SHRI VAISHNAV INSTITUTE OF TECHNOLOGY AND SCIENCE, INDORE

DATA SCIENCE

with Python

" GAMING ARCADE "

TRAINING PROJECT REPORT

Submitted By:

SAMRADHI GOYAL

17100BTBDAI01639

Computer Science & Engineering

In Bachelor of Technology

July – December 2020

SVVV, Indore

A REPORT OF TWENTY ONE DAYS INDUSTRIAL TRAINING

AT

WebTek Labs Pvt. Ltd.

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE AWARD

OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE & ENGINEERING

July – December 2020

Submitted by:

SAMRADHI GOYAL

17100BTBDAI01639

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

SVVV, INDORE

JULY-DECEMBER 2020

CANDIDATES' DECLARATION

We hereby declare that we have undertaken industrial training at "WEBTEK LABS PVT. LTD." during a period from 27th July to 16th August in partial fulfillment of requirements for the award of the degree of B.Tech (COMPUTER SCIENCE & ENGINEERING) at SHRI VAISHNAV INSTITUTE OF TECHNOLOGY AND SCIENCE, INDORE. The work which is being presented in the training report submitted to the Department of COMPUTER SCIENCE & ENGINEERING at SHRI VAISHNAV INSTITUTE OF TECHNOLOGY AND SCIENCE, INDORE is an authentic record of training work.

Samradhi Goyal

(Name) (Signature)

ACKNOWLEDGEMENT

We are grateful to several people for their advice and support during the time of completing our project work. First and foremost, our regards go to **Ms. Mousita Dhar**, the mentor of our project for providing us with valuable support and necessary help whenever required and also helping us explore new technologies with the help of her technical expertise. Her direction, supervision and constructive criticism were indeed the source of inspiration for us.

We forward our sincere gratitude to all **teaching and non-teaching staff** of the Computer Science & Engineering department, SVVV Indore for providing us with the necessary information and their kind co-operation, and the opportunity to undergo the training.

We would like to thank our parents and family members, our classmates and our friends for their motivation and their valuable suggestions during the project. Last, but not the least, we thank all those people, who have helped us directly or indirectly in accomplishing this work. It has been a privilege to study at SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE.

CERTIFICATE OF APPROVAL

The project "GAMING ARCADE" made by the efforts of POOJA KHANDEKAR is hereby approved as a creditable study for the Bachelor of Technology in COMPUTER SCIENCE & ENGINEERING and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

_____

Ms. Mousita Dhar

(Project In-charge)

1.1. PYTHON

1.1.1. About Python:

- Python is a high-level, general-purpose, open-source, strongly-typed programming language. The language provides constructs intended to enable clear programs on both a small and large scale.
- Python was created By Guido van Rossum.
- The Python Software Foundation (PSF) is the organization behind Python.

1.1.2. Python versions:

- First released in 1991.
- Python 2.0 was released on 16 October 2000
- Python 3.0 was released on 3 December 2008

1.1.3. Current Versions:

- 3.6.3
- 2.7.14

1.1.4. Python features:

Some of the features of Python include:

• Easy to understand

• Dynamic

• Object-oriented

• Multipurpose

• Strongly typed

• Open Sourced

1.1.5. Python usage Domains:

• Data Analysis

• Machine Learning

• Internet Of Things

• GUI Development

• Image processing

• Data visualization

• Game Development

1.2. IDLE

IDLE is an integrated development environment for Python, which has been bundled with the default implementation of the language.

1.2.1. Anaconda:

Anaconda is an open-source Distribution for data science and machine learning using python. It includes hundreds of popular data science packages and the conda package and virtual environment manager for Windows, Linux, and macOS. Conda makes it quick and easy to install, run, and upgrade complex data science and machine learning environments like scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Service Machine Learning AMIs and Anaconda for Microsoft on Azure and Windows.

1.3. Packages

1.3.1. Numpy:

NumPy is the fundamental package for scientific computing with Python.

It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

1.3.2. Pandas:

Pandas is an open-source, BSD-licensed library providing high performance, easy-to-use data structures and data analysis tools for the Python programming language. The Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables and time series.

1.3.3. Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

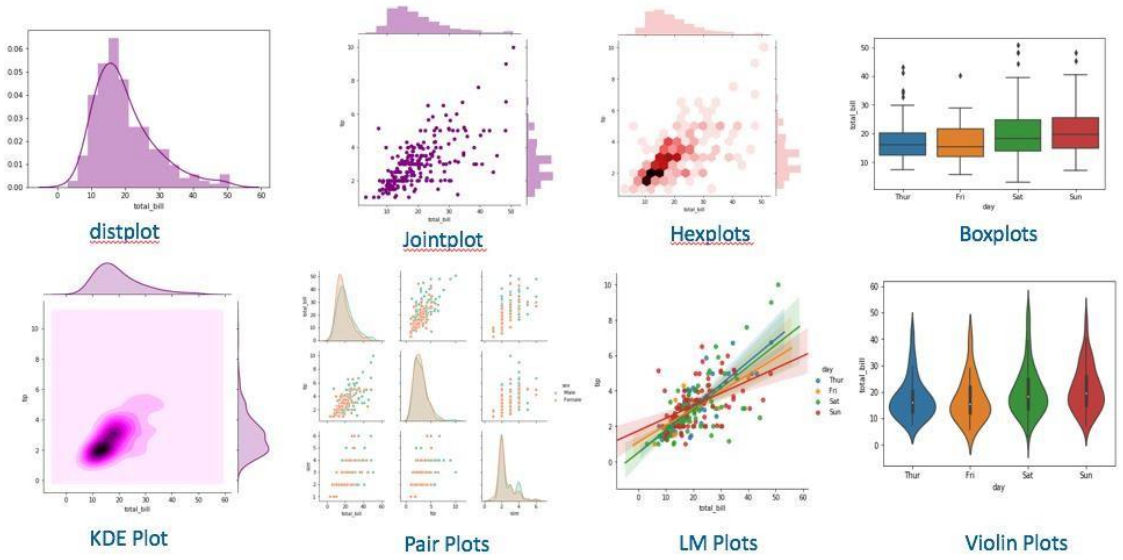


Figure 1: Example of Graphs offered by seaborn

1.3.4. Matplotlib:

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

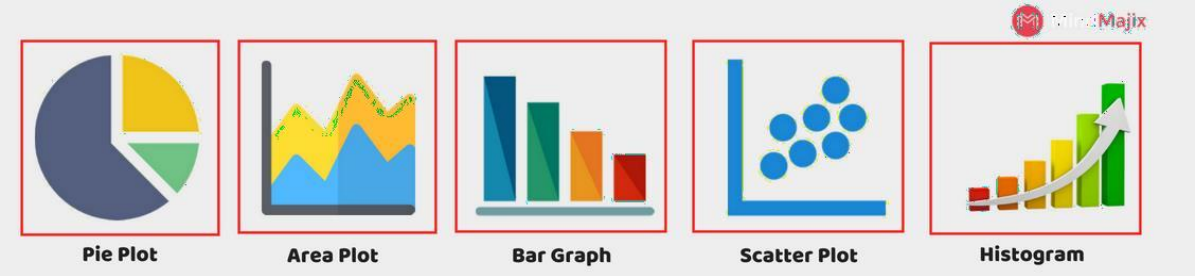


Figure 2: Example of Graphs offered by Matplotlib

1.3.5. SciKitlearn:

Scikit-learn provides machine learning libraries for python. Some of the features of Scikit-learn includes:

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib

1.3.6. Pygame:

Pygame is a Python wrapper module for the SDL multimedia library. It contains python functions and classes that will allow the programmer to use SDL's support for playing CD-ROMs, audio and video output, keyboard, mouse, and joystick input.

1.3.7. Django:

Django is a high-level Python Web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of Web development, so the programmer can focus on writing their app without needing to reinvent the wheel. It's free and open source.

1.3.8. Keras:

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, R, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

1.3.9. Tensorflow:

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.

1.3.10. OpenCV:

OpenCV is a library of programming functions mainly aimed at real-time computer vision. Originally developed by Intel, it was later supported by Willow Garage then Itseez. The library is cross-platform and free for use under the open-source BSD license.

2.1. COLLECTING DATA FROM KAGGLE

Kaggle is a platform for predictive modeling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowdsourcing approach relies on the fact that countless strategies can be applied to any predictive modeling task and it is impossible to know beforehand which technique or analyst will be most effective. On 8 March 2017, Google announced that it was acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyze cell images and identify nuclei.

2.2 DATA SCIENCE

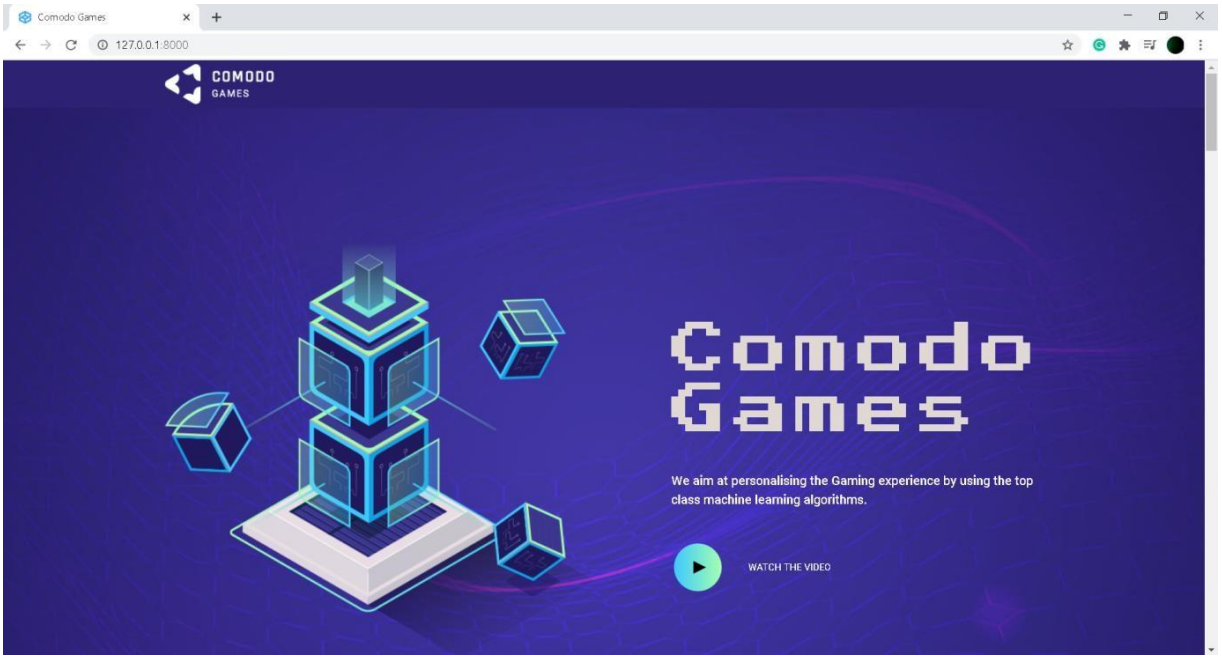
Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data in various forms, both, structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning, and their related methods" to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner JiGray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational, and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modeling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition of suitable curriculum contents. Because of the current popularity of this term, there are many "advocacy efforts" surrounding the field. To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

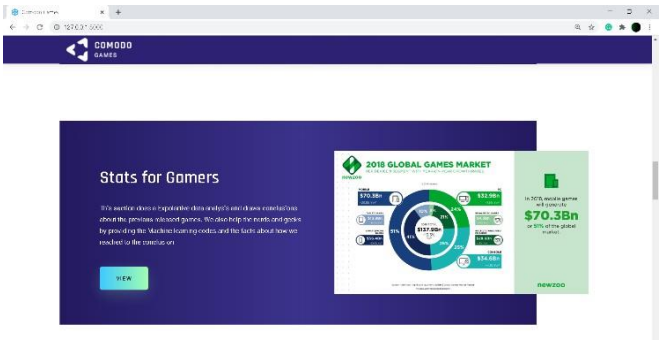
3. SOURCE CODE & OUTPUT

The project was divided into 3 modules, namely: Rock Paper and Scissors, Hangman, and Exploratory Data Analysis of Past Gaming Data.

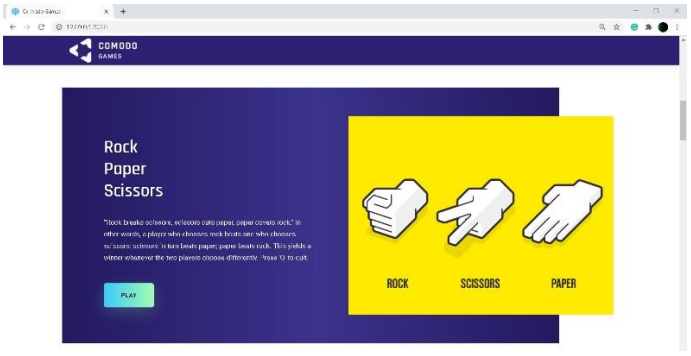
3.1. Front End:



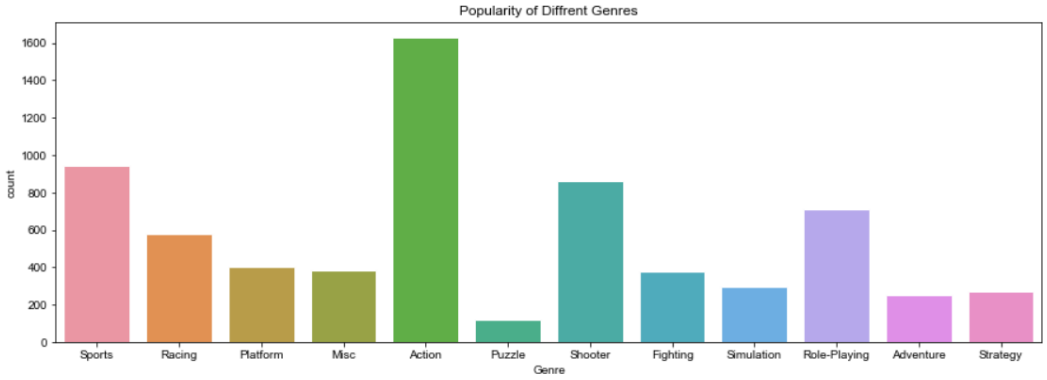
Home screen of the gaming website



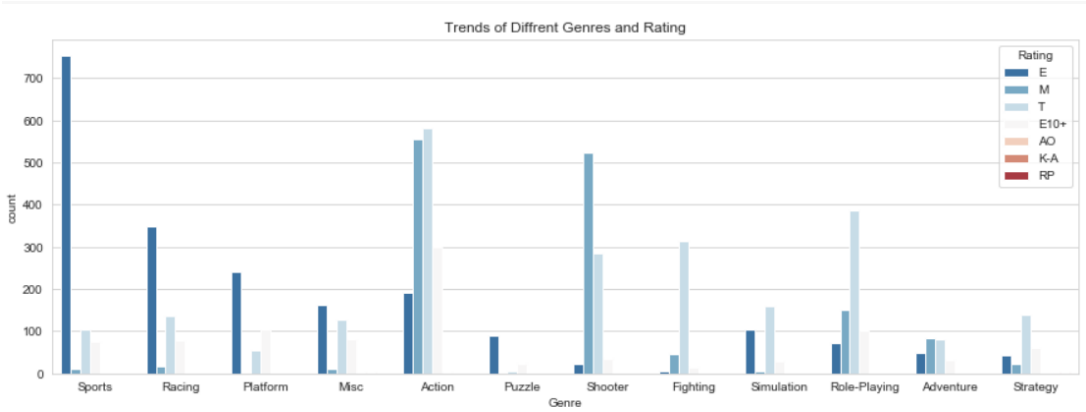
User options on the website



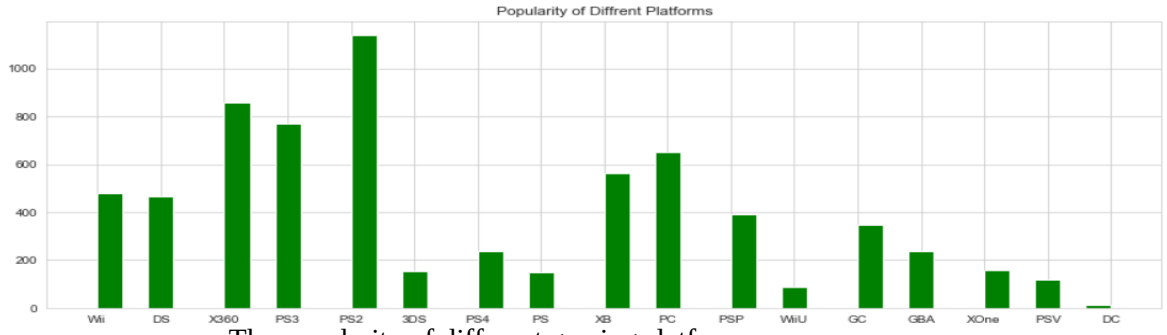
3.2. Exploratory Data Analysis



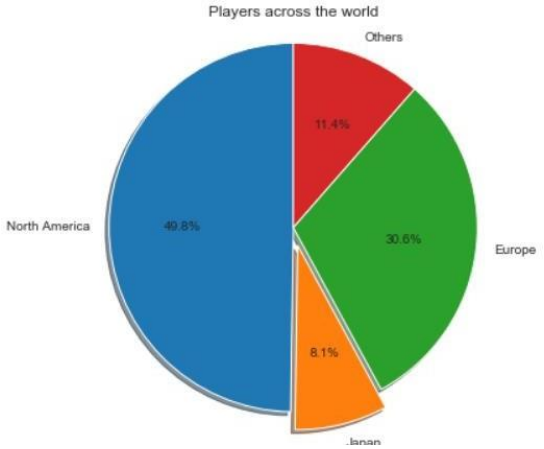
The popularity of different gaming genre



Trends of different genres vs rating

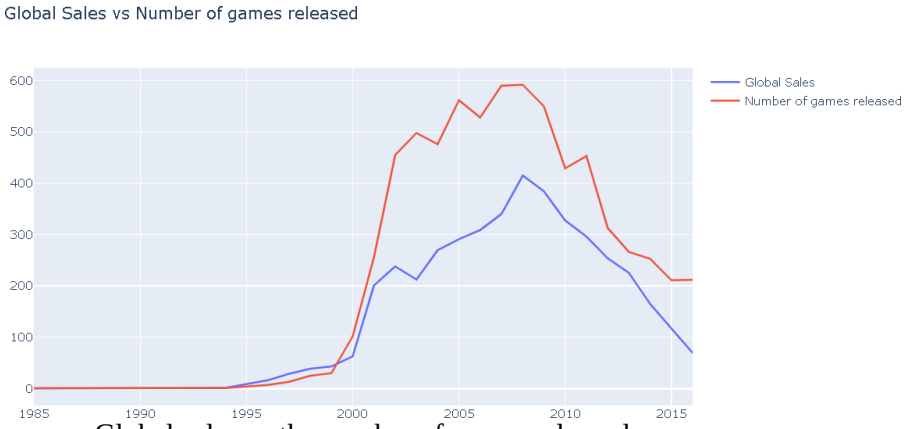


The popularity of different gaming platforms

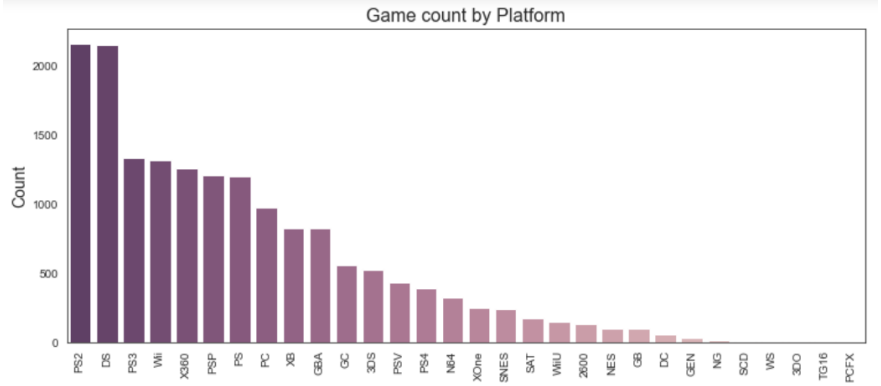


Players across the world

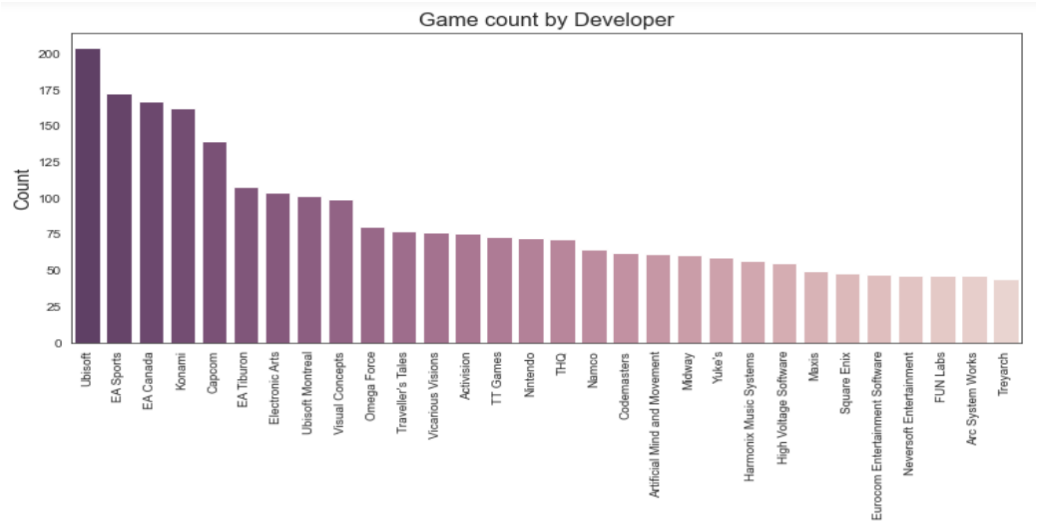
Sales of Different Genres with respect to the platform



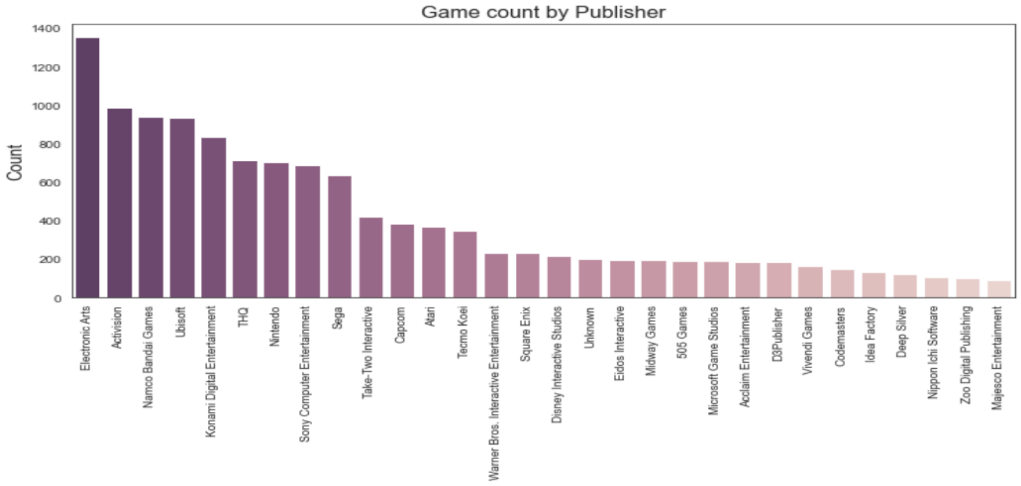
Global sales vs the number of games released



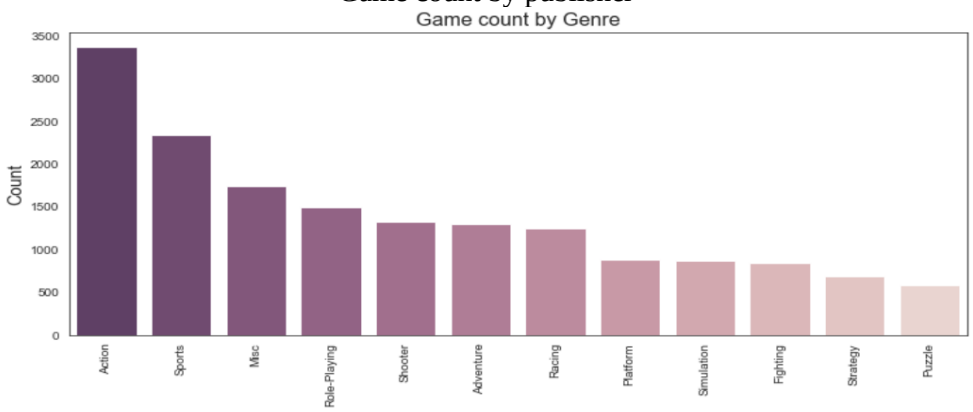
Game count by Platform



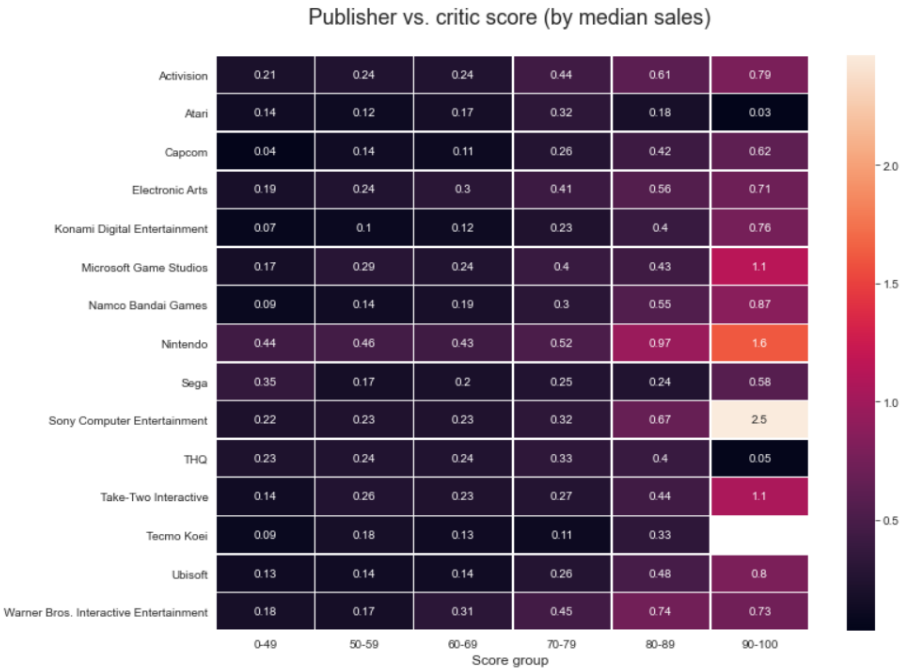
Game count by developer



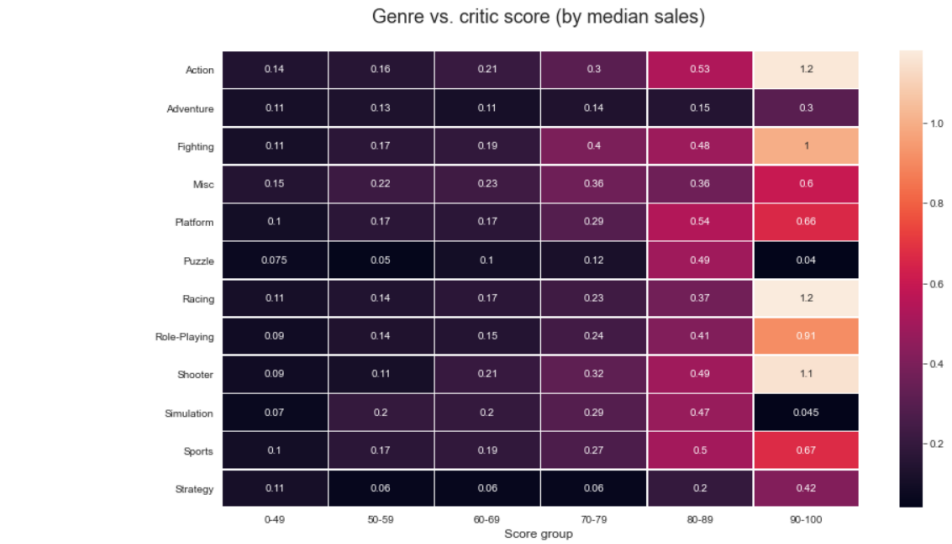
Game count by publisher



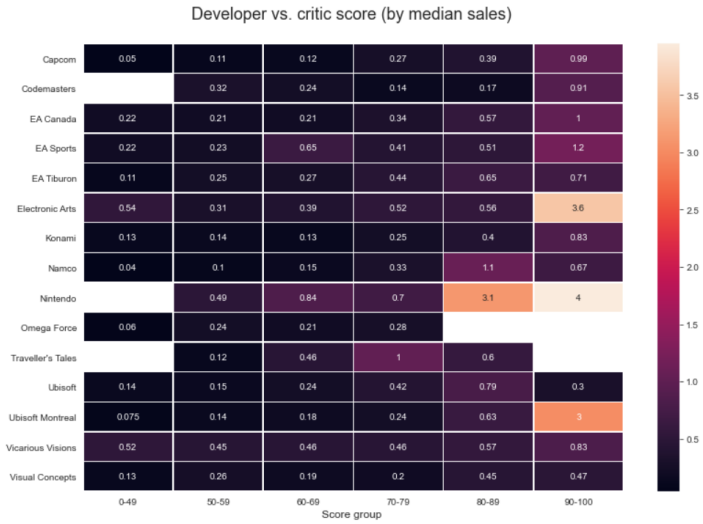
Game count by genre



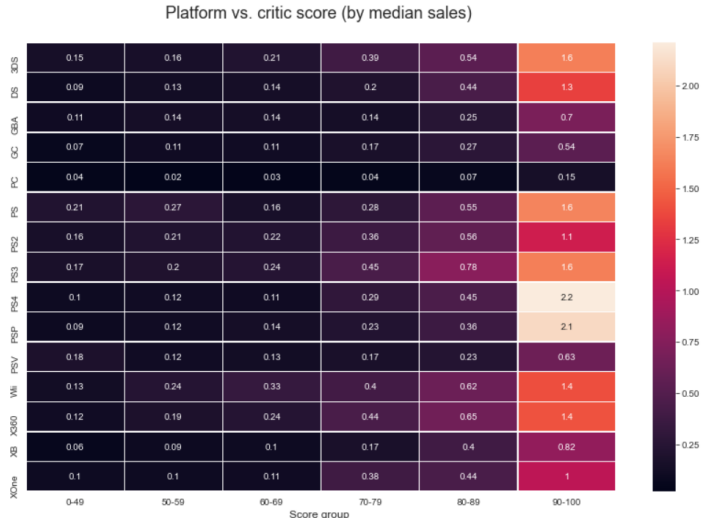
Publisher vs critic Score



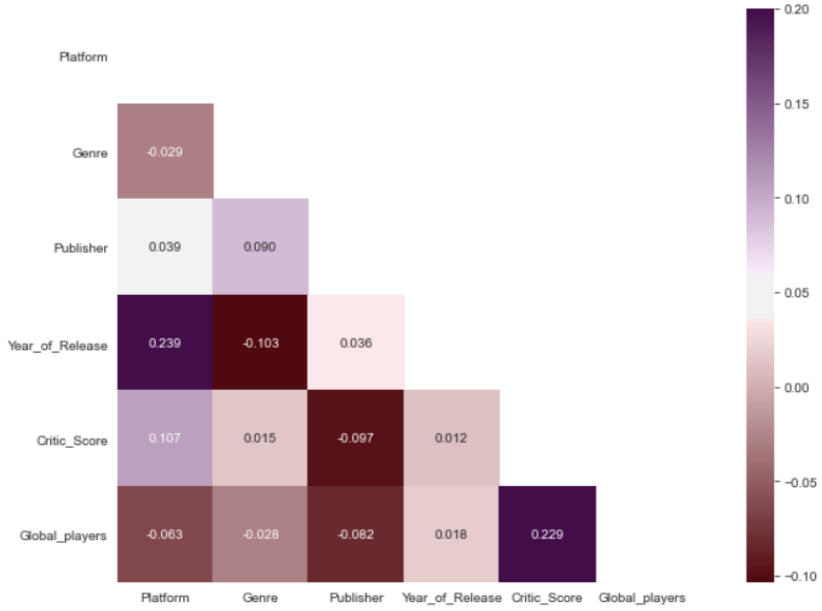
Genre vs Critic Score



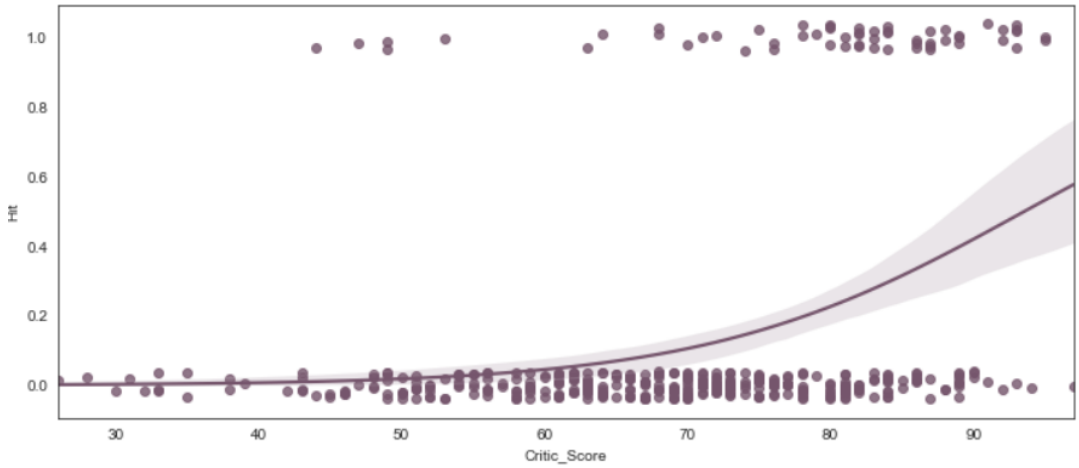
Developer vs critic score



Platform vs critic score



Dataset Parameters relations



Hits vs critic scores

In [44]:
```
radm = RandomForestClassifier(random_state=2).fit(Xtrain, ytrain)
y_val_1 = radm.predict_proba(Xtest)
print("Validation accuracy: ", sum(pd.DataFrame(y_val_1).idxmax(axis=1).values
                                   == ytest)/len(ytest))
```

```
Validation accuracy:  0.8692057128539213
```

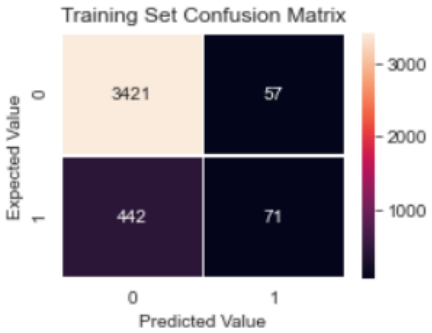
In [45]:
```
log_reg = LogisticRegression().fit(Xtrain, ytrain)
y_val_2 = log_reg.predict_proba(Xtest)
print("Validation accuracy: ", sum(pd.DataFrame(y_val_2).idxmax(axis=1).values
                                   == ytest)/len(ytest))
```

```
Validation accuracy:  0.8749686795289401
```

Validation Accuracies for different models

	precision	recall	f1-score	support
0	0.89	0.98	0.93	3478
1	0.55	0.14	0.22	513
accuracy			0.87	3991
macro avg	0.72	0.56	0.58	3991
weighted avg	0.84	0.87	0.84	3991

Classification report



Training set confusion matrix

Out[57]:

	Name	Platform	Hit_Probability
0	Titanfall 2	PS4	0.621781
1	Overwatch	PS4	0.618340
2	Dishonored 2	PS4	0.613515
3	Battlefield 1	PS4	0.601963
4	Fast Racing Neo	WiiU	0.596052
5	Kirby: Planet Robobot	3DS	0.571116
6	Dishonored 2	XOne	0.490415
7	Battlefield 1	XOne	0.457534
8	Titanfall 2	XOne	0.457534
9	Plants vs. Zombies: Garden Warfare 2	PS4	0.457451

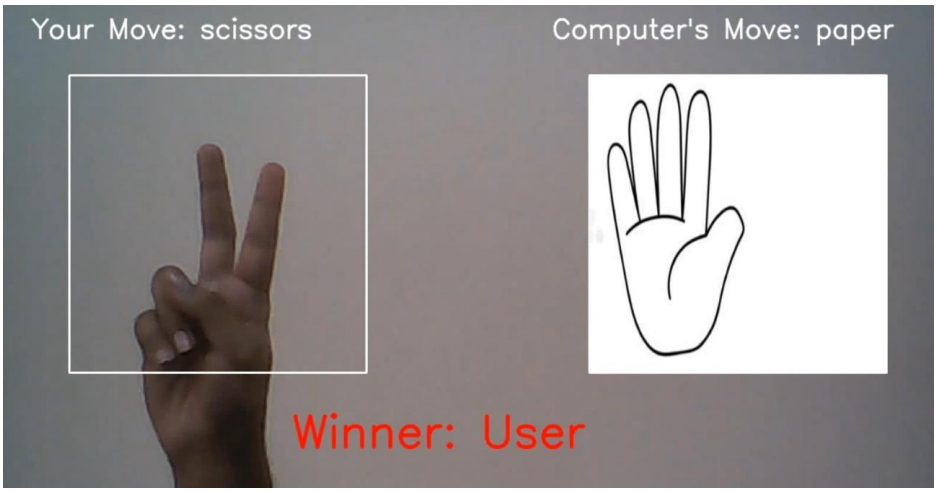
Top 10 games with highest probability of being hits according to model

Out[59]:

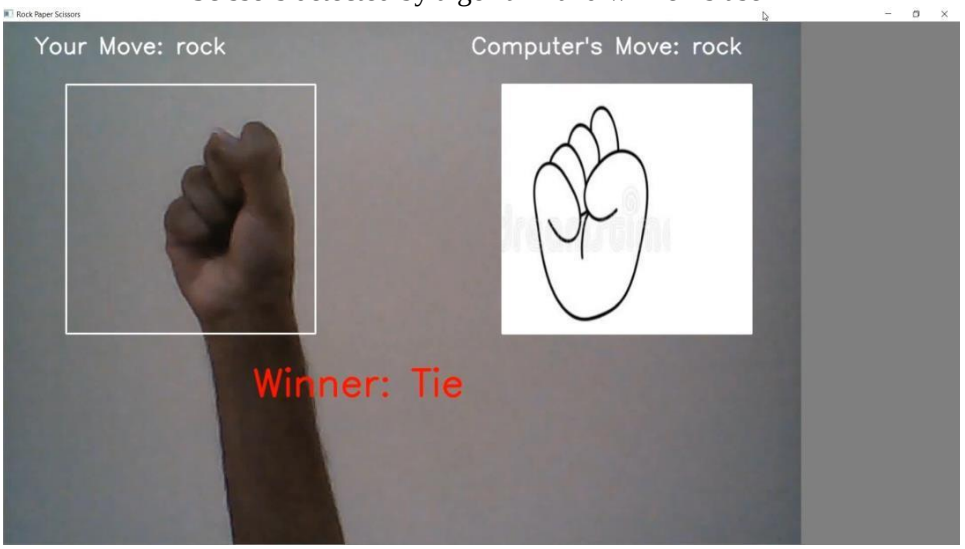
	Name	Platform	Hit_Probability
0	RollerCoaster Tycoon World	PC	0.001646
1	Bus Simulator 16	PC	0.002060
2	The Technomancer	PC	0.002304
3	Homefront: The Revolution	PC	0.002714
4	Pro Evolution Soccer 2017	PC	0.002915
5	Dino Dini's Kick Off Revival	PS4	0.003020
6	Pro Cycling Manager 2016	PC	0.003239
7	Sherlock Holmes: The Devil's Daughter	PC	0.003676
8	Agatha Christie: The ABC Murders	PC	0.003887
9	Dead or Alive Xtreme 3: Fortune	PS4	0.005849

Top 10 games with highest probability of being flops according to model

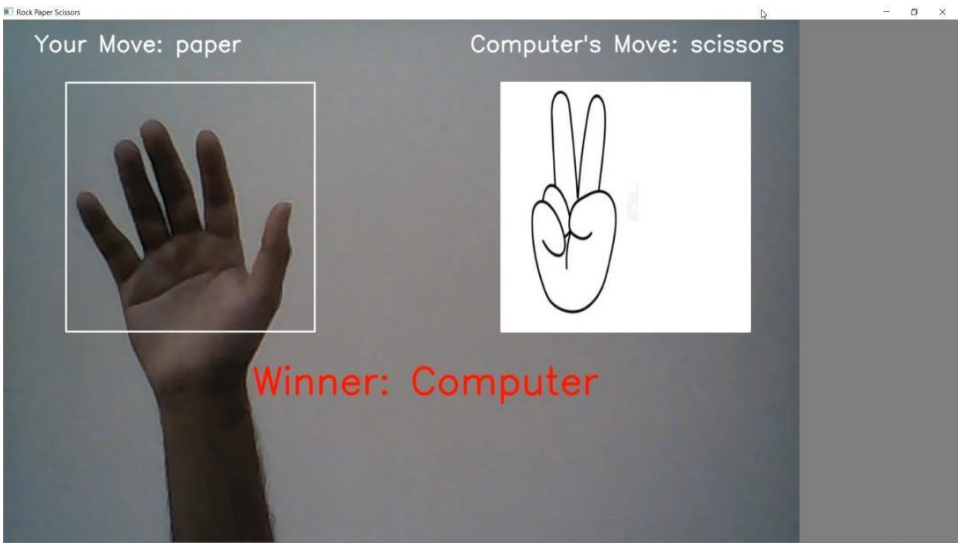
3.3. Rock Paper Scissors:



Scissors detected by algorithm and winner is user

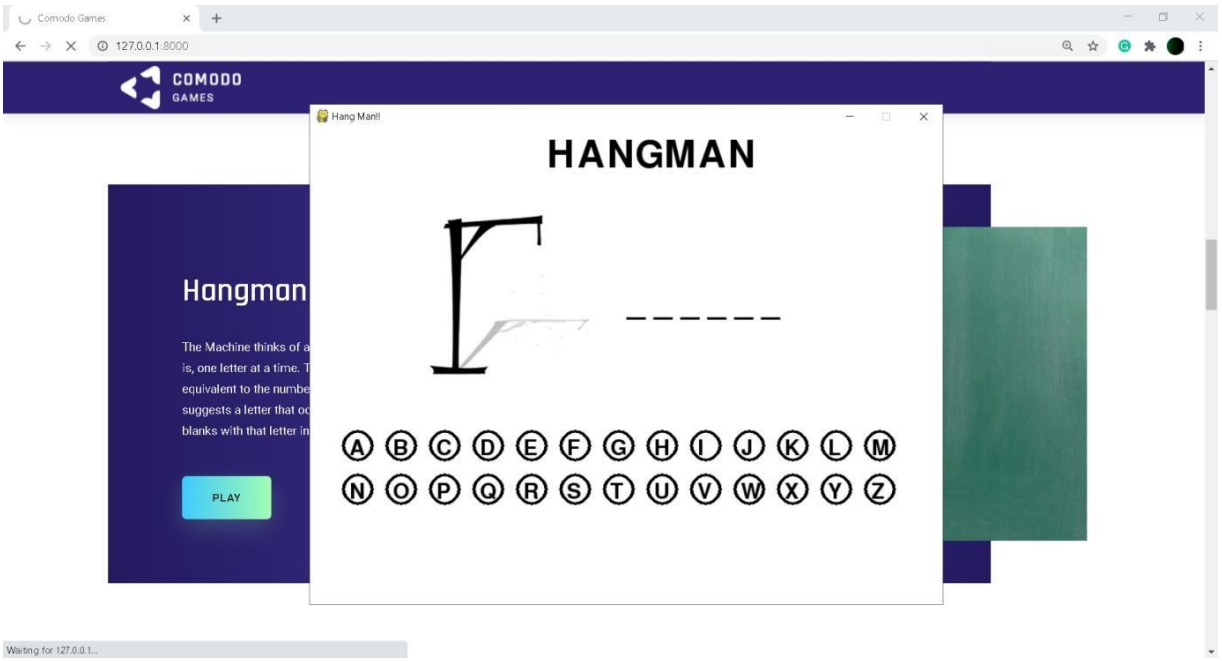


Rock detected by algorithm and it’s a tie

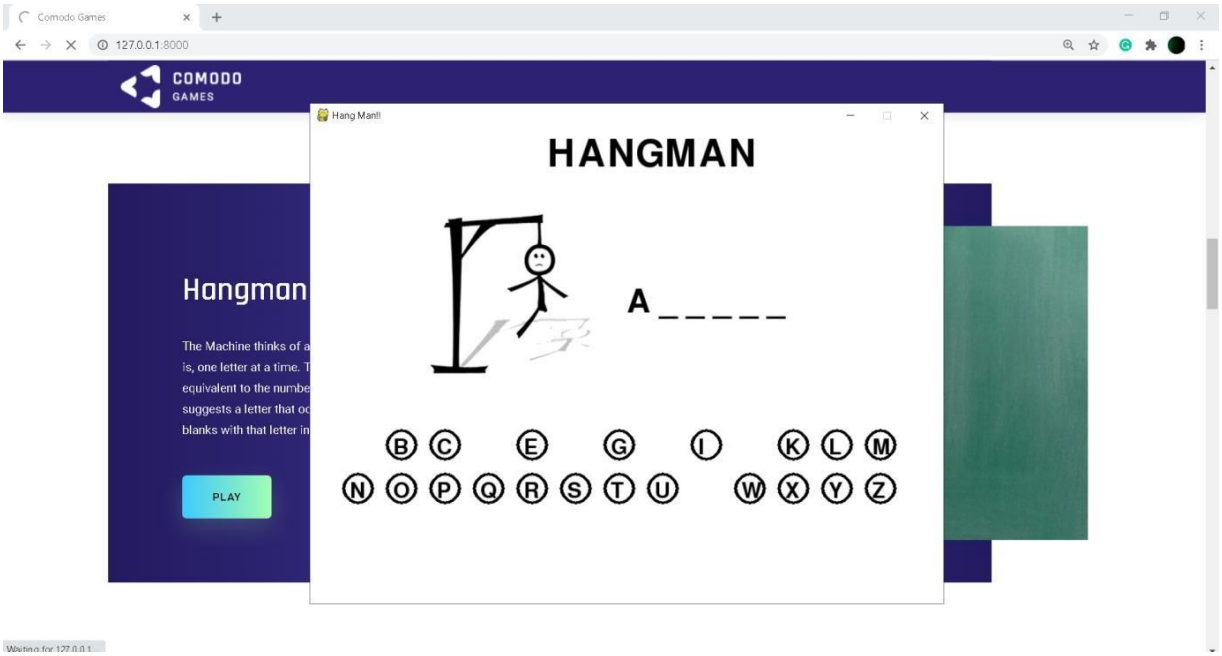


Paper detected by algorithm and winner is the computer (algorithm)

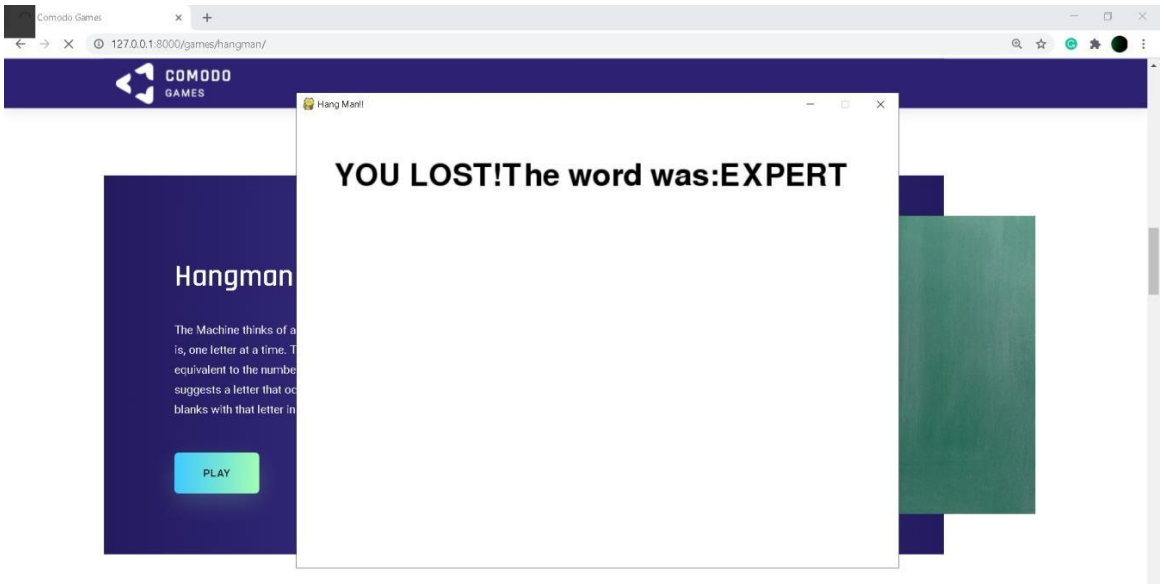
3.4. Hangman



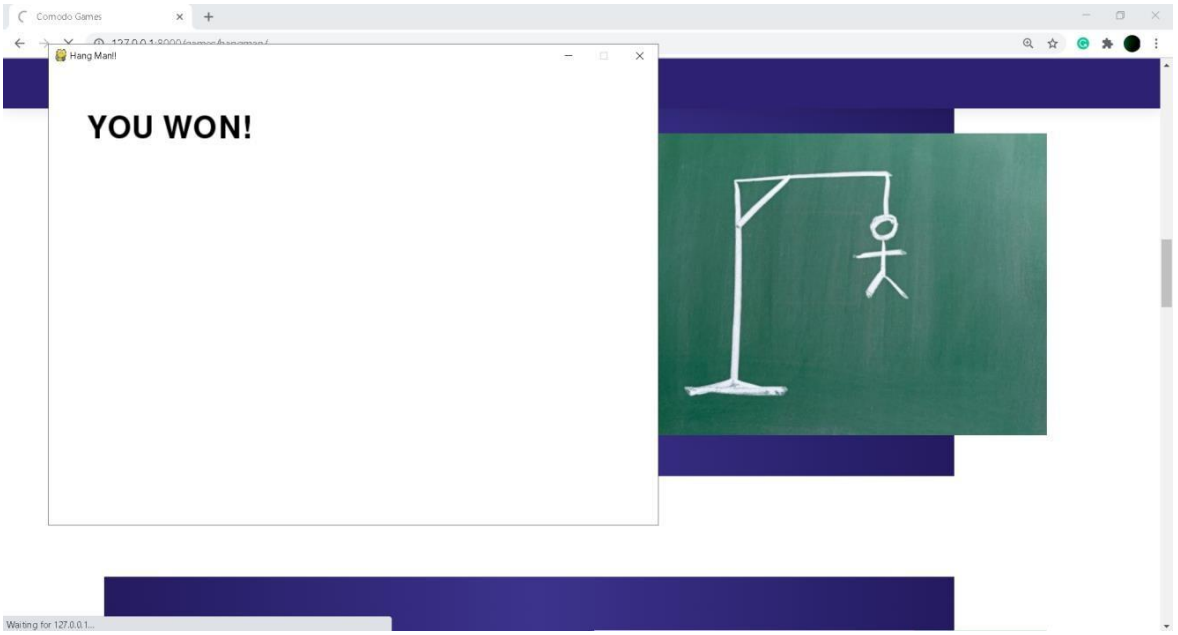
The initial view of Hangman



Hangman’s body being completed with each wrong move



Window if a player loses the game



Window if player wins the game

4. RESULTS AND DISCUSSION

4.1. Results

The prediction done by the rock paper scissors algorithm is 88% correct.

The prediction done in exploratory data analysis is done using two models, namely Random Forests and logistic regression and they give accuracy of 87.49% and 86.92%. average accuracy is 87%.

4.2. Discussion

The analysis of past games shows that

RPGs like Assassin's creed are by far the most popular games bought and played. It is, of course, one of the greatest RPG cum Strategy games released by Ubisoft on multiple platforms. It combines RPG, Adventure, and Strategy.

Call of Duty and Battlefield are by far the most popular shooters which are bought and played a lot.

There is a shift in the scores of critics compared to users.

Usually, it is found that critics do not buy the games and the critic scores are largely based on online viewing of game-plays.

Therefore a game that may be great when playing on your own might not be so appealing when game-plays are the mode of judgment.

Japanese players constitute a large portion of global players one reason could be 2 of the major gaming platforms, namely Sony and Nintendo are Japanese companies.

The predictive algorithm curated using random forest and logistic regression, describes the top 10 most popular games and top 10 least popular games.

Shooter games on X360 are widely popular

Action games are the most popular genre.

5. CONCLUSION

The idea behind this project was to inculcate the basic pillars of data science namely, data analysis, Machine learning, and python programming to create a website that can serve as a single user gaming interface.

The website was divided into three modules, namely:

1. Hangman

In the Hangman game, the program gives a random word from a list of words and the gamer predicts the word. A quirky user interface helps the game to be more enjoyable. The game is made purely in python, including the user interface. The gaming module “pygame” is used here to create the game.

2. Rock Paper Scissors

The Rock paper scissors game enables the user to play the game with real-time hand gestures. The computer generates a random gesture to play. The hand gestures made by the user are segregated in three categories: Rock, paper, and scissors based on which the winner is predicted. We used Keras to train a model on hand gestures into classes of rock, paper, and scissors. The model was saved as an ‘h5’ file. Using this, we could predict the hand gestures of users. The hand gestures were captured and used for playing the game using the video support libraries in OpenCV.

3. Exploratory Data Analysis

Exploratory data analysis is used to find the trends of the gaming world based on past data (1980 to 2015). Tickers such as genre, publisher, Global players, rating, Critic scores, and user scores help to find trends like: Popularity of genres, ratings based on genres, games released along the years, players across different platforms and nations, Sales of different genres according to the platforms, global sales vs. number of games released in different years can be seen. This can help gamers to choose the next best game according to their taste and the companies to cater to the needs of the gaming community. The predictive analysis, predicts the popularity (global sales) of various games in relation to the above specified tickers to in the coming year.

6. REFERENCES

- https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings
- https://www.python.org/
- https://anaconda.org/anaconda/python/
- http://www.numpy.org/
- https://matplotlib.org/
- http://scikit-learn.org/
- https://pandas.pydata.org/
- https://www.djangoproject.com/
- https://www.pygame.org/news
- https://github.com/SouravJohar/rock-paper-scissors
- https://www.youtube.com/watch?v=UEO1B_llDnc
- https://www.youtube.com/watch?v=W6cjx7t39d4