Project proposal

Addis Ababa institute of technology – Center of Information Technology and Scientific Computing

Big Data Modeling and Management System

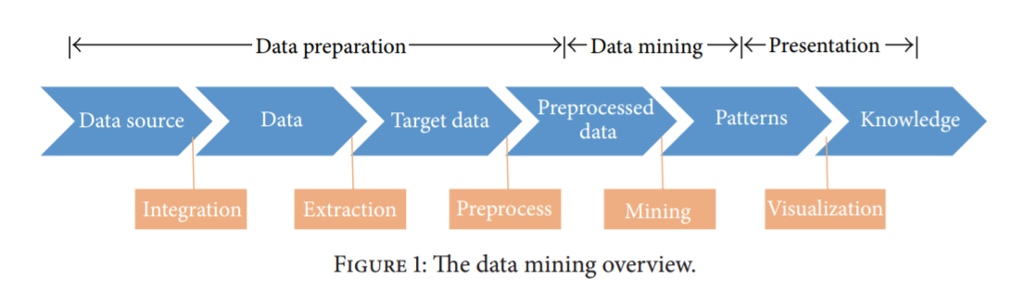
Group Members:

1. Absalat Dawit ATR/8266/08
2. Dagnachew Tesfaye ATR/4262/08
3. Leykun Yilma ATR/4380/08
4. Nahom Bekele ATR/2892/08

# **ABSTRACT**

The success rate of a movie depends on people’s personal perception and opinions. In real world prediction models and mechanisms can be used to predict the success of a movie. This proposed work aims to develop a system based upon data mining techniques that may help in predicting the success of a movie in our society in advance thereby reducing certain level of uncertainty. In this project, we apply data mining technique and machine learning algorithms using R software to predict the success and failure of movie based on several attributes. This project also shows the power of predictive and prescriptive data analytics for information systems to aid movie business decisions. This model also helps to find out the review of the new movie.

1. **Business Understanding**
   1. **Introduction**

Movies are one of the biggest entertainment sectors in the world. A movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics’ review, ratings for the movie, release year of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much revenue a particular movie will be generating. However, by analyzing the revenues generated by previous movies in our country, a model can be built which can help us predict the expected revenue for a particular upcoming movie. Such a prediction could be very useful for the movie studios which will be producing the movie so they can decide on different expenses like artist compensations, advertising of the movie, promotions in various cities, etc. accordingly. Plus, it allows investors to predict an expected return-on-investment (ROI). Also, it will be useful for many movie theatres to estimate the revenues they would generate from screening a particular movie. To expend this business further we need the technology through which we can predict the success rate of the movie. If we were able to predict the movie success rate in the correct manner then it will be easy for the businessman to get higher profit from it and also if the prediction shows the success rate is low of certain movie then it helps those businessmen to improve the content of the movie so that they can get higher revenue from it. success rate of movies, models and mechanisms can be used to predict the success of a movie. It will help the business significantly. stakeholders such as actors, producers, directors etc. can use these predictions to make more informed decisions. They can make the decision before the movie release. This proposed work aims to develop a model based upon the data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. The excellent way to find detailed information about almost every film ever made is through IMDb. Vast amount of data, which contains much valuable information about general trends in films. Data mining techniques enable us to uncover information which will both confirm or disprove common assumptions about movies, and also allow us to predict the success of a future film given select information about the film before its release. So here we are developing the software for data analytics through which we can predict the success rate of the movie which high accuracy. We will follow this step to 

**1.2 Objective**

The Objective of the project is to build a predicting model. This model will use the data that we will collect to predict if a movie will be a “hit”, “flop” or “average” based on it features. The features that we will use is genre, actor, actress, director, rating and attendance.

The success or failure of a movie is not necessarily dependent only on the factors that we will be considering for this project. For that reason, the outcome of the analysis could certainly be off by some degree because of these parameters. Considering the unpredictability of the success of a given movie, this project can be considered successful if it provides a successful prediction for the majority of the time.

* 1. **Data Mining goal**

This project had as a main goal to develop a model, able of predicting the box office financial success of a certain set of movies through specific variables and historical data. It was possible to conclude that the percentage of success of the cinematographic revenue prediction is quite different based on the typology of the dependent variable used in the study.

The primary software that we will be using to manipulate and test our data will be RStudio, which is a free and open-source IDE for the R programming language and we will use python to run the k means and linear regression.

1. **Data Understanding**

**2.1 Collect Data**

We will be collecting data from local cinemas. This data will contain features such as the genre of movies these people prefer, the actresses/actors they like, the possibility of social media affecting their choices and so on. We might also consider information from movie rating sites such as IMDb.

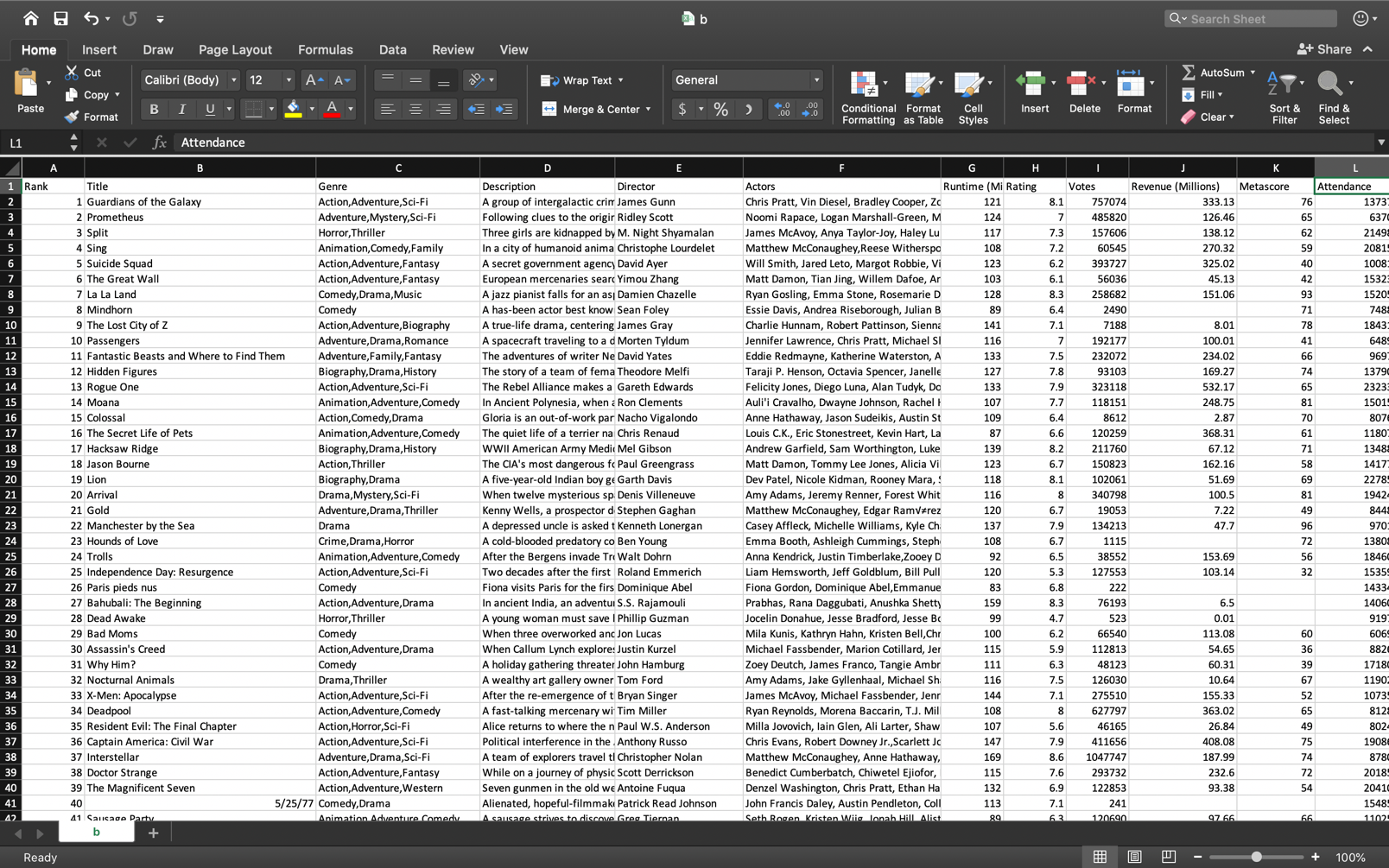
**2.2 Describe Data**

Our data initially has 12 columns and 1000 rows. The columns are:

* Rank
* Title
* Genre
* Description
* Director
* Actors
* Runtime (Minutes)
* Rating
* Votes
* Revenue
* MetaScore
* Attendance

**2.3 Data Format**

The data we initially collected looks like:



*Figure 1: Data initially collected.*

There are a lot of factors that can affect the success of a movie, Like gender of a person and age of a person, but since the data we obtained doesn’t show that, we will discard it.

The data obtained from the collection stage is of mixed format. Some of the fields contain strings such as names of actors and actresses, while some contain numeric replacement for strings (for example, genre where it is substituted by a finite set of integers), etc. We will focus on the five attributes which we think will help us determine the success of a movie efficiently. The attributes are:

|  |  |  |
| --- | --- | --- |
| Variable | Description | Data Type |
| Actor/Actress | Name of actress/actor in a movie | Character |
| Genre | Genre of the movie | Character |
| Director | Name of Director of the movie | Character. |
| Rating | Rating of the movie | Num |
| Attendance | Number of people who watched the movie | Num |

**2.4 Verify Data Quality**

We will verify that the data we got is correct and includes every attribute we want before we move to the next step. The proposed methodology deals with different stages of the project which consists of data collection, data preprocessing, generating training and testing dataset, model generation, prediction and outcomes. These stages will lead us into building the system we aim to predict the success of an upcoming movie in our society.

1. **Data Preparation**
   1. **Data Selection**

We will be collecting data from local cinemas. The data we will collecting from them will be features like, what genre of movies they like, which actress/actor they like, if social media affects their movie choices and so on. We will also consider information from movie rating sites like IMDb. We will select the data that we think consist the information we need to the fullest.

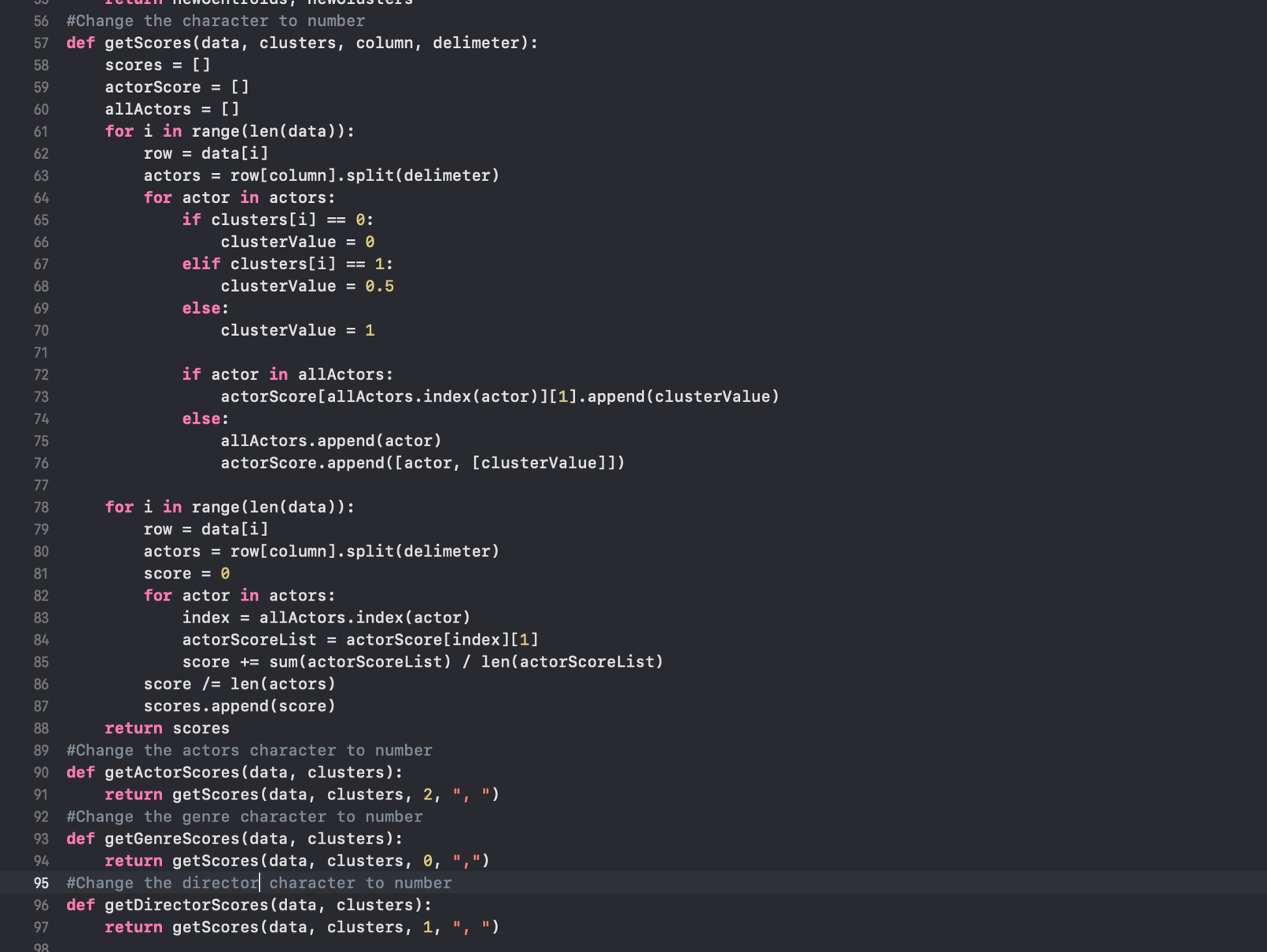


* 1. **Data Cleaning**

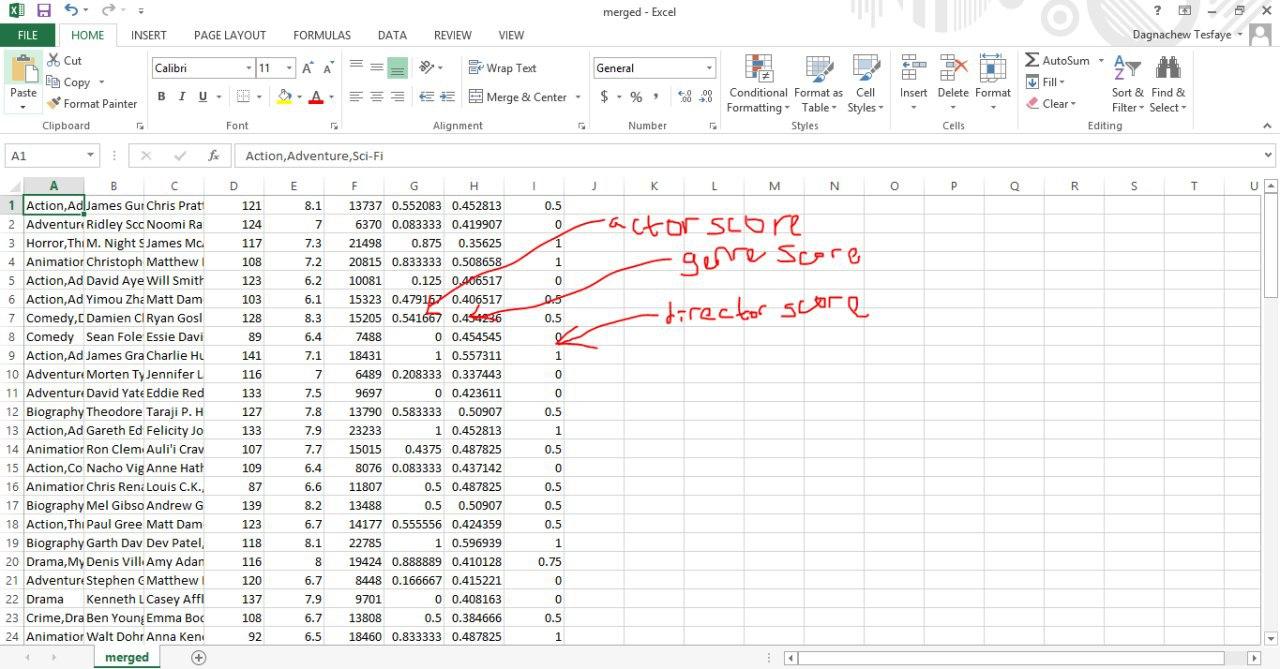
At this stage we will perform pre-processing methods like data cleaning, where we get rid of any information we don’t need for this project, and data transformation, where we save our data in a CVS or excel form to help us manipulate it easily later on. We will also remove data consist missing information. And construct the data into a format we want.

**3.3 Integrate Data**

Since we have some data that is character, we had to change it into integer. We do this by:

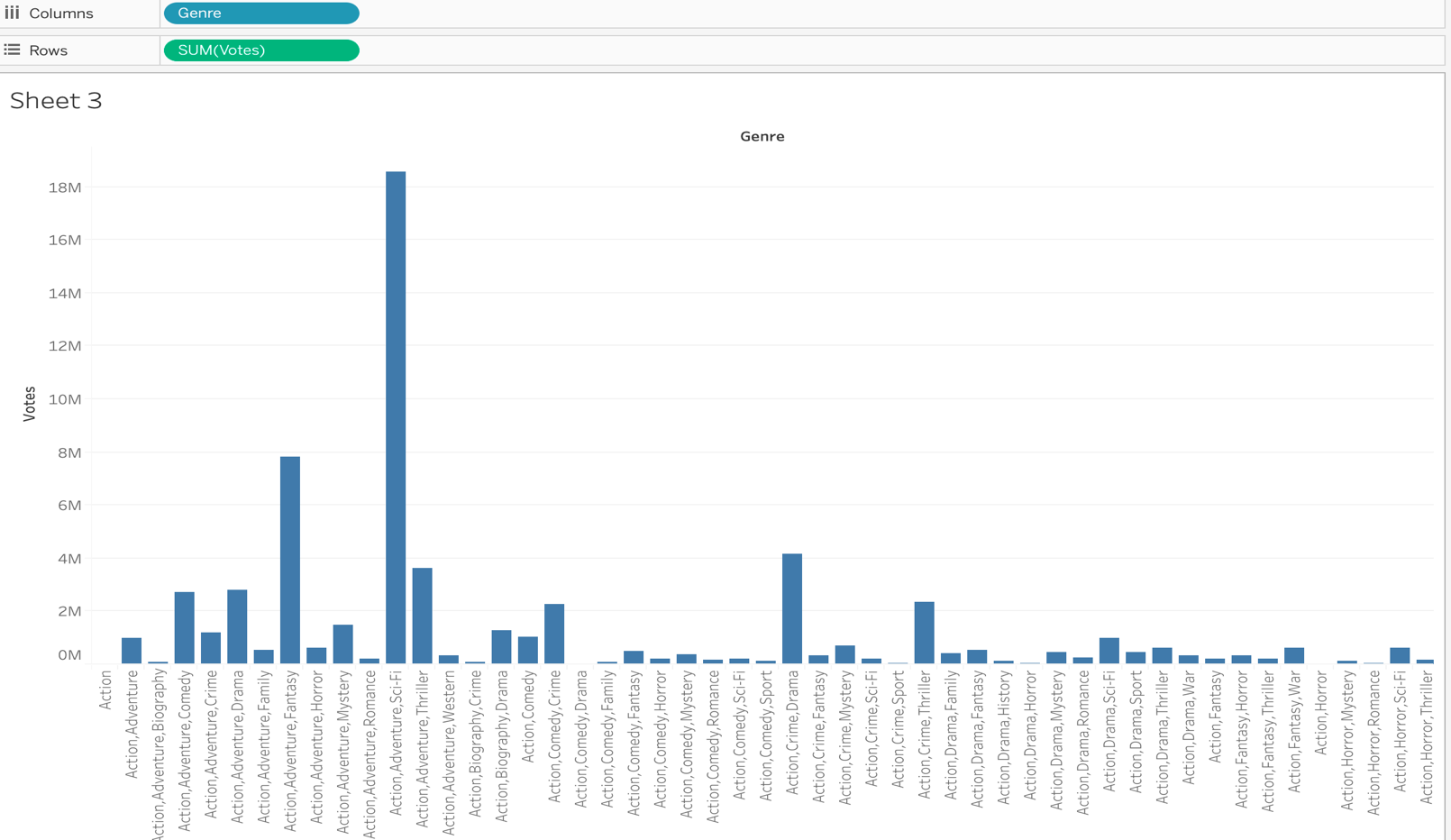
****

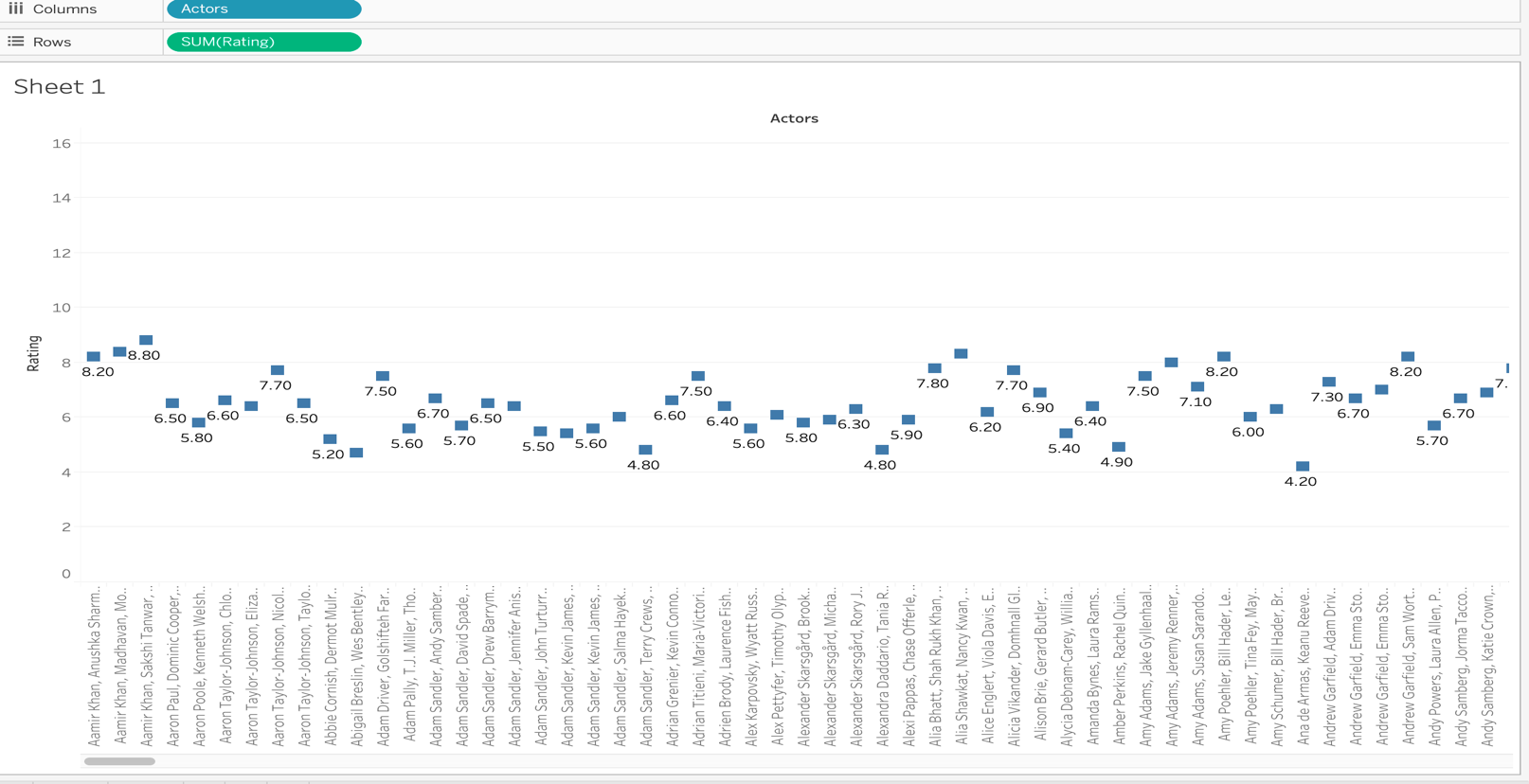
After that we get a data whose character has been changed to number:



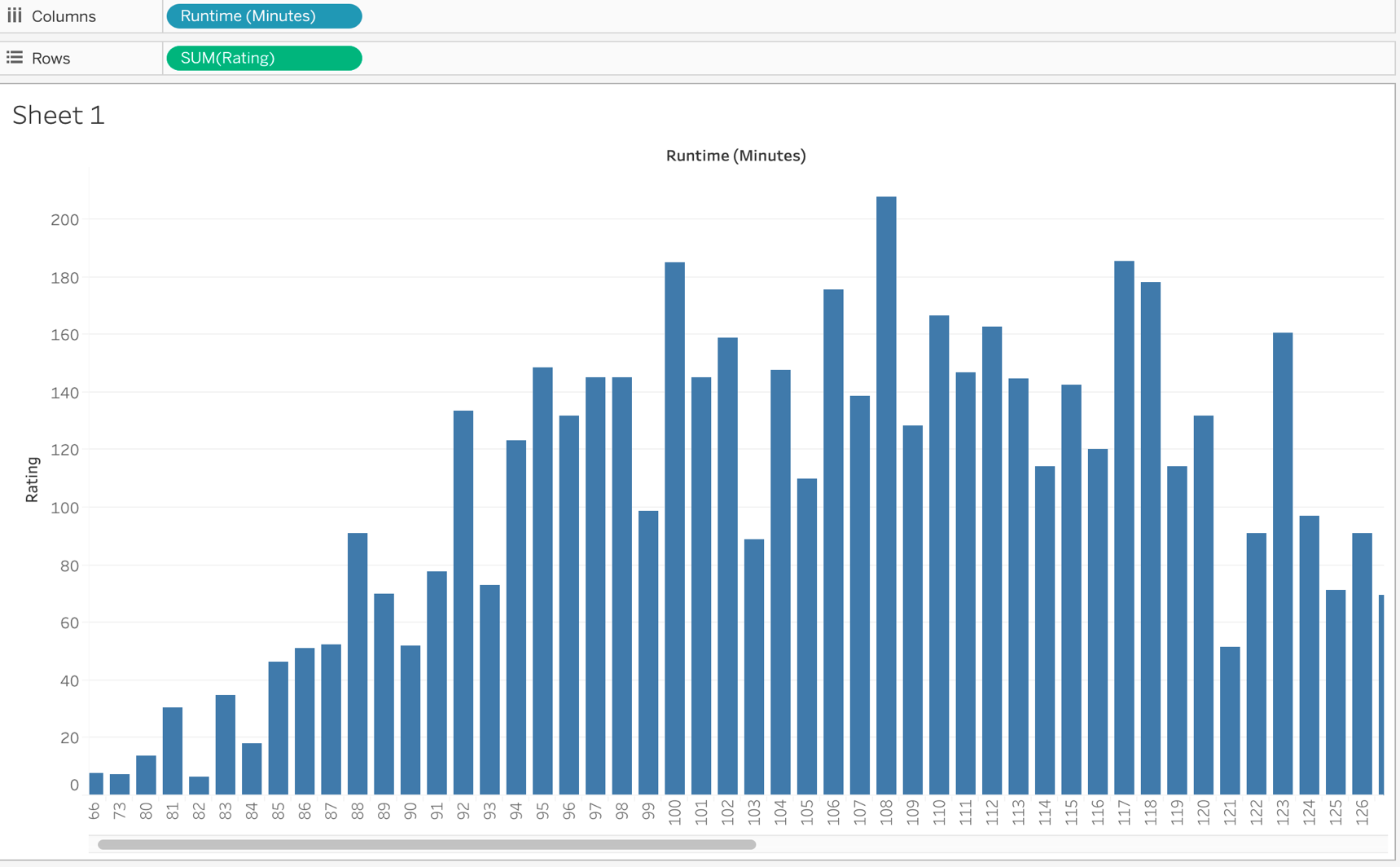
* 1. **Data Analysis and Formatting Data**

At this stage we will explore our data and make sure we have the data that we want for our project. Then we will also verify that the data we have currently is the data we need**.** Those datasets will then bedivided into training dataset and test dataset which contains the classes like Hit, Flop and Average and predicting variables like actor, actress, composer, genre, director producer and music director to help us with the prediction stage that comes after.









1. **Modeling** 
   1. **Selection of Modeling Technique**

We will be using clustering model based on the attendance of the movie goers. K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

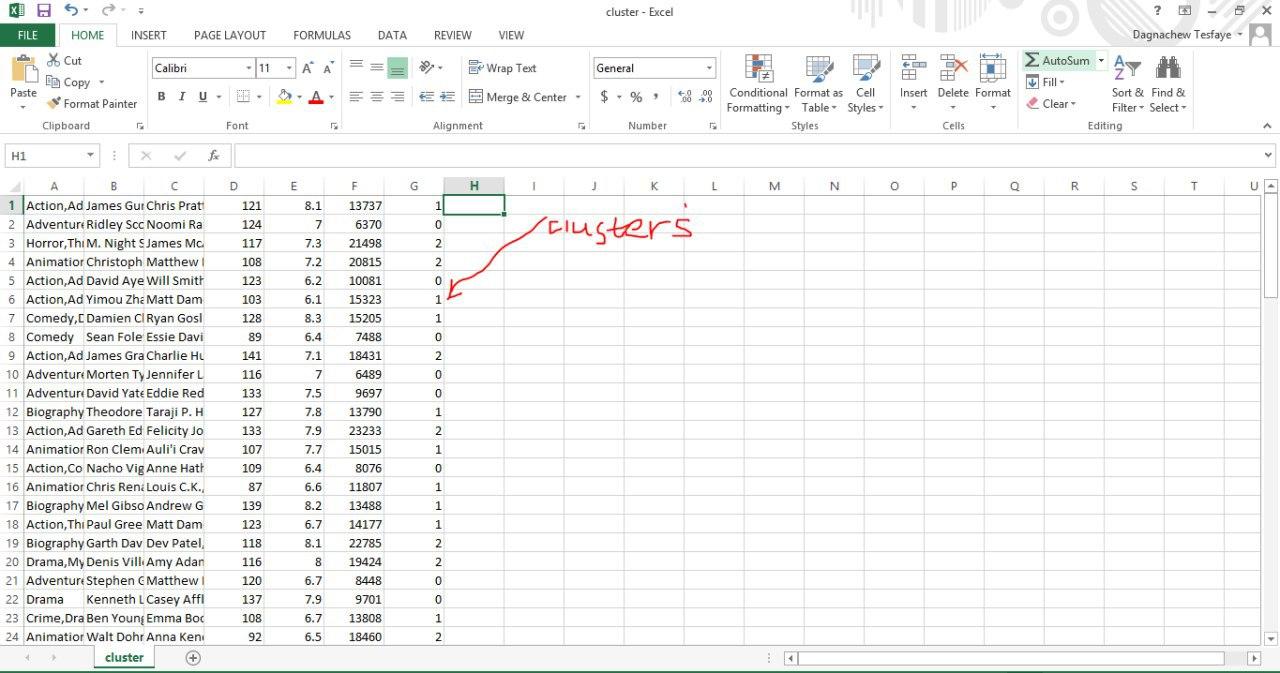
1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

We will divide them into three clusters. “Hit”, “Flop” or “Average”. We implemented the K means clustering as follows:



The output was will be



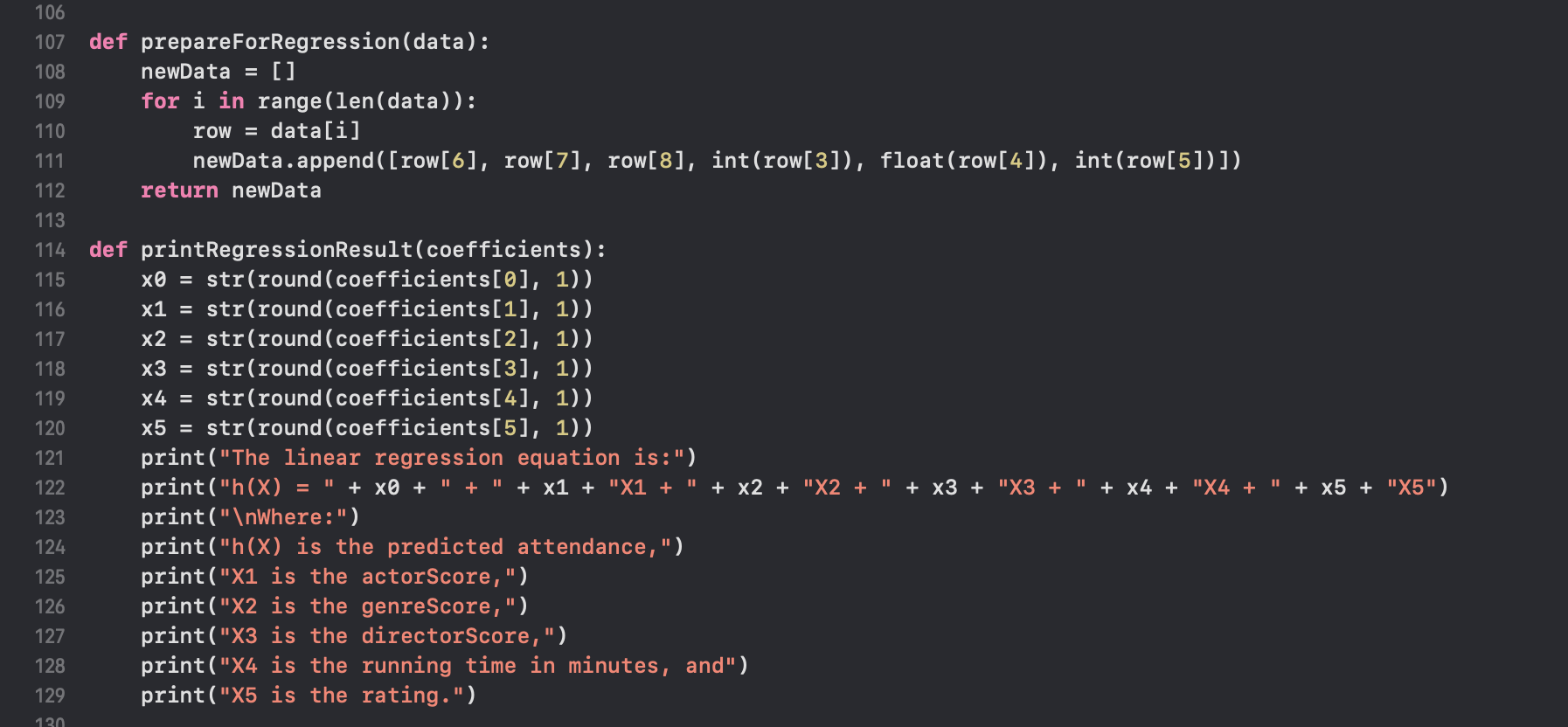
* 1. **Generate testing and training data**

Training dataset is a set of attributes used to fit the parameters of the model. The model like naive Bayes classifier is trained on the training dataset using supervised learning method like gradient descent or stochastic gradient descent. In practice, the training dataset often consist of pairs of an input vector and the corresponding output vector (or scalar), which is commonly denoted as the target. The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation. Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. When the data in the test dataset has never been used in training like cross-validation, the test dataset is also called a holdout dataset. We will divide the data we have into training and testing data. We will use the training to train our model and we will test them using the testing data.

**4.3 Model Creation and assessment**

We will use linear regression to create our prediction model. Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

To train out dataset we used the following code:



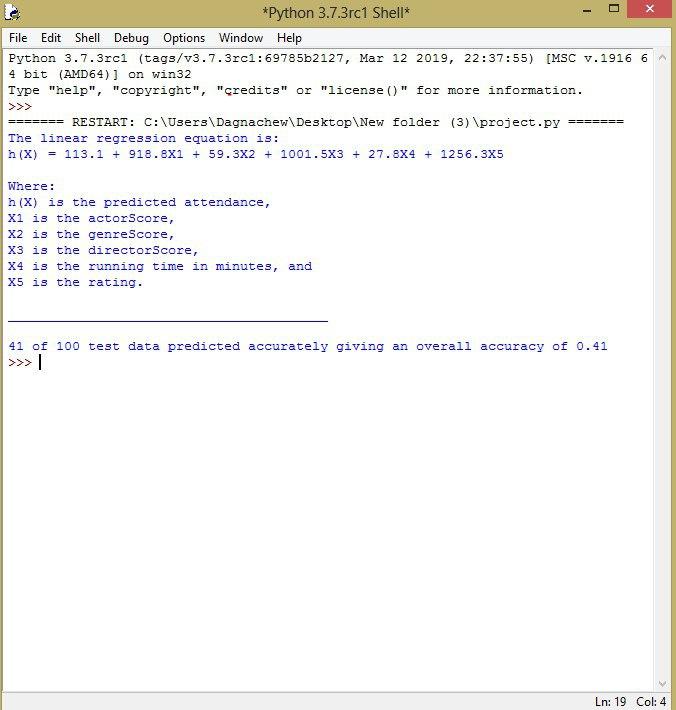
Next what we did was we wrote a code to test our dataset:



1. **Evaluation**

**5.1 Evaluate DM process**

We will evaluate our working by measuring the accuracy and residual error.After training our dataset and testing it we have found out a 45% accuracy. Meaning we predict if the movie is a success accurately 45% of the time. Below you can see the result:

****

1. **Deployment**

**6.1. Monitoring and Maintenance**

The Monitoring Maintenance Lifecycle (MML) is a monitoring development process to reduce maintenance costs and increase reliability of IT infrastructure concerning service recovery related problems. It is based on the classical Waterfall Model.

Monitoring Maintenance Lifecycle are methods and standards for improving and mastering maintenance [processes](https://en.wikipedia.org/wiki/Industrial_process), supporting processes and management processes throughout the monitoring lifecycle.

As we go on, we will try to increase our accuracy with a more precise dataset. Working to get a better outcome is our goal.

**Conclusion**

In this project we are just trying to determine if there is any association between different attributes present in our dataset. Here, our main aim is to find association between different attributes that is used a scoring system and how we can use this association for prediction. As a result, we found that attributes like actors/actress, director, genre and rating is strongly positive relationship between rating and genre. And we can also conclude that rating is best predictor of audience scores. Thus, we can predict our movies success on the basis of ratting. In future, we can add many attributes as our predictors and build model for that attributes to perform prediction. Here, we can assume that if we have movie gross score and movie net profit along with movie manufacturing cost, then we can build a stronger model for movie success prediction. In future, we can apply other machine learning algorithms for movie success prediction.

# **REFERENCE**

1. *Kumar .S, Metha .A, Pal .J, Movie Success Prediction using Data Mining For Data Mining and Business Intelligence(ITA5007) of Master of Computer Application, 2019*
2. *Javaria Ahmad,Prakash Duraisamy,Amr Yousef,Bill Buckles, “Movie Success Prediction Using Data Mining”, in 8th International Conference On Computing, Communication And Networking Technologies (ICCCNT),3 -5 July, 2017.*
3. *Antara Upadhyay, 2Nivedita Kamath, 3Shalin Shanghavi, 4Tanisha Mandvikar, 5Pranali Wagh, Movie Success Prediction Using Data Mining, 2018*
4. *K Meenakshi , G Maragatham , Neha Agarwal and Ishitha Ghosh, A Data mining Technique for Analyzing and Predicting the success of Movie, 2018*