# Data Engineer Interview Task

You should have been given zip file containing two folders:

`/policies`
~10k jsons of UK and Swedish policies that were extracted from a Mongo database. They contain information about one policy each.

`/supporting_docs`
An excel of the finance teams targets for UK and Sweden.

# Tasks

In answering the below you're free to work locally or in a cloud environment. If you want to work on GCP we can give you access to a new project, just let us know. Please spend only a few hours on the task, we are not looking for a complete solution/answer.

## 1. Process input files into CSVs

Extract the data into CSV files to act as a toy data warehouse. Please create at least `policies.csv` and `pets.csv`. Some relevant fields in the jsons are pointed out below but include anything you need to answer the reporting questions in Task 2.

**For discussion at interview:**

- How do you connect the tables?
- In a cloud environment what storage method and structure/layout would you use for data like this?

**Dealing with the finance team's Excel file**
They update the sales targets roughly once a month and the file may change format in between. They're nice people and amenable to pasting the targets into a different tool or file format. But don't expect them to do any coding or anything too technical.

**For discussion at interview:** How could we enable them to update the targets in a cloud data warehouse themselves?

## 2. Reporting

Using the CSVs you've created analyse the data to create the metrics below. Provide an output of charts, tables etc. as you see fit.

- Compare sales value by month against finance targets. Split by country.
- Were UK sales above or below target in July? What was the difference?
- What is the most common cat breed overall?
- What was the most common dog breed sold in UK in May and how many of that breed where there?
- Sweden average order value in October was expected to be around 4000 Krona. Why was it noticeably different?

**For discussion at interview:** What tool would you use so business users can easily view dashboards of metrics like these (sales, product mix, pet info, etc)?

## 3. Processing new data and updating the metrics

**For discussion at interview:** In production new or updated jsons are added to a cloud storage bucket every 10 minutes. We need to be able to provide the updates on the metrics to business every few hours during the day.

- How can the newly added data be processed?
- How can the dashboards the business use be regularly updated?

## 4. Any other insights or ideas?

Feel free to surprise us with anything else you think is interesting relating to this data or its manipulation or how you visualise it.

## Supporting Information

Useful fields in the mongo jsons and some explanations:

```
{
    "account_ref": "BBM",  # will be "BBM for UK or "BBM-SE" f
or Swedish polices
```

```
    "created_at": 1564677846000,  # When policy was purchased. UTC timestamp in milliseconds.
    "data": {
        "insured_entities": [{}],  # Pets on a policy
        "policy_holders": [{}],  # Owner information
        "sales_channel": "phone"  # phone or web sale
    },
    "products": [
        {
            "name": "MoneyBack",  # Name of product sold
            "price": {
                "annual": {
                    "amount": 46921  # Cost of policy in pennies in local currency. Eg as the account_ref for this one is "BBM", this policy cost 469.21 GBP.
                }
            }
        }
    ],
    "status": "ON_RISK",  # Current status. Becomes ON_RISK straight after purchase.
    "uuid": "0002af32-aadb-43f2-b2d6-a4d5ebf2bde3"  # Unique ID for policy
}
```