Assessing the Long Term Impacts of Opening Capital Bikeshare Stations on Traffic Volume

Ibadat Jarg[1] **and Helen Wang**[1]

Capital Bikeshare was first implemented in the metropolitan DMV area in 2010 as an alternative transit option for residents. In our study, we assessed the impact of opening such a station on local traffic volume. Using traffic APIs and station coordinates, we obtained Annual Average Daily Traffic (AADT) from monitors around each Bikeshare station between 2007-2019. Using 2013 as the reference year, we limited our sample to the 78 stations that opened that year and evaluated change in AADT 6 years pre- and post-station opening. Using OLS, we observed no significant effects, however, panel data analysis with entity effects found significant reductions in traffic volume, with opening a station being linked to a 9 perecent reduction in AADT. A two-way fixed effect model with time effects included, on the other hand, suggested that opening a station did not affect traffic volume. As such, the current results of the study are inconclusive and require further investigation. Limitations of the study include the need to control for confounding factors such as population density and public transit infrastructure changes. The scarcity of the granular geolocation and time data needed for such changes, however, remain prominent challenges for future work.

Data Science | Public Transit | Bike Infrastructure | Washington DC

Capital Bikeshare is a bicycle-sharing system first started on September 20, 2010 [4]. Opening with approximately 1,100 bikes across 114 stations, it has since expanded to encompass over 7000+ bikes across 700 stations, with over 27 million trips having been taken since 2010 [4, 5, 2]. As of 2024, Capital Bikeshare stations are available in 8 different jurisdictions, including the District of Columbia, Arlington, Montgomery, Prince George's County, Fairfax, Alexandria, and Falls Church [1]. Since its opening, Capital Bikeshare has touted itself as an affordable, eco-friendly transit alternative–an attractive prospect given that Washington D.C. is reportedly the 55th most congested city in the world and 2nd most congested city in the United States [6]. Whether Capital Bikeshare does affect traffic congestion, however, remains to be seen.

Related Work. Past work on the effect of bicycle-sharing systems has been fairly ambiguous. While some work has suggested that bicycle-sharing systems increase sustainable mobility (reducing greenhouse gas emissions while improving congestion), others have noted that such systems have little impact on small-to-medium cities or in car-dominant territories [3]. In large urban centers with less car usage, bicycle-sharing systems have also been found to have negligible impacts on overall emissions [3]. In the context of Capital Bikeshare specifically, one past study examined the effects of its implementation on traffic and found that it resulted in a 4 percentage reduction in congestion. The study itself, however, was limited to data available between 2010 and 2012. In our study, we expand on past work by exploring the impact of opening a

Author affiliations: [1]Georgetown University Mc-Court School of Public Policy

Authors contributed equally to this work

Capital Bikeshare station on traffic volume between a 12-year period. In assessing trends in traffic volume six years before and after opening, we examine the long term effects that opening a Capital Bikeshare station may have on local traffic volume.

Data

Data Sources. Four data sources were used to acquire data for this project:

Capital Bikeshare Locations' dataset from Open Data DC. This dataset is publicly available on Open Data DC and was originally created as part of the DC Geographic Information System (DC GIS). It contains data on 748 different Capital Bikeshare stations across the DC metro area, providing information on the station locations (including address and coordinates) as well as individual station features, such as the number of docks and bikes available, general capacity, region, etc. (see **Fig 1** for locations of Bikeshare Stations).

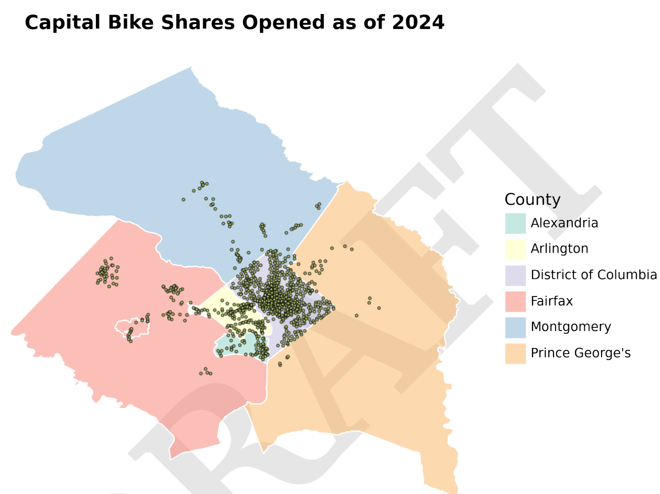

Fig. 1. All the bikeshare stations present as of 2024

Trip History and Expansion Plan Data from the Capital Bikeshare website. The Capital Bikeshare website provides data on all trips taken per year since its opening. For our project, we downloaded data collected on trip history between 2010 and 2020 to deduce the opening year for all active stations. The trip history data contained variables such as the duration of the trip, start data, end state, start station, end station, etc. The data was preprocessed according to the Capital Bikeshare website, with any trips taken by staff or from testing stations removed. Any trips that lasted less than 1 minute were also pre-filtered by Capital Bikeshare to account for false starts or re-docking attempts. In addition, we also downloaded their expansion plan data directly from their website. The dataset included a list and coordinates of potential future stations. While the files were provided in KML format, we converted them to .xlsx for ease of use.

Traffic Counts - Historic AADT by Count Station' dataset from the Regional Transportation Data Clearinghouse. To obtain traffic volume data, we pulled a dataset created by the National Capital Region Transportation Planning Board from the Regional Transportation Data Clearinghouse. The dataset contains Annual Average Daily Traffic (AADT) estimates from counting stations within the District of Columbia, Maryland, Virginia, and West Virginia between the years of 1986 and 2022. Counts after 2006 were based on counts taken at the count station location for all regions. Counts prior to 2006, 2000, and 1997 for DC, Maryland, and Virginia respectively were based on traffic volumes reported from regional sources.

TIGER/Line Shapefiles from the United States Census Bureau. We utilized shape files from the US Census Bureau to get the legal boundaries of the metropolitan DMV area. These files were used to inform all geographic visualizations generated for this project. We specifically utilized the 2024 County and Equivalent national shape files as well as the 2024 Census Tract shape files for the District of Columbia.

Data Acquisition. Capital Bikeshare location and trip history data were directly downloaded from the Open Data website and the Capital Bikeshare website as .geojson and .csv files respectively. Capital Bikeshare station coordinates were then used to query the 'Traffic Counts - Historic AADT by Count Station' dataset API as we were interested in identifying traffic volume around each station. Using the coordinates, we grabbed all traffic volume data from count stations that were within a 500-meter distance of each Capital Bikeshare location.

Data Information . As our main variable of interest was the impact of opening a Capital Bikeshare station, we utilized the number of newly opened stations per year as our main determining factor for our reference year (see **Fig. 2**). 2020 and 2023 reported the highest number of new stations at 95 and 79 stations respectively. However, due to the extenuating circumstances of COVID-19 during those time periods, we elected to choose 2013 (at 78 new stations opened) as our reference year. Similarly, we limited our traffic volume data to 2007-2019, six years before and after 2013, to avoid introducing COVID-19 as a potential confounder in our analysis.

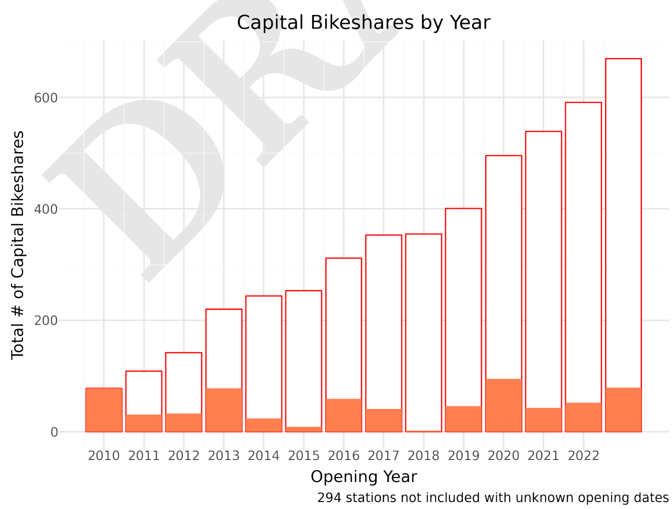

Fig. 2. Showcases growth of Capital Bikeshare stations since opening in 2010

Limitations of the Data . There are some key limitations to our data, namely:

Opening Year for Capital Bikeshare was constructed using trip history. Our raw dataset lacked a variable containing information on the opening year of each bikeshare station. While we were able to deduce the opening years for each station using the trip history data available, there may still be some margin of error not accounted for.

Missing Data for Traffic Volume. Furthermore, we are missing traffic volume data more for certain years than others—namely, we have more missing data between 2007-2009 than other years (see **Fig 3**). This discrepancy may introduce greater variance for certain years.

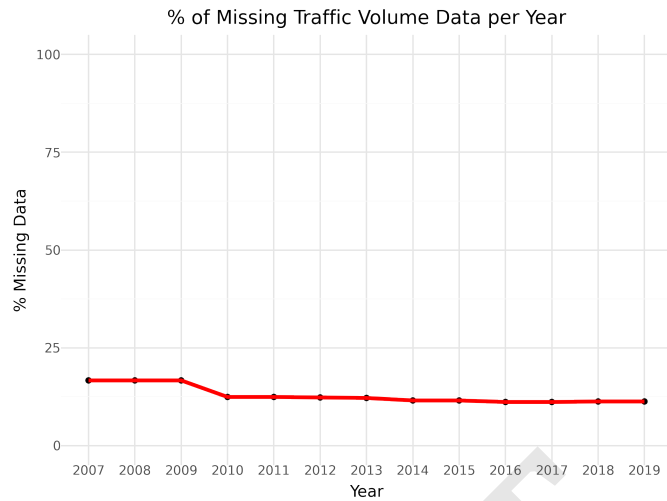

Fig. 3. Represents the proportion of Traffic Volume data missing for each year

Methods

Data Cleaning. Using `01_cleaning_capital_bikeshare.ipynb`, we cleaned the Capital Bikeshare data. To do so, we first obtained the Capital Bikeshare location data through its .geojson file and extracted the data in the key "features" before converting it to a pandas dataframe. We then consolidated all the individual yearly and monthly trip history .csv files into one pandas dataframe using a loop. We cleaned the `Start Date` variable by filling in missing data with `NaN` before grouping the data by `station name` and `Start Date`. We then computed the lowest `Start Date` value for each station to get the approximate opening year. The newly created `Opening Year` column was then merged with the Capital Bikeshare location data using their Station Name.

To obtain the corresponding traffic volume data for each station, we utilized `02_cleaning_traffic.ipynb`. We created two functions, `traffic_query()` and `clean_traffic()`. The first function, `traffic_query()` took in the coordinates for a station and queried the 'Traffic Counts - Historic AADT by Count Station' API for traffic volume. We set the API parameters such that it would return (1) only the AADT variables for 2007-2019 and (2) all the traffic volume data from the count stations that were within 500 meters of the Capital Bikeshare coordinates given as arguments. We chose 500 meters as we believed that would allow us to effectively evaluate the impact that Capital Bikeshare would have on traffic. The `traffic_query()` function would then evaluate if the API connection was successful before returning a .json file. The second function, `clean_traffic()` was utilized to extract important variables from the .json file and return a list of dictionaries.

Using the two functions, we ran a loop through all the capital bikeshare station coordinates. We used `traffic_query()` to query for traffic volume for each station and then `clean_traffic()` to clean the data and append the results on a list. We converted the data into a pandas dataframe before aggregating to get the mean traffic volume for each year. As such, each station was represented as a singular row, with the traffic volume for each year representing the mean traffic volume of that year across all the count stations within 500 meters. This loop was repeated with both all the opened stations as well as proposed stations in the expansion plans. Indicator columns were added to indicate if stations opened in 2013 as well as to indicate if they were proposed or opened stations. Data for opened stations and proposed stations were merged into one dataframe using `03_merging_data.ipynb` before being exported as `final_data.csv`. For most of our analyses, we filtered our dataset to include only stations that opened in 2013, leaving us with a sample of 78 stations. A snapshot of the completed dataframe can be seen in **Fig. 4**

id	long	lat	open_year	name	2007	2008	2009	2010	2011	2012
08249ef2-1f3f-11e7-bf6b-3863bb334450	-77.041779	38.905067	2010.0	18th & M St NW	16554.625000	17207.750000	15743.375000	14952.545455	14570.454545	14663.818182
082544b7-1f3f-11e7-bf6b-3863bb334450	-77.077078	38.943837	2013.0	39th & Veazey St NW	17128.666667	17043.000000	16884.666667	16736.666667	16823.000000	17252.666667
082524a2-1f3f-11e7-bf6b-3863bb334450	-77.086063	38.893237	2020.0	N Veitch St & Key Blvd	14604.333333	14618.333333	14166.666667	12773.333333	12804.666667	12719.666667

Fig. 4. Structure of final cleaned dataframe

Visualizations. Because we had access to geolocation data, we elected to utilize maps and various data visualizations to represent our findings. Using the TIGER/Line Shapefiles, we were able to construct maps of the metropolitan DMV area using the `plotnine` package. Using coordinates, we then plotted the individual Capital Bikeshare stations and utilized different aesthetics to represent different factors, such as changes in traffic volume. Most of our visualizations were created using `05_visualizations.ipynb`. To represent the relationships assessed in our statistical models, we graphed both our raw data and our fitted values by Year vs Traffic Volume. Those visualizations were constructed in `04_exploratory_data_analysis.ipynb`.

Statistical Analysis

Ordinary Least Square (OLS) Models. Using the `statsmodel` package, we ran two separate OLS models, one with data subsetted for before 2013 and one with data subsetting for after 2013. For each subset, we ran a regression with `Year` as the independent variable. For the dependent variable, we used a measure that captured the average percentage change in traffic volume from 2013. We omitted any point that was more than a 50 percentage change to control for the effect of outliers. We utilized the following OLS model for our analysis:

$$\log(TrafficVolume)_i = \beta_0 + \beta_1 Year_i + \epsilon_i$$

Panel Data Analysis: One-way Fixed Effect Models. Because we were examining individual stations across time, we elected to conduct panel data analysis due to its appropriateness for our given data structures. Running a panel analysis would allow us to see the effect of opening a bikeshare on a per unit level (each individual bikeshare station) and across time (from the years 2007-2019). We transformed our data from wide to long-form, using used `Year` and `id` as our indices. We also logged our traffic volume variable to account for differences in baseline traffic volume across stations.

We then elected to run fixed effect models rather than random effects models because we suspected that the unobserved factors (infrastructure, population) in the vicinity of bikeshare locations are correlated with our explanatory variable (station oppening). To conduct the analysis we utilized the `linearmodel` library in python and imported the `Panel.OLS` function. We first ran a one-way fixed effects model with Logged Traffic Volume as the dependent variable and a dummy variable representing whether it was before or after the station opened. We included entity effects in the model to control for variations within stations, resulting in the following:

$$\log{(TrafficVolume)_{it}} = \beta_0 + \beta_1 StationOpen + \upsilon_i + \epsilon_{it}$$

Panel Data Analysis: Two-way Fixed Effect Models. In our second model, we ran two-way fixed effects models using the same dependent and independent variables as before. This time, we controlled for both entity and time effects, allowing us to control for variations both within stations and within different years. Our resulting model can be seen below:

$$\log{(TrafficVolume)_{it}} = \beta_0 + \beta_1 StationOpen + \upsilon_i + \epsilon_i + \epsilon_{it}$$

Results

Descriptive Statistics. We were able to graph the average traffic volume in the DMV across our period of study (see **Fig. 5**). We see a significant drop in 2010, notably the year that BikeShare opened. Since then, the overall traffic levels in the DMV have remained relatively stable.

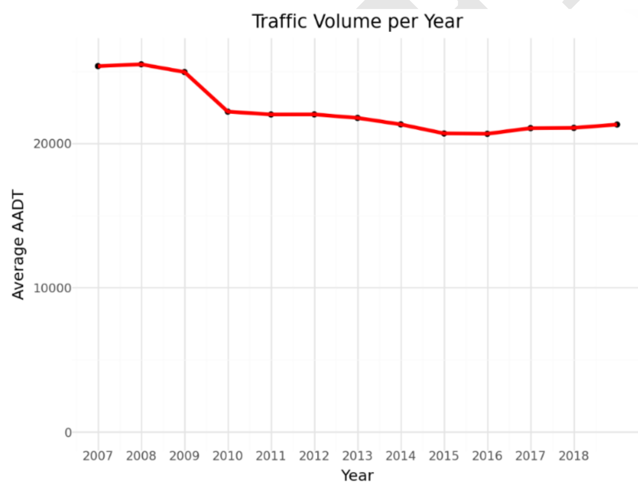

Fig. 5. Change in Traffic Volume across time

We can also observe that the average percentage difference in traffic volume post 2013 is lower than prior to it. We see this in **Fig.6**.

Regression Analysis. Our results focused on statistical analyses that aim to establish the causal effect of opening a Bikeshare location (X) on traffic volume (Y). For our OLS models, the results were not statistically significant at all, with an extremely high p-value. This is not surprising considering we did not control for several important variables such as population density in our models. This omission of important variables makes endogeneity a real concern for our model as well, given the likelihood of omitted variable bias.

The OLS results do not give us any meaningful insights into the relationship between Capital Bikeshare station opening and traffic, and may indicate that there is no effect on traffic when opening a bikeshare in the vicinity.

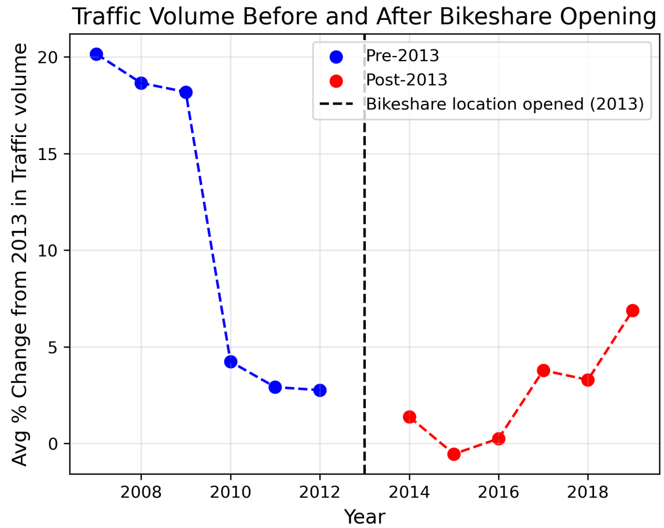

Fig. 6. Percent Change in Traffic Volume across Time

Pre-2013 Results

Dep. Variable:	Percent Change	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.004
Method:	Least Squares	**F-statistic:**	0.09632
Date:	Tue, 17 Dec 2024	**Prob (F-statistic):**	0.757
Time:	14:18:13	**Log-Likelihood:**	-794.16
No. Observations:	209	**AIC:**	1592.
Df Residuals:	207	**BIC:**	1599.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	578.2381	1853.502	0.312	0.755	-3075.924	4232.400
Year	-0.2860	0.922	-0.310	0.757	-2.103	1.531

Omnibus:	12.301	**Durbin-Watson:**	2.148
Prob(Omnibus):	0.002	**Jarque-Bera (JB):**	22.220
Skew:	0.282	**Prob(JB):**	1.50e-05
Kurtosis:	4.494	**Cond. No.**	4.96e+06

Post-2013 Results

Dep. Variable:	Percent Change	**R-squared:**	0.003
Model:	OLS	**Adj. R-squared:**	-0.002
Method:	Least Squares	**F-statistic:**	0.5169
Date:	Tue, 17 Dec 2024	**Prob (F-statistic):**	0.473
Time:	14:18:12	**Log-Likelihood:**	-785.68
No. Observations:	207	**AIC:**	1575.
Df Residuals:	205	**BIC:**	1582.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	1331.8318	1856.072	0.718	0.474	-2327.607	4991.270
Year	-0.6622	0.921	-0.719	0.473	-2.478	1.154

Omnibus:	29.278	**Durbin-Watson:**	1.878
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	144.962
Skew:	-0.315	**Prob(JB):**	3.33e-32
Kurtosis:	7.051	**Cond. No.**	4.97e+06

Panel Data Analysis. Faced with the inconclusive results of our OLS analysis and its lack of robustness, we were prompted to recontextualize our data into the format of panel data. We ran two models. Our first model only did entity level effects and our second one did entity and time level effects, essentially a difference in difference model.

Entity Level Results. We observed a highly statistically significant result in our entity level model model. Our model indicated that opening a bikeshare location on average reduced yearly traffic volume by approximately 9 percent.

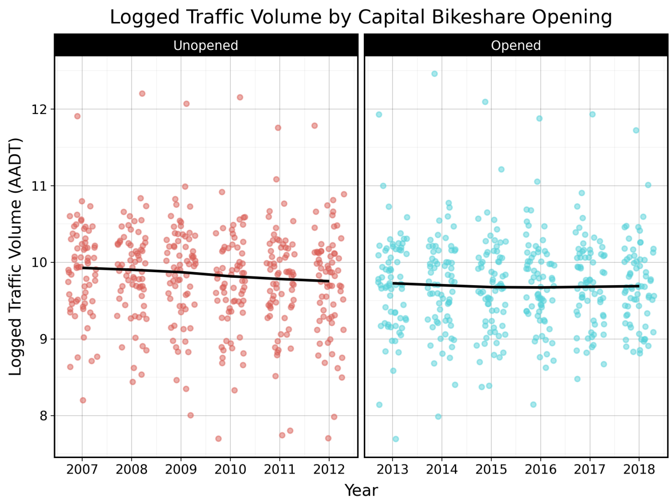

Fig. 7. Logged Change in AADT over Time

However, it is unlikely that this model is particularly robust either. To compensate for any time trends, we also used an entity and time level approach in the fixed effects model.

Fixed Effects - Entity level

Dep. Variable:	log_traffic_volume	**R-squared:**	0.0771
Estimator:	PanelOLS	**R-squared (Between):**	-0.0095
No. Observations:	837	**R-squared (Within):**	0.0771
Date:	Tue, Dec 17 2024	**R-squared (Overall):**	-0.0095
Time:	14:14:16	**Log-likelihood**	343.07
Cov. Estimator:	Unadjusted		
		F-statistic:	63.965
Entities:	70	**P-value**	0.0000
Avg Obs:	11.957	**Distribution:**	F(1,766)
Min Obs:	9.0000		
Max Obs:	12.000	**F-statistic (robust):**	63.965
		P-value	0.0000
Time periods:	12	**Distribution:**	F(1,766)
Avg Obs:	69.750		
Min Obs:	69.000		
Max Obs:	70.000		

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
opened	-0.0929	0.0116	-7.9978	0.0000	-0.1157	-0.0701

F-test for Poolability: 144.36
P-value: 0.0000
Distribution: F(69,766)

Included effects: Entity

Entity and Time level results . When we re-ran the model to compensate for time level variation, we see the following results:

Fixed Effects - Entity and Time level

Dep. Variable:	log_traffic_volume	**R-squared:**	2.744e-08
Estimator:	PanelOLS	**R-squared (Between):**	-2.866e+17
No. Observations:	837	**R-squared (Within):**	-9.674e+20
Date:	Tue, Dec 17 2024	**R-squared (Overall):**	-5.683e+17
Time:	14:14:16	**Log-likelihood**	373.62
Cov. Estimator:	Unadjusted		
		F-statistic:	2.072e-05
Entities:	70	**P-value**	0.9964
Avg Obs:	11.957	**Distribution:**	F(1,755)
Min Obs:	9.0000		
Max Obs:	12.000	**F-statistic (robust):**	2.072e-05
		P-value	0.9964
Time periods:	12	**Distribution:**	F(1,755)
Avg Obs:	69.750		
Min Obs:	69.000		
Max Obs:	70.000		

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
opened	-1.041e+10	2.286e+12	-0.0046	0.9964	-4.498e+12	4.477e+12

F-test for Poolability: 132.73
P-value: 0.0000
Distribution: F(80,755)
Included effects: Entity, Time

Results gave us a coefficient value that is effectively 0, with an extremely high p-value, making the results unusable. This result once again highlights the weakness in our dataset, being unable to control for several important variables.

Discussion and Conclusions

The results of our study have been largely inconclusive but it marks a useful starting point for future public infrastructure analysis. Our statistical results were not robust or particularly insightful and thus do not give us much causal explanatory power.

Yet our descriptive statistics show us that the years that Bikeshare opened actually was the lowest level of congestion in our time period of study. This would indicate that opening a bikeshare station can at least reduce traffic in the year that it is opened. This can be inferred as a large amount of initial enthusiasm for bikeshare adoption, which slowly tapers off as time goes on.

We believe that Capital Bikeshare must expand further, not only to reduce the high levels of congestion that the DMV can have, but also as a public good to avoid car dependence.

Policy Proposals. While our results are inconclusive at best, from our descriptive statistics we see that volume does tend to decrease when a bikeshare opens in that given year. Thus we feel that if there is any insight that can be drawn from this study, it is that stations should be built in areas that have high levels of traffic volume.

Capital Bikeshare shares locations of proposed future bikeshare locations on their website and we extracted the coordinates and traffic volume in 2019 for all the proposed locations. We recommend that the proposed stations with the highest levels of traffic volume should be prioritized in future DC infrastructure planning. We constructed a map of the highest traffic density locations in DC to better visualize the dispersion of these proposed locations (see **Fig 8**).

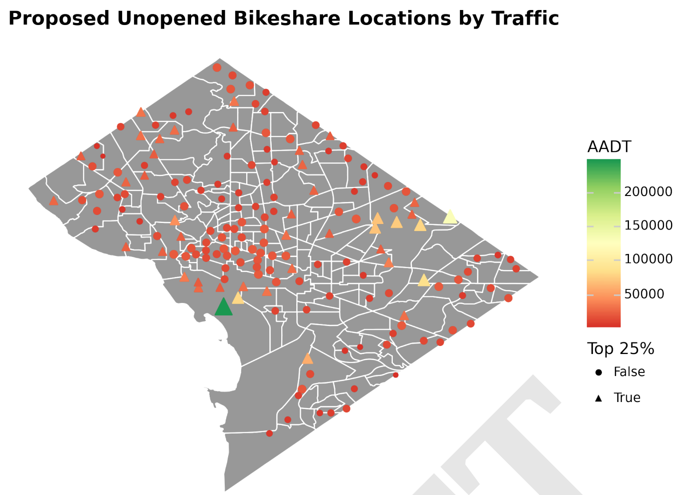

Fig. 8. Prospective Capital Bikeshare Stations by Traffic Volume

Limitations and Further Research. The regression results do not give us much to hint at any type of causal relationship between a bikeshare location opening and traffic volume. To make our regression robust, we will need population data geotagged across our time period of our study. Another feasible data point for the dataset would be to include the number of trips that took place in a given year at a certain bikeshare location.

Further research can also look at the other impacts of bikeshare on the local vicinity. We wanted to use pollution data but currently there exists no such data at the level of granularity that we require. Other variables such housing prices, mobility metrics etc. can be measured.

References

[1] *About Company & History*. en. URL: http://ride.capitalbikeshare.com/about (visited on 11/15/2024).

[2] *Capital Bikeshare — ddot*. URL: https://ddot.dc.gov/page/capital-bikeshare (visited on 12/17/2024).

[3] Cyrille Médard de Chardon. “The contradictions of bike-share benefits, purposes and outcomes”. In: *Transportation Research Part A: Policy and Practice* 121 (Mar. 2019), pp. 401–419. ISSN: 0965-8564. DOI: 10.1016/j.tra.2019.01.031. URL: https://www.sciencedirect.com/science/article/pii/S0965856417316099 (visited on 12/17/2024).

[4] *Press Kit*. en. URL: http://ride.capitalbikeshare.com/press-kit (visited on 12/17/2024).

[5] *System Data — Capital Bikeshare*. en-US. URL: https://capitalbikeshare.com/system-data (visited on 12/16/2024).

[6] *Washington traffic report — TomTom Traffic Index*. URL: https://www.tomtom.com/traffic-index/washington-traffic/ (visited on 12/17/2024).