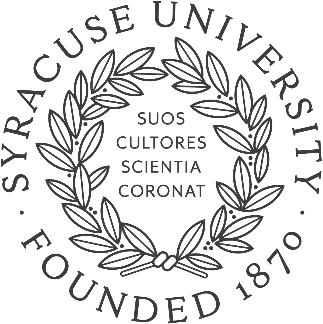
Kobi Wiseman

Syracuse University | The school of information studies

netid: yvaisman | SUID: 543673285

Portfolio Milestone

applied data science program, APRIL 2021



**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **1** | **Introduction** | **2** |
| **2** | **The Program’s Learning Goals** | **2** |
| 2.1 | Description of the Major Practice Areas in Data Science | 5 |
| 2.2 | Collecting and Organizing Data | 5 |
| 2.3 | Identifying Patterns in Data via Visualization, Statistical Analysis, and Data Mining | 5 |
| 2.4 | Developing Alternative Strategies Based on the Data | 6 |
| 2.5 | Developing a Plan of Action to Implement the Business Decisions Derived from the Analysis | 6 |
| 2.6 | Demonstrating Communication Skills Regarding Data and Its Analysis for Managers, IT Professionals, Programmers, Statisticians, And Other Relevant Professionals in Their Organization | 7 |
| 2.7 | Synthesizing the Ethical Dimensions of Data Science Practice | 7 |
| **3** | **Conclusion** | **8** |
| **4** | **Appendix** | **9** |

# **Introduction**

### This paper aims to encapsulate my overall learning achievements which have been acquired during the Master of Science in Applied Data Science program I have been participating between August 2019 and May 2021. For each learning goal of the program, this document defines the goal, then referring one of the classes’ group projects I completed, as an example of how this goal has been accomplished during that project.

# **The** **Program’s Learning Goals**

### Defining the level of the program’s success throughout the curriculum, seven learning goals have been defined to outline the extent of which each student has successfully acquire both theoretical knowledge and practical tools for being a ‘ready-on’ data scientist in the job market.

### The program’s learning goals have been outlined as: (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 analysis; (6) Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization; and (7) Synthesize the ethical dimensions of data science practice. This portfolio will go over each one of these goals, define it and provide a case study example of how this goal has been implemented in a data science project. Two exceptions will be the first and the last goals, which will be discussed in a higher level.

### For referring examples to these goals, four of the classes which I have obtained will be used, as described in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class Code** | **Class Title** | **Project’s Data Description** | **Individual / Group Project** | **Program Used** |
| IST 687 | Introduction to Data Science | Airlines Customer Experience:  The data, provided by the instructors, included information about airlines passengers’ customer experience. The features were consisted of both a passenger’s details (age, gender, and airline membership information) and that passenger’s related flight details (origin, destination, delay in departure, delay at arrival). In addition, each customer’s Net Promoter Score was included as a target variable. The goal for these two projects is to predict what factors shall increase the passengers’ tendency to become more ‘promoter’ and less ‘neutral’ and ‘detractors’ (these statuses are a function of the Net Promoter Score). | Group | R |
| IST 707 | Data Analytics | Group | Orange |
| IST 652 | Scripting for Data Analysis | Classic Video Games Sales:  This data, which was chosen by me, focused on revenues of classic video games. The features in the data included information about each game itself (game type, genre, and distribution year) and its revenues, divided by areas in the world (US, Japan, Europe, and others). In this project I conducted both supervised model for finding the most influential factors for determining a game’s revenues, and unsupervised model to identify different patterns among those video games. | Individual | Python |
| IST 718 | Big Data Analytics | Used Vehicles Buying:  Chosen by our team, this data included information about listings of used vehicles across the United States. The data file included many features regarding the vehicles’ characteristics (manufacturer, car type, engine, fuel consumption, gear type etc.) and the listing itself (posting date, number of previous owners and offered price). Using various different models, we were able to predict listing price and number of days to remain on the market with supervised models, as well as identifying patterns within the listings using unsupervised models. | Group | PySpark |

## Description of the Major Practice Areas in Data Science

### As data rapidly become cheaper and more available to collect and store, having the data a key factor in nearly every organization within every possible industry was simply a matter of time. The desire of obtaining the data, maintaining its data integrity and ready to be used for business analysis occurs very often as result of intense competition, but also could be a result of universal humanity needs, such as the immediate necessity of COVID19 tests and vaccines.

### Reviewing my entire courses retrospectively in terms of homework assignments and final projects altogether, I came across with so many various business topics, from orange juice sales, throughout airlines customer experience, evaluating diamonds and even detecting sarcasm among different documents. Having this tremendous selection of topics, including the examination of the Titanic passengers dataset which been conducted in multiple classes, is just the tip of the iceberg (not a random choice of words) when it comes to how various could practice areas in data science get.

### 

## Collecting and Organizing Data

### Every cutting-edge business insight relies on data. Furthermore, the data must be reliable in its accuracy so we could take these insights seriously and bring its recommendations into actions. The overall process of maintaining a clean and accurate data is also known as ‘Data Integrity’. Therefore, there is a massive importance for not only importing in the data, but also carefully read and examine it before moving forward to statistical analysis and model developing. The most common examples of this process would be detecting and removing duplicate instances (identical rows in the data), missing values (and replacing them with an alternative value) and outliers.

### Working on the airlines’ customer experience data during the Introduction to Data Science final project, our team came across some data integrity issues we had to address prior to proceeding with EDA and modeling. While cleaning the data, we noticed that some of the attributes contain a few missing values. After carefully viewing the distribution of each one of these variables, we decided as team to replace these missing values with each attribute’s median value. By doing so, we were able to keep those instances in the data without causing any drastic changes in the distribution.

### 

## Identifying Patterns in Data via Visualization, Statistical Analysis, and Data Mining

### Once our data is clean and reliable, conducting EDA (Exploratory Data Analysis) is one essential step needed prior to model selection and execution. Viewing the distribution (along with other pre-modeling features such as correlations between variables) of each attribute in the data allows us to examine the ‘behavior’ of the data and possibly re-scope our modeled population and our definition of target variable (if exists). EDA also bolsters many of the model’s interpretations.

### A good application of identifying data and confirm its conclusions in the model output, occurred while working on my final project of Scripting for Data Analysis class. Outputting the descriptive statistics, I have identified two main market segments – one in the United States and the other in Japan. This finding was confirmed later while I conducted k-means clustering model and received two separate clusters which one of them is mainly characterized by American customers which were interested in one specific game genre, and another cluster of Japanese gamers who were into other genre of video games.

### 

## Developing Alternative Strategies Based on the Data

### Oftentimes, our original plan of modeling might be shifted into different directions once we examine specific patterns during the EDA phase. Whether it is the focus of the data population, different algorithm or altered features selection, a good data scientist shall be dynamic and adjust his/her game plan when required.

### Working on our group project for Big Data Analytics class, we have encountered the necessity of having a major change needed to be performed prior to the modeling stage. Viewing the descriptive statistics, we noticed that some of the listed vehicles are priced very highly, as they were considered luxury vehicles, production of high-end firms such as Lamborghini and Maserati. This finding forced us to re-think about what should be our focus for this model, since including all these vehicles might decrease model’s prediction power. Ultimately, we have decided to omit those instances out of the data to predict more common, average-priced listed vehicles.

### 

## Developing a Plan of Action to Implement the Business Decisions Derived from the Analysis

### Plan of action is the cherry on the top of the data mining’s whipped cream. Unlike standard ad-hoc reports which mostly do not require more than straight-forward descriptive statistics, data science applications allow interpreting the model and convey its output into an intuitive business language. Indeed, most of the executive have great management and business capabilities, as well as tremendous knowledge about their own business or domain. However, many of them lack the background of statistics and machine learning. Therefore, transforming the model results into ‘everyday’ jargon, including forming some of these results into action items – is essential for modeling project to comply with business goals and processes.

### Data Analytics final project uses a great example of modeling and producing action items from the models’ results. Using both supervised and unsupervised learning models, our team was able to provide action items which should assist ‘detracting’ and ‘neutral’ customers to become ‘promoters’, while making ‘promoter’ customers to maintain this level of satisfactory. These actions items reflected as straight-forward bullets, such as: ‘customers with ‘silver’ or ‘gold’ status should be rewarded with some loyalty program. By doing that, having these customers enjoy different perks while their flight is being delayed, will significantly reduce their dissatisfactory rate’.

### 

## Demonstrating Communication Skills Regarding Data and Its Analysis for Managers, IT Professionals, Programmers, Statisticians, And Other Relevant Professionals in Their Organization

### Managing complexed data tasks under tremendously competitive business environment, being a good data scientist requires much more than just writing a decent code. The ability to convey the data scientist’s model outputs to other colleagues in similar areas emphasized the data scientist’s level of understanding of both theoretical and practical matters. Although mastering of writing ‘for’ loops is always a good skill to have, an excellent data scientist knows not only ‘how’, but also ‘*why*’.

### My individual final project of Scripting for Data Analysis class included a requirement of a video presentation, which perfectly demonstrates my capabilities of communicating my entire data process, all the way from import the data thru interpreting the results and making business conclusions.

### *A link this video will be available along with the entire content included in my GitHub sub-folders at the final submission phase.*

### 

## Synthesizing the Ethical Dimensions of Data Science Practice

### Data represents lots of knowledge, and as the old say goes – ‘knowledge is power’. With this power comes lots of responsibility, as this information might go the wrong hands. In many cases, data files include some individuals (and businesses) sensitive information which might be used for negative purposes, such as targeted advertisement, phishing and even blackmailing. Therefore, all the possible precaution measures to make sure that there will be no privacy data leakage outside the organization or any data project.

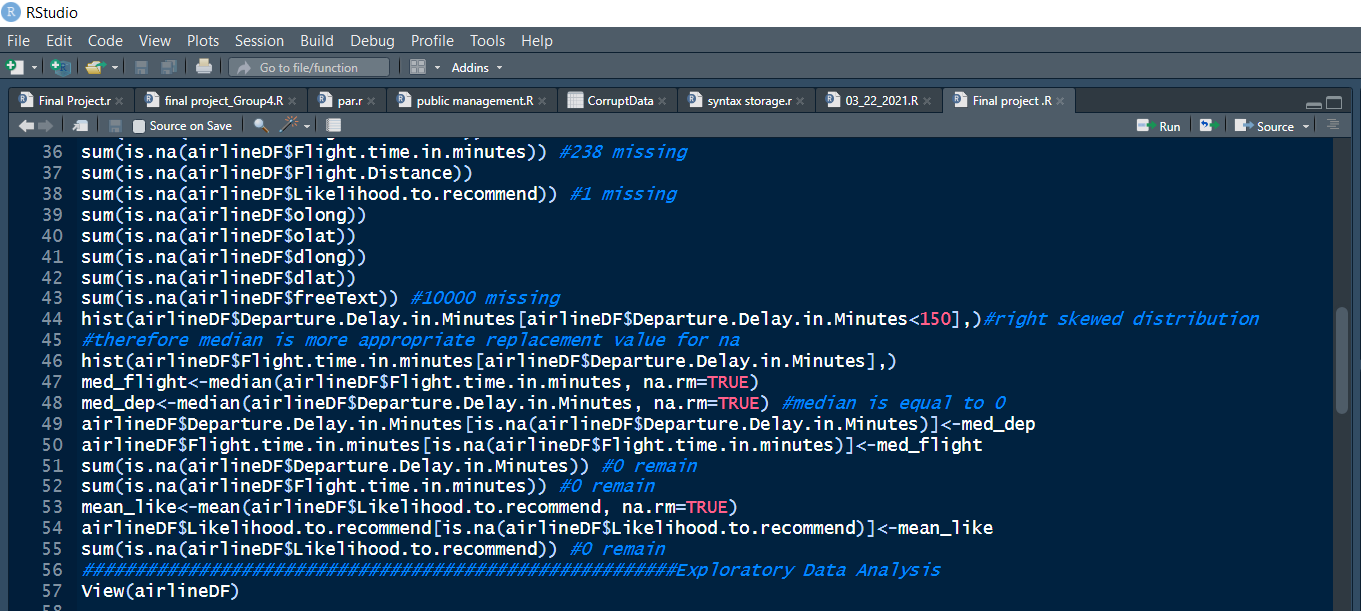
### Throughout my entire final projects, I could not recall any project which included a drastic data decision that addresses a privacy issue. The closest example I can think of would be dealing with the airline data for both Introduction to Data Science and Data Analytics classes, where the attributes of individuals’ personal information (first and last name) were omitted from the analysis for the sake of this privacy issue.

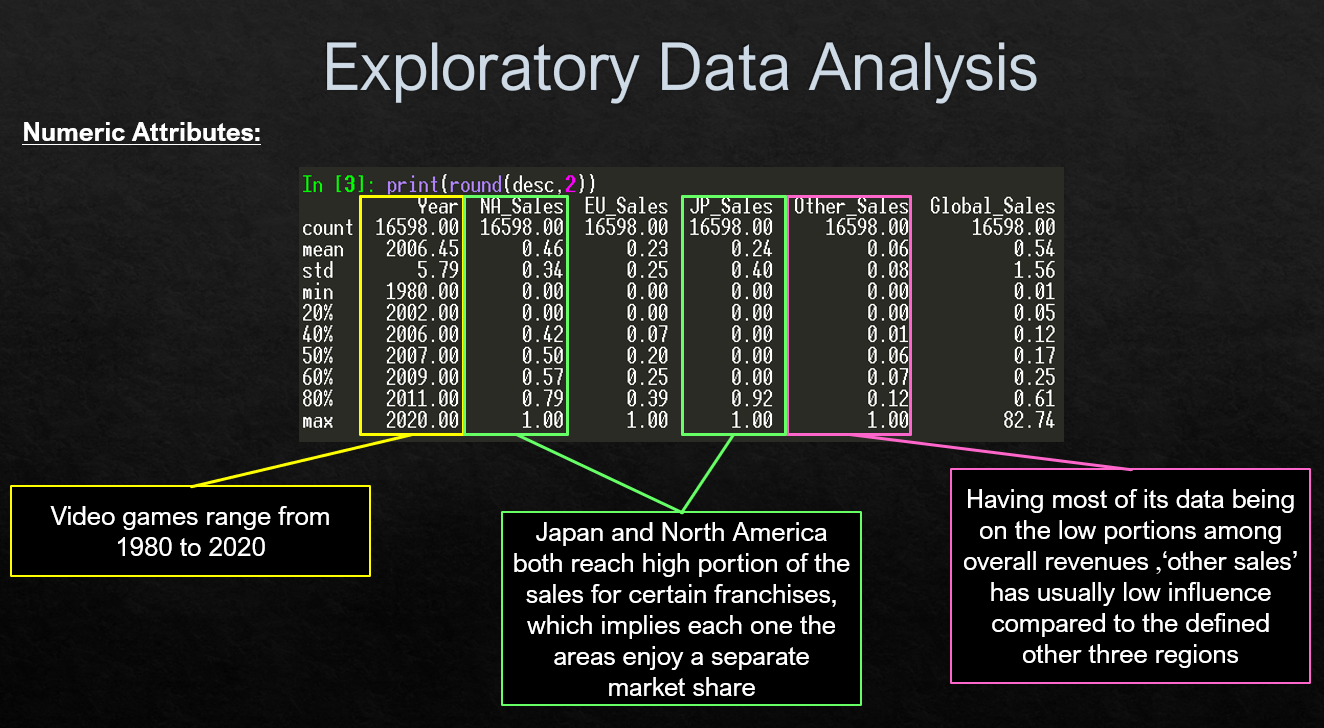
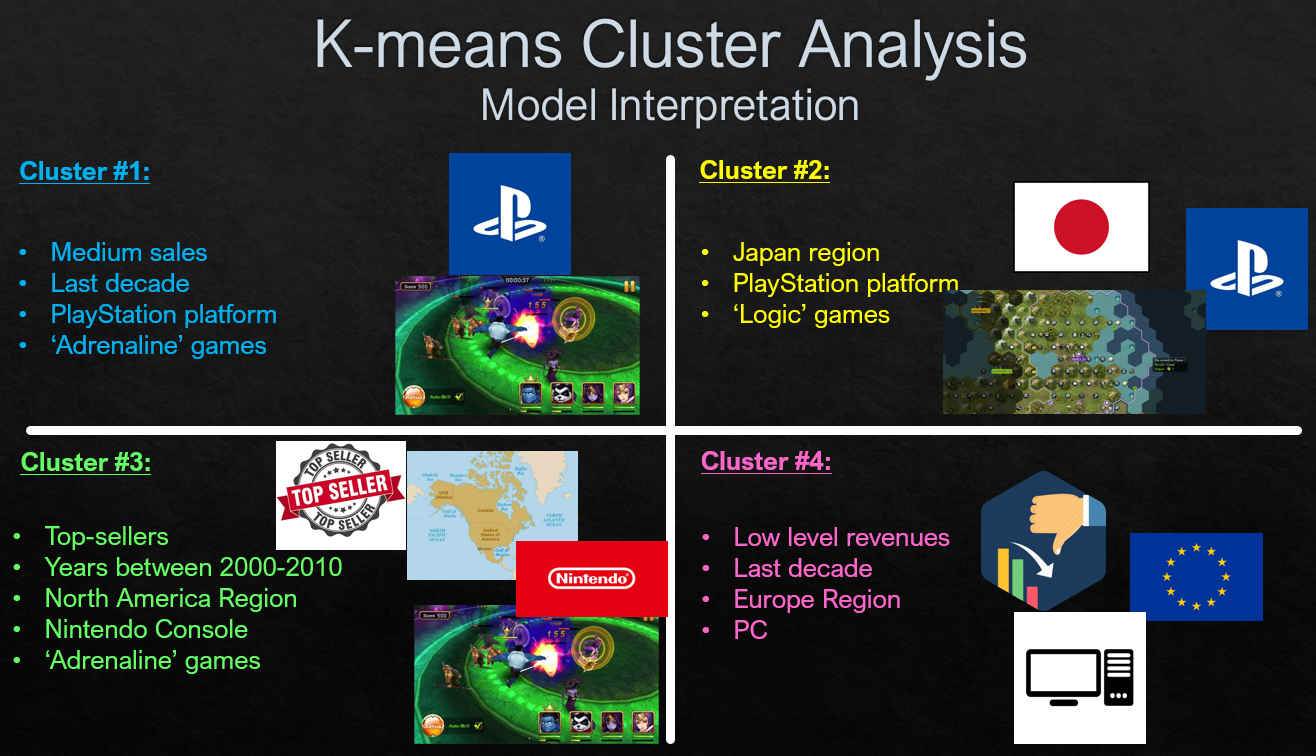
# **Conclusion**

### Reflecting this program as a whole, I would describe this program as an ensemble of ‘Ah-ha’ moments. During my recent role as a data scientist in a credit card firm in Israel, I used to conduct many data operations without truly know why those things are done. As the curriculum goes on, I keep producing many ‘clicks’ which reflect my realization of the rationale behind many data decisions and actions.

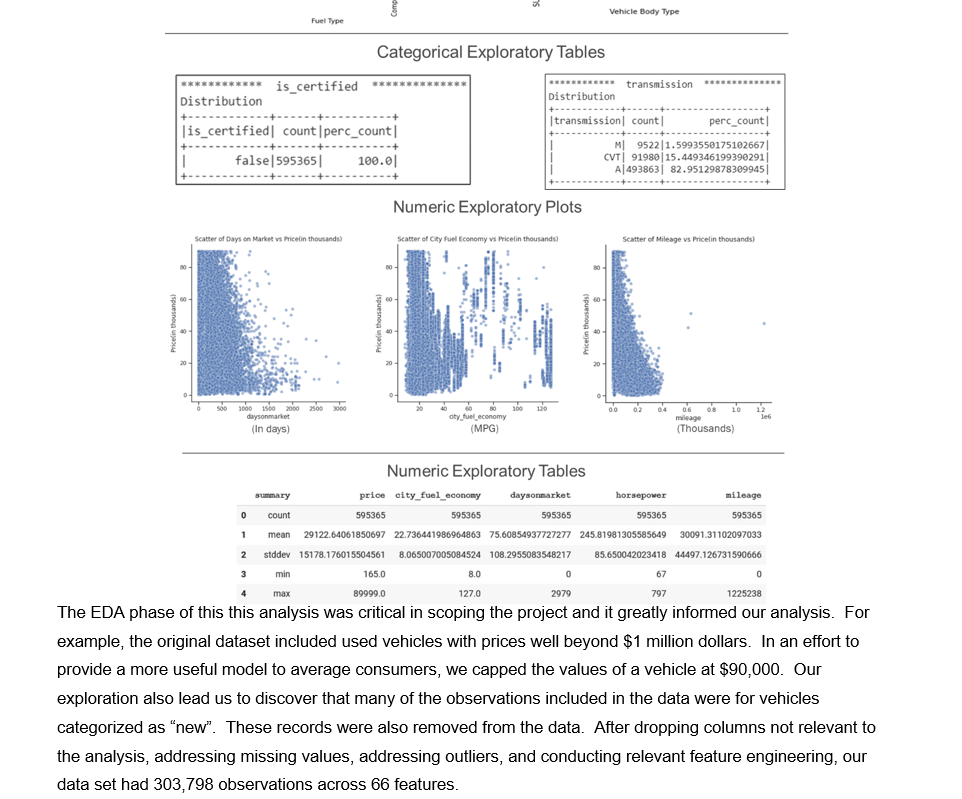
### Finalizing this portfolio, I would like to use this opportunity to thank to the entire program’s faculty members, instructors and any person who contributed by forming content into this curriculum. Thanks to this program, I am now able to not only perform tasks with technical mastery, but also deeply understanding the reasoning of each part of the data process.

# **Appendix**

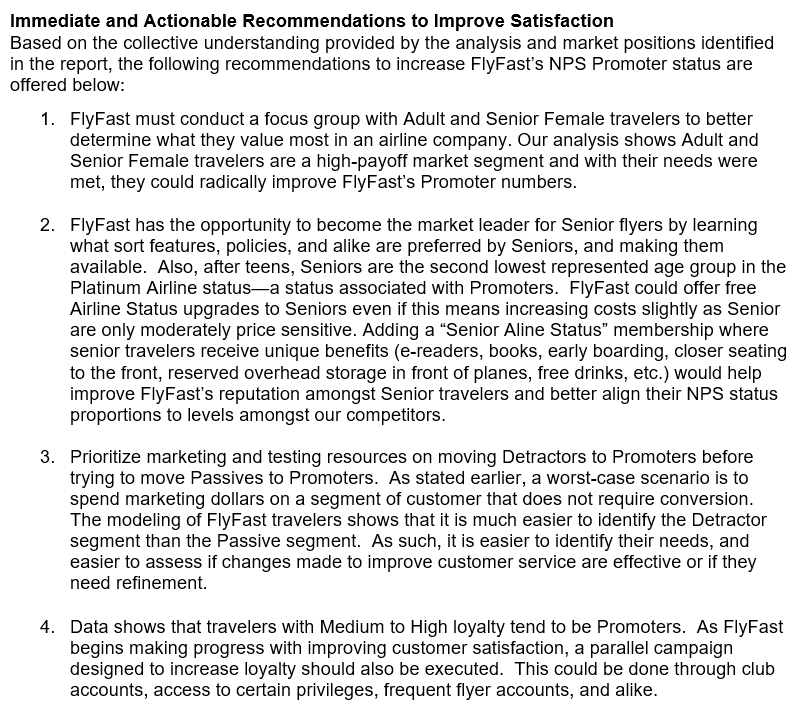
Section 2.2: taken from IST687 final project's code; while collecting and organizing data, detecting missing values within the data and replacing it with each attribute's median.



Section 2.3: taken from IST 652 final project's presentation: detecting patterns using exploratory data analysis, reinforced by further model interpretation.



Section 2.4: taken from IST 718 final project's paper; Strategy of what data to be used for modeling has been shifted thanks to early exploratory data analysis.

Section 2.5: taken from IST 707 final project's paper; A plan of action is outlined for the business decision makers, based on findings from model interpretations.

<https://youtu.be/hu9kg7d4ya0>

Section 2.6: link to my IST 652 final projects video presentation; demonstrating my skills of communicating data and conveying it to business matters.