Divvy Bike Sharing
Chiratidzo Sanyika

Introduction

Divvy is a bike-sharing system that operates across Chicago and Evanston. There are various stations throughout these areas where people can rent a bicycle and pay according to the duration of their rental. There are also different subscriptions that may be utilized by more frequent Divvy riders. Divvy bikes can be used as a form of recreation or transportation. This work seeks to analyze the ride patterns of Divvy bike users and track Divvy bike movement around Chicago and Evanston.

Dataset

The Divvy dataset (*divvy*) consists of Divvy trips taken from 2013 to 2020 available in the Chicago Data Portal: https://data.cityofchicago.org/Transportation/Divvy-Trips/fg6s-gzvg/data_preview. The data has 21.2 million rows and 18 columns. Each row is one trip. For each trip, the trip ID, start time, stop time, bike ID, trip duration, from station ID, from station name, user type, gender, birth year, from latitude, from longitude, from location, to latitude, to longitude and to location are provided. The times are provided in the format MM/DD/YYYY HH:MM:SS AM or PM. Trip duration is provided in seconds. “From station” indicates where the trip began, and “to station” indicates where the trip ended. User type is either “customer,” “dependent” or “subscription.” “From location” is the coordinate pair for the location where the Divvy trip began. “To location” is the coordinate pair for the location where the trip ended.

Methods

I performed data cleaning before I could start manipulating it. All the column names were changed to have only lowercase letters and any spaces in the column names to underscores using the clean_names() function in the ‘janitor’ library. From there, all dates and times needed to be converted to datetime objects to ensure calculations could be performed on them. After doing this, the data was ready to be analyzed.

Missing Gender. The rides that had a missing value under the gender variable and possible reasons were analyzed. The dataset was filtered for rows that had a missing value for gender, then they were counted. This provides the total number of entries that have missing values. This value was then divided by the total entries (number of rows) in *divvy* to get the proportion of entries without gender listed. The types of users were also analyzed to look for reasoning as to why certain entries did not state gender. To do this, *divvy* was filtered for entries with missing values for gender, grouped by user type and the count of users of each user type was calculated.

Average Rides Per Day. The average rides per day were calculated and stratified by user type and then by gender. To do this, a variable (“start_date”) was added to *divvy* which extracted the date section of the start time without the time of day. This was then grouped by start date and user type, and the number of rides per date and user type was calculated. From here, a variable (dayofweek) that extracted the day of the week for each date was extracted. This new dataframe was grouped by day of week and user type, and the average number of rides per day was calculated for each group. A line graph was used to demonstrate whether there was a pattern in which day of the week types of users use Divvy bikes. This process was repeated for gender. To do this, rides with missing gender entries were excluded, and the process was repeated by replacing user type with gender in the analysis. A line graph was used to demonstrate whether there was a pattern in which day of the week males and females use Divvy bikes.

Average Rider Time of Day. The average number of riders at different times of the day throughout the week were calculated for male and female Divvy users. To do this, *divvy* was first filtered for rides without missing entries for gender. A variable ("start_date") was added to *divvy* which extracted the date section of the start time without the time of day, and a variable ("s_time") was added to extract the time of day (the hour in 24-hr time). This dataset was then grouped by start date, gender and time of day, and the number of rides for each group was counted. A variable ("dayofweek") was then added to extract the day of the week for each date. The dataset was then grouped by day of week, time of day and gender, and the average number of riders for each group was calculated. This dataset was then ungrouped to ensure a heatmap could be created. For the male entries, the new dataset was filtered for the entries that listed "male" as the gender and a heatmap was created to show the average number of male riders per day at different times of the day. The same was then done for female riders. A second heatmap was created for female riders with the same range as the male heatmap in order to make a comparison.

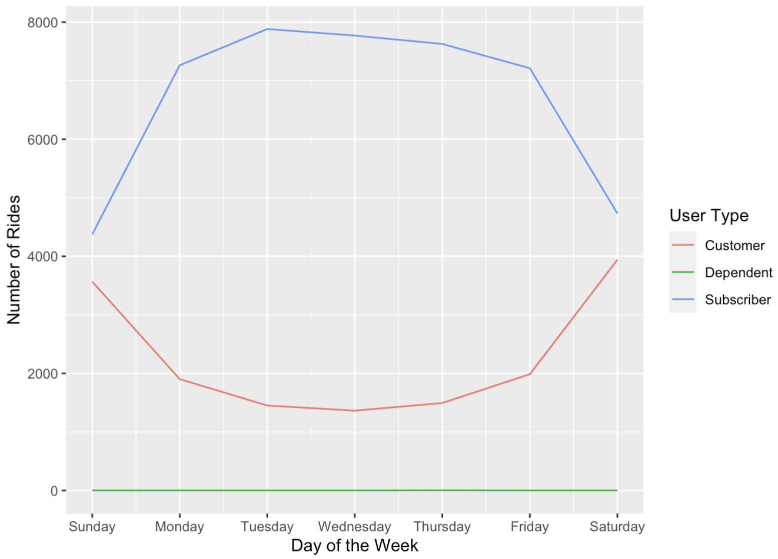
Busy Month Stations. The difference in bikes taken and returned to stations were analyzed during the month where the most Divvy rides were taken. First, a variable ("start_month") extracting the month from the start date was added. Another was added for the stop date. *Divvy* was grouped by the start month variable. *Divvy* was grouped by stop month and number of rides per month was calculated separately to compare whether the same start and stop month had the most rides. Divvy was filtered for the busiest start month and the number of bikes that start at each station was calculated and set to a data frame. The same was done for bikes that stop at each station and another data frame was created. These data frames were joined by station name and the difference between bikes taken and bikes returned to each station were calculated in a new variable. The stations with the biggest positive differences were stations that had more bikes taken than returned in that month. The stations with the negative differences were stations that had more bikes returned than taken in that month.

Bike Not Used. The proportion of bikes that have their first ride in a specific year and are not utilized in the years that follow was analyzed. To do this, divvy was grouped by bike id and the first year and last year that each bike was used were added as variables. The year 2019 was removed because the Divvy data ends in 2019 so there is no information about bike use in the next year. This new dataset was set to the variable bike_life. The bike_life was then grouped by the first year that the bike was used and the count of each bike first used in each year was calculated and this data frame was set to a variable new_year. Bike_life was filtered for years where the bike was first and last used in the same year. This was then grouped by the first year that the bike was used and the count of bikes in each year was calculated. The new_year dataframe was then joined to this dataframe by the "first_year" and the proportion of bikes only used within the same year was calculated for each year by dividing the number of bikes only used in one year by the total number of bikes first used within that year.

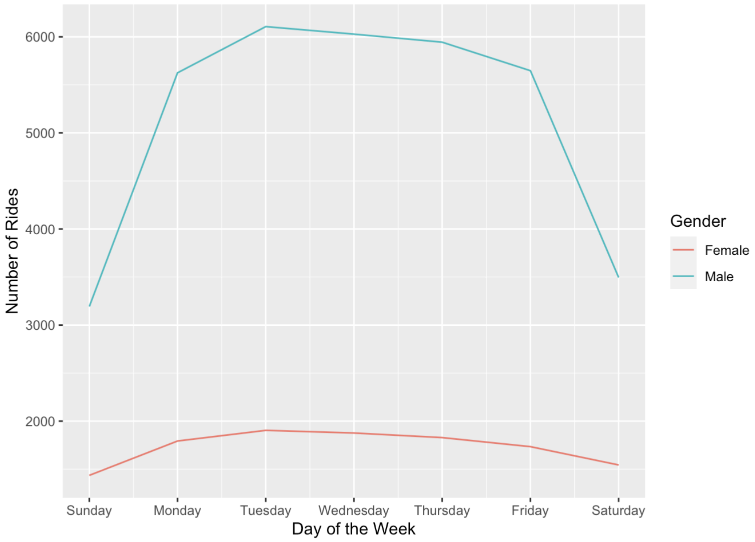
Results

Missing Gender. The rides that had missing values under the gender variable accounted for 23.04% of the data entries. Due to this large percentage of missing values, these rides were further analyzed. Of the rides that did not specify gender, 99.3% of users were customers, 0.7% of users were subscribers and less than 0.1% of users were dependents. A potential reason for customers accounting for a majority of the missing gender information is that customers can buy passes as needed at a Divvy station and are not required to have an account, whereas subscribers will have an account where they can enter their personal information.

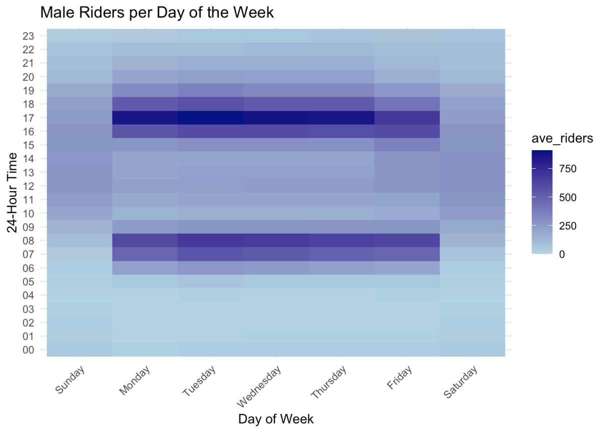
Average Rides Per Day. A pattern was observed for both graphs. Customers tend to start rides during the weekend and their use decreases as it gets closer to the middle of the week. This may be because customers buy passes as needed so their demographic could increase over the weekend due to tourists coming into the city on the weekends and more people using Divvy bikes for recreational use. The opposite is observed for subscribers. Subscribers tend to take rides during the work week and less during the weekend. This may be due to using Divvy bikes for transportation to and from work during the work week. Dependent's use of Divvy bikes is consistently low throughout the week compared to the other user types. When looking into this more, it was discovered that dependents make up less than 1% of users in the dataset.

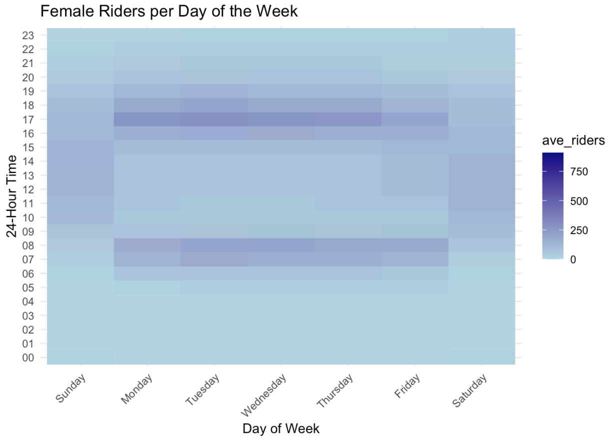


The trend in Divvy use between males and females showed similar trends in that use was highest during the work week and decreased on the weekends. The use during the work week versus on the weekends by males had more of a significant difference than that of females.

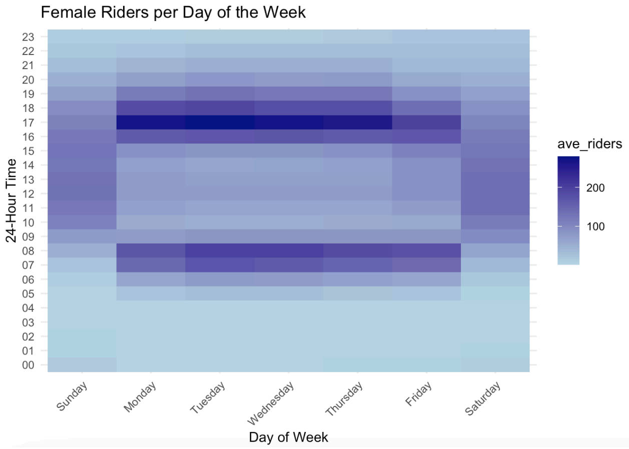


Average Rider Time of Day. The male and female heat maps showed similar patterns with most rides starting during the work week in the morning and in the evening. This supports my inference drawn from *Average Rides Per Day* that the increased riders during the work week (for subscribers and rides where gender is listed) are due to people using Divvy bikes mostly to get to and from work. During the weekend, rides are mostly taken in the middle of the day, approximately between the hours of 9am and 4pm.



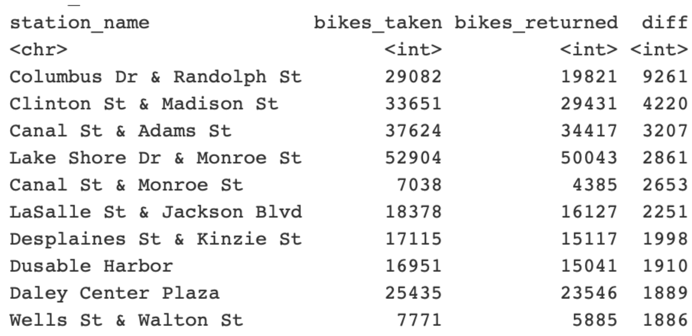


*these two graphs above have the same range on the legend so the male heatmap can be compared against the female heatmap

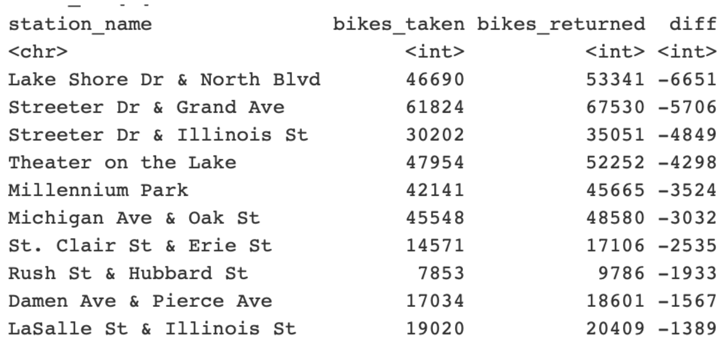


*this graph has a smaller legend because the range of the average number of female riders is much smaller than the range of the average number of male riders

Busy Month Stations. The busiest month for Divvy rides is August. The stations that had more bikes taken than returned in that month were:


```
station_name               bikes_taken bikes_returned  diff
<chr>                            <int>          <int> <int>
Columbus Dr & Randolph St        29082          19821  9261
Clinton St & Madison St          33651          29431  4220
Canal St & Adams St              37624          34417  3207
Lake Shore Dr & Monroe St        52904          50043  2861
Canal St & Monroe St              7038           4385  2653
LaSalle St & Jackson Blvd        18378          16127  2251
Desplaines St & Kinzie St        17115          15117  1998
Dusable Harbor                   16951          15041  1910
Daley Center Plaza               25435          23546  1889
Wells St & Walton St              7771           5885  1886
```

The stations that had more bikes returned that taken that month were:



```
station_name                bikes_taken bikes_returned  diff
<chr>                             <int>          <int> <int>
Lake Shore Dr & North Blvd        46690          53341 -6651
Streeter Dr & Grand Ave           61824          67530 -5706
Streeter Dr & Illinois St         30202          35051 -4849
Theater on the Lake               47954          52252 -4298
Millennium Park                   42141          45665 -3524
Michigan Ave & Oak St             45548          48580 -3032
St. Clair St & Erie St            14571          17106 -2535
Rush St & Hubbard St               7853           9786 -1933
Damen Ave & Pierce Ave            17034          18601 -1567
LaSalle St & Illinois St          19020          20409 -1389
```

Bike Not Used. Only a small proportion of bikes used for the first time each year are not used in following years. The proportions are 0.3% for 2013, 1.02% for 2015, 0.56% for 2016 and 1.09% for 2017. This could be because most bikes are repaired or revamped as riders can report if there is a problem with a bike. Some possible reasons for bikes not being utilized in future could be theft or damage beyond repair.

Conclusions and Future Work

It can be concluded that Divvy bike usage is most popular at different times of the week depending on the user. This may be a way that Divvy has consistent business throughout the week. The busiest time of year for Divvy is in August which is understandable because it is during the summer, and there are standout popular Divvy stations for starting rides and ending rides.

It would be interesting to explore data on Divvy bike-sharing that also included a variable for bike type and for bike repairs. The use of electric vs traditional (or acoustic) bicycles could cause for an interesting analysis. It would also be interesting if for the rides by subscribers, a subscriber ID was also included to analyze the ride patterns of individuals. This would need to be done ethically to ensure the individuals' privacy.

Appendix

All code can be found in the following GitHub Repository:
https://github.com/csanyika/DivvyChicago