Alejandro Mora

Syracuse University | Syracuse, New York | SUID: 445623874 | Email: Almora@syr.edu

Portfolio Milestone M.S. in Applied Data Science

Data Science Project Analysis

**Table of Contents:**

1. Introduction (Page 2)
2. Purpose (Page 2)
3. Projects
   1. Looking at Data using visualization (Page 3)
   2. Analyzing Text using Natural Language Processing (Page 7)
   3. Apply Data Analysis methods to data (Page 10)
   4. Using Data Analysis in Big Data(Regression) (Page 14)
   5. Using Data Analysis in Big Data(Time series) (Page 17)
4. References (Page 20 )

**Introduction**

The program of Applied Data Science here at Syracuse University teaches all that is necessary to understand and apply the knowledge of Data Science into real world situations and help solve business problems. The program helps you learn key topics while teaching you how to apply them instead of the traditional, read from a book and memorize definitions. With applications to scenarios and problems, the students can better learn the area of study in each course taken and build experience. While learning in an environment with other students, help can be taken from minds across a variety of backgrounds and gaining knowledge in different fields in general. While in the program, I have learned from not only the professors, but other professionals within the courses themselves. Data Science in general consists of a mix of topics and areas of expertise, all of which can be taught through this program. During the program I have learned many valuable skills for the workforce and have worked on projects to develop these skills needed for applicable problems today.

**Purpose**

The motivation for this paper is to analyze projects I have created within the program, to better understand the process of Data Science and display what I have accomplished throughout while displaying my knowledge within the field. Throughout the entirety of the program in all my courses, there have been many assignments and projects created that showcase knowledge, however there were several ones that have stood out as far as the Data Science process. Collecting and organizing data, identifying patterns using visualizations and data mining tools, and trying to find alternate strategies after data has been analyzed; all have been done in projects.

Looking at Data Using Visualization

**Project Description and Goals**

Information visualization is strong tool to learn while analyzing data within the Data Science field. It is strong because it helps display information to everybody, regardless of knowledge of analytics. Visualizations should be easy to read and interpretable at all levels and must be able to tell a story without background knowledge of a specific area. Creating poster boards of information on data sets is very helpful in doing this and builds knowledge of creating a story to tell. In this project, I have created a poster describing how music has changed over time and I analyzed what is the most popular music in today’s time. My goal in this project was to extract data from a source, create questions I wanted to answer with the data, and explain the answers to those questions using visualizations within one poster board with text to help. The poster had to have a purpose as to what I was trying to solve and why the information was valuable to myself and an audience. The audience for this specific study is music lovers and music researchers around the industry as the data is from Spotify. If you are part of the music industry, research needs to be done to see what is doing well and what is not as far as music that is being released. Music labels and production industries can see what is popular to take advantage of the listeners’ interests.

**Data**

To start the poster, a question of what data to use was in order as I had to pick something that would provide a valuable analysis. As a person that is interested in music, I decided to go that route and find something revolving the area of interest. From the data source, Kaggle, a data set was imported from a creator that used Spotify and its web API to extract a large amount of data from songs throughout the century. The data ranged from 1920-2020 and contained great information about songs regarding, popularity scores and attributes of songs that Spotify has created from their API. There is a total of over 174,000 songs and each observation contained 19 variables to give depth to analysis. The data variables that were largely looked at were the artists, release dates, popularity scores, song attributes, and maturity ratings.

**Questions and Methods**

In analysis, the main question at hand was what drives music’s’ popularity and how is music changing? From that, four questions can be derived to be more specific and can be solved using the data:

-What song traits are most common over the past century?

-Has music popularity gone up or down?

-Do explicit ratings affect popularity?

-Who are the most popular artists that are streamed over the past century?

While using the programming language ‘R’ to help in creating visualizations, a couple of libraries were used to help without using the base R functions(libraries: ggplot2 and Vioplot). With the creation of the plots in R, all plots were exported into Adobe Illustrator to help with color schemes, sizing, and formatting to make the plots more easily readable for the final poster. This was done to help with the hardships that R has in formatting and its lack of functionality within that area. The poster itself was created and edited in Adobe Illustrator while using the rules of proper visualization techniques as far as the color schemes and layout to allow any audience to examine the results with ease and make the poster appealing. Five charts/plots were created using R: a line chart, histogram, scatter plot, violin plot, and a bar plot. All five plots help answer the four questions presented.

A picture containing text

Description automatically generated

**Spotify and the Progression of Music** (Figure 1)

Five single and multivariable charts showing exploratory data analysis of the song dataset from Spotify in a poster format.

**Analysis and Results**

When looking at song traits using a line chart, the five most easy to read song characteristics displayed, have a strong change over time based on averages of the songs present within the data. This shows that music has strongly shifted away from its strong instrumental presence and vocals while staying the same in danceability and increasing in energy levels. Music popularity has risen over time in popularity and the amount of songs that being produced. More people are listening to music, but they also most interested in recent music compared to older music. On average, explicit content tends to have more popularity than clean music showing that people would rather have vulgar music nowadays. Lastly, many older artists still hold their popularity amongst artists, but some newer ones are coming up on the scale.

**Learning Outcomes/Skills**

The main learning outcome of this project was to learn how to display information in a way that can be read by a majority audience. Being able to display information visually to display results is a strong skill to yield in Data Science. If plots on posters do not match a scheme, layout, or purpose then the charts can be useless as people will turn away from the results and seem meaningless. With the skills of collecting data, creating charts using programming languages such as R, organizing visualizations, and using Illustrator to enhance plots, I can continue progressing in the Data Science field. Some people have trouble displaying results to reach the audience at large and only find ways to display information to the smaller group that will understand. I can strongly say that I can gather data and answer questions based on the data while also presenting the data. A thing to work on is the ability to recognize patterns better and determine the type of plots to use for certain situations to make sure the data is presented correctly. Multiple edits had to be made throughout the process as multiple plots were not answering the questions being asked.

Analyzing Text Using Natural Language Processing(NLP)

**Project Description and Goals**

Moving into text analysis, my goal for this project was to see how well a program can be improved to determine contact information within a body of text; the most True positive recognitions is the desired outcome. Being able to use and recognize regular expression usage within programming languages is also valuable in the Data Science field. Regular Expressions are used to help determine pieces of text that might be good for company usage in day to day processes or bad information for products and business processes. Natural Language Processing is also being used and improved on every day and really helps with data extraction, Artificial intelligence, and privacy. All of which are important in the field. For this project, finding contacts within emails(contact emails and phone numbers) is important when dealing with privacy and helping companies, determine who and what needs to be selected without manually screening emails for contacts. Throughout the process, it took time and patience as it can be struggling topic but can be improved with time and practice. My other goal within the project was to gain that experience with working with regular expressions to help evolve my career portfolio and gain an edge in NLP to use for my future career.

**Data**

The data was presented in notepad text format to represent the email data and programs were written in Python to search the files individually and display a resulting email and phone number. A snippet from a sample piece of data was formatted as follows:

<address>engler WHERE stanford DOM edu</address>

This displays an email address within the text and regular expressions has to recognize that line of text and format it into a format of email that is recognizable.

**Methods**

In the Python program itself, premade functions were used to help with recognizing the correct format from a golds list and comparing the pre-formatted data used with the regular expression created. The results were displayed with a count of True Positive, False Positive, and False Negative comparison results. Each pattern created was appended to a set variable to use with the functions and can be whatever pattern was deemed reasonable to recognize the emails and phone numbers. The format of regular expression that was a base used, was in this format:

epatterns.append('([A-Za-z.]+)@([A-Za-z.]+)\.edu') #orignal line

ppatterns.append('(\d{3})-(\d{3})-(\d{4})') #original line

All patterns were worked off of these base patterns to help recognize any format given within the text files. Using an online tool to create patterns in regular expressions, multiple patterns were created and appended to recognize as many as possible.

**Analysis and Results**

Out of 119 possible correct recognitions all together, the final results of created patterns for both emails and phone numbers were 104 True positives, 2 False Positives and 13 False Negative results. The results were not as best as one would want, as for a company, that is 15 contacts that are not accounted for and can be costly depending on what the contacts are being used for. Many forms of emails were being recognized as other words or patterns were being overwritten by other patterns and passing through emails that were supposed to be recognized. This is a process of being a natural language programmer to determine what patterns are useless or what patterns are being used twice which can overlap and waste memory space.

**Learning Outcomes/Skills**

The biggest take from this project was that using regular expressions which is a small part of Data Science, can be quite challenging but in every mean, very important. Getting the correct format for patterns and installing them into data can mean everything when using it for company use. For this project, there was a check list of the correct formatting that be re-run and constantly used to try to solve for the correct format, whereas with a large dataset of actual contacts, there is no correct data set to put checks in. The patterns must match a whole set at large not knowing the correct address. If there are too many False Positives in your patterns, there might be many contacts listed, but many emails or numbers can be wrong and therefore invalid for use. The learning outcome is to take whatever emails and numbers that weren’t recognized and study them to create patterns later on to recognize them and debug the programs. Like the other parts of Data Science, natural language processing is a desired skill to learn but will take practice to master and I can use the skills I have now to practice more and be able to create patterns that can find contact information more efficiently.

Apply Data Analysis Methods to Data

**Project Description and Goals**

Moving onto more statistical modeling, for this project I worked with a fellow student within the M.S. in Data Science program here at Syracuse. My partner, Pedro Pereira, and I worked on a data set and created multiple models to solve questions created by us. Throughout the project, there was the gathering of data, cleaning data, exploratory analysis, statistical modeling, and analysis with recommendations. All of these steps provide the steps required for a Data Science project. Our data was large and was analyzed by us both throughout a 6 week span. Using our combined knowledge, we were able to mix programming languages and work off of each other’s skill sets. Whatever he was strong at, he would provide lots of input in that area and vice versa. Our goal for the project was to develop our skills in performing and presenting a Data Science project while representing our findings with valid models and reasoning. This project helped us both in developing our Data Science skills to get us further in the program and careers. I will be going over the parts I contributed most on the project and areas I can work on for the future.

**Data**

The data used for this project was accident data within the United States, the data was found on Kaggle and the user created the data using MapQuest and Bing API from the span of January 2016 to June 2020. Each observation was provided with 49 variables which led to us looking at a lot of data to clean and separate for modeling. However with all the data, this led to creating a lot of plots for exploratory analysis and modeling options.

**Questions and Methods**

The main question being asked by us during the data exploration was what causes these accidents? We split up the questions to be listed as this

-What correlations are present in physical environment factors in accidents?

-Can we predict accident severity?

-Can we predict accident duration?

We wanted to know what caused the most accidents, how bad can accidents get based on factors, and how long do accidents last to avoid future disruptions to traffic. We were trying to get on the side of government officials in accident prevention, whether it be highway or local level, and see what can be done to avoid the worst. We used Association rule mining, Naïve Bayes, Random Forest classifiers and Decision trees to help solve our problems. The models I was responsible for were the association rule mining and Naïve bayes models for predictions.

**Analysis and Results**

When looking at association rule mining, the question trying to be answered was what physical factors are present in the surrounding area of an accident. The data set provided a mix of factors available for each observation with the API and rules can be made to see which most are most closely together. With the large data set, almost 50000 rules could be created with high confidence and lift.



**Association Rules of Car Accidents** (Figure 2)

Matrix showing almost 50000 association rules for car accidents with a scale based on lift value. Gathered from page 10 of “Car Accidents in the U.S.” (Mora, Pereira 2020).

## lhs rhs support confidence lift

## [1] {Traffic\_Signal=False,

## Traffic\_Calming=False,

## Stop=False,

## Station=False,

## Railway=False,

## No\_Exit=False,

## Give\_Way=False,

## Crossing=False,

## Amenity=False} => {Junction=True} 0.2626 0.9999 3.694

## [2] {Traffic\_Signal=False,

## Stop=False,

## Station=False,

## Railway=False,

## No\_Exit=False,

## Give\_Way=False,

## Crossing=False,

## Bump=False,

## Amenity=False} => {Junction=True} 0.2626 0.9988 3.690

**Top 2 Association Rules of Car Accidents** (Figure 3)

Table of top 2 association rules created using the car accident data, shows variables associated, lift, support, and confidence. Gathered from page 11 of “Car Accidents in the U.S.” (Mora, Pereira 2020).

With he rules sorted by highest confidence and lift, whenever there is junction present, or an area of changing directions on a road, there was no traffic signal present, no stop signage, or traffic calming signals in the area of an accident. This is pretty big considering the instances of this rule occurred around 280000 times for the firs rule and around the same for the second rule. There were many accidents were drivers were changing directions and no traffic help was present.

Looking at Naïve Bayes, we were trying to predict severity of accidents based on many factors, including physical environments, and the physical environment in general. The more we know about how dangerous some areas can be, the more precautions that can be placed in those areas. Our model was only able to come back 73% accuracy in predicting the severity of accidents based on those factors and had a very low Kappa. This could be because of the high volume of data but it was not the best accuracy when predicting severity.

In all, our conclusion for the models was that more traffic signals need to be present in as many junctions as possible as they are the most vulnerable to accidents. During our other models, it was being said that there is longer clean up of accidents during inclement weather so there would just have to be more training in cleaning under bad weather conditions to not disturb the flow of traffic for how bad it already might be.



**Naïve Bayes Matrix** (Figure 4)

Matrix of accuracy regarding a Naïve Bayes prediction on severity of car accidents. Gathered from page 15 of “Car Accidents in the U.S.” (Mora, Pereira 2020).

**Learning Outcomes/Skills**

While doing this project, my partner and I learned a lot from each other when looking at the data. Coming up with a problem, gathering the data, cleaning the data, performing exploratory analysis, creating models, and solving for a solution was done with teamwork. In saying that there were areas that obviously were not worked on the most by both of us as one or the other took responsibility for it. He was mostly leading the way for data cleaning as he was more experienced in Python at the time, and it had more capabilities to clean the data easier. Using some packages in R, I was able to make some charts and plots easier as I can perform more functions with R rather than Python. Some models were also more accessible or were just better performed in R when I worked on the association rules and Naïve Bayes. There was more room for editing parameters in R for this data set. In all, I feel like for me I can learn more about data cleaning and exploratory analysis. The data cleaning in a Data Science project is probably the most important because in order to have good models, you need valid data. Performing more projects like this can help in getting better at the Data Science process as I will talk about in the next two projects.

Using Data Analysis in Big Data(Regression)

**Project Description and Goals**

Progressing into more detailed Data Science projects and dealing more with Big Data, I will talk about a project that used the Data Science process while working by myself to solve for the solution of the problem at hand. Using a linear regression model, the problem to be solved was to determine a salary of a football coach within any division 1 program across the NCAA. The goal of the project was to look at the big picture of things and see, what can be used to solve for this other than one piece of data. Multiple data sets will have to gathered to get a more accurate prediction and sources will have to be found without an area to start. The end result from the project wasn’t an exact answer, however it is teaching you how to do research and determine what is valid for modeling and what is not. If working for a school’s board of athletics, seeing how much a coach is worth at your school is important to know so not too much money is spent in that area of budget. There was a result listed for the Syracuse head coach, but it was a side goal to see how to implement the model and see if it is possible to solve for one specific result if asked to do so.

**Data**

With having one data set containing coach information and current salary throughout the NCAA, one data set will have to contain football program records from previous years, one will have to include graduation ratings within the football programs for each school, and one will have to take the school’s stadium capacity into account to develop a well rounded model base. Some schools had to be dropped as not enough information was present, like salary being cut out or other information being private so they would skew the models. Variables including the ranks, wins, loses, conferences were also valuable in having in the data for later analysis.

**Methods**

To start analyzing the variables at hand, some correlation matrixes can help determine what will worth looking at and what is not significant, comparing conferences to other variables can help in seeing what is going on within different categories of schools. An all around correlation comparison of Total pay of coaches and all other variables is also viable to see if some are stronger than others.



**Matrix Correlation** (Figure 5)

Correlation plot of variables against Total Pay. Gathered from page 6 of “Coaches’ Salaries” (Mora 2021).

Some variables had to be taken out of the linear regression model as some variables are too closely related by nature, or in other words are directly related to the pay itself. Like “bonus” or “coach”. Three models were created using different variables to see which one was showing a better R-squared value and AIC value. A model with conference, stadium capacity, SRS rating(Average Point Differential and Strength of Schedule), and Rank were used to determine salary predictions.

**Analysis and Results**

Using that model to determine a predicted salary, the Syracuse head coach was inputted into the model and resulted in a $3,450,656 salary for being in the ACC. Looking at a change of variables, if Syracuse were in the Big-10 conference, the pay would increase to $3,618,445 for the football programs head coach salary. Schools that were removed during this prediction were the independent schools, the reason that they were removed it because most of these schools lacked a coach’s total pay information. These schools could have highly skewed that salary prediction. Even though multiple data sets were used, and many variables were taken out of the regression, it is good to analyze as much information as possible to make a better prediction rather than selecting one data set and just “accepting” the results. More data could have also been included to help increase the accuracy of the model, things like merchandise sales, city populations and maybe even income levels within the area of the schools can help in determining salary levels for coaches.

**Learning Outcomes/Skills**

Working on this project greatly enhanced my skills in the Data Science skills as a whole. Since working on my own, I was able to perform all steps by myself which gave me the experience to work on things that I would recently struggle with. From before it was the data cleaning working in Python. All steps were done in Python and it forced me to get more comfortable with the language. It was the best program, and it can be shortened but with mastery, comes practice which I have gotten and could get more with more projects. My analytical skills have surely heightened as I can figure out if information is missing in solving problems or I can recognize that some information can be added to increase the project’s performance.

Using Data Analysis in Big Data

(Time Series)

**Project Description and Goals**

Looking more into the process of a Data Science project, another project can be seen that I have worked on to display the progress on improving my all around skills. For this project, looking at house pricing data, the goal is to see what areas or zip codes within the United States are worth investing in for property. If a company is looking to invest, they want the best return on their investments and want to know where to go. This project uses time series analysis to look at listing throughout many years of data and using down-sampling techniques, zip codes can be broken down to find the best areas. My goal revolving around this project was to also improve my knowledge in data exploration, cleaning, and modeling to help solve real world problems.

**Data**

The data used is from Zillow, it is their housing data from real estate listings that they have made public for analysis. It is for Single Family Residence listings only and has variables that include zip codes, years, and prices. When breaking it down to a specific area of houses, I took the top 25% of the listings based on their given rank by Zillow. I then down-sampled again to the top 10% of that data based on the return values since they were first listed and decided to analyze the top 10 listings from that list. The 10 listings were based on how high the returns were and listings were removed to only have one listing per city(whichever was highest). This was to avoid city bias and not having all the listings in one city.

Top 10 areas/zip codes:

- Los Angeles (90027)

- Oakland (94610)

- New York (11216)

- San Diego (92104)

- Palm Springs (92264)

- Jersey City (07302)

- Philadelphia (19125)

- Atlanta (30310)

- Alameda (94501)

- Boston (02125)

**Top 10 Areas/Zip Codes** (Figure 6)

Top 10 areas based on ranking from Zillow. Gathered from page 3 of “Zillow House Listings” (Mora 2021).

**Methods**

For time series analysis, an additive time series model was used in Python. The library used was “Prophet” created by Facebook. To get good results, the confidence interval was set to 95% and was also set to predict 60 months ahead of the last date to get a 5 year prediction.



**Time Series Plot** (Figure 7)

Time series showing predictions of top 10 areas. Gathered from page 5 of “Zillow House Listings” (Mora 2021).

**Analysis and Results**

To determine which zip codes were getting a better prediction, the difference predicted outcome of price amongst the trend line was the qualifying factor. The zip codes with the high trend difference or higher percent changes, were taken into consideration for investment opportunity. The three options were Los Angeles, California(90027), Alameda, California(94610), and Brooklyn, New York(11216). This analysis and result are subjective. My analysis revolved around down-sampling and choosing what factor makes the zip code the “best” option. Anyone can run the time series model but pick different areas based on different factors. This is what makes the process of Data Science interesting as many people can run the same experiment but receive different results.

**Learning Outcomes/Skills**

Something to learn about this project is that even though received an answer as to which areas are better to invest in, there is still the area of study to improve on. My results in no way are exact or the best, but they are proven by what I’m trying to solve and the process I took. I could have done a whole population of zip codes in the time series analysis and predicted on the whole United States, however I decided to down-sample which some will, and some will not. My weakness for this project was that I was too focused on finding something too specific when I could have looked at the bigger picture and maybe found different areas that were probably better to invest. This is still all part of the learning process and something that I learned to look out for in the next attempt at something like this; I need to look at all options to see which can be better. Performing different models on the same data can help in improving results.

**References to Projects:**

**Looking at Data using visualization (Page 3)**

<https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>

**Analyzing Text using Natural Language Processing (Page 7)**

regex101.com

**Apply Data Analysis Methods to Data (Page 10)**

Administration, N. (n.d.). FARS Encyclopedia. Retrieved November 15, 2020, from https://wwwfars.nhtsa.dot.gov/Main/index.aspx

Brown, H., Jr. (2018). Leading Causes of Car Accidents With Statistics. Retrieved November 12,

2020, from <https://www.jdsupra.com/legalnews/leading-causes-of-car-accidents-with-> 60370/

Dataset:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath.“A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu,

and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International

Conference on Advances in Geographic Information Systems, ACM, 2019.

Fellow Student Contributor: Pedro Pereira, [ppereira@syr.edu](mailto:ppereira@syr.edu)

**Using Data Analysis in Big Data(Regression) (Page 14)**

Coaches Data set and Salaries:

https://github.com/2SUBDA/IST\_718/blob/master/Coaches9.csv

Graduation Rates:

https://www.ncaa.org/about/resources/research/shared-ncaa-research-data

Stadium Size:

https://github.com/gboeing/data-visualization/blob/master/ncaa-football-stadiums/data/stadiums-geocoded.csv

Records:

https://www.sports-reference.com/cfb/years/2018-ratings.html

**Using Data Analysis in Big Data(Time Series) (Page 17)**

**Zillow Housing Data:**

https://files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv

**Variable Dictionary Help:**

https://sls.gatech.edu/sites/default/files/documents/Toolkit-Docs/hong\_fieldguide\_zillow.pdf