title: "Video Game Sales Analysis"

## Contents

1 Introduction ......................................................................... 1\

2 Methods and Analysis .............................................................. 2-9\

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.1 Data Generation and Cleaning ............................................... 2-3\

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.2 Data Visualization and Exploration .......................................... 3-9 \

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.3 Data Modeling .................................................,............ 9-16\

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.3.1 kNN Model ........................................................... 10-11\

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.3.2 Naive Bayes Model ................................................... 12-13\

&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;&nbsp; 2.3.3 Random Forests Model ................................................ 13-16\

3 Results and Conclusions ........................................................... 17-19\

4 Reference ........................................................................... 19\

5 Appendix - Environment ............................................................ 19\

## 1 Introduction

As part of the HarvardX's Data Science Professional Certificate Program, students get the opportunity to perform formal data science analyses in our Capstone course using the knowledge gained from our coursework.\

We have selected video games sales for our analysis. Since their debut, video games have exploded in popularity through the years. This year video games in the U.S alone, has produced on average over 1 billion dollars in sales per month as reported on www.statista.com. Being such a significant part of the market share, it is exciting to dive in and learn more about the video game business.\

The main objective of our project is to apply machine learning techniques that go beyond standard linear regression using a publicly available data set to solve a problem of our choice. We will generate our data set, interpret the data, perform various algorithms/modelings and present our insights. We will review which video games were popular, explore and look for trends and develop some prediction models for video game ranking and sales.\

## 2 Methods and Analysis

2.1 Data Generation and Cleaning\

The video games sales data was downloaded from Kaggle (https://www.kaggle.com/gregorut/videogamesales) as a vgsales.csv file. We imported the data programmatically with the R script below. We will use the below script to generate and prepare this vgsales data. We have cleaned the data by removing the rows that were not applicable from the 'Year' and 'Publisher' columns.

# Video games sales data set:

# https://www.kaggle.com/gregorut/videogamesales

# Manually downloaded and included the vgsales.csv file

# Use the below code to import programmically or use R Studio to import the data

# View the structure of the data set

# View data summarys

The video game sales data were from these 5 areas: North America, Europe, Japan, Other and Global. There are 576 game publishers and 31 platforms.