Exploring bike sharing scheme journeys and its customer base over an average day in New York

Vaida Gulbinskaite
Visual Analytics - City University of London

ABSTRACT: In this data science piece I will try to explore New York City through visual analytics of bike journeys, taken during the first Wednesday of September 2016. I will also look at consumer attributes and cycling behavior. In order to produce less visually cluttered maps, I will use aggregated movement graphs in order to visualize the data. It was observed that there were very few 24-hour pass holders or 7-day pass holders compared to annual subscribers, who used the scheme. There were also significantly more men who used the scheme compared to women. Findings from aggregated movement graphs segmented by time intervals showed that cycling mobility corresponds with the general pattern of regular human mobility. By investigating maps segmented by consumer characteristics, a few differences were observed.

1. Motivation and research question

Origin destination data consists of a few key features; The start point of the journey, the end point of the journey and time each event occurred. It has 3 dimensions: 1. Spatial, 2. Temporal 3. Attributive. Some popular data sources could be transport pick up and drop off locations, social media activity, people's movement from and to an event or a migration pattern between countries.

With the continuous availability of origin destination data of regular human mobility, researchers around the world have started studying these patterns. For example, by looking at twitter data in Greater London, Landesberger et al 2016 have identified that mass mobility behaviour corresponded with general patterns of human mobility. Beecham and Wood 2014 looked at group cycling behaviours of London Cycle Hire Scheme. Researchers identified that people are more likely to cycle in groups at weekends, late evenings and lunchtimes, within the pleasant areas of the city, with people that know each other. Howard and Burns 2014 looked at cycling commuter behaviour in the Phoenix Metropolitan area and identified that cyclists adjust their routes to use the available street bicycle facilities, and suggest that policy makers should concentrate on linking bicycle facilities across jurisdictions.

The New York City Bike Share Scheme has released their first publically available dataset in 2015. Which was about a year after the scheme was rolled out. Since the scheme is around for a few years now and the initial hype could be over I would like to look at the current state of the NYC bike travel and in this visual analytics project, I will investigate a typical day in the New York City in order to see where New Yorkers travel to work, head home and travel for leisure. This will be done through investigating 'citibike' - New York's bike sharing scheme and comparing results between types of consumers (Subscriber vs Customer) and gender of the cyclist.

2. Data source and treatment

2.a. Data source

The data was taken from the New York's bike sharing scheme called 'citibike' (NYC Bike Share Scheme. 2014). I have selected the most recent data available, which contains the bike trip origin and destination (as well as age, gender and type of bike scheme) throughout

September 2016. I have then looked further into the first Wednesday of the month (7th of September). This date was chosen as Wednesday falls within the middle on the week and therefore should represent a 'regular working day'.

2.b Software used

In order to clean and structure the data, python's math, numpy, scipy and pandas libraries were used. For the initial investigations and visual outlier identification, data was loaded into Tableau.

2c. Data treatment

2.c.1. Data Restructuring

The data only had a bike id, which (as I found out by filtering data in python) could represent multiple journeys, taken by multiple people, Therefore I have decided to index the data by assigning a 1-6 digit number to each journey taken within the month.

Before the data could be uploaded to any software for analysis, it had to be re-structured to fit a long format data shape. In order to do so, a few steps needed to be taken. Firstly, the indexed data was split into two datasets; one for origin and another one for destination. Secondly, I have indexed each dataset by creating a type of point variable, with the following values: a. Origin for the origin dataset and b. Destination for the destination dataset. The two datasets were then merged together to fit origin and destination coordinates, dates, stations and their identifiers into single columns for each dimension/attribute. Lastly, data was then sorted by the newly created identifier and type of journey so each origin destination pair appear: 1. next to each other by id 2. All destination values appear after all origin values.

2.c.2. Outliers

There were a few types of outliers identified in the dataset. Spatial outliers – By looking at the data mapped in Tableau, I have identified that there were some unrealistic bike journeys taken (as some of them ended in the Atlantic Ocean or New Jersey). In order to remove these, a new distance field was created in python by using latitude and longitude of origin and destination. I have noticed that there were journeys exceeding 8,000 kilometers. By creating box plots I have identified some obvious outliers and therefore removed any journeys that have exceeded 10 kilometers, as these could be either done for sport (e.g. training for a cycling event) or in collaboration with public transport. I have also removed any journeys that had the same origin and destination point or any travels that did not exceed 500metres.

Attributive outliers – one of the attributes, recorded in the data was self-reported age of the person, using the bike. I have identified that there was a small amount of people whose age exceeded 100 significantly (according to the data, the oldest participant in the bike scheme was 131 years old). In order to eliminate any participants that have potentially falsely reported their age, I have decided to only look at people between 16 - 70 years old.

3. Tasks and approach

3.1. Investigate people participating in biking scheme.

In this section I wish to look into the type of customer that uses the biking scheme. Through a series of graphs, I will look into descriptive statistics of the consumer of the cycling scheme.

3.2. Investigating the difference in travel behavior

3.2.1. Differences between time segments

This segment will look at differences between cycling journeys between time segments. In order to investigate the data further and reduce the processing load for the computer, I have segmented the time intervals, in accordance to New York's public transport fares NYC Bike Share Scheme. 2014 and similar to Landesberger et al 2016 Greater London temporal clusters. I will investigate the following time intervals: 1. Morning off-peak hours (00:00 – 06:00 am); 2. Morning peak hours (6:00am – 10:00 am); 3. Mid-day off-peak hours (10:00am - 4:00pm); 4. Evening peak (4:00pm – 8:00pm); 5. Evening off- peak hours (8:00pm – 23:59pm). In this section, only trips that started and ended within set time interval were selected.

3.2.2. Differences between gender and user type

In this segment I aim to investigate the difference in bike journeys, taken between user types: 1. Subscriber - person paying annual membership 2. Customer - Person that has bought a 24-hour or 7-day pass. I will also look into the differences in travel behavior between genders.

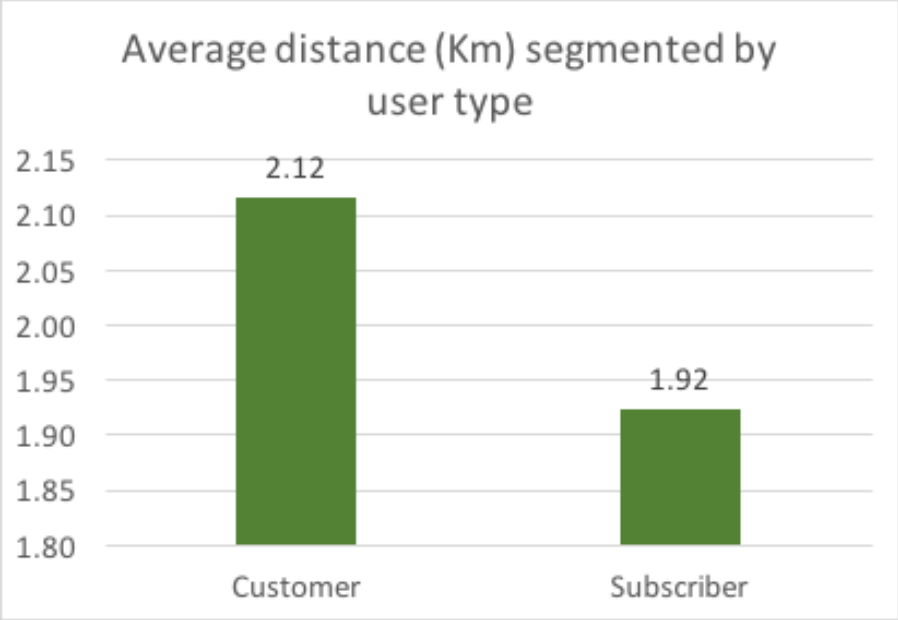
3.2.3 Computation technique

Data was processed using Visual analytics software, called V-Analytics. In order to present the data, the New York city Area was divided into Voronoi Polygons, which then were used in order to aggregate the moves between them. The moves were aggregated based on number of visits to each polygon, number of different visitors to each polygon, number of start and end trips, number of moves, their speeds, durations and lengths. Moves were also temporally aggregated into one hour intervals. Any insignificant moves or moves within the polygon were removed in order to reduce noise and clutter.

4. Analytical steps

4.1 User type characteristics

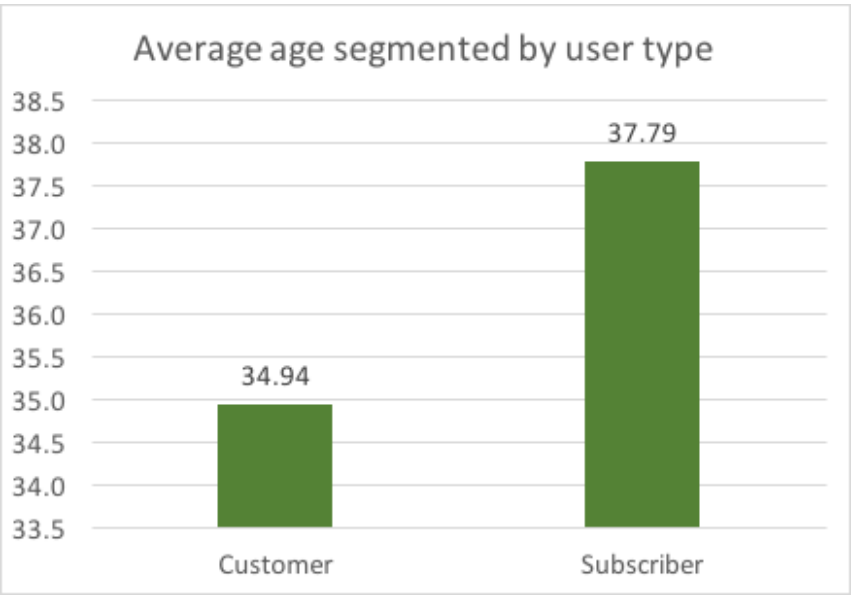
There were only 378 trips taken by customers and 54,106 trips taken by subscribers. On average, customer's journeys were only 200 meters longer than subscriber distance (Graph 1). The average customer age was 2 years and 10 months lower than subscriber age. In terms of journeys taken throughout the day, the number of journeys peaked at around 8am and 6 pm for subscribers (corresponds with peak times identified using NYC public transport information)



Graph 1: *Average trip distance segmented by user type*

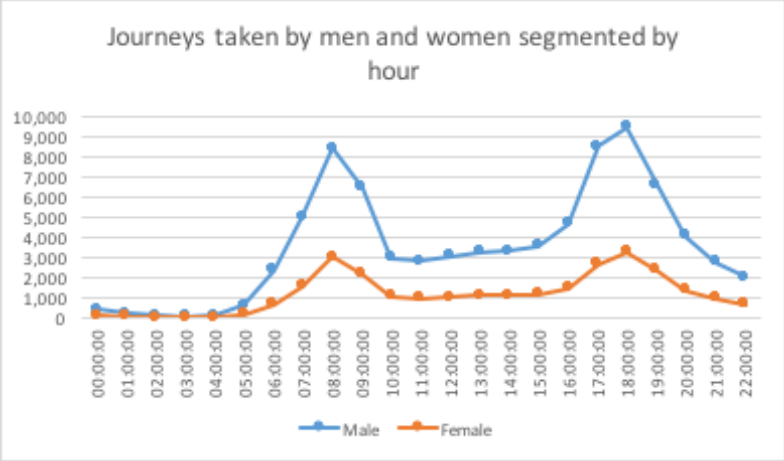
4.2. Gender Characteristics

There were 40,846 men and 13,456 women who participated in the bike sharing scheme. Graph 4 shows that there is no difference between the peaked number of journeys between men and women, similar to the Subscribers, trips for both genders peaked at around 8 am and 6pm. On average, women's journeys were 202 meters longer than men's journeys. Women who used the bike sharing scheme were on average 1 year and 6 months younger than men.

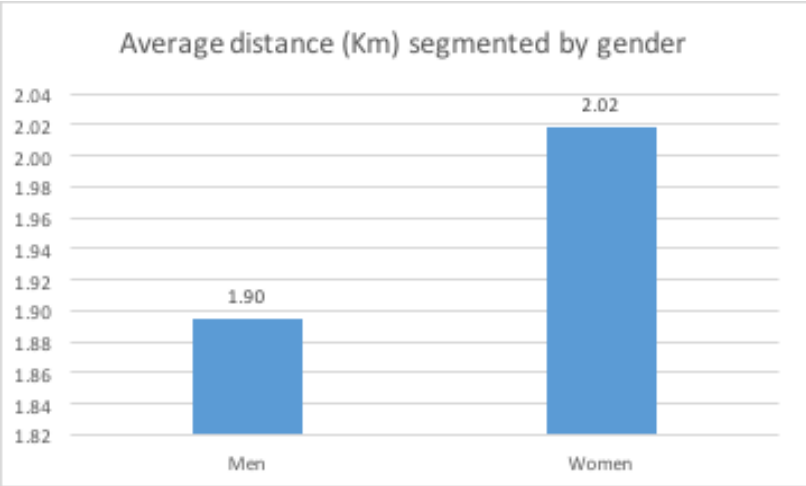


Graph 2: *Average age segmented by user type*

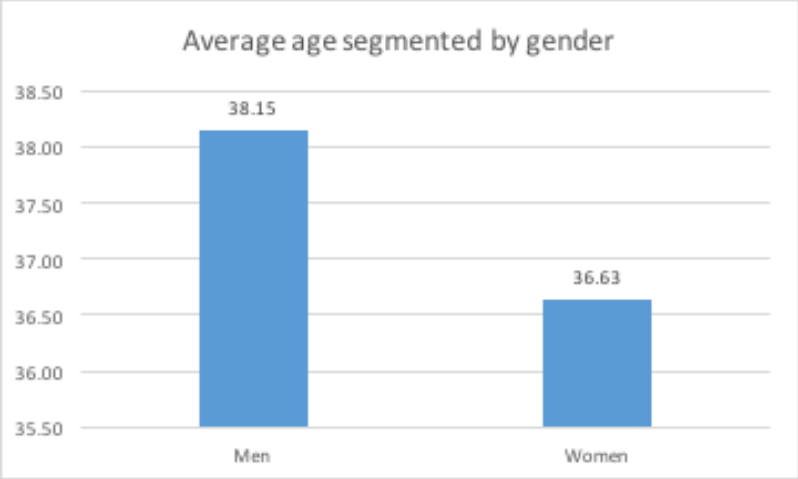
Graph 3: *Hourly segmentation of journeys taken by subscribers*



Graph 4: *Hourly segmentation of journeys taken by men and women*



Graph 5: *Average trip distance segmented by gender*



Graph 5: *Average age segmented by gender*

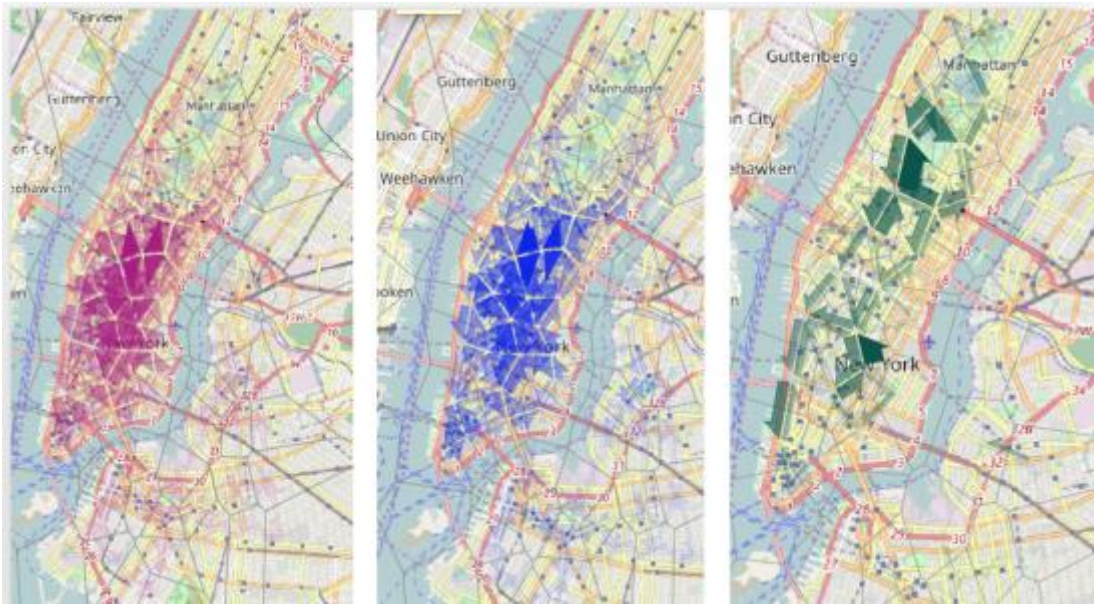
4.3 Differences in trips taken

4.3.1. Map legend

The width and brightness of the arrow represent the density of movements between Voronoi Polygons (the brighter and wider the line, the more trips occurred within the defined regions)

4.3.2 Segmented by user type

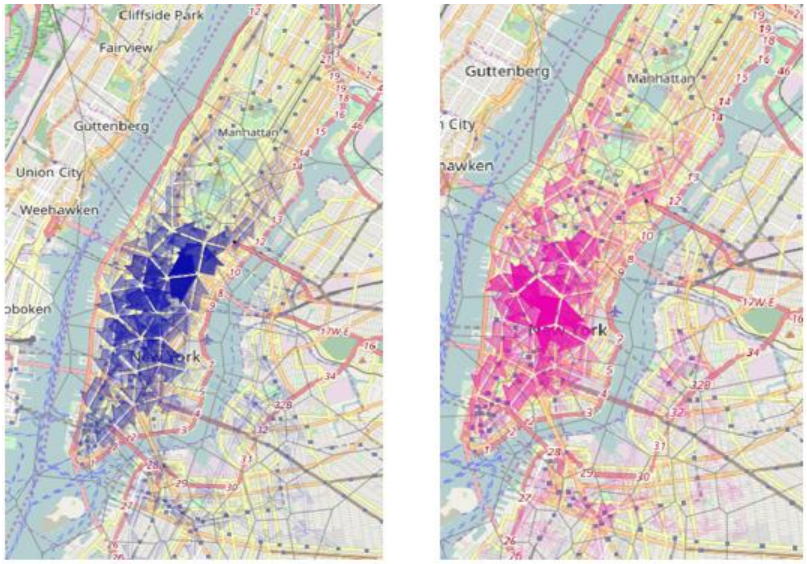
In the first map (Map 1), I have segmented the bike journeys taken by the type of user. Map 1 shows that there is no visible difference between all bike share scheme participants and Subscribers to the scheme. This could suggest that subscribers use cycling for more day-to-day operations. (e.g. heading to/from work, traveling for food or shopping). However, when looking at the Customers (24-hr or 7-day pass holders), it is clear that journeys are mostly concentrated around the Central Park area as well as around the epicenter of Manhattan. This could be due to the fact that there is a significant amount of tourists using the Customer scheme and therefore they visit tourist attractions (e.g. take a ride in Central Park or visit shopping districts or other famous landmarks). On the other side, this effect might be visible only due to the fact that there were significantly less trips made by Customers compared to subscribers.



Map 1: *Aggregated movements segmented by user type (Left to right) 1. All movements (Purple); 2. Subscriber movements (Blue) 3. Consumer movements (Teal).*

4.3.3. Segmentation by gender

In the second map (Map 2), this shows the segmented gender bike travel behavior. It appears that although both men and women travel to Manhattan, men's bike journeys seem to be more spread out whereas women's are very dense around Midtown, Chelsea and East Village. It also shows that there is more significant movement around the Central Park area for women (outgoing and incoming movements from living the Upper East Side and Upper West Side), whereas this is not as visible for men, as they travel within the Midtown and the Civic Centre. This could indicate that women cycle around New York for leisure and men chose more routine travels.



Map 2: *Aggregated movements segmented by gender (Left to right) 1. Male (Blue); 2. Female (Pink)*

4.3.4. Segmentation by time interval

The third map (Map 3) shows aggregated movements segmented by time interval. The first map of the series (1st map to the left) represents the morning off-peak time interval. Although there was not a lot of movement during these hours, movement from the upper areas of Manhattan to the lower areas of Manhattan could be observed. The movement is especially visible at around the East 3rd Street and the Grace Church memorial house. The Morning peak interval is represented in the second map from the left. It shows some significant incoming flows from upper Manhattan and Brooklyn. This could signify that during these hours people usually travel to work. There is a significant density art around the Grand Central and Penn Street station, perhaps people travel there to get public transport for any further work journeys they might need to take. The third map from the left represents afternoon off-peak hours. It shows significant movement towards the Midtown and Chelsea areas of Manhattan, with an increased number of flows around Greely Square, which is within a close proximity to the station as well as hotels and shopping locations (e.g. The Manhattan Mall). This could suggest

that people travel there during their lunch hour at work (as there are places to eat around too) or for leisure if they are not at work. The 4th map from the left shows the evening peak time interval. Compared to the two off-peak maps, it shows an increase of travel to Brooklyn and towards the Upper Manhattan. This could suggest that this is when people head back home from work. There is also an increased amount of movement in Chelsea and Midtown. There is also a significant movement between the Maddison Square park and the Union Square park and between. Lastly, the 5th map from the left represents the Night off-peak interval. There is an increased amount of movements near the NYPD 10 PCT, Hywatt Square (Close to Union park), Thompkins Square and around New York Penn station, Siempre Verde Garden. Other than NYPD and Penn station and Hywatt Square, these locations could be popular for leisure or going out.

Map 3: *Aggregated movements segmented by time interval (Left to right) 1. Morning off-peak (Blue); 2. Morning peak (Red); 3. Afternoon off-peak(Teal); 4. Evening peak(Purple); 5 Night off-peak (Pink)*

5. Findings

In this project I have looked at the variations of bike journeys taken in New York City as well as some characteristics of people who take them. By looking at user characteristics I have identified that there are significant variations between the number of people who subscribe to bike sharing schemes versus people who buy 24-hour passes and 7-day passes. I have also identified that significantly more men cycle around the New York city compared to women. I have also looked at the distribution around the number of journeys taken, segmented by time. I have identified that both for men, women and subscribers, the number of journeys taken increases around 8am and 6pm, which suggests that New Yorkers use bike share scheme to cycle to work. I have also identified that the average distance New Yorkers and tourist travel is about 2 km. The average age varies between 37 – 38 years between men and women and 35 - 38 years between subscribers and customers.

By creating aggregated movement maps I have identified that there is a difference in cycling behavior between types of users: as subscribers use the scheme for day-to-day operations (e.g. heading to work) whereas customers travel to well-known New York attractions and leisure activities. These could also be attributed to tourists. By looking at the difference between men and women, I have identified that journeys are more spread out around Manhattan for men, however, women travel around Chelsea and East Village and Midtown more densely than any other areas of Manhattan. There is also a significant amount of movement around the central park, suggesting that women might cycle more for leisure than men.

Lastly I looked at bike journeys segmented by time interval. I have identified that during peak times, mobility from and to living areas increases, which suggests people cycle to work. On the other side, during off-peak hours' people potentially travel more for leisure, hence the increase of mobility around famous landmarks, shopping locations and eating locations. Findings correspond with the general knowledge human mobility behavior; findings also are similar to Landesberger et al 2016 London Twitter study.

6. Critical reflection

6.1. Consumer recommendations

By looking at consumer information, I have identified a few interesting conclusions: 1. Significantly more men use the scheme than women: 2. There are only very few pass holders (Consumers). Women could perceive cycling in the city as more dangerous, compared to men. In order to attract women, the NYC authorities could try to make cycling in the city safer by increasing the number of cycling roads. In order to attract tourists to use the scheme, New York city could increase marketing more in places such as airports, train stations and tourist information centers. Another reason why fewer tourists might choose to use cycling schemes is possibly the scheme is not as user-friendly for a new-comer. Perhaps the NYC authorities could look into making the system easier. In order to attract more people to sign up to the cycling scheme, the NYC could also run a campaign that would clearly outline benefits of cycling and provide facts about how much CO2 is being reduced by people who have already signed up and how much more could it be reduced if people kept signing up for the scheme at the same (or increased) rate (machine learning techniques could be used to predict this). The average age of people signing up for the scheme were in their 30s. The NYC could possibly attract more students by creating higher student discounts and more bike stations around popular student living locations and University/College/High School.

6.2. Implementation in urban planning

Although I looked at cycling behavior of one day in New York City, I have already identified some trends to where people go, when they travel and what customers might expect to see when they travel to a certain destination. I believe these methods could aide in urban planning and development. By looking at where people travel at peak times over a longer period of time, it could be more prominently determined where they work and where they live. Perhaps there could be more locations to eat around the areas where people work, so during lunch hours New Yorkers would not need to travel as far to get a desired snack. On the other side, by looking at up-coming living areas, new community centers, playgrounds, activity centres (gyms) or even schools could be created in order to not only bring communities closer together but also help new parents with the burden of taking their kids to school potentially on the other side of the district.

6.3. Further research

Due to the processing power of my machine, I could not look at a larger sample of data, however I believe it would be interesting to see if cycling behavior changes within a month or even between seasons (e.g. perhaps people use the biking scheme significantly more in summer compared to winter). It would also be interesting to look at cycling behavior in comparison to weather on a smaller scale, e.g. combining weather data (rainfall, temperature, percentage of sunshine visible) with cycling data in order to see if people chose to cycle more when it's sunny than when the sun is covered with dark clouds. This could also be a marketing opportunity for the NYC, as they could attract more people to sign up for the scheme when they feel like they

want to cycle to their destination rather than take the famous yellow Taxi or call an Uber or Lyft. Cycling data could also be compared with Taxi/Uber pick up and drop off location data. It would be interesting to see if there are times when bike journeys of the same length could take significantly less time than a taxi ride. It could also be interesting to create a similar study to Beecham and Wood 2014 in order to see if group mobility in NYC corresponds with group mobility in London. Lastly, different visualization techniques could be applied in order to represent the data, for example Landesberger et al 2016 Mobilitygraphs, which aim to spatially and temporally aggregated origin destination data in order to present easily viewed graphs.

7. ACKNOWLEDGMENT

I would like to thank Gennady Andrienko for helping me to create a visual analytics app for V-Analytics in order to conduct the analysis as well as giving me a few pointers on the size of the data.

8. References

Von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G. & Kerren, A. (2016). *MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via SpatioTemporal Graphs and Clustering.* IEEE Transactions on Visualization and Computer Graphics, 22(1), pp. 11-20. doi: 10.1109/TVCG.2015.2468111

Beecham, R. & Wood, J. (2014). *Characterising group-cycling journeys using interactive graphics.* Transportation Research Part C: Emerging Technologies, 47(2), pp. 194-206. doi: 10.1016/j.trc.2014.03.007

Howard C, & Burns E. (2014) *Cycling to work in Phoenix: route choice, travel behavior, and commuter characteristics.* Trans Res Record. 2001;1773(1):39–46. doi: 10.3141/1773-05.

NYC Bike Share Scheme. 2014. *CitiBike*. Available at: https://www.citibikenyc.com/system-data. [Accessed 1 January 2017].