COGS9: Introduction to Data Science
Final Project
Due date: 2023 December 15 23:59:59 (Friday)
Grading: 10% of overall course grade. 40 points total.
Completed as a group. One submission per group on Gradescope.

Group Member Information:

Please read the COGS 9 team policies to best understand how to approach group work and to understand what the expectations are of you in COGS 9.

First Name	Last Name	PID
Dibyesh	Sahoo	A16685980
Ryan	Ho	A17003168
Pranav	Rebala	A16823283
Saksham	Rai	A16228640
Siddhant	Bhagat	A16222672

Question (2 pts)

Clearly state the specific data science question you're interested in answering. This question can be the same as what you submitted for your project proposal. Alternatively, you can edit your original question or change your topic completely.

How has the popularity based on sales of video game genres changed over time with the increasing computing power of gaming platforms?

Hypothesis (2 pts)

Write down your group's hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). You should incorporate the feedback you received on your proposal.

As gaming platforms have evolved, their increased computational capabilities have had a profound influence on game design and genre popularity. We hypothesize that genres such as Role Playing Games (RPGs), Open World games, and First Person Shooters (FPS) have seen a rise in popularity in tandem with the release of more powerful consoles. Iconic titles like "The Grand Theft Auto Series", "God of War", and "Call of Duty" epitomize this, as they require advanced hardware for optimal performance. Our hypothesis is rooted in the observation that with each technologically advanced console launch, such as the PlayStation 5 and Xbox One, there has been a surge in games that fully harness their computational power. We aim to explore the correlation between the advancement of console technology and the increased popularity of certain game genres.

Background Information (3 pts)

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources are fine.

Over time, many different gaming platforms have been developed with the intention of allowing users to play games from the comfort of their own homes. Many would consider the release of the Atari 2600 in 1977 as the beginning of an era of home console gaming that has lasted to this day, paving the way for the future release of iconic platforms such as the Nintendo NES, Sony Playstation, or Microsoft XBOX. In correlation to the rapid increase in computational power in CPU's, and subsequently GPU's, gaming platforms have also rapidly increased in their ability to perform large amounts of operations incredibly quickly. This has arisen as companies constantly seek to offer a more powerful gaming machine to feed the increasing complexity and variety of new games. While FLOPs may not be the most perfect way to measure performance, it can show a general trend for the increase in console performance. FLOPs stands for floating point operations per second which is a type of measurement for how much computing the computer can do in a given time. For example, the Sega Dreamcast was measured to have a performance of 1.4 GFLOPs, while the Playstation 4 was measured at 1.843 TFLOPs. These consoles were released in 1998 and 2013 respectively, showing an increase of 3 orders of magnitude in this specific performance metric over 15 years.

There are many video game genres such as shooters, platformers, and strategy with new genres and subgenres frequently coming out. Each genre has its appeal for different gamers – some want to explore, some want to relax, and others want to use their decision-making skills. Different genres were popular at different times for different platforms. For example, platformers were popular in the mid-80s because they utilized 2D graphics, however once the next generation of consoles were able to produce 3D graphics, a new genre was born: the RPG. MMO games came to light once gaming platforms could connect with each other. Other genres like racing and shooter genres came along over time. So over time, the graphic intensity of video games increases as new popular genres can appeal to the players in a way that was not previously possible due to computing power.

1. https://sugargamers.com/video-game-genres-through-the-years/
2. https://www.bluent.net/blog/evolution-of-gaming/
3. https://www.gamespot.com/gallery/console-gpu-power-compared-ranking-systems-by-flop/2900-1334/#1

4. https://www.ign.com/articles/2009/03/31/a-history-of-gaming-in-nine-influential-genres
5. https://www.wired.com/story/evolution-of-game-console-design-america/

Data (2 pts)

Include a description of the perfect dataset you would need to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the dataset's limitations and how it differs from your ideal dataset. If you collected your own data, explain what information you collected, from whom you collected it, and a link to the data.

The perfect dataset that we want has information about different video games and their release date, console, genre, number of sales, and user ratings. We want to collect the release number of sales and user ratings to determine the popularity of the game in order to compare it to other games. We also want the release date of every game in order to track the popularities over time and observe the trends. Lastly we want data for the console and genre of each video game to analyze how the release of new consoles could affect the popularity of certain genres.

The perfect dataset should have a few thousand observations in order to capture the trends in popularity over time. The data should range from the 1990s to today and contain all the popular consoles like Xbox, Playstation, and Wii to get a better spread. There should be at least 10 games per genre per year and at least 30 user reviews per game to make sure that the data accurately represents reality and is not biased by sampling.

A dataset we found online is linked here:
https://www.kaggle.com/datasets/sidtwr/videogames-sales-dataset?select=XboxOne_GameSales.csv

The dataset contains all of the features we need like games, consoles, release year, genre, sales, user score and critic score. There are 11,563 different observations listed in the dataset. One limitation is that a lot of the review data is missing so we don't have the user score for many games. Another limitation is that a lot of games only have less than 5 reviewers so the average score is easily skewed and not well representative of the game. Also, a lot of the less popular genres of the games are sorted into a miscellaneous category which makes it so we can only look at the few most popular genres and not all of them.

Ethical Considerations (3 pts)

Read the data science ethics checklist from lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Team Bias, Sampling Bias, Data Bias, Consent, Data Privacy / Ownership, Algorithmic Bias / Discrimination, Transparency, Unintended Consequences, Continued Monitoring / Accountability. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered.

Ethical considerations will be a constant and integral part of our project throughout the quarter as we aim to abide by most ethics outlined in the lecture to produce bias free analysis and conclusions

- During the discussion, all of us showed enthusiasm for video games. This might result in each of us having different preconceived notions and perspectives. To evaluate our own biases and minimize them, we plan to conduct peer reviews and feedback at every step of the project
- We plan to inspect the dataset for existing biases relating to mainstream games and companies. While sampling, we aim to be careful to have less mainstream games that might not have significant sales but could potentially influence genre popularity
- The data collected in the dataset might have inherent biases as reasoned above. To eliminate or minimize this bias, we plan to cross-validate our data with other sales data and indicators of popularity
- Consent should not be a concern with respect to our project as all the data alongside with the critic score is publicly available, implying consent. We still aim to abide by the terms and conditions of the data set as per the license wherever possible. Also since this dataset is put together by someone on kaggle, we will always respect its privacy and ownership
- The choice of algorithms, variable, and techniques during our data analyses of video game genre’s popularity over time, we will always be careful to not introduce any kind of bias that might represent a certain group over another
- The dataset itself, the code implemented for data analysis, the decisions made at every step of the process will be documented in detail and will explain our design and implementation choices to provide complete transparency
- We plan to continuously monitor our work and peers by providing constant feedback and reviews
- We aim to represent all genres of video games to avoid any sort of discrimination
- Any re-usable or scalable code for this project will be made publicly available and we also plan to be accountable for our analyses and the conclusions drawn
- Lastly, we shall respect the members of the team, support them, and provide help wherever needed to foster a healthy environment for data analyses

Analysis Proposal (15 pts)

Here, you will propose how you would use and analyze data to answer your question(s) of interest. You are neither expected nor encouraged to carry out the analyses to answer your question(s). You will describe, in detail, what you would need to do to prepare your dataset for analysis (data wrangling) and what type of analysis you would do to answer your question(s). Explain which how your proposed methods / approaches would allow you to interpret the results from this analysis. We are looking for the correct conceptual understanding and application of ideas discussed in class, not specific and technical implementations. For example, if you are applying machine learning to some categorical data, it's important to specify whether you will be performing regression or classification. If you are unsure about the details of anything above, ask on Piazza, come to office hours, and/or do further research on your own (Stack Exchange, Google, Wikipedia, etc.).

Specifically, you are required to incorporate *at least four different methods*, exploring ideas from a combination of:

- Data Collection (web scraping, APIs, etc.)
- Data Wrangling
- Descriptive & Exploratory Data Analysis (summary stats, correlation, etc.)
- Data Visualization
- Statistical Analysis (Inference, A/B testing, etc.)
- Predictive Analysis (machine learning, classification, regression, etc.)
- Text Analysis (Sentiment Analysis, TF-IDF, etc.)
- Geospatial Analysis (choropleth maps, geospatial statistics, etc.)

Method 1: Data Collection

The first step of the project is to collect high quality data to use for our analysis to get the best possible results. The specific things we are looking for in the ideal dataset are: "Game_Title (or Game_Genre), "Year_of_Release ""Global_Sales" and a component of some form of user/critique rating. We want the genre and sales amount for every game in the dataset. However, the user and critique scores are something that we will only consider if our initial analysis that measures game popularity via sales does not give a sound result. This is so because, while user/critique ratings are a more accurate representation of the general "appeal" of the game amongst the population, they are highly subjective and have internal discrepancies regarding the review of a game, from different game review websites such as Metacritic, IGN, Gamestop etc. Whereas, direct sales over a long period of time in years is a pretty straightforward metric of indicating if a gem was well received or flopped. We also want as many games as possible to get the best overall view of the trends in the data. We decided to download the dataset mentioned under the **Data** section from Kaggle that was scraped from Metacritic since we know that website is used a lot for video games and can give us a lot of data. Since the data is stored in a csv file, we can download it and utilize python libraries for prepare it for analysis.

Method 2: Data Wrangling

After identifying our ideal dataset based on our requirements, the next step would be **Data wrangling** on the collected data.

- For our analysis, we concern ourselves with only specific fields in the dataset, namely: “Year_of_Release”, “Genre” and “Total_Global_Sales”. We will filter the original data set based on these columns and store them in a new dataframe. Next, we group by separate game titles under a common “genre” and store the total global sales for that genre For our purposes, we only care about the genre, total sales, and the year published. So we want to select those columns to create a new table that only includes those.
- For the total sales, we want to add up the sales of games that have the same genre and year. This will give us a table that will look something like Table - 2 below. This provides us with a table with only the fields we need to do initial stages of EDA. From this, we can perform analysis to help answer our question.
- In addition to this, as discussed on the **Data** section, some specific games fall under no actual genre like “Sports”, “Action” etc., hence they fall under the category “miscellaneous”. This mostly includes the non-popular genres, all clubbed into one category. Since there are not a lot of games in this group and no clear distinction of genres to draw insights from, we will drop all the rows where Genre == “Miscellaneous”.
- In addition to this, some games/ genres have missing values for specific sales data of different regions. This could just be one of missing values for certain geographical regions, or it could be that certain games are banned in certain regions around the world, during that particular year. However, we take care of that by summing up all regional sales under one column - “Global_Sales”. This at least gives a good ballpark number about the popularity of a certain game or genre during that year.

Sample original Table (Table - 1)

Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
Wii Sports	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53
Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81

Sample Data Wrangled Table (Table - 2)

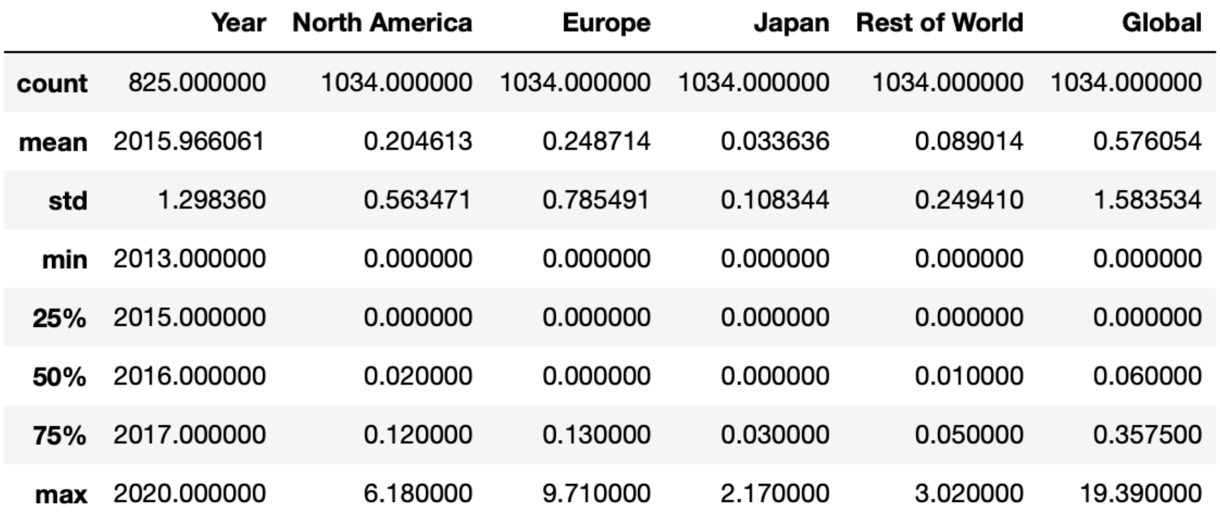
Year_of_Release	Genre	Total_Global_Sales
2006	Sports	182.53
2004	Action	120.81

Method 3: Exploratory Data Analysis

To discern potential correlations between genre types, release years, and sales, an Exploratory Data Analysis (EDA) will be conducted. Firstly, we want to identify outliers that might influence our analytical approach. Secondly, we want to derive valuable insights such as average sales per genre per year, the frequency of game releases categorized by genre and year, and the proportional contribution of each genre to overall sales. Visualization techniques will be employed to facilitate this analysis.

The initial step involves scrutinizing the data distribution to pinpoint outliers. These outliers could represent instances where a single game within a specific genre drastically influences the sales data. Following this, the data will be graphed using line plots to track genre popularity trends over time. This will offer insights into genres that demand greater computational resources, given the historical evolution of gaming hardware and the subsequent advancements in computational capabilities. Identifying these trends will illuminate how certain genres have surged or waned in popularity as computing power has advanced over the years.

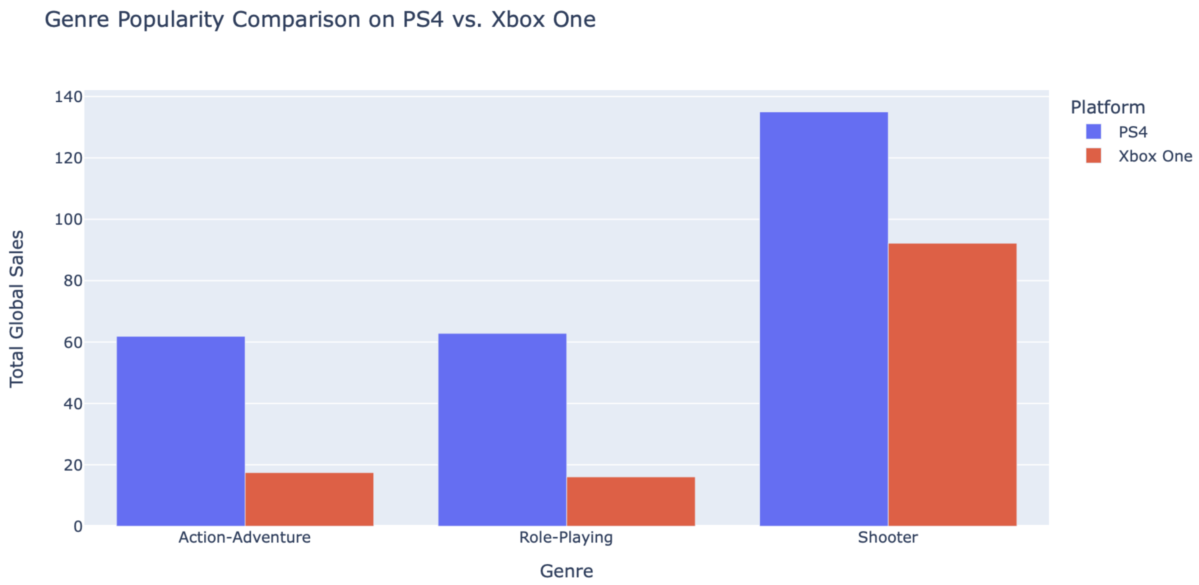
Sample data looking at the distribution of the original dataset



	Year	North America	Europe	Japan	Rest of World	Global
count	825.000000	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000
mean	2015.966061	0.204613	0.248714	0.033636	0.089014	0.576054
std	1.298360	0.563471	0.785491	0.108344	0.249410	1.583534
min	2013.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2015.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2016.000000	0.020000	0.000000	0.000000	0.010000	0.060000
75%	2017.000000	0.120000	0.130000	0.030000	0.050000	0.357500
max	2020.000000	6.180000	9.710000	2.170000	3.020000	19.390000

Method 4: Data Visualization

For our data visualization, we would like to begin with a line graph that displays the average sales per year of each genre over time. Each line in the line graph represents a single genre, and each point on the lines represents the average sales for that genre. The y-axis would be average sales and the x-axis would be each year that we have data from. A line graph was chosen because it is able to clearly indicate any trends in the data cleanly. In addition, we would like to highlight the distribution of sales within each genre by creating a graph containing multiple box plots, with one for each genre. This will allow us to see the spread of the data and provide more perspective than just looking at the averages. Additionally, we would like to aggregate together the entire sales from each genre to generate totals that will be inputted into a pie chart. This will clearly visualize which genres are more popular than others by the share of the pie chart that they take up. Finally, another form of visualization that would be helpful to compare the total sales from each genre together would be a bar graph. The bar graph will also additionally allow us to clearly separate each genre by gaming platform as well, with different colors for each gaming platform. Bar graphs are useful because the height of the bars clearly indicates the magnitude of total sales in a different way from pie charts, where the size of the slices of the pie may be difficult to differentiate in size.



This figure shows a sample graph indicating the total sales of three different genres over all time, depicted as a bar graph.

Method 5: Predictive Analysis

We want a model that can predict the amount of sales that a game has given its release year and genre. We will use multiple linear regression because we have 2 features of year and genre. We want to use regression because we want to predict a numerical value of sales rather than a categorical value. Using this regression model, we are able to predict the sales of any game within our training data's time period (1980-2016) and genres. We can implement our linear regression prediction using machine learning tools like sk-learn in python. We will first do one hot encoding on the genre column because it is a categorical variable and we need to make it numerical. We will also introduce new features like the age of a game platform because it might affect game sales. We want to use multiple linear regression because it is very good at taking multiple columns and finding trends in the data. We want to use a 80-20 train test split to create the model and test to make sure it makes good predictions. We train the model on 80% of our data and find the RMSE on the testing data to determine the model's performance. This is a good metric because the model tests on unseen data so we can guess how well it will work on new data outside of our dataset. To finish our model, we can retrain it on all of our combined data and it is then ready to be used.

Discussion (10 pts)

Interpretation of results:

In the endeavor of doing our proposed analysis on the topic, we hope to find concrete proof of correlation between increasing computing power and game genre popularity.

- We projected that our analysis would likely reveal trends indicative of a relationship between the technological evolution of gaming platforms and shifts in genre popularity, particularly in genres that demand higher computational power, such as RPGs, FPS, and Open World games. However, our conclusions would require careful interpretation.
- The nature of our data, primarily derived from sales figures, presents limitations. Sales data, while quantifiable, might not fully capture the nuances of a genre's popularity or gamer preferences, particularly for less mainstream or indie games, which could lead to sampling bias.
- Moreover, the potential for confounding variables affecting game popularity exists. Increased sales might correspond with marketing efforts or social phenomena rather than purely with platform capabilities. The data might also reflect a regional preference bias, given that our dataset is more reflective of Western gaming markets.To account for this, we would consider integrating data on marketing spend and social media engagement to see how these factors interact with sales figures and platform capabilities.
- Our dataset's scope, from the 1990s to the present, encompassing platforms like Xbox, PlayStation, and Wii, is extensive yet not exhaustive. It omits less popular genres and platforms, possibly skewing genre popularity analyses.This involves popular non-popular console titles such Legend Of Zeda (Nintendo), Star Wars (PC).

To address the limitations and biases in our project, a strategic and multifaceted approach is essential:

- Firstly, diversifying our dataset is crucial. We should include not only mainstream games but also indie and niche titles, and integrate global sales data. This approach provides a more comprehensive view of genre popularity, mitigating regional and mainstream biases. Additionally, we can consider incorporating qualitative data from player reviews and social media to capture nuances beyond sales figures.

- Secondly, it's vital to account for confounding variables. Analyze factors such as marketing strategies and social trends that might influence game popularity. By integrating data on marketing spend and social media engagement, we can discern whether popularity trends are truly due to platform capabilities or other external factors.
- Regarding ethical considerations, maintaining an unbiased and respectful research environment is key. Regular peer reviews and feedback will help mitigate personal biases and ensure balanced perspectives. Documenting all aspects of your research process enhances transparency and credibility. Moreover, ensuring that the analysis represents all game genres prevents discrimination, contributing to a more inclusive understanding of the gaming industry.

Group Participation (3 pts)

Include one paragraph briefly outlining the contribution of each group member throughout the quarter while working on this project. Each of you must also fill out the survey (link provided toward the end of the quarter) about individual and group participation. **The results of this survey can negatively impact an individual's final grade if the group provides evidence that one member did not contribute to the project**. (3 pts)

- **Saksham Rai**: Wrote the data wrangling, data collection and contributed to writing the research question, hypothesis and the Discussion Section **->** Interpretation of results, Potential Limitations.
- **Siddhant Bhagat**: Discussion Section **->** How did we address them, societal implications and ethical implications. Wrote the ethical consideration section too. Exploratory Data Analysis statistics table. Data Visualization graph and information
- **Ryan**: First part of Background Info, Data Visualization
- **Dibyesh**: Second part of Background Info, Data Wrangling, Exploratory Data Analysis
- **Pranav**: Data and Analysis Proposal (Data collection, predictive analysis)