**Problem-1:**

For the first set of milestones on your final project, meet as a group and complete the following

    a)   Record minutes of the discussion

**Date:** 26th October 2016

**Time:** 5.30 p.m. to 8.00 p.m.

**Agenda:**

> ➢ General discussion about datasets for the project and to decide about work allocation for the upcoming tasks.

**Discussion:**

1. The group liaison was chosen to be Pradeep. Team members are Arshita, Meghana and Daniel

2. Pradeep presented a dataset on '***Divvy Bikes***' and discussed about the feasibility of choosing this dataset for the final project and how these data are already visualized.

3. Ashrita presented the dataset '***Urban housing affordability***'. Discussed whether all the visualizations expected out of project could be produced from the dataset and realized that geospatial visualization was not possible because of no geographical data.

4. Daniel presented the datasets '***Automotive Data***' and '***Minst Dataset***' (posted in discussion board). Discussed about how the Automotive dataset is a huge, interesting dataset but would require knowledge about Machine Learning and Neural Networks.

5. ***Decided that the Divvy dataset is the most suitable for the project*** because all the visualizations expected for course project can be achieved with it with good learning.

6. Created a GitHub repository for the project named 'CSC-465-Data-Visualization-Divvy-Final-Project', all project artifacts will be here for Prof/team reference.

7. Decided to meet every Monday, Wednesday and Friday for project meeting.

**Work Allocation:**

1. ***Daniel***: R-mark down and choose a visualization method to work upon
2. ***Meghana:*** Minutes of Meeting and choose a visualization method to work upon
3. ***Ashrita:*** Data Validation by Ashrita and choose a visualization method to work upon
4. ***Pradeep:*** Project management and choose a visualization method to work upon

    b)   Choose a group liaison who will be responsible for interaction with me on group project issues

       PMADS's group liaison will be ***Pradeep Sathyamurthy***

c) Perform a cursory analysis on your data and decide on two distinct avenues of investigation for your data, then divide up the group into two subgroups that will investigate each direction. Think about how you might exploit the following types of visualizations for your project

- Basic exploratory visualizations, like we've covered in class so far (distributions, time series, etc.)
- 2D visualizations such as heat-maps, or rose plots (see next week's lecture)
- Geospatial visualization
- Network and cluster visualization (we will be talking about these in coming weeks)
- Volumetric visualization

We have sub-divided our team into two groups:
1. **Sub-Group-1:** Pradeep and Meghana
2. **Sub-Group-2:** Ashrita and Daniel

Responsibilities for these sub-groups are explained below:

➤ *Sub-Group-1* will have a primary responsibility of Basic Exploratory Visualization, Network or cluster visualization.

➤ *Sub-Group-2* will have a primary responsibility of 2D visualization and volumetric visualization

➤ *Both Sub-group-1&2* will work together in rendering one geospatial visualization as per current scope.

➤ *As an extra mile*, we are looking for collating weather data and merge it with current Divvy dataset in order to do a parallel geospatial analysis of bicycle usage based on weather condition.

d) Each subgroup should begin by searching the net for **ways that other people have visualized your kind of data in the past**. The idea here is to get inspiration for your visualizations and to help you get ideas about what types of visualizations you might be interested in investigating.

➤ In order to tackle this project, we would like to take insights from various research papers and work which was already done in past on this dataset and replicate to render similar result for our dataset as well.

➤ Each one of us in team have went through numerous websites and research papers in this well and below are few reference which we felt will be a good head start for this project:

- *http://www.redeyechicago.com/news/redeye-divvy-data-challenge-entry-26-20150330-htmlstory.html*

- *http://www.johnsavage.net/divvy.html*

- *http://transitized.com/divvy-vs-transit/*

- We have our collection of research papers downloaded for reference in our project git hub page for your kind reference.

**Problem-2:**

Prepare a first in-depth visualization of your data for the final project. You may use the graphs that your group produced in 1) as a start, or these graphs may suggest other directions that you decide to take. You should also spend some time working with basic summary views of the data and the fields. Remember that one of your visualizations for the project may be a collection of more basic charts that explores the scope of the data, its fields, and how those fields are related. **Note that if you do choose to do this for one of your visualizations, you should spend some time thinking about layout and structuring this display to best show the summary statistics**. But then build on these first simple looks at the data.

For this problem, you should make a good start on each of your visualizations, though if you wish as a group, you may decide to concentrate more on two or three at this point, leaving the remaining visualization in a sketchier "proof of concept" form for the moment. Be clear in your report if this is what you are doing. As you are building your graphs, use the concepts from class and the final project requirements posted to D2L as guidelines.

Remember that visualization is an iterative process, so you should be prepared to go through at least two or three drafts of your graphs. Create the graph and then analyse the graph for its content and communication. Then revise the graph appropriately. When you submit them for the final, your graph or graphs should be clean, clutter free with all necessary scales and legends to indicate the data being graphed. They should adhere to the design criteria we've covered in class.

Your answer to this question should contain a write-up that explains the direction you are taking with each of your visualizations. Include a description of how you revised your graph/graphs that explains the revision in terms of how it improved in either an exploratory or explanatory way.

**Exploratory Visualization:**

Real-world data is known to be imperfect, suffering from various forms of defects such as sensor variability, estimation errors, uncertainty, human errors in data entry, and gaps in data gathering, etc., Since our Divvy dataset is an information obtained from network Kiosk we feel it is better to do some exploratory data analysis. Main focus as part of this analysis is to visualize the hypothesis in data.
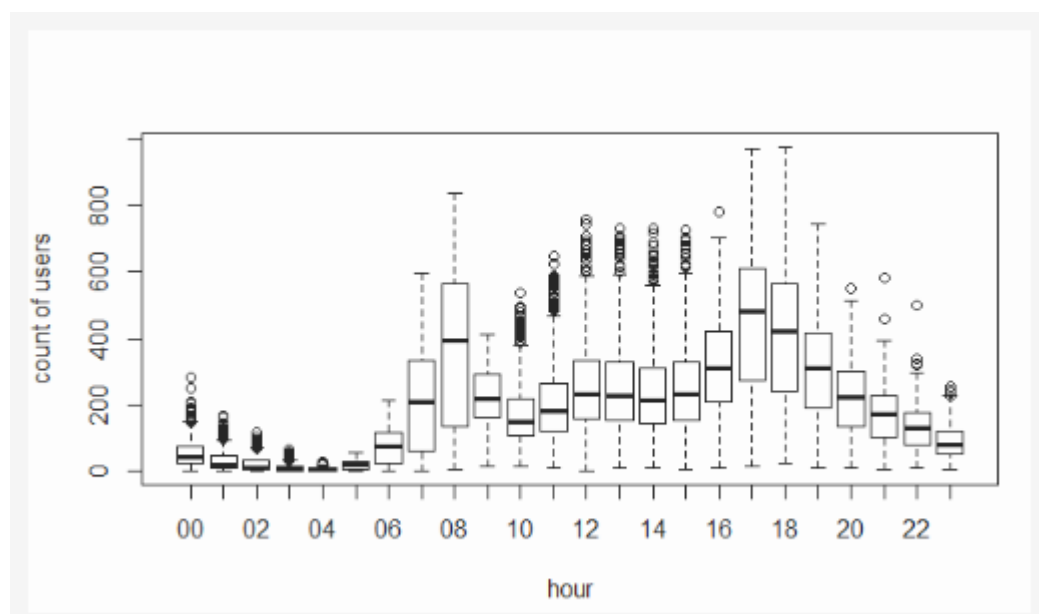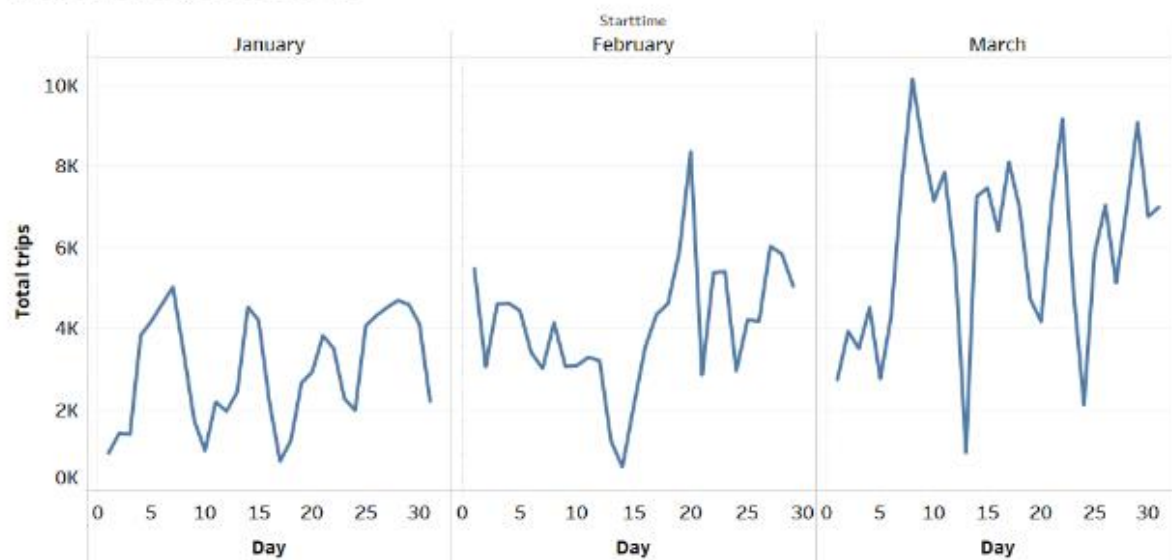
**2D Visualization:**

As part of this we are trying to explore the busiest month in last two quarter and days when the usage of DIVVY bikes is more in form of heat map. We also like to sort few correlations between variables and would like to provide an inference from it. A sample Location ID vs Docking Capacity plot is shown below, more intrinsic 2D visualization will be presented in next group work:

## Docking capacity in each Location
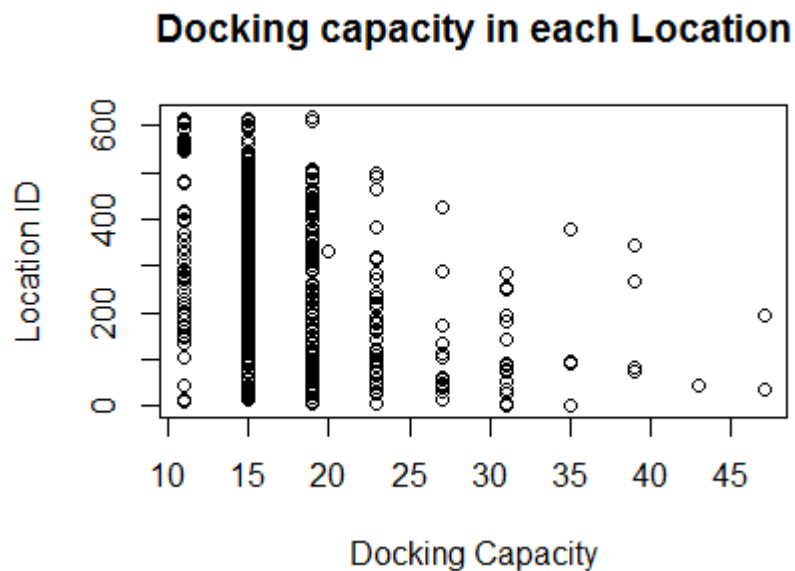


Riding pattern of Customer vs Subscribers will be shown in form of line graph as below, this is not an exact representation of data which we are trying to visualize, once the data is completely cleaned and processed we will render the final graph which will be submitted as part of GW2:

Linegraph of days vs total trips



The trend of sum of Number of Records for Starttime Day broken down by Starttime Month.

**Network and Cluster Visualization:**

Visualization that will show the most travelled path of commuters will be visualized from this graph, our sample graph for this visualization would look like below, final report will carry a high definition image submission with necessary source code in place for the Divvy data:
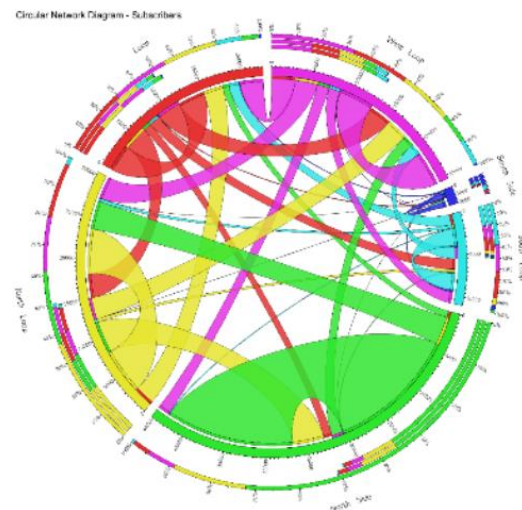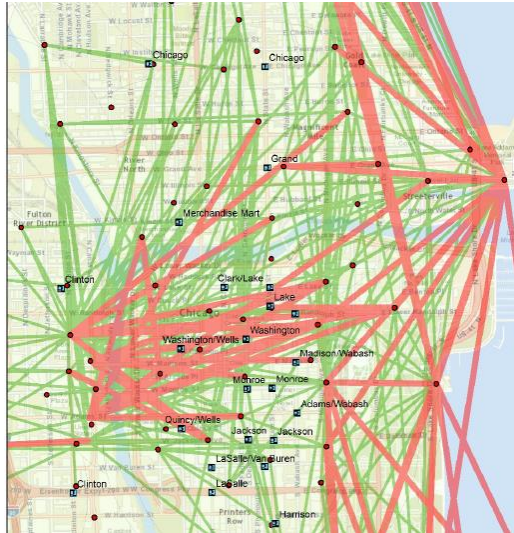


**Geo-spatial Visualization:**

We will work on few geo-spatial graphs to visualize this data, below image shows the map of Chicago. We can see from this map that the non-subscribers (blue representation) are the customers who take/ride the divvy bikes for longer duration along the lake. They also usually originate near the loop. Whereas the subscribers use the bikes for lesser duration comparatively and are spread inside the city. So, we can deduce that may the non-subscribers are mostly the tourists who take the bike usually from the loop and ride along the lake front trail.

types of customers and the region where bikes used
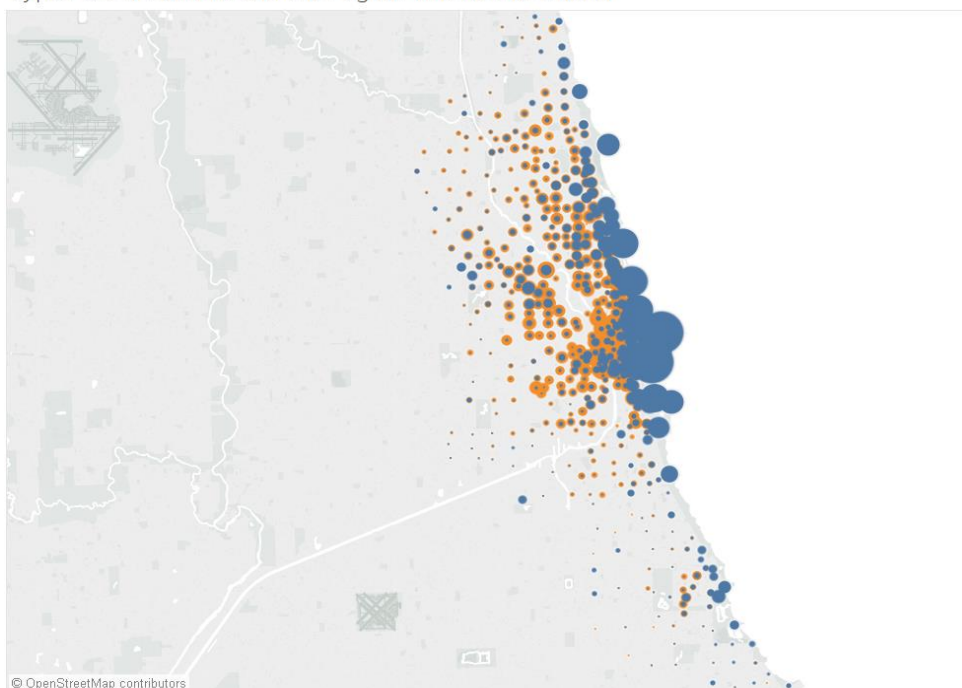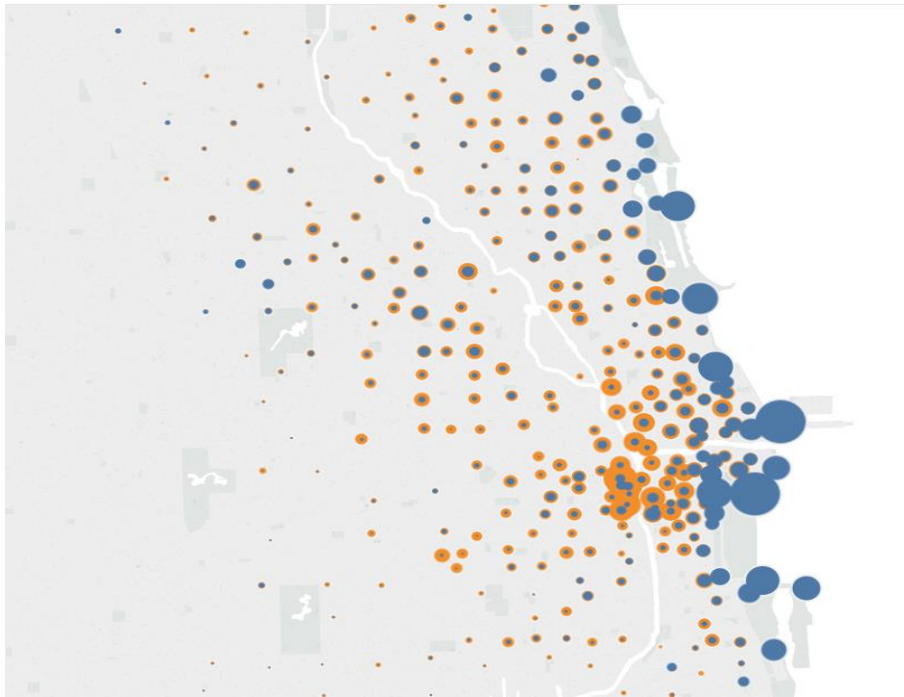
Again this is not an exact final representation of data which we are trying to visualize, once the data is completely cleaned and processed we will render the final graph which will be submitted as part of GW2.