Process and Rationale of Exploratory Analytical Methods

Jessica Chipera

What is EDA

Exploratory Data Analysis (EDA) is a method of using statistics and visualization techniques to explore and summarize essential elements of data.

It is used before engaging in a data analysis or machine learning project.

There are several approaches, including:

- Univariate
- Bivariate
- Multivariate
- (Martinez et all, 2017).

Purposes of EDA

- Determine the underlying structure of data
- Detect the most optimal approach for treating missing values, outliers, and other anomalies
- Discovering the shape of the data
- Identification of significant correlations and relationships in the data
- Gain a better understanding of the data before making assumptions about what the data means
- (Martinez et all, 2017).

How to Use EDA

- Exploratory Data Analysis is done using Python or R. Some useful Python libraries include (Mukhiya & Ahmad, 2020):
 - Numpy
 - Matplotlib.Pyplot
 - Scipy
 - Pandas
 - Seaborn
 - Sklearn (various forms, depending on the data)
 - Statsmodels
- The first step is to evaluate the data and determine what kinds of variables are present. The df.info() function in Python will be useful.
- Next, clean the data. Check for duplicates and NaN values.

How to Use EDA (Continued)

- The describe function will be helpful in determining the mean, standard deviation, IQR values, and min and max of the data.
- Then, determine the shape of the data. This can be done using a histogram or several histograms.
- Sometimes it is useful to use a strip plot to further examine the shape of data as it relates to different dimensionality.
- Next, you can identify significant correlations and relationships with a heatmap. Variables that show a high correlation or high inverse correlation can be analyzed with a scatterplot.
- A boxplot should be used to determine the extent of outliers.

• (Martinez et all, 2017; Mukhiya & Ahmed, 2020).

Univariate Non-Graphical

- Univariate analysis is the simplest form of data analysis in which only one variable is analyzed.
- Like other statistics, it can be either inferential or descriptive in nature.
- Sometimes univariate analysis can generate misleading results if multivariate analysis would have been more appropriate for the data.
- Some non-graphical univariate analysis results could be standard deviation, mode, median, mean, or interquartile range.

• (Martinez et all, 2017; Mukhiya & Ahmed, 2020).

Univariate Graphical

- Graphs can be used for univariate analysis. Some of these graphs could be:
 - Pie charts
 - Bar charts
 - Histograms
 - Box plots
 - Strip plots
- The chart provided shows a hypothetical sample of clients per insurance company for a Financial Services company.
- (Martinez et all, 2017; Mukhiya & Ahmed, 2020).



Multivariate Non-Graphical

- Multivariate analysis looks for trends in multiple variables
- Studies multiple variables simultaneously
- More informative than univariate analysis
- Like other statistics, it can be either inferential or descriptive in nature.
- Non-graphical multivariate analysis results would ordinarily be organized into tables or lists.
- (Martinez et all, 2017; Mukhiya & Ahmed, 2020).

Multivariate Graphical

- Graphs can be helpful in multivariate analysis. The same graphs that were introduced in univariate analysis can be used.
- However, there are some additional charts that can be useful in multivariate analysis.
- The chart provided is a heatmap which shows correlations between variables. Black boxes are a strong negative correlation, and orange boxes are a strong positive correlation.
- (Uhler, 2020).



Multivariate Graphical (Continued)

- Scatter plots are also extremely useful in multivariate analysis.
- In addition to revealing some outliers in the lower right corner, this scatter plot shows a cluster of grocery store customers who buy both wine and meat in the same visit.

• (Uhler, 2020).



EDA in Data Science: An Example

- Data Scientists use Exploratory Data Analysis to look for relationships and correlations in data that could help train machine learning models or test hypotheses.
- Say you own a grocery store and you want a data scientist to help you with your marketing.
- A heatmap might show a high correlation between purchases of meat and purchases of wine and a high negative correlation between purchases of meat products and internet sales.
- In fact, maybe this data reveals that your website is a problem because web visitors only buy discounted goods.
- What's worse is that your data scientist discovers that most people visit your website five times per month. That's lost revenue on more expensive products!

EDA in Data Science: Continued



EDA in Data Science: Continued



ID -	- 1	2.8e-05	0.013	0.0024	-0.0026	-0.047	-0.023	0.0046	0.0044	-0.024	0.0076	-0.013	-0.037	-0.019	-0.0034	-0.015	-0.0074	-0.036	-0.025	-0.0075	-0.022	-0.013	0.034	-0.022
Year_Birth ·	2.8e-05	1	-0.16	0.23	-0.35	-0.02	-0.16	-0.018	-0.031	-0.042	-0.018	-0.062	-0.061	-0.15	-0.12	-0.13	0.12	0.062	-0.061	0.0071-	0.0059	-0.014	-0.03	0.021
Income -	0.013	-0.16	1	-0.43	0.019	-0.004	0.58	0.43	0.58	0.44	0.44	0.33	-0.083	0.39	0.59	0.53	-0.55	-0.016	0.18	0.34	0.28	0.083	-0.027	0.13
Kidhome -	0.0024	0.23	-0.43	1	-0.036	0.0088	-0.5	-0.37	-0.44	-0.39	-0.37	-0.35	0.22	-0.36	-0.5	-0.5	0.45	0.015	-0.16	-0.21	-0.17	-0.08	0.04	-0.08
Teenhome -	0.0026	-0.35	0.019	-0.036	1	0.016	0.0048	-0.18	-0.26	-0.2	-0.16	-0.022	0.39	0.16	-0.11	0.051	0.13	-0.043	0.039	-0.19	-0.14	-0.012	0.0031	-0.15
Recency -	-0.047	-0.02	-0.004	0.0088	0.016	1	0.016	-0.0043	0.023	0.0011	0.023	0.017	-0.0011	-0.011	0.025	0.0008	-0.021	-0.033	0.019	0.00013	-0.019 -	0.0079	0.013	-0.2
MntWines -	-0.023	-0.16	0.58	-0.5	0.0048	0.016	1	0.39	0.56		0.39	0.39	0.011	0.54	0.64	0.64	-0.32	0.062	0.37	0.47	0.35	0.21	-0.039	0.25
MntFruits -	0.0046	-0.018		-0.37	-0.18	-0.0043	0.39	1				0.39	-0.13	0.3			-0.42	0.015	0.01	0.22	0.19	-0.01	-0.0052	0.13
MntMeatProducts	0.0044	-0.031	0.58	-0.44	-0.26	0.023		0.54	1	0.57		0.35	-0.12	0.29	0.72		-0.54	0.018	0.1	0.37	0.31	0.032	-0.023	0.24
MntFishProducts	-0.024	-0.042	0.44	-0.39	-0.2	0.0011				1		0.42	-0.14	0.29			-0.45 (0.00036	50.017	0.2	0.26	0.0053	-0.021	0.11
AntSweetProducts	0.0076	-0.018		-0.37	-0.16	0.023	0.39				1	0.37	-0.12	0.35			-0.42	0.0015	0.029	0.26	0.24	0.01	-0.022	0.12
MntGoldProds -	-0.013	-0.062	0.33	-0.35	-0.022	0.017	0.39	0.39	0.35		0.37	1	0.049			0.38	-0.25	0.12	0.022	0.18	0.17	0.05	-0.031	0.14
mDealsPurchases ·	-0.037	-0.061	-0.083	0.22	0.39	-0.0011	0.011	-0.13	-0.12	-0.14	-0.12	0.049	1	0.23	-0.0086	0.069	0.35	-0.023	0.016	-0.18	-0.12	-0.034	0.00042	0.0022
umWebPurchases -	-0.019	-0.15	0.39	-0.36	0.16	-0.011		0.3	0.29	0.29	0.35		0.23	1	0.38		-0.056	0.042	0.16	0.14	0.16	0.035	-0.016	0.15
CatalogPurchases -	0.0034	-0.12	0.59	-0.5	-0.11	0.025			0.72				-0.0086	0.38	1	0.52	-0.52	0.1	0.14	0.32	0.31	0.1	-0.02	0.22
ImStorePurchases -	-0.015	-0.13	0.53	-0.5	0.051	0.0008	0.64	0.46	0.48	0.46	0.45	0.38	0.069	0.5	0.52	1	-0.43	-0.068	0.18	0.21	0.18	0.085	-0.017	0.039
mWebVisitsMonth -	0.0074	0.12	-0.55	0.45	0.13	-0.021	-0.32	-0.42	-0.54	-0.45	-0.42	-0.25	0.35	-0.056	-0.52	-0.43	1	0.061	-0.032	-0.28	-0.19	0.0035	0.02	-0.004
AcceptedCmp3 ·	-0.036	0.062	-0.016	0.015	-0.043	-0.033	0.062	0.015	0.018	0.00036	0.0015	0.12	-0.023	0.042	0.1	-0.068	0.061	1	-0.08	0.08	0.095	0.059	0.0084	0.25
AcceptedCmp4	-0.025	-0.061	0.18	-0.16	0.039	0.019	0.37	0.01	0.1	0.017	0.029	0.022	0.016	0.16	0.14	0.18	-0.032	-0.08	1	0.31	0.25	0.3	-0.028	0.18
AcceptedCmp5	0.0075	0.0071	0.34	-0.21	-0.19	0.00013		0.22	0.37	0.2	0.26	0.18	-0.18	0.14	0.32	0.21	-0.28	0.08	0.31	1	0.4	0.21	-0.0094	0.33
AcceptedCmp1 ·	-0.022	-0.0059	0.28	-0.17	-0.14	-0.019	0.35	0.19	0.31	0.26	0.24	0.17	-0.12	0.16	0.31	0.18	-0.19	0.095	0.25	0.4	1	0.16	-0.025	0.29
AcceptedCmp2 ·	-0.013	-0.014	0.083	-0.08	-0.012	-0.0079	0.21	-0.01	0.032	0.0053	0.01	0.05	-0.034	0.035	0.1	0.085	-0.0035	0.059	0.3	0.21	0.16	1	-0.011	0.16
Complain -	0.034	-0.03	-0.027	0.04	0.0031	0.013	-0.039	-0.0052	-0.023	-0.021	-0.022	-0.031	0.00042	-0.016	-0.02	-0.017	0.02	0.0084	-0.028	-0.0094	-0.025	-0.011	1	-0.0017
Response -	-0.022	0.021	0.13	-0.08	-0.15	-0.2	0.25	0.13	0.24	0.11	0.12	0.14	0.0022	0.15	0.22	0.039	-0.004	0.25	0.18	0.33	0.29	0.16	-0.0017	1
	0	'tear_Birth -	Income -	Kidhome -	Teenhome -	Recency -	MntWines -	MntFruits -	MntMeatProducts -	MntFishProducts -	MntSweetProducts -	MntGoldProds -	NumDealsPurchases -	NumWebPurchases -	umCatalogPurchases -	NumStorePurchases -	NumWebVisitsMonth -	AcceptedCmp3 -	AcceptedCmp4 -	AcceptedCmp5 -	AcceptedCmp1 -	AcceptedCmp2 -	Complain -	Response -

This is a heatmap similar to what a data scientist could find when conducting Multivariate Analysis for our hypothetical grocery store problem.

Take a minute to notice the black boxes indicating a strong negative correlation, and the orange boxes indicating a strong positive correlation.

(Uhler, 2020.)

-1.0

- 0.8

- 0.6

- 0.4

- 0.2

- 0.0

- -0.2

- -0.4

EDA in Data Science: Continued

- Perhaps the data scientist will suggest that you put a meat-wine pairing algorithm on your website to remind website visitors that they want wine and meat.
- The website could also include recipes
- Or potentially a recommendation algorithm similar to what Amazon has, to suggest goods to customers based on their past purchases.
 - You bought diapers one month ago. Do you need more?
 - Also, do you need more formula?
 - It's spring. Do you need Benadryl for your allergies?

References

- 365 Careers. (2018). Statistics for data science and business analysis. Packt Publishing.
- Martinez, W. (2017). Exploratory data analysis with MATLAB. CRC Press, Taylor & Francis Group.
- Mukhiya, S. K., & Ahmed, U. (2020). Hands-on exploratory data analysis with Python. Packt Publishing.
- Uhler, C. (2022). Exploratory Data Analysis for Customer Segmentation: Grocery Store Case Study. MIT.