I decided to write my paper about the EDA process because it is important for me to understand the foundation of this process and some of the history behind it.

**What is exploratory data analysis?**

According to IBM, Exploratory Data Analysis (EDA) is a method commonly “…used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.”

**Why does it need to be done and is it a step in the process that can be skipped?**

EDA should always be done as it helps to look at the data before making any assumptions. This assists in the early identification of potential data collection errors, missing data points, outliers, etc. It also helps identify variables and their characteristics before starting a modelling phase. The process should never be skipped

**Is there a particular order to when you do EDA or is it ever “really” done?**

Think Stats (Downey, 2014) summarizes the steps followed by his process as follows:

* Importing and cleaning the data sets is the first big effort after obtaining the data. It includes reading, cleaning, and transforming without compromising the integrity of the data.
* Single variable exploration: examining one variable at a time, looking for common statistical measures like mean, distributions, etc.
* Pair-wise explorations: analyzing pairs of variables at a time to identify correlations using scatter plots and linear fits, for example.
* Multivariate analysis: If apparent relationships are identified in the previous step, one can proceed to do regression to identify possible more complex relationships.
* Estimation and hypothesis testing: at this point, a study of effect of independent variables over depend variables can be analyzed, keeping in mind variability, validity, and reliability.
* Visualization: this is used to find relationships and effects between variables but also to communicate results.

These steps are not exclusive from one another, meaning that you can do visualizations while working on single variable exploration, for example. You could also go back to the first step and do some more cleaning before you move to the multivariate analysis. In this sense, EDA is considered as a way to begin analysis, but it probably never ends.

**Who started EDA and what are some examples of it being done pre-computer era?**

Exploratory Data Analysis was first developed by a Mathematician named John Tukey around the year 1970. Tukey defined data analysis in 1961 as: "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."

Many EDA ideas can be traced back to earlier authors, where some of the early calculations were done “by hand”:

* *Francis Galton* emphasized order statistics and quantiles very early in the 20th century.
* *Arthur Lyon Bowley, in* 1920, used precursors of the stemplot and five-number summary (Bowley actually used a "seven-figure summary", including the extremes, deciles and quartiles, along with the median he defines "the maximum and minimum, median, quartiles and two deciles" as the "seven positions".
* *Andrew Ehrenberg* articulated a philosophy of data reduction in the year 2000 in a book of the same name.
* The Open University course Statistics in Society (MDST 242), took the above ideas and merged them with Gottfried Noether's work, which introduced statistical inference via coin-tossing and the median test.

The introduction of computational tools like R, Python, SPSS, and many others have definitely facilitated the EDA process allowing statistical computations and plots to be generated in minutes.

**References:**

IBM Cloud Education. August 25, 2020. *Exploratory Data Analysis.* <https://www.ibm.com/cloud/learn/exploratory-data-analysis>

Cook, D. and Swayne, D.F. (with A. Buja, D. Temple Lang, H. Hofmann, H. Wickham, M. Lawrence). December 12, 2007. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi.* Springer.

Downey Allen B. *Think Stats: Exploratory Data Analysis*. October 2014. O’Reilly. Second Edition.

https://en.wikipedia.org/wiki/Exploratory\_data\_analysis#cite\_note-2