Film Box Office Gross Prediction Tool Capstone Project

Western Governors University

Name: Joshua James

Student ID: 011101737

Date: 12/17/23

Table of Contents

[**Prompt A** 4](#_Toc25684)

[Letter of Transmittal 4](#_Toc25685)

[Project Recommendation 6](#_Toc25686)

[Problem Summary 6](#_Toc25687)

[Application Benefits 6](#_Toc25688)

[Application Description 6](#_Toc25689)

[Data Description 8](#_Toc25690)

[Objective and Hypothesis 8](#_Toc25691)

[Methodology 9](#_Toc25692)

[Funding Requirements 9](#_Toc25693)

[Stakeholders Impact 11](#_Toc25694)

[Data Precautions 12](#_Toc25695)

[Developer Expertise 13](#_Toc25696)

[**Prompt B** 14](#_Toc25697)

[Project Proposal 14](#_Toc25698)

[PROBLEM STATEMENT 14](#_Toc25699)

[CUSTOMER SUMMARY 14](#_Toc25700)

[EXISTING SYSTEM ANALYSIS 15](#_Toc25701)

[DATA 15](#_Toc25702)

[PROJECT METHODOLOGY 16](#_Toc25703)

[PROJECT OUTCOMES 18](#_Toc25704)

[IMPLEMENTATION PLAN 18](#_Toc25705)

[EVALUATION PLAN 18](#_Toc25706)

[RESOURCES AND COSTS 19](#_Toc25707)

[TIMELINE AND MILESTONES 20](#_Toc25708)

[**Prompt C** 22](#_Toc25709)

[Application Files 22](#_Toc25710)

[**Prompt D** 23](#_Toc25711)

[Post-implementation Report 25](#_Toc25712)

[Project purpose 25](#_Toc25713)

[Datasets 25](#_Toc25714)

[Data product code 25](#_Toc25715)

[Hypothesis verification 27](#_Toc25716)

[Effective visualizations and reporting 28](#_Toc25717)

[Accuracy analysis 29](#_Toc25718)

[Application testing 34](#_Toc25719)

[**Appendices** 34](#_Toc25720)

[Installation Guide 34](#_Toc25721)

[User Guide 36](#_Toc25722)

[Summation of Learning Experience 37](#_Toc25723)

[References 39](#_Toc25724)

# Prompt A

## Letter of Transmittal

Subject: Film Box Office Gross Prediction Tool

Fred’s Film Productions

Hollywood, CA 90028

Dear Fred’s Film Productions,

With the enormous costs associated with making a film, it is vital to be able to accurately predict whether a film will be a success or not upon release into theaters. Even one failure at the box office can lead to financial ruin for a smaller studio such as Fred’s Film Productions, especially in the face of the ever-increasing popularity of streaming services leading to lower theater attendance. I am writing this letter of transmittal to propose a solution to create a tool that uses machine learning to accurately predict a film’s box office gross. This would provide a vital tool that would help decrease the risk of releasing a box office failure and allow for better management of budgets, marketing costs, and the allocation of resources.

To explain further, I propose to create a film box office gross prediction tool that uses a machine learning algorithm to accurately predict a film’s box office gross based on its budget. I estimate that this tool would cost approximately $11,750, with an additional monthly cost of $4,250, and take approximately 14 weeks to design, test, and deploy. I believe the cost of this tool could be easily recouped when used to plan for future film budgets, marketing costs, and other resources.

As a soon-to-be Computer Science graduate from Western Governor’s University (WGU), I have the education and programming experience to create this solution to your studio’s prediction problem. During my time at WGU, I used machine learning and the Python programming language to develop projects similar to this. In addition, I am a big fan of your studio’s recent films and would passionately develop a tool to help your studio maximize its film profits and revenue.

Thank you for taking the time to read this letter. I am excited to have Fred’s Film Productions review our proposal and look forward to hearing from you soon.

Sincerely,

Joshua James

WGU Computer Science Major

## Project Recommendation

### Problem Summary

Fred’s Film Productions is a small, yet up-and-coming film studio that distributes its films to theaters, before eventually going to streaming and disc services (DVD/Blu-Ray). Currently, Fred’s Film Productions does not have a way to accurately predict a film’s box office gross. Their current method involves studio leadership making an “educated guess” based on their intuition, knowledge, and personal experiences. Fred’s Film Productions recognizes how this is a very subjective and inconsistent way to attempt to make a film’s box office gross prediction and how it can easily lead to budgeting and marketing cost errors that negatively impact the studio. I propose to develop an easy-to-use tool, along with an associated installation and user guide, that uses a machine learning algorithm and historical film budget and box office gross data, to make more accurate and data-driven film box office gross predictions than the current method being used by the studio’s leadership. By being able to more accurately predict a film’s box office gross, Fred’s Film Productions will be able to better maximize profits and cost savings, due to being able to plan its film budgets, marketing costs, and resources much more effectively.

### Application Benefits

My proposed film box office prediction tool would provide Fred’s Film Productions with a solution to their box office gross prediction problem. Instead of a subjective, user-biased guess made by studio leadership, my proposed solution would provide an objective, data-driven way to make more accurate predictions. This would benefit Fred’s Film Productions in multiple ways, including:

* **Budget Planning**
  + Using the proposed tool will allow studio leadership to find the predicted minimum budget needed for a film to produce a box office gross profit.
* **Marketing Costs**
  + By having the predicted box office gross amount, studio leadership can manage marketing costs more effectively. For example, leadership could decide to move some of the marketing funding from films that are predicted to do poorly at the box office, to films that are predicted to do better financially at the box office.
* **Better Resource Allocation**
  + Based on a film’s predicted box office gross, leadership could also reallocate resources other than budget or marketing, such as staffing, equipment, and writers, to films that are predicted to do better at the box office.
* **More Informed Decision Making**
  + Having more accurate information will allow studio leadership to make more informed, data-driven decisions.
* **Ease of Use**
  + My proposed box office prediction tool does not require a user to have extensive knowledge of a film, or technical expertise to use. It only needs a film’s budget amount to make an accurate box office gross prediction.

### Application Description

My proposed film box office prediction tool will be developed using the Python programming language and will use a linear regression machine learning algorithm and historical movie budget and box office gross data, to accurately predict a future film’s box office gross based on its budget. In other words, by using machine learning and historical budget and box office gross data to train my proposed film box office prediction tool, it will then be able to make an accurate prediction for a film’s box office gross based on its budget.

### Data Description

To train my proposed box office gross prediction tool, a dataset will need to be obtained. For this project, the movie industry dataset that will be used to train the proposed film box office gross prediction tool will be obtained from Kaggle.com. This dataset is in a CSV file format and contains 7,668 films, from 1980 to 2020. Each film includes associated information such as the film’s title, genre, director, budget, box office gross, etc. The dataset will be cleaned by removing films that do not have budget or box office gross data before being used by a linear regression machine learning algorithm to train my proposed film box office gross prediction tool. Please note, that in this proposed project, the box office gross data is the dependent variable, while the film’s budget data is the independent variable. In the future, by updating the movie industry dataset with movie data from movies that have come out since the last update, the tool will be able to continue to provide accurate and relevant box office gross predictions.

### Objective and Hypothesis

By using my proposed film box office gross prediction tool, Fred’s Film Productions will see an increase in its revenue and profitability. Being able to more accurately forecast a film’s box office gross, Fred’s Film Productions will be able to make much more informed, data-driven decisions regarding budgets, marketing costs, and resources, which in turn, will lead to an increase in revenue and profitability. Upon deployment to its live environment, the tool’s predictions will be tracked and compared to the film’s eventual box office gross to rate its accuracy to ensure it performs above Fred’s Film Productions current method, in which studio leadership makes an “educated guess” based on their intuition, knowledge, and personal experiences. In addition, studio revenue and profitability will also be monitored and tracked, to ensure the tool’s introduction is leading to an increase in both.

### Methodology

The development of my proposed box office gross prediction tool will use the commonly used Waterfall methodology. The Waterfall methodology is a linear, sequential approach to software development, that typically has an established timeline and fixed requirements that are unlikely to change. This methodology typically involves six steps:

**Requirements:** During the requirements phase, I will meet with Fred’s Film Productions to gain a clear understanding of their requirements for the box office prediction tool.

**Design:** During the design phase, solutions to the requirements are created and designed.

**Implementation:** During the implementation phase, one of the design solutions will be selected to be developed.

**Verification:** During the verification phase, the implementation created during the previous phase, is tested and verified to meet the requirements established in the requirements phase.

**Deployment:** The box office prediction tool is deployed into its live environment, which for this project, would be the studio leadership members.

**Maintenance:** After deployment, the box office prediction tool will remain to be monitored to ensure of its accuracy, to fix any issues that may arise, and to periodically update and upgrade the tool, by updating the film dataset or machine learning algorithm.

In addition, there are multiple advantages to using the Waterfall methodology which I believe will be beneficial to this project. Some of these benefits include:

* **Clear and Intuitive Structure:** Using a Waterfall methodology means that every project has to go through the same sequence of six phases (requirements, design, implementation, verification, deployment, and maintenance), with each phase needing to be completed fully before moving on to the next phase. This creates a clear, sequential, and defined set of steps or phases for a project that a team can easily and intuitively understand. In addition, due to each phase needing to be completed fully before moving to the next phase, any problems that arise during development can be easily and quickly identified.
* **Early Determination of End Goal or Product:** With the Waterfall methodology, the end goal or product, in this case, a box office gross prediction tool, is determined during the initial planning phase. This allows for a team to have clear, defined goals from the start of the project, which helps to lessen the potential for a project to accidentally complete unnecessary and time-wasting tangents.
* **Ease of Management and Stability:** The Waterfall methodology is easy to manage and stable due to its linear, sequential approach, and the need to complete each phase fully before moving on to the next. This allows the methodology to be easier to track progress, identify bottlenecks or delays, and stay on an agreed-upon timescale, for each step or phase of a project.
* **Fewer Financial Surprises:** By using the Waterfall methodology, it is easier to more accurately predict the financial cost of the project, due to the project and each phase, being defined during the initial planning phase.

### Funding Requirements

I anticipate that my proposed film box office gross prediction tool will require $11,750, with an additional monthly cost of $4,250, to fully fund. The majority of the cost is related to the labor costs associated with the design and development of the tool, which I estimate to be a total of $10,000. In addition, software testing costs are estimated to be approximately $1,250. The monthly cost of $4,250 is related to server and machine learning costs to train and tune the tool. Please note, however, that the development tools needed for this project are free, so no additional costs are needed there. In summary, I estimate an initial cost of $11,750 to develop and deploy the tool, with an additional monthly cost of $4,250 for server and machine learning services.

### Stakeholders Impact:

If my proposal is accepted, then while Fred’s Film Productions will initially see a negative financial impact to fund the project, once the tool is deployed, the studio will quickly recoup this cost and see multiple benefits, including:

* Using the proposed tool will allow studio leadership to find the predicted minimum budget needed for a film to produce a box office gross profit. They can then reallocate budget costs as needed to ensure profitability.
* By having the predicted box office gross amount, studio leadership can manage marketing costs more effectively. For example, leadership could decide to move some of the marketing funding from films that are predicted to do poorly at the box office, to films that are predicted to do better financially at the box office.
* Based on a film’s predicted box office gross, leadership could also reallocate resources other than budget or marketing, such as staffing, equipment, and writers, to films that are predicted to do better at the box office.
* Having more accurate information will allow studio leadership to make more informed, data-driven decisions.

### Data Precautions

The movie industry dataset used for training is publicly available through Kaggle.com and requires no additional data precautions. However, the tool's development and Fred Film Productions' film budget information will be kept secure and confidential from competitors.

### Developer Expertise

As a soon-to-be graduating Computer Science major from Western Governors University (WGU), I would make an excellent candidate to develop and deploy my proposed film box office gross prediction tool. During my time at WGU, I have created multiple projects similar to this, using Python, the proposed programming language for this project, and other programming languages. I have also learned project management, design, and development processes, as well as acquired an ITIL Foundation Certification, which will ensure this project is completed on time and within budget. In addition, I also have experience and have learned multiple best practices in providing the proper documentation, such as including an installation guide and user guide, and end-of-project reporting that will be beneficial when using the tool. Based on my time, experience, and projects completed at WGU, I believe I would be more than capable of completing this proposal for Fred’s Film Productions within schedule and budget.

# Prompt B

## Project Proposal

### PROBLEM STATEMENT

With the enormous costs associated with making a film, it is vital to be able to accurately predict whether a film will be a success or not upon release into theaters. Even one failure at the box office can lead to financial ruin for a smaller studio such as Fred’s Film Productions, especially in the face of the ever-increasing popularity of streaming services leading to lower theater attendance. While Fred’s Film Productions would like to capitalize on its recent film popularity, studio leadership is hesitant to increase future film budgets due to the increased risk and negative impact a negative-grossing movie would have on the studio. Currently, box office gross amounts are predicted by studio leadership, based on their intuition, knowledge, and personal experiences. However, Fred’s Film Productions recognizes this process is inherently biased and subjective, leading to a high risk of inaccuracy. My proposal would create a tool that uses machine learning to provide studio leadership with a much more subjective, accurate, informed, and data-driven box office gross prediction. This would lead to studio leadership being able to make more informed decisions regarding budget, marketing, and resource allocations, thus lowering the risk of a potential box office failure, and increasing savings, film profits, and studio revenue.

### CUSTOMER SUMMARY

Fred’s Film Productions is a small, yet up-and-coming movie studio, that has seen its recent films become increasingly more popular and attracting a national audience. Due to this influx of popularity, the studio leadership is planning to increase the budgets of its future film releases, however, as described above, they are concerned about the increased risk and impact a negative grossing film release would cause on their up-and-coming studio. My proposed film box office prediction tool would allow for the studio leadership to make more data-driven, informed decisions regarding budget, resource, and marketing cost allocations, thus lowering the risk of a potential box office failure. In addition, my proposed tool would be very user-friendly, allowing even non-technical members of the studio leadership to use it with ease.

### EXISTING SYSTEM ANALYSIS

Currently, Fred’s Film Productions relies on its studio leadership to manually evaluate and make box office gross predictions based on their intuition, knowledge, and experiences. Fred’s Film Productions understands this is a high-risk, biased, and subjective process, leading to inconsistent and inaccurate predictions. By accepting my proposal, the proposed film box office gross prediction tool will use machine learning and the vast amount of available historical film budget and box office data, to create more accurate, unbiased, data-driven box office gross predictions. This will lead to studio leadership being able to make more informed and less risky decisions. In addition, if my proposal is accepted, studio leadership will no longer need to perform manual evaluations of predicting box office grosses, thus freeing up their time to perform other important tasks.

### DATA

To train my proposed box office gross prediction tool, a dataset will need to be obtained. For this project, the movie industry dataset that will be used to train the proposed film box office gross prediction tool will be obtained from Kaggle.com. This dataset is in a CSV file format and contains 7,668 films, from 1980 to 2020. Each film includes associated information such as the film’s title, genre, director, budget, box office gross, etc. The dataset will be cleaned by removing films that do not have budget or box office gross data before being used by a linear regression machine learning algorithm to train my proposed film box office gross prediction tool. Please note, that in this proposed project, the box office gross data is the dependent variable, while the film’s budget data is the independent variable. In the future, by updating the movie industry dataset with movie data from movies that have come out since the last update, the tool will be able to continue to provide accurate and relevant box office gross predictions.

### PROJECT METHODOLOGY

The development of my proposed box office gross prediction tool will use the Waterfall methodology. The Waterfall methodology is a linear, sequential approach to software development, that typically has an established timeline and fixed requirements that are unlikely to change. There are multiple advantages to using the Waterfall methodology, some of which include:

* **Clear and Intuitive Structure:** Using a Waterfall methodology means that every project has to go through the same sequence of six phases (requirements, design, implementation, verification, deployment, and maintenance), with each phase needing to be completed fully before moving on to the next phase. This creates a clear, sequential, and defined set of steps or phases for a project that a team can easily and intuitively understand. In addition, due to each phase needing to be completed fully before moving to the next phase, any problems that arise during development can be easily and quickly identified.
* **Early Determination of End Goal or Product:** With the Waterfall methodology, the end goal or product, in this case, a box office gross prediction tool, is determined during the initial planning phase. This allows for a team to have clear, defined goals from the start of the project, which helps to lessen the potential for a project to accidentally complete unnecessary and time-wasting tangents.
* **Ease of Management and Stability:** The Waterfall methodology is easy to manage and stable due to its linear, sequential approach, and the need to complete each phase fully before moving on to the next. This allows the methodology to be easier to track progress, identify bottlenecks or delays, and stay on an agreed-upon timescale, for each step or phase of a project.
* **Fewer Financial Surprises:** By using the Waterfall methodology, it is easier to more accurately predict the financial cost of the project, due to the project and each phase, being defined during the initial planning phase.

### PROJECT OUTCOMES

The main outcome of the proposed project is the creation of a user-friendly film box office prediction tool, in which Fred’s Film Productions’ studio leadership would be able to input a film’s budget and they would be provided with a box office gross prediction. In addition to this tool, three data visualizations will also be included to assist studio leadership in gaining a better understanding of the dataset used for the machine learning training. Finally, a user guide and an installation guide will also be included, to ensure studio leadership will be able to install and use the tool fully.

### IMPLEMENTATION PLAN

The development of my proposed box office prediction tool involves using machine learning. The implementation of the machine learning solution will follow the commonly used SEMMA methodology (1).

* **Sample:** A representative sample of film industry data, including budget and box office gross information, will be extracted.
* **Explore:** The film industry data will be analyzed to help identify trends and groupings that will assist in making accurate box office gross predictions.
* **Modify:** The film industry data is increased and decreased, and organized into groups and sub-groups, to test which makes the most accurate box office gross predictions.
* **Model:** The film industry data, specifically the budget and box office gross data, will be used to train the box office gross prediction tool using a linear regression machine learning algorithm.
* **Assess:** The box office prediction tool will be tested to ensure it is meeting the agreed-upon objectives.

### EVALUATION PLAN

Throughout the project, verification methods will be used during each stage of development. During the requirements phase, I will meet with studio leadership to create clear and well-understood requirements for the project. During the design phase, once solutions to the requirements are created and designed, I will meet once again with studio leadership to ensure the selected solution meets their approval. In the implementation phase, the code will be reviewed and optimized to ensure it meets the requirements and design agreed upon in the earlier phases. Software testing will occur next during the verification phase, in which any errors or defects will be fixed before moving on to the next phase. Upon deployment to its live environment, I will ensure that the tool is working properly for each member of the studio’s leadership. Maintenance will be performed periodically to ensure the dataset is up to date, and that the machine learning algorithm is optimized. After deployment, I will follow up with studio leadership to conduct additional testing to ensure the tool is meeting the intended objectives and requirements. Feedback will be implemented, and updates to the tool will be made if needed. In addition, standard end-of-project documentation and reporting will be provided to the studio.

### RESOURCES AND COSTS

|  |  |  |
| --- | --- | --- |
| **Resource** | **Description** | **Cost** |
| Design and Development | Labor costs include the hiring of a project manager, UI/UX designers, and software developers. | $10,000 |
| Development Tools | The cost of the IDE and development tools used to develop the proposed film box office gross predictor. | Free |
| Machine Learning Services and Server | Server and machine learning services monthly cost that includes the training and tuning of the film box office gross predictor. | $4,250 per month |
| Software Testing | The overall cost for software testing. | $1,250 |
|  | **Total** | $11,750 |
|  | **Monthly Cost** | $4,250 |

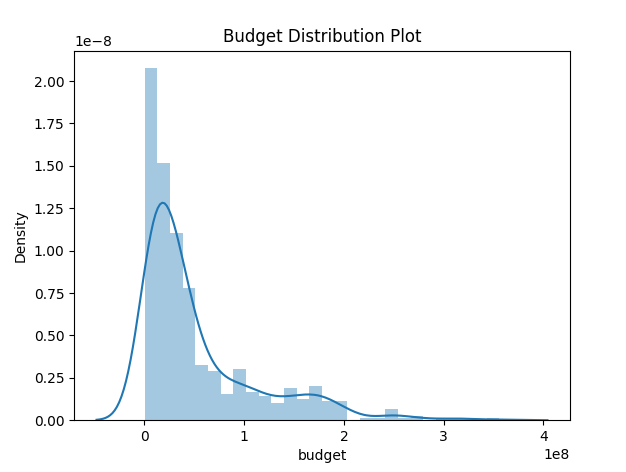
### TIMELINE AND MILESTONES

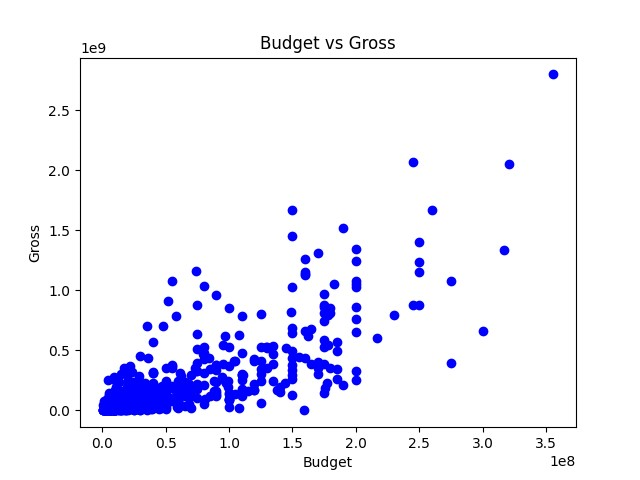
While the project timeline may change based on the proposal acceptance date, team size, budget, etc., please see below for the current projected timeline for the proposed film box office prediction tool:

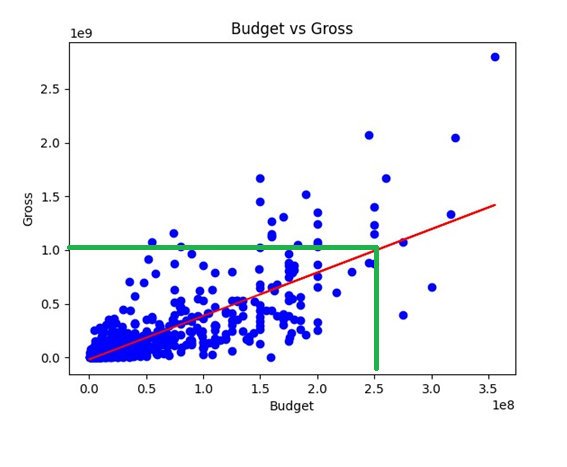
|  |  |  |
| --- | --- | --- |
| **Date** | **Description** | **Length** |
| January 1, 2024 | The film Box Office Gross Predictor proposal is accepted. Project official start date. | 1 Day |
| January 2, 2024 | Collect, prepare, and clean the film industry dataset. | 2 Weeks |
| January 17, 2024 | Design and develop film box office gross predictor tool. | 3 Weeks |
| February 18, 2024 | Use the film industry dataset for training and testing. | 2 Weeks |
| April 1, 2024 | User testing is performed and feedback is obtained. | 2 Weeks |
| April 16, 2024 | Based on testing and feedback received, the film box office gross predictor tool is updated and optimized. | 2 Weeks |
| May 1, 2024 | The film box office gross predictor tool is deployed to studio leadership. Its performance will be reviewed and evaluated to ensure it is functioning as intended. | 2 Weeks |
| May 16, 2024 | End-of-project reporting is performed. | 1 Week |
| May 24, 2024 | The project is officially completed. | 1 Day |

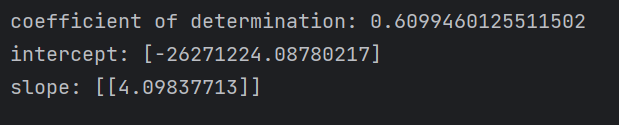
# Prompt C

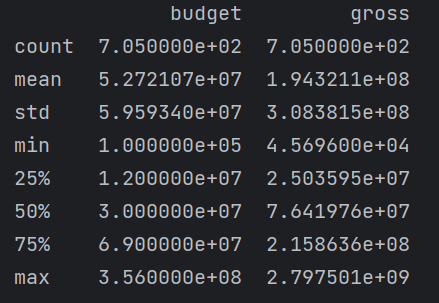
## Application Files

* Please refer to the zipped file named “C964Task2” included in this submission, to use the box office gross prediction tool. Information regarding the installation and use of the tool can be found in “Appendices: Installation Guide” and “Appendices: User Guide”.
* Please see the visualizations below. More information about each visualization can be found in “Prompt D: Effective Visualizations and Reporting”.

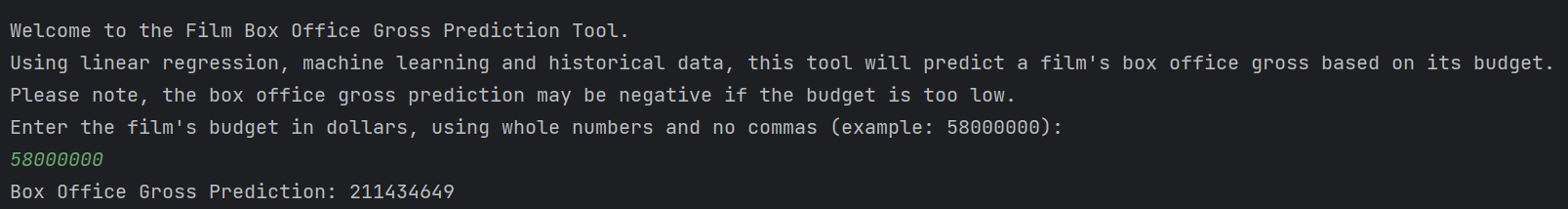








* The below “User Interface” allows a user to input a film’s budget, in which the tool outputs a “Box Office Gross Prediction” based on the budget entered.



# Prompt D

## Post-implementation Report

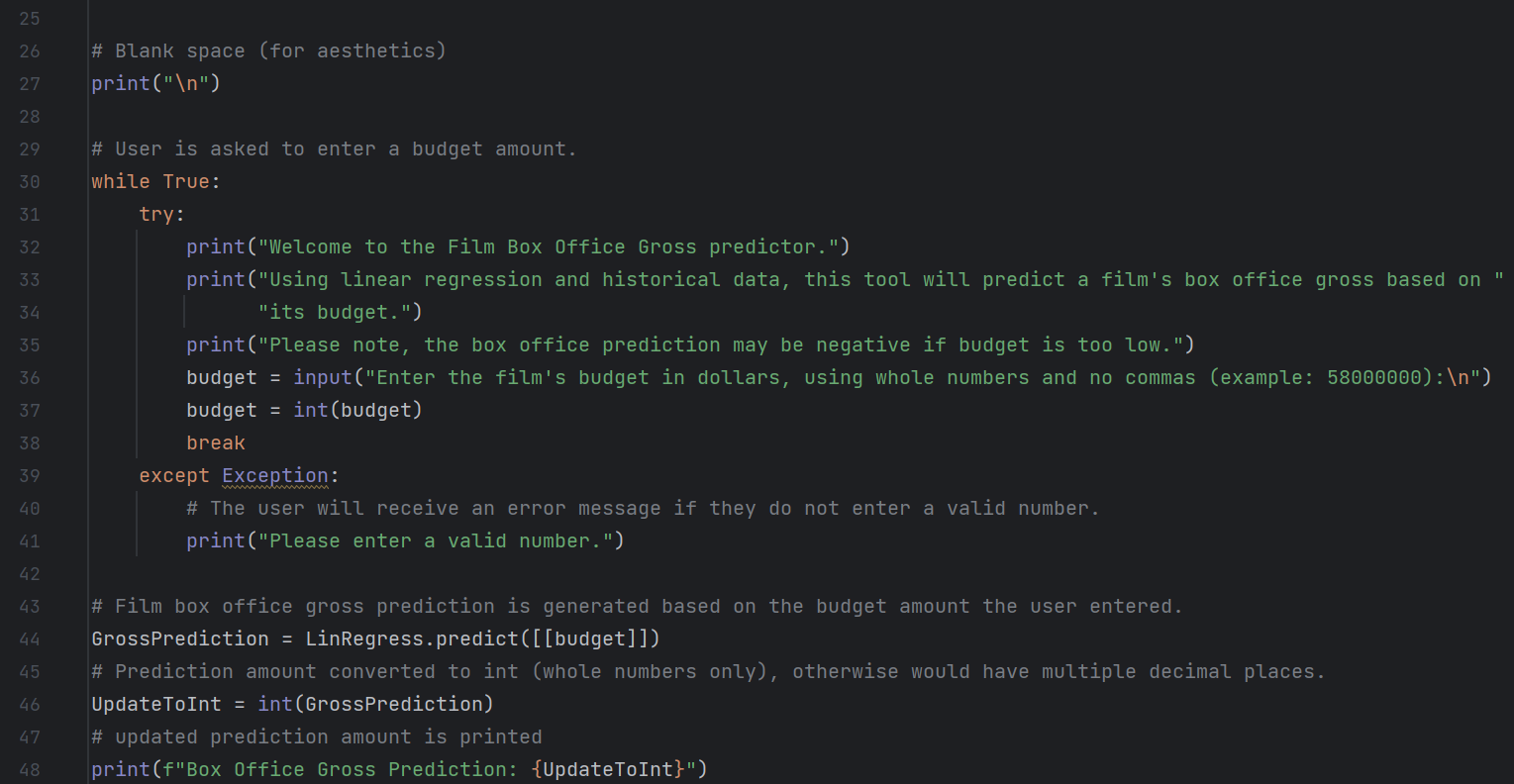
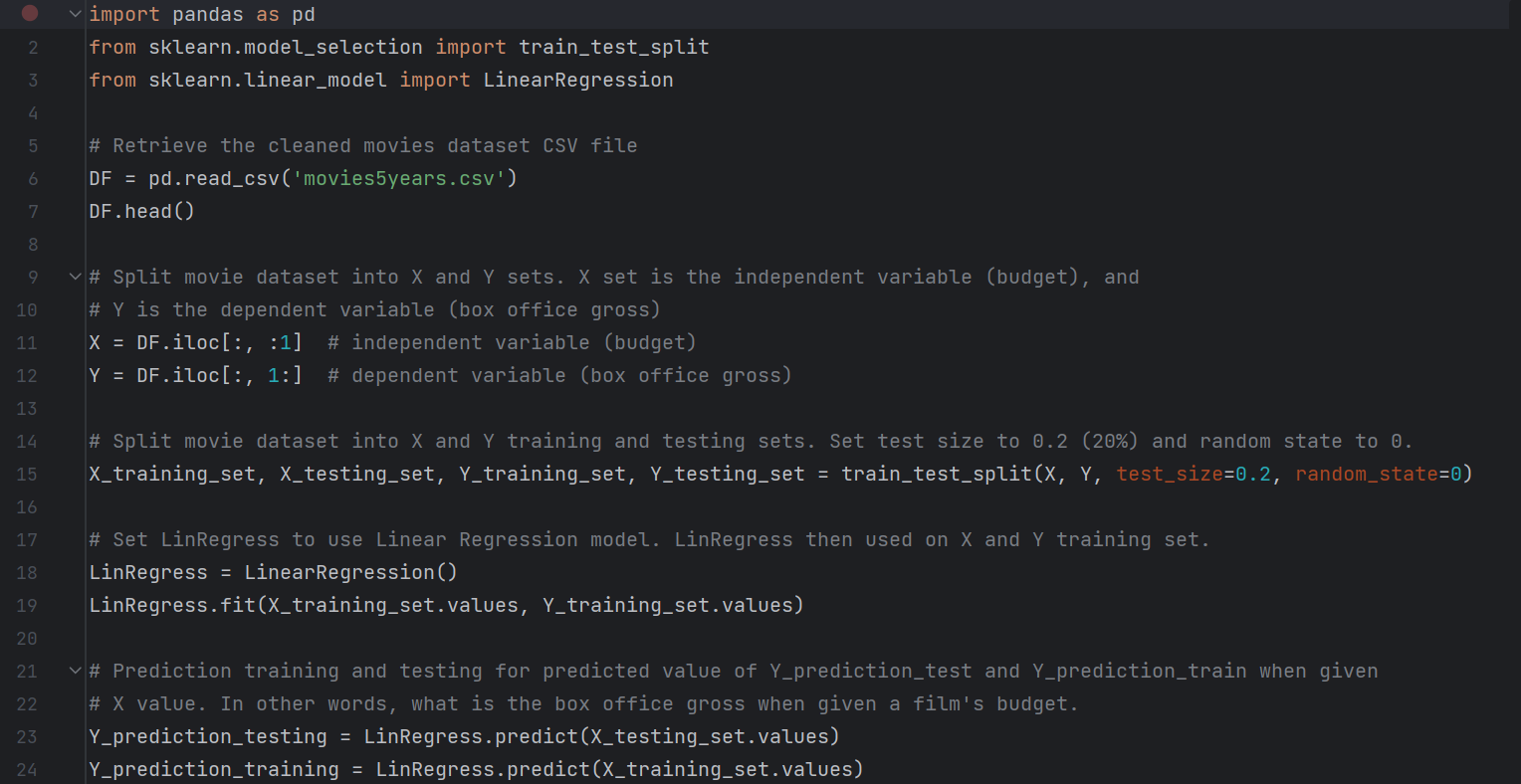
### Project purpose

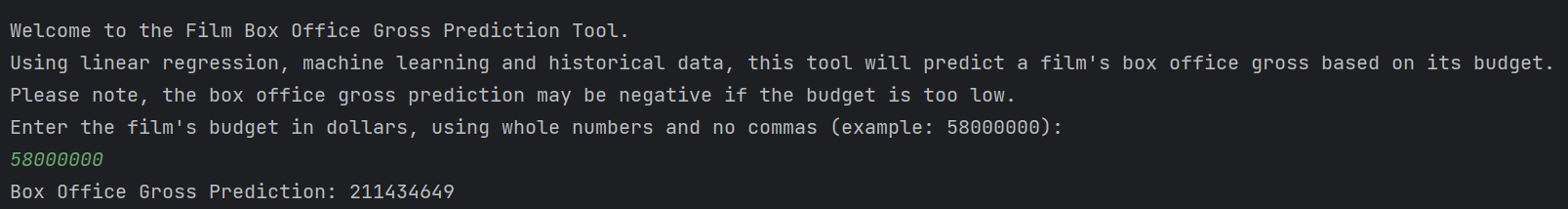
With the enormous costs associated with making a film, it was vital to be able to accurately predict whether a film would be a success or not upon release into theaters. Even one failure at the box office could have led Fred’s Film Productions to financial ruin, especially in the face of the ever-increasing popularity of streaming services leading to lower theater attendance. The purpose of this project was to create a tool that would allow Fred’s Film Productions to more accurately predict a film’s box office gross. Before the tool’s development and deployment, Fred’s Film Productions relied on studio leadership to make an “educated guess” based on their intuition, knowledge, and personal experiences, which was understood to be an inherently biased and risky process. I was able to develop and deploy an easy-to-use film box office gross prediction tool that uses a linear regression machine learning algorithm and historical film budget and box office gross data, to make more accurate, more informed, and data-driven film box office gross predictions based on a film’s budget.

### Datasets

For this project, a movie industry dataset was obtained from Kaggle.com and was used to train the film box office gross prediction tool. The movie industry dataset was in a CSV file format and contained 7,668 films, from 1980 to 2020. Each film included associated information such as the film’s title, genre, director, budget, box office gross, etc. The dataset was cleaned by first removing films that did not have budget or box office gross data. The dataset was then additionally cleaned by removing films from 2020. Films from 2020 were removed due to being considered an outlier year based on the year occurring during the COVID-19 pandemic, which severely impacted film releases and attendance. Next, after performing accuracy tests on different datasets, the dataset that included the most recent 5 years of data, excluding 2020, was selected due to being the best balance of prediction accuracy and being a large enough dataset to provide variation. Finally, to ensure optimal efficiency, all irrelevant data, such as the film’s title, director, genre, etc. was removed, and only the relevant budget and box office gross data was kept. After the cleaning was completed, the dataset was used by a linear regression machine learning algorithm to train the film box office gross prediction tool. This was performed by splitting the dataset into two sets, a training set and a testing set. The training set was used to train the tool, while the testing set was then used to test the tool’s prediction abilities. This training ensured the tool was able to make predictions on brand-new data. In the future, it is recommended to keep the dataset updated with the most recent movie data to ensure the tool’s predictive abilities remain accurate and relevant.

### Data product code





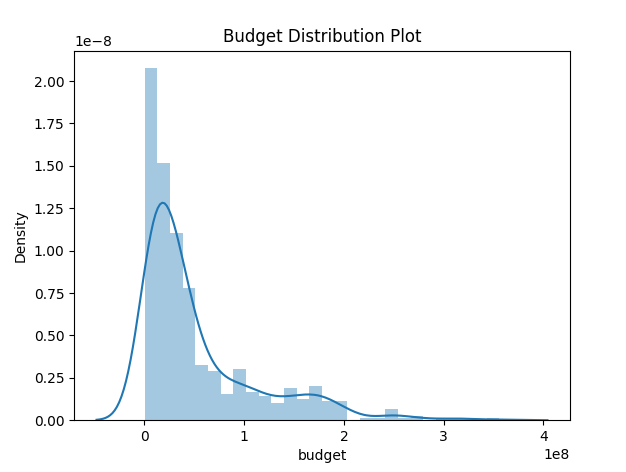
The data product code begins by importing the “pandas” library, as well as importing “train\_test\_split” and “LinearRegression” from “sklearn”. Next the cleaned movie dataset CSV file “movies5years.csv” is retrieved and set as a data frame (DF). The movie dataset is then split into X and Y sets. The X set is the independent variable, which is represented for this project as a film’s budget, while the Y set is the dependent variable, which is represented for this project as a film’s box office gross. The X and Y sets are then split into training and test sets, with the training sets representing 80% of the dataset, and the test sets representing 20% of the dataset. A Linear Regression model is used on the training sets to train the model, while the test sets are tested using the “predict” method. When the user enters a film’s budget amount, it is this trained model that is used to “predict” the film’s box office gross based on the entered budget.

Please refer to the zipped file named “C964Task2” included in this submission, to view the box office prediction tool code. Information regarding the installation and use of the tool can be found in “Appendices: Installation Guide” and “Appendices: User Guide”.

### Hypothesis verification

The project’s hypothesis was based on Fred’s Film Productions seeing an increase in its revenue and profitability after the introduction of a film box office gross prediction tool that uses machine learning to make more accurate predictions than the current method used by the studio. Upon deployment of this tool, box office gross predictions were monitored for accuracy, in which the film box office gross prediction tool, using a linear regression machine learning algorithm and historical budget and box office gross data, significantly outperformed by 60% the previously used method, in which studio leadership would make educated guesses based on their intuition, knowledge, and personal experiences. In addition, studio revenue and profitability were also tracked and saw a dramatic positive increase of 85% starting 30 days after the tool was deployed. Finally, through follow-up interviews and surveys with studio leadership, positive feedback was received regarding the tool’s functionality, and ease of use, even for non-technical users.

### Effective visualizations and reporting

**Visual 1: Budget Distribution (histogram/line)**

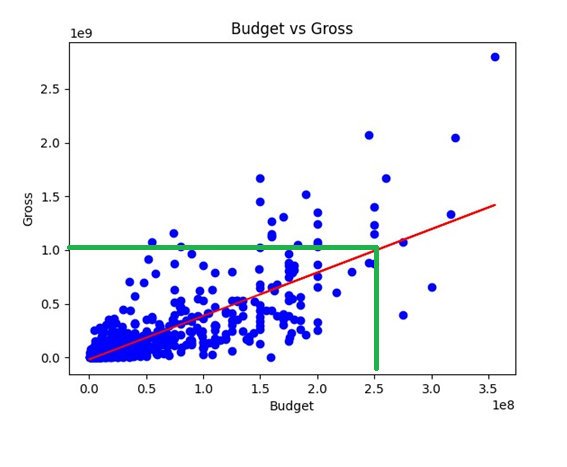
Using the “seaborn” library to create a distribution plot, Visual 1 is a histogram with a line that shows the distribution of budget amounts in the movie industry dataset used for this project. This is useful as it allows a user to see visually what budget amounts are most common. For this graph, we can see the majority of budgets fell under $100 million, with the most common budget amount being approximately $10 million.

### Visual 2: Budget versus Gross

### 

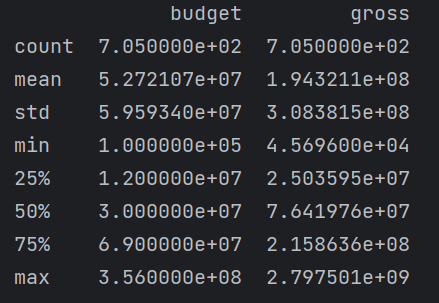
Using the “matplotlib” library to create a scatterplot, Visual 2 shows the relationship between a film’s budget and its associated box office gross. For example, we can see that the film in the database with the biggest budget had a budget of approximately $356 million and grossed approximately $2.8 billion.

### Visual 3: Budget versus Gross with Line of Best Fit



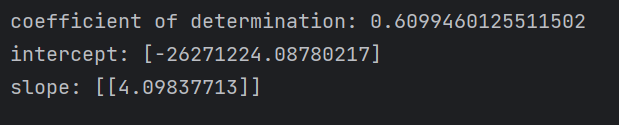
Visual 3 is the same as Visual 2, except it now includes the line of best fit (in red). The line of best fit, otherwise known as a linear regression line or simply, a regression line, is a straight line that best represents the data on the scatterplot graph. The green line visually represents how the line of best fit could be used to estimate a film’s gross by its budget. For example, the green line in Visual 3 shows how a film with a $250 million budget would be estimated to gross $1 billion.

### Visual 4: Describe Method



Using the “pandas” library, the “describe” method was used to generate various statistics about the budget and box office gross including the count (the number of non-empty values), the mean (average), the standard deviation, the minimum value, the maximum value, and the 25th, 50th, 75th percentiles. In visual 2, we were able to visually estimate the maximum budget and gross amount, which we can now confirm by viewing the “max”, which we can see as the maximum budget value is $356 million, and the maximum gross value is approximately $280 billion.

### Visual 5: Coefficient of determination, the Y-intercept, and the slope



Finally, I was also able to determine the coefficient of determination, the Y-intercept, and the slope. The coefficient of determination, or R-squared score, measures the accuracy and efficiency of the model, with a score of “1” representing 100% accuracy. The Y-intercept and slope can be used in the simple linear regression equation “Y = Mx + C”, where “M” is the slope, “C” is the Y-intercept, “X” is the independent variable (the film’s budget), and “Y” would be the dependent variable (the film’s predicted gross).

### Accuracy analysis



In regression, the coefficient of determination, or R-squared score measures how well the regression predictions approximate the real data points. An R-squared score measuring 1 indicates 100 percent of the predictions were correct, however, a high score such as that is more related to when researchers are working with items such as molecules, materials, or atoms, whose properties and behavior are known and predictable. When working in the social sciences, in which the prediction of movie grosses based on their budget would be based, an R-squared score measuring above 0.5, or 50 percent, is considered to be acceptable (2). The film box office gross prediction tool has a coefficient of determination of approximately 0.6099, or 61 percent, indicating it is acceptable. This is especially true when considering this tool uses a simple linear regression algorithm, which predicts a dependent variable, in this case, a film’s box office gross, based on a single independent variable, which in this case is represented by a film’s budget. The coefficient of determination could potentially be increased if a multiple linear regression algorithm were used, which would involve predicting a film’s box office gross based on multiple independent variables, such as budget and genre, instead of just budget.

### Application testing

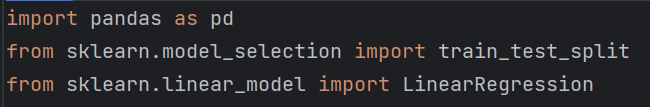
For this project, unit testing was performed throughout development to ensure the tool was performing as intended. Unit testing is when testing is performed on the individual units to ensure they are working properly, before testing the whole system together. By testing this way, any errors or issues were found and fixed sooner, than compared to waiting for the whole system to come together before beginning testing. Based on the testing performed, the film box office gross prediction tool was confirmed as working as intended and provided acceptable box office gross predictions.

# Appendices

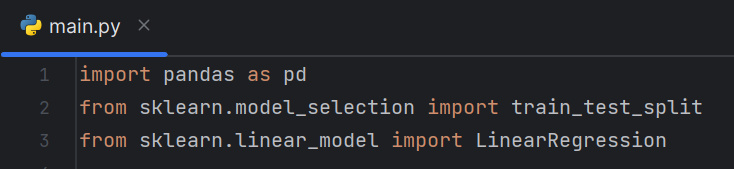
## Installation Guide

Prerequisites:

* An IDE that supports Python 3.9. For example, PyCharm Community Edition 2023.1.3 was used to develop this project and can be downloaded here: <https://www.jetbrains.com/guide/python/tutorials/getting-started-pycharm/installation-and-setup/>
* If an IDE is selected other than PyCharm, please ensure the following libraries are supported:
  + Import “pandas” as “pd”
  + From “sklearn.model\_secition” import “train\_test\_split”
  + From “sklearn.linear\_model” import “LinearRegression”



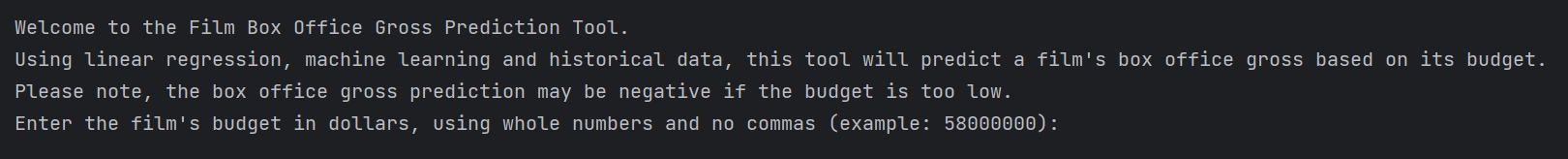
1. Please install the selected IDE. If PyCharm was the IDE selected to be used, an installation guide can be found here: <https://www.jetbrains.com/help/pycharm/installation-guide.html#standalone>
2. Please “extract all” (“unzip”) the “C964Task2” zipped folder contents included in this submission. Please take note of where the unzipped folder contents are located.
3. Using the selected IDE, please open the “C964Task2” unzipped file folder.
4. Once opened, please select “main.py”.
5. Please install the following from the first three lines of “main.py”:
   * Import “pandas” as “pd”
   * From “sklearn.model\_selection” import “train\_test\_split”
   * From “sklearn.linear\_model” import “LinearRegression”



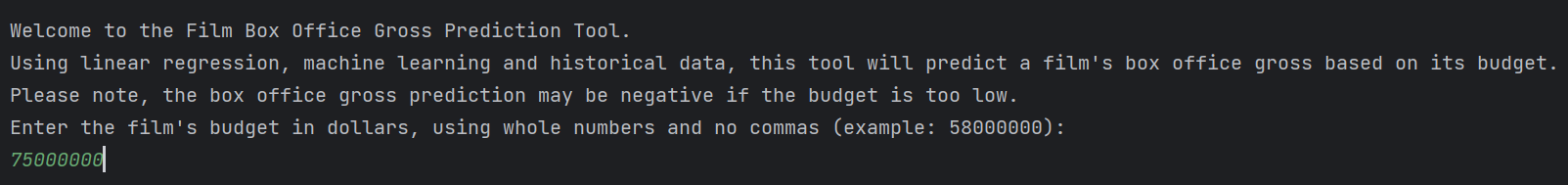
1. Congratulations! The installation steps have been completed. Please refer to the “User Guide” for the next steps.

## User Guide

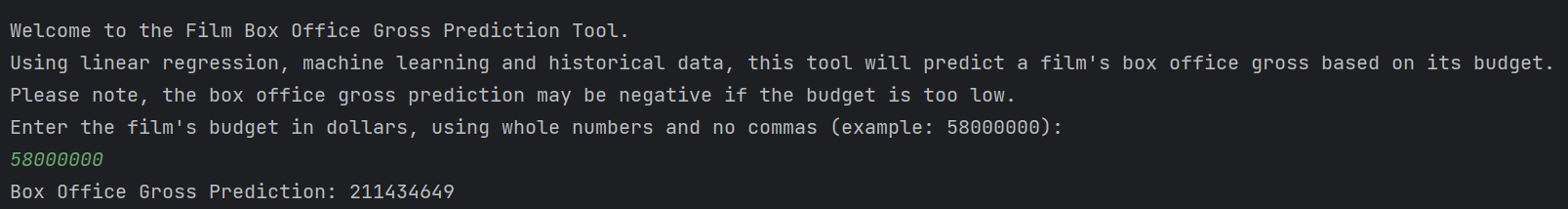
1. Please ensure all steps are completed from the “Installation Guide” first.
2. Using an IDE that supports Python 3.9, please open the “C964Task2” unzipped file folder.
3. Select and run “main.py”.
4. The user will see a “Welcome to the Film Box Office Gross Prediction Tool” message, with a short explanation of the tool’s function.



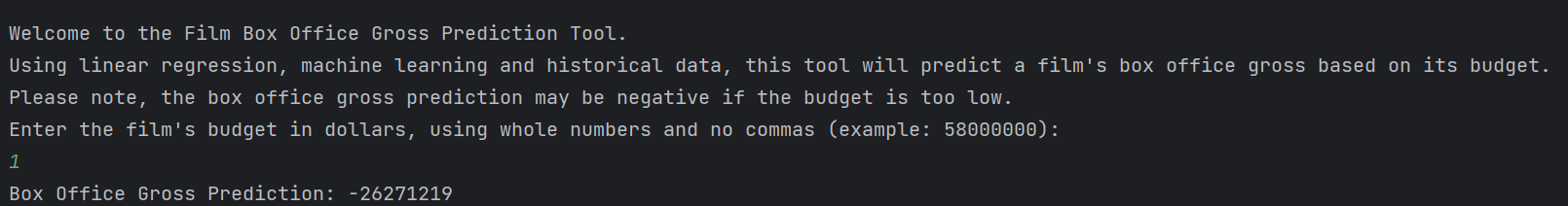
1. The user is asked to enter a film’s budget, using whole numbers and no commas. For example, $58,000,000 ($58 million), would be entered as “58000000”.



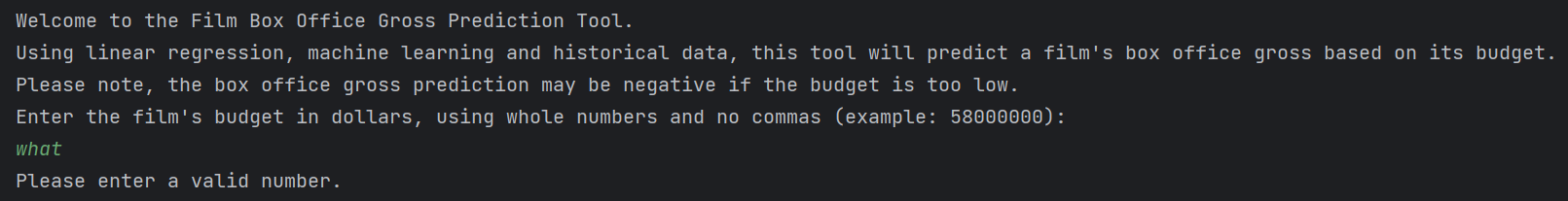
1. Once the film’s budget is entered, the tool will return the “Box Office Gross Prediction” amount. For example, as seen in the screenshot below, a film with a budget of $58 million, is expected to have a box office gross of $211,434,649.



1. Please note, that if the budget entered is too low, a negative expected box office gross will display. For example, as seen in the screenshot below, a film with a budget of $1, is expected to have a box office gross of negative $26,271,219.



1. If a user enters invalid characters such as a non-whole number or letters, they will see a “Please enter a valid number” message and the tool will start over.



## Summation of Learning Experience

This capstone project truly is a culmination of the WGU Computer Science bachelor’s program, as to complete the project, skills, and knowledge needed to be used from all aspects of the program. Examples of this include:

* How to format and propose a formal project.
* Design a use case for a tool/application.
* Use coding to develop the tool and application, including incorporating a machine learning algorithm.
* Use Project management, methodologies, and best practices to complete the project efficiently and effectively.
* How to format and complete formal end-of-project reporting.
* Independent research into using various libraries such as “sklearn” and “matplotlib” to learn more about machine learning and visualizations.

While I am proud of what I have learned so far, and how I am on the verge of graduating, I understand that there is still a lot to learn to become as proficient of a Computer Scientist as I desire to be. This is why the most important concept that I will take from the capstone project, is that even once I have completed my Computer Science degree from WGU, I can continue to develop and learn through independent research and study, using my current tools and technology. Overall, this project and the WGU Computer Science bachelor’s program have provided me with the skills and knowledge to confidently start a career in the Computer Science field.

## References

1. “Data Mining Process: Models, Process Steps & Challenges Involved.” Software Testing Help, <https://www.softwaretestinghelp.com/data-mining-process/#Data_Mining_Models>. Accessed 13 Dec 2023.
2. “The acceptable R-square in empirical modeling for social science research.” Munich Personal RePEc Archive, <https://mpra.ub.uni-muenchen.de/115769/1/MPRA_paper_115769.pdf>. Accessed 15 Dec 2023.