# Data Intake Report

Name: G2M insight for Cab Investment firm
Report date: 05/14/2014
Internship Batch: LISUM33
Version: 1.0
Data intake by: Jackson Taylor
Data intake reviewer:
Data storage location:

**Tabular data details:**

**Cab_Data.csv**

| | |
|---|---|
| **Total number of observations** | 359392 |
| **Total number of files** | |
| **Total number of features** | 7 |
| **Base format of the file** | .csv |
| **Size of the data** | 20663 KB |

**City.csv**

| | |
|---|---|
| **Total number of observations** | 20 |
| **Total number of files** | |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 1 KB |

**Customer_ID.csv**

| | |
|---|---|
| **Total number of observations** | 49171 |
| **Total number of files** | |
| **Total number of features** | 4 |
| **Base format of the file** | .csv |
| **Size of the data** | 1027 KB |

**Transaction_ID.csv**

| | |
|---|---|
| **Total number of observations** | 440098 |
| **Total number of files** | |
| **Total number of features** | 3 |
| **Base format of the file** | .csv |
| **Size of the data** | 8788 KB |

**Proposed Approach:**

- Mention approach of dedup validation (identification):
  I checked for duplicate (Transaction ID) for Cab_Data.csv (non found)
  I checked for duplicate (Transaction ID) for Transaction_ID.csv (non found)
  I checked for duplicate (Customer ID) for Customer_ID.csv (non found)

- Mention your assumptions (if you assume any other thing for data quality analysis):
  In Cab_Data.csv, for each transaction the profit for that transaction is (Price Charged – Cost of the Trip).
  Date of travel is number of days starting from January 1, 1900.
  In the City.csv, the number of Users is the total number of cab users in the city including Pink cab and Yellow cab as well as other cab companies.
  A ride can have multiple transactions. The cost is adjusted to only count the cost once for over lapping routes for multiple people. The charge varies.