Western Governors University

D212 - Data Mining II - Principal Component Analysis

Shane Boyer

Scenario 1

One of the cross-sell factors is customer relationship management this identifies a company's long term profitability. A understanding by customers. When a company can better understand its customer characteristics, it is better able to target products and services to customers that are most profitable to the company. Customer Retention is a key factor.

We are asked to complete business analytics report on the Churn variable. We will be investigating the backgrounds of our customers. You have been asked to use principal component analysis (PCA) to analyze customer data to identify the principal variables of your customers, ultimately allowing better business and strategic decisions.

Part I: Research Question

A. Describe the purpose of this data mining report by doing the following:

- Propose one question relevant to a real-world organizational situation that you will answer by using principal component analysis (PCA)

What factors drive customer churn in the company's customer base?

- Define one goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

One goal of this analysis is to reduce the dataset to a smaller number of features to speed up future analysis and prevent overfitting.

Part II: Method Justification

B. Explain the reasons for using PCA by doing the following:

- Explain how PCA reduces the dimensions of the original dataset.

PCA is a dimensionality reduction technique that reduces the number of features in a dataset by combining features that are highly correlated. The principal component is a matrix number of features that explain the majority of the variance in the dataset. PCA allows for the reduction of the number of features while retaining most of the variance in the dataset. The eigenvectors with the highest eigenvalues are the principal components of the dataset.

- Summarize one assumption of PCA.

PCA assumes that data variance is the same as common variance, or change between. = PCA also requires features are continuous or ordinal. PCA will not accept categorical features.

Part III: Data Preparation

C. Perform data preparation for the chosen dataset by doing the following:

- Identify the continuous variables that you will need to utilize to answer the PCA question proposed in part A.

```
continuous_cols = churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
```

- Standardize the continuous dataset variables identified in part C. Include a copy of the dataset cleaned.

```
churn_df.describe() # Check for data ranges
churn_df.info()
churn_df.dtypes
churn_df.shape
churn_df.isnull.sum()
churn_df.duplicated()
churn_df.drop_duplicate(inplace=True)
churn_df.shape
churn_df.dtypes.value_counts()
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

```
churn_df.describe()
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts()
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

D. Perform data preparation for the chosen dataset by doing the following:

- Identify the continuous variables that you will need to utilize to answer the PCA question proposed in part A.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the continuous variables that you will need to utilize to answer the PCA question proposed in part A.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the principal components to be retained in the dataset with a cumulative variance of at least 75%.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the variance of each of the principal components identified in part D2.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the variance of each of the principal components identified in part D2.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the principal components to be retained in the dataset with a cumulative variance of at least 75%.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

Sources

No external sources used

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the variance of each of the principal components identified in part D2.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the principal components to be retained in the dataset with a cumulative variance of at least 75%.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 InternetService 9464 non-null   object
 30 Phone           8989 non-null   object
 31 Multiple        10000 non-null  object
 32 OnlineSecurity  9464 non-null   object
 33 OnlineBackup    9464 non-null   object
 34 DeviceProtection  9464 non-null  object
 35 TechSupport     9464 non-null   object
 36 StreamingTV     9464 non-null   object
 37 StreamingMovies 9464 non-null   object
 38 PaperlessBilling  10000 non-null  object
 39 Tenure          9996 non-null   float64
 40 MonthlyCharge   10000 non-null  float64
 41 Bandwidth_GB_Year  10000 non-null  float64
 42 Item1           10000 non-null  int64
 43 Item2           10000 non-null  int64
 44 Item3           10000 non-null  int64
 45 Item4           10000 non-null  int64
 46 Item5           10000 non-null  int64
 47 Item6           10000 non-null  int64
 48 Item7           10000 non-null  int64
 49 Item8           10000 non-null  int64
dtypes: object(24), int64(17), float64(9)
memory usage: 3.8 MB
```

- Identify the variance of each of the principal components identified in part D2.

```
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.isnull.sum()
churn_df.dtypes
churn_df.dtypes.value_counts
churn_df.select_dtypes(include=['object', 'category', 'bool'])
churn_df.select_dtypes(exclude=['object', 'category', 'bool'])
churn_df.shape
churn_df.to_csv('churn_clean.csv')
churn_df = pd.read_csv('churn_clean.csv')
churn_df.shape
churn_df.dtypes
churn_df.head()
churn_df.tail()
churn_df.sample(5)
churn_df.nunique()
churn_df.info()
```

```
<class RangeIndex>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50):
 #  Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0  Unnamed: 0      10000 non-null  int64
 1  Customer_id     10000 non-null  object
 2  Interaction     10000 non-null  object
 3  UID             10000 non-null  int64
 4  City            10000 non-null  object
 5  State           10000 non-null  object
 6  County          10000 non-null  int64
 7  Zip             10000 non-null  int64
 8  Lat             10000 non-null  float64
 9  Lng             10000 non-null  float64
 10 Population      10000 non-null  int64
 11 Area            10000 non-null  object
 12 Time_zone       10000 non-null  object
 13 Job             10000 non-null  object
 14 Children        7412 non-null   float64
 15 Age             7414 non-null   float64
 16 Education       10000 non-null  object
 17 Employment      10000 non-null  object
 18 Marital         10000 non-null  object
 19 Gender          10000 non-null  object
 20 Churn           10000 non-null  object
 21 Outage_sec_perweek  10000 non-null  float64
 22 Email           10000 non-null  int64
 23 Contacts        10000 non-null  int64
 24 Yearly_equip_failure  10000 non-null  int64
 25 Techie          7763 non-null   object
 26 Contract        10000 non-null  object
 27 Port_modem      10000 non-null  object
 28 Tablet          10000 non-null  object
 29 Intern
```

```
>>> 1
```

	Unique_id	Applicant_id	Investment_ad	Age	Income	Savings_rate_personal	Small	Customer	Early_stage_Active	Membership
0	1	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
1	2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
2	3	0.000000	0.861000	0.571000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
3	4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
4	5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
5	6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
6	7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000

```
>>> customer_df.to_csv("customer_clean.csv")
```

```
>>> numeric_cols = [
    'Age', 'Income', 'Savings_rate_personal',
    'Small', 'Customer', 'Early_stage_Active', 'Membership']

>>> customer_df[numeric_cols].describe()
```

```
       Age  Income  Savings_rate_personal       Small  Customer  Early_stage_Active  Membership
count  0.0     0.0                    0.0  0.000000  0.000000                0.000       0.000
mean   NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
std    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
min    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
25%    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
50%    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
75%    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
max    NaN     NaN                    NaN  0.000000  0.000000                0.000       0.000
```

```
>>> from sklearn.decomposition import PCA
>>> from sklearn.preprocessing import StandardScaler
>>> numeric = customer_df[numeric_cols].dropna()
>>> scaled = StandardScaler().fit_transform(numeric)
>>> pca = PCA(n_components=2)
>>> pc = pca.fit_transform(scaled)
>>> print("Explained variance:", pca.explained_variance_ratio_)
Explained variance: [nan nan]
>>> import matplotlib.pyplot as plt
>>> plt.scatter(pc[:, 0], pc[:, 1], alpha=0.5)
>>> plt.xlabel("PC 1 (28.2%)")
>>> plt.ylabel("PC 2 (22.1%)")
>>> plt.title("Customer Segmentation — PCA")
>>> plt.show()
```



```
>>> import matplotlib.pyplot as plt, seaborn
>>> sns.scatterplot(x=pc[:, 0], y=pc[:, 1], hue=customer_df.loc[numeric.index, "Income"])
>>> plt.show()
```



```
>>> import matplotlib.pyplot as plt, seaborn
>>> sns.scatterplot(x=pc[:, 0], y=pc[:, 1], hue=customer_df.loc[numeric.index, "Income"])
>>> plt.show()
```

```
/Library/Java/JavaVirtualMachines/jdk-17.0.2.jdk/Contents/Home/bin/java -Dfile.encoding=UTF-8 -classpath /Users/ryanmcdonough/IdeaProjects/untitled/out/production/untitled com.ryanmcdonough.untitled.Main
```

- To summarize the results of your data analysis:

The analysis shows that we can generally group the customers in this dataset with 2 principal components. The top 4 principal components explain 100% of the variance in the data. Those statistics are the first and second principal components with component 1 being variance and component variance is quick with further analysis. PCA structure's only resulted in remaining 1/2nd of the components in a dataset. Principal components are a linear combination of the original variables in the dataset. The principal components are chosen to have maximum variance and distribution of components, overlapping in a clear method the PC's in a good choice for this dataset and manual dimensionality reduction through mega. Items in PCA are previously used on this an would be better suited.

Sources

No external sources used