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.

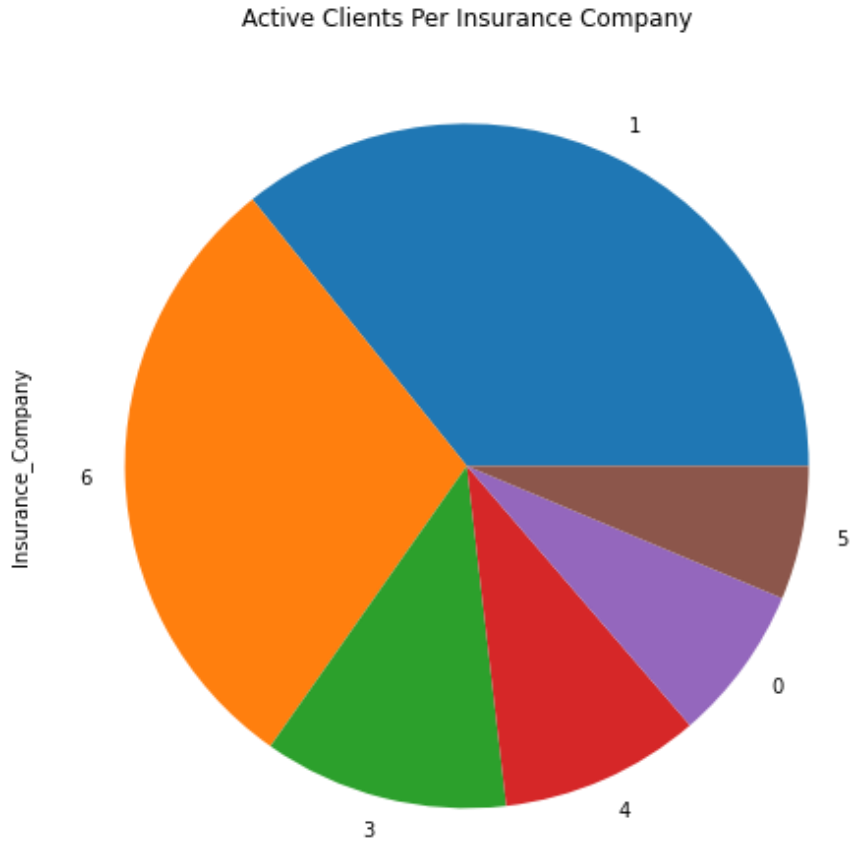
- (Martinez et all, 2017; Mukhiya & Ahmed, 2020).

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

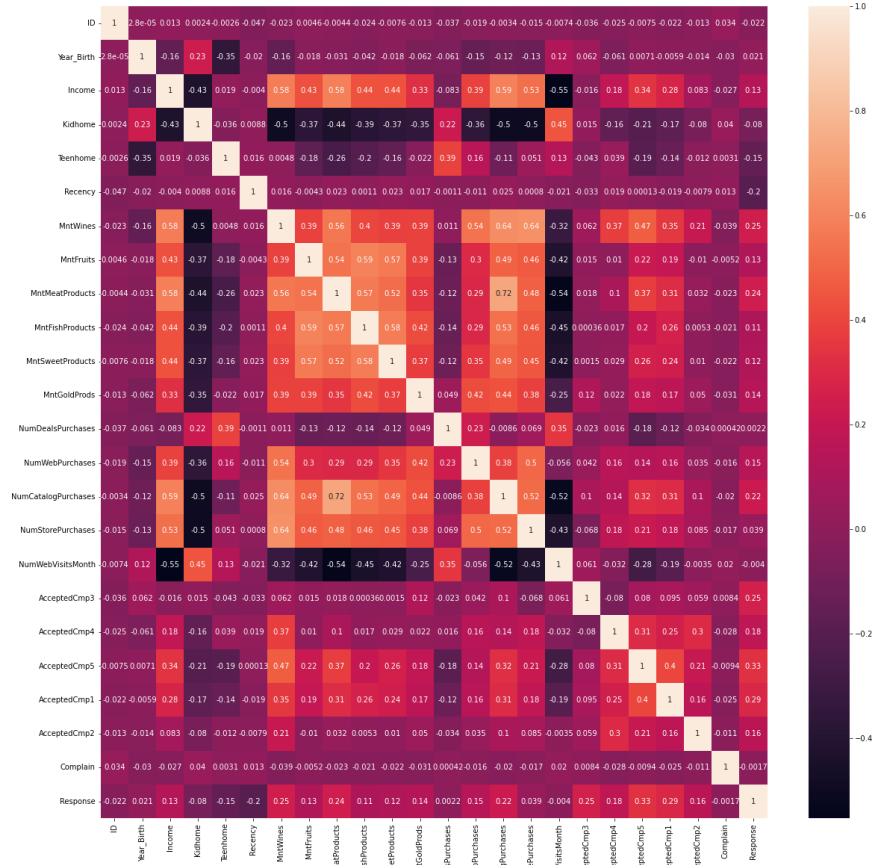
- Multivariate analysis looks for trends in multiple variables
- Studies multiple variables simultaneously
- More informative than univariate analysis
- Like other statistics, it can be either inferential or descriptive in nature.

- Non-graphical multivariate analysis results would ordinarily be organized into tables or lists.

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

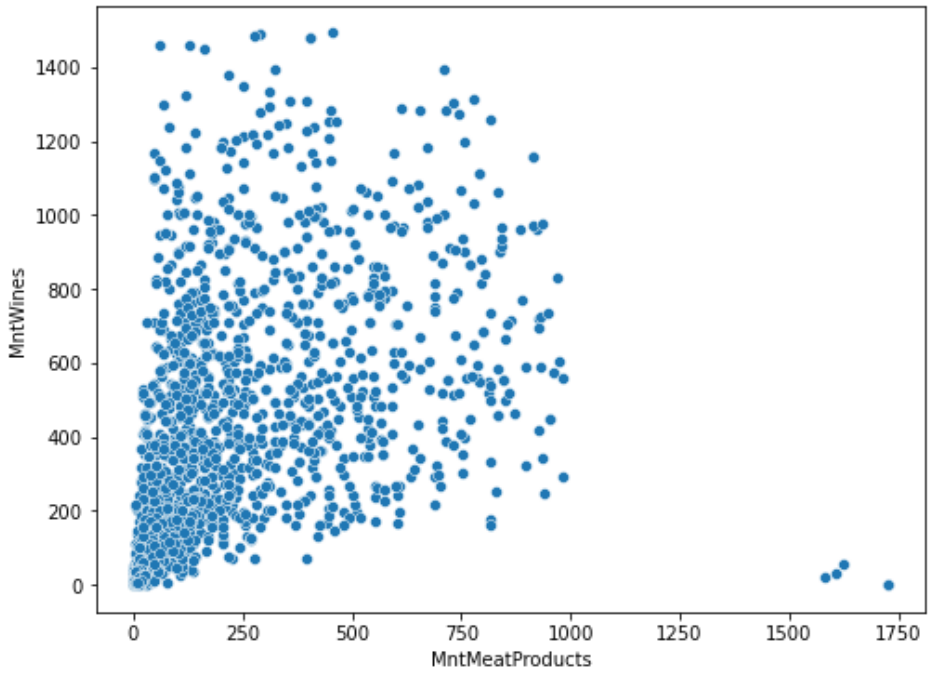
Multivariate Graphical

- 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)

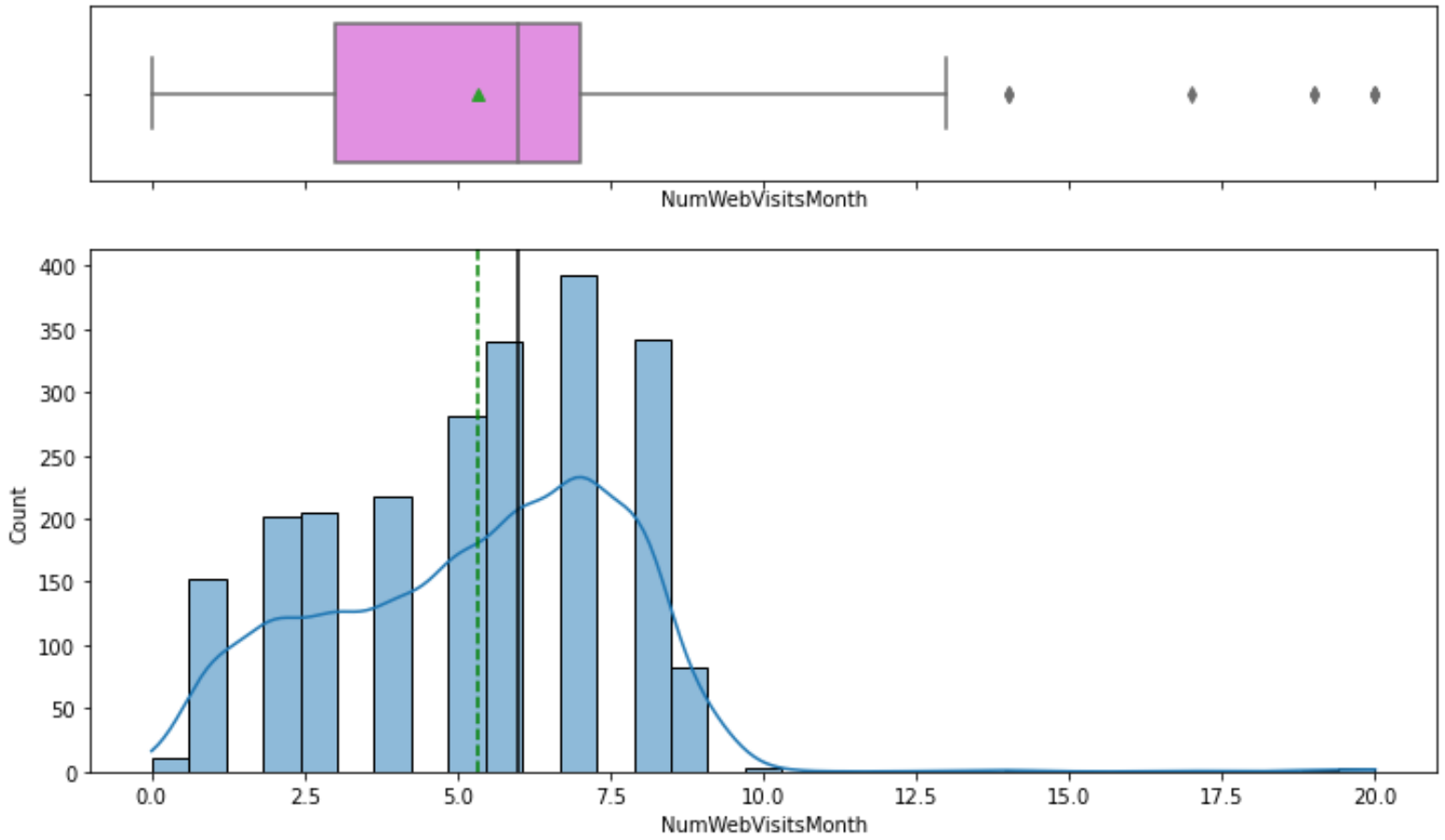
- Scatter plots are also extremely useful in multivariate analysis.
- In addition to revealing some outliers in the lower right corner, this scatter plot shows a cluster of grocery store customers who buy both wine and meat in the same visit.
- (Uhler, 2020).



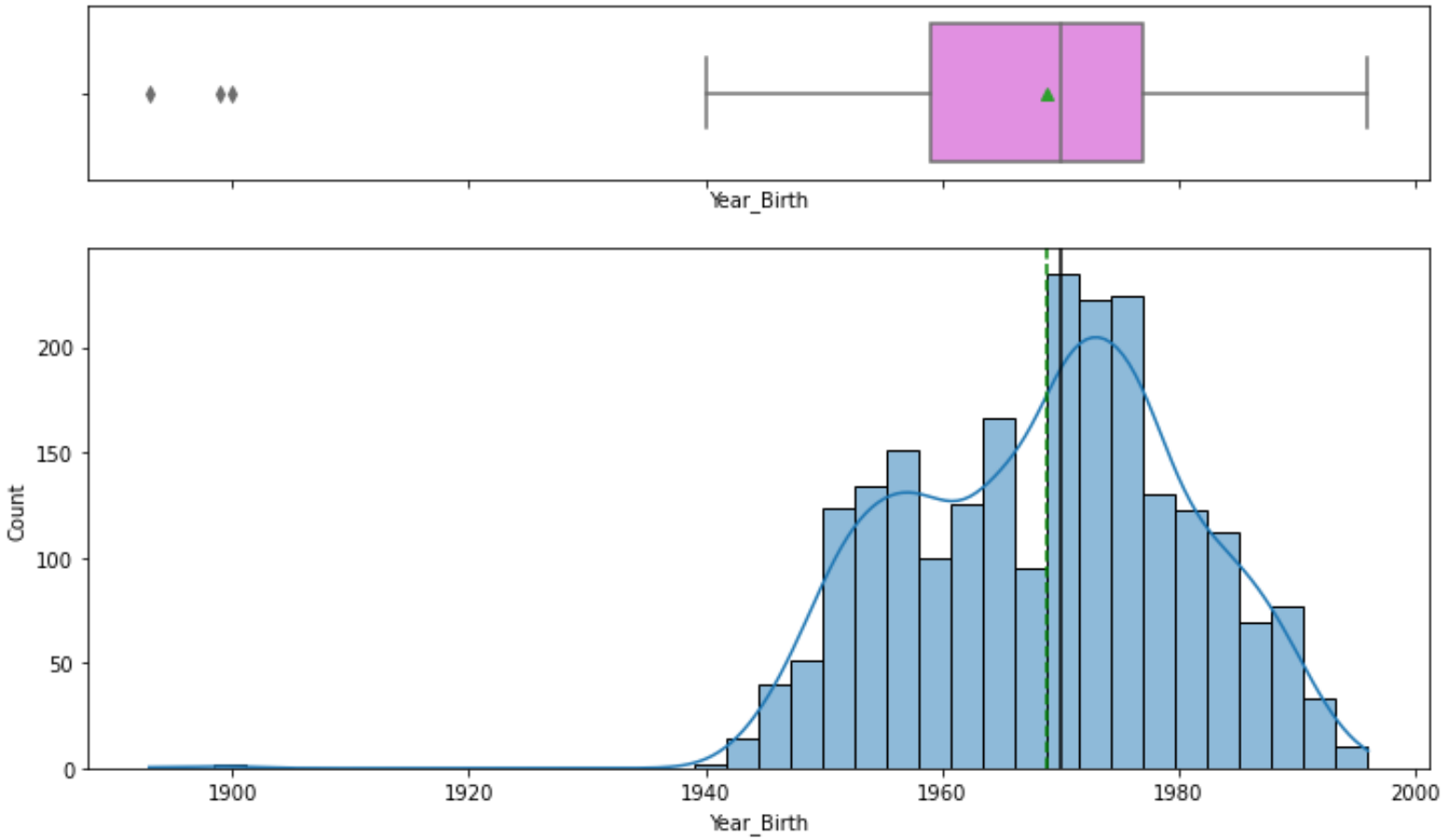
EDA in Data Science: An Example

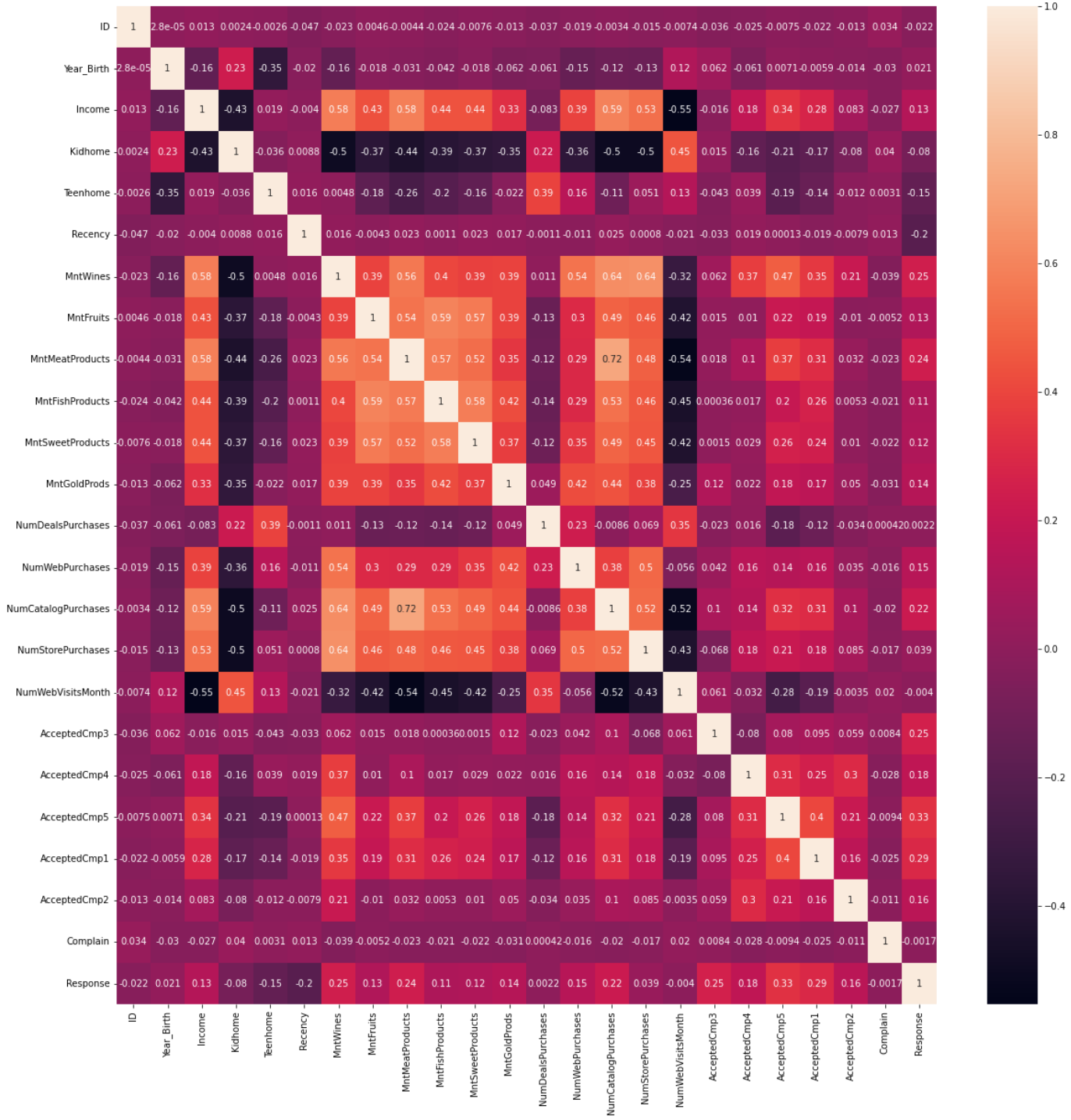
- Data Scientists use Exploratory Data Analysis to look for relationships and correlations in data that could help train machine learning models or test hypotheses.
- Say you own a grocery store and you want a data scientist to help you with your marketing.
- A heatmap might show a high correlation between 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





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.)

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.