Data Science Take Home Assignment

Goal

In silicon valley it is a common practice for tech companies to use shuttle buses to ferry their employees from home to the office. The goal of this exercise is to figure out the optimal stops for a bus shuttle. The company is based in Mountain View and the shuttle provides transportation for employees based in San Francisco.

With the explosion of user location data, data science can be used to optimize many of the services cities offer to their citizens. Transportation optimization is an example of that, but there are so many other possible applications. All this often goes under the (buzz name) "smart city" and it is one of the most interesting future applications of data science.

Challenge Description

Company XYZ has decided to offer a shuttle bus to help its employees commute from San Francisco to Mountain View. The city of San Francisco has given the company a list of potential bus stop locations to choose from and asked to not have more than 10 stops within the city.

You have been given the home address of all employees interested in taking the shuttle and asked to come up with the ten most efficient stops. While you have been given a certain freedom in defining what is "efficient", the general consensus within the company is that the most efficient way to select the bus stops is to minimize the overall walking distance between employee homes and the closest bus stop.

Estimating all possible 10 stop combinations would require a lot of time (how many combinations would that be?). Therefore, your boss is fine with simplifying the problem and returning 10 stops that have a high probability of being the best stops.

You should write an algorithm that returns the best 10 stops in your opinion. Also, please explain the rationale behind the algorithm.

Data Information

We have included a zip file called "Bus.zip" that contain 2 tables worth of data:

> "Potentail_Bust_Stops" - this is the list of potential bus stops given by San Francisco to the company. It is a list of intersections in the city.

Columns:

- **Street_One** : one of the two streets intersecting
- **Street_Two** : the other street intersecting

> "Employee_Addresses" - the home address of each employee interested in taking the shuttle.

Columns:

- **address** : employee address
- **employee_id** : employee id, unique by employee

Example

> Let's check the first bus stop candidate:

head (Potentail_Bust_Stops, 1)

Column Name	Value	Description
Street_One	MISSION ST	One of the two streets intersecting is Mission St
Street_Two	ITALY AVE	The second street is Italy Ave

That is, the bus stop candidate is [here](#)

```
    Let's now check where one employee lives
```

head (Employee_Addresses,1)

Column Name	Value	Description
address	98 Edinburgh St, San Francisco, CA 94112, USA	She lives in Edinburgh St
employee_id	206	Her employee id is 206

That is, the employee lives here