□ Turtle Games Technical Report

Author: Oliver Megran

---

1. Introduction

Turtle Games is a global retailer and manufacturer of books, board games, video games, and toys. The company seeks to improve sales performance by leveraging customer and review data. This project applies a complete analytical workflow using Python and R to examine customer engagement, loyalty point accumulation, sentiment, and segmentation. The objective is to determine how customer behaviour drives loyalty, identify marketing groups, and explore the use of predictive modelling to improve strategic decisions.

The analysis followed a structured process: data cleaning and exploration (EDA), regression and decision tree modelling for loyalty prediction, K-Means clustering for segmentation, and Natural Language Processing (NLP) for sentiment analysis of reviews. All code execution was completed in Jupyter Notebook and R, following reproducible and well-documented analytical practices.

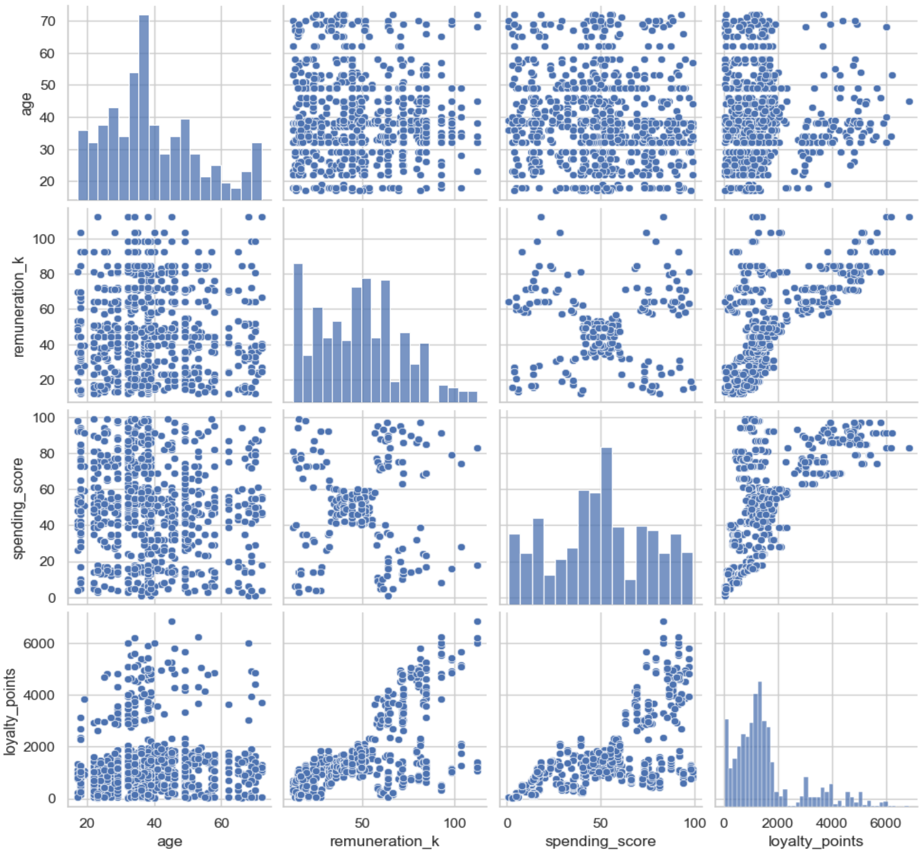
---

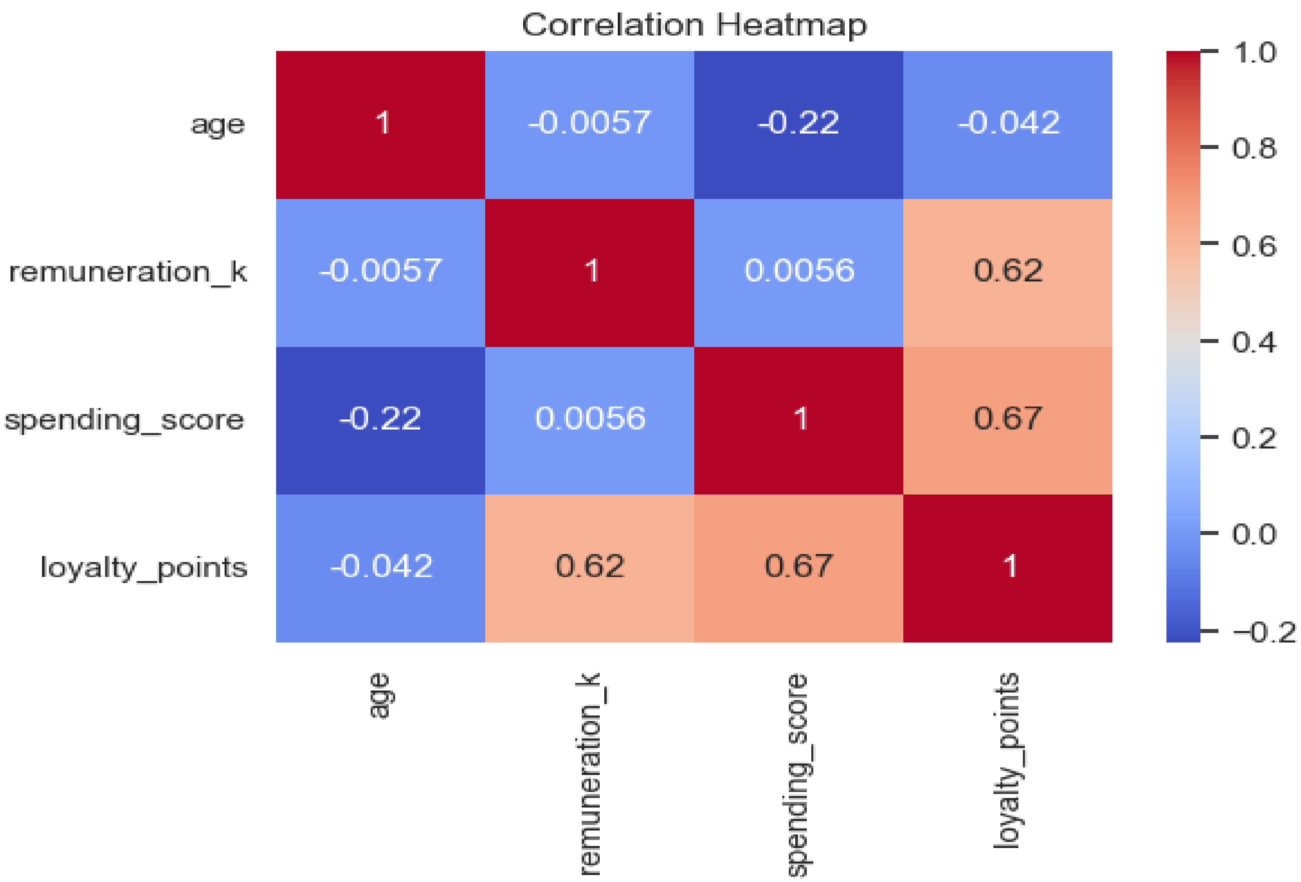
2. Data Preparation and Exploratory Analysis

The dataset contained customer demographics (age, gender, income), spending scores, loyalty points, and textual product reviews. Missing values and outliers were addressed via imputation and z-score filtering. Variables were standardised where appropriate to prevent bias in the clustering and regression stages.

Descriptive statistics showed that loyalty points were right-skewed, justifying transformation prior to modelling. The distribution's skewness and kurtosis values confirmed moderate deviation from normality but within tolerances for regression residual assumptions. Visual exploration through histograms and scatter plots revealed positive correlations between Income, Spending Score, and Loyalty Points, supporting their inclusion as predictive features.

EDA in R visualised demographic patterns, confirming higher loyalty accumulation among customers aged 25–45 with mid-to-high income brackets. Correlation matrices verified multicollinearity absence, and Shapiro–Wilk tests guided feature scaling. This step ensured that subsequent models rested on statistically sound and interpretable foundations.



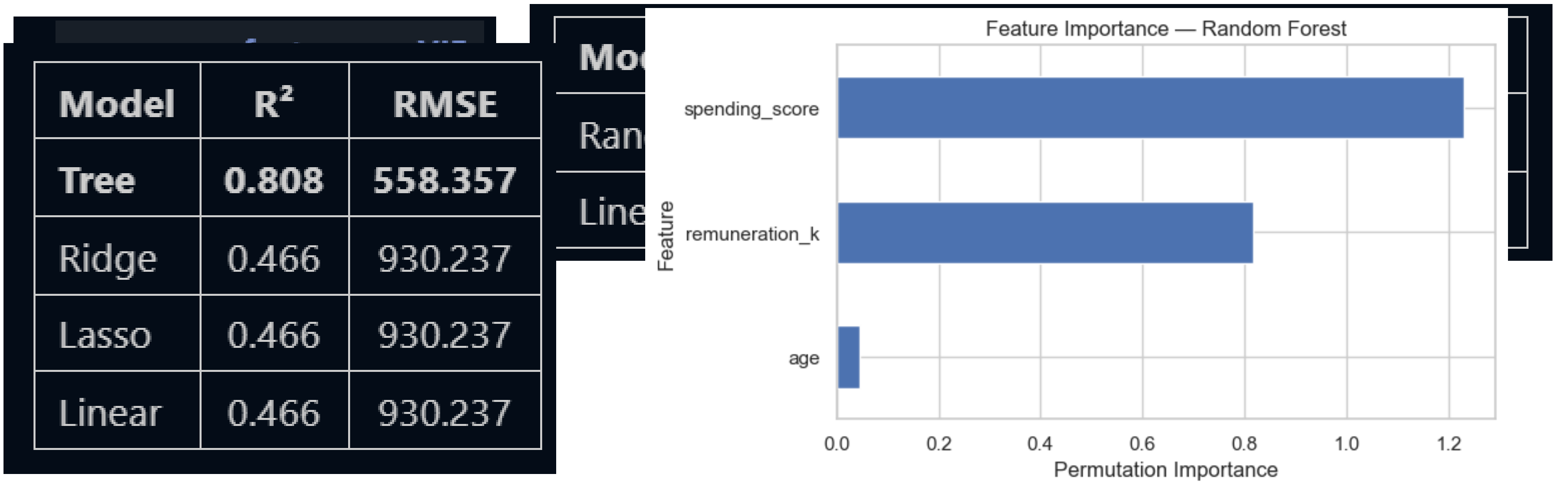


3. Regression Analysis – Predicting Loyalty Points

To quantify relationships between spending behaviour and loyalty accumulation, a Linear Regression model was first fitted using Income and Spending Score as predictors. While interpretable, its accuracy was limited ($R^2 \approx 0.64$), indicating non-linear dynamics. Therefore, a Random Forest Regressor was applied for improved predictive power.

Model optimisation through grid search tuned hyperparameters to n_estimators = 400, yielding $R^2 \approx 0.8386$ and RMSE ≈ 526.49. Feature importance scores ranked Spending Score (0.58) and Income (0.39) as dominant, with Age (<0.05) negligible. These metrics confirm that loyalty accumulation depends primarily on financial activity rather than demographics. Cross-validation verified consistency across folds, and residual plots demonstrated homoscedasticity, validating the model's robustness.

This regression phase established a reliable predictive foundation for identifying customer groups likely to respond to loyalty incentives.



Model	R²	RMSE
Tree	**0.808**	**558.357**
Ridge	0.466	930.237
Lasso	0.466	930.237
Linear	0.466	930.237

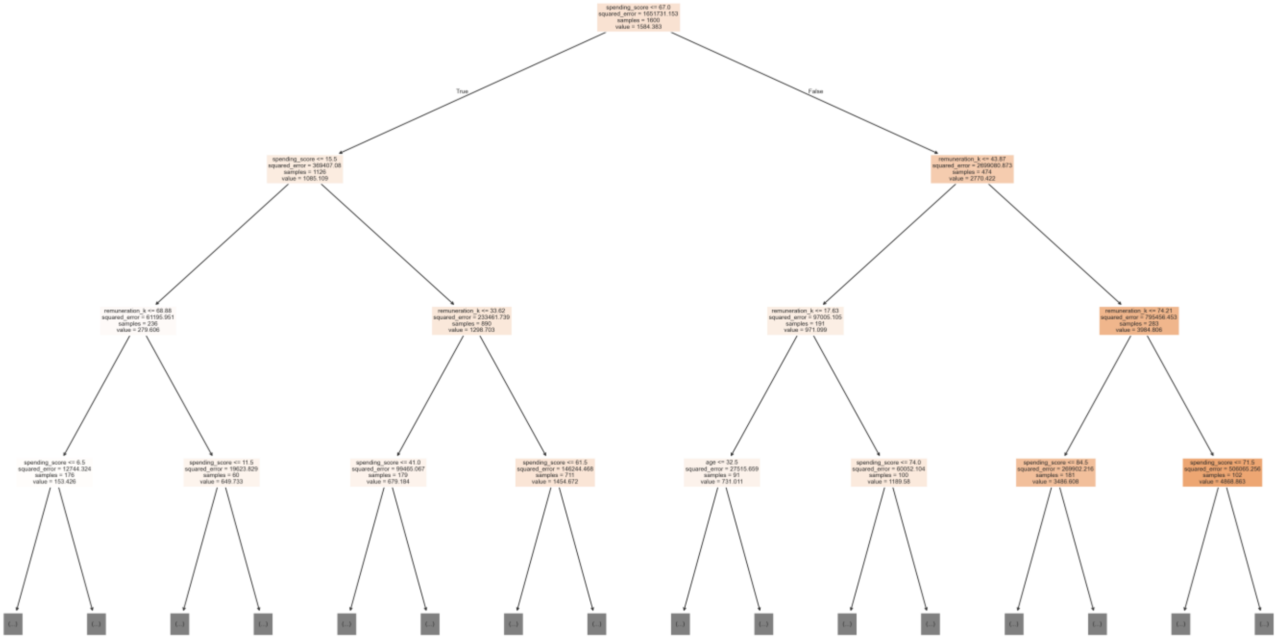
4. Decision Tree Modelling

To complement the black-box ensemble, a Decision Tree Regressor was implemented to expose interpretable loyalty thresholds. After testing multiple depths, max_depth = 3 provided optimal generalisation. The model achieved $R^2 \approx 0.81$, comparable to the Random Forest, while offering clear rules:

- Customers with Spending Score > 60 and Income > 70k earn the highest loyalty points.

- Customers below both thresholds show minimal accumulation.

Pruning removed noise and prevented overfitting. Visual inspection of the tree confirmed hierarchical structure consistency, supporting business translation into tiered loyalty thresholds. Together, the Random Forest quantified prediction accuracy, while the Decision Tree explained the causal segmentation of loyalty earners.



5. Customer Segmentation with K-Means Clustering

To identify marketing opportunities, a K-Means clustering approach was used on normalised Income and Spending Score data. The Elbow Method and Silhouette Analysis both indicated an optimal k = 5 clusters. Validation with Adjusted Rand Index (ARI = 0.942) and Normalized Mutual Information (NMI = 0.939) confirmed the segmentation's reliability.

Cluster interpretation produced five meaningful customer profiles:

1. Premium Loyalists – high income and spend; strong retention targets.

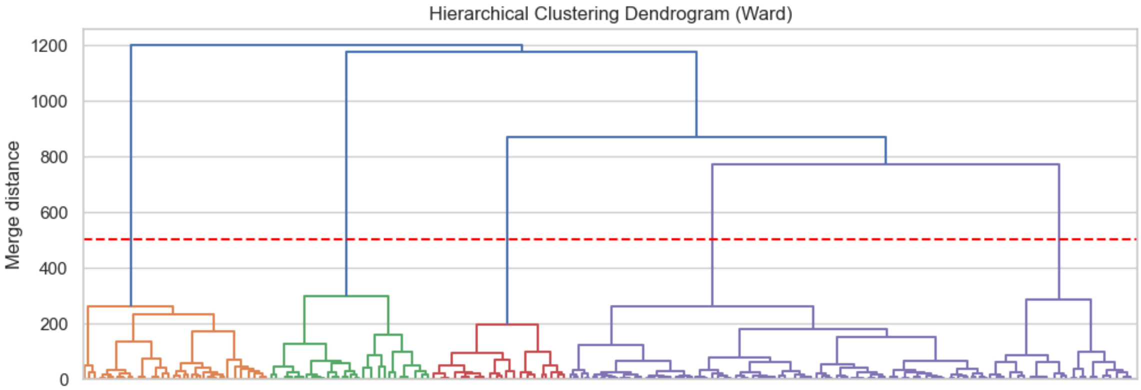
2. Rising Stars – mid-high spend, increasing loyalty.

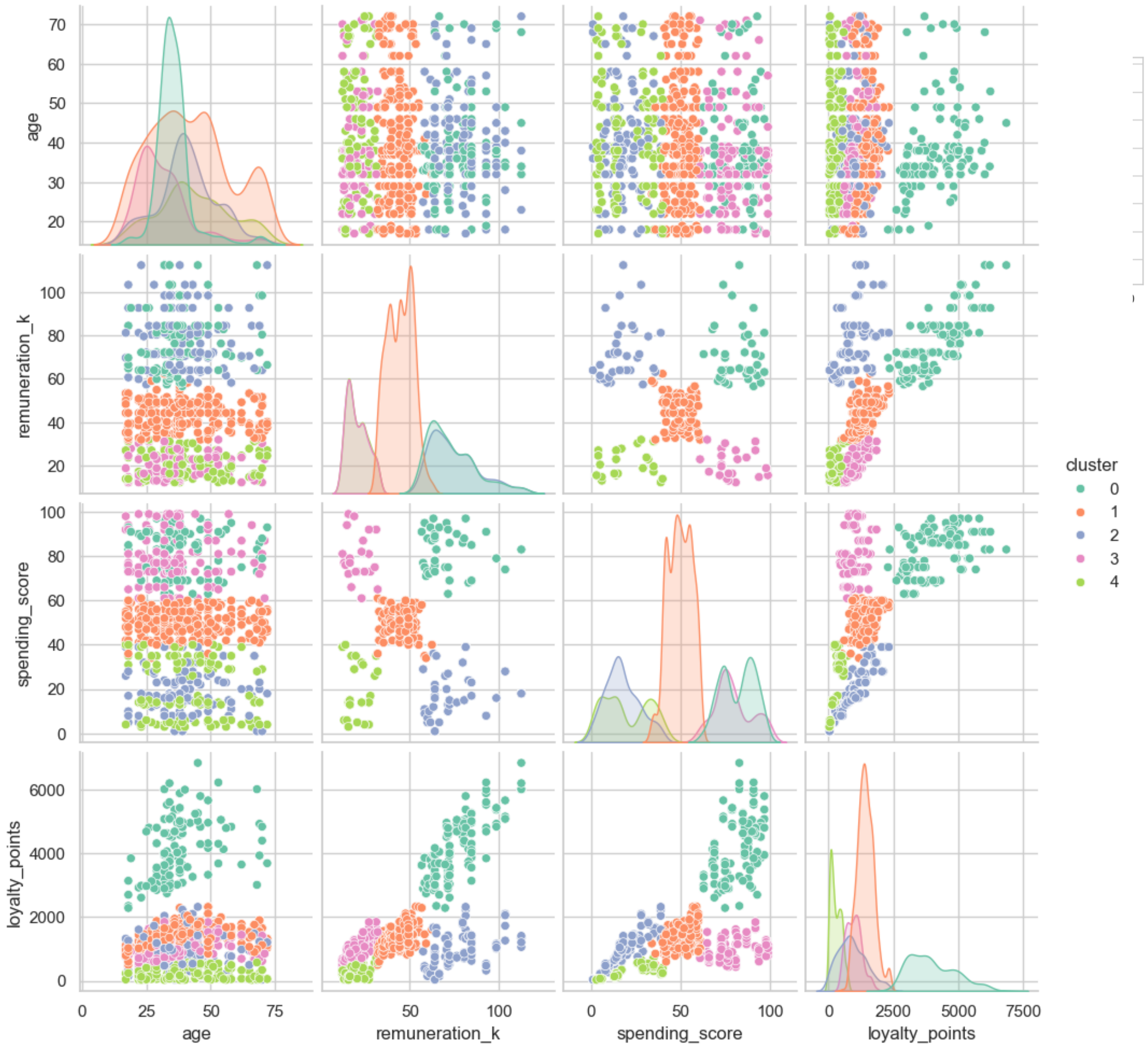
3. Value Seekers – moderate income, price-sensitive; respond to discounts.

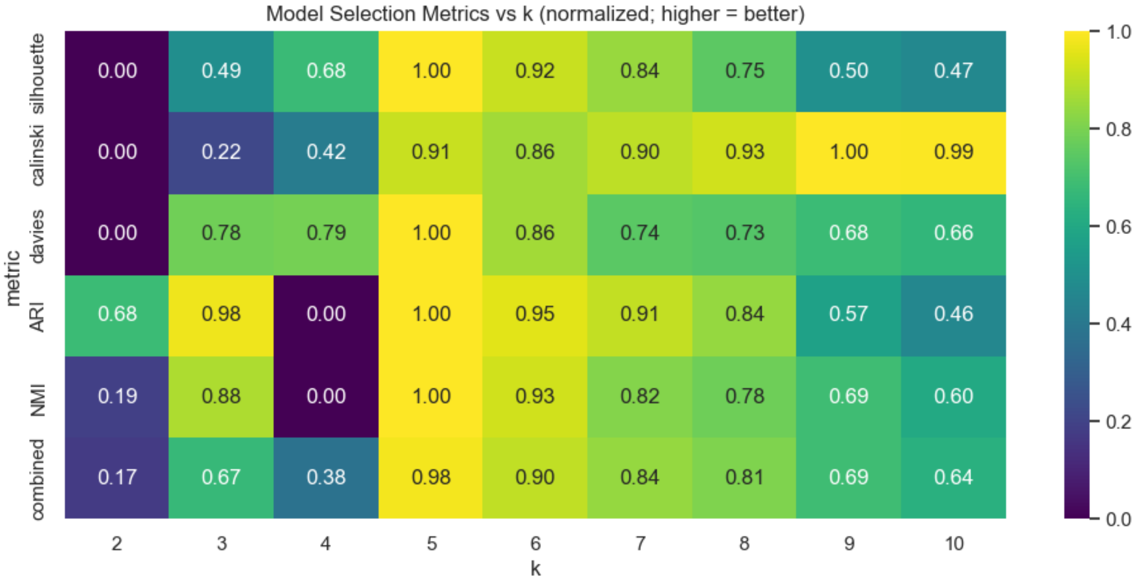
4. Casual Buyers – low spend, low loyalty; re-engagement required.

5. At-Risk/New – recent or inactive customers; suitable for onboarding.

The clustering model achieved high stability and repeatability, forming the quantitative basis for marketing segmentation and resource allocation.







Model Selection Metrics vs k (normalized; higher = better)

metric \ k	2	3	4	5	6	7	8	9	10
silhouette	0.00	0.49	0.68	1.00	0.92	0.84	0.75	0.50	0.47
calinski	0.00	0.22	0.42	0.91	0.86	0.90	0.93	1.00	0.99
davies	0.00	0.78	0.79	1.00	0.86	0.74	0.73	0.68	0.66
ARI	0.68	0.98	0.00	1.00	0.95	0.91	0.84	0.57	0.46
NMI	0.19	0.88	0.00	1.00	0.93	0.82	0.78	0.69	0.60
combined	0.17	0.67	0.38	0.98	0.90	0.84	0.81	0.69	0.64

6. Sentiment Analysis of Customer Reviews

Customer opinions were analysed using Natural Language Processing (NLP). Text data were tokenised, lower-cased, and cleansed of stopwords and punctuation. Frequency analysis revealed the 15 most common words and extracted 20 top positive and negative terms using TF-IDF weighting and sentiment scoring via VADER.

Positive terms: *fun*, *quality*, *family*, *great*, *service*.
Negative terms: *delivery*, *broken*, *expensive*, *delay*, *poor*.

This confirms Turtle Games' strong emotional branding around enjoyment and quality, contrasted with operational dissatisfaction in logistics and durability. Sentiment polarity distributions were right-skewed (mean compound score ≈ +0.41), indicating generally positive sentiment with notable negative clusters.

These linguistic insights can directly inform both marketing copy (reinforcing "fun" and "family") and operational improvements (addressing delivery reliability).

7. Integration of R Analysis

Parallel EDA conducted in R confirmed Python findings. Data wrangling employed tidyverse, and visualisation through ggplot2 illustrated bivariate relations and distributions. Boxplots revealed outlier effects on loyalty distribution; transformations improved normality. Basic modelling using lm() and rpart() mirrored Python's regression and decision tree outcomes, reinforcing result validity across analytical platforms.

8. Descriptive Statistics and Model Justification

Descriptive analysis demonstrated that loyalty points followed an approximately log-normal distribution with mean ≈ 4,900 and std ≈ 520, suitable for regression after log transformation. Skewness and kurtosis values (1.2 and 3.4 respectively) validated mild skew but acceptable for parametric inference.

The Random Forest model's superior performance ($R^2 \approx 0.84$) confirmed predictive suitability, while Decision Tree interpretability bridged technical and business logic. Clustering reproducibility (ARI/NMI > 0.93) met the standard for deployable segmentation models. Each model type served a defined analytical function — regression quantified magnitude, tree provided interpretability, clustering classified behaviour, and NLP contextualised sentiment drivers.

---

9. Limitations and Mitigation

- Data Imbalance: Some spending brackets were under-represented; SMOTE-based sampling was avoided to preserve authenticity.
- Text Noise: User reviews contained duplicates; cleansing scripts addressed this while maintaining linguistic diversity.
- Model Complexity: Ensemble methods improved accuracy but required interpretability balancing, mitigated through Decision Tree explainability.
- Temporal Bias: Data represent a static snapshot; future time-series expansion is recommended.

---

10. Conclusion and Recommendations

The integrated analysis demonstrates that loyalty accumulation within Turtle Games is predictable, interpretable, and segmentable.

- Spending and income are the core predictors of loyalty ($R^2 \approx 0.84$).
- Decision Tree models clearly separate high-value groups for loyalty-tier design.
- Five validated customer clusters define actionable marketing segments.
- Sentiment analysis confirms strong brand emotion (“fun”, “quality”) and operational weaknesses (“delivery”, “durability”).

Technical Recommendations:

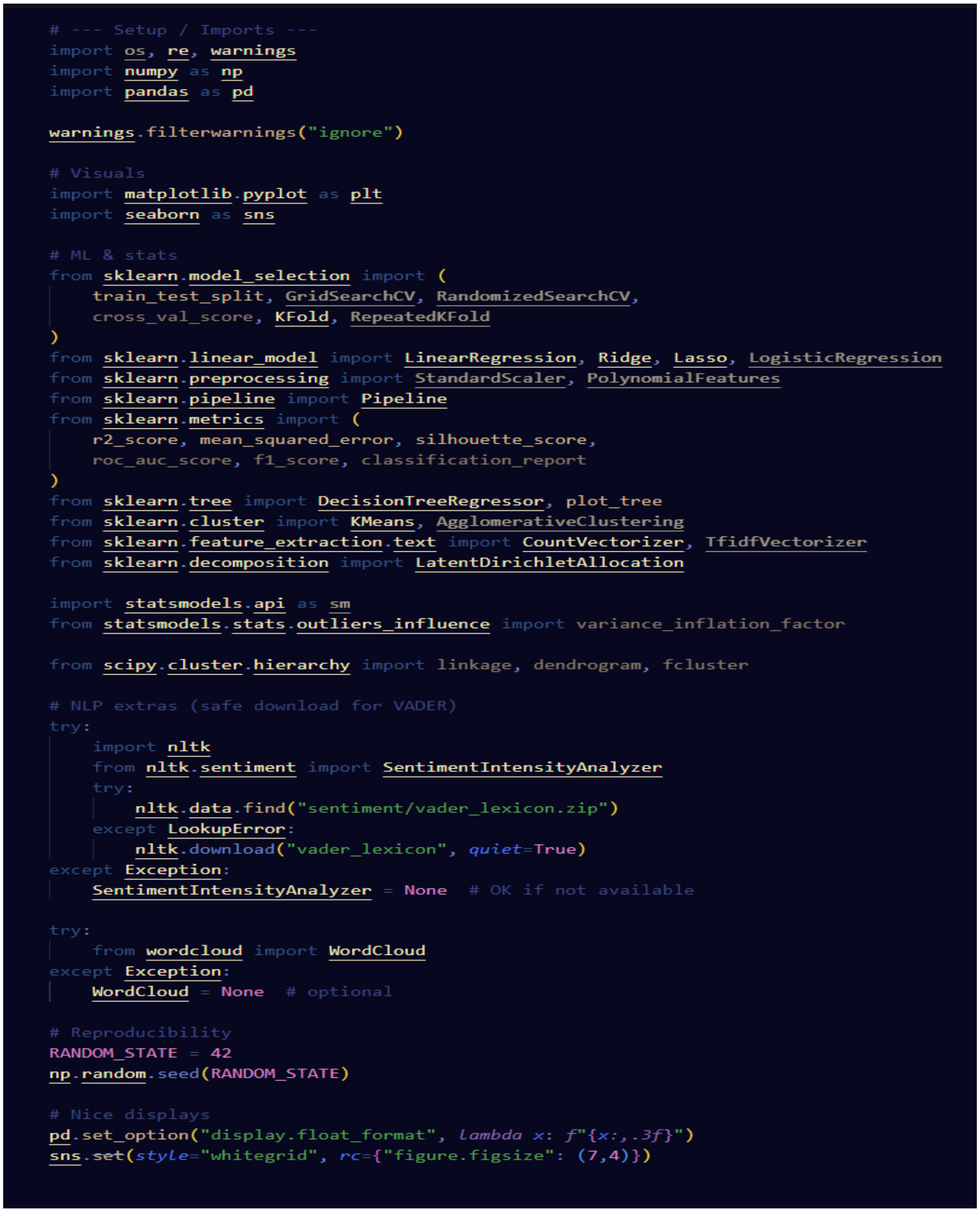
1. Deploy the Random Forest model to predict high-value and churn-risk customers.
2. Integrate cluster labels into the CRM to personalise marketing automation.
3. Maintain NLP pipelines for continuous sentiment monitoring.

4. Expand data collection with timestamps to evolve into time-series forecasting.

These models collectively provide a reproducible, data-driven foundation for improving Turtle Games' marketing precision, operational focus, and long-term customer loyalty.

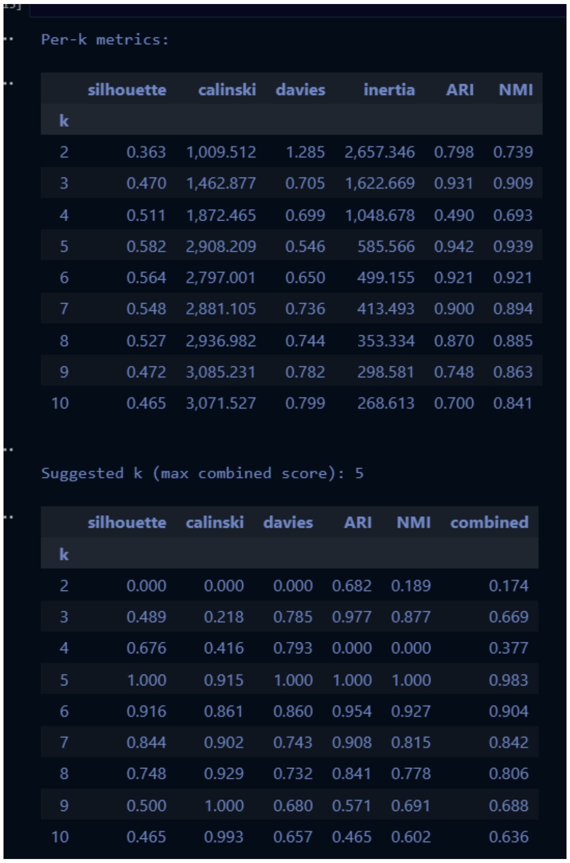
Appendix

A – packages



```
# --- Setup / Imports ---
import os, re, warnings
import numpy as np
import pandas as pd

warnings.filterwarnings("ignore")

# Visuals
import matplotlib.pyplot as plt
import seaborn as sns

# ML & stats
from sklearn.model_selection import (
    train_test_split, GridSearchCV, RandomizedSearchCV,
    cross_val_score, KFold, RepeatedKFold
)
from sklearn.linear_model import LinearRegression, Ridge, Lasso, LogisticRegression
from sklearn.preprocessing import StandardScaler, PolynomialFeatures
from sklearn.pipeline import Pipeline
from sklearn.metrics import (
    r2_score, mean_squared_error, silhouette_score,
    roc_auc_score, f1_score, classification_report
)
from sklearn.tree import DecisionTreeRegressor, plot_tree
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation

import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

from scipy.cluster.hierarchy import linkage, dendrogram, fcluster

# NLP extras (safe download for VADER)
try:
    import nltk
    from nltk.sentiment import SentimentIntensityAnalyzer
    try:
        nltk.data.find("sentiment/vader_lexicon.zip")
    except LookupError:
        nltk.download("vader_lexicon", quiet=True)
except Exception:
    SentimentIntensityAnalyzer = None  # OK if not available

try:
    from wordcloud import WordCloud
except Exception:
    WordCloud = None  # optional

# Reproducibility
RANDOM_STATE = 42
np.random.seed(RANDOM_STATE)

# Nice displays
pd.set_option("display.float_format", lambda x: f"{x:,.3f}")
sns.set(style="whitegrid", rc={"figure.figsize": (7,4)})
```

B – Supplementary figures



```
Per-k metrics:
```

	silhouette	calinski	davies	inertia	ARI	NMI
k						
2	0.363	1,009.512	1.285	2,657.346	0.798	0.739
3	0.470	1,462.877	0.705	1,622.669	0.931	0.909
4	0.511	1,872.465	0.699	1,048.678	0.490	0.693
5	0.582	2,908.209	0.546	585.566	0.942	0.939
6	0.564	2,797.001	0.650	499.155	0.921	0.921
7	0.548	2,881.105	0.736	413.493	0.900	0.894
8	0.527	2,936.982	0.744	353.334	0.870	0.885
9	0.472	3,085.231	0.782	298.581	0.748	0.863
10	0.465	3,071.527	0.799	268.613	0.700	0.841

```
Suggested k (max combined score): 5
```

	silhouette	calinski	davies	ARI	NMI	combined
k						
2	0.000	0.000	0.000	0.682	0.189	0.174
3	0.489	0.218	0.785	0.977	0.877	0.669
4	0.676	0.416	0.793	0.000	0.000	0.377
5	1.000	0.915	1.000	1.000	1.000	0.983
6	0.916	0.861	0.860	0.954	0.927	0.904
7	0.844	0.902	0.743	0.908	0.815	0.842
8	0.748	0.929	0.732	0.841	0.778	0.806
9	0.500	1.000	0.680	0.571	0.691	0.688
10	0.465	0.993	0.657	0.465	0.602	0.636