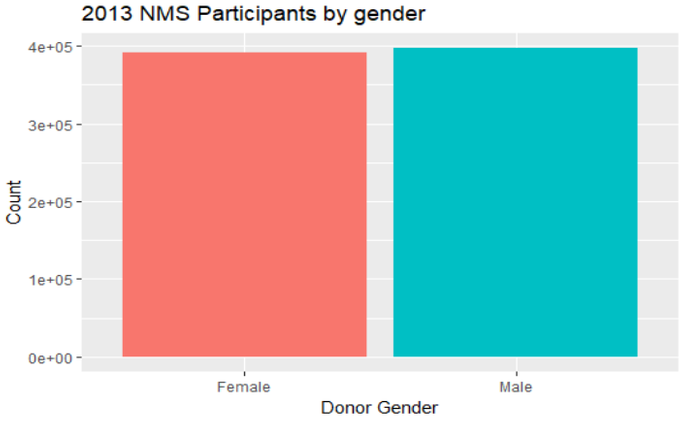
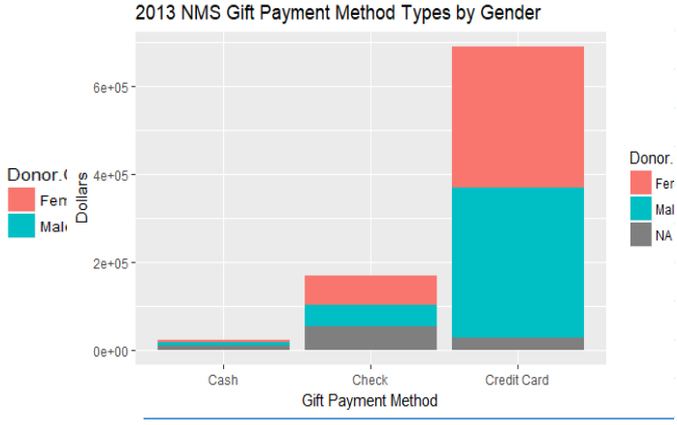
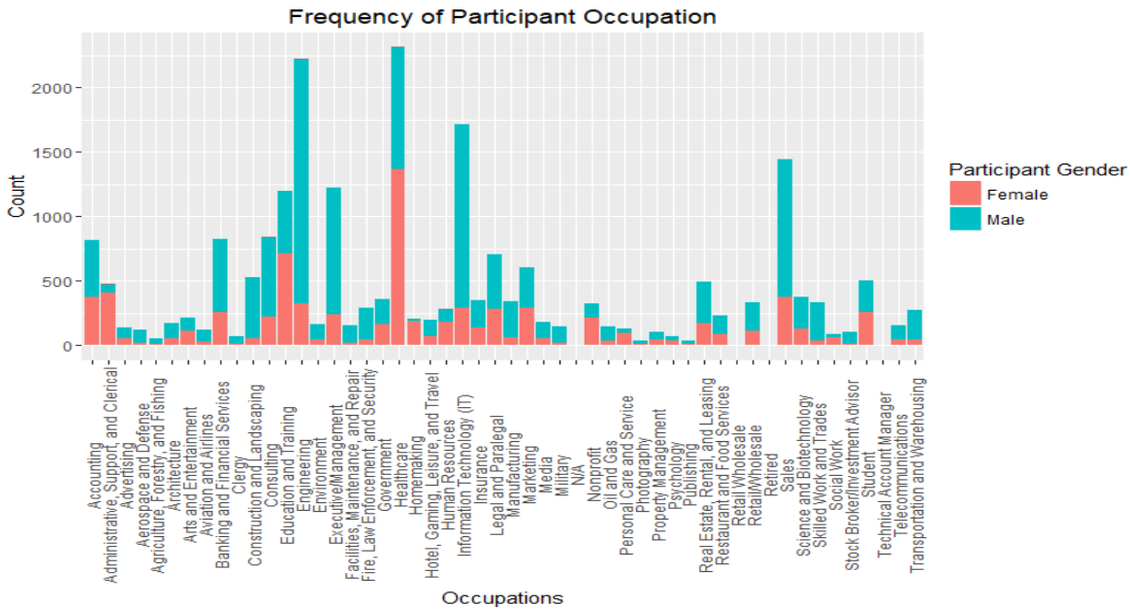
Week	Activity	Comments and Observations	Resources
Week 2			
	Choosing a Data Set	The plan is to use the BikeTown data that we will receive as part of our project in ETM 538 Data Warehousing, performing two different analyses using distinct methodologies for different parts of the data.	ETM 538 Dr. Feiling advised us that our team would be the one taking on the BikeTown Project.
Week 3			
	Email	We corresponded with Dr. Balkan to obtain permission to use the data set for both class projects. We received word this evening that we needed to find an original dataset.	
	Backup Plan	We had already discussed a back up Plan if we could not use the same dataset. So Konrad set out to see if there was a dataset we could obtain from his place of employment.	
	Project Proposal	We were running behind submitting a project proposal. We learned that obtaining a dataset can sometimes be quite difficult and/or time consuming.	
	Choosing a Dataset	We immediately began discussing viable options with rich enough data to really give us a chance to showcase the skills we would be learning.	
	TUN	We had all read the information about the Data Challenge that Dr. Balkan mentioned in class. It was discussed later that the best alternative would be the TUN challenge for several reasons. The specific business questions asked gave us a framework for research and analysis, we also had not just one but several datasets to work with available now. We were all excited by the challenge the Data Challenge vs. the Analytics Challenge posed by TUN. The potential to expand our skill set beyond the class.	www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge/Business-Questions/

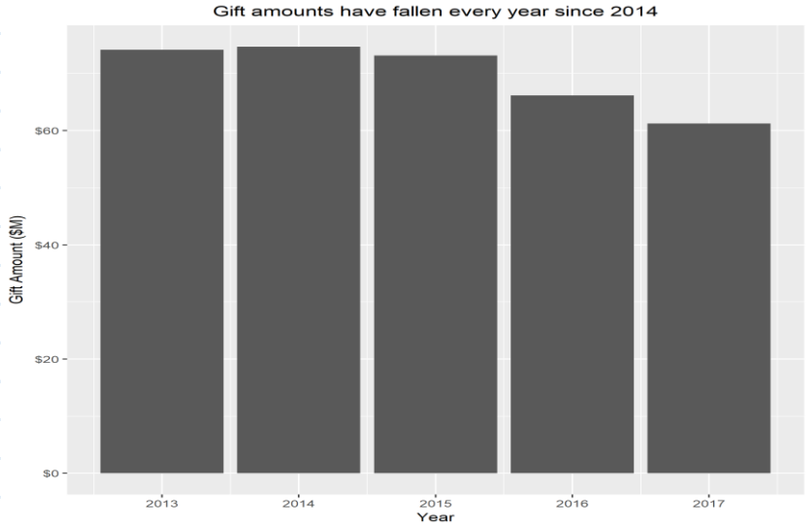
	Workspaces	Jordan created a workspace on Trello so that we could easily see what need to be done, what each team member was working on and what has been completed. We did this in hopes of being efficient. We also created us a chat channel that we would use to communicate about the project. He also opened us a GitHub page which is an online all-in-one service that allowed us to share cleaned data sets, Visuals and code.	www.trello.com www.slack.com www.github.com
	NMSS	Gave us data on their 'Bike MS' event we had a primary domain for the term!	https://secure.nationalmssociety.org/site/SPageNavigator/BIKE_HOM_splash.html
	In-Person Meeting	Having decided on a project-- we discussed-- then began the writing process on our project proposal.	EB Lobby
	Helping with Code	Jordan shared a link to Fundamentals of Data Visuals.	http://serialmentor.com/dataviz/visualizing-data-mapping-data-onto-aesthetics.html
	Pre-processing	We discussed and decided that Jordan would do an initial survey of the datasets to get an idea of how best to distribute the work equally. In the meantime the rest of us would begin to read again all project specifications, study the datasets, all info and reports. We would also make certain to have all the datasets downloaded before we meet again.	http://www.teradatauniversitynetwork.com/Community/Student-Competitions/2018/Data-Challenge/Datasets/
	Google Drive	Ramya created a shared file on Google Drive while we all began pre-processing the data	
	Prepare Visuals	Although the data needed a lot more cleaning two rough Visuals were created so that we'd have something produced.	





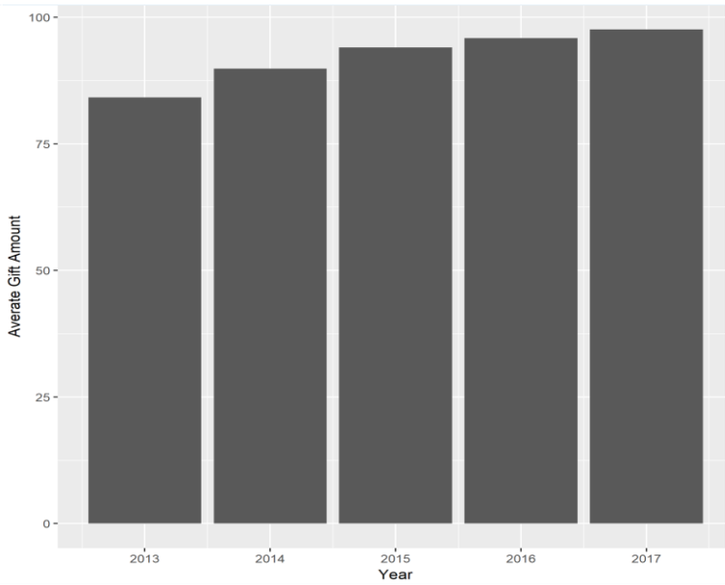
	Online Mini-Meeting	In addition to choosing our datasets we set goals to have our Visuals done by the end of the weekend	
	Exchanging Knowledge	Those with more experience in R guided those with less. An exchange of cheat sheets, advice on functions to call etc.	www.rstudio.com-data-wrangling-cheatsheet.pdf - rmarkdown-reference.pdf ggplot2-cheatsheet.pdf lineardata.net/the-ultimate-guide-to-cleaning-data-in-r
Week 4			
	Cleaning continues	Some of the cities had states attached to them and at different levels. For example- City-Athens is one entry " Athens AL" is another entry. It was decided that trimming after the space would be the best course.	
	Data Viz	Preliminary data viz continued as a some datasets were cleaned. We helped each other with some of the finer points of Visuals i.e. how to adjust axis labels and the like.	
			
	Coding Discussions for Cleaning	The DPLYR function was discussed: its use as well as syntax.	
	Sharing Info	Ramya posted an invitation to hear Emrah Cimren from Nike Analytics speak at PSU.	http://meetu.ps/e/DCxcK/rzR1H/a

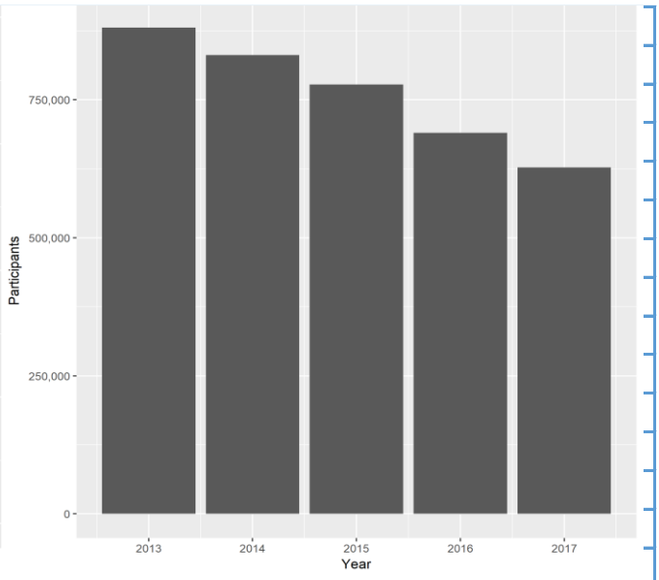
Project Proposal	Project proposal was worked on and shared through Google Drive. A general write-up describing each of the tables was created and shared on the drive. We reviewed homework problem #3 and used it as a guide for our work.
Cleaning continues	We continued to work on cleaning an pre-processing all the datasets
Coding Discussions for Cleaning	We were stumped as to how to format the date and time using R. We were manipulated dates in Excel using R code as a work around. {as.date} and {date_cols} were explored.
Cleaning continues	As the cleaning continued we were creating Visuals which made it much easier to glean insights.
Participant Data	After obtaining the average money raised from participants from each state the next thing to do was sort by amount. Which we would end up using the order() function.
R Markdown	The newest version of the data cleaning markdown went up on GitHub. As we finish cleaning files we'd either send an R/Rmarkdown file of our work, or add to the markdown file and submit a pull request. Hoping we can use the markdown files as part of our notebook for the class.
Data Visuals	Falling gift amounts, decreased participation and increasing gift per participant were graphed and added to GitHub



Average gift per participant has increased every year since 2013

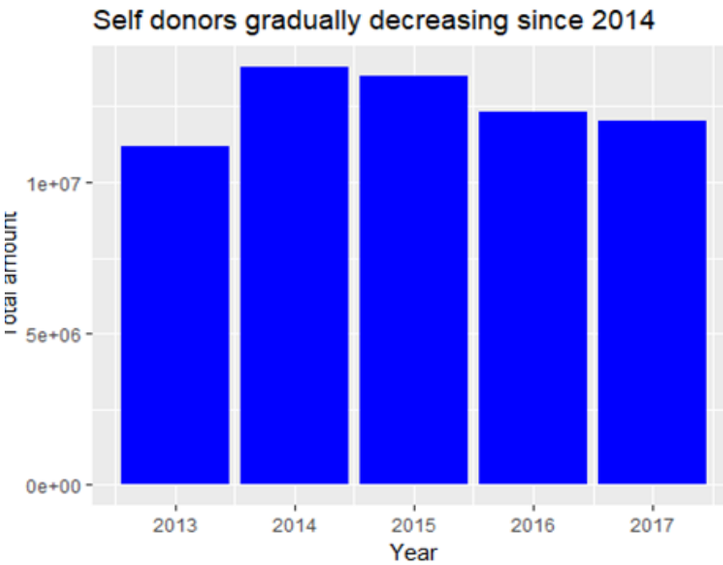
Number of participants has fallen every year since 2013

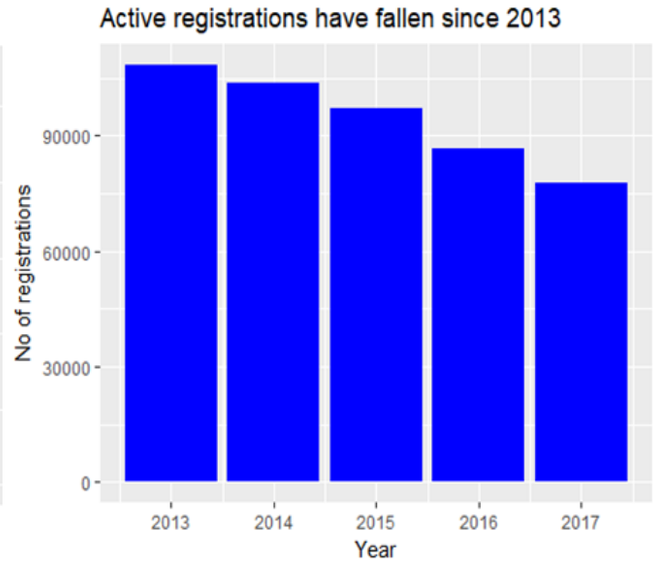




	Cleaning continues	We want to clean the entire data before we start thinking about what and how much to sample. Completed Bike Events cleaning, markdown file sent to email for inclusion. .....Right now as we're trying to get all of our data formatted, we want to keep everything we possibly can. Depending on which route we take down the road we'll deal with NAs differently. For the EDA I just wanted some quick numbers on median and sum of donations. In that case, having NAs will skew the calculations.
	Schedule Meeting	Everyone available to meet tomorrow on campus and plan out the next steps for the BikeMS project? Probably 1-2 hours maximum? By tomorrow we can start asking questions of the data so let's meet up and figure out which questions to start addressing first, and partition who's going to start answering which questions. There are a couple ideas on which outside datasets we could bring in and use. Meeting confirmed and study room reserved for 1-2 tomorrow at the library. If we end up needing more time we can probably move over to EB or FAB

Coding Discussions for Cleaning	When I convert my date column to date in R it is changing the dates to future dates. Hey guys my str function is not outputting variable names. Just V1, V2,…{I figured it out and it was so simple. My headers were in row 2 it read row one as header. A simple add to the import line skip = and it will skip however many rows but because header defaults to true it becomes the new header. I'm not even using na.rm -- I'm keeping every possible observation for now (and that includes observations with NAs) Use library(readr) and read_csv()	http://lineardata.net/the-ultimate-guide-to-cleaning-data-in-r/
Data Visuals	We looked at historical visualizations: the number of self donors over time, active registration numbers over time, and total amount raised by participants.	

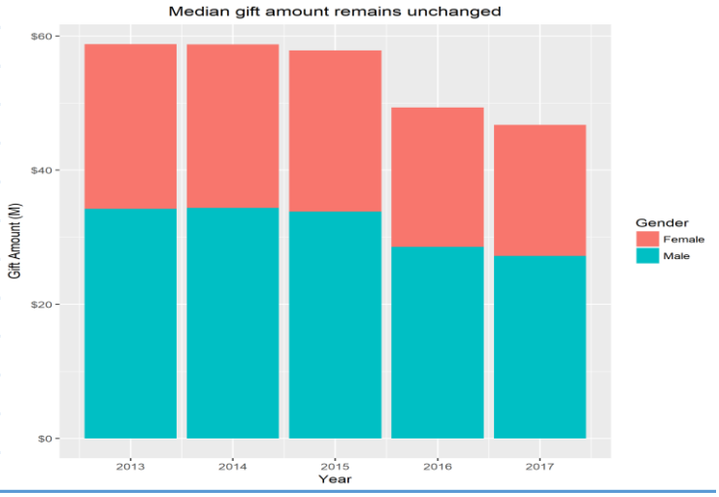


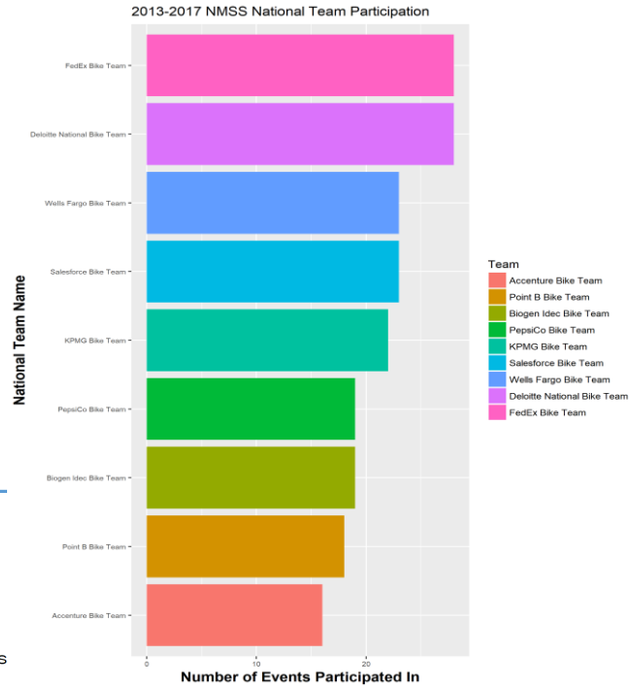


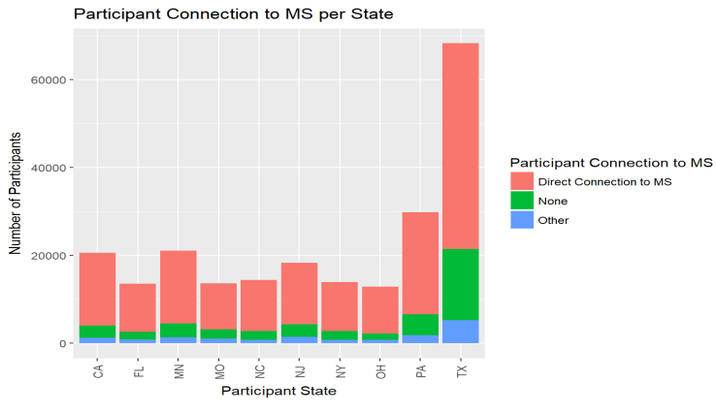
Things to Think About Before Our Meeting Tomorrow	What kinds of questions do you want to ask and answer about the data? What kinds of graphs about the data you want to make (these can answer the questions from above, or be in addition to)? What kinds of outside datasets you'd want to bring in to help answer the questions? There was a lot of good input and idea generating content to add to our Google Doc before we even had our meeting.

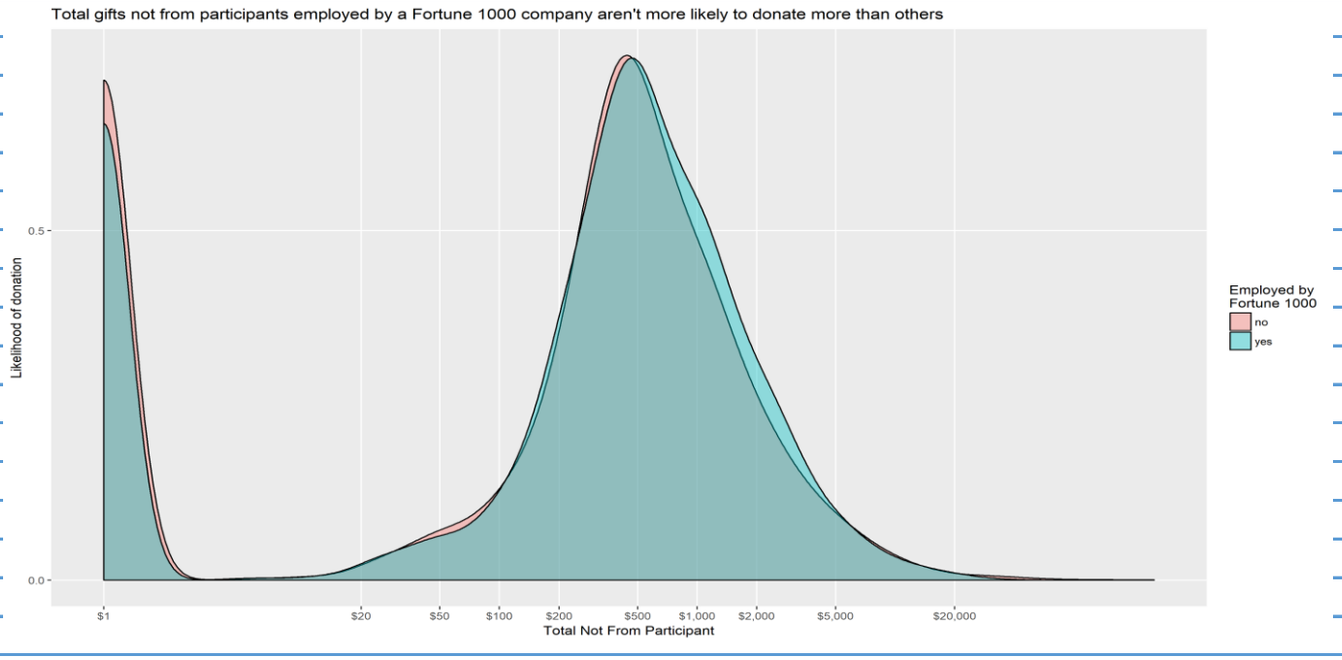
	Meet at PSU Library study room	Many questions were raised and discussed…. Bring in statewide MS incidence data and population data to see if states that have a higher incidence of MS also have more participants than average…. could this help the NMSS direct advertising or fundraising efforts in certain states? Konrad suggested comparing our data with a list of Fortune 1000 companies to see if there was a relationship...etc.	
	Data Visuals	We continued to muck our way through to complete all of the cleaning and pre-processing. We even began to refine our models and their accompanying images.	
	Sharing Info	As we continued to attempt to answer the posed business questions other questions came up. There were a lot unexplainable things we were seeing in the data. For example, from the data we have it seems a lot of people are donating but not participating? missing data producing strange gaps in the graphs.	https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
	Submission	We made sure to save our code as RDS files which were easier to deal with on GitHub and have our cleaned datasets uploaded. All images that had been decided upon were also collected on GitHub then merged it all together into the week 3 project submission.	
Week 5			
	Analysis	We spent more time cleaning the data, re-factoring variables, and dealing with misspellings in the data (e.g. some of the thousands of donors submitted their genders as “iale” or “feiale” which were re-factored as “male” or female”).	

	We focused primarily on our next steps regarding the business questions this week. The question became any ideas on where to go from here with the NMSS data? We worked more on the summary statistics of the donations data since it was the largest of all the datasets. The idea being to make the information more 'digestible'. We built out some different "personas" for our customer base which is usually a useful activity in marketing. Are we interested in the split in donations between credit card and cash/check donations?	
Next Steps	We made more graphs of participant occupations and employers and what the donations looked like. Konrad started working on the employers of participants., in trying to answer which companies have large teams and what the potential opportunities are for other companies not represented. For this, he drew in data from the Fortune 1000 to look at which participants are employed by the largest American companies. He also found a list of 20,000 unique up-and-coming employers as rated by the INC 5000 over the last 10 years. However, a big issue was that company names were misspelled or not quite aligned-- Salesforce and Salesforce.com are two different employer levels but belong grouped together. He spent some time trying to find the best ways of fuzzy matching strings that are similar but aren't exactly identical. Are we interested in if there's a difference in how much people donate by their 'connection to MS' IE relative/parent/friend has MS?	www.forbes.com/companies/national-multiple-sclerosis-society

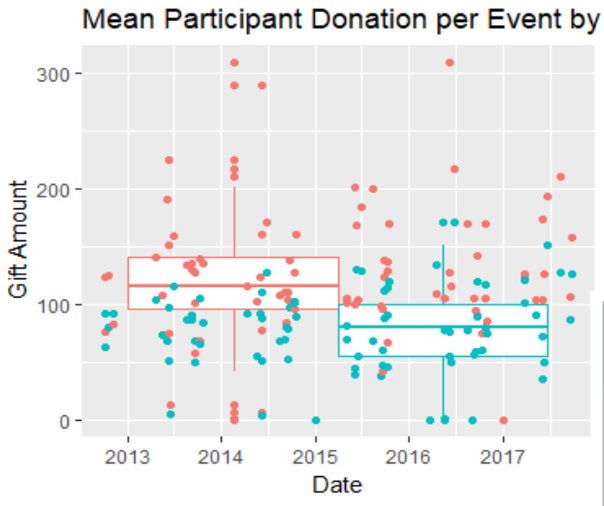

Median gift amount remains unchanged
$60
$40
$20
$0
Gift Amount (M)
2013
2014
2015
2016
2017
Year
Gender
Female
Male

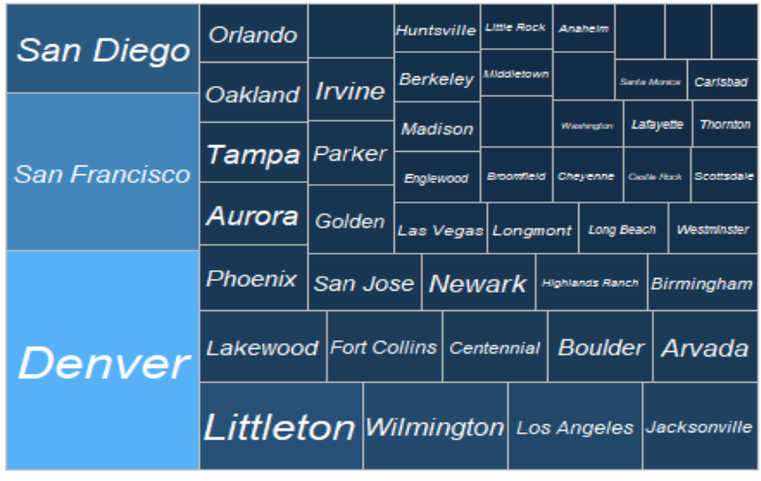

2013-2017 NMSS National Team Participation
FedEx Bike Team
Deloitte National Bike Team
Wells Fargo Bike Team
Salesforce Bike Team
KPMG Bike Team
PepsiCo Bike Team
Biogen Idec Bike Team
Point B Bike Team
Accenture Bike Team
National Team Name
0
10
20
Number of Events Participated In
Team
Accenture Bike Team
Point B Bike Team
Biogen Idec Bike Team
PepsiCo Bike Team
KPMG Bike Team
Salesforce Bike Team
Wells Fargo Bike Team
Deloitte National Bike Team
FedEx Bike Team

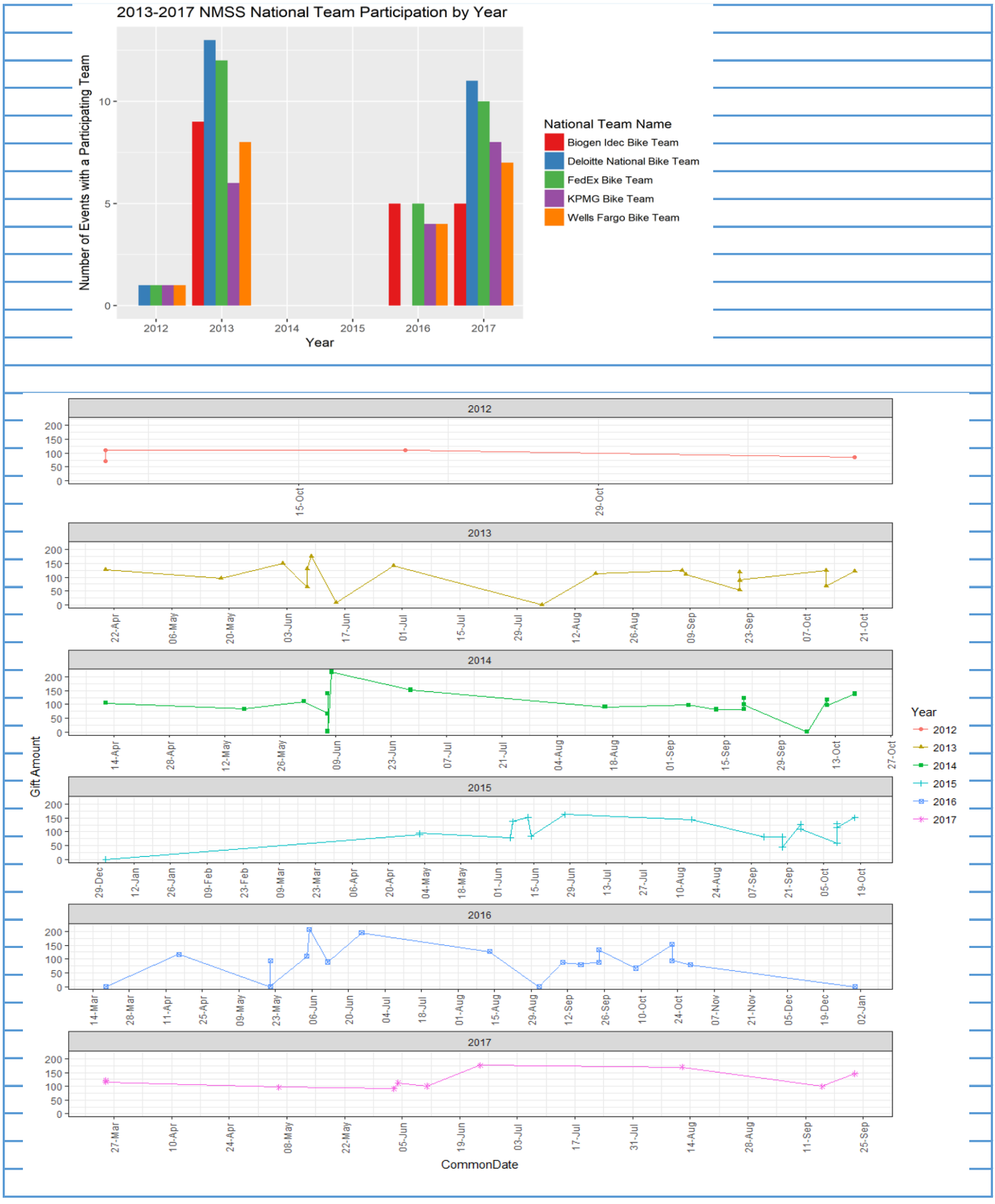

Participant Connection to MS per State
60000
40000
20000
0
Number of Participants
CA
FL
MN
MO
NC
NJ
NY
OH
PA
TX
Participant State
Participant Connection to MS
Direct Connection to MS
None
Other


Total gifts not from participants employed by a Fortune 1000 company aren't more likely to donate more than others
0.5
0.0
Likelihood of donation
$1
$20
$50
$100
$200
$500
$1,000
$2,000
$5,000
$20,000
Total Not From Participant
Employed by Fortune 1000
no
yes

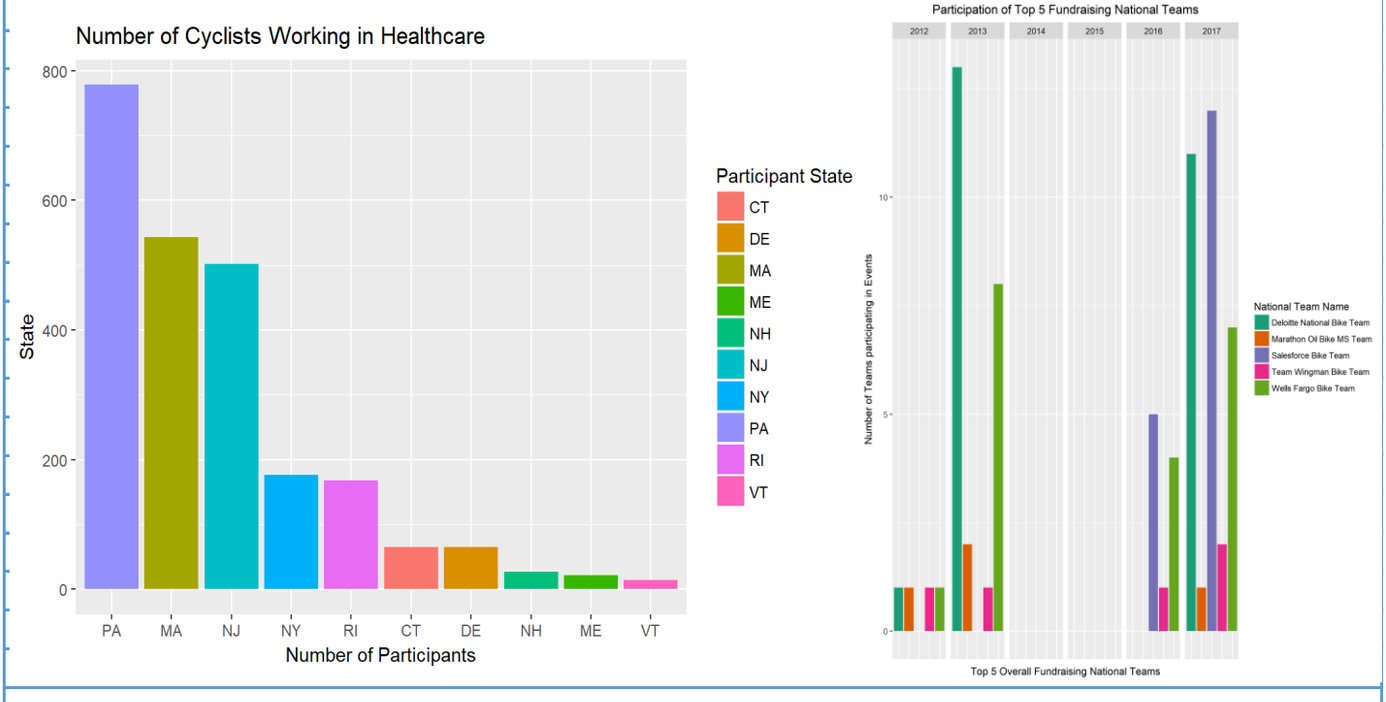
Week 6		
Analysis Continues	For example, Konrad continued with the fuzzy matching, settling on using the Jaro-Winkler similarity between strings as the best way to match employer names. He figured out how to aggregate the companies, and find potential matches in the Fortune 1000 and the INC 5000, though it took a lot of time and computational power (running the final iteration took about six hours of computation by his laptop). He then went through the lists and made sure the right companies were matched with each-other. This required some manual alteration as well; for example, grouping the USPS and the United States Postal Service. He then performed some analysis of the companies that had the most visitors.	
Data Viz link	ggplot2()	http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html





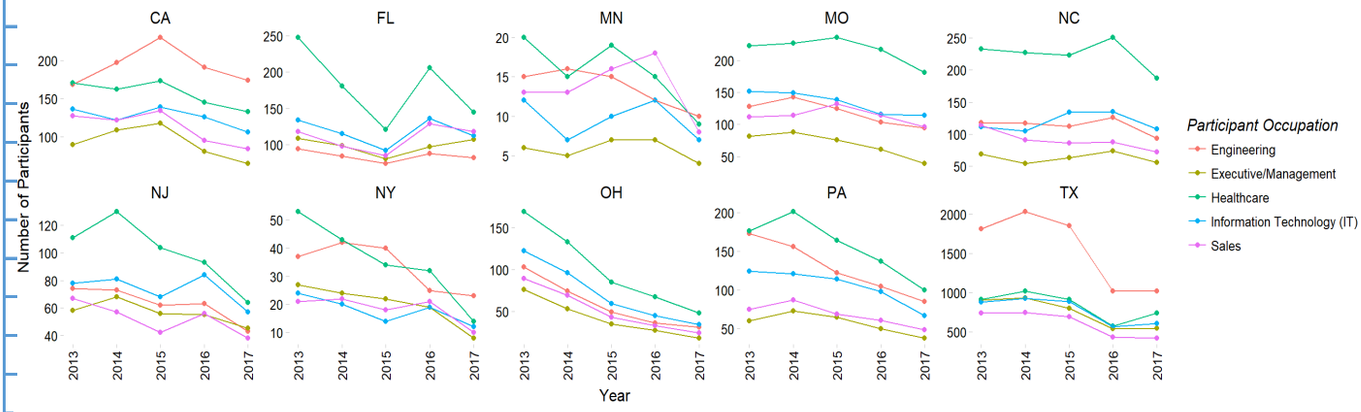

2013-2017 NMSS National Team Participation by Year
Number of Events with a Participating Team
Year
National Team Name
Biogen Idec Bike Team
Deloitte National Bike Team
FedEx Bike Team
KPMG Bike Team
Wells Fargo Bike Team
Gift Amount
CommonDate
Year
2012
2013
2014
2015
2016
2017

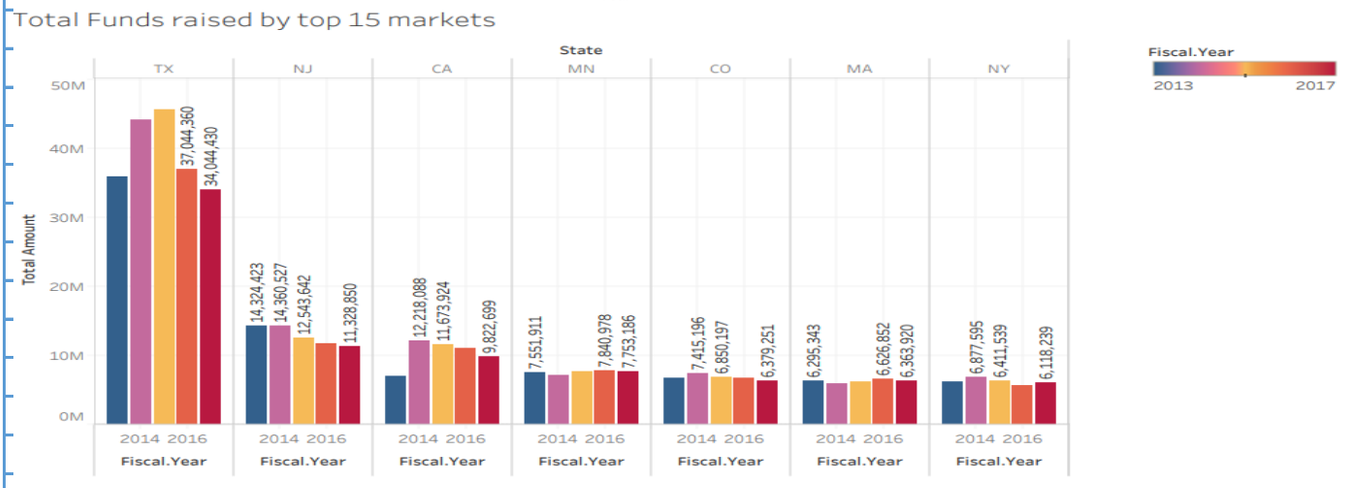
Week 7		
	The rough draft was due at the end of this week	Our time was spent preparing and updating this document. Konrad reorganized some of his previously created codes and Visuals so the workflow was more logical and interpretable. We discussed and decided that it would be best to present in person.
	Healthcare	At this point we were nearly done answering all of the posed business questions but there was still this question of did we really find all the opportunities for growth. Many thoughts and idea were passed ultimately it was Jordan's extensive work with the Participant data that leads us to an additional opportunity for growth. This work pulled together many of the conclusions we had drawn along the way.
	Creating Visuals	Additional charts and data Visuals were made, to summarize a lot of work and information into a concise format. We wanted to of course add the new work that Jordan had done. We put together a PowerPoint of our work thus far summarized and visualized for presentation.

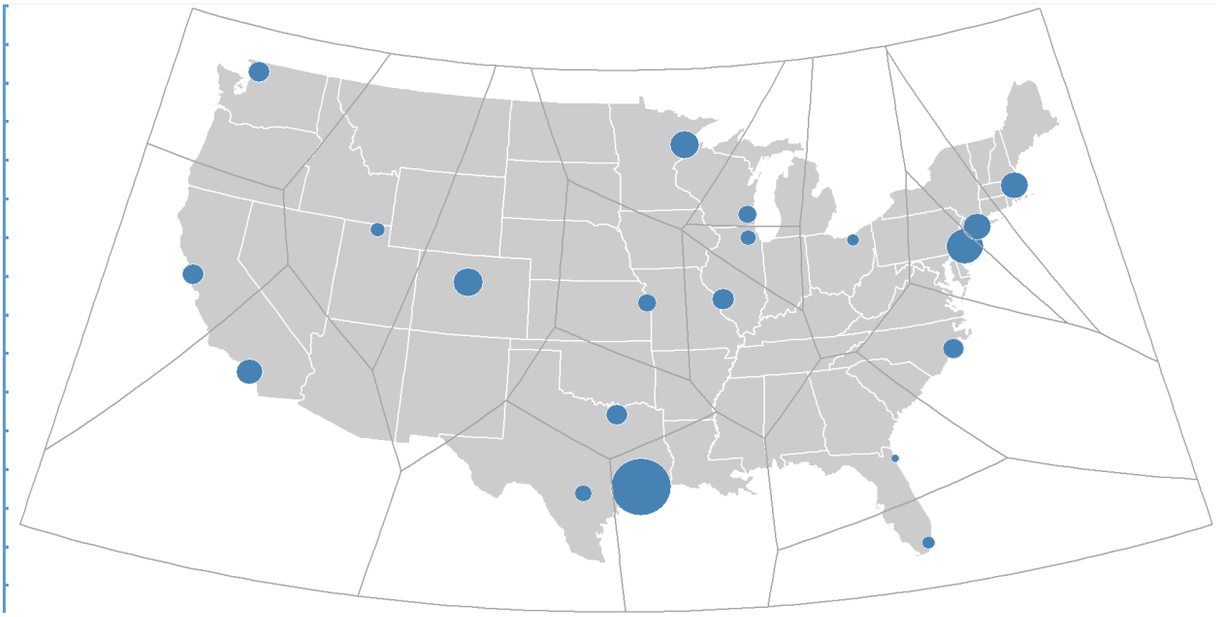


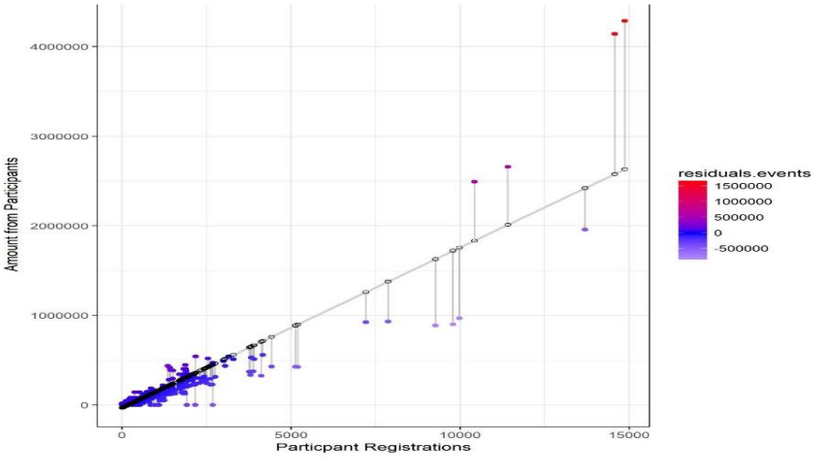
	In-person meeting	We met to discuss the presentation of our rough draft however, most communication was through online chat. Once the rough draft and PowerPoint were complete we continued to discuss possible relationships amongst the corporate teams and how we could answer the business questions.
	Email	Quite a bit of time was spent trying to confirm numbers since we found some of the datasets had conflicting information after cleaning and preprocessing. Donations data had more participants than the Events data and some events from Bike Events were not in Donations data, so Jordan sent an email to NMSS on behalf of the team to gain clarity in this discrepancy. We emailed TUN to ask for some clarity regarding the last of the 'First Priority' business question.
Week 8		
	Email Replies	Received responses to our emails providing the clarity needed to perform the best possible analysis leading ultimately to answers to their key business questions.
	Presentation De-briefing	This week we went straight to work on incorporating all the suggestions made by Dr. Balkan. As well as additional work inspired by the feedback.
	Research	Research was conducted as to the prevalence of MS in the US and found to the Midwest and east. Ramya concluded on extensive research on each of the top competing events. We continued to each add to the draft our findings and any accompanying Visuals.

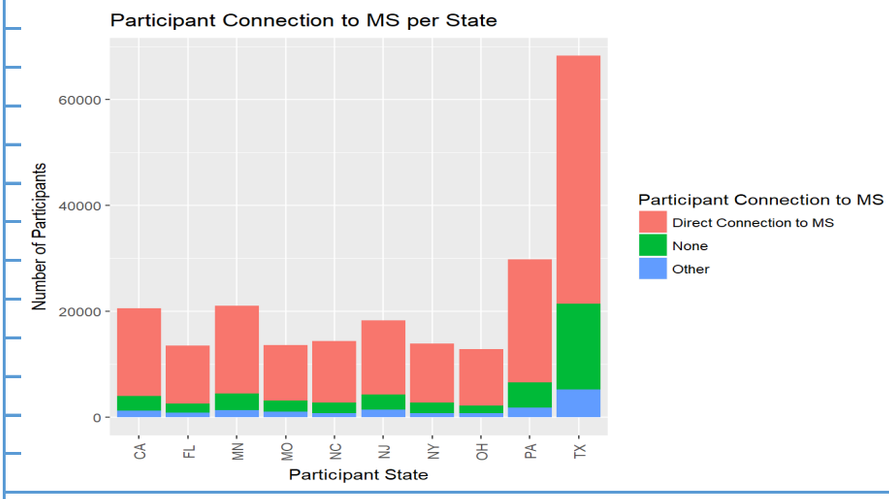
Week 9		
	This week we would be presenting our project at the end of the week.	New Visuals were created, the new analyses were summarized and organized. Most of our time was sent fine tuning the work that was done. We edited, polished and improved the questions previously answered. The PowerPoint was put together, the '2nd Priority' business answers were completed. Continued work on the final paper. Editing and polishing. Project presentation, received feedback.








4000000
3000000
2000000
1000000
0
Amount from Participants
0
5000
10000
15000
Particpant Registrations
residuals.events
1500000
1000000
500000
0
-500000


Participant Connection to MS per State
60000
40000
20000
0
Number of Participants
CA
FL
MN
MO
NC
NJ
NY
OH
PA
TX
Participant State
Participant Connection to MS
Direct Connection to MS
None
Other

Week 10		
	Challenge Submission Discussion	After our presentation we plan to attend a QA session with TUN regarding the Data Challenge and will meet with Dr. Balkan to assist us in creating good questions during the teleconference.