FORE SCHOOL OF MANAGEMENT

MLM-2

PROJECT 1 REPORT

CUSTOMER SEGMENTATION FOR INTERSTELLAR TOURISM

Submitted To: **Prof. Amarnath Mitra**

Submitted By:

Adhyatik (311063)

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Prof. Amarnath Mitra for his exceptional guidance and mentorship throughout the learning process of machine learning concepts. His profound knowledge, clear explanations, and dedication to teaching have been instrumental in my understanding and application of these concepts.

Prof. Amarnath Mitra's teachings have been invaluable in shaping my ability to analyze and interpret data effectively. His expertise in unsupervised machine learning techniques, particularly clustering algorithms like K-Means, has been crucial in conducting this customer segmentation analysis for interstellar tourism.

I am grateful for the engaging and insightful lectures, where Prof. Mitra provided real-world examples and practical applications of machine learning concepts. His ability to break down complex topics into understandable components has been truly remarkable.

Furthermore, I would like to acknowledge Prof. Mitra's unwavering support and encouragement throughout this project. His invaluable feedback and suggestions have greatly contributed to the successful completion of this analysis and the generation of actionable insights.

I extend my heartfelt appreciation to Prof. Amarnath Mitra for imparting his knowledge and fostering a stimulating learning environment. His teachings have not only equipped me with the necessary skills but have also ignited a passion for exploring the vast potential of machine learning in solving complex business problems.

EXECUTIVE SUMMARY

This report presents the findings of a customer segmentation analysis for the interstellar tourism industry, leveraging unsupervised machine learning techniques. The analysis aimed to identify distinct customer segments, determine their characteristics, and provide actionable insights to enhance marketing strategies, operations, and customer experiences.

The analysis utilized the K-Means clustering algorithm on a comprehensive dataset of approximately 500,000 space travel records. The optimal number of clusters was determined to be three, based on the Silhouette Score metric. Cluster analysis techniques, including the Kruskal-Wallis test for categorical variables and one-way ANOVA for non-categorical variables, were employed to characterize each cluster.

The key findings revealed three distinct customer segments with varying preferences and behaviors:

Cluster 0: Cost-conscious customers preferring economy travel class, short-duration stays, and nearby destinations, attracted by affordable and accessible options.

Cluster 1: Adventure-seeking customers balancing affordability with a willingness to travel longer distances for immersive experiences, valuing unique destinations and efficient transportation modes.

Cluster 2: Premium customers with a propensity for luxury experiences, advanced transportation modes, and extended stays, seeking exclusivity and personalized services.

Significant contributing variables included Travel Class, Star System, Transportation Type, Distance to Destination, and Duration of Stay.

Based on these insights, the report provides tailored recommendations for each cluster, encompassing targeted marketing strategies, transportation and fleet management, service and operational optimization, product development and innovation, and strategic partnerships and collaborations.

By implementing these recommendations, the space travel and tourism industry can effectively cater to the diverse needs and preferences of its customer base, enhance customer satisfaction, optimize resource allocation, and gain a competitive edge in the rapidly evolving interstellar tourism market.

1.) OBJECTIVES

1.1) Customer Segmentation for Targeted Marketing Using Unsupervised Machine Learning Clustering Algorithms.

1.2.) Identify Number Of Optimal Segments.

1.3.) Determining the Characteristics of each Cluster.

2.) DESCRIPTION OF DATA

2.1. Data Source, Size & Shape

2.1.1. Data Source (Website Link):-

https://www.kaggle.com/datasets/anthonytherrien/interstellar-travel-customer-satisfaction-analysis

2.1.2. Data Size:- 74.6 MB

2.1.3. Data Description & Dimensions:- This dataset, titled "Interstellar Travel Customer Satisfaction Analysis," provides a comprehensive view of customer experiences in interstellar space travel. It encompasses approximately **500,000** records, each representing an individual space travel experience. The primary objective of this dataset is to understand and predict the Customer Satisfaction Score, a key indicator of service quality and customer experience in the burgeoning field of interstellar tourism and travel.

Number Of Observations Taken For Analysis:- **1,00,192**
Number Of Variables:- **18**

2.2. Description of Variables

2.2.1. Index Variable(s): Row ID

2.2.2. Categorical Variables or Features (CV):- 10

1. **Gender**: Gender of the traveler.
2. **Occupation**: Occupation of the traveler, such as Colonist, Tourist, Businessperson, etc.

3. **Travel Class**: Class of travel, e.g., Business, Economy, Luxury.
4. **Destination**: Interstellar destination.
5. **Star System**: Star system of the destination.
6. **Purpose of Travel**: The primary purpose of travel, e.g., Tourism, Research, Colonization.
7. **Transportation Type**: Type of transportation, e.g., Warp Drive, Solar Sailing, Ion Thruster.
8. **Special Requests**: Any special requests made by the traveler.
9. **Loyalty Program Member**: Indicates if the traveler is a member of a loyalty program.
10. **Month**: Month of travel.

2.2.2.1. Categorical Variables or Features - Nominal Type:- 10

1. **Gender**
2. **Occupation**
3. **Travel Class**
4. **Destination**
5. **Star System**
6. **Purpose of Travel**
7. **Transportation Type**
8. **Special Requests**
9. **Loyalty Program Member**
10. **Month**

2.2.2.2. Categorical Variables or Features - Ordinal Type: 0

There are no Categorical Variables of Ordinal Type in this Dataset.

2.2.3. Non-Categorical Variables or Features: 8

1. **Age:** Age of the traveler.
2. **Distance to Destination (Light-Years):** The distance to the destination measured in light-years.
3. **Duration of Stay (Earth Days):** Duration of stay at the destination in Earth days.

4. **Number of Companions:** The number of companions accompanying the traveler.
5. **Price (Galactic Credits):** Price of the trip in Galactic Credits.
6. **Booking Date:** Date when the trip was booked.
7. **Departure Date:** Date of departure.
8. **Customer Satisfaction Score:** Indicator of service quality and customer experience.

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count & Relative Frequency Statistics

1. **Gender**

Gender	Count	Relative Frequency
Male	58403	0.5829108112424146
Female	41789	0.41708918875758544

2. **Occupation**

Occupation	Count	Relative Frequency
Tourist	16768	0.16735867135100607
Businessperson	16756	0.1672389013094858
Scientist	16746	0.16713909294155221
Other	16666	0.1663406259980837
Colonist	16655	0.16623083679335676
Explorer	16601	0.1656918716065155

3. **Travel Class**

Travel Class	Count	Relative Frequency
Economy	54967	0.5486166560204407
Business	31828	0.3176700734589588
Luxury	13397	0.13371327052060045

4. **Destination** | **Count** | **Relative Frequency**

Destination	Count	Relative Frequency
Alpha Centauri	9265	0.09247245289045034
Zeta II Reticuli	9187	0.09169394762056851
Proxima Centauri	9167	0.09149433088470137
Gliese 581	9146	0.09128473331204089
Tau Ceti	9146	0.09128473331204089
Trappist-1	9127	0.0910950974129671
Barnard's Star	9101	0.09083559565633982
Kepler-22b	9019	0.09001716703928457
Epsilon Eridani	9019	0.09001716703928457
Lalande 21185	8967	0.08949816352603002
Exotic Destination 9	986	0.00984110507824976
Exotic Destination 5	919	0.009172389013094858
Exotic Destination 4	914	0.009122484829128074
Exotic Destination 3	910	0.009082561481954647
Exotic Destination 2	901	0.008992733950814437
Exotic Destination 8	898	0.008962791440434366
Exotic Destination 1	893	0.008912887256467583
Exotic Destination 7	880	0.008783136378153944
Exotic Destination 6	879	0.008773155541360588
Exotic Destination 10	868	0.008663366336633664

5. **Purpose Of Travel** | **Count** | **Relative Frequency**

Purpose Of Travel	Count	Relative Frequency
Research	20352	0.2031299904183967
Other	20060	0.2002155860747365
Tourism	19935	0.1989679814755669
Colonization	19926	0.1988781539444267
Business	19919	0.1988082880868732

6. **Transportation Type** | **Count** | **Relative Frequency**

Transportation Type	Count	Relative Frequency
Solar Sailing	25154	0.25105796870009583
Warp Drive	25085	0.2503692909613542
Ion Thruster	24977	0.24929136058767168
Other	24976	0.2492813797508783

7. Special Requests

Special Requests	Count	Relative Frequency
Window Seat	20139	0.2010040721814117
Extra Space Suit	20101	0.20062480038326413
Special Meal	20017	0.19978641009262216
None	19991	0.1995269083359949
Other	19944	0.1990578090067071

8. Loyalty Status

Loyalty Status	Count	Relative Frequency
Yes	52117	0.520171271159374
No	48075	0.479828728840626

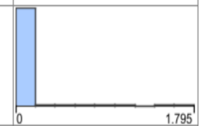
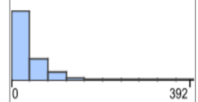
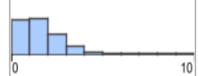
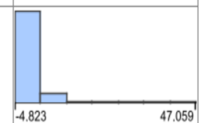
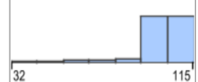
9. Month

Month	Count	Relative Frequency
1	8744	0.08727243692111146
12	8596	0.08579527307569466
8	8548	0.08531619290961354
5	8544	0.08527626956244011
10	8525	0.08508663366336634
7	8496	0.084797189396359
3	8485	0.08468740019163207
4	8281	0.08265130948578729
9	8110	0.08094458639412329
11	8067	0.08051541041200894
6	8044	0.08028585116576174
2	7744	0.07729160012775471

2.3.2. Descriptive Statistics: Non-Categorical Variables or Features

2.3.2.1. Measures of Central Tendency

2.3.2.2. Measures of Dispersion

Row ID	S Column	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewness	D Kurtosis	D Overall ...	I No. mis...	I No. NaNs	I No. +∞s	I No. -∞s	D Median	I Row co...	Histogram
Age	Age	0	98	31.184	19.677	387.188	0.723	-0.127	3,124,363	0	0	0	0	28	100192	
Distance to D...	Distance to ...	0	1,795.31	8.302	23.635	558.637	19.638	841.539	831,550.93	26	0	0	0	2.7	100192	
Duration of St...	Duration of ...	0	391.5	34.067	33.317	1,110.015	1.85	4.557	3,412,536	22	0	0	0	22	100192	
Number of Co...	Number of C...	0	10	1.108	1.095	1.199	1.108	1.577	111,043	1	0	0	0	1	100192	
Price (Galactic...	Price (Galact...	-4,822.526	47,059.009	1,025.638	1,366.348	1,866,908.213	4.843	57.855	102,699,18...	60	0	0	0	630.696	100192	
Customer Sati...	Customer S...	32.25	115	101.627	9.304	86.56	-2.372	7.715	10,178,498.88	37	0	0	0	102	100192	

2.3.2.3. Correlation Statistics

Correlation Matrix

Row ID	D Age	D Distance to Destina...	D Duration of Stay (...	D Number of Comp...	D Price (Galactic Cre...	D Customer Satisfac...
Age	1.0	0.0012018622331481...	-3.268359978797662...	-0.07617321187515...	-0.13291710909193...	-0.19332349866477...
Distance t...	0.00120186...	1.0	0.06848002091921744	0.001791694565926...	0.5703616094405309	-0.05419203578003...
Duration o...	-3.2683599...	0.06848002091921744	1.0	-0.06400337965526...	0.049832633893204...	-0.00128230287324...
Number of...	-0.0761732...	0.0017916945659265...	-0.064003379655267...	1.0	0.008712973293811...	-0.02143670845884...
Price (Gala...	-0.1329171...	0.5703616094405309	0.04983263389320477	0.008712973293811...	1.0	0.090895137117804...
Customer ...	-0.1933234...	-0.05419203578003211	-0.001282302873247...	-0.02143670845884...	0.090895137117804...	1.0

Row ID	S First col...	S Second column name	D Correlation value	D p value	I Degrees...
Row ID	Age	Distance to Destination (Light-Years)	0.0012018622331481806	0.7036319749300362	100190
Row1	Age	Duration of Stay (Earth Days)	-3.268359978797662E-4	0.9176038447543221	100190
Row2	Age	Number of Companions	-0.07617321187515685	8.22863333694245...	100190
Row3	Age	Price (Galactic Credits)	-0.13291710909193286	0.0	100190
Row4	Age	Customer Satisfaction Score	-0.19332349866477952	0.0	100190
Row5	Distance to ...	Duration of Stay (Earth Days)	0.06848002091921744	0.0	100190
Row6	Distance to ...	Number of Companions	0.0017916945659265738	0.5706322419283532	100190
Row7	Distance to ...	Price (Galactic Credits)	0.5703616094405309	0.0	100190
Row8	Distance to ...	Customer Satisfaction Score	-0.05419203578003211	4.78262973117410...	100190
Row9	Duration of ...	Number of Companions	-0.06400337965526706	1.94679859254818...	100190
Row10	Duration of ...	Price (Galactic Credits)	0.04983263389320477	0.0	100190
Row11	Duration of ...	Customer Satisfaction Score	-0.0012823028732476...	0.6848278575279976	100190
Row12	Number of C...	Price (Galactic Credits)	0.008712973293811595	0.00581658376333...	100190
Row13	Number of C...	Customer Satisfaction Score	-0.021436708458846328	1.15219608378159...	100190
Row14	Price (Galact...	Customer Satisfaction Score	0.09089513711780459	0.0	100190

corr = -1
corr = +1
corr = n/a
Age
D...
D...
N...
P...
C...
Age
"Age" (1/6)
Distance to Desti...
Duration of Stay ...
Number of Comp...
Price (Galactic Cr...
Customer Satisfa...

3.) ANALYSIS OF DATA

3.1. Data Pre-Processing

3.1.1. Missing Data Statistics and Treatment

3.1.1.1 Missing Data Statistics:

Field-wise Analysis:

1. **Distance To Destination :** 26 missing values
2. **Duration Of Stay :** 22 missing values
3. **Number Of Companions :** 1 missing values
4. **Price(Galactic Credits) :** 60 missing values
5. **Customer Satisfaction Score:** 37 missing values

Distance To Destination, Duration Of Stay, Number Of Companions, Price, Customer Satisfaction Score All of these are NON-CATEGORICAL Variables.

Record-wise Analysis:

1. Total missing records: 146

3.1.1.2 Missing Data Treatment:

1. Since all the missing values are non-categorical, the treatment will focus solely on non-categorical fields.
2. Treatment of 'price' Variable: As 'price' is a non-categorical variable in numerical format, the Simple Imputer method is applied to all fields containing null values. The imputation strategy is 'mean', which replaces missing values with the mean value in each column. Similar Treatment is done for all the variables having missing values as all of them are Non-Categorical and have numerical values. Following the imputation process, no columns retain null values.

3.1.1.3 Missing Data Exclusion:

1. Deletion of Records: If any Record(Rows) contains more than 50% of null values those Records are removed from the dataset.
2. Variable Deletion: If any Variable(Columns) contains more than 50% of null values persist, the corresponding variables are considered for removal.
3. Following the implementation of deletion procedures, the dataset achieves completeness.

3.1.2. Numerical Encoding of Categorical Variables or Features

Since all categorical variables in the dataset are nominal, we apply label encoding to transform them into numerical representations and the Encoding Schema is Alphanumeric Order.

3.1.3. Outlier Statistics and Treatment

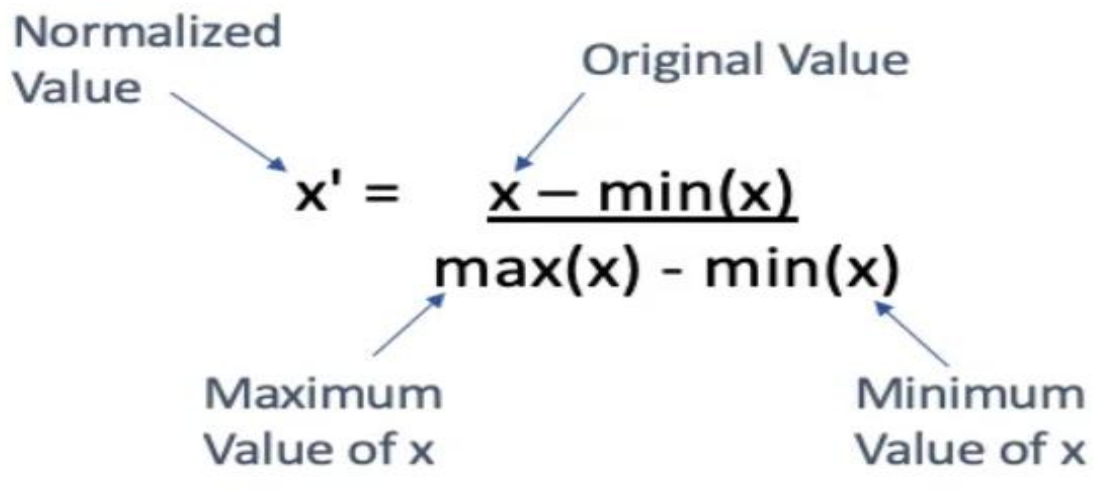
3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features

Row ID	S Outlier c...	I Member ...	I Outlier c...	D Lower bo...	D Upper bo...
Row0	Age	100192	922	-24.5	83.5
Row1	Distance to ...	100192	10948	-8.69	17.07
Row2	Duration of ...	100192	5158	-43	101
Row3	Number of ...	100192	235	-3	5
Row4	Price (Galac...	100192	7295	-1,268.819	2,832.369
Row5	Customer S...	100192	6289	89.5	117.5

3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

1. Identification of Outliers: Box plots are utilized to visually inspect the presence of outliers within the dataset.
2. Outlier Detection: Outliers are identified in the following columns: ‘Distance To Destination’, ‘Duration Of Stay’, ‘Number Of Companions’, ‘Price’, ‘Customer Satisfaction Score’
3. Outlier Treatment: Min-Max Scaling normalization technique is applied to address outliers in the identified columns. Min-Max Normalization transforms x to x’ by converting each value of features to a range between **0 and 1**, and this is also known as **(0–1) Normalization.** If the data has negative values the range would have been between **-1 and 1**.

   The formula for Min-Max Normalization is:



Formula for Min-Max Normalization

4. Min-Max Scaling normalizes the data and removes outliers. We utilize the interquartile range (IQR) from the 25th to the 75th percentile.

3.1.4. Data Bifurcation: Training & Testing Sets

1. The dataset is partitioned into two subsets: training and testing datasets.
2. Training dataset is 80% of complete data.
3. Testing dataset is 20% of complete data.

3.2. Data Analysis

3.2.1 Unsupervised Machine Learning Clustering Algorithm: K-Means

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It is an unsupervised learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

The working of the K-Means algorithm is explained in the below steps:

- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid of each cluster.
- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.

3.2.2 Clustering Model Performance Evaluation: Silhouette Score

The silhouette coefficient or silhouette score k-means is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).

- Select a range of values of k (say 1 to 10).
- Plot Silhouette coefficient for each value of K.

The equation for calculating the silhouette coefficient for a particular data point:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- S(i) is the silhouette coefficient of the data point i.
- a(i) is the average distance between i and all the other data points in the cluster to which i belongs.
- b(i) is the average distance from i to all clusters to which i does not belong.

We will then calculate the average_silhouette for every k.

$$\text{AverageSilhouette} = \text{mean}\{\text{S(i)}\}$$

Then plot the graph between average_silhouette and K.

Points to Remember While Calculating Silhouette Coefficient:

- The value of the silhouette coefficient is between [-1, 1].
- A score of 1 denotes the best, meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1. Values near 0 denote overlapping clusters.

3.2.3 Cluster Analysis: Base Model (K-Means)

3.2.3.1 Cluster Analysis with Categorical Variables or Features: Kruskal-Wallis Test

Categorical Variables: Kruskal-Wallis Test:

- Assumptions: The Kruskal-Wallis test is a non-parametric test and does not require the assumption of normality or homogeneity of variances. However, it assumes that the observations are independent and the dependent variable (categorical variable) is measured on at least an ordinal scale.
- Effect Size: In addition to the p-value, it is crucial to examine the effect size to gauge the practical significance of the differences between clusters, as statistical significance alone does not imply practical relevance.
- Post-hoc Tests: If the Kruskal-Wallis test indicates significant differences across clusters, post-hoc tests can be performed to identify which specific pairs of clusters differ significantly for the given categorical variable.
- Visualization: Visualizations such as bar plots, mosaic plots, or alluvial diagrams can aid in interpreting the distribution and patterns of categorical variables within and across clusters.

3.2.3.2 Cluster Analysis with Non-Categorical Variables or Features: Analysis of Variance (ANOVA)

Non-categorical Variables: One-way ANOVA:

- Assumptions: One-way ANOVA assumes normality of the dependent variable (non-categorical variable) within each cluster and homogeneity of variances across clusters. If these assumptions are violated, robust alternatives such as the Welch's ANOVA or non-parametric tests like the Kruskal-Wallis test can be considered.
- Effect Size: Similar to the Kruskal-Wallis test, effect size measures should be reported alongside the p-value to assess the practical significance of the differences between cluster means.
- Post-hoc Tests: When the ANOVA indicates significant differences, post-hoc tests (e.g., Tukey's HSD, Fisher’s LSD for unequal variances) are required to identify which specific pairs of clusters differ significantly for the given non-categorical variable.
- Assumptions Validation: Before conducting ANOVA, it is essential to validate the assumptions of normality and homogeneity of variances (e.g., Levene's test, Bartlett's test). If assumptions are violated, appropriate transformations or non-parametric alternatives may be required.
- Visualization: Box plots, violin plots, or interaction plots can help visualize the distribution and variability of non-categorical variables within and across clusters, aiding in the interpretation of the ANOVA results.

Multiple Testing Correction: When conducting multiple tests (e.g., testing multiple categorical and non-categorical variables), it is crucial to control the family-wise error rate (FWER) or the false discovery rate (FDR) to avoid an inflated Type I error (false positives). Techniques such as Bonferroni correction, Holm-Bonferroni method, or the Benjamini-Hochberg procedure can be employed to adjust the significance level accordingly.

Cluster Profiling: After identifying the distinguishing features and characteristics of each cluster, it is essential to profile the clusters by combining the insights from both categorical and non-categorical variables. This profiling can reveal patterns, preferences, or behaviors specific to each cluster, which can be leveraged for targeted marketing, product development, or operational strategies in the space travel and tourism industry.

Domain Knowledge Integration: Throughout the analysis process, it is crucial to collaborate with domain experts or stakeholders to ensure the interpretation of the cluster characteristics aligns with the business context and objectives. Their insights can validate the meaningfulness of the identified patterns, suggest alternative explanations, or highlight potential limitations or biases in the data or analysis methods.

Iterative Refinement: Cluster analysis is often an iterative process, where the initial results may prompt further exploration, variable selection, or adjustments to the clustering algorithm parameters. Continuous refinement and validation with domain experts can lead to more robust and actionable insights from the clustering analysis.

4.) OBSERVATIONS

4.1. Optimal Number of Clusters: Base Model (K-Means)

The Optimal Number Of Segments Or Clusters Comes Out To Be 3 Which Is a Non-Trivial Set Of Cluster When Unsupervised Machine Learning K-Means Clustering Algorithm Is Performed on the dataset of The Interstellar Space Travel And Tourism Because The Silhouette Score for K=3 is the Highest among all the Values for K(K=2,3,4,5) Performed And the Overall Mean For K=3 is 0.586 Which is the Highest Among all the Iterations.

For K=2

Row ID	D Mean Silhouette Co...
cluster_0	0.649
cluster_1	0.604
Overall	0.63

For K=3

Row ID	D Mean Silhou...
cluster_0	0.637
cluster_1	0.507
cluster_2	0.604
Overall	0.586

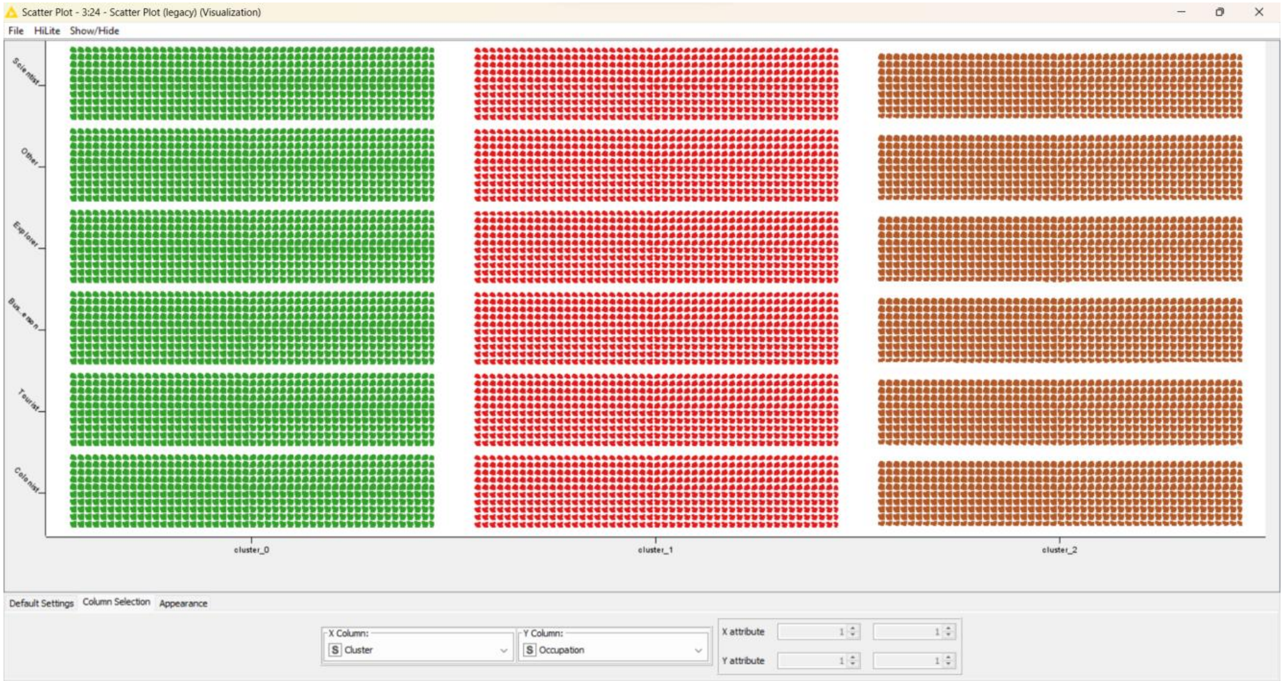
For K=4

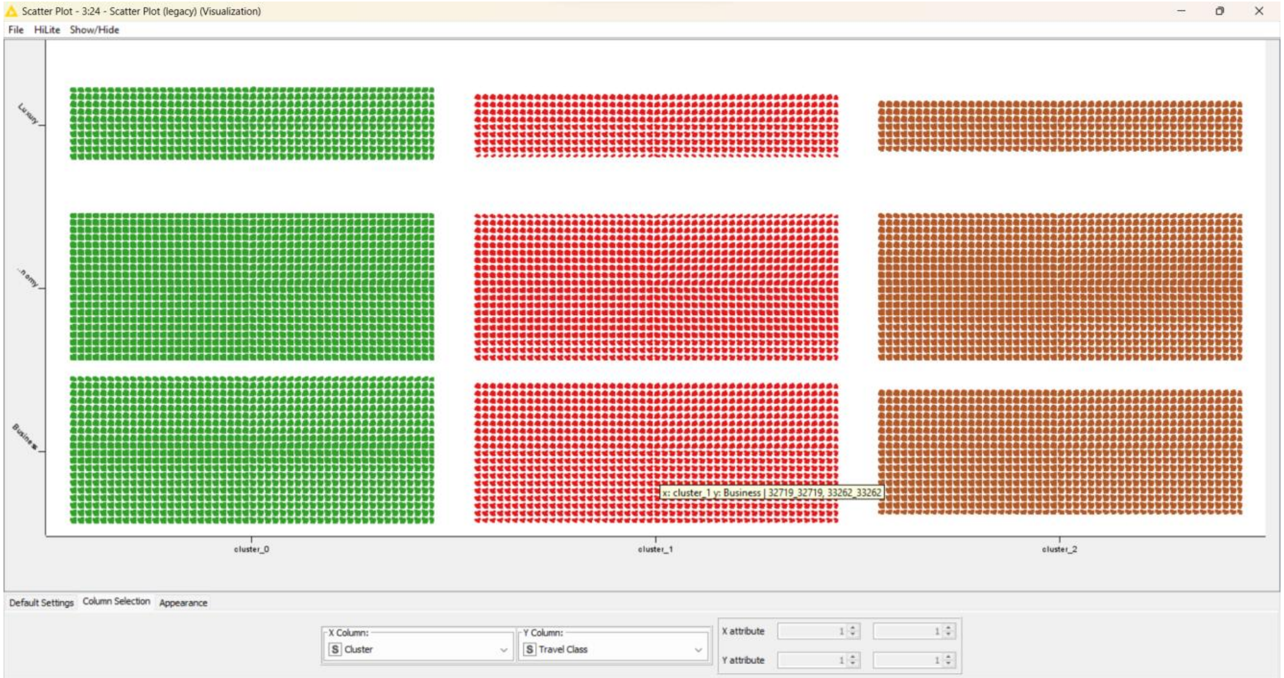
Row ID	D Mean Silhoue...
cluster_0	0.629
cluster_1	0.502
cluster_2	0.502
cluster_3	0.606
Overall	0.562

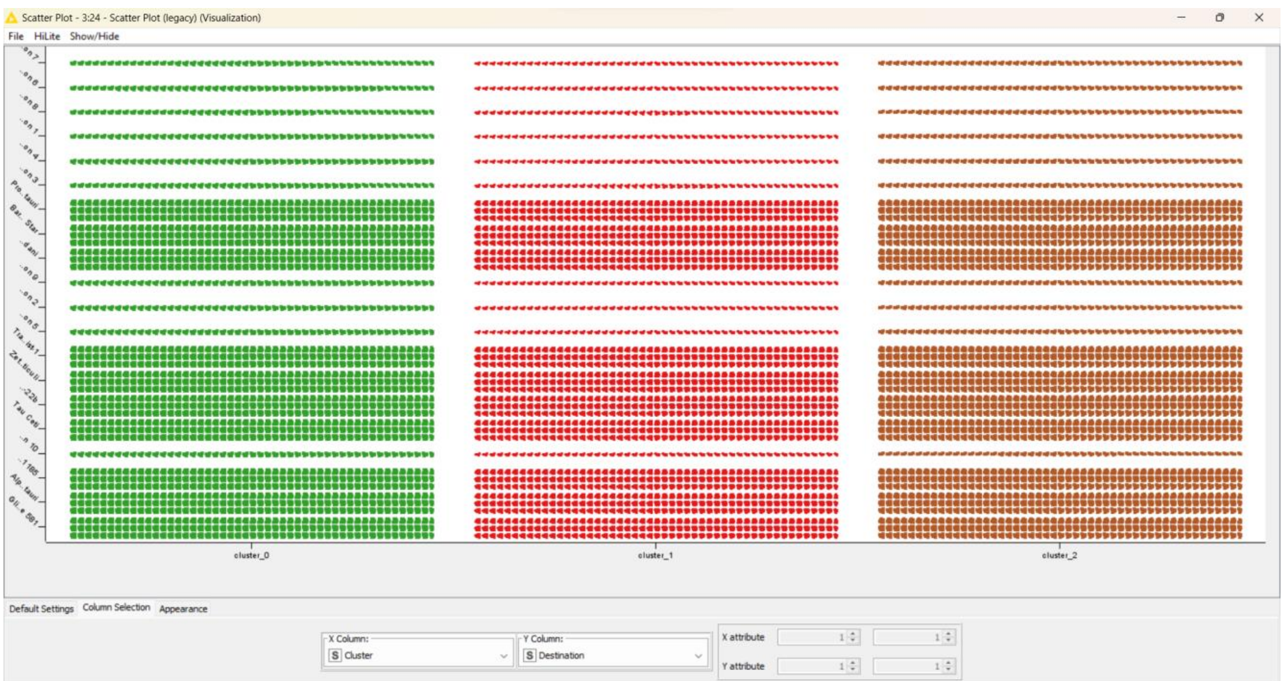
For K=5

Row ID	Mean Silhoue...
cluster_3	0.62
cluster_0	0.495
cluster_1	0.498
cluster_2	0.503
cluster_4	0.602
Overall	0.545

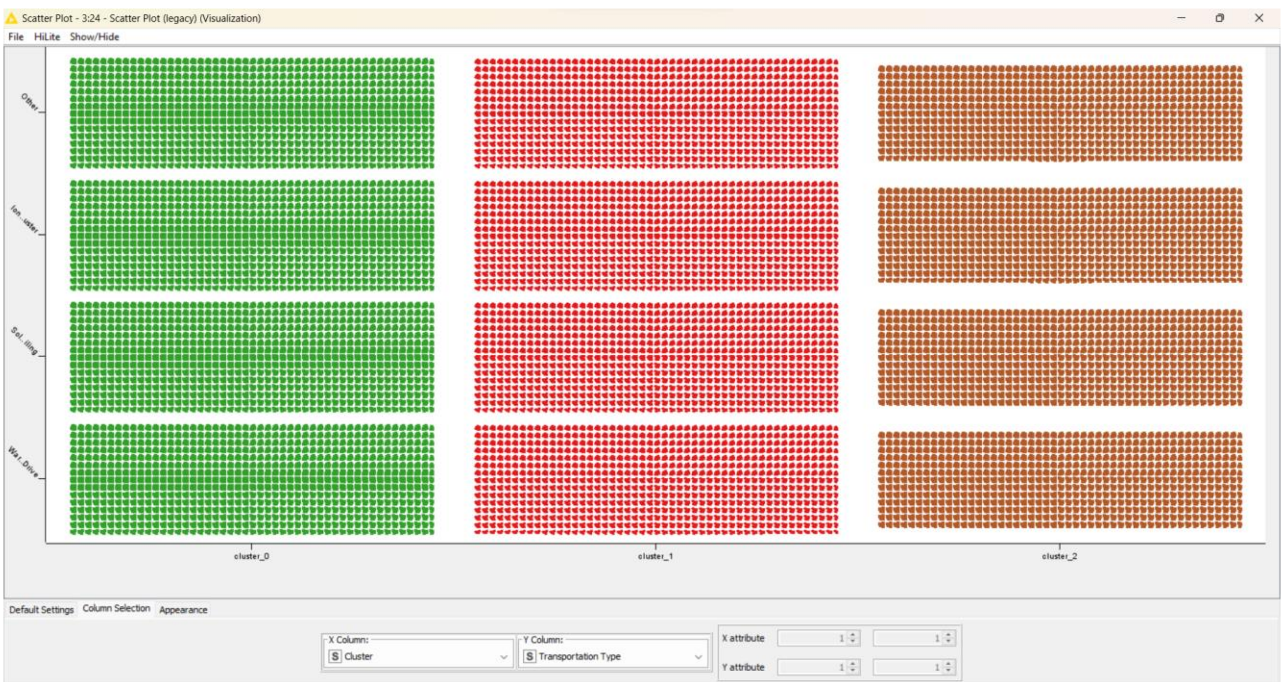
4.2 SCATTER PLOTS




Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
x: cluster_1 y: Business | 32719_32719, 33262_33262
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Travel Class
X attribute
Y attribute

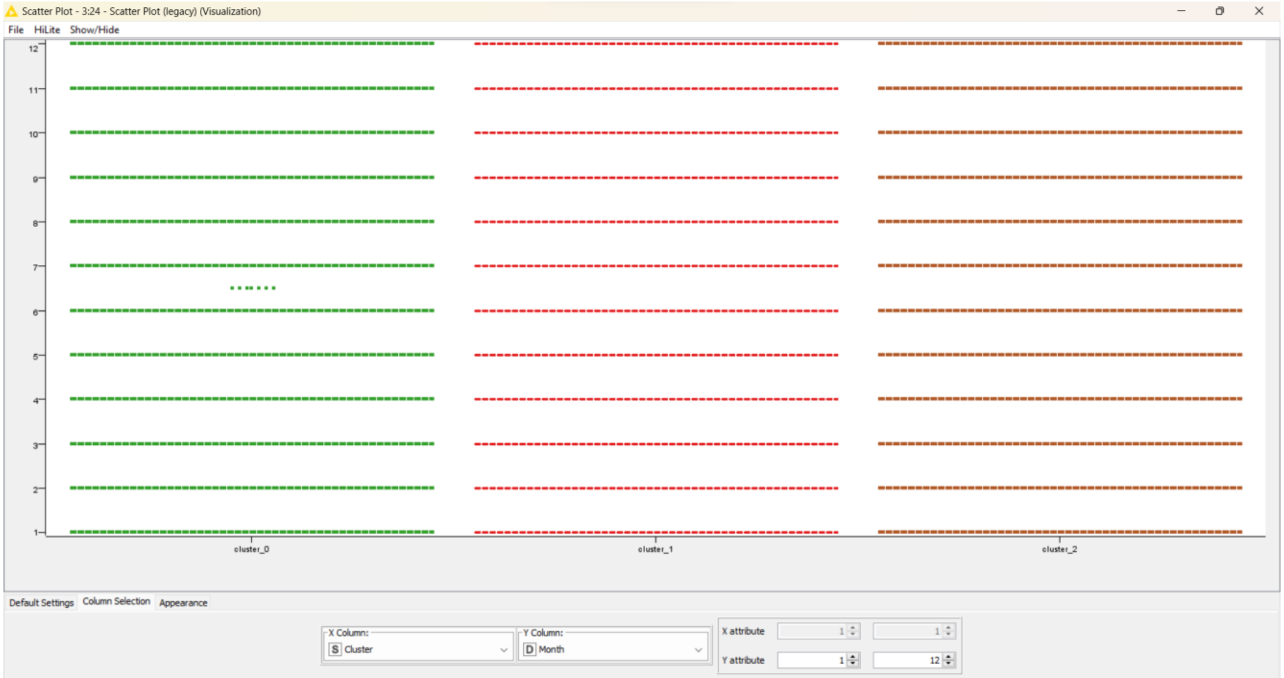

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Destination
X attribute
Y attribute

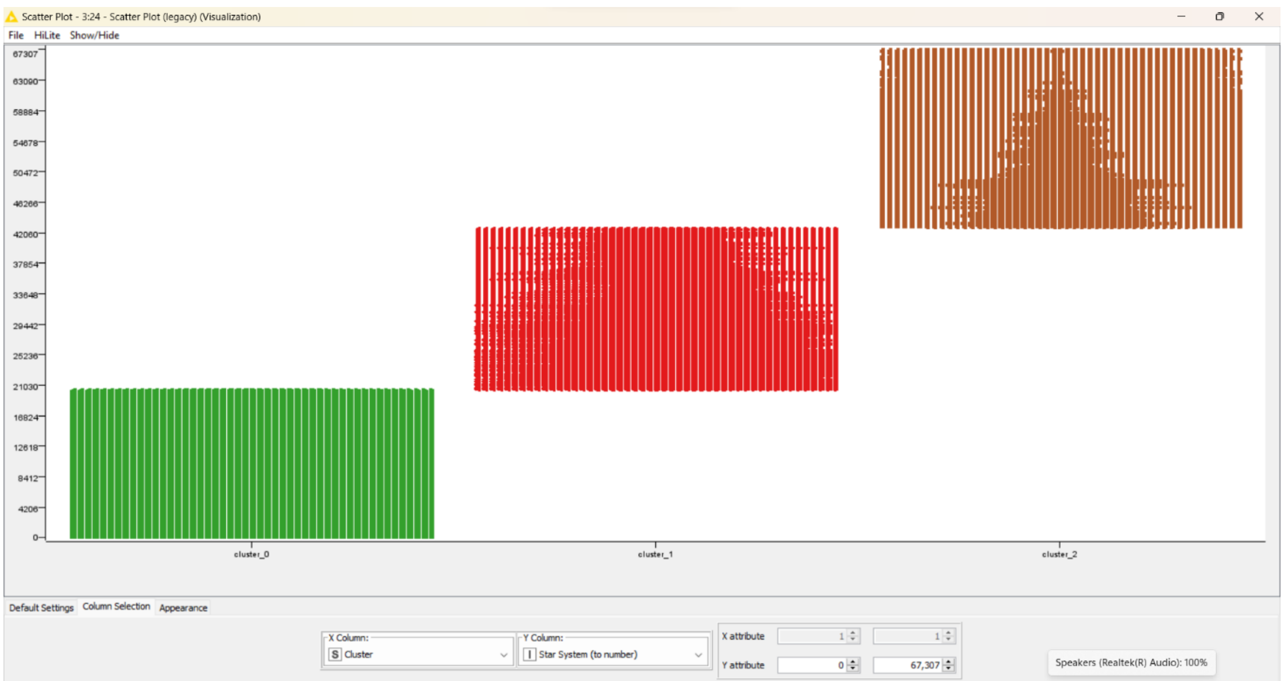

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
S Purpose of Travel
X attribute
Y attribute

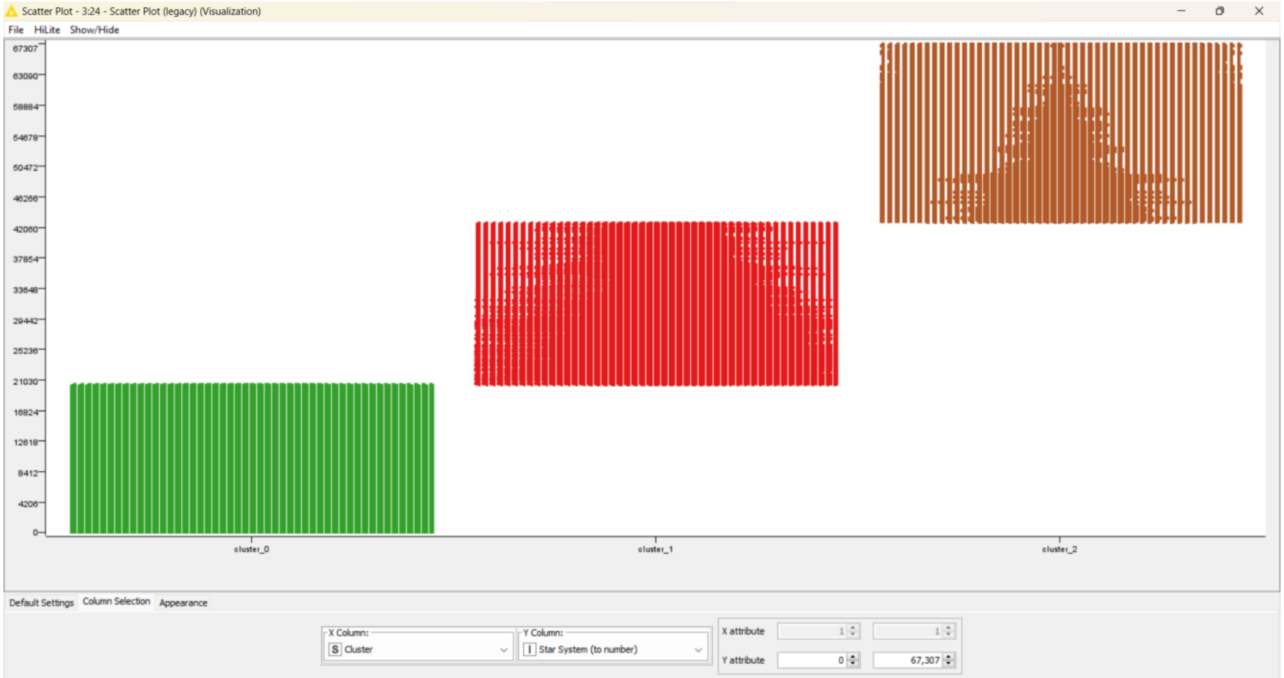

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
S Transportation Type
X attribute
Y attribute

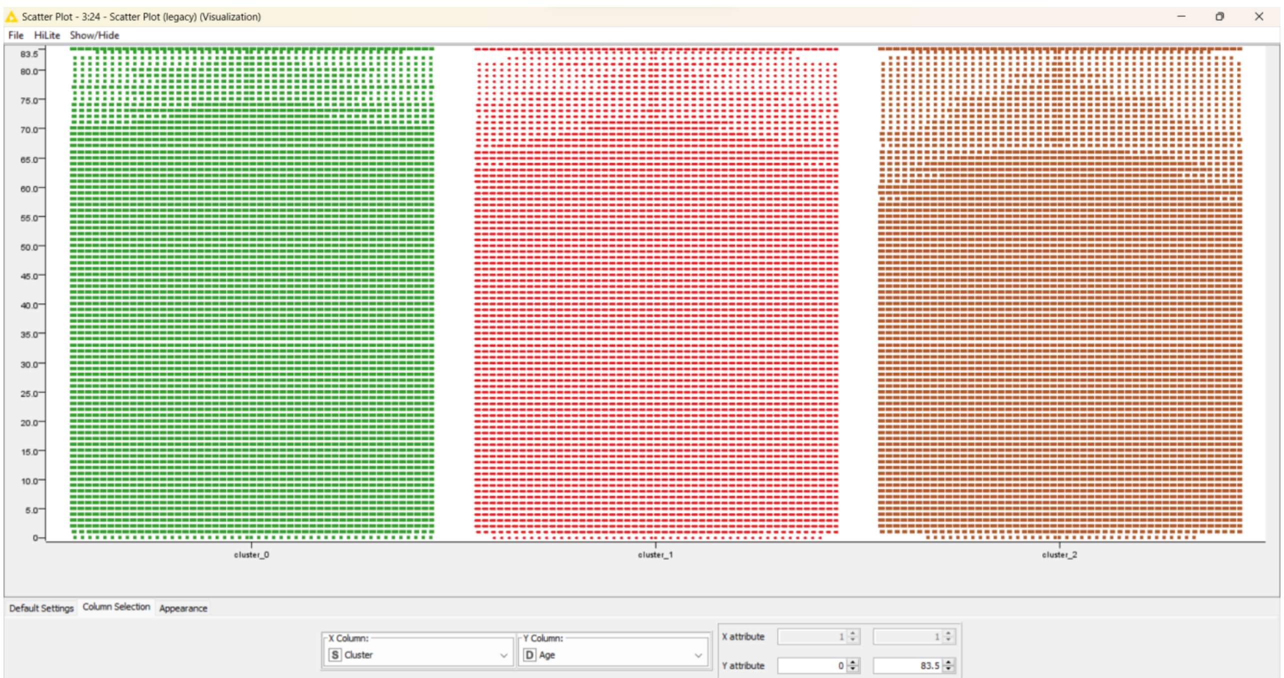

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Special Requests
X attribute
Microsoft Store

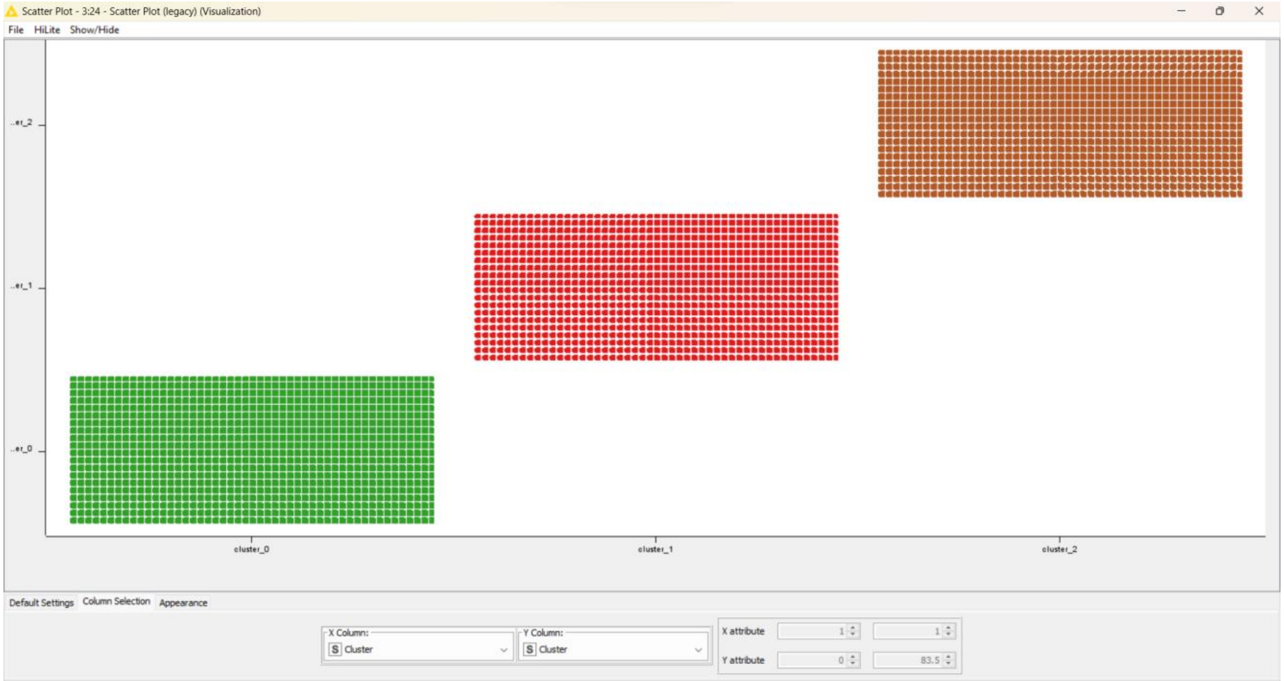

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
Yes
No
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Loyalty Program Member
X attribute
Y attribute

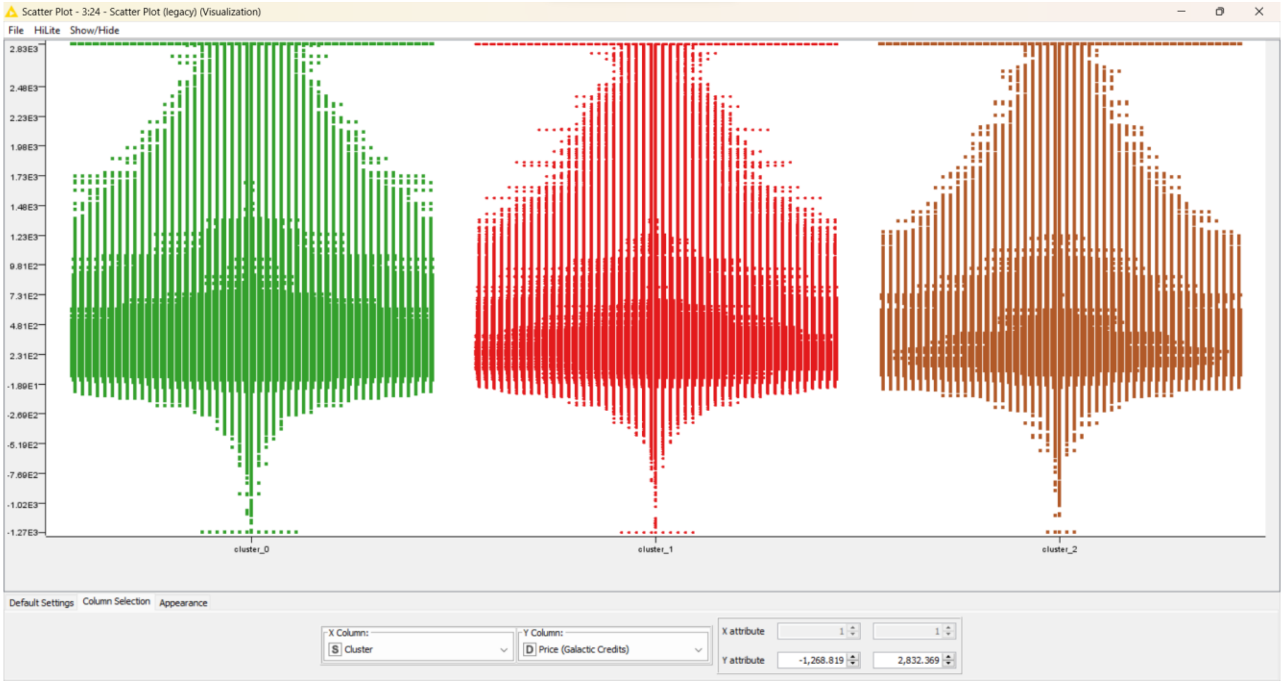

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column: Cluster
Y Column: Month
X attribute
Y attribute

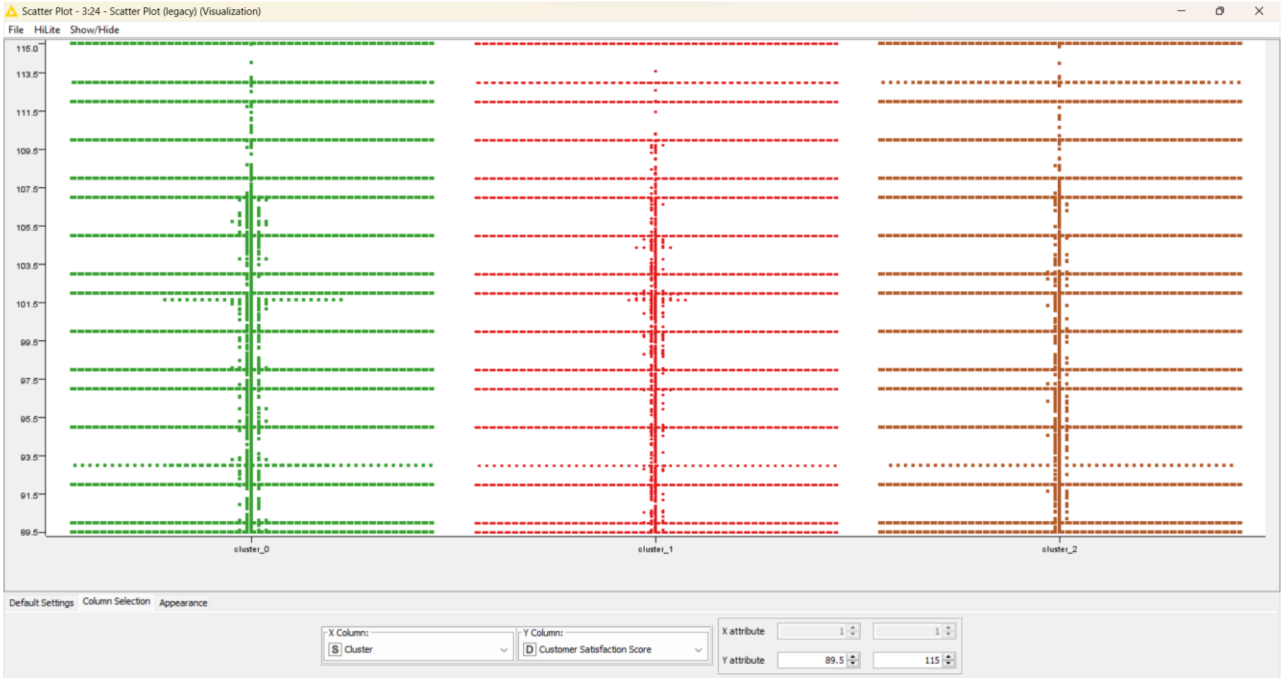

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column: Cluster
Y Column: Star System (to number)
X attribute
Y attribute
Speakers (Realtek(R) Audio): 100%


Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
67307
63090
58884
54678
50472
46266
42060
37854
33648
29442
25236
21030
16824
12618
8412
4206
0
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
I Star System (to number)
X attribute
Y attribute
67,307

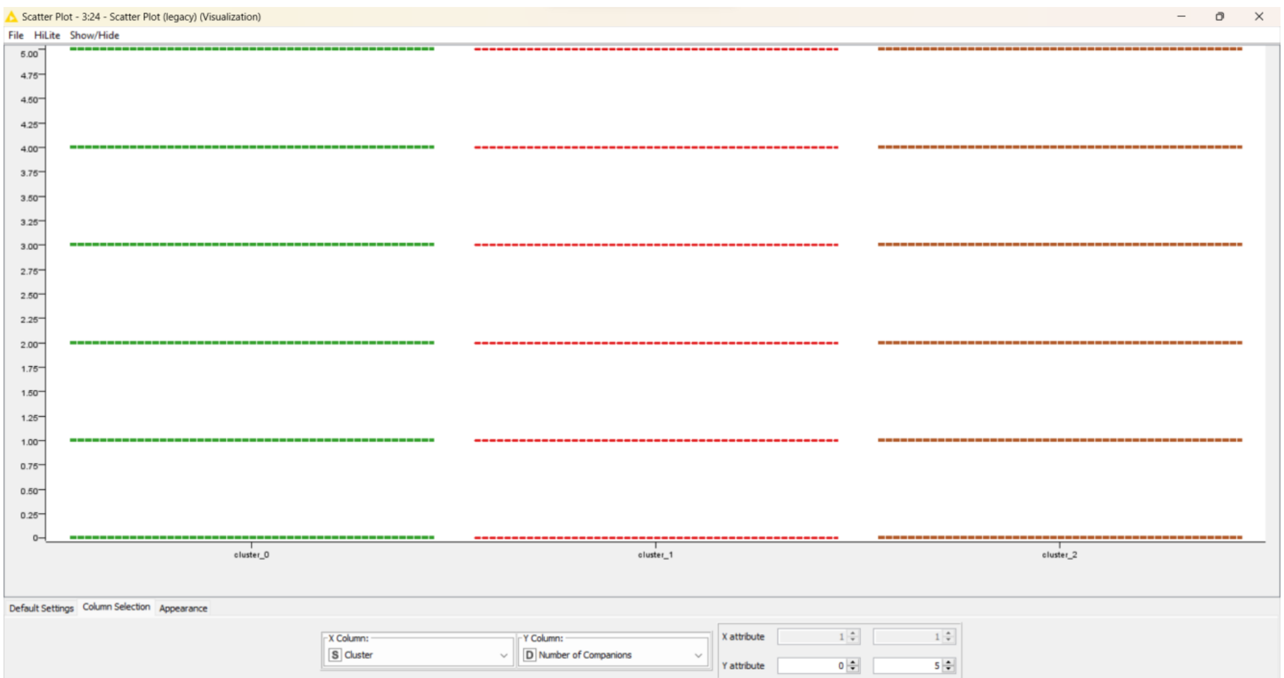

Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
83.5
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
D Age
X attribute
Y attribute
83.5


Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Cluster
X attribute
Y attribute
83.5


Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
2.83E3
2.48E3
2.23E3
1.98E3
1.73E3
1.48E3
1.23E3
9.81E2
7.31E2
4.81E2
2.31E2
-1.89E1
-2.69E2
-5.19E2
-7.69E2
-1.02E3
-1.27E3
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Price (Galactic Credits)
X attribute
Y attribute
-1,268.819
2,832.369



Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
D Customer Satisfaction Score
X attribute
Y attribute
89.5
115



Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
S Cluster
Y Column:
D Number of Companions
X attribute
Y attribute
0
5


Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Duration of Stay (Earth Days)
X attribute
Y attribute
101


Scatter Plot - 3:24 - Scatter Plot (legacy) (Visualization)
File HiLite Show/Hide
cluster_0
cluster_1
cluster_2
Default Settings Column Selection Appearance
X Column:
Cluster
Y Column:
Distance to Destination (Light-Years)
X attribute
Y attribute
17.07

4.3 Determining Characteristics From Cluster Analysis: Base Model (K-Means)

4.3.1 Categorical Variables or Features: Contributing or Significant | Non-Contributing or Non-Significant

The k-means clustering algorithm, coupled with the Kruskal-Wallis test, has proven to be an effective approach for identifying the significant categorical variables that contribute to the characterization of the three clusters in the space travel and tourism dataset. By analyzing the distributions of these variables across the clusters, valuable insights can be gained to inform business strategies and enhance customer experiences.

Contributing or Significant Categorical Variables:

1. **Travel Class:**

    - The significance of the Travel Class variable suggests that the preferences and expectations of customers regarding service levels, amenities, and comfort during their space travel experience differ across the identified clusters.
    - For instance, one cluster may predominantly consist of customers who prioritize luxury and high-end services, while another cluster may be more cost-conscious, opting for economy or basic travel classes.
    - By understanding these preferences, space travel companies can tailor their offerings, pricing strategies, and on-board experiences to better cater to the distinct needs of each cluster.
    - Additionally, targeted marketing campaigns and promotions can be designed to highlight the unique selling points and value propositions that resonate with each customer segment.

2. **Star System**:

- The significance of the Star System variable highlights the varying interests and motivations of customers when choosing their celestial destinations.
- Some clusters may gravitate towards popular or well-known star systems, while others may seek out more exotic or off-the-beaten-path destinations.
- This information can guide space travel companies in optimizing their travel routes, allocating resources based on demand patterns, and developing immersive experiences or educational programs tailored to the specific interests of each cluster.
- Collaborations with astrophysicists, astronomers, and planetarium experts can enhance the educational and exploratory aspects of space travel, catering to the curiosity and intellectual pursuits of certain customer segments

  .

3. **Transportation Type**:

- The significance of the Transportation Type variable suggests that customers have varying preferences and requirements when it comes to the mode of transportation used for their space travel.
- Some clusters may prioritize speed and efficiency, favoring advanced spacecraft or shuttles, while others may value the novelty and experience of traveling in more traditional or historic transportation modes.
- Space travel companies can leverage this information to optimize their transportation fleets, allocate resources for maintenance and upgrades, and potentially invest in developing new transportation technologies tailored to the specific needs and preferences of each cluster.
- Additionally, partnerships with transportation manufacturers and research institutions can drive innovation and enhance the overall travel experience for customers across all clusters.

Non-Contributing or Non-Significant Categorical Variables:

While the Kruskal-Wallis test identified seven categorical variables as non-contributing or non-significant in differentiating the three clusters, it is crucial to acknowledge that these variables should not be dismissed entirely. There are several considerations and potential avenues for further exploration:

1. **Interaction Effects**:

    - The non-contributing variables may exhibit significance when analyzed in combination with other variables or when considering interaction effects.
    - For example, the variable "Occupation" may not be significant on its own, but when combined with "Travel Class" or "Purpose of Travel," it could reveal interesting patterns or preferences specific to certain customer segments.
    - Exploring these interaction effects through advanced statistical techniques, such as two-way or higher-order ANOVA, or regression models with interaction terms, can uncover additional insights and nuances in the data.

2. **Alternative Clustering Techniques**:

    - The k-means algorithm is a widely used clustering technique, but it may not be the most suitable method for all datasets or variable types.
    - Alternative clustering algorithms, such as hierarchical clustering, DBSCAN, or model-based clustering, may reveal different patterns or highlight the significance of variables that were previously deemed non-contributing.
    - Experimenting with different clustering techniques and comparing the results can provide a more robust understanding of the underlying structure and relationships within the data.

3. **Domain Knowledge and Business Objectives**:

    - While the statistical tests may deem certain variables as non-contributing, these variables should be evaluated in the context of domain knowledge and business objectives.
    - For instance, the "Loyalty Program Member" variable may not contribute significantly to the cluster characterization, but it could be a critical factor in developing customer loyalty programs and retention strategies.
    - Similarly, the "Gender" variable, although non-significant in this analysis, may be relevant for targeted marketing campaigns, product design, or addressing specific needs and preferences of different genders in the space travel industry.

4. **Data Quality and Completeness**:

    - The non-contributing nature of certain variables may be influenced by data quality issues, such as missing values, inconsistencies, or measurement errors.
    - Addressing these data quality concerns through data cleaning, imputation techniques, or collecting additional data can potentially reveal previously undetected patterns or relationships.

5. **Outlier Analysis and Robustness Checks**:

    - Outliers or extreme values in the data can potentially influence the results of the statistical tests and clustering algorithms.
    - Conducting robust outlier analysis and implementing appropriate techniques for handling outliers can improve the reliability and interpretability of the cluster characterization.

Travel Class

Row ID	D H-Value	D p-value	D Mean Rank of Group cluster_0	D Median Rank of Group cluste...	D Mean Rank of Group cluster_1	D Median Rank of Group cluste...	D Mean Rank of Group cluster_2	D Median Rank of Group cluste...
Row0	7.526	0.023214678441555003	39,977.03	47,481	40,361.57	47,481	39,887.734	47,481

Star System

Row ID	D H-Value	D p-value	D Mean Rank of Group cluster_0	D Median Rank of Group cluste...	D Mean Rank of Group cluster_1	D Median Rank of Group cluste...	D Mean Rank of Group cluster_2	D Median Rank of Group cluste...
Row0	70,692.895	0.0	15,850	15,850	44,720.5	44,720.5	68,947.5	68,947.5

Gender

Row ID	D H-Value	D p-value	D Mean Rank of Group cluster_0	D Median Rank of Group cluste...	D Mean Rank of Group cluster_1	D Median Rank of Group cluste...	D Mean Rank of Group cluster_2	D Median Rank of Group cluste...
Row0	0.451	0.7981524986074076	40,112.281	56,773	40,009.567	56,773	40,105.453	56,773

Transportation Type

Row ID	D H-Value	D p-value	D Mean Rank of Group cluster_0	D Median Rank of Group cluste...	D Mean Rank of Group cluster_1	D Median Rank of Group cluste...	D Mean Rank of Group cluster_2	D Median Rank of Group cluste...
Row0	5.906	0.05217726711961126	39,953.741	30,077	40,353.859	50,151.5	39,929.635	30,077

4.3.2. <u>Non-Categorical Variables or Features: Contributing or Significant | Non-Contributing or Non-Significant</u>

In the context of the space travel and tourism dataset, the characterization of the three clusters derived from the k-means clustering algorithm was further enriched by analyzing the non-categorical variables using one-way ANOVA (Analysis of Variance). This statistical technique assesses whether the means of a continuous variable differ significantly across the identified clusters. The analysis was conducted at a 90% confidence level, with a significance threshold of 0.05 for the p-values.

The choice of a 90% confidence level, instead of the more commonly used 95% confidence level, was made to increase the statistical power of the analysis and to capture potentially relevant differences in means that may have been overlooked at a stricter significance level. In exploratory data analysis and cluster characterization, it is often beneficial to adopt a more lenient significance threshold to uncover potential patterns or relationships that may warrant further investigation or validation.

Contributing or Significant Non-Categorical Variables:

The one-way ANOVA results revealed that two non-categorical variables exhibited statistically significant differences in their means across the three clusters. These variables were:

1. **Distance to Destination (Light-Years)**
2. **Duration of Stay (Earth Days)**

Distance to Destination (Light-Years):

- The Distance to Destination variable represents the spatial distance, measured in light-years, between the point of origin and the celestial destination visited by the customers.
- The significant difference in the means of this variable across the clusters (p-value < 0.05) suggests that certain customer segments have distinct preferences or constraints regarding the travel distance they are willing to undertake.

- This information can guide space travel companies in optimizing their travel routes, allocating resources for longer or shorter journeys, and developing appropriate pricing strategies based on the distance factor.
- Additionally, it can inform the design and development of spacecraft or transportation modes optimized for different distance ranges, ensuring efficient and comfortable travel experiences for each cluster.

Duration of Stay (Earth Days):

- The Duration of Stay variable represents the length of time, measured in Earth days, that customers spend at their celestial destination.
- The significant difference in the means of this variable across the clusters (p-value < 0.05) indicates that different customer segments have varying preferences or requirements regarding the duration of their stay at the destination.
- Some clusters may prefer shorter stays, perhaps due to time constraints or a desire for a quick getaway, while others may opt for extended stays, allowing for more immersive experiences and exploration.
- This information can assist space travel companies in tailoring their accommodation offerings, activity packages, and on-site experiences to cater to the distinct needs and preferences of each cluster.
- It can also inform resource allocation and staffing strategies at the destination facilities to ensure optimal service delivery and customer satisfaction.

Non-Contributing or Non-Significant Non-Categorical Variables:

The remaining six non-categorical variables – Age, Number of Companions, Price (Galactic Credits), Booking Date, Departure Date, and Customer Satisfaction Score – were found to be non-contributing or non-significant in differentiating the three clusters based on the one-way ANOVA results. Their p-values exceeded the 0.05 significance level, indicating that the means of these variables did not differ significantly across the clusters.

It is important to note that while the p-values for these variables were not statistically significant at the chosen 0.05 level, some variables exhibited an F-value greater than 1, suggesting potential differences in means across the clusters. However, these differences may be due to random chance or other factors, and further investigation or alternative statistical techniques may be required to assess their true significance.

Interpretation and Considerations:

1. **Effect Size and Practical Significance**:
   - In addition to the p-values, it is crucial to examine effect size measures, such as eta-squared or omega-squared, to gauge the practical significance of the differences in means between the clusters.
   - Effect sizes provide an understanding of the magnitude of the observed differences, ensuring that statistically significant results also translate to practical relevance in the context of the space travel and tourism industry.

2. **Assumption Validation**:
   - One-way ANOVA relies on several assumptions, including normality of the variable within each cluster and homogeneity of variances across clusters.
   - Violation of these assumptions can affect the validity of the results, and appropriate remedial measures, such as data transformations or robust alternatives (e.g., Welch's ANOVA or non-parametric tests), may be required.

3. **Post-hoc Tests**:
   - When the one-way ANOVA indicates significant differences in means, post-hoc tests (e.g., Tukey's HSD, Bonferroni correction) are necessary to identify which specific pairs of clusters differ significantly for the given non-categorical variable.
   - These pairwise comparisons provide valuable insights into the nature and direction of the differences, facilitating more targeted strategies and decision-making.

4. **Interaction with Categorical Variables**:
- The interpretation of the contributing and non-contributing non-categorical variables should be made in conjunction with the findings from the categorical variable analysis.
- Exploring potential interactions between categorical and non-categorical variables can uncover more nuanced patterns and relationships, leading to a more comprehensive understanding of the cluster characteristics.

5. **Domain Knowledge and Business Context**:
- As with the categorical variables, the interpretation of the non-categorical variable analysis should be guided by domain knowledge and the specific business objectives of the space travel and tourism industry.
- Collaboration with industry experts, stakeholders, and business analysts is crucial to ensure that the insights derived from the statistical analysis are actionable, practical, and aligned with the overall goals of enhancing customer experiences and optimizing operations.

Justification for Using a 90% Confidence Level:

In the context of exploratory data analysis and cluster characterization, the choice of a 90% confidence level (corresponding to a significance level of 0.10) can be justified based on the following considerations:

1. **Increased Statistical Power**:
- By using a less stringent significance level of 0.10 instead of the more commonly used 0.05, the statistical power to detect potentially relevant differences in means across clusters is increased.
- This approach reduces the likelihood of committing a Type II error (failing to reject a null hypothesis that is actually false), which is particularly important in exploratory analyses where the goal is to uncover patterns and relationships that may warrant further investigation.

2. **Balancing Type I and Type II Errors**:
- While a higher significance level (e.g., 0.10) increases the risk of committing a Type I error (rejecting a null hypothesis that is true), it simultaneously decreases the risk of committing a Type II error.

- In exploratory analyses, it is often preferable to err on the side of caution and consider potentially relevant differences, even if they may be due to chance, rather than overlooking potentially valuable insights.

3. **Exploratory Nature of the Analysis**:
   - The cluster characterization process is inherently exploratory, aimed at identifying patterns and relationships that can inform business strategies and decision-making.
   - By adopting a more lenient significance level, the analysis can uncover potential avenues for further investigation, hypothesis testing, or validation using additional data or more rigorous statistical methods.

4. **Domain Knowledge and Business Priorities**:
   - The choice of the confidence level should also be guided by domain knowledge and the specific priorities and objectives of the space travel and tourism industry.
   - In certain contexts, it may be more crucial to identify potentially relevant differences, even if they are marginally significant, rather than risking overlooking important insights that could provide a competitive advantage or enhance customer experiences.

It is important to note that the interpretation of the results obtained using a 90% confidence level should be done with caution and in conjunction with effect size measures and domain knowledge. The identified significant differences should be viewed as preliminary findings that require further validation, replication, or hypothesis testing using more stringent significance levels or alternative statistical techniques.

By integrating statistical rigor with domain expertise and business priorities, space travel companies can leverage the insights derived from the cluster analysis to make informed decisions, optimize operations, and deliver exceptional experiences tailored to the distinct preferences and characteristics of their diverse customer segments.

One-way analysis of variance (ANOVA)

Descriptive Statistics

Confidence Interval (CI) Probability: 90.0%

	Group	N	Missing	Missing Group	Mean	Std. Deviation	Std. Error	CI (Lower Bound)	CI (Upper Bound)	Minimum	Maximum
Age	cluster_0	31699	0	0	0.3734	0.2338	0.0013	0.3712	0.3756	0.0	1
Age	cluster_1	26042	0	0	0.3743	0.2351	0.0015	0.3719	0.3767	0.0	1
Age	cluster_2	22412	0	0	0.3713	0.2344	0.0016	0.3687	0.3739	0.0	1
Age	Total	80153	0	0	0.3731	0.2344	0.0008	0.3717	0.3745	0.0	1
Distance to Destination (Light-Years)	cluster_0	31699	0	0	0.2977	0.3214	0.0018	0.2948	0.3007	0.0006	1
Distance to Destination (Light-Years)	cluster_1	26042	0	0	0.305	0.3268	0.002	0.3017	0.3084	0.0	1
Distance to Destination (Light-Years)	cluster_2	22412	0	0	0.3003	0.3229	0.0022	0.2967	0.3038	0.0006	1
Distance to Destination (Light-Years)	Total	80153	0	0	0.3008	0.3236	0.0011	0.2989	0.3027	0.0	1
Duration of Stay (Earth Days)	cluster_0	31699	0	0	0.3205	0.2798	0.0016	0.3179	0.323	0.0	1
Duration of Stay (Earth Days)	cluster_1	26042	0	0	0.3214	0.2823	0.0017	0.3185	0.3242	0.0	1
Duration of Stay (Earth Days)	cluster_2	22412	0	0	0.3255	0.2836	0.0019	0.3224	0.3286	0.0	1
Duration of Stay (Earth Days)	Total	80153	0	0	0.3222	0.2817	0.001	0.3205	0.3238	0.0	1
Number of Companions	cluster_0	31699	0	0	0.222	0.2165	0.0012	0.22	0.224	0.0	1
Number of Companions	cluster_1	26042	0	0	0.221	0.2157	0.0013	0.2188	0.2232	0.0	1
Number of Companions	cluster_2	22412	0	0	0.2205	0.2172	0.0015	0.2182	0.2229	0.0	1
Number of Companions	Total	80153	0	0	0.2213	0.2164	0.0008	0.22	0.2225	0.0	1
Price (Galactic Credits)	cluster_0	31699	0	0	0.5268	0.2054	0.0012	0.5249	0.5287	0.0	1
Price (Galactic Credits)	cluster_1	26042	0	0	0.5285	0.2066	0.0013	0.5264	0.5306	0.0	1
Price (Galactic Credits)	cluster_2	22412	0	0	0.5271	0.204	0.0014	0.5249	0.5294	0.0	1
Price (Galactic Credits)	Total	80153	0	0	0.5274	0.2054	0.0007	0.5262	0.5286	0.0	1

Customer Satisfaction Score	cluster_0	31699	0	0	0.5153	0.2397	0.0013	0.5131	0.5175	0.0	1
Customer Satisfaction Score	cluster_1	26042	0	0	0.5147	0.2401	0.0015	0.5122	0.5171	0.0	1
Customer Satisfaction Score	cluster_2	22412	0	0	0.5187	0.2423	0.0016	0.516	0.5213	0.0	1
Customer Satisfaction Score	Total	80153	0	0	0.516	0.2406	0.0008	0.5146	0.5174	0.0	1

Levene Test

The Levene Test is used to test for the equality of variances.

	F	df 1	df 2	p-Value
Age	0.5137	2	80150	0.5983
Distance to Destination (Light-Years)	7.1826	2	80150	0.0008
Duration of Stay (Earth Days)	4.6344	2	80150	0.0097
Number of Companions	0.3154	2	80150	0.7295
Price (Galactic Credits)	3.4141	2	80150	0.0329
Customer Satisfaction Score	2.5645	2	80150	0.077

ANOVA

	Source	Sum of Squares	df	Mean Square	F	p-value
Age	Between Groups	0.1121	2	0.056	1.0199	0.3606
Age	Within Groups	4,403.7458	80150	0.0549		
Age	Total	4,403.8579	80152			
Distance to Destination (Light-Years)	Between Groups	0.7653	2	0.3826	3.6538	0.0259
Distance to Destination (Light-Years)	Within Groups	8,393.4155	80150	0.1047		
Distance to Destination (Light-Years)	Total	8,394.1807	80152			
Duration of Stay (Earth Days)	Between Groups	0.3594	2	0.1797	2.265	0.1038
Duration of Stay (Earth Days)	Within Groups	6,359.1504	80150	0.0793		
Duration of Stay (Earth Days)	Total	6,359.5098	80152			
Number of Companions	Between Groups	0.0321	2	0.016	0.3423	0.7101
Number of Companions	Within Groups	3,754.845	80150	0.0468		
Number of Companions	Total	3,754.8771	80152			
Price (Galactic Credits)	Between Groups	0.0441	2	0.0221	0.5234	0.5925
Price (Galactic Credits)	Within Groups	3,380.3072	80150	0.0422		
Price (Galactic Credits)	Total	3,380.3513	80152			
Customer Satisfaction Score	Between Groups	0.223	2	0.1115	1.927	0.1456
Customer Satisfaction Score	Within Groups	4,638.2565	80150	0.0579		
Customer Satisfaction Score	Total	4,638.4795	80152			

5. MANAGERIAL INSIGHTS

Cluster Segmentation and Targeting:

1. The identification of three distinct clusters suggests the presence of heterogeneous customer segments within the space travel and tourism market. This segmentation allows for targeted marketing strategies and tailored offerings to cater to the specific preferences and behaviors of each cluster.

2. **Cluster 0**: Highly Economy Class Dominated, Solar Sailer (Transportation Type), Michael Street (Star System), low mean duration of stay, low distance to destination

1. This segment represents highly price-sensitive customers seeking affordable, short-duration space travel experiences within nearby destinations.
2. Marketing efforts should emphasize budget-friendly packages, highlighting the cost-effectiveness of Solar Sailer transportation and the accessibility of the Michael Street star system.
3. Promotions can focus on quick getaways or weekend escapes, catering to the preferences for low duration of stay and proximity to the destination.

Cluster 0 represents a cost-conscious segment, as evidenced by their preference for the economy travel class. This segment is likely to be price-sensitive and may prioritize value-for-money experiences. Managerial strategies for this cluster could involve competitive pricing, package deals, or promotional offers to attract and retain these customers.

3. **Cluster 1**: Standard Economy Class Dominated, Ion Thruster (Transportation Type), Michael Mission (Star System), medium mean duration of stay, high distance to destination

1. This segment exhibits a balance between affordability and a willingness to travel longer distances for more immersive experiences.
2. Marketing strategies should highlight the unique selling points of the Michael Mission star system, positioning it as an off-the-beaten-path destination for adventure seekers.

3. Emphasize the capabilities of Ion Thruster transportation for efficient long-haul travel while maintaining cost-effectiveness.
4. Offer modular packages options to cater to varying preferences for duration of stay within this segment.

Cluster 1 exhibits a preference for economy travel class and longer distances to destinations, suggesting a segment that values affordable exploration and adventure. Marketing campaigns highlighting unique or off-the-beaten-path destinations and immersive experiences could resonate well with this cluster.

4. **<u>Cluster 2</u>**: Less Economy Class Dominated, Warp Drive (Transportation Type), John Turnpike (Star System), high mean duration of stay, medium distance to destination

1. This segment represents customers with a higher propensity for premium experiences and a willingness to invest in luxury offerings, including advanced transportation and extended stays.
2. Marketing campaigns should emphasize the exclusivity, luxury, and cutting-edge technology of Warp Drive transportation, positioning it as a premium mode of travel.
3. Highlight the John Turnpike star system as a destination suitable for extended stays, offering curated activities, personalized experiences, and luxury accommodations.
4. Collaborate with high-end hospitality brands and activity providers to create exclusive packages tailored to this segment's preferences.

Cluster 2, characterized by a preference for economy travel class, extended stays, and medium distances, represents a segment that seeks more immersive and leisurely experiences. Offering tailored accommodation packages, curated activities, and on-site amenities catering to longer stays could appeal to this segment.

Transportation and Fleet Management:

1. Ensure a balanced allocation of transportation modes across the clusters, prioritizing Solar Sailers for Cluster 0, Ion Thrusters for Cluster 1, and Warp Drive spacecraft for Cluster 2.
2. Optimize maintenance schedules and resource allocation based on the varying distance and duration preferences of each cluster.

Service and Operational Optimization:

1. The identification of preferred transportation types within each cluster (Solar Sailer, Ion Thruster, and Warp Drive) provides insights into the desired mode of travel for different customer segments. This information can guide fleet management, resource allocation, and investment decisions to ensure the availability of the preferred transportation options for each cluster.

2. The varying duration of stay preferences across the clusters necessitates adjustments in accommodation offerings, staffing levels, and on-site resources at the destination facilities. For clusters with shorter stay preferences (Cluster 0), efficient check-in/check-out processes and streamlined services may be prioritized. Conversely, for clusters with longer stay preferences (Cluster 2), more comprehensive amenities, activity programs, and personalized services could enhance the customer experience.

3. The identification of preferred star systems or destinations within each cluster (Michael Street, Michael Mission, and John Turnpile) can inform route planning, destination marketing, and the development of tailored experiences or educational programs specific to each location.

3. The observed differences in distance preferences across the clusters can guide decisions regarding travel route optimization, spacecraft design (for long or short-haul travel), and pricing strategies based on travel distance.

Product Development and Innovation:

1. The cluster insights can drive innovation and new product development within the space travel and tourism industry. For example, clusters with preferences for extended stays may benefit from the development of space-based accommodation facilities or orbital resorts, providing unique and immersive experiences.

2. The identification of transportation type preferences can spur research and development efforts to enhance existing modes of space travel or explore new propulsion systems tailored to the needs of different customer segments.

4. Cross-cluster analysis and trend monitoring can help identify emerging preferences or shifts in customer behavior, enabling proactive adaptation and the introduction of new offerings or services to stay ahead of the competition.

Strategic Partnerships and Collaborations:

1. Partnerships with transportation manufacturers, aerospace companies, or research institutions can facilitate the development of preferred transportation modes or the optimization of existing fleets based on the identified cluster preferences.

2. Collaborations with destination management organizations, local tourism authorities, or educational institutions can enhance the on-site experiences and educational offerings tailored to the specific interests and preferences of each cluster.

4. Strategic alliances with accommodation providers, activity organizers, or hospitality companies can ensure the availability of suitable options and seamless experiences for customers with varying duration of stay preferences.

Group table - 3:37 - GroupBy

File Edit Hilite Navigation View

Table "default" - Rows: 3 Spec - Columns: 13 Properties Flow Variables

Row ID	S Cluster	S Unique concatenate with count*(Travel Class)	D Mean(Duration of Stay (Earth Days))	D Standard deviation(Duration of St...	S Unique concatenate with count*(Transportation Type)	D Mean(Distance to Destination (Lig...	D Standard deviation (Distance to Des...	I Unique count*(Star System)	I Unique count*(Transportation Type)	S Mode*(Travel Class)	S Mode*(Star System)	S Mode*(Transportation Type)	I Unique count*(Travel Class)
Row0	cluster_0	Luxury(4226), Economy(17295), Business(10178)	32.366	28.26	Ion Thruster(7919), Warp Drive(7961), Solar Sailing(7988), Other(7831)	5.083	5.487	17886	4	Economy	Michael Street	Solar Sailing	3
Row1	cluster_1	Economy(14277), Luxury(3595), Business(8170)	32.457	28.508	Solar Sailing(6474), Warp Drive(6391), Other(6566), Ion Thruster(6611)	5.207	5.579	18926	4	Economy	Michael Mission	Ion Thruster	3
Row2	cluster_2	Business(7198), Luxury(2917), Economy(12297)	32.876	28.646	Ion Thruster(5496), Warp Drive(5663), Other(5592), Solar Sailing(5661)	5.126	5.512	20167	4	Economy	Joseph Turn...	Warp Drive	3