**Dataset QA**

**File:**

**NY Weather Data - Apr 14 - Jul 14**

Excel

|  |  |
| --- | --- |
| Step | Action |
| 1 | Modify headers to match columns |
| 2 | Add Month / Full date columns |
| 3 | Remove unneeded repeat header rows |

Python (script included)

|  |  |
| --- | --- |
| Step | Action |
| 1 | Drop unneeded date columns |
| 2 | Replace T in precipitation (trace) to zero |
| 3 | Convert weather events (rain, fog, thunderstorm, snow) into separate binary indicator columns |

**Files:**

**uber-raw-data-apr14.csv**

**uber-raw-data-may14.csv**

**uber-raw-data-jun14.csv**

**uber-raw-data-jul14.csv**

Python (script included)

|  |  |
| --- | --- |
| Step | Action |
| 1 | Selected files based on months that matched weather (April, May, June, July) |
| 2 | Union the 4 files together (concat) |
| 3 | Add pickup Id column |
| 4 | Convert Date/Time into date data type |
| 5 | Create columns for each date element (month, day, hour) |
| 6 | Use left join on date to merge weather data with uber date |

**Final Dataset (5 files combined)**

2,676,915 Rows

35 Columns

0 Missing values

Tableau Transformations

|  |  |
| --- | --- |
| Step | Action |
| 1 | Use fixed LODs to calculate daily / hourly pickup counts |
| 2 | Used corr function to calculate correlations |
| 3 | Created binary indicators for rain and warmer than avg day events |
| 4 | Exported data aggregated by day to use for ML model experiments |