The Value of Exploratory Data Analysis

<= Previous post
Next post =>

Tags: Data Analysis, Data Exploration, Data Visualization, SVDS

In this post, we will give a high level overview of what exploratory data analysis (EDA) typically entails and then describe three of the major ways EDA is critical to successfully model and interpret its results.

By Chloe Mawer, Silicon Valley Data Science.

From the outside, data science is often thought to consist wholly of advanced statistical and machine learning techniques. However, there is another key component to any data science endeavor that is often undervalued or forgotten: exploratory data analysis (EDA). At a high level, EDA is the practice of using visual and quantitative methods to understand and summarize a dataset without making any assumptions about its contents. It is a crucial step to take before diving into machine learning or statistical modeling because it provides the context needed to develop an appropriate model for the problem at hand and to correctly interpret its results.

With the rise of tools that enable easy implementation of powerful machine learning algorithms, it can become tempting to skip EDA. While it's understandable why people take advantage of these algorithms, it's not always a good idea to simply feed data into a black box—we have observed over and over again the critical value EDA provides to all types of data science problems.

EDA is valuable to the data scientist to make certain that the results they produce are valid, correctly interpreted, and applicable to the desired business contexts. Outside of ensuring the delivery of technically sound results, EDA also benefits business stakeholders by confirming they are asking the right questions and not biasing the investigation with their assumptions, as well as by providing the context around the problem to make sure the potential value of the data scientist's output can be maximized. As a bonus, EDA often leads to insights that the business stakeholder or data scientist wouldn't even think to investigate but that can be hugely informative about the business.

In this post, we will give a high level overview of what EDA typically entails and then describe three of the major ways EDA is critical to successfully model and interpret its results. Whether you are a data scientist or the consumer of data science, we hope after reading this post that you will know why EDA should be a key part of the way data science operates in your organization.

What is EDA?

While aspects of EDA have existed as long as data has been around to analyze, John W. Tukey, who wrote the book Exploratory Data Analysis in 1977, was said to have coined the phrase and developed the field. At a high level, EDA is used to understand and summarize the contents of a dataset, usually to investigate a specific question or to prepare for more advanced modeling. EDA typically relies heavily on visualizing the data to assess patterns and identify data characteristics that the analyst would not otherwise know to look for. It also takes advantage of a number of quantitative methods to describe the data.

EDA usually involves some combination of the following methods:

- Univariate visualization of and summary statistics for each field in the raw dataset (see figure 1)

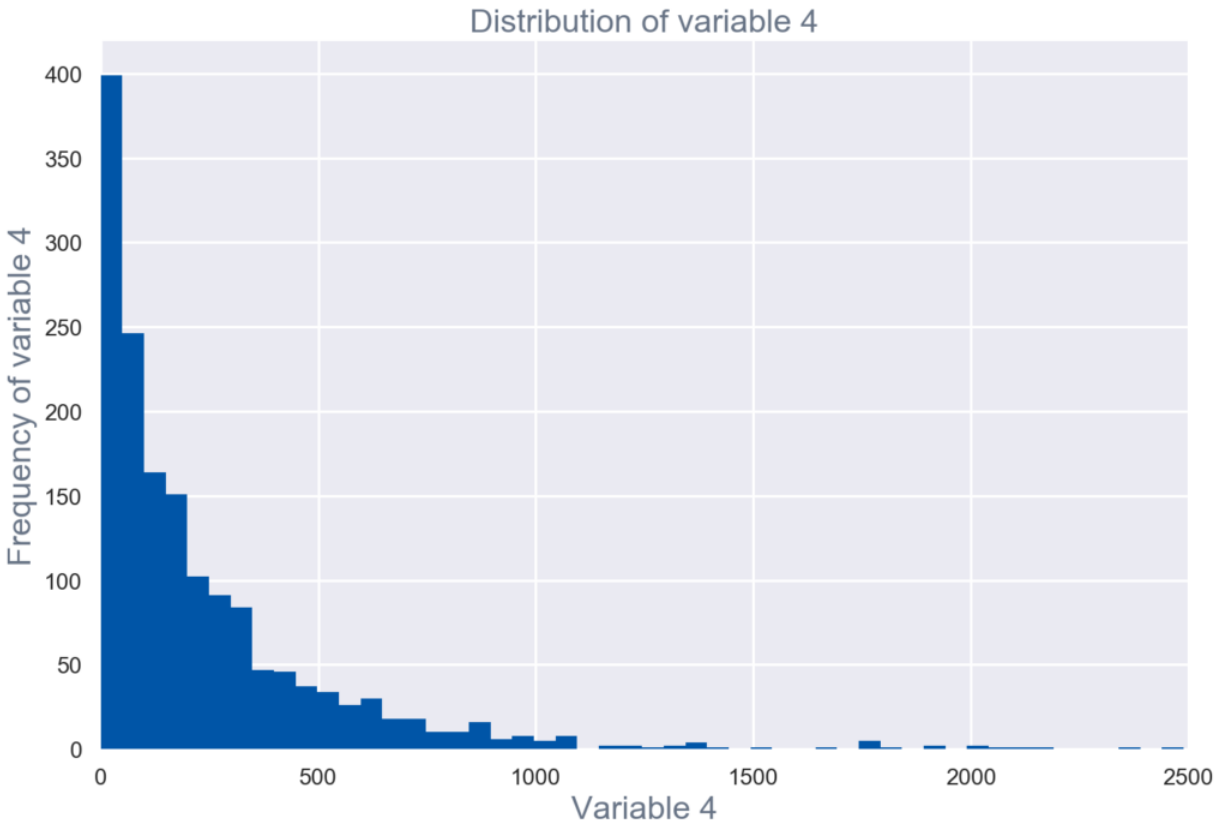


Figure 1

- Bivariate visualization and summary statistics for assessing the relationship between each variable in the dataset and the target variable of interest (e.g. time until churn, spend) (see figure 2)

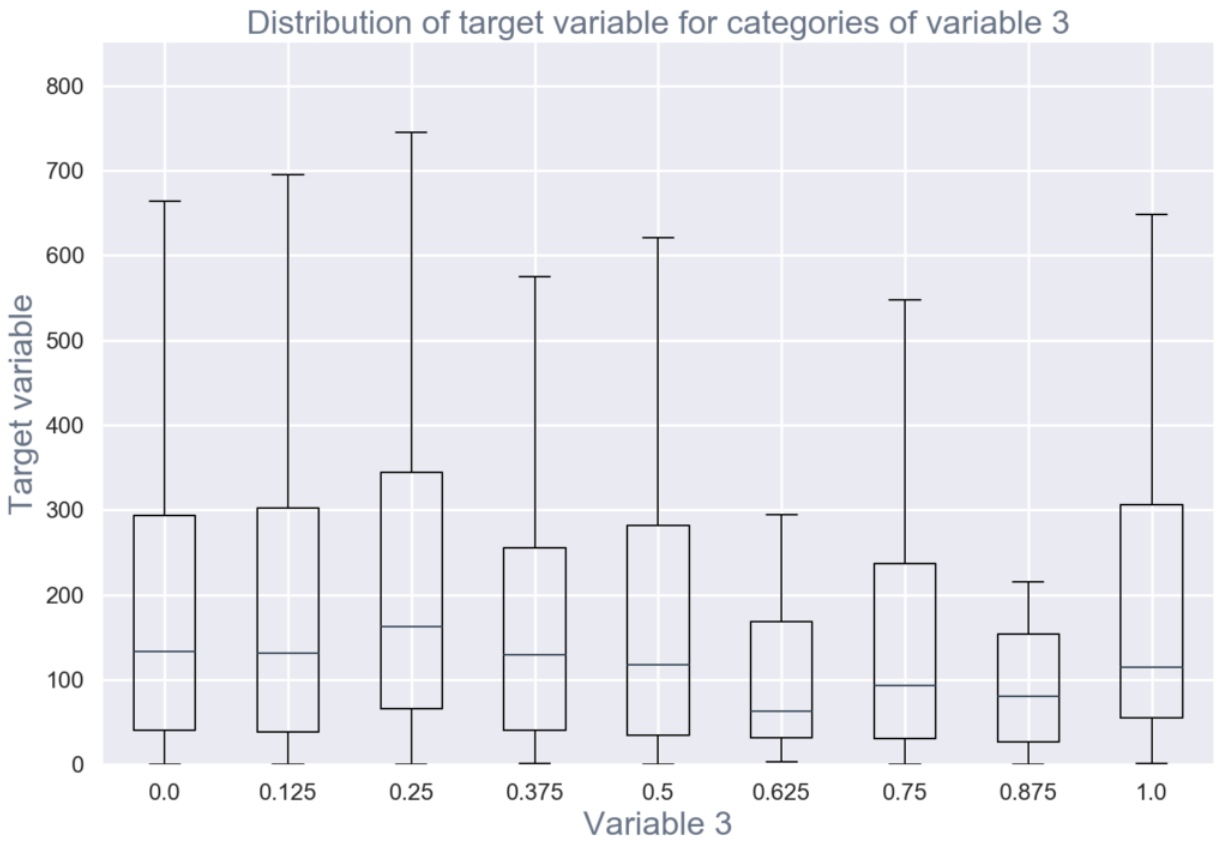


Figure 2

- Multivariate visualizations to understand interactions between different fields in the data (see figure 3).

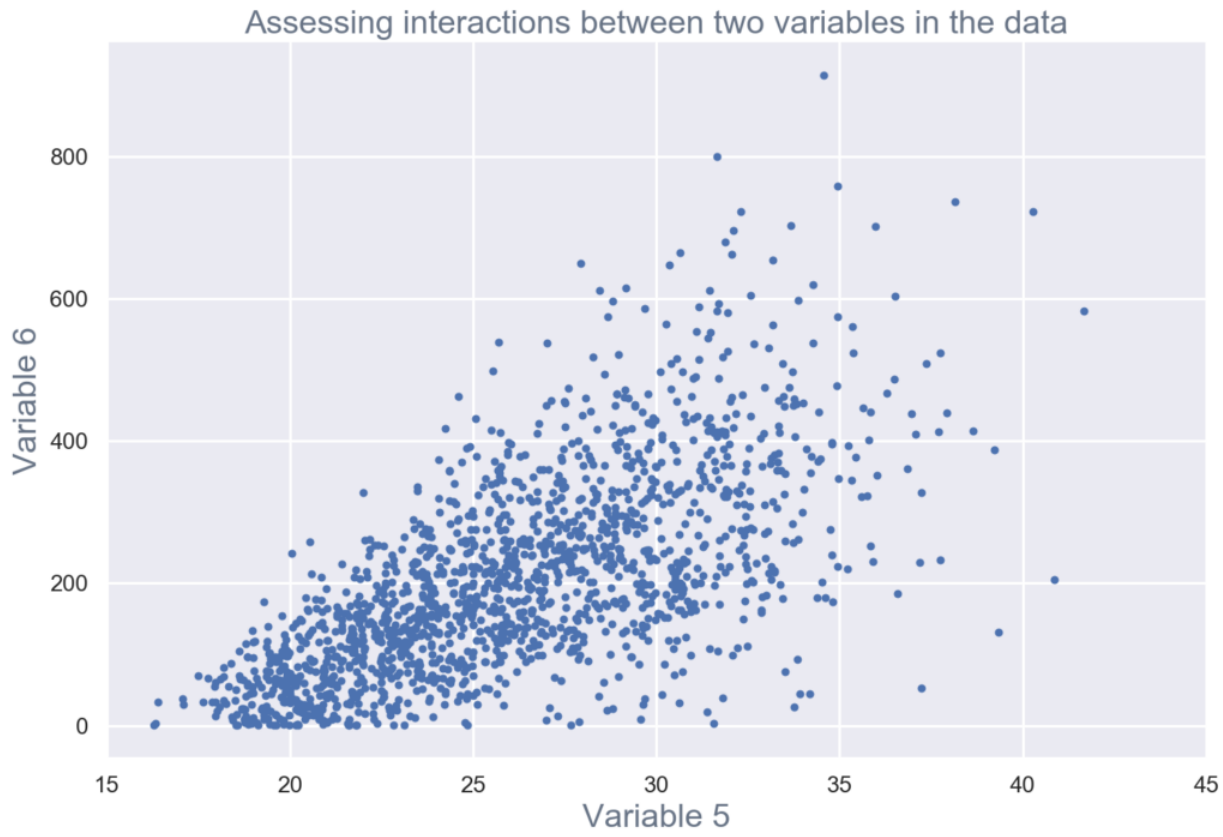


Figure 3

- Dimensionality reduction to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data
- Clustering of similar observations in the dataset into differentiated groupings, which by collapsing the data into a few small data points, patterns of behavior can be more easily identified (see figure 4)

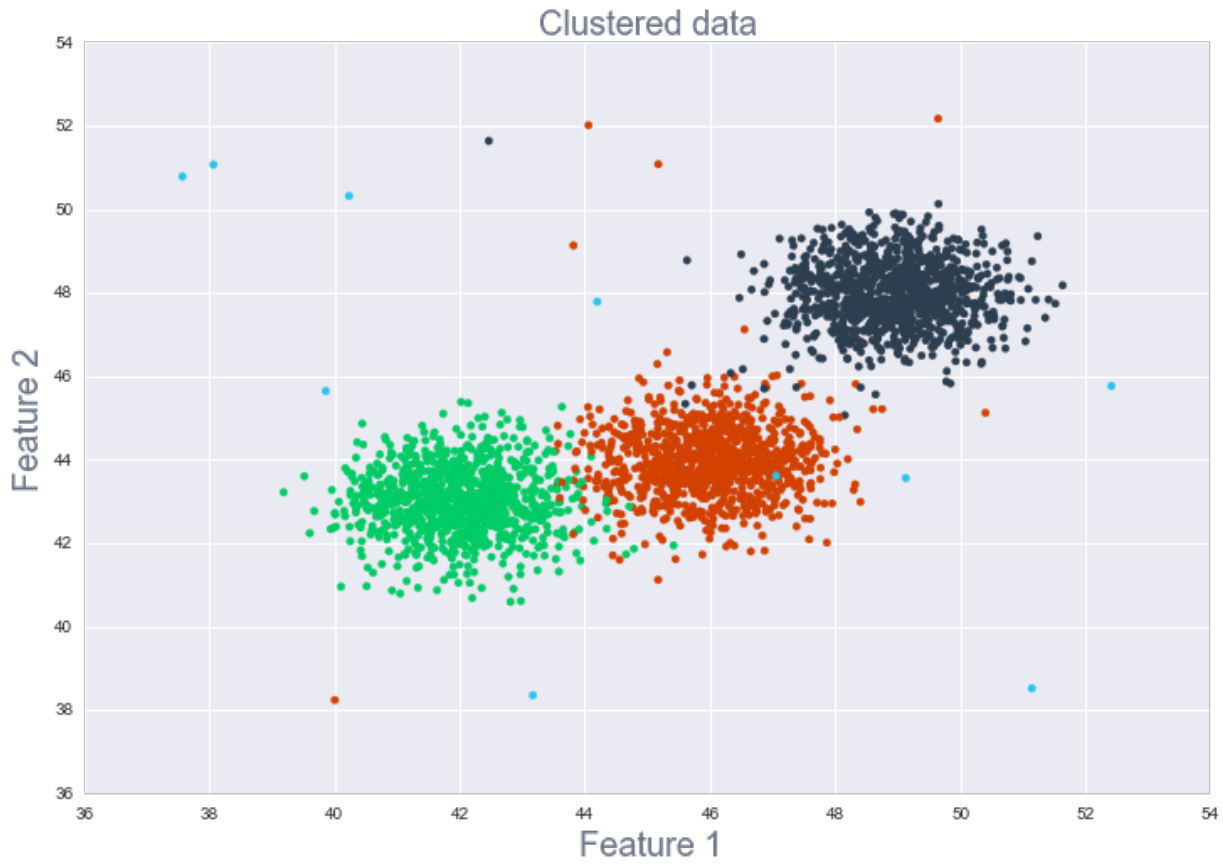


Figure 4

Through these methods, the data scientist validates assumptions and identifies patterns that will inform the understanding of the problem and model selection, builds an intuition for the data to ensure high quality analysis, and validates that the data has been generated in the way it was expected to.

Validating assumptions and identifying patterns

One of the main purposes of EDA is to look at the data before assuming anything about it. This is important, first, so that the data scientist can validate any assumptions that might have been made in framing the problem or that are necessary for using certain algorithms. Second, assumption-free exploration of the data can aid in the recognition of patterns and potential causes for observed behavior that could help answer the question at hand or inform modeling choices.

Often there are two types of assumptions that can affect the validity of analysis: technical and business. The proper use of certain analytical models and algorithms relies on specific technical assumptions being correct, such as no collinearity between variables, variance in the data being independent of the data's value, and whether data is missing or corrupted in some way. During EDA, various technical assumptions are assessed to help select the best model for the data and task at hand. Without such an assessment, a model could be used for which assumptions are violated, making the model no longer applicable to the data in question and potentially resulting in poor predictions and incorrect conclusions that could have negative effects for an organization. Furthermore, EDA helps during the feature engineering stage by suggesting relationships that might be more efficiently encoded when incorporated into a model

a model, the data scientist knows each type of assumption that must be valid for its use and can go about systematically checking them. Business assumptions, on the other hand, can be completely unrecognized and deeply entangled with the problem and how it is framed. Once, we were working with a client who was trying to understand how users interacted with their app and what interactions signaled likely churn. Deeply embedded in their framing of the problem was their assumption that their user base was composed of, say, experienced chefs looking to take their cooking to the next level with complex recipes. In fact, the user base was composed mostly of inexperienced users trying to find recipes for quick, easy-to-make meals. When we showed the client the assumption that they had been building their app upon was misinformed, they had to pivot and embark on understanding a whole new set of questions to inform future app development.

While validating these technical and business assumptions, the data scientist will be systematically assessing the contents of each data field and its interactions with other variables, especially the key metric representing behavior that the business wants to understand or predict (e.g. user lifetime, spend). Humans are natural pattern recognizers. By exhaustively visualizing the data in different ways and positioning those visualizations strategically together, data scientists can take advantage of their pattern recognition skills to identify potential causes for behavior, identify potentially problematic or spurious data points, and develop hypotheses to test that will inform their analysis and model development strategy.

Building an intuition for the data

There is also a less concrete reason for why EDA is a necessary step to take before more advanced modeling: data scientists need to become acquainted with the data first hand and develop an intuition for what is within it. This intuition is especially important for being able to quickly identify when things go wrong. If, during EDA, I plot user lifetime versus age and see that younger users tend to stay with a product longer, then, I would expect whatever model I build to have a term that would result in increased lifetime when age is decreased. If I train a model that shows different behavior, I would quickly realize that I should investigate what is happening and make sure I didn't make any mistakes. Without EDA, glaring problems with the data or mistakes in the implementation of a model can go unnoticed for too long and can potentially result in decisions being made on wrong information.

Validating that the data is what you think it is

In the days of Tukey-style EDA, the analyst was typically well aware of how the data they were analyzing was generated. However, now as organizations generate vast numbers of datasets internally as well as acquire third-party data, the analyst is typically far removed from the data generation process. If the data is not what you think it is, then your results could be poorly affected, or worse, misinterpreted and acted on.

One example of a way data generation can be misinterpreted and cause problems is when data is provided at the user level but is actually generated at a higher level of granularity (such as for the company, location, age group the observation is a part of). This situation results in data being the same for otherwise disparate users within a group.

Let's look at Company A's situation. Company A is trying to predict which of its users would subscribe to

first thought was unnecessary for achieving their desired results. The results show them that the subscribers being predicted for were part of larger corporate accounts who controlled what products their employees subscribed to. This control meant that users could look exactly the same in the data in every way but have different target outcomes, meaning that the individual-level data had little ability to inform predictions. Not only did EDA in this case expose technical problems with the approach taken so far but it also showed that the wrong question was being asked. If the subscriber's' behavior was controlled by its organization, there was no business use to targeting subscribers. The company needed to target and thus predict new product subscriptions for corporate accounts.

Other examples that we have seen where data generation process has been wrongfully assumed:

- Data is generated the same across versions of a product or across platforms.
- Data is timestamped according to X time zone or the same across time zones.
- Data is recorded for all activity but is only recorded when a user is signed in.
- Identifiers for users remain constant over time or identifiers are unique.

Bio: **Chloe Mawer** comes from a background in geophysics and hydrology, and is well-versed in leveraging data to make predictions and provide valuable insights. Her experience in both academic research and engineering makes her capable of tackling novel problems and creating practical, effective solutions.

Original. Reposted with permission.

Related:

- 5 Steps for Advanced Data Analysis using Visualization
- What makes a good data visualization – a Data Scientist perspective
- Bokeh Cheat Sheet: Data Visualization in Python

<= Previous post
Next post =>

Top Stories Past 30 Days

Most Popular

1. **Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data Science**
2. **Introduction to Time Series Analysis in Python**

Most Shared

1. **Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data Science**
2. **Introduction to Time Series Analysis in Python**

4. **Machine Learning from Scratch: Free Online Textbook**
5. **Autograd: The Best Machine Learning Library Youre Not Using?**
6. **How I Consistently Improve My Machine Learning Models From 80% to Over 90% Accuracy**
7. **The Best Free Data Science eBooks: 2020 Update**

4. **Machine Learning from Scratch: Free Online Textbook**
5. **The Best Free Data Science eBooks: 2020 Update**
6. **Deep Learning's Most Important Ideas**
7. **Online Certificates/Courses in AI, Data Science, Machine Learning from Top Universities**

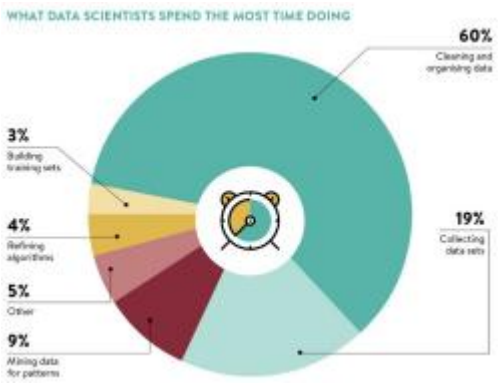
Latest News

- Cartoon: Cloud Dating
- DOE SMART Visualization Platform 1.5M Prize Challenge
- Optimizing the Levenshtein Distance for Measuring Text ...
- Deep Learning for Virtual Try On Clothes – Challe...
- Fast Gradient Boosting with CatBoost
- Machine Learning's Greatest Omission: Business Le...

Top Stories
Last Week

Most Popular

1. **A step-by-step guide for creating an authentic data science portfolio project**



2. **Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data Science**
3. **10 Best Machine Learning Courses in 2020**
4. **Strategies of Docker Images Optimization**
5. **The Best Free Data Science eBooks: 2020 Update**
6. **How LinkedIn Uses Machine Learning in its Recruiter Recommendation Systems**
7. **Introduction to Time Series Analysis in Python**

1. **10 Best Machine Learning Courses in 2020**
2. **Free Introductory Machine Learning Course From Amazon**
3. **How LinkedIn Uses Machine Learning in its Recruiter Recommendation Systems**
4. **A step-by-step guide for creating an authentic data science portfolio project**
5. **Annotated Machine Learning Research Papers**
6. **A Guide to Preparing OpenCV for Android**
7. **A step-by-step guide for creating an authentic data science portfolio project**

More Recent Stories

- Machine Learning's Greatest Omission: Business Leadership
- fastcore: An Underrated Python Library
- How to ace the data science coding challenge
- Text Mining with R: The Free eBook
- Top tweets, Oct 7-13: Every DataFrame Manipulation, Explain...
- Deep Learning Design Patterns
- Free From MIT: Intro to Computational Thinking and Data Science
- Goodhart's Law for Data Science and what happens when a meas...
- Getting Started with PyTorch
- KDnuggets 20:n39, Oct 14: A step-by-step guide for creating...
- Top September Stories: Free From MIT: Intro to Computer Scienc...
- SIAM launches activity group, publications for data scientists
- The Future of Fake News
- Software Engineering Tips and Best Practices for Data Science
- Uber Open Sources the Third Release of Ludwig, its Code-Free M...
- 5 Best Practices for Putting Machine Learning Models Into Prod...
- How to be a 10x data scientist
- Top Stories, Oct 5-11: A step-by-step guide for creating an au...
- Exploring The Brute Force K-Nearest Neighbors Algorithm
- Annotated Machine Learning Research Papers