**Data Science Recruiting Assessment**

**Introduction**

Last week, we got a call from an existing client. They have a new dataset and want to know if it can help them understand their current data a bit better. The dataset contains information surrounding airport delays, and if the information proves valuable, potential patterns in the data can help our client decide which airports to avoid and which carriers to book for future travel.

The data came to us from the client in JSON format, but some fields did not correctly load. We have to figure out how to load in the data, sift through it, and summarize what we think is important for the client to know. Below are some prompts for addressing the problem at hand, but the client also trusts us to come up with new ideas and summary options for their dashboard. This is where we have to present the most important information while also bringing creativity to the table.

You have 10 Days to send back the results (and code) via email (if the tenth day falls on a weekend, the next business day is adequate). Note that you will be evaluated on your code quality which includes efficient code, good commenting practices, and anything you deem necessary to have as part of the good development process. If you get called to a different project or a new team member starts, your team needs to be able to understand where you are and what you have found so far.

You may use any coding language to complete this task, but at minimum you should be able to use Python to 1) load in this format of data and 2) extract information based on desired key or value. If using a language other than Python for the rest of the task, please include a Python script with examples of these two tasks in addition to the main task. Additionally, you do not have to use one language for the entirety of the task.

Use the prompts below as a guide but use your imagination. What can we find out about this dataset that will help our client? Write up your findings as if you will present them to the client directly and be prepared to explain in detail the code and methods you used.

Don’t hesitate to email me ([orly.olbum@inpleo.com](mailto:orly.olbum@inpleo.com)) with any questions while you are working on the assessment!

**First Glance**

1. How many airports are present in this dataset?
2. What is the time range of the dataset?
3. Choose two fields that should be numeric and find the min, max, mean, and median across the whole dataset.
4. Choose a field that should be numeric and plot the value over time.
5. Create a frequency table for how many of each airline are hosted at each airport in the dataset.
   1. Which airport has the most carriers in this dataset?
   2. Which airport has the least?
   3. How many of the airports in this dataset are host to ‘Northwest Airlines’?
6. Which other fields might be of interest to the client?
   1. How do they contribute to solving the problem outlined above?
   2. What types of statistics might you want to calculate to describe these fields to the client?

**Data Manipulation**

1. Create a dictionary that has ‘Airport/Code’ as the keys and ‘Carriers/Names’ and ‘Carriers/Total’ as values; export this dictionary as a json file.
2. Create a dataframe that has Airport, ‘Label’ (year/month), and ‘# of Delays’ (‘Late Aircraft’) as the values.
   1. Plot the # of Delays due to late aircraft over time.
   2. Which year/month had the most delays due to late aircraft? How about the least?
   3. Which airport had the most delays all of 2010?
3. Using the excel file ‘airline-codes’, make a list of the airlines present in this dataset along with their corresponding codes.

**Analysis**

1. Determine if there is a significant difference between the average number of flights delayed each month and the average number of flights cancelled each month.
2. Extract all fields in the ‘Minutes Delayed’ section by month.
   1. Split the year into four seasons: Dec-Feb, March-May, June-Aug, Sept-Nov.
   2. Is there a relationship between minutes delayed (by category) and season?
3. If a flight at ATL has a delay, what is the probability that the delay is due to…
   1. Weather?
   2. Late Aircraft?