**Project**

|  |  |
| --- | --- |
| **Project Name** | **Team Members** |
| ETL Project - Capturing Sales Data vs Weather Data Impact | **Prajakta Gaikwad** |
| **Isaac Muck** |
| **Allen Chang** |
| **Jean Carlo Camacho** |

**E**xtract

|  |  |  |
| --- | --- | --- |
| **Data** | **Data Source** | **Data Format** |
| NYC Weather Data (2016) | Kaggle | CSV |
| NYC Store Sales Data (2016) | BBB | CSV |

**T**ransform

|  |  |
| --- | --- |
| **Data** | **Transformation Needed** |
| NYC Weather Data (2016) | In the weather data, cells with no data are represented with the string 'T'. We replaced these with empty values. |
| The dates are in 2 different formats (M/D/Y and D-M-Y). We converted all dates to a standard datetime format. |
| We replaced the spaces in some column headers with underscores. |
| Converted object types to float where appropriate. |

|  |  |
| --- | --- |
| **Data** | **Transformation Needed** |
| NYC Store Sales Data (2016) | Data included SALES and RETURNS, we had to filter for SALES data only. |
| There were fields with ‘?’ in them for null values, those fields had to get converted to ‘0’. |
| Data types needed to be adjusted to keep the data uniformed. |

**L**oad

MYSQL databases were used throughout the process. We loaded both CSV files into

Python data frames and then did the cleanup using Pandas. SQL Alchemy was used to make a connection to the SQL databases. We merged both data frames and inserted the results into the SQL database created.