**Data Science Program Final Project**

**Executive Summary**

To outline the steps taken to complete analysis on movie finance and rating data to determine correlation with rating, genre, and movie profits.

**Business Objectives**

To determine whether movie genre or budget correlate with the user rating or boxoffice success of the movie.

**Background**

I am a huge movie fan. Spent years in front of the old tube tv with my mom watching old movies and staying up late past bedtime with my sister watching Poltergeist and cheesy b-movies. I’m always a little surprised by my friends choices in favorite movies and wanted to see what movies could be considered the popular of all time (user ratings) versus which ones actually were the most successful in theaters.

**Scope**

To highlight the skills that we have learned throughout the Data Science program. We will be using R, Python, Tableau, MS Suite, and other programs to wrangle, analyze, and visualize the movie datasets made available by InformationIsBeautiful.net and the Movie Dataset from Kaggle.

**Functional requirements**

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Nulls and unusable columns should be removed. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: The team will familiarize themselves 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 the team has 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: Collaborating with school leaders, the Team 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**

The developer is Ian Franklin. I will need to work diligently for this project to succeed. I will be in contact once a week via Zoom or Slack to problem-solve or to check in on work progresses. Once a week, I will review the past week workload and plan out the next week.

Once a week, I will meet with my mentor and instructors. I will be prepared to ask questions and seek guidance for the next steps.

I 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 cancer and risk factors. 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 are any interesting findings.

Week 3: Continue to analyze and complete any additional wrangling that may arise.

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. We should practice presenting at least a couple times.

**Other requirements**

All programs used should be free of charge.

**Assumptions**

The software programs and platforms we use should be available, up-to-date, and not broken.

**Limitations**

If something should come up for during this six-week period, the project may be delayed. If the instructor or mentor have scheduled or unscheduled time-off, the project may be delayed as well. The team may experience a roadblock in their work, which may push back the completion date.

**Risks**

The risks that may arise are such like natural disasters, power outages, family emergencies or broken software/hardware. The team is eager to complete the program so there should be no motivation issues.