

IDA Data Science Division Hands-on Interview

MRT Trips: Introduction

You are provided with the following dataset:

- **mrt_trips_sampled.csv**: A list of MRT train journeys on a particular day in Singapore.

Each line in the file consists of five fields: (a) a unique journey ID, (b) destination station, (c) time when the commuter tapped out, (d) origin station and (e) time when the commuter tapped in.

Your task is to write programs to answer the following questions:

Q1: What is the underlying MRT network according to the dataset provided? You should derive the network using only the dataset provided, without referring to other resources, such as the MRT network in Singapore from SMRT, Wiki, Datamall, etc.

Q2: Given an origin and destination station pair (say Station A and Station B), how many trains travelled from Station A to Station B in the given dataset?

Q3: Suppose a commuter wants to board a train at Clementi Station at 9:25am and alight at Kallang Station. Can you determine the number of people in the train at every station along the path from Clementi to Kallang starting from 9:25am?

This exercise should take you no more than a weekend. If you need any clarification, please don't hesitate to ask us.

Some of these questions are **difficult**. Hence, it would be a good idea to break up the problem into easier sub-problems. If you are not able to solve the problem, do show us your plan of how to solve the problem, as well as any intermediate results and code you might have worked out.

Deliverables

You are to deliver the following:

1. **Any code you wrote to answer the questions.** We should be able to run your code, and it should complete within a reasonable runtime. Please provide instructions for running and using your code on our systems, assuming some level of technical proficiency.
2. **A PowerPoint presentation outlining your approach to the problems and results you obtained.** Please include well-chosen charts to communicate data insights and messages as intuitively as possible. You will need to present this deck of slides when you meet with us for the face-to-face interview. This presentation should be no longer than 20 minutes.

Please send us the code and PowerPoint slides **by noon of the day before your face-to-face interview**.

Dataset

The dataset is available for download at this link:

https://drive.google.com/a/data.gov.sg/file/d/0B_n63NVfLDP-NjNnc1dNWGhLVVv/view?usp=sharing