**Open House 2019**

**Database Discrepancy**

**Report**

**Table Of Contents**

1. **Introduction…………………………………………… 2**
2. **Objective….…………………………………………… 2**
3. **Types Of Discrepancy…….……………………….... 3 -**
   1. **Incorrect Time Stamp…………………………………… 3 - 6**
   2. **Invalid Arrival And Departure Time…………………… 4**
   3. **Tour Cancellation With Inputs……………………. 4**
   4. **Extra SIT Stop Record…………………………….. 5**
   5. **Missing End Time…………………………………... 5**
   6. **Others………………………………………………... 6**
4. **Innovations Of Current Data …………………...…. 7 - 9**
   1. **Duration Of Each Stop..…………………………………. 7 - 9**

**1. Introduction**

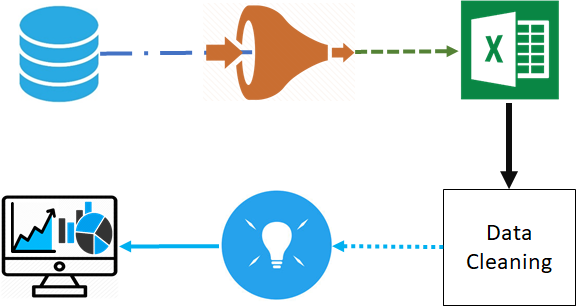
With the use of the “Database To Excel App”, I converted the Open House 2019 SQLite database into an Excel Sheet and was tasked to use the Open House 2019 database to create meaningful visualisations inside TIBCO Spotfire.

While doing so however, I met an obstruction that hindered me from using the data source properly. This obstruction consisted of multiple discrepancies created during the deployment of the Open House 2019 program.

**2. Objective**

In this report, I will be identifying these discrepancies. The discrepancies will be shown after converting the database into an Excel Sheet file format.

The cleaned data will then be used inside TIBCO Spotfire to create visualisations.



**3. Types Of Discrepancy**

Most of the errors are where inputs were either entered wrongly or registered in the wrong sequence. This resulted in the anomalous data in the database that can cause inaccuracy when consolidating the data to be used for analytical purposes.

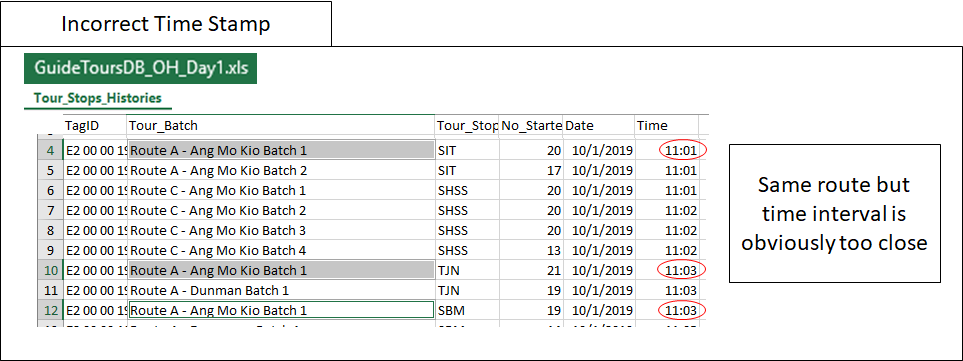
When viewing the Excel Sheet data, you can find the following errors:

1. Incorrect Time Stamp
2. Missing End Time
3. Tour Cancellation With Inputs
4. Extra SIT Stop Record
5. Others

Note: Although the list above is a compilation of all errors, there may be other unknown errors in the data sheet. Also, every listed error belongs to their own data table/ Excel worksheet.

**3.1 Incorrect Time Stamp**

Viewing the data, it was observed that certain time stamps were too close to each other, indicating that it was obviously incorrect. Below is a sample of the described data.



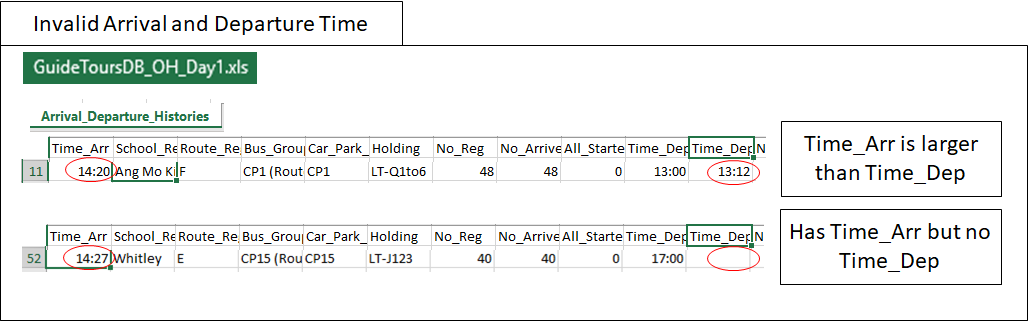
Another type of incorrect time stamp could also be that the tour group decided to follow a different order of stops, causing an irregular set of data.

Example: Stop order B,C,A instead of A,B,C.

To clean the data, assumptions have to be made to estimate the duration for each stop and add into the time. The estimated duration had to be referred with other data entries to ensure that the data is as accurate and reasonable as possible.

**3.2 Invalid Arrival And Departure Time**

The invalid start and end time is similar to the above example, where the human error is an incorrect or absence of an input for the designated data slot. This can cause confusion when creating analytic visuals due to inconsistency of the data.

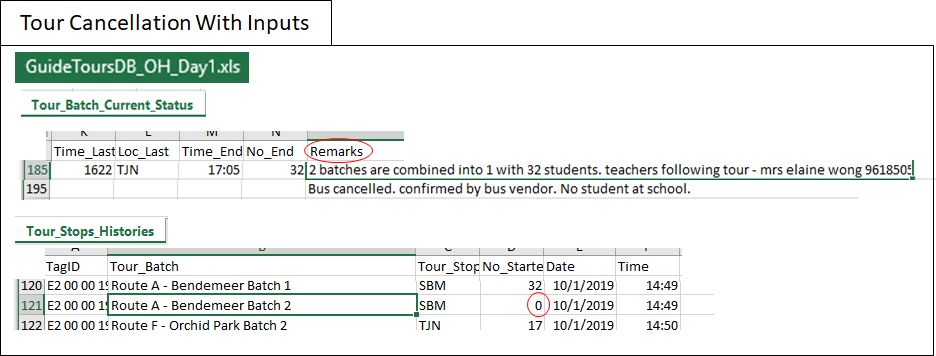


**3.3 Tour Cancellation With Inputs**

This error arises due to the accidental registration of a tour stop when a tour is cancelled. This can cause inaccuracy to the TIBCO Spotfire visualisations due to how the visuals functions.

By looking carefully into the Excel Sheet, you can spot remarks of user inputs, stating any additional comments like “Bus Cancelled” or “Tour combined”. This gives an indication that the current tour group was cancelled.

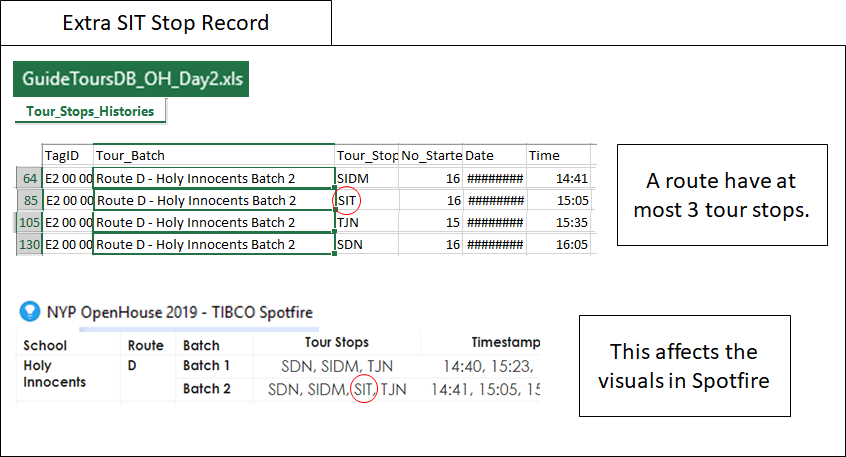
However, although the tour was cancelled, it was still able to pick up an input to register a tour stop history. This can cause inaccuracy as the erroneous data row can create a large valued anomaly, affecting the average result.



**3.4 Extra SIT Stop Record**

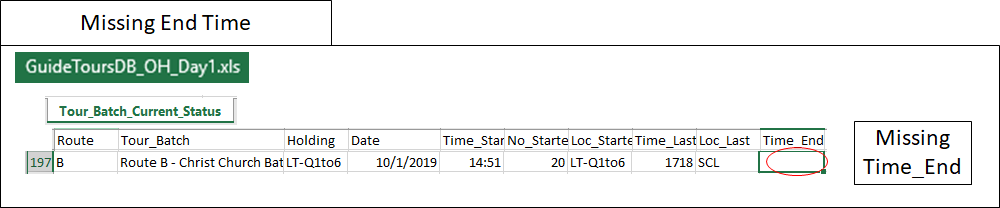
Occasionally, certain tours will have an additional SIT stop recorded into the tour stop history records. This is because in SIT, there are sensors that will automatically pickup signals to register a SIT stop.

This extra tour stop history will affect the visualisation of TIBCO Spotfire to display the wrong information. Below is an example of the extra tour stop record.



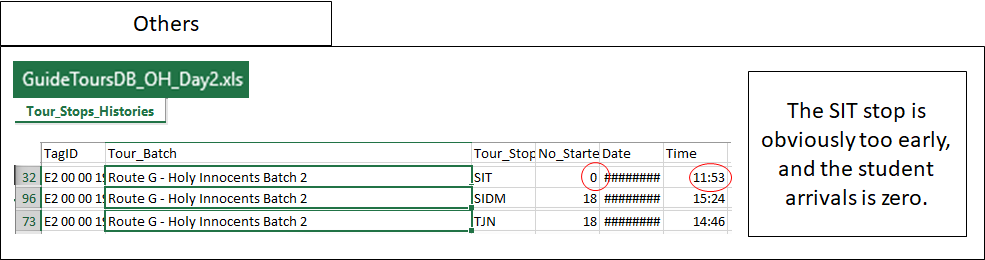
**3.5 Missing End Time**

Tour\_End is when they arrive back at the meeting location. Below is an example of a missing timestamp for the final arrival back to their release location.



**3.6 Others**

Other bugs with little occurrence will be listed under this category.



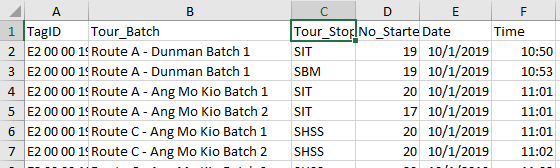
**4. Innovations Of Current Data**

In this section of the report, I will be sharing ideas on how to create more valuable data with the existing OH 2019 databases. The ideas shown will be using Excel Sheet.

In the current database, there are limitations that imposes restrictions on what can be used for analytical research. An example would be that the database does not contain information like duration of each stop. If a column exists for this dataset, more analytical insights can be induced from the given data. Hence, the purpose of this section is to give innovative ideas on how to overcome such limitations and restrictions.

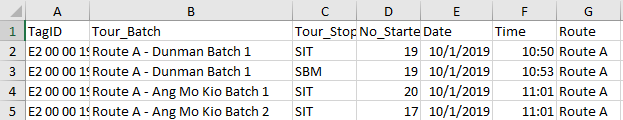
**4.1 Duration Of Each Stop**

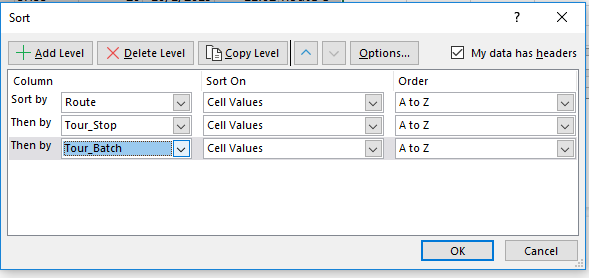
To create the duration of each stop, we must first identify the end time of each stop. The current datatable houses only the start time of each stop.



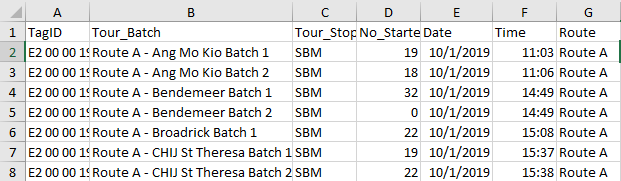
With the given current datatable, it is possible to induce the end time. Studying the Excel Sheet carefully and understanding the available data it provides, each stop can be derived by finding the order of stops per route, then using the start time of the next stop as the end time of the previous stop. The last stop however, will require additional steps as the end time of the last stop will be the time the group reaches the meeting location, which is from another table.

By using the tools of your working software, create a column for route. Then order the data by route by tour stop by tour batch. This will then allow for an easier time creating the end time and duration for each row.





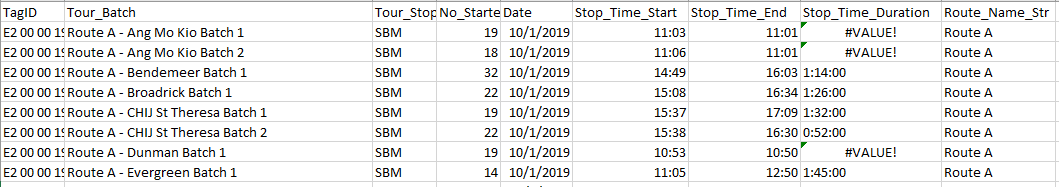
After creating the route column and sorting the data, you will see a similar table below.



With the last few steps, create two other columns for time end and stop duration. Then, with reference to the route info for stops in each route, fill in the time end by adding the data with the next stop start time.

After filling all the end time for each row, make a simple script to create an algorithm to take the difference between the end and start time of each stop. The resulting data will be served as the duration of each stop.

Your final datatable result should look similar to this.



Note that with the existing discrepancies mentioned, certain errors can occur. The end time may be larger than the start time. This can cause the stop duration to be invalid as it records as a negative timespan value.