Data Fellowship: Data Analytics Hackathon

Introduction

It is time for the first Hackathon - 6 hours of data crunching and analysis in teams! It's a chance to put everything you've learnt so far into practice, experience the realities of data analytics, and test out new ideas in an experimental environment.

Task

You will be provided with several datasets from different sources that cover a range of topics. You are also welcome to bring your own data that you would like to analyse if your organisation's data-sharing rules permit that.

In the time provided, create a data product using some of the skills you have learnt/practised over the past months. You are free to choose the dataset, aspect, and methods that most interest you; you can even combine multiple datasets if you wish.

Guidelines

- It's a good idea to explore the data available early on; having a strong understanding of the dataset's structure and contents will be invaluable for modelling (if appropriate).
- All the datasets will need some level of data cleaning before modelling can begin. Some will be much cleaner than others.
- It's best to work in pairs/small groups.
- A hackathon isn't about creating a finished, polished product; it's about learning, exploring, and collaborating together.

Deliverables

At the end of the hackathon, you will be asked to prepare a 10-minute presentation to show your work to the rest of the group. You can use the STAR framework to structure your presentation.

Project Briefs and Datasets

We have chosen datasets that allow you to practise the skills covered so far. While the data used in the hackathons might not always be directly relevant to your day-to-day work, the skills, experience, and new ideas you garner from working with them will have wider applications.

You have two options to approach this challenge: In **Section A. (Project Briefs)**, there are five different project briefs you can choose from. You might need to explore potential data sources and identify datasets that are relevant for your analysis. You can modify the brief to better suit your data problem.

In **Section B. (Alternative Datasets)**, there are five dataset options. Feel free to explore the datasets and create your own data questions.

A. Project Briefs

1. The CMO for a travel company is looking to target 25-35-year-olds who are concerned about climate change. The company has set itself the goal of reducing air travel for customers by 37% in 2021. The executive sponsor has asked for your recommendation on the top 3 destinations for summer 2021, that does not involve a flight, based on
    - Average monthly temperature above 22 degrees Celsius
    - Average monthly rainfall less than 60mm
    - Less than 4hrs from London

    Potential datasets:
    NOAA
    Weatherbase

2. The CTO of an on-demand delivery startup based in Greater London is looking to open the company's first operational facility. The company uses a fleet of electric scooters to provide a "last-mile" delivery service to Fortune 500 companies in the city (e.g. important packages, laundry service, the CEO's sushi order). The executive sponsor has asked you to make a recommendation about where to open the first location, based on the following criteria:
    - Lowest rent
    - Within 20 minutes of the City and Canary Wharf
    - Low emission zone

    Potential datasets:
    Open Postcode Geo
    Commercial and Industrial Property Stats

3. The CEO of the Department of Health is putting together a spending recommendation for 2021. The department is particularly concerned about rising levels of obesity, access to nutrition and air quality across Greater London. You have been asked to put together a report analysing the top health issues per borough, based on the following criteria:
    - Which boroughs should receive more investment?
    - How would you prioritise spending across the city?
    - What health issues should the Dept. of Health focus on?

    Potential datasets:
    London Datastore

4. A travel company is concerned about the potential effect of climate change on future bookings. Their major concerns centre around changes in average/maximum temperatures as well as the amount of rainfall experienced. Their current top five destinations are:
    - Sydney, Australia
    - Venice, Italy
    - Paris, France
    - The Dead Sea, Jordan
    - Malmo, Sweden

    The COO has asked you to investigate whether these two types of climate change

will affect their ability to offer bookings to these destinations.

Potential datasets:
NOAA
Weatherbase

Extend your Project - Are there new territories to target, based on the climate changes identified?

5. The QA Officer of NHS England is concerned that the future provision of Mental Health services will be compromised due to the currently used, but under-estimating, formula:

   New funding = Old funding * 1.012

   You have been asked to determine which of the London boroughs are likely to require the largest funding increases over a five year period.

   Potential datasets:
   Prescribing data
   Mental Health Statistics

   Extend your Project - how would your recommendations differ if you were considering physical health?

B. Alternative Datasets

1. Olympics

   This dataset provides an opportunity to ask questions about how the Olympics have evolved over time, including questions about the participation and performance of athletes, different nations, and different sports and events.

2. Netflix Movies and TV Shows

   Some of the interesting questions which can be asked of this dataset
   a. Understanding what content is available in different countries
   b. Identifying similar content by matching text-based features
   c. Network analysis of Actors / Directors and find interesting insights
   d. Is Netflix increasingly focusing on TV rather than movies in recent years?

3. COVID-19 World Vaccination Progress

   Some potential questions to be asked:
   a. Which country is using what vaccine?
   b. In which country the vaccination programme is more advanced?
   c. Where are the most/least people vaccinated per day? In absolute terms, or as a proportion of the entire population?

4. The Museum of ModernArt (MoMA) Collection

   Artworks Dataset

   The Artworks dataset contains 138,151 records, representing all of the works that have been accessioned into MoMA's collection and catalogued in our database. It includes basic metadata for each work, including title, artist, date made, medium, dimensions, and date acquired by the Museum. Some of these records have incomplete information and are noted as "not Curator Approved."

   Artists Dataset

   The Artists dataset contains 15,222 records, representing all the artists who have work in MoMA's collection and have been catalogued in our database. It includes basic metadata for each artist, including name, nationality, gender, birth year, death year, Wiki QID, and Getty ULAN ID.

5. Chinook Database - SQL Database.

   The Chinook data model represents a digital media store, including tables for artists, albums, media tracks, invoices and customers. Media related data was created using real data from an iTunes Library. Customer and employee information was manually created using fictitious names, addresses that can be located on Google maps, and other well-formatted data (phone, fax, email, etc.). Sales information is auto-generated using random data for a four year period.

   You can access this database using pgadmin, similarly to the Iowa liquor database we used in Bootcamp. Please follow the instructions below:

   URL: http://delivery-pgadmin.multiverse.io/
   Username: pgadmin4@docker
   Password: pgadmin4