# INTERMEDIATE STATUS REPORT

# BIKE SHARE TREND ANALYSIS

**Bhakti Raichura: 016020628**
**Khushee Thakker: 015271529**
**Shreya Srirama: 016029845**
**Varsha Srinivasan: 016001544**

## 1.Progress towards the goal achieved so far

### Phase 1 : Data selection and proposal (Status- Complete)

**Project data set used:** The data set describes the bike sharing trip data in the city of Seattle for a period of 3 years

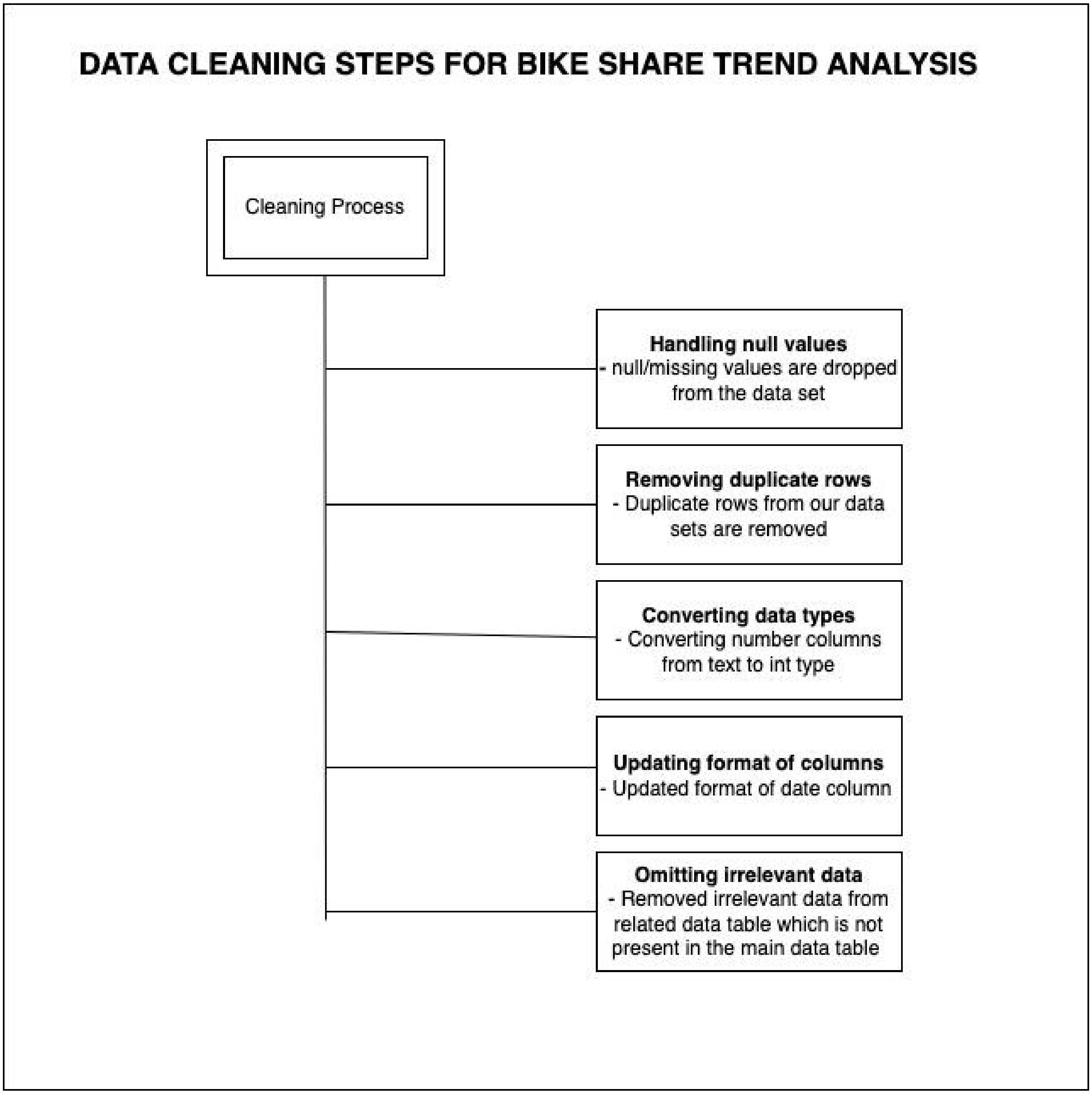
https://data.seattle.gov/Community/Pronto-Cycle-Share-Trip-Data/tw7j-dfaw

### Phase 2 : Scope Finalization (Status- Complete)

The analysis to be performed on the data set will include but not be limited to  the following

- Time of the day most users ride the cycle
- Weather analysis on the data- Impact of weather conditions- wind speed, humidity, rain etc. on the trip data (can also be bike share count season wise). This helps in fixing price range based on that
- Impact of weekends and weekdays on users bike riding patterns
- Location of stations which can influence the demand of bikes
- Locations to which users commute more on a daily basis
- Most used routes which can be used to construct lanes/roads based on that

## Phase 3 : Data Cleaning and Normalization (Status- Complete)



DATA CLEANING STEPS FOR BIKE SHARE TREND ANALYSIS

Cleaning Process

**Handling null values**
- null/missing values are dropped from the data set

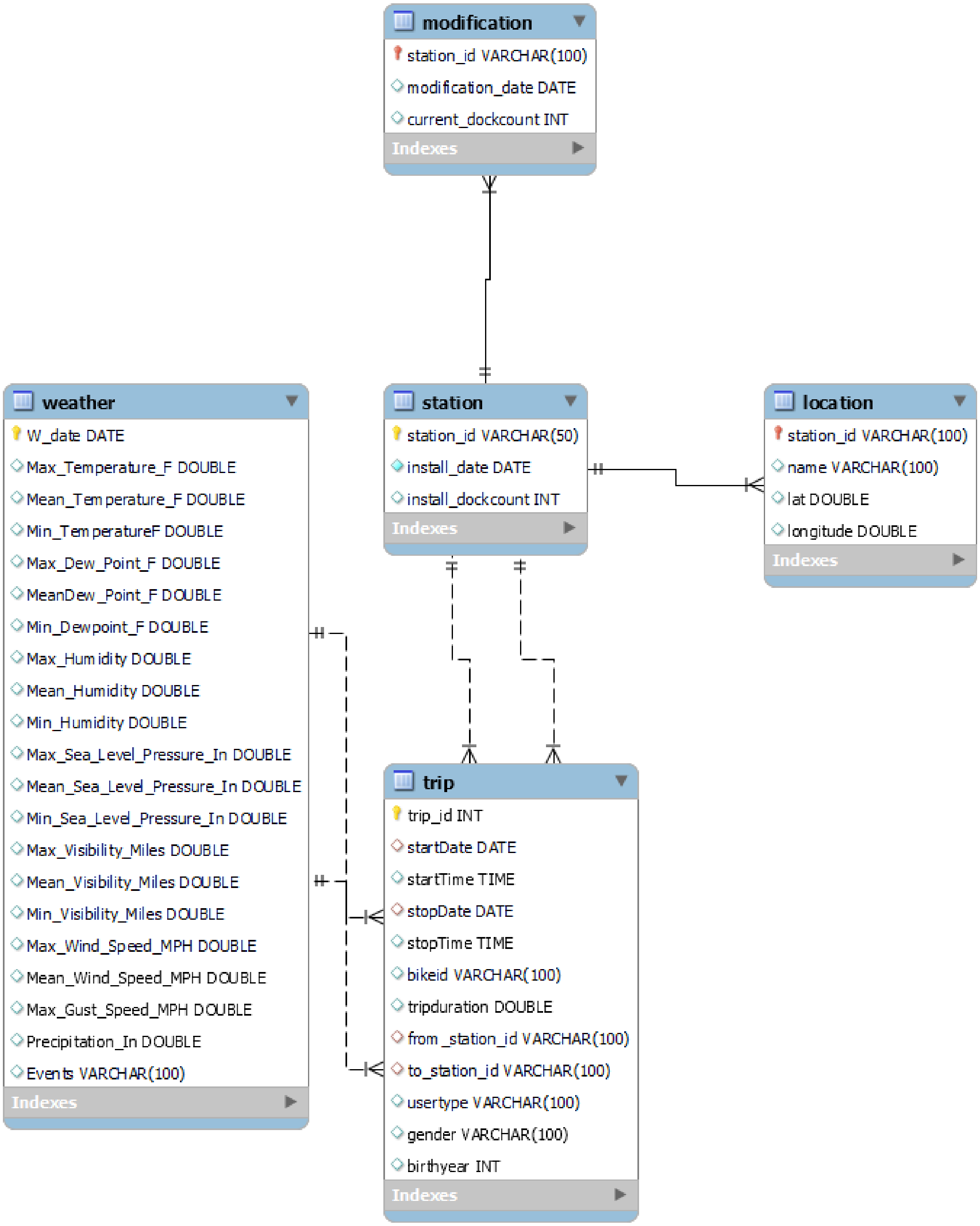
**Removing duplicate rows**
- Duplicate rows from our data sets are removed

**Converting data types**
- Converting number columns from text to int type

**Updating format of columns**
- Updated format of date column

**Omitting irrelevant data**
- Removed irrelevant data from related data table which is not present in the main data table

## ER Diagram


modification
station_id VARCHAR(100)
modification_date DATE
current_dockcount INT
Indexes
weather
W_date DATE
Max_Temperature_F DOUBLE
Mean_Temperature_F DOUBLE
Min_TemperatureF DOUBLE
Max_Dew_Point_F DOUBLE
MeanDew_Point_F DOUBLE
Min_Dewpoint_F DOUBLE
Max_Humidity DOUBLE
Mean_Humidity DOUBLE
Min_Humidity DOUBLE
Max_Sea_Level_Pressure_In DOUBLE
Mean_Sea_Level_Pressure_In DOUBLE
Min_Sea_Level_Pressure_In DOUBLE
Max_Visibility_Miles DOUBLE
Mean_Visibility_Miles DOUBLE
Min_Visibility_Miles DOUBLE
Max_Wind_Speed_MPH DOUBLE
Mean_Wind_Speed_MPH DOUBLE
Max_Gust_Speed_MPH DOUBLE
Precipitation_In DOUBLE
Events VARCHAR(100)
Indexes
station
station_id VARCHAR(50)
install_date DATE
install_dockcount INT
Indexes
location
station_id VARCHAR(100)
name VARCHAR(100)
lat DOUBLE
longitude DOUBLE
Indexes
trip
trip_id INT
startDate DATE
startTime TIME
stopDate DATE
stopTime TIME
bikeid VARCHAR(100)
tripduration DOUBLE
from_station_id VARCHAR(100)
to_station_id VARCHAR(100)
usertype VARCHAR(100)
gender VARCHAR(100)
birthyear INT
Indexes

## 2. Findings / Results so far

- Found duplicate rows in datasets (including 50K rows duplicated in trip data set)
- Worked on missing data and redundancies
- We found few additional station ids (like pronto shop, pronto shop 2) in the trip table which were missing in the station table
- Pronto bike share system was launched on 13th Aug 2014. Major of the station dock was installed from the month august 2014 to december 2014. In later years there were new docks installed and some docks were completely uninstalled
- In our data, ‘usertype’ has 2 possible values
    - o "**Short-Term Pass Holder**" is a rider who purchased a 24-Hour or 3-Day Pass
    - o "**Member**" is a rider who purchased a Monthly Subscription, an Annual Subscription, or a Special Pass.
- We found that the user details (gender and birth year) was only collected to the usertype ‘member’ and not for ‘Short-Term Pass Holder’
- Usage of python in our project:
    - o Loaded the data into a data frame and cleaned our dataset
    - o Connected to MySQL and created database **Cycle** and required tables

## 3. Difficulties being encountered and how you plan to resolve them

- Data is denormalized
    - o Solution: Data normalization to remove redundancies. Station table has multiple themes making it challenging to analyze to derive insights
    - o Final tables to be used: Station, trip, weather, location, modification table (Our database is in 3NF)
- Multiple and unsupported date formats in trip and weather tables. Hence had to format the dates in these tables
- Multiple Null values in the weather table had to replace them with ‘0’ for integer data type column and ‘Null’ for string data type column
- The trip table had a ‘datetime’ column whereas the weather table had only a ‘date’ column.That made it difficult to associate these 2 tables. To resolve this difficulty we split the ‘datetime’ column in trip into 2 separate columns i.e Date and Time. So we could create the relationship between these 2 tables using startDate.

## 4. Remaining tasks

- Further analysis required on the data set. The below table shows a summary of all the tasks:

| Topic | Status |
|---|---|
| Data Selection and proposal | Complete |
| Scope Finalization | Complete |
| Data Cleaning | Complete |
| Analyzing data set | In Progress (Python and MySQL analysis) |
| Data Visualization | To be completed (Tableau) |
| Documentation | To be completed |

## 5. Any others that you think is relevant

- Data connection in Python and MySQL to find correlations and further analysis
- Interactive dashboards to help to represent our analysis that is done using Tableau
- Code repository to be made available on GitHub for the above analysis