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 [2]:
```
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 [3]:
```
df = pd.read_csv("Dataset/games_cleaned.csv")
```

In [4]:
```
print(f"DataFrame shape: {df.shape}")
print(f"DataFrame columns: {df.columns.tolist()}")
```

```
DataFrame shape: (66272, 550)
DataFrame columns: ['AppID', 'Name', 'Peak CCU', 'Required age', 'Price', 'DLC co
unt', 'About the game', 'Windows', 'Mac', 'Linux', 'Achievements', 'Recommendatio
ns', 'Average playtime two weeks', 'Median playtime forever', 'Developers', 'Publ
ishers', 'Game Age (Days)', 'is_indie', 'Owner range', 'Total Reviews', 'Review R
atio', 'Num Languages', 'Is English Supported', 'Tags_Singleplayer', 'Tags_Action
', 'Tags_Casual', 'Tags_Adventure', 'Tags_2D', 'Tags_Strategy', 'Tags_Simulation'
, 'Tags_RPG', 'Tags_Puzzle', 'Tags_Atmospheric', 'Tags_3D', 'Tags_Early Access',
'Tags_Pixel Graphics', 'Tags_Story Rich', 'Tags_Colorful', 'Tags_Exploration', 'T
ags_Cute', 'Tags_First-Person', 'Tags_Arcade', 'Tags_Multiplayer', 'Tags_Fantasy'
, 'Tags_Funny', 'Tags_Shooter', 'Tags_Horror', 'Tags_Retro', 'Tags_Platformer', '
Tags_Anime', 'Tags_Family Friendly', 'Tags_Sci-fi', 'Tags_Action-Adventure', 'Tag
s_Relaxing', 'Tags_Female Protagonist', 'Tags_Difficult', 'Tags_Third Person', 'T
ags_VR', 'Tags_Survival', 'Tags_Top-Down', 'Tags_Open World', 'Tags_Stylized', 'T
ags_Controller', 'Tags_Combat', 'Tags_Great Soundtrack', 'Tags_Comedy', 'Tags_2D
Platformer', 'Tags_Visual Novel', 'Tags_FPS', 'Tags_Violent', 'Tags_Mystery', 'Ta
gs_Co-op', 'Tags_Dark', 'Tags_Physics', 'Tags_Realistic', 'Tags_Minimalist', 'Tag
s_Free to Play', 'Tags_Cartoony', 'Tags_Psychological Horror', 'Tags_Choices Matt
er', 'Tags_Point & Click', 'Tags_Linear', 'Tags_Gore', 'Tags_PvP', 'Tags_Sandbox'
, 'Tags_Multiple Endings', 'Tags_Sports', 'Tags_Side Scroller', 'Tags_Space', 'Ta
gs_Rogue-like', 'Tags_Old School', 'Tags_PvE', 'Tags_Tactical', 'Tags_Rogue-lit
e', 'Tags_Building', 'Tags_Puzzle-Platformer', 'Tags_Management', 'Tags_Character
Customization', 'Tags_Hidden Object', 'Tags_Hand-drawn', 'Tags_Action RPG', 'Tags
_Sexual Content', 'Tags_Magic', 'Tags_Logic', 'Tags_Racing', 'Tags_Local Multipla
yer', 'Tags_Nudity', "Tags_Shoot 'Em Up", 'Tags_3D Platformer', 'Tags_Procedural
Generation', 'Tags_Turn-Based Strategy', '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 Tactics', 'Tags_Walking Si
mulator', 'Tags_Drama', 'Tags_Interactive Fiction', 'Tags_Choose Your Own Adventu
re', 'Tags_Hack and Slash', 'Tags_Action Roguelike', "Tags_1990's", 'Tags_Resourc
e Management', 'Tags_Zombies', 'Tags_Local Co-Op', 'Tags_Score Attack', 'Tags_JRP
G', 'Tags_Dungeon Crawler', 'Tags_Replay Value', 'Tags_Dark Fantasy', 'Tags_Surre
al', 'Tags_War', 'Tags_Historical', 'Tags_Immersive Sim', 'Tags_Turn-Based', 'Tag
s_Post-apocalyptic', 'Tags_Top-Down Shooter', 'Tags_Nature', 'Tags_Emotional', 'T
ags_Base-Building', 'Tags_Stealth', 'Tags_Isometric', 'Tags_Education', 'Tags_Rom
ance', 'Tags_Text-Based', 'Tags_Fast-Paced', 'Tags_Massively Multiplayer', 'Tags_
Short', 'Tags_Card Game', 'Tags_Abstract', 'Tags_Classic', 'Tags_Clicker', 'Tags_
RTS', 'Tags_Precision Platformer', 'Tags_Military', 'Tags_1980s', 'Tags_2.5D', 'T
ags_Third-Person Shooter', 'Tags_Tower Defense', 'Tags_Investigation', 'Tags_Boar
d Game', 'Tags_Detective', 'Tags_Memes', 'Tags_RPGMaker', 'Tags_Narration', 'Tags
_Dating Sim', 'Tags_Robots', 'Tags_Aliens', 'Tags_Cyberpunk', 'Tags_Perma Death',
'Tags_Tabletop', 'Tags_Driving', 'Tags_Cinematic', 'Tags_Life Sim', 'Tags_Dark Hu
mor', 'Tags_Economy', 'Tags_Arena Shooter', 'Tags_Time Management', 'Tags_VR Only
', 'Tags_Mature', 'Tags_4 Player Local', 'Tags_Flight', 'Tags_Psychological', 'Ta
gs_Thriller', 'Tags_Strategy RPG', 'Tags_Real Time Tactics', 'Tags_Demons', 'Tags
_City Builder', 'Tags_Experimental', "Tags_Beat 'em up", 'Tags_Conversation', 'Ta
gs_Psychedelic', 'Tags_Runner', 'Tags_Fighting', 'Tags_Metroidvania', 'Tags_LGBTQ
+', 'Tags_Tactical RPG', 'Tags_Wargame', 'Tags_Nonlinear', 'Tags_Supernatural', '
Tags_Music', 'Tags_Team-Based', 'Tags_Collectathon', 'Tags_Tutorial', 'Tags_Compe
titive', 'Tags_Lore-Rich', 'Tags_Level Editor', 'Tags_Artificial Intelligence', '
Tags_Idler', 'Tags_Comic Book', 'Tags_Loot', 'Tags_Automobile Sim', 'Tags_Twin St
ick Shooter', 'Tags_Party-Based RPG', 'Tags_Dystopian', 'Tags_Parkour', 'Tags_Mod
ern', 'Tags_Destruction', 'Tags_Grid-Based Movement', 'Tags_2D Fighter', 'Tags_So
uls-like', 'Tags_Hentai', 'Tags_Match 3', 'Tags_Cats', 'Tags_Alternate History',
'Tags_Rhythm', 'Tags_Deckbuilding', 'Tags_CRPG', 'Tags_Inventory Management', 'Ta
gs_Moddable', 'Tags_Crime', 'Tags_Mythology', 'Tags_Space Sim', 'Tags_Card Battle
r', 'Tags_Beautiful', 'Tags_Soundtrack', 'Tags_Grand Strategy', 'Tags_World War I
I', 'Tags_Philosophical', 'Tags_Dark Comedy', 'Tags_Science', 'Tags_Noir', 'Tags_
Mystery Dungeon', 'Tags_3D Fighter', 'Tags_Character Action Game', 'Tags_Lovecraf
tian', 'Tags_NSFW', 'Tags_Word Game', 'Tags_Split Screen', 'Tags_Swordplay', 'Tag
```

s_Colony Sim', 'Tags_Farming Sim', 'Tags_Automation', 'Tags_Mouse only', 'Tags_Cr
eature Collector', 'Tags_e-sports', 'Tags_6DOF', 'Tags_Dragons', 'Tags_Voxel', 'T
ags_3D Vision', 'Tags_Vehicular Combat', 'Tags_Solitaire', 'Tags_Bullet Time', 'T
ags_Hero Shooter', 'Tags_Mechs', 'Tags_Battle Royale', 'Tags_Agriculture', 'Tags_
Capitalism', 'Tags_Parody', 'Tags_Combat Racing', 'Tags_Open World Survival Craft
', 'Tags_Spectacle fighter', 'Tags_America', 'Tags_Design & Illustration', 'Tags_
Blood', 'Tags_Time Manipulation', 'Tags_Gun Customization', 'Tags_Class-Based', '
Tags_Sokoban', 'Tags_MMORPG', 'Tags_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_Underground', 'T
ags_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_Politics', 'Tags_Escape Room', 'Tags_Real-Time with Pause',
'Tags_Faith', 'Tags_Naval', 'Tags_Vampire', 'Tags_Minigames', 'Tags_Software', 'T
ags_Political Sim', 'Tags_Western', 'Tags_Superhero', 'Tags_Narrative', 'Tags_Par
ty Game', 'Tags_Transportation', 'Tags_Trading Card Game', 'Tags_Assassin', 'Tags
_Immersive', 'Tags_Action RTS', 'Tags_Illuminati', 'Tags_Cozy', 'Tags_Touch-Frien
dly', 'Tags_MOBA', 'Tags_Time Attack', 'Tags_Typing', 'Tags_Game Development', 'T
ags_Asynchronous Multiplayer', 'Tags_Trivia', 'Tags_Trains', 'Tags_On-Rails Shoot
er', 'Tags_Roguelike Deckbuilder', 'Tags_Cold War', 'Tags_Party', 'Tags_Snow', 'T
ags_Animation & Modeling', 'Tags_Traditional Roguelike', 'Tags_Archery', 'Tags_Of
froad', 'Tags_Heist', 'Tags_Naval Combat', 'Tags_Diplomacy', 'Tags_Football', 'Ta
gs_Music-Based Procedural Generation', 'Tags_Soccer', 'Tags_Villain Protagonist',
'Tags_Kickstarter', 'Tags_Sailing', 'Tags_Mars', 'Tags_Wholesome', 'Tags_Chess',
'Tags_Foreign', 'Tags_Experience', 'Tags_Nostalgia', 'Tags_Gambling', 'Tags_Seque
l', 'Tags_Sniper', 'Tags_GameMaker', 'Tags_Boxing', 'Tags_Horses', 'Tags_Episodic
', 'Tags_World War I', 'Tags_Golf', 'Tags_Spelling', 'Tags_Jet', 'Tags_Unforgivin
g', 'Tags_Motorbike', 'Tags_Outbreak Sim', 'Tags_Transhumanism', 'Tags_Werewolves
', 'Tags_Rome', 'Tags_Pinball', 'Tags_Bikes', 'Tags_Farming', 'Tags_Silent Protag
onist', 'Tags_Epic', 'Tags_Roguevania', 'Tags_Spaceships', 'Tags_Basketball', 'Ta
gs_Crowdfunded', 'Tags_Asymmetric VR', 'Tags_Submarine', 'Tags_Medical Sim', 'Tag
s_Audio Production', 'Tags_LEGO', 'Tags_Video Production', 'Tags_Social Deduction
', 'Tags_Games Workshop', 'Tags_Mini Golf', 'Tags_Vikings', 'Tags_Based On A Nove
l', '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_Gaming', 'Tags_Mod', '
Tags_Motocross', 'Tags_Lemmings', 'Tags_Intentionally Awkward Controls', 'Tags_Ph
oto Editing', 'Tags_Skateboarding', 'Tags_Hockey', 'Tags_Cycling', 'Tags_Boomer S
hooter', 'Tags_Web Publishing', 'Tags_Instrumental Music', 'Tags_Jump Scare', 'Ta
gs_Skating', 'Tags_Bowling', 'Tags_Skiing', 'Tags_Football (Soccer)', 'Tags_Track
IR', 'Tags_Rock Music', 'Tags_Boss Rush', 'Tags_Snowboarding', 'Tags_Musou', 'Tag
s_8-bit Music', 'Tags_BMX', 'Tags_Documentary', 'Tags_Job Simulator', 'Tags_Maste
rpiece', 'Tags_Voice Control', 'Tags_Electronic', 'Tags_ATV', 'Tags_Well-Writte
n', 'Tags_Cricket', 'Tags_Football (American)', 'Tags_Lara Croft', 'Tags_Hardware
', 'Tags_Reboot', 'Tags_Benchmark', 'Tags_Volleyball', 'Tags_Feature Film', 'Tags
_Steam Machine', 'Tags_Shop Keeper', 'Tags_Mahjong', 'Tags_Birds', 'Tags_Extracti
on Shooter', 'Tags_Rugby', 'Tags_Coding', 'Tags_Fox', 'Tags_Dwarf', 'Tags_Hobby S
im', 'Tags_Tile-Matching', 'Tags_Batman', 'Tags_Elf', 'Tags_Snooker', 'Genres_Act
ion', 'Genres_Casual', 'Genres_Adventure', 'Genres_Simulation', 'Genres_Strategy'
, 'Genres_RPG', 'Genres_Early Access', 'Genres_Free to Play', 'Genres_Sports', 'G
enres_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 Tr
aining', 'Genres_Web Publishing', 'Genres_Photo Editing', 'Genres_Accounting', 'G
enres_Movie', 'Genres_Documentary', 'Genres_Episodic', 'Genres_Short', 'Genres_Tu
torial', 'Genres_360 Video', 'Genres_Free To Play', 'Categories_Single-player', '

```
Categories_Steam Achievements', 'Categories_Steam Cloud', 'Categories_Full contro
ller support', 'Categories_Multi-player', 'Categories_Steam Trading Cards', 'Cate
gories_Partial Controller Support', 'Categories_PvP', 'Categories_Co-op', 'Catego
ries_Steam Leaderboards', 'Categories_Online PvP', 'Categories_Remote Play Togeth
er', 'Categories_Shared/Split Screen', 'Categories_Online Co-op', 'Categories_Sha
red/Split Screen PvP', 'Categories_Stats', 'Categories_Family Sharing', 'Categori
es_Shared/Split Screen Co-op', 'Categories_Remote Play on TV', 'Categories_Cross-
Platform Multiplayer', 'Categories_Includes level editor', 'Categories_Steam Work
shop', 'Categories_In-App Purchases', 'Categories_Captions available', 'Categorie
s_Remote Play on Tablet', 'Categories_MMO', 'Categories_Remote Play on Phone', 'C
ategories_LAN PvP', 'Categories_LAN Co-op', 'Categories_VR Only', 'Categories_VR
Support', 'Categories_Commentary available', 'Categories_Tracked Controller Suppo
rt', 'Categories_Valve Anti-Cheat enabled', 'Categories_Steam Turn Notifications'
, 'Categories_VR Supported', 'Categories_Includes Source SDK', 'Categories_SteamV
R Collectibles', 'Categories_HDR available', 'Categories_Mods', 'Categories_Mods
(require HL2)', 'Developers freq', 'Publishers freq']
```

In [5]:
```
columns = df.columns.tolist()
print(columns)
```

['AppID', 'Name', 'Peak CCU', 'Required age', 'Price', 'DLC count', 'About the ga
me', 'Windows', 'Mac', 'Linux', 'Achievements', 'Recommendations', 'Average playt
ime two weeks', 'Median playtime forever', 'Developers', 'Publishers', 'Game Age
(Days)', 'is_indie', 'Owner range', 'Total Reviews', 'Review Ratio', 'Num Languag
es', 'Is English Supported', 'Tags_Singleplayer', 'Tags_Action', 'Tags_Casual', '
Tags_Adventure', 'Tags_2D', 'Tags_Strategy', 'Tags_Simulation', 'Tags_RPG', 'Tags
_Puzzle', 'Tags_Atmospheric', 'Tags_3D', 'Tags_Early Access', 'Tags_Pixel Graphic
s', 'Tags_Story Rich', 'Tags_Colorful', 'Tags_Exploration', 'Tags_Cute', 'Tags_Fi
rst-Person', 'Tags_Arcade', 'Tags_Multiplayer', 'Tags_Fantasy', 'Tags_Funny', 'Ta
gs_Shooter', 'Tags_Horror', 'Tags_Retro', 'Tags_Platformer', 'Tags_Anime', 'Tags_
Family Friendly', 'Tags_Sci-fi', 'Tags_Action-Adventure', 'Tags_Relaxing', 'Tags_
Female Protagonist', 'Tags_Difficult', 'Tags_Third Person', 'Tags_VR', 'Tags_Surv
ival', 'Tags_Top-Down', 'Tags_Open World', 'Tags_Stylized', 'Tags_Controller', 'T
ags_Combat', 'Tags_Great Soundtrack', 'Tags_Comedy', 'Tags_2D Platformer', 'Tags_
Visual Novel', 'Tags_FPS', 'Tags_Violent', 'Tags_Mystery', 'Tags_Co-op', 'Tags_Da
rk', 'Tags_Physics', 'Tags_Realistic', 'Tags_Minimalist', 'Tags_Free to Play', 'T
ags_Cartoony', 'Tags_Psychological Horror', 'Tags_Choices Matter', 'Tags_Point &
Click', 'Tags_Linear', 'Tags_Gore', 'Tags_PvP', 'Tags_Sandbox', 'Tags_Multiple En
dings', 'Tags_Sports', 'Tags_Side Scroller', 'Tags_Space', 'Tags_Rogue-like', 'Ta
gs_Old School', 'Tags_PvE', 'Tags_Tactical', 'Tags_Rogue-lite', 'Tags_Building',
'Tags_Puzzle-Platformer', 'Tags_Management', 'Tags_Character Customization', 'Tag
s_Hidden Object', 'Tags_Hand-drawn', 'Tags_Action RPG', 'Tags_Sexual Content', 'T
ags_Magic', 'Tags_Logic', 'Tags_Racing', 'Tags_Local Multiplayer', 'Tags_Nudity',
"Tags_Shoot 'Em Up", 'Tags_3D Platformer', 'Tags_Procedural Generation', 'Tags_Tu
rn-Based Strategy', 'Tags_Survival Horror', 'Tags_Cartoon', 'Tags_Futuristic', 'T
ags_Bullet Hell', 'Tags_Medieval', 'Tags_Turn-Based Combat', 'Tags_Online Co-Op',
'Tags_Crafting', 'Tags_Turn-Based Tactics', 'Tags_Walking Simulator', 'Tags_Drama
', 'Tags_Interactive Fiction', 'Tags_Choose Your Own Adventure', 'Tags_Hack and S
lash', 'Tags_Action Roguelike', "Tags_1990's", 'Tags_Resource Management', 'Tags_
Zombies', 'Tags_Local Co-Op', 'Tags_Score Attack', 'Tags_JRPG', 'Tags_Dungeon Cra
wler', 'Tags_Replay Value', 'Tags_Dark Fantasy', 'Tags_Surreal', 'Tags_War', 'Tag
s_Historical', 'Tags_Immersive 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-Bas
ed', 'Tags_Fast-Paced', 'Tags_Massively Multiplayer', 'Tags_Short', 'Tags_Card Ga
me', 'Tags_Abstract', 'Tags_Classic', 'Tags_Clicker', 'Tags_RTS', 'Tags_Precision
Platformer', 'Tags_Military', 'Tags_1980s', 'Tags_2.5D', 'Tags_Third-Person Shoot
er', 'Tags_Tower Defense', 'Tags_Investigation', 'Tags_Board Game', 'Tags_Detecti
ve', 'Tags_Memes', 'Tags_RPGMaker', 'Tags_Narration', 'Tags_Dating Sim', 'Tags_Ro
bots', 'Tags_Aliens', 'Tags_Cyberpunk', 'Tags_Perma Death', 'Tags_Tabletop', 'Tag
s_Driving', 'Tags_Cinematic', 'Tags_Life Sim', 'Tags_Dark Humor', 'Tags_Economy',
'Tags_Arena Shooter', 'Tags_Time Management', 'Tags_VR Only', 'Tags_Mature', 'Tag
s_4 Player Local', 'Tags_Flight', 'Tags_Psychological', 'Tags_Thriller', 'Tags_St
rategy RPG', 'Tags_Real Time Tactics', 'Tags_Demons', 'Tags_City Builder', 'Tags_
Experimental', "Tags_Beat 'em up", 'Tags_Conversation', 'Tags_Psychedelic', 'Tags
_Runner', 'Tags_Fighting', 'Tags_Metroidvania', 'Tags_LGBTQ+', 'Tags_Tactical RPG
', 'Tags_Wargame', 'Tags_Nonlinear', 'Tags_Supernatural', 'Tags_Music', 'Tags_Tea
m-Based', 'Tags_Collectathon', 'Tags_Tutorial', 'Tags_Competitive', 'Tags_Lore-Ri
ch', 'Tags_Level Editor', 'Tags_Artificial Intelligence', 'Tags_Idler', 'Tags_Com
ic Book', 'Tags_Loot', 'Tags_Automobile Sim', 'Tags_Twin Stick Shooter', 'Tags_Pa
rty-Based RPG', 'Tags_Dystopian', 'Tags_Parkour', 'Tags_Modern', 'Tags_Destructio
n', 'Tags_Grid-Based Movement', 'Tags_2D Fighter', 'Tags_Souls-like', 'Tags_Henta
i', 'Tags_Match 3', 'Tags_Cats', 'Tags_Alternate History', 'Tags_Rhythm', 'Tags_D
eckbuilding', 'Tags_CRPG', 'Tags_Inventory Management', 'Tags_Moddable', 'Tags_Cr
ime', 'Tags_Mythology', 'Tags_Space Sim', 'Tags_Card Battler', 'Tags_Beautiful',
'Tags_Soundtrack', 'Tags_Grand Strategy', 'Tags_World War II', 'Tags_Philosophica
l', 'Tags_Dark Comedy', 'Tags_Science', 'Tags_Noir', 'Tags_Mystery Dungeon', 'Tag
s_3D Fighter', 'Tags_Character Action Game', 'Tags_Lovecraftian', 'Tags_NSFW', 'T
ags_Word Game', 'Tags_Split Screen', 'Tags_Swordplay', 'Tags_Colony Sim', 'Tags_F
arming Sim', 'Tags_Automation', 'Tags_Mouse only', 'Tags_Creature Collector', 'Ta

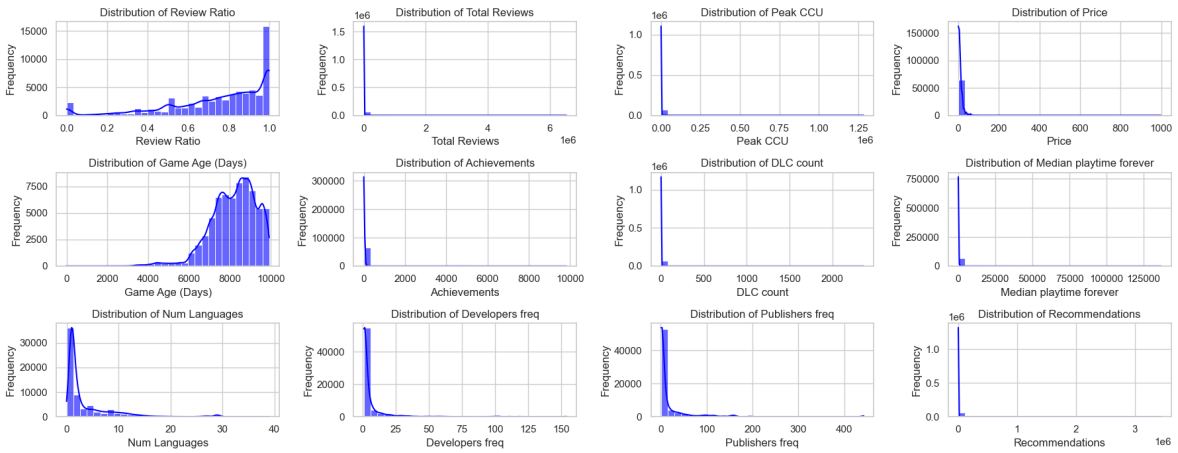
gs_e-sports', 'Tags_6DOF', 'Tags_Dragons', 'Tags_Voxel', 'Tags_3D Vision', 'Tags_
Vehicular Combat', 'Tags_Solitaire', 'Tags_Bullet Time', 'Tags_Hero Shooter', 'Ta
gs_Mechs', 'Tags_Battle Royale', 'Tags_Agriculture', 'Tags_Capitalism', 'Tags_Par
ody', 'Tags_Combat Racing', 'Tags_Open World Survival Craft', 'Tags_Spectacle fig
hter', 'Tags_America', 'Tags_Design & Illustration', 'Tags_Blood', 'Tags_Time Man
ipulation', 'Tags_Gun Customization', 'Tags_Class-Based', 'Tags_Sokoban', 'Tags_M
MORPG', 'Tags_Steampunk', 'Tags_Hex Grid', 'Tags_Addictive', 'Tags_God Game', 'Ta
gs_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_Underground', 'Tags_Satire', 'Tags_Rea
l-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_D
ynamic Narration', 'Tags_Cult Classic', 'Tags_Dinosaurs', 'Tags_Fishing', 'Tags_P
olitics', 'Tags_Escape Room', 'Tags_Real-Time with Pause', 'Tags_Faith', 'Tags_Na
val', 'Tags_Vampire', 'Tags_Minigames', 'Tags_Software', 'Tags_Political Sim', 'T
ags_Western', 'Tags_Superhero', 'Tags_Narrative', 'Tags_Party Game', 'Tags_Transp
ortation', 'Tags_Trading Card Game', 'Tags_Assassin', 'Tags_Immersive', 'Tags_Act
ion RTS', 'Tags_Illuminati', 'Tags_Cozy', 'Tags_Touch-Friendly', 'Tags_MOBA', 'Ta
gs_Time Attack', 'Tags_Typing', 'Tags_Game Development', 'Tags_Asynchronous Multi
player', 'Tags_Trivia', 'Tags_Trains', 'Tags_On-Rails Shooter', 'Tags_Roguelike D
eckbuilder', 'Tags_Cold War', 'Tags_Party', 'Tags_Snow', 'Tags_Animation & Modeli
ng', 'Tags_Traditional Roguelike', 'Tags_Archery', 'Tags_Offroad', 'Tags_Heist',
'Tags_Naval Combat', 'Tags_Diplomacy', 'Tags_Football', 'Tags_Music-Based Procedu
ral Generation', 'Tags_Soccer', 'Tags_Villain Protagonist', 'Tags_Kickstarter', '
Tags_Sailing', 'Tags_Mars', 'Tags_Wholesome', 'Tags_Chess', 'Tags_Foreign', 'Tags
_Experience', 'Tags_Nostalgia', 'Tags_Gambling', 'Tags_Sequel', 'Tags_Sniper', 'T
ags_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_Crowdfunded', 'Tags
_Asymmetric VR', 'Tags_Submarine', 'Tags_Medical Sim', 'Tags_Audio Production', '
Tags_LEGO', 'Tags_Video Production', 'Tags_Social Deduction', 'Tags_Games Worksho
p', 'Tags_Mini Golf', 'Tags_Vikings', 'Tags_Based On A Novel', 'Tags_Movie', 'Tag
s_Electronic Music', 'Tags_Ambient', 'Tags_Baseball', 'Tags_Software Training', '
Tags_Wrestling', 'Tags_360 Video', 'Tags_Dungeons & Dragons', 'Tags_Tennis', 'Tag
s_Warhammer 40K', 'Tags_Pool', 'Tags_Gaming', 'Tags_Mod', 'Tags_Motocross', 'Tags
_Lemmings', 'Tags_Intentionally Awkward Controls', 'Tags_Photo Editing', 'Tags_Sk
ateboarding', 'Tags_Hockey', 'Tags_Cycling', 'Tags_Boomer Shooter', 'Tags_Web Pub
lishing', 'Tags_Instrumental Music', 'Tags_Jump Scare', 'Tags_Skating', 'Tags_Bow
ling', '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 C
ontrol', 'Tags_Electronic', 'Tags_ATV', 'Tags_Well-Written', 'Tags_Cricket', 'Tag
s_Football (American)', 'Tags_Lara Croft', 'Tags_Hardware', 'Tags_Reboot', 'Tags_
Benchmark', 'Tags_Volleyball', 'Tags_Feature Film', 'Tags_Steam Machine', 'Tags_S
hop Keeper', 'Tags_Mahjong', 'Tags_Birds', 'Tags_Extraction Shooter', 'Tags_Rugby
', 'Tags_Coding', 'Tags_Fox', 'Tags_Dwarf', 'Tags_Hobby Sim', 'Tags_Tile-Matchin
g', 'Tags_Batman', 'Tags_Elf', 'Tags_Snooker', 'Genres_Action', 'Genres_Casual',
'Genres_Adventure', 'Genres_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', 'Gen
res_Design & Illustration', 'Genres_Animation & Modeling', 'Genres_Nudity', 'Genr
es_Sexual Content', 'Genres_Education', 'Genres_Video Production', 'Genres_Game D
evelopment', 'Genres_Audio Production', 'Genres_Software Training', 'Genres_Web P
ublishing', 'Genres_Photo Editing', 'Genres_Accounting', 'Genres_Movie', 'Genres_
Documentary', 'Genres_Episodic', 'Genres_Short', 'Genres_Tutorial', 'Genres_360 V
ideo', 'Genres_Free To Play', 'Categories_Single-player', 'Categories_Steam Achie
vements', 'Categories_Steam Cloud', 'Categories_Full controller support', 'Catego

```
ries_Multi-player', 'Categories_Steam Trading Cards', 'Categories_Partial Control
ler Support', 'Categories_PvP', 'Categories_Co-op', 'Categories_Steam Leaderboard
s', 'Categories_Online PvP', 'Categories_Remote Play Together', 'Categories_Share
d/Split Screen', 'Categories_Online Co-op', 'Categories_Shared/Split Screen PvP',
'Categories_Stats', 'Categories_Family Sharing', 'Categories_Shared/Split Screen
Co-op', 'Categories_Remote Play on TV', 'Categories_Cross-Platform Multiplayer',
'Categories_Includes level editor', 'Categories_Steam Workshop', 'Categories_In-A
pp Purchases', 'Categories_Captions available', 'Categories_Remote Play on Tablet
', 'Categories_MMO', 'Categories_Remote Play on Phone', 'Categories_LAN PvP', 'Ca
tegories_LAN Co-op', 'Categories_VR Only', 'Categories_VR Support', 'Categories_C
ommentary available', 'Categories_Tracked Controller Support', 'Categories_Valve
Anti-Cheat enabled', 'Categories_Steam Turn Notifications', 'Categories_VR Suppor
ted', 'Categories_Includes Source SDK', 'Categories_SteamVR Collectibles', 'Categ
ories_HDR available', 'Categories_Mods', 'Categories_Mods (require HL2)', 'Develo
pers freq', 'Publishers freq']
```

Numerical Predictors

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

In [6]:
```
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
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())
```



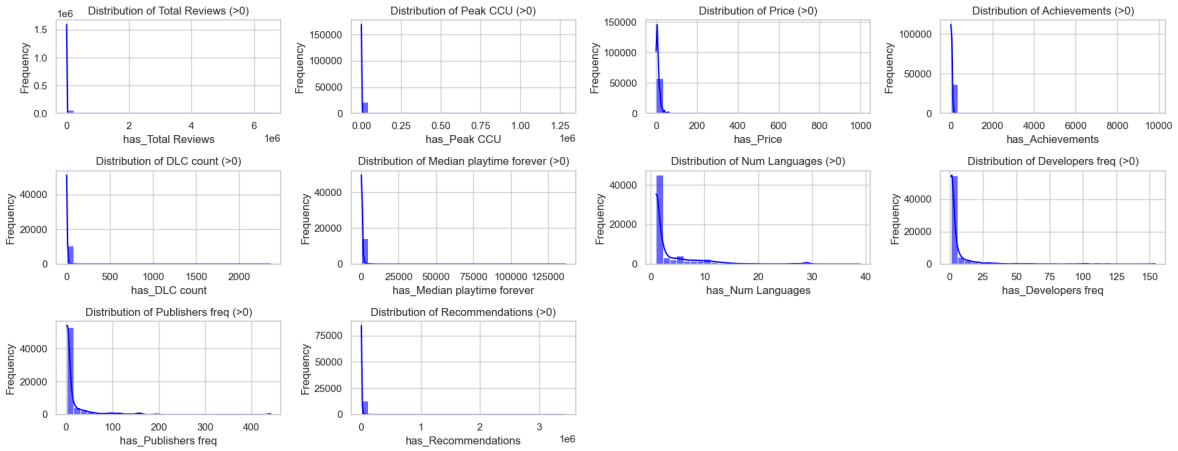
Summary Statistics for Numerical Predictors:

	Review Ratio	Total Reviews	Peak CCU	Price	Game Age (Days)	Achieveme
count	66272.000000	6.627200e+04	6.627200e+04	66272.000000	66272.000000	66272.000(
mean	0.756011	1.437623e+03	2.353899e+02	7.963390	8177.432068	24.052(
std	0.248087	3.188244e+04	9.314509e+03	10.622066	1102.109736	193.902(
min	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000(
25%	0.634389	5.000000e+00	0.000000e+00	1.990000	7495.000000	0.000(
50%	0.821744	2.100000e+01	0.000000e+00	4.990000	8324.000000	6.000(
75%	0.959184	1.130000e+02	1.000000e+00	9.990000	8992.000000	21.000(
max	1.000000	6.531097e+06	1.284268e+06	999.000000	9924.000000	9821.000(

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 [7]:
```
skewed_col = ['Total Reviews', 'Peak CCU', 'Price', 'Achievements', 'DLC count',

# 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)

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'
    plt.title(f'Distribution of {col} (>0)')
    plt.xlabel(f'has_{col}')
    plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```

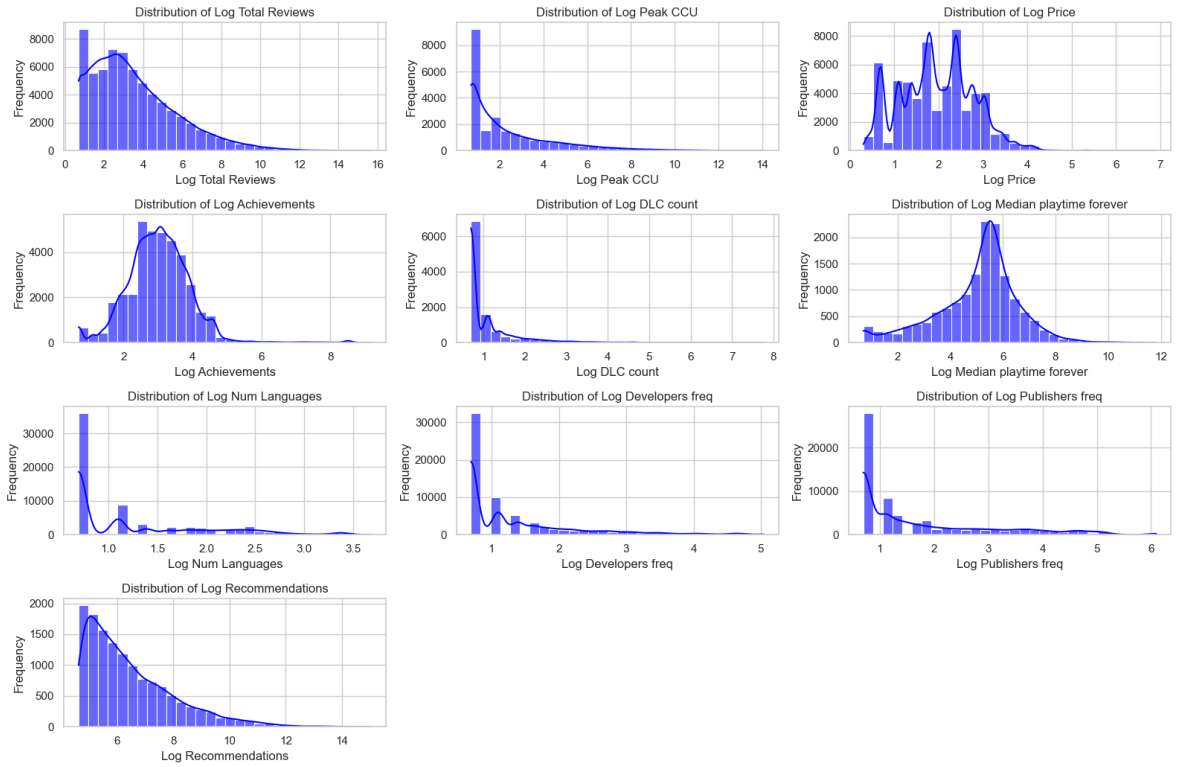


As seen here, it still quite heavily right skewed.

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 [8]:
```
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, col
    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())
```

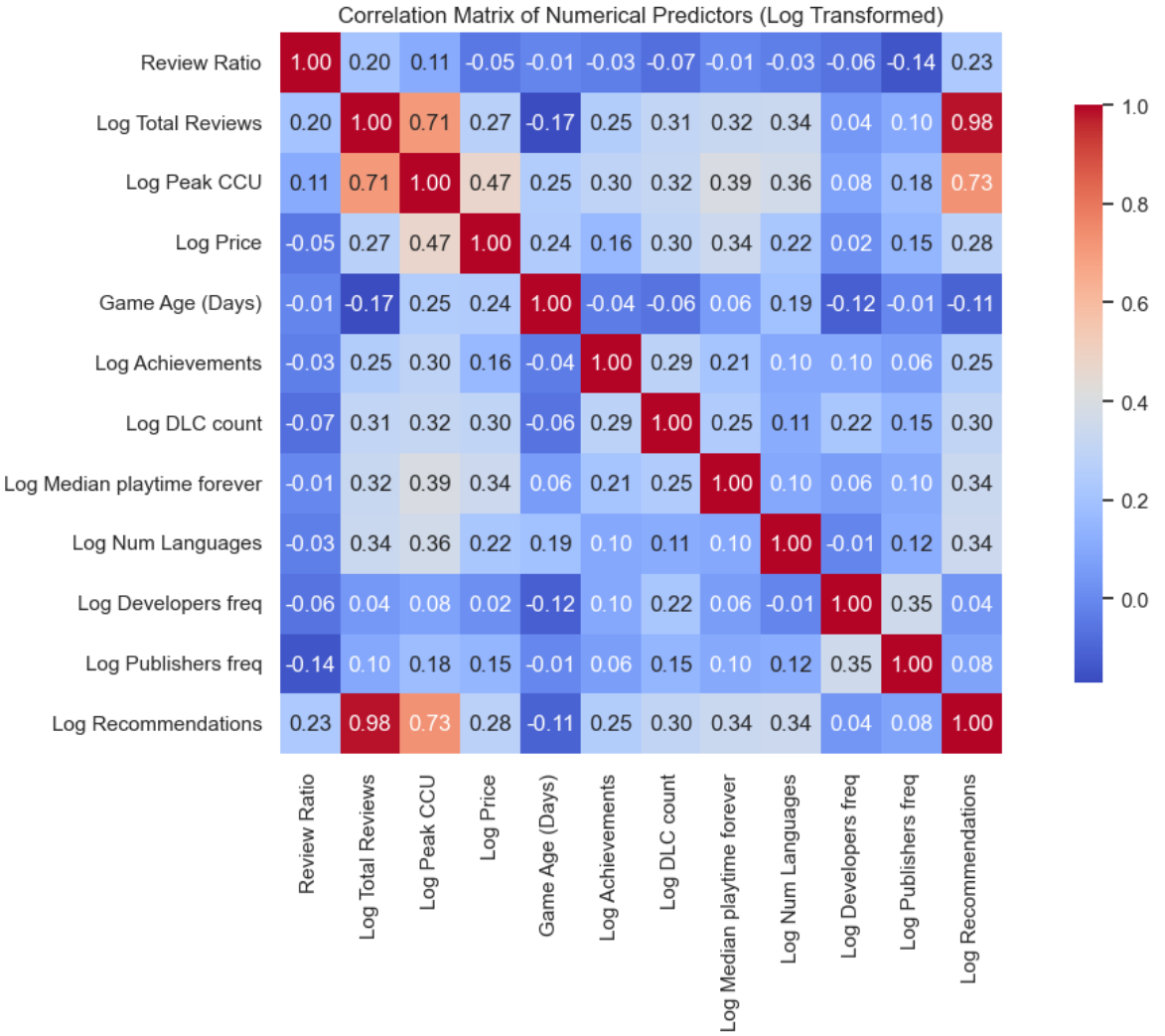


	Total Reviews	Peak CCU	Price	Achievements	DLC count	Me playt for
count	6.627200e+04	6.627200e+04	66272.000000	66272.000000	66272.000000	66272.00(
mean	1.437623e+03	2.353899e+02	7.963390	24.052677	0.644797	106.30!
std	3.188244e+04	9.314509e+03	10.622066	193.902606	15.499308	1267.64!
min	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	0.00(
25%	5.000000e+00	0.000000e+00	1.990000	0.000000	0.000000	0.00(
50%	2.100000e+01	0.000000e+00	4.990000	6.000000	0.000000	0.00(
75%	1.130000e+02	1.000000e+00	9.990000	21.000000	0.000000	0.00(
max	6.531097e+06	1.284268e+06	999.000000	9821.000000	2366.000000	136629.00(

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 [9]:
```
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
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 transform
cor_matr = new_df[new_numical_col].corr()

# Plotting the correlation matrix of numerical predictors after log transformati
plt.figure(figsize=(12, 8))
sns.heatmap(cor_matr, annot=True, fmt=".2f", cmap='coolwarm', square=True, cbar_
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	L
Review Ratio	1.000000	0.203457	0.109366	-0.048762	-0.009498	-0.029771	-0
Log Total Reviews	0.203457	1.000000	0.707803	0.271099	-0.168698	0.253570	0
Log Peak CCU	0.109366	0.707803	1.000000	0.472686	0.251193	0.295921	0
Log Price	-0.048762	0.271099	0.472686	1.000000	0.243926	0.160714	0
Game Age (Days)	-0.009498	-0.168698	0.251193	0.243926	1.000000	-0.036673	-0
Log Achievements	-0.029771	0.253570	0.295921	0.160714	-0.036673	1.000000	0
Log DLC count	-0.073591	0.308992	0.319430	0.300923	-0.063655	0.287506	1
Log Median playtime forever	-0.008537	0.324558	0.393869	0.341059	0.055484	0.209354	0
Log Num Languages	-0.029964	0.335962	0.364955	0.215668	0.192039	0.098757	0
Log Developers freq	-0.061416	0.043589	0.075250	0.017387	-0.118043	0.097635	0
Log Publishers freq	-0.141447	0.098684	0.177273	0.150570	-0.007614	0.064762	0
Log Recommendations	0.231703	0.984320	0.729956	0.280943	-0.108683	0.254534	0

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

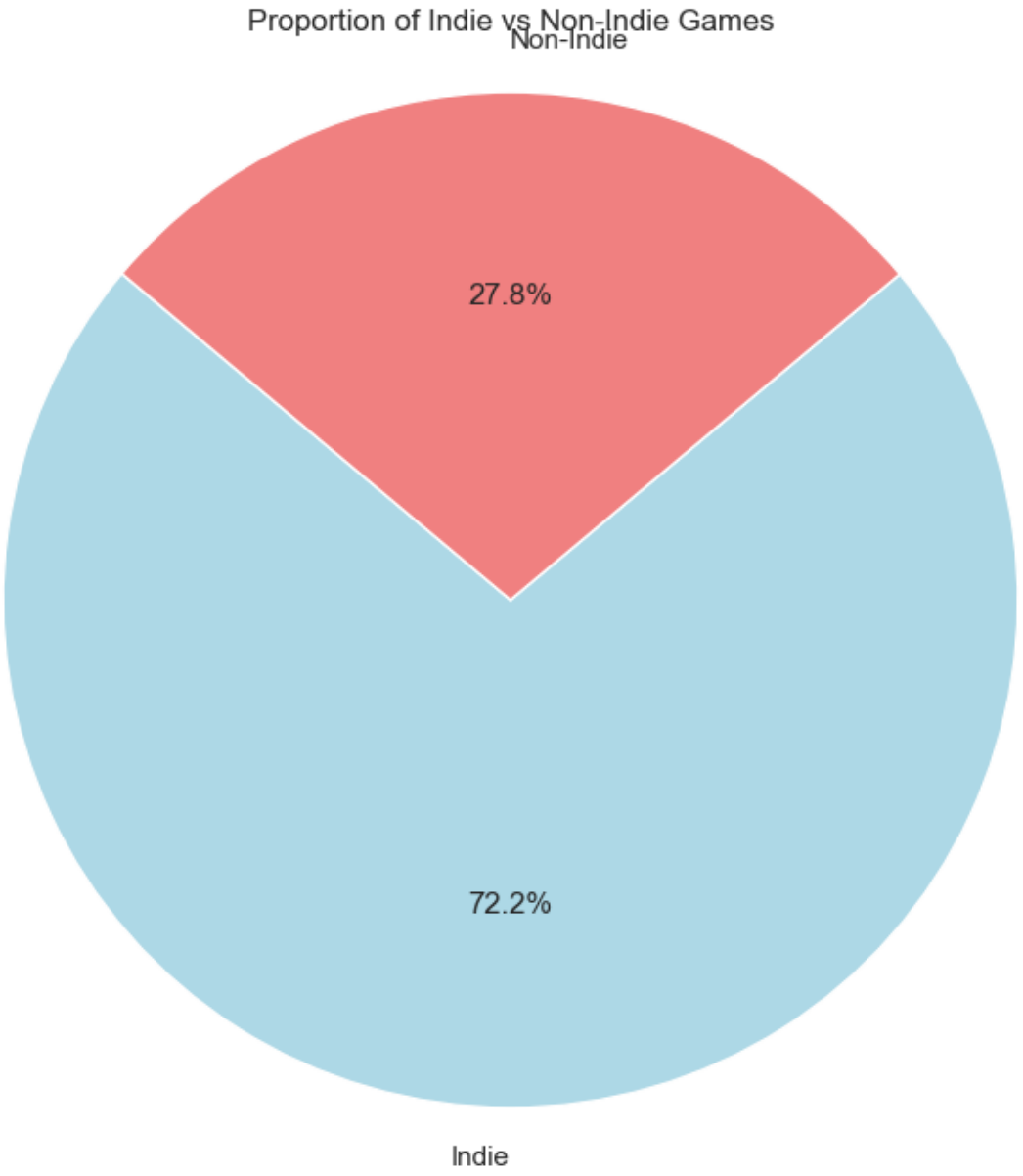
Therefore, the only useful numerical predictor we can utilize for providing **recommendations** here is **Peak CCU**.

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 [10]:
```
# pie chart of indie vs non-indie games
indie_count = df['is_indie'].value_counts()
```

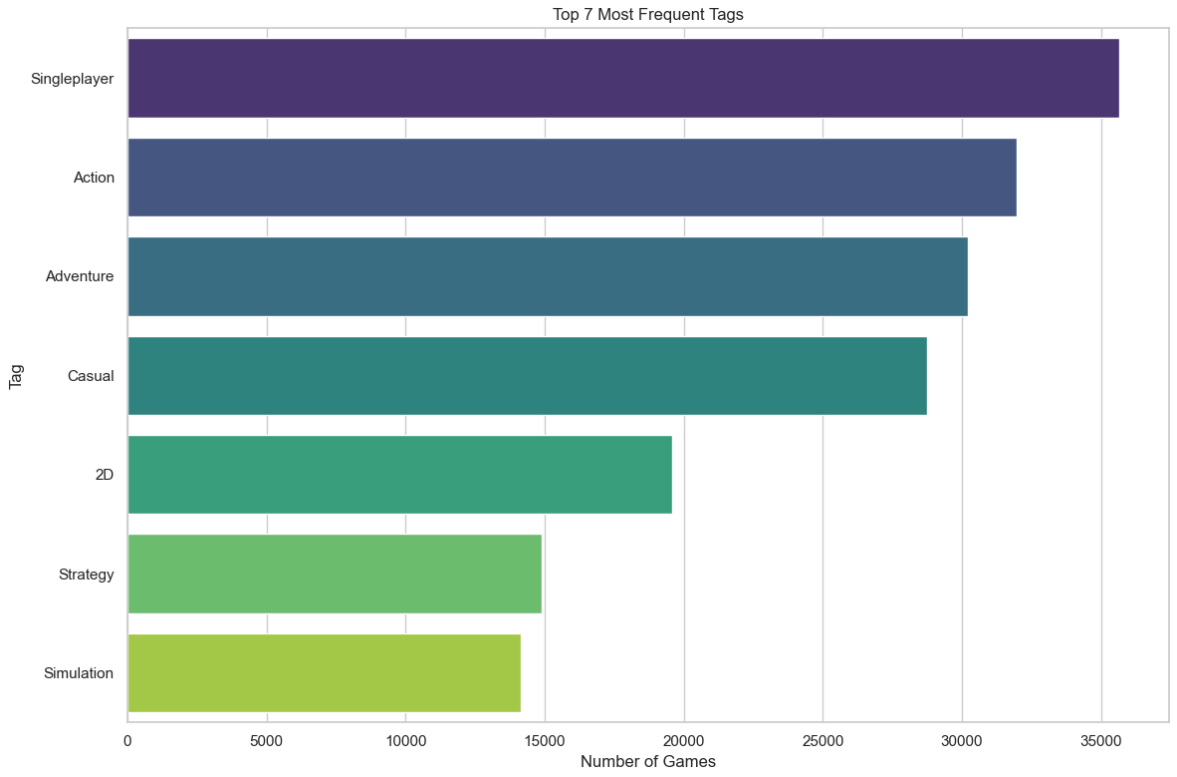
```
indie_count.index = ['Indie', 'Non-Indie']
plt.figure(figsize=(8, 8))
plt.pie(indie_count, labels=indie_count.index, autopct='%1.1f%%', startangle=140
plt.title('Proportion of Indie vs Non-Indie Games')
plt.axis('equal')
plt.show()
```

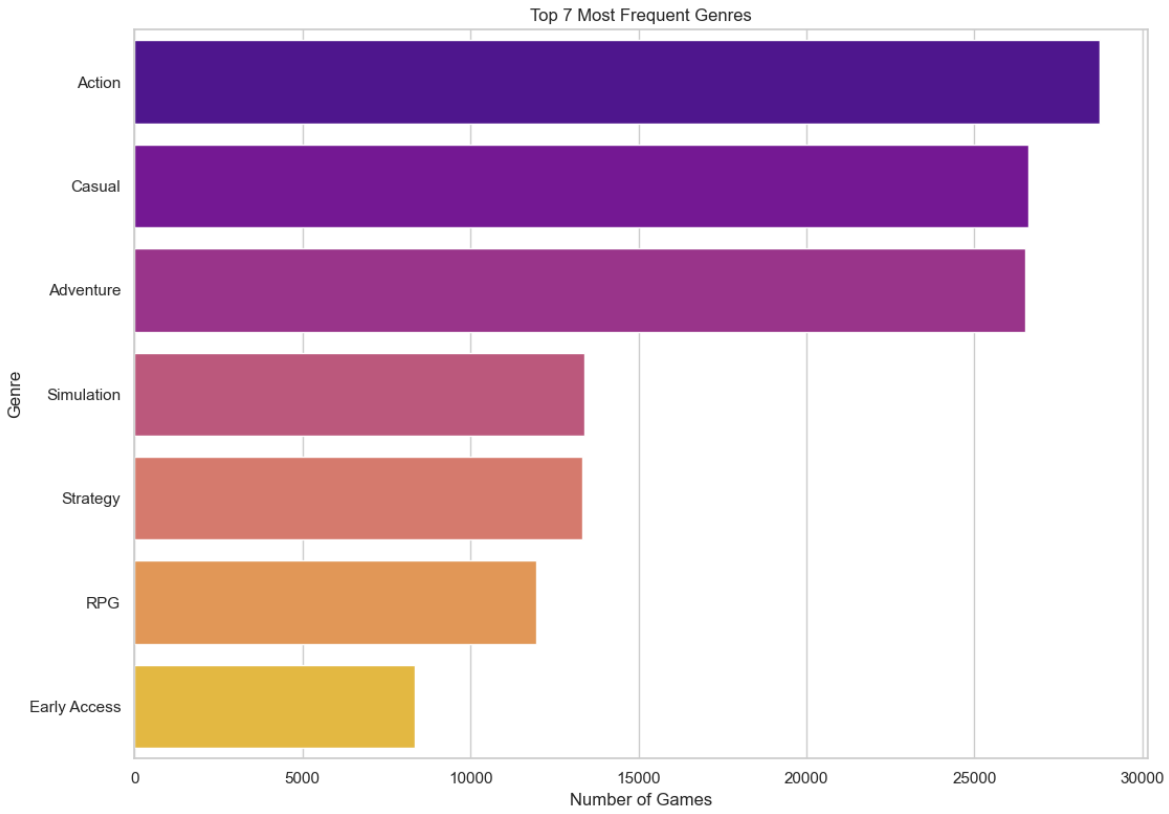


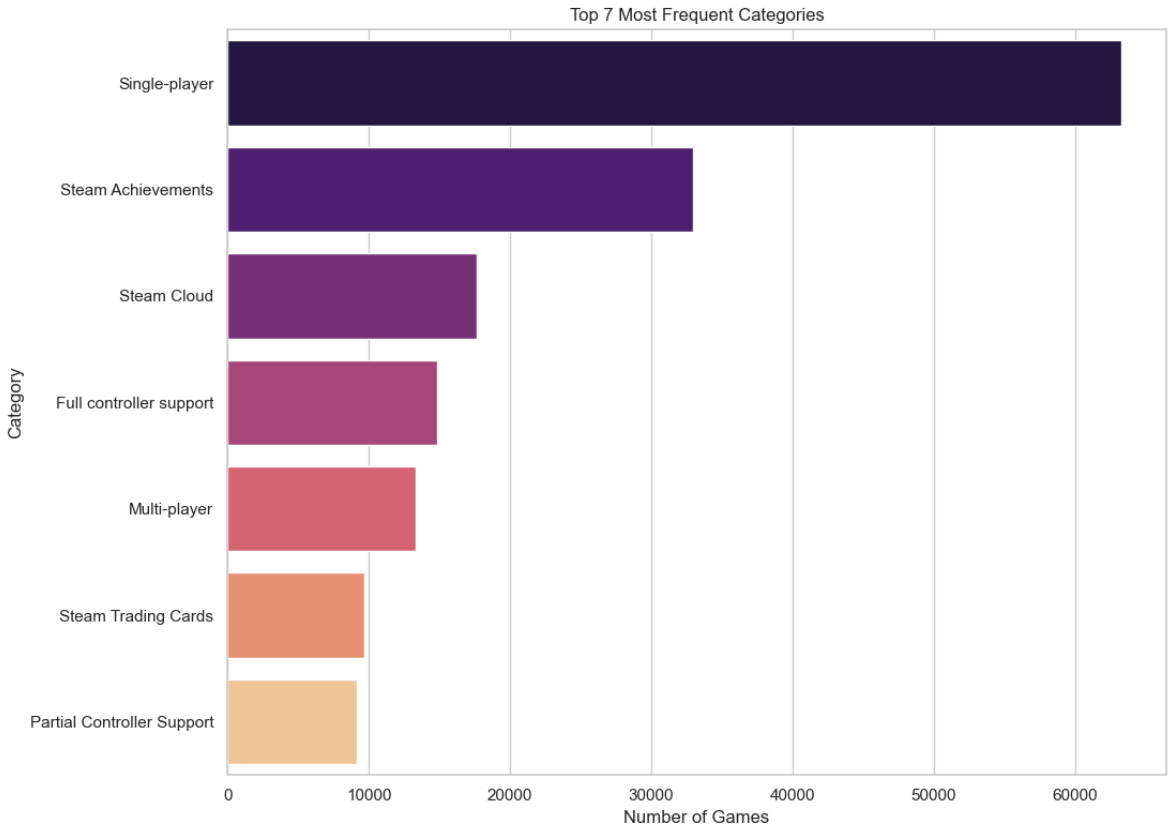
The graph actually shows that the majority games are indie games, meaning games developed by independent developers. Our recommendation model will have to take into account this information when recommending games.

In [22]:
```
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=
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[:
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', h
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.i
plt.legend([],[], frameon=False)
plt.show()

plt.figure(figsize=(12, 8))
sns.barplot(y=category_freq.index[:N], x=category_freq.values[:N], palette='magm
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 catego
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    35664
Tags_Action          31981
Tags_Adventure       30231
Tags_Casual          28736
Tags_2D              19593
dtype: int64
Top 5 Genres:
```

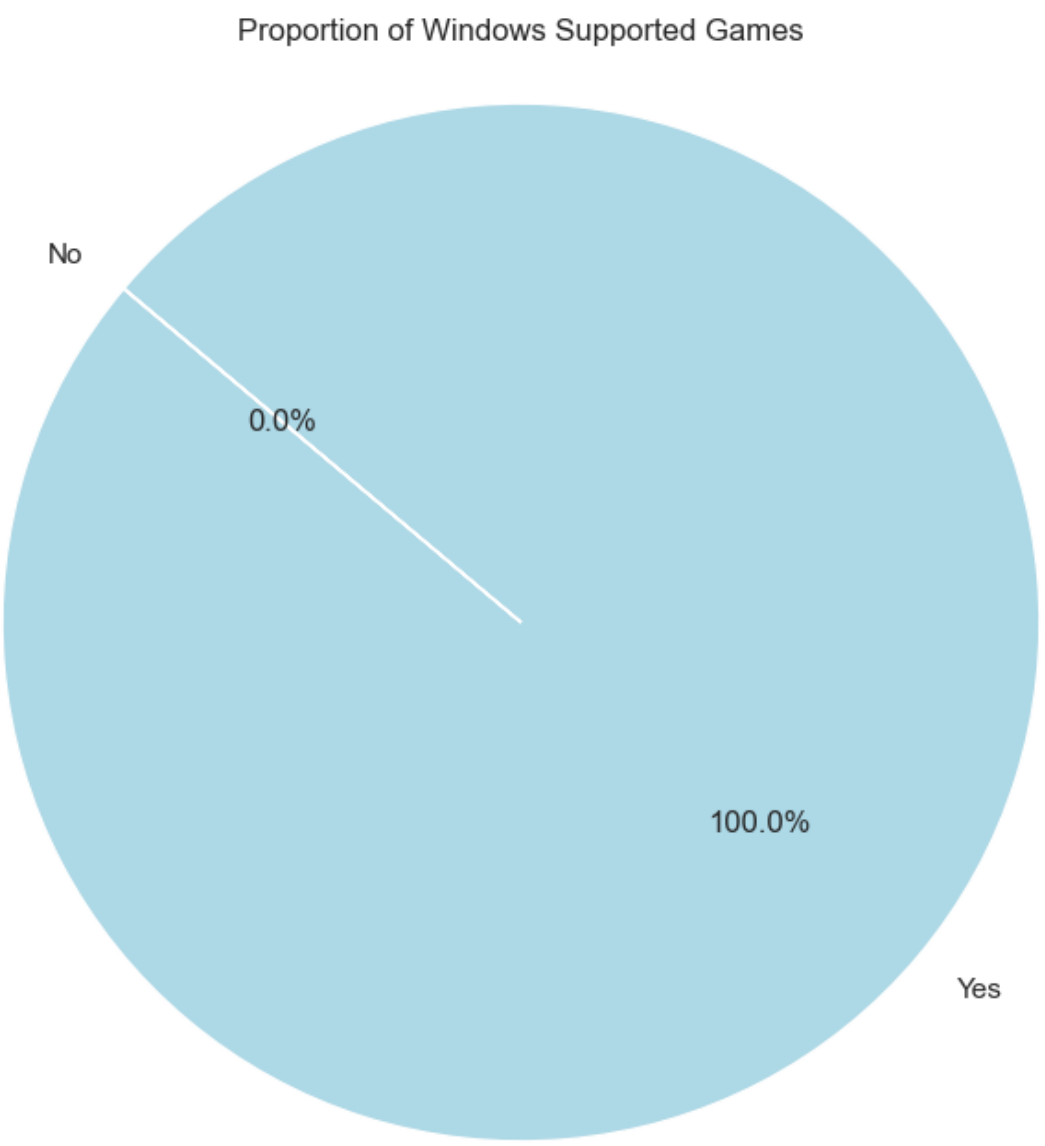
```
Genres_Action         28731
Genres_Casual         26622
Genres_Adventure      26514
Genres_Simulation     13399
Genres_Strategy       13330
dtype: int64
Top 5 Categories:
Categories_Single-player               63281
Categories_Steam Achievements          32989
Categories_Steam Cloud                 17647
Categories_Full controller support     14824
Categories_Multi-player                13286
dtype: int64
```

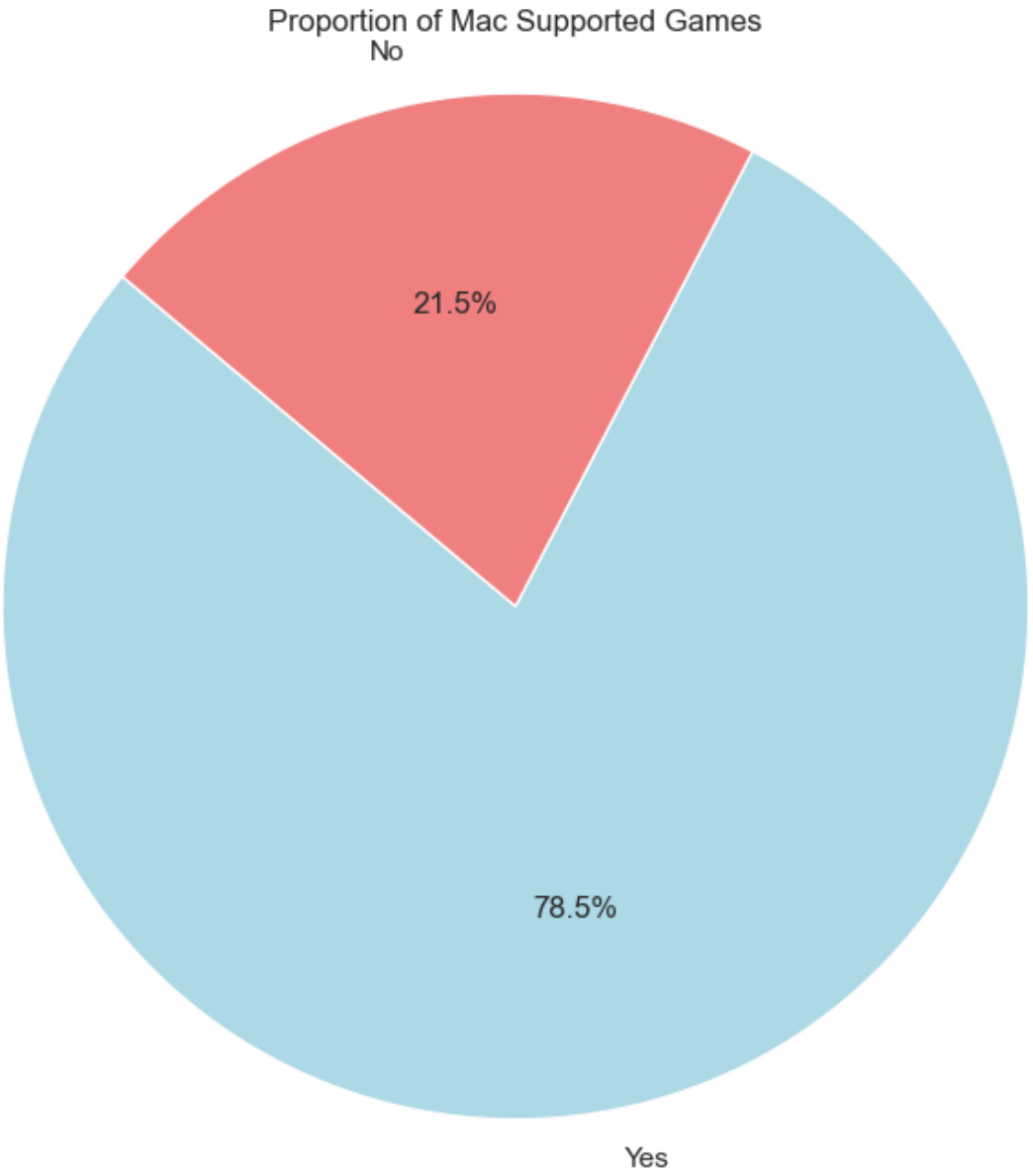
The plots show that 'Singleplayer', 'Adventure', 'Casual', and 'Action' are extremely common tags and genres, and 'Steam Achievements' and 'Steam Cloud' are very common categories.

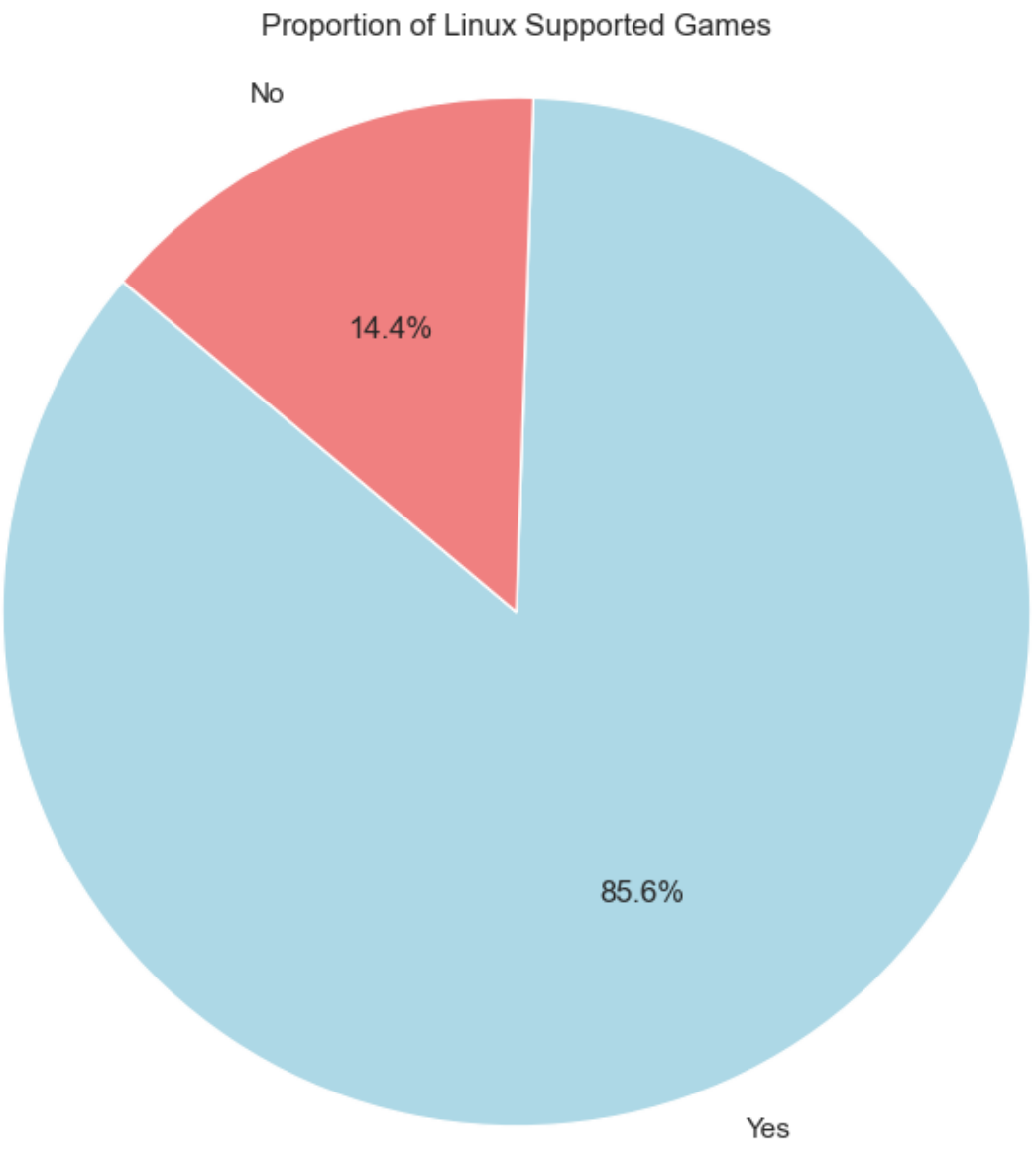
This indicates a large portion of the dataset consists of single-player indie games, often with casual or action elements. These common features means our recommendation models need to leverage the less common, more specific tags/genres/categories effectively to provide diverse and relevant suggestions.

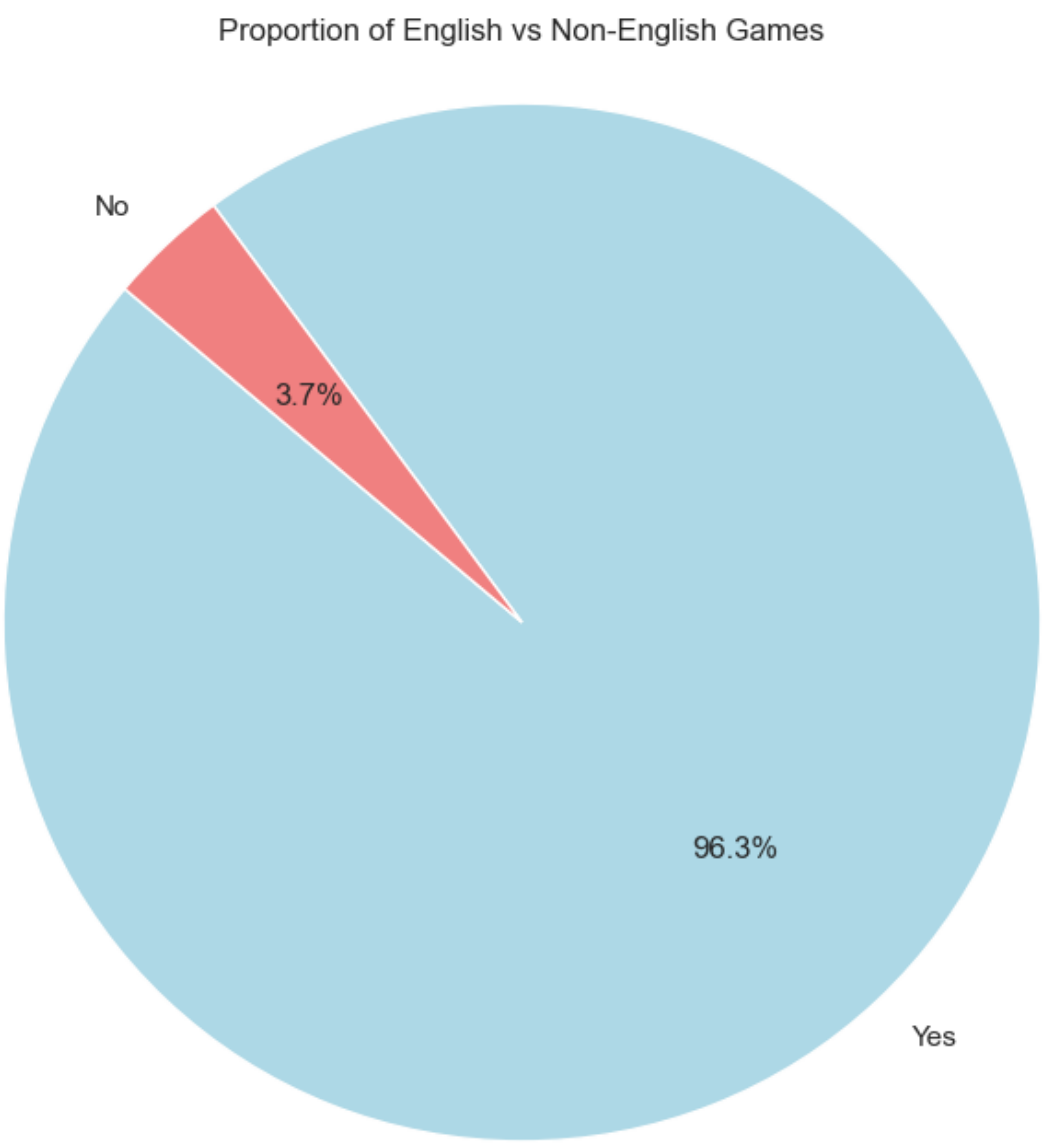
Next, we want to find out more about the distributions of game languages and platform supports in steam games

In [12]:
```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
    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
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

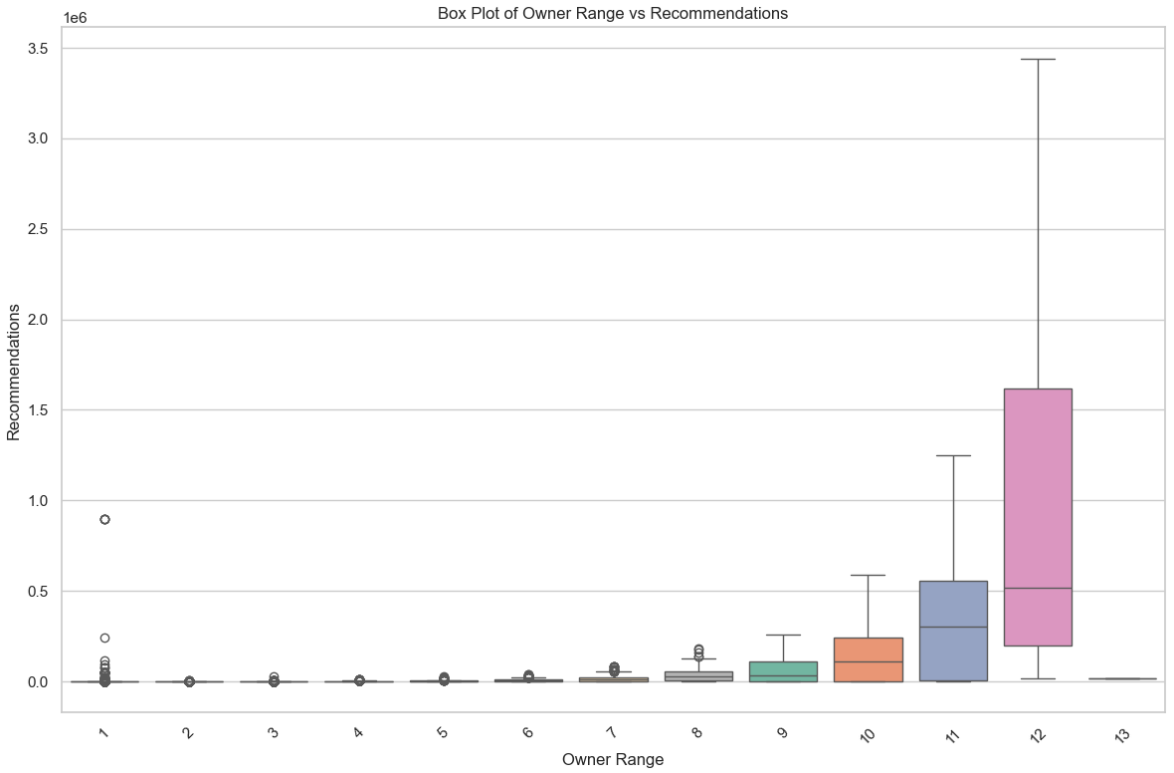



```
1    0.962503
0    0.037497
Name: Is English Supported, dtype: float64
True     0.999834
False    0.000166
Name: Windows, dtype: float64
False    0.784645
True     0.215355
Name: Mac, dtype: float64
False    0.855806
True     0.144194
Name: Linux, dtype: float64
```

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 [13]:
```
plt.figure(figsize=(12, 8))
sns.boxplot(x='Owner range', y='Recommendations', data=df, palette='Set2', hue='
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%	
1	47857.0	1.294155e+02	8.403412e+03	0.0	0.0	0.0	0.00	8
2	7909.0	1.439139e+02	2.785545e+02	0.0	0.0	0.0	205.00	
3	3915.0	4.057430e+02	7.607988e+02	0.0	0.0	158.0	561.50	
4	2566.0	8.979961e+02	1.242282e+03	0.0	0.0	447.5	1284.25	
5	2129.0	2.245640e+03	2.918576e+03	0.0	0.0	1291.0	3240.00	
6	900.0	6.026110e+03	6.593285e+03	0.0	471.0	4333.5	9122.25	
7	516.0	1.480264e+04	1.582332e+04	0.0	1856.0	10119.5	22343.50	
8	326.0	3.657755e+04	3.580609e+04	0.0	6402.0	28127.0	55573.00	1
9	90.0	6.597726e+04	7.461084e+04	0.0	1857.5	35424.0	112969.50	2
10	37.0	1.602234e+05	1.748687e+05	0.0	2469.0	110892.0	243925.00	5
11	21.0	3.384199e+05	3.667980e+05	1000.0	7699.0	305370.0	553709.00	12
12	5.0	1.157720e+06	1.419914e+06	14410.0	196597.0	519578.0	1616422.00	34
13	1.0	1.430000e+04	NaN	14300.0	14300.0	14300.0	14300.00	

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 [14]:
```
text_corpus = " ".join(str(review) for review in df['About the game'].fillna('')

# 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 s
text_corpus = re.sub(r'\s+', ' ', text_corpus).strip() # Remove extra spaces

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

wordcloud = WordCloud(stopwords=custom_stopwords, background_color="white", max_
plt.figure(figsize=(15, 7))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.title('Most Common Words in "About the game" Descriptions')
plt.show()

# Hellos
```

In [15]:
```
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))
```

```
Top 10 Most Common Words in 'About the game' Descriptions:
```

```
different     33783
time          32749
one           30289
unique        27512
find          27074
story         25445
levels        25124
experience    23773
mode          23518
enemies       23212
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

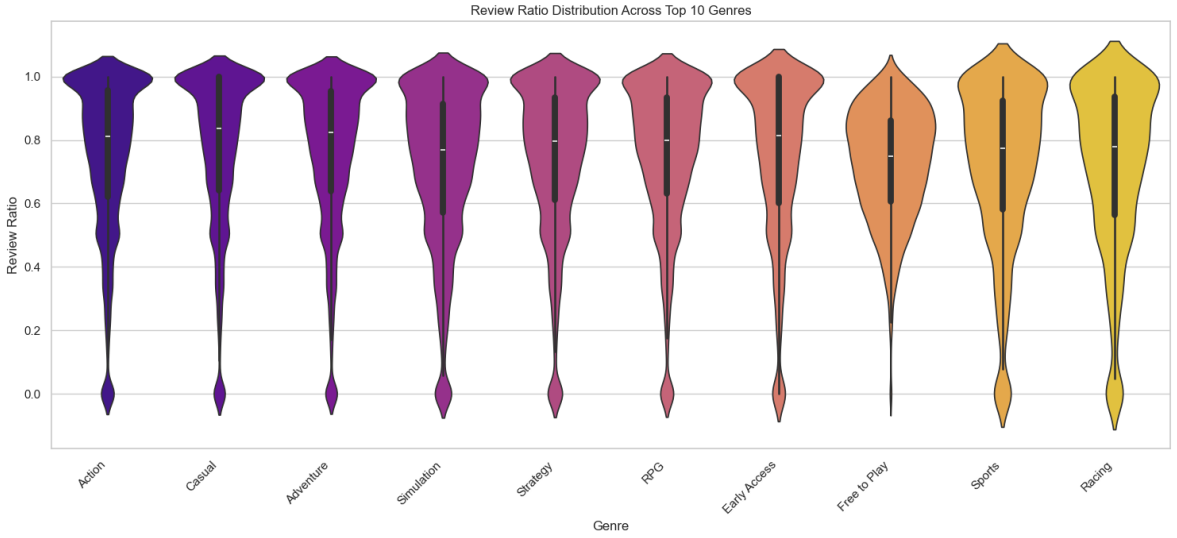
In general, the presence of these words indicates that the game is rich with terms describing the game mechanics, game structure, key features, selling points and player experience which is useful for our content-based models (Cosine Similarity and KNN), and that TF-IDF should be able to effectively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

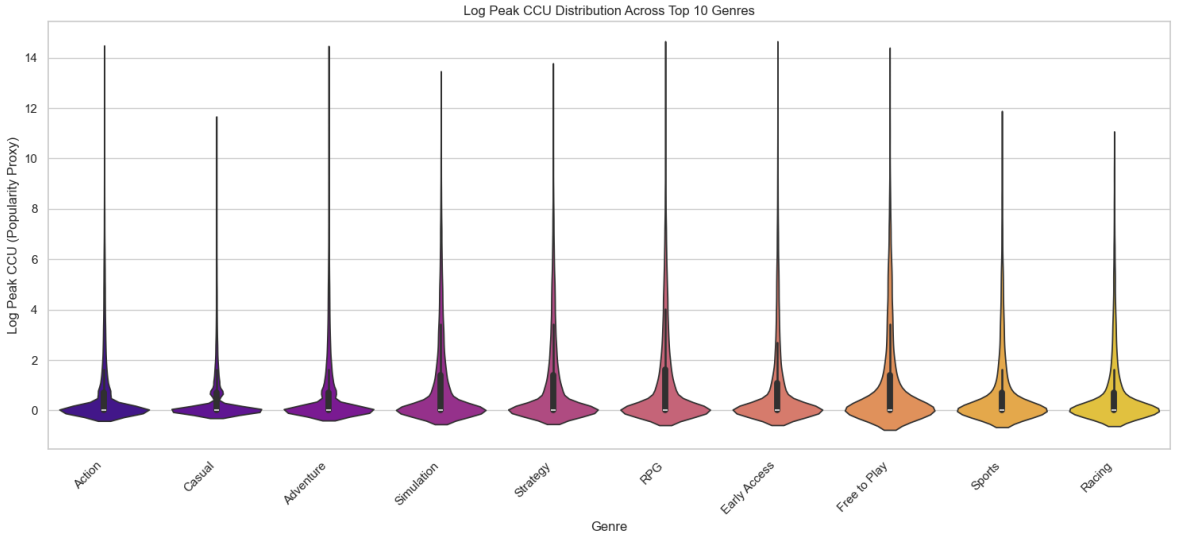
Cross-Feature Analysis

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 [18]:
```
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 le

melted_df = pd.melt(df_top_genres,
                    id_vars=['Review Ratio', 'Log Peak CCU'], # Use the log-tran
                    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', hu
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', hu
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 [17]:
```
# 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.