1. **Intro**
    1. To cluster customers into segments to optimize the significance of each customer to the business. This may help modify products according to customers specific needs and help businesses address the concerns of different user cohorts
2. **Data Cleansing Approaches**
    1. Basic EDA - check for missing data (Null), data check to evaluate which columns would be valuable to include
    2. Add any additional variables that would be valuable for the model:
        1. Total ($) Spent
        2. Cleanse data grouping on user attributes: education, household size, marital status
        3. Remove unnecessary attributes such as: costcontract, ID, etc.
        4. Convert categorical variables using LabelEncoder to be consumable for K-Means
    3. Remove outliers:
        1. Remove age > 90
        2. Income > 75%
3. **Initial Findings (before training):**
    1. income and num of visit per month is negatively correlated
    2. high correlation between wine, fruit, meat
    3. low correlation of complain to any of the purchases (or income)
4. **Data Modeling Approaches**
    1. Standard Scaling
        1. K-Means is distance based algorithms that's affected by the scale of the variable
        2. If we print - the mean of the variables in the different cluster, we can see that the income range is wider than other variables - if unscaled, our cluster will be more dependent on income variable vs others.
    2. Finding the optimal K:
        1. Elbow Method / Inertia Curve: The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.
        2. As seen in the graph, the inertia curve starts to flatten out around k = 3
    3. Success Metrics
        1. Silhouette Score - measures how well the clusters are separated
        2. Inertia
5. **Findings**
    1. Cluster 0 - Low Intent
        1. Low Spending, across all income bracket
        2. Low Deals Purchases
        3. Kids

b. Cluster 1 - High Intent
   i. High Spending, High Income
   ii. Medium Deals Purchases
   iii. No Kids
   iv. Slightly Younger Demographic

c. Cluster 2 - Medium Intent
   i. Medium Spending, medium bracket
   ii. High Deals Purchases
   iii. Kids
   iv. Older Demographic
   v. Bigger household size

6. **Next Steps / Limitations:**

   a. Cluster is not as tight and separated as expected, potentially due to the similarities in a lot of the input variables that still have high correlations with income.

   b. Try out different method of unsupervised machine learning and couple this with a supervised learning to model out how to best maximize revenue / business goals.