

# Data Intake Report

Name: Yellow V. Pink Cab

Report date: June 14, 2022

Internship Batch: LISUM10

Version:<1.0>

Data intake by: William Harvey

Data intake reviewer:

Data storage location: <https://github.com/DataGlacier/DataSets>

## Tabular data details:

<b>Total number of observations</b>	359392
<b>Total number of files</b>	4
<b>Total number of features</b>	15
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	46 MB

When I began to clean the data, I first checked for NaN values, which was only one row, so then I removed that row. I knew that the only place duplicates could exist was if there were multiple of the same Transaction ID, and after further exploration no duplicates were found.

Trip duration is not known, so even though there are some outliers in the Price Changed column, I decided to keep those outliers. The profit column I calculated was simply subtracting the Price Charged columns and the Cost of Trip columns, everything else was kept at 0.