OPTIMIZING RIDE SHARING ALLOCATION

Analysis on Ridership for Transportation Network Providers(TNP) In Chicago

PRESENTED BY:
AMILY HUANG, DUO ZHOU AND YIHENG ZHU

MEET THE TEAM

Duo Zhou

Amily Huang

Yiheng Zhu

zd0009@uchicago.edu

yunh@uchicago.edu

yihengzhu@uchicago.edu

CONTENT

EXECUTIVE SUMMARY

EXECUTIVE SUMMARY

- Our goal is to optimize vehicle allocation according to ridership demand and give business insights and recommendations to TNP
- Use weather, sport events, crime, and census data to analyze customer behavior and to give precise recommendations
- We focused on relational database system which can efficiently load, store, and extract data from different sources for analysis (OLAP)

BUSINESS USE CASE

Actor	Incentive	Business Use
TNP	• To better understand customer behavior • Vehicle allocation Optimization • Improve customer service management	Understand how different factors impact ridership and customer behavior
Driver	• Maximize income per unit time	Understand how tips vary in different circumstances
Customer	• Time-efficient and safer service	Improvement of service experience with better vehicle allocation and customer service

DATA INGESTION & PREPARATION

SOLUTION OVERVIEW: DATA / TOOLS

Main dataset:

Transportation Network Providers
- Trips by Location, Distance, Day/Time, Tips, etc.

Supporting dataset:

Weather
- Historical Weather Conditions, Short-term Forecasts

Geography
- Community area boundaries in Chicago

Sports Event
- Chicago sports team (NBA,NFL,MLB) schedules

Census & Crime
- Income, Education, Crime Rate by Chicago Community Area (CCA)

Data Connectors:

City of Chicago Data Portal
- CSV batch download

ESPN
- Python web scraping

NoAA
- API

Data Processing/ Storage:

Python
- Data Cleaning and Processing

MySQL(GCP)
- Data storage and model design

Excel
- Data Cleaning and Processing

UChicago RCC
- Cloud Computing for Large Datasets

Visualizations:

Python
- matplotlib
- seaborn
- ggplot

Tableau

Analytics:

Python
- sklearn
- fbprophet
- pandas
- scipy
- numpy

DATA PREPARATION

Data Source	Format and Size	Processed Data That Meet Analytical Needs	Platform and Tools Used
Chicago Data Portal: Transportation Network Providers	30 GB (129 Million Rows, 21 Columns) Structured CSV File	Ridership, Avg Traveled Distance, Avg Tips and Number of Pooled trips Group by CCA, Date and Time	Python, MySQL GCP and RCC
Chicago Data Portal: Boundaries - Chicago Community Area (CCA)	1.92 MB Structured CSV File	MULTIPOLYGON Data by CCA For Tableau	Python and Excel
Census Data By CCA	1.1 MB Structured CSV File	Education, Age, Income, Population & Unemployment Rate etc. by CCA	Python, MySQL
Chicago Data Portal: Crimes	16 GB (7.12 Million Rows, 22 Columns) Structured CSV File	Total Number of Crimes By CCA	Python, MySQL GCP and RCC
ESPN	2 MB Unstructured Data: From Web Scraping	Only the home game dates and location	Python and Excel
National Centers for Environmental Information	5.5 MB Structured CSV File	Avg daily temp, total daily precipitation and avg daily wind speed by date.	Python and Excel
Wikipedia: Community Areas in Chicago	0.1 MB Unstructured Data: From Web Scraping	Chicago community area code	Python and Excel

DATA MODELING & DESIGN

DATA MODELING

Compiling Data into Tables

- Use MySQL Workbench to create our dimensional tables
- Sports and Weather tables are indexed by Date (primary key)
- Census and Crime tables are indexed by Chicago Community Areas(CCA) (primary key)
- Date and CCA are both linked to the fact table as foreign key

Data Transformations

- Data are transformed into a rows and columns format with appropriate data type
- Main dataset ridership measures are aggregated and grouped by CCA, date and time
- Sports schedule datasets, and weather datasets are aggregated and grouped by date
- Census and Crime datasets are aggregated and grouped by CCA
- CCA Geographic boundaries data are transformed to meet tableau virtualization requirement

Data Mapping

- One main dataset (fact table with ridership measures) and four supporting datasets (dim tables)
- *Star type dimensional model* is adapted by linking 4 dimensional datasets to the main fact dataset using either DATE or CCA
- *Note: Ridership By Hours of A Day is an independent analytical entity that provides additional business insights on ridership, tip and Shared Trips*

DESIGN CONSIDERATIONS

Data Types

- id: INT
- Date: DATE
- CCA: INT
- MEAN & MEDIAN attributes: DOUBLE
- Others: INT

Dealing with NAs

Weather:

- Drop NA rows

Main Dataset:

- Fill NA with 0

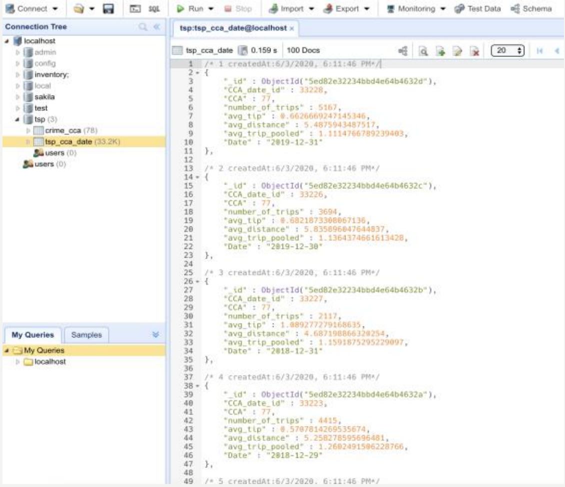
Using Dimensional Tables

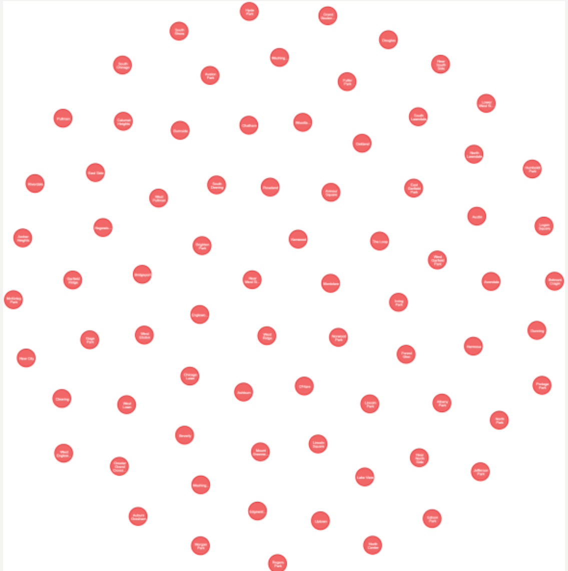
- Maintain historical information for all dimensions
- Less processing time and higher performance

Expected Output of Data Analysis (Data Quality Metrics)

- The relationship between census attributes and fact measures
- The impact on fact measures from different weather factors
- The impact on fact measures during major sports event
- The relationship between public safety and fact measures

NOSQL CONSIDERATION





MongoDB

- Can create document based OLTP database with each real-time trip as a transaction
  - [JSON File] Each object contains measures from fact table and sub-arrays with information from dimension tables

Neo4j

- Can create graphic based OLTP with each real-time trip as a transaction
  - [Graphic Nodes] Each node contains information of a trip or a dimension

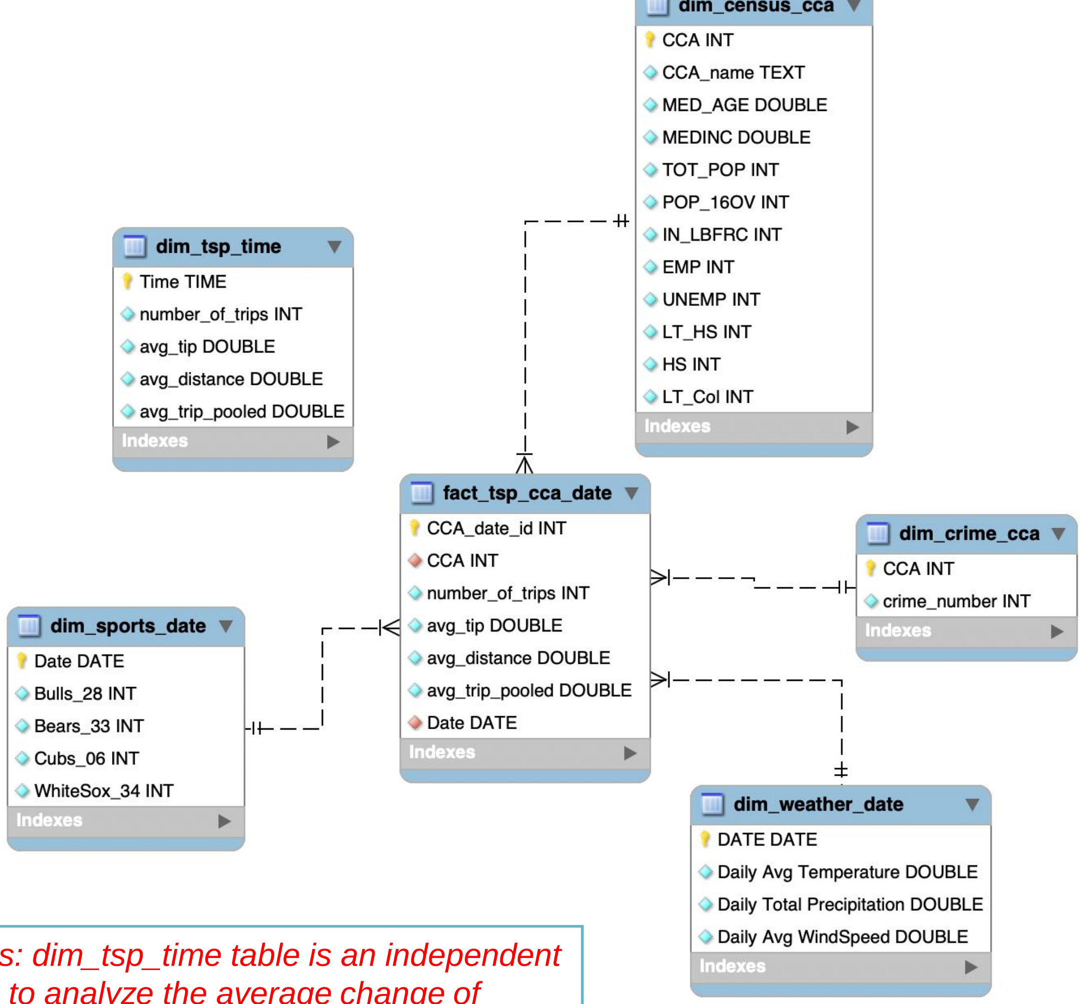
Advantages

- Easily store new data (update quarterly)
- Flexible Schema

DATA QUALITY DIMENSION

- ✓ Completeness: Missing values in weather dataset treated as zero
- ✓ Validity: Data is transformed into fact and dimensions to meet our analytical need
- ✓ Uniqueness: No duplicated data
- ✓ Consistency: Data format is consistent throughout the database
- ✓ Timeliness: Data represent reality in time as data consist all the entries over the period considered
- ✓ Accuracy: Data is simply aggregated by summing and averaging over locations and dates, and this transformation can represent reality

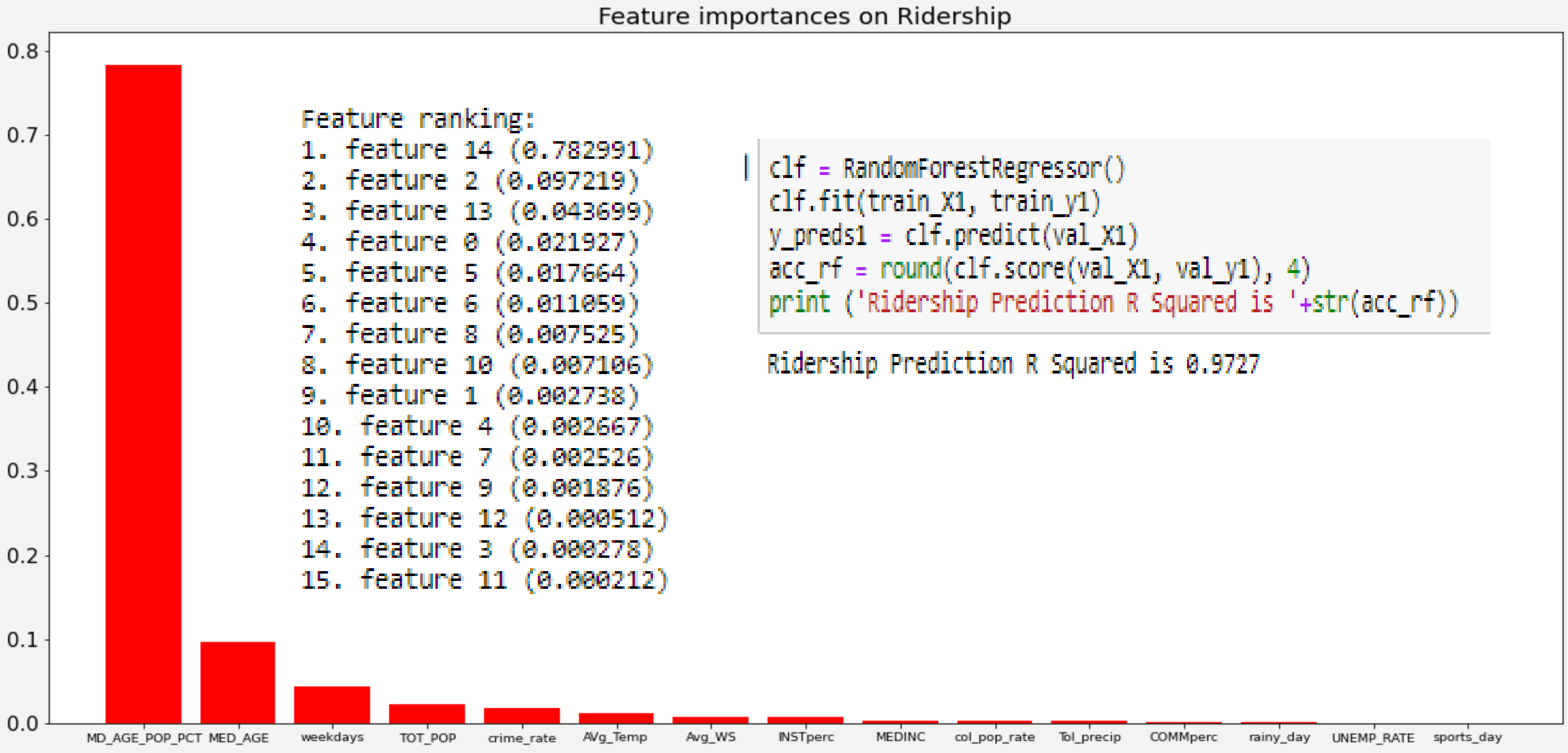
ENHANCED ENTITY RELATIONSHIP DIAGRAM



Notes: dim_tsp_time table is an independent table to analyze the average change of measures from hour to hour within a day.

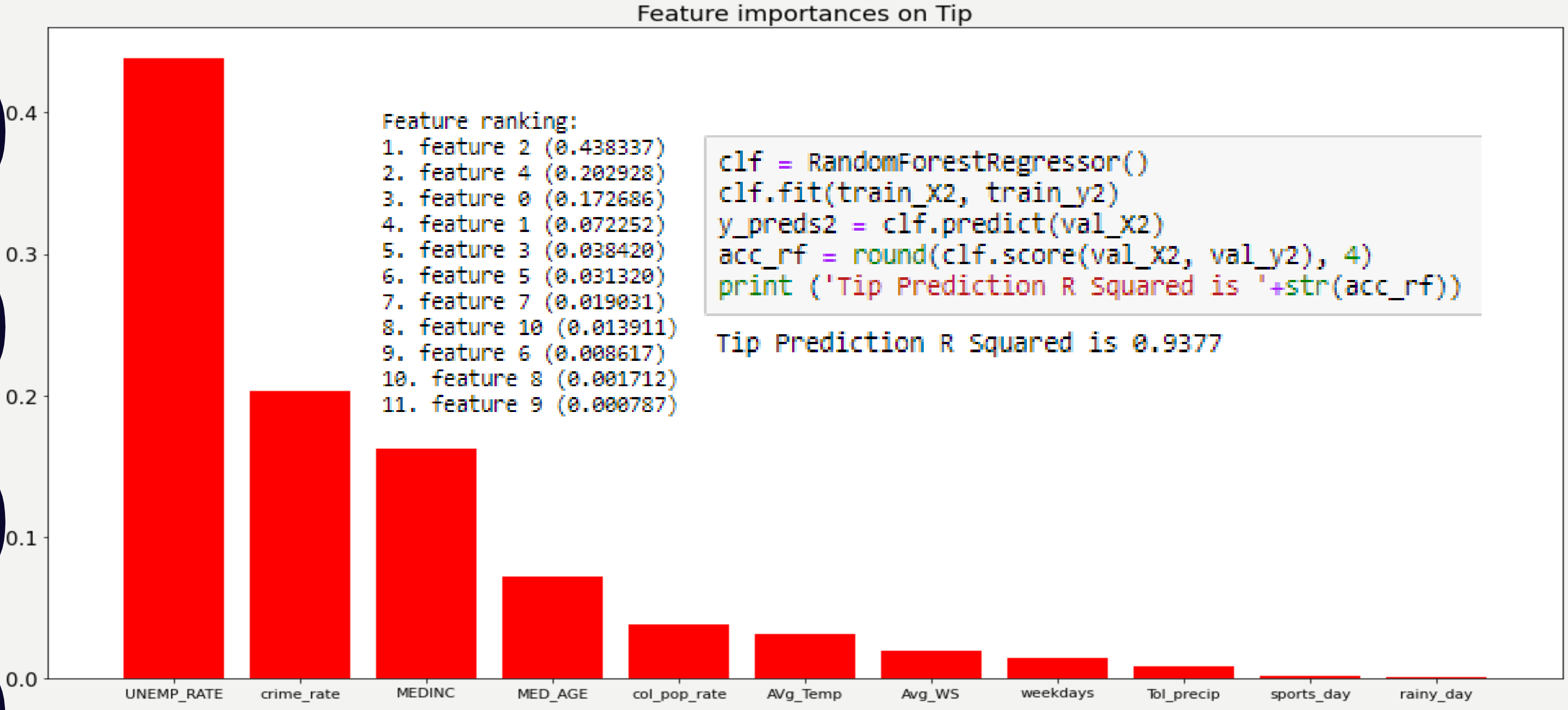
ANALYTICS & RESULTS

Random Forest Regressor



Important Features

Random Forest Regressor

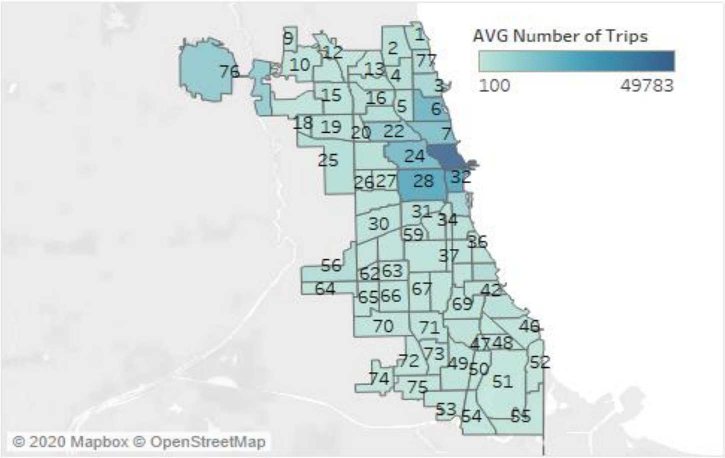



```
clf = RandomForestRegressor()
clf.fit(train_X2, train_y2)
y_preds2 = clf.predict(val_X2)
acc_rf = round(clf.score(val_X2, val_y2), 4)
print ('Tip Prediction R Squared is '+str(acc_rf))
```

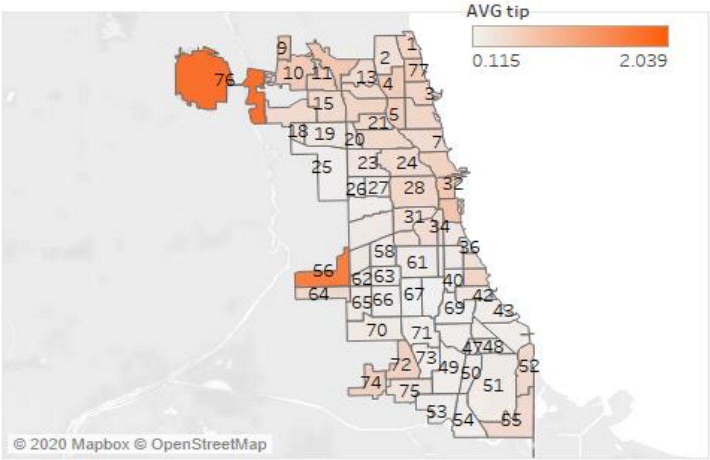
```
Tip Prediction R Squared is 0.9377
```

Important Features

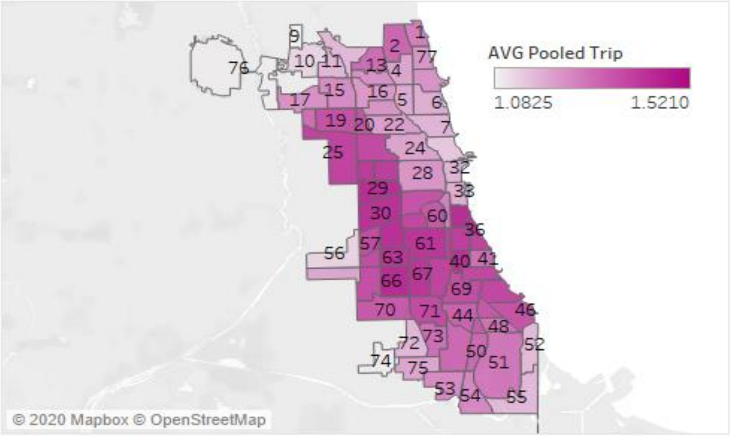
Ridership Distribution By Chicago Community Area



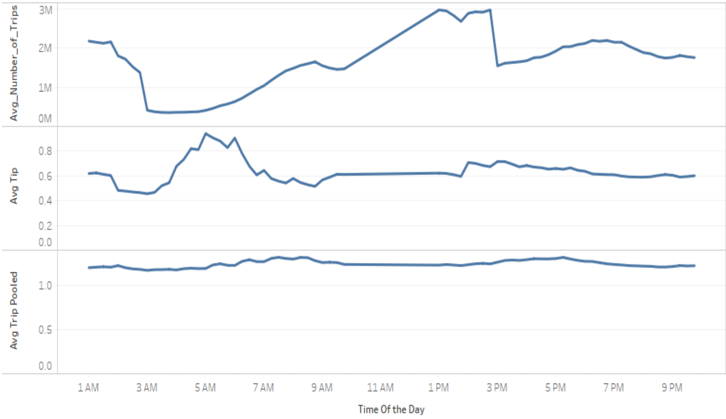
Tip by Chicago Community Area



Pooled Trip By Chicago Community Area

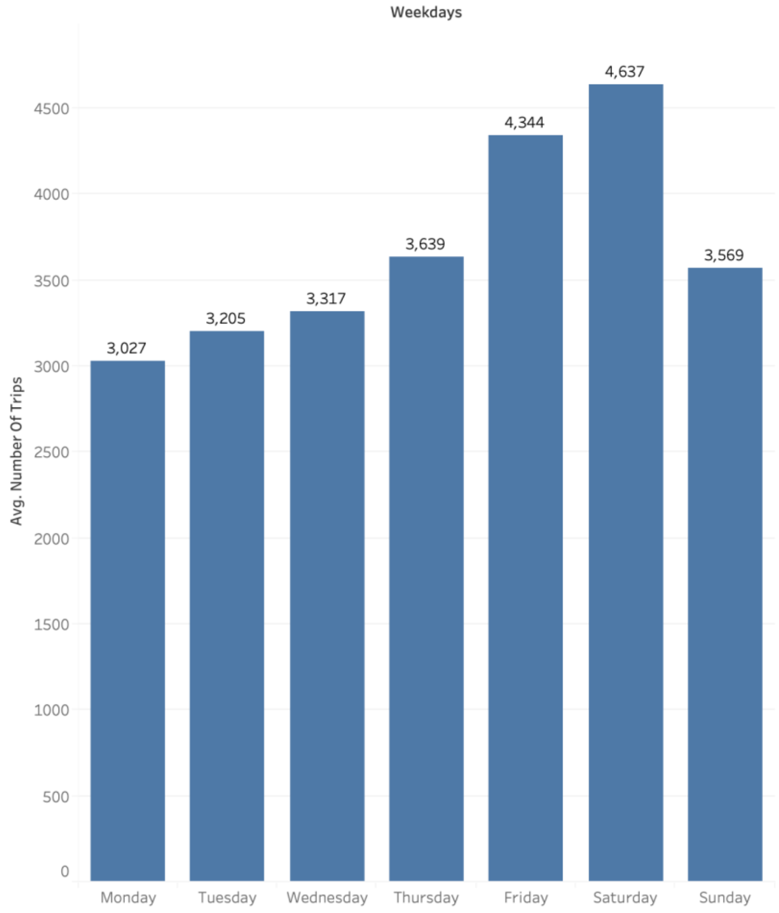


Average Measures Change Over 24 Hr Period

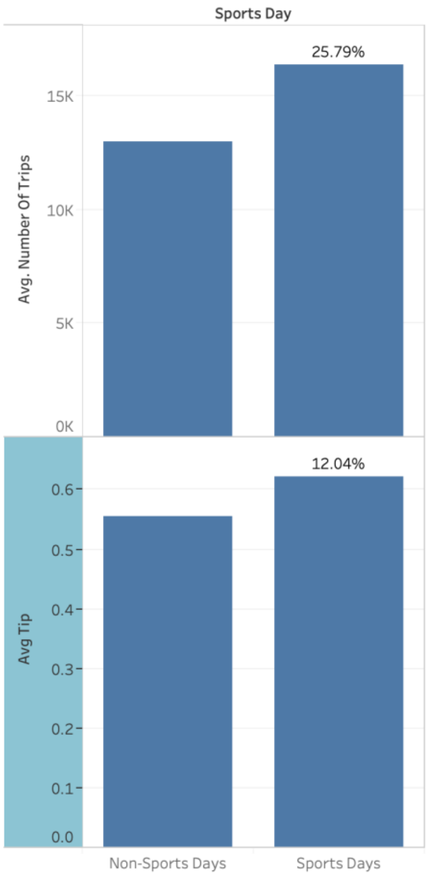


The trends of average of Number Of Trips, average of Avg Tip and average of Avg Trip Pooled for Time.

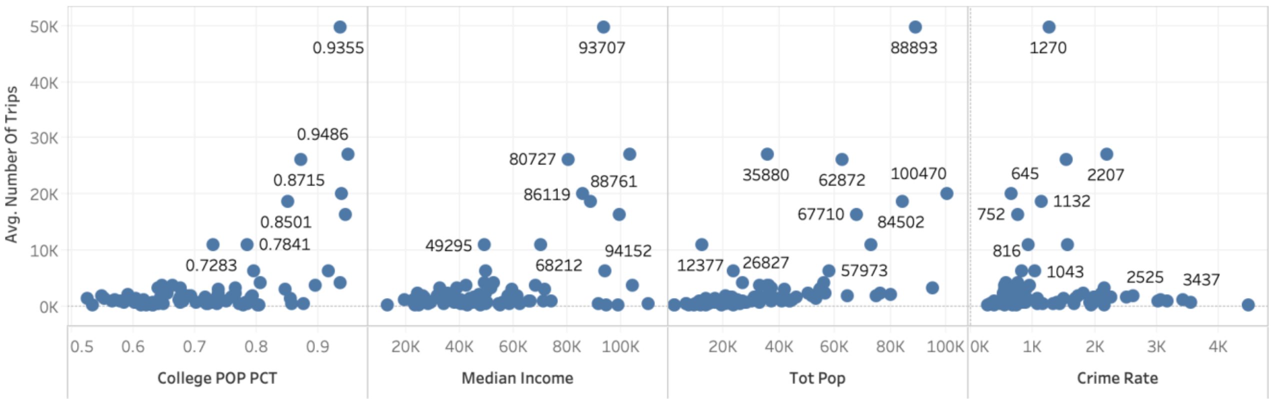
Ridership vs Weekdays



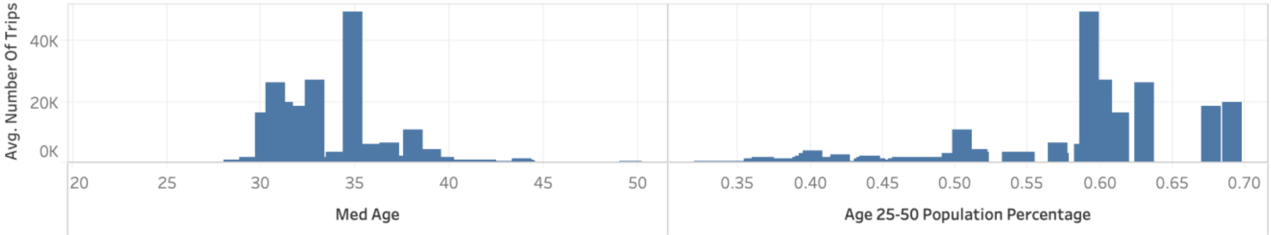
Effect of Sports Events on Ridership and Tip



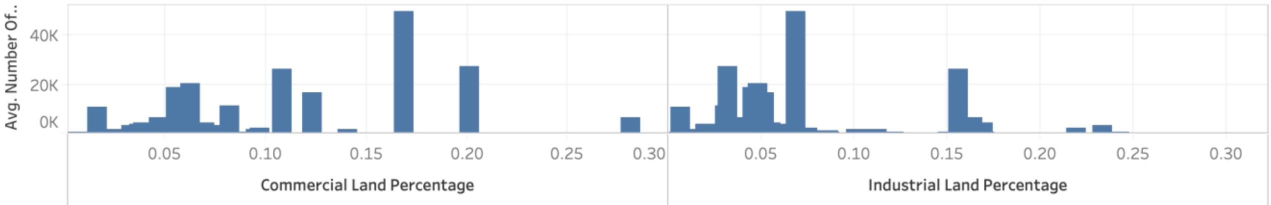
Education/Income/Tot population/Crime Rate on Ridership



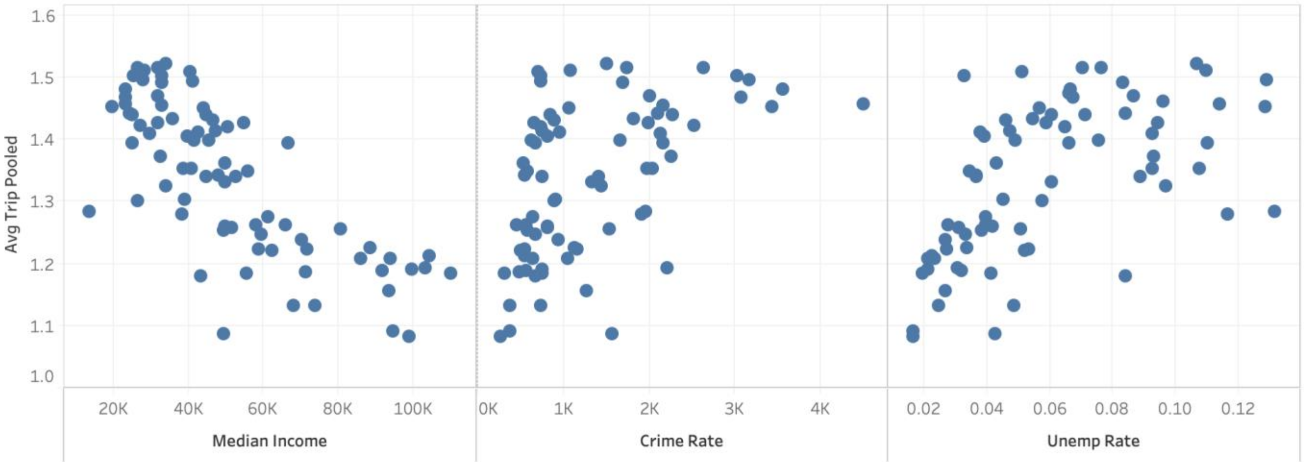
Age on Ridership



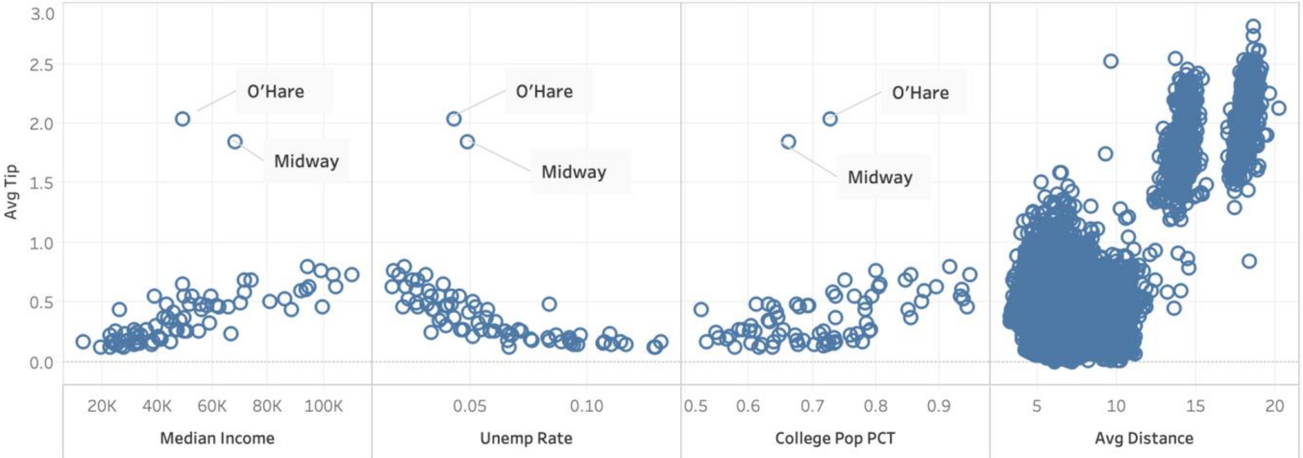
Ridership vs Commercial/Industrial Land Percentage

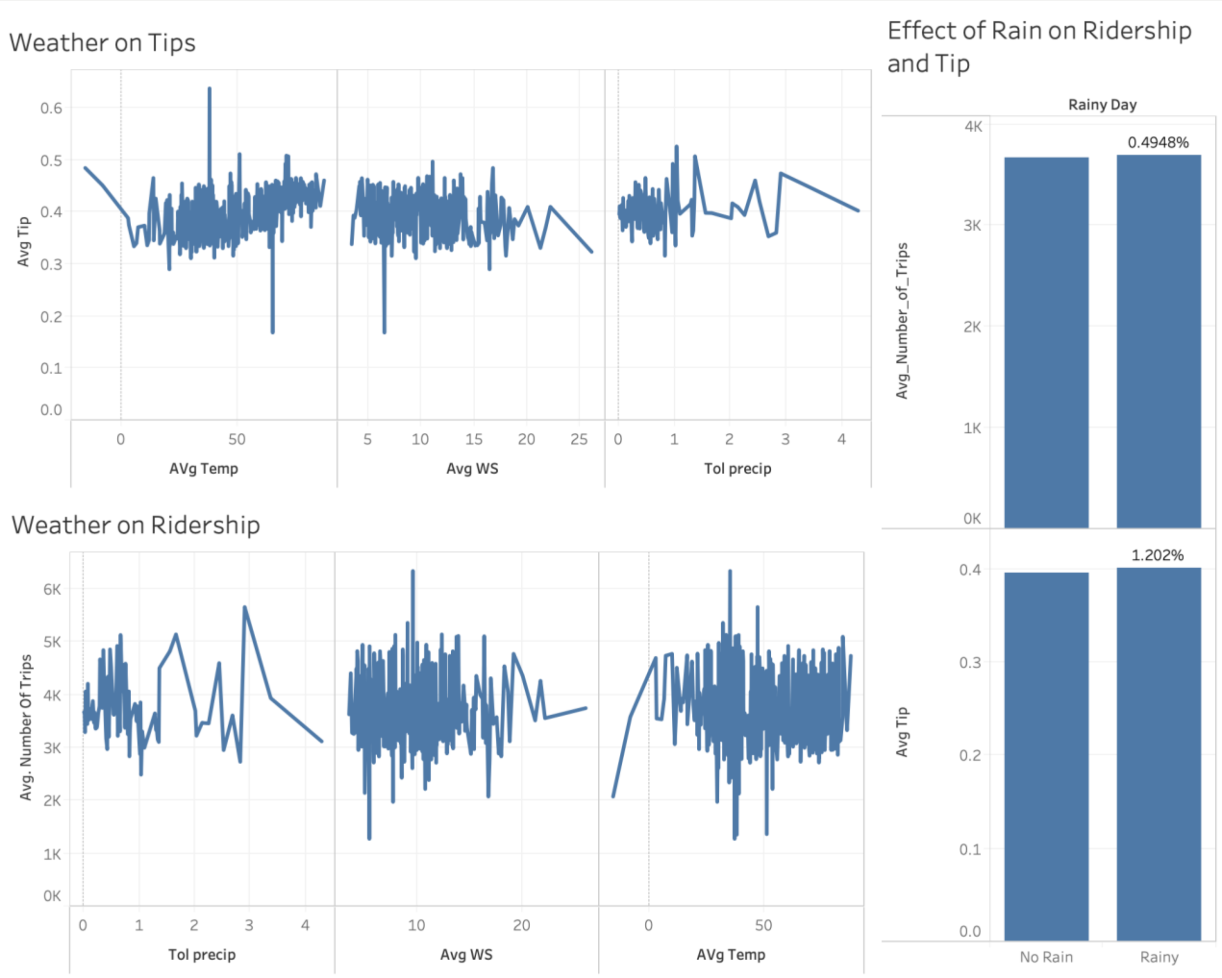


CrimeRate/Unemp/MedINC on Pooled Trips



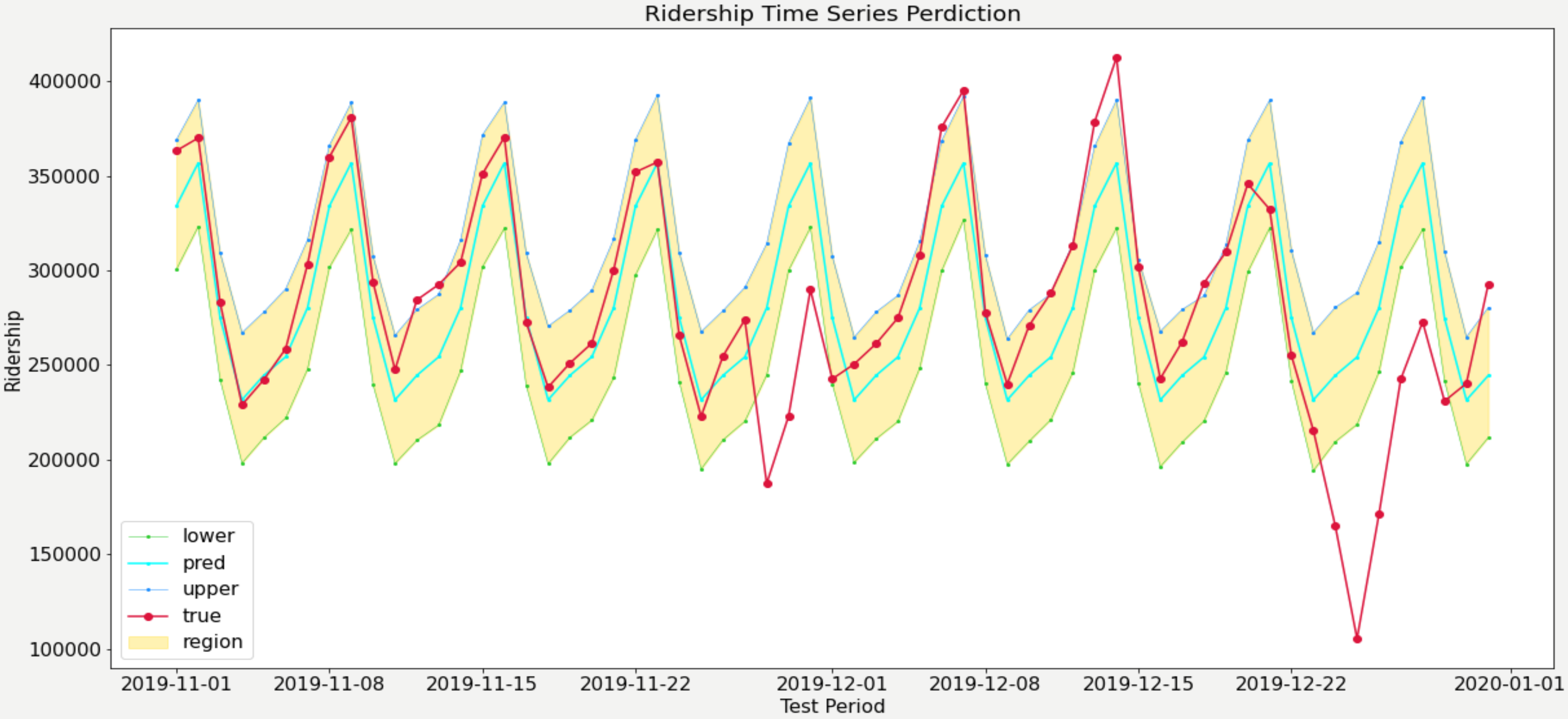
Tips




Weather on Tips
Avg Tip
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
50
AVg Temp
5
10
15
20
25
Avg WS
0
1
2
3
4
Tol precip
Weather on Ridership
Avg. Number Of Trips
6K
5K
4K
3K
2K
1K
0K
0
1
2
3
4
Tol precip
10
20
Avg WS
0
50
AVg Temp
Effect of Rain on Ridership and Tip
Rainy Day
Avg_Number_of_Trips
4K
3K
2K
1K
0K
0.4948%
Avg Tip
0.4
0.3
0.2
0.1
0.0
1.202%
No Rain
Rainy

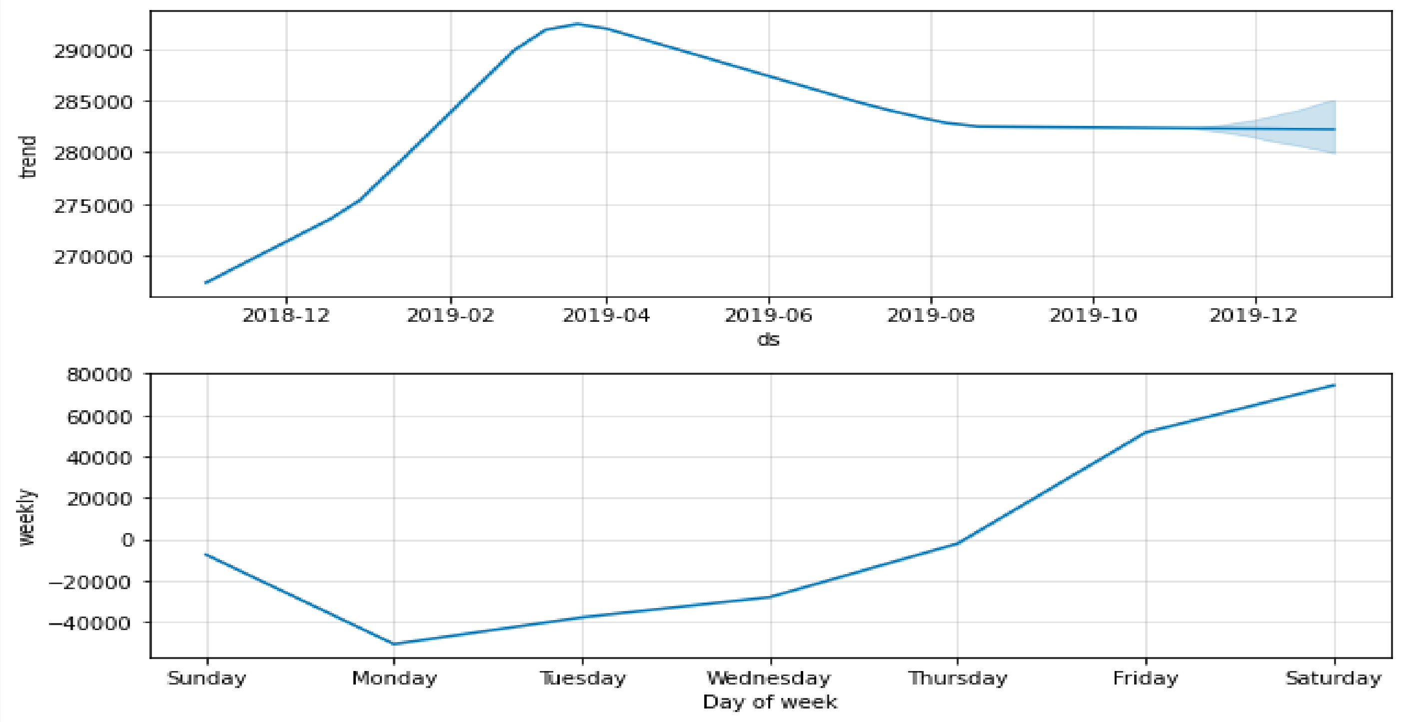
Time Series Forecast

(Facebook Prophet)



Time Series Forecast

(Facebook Prophet)



- Facebook Prophet produces same weekly pattern as our analysis

CONCLUSION
& RECOMMENDATIONS

CONCLUSION / RECOMMENDATION

Ridership and Tips are independent of weather.

Airport customers tend to tip exceptionally better and travel longer distances.

CCA with better economic and education statistics tend to utilize more services and tip better.

- TNP companies should allocate more vehicles in the business and college districts.

Ridership increases towards weekend and peaks on Fridays and Saturdays.

Ridership also increases through morning and peaks in the mid-day.

- Encourage drivers to participate more during weekends and mid-day by lowering commission fee.

CONCLUSION / RECOMMENDATION

Customers between age 25 and 40 are the biggest base for TNP.

- Marketing Campaign should target on this age group
- Increasing TNP awareness among other age groups.

On sports days, ridership increase about 25% in the home game CCAs.

- Encourage drivers to move to sports home arena CCAs on sports days

Neighborhoods with higher crime rates have worse economic and educational status, people tend to do more pooled trips.

- Increase wait time limit in high crime Neighborhoods for customer safety

FUTURE WORK/LESSON LEARNED

FUTURE WORK/LESSON LEARNED

Ridership forecast can be more accurate with data from 2020 taking the consideration of Covid-19.

Ridership and Tips should be analyzed with more attributes, like public transportation distribution and major facilities' location, etc.

Our initial intuition does not match the analysis result. Machine Learning and Statistical Analysis should be applied with more attributes to select a better set of predictors.

During discovery and data preparation phase, a closer analysis should be applied to determine whether there is enough information in the data to meet our analytical goals.

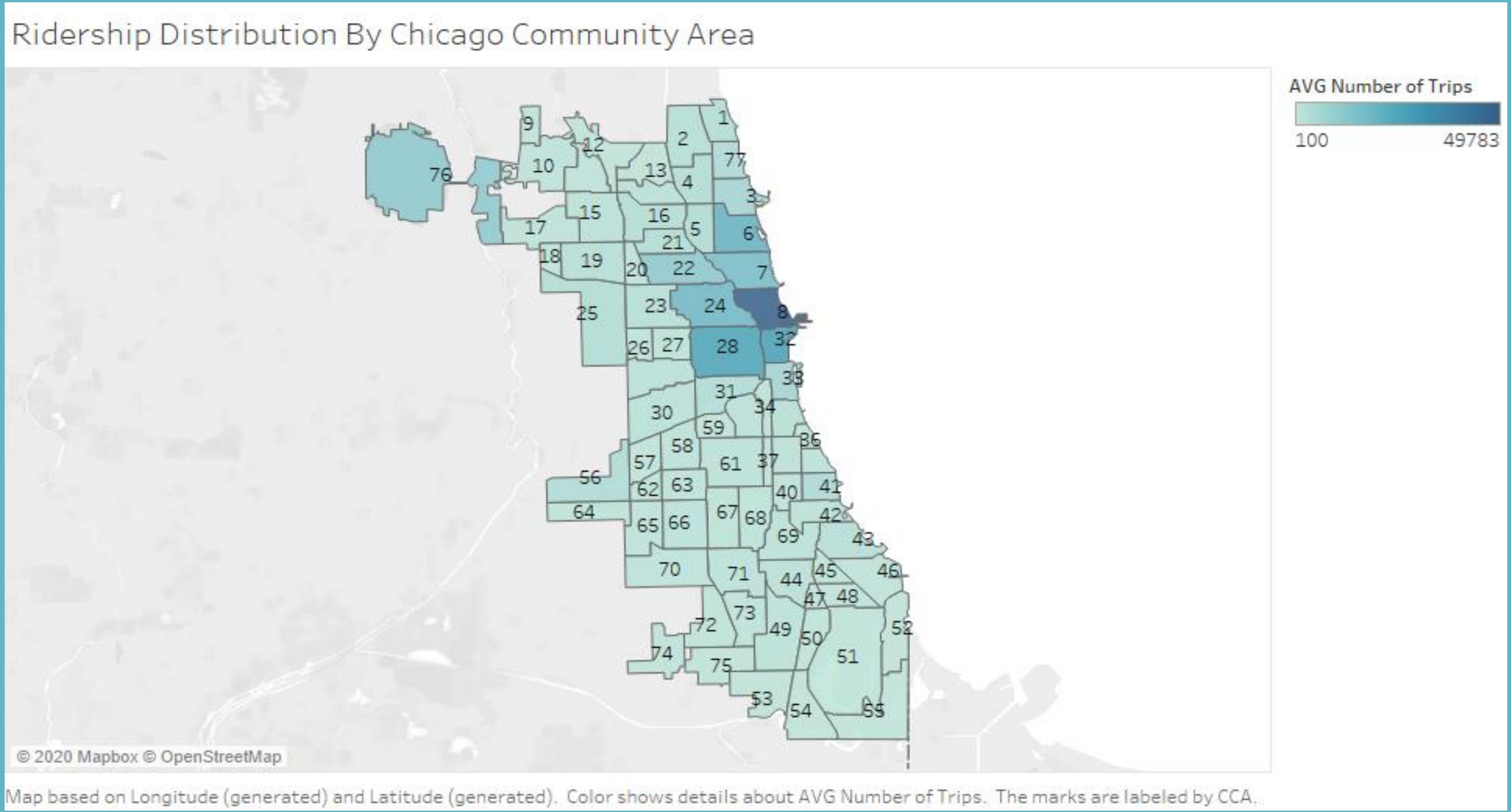
REFERENCE

- Jonathan Levy. (2018). Transportation Network Providers – Trips [Data File]. Available from Chicago Data Portal: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p
- Menne, Matthew J., Durre I., Korzeniewski B., McNeal S., Thomas K., Yin X., Anthony S., Ray R., Vose R., Gleason B., and Houston T. (2012). Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. [indicate subset used]. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ [access date]. Available from National Centers for Environmental Information web site: https://www.ncdc.noaa.gov/cdo-web/datasets
- City of Chicago. (2013). Boundaries - Community Areas (current). Available from Chicago Data Portal Web Site: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6
- Chicago Police Department. (2019). Crimes - 2001 to present [Data File]. Available from Chicago Data Portal Web Site: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2
- Chicago Metropolitan Agency for Planning. (2015). Community Data Snapshots Raw Data, June 2019 Release [Data File]. Available from CMAP Data Hub Web Site: https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data
- 2018-19 Chicago Bulls Schedule. (n.d.). Retrieved from https://www.espn.com/nba/team/schedule/_/name/chi/season/2019/seasontype/1
- Chicago Bears NFL - Bears News, Scores, Stats, Rumors & More. (n.d.). Retrieved from https://www.espn.com/nfl/team/_/name/chi/chicago-bears
- Chicago Cubs Baseball - Cubs News, Scores, Stats, Rumors & More. (n.d.). Retrieved from https://www.espn.com/mlb/team/_/name/chc/chicago-cubs
- Chicago White Sox Baseball - White Sox News, Scores, Stats, Rumors & More. (n.d.). Retrieved from https://www.espn.com/mlb/team/_/name/chw/chicago-white-sox

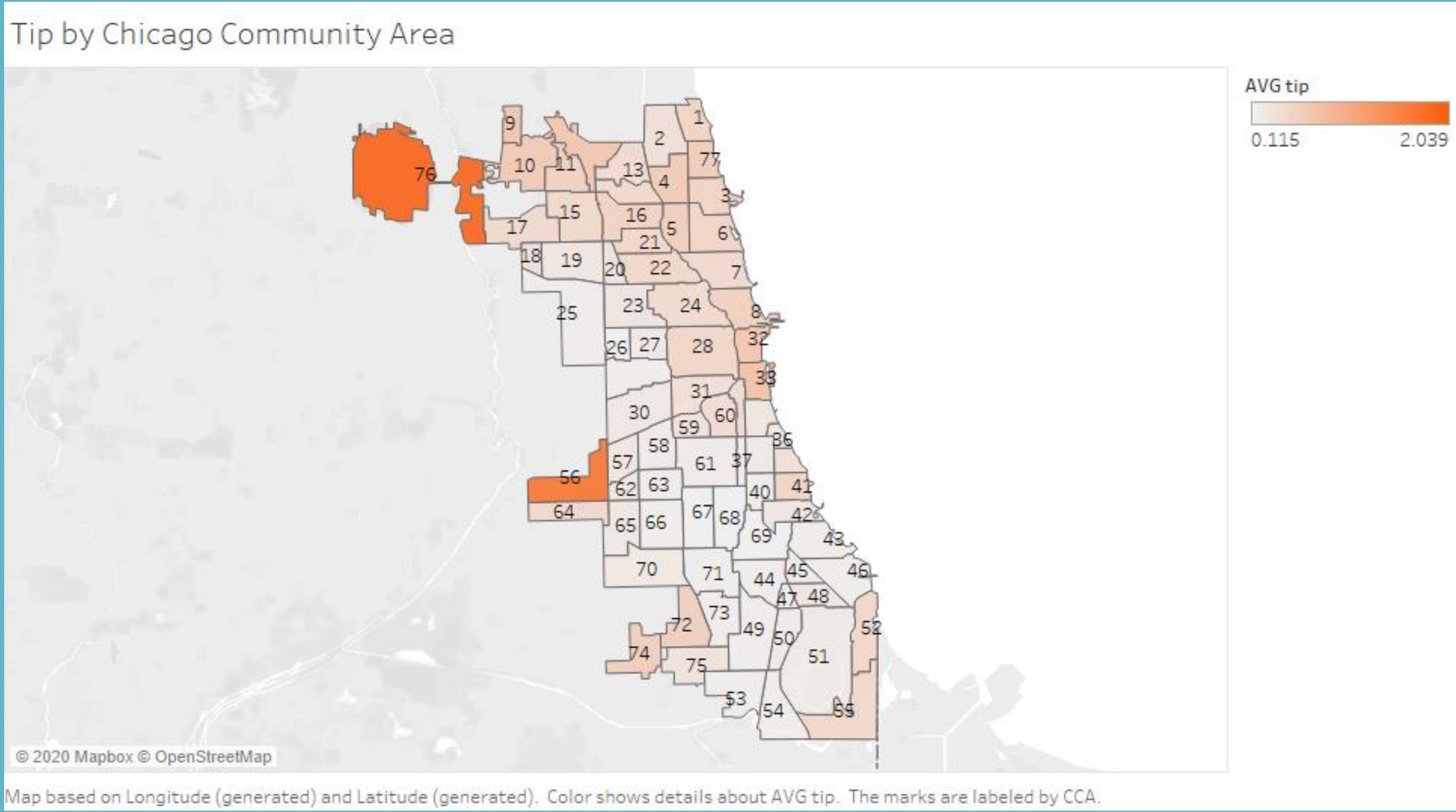
THANK YOU!

Q/A

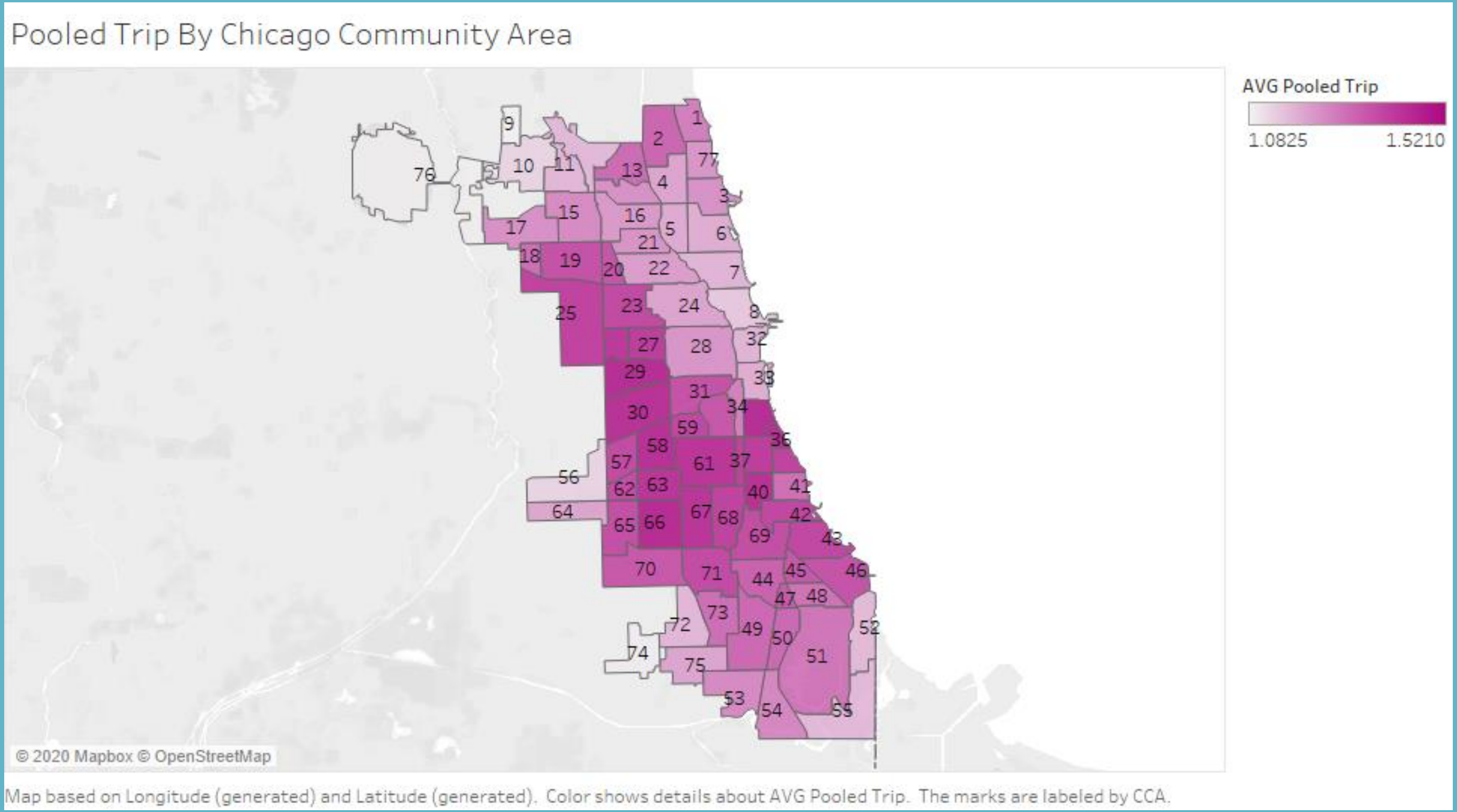
APPENDIX



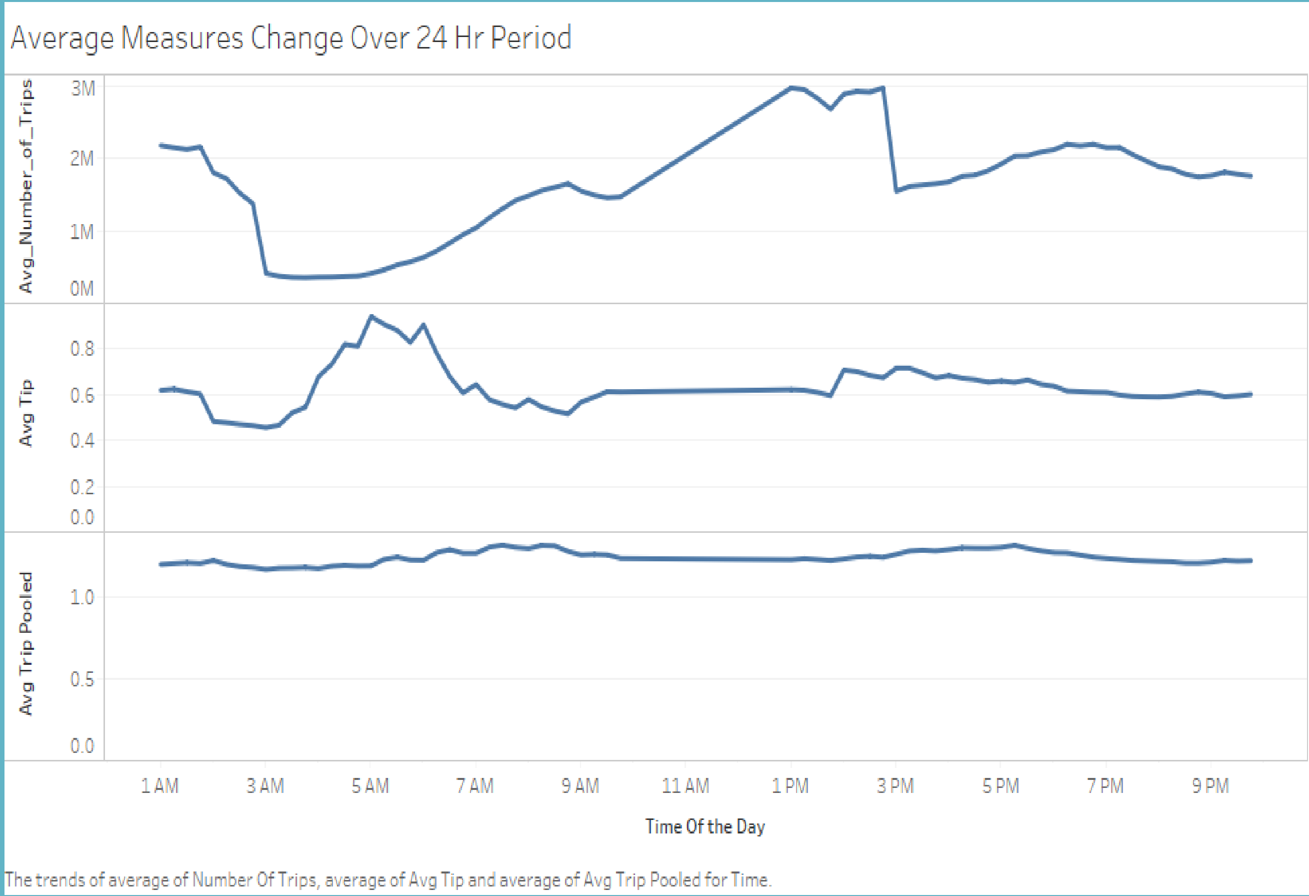
- People in business districts and surrounding neighborhoods tend to utilize the service more.



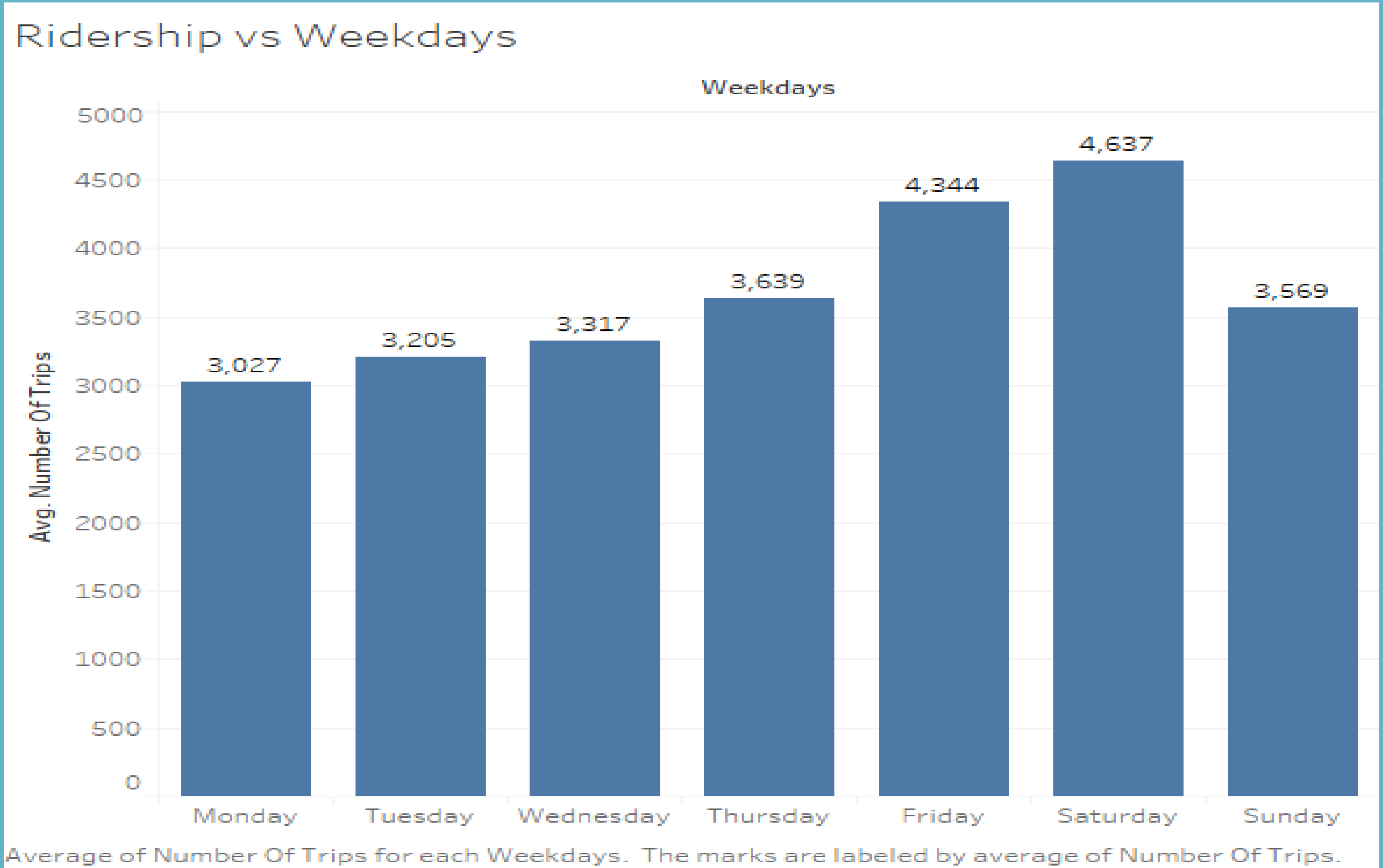
- Tip are higher in CCA where O'hare international airport and Midway airport are located
- Neighborhoods with higher income tip better



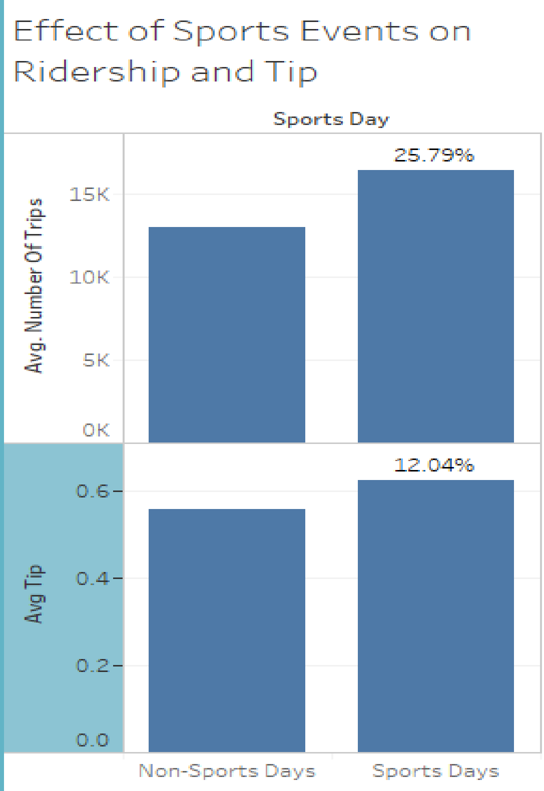
- Carpool are used more in west and south lower income neighborhoods of Chicago.



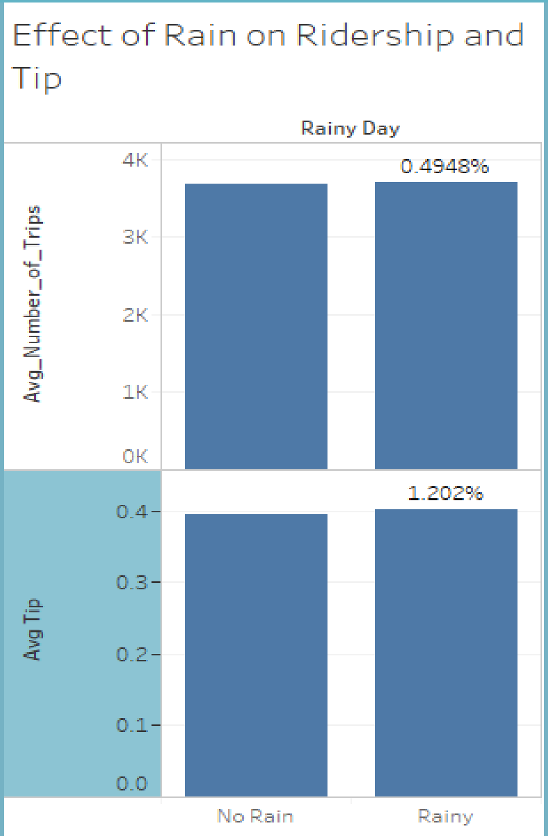
- Higher number of trips at daytime
- Rapid decline from 2-3pm might due to rush hour or data gap
- Higher tips from 5-7am, reason might be morning arrival or departure flight
- No significant difference for carpool over 24hr period



- Friday and Saturday have the most ridership due to the coming of weekends.

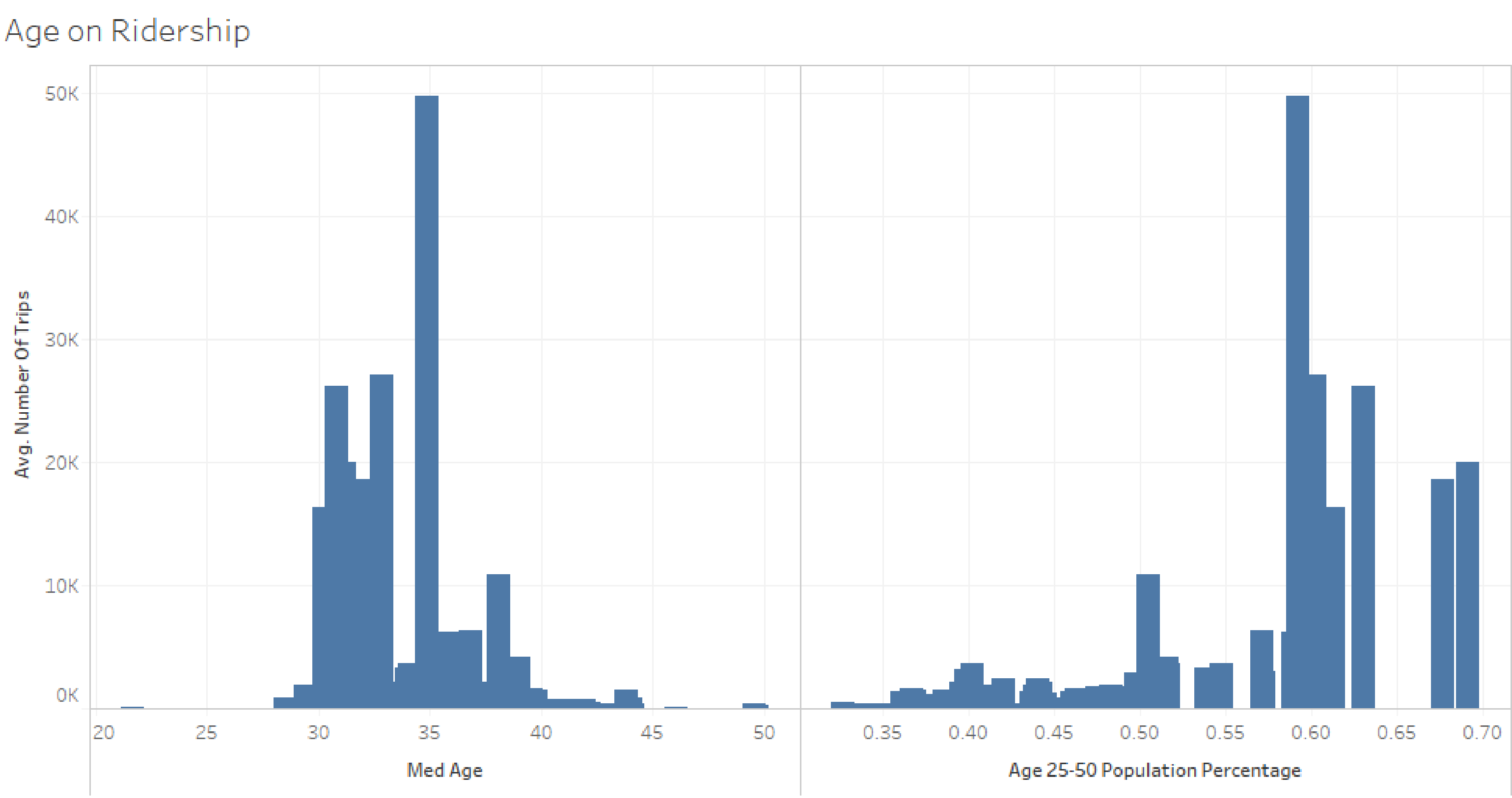


Average of Number Of Trips and average of Avg Tip for each Sports Day. For pane Average of Number Of Trips: The marks are labeled by % Difference in Avg. Number Of Trips. For pane Average of Avg Tip: The marks are labeled by % Difference in Avg. Avg Tip. The data is filtered on CCA1, which keeps 6, 28, 33 and 34.



Average of Number Of Trips and average of Avg Tip for each Rainy Day. For pane Average of Number Of Trips: The marks are labeled by % Difference in Avg. Number Of Trips. For pane Average of Avg Tip: The marks are labeled by % Difference in Avg. Avg Tip.

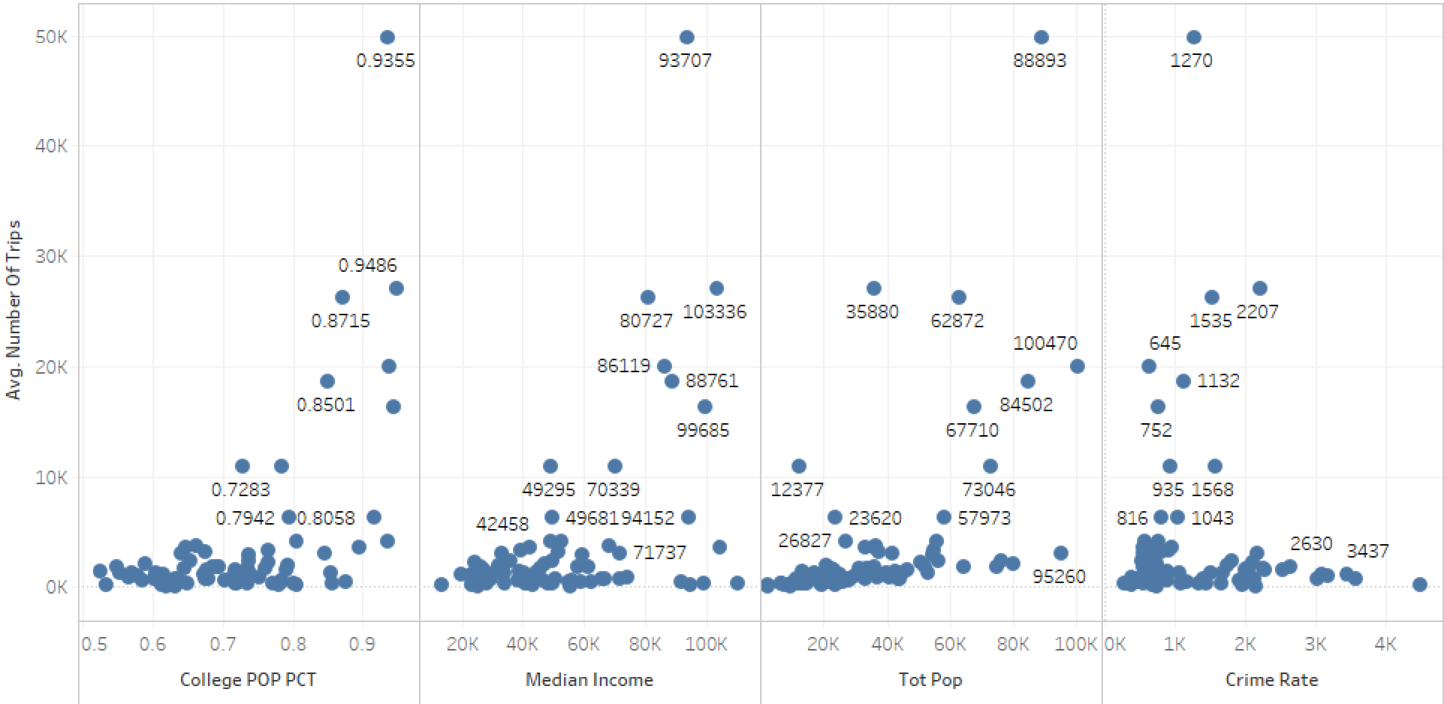
- Higher Number of Trips and Tips during sports event days
- No significant difference in Number of trips and Tips among rainy day



The plots of average of Number Of Trips for Med Age and MD_AGE_POP_PCT.

- Most customers are at the age between 25 to 40

Education/Income/Tot population/Crime Rate on Ridership



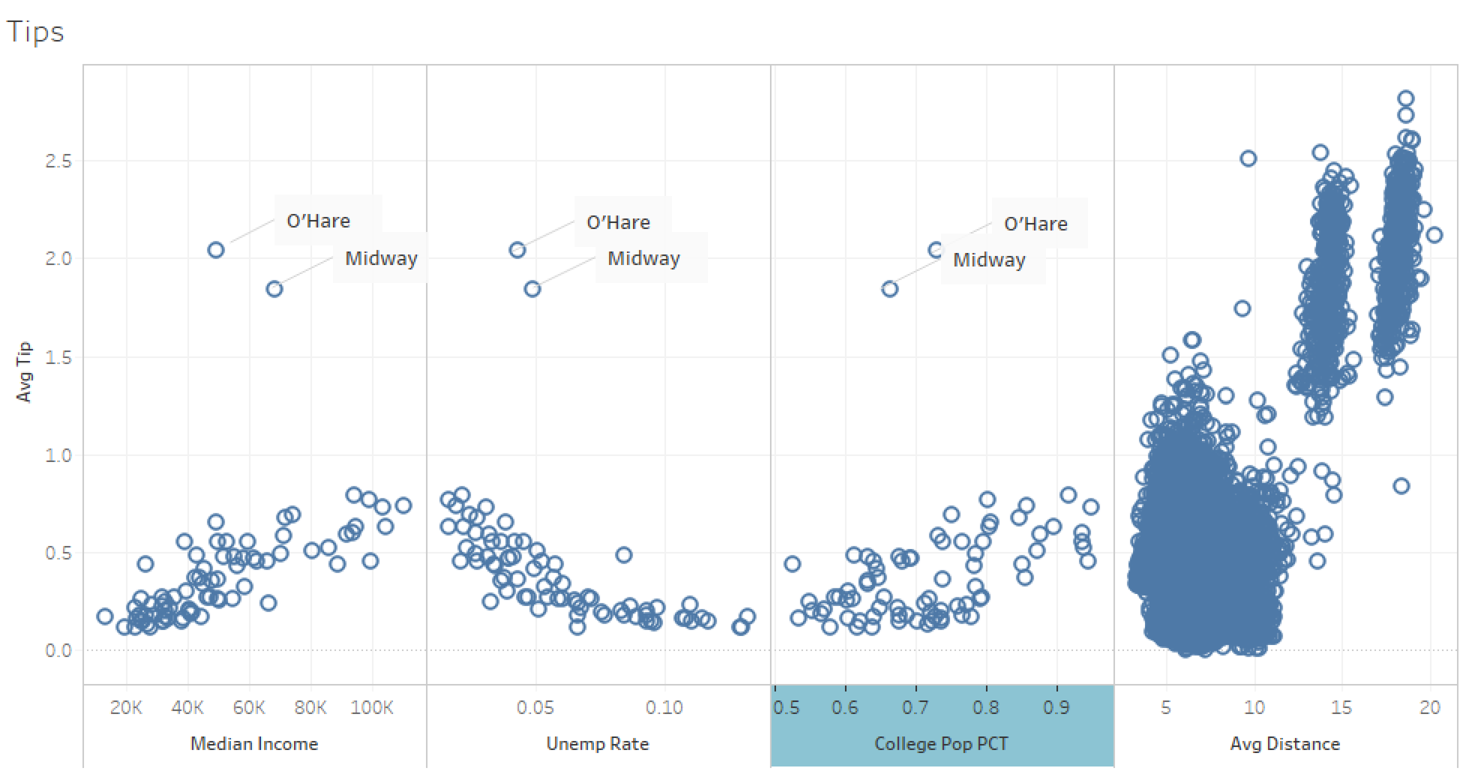
The plots of average of Number Of Trips for Col Pop Rate, Medinc, Tot Pop and Crime Rate. For pane Col Pop Rate: The marks are labeled by Col Pop Rate. For pane Medinc: The marks are labeled by Medinc. For pane Tot Pop: The marks are labeled by Tot Pop. For pane Crime Rate: The marks are labeled by Crime Rate.

Higher Education -> More Trips

Higher Population -> More Trips

Higher Income -> More Trips

Higher Crime Rate -> Fewer Trips



The plots of average of Avg Tip for Medinc, Unemp Rate, Col Pop Rate and Avg Distance.

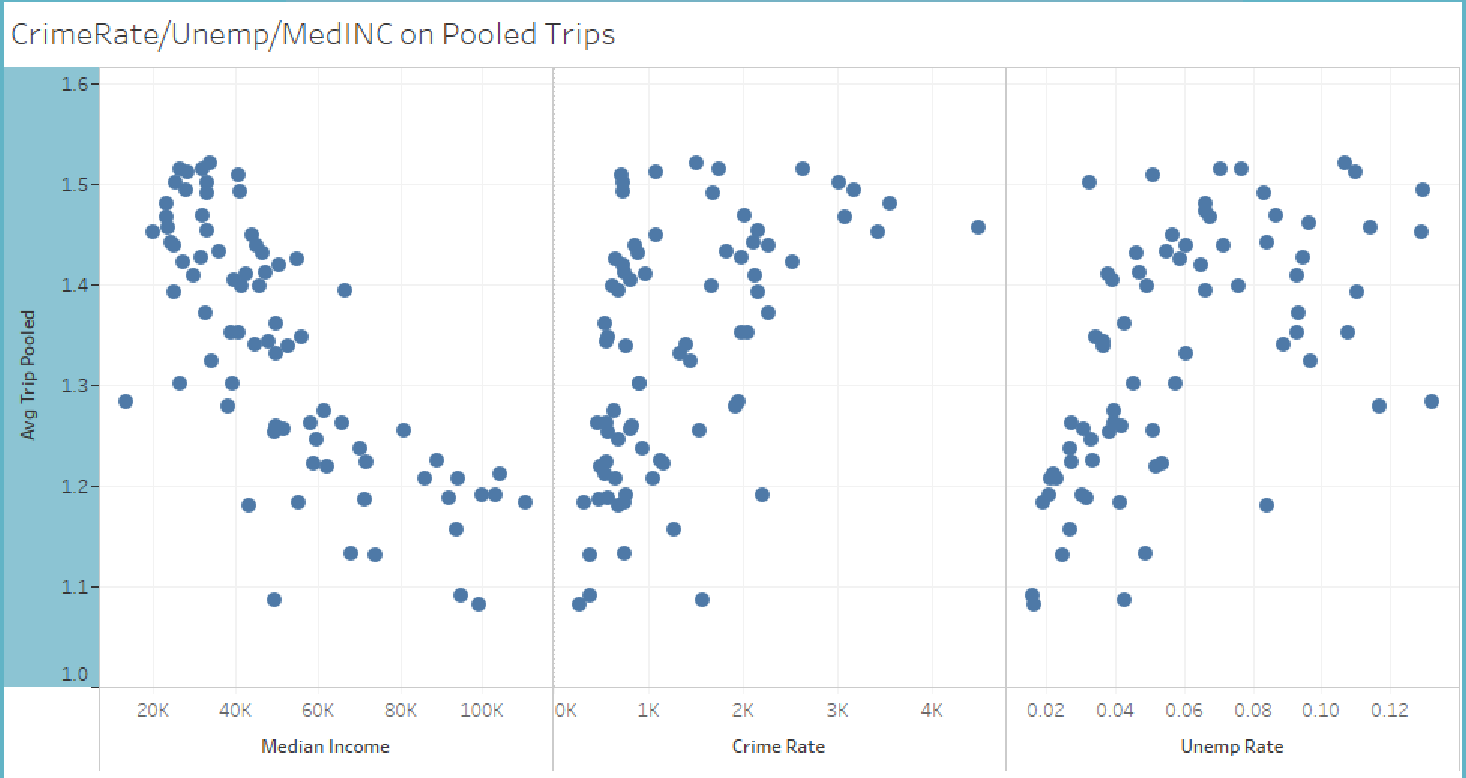
Higher Income -> Higher Tips

Higher Education -> Higher Tips

Higher unemployment rate -> Less Tips

Longer Distance -> Higher Tips

Explaining higher tips in O'Hare and Midway airport

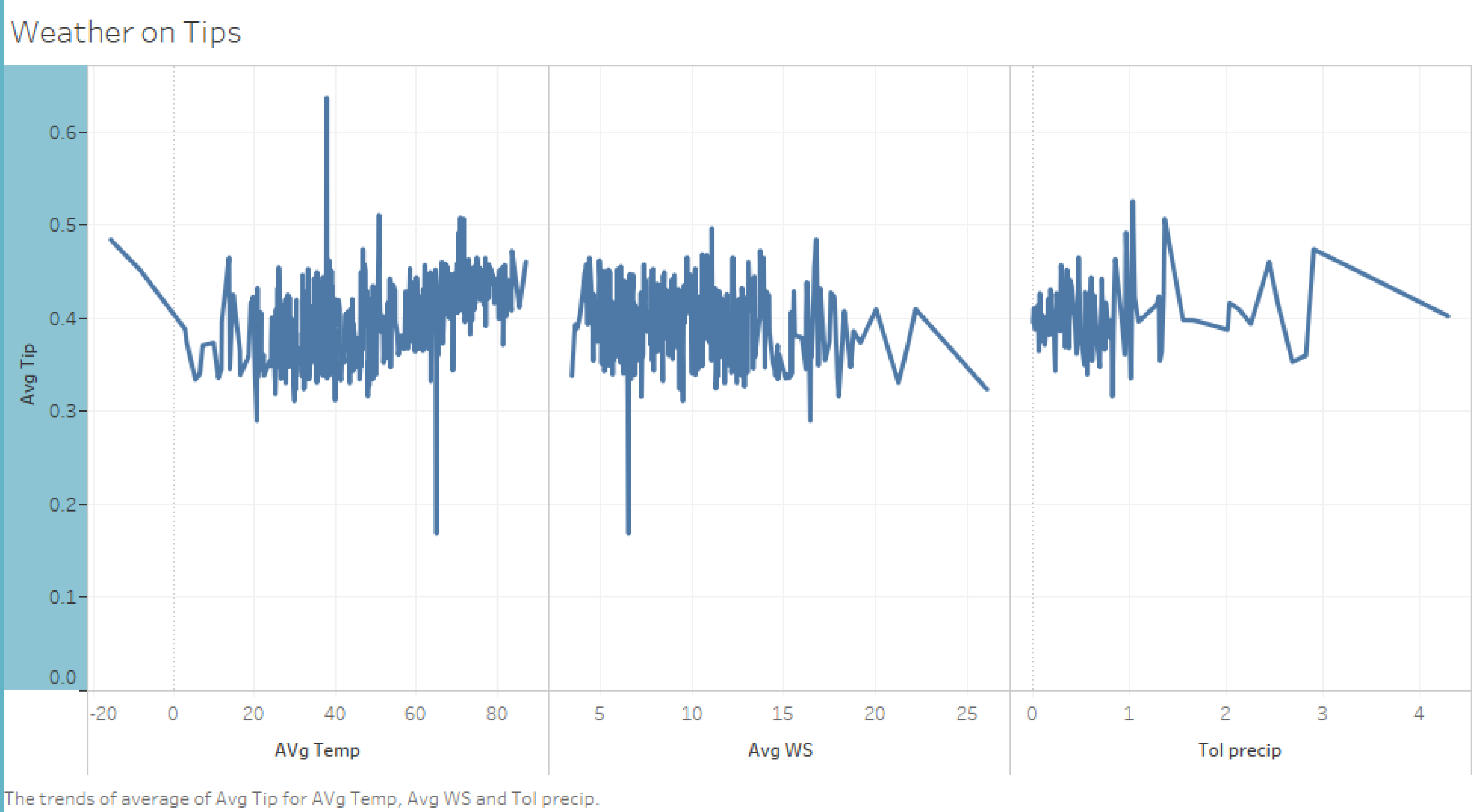


The plots of average of Avg Trip Pooled for Medinc, Crime Rate and Unemp Rate.

Lower Income -> More Carpool

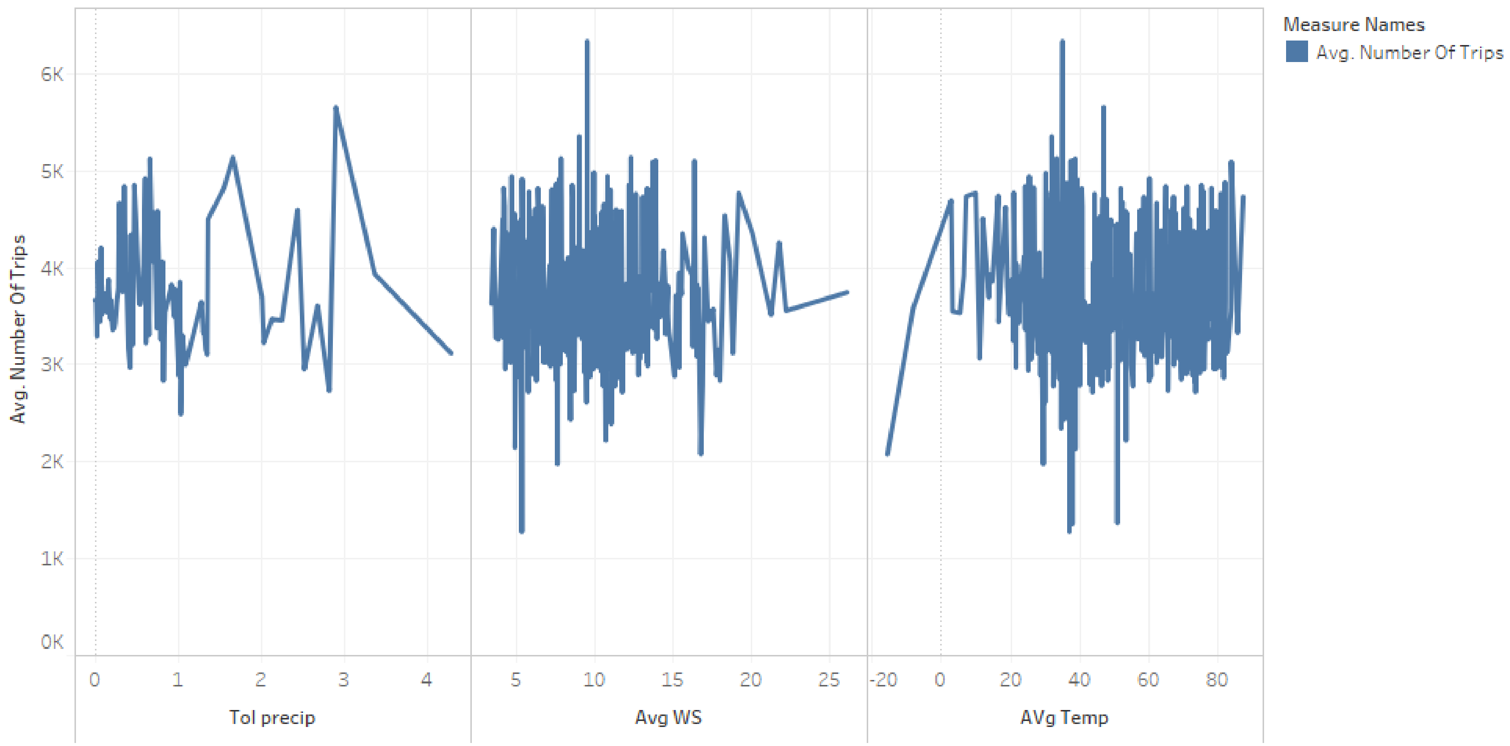
High unemployment rate -> More Carpool

Higher Crime Rate -> More Carpool



- No significant difference among avg temperature, avg windspeed and total precipitation over tip.

Weather on Ridership



The trends of Avg. Number Of Trips for Tol precip, Avg WS and AVg Temp. Color shows details about Avg. Number Of Trips.

- No significant difference among avg temperature, avg windspeed and total precipitation over number of trips.