Exploratory Data Analysis What is it

Matthew Cornelius

Bellevue University

Abstract

In the duration of this paper I will be asking the question what is exploratory data analysis? When asking this question there are also key elements that I wish to pose that fall under the umbrella of needed knowledge on the subject to better understanding. I will be asking why is exploratory data done in the first place? During the process of exploratory data science, are there any steps that can be skipped and still give the accurate results as the complete process? Is there a certain order to the madness of the exploratory data science? I will attempt to gain knowledge of who is the originator of the exploratory data analysis process. I know this may be a lot of information to take in, but my wish is that I have broken the subject matter down in simple enough terms that a general audience has some understanding of the content, and if interested may wish to pursue further research as need.

Exploratory data analysis also known as EDA is a component of statistics, that is often used to analyze data sets. This is done to summarize the data set that is being used, visual representations are often used in unison to give a clearer interpretation of what is to be conveyed. In theory EDA is able to tell what can’t be normally be seen in data by using more traditional modeling. So essentially this is the reason an individual would use exploratory data analysis in statistics to explain data as opposed to other methods.

An American mathematician by the name of John Tukey was the first to promote the exploratory data analysis methods. I found online a quote from John Tukey were he attempted to define what data analysis is, the website goes to say, “Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.”( Exploratory data analysis, n.d.).

A version of the steps that I have found in the exploratory data analysis process goes as followed, step one which also know as first approach. Step two which is analyzing categorical variables, step three analyzing numerical variable. And lastly step four which is analyzing numerical and categorial at the same time. Before we start I will state that I have found during my research that there are two different types of analysis results. What I read goes to define the two as followed, “They can be two: informative or operative. Informative - For example plots, or any long variable summary. We cannot filter data from it, but give us a lot of information at once. Most used on the **EDA** stage. Operative - The results can be used to take an action directly on the data workflow (for example, selecting any variables whose percentage of missing values are below 20%). Most used in the **Data Preparation** stage.” (Casas, 2018)

In step one also known as first approach to data, this is where you would look at the data and observe the rows and variables. By doing this your aim is to get the gist of what the overall data is about, you are looking for data types, zeros, infinite numbers, etc. When browsing the data in this manner there are a few key points to remember. You should be asking yourself, are the variables that you see in the needed data type and are there any that have an unusually high cardinality variable. Step two is analyzing the categorical variables, this is where you are trying to determine if the categories within the variable make sense. Step three which is analyzing numerical variables is used to determine if the there is any unbalance in the variables and to insure checks for outliers that may throw of the data set findings. And step four which is where you combine steps two and three is called analyzing numerical and categorical at the same time, here is where your main goals are to fact check for the min and max values, and check the overall distribution of the data. The four steps here are on a single version of the many that you could find when it comes to EDA, so there is always room for interpretation on what the steps are and how they should flow in priority.

References

Casas, P. (2018, August 24). Exploratory Data Analysis in R (introduction). Retrieved from

https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/.

Exploratory data analysis. (2019, November 14). Retrieved from

https://en.wikipedia.org/wiki/Exploratory\_data\_analysis.