Peter Dolan

DS-401

Professor Prince-Nelson

March 11, 2025

Reflection Essay

Outline

For this data science business analytics minor, I have taken a couple of courses that relate to business topics and more importantly data visualization and interpretation. Over the last three years I have learned about financial statements in Accounting 101, the basics of data interpretation in INTR 202, SQL based databases in BUS 315, a general working knowledge of python from CSCI 111, and an introduction to data interpretation for mind analytics in CBSC 309. While working through this minor I have also been completing a mathematics major, which has some overlap in required courses. Specifically, courses in Linear Algebra, Statistics, and Abstract Algebra. For the purpose of this reflection, I will be going through three projects I have completed over this period. Two of these projects are from the course titled INTR 202 with Professor Hu and the other project is from BUS 315 with Professor Keri Larson.

Project 1

The first project I plan to include in my data science portfolio is a midterm project I completed in INTR 202. This project was done in pairs, and I worked with fellow student Hyatt Sbar. The main teaching goal of this project was to get a working proficiency with the coding language R, as this was the first big project of the class. For this assignment we created a survey for other W&L students to complete that asked about certain demographics and familiarity with drinking games. For the survey we used a seven point Likert scale to determine each participants familiarity with seven different drinking games and then asked each student to rank each drinking game by favorite to least favorite, and most played to least played. We had about 41 usable responses to the survey with made up our dataset. This data was collected anonymously over Google's survey platform, no names or identifiable information was collected. The findings of this projected showed that beer pong was the most popular game by a substantial margin, with a confidence interval of 70-90%. Beer ball had the most variety of answers when addressing familiarity. Reflecting on this project, there were definitely some limitations with the survey, the ratio of male to female participants is almost 2:1, and only 41 people took the survey. Thus, these findings are not representative of the student body as a whole. Otherwise, the skills practiced and learned for this project are statistics in R, working with ggplot, understanding confidence interval, and learning how to produce a useable survey. Overall, this project was a success and taught me a working knowledge of R that I still use today.

Project 2

The second project I plan to include in my portfolio is also from INTR 202 and was also completed with Hyatt Sbar. In this project we looked at the relationship between win rate in boxing matches depending on stance and knock out rate. We collected our data from Kaggle.com which has a real possibility of having large amounts of bias, as this data includes amateur boxers in the 2,760 observations. However, this was the best data we could find at the time. In our analysis we looked at the major statistical metrics like mean, standard deviation, and range. We conducted an ANOVA test in R, of wins and KO rate by stance. Finding that there is a significant difference in mean of wins by stance and mean of KO rate by stance. Therefore, we looked at the multiple regression of win rate depending on stance and KO rate. Finding that southpaw stance and KO rate had very small p values while orthodox stance and KO rate has a much higher p value significantly higher than .05. Altogether we found that southpaw stance has a positive correlation to win percentage. For this project I learned how to mutate data frames, practiced visualizing different statistics with ggplot, leaned how to conduct an ANOVA test and a multiple regression to show correlation.

Project 3

The third project I plan to include in my data science portfolio is taken from the BUS 315 course with Keri Larson. For this project I worked with 2 other classmates to create an SQL database that could be implemented for a pharmacy type store like CVS or Walgreens. This project was completed using SQL workbench, which we used to create a graph symbolizing categories in different tables and each tables relationship with other tables in the data frame. This project was more conceptual than I expected as we had to take be careful in which variables were used as primary keys or the unique identifier across the set. There was also the task of making sure each variable was stored correctly so that when queried it would output in the right format and not change over time. For example, making sure age is stored as date of birth. To help with this aspect we created a data dictionary that breaks down key variables describes their purpose and defines the way in which it is stored and why. The project was a success, in which I learned how to create a SQL database, understand how to store certain variables correctly, query using the correct relationship, and an overall proficiency in writing SQL code.

Overall, throughout my journey in the Data Science Minor I started with very little knowledge of R and working with data in such a platform. Through projects in INTR 202, mentioned above, I honed in on working with large datasets, manipulating the data and visually representing data in way in which a layperson might understand. With the project from BUS 315 about creating a database, I learned how to represent large amounts of data in a way that can be organized to a diverse audience and clearly applicable in a real-world scenario. The skills I learned in these projects have been crucial to my development in other classes, for example working with python in CSCI 111 and its similarity to R. This skill also carried over to the class I took on statistics which worked with R in a way that I wasn't familiar with. In statistics I wrote a lot of functions and used formulas to calculate odds in blackjack rather than working with large datasets to interpret results. In my current class CBSC 309, almost all the work is done in R using large datasets and looking through visualizations and factor analysis to create simple but accurate linear regressions. This CBSC class operates at a much higher level of R knowledge, and I owe most of my success to the basic skills I learned in INTR 202 and the boxing project and student drinking survey project I completed in that class.