**Data Science Program Final Project**

**Executive Summary**

At the end of the Data Science program, students are required to complete a final project of their choice. They are given six weeks to work on the project. Oftentimes, they will be paired up with another fellow-student.

This document is dedicated to Chamyia and Joshua. We intend to answer 2 questions: What video game has the most play time and how many games does a user typically have?. We will present our answer to these questions as well as any interesting findings that may present themselves through data analysis in a powerpoint. Our main program of use will be python. We have chosen a data set on video games from Kaggle. Using the tools learned during the course we will do data wrangling, analysis and a presentation to present our findings. We will work together over zoom and text during the 6 weeks. Outside of the normal risks with technology the risks are minimal.

**Business Objectives**

To showcase the skills that we have acquired through the Data Science program. They will be using R, Python, Tableau and other programs to wrangle, analyze, and visualize the “Steam Video Game” dataset made available by Tamber on Kaggle.

At the end of the project, we will be able to explain our work in layman’s term, and present our findings to the students, faculty, staff, and potential employers, along with other interested parties via Zoom.

**Background**

As a way to activate and put practical use to what the students have learned, doing a final project is a good way to demonstrate that.

We have chosen the “Steam Video Game” dataset because we are both interested in video games, and steam. They hope to glean insight from this document to discover: What video game has the most play time and how many games does a user typically have?

**Scope**

We will be using the software taught in the program to complete the project. They will be intentional on using tools of their interest or tools that may aid finding a job. They may choose to use additional software/tools, but that is not required.

**Functional requirements**

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Columns and unusable columns should be removed. As the dataset is fairly large, we should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: We will familiarize ourselves with the dataset. They should have a good understanding of what each column means, and how the values are measured. They will brainstorm on questions to ask, and what they might gather from the dataset. Then, they will identify the proper functions to create models, predictions, etc.

Data Visualization: Once we have a comprehensive understanding of and insight gathered from the dataset, they will work on visualizing the findings. They may decide to use Tableau or other graphing programs, and compile the visuals and texts in a Power Point slideshow.

Presentation: Working with school leaders, we will schedule a time to present their findings via Zoom. They should be able to communicate in a clear and easy-to-understand manner. The presentation should be kept around 20 minutes. They should be dressed professionally for this occasion.

**Personnel requirements**

We are the two developers. They will need to work closely for this project to succeed. They will touch base once a day via Zoom or Slack to problem-solve or to check in on work progresses. Once a week, they will review the past week workload and plan out the next week. They will take turns being the scrum master, and report their progress to their instructor (Product Owner.)

Once a week, they will meet with their instructor. They should be prepared to ask questions and seek guidance for the next steps.

They may also consult with their coding mentor.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate ourselves on Steam Game user data. Set up Github.

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to see if there is any interesting findings.

Week 3: Modeling through tableau to visualize the data

Week 4: Review and validate findings from the previous week, and draw insights/conclusions.

Week 5: Compile findings into a Power Point slideshow. Go over it with their instructor and friend/family member to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is delightful on the eyes.

Week 6: Make final touches to the Power Point presentation. We should not attempt to come up with a brand-new analysis. There will not be enough time to verify their findings. They should practice presenting at least a couple times with the two of them, and at least once with their instructor.

**Other requirements**

All programs and data used should be free of charge.

**Assumptions**

The software programs and platforms we will use are readily available, up-to-date, and not broken.

**Limitations**

Possible limitations include scheduling amongst team members and the instructor/mentor. Another limitation may include roadblocks in our work due to unforeseen errors while programming reducing time spent on completing important tasks.

**Risks**

The risks that may arise are internet outages, power outages, family emergencies or broken software/hardware. Another risk is that Joshua is not from the program so there could be a lack of interest later in the project. However both Joshua and Chamyia intend to use this project as something to send to future employers so that risk solves itself because of Joshua’s need for a better paying data science job. The risk of this project being incomplete are unlikely aside from family emergencies.