Market Research

Understanding Netflix Customers to Suggest New Content

By

Pietro Godowicz

Ashkan Lotfollahzadeh

Enrique Marenco

Francesco Genovese

Glossary

1. Introduction……3
2. Survey Description……4
3. Questionnaire……..5
    a. Screening Questions……5
    b. Socio-Demographic Questions……6
    c. Core Questions ……6
    d. Behavioral Questions……8
4. Data Observation ……..10
    a. Principal Component Analysis……..10
    b. Relationship of PCA and SVD of Data Matrix……..11
    c. Size-effect Theory……..13
5. Code Explanation……...14
6. Cluster Analysis……20
7. Chi-Suare Analysis……22
8. T-test on Core Components……22
9. Cluster Description ……...25
10. Conclusion ……27
11. References……..28

Introduction

Netflix, the world's leading streaming entertainment service, has changed the way people consume television and movies. With over 208 million subscribers worldwide, it has become a cultural phenomenon that has greatly impacted the entertainment industry. This research paper presents a study aimed at gaining a deeper understanding of Netflix users and identifying opportunities to create value for both the company and its content. A survey was conducted to gather information about the preferences, habits, and behaviors of Netflix users. Cluster analysis and T-tests were used to analyze the data and draw conclusions about the user base. The results of the study provide valuable insights into the needs and desires of Netflix users, and suggest a number of strategies that the company can use to enhance the user experience and increase the value of its offerings. The findings of this research highlight the importance of considering the perspectives of the audience in the decision-making process, and demonstrate the usefulness of cluster analysis and T-tests as methods for understanding user behavior. This study is expected to contribute to the body of knowledge in the field of streaming entertainment and to inform the ongoing efforts of Netflix to create value for its customers and the company.

Survey Description

A survey was conducted to gather information about the preferences, habits, and behaviors of Netflix users in comparison to other streaming services users. The survey consisted of four sections, including the Screening Section, Socio-Demographic Questions, Core Questions, and Behavioural Questions. The Screening Section was designed to discern answers from Netflix users and those who prefer other services, and the Socio-Demographic Questions aimed to understand the characteristics of the sample in terms of age, gender, education, occupation, marital status, and geographical origin.

The Core Questions were aimed at collecting the personal opinions of the respondents regarding the situations in which Netflix is more appreciated and the factors that influence the choice of a streaming service over another. The Behavioural Questions aimed at understanding the typical habits of the respondents related to the use of these platforms, such as the monthly amount paid, preferred episode or series length, and preferences between movies or series.

The results of the survey provided valuable insights into the needs and desires of Netflix users and other streaming services users, and the information gathered through the Socio-Demographic Questions and Behavioural Questions allowed for a more precise analysis of the sample. The findings of the survey demonstrate the importance of considering the perspectives of the audience in the decision-making process and provide valuable information that can inform the ongoing efforts of Netflix to create value for its customers and the company.

Questionnaire

Screening Questions:

1. Do you watch Netflix?
    - Yes
    - No
2. Why don't you watch Netflix?
    - I use other streaming platforms
    - I prefer to watch cable TV
    - I prefer to spend money and time for other kind of entertainments
    - Other

Socio-Demographic Questions:

3. Sex:
    - Male
    - Female
4. How old are you?
    - 18 or younger
    - 19 - 25
    - 26 - 39
    - 40 or older
5. What is your degree of education?
    - High School or lower
    - Bachelor's Degree
    - Master's Degree or higher
6. Which is your area of study/business?
    - Management and Administration
    - Arts and Humanities
    - Science and Technology
    - Healthcare
    - Education
    - Other
7. Are you in a relationship?
    - Yes
    - No
8. Where are you from?
    - North America

- Africa
- Europe
- Asia
- Oceania

Core Questions

9. Spend quality time with my family and loved ones.(How much do you enjoy this activity when you are with your family or friend)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

10. I use the platform to relax. (How important is it for you to have something to watch while you relax)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

11. I watch it to pass the time. (How important is it to you to have a streaming service when you are bored/don't know what to do)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

12. I want to watch quality content with quality productions. (How important is the quality of contents for you?)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

13. I choose the streaming service considering the amount of content. (How important is it for you to have the possibility of choosing between a large number of movies and series?)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

14. I enjoy the availability of contents in different languages and flexibility of translated subtitles. (How important is it for you to have translations and subtitles in many different languages?)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

15. I like to discover some new exciting contents. (How important is it for that your streaming service suggests to you contents that you could enjoy or that it is updated regularly with new contents)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

16. I want to have something to watch while I perform other activities (Eating, exercising..)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

17. I choose the services considering how easy it is to use. (How much the usability influences your choice of a streaming platform)

1	2	3	4	5	6	7
○	○	○	○	○	○	○

Behavioral Questions

18. How often do you use your streaming service?
    - Once a week or less
    - 2 to 4 times a week
    - 5 or more times per week
    - Everyday

19. What do you prefer to watch?
    - Movies
    - Series

20. Your ideal series is composed of…
    - 1 season
    - 2 - 5 seasons
    - 6 - 9 seasons
    - 10+ seasons

21. In your ideal series the episodes are…
    - Short (less than 30 minutes)
    - Long (30 minutes or more)

22. Which is your favorite category to watch?
    - Action
    - Adventure
    - Animation
    - Comedy
    - Drama
    - Documentary
    - Fantasy
    - Historical
    - Horror
    - Romantic
    - Scientific
    - Thriller
    - Western
    - Other

23. For how long have you used your streaming platform?
    - up to 1 year
    - 1 - 3 years
    - 3 - 5 years
    - more than 5 years

24. How many streaming platforms do you watch?
    - only one
    - 2 or 3
    - more than 3

25. How much money do you spend on streaming services a month?
    - Somebody else pays for me
    - from 0$ to 6$
    - from 7$ to 13$
    - from 13$ to 20$
    - more than 20$

Data Observation

Following the completion of the survey, it was determined that only the participants who were users of Netflix would be considered for analysis. The resulting dataset consisted of 255 observations. Once the data was collected, the initial step was to rename the opinion variables, which were the results from the core questions. These variables were renamed as d10_1, d10_2, and so forth, up to d10_9.

- "Spend quality time with my family and loved ones." (d10_1)
- "I use the platform to relax."(d10_1)
- "I watch it to pass the time"(d10_1)
- "I want to watch quality contents with quality productions"(d10_1)
- "I choose the streaming service considering the amount of contents"(d10_1)
- "I enjoy the availability of contents in different languages and flexibility of translated subtitles"(d10_1)
- "I like to discover some new exciting contents"(d10_1)
- "I want to have something to watch while I perform other activities (Eating, exercising, etc)"(d10_1)
- "I choose the services considering how easy it is to use"(d10_1)

Principal Component Analysis

The Principal Component Analysis (PCA), is a multivariate technique for examining relationships among several quantitative variables related to the Singular Value Decomposition of data matrix. Indeed, finding principal components starts with SVD of covariance matrix or correlation matrix. In case we use a covariance matrix, we usually use centered variables (subtraction of mean from the data). In the case of correlation matrix, entries are also divided by their corresponding variance which is called standardizing and it is used when there is different scaling for each variable. This transformation specially is useful for survey analysis regarding the fact that setting a scale for sentiments of individuals is impossible and we have the problem of different scales among respondents. Using PCA as a multivariate analysis method, has usually two main applications to analyze numeric data. First, providing a 2-dimensional display of the data by choosing principal components (scores), corresponding to two largest eigenvalues, to plot the data. Second, dimension reduction for analyzing the data by choosing a certain number of PC's considering their eigenvalues magnitude. In other words, the results of the PCA give us the possibility to reduce the variables in our analysis [4],[2].

Relationship of PCA and SVD of a data matrix [1],[2]**:**

If A is an $n \times n$ matrix, in order to find eigenvalue λ and an associated eigenvector v, it must be the case that $Av = \lambda v$ and this is equivalent to the homogeneous system.

$$(A - \lambda I)v = 0$$

If λ is a root of $p(\lambda) = det(A - \lambda I)$, it is an eigenvalue of A and if v is a nonzero column vector satisfying $Av = \lambda v$, it is an eigenvector of A[1].

The singular value decomposition of a matrix $X_{n\times p}$ is:

$$X_{n\times p} = U_{n\times n} S_{n\times p} V_{p\times p}$$

In which, U contains left singular vectors of X (eigenvectors of XX^T) and V contains right singular vectors of X (eigenvectors of $X^T X$), and [1]:

$$S = \begin{bmatrix} s_1 = \sqrt{\lambda_1} & & 0 & \\ & s_2 = \sqrt{\lambda_2} & & \\ & & \ddots & \\ & 0 & & s_p = \sqrt{\lambda_p} \\ \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots\cdots\cdots & \cdots\cdots\cdots \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

When we do the decomposition on covariance or correlation matrix of our data matrix, we are operating on a symmetric matrix, as a result left and right singular vectors are eigenvectors of the matrix.

Let the data matrix $X_{n\times p}$, where n is the number of respondents and p is the number of variables. Let us assume that it is centered, i.e. column means have been subtracted and are now equal to zero. Then the $p \times p$ covariance matrix C is given by $C = X^T X/(n-1)$ It is a symmetric matrix, its right and left eigenvectors are equal so its singular value decomposition is:

$$C = VLV^T$$

where V is a matrix of eigenvectors (each column is an eigenvector) and L is a diagonal matrix with eigenvalues λ_i in the decreasing order on the diagonal. The eigenvectors are called principal axes or principal directions of the data. Projections of the data on the principal axes are called principal components, also known as PC scores; these can be seen as new, transformed, variables. The j-th principal component is given by j-th column of XV. The coordinates of the i -th data point in the new PC space are given by the i-th row of XV.

If we now perform singular value decomposition of X, we obtain a decomposition $X=USV^T$,where U is a unitary matrix and S is the diagonal matrix of singular values s_i. From here one can easily see that $C = VSU^TUSV^T/(n-1) = V\frac{S^2}{n-1}V^T$,meaning that right singular vectors V are principal directions and that singular values are related to the eigenvalues of covariance matrix via $\lambda_i = s_i^2/(n-1)$. Principal components are given by $XV=USV^TV=US$.

.

Followings should be considered about the PCA and the data matrix:

1. To summarize: If $X=USV^T$, then columns of V are principal directions (axes). Columns of US are principal components ("scores"). Singular values are related to the eigenvalues of the covariance matrix via $\lambda_i = s_i^2/(n-1)$. Eigenvalues λ_i show variances of the respective PCs.

2. The above is correct only if X is centered. Only then is covariance matrix equal to $X^TX/(n-1)$. The above is correct only for X having samples in rows and variables in columns. If variables are in rows and samples in columns, then U and V exchange interpretations. If one wants to perform PCA on a correlation matrix (instead of a covariance matrix), then columns of X should not only be centered, but standardized as well, i.e. divided by their standard deviations.

3. To reduce the dimensionality of the data from p to $k < p$, select k first columns of U, and $k \times k$ upper-left part of S. Their product U_kS_k is the required $n \times k$ matrix containing first k PCs. Further multiplying the first k PCs by the corresponding principal axes C_k^T yields $X_k=U_kS_kV_k^T$ matrix that has the original $n \times p$ size but is of lower rank (of rank k) so we have our data in a lower dimension.

4. Strictly speaking, U is of $n \times n$ size and V is of $p \times p$ size. However, if $n > p$ then the last $n - p$ columns of U are arbitrary (and corresponding rows of S are constant zero); one should therefore use an economy size (or thin) SVD that returns U of $n \times p$ size, dropping the useless columns. For large $n \gg p$ the matrix U would otherwise be unnecessarily huge. The same applies for an opposite situation of $n \ll p$.

Size-effect theory(required transformations of data before PCA)[3]**:**

Size-effect has been recognized as a phenomenon in PLS-PM (Partial Least Square Path Modeling) as an estimation approach for SEM (Structural Equation Model). PLS-PM is also known as a component based estimation approach. The basic idea of SEM is that complexity inside a system can be explained taking into account a network of causal relationships (imagine a pattern represented by graphs), defined according to a theoretical model, linking latent complex concepts, called Latent Variables (LV)(that is what we are going to find after analysis), each measured by several observed indicators usually defined as Manifest Variables (MV)(which here is our collected data). The main aim of component-based methods like PLS-PM is to provide an estimation of the latent variables in the model in such a way that they are able to properly explain the causal relationships defined by the path diagram structure and, at the same time, the most representative of each corresponding block of manifest variables. In PLS path modeling a priori knowledge is incorporated in the algorithm. At least at the theoretical level, the MVs one block can always be built in a way that they are all positively correlated. On practical data this condition has to be checked. A block is essentially unidimensional if the first eigenvalue of the correlation matrix of the block MVs is larger than 1 and the second one is smaller than 1, or at least very far from the first one.

In opinion surveys, evaluation of intangible constructs (latent variables) is conditioned to the perception of the measurement scale suggested to respondents. Level of involvement of the respondent in fact is often the real cause of the first principal component, and therefore the PLS-PM assumption of unidimensionality is only mathematically verified.

The presence of a size-effect generating a strong first principal component in the data is related only to the perception of the opinion scale. The so-called ”size factor” Level of involvement of the respondents tends to be the real cause of the first principal component, and therefore the PLS-PM assumption of unidimensionality is verified because of the presence of a large size effect. Many of the operative applications of PLS-PM to customer satisfaction surveys analyze and model therefore not the true latent component of satisfaction, but an aspect of perception of the scale to measure such satisfaction.

Using a particular transformation of the data it is possible to eliminate size-effect but after this transformation the probability that the hypothesis of unidimensionality is verified becomes much lower. It means that with lower probability we can be sure about the practical value of latent variables or relationships among manifest variables after analysis and hypothesis testing.

One technical solution should be the elimination from the classification procedure of factorial coordinates on the first factor and then proceed with the analysis of the data structure ”cleaned” of the size-effect.

$$k_{ij} = 0 \quad if\ x_{ij} = x_{avg}$$

$$k_{ij} = \frac{\left(x_{ij} - x_{min}\right)}{\left(x_{avg} - x_{min}\right)} \quad if\ x_{ij} < x_{avg}$$

$$k_{ij} = \frac{\left(x_{ij} - x_{avg}\right)}{\left(x_{max} - x_{avg}\right)} \quad if\ x_{ij} > x_{avg}$$

The result will be given by a new variable k_j which will have the range of variation -1, 1. It is evident that if the average is perfectly in the midrange of the used scale, the re-coding function will be the same overall the range maximum - minimum.

An important property of this type of scaling transformation is that all respondents, in the new system, have a vector of opinions between -1 and +1, centered on zero corresponding to their own individual average.

Code Explanation

Transformation of data to cleaned size-effect:

Based on previous explanations, the procedure that we implemented declares a new dataset called **Project_new**; this dataset has the columns of average (**avgi**), minimum (**mini**), and maximum (**maxi**) of the variables **d10_1** to **d10_9**. The procedure also declares two arrays: **p1** with the values of **d10_1** to **d10_9**, and **p2** with **new_1** to **new_9**. A **do** loop is performed over the **p2** array,where for each element:

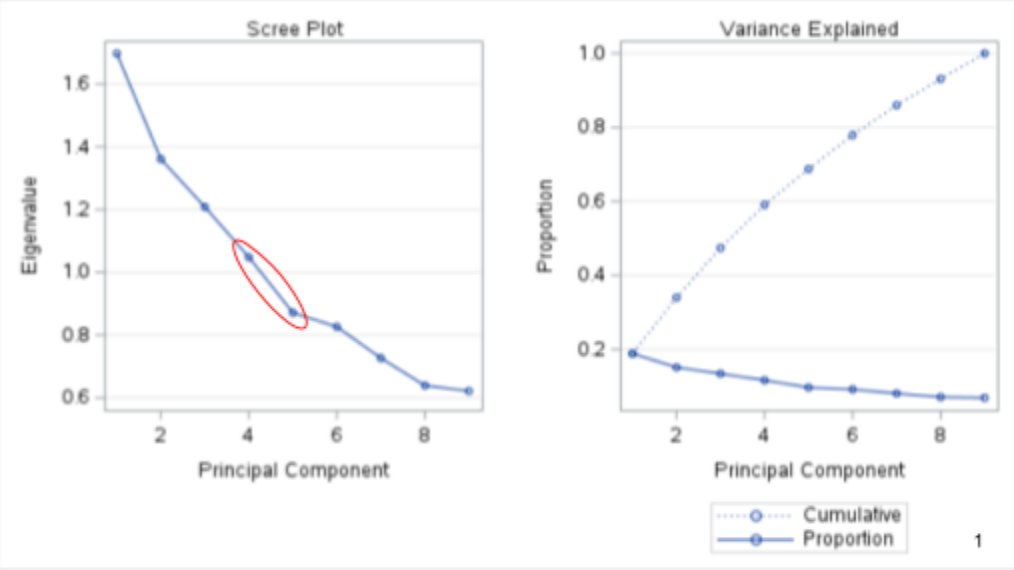
- If the corresponding element in **p1** is greater than **avgi**, **p2** is set to the value of **(p1-avgi)/(maxi-avgi)**.
- If the corresponding element in **p1** is less than **avgi**, **p2** is set to the value of **(p1-avgi)/(avgi-mini)**
- If the corresponding element in **p1** is equal to **avgi**, **p2** is set to 0

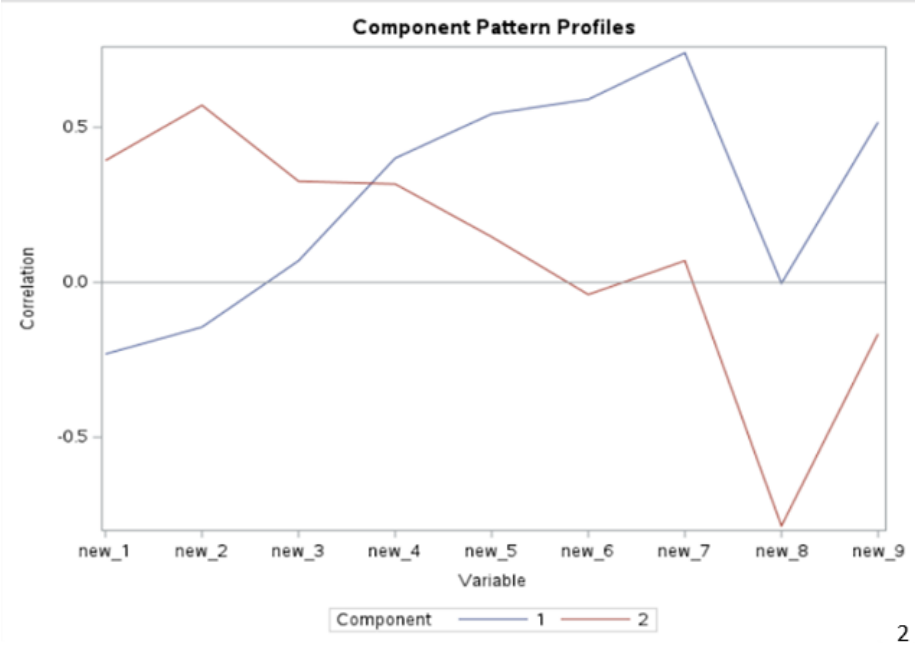
- If the corresponding element in **p1** is missing (. or **null**), **p2** is set to 0. The purpose of the code is to normalize the values of **d10_1** to **d10_9** and store them in **new_1** to **new_9.**

The new variables **new_1** to **new_9** are the variables on which the mentioned size-effect transformation has been done.

Implementation of PCA on data and selection of PC's:

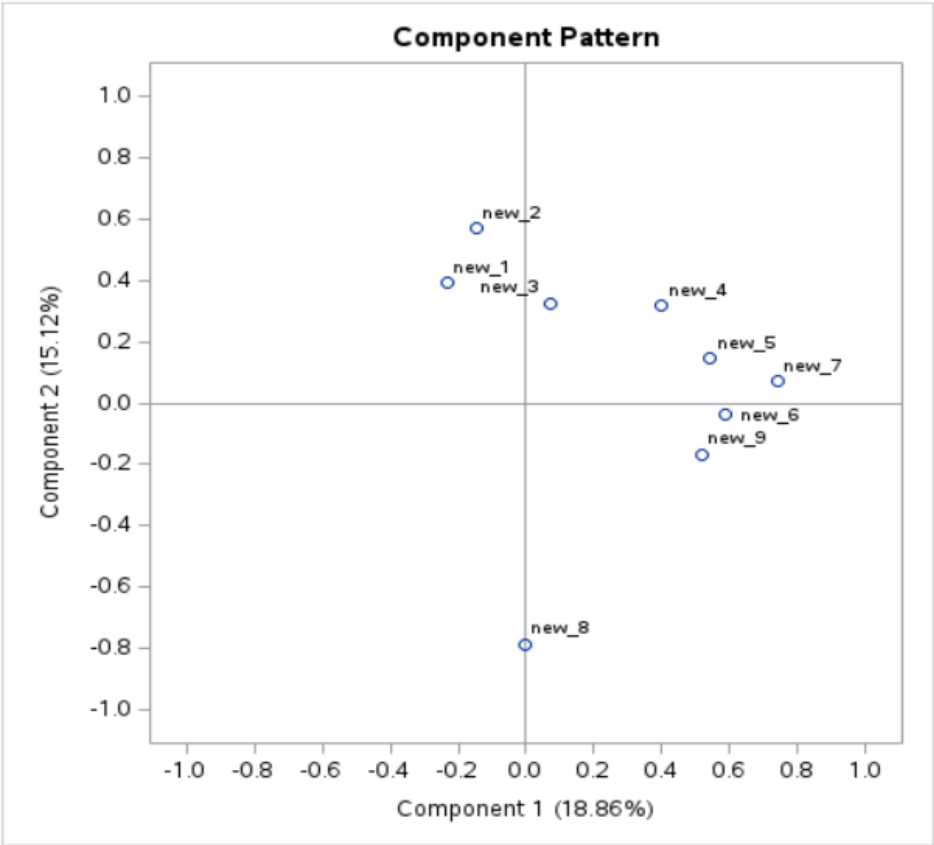
In this stage we use the proc **princomp** command to implement PCA on our data and store results in the new table **coord_new.** Based on the below graph we notice a big change (point of inflection) between two eigenvalues of the data correlation matrix. This observation is the reason to retain only the first 4 principal components for analyzing the data.The reason behind this choice is the fact that the built matrix by means of smallest eigenvalues and their corresponding eigenvectors would have small impact on values of the main data correlation matrix.



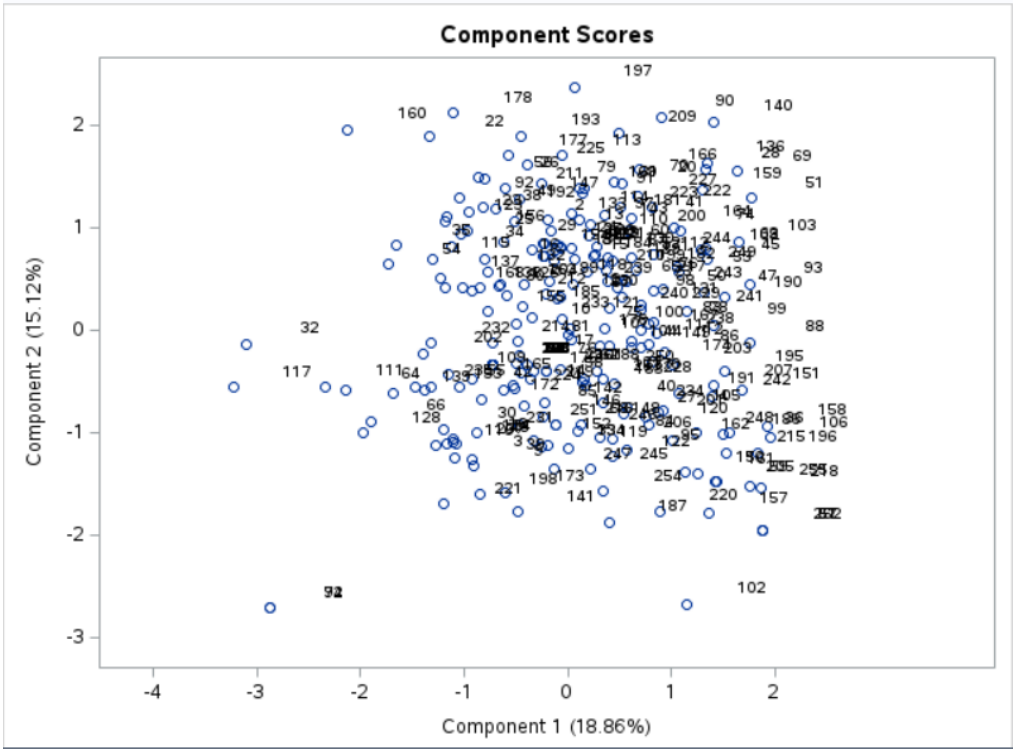


2

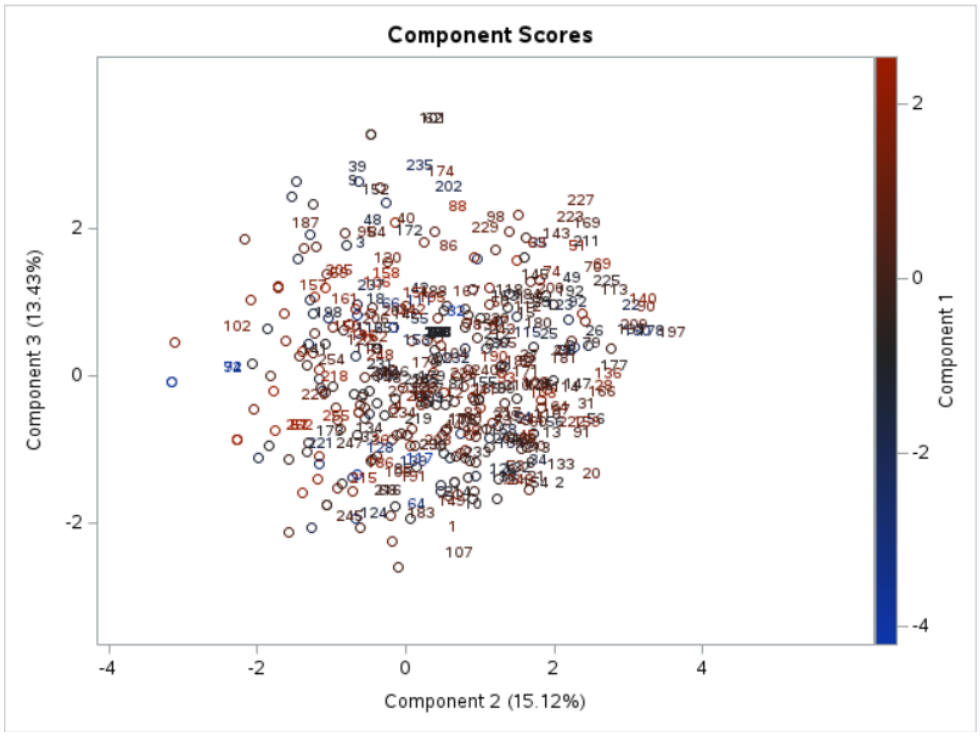
Correlation of the first two PC's with variables, the first PC has a positive correlation with about 60% of the data. This plot generally displays correlation of the two first PC's and variables.

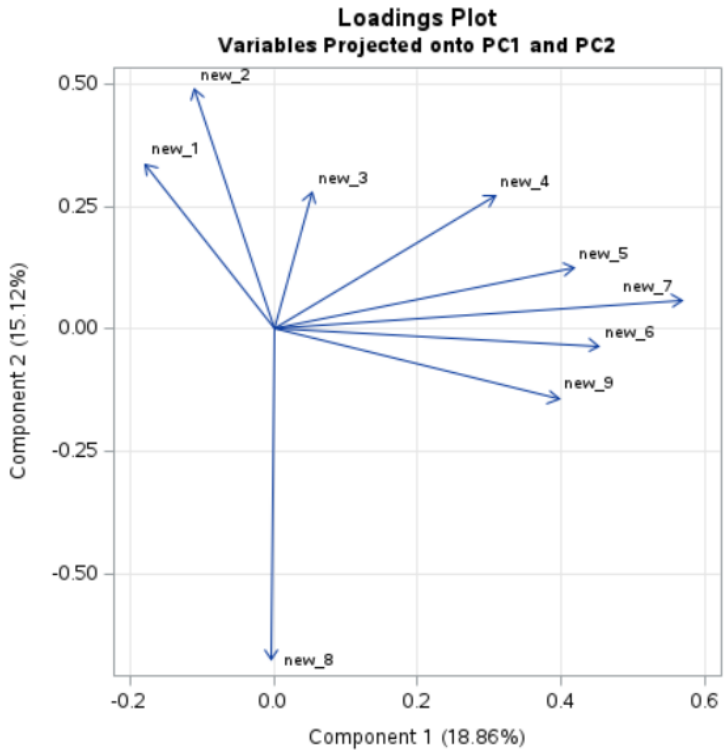


Location of variables relative to each other based on their projection on 2-dimensional space of the first two PC's.



Presentation of data on 2-dimension using the two first PC's and 3-dim PC one as color.





The loadings plot shows the relationship between the PCs and the original variables.

We notice that by considering 4 values out of 9, which are the one with an Eigenvalue higher than 1, we explain 59.06% of our data correlation. For this reason our approach is to base our principal component analysis on only 4 variables because they can explain the majority of the variance, reducing the number of dimensions.

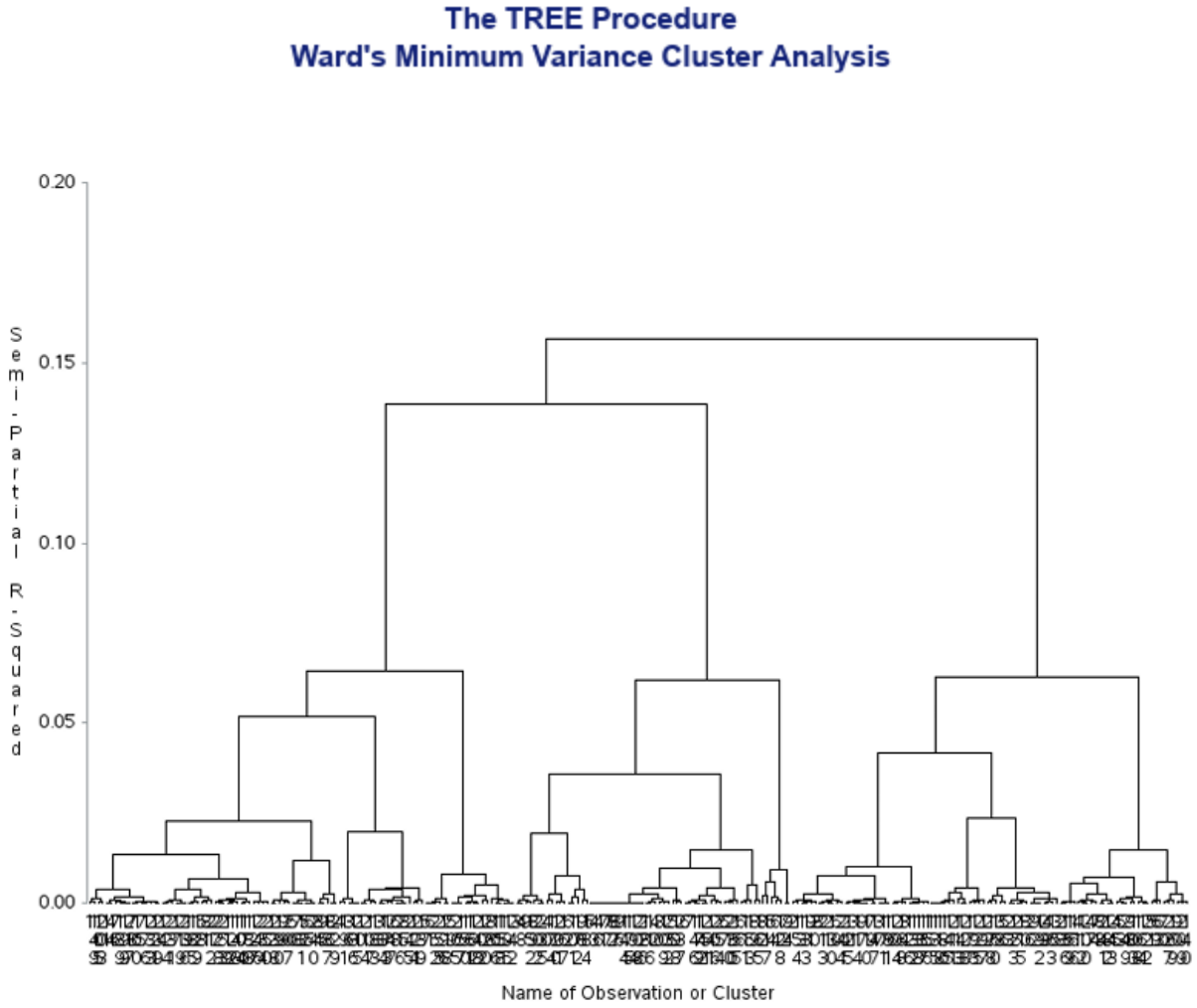
Correlation Matrix

	new_1	new_2	new_3	new_4	new_5	new_6	new_7	new_8	new_9
new_1	1.0000	0.0801	-.0668	0.0406	-.0984	-.0464	-.1161	-.1615	0.0089
new_2	0.0801	1.0000	0.1844	-.0764	-.0867	0.0037	0.0022	-.1837	-.0425
new_3	-.0668	0.1844	1.0000	-.0345	-.0098	-.0191	0.1040	-.0600	0.0250
new_4	0.0406	-.0764	-.0345	1.0000	0.1331	0.0966	0.1590	-.1745	0.0626
new_5	-.0984	-.0867	-.0098	0.1331	1.0000	0.1318	0.2312	-.1332	0.0810
new_6	-.0464	0.0037	-.0191	0.0966	0.1318	1.0000	0.2772	0.0419	0.1691
new_7	-.1161	0.0022	0.1040	0.1590	0.2312	0.2772	1.0000	-.0052	0.2513
new_8	-.1615	-.1837	-.0600	-.1745	-.1332	0.0419	-.0052	1.0000	0.1145
new_9	0.0089	-.0425	0.0250	0.0626	0.0810	0.1691	0.2513	0.1145	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.69762593	0.33667962	0.1886	0.1886
2	1.36094631	0.15254511	0.1512	0.3398
3	1.20840120	0.16018819	0.1343	0.4741
4	1.04821301	0.17688781	0.1165	0.5906
5	0.87132519	0.04499196	0.0968	0.6874
6	0.82633323	0.09986170	0.0918	0.7792
7	0.72647153	0.08734910	0.0807	0.8599
8	0.63912244	0.01756128	0.0710	0.9309
9	0.62156115		0.0691	1.0000

The next function we implemented is the **proc cluster** procedure used for clustering the observations in a data set based on the similarity of their variables. The code specifies the following options: **data=coord_new**: The input data set is named **coord_new**. **method=ward**: The clustering method used is "Ward's method", which aims to minimize the variance of distances between the newly formed cluster and its children. **outtree=tree_new**: The output of the procedure will be stored in a data set named **tree_new**. The var statement defines the variables to be used in the clustering, in this case, **prin1-prin4** and **id**. The second procedure, **proc tree**, creates a dendrogram, which is a graphical representation of the hierarchical clustering solution obtained from the previous procedure. The run statement is used to execute the procedure. The result of this procedure is a dendrogram, which is a diagram that represents a tree and in this case is used to represent hierarchical clustering by showing all the possible clusters that can be created by splitting the data that we collected. Based on the cut line we chose 4 clusters to analyze.

Cluster Analysis



Cluster analysis is a statistical technique used in market research surveys to identify homogeneous groups of individuals or objects, also known as **clusters**.The goal of cluster analysis is to group samples with similar characteristic and divide the samples with less in common this technique is based on the principle of minimizing the within-group variation and maximizing the between-group variation.

In the performed research, cluster analysis is used to identify subgroups within a population that share similar characteristics, behaviors, or preferences. There can be more ways to group the samples in the same survey starting from the simplest ones like sex, age, geographical, ecc, and so only by understanding the sample you can have a better cluster. Moreover there isn't a right algorithm to do clustering, but it depende of the sample you have. The cluster Analysis is useful in many subject matters, it can be used to develop targeted marketing strategies, improve product design, and understand consumer behavior.

With the following procedure we created 4 clusters, this was the starting number from which the analysis has started.

```
proc tree data=tree_new noprint nclusters=4 out=cluster_new;

id id;

run;
```

With this procedure we create a new table called **project_new_1** that merges **project_new** and **cluster_new** tables by id and before we sort the id of both.

```
proc sort data=project_new; by id; run;

proc sort data=cluster_new; by id; run;

data project_new_1; merge project_new cluster_new;

by id;

Run;
```

The code performs a cluster analysis on a data set called "Project" to form 4 clusters. The steps involved in the analysis include:

- standardizing the data using the mean and minimum/maximum values,
- performing a principal component analysis (PCA) on the standardized data,
- clustering the data based on the PCA results using the Ward method,
- creating a dendrogram to visualize the resulting clusters,
- merging the cluster assignments with the original data, and
- conducting chi-squared tests for independence between the cluster assignments and the categorical variables in the data set.

The next step is to do the chi-squared test for each Socio-Demographic and Behavioral component to understand how correlated is the fact of being part of one specific cluster and the fact of having a particular socio-demographic or behavioral characteristic. This is done by comparing the chi-squared p-value with an **alpha** value that we set at **0.05.** This has been chosen because it's one of the most used value for the alpha value and since we have a small amount of data, choosing a 0.1 value would put us in the situation in which we could do a type I error, which means that we could consider some variables significant even if they weren't, at the same time choosing a value of 0.01 would be too specific and so make a type 2 error.

Chi-Squared Analysis

The results with a 4 cluster split resulted in only one of our behavioral components to be statistically significant, for this reason, we chose to increase the number of **clusters** to **5** . This leads to better results and to the possibility of considering 3 variables as significant. These variables are the answers to the questions:

1. How often do you use your streaming service?
2. Where are you from?
3. Which is your area of study/business?

This means that we can say that the fact that people with these same characteristics belong to the same cluster is not randomic and that these variables can influence how respondents are positioned in the different groups formed with the dendrogram.

The second value that is important to be checked is the Cramer's V value, in this case all of our variables have a value around 0.17, this might mean that these variables have only a limited effect on how respondents are positioned in the different groups formed with the dendrogram.

When we do the Chi squared analysis we are analyzing behavioral or sociodemographic variables, we are in particular considering how correlated is the fact of being present in the first cluster is with the fact of for example being a male or having studied something or preferring movies rather than series. To do this we have to set an alpha value, which will be 0.05 because it's the most common one, and we will have to compare it to the p-value of the chi squared. If the p value is smaller than 0.05 we can say that this correlation is statistically significant.

T-Test on Core Components

Another analysis that we are doing on our clusters is the t-test which it's done on the core components. This will consider the mean of the total population and the mean of each cluster so we will know cluster components' preferences and opinions. And this is the function implemented by the macro command.

The results of the t-test include the t-value and the associated probability. The t-value represents the difference between the mean of the observations for the current cluster and the mean of the entire population, divided by the standard error of the difference. The probability represents the probability that the difference between the two means is due to chance, i.e., that there is no true difference between the two means. If the probability is below a certain level of significance, in our case 0.05, it can be concluded that there is a statistically significant difference between the two means.

The method used in this t test is Satterthwaite's test, which is used to perform a t-test for comparison between two means when the variances between the two populations are different.

This method performs a correction to the standard error that takes into account the differing variances between the two populations, and makes the test more accurate and reliable.

(The Satterthwaite approximation is **a formula used in a two-sample t-test for degrees of freedom**. It's used to estimate an "effective degrees of freedom" for a probability distribution formed from several independent normal distributions where only estimates of the variance are known)

The TTEST Procedure

Variable: new_1

CLUSTER	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		64	0.0217	0.6959	0.0870	-1.0000	1.0000
6		255	-0.00695	0.7729	0.0484	-1.0000	1.0000
Diff (1-2)	Pooled		0.0286	0.7582	0.1060		
Diff (1-2)	Satterthwaite		0.0286		0.0995		

CLUSTER	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		0.0217	-0.1521	0.1955	0.6959	0.5928	0.8428
6		-0.00695	-0.1023	0.0884	0.7729	0.7111	0.8465
Diff (1-2)	Pooled	0.0286	-0.1799	0.2372	0.7582	0.7035	0.8222
Diff (1-2)	Satterthwaite	0.0286	-0.1687	0.2260			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	317	0.27	0.7871
Satterthwaite	Unequal	105.53	0.29	0.7741

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	254	63	1.23	0.3223

This is the output of the t test for the first variable and the first cluster. The first two tables show descriptive statistics for the first cluster and the population (which is cluster 6). The statistics include the number of observations (N), mean, standard deviation, standard error of the mean, minimum, and maximum values, while the "Diff (1-2)" rows show the difference in mean between the two clusters, with the 95% confidence interval, with respect to the two methods.

The third table shows the results of two t-tests. The first test uses the "Aggregation" method and the second uses the "Satterthwaite" method. The "Valore t" (t value) column represents the difference between the means of the two clusters divided by the standard error of the difference.

The "Pr > |t|" (probability) column shows the probability that the difference between the two means is due to chance.

The fourth table shows the results of a test for equal variances between the two clusters. The "Metodo" column indicates the method being used (Folded F-test). The "Valore F" (F value) represents the ratio of variance between the two clusters. The "Pr > F" (probability) column shows the probability that the variances are equal.

In this case since the **"Pr > |t|" value is larger than 0.05**. We can conclude that for the first cluster, the characteristic of subjects belonging to it, of having a higher preference, on average of 0.29, of watching their streaming platform when they are with their family or friends **is not statistically significant**.

The following variables are the ones that present a statistically significant "Pr > |t|" value for each cluster.

CLUSTER 1: new_4(-5.8 o -0.5779), new_5(-7.01 o -0.6471), new_7(-3.55 o -0.3482), new_8 (4.8 o 0.4867)

CLUSTER 2: new_1(-14.8 o -0.9359), new_2(-3.46 o -0.4789), new_5(8.64 o 0.6637), new_6(7.63 o 0.5651), new_7((6.99 o 0.5791), new_8((8.61 o 0.9420), new_9(12.47 o 0.8599)

CLUSTER 3: new_1(-2.01 o -0.1961), new_2(-5.96 o -0.4564), new_3(-2.71 o -0.2398), new_4(3.63 o 0.2287), new_5(2.74 o (0.2117)

CLUSTER 4: new_1(3.93 o 0.3901), new_2(4.17 o 0.4056), new_3((2.18 o 0.2135), new_4(2.28 o 0.1683), new_6(-2.87 o -0.2785), new_8(-8.65 o -0.6503), new_9(-5.01 o -0.4047)

CLUSTER 5: new_1(2.10 o 0.2976), new_2(2.34 o 0.3420), new_3 (2.47 o 0.3592), new_4 (2.97 o 0.2853), new_5 (3.92 o 0.4084), new_6 (5.01 o 0.4434), new_7(7.08 o 0.5473), new_8(-8.35 o -0.7479), new_9 (4.80 o 0.5831).

Cluster Description

To describe our cluster we will have to put together information we gained in both chi squared analysis and t test. From the chi squared analysis we could consider that only the frequency with which our population watches streaming platforms, geographical origins and the area of study or work are statistically significant aspects from behavioral and socio-demographic variables, and even if the Cramer's V value is low we can anyway observe some characteristics of our clusters.

Talking about Geographical Origins, we can notice that in **CLUSTER 2** all the respondents come from America and none from Europe or Asia, while **CLUSTER 4** has the majority of components coming from Europe. **CLUSTER 5** is also mainly populated by Americans with only a few subjects from Europe and Asia.

With regard to area of study and business we can notice that **CLUSTER 1** is mainly populated by subjects that work or study in the Management and Administration sector, same for **CLUSTER 2** in which we also notice a high percentage of Science and Technology. Education and Arts and Humanities mainly appear in **CLUSTER 4** with a good component of Science and Technology workers. In **CLUSTER 5** there's again a large portion of subjects in M&A and a large portion as well of subjects that work in areas that are out of the ones that we choose to consider in the survey.

Considering the frequency with which the service is used, data is distributed in a homogenous way, the only exception is that **CLUSTER 4** is mainly composed by people who use the service less than 4 times a week.

From the Core Components side we can assert that:

- CLUSTER 1 **"Just give me something to watch"** is composed of subjects who work in the Management and Administration area and these subjects have significantly lower preferences with regard to quality, quantity, discovering new contents and subscribing to a service to have something to watch while performing other activities.
- CLUSTER 2 **"American committed learning"** is composed of exclusively people from America, these subjects mainly are working in the M&A and Scientific sector and their opinion tends to be highly against the fact of watching netflix with loved ones compared to the total population and to relax, while preferring to choose between a large amount of contents, watching contents in different languages, discovering new contents, have something to watch while they are eating or exercising and consider the easy usability of the platform with a strong preference regarding the last one.
- CLUSTER 3 is a cluster in which behavioral and socio-demographic components don't show particularly specific characteristics and even preferences regarding core components don't show highly relevant differences from the population's mean with respect to other clusters.

- CLUSTER 4 **“European social enrichment”**is the one in which the majority of the components come from Europe and this is an important information since the majority of our population comes from America, this cluster is also populated with the majority of the subjects working in Education, Arts and Humanities and Science and Technology and these are the users that tend to use their platform less frequently. The opinions of people in this cluster tends to be strongly against watching movies or series while performing other activities and against choosing the service considering how easy it is to use. They tend to use it to relax and with friends and look for the quality of the contents.
- CLUSTER 5 **“The demanding American”** has a majority of American people who enjoy watching contents in different languages, discovering new contents and they are not interested in watching contents while performing other activities. With respect to other opinion components these subjects tend to have higher preferences compared to the whole population.

Conclusions

Based on the findings of this research, the author recommended investing in the creation of new content that is specifically designed to meet the needs of these consumer segments. This can include instructional or educational content that is perfect for those who want to learn while they exercise or eat, as well as content that is designed to be watched and discussed together as a group. This type of content will not only meet the demands of these consumer segments, but it will also provide Netflix with a competitive advantage, as it will offer content that is not available on other platforms.

Market research and consumer insights have revealed that there are four key consumer segments that are seeking unique and diverse content offerings from Netflix. These segments are: (1) people who want to watch entertainment while engaging in other activities (2) Americans who seek new and educational content to learn new things, (3) Europeans watch movies and later discuss them, and (4) Americans who demand high-quality Netflix content. These consumers are searching for content that is not only entertaining but also interactive and educational.

For the first segment, it is recommended that Netflix invest in visually stimulating shows that do not require a lot of mental engagement, such as "The Great British Bake-Off" or "Neo Tokyo." Shows like the suggested ones tend to be visually appealing and don’t require to much context to know what is happening in the scene. These types of shows are perfect for passive consumption while engaging in other activities.

For the second segment, it is proposed that Netflix invest in more in-depth educational shows like "Space" or "Explained," which requires the audience to be engaged for longer periods of time. These shows provide new information and perspectives, making them ideal for those who are seeking to learn something new.

For the third segment, consisting mainly of European audiences seeking original content in various languages, it is recommended that Netflix invest in content that is culturally enriching and provides insight into the social issues of other cultures, such as "Kalifat" or "La Casa de las Flores."

Finally, for the fourth segment, Netflix may want to continue investing in high-quality content that appeals to American audiences. Shows such as "House of Cards" or "Dahmer" that revolve around recent controversy, political topics or racial crimes, while using state of the art cinematography and top-listed Hollywood actors, tend to resonate with American audiences and contribute to Netflix's popularity.

References:

1. **Ford, William. *Numerical linear algebra with applications: Using MATLAB*. Academic Press, 2014.**
2. **https://stats.stackexchange.com/**
3. **Furio Camillo, Valentina Adorno. "PLS-PM in opinion surveys when respondent minds generate size-effect (involvement axis) in the data".**
4. **SAS OnlineDoc: Version 7-1,(Chapter 47: The PRINCOMP Procedure)**