Bike-Sharing Network Analysis

- Python codes on Jupyter notebook -

Photo credit: uptown minneapolis | Courtney Coherent (wordpress.com)

Introduction

In this mini project, we will look at bike-sharing data from Pronto Cycle Share, a short-lived public bike sharing company operated in Seattle, U.S. between October 2014 and August 2016.

The bike sharing dataset was sourced from Kaggle: Cycle Share Dataset | Kaggle

Thanks to the generosity of some bike sharing companies, bike sharing data could often be found online for public download. This enables any interested parties to conduct researches which potentially can help in improving the bike sharing systems. Although most of these datasets does not contain individually identifiable data that enables one to conduct in-depth travel pattern analysis of different riders clusters, we may still be able to gain some understanding about the general travel pattern and habits of riders in a particular region, and possibly make comparison between different bike sharing systems then form a general view of bike sharing behaviour, good and bad practice etc.

Introduction (cont.)

So, given the attributes in this dataset, I will demonstrate some possible statistical, temporal and network analysis that most of us as a layman, can attempt. Specifically, we will focus on these questions:

1. How does the overall bike network look like? Where do riders usually travel to and from? How does it change over time? Are we able to derive insights about the citizens' travel habits from this trend?

2. Pronto set two types of pricing for two different groups – members and short-term pass holders. How are their ride patterns different? Could there be improvement in pricing?

3. We expect weather impacts ridership. How susceptible are different groups of Pronto riders to the adverse weather conditions?

Background about Pronto Cycle Share

The bike share system was launched in October 2014 in Seattle, U.S. and had approximately 500 bikes in 50 stations.

In 2015, it ran into major funding issues after the City of Seattle put any further fundraising on hold while awaiting council approval to purchase the system.

The Seattle City Council eventually bought the system in March 2016. However, the system continued to suffer from less than expected ridership, revenue and lack of funding.

In early 2017, the City Council decided to shut down its operation and divert the funding to other transportation programs.

Reference: https://en.wikipedia.org/wiki/Pronto_Cycle_Share#cite_note-4

Dataset Description

The dataset contains individual trips, stations locations and daily weather information, in the period of October 2014 to August 2016:

Trips Data
Trip ID
Start date, time
Stop date, time
Bike ID
From station ID, name
To station ID, name
User type
Gender
Birth year

Stations Data
Station ID
Station name
Latitude
Longitude
Install date
Install dock count
Current dock count
Modification date

Weather Data
Date
Temperature
Dew Point
Humidity
Sea Level Pressure
Visibility
Wind Speed
Gust Speed
Precipitation
Events (e.g. rain)

Data Pre-processing

The major data pre-processing steps performed before constructing the network:

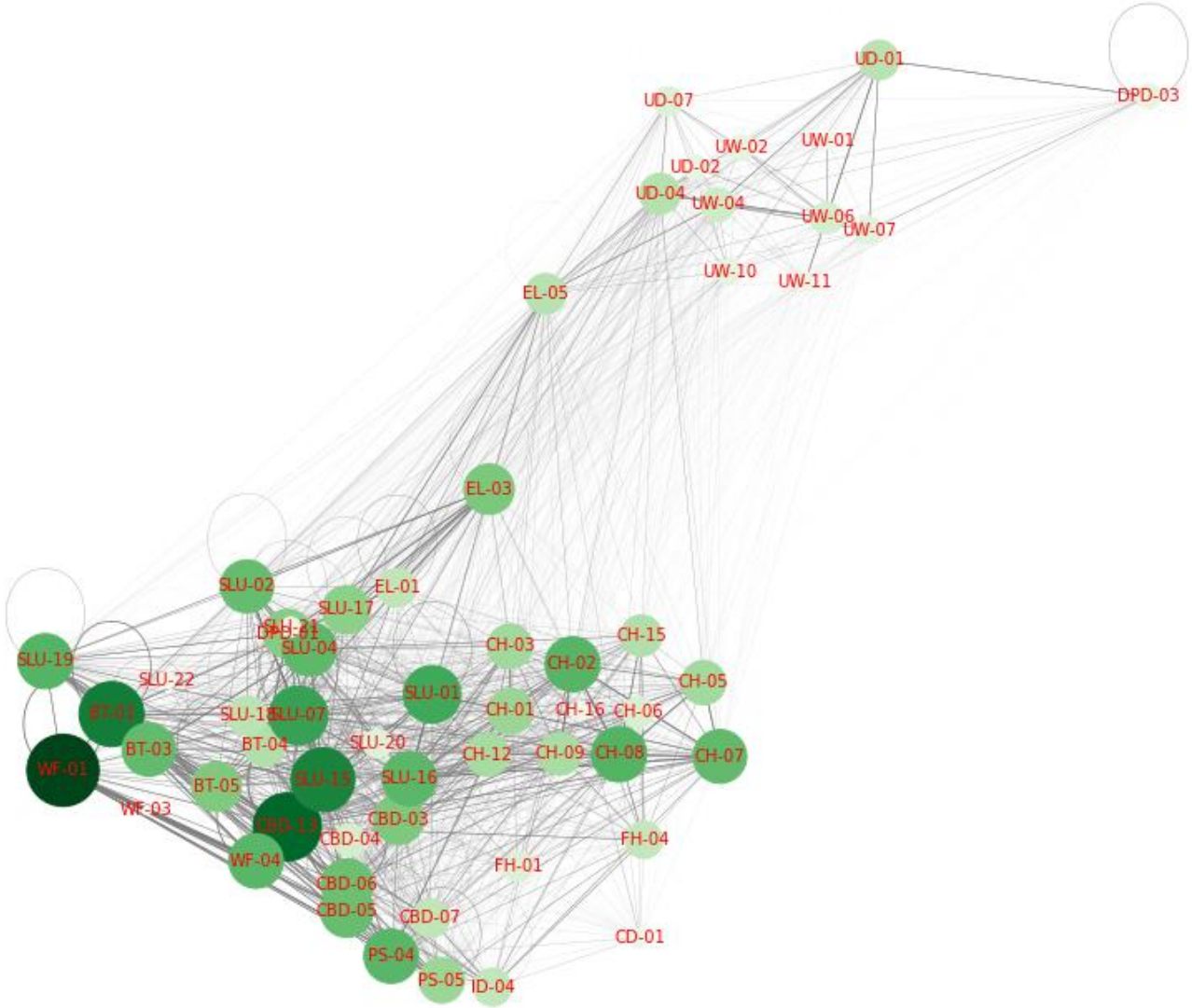
- Summing up the number of trips by each quarter
- Removing three of the nodes that have very few trips and do not have station locations data

Constructing the Network

NetworkX package in Python was used to create and study the network structure.

From personal experience, having briefly touched on igraph – another similar popular network analysis package used by scholars, I found NetworkX being easier to use in Python, mainly because the official documentation is more comprehensive and there are more guides and solutions when searching NetworkX-python on Google or Stack Overflow. It seems R users favour igraph though.

Link to NetworkX guide:
https://networkx.org/documentation/stable/reference/introduction.html



Constructing the Network

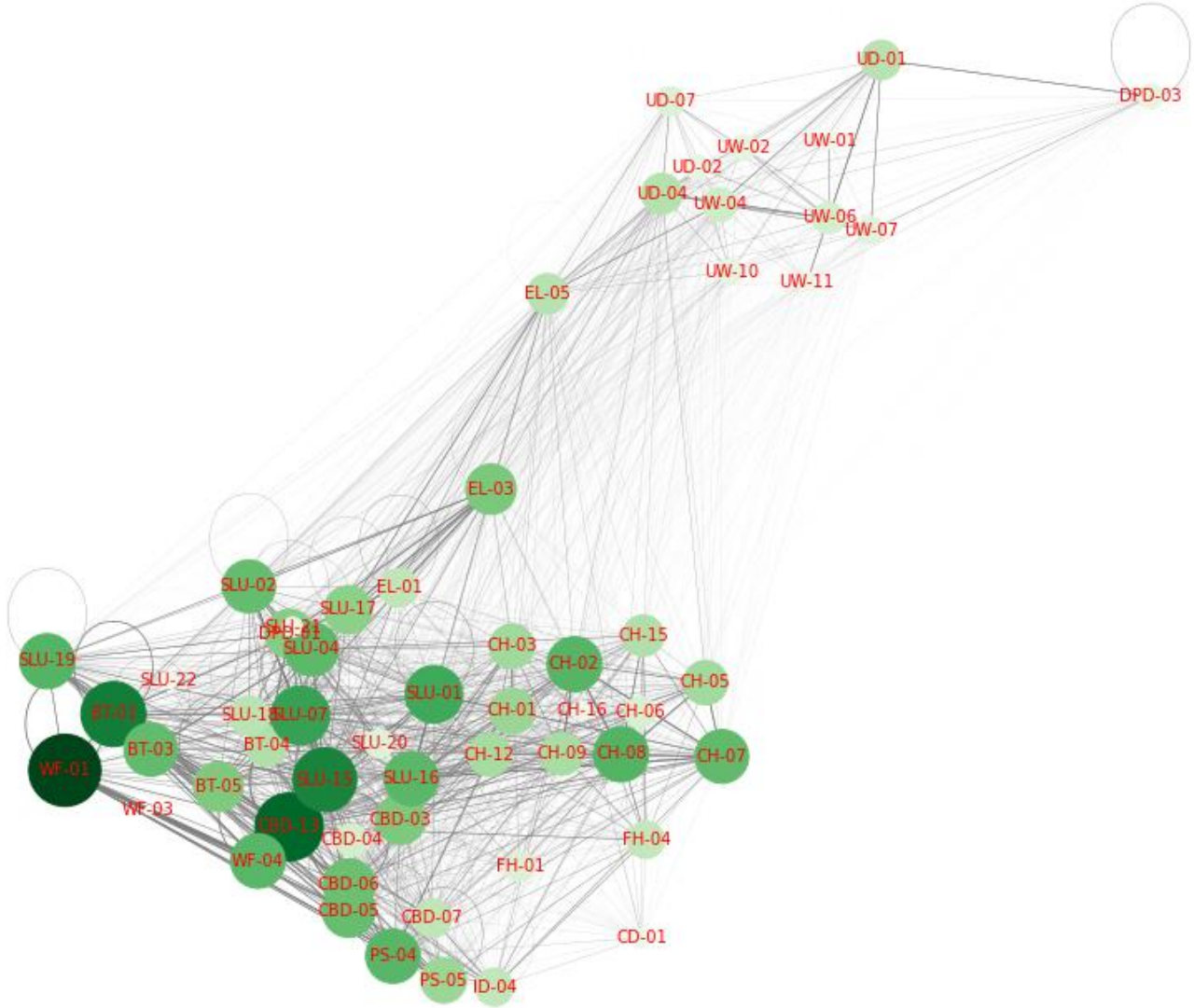
Nodes

- Bike stations

Edges

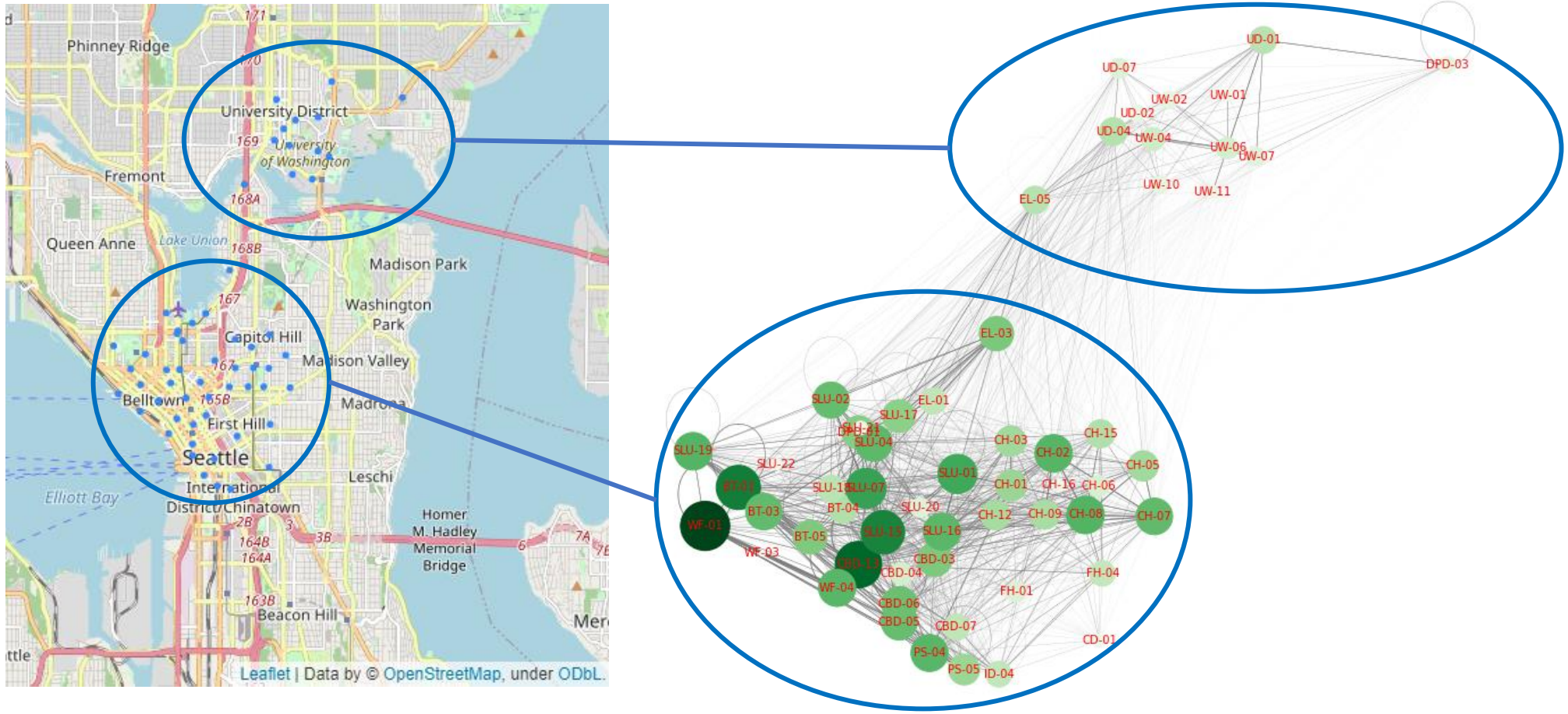
- Directed trips between stations
- Weighted by the number of trips

There are 58 nodes and 2930 edges in the network during the full operation period.



Constructing the Network

The node position in this network visualization is set by the longitude and latitude values of the bike stations. The values correspond to the actual locations in Seattle downtown and the University District:



Map source: https://www.kaggle.com/code/skantbksc/network-analysis-of-pronto-cycle-network/notebook

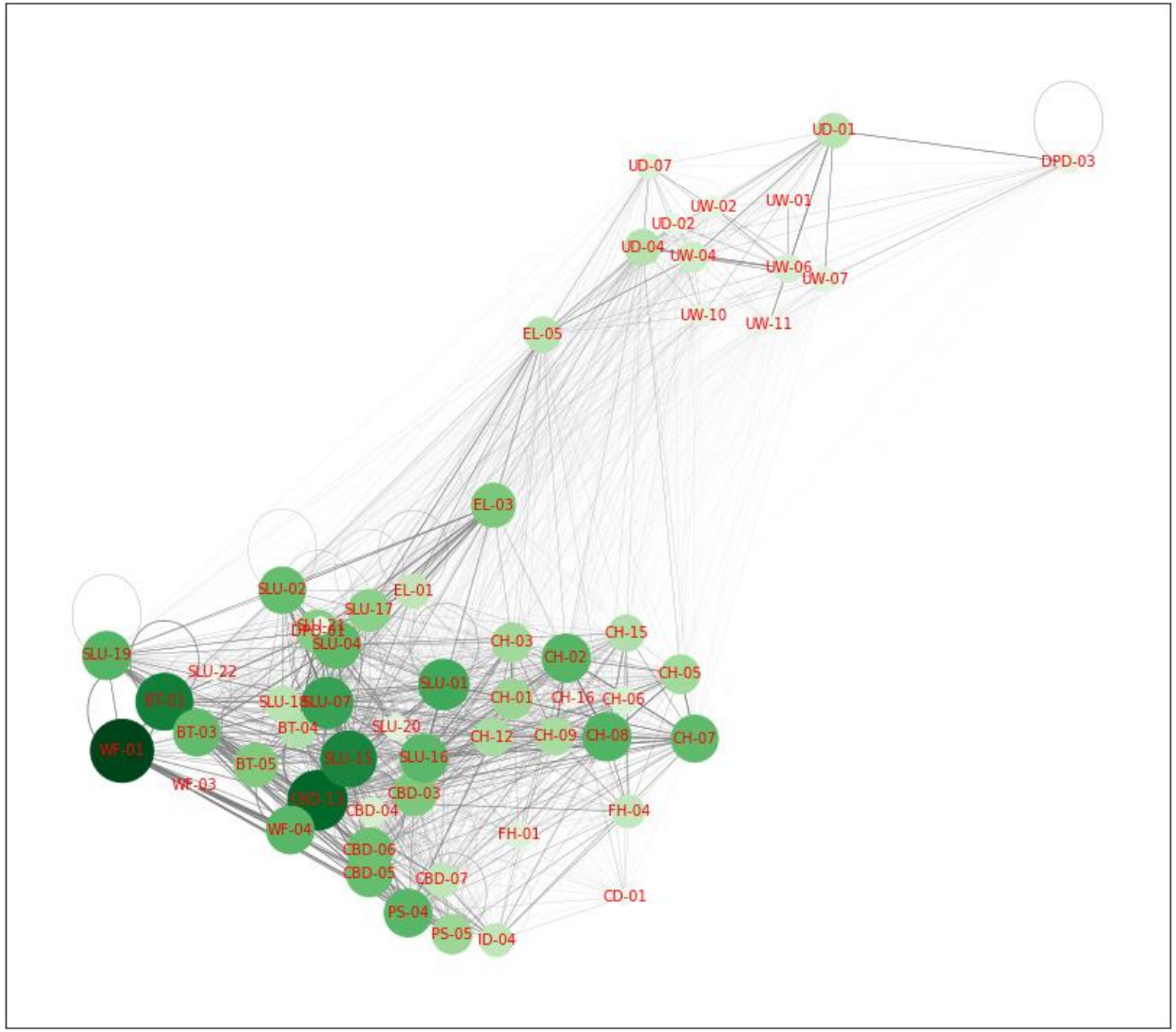
Network Analysis

The colour and size of the nodes correspond to the weighted degree centrality of the stations.

Since the weight is defined as the number of trips between each station, the weighted degree centrality here is basically the total number of trips involving the stations as start point or end point.

Noticed that WF-01 and CBD-13, in darkest colour, are the busiest stations. University District stations at the top have fewer trips.

WF-01 is next to a pier and at the waterfront, which is logical as a busier spot.

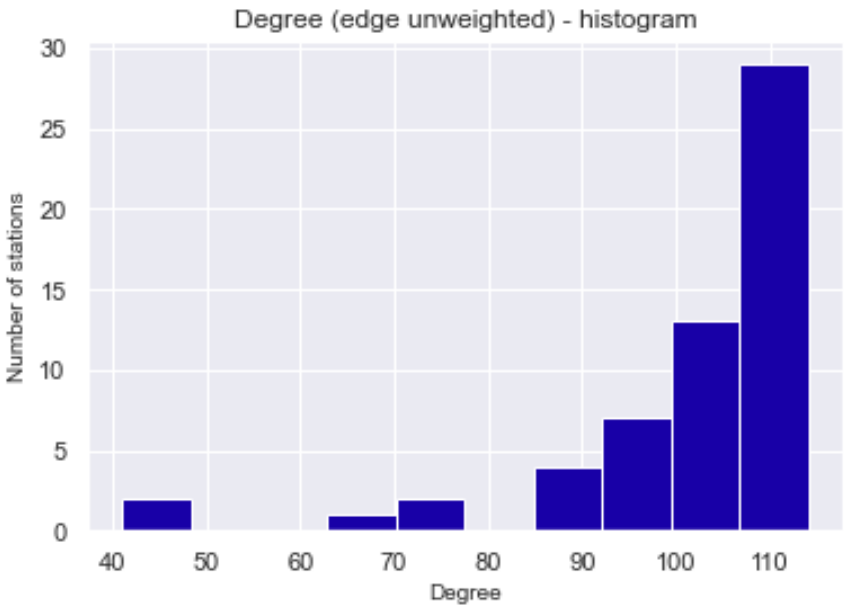


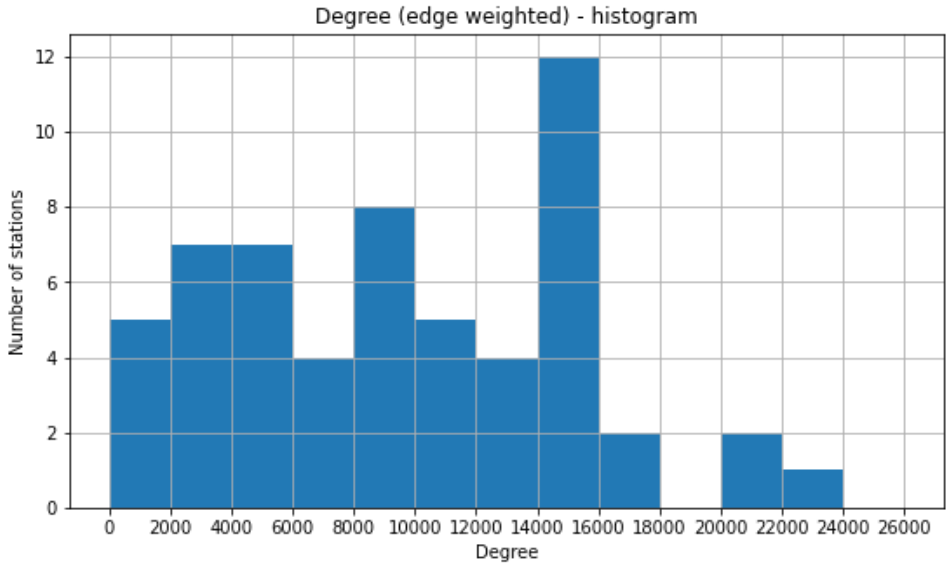
Network Analysis

Weighted degree centrality vs **Unweighted degree centrality**

The unweighted degree represents the number of connections each station has. A connection exist even if there is only a single trip. This is not conductive to further analysis as many stations have at least one trip with each other, as reflected by the right skewed degree histogram. The network is almost fully connected in this sense.

In weighted degree histogram, we can see that some stations had less than 2000 trips while a few had more than 20000 during the period. The distribution appeared relatively uniform.





Riders User-type and Pricing

Member and Short-Term Pass Holder

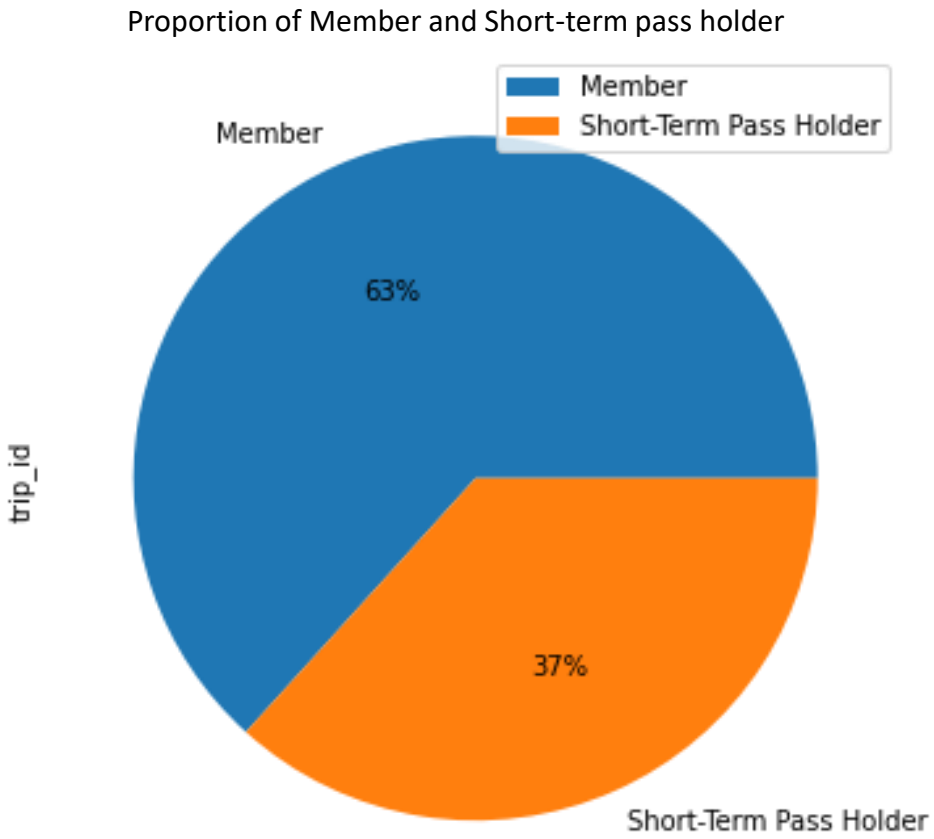
According to Pronto's tumblr blog, the pricing for bike share is as follows:

- 24-hour access pass: $8
- 3-day access pass: $16
- Annual membership: $7.95/mo or $85 upfront

The above passes included unlimited trips up to 30 minutes each.

Trips longer than 30 minutes incur overtime fees:

Overtime Fees	
30 – 60 minutes	$2 additional charge
Each additional ½ hour	$5 additional charge



Reference: https://prontocycleshare.tumblr.com/post/101094854343/a-note-on-pricing

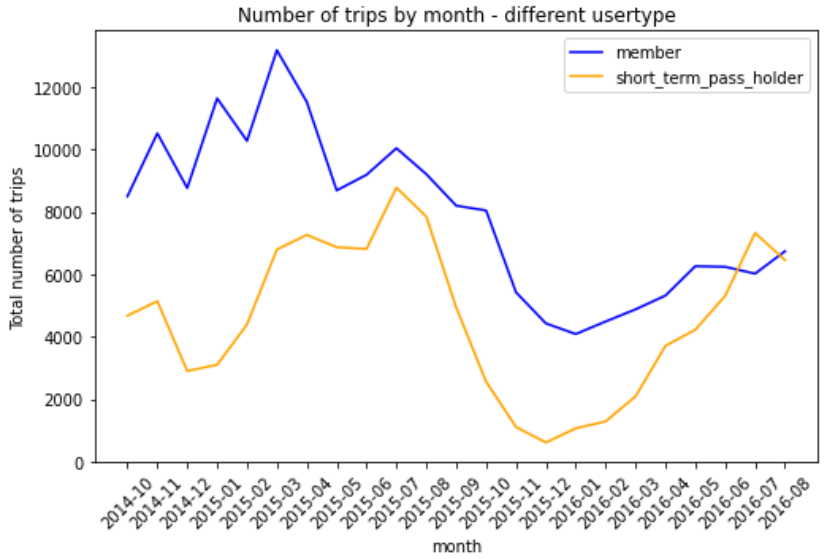
Ridership Over the Operation Period

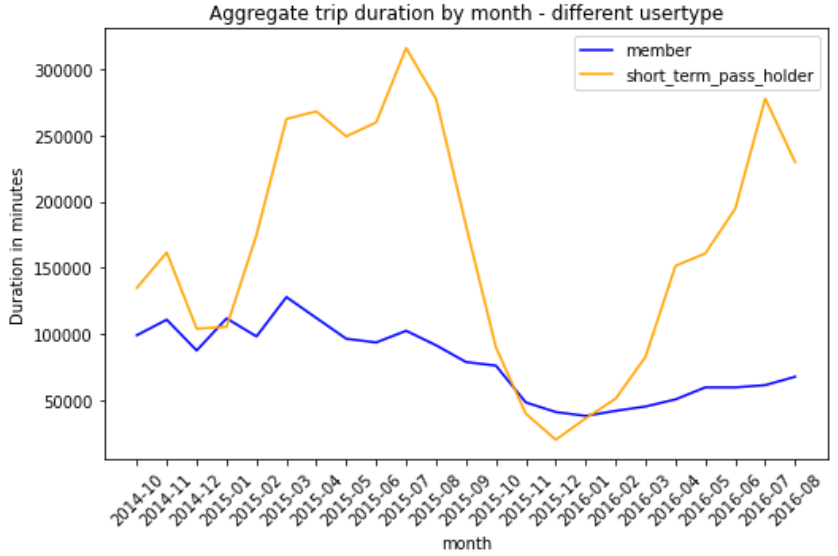
Number of trips and duration

- Member appears to make more trips than short-term pass holders. However, short-term pass holders recorded higher total trip duration. This may be due to the pricing structure of the passes and the purpose of biking.
- Ridership of short-term pass holders has higher fluctuation than members.

Seasonal fluctuation

- There appears some seasonal fluctuation in ridership due to weather.
- However, there was much lower ridership in 2015 winter than in 2014 winter when the bike share just started operation, and ridership did not return to 2014 initial level after winter. Why is this so?





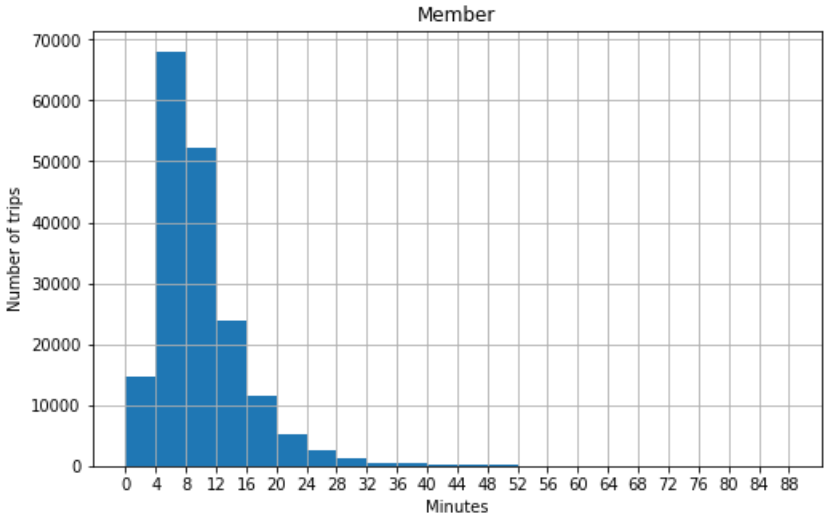
Trip Duration of Members vs Short-Term Pass Holders

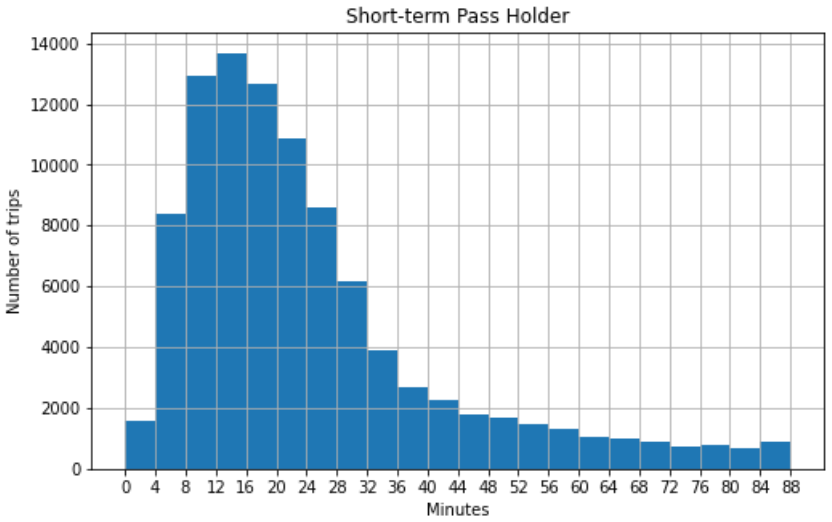
Members

Most are short trips within 4-12 minutes. Riders appear to be quite aware of the extra charges beyond 30 minutes as there are very few trips beyond this 30 min cut-off point.

Short-term pass holders

Overall trip durations are longer, with more riders going beyond 12 minutes. There are quite some riders ride beyond 30 minutes. The short-term pass holders are either, much more willing to pay (e.g. tourists), or not as aware about the extra charges beyond 30 min.

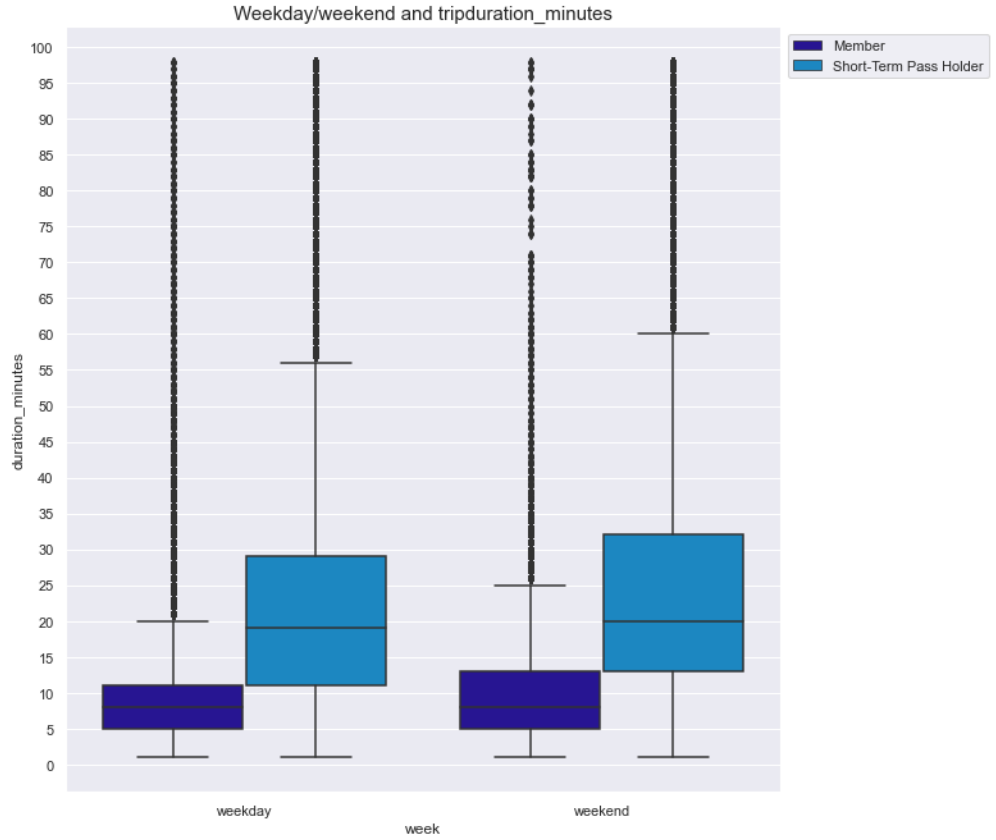


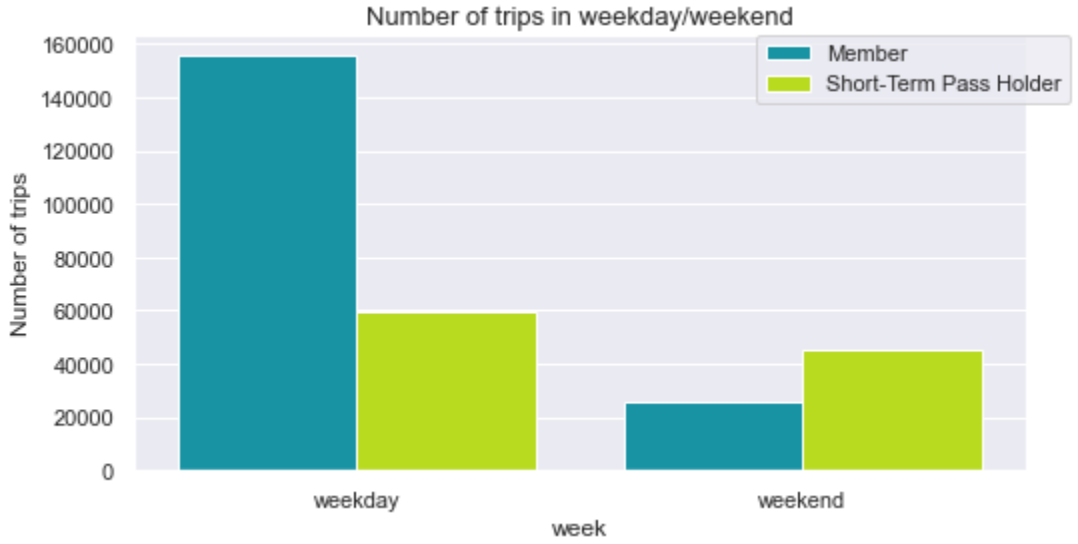


Trip Duration of Weekday vs Weekend

For both member and non-members, trip duration is just slightly longer in weekends as compared to weekdays.

But there are more short-term pass holders than members using the bikes during weekends. This further confirmed that members tend to use the bikes for regular commutes while short-term pass holders use it for leisure purpose.

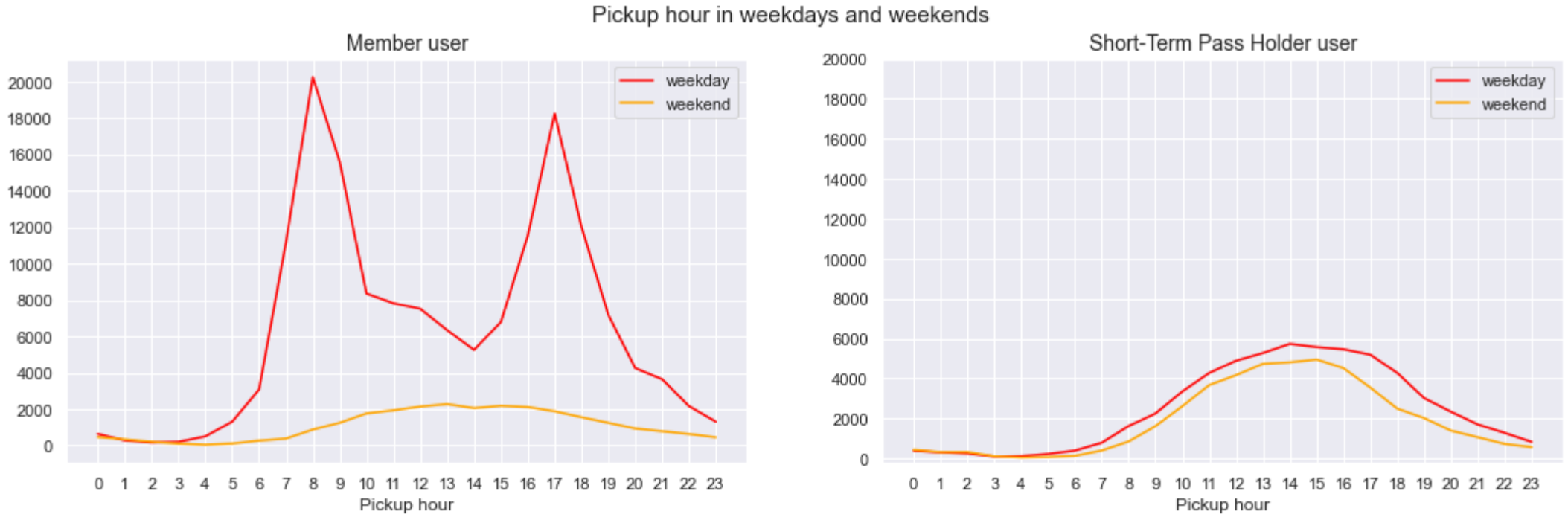




Pickup Hour of Members vs Short-Term Pass Holders

Members riding the bikes for commute is also reflected on the pickup hours, with 8 to 9 am and 5 to 6 pm being the peak hours.

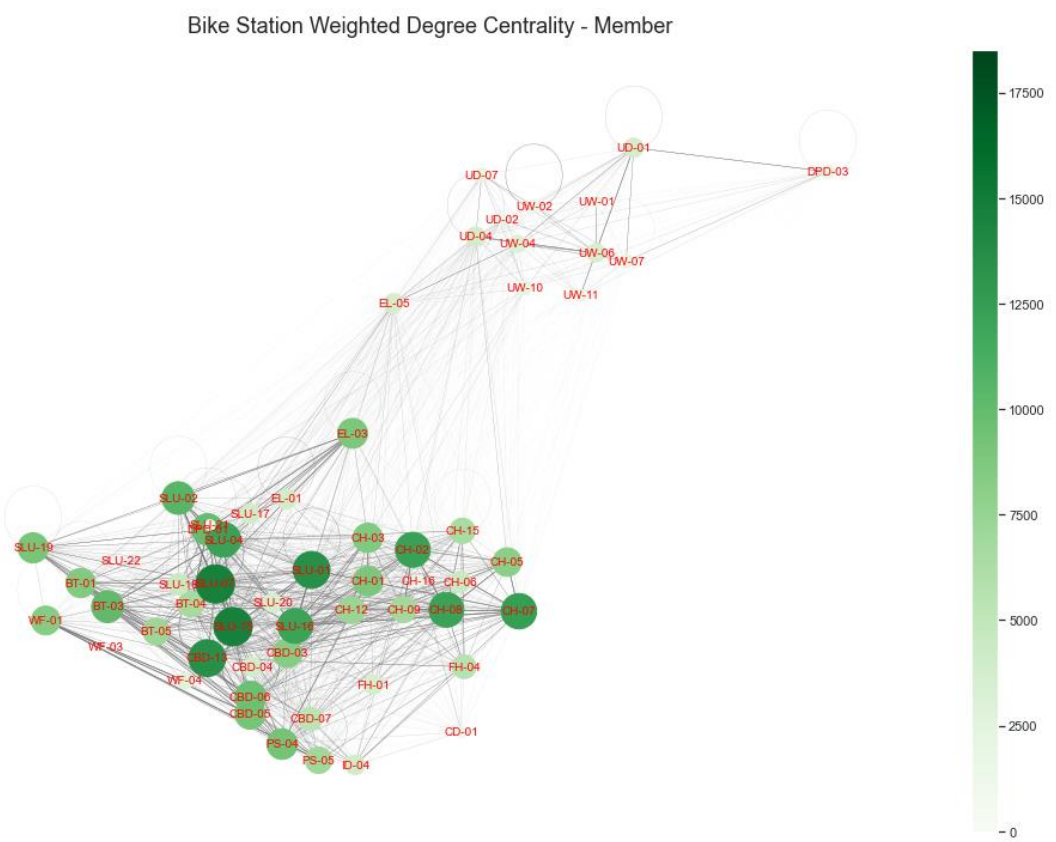
For members in weekends, or for short-term pass holders in general, the pickup hours do not have such rush-hour peak pattern.

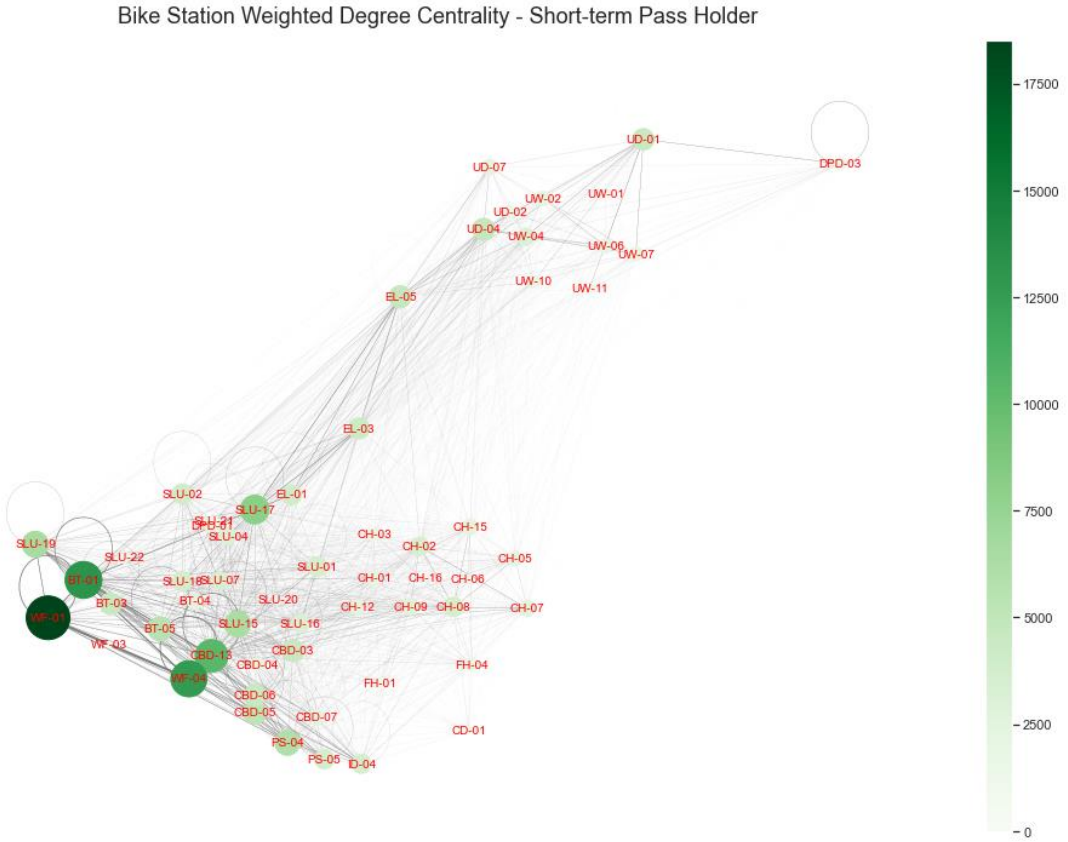


Hotspots for Members and Short-Term Pass Holders

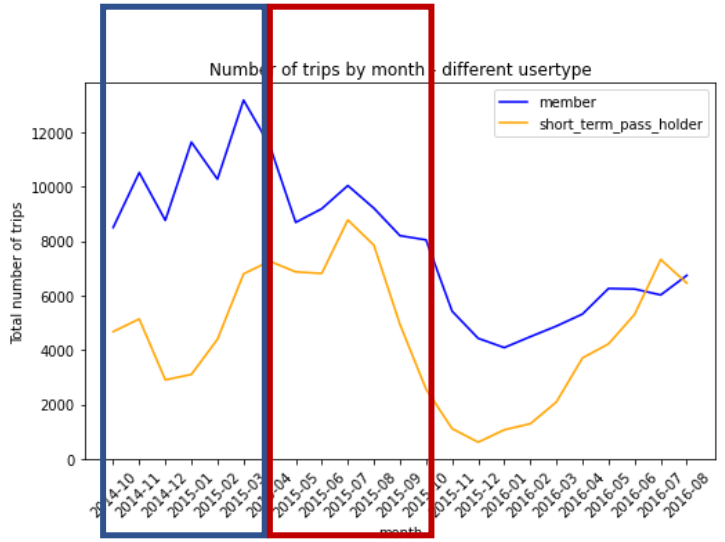
Members biked around most of the stations in the downtown area.

Short-term pass holders focused on a few leisure locations that members not as frequently going, for example WF-01 Pier69 and WF-04 Seattle Aquarium, which are along the waterfront; and SLU-17 Lake Union Park. Furthermore, they had more trips that start and end at the same stations (the circular edges), reflecting the non-commute purpose.



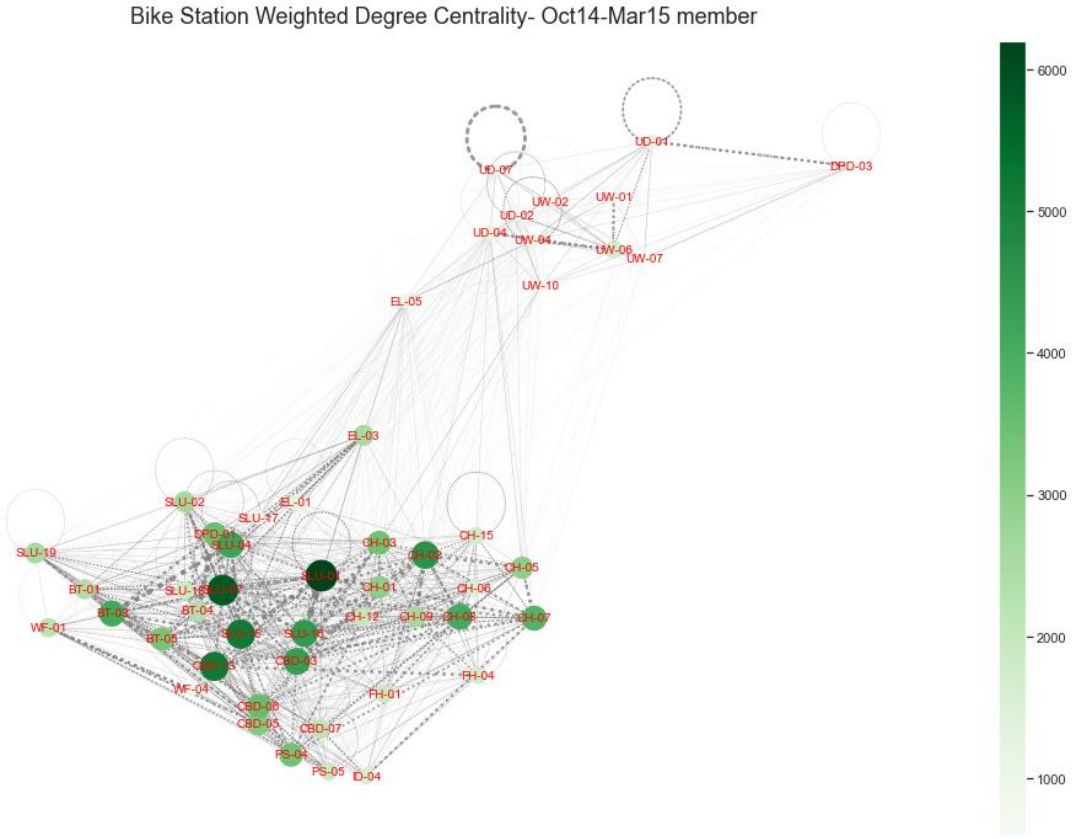


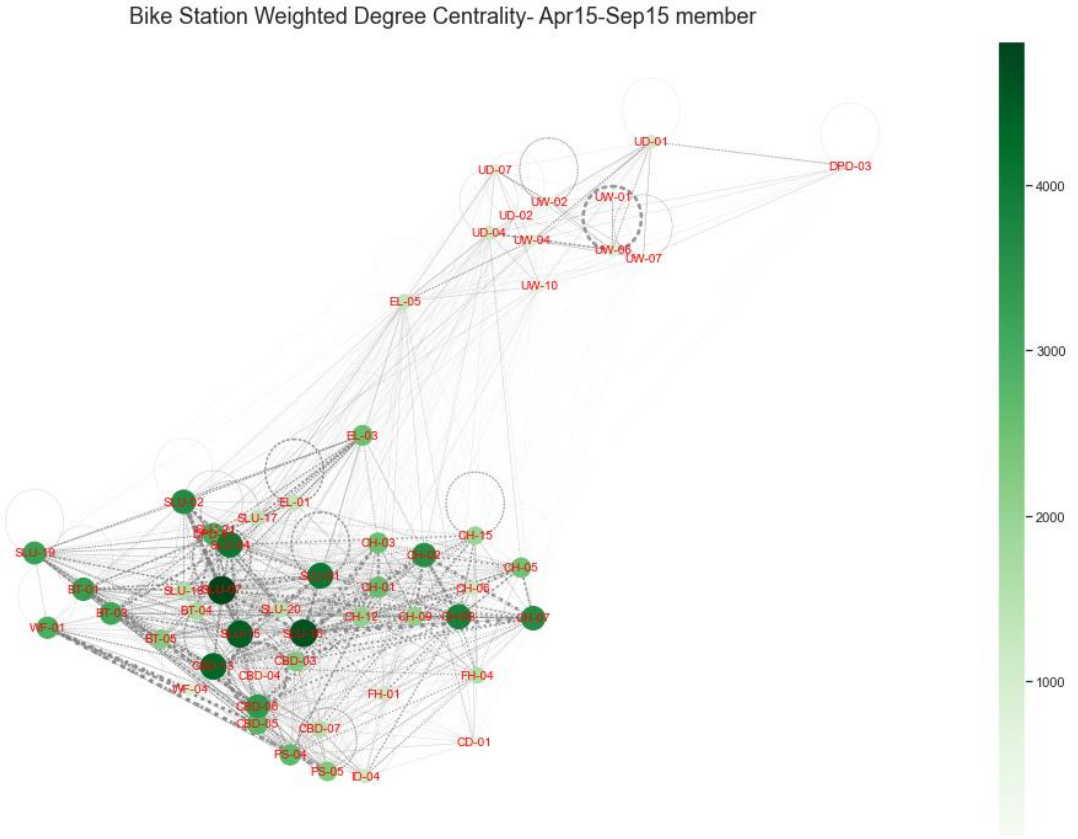
Hotspots for Members – Temporal Analysis



Are there any changes in network changes over different time periods? Here we look at the network for members in two different six-month periods.

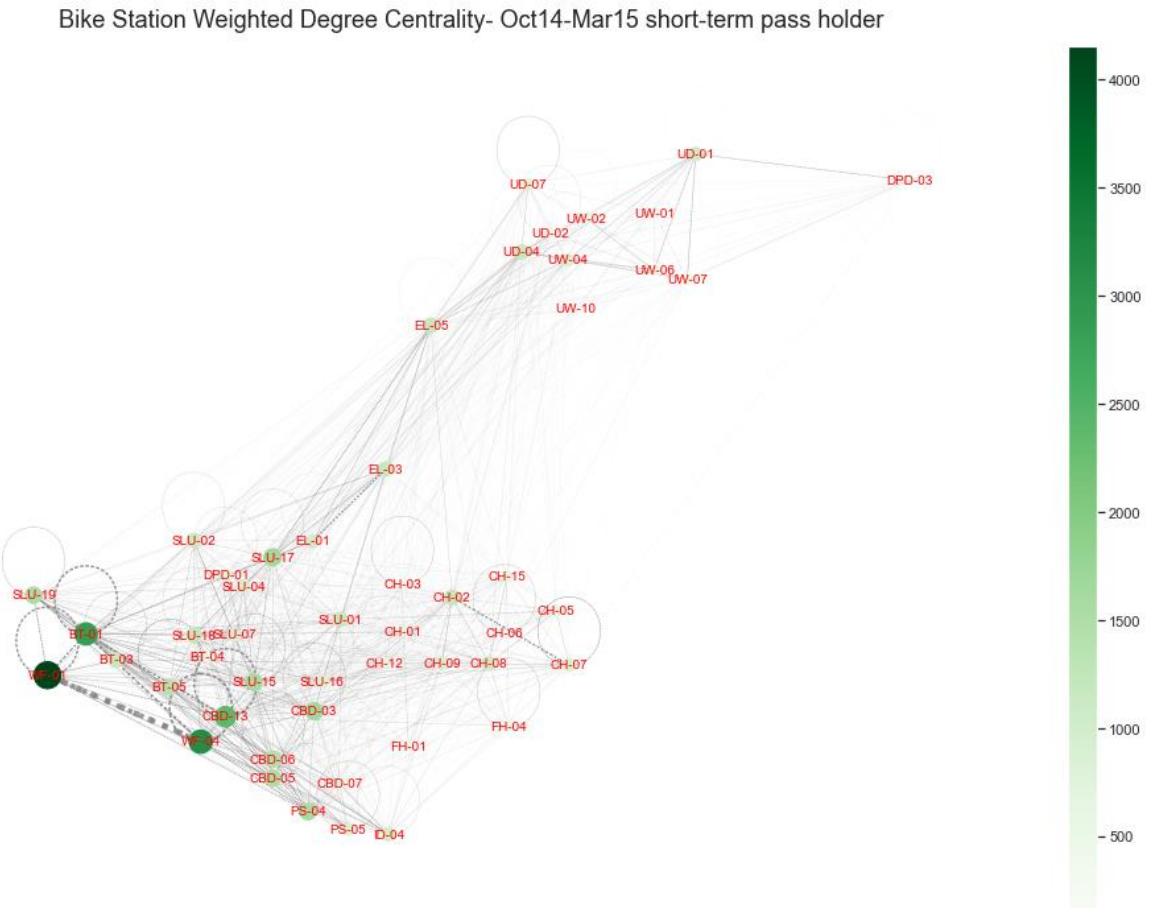
Noticed that during the initial launch period, the most popular station is SLU-01 but it lost popularity over time (why?). However, the darker green colour in more nodes in the latter period indicates that riders had more balanced use of the bike network overall.

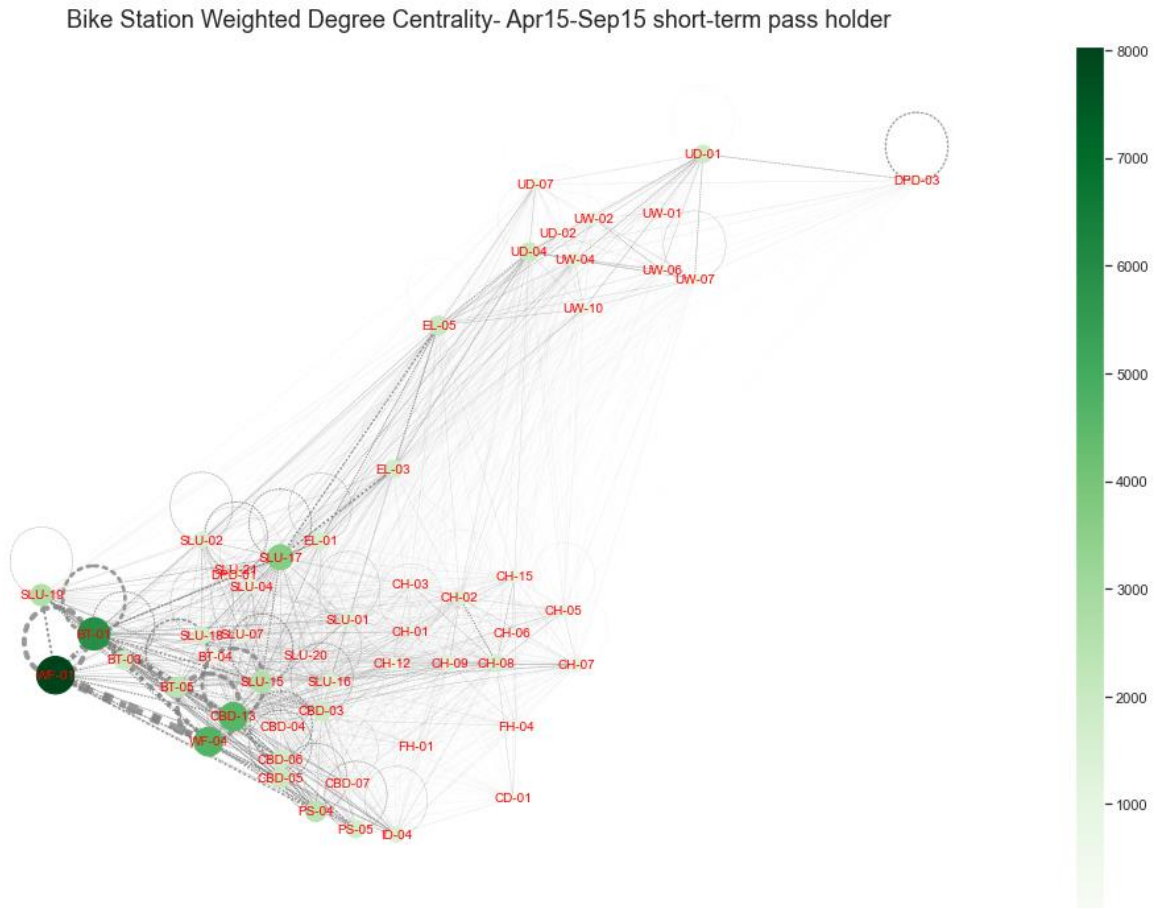




Hotspots for Short-Term Pass Holders – Temporal Analysis

For short-term pass holders, there are not much changes in the hotspots, which are the stations near the waterfront and the park. But the bigger nodes and denser edges do indicate an increase in ridership in the second 6-month period.



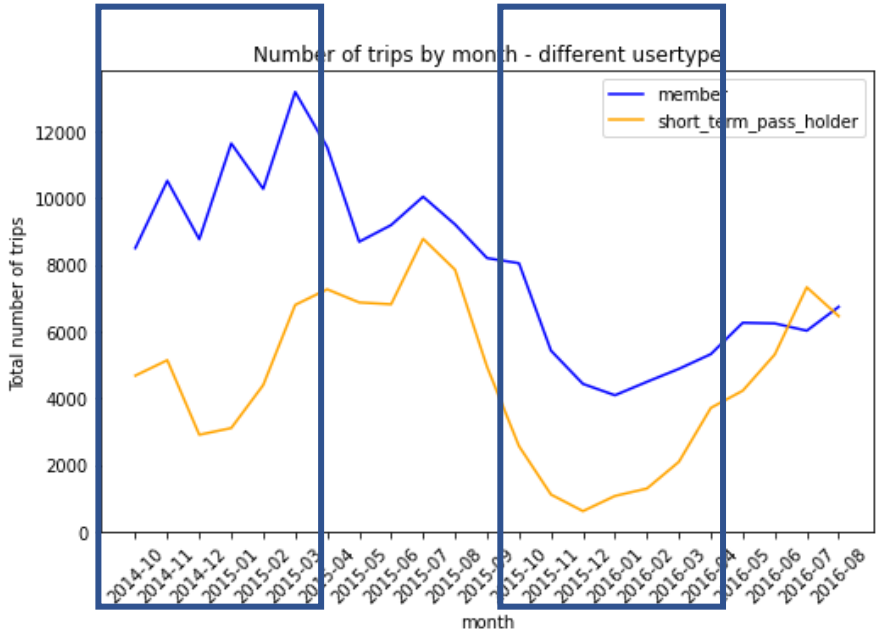


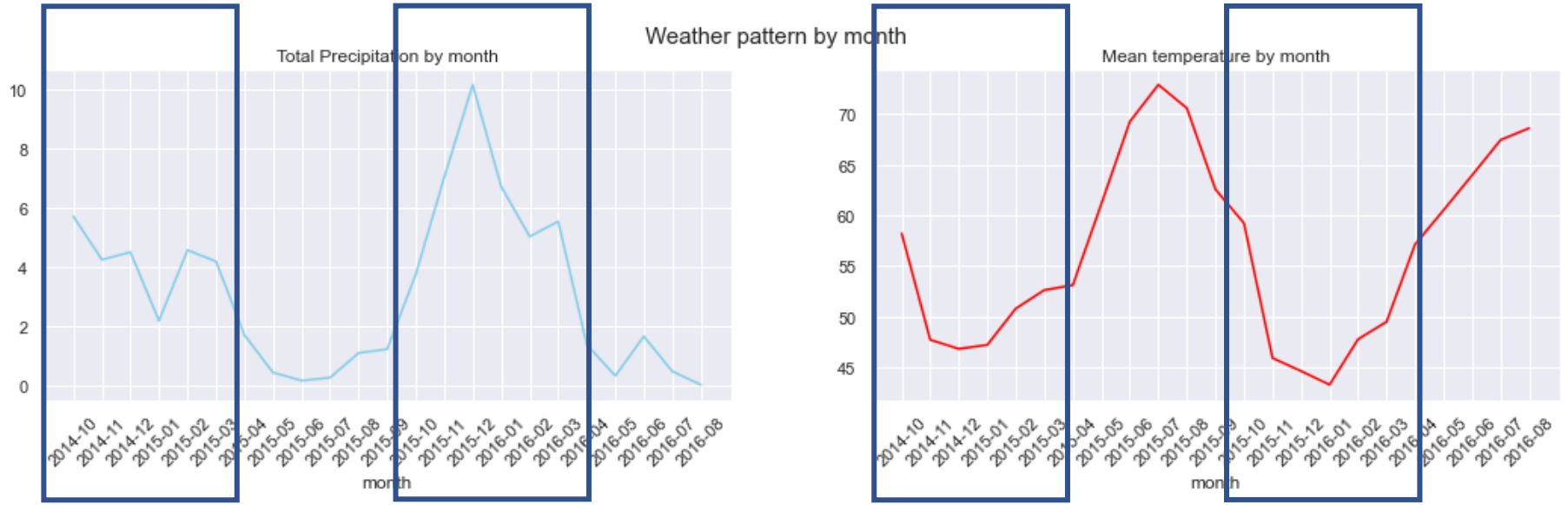
Weather and Ridership

The weather pattern graphs show that 2015 winter has higher total precipitation and lower mean temperature than 2014 winter. These appear as negative factors to the lower ridership in the second winter for Pronto.

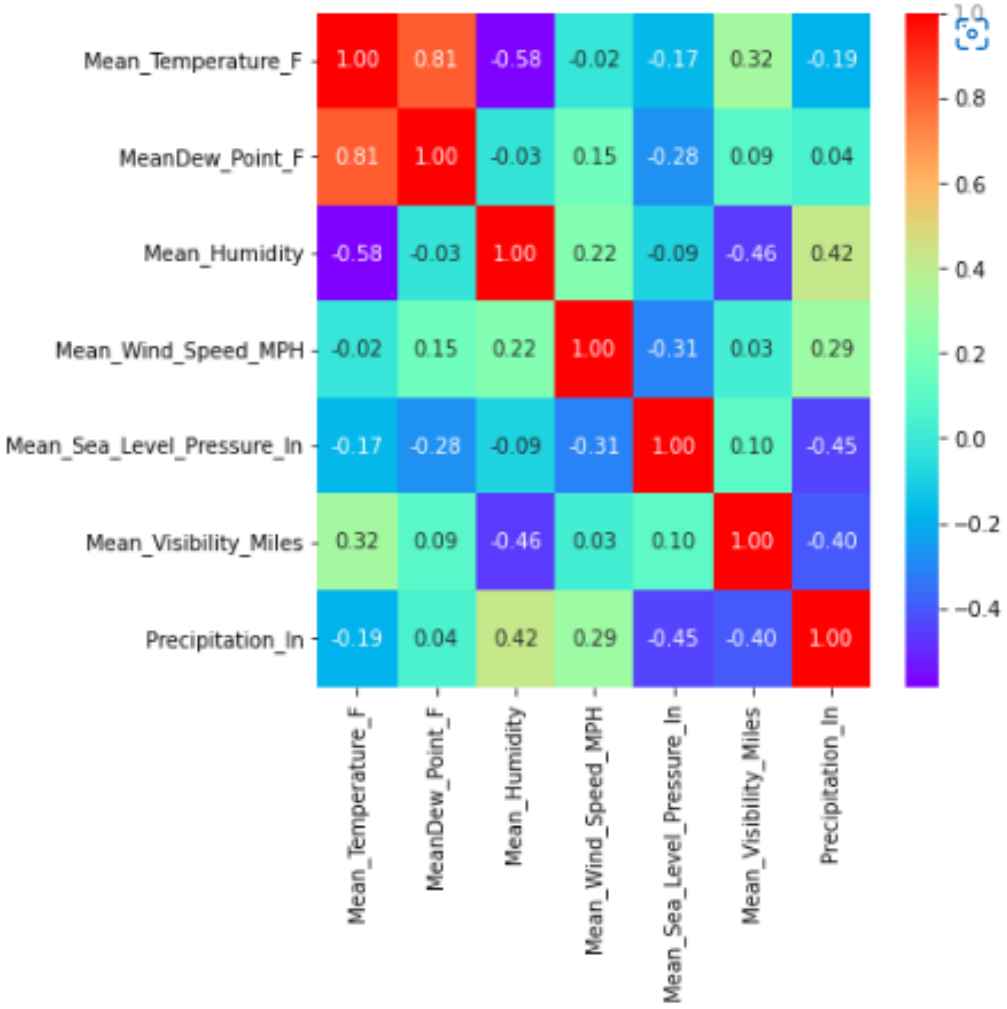
How large is the correlation between different weather conditions and ridership?

Let’s try some regression analysis.





Correlations between different weather features



Before running the regression, let's check the correlations first.

Some weather features correlated with each other in medium to high extent:

- Mean dew point vs Mean temperature +0.81
- Mean humidity vs Mean temperature -0.58
- Mean humidity vs Mean visibility -0.46
- Mean humidity vs Precipitation +0.42
- Mean sea-level pressure vs Precipitation -0.45

If we include all the features, there would be multicollinearity issue, which make the model inaccurate.

Therefore, we select only mean temperature, mean wind speed, mean visibility and precipitation as main features in the regression.

Correlation between Weather And Ridership

Hypothesis

Lower temperature (implying winter), lower visible distance, higher wind speed or higher precipitation, may discourage people travelling by bikes.

Results

Except for visibility distance*, regression results are expected.

- Higher temperature correlates with higher ridership.
- Higher mean wind speed and precipitation correlates with lower ridership.
- Precipitation appear to discourage ridership the most.

Weather features	Coefficient	P-value
Mean_Temperature_F	6.4203	0.000
Mean_Wind_Speed_MPH	-13.5448	0.000
Precipitation_In	-193.3012	0.000

*coefficient is not statistically significant when a four feature model including mean visibility is run.

Which sub-groups are more resilient to weather fluctuation

Member vs Short-term pass holder

- The explainability of weather conditions to daily total number of trips is higher for short-term pass holders, as reflected by higher R-square value.
- Members are more resilient to changes in temperature and precipitation, as reflected by lower coefficients value.

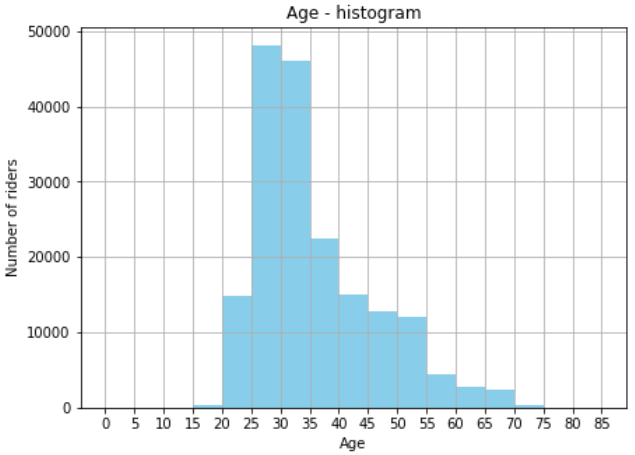
User type	R-square	Mean_Temperature_F	Mean_Wind_Speed_MPH	Precipitation_In
Member	0.065	1.0874 (0.031)	-7.3512 (0.000)	-83.2361 (0.000)
Short-term pass holder	0.316	5.3266 (0.000)	-6.2446 (0.000)	-109.2605 (0.000)
All	0.280	6.4203 (0.000)	13.5448 (0.000)	193.3012 (0.000)

(P-values of coefficients) in bracket

Which sub-groups are more resilient to weather fluctuation

Different age groups (members only)

- Age data is not available for short-term pass holders unfortunately.
- Impact of windspeed and precipitation on daily number of trips is larger for more mature age groups of 26-60 as compared to 16-25.
- Though, overall R-square is low as there are more factors other than weather influencing members' ridership.



Age Group	R-square	Mean_Temperature_F	Mean_Wind_Speed_MPH	Precipitation_In
16-25	0.064	0.2667 (0.000)	-0.8489 (0.000)	-7.8758 (0.001)
26-30	0.051	0.1512 (0.315)	-2.072 (0.000)	-22.7932 (0.001)
31-40	0.067	0.4235 (0.013)	-2.4970 (0.000)	-27.1320 (0.001)
41-60	0.056	0.2341 (0.085)	-1.7435 (0.001)	-22.3795 (0.000)
61-90	0.032	0.0120 (0.576)	-0.1898 (0.021)	-3.0545 (0.002)

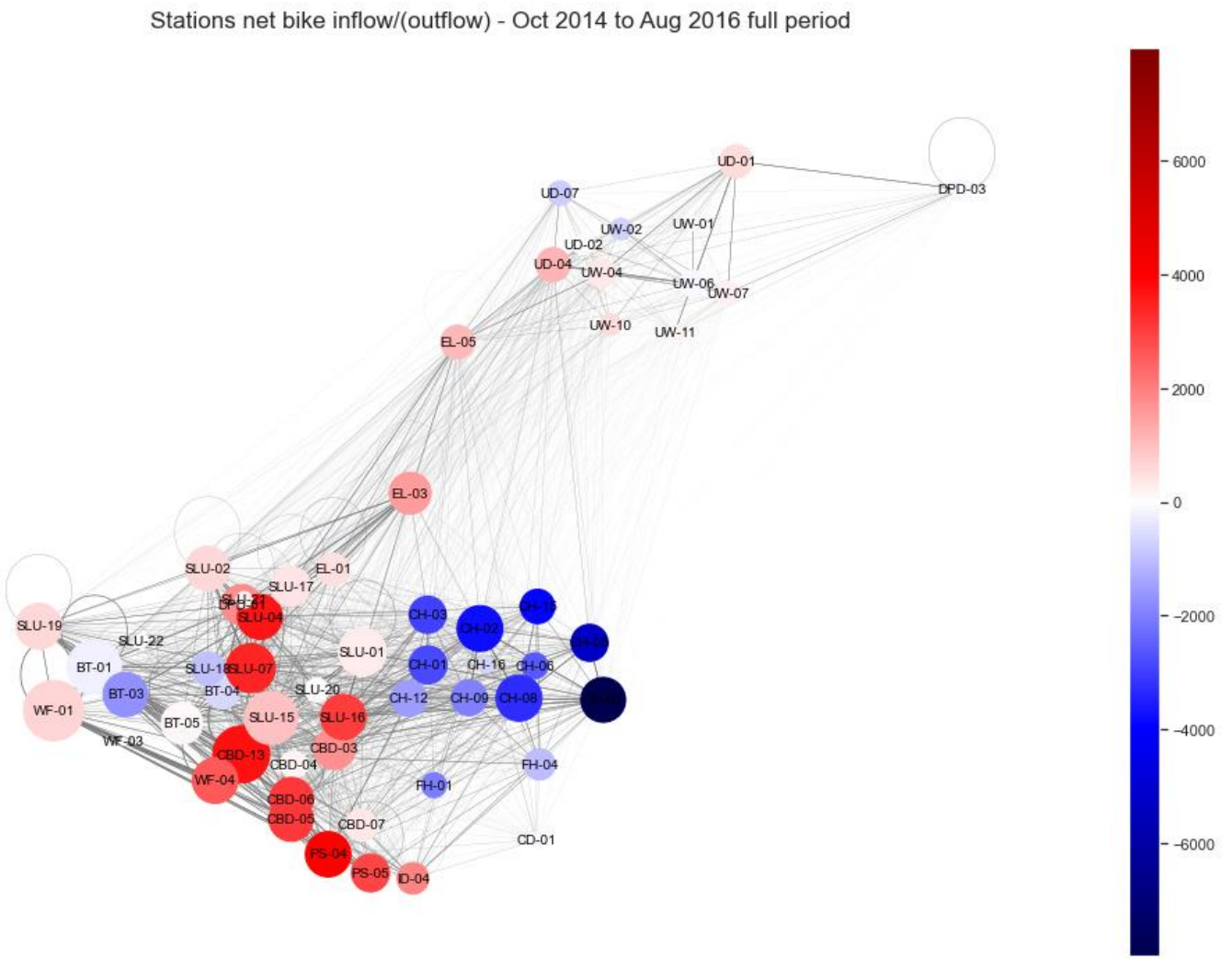
(P-values) in bracket

Stations net bike inflow / (outflow)

As with any other bike share systems, there could be imbalances in bikes inflow and outflow at any stations, and therefore bike sharing companies need to move the bikes around by truck regularly.

In this visualization, the size of nodes still represents the weighted degree centrality of the stations, while the colour represents the net inflow or outflow throughout the whole period of operation.

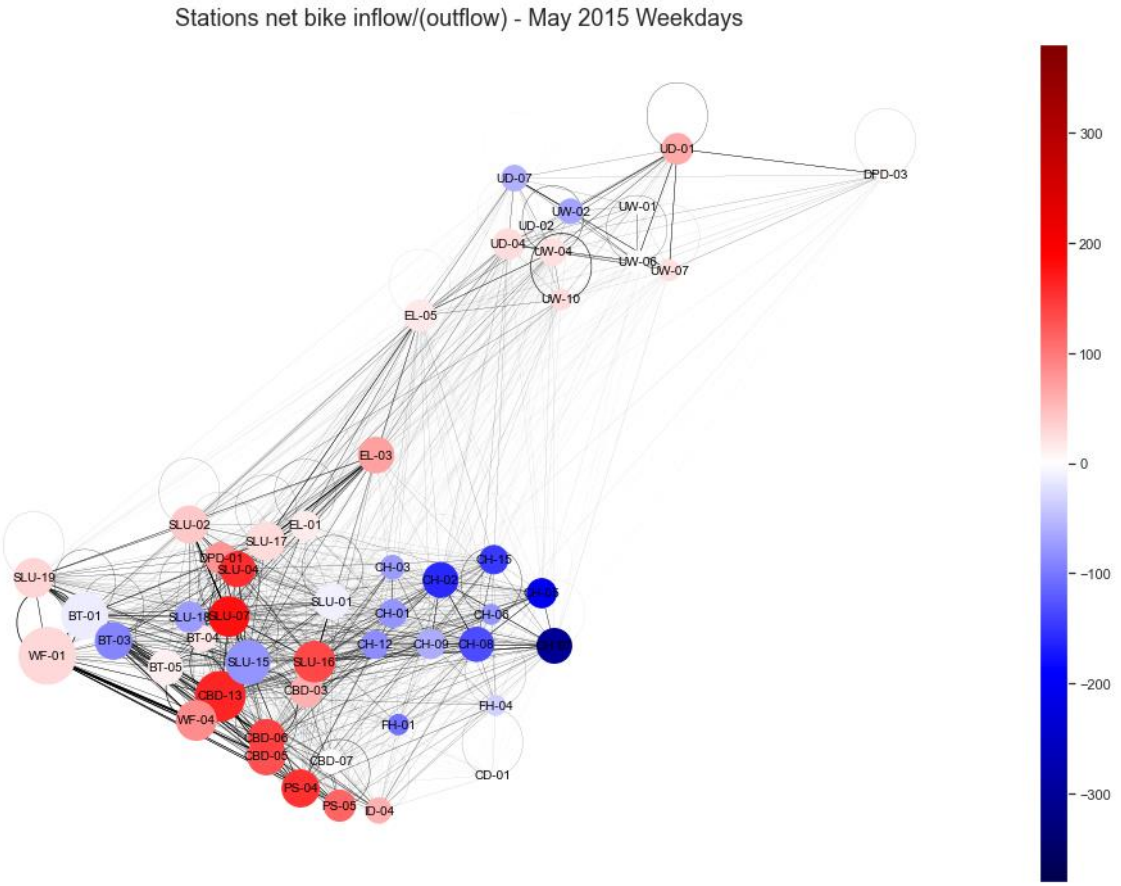
We can see that the CH (Capitol Hill) cluster had a heavy net outflow while the CBD and SLU cluster had net inflow.

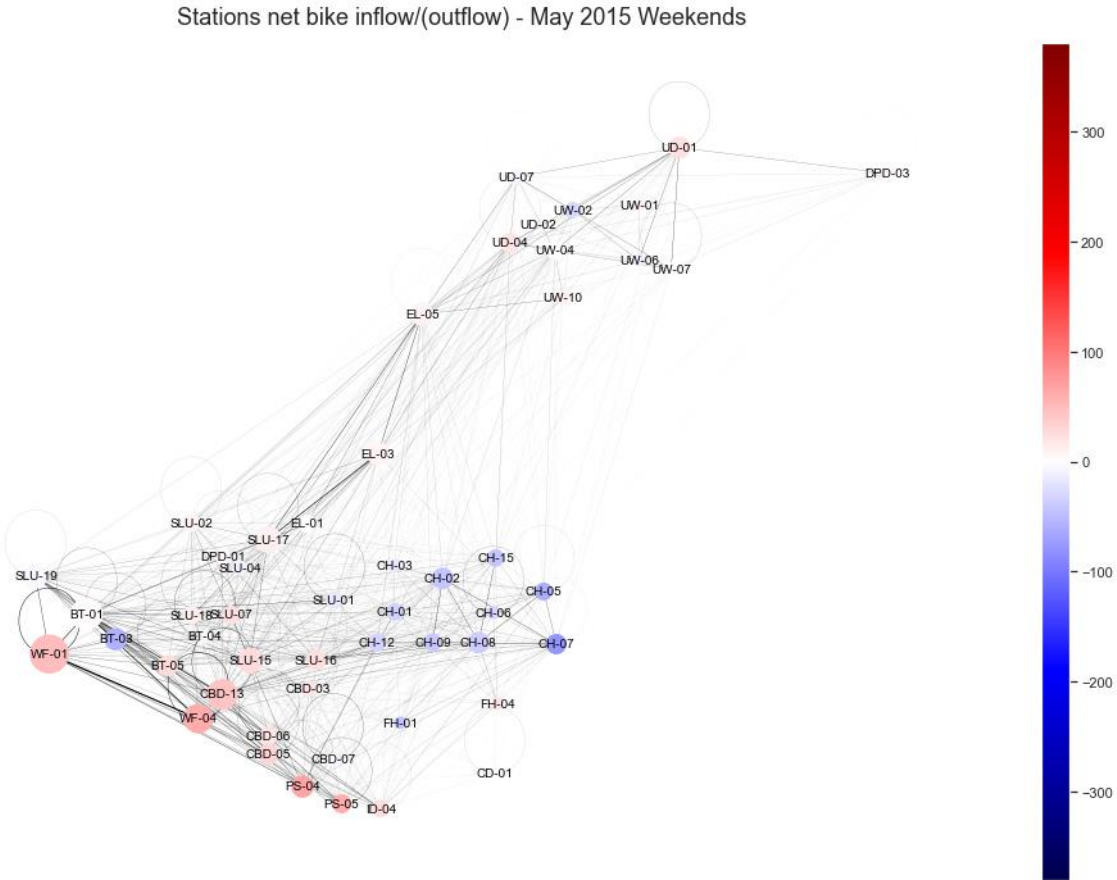


Stations net bike inflow / (outflow) – A specific *month*

The pattern isn't too much different when we look at a specific month, say May 2015.

Comparing May 2015 weekdays and weekends, although overall ridership is much lower, we can still see the net outflow at CH stations and net inflow at some of the CBD and SLU stations consistently.

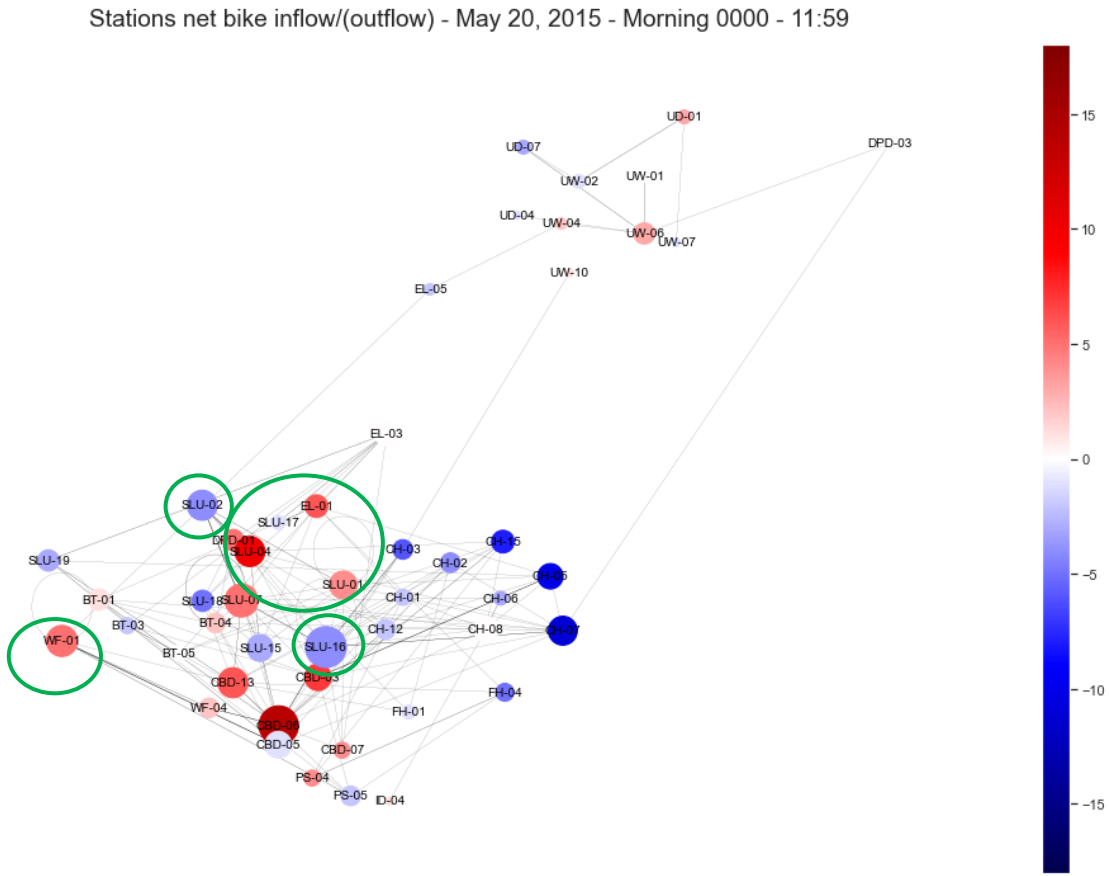


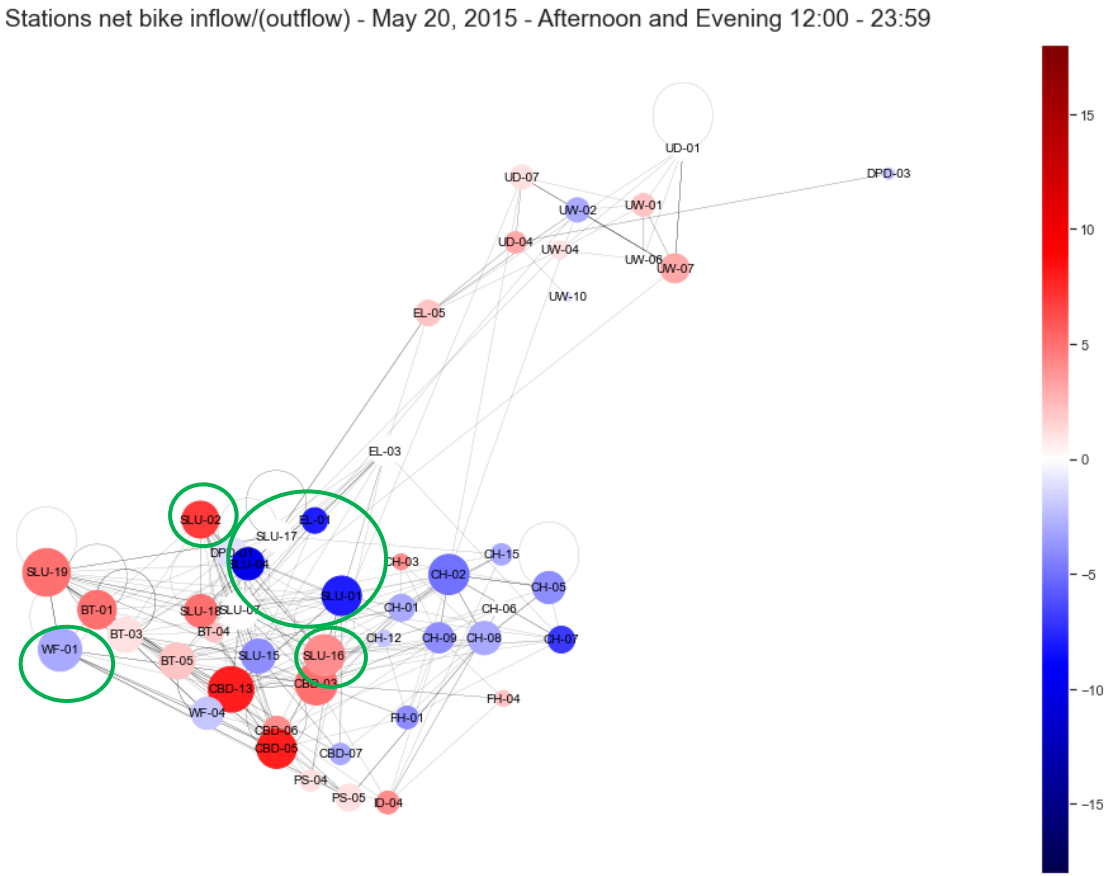


Stations net bike inflow / (outflow) – A specific *day*

Comparing morning (00:00 – 11:59) and evening (12:00 – 23:59) period of a specific weekday:

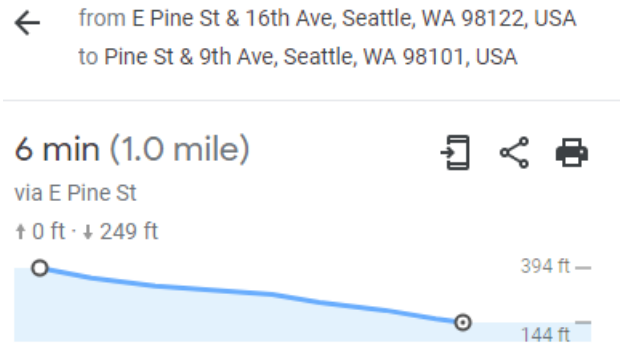
Some stations, as circled in green, have opposite flow in morning and afternoon hours. However, the CH stations remained having a net outflow throughout the day. So the flow isn't necessarily in opposite direction in morning and evening hours (i.e. maybe people don't travel with bike for both trip to work and back home).





Observations from the highest net outflow station CH-07

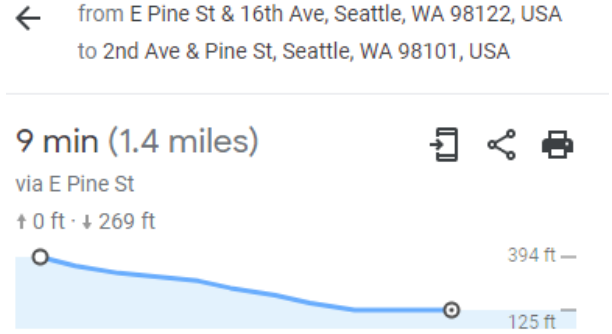
CH-07 to SLU-16



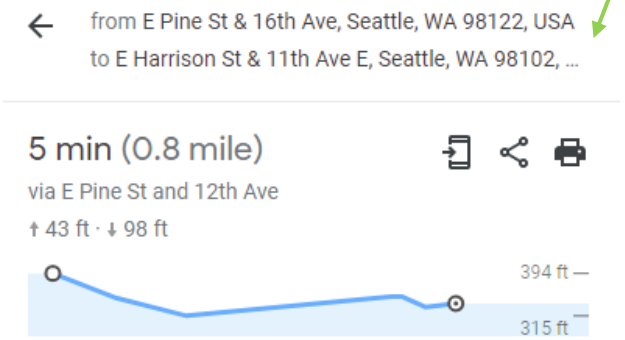
Top 10 destinations from CH-07 (full period)

from_station_id	to_station_id	nbr_of_trips
CH-07	SLU-16	1040
CH-07	CH-02	944
CH-07	SLU-01	749
CH-07	CBD-13	725
CH-07	CH-05	712
CH-07	CH-08	459
CH-07	CBD-03	446
CH-07	SLU-07	418
CH-07	PS-04	393
CH-07	CH-09	386

CH-07 to CBD-13



CH-07 to CH-02



Top 10 origins to CH-07 (full period)

from_station_id	to_station_id	nbr_of_trips
CH-05	CH-07	653
CH-02	CH-07	489
CH-07	CH-07	289
CH-15	CH-07	230
CH-08	CH-07	143
CH-09	CH-07	138
SLU-16	CH-07	131
ID-04	CH-07	130
FH-01	CH-07	84
CBD-13	CH-07	72

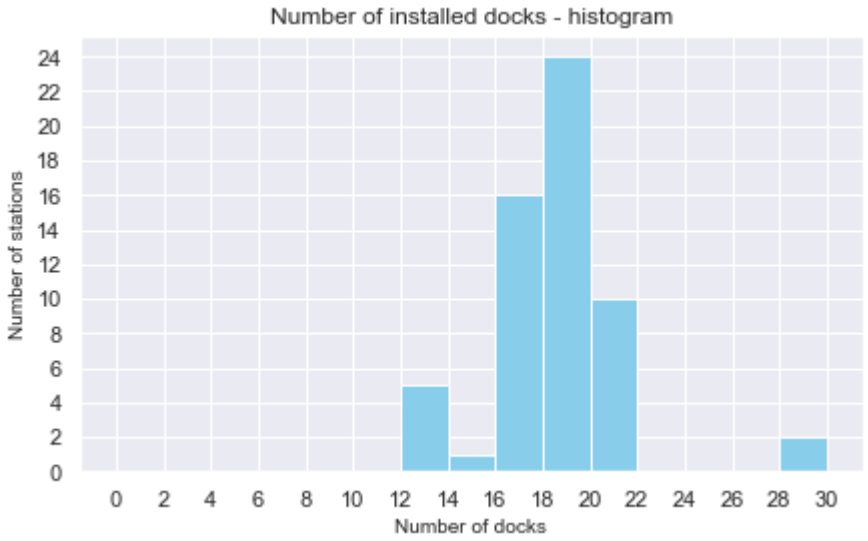
We suspect altitude of station locations could be a factor influencing the imbalance. Here we picked CH-07, the highest net outflow stations and checked from Google map the biking altitude changes between CH-07 and three top destinations. Noticed that, of the highest imbalance that occurred with SLU-16, the altitude change is 249 ft, whereas with CH-02, a less imbalanced combination and within the same district, the altitude change is only 98 ft.

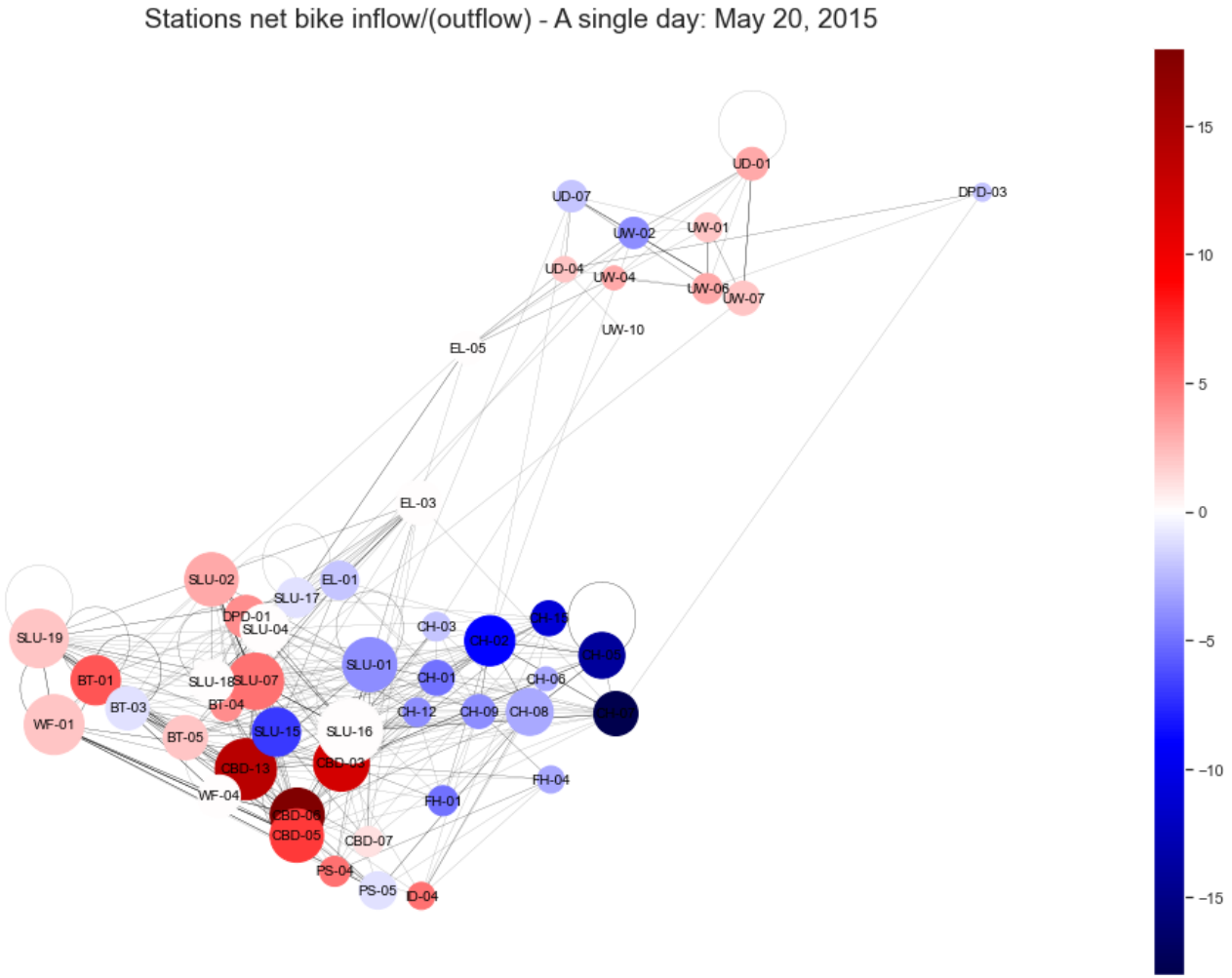
Hence, CH (Capitol Hill) is likely more hilly, and this confirmed again people do avoid riding uphill.

Stations net bike inflow / (outflow) – A specific day

The pattern reflected the need of very frequent rebalancing

As the number of docks in each station was in the range of 12 to 22 only, while the max net inflow or outflow reached more than 15 in a single day, the bike share company probably had to move the bikes in every 1 or 2 days, or else some bike stations in the Capitol Hill area would be out of bikes and some stations in the CBD area would be completely full.





Summary of Observations

- Over time, Pronto member utilised this system more evenly but overall ridership fell.
- Weather affects ridership just as we expected.
  - Members appear to be more resilient to adverse weather than short-term pass holders, likely because members had already paid the annual membership fee.
  - Younger group of 16-25 age may be slightly more immune to increasing wind speed and precipitation than older age groups.
- Short-term pass holders focused on the few stations that are near waterfront and park throughout the whole period, as they did leisure rides around those areas.
- In general, members and short-term pass holders ride differently in terms of duration, pickup hours and locations. The difference mainly reflected the different purpose, i.e. regular commute or leisure.
  - This prompted the question – is the pricing competitive? It seems there are only long term members who use the system for commute, or leisure riders in the form of short-term pass holders. $8 appears not appealing to occasional riders who do not prefer committing to an annual membership but still want very short rides of 4-12 minutes for commute purpose. If lower price can be offered for shorter duration rides, the system may capture these riders.

Summary of Observations (cont.)

- The consistent net outflow in Capitol Hill likely confirmed common understanding that riders would avoid riding uphill.
- The imbalance in inflow and outflow in some stations, as relative to the docks available, imply that the bike company need to rebalance the bikes as frequent as once every one or two days.

Questions that cannot be answered by looking at this dataset alone

To comment on the effectiveness of a bike sharing system, or to understand why some systems succeeded and some failed, there are more questions to ask though -

1. Did the bike sharing network have enough coverage of where the people would travel to? If the bikes cannot access to points people actually go, there is no point to get on a bike but to use other transportation instead. Since we are studying the existing nodes and edges only, we cannot conclude about the potentials beyond them.

2. Were the bike stations placed at a convenient place (e.g. just near the railway station entrance or next to the bus stop)?

3. Is the overall environment suitable for biking – we found out weather could be a negative factor for Pronto in certain months and riders avoid riding uphill. How about the availability and quality of bike lanes?

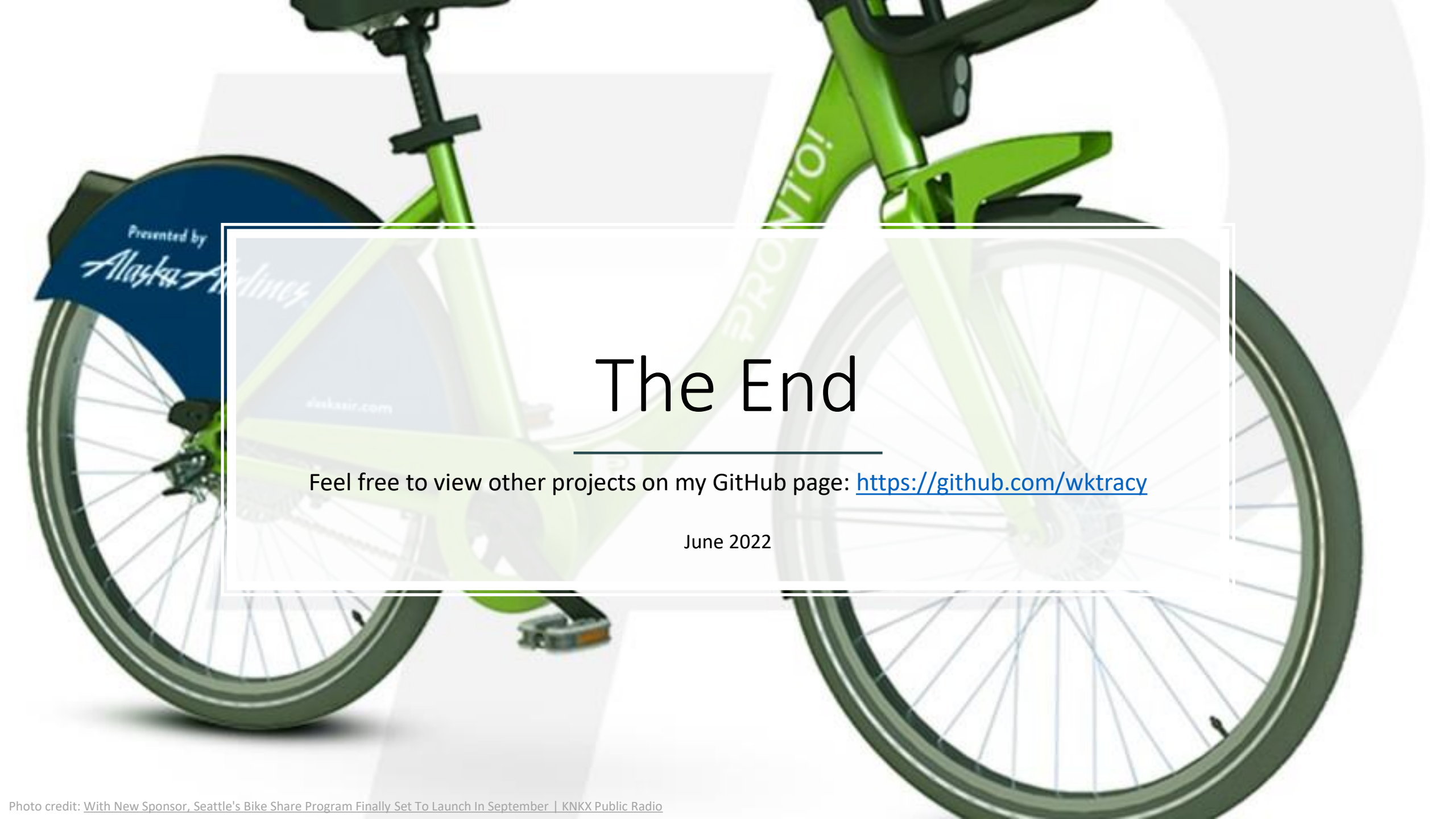
4. Is the pricing competitive as compared to other transportation mode?

5. Just now we discussed about the imbalances in bike inflow and outflow at some stations. Did the bike company rebalance the bikes frequent enough?

Questions that cannot be answered by looking at this dataset alone (cont.)

Hence, in analysing the success potential or effectiveness of a bike share system, we may include various perspectives, for instance –

- The existing usage pattern and demographics of users (which we have demonstrated).
- The impact of natural factors, i.e. weather and slope (we confirmed that they do matter).
- The population density of the area.
- The physical infrastructures in the city on whether they enable a bike-friendly environment.
- The quality of bikes and most importantly, their availability when people need them (rebalancing).
- The coordination or interaction with other public transportation in the area.
- Pricing, as relative to competitive transportation cost.



The End

Feel free to view other projects on my GitHub page: https://github.com/wktracy

June 2022

Photo credit: With New Sponsor, Seattle's Bike Share Program Finally Set To Launch In September | KNKX Public Radio