Chaeyon Jang

DS-40-01

Sybil Prince-Nelson

Data Science Portfolio Reflection Essay

Classes I took for the Data Science Business Analytics minor are: Accounting 100 (ACC 100), Statistics (INTR 202), Responsible AI (BUS 195A), Fundamentals of Programming I and II (CS 111 and 112), Business Analytics (BUS 316), and Accounting Information Systems (ACC 310). The main skills and insights I gained from these classes were critical thinking, problem-solving, data restricting and organizing, data interpretation, and qualitative evaluation of the project results and implications. I think my BUS 316 and ACC 310 helped develop my data science coding and technical skills that would be relevant in data organization and restriction. INTR 202 helped me gain a high-level understanding of statistical concepts to further explain the meaning behind different results through both visual and quantitative representations. CS 111 and 112 helped nurture my problem-solving skills by teaching me to be creative and flexible in how to answer queries in different ways. BUS 195A was by far the most qualitative class that developed my understanding of the ethical concerns of AI and the future of technological innovation. I think this class made me very cognizant of the ethical considerations of data privacy and collection and how to analyze confidential data responsibly. Furthermore, understanding the feasibility of certain projects and if they can be reproduced is an important lesson I learned through the portfolio class. I never really considered the reproducibility of projects especially because they are given under the framework of it being used once for a class project. The portfolio class put in perspective how these projects can be used and remade to create new projects that can create discoveries. Throughout the wide variety of classes in the minor, I gained technical skills that can be useful in the future while also gaining unique and deep insights on the impact data science and technology have in our lives.

Originally, I intended to be a Computer Science minor but found that beyond the fundamental classes, the classes after were not as applicable to what I wanted to take away and apply to my career in finance. I was taking a stats course at the time, and my professor introduced the topic of data science and explained what it was. A lot of my peers at the time were interested in data science, which supported my interest and ultimately led to me declaring the minor. Initially, I did not have any data science skills; I took the fundamentals of programming and the interdisciplinary statistics course. I had a basic understanding of what I was supposed to do but had no exposure to the objectives and overall data science field.

Projects that are included in my portfolio are BUS 316, Business Analytics, and ACC 310, Accounting Information Systems. BUS 316: Business Analytics - EV Charging Analysis, BUS 316: Business Analytics - Business Analytics with R, and ACC 310: Accounting Information Systems - Accounting Analytics are the three projects I would like to discuss in detail. These projects were chosen because they represent a culmination of the large technical growth I experienced while in the minor.

The Business Analytics with R project aims to analyze European Union product sales data using R programming. The goal is to filter relevant data, conduct exploratory data analysis, and visualize total sales by month. The objectives of the project were learning how to use both SQL and R on a single system and effectively restrict, organize, and read data.

The EV Charging Analysis aims to understand the data sets and what each prompt means in aggregate and experiment with a new language (R). This project explores an electric vehicle (EV) dataset to analyze patterns and trends.

The Accounting Analytics project focuses on payroll and employee wage analysis using SQL queries. The goal is to extract, clean, and analyze payroll data for different job roles and periods. A large focal point of the project was to combine two very unorganized data sets and clean them up, join them into new data sets, and fulfill various prompts to effectively understand the company's operations to give the manager suggestions and recommendations on how to improve the data system of the particular company.

Across all the projects, notable skills and insights gained were critical thinking skills, data exploration, and organization, learning the importance of easy code interpretation, understanding the outputted results and visualizations, and considering the ethical implications of each project and what could've been improved.

Regarding critical thinking, I learned new problem-solving skills that helped create the correct queries that will output the correct answers. Data exploration was a new skill I learned that helped me understand the data set to be able to reference the correct tables and parameters. One of the most important skills I learned throughout the projects is data organization. Cleaning up data is a key skill that I gained; identifying errors by finding incorrect, corrupted, or duplicate data, correcting errors by replacing, modifying, or deleting the affected data, filling in missing values by adding missing data to the dataset, standardizing data to make sure data is formatted consistently, and validating data by confirming that the data is accurate and consistent are all important steps needed to ensure that the coding itself will be easier and more understandable to other users. Something that follows in a similar vein is learning the importance of easy code interpretation. Implementing easy-to-understand code helps other users and myself (the creator) identify errors and makes it easier for other users to edit. Understanding output results and visualizations is crucial to be able to take the main points from the dataset to come to conclusions. Another key insight and skill that I have learned over my time is considering different ethical implications of the projects and what could've been. It is necessary to always contemplate whether or not the data has been compliant with regulation and privacy laws. By continually weighing how future projects and data gathering can be improved, the quality and trustworthiness of the data will be higher.

There are many strengths of the EV Charging Analysis project. The data directly experiments with real data, which gives students examples of what future data science projects will be like. We used various parameters to understand different aspects of the data set, helping us learn skills that would come to aid us when trying to understand what kind of data we are working with. Comprehensive data exploration where the project utilizes exploratory R functions like str() and head() to provide a structured understanding of the dataset before analysis. The incorporation of R packages like tidyverse, ggplot2, and fosdata ensures efficient data manipulation and visualization. The project outlines a clear analytical approach where the project follows a logical workflow, starting with loading and structuring the data before conducting analysis. Inline Code for Documentation is prevalent, where the project includes inline R code to ensure transparency in reporting key statistics like row and column counts. The project applies appropriate filtering and summarization techniques to derive meaningful insights from the dataset.

Some limitations include a small data pool, with only data coming from one location/ company, and limited discussion on findings, where the project could provide a more detailed interpretation of key findings and their significance. Potential data cleaning oversights are evident, where the project does not explicitly mention steps taken to handle missing or inconsistent data, which could impact the accuracy of results. There is a lack of comparative analysis where the project does not compare findings across different datasets or provide insights into trends over time. There is a lack of consideration of external factors. The analysis appears to be solely based on the dataset without integrating external business or economic factors that could influence results. Finally, there is limited error handling. The project does not explicitly outline error-handling mechanisms for common issues like NA values, incorrect data types, or outliers.

In regards to the Business Analytics with R project, some strengths are the combined use of R and SQL on R Studio Workbench, the utilization of real-world data (e.g., European sales and university salary data) to develop analytical skills, and the demonstration of proficiency in R programming, data cleaning, and visualization. Furthermore, the project provides clear documentation and references for functions and commands used. The project also addresses multiple aspects of data processing, including filtering, summarizing, and graphing data. The encouragement of problem-solving through debugging and troubleshooting errors is another strength added to the project.

Limitations to the Business Analytics with R project are that invalid data within the dataset may create misrepresented visualizations and limited discussion of insights derived from data analysis beyond the technical implementation. The focus is primarily on execution rather than interpretation, which might reduce the business impact of the findings.

Accounting Analysis has many strengths as well as limitations. The accounting analysis project is a well-organized query execution with structured commands. The project ensures clarity in data presentation by displaying relevant tables and breakdowns. There is a strong data-driven approach – the project effectively utilizes SQL and R to analyze salary data, ensuring an evidence-based decision-making process. Comprehensive data exploration is a strength of the project. It includes exploratory data analysis using various R functions, ensuring a thorough understanding of salary distribution. Another strength of the accounting analytics project is its clear methodology – The use of structured queries, filtering, and data transformations provides a replicable and transparent analytical framework. A unique strength of this project that isn’t present in the previous two is that the project prompts practical business insights where the project aims to identify patterns in wages, department trends, and potential discrepancies, providing useful insights for financial decision-making.

The accounting analytics project was very helpful in learning many interpretive skills, but it lacks a deeper analysis of payroll trends, patterns, or business implications (the more numbers-based results) compared to the other two projects. The project primarily presents queries rather than interpreting the broader impact of payroll structures on organizational efficiency. I think this project could also benefit from additional visualization techniques for better insights into simple tables and outputs. In addition, there are some data scope restrictions where the analysis is limited to a specific time frame, which might not fully capture long-term trends in salary distributions. Potential data quality issues are also a concern. The project relies on the accuracy of salary datasets, which could contain missing or inconsistent entries that affect conclusions. Another concern is that the project integrates external data sources (such as EU country lists), which could introduce inconsistencies if not properly maintained.

If Computer Science is more about solving and creating algorithms, data science, to me, is more useful in the sense that it helps users understand different types of data, clean and organize large datasets, and interpret them to answer questions specific to the data. Throughout my data science classes, I learned a lot about what to look for when determining if a data set is organized, what I should prioritize when the data is not cleaned up, and how to utilize different languages to restrict, reframe, and analyze the given data. I think that as an accounting major, my skills as a data science minor have allowed me to be able to look at different values that are used to interpret different companies and its financial health, for example. This minor tested how to communicate results for diverse datasets and audiences so that the audience can take away the main points and leave with an important lesson. Going forward, the main takeaways that I would like to apply to my career in finance are the lessons of data cleaning, the importance of readability in both code and presentation of visual outputs, and interpretation of these results to create a substantial argument or claim. The minor overall has been very helpful in developing interpretation skills, data organization skills, and critical thinking to help benefit my future.