Customer Segmentation for Personalized Communications

Amin Khodkar

Holland America Line

CARNIVAL CORPORATION&PLC

Data Overview

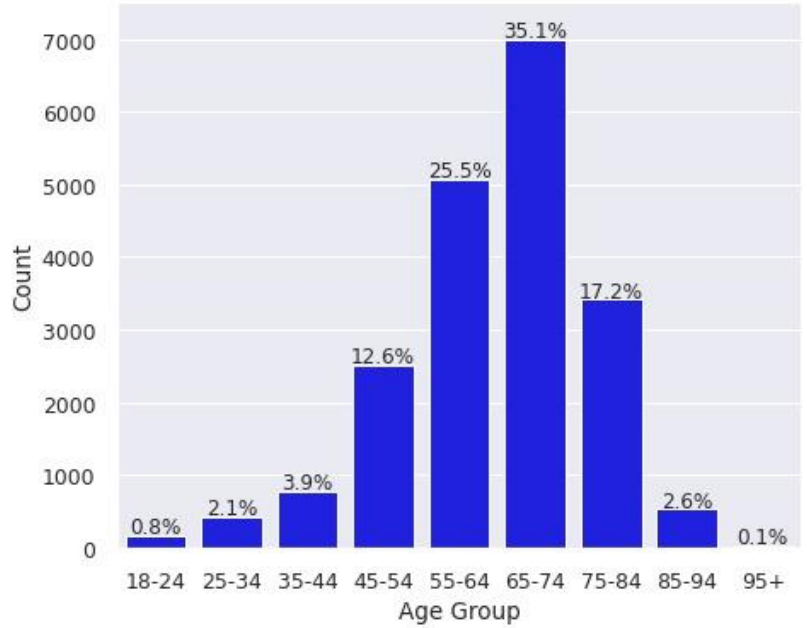
- 19925 records (bookings) and 17 features: Age, region, income, duration, subtrade, Experian mosaic group and type, etc.
- 1 datetime variable, 5 numerical variables and 11 categorical variables.
- It contains the data of a full year: December 2017 to November 2018.
- General attributes of the guests:
  - The median age of guests is about 70!
  - More than 1/3 are empty-nesters.
  - More than 1/3 come from the US West Coast.
  - More than 90% are repeating customers.
  - Only 5% of bookings are with children.
  - Prestige Positions (> 22%) and Empty-Nest Adventure (> 11%) are, respectively, the most frequent mosaic group and type in the data.

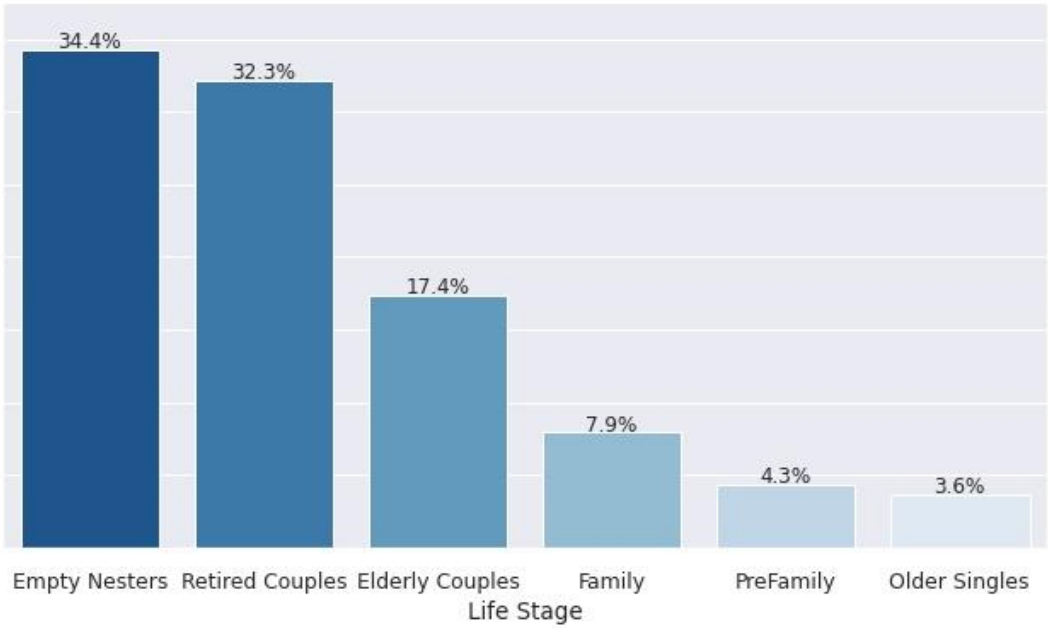
Data Preprocessing

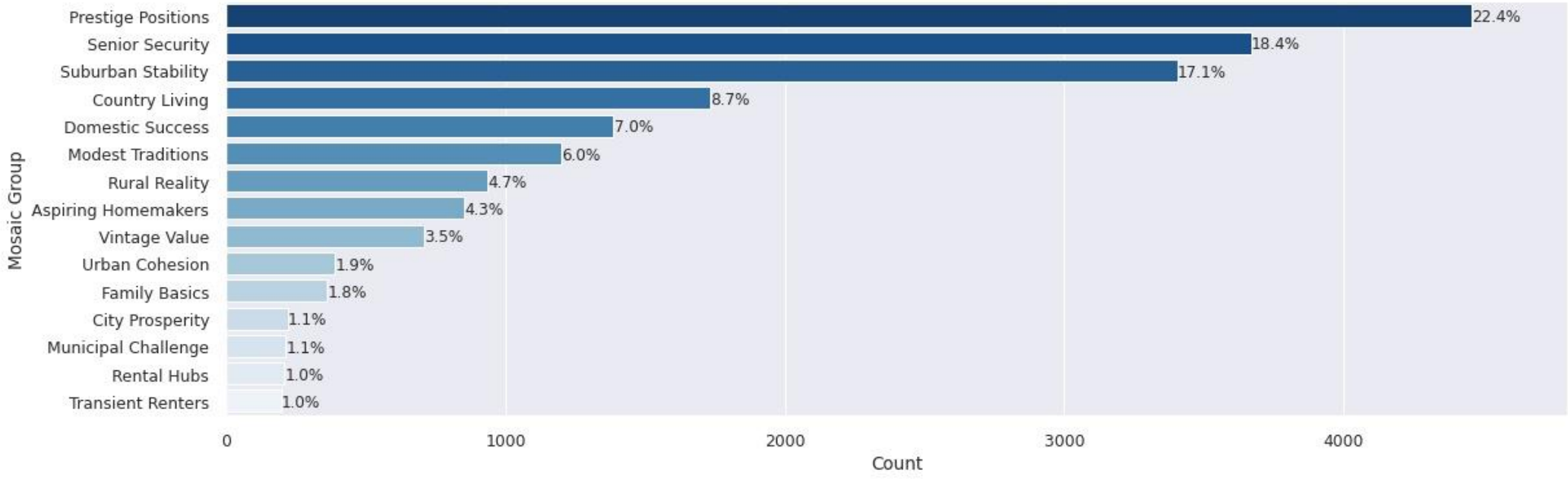
- 4 duplicated entries
- ~ 6.5% of weekly incomes were missing: The median income of the corresponding age group was chosen for imputation.
- Categorical variables with natural order were converted into numerical variables: Age group and cabin meta.
- Month was extracted as our chronical feature.
- Unclassified, staff and unknown cabins were labeled as undetermined.
- Cosmetic changes and shortening of the names of some columns and classes

EDA: Univariate Analysis

Frequency distributions of age, life stage and mosaic group

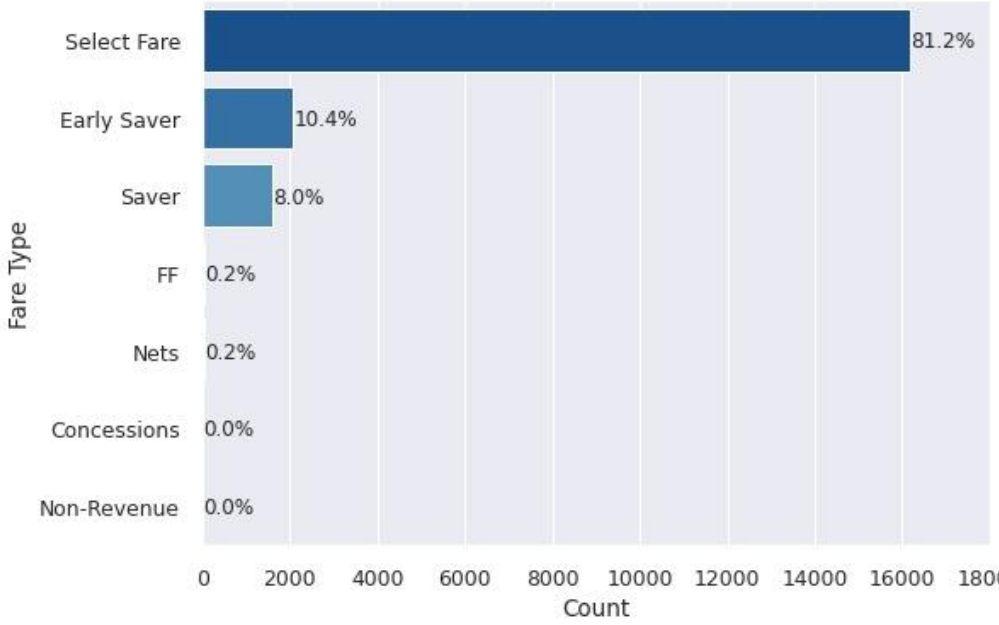


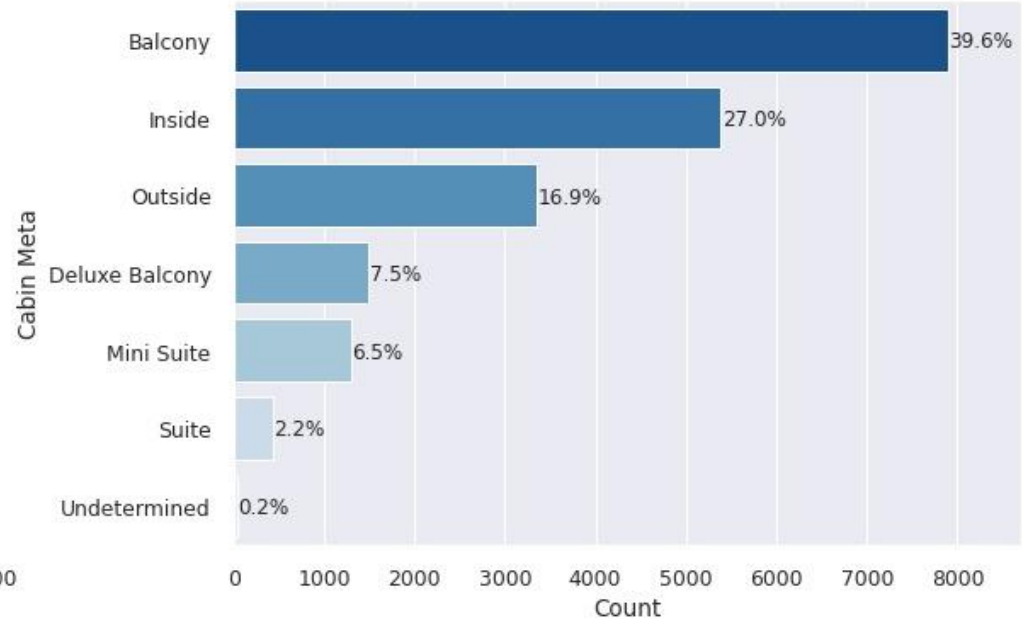


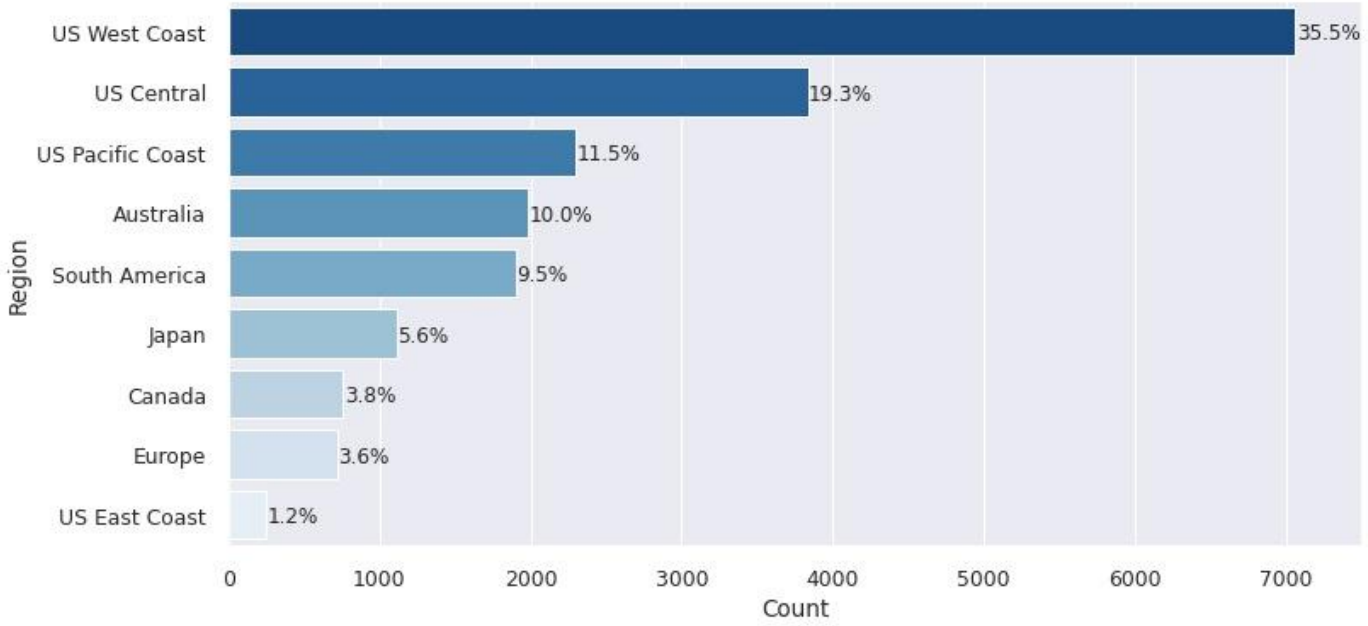


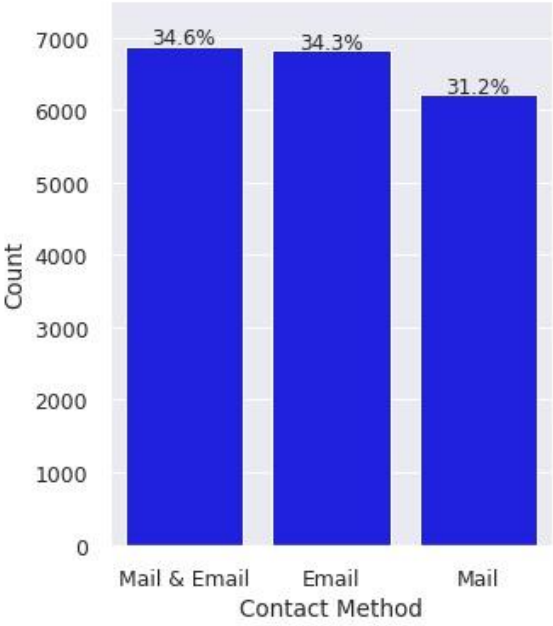
EDA: Univariate Analysis

Frequency distributions of fare type, cabin type, region and method of contact



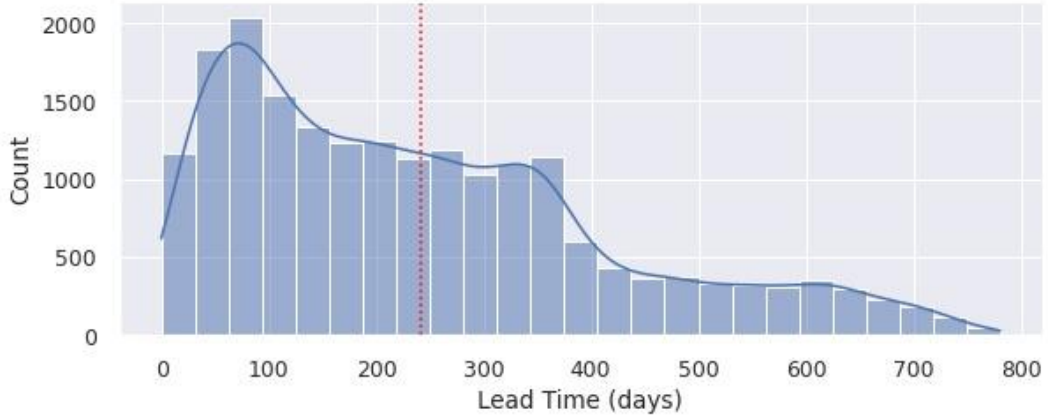


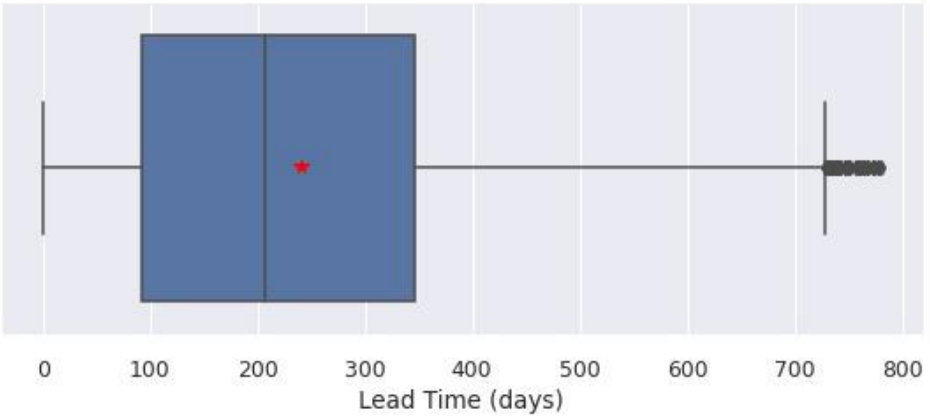


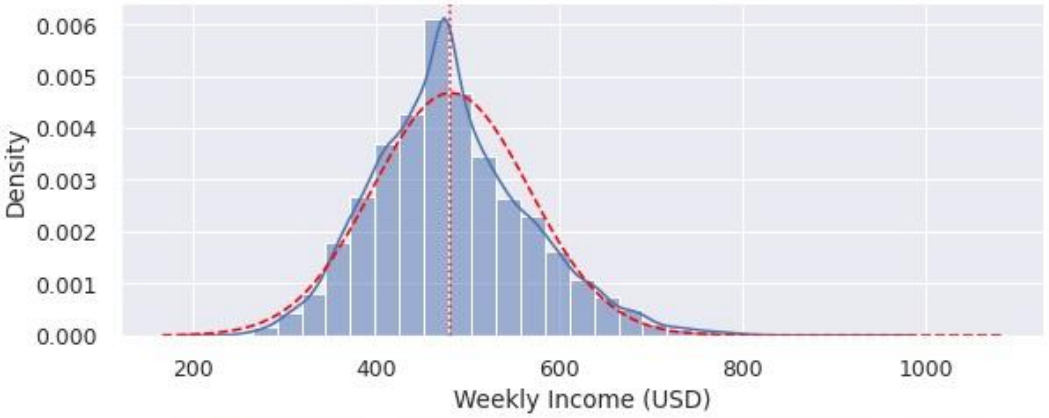


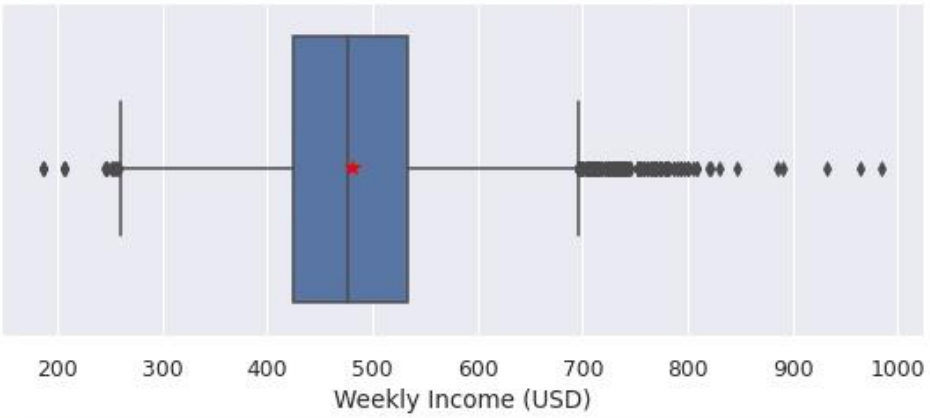
EDA: Univariate Analysis

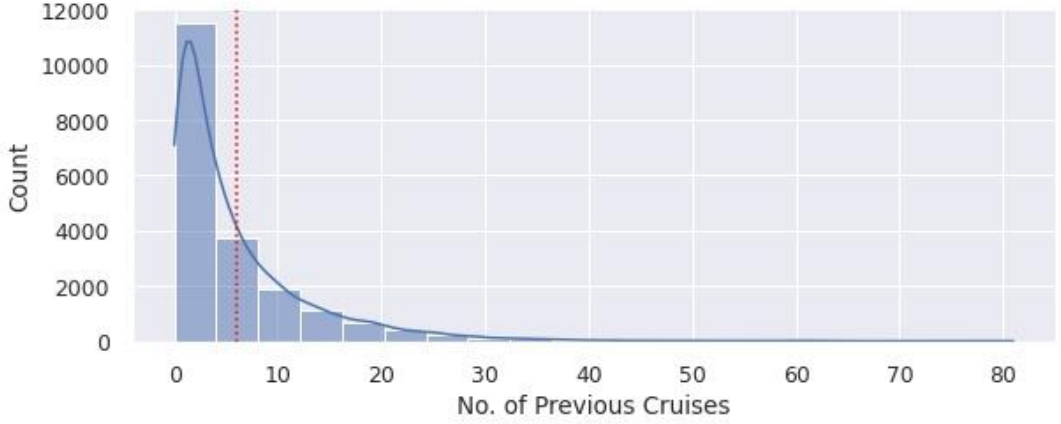
Distributions of lead time, weekly income and number of previous cruises

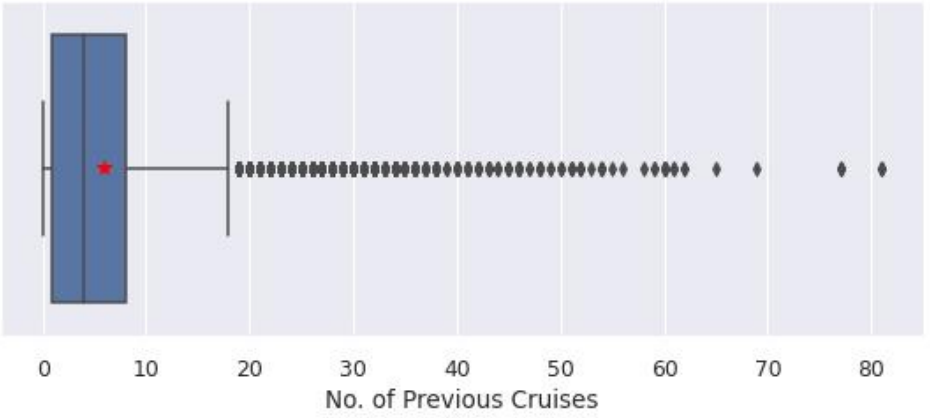






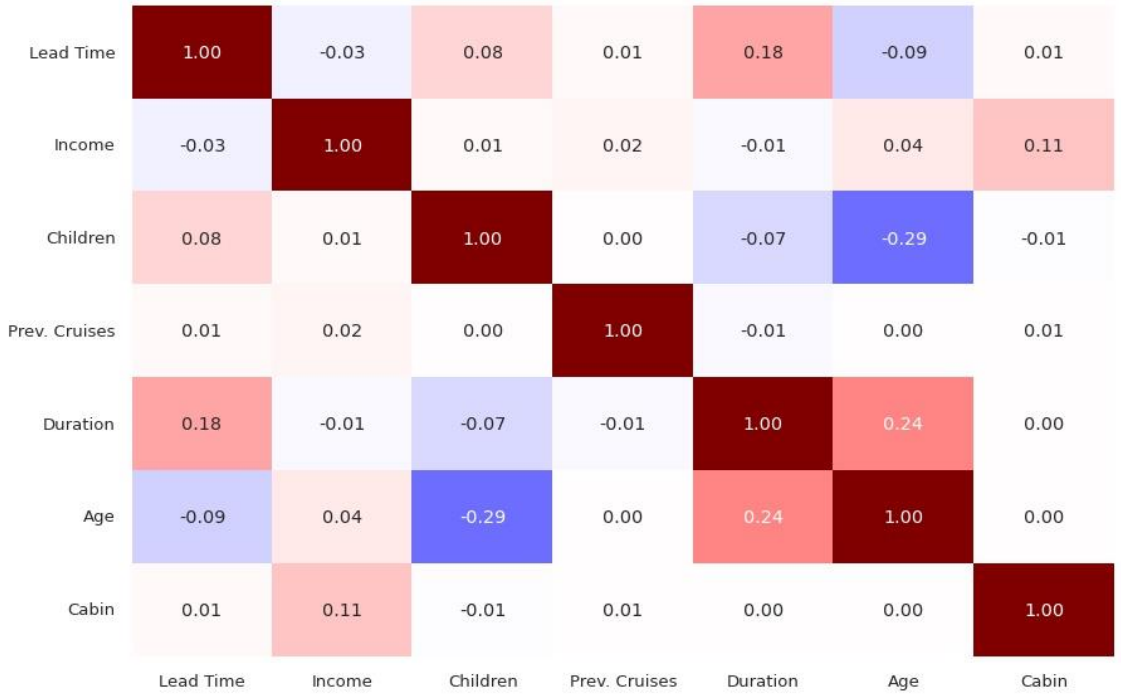




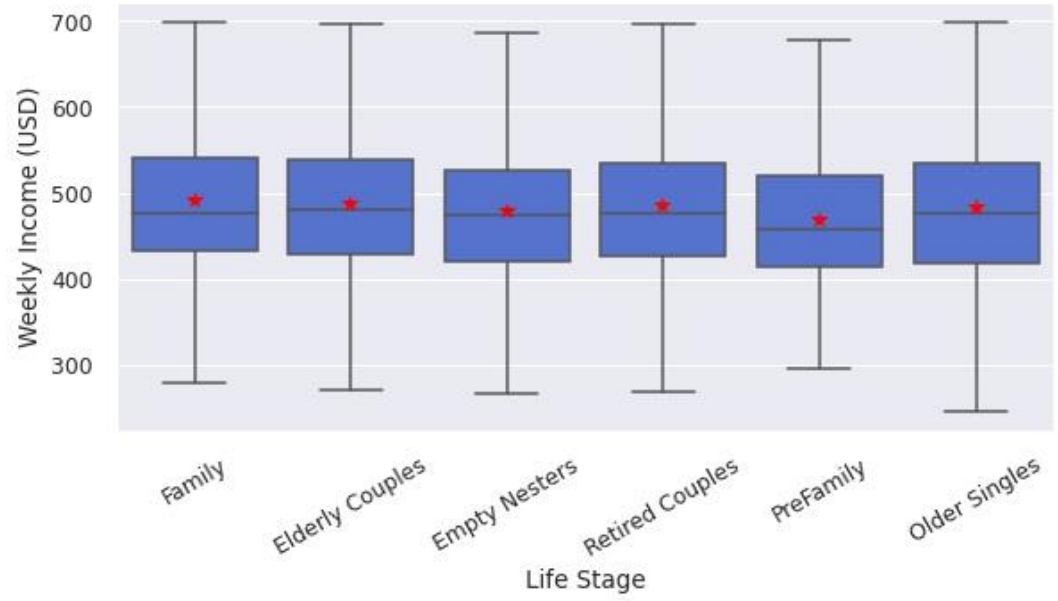


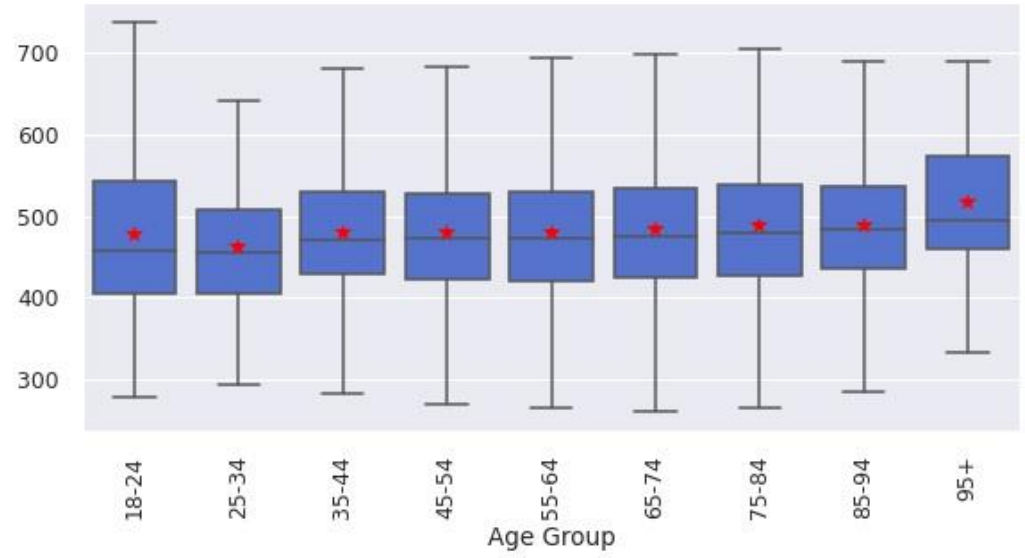
EDA: Bivariate Analysis

Correlation heatmap of numerical variables:



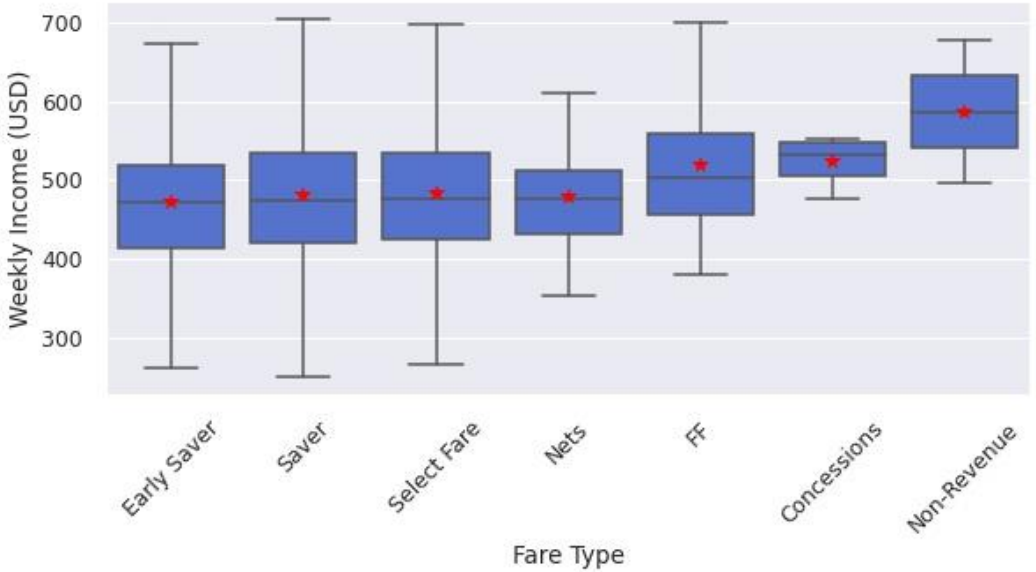
Variation of weekly income among life stages and age groups

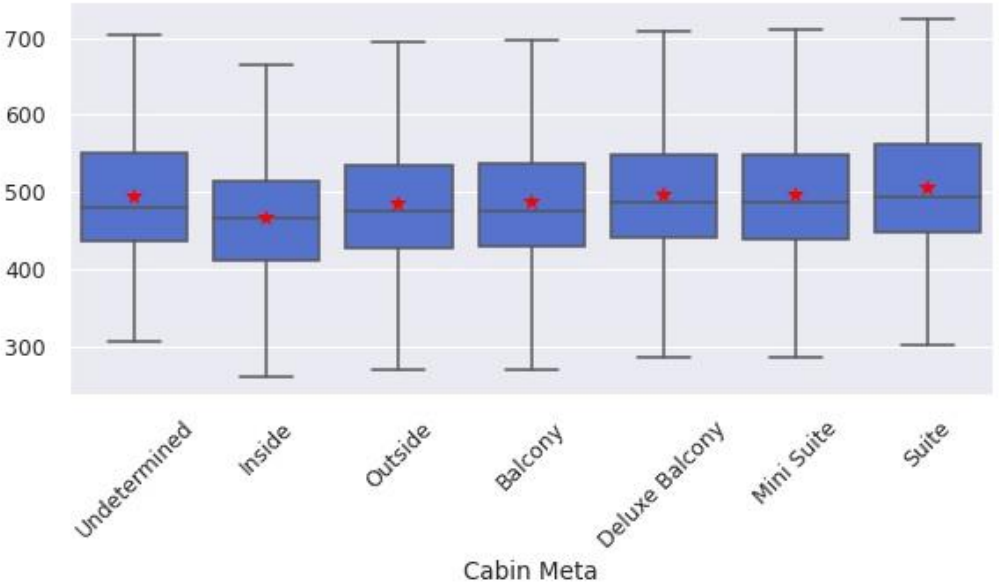




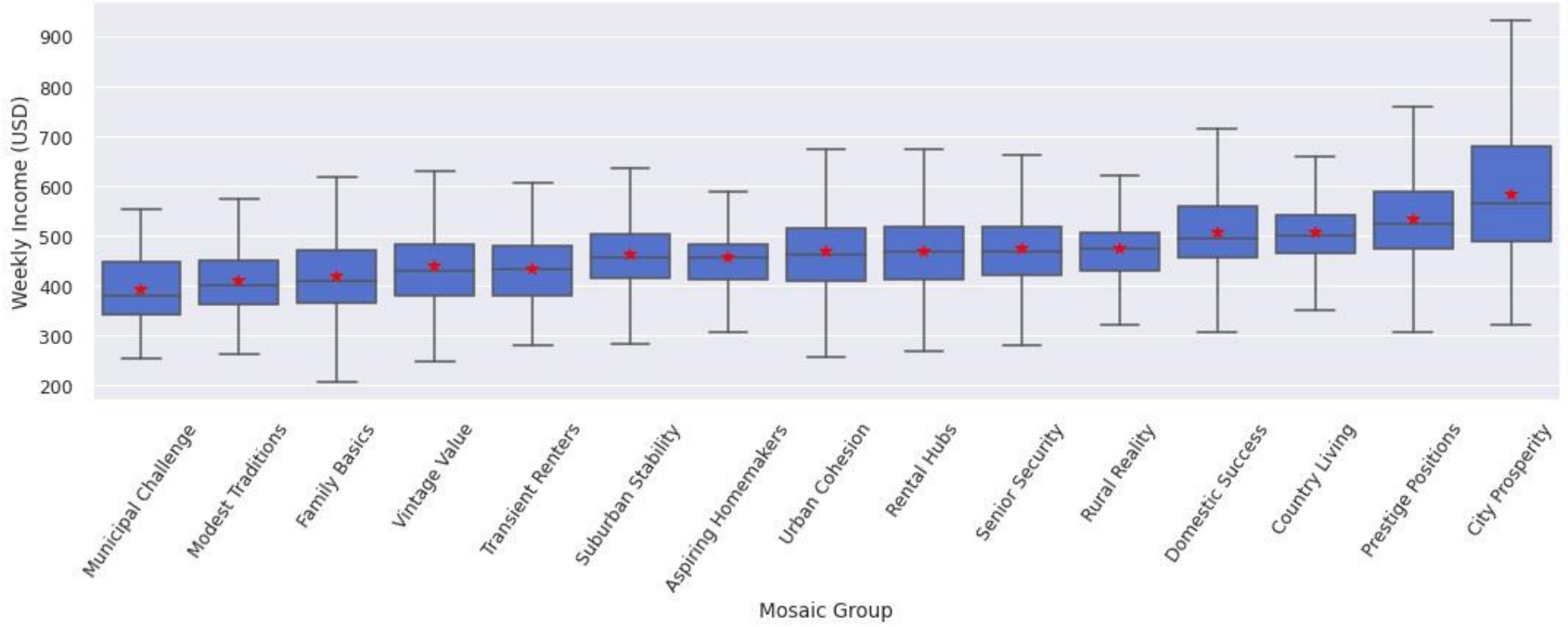
EDA: Bivariate Analysis

Variation of weekly income with fare and cabin types



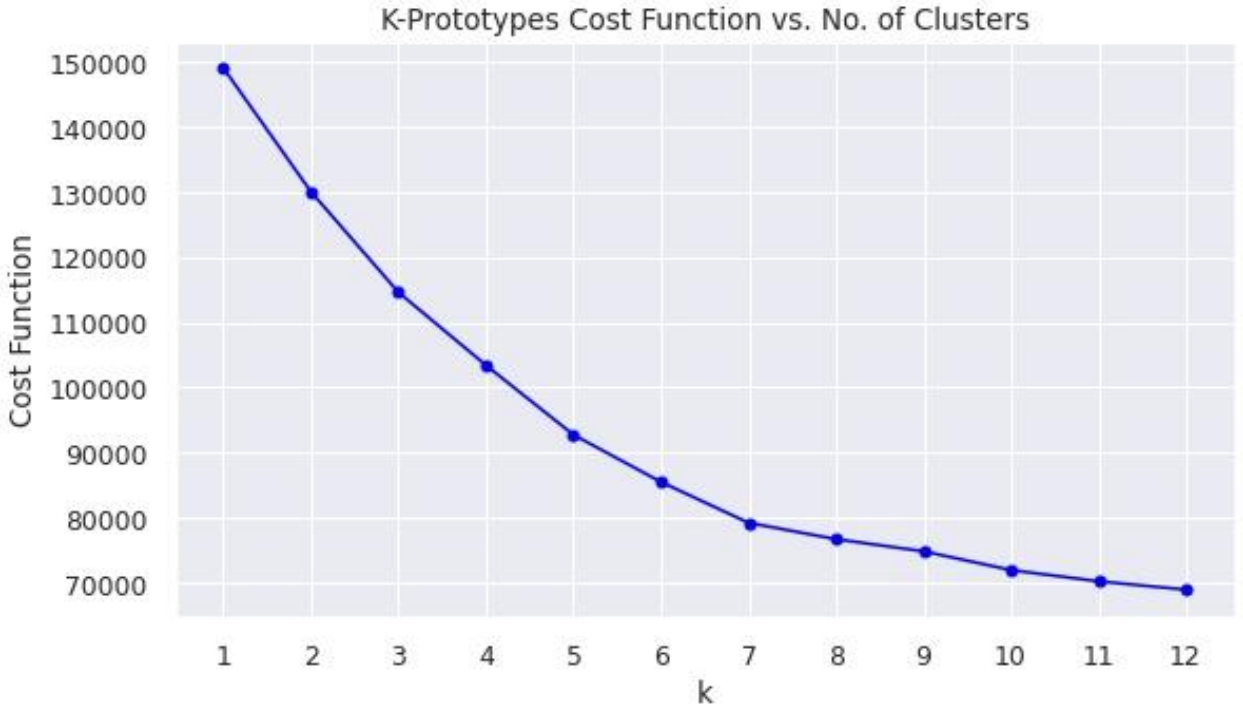


Variation of weekly income across mosaic groups



Segmentation Methodology

- Presence of too many categorical variables renders k-means ineffective.
- k-prototypes clustering, relying on different dissimilarity measures for numerical and categorical variables, proves far more suitable for the current data.
- Documentation: https://kprototypes.readthedocs.io/en/latest/api.html



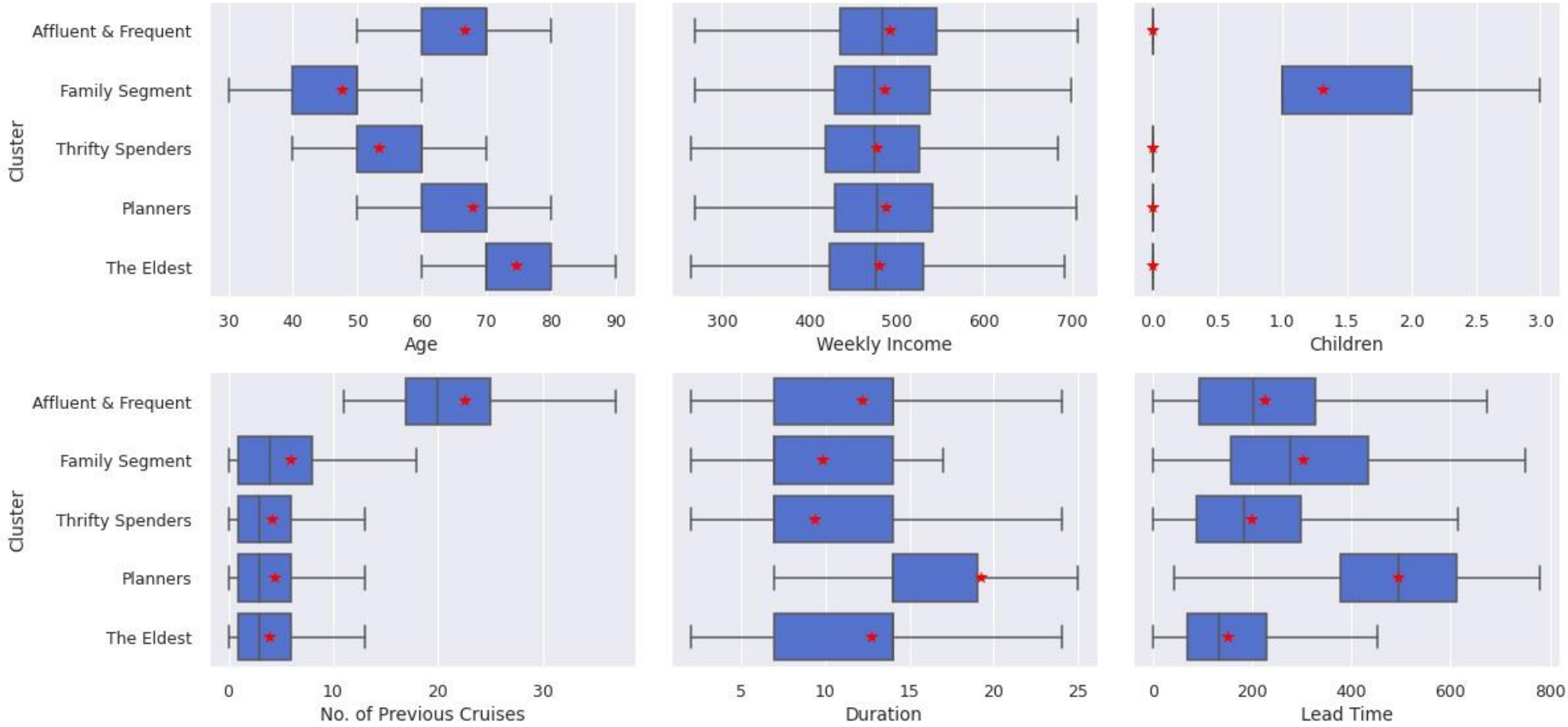
Segmentation Results

- $k = 1$: Affluent and frequent
- $k = 2$: Family segment
- $k = 3$: Thrifty spenders
- $k = 4$: Planners
- $k = 5$: The eldest

Segment	Age	Weekly Income	Children	# Previous Cruises	Duration	Lead Time	Percentage
1	66.8	493	0	22.6	12.2	226	9.78
2	47.6	486	1.32	5.97	9.85	305	4.78
3	53.5	477	0	4.26	9.41	200	31.2
4	67.9	488	0	4.48	19.2	494	17.5
5	74.6	481	0	3.96	12.8	153	36.7

Segmentation Results

Distribution of numerical features w.r.t. clusters

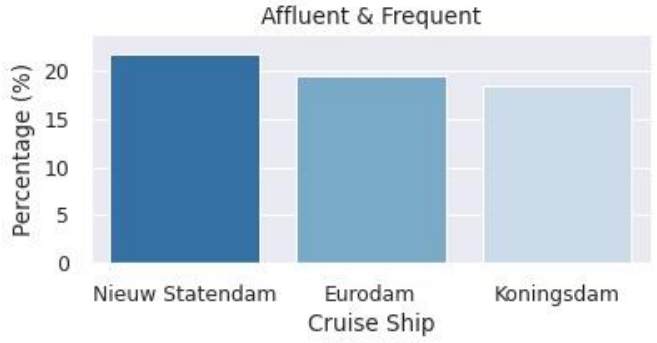


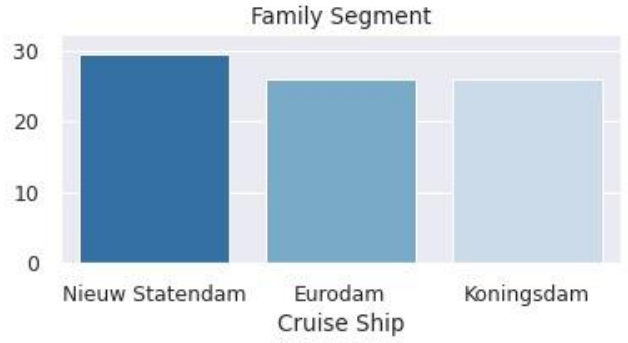
Segmentation: Results

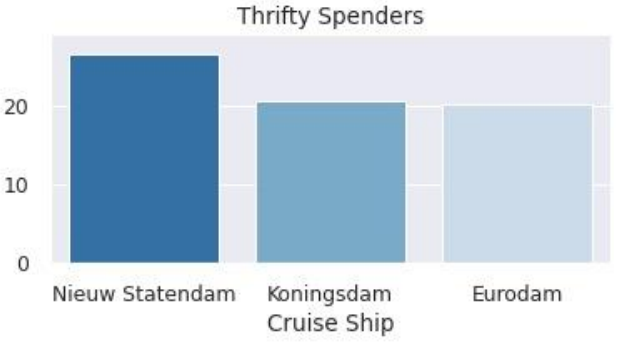
Categorical centroids of various segments

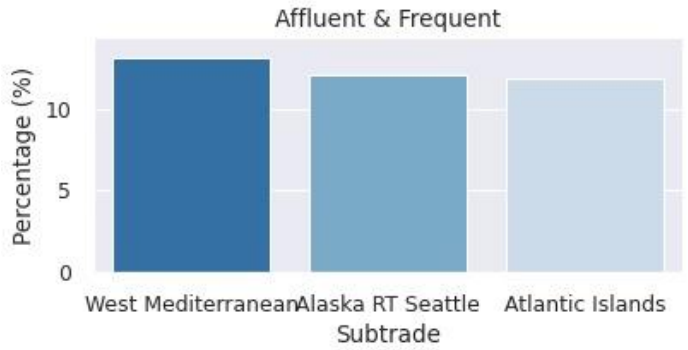
Segment	Mosaic Group	Mosaic Type	Region	Life Stage
1	Prestige Positions	Empty-Nest Adventure	US West Coast	Retired Couples
2	Domestic Success	Mid-Career Convention	US West Coast	Family
3	Suburban Stability	Fledgling Free	US West Coast	Empty Nesters
4	Prestige Positions	Empty-Nest Adventure	US West Coast	Retired Couples
5	Senior Security	Classic Grandparents	US West Coast	Retired Couples

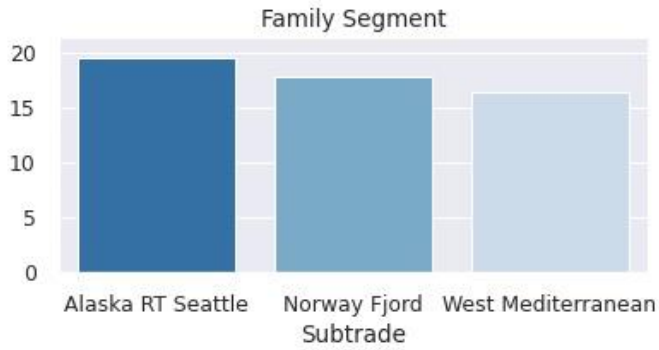
Segmentation Results: Most Frequent Classes of Each Segment

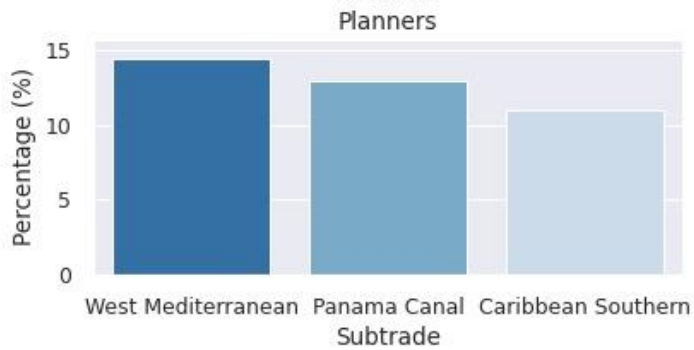


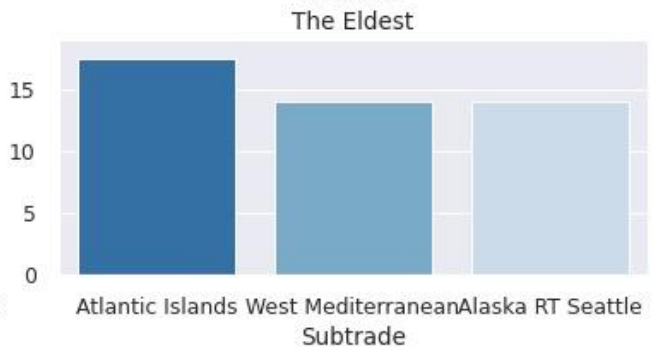




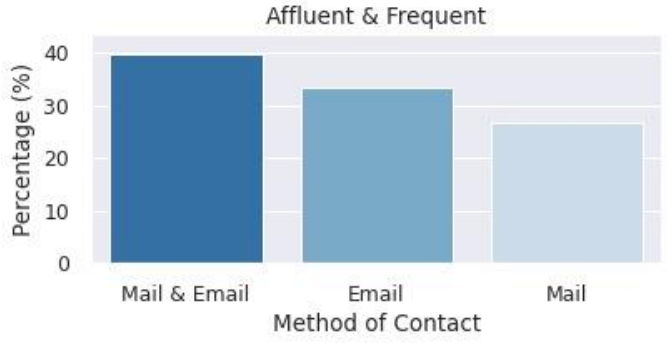


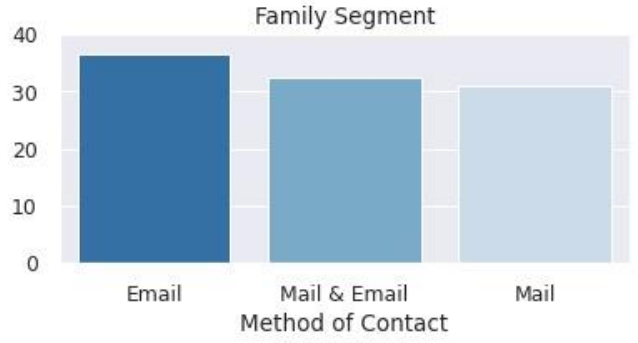


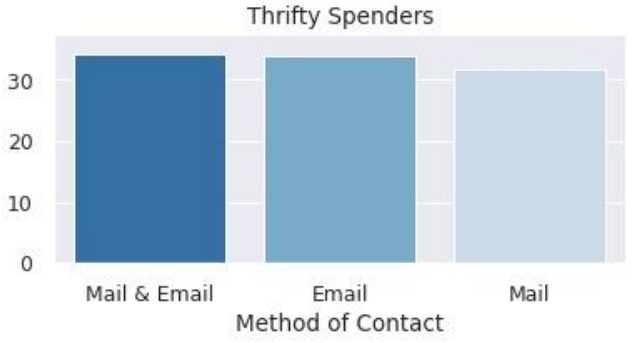


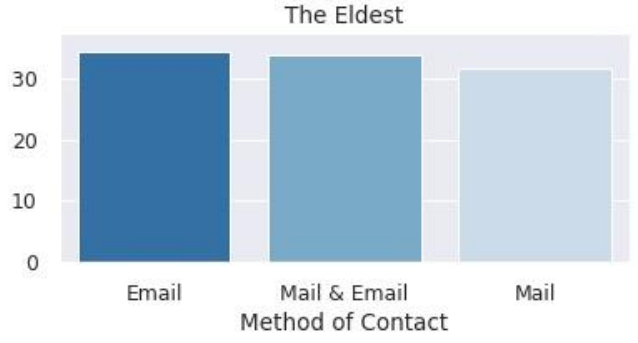


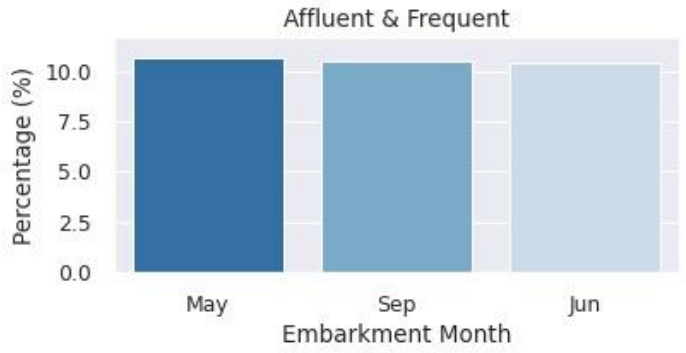
Segmentation Results: Most Frequent Classes of Each Segment

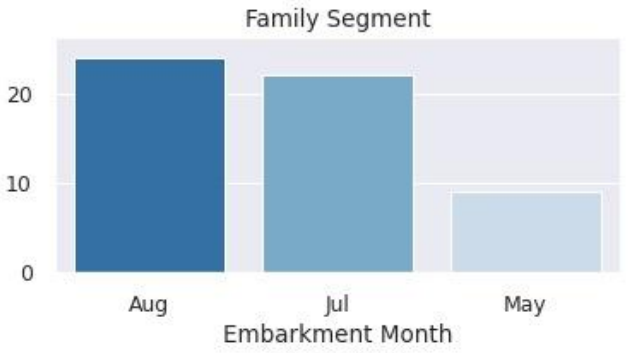


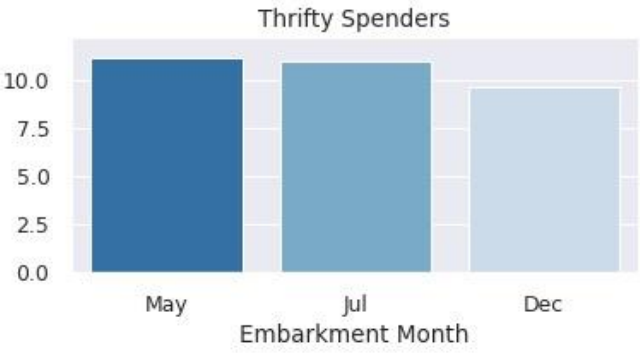


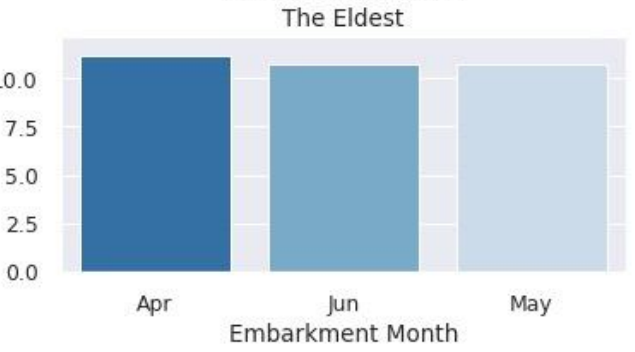












Business Insights

- The business appear to be mostly fueled by socially prestigious and financially secure elderlies with no child around.
- Information revealed by each segment can be employed to approach guests at the right time, using the right method, and to anticipate the duration, destination and other aspects of their trips.
- Test & learn and hypothesis testing should be conducted on each segment separately, as various changes/messages would alter the engagement/conversion rates of each segment differently.
- A well-designed message should account for the average age group, income, desired travel month, method of contact, etc. of the segment a certain guest falls in.
- The consistency between the Experian mosaic-based grouping and the present segmentation (based on attributes such as age, income and life stage) is interesting.