### **Project Topic: Real-time analysis of online shoppers’ purchasing intention using Dimension Reduction And Clustering**

### Group Members: Tianxiang Chen, Xuewei Zhong

1. **Introduction**

This research paper aims to analyze real-time online shopper behavior by using a mix of dimensionality reduction and clustering techniques. We work with a dataset that includes information about shopper interactions, demographics, and transactions. The first step is to simplify the complex data structure with methods like Principal Component Analysis (PCA) and factor analysis. After reducing the dimensions of the data, we will use clustering techniques -–– K-means to sort the online shoppers into two main groups: those who generate revenue and those who do not. The study will focus on finding clear behavioral patterns and characteristics within these groups, providing insights on how to engage and convert customers more effectively.

1. **Goals**
2. **Feature Extraction and Dimensionality Reduction:**

Use **PCA** and **factor analysis** to cut down the dataset’s complexity, keeping important features that significantly influence shopper behavior. Check other dimension reduction tools for their ability to make the data easier to understand while keeping its value.

1. **Clustering Analysis:**

Use clustering algorithms **K-means** to divide online shoppers into two key groups: those who help generate revenue and those who do not. Study the clusters to identify the main differences and patterns, such as demographics, browsing behavior, and purchase history, that separate revenue-generating shoppers from the rest.

1. **Pattern Identification and Strategic Insights:**

Closely examine each cluster to find common patterns and behaviors linked with generating revenue and those not leading to sales. Offer strategic advice on how companies can customize their marketing and operational strategies to target and convert different types of online shoppers effectively.

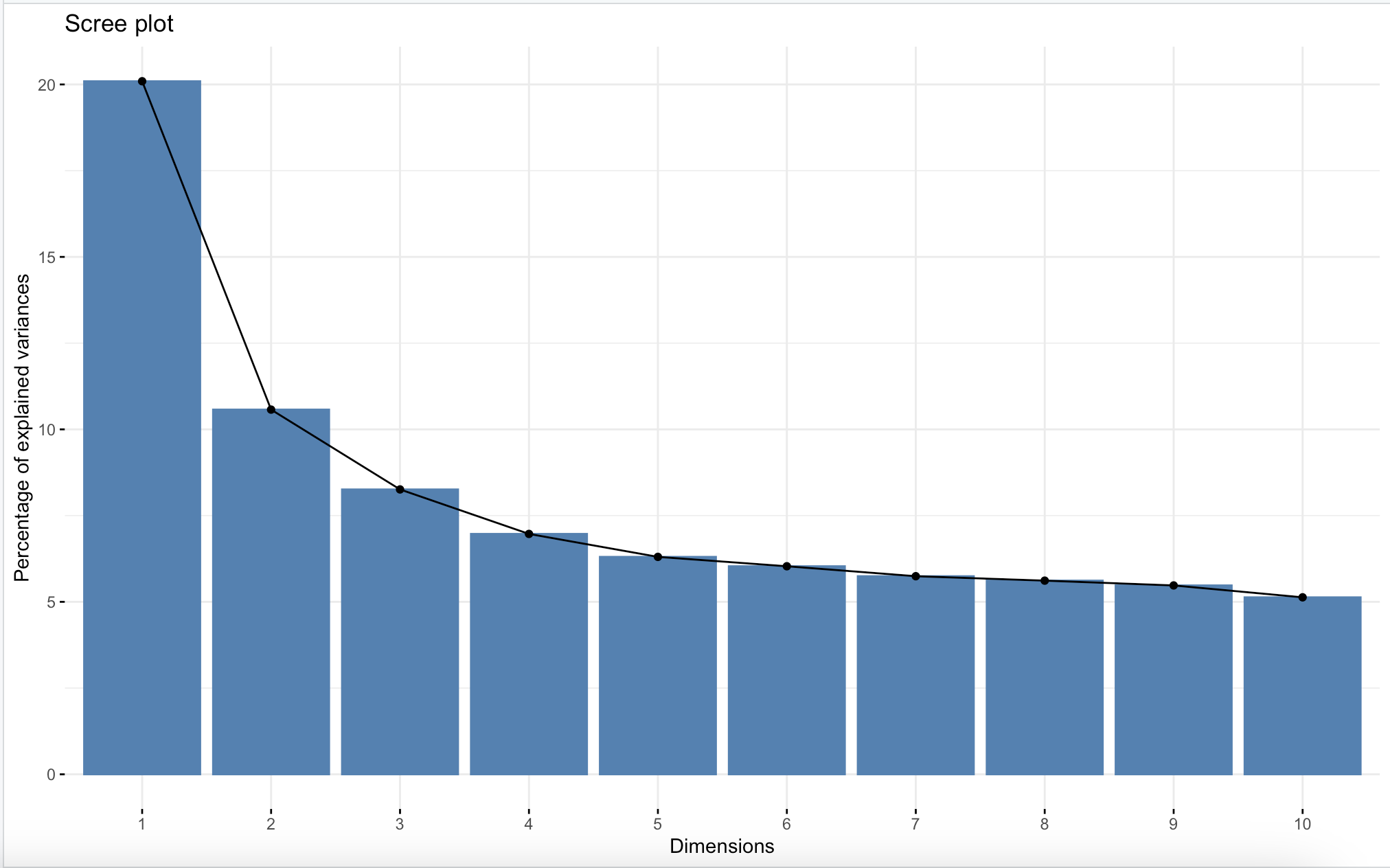
1. **Visualization and Interpretation:**

Use visualization tools to clearly present the results, highlighting differences and similarities between the clusters. Interpret the visual data to provide practical recommendations for business strategy and customer relationship management.

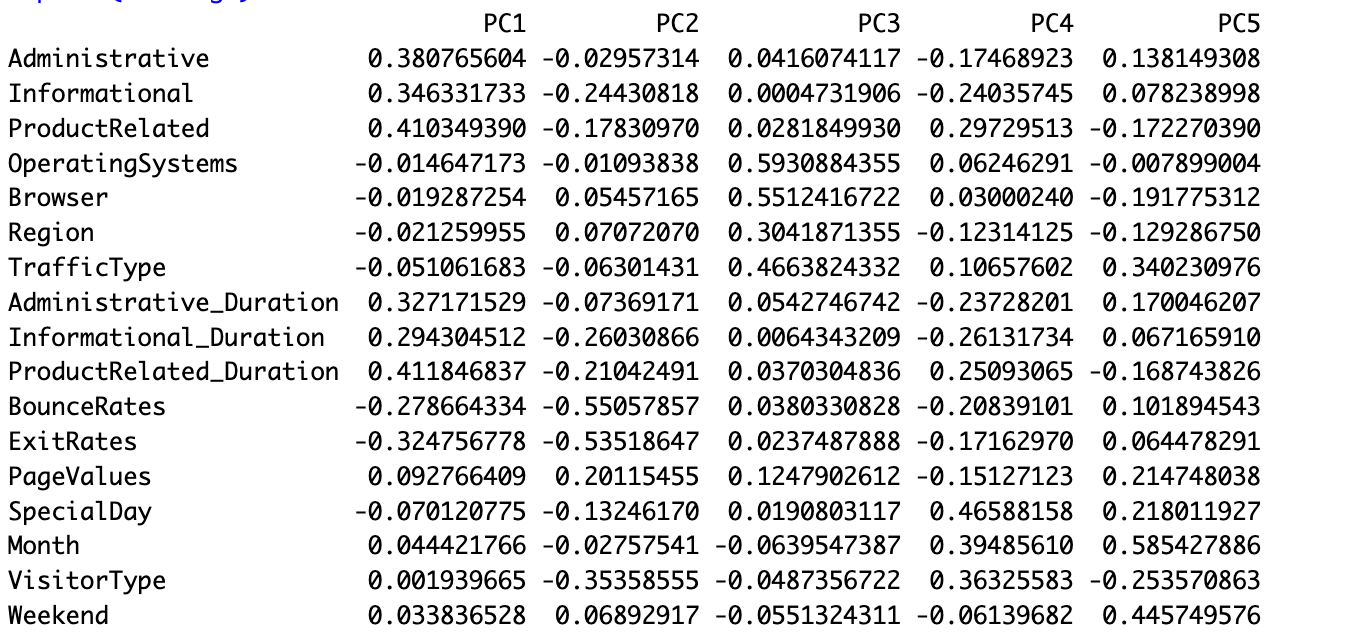
**III. Dataset Description**

| **Variable Name** | **Type** | **Description** |
| --- | --- | --- |
| Administrative | Integer | Administration Page visited by the visitor in that session |
| Administrative\_Duration | Integer | total time spent in Administration Page |
| Informational | Integer | Informational Page visited by the visitor in that session |
| Informational\_Duration | Integer | total time spent in Informational Page |
| ProductRelated | Integer | ProductRelated Page visited by the visitor in that session |
| ProductRelated\_Duration | Continuous | total time spent in ProductRelated Page |
| BounceRates | Continuous | The percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. |
| ExitRates | Continuous | The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. |
| PageValues | Integer | The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. |
| SpecialDay | Integer | The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. |
| Month | Categorical | Month |
| OperatingSystems | Integer | OperatingSystems |
| Browser | Integer | Browser Type |
| Region | Integer | Region |
| TrafficType | Integer | TrafficType |
| VisitorType | Categorical | visitor type as returning or new visitor |
| Weekend | Binary | whether the date of the visit is weekend |
| Revenue | Binary | Indicate either have revenue |

**IV. PCA**

* Looking for components that explain 80% of the variance

We found that first 10 principal components explain 80% of the variance

* Variable Loading 
* PCA Loadings Summary

1. PC1: User Engagement

Positive Loadings: ProductRelated, ProductRelated\_Duration, Administrative, Informational, Administrative\_Duration, Informational\_Duration.

This component appears to capture overall user engagement with the site, reflecting activities related to browsing products and accessing information.

Negative Loadings: BounceRates, ExitRates.

These indicate that higher scores on this component are associated with lower bounce and exit rates, suggesting better user retention or more successful engagement.

1. PC2: User Retention Metrics

Negative Loadings: BounceRates, ExitRates, VisitorType.

This component mainly captures aspects of user retention, with a strong inverse relationship indicating that high scores may be associated with features that contribute to lower bounce and exit rates, and different types of visitors might have distinct behaviors concerning these metrics.

1. PC3: Technical Aspects

Positive Loadings: OperatingSystems, Browser, TrafficType.

This component is strongly associated with technical aspects of the user’s access mode, possibly reflecting differences in user experience and site interaction based on the operating system, browser type, and traffic sources.

1. PC4: Seasonal and Event Influence

Positive Loadings: SpecialDay, Month.

Indicates a capture of seasonal or event-specific variations, possibly reflecting user behavior changes during special days or months (e.g., holidays or sales events).

Negative Loadings: Administrative\_Duration, Informational\_Duration.

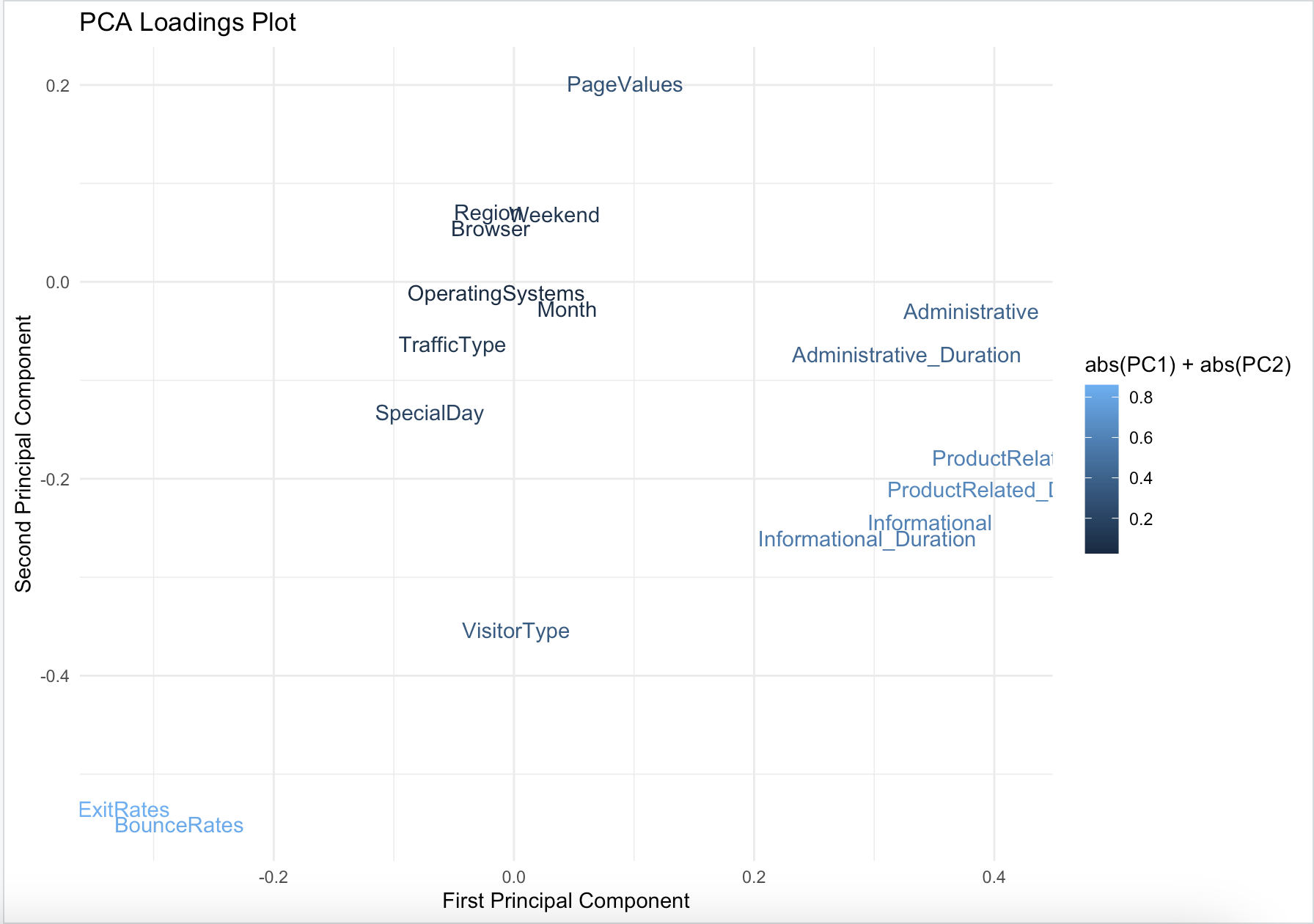
Suggests that during certain times (likely corresponding to positive loadings on SpecialDay and Month), the duration of administrative and informational activities decreases.

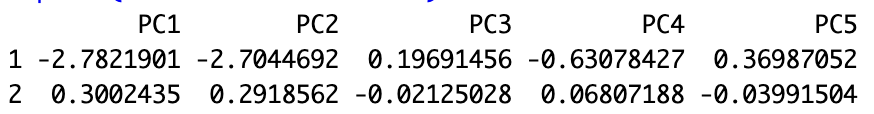
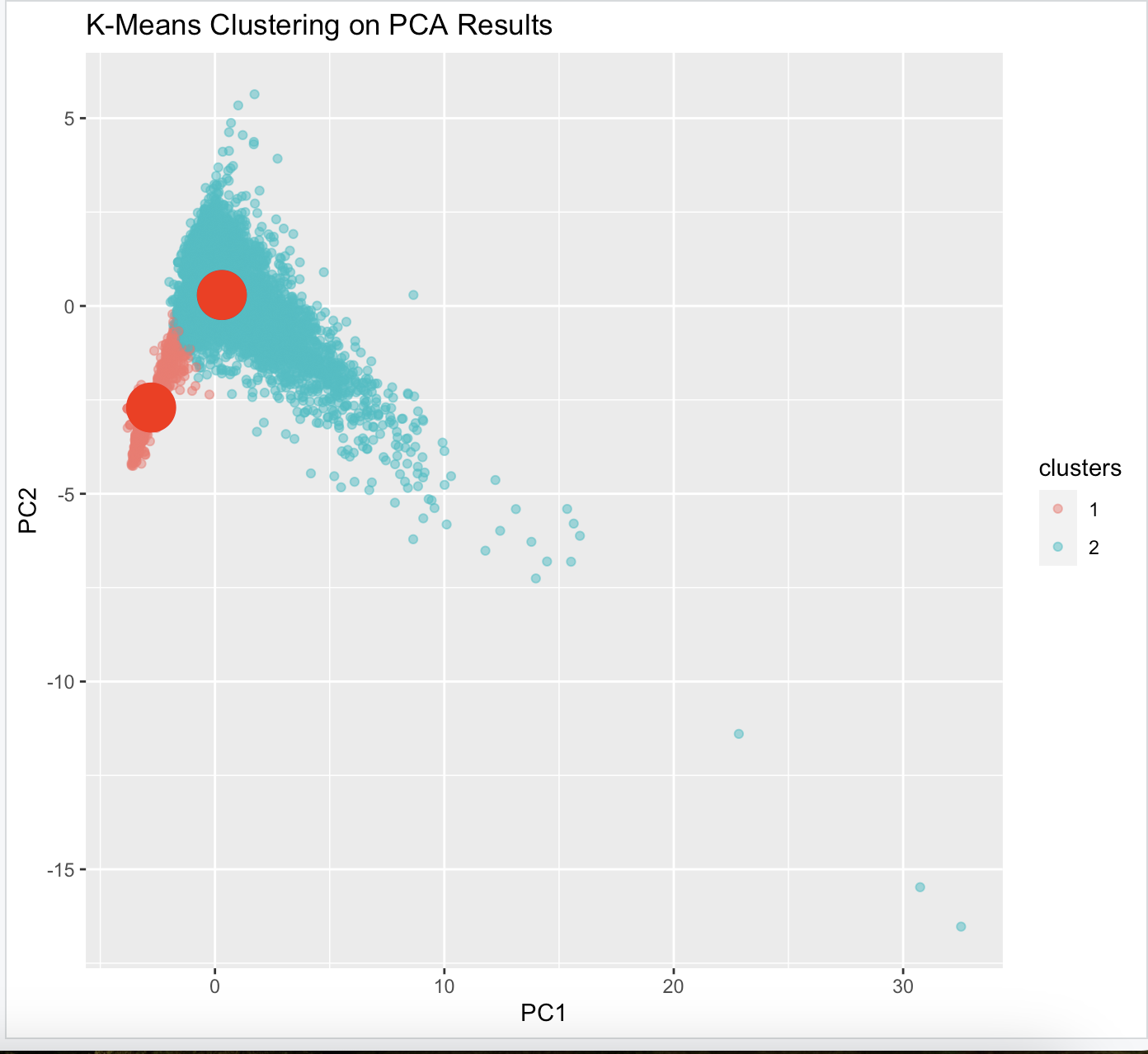
1. PC5: Time-Based Variability

Positive Loadings: Month, SpecialDay, Weekend.

This component reflects variations across different times, such as months and special days, highlighting changes in user behavior during weekends compared to weekdays.

* Plotting the first two principal components



* Implement the k-means using pca reduction
* Center coordinate under each Principal Component 
* Visualize the Clustering

Size of each group

Non-revenue: 11129 (2 or Blue)

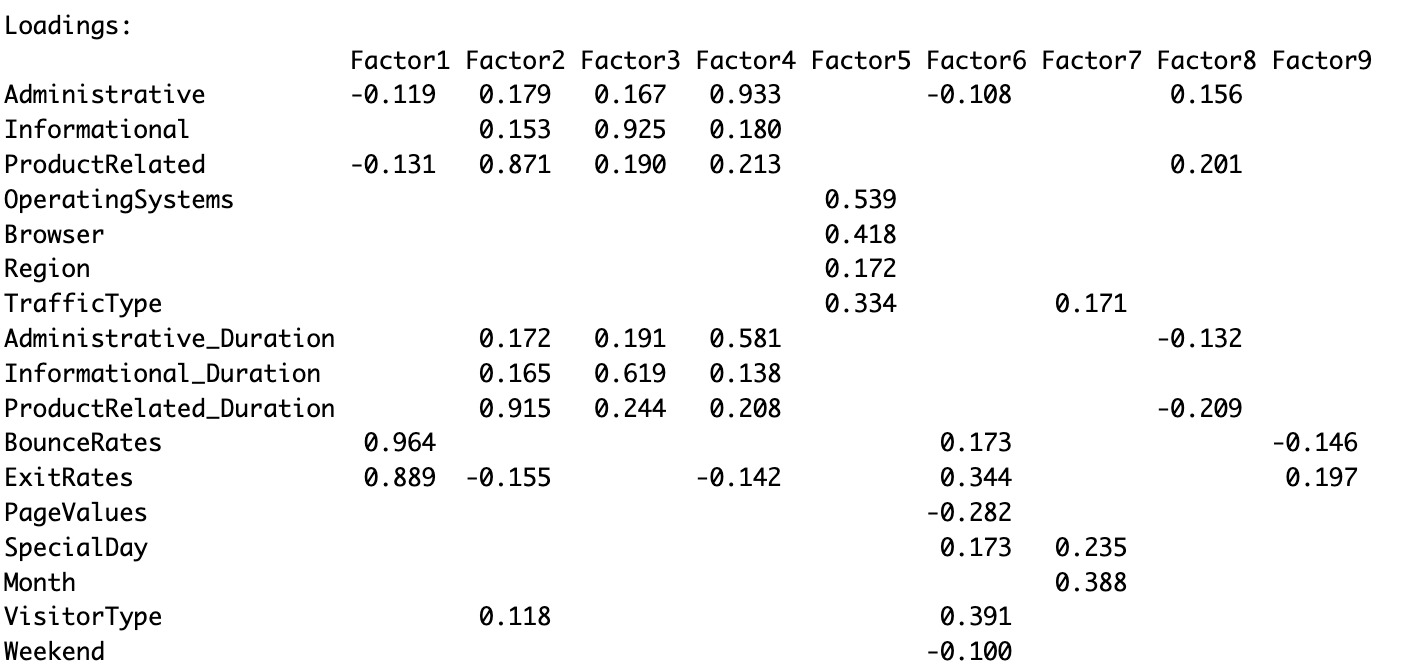
Revenue : 2030 (1 or Red)

Cluster Analysis Based on Revenue:

Group 1 (Non-Revenue, Red): This cluster is characterized by users who are less engaged and have higher bounce and exit rates as indicated by the PCA loadings on PC1 and PC2. The tight clustering around the center in the PCA plot suggests that these users have a more uniform pattern of lower engagement and possibly less complex interactions with the site.

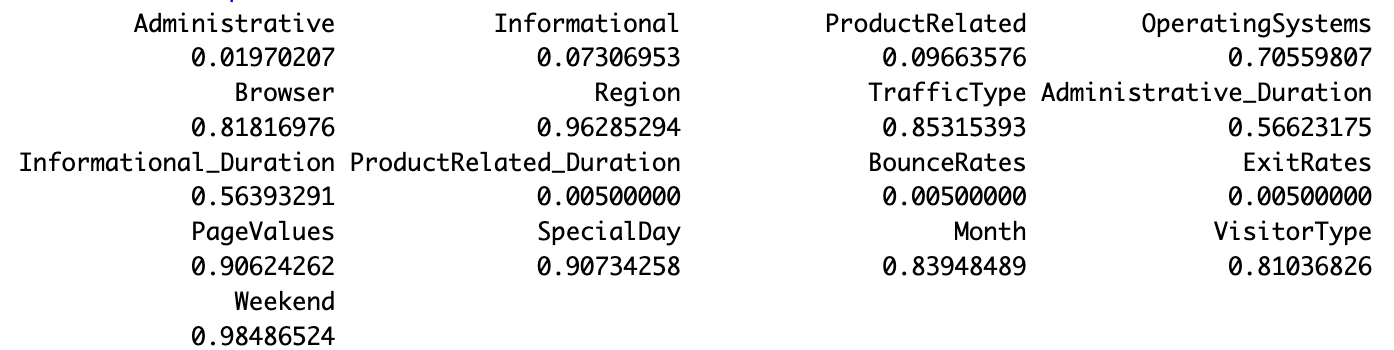
Group 2 (Revenue, Blue): This cluster includes users who exhibit a higher level of engagement with the site, which translates into revenue generation. The spread along PC1 in the PCA plot indicates varying levels of engagement and interaction complexity among these users. Their positioning along PC2 suggests better user retention metrics compared to Group 1.

**V. Factor Analyse**

* Loading

The factor loadings show how each variable contributes to the nine factors identified. High loadings (close to ±1) indicate strong associations between variables and specific factors. For instance:

1. Factor1 has a high positive loading with BounceRates and ExitRates, suggesting this factor could represent overall site engagement or user experience.
2. Factor2 is heavily influenced by ProductRelated and ProductRelated\_Duration, indicating a factor related to the depth of interaction with products.
3. Factor3 primarily relates to Informational and Informational\_Duration, which might capture the informational engagement of users.

* Uniquenesses

1. Low Uniqueness (≤ 0.1):

Variables: Administrative, Informational, ProductRelated, BounceRates, ExitRates, ProductRelated\_Duration

Implication: These variables have low uniqueness values, indicating that a significant portion of their variance is explained by the common factors derived in the factor analysis. This suggests that these variables are well represented by the factor model and are closely related to the underlying latent structures in the data.

1. Moderate Uniqueness (0.1 - 0.6):

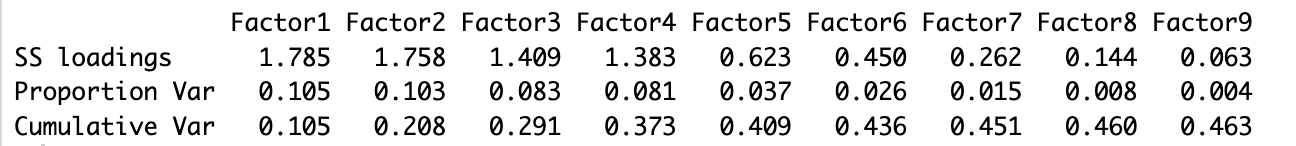
Variables: Administrative\_Duration, Informational\_Duration

Implication: These variables have moderate uniqueness values, meaning that while some of their variance is explained by the common factors, a substantial amount remains unique. This suggests that these variables are influenced by the factors but also have individual characteristics not captured by the common latent factors.

1. High Uniqueness (> 0.6):

Variables: OperatingSystems, Browser, Region, TrafficType, PageValues, SpecialDay, Month, VisitorType, Weekend

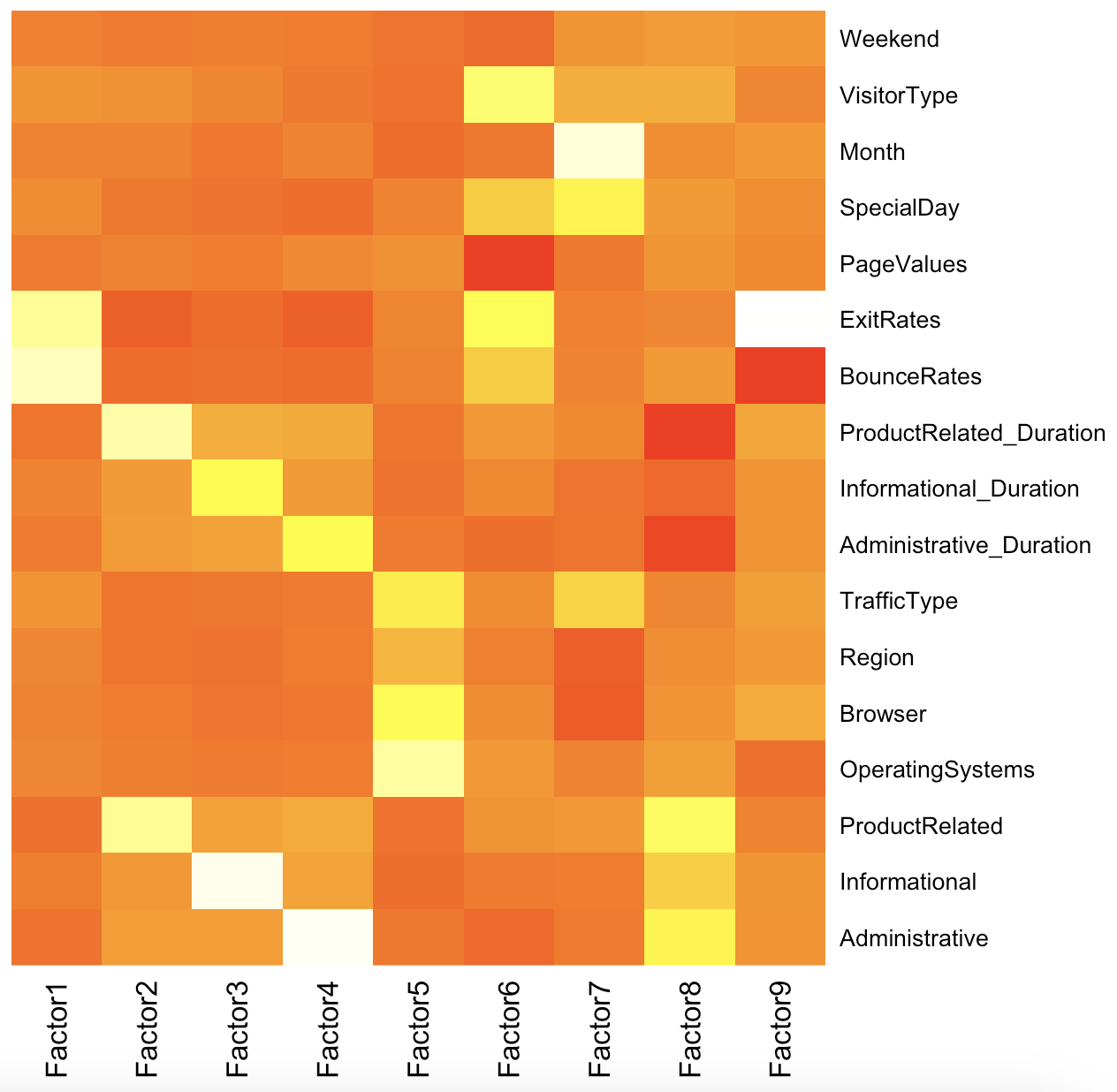
Implication: These variables have high uniqueness values, indicating that most of their variance is not explained by the common factors. This suggests that these variables do not correlate strongly with the underlying factors identified in your analysis and may be influenced by other factors not included or captured in this factor model.

* Sum of Squared Loadings and Variance

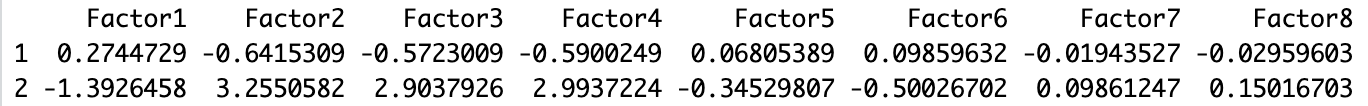
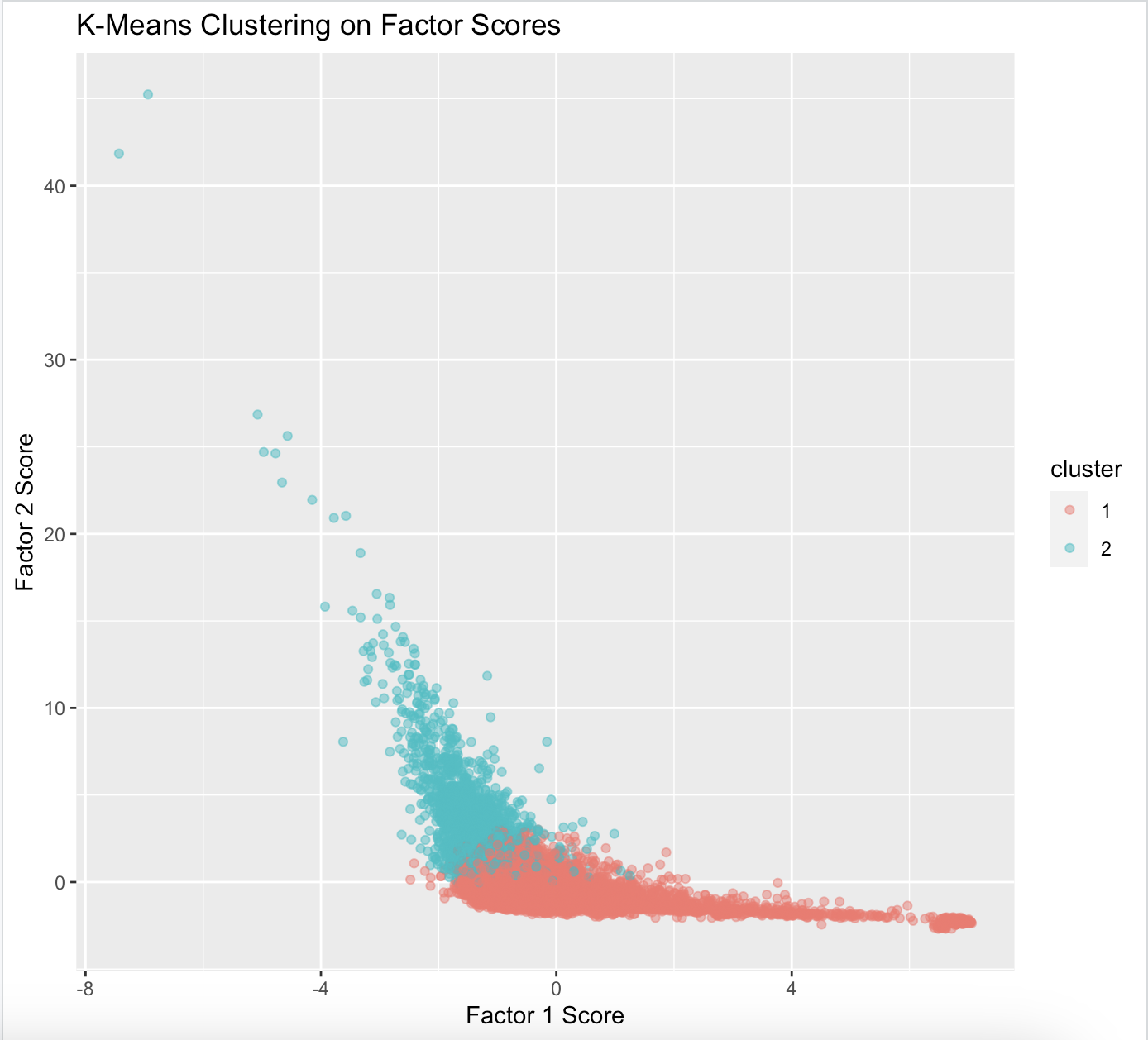
Sum of Squared Loadings indicate the variance accounted for by each factor. Factor1 is the most significant with an SS loading of 1.785, suggesting it explains more variance in the dataset than the other factors.

Cumulative Variance adds up the proportion of variance explained as more factors are considered. By Factor4, about 37.3% of the variance in the data is explained, reaching 46.3% by Factor9.

* Heatmap



Implement k-means clustering on the results from a factor analysis

* Center coordinate under each factor
* Visualize the Clustering 

Size of each group

Non-revenue : 10300 (2 or Red)

Revenue : 2030 (1 or Blue)

* Cluster Analysis Based on Revenue:

1. Group 1 (Red):

This cluster spans a broad range of Factor 1 scores and generally lower Factor 2 scores. The spread along Factor 1 indicates varying levels of poor user engagement and retention. The lower scores on Factor 2 suggest less depth in product interaction.

This group's characteristics could indicate users who visit the site but either bounce off quickly without engaging deeply with the content or have unsatisfactory experiences leading to early exits.

1. Group 2 (Blue):

This cluster spans a broad range of Factor 2 scores and generally lower Factor 1 scores.This indicates that these users engage more deeply with product content, potentially exploring multiple pages or spending more time on product-related activities, correlating with revenue generation.

The variation in Factor 1 scores, even if lower on average than Group 1, might still indicate some issues with user experience that could be further optimized to enhance revenue generation.

**VI. Conclusion**

* **Conclusion**

The combined use of PCA and FA in our project provided a comprehensive analysis framework. PCA offered a broad overview by identifying principal components that capture significant data variance, which streamlined the clustering process. FA complemented this by exposing deeper, latent structures within the data, enabling a more nuanced understanding of behavioral drivers.

This strategic application of both techniques enabled precise segmentation of online shoppers into revenue-generating and non-revenue groups, highlighted by distinct behavioral patterns identified through PCA and FA. The insights gained from this analysis are instrumental for online retailers looking to tailor marketing strategies, enhance customer engagement, and optimize conversion rates.

**Reference**

Blaufuks, W. (2022, March 28). FAMD: How to generalize PCA to categorical and numerical data. Medium. https://towardsdatascience.com/famd-how-to-generalize-pca-to-categorical-and-numerical-data-2ddbeb2b9210

Kalantan, Z. I., & Alqahtani, N. A. (2019). A study of principal components analysis for mixed data. International Journal of Advanced and Applied Sciences, 6(12), 99–104. https://doi.org/10.21833/ijaas.2019.12.012

Dunn, J. C. (1974). "Well-separated clusters and optimal fuzzy partitions". Journal of Cybernetics. 4 (1): 95–104. doi:10.1080/01969727408546059.

Tibshirani, R., Walther, G., & Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic". Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423. doi:10.1111/1467-9868.00293

van, M. (n.d.). Introduction to the gower package - the comprehensive R archive network. https://cran.r-project.org/web/packages/gower/vignettes/intro.pdf