**Estimation of Bike Ridership for Lexington Kentucky**

CE 599

Spring 2018

Daniel Francis

Submitted: 5/2/2018

Word Count: 5032 words text + 3 figures\*250 + 3 tables\*250 = 6,532

**Table of Contents:**

[**Abstract** 1](#_Toc512974306)

[**Introduction** 1](#_Toc512974307)

[**Literature Review** 1](#_Toc512974308)

[**Description of Study Area** 2](#_Toc512974309)

[**Methodology** 3](#_Toc512974310)

[**Data Acquired** 3](#_Toc512974311)

[**Geographic Grid Made** 5](#_Toc512974312)

[**Data Applied to Grid** 6](#_Toc512974313)

[LEHD data applied to points 6](#_Toc512974314)

[ACS data merged with block data for Fayette County 6](#_Toc512974315)

[Points with LEHD data merged with Blocks with ACS data 7](#_Toc512974316)

[**Model Applied to Grid** 7](#_Toc512974317)

[**Heat Map Made from Geographic Grid** 8](#_Toc512974318)

[**Results and Next Steps** 10](#_Toc512974319)

[**References** 12](#_Toc512974320)

**Table of Figures:**

[Figure 1: Area of Lexington Model was applied to 2](#_Toc512974385)

[Figure 2: Created Grid Network to be overlain on Lexington 5](#_Toc512974386)

[Figure 3: Heat map of Estimated Ridership for Lexington 9](#_Toc512974387)

**Table of Tables:**

[Table 1: ACS Data Used 4](#_Toc512974396)

[Table 2: LEHD Data Used 4](#_Toc512974397)

[Table 3: Data and Coefficients used with Model 8](#_Toc512974398)

# Abstract

This paper uses a Bike Ridership Estimation model in order to estimate the bike ridership in Lexington, Kentucky as well as to determine the best areas in which to install bike sharing stations. The model used for this estimation was created based on three other cities that are similar in size and demographic makeup to Lexington. These 3 cities also already have bike sharing implemented, so the goal of the model is to take the characteristics of the bike sharing in those cities, and apply it to Lexington. The main output of this process, as of the writing of this paper is a heat map of Lexington. This heat map shows “hot spots”, which are areas where the model predicts the most ridership will occur which indicates that these areas would be good candidates for a bike station. This heat map is intended to be used as a tool for the implementation of bike sharing stations due to the fact that the city of Lexington is currently investigating using bike sharing, but is unsure of where to install the stations. It is thought that the heat map created here will be the answer to the question of “where should the stations go?”

# Introduction

Bike sharing is not a new trend, but has instead been around since the 1960s. In 1965, a group of activists in Amsterdam introduced the “Witte Fietsen”, or White Bikes, which were bikes painted white and which were left unlocked around the city for anybody to use (*2*). Then in 1995, Copenhagen introduced Copenhagen’s Bycykeln, or City Bikes, which allowed people to access sturdy bikes at specific locations around the city via coin-operated systems (*2*). From then on, bike sharing evolved into what is currently, which is an efficient means for tourists/citizens to get around a city without contributing to vehicular congestion through the use of rented bikes. A multitude of large cities (New York City, Washington DC, etc.) already have bike sharing implemented, with a number of similar sized cities and smaller cities following suit. However, for cities that don’t already have bike sharing implemented, the question of where to place the bike stations (if it is decided to utilize bike sharing) becomes problematic. This is because for a bike sharing station to actually be efficient, people need to actually use it. So when it comes to determining where to actually place the station, characteristics such as proximity to recreational activities, proximity to high employment areas, age of the populous, etc. need to be taken into account in order to better determine where to place the stations. As long as these characteristics are taken into account, and modeled properly, then it is very easy to determine where the stations should be placed. That being said, that is the main goal of this paper and the research that accompanies this paper. As discussed previously, Lexington, Kentucky is interested in using bike sharing stations, but the question of where to place the stations is hard to answer as of now. Because of this, it is hoped that the estimated model and the produced outputs from the model (primarily the heat map of Lexington) will be beneficial tools that will aid in the answering of this question.

# Literature Review

The research within this paper builds upon a correlated paper, which the model used in this research comes from. The correlated paper, “Bikeshare Ridership Model for Station-Level Forecasting” (*1*), looked at three cities that already have bike sharing implemented, and that are similar in size and sociodemographic make up to Lexington, and estimated a model based on ridership characteristics within these cities. These three cities are as follow:

* Chattanooga, Tennessee
* Boulder, Colorado
* Columbus, Ohio

The ridership model estimation from these three cities is what created the model that was used to carry out the bike ridership estimation for Lexington, which will be discussed in detail in the following paper.

# Description of Study Area

As stated previously, the area where the estimated model will be applied is Lexington, KY. Lexington lies within Fayette County, Kentucky, but as of the writing of this paper, only the city will have the model applied to it. The population of Lexington, as of the writing of this paper, is approximately 318,449 people (*3*) with approximately 49% of the population being 34 years of age or younger (which is simply an age range that is thought to utilize riding bikes as means of transportation the most) (*4*). In addition to the population characteristics, Lexington currently has around 65 miles of bike lanes within the city (*5*) which indicates that the city is a bike friendly city and bike sharing would most likely be successful (assuming the bike stations are placed in the appropriate areas). It is thought that future projects may look at the entire county to see bike ridership estimates, but for now the city is the primary focus. The area of Lexington, shown in Figure 1 (below), is the area in which the model was applied and which the heat map will be applied to.

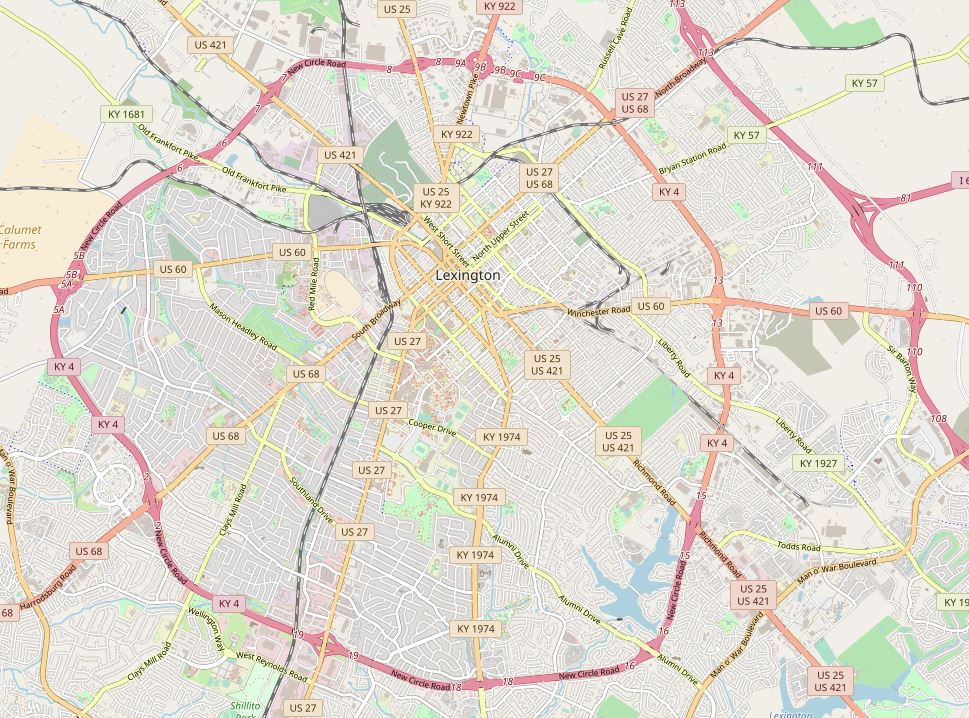


Figure 1: Area of Lexington Model was applied to

The modeled area includes the University of Kentucky, Fayette Mall, The Summit, Hamburg Pavilion, a large amount of residential areas, and a large amount of areas with high employment. The point of making sure all of these areas were included in the study area was to attempt to provide the model with enough varying characteristics in order to actually properly model the ridership that would be seen in Lexington. With these characteristics known and the study area known, the modeling process was able to be carried out, which will be discussed in depth in the following sections.

# Methodology

The following is simply an all-encompassing discussion of the methodology used during this research, but each step will be discussed in further detail in the following sections. The first step in the modeling process was to acquire the same sets of data that were used during the model estimation and creation process. For this, data from the American Census Survey (ACS), Longitudinal Employer Household Dynamics (LEHD), and yearly weather were pulled in. In addition to these datasets, the population density of Lexington was also determined (using census block data and population data) in case that data was required for the model estimation of Lexington. With the same datasets used in the model creation attained for Lexington, the model could then be applied to Lexington. However, the first step was to create a means of actually seeing where the best places to place a station would be. To do this, a geographic grid of points was created and overlain on Lexington which would then be used to create a heat map of ridership. With the geographic grid made, the data that was already acquired was then applied to each of the points in order for the points to be able to represent the ridership in order to be used as a heat map. With the points now representing the data as well as the estimated ridership, a heat map of the points was then created. This heat map showed the ridership estimated by the model by representing areas with more ridership as a more intense “heat” and areas with less ridership as a less intense “heat”. The heat map was then investigated to determine if the highest ridership areas made sense and was also investigated to see why areas which were thought to have been high ridership were represented as low ridership areas. After the investigation into the heat map was concluded, the model was then looked at to see how it could be improved and how future research in this area could benefit from an improved model. As discussed at the beginning of this section, each of the mentioned steps will now be further expound upon in an attempt to provide a much more in depth understanding of what actually went into each step and what came out of each step.

## **Data Acquired**

As stated previously, numerous sets of data were acquired for the model estimation of Lexington to be carried out. The first set of data was the census block groups for the state of Kentucky that would be used in the assigning of data to the geographic grid (which will be discussed in later sections). The second set of data that was acquired was 2015 5-year estimates ACS demographic data by census block for Fayette County, Kentucky. The following table (Table 1), shows which sets of data were taken from ACS as well as which attributes within each data were used in this process:

|  |  |  |
| --- | --- | --- |
| ACS Table # | Data | Attribute(s) Used |
| B01003 | Total Population | Total Estimate |
| B19013 | Median Household Income in past 12 Months | Total Estimate |
| B25001 | Number of Household Units | Total Estimate |
| B15012 | Population 25 years or over with Bachelor's Degree | Total Estimate |
| B15003 | Educational Attainment for Population over 25 | Estimate of Master's, Professional, or Doctorate degrees |
| B08301 | Means of Transportation to Work | Estimate of People that Use Transit, Walking, or Biking to get to work |

Table 1: ACS Data Used

The third set of data that was attained was employment factors from LEHD Origin-Destination Employment Statistics (LODES) data for the state of Kentucky. As above, the following table (Table 2) shows which sets of data were taken from LEHD as well as whether or not the data pertained to Workplace Area Characteristics (WAC) or Residence Area Characteristics (RAC):

|  |  |
| --- | --- |
| Data Acquired | WAC or RAC |
| Total Number of Jobs | RAC |
| Number of Jobs in Healthcare and Education | WAC |
| Number of Jobs in Recreation and Accommodation | WAC |
| Number of Jobs in Retail Trade | WAC |

Table 2: LEHD Data Used

As well as the above data, the population density for Fayette County was calculated using the total population data from ACS and the data from the census blocks for land area. This data was used as a means to investigate whether or not the population density of a certain area actually had a large impact on the bike ridership estimates and will be discussed in the “Data Applied to Grid” section. Finally, the last bit of data that was acquired was meteorology data for Fayette County (average rainfall and average temperature) which was used as a means to determine whether or not the weather any effects on the bike ridership estimates. As with the population density, the way which this data was used with the model will be discussed more in depth in the “Data Applied to Grid” section.

It is worth mentioning that not all of the aforementioned data was used in the model estimation of ridership for Lexington in the end. The reasons for the selection of which data was used will not be discussed in depth in this paper because there is a correlated paper that discusses the actual model creation for the model that was used on Lexington. In the correlated paper, the model estimation steps, the data used, and the actual model creation will be discussed in depth. Because of this, it is recommended that parties interested in the model creation steps of this research be directed to that correlated paper. In terms of how the aforementioned data will be used with this paper, the ways in which the selected data for the model estimation was used will be discussed in the following sections (primarily the “Data Applied to Grid” section).

## **Geographic Grid Made**

As mentioned previously, a geographic grid of points was overlain on Lexington in order to create the final product (a heat map of estimated ridership in Lexington). The sole purpose of this section is to provide an overview of the assumptions made and the steps taken while creating this geographic grid. The area of Lexington shown in Figure 1 is the area that the grid was created for. Since that areal map was taken from OpenStreetMap (*6*), it also came with associated latitudes and longitudes for the “squared” view of Lexington (i.e. the 2 lats and 2 longs that created the square). These latitudes and longitudes were what were used to create all of the unique points within the actual grid. An arbitrary value of 1000 feet was assigned as the spacing between each of these points (since it is assumed most people are willing to walk ¼ mile for a bike station, the 1000 foot spacing is well within that walking distance). This 1000 foot spacing was then converted to decimal degrees in order to actually be used with the lats and longs and to create the properly spaced points on a map. The conversion from feet to decimal degrees was done using the knowledge that 1 degree is approximately equal to 69 miles and was used as follows:

With the spacing between points known (), the points could actually be created. Using the discussed spacing, a grid containing 2,541 points was created, which is shown in the following figure (Figure 2):

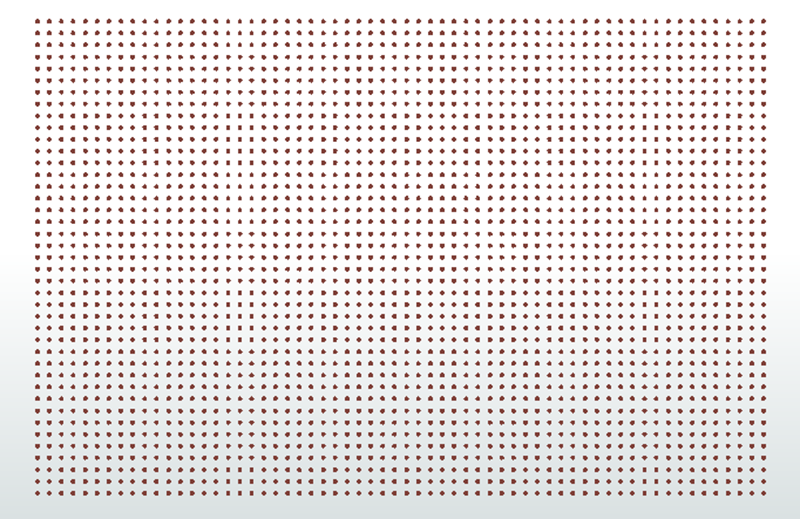


Figure 2: Created Grid Network to be overlain on Lexington

It is worth mentioning that at this point in the research, the above grid has no actual bearing on the ridership estimates that will be attained from applying the actual model to the city. This is because the above points do not have any data associated with them, but are simply just geographic points on a map. In order to make the above grid actual be ready to be used with the model estimation, the acquired data (discussed previously) has to actual be applied to each of the points. The ways in which the data was applied to each of the points will be discussed in detail in the following section.

## **Data Applied to Grid**

As mentioned in the previous section, the newly created geographic grid (at this point in the research) was simply that, a geographic grid. For it to be useful for the modeling process, and the determination of bike sharing station locations, it needed to have data for the area associated with it. The ways in which the data was associated with the grid are as follows:

### LEHD data applied to points

The LEHD data, which pertains to employment factors in the area, was the first thing to be applied to the points. The way in which this was done was by first importing the newly created points into the program “Python” and creating buffers around each of the points. The buffers were then set to be half of the spacing of the points (so in this case 500 feet). The reasoning behind the size of the buffers is because each point wouldn’t pull in people from the entire 1000 foot spacing, but would instead pull in people up to half the distance to the next point. Since all the points within the grid had an equal spacing (1000 feet), the 500 foot buffers were assumed to be properly representative and effective. With the spacing of the buffers established, the points and buffers could then be linked with the LEHD data. The way in which the points/buffers interacted with the LEHD data was in “cookie cutter” type sense. What this means is that wherever the LEHD data was “within” the buffer, that data was applied to the point for that buffer and so on for all of the points within the area. Any LEHD data that was outside of the buffers was disregarded because it was thought to have negligible effect on that point. By doing this process, each of the 2,541 points within the grid was assigned unique sets of LEHD data which in turn made it so each of those points would be able to be representative of different aspects of ridership estimates. With the LEHD data now applied to each of the points, it was then time to move on to applying the ACS data to the points, which took a different technique to get done.

### ACS data merged with block data for Fayette County

As with the LEHD data, the ACS data needed to also be applied to the points in order to be used to turn the points into a tool for the determination of bike sharing station locations. However, the ACS data could not be applied directly to the points as the LEHD data was. The ACS data had to instead be applied to the census block groups for Fayette County and was then applied to the points. The way in which this was done was the ACS data for the state of Kentucky was filtered down to only represent Fayette County (the code for Fayette County in the KY ACS data is 067) and then each dataset was filtered down to represent only the data used in Table 1. With the data filtered to represent only Fayette County and only the data desired, it was then associated with the census block groups. As with the LEHD data, the ACS data and the census blocks were pulled into Python and were then linked together on the grounds of similar Geographic Identifiers (GEOID in the block data and GEO.id2 in the ACS data). With the datasets being linked together on the grounds of similar Geographic Identifiers, it ensured that the proper data from ACS was applied to the proper areas within Lexington, which is vital in ensuring an accurate model is produced through the use of the data. Now with the ACS data linked to the block data, this new set of data could be applied to the LEHD/points dataset previously made.

### Points with LEHD data merged with Blocks with ACS data

With the LEHD data applied to the points and the ACS data applied to the census block groups, these two datasets could then be linked together in order to create the final dataset that the model was applied to. The way in which the datasets were linked, however, is somewhat different than how the other two sets were linked. As with the other data linking, these datasets were pulled into python, but instead of directly linking the data, a spatial join was instead carried out. The way in which a spatial join works is that it looks at points within polygons and either applies the data to the points within the polygons or to the points intersecting the polygons. In this case, the LEHD/point data acted as the points and the ACS/Block data acted as the polygons. So, the spatial join applied the ACS data to whatever points fell within a block (i.e. a polygon) with there being differing blocks for different areas of Fayette County. In doing this, it ensured that the points within the grid actually had the proper ACS data applied to it for where that point lied within Lexington.

Now, with all of the data acquired linked together and applied to the points within the geographic grid, the model could then be applied to the grid (i.e. the data within the grid) and the ridership estimates for the area could be attained. The application of the model to the grid network is discussed in the following section.

## **Model Applied to Grid**

With all of the acquired data being applied to the points, the model could then be applied to the points and the ridership at each of the points could be estimated (which is the overall goal of this research). As stated in the beginning of this paper, the creation of the model will not be discussed within this paper, but the coefficients from that model creation will be what are used for the model estimation for this research. Once again, the reader is encouraged to read the correlated paper, in which the model creation is discussed and explained, in order to gain a better understanding of the values used in this research and to understand where the model used here actually originated. With that being said, the following table (Table 3) shows which data was used in the ridership estimation as well as the associated coefficients from the model estimation:

|  |  |
| --- | --- |
| Term | Estimate |
| Intercept |  |
| Average Rain |  |
| Average Temp |  |
| Number of Jobs in Healthcare and Education |  |
| Total Number of Jobs |  |
| Number of Jobs in Recreation and Accommodation |  |
| Number of Jobs in Retail Trade |  |
| Total Population |  |
| Educational Attainment for Population over 25 |  |
| Means of Transportation to Work |  |
| Average Rain\*(Means of Transportation to Work) |  |
| Median Household Income in past 12 Months |  |
| Median Income\*(Means of Transportation to Work) |  |

Table 3: Data and Coefficients used with Model

With respect to the data used with the model, obviously not all of the acquired data was used. This is not because the data was ignored per se, but instead this was the data that was found to have the largest effects on the 3 cities that the model was estimated for. However, having the rest of the data that was not used with the above model ready to go is not a bad thing. If it is decided that the model needs to be “tinkered” with, then that means that the extra data is already ready to go and can be very easily utilized with a new run of the model if that is desired. So, in the end it is specified that not all data discussed in this paper was used in the ridership estimation that was carried out, but the data not used could possibly be used in future modeling attempts.

The outputs from the model estimation are too large to include within this (i.e. there are 2541 estimates, which is far too many to include in a paper). However, the maximum monthly ridership, minimum monthly ridership, and average monthly ridership will be given to give an idea of what the model estimated in terms of ridership for the area:

It is worth noting that these values (aside from the average) are not for the area as a whole, but are indicative solely of points within the area. So, some of the points within the area (i.e. the grid) may have ridership nearer to the max while some may have ridership nearer to the minimum. This difference in ridership estimation will be what is used to create the heat map for the area and will in turn be what is used to create a tool to say where the best locations for the bike sharing stations are. The heat map and the discussion of where to put the bike stations (in accordance with this model run) will be discussed in the following section.

## **Heat Map Made from Geographic Grid**

As discussed numerous times, the model outputs from the ridership estimation was used, along with the geographic grid over Lexington, to create a heat map of ridership estimates to be used as a tool for determining bike share station locations. Before the actual heat map is discussed, the way in which the heat map was created will be discussed to provide a background for where the heat map actually came from. It has been discussed that each point within the geographic grid had datasets applied to it in order to carry out the ridership estimation at each and every point in the grid over Lexington. With each point having data applied, the model was then ran for each point so that each point would end up having its own unique ridership estimate. The estimates at each point are what were used to create the heat map. Once the points had ridership estimates, the grid was pulled into arcmap and was again overlain over Lexington, with the ridership estimates being in the attribute table for the grid. Once the grid was overlain, arcmap was used apply a color scale to each of the points, with red being low ridership estimates (i.e. the 1.23 riders per month) and green being high ridership estimates (i.e. the 215.93 riders per month). The color then scales from red to green throughout the map to show riderships that are lower or higher than the max or min ridership. The created heat map from this model estimation can be seen in the following figure (Figure 3):

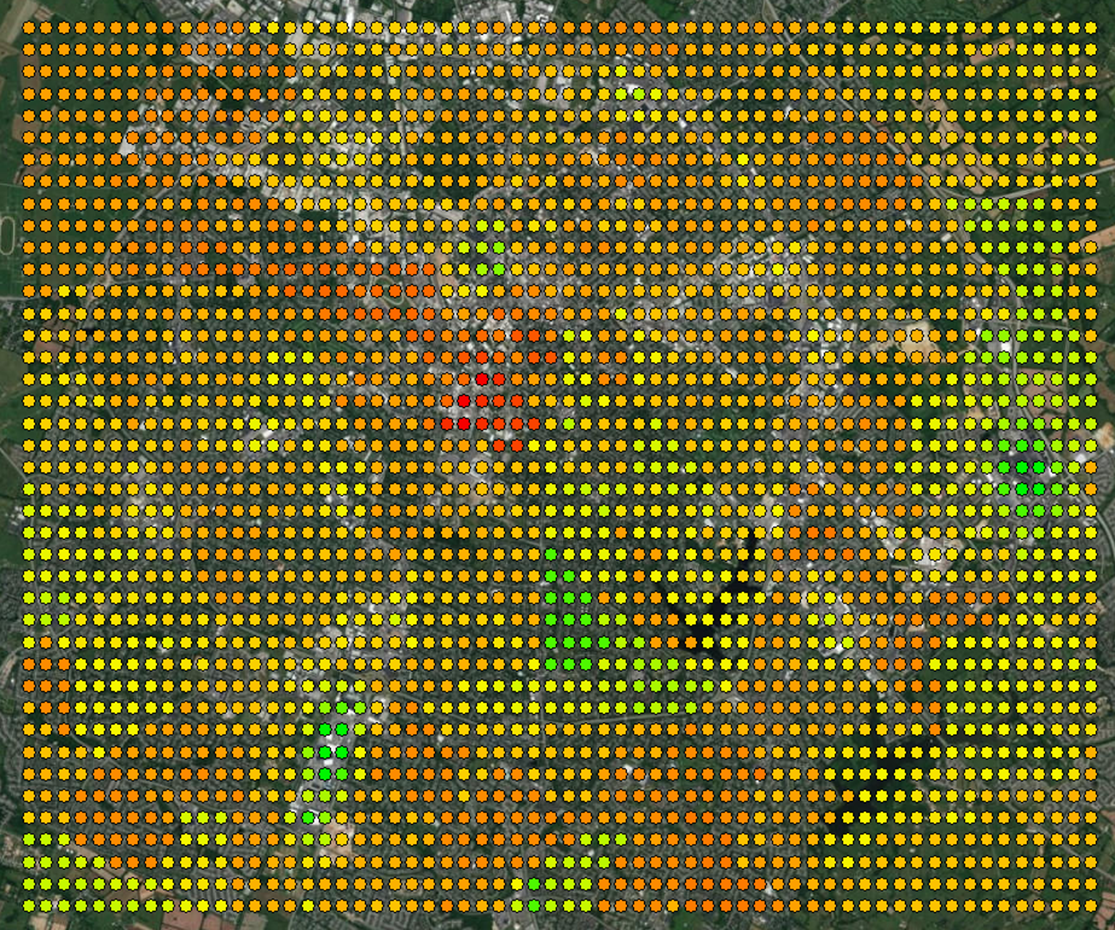


Figure 3: Heat map of Estimated Ridership for Lexington

As stated previously, the green dots in the above heat map represent higher ridership estimates, the red dots represent lower ridership estimates, and the rest of the dots (from orange to yellow) represent the midrange of ridership with orange being lower and yellow being higher.

The purpose behind the creation of this heat map is to have a tool that can be used in the determination of the best locations for bike share stations to be placed. In the terms of this heat map, the following locations are most likely the best areas for a bike share station, or stations, to be placed (with respect to this model estimate):

* Fayette Mall/The Summit/Shilito Park Area (bottom left grouping of green dots)
* Tate’s Creek Area (middle grouping of green dots)
* Hamburg Mall Area (right grouping of green dots)

There is one area of the heat map that is shown to have low ridership, but it is thought that this area should actually have a high ridership, if not the highest of the area. This area is that of the University of Kentucky and Downtown Lexington. It is thought that this area should have a much higher ridership than it does due to the fact that the University of Kentucky is larger and bike sharing would most likely be utilized by the students and because there is a lot of recreational activities, which typically go hand in hand with bike ridership. Because of this, it is thought that something is being misrepresented (or over represented) in the model and will most likely require some further investigation, but that investigation will be discussed in the “Results” section of this paper. All in all, this heat map serves as a very good proof of concept for the creation of a tool to aid in the determination of bike share station locations. It obviously needs to be refined, but the fact that it can make it much easier (with a proper model) to determine the locations will most likely make it a useful tool in the future.

# Results and Next Steps

The overall goal of this research was to utilize the model from the correlated paper to estimate bike ridership for Lexington and to create a tool that utilizes those estimates to provide a means of determining where to place bike stations. From the model estimation for Lexington, the following information about ridership was attained:

As discussed, the model estimation was too large to include within this paper, so the three previous pieces of information about the estimates were given in order to give the reader some idea about the overall estimation. However, the heat map gives a better idea of the ridership due to the fact that it is easy to tell which areas have higher ridership and which have lower ridership (with green being high ridership, red being low ridership, and the scale from orange to yellow being lower to higher ridership). So, overall, the goal for this research has been achieved since the heat map was created and it can be used to determine where to put bike stations (with respect to this model run). Also in respect to this model estimation, the heat map points to the following areas as areas to place bike stations:

* Fayette Mall/The Summit/Shilito Park Area
* Tate’s Creek Area
* Hamburg Mall Area

These areas make sense in terms of ridership, however, there are some inconsistencies with the model that need to be addressed. The model, and thus the heat map, is representing areas such as the University of Kentucky and Downtown Lexington (areas expected to have high ridership) as being low ridership areas. This most likely means that the model is simply under or over representing characteristics of the area, or there is some missing data that would most likely refine these problem areas. These problem areas will be refined in future runs of the model which should then end up with these areas being represented as high ridership (as they are thought to be).

In addition to refining the data already in the model, it is thought that another set of data could be added to the dataset used by the model. This set of data pertains to the proximity of the points within the grid to bike lanes within Lexington. As stated, there are approximately 65 miles of bike lanes in the city, which means that a number of the points within the grid will actually be close to these bike lanes. The proximity of the points to the bike lanes simply means that it would most likely result in a higher ridership for that point due to ease of mobility for riding the bike around in the bike lanes (rather than on sidewalks or in vehicular lanes). As with the refining of the data already in the model, this data will be looked at in future runs of the model to see if it has any bearing on the ridership of the area (which the assumption as of now is that it will).

In summation, the purpose of this research was to apply the estimated model from the correlated paper and to create a tool to be used to determine station locations. That purpose was achieved, with some obvious inconsistencies and problems. It is because of this that this research will be returned to in the future and the model will be “tinkered” with to see how the model estimations vary by model run. However, the proof of concept has been shown by this research: The model can be ran, ridership can be attained, and a tool for determining where to place stations (the heat map) can be made.

# References

1. Sagar, S. (2018). “Bikeshare Ridership Model for Station-Level Forecasting”
2. *The Bike-Share Boom.* City Lab.

<https://www.citylab.com/city-makers-connections/bike-share/>

1. *US Census Bureau.* Census Bureau.

< https://www.census.gov/ >

1. *Demographics.* Commerce Lexington. <http://www.commercelexington.com/demographics.html>
2. *Lexington adding 20 miles of bike lanes by end of 2015.* Lexington Herald Leader. 2015. <http://www.kentucky.com/news/local/counties/fayette-county/article44603070.html>
3. *Map of Lexington.* Open Street Map

<https://www.openstreetmap.org>