# Boats: A segmentation case

## MLO Group Assignment

Group 19: Mingxiang Xia, Filippo Schieratti, Rohan Ahuja, Tom Benabou, Amrithaa Yelagani

# Problem definition

## 01 What is the problem that Mary is trying to address?

In order to reverse the crisis of CreeqBoat company and realize the growth, Mary and her team intend to understand consumers' habits to complete market segmentation, so as to be able to provide tailored marketing strategies.

## 02 Why is she using PCA, clustering and classification? Why in this order?

**PCA**: Q1-29 reflect consumers' attitudes towards boating. There are 29 features and PCA can reduce the number of features without losing too much information.

**Clustering**: We cluster those customers sharing similar needs and demographic information together

.

**Classification**: In order to set tailored marketing strategies, we intend to understand the brand equity that customers value. In the classification, the "features" would be brand equities and the "label" is their final decision (buy or not).

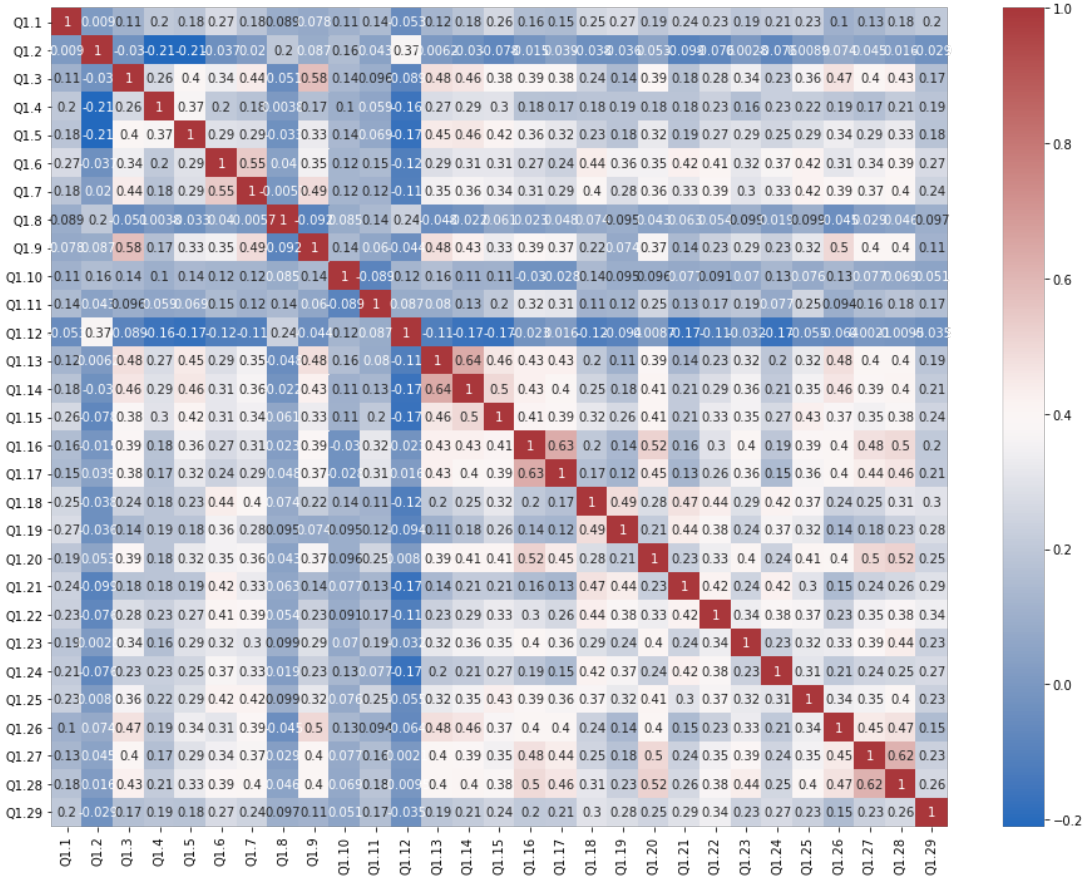
## 03 Which questions from the survey is she using as input data for each method?

**PCA**: Q1, PCA method only to capture maximum information from the numerical feautures and reduce dimensions

**Clustering**: Q2-15, no labels available

**Classification**: Q16-18. We apply Q16 as features, Q17 or Q18 as labels

# Correlation Matrix




## PART B – QUESTION 1

- Are the answers to these questions correlated?
- Do the correlated questions make sense if you look at the questions in Appendix 1?
- Explain why this could be useful in a marketing survey.

In order to find highly correlated questions, we used **correlation > 0.62**, and found the following: **{ Q13 & Q14 },{ Q16 & Q17 }**

**Q13** says that when buying a boat, I tend to buy the latest and greatest

**Q14** says that when buying accessories for my boat, I tend to buy the latest and greatest

These two questions are highly related because they overlap in their scope.

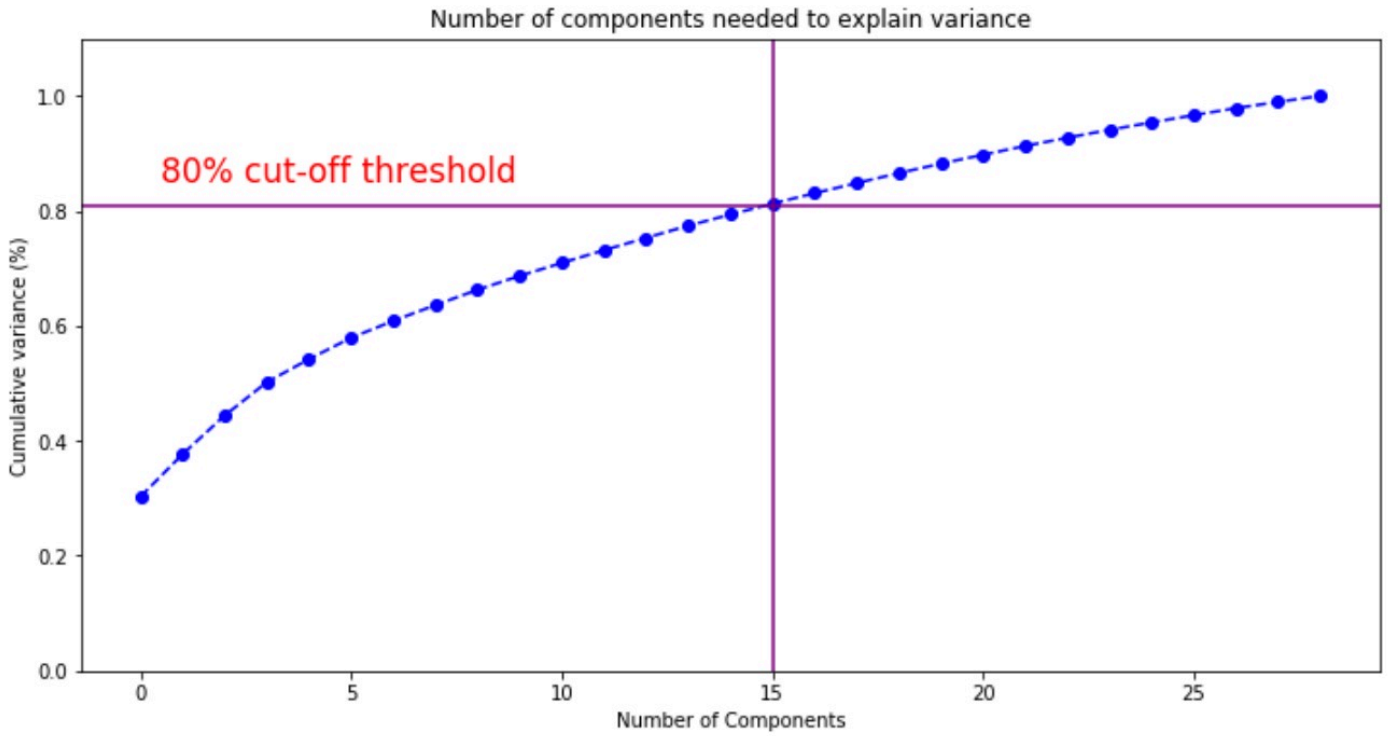
Correlation analysis method is very important in the marketing survey as:

- It helps define the overlapped and redundant questions, this helps retain the most effective questions and condense the length of the survey.
- To check the reliability of the survey responses for highly correlated questions.

```
corr_matrix = X.corr().abs()
high_corr_var=np.where(corr_matrix>0.62)
high_corr_var=[(corr_matrix.columns[x],corr_matrix.columns[y]) for x,y in zip(*high_corr_var) if x!=y and x<y]
high_corr_var
```

```
[('Q1.13', 'Q1.14'), ('Q1.16', 'Q1.17')]
```

# PCA Modelling




2. Do you need to scale the data? Why/Why not?

We use scaling only if there are varied ranges in the numerical features. For example, one feature is in floats and the other feature is in 10000's.
Since all the 29 features in our dataset have range from 1-5, scaling is not required

3. Fit a PCA model to the data. How many new features would you need and why?

In order to capture at least 80% of variance or information, we picked 15 components.

# PCA Interpretation

```
sparse_pca=dcp.SparsePCA(alpha=5,n_components=5)
sparse_pca.fit(X)
data_pca = sparse_pca.fit_transform(X)
```

```
loadings=sparse_pca.components_
df=pd.DataFrame(data=[loadings[0], loadings[1], loadings[2],loadings[3],loadings[4]],
df
```

| | Q1.1 | Q1.2 | Q1.3 | Q1.4 | Q1.5 | Q1.6 | Q1.7 | Q1.8 | Q1.9 | Q1.10 | ... | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| z1 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... | -0.391 |
| z2 | 0.000000 | 0.684598 | 0.000000 | -0.089452 | -0.111346 | 0.000000 | 0.000000 | 0.198289 | 0.000000 | 0.180564 | ... | 0.000 |
| z3 | 0.230905 | 0.000000 | 0.000000 | 0.084618 | 0.016608 | 0.392626 | 0.300802 | 0.022149 | 0.000000 | 0.076787 | ... | 0.006 |
| z4 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -0.087316 | ... | 0.000 |
| z5 | 0.000000 | 0.000000 | 0.425499 | 0.089748 | 0.278080 | 0.020721 | 0.144530 | 0.000000 | 0.452439 | 0.051246 | ... | 0.000 |

5 rows × 29 columns

## 4. Interpret the loadings obtained. What does each new feature mean at a high level?

Loadings represent the coefficient values for each questions from Q1-Q29. Which implies, greater is the absolute value in every row for all the principal components ( Z1 – Z5 ), the corresponding question is most impactful and the particular feature is most relevant to the content of the question. Table below shows each components interpretation:

| Component | Coefficient | Corresponding question |
|---|---|---|
| Z1 | 0.451062 Q1.27 | Boating is the number one thing I do with my spare time |
| Z2 | 0.684598 Q1.2 | When buying a boat, getting the lowest price is more important than the boat brand |
| Z3 | 0.392626 Q1.6 | Owning a boat is a way of rewarding myself for my hard work |
| Z4 | 0.995014 Q1.11 | I tend to perform minor boat repairs and maintenance on my own |
| Z5 | 0.452439 Q1.9 | I see my boat as a status symbol |