**General instructions:**

- This task is intended to be a general DS assessment. If you have applied for an Operational Research role, please contact your recruiter.
- Please, explain any step or though that you think may be important to evaluate your task.
- The expected programming language is **python**

- For the sake of the review, we **strongly prefer** to receive back a jupyter notebook containing all the code, comments and thoughts. This notebook should work from end to end, so we can `restart and run all` or go through it, cell by cell, if we needed to do so.

# TESCO STORES Dataset

At Tesco, the location of a retail store plays a huge role in its commercial success. Our Stores Team use various data sources to better understand the potential of candidate locations for new stores in the UK. They need data science help in designing a model that can predict the future sales **[normalised_sales]** of a store based on location characteristics. Your task is to examine the provided dataset and answer the questions below.

Dataset files

- `tesco-dataset/train.csv`
- `tesco-dataset/test.csv`

Columns

- `location_id`: id of Tesco property location
- `normalised_sales`: normalised sales value of Tesco store
- `crime_rate`: crime rate in the area
- `household_size`: mean household size in the area
- `household_affluency`: mean household affluency in the area
- `public_transport_dist`: index of public transport availability in the area
- `proportion_newbuilds`: proportion of newly built property in the area
- `property_value`: average property value in the area
- `commercial_property`: percentage of commercial properties in the area
- `school_proximity`: average school proximity in the area
- `transport_proximity`: proximity of different transport modes
- `new_store`: new Tesco store opened recently
- `proportion_nonretail`: proportion of non-retail commercial properties in the area
- `competitor_density`: density of competitor retailers
- `proportion_flats`: proportion of blocks of flats in the area
- `county`: county code of the area

# Q1

Before diving into the modelling, you are given the dataset and the Stores Team expect you to come back with an analysis of the data and any concerns you may have about it. They would also like to know which other information you think would be useful to collect for future developments.

# Q2

Build a model that can predict store sales based on the provided area features. Please show how you developed the model and report how well your model is performing. ***Constraint:*** Please use Random Forest as the model family to solve this problem.

# Q3

The dataset contains a test set of potential store locations. Use your developed model to predict the sales value in these areas and explain what recommendations you would give to the Stores Team to use it. Use any tools that may help you to share your findings with product owners and other non-technical decision makers in the team. Complete this task by explaining how you would improve the current results.

---

# Masked Dataset

You are given the following small dataset, which has been completely masked for privacy reasons. Please train the best model you can come up with to predict the target variable `y` based on the features `x1` and `x2`. Explain every step you take.

Assuming that this model will be used for making decisions involving important sums of money, provide any comments that you think you should be giving to the business as a technical expert.

Dataset files

- `masked_dataset/train.csv`