Western Governors University
School of Technology, College of IT
Master of Science, Data Analytics

D205 Performance Assessment

which factors are the highest contributors to customer churn?

Tyson Biegler
Student ID: 012170282
D206 - Data Cleaning
June 19, 2024

Part I: Research Question

A: *Description of the research question.*

The data dictionary states that retaining customers is often the number 1 goal for telecom providers. This particular company is interested in predicting which customers are at high risk for churn. To determine the individual customers at risk, I first need to determine the factors that lead to churn. I will seek to discover **which factors are the highest contributors to customer churn.**

B: *Describing all the variables.*

This data set consists of 52 variables about customer demographics, subscription status of various services, and a range of responses to an 8-question survey question.

The data dictionary states that this CSV consists of 50 columns. The additional two columns displayed in the CSV output data below are due to the 1st column being automatically added as a row names column, and the "Lat, Lng" column seems to have been separated into two individual columns.

This table displays the column name, data type, an example of the data, and a brief description of the data in the columns.

Column Name	**Data Types**	**Example**	**Description**
...1	Index	“1”	This is the row names column that was added automatically after using read_csv
CaseOrder	Index	“1”	CaseOrder is an index column.
Customer_id	Qualitative, col_character()	“K409198”	This column serves as a unique customer identifier
Interaction	Qualitative, col_character()	“aa90260b-4141-4a24-8e36-b04ce1f4f77b”	This is a unique identification number for customer transactions.
City	Qualitative, col_character()	“Point Baker”	This is the customer's city.
CaseOrder	Quantitative, col_character()	“AK”	This is the abbreviated state the customer lives in.
County	Qualitative, col_character()	“Prince of Wales-Hyde”	The customer's county.
Zip	Qualitative, Numeric	“99927”	The customer's zip code.
Lat	Qualitative, Numeric	“56.25100”	This is the latitude part of the customer's GPS coordinates.
Lng	Qualitative, Numeric	“-133.37571”	This is the longitude part of the customer's GPS coordinates.

Population	Qualitative, Numeric	“38”	The population within 1 mile of the customer
Area	Qualitative, col_character()	“Urban”	The type of area in which the customer lives, based on census data.
Timezone	Qualitative, col_character()	"America/Sitka"	The timezone of the customer’s residence.
Job	Qualitative, col_character()	"Environmental health practitioner"	The customer or invoiced person’s job
Children	Quantitative, Numeric	“1”	The number of kids in the customer's household.
Age	Quantitative, Numeric	“68”	The age of the customer when they signed up for services.
Education	Qualitative, col_character()	“Master’s Degree”	The customer’s highest level of education.
Employment	Qualitative, col_character()	"Part-Time"	The employment status of the customer.
Income	Quantitative, Numeric	“28561.99”	This is the customer’s annual income.
Marital	Qualitative, col_character()	"Widowed"	The customer’s marital status.
Gender	Qualitative, col_character()	“Male”	The customer’s gender.
Churn	Qualitative, col_character()	“No”	This column states if the customer has churned within the past month.
Outage_sec_perweek	Quantitative, Numeric	“6.972566”	This is the average number of outages, in seconds, in the customer’s neighborhood.
Email	Quantitative, Numeric	“10”	The total number of emails sent to the customer over the last year.
Contacts	Quantitative, Numeric	“0”	The number of times a customer had to contact customer support.
Yearly_equip_failure	Quantitative, Numeric	“1”	The number of times a customer’s equipment needed to be replaced due to failure.
Techie	Qualitative, col_character()	“No”	Whether or not the customer would consider themself to be technologically inclined.
Contract	Qualitative, col_character()	“One year”	The customer’s contract duration.

Port_modem	Qualitative, col_character()	"Yes"	This question determines if the customer has a portable modem or not.
Tablet	Qualitative, col_character()	"Yes"	Does the customer own any type of tablet?
InternetService	Qualitative, col_character()	"Fiber Optic"	The type of internet the customer is subscribed to
Phone	Qualitative, col_character()	"Yes"	Does the customer subscibe to phone service?
Multiple	Qualitative, col_character()	"No"	Does the customer have multiple phone lines?
OnlineSecurity	Qualitative, col_character()	"Yes"	Does the customer subscribe to online security?
OnlineBackup	Qualitative, col_character()	"Yes"	Does the customer subscribe to an online backup service?
DeviceProtection	Qualitative, col_character()	"No"	wheather or not the customer subscribed to the device protection service.
TechSupport	Qualitative, col_character()	"No"	Whether or not the customer subscribes to technical support services
StreamingTV	Qualitative, col_character()	"No"	Does the customer subscribe to the streaming TV service?
StreamingMovies	Qualitative, col_character()	"Yes"	Does the customer subscribe to streaming movie services?
PaperlessBilling	Qualitative, col_character()	"Yes"	Does the customer have paperless billing setup?
PaymentMethod	Qualitative, col_character()	"Credit Card (automatic)"	The column records the customer's payment method
Tenure	Quantitative, Numeric	"6.795513"	The length of time since the customer became a customer. Recorded in months.
MonthlyCharge	Quantitative, Numeric	"171.44976"	The average monthly charge for services.
Bandwidth_GB_Year	Quantitative, Numeric	"904.5361"	The average amount of data used by the customer in a year. Recorded in GB
item1	Quantitative, Numeric	"5"	Survey response to 'Timely response'
item2	Quantitative, Numeric	"5"	Survey response to 'Timely fixes'
item3	Quantitative, Numeric	"5"	Survey response to 'Timely replacements'
item4	Quantitative, Numeric	"3"	Survey response to 'Reliability'

item5	Quantitative, Numeric	"4"	Survey response to 'Options'
item6	Quantitative, Numeric	"4"	Survey response to 'Respectful responses'
item7	Quantitative, Numeric	"3"	Survey response to 'Courteous exchange'
item8	Quantitative, Numeric	"4"	Survey response to 'Evidence of active listening'

Part II: Data-Cleaning Plan

C: *Plan for cleaning the data.*

To assess the quality of the data set, I did the following:

1. Checked for missing or Na values, outliers, and duplicates using histograms, box plots, and functions like duplicated() and is.na()
2. Determined whether the re-expression of variables was needed by looking at the data type and an example of the data.
3. I looked for inconsistencies between the data dictionary and the data set by inspecting the data types of each variable and its corresponding definition in the dictionary. I checked for misspellings by checking for unique values.
4. I added an index column (Larose & Larose, 2019) and removed the row names column.

C2: *Justification of cleaning approach*

Because my question seeks to determine the best-contributing factors to customer churn, I will need to address the issues noted above (NA values, duplicates, misspellings, outliers, and incorrect data types) before I can begin to make any assumptions about the data.

Some columns have large amounts of missing values and outliers, so appropriately handling these will be essential to creating meaningful conclusions. Similarly, many of the columns in this data set are categorical. They would be of more use as another data type that can have statistical methods applied and visualized in charts to help explain their data. Gender, for example, is currently a character type but would be far better utilized as a factor. Lastly, there doesn't appear to be a consistent naming convention among the variables, so I will need to address that too.

C3: *Justification of programming language*

According to an article on the Datacamp blog (Canales Luna, 2022), python has outranked R in popularity in recent years. However, I utilized R programming language for this assessment via R-studio. I chose to use R for this project because I have some experience with Python. Still, I want to become more familiar with R. The WGU data cleaning module section 1, lesson 1 states, "A good analyst knows either Python or R, but a great data analyst knows both." (WGU, n.d.)

In addition, R is a statistical programming language. Therefore, I can access packages specifically created to easily handle complex statistical tasks, such as the principal component analysis (PCA).

C4: *Annotated code script file*

The uploaded files will include my R script file and the cleaned CSV file.

I will use' Tidyverse' packages to access **ggplot2**, **dplyer, plyr**, and **stringr**. I will also use **factoextra** for the PCA.

Part III: Data Cleaning

D1 - D5:

Dealing with NA values.

I found eight columns with NA values. They include ‘Children,’ ‘Age,’ ‘Income,’ ‘Techie,’ ‘Phone,’ ‘TechSupport,’ ‘Tenure,’ and ‘Bandwidth_GB_Year.’

‘Children,’ ‘Income,’ ‘Tenure,’ and ‘Bandwidth_GB_Year’ all had their Na values imputed with the median. I chose to impute with the median because ‘Children’ and ‘Income’ were both right-skewed, and ‘Tenure’ and ‘Bandwidth_GB_Year’ were bimodal. However, the mean was used for imputation on the ‘Age’ variable due to having had a uniform distribution.

I utilized imputation of the mode for the categorical variables like ‘Phone,’ ‘TechSupport,’ and ‘Techie.’

Dealing with outliers.

I determined having more than 6 children would be defined as an outlier based on a boxplot diagram. I used imputation to replace the outliers with the mean.

The income variable contained a large portion of outliers, equaling 7.58%. I determined the outliers by looking for the value of the upper whisker in a boxplot diagram. By doing this, I determined that any income value of $78272.96 and above would be considered an outlier. I retained the outliers in the income column because the range between the median and the higher end of the income range was so great that imputation would skew the results, and the income column would no longer be accurate. The median ‘Income’ value is $33186.80. So if the outliers are defined as any value greater than $78272.96, then the 758 values from $78272.96 to 258900.70 would be imputed to being just a fraction of the actual incomes of these customers. Similarly, ‘Population’ contained 9.37% of its values as outliers. I determined the outliers by looking at the value of the upper whisker.

The Email variable contained outliers that were under four and over 20. This can be seen by viewing a boxplot. I imputed the mean to deal with the outliers because the distribution before imputation was normal. I used the mean because the outliers were not so extreme as to require a median imputation.

‘Contacts,’ ‘Yearly_equip_failure,’ and ‘Children’ were right-skewed and contained minimal outliers. ‘Contacts’ only contained eight outliers, which represents 0.08%. I imputed the outliers with the median due to the right-skewed distribution. Likewise, ‘Yearly_equip_failure’ contained less than 1% of outliers, and ‘Children’ contained 4.51%. These, too, were imputed using the median due to the right-skewed distribution.

Other data quality issues.

I discovered that ‘Zip’ was a numeric type but needed to be changed to a character due to automatically dropping leading zeros as numeric. First, I converted Zip to a character and then trimmed the white space that may have existed from the front and back. I added a zeros padding to the Zip variable to a width of 5 digits.

As the data dictionary mentions, ‘Gender’ should include ‘Male,’ ‘Female,’ and ‘Nonbinary.’ However, at the moment, ‘Gender’ contains ‘Male,’ ‘Female,’ and ‘Prefer not to answer.’ After updating the values for ‘Gender,’ I converted the data type to a factor because it only has three levels. It is much more helpful for building charts and conducting statistical operations as a factor rather than a character type. In addition, I renamed the value ‘bank transfer(automatic)’ from the ‘PaymentMethod’ column to ‘bank (automatic bank transfer)’ to match the data dictionary.

I used ordinal encoding to convert the education types from text to a number relevant to the degree level. For example, ‘No Schooling Completed’ is equivalent to 12, the numeric value, and ‘Doctorate Degree’ is equivalent to 1, the highest value. I stored the results in a new column named ‘Edu_num’ and converted it to a numeric type.

The names of the survey responses were not intuitive, being named ‘item1’ though ‘item8.’ For clarity, I renamed the survey response columns to reflect their definition in the data dictionary. The following are the new names for these survey response columns:

Item1: Timely_response
Item2: Timely_fixes
Item3: Timely_replacements
Item4: Reliability
Item5: Options
Item6: Respectful
Item7: Courteous
Item8: Active_listening

I converted the Yes, and No values from 'Churn,' 'Techie,' 'Port_modem,' 'Tablet,' 'Phone,' 'Multiple,' 'OnlineSecurity,' 'OnlineBackup', 'DeviceProtection,' 'TechSupport,' 'StreamingTV,' 'StreamingMovies,' and 'PaperlessBilling' to a logical type as TRUE or FALSE. Doing this clarifies if the customer is subscribed to any particular service.

I rounded MonthlyCharge, and Income to two decimal points because they represent monetary values. However, I also rounded Tenure to two decimal places because it measures a time frame. Keeping Tenure to at least two decimal places ensures clarity for database users.

I determined that for simplicity, it makes sense to have the average outage in seconds for a neighborhood measured in whole numbers as it seems unnecessarily specific to measure the seconds up to hundredths or even thousandths of a second. Due to my experience installing and maintaining telecom equipment, the bandwidth in GB is typically represented by whole numbers. Therefore Bandwidth_GB_Year and Outage_sec_perweek were rounded to the nearest whole number, whereas

Age was rounded down to the nearest whole number. Likewise, Age was rounded down to the nearest whole number. Analyzing age in groups rather than individually is usually more helpful or insightful when dealing with Ages. This means that representing customer ages to multiple decimal places seems unhelpful.

Due to the above reasoning, I added an ‘Age_groups’ column that groups the customers into six different age groups. Due to the large number of customers, it will be more helpful to know, for example, that most customers are between the ages of 45 and 54 than it is to know how many customers are precisely 26 or 52. If the number of customers was far less, then knowing the exact ages might be helpful, but not at this scale. Lastly, I converted the age groups to a factor.

I created a similar grouping of incomes. The ‘Income_groups’ column increments by $20,000 and contains 13 groups. I converted these to a factor as well. Regarding the ability to extract meaningful insights from the data, it is more beneficial to know that most customers' income falls between $20,000 and $40,000 than knowing how many customers make exactly $62,254.56 per year, for example.

Lastly, I added a column that contains the sum of all survey scores for a customer. Customers with a higher score appear to be more satisfied with the services they receive. Likewise, customers with a low score are less satisfied. The reason for this column is to identify customers who are unsatisfied with services quickly and at a glance. Determining the value that signifies what is statistically considered ‘unsatisfied’ would need to be calculated to create any significant insights from these columns.

Due to having assigned a column earlier as the index, I removed the column named ‘... 1’ because it was automatically generated as a row names column and was not needed.

The last data quality issue I addressed is the naming convention inconsistencies. All the columns used a different naming convention in that they contained ‘_’ or a mixture of upper and lower case letters. I converted the columns to the same naming convention for readability to remedy this. Instead of InternetService, we have Internet_service, or Outage_sec_perweek, which is now Outage_sec_per_week.

D6 - D7: *Summarizing the limitations of the data cleaning process.*

One limitation that could be a topic of concern is the fact that the term ‘non-binary’ can be an exclusive term of identification like ‘male’ or ‘female,’ or it can mean, according to Webster's dictionary, ‘not restricted to two things or parts’ (Merriam-Webster, 2024). By that definition, anyone who does not consider themselves male or female exclusively would be regarded as non-binary. However, not everyone who falls under the umbrella term ‘non-binary’ would consider themselves to be ‘non-binary.’ If changing the term ‘prefer not to answer’ to ‘non-binary’ increases inclusivity to others outside of the gender binary, then perhaps using the term ‘Other’ would feel more inclusive.

Another limitation of the cleaning process involves the ‘Interaction’ variable. The data dictionary states that it is a unique identifier relating to transactions, tech support, and sign-ups. However, how this identifier is created or used is unclear, so the cleaning process for that variable remains unclear.

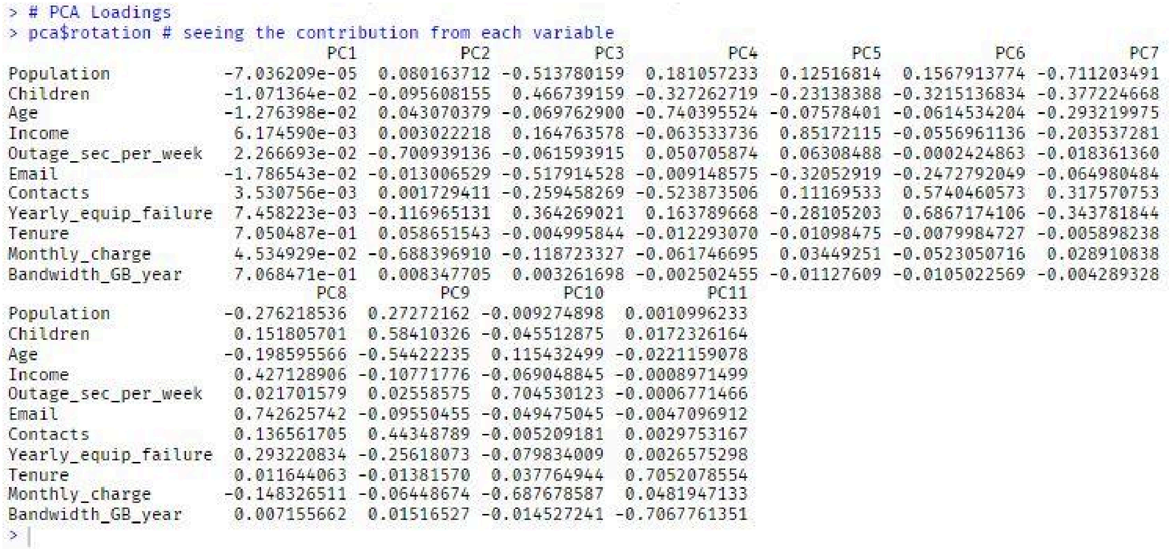
Lastly, there is a lack of clarity in the measurement of GB usage and seconds of outage. In my cleaning process, I rounded these variables to two seminal places. Without knowing the measurement scale, it is difficult to tell whether tracking seconds or GB to the fourth or fifth decimal point will create meaningful insights. For example, is there a company goal in which +0.0025 seconds of outage could be significant? Or is this seemingly minuscule amount of time considered negligible by the company or the telecom industry at large? These questions could likely be answered if an analyst, in this situation, had access to stakeholders.

E: *Applying Principal Component Analysis (PCA)*

E1: *Identifying the number of principal components*

I applied principal component analysis to 'Population,' 'Children,' 'Age,' 'Income,' 'Outage_sec_per_week,' 'Email,' 'Contacts,' 'Yearly_equip_failure,' ‘Tenure,’ ‘Monthly_charge,’ and ‘Bandwidth_GB_year’ resulting in 11 principal components. PCA requires quantifiable numeric data. These 11 variables are all of the continuous and discrete variables in the data set.

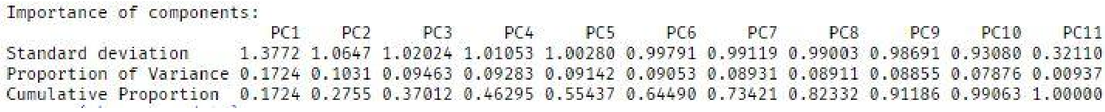
The following is the PCA loadings matrix.



```
> # PCA Loadings
> pca$rotation # seeing the contribution from each variable
                               PC1          PC2          PC3          PC4         PC5           PC6          PC7
Population           -7.036209e-05  0.080163712 -0.513780159  0.181057233  0.12516814  0.1567913774 -0.711203491
Children             -1.071364e-02 -0.095608155  0.466739159 -0.327262719 -0.23138388 -0.3215136834 -0.377224668
Age                  -1.276398e-02  0.043070379 -0.069762900 -0.740395524 -0.07578401 -0.0614534204 -0.293219975
Income                6.174590e-03  0.003022218  0.164763578 -0.063533736  0.85172115 -0.0556961136 -0.203537281
Outage_sec_per_week   2.266693e-02 -0.700939136 -0.061593915  0.050705874  0.06308488 -0.0002424863 -0.018361360
Email                -1.786543e-02 -0.013006529 -0.517914528 -0.009148575 -0.32052919 -0.2472792049 -0.064980484
Contacts              3.530756e-03  0.001729411 -0.259458269 -0.523873506  0.11169533  0.5740460573  0.317570753
Yearly_equip_failure  7.458223e-03 -0.116965131  0.364269021  0.163789668 -0.28105203  0.6867174106 -0.343781844
Tenure                7.050487e-01  0.058651543 -0.004995844 -0.012293070 -0.01098475 -0.0079984727 -0.005898238
Monthly_charge        4.534929e-02 -0.688396910 -0.118723327 -0.061746695  0.03449251 -0.0523050716  0.028910838
Bandwidth_GB_year     7.068471e-01  0.008347705  0.003261698 -0.002502455 -0.01127609 -0.0105022569 -0.004289328
                              PC8         PC9         PC10          PC11
Population           -0.276218536  0.27272162 -0.009274898  0.0010996233
Children              0.151805701  0.58410326 -0.045512875  0.0172326164
Age                  -0.198595566 -0.54422235  0.115432499 -0.0221159078
Income                0.427128906 -0.10771776 -0.069048845 -0.0008971499
Outage_sec_per_week   0.021701579  0.02558575  0.704530123 -0.0006771466
Email                 0.742625742 -0.09550455 -0.049475045 -0.0047096912
Contacts              0.136561705  0.44348789 -0.005209181  0.0029753167
Yearly_equip_failure  0.293220834 -0.25618073 -0.079834009  0.0026575298
Tenure                0.011644063 -0.01381570  0.037764944  0.7052078554
Monthly_charge       -0.148326511 -0.06448674 -0.687678587  0.0481947133
Bandwidth_GB_year     0.007155662  0.01516527 -0.014527241 -0.7067761351
> |
```

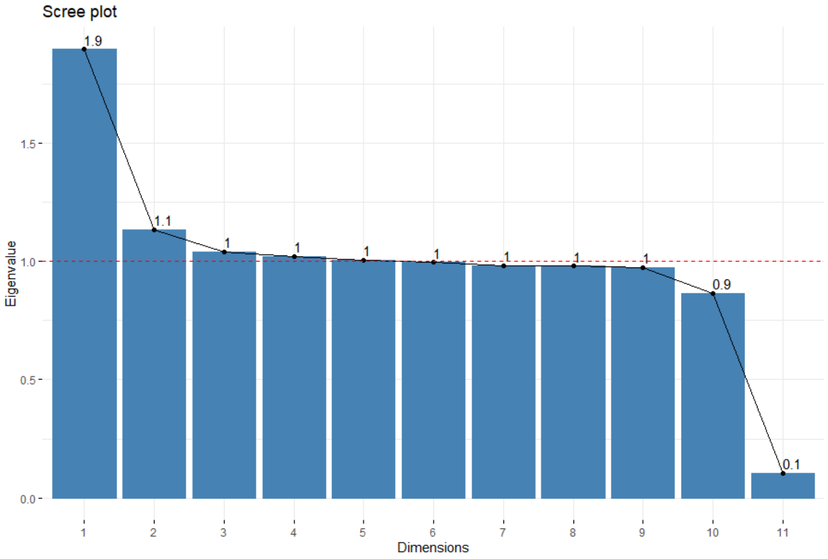
E2: *Justifying the reduced PCA components.*

The elbow method indicates that the first two components, in this case, contribute the most variance. However, these two components only account for a cumulative variance of 27.55% based on the **summary(PCA)** results shown below. Because PCA reduces the dimensionality of the data while maintaining as much variance as possible, I have chosen the Kaiser method to determine which principal components to select. The Kaiser method seems more appropriate in this situation due to the low cumulative variance resulting from the elbow method.

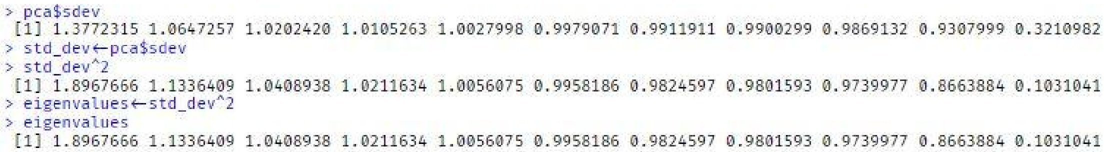


```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     1.3772 1.0647 1.02024 1.01053 1.00280 0.99791 0.99119 0.99003 0.98691 0.93080 0.32110
Proportion of Variance 0.1724 0.1031 0.09463 0.09283 0.09142 0.09053 0.08931 0.08911 0.08855 0.07876 0.00937
Cumulative Proportion  0.1724 0.2755 0.37012 0.46295 0.55437 0.64490 0.73421 0.82332 0.91186 0.99063 1.00000
```

One potential issue with the Kaiser method is that all principal components above one eigenvalue should be selected. The Screeplot below shows that PC1 - PC9 are all greater than or equal to one eigenvalue. Resulting in a cumulative variance of 91.186%. However, when eigenvalues are calculated manually, the actual principal components above one eigenvalue are PC1 through PC5, accounting for only 55.437 % variance. Therefore, I have chosen to apply the Kaiser method using the results from the scree plot to capture the most variance while selecting only nine of the eleven principal components.



The calculated eigenvalues mentioned above are displayed below. First, I saved the standard deviations from the PCA function as ‘std_dev.’ Then I squared the standard deviations resulting in the manually calculated eigenvalues for each principal component (StatQuest with Josh Starmer, 2017).



```
> pca$sdev
 [1] 1.3772315 1.0647257 1.0202420 1.0105263 1.0027998 0.9979071 0.9911911 0.9900299 0.9869132 0.9307999 0.3210982
> std_dev←pca$sdev
> std_dev^2
 [1] 1.8967666 1.1336409 1.0408938 1.0211634 1.0056075 0.9958186 0.9824597 0.9801593 0.9739977 0.8663884 0.1031041
> eigenvalues←std_dev^2
> eigenvalues
 [1] 1.8967666 1.1336409 1.0408938 1.0211634 1.0056075 0.9958186 0.9824597 0.9801593 0.9739977 0.8663884 0.1031041
```

E3: *How would an organization benefit from PCA?*

The data dictionary suggests that predicting customer churn is often a top priority for many telecom companies. PCA could benefit this telecom company by assisting analysts in identifying underlying factors by reducing the dataset's dimensions into just a few components. This reduction allows analysts to uncover relationships and patterns within the data. In this specific dataset, PCA revealed a strong relationship between 'Tenure' and 'Bandwidth_GB_per_year,' significantly influencing PC1. Therefore, customers with higher scores in PC1 are likely to exhibit similar patterns in tenure and bandwidth consumption. If these customers are churning at a high rate, addressing these factors could mitigate the churn rate.

Part IV. Supporting Documents

F: The Panapto video link will be included with the submitted files.

G. Web sources

Code Sources:

1. cut. (n.d.). In R Documentation. Retrieved June 20, 2024, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut

2. ns-dblcolon. (n.d.). In R Documentation. Retrieved June 30, 2024, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/ns-dblcolon

3. sprintf. (n.d.). In Statistics Globe. Retrieved June 30, 2024, from https://statisticsglobe.com/sprintf-r-function-example

4. trimws. (n.d.). In R Documentation. Retrieved June 27, 2024, from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/trimws

Other Sources:

1. Larose, C. D., & Larose, D. T. (2019). Data Science Using Python and R (p. 33). Wiley

2. Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley. pg 33 7/2/24

3. Canales Luna, J. (2022, February 23). Python vs. R for Data Science: What's the Difference? DataCamp. Retrieved June 15, 2024, from https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

4. Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R (p. 33). Wiley

5. Merriam-Webster. (2024). Nonbinary. Merriam-Webster.com. Retrieved July 7, 2024, from https://www.merriam-webster.com/dictionary/nonbinary

6. StatQuest with Josh Starmer. (2017, October 12). PCA main ideas simply explained [Video]. YouTube. https://www.youtube.com/watch?v=FgakZw6K1QQ

7. Western Governors University (WGU). (n.d.). Welcome to Data Cleaning. Retrieved June 16, 2024, from Data Cleaning | WGU-CGP-OEX

8. WGU Courseware. (2024). Data Cleaning. In Lesson 7: How to Perform PCA in R. Retrieved June 27, 2024, from https://apps.cgp-oex.wgu.edu/wgulearning/course/course-v1:WGUx+OEX0026+v02/block-v1:WGUx+OEX0026+v02+type@sequential+block@2b5f23c5dad64357b352728993788677/block-v1:WGUx+OEX0026+v02+type@vertical+block@802622e6308a4ba9957be2a69aee38aa