Modelling Bike Sharing Data

Hubert Luo

December 8, 2019

The bike sharing data is pulled from publicly available sources for Chicago, New York, and Washington DC in 2016.

Contents

1 Preliminary Data Analysis

1.1 Demographic Information

1. Below is the distribution of male to female riders for chicago.csv.

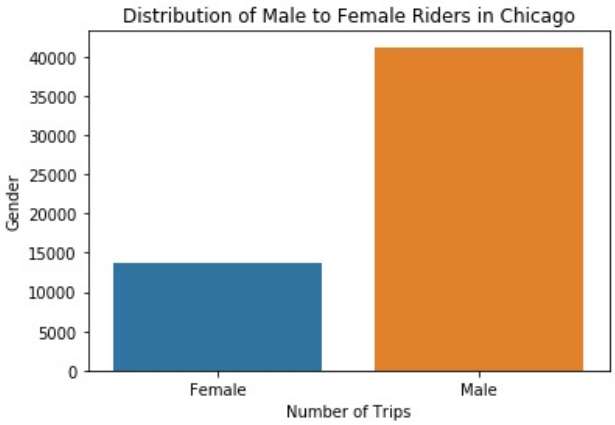


Figure 1: Distribution of Male to Female Riders in Chicago

2. Below is the distribution of the gender column for ny.csv.

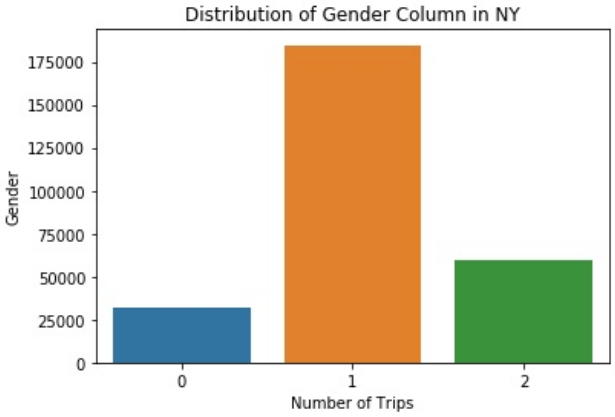
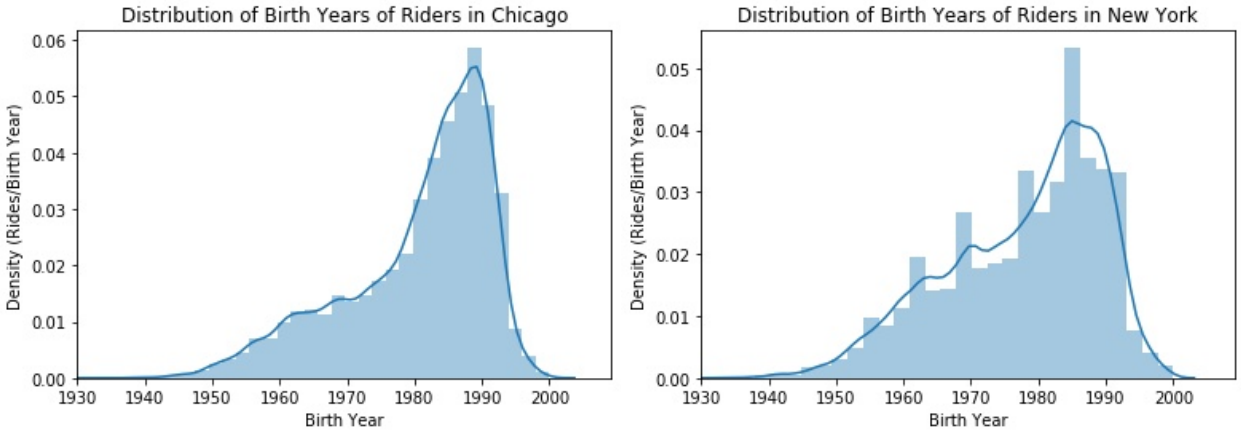


Figure 2: Distribution of Male to Female Riders in New York

3. Given the results in Chicago, the NY dataset likely maps 1 to male, 2 to female, and 0 to unspecified. This is because the relative proportion of male riders in Chicago is most similar to that of 1 in the NY dataset, and the relative proportion of female riders in Chicago is most similar to that of 2 in the NY dataset, thus leaving 0 for unspecified.

4. On the next page are the distributions of the birth years of bike renters in Chicago and NY.



(a) Birth Years of Riders in Chicago (b) Birth Years of Riders in New York

Figure 3: Distributions of Birth Years in Chicago and New York

5. Intuitively, the expectation would be for most of the people to use bikes to get to work so the distribution of birth years would be expected to be highest for the peak workforce ages, i.e., people in their late 20s and early 30s. As a result, we would expect peaks in the mid-to-late-1980s. This is reflected in the graphs above for both Chicago and New York, which have peaks around 1990.

   Data which could be removed are any birth years which are missing and any birth years less than 1930 - although it is certainly still plausible for riders to be in their 90s or older, their data for variables such as ride time may have a lot of outlier values for example.

1.2 Rental Times

1. Below are plots of the three distribution of trip duration in minutes across all three cities.

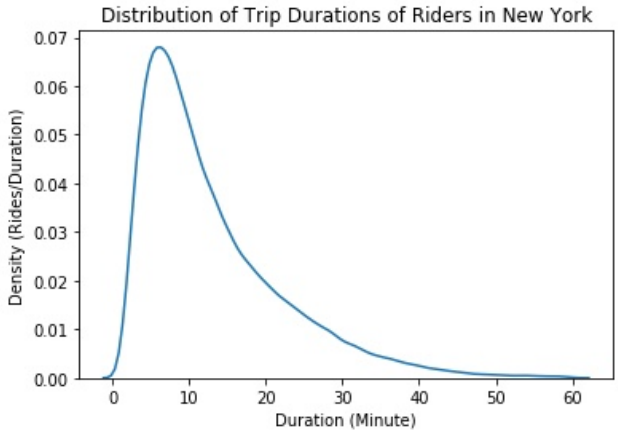
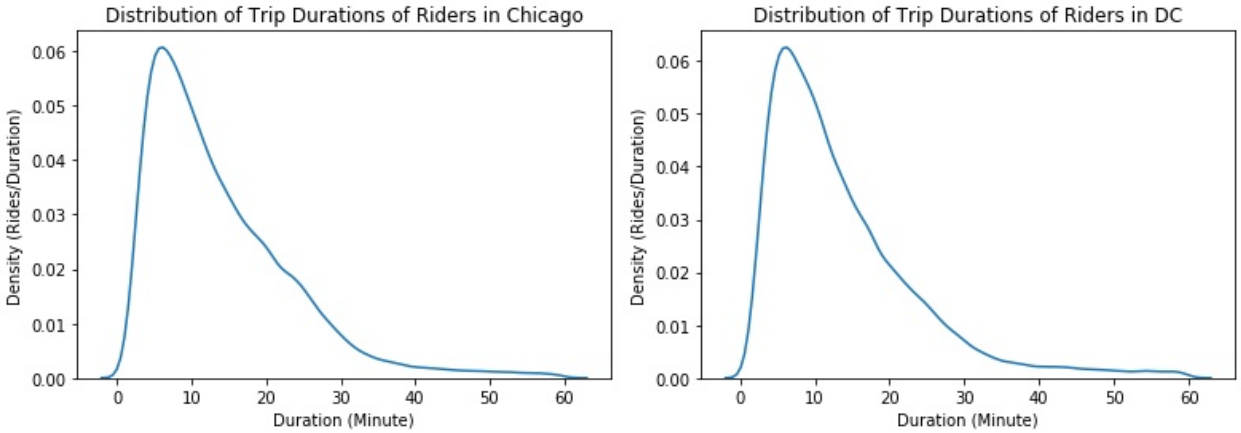


Figure 4: Distribution of Trip Durations in New York



(a) Trip Durations in Chicago (b) Trip Durations in DC

Figure 5: Distributions of Trip Durations in Chicago and DC

2. Yes, the plots generated above are useful as it clearly demonstrates the shapes of the distribution. Note that for the plots above, this was only possible after the x-axis was limited to only span from 0 to 60 minutes, i.e., only including trips of 1 hour or less.

3. Below are plots of the start time of trips split by hour for all three cities

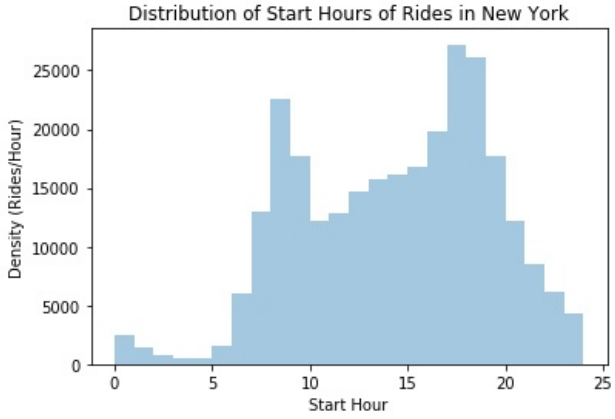
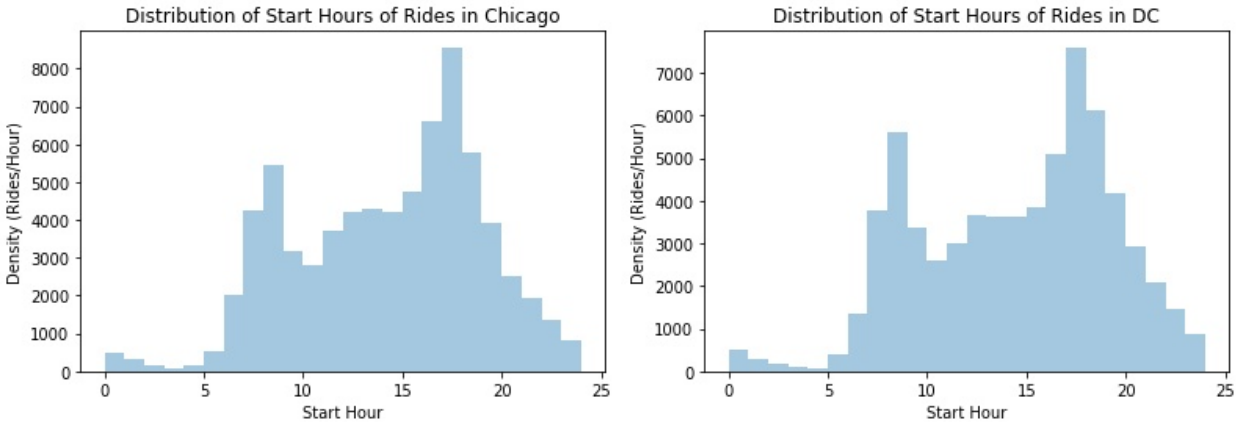


Figure 6: Distribution of Start Time of Trips in New York

See next page for distributions of start times of trips in Chicago and DC.



(a) Start Time of Trips in Chicago (b) Start Time of Trips in DC

Figure 7: Distributions of Start Time of Trips in Chicago and DC

4. Intuitively, the start hours of rides would typically be expected to be in the morning and evening rush hours. This generally lines up with the distributions of start hours of rides in all of the three cities, with peaks around 7-9 AM for the morning rush hour, and 4-7PM in the evening rush hour.

1.3 Further Exploration

In this section you should:

1. Below are visualizations of the distributions of user types, the number of trips from each station, and the number of trips using each bike.

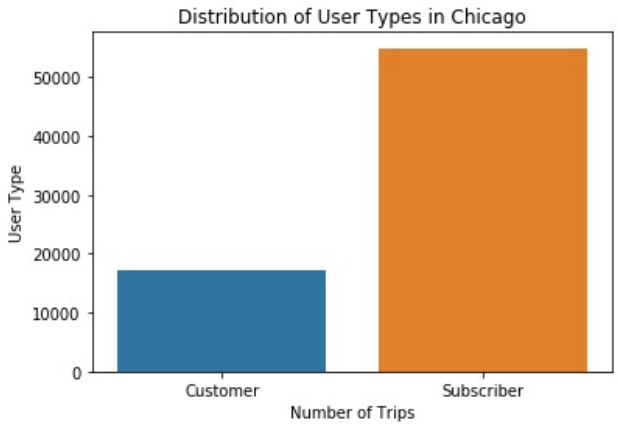


Figure 8: Distribution of User Types in Chicago

From this graph, it is clear that most of the rides are taken by subscribers, i.e., riders who have committed to bike-sharing platforms, rather than just customers.

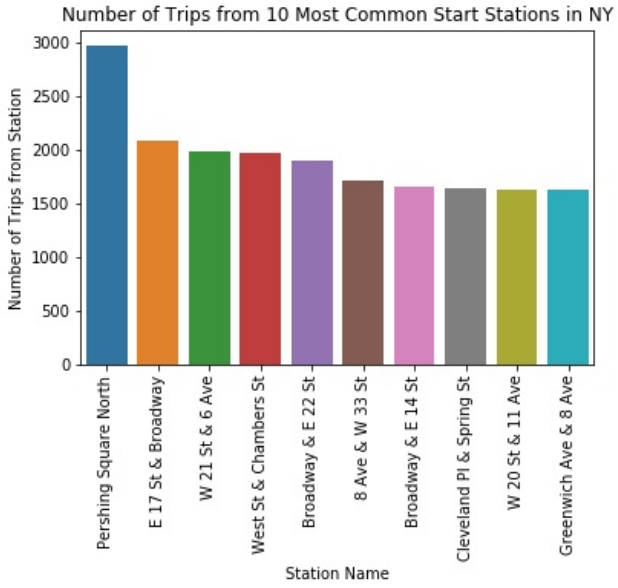


Figure 9: Number of Trips from 10 Most Common Start Stations in New York

From this graph, it is evident Pershing Square North is the most common station to start a trip from in New York. The other stations where a large number of rides originate are generally along the rivers, suggesting commuters are using bikes to get to work after starting farther away from the city centre.

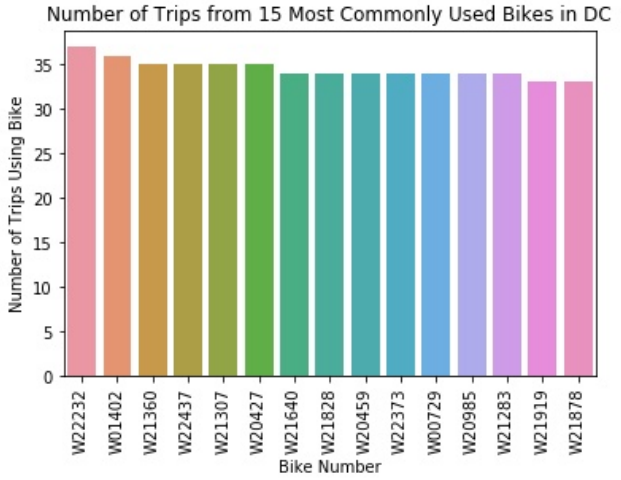
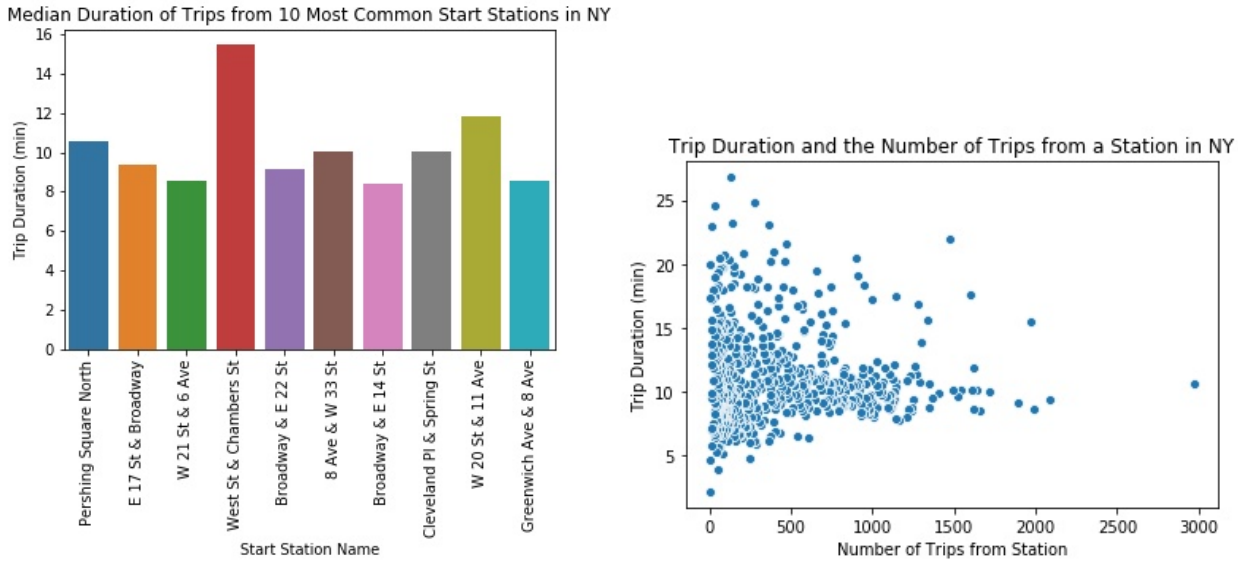


Figure 10: Number of Trips from 15 Most Commonly Used Bikes in DC

The most commonly used bikes in DC were used more than 30 times, with the most-used bike, WW22232, being used 36 times.

2. Below, the relationship between the trip duration and the station the trip originates from is examined in a bit more detail. Compared with the average trip duration of 11.19 minutes, most of these stations have below-average trip durations, suggesting they are frequently used by commuters on shorter trips. This supports why these

particular stations have a high number of trips, as a large number of commuters come by every day on their daily commute.



(a) Median Duration of Trips from 10 Most Common Start Stations in New York

(b) Scatterplot of the Trip Duration and the Number of Trips from a Station in New York

Figure 11: Relationship between the Trip Duration and the Number of Trips from a Station in New York

The scatterplot above is limited to only stations from which 5 or more trips originate and demonstrates that for some of the more infrequent trips, there are extremely high trip durations. Possible explanations are that the median trip duration is more likely to be affected by outlier trips, either extremely short or extremely long.

3. A possible hypothesis that could be tested is the more common it is to start a ride from a particular station, the shorter the trip duration. This could be tested using a null hypothesis that there is no effect of how frequently a trip starts from a station and how long a median trip from that station is. The alternative hypothesis is that stations where more trips start from have longer median trip durations. A possible test statistic would be the median trip duration from a station subtracted by the average trip duration for all the stations - larger values of this test statistic would support the alternative hypothesis over the null hypothesis. In order to carry out such a hypothesis test, median trip durations would have to be simulated following a null distribution centred at the average trip duration for all the stations.

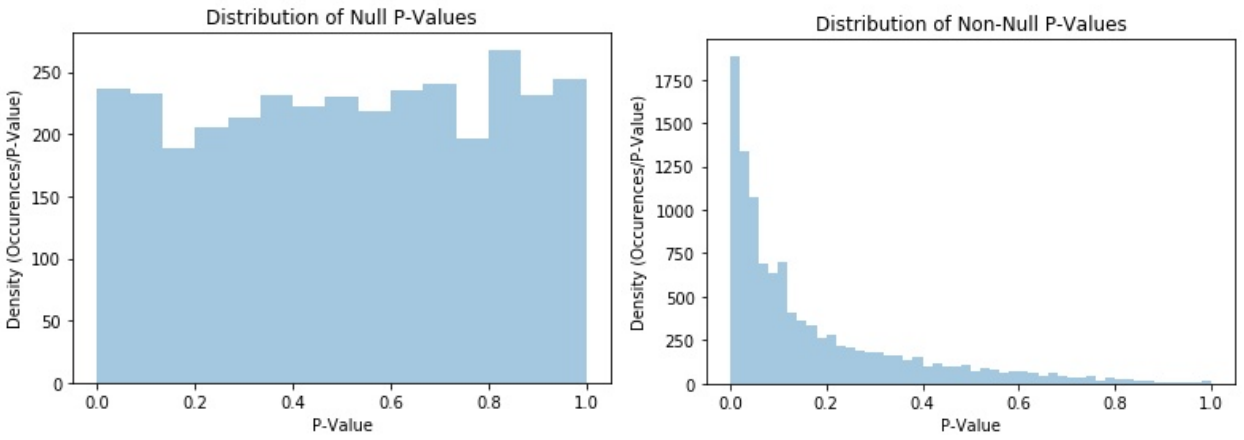
2 Hypothesis Testing

1. The logistic regression model

$$\mathbb{P}(Y_i = 1|X_i) = \frac{1}{1 + e^{-\theta^\top X_i}},$$

was fit using the data from the first split S_1. The optimal θ_* was found to be $\theta_* = (-0.00901922, 0.21235621, -0.23708859)^t$.

2. Below are two histograms: one of null p-values ($Y_i = 0$, casual riders) and one of non-null p-values ($Y_i = 1$, non-casual riders). The distribution of null p-values looks



(a) Null P-Values (b) Non-Null P-Values

Figure 12: Comparison of the distributions of P-Values for Casual Riders and Subscribers

roughly uniform, while the distribution of null p-values is right-skewed with a long right tail, with most p-values being around zero - it appears to look roughly exponential.

3. Running the Benjamini-Hochberg algorithm under level 0.2 on the p-values for the S_2 split of the data resulted in a false discovery proportion (FDP) of 0.0477 and sensitivity of 0.407.

   The average sensitivity is 0.341 over these 200 trials. The average false discovery proportion over these same 200 trials is 0.047, which is less than 0.2. This is because the Benjamini-Hochberg procedure controls the FDR, the expected false discovery proportion to be less than the specified α bound of 0.2.

2.1 Improving the Average Sensitivity

The steps taken to improve the sensitivity were to first add a new feature, the birth year, as this could be another plausible factor in determining the usertype of the rider. Then, all the features were standardized to have mean zero and standard deviation of 1 - this would allow for more accurate comparison between the features, since these variables are all on different scales - without normalizing, a birth year of 1990 for example may be unfairly weighted a lot more heavily than a start hour of 10 simply because they are of much different magnitudes. Only looking at the standard units ensures an equal comparison between the different variables. Finally, these standardized variables were then winsorized, i.e., measurements of magnitude greater than 4 were set to positive or negative four respectively. This was done to control the effect of outliers - these measurements more than 4 standard deviations away from the mean were extreme outliers relative to the other data points. For example, some

riders had a trip duration more than 20 standard deviations away from the mean. These implausible duration played an outstretched effect on the model being used.

After these steps were taken, the average False Discovery Proportion across 200 trials was 0.048 and the average sensitivity across 200 trials was 0.472, a 38% improvement from the previous average sensitivity of 0.341.

3 Gaussian Mixture Models of Trip Durations

1. Subscribers are regular users and hence are more likely to be everyday commuters, i.e., more likely to use bikes for relatively longer distances every day, rather than casual customers who are more likely to be tourists are just using the bikes for short-distance leisure. This intuition is supported by the plot below comparing the distributions of trip duration for customers and subscribers.

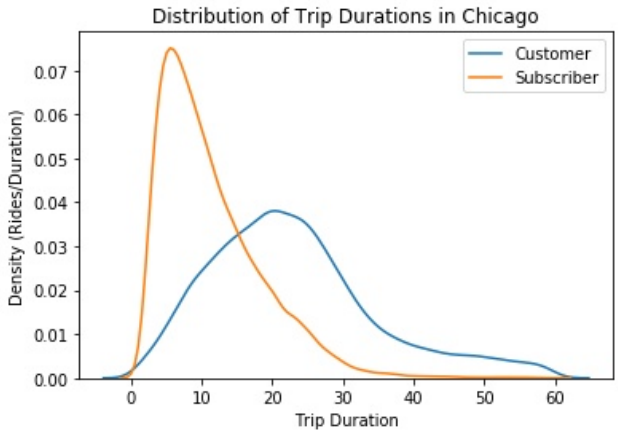


Figure 13: Distributions of Trip Duration for Customers and Subscribers

2. The Expectation-Maximization (E-M) Algorithm was then used to learn a mixture of two Gaussians that describes the distribution of the durations of trips less than an hour in length in Chicago. Results are below for the first initialization where π_0 is the proprotion of riders which are just customers, while μ_0 and σ_0 are their mean and standard deviation respectively. On the other hand, μ_1 and σ_1 are the mean and standard deviation respectively of the subscribers - the proportion of subscribers is π_1.

   However, running this multiple times from different initializations does in fact result in significant change to the estimates depending on the initialization. For example, using an initalization where the two distributions were presumed to be equal, i.e., $pi_0 = \pi_1$ and $\mu_0 = \mu_1 = 30$, the following is a table of the results:

3. The second normal distribution with a relatively lower average of 8.82 minutes and standard deviation of 7.26 captures the distribution of the trip duration of subscribers. For each customer, in the dataset, the posterior probability of that customer coming from this distribution was then calculated (see code Appendix).

Parameter	Estimate
π_0	0.303
μ_0	25.827
σ_0	11.787
π_1	0.697
μ_1	8.820
σ_1	7.258

Table 1: Expectation-Maximization Algorithm Results for Two Gaussians Using Optimal Initialization

Parameter	Estimate
π_0	0.175
μ_0	30.526
σ_0	7.258
π_1	0.825
μ_1	10.450
σ_1	11.787

Table 2: Expectation-Maximization Algorithm Results for Two Gaussians Using Uniform Initalization (Sub-Optimal)

4. Designing a classifier which classifies a customer as a Subscriber if their posterior probability is greater that 0.5, the accuracy of such a predictor is 77.51%, with an error of 22.49% on the true user types in the Chicago dataset.

5. The classifier performs comparably for the New York and DC datasets, with accuracies for all three cities between 76% and 79% even though the model was trained only on the Chicago dataset. In fact, the classifier surprisingly has even higher accuracy on the DC dataset than it does on the Chicago dataset.

City	Classifier Accuracy (%)
Chicago	77.51
New York	76.31
DC	78.13

Table 3: Accuracy of Classifier on Various Cities

4 Causality and Experiment Design.

4.1 Using 2-Stage Least Squares to Estimate the Effect of Weather on the Number of Bike Rentals

4.1.1 The Causal Model

1. Below is the graph of the causal model, with arrows between variables to denote causality among the following variables: temperature, `weathersit`, humidity, and the number of rentals.

   The instrumental variable is the humidity, the explanatory variables are the temperature and `weathersit`, and the response variable is the number of bike rentals.

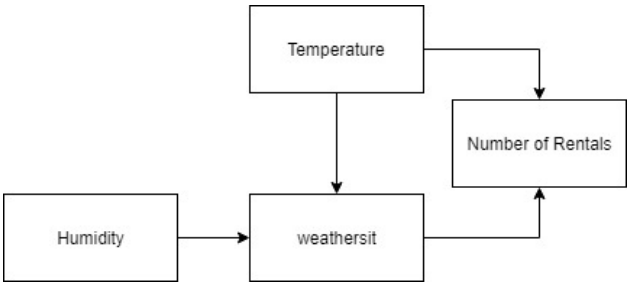
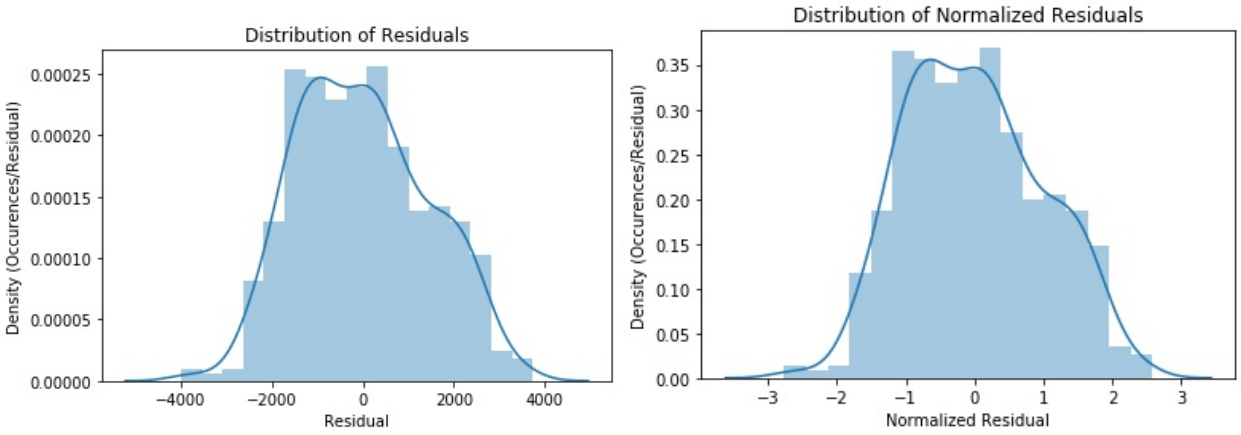


Figure 14: Causal Model Diagram

2. The assumptions for doing a 2-stage least squares are that the residuals are normally distributed, the instrumental variable must be correlated with the explanatory variable, and the instrumental variable cannot be correlated with the residuals given the other covariates. Note that generally 2-stage least squares is used because of a correlation between the explanatory variable and the residuals.

   The graphs below demonstrate that the residuals are roughly normally distributed. The normalized residuals are included since they are required later on to calculate the correlations between the residuals and other variables.



(a) Distribution of Residuals

(b) Distribution of Normalized Residuals

Figure 15: Distributions of Residuals and Normalized Residuals

The correlation between the normalized residuals and the normalized instrumental variable is extremely small, at just 5.69 x 10^{-16} and thus the residuals and instrumental variables are uncorrelated.

The correlations between the normalized instrumental variable and the normalized explanatory variables are 0.13 for the humidity and 0.59 for the `weathersit`. Therefore, the instrumental variable is correlated with both explanatory variables.

4.1.2 2 Stage Least Squares

1. The 2-stage least squares procedure is used when there is a correlation between the explanatory variables and the residuals, which will mean ordinary least squares is not possible as there will be a biased prediction. As a result, an instrumental variable correlated with the explanatory variable but not correlated with the residuals is used. This instrumental variable only affects the response variable through its effect on the explanatory variables. To implement this 2SLS, there are two linear regression. First, the instrumental variable and one of the explanatory variables is used to predict the other explanatory variable, in this case the humidity and the temperature is used to predict the `weathersit` variable. The second regression uses one of the original explanatory variables and the other predicted explanatory variable to predict the response variable, in this case the temperature and predicted `weathersit` variables are used to predict the number of bike rentals.

2. Below are two tables summarizing the two regressions performed using the 2-stage least squares

OLS Regression Results

Dep. Variable:	weathersit	R-squared:	0.388
Model:	OLS	Adj. R-squared:	0.387
Method:	Least Squares	F-statistic:	231.0
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	2.07e-78
Time:	22:49:53	Log-Likelihood:	-413.30
No. Observations:	731	AIC:	832.6
Df Residuals:	728	BIC:	846.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2082	0.079	2.629	0.009	0.053	0.364
temp	-0.5919	0.087	-6.804	0.000	-0.763	-0.421
hum	2.3578	0.112	21.088	0.000	2.138	2.577

Omnibus:	117.669	Durbin-Watson:	1.574
Prob(Omnibus):	0.000	Jarque-Bera (JB):	310.375
Skew:	0.826	Prob(JB):	4.01e-68
Kurtosis:	5.732	Cond. No.	10.7

(a) Predicting `weathersit`

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.427
Model:	OLS	Adj. R-squared:	0.425
Method:	Least Squares	F-statistic:	271.0
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	1.06e-88
Time:	22:52:23	Log-Likelihood:	-6366.3
No. Observations:	731	AIC:	1.274e+04
Df Residuals:	728	BIC:	1.275e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2877.9756	300.835	9.567	0.000	2287.367	3468.584
weathersit_predicted	-1057.2982	163.191	-6.479	0.000	-1377.679	-736.918
temp	6261.1390	302.680	20.686	0.000	5666.909	6855.369

Omnibus:	14.340	Durbin-Watson:	0.401
Prob(Omnibus):	0.001	Jarque-Bera (JB):	9.801
Skew:	0.154	Prob(JB):	0.00744
Kurtosis:	2.524	Cond. No.	13.3

(b) Predicting the Number of Bike Rentals

Figure 16: Summaries of Regressions Used In 2-Stage Least Squares Regression

The treatment affect of `weathersit` on the total number of bike rentals is -1057.30, which means the higher the value of the `weathersit`-predicted variable, the lower the number of bike rentals. This makes a lot of intuitive sense as high values of the `weathersit`-predicted variable mean there is bad weather. Thus, bad weather often happens along with fewer bike rentals. Specifically, for each increase in the severity of the weather, the number of bike rentals is expected to drop by roughly 1,057 bike rentals.

3. The treatment effect of `weathersit` on the number of casual bike rentals is -303.38, which means the higher the value of the `weathersit`-predicted variable, the lower the number of casual bike rentals. This again makes a lot of intuitive sense as high values of the `weathersit`-predicted variable mean there is bad weather. Thus, bad weather often happens along with fewer casual bike rentals. Specifically, for each increase in the severity of the weather, the number of casual bike rentals is expected to drop by roughly 303.

   The treatment effect of `weathersit` on the number of registered bike rentals is -753.92, which means the higher the value of the `weathersit`-predicted variable, the lower the number of registered bike rentals. This also makes a lot of intuitive sense as high values of the `weathersit`-predicted variable mean there is bad weather. Thus, bad weather often happens along with fewer registered bike rentals. Specifically, for each increase in the severity of the weather, the number of registered bike rentals is expected to drop by roughly 754.

4.1.3 Discussion

In this section you should discuss the following:

1. The question being asked was to evaluate the effect of the weather on the number of bike rentals among both casual and registered riders. This was tested by looking at the treatment effect of a variable that represents the state of the weather using a two-stage least squares regression due to pre-existing relationships between the explanatory variables and the residuals of such a model. The results of the analysis were that the

worse the weather, generally the fewer bike rentals there are among both casual and registered riders.

2. The treatment effect estimates were -303.38 for the casual bike rentals and -753.92 for the registered bike rentals. For both registered and casual bike rentals, the interpretation of these effects is that worse weather occured along with a drop in bike rentals. The magnitude of the treatment effect is clearly higher for registered bike rentals than it is for casual bike rentals. One reason this might be is that there are a lot more registered bike rentals than there are casual bike rentals - an increase in the severity of the weather would naturally lead to a larger decrease of registered bike rentals (754 in this case) than a decrease of casual bike rentals (303) simply due to the fact there are many more registered bike rentals than casual bike rentals. To more accurately compare the effect of weather on the casual/registered bike rentals, a possible option would be to normalize the data first so the magnitude of the numbers are actually comparable.

3. The model chosen above is relatively applicable to this problem, although there are variables unaccounted for from the causal graph that might be missing, for example the day of the week and whether the day is a holiday. These are just a few examples of confounding variables which likely also effect the number of bike rentals and are unaccounted for in the data. Related specifically to weather, variables like the air quality that day would also be potential candidates for inclusion in the causal graph. In terms of missing arrows, the humidity and temperature might have some interdependence between the two of them, so a more comprehensive model could involve some arrows going between those two variables.

4. If one were to design a new study and collect new data to test the effect of adverse weather on the number of rentals, the study would first try to limit the number of confounding variables. Therefore, data would only be collected on regular weekdays, for example every Wednesday, on the number of bike rentals among both casual and registered riders. Then, that date's temperature, humidity, air quality, and weather rating would be collected to use in the study to present a more comprehensive view of the weather on that date. This approach would not only reduce the potentially misleading results involved by not accounting for the day of the week, but also allow for more comprehensive conclusions.

5 Multi-Armed Bandits

5.1 Formalizing as a Multi-Armed Bandits Problem

In order to formalize the problem as a multi-armed bandits problem, the arms are each intersection, while the rewards are the number of promotional flyers that are handed out. The rewards would be modelled as sub-Gaussian, since most intersections likely have the same number of people walking through every day. They would not be bounded, since there is no limit to the number of people who could walk by an intersection assuming there is

an inexhaustible supply of flyers. The time horizon is the number of days for which the promotion will take place.

The modelling assumption is that the rewards for each arm are assumed to have fixed independent probability distributions, i.e., a stochastic environment. In addition, in this specific situation, the rewards are expected to be sub-Gaussian as explained in the earlier paragraph and to be non-negative, i.e., greater than or equal to zero. The non-negative rewards is reasonable, although the stochastic assumption might not hold since the number of flyers distributed may be affected by a number of different outside variables - for example, the weather of that day, the day of the week, the tome of day, and whether or not flyers had already been distributed at that intersection. These assumptions can be tested by gathering data on the intersection for different days of the week, i.e., seeing how many people visit the intersection between noon and 1pm on a sunny Monday compared to the same intersection between noon and 1pm on a rainy Monday.

The notion of regret being considered is the difference between the number of flyers actually distributed and the maximum number of flyers which could be distributed during that time period. If given perfect knowledge, the optimal strategy would be to go and distribute strategies from the intersection with the greatest reward every single day.

5.2 Simulate Upper Confidence Bound (UCB) Strategy Using Past Data

5.2.1 Implementation and Results

The Upper Confidence Bound (UCB) algorithm was then used to determine the best locations with past data, using the datsets `dc.csv`, `chicago.csv`, and `ny.csv`. In order to run the algorithm, it is first necessary to get the number of rides which begin and end at each station in the dataset. Then, at each station, there is a normal distribution with mean equal to the average number of daily rides at each station and standard deviation equal to the standard deviation of the number of daily rides at each station.

At each iteration of the simulation one of the stations is selected and an estimated number of rides is randomly generated from its corresponding distribution as described earlier. In order to choose which station to sample from, the UCB algorithm needs to be used in this situation.

This decision is based on previous history having pulled from different arms - intuitively, the algorithm wants to balance choosing from a station where it already knows it has a relatively high average number of rides while also making sure it explores other stations to see whether one of them might actually be better. The UCB algorithm threads this delicate balance between exploitation and exploration by examining the sample means and sample standard deviations of each station, and then choosing the station with the highest upper confidence bound, i.e., the highest sum of sample mean and a slightly modified sample standard deviation, which actually represents half of the width of the confidence interval. Specifically, the formula used for upper bound is:

$$C_a(T_a(t), \delta(t)) = \begin{cases} \infty & : T_a(t) = 0 \\ \hat{\mu}_a(t) + \sqrt{\frac{4\sigma^2}{T_a(t)} \log t} & : T_a(t) > 0 \end{cases}$$

In this case, if the intersection has not yet been visited, the upper bound is set to ∞ in order to force the algorithm to visit that intersection at least once. Afterwards, the upper bound is treated as the sum of the sample mean and a modified variant of the sample standard deviation. This second term in the summation represents half the width of the confidence interval and intuitively, as time goes on if the algorithm keeps on only picking one intersection, the width of that interval will decrease, which may help ensure the algorithm is only exploiting an intersection which is significantly better than the other intersections.Note that here the algorithm assumes the sub-Gaussianity of the rewards, as Chernoff/Hoeffding bounds in implementing the UCB algorithm.

The associated parameters for the distributions from which the estimated number of riders is generated from are instantiated using the means and standard deviations of the average number of rides at each of the 10 most-visited stations as described earlier. The regret is instantiated to be zero initially as it is before any decisions on which intersection to visit are made. The number of visits to each station is obviously zero when initiated.

The plots below illustrate the regret of the UCB algorithm over time for Chicago, as well as the number of total pulls of each arm over time, and finally the true/estimated means and upper confidence bounds for each arm over time.

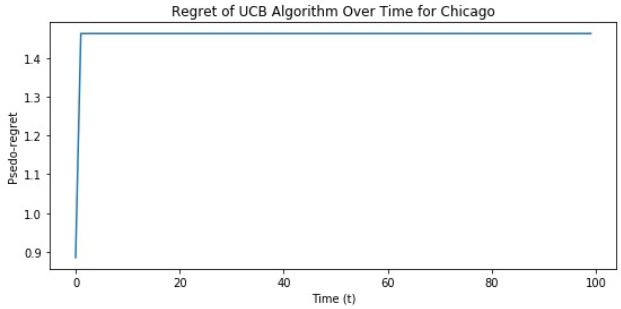
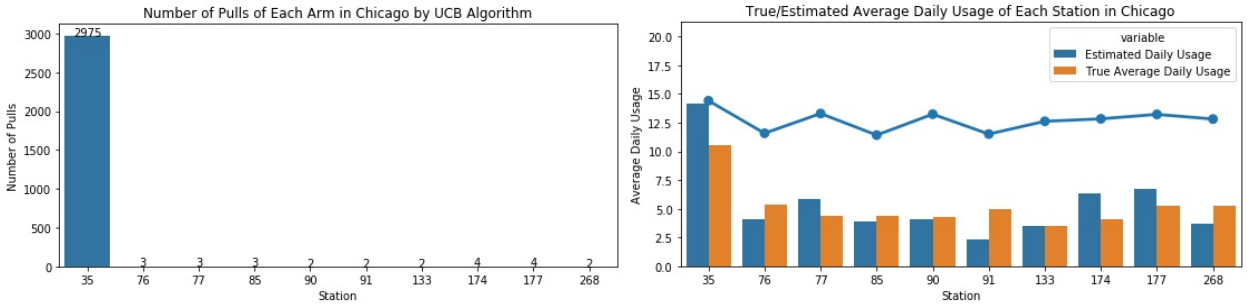


Figure 17: Regret of the UCB Algorithm for Chicago



(a) Number of Pulls of Each Arm

(b) Average Daily Usage

Figure 18: Number of Visits to Each Station by UCB Algorithm and True/Estimated & Upper Confidence Bound of Average Daily Usage of Each Station in Chicago

Note that for the graph on the right above, the blue bars represent the estimated average daily usage of each station in Chicago, while the orange bars represent the actual daily usage of each station and the blue line represents the upper confidence bound.

It is clear for Chicago that the algorithm was able to correctly identify the most-visited station in Chicago, station 35, after about 15 days had passed. This was identified so quickly because station 35 had significantly more visitors than the other stations in Chicago, as evidenced by how much larger the true average daily usage of station 35 was relative to the others. Almost exclusively visiting station 35 after the first few days, the upper confidence bound for station 35 ended up being extremely close with the estimated daily average.

The analogous three plots are repeated below for New York. Specifically, the three plots below are the regret of the UCB algorithm over time for New York, as well as the number of total pulls of each arm over time, and finally the true/estimated means and upper confidence bounds for each arm over time.

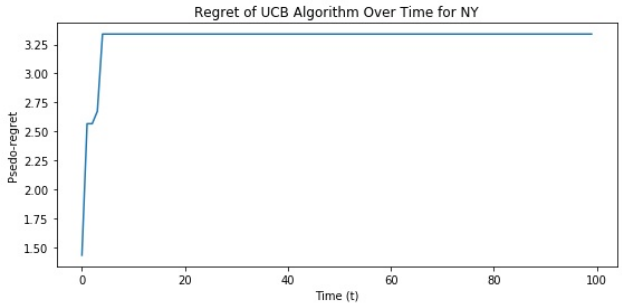
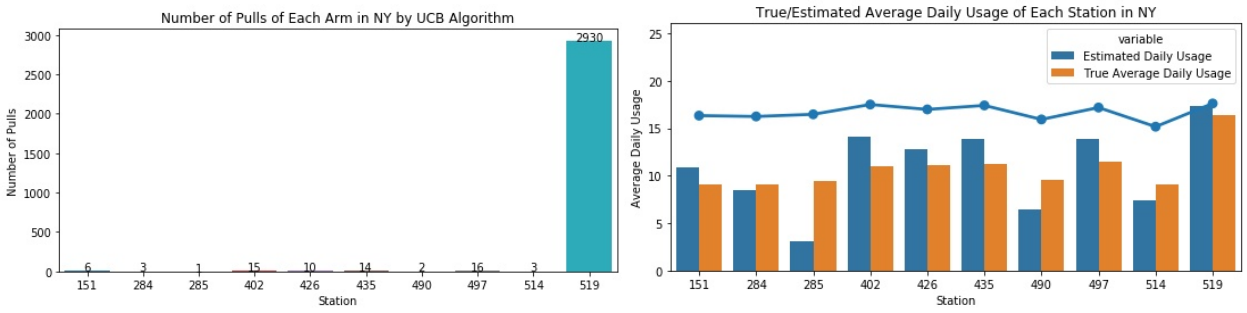


Figure 19: Regret of the UCB Algorithm for New York

Note that again for the graph on the right on the next page below, the blue bars represent the estimated average daily usage of each station in New York, while the orange bars represent the actual daily usage of each station and the blue line represents the upper confidence bound.



(a) Number of Pulls of Each Arm (b) Average Daily Usage

Figure 20: Number of Visits to Each Station by UCB Algorithm and True/Estimated & Upper Confidence Bound of Average Daily Usage of Each Station in NY

It is clear for New York that the algorithm was able to correctly identify the most-visited station in New York, station 519, after about 60 days had passed. It took relatively longer to identify the most-frequented station in New York compared to how long it took for Chicago since there were multiple stations in New York with similar levels of usage - this is demonstrated by the relatively constant upper confidence bound, as represented by the blue line. Almost exclusively visiting station 519 after the first 60 days, the upper confidence bound for station 519 ended up being extremely close with the estimated daily average.

The analogous three plots are repeated below for Washington DC. Specifically, the three plots below are the regret of the UCB algorithm over time for Washington DC, as well as the number of total pulls of each arm over time, and finally the true/estimated means and upper confidence bounds for each arm over time.

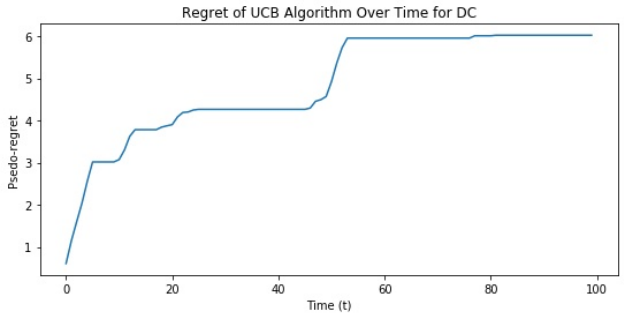
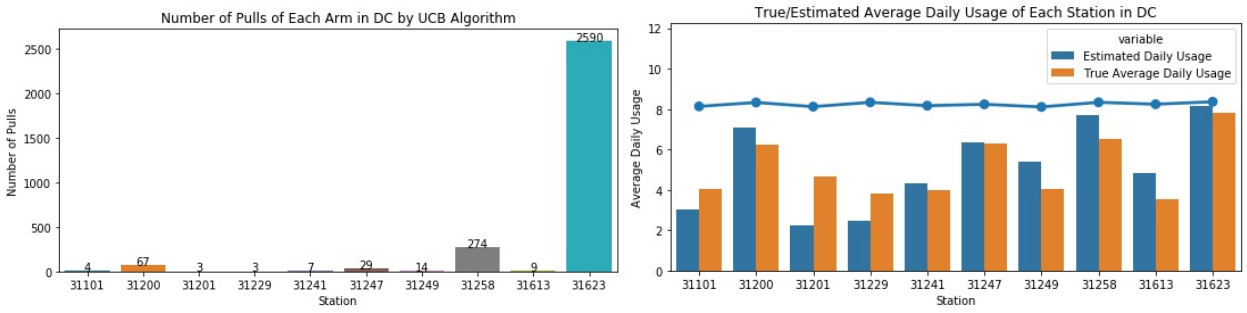


Figure 21: Regret of the UCB Algorithm for DC

Note that again for the graph on the right on the next page below, the blue bars represent the estimated average daily usage of each station in New York, while the orange bars represent the actual daily usage of each station and the blue line represents the upper confidence bound.



(a) Number of Pulls of Each Arm

(b) Average Daily Usage

Figure 22: Number of Visits to Each Station by UCB Algorithm and True/Estimated & Upper Confidence Bound of Average Daily Usage of Each Station in DC

It is clear for Washington DC that the algorithm was able to correctly identify the most-visited station in DC, station 31623, after about 25 days had passed. It took relatively longer to identify the most-frequented station in DC since there were multiple stations in DC with similar levels of usage - this is demonstrated by the relatively constant upper confidence bound, as represented by the blue line. Compared to other cities, another non-optimal station was visited relatively more often, in this case station 31258. Almost exclusively visiting station 31623 after the first 25 days, the upper confidence bound for station 31623 ended up being extremely close with the estimated daily average.

5.2.2 Discussion

The order of which arms to pull first does not matter - this is because the UCB algorithm does not take into account the order in which the arms are pulled, only the number of times each arm has been pulled. In addition, the rewards of each arm are assumed to be generated from indepdendent distributions so drawing from one arm will not affect the value drawn from another arm. However, because the estimated rewards for each arm are randomly generated from a distribution, the results of the simulation changed when the algorithm was re-run with the stations sorted in a particular order, not because of the ordering of the stations but instead because of the randomness in generating the rewards.

An adaptive strategy that could beat the no-regret strategy of always sending your employee to the location with the highest average number of rides is to send an employee to different locations depending on the day of the week. For example, a particular station might have a high number of commuters on weekdays but few number of visitors on weekends. As a result, sending the employee there on weekends would not make much sense. By adapting which station to send the employee to depending on the day of the week, there would be a higher number of people that could be reached than the no-regret strategy discussed earlier.

5.3 Takeaways

The simulation of the problem is somewhat applicable to the problem posed in Section 1.1 - although it is not exactly the same, there are some parallels for example representing an intersection of the city by a nearby bike station. The simulation is suitable for answering

whether it is worth investing in the promotional program in your city by identifying which general areas are likely to have a high number of bike riders, which could be a reasonable proxy for the number of visitors to an intersection. The simulated experiment is lacking first of all since not all visitors going by an intersection ride a bike. In addition, there is difficulty in getting information about a specific street corner, since it only has data on the number of bike rides to and from a station.

The multi-armed bandits formulation to the task at hand in Section 1.1 is not entirely applicable. This is because the number of flyers distributed at each intersection may vary depending on how many times that intersection has already been visited - for example, if the same intersection is being flyered every day the people walking by each day are likely also the same people and thus will not take a flyer after having already received one. In addition, whether or not flyers have been distributed at nearby intersections may in fact influence the number of flyers distributed at the intersection in question, a violation of the independent rewards assumption.

Ultimately, the UCB algorithm would be recommended to adaptively place the person handing out promotional flyers. That being said, a modified version would have to be used to take into account the issues discussed above and revealed in the simulation. Instead of just visiting the same intersection once it has been identified to be the most-visited, there would need to be more effort to visit other intersections farther away which might not have as many visitors on the surface, but also have not yet been over-exposed to flyers. Again, different intersections should be visited on different days depending on the specific past data on the number of visitors at each intersection depending on the day of the week.

6 Privacy Concerns

6.1 Exploratory Analysis

Below is a plot of the number of females and the number of males in the leaked dataset

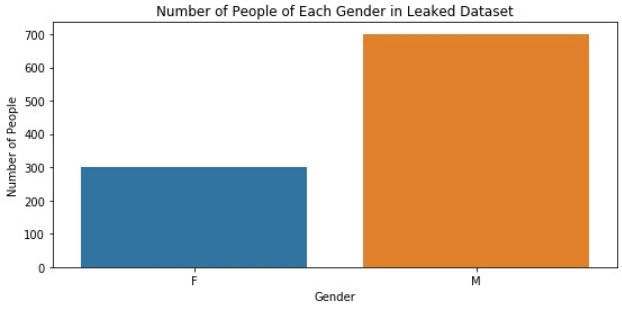


Figure 23: Number of People of Each Gender in Leaked Dataset

The distribution of the number of people of each gender in the leaked dataset does not appear to be uniform, as there are many more males than females in the leaked dataset. This distribution of gender is similar with what was observed in Part 1, when there were many more male than female riders.

Below are plots of the number of people born in each month and year.

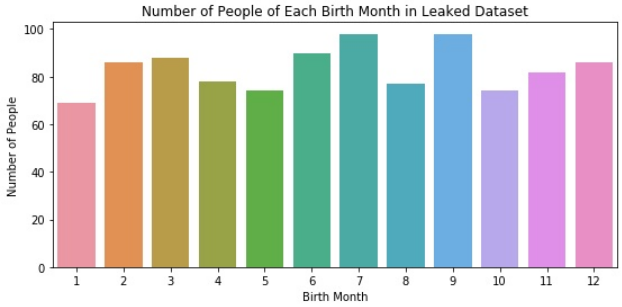


Figure 24: Number of People of Each Birth Month in Leaked Dataset

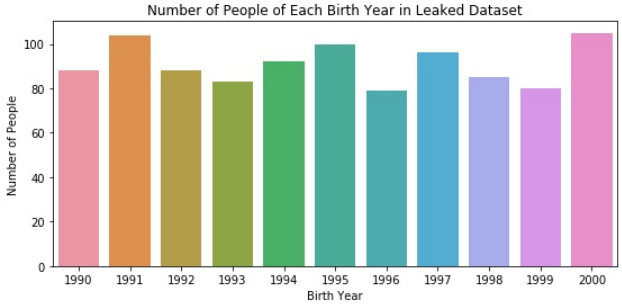


Figure 25: Number of People of Each Birth Year in Leaked Dataset

The distribution of birth months and birth years in the leaked dataset both appear to be roughly uniform, although of course some birth months or birth years have more people than others as would be expected due to random chance. The distribution of the birth years in the leaked dataset differs sharply from that of the birth years in the datasets of bike riders for each city that was observed in Part 1, which instead of having a roughly uniform distribution from 1990 to 2000 had a non-symmetric, uni-modal distribution with a long left tail that contained birth years as early as 1940 for example.

6.2 Simple Proof of Concept

There were 43 users from the leaked dataset can be isolated just from just three attributes, their gender, birth month, and birth year. This was done by finding there were 43 unique combinations of gender, birth month, and birth year which had only a single person in the leaked dataset, thus setting a one-to-one correspondence for those 43 identifiable users between the Berkeley and leaked datasets.

Below on the following page is a plot of the number of females and the number of males in the set of identifiable users.

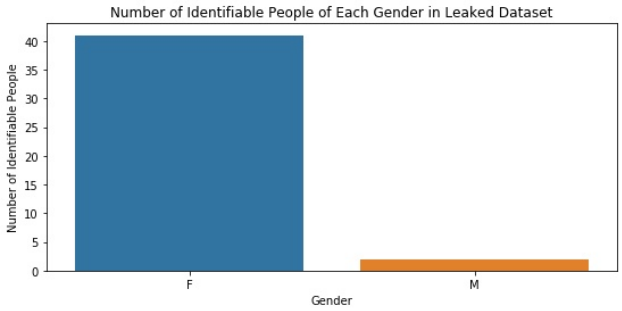


Figure 26: Number of Identifiable People of Each Gender in Leaked Dataset

The distribution of identifiable people of each gender in the leaked dataset demonstrates a lot more females than males. Therefore, this does not match the distribution of gender in the overall leaked dataset since that dataset has a lot more males than females. This difference between the distributions is due to the fact that since there are a lot fewer females, females are much more easily identifiable as it is a lot less common for there to be multiple females which share a birth month/year.

Below is a plot of the distribution of birth months for identifiable users.

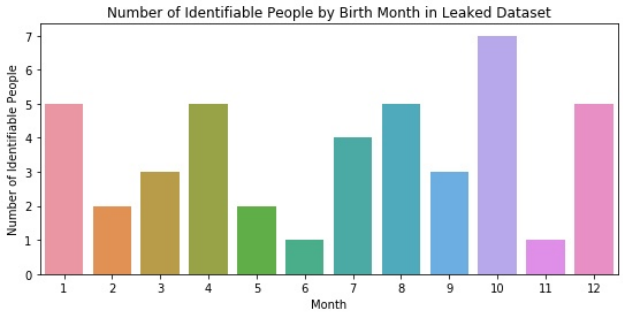


Figure 27: Number of Identifiable People of Each Birth Month in Leaked Dataset

The distribution of identifiable people of each birth month does not seem uniform like the distribution of birth month in the overall leaked dataset. This is due to the fact that there are a lot fewer people, and thus there will be more noticable fluctuation in the number of people born in each month even if both underlying distributions are actually uniform.

The table of identifiable users was then merged with the Berkeley dataset to get the number of scooter rentals each identifiable user had done. As an example, the identifiable users with the top-five number of scooters rented is displayed on the table on the following below:

Identifiable User	Number of Rentals
Sophia Hall	97
Harper Harris	92
Ella Lee	92
Abigail Campbell	90
Emily Martin	86

Table 4: Top 5 Number of Scooters Rented by Identifiable Users

6.3 A More Elaborate Attack

The estimates of $p_1 = 0.0966$ and $p_2 = 0.278$ were generated by looking at the identifiable bike rides. p_1 was simply set to be equal to the proportion of trips with different start and home zip codes among all trips by identifiable users, while p_2 was equal to the proportion of trips with different end and home zip codes among all trips by identifiable users. Note that the Maximum Likelihood Estimate of the proportion of a Bernoulli is simply the sample mean, here the sample proportion as described above.

To get the 95% confidence intervals for these estimated p_1 and p_2 values, first it was necessary to get the standard deviation of the estimates. This was obtained by using a bootstrap - first, a new re-sampled table was created by drawing from the original table with replacement to get new estimates for p_1 and p_2. Then, this entire process was repeated 1,000 times to get 1,000 estimates for p_1 and p_2. The standard deviations of these estimates is used to approximate the actual standard deviation of the original estimate. As a result, the 95% confidence interval for p_1 is $0.0966 \pm 1.96 * 0.00643 = (0.0840, 0.109)$ and a 95% confidence interval for p_2 is $0.278 \pm 1.96 * 0.00958 = (0.259, 0.297)$.

There are 720 theoretically identifiable users in the leaked dataset using the gender, birth month, birth year, and home zip code of each user. This assumes that the Berkeley dataset had the address of each user, and the list of theoretically identifiable users was created by generating all the possible combinations of gender, birth month, birth year, and home zip code of each user and seeing how many users there were in each unique combination. The categories with only 1 such user were deemed to be unique, as there was a theoretically identifiable user in each of those categories.

The algorithm that will be implemented first predicts the home zip code of the rider of each trip in the Berkeley dataset. If the start and end zip codes are the same, the predicted home zip code is that common shared zip code. If they differ, then the algorithm has to choose which of the two zip codes to set as the home zip code. Assuming the estimated values of p_1 and p_2 above are actually equal to the true values, then approximately 9.66% of the time, the start zip code will be different from the home zip code, while the end zip code will be different from the home zip code approximately 27.8% of the time.

Therefore, with probability $(p_1)(1\text{-}p_2)$, the given start zip code is not the home zip code and the end zip code is actually the home zip code. Likewise, with probability $(1\text{-}p_1)(p_2)$, the start zip code is actually the home zip code and the end zip code is not the home zip code. In this situation, the home and end zip codes differ so these are the only to situation which will be considered. Thus, with probability $\frac{p_1(1-p_2)}{p_1(1-p_2)+(1-p_1)p_2} \approx 0.217$ the algorithm will

take the end zip code as the home zip code. The rest of the time, i.e., with probability $\frac{p_2(1-p_1)}{p_1(1-p_2)+(1-p_1)p_2} \approx 0.783$ the algorithm will take the start zip code as the end zip code.

After generating those predicted home zip codes, the algorithm then matches each row in the Berkeley dataset with the leaked dataset using the combinations of gender, birth month, birth year, and home zip code. This resulted in a perfect one-to-one match for all the theoretically identifiable users. For the non-theoretically identifiable users, the algorithm randomly selected a user in the leaked dataset which met the desired combination of gender, birth month, birth year, and home zip code.

6.4 Takeaways

The findings above demonstrate that all the theoretically identifiable users can in fact have their rides identified with a relatively high degree of certainty. Even though the zip code of each rider's home is not given directly, it can be derived with reasonable confidence given the start and end locations of the trip.

With the already-released data, it would be advisable to try and limit the ease with which it can be accessed, as personal information linking each person with their zip code for example should not be freely available to everyone. In addition, it my be advisable to look into removing more identifiable information such as the birth month, which did not add a lot of value when analyzing the dataset, i.e., in predicting which users were registered and which users was casual, but had an abnormally large effect in identifying the users given the leaked data.

In releasing future datasets, just removing overt Personal Identifying Information (PII) would not be enough - more needs to be done to potentially remove other information which can aid in identifying users. Currently the form of the data has too much potentially identifiable information. Therefore, data like the birth month could be removed in the future to make it more difficult to link users in this dataset with leaked datasets.