Report

Principal Component Analysis

The data we have was high dimensional data with multiple columns which led to difficulty in organizing and analyzing the various components. We thus used PCA to reduce the dimensionalty of our data.

First step was to clean and filter out the data which was not required. We removed the categorical variables as they were increasing the complexity of the data. Further, we removed the columns with zero variance and the rows with missing values. Following is the head of the cleaned data:

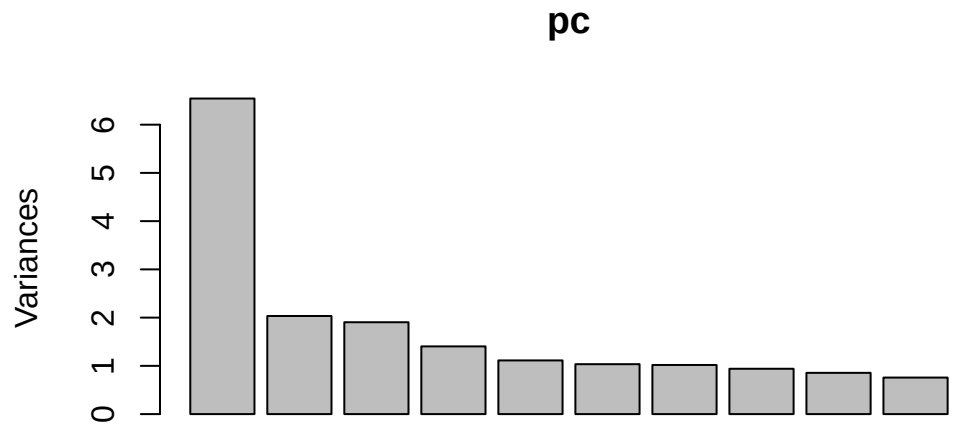
```
# A tibble: 5 x 24
     ID Year_Birth Income Kidhome Teenhome Recency MntWines MntFruits
  <dbl>      <dbl>  <dbl>   <dbl>    <dbl>   <dbl>    <dbl>     <dbl>
1  5524       1957  58138       0        0      58      635        88
2  2174       1954  46344       1        1      38       11         1
3  4141       1965  71613       0        0      26      426        49
4  6182       1984  26646       1        0      26       11         4
5  5324       1981  58293       1        0      94      173        43
# i 16 more variables: MntMeatProducts <dbl>, MntFishProducts <dbl>,
#   MntSweetProducts <dbl>, MntGoldProds <dbl>, NumDealsPurchases <dbl>,
#   NumWebPurchases <dbl>, NumCatalogPurchases <dbl>, NumStorePurchases <dbl>,
#   NumWebVisitsMonth <dbl>, AcceptedCmp3 <dbl>, AcceptedCmp4 <dbl>,
#   AcceptedCmp5 <dbl>, AcceptedCmp1 <dbl>, AcceptedCmp2 <dbl>, Complain <dbl>,
#   Response <dbl>
```

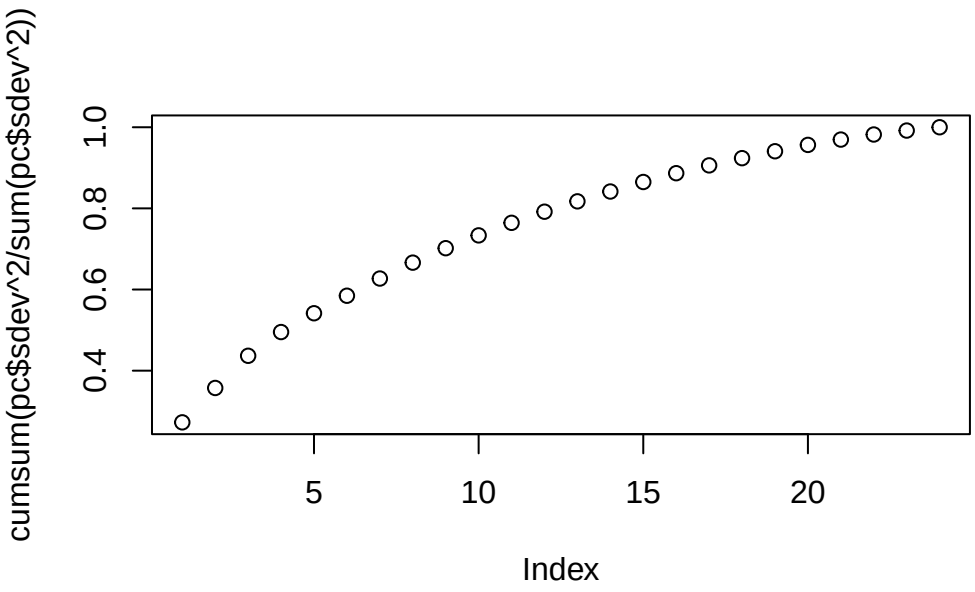
As we can see, there are 24 columns even after cleaning the data, which shows the high dimensionalty.

Then, we performed PCA which gave the following result:

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5578 1.42624 1.38007 1.18466 1.05487 1.01737 1.00901
Proportion of Variance 0.2726 0.08476 0.07936 0.05848 0.04636 0.04313 0.04242
Cumulative Proportion  0.2726 0.35735 0.43670 0.49518 0.54154 0.58467 0.62709
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.96948 0.92476 0.87041 0.86141 0.81159 0.78256 0.76318
Proportion of Variance 0.03916 0.03563 0.03157 0.03092 0.02744 0.02552 0.02427
Cumulative Proportion  0.66625 0.70189 0.73345 0.76437 0.79182 0.81733 0.84160
                          PC15    PC16    PC17   PC18    PC19    PC20   PC21
Standard deviation     0.75005 0.72085 0.68036 0.6537 0.63905 0.61263 0.5629
Proportion of Variance 0.02344 0.02165 0.01929 0.0178 0.01702 0.01564 0.0132
Cumulative Proportion  0.86504 0.88669 0.90598 0.9238 0.94080 0.95644 0.9696
                          PC22    PC23    PC24
Standard deviation     0.54660 0.48427 0.44205
Proportion of Variance 0.01245 0.00977 0.00814
Cumulative Proportion  0.98209 0.99186 1.00000
```

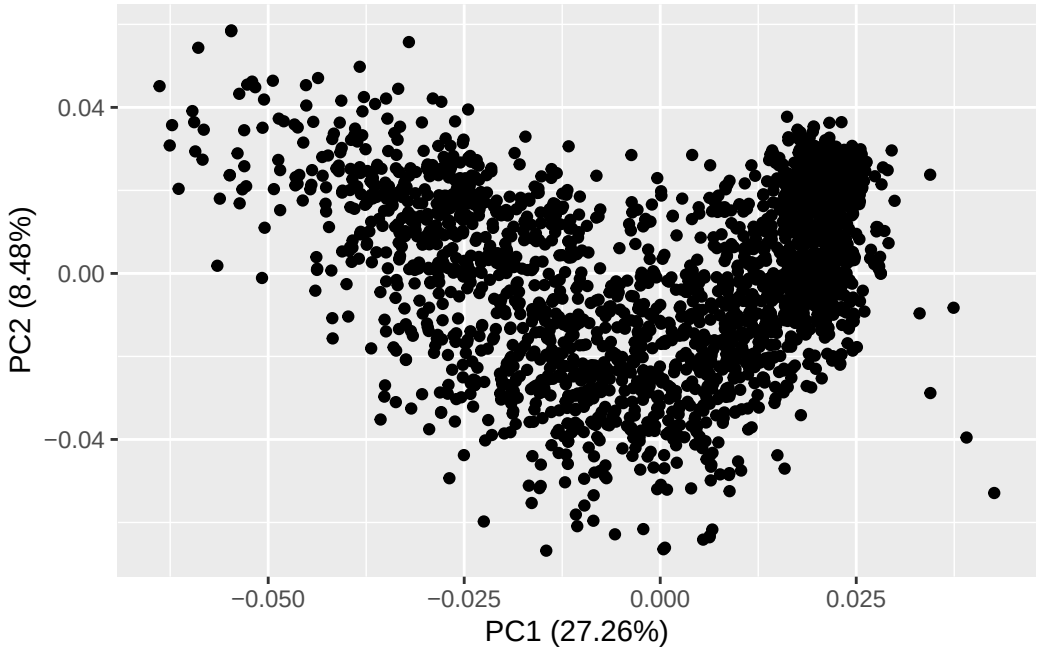
Below is the plot of the PCA analysis and the cumulative variance of the components:





Thus, around 13 components are able to explain 80% variability in the data.

Below is a plot of the relationship between the first two components after PCA.



Thus, with the help of PCA we can reduce the data with 24 columns to upto 13 columns and still explain 80% variability in the data.