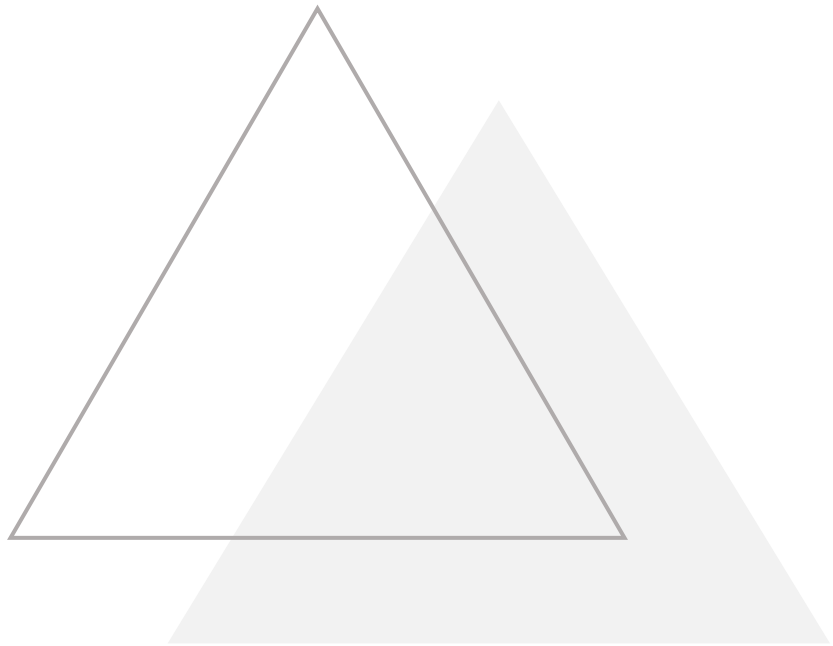
# Optimizing Ride Sharing Allocation

*Analysis on Ridership Allocation for Transportation Network Providers(TNP) in Chicago*

# Agenda



- Executive Summary
- Data Ingestion & Preparation
- Data Modeling & Design
- Analytics & Results
- Conclusion & Recommendations
- Future Work



# 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 focus on relational database system that 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<br>• Vehicle allocation Optimization<br>• 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 and 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**

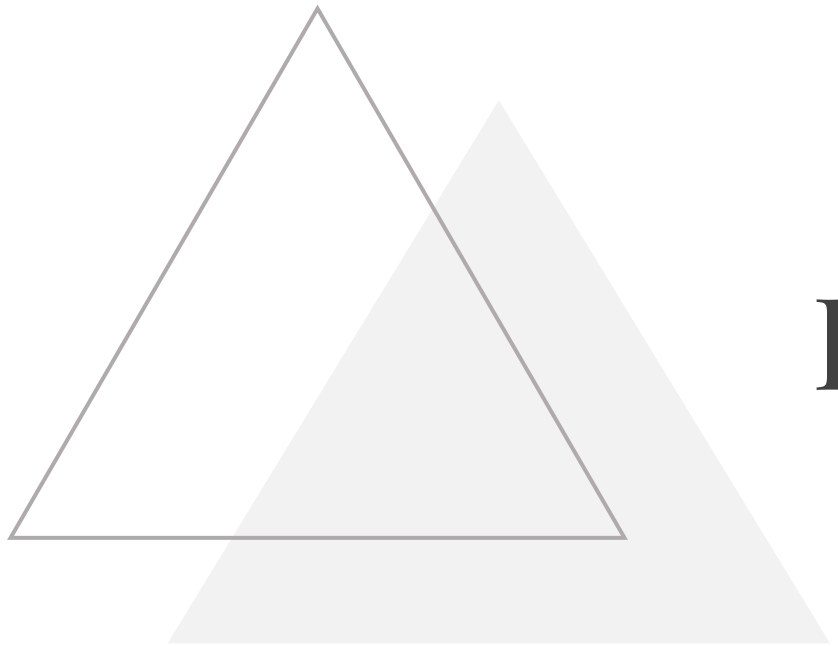
## Analytics

**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 | 70 GB<br>(129 Million Rows, 21 Columns)<br>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<br>Structured CSV File | MULTIPOLYGON Data by CCA For Tableau | Python and Excel |
| Census Data By CCA | 1.1 MB<br>Structured CSV File | Education, Age, Income, Population & Unemployment Rate etc. by CCA | Python, MySQL |
| Chicago Data Portal: Crimes | 16 GB<br>(7.12 Million Rows, 22 Columns)<br>Structured CSV File | Total Number of Crimes By CCA | Python, MySQL GCP and RCC |
| ESPN | 2 MB<br>Unstructured Data:<br>From Web Scraping | Only the home game dates and location | Python and Excel |
| National Centers for Environmental Information | 5.5 MB<br>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<br>Unstructured Data:<br>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 table
- 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:** DOUBLE
- **Others:** INT

## Dealing with NAs

- **Main Dataset:** Fill NA with 0
- **Weather:** Drop NA rows

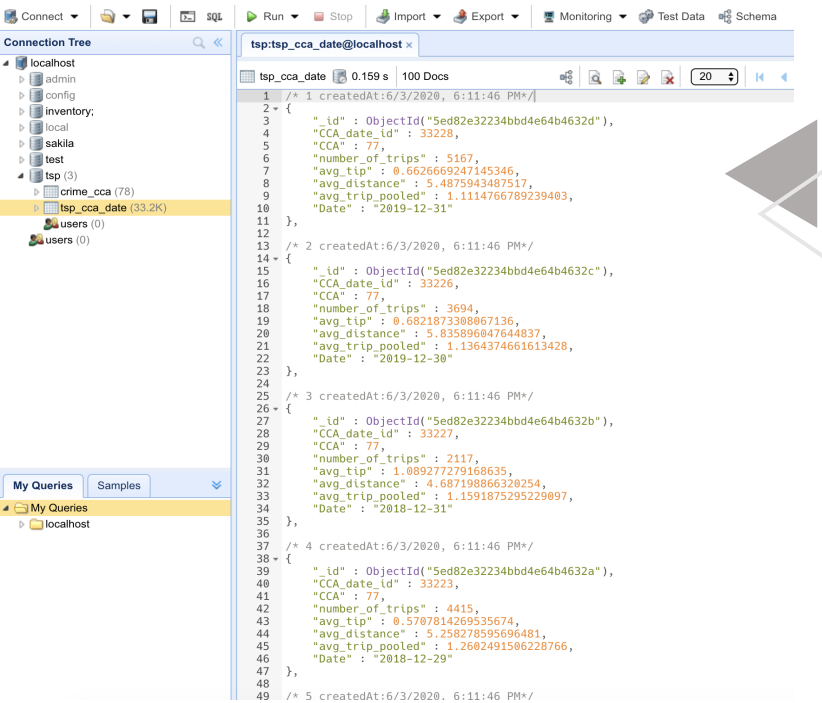
## Dimensional Tables

- Maintain historical information for any dimension
- Less processing time and higher performance

## Expected Output of Data Analysis

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

# NoSQL Considerations

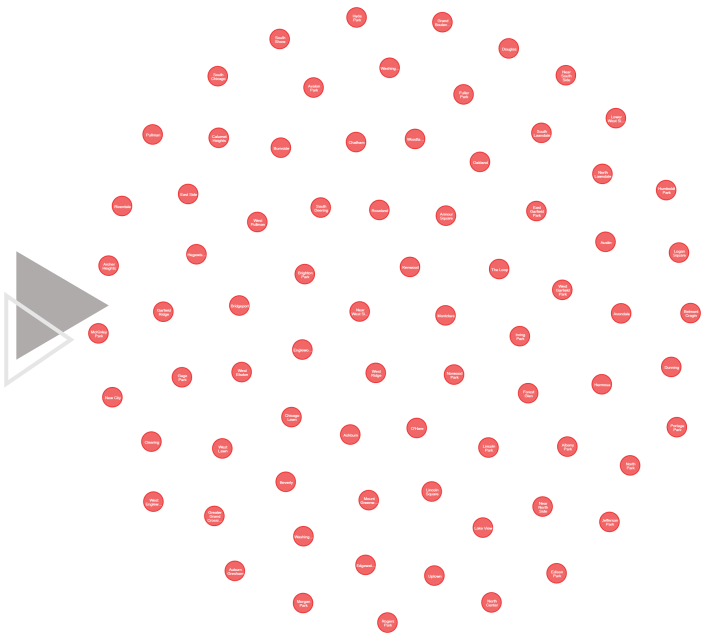



## MongoDB

- 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



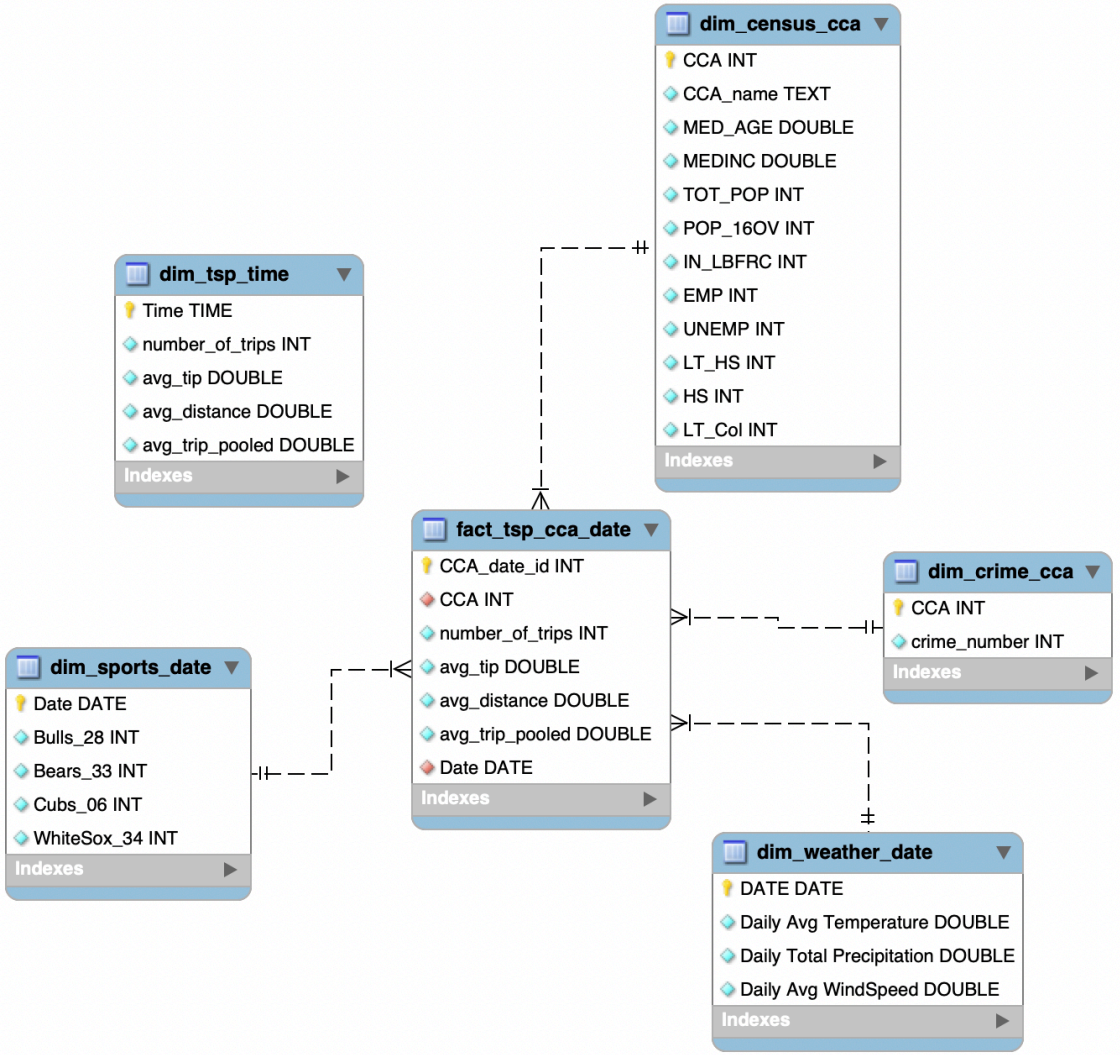
## Advantages

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

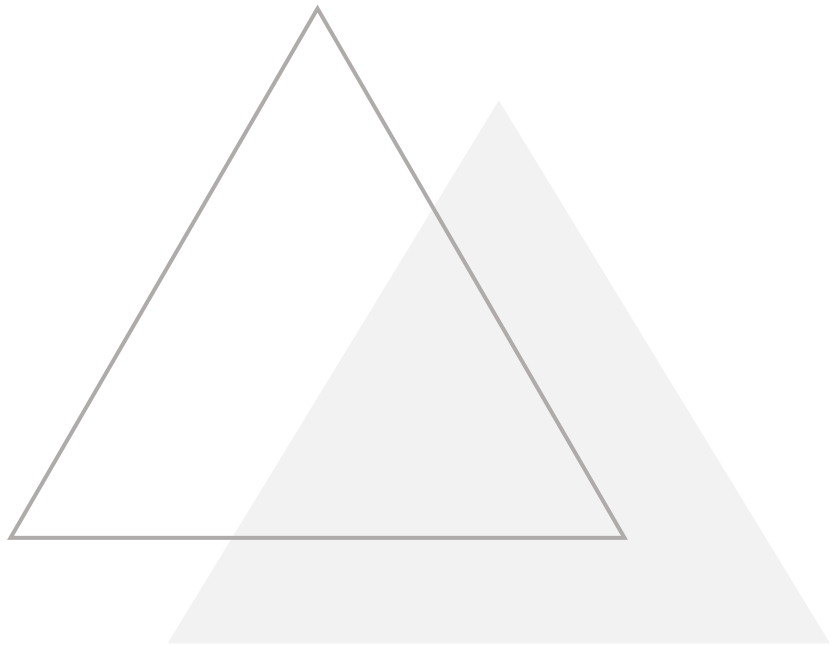
# Data Quality Dimension

- ✓ **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

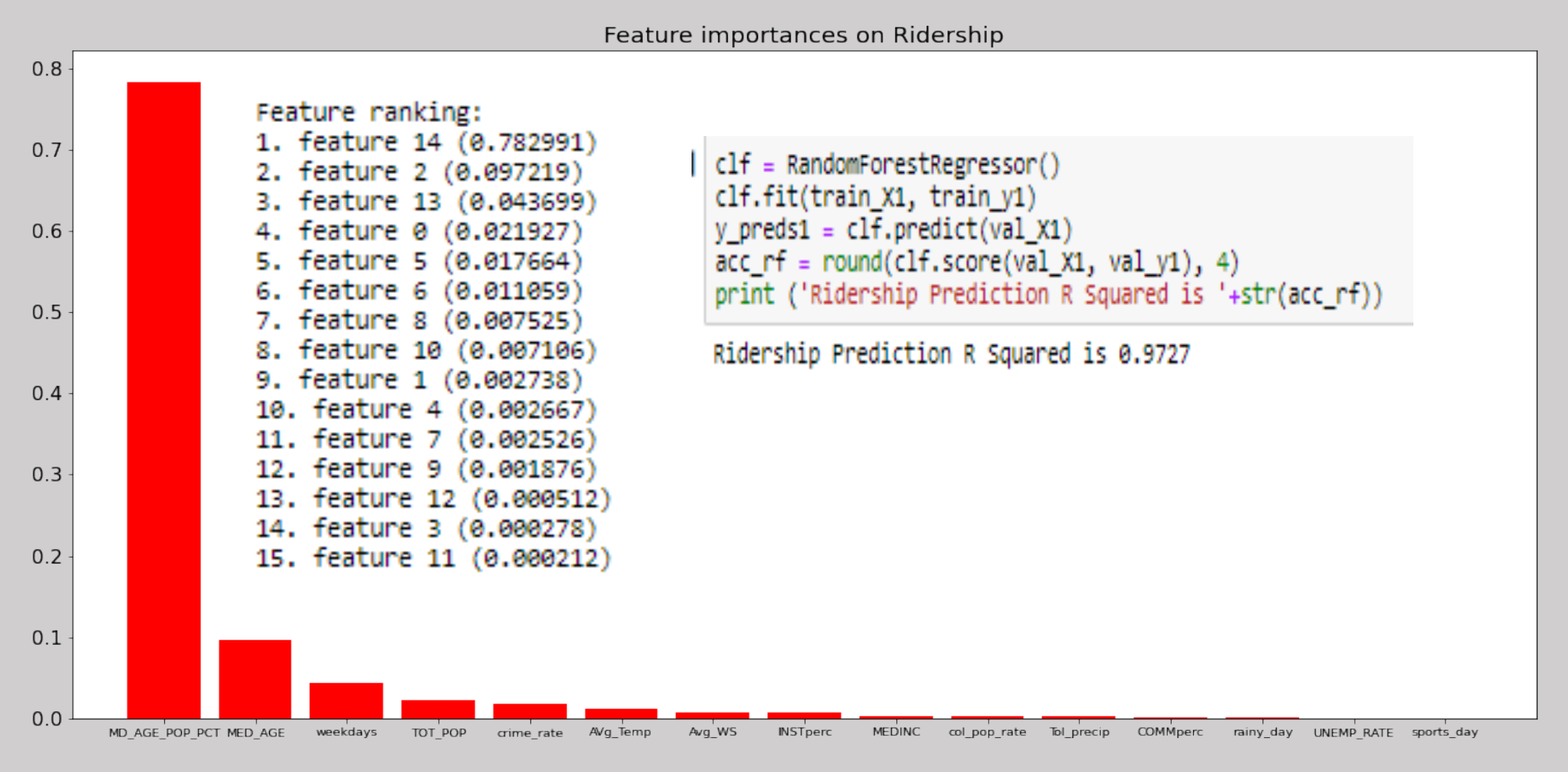



Notes: dim_tsp_time table is an independent table to analyze the average change of measures from hour to hour within a day



# Results & Analytics

# Random Forest Regressor

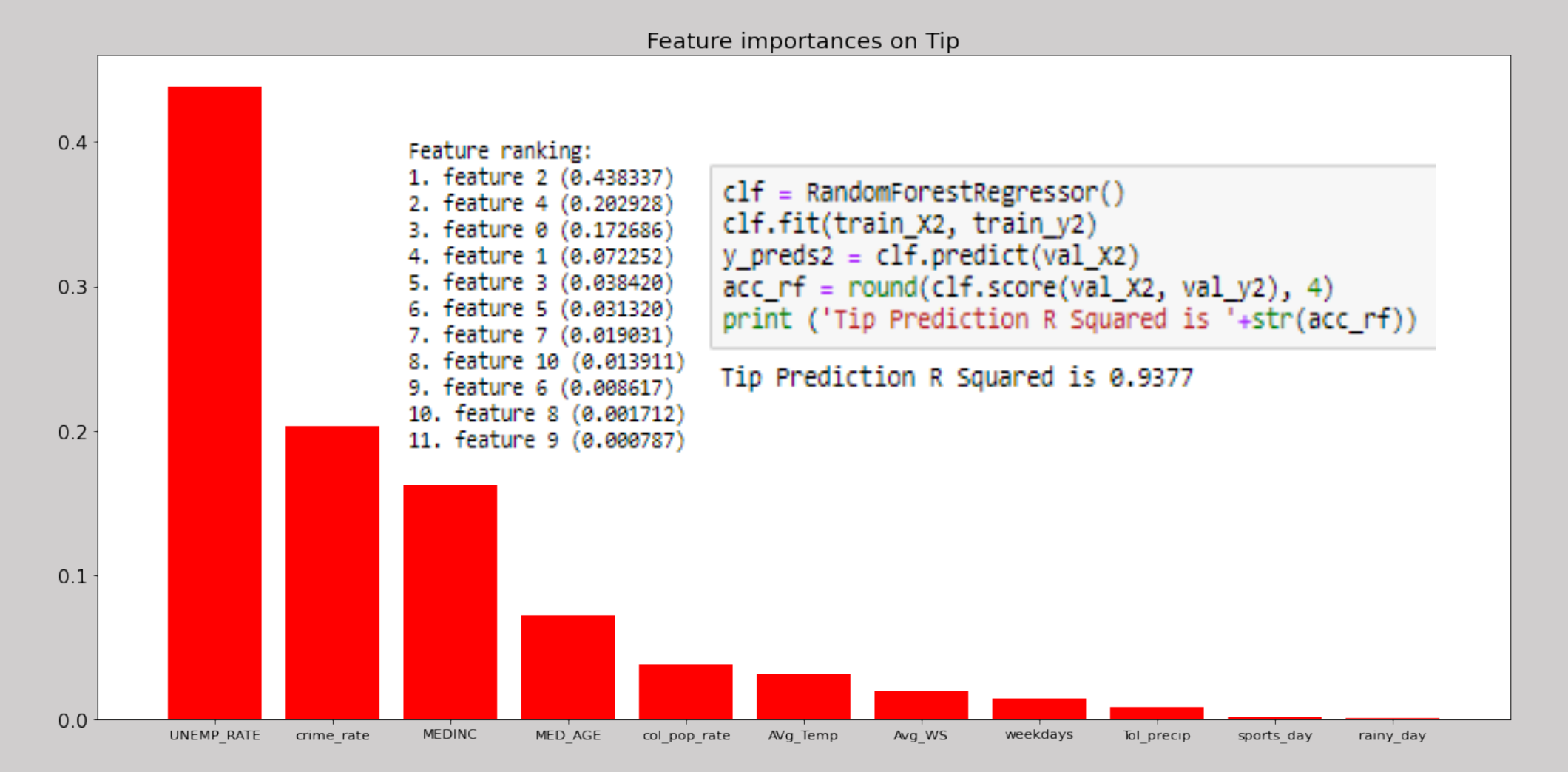



```
Feature ranking:
1. feature 14 (0.782991)
2. feature 2 (0.097219)
3. feature 13 (0.043699)
4. feature 0 (0.021927)
5. feature 5 (0.017664)
6. feature 6 (0.011059)
7. feature 8 (0.007525)
8. feature 10 (0.007106)
9. feature 1 (0.002738)
10. feature 4 (0.002667)
11. feature 7 (0.002526)
12. feature 9 (0.001876)
13. feature 12 (0.000512)
14. feature 3 (0.000278)
15. feature 11 (0.000212)
```

```
clf = RandomForestRegressor()
clf.fit(train_X1, train_y1)
y_preds1 = clf.predict(val_X1)
acc_rf = round(clf.score(val_X1, val_y1), 4)
print ('Ridership Prediction R Squared is '+str(acc_rf))
```

```
Ridership Prediction R Squared is 0.9727
```

# Random Forest Regressor

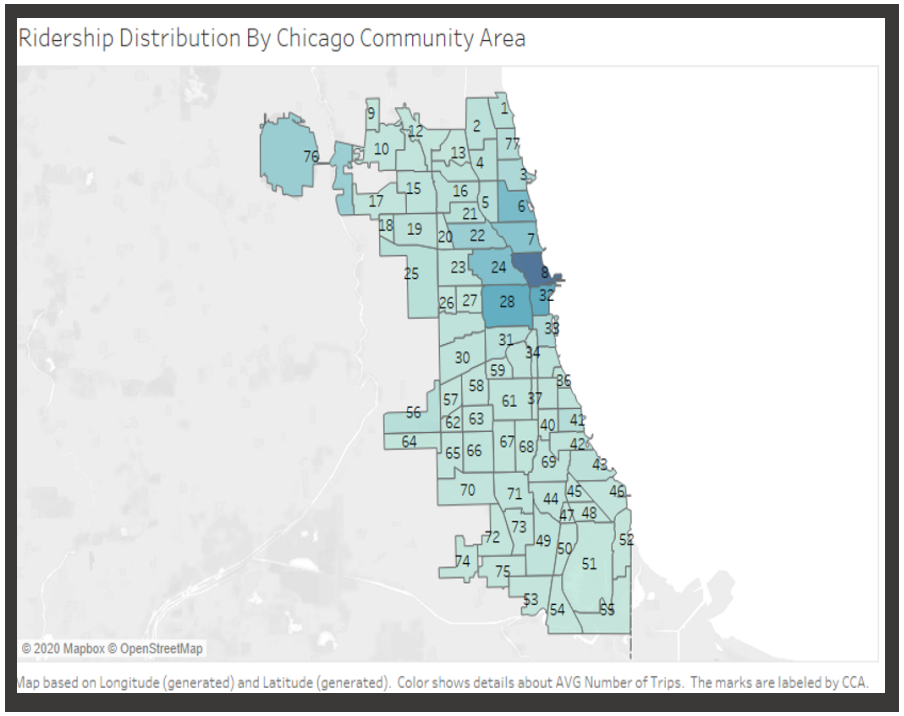



```
Feature ranking:
1. feature 2 (0.438337)
2. feature 4 (0.202928)
3. feature 0 (0.172686)
4. feature 1 (0.072252)
5. feature 3 (0.038420)
6. feature 5 (0.031320)
7. feature 7 (0.019031)
8. feature 10 (0.013911)
9. feature 6 (0.008617)
10. feature 8 (0.001712)
11. feature 9 (0.000787)
```

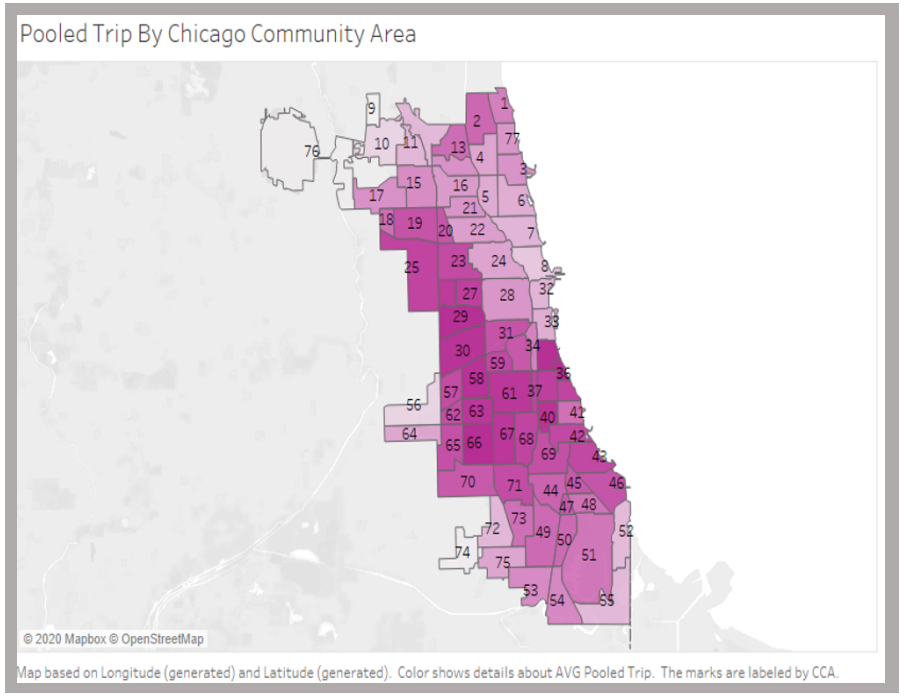
```
clf = RandomForestRegressor()
clf.fit(train_X2, train_y2)
y_preds2 = clf.predict(val_X2)
acc_rf = round(clf.score(val_X2, val_y2), 4)
print ('Tip Prediction R Squared is '+str(acc_rf))
```

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

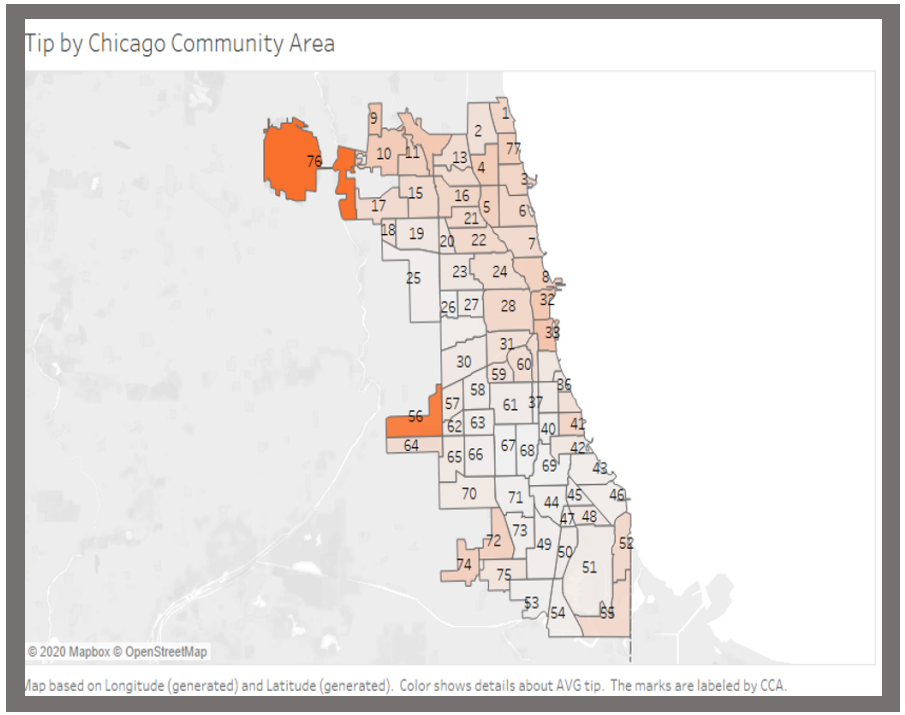
Community Areas with high ridership are business districts and surrounding areas



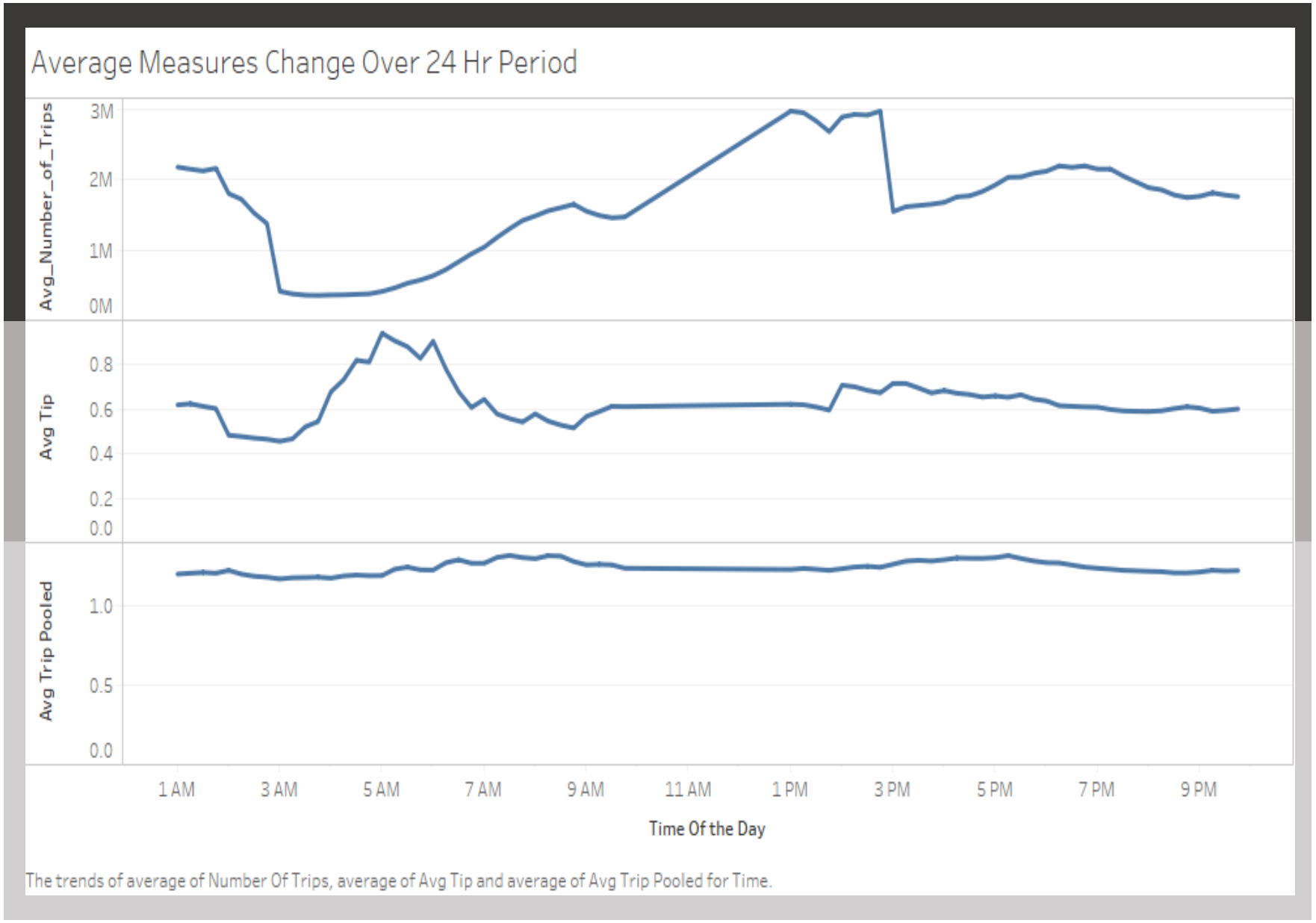

Carpool are used more in west and south lower income neighborhoods of Chicago



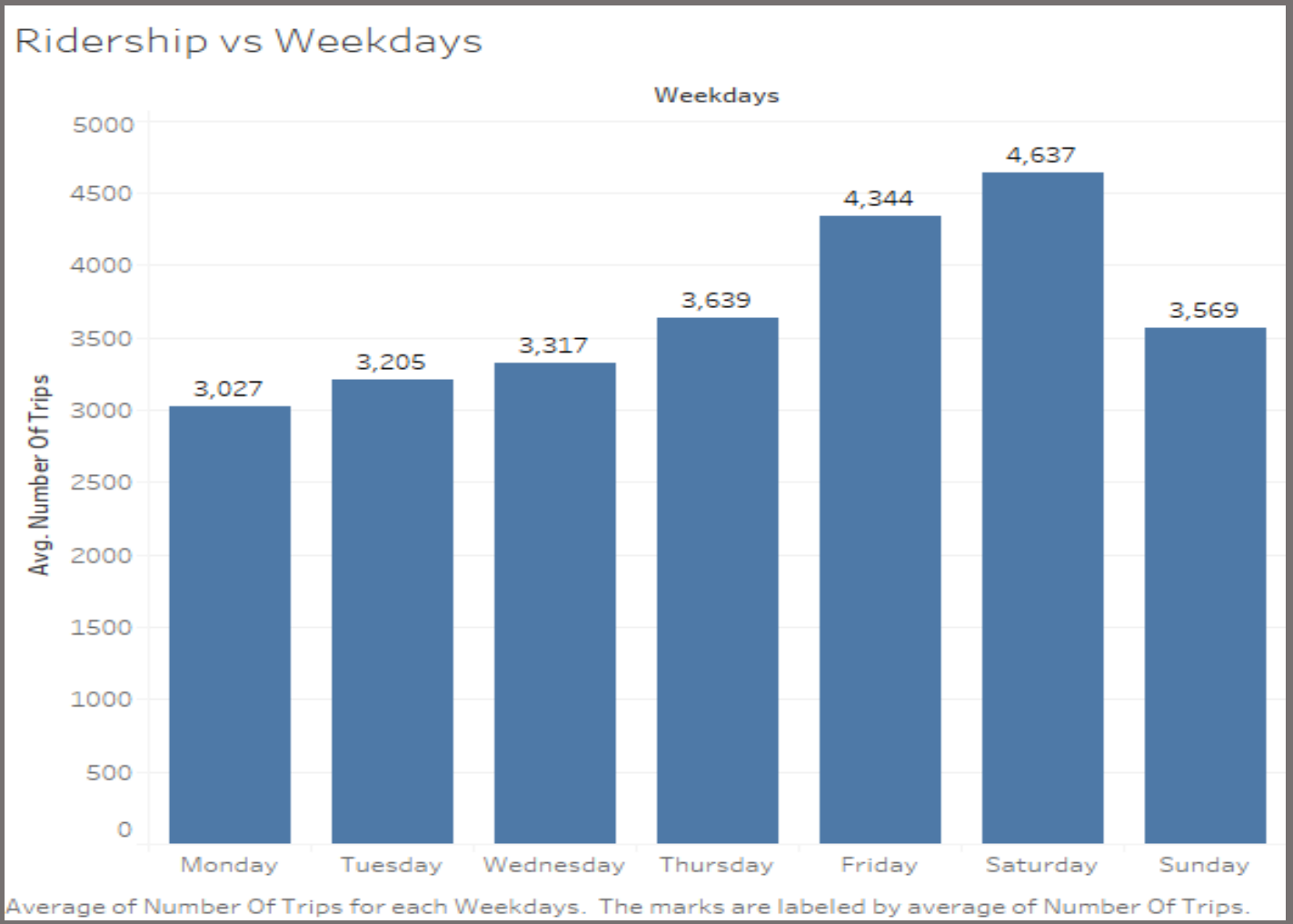

Tip are higher in CCA where O'hare international airport and Midway airport are located

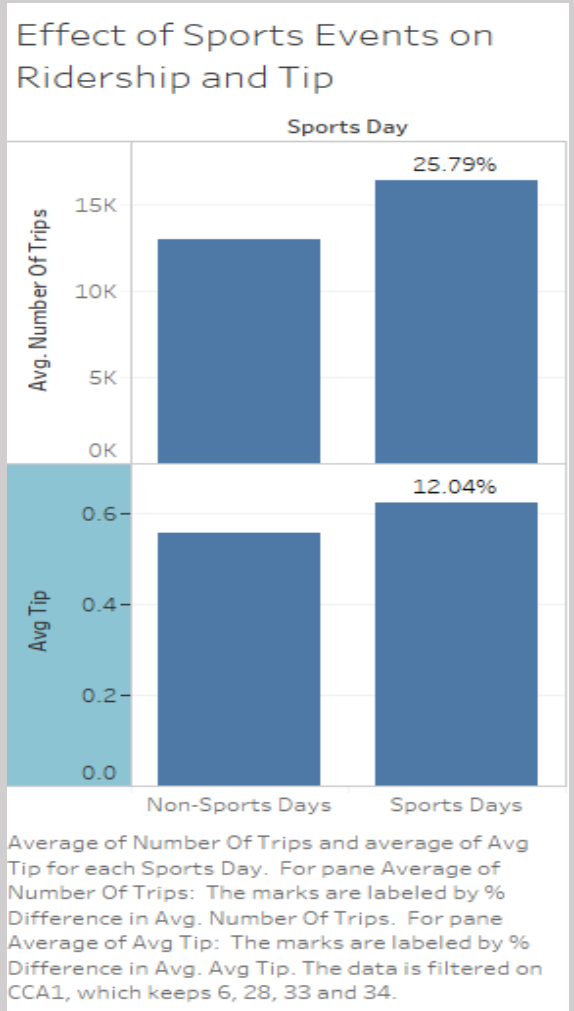



Neighborhoods with higher income tip better



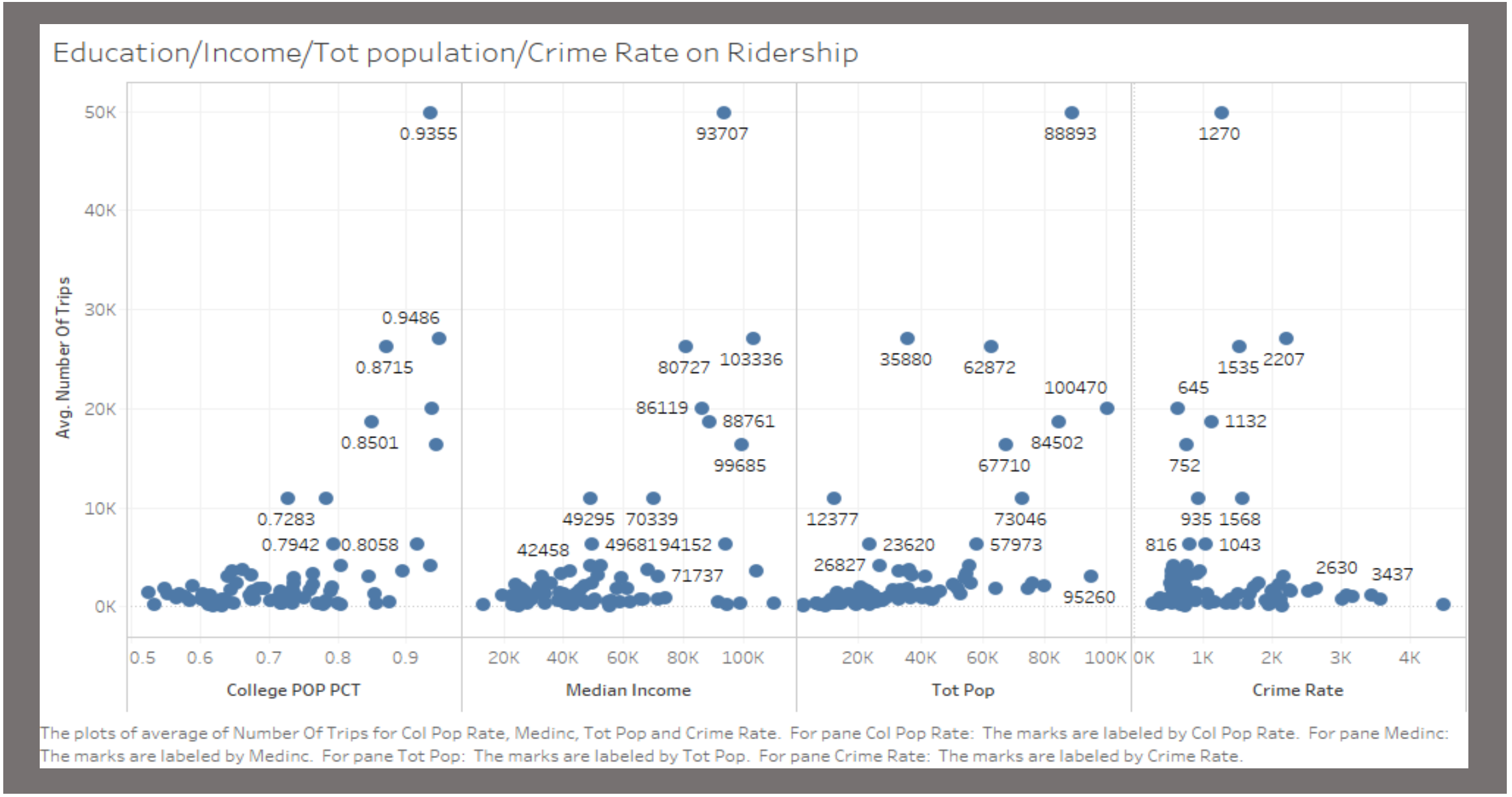

- Higher number of trips at daytime
- Rapid decline from 2-3pm might due to rush hour or data gap
- Higher tips from 5-7am, reason might be morning arrival or departure flight
- No significant difference for carpool over 24hr period




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

Higher Number of Trips and Tips during sports event days

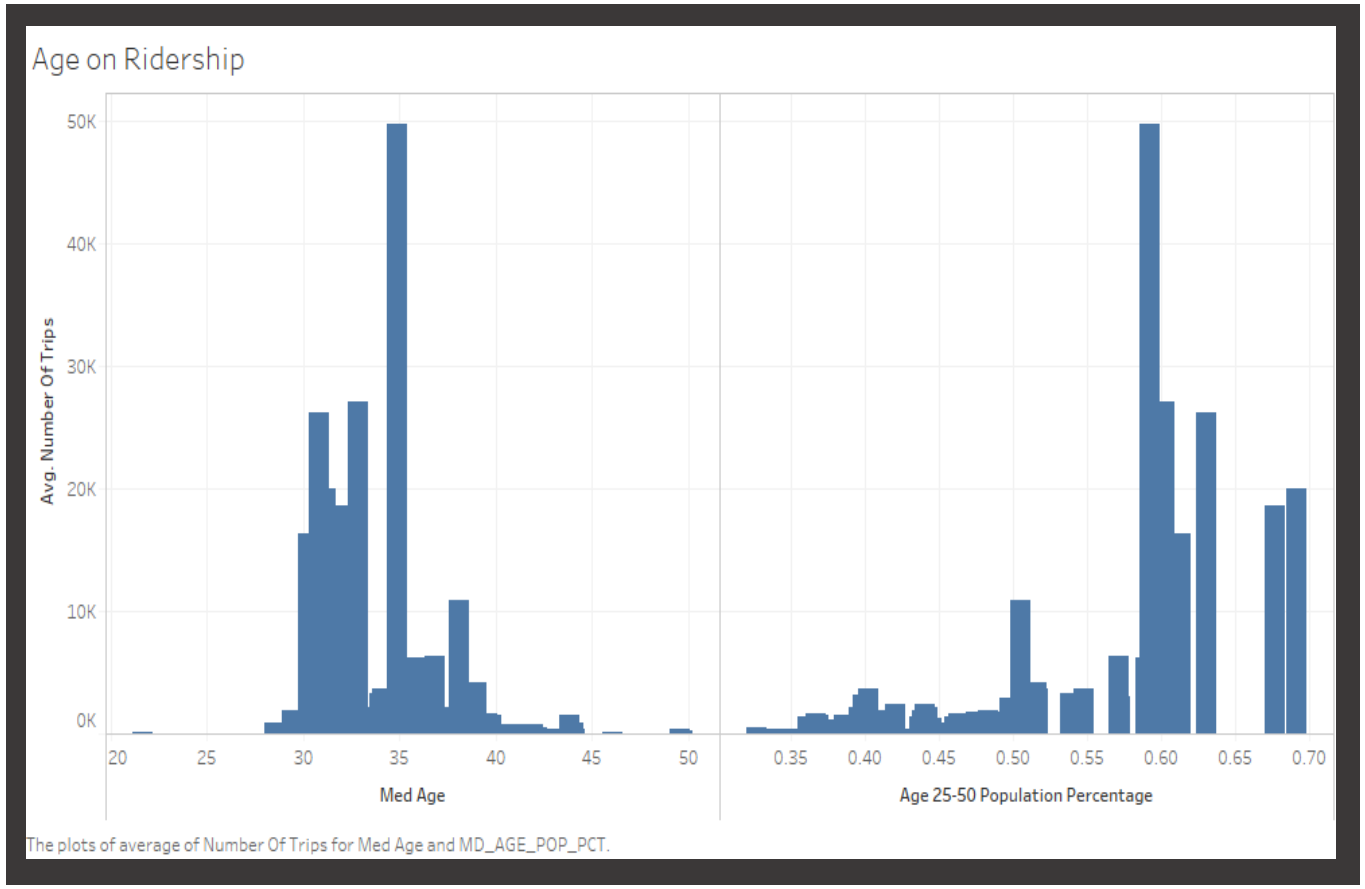



Higher Education
↓
More Trips

Higher Income
↓
More Trips

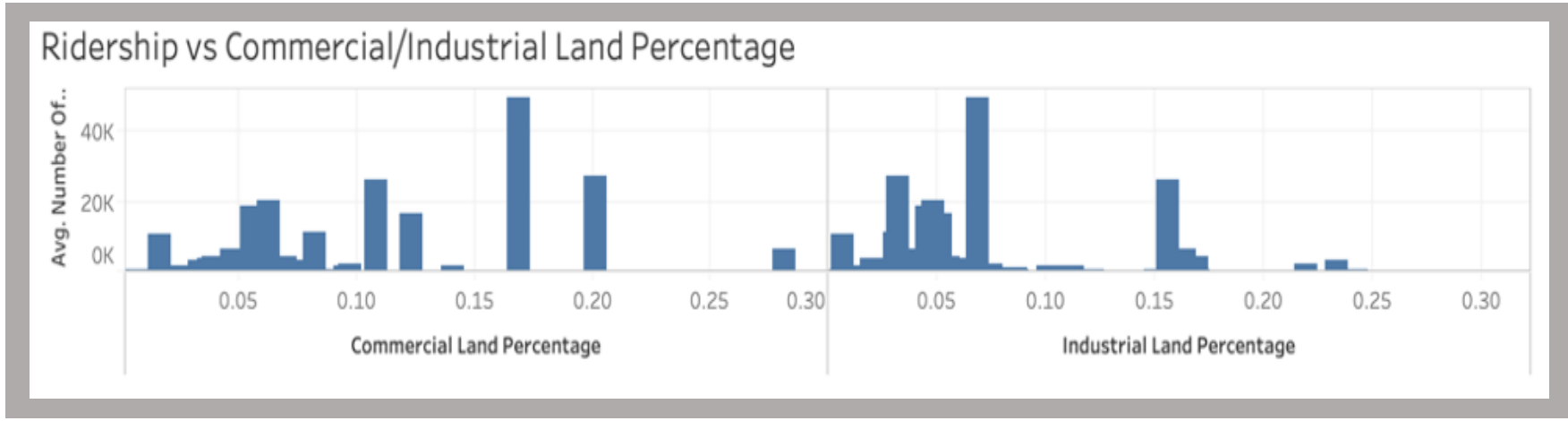
Higher Population
↓
More Trips

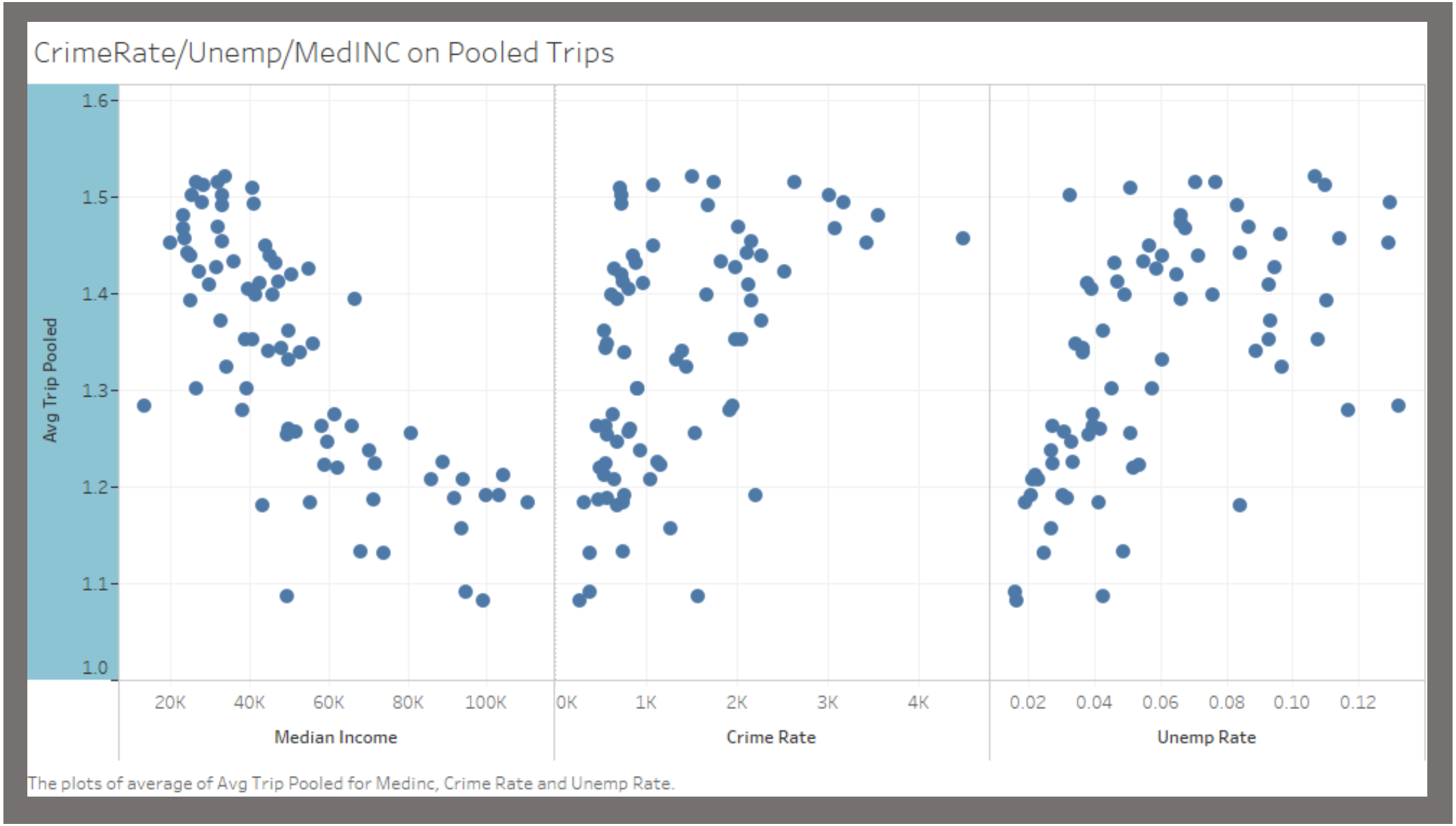
Higher Crime Rate
↓
Fewer Trips




Most customers are those age between 25 to 40

No significant difference in Number of trips and commercial / industrial land percentage

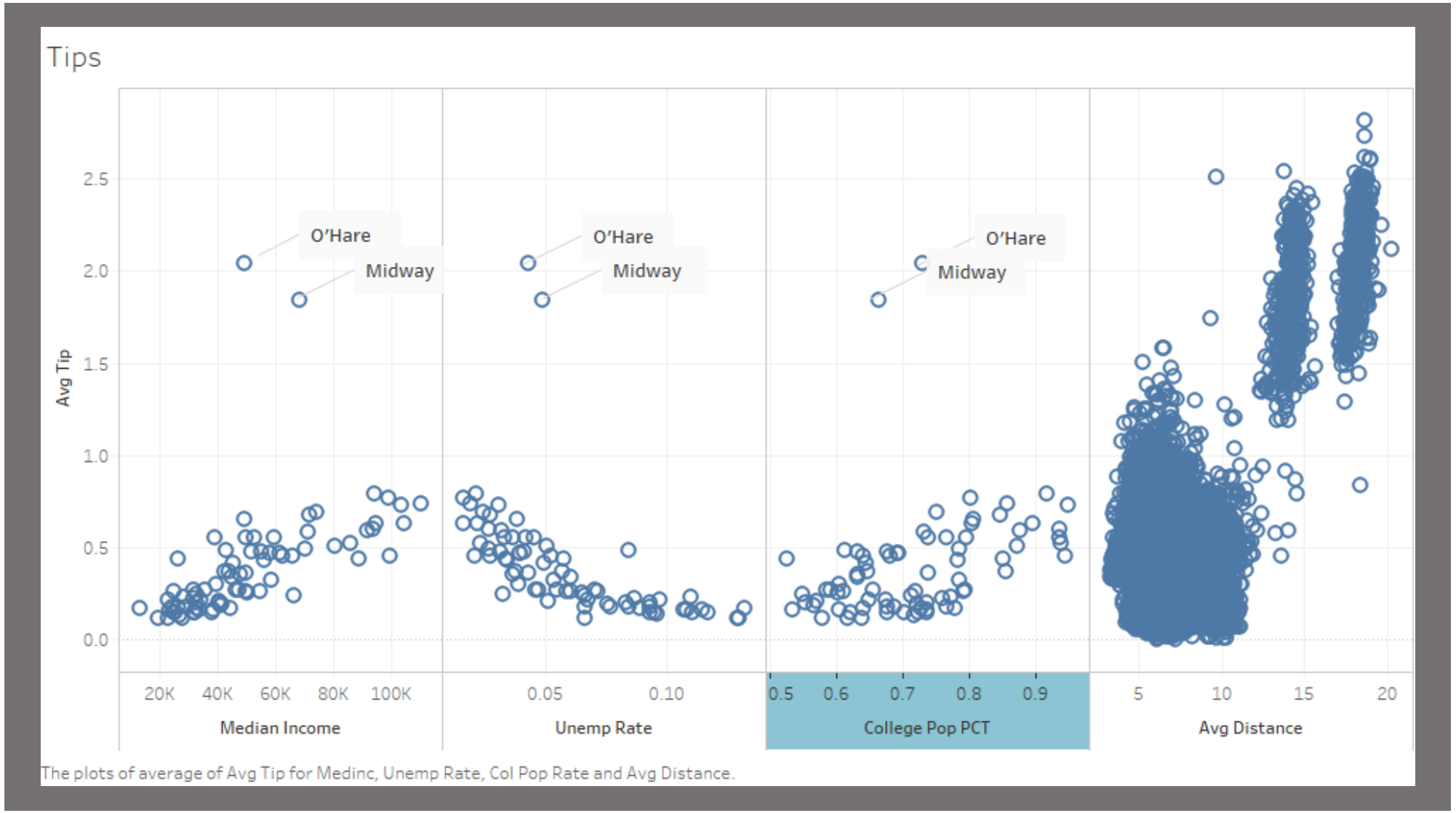





Lower Income
↓
More Carpool

Higher Crime Rate
↓
More Carpool

High unemployment rate
↓
More Carpool



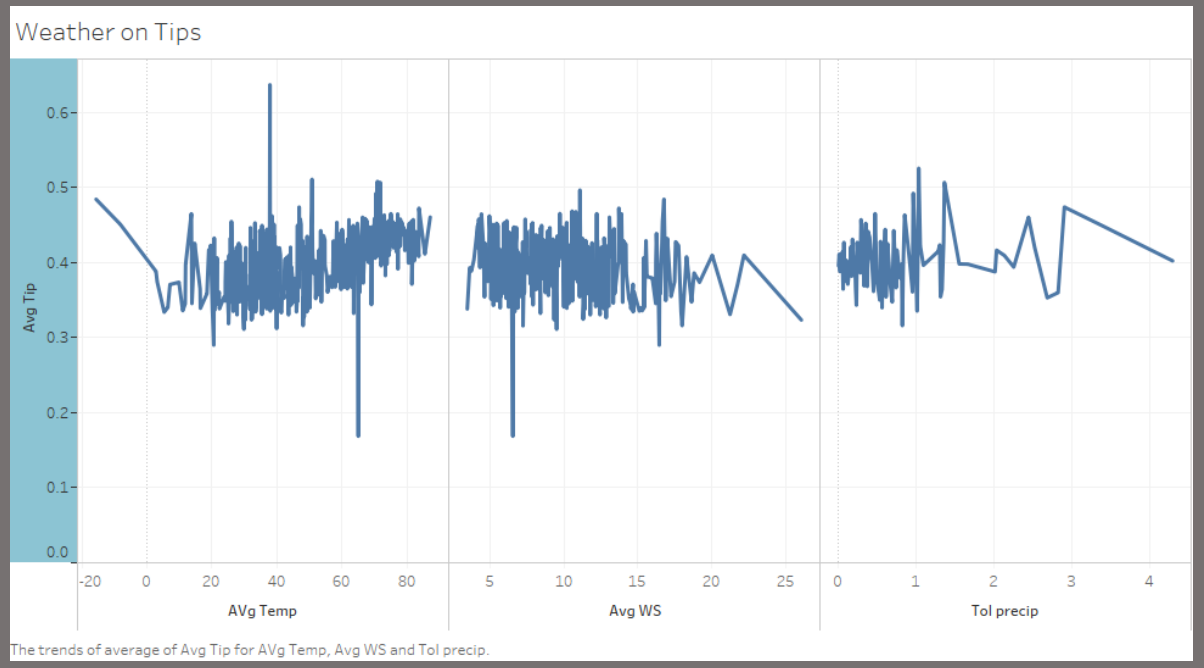

Higher Income
↓
Higher Tips

Higher unemployment rate
↓
Less Tips

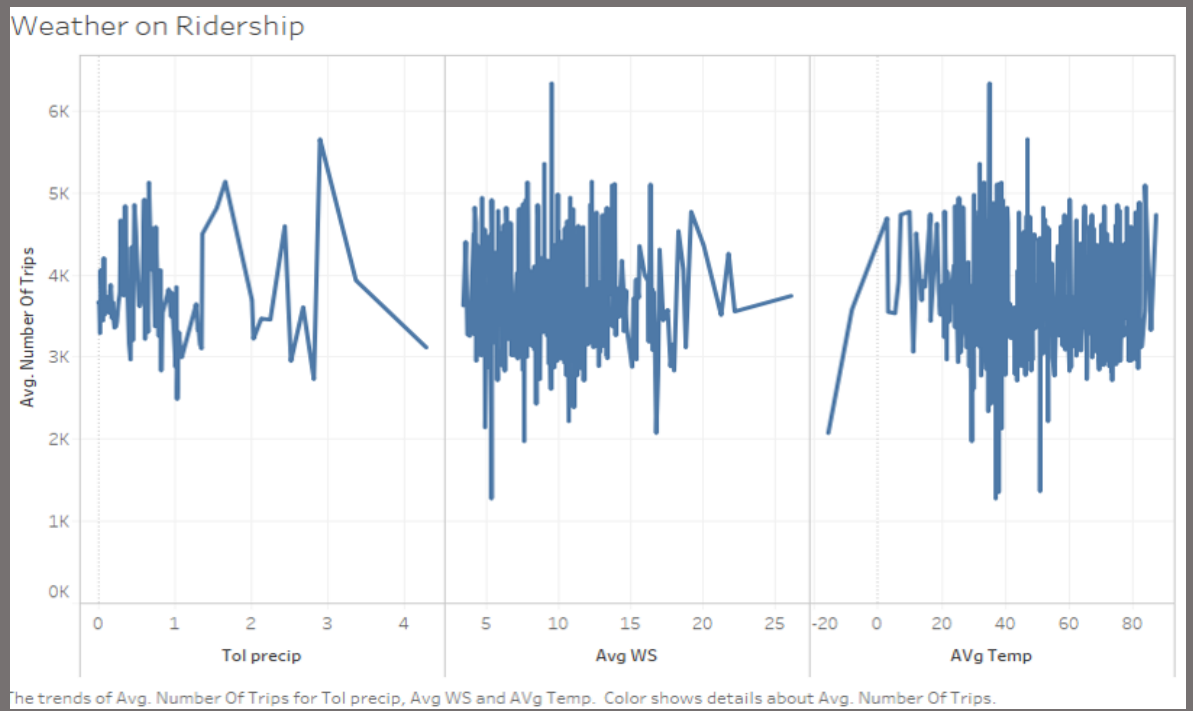
Higher Education
↓
Higher Tips

Longer Distance
↓
Higher Tips

Explaining higher tips in
O'Hare and Midway airport

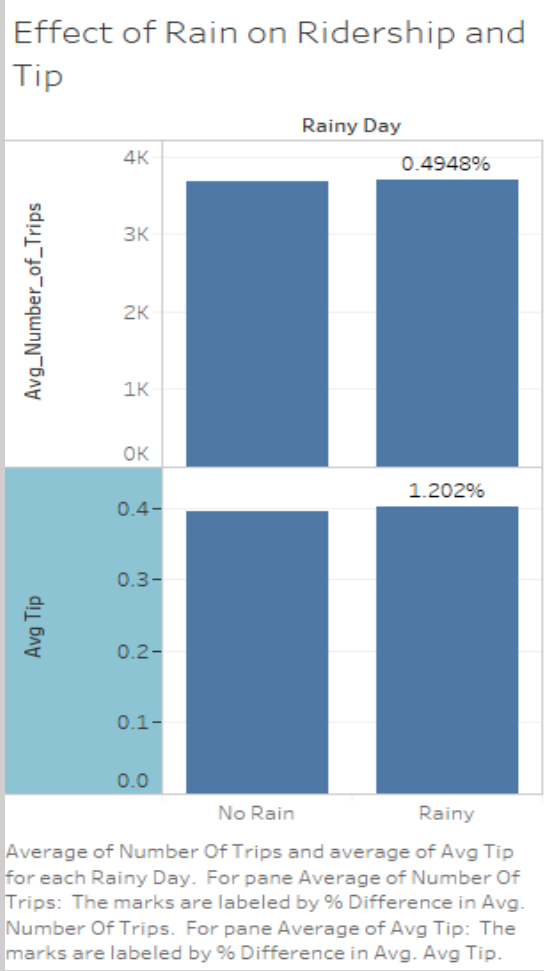



The trends of average of Avg Tip for AVg Temp, Avg WS and Tol precip.



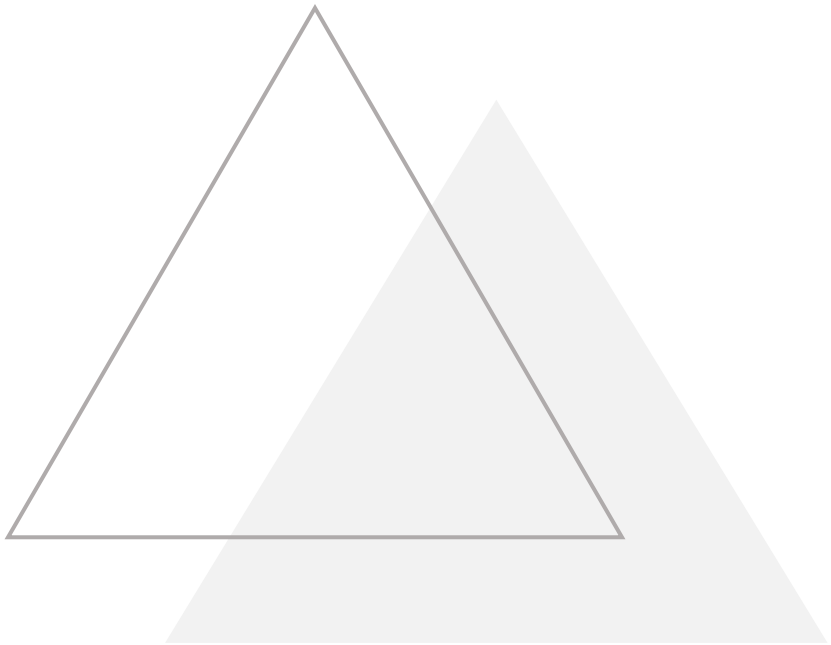

The trends of Avg. Number Of Trips for Tol precip, Avg WS and AVg Temp. Color shows details about Avg. Number Of Trips.

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




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.

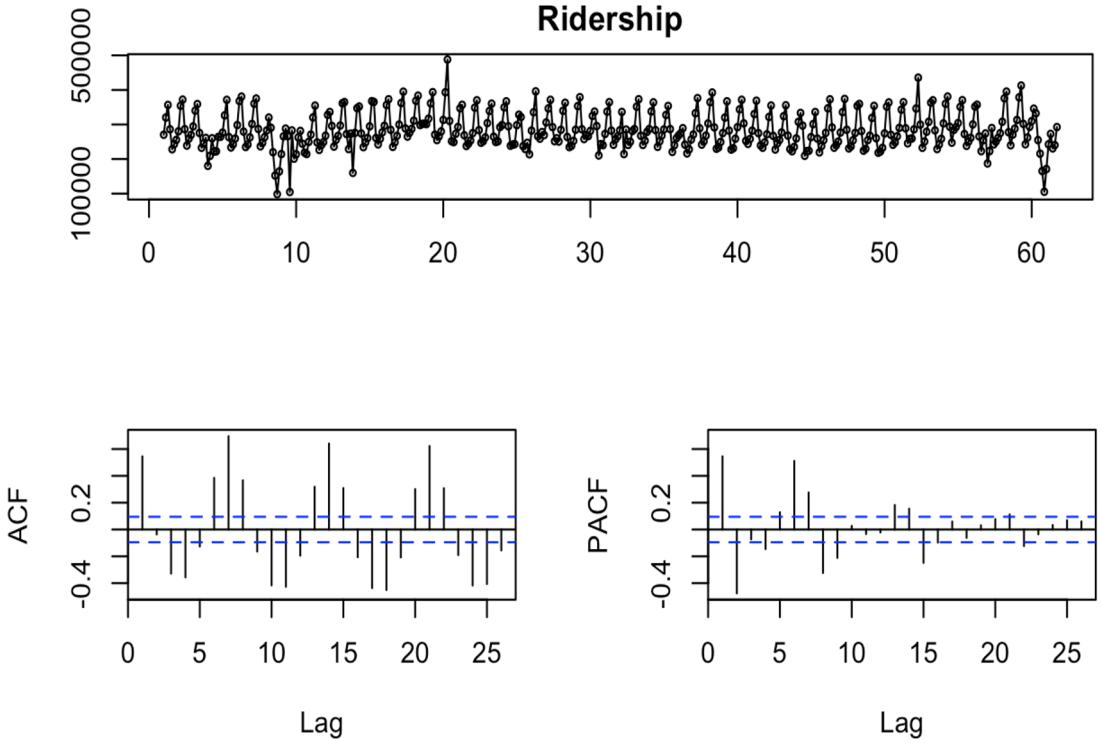
No significant difference in Number of trips and Tips among rainy day

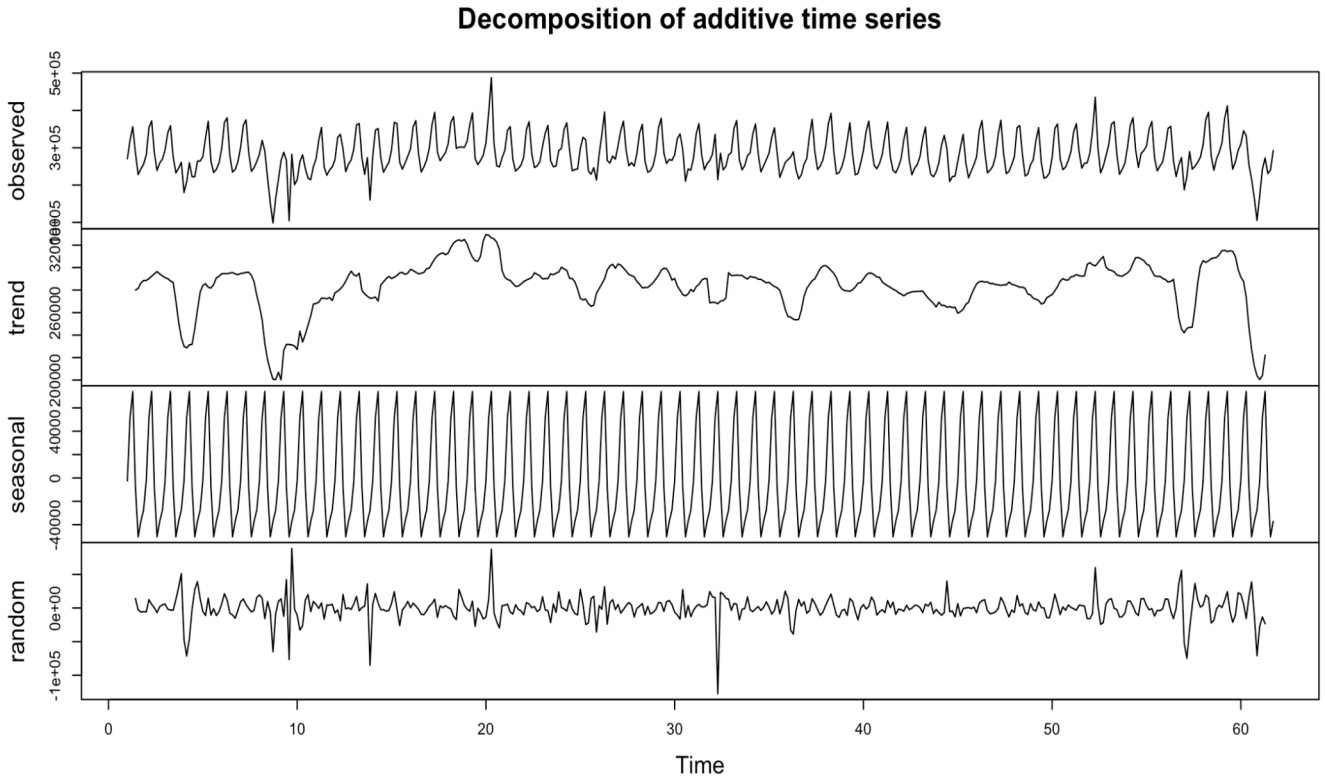


# Time Series Forecast

# Main Characteristics of Time Series

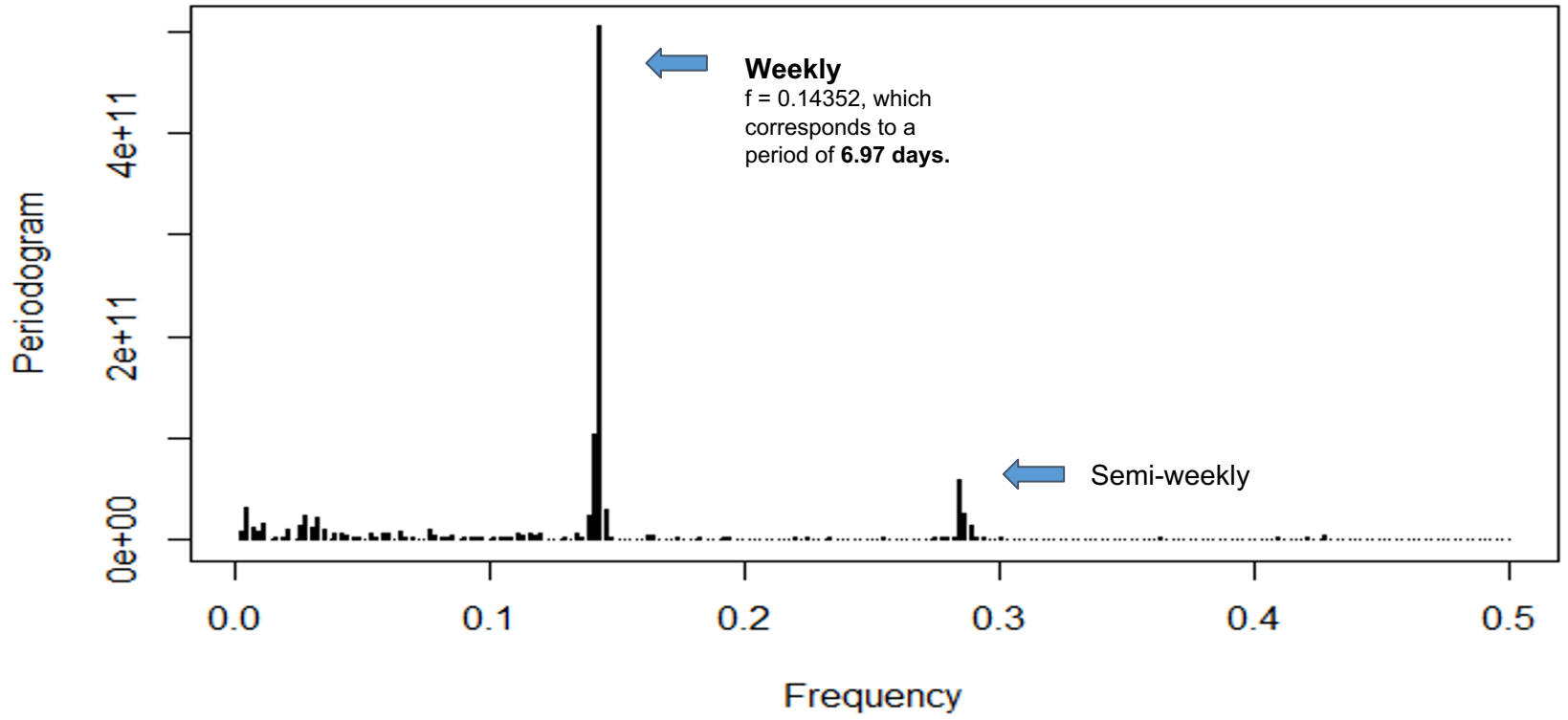
- Clear weekly seasonal pattern
- No need for Box-Cox
- No clear trend
- ADF Test - Data points are non-stationary



# Main Characteristics of Time Series

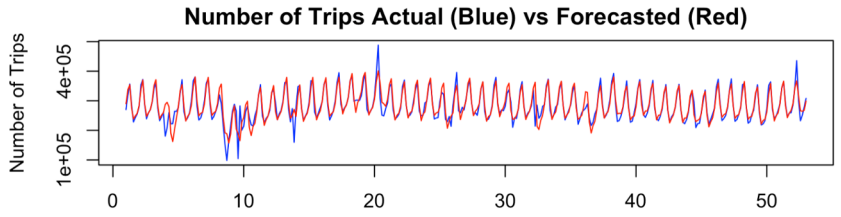
- Our spectral analysis shows that, by far, the most significant seasonal component is weekly



# Model Fitting – Holt Winters

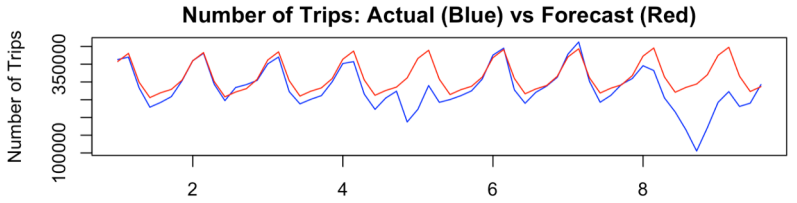
## Modelling Phase Outline

- Weekly seasonality was modeled by setting the TS object with Frequency = 7
- Seasonality was also modeled as Additive after observing that Multiplicative yielded higher AICc
- The resulting smoothing parameters are:
  - Alpha (level) = 0.452
  - Beta (trend) = 0.003
  - Gamma (seasonality) = 1e-4

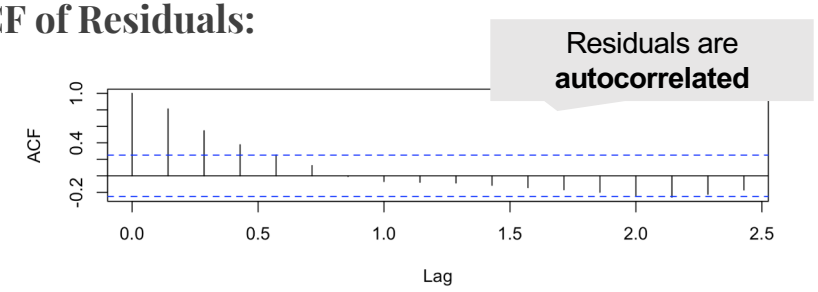



## Results on Test Data

- RMSE: 55,765
  MAE: 35,141




**ACF of Residuals:**



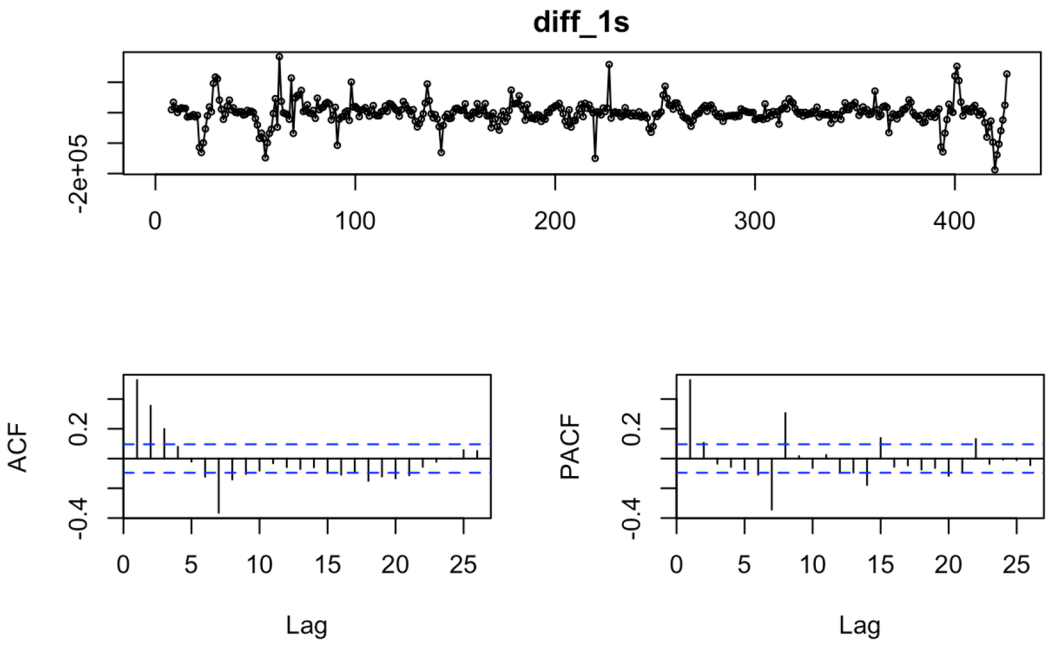

**Ljung-Box Test:**

Residuals are NOT independent (p-value = 2.31e-10)

# Model Fitting – sARIMA

**Modelling Phase Outline**

- Order one of seasonal differencing




- Model Estimation - “Maximum Likelihood Estimation”
- Model Combinations:

  sARIMA(1,0,0)(0,1,1)[7]

  sARIMA(2,0,0)(0,1,1)[7]

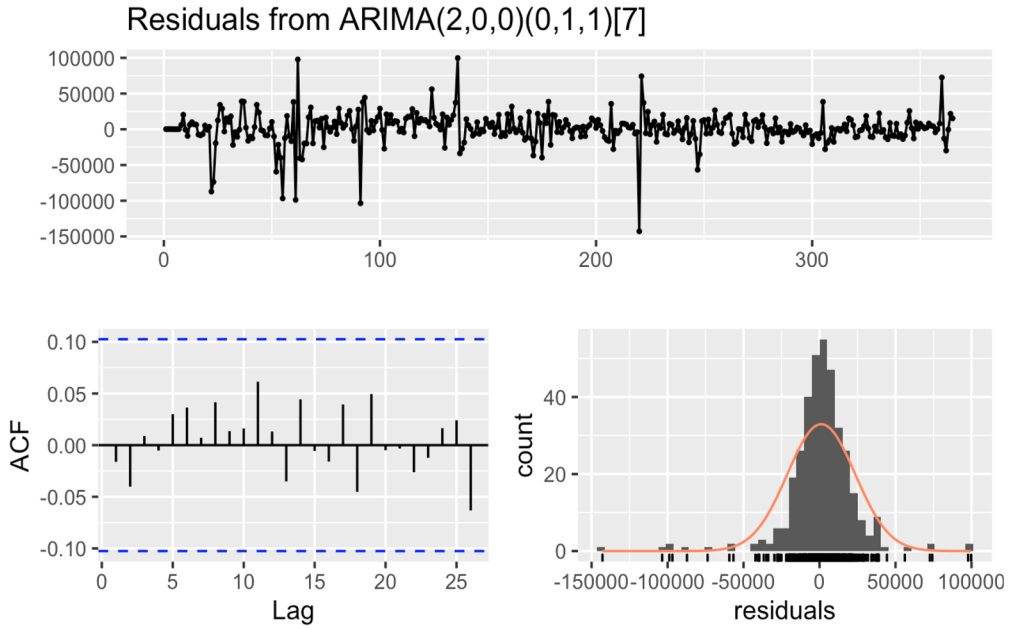
  ARIMA(2,0,1) - auto.arima

| | AICc | BIC |
|---|---|---|
| **Model 1** | 8246.2 | 8257.8 |
| **Model 2** | 8226.1 | 8241.5 |
| **Model 3** | 8746.8 | 8766.1 |

# Model Fitting – sARIMA

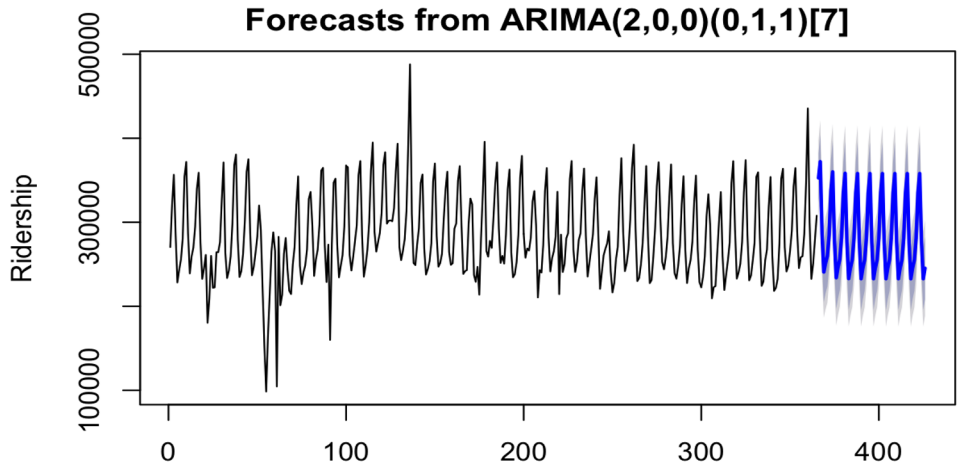
## Model Diagnostic

- Model 1: Residuals of are not fully white noise
- Model 2: Residuals of are white noise
- Model 3: Residuals of are not fully white noise




## Forecast

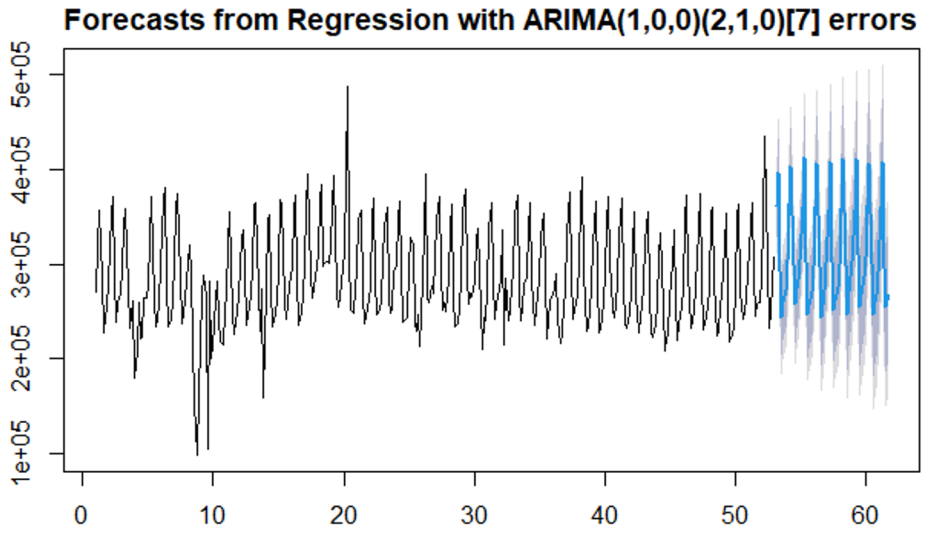
- **Simple Forecasting Methods (Benchmark)**
  - Average Method
  - RMSE: 51407
  - MAE: 40426
- **Forecast using Model 2**
  - RMSE: 42979.81
  - MAE: 30065.31



# Model Fitting – Regression w/ auto.arima errors

## Modelling Phase Outline

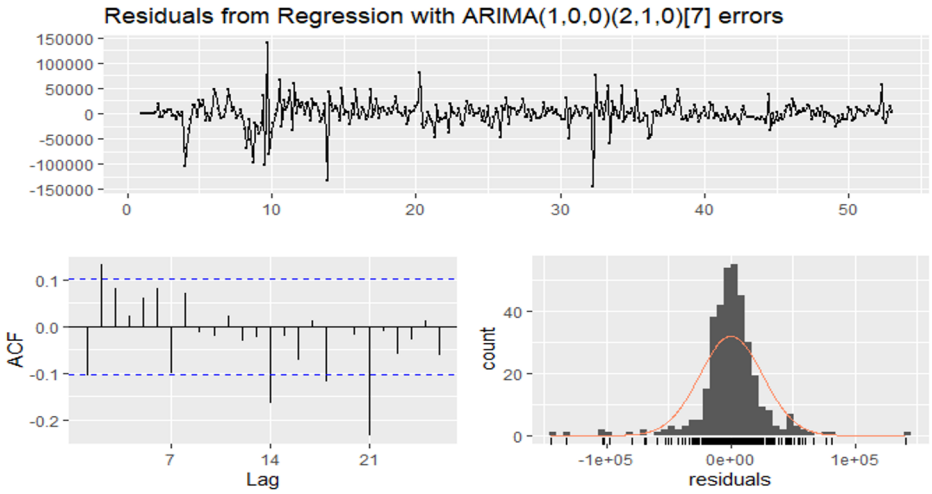
- Fitted using auto.arima
- Resulted in ARIMA(1,0,0)(2,1,0)[7] for errors
- Residuals are not considered time-independent




## Results on Test Data

- RMSE: 50,484
- MAE: 30,115

**ACF of Residuals:**



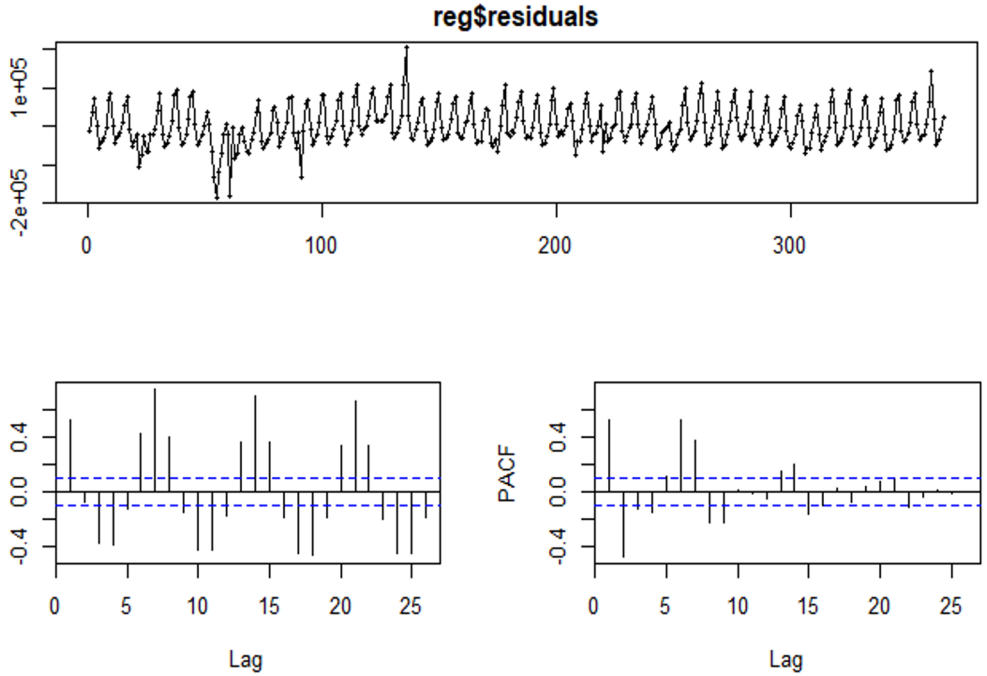

**Ljung-Box Test:**

Residuals are NOT independent (p-value = 1.559e-05)

# Model Fitting – Regression w/ ARIMA errors

**Modelling Phase Outline**

- Fitted linear regression to analyze errors:




- As with previous sARIMA model, we identified an AR model with seasonal component most likely at 7
- We will fit the sARIMA for the errors given the best previous sARIMA model

# Model Fitting – Regression w/ ARIMA errors

**Results on Test Data**

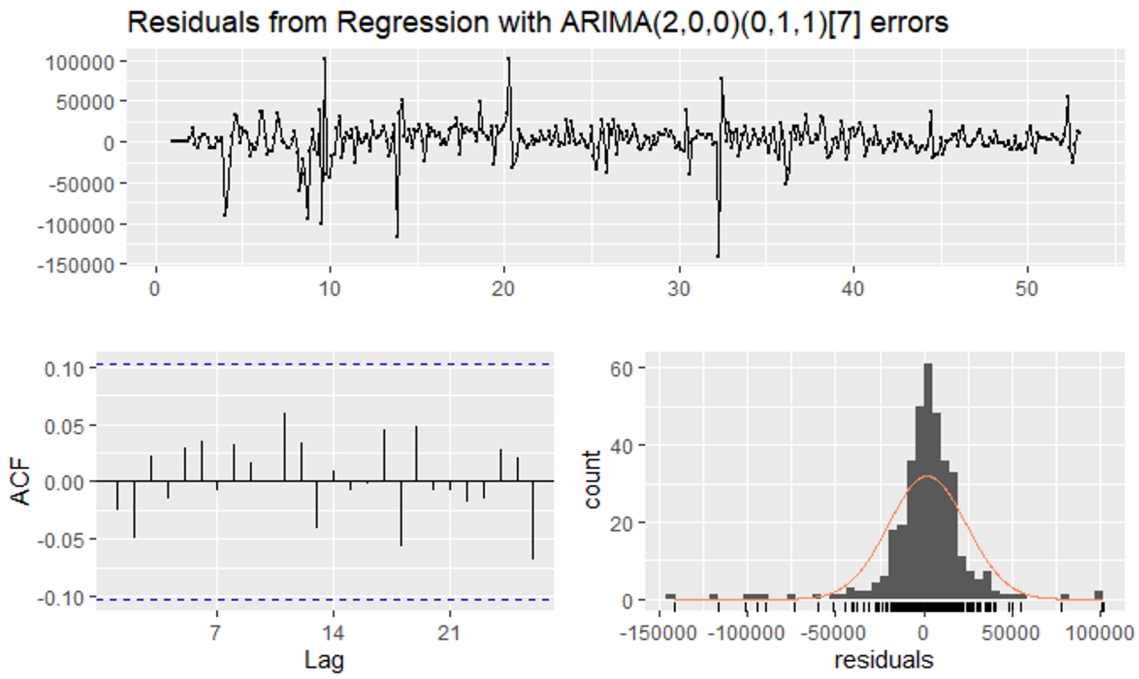
- RMSE: 42,678

  MAE: 27,021

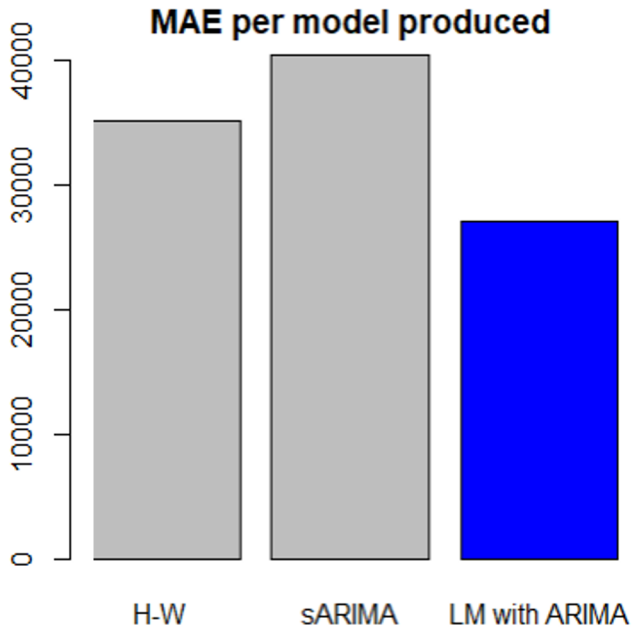
**Ljung-Box Test:**

Residuals ARE independent (p-value = 0.7368)

**ACF of Residuals:**



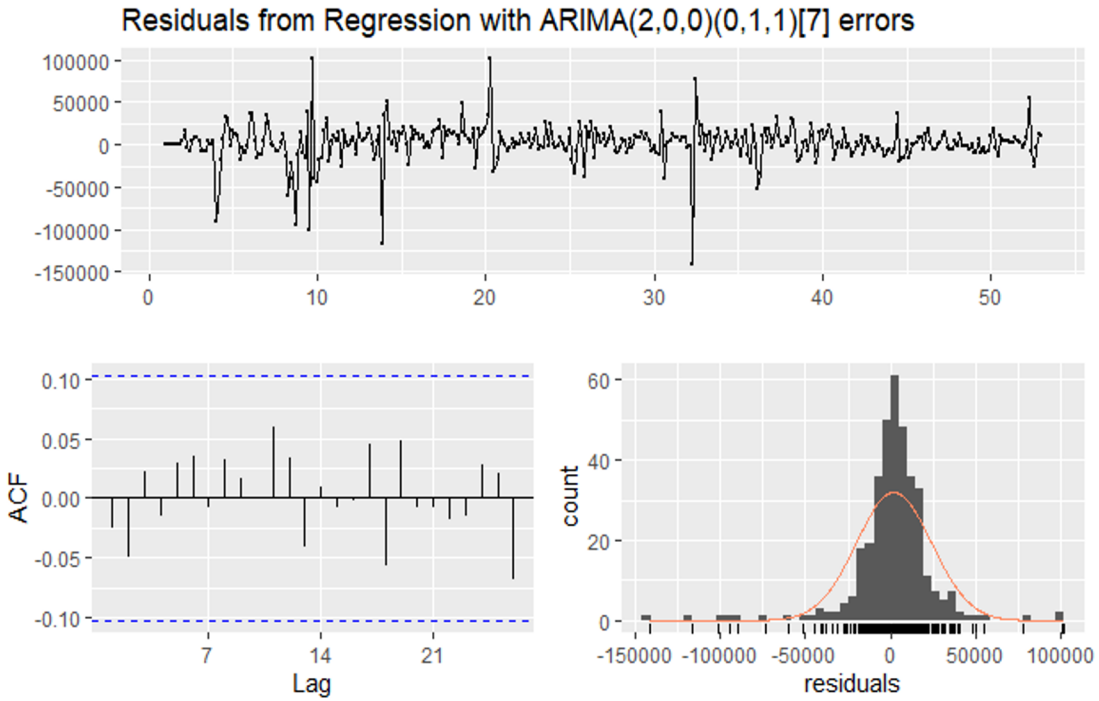
# Model Selection




- Not surprisingly, LM with ARIMA was the best models because it considered the weather variables that we would expect to influence ridership

# Model Evaluation

**ACF of Residuals:**

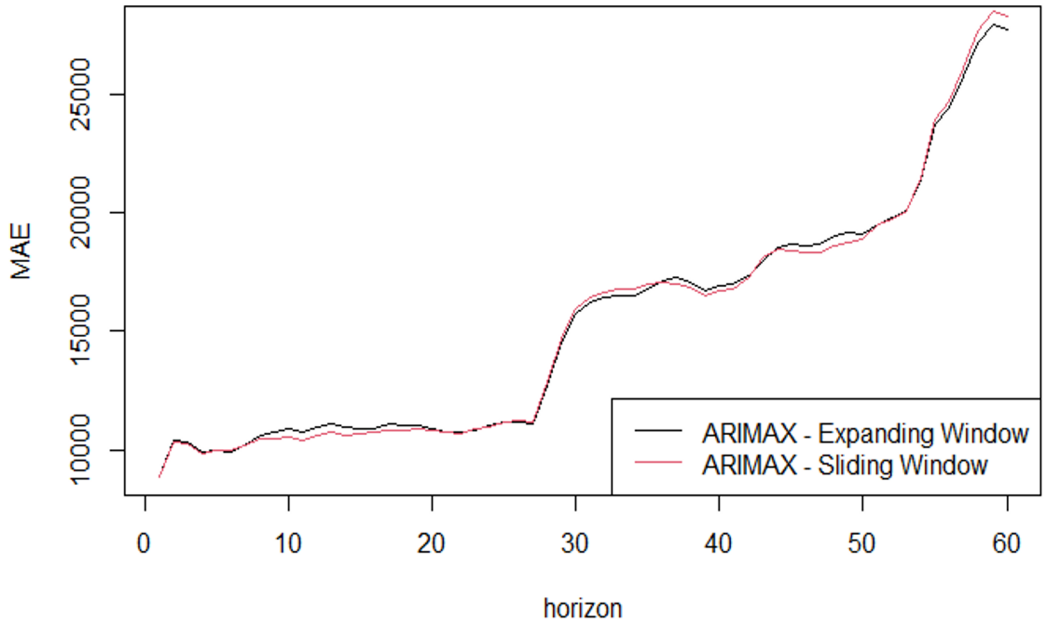



**Ljung-Box Test:**

Residuals ARE independent (p-value = 0.7368)

**Cross-validation**

Errors are stable until holiday periods are introduced, the two sharp rises of errors were caused by Thanksgiving and Christmas



# 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

# References


- 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