CSC-465 PROJECT SUBMISSION

Data Visualization on
Divvy Bike Dataset
2016

by

Team: PMAD

Ashrita Nyna Mannepalli

Daniel Glownia

Meghana Seggem

Pradeep Sathyamurthy

Under the guidance of: Dr. Eli T. Brown

DePaul University

21st November 2016

Contents

Project Summary

Introduction

Divvy Bike sharing is an innovative transportation system located in the City of Chicago operated by franchise [1] for Chicago Department of Transportation. It is ideal for short distance point-to-point trips providing users the ability to pick up a bicycle at any self-serve bike-station and return it to any other bike station located within the system's service area. Our project focus on visualizing the Divvy data and find hidden values and insights from them using modern visualization techniques available and taught as part of this course.

Project Scope

This dataset consists of Divvy Bike usage from January 2016 till June 2016. Thus, we have ***scoped our visualization only for these 6 months*** for which the data is available from Divvy Bike website[4]. Since, we do not want to get influenced on already visualized dataset of Divvy for past years, we as a team decided to work on this latest 6-month dataset shared by Divvy to get a fresh perspective on current trend.

As an extra mile, members in our team built a linear regression model by fusing the Divvy dataset with other datasets like weather, wind speed, etc., and found that ***temperature had a significant correlation of 0.39*** compare to other factors like wind speed and Humidity which had significantly less correlation [2]. Hence, we decided to scope our extra mile to visualize Divvy Bike Usage based on weather in Chicago from Jan-Jun 2016 for which our main visualization is scoped. We were also interested in visualizing the ***Divvy Ride behaviour between Male and Female*** as our final regression model [3] showed a high usage by Male compare to female and even their age being a significant contributor.

Dataset Description

Our Divvy dataset [4] was derived from the official Divvy website. Dataset consisted of totally 5 CSV files with 2 major classifications:

1. **Divvy Trips Details:** we had 4 separate CSV files for 6 months of Divvy Trips data constituting ***14,69,740*** trip details with below sets of features:

SI.NO	Variable	Description
1	trip_id	ID attached to each trip taken
2	starttime	day and time trip started, in CST
3	stoptime	day and time trip ended, in CST
4	bikeid	ID attached to each bike
5	tripduration	time of trip in seconds
6	from_station_name	name of station where trip originated
7	to_station_name	name of station where trip terminated
8	from_station_id	ID of station where trip originated
9	to_station_id	ID of station where trip terminated
10	usertype	"Customer" is a rider who purchased a 24-Hour Pass; "Subscriber" is a rider who purchased an Annual Membership
11	gender	gender of rider
12	birthyear	birth year of rider

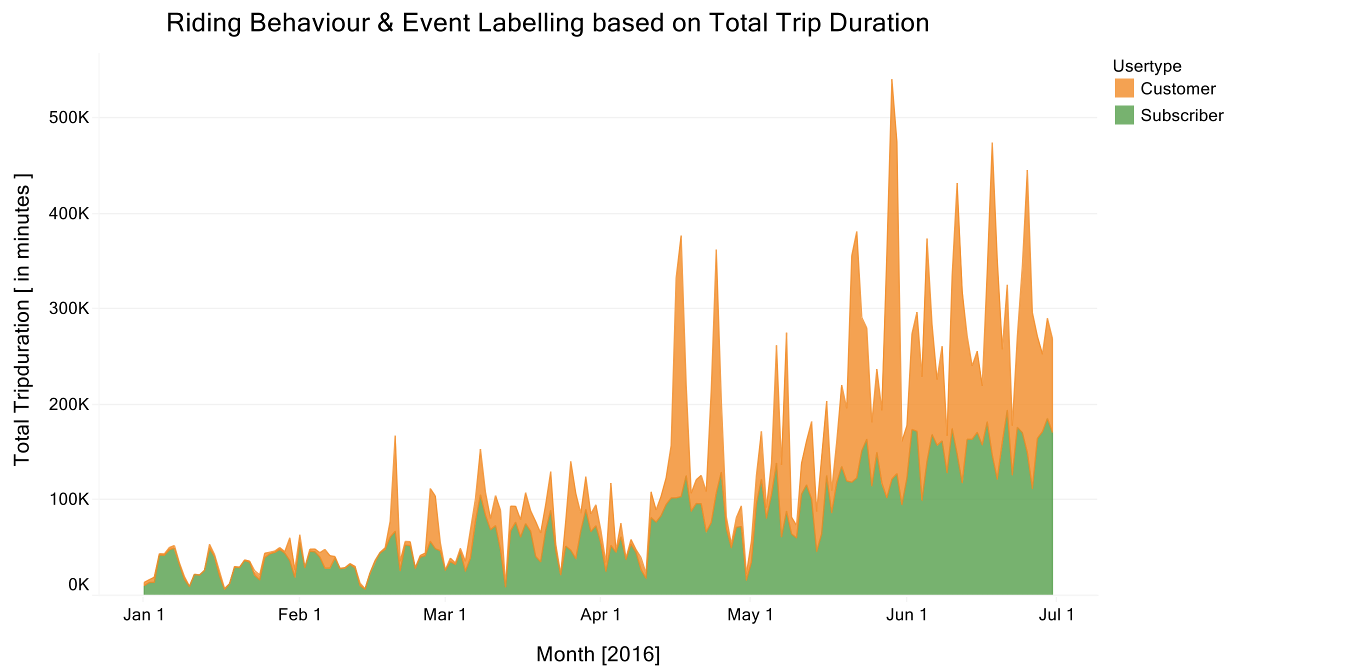
2. **Divvy Station Details:** we had a CSV file with ***535*** records having details about all Divvy stations in Chicago

SI.NO	Variable	Description
1	id	ID attached to each station
2	name	station name
3	latitude	station latitude
4	longitude	station longitude
5	dpcapacity	number of total docks at each station as of 6/30/2016
6	online_date	date the station went live in the system

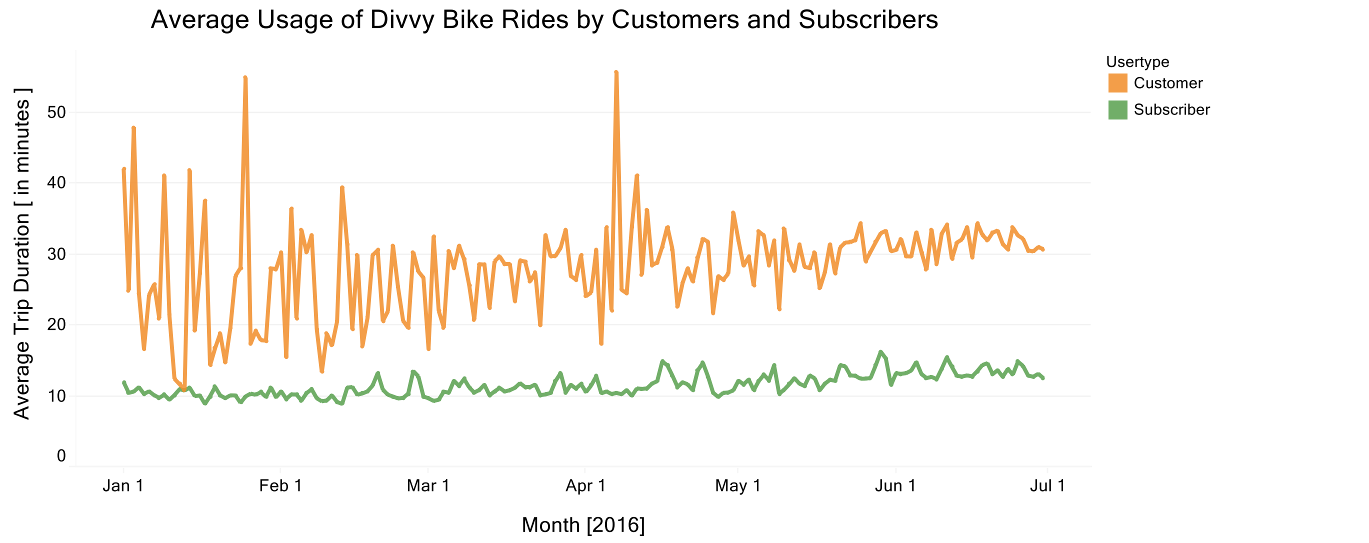
Final Visualizations for Grading

Visualization-1 – Time Series

1. ***Area chart to visualize Event Labelling* with x axis mapped to continuous days labelled by Months and Y axis mapped to Total Trip Duration:**



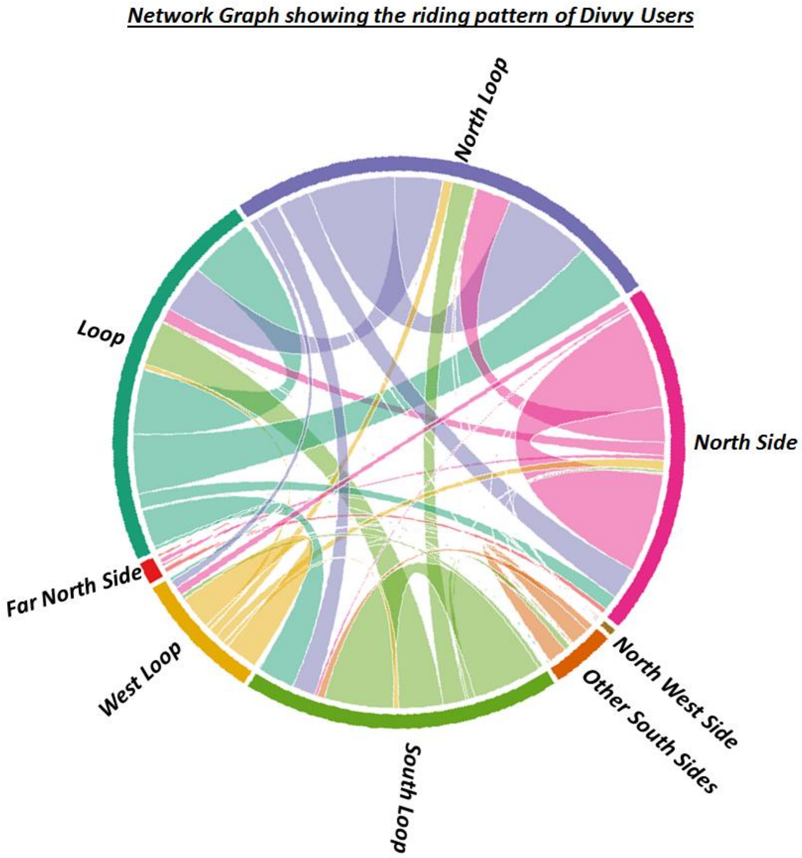
2. ***Line graph to study Bike usage* by Customer and Subscribers with x axis mapped to continuous days labelled by Months and Y axis mapped to Average Trip Duration.**



- Event labelling [5] is the ***process of marking events*** in unlabelled data. Traditionally, this is done by involving one or more human experts through an expensive and time-consuming task
- We wanted to experiment on ***how visualization can optimize the task of event labelling*** as part of these time-series graphs***. For e.g***. ***highest peak in above area chart represent the day of Memorial Day parade***.
- Particularly we found ***that Area chart is the most optimized graph for event labelling*** as it represents the high and low peaks (usage) considering the difference in population from time period to time period.
- We also wanted to visualize how Divvy pricing [6] alters the ***usage behaviour between Subscriber and customer***, for this ***line graph was the best*** representation which avoid distractions due to population difference between Customer and Subscribers. ***Refer details***[D] ***and insights***[I] ***for more information.***

Visualization-2 – Chord Diagram/Network Plot [Built in R]

***Chord Diagram to visualize the riding pattern** of Divvy Users with each coloured arc representing the Source/base station where ride started:*



- Chord Diagram is a form of ***network diagram used to represent a data that holds navigation pattern*** by having a source and a particular destination.

- In above chord diagram ***each coloured arc represent the base station***. This is the point from where the trip starts. ***For e.g.*** green colour arc represent the rides which started from South Loop, we see users navigating to various destinations like Loop, North Loop and we could prominently see most user to drive within south loop allot. Thus, ***each arc represent the variable from_station_id in a consolidated form present in the dataset***.

- Thickness in the above network diagram represents the total number of trips made by users in that route. It shows, ***users normally hire Divvy to commute within the same sides*** (locality with shorter distance) rather than using Divvy for long travel commute. From above graph we could infer that ***Divvy is highly used to commute within downtown*** when compared to other sides.

- High navigation *or* ***busy route with respective to Divvy seems to be Loop and North Loop***

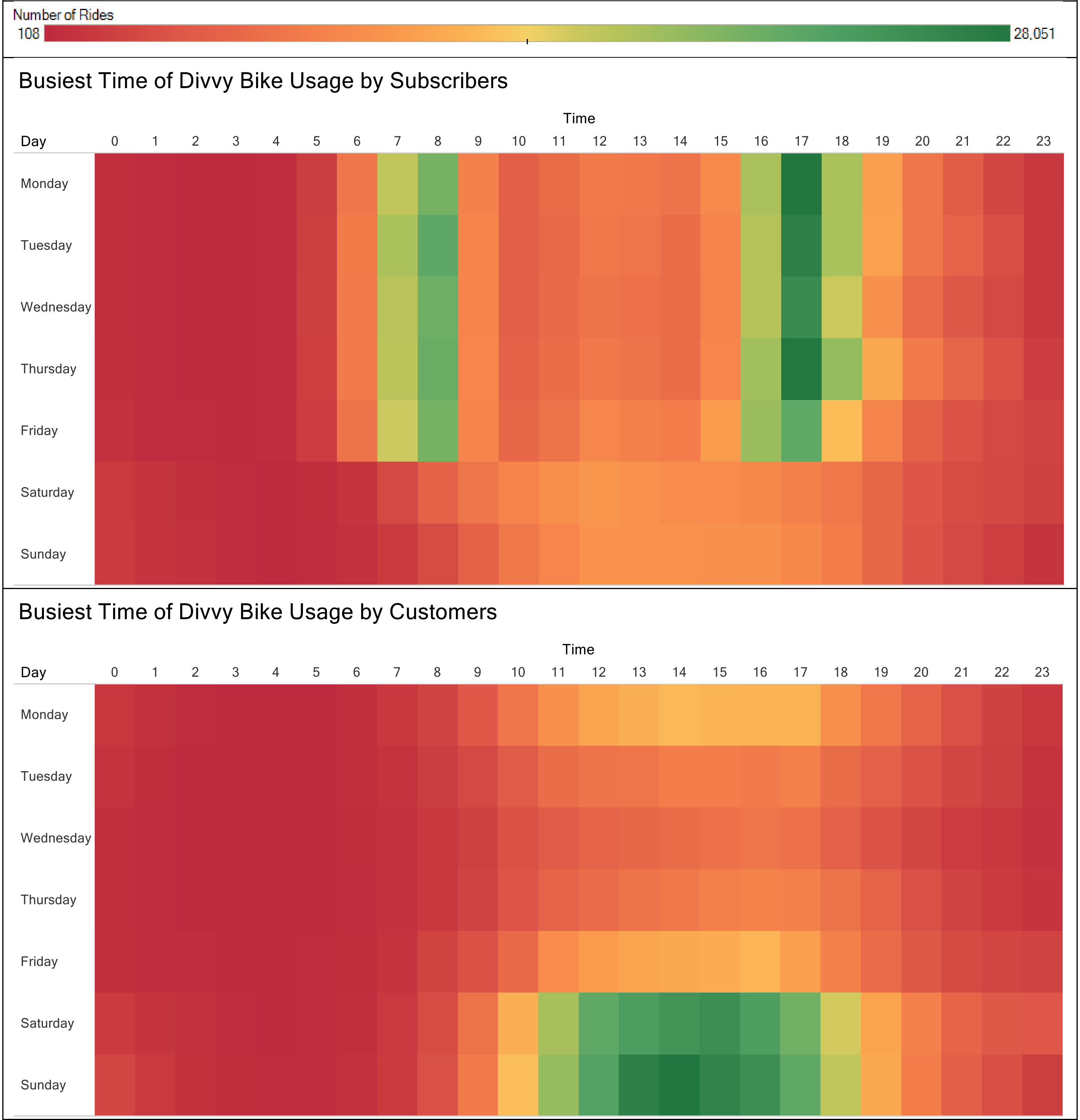
- Least navigation or ***less busy route with respective to Divvy seems to be North West Sides, Far North Sides***.

- Since the usage pattern between Subscribers and Customers where almost resembling the same, we represented one chord diagram to show the overall riding pattern of the Divvy users.

- This was the ***highest possible resolution we could make for a chord diagram that represent 14.69 million records***, further more R-studio is distorting the whole visualization. ***Refer details[D] and insights[I] for more information.***

Visualization-3 – Heat Map

Heat map to visualize Busiest Hours of Divvy Bikes usage by Subscribers and Customers with x axis mapped to Hours in a day and Y axis mapped to Days in a week.



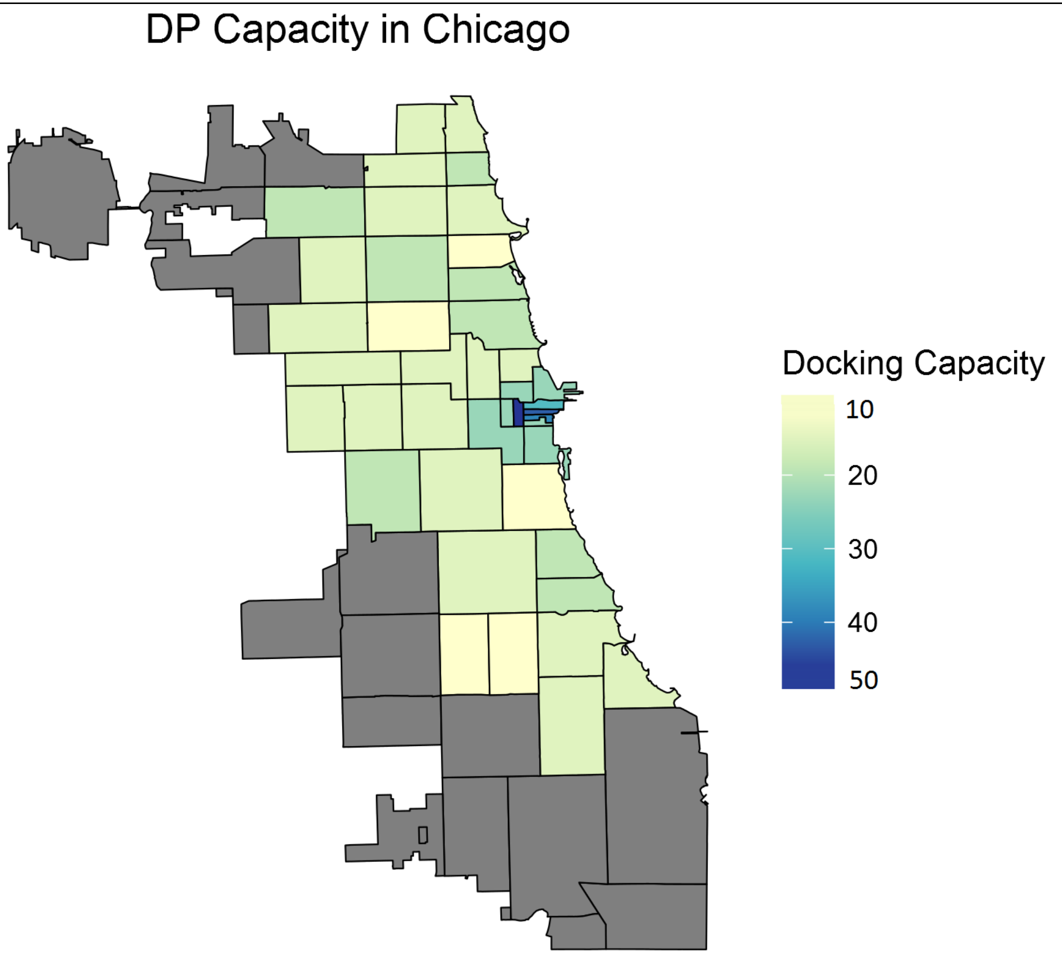
Busiest Time of Divvy Bike Usage by Subscribers

Busiest Time of Divvy Bike Usage by Customers

- ***Subscribers*** use frequently between ***7-8AM and 4-6PM on weekdays***, while ***customers*** use frequently during ***weekends between 11AM until 5PM***.
- Divergent color palate was used to clearly differentiate between ***High, Moderate and Low usage of Divvy Bikes*** for a particular hour by user.
- Midpoint is pinned to represent moderate usage of Divvy Bikes to visualize the usage between Day and Night times. ***With just a single color it was hard to differentiate the amount of Moderate usage of Divvy Bikes by users*** and hence Divergent color was our best bet.
- We observe a ***moderate usage of Divvy bikes*** between ***9AM until 4PM*** by customers on Monday and Friday, while other weekdays have less usage, So ***Divvy can do some promotion for Tuesday – Thursday to increase the bike usage on these days***. ***Refer details[D] and insights[I] for more information.***

Visualization-4 – Choropleth [Built in R]

Choropleth to visualize the docking capacity in Divvy Station at various Chicago's neighbourhood

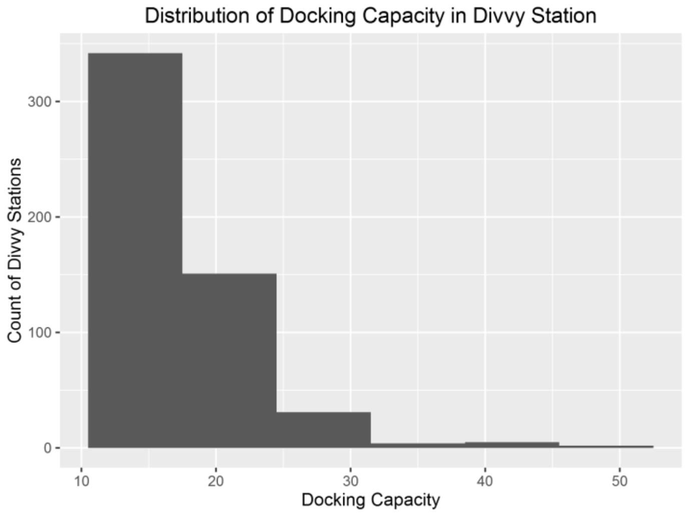


- Choropleth was our best option to visualize the importance given to each Chicago's neighbourhood in implementing Divvy bike sharing program.

- In ***the above visualization we represent the proportion of docking capacity in each Chicago neighbourhood*** based on the usage pattern visualized as part of network diagram above.

- Since ***Divvy bikes sharing holds Blue Color as their trade mark, we wanted to use sequential Blue color gradient*** (***"YlGnBu"***[8] ***which is printer, colour-blind and photo-copy safe***) to represent the docking capacity of Divvy at each neighbourhood.

- From above graph, we could clearly infer that ***area around loop have stations with higher Docking capacity. To be precise area around Navvy Pier has biggest Divvy station with a capacity of 47 docks.*** No wonder as it is one of the important tourist attraction.

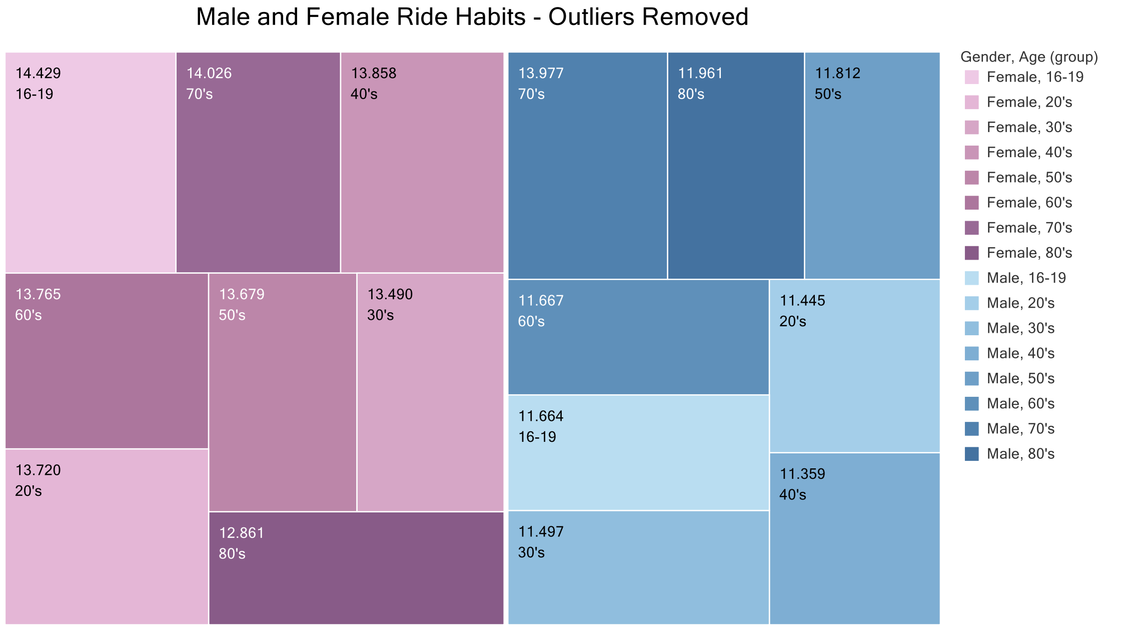


- Each docking station has ***minimum of 10 docks to the maximum of 47 docks*** in least and most busiest areas respectively, ***same is represented in histogram as well above***. ***Refer details***[D] ***and insights***[I] ***for more information.***

Extra Miles

Extra Mile Visualization-1 – Tree Map

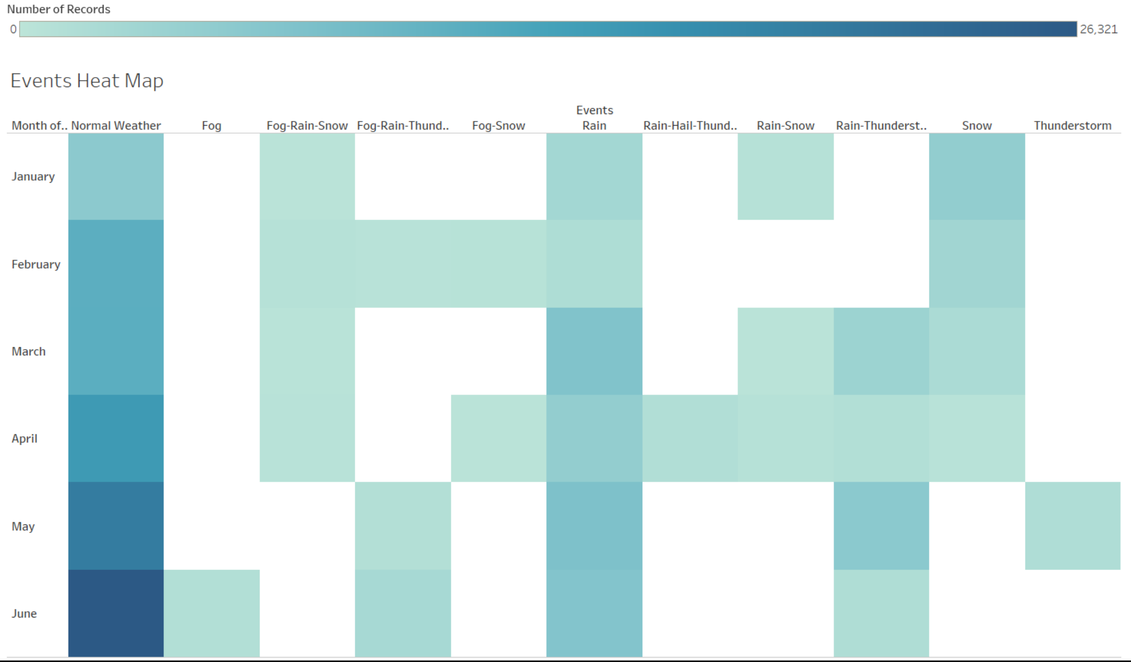
Tree Map to visualize the Male and Female ride behaviour – Binning Age Groups based on average ride time (in minutes)



- Tree map are best to represent the Male and Female ride habit as it represents the hierarchical data with respective to age.
- Built by ***binning*** the discrete ***age*** of subscribers in to age groups. For e.g. 20 to 29 years old as 20's & so on
- Average ride time of Female is more than Male which can be directly correlated with stamina factor b/w genders. ***Refer details[D] and insights[I] for more information.***

Extra Mile Visualization-2 – Heat Map

Heat Map to Visualize the weather impact on Divvy Bike Usage from Jan-Jun 2016:

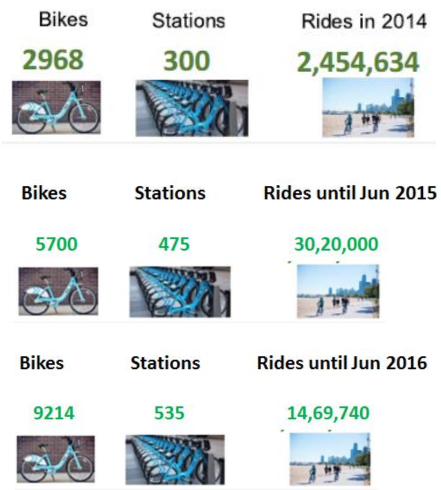


- Temperature has ***positive correlation (0.39)*** with Divvy bike usage based on linear regression model built
- Hence we took weather data from weather forecasting website[7] and fused with Divvy dataset to see the impact.
- We observed high usage during normal weather, ***Snow affects the usage more than rain*** and wind speed has no correlation with Divvy bike usage. ***Refer details[D] and insights[I] for more information.***

Detailed Report on Each Visualization

Basic Statistics on Divvy Dataset

- Our Dataset is obtained from the official Divvy website[4].
- Dataset consist of 2 CSV files explaining about trip details and Divvy location.
- From the basic statistics, we could figure out that ***until June 2016*** Divvy had ***5837 bicycles*** in circulation within Chicago city with ***535 stations*** in Chicago neighbourhood.
- Total number of ***rides*** until June 2016 was ***14, 69, 740***. ***As part of this project we started with exploratory data analysis*** to study the behaviour of each variable to get some direction on Visualization.
- We could clearly see the growth of Divvy being 3 times in Bike count, 2 times in Station and Number of Rides being increasing from 2014 until now. This is ***clear indication of success of Divvy Bike Program in Chicago***.



Exploratory Data Analysis on Divvy Dataset

- As part of EDA we tried to plot histogram , box-plot, pie &bar for each variable and build correlation matrix to understand the variables in the dataset. A snapshot of the exploratory analysis done is shown in Appendix [9]
- Key take away as part of the whole EDA done on dataset was, ***any indepth analysis needs to be done based on Subscriber and Customer seperately***.
- Subscribers are Divvy users with yearly membership constituting 78% of ride within Chicago. While Customers are Divvy users with 24hrs day pass who constitute 22% of total divvy rides. Based on this propotion found as part of EDA ***we where very careful so that Population of Subscribers doesn't shadows Customer behaviour***.
- A scatter plot between Bike Usage vs Trip Start Time gave a picture that most busiest ***divvy usage time is between 7 to 9Am and 4 to 6 Pm***. However, this was a consolidated behaviour, so ***we where interested to view how it is based on user type***.
- We also created a LR model (shortened the equation for display) ***Count = 2.6 + seasons + 0.12*holiday1 + 0.23*workingday1 + 0.02*temp + 0.004*Casual – hours_2_6 + hour_7_24 + days_Sat_Sun – day_Mon_Thr + 0.23*gender_male***

Time Series Graphs on Divvy Dataset

<u>Event Labelling:</u>

- ***Event labelling*** is the process of ***marking events in unlabelled data***. We studied an application of Analytics in Event Labelling in a research paper [5] and hence thought to try how a visualization can optimize this process.
- For this we chose Area map which we thought would well represent the Bike Usage considering the change in population between Customer and Subscriber population density. To our surprise we got a great insight.
- All ***spike*** [both high and low] in graph-1[G] clearly ***represent an event happening in Chicago loop***.
- We would like to highlight few, the ***highest peak*** you see in in graph-1[G] represent the ***Chicago memorial day parade on May 28th <u>in loop area</u>***. ***Second higher peak in same graph represent Chicago Pride Fest on June 18th <u>in loop area</u>***. This way we feel such kind of rich visualization is useful to find great insights from data. However, this is just a snapshot of what we experimented.

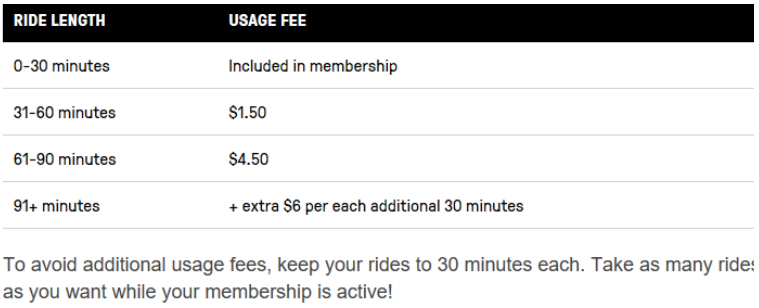
Pricing impact on Divvy Bike Usage between Customer and Subscribers:

We see pricing[6] has a definite impact on Divvy bike usage. In order to visualize this we plotted a line graphs based on Average Trip Duration as part of time series graph-2[G].

From graph-2[G] time series line graph it is more evident now why ***Subscribers average trip duration is between 10 to 15 mints***. This is ***because subscribers have additional usage fee for rides exceeding 30 mints***.

However, for ***customers*** there is no restriction as such and hence we could clearly see there ***average ride time is between 20 to 40 mints***.

Thus, EDA helped in assisting to make a study separately for Customer and Subscriber and we see its result with this riding pattern.



RIDE LENGTH	USAGE FEE
0-30 minutes	Included in membership
31-60 minutes	$1.50
61-90 minutes	$4.50
91+ minutes	+ extra $6 per each additional 30 minutes

To avoid additional usage fees, keep your rides to 30 minutes each. Take as many ride as you want while your membership is active!

Stories Behind:

- ***Subscribers usage of Divvy Bikes are more when compare to customers. However, average ride time of Customer is more when compare to subscribers.*** It can be hypothesized that Subscribers are users who commute to work place and Universities from their residence/CTA situated in and around loops. While Customers are tourist and other users who try to explore Chicago city.

Chord Diagram on Divvy Dataset

About the Graph:

After ***experimenting the Event Labelling*** we see most of the events associated with high peaks normally happen in Loop and hence the Divvy Usage in and around that area. So ***we wanted to visualize how does the riding pattern looks like with the dataset we have***. We found Chord diagram will be the best choice to represent this navigation.

This ***needed huge preparation in data***. Since raw dataset had around 535 nodes whose visualization was not appealing. Hence we mapped all ***535 location to respective 72 distinct Zip code***[10] and from that we ***further mapped each zip codes to 8 Sides***[11] as shown in chord diagram [G]. Each arc represent the base from which a trip starts and each strip shows the number of trip towards destination.

Stories Behind:

- From above graph we could infer that ***Divvy is highly used to commute within downtown*** when compared to other sides.
- Thickness in the above network diagram represents the total number of trips made by users in that route. It shows, ***users normally hire Divvy to commute within the same sides*** (locality with shorter distance) rather than using Divvy for long travel commute.
- High navigation *or* ***busy route with respective to Divvy seems to be Loop and North Loop***
- Least navigation or ***less busy route with respective to Divvy seems to be North West Sides, Far North Sides***.
- Thus, ***we suggest Loop and North Loop needs much more new Divvy Stations with further increase in Docking Capacity***.
- Divvy must ***concentrate actively navigate their bicycles in Loop and North Loop to meet the users demand*** and thereby increase their productivity.
- This also reveals, ***bicycles at these (Loop & North Loop) stations are used more and hence needed frequent maintenance*** to keep us the user experience sustaining.

Heat Map on Divvy Dataset

About the Graph:

Based on Time Series graph and Chord Diagram ***we came to know that Divvy usage increases whenever there are events in loop*** and further our ***chord diagram confirms that most of the navigation is happening with in Loop*** and North Loop.

With this information, ***we were interested to visualize at what time the Divvy usage increases***. With basic EDA it was visualized that morning 7 to 9 Am and evening 4 to 6Pm has a high usage. However, since ***EDA suggest us to visualize the Customer and Subscriber behaviour separately***, we wanted to use heat map which can best represent such pattern in usage of Divvy bikes.

We ***wanted to study the Divvy usage based on 3 levels High, Moderate and Low.*** And ***hence used divergent colour pallet*** with midpoint representing the moderate usage of Divvy Bike. This was very important for us to see how much moderate usage of Divvy Bikes is happening in day times in order to promote the usage further.

We designed Heat map[G] for Subscribers and Customers separately and as expected, we ***saw a difference in usage of Divvy by these two different user types***.

Stories Behind:

- ***Subscribers*** use frequently between ***7-8AM and 4-6PM on weekdays***
- While ***customers*** use frequently during ***weekends between 11AM until 5PM***.
- Without doing proper EDA and initial analysis on dataset, we could have not got this insight.
- ***This proves the hypothesis true, subscribers use Divvy to commute for their work place and Universities***.
- ***While Customers use Divvy to explore city*** in the Day time.
- We also observe a ***moderate usage of Divvy bikes*** between ***9AM until 4PM*** too on Monday and Friday, while other weekdays have less usage, So ***Divvy can do some promotion for Tuesday – Thursday*** to increase the bike usage.
- Also, we suggest Divvy must ***concentrate in moving around their bicycles more actively between the working hours (6-9AM & 4-7PM)*** sighted above to meet the user demands.

Choropleth on Divvy Dataset

After, we studied the Divvy Bike Usage, pattern of rides within loop and busy hours of Divvy bike usage, we wanted to visualize how Divvy manages the availability of Bikes in each Divvy station. So, we ***wanted to visualize the docking capacity in each divvy station across Chicago neighbourhood***.

For this, we ***decided that Choropleth[G] would be a right choice*** and hence started to work on it.

Since ***Divvy bikes sharing holds Blue Color as their trade mark, we wanted to use sequential Blue color gradient*** (***"YlGnBu"[8] which is printer, colour-blind and photo-copy safe***) to represent the docking capacity of Divvy at each neighbourhood.

There was a ***considerable amount of time spent in finding the right shape file for Divvy dataset***. We finally found shape file with respective to zip-code useful.

This ***needed various binning on data in order to match each location with its respective zip codes*** which are present in shape file.

Once, data was transformed as per shape file chosen, ***considerable time was spent in fortifying and merging the shape file with my Divvy Dataset***. Once this was done, we where able to plot choropleth based on DP Capacity.

Stories Behind:

- We suggest Divvy should ***establish new docking stations in Far Southeast and Far Southwest sides*** of the city (*areas under the bottom of above graph*)
- ***These are the areas even with non-frequent CTA lines*** and thus setup of Divvy Station in such area will definitely help people who are residing in those area.
- Divvy can also concentrate in setting up ***free Docking or low cost docking stations in O'Hare airport***. This would help transit passengers to commute from once terminal to the other.
- I do understand that airport is running monorails to connect terminals, however such Divvy setup will definitely ***help many local employees and would also show the importance given*** by state personals in encouraging such renewable transportation and working ***towards climatic control measures to the world***.

Male and Female Ride Behaviour

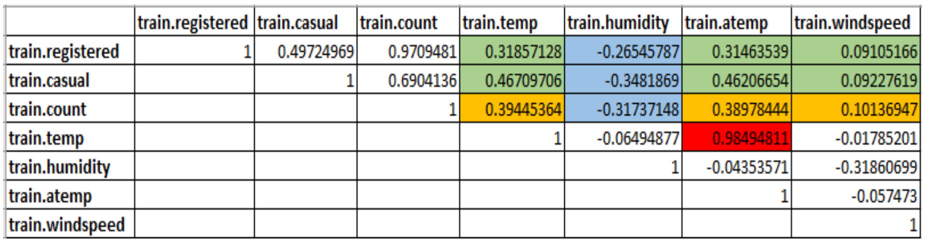
In team we were ***very interested to see if stamina factor between each gender has any correlation with Divvy dataset***. Since it is challenging to derive a dataset to compute correlation with stamina level, we wanted to hypothesize it using visualization techniques.

Hence, we ***chosen a Hierarchical tree map***[G] to visualize it. In order to build this visualization, we re-engineered the dataset by introducing a ***new feature called age group derived from Age of the Subscriber***. Using this, we plotted a hierarchical tree map based on Average Ride Time.

Visualization rendered out of it ***proved the hypothesis true***. We ***observed Female average ride time is more when compare to Male.*** We where also surprise to see that Divvy had few subscriber who are even 70 and 80 years older driving Bikes. We examined if it is a outlier, but we found that such trips where very less and non-frequent, so there is a ***huge chance that senior citizens do use Divvy Bikes at times***.

Weather impact on Divvy Dataset

We were interested in collating various data like weather, traffic, pollution, etc., to study the impact of them on Divvy bike usage. ***Due to time constraint we scope our explanation only with respective to temperature***. When we merged temperature data along with Divvy dataset and created a correlation matrix below was the result obtained:



	train.registered	train.casual	train.count	train.temp	train.humidity	train.atemp	train.windspeed
train.registered	1	0.49724969	0.9709481	0.31857128	-0.26545787	0.31463539	0.09105166
train.casual		1	0.6904136	0.46709706	-0.3481869	0.46206654	0.09227619
train.count			1	0.39445364	-0.31737148	0.38978444	0.10136947
train.temp				1	-0.06494877	0.98494811	-0.01785201
train.humidity					1	-0.04353571	-0.31860699
train.atemp						1	-0.057473
train.windspeed							1

Looking at the above matrix, we found that temperature is highly correlated (0.39) when compare to other weather factors like humidity, feel like temp (ATEMP), etc., We also observed wind speed has no correlation with Divvy Bike usage. So in order to visualize this we used heat map[G] for weather vs Divvy Bike usage and derived below stories:

Stories Behind:

- ***Summer*** is the season which has ***highest usage*** of Divvy Bikes
- ***Winter*** is the season which has ***lowest usage*** of Divvy Bikes
- ***Spring*** sees a ***moderate usage*** of Divvy Bikes by its users.
- So, Divvy ***can plan for more promotions during Winter and Spring season to attract more users***. However, Divvy must also concentrate on safety measure in using Divvy bikes at these seasons accordingly.
- We also observe Snowy days has more impact in Divvy usage when compare to Rainy days.

Insights

1. Usage of Divvy increases every year, this shows the success of program in Chicago

2. Divvy Bike usage by Subscribers is completely different when compare to Customers

3. Area plot on Divvy Dataset helps in optimizing Event Labelling[5]

4. Average ride time of Customer is more compare to Subscribers due to pricing structure

5. Busiest route of Divvy is Loop and North Loop. Least busy route of Divvy is North West Side and Far North Side.

6. Divvy should concentrate on actively navigating their bicycles in loop and North Loop and also actively maintain bikes in these areas as they are subjected to frequent usage.

7. Subscribers use Divvy mostly in weekdays to commute office between 7-8 Am and 4-6Pm

8. Customers use Divvy mostly in weekends for recreations between 11Am until 5Pm

9. Divvy can launch promotions for Tuesdays, Wednesdays and Thursdays as these days seem to have less than moderate usage of Divvy Bikes by customers

10. Divvy should establish new docking stations in Far Southeast and Far Southwest sides where there is less CTA frequency

11. Divvy can also concentrate in setting up free Docking or low cost docking stations in O'Hare airport for local usage

12. We observed Female average ride time is more when compare to Male, stamina plays a significant role in Bike rides. Divvy can introduce women friendly bicycles which will make them ride it much more easily

13. Divvy bikes have subscribers greater than 80 years who use Divvy Bikes, Divvy can give promotions and advertise them

14. Divvy should plan for more promotions during Winter and Spring season to attract more users. However, Divvy must also concentrate on safety measure in using Divvy bikes at these seasons accordingly

R- Codes used for Project

R-Code to render Chord Diagram:

```
###########################################################################################################
# Author: Pradeep Sathyamurthy
# Team Mates: Ashrita, Meghana, Daniel
# Guiding Professor: Dr. Eli T Brown
# Course: CSC-465
# Project: Final Course Project for CSC-465, visualizing DIVVY dataset
# Part-1: Plotting Chord Diagram for visualize the driving pattern of Divvy users in Chicago using Divvy Bikes
# Date Created: 07-Nov-2016
# Date Last Modified: 20-Nov-2016
###########################################################################################################

install.packages("dplyr")
install.packages("circlize")
require(dplyr)
require(circlize)

# Create Fake Flight Information in a table
setwd("D:/Courses/CSC465 - Tableau - Data Visualization/Trails/Network Plot")
data.divvy_04 <- read.csv("Divvy_Edges_Sub_Dt_Dataset.csv",na.strings = NULL, stringsAsFactors = FALSE)

set.seed(2400000)
divvy_subscriber <- data.divvy_04[which(data.divvy_04$Type=="Subscriber"),]
divvy_Customer <- data.divvy_04[which(data.divvy_04$Type=="Customer"),]

origin_subs <- divvy_subscriber$Source
dest_subs <- divvy_subscriber$Target
origin_cust <- divvy_Customer$Source
dest_cust <- divvy_Customer$Target

df_subs = data.frame(origin_subs, dest_subs)
df_cust = data.frame(origin_cust, dest_cust)

# Create a Binary Matrix Based on mydf
matrix_subs <- data.matrix(as.data.frame.matrix(table(df_subs)))
matrix_cust<- data.matrix(as.data.frame.matrix(table(df_cust)))

# create the objects you want to link from to in your diagram
from_subs <- rownames(matrix_subs)
to_subs <- colnames(matrix_subs)
from_cust <- rownames(matrix_cust)
to_cust <- colnames(matrix_cust)

# Create Diagram by suppling the matrix
par(mar = c(1, 1, 1, 1))
grid.col2 = c("#e41a1c","#1b9e77","#7570b3","#e7298a","#a6761d","#d95f02","#66a61e","#e6ab02")

# Sort Link on Sector, this is very clear
# Direction connection
chordDiagram(matrix_cust, grid.col = grid.col2,annotationTrack = "grid", self.link = 2, link.border = 0)
circos.clear()

# Customize sector labels
circos.trackPlotRegion(track.index = 1, panel.fun = function(x, y) {
  xlim = get.cell.meta.data("xlim")
  ylim = get.cell.meta.data("ylim")
  sector.name = get.cell.meta.data("sector.index")
  circos.text(mean(xlim), ylim[1], sector.name, facing = "clockwise",
        niceFacing = TRUE, adj = c(0, 0.25))
}, bg.border = NA) # here set bg.border to NA is important
```

R-Code to render Choropleth Graph:

```
################################################################################################################
# Author: Pradeep Sathyamurthy, Daniel Glownia
# Team Mates: Ashrita, Meghana, Daniel
# Guiding Professor: Dr. Eli T Brown
# Course: CSC-465
# Project: Final Course Project for CSC-465, visualizing DIVVY dataset
# Part-1: Gathering right shape file for Divvy Dataset and fortifying DIVVY dataset with it
# Part-2: Plotting Choropleth for Chicago City based on DP Capacity and when each station went live
# Date Created: 21-Oct-2016
# Date Last Modified: 20-Nov-2016
################################################################################################################

# Setting up the project directory
setwd("D:/Courses/CSC465 - Tableau - Data Visualization/Trails/Choroplath")
install.packages("maptools")
install.packages("rgeos")
install.packages("Cairo")
install.packages("proto")
install.packages("ggmap")
install.packages("scales")
install.packages("RColorBrewer")

require(ggplot2)
require(rgeos)
require(maptools)
require(Cairo)
require(ggmap)
require(scales)
require(RColorBrewer)
set.seed(8000)

# Reading the shape file
chicago_neighbor.shp <- readShapeSpatial("geo_export_f83f4fda-cb0e-47e2-9ffa-859f2b78e325.shp")

# Checking its class
class(chicago_neighbor.shp)
# Checking the names associated with shape file
names(chicago_neighbor.shp)
# Checking the valus into it
print(chicago_neighbor.shp$objectid)
print(chicago_neighbor.shp$zip)
##create (or input) data to plot on map
num.neighbor <- length(chicago_neighbor.shp$objectid)

# Getting DIVVY dataset in
mydata_chi <- read.csv("Divvy_Stations_2016_Q1Q2_Modified.csv", stringsAsFactors = FALSE)
head(mydata_chi)
print(mydata_chi$shp_id)
print(chicago_neighbor.shp$objectid)

# fortify shape file to get into dataframe, one of the piece of code which took considerably long time
neigh.shp.f <- ggplot2::fortify(chicago_neighbor.shp, region = "objectid")
class(neigh.shp.f)
head(neigh.shp.f)

#merge with coefficients and reorder
merge.shp.coef3 <- merge(neigh.shp.f, mydata_chi, by="id", all.x=TRUE)
final.chi.plot <- merge.shp.coef3[order(merge.shp.coef3$order), ]

# ggplot for plotting choropleth using geom_polygon plottiing Choropleth filled with dpcapacity
c1 <- ggplot() +
  geom_polygon(data = final.chi.plot,
           aes(x = long, y = lat, group = group, fill = dpcapacity),
           color = "black", size = 0.25) +
  coord_map()+
  scale_fill_distiller(name="Docking Capacity", palette = "YlGnBu",trans="reverse", breaks = pretty_breaks(n = 5))+
  theme_nothing(legend = TRUE)+
  labs(title="DP Capacity in Chicago")

# Histogram to assist the choropleth plotted for visualization
c2 <- ggplot(data = mydata_chi,aes(dpcapacity)) + geom_histogram(binwidth = 7)+
  labs(x="Docking Capacity", y="Count of Divvy Stations") +
  ggtitle("Distribution of Docking Capacity in Divvy Station")

# saving the file to local working directory
ggsave(c1, file = "final_divvy_choroplath.png", width = 6, height = 4.5, type = "cairo-png")
ggsave(c2, file = "Divvy_DP_Distribution.png", width = 6, height = 4.5, type = "cairo-png")
```

Timetable

Phases	Description of Work	Start and End Dates
Phase One	Obtaining Dataset	18-Oct to 24-Oct 2016
Phase Two	Performing EDA on Divvy Dataset	25-Oct to 31-Oct-2016
Phase Three	Concentration on Event Labelling Research	01-Nov to 07-Nov-2016
Phase Four	Designing Time Series, Choropleth	01-Nov to 07-Nov-2016
Phase Five	Designing Heat Map, Network Graph	08-Nov to 14-Nov-2016
Phase Six	Extra Mile with Weather, Neural Networks	08-Nov to 14-Nov-2016
Phase Seven	Testing and Graph Validation	15-Nov to 20-Nov-2016
Phase Eight	Report Writing and Review	15-Nov to 20-Nov-2016
Phase Nine	Deliverable Submission	21-Nov-2016

Key Personnel

Team Liaison	Pradeep Sathyamurthy
Team Members	Ashrita, Daniel and Meghana
Professor	Dr. Eli T. Brown
Project for	CSC-465
Target Team	DePaul CDM

Deliverables

Final Report	PMAD_Final_Divvy_Dataset_Viz.pdf
Raw_Divvy_Dataset	Divvy_Trips_2016_01_06.csv, Divvy_Stations_2016_Q1Q2.csv, README.txt
Processed_Divvy_Dataset	divvy_merge.txt, Divvy_Stations_2016_Q1Q2_Modified.csv, Divvy_Edges_Sub_Dt_Dataset.csv, Street_Neighborood.xlsx
R	Prady_Divvy_Choropleth.R, geo_export_f83f4fda-cb0e-47e2-9ffa-859f2b78e325.shp, Divvy_Stations_2016_Q1Q2_Modified.csv; Prady_Chord_Dig_District.R, Divvy_Edges_Sub_Dt_Dataset.csv
Tableau	Divvy_Heat_Map.twb, Divvy_Heat_Map.twbx
Images	Graphs_Divvy_Dataset.rar
Team Members Summary	Ashrita_Individual Summary.docx Daniel_Individual Summary.docx Meghana_Individual Summary.docx

Project related reference links

[1] Detail about franchise Motivate which runs Divvy program https://en.wikipedia.org/wiki/Motivate_(company)

[4] Divvy Data set https://www.divvybikes.com/system-data

[5] Event Labelling Research Paper http://link.springer.com/article/10.1007/s13748-013-0040-3

[6] Divvy Bike Pricing Details https://www.divvybikes.com/pricing/annual

[7] Weather Data https://www.wunderground.com/history/airport/KORD/2016/6/28/MonthlyHistory.html?req_city=&req_state=&req_statename=&reqdb.zip=&reqdb.magic=&reqdb.wmo=&MR=1

[8] website to choose color palette http://colorbrewer2.org/#type=sequential&scheme=YlGnBu&n=3

[10] Zip Code Mapping http://www.zipmap.net/Illinois/Cook_County/Chicago.htm

[11] Sides Mapping http://www.seechicagorealestate.com/chicago-zip-codes-by-neighborhood.php

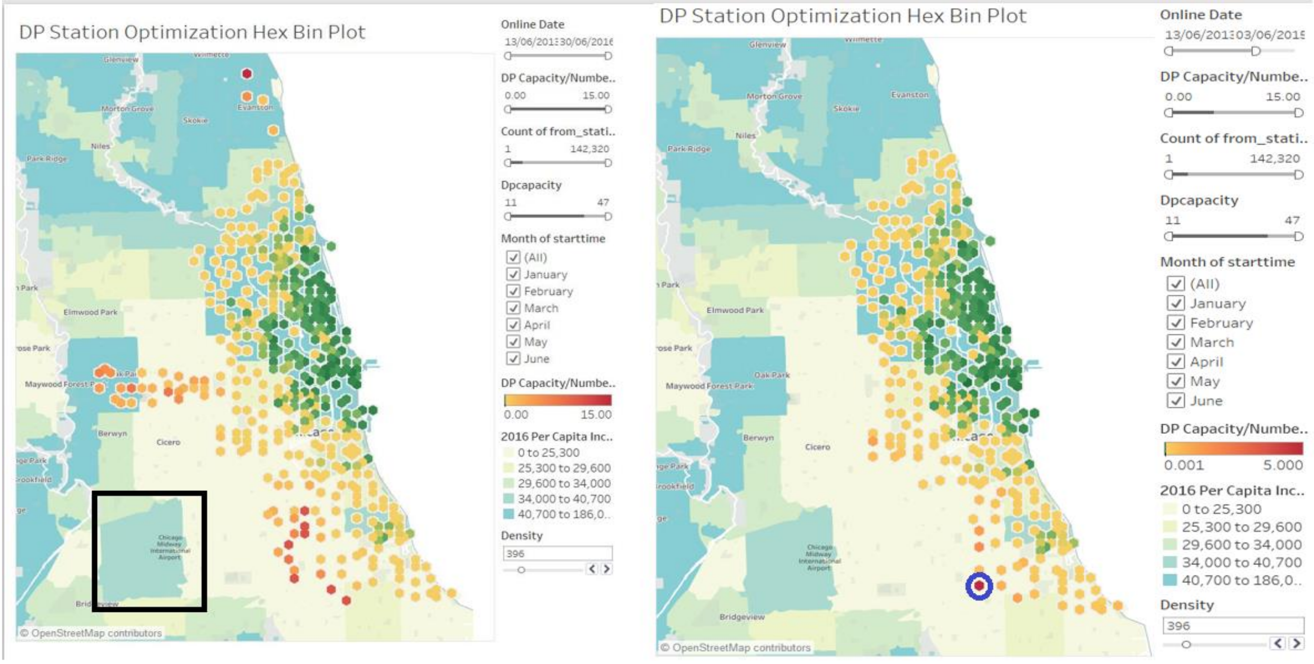
Appendix – Team Members Summary

Ashrita Contribution

I (Ashrita) had the reponsibility to find the relationship between the Subscribers and the Customers of the Divvy Cycles and the complete ownership of the collaboration of the weather data for 6 months to the 6 months Divvy Dataset.

I chose to make the data speak through Vizualizations in Tableau.

After Pradeep, Daniel and Meghana came up with their initial graphs I/we learned the various Divvy docks, their capacity and their location, who had the major share of the riding time (male/female) and also (Customers/Subscribers). So I wanted to see if income played any role in the pacing of the docks at various regions and if there were any docks which were highly strained and also scarsely used. I came up with the bel,ow geo-spatial vizualization.


DP Station Optimization Hex Bin Plot

Online Date
13/06/201:30/06/2016

DP Capacity/Numbe..
0.00 15.00

Count of from_stati..
1 142,320

Dpcapacity
11 47

Month of starttime
☑ (All)
☑ January
☑ February
☑ March
☑ April
☑ May
☑ June

DP Capacity/Numbe..
0.00 15.00

2016 Per Capita Inc..
0 to 25,300
25,300 to 29,600
29,600 to 34,000
34,000 to 40,700
40,700 to 186,0..

Density
396

DP Station Optimization Hex Bin Plot

Online Date
13/06/201:03/06/2015

DP Capacity/Numbe..
0.00 15.00

Count of from_stati..
1 142,320

Dpcapacity
11 47

Month of starttime
☑ (All)
☑ January
☑ February
☑ March
☑ April
☑ May
☑ June

DP Capacity/Numbe..
0.001 5.000

2016 Per Capita Inc..
0 to 25,300
25,300 to 29,600
29,600 to 34,000
34,000 to 40,700
40,700 to 186,0..

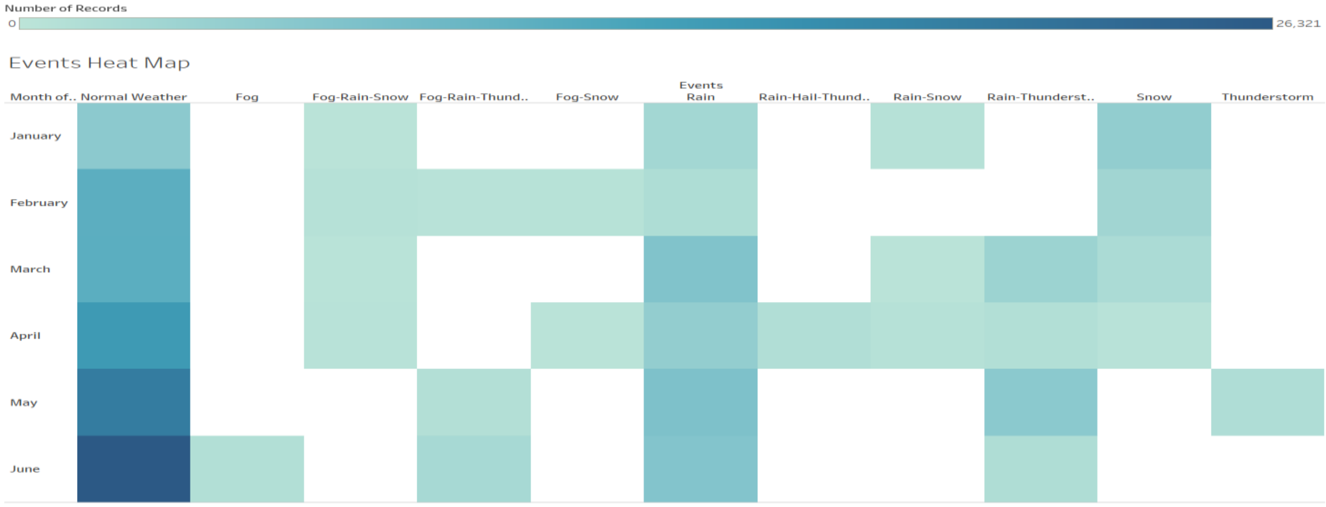
Density
396

The above Hex plot/ plots are for the same data for a different time duration. Meaning, this map shows the ratio (Docking capacity/ Number of rides from that station). The higher the ratio the lower is the usage and more towards red is the color. Similarly, lesser the ratio, more is the usage and darker is the Green color. The hex also contains the date from when the station went live. We have the stations dating back to June 2013 till June 2016. We know that if the station (divvy dock) is latest establishment then the bikes usage would be way less as compared to the older ones as it is the new customer base. So just for the sake of better comparison I have removed the docking stations that went online after June 2015 in the rightmost graph. We can see that as we travel south from the loop the hex tend to move towards red. Specifically the one rounded in blue was set up in April 2015 has a ratio or 5, which means despite being set up in spring of 2015 the usage is way less. In such a scenario, may be Divvy should consider moving some bikes from here to the heavy traffic area docks of the Divvy (like loop and near the lake shore drive area).

Suggestion: From the analysis, we found that the Divvy subscribers are mostly young male who earn well. So, considering that, we can see in the above graph that Chicago Midway Airport region (highlighted in black) has population with decent per capita income, so maybe Divvy should consider their next venture there.

Since we realised that the customers ride for a longer duration and along the lake front whereas the Subscribers ride for comparatively lesser duration and are spread across the city, my curiosity increased and wanted to know what were the peak hours for the bike riding. With the help of the heat map I was able to come up with a visualization which projected that the bike rides went high twice a day: In the morning 7:00 am to 8:00 am and In the evening between 4:00 to 6:00 pm. This was true across the months from January till June but the density increases as the year reaches to June from January. (***Refer to heat map for Busiest time of the day and busiest of 6 months***)

Since across the months the density (no. of rides) was increasing as the year was progressing I wanted to see if weather was a factor determining the rides. So, came up with another heatmap showing the different attributes of the weather across the 6 months duration:



The above heat map shows the various climatic condition across the first 2 quarters of 2016. The white spaces are for no data.

For example, we can see that, April had Normal weather, Fog-Rain-Snow, Fog-Rain-Thunder, Rain, Rain-Hail-Thunder, Rain-Snow, Rain Thunderstorm and Snow. So, during April the highest rides were recorded in the normal weather (as expected) and the next higher rides were during the events of rain. So, rain is not a big hindrance to the bike riders.

We can see across all the months that during rain, snow and rain-thunderstorm the number of rides were next to the normal weather number of rides. Also, as expected as the temperature rises the number of rides increases, i.e., compared to January, June had the higher number of rides.

Suggestion to Divvy: Cold temperatures do bring down the number of rides, but there are people who would still prefer to ride during those months. To attract more of such people, maybe they can promote by reducing the subscribing price or having a quarter passes for those who would want to ride during those seasons.

Reflection Summary:

The data speaks aloud through visualizations and allows us to form an opinion, find relationships and learn deeper aspects of any given dataset. For example, Heat map gives out the information using the density aspect, various methods of plotting can be implemented on a geospatial map (from scatter to line to hex to choropleth etc.) depending on what is there and what must be projected. It is just amazing to see the hybrid and cross breed graphs which involve various features in one and their correlations. The basic graphs like scatter plots, bar graphs, histograms etc. have their importance but are limited in their projections as compared to the complex ones, where these not only are easily comprehendible but also give out load of information. In other words, “Visualizations” is the language between a data Scientist and Business or any end user. They allow the Analyst to put his numbers to picture and the end user or strategist to understand them with ease and take the further desired actions.

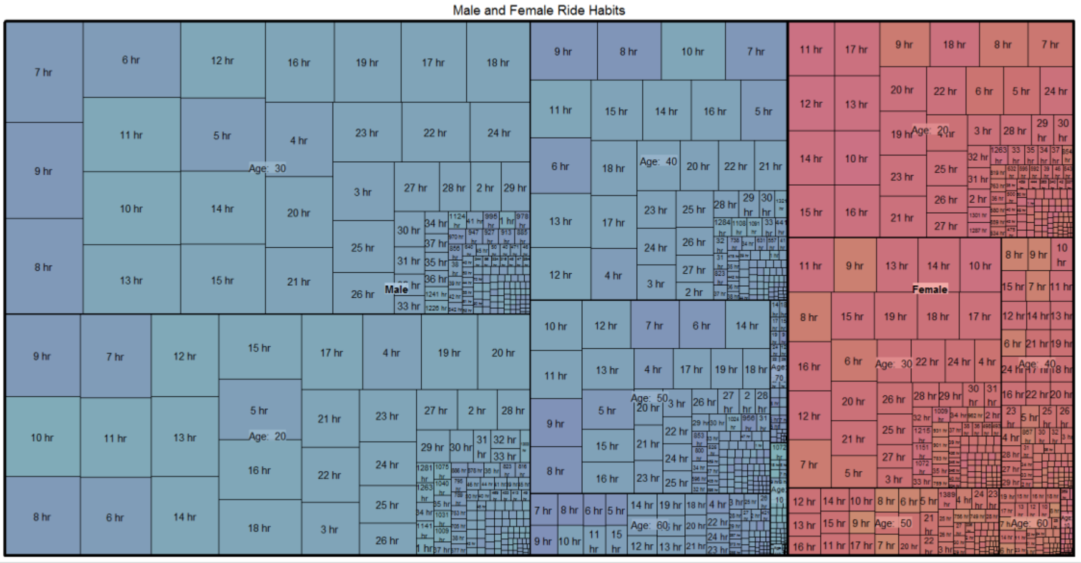
Apart from the visualizations, forming the group and working together was fun and lot of learning. It is very interesting to see how each person has a different perspective to a single point and are right to an extent. The discussions over the color palate, which Viz to choose, what to project etc. were all very informative and productive interactions. I learned a lot as everyone in my team comes from a different background and hence brought a lot to the table. When I was asked to present in the class, I gladly accepted as I am told that it is my strongest suit ☺. All the decisions were collective, collaborative and with everyone’s consent.

Above everything, thank you Professor for a great quarter and making us go through this team activity, I personally enjoyed and learnt a lot.

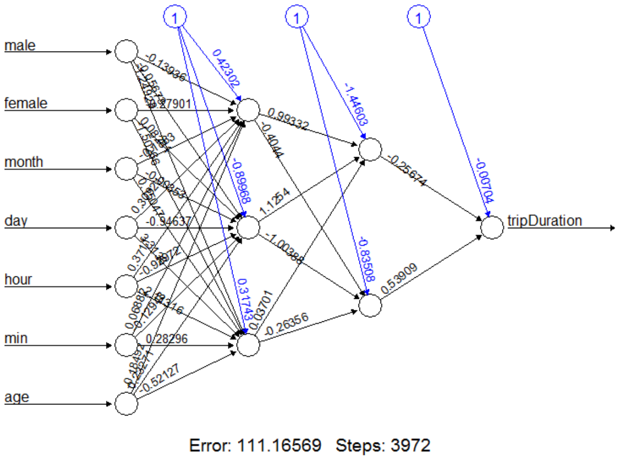
Daniel Contribution

Data Visualization - Daniel Glownia, Final Project Summary

Below is the first visualization that I have contributed to the team. This is a tree map of Male and female ride habits. The visualization was created to understand how long people take out the bikes. The colours indicate the categorical nature of the data and show similarities between the components. I learned one can transform the data to make it more categorical and as a result, create a tree map. I did this by group the ages and rounding the trip duration to the nearest 10th.



The next visualization is a diagram of a neural network. Our group did not choose to use this visualization because it was a simple network plot and used the built in nn plot(). Even though our group did not choose to use this visualization, I learned a lot from the process of making the visualization. I learned about the structure of a neural network. I also learned that a network plot would have been a good candidate for this processes. In addition, I learned that to have an affective visualization of a model that is difficult to understand, the visualization must tell and story that is relatable to the user. This is why my original visualization was not chosen for the final project submission. I represented the structure of the neural network but didn't tell and interesting story about the data with regards to the visualization.

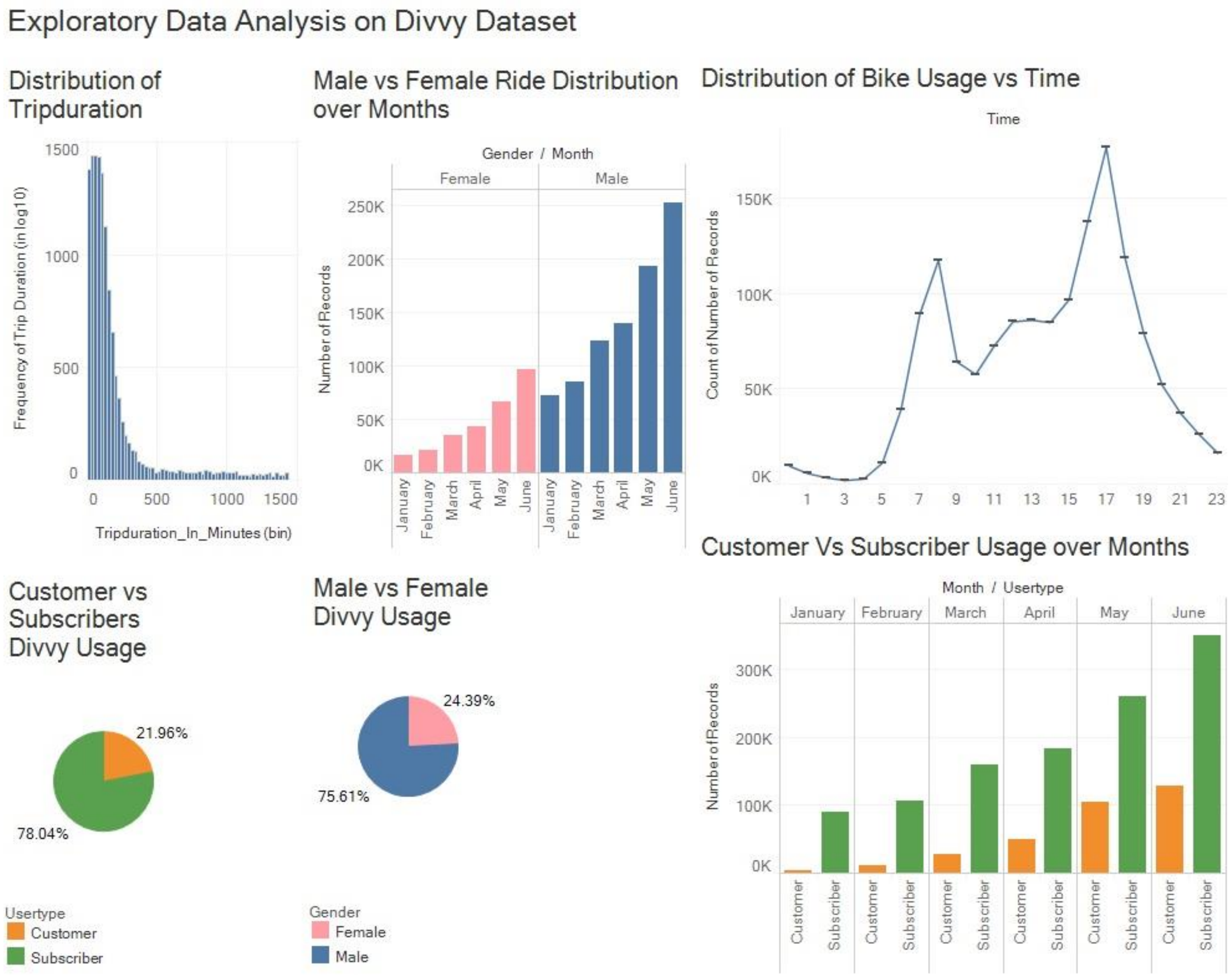


Lastly, I played the role of team member and R Developer. I created many of my visualizations in R and often collaborated with the team if we needed to finish a visualization in R. I also created the git hub and many collaboration tools for the team to use do that we could work on the project remotely.

Meghana Contribution

Exploratory Analysis of Divvy dataset – Individual Report - Meghana

I have taken up the responsibility for doing exploratory data analysis on the Divvy dataset for the first two quarters of 2016. I tried finding out any interesting patterns in the dataset by plotting all the variables, in different combinations which were analysed and worked upon by me and my groupmates. I had also taken up the responsibility of drafting out the minutes of meeting. In addition to the area/density graph and the time series graph in the report, I plotted the below graphs as part of the exploratory analysis.



- The most important take-away that we got from analysing the exploratory graphs is that ***studying the riding patterns and trends of customers and subscribers separately can bring out interesting insights from the data***. From the above graphs, it can be observed that the customers and the subscribers have differing riding patterns. Hence we focused more on the analysing that aspect of the data, taking into account the attributes such as weather, per capita income, age, trip duration, etc., and how they affected the riding patterns as a whole.
- Taking a look at the 'Distribution of Trip duration' graph, it can be observed that the trip duration of a large number of users, irrespective of them being subscribers or customers, is shorter and very few users are riding the bikes for longer periods of time. This has been correlated to the graph 'Distribution of Bike usage versus Time', where it can be observed that the most number of rides have been in the morning from 7 to 9 and in the evening from 4 to 5. This graph gave us the idea of further analysing the busiest time of Divvy bike usage among the users which has been dealt by Ashrita (Heat map).

- This idea also lead us to analyse the riding patterns of the subscribers and customers individually. The pie chart 'Customer versus Subscribers Divvy usage', gives an outlook on how the number of rides are divided among the customers and subscribers. The graph 'Customer versus Subscriber usage over Months' also drives home the same point as the pie chart that ***subscribers are taking more rides than customers***. The graph follows a pattern where the usage increased linearly from January to June.
- The pie chart 'Male versus Female Divvy usage' depicts how the riding patterns have been distributed among the subscribers according to their gender. It can be observed that the ***usage of divvy by females is substantially less when compared to males***. The graph 'Male versus Female Ride Distribution over Months' serves the same purpose as the pie chart in terms of who is riding more; but here the riding patterns can be seen clearly for each month. The ***riding patterns of both males and females seems to be the same, the only difference being in terms of the number of rides***. Further analysis regarding the riding habits of males and females according to their age and trip duration has been taken up by Daniel and can be seen in the report (Tree maps).
- ***The number of rides seem to increase linearly from January to June***, and subsequently showing that there are maximum number of rides in the month of June. This lead us to the idea that the riding behaviour may be affected by weather. This aspect of the data has been further analysed by Ashrita as can be seen in the report.
- The riding patterns have also been analysed from the perspective on how the docking stations are located all around Chicago and how that affects the users. This specific task has been taken up by Pradeep, seen in the report (Choropleth and Network graph).
- Hence we took the ideas from exploratory graphs, further analysed and built the subsequent graphs based on what we considered were interesting patterns in the Divvy dataset.

Pradeep Sathyamurthy

I am the group liaison for team PMAD. I basically designed below two graphs for this project in R:

1. **Choropleth**
2. **Chord Diagram**

I addition referred many research papers and visualization reference and tried to help our team to build a great visualization through which we were able to derive great insights. I also took over the responsibility of this final report writing.

I had an opportunity to work with a great team composing of Meghana, Ashrita and Daniel. Meghana and Ashrita where Tableau experts in team while Daniel and I brought expertise from R

Meghana was assisting me in data clean up and merging activities and actively participating in graph validation and testing. She kick started the project with her EDA. Though this is her first quarter, she proved herself a professional in working with Tableau and EDA research. She was very instrumental in assisting me with various activity in building this final report.

Ashrita was very instrumental in building complex graphs through tableau and she brought great ideas which helped us to build great visualization sighted above. She was also instrumental in getting the weather data and she along with Daniel went an extra mile in validating Divvy usage based on temperature.

Daniel on the other hand worked very hard in building the tree map to observe the Male vs Female riding behaviour which helped us to get a great insight on data showing usage of senior citizens in Divvy program. His research towards neural network was also mind blowing. Due to time constraint we were not able to train and test them. Neural network was built on purpose initially to find an optimized model to find the trip duration and from there on making its output used in interactive dashboard to keep up the user demand across the station and improving the navigation mechanism of Divvy bikes across Divvy Stations. Due to time constraints, he was successful in building the neural network model and we are planning to continue this research in our winter breaks.