Global
Data &
Content
Solera

DATA SCIENTIST:

TAKE-HOME TEST EXERCISE

In this document we describe a real world data set. The dataset aggregates data transmitted from a number of cars while they are driving around. Towards the end of the document there are questions for you to answer about the dataset.

1. Dataset Description

The dataset contains a row for each trip taken by a device user. The dataset contains the following self-explanatory variables (among others):

device_key This is a unique identifier for the device transmitting the data.

start_point_latitude This is the latitude of starting point of the trip. There is also an end point latitude.

start_point_longitude This is the longitude of the starting point of the trip. There is also an end point longitude.

start_point_timestamp This is the unix timestamp of the beginning of the trip.

2. Questions

For the below questions, perform the task required in Python or R. Please supply the code implementing your solutions.

1. The data may be dirty. Clean it.
2. Suggest some predictive analytics that can be done on this dataset.
3. Implement your chosen piece of predictive analytics.
4. What are the three top ways you would have improved this project, had you had more time?