MISM6210 Project 2

Michael Conan

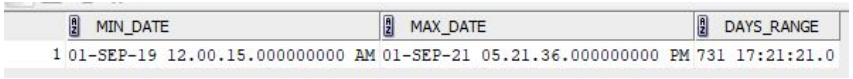
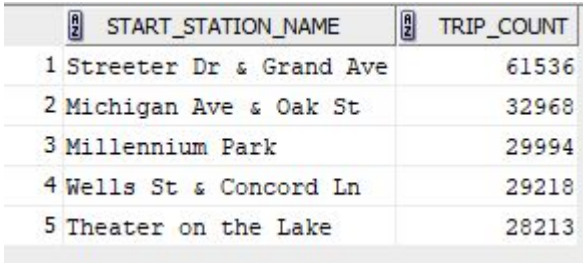
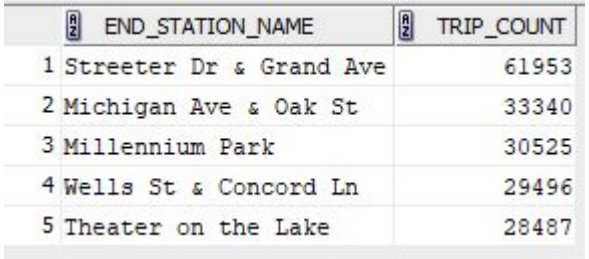
Why we ride

Divvy can reach its new member goals by targeting chicago commuters, BUT
Bad experiences finding bikes or open docks limit new membership.

How can Divvy fix its bike rebalancing challenge?

Oracle SQL Analysis

Divvy Bikeshare

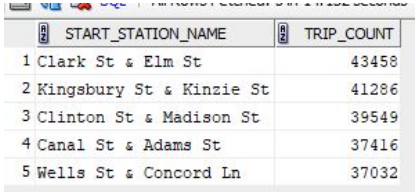
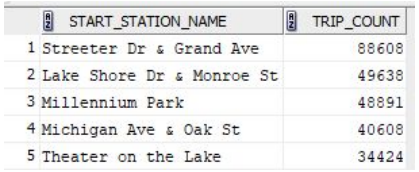
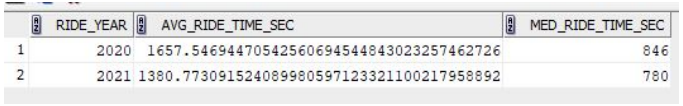
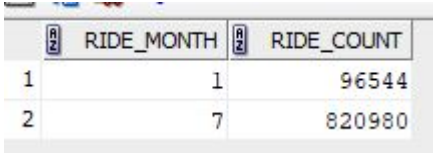
Prompt	Query	Result
Over what period of time does the provided Divvy data cover?	*-- Range of dates in table data* **Select** MIN(STARTED_AT) min_date, MAX(ENDED_AT) max_date, MAX(ENDED_AT) - MIN(STARTED_AT) days_range **From** C##MISM6210.DIVVY;	
Considering only rides that started in 2021, which 5 stations are most used for ride starts?	*-- Top 5 2021 start stations* **Select** * **From** (**Select** START_STATION_NAME, Count(*) trip_count **From** C##MISM6210.DIVVY **Where** START_STATION_NAME is not Null **And** Extract(Year from STARTED_AT) = 2021 **Group By** START_STATION_NAME **Order By** trip_count desc) trips **Where** rownum < 6;	
...What about trip ends?	*-- Top 5 2021 end stations* **Select** * **From** (**Select** END_STATION_NAME, Count(*) trip_count **From** C##MISM6210.DIVVY **Where** END_STATION_NAME is not Null **And** Extract(Year from STARTED_AT) = 2021 **Group By** END_STATION_NAME **Order By** trip_count desc) trips **Where** rownum < 6;	

	MIN_DATE	MAX_DATE	DAYS_RANGE
1	01-SEP-19 12.00.15.000000000 AM	01-SEP-21 05.21.36.000000000 PM	731 17:21:21.0

	START_STATION_NAME	TRIP_COUNT
1	Streeter Dr & Grand Ave	61536
2	Michigan Ave & Oak St	32968
3	Millennium Park	29994
4	Wells St & Concord Ln	29218
5	Theater on the Lake	28213

	END_STATION_NAME	TRIP_COUNT
1	Streeter Dr & Grand Ave	61953
2	Michigan Ave & Oak St	33340
3	Millennium Park	30525
4	Wells St & Concord Ln	29496
5	Theater on the Lake	28487

SQL Analysis 1/2

Prompt	Query	Result
What are the Top 5 trip start stations for members?	*-- Top 5 start stations for members* **Select** * **From** (**Select** START_STATION_NAME, Count(*) trip_count **From** C##MISM6210.DIVVY **Where** START_STATION_NAME is not Null **And** MEMBER_CASUAL = 'member' **Group By** START_STATION_NAME **Order By** trip_count desc) trips **Where** rownum < 6;	 START_STATION_NAME | TRIP_COUNT 1 Clark St & Elm St | 43458 2 Kingsbury St & Kinzie St | 41286 3 Clinton St & Madison St | 39549 4 Canal St & Adams St | 37416 5 Wells St & Concord Ln | 37032
...and top 5 trip start stations for casual riders?	*-- Top 5 start stations for casual riders* **Select** * **From** (**Select** START_STATION_NAME, Count(*) trip_count **From** C##MISM6210.DIVVY **Where** START_STATION_NAME is not Null **And** MEMBER_CASUAL = 'casual' **Group By** START_STATION_NAME **Order By** trip_count desc) trips **Where** rownum < 6;	 START_STATION_NAME | TRIP_COUNT 1 Streeter Dr & Grand Ave | 88608 2 Lake Shore Dr & Monroe St | 49638 3 Millennium Park | 48891 4 Michigan Ave & Oak St | 40608 5 Theater on the Lake | 34424
What was the mean ride duration in 2020? What was the median ride duration in 2020? What was the mean and median ride duration so far in 2021?	*-- Mean and median ridetimes in 2020 and 2021* **Select** Extract(Year from STARTED_AT) ride_year, avg(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60*60 avg_ride_time_sec, median(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60*60 med_ride_time_sec **From** C##MISM6210.DIVVY **Where** Extract(Year from STARTED_AT) > 2019 **Group By** Extract(Year from STARTED_AT);	 RIDE_YEAR | AVG_RIDE_TIME_SEC | MED_RIDE_TIME_SEC 1 2020 | 1657.5469447054256069454484302325746 2726 | 846 2 2021 | 1380.7730915240899805971233211002179 58892 | 780
How many rides occurred in July 2021? How many rides occurred in January 2021?	*-- January and July ride counts* **Select** Extract(Month from STARTED_AT) ride_month, Count(*) ride_count **From** C##MISM6210.DIVVY **Where** Extract(Year from STARTED_AT) = 2021 **and** Extract(Month from STARTED_AT) in (1, 7) **Group by** Extract(Month from STARTED_AT);	 RIDE_MONTH | RIDE_COUNT 1 | 1 | 96544 2 | 7 | 820980

SQL Analysis 2/2

Dataset Selection

Divvy Bikeshare

Dataset	Query	Purpose
Summary of Divvy rides by member type and start / end station combination	*-- Ride summary by station combination* **Select** MEMBER_CASUAL, START_STATION_NAME, END_STATION_NAME, sum(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60 total_ride_time, avg(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60 avg_ride_time, count(*) ride_count **From** C##MISM6210.DIVVY **Group by** MEMBER_CASUAL, START_STATION_NAME, END_STATION_NAME;	This summary of the data will allow us to compare Divvy ride statistics to relevant station, location and population data
Divvy Station Master Data	[Table Provided]	The station master data provides addresses and dock counts to analyze station capacity and population served
Chicago Commuter Survey	[Table Provided]	The Chicago commuter survey provides a population of cyclists to compare to proximate Divvy ride volume and station capacity

Cyclist Population Location Analysis

Dataset	Query	Purpose
Daily summary of Divvy rides by member type	-- *Daily ride summary by member type* **Select** MEMBER_CASUAL, to_char(STARTED_AT, 'yyyy-mm-dd') ride_date, sum(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60*60 total_ride_time, avg(cast(ENDED_AT as date) - cast(STARTED_AT as date))*24*60*60 avg_ride_time, count(*) ride_count **From** C##MISM6210.DIVVY **Group by** MEMBER_CASUAL, to_char(STARTED_AT, 'yyyy-mm-dd');	This summary of the data will allow us to compare Divvy ride statistics to weather statistics for the given day
Chicago Weather Data	[Table Provided]	The Chicago weather data provides information about temperature, wind, and rain to correlate with ride statistics

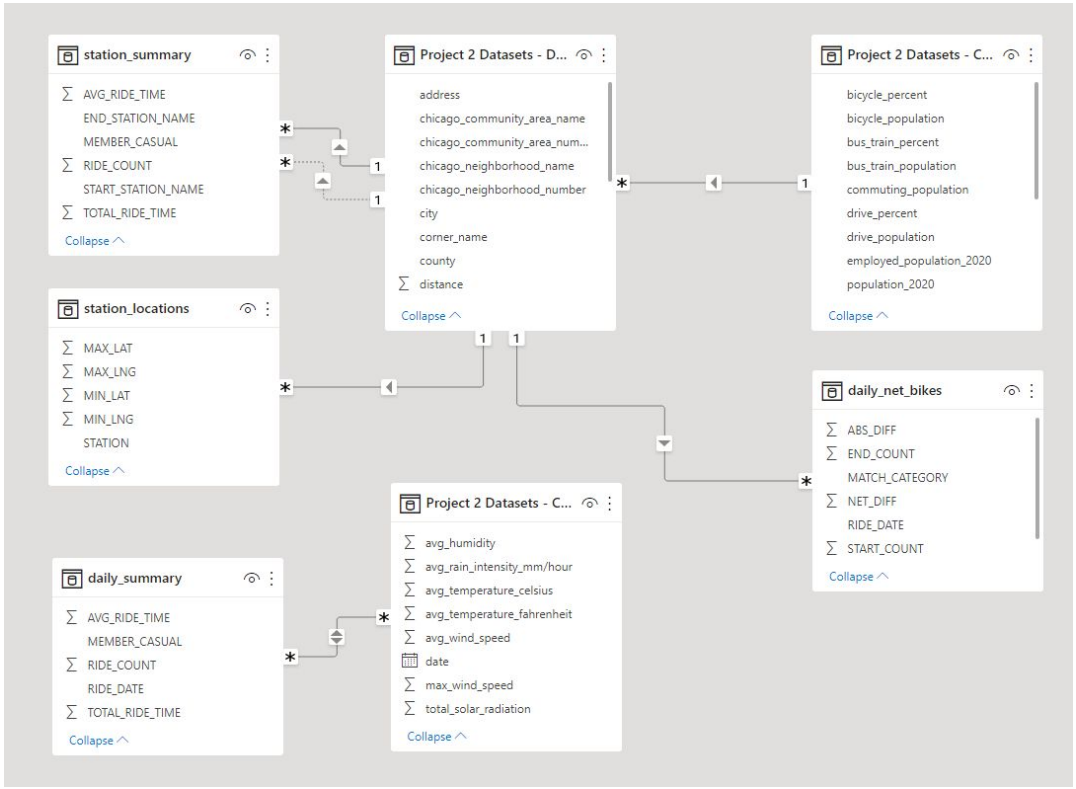
Cyclist Population Location Analysis

Dataset	Query	Purpose
Summary of Divvy station start and end ride volume by day	*-- Daily Combined summary of bike movements* **Select** Case When st.ride_date is Null Then ed.ride_date Else st.ride_date End ride_date, Case When st.START_STATION_NAME is Null Then ed.END_STATION_NAME Else st.START_STATION_NAME End station_name, nvl(st.start_count,0) start_count, nvl(ed.end_count,0) end_count, (nvl(st.start_count,0) - nvl(ed.end_count,0)) net_diff, abs(nvl(st.start_count,0) - nvl(ed.end_count,0)) abs_diff, Case When nvl(st.start_count,0) - nvl(ed.end_count,0) = 0 Then 'Net Even' When nvl(st.start_count,0) - nvl(ed.end_count,0) > 0 Then 'Net Start' Else 'Net End' End match_category **From** (**Select** to_char(STARTED_AT, 'yyyy-mm-dd') ride_date, START_STATION_NAME, count(*) start_count **From** C##MISM6210.DIVVY **Where** START_STATION_NAME is not Null **Group By** to_char(STARTED_AT, 'yyyy-mm-dd'), START_STATION_NAME) st **Full Outer Join** (**Select** to_char(STARTED_AT, 'yyyy-mm-dd') ride_date, END_STATION_NAME, count(*) end_count **From** C##MISM6210.DIVVY **Where** END_STATION_NAME is not Null **Group By** to_char(STARTED_AT, 'yyyy-mm-dd'), END_STATION_NAME) ed **On** st.START_STATION_NAME = ed.END_STATION_NAME **and** st.ride_date = ed.ride_date **Order By** abs_diff desc;	This summary of Divvy rides allows us to observe the net movement of bikes, which can be compared to dock counts for surplus or shortages and help inform required redistribution
Divvy Station Master Data	[Table Provided]	The station master data provides dock counts to compare to bike movements

Relationship Summary

The datasets were related by 3 main keys:

1. Station Name
2. Zip Code
3. Ride Date

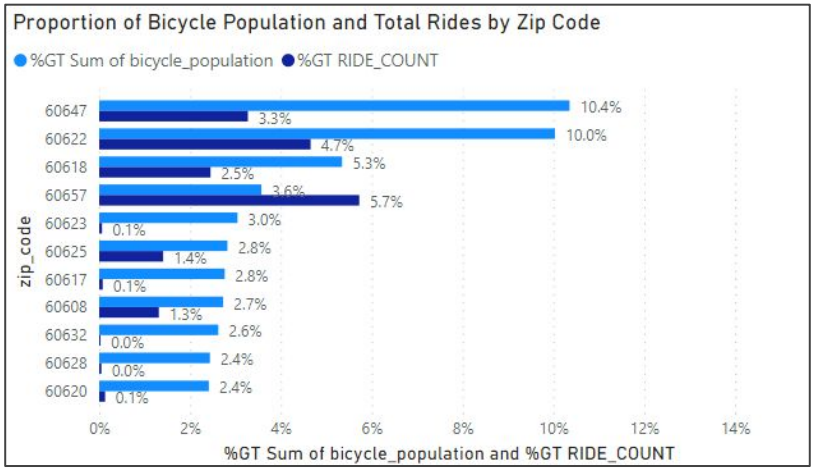


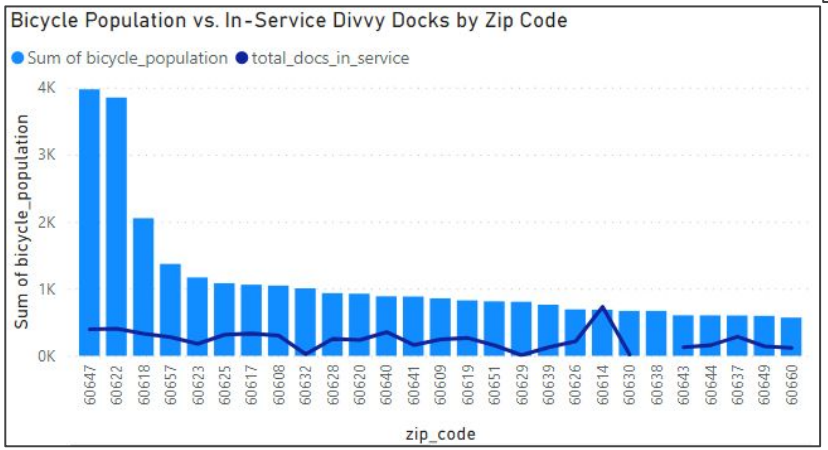
Power BI Visalization

Divvy Bikeshare

Bring the Bikes to the People

As shown in the chart on the right, zip codes with a large population of cyclists do not consume an equivalent proportion of Divvy bike rides. **Why might this be?**

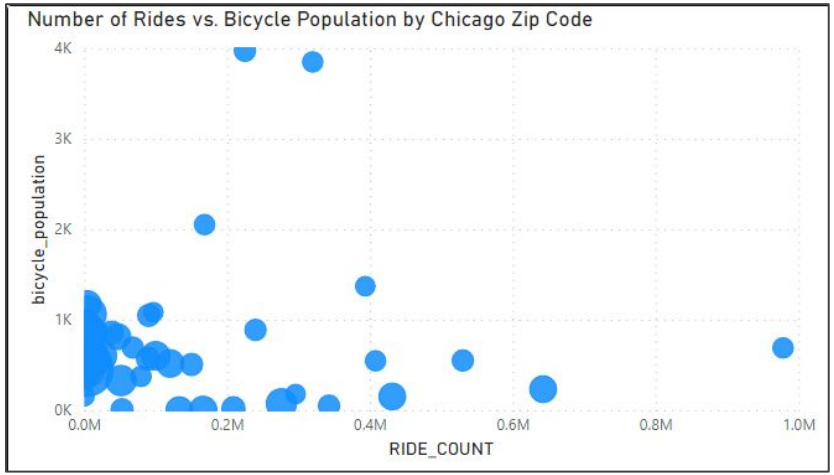


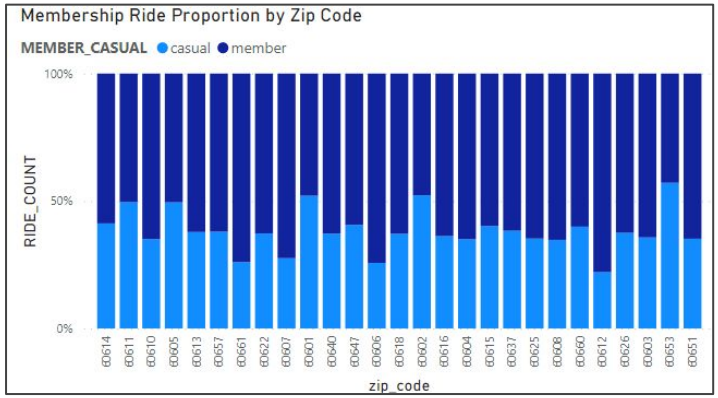


The chart on the left shows that these same zip codes with the highest proportion of the cycling population actually have an equal or lesser number of available Divvy bike docks. With more cyclists and fewer docks, they are less likely to be satisfied with Divvy's service and to become or remain members.

So Many Cyclists, So Few Rides

To further validate this mismatch in the population, I reviewed the absolute relationship between cyclist population and total rides in a zip code. While this is a vague positive correlation, it is much weaker than expected.

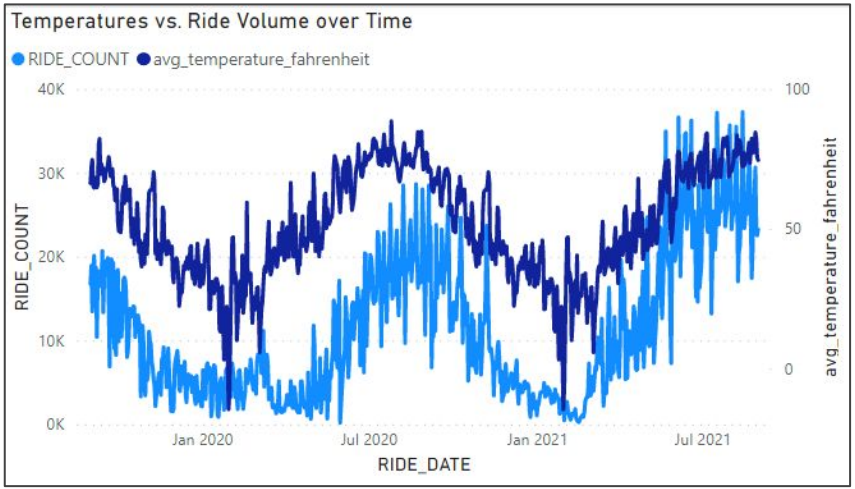


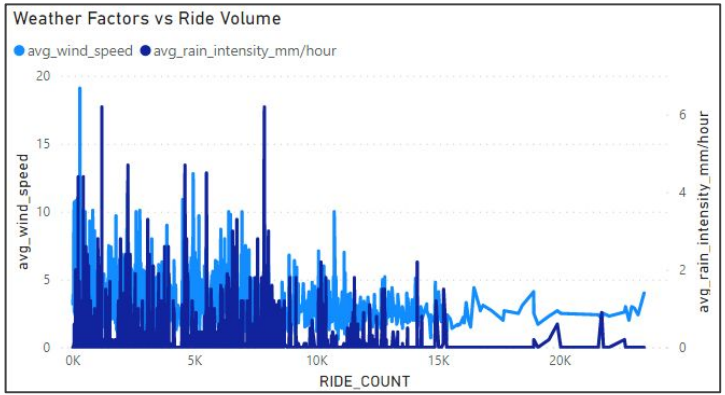


One final validation of this theory is the proportion of rides by members within a zip code - many of the locales with the highest bicycle population still trend low in proportion of member rides.

When it rains, it pours

Outside of the mismatch between bike placement and rider population, Divvy ride volumes are significantly influenced by seasonality and weather conditions. In the chart on the right, we see large drops in rides during colder winter months.

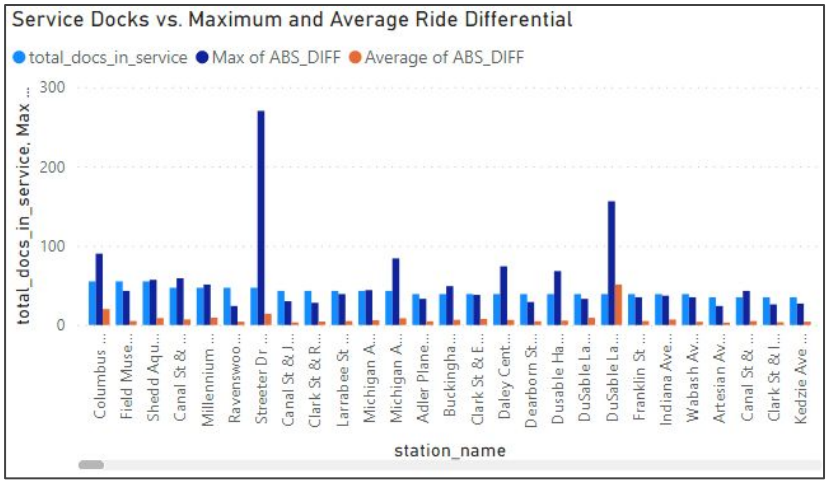


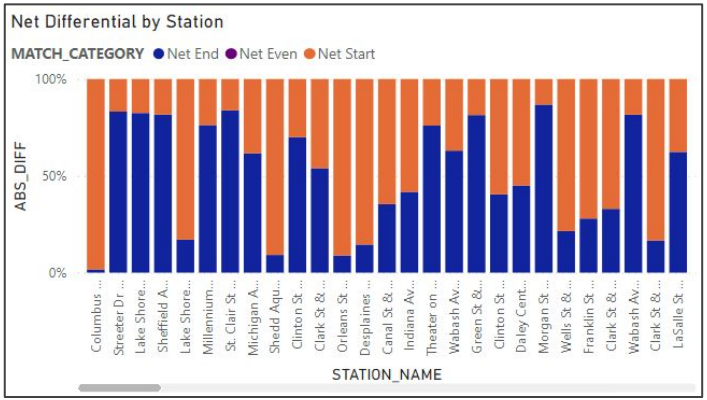


In addition to pure temperature correlation, it is clear from the chart at left that higher wind speed and rain intensity are correlated with lower ride volume. While Divvy certainly can't influence the weather, they can plan to increase efforts to rebalance bike inventory on fair weather days.

Time to get moving

Divvy's bike equation is composed of 2 key factors: having bikes stationed near their user base and keeping the bikes distributed across those stations. By summarizing ride starts and ends, we can clearly see imbalances across stations.





The chart above shows maximum and average start / end difference compared to station capacity, highlighting mismatches. On the left, it is clear that many stations are most frequently starting or ending destinations. Divvy can target the stations with the largest imbalances to redistribute the bikes.

Let’s Ride

In summary, my analysis has revealed the following:

- Divvy stations and bike docks in service are not distributed in the same locations as the self-identified bicycling population
- Divvy rides are highly correlated with seasonality, primarily due to the impact of temperature drops and increases in wind and rain intensity
- Net ride differentials often exceed station dock capacity, which will result in riders without bikes or bikers without docks to drop their bikes

Divvy should develop a forecasting model for bike usage based on historical ride data, increase station capacity near key cycling bases, strategically redistribute bikes and offer promotions on annual memberships.