DIVY

Group 4

Jane Liu / Mia Song / Minh Vo

Nida Ulhaq Fitriyah / Rolamjaya Hotmartua /Sakshi Shende

Agenda

- Executive Summary
- Business Case
- Methodology
- Data Profiling
- Data Model
- Insights
- Summary
- References

Executive Summary

Background

- Divvy is a Chicago bike-sharing system launched in 2013 by Motivate and purchased by Lyft in 2019.
- The primary bike-share system in Chicago, with over 400,000 new riders and 5.5 million rides taken in 2021

Status Quo

Goal: CDOT and Lyft aim to make Divvy available for all Chicagoans

Status Quo: Launched e-bike in 2020 and electrified stations in 2021

Project Objectives

- Goals: To assist Divvy's plan on increasing overall bike availability for Chicagoans and attracting more membership riders.
- Plans:
  - Provide analysis and visualization
  - Give recommendations to Divvy ridership and bike allocation

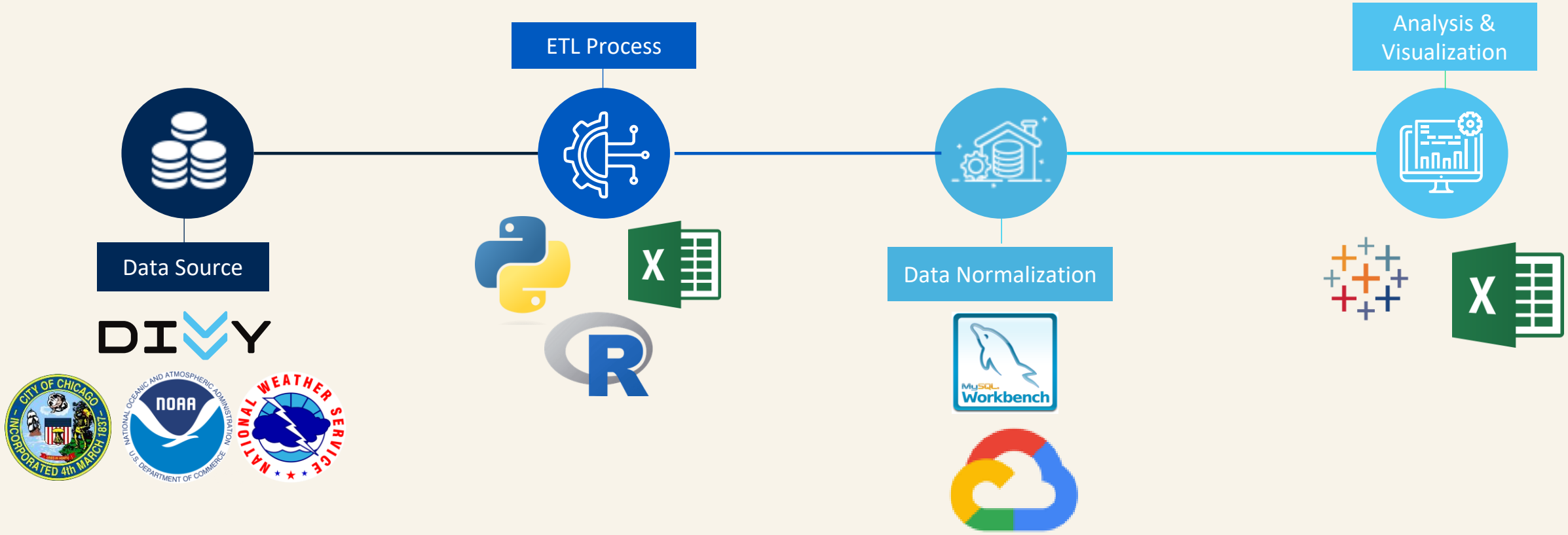
Problem Statements

- Station locations and bicycle allocation
- Current ridership by rider groups

- Weather influence on the ridership: temperature, precipitation, snow

- User biographical, demographic and membership information
- User behavior about peak time, service duration
- User preference between bike types

Data Pipeline



Data Preparation

Data Set	Trip	Weather	Population	Station
Data Source	Divvy	National Weather Service	Chicago Data Portal	
Attributes	ride type, member type, service time, departure and arrival station and its geography	max and min temperature, precipitation, snow depth	population by gender and age	Station name, bicycle dock information (number, service status), location of stations
Data Period	2020.11 ~ 2021.10	2020.11 ~ 2021.10	2018	Current (2022)
Data Size (RXC)	11.5 M X 13	763 X 13	121 X 26	1420 X 8
Missing Values	There are missing values in some columns such as start_station and end_station	N/A	There are missing values for detailed data per zip code for some age ranges	N/A
Anomalies	N/A	There are non-numeric values i.e., T for precipitation	N/A	Duplicate Station ID

Data Cleansing

Case	Null	Abnormal	New
Definition	Data is non-existing where it should have been	Data is existing but clearly contains wrong information	Data is calculated using existing ones, making new column additions
Example	Missing station names	1) Station ID Inconsistency (duplication in ID) 2) Rental starting time later than rental ending time	1) Trip duration: ending - starting time 2) Trip distance: end - start location 3) Zip code: latitude + longitude
Problem	Losing data	Inaccuracy in analysis	Technical knowledge required
Strategy	1) Trying to fill the blanks as much as possible 2) Removing nulls if insignificant	1) Making the right alternative decisions 2) Removing anomalies if insignificant	1) Identifying new columns necessary 2) Utilizing programs and supporting functions

Data Cleansing

- **Null**
  - **Missing Station Name:** Found and Filled by looking up the available latitude and longitude data to station table. Some trips that have no latitude and longitude data is removed (less than 1% of the data – insignificant).
  - **Zip Code Not Provided:** Calculated from latitude and longitude data from station table using package in python
  - **Missing Population Data for Some Age Ranges:** Exclude the non-related age ranges from our analysis.
- **Abnormal**
  - **Station ID Inconsistency:** Ignore the original station ID. Make surrogate key. Use the name of station as the main reference. There are duplicates but can be solved by seeing the string similarity. No station data is removed.
  - **Starting Time is later than Ending Time:** Removing abnormalities because they take only 0.0003% of the data set.
- **New**
  - **Calculate the Duration**: Changing starting and ending time's data type into date using as.POSIXct() and applied difftime() in R.
  - **Calculate the Distance:** Using latitude and longitude data of start and end station, distance is calculated using haversine function. Haversine function calculated distance by the fact that earth is round in Python.

Data Assumptions

- **Gender and age:** Have similar(same) traits with the population in the bicycle station's location
- **Estimated Price:** calculated by $1 for unlock and $0.39/minute(casual) and $1 for unlock and $0.16/minute(member)
- **Weather:** Use weather data measured on a daily basis
- **Not Null:** for end_station_key, end_time, duration, estimate_price
- **Zip Code:** created for each station by combining latitude and longitude in the data cleansing stage

Data Model: Normalizing process

- Normalizing table containing trip data: separate several attributes into their own table.

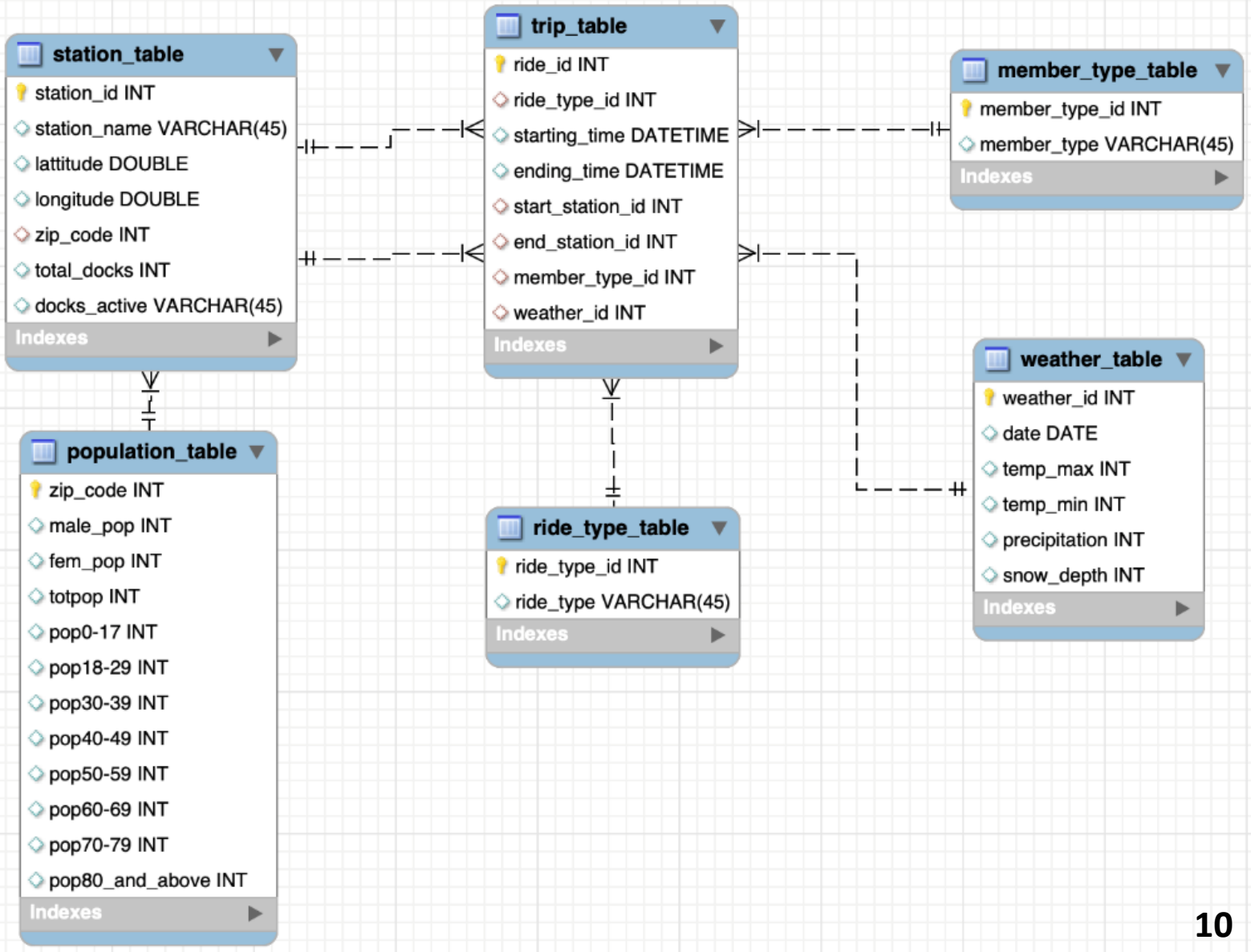
Name column of Raw Trip Data Table

[1]	[2]	[3]	[4]	[5]	[6]
ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id

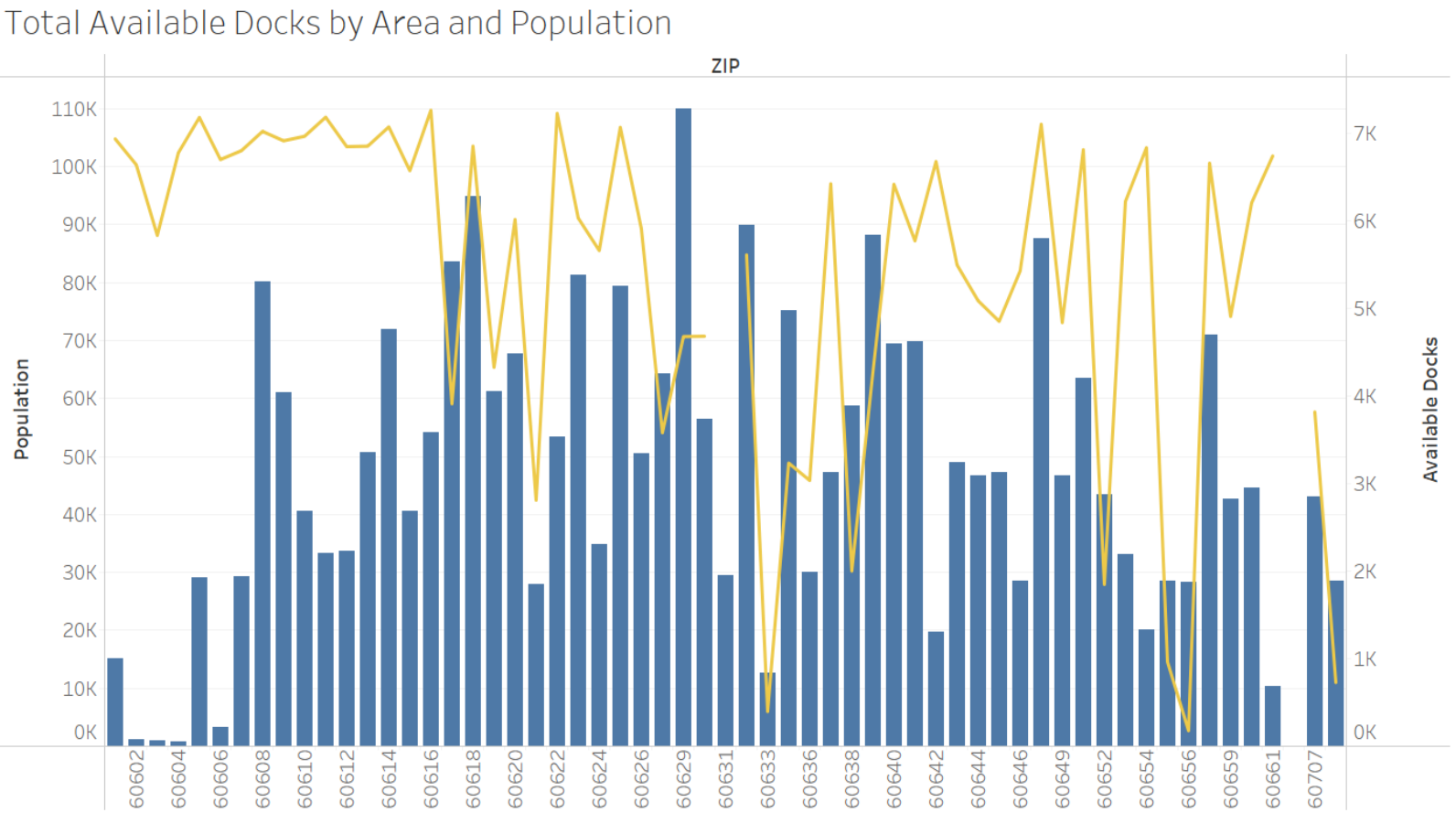
[7]	[8]	[9]	[10]	[11]	[12]	[13]
end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual

- Attribute [5], [7], [9] - [12] is grouped into station table, matched and supplemented by the station table from data source.
- Attribute [6] and [8] become FK refer to station table.
- Attribute [2] and [13] are also separated and created their own table.
- Station Table -> + column: Zip Code -> connect to Population Table.
- New Normalized Trip Table -> + column: Weather ID -> connect to Weather Table.

Data Model: Relational

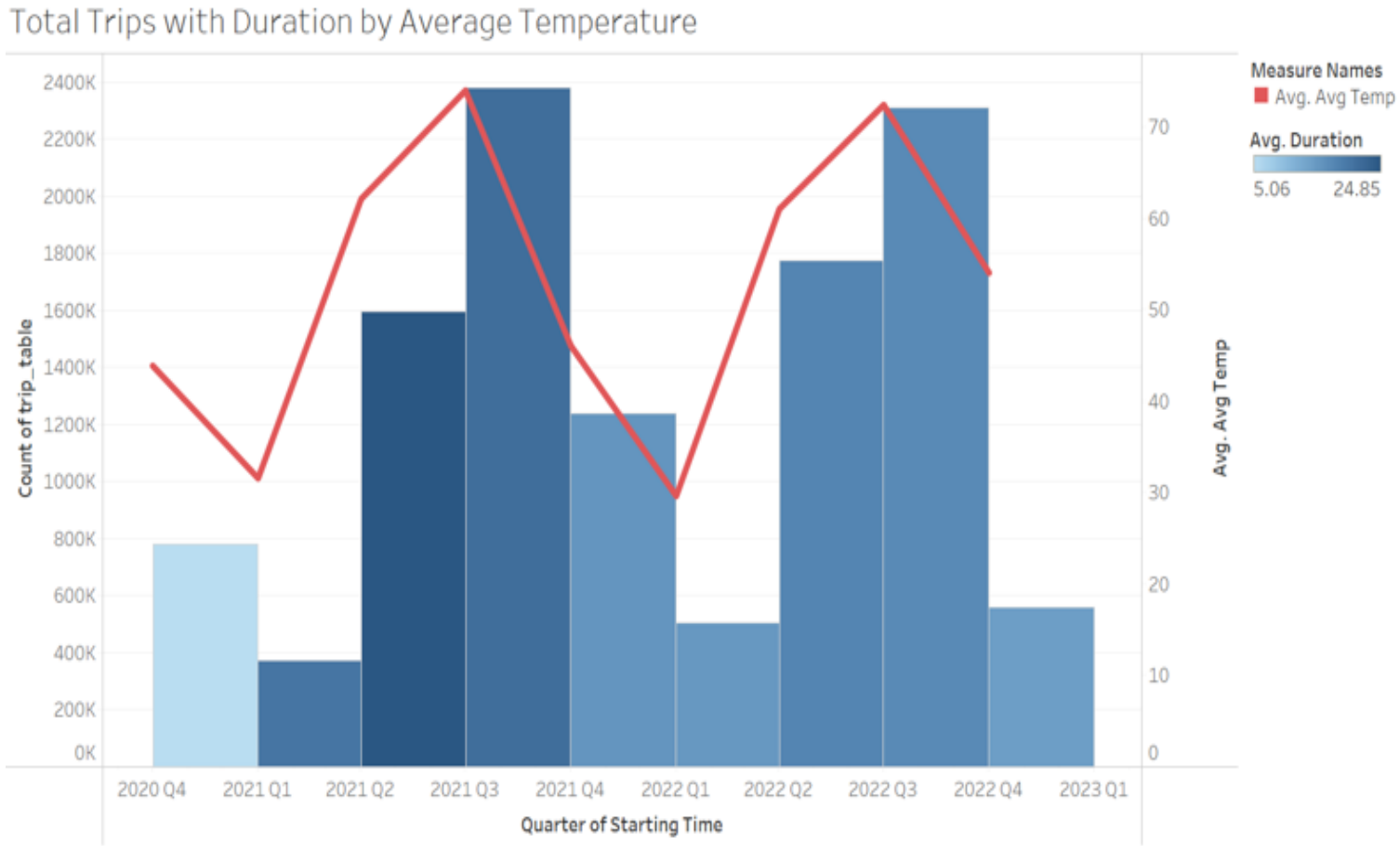


Insights: Population



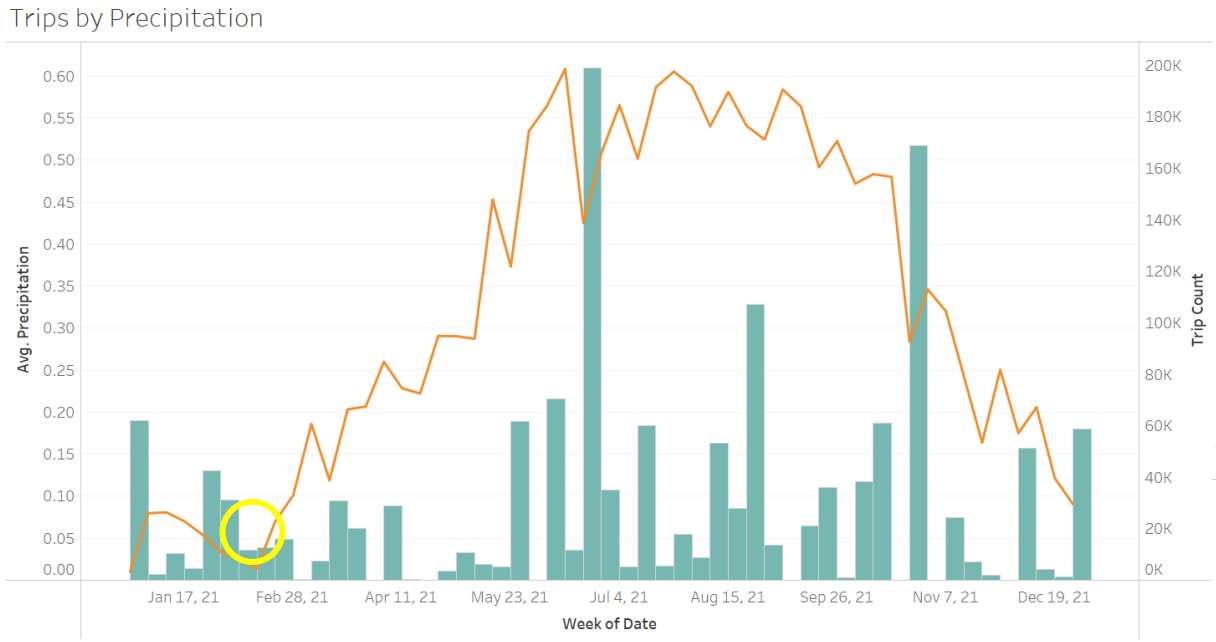
- We found several places with notable discrepancies.
- 60614 is close to downtown Chicago, and Lincoln Park and the Lincoln Park
- A larger floating population is expected hence the number of available bikes is relatively greater.
- **Promotions for non-members are recommended for areas similar to 60614.**

Insights: Weather

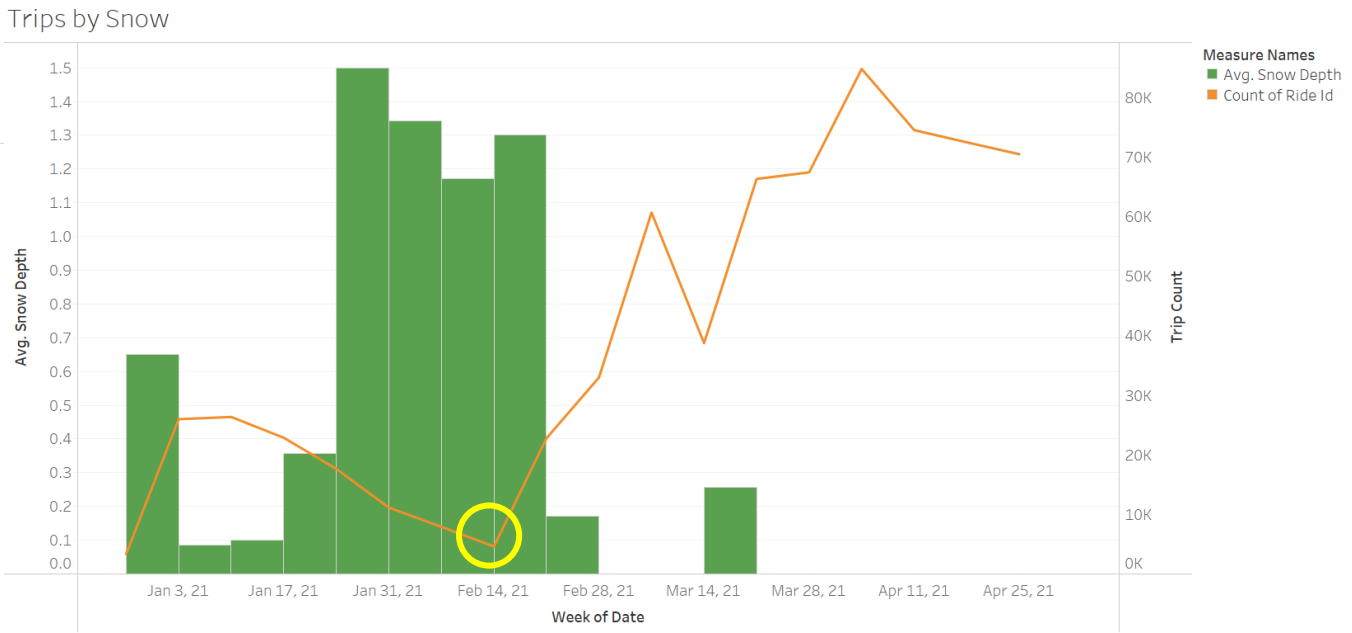


- In Q3, most members used Divvy (over 2M trips), while members used Divvy the longest in Q2. (around 20 min).
- Weather played a crucial role to determine the number of trips and trip duration.
- Increasing the duration of the trip can make the service more profitable.
- **Promoting the service in Q2 to extend trip duration such as rental rates discount, therefore, can be a valid strategy for the business.**

Insights: Weather

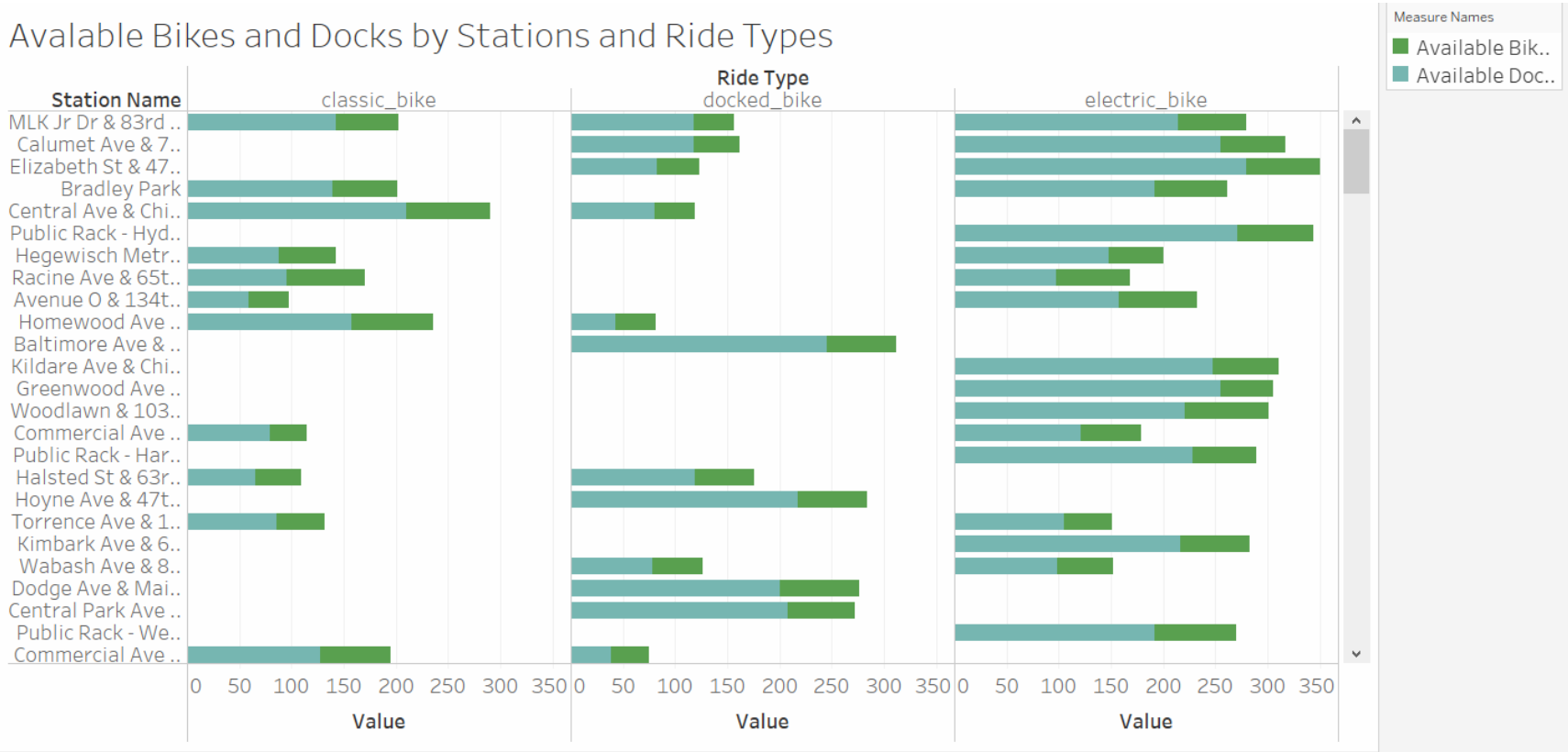


- What keeps Divvy customers from riding is snow, not rain!
- No significant slowdown was observed during the rainy weeks of 2021. However, the number of trips dropped considerably during heavy snowfall in February of that year.



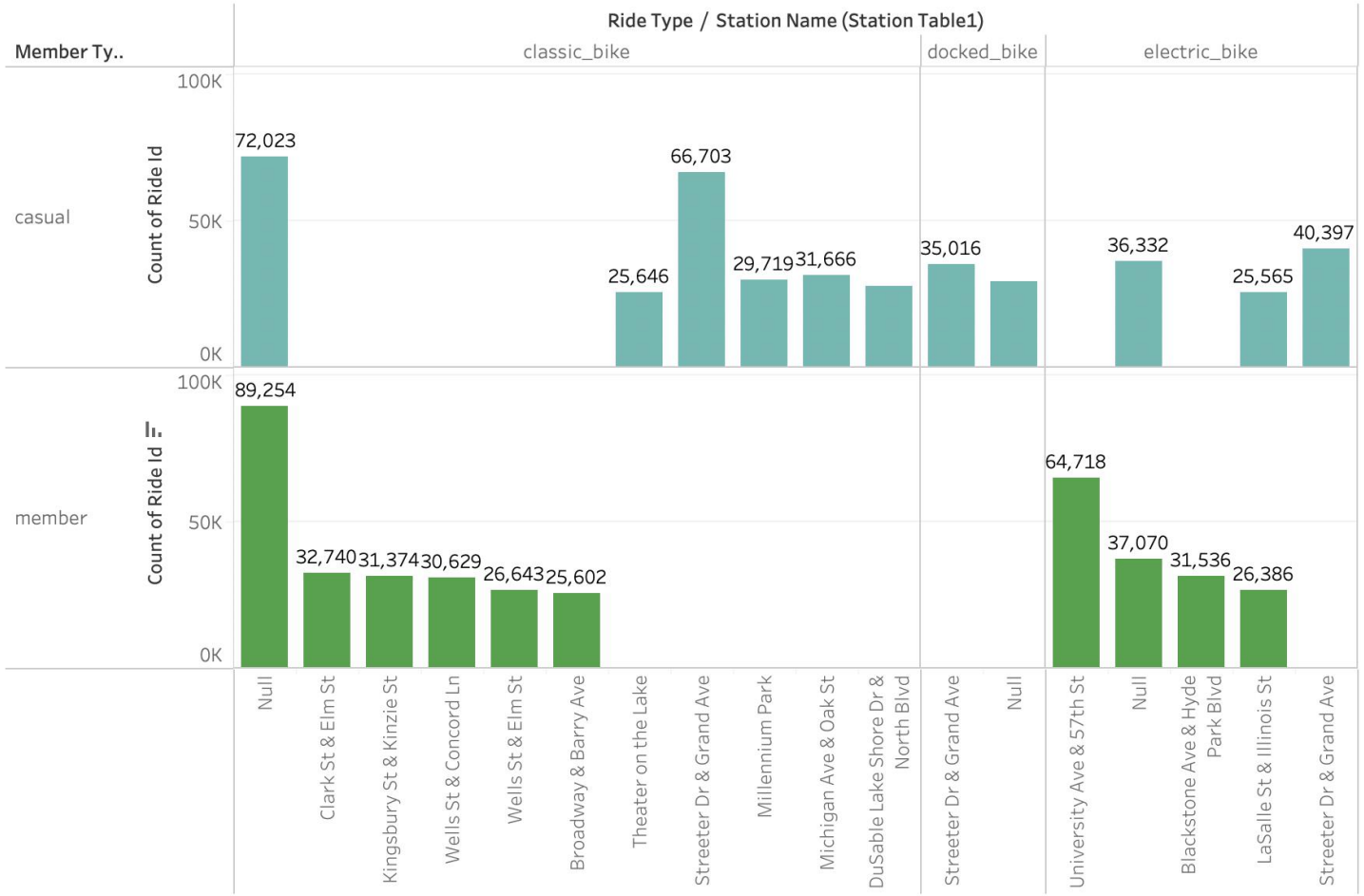
- **Business plans can be two tracks:**
  - **Maximize profits in good weather**
  - **Maintain and repair the service in bad weather, preparing for the next season.**

Insights: Top Busiest Stations



- Top busiest start stations are indicated by **35 and 81 total available bikes.**
- Top busiest start stations have sufficient available docks for bikes.
- **Divvy should pay special attention to the stations where docks are insufficient to park all bikes.**

Insights: Top Busiest Stations



- Top busiest start stations are indicated by **number of riders >25k.**
- Member vs Casual riders have different favorite stations.
- **Divvy should develop different marketing strategies at distinct station to acquire different users (member vs casual).**

Summary

Recommendations & Future Works:

- Rearrange bike allocation (at busy stations and stations with insufficient available docks)
- Increase membership riders with promotion programs.
- More in-depth analysis of ridership at popular stations like total rides on weekday vs. weekend, etc.

Lessons learned:

- Analysis cannot be done without prepared datasets despite the right business questions.
- Design data pipeline based on data used and its size.
- Data comes from various sources and might have inconsistent formats.
- Normalization process for large size of data can be slow.

THANK YOU

References

City of Chicago Official Website

- https://www.chicago.gov/city/en/depts/cdot/provdrs/bike/news/2018/june/divvy-celebrates-five-years-in-chicago.html
- https://www.chicago.gov/city/en/depts/cdot/provdrs/future_projects_andconcepts/news/2022/may/divvy-becomes-first-u-s--bikeshare-system-to-incorporate-ebike-c.html)

Raw Datasets:

- Divvy Trip Data: https://divvy-tripdata.s3.amazonaws.com/index.html
- Divvy Station: https://data.cityofchicago.org/Transportation/Divvy-Bicycle-Stations/bbyy-e7gq/data
- Weather Data: https://www.weather.gov/wrh/climate?wfo=lot
- Population: https://data.cityofchicago.org/Health-Human-Services/Chicago-Population-Counts/85cm-7uqa