

# Data Intake Report

Name: Pink and Yellow Cab Analysis

Report date: 11/03/2021

Internship Batch: [LISP01](#)

Version:<1.0>

Data intake by: Rahma Mahjoub Abker Habeeb

Data intake reviewer : ...

Data storage location: [https://Rahma\\_Mahjoub@bitbucket.org/Rahma\\_Mahjoub/week2.git](https://Rahma_Mahjoub@bitbucket.org/Rahma_Mahjoub/week2.git)

## Tabular data details:

<b>Total number of observations</b>	359392
<b>Total number of files</b>	4
<b>Total number of features</b>	19
<b>Base format of the file</b>	.csv
<b>Size of the data</b>	52.4+ MB

## Proposed Approach:

The following steps have been done:

1. Loading the four database files.
2. Printing the number of columns and rows for each database file.
3. Looking for null values.
4. Printing the data types and capacity for each database file.
5. Merging the four files into a single file (master data).
6. Applying feature extraction on the master data and the results was two features ( Number of travel and Profit).
7. The column of (date of travel) was divided into (year, month and day).
8. Renaming some columns and changing the data type.
9. Some statistics was applied on the data.
10. Finding the relationship between the numeric and categorical variables.
11. Detecting the outliers and duplicated rows.
12. Saving the data after preprocessing into a csv file.
13. Multiple hypothesis were created.
14. lastly the model was created by using two algorithms which were (ARIMA and Prophet ), with comparing results for each.
15. The profit forecasting was done for two years.