VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report
On

“Exploratory Data Analysis on Steam Games”

Submitted in partial fulfillment of the requirements for the VII Semester of degree of **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

by

Vishal Veeraraj Shetty
1RN19IS178

Under the Guidance of

Ms. Shwetha G N
Asst. Professor
Department of ISE

ESTD:2001
An Institute with a Difference

Department of Information Science and Engineering

RNS Institute of Technology

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru-560098

2022-2023

RNS INSTITUTE OF TECHNOLOGY

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru - 560098

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

ESTD : 2001
An Institute with a Difference

CERTIFICATE

Certified that the internship work entitled ***"Exploratory Data Analysis on Steam Games"*** has been successfully completed by **Vishal Veeraraj Shetty (1RN19IS178),** a bonafide student of **RNS Institute of Technology, Bengaluru** in partial fulfillment of the requirements for the award of degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during academic year **2022-2023**. The internship report has been approved as it satisfies the academic requirements in respect of project work for the said degree.

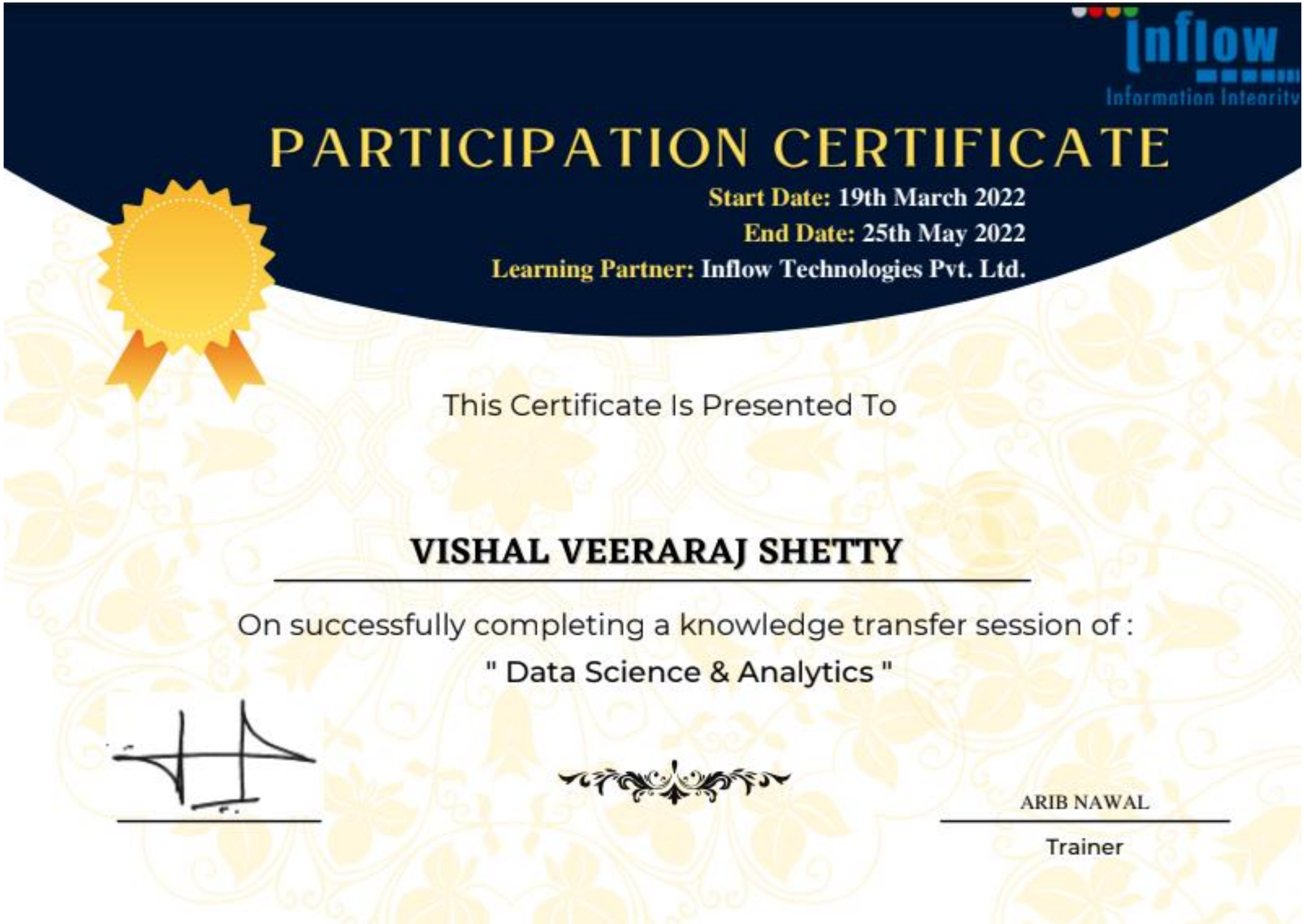
Ms. Shwetha G N Internship Guide Associate Professor Department of ISE	**Dr. R Rajkumar / Mr. Pramoda R** Internship Coordinators Associate Professor Department of ISE	**Dr. Suresh L** Professor & HoD Department of ISE RNSIT

External Viva

Name of the Examiners **Signature with Date**

1. ______________________________ **1.** ________________________

2. ______________________________ **2.** ________________________



Inflow
Information Integrity

PARTICIPATION CERTIFICATE

Start Date: 19th March 2022
End Date: 25th May 2022
Learning Partner: Inflow Technologies Pvt. Ltd.

This Certificate Is Presented To

VISHAL VEERARAJ SHETTY

On successfully completing a knowledge transfer session of :

" Data Science & Analytics "

ARIB NAWAL

Trainer

ABSTRACT

Exploratory data analysis allows business analysts to draw conclusions and make business decisions.

In the world of business, each decision can be either monumental or catastrophic in terms of outcomes.

Since they run on profit, performing analysis on the domain from past data can help managers make better business decisions and reduce overall risk as well as maximize profit.

Depicting most of the trends through graphs and other analytical tools helps us understand the flowthrough of the industry and much more.

The final result of pictorial representation of data gives a clear-cut solution towards future changes and enhancements required as per industry desires.

ACKNOWLEDGMENT

At the very onset I would like to place our gratefulness to all those people who helped me in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is they who guided me in the right direction.

First of all, I would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, I express sincere gratitude to our beloved Principal **Dr. M K Venkatesha,** for providing us all the facilities.

We are extremely grateful to our own and beloved Professor and Head of Department of Information science and Engineering, **Dr. Suresh L**, for having accepted to patronize me in the right direction with all her wisdom.

We place our heartfelt thanks to **Ms. Shwetha G N,** Professor and HOD, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

I thank **Mr. Arib Nawal, Inflow Technologies**, for providing the opportunity to be a part of the Internship program and having guided me to complete the same successfully.

I also thank our internship coordinators **Dr. R Rajkumar,** Associate Professor, and **Mr. Pramod R,** Assistant Professor, Department of Information Science and Engineering. I would thank my friends for having supported me with all their strength and might. Last but not the least, I thank my parents for supporting and encouraging me throughout. I have made an honest effort in this assignment.

VISHAL VEERARAJ SHETTY
1RN19IS178

TABLE OF CONTENTS

List of Figures

List of Abbreviations

EDA	- Exploratory Data Analysis
NLP	- Natural Language Programming
Df	- Data Frame
Pd	- Pandas
Re	- Regular Expression
API	- Application Programming Interface
CSV	- Comma Separated Values

CHAPTER 1

INTRODUCTION

The goal of this project is to analyze and make conclusions on one of the fastest growing industry in the entertainment sector: The video game industry.

We will be analyzing data taken from **Steam platform** as it is the market leader in the industry.

This is performed by taking the raw steam datasets from Kaggle for implementation.

Exploratory data analysis is the process where we clean, analyse and edit the dataset after which it is passed into a visualization software I.e. Tableau to create graphs for enhanced readability.

EDA provides us the opportunity to help us understand the dataset better by the tools it allows us to use, hence making data analysis much simpler.

Cleaning and performing a set of data analytic operations provides a much better understanding towards specific trends and market specific environment.

CHAPTER 2

LITERATURE REVIEW

Big Data Analytics: A review

Authors: Nada Elgendy and Ahmed Elragal, German University in Cairo(GUC), Egypt.

Description:

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques.

Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets.

Data Analytics: A literature review perspective

Author: Sarah Al-Shiakhli, Luleå University of Technology, Sweden.

Big data is currently a buzzword in both academia and industry, with the term being used to describe a broad domain of concepts, ranging from extracting data from outside sources, storing and managing it, to processing such data with analytical techniques and tools.

This thesis work thus aims to provide a review of current big data analytics concepts in an attempt to highlight big data analytics' importance to decision making.

CHAPTER 3

ANALYSIS

3.1 INTRODUCTION

The agenda of performing an exploratory data analysis on the given dataset is to present the data in the form of various pictorial aspects of the information given in the above mentioned dataset.

We are provided with two datasets from the gaming company STEAM and these datasets are used for performing various data analysis on them.

The SteamSpy dataset has a set of unnecessary data present in it which requires us to clean and segregate the un-needed data from the needed data.

These datasets are then merged using specific merge functions provided by python and we perform Data Visualization on them.

3.2 Requirement specification

3.2.1Requirements

3.2.1.1 Software Requirements

Operating System : Windows 10

IDE : Google Collaboratory

Tools/Technology : Python, Matplotlib, NumPy, Tableau

3.2.1.2 Hardware Requirements

Processor : Any processor above 500Mhz

RAM : 512 Mb

Hard Disk : 4GB

Input Device : Standard Keyboard and Mouse

Output Device : VGA and High Resolution Monito

CHAPTER 4

SYSTEM DESIGN

4.1 INTRODUCTION

The entire EDA process involves 4 procedures or steps mentioned below:

1. Data Collection
2. Data Cleaning
3. Data Pre-processing
4. Data Visualization

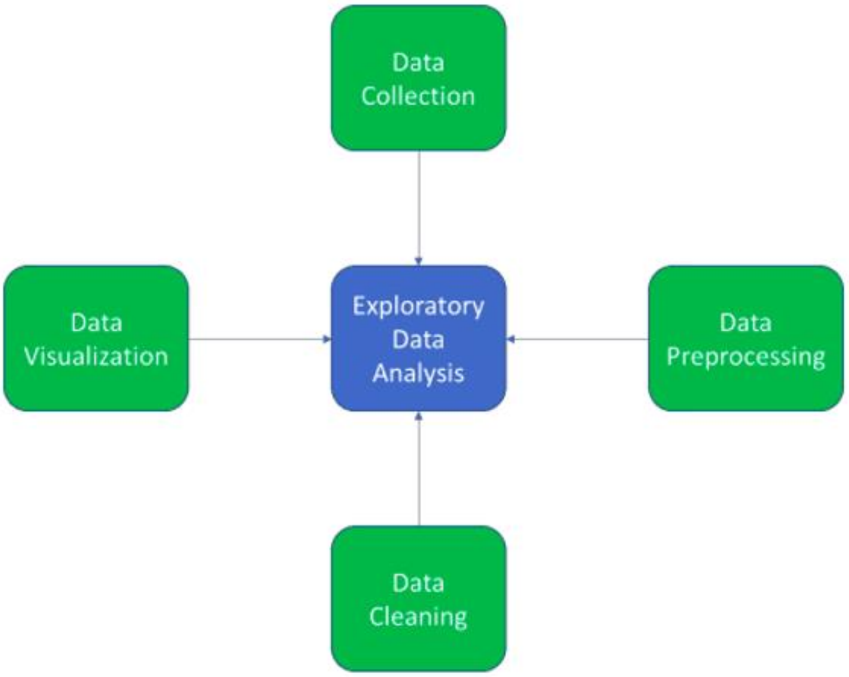


Fig 4.1.1 System Design Flowchart

CHAPTER 5

DETAILED DESIGN

We begin the process by performing the first step:

Gathering raw data via Kaggle.

Two datasets are generated due to use of two APIs –

Steam and SteamSpy Api.

After the generation of the datasets, we import them into a python environment by using Google Collaboratory and begin the second step :

We use Pandas library to create the data-frame and perform the data cleaning process such as removing null and duplicate values.

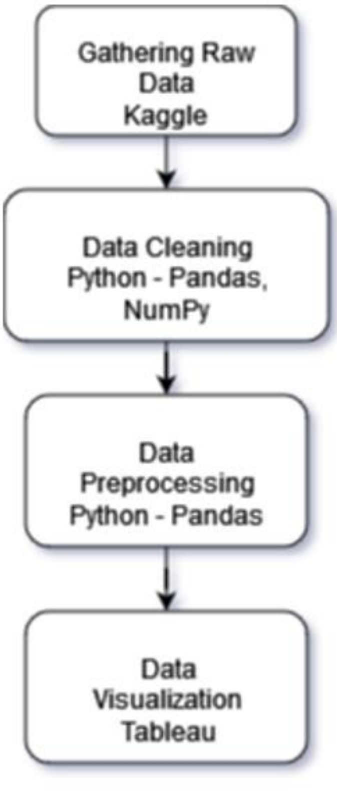

Gathering Raw Data Kaggle
Data Cleaning Python - Pandas, NumPy
Data Preprocessing Python - Pandas
Data Visualization Tableau

Fig 5.1 Detail Design Flowchart

5.1 Gathering Raw Data Kaggle

We have downloaded two data sets from Kaggle which contains games provided by a platform named “Steam”.

They are namely Steam and SteamSpy.

The SteamSpy contains numerical data like owners, number of users, tags etc. where as Steam dataset contains data like release dates, revenue etc.

After gathering the dataset, we import the csv file into google collaboratory for further processing.

5.2 Data Cleaning, Numpy and Pandas

Once we have imported the csv file into the google collaboratory, we perform data cleaning on the data set.

Data cleaning plays an important role in EDA because the visualization will provide more accurate details for future data analysis.

We have used different libraries like numpy and pandas for performing operations for cleaning the data set.

Once the data has been cleaned, we move on to the next process which is data pre-processing.

5.3 Data Pre-Processing

In this step, we pre-process the data in such a way that it becomes more readable and understandable to the data analyst and provides better and accurate results and hence the report can be made with accurate details pertaining to the results.

We have performed data pre-processing on the data by processing two columns related to price which are redundant and unnecessary, hence we have made it into a single column which is now much easier to create visualization to the data analyst and can now make specific reports.

We have also made the release date into a better understanding format which now only previews the exact reveal date without any extra unneeded information.

5.4 Data Visualization, Tableau

After all the cleaning and pre-processing, we have finally achieved a cleaned data set and it is now ready for visualization.

We have used a visualization tool named tableau which imports the cleaned csv file and depicts the trends and flow through charts like pie charts, bar charts, line graphs etc.

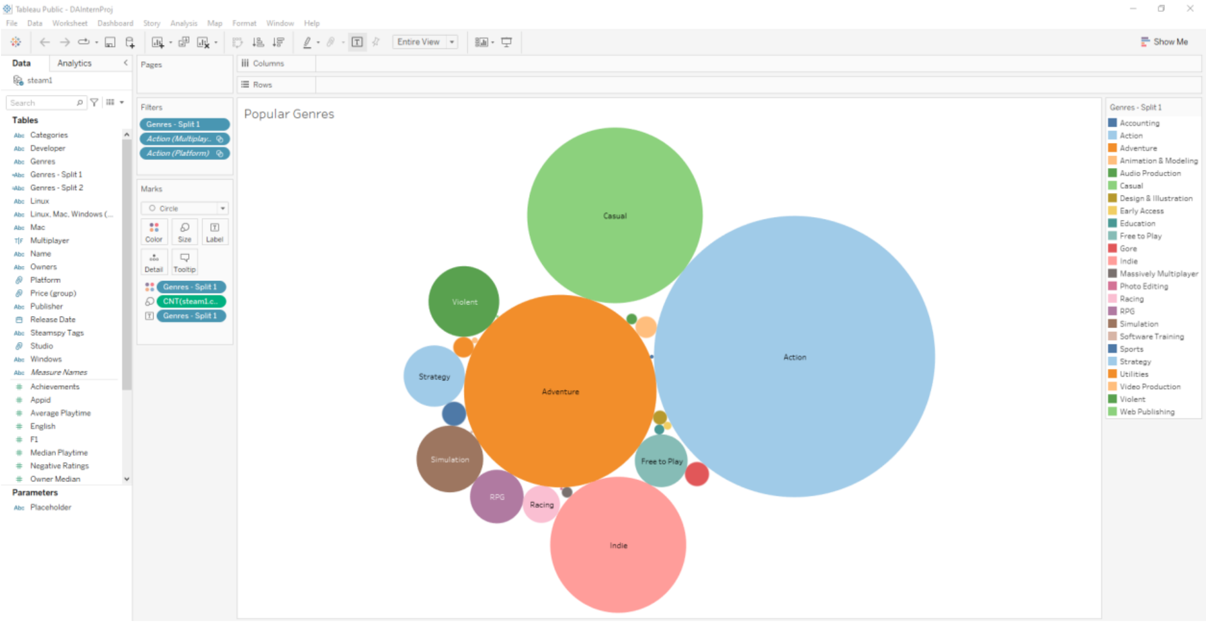


Fig 5.4.1 Tableau UI

CHAPTER 6

TESTING

System testing is actually a series of different tests whose primary purpose is to fully exercise the computer-based system. Although each test has a different purpose, all work to verify that all the system elements have been properly integrated and perform allocated functions. The testing process is actually carried out to makesure that the product exactly does the same thing what is supposed to do.

In the testing stage following goals are tried to achieve: -

- To affirm the quality of the project.
- To find and eliminate and residual errors from previous stages.
- To validate the software as solution to the original problem.
- To provide operational reliability of the system.

We have performed few tests on a duplicate data set in order to check whether they are in working state or not. Once the tests have been performed on the data set and they turn out to be positive, the following operations are applied to the actual data set. The initial tests are done in order to check whether any minor error exists in the data set which could hamper the quality of the project in the future, hence the tests have been done to eliminate these errors and provide reliability of the system.

CHAPTER 7

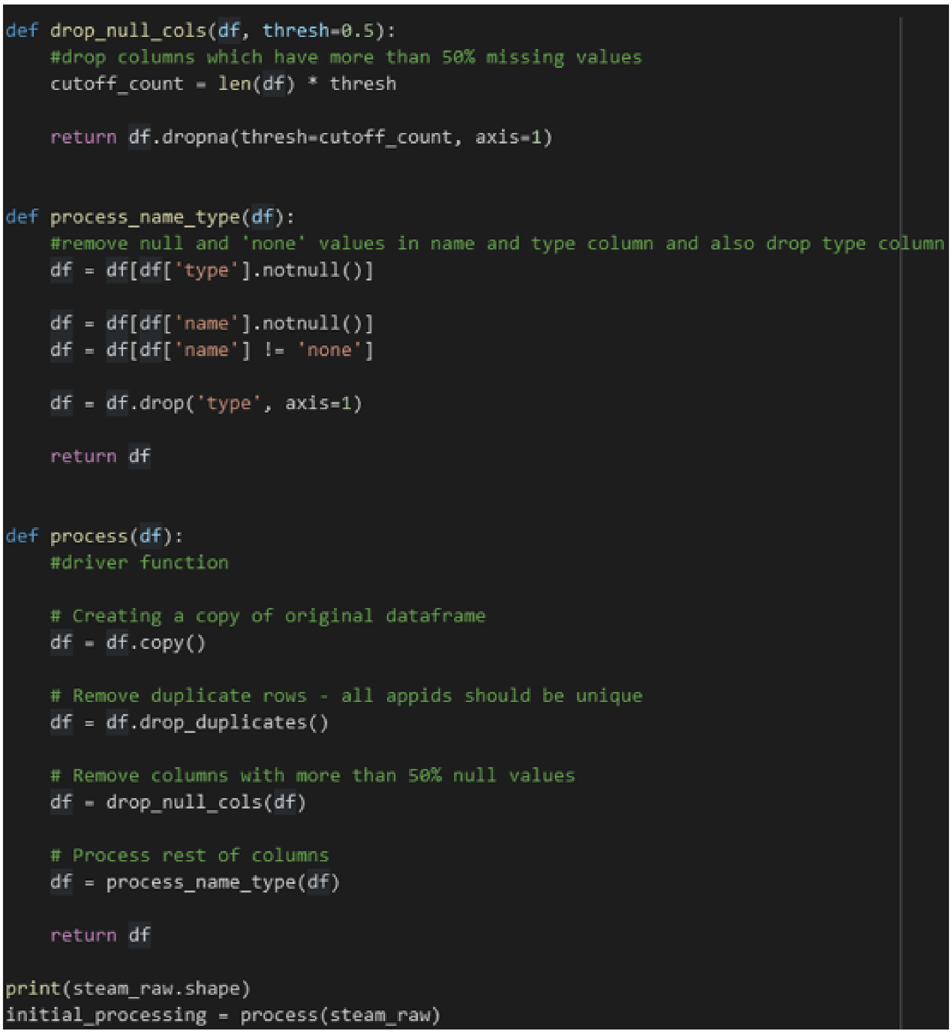
IMPLEMENTATION DETAILS

6.1 INTRODUCTION

We have implemented various functions available in python to help us clean our dataset with much ease.

We have first removed those columns which have insufficient data in it and do not provide any use to data analysis.

The drop keyword is used here to perform the operation.



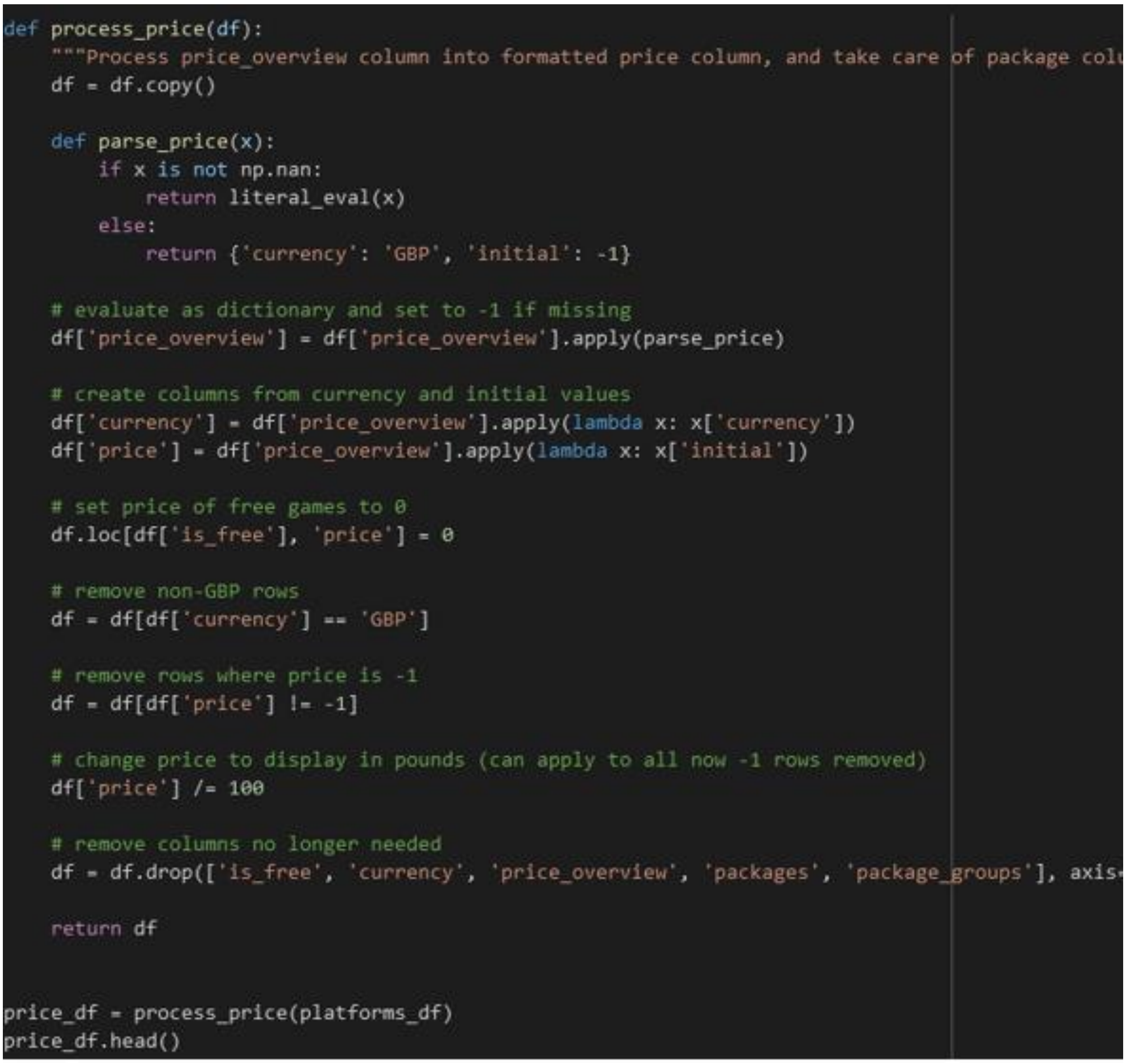
```
def drop_null_cols(df, thresh=0.5):
    #drop columns which have more than 50% missing values
    cutoff_count = len(df) * thresh

    return df.dropna(thresh=cutoff_count, axis=1)


def process_name_type(df):
    #remove null and 'none' values in name and type column and also drop type column
    df = df[df['type'].notnull()]

    df = df[df['name'].notnull()]
    df = df[df['name'] != 'none']

    df = df.drop('type', axis=1)

    return df


def process(df):
    #driver function

    # Creating a copy of original dataframe
    df = df.copy()

    # Remove duplicate rows - all appids should be unique
    df = df.drop_duplicates()

    # Remove columns with more than 50% null values
    df = drop_null_cols(df)

    # Process rest of columns
    df = process_name_type(df)

    return df

print(steam_raw.shape)
initial_processing = process(steam_raw)
```

Fig 6.1.1 Removal of Columns

There are two columns namely “is_free” and “price_overview” .

These columns contain similar data and need to be processed into a single meaningful column for better analysis and learning.

Hence, the following snippet shows how the columns are made into a single column named “price”



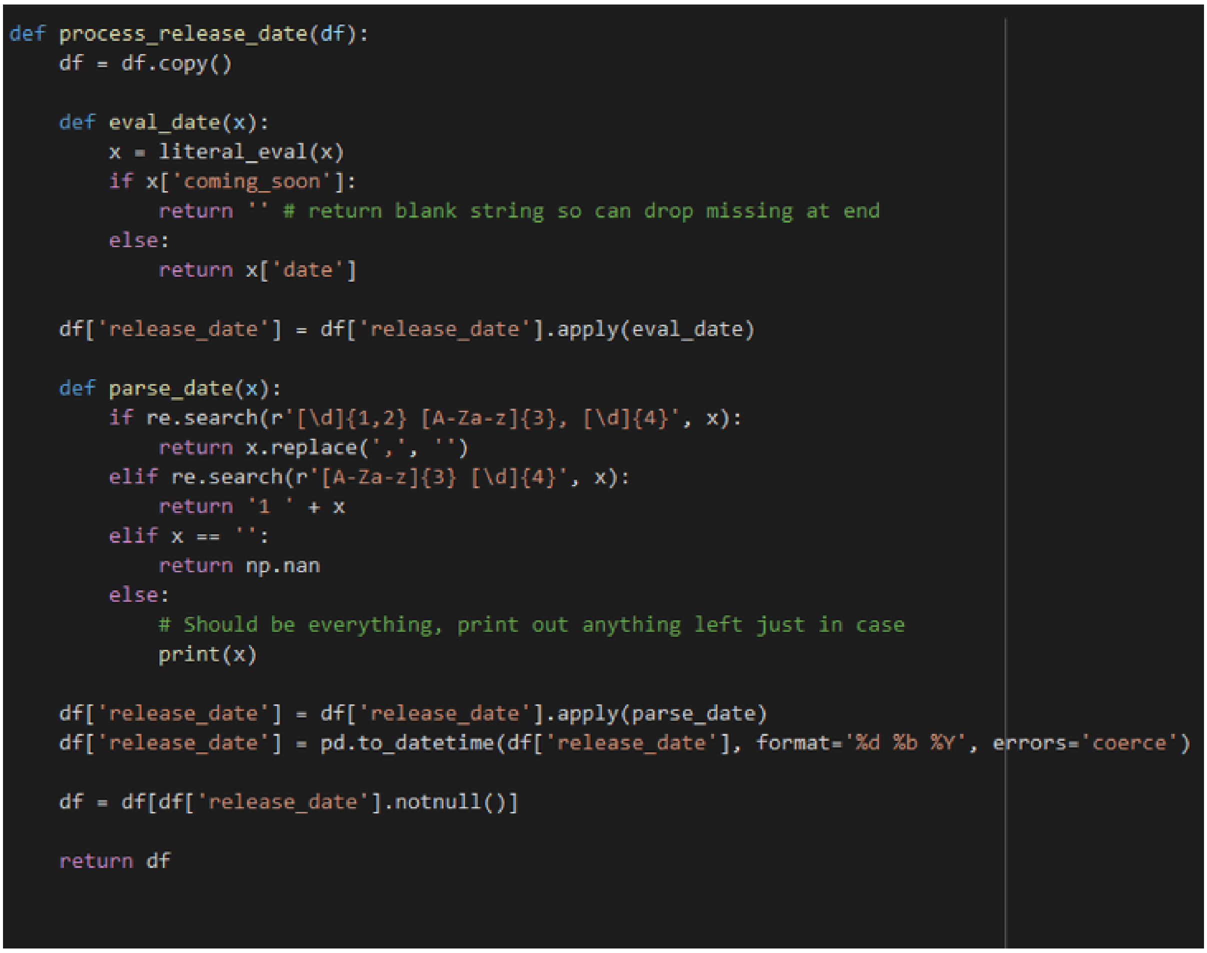
```
def process_price(df):
    """Process price_overview column into formatted price column, and take care of package colu
    df = df.copy()

    def parse_price(x):
        if x is not np.nan:
            return literal_eval(x)
        else:
            return {'currency': 'GBP', 'initial': -1}

    # evaluate as dictionary and set to -1 if missing
    df['price_overview'] = df['price_overview'].apply(parse_price)

    # create columns from currency and initial values
    df['currency'] = df['price_overview'].apply(lambda x: x['currency'])
    df['price'] = df['price_overview'].apply(lambda x: x['initial'])

    # set price of free games to 0
    df.loc[df['is_free'], 'price'] = 0

    # remove non-GBP rows
    df = df[df['currency'] == 'GBP']

    # remove rows where price is -1
    df = df[df['price'] != -1]

    # change price to display in pounds (can apply to all now -1 rows removed)
    df['price'] /= 100

    # remove columns no longer needed
    df = df.drop(['is_free', 'currency', 'price_overview', 'packages', 'package_groups'], axis=

    return df


price_df = process_price(platforms_df)
price_df.head()
```

Fig 6.1.2 Code snippet for processing the price column.

In the original dataset, the “release_date” column consists of certain descriptions which are not clearly understandable for data analysis.

Hence, we decided to process the release date column for filtering out only the date of release for easier data analysis.

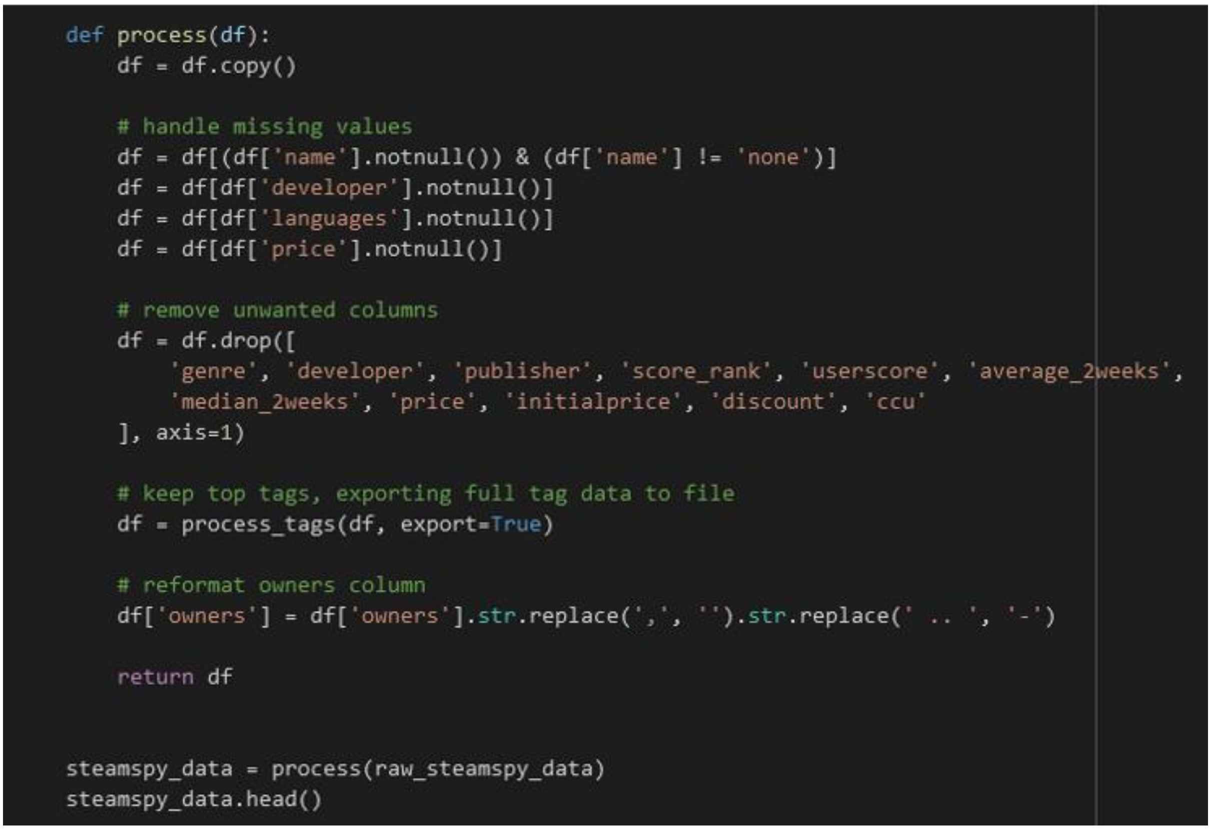
With the help of Python regular expressions, we were able to achieve the filtering of the release date.



```python
def process_release_date(df):
    df = df.copy()

    def eval_date(x):
        x = literal_eval(x)
        if x['coming_soon']:
            return '' # return blank string so can drop missing at end
        else:
            return x['date']

    df['release_date'] = df['release_date'].apply(eval_date)

    def parse_date(x):
        if re.search(r'[\d]{1,2} [A-Za-z]{3}, [\d]{4}', x):
            return x.replace(',', '')
        elif re.search(r'[A-Za-z]{3} [\d]{4}', x):
            return '1 ' + x
        elif x == '':
            return np.nan
        else:
            # Should be everything, print out anything left just in case
            print(x)

    df['release_date'] = df['release_date'].apply(parse_date)
    df['release_date'] = pd.to_datetime(df['release_date'], format='%d %b %Y', errors='coerce')

    df = df[df['release_date'].notnull()]

    return df
```

Fig 6.1.3 Processing release date using RegEx

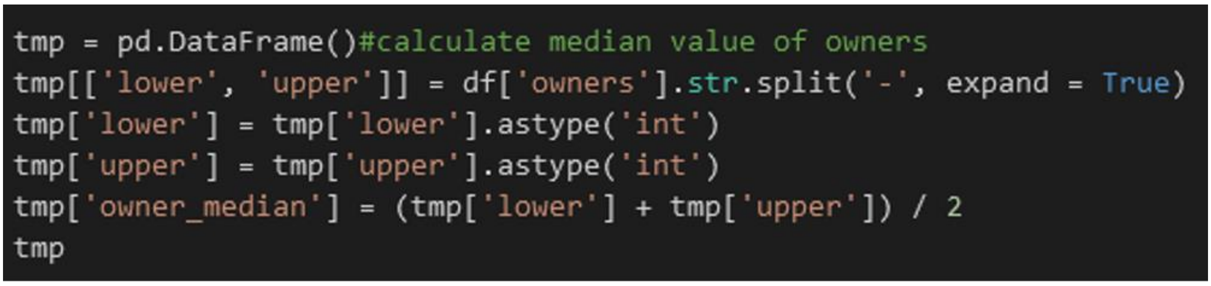
There are some columns which are unneeded for analysis, hence it was ideal to keep 5 tags to a game and drop the remaining tags for easier data analysis.



```python
def process(df):
    df = df.copy()

    # handle missing values
    df = df[(df['name'].notnull()) & (df['name'] != 'none')]
    df = df[df['developer'].notnull()]
    df = df[df['languages'].notnull()]
    df = df[df['price'].notnull()]

    # remove unwanted columns
    df = df.drop([
        'genre', 'developer', 'publisher', 'score_rank', 'userscore', 'average_2weeks',
        'median_2weeks', 'price', 'initialprice', 'discount', 'ccu'
    ], axis=1)

    # keep top tags, exporting full tag data to file
    df = process_tags(df, export=True)

    # reformat owners column
    df['owners'] = df['owners'].str.replace(',', '').str.replace(' .. ', '-')

    return df


steamspy_data = process(raw_steamspy_data)
steamspy_data.head()
```

Fig 6.1.4 Code snippet for dropping unneeded tags

We have also calculated the median value of the owner in the following snippet.



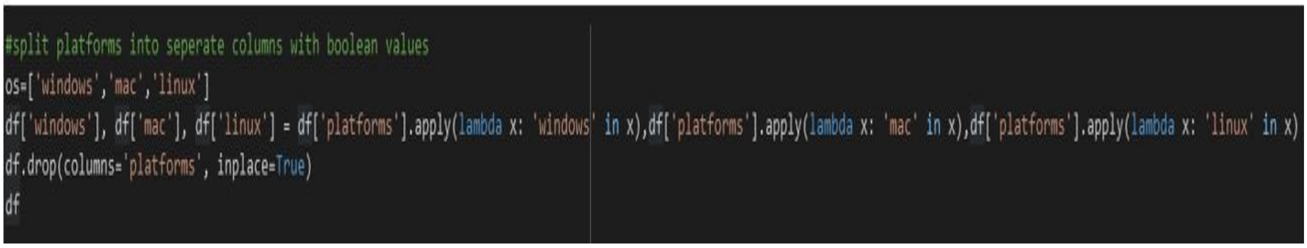
```python
tmp = pd.DataFrame()#calculate median value of owners
tmp[['lower', 'upper']] = df['owners'].str.split('-', expand = True)
tmp['lower'] = tmp['lower'].astype('int')
tmp['upper'] = tmp['upper'].astype('int')
tmp['owner_median'] = (tmp['lower'] + tmp['upper']) / 2
tmp
```

Fig 6.1.5 Calculating owner median

In the original dataset, the platforms are not separated individually are all placed under the same column name “platforms”.

This becomes a tedious task to depict games based on platforms.

Hence, we have segregated each platform to different columns to make the data analysis much easier.



```
#split platforms into seperate columns with boolean values
os=['windows','mac','linux']
df['windows'], df['mac'], df['linux'] = df['platforms'].apply(lambda x: 'windows' in x),df['platforms'].apply(lambda x: 'mac' in x),df['platforms'].apply(lambda x: 'linux' in x)
df.drop(columns='platforms', inplace=True)
df
```

Fig 6.1.6 Splitting platforms individually

CHAPTER 8

RESULTS

After all the data has been successfully cleaned, we have imported the cleaned dataset into a data analytics software called TABLEAU which shows the results in a pictorial representation format.

The following Dashboard in tableau shows the various trends and revenue made in the gaming industry across the world.

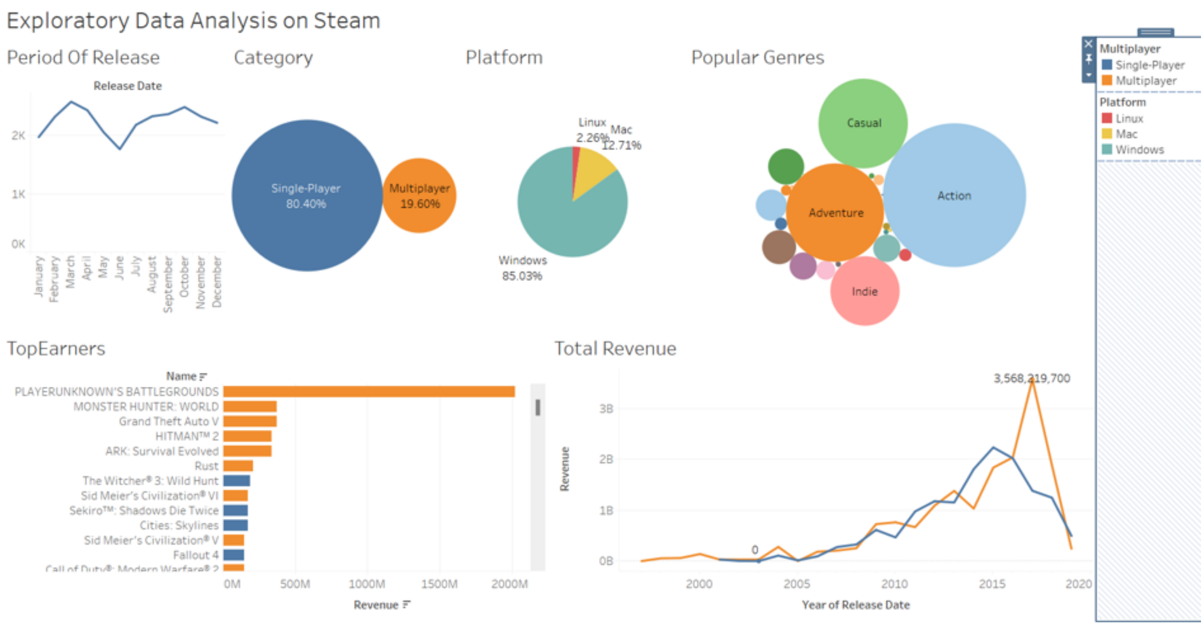


Fig 8.1 Tableau Dashboard representing various charts

The graph below shows the highest revenue made is from a famous multiplayer game called "PUBG" which has made around 2 billion pounds in sales followed by monster hunter.

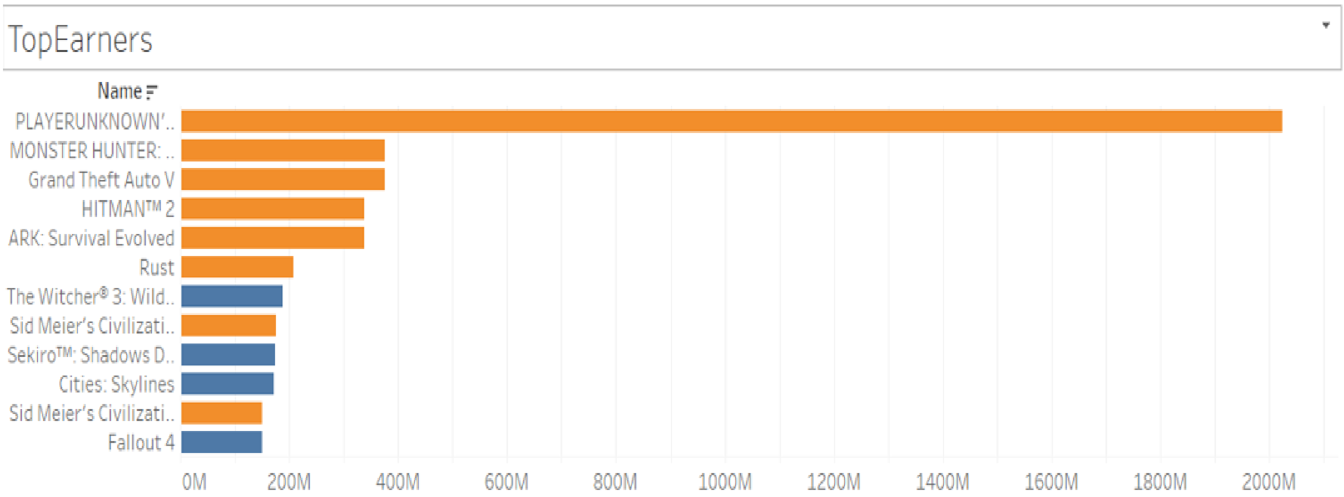


Fig 8.2 Bar Graph representing revenue made

From the line graph below, we can observe that the video games industry has been significantly growing over the years with highest being in 2018 with 3.5 Billion pounds outperforming industries like the music industry.

We can also see that *multiplayer* games are the ones bringing in the most revenue.

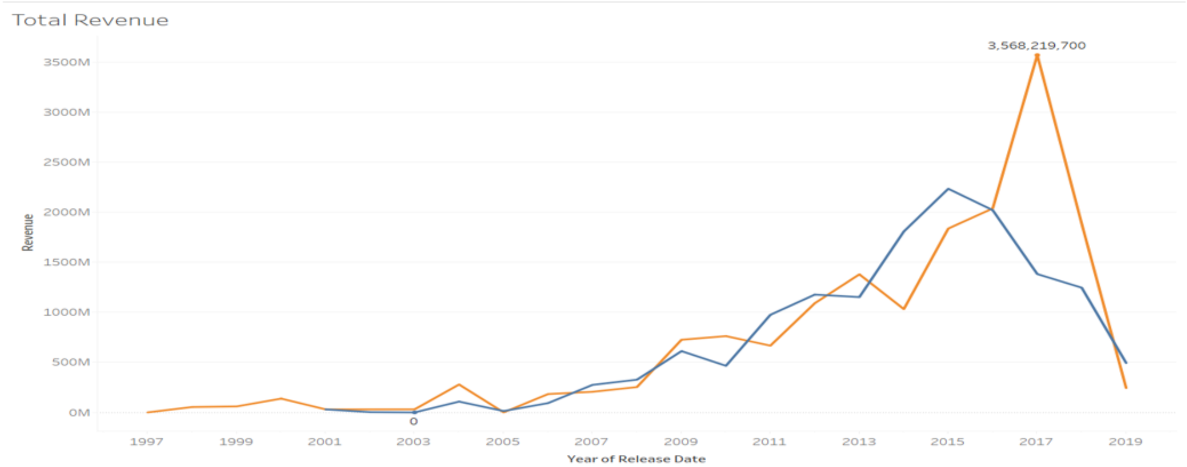


Fig 8.3 Line graph representing revenue v/s year or release

From the bubble chart below, we can see that 80% of most games are *single-player* with *multiplayer* being in the minority despite it bringing in more revenue.

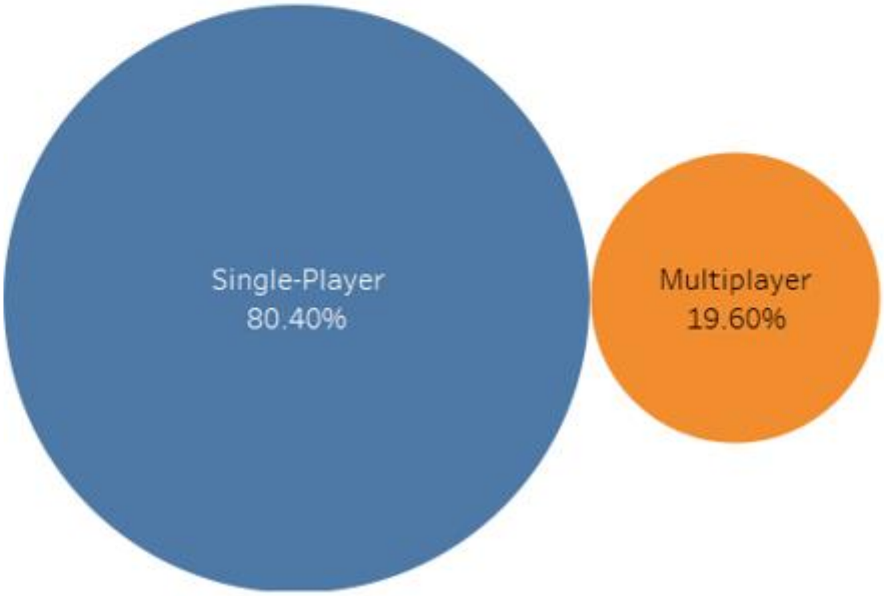


Fig 8.4 Bubble chart representing type of games.

CHAPTER 9

CONCLUSION AND FUTURE WORK

9.1 CONCLUSION

- **From the data, we can conclude that:**
    - Windows is the most preferred platform for releasing video games.
    - Top 5 earners in the industry are all multiplayer games despite single-player games being more in number.
    - Action and Adventure are the most popular genres among users.
    - Video game industry has been growing at a fast pace and is expected to overtake older industries like the music and movie industry making it the top performing industry in the entertainment sector.

9.2 FUTURE ENHANCEMENTS

- Creating a system that involves automatic data collection via Steam and SteamSpy APIs.
- Creating a frontend page and a backend system to provide real-time analytics of various stats.
- Training ML models to generate future predictions and analyze data from a different perspective.

CHAPTER 10

REFERENCES

[1] https://pandas.pydata.org/pandas-docs/stable/reference

[2] https://www.kaggle.com/datasets/nikdavis/steam-store-raw

[3] https://steamapi.xpaw.me/#

[4] https://partner.steamgames.com/doc/webapi

[5] https://stackoverflow.com/