Predicting Our Next Successful Game

WGU C964 Task 2

James Clair

#000847594

2/1/2021

Table of Contents

[Letter of Transmittal 2](#_Toc104715413)

[Project Proposal 2](#_Toc104715414)

[A. Problem Summary 2](#_Toc104715415)

[A.1. Organizational Need 3](#_Toc104715416)

[A.2. Context and Background 3](#_Toc104715417)

[A.3. Outside Works Review 3](#_Toc104715418)

[A.4. Solution Summary 4](#_Toc104715419)

[A.5. Machine Learning Benefits 5](#_Toc104715420)

[B. Machine Learning Project Outline 5](#_Toc104715421)

[B.1. Scope 5](#_Toc104715422)

[B.2. Goals, Objectives, and Deliverables 5](#_Toc104715423)

[B.3. Standard Methodology 6](#_Toc104715424)

[B.4. Projected Timeline 7](#_Toc104715425)

[B.5. Funding Requirements 8](#_Toc104715426)

[B.6. Evaluation Criteria 8](#_Toc104715427)

[C. Machine Learning Solution Design 8](#_Toc104715428)

[C.1. Hypothesis 8](#_Toc104715429)

[C.2. Selected Algorithm 8](#_Toc104715430)

[C.3. Tools and Environment 9](#_Toc104715431)

[C.4. Performance Measurement 9](#_Toc104715432)

[D. Description of Data Sets 10](#_Toc104715433)

[D.1. Data Source 10](#_Toc104715434)

[D.2. Data Collection Method 10](#_Toc104715435)

[D.3. Quality and Completeness of Data 11](#_Toc104715436)

[D.4. Precautions for Sensitive Data 11](#_Toc104715437)

[References 11](#_Toc104715438)

[Executive Letter 12](#_Toc104715439)

[Business Vision 12](#_Toc104715440)

[Quickstart Guide 12](#_Toc104715441)

# Letter of Transmittal

WCVG is at a turning point, we recently completed Pumpkin Slaughter 9000 and are proud of its recent success, but it won't last forever. Pumpkin Slaughter 9000 successes will sooner or later stop paying out and we have a company that's sole purpose is to create successful video games, hence the name. To ensure we continue to succeed as a company, we have a choice to make, what do we create next? And how do we confirm we are making the right choice?

The WCVG data science team believes the answer to both of those questions is data. First, we can perform a statistical analysis of video game industry sales data to help us inform our decision. Assuming the data is accurate and fits our model, we can then use a machine learning model to make highly accurate predictions on which type of game would result in the highest global sales. Or in other words, the type of game that is most likely to bring us success if we were to develop it.

In the event the data doesn’t fit our model, we should still now have tons more industry sales knowledge to help inform our decision, for example: “What type of games performed the best over X number of years?” and “What platforms have performed the best?”. Making our project a win-win scenario and leading to valuable insights regardless of the results.

A project proposal has been compiled below this letter and includes the following details and resources:

* Summary
* Outline
* Solution Design
* Description of Data

I do hope that you find this project helpful to the success of WCVG and look forward to working together. Please feel free to contact me at any time regarding this project.

Thank You,

James Clair

WCVG Data Scientist

# Project Proposal

### A. Problem Summary

We Create Video Games (WCVG) will soon be finishing and releasing our latest video game, Pumpkin Slaughter 9000. With PS9000’s release fast approaching, WCVG would like to start looking to the future and determining what our next big game will be. This proposal will describe a strategy for deciding what game we want to develop while maximizing the game’s potential success.

#### A.1. Organizational Need

With our recent successes, one of the biggest problems that WCVG is faced with is deciding what next to do with our time and development resources. We want to pick a game that will maximize our profits, expand our brand awareness, and crush the top charts. There is a lot of guesswork involved in picking a new game to design and develop. But in the information age, we can take a lot of the guessing out of that work. And leverage our data science team to help make a data-driven decision that will highly increase our chances of picking a successful game.

#### A.2. Context and Background

Many properties or factors could contribute to a video game’s success, like platform, countries it was available, and type of game. The data science team proposes we analyze a dataset containing the highest-grossing video games of all time and look for factors that are most likely to increase a game’s chance to succeed. Once these factors are identified and we have had a chance to explore their relationship, we intend to pick an appropriate model and leverage the power of machine learning to help us predict what kind of game will bring us the highest chances of success. These predictions will help WCVG decide what game it will invest in developing next.

#### A.3. Outside Works Review

1. After researching, one of the best sources of data we have found was scraped from [VGChartz](https://www.vgchartz.com/). VGChartz is a well-recognized source of video game sales data (<https://en.wikipedia.org/wiki/VGChartz>). The scraped VGChartz data can be found on Kaggle (<https://www.kaggle.com/gregorut/videogamesales>) and is formatted as a CSV, so it should be easy to ingest and pre-process. The data contains information about video games that sold greater than 100,000 copies, including Platform (i.e., Wii, Xbox, or PC) the game was sold on, Genre, the year it was made, and regional sales totals. We can use this sales data as our measure of success and examine the rest of the features to see how well they can predict increased chances of total sales. (Smith, Gregory. 2016, October)
2. To help us better understand the data we found a great analysis of the above VGChartz dataset: <https://datascience.fm/video-game-sales-analysis/>. The analysis starts by posing some important questions like "Which region has performed the best in terms of sales?" and “What are the top 10 games currently making the most sales globally?” that can help us determine potential predicting features like region or outliers that may skew the data. The research paper uses many of the tools that we also plan to use for our project. Pandas is used to load the data into a python data structure making it easy to manipulate. The Plotly library is then used to help visualize the data for further analysis resulting in some very easy-to-read bar graphs showing the spread of the data. Some interesting conclusions that come out of the paper are that North America has the highest average sales, Xbox 360 is the top preferred platform in NA, and the topmost played game is Wii Sports. (Narnauli, Lehak. 2021, June)
3. Another great analysis source on the VGChartz dataset was found here: <https://medium.com/analytics-vidhya/a-data-driven-exploration-of-video-games-sales-and-scores-3c77f1c6573c>. This paper doesn’t set out to answer specific questions about the data but looks to see what types of general trends can be obtained from the dataset. Some valuable questions are posed like “How have video game sales changed over time?”, “What are the most popular video game consoles and their best-selling games?”, and what are the most popular games? Again, this paper uses Pandas for loading the data into an easy-to-use python data structure. Tableau is used as the main visualization tool for creating some stunning and useful line, bar, and heat map graphs. One very interesting trend is that around 2008 there was a peak in video game sales and until about 2015 (the end of the dataset) there has been a steady decline. The most popular console was PS2 which should be investigated further during our analysis as this differs from the top-selling platform of Xbox 360 in the first research paper. (Norris, Devin. 2021, February)

#### A.4. Solution Summary

We know that the dataset holds useful insights, and we can analyze them to help the company make an informed choice of our next game, however, with machine learning we intend to predict, with a much higher degree of accuracy, what game will do the best. Once candidate predictive factors have been identified, will clean the data by removing outlying factors, adjusting for orders of magnitude, and generalizing the data. After the data is prepped, we can do some manual analysis to help inform the selection and design of the model, for instance, this stage can be used to evaluate a linear or logistic regression. Now that we have our model, we will want to choose an objective function that will be used to measure how accurately the model’s outputs match the desired outcome. A chosen optimization algorithm will then mutate the objective function's parameters as we train and test the model until we have produced the factors with the highest degree of predictive accuracy. These results are what the data science team will use to predict and drive our recommendation for WCVG’s next big game.

#### A.5. Machine Learning Benefits

We've chosen machine learning for this project because it is a very powerful tool for helping companies make business decisions. Machine learning can significantly increase the chances that an impactful business decision will successfully achieve the desired goal. Our desired goal is to choose a game that will have the highest chances of success and reduce the risk of choosing a game that may fall short of the mark hurting our company and subsequently our employees. By analyzing a large dataset from successful video games, we can use machine learning to help us predict what kind of video game is most likely to do well in the future with a much higher degree of accuracy and performance than a human.

### B. Machine Learning Project Outline

#### B.1. Scope

In-scope:

* Research, manual analysis, and data preprocessing are used to help select and refine our model.
* Select and refine a Machine Learning model.
* Game type recommendation based on the results of our model.

Out-of-scope:

* Analysis of datasets other than the selected VGChartz dataset.
* The decision on what game WCVG creates next will be made by leadership.

#### B.2. Goals, Objectives, and Deliverables

##### Goals

* Develop a machine learning model to predict what type of video game is most likely to succeed with a high degree of accuracy.
* Use the machine learning model to predict what type of video game is most likely to succeed with a high degree of accuracy.

##### Objectives

* **Objective 1**: Preprocess and clean the data of outlying or non-essential factors that would skew the intended results.
* **Objective 2**: Analyze the data with statistical tools to help select our model and algorithm
* **Objective 3**: Select an objective function to measure the model’s accuracy and choose an optimization algorithm that will mutate objective function parameters as process our dataset through the model.
* **Objective 4**: A prediction on what game or game type will bring WCVG the most success is made.

##### Deliverables

* **Objective 1**:
  + A clean and generalized dataset including only the necessary factors that can be fed through our model.
* **Objective 2**:
  + Our model and Neural Network dimensions are known and documented.
* **Objective 3**:
  + The objective function is selected.
  + The optimization algorithm is selected.
* **Objective 4**:
  + The Data Science Team presents our prediction to leadership.

#### B.3. Standard Methodology

To guide our development of this project we will follow the SEMMA methodology:

• **Sample**:The sampling phase has already been completed and is discussed in section A.3. Outside Works Review. In summary, we will be using the following Kaggle dataset <https://www.kaggle.com/gregorut/videogamesales>. We picked this dataset because it should be easy to preprocess, contains data applicable to our use case, it’s not too large or small, from a reputable source, and have a format that should be easy to ingest and pre-process.

• **Explore**: The Explore phase aligns with our first objective (above). We will pre-process and visualize the data to get a better understanding of how we can model our data in the Modify phase.

• **Modify**: The Modify phase aligns with our second objective, wherein we will begin to determine and select the appropriate model for our data now that it has been preprocessed.

• **Model**: The Model phase aligns with our third objective. We will begin to tweak and tune our model(s) to produce the desired results.

• **Assess**: The Assess phase aligns with our last objective. The data science team will assess the model's results and predict what game will most likely succeed.

(SEMMA. 2021, August. In Wikipedia.)

#### B.4. Projected Timeline

4/18 – Objective 1 deliverables delivered.

5/2 – Objective 2 deliverables delivered.

5/16 – Objective 3 deliverables delivered.

6/30– Objective 4 deliverables delivered.

##### Sprint Schedule

|  |  |  |  |
| --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Tasks** |
| 1 | 4/5/2022 | 4/18/2022 | A clean and generalized dataset including only the necessary factors that can be fed through our model. |
| 2 | 4/19/2022 | 5/2/2022 | Our model and Neural Network dimensions are known and documented. |
| 3 | 5/3/2022 | 5/16/2022 | The objective function and optimization algorithms are selected. |
| 4 | 5/17/2022 | 6/30/2022 | The Data Science Team presents the prediction to leadership. |

#### B.5. Funding Requirements

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Human | One Data Scientist w/ avg salary of 100,000 | (100,000/52) \* 8 **= $15,384.00** |
| macOS Laptop | Needed for developing and running Machine Learning Model | **$3,899.00** |
|  | **Total** | **$19283.00** |

#### B.6. Evaluation Criteria

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Prediction Accuracy | Greater than 80% |

### C. Machine Learning Solution Design

#### C.1. Hypothesis

WCVG is ready to choose the next game it will develop. Given data on the industry's past games and their success factors, the data science team can predict with greater than 90% accuracy what type of game will bring WCVG the highest success.

#### C.2. Selected Algorithm

We intend to use a neural network; however, our choice of an algorithm may change after our Exploration phase and Modification phase once we have a better understanding of our data and the best algorithm for it.

##### C.2.a Algorithm Justification

We are less concerned with performance and more concerned with the accuracy of our results. We aren’t sure whether our data will be linear or how simple the relationship between predictors and the results(interoperability) is. There are only a few factors compared to the number of observations.

##### C.2.a.i. Algorithm Advantage

During pre-processing, we may need to rely on unsupervised learning methods like K-means clustering to help group and understand our data, however, our objective is to predict the next big game. Supervised learning is best suited for this type of problem. Our input data will consist of one or more finite features, like the type of game and the region, and predicting a continuous quantity like total predicted sales. Total sales could be an infinite amount like 200 copies or 1,000,000 copies depending on the input parameters. This is the type of problem that a regression algorithm is best suited for. For this reason, we will be using a supervised learning regression algorithm for our machine learning model. Supervised learning regression algorithms have the advantage of accurately predicting outcomes because they have a well-defined training phase. During the training phase, the algorithm will iterate over the input data with a direct feedback loop containing information like how much closer or further away the last iteration got from the intended goal. (Edureka Jan, 2019)

##### C.2.a.ii. Algorithm Limitation

We will likely use a linear or logistic regression algorithm depending on the shape of our input data. If the data is not linear, it won't make much sense to attempt to predict high total sales using a linear regression algorithm, because its accuracy may not be very high, and it would make more sense to use logistic regression. On the other hand, if the data is relatively linear a linear regression algorithm can be very accurate and with the potential added benefit of better performance and simplicity. (Edureka Jan, 2019)

#### C.3. Tools and Environment

Research and development will take place in a Jupyter Notebook. Processing of the data will take place in Python 3.8.7 primarily relying on Pandas, Numpy, Matplotlib, and Tensorflow libraries to help perform common data science tasks.

#### C.4. Performance Measurement

The performance of the algorithm will be measured by the time it takes to complete and its effectiveness will be measured by determining how close it can predict actual values. This can be accomplished using test data. A subset of our original dataset will be kept for comparison and to ensure the results are accurate.

### D. Description of Data Sets

#### D.1. Data Source

Our data will sample will be obtained from Kaggle, a well-known repository for good clean datasets. Our sample can be found here, <https://www.kaggle.com/gregorut/videogamesales>. Which was sourced from VGChartz, a well-recognized source of video game sales data.

Another key differentiator for using this dataset is that other folks have published quite in-depth works doing statistical analysis of data:

1. <https://datascience.fm/video-game-sales-analysis/>
2. <https://medium.com/analytics-vidhya/a-data-driven-exploration-of-video-games-sales-and-scores-3c77f1c6573c>

Both research papers use Pandas and Numpy, two python packages that we will also be using in our project to help us pre-process and analyze the data. They also provide tons of valuable insights into the data like the country with the highest average sales, most preferred platforms, and highest-grossing or most played games. While we will be doing our analysis of the data these will be great sources to help speed up the pre-requisites of pre-processing the data, identifying predicting factors, and picking our machine learning model. (Narnauli, Lehak. 2021, June; Norris, Devin. 2021, February)

#### D.2. Data Collection Method

Our data is taken from a Kaggle dataset. The Kaggle dataset is formatted as CSV which can be retrieved and loaded for processing with the following steps:

1. Download the CSV by following this link: <https://www.kaggle.com/gregorut/videogamesales/download>
2. Unzip the file to “vgsales.csv”
3. The file can now be loaded into a python pandas data frame with the following code:

import pandas as pd

df = pd.read\_csv(‘vgsales.csv’)

(Norris, Devin. 2021, February)

##### D.2.a.i. Data Collection Method Advantage

The advantage to Kaggle is the data has already been scraped from the source and formatted to make it easier to retrieve, load, and pre-process. The CSV file can be downloaded on any OS, and it should be easy to ingest and pre-process using our Python libraries. Kaggle also includes discussions, descriptions, and usability scores to help pick well-curated datasets that fit your use case. The dataset has been processed and researched by many data scientists before us and contains over 16,500 games worth of data. It includes some very valuable features that will apply directly to our use case: Name, platform, overall sales per region, and the year of the game’s release. (Smith, Gregory. 2016, October)

##### D.2.a.ii. Data Collection Method Limitation

We are downloading the data manually from Kaggle, which can be slow, error-prone, and not easy to update. If the CSV is updated there isn’t a mechanism to detect that and re-run the algorithm, nor a great. Since we are doing point-in-time analytics of the data, having to manually download, and load the data in python, and not being able to detect updates and respond to updates are acceptable trade-offs for the scope of our project.

#### D.3. Quality and Completeness of Data

Our data is from a point in time that was 5 years ago. Trends that affected video game sales in the past 5 years are not available and may skew our results. The usability score is a bit low for Kaggle as well, which may mean it requires a lot of pre-processing to account for missing or non-standardized values. During pre-processing the data, we will be accounting for missing values, standardizing numeric scales to account for outliers, reformatting data so that it is easier to work with, and removing misleading factors. (Smith, Gregory. 2016, October)

#### D.4. Precautions for Sensitive Data

By using a publicly available data set we have removed the need to worry about protecting sensitive data from the sample we have obtained. However, the results of our research and the predictions made should be kept internal to WCVG as they could be damaging to the company in a competitor’s hands.

### References

1. Smith, Gregory. (2016, October). Video Game Sales. Kaggle. <https://www.kaggle.com/gregorut/videogamesales>
2. Narnauli, Lehak. (2021, June). Video Game Sales Analysis. Datascience.fm. <https://datascience.fm/video-game-sales-analysis/>
3. Norris, Devin. (2021, February). A Data-Driven Exploration of Video Games — Sales and Scores. The Medium.
4. <https://medium.com/analytics-vidhya/a-data-driven-exploration-of-video-games-sales-and-scores-3c77f1c6573c>
5. SEMMA. (2021, August). In Wikipedia. <https://en.wikipedia.org/wiki/SEMMA>
6. Edureka (Jan, 2019). Supervised vs Unsupervised vs Reinforcement Learning | Data Science Certification Training | Edureka. Youtube. <https://www.youtube.com/watch?v=xtOg44r6dsE>

# Executive Letter

## Problem Description

With our recent successes, one of the biggest problems that WCVG is faced with is deciding what next to do with our time and development resources. We want to pick a game that will maximize our profits, expand our brand awareness, and crush the top charts. There is a lot of guesswork involved in picking a new game to design and develop. But in the information age, we can take a lot of the guessing out of that work. And leverage our data science team to help make a data-driven decision that will highly increase our chances of picking a successful game.

## Customers

The data science team will be serving the broader WCVG company by providing support for deciding what type of game will be developed next. In the short term, the decision will directly impact the data science, sales, and engineering teams. In the long-term, it will impact all organizations and departments within WCVG as this decision affect our future revenue streams.

## Data

Our data will sample will be obtained from Kaggle, a well-known repository for good clean datasets. Our sample can be found here, <https://www.kaggle.com/gregorut/videogamesales>. Which was sourced from VGChartz, a well-recognized source of video game sales data.

## Methodology

To guide our development of this project we will follow the SEMMA methodology:

• **Sample**:The sampling phase has already been completed and is discussed in section A.3. Outside Works Review. In summary, we will be using the following Kaggle dataset <https://www.kaggle.com/gregorut/videogamesales>. We picked this dataset because it should be easy to preprocess, contains data applicable to our use case, it’s not too large or small, from a reputable source, and have a format that should be easy to ingest and pre-process.

• **Explore**: The Explore phase aligns with our first objective (above). We will pre-process and visualize the data to get a better understanding of how we can model our data in the Modify phase.

• **Modify**: The Modify phase aligns with our second objective, wherein we will begin to determine and select the appropriate model for our data now that it has been preprocessed.

• **Model**: The Model phase aligns with our third objective. We will begin to tweak and tune our model(s) to produce the desired results.

• **Assess**: The Assess phase aligns with our last objective. The data science team will assess the model's results and predict what game will most likely succeed.

(SEMMA. 2021, August. In Wikipedia.)

## Deliverables

The data science team will be delivering a data product that will contain statistical analysis of industry sales data and a machine-learning algorithm to predict the global sales of a video game given a number of features.

* **Objective 1**:
  + A clean and generalized dataset including only the necessary factors that can be fed through our model.
* **Objective 2**:
  + Our model is known and documented.
* **Objective 3**:
  + The objective function is selected.
  + The optimization algorithm is selected.
* **Objective 4**:
  + The Data Science Team presents our prediction to leadership.

## Implementation Plan

We know that the dataset holds useful insights, and we can analyze them to help the company make an informed choice of our next game, however, with machine learning we intend to predict, with a much higher degree of accuracy, what game will do the best. Once candidate predictive factors have been identified, will clean the data by removing outlying factors, adjusting for orders of magnitude, and generalizing the data. After the data is prepped, we can do some manual analysis to help inform the selection and design of the model, for instance, this stage can be used to evaluate a linear or logistic regression. Now that we have our model, we will want to choose an objective function that will be used to measure how accurately the model’s outputs match the desired outcome. A chosen optimization algorithm will then mutate the objective function's parameters as we train and test the model until we have produced the factors with the highest degree of predictive accuracy. These results are what the data science team will use to predict and drive our recommendation for WCVG’s next big game.

## Success Criteria and Verification

Our model’s prediction accuracy will be compared against our success criteria to evaluate whether we were able to meet our objective.

|  |  |
| --- | --- |
| **Objective** | **Success Criteria** |
| Prediction Accuracy | Greater than 80% |

## Cost

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Human | One Data Scientist w/ avg salary of 100,000 | (100,000/52) \* 8 **= $15,384.00** |
| macOS Laptop | Needed for developing and running Machine Learning Model | **$3,899.00** |
|  | **Total** | **$19283.00** |

## Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sprint** | **Start** | **End** | **Dependencies** | **Tasks** |
| 1 | 4/5/2022 | 4/18/2022 | None | A clean and generalized dataset including only the necessary factors that can be fed through our model. |
| 2 | 4/19/2022 | 5/2/2022 | Task 1 | Our model is known and documented. Our model is known and documented. |
| 3 | 5/3/2022 | 5/16/2022 | Task 2 | The objective function and optimization algorithms are selected. |
| 4 | 5/17/2022 | 6/30/2022 | Task 3 | The Data Science Team presents the prediction to leadership. |

# Business Requirements

The following document is a summary of our business requirements and the objectives that will allow us to achieve them.

##### Goals

* Develop a machine learning model to predict what type of video game is most likely to succeed with a high degree of accuracy.
* Use the machine learning model to predict what type of video game is most likely to succeed with a high degree of accuracy.

##### Objectives

* **Objective 1**: Preprocess and clean the data of outlying or non-essential factors that would skew the intended results.
* **Objective 2**: Analyze the data with statistical tools to help select our model and algorithm
* **Objective 3**: Implement a Machine Learning algorithm.
* **Objective 4**: A prediction on what game or game type will bring WCVG the most success is made.

# Quickstart Guide

## Prerequisites:

* Jupyter installed, see: <https://jupyter.org/install>
* The following python packages:

Graphical user interface, text, application, email

Description automatically generated

## Important Files

* Wgu-capstone-project.ipynb: The Jupyter notebook containing the data product used to predict our next successful game along with supporting visualizations, statistical analysis, and Machine Learning results.
* Vgsales.csv: This is the vgchartz sales dataset retrieved from Kaggle: <https://www.kaggle.com/datasets/gregorut/videogamesales>
* This document.

## Usage:

1. From a terminal, change your working directory to this project’s directory.
2. Run `jupyter notebook`, this will start the jupyter notebook server:

Text

Description automatically generated

1. As shown in the screenshot above, the server is listening on `localhost:8888` and it should also open a bowser listing the contents of the projects:

Graphical user interface, application, Teams

Description automatically generated

1. Click on `wgu-capstone-project.ipynb`, which should take you to the jupyter notebook which serves as both the data product and analysis.
2. Click on Kernel > Restart & Run All to load the data, normalize the data, populate graphs, and run our machine learning algorithms. Test and Train datasets are randomly selected on a 70/30 split so re-running this step will vary the outcome of the regressions.