DATA ANALYSIS- TAGS AND SEGMENTATION

KENDALL STARCEVICH, JACK UPTON
DRAKE INSURANCE INNOVATION LAB

Drake UNIVERSITY | Zimpleman College of Business

drake.edu/zimpleman

GOALS

1. **Third Party Data:** Analyze customer data to identify patterns related to third party and publicly available data
2. **Purchase/Product:** Explore customer segmentation based on purchasing behavior, state, product, and user demographic data (gender, age)
3. **Claims:** Study insurance claims data to identify the type of claims and the time of claims (e.g. is there an increase during a specific time?)
4. **Proposed Tags/Segments:** Develop data-driven tags and segments that allow personalized communication and provide recommendations.

STEPS

Build a KMEANS model in Snowflake using third party data.

Use clustering method to **explore key dimensions** in Purchase, Product, Claims (Behavior. Geographic, Psychological)

Review Third-Party Machine Learning Models. How and why was it different from our clustering and segmentation?

Using our clusters and segmentation, which fields most affect our target variables of **persistency and engagement?**

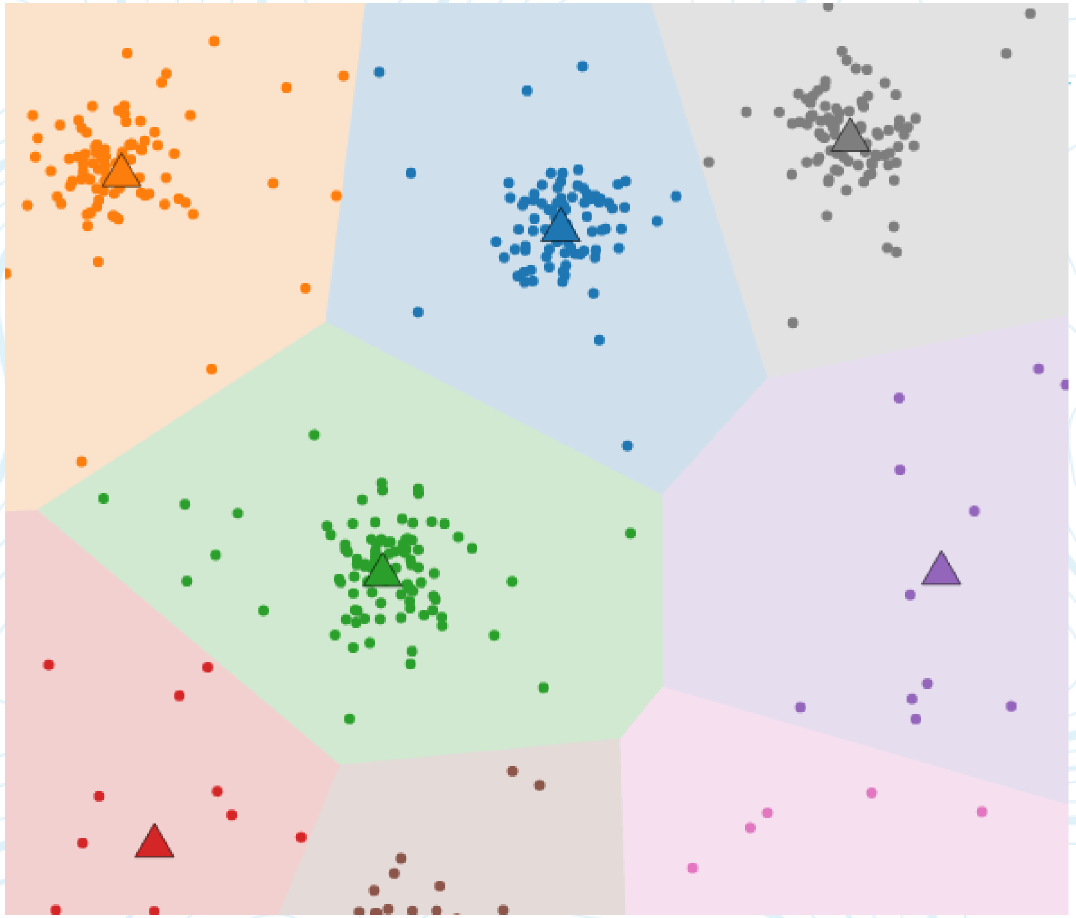
INTRO TO KMEANS

The algorithm will categorize the items we choose into **k groups** or **clusters** of similarity. To calculate that similarity, we will use the Euclidean Distance as a measurement.

1. We randomly initialize k points, called **centroids.**

2. We categorize each item to its closest centroid, and we update the centroid coordinates, which are the **averages** of the items categorized in that cluster so far.

3. We repeat the process for a given number of iterations and at the end, we have our **clusters**.



KMEANS STEPS

IMPORT DATA

In a Snowflake worksheet, we imported a SQL table of our data and converted it to a Pandas data frame to write the algorithm in Python.

CLEAN/FILTER DATA

All datatypes need to be converted to numbers (int, float, etc.) for the algorithm.

Also, all NA’s need to be dealt with (dropped, replaced with mean/mode).

ASSESS MODEL

Use a heatmap to show how correlated each pair of features is to eliminate unhelpful features. Use the elbow method to find optimal k value. Find silhouette score to assess the effectiveness of the clusters.

RUN THE ALGORITHM

Use Principal Component Analysis (PCA) for a fast, global overview of the data or t-Distributed Stochastic Neighbor Embedding (t-SNE) to discover finer, non-linear patterns in the data.

SELECTED ATTRIBUTES

Binary Columns (Yes or No):

1. ‘Do it yourself’
2. Avid reader
3. Internet Buyer
4. Interest in arts and craft
5. Interest in charity
6. Pet owner
7. Owns investments
8. Interest in sports
9. Interest in healthy living
10. Interest in outdoors

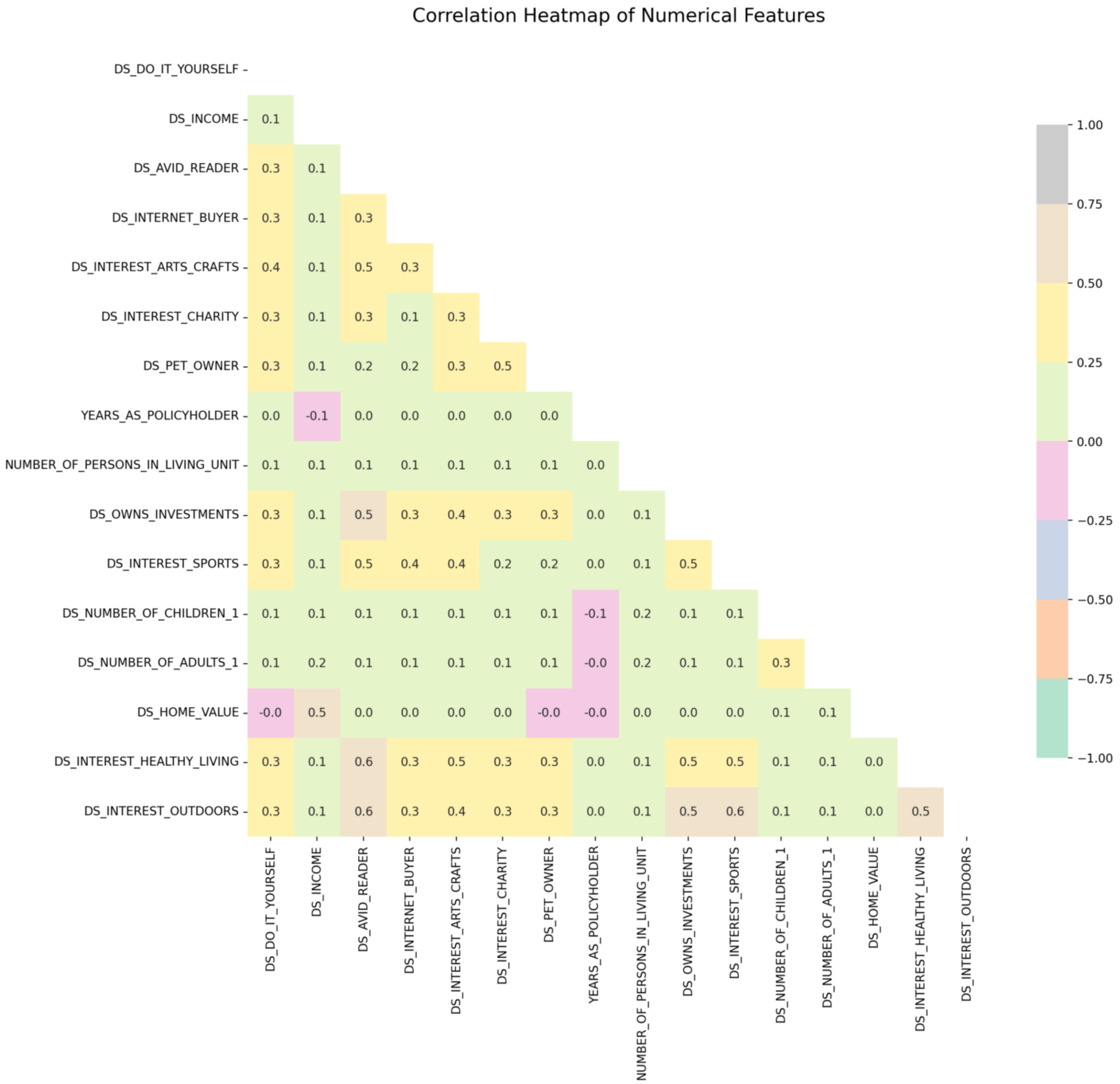
SELECTED ATTRIBUTES

Numerical:

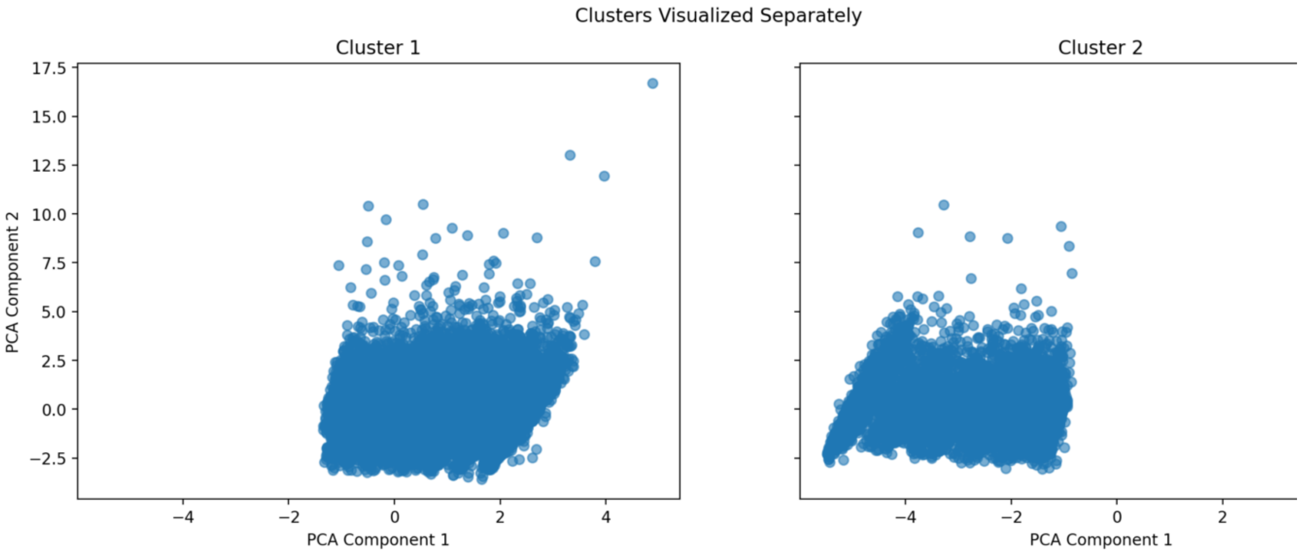
13. Income
14. Years as policy holder
15. Number of persons in living unit
16. Number of children
17. Number of adults

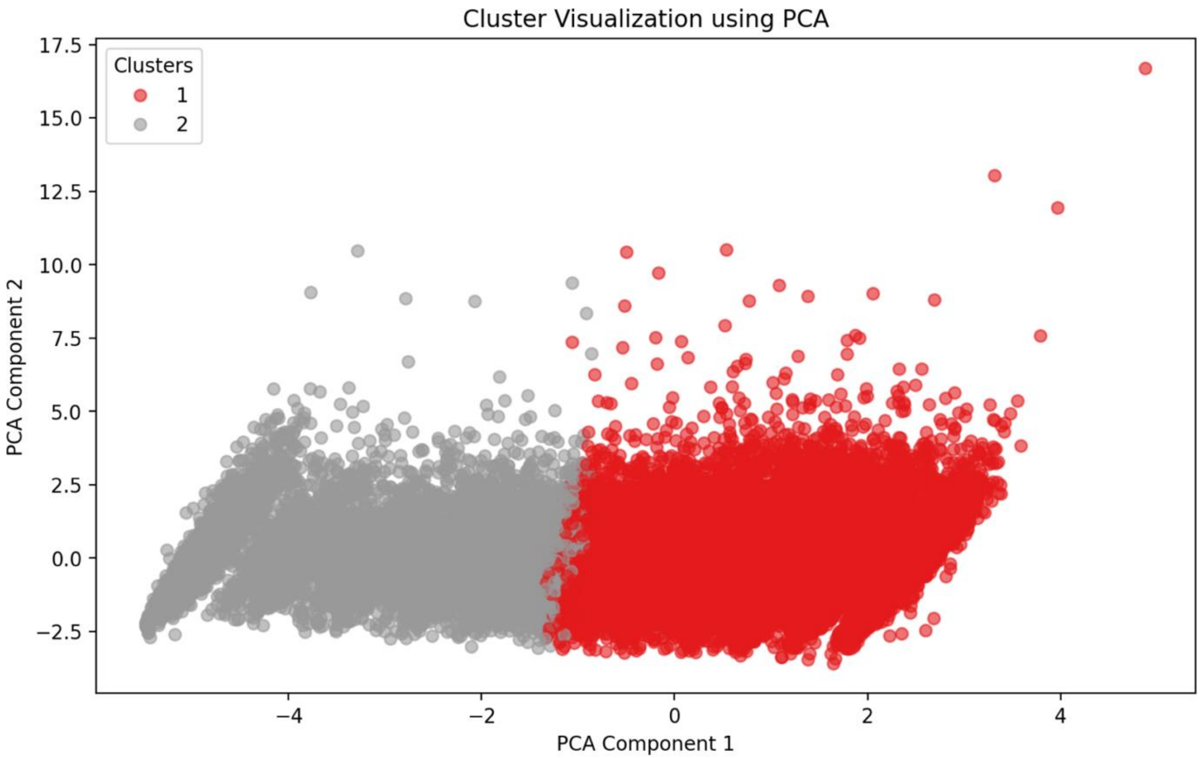
Categorical:

18. Savvy Researchers
19. Deal Seekers
20. Mobile App Users
21. Family Position
22. Highest Education
23. Carrier Route

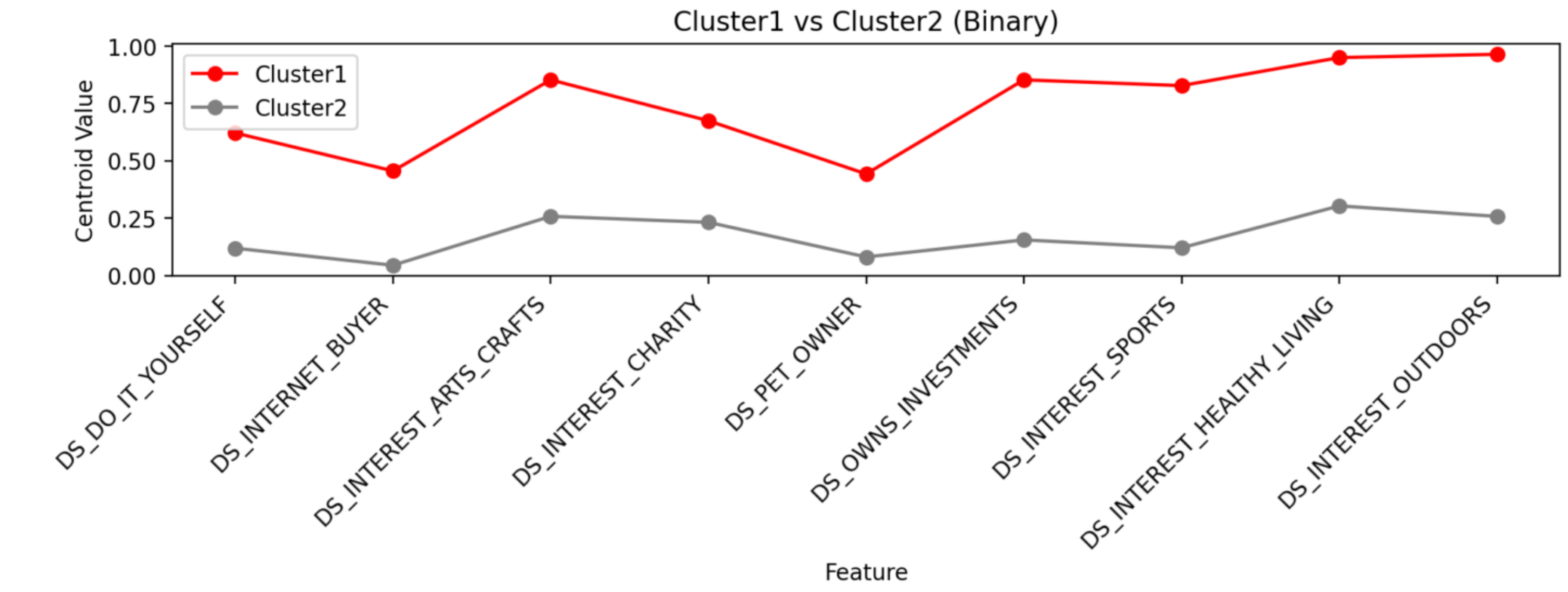

Correlation Heatmap of Numerical Features
DS_DO_IT_YOURSELF
DS_INCOME
DS_AVID_READER
DS_INTERNET_BUYER
DS_INTEREST_ARTS_CRAFTS
DS_INTEREST_CHARITY
DS_PET_OWNER
YEARS_AS_POLICYHOLDER
NUMBER_OF_PERSONS_IN_LIVING_UNIT
DS_OWNS_INVESTMENTS
DS_INTEREST_SPORTS
DS_NUMBER_OF_CHILDREN_1
DS_NUMBER_OF_ADULTS_1
DS_HOME_VALUE
DS_INTEREST_HEALTHY_LIVING
DS_INTEREST_OUTDOORS
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

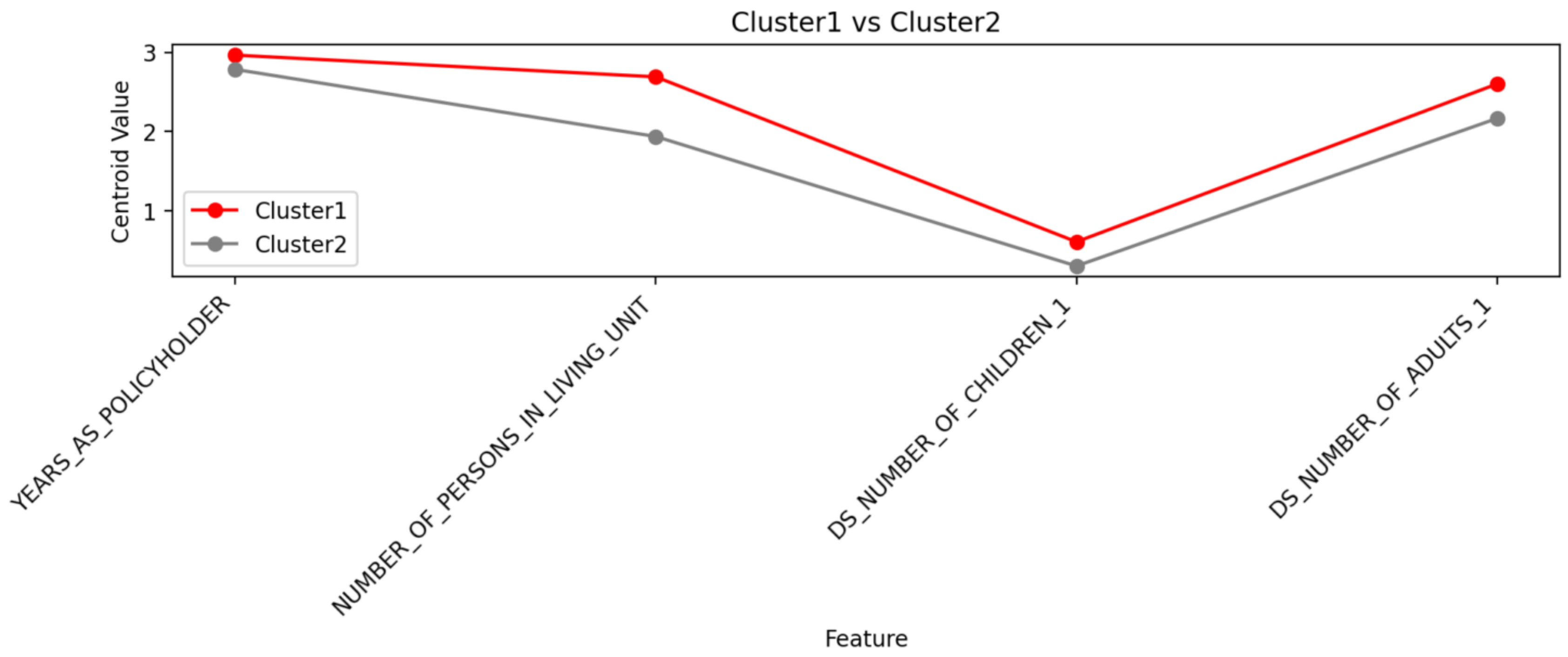
INITIAL FINDINGS

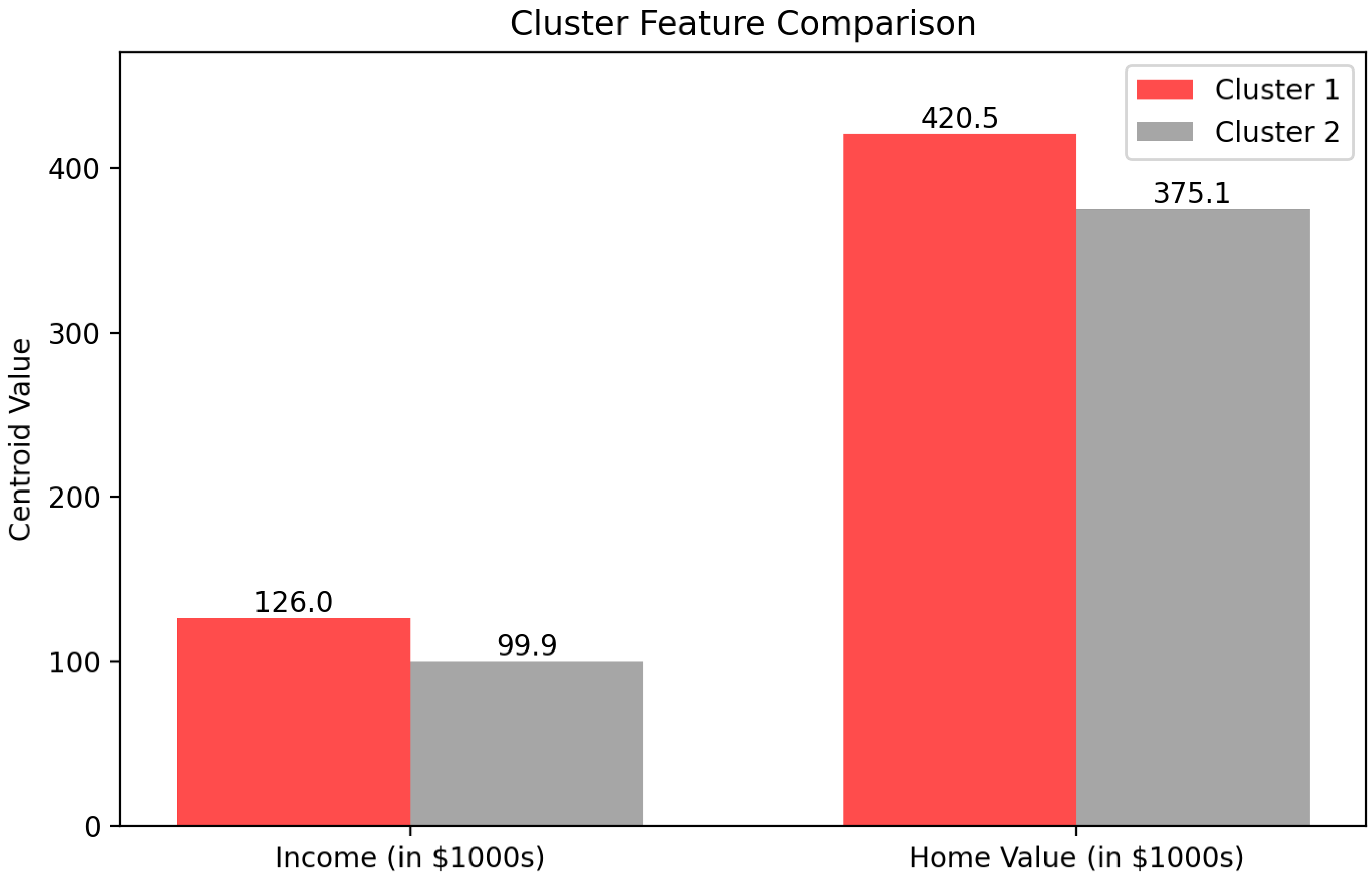




Silhouette score is only 0.2171, that is why there is overlap in the clusters – the model is not super strong


Cluster1 vs Cluster2 (Binary)
Centroid Value
1.00
0.75
0.50
0.25
0.00
Cluster1
Cluster2
DS_DO_IT_YOURSELF
DS_INTERNET_BUYER
DS_INTEREST_ARTS_CRAFTS
DS_INTEREST_CHARITY
DS_PET_OWNER
DS_OWNS_INVESTMENTS
DS_INTEREST_SPORTS
DS_INTEREST_HEALTHY_LIVING
DS_INTEREST_OUTDOORS
Feature


Cluster1 vs Cluster2
Centroid Value
3
2
1
Cluster1
Cluster2
YEARS_AS_POLICYHOLDER
NUMBER_OF_PERSONS_IN_LIVING_UNIT
DS_NUMBER_OF_CHILDREN_1
DS_NUMBER_OF_ADULTS_1
Feature


Cluster Feature Comparison
Cluster 1
Cluster 2
Centroid Value
0
100
200
300
400
126.0
99.9
420.5
375.1
Income (in $1000s)
Home Value (in $1000s)

	DS_HOME_VALUE						
Cluster	mean	std	min	25%	50%	75%	max
1	420,506.6224	326,923.9084	0	242,524	356,278	511,995	5,987,404
2	375,263.9512	350,374.0747	0	181,791	322,982	487,371.5	5,557,751

	DS_INCOME						
Cluster	mean	std	min	25%	50%	75%	max
1	126,020.1435	62,129.2412	0	91,747	117,723	152,929	1,491,421
2	99,914.2247	62,990.2855	0	55,000	99,328	131,000	770,913

Cluster 1

- Family position is Grandmother
- Highest Education is College
- Population of 25,909 people

Cluster 1 and Cluster 2

- Low Digital Display Usage
- Low Deal Seekers
- Not mobile app users
- Rural Carrier Routes

Cluster 2

- Medium Savvy Researchers
- Population of 8,383 people

SHIFTING RESEARCH EFFORTS

PHYSICAL WELLNESS

- Initial research focused on key customer dimensions and segments in relation to physical well-being

RESEARCH FINDINGS

- Recognized the significant impact of loneliness on senior mental health.

ADDRESSING LONELINESS

- Shifted focus to explore factors contributing to social connection and potential loneliness.

MEASURING SOCIAL INTERACTION

Dwelling Data Exploration

- Explored "dwelling types" as a proxy for living situation (alone vs. with family).
- Investigated pet ownership as a potential indicator of companionship.
- Examined hobbies that could include social interaction.

Questions and Challenges

- Accuracy and interpretability of "dwelling types" data:
  - Does "4+ people" indicate a senior living facility
  - Frequency of social interaction within the dwelling
  - Proximity to family and existing family relationships?
- Difficulty quantifying social interaction based on available data.
- How can we effectively measure social connection and its impact?

NEXT STEPS

Now that we've developed clusters to analyze customer behavior, product usage, and claims data, we can expand our approach to uncover meaningful segments for personalized engagement.

1. **Centroid Analysis and Refinement**
2. **Dwelling Type Analysis**
3. **DIIL vs Third Party Cluster Insights**
4. **Collaborate with KPI Group on Persistency and Claim Frequency**