Course 3 Assignment

Predicting future outcomes

Damian Ferguson

20th April 2024

Background

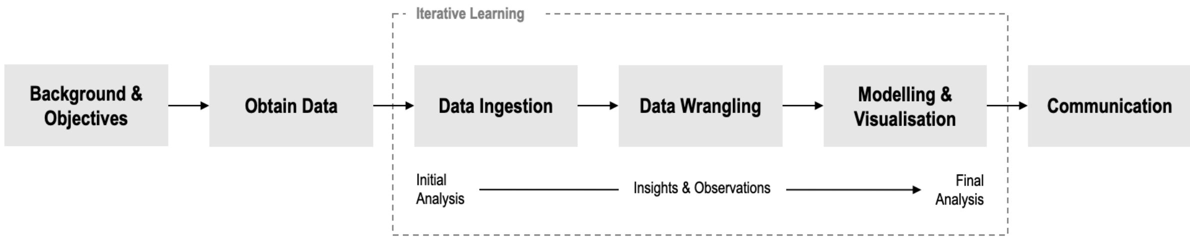
Turtle Games is a global manufacturer, retailer and reseller of its own and other companies' games and toys. The company collects sales and customer review data. It wants to use this data to support its' objective of growing sales.

Turtle Games has developed a set of questions rand objectives relating to:

- Customer engagement with loyalty points
- Creation of prediction models to provide insight into customer loyalty points
- Customer segmentation for targeted marketing campaigns
- Use of text-based reviews to inform marketing decisions

Turtle Games key metric is loyalty points.

Analytical approach



Data Ingestion

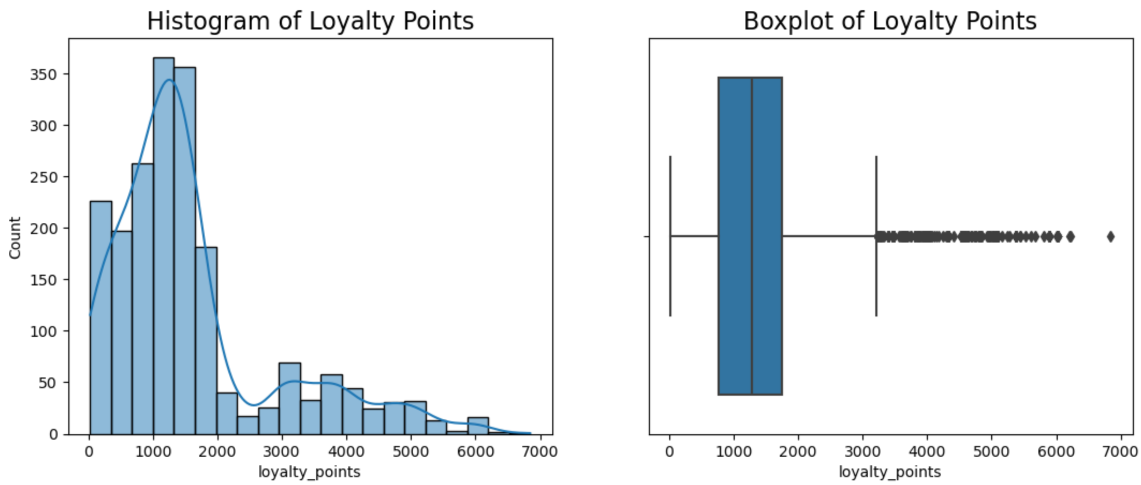
The data was imported from CSV files to a dataframes. Field names, data types and null values were checked; the dataframe head was inspected; descriptive statistics viewed; and value counts executed on categorical variables. The data was explored in tabular and graphical formats.

The data represents 2,000 product reviews. 200 different products are referenced, averaging 10 reviews per product (range = 8 to 13). There is no unique identifier of customer. Reviews may have been submitted by the same customer for different products. This may impact statistical analysis and normality of data.

Observations:

- 2000 rows, 11 columns (1 float, 4 integers, 6 objects)
- no null values
- '*spending_score (1-100)*' and '*remuneration (k£)*' headings are non-standard
- language and platform hold single values
- first character of values in education column inconsistent case
- numeric product data represents categorical information; not a suitable continuous variable
- summary and review columns are text
- evidence of relationships between spending score, remuneration and age subgroups with loyalty points
- evidence of clustering between remuneration and spending score
- outliers only found in loyalty points which is right-skewed and appears non-normal

Loyalty Points Normality




```
Loyalty Points Tests for Normality:
Skew = 1.4637
Kurtosis = 1.7088
Shapiro Wilk statistic = 0.8431
Shapiro Wilk p-value = 0.0
```

Loyalty points confirmed as not normally distributed:

- Shapiro Wilk test, p-value < 0.05
- right-skewed, skew = 1.46
- platykurtic, kurtosis = 1.70

No contextual evidence to justify removal of outliers. Five data transformations tested with outliers (Appendix A), none of these producing normal distributions. Transformations were repeated with outliers removed (Appendix B), none producing normal distributions. Decision to **proceed '*at risk*' using the original data**:

- Non-normality may be underlying population characteristic
- Ensure model residuals are normal to mitigate
- Original data keeps modelling simple, allowing easy re-run of analysis with larger data sets in the future

Data Wrangling (Initial)

The following actions taken:

- language and platform columns deleted
- '*spending_score (1-100)*' and '*remuneration (k£)*' renamed
- values in education column updated to ensure first character is upper case
- new column age group created (Appendix C)

Further data wrangling was conducted during modelling and visualisation.

Modelling & Visualisation

Modelling Methods & Evaluation

Simple Linear Regression (SLR); Multiple Linear Regression (MLR); and Decision Tree Regressor (DTR) were used for **predictive modelling**. Each were trained / tested using 80%/20% split of the data. Models were evaluated using goodness of fit and normality of residuals.

For SLR and MLR models, p-values were checked for significance (< 0.05) and R^2 / R^2_{Adj} used to evaluate goodness of fit. MLR models were checked for multicollinearity and homoscedasticity.

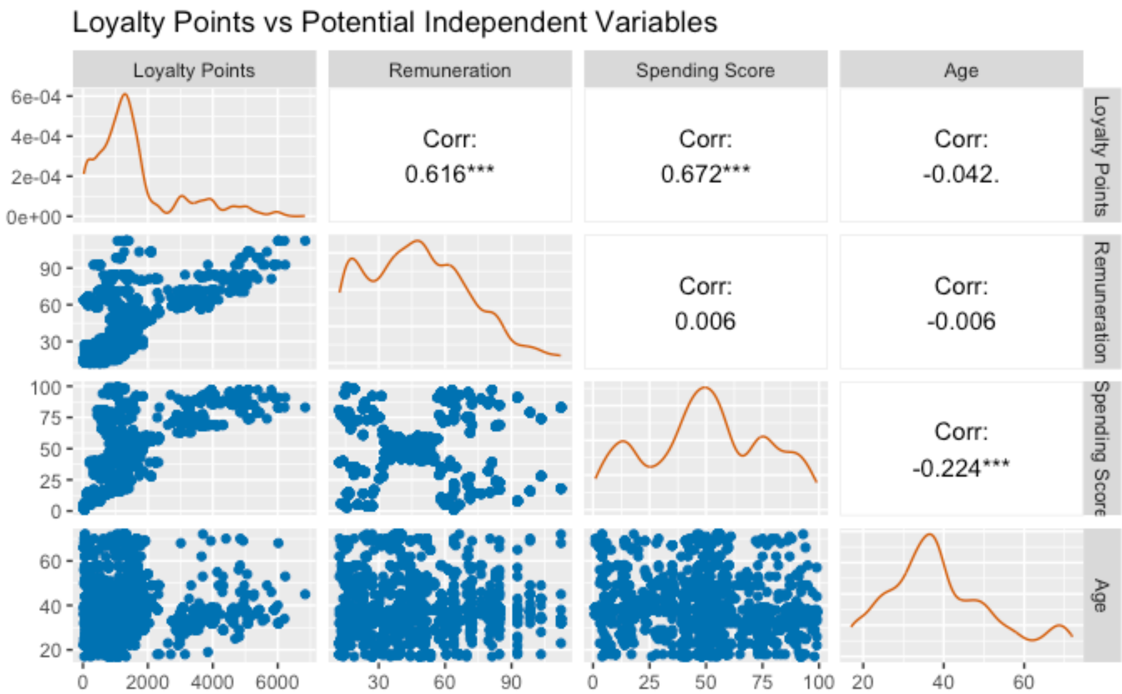
MSE and MAE were used to evaluate DTR goodness of fit. Feature importance, tree depth and samples per leaf were used to post-prune models.. Gender, education and age group were converted to numerical formats for DTR analysis.

K-Means was used for **cluster analysis**. Elbow and Silhouette methods determined optimal number of clusters. Most balanced cluster sizes and logical cluster distribution were used to select the best model.

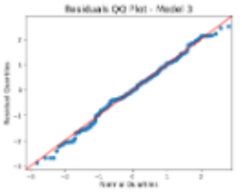
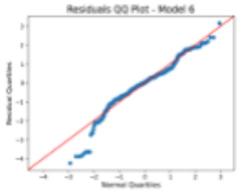
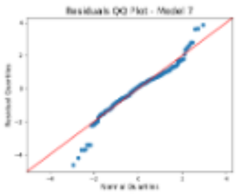
WordClouds, polarity and subjectivity scores were used to **analyse sentiment**. Data was converted to lower case; punctuation removed; tokenised and stop words removed. Duplicate values were retained to provide a true depiction of sentiment.

Predictive Modelling (Overall)

Remuneration and spending score had significant correlations with loyalty points.



Seven models were evaluated: two SLR models, one MLR and four DTR (Appendix D and E). The best models are shown below.

		Model 3	Model 6	Model 7
Type		Multiple Linear Regression	Decision Tree Regressor	Decision Tree Regressor
Dependent Variable (Y)		loyalty_points	loyalty_points	loyalty_points
Independent Variables (X)		remuneration spending_score	remuneration spending_score	remuneration spending_score
Tree Depth		*Not Applicable*	2	3
Number of Leaves		*Not Applicable*	4	6
Min Samples per Leaf		*Not Applicable*	150	125
Goodness of Fit Measures	R^2	0.830	*Not Applicable*	*Not Applicable*
	MSE	300,944	272,344	153,560
	MAE	430	377	294
Residuals Plot				
Evaluation		**Recommended Model** Best combination of goodness of fit and normaility of residuals	**Potential Model** Slightly better goodness of fit vs Model 3. Residuals concern at extremes.	**Potential Model** Much better goodness of fit vs Model 3. Residuals concern at extremes.

Model 3 was selected as the best model. MSE and MAE were used to assess goodness of fit between MLR and DTR models. Normality of the residuals and model simplicity were also assessed. Multicollinearity and homoscedasticity were not present (Appendix F).

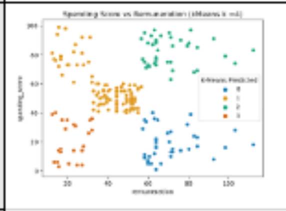
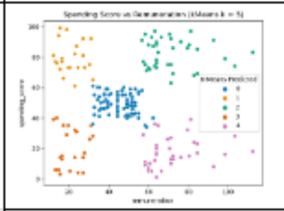
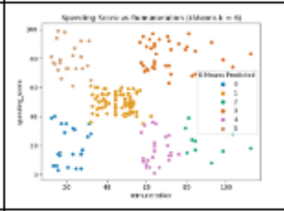
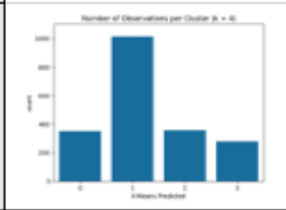
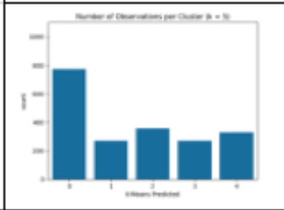
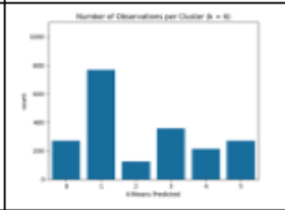
Regression equation is shown below with a table demonstrating the impact of increasing each independent variable by a value of 10 independently and combined.

Loyalty Points = -1,700.32 + 34.33 Remuneration + 32.64 Spending Score

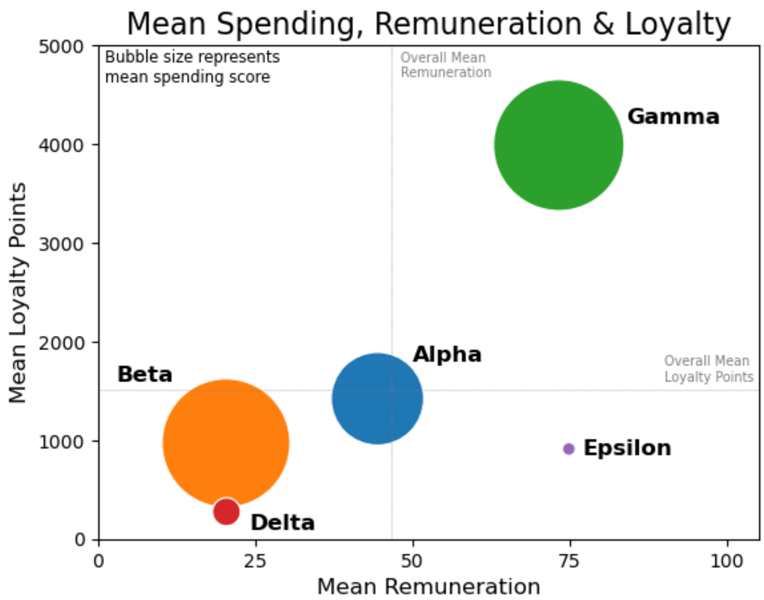
Remuneration		Spending Score		>>>	Loyalty Points	Impact
Baseline	48	Baseline	50	>>>	1,580	*NA*
Baseline + 10	58	Baseline	50	>>>	1,923	343
Baseline	48	Baseline + 10	60	>>>	1,906	326
Baseline + 10	58	Baseline + 10	60	>>>	2,250	670

Cluster Analysis

Scatterplot of remuneration vs spending score indicated clustering. Three models were evaluated with 4, 5 and 6 clusters (Appendix G). The models are shown in the table below.

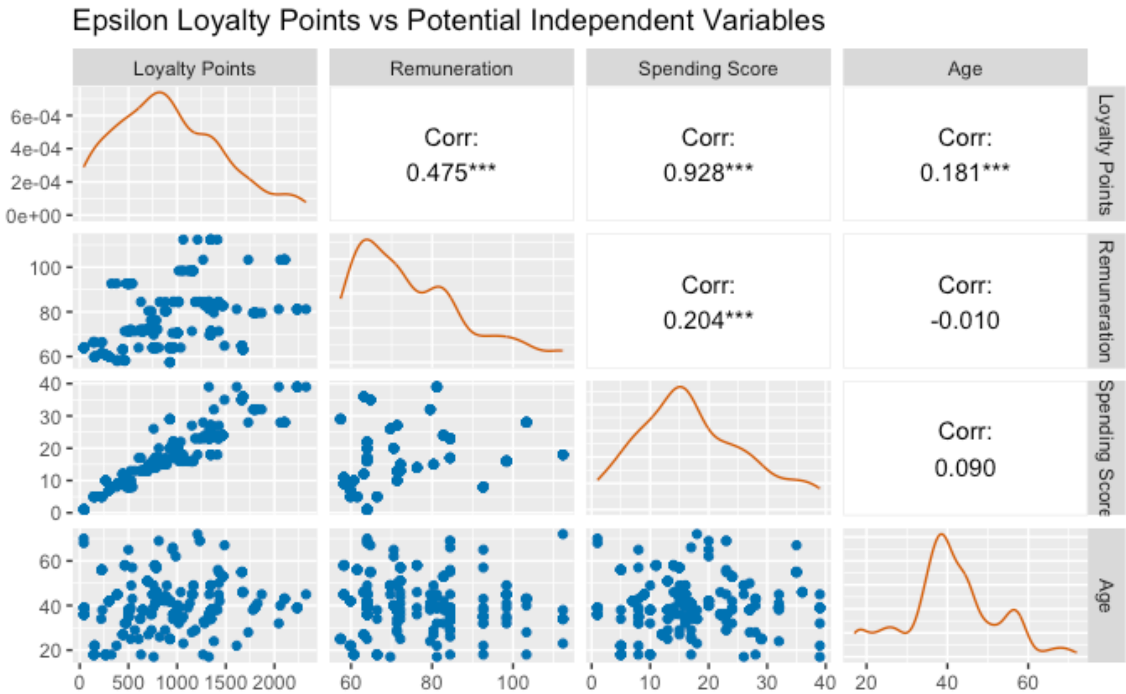
	Model 1	Model 2	Model 3
Number of Clusters	k = 4	k = 5	k = 6
Variables	remuneration spending_score	remuneration spending_score	remuneration spending_score
Cluster Plot			
Observations by Cluster			
Evaluation	**Discard** One large cluster ~1000 observarions; remaining 3 clusters reasonably balanced number of observations; clusters look illogical with size and spread of Cluster 1	**Recommended Model** One large cluster ~750 observarions; remaining 4 clusters reasonably balanced number of observations; clusters look most logical	**Discard** One large cluster ~775 observarions; remaining 5 clusters unbalanced number of observations; clusters look illogical notably Cluster 2

Model 2 was selected for further analysis. Clusters were given meaningful names. The relationship between clusters, loyalty points and spending score plotted.

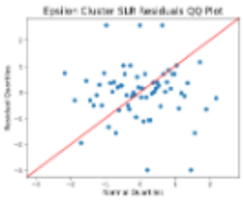
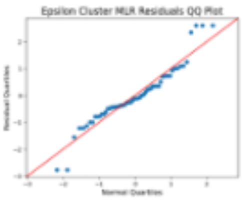


The Epsilon cluster has high remuneration like Gamma, but the lowest spending score and lowest loyalty points of the clusters. Turtle Games cannot affect remuneration, but it can affect customer spend through marketing.

The Epsilon cluster was selected for predictive modelling due to high correlation between loyalty points and spending score.



Two models were evaluated.

		Model 1	Model 2
Type		Simple Linear Regression	Multiple Linear Regression
Dependent Variable (Y)		loyalty_points	loyalty_points
Independent Variables (X)		spending_score	spending_score remuneration
Goodness of Fit Measures	R^2	0.868	0.946
	MSE	50,359	17,472
	MAE	161	92
Residuals Plot			
Evaluation		**Discard** Best combination of goodness of fit and normaility of residuals	**Recommended Model** Best combination of goodness of fit and normaility of residuals

Model 2 was selected as the best model (Appendix H).

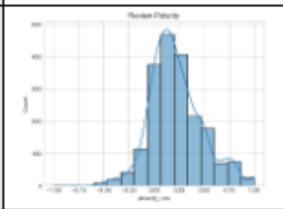
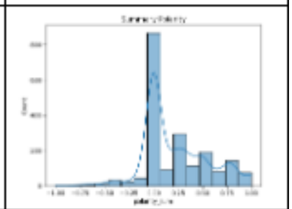
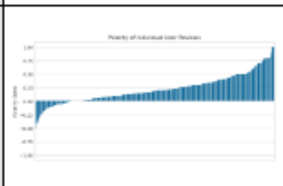
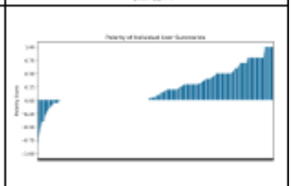
Regression equation is shown below with a table demonstrating the impact of increasing spending score in increments of 10 while holding remuneration constant.

Loyalty Points = -808.02 + 11.51 Remuneration + 49.53 Spending Score

Remuneration		Spending Score		>>>	Loyalty Points	Impact
Hold constant at	75	Increase in	10	>>>	549	NA
Epsilon mean	75	increments of 10	20	>>>	1,044	495
	75		30	>>>	1,539	495
	75		40	>>>	2,035	495
	75		50	>>>	2,530	495
	75		60	>>>	3,025	495
	75		70	>>>	3,520	495
	75		80	>>>	4,016	495
	75		90	>>>	4,511	495

Sentiment Analysis

Sentiment analysis was conducted on review and summary data using NLP methods. The results are displayed in the table below.

	Review	Summary
WordCloud		
Histogram of Polarity Scores		
Individual Polarity Scores (lowest to highest)		
Sentiment	Positive 21.7%	Positive 21.9%
Positive	81%	53%
Neutral	4%	38%
Negative	15%	10%
Evaluation	Sentiment is positive but could be better. Majority of reviews are positive, very few are neutral, small amount negative	Sentiment is positive but could be better. Majority of reviews are positive, large amount are neutral, smaller amount negative

Review and summary demonstrate 22% positive sentiment. Review data is subjective at 52%, summary data is moderately subjectivity at 38%. This should be interpreted as an overall perspective. Further analysis is required with more product datal. No relationships were identified between sentiment scores and other variables.

Recommendations

Use the overall (general) predictive model for setting marketing targets and prioritising budget allocations.

Conduct A/B testing of marketing or discounting programs on the Epsilon cluster to increase spending score and loyalty points acquisition.

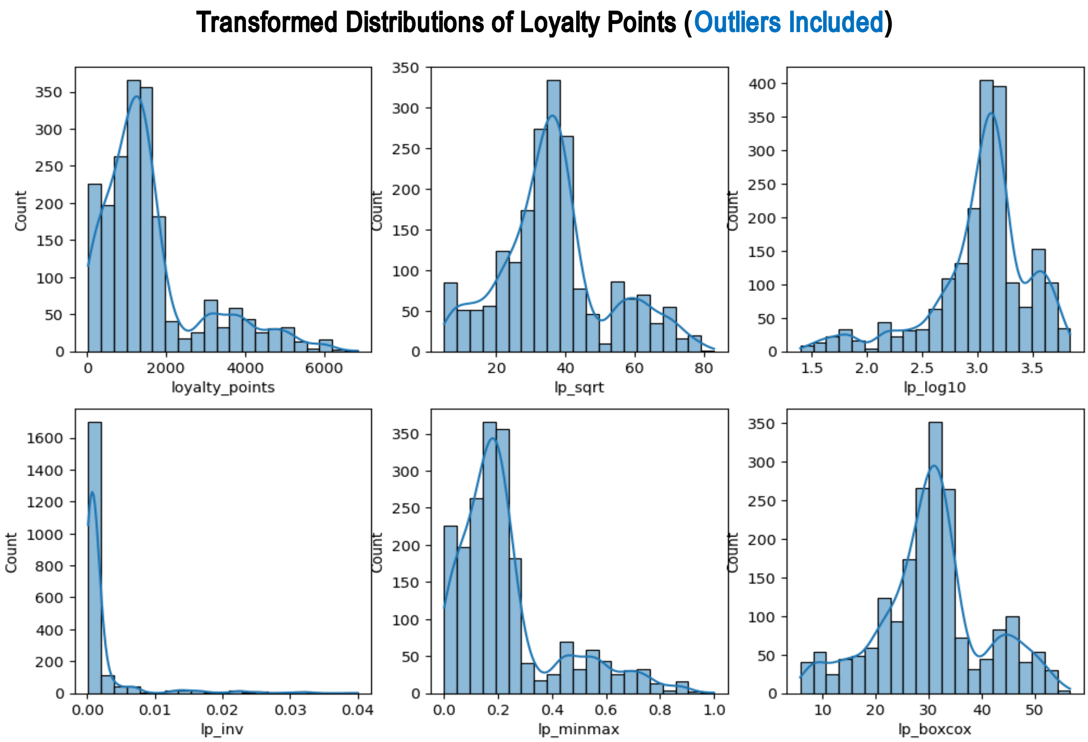
Replicate the Epsilon predictive modelling for other clusters. Identify other cause and effect relationships to increase spending score and acquisition of loyalty points.

Gain access to a larger data set to improve predictive modelling and increase the amount of product related data enabling product level insights and actions to be developed.

Benchmark sentiment analysis against competitors. Is current sentiment being good enough? Does it need to be improved in a targeted or holistic way. Valuable insight could be gained on products, markets and competitors.

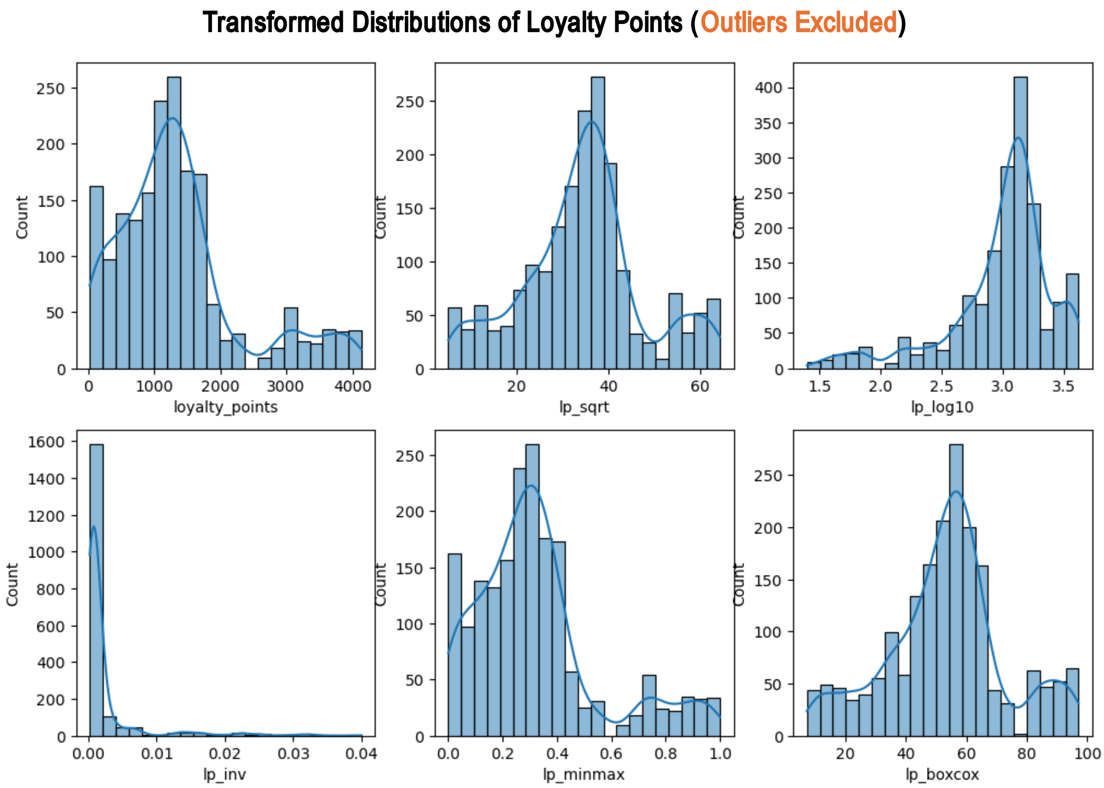
APPENDIX

A. Loyalty points transformations analysis including outliers.



Baseline		Best Candidate Transformations			
Loyalty Points Untransformed		SQRT Transformation		Box Cox Transformation	
Shapiro Wilk p-value	1.24E-40	Shapiro Wilk p-value	8.62E-23	Shapiro Wilk p-value	2.35E-19
Skew	1.4637	Skew	0.4543	Skew	0.0026
Kurtosis	1.7088	Kurtosis	0.1439	Kurtosis	0.1230

B. Loyalty points transformations analysis excluding outliers.



Baseline No Outliers		Best Candidate Transformations No Outliers			
Loyalty Points Untransformed		SQRT Transformation		Box Cox Transformation	
Shapiro Wilk p-value	1.49E-34	Shapiro Wilk p-value	1.22E-19	Shapiro Wilk p-value	1.69E-19
Skew	1.1530	Skew	0.0857	Skew	-0.0240
Kurtosis	1.0429	Kurtosis	0.0630	Kurtosis	0.0775

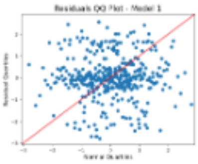
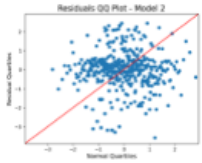
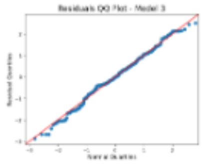
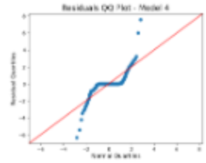
C. Creation of new column age group

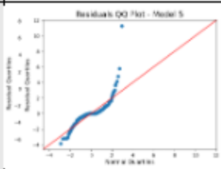
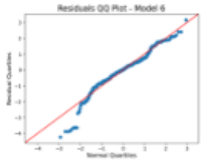
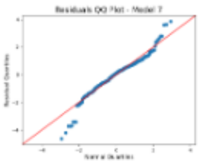
```
# Create a list of conditions.
conditions = [
    (reviews['age'] < 30),
    ( (reviews['age'] >= 30) & (reviews['age'] < 40) ),
    ( (reviews['age'] >= 40) & (reviews['age'] < 50) ),
    ( (reviews['age'] >= 50) & (reviews['age'] < 60) ),
    ( (reviews['age'] >= 60) & (reviews['age'] < 70) ),
    (reviews['age'] >= 70)
    ]

# Create a list of the values to assign for each condition.
values = ['30 & Below', '30-39', '40-49', '50-59', '60-69', '70 & Over']

# Create a new column and use np.select to assign values.
reviews['age_group'] = np.select(conditions, values)

# Check the change worked.
print(reviews['age_group'].value_counts())
reviews[['age', 'age_group']].sample(n = 5)
```

```
age_group
30-39         730
30 & Below    510
40-49         360
50-59         200
60-69         140
70 & Over      60
Name: count, dtype: int64
```

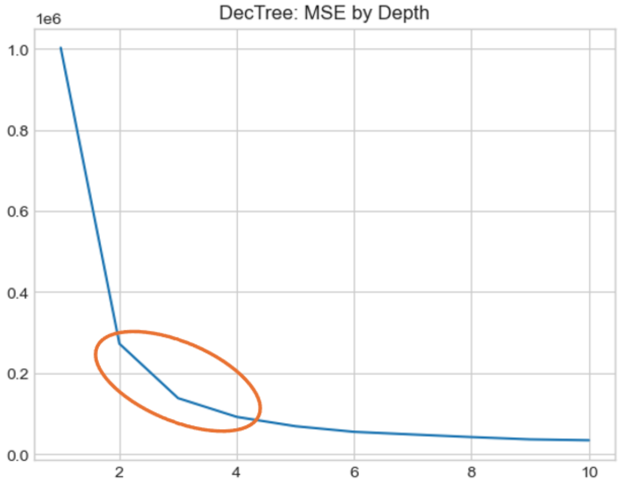
	age	age_group
21	27	30 & Below
1976	57	50-59
1040	67	60-69
1910	67	60-69
317	49	40-49

D. Evaluation of potential predictive models

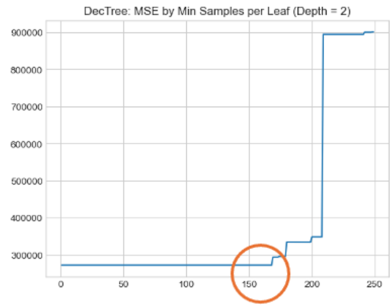
		Model 1	Model 2	Model 3	Model 4
Type		Simple Linear Regression	Simple Linear Regression	Multiple Linear Regression	Decision Tree Regressor
Dependent Variable (Y)		loyalty_points	loyalty_points	loyalty_points	loyalty_points
Independent Variables (X)		remuneration	spending_score	remuneration spending_score	remuneration spending_score education_num age_group_num gender_Male
Tree Depth		*Not Applicable*	*Not Applicable*	*Not Applicable*	19
Number of Leaves		*Not Applicable*	*Not Applicable*	*Not Applicable*	538
Min Samples per Leaf		*Not Applicable*	*Not Applicable*	*Not Applicable*	1
Goodness of Fit Measures	R^2	0.394	0.448	0.830	*Not Applicable*
	MSE	1,106,064	865,342	300,944	8,664
	MAE	748	652	430	36
Residuals Plot					
Evaluation		**Discard** Insufficient variation explained by the independent variabel	**Discard** Insufficient variation explained by the independent variabel	**Recommended Model** Best combination of goodness of fit and normaility of residuals	**Discard** Best goodness of fit. Residuals follow a pattern and not normal. Complex model, concern of overfitting.

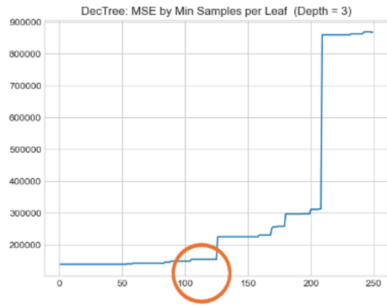
		Model 5	Model 6	Model 7
Type		Decision Tree Regressor	Decision Tree Regressor	Decision Tree Regressor
Dependent Variable (Y)		loyalty_points	loyalty_points	loyalty_points
Independent Variables (X)		remuneration spending_score	remuneration spending_score	remuneration spending_score
Tree Depth		18	2	3
Number of Leaves		196	4	6
Min Samples per Leaf		1	150	125
Goodness of Fit Measures	R^2	*Not Applicable*	*Not Applicable*	*Not Applicable*
	MSE	26,098	272,344	153,560
	MAE	83	377	294
Residuals Plot				
Evaluation		**Discard** Second best goodness of fit. Residuals follow a pattern and not normal. Complex model, concern of overfitting.	**Potential Model** Slightly better goodness of fit vs Model 3. Residuals concern at extremes.	**Potential Model** Much better goodness of fit vs Model 3. Residuals concern at extremes.

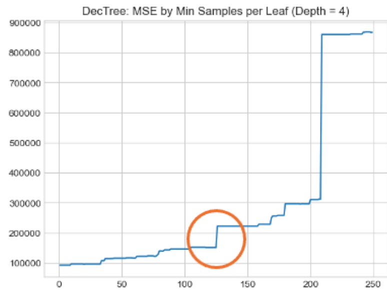
E. Decision tree optimisation



Decision tree depth of 2-4 levels is optimal as there is minimal gain after 4.







At depth = 2, min samples per leaf has no impact on MSE until ~170

At depth = 3, min samples per leaf has no impact on MSE until ~125

At depth = 4, min samples per leaf has a gradual impact on MSE until ~125

F. Evaluation of the recommended overall predictive model (Model 3, MLR)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          loyalty_points   R-squared:                       0.830
Model:                             OLS   Adj. R-squared:                  0.830
Method:                  Least Squares   F-statistic:                     3895.
Date:                 Sat, 20 Apr 2024   Prob (F-statistic):               0.00
Time:                         10:20:03   Log-Likelihood:                -12307.
No. Observations:                 1600   AIC:                         2.462e+04
Df Residuals:                     1597   BIC:                         2.464e+04
Df Model:                            2
Covariance Type:             nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const          -1700.3237     39.588    -42.950      0.000   -1777.974   -1622.674
remuneration      34.3346      0.574     59.838      0.000      33.209      35.460
spending_score    32.6439      0.510     63.947      0.000      31.643      33.645
==================================================================================
Omnibus:                        2.977   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.226   Jarque-Bera (JB):                2.923
Skew:                           0.075   Prob(JB):                        0.232
Kurtosis:                       3.147   Cond. No.                         220.
==============================================================================
```

Adj Rsq indicates the model explains 83% of the variation in the dependent variable '*loyalty_points*'

p-values < 0.05 indicate '*remuneration*' and '*spending_score*' are both statistically significant independent variables

```
# Check for multi-colinearity
# Add a constant.
X_temp = sm.add_constant(X_train)

# Create an empty DataFrame.
vif = pd.DataFrame()

# Calculate the 'vif' for each value.
vif['VIF Factor'] = [variance_inflation_factor(X_temp.values,
                                               i) for i in range(X_temp.values.shape[1])]


# Create the feature columns.
vif['features'] = X_temp.columns

# Print the values to two decimal points.
print(vif.round(2))
```

```
   VIF Factor        features
0        8.91           const
1        1.00    remuneration
2        1.00  spending_score
```

VIF < 3 indicate no presence of multicollinearity

```
# Check for heteroscedasticity
# Run the Breusch-Pagan test function on the model residuals and x-variables.
test = sms.het_breuschpagan(model.resid, model.model.exog)

# Print the results of the Breusch-Pagan test.
terms = ['LM stat', 'LM Test p-value', 'F-stat', 'F-test p-value']
print(dict(zip(terms, test)))
```

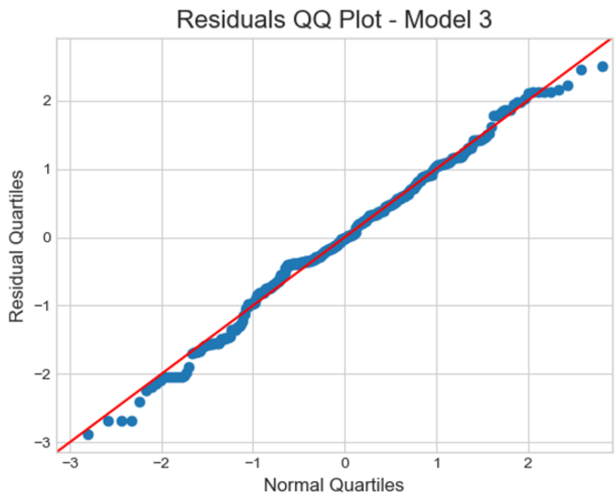
```
{'LM stat': 45.09301971750066, 'LM Test p-value': 1.6150098749727358e-10, 'F-stat': 23.156868353554195, 'F-test p-v
alue': 1.2194606385152372e-10}
```

p-values < 0.05 for LM and F tests indicate that homoscedasticity is not present

```
# Evaluate the model using the test data.
print('MODEL 3 Evaluation (Test Data)')
print('MAE:', round(metrics.mean_absolute_error(y_test, y_pred), 0))
print('MSE:', round(metrics.mean_squared_error(y_test, y_pred), 0))
```

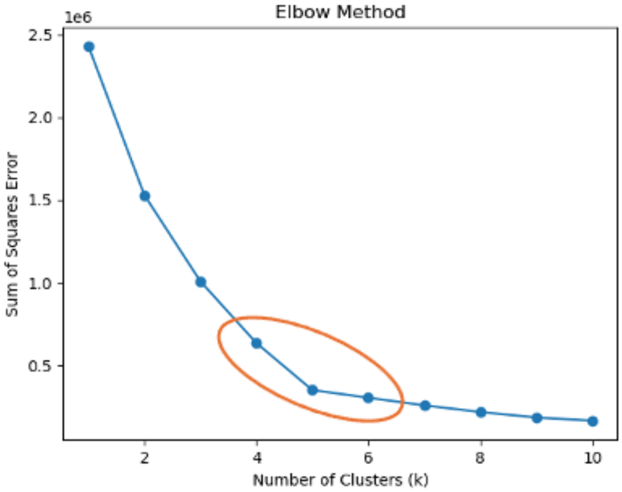
```
MODEL 3 Evaluation (Test Data)
MAE: 430.0
MSE: 300944.0
```

Additional goodness of fit measures to ~ compare with decision tree models

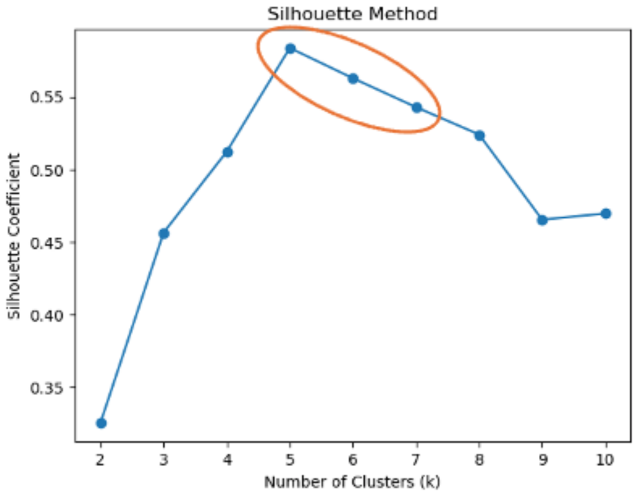


Residuals are closely packed along the red line (with some slight deviations at the extremes) which indicates they are normally distributed

G. Optimal cluster size for K-Means



Elbow indicates 4-6 clusters, most likely 5; optimal as there is minimal gain after 6.



Silhouette method indicates 5-7 clusters, most likely 5

H. Evaluation of the recommended Epsilon cluster predictive model (Model 3, MLR)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         loyalty_points   R-squared:                       0.947
Model:                            OLS   Adj. R-squared:                  0.946
Method:                 Least Squares   F-statistic:                     2321.
Date:                Sun, 21 Apr 2024   Prob (F-statistic):          5.70e-167
Time:                        22:02:17   Log-Likelihood:                -1651.5
No. Observations:                 264   AIC:                             3309.
Df Residuals:                     261   BIC:                             3320.
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           -808.0208     44.273    -18.251      0.000    -895.199    -720.843
remuneration      11.5131      0.587     19.625      0.000      10.358      12.668
spending_score    49.5273      0.824     60.075      0.000      47.904      51.151
==============================================================================
Omnibus:                      112.132   Durbin-Watson:                   1.940
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              679.151
Skew:                          -1.592   Prob(JB):                    3.34e-148
Kurtosis:                      10.183   Cond. No.                         445.
==============================================================================
```

Adj Rsq indicates the model explains 95% of the variation in the dependent variable '*loyalty_points*'

p-values < 0.05 indicate '*remuneration*' and '*spending_score*' are both statistically significant independent variables

```
# Check for multi-colinearity
# Add a constant.
X_temp = sm.add_constant(X_train)

# Create an empty DataFrame.
vif = pd.DataFrame()

# Calculate the 'vif' for each value.
vif['VIF Factor'] = [variance_inflation_factor(X_temp.values,
                                               i) for i in range(X_temp.values.shape[1])]


# Create the feature columns.
vif['features'] = X_temp.columns

# Print the values to two decimal points.
print(vif.round(2))
```

```
   VIF Factor        features
0       32.19           const
1        1.04    remuneration
2        1.04  spending_score
```

VIF < 3 indicate no presence of multicollinearity

```
# Check for heteroscedasticity
# Run the Breusch-Pagan test function on the model residuals and x-variables.
test = sms.het_breuschpagan(model.resid, model.model.exog)

# Print the results of the Breusch-Pagan test.
terms = ['LM stat', 'LM Test p-value', 'F-stat', 'F-test p-value']
print(dict(zip(terms, test)))
```

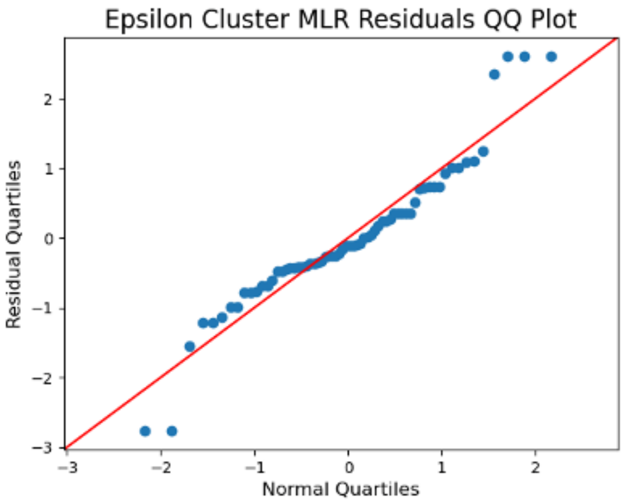
```
{'LM stat': 28.015369937219024, 'LM Test p-value': 8.251629387408348e-07, 'F-stat': 15.492550883324072, 'F-test p-v
alue': 4.385435405398325e-07}
```

p-values < 0.05 for LM and F tests indicate that homoscedasticity is not present

```
# Evaluate the model using the test data.
print('MODEL 3 Evaluation (Test Data)')
print('MAE:', round(metrics.mean_absolute_error(y_test, y_pred), 0))
print('MSE:', round(metrics.mean_squared_error(y_test, y_pred), 0))
```

```
MODEL 3 Evaluation (Test Data)
MAE: 92.0
MSE: 17472.0
```

Additional goodness of fit measures to ~ compare with decision tree models



Residuals are closely packed along the red line (with some slight deviations at the extremes) which indicates they are normally distributed