Project Overview

You just joined the data department of the London Borough of Camden Council as a junior data analyst.

Your employer's logo

The Parks and Open Spaces Office within the Council have approached the Data Department and identified some data that they think will support upcoming projects as part of the council’s new data initiative. The data includes a list of trees in the borough. They hope that this data will support the following projects:

1. Public Tree Data
   Provide a list of all trees in the borough, which can be downloaded by the public for free from the Camden Council website. The Borough of Camden is committed to the “opendata” initiative and are interested in opening their data as much as possible to its citizens and to the world.

2. Tree Walks Brochures
   Create a series of “Tree Walks” brochures, each one of which is a guided walk around an individual park showing a map of the location of interesting trees.

3. Environment Report
   Put up a single page on the Camden Council website showing the total carbon and pollution benefit provided by all their trees. This will also show information about trees removed, trees planted and the net carbon and pollution impact of this activity.

The team are unsure as to whether the data they have can be used to deliver these projects. Your job is to evaluate the data and determine the extent to which this is possible.

You receive the following email from Astrid, the data manager in your department.

From: Astrid, Data Manager
To: You

Good Morning,

I hope you are settling well in your new role. As you know, the council's new data initiative is big news for our data team! I'm really interested in moving quickly to show our enthusiasm!

I've gathered some data that I think will do the job for us. Please take a look at it and let me know if it can. I have three data files:

1. Trees Data
   A list of all the trees we have in the borough (both their location and types). This is an Excel file which I downloaded from the Council website.

2. Environmental Data
   An extract from our council assets database which lists all the trees and gives some environmental data about each tree. This is a csv file.

3. Common Names Data
   A list of scientific tree names and their matching common names. I got this from a horticultural website and it's in json format. I had to write some code to scrape it from their website. I hope they don't mind!

I've attached the data.

Please record your work in a Jupyter notebook so that your work is maintainable and reproducible by people who will work on this project after you.

Shout if you need anything.

Thanks!
Astrid.

Attachments:

- tree_common_names.json
- camden_trees.xlsx
- camden_trees_environmental.csv

Task 1: Document the Data Sources

After reading the email, you download the files and sensibly think that the best place to start is to list some of the key details of the data sources you have available. These data files are, after all, going to be the focus of your work. On the company document repository you find a template called Data Source List and it looks to be just the template for the job! You realize that the information requested in the template is really important to build awareness and clarity around what data you have, where it comes from and the extent to which you will be permitted to use it.

You therefore read Astrid's email again closely to use the information she provided to complete the Data Source List.

At the end of task 1, you should:

- Have completed the data source list to help you understand where the data comes from and if there are any problems or limitations with its usage.

Task 2: Load Data and Perform Initial Exploration

Lucas, your teammate who sits next to you, looks over your shoulder and asks you what you're working on. You tell him, and he suggests you make an initial exploration of the data to get to know it a bit better. He explains that you won't be able to provide the information Astrid needs if you don't! The more you understand, the more success you will have with your analysis.

Lucas decides to help you by creating a template Jupyter Notebook for you. It is heavily guided and includes all of the steps to go through. He has even done the first few steps for you as an example.

He sends it to you, asks you to open it, rename it, and then get to work!

You can use Google Colab to open the template Jupyter Notebook. Colab is a fully online Jupyter Notebook environment. Take a look at this video to learn more and read these quick-start notes. Alternatively, if you prefer, you can install Anaconda on your computer as described in the "Learn Python Basics for Data Analysis" course.
At the end of task 2, you should:

- Have a Jupyter Notebook up and running.
- Know the overall size of the data sets and the data types of the columns.

Task 3: Further Inspect the Data Sets

After completing task 2 in the Jupyter Notebook, you go back to Lucas and explain that while you better understand the data, you still don't feel *intimate* with it.

> Lucas: "Ah, that was just the first step. You really need to take a look at each column and understand the values that it contains. We sometimes call them variables or features."

He encourages you to move on to the next steps in the Jupyter Notebook and continue inspecting the data! He sends you this article to help you understand the different variable types.

Later that day, Astrid stops by your desk and congratulates you on getting on top of the data so quickly. In your head, you thank Lucas for his help!

> Astrid: Great! I see you have started exploring the data. I'm a bit worried that some of it isn't up-to-scratch. Can you take a look? Sorry, got to run, I'm late for a meeting, I'll check on you later.

Astrid sprints out, leaving you to figure out what she means by not “up-to-scratch”. Surely, the data is all there and should be accurate. Otherwise, why would she have shared it? □ Thankfully, Lucas steps in and shares some more of his wisdom!

Lucas: Don't worry! It’s normal to dig around in the data set and see how reliable it is as a first step. We always do this. It seems that the first thing to do is to look at missing values, outliers and duplicates.

Luckily, I was pretty sure Astrid would want you to do this, and the next steps in the Jupyter Notebook show you how. You’ll need to start by looking at where there are nulls and 0 values in the data. Then do some boxplots to find out if there are any outliers. Finally, for the duplicates, try to find an ID column in the data and see if the IDs occur more than once.

And it’s best to get in the habit of noting anomalies and other observations like this. I recommend that you log each anomaly you find throughout the project and how you found it. I know I sometimes skip over this in order to dive into the data set and get some results, but we often get asked if there were any issues, and you don’t want to be caught off guard as I have been in the past.

At the end of task 3, you should:

- Be aware of the actual values within each column.
- Be aware of the types of variables you are dealing with.
- Feel more intimate with the data!

Task 4: Identify Missing Values

Lucas helps you get started on your first verification: missing values! He further explains more about them:

> Lucas: Missing values can indicate data quality issues. Maybe the data got lost or deleted, or perhaps it wasn’t there in the first place. These missing values might give you problems when you perform certain data analyses. If there are too many missing values, you won’t be able to draw conclusions confidently. A null value is a missing value in a column. And sometimes (but not always) 0 values are also effectively null. For example, a tree with 0 height is incorrect, so you can treat it as missing. Again, this is standard practice, so that’s the next step in the Jupyter Notebook. Keep up the good work!

By the end of task 4, you should:

- Know which columns have missing values and how big the problem is.

Task 5: Identify Outliers in the Trees Dimensions

You go back to Lucas and ask him about the second verification: outliers. You need clarification on using boxplots to identify outliers. And, for that matter, what exactly are outliers? □

> Lucas: Outliers are values that are so unusual they could be incorrect. But you have to look at them in the context of your understanding of the data. For example, a 100-meter tall tree would make sense in California, but not in Camden! I’ll send you some documentation that helped me understand how boxplots can show outliers. Give it a read.

He sends this link and reassures you that this is the next task in the Jupyter Notebook he created. However, this time he hasn’t given you any example code, and you realize that it’s up to you to work out how to do it based on the documentation links provided!

By the end of task 5, you should:

- Know if there are any outliers in the tree heights, spreads or diameters.

Task 6: Identify Duplicates in the Trees Data Set

Feeling pleased with yourself, you return to Lucas, asking for a bit of clarification on the third and final verification: spotting duplicates in data.

> Lucas: Well, duplicates mean we’ve somehow got the same data twice in a data set. It could be because someone typed it in twice, or maybe there was another human or technical error. Either way, we need to look for values that should be unique but aren’t. Look for an ID column (it might be called trees_id or identifier or something). Usually, this should be unique in the main data set. So see if you have the same value twice by counting the number of occurrences of each value in that column. If so, you have duplicates! Keep following the instructions in the notebook!

By the end of task 6, you should:

- Know how many duplicate trees you have.

Task 7: Identify Geolocation Issues

The following day, as you arrive at work, you receive another email from Astrid.

From: Astrid
To: you

Hello,

Sorry I had to rush off yesterday.

As you have seen, the data set has several variables that relate to the localization of each of the trees. We can work with latitude and longitude or with easting and northing coordinates.

> If you are not familiar with easting and northing coordinates, please take the time to read this Wikipedia article. The easting and northing system of coordinates is very convenient. It allows you to plot all the trees with a simple scatter plot.
>
> That said, we have been told that some trees in the data set are actually not located in Camden! Can you please identify them?
>
> Thanks!
> Astrid

Once again, Lucas came to the rescue and predicted that this would come up. All of the tasks under “task 7” in the Jupyter Notebook include precise tasks to meet Astrid’s request.

Finally, he sends you one additional piece of advice:

Lucas: Don't forget to write down your decisions in the Jupyter Notebook and results along the way; they are bound to come in handy if someone asks you to explain how you got to certain results. If you are anything like me, you might forget a few things.

By the end of task 7, you should:

- Have a list of trees that are not actually in Camden.

Task 8: Identify Unmatched Data

Later that day, you meet with Astrid. She explains that it's crucial that a data analyst not lose sight of the actual requirements. And she confesses that she might have been guilty of that herself! She reminds you (and herself) of two of the two initiatives that are still worrying her:

- The Tree Walks Brochures
- The Environment Report

Astrid: My concern is that the data we have won't be sufficient to deliver on these initiatives. The environmental data should match with the trees data, but I haven't checked. Also, the common names data was scraped from a website, so I don't know whether they have the same scientific names. Can you please run some checks?

You notice that Lucas is heading out for lunch, but before you panic, you check the next steps in the Jupyter Notebook and realize that you are all set up to run final checks!

By the end of task 8, you should:

- Identify any mismatches between the trees and environmental data set.
- Identify any mismatches between the trees and common trees data set.

Task 9: Report back on your process

As you leave at the end of the day, full of ideas of other things you could further explore in that data set, you bump into Astrid, who is about to leave.

Astrid: I hope you had a good day! Can you please send me what you have done so I can use it for the data team meeting tomorrow afternoon? I need your Data Quality Report and also a table showing how the data maps to the three requirements. I'll send you a template.

And with that, she quickly forwards the Data Requirements template, and she's gone! Phew, you think to yourself, Lucas was right; you would be asked about decisions you made related to quality!

It's a bit late to start now, but you make a note of these new two documents Astrid would like you to work on to get the ball rolling first thing tomorrow!

Here are a few tips to help you create these two documents.

- For the Data Quality Report: save the Jupyter Notebook as an IPYNB file in its executed form. This will form the report. Remember that the notebook must be executable in one try, and make sure that all your analyses and modifications are well documented in markdown cells. Highlight each issue you find with a clear comment.

- For the data requirements: use the information you gained from your analysis to complete the provided template. When considering requirements for any data project, understanding what data is used to satisfy what requirements will help identify gaps and provide vital information required to streamline or automate the data acquisition task. There is some guidance in the document to help you.

Astrid gets back to you and has follow-up questions about how you did your work. She realizes that Lucas spent some time training you and wants you to document certain steps for new data analysts joining the team!

She asks you to create a series of slides, which answer questions about data analysis lifecycle, data requirements and data quality. She sends you them in an email.

At the end of task 9 you should have:

- Created your presentation which includes at least one slide for each of the topics in Astrid's email.

Good luck!

Deliverables

You will need to deliver:

1. The Jupyter Notebook, which will also serve as the Data Quality Report
2. Data Source List
3. Data Requirements
4. A project review presentation that you could use to brief a junior data analyst at a later stage

To make it easier for your work to be reviewed by the jury, upload all the project deliverables to the platform in a zip folder called 'Project_title_LastName_FirstName'. Use the following naming convention for each of your deliverable: LastName_FirstName_number of deliverable_name of deliverable__your start date of project. This is how it should be named:

- LastName_FirstName_1_jupyternotebook_mmyyyy
- LastName_FirstName_1_datasourcelist_mmyyyy
- etc.

For example, the deliverable here would be named: Smith_Mary_1_jupyternotebook_042022

Project Presentation

Your oral presentation will last 30 minutes. During the oral presentation, your assessor will play the role of Astrid. The assessor will challenge your decisions, so be prepared to defend your work. The presentation will be structured as follows:

- Presentation of Jupyter Notebook, Data Source List and Data Requirements (10 minutes)
    - Walk through the deliverables you have prepared and explain your findings.
- Presentation of project review to Astrid (5 minutes)
    - Explain your approach.
- Discussion (10 minutes)
    - Playing the role of Astrid, the assessor will ask you questions about your methodology and your deliverables, for example:

        - What are the key data issues that could impact our ability to use this data for the three initiatives?
        - How big is the problem?
        - What actions can we take to improve our data?
- Debrief (5 minutes)
    - At the end of the session, the assessor will stop playing the role of Astrid so that you can debrief together.

Your presentation should last 15 minutes (+/- 5 minutes). Respecting presentation time requirements is important in professional environments. In consequence, if your presentation is under 10 minutes or over 20 minutes, you may be asked to redo the assessment.

Skills

- Describe principles of data, including open, public, administrative and research data
- Identify data sources and the risks, challenges to combination
- Apply principles of data classification within data analysis activity
- Identify and escalate quality risks in data analysis with suggested mitigation/resolutions
- Undertake customer requirements analysis and implement findings in planning and outputs
- Describe the data life cycle and the steps involved in carrying out routine tasks