Chicago Divvy Usage Summer of 2017

Nathan Lang
Marquette University
Milwaukee, Wisconsin
nathan.lang@marquette.edu
https://github.com/Nathanlang14/DivvyBikeSharingAndEthics

ABSTRACT

UPDATED—October 18, 2018. The aim of this paper goes into detail about the trends of Divvy bicycle riders in the city of Chicago during the summer of 2017. The dataset was courtesy of Kaggle from Divvy reports as well as public Chicago weather data. I then analyzed the trends with respect to which days they rode, which stations were most common, as well as how long they rode for. After finding who the average user is, I used a Logistic Regression model to predict whether the duration of a bicycle ride was over 10 minutes given parameters: day of the week, station, time, temperature, and weather. The analysis was then reflected upon using deon ethics checklist to understand the social and ethical implications of the analysis and future work.

KEYWORDS

Bike Sharing, Sharing Economy, Chicago Transportation, Divvy, Deon Ethics

ACM Reference Format:
Nathan Lang. 2018. Chicago Divvy Usage Summer of 2017. In *Proceedings of Social and Ethical Implications of Data (Marquette '18),* Nathan Lang (Ed.). ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

This dataset is a curated dataset from Kaggle, meaning it was preprocessed and cleaned. Although the original dataset spans from 2014 to 2017 everyday included, I chose to focus on the most recent summer (2017) as a subset. All the rides tracked in this dataset were done by a âĂIJsubscriberâĂİ which is defined by anyone who purchased an annual membership. Riding a bike in Chicago is most consistent during the summer due to weather constraints. So, to get a more diverse understanding of the effects of different parameters, I chose this subset to focus on specifically. There were roughly 1.2 million different bike rides taken from June to August and the dataset includes time and date taken, trip duration, temperature, events, station to and from, and number of bikes each station holds. Events can be described as cloudy, rain or snow, thunderstorms, clear, and not clear. 93% of the days were cloudy with 4% clear, 2% rain or snow, and 1% not clear. Since the data comes from Divvy themselves and released publicly, we can assume the data is reliable. Because the data consists of annual members, it is skewed towards people who use it as work transportation and people who live in the city as only people who use it often enough will purchase an annual membership.

Marquette '18, October 2018, Milwaukee, Wisconsin USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 INITIAL OBSERVATIONS

As mentioned before, our dataset consists of Divvy subscriber bicycle usage in the summer months (June, July, August) of 2017. Bike sharing has become a widely spread thing across America in metropolitan areas due to the increasing amount of people and traffic. Bike sharing provides a faster than walking and cheaper than ride sharing opportunity for many to get to work/around the city or to ride casually. Divvy is Chicagoś main bike sharing company. I first wanted to understand the locations of Divvy stops and how the bike company placed their stations. Following, the understanding of which stations were hot spots during the summer as well as which hours. This allowed me to depict how subscribers are using their bikes and where.

Bike sharing has also become popular as an alternative to having to drive in traffic during rush hours, both morning and evening. Although there seems to be a limit in how long people tend to ride. It seems there are two rides, to work or school and casual bike ride to traverse the city. Because both are annual members, we can assume many who are using to go to work or school are also using other kinds of transportation besides a car. On the other hand, the longer rides are more for people who live in the city and want to explore it more, exercise, or even run errands.

Before modeling, some initial observations and visualizations are needed to better understand the dataset.

2.1 Location of Stations

Divvy tends to strategically put their bike stations near other means of transportation such as bus stops and train stations. They understand that their target audience are people who are not using a car to get around, thus living in the city or traveling into the city using a train or bus. In Figure 1, a map of the public transportation in Chicago can be seen, while comparing to Figure 2 we can see a similar trend.It seems that Divvy wanted to put stops surrounding public transportation stops within a certain amount of distance. The further into the downtown area the map goes, the more dense the amount of stations is.

2.2 Frequency of Stations

When people use city transportation i.e. bus or train, the stops are usually distant from one another, bike sharing companies target

Figure 1: Chicago Public Transportation Route Map

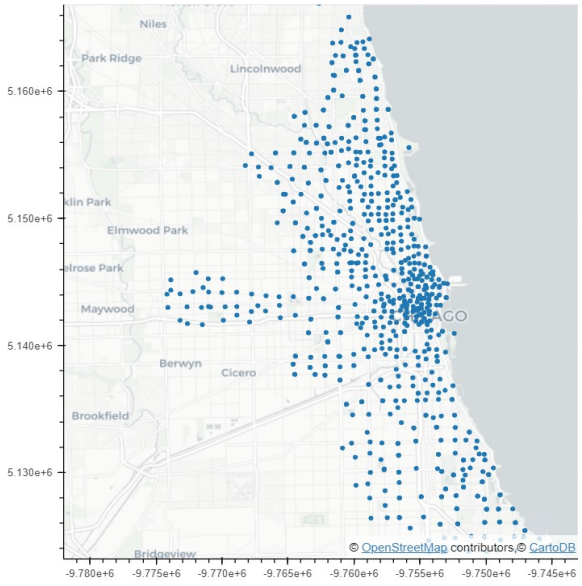

Figure 2: Divvy Stations Map

these sections to ensure the alternative to getting your destination quickly without having the expense of ride sharing. Often in metropolitan areas, a bike will be faster than a car during certain days and hours, like rush hours in the morning and evening. This is why many people tend to use public transportation to get into the city and then walk, or in this case bike, to get to their locations. In Table 1, the locations that were the most common starting point for Divvy bike rentals were also the the top three for the location of ending. Does this contextually make sense? Yes. These locations are the closest to Union Station in downtown Chicago where all the

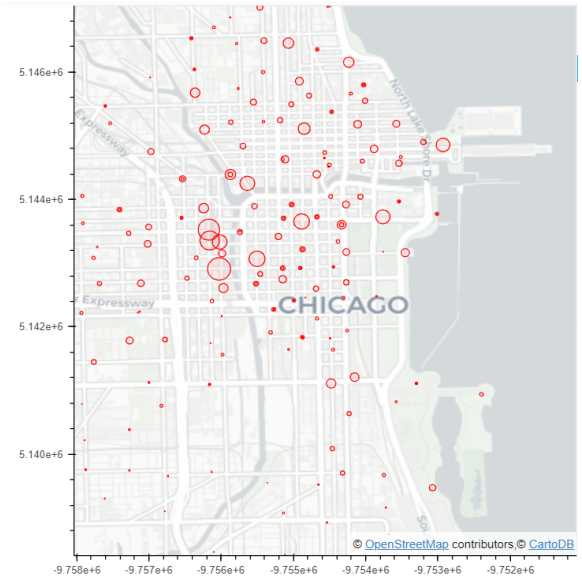

Figure 3: Frequency of Divvy Stations by Bike Ride Volume

trains come in along with a common bus stop. This suggests that the previous claim that people who use public transportation is the target audience for Divvy annual subscription. Figure 3 depicts an increasing radius for the amount of bike rides started at the location of the origin. This map shows more of where common locations are with respect to Chicagoś downtown landscape.

Table 1: Frequency of Starting Station

Location of Station	Frequency of Bike Rides
Canal St and Adams St	17,158
Clinton St and Washington Blvd	15,663
Clinton St and Madison St	15,127

2.3 Daily and Hourly Trends

So, in the previous subsections I was able to determine that many of the subscribers are people who I am claiming are working or going to school in the city during the week. On an average week will come in from a bus or train in the morning rent a bike and the reverse that on the way back by returning the bike and taking bus or train home. Now, will this claim uphold when analyzing daily and hourly trends? In fact, it does. During the week there is roughly 175,000 bike rides done a day, while on the weekend there is roughly 40% less with 125,000 bike rides done. Similarly, the peak hour is 8am in the morning with roughly 100,000 rides, and then has another large peak at 5pm having the global maximum for an hour with over 140,000 rides with 4pm and 6pm also above 100,000 rides in this range of dates from June to August of 2017 which is shown in Figure 4. While people are at work or school, the bike rental average for subscriber usage is around 60,000 rides which is much less than during peak hours.

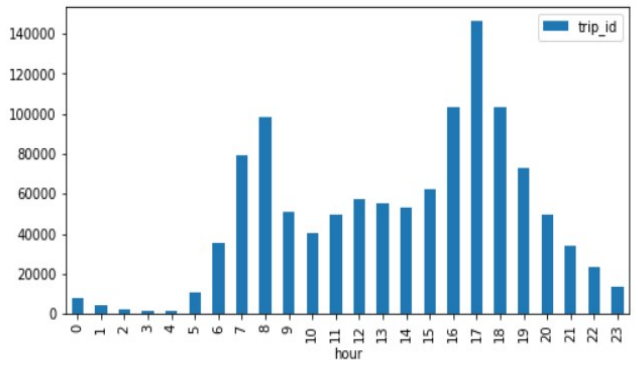

Figure 4: Frequency of Bike Rides per Hour

2.4 Initial Conclusion

It is apparent that weekday vs weekend, hour of day, and location all affect the usage of bikes for Divvy annual subscribers. This consumer base seems to be people who use public transportation to get into the city and then use Divvy bikes to travel to their specific location. Likewise, they will then use a Divvy back to the public transportation stations, namely Union Station specifically, to take that means back to their hometown.

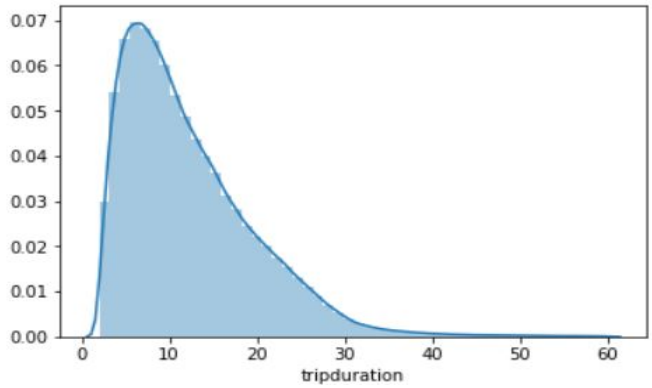

Figure 5: Distribution of Trip Duration in Minutes

3 MODELING

Now, that we know what affect Divvy bike usage, the question is whether we can predict if a bike ride duration will be under or above 12.25 minutes, the mean of ride duration during the summer of 2018. 12.25 was also chosen because if the average biker goes at a pace of 9.6 mph and we believe a 2 mile radius is the average a person will travel on bike to get to work, then the time it takes is roughly the mean we are using. Trip duration count vs trip duration is a right skewed graph which is seen as a distribution plot in Figure 5.

So, I chose a Logistic Regression Model and a Random Forest Classifier to predict this question. The parameters the model took in are date and time (month, week, day, hour), temperature, events, and which station it started at.

3.1 Data Preprocessing

Although there are many categorical variables, the use of dummy variables allows for manipulation. For instance, there are roughly 600 stations that are all names. So, to make this usable in the models, I took the stations that had less than 9,000 bike rental starts and then grouped them into other. Leftover were 13 significant stations and then an other bin. The stations were then made into dummy variables leaving one out to eliminate the possibility for multicollinearity.

Events were previously classified as cloudy, clear, not clear, thunderstorms, and rain or snow. In our context, if it is not clear then bike riding will be effected irregardless of what the exact kind of clear it is. Therefore, events was transformed into a binary 'clear' and 'not clear'.

The last transformation was on duration which was turned from a minute value to a binary with 1 being greater than the mean and 0 being less than or equal to mean.

The other variables used were month, week, day, hour, and temperature.

3.2 Modeling and Analysis

The first model I used was a Logistic Regression model with a test size = 0.3. This model produce an accuracy score of 60.2%. From the confusion matrix in Figure 6, we can see the model had a very small count of false negatives. On the other hand, false positives was very large. Why could this be? Well, one of the logistic regression assumptions is that the y values, duration in this space, come from a normal distribution. As stated previously, the trip duration is a right skewed distribution, reference Figure 5. I also used a Random Tree Classifier to model this hypothesis, and it had similar results with an accuracy of 60%

		Predicted	
		No	Yes
Actual	No	201974	3050
	Yes	134879	7427

Figure 6: Confusion Matrix for Logistic Regression

4 DEON CHECKLIST

The Deon Checklist provides an opportunity for data scientists to do a ethics check on each step of a machine learning, artificial intelligence, and other likewise projects. I will now analyze my process through this lens.

4.1 Data Collection

The data comes from the company which the project focuses on, Divvy. Divvy is Chicagoś bike share company. The data they release does not include personally identifiable information, although they do collect more personal data for the companies database. There are some biases that could exist in this data. For instance, the data at focus is only from people who are annual subscribers to the service

during the time frame of June to August of 2017. Therefore, some of Divvyś customer base is left out i.e. people who use it only once when they are visiting and exploring Chicago. Unlike other sharing economy companies, Divvy does provide locations where people can use cash to get a membership. Another bias is the location of Divvy bikes, they are located more in wealthy areas in downtown Chicago opposed to ethnic communities near the city, thus limiting their diversity of customers. When Divvy releases their data they minimize the personal data to only gender and birth year.

4.2 Data Storage

This section is more difficult to answer as much of this is up to Divvy on how they store their data. Although on their website they do provide ways to remove your personal data from their databases.

4.3 Analysis

As mentioned in the data collection one of the major biases in this data is that it is from one summer with only subscribers information. The variables used can be imbalanced. For instance, in the heart of downtown Chicago there are many Divvy bike stations within a few blocks of each other. Some may receive a data point for being the starting or ending location for a ride, but this could be solely based on a persons personal decision. For instance somebody returning it two blocks from the normal one they go to because they wanted to grab coffee that day. When the data was then processed, that point might fall under my 'other' location for my models. Furthermore, the weather data is an aggregation of that day so, it could be misleading because a day with thunderstorms could be clear during the rush hours and thunderstorms during the day. This instance could be misclassified as a not clear day when in reality it functions as a clear day. Similarly, the temperature of the day in Chicago fluctuates so much and cannot always truly represent the temperature outside i.e. humidity and 'The Windy City'.

The visualization used were used to honestly represent the underlying data. None of the figures were very complex, but rather initial observational plots. No personal data was plotted at all. Lastly, the process of generating this analysis are well documented and reproducible if there are issues in the future.

4.4 Modeling

I chose to leave out gender and there were many more males than female users in this dataset, but I felt this would unfairly discriminate.As stated previously, many of the subscribers are students or working people thus the length of duration should not depend on gender.

The data provided does not include age, racial background, or socio-economic status so I was not able to test the fairness across groups. Although I think that would be an interesting expansion on this analysis in regards to future work and Divvyś ethical placements of stations.

Similarly, the models were only trained using basic time stamp data, weather, and location. There are endless metrics that could be used to predict usage, but not as many that would help predicting duration of bike ride.

Both models are simplistic in terms of explainability. The goal was to determine if the duration was a binary classifier of greater than the mean of bike duration, or less than or equal to. Therefore, we can explain why the model made said decisions.

I believe this paper gives a thorough explanation of the limitations the models have and so they can be understood.

4.5 Deployment

While this model will not be used by any company or deployed, it can still be harmed by concept drift and have unintended usage.

For instance, as the years go on Divvy will inevitably change/add locations to the city of Chicago. This will affect the 'other' category which refers to the less popular stations. For instance, right now some stations might have a lot of interest, but are constantly running out of bikes to rent. In this case, the location will not obtain any more data points even though it is a popular location and needs more bikes. To this point, Divvy may shift the number of bikes a location offers which in turn would change the popularity of a station. To limit concept drift, a report like this can be done every quarter or fiscal year.

As for unintended use, the conclusions made from these models cannot do any physical harm, although it can be misused to understand Divvy as a whole. Divvy has two bodies of consumers. The subscribers who are people who purchase annual subscription to use Divvyś bikes and one time users. This paper is a representation of Divvyś subscribers not as a whole. Furthermore, the analysis done on if a ride will be longer or shorter than 12.25 minutes, was done to get a better understanding of how people are using these bikes. Many times people who go longer than that are using it for other means than just getting to and from work.

Lastly, although the intended use is to understand the trends surrounding subscribers, it is important to understand the time frame. With it being summer, many people take half days to enjoy their summer and/or have days off for holidays, etc. Also, if they are students, summer usage would be different than when they are going to class. If they live in the city, they may now use it to just explore or run errands. Similarly, subscribers who live in the city may use it as the means of transportation in the city for summer occasioned things that happen in Chicago i.e. marathons, music festivals, air and water show, etc. All of these are common usage of Divvy bikes that will affect the data at hand. So, it is important to note that a subscribers habits may change during the summer months.

4.6 Checklist Improvements

As a whole the deon checklist provides an ethical lens to analyze you project process through. Unfortunately, right now it is guided more towards in-house company projects. I believe that it could be more general in light of using other peoples data. For instance, in this case I did not personally record the data or through a company I work for, but rather obtained a public data set online. I am able to understand only as much as the company wants about the data collection and even less about the data storage. Only my personal analysis is able to reverse engineer the ethics behind the data they obtain. Furthermore, data storage is very limited. It is difficult to have companies be outspoken about how they store their data. The checklist can focus more on the impacts of which data is being stored and which data should be deleted. Should is a

subjective term of course, but it is better than not touching upon the data storage. Lastly, not all projects are deployed, namely the projects in academic world and research based. Deployment could be more tuned to an overarching what can be the outcome of your model progressing over time, what is the intended use and how can you prevent unintended use, and what are future adjustments. Removing Roll back would be step one, in my opinion.

5 IMPLICATIONS

I was able to understand what kind of people are using the Divvy bikes in Chicago as an annual subscriber and speculate how the rides are being used. First major conclusion was that the average user for our dataset was a working man that uses city transportation (bus or train) to a central hub where they then use a Divvy bike to get to work around 8am start and do the reverse around 4-6pm with 5pm being the most common.

Some future work I would like to dive deeper into include understanding the amount of bikes provided at each location and supply vs demand for the locations, as well as increased ways to provide opportunities for lower-income people of color can use bike sharing programs.

I did some very basic analysis on the frequency of stops, but it would be interesting to exam further which stations are popular. Instead of focusing on location alone, I would focus on the surrounding environment. This would provide further insight into how people are using the bikes, are they stopping only at bus stops/train stations or near shops or food, office buildings, the beach or other landmarks in downtown Chicago, etc.

While still following the deon checklist, I would want to learn more about the lower-income individuals who are using each bike. By leaving out personal information, making sure to get consent, and keep bias out of it, this data could further the understanding of the usage of bikes. Some questions that could be answers are are they using it as means of transportation instead of buying a car? Are they traveling further to get to their jobs, or run errands? Do they use it for fun or only for means of transportation? When money is a worry, people tend to track every dollar thus these questions will provide insight into the lifestyle of lower-income individuals. While this information can be misused by outside factors, for social implications, it is important to understand how every subgroup uses the sharing ecosystem.

New York City has a similar bike sharing company to Divvy named Citi Bike, evidently. Researchers from New York University performed a study on Citi Bikes impact and trends in New York City. They found their average trips were between 1 to 1.5 miles and are more than 5 minutes faster and 10 dollars cheaper than a taxi[4]. It is apparent that people are not only using bike sharing to save money but also move throughout traffic intensive urban cities such as New York City, Chicago, Los Angeles, etc. It seems as if the usage of bike sharing is exponential. With this comes more users and more data. It is important to understand when to let go of data and put a data retention plan in place. Furthermore, technological advancement of these bikes is inevitable. For instance, tracking the bike and linking that with a person to understand individuals lifestyles. This crosses a line if analysis is done. On the deon ethics checklist, privacy in analysis would not be upheld. Likewise, it could be seen as lack of informed consent or unintended use if these companies claim to be tracking the bike to make sure it is not stolen.

Researchers at Portland State University studied the bike sharing system with regards to demographics in âĂIJEvaluating Efforts to Improve the Equity of Bike Share SystemsâĂİ [2]. In this paper, they conducted a survey in Philadelphia, PA; Chicago, IL; and Brooklyn, NY as they are Better Bike Share Partnership (BBSP) outreach areas. The survey was conducted to get an understand about how lower-income people of color. They found that only 2 percent of lower-income people were bike share members, while 10 percent of higher-income white residents were members.

Now for Chicago as whole, the racial wealth gap is far worse than the countryś average. This suggests that for the dataset in this paper almost all of the users are white. The Chicago Tribune stated that the median income of whites in Chicago is $70,960 compared with $41,188 for Latinos and $30,303 for blacks [1]. So, we can assume that Divvy users are white people who have the means to purchase yearly memberships because 48% of lower-income people of color said that they are not using bike sharing services because of cost [2]. Overall, it is apart lower-income people of color face more barriers to bike sharing. Even though many bike sharing companies sell their selves as inexpensive alternatives, there is still a large racial gap due to the cost of yearly service. The social implication of this is that Divvy and other bike sharing services will continue to put locations at spots higher-income white people will be. Therefore, there will be a continued lack of options in the lower-income neighborhoods. The gap between the usage will continue to grow. This is an ethical problem as well because this limits the range that somebody in lower-income communities can travel without some form of transportation. Of course, walking is an option, but I believe bike sharing is not an amenity, but rather a necessary alternative.

	Initial rollout	First expansion	Second expansion	*Total*[h]
Divvy stations	300	175	107	***582***
Service area (mi²)[a]	31.5	30.6	19.2	***74.4***
Station density (mi²)	9.52	5.72	5.57	***7.82***
Communities[b]	21	35	24	***47***
Population density (mi²)	20,761	18,495	13,298	***17,671***
Station distance (mi)[c]	0.24	0.41	0.46	***0.33***
Train stations[d]	84	68	39	***191***
Bus stops[e]	2,549	2,260	1,377	***6,186***
% Non-white[f]	42.2	62.5	76.8	***57.8***
% Unemployed[g]	5.1	6.8	9.8	***9.3***

Notes: (a) Combined area of non-overlapping ¼ mile buffers from Divvy stations; (b) Number of communities that either intersect or are completely within service area; (c) Average minimum distance to closest Divvy station by service area; (d) Chicago Transit Authority (CTA) L train and Metra commuter train stops; (e) CTA and PACE suburban bus stops; (f) ACS, 2011-2015 5-year estimates, nonwhite and non-Latino; (g) ACS, 2011-2015 5-year estimates, population 16 years of age and older in the labor force; (h) attributes reported under *total* service area reflects data from all three station cohorts with no overlap (i.e., the present characteristics of the system service area at the time of this writing).

Figure 7: DePaul University Table

The Chaddick Institute for Metropolitan Development at DePaul University also analyzed Divvy Bike Sharing in Chicago. As seen in Figure 7 the number of stations near non-white communities was expanding as the second expansion was done in July 2017- June 2017. The paper also says, over a third were located within lower-income communities [3]. This suggests that the dataset I analyzed might be biased now, but the changes are already in motion/completed. While this expansion did include Divvy expanding further outside

of just the city, it is important to note that they are acknowledging the social and ethical implications of leaving these communities out that I previously mentioned.

6 CONCLUSIONS

It is clear that bike sharing can be a reliable means of transportation in overpopulated and traffic prone cities. Many large cities already have some form of bike sharing company in their city and are continuously expanding. In Chicago, the company is Divvy. Given Divvy subscriber data, people who bought an annual subscription, combined with weather data in the summer months of 2017, I was able to get a better understanding of Divvy and their consumers.

During the week there was roughly a 40% increase in bike usage opposed to the weekends. Moreover, there was a bimodal distribution for the hourly trends with spikes at 8am and 5pm. These two statistics suggested that they were business people who are commuting to and from work. To test this theory, I looked at a map depicting the frequency of locations. From this map, I was able to notice that the most frequent stations were those near Union Station and other public transportation hubs.

After an initial understanding of where and how people are using the bike sharing service, I wanted to predict the if a ride duration was under or over 12.25 minutes, the mean of all ride durations in the summer of 2017. By using a logistic regression model and random tree classifier model with parameters month, week, day, hour, temperature, weather events, and starting station, I was able to predict ride duration of under or over 12.25 minutes at a 60% accuracy. There are two main limitations to these models. Firstly, the trip duration is not from a Normal distribution, but rather a right skewed distribution. Furthermore, because many variables are categorical, I had to create dummy variable. I chose to break the starting station into dummy variables with the top 13 and an 'other' bin. While there were roughly 600 stations, I understood the computation power of my personal computer can only handle so many variables thus I used the 'other' bin to limit the computing power needed.

Lastly, I took my project from start to finish and looked at it in an ethical lens. I began by using the deon ethics checklist to evaluate every step of the way and the ethics involved. The checklist helped me realize areas I can improve as well as Divvy to ensure the data in handled correctly. Then, I took these thoughts a step further by analyzing the social and ethical implications of my project and bike sharing as a whole. Lastly, I took my project from start to finish and looked at it in an ethical lens. I began by using the deon ethics checklist to evaluate every step of the way and the ethics involved. The checklist helped me realize areas I can improve as well as Divvy to ensure the data in handled correctly. Then, I took these thoughts a step further by analyzing the social and ethical implications of my project and bike sharing as a whole.

REFERENCES

[1] Gail MarksJarvis. 2017. Chicago's racial wealth gap far worse than U.S. average, report finds. *Chicago Tribune* (Jan. 2017). http://www.chicagotribune.com/business/ct-chicago-racial-wealth-divide-0131-20170130-story.html

[2] John MacArthur Nathan McNeil and Jennifer Dill. 2018. Breaking Barriers to Bike Share: Lessons on Bike Share Equity. *National Institute for Transportation and Communities* (Feb. 2018), 1–5. https://nitc.trec.pdx.edu/research/project/884

[3] C. Scott Smith and Riley O'Neil. 2018. Exploring the social, spatial and temporal performance of bikesharing in a period of growth and expansion. *Chaddick Institute for Metroploitan Development at DePaul University* (Feb. 2018). https://las.depaul.edu/centers-and-institutes/chaddick-institute-for-metropolitan-development/research-and-publications/Documents/ChaddickInstitute_DivvyReport_Feb2018.pdf

[4] Henry Chan Marc Postle Stanislav Sobolevsky, Ekaterina Levitskaya and Constantine Kontokosta. 2018. Impact of Bike Sharing In New York City. (2018). arXiv:arXiv:1808.06606