DATA SCIENCE CHALLENGE

1.1 OVERVIEW

Thanks for your interest in Solution BI's Data Science internship . We've put together a data science problem to test your abilities on real data. We expect you to devote 2 days dedicated to complete this task. Please return your solution within 2 days of receiving this document. If you use any non-standard libraries, please point to public repositories or package them with your submission so that we can install them. Your submission should include a detailed report explaining your reasoning and any assumptions that you have made in the process of solving the problem in addition to the python source code and the results output files.

1.2 DATA

flight.csv – File attached

> The file contains about 112000 lines and 4 comma-separated columns. The first line of the file is a header. The first column is a unique flight identifier. Each line after the header contains a date followed by flight arrival information, including number of passengers on board and flight duration (in hours). The data are available for one and only destination.

weather.csv – File attached

> The file contains about 7600 lines and 4 comma-separated columns. The first line of the file is a header. Each line after the header contains a date followed by weather information, including mean temperature (in degrees Celsius), maximum wind speed (in m/s) and total precipitation (in mm). The weather is available solely for the only destination city.

Dates are represented in POSIX strftime format "%Y-%m-%d" and range from January 1, 1995 ("1995-01-01") to January 1, 2016 ("2016-01-01").

1.3 PROBLEM STATEMENT

This challenge is composed of four parts.

Using the data in flight.csv and weather.csv, build a new data set in which entries extend the information in flight.csv with weather information.

1. In the resulting dataset, identify and remove entries with outliers in flight duration, flight occupancy or weather data.
2. Using the resulting data set of step 2, predict the number of passengers travelling for the entries which have missing `number_passengers` column. The output must be written in the same comma-separated line format as flights.csv.
3. Using the resulting data set of step 2, and the aircraft specifications in Table 1, what fleet (number of aircrafts of each type) would you recommend to the airline servicing these flights? *Note: This question admits different answers. Use your logic, analytics skills and creativity to provide a reasonable recommendation.*

Table 1 - Aircraft Specifications (see item 3)

Aircraft Type	Aircraft Specification	
	Maximum Number of Passengers	Flight Autonomy (in hours)
A	100	1.5
B	50	2.5
C	150	6.0

Please provide any code you created to complete this task and also a report detailing:

- Any data preprocessing and exploratory data analysis you may have performed.
- Any assumptions you have made.
- Your methodologies and results.

1.4 Evaluation

We are giving you this challenge to help us understand how you approach new problems. Our expectations for this challenge are calibrated by your level of experience. We are most interested in how you think, so please carefully explain the choices you make in solving the problem.

In evaluating your solution, we are looking for the following:

1. **Communication skills**: Are the models clearly explained?
2. **Appropriateness of methods**: Are the methods used appropriate for the problems?
3. **Coding ability**: Is the code clear, organized and well-documented?

Your code and report are only used for assessing how well suited you are for this role.

We hope you enjoy this exercise! We look forward to reading your report.