EDA made easy

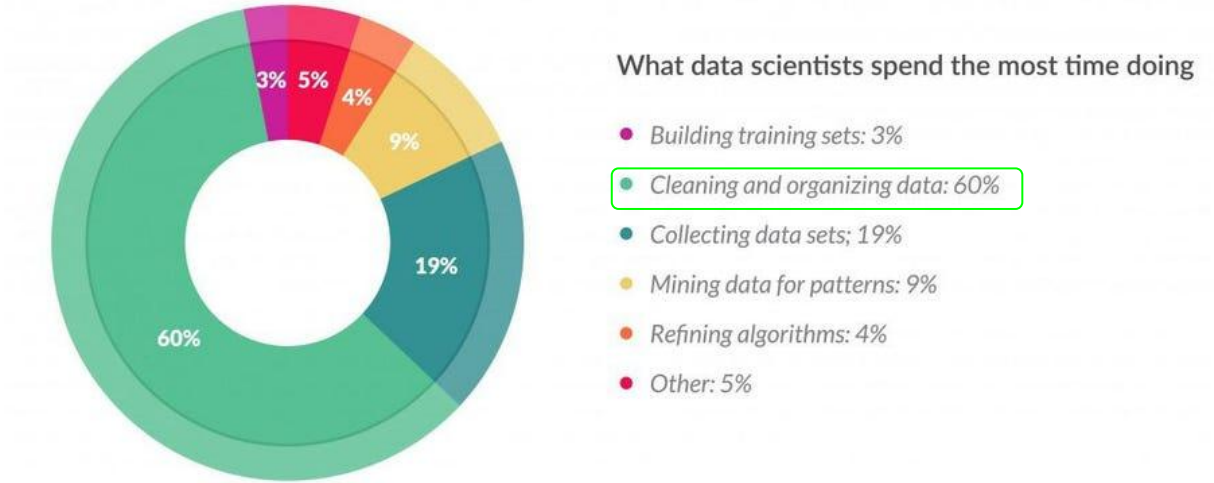
EDAhelper - a Python package

Group 5: Vera, Rowan, Jennifer & Steven

What is EDA and what is the problem?

EDA stands for *Exploratory Data Analysis*

- 60% = “Cleaning and organizing data”
- 70% of data scientists think “least enjoyable.”



1. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=7b1196b66f63

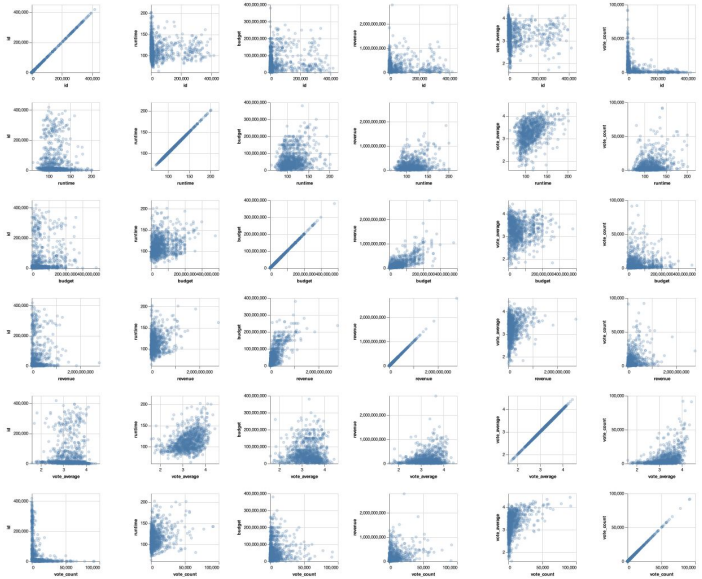
The Value of EDAhelper

Without **EDAhelper**

- Many lines of code
  - Error-prone
  - Time-consuming
- R has an easy package called GGally, but no well-known Python package.

With **EDAhelper**

- One line of code

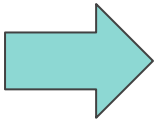


Preprocessing

```
preprocess(file_path, **kwarg)
```

BEFORE:

- ➔ Which function to read data?
- ➔ Need to check missing values?
- ➔ Which method for imputation?



AFTER:

- **One line of code**
- **Capable of standard imputation**
- **Compatible with other functions:**

  pandas.read_csv()

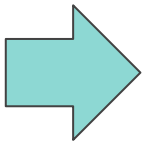
  pandas.read_excel()

  pandas.read_pickle()

Column Stats

`column_stats(data, columns)`

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



	Column	Count	Mean	Median	Mode	Q1	Q3	Var	Stdev
0	sepal_width	150.0	3.057	3.00	3.0	2.8	3.3	0.190	0.436
1	petal_length	150.0	3.758	4.35	1.4	1.6	5.1	3.116	1.765

Summary statistics

Correlation

	sepal_width	petal_length
sepal_width	1.00000	-0.42844
petal_length	-0.42844	1.00000

Covariance matrices

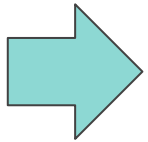
	sepal_width	petal_length
sepal_width	0.189979	-0.329656
petal_length	-0.329656	3.116278

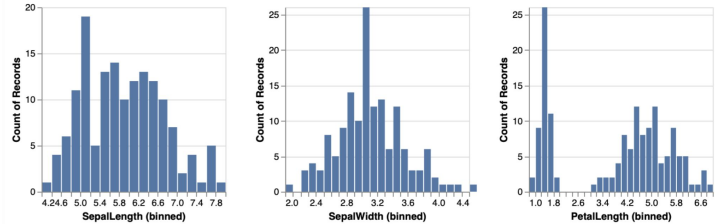
- **Select any number of numeric columns**
- **Get summary statistics, correlation, and covariance matrix instantly**

Histograms

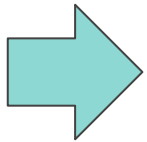
```
plot_histogram(df, columns, num_bins)
```

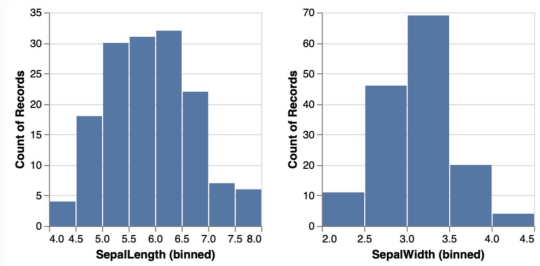
- **Fast**: plot all numeric columns in one line





- **Flexible**: choose specific columns and number of bins

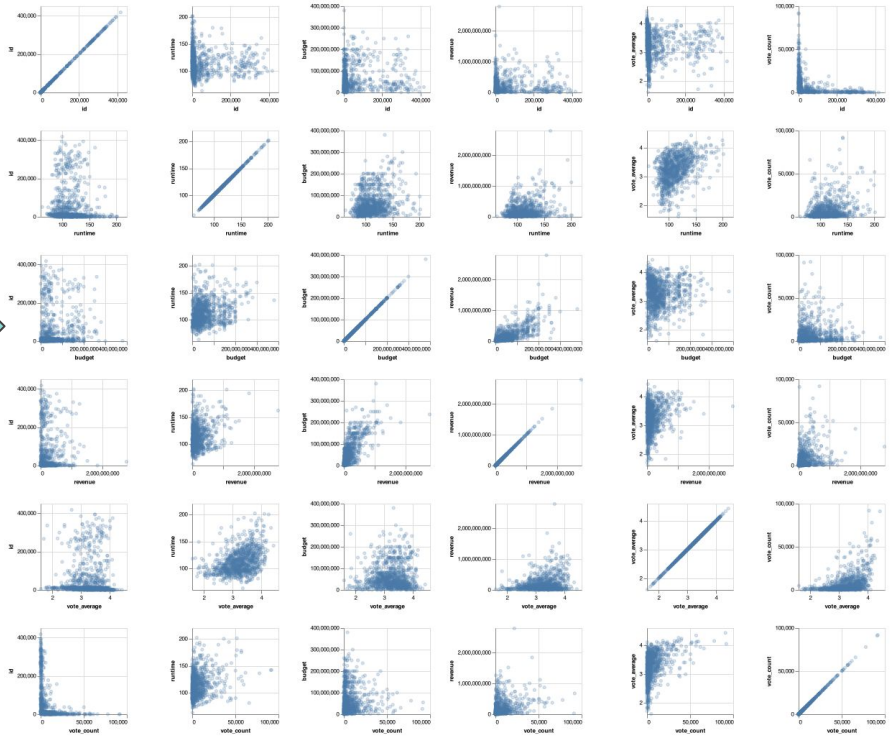




Matrix of Numeric Plots

```
numeric_plots(df)
```

- Select only numeric columns
- Generate a matrix of scatter plots



EDAhelper - Summary

Data Processing	Summary Statistics	Visualization
`preprocess(file_path, **kwarg)`	`column_stats(data, columns)`	`plot_histogram(df, columns, num_bins)`
		`numeric_plots(df)`

Key benefits:

- Four one-line function calls covering EDA. Better readability.
- Life is short - Less time coding. More time on the enjoyable things.

EDAhelper

To find out more:
https://github.com/UBC-MDS/EDAhelper/