School of Information Studies
SYRACUSE UNIVERSITY

M.S Applied Data Science

Portfolio Milestone

Maya Mileva

SUID: 489678878

https://github.com/MayaMM99/MSADS_Portfolio

1. Introduction

The MS of Applied Data Science program at Syracuse University provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using various tools and techniques. Data scientists are a new breed of analytical data expert who have the technical skills to solve complex problems – and the curiosity to explore what problems need to be solved. They're part mathematician, part computer scientist and part trend-spotter. Ranked as Glassdoor's No. 1 best job of 2016, 2017, and 2018, data scientists are critical to the success of any organization. As the data science field evolves, the demand for analytics skills continues to grow. Employers are actively seeking candidates with the advanced technical expertise to make data-driven decisions.

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. Data science is related to data mining and big data and is a concept to unify statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data. It employs techniques and theories drawn from many fields such as mathematics, descriptive statistics, computer science and programming, statistical modeling, database technologies, data modeling, artificial intelligence and learning, text mining, natural language processing, visualization, and predictive analytics. Some of the applications that build upon the concepts of Data Science are Fraud and Risk Detection, Healthcare, Internet Search, Targeted Advertising, Website Recommendations, Advanced Image Recognition, Speech Recognition, Airline Route Planning, Gaming and Augmented Reality.

Through the Applied Data Science program, students learn to collect and organize data, identify patterns in data via visualization, statistical analysis, and data mining. The curriculum offers an innovative blend of information science and management principles. The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

I started my MS of Applied Data Science program at Syracuse University in January 2019 and anticipate graduation in June 2020.

2. Courses

2.1 IST 687 Introduction to Data Science

2.1.1 Project Description

Introduction to Data Science course provided introduction to data science, with applied examples of data collection, processing, transformation, management and analysis. Key concepts related to data science were explored, including applied statistics, information visualization, text mining and machine learning. R, the open source statistical analysis and visualization system, was used throughout the course. Supervised and unsupervised machine learning techniques were introduced. The focus was on structured data, using R (e.g., support vector machines, association rules mining) in conjunction with learning the full life cycle of data science.

The final team project was performed in RStudio and dealt with a dataset containing aggregated information on US K-12 Education since 1992. Given that we had multiple options on the ways we could look at US K-12 Education through different data sets, we decided to reshape and model our own data sets. We came up with general questions that had driven our investigation. Our team believed that an appropriate use of data and data analysis in education could change its course for years to come. With this in mind, the purpose of this project was to identify relationships among expenditure, enrollment, and student achievement (math and reading scores) in public schools in different states and years in order to advance some actionable insights in particular for the state of New York. In order to have a better point of observation for New York state, we examined two other states in particular, Massachusetts (the state with the highest total average score for math and reading in the US), and Florida (the state with the lowest total average score for math and reading in the US).

We first decided to analyze the full dataset after cleaning it, then we reorganized the full dataset into subsets. We decided to exclude Alaska, District of Columbia, and Hawaii from the list of states composing the US in order to work on the lower 48 states. We used the tapply() and aggregate() functions to compare means of the various columns by category. We started visualizing the data through maps with ggplot (ggplot2, ggmap and grid.arrange packages). We visualized different color-coded maps for the variables Primary/High school populations and Math/Reading scores for 8th grade.

Fig . 1 Primary school/High school populations

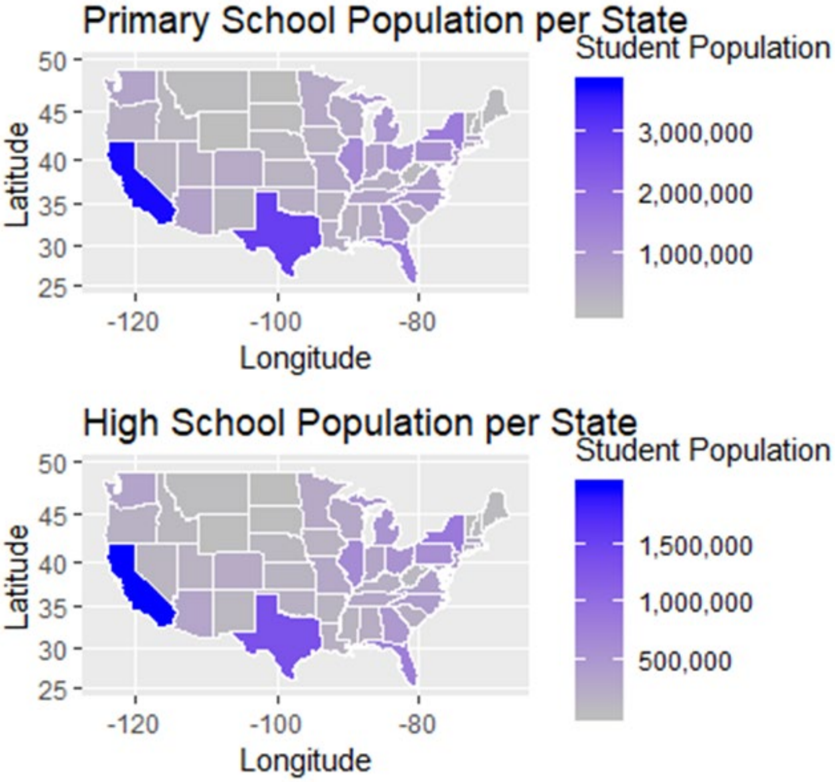
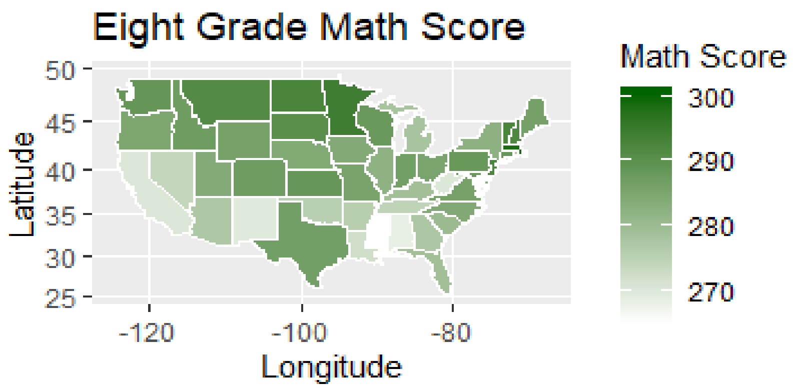
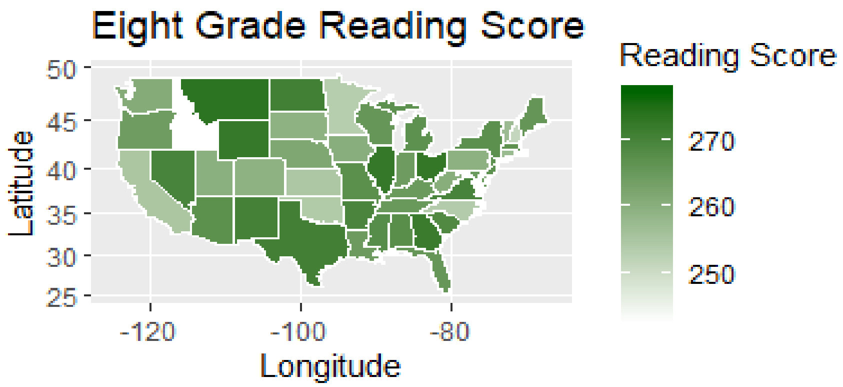


Fig. 2 Math/Reading scores for 8th grade





Creating these and other visualizations (see below) was beneficial to make things clearer and easier to understand, especially with such a large and multi-dimensional dataset as ours. We calculated average math and reading score for states too. We were curious to see which state and school has the most dropouts.

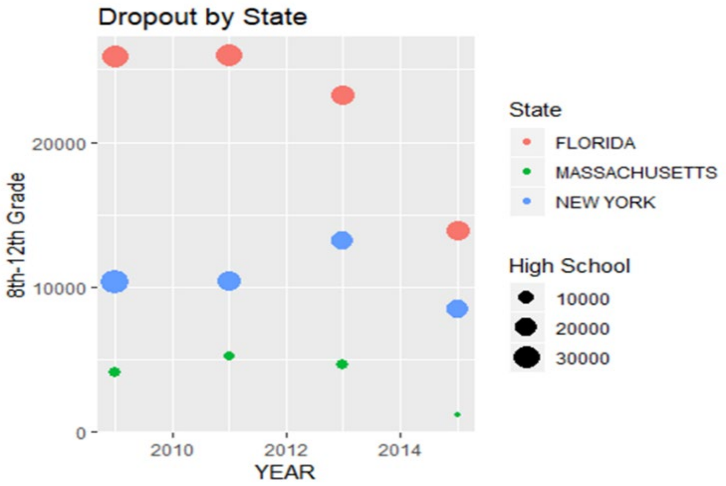


Fig. 3

In our analysis, we focused on finding possible factors that could affect student retention/attrition in public high schools. We observed different intensity in the examined correlations in different states using linear modeling. The findings lead to confirm a general principle that grounds our actionable suggestions at the end of the project (see the summary below): *there is not one combination of factors that works similarly for all states. Each state can have different correlations among the following three possible ones.*

COMBINATION	STRONG CORRELATION	WEAK CORRELATION	STATE
1	Retention ~ Instruction exp.	Retention ~ Scores	NY
2	Retention ~ Scores	Retention ~ Instruction exp.	MA
3	Retention ~ Instruction exp. Retention ~ Scores		FL

The Support Vector Machine was another method used to observe correlations among variables. Overall, the SVM, which we operated with the ksvm function in RStudio, confirmed the results returned through linear modeling.

Along with answering the business questions we provided actionable suggestions to the future administrations of New York state:

1. Always keep in mind the general principle according to which: there is not one combination of factors that works similarly for all states. Each state can have different correlations among different factors.
2. Comparing New York state with other two states, we observed that for improving student retention, student grades in quantitative skills as well as humanities-oriented areas of study are not as crucial as instruction expenditure.
3. The state of New York might want to consider redistributing expenditure in order to have higher instruction expenditure to have higher retention rate.

2.1.2 Reflection & Learning Goals

Considering IST 687 introduction course, it gave us solid ground for our further knowledge. In the course, we learned to setup R in R-Studio, R Markdown and install/load packages. We developed skills for manipulating and transforming data, such as accounting for missing values. We manipulated columns, vectors and data frames. We learned to inspect the data with commands such as str() and summary(), dim(). In learning to analyze distributions, we learned how to generate random distributions of different shapes using commands, such as rnorm(). We learned to display descriptive statistics, quantiles, and make histograms and boxplots, of the distributions. We learned about sampling in R, with commands such as rep() and sample() and making inferences about populations from the samples. Different build-in data set were introduced. We learned to make visualizations with the "ggplot2" library and looked deeper into ggmap() for creating geographical maps. We also used geocode() function to lookup coordinates and add points to a map. Correlation was analyzed via scatter plots and linear predictive modeling was introduced. Unsupervised machine learning algorithm, association rule mining (AR) as well as supervised machine learning algorithm named Support Vector Machine (SVM) and Naïve Bayes (NB) were part of that course too. Text mining and sentiment analysis were used to analyze Martin Luther King Jr speech, 'I have a dream'.

The final project provided an opportunity to implement everything learned in the course. It contributed to the successful application of the learning goals through the exercise of collecting and organizing data, as well as the identification of patterns using statistical analysis. These observations are leveraged to reveal insights. Actionable suggestions were made for implementing business decisions. Graphs were created for better communication.

2.2 IST 707 Data Analytics

2.2.1 Project Description

The Data Analytics course introduced data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. We dove into the various approaches and algorithms of machine learning and received hands-on experience with R and open-source software packages.

For the final project, my teammate and I analyzed a dataset of intakes and outcomes from Austin Animal Shelter to understand animal adoption trends, including which attributes of animals result in a higher likelihood of adoption. The Austin Animal Center is the largest no-kill animal shelter and shelters and protects over 18,000 animals each year. According to the data portal, over 90% of animal outcomes are adoptions, transfers to other shelter partners or returning lost pets to owners. We tried to identify a predictable pattern or a visible trend to shelter pet outcomes, find the features that are the best determinant for animal shelter outcomes and the features that are the best determinant for whether or not the animal is adopted. Our final goal was to predict whether or not an animal will be adopted based on characteristics Austin Animal shelter can identify upon intake.

After the data set was clean and preprocessed, redundant variables removed, we were left with 21 variables to be used for exploratory analysis and building the predictive models: breed, color, outcome type, age upon outcome etc. Due to the big variety we had to organize and split the color and breed variables. Through descriptive statistics and visualizations, we look at the overall numbers of the animals being admitted into the shelter and the distributions of their outcome based on animal type.

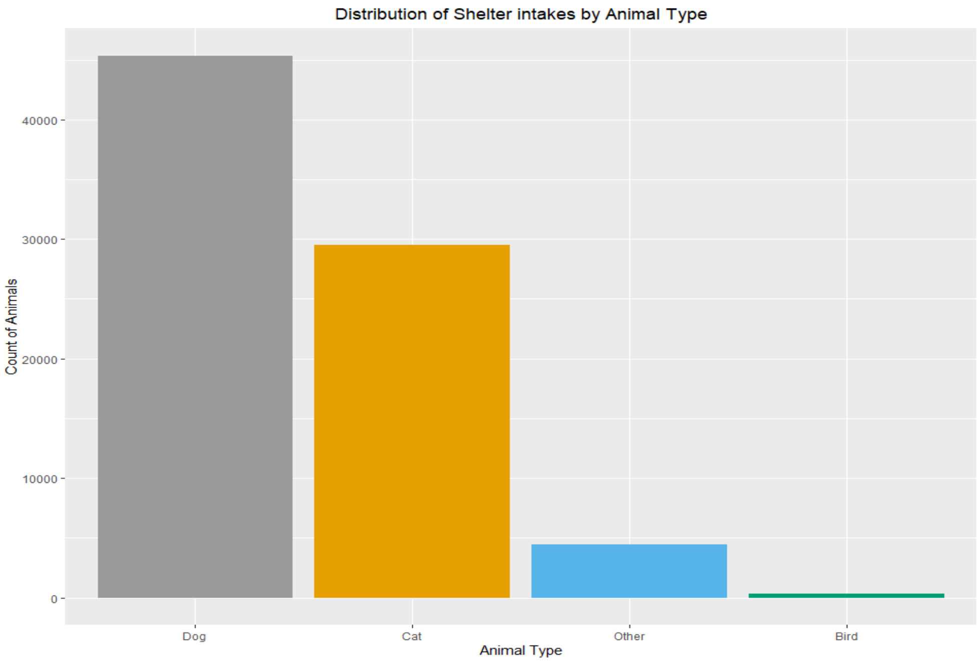


Fig.1

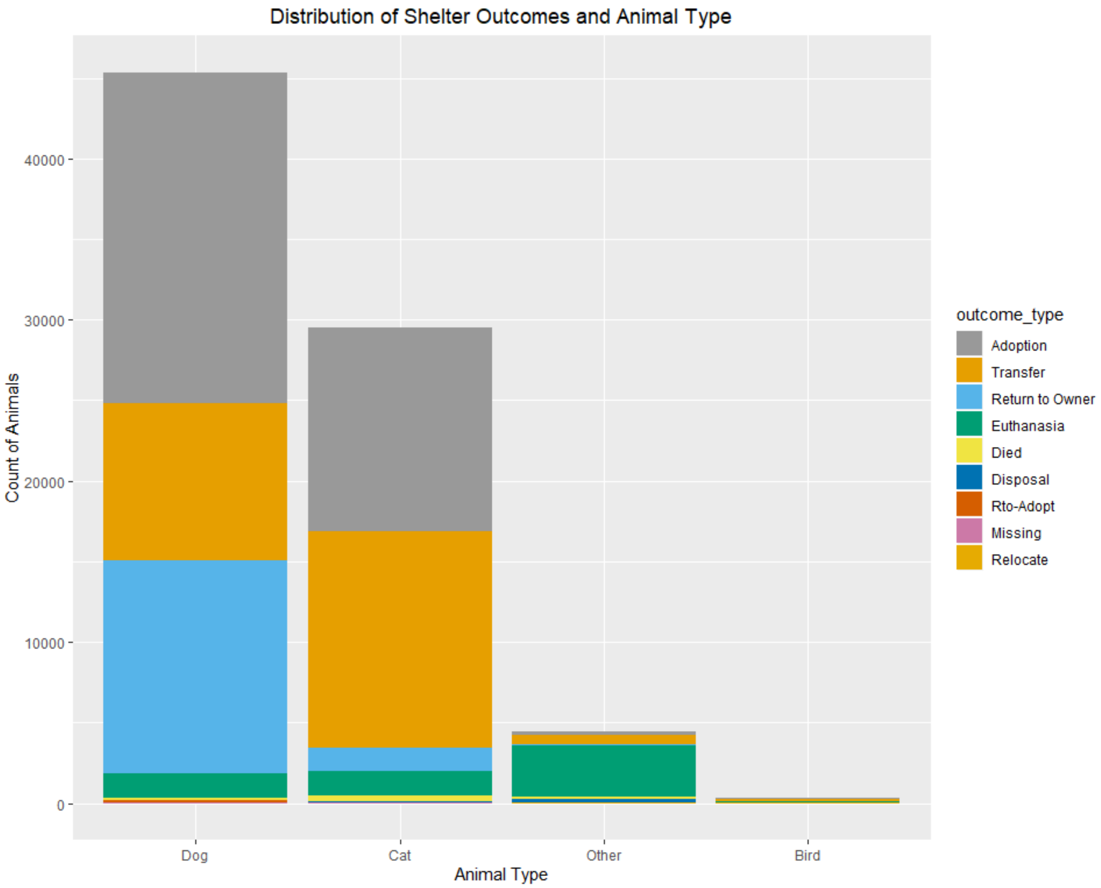


Fig. 2

We discovered that more dogs are returned to owners than cats, but more dogs overall are adopted than cats. And cats of all animal types get transferred more often. We further break down this distribution into the individual populations to see the percentages of outcomes by animal type. We also explored whether the age had an impact on outcome and if there’s a relation to the animal color. It was interesting to find out that there was drop off in adoptions and transfers for dogs after about 15 years of age. Though 15 years and older continue to be returned to owners. Additionally, there did not seem to be a specific pattern with regards to animal color. Majority of cats and dogs adopted had been spayed or neutered regardless of age or sex. We take a look at the intakes and outcomes over time in terms of day of week, hour of day and month of the year. The results showed that the number of intakes on each of day the week is almost consistent. Saturday had the highest number but having said that, Monday wasn’t far behind and so we couldn`t really conclude that the animals intook on Saturday were dropped in by their owners. The adoptions tent to peak around 5pm with 6pm following closely behind. May and June had the highest number of intakes.

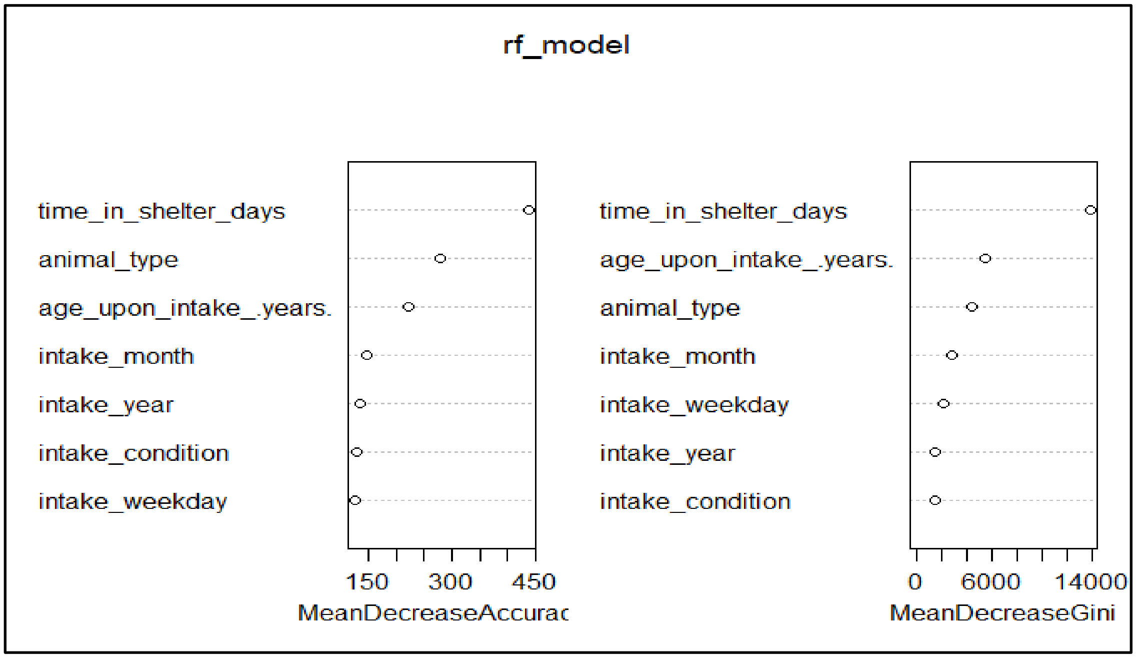
Probably people had gone for their summer vacations and so had to leave their pets at the shelter home and claim them once they were back.

Seeing as more dogs get adopted than cats, we expected the most popular breed to be a dog breed.

	primary_breed	freq
121	**Domestic Shorthair**	**23814**
251	**Pit Bull**	**6864**
190	**Labrador Retriever**	**6260**
87	**Chihuahua Shorthair**	**5726**
150	**German Shepherd**	**2612**
120	**Domestic Medium Hair**	**2464**

Interestingly the domestic shorthair shows up as the most popular breed, but it is actually really common cat breed not a dog breed. Dogs are more popular when it comes to adoptions: 74.63% adoption rate vs cats: 47.89%. Also, looked at which day of the week has the most adoptions and other outcome types. Adoptions peaked on Saturdays and Sundays but remained fairly consistent on the remaining weekdays ranging from ~3500-4,250 adoptions. The other outcome types took a dip on the weekends and also showed consistent levels across weekdays.

Four models were used to predict the outcomes of the animal with a goal of seeing which variable would lead to higher adoptions rates: Support Vector Machine, Decision Tree, Naive Bayes and Random Forest. Before generating the models, we use information Gain function of the RWeka package to rank the features. Time in the shelter, sex upon outcome, primary breed, age upon outcome and intake type were the ones, with largest weights. Some discretization of the time-based variables was performed to limit the buckets time to 10 based on frequency of occurrence. SVMs assume that the data it works with is in a standard range, usually either 0 to 1, or -1 to 1 (roughly). So, the normalization of feature vectors prior to feeding them to the SVM is very important. Primarily splitting into training and test sets based on a 75/25 split and cleaning those train and test sets to remove non-zero variance were performed. Lastly, we created dummy variables and re-normalized the data for the remaining categorical values. The tune.svm function was used to iterate between cost and gamma values of 0.1, 1.0, and 10. The tune.svm function also applies a 10-fold cross validation procedure to the model. The optimal parameters based on the variables chosen were Cost = 10 Gamma = 1. The confusion matrix showed an accuracy of 65.58% and varying levels of specificity and sensitivity. A decision tree was built using the normalized train and test dataset along with removal of near zero variance values. The full tree was first built, and post pruned at cp = 0, 0.1, 0.001, 0.0001 and 0.00001. Ultimately the cp of .0001 showed the best improvement in model accuracy. The best accuracy was 70.94% which was an improvement over the best SVM model. The Sensitivity and Specificity values were similar to those from the SVM model. A Naive Bayes algorithm-based model was attempted but showed the worst performance of all models generated. The random forest model was run using the default ntree setting of 500 and showed the best performance of all models. A second run was attempted at ntree = 800 but did not show any improvement in accuracy. As the ntree = 800 model took longer to run than the ntree = 500 model and showed no improvement, the ntree parameter was reset to 500.



The variable importance plot above was consistent with the feature ranking table with regards to the most important variable time in shelter days.

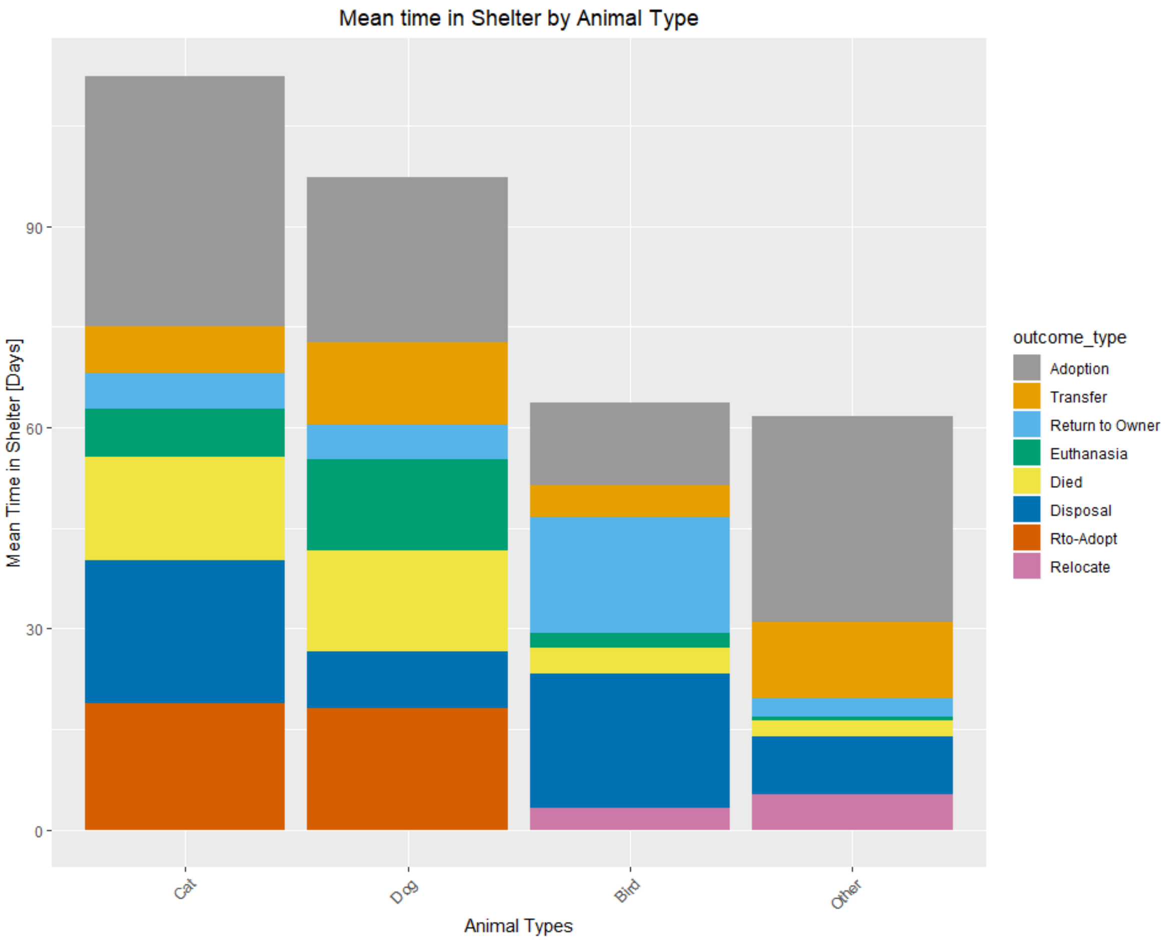


Fig. 3

Based on the models evaluated, the random forest algorithm presented the best accuracy at 73% which was significantly better than random chance and selecting the most probable outcome. Based on this model and its rating of variable importance we saw that time spent in the shelter has a strong link to animal outcomes. However, we could not make an argument that animals spending more or less time in the shelter will lead to higher adoptions, we must rely on other variables such as animal type and age. Here we saw a better link between the characteristics of the animal and the outcome. Focusing on these variables was supported by the exploratory analysis that shows dogs are generally adopted in greater numbers than cats and that both cats and dogs see a significant drop off in adoptions after 15 years.

Areas for future improvement of these models included grouping common breeds and gathering additional data across multiple animal shelters as there may be a geographical component over which pets are popular in certain regions. This could lead to additional recommendations regarding where to transfer cats in order to improve their adoption rates.

2.2.2 Reflection & Learning Goals

The Data Analytics course thought us a lot and was one on my favorite. After the ground base, established in IST 687, we continued to develop our skills. We learned about classification vs clustering. We discussed data types and data quality issues, such as outliers, missing values, duplicate data. We talked about business application for association rule (AR) mining, and metrics to evaluate its strength. We applied the apriori method in R and learned how to identify rules with high lift and confidence which at the same time have relatively good support, and made recommendations, based on the discovered rules. We learned about partitional and hierarchical clustering analysis, and the various methods used today such as k-means. We learned to evaluate the results with some criterion, e.g., minimizing the sum of square errors. We learned about Euclidean and Manhattan distance measures and the cosine similarity. We learned how to prepare the data to be clustered. Decision Tree classification was introduced. We learned to measure how well a given attribute separates the training examples with information gain and gain ratio, and to measure impurity with entropy, also the problem of overfitting (can be caused by noise and insufficient samples in the data) and pruning reducing the decision tree complexity) was discussed. We learned to evaluate models based on hold-out and cross-validation tests. We learned to detect errors with a confusion matrix and measure precision, recall, and F-measure. We learned that having not enough data can affect prediction accuracy, so semi-supervised learning, active learning, or crowdsourcing can be implemented to account for this. We used the Naïve Bayes (NB) classifier and about the Bayes' theorem of calculating probabilities based on posterior observations. We learned about the K-Nearest Neighbor (K-NN) algorithm, and it, unlike decision trees, can be used when the decision function to be learned is very complex. Since it does not use a linear boundary, it differs from NB in that its boundary has no defined shape. The disadvantages are that is sensitive to noisy data, and computationally expensive. We learned the SVM is an algorithm that can solve both linearly separable and inseparable problems with kernel functions. We learned about random forests and why ensemble works.

The final project provided an opportunity to show our understanding of everything learned during the term in that class, contributed to the successful application of the learning goals through the development of alternative strategies based on the data, and the communication of observations which translate to actionable insights. Data was organized and examined. Data mining was also used in conjunction with visualization to identify patterns in the data for use in classification tasks. Successful presentation of the project report demonstrated communication skills regarding the data and its analysis.

2.3 MBC 638 - Data Analysis and Decision Making

2.3.1 Project Description

In the Data Analysis and Decision Making course, we used the six sigma framework, DMAIC, or Define, Measure, Analyze, Improve, and Control, a data-driven improvement cycle for improving, optimizing and stabilizing business processes and designs in order to make better business decisions.

Throughout the course, I worked on a process improvement project. The project involved following the DMAIC framework to improve a personal process. For the “define” stage I declared by business problem: I needed more quality time to study for my class MBC638. . I hadn’t been in school for 5 years, so my life has changed a lot – work, family, personal responsibilities came along. I needed to change my everyday schedule in order to commute my new school classes and assignments. Not having enough time to study could have led to bad grades and not completing the class. My goal was spending 20% more quality time a week studding for my class MBC 638. I wanted to improve the way a study, make most of my time, by eliminating all the distractions. I had to create project plan and process map that had to be improved (Fig.1). The Measurement process included taking notes every day, started from the moment I woke up every morning and measured my whole day – where, when and what time I spent for different activities and continued for 28 days. I created a Thought Process Map representing my thought process (questions, related actions and related decisions). Asking those questions gave me clarity on my process and where to look for the actual problem. With 95% confidence and margin of error 40, I needed at least 24 samples for my project. In the “analyze” phase, hypothesis testing and regression analysis was used to discover the root causes, identify/prioritize critical inputs, and determine the inputs impact on the output. In the “improve” process, a new process map was drawn based on the analysis. In the last phase, “control,” a time-series control chart was created. Additionally, in the control phase, a moving average model was produced to forecast results from the improved process (Fig.2). I was happy to see the results from the final changes made during the control phase. The weekly average study time increase from 91 minutes to 190 minutes daily.

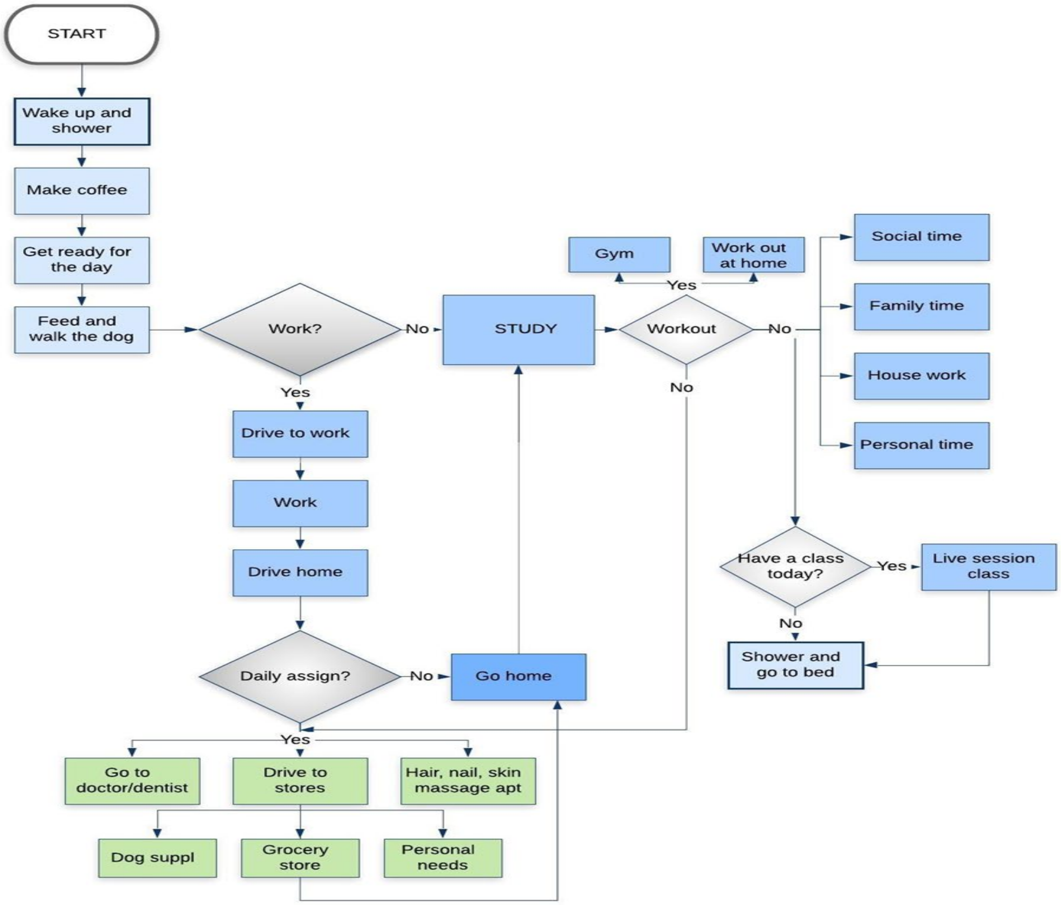


Fig. 1

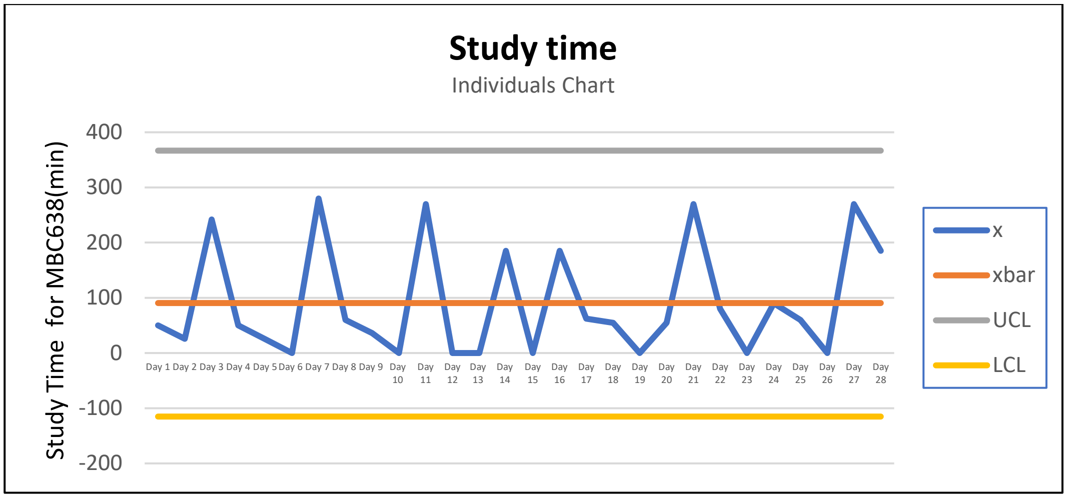
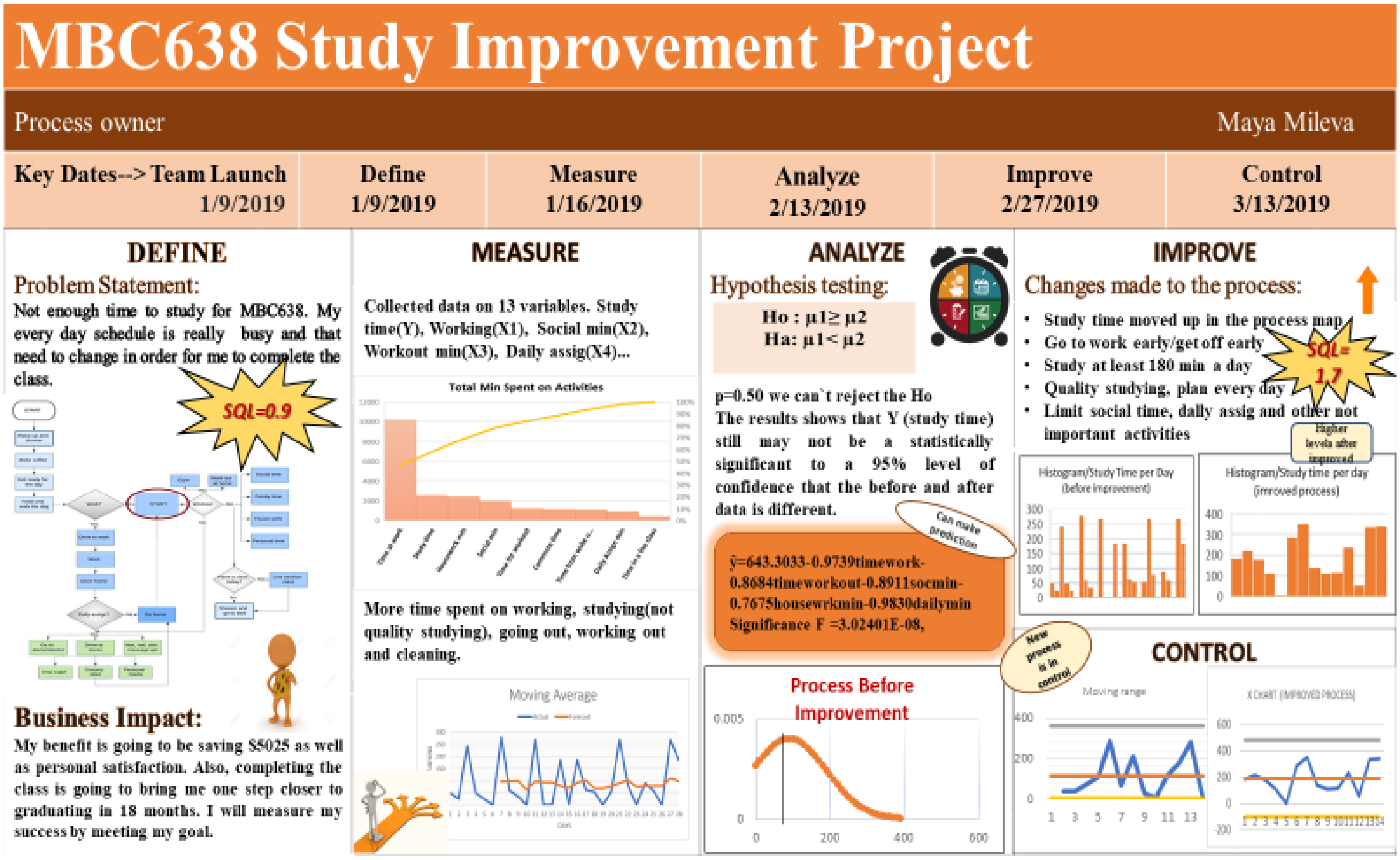


Fig. 2



MBC638 Study Improvement Project

Process owner Maya Mileva

Key Dates--> Team Launch	Define	Measure	Analyze	Improve	Control
1/9/2019	1/9/2019	1/16/2019	2/13/2019	2/27/2019	3/13/2019

DEFINE

Problem Statement:

Not enough time to study for MBC638. My every day schedule is really busy and that need to change in order for me to complete the class.

SQL=0.9

Business Impact:

My benefit is going to be saving $5025 as well as personal satisfaction. Also, completing the class is going to bring me one step closer to graduating in 18 months. I will measure my success by meeting my goal.

MEASURE

Collected data on 13 variables. Study time(Y), Working(X1), Social min(X2), Workout min(X3), Daily assig(X4)...

More time spent on working, studying(not quality studying), going out, working out and cleaning.

ANALYZE

Hypothesis testing:

Ho : µ1≥ µ2
Ha: µ1< µ2

p=0.50 we can`t reject the Ho
The results shows that Y (study time) still may not be a statistically significant to a 95% level of confidence that the before and after data is different.

Can make prediction

ŷ=643.3033-0.9739timework-0.8684timeworkout-0.8911socmin-0.7675housewrkmin-0.9830dailymin
Significance F =3.02401E-08,

IMPROVE

Changes made to the process:

- Study time moved up in the process map
- Go to work early/get off early
- Study at least 180 min a day
- Quality studying, plan every day
- Limit social time, daily assig and other not important activities

SQL= 1.7

Higher levels after improved

CONTROL

New process is in control

Fig. 3 Final Poster

2.3.2 Reflection & Learning Goals

I this course we learned about the fundamentals of statistics and the DMAIC framework. In the “define” phase, we learned to clearly define the business problem, output measure, customer, scope, goals, and resources. We discussed the output formula, descriptive statistics, as well as the different types of data (continuous vs discrete). We learned to measure processes with the Sigma Quality Level (SQL). We about various tools, such as thought process maps, SIPOC, cause and effect diagrams, Paretto charts, histogram/frequency charts, and trend charts and time plots to help define the process. In the “measure” phase, we learned how to the validate the measurement system and collect baseline data. We discussed mapping a process/value-stream, forms of waste, measurement error, reproducibility, and repeatability. We learned about the importance of operational definitions, and the use of Kappa in measuring agreement in a measurement system. We also talked about using data stratification trees. In the "analyze" phase, we learned to describe and present the data to discover the root cause(s), identify/prioritize critical inputs, and determine the inputs impact on the output. We discussed common distributions and used hypothesis and Chi-square testing on samples to perform inferential statistics and describe the likelihood of an event occurring. We discussed confidence intervals and how sample size is related. We used regression analysis to identify relationships between the dependent and predictor variables. We performed simple and multiple linear regression, and calculated correlation among variables. In the "improve" stage, we learned to implement a solution, run a pilot, evaluate the results, and complete a hypothesis test. We used solution selection matrices to be assessing the positive impact of each proposed solution. Lastly, in the "control" stage, we learned to implement process changes and controls, verify expected performance was achieved, and monitor performance to sustain new levels. We used Xbar/R and ImR control charts, and time series analysis, to include first order autoregressive (AR1), moving average forecast, and exponential smoothing models.

The final assignment I developed a plan of action to implement the business decisions derived from the analyses in the personal improvement project. Collecting and organizing data, created me by me was really interesting. Analyzing that data, the level of measurement error had to be considered. I was able to identify the main problem, dig deeper to the cause of it and eliminate it. Pareto chart results helped identify the main spending time activities and later I focus on them and I found I way to remove or minimize them. Changing my priorities was the key point. Chi square showed that I actually spend more time studying during the weekend. New strategy had to be created and implemented based on the data analysis. The result was new improved process. Recommendations for further action were made by adding more variables into the analysis. The final poster had the idea of summarizing the process. The main steps were successfully visualized sowing the main breaking points.

The kills gained from this course were used through the whole Applied Data Science Program and in my personal and professional life. The skill of identifying a problem and getting to the reason behind it is essential. The ability to create a plan in time frame and implement it successfully is important.

2.4 IST 719 Information Visualization

2.4.1 Project Description

The Information Visualization course provided a broad introduction to data visualization for information professionals. We developed a portfolio of resources, demonstrations, recipes, and examples of various data visualization techniques through the use of R programming language and Adobe Illustrator. The skills we learned include using data cleaning techniques, developing custom plots, visually exploring data, using design concepts to visually communicate the story in the data, and discussing issues related to the ethics of data visualization. Conceptual themes were presented alongside technical aspects of data visualization.

For the final project, I created a visualization poster and presented the work during a live session. The dataset I chose was Employee Attrition dataset from Kaggle. The dataset includes features like Age, Education, Job Satisfaction, Years at the company, Years in the current role etc. Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. As such, there is great business interest in understanding the drivers of and minimizing staff attrition.

I explored the factors that lead to employee attrition and created visualizations using R and Adobe Illustrator. The goal of the project was to answer my business questions with visualizations and create a story, that can be easily accumulated by anyone. I wanted to discover the average job satisfaction by attrition status, what was the attrition by education level, was the income reason for the employee to leave and was overtime important. One of the factors that I was curious about was the behavioral difference between generations (Fig. 1). Fig. 2 displays the age distribution in the observed company.

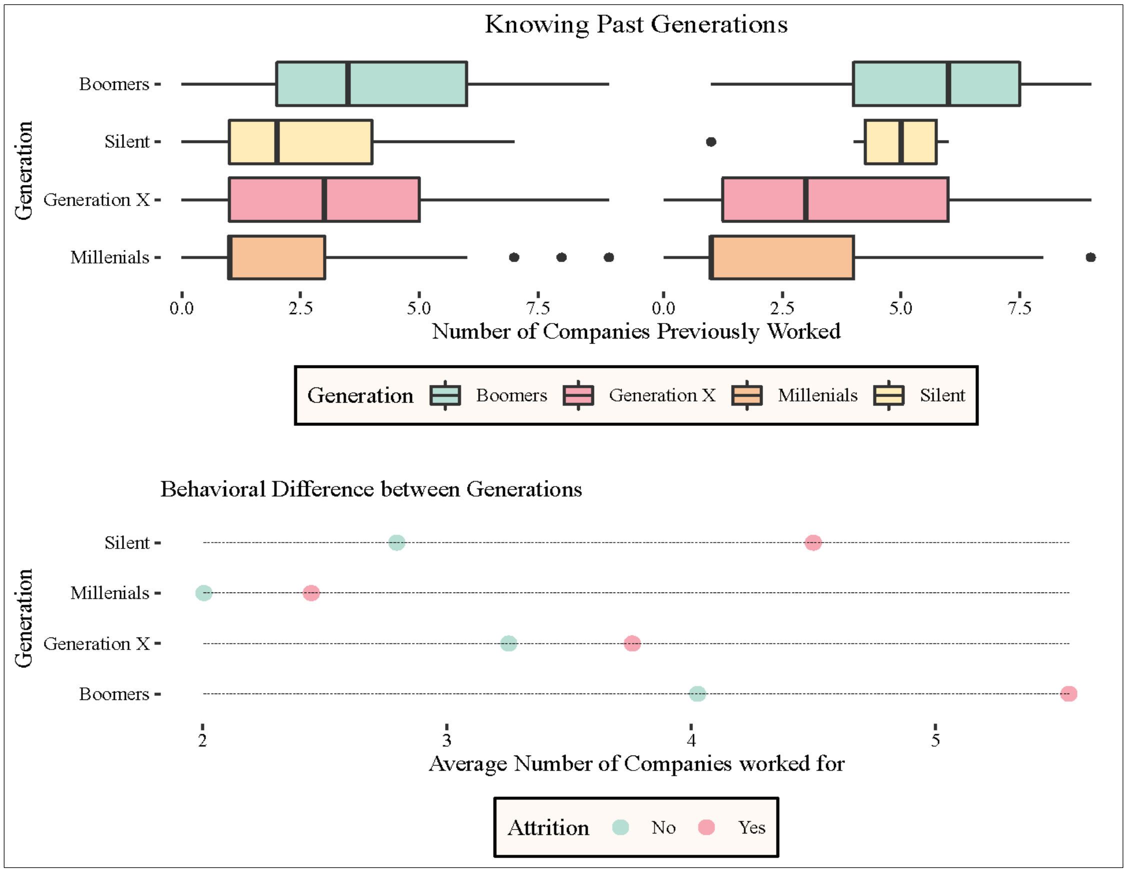

Knowing Past Generations
Generation
Boomers
Silent
Generation X
Millenials
0.0
2.5
5.0
7.5
Number of Companies Previously Worked
Generation
Boomers
Generation X
Millenials
Silent
Behavioral Difference between Generations
2
3
4
5
Average Number of Companies worked for
Attrition
No
Yes

Fig. 1

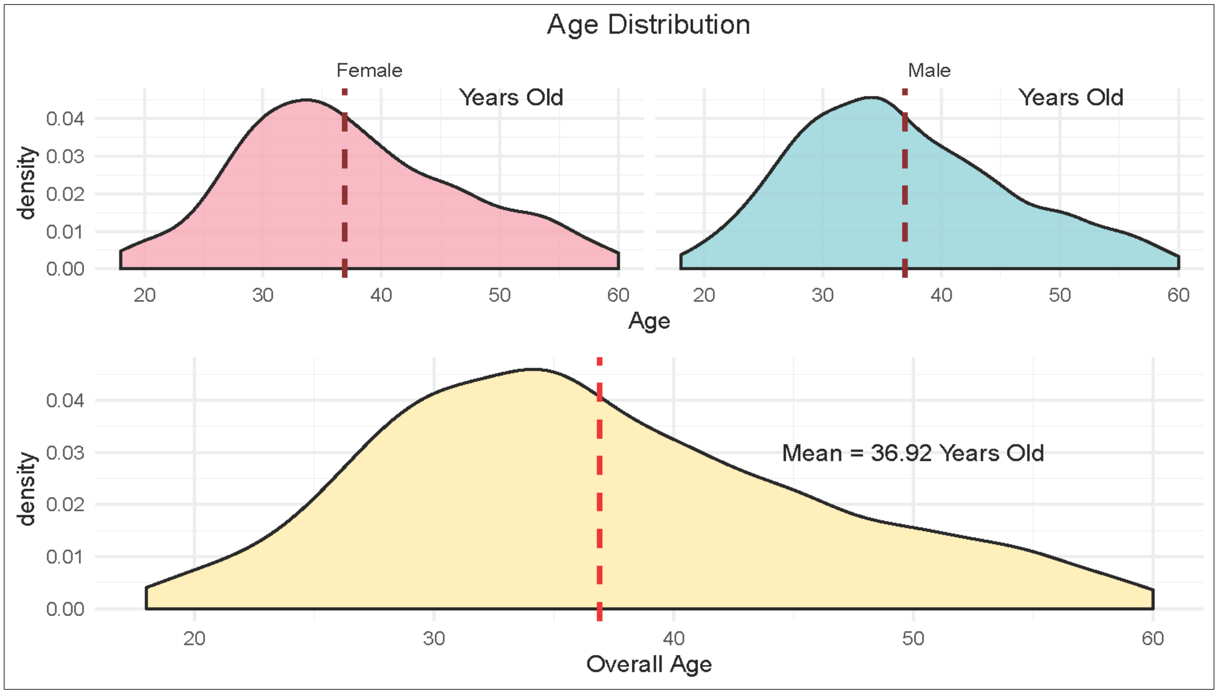
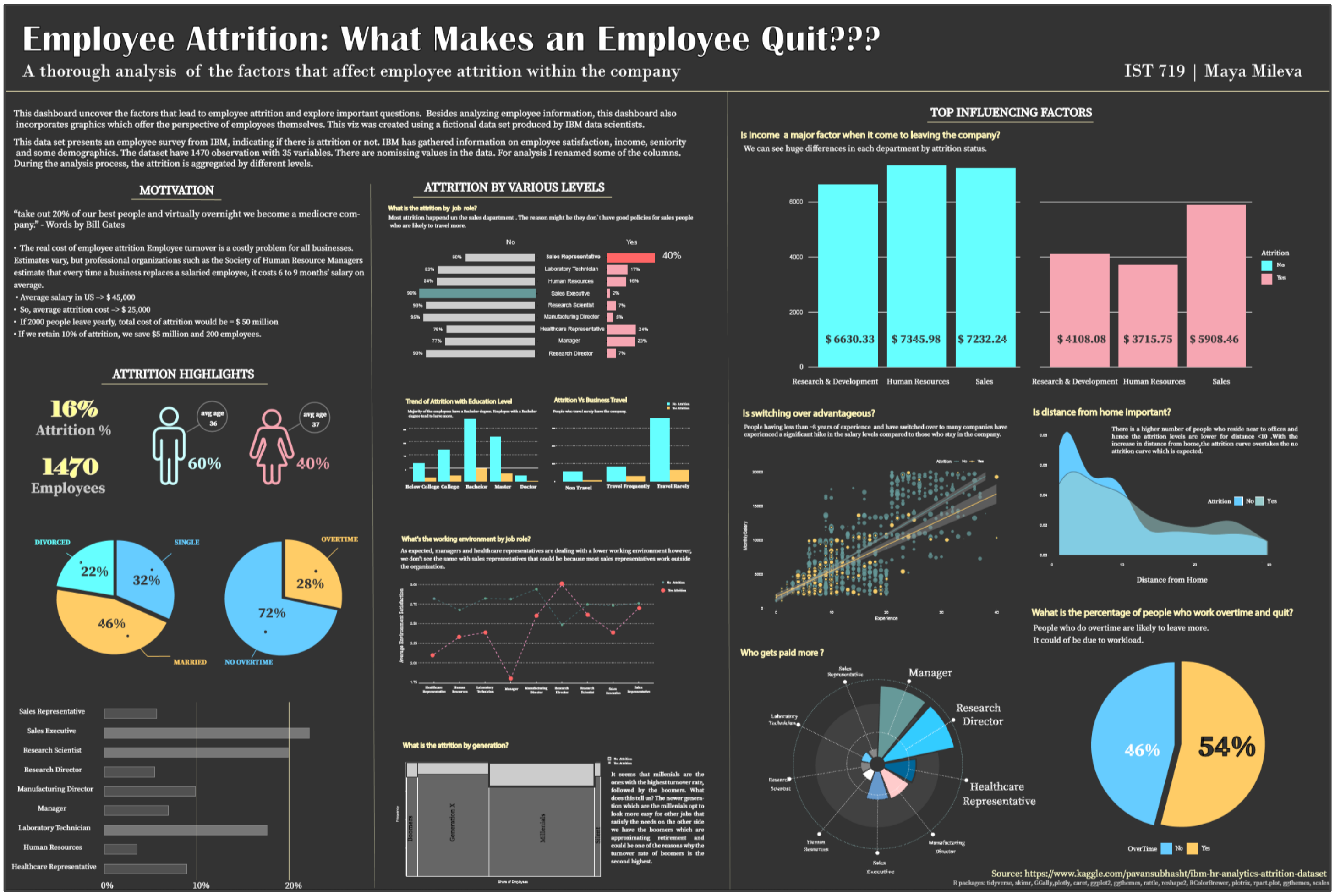

Age Distribution
Female
Male
Years Old
Years Old
density
0.00
0.01
0.02
0.03
0.04
20
30
40
50
60
Age
Mean = 36.92 Years Old
Overall Age

Fig. 2

All my skills developed in the previous classes was used here: I had to clean and preprocess my data and get it ready to be visualized. The poster was designed to provide a hierarchal flow. The rule of thirds alignment was used, along with adequate spacing for rest of eye. The overall font was selected to match the subject of the paper. A mix of single and multidimensional plots were used in order to answer the questions asked. The poster supports the "3 distances" of audience. The top influencing factor was added to the top right to attract audience across the room and hint about the topic of the poster. Colored charts, along with the mid-sized headings were added to intrigue the middle-distance audience. The granular details offered below the title and surrounding the charts are offered for audience up-close. Some of the analytics offered are showing the main insights of the attrition problem.

2.4.2 Reflection & Learning Goals

Fig.3 Final Poster



Employee Attrition: What Makes an Employee Quit???

A thorough analysis of the factors that affect employee attrition within the company

IST 719 | Maya Mileva

This dashboard uncover the factors that lead to employee attrition and explore important questions. Besides analyzing employee information, this dashboard also incorporates graphics which offer the perspective of employees themselves. This viz was created using a fictional data set produced by IBM data scientists.

This data set presents an employee survey from IBM, indicating if there is attrition or not. IBM has gathered information on employee satisfaction, income, seniority and some demographics. The dataset have 1470 observation with 35 variables. There are nomissing values in the data. For analysis I renamed some of the columns. During the analysis process, the attrition is aggregated by different levels.

MOTIVATION

"take out 20% of our best people and virtually overnight we become a mediocre company." - Words by Bill Gates

- The real cost of employee attrition Employee turnover is a costly problem for all businesses. Estimates vary, but professional organizations such as the Society of Human Resource Managers estimate that every time a business replaces a salaried employee, it costs 6 to 9 months' salary on average.
- Average salary in US –> $ 45,000
- So, average attrition cost –> $ 25,000
- If 2000 people leave yearly, total cost of attrition would be = $ 50 million
- If we retain 10% of attrition, we save $5 million and 200 employees.

ATTRITION HIGHLIGHTS

ATTRITION BY VARIOUS LEVELS

What is the attrition by job role?

Most attrition happend un the sales department . The reason might be they don`t have good policies for sales people who are likely to travel more.

Trend of Attrition with Education Level

Majority of the employees have a Bachelor degree. Employee with a Bachelor degree tend to leave more.

Attrition Vs Business Travel

People who travel rarely leave the company.

What's the working environment by job role?

As expected, managers and healthcare representatives are dealing with a lower working environment however, we don't see the same with sales representatives that could be because most sales representatives work outside the organization.

What is the attrition by generation?

It seems that millenials are the ones with the highest turnover rate, followed by the boomers. What does this tell us? The newer generation which are the millenials opt to look more easy for other jobs that satisfy the needs on the other side we have the boomers which are approximating retirement and could be one of the reasons why the turnover rate of boomers is the second highest.

TOP INFLUENCING FACTORS

Is Income a major factor when it come to leaving the company?

We can see huge differences in each department by attrition status.

Is switching over advantageous?

People having less than ~8 years of experience and have switched over to many companies have experienced a significant hike in the salary levels compared to those who stay in the company.

Is distance from home important?

There is a higher number of people who reside near to offices and hence the attrition levels are lower for distance <10 .With the increase in distance from home,the attrition curve overtakes the no attrition curve which is expected.

Who gets paid more ?

Wahat is the percentage of people who work overtime and quit?

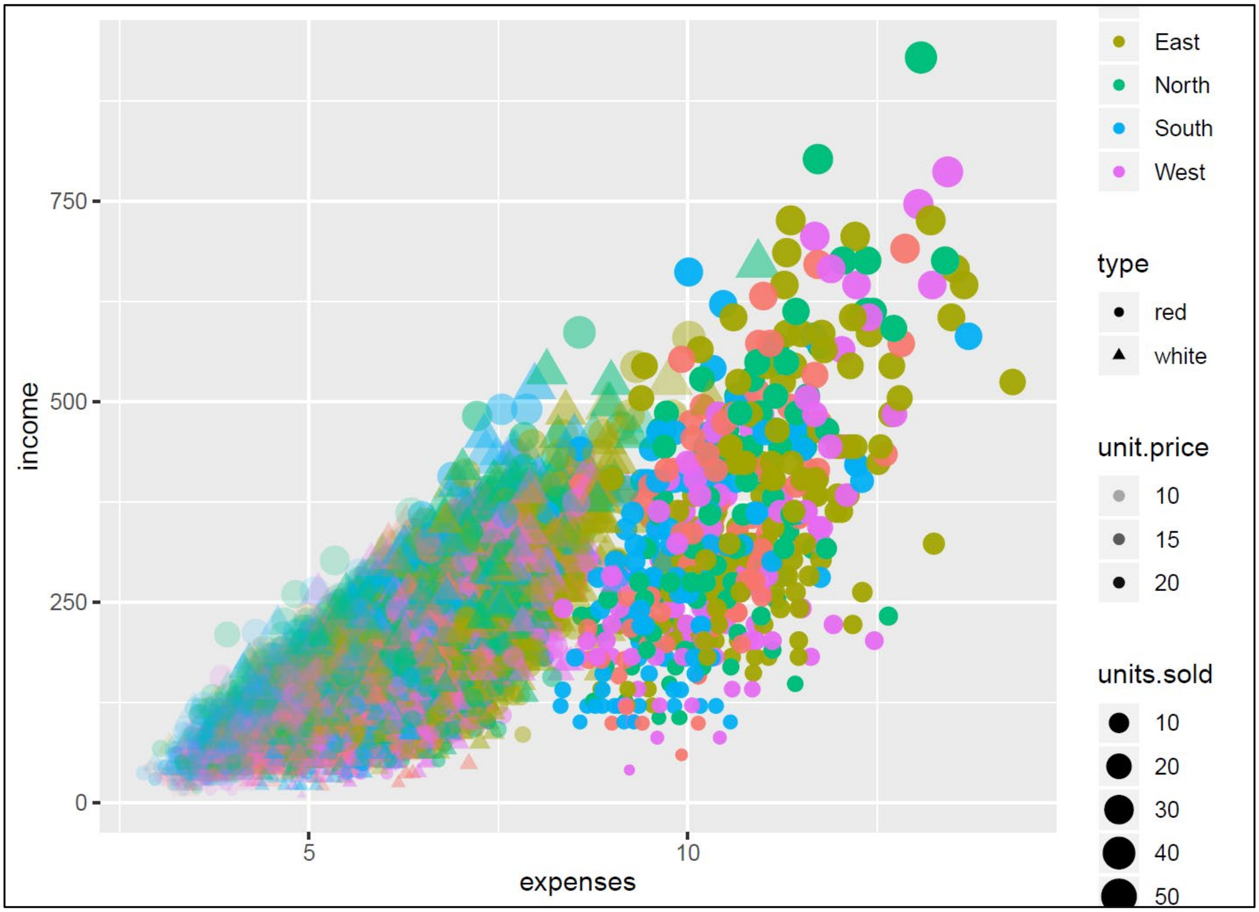
People who do overtime are likely to leave more.
It could of be due to workload.

Source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

R packages: tidyverse, skimr, GGally,plotly, caret, ggplot2, ggthemes, rattle, reshape2, RColorBrewer, plotrix, rpart.plot, ggthemes, scales

In this course we learned how to use R to do basic data cleaning and preparation on a wide range of data sets with functions such as the aggregate/apply functions and the tidyr package. We identified stories in data sets through exploration using R to create appropriate rough plots to identify distributions and relationships in the data. We built pie charts, bar charts, histograms, line, scatter, and box plots. In addition to using the built-in R functions to create these plots, we also were able to draw them ourselves by drawing shapes. We built multi-dimensional plots. We created rich visual artifacts that communicated data stories, as well as how to identify the optimal type of visualization to minimize the viewer cognitive overload and maximize image interpretability. We discussed different color schemes and what is the meaning of colors and how to use them. We were taught new ways of showing distribution and relationships in the data.

Fig.4



We used Adobe Illustrator to enhance our plots from R. We learned to use vector images, instead of raster images, for more flexibility in Illustrator, and to avoid lack of image quality.

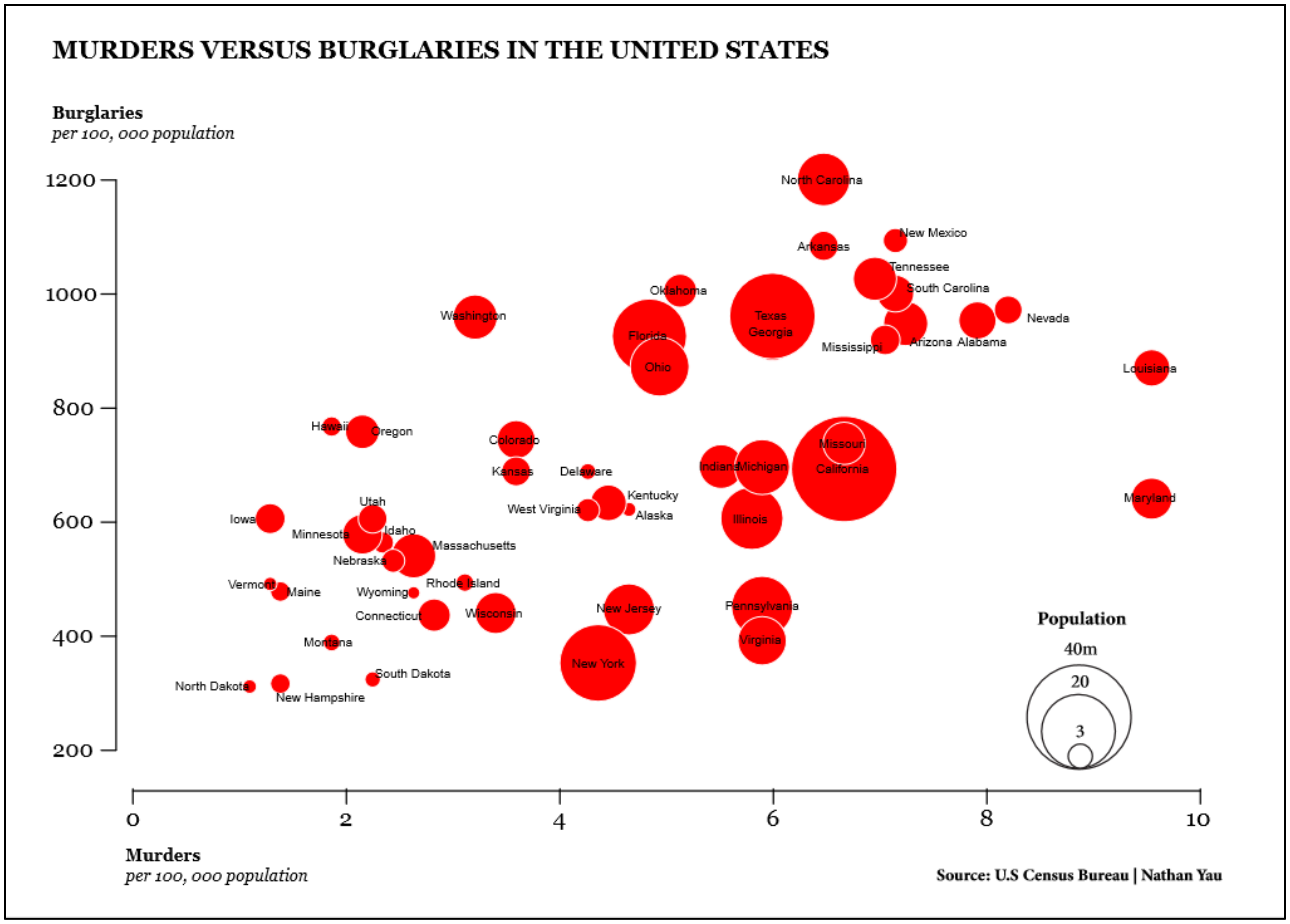
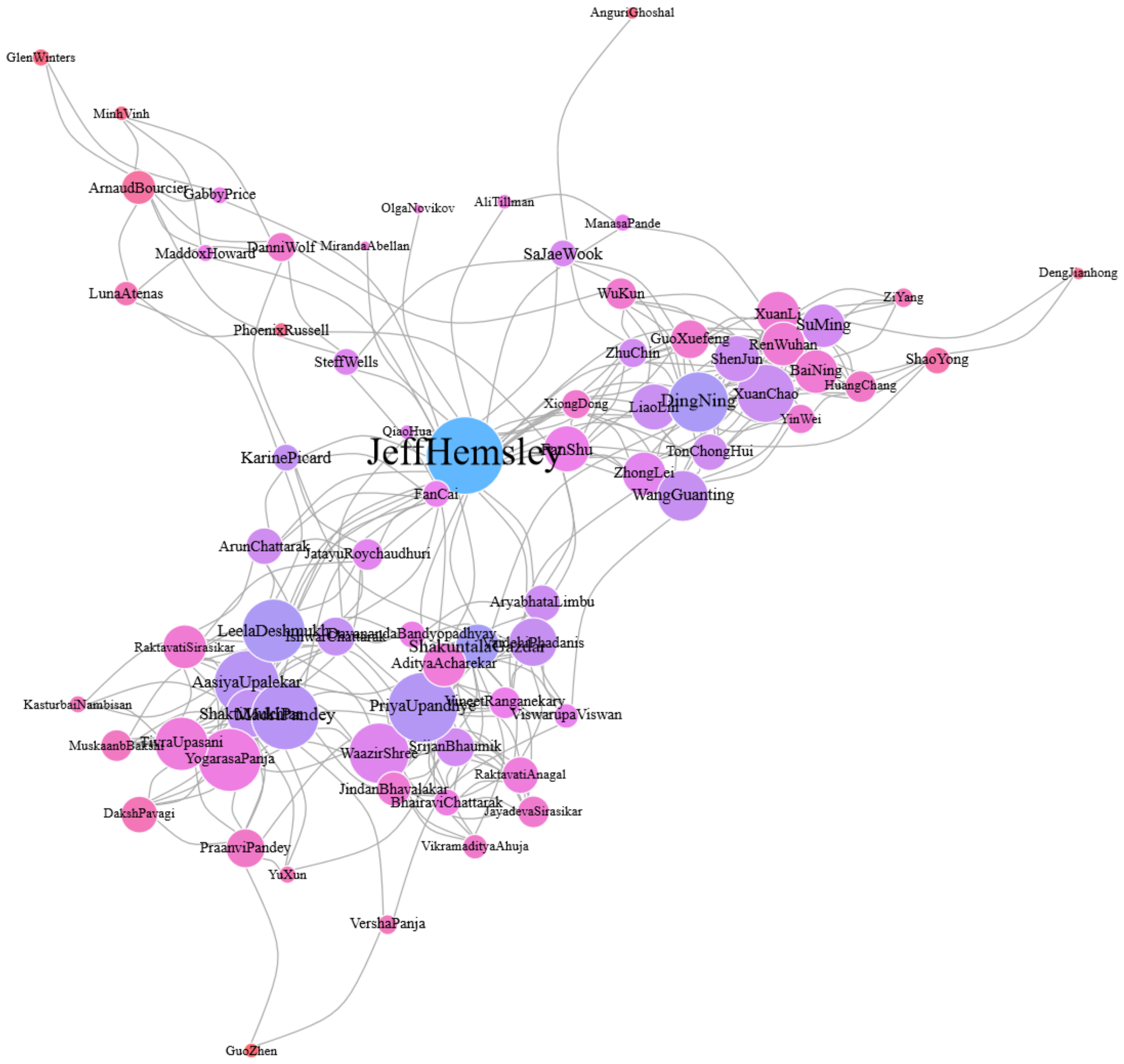


Fig. 5

Layouts were introduced together with grid structures and hierarchy. We learned about the use of composition, to include the use of the Rule of Thirds as well as the Golden Ratio to enforce flow and the use of lines to point to an area of focus. Font styles and fonts sizes were also discussed. We analyzed social media data, learned got to draw a map, control the zoom and add points at certain locations. We discussed ways to study relationships in social media to discover who is important or who is influential, observing the degree, betweenness, and closeness centrality measures. Node structures and network maps were discussed. In the graph below the size of the node is based on how many links the person has, betweenness is captured by how big the name is, and the color captures how close nodes are to other nodes.



We learned how to identify appropriate audience and bring an ethic-based perspective to the development and interpretation of visualizations. We created some 3-D images using the RGL package:

mileva_index.html

We plotted a network map in 3-D and created various interactive dashboard apps using the shiny package.

The final poster provided an opportunity to combine and use all the skills gained through the term. Collecting and organizing the data was just the first step. Understanding of the patterns and showing them with the right visualizations, colors and techniques was the main part. The right communication was the key here. I was able to target audience from different spots in the room and create a visual hierarchy to tell a story. The poster created, was readable for wide variety of audience.

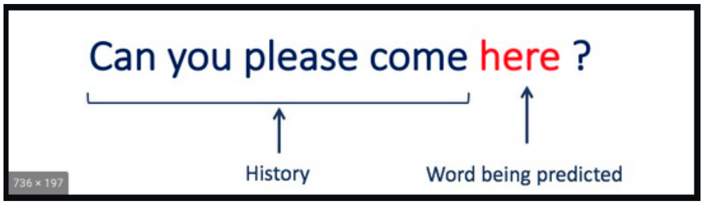
2.5 IST 736 Text Mining

2.5.1 Project Description

The Text Mining course introduced concepts and methods for gaining insight from a large amount of text data. The main goal of this course was to increase student awareness of the power of a large amount of text data and the computational methods used for finding patterns in large text corpora. It introduced text mining technologies rooted in machine learning, natural language processing and statistics. It also showcased the applications of text mining technologies in information organization and access, business intelligence, social behavior analysis and digital humanities.

As a final team project, we worked on building text predicting application for movie reviews with sentiment analysis in real-time. The dataset used for this project was the Large Movie Review dataset provided from Stanford University - dataset for binary sentiment classification containing a set of 25,000 highly polar movie reviews for training, and 25,000 for testing. The goal for this application was to predict the sentiment level (positive or negative) of a movie review in real-time with the ability to recommend the next word that a user is going to type.

N-gram models were built for identifying following word in the sentence.

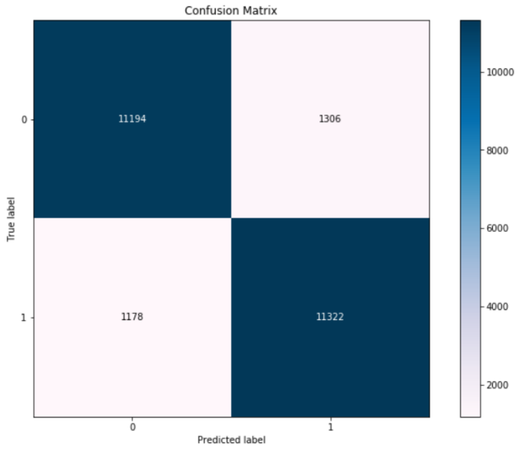


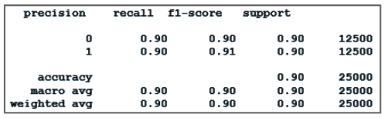
In order to make valid predictions, supervised machine learning models were applied to the labeled data. . Both MNB and SVM algorithms were used to predict sentiment value on the labeled data. The reviews were vectorized as word count, binary, and TF-IDF to compare models. In addition to frequency, the minimum count needed to be recorded as a vector was configured to find the optimal value. Tokenization with and without all lowercase letters, and the collection of unigrams, bigrams, and both unigrams and bigrams together were collected. Once an instance of the vectorizer class was created, the fit() function was called in order to learn a vocabulary from the documents. Next, the transform() function was applied to the documents to encode each as a vector. An encoded vector was returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. After the vectorization process was complete, the MNB and SVM algorithms were applied to the feature sets. While only the default parameters were accepted with the MNB algorithm, the SVM algorithm cost, or c parameter, was set from 1 to 10 to test model accuracy. Once the best model was discovered, the entire labeled set was applied to the model for training and used to predict sentiment on the unlabeled movie review data.

The most indicative words from the models were discovered.

SVM

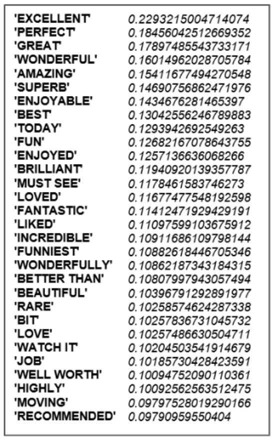
Accuracy: **0.90064**





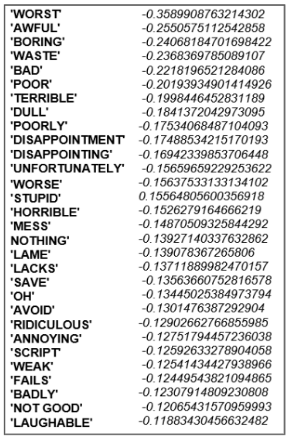
```
              precision    recall  f1-score   support

           0       0.90      0.90      0.90     12500
           1       0.90      0.91      0.90     12500

    accuracy                           0.90     25000
   macro avg       0.90      0.90      0.90     25000
weighted avg       0.90      0.90      0.90     25000
```

Top Positive Features



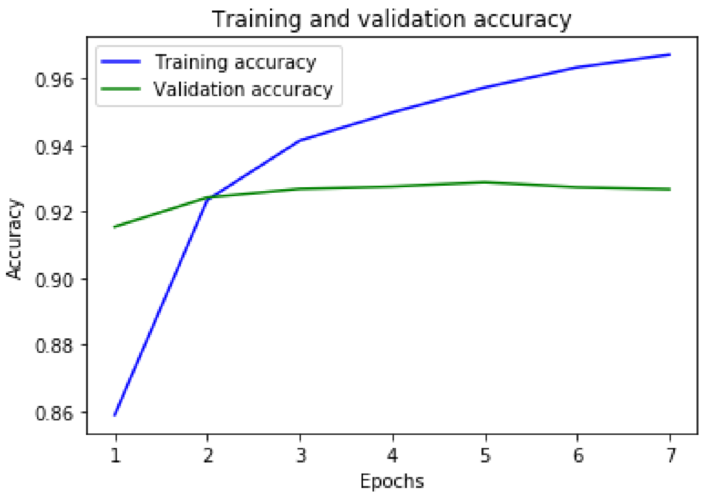
'EXCELLENT'	0.2293215004714074
'PERFECT'	0.18456042512669352
'GREAT'	0.17897485543733171
'WONDERFUL'	0.16014962028705784
'AMAZING'	0.15411677494270548
'SUPERB'	0.14690756862471976
'ENJOYABLE'	0.1434676281465397
'BEST'	0.13042556246789883
'TODAY'	0.1293942692549263
'FUN'	0.12682167078643755
'ENJOYED'	0.1257136636068266
'BRILLIANT'	0.11940920139357787
'MUST SEE'	0.1178461583746273
'LOVED'	0.11677477548192598
'FANTASTIC'	0.11412471929429191
'LIKED'	0.11097599103675912
'INCREDIBLE'	0.10911686109798144
'FUNNIEST'	0.10882618446705346
'WONDERFULLY'	0.10862187343184315
'BETTER THAN'	0.10807997943057494
'BEAUTIFUL'	0.10396791292891977
'RARE'	0.10258574624287338
'BIT'	0.10257836731045732
'LOVE'	0.10257486630504711
'WATCH IT'	0.10204503541914679
'JOB'	0.10185730428423591
'WELL WORTH'	0.10094752090110361
'HIGHLY'	0.10092562563512475
'MOVING'	0.09797528019290166
'RECOMMENDED'	0.09790959550404

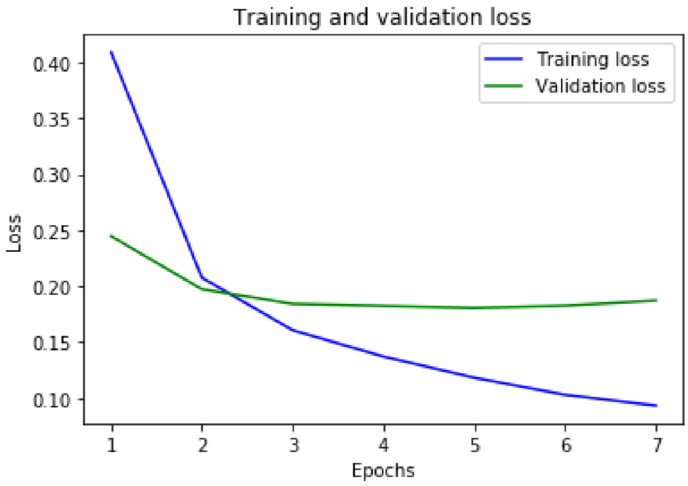
Top Negative Features



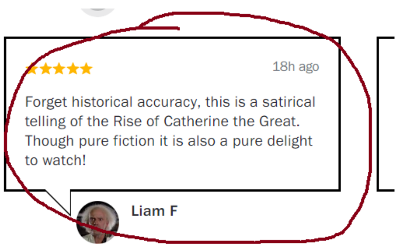
'WORST'	-0.3589908763214302
'AWFUL'	-0.2550575112542858
'BORING'	-0.24068184701698422
'WASTE'	-0.2368369785089107
'BAD'	-0.2218196521284086
'POOR'	-0.20193934901414926
'TERRIBLE'	-0.1998446452831189
'DULL'	-0.1841372042973095
'POORLY'	-0.17534068487104093
'DISAPPOINTMENT'	-0.17488534215170193
'DISAPPOINTING'	-0.16942339853706448
'UNFORTUNATELY'	-0.15659659229253622
'WORSE'	-0.15637533133134102
'STUPID'	0.15564805600356918
'HORRIBLE'	-0.1526279164666219
'MESS'	-0.14870509325844292
NOTHING'	-0.13927140337632862
'LAME'	-0.139078367265806
'LACKS'	-0.13711889982470157
'SAVE'	-0.13563660752816578
'OH'	-0.13445025384973794
'AVOID'	-0.1301476387292904
'RIDICULOUS'	-0.12902662766855985
'ANNOYING'	-0.12751794457236038
'SCRIPT'	-0.12592633278904058
'WEAK'	-0.12541434427938966
'FAILS'	-0.12449543821094865
'BADLY'	-0.12307914809230808
'NOT GOOD'	-0.12065431570959993
'LAUGHABLE'	-0.11883430456632482

Other models included in this project were k-mean clustering, topic modeling, neural network.



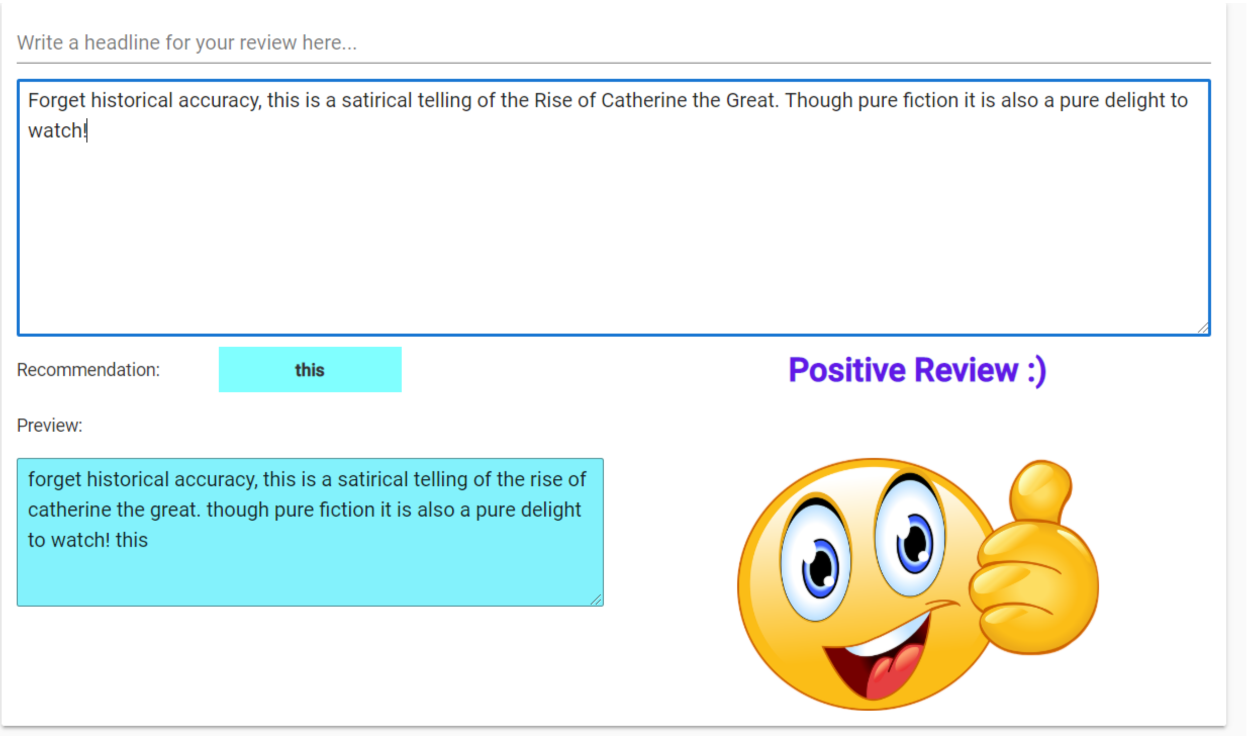


The final application was hosted on the Anvil website.


★★★★★ 18h ago

Forget historical accuracy, this is a satirical telling of the Rise of Catherine the Great. Though pure fiction it is also a pure delight to watch!

Liam F


Write a headline for your review here...

Forget historical accuracy, this is a satirical telling of the Rise of Catherine the Great. Though pure fiction it is also a pure delight to watch!

Recommendation: this

Positive Review :)

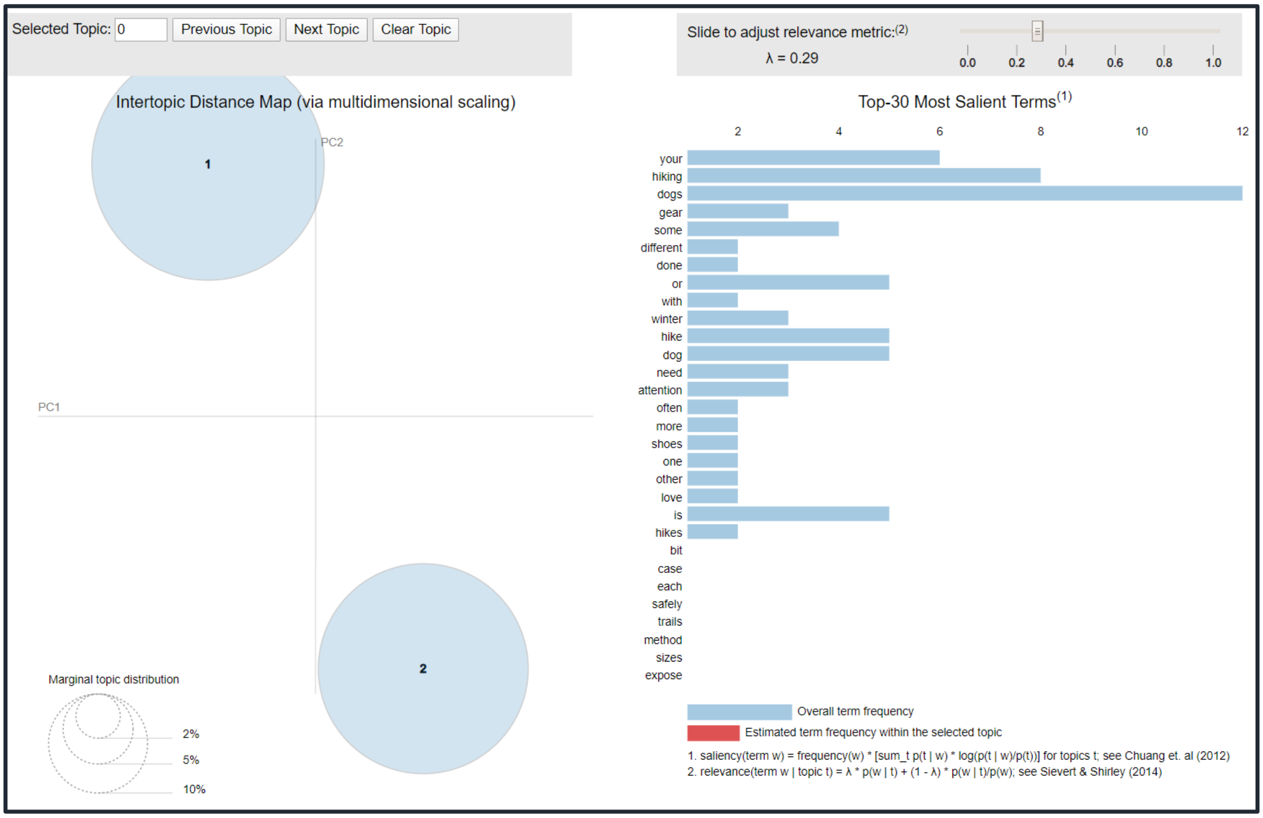
Preview:

forget historical accuracy, this is a satirical telling of the rise of catherine the great. though pure fiction it is also a pure delight to watch! this

2.5.2 Reflection & Learning Goals

The Text Mining course was one of the most beneficial in the whole program. We started by looking at the vectorization process, regarding both what to count and how to count that would be best for the goal. This included the stemming, and whether we want to eliminate important linguistic information. We looked into the application of the Multinomial Naïve Bayes (MNB) machine learning algorithm for text mining. For an assignment, I applied NLTK's NB model to customer reviews for sentiment analysis and for lie detection. I used a 10-fold test for evaluation. While stemming did not provide better results, the filtering of stop-words did. An additional step of allowing just POS adjectives words as features did not yield better results. An information gain test showed four words that most influence sentiment value in the dataset are "amazing," "terrible," "took," and "best."

We used the Amazon Mechanical Turk website to collect manual sentiment classifications on restaurant reviews. To determine each worker's quality of work, Kappa values were calculated and compared to the ground truth. We compared the Benoulli (BNB) and Multinomial Naïve Bayes algorithm. In an assignment, I used both to perform sentiment and lie detection analysis on customer reviews. I used the words were tokenized as binary for BNB and both normal and term frequency-inverse document frequency (TF-IDF) for MNB. Unigrams and bigrams were tested under each algorithm, along with stop-words, minimum frequency, and conversion to lowercase. We used evaluation methods to evaluate machine learning prediction accuracies. In an assignment, I applied the MNB and SVM algorithms, with different options, to movie reviews and compared models in a confusion matrix and calculated precision and recall. Word gain was also used to report the most informative words. We learned about topic modeling and used Latent Dirichlet Allocation (LDA) to summarize the main topics of the floor debate of the 110th Congress. I used MALLET with different topic sizes and compared the relevancy of the results until an optimal size was achieved. After reducing the topic size to 45, since there usually are about 40-50 topics per floor, and removing stop-words, the best results were achieved.



Selected Topic: 0

Slide to adjust relevance metric:(2)

λ = 0.29

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

This project contributed to the successful execution of the learning objectives where data was collected and organized and cleaned, data mining techniques were leveraged to reveal patterns from within the data, which were further highlighted using visualization tools. Communication skills were displayed in the translation of the insights into easily consumable information. The ethical dimensions of data science practice were also considered in the collection of reviews.

3. Conclusion

As a student of the MS Applied Data Science program at Syracuse University, I have been exposed to a broad range of areas related to data science. I've collected data, either by using the DMAIC process to define and measure data in the process improvement project in MBC 638, by collecting missing information for all the states in IST 687, to using a variety of tools to import, clean, and transpose various types of data files in all of the courses taken in the program. I have identified patterns in data by using R to create visualizations such as bar plots, pie charts, box plots, scatter plots, line charts, geographical maps, and network maps in IST 719. I was able to apply data mining techniques by performing statistical analysis on many assignments to observe distributions and relationships. Current project I am working on is an application, that is going to predict the sentiment towards a movie, by typing the review. Machine learning was applied to various datasets, such as to animal adoption in IST 707, predict sentiment analysis on movie review data for classes like Text Mining and Natural Language Processing, and identify the anonymous authors of the Federalist Papers by applying clustering and decision tree models in IST 687. I was able to develop alternative strategies based on the data from the final project in MBC 638, as well as give recommendation for adopting animals. I have acquired communication skills as a data scientist, either through IST 772 (Statistical Methods in information science) by using language that accurately describes the uncertainty of inferential statistics, or through live-session presentations in IST 707, IST 687, and IST 719. Through the attrition project in IST 719, I was able to target audience from different spots in the room and create a visual hierarchy to tell a story.

We held multiple discussions on ethics in the various classes. Privacy issues, for example, could be violated in the health care industry when predicting when a patient will come for a visit due to health issues. Another discussion we had on ethics was on machine learning bias, and how data scientists should be cognizant of this. An example is that when facial recognition is more likely to recognize Caucasians than African Americans.

My journey in Syracuse University has been greatly appreciated. The skill I gained these 18 months has helped me in succeed in my current career and I cannot wait to see what is next. It wasn`t always easy, but at the end I know I am ready to face all the new challenges as a data scientist.

Syracuse University's School of Information Studies provides students the opportunity to synthesize the collection, management, and analysis of data, as well as the delivery of actionable insights using various data science techniques. Skills learned in the program have developed a multifaceted approach to solving structured and unstructured data problems, it has also cultivated strategies that improve organizational efficiency. The program has fostered a practice of transparency, reproducibility, and ethical data management which promotes integrity and credibility within an organization's analytics team. Using the methods learned at the School of

Information Studies, data scientists are equipped with the ability to tackle a wide range of problems and the resources to explain observations to a variety of stakeholders and business professionals.

4. References

School of Information Studies, Syracuse University. (2019, December). MS in Applied Data Science.

Retrieved April 15th , 2020 from https://ischool.syr.edu/academics/graduate/masters-degrees/ms-in-applied-data-science

Austin Animal Center Shelter Intakes and Outcomes (80,000 Shelter Animal Intakes and Resulting Outcomes)

Retrieved Nov 7th, 2019 from https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes

Analytics Employee Attrition & Performance

Retrieved Jul 16th, 2019 from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset