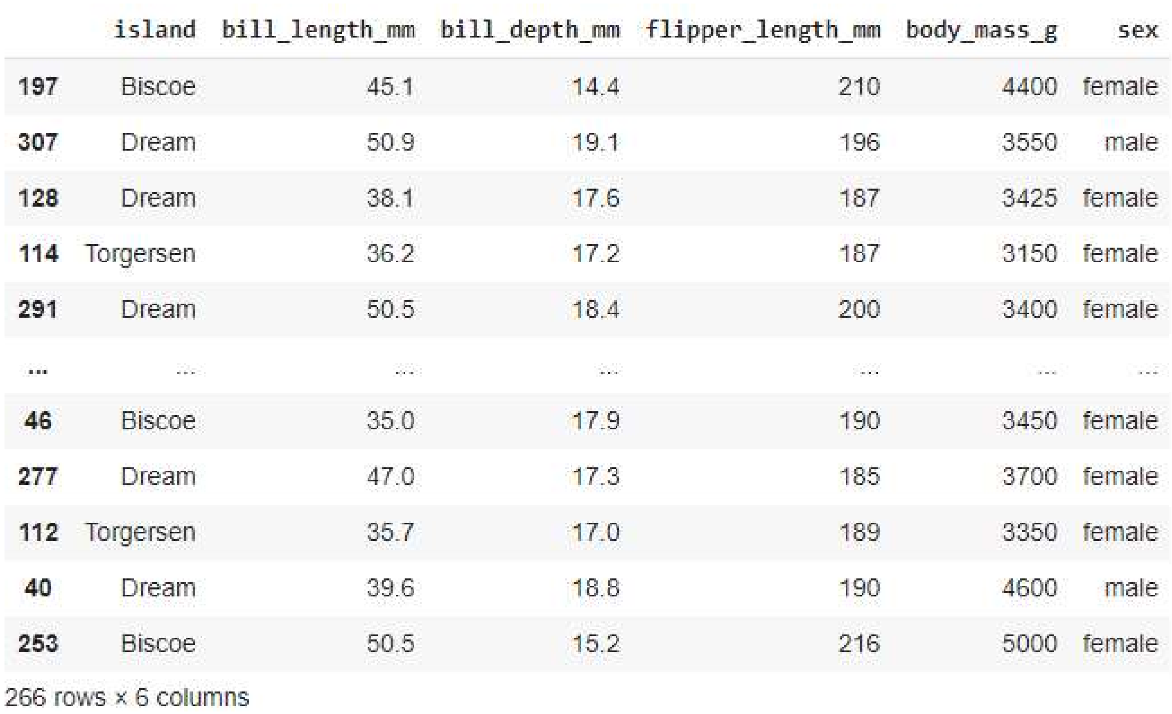
[Only including dataset "DataFrame"]

■ **Squares** are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **assymmetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• **Circles** are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.

Now Lets Move Ahead and create a comparison report of our Train versus Test Dataset,

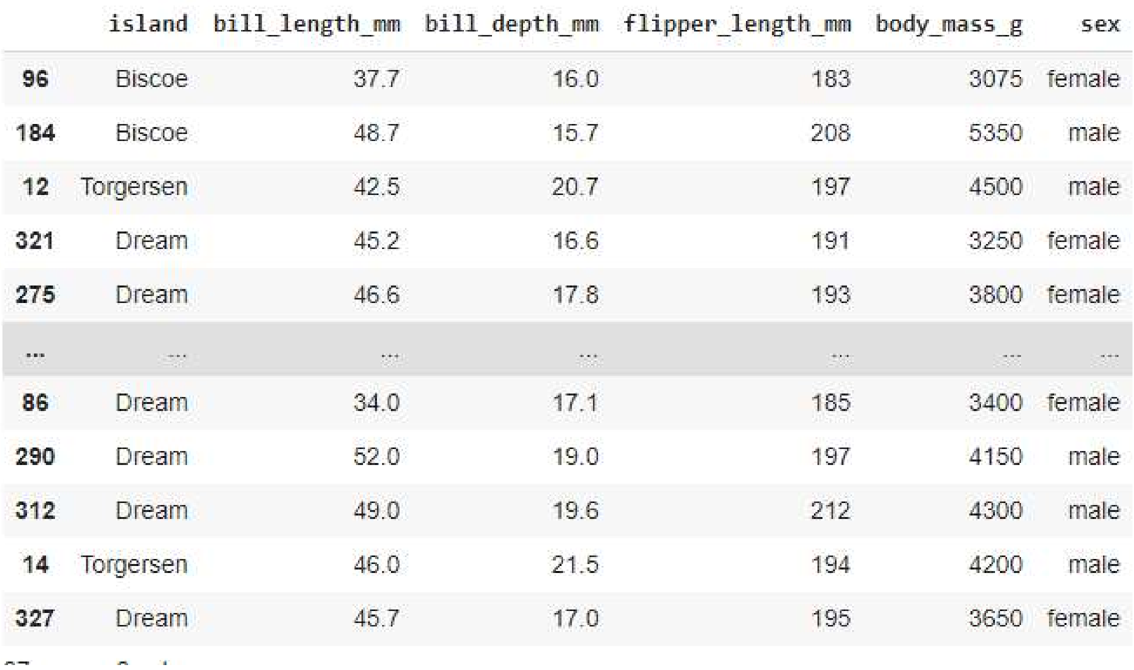
```
#splitting into train and test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2)



#displaying Train set
X_train
```



	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
197	Biscoe	45.1	14.4	210	4400	female
307	Dream	50.9	19.1	196	3550	male
128	Dream	38.1	17.6	187	3425	female
114	Torgersen	36.2	17.2	187	3150	female
291	Dream	50.5	18.4	200	3400	female
...	...	...	...	...	...	...
46	Biscoe	35.0	17.9	190	3450	female
277	Dream	47.0	17.3	185	3700	female
112	Torgersen	35.7	17.0	189	3350	female
40	Dream	39.6	18.8	190	4600	male
253	Biscoe	50.5	15.2	216	5000	female

266 rows × 6 columns

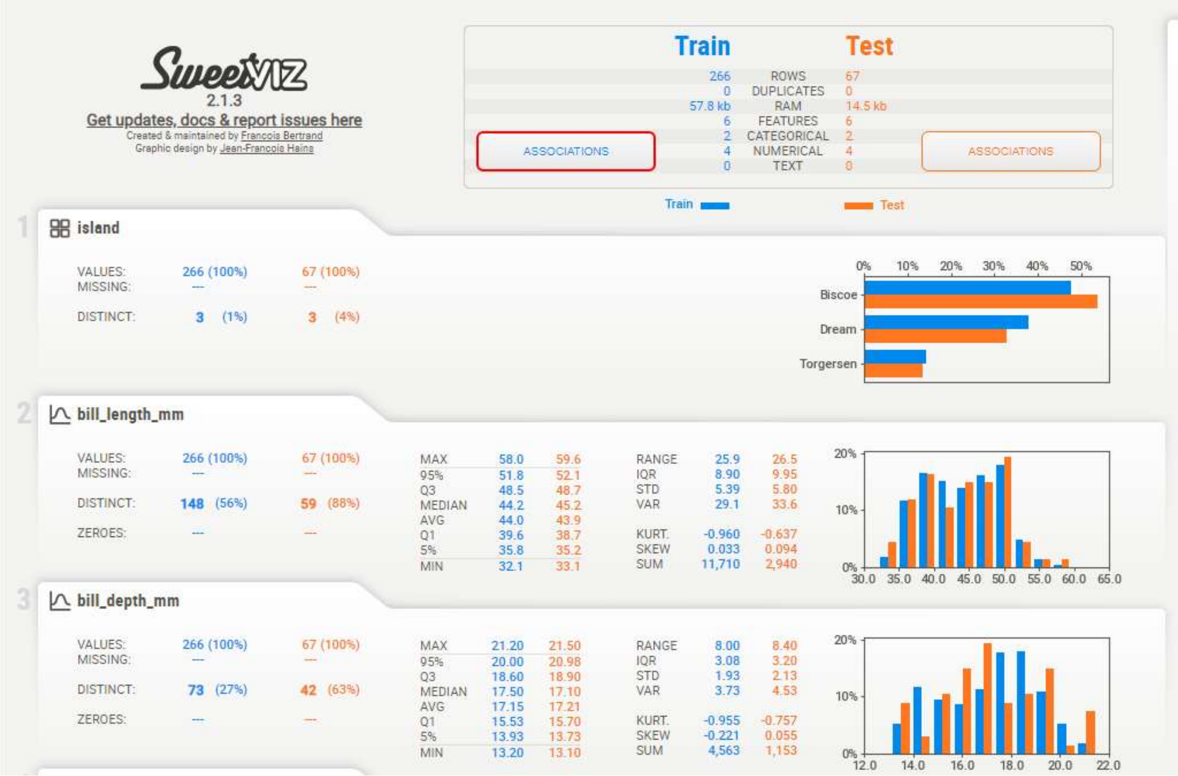
```
#displaying Test set
X_test
```



	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
96	Biscoe	37.7	16.0	183	3075	female
184	Biscoe	48.7	15.7	208	5350	male
12	Torgersen	42.5	20.7	197	4500	male
321	Dream	45.2	16.6	191	3250	female
275	Dream	46.6	17.8	193	3800	female
...	...	...	...	...	...	...
86	Dream	34.0	17.1	185	3400	female
290	Dream	52.0	19.0	197	4150	male
312	Dream	49.0	19.6	212	4300	male
14	Torgersen	46.0	21.5	194	4200	male
327	Dream	45.7	17.0	195	3650	female

Creating the comparison report between Train versus Test,

```
import sweetviz as sv
compare_report = sv.compare([X_train, 'Train'], [X_test,
'Test'])
compare_report.show_html('compare.html',
open_browser=False)

#displaying results
import IPython
IPython.display.HTML('compare.html')
```



SweetViz 2.1.3

Get updates, docs & report issues here

Created & maintained by Francois Bertrand

Graphic design by Jean-Francois Hains

Train		Test
266	ROWS	67
0	DUPLICATES	0
57.8 kb	RAM	14.5 kb
6	FEATURES	6
2	CATEGORICAL	2
4	NUMERICAL	4
0	TEXT	0

1 island

VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	3 (1%)	3 (4%)

2 bill_length_mm

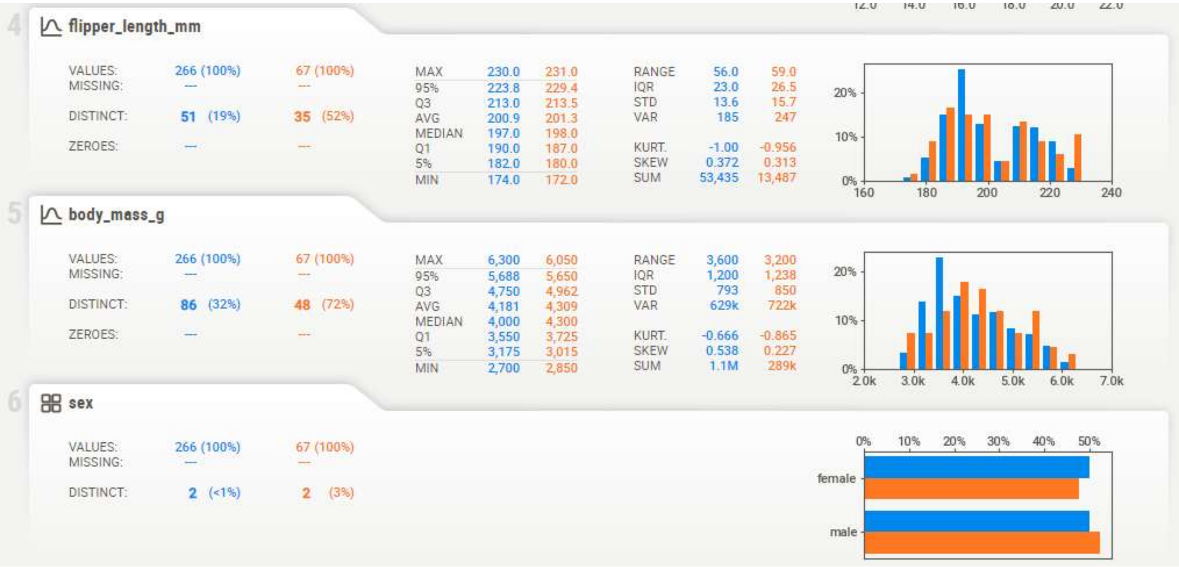
VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	148 (56%)	59 (88%)
ZEROES:	---	---

MAX	58.0	59.6	RANGE	25.9	26.5
95%	51.8	52.1	IQR	8.90	9.95
Q3	48.5	48.7	STD	5.39	5.80
MEDIAN	44.2	45.2	VAR	29.1	33.6
AVG	44.0	43.9			
Q1	39.6	38.7	KURT.	-0.960	-0.637
5%	35.8	35.2	SKEW	0.033	0.094
MIN	32.1	33.1	SUM	11,710	2,940

3 bill_depth_mm

VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	73 (27%)	42 (63%)
ZEROES:	---	---

MAX	21.20	21.50	RANGE	8.00	8.40
95%	20.00	20.98	IQR	3.08	3.20
Q3	18.60	18.90	STD	1.93	2.13
MEDIAN	17.50	17.10	VAR	3.73	4.53
AVG	17.15	17.21			
Q1	15.53	15.70	KURT.	-0.955	-0.757
5%	13.93	13.73	SKEW	-0.221	0.055
MIN	13.20	13.10	SUM	4,563	1,153



4 flipper_length_mm

VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	51 (19%)	35 (52%)
ZEROES:	---	---

MAX	230.0	231.0	RANGE	56.0	59.0
95%	223.8	229.4	IQR	23.0	26.5
Q3	213.0	213.5	STD	13.6	15.7
AVG	200.9	201.3	VAR	185	247
MEDIAN	197.0	198.0			
Q1	190.0	187.0	KURT.	-1.00	-0.956
5%	182.0	180.0	SKEW	0.372	0.313
MIN	174.0	172.0	SUM	53,435	13,487

5 body_mass_g

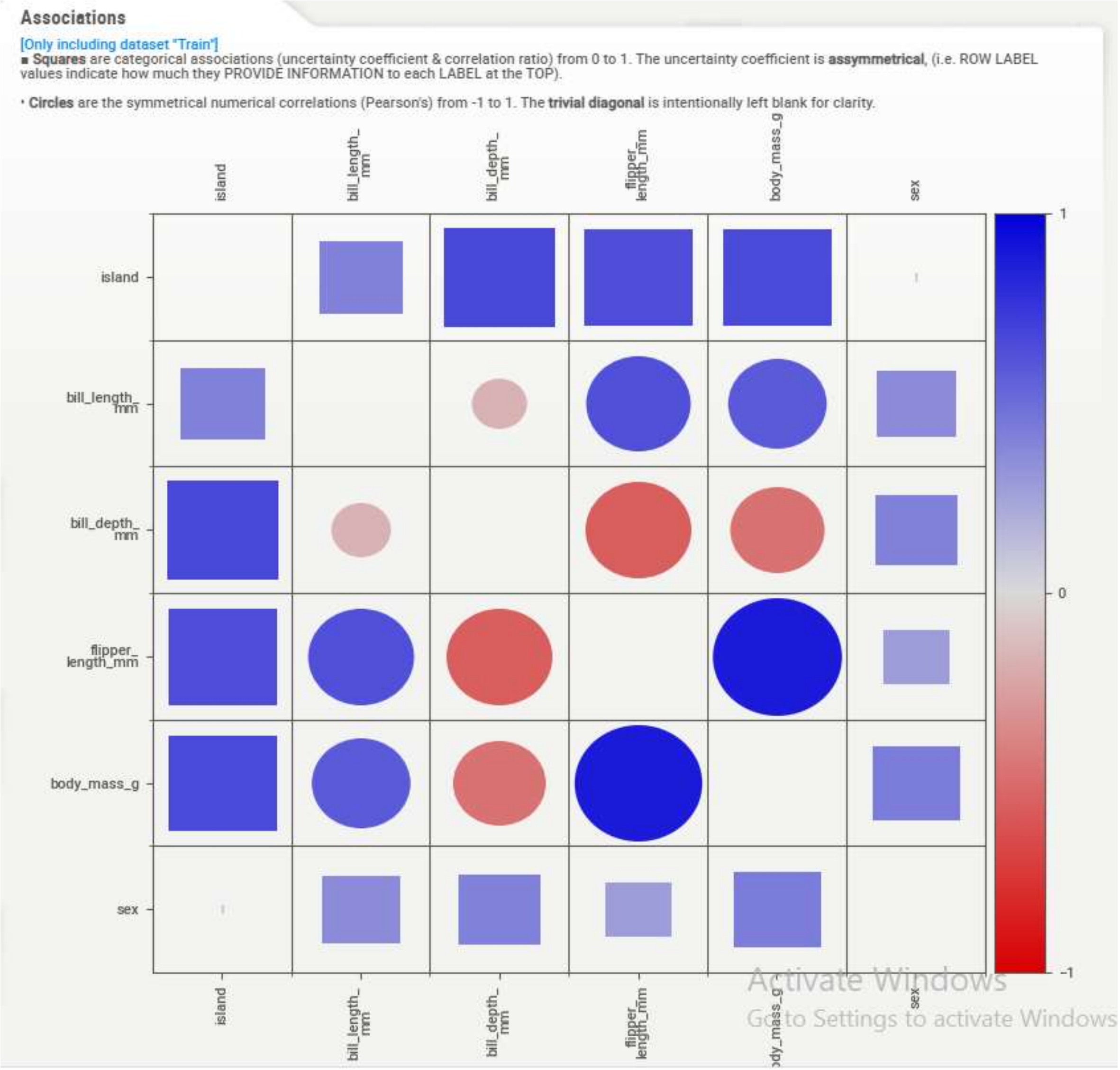
VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	86 (32%)	48 (72%)
ZEROES:	---	---

MAX	6,300	6,050	RANGE	3,600	3,200
95%	5,688	5,650	IQR	1,200	1,238
Q3	4,750	4,962	STD	793	850
AVG	4,181	4,309	VAR	629k	722k
MEDIAN	4,000	4,300			
Q1	3,550	3,725	KURT.	-0.666	-0.865
5%	3,175	3,015	SKEW	0.538	0.227
MIN	2,700	2,850	SUM	1.1M	289k

6 sex

VALUES:	266 (100%)	67 (100%)
MISSING:	---	---
DISTINCT:	2 (<1%)	2 (3%)

Comparative Association,



Associations

[Only including dataset "Train"]

■ **Squares** are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **assymmetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• **Circles** are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.

As we can observe, SweetViz holds the capacity to give us a detailed comparison and EDA to help observe and discover insights from our dataset processed.

EndNotes

This article tried to understand the essence of performing Exploratory Data Analysis and answered why it is important. We also got to know how to create an interactive EDA report with just a few lines of code using SweetViz. Finally, I recommend that the reader even further explore the SweetViz library using complex datasets and analyse the insights.

The following implementation can be found as a colab notebook, using the link here. (https://colab.research.google.com/drive/1wRSLf_UPEf_NnFmXxkShmh2DYAvVKPmV?usp=sharing)

Happy Learning!

References

- Official SweetViz GitHub Repo (https://github.com/fbdesignpro/sweetviz)
- Using SweetViz In Python (https://pypi.org/project/sweetviz/)

- Performing EDA In Sweetviz (https://morioh.com/p/c2676dee031a)

(https://analyticsindiamag.com/author/victor-deyanalyticsindiamag-com/)

Victor is an aspiring Data Scientist & is a Master of Science in Data Science & Big Data Analytics. He is a Researcher, a Data Science Influencer and also an Ex-University Football Player. A keen learner of new developments in Data Science and Artificial Intelligence, he is committed to growing the