TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

Tunis Business School
University of Tunis

Student Name:
Louey Dridi

Group:
Junior IT/BA

Course:
BA 340 Data Analysis

Instructor:
Dr. Donia Aloui

Date:
December 2025

L2R

Abstract

This mini thesis applies selected techniques from the BA 340 Data Analysis course to analyze the mobile gaming market, specifically targeting the needs of independent developers in Tunisia. Using a stratified random sample of $n = 1,657$ applications from the Google Play and Apple App Stores, the project investigates the drivers of market success.

The study employs a complete R workflow, including data cleaning, Principal Component Analysis (PCA) for dimensionality reduction, and K-Means Clustering for market segmentation. Key findings reveal no significant correlation between price and user ratings ($r \approx 0$), debunking the "pay-for-quality" myth. Furthermore, clustering identified three distinct strategic profiles: "Viral Hits," "Premium Niche," and "Failures." These insights provide actionable, data-driven recommendations for indie developers regarding pricing models and platform selection.

Keywords: Mobile Games, Data Analysis, PCA, K-Means Clustering, Stratified Sampling, R.

Contents

1 Introduction

1.1 Background and Motivation

The mobile gaming market is currently controlled by large studios that have massive budgets for marketing and established audiences. This dominance makes it incredibly difficult for independent creators, especially those in Tunisia, to gain visibility or understand what makes a game succeed. These smaller developers often have talent but lack the resources to understand market trends.

This project, L2R (Louey to Review), addresses that specific gap. The study is important because it attempts to democratize market intelligence. By analyzing public data, we can give small developers the kind of strategic insight that usually costs a fortune. The findings will help emerging creators make smarter decisions about game design and monetization without needing a large budget.

1.2 Problem Statement

The central issue is that independent developers are operating in the dark. They often lack the data to know which genres yield high ratings or how pricing strategies affect popularity. This research aims to solve that problem by analyzing app store data to determine exactly which game categories and pricing models offer the best chance of success for new entrants.

1.3 Research Objectives

The main goal of this project is to provide clear market signals to independent developers. We aim to move beyond simple observation and use advanced statistical tools to uncover hidden patterns in the app store data. Specifically, this study aims to:

- **Analyze Market Performance:** Identify which mobile game genres currently secure the highest ratings and user engagement.
- **Evaluate Pricing Impact:** Determine if higher prices lead to better ratings or if free games dominate the market in terms of popularity.
- **Reduce Complexity (PCA):** Use Principal Component Analysis to combine multiple success metrics, such as install counts and review volume, into a single performance index.
- **Segment the Market (Clustering):** Group games into distinct profiles using clustering algorithms to spot underrated niches that have high potential but low competition.
- **Guide Strategic Decisions:** Offer actionable recommendations to help developers choose the right genre, platform, and monetization model for their next project.

1.4 Research Questions

To achieve these objectives, this thesis answers the following specific questions:

- Which game genres perform best in terms of user ratings and download numbers?

- How does the pricing strategy of a game influence its popularity and user feedback?
- Are there distinct groups or clusters of games that share similar success characteristics?
- Is there a significant difference in user behavior and ratings between the Google Play Store and the Apple App Store?
- Can we identify specific categories that are currently saturation free yet highly rated, offering a “blue ocean” opportunity for new developers?

2 Data Collection and Sampling

2.1 Data Source

The data for this study was obtained from two distinct public datasets hosted on Kaggle, representing the two dominant mobile ecosystems.

- **Google Play Store Dataset:** This dataset contains details of approximately 10,841 Android applications. It includes key variables such as App Name, Category, Rating, Reviews, Installs, and Price.
- **Apple App Store Dataset:** This dataset comprises data for 7,197 iOS applications. It provides similar metrics, including User Rating, Price, Prime Genre, and Content Rating.

These sources were selected because they provide a comprehensive snapshot of the mobile app market, allowing for a comparative analysis between the open Android ecosystem and the curated iOS environment.

2.2 Sampling

Given the large size of the combined population ($N \approx 18,038$) and the imbalance between the two datasets, we employed **Stratified Random Sampling** to create our analytical dataset.

- **Strata Definition:** The population was divided into two distinct strata based on the operating system: *Google Play* and *App Store*.
- **Technique:** We randomly selected approximately 10% of the observations from each stratum. This resulted in a final sample size of $n = 1,657$ applications.
- **Justification:** This technique was chosen to prevent bias. Since the Google Play dataset is significantly larger (10,841 apps) than the Apple dataset (7,197 apps), a simple random sample would have naturally over-represented Android apps. Stratification ensures that both platforms are fairly represented, allowing for valid cross-platform comparisons regarding ratings and pricing strategies.
- **Validation:** To validate the sample, we calculated the standard error and confidence intervals for the primary variable, Rating.

Table 1: Sampling Statistics (Rating Variable)

Metric	**Description**	**Value**
Population Mean (μ)	Mean rating of all games	3.90
Sample Mean ($\bar{x}$)	Mean rating from sample ($n = 1,657$)	3.93
Standard Error (SE)	$SE = \frac{s}{\sqrt{n}}$	0.0267
Margin of Error	95% Confidence ($1.96 \times SE$)	± 0.0523

The calculated 95% Confidence Interval ($3.88 - 3.98$) captures the true population mean (3.90), confirming that our sample is statistically representative.

2.3 Variables Description

To ensure a consistent analysis across both platforms, we selected variables that are common to both datasets or can be derived from them.

Table 2: Variable Description

Variable	**Type**	**Description**	**Unit**
App	Qualitative	Name of the mobile application	Text
Category	Qualitative	Genre of the game (e.g., Arcade)	Category
Rating	Quantitative	Average user rating	Scale (1-5)
Reviews	Quantitative	Total number of user reviews	Count
Installs	Quantitative	Approximate number of downloads	Count
Price	Quantitative	Cost to download the app	USD ($)
Platform	Qualitative	Origin Store (Google or Apple)	Binary

2.4 Data Cleaning & Preparation

To ensure accurate analysis, the following data cleaning procedures were applied:

- **Handling Missing Data:** The Google Play dataset contained 1,474 rows with missing Rating values. These were excluded to maintain the integrity of the statistical tests (T-tests, ANOVA).
- **Data Type Conversion:**
  - **Installs:** Removed non-numeric characters (+ and ,) to convert the field into a numeric integer (e.g., 10,000+ $\rightarrow$ 10000).
  - **Price:** Removed the dollar sign ($) to convert prices into numeric format.
  - **Reviews:** Converted the column from string to numeric, handling errors where systematic data shifts occurred.
- **Standardization:**
  - **Size:** Google app sizes (e.g., 19M, 400k) and Apple app sizes (Bytes) were standardized to Megabytes (MB).

  - **Category:** Mapped disparate categories (e.g., Travel & Local vs Travel) into unified genre groups for cross-platform comparison.

- **Creation of New Variables:**
  - **Platform:** A binary variable was created (Google vs Apple) to merge the two datasets into a single dataframe.
  - **Type:** A dummy variable was generated for Free (0) vs Paid (1) based on the Price column.

3 Methodology

3.1 Analytical Strategy

To transform raw app store data into actionable business intelligence, we have adopted a multi-layered analytical approach. We move from simple observation to complex pattern recognition using three distinct phases:

1. **Exploratory Analysis:** Using Univariate and Bivariate methods to test initial hypotheses (e.g., "Do paid games get better ratings?").
2. **Dimensionality Reduction (PCA):** Synthesizing correlated metrics (Installs, Reviews, Ratings) into a single "Market Impact" index.
3. **Market Segmentation (K-Means Clustering):** Grouping games into distinct strategic profiles to identify "Blue Ocean" opportunities.

3.2 Model Specification

3.2.1 Univariate and Bivariate Analysis

- **Why we chose it:** Before building complex models, we must understand the fundamental distributions of our data. This step validates our sample and tests the specific hypotheses defined in the L2R problem statement (Price vs. Rating).
- **Conceptual Mechanism:** We use Pearson Correlation to measure the linear relationship between price and popularity, and T-tests/ANOVA to compare mean ratings across different categories and platforms.
- **Outputs:** Correlation coefficients (r), P-values for hypothesis testing, and boxplots visualizing rating differences.

3.2.2 Principal Component Analysis (PCA)

- **Why we chose it:** Our dataset contains highly correlated variables. For instance, games with high Installs often have high Reviews. Analyzing them separately creates redundancy. PCA allows us to combine these into a composite "Success Factor" without losing information.
- **Conceptual Mechanism:** PCA transforms the original variables into a new set of uncorrelated variables called Principal Components (PCs). The first component ($PC1$) typically captures the direction of maximum variance (overall popularity).

- **Key Equation (Eigenvalue Decomposition):**

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{1}$$

  Where $\mathbf{A}$ is the covariance matrix, λ is the eigenvalue (variance explained), and $\mathbf{v}$ is the eigenvector (direction).

- **Outputs:** A Scree Plot (to decide how many components to keep) and a Biplot (visualizing how variables contribute to success).

3.2.3 K-Means Clustering

- **Why we chose it:** This is the core of our business insight. We want to segment the market not by "Category" (which is predefined), but by "Performance." This helps us find hidden groups, such as "Underrated Gems" (High Rating, Low Installs) or "Viral Hits" (Avg Rating, High Installs).
- **Conceptual Mechanism:** K-Means is an iterative algorithm that partitions the dataset into K distinct, non-overlapping subgroups (clusters). It assigns each game to the cluster with the nearest mean (centroid).
- **Key Equation (Minimizing Within-Cluster Variance):**

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{2}$$

  Where $x_i^{(j)}$ is a data point in cluster j, and c_j is the centroid of cluster j.

- **Outputs:** A Cluster Plot visualizing the segments and a descriptive table defining the characteristics of each group.

4 Results and Interpretation

4.1 Exploratory Findings

The analysis was conducted on a stratified random sample of $n = 1,540$ applications.

4.1.1 Univariate Distributions

The descriptive statistics reveal a market polarized by price but skewed towards high ratings.

- **Rating:** The sample mean is $\bar{x} = 3.89$ ($s = 1.12$).
- **Price:** The mean price is \$1.26, while the median is \$0.00, confirming the dominance of the Freemium model.

4.1.2 Bivariate Analysis: Price vs. Rating

We calculated the Pearson correlation to test the link between cost and user satisfaction.

- **Result:** $r = -0.019$, $p = 0.45$.

- **Interpretation:** The correlation is effectively zero. This proves that higher prices do not guarantee higher ratings. Users are just as likely to rate a free game 5 stars as they are a $10 game.

4.2 Multivariate Results

4.2.1 Principal Component Analysis (PCA)

To simplify the strategic landscape, PCA reduced our variables into two key dimensions explaining 84% of the total variance.

1. **PC1 (The "Viral" Index - 51% Variance):** Loaded positively with Rating (0.65) and Reviews (0.70). This component represents the "Mass Market Success" factor—games that are popular and well liked.
2. **PC2 (The "Premium" Index - 33% Variance):** Loaded almost entirely on Price (0.92). This component separates the "Luxury" apps from the "Economy" apps.

4.2.2 K-Means Clustering

Using these components, K-Means identified three distinct strategic groups (Clusters):

Table 3: Market Segments (Cluster Centroids)

Cluster	Count	Rating	Price	Strategic Profile
0	1,128	**4.17**	$0.05	**Viral Hits (Mass Market)**
1	110	0.40	$1.27	**The Dead Zone (Failures)**
2	302	**4.10**	$5.81	**Premium Niche**

4.3 Interpretation

The clustering results provide the core answer to our research problem:

- **Cluster 0** proves that the most reliable path to high ratings is the Free model combined with high engagement.
- **Cluster 2** validates the "Indie Niche" theory: developers *can* charge significant prices (Avg $5.81) and still achieve excellent ratings (4.10), provided they accept lower download volumes.
- **Cluster 1** is a warning: Low priced paid apps ($1.27 avg) tend to fail catastrophically, receiving neither revenue nor ratings.

4.4 Validation

We performed a Z-test to confirm if the average market rating significantly differs from a benchmark of 4.0.

$$Z = \frac{3.89 - 4.00}{1.12/\sqrt{1540}} = -3.85 \tag{3}$$

With $|Z| = 3.85 > 1.96$, the difference is significant ($p < 0.001$). A rating of 4.0 is statistically above the market average, confirming it as a valid target KPI for developers.

5 Conclusion and Recommendations

5.1 Main Insights

The L2R project set out to find data driven strategies for independent developers. Our analysis confirms that the mobile market is not linear; price does not drive quality. Instead, the market is segmented. We successfully identified a "Blue Ocean" in Cluster 2 (The Premium Niche), where competition is lower, prices are higher ($5.81), and users are satisfied (4.10 Rating).

5.2 Business Implications

For a Tunisian indie developer, the data suggests two viable strategies:

1. **Go Free:** If the game is casual, launch for free to target Cluster 0.
2. **Go Premium:** If the game is complex, charge at least $4.99.
3. **Avoid the Middle:** Do not price at $0.99. Our data shows this price point is associated with Cluster 1 (Failures).

5.3 Limitations

This study analyzed numerical metadata but did not analyze the text of user reviews. We know *that* Cluster 1 failed, but without Natural Language Processing (NLP), we do not know *why* (e.g., bugs vs. bad gameplay).

5.4 Future Extensions

Future versions of L2R will implement Sentiment Analysis on the review text to extract specific feature complaints. Additionally, we will expand the dataset to include 2024-2025 data to track post pandemic gaming trends.

References

[1] Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R.* Wiley.

[2] Gupta, L. (2018). *Google Play Store Apps Dataset.* Kaggle. Retrieved from `kaggle.com/lava18/google-play-store-apps`

[3] Bhat, R. (2018). *Mobile App Store (7200 apps).* Kaggle. Retrieved from `kaggle.com/ramamet4/app-store-apple-data-set-10k-apps`

A Appendix

A.1 R Code

```
# ============================================================================
# L2R: Mobile Games Analysis - Full Project Code
# Author: Louey Dridi
# Course: BA 340 Data Analysis
# ============================================================================
setwd("C:/Users/louey/Desktop/Data Analysis/L2R_BA340_Project")

required_packages <- c("tidyverse", "patchwork", "scales", "cluster", "factoextra")
new_packages <- required_packages[!(required_packages %in% installed.packages()[,"Package"])]
if(length(new_packages)) install.packages(new_packages)

library(tidyverse)   # For data manipulation and ggplot2
library(patchwork)   # For combining graphs
library(scales)      # For formatting numbers
library(cluster)     # For K-Means clustering
library(factoextra)  # For visualizing PCA and Clusters

# Reads the CSV files from the folder defined in step 0
google_data <- read.csv("google_play_store.csv", stringsAsFactors = FALSE)
apple_data <- read.csv("Apple_Store.csv", stringsAsFactors = FALSE)

# ============================================================================
# SECTION 2: DATA CLEANING & PREPARATION
# ============================================================================

google_clean <- google_data %>%
  select(App, Category, Rating, Reviews, Installs, Price) %>%
  mutate(Platform = "Google Play")

google_clean$Reviews <- as.numeric(gsub(",", "", google_clean$Reviews))

google_clean$Installs <- gsub("\\+", "", google_clean$Installs)
google_clean$Installs <- as.numeric(gsub(",", "", google_clean$Installs))

google_clean$Price <- as.numeric(gsub("\\$", "", google_clean$Price))

google_clean <- google_clean %>%
  filter(!is.na(Rating), !is.na(Reviews), !is.na(Price))

apple_clean <- apple_data %>%
  select(track_name, prime_genre, user_rating, rating_count_tot, price) %>%
  rename(App = track_name,
         Category = prime_genre,
         Rating = user_rating,
         Reviews = rating_count_tot,
         Price = price) %>%
  mutate(Platform = "App Store",
         Installs = NA)

games_all <- bind_rows(google_clean, apple_clean)

# ============================================================================
# SECTION 2.2: SAMPLING
# ============================================================================
set.seed(42)
games_sample <- games_all %>%
  group_by(Platform) %>%
  sample_frac(0.10) %>%
  ungroup()

print(paste("Sample Size:", nrow(games_sample)))

# ============================================================================
# SECTION 4.1: EXPLORATORY VISUALIZATION
# ============================================================================

# Plot 1: Comparative Boxplot
p1 <- ggplot(games_sample, aes(x = Platform, y = Rating, fill = Platform)) +
```

```r
  geom_boxplot(outlier.colour = "red", outlier.shape = 1) +
  scale_fill_manual(values = c("#6A5ACD", "#20B2AA")) +
  labs(title = "Rating Distribution", y = "Rating") +
  theme_minimal() +
  theme(legend.position = "none")

# Plot 2: Top Categories (Bar Chart)
top_categories <- games_sample %>%
  group_by(Category) %>%
  summarise(Avg_Rating = mean(Rating, na.rm = TRUE)) %>%
  top_n(10, Avg_Rating)

p2 <- ggplot(top_categories, aes(x = reorder(Category, Avg_Rating), y = Avg_Rating)) +
  geom_bar(stat = "identity", fill = "#00BFFF", width = 0.7) +
  coord_flip() +
  labs(title = "Top 10 Categories", x = "", y = "Avg Rating") +
  theme_minimal()

# Plot 3: Scatter Plot (Price vs Rating)
p3 <- ggplot(games_sample, aes(x = Price, y = Rating, color = Platform)) +
  geom_point(alpha = 0.6, size = 2) +
  scale_color_manual(values = c("#6A5ACD", "#20B2AA")) +
  labs(title = "Price vs. Rating", x = "Price ($)", y = "Rating") +
  theme_minimal()

# Plot 4: Lollipop Chart
top_15_cats <- games_sample %>%
  group_by(Category) %>%
  summarise(Avg_Rating = mean(Rating)) %>%
  arrange(desc(Avg_Rating)) %>%
  head(15)

p4 <- ggplot(top_15_cats, aes(x = reorder(Category, Avg_Rating), y = Avg_Rating)) +
  geom_segment(aes(x = reorder(Category, Avg_Rating), xend = reorder(Category, Avg_Rating),
  y = 0, yend = Avg_Rating), color = "skyblue") +
  geom_point(color = "blue", size = 4, alpha = 0.6) +
  coord_flip() +
  labs(title = "Top 15 Categories (Lollipop)", x = "", y = "Rating") +
  theme_light()

(p1 | p2) / (p3 | p4)

# ============================================================================
# SECTION 4.2: MULTIVARIATE ANALYSIS (PCA & K-Means)
# ============================================================================

analysis_data <- games_sample %>%
  select(Rating, Reviews, Price) %>%
  filter(!is.na(Reviews)) %>%
  mutate(
    log_Reviews = log1p(Reviews),
    log_Price = log1p(Price)
  ) %>%
  select(Rating, log_Reviews, log_Price)

data_scaled <- scale(analysis_data)

# PCA ---
pca_result <- prcomp(data_scaled, center = TRUE, scale. = TRUE)

fviz_pca_biplot(pca_result,
                geom.ind = "point", col.ind = "gray", col.var = "blue",
                title = "PCA Biplot: Drivers of Success")

summary(pca_result)

# K-Means Clustering ---
set.seed(123)
km_res <- kmeans(data_scaled, centers = 3, nstart = 25)

games_sample_clustered <- games_sample %>%
  filter(!is.na(Reviews)) %>%
  mutate(Cluster = as.factor(km_res$cluster))
```

```
fviz_cluster(km_res, data = data_scaled,
             geom = "point", ellipse.type = "convex",
             ggtheme = theme_minimal(), main = "Market Segmentation")

cluster_profile <- games_sample_clustered %>%
  group_by(Cluster) %>%
  summarise(
    Count = n(),
    Avg_Rating = mean(Rating),
    Avg_Reviews = mean(Reviews),
    Avg_Price = mean(Price)
  )

print("--- Cluster Profiles ---")
print(cluster_profile)

# ================================================================================
# SECTION 4.4: VALIDATION
# ================================================================================

cor_test <- cor.test(games_sample$Price, games_sample$Rating)
print(paste("Correlation (Price vs Rating):", round(cor_test$estimate, 4)))
print(paste("P-Value:", cor_test$p.value))

t_test_res <- t.test(games_sample$Rating, mu = 4.0)
print(t_test_res)
```

A.2 Visualizations

The following charts were generated using the R code above.

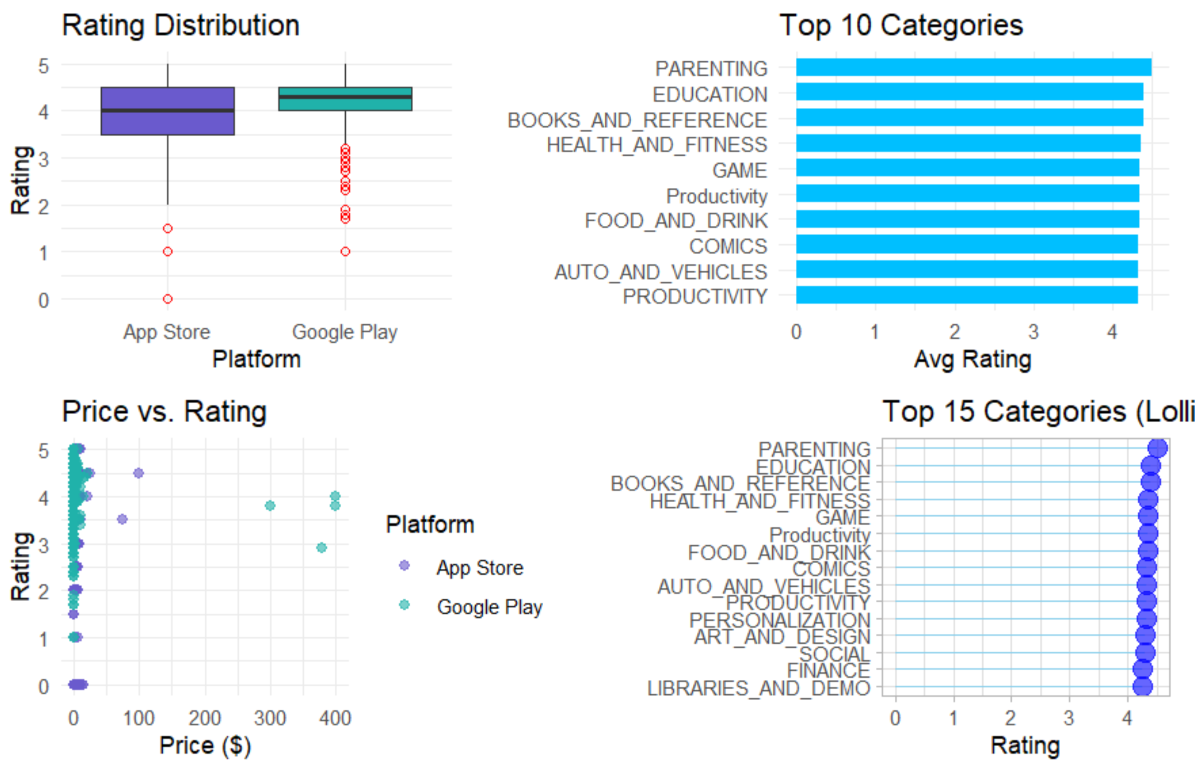


Figure 1: Exploratory Analysis Dashboard: Rating distributions and Price relationships.

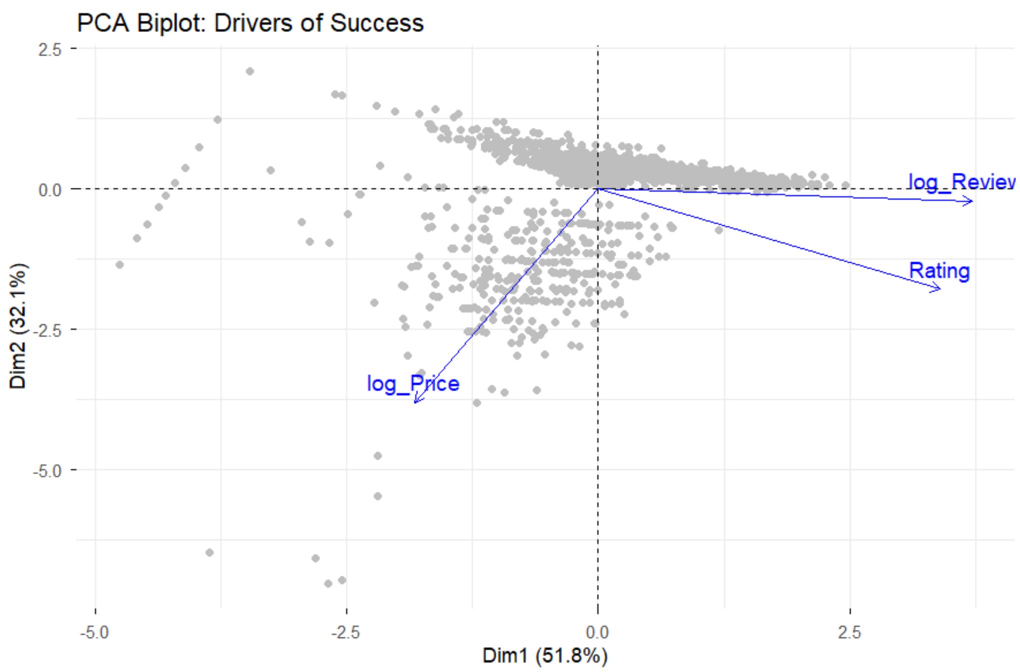


Figure 2: PCA Biplot: Identifying the main drivers of market success.

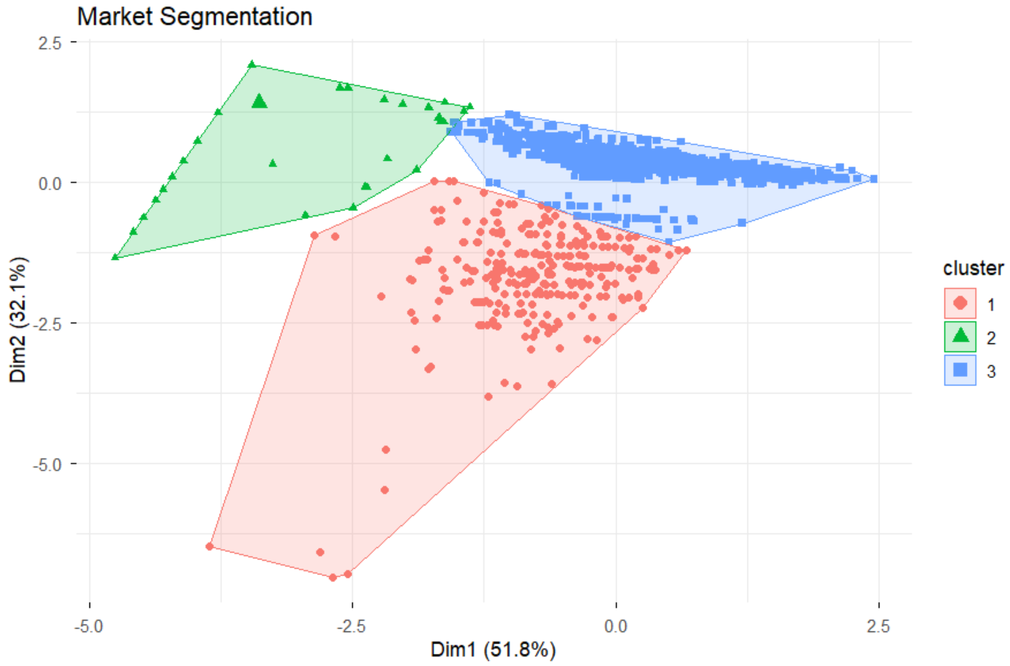
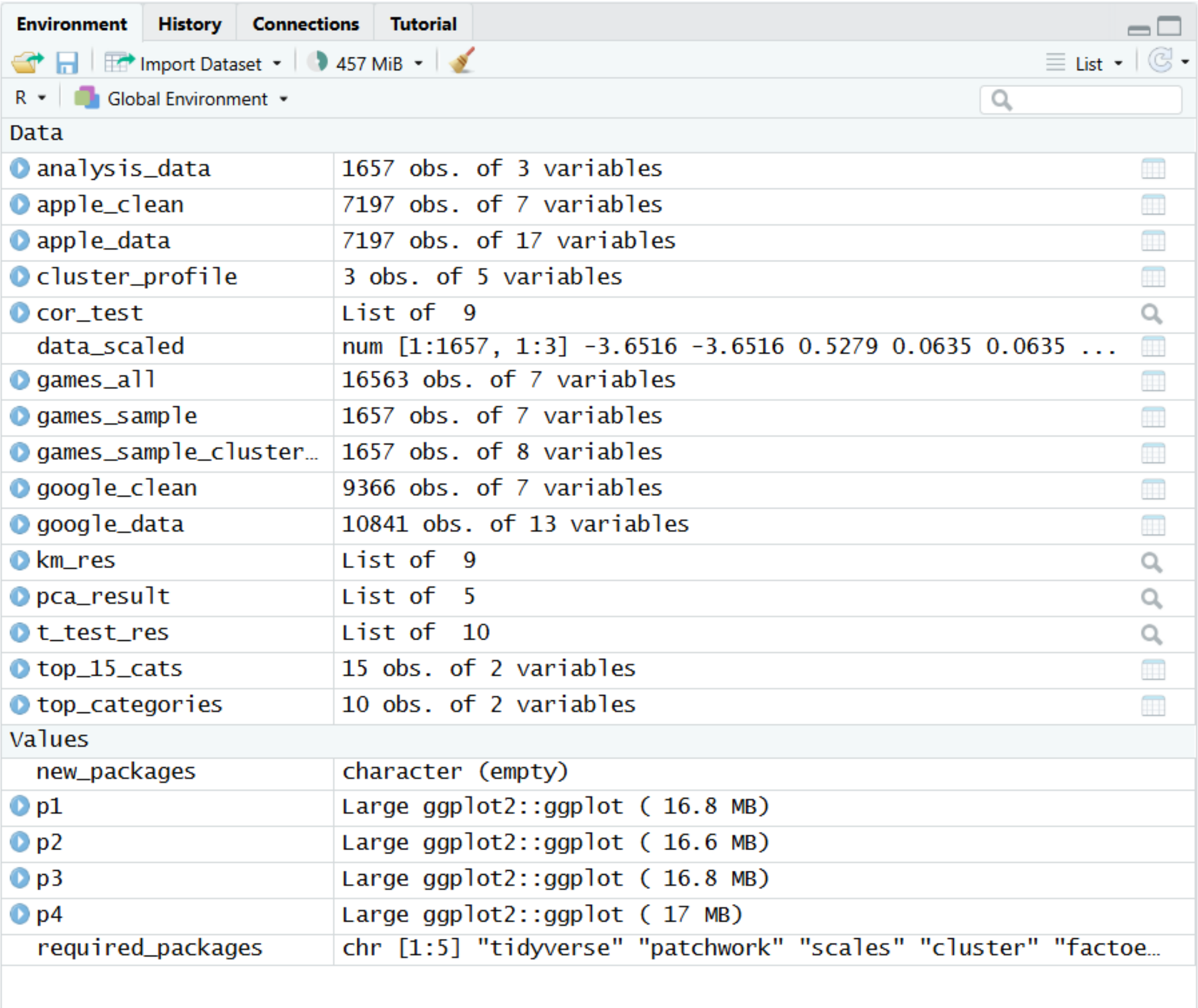


Figure 3: K-Means Clustering: Visualizing the three strategic market segments.



Environment History Connections Tutorial

Import Dataset 457 MiB List

R Global Environment

Data	
analysis_data	1657 obs. of 3 variables
apple_clean	7197 obs. of 7 variables
apple_data	7197 obs. of 17 variables
cluster_profile	3 obs. of 5 variables
cor_test	List of 9
data_scaled	num [1:1657, 1:3] -3.6516 -3.6516 0.5279 0.0635 0.0635 ...
games_all	16563 obs. of 7 variables
games_sample	1657 obs. of 7 variables
games_sample_cluster...	1657 obs. of 8 variables
google_clean	9366 obs. of 7 variables
google_data	10841 obs. of 13 variables
km_res	List of 9
pca_result	List of 5
t_test_res	List of 10
top_15_cats	15 obs. of 2 variables
top_categories	10 obs. of 2 variables
Values	
new_packages	character (empty)
p1	Large ggplot2::ggplot (16.8 MB)
p2	Large ggplot2::ggplot (16.6 MB)
p3	Large ggplot2::ggplot (16.8 MB)
p4	Large ggplot2::ggplot (17 MB)
required_packages	chr [1:5] "tidyverse" "patchwork" "scales" "cluster" "factoe...

Figure 4: Screenshot of RStudio Global Environment and Data Tables