AI- LABS

OPEN-SOURCE TOOLS, TECHNOLOGIES, AND PLATFORMS

DATA EXPLORATION

Abstract

Instructions for a hackathon centred around problems using an open solution platform

AI TECH UK

20th April 2023

Table of Contents

Before you start

Think about what you hope to get out of this hackathon. Do you want to focus on just one technology and get the best model you possibly can, or do you want to understand all the technologies? Either is perfectly valid.
Remember learning is a journey not a destination, you can get a working model in a short period of time but what will you have learnt if you just stop there?

Overview of Problem Uses

There are four areas of AI that can be explored as part of the AI Tech UK's – open-source Hackathon offering coding in Python. These are:

01 Data Exploration
02 Machine Learning Algorithms and Applications
03 Computer Vision
04 Natural Language Processing

Each area will have a series of problems with associated notebooks and data files. For each area a structured approach is provided detailing the objective, dataset, algorithms and models and accompanying libraries, useful links, success criteria and in some cases what to do next (a question, a discussion, use of your own data, refer to a specific link/lab/task, or simply reading further resources of your choice).

01 Data Exploration

Introduction

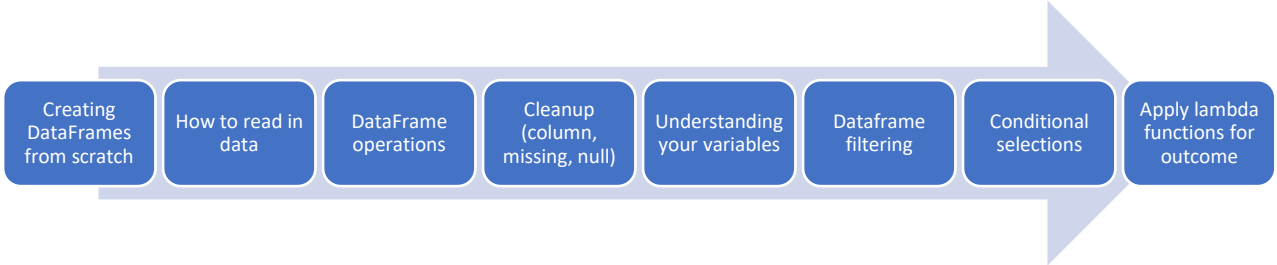
Data exploration is the initial step in data analysis, where users explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest.
There are a series of problems for data exploration starting off with some very simple data, wot working with files to and working through problems that have large dataset and more data issues. These include:

1. Simple Data Exploration Part I and Part II
2. Data exploration – Football Dataset
3. Data exploration – Automobile
4. Data exploration – Supermarket Analysis

1 Simple Data Exploration Part I and Part II

Objective

In data exploration- Part 1 and Part II we focus on the form of the data, and how to get it ready so that you can start cleaning it, transforming, and analysing it.
In Part 1 we work on the core components of pandas that are series and dataframes. You will use the Part1 Notebook file to complete the tasks of the Part II – starting from step 3 onwards – dataframe operations.



Approach

For example, say you want to explore a dataset stored in a CSV on your computer. Pandas will extract the data from that CSV into a DataFrame — a table, basically — then let you do things like:

- Calculate statistics and answer questions about the data, like
    - What's the average, median, max, or min of each column?
    - Does column A correlate with column B?
    - What does the distribution of data in column C look like?
- Clean the data by doing things like removing missing values and filtering rows or columns by some criteria
- Visualize the data with help from Matplotlib. Plot bars, lines, histograms, bubbles, and more.
- Store the cleaned, transformed data back into a CSV, other file, or database.

Dataset

For Part I you will use a dictionary to store your simple data of purchases of fruit. In Part II you will use the same dictionary data but from Purchases.csv - a simple Excel csv file with 3 columns and 4 rows. This is the same data in Part 1 but stored in a file, but you will read from this file. You will work with the IMDB-Movie-Data.csv file to carry out the dataframe operations, clean the data carry out some analysis based on conditional sections and applying functions.

Libraries

Library used is Pandas – a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool.

Algorithms and models

As this is a simple data exploration exercise – you will use carry out a range of dataframe operation using a range of methods to clean the data and then to filter against conditions using a lambda python function.

Success criteria.

1. To be able to create a dataframe from imported data from a file
2. To be able to carry out a range of dataframe operations to get a summary of the data
3. To understand the role of variables in working with the data to analyse it further.
4. To be able to filter or query data like a database but using python functions to retrieve results

Useful links

- The ultimate guide to cleaning data
- Data exploration tools and techniques

- Comprehensive guide to data exploration

What to do next

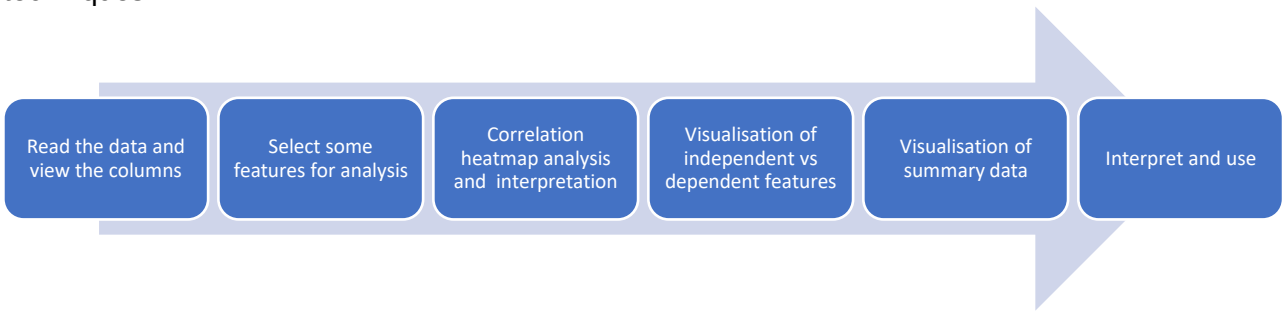
To review the useful links and explore data in problem domains with the other labs where their business rules to be followed to explore and make sense of the data.

2 Data exploration – Football Dataset

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check # assumptions with the help of summary statistics and graphical representations. We will use a range of libraries and techniques to explore data and consider the efficiency of getting the results. In this lab, some of the steps earlier will be grouped together.

Objective

To explore football match data to make financial decisions.
As football is a global multi-billionaire business making decisions on who should play for what the club, and matches is based on the players behaviour during the match. This data is based on text – video analysis will provide further explorative analysis and support decision making and find some reason why a player has the score they have. Here we will introduce a range of visualisation techniques.



Approach

Here descriptive summaries are created to understand the football data and focus on specific variables to analyse their relationship or correlation with other variables. Visualizations are used to present the data at various interim stages and results,.

Dataset

'Footballmatchdata' is a dataset with 89 columns of 18206 rows of data. This has a lot of data. It has information about football players, their personal and financial information, player behaviour and overall performance scoring and salary equivalence. As football is a global multi-billionaire business making decisions on who should play for what the club, matches and players

Libraries

Libraries used are:

- Pandas – a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool.
- NumPy - to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices, and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.
- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data.
- For efficiency **Pandas profiling** is an open source **Python** module with which we can quickly do an exploratory data analysis with just a few lines of code.

Algorithms and models

The technique of correlation is used to compare variables, to ascertain any relationship. Here Seaborn correlation heatmaps present the same information in a visually appealing way. What more: they show in a glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems.
Heatmap have various arguments one of which is the colour palette customise to show stronger correlation (normally darker) and weaker correlation are lighter shades (but this can vary). For example, you will see a score of 0.94 identifying a strong correlation between dribbling and ball control which is a known football fact, but the data also states this.
Use of visualization techniques such as cat plot, box plots and joint plots to view independent vs dependent variables. For example, to visualise the overall score of a player (independent) and their age (dependent) or the player position (dependent).

Success criteria

1. To be able to generate descriptive analytics and selective summaries of the data.
2. To utilise the correlation technique to create a starting point to select features for further analysis
3. To create visualization of various combinations of independent variables (sometimes summarised) vs dependent variable to discover patterns to spot anomalies and to create a plan for further analysis.

Useful links

Data Visualization in Python with matplotlib, Seaborn, and Bokeh

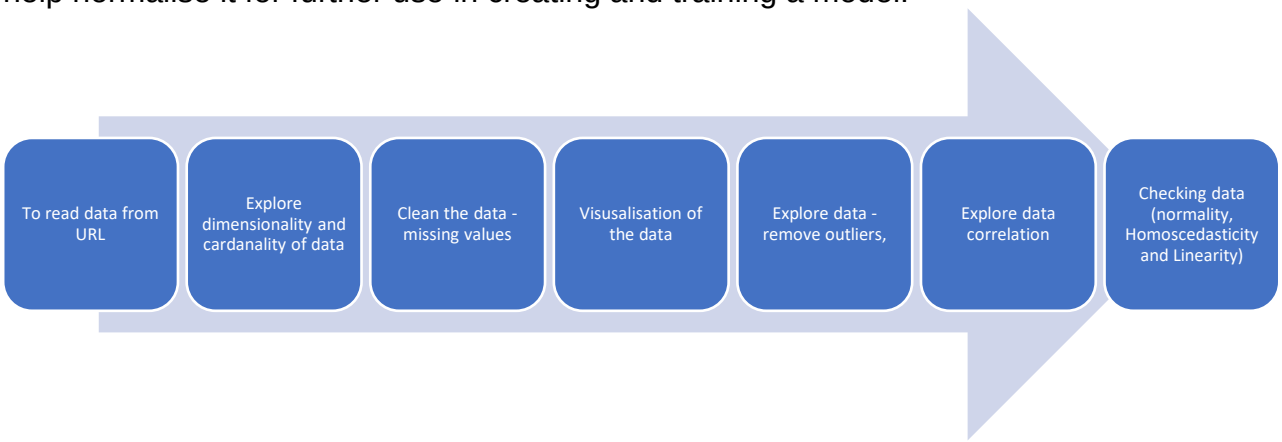
What to do next

To explore data based on problem domains and specific business rules using more techniques and specific algorithms before training the data.

3 Data exploration – Automobile Dataset

Objective

In this data exploration and visualization lab the goal is data discovery, from an automobile dataset which is used for several business functions from procurement to transactions to customer support. Here more refined statical techniques are used to understand the data, to help normalise it for further use in creating and training a model.



Approach

This utilises a pipeline of tasks to support the data exploration and preparation for the next stage of creating of models.

Dataset

The Auto MPG sample data set is a collection of 398 automobile records from 1970 to 1982. This not a large dataset but used for data discovery.
It contains attributes like car name, MPG, number of cylinders, horsepower, and weight.
This is a great sample data set to explore and visualize using python.

Libraries

- OS – a library to support working with files and their storage folders.
- Pandas – a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool.
- NumPy - to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices, and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.
- Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data.
- For efficiency **Pandas-profiling** is an open-source **Python** module with which we can quickly do an exploratory data analysis with just a few lines of code.
- SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics, and many other classes of problems.

Algorithms and models

Use of aggregation(mean, min, max) to group data. To create a range of custom functions to find missing and duplicate data. Use of visualization to view the data and establish data such as outliers that should be excluded using multivariate analysis. To visualise pair-to-pair relationships and correlation to analyse the data. The use of SciPy statistical methods to support the error rates and to normalise the individual features and grouped features for distribution visualisations using normality, normality, Homoscedasticity and Linearity.

Success criteria.

1. To be able to create custom functions to explore variables
2. To be able to clean the data using a range of techniques and appropriate for the business rules
3. To create visualisation of all the attributes
4. To create a correlation of features for analysis
5. Data exploration by excluding data (outliers) by multivariate analysis.
6. To explore and check/test the data using statistical functions to support distribution visualisations and error rates.

Useful links

- Missing values
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html

What to do next

- Other data quality issues such as: Inconsistent formats, Features importance, Misleading correlation, or data dependency; Data imbalance
- 9 data quality issues than can side-line AI projects
- Challenges for Data Governance and Data Quality in a Machine Learning Ecosystem

4 Data exploration – Supermarket Sales Analysis

This involves Supermarket sales exploration and analysis from different perspectives. The results from this are used to consolidate ideas for features analysis and answer specific questions. For example, 'does customer type influences the sales'.

Objective

Approach

This a pipeline of tasks to analyse the data from each perspective and then to establish key variables for feature analysis and prediction.

Dataset

Upload the dataset – 'Supermarket- sales' which containing information on sales, products, payment, and customer perspectives. Contains 1000 rows with 17 columns.

Libraries

- OS – a library to support working with files and their storage folders.
- Pandas – a fast, powerful, flexible, and easy to use open-source data analysis and manipulation tool. NumPy - to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices, and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices. Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data.

Algorithms and models

Here a range of techniques are used including aggregation, summaries, data manipulation based on object types, unique value selection, visualisation plots of both individual, group and pair-wise features and finally filtering data based on different level of data for example – customer type and sales.

Success criteria.

- To be able to create descriptive analytics and visualizations for each perspective with reference to the business rules and individual meta data of the variables
- To be able to create aggregations and visualisations to support specific queries
- To collate aggregate data and feature variable data to establish a query for analysis.

Useful links

Creating a Simple Recommender System in Python using Pandas

What to do next

To explore working on a classification, and clustering techniques on the data.