Data Analysis

The results of my data analysis provided me with rather surprising results. I used web scraping to pull data from steam store page, most popular games. I first tried

picking randomly but it was inefficient and worse result, even with adjustments i mentioned, so i stuck with most popular games. First I measured the correlation

between the number of reviews(which indicates number of units sold) and positive reviews. Second, in the same code, I measured the correlation between the price of a game and number of reviews(units sold). This is the code I used to conduct this experiment:

import requests

from bs4 import BeautifulSoup

import time #to avoid overloading servers which can violate policies

import matplotlib.pyplot as plt

import re

import numpy as np

base\_url = "https://store.steampowered.com/search/?ignore\_preferences=1&category1=998&ndl=1&page={}"

all\_games = []

# scrape 10 pages

for page in range(1, 10):

print(f"Scraping page {page}...")

url = base\_url.format(page) #scrape each page

response = requests.get(url)

soup = BeautifulSoup(response.text, "html.parser")

games = soup.find\_all("a", class\_="search\_result\_row") #search for for all the <a> tags

for game in games:

# get title

title\_tag = game.find("span", class\_="title")

title = title\_tag.text.strip() if title\_tag else "Unknown"

# release date

date\_tag = game.find("div", class\_="search\_released")

release\_date = date\_tag.text.strip() if date\_tag else "Unknown"

# review summary

review\_tag = game.find("span", class\_="search\_review\_summary")

review\_text = review\_tag["data-tooltip-html"] if review\_tag and "data-tooltip-html" in review\_tag.attrs else "No reviews"

# review percentage and count

review\_percent = None

review\_count = None

if "of the" in review\_text:

match = re.search(r"(\d+)% of the ([\d,]+)", review\_text)

if match:

review\_percent = int(match.group(1))

review\_count = int(match.group(2).replace(",", ""))

price\_tag = game.find("div", class\_="discount\_final\_price")

price = None

if price\_tag:

price\_text = price\_tag.text.strip()

price\_text = price\_text.replace("Free", "0").split("$")[-1].strip()

try:

price = float(price\_text.replace(",", "."))

except:

pass # Ignore if price format is weird

# Save all info

all\_games.append({

"title": title,

"release\_date": release\_date,

"review\_percent": review\_percent,

"review\_count": review\_count,

"price": price

})

time.sleep(0.2) # don't overload the server

# Keep only 2020 games with reviews

games\_2020 = [g for g in all\_games if "2024" in g["release\_date"] and g["review\_percent"] is not None and g["review\_count"] is not None]

x = [g["review\_count"] for g in games\_2020]

y = [g["review\_percent"] for g in games\_2020]

plt.figure(figsize=(10, 6))

plt.scatter(x, y, alpha=0.6, color="mediumseagreen")

plt.xscale("log")

plt.title("Steam Games (2020): Positive Review % vs Review Count")

plt.xlabel("Review Count")

plt.ylabel("Positive Review %")

plt.grid(True)

plt.tight\_layout()

plt.show()

corr = np.corrcoef(x, y)[0, 1]

print(f" Correlation (Review Count vs Positive %): {corr:.4f}")

games\_with\_price = [g for g in games\_2020 if g["price"] is not None]

x\_price = [g["price"] for g in games\_with\_price]

y\_reviews = [g["review\_count"] for g in games\_with\_price]

plt.figure(figsize=(10, 6))

plt.scatter(x\_price, y\_reviews, alpha=0.6, color="coral")

plt.yscale("log") #we need log scale to have a proper graph

plt.title("Steam Games: Price vs Review Count")

plt.xlabel("Price ($)")

plt.ylabel("Review Count")

plt.grid(True)

plt.tight\_layout()

plt.show()

price\_corr = np.corrcoef(x\_price, y\_reviews)[0, 1]

print(f" Correlation (Price vs Review Count): {price\_corr:.4f}")

this code yielded me interesting results:

ekran görüntüsü, çizgi, metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, metin, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

First graph yields a correlation coefficient of 0.22, which shows a weak correlation between the positive reviews and purchase numbers, which is as expected. So our alternative hypothesis H₁: Review scores have a correlation with financial success is true and we reject the null hypothesis. I found something very interesting on the second test. As it turns out (contrary to what I expected) , as game prices go up, sales numbers also go up. Graph shows 0.11 correlation coefficient. I think this is caused as a result of more pricey games being more worked on and more premium. Turns out people chose quality over low prices in video games. So again we reject the null hypothesis H₀: Game price has no effect on revenue generation.

Second test I conducted aimed to see if there is a correlation between developer studios country of origin and sales numbers. I could not Access the developer directly from the store we page so i had to individually go through the gamesi which took waaaay longer, and i multiplied the price of the game with number of sales (number of reviews\*50) to find the gross revenue generated. I could not plot like the previous ones so I listed the top 10 most developers according to revenue generation.

metin, ekran görüntüsü, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

2 in UK

1 in Sweden

1 in Japan

4 in US

1 in Canada

1 in Czechia

The developers origins show that 40% of studios are in US but this is expected since they are the ones doing most of the investment compared to other countries. The rest of the countries in the list are well developed countries as well (from western europe) so this time again, we reject the null hypothesis H₀: The country of origin has no influence in the financial success of a game.