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:**

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:**

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, as well as 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:**

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**:

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:**

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).

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

# Part III: Data Cleaning

## **D.**

### 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, I can say that 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.

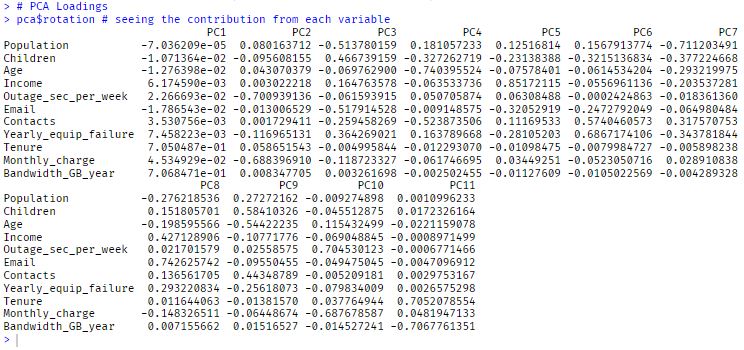
Lastly, I added a column that contains the sum of all survey scores for a customer. Customers with a higher score appear to be more satisfied with the services they receive. Likewise, customers with a low score would appear to be less satisfied. The reason for this column is to identify customers who 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.

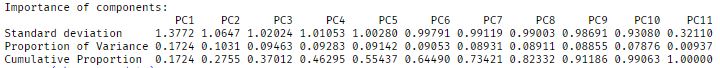
## 

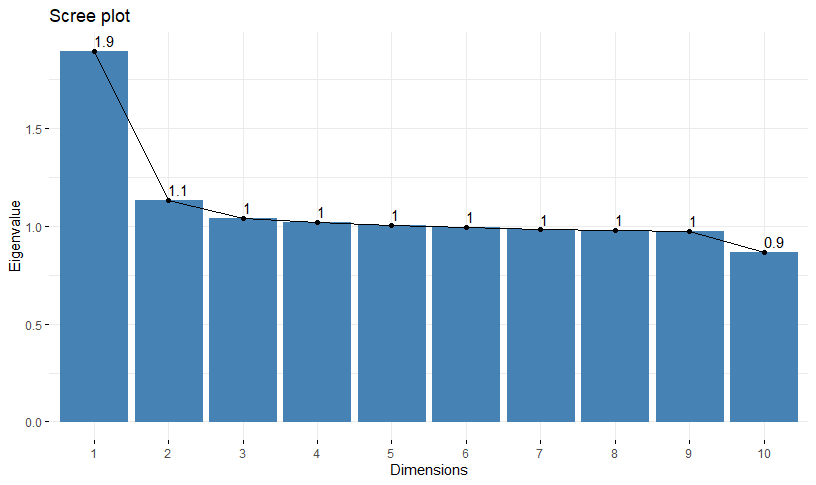
### Principal Component Analysis (PCA)

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.   


* 1. Justify the reduced number of the principal components and include a screenshot of a scree plot.

asdfasdf





Describe how the organization would benefit from the use of PCA.

asdfasdf

**Part IV. Supporting Documents**

## [Panopto video](https://wgu.hosted.panopto.com/Panopto/Pages/Home.aspx)

## **G.** Web sources

**Code sources:**

1. <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut> 6/20/24444
2. <https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/trimws> 6/27/24
3. <https://statisticsglobe.com/sprintf-r-function-example> 6/30/24
4. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/ns-dblcolon 6/30/24
5. https://apps.cgp-oex.wgu.edu/wgulearning/course/course-v1:WGUx+OEX0026+v02/block-v1:WGUx+OEX0026+v02+type@sequential+block@70ce00423e2f4a88b19b30b6b09078c1/block-v1:WGUx+OEX0026+v02+type@vertical+block@300252c5cde84d709622d8b659f5630d 7/2/24

**Other sources:**

1. 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>
2. Western Governors University (WGU). (n.d.). *Welcome to* *Data Cleaning*. Retrieved June 16, 2024, from [Data Cleaning | WGU-CGP-OEX](https://apps.cgp-oex.wgu.edu/wgulearning/course/course-v1:WGUx+OEX0026+v02/block-v1:WGUx+OEX0026+v02+type@sequential+block@ec44f144ae394de79027cd6d2a17bb72/block-v1:WGUx+OEX0026+v02+type@vertical+block@bf0e3787d10542fd9d0945eebae4caa3)
3. 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>
4. Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley. pg 33 7/2/24