MUSA 620 Term Project Draft

Group Name: A plus^3

Group Member: Zixuan Xu, Yuchen Zhang, Yuyang Yin

Evaluate the New Deal of Expanding the Divvy Stations to Far South and West Side of Chicago City

1. Introduction

In large cities, bike-share becomes an increasingly popular green transportation alternative for the city to ameliorate issues like traffic congestion, parking difficulty, and degradation of the aesthetic urban environment. More and more citizens and tourists choose to use bikes as an ideal option to commute or the last mile alternative to connect to public transit networks. However, it is also critical to make sure that the expansion of the bike-share system is non-rivalry and should be under control based on the users' demands. One possible solution is to ensure that the bikes are embedded in some relatively high-demand regions.

The City of Chicago has embraced its citywide bike-share system, Divvy, since 2011. Currently, Divvy operates over 6,000 bikes and 609 stations across Chicago, providing Chicagoans with a convenient, healthy, and affordable transportation alternative. As shown in Figure 1, most of the Divvy's bike-share stations are around the CBD area, which has the highest density of bike-share stations. However, several neighborhoods, such as Morgan Park and Ashburn, still do not popularize the bike-share service.

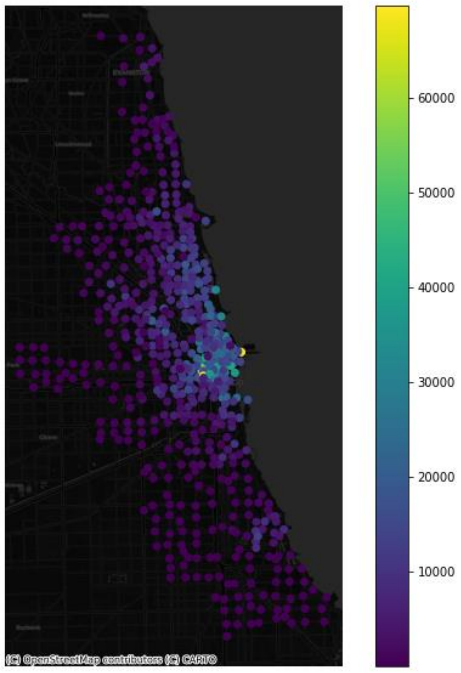


Figure 1. Trip counts of each station

In April 2019, the City Council of Chicago approved that the new Divvy contract, which will add 10,500 publicly-controlled electric-assist bikes to expand the system citywide. This time the city will also try to broaden the Divvy service area on the South and West sides of Chicago city for over eight years[i]. Expanding the Divvy system is an essential step to embracing the vision of an active multi-modal transportation network for the future of Chicago in an equitable manner. However, we should also avoid the typical failure of bike-share, such as over-supply.

Therefore, we carry out a project to help the city officials evaluate the new deal of expanding the Divvy station to the far south and west side of Chicago city. Spontaneously, we will offer suggestions on the possible regions for the Divvy where have relatively high demands. Meanwhile, we will also provide a dashboard for users to evaluate the performance level for each existing station, with interactive functions to help them explore their areas of interest.

2. Methodology

We intend to build linear regression models with diverse valid factors to predict the demand for the bike-share service in some specific areas and also identify the proposed new Divvy stations in the city of Chicago.

We will separate the entire process into three steps: data wrangling and feature engineering, building the first linear regression model to predict the trip counts of each station (to test performance of existing indicators), and building the second regression model on a fishnet of Chicago to predict the future demand for bike-share services across the whole city.

Users' behaviors and attitudes will mainly influence their demands of the bike-share service. The socio-economic profile of the user, the spatial locations of the user's origin and destination, and the availability of the bike-share system will drive users' attitudes on their needs of the bike-share service across the city of Chicago. Therefore, our model will include factors diversifying from demographic indicators to spatial amenities.

The criteria to choose potential indicators is based on commonly used demographic and socio-economic indicators that primarily relate to citizens' healthy, sustainable, and affordable work-life styles in the city. Simultaneously, we will include distance to public infrastructures and facilities since the distance to nearby local amenities will also influence demand for bike-share services.

To propose new Divvy stations with high demand, we will also create a fishnet to cover the entire city of Chicago. After these steps, we will spatial join all the data back to the primary datasets, based on their geospatial locations within the fishnet.

Before building the models, we will use the correlation plot to select independent variables that highly associate with the dependent variable but not correlate to each other. Then we will use linear regression to build models.

After these steps, we will test the entire dataset and judge the results by its adjusted R-square, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Meanwhile, the k-fold cross-validation could help ensure the generalizability of the model.

3. Data Description

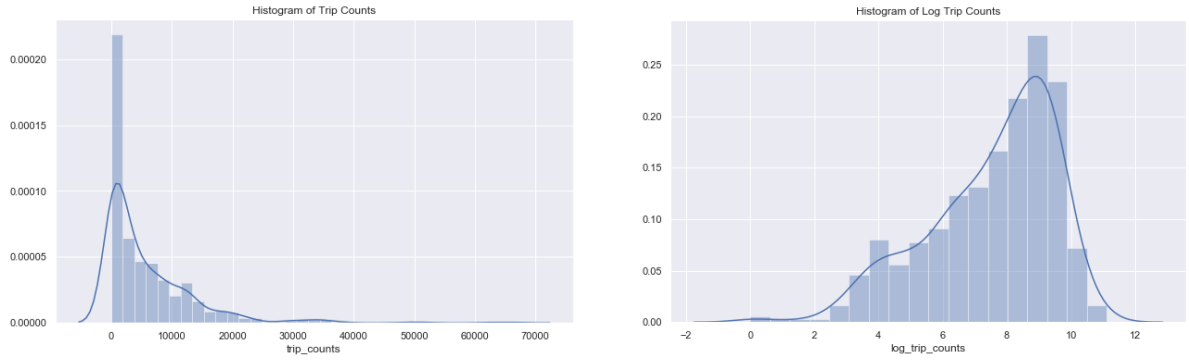
According to the methodology, we are going to utilize the historical Divvy bike-share trips from the Chicago Data Portal API as our primary working data set, which has more than 17 million rows. Other than the bike-share trips, we will also use Divvy bike-share station data set to help us figure out the station locations within the city. We will also join other relevant demographic data sets, such as census data, for demographic characteristics. For spatial influences, we will use the Chicago neighborhood boundary and transportation infrastructure and amenities from the Open Street Map, to demonstrate the importance of spatial structures on bike-share demand.

3.1 Main Datasets Overview

The primary dataset that we use in this project combines the two Divvy's datasets mentioned above, which are both available on the Chicago Data Portal API. Based on the '*start_time'* attribute, we query the data from 2018 until now. Interested in the demands of bike-share service at station level, we calculate and join the total trip counts since 2018 to each station.

3.2 Dependent Variable

For the dependent variable, we will use the total bike-share trip counts in each specific area in the city of Chicago, since 2018, which is from the Divvy bike-share trip dataset.



(a) Histogram of trip counts

(b) Histogram of logged trip counts

Figure 2. Histogram of dependent variables

We can observe that the original trip counts are left-skewed (Figure 2(a)). In order to remove the systematical residual changes and linearize the relationship between the dependent variable and predictors, we will log-transform the trip counts of each station. The transformed histogram is shown below.

It is evident that the log-transformed trip counts is more normally distributed than the original one (Figure 2(b)). Thus, we choose the log-transformed trip counts of each station as the dependent variable of our linear regression model.

3.3 Independent Variables

The independent variables are separated into five categories: demographic indicators, attraction indicators, transportation network indicator, regional indicators, physical indicator and spatial lag indicator. All these independent variables will be included in the regression model to examine their correlation to the dependent variable.

For transportation network indicators, we use the intersections from OpenStreetMap to calculate each grid cell's average distance to the nearest ten intersections and apply the log-transform algorithm. The attraction indicators and spatial lag indicator use the same calculation method mentioned above, showing the log-transformed distance from the station to the nearest amenities. All these indicators could represent the spatial structure of each grid cell in the fishnet.

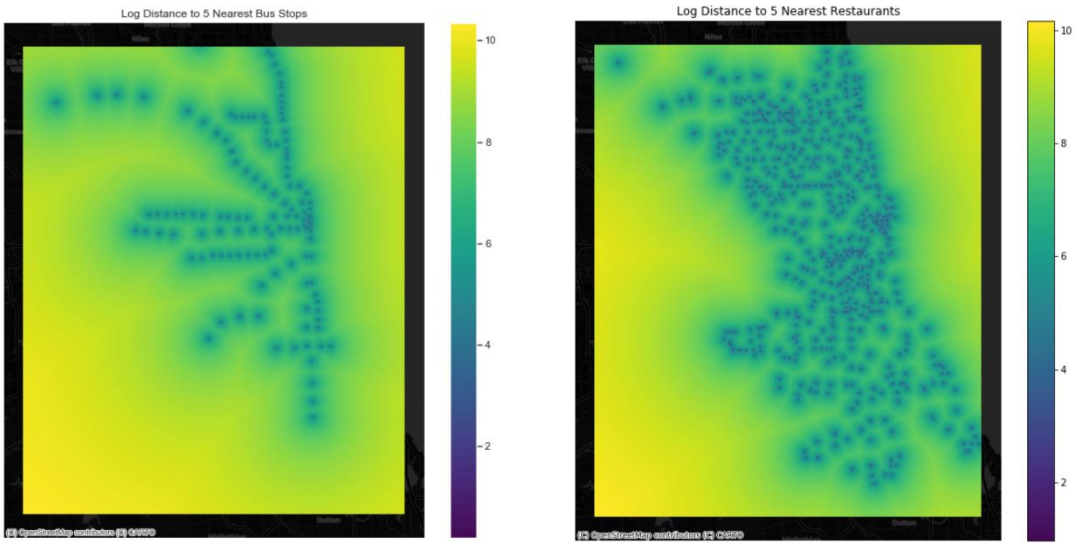
Table 1: Overview of all independent variables

Category	Variable Name	Type	Definition
Demographic indicators	Log_medHHInc (removed)	Numeric	Log-transformed median household income in census tract
	Percent_walk	Numeric	Percent of people walk to work in census tract
	Percent_public	Numeric	Percent of people taking public transits to work in census tract
	Percent_male	Numeric	Percent of males in census tract
	Percent_bachelors	Numeric	Percent of people get bachelor degree in census tract
Transportation Network indicator	Log_Dist2Intersection_10	Numeric	Log-transformed distance to 10 nearest intersections
Regional indicators	neighborhoods	Categories	The neighborhood that the trip started was in.
	ifCBD	Numeric	Whether the trip started WITHIN the CBD area
Attraction indicators	Log_Dist2Library	Numeric	Log-transformed distance to 1 nearest library
	Log_Dist2Restaurant_5	Numeric	Log-transformed distance to 1 nearest restaurant

	Log_Dist2AffordableHousing_5	Numeric	Log-transformed distance to 5 nearest affordable housings
	Log_Dist2Park_5	Numeric	Log-transformed distance to 5 nearest parks
	Log_Dist2Mural	Numeric	Log-transformed distance to 1 nearest mural wall
	Log_Dist2Landmark	Numeric	Log-transformed distance to 1 nearest landmark location
	Log_Dist2BusStop_5	Numeric	Log-transformed distance to 5 nearest bus stops
	Log_Dist2TransStation (removed)	Numeric	Log-transformed distance to 1 nearest train stations
	Log_Dist2BikeRack_5	Numeric	Log-transformed distance to 5 nearest bike racks
	Log_dist2PedStreet (removed)	Numeric	Log-transformed distance to 1 nearest pedestrian street
	Log_Dist2College	Numeric	Log-transformed distance to 1 nearest college (from OSM)
	Log_Dist2University (removed)	Numeric	Log-transformed distance to 1 nearest university (from OSM)
	Log_Dist2ArtsCenter	Numeric	Log-transformed distance to 1 nearest arts center (from OSM)
	Log_Dist2CommunityCenter	Numeric	Log-transformed distance to 1 nearest community center (from OSM)
	Log_Dist2Theatre	Numeric	Log-transformed distance to 1 nearest Theater (from OSM)
	Log_Dist2Market	Numeric	Log-transformed distance to 1 nearest Market place (from OSM)
Spatial Lag indicator	Log_loggedTrips_5	Numeric	Distance to the nearest 5 starting stations of the trips from *Station_Trip* dataset
Physical indicator	Total_docks	Numeric	The total number of bike-share docks in the area

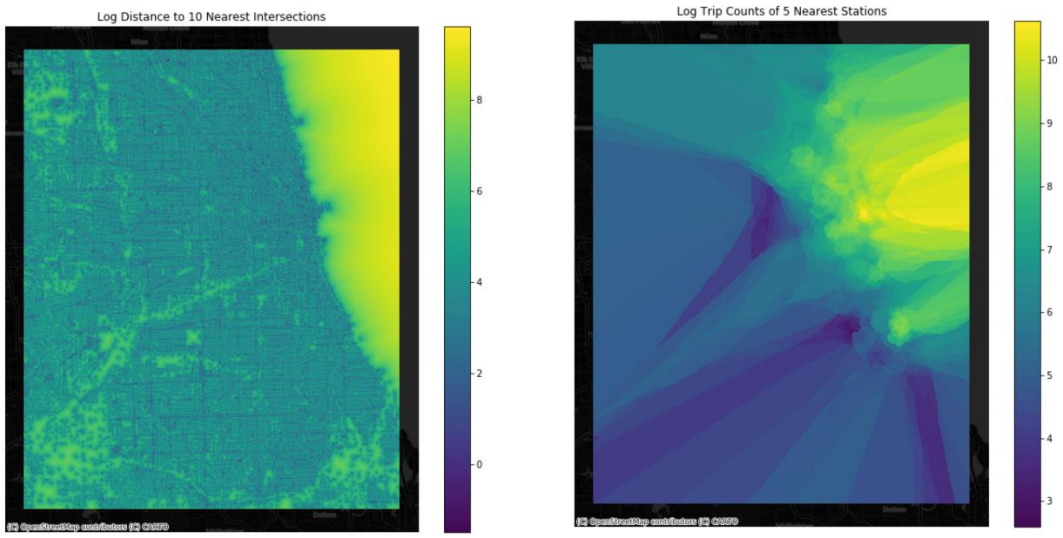
3.4 Data Exploration

3.4.1 Log-transformed distance to amenities



(a) Five nearest bus stops

(b) Nearest restaurants



(c) Ten Nearest Intersections

(d) Five nearest stations

Figure 3. Map of log-transformed distance to amenities

Compared to restaurants, bus stops are more evenly located across the whole city (Figure 3(a)).

3.4.2 Scatterplots of independent variables against dependent variable

Below shows 4 linear regression scatterplots bewteen log_trip_counts and log_laggedTrips_5, log_Dist2BusStop_5, log_Dist2Resraurant_5, and log_Dist2Theatre (Figure 4). We can see that log_laggedTrips_5 and log_Dist2Theatre have stronger linear relation with log_trip_counts, which shows relative strong correlation. While log_Dist2BusStops_5 and log_Dist2Resaurant_5 are more scattered and have lower correlation.

(a) Log transformed nearest 5 stations (b) Log transformed nearest 5 bus stations

(c) Log transformed nearest 5 restaurants (d) Log transformed nearest threatre

Figure 4. Scatterplots of independent variables

3.4.3 Correlation plots of independent variables

To avoid multicollinearity in the model, we conduct the pairwise correlation test between potential explanatory variables. Blue represents a negative correlation between independent variables, while the red box represents a positive relationship. The darker the color, the higher the correlation.

From the correlation plot (Figure 5), it is obvious that the correlation is strong *bewteen log_Dist2BusStop_5 and log_Dist2TrainStation, log_Dist2BikeRack_5 and log_Dist2PedStreet, log_Dist2College and log_Dist2University, and percent_bachelors and log_medHHInc*.

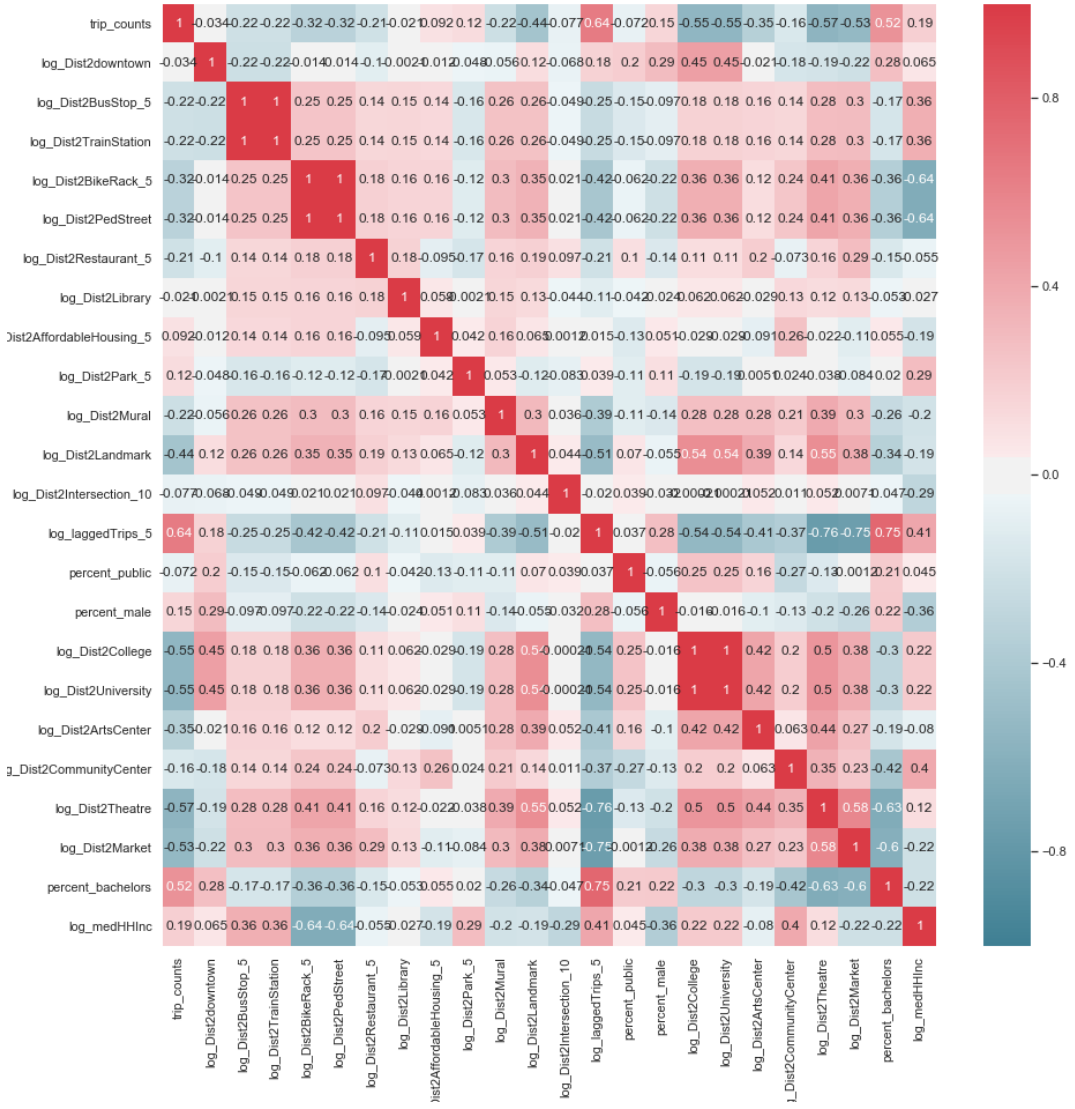


Figure 5. Correlation plot of all numeric independent variables

Taking the correction of independent variables with the dependent variable into consideration, we removed *log_Dist2TrainStation, log_Dist2PedStreet, log_Dist2University, and log_medHHInc* from the model (Figure 6). However, we will still include *log_laggedTrips_5* in our model because it might be inherently correlated with *long_trip_count* due to their inner nature.

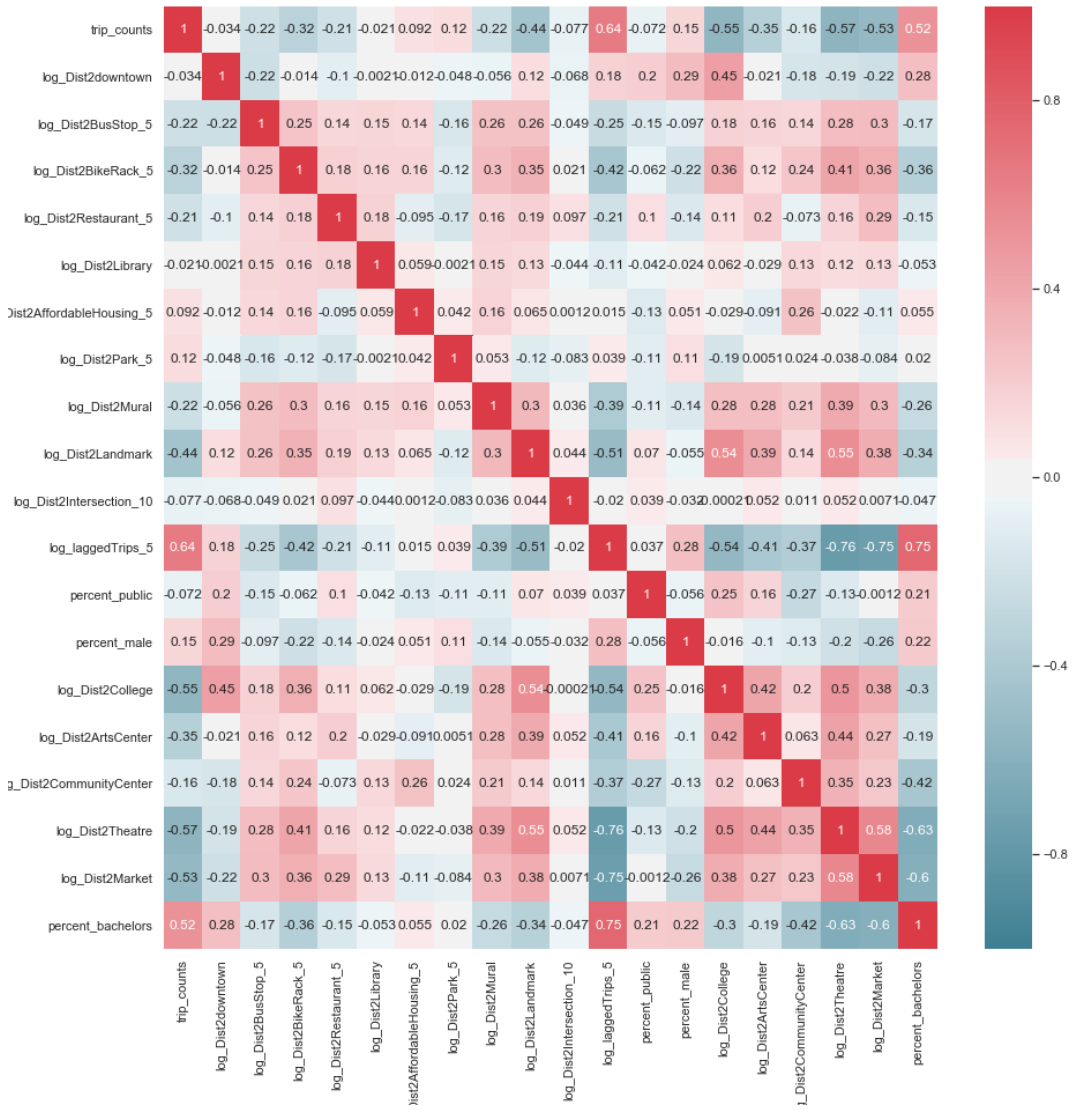


Figure 6. Correlation plot of final used independent variables

4. Model Built on Station Level

4.1 Model Results

The regression model that we build includes one category data and 21 numeric data. When making the pipeline, we will split the datasets into 70% train set and 30% test set, by using 20 estimators to run in a reasonable time, and 3-fold to do the cross-validation.

We get the overall valuation of our best random forest model is 0.81, indicating that 81.0% of the variations in the dependent variable can be explained by our model, which is a relatively well-performing regression model. Therefore, this model could quick effectively help visualize the overall pattern of performance of bike-share service in the city of Chicago due to high R-square. Our model successfully captured the general trend of trip counts within CBD and outside CBD.

Meanwhile, the model does a great job of predicting the trip counts outside the CBD. For the stations with the highest trip counts in CBD, the anticipated results are relatively more accurate, compared to the busiest stations outside CBD.

4.2 Prediction Results

To evaluate the performance level of the regression model, the R-squared is not sufficient. 3-fold cross-validation is used to test the estimate accuracy of the model. The MAE is 2,448, and the MAPE is 150%. The result claims that we should expect an average difference of 2,448 between the predicted and actual trip counts (Figure 7). Nevertheless, many reasons can affect prediction accuracy, such as the total counts of datasets.

Using this model, we evaluate the performance level of our current independent variables. The next step is to use the same indicators for a regression model built on the fishnet of Chicago, which can predict future bike-share demands for areas without an existing station.

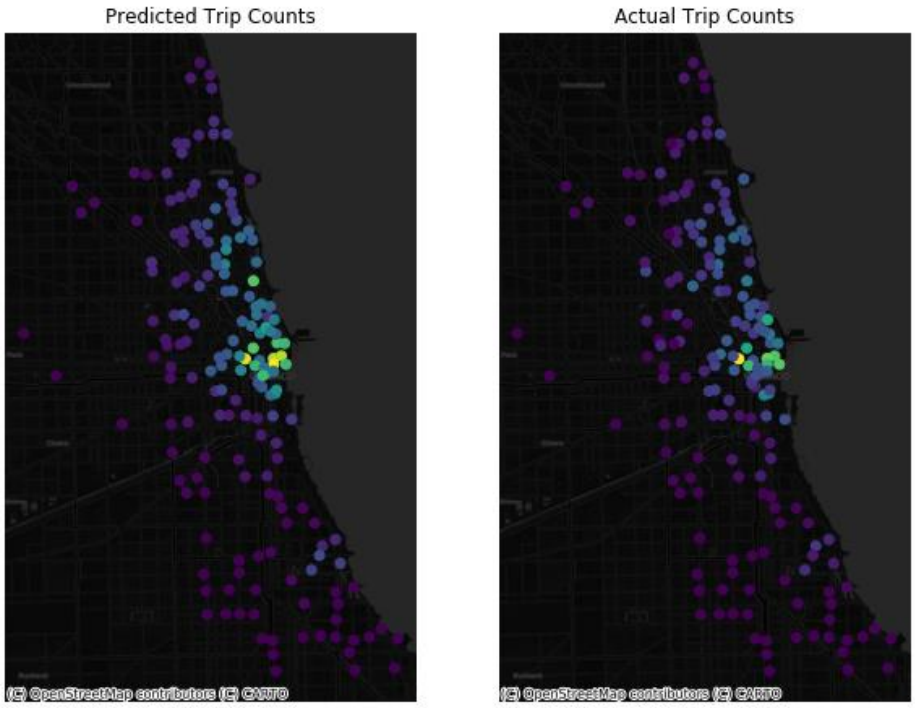


Figure 7. Comparison of predicted and actual trip counts

5. Model Built on Fishnet for Prediction

5.1 Create Fishnet

To propose new bike-share stations, we need to create specific grids covering the entire city of Chicago to hold data with existing stations and without a station.

The principle to create the fishnet is that each grid will only contain one bike-share station at maximum, to help us pick out specific areas to locate new stations. So, the eventual size for each grid cell is 110m*110m.

5.2 Model Preparation

We join all data, such as trip_counts, station datasets, different indicator datasets back to the fishnet by geospatial join.

Then we separate the entire new datasets into two parts: grid with stations as training data set and grid cell without station as predicting data set.

5.3 Predictive Model and K-fold Cross-Validation

The predictive model is built on the training data set, which contains all grid with stations. The data set is split into a training set of 70% and a test set of 30%.

The R-square of this predictive model is 0.8533, indicating that 85.33% of the variations in the dependent variable can be explained by our model, which is a well-performing regression model.

To evaluate the performance of the regression model, before predicting the new stations, we should also test the accuracy of the test data. The result shows that the average absolute error is 2,517, and the MAPE is 109.6%. Even though the accuracy results might not be ideal, our model still make lots of progress compared to the first regression model and can qualitatively represent the demand for bike-share services in a specific area.

6. Result for Predicted Cells without a Station

6.1 Predicted Patterns

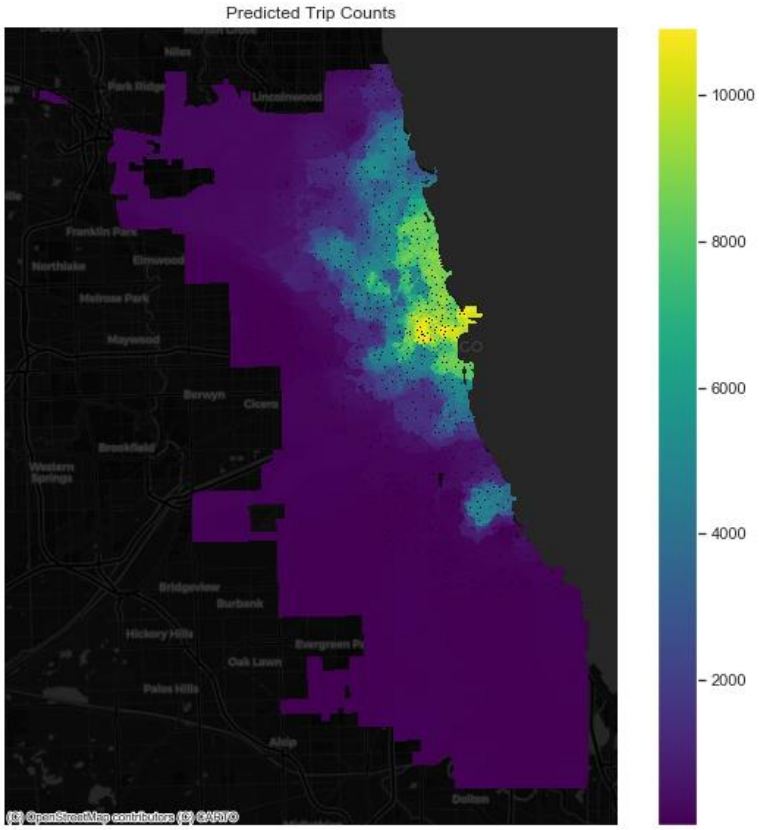


Figure 8. Predicted Divvy stations across the city of Chicago

With the regression model, we predict the bike-share trip counts for each area in the city of Chicago since 2018, under the assumption that the decision-making process in choosing new stations remains the same (Figure 8). Not surprisingly, all the proposed stations are condensed in CBD or just outside of it, where there have the most population density and many amenities nearby. According to the prediction results, a station in the CBD can have more than 6,000 trips. However, the number will gradually decline when the station is further away from the CBD. Stations close to the urban boundary of Chicago have less than 1,000 trips.

6.2 Conclusion

Through the study, we identify the proposed new Divvy bike-share stations across the city of Chicago. The highest demand of bike-share services occur in the CBD and adjacent areas, due to the dense amenities and higher demographic profile. As the spatial lag of the existing bike-share stations turns out to be the most statistically significant indicator in our model, the grid cells near the existing busiest bike-share stations tend to have a much higher demand for bike-share services. Therefore, we recommend the city to keep implementing new Divvy stations in the CBD areas because of the high demands of the bike-share system.

However, the city of Chicago provides bike-share services not only for the profitability but also to promote a more accessible and equitable mode of transit across the whole city, especially in lower-income neighborhoods located near the city's urban boundary. Even though the model shows a relatively low demand of bike-share services in these neighborhoods, there are still need for new provisions of such services for better social well-beings. This is also the reason why the city will propose more Divvy bikes and stations in the far south and west of Chicago.

The model's accuracy is vitally impacted by the current limitations of our study. Firstly, the existing stations are mainly concentrated in the CBD of Chicago, causing the issue that the spatial lag could rarely represent the needs of the bike-share services in neighborhoods far from the CBD. It is hard to use spatial lag to propose new stations that did not exist yet. Secondly, the prediction is made under the assumption that the training set and the test set were similar in terms of distribution. However, since we only have roughly 600 stations as the sample to build our model on, the random split of data set may exhibit a great bias. Thirdly, to promote a more equitable and affordable bike-share system, we shall include and prioritize social-equity indicators in our future models.

To conclude, our model could propose quite well on new stations in CBD and adjacent areas. It is still too early to validate the future stations in further west and south region of the city without any data support. However, the new deal of the city of Chicago will help achieve social equity and render new transportation alternatives in those lower-income communities with high demand in the far west and south parts of the city. Although our model does not do a great job of accurately predicting the bike-share demand near the urban boundary, it could become a useful reference by identifying the most suitable area to build new stations in the far west and south parts of Chicago with further data support after the new deal starts to operate and resources are allocated in these neighborhoods.

---

[i] Greenfield, John. April 2019. Divvy Deal Passes, and Uber Fails to Buy the Future of Chicago Bike-Share. https://chi.streetsblog.org/2019/04/10/divvy-deal-passes-ubers-cash-fails-to-buy-the-future-of-chicago-bike-share/