**20 Thieves Solution by Syed**

**Initial Scenario**

I received email from Tim Brewer at 11:42 and after reading the email and problem readme file. I was up to solve the problem by 12:00 PM. Now 48 hours mean, I must submit it before 12:PM Friday. That’s great. I liked the problem and I liked the approach taken by freelancer.com.

**My Approach Log**

* Initially I just skimmed through the problem to get an idea of what the problem is all about. And after starting writing this solution file, I got back to readme file, but this time to read it with precision. It is important to understand the problem first rather than thinking about the solution. My intial analysis of the problem can be found under Initial Problem Analysis Heading below.
* After completing the initial problem analysis, I moved to the data analysis part. This doesn’t mean programmatically analyzing data, rather it is about reading and understanding the data. The details of initial data analysis can be found under Initial Data Analysis heading below.
* After completing the initial analysis of the data and writing email to Tim Brewer, I moved forward with exploratory data analysis of data to get insight of the data and to identify the need to data preprocessing. Details are available below under initial exploratory data analysis heading.

**Initial Problem Analysis**

Few important observations.

* There is at-least one thief (Yes, it is obvious, but I preferred to mark it down)
* Scanner only scan the name and DOB for anyone who enters the club. (Not who left, be careful)
* Ranked list of up-to 20 suspects is needed. From most suspicious to least suspicious. There could be fewer. (Ranking problem)
* Post mid-night time will be considered as previous day
* We have visitor and theft logs
* Freelancer.com need details of how I approached and solved the problem and the problems I faced during the journey. (This is why I have created this document, I hope this will help)

**Initial Data Analysis**

Initial observations about the data:

* theft\_log.csv contains only the dates about the theft. (No time information is provided. This is lack of information, but we data scientist should work it out)
* visitor\_log.csv contains only the visit\_date, name, and dob. (Again we only have the day information. We only know that individual entered the place on the day, but we don’t know either he/she left on the same day or not)
* The joining attribute between visit and theft is date.
* Before moving forward, inquire from Tim Brewer that should be we asking question related to problem? Just to make sure we don’t commit any sins of 09 sins of data mining.

**Initial Exploratory Data Analysis**

* Before, I start performing the exploratory data analysis, I preferred to get the data into a database. Well, it is not necessary. Many data scientist can process the same data using MS Excel or directly using any data mining tools. But I prefer to work using SQL. And it really help us while working with large datasets. Other tools often start crying when data size grows. I used SQL Server 2016, again only for my personal preference. I created a database nct, short for nightclubthieves. Created two tables each for theft\_log and visitor\_log. Then I bulk inserted data into tables. The queries for this step are provided as initialbulkinsert.sql with this solution.
* I started with theft\_log.csv. First I decided to analysis the frequency of theft in each month.