DS First Project

Exploratory Data Analysis

In the world of data analysis, gaining a deep understanding of the data at hand is essential before diving into any modeling or decision-making processes. This is where Exploratory Data Analysis (EDA) comes into play. EDA is a crucial step in the data analysis pipeline that allows us to uncover patterns, identify trends, detect anomalies, and gain valuable insights from the data.

EDA involves the application of various statistical and visualization techniques to examine and explore the characteristics of a dataset. By systematically examining the data, EDA aims to uncover hidden patterns, relationships, and underlying structures that may not be immediately apparent. It helps data scientists and analysts to better understand the nature of the data and make informed decisions based on their findings.

One of the primary objectives of EDA is to summarize the main characteristics of the data, such as its distribution, central tendencies, variability, and potential outliers. It provides a foundation for subsequent statistical modeling, hypothesis testing, and machine learning tasks. EDA is often considered a precursor to more advanced analyses as it helps researchers identify potential pitfalls, assess the quality of the data, and determine the appropriate methodologies to employ.

Exploratory Data Analysis employs a wide range of techniques and tools, including summary statistics, data visualization, data transformation, dimensionality reduction, and correlation analysis, among others. These methods help analysts gain an intuitive understanding of the data, identify interesting patterns or trends, and formulate hypotheses for further investigation.

Moreover, EDA allows researchers to validate assumptions, evaluate the suitability of data for specific analyses, and address potential data quality issues, such as missing values, data inconsistencies, or data entry errors. It helps in the process of data cleaning and preparation, ensuring that the subsequent analyses are conducted on reliable and accurate data.

In summary, Exploratory Data Analysis plays a fundamental role in the data analysis process by providing a comprehensive understanding of the data's characteristics and uncovering valuable insights. It serves as a powerful tool for data scientists, statisticians, and analysts to explore, visualize, and interpret the data before embarking on more advanced analyses. Through EDA, we can unlock the potential of data, make informed decisions, and extract actionable knowledge from complex datasets.

In this assignment, you will identify a dataset of interest and perform an exploratory analysis to better understand the shape & structure of the data, investigate initial questions, and develop preliminary insights & hypotheses. Your final submission will take the form of a report consisting of captioned visualizations that convey key insights gained during your analysis.

Step 1: Data Selection

First, you will pick a topic area of interest to you and find a dataset that can provide insights into that topic. To streamline the assignment, we've pre-selected a number of datasets for you to choose from.

However, if you would like to investigate a different topic and dataset, you are free to do so. If working with a self-selected dataset, please check with the course staff to ensure it is appropriate for the course. Be advised that data collection and preparation (also known as *data wrangling*) can be a very tedious and time-consuming process. Be sure you have sufficient time to conduct exploratory analysis, after preparing the data.

After selecting a topic and dataset – *but prior to analysis* – you should write down an initial set of **at least three questions** you'd like to investigate.

Part 2: Exploratory Visual Analysis

Next, you will perform an exploratory analysis of your dataset using a visualization tool such as Plotly (https://plotly.com/python/basic-charts/) . You should consider two different phases of exploration.

In the first phase, you should seek to *gain an overview* of the shape & stucture of your dataset. What variables does the dataset contain? How are they distributed? Are there

any notable data quality issues? Are there any surprising relationships among the variables? Be sure to also perform "sanity checks" for patterns you expect to see!

In the second phase, you should investigate your initial questions, as well as *any new questions* that arise during your exploration. For each question, start by creating a visualization that might provide a useful answer. Then refine the visualization (by adding additional variables, changing sorting or axis scales, filtering or subsetting data, *etc.*) to develop better perspectives, explore unexpected observations, or sanity check your assumptions. You should repeat this process for each of your questions, but feel free to revise your questions or branch off to explore new questions if the data warrants.

Recommended Data Sources

To get up and running quickly with this assignment, we recommend exploring one of the following provided datasets:

- World Bank Indicators, 1960–2017. The World Bank has tracked global human developed by indicators such as climate change, economy, education, environment, gender equality, health, and science and technology since 1960. The linked repository contains indicators that have been formatted to facilitate use with Tableau and other data visualization tools. However, you're also welcome to browse and use the original data by indicator or by country. Click on an indicator category or country to download the CSV file.
- Chicago Crimes, 2001–present (click Export to download a CSV file). This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
- Daily Weather in the U.S., 2017. This dataset contains daily U.S. weather measurements in 2017, provided by the NOAA Daily Global Historical Climatology Network. This data has been transformed: some weather stations with only sparse measurements have been filtered out. See the accompanying weather.txt for descriptions of each column.
- Social mobility in the U.S.. Raj Chetty's group at Harvard studies the factors that contribute to (or hinder) upward mobility in the United States (i.e., will our children

earn more than we will). Their work has been extensively featured in The New York Times. This page lists data from all of their papers, broken down by geographic level or by topic. We recommend downloading data in the CSV/Excel format, and encourage you to consider joining multiple datasets from the same paper (under the same heading on the page) for a sufficiently rich exploratory process.

- The Yelp Open Dataset provides information about businesses, user reviews, and more from Yelp's database. The data is split into separate files (*business*, *checkin*, *photos*, *review*, *tip*, and *user*), and is available in either JSON or SQL format. You might use this to investigate the distributions of scores on Yelp, look at how many reviews users typically leave, or look for regional trends about restaurants. Note that this is a large, structured dataset and you don't need to look at all of the data to answer interesting questions. In order to download the data you will need to enter your email and agree to Yelp's Dataset License.

General Notes

- One of the best resources to learn data manipulation with pandas (In my opinion):

Study guide - Data with Python

Teaching page of Afshine Amidi, Instructor at MIT.

https://www.mit.edu/~amidi/teaching/data-science-tools/study-guide/data-manipulation-with-python/?fbclid=IwAR2WJrvwWzNezX-CbSshx6DG-qHohrLODiLtgBmC6ityc-spkJp3eTX4dCE#main-concepts

However you are free to learn from any resource.

- For Visualization you are free to use any package but it’s good to learn plotly

Reasons Why Plotly Is The Best Visualization Library

Once you go Plotly, you can't go back Data Visualization is a technique used, in which the analyst is able to examine the data in a graphical format in order to obtain additional insights, regarding

https://www.linkedin.com/pulse/reasons-why-plotly-best-visualization-library-shripad-bhosale/

- On the journey of Exploratory Data Analysis (EDA), it's common to come across various notebooks or examples that showcase how others have worked with similar

datasets. While it can be beneficial to explore these resources and learn from the approaches of experienced analysts, it's crucial to remember that simply copying and pasting code without understanding it undermines your learning process

- Regarding submission file, you are requested to submit a notebook which must have a short conclusion for each step (under each piece of code) and general conclusion by the end of you work (an overall conclusion)
- **You are free to change the data you want to investigate however answering three questions is a must**