## Chicago Divvy Bicycle Sharing Data (source Kaggle.com)

Data come from Chicago Divvy bicycle sharing system as well as the weather information in Chicago. The dataset used in this class contains a sample of about 730K trips taken during 2017.

The variables in the dataset are as follows:

- **Trip ID's**
- **Month**
- **Day**
- **User type** ("Customer" is a rider who purchased a 24-Hour Pass with unlimited 3-hour rides; "Subscriber" is a rider who purchased an Annual Membership allowing unlimited 45-minute rides)
- **Gender** (F, M)
- **Starttime**: day and time trip started, in CST
- **Starthour**: hour trip started in military time (0-24)
- **Stoptime**: day and time trip ended, in CST
- **Tripduration**: time of trip in minutes
- **from_station_id**: ID of station where trip originated
- **from_station_name**: name of station where trip terminated
- **latitude_start**: start station latitude
- **longitude_start**: start station longitude
- **dpcapacity_start**: number of total docks at start station
- **to_station_id**: ID of station where trip ended
- **to_station_name**: name of station where trip ended
- **latitude_end**: end station latitude
- **longitude_end**: end station longitude
- **dpcapacity_end**: number of total docks at end station
- **temperatures** (F degrees)
- **humidity** (0-100%)
- **visibility**: 0 (poor) to 10 (optimal)
- **wind speed** (miles per hour)
- **Precipitation** (inches)
- **Rain** (0 no, 1 yes)
- **Conditions** (weather conditions: clear, drizzle, fog, haze, heavy rain, etc..)

### Possible research questions:

1) What's the busiest day of the week?
2) What's the busiest time of the year?
3) What's the typical duration of a trip?
4) What about the shortest or longest?
5) How does temperature affect use of Divvy bikes?
6) How about when it rains? Do people ride bikes in the rain?
7) In general how do weather conditions affect the use of Divvy bikes?

8) Are women and men using the bikes differently?
9) Is there a different pattern of usage between customers (occasional users) and subscribers?
10) What are the areas in Chicago where people rent Divvy bikes most often? Is it the same for occasional users or subscribers? Does it change in the morning or at night?
11) Is there a seasonal pattern in the use of Divvy bikes during the day?

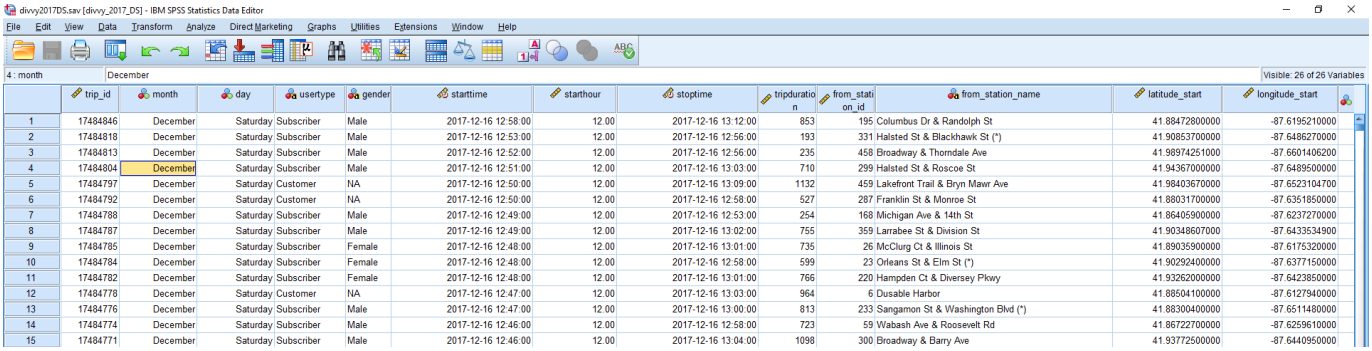
## 1. Dataset Preparation

Let's import the file into SPSS

- Open SPSS on your desktop
- Select File > Import Data > CSV data and upload the "divvy2017DS.csv" dataset

## 2. A glimpse of the data

Let's take a look at the snapshot of top rows in the Divvy dataset




| | trip_id | month | day | usertype | gender | starttime | starthour | stoptime | tripduration | from_station_id | from_station_name | latitude_start | longitude_start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 17484846 | December | Saturday | Subscriber | Male | 2017-12-16 12:58:00 | 12.00 | 2017-12-16 13:12:00 | 853 | 195 | Columbus Dr & Randolph St | 41.88472800000 | -87.6195210000 |
| 2 | 17484818 | December | Saturday | Subscriber | Male | 2017-12-16 12:53:00 | 12.00 | 2017-12-16 12:56:00 | 193 | 331 | Halsted St & Blackhawk St (*) | 41.90853700000 | -87.6486270000 |
| 3 | 17484813 | December | Saturday | Subscriber | Male | 2017-12-16 12:52:00 | 12.00 | 2017-12-16 12:56:00 | 235 | 458 | Broadway & Thorndale Ave | 41.98974251000 | -87.6601406200 |
| 4 | 17484804 | December | Saturday | Subscriber | Male | 2017-12-16 12:51:00 | 12.00 | 2017-12-16 13:03:00 | 710 | 299 | Halsted St & Roscoe St | 41.94367000000 | -87.6489500000 |
| 5 | 17484797 | December | Saturday | Customer | NA | 2017-12-16 12:50:00 | 12.00 | 2017-12-16 13:09:00 | 1132 | 459 | Lakefront Trail & Bryn Mawr Ave | 41.98403670000 | -87.6523104700 |
| 6 | 17484792 | December | Saturday | Customer | NA | 2017-12-16 12:50:00 | 12.00 | 2017-12-16 12:58:00 | 527 | 287 | Franklin St & Monroe St | 41.88031700000 | -87.6351850000 |
| 7 | 17484788 | December | Saturday | Subscriber | Male | 2017-12-16 12:49:00 | 12.00 | 2017-12-16 12:53:00 | 254 | 168 | Michigan Ave & 14th St | 41.86405900000 | -87.6237270000 |
| 8 | 17484787 | December | Saturday | Subscriber | Male | 2017-12-16 12:49:00 | 12.00 | 2017-12-16 13:02:00 | 755 | 359 | Larrabee St & Division St | 41.90348607000 | -87.6433534900 |
| 9 | 17484785 | December | Saturday | Subscriber | Female | 2017-12-16 12:48:00 | 12.00 | 2017-12-16 13:01:00 | 735 | 26 | McClurg Ct & Illinois St | 41.89035900000 | -87.6175320000 |
| 10 | 17484784 | December | Saturday | Subscriber | Female | 2017-12-16 12:48:00 | 12.00 | 2017-12-16 12:58:00 | 599 | 23 | Orleans St & Elm St (*) | 41.90292400000 | -87.6377150000 |
| 11 | 17484782 | December | Saturday | Subscriber | Female | 2017-12-16 12:48:00 | 12.00 | 2017-12-16 13:01:00 | 766 | 220 | Hampden Ct & Diversey Pkwy | 41.93262000000 | -87.6423850000 |
| 12 | 17484778 | December | Saturday | Customer | NA | 2017-12-16 12:47:00 | 12.00 | 2017-12-16 13:03:00 | 964 | 6 | Dusable Harbor | 41.88504100000 | -87.6127940000 |
| 13 | 17484776 | December | Saturday | Subscriber | Male | 2017-12-16 12:47:00 | 12.00 | 2017-12-16 13:00:00 | 813 | 233 | Sangamon St & Washington Blvd (*) | 41.88300400000 | -87.6511480000 |
| 14 | 17484774 | December | Saturday | Subscriber | Male | 2017-12-16 12:46:00 | 12.00 | 2017-12-16 12:58:00 | 723 | 59 | Wabash Ave & Roosevelt Rd | 41.86722700000 | -87.6259610000 |
| 15 | 17484771 | December | Saturday | Subscriber | Male | 2017-12-16 12:46:00 | 12.00 | 2017-12-16 13:04:00 | 1098 | 300 | Broadway & Barry Ave | 41.93772500000 | -87.6440950000 |

The data set contains 26 variables

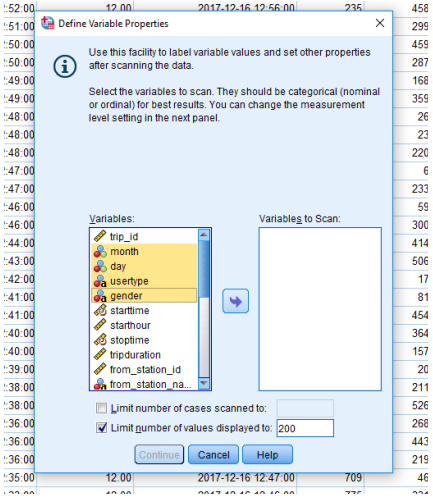
**EXERCISE: Identify the type (nominal or scale) for the following variables: Starthour, humidity, rain, latitude_start, conditions**

## 3. Let's check if there are invalid values

Let's take a look at the variable properties. We want to find out if there are problems with the data (e.g. invalid values or missing records). We will use the "Define Variable Properties" function in SPSS:

Select "Data > Define Variable Properties … " in the top Menu to open the app. Move the variables to be analyzed into the Variables to Scan box, and click Continue.

A new window will open up with information about each variable. We can use the window to change the variables properties, for instance identify missing values. See example in class.



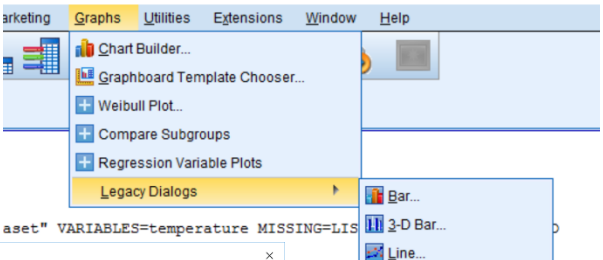

**EXERCISE: Analyze the properties for starthour, humidity, precipitation and rain**

## 4. Let's get some insights about the data

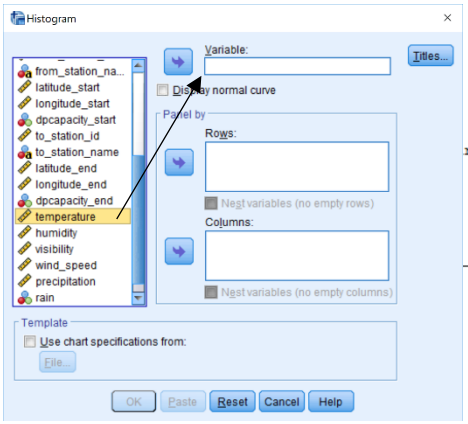
We can start by analyzing the distributions of individual variables. The **distribution** of a variable tells us what values it takes and how often it takes those values.

QUESTIONS: Do many people rent bikes when it is below 32F degrees? Or above 100F degrees? What's the range of temperatures associated to higher numbers of trips?

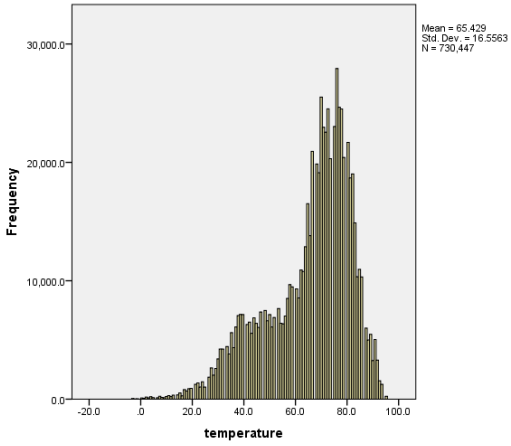
The distribution of a variable is displayed by a **histogram**. In SPSS we can create a histogram by selecting Graphs > Legacy Dialogs > Histogram.




In the Histogram window, move the variable to be analyzed to the "Variables" box and click ok




The following histogram will be created. The range of temperatures is divided into small intervals and each bar represents the number of observations (trips) with temperatures that fall in that interval.




The histogram shows a peak between 60F and 85F, indicating that divvy bike rentals are high during warm days with temperatures between 60F and 85F. The plot also shows a smaller "bump" for temperatures between 40 and 60 degrees. Very few people rent bikes when it is below 32F.

**EXERCISE: Analyze the distribution of starthour and tripduration and answer the following questions:**

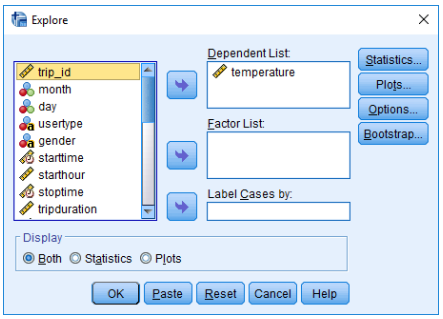
1) **During which hours are people more likely to rent bikes?**
2) **Do people rent bikes in the middle of the night (from midnight to 4am)?**
3) **What's the typical duration of a Divvy trip?**
4) **Are most rides below 45-minutes?**

## 5. Distributions can be described by summary statistics

1. Center: *where most data fall around*
   Measures:
   - Average ave=$(x_1+x_2+x_3+\ldots+x_n)/n$
   - Median defined as the value $x_i$ s.t. 50% of points fall below it
2. Variability: *indicating noise in data*
   Measures:
   - Variance VAR= $((x_1\text{-ave})^2+(x_2\text{-ave})^2+(x_3\text{-ave})^2+\ldots+(x_n\text{-ave})^2)/n$
   - Range =max-min
3. Outliers: *unusual observations that are far from the rest*
   Points that lie far from the center of the distribution. Possible indicators are points that fall under the tails of the distribution, say below the 1st percentile or above the 99th percentile.

QUESTIONS: What's the median temperature in the Divvy trips sample? What's the min and max temperature? What are the highest temperatures observed in 5% or more trips (95th percentile)?

In SPSS, select Analyze > Descriptive Statistics > Explore… . Move the variable(s) that you want to analyze into the Dependent List box.




Click on the "Statistics…" button to select the statistics to display and on the "Plots…" button to select the plots (boxplot, histogram &/or stemplot).

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| temperature | Mean | | 65.429 | .0194 |
| | 95% Confidence Interval for Mean | Lower Bound | 65.391 | |
| | | Upper Bound | 65.467 | |
| | 5% Trimmed Mean | | 66.324 | |
| | Median | | 70.000 | |
| | Variance | | 274.112 | |
| | Std. Deviation | | 16.5563 | |
| | Minimum | | -2.9 | |
| | Maximum | | 95.0 | |
| | Range | | 97.9 | |
| | Interquartile Range | | 21.1 | |
| | Skewness | | -.887 | .003 |

**Percentiles**

| | | Percentiles | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| Weighted Average (Definition 1) | temperature | 33.100 | 39.900 | 55.900 | 70.000 | 77.000 | 82.900 | 86.000 |
| Tukey's Hinges | temperature | | | 55.900 | 70.000 | 77.000 | | |

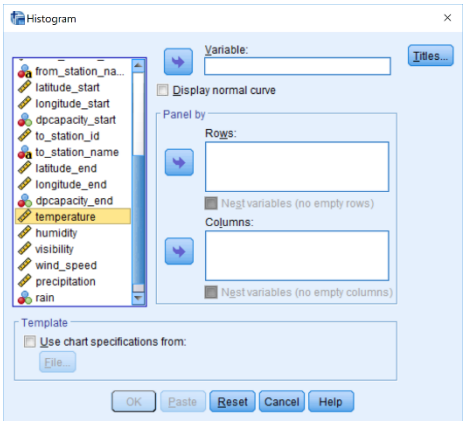
**EXERCISE: What's the average duration in minutes of a divvy bike trip? What's the minimum duration? What's the top 25% trip duration?**

## 6. We often want to learn about variable interactions, and how patterns change within different groups of observations

QUESTION: Are the occasional customers more like to rent a bike when it is warmer?

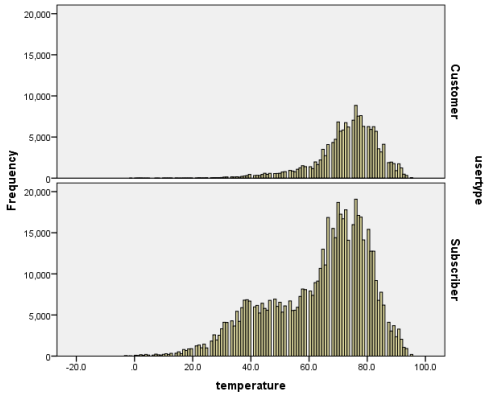
This question requires us to analyze the interaction between “temperature” and “usertype”. We can build two separate histograms for the temperatures of trips taken by the two user types.

In SPSS, select Graphs > Legacy Dialogs > Histogram from the top menu. In the histogram box, move “temperature” to the Variables box, and move the grouping variable “usertype” to the “Rows” box.



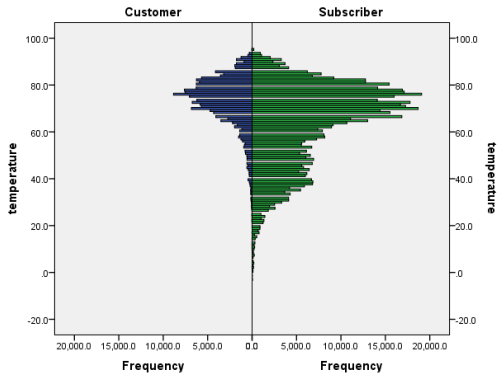

The two histograms display the distributions of temperatures for trips taken by occasional users (customer) or by subscribers.

The distributions are not the same. The trips taken by occasional users (customer) have temperatures in the higher range, while the subscribers’ trips are taken in a wider range of temperatures with a significant number of trips

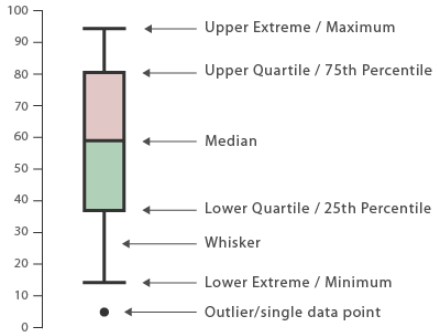



The image below is another option called Population Pyramid showing the two histograms on the same axis.

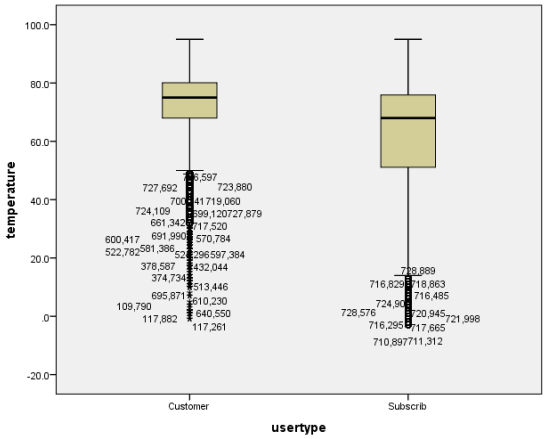
This is obtained in SPSS using the Chart Builder feature. Select Graphs> Chart Builder to open the interactive window to create the graph.



**Boxplots** are very useful tools to compare distributions of values in different groups of observations, as distributions are displayed side by side.




**Boxplot of temperatures for customers and subscribers**

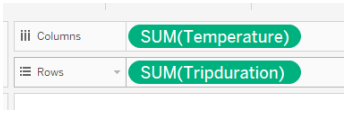



**EXERCISE: Are Divvy trips shorter in time when it rains? Is there a difference in trips depending on the day of the week? Do occasional users ride for a longer time than subscribers do?**

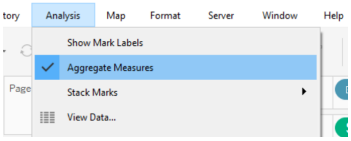
## 7. Scatterplots

We're going to transition to Tableau for SPSS now because we're going to talk about data visualization and this is a purpose-built tool. I'm going to give a live tour of the software to you, but I also want to point out that you can get this software for yourself for free as a student from Tableau's website. They have excellent video tutorials that will show you many more features than we can cover today. https://www.tableau.com/academic/students

So far we looked at relationships between numeric variables and categorical ones, but what about relationships between numerical variables. As we saw in the presentation, scatterplots are typically the right choice. To create a single scatterplot in Tableau, drag a numerical variable to the *Columns* mapping, and one to the *Rows* mapping. For example, you could use *Tripduration* and *Temperature*:

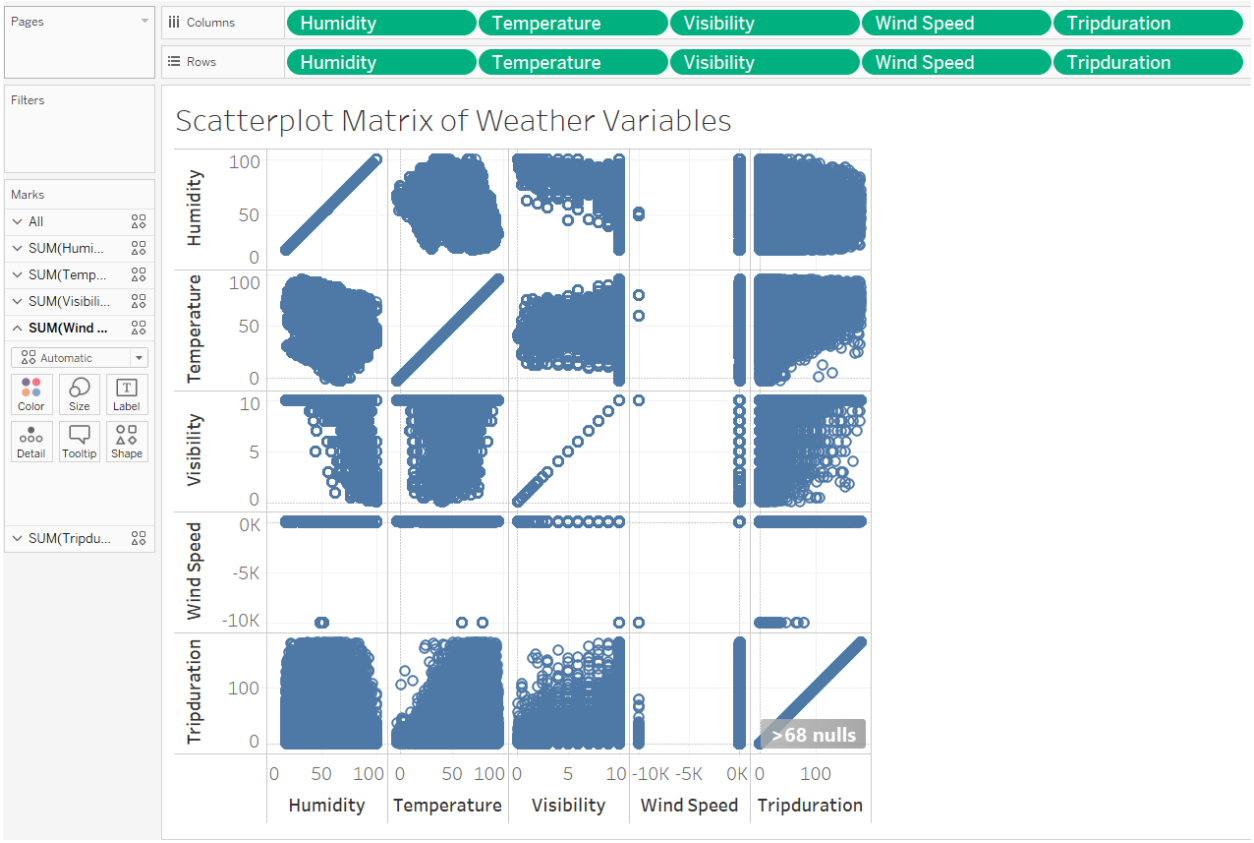



Notice that there is only one dot on your graph. This is because of aggregation. Aggregation means using a single number to represent multiple data points. Typically, summary statistics are used as the aggregations measures. In this case, Tableau has used the sum to create an aggregate measure of the temperature. To change this, we unselect *Aggregate Measures* from the *Analysis* menu.



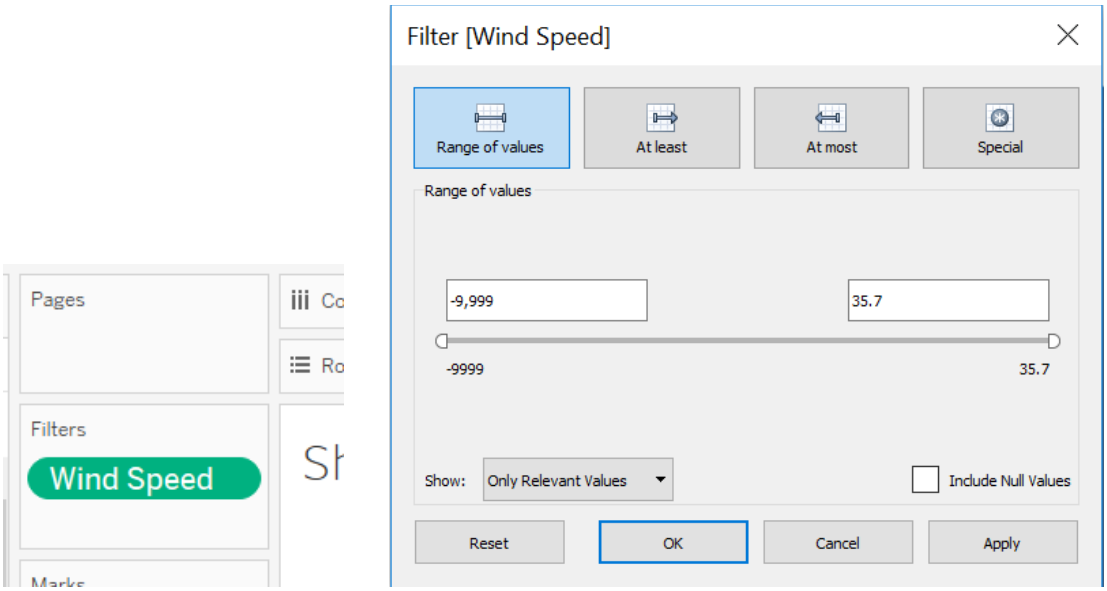

It's a pretty crowded scatterplot, but it does show us that there are longer trips being taken when the temperature falls within a certain range. If we want to see multiple relationships at the same time, we

can add more variables to the lists:



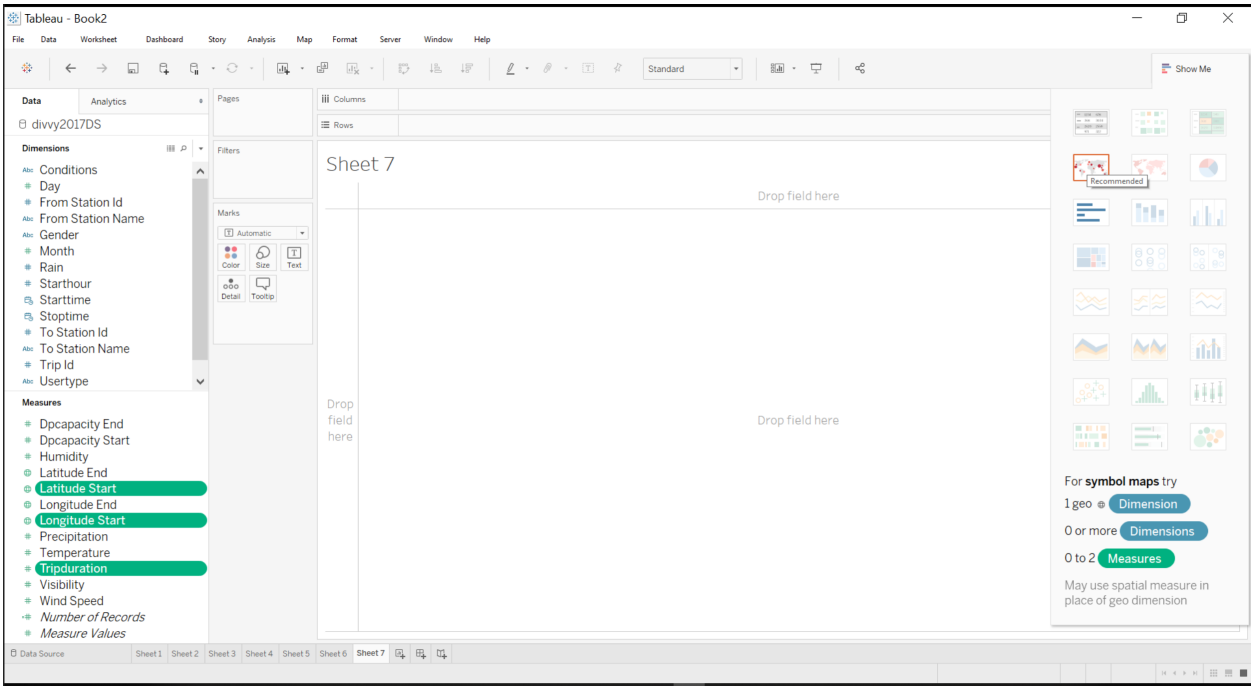

This view is called a scatterplot matrix. **Why are the graphs along the diagonal of this grid of scatterplots so bland? Which weather variables have an apparent relationship to the trip duarion? Which two weather variables are related?**

Right away we notice something is funny with *Wind Speed*. Why is it reaching nearly -10,000 a few times? We can add a filter on this variable to remove this issue by dragging *Wind Speed* to the filters box and then filtering out the negative values with the slider.



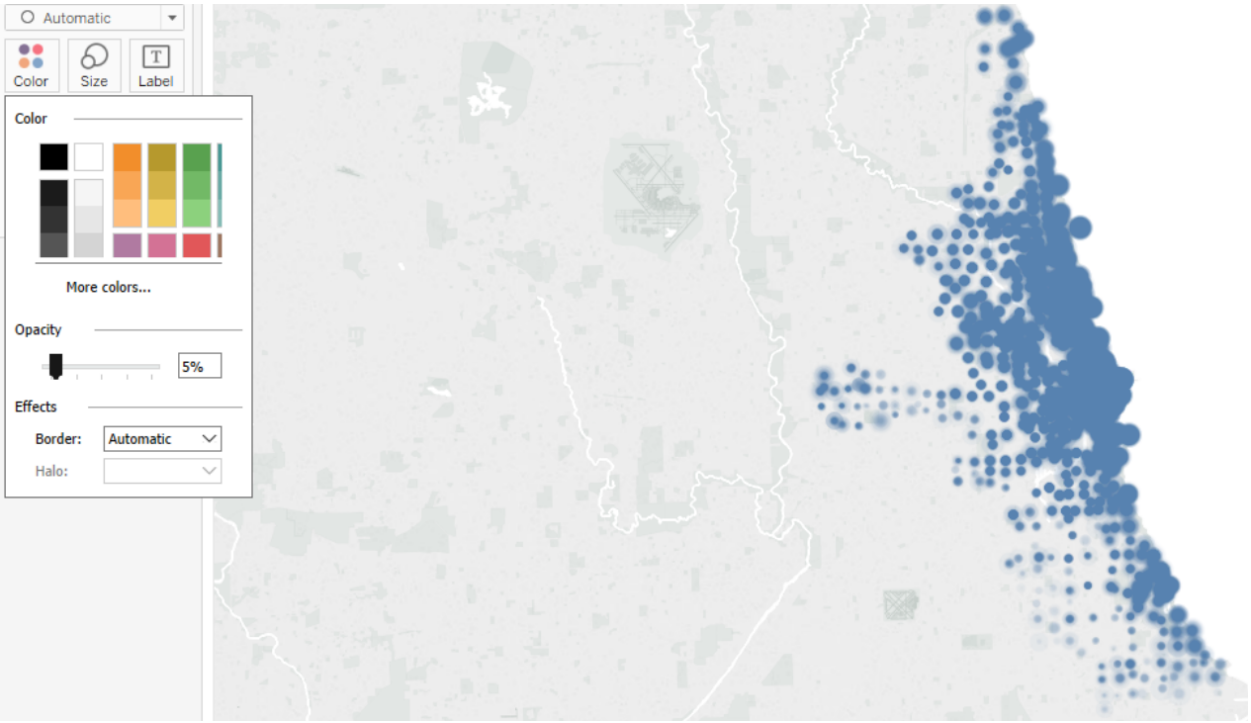

## 8. Mapping

Tableau also makes it easy to see geographic patterns. We'll use the *Show Me* feature of Tableau to get a map of where all the trips started from. By holding down the control button on the keyboard, you can select multiple variables on the left. When the selection matches a type of graph in the Show Me menu, you can click it.




You'll need to turn off aggregation. Notice that there are a lot of dots on top of each other because it is graphing one dot per trip. There is a neat trick for dealing with this called transparency. The transparency of a color is how much you can see through it. By making the color of a dot highly transparent (low opacity) you can see right through it to the next dot. Only when multiple dots are on

top of each other do we see the full blue. Click the color setting in the *Marks* panel and you can change the setting.

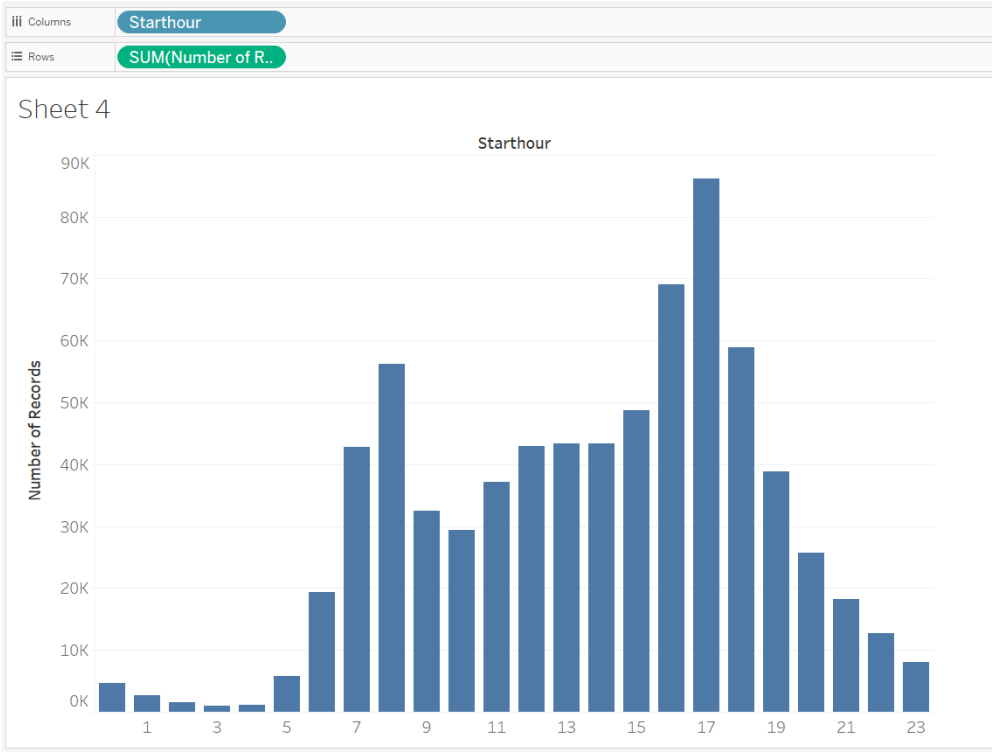



Add a filter based on *To Station Name* so you can see where people started from and how long their trips were when going to certain places in the city. The locations list includes Adler Planetarium, the Field Museum, and the Museum of Science and Industry.

## 9. Time plots

Let's get a sense for how time of day affects ridership. There is a variable called *Starthour* that tells us at what hour of the day a ride started. There are 24 discrete values for this, but Tableau still thinks it is simply a numeric variable. If you try to use it to create a graph with *Number of Records* to see how many rides are taken at different times of day, you will end up with a scatterplot of sorts. Right click the

variable and select *Convert to Dimension* to change it. Now we can easily create a bar graph.




We have the actual time of each ride in the variable *Starttime*, so we could actually look for a smoother view based on time. If you create the graph with this variable, you'll notice the graph doesn't look how you'd expect because of what Tableau assumes you mean when mapping a time variable. Here is a view of the variable mapping to Column:

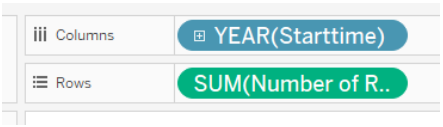
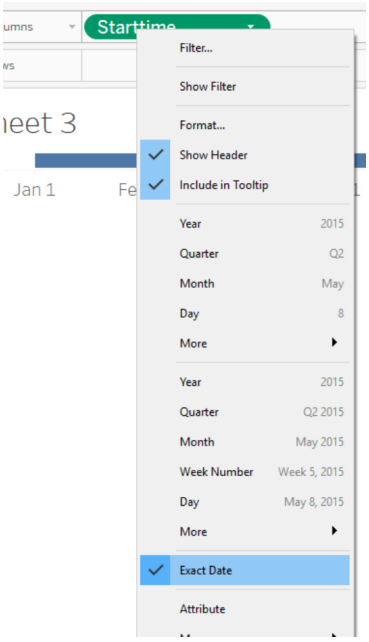
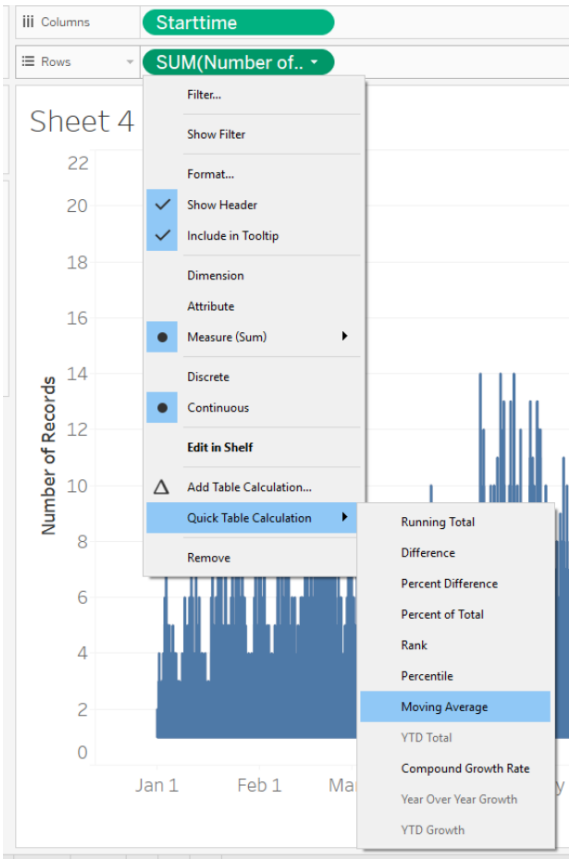


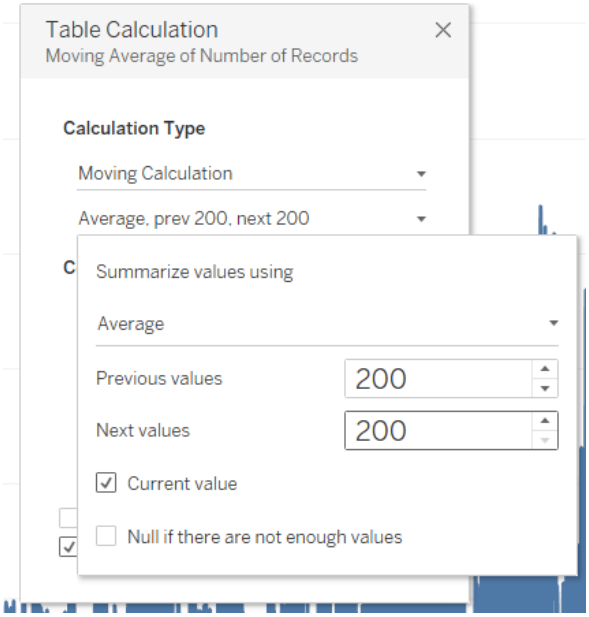

Tableau has extracted the year from the *Starttime* variable. Since these rides in our data are all the same year, that's not very interesting. If we click the dropdown menu for the bean there are a number of options like using month or hour instead, but let's start with the actual time by choosing *Exact Date*.




Now we get quite an unwieldy plot. There are so many points plotted in this range that we need smoothing. To smooth this out by averaging out variation, use the dropdown button on *Number of Records* and under *Quick Table Calculation* choose *Moving Average.*

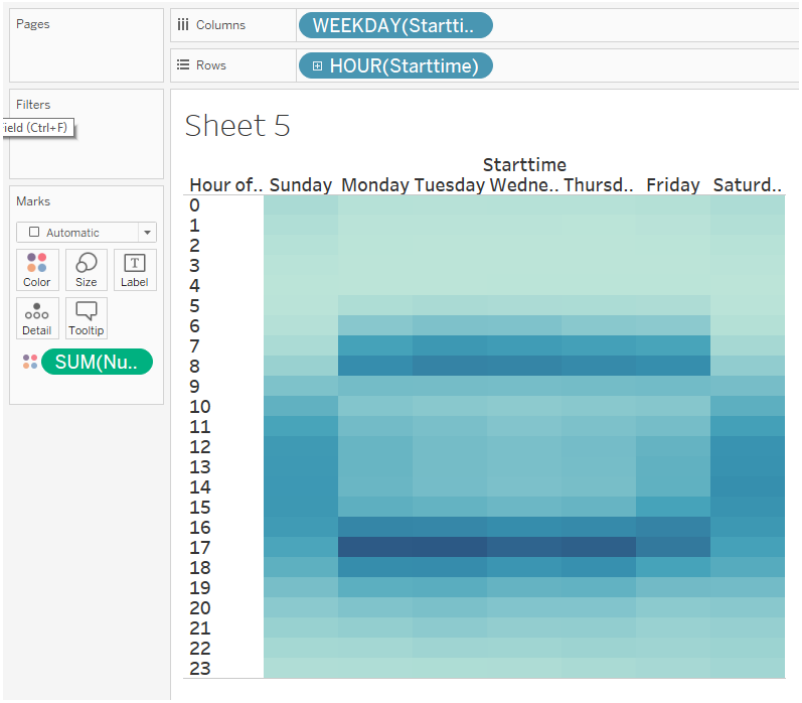


Go back to the menu to *Edit Table Calculation* and change the settings for how wide the smoothing window is (making these numbers bigger will make the line less detailed but more smooth).




The data is pretty dense, but we can now see that while there is mostly short periodic signal probably from days and weeks, there are a couple odd spikes that survived the smoothing. Why would there be spike in usage in the Winter? We can get a bit more context by drilling into the data. Mapping a variable to *Detail* means when we mouseover a point on the graph, we will see the value. **Map some of the variables to *Detail* and try to figure out what caused the usage spkes.**

Rather than look at a single time horizon as a line graph, we might like to see more sophisticated time patterns. One technique for that is a heatmap. We lay out time spatially and use color to encode high or low values. *Heatmap* is one of the options in the *Show Me* bar, but you can also get it by using categorical variables for column and row with a numeric variable for color. Try mapping the hour and weekday of the *Starttime* to columns and rows, and the *Number of Records* to color. You will need to use *Starttime* and then use the dropdown to select different parts of the time to get the different mapping.s




Use a filter to see if this is different from how people use Divvy when it's raining. Aside from just dragging *Rain* to filter, you can right click it in the *Filters* tray and choose *Show Filter* to get an interactive control on the right.