# Exploratory Data Analysis (EDA)

This notebook is responsible for exploring the data on the cleaned steam games dataset. The goal is to understand the distributions of key variables, identify relationships between features and gather insights for providing recommendations for models.

We would also like to test whether content features (like genre/tags) can be potentially used to describe metadata features (like popularity/rating).

In [ ]:

```python
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np

from wordcloud import WordCloud, STOPWORDS
import re

sns.set(style="whitegrid")
```

In [9]:

```python
df = pd.read_csv("Dataset/games_cleaned.csv")
```

In [10]:

```python
print(f"DataFrame shape: {df.shape}")
print(f"DataFrame columns: {df.columns.tolist()}")
```

```
DataFrame shape: (66249, 550)
DataFrame columns: ['AppID', 'Name', 'Peak CCU', 'Required age', 'Price', 'DLC count', 'About
the game', 'Windows', 'Mac', 'Linux', 'Achievements', 'Recommendations', 'Average playtime two
weeks', 'Median playtime forever', 'Developers', 'Publishers', 'Game Age (Days)', 'is_indie',
'Owner range', 'Total Reviews', 'Review Ratio', 'Num Languages', 'Is English Supported', 'Tags
_Singleplayer', 'Tags_Action', 'Tags_Casual', 'Tags_Adventure', 'Tags_2D', 'Tags_Strategy', 'T
ags_Simulation', 'Tags_RPG', 'Tags_Puzzle', 'Tags_Atmospheric', 'Tags_3D', 'Tags_Early Access'
, 'Tags_Pixel Graphics', 'Tags_Story Rich', 'Tags_Colorful', 'Tags_Exploration', 'Tags_Cute',
'Tags_First-Person', 'Tags_Arcade', 'Tags_Multiplayer', 'Tags_Fantasy', 'Tags_Funny', 'Tags_Sh
ooter', 'Tags_Horror', 'Tags_Retro', 'Tags_Platformer', 'Tags_Anime', 'Tags_Family Friendly',
'Tags_Sci-fi', 'Tags_Action-Adventure', 'Tags_Relaxing', 'Tags_Female Protagonist', 'Tags_Diff
icult', 'Tags_Third Person', 'Tags_VR', 'Tags_Survival', 'Tags_Top-Down', 'Tags_Open World', '
Tags_Stylized', 'Tags_Controller', 'Tags_Combat', 'Tags_Great Soundtrack', 'Tags_Comedy', 'Tag
s_2D Platformer', 'Tags_Visual Novel', 'Tags_FPS', 'Tags_Violent', 'Tags_Mystery', 'Tags_Co-o
p', 'Tags_Dark', 'Tags_Physics', 'Tags_Realistic', 'Tags_Minimalist', 'Tags_Free to Play', 'Ta
gs_Cartoony', 'Tags_Psychological Horror', 'Tags_Choices Matter', 'Tags_Point & Click', 'Tags_
Linear', 'Tags_Gore', 'Tags_PvP', 'Tags_Sandbox', 'Tags_Multiple Endings', 'Tags_Sports', 'Tag
s_Side Scroller', 'Tags_Space', 'Tags_Rogue-like', 'Tags_Old School', 'Tags_Tactical', 'Tags_P
vE', 'Tags_Rogue-lite', 'Tags_Building', 'Tags_Puzzle-Platformer', 'Tags_Management', 'Tags_Ch
aracter Customization', 'Tags_Hidden Object', 'Tags_Hand-drawn', 'Tags_Action RPG', 'Tags_Sexu
al Content', 'Tags_Magic', 'Tags_Logic', 'Tags_Racing', 'Tags_Nudity', 'Tags_Local Multiplayer
', "Tags_Shoot 'Em Up", 'Tags_3D Platformer', 'Tags_Procedural Generation', 'Tags_Turn-Based S
trategy', 'Tags_Survival Horror', 'Tags_Cartoon', 'Tags_Futuristic', 'Tags_Bullet Hell', 'Tags
_Medieval', 'Tags_Turn-Based Combat', 'Tags_Online Co-Op', 'Tags_Crafting', 'Tags_Turn-Based T
actics', 'Tags_Walking Simulator', 'Tags_Drama', 'Tags_Interactive Fiction', 'Tags_Choose Your
Own Adventure', 'Tags_Hack and Slash', 'Tags_Action Roguelike', "Tags_1990's", 'Tags_Resource
Management', 'Tags_Zombies', 'Tags_Local Co-Op', 'Tags_Score Attack', 'Tags_JRPG', 'Tags_Dunge
on Crawler', 'Tags_Replay Value', 'Tags_Dark Fantasy', 'Tags_Surreal', 'Tags_War', 'Tags_Histo
rical', 'Tags_Immersive Sim', 'Tags_Turn-Based', 'Tags_Post-apocalyptic', 'Tags_Top-Down Shoot
er', 'Tags_Nature', 'Tags_Emotional', 'Tags_Base-Building', 'Tags_Stealth', 'Tags_Isometric',
'Tags_Education', 'Tags_Romance', 'Tags_Text-Based', 'Tags_Fast-Paced', 'Tags_Massively Multip
layer', 'Tags_Short', 'Tags_Card Game', 'Tags_Abstract', 'Tags_Classic', 'Tags_Clicker', 'Tags
_RTS', 'Tags_Precision Platformer', 'Tags_Military', 'Tags_1980s', 'Tags_2.5D', 'Tags_Third-Pe
rson Shooter', 'Tags_Tower Defense', 'Tags_Investigation', 'Tags_Board Game', 'Tags_Detective'
, 'Tags_Memes', 'Tags_RPGMaker', 'Tags_Narration', 'Tags_Dating Sim', 'Tags_Robots', 'Tags_Ali
ens', 'Tags_Cyberpunk', 'Tags_Perma Death', 'Tags_Tabletop', 'Tags_Driving', 'Tags_Cinematic',
'Tags_Life Sim', 'Tags_Dark Humor', 'Tags_Economy', 'Tags_Arena Shooter', 'Tags_Time Managemen
t', 'Tags_VR Only', 'Tags_Mature', 'Tags_4 Player Local', 'Tags_Flight', 'Tags_Psychological',
'Tags_Thriller', 'Tags_Strategy RPG', 'Tags_Real Time Tactics', 'Tags_Demons', 'Tags_City Buil
der', 'Tags_Experimental', "Tags_Beat 'em up", 'Tags_Conversation', 'Tags_Psychedelic', 'Tags_
Runner', 'Tags_Fighting', 'Tags_Metroidvania', 'Tags_LGBTQ+', 'Tags_Tactical RPG', 'Tags_Warga
me', 'Tags_Nonlinear', 'Tags_Supernatural', 'Tags_Music', 'Tags_Team-Based', 'Tags_Collectatho
n', 'Tags_Tutorial', 'Tags_Competitive', 'Tags_Lore-Rich', 'Tags_Level Editor', 'Tags_Artifici
al Intelligence', 'Tags_Idler', 'Tags_Comic Book', 'Tags_Loot', 'Tags_Twin Stick Shooter', 'Ta
gs_Automobile Sim', 'Tags_Party-Based RPG', 'Tags_Dystopian', 'Tags_Parkour', 'Tags_Modern', '
Tags_Destruction', 'Tags_Grid-Based Movement', 'Tags_2D Fighter', 'Tags_Souls-like', 'Tags_Hen
tai', 'Tags_Match 3', 'Tags_Cats', 'Tags_Alternate History', 'Tags_Rhythm', 'Tags_Deckbuilding
', 'Tags_CRPG', 'Tags_Inventory Management', 'Tags_Moddable', 'Tags_Crime', 'Tags_Mythology',
'Tags_Space Sim', 'Tags_Card Battler', 'Tags_Beautiful', 'Tags_Soundtrack', 'Tags_Grand Strate
gy', 'Tags_World War II', 'Tags_Philosophical', 'Tags_Dark Comedy', 'Tags_Science', 'Tags_Noir
', 'Tags_Mystery Dungeon', 'Tags_3D Fighter', 'Tags_Character Action Game', 'Tags_Lovecraftian
', 'Tags_NSFW', 'Tags_Word Game', 'Tags_Split Screen', 'Tags_Swordplay', 'Tags_Colony Sim', 'T
ags_Farming Sim', 'Tags_Automation', 'Tags_Mouse only', 'Tags_Creature Collector', 'Tags_e-spo
rts', 'Tags_6DOF', 'Tags_Dragons', 'Tags_Voxel', 'Tags_3D Vision', 'Tags_Vehicular Combat', 'T
ags_Solitaire', 'Tags_Bullet Time', 'Tags_Hero Shooter', 'Tags_Mechs', 'Tags_Battle Royale', '
Tags_Agriculture', 'Tags_Capitalism', 'Tags_Parody', 'Tags_Combat Racing', 'Tags_Open World Su
rvival Craft', 'Tags_Spectacle fighter', 'Tags_America', 'Tags_Design & Illustration', 'Tags_B
lood', 'Tags_Time Manipulation', 'Tags_Gun Customization', 'Tags_Class-Based', 'Tags_Sokoban',
'Tags_MMORPG', 'Tags_Steampunk', 'Tags_Hex Grid', 'Tags_Addictive', 'Tags_God Game', 'Tags_Pol
itical', 'Tags_FMV', 'Tags_Conspiracy', 'Tags_Martial Arts', 'Tags_Ninja', 'Tags_Gothic', 'Tag
s_Co-op Campaign', 'Tags_Pirates', 'Tags_Otome', 'Tags_Tanks', 'Tags_Auto Battler', 'Tags_Trad
ing', 'Tags_Underground', 'Tags_Satire', 'Tags_Real-Time', 'Tags_Utilities', 'Tags_Quick-Time
Events', 'Tags_Dog', 'Tags_Mining', 'Tags_Cooking', 'Tags_Time Travel', 'Tags_Looter Shooter',
'Tags_4X', 'Tags_Remake', 'Tags_Underwater', 'Tags_Hacking', 'Tags_Hunting', 'Tags_Programming
', 'Tags_Dynamic Narration', 'Tags_Cult Classic', 'Tags_Dinosaurs', 'Tags_Fishing', 'Tags_Poli
tics', 'Tags_Escape Room', 'Tags_Real-Time with Pause', 'Tags_Faith', 'Tags_Naval', 'Tags_Vamp
ire', 'Tags_Minigames', 'Tags_Software', 'Tags_Political Sim', 'Tags_Western', 'Tags_Superhero
', 'Tags_Narrative', 'Tags_Party Game', 'Tags_Transportation', 'Tags_Trading Card Game', 'Tags
_Assassin', 'Tags_Immersive', 'Tags_Action RTS', 'Tags_Illuminati', 'Tags_Cozy', 'Tags_Touch-F
riendly', 'Tags_MOBA', 'Tags_Time Attack', 'Tags_Typing', 'Tags_Game Development', 'Tags_Async
hronous Multiplayer', 'Tags_Trivia', 'Tags_Trains', 'Tags_On-Rails Shooter', 'Tags_Roguelike D
eckbuilder', 'Tags_Cold War', 'Tags_Party', 'Tags_Snow', 'Tags_Animation & Modeling', 'Tags_Tr
aditional Roguelike', 'Tags_Archery', 'Tags_Offroad', 'Tags_Heist', 'Tags_Naval Combat', 'Tags
_Diplomacy', 'Tags_Football', 'Tags_Soccer', 'Tags_Villain Protagonist', 'Tags_Music-Based Pro
cedural Generation', 'Tags_Kickstarter', 'Tags_Sailing', 'Tags_Mars', 'Tags_Wholesome', 'Tags_
Chess', 'Tags_Foreign', 'Tags_Experience', 'Tags_Nostalgia', 'Tags_Gambling', 'Tags_Sequel', '
Tags_Sniper', 'Tags_GameMaker', 'Tags_Boxing', 'Tags_Horses', 'Tags_Episodic', 'Tags_World War
I', 'Tags_Golf', 'Tags_Spelling', 'Tags_Jet', 'Tags_Unforgiving', 'Tags_Motorbike', 'Tags_Outb
reak Sim', 'Tags_Transhumanism', 'Tags_Werewolves', 'Tags_Rome', 'Tags_Pinball', 'Tags_Bikes',
'Tags_Farming', 'Tags_Silent Protagonist', 'Tags_Epic', 'Tags_Roguevania', 'Tags_Spaceships',
'Tags_Basketball', 'Tags_Asymmetric VR', 'Tags_Submarine', 'Tags_Crowdfunded', 'Tags_Medical S
im', 'Tags_Audio Production', 'Tags_LEGO', 'Tags_Video Production', 'Tags_Social Deduction', '
Tags_Games Workshop', 'Tags_Mini Golf', 'Tags_Vikings', 'Tags_Based On A Novel', 'Tags_Movie',
'Tags_Electronic Music', 'Tags_Ambient', 'Tags_Baseball', 'Tags_Software Training', 'Tags_Wres
tling', 'Tags_360 Video', 'Tags_Dungeons & Dragons', 'Tags_Tennis', 'Tags_Warhammer 40K', 'Tag
s_Pool', 'Tags_Gaming', 'Tags_Mod', 'Tags_Motocross', 'Tags_Lemmings', 'Tags_Intentionally Awk
ward Controls', 'Tags_Photo Editing', 'Tags_Skateboarding', 'Tags_Hockey', 'Tags_Cycling', 'Ta
gs_Boomer Shooter', 'Tags_Web Publishing', 'Tags_Instrumental Music', 'Tags_Jump Scare', 'Tags
_Skating', 'Tags_Bowling', 'Tags_Skiing', 'Tags_Football (Soccer)', 'Tags_TrackIR', 'Tags_Rock
Music', 'Tags_Boss Rush', 'Tags_Snowboarding', 'Tags_Musou', 'Tags_8-bit Music', 'Tags_BMX', '
Tags_Documentary', 'Tags_Job Simulator', 'Tags_Masterpiece', 'Tags_Voice Control', 'Tags_Elect
ronic', 'Tags_ATV', 'Tags_Well-Written', 'Tags_Cricket', 'Tags_Football (American)', 'Tags_Lar
a Croft', 'Tags_Hardware', 'Tags_Reboot', 'Tags_Benchmark', 'Tags_Volleyball', 'Tags_Feature F
ilm', 'Tags_Steam Machine', 'Tags_Shop Keeper', 'Tags_Mahjong', 'Tags_Birds', 'Tags_Extraction
Shooter', 'Tags_Rugby', 'Tags_Coding', 'Tags_Fox', 'Tags_Dwarf', 'Tags_Hobby Sim', 'Tags_Tile-
Matching', 'Tags_Batman', 'Tags_Elf', 'Tags_Snooker', 'Genres_Action', 'Genres_Casual', 'Genre
s_Adventure', 'Genres_Simulation', 'Genres_Strategy', 'Genres_RPG', 'Genres_Early Access', 'Ge
nres_Free to Play', 'Genres_Sports', 'Genres_Racing', 'Genres_Massively Multiplayer', 'Genres_
Violent', 'Genres_Gore', 'Genres_Utilities', 'Genres_Design & Illustration', 'Genres_Animation
& Modeling', 'Genres_Nudity', 'Genres_Sexual Content', 'Genres_Education', 'Genres_Video Produ
ction', 'Genres_Game Development', 'Genres_Audio Production', 'Genres_Software Training', 'Gen
res_Web Publishing', 'Genres_Photo Editing', 'Genres_Accounting', 'Genres_Movie', 'Genres_Docu
mentary', 'Genres_Episodic', 'Genres_Short', 'Genres_Tutorial', 'Genres_360 Video', 'Genres_Fr
ee To Play', 'Categories_Single-player', 'Categories_Steam Achievements', 'Categories_Steam Cl
oud', 'Categories_Full controller support', 'Categories_Multi-player', 'Categories_Steam Tradi
ng Cards', 'Categories_Partial Controller Support', 'Categories_PvP', 'Categories_Co-op', 'Cat
egories_Steam Leaderboards', 'Categories_Online PvP', 'Categories_Remote Play Together', 'Cate
gories_Shared/Split Screen', 'Categories_Online Co-op', 'Categories_Shared/Split Screen PvP',
'Categories_Stats', 'Categories_Family Sharing', 'Categories_Shared/Split Screen Co-op', 'Cate
gories_Remote Play on TV', 'Categories_Cross-Platform Multiplayer', 'Categories_Includes level
editor', 'Categories_Steam Workshop', 'Categories_In-App Purchases', 'Categories_Captions avai
lable', 'Categories_Remote Play on Tablet', 'Categories_MMO', 'Categories_Remote Play on Phone
', 'Categories_LAN PvP', 'Categories_LAN Co-op', 'Categories_VR Only', 'Categories_VR Support'
, 'Categories_Commentary available', 'Categories_Tracked Controller Support', 'Categories_Valv
e Anti-Cheat enabled', 'Categories_Steam Turn Notifications', 'Categories_VR Supported', 'Cate
gories_Includes Source SDK', 'Categories_SteamVR Collectibles', 'Categories_HDR available', 'C
ategories_Mods', 'Categories_Mods (require HL2)', 'Developers freq', 'Publishers freq']
```

In [11]:

```python
columns = df.columns.tolist()
print(columns)
```

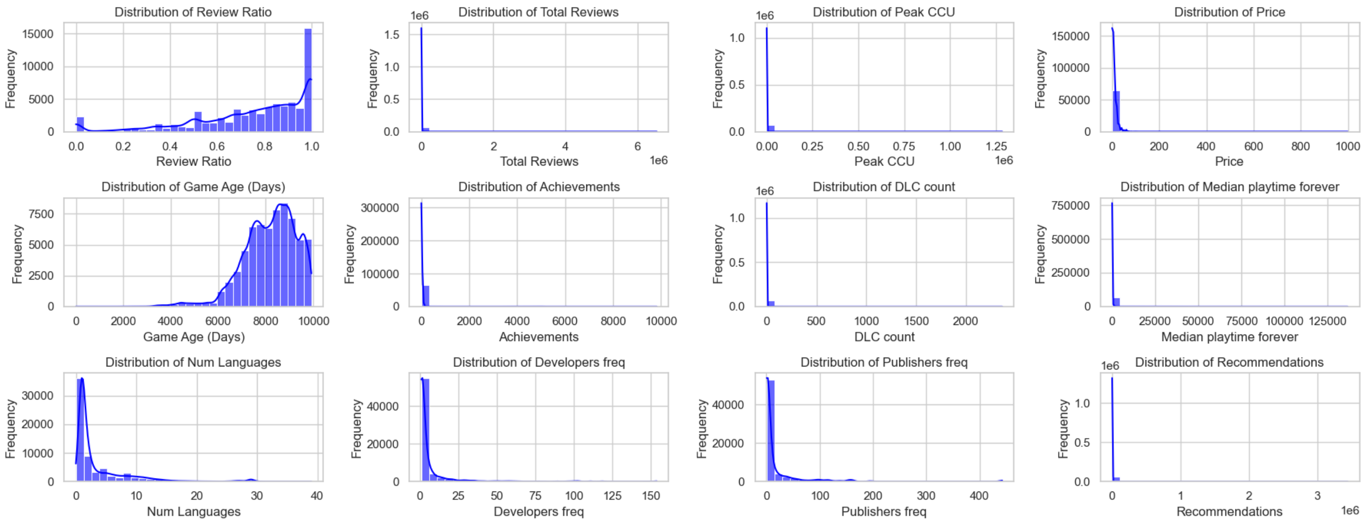
```
['AppID', 'Name', 'Peak CCU', 'Required age', 'Price', 'DLC count', 'About the game', 'Windows
', 'Mac', 'Linux', 'Achievements', 'Recommendations', 'Average playtime two weeks', 'Median pl
aytime forever', 'Developers', 'Publishers', 'Game Age (Days)', 'is_indie', 'Owner range', 'To
tal Reviews', 'Review Ratio', 'Num Languages', 'Is English Supported', 'Tags_Singleplayer', 'T
ags_Action', 'Tags_Casual', 'Tags_Adventure', 'Tags_2D', 'Tags_Strategy', 'Tags_Simulation', '
Tags_RPG', 'Tags_Puzzle', 'Tags_Atmospheric', 'Tags_3D', 'Tags_Early Access', 'Tags_Pixel Grap
hics', 'Tags_Story Rich', 'Tags_Colorful', 'Tags_Exploration', 'Tags_Cute', 'Tags_First-Perso
n', 'Tags_Arcade', 'Tags_Multiplayer', 'Tags_Fantasy', 'Tags_Funny', 'Tags_Shooter', 'Tags_Hor
ror', 'Tags_Retro', 'Tags_Platformer', 'Tags_Anime', 'Tags_Family Friendly', 'Tags_Sci-fi', 'T
ags_Action-Adventure', 'Tags_Relaxing', 'Tags_Female Protagonist', 'Tags_Difficult', 'Tags_Thi
rd Person', 'Tags_VR', 'Tags_Survival', 'Tags_Top-Down', 'Tags_Open World', 'Tags_Stylized', '
Tags_Controller', 'Tags_Combat', 'Tags_Great Soundtrack', 'Tags_Comedy', 'Tags_2D Platformer',
 'Tags_Visual Novel', 'Tags_FPS', 'Tags_Violent', 'Tags_Mystery', 'Tags_Co-op', 'Tags_Dark', 'T
ags_Physics', 'Tags_Realistic', 'Tags_Minimalist', 'Tags_Free to Play', 'Tags_Cartoony', 'Tags
_Psychological Horror', 'Tags_Choices Matter', 'Tags_Point & Click', 'Tags_Linear', 'Tags_Gore
', 'Tags_PvP', 'Tags_Sandbox', 'Tags_Multiple Endings', 'Tags_Sports', 'Tags_Side Scroller', '
Tags_Space', 'Tags_Rogue-like', 'Tags_Old School', 'Tags_Tactical', 'Tags_PvE', 'Tags_Rogue-li
te', 'Tags_Building', 'Tags_Puzzle-Platformer', 'Tags_Management', 'Tags_Character Customizati
on', 'Tags_Hidden Object', 'Tags_Hand-drawn', 'Tags_Action RPG', 'Tags_Sexual Content', 'Tags_
Magic', 'Tags_Logic', 'Tags_Racing', 'Tags_Nudity', 'Tags_Local Multiplayer', "Tags_Shoot 'Em
Up", 'Tags_3D Platformer', 'Tags_Procedural Generation', 'Tags_Turn-Based Strategy', 'Tags_Sur
vival Horror', 'Tags_Cartoon', 'Tags_Futuristic', 'Tags_Bullet Hell', 'Tags_Medieval', 'Tags_T
urn-Based Combat', 'Tags_Online Co-Op', 'Tags_Crafting', 'Tags_Turn-Based Tactics', 'Tags_Walk
ing Simulator', 'Tags_Drama', 'Tags_Interactive Fiction', 'Tags_Choose Your Own Adventure', 'T
ags_Hack and Slash', 'Tags_Action Roguelike', "Tags_1990's", 'Tags_Resource Management', 'Tags
_Zombies', 'Tags_Local Co-Op', 'Tags_Score Attack', 'Tags_JRPG', 'Tags_Dungeon Crawler', 'Tags
_Replay Value', 'Tags_Dark Fantasy', 'Tags_Surreal', 'Tags_War', 'Tags_Historical', 'Tags_Imme
rsive Sim', 'Tags_Turn-Based', 'Tags_Post-apocalyptic', 'Tags_Top-Down Shooter', 'Tags_Nature'
, 'Tags_Emotional', 'Tags_Base-Building', 'Tags_Stealth', 'Tags_Isometric', 'Tags_Education',
 'Tags_Romance', 'Tags_Text-Based', 'Tags_Fast-Paced', 'Tags_Massively Multiplayer', 'Tags_Shor
t', 'Tags_Card Game', 'Tags_Abstract', 'Tags_Classic', 'Tags_Clicker', 'Tags_RTS', 'Tags_Preci
sion Platformer', 'Tags_Military', 'Tags_1980s', 'Tags_2.5D', 'Tags_Third-Person Shooter', 'Ta
gs_Tower Defense', 'Tags_Investigation', 'Tags_Board Game', 'Tags_Detective', 'Tags_Memes', 'T
ags_RPGMaker', 'Tags_Narration', 'Tags_Dating Sim', 'Tags_Robots', 'Tags_Aliens', 'Tags_Cyberp
unk', 'Tags_Perma Death', 'Tags_Tabletop', 'Tags_Driving', 'Tags_Cinematic', 'Tags_Life Sim',
 'Tags_Dark Humor', 'Tags_Economy', 'Tags_Arena Shooter', 'Tags_Time Management', 'Tags_VR Only
', 'Tags_Mature', 'Tags_4 Player Local', 'Tags_Flight', 'Tags_Psychological', 'Tags_Thriller',
 'Tags_Strategy RPG', 'Tags_Real Time Tactics', 'Tags_Demons', 'Tags_City Builder', 'Tags_Exper
imental', "Tags_Beat 'em up", 'Tags_Conversation', 'Tags_Psychedelic', 'Tags_Runner', 'Tags_Fi
ghting', 'Tags_Metroidvania', 'Tags_LGBTQ+', 'Tags_Tactical RPG', 'Tags_Wargame', 'Tags_Nonlin
ear', 'Tags_Supernatural', 'Tags_Music', 'Tags_Team-Based', 'Tags_Collectathon', 'Tags_Tutoria
l', 'Tags_Competitive', 'Tags_Lore-Rich', 'Tags_Level Editor', 'Tags_Artificial Intelligence',
 'Tags_Idler', 'Tags_Comic Book', 'Tags_Loot', 'Tags_Twin Stick Shooter', 'Tags_Automobile Sim'
, 'Tags_Party-Based RPG', 'Tags_Dystopian', 'Tags_Parkour', 'Tags_Modern', 'Tags_Destruction',
 'Tags_Grid-Based Movement', 'Tags_2D Fighter', 'Tags_Souls-like', 'Tags_Hentai', 'Tags_Match 3
', 'Tags_Cats', 'Tags_Alternate History', 'Tags_Rhythm', 'Tags_Deckbuilding', 'Tags_CRPG', 'Ta
gs_Inventory Management', 'Tags_Moddable', 'Tags_Crime', 'Tags_Mythology', 'Tags_Space Sim', '
Tags_Card Battler', 'Tags_Beautiful', 'Tags_Soundtrack', 'Tags_Grand Strategy', 'Tags_World Wa
r II', 'Tags_Philosophical', 'Tags_Dark Comedy', 'Tags_Science', 'Tags_Noir', 'Tags_Mystery Du
ngeon', 'Tags_3D Fighter', 'Tags_Character Action Game', 'Tags_Lovecraftian', 'Tags_NSFW', 'Ta
gs_Word Game', 'Tags_Split Screen', 'Tags_Swordplay', 'Tags_Colony Sim', 'Tags_Farming Sim', '
Tags_Automation', 'Tags_Mouse only', 'Tags_Creature Collector', 'Tags_e-sports', 'Tags_6DOF',
 'Tags_Dragons', 'Tags_Voxel', 'Tags_3D Vision', 'Tags_Vehicular Combat', 'Tags_Solitaire', 'Ta
gs_Bullet Time', 'Tags_Hero Shooter', 'Tags_Mechs', 'Tags_Battle Royale', 'Tags_Agriculture',
 'Tags_Capitalism', 'Tags_Parody', 'Tags_Combat Racing', 'Tags_Open World Survival Craft', 'Tag
s_Spectacle fighter', 'Tags_America', 'Tags_Design & Illustration', 'Tags_Blood', 'Tags_Time M
anipulation', 'Tags_Gun Customization', 'Tags_Class-Based', 'Tags_Sokoban', 'Tags_MMORPG', 'Ta
gs_Steampunk', 'Tags_Hex Grid', 'Tags_Addictive', 'Tags_God Game', 'Tags_Political', 'Tags_FMV
', 'Tags_Conspiracy', 'Tags_Martial Arts', 'Tags_Ninja', 'Tags_Gothic', 'Tags_Co-op Campaign',
 'Tags_Pirates', 'Tags_Otome', 'Tags_Tanks', 'Tags_Auto Battler', 'Tags_Trading', 'Tags_Undergr
ound', 'Tags_Satire', 'Tags_Real-Time', 'Tags_Utilities', 'Tags_Quick-Time Events', 'Tags_Dog'
, 'Tags_Mining', 'Tags_Cooking', 'Tags_Time Travel', 'Tags_Looter Shooter', 'Tags_4X', 'Tags_R
emake', 'Tags_Underwater', 'Tags_Hacking', 'Tags_Hunting', 'Tags_Programming', 'Tags_Dynamic N
arration', 'Tags_Cult Classic', 'Tags_Dinosaurs', 'Tags_Fishing', 'Tags_Politics', 'Tags_Escap
e Room', 'Tags_Real-Time with Pause', 'Tags_Faith', 'Tags_Naval', 'Tags_Vampire', 'Tags_Miniga
mes', 'Tags_Software', 'Tags_Political Sim', 'Tags_Western', 'Tags_Superhero', 'Tags_Narrative
', 'Tags_Party Game', 'Tags_Transportation', 'Tags_Trading Card Game', 'Tags_Assassin', 'Tags_
Immersive', 'Tags_Action RTS', 'Tags_Illuminati', 'Tags_Cozy', 'Tags_Touch-Friendly', 'Tags_MO
BA', 'Tags_Time Attack', 'Tags_Typing', 'Tags_Game Development', 'Tags_Asynchronous Multiplaye
r', 'Tags_Trivia', 'Tags_Trains', 'Tags_On-Rails Shooter', 'Tags_Roguelike Deckbuilder', 'Tags
_Cold War', 'Tags_Party', 'Tags_Snow', 'Tags_Animation & Modeling', 'Tags_Traditional Roguelik
e', 'Tags_Archery', 'Tags_Offroad', 'Tags_Heist', 'Tags_Naval Combat', 'Tags_Diplomacy', 'Tags
_Football', 'Tags_Soccer', 'Tags_Villain Protagonist', 'Tags_Music-Based Procedural Generation
', 'Tags_Kickstarter', 'Tags_Sailing', 'Tags_Mars', 'Tags_Wholesome', 'Tags_Chess', 'Tags_Fore
ign', 'Tags_Experience', 'Tags_Nostalgia', 'Tags_Gambling', 'Tags_Sequel', 'Tags_Sniper', 'Tag
s_GameMaker', 'Tags_Boxing', 'Tags_Horses', 'Tags_Episodic', 'Tags_World War I', 'Tags_Golf',
'Tags_Spelling', 'Tags_Jet', 'Tags_Unforgiving', 'Tags_Motorbike', 'Tags_Outbreak Sim', 'Tags_
Transhumanism', 'Tags_Werewolves', 'Tags_Rome', 'Tags_Pinball', 'Tags_Bikes', 'Tags_Farming',
'Tags_Silent Protagonist', 'Tags_Epic', 'Tags_Roguevania', 'Tags_Spaceships', 'Tags_Basketball
', 'Tags_Asymmetric VR', 'Tags_Submarine', 'Tags_Crowdfunded', 'Tags_Medical Sim', 'Tags_Audio
Production', 'Tags_LEGO', 'Tags_Video Production', 'Tags_Social Deduction', 'Tags_Games Worksh
op', 'Tags_Mini Golf', 'Tags_Vikings', 'Tags_Based On A Novel', 'Tags_Movie', 'Tags_Electronic
Music', 'Tags_Ambient', 'Tags_Baseball', 'Tags_Software Training', 'Tags_Wrestling', 'Tags_360
Video', 'Tags_Dungeons & Dragons', 'Tags_Tennis', 'Tags_Warhammer 40K', 'Tags_Pool', 'Tags_Gam
ing', 'Tags_Mod', 'Tags_Motocross', 'Tags_Lemmings', 'Tags_Intentionally Awkward Controls', 'T
ags_Photo Editing', 'Tags_Skateboarding', 'Tags_Hockey', 'Tags_Cycling', 'Tags_Boomer Shooter'
, 'Tags_Web Publishing', 'Tags_Instrumental Music', 'Tags_Jump Scare', 'Tags_Skating', 'Tags_B
owling', 'Tags_Skiing', 'Tags_Football (Soccer)', 'Tags_TrackIR', 'Tags_Rock Music', 'Tags_Bos
s Rush', 'Tags_Snowboarding', 'Tags_Musou', 'Tags_8-bit Music', 'Tags_BMX', 'Tags_Documentary'
, 'Tags_Job Simulator', 'Tags_Masterpiece', 'Tags_Voice Control', 'Tags_Electronic', 'Tags_ATV
', 'Tags_Well-Written', 'Tags_Cricket', 'Tags_Football (American)', 'Tags_Lara Croft', 'Tags_H
ardware', 'Tags_Reboot', 'Tags_Benchmark', 'Tags_Volleyball', 'Tags_Feature Film', 'Tags_Steam
Machine', 'Tags_Shop Keeper', 'Tags_Mahjong', 'Tags_Birds', 'Tags_Extraction Shooter', 'Tags_R
ugby', 'Tags_Coding', 'Tags_Fox', 'Tags_Dwarf', 'Tags_Hobby Sim', 'Tags_Tile-Matching', 'Tags_
Batman', 'Tags_Elf', 'Tags_Snooker', 'Genres_Action', 'Genres_Casual', 'Genres_Adventure', 'Ge
nres_Simulation', 'Genres_Strategy', 'Genres_RPG', 'Genres_Early Access', 'Genres_Free to Play
', 'Genres_Sports', 'Genres_Racing', 'Genres_Massively Multiplayer', 'Genres_Violent', 'Genres
_Gore', 'Genres_Utilities', 'Genres_Design & Illustration', 'Genres_Animation & Modeling', 'Ge
nres_Nudity', 'Genres_Sexual Content', 'Genres_Education', 'Genres_Video Production', 'Genres_
Game Development', 'Genres_Audio Production', 'Genres_Software Training', 'Genres_Web Publishi
ng', 'Genres_Photo Editing', 'Genres_Accounting', 'Genres_Movie', 'Genres_Documentary', 'Genre
s_Episodic', 'Genres_Short', 'Genres_Tutorial', 'Genres_360 Video', 'Genres_Free To Play', 'Ca
tegories_Single-player', 'Categories_Steam Achievements', 'Categories_Steam Cloud', 'Categorie
s_Full controller support', 'Categories_Multi-player', 'Categories_Steam Trading Cards', 'Cate
gories_Partial Controller Support', 'Categories_PvP', 'Categories_Co-op', 'Categories_Steam Le
aderboards', 'Categories_Online PvP', 'Categories_Remote Play Together', 'Categories_Shared/Sp
lit Screen', 'Categories_Online Co-op', 'Categories_Shared/Split Screen PvP', 'Categories_Stat
s', 'Categories_Family Sharing', 'Categories_Shared/Split Screen Co-op', 'Categories_Remote Pl
ay on TV', 'Categories_Cross-Platform Multiplayer', 'Categories_Includes level editor', 'Categ
ories_Steam Workshop', 'Categories_In-App Purchases', 'Categories_Captions available', 'Catego
ries_Remote Play on Tablet', 'Categories_MMO', 'Categories_Remote Play on Phone', 'Categories_
LAN PvP', 'Categories_LAN Co-op', 'Categories_VR Only', 'Categories_VR Support', 'Categories_C
ommentary available', 'Categories_Tracked Controller Support', 'Categories_Valve Anti-Cheat en
abled', 'Categories_Steam Turn Notifications', 'Categories_VR Supported', 'Categories_Includes
Source SDK', 'Categories_SteamVR Collectibles', 'Categories_HDR available', 'Categories_Mods',
'Categories_Mods (require HL2)', 'Developers freq', 'Publishers freq']
```

## Numerical Predictors

We begin by examining key numerical features that are relevant to game characteristics and popularity

In [12]:
```
numerical_col = [
    'Review Ratio', 'Total Reviews', 'Peak CCU',
    'Price', 'Game Age (Days)',
    'Achievements', 'DLC count',
    'Median playtime forever',
    'Num Languages',
    'Developers freq', 'Publishers freq', 'Recommendations'
    ]
```

```
plt.figure(figsize=(6 * 3, 2 * 3.5))

# Plotting histograms for log-transformed numerical predictors before and after log transforma
for i, col in enumerate(numerical_col):
    plt.subplot(3, 4, i + 1)
    sns.histplot(df[col], kde=True, bins=30, color='blue', alpha=0.6)
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
plt.tight_layout()
plt.show()

print("\nSummary Statistics for Numerical Predictors:")
display(df[numerical_col].describe())
```



```
Summary Statistics for Numerical Predictors:
```

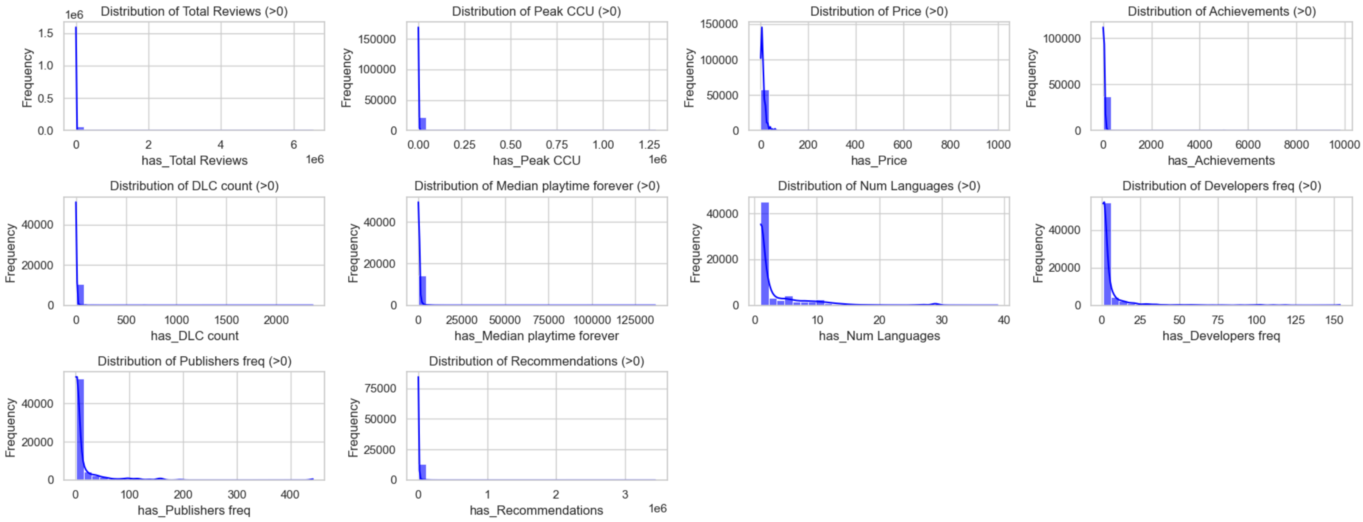
| | Review Ratio | Total Reviews | Peak CCU | Price | Game Age (Days) | Achievements | DLC cou |
|---|---|---|---|---|---|---|---|
| **count** | 66249.000000 | 6.624900e+04 | 6.624900e+04 | 66249.000000 | 66138.000000 | 66249.000000 | 66249.0000 |
| **mean** | 0.756041 | 1.438099e+03 | 2.354716e+02 | 7.964666 | 8178.903248 | 24.058567 | 0.6450 |
| **std** | 0.248092 | 3.188796e+04 | 9.316125e+03 | 10.623297 | 1101.415989 | 193.935879 | 15.5019 |
| **min** | 0.000000 | 0.000000e+00 | 0.000000e+00 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| **25%** | 0.634441 | 5.000000e+00 | 0.000000e+00 | 1.990000 | 7498.000000 | 0.000000 | 0.0000 |
| **50%** | 0.821904 | 2.100000e+01 | 0.000000e+00 | 4.990000 | 8325.000000 | 6.000000 | 0.0000 |
| **75%** | 0.959184 | 1.130000e+02 | 1.000000e+00 | 9.990000 | 8993.000000 | 21.000000 | 0.0000 |
| **max** | 1.000000 | 6.531097e+06 | 1.284268e+06 | 999.000000 | 9924.000000 | 9821.000000 | 2366.0000 |

Overall, it can be seen that Total Reviews, Peak CCU, Price, Achievements, DLC count, Median Playtime forever, Num Languages, Developers_freq, Publishers_freq and Recommendations are all very heavily right skewed due to the very high number of low values (Like no DLC, no Achievements, ...) in the dataset

To check if that is the case, we'll create new features has_X to seperate the 0s from the other values and plot it seperately

In [13]:
```
skewed_col = ['Total Reviews', 'Peak CCU', 'Price', 'Achievements', 'DLC count', 'Median playt

# Add has_x for each skewed column
for col in skewed_col:
    df[f'has_{col}'] = df[col].apply(lambda x: 1 if x > 0 else 0)
```

```python
plt.figure(figsize=(6 * 3, 2 * 3.5))

# Plot histogram for only columns that has x
for i, col in enumerate(skewed_col):
    plt.subplot(3, 4, i + 1)
    sns.histplot(df[col][df[f'has_{col}'] == 1], kde=True, bins=30, color='blue', alpha=0.6)
    plt.title(f'Distribution of {col} (>0)')
    plt.xlabel(f'has_{col}')
    plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



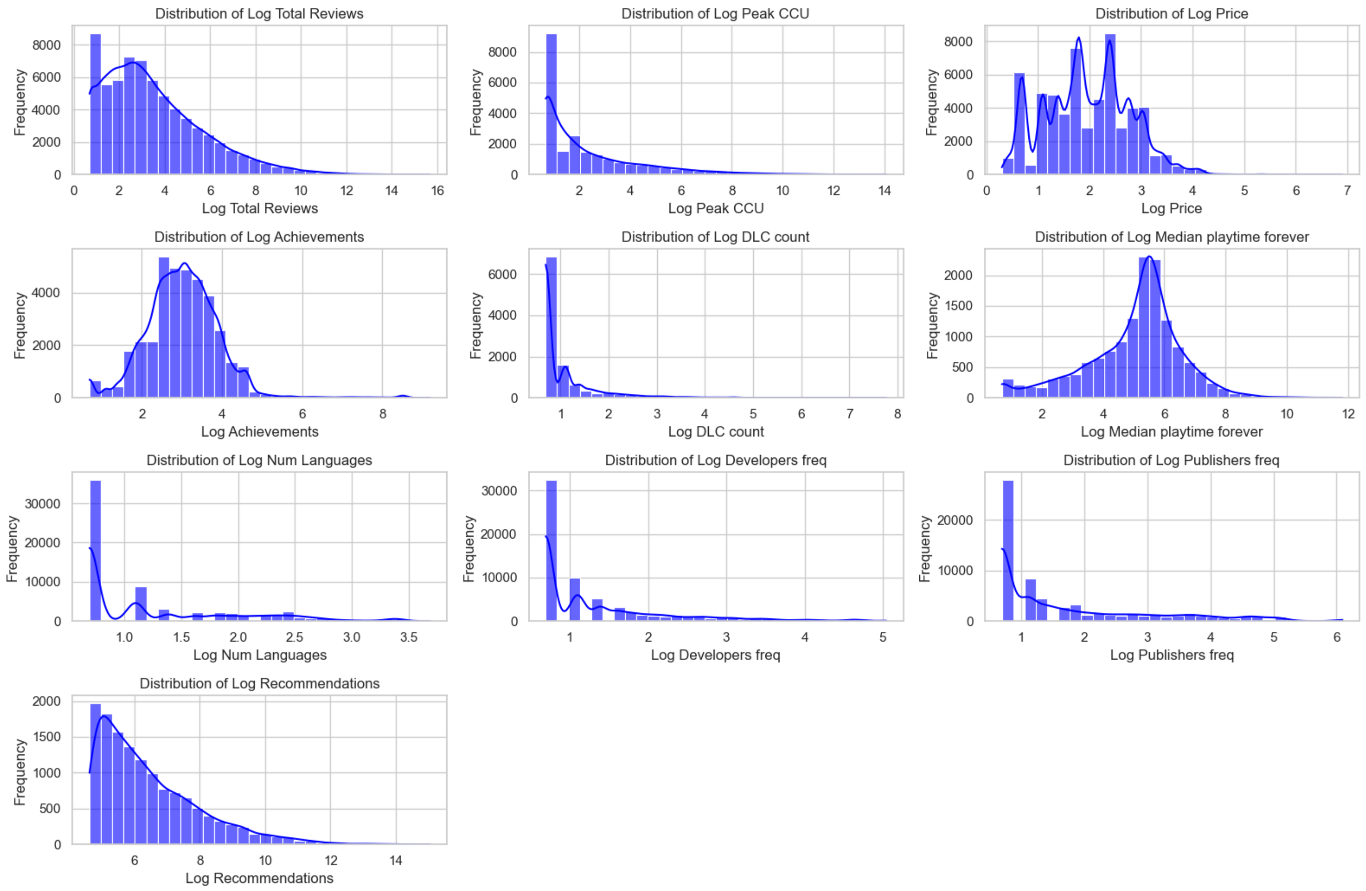
As seen here, it still quite heavily right skewed.

To resolve this, we will be applying log1p (log(1 + x)) on these columns. Doing so will heavily compress the high ends and avoid errors with zero values.

However, 0 will be treated very differently from other small positive numbers, example (log(1+0)=0, but log(1+1=0.69). But is acceptable here since most our 0 are distinct states like no players, or no DLC.

In [14]:

```python
for col in skewed_col:
    df[f'Log {col}'] = np.log1p(df[col])

plt.figure(figsize=(4 * 4, 3 * 3.5))

for i, col in enumerate(skewed_col):
    plt.subplot(4, 3, i + 1)
    sns.histplot(df[f"Log {col}"][df[f'has_{col}'] == 1], kde=True, bins=30, color='blue', alp
    plt.title(f'Distribution of Log {col}')
    plt.xlabel(f'Log {col}')
    plt.ylabel('Frequency')
plt.tight_layout()

plt.show()
display(df[skewed_col].describe())
```



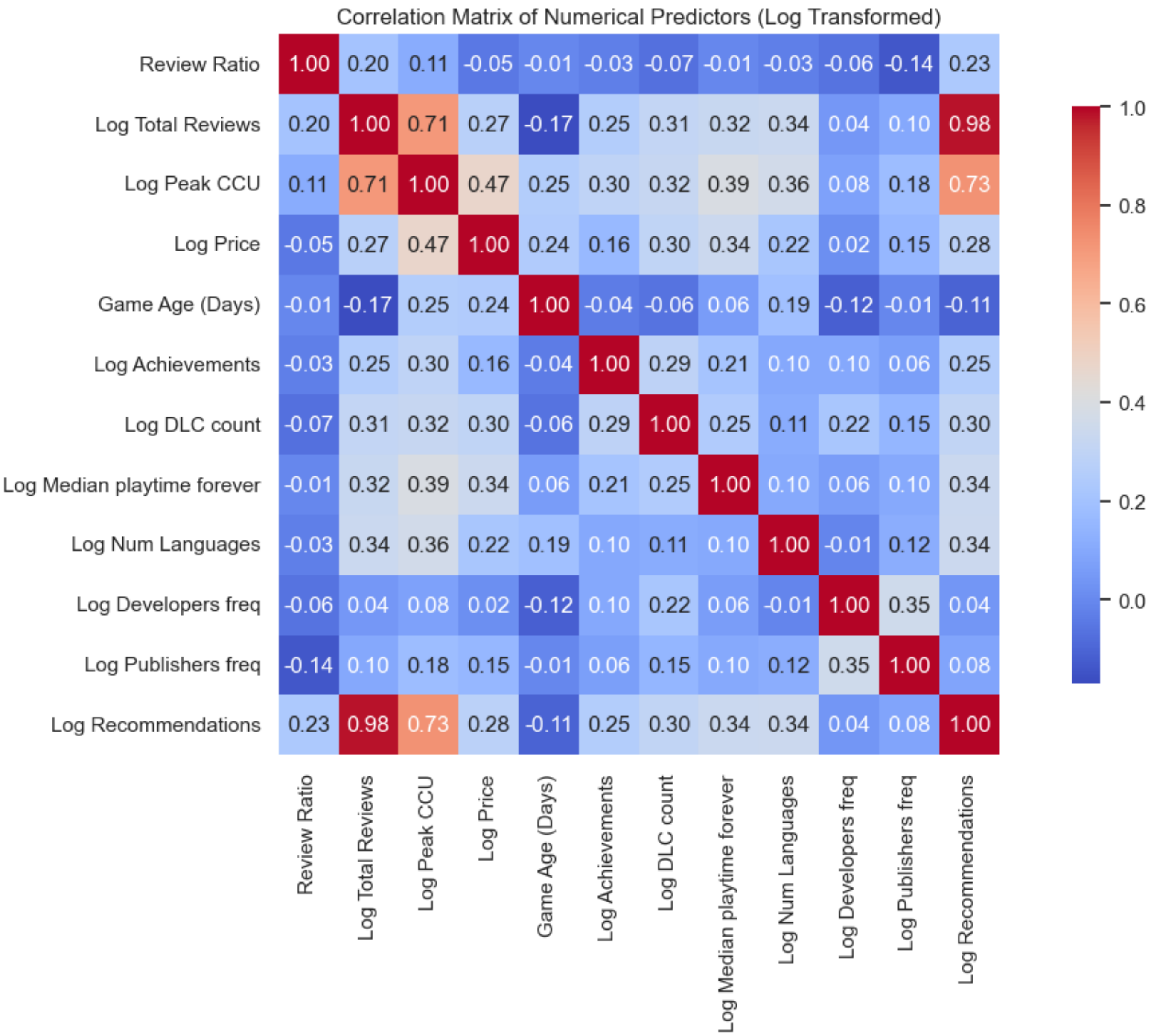
| | Total Reviews | Peak CCU | Price | Achievements | DLC count | Median playtime forever | N Langua |
|---|---|---|---|---|---|---|---|
| **count** | 6.624900e+04 | 6.624900e+04 | 66249.000000 | 66249.000000 | 66249.000000 | 66249.000000 | 66249.000 |
| **mean** | 1.438099e+03 | 2.354716e+02 | 7.964666 | 24.058567 | 0.645006 | 106.335598 | 3.612 |
| **std** | 3.188796e+04 | 9.316125e+03 | 10.623297 | 193.935879 | 15.501994 | 1267.863825 | 4.961 |
| **min** | 0.000000e+00 | 0.000000e+00 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000 |
| **25%** | 5.000000e+00 | 0.000000e+00 | 1.990000 | 0.000000 | 0.000000 | 0.000000 | 1.000 |
| **50%** | 2.100000e+01 | 0.000000e+00 | 4.990000 | 6.000000 | 0.000000 | 0.000000 | 1.000 |
| **75%** | 1.130000e+02 | 1.000000e+00 | 9.990000 | 21.000000 | 0.000000 | 0.000000 | 4.000 |
| **max** | 6.531097e+06 | 1.284268e+06 | 999.000000 | 9821.000000 | 2366.000000 | 136629.000000 | 39.000 |

The skewness has improved slightly, although still slightly skewed and imbalanced for some, next, we'll be plotting a correlation matrix among the predictors, to see which is best suitable for describing the number Recommendations a game gets.

In [15]:

```
new_numical_col = []
# create a new df where only all has_x columns are 1

# Creating a new df where only have rows that has all the values > 0, this is to ensure the co
new_df = df.copy()
for col in numerical_col:
    if col in skewed_col:
        new_df = new_df[new_df[f'has_{col}'] == 1]
        new_numical_col.append(f'Log {col}')
    else:
        new_numical_col.append(col)

# Calculating the correlation matrix of numerical predictors after log transformation
cor_matr = new_df[new_numical_col].corr()
```

```
# Plotting the correlation matrix of numerical predictors after log transformation
plt.figure(figsize=(12, 8))
sns.heatmap(cor_matr, annot=True, fmt=".2f", cmap='coolwarm', square=True, cbar_kws={"shrink":
plt.title('Correlation Matrix of Numerical Predictors (Log Transformed)')
plt.tight_layout()
plt.show()

display(cor_matr)
```



| | Review Ratio | Log Total Reviews | Log Peak CCU | Log Price | Game Age (Days) | Log Achievements | Log DLC count | L Med playti fore |
|---|---|---|---|---|---|---|---|---|
| **Review Ratio** | 1.000000 | 0.203457 | 0.109366 | -0.048762 | -0.009396 | -0.029771 | -0.073591 | -0.0085 |
| **Log Total Reviews** | 0.203457 | 1.000000 | 0.707803 | 0.271099 | -0.169572 | 0.253570 | 0.308992 | 0.3245 |
| **Log Peak CCU** | 0.109366 | 0.707803 | 1.000000 | 0.472686 | 0.251939 | 0.295921 | 0.319430 | 0.3938 |
| **Log Price** | -0.048762 | 0.271099 | 0.472686 | 1.000000 | 0.244493 | 0.160714 | 0.300923 | 0.3410 |
| **Game Age (Days)** | -0.009396 | -0.169572 | 0.251939 | 0.244493 | 1.000000 | -0.037372 | -0.063784 | 0.0566 |
| **Log Achievements** | -0.029771 | 0.253570 | 0.295921 | 0.160714 | -0.037372 | 1.000000 | 0.287506 | 0.2093 |
| **Log DLC count** | -0.073591 | 0.308992 | 0.319430 | 0.300923 | -0.063784 | 0.287506 | 1.000000 | 0.2457 |
| **Log Median playtime forever** | -0.008537 | 0.324558 | 0.393869 | 0.341059 | 0.056684 | 0.209354 | 0.245719 | 1.0000 |
| **Log Num Languages** | -0.029964 | 0.335962 | 0.364955 | 0.215668 | 0.191016 | 0.098757 | 0.113189 | 0.1001 |
| **Log Developers freq** | -0.061416 | 0.043589 | 0.075250 | 0.017387 | -0.118148 | 0.097635 | 0.216068 | 0.0633 |
| **Log Publishers freq** | -0.141447 | 0.098684 | 0.177273 | 0.150570 | -0.006976 | 0.064762 | 0.149758 | 0.1026 |
| **Log Recommendations** | 0.231703 | 0.984320 | 0.729956 | 0.280943 | -0.109399 | 0.254534 | 0.301754 | 0.3351 |

What we can find here is that there is actually not a lot of strong variables that can describe Recommendations.

The 2 strongest variables are Log Total Reviews at 0.98 and Peak CCU at 0.73. Total Reviews is not suitable to be used here since whether a game is recommended is just dependent on total number reviews and values are roughly the same.

Peak CCU is useful however, as it suggests games that has high number of concurrent users at one point, showing it's popularity, is more likely to recommend the games. While other numeric variables are only correlated by around 0.3

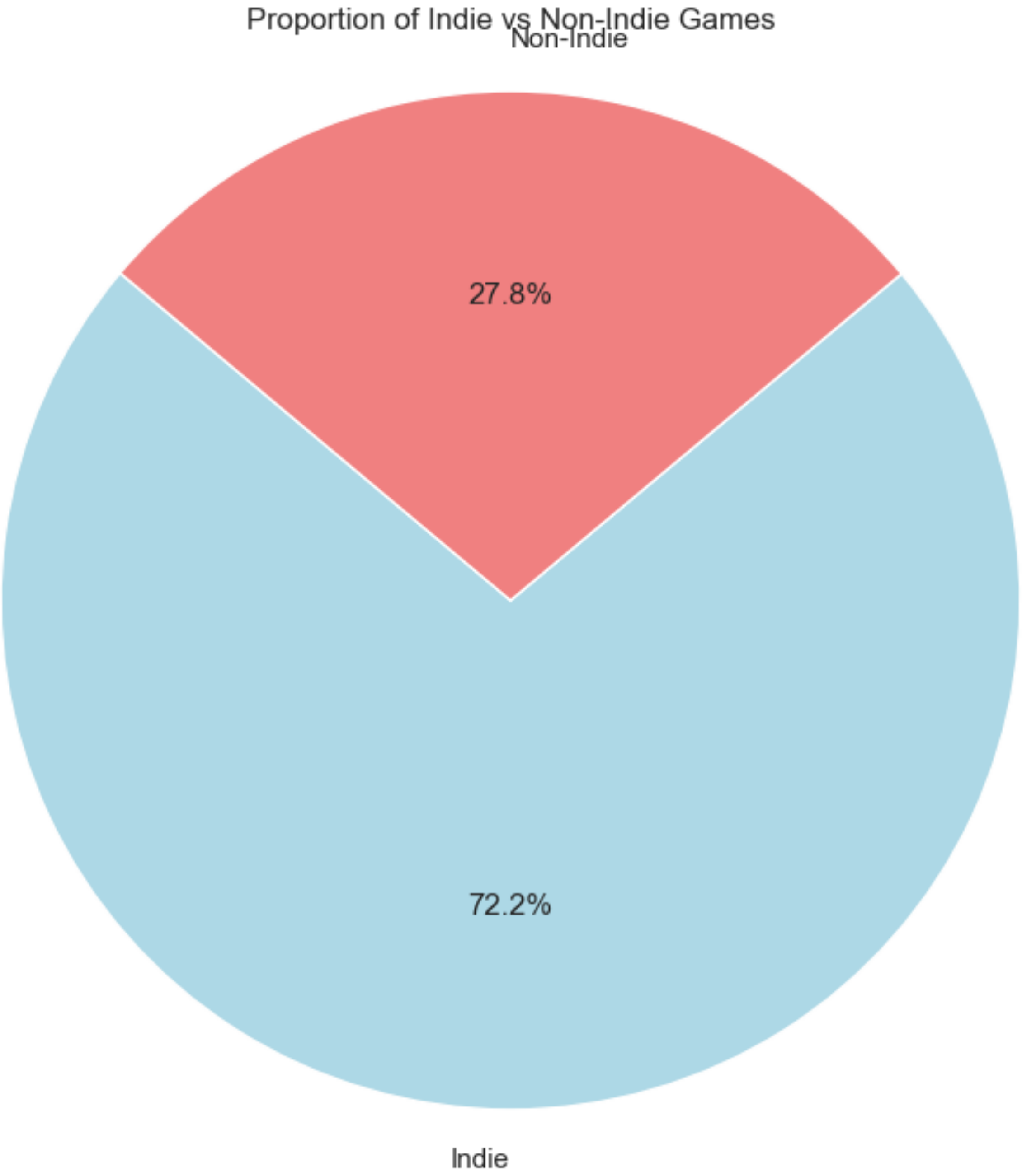
Therefore, the only useful numerical predictor we can utilize for providing **recommendations** here is **Peak CCU**.

## Categorical Predictors

Moving on to categorical variables, we want to start by identifying patterns in the tags, genres and categories of games starting off with frequencies of indie and non-indie games and tags

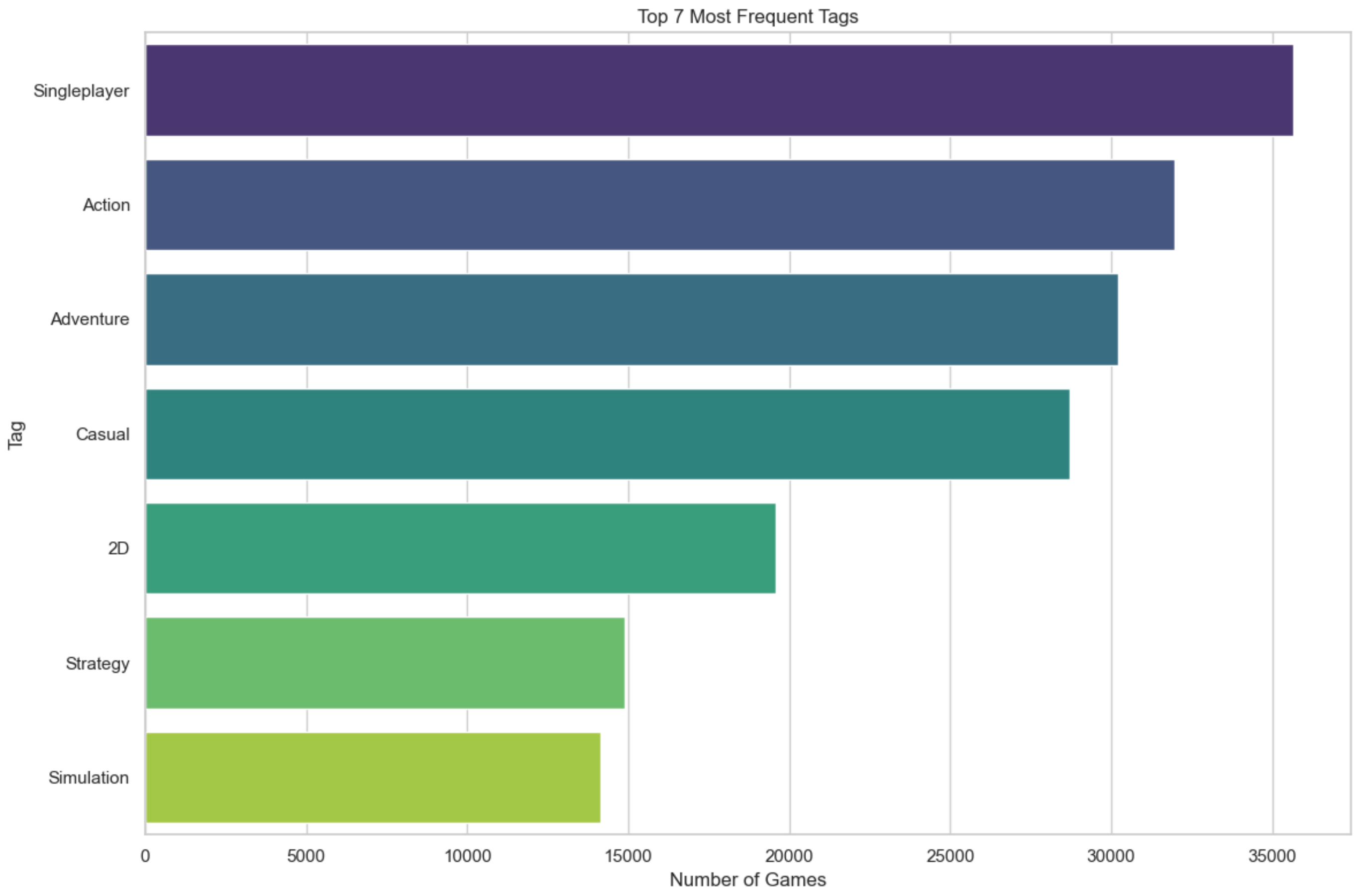
In [16]:

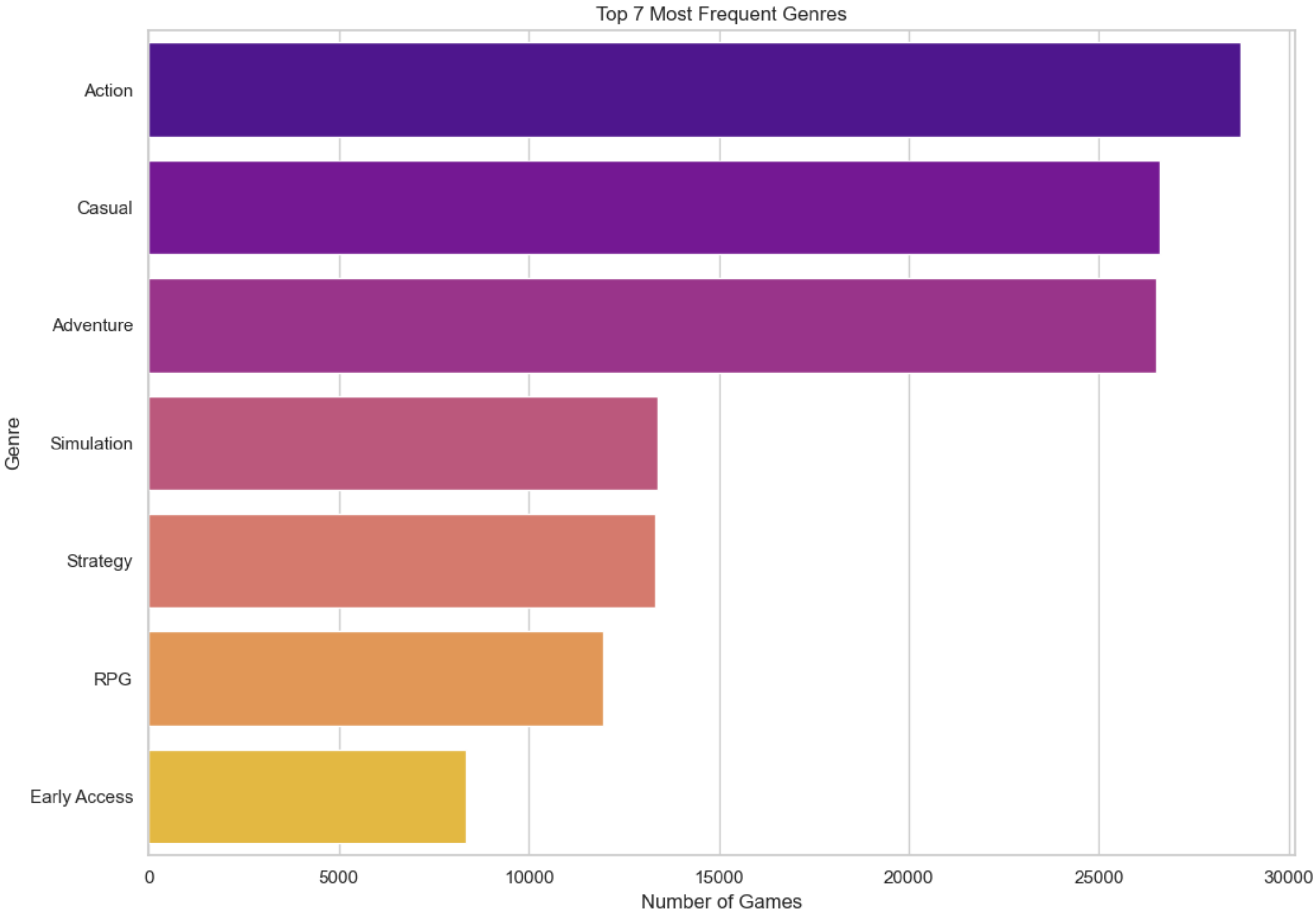
```python
# pie chart of indie vs non-indie games
indie_count = df['is_indie'].value_counts()
indie_count.index = ['Indie', 'Non-Indie']
plt.figure(figsize=(8, 8))
plt.pie(indie_count, labels=indie_count.index, autopct='%1.1f%%', startangle=140, colors=['lig
plt.title('Proportion of Indie vs Non-Indie Games')
plt.axis('equal')
plt.show()
```

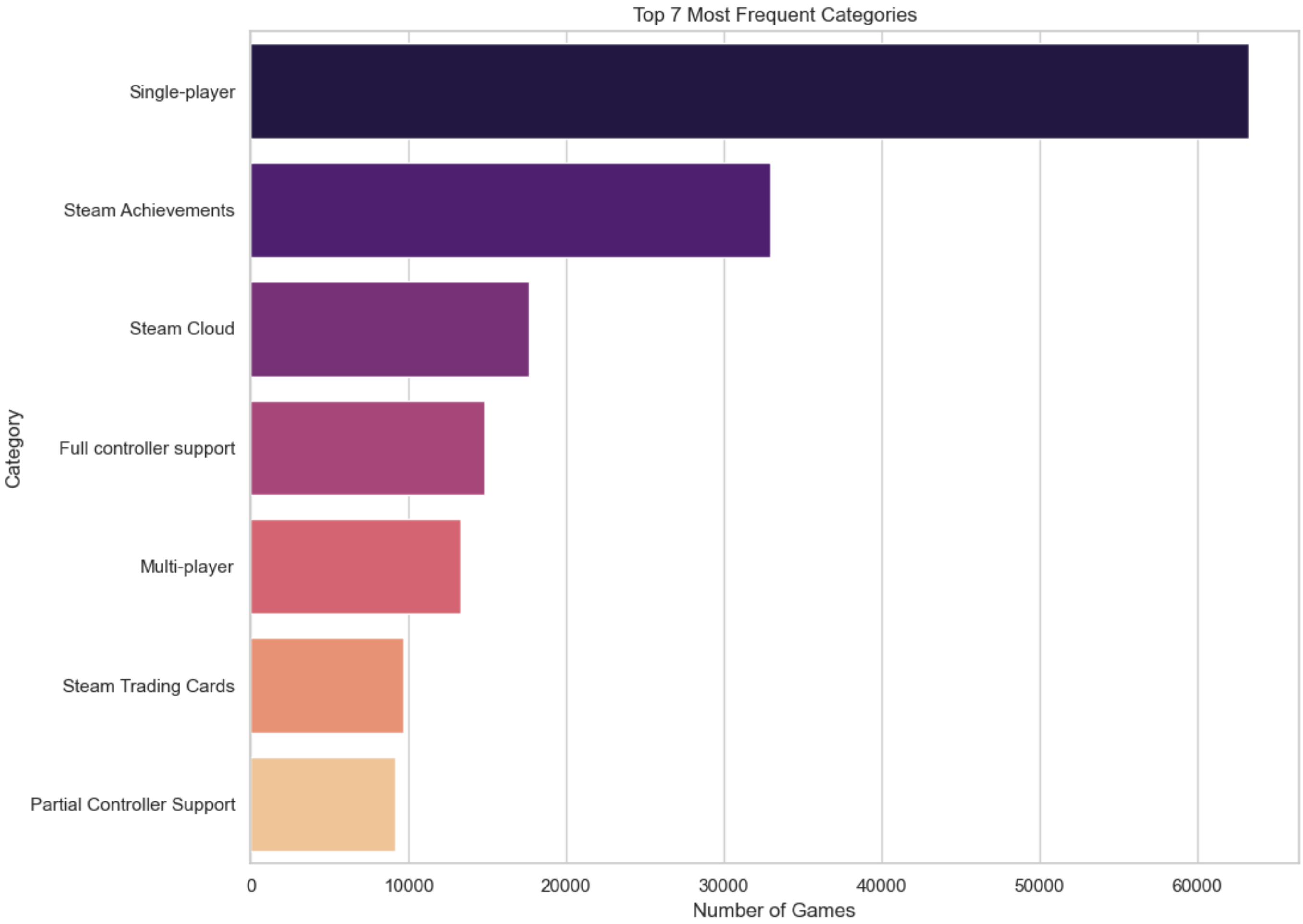


The graph actually shows that the majority games are indie games, meaning games developed by independent developers. Our recommendation model will have to take into account this information when recommending games.

In [17]:
```
tag_cols = [col for col in df.columns if col.startswith('Tags_')]
genre_cols = [col for col in df.columns if col.startswith('Genres_')]
category_cols = [col for col in df.columns if col.startswith('Categories_')]

# Get number of each tags, genres, and categories sorted by frequency
tag_freq = df[tag_cols].sum().sort_values(ascending=False)
genre_freq = df[genre_cols].sum().sort_values(ascending=False)
category_freq = df[category_cols].sum().sort_values(ascending=False)

N = 7 # Number of top tags, genres, and categories to display

plt.figure(figsize=(12, 8))
sns.barplot(y=tag_freq.index[:N], x=tag_freq.values[:N], palette='viridis', hue=tag_freq.index
plt.title(f'Top {N} Most Frequent Tags')
plt.xlabel('Number of Games')
plt.ylabel('Tag')
plt.yticks(ticks=range(N), labels=[tag.split('_')[1] for tag in tag_freq.index[:N]])
plt.tight_layout()
plt.legend([],[], frameon=False) # Turn off legends since labels shown in y axis
plt.show()
plt.figure(figsize=(12, 8))
sns.barplot(y=genre_freq.index[:N], x=genre_freq.values[:N], palette='plasma', hue=genre_freq.
plt.title(f'Top {N} Most Frequent Genres')
plt.xlabel('Number of Games')
plt.ylabel('Genre')
plt.tight_layout()
plt.yticks(ticks=range(N), labels=[genre.split('_')[1] for genre in genre_freq.index[:N]])
plt.legend([],[], frameon=False)
plt.show()

plt.figure(figsize=(12, 8))
sns.barplot(y=category_freq.index[:N], x=category_freq.values[:N], palette='magma', hue=catego
plt.title(f'Top {N} Most Frequent Categories')
plt.xlabel('Number of Games')
plt.ylabel('Category')
plt.tight_layout()
plt.yticks(ticks=range(N), labels=[category.split('_')[1] for category in category_freq.index[
plt.legend([],[], frameon=False)
plt.show()

print("\nTop 5 Tags:")
display(tag_freq.head(5))
print("\nTop 5 Genres:")
display(genre_freq.head(5))
print("\nTop 5 Categories:")
display(category_freq.head(5))
```







```
Top 5 Tags:
Tags_Singleplayer    35659
Tags_Action          31965
Tags_Adventure       30224
Tags_Casual          28725
Tags_2D              19592
dtype: int64
Top 5 Genres:
```

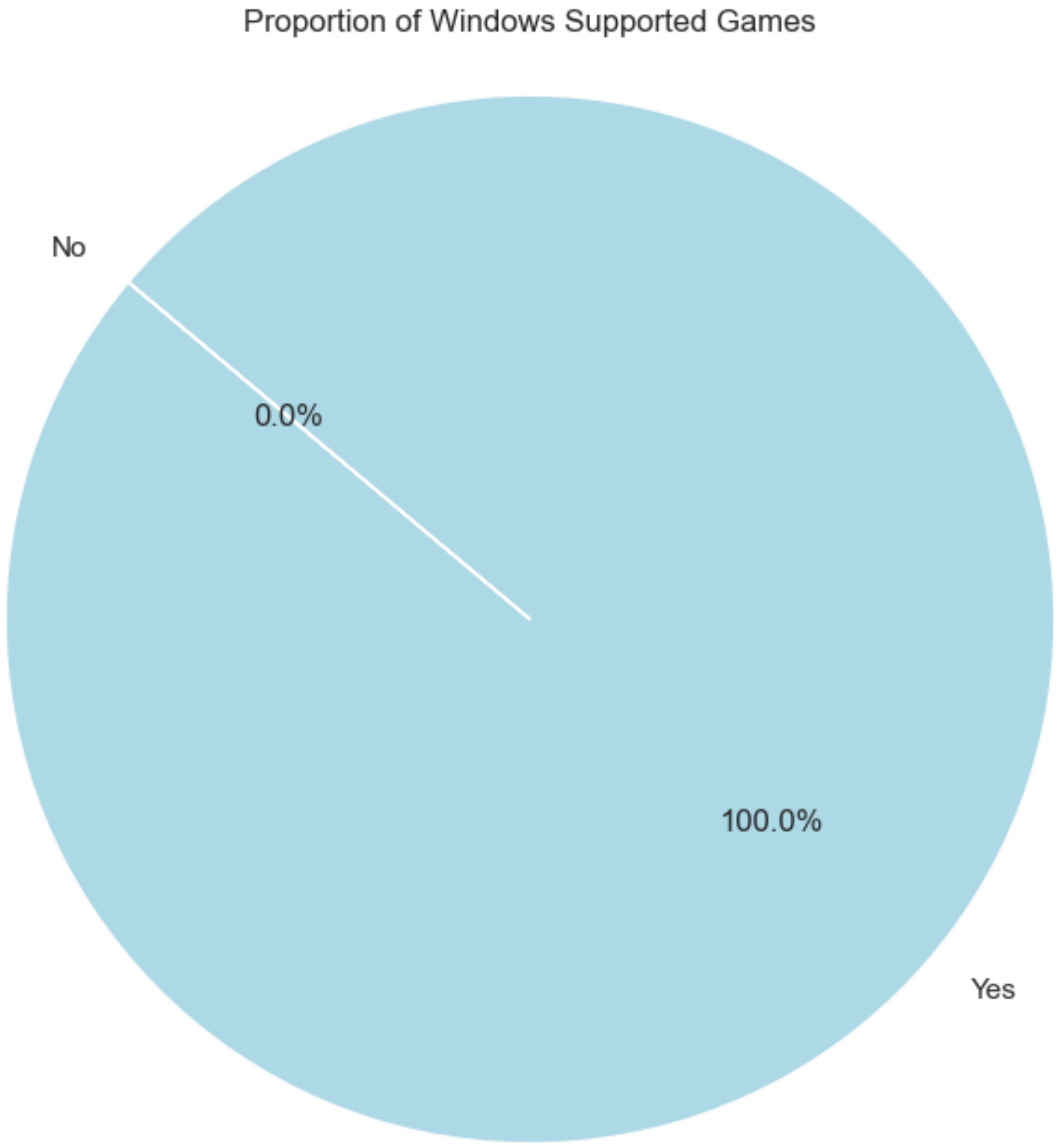
```
Genres_Action        28717
Genres_Casual        26613
Genres_Adventure     26507
Genres_Simulation    13395
Genres_Strategy      13328
dtype: int64
Top 5 Categories:
Categories_Single-player                 63260
Categories_Steam Achievements            32980
Categories_Steam Cloud                   17647
Categories_Full controller support       14820
Categories_Multi-player                  13280
dtype: int64
```

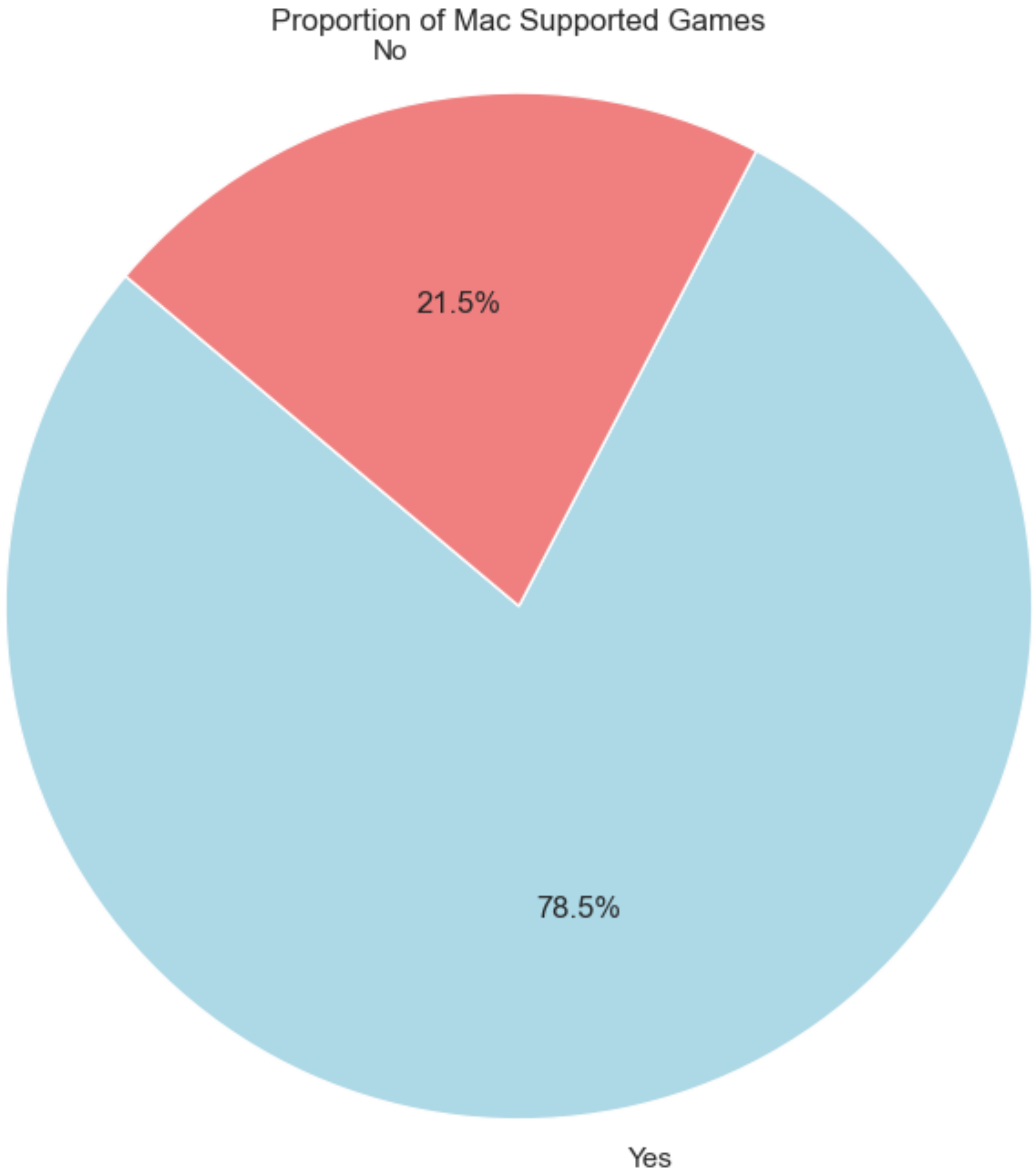
The plots show that 'Singleplayer', 'Adventure', 'Casual', and 'Action' are extremely common tags and genres, and 'Steam Achievements' and 'Steam Cloud' are very common categories.

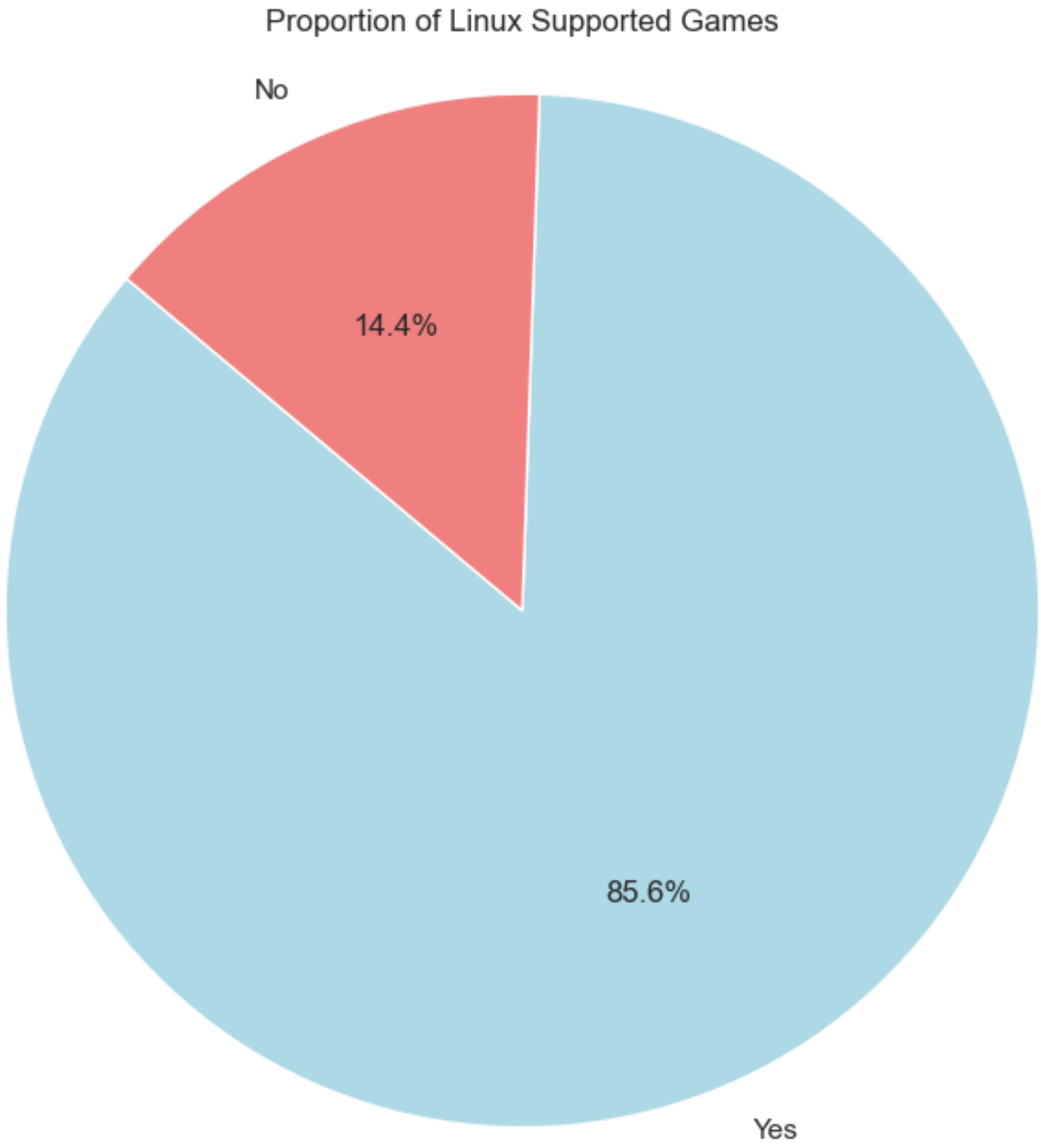
This indicates a large portion of the dataset consists of single-player indie games, often with casual or action elements. These common features means our recommendation models need to leverage the less common, more specific tags/genres/categories effectively to provide diverse and relevant suggestions.

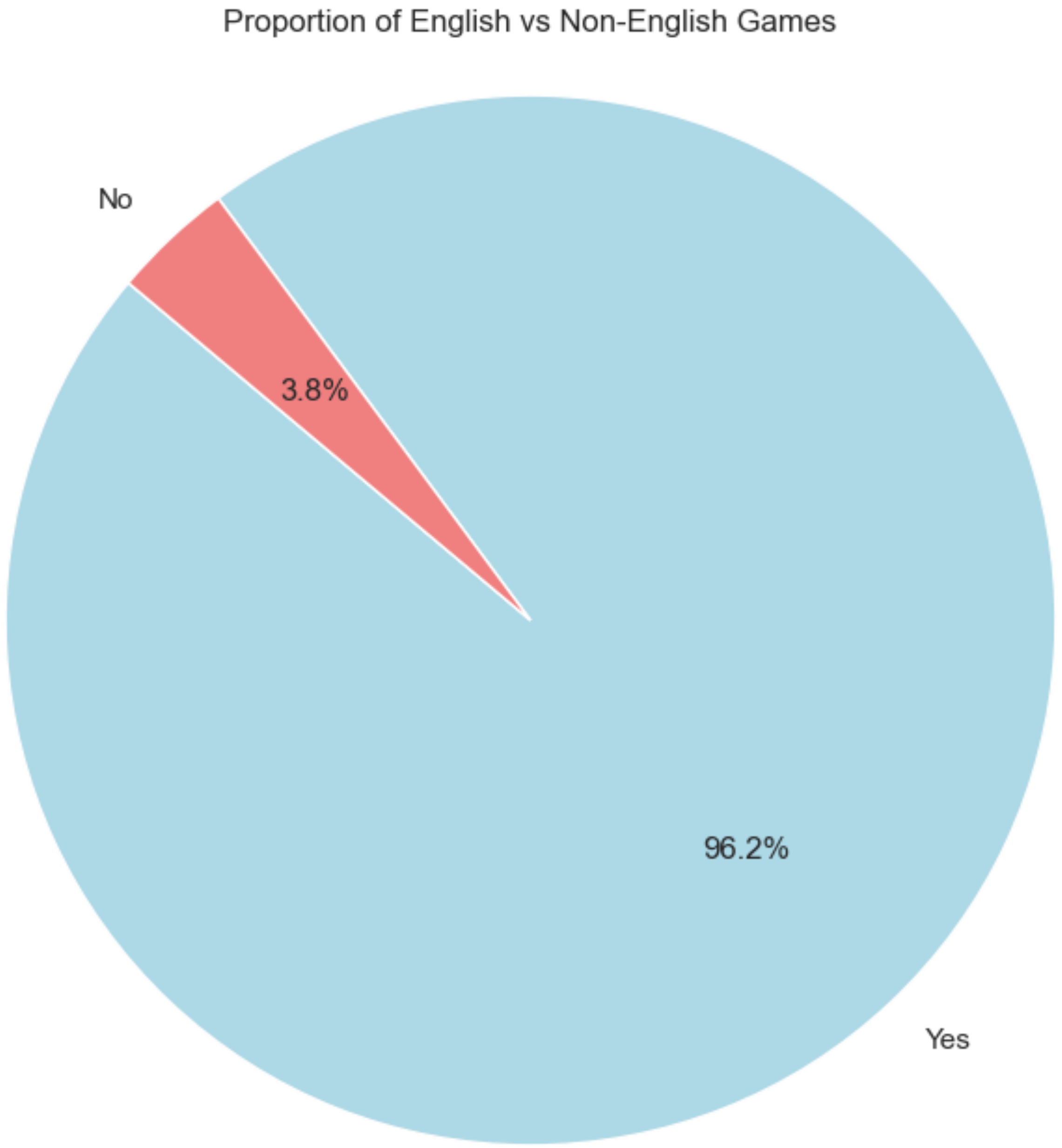
Next, we want to find out more about the distributions of game languages and platform supports in steam games

In [18]:
```python
# pie chart of indie vs non-indie games
cate = ['Windows', 'Mac', 'Linux']
for cat in cate:
    plt.figure(figsize=(8, 8))
    indie_count = df[cat].value_counts()
    indie_count.index = ['Yes', 'No']
    plt.pie(indie_count, labels=indie_count.index, autopct='%1.1f%%', startangle=140, colors=[
    plt.title(f'Proportion of {cat} Supported Games')
    plt.axis('equal')
    plt.show()

indie_count = df['Is English Supported'].value_counts()
indie_count.index = ['Yes', 'No']
plt.figure(figsize=(8, 8))
plt.pie(indie_count, labels=indie_count.index, autopct='%1.1f%%', startangle=140, colors=['lig
plt.title('Proportion of English vs Non-English Games')
plt.axis('equal')
plt.show()

print(df['Is English Supported'].value_counts(normalize=True))
print(df['Windows'].value_counts(normalize=True))
print(df['Mac'].value_counts(normalize=True))
print(df['Linux'].value_counts(normalize=True))
```


Proportion of Windows Supported Games
No
0.0%
100.0%
Yes


Proportion of Mac Supported Games
No
21.5%
78.5%
Yes


Proportion of Linux Supported Games
No
14.4%
85.6%
Yes



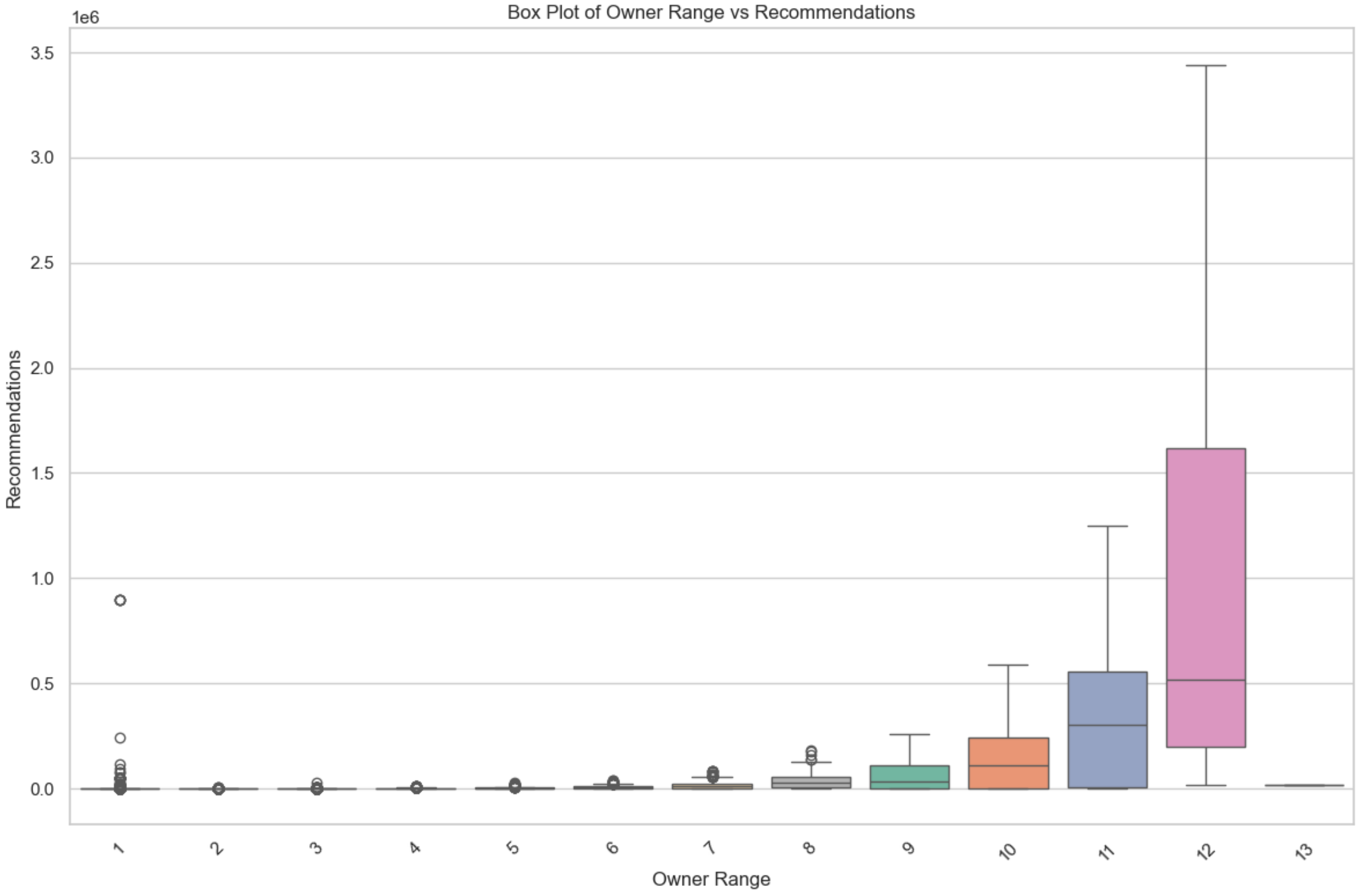
```
Is English Supported
1    0.96249
0    0.03751
Name: proportion, dtype: float64
Windows
True     0.999834
False    0.000166
Name: proportion, dtype: float64
Mac
False    0.784601
True     0.215399
Name: proportion, dtype: float64
Linux
False    0.855787
True     0.144213
Name: proportion, dtype: float64
```

As expected, a very high proportion of the game supports english. Majority of games are in english, in fact, 99.99% of games support windows. Since the proportion is so one-sided, the recommendation model should not put high focus on the language or the platform.

Lastly, we want to check if the number of estimated owners affect the recommendations, meaning more owners = more recommendations to see if owner range is useful in our model.

In [19]:
```
plt.figure(figsize=(12, 8))
```

```python
sns.boxplot(x='Owner range', y='Recommendations', data=df, palette='Set2', hue='Owner range')
plt.title('Box Plot of Owner Range vs Recommendations')
plt.xlabel('Owner Range')
plt.ylabel('Recommendations')
plt.xticks(rotation=45)
plt.legend([],[], frameon=False)
plt.tight_layout()
plt.show()

display(df.groupby('Owner range')['Recommendations'].describe())
```



| Owner range | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **1** | 47841.0 | 1.294546e+02 | 8.404817e+03 | 0.0 | 0.0 | 0.0 | 0.00 | 899838.0 |
| **2** | 7905.0 | 1.439867e+02 | 2.786061e+02 | 0.0 | 0.0 | 0.0 | 205.00 | 7558.0 |
| **3** | 3912.0 | 4.060542e+02 | 7.610075e+02 | 0.0 | 0.0 | 158.0 | 562.25 | 29376.0 |
| **4** | 2566.0 | 8.979961e+02 | 1.242282e+03 | 0.0 | 0.0 | 447.5 | 1284.25 | 10098.0 |
| **5** | 2129.0 | 2.245640e+03 | 2.918576e+03 | 0.0 | 0.0 | 1291.0 | 3240.00 | 27540.0 |
| **6** | 900.0 | 6.026110e+03 | 6.593285e+03 | 0.0 | 471.0 | 4333.5 | 9122.25 | 38674.0 |
| **7** | 516.0 | 1.480264e+04 | 1.582332e+04 | 0.0 | 1856.0 | 10119.5 | 22343.50 | 84030.0 |
| **8** | 326.0 | 3.657755e+04 | 3.580609e+04 | 0.0 | 6402.0 | 28127.0 | 55573.00 | 181183.0 |
| **9** | 90.0 | 6.597726e+04 | 7.461084e+04 | 0.0 | 1857.5 | 35424.0 | 112969.50 | 259417.0 |
| **10** | 37.0 | 1.602234e+05 | 1.748687e+05 | 0.0 | 2469.0 | 110892.0 | 243925.00 | 591701.0 |
| **11** | 21.0 | 3.384199e+05 | 3.667980e+05 | 1000.0 | 7699.0 | 305370.0 | 553709.00 | 1247051.0 |
| **12** | 5.0 | 1.157720e+06 | 1.419914e+06 | 14410.0 | 196597.0 | 519578.0 | 1616422.00 | 3441592.0 |
| **13** | 1.0 | 1.430000e+04 | NaN | 14300.0 | 14300.0 | 14300.0 | 14300.00 | 14300.0 |

Owner range represents the estimated number of owners for a game, the values are assigned as such

1->0-20000

2->20000-50000

3->50000-100000

4->100000-200000

5->200000-500000

6->500000-1000000

7->1000000-2000000

8->2000000-5000000

9->5000000-10000000

10->10000000-20000000

11->20000000-50000000

12->50000000-100000000

13->100000000-200000000

The graph is hard to visualize, so the summary statistics is provided to better show the values in the box plot, from the table, we can see there is good seperation between the owner ranges, where more people, on average, will recommend the game if more people play it. However, looking at the Inter-quartile range, there is significant overlap in some classes, making it unreliable for predicting the recommendations.

## Text Predictors

Finally, we will analysing description of the games, to see if there any patterns that could be useful for recommending games

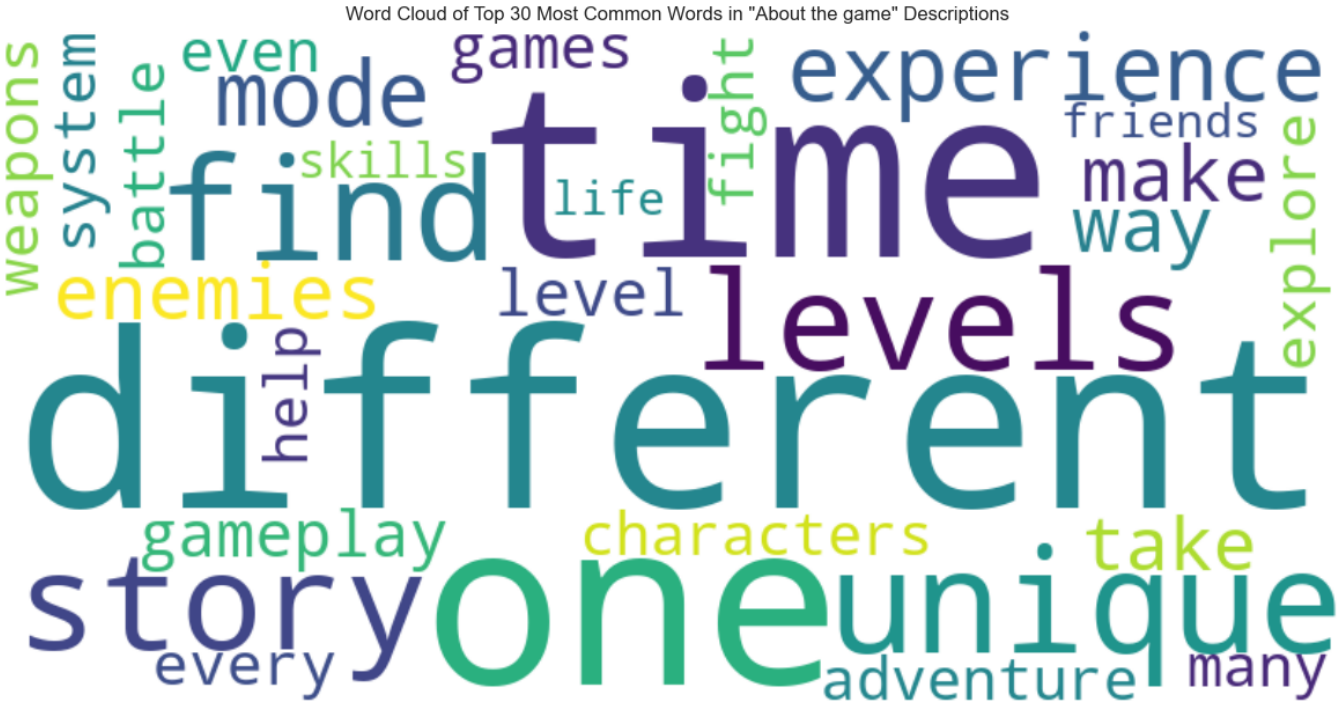
In [24]:

```
text_corpus = " ".join(str(review) for review in df['About the game'].fillna('')) # Concatenat

# Cleaning the text corpus
text_corpus = text_corpus.lower() # Convert to lowercase
text_corpus = re.sub(r'[^a-z0-9\s]', '', text_corpus) # Remove punctuation and special charact
text_corpus = re.sub(r'\s+', ' ', text_corpus).strip() # Remove extra spaces

custom_stopwords = set(STOPWORDS)
custom_stopwords.update(['game', 'player', 'players', 'feature', 'features', 'world', 'new', '
```

In [25]:

```
word_freq = pd.Series(text_corpus.split()).value_counts()
word_freq = word_freq[~word_freq.index.isin(custom_stopwords)]
print("\nTop 10 Most Common Words in 'About the game' Descriptions:")
display(word_freq.head(10))

word_freq = word_freq[:30]
plt.figure(figsize=(12, 8))
wordcloud = WordCloud(width=800, height=400, background_color='white', colormap='viridis', sto
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Top 30 Most Common Words in "About the game" Descriptions')
plt.tight_layout()
plt.show()
```

```
Top 10 Most Common Words in 'About the game' Descriptions:
different     33761
time          32743
one           30276
unique        27502
find          27060
story         25430
levels        25118
experience    23768
mode          23516
enemies       23195
Name: count, dtype: int64
```



The words highlight different aspects and genres of the games.

Mode could point to different ways the game is played

time could refer to game length, real time or turn based aspects or similar mechanics

Different, unique, new could just be marketing terms for game mechanics

find, make, take describes actions players can make in the game

one could be describing the number of players in the game

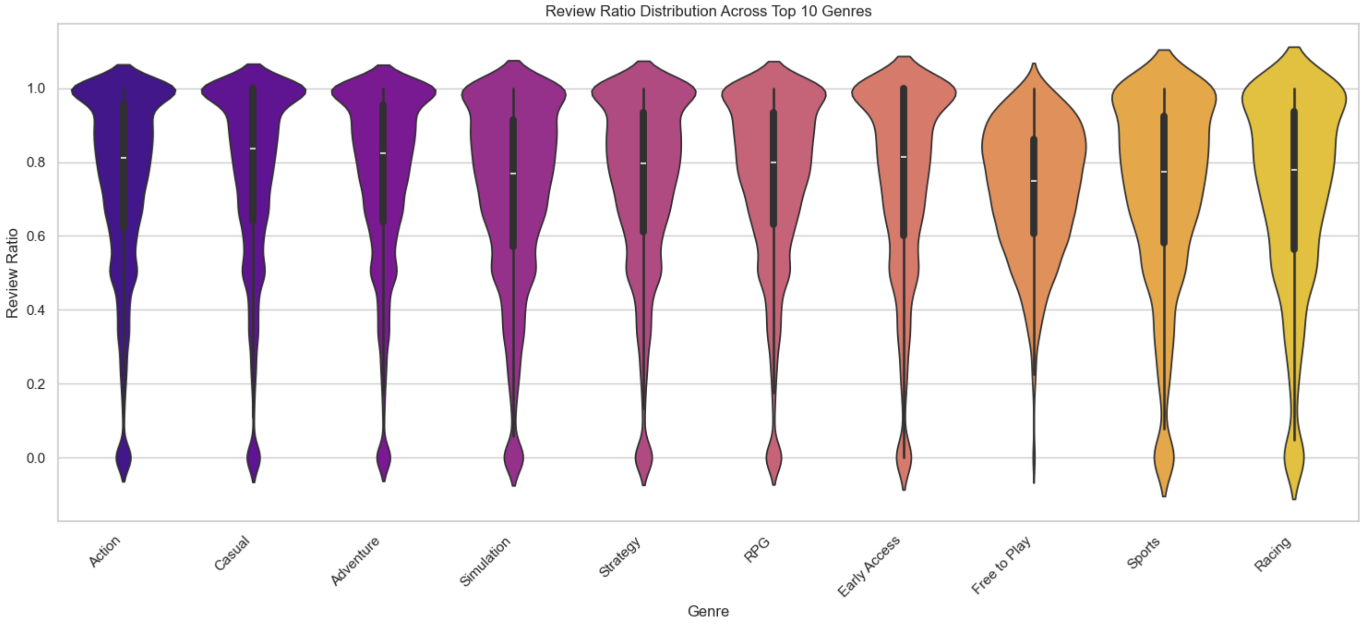
In general, the presence of these words indicates that the game is rich with terms describing the game mechanics, game structure, key features, selling points and player experience which is useful for our content-based models (Cosine Similarity and KNN), and that TF-IDF should be able to effectively capture the similarities between games and described features.

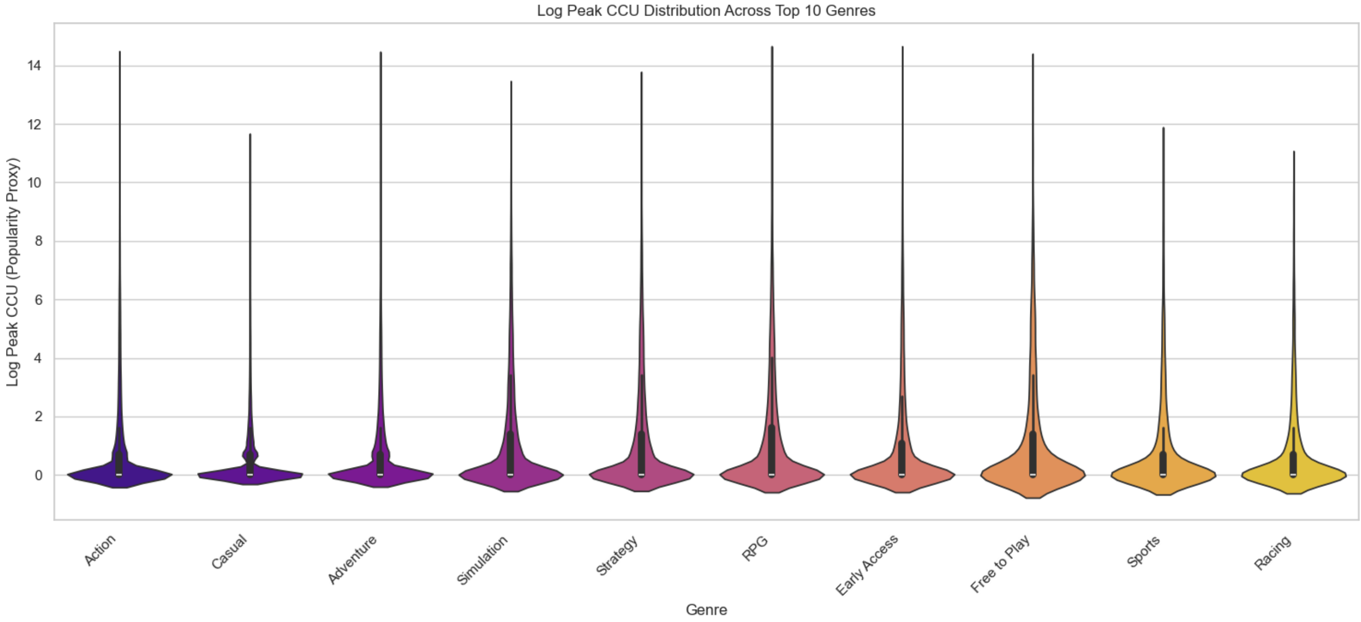
## Cross-Feature Analysis

Now, we would like to explore relationships genres and key metadata like Peak CCU to help us understand if certain types of games inherently attract more players.

```
In [22]: top_n_genres = genre_freq.head(10).index.tolist() # Get top 10 genres
         df_top_genres = df[df[top_n_genres].any(axis=1)].copy() # Filter rows with at least one of the
```

```
melted_df = pd.melt(df_top_genres,
                    id_vars=['Review Ratio', 'Log Peak CCU'], # Use the Log-transformed Peak C
                    value_vars=top_n_genres,
                    var_name='Genre',
                    value_name='Is_Genre')

melted_df = melted_df[melted_df['Is_Genre'] == 1]
melted_df['Genre'] = melted_df['Genre'].str.replace('Genres_', '')
# Plot Review Ratio vs Top Genres
plt.figure(figsize=(15, 7))
sns.violinplot(x='Genre', y='Review Ratio', data=melted_df, palette='plasma', hue='Genre', leg
plt.title('Review Ratio Distribution Across Top 10 Genres')
plt.xticks(rotation=45, ha='right')
plt.ylabel('Review Ratio')
plt.xlabel('Genre')
plt.tight_layout()
plt.show()

# Plot Log Peak CCU vs Top Genres
plt.figure(figsize=(15, 7))
sns.violinplot(x='Genre', y='Log Peak CCU', data=melted_df, palette='plasma', hue='Genre', leg
plt.title('Log Peak CCU Distribution Across Top 10 Genres')
plt.xticks(rotation=45, ha='right')
plt.ylabel('Log Peak CCU (Popularity Proxy)')
plt.xlabel('Genre')
plt.tight_layout()
plt.show()
```





Unfortuantely both of these plots are not very informative. The violin plots show that the distributions of Review Ratio and Log Peak CCU are quite similar across the top 10 genres. This suggests that there is no significant difference in player reception or popularity based on genre classification.

Overall, it suggest within this data, the genre of a game is not a strong predictor of it's general popularity (Peak CCU) or overall player reception (Review Ratio), and that content features like genre and tags and metadata features like popularity and rating are not strongly correlated.

It implies that content features (like genre/tags) and metadata features (like popularity/rating) might be more independent than initially assumed. They capture different aspects of a game, which makes a hybrid approach (based on content and based on metadata) potentially viable.

We save the new dataset with more features like has_x and log of skewed vairables to another dataset

```
In [23]: # Saving DF for ML to use
         df.to_csv("Dataset/games_eda.csv", index=False)
```

# Conclusion

Firstly, there is rich text contents in the dataset, the game descriptions ('About the game') are detailed and cover specific aspects about the game, making text analysing promising for finding similar content.

Secondly, metadata are very heavily right skewed, which would require log transformation for better analysis and analysis. Log Peak CCU showed strong correlation with Recomendations, indicating game popularity could be a relevant factor for recommending games. Other features also showed some correlations, but overall, it suggests that incorporating numerical metadata, particularly popularity and engagement metrics, could enhance recommendations.

Thirdly, there are a lot of common categories. The dataset is dominated by indie, singleplayer, action, adventure games and categories like steam achievements. They are able describe a lot of games, as a result, alone they're pretty weak in differentiating games. To provide better recommendations, most likely will need to leverage on a combination of these features with less frequent tags/genres/categories. The column is also one-hot encoded to better prepare for modeling.

Finally, the feature types are independent, the cross-feature analysis actually showed that the content data (genres, tags) are actually not very good predictors for metadata (popularity, prices), this implies that content features (like descriptions, tags, genres) and metadata capture different and complementary aspects of a game.

## Model Recommendation

Based on our EDA findings, we decided to implement two distinct approaches:

Model 1: TF-IDF + Cosine Similarity (Content-based)

This method directly leverages the rich textual information in 'About the game' descriptions. TF-IDF is suited for capturing the importance of descriptive term, and cosine similarity is effective for comparing the results in high-dimensions spare vectors (Many features and lots of missing data/0). Overall the model should be a good baseline model based on content similarity expressed through the description.

Model 2: K-Nearest Neighbors (KNN) on Combined Features (Hybrid) Since numerical (popularity, engagement), categorical (tags/genre) and text all showed to be potentially useful on their own but independent. This model attempts to make use of all of them. By combining TF-IDF vectors, scaled numerical features and one-hot encoded categorical features into one feature space, K nearest Neighbour can attempt to find games that are "close" across all of these dimensions. This approach tests if integrating diverse metadata leads to different and potentially more relevant recommendations.