

D206 – DATA CLEANING

Performance Assessment

Data Cleaning the Telecommunications Churn Scenario Dataset

Mike Hallenstein

WGU - MSDA

[fhallen@wgu.edu](mailto:fhallen@wgu.edu)

Student ID: 010636300

**Part I**

# Research Question

For the performance assessment I chose the Telecommunications Churn dataset because I had used it previously in D205 and felt comfortable with it. The data dictionary for the Telecommunications Churn dataset defined “Customer ‘churn’ as the percentage of customers who stopped using a provider’s product or service”. The scenario directs the data analyst to clean the data to prepare it for further exploration, to identify trends in the data, and to be able to “compare key metrics. The Telecommunications Churn scenario provided that the industry can experience churn rates “as high as 25%” and that “It costs 10 times more to acquire a new customer than to retain an existing one.” The raw data used in this assignment came from the D206 Definitions and Data Files https://web5.wgu.edu/aap/content/d206-ema.html (‘churn\_raw\_data.csv’).

After looking at the scenario the research question I decided on is- What are the probable root causes of customer churn?

# Required Variables and Examples

The dataset for the chosen scenario is the churn\_raw\_data.csv and Data Cleaning Churn Data Consideration and Dictionary files. The churn\_raw\_data.csv file contains data formatted into 52 columns and 10,000 rows. The unadjusted variables from the unprocessed data set are as follows:

*Unnamed:0, CaseOrder, Customer\_id, Interaction, City, State, County, Zip, Lat, Lng, Population, Area, Timezone, Job, Children, Age, Education, Employment, Income, Marital, Gender, Churn, Outage\_sec\_perweek, Email, Contacts, Yearly\_equip\_failure, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod, Tenure, MonthlyCharge, Bandwidth\_GB\_Year, item1, item2, item3, item4, item5, item6, item7, item8.*

The dataset contains 10,000 rows of records of customer data. This data includes demographic data such as geographic location, income, and number of children in the household. The data set includes service-related variables and contains responses to a customer survey. Most relevant to the research question in Part A is whether a customer has kept or left the company’s services during the month which is captured in the “churn” variable.

The variables provided in the data set will be used to perform data science tasks such as description, estimation, classification, clustering, prediction, and association (Larose & Larose, 2019) pertaining to customer “churn” , and include:

* + Customer services such as phone, internet, multiple lines, streaming services used, online security, and online backup.
  + Customer account information such as tenure, payment method,

and data usage year-to-date.

* + Customer demographic information such as age, education, employment, income, marital status, and gender.
  + Responses from a customer survey with importance weightings on a scale of 1-8.

The data set includes both numerical data, both discrete and continuous, and categorical data including ordinal and nominal.

# The Data:

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

The data consists of both numeric and string data, the missing data consists of ‘NA’ or “NaN’ values where either through data entry or collection error the information is missing. To clean the data set I’ll need to first identify where missing values are located and then depending on whether data is missing at random, and type of data, determine how to replace or impute the missing values. I will use techniques described in the

WGU provided Textbook and datacamp Python courses to accomplish the data cleaning.

Definition of Variables in Dataset

* CaseOrder: Placeholder variable to preserve the original order of the data
* Customer\_id: Unique customer ID
* Interaction, UID: Unique IDs related to customer transactions, technical support, and sign-ups
* City, State, County, Zip: Customer's city, state, county, and zip code of residence
* Lat, Lng: GPS coordinates of customer residence
* Population: Population within a mile radius of the customer's location
* Area: Type of area (rural, urban, suburban)
* TimeZone: Time zone of customer's residence
* Job: Customer's reported job or the job of the invoiced person
* Children: Number of children in the customer's household
* Age: Age of the customer
* Education: Highest degree earned by the customer
* Employment: Employment status of the customer
* Income: Annual income of the customer
* Marital: Marital status of the customer
* Gender: Customer's self-identified gender

Other variables in the dataset include:

* Churn: Indicates whether the customer discontinued service within the last month (yes or no)
* Outage\_sec\_perweek: Average number of seconds per week of system outages in the customer's neighborhood
* Email: Number of emails sent to the customer in the last year (marketing or correspondence)
* Contacts: Number of times the customer contacted technical support
* Yearly\_equip\_failure: Number of times the customer's equipment failed and had to be reset or replaced in the past year
* Techie: Indicates whether the customer considers themselves technically inclined (yes or no)
* Contract: Contract term of the customer (month-to-month, one year, two years)
* Port\_modem: Indicates whether the customer has a portable modem (yes or no)
* Tablet: Indicates whether the customer owns a tablet (yes or no)
* InternetService: Customer's internet service provider (DSL, fiber optic, None)
* Phone: Indicates whether the customer has a phone service (yes or no)
* Multiple: Indicates whether the customer has multiple lines (yes or no)
* OnlineSecurity: Indicates whether the customer has an online security add-on (yes or no)
* OnlineBackup: Indicates whether the customer has an online backup add-on (yes or no)
* DeviceProtection: Indicates whether the customer has device protection add-on (yes or no)
* TechSupport: Indicates whether the customer has a technical support add-on (yes or no)
* StreamingTV: Indicates whether the customer has streaming TV (yes or no)
* StreamingMovies: Indicates whether the customer has streaming movies (yes or no)
* PaperlessBilling: Indicates whether the customer has paperless billing (yes or no)
* PaymentMethod: The customer's payment method (electronic check, mailed check, bank transfer, credit card)
* Tenure: Number of months the customer has stayed with the provider
* MonthlyCharge: The average monthly charge to the customer
* Bandwidth\_GB\_Year: The average amount of data used in GB by the customer in a year

The dataset also includes variables representing responses to an eight-question survey, where customers rated the importance of various factors on a scale of 1 to 8. The survey variables are labeled as Item1 to Item8, representing factors such as timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange, and evidence of active listening.

Top of Form

Bottom of Form

**Part II**

# C1. Data Cleaning Plan-Detection

My data cleaning plan is to back-up the data locally and to a GitHub repository and then import the data using Python into a Jupyter Notebook in JupyterLab.

My plan to clean the churn\_raw\_data.csv data included:

* Reexpression of categorical data where appropriate.
* Use exploratory data analysis to create visualizations using techniques such as histograms and box plots (Larose & Larose, 2019).
* Display and treat missing values by locating columns with missing data and showing the number of values missing. Use imputation or other techniques to replace missing values for variables with missing data. (Larose & Larose, 2019).
* Computing standard deviation to assist in the identification of outliers. (Larose & Larose, 2019).
* Search for duplicate rows or columns. I will use Python to identify any duplicate rows or columns and remove them if necessary (Larose & Larose, 2019).
* Defining outliers using z-scores for quantitative variables. This is a standard method to identify outliers. (Larose & Larose, 2019).

# C3. Programming Language and Libraries

Python and R are both popular programming languages for data cleaning and analysis, and each has strengths. The main reason I chose Python over R for data cleaning is that it’s a widely used object-oriented general-purpose language. It has a large ecosystem of libraries and frameworks and libraries such as Pandas, NumPy, and scikit-learn that excel in data manipulation and cleaning. Pandas provide a rich set of functionalities for handling structured data, including missing data imputation, data transformation, merging, reshaping, and more. It allows data analysts to perform complex data-cleaning operations efficiently.

Another reason I chose to use the Python programming language for the data cleaning task is because I wanted to learn Python and JupyterLab. I set up accounts with Anaconda.org and Anaconda Navigator and hadn’t had a chance to use them. During the Google Data Analytics Certificate program that I completed just before entering the master’s program, I used R and R Studio. I found R interesting but for this course, I wanted to try out Python.

Through some brief exposure during a Coursera course, I saw that Jupyter Notebooks provide a convenient way to run code visually, and it was easy to add notes and context using Markdown. Also, Jupyter Notebooks allow me to show the Python code and graphic visualizations along with the documentation in markdown which will make the process easier to follow for other data analysts, evaluators, etc.

I installed and imported the Python packages and libraries that allowed me to perform the complex data science tasks required to complete the performance assessment. These included:

* Pandas – “to import, manipulate, and analyze data” (Trotta, 2022);
* Matplotlib - a comprehensive library for creating static, animated, and interactive visualizations in Python(Trotta, 2022). Useful to plot charts;
* Seaborn - a Python data visualization library based on matplotlib providing high-level interface for drawing attractive and informative statistical graphics for advanced plots(Trotta, 2022);
* Scikit-learn(sklearn) – for “efficient tools for predictive data analysis” using machine learning models (Trotta, 2022).
* NumPy - to work with arrays ;
* Missingno- “simple-to-use Python library that provides a series of visualizations to understand the presence and distribution of missing data within a pandas data frame. This can be in the form of either a bar plot, matrix plot, heatmap, or a dendrogram” (McDonald, 2021); and
* tabulate- used to turn tabular data into plain text tables. (Matalka,2021)

# C4. Code used for detecting duplicates, missing values, and outliers

# Detect Duplicates:

# churn\_df.duplicated()

# duplicate\_rows = churn\_df.loc[churn\_df.duplicated()]

# print(duplicate\_rows)

# Detect Missing Values:

# churn\_df.isnull().sum()

# cols=churn\_df.columns

# msno.bar(churn\_df[cols[:20]], fontsize=8)

# msno.bar(churn\_df[cols[20:]],fontsize=8)

# msno.matrix(churn\_df, fontsize = 8, labels=True)

# plt.title('Missing data matrix')

# plt.show()

# Detect Outliers

# churn\_df['z\_score\_income']=stats.zscore(churn\_df['Income'])

# churn\_df[['Income','z\_score\_income']].head()

# churn\_df['z\_score\_age']=stats.zscore(churn\_df['Age'])

# churn\_df[['Age','z\_score\_age']].head()

# churn\_df['z\_score\_children']=stats.zscore(churn\_df['Children'])

# churn\_df[['Children','z\_score\_children']].head()

# churn\_df['z\_score\_outage']=stats.zscore(churn\_df['Outage\_sec\_perweek'])

# churn\_df[['Outage\_sec\_perweek','z\_score\_outage']].head()

# churn\_df['z\_score\_email']=stats.zscore(churn\_df['Email'])

# churn\_df[['Email','z\_score\_email']].head()

# churn\_df['z\_score\_contacts']=stats.zscore(churn\_df['Contacts'])

# churn\_df[['Contacts','z\_score\_email']].head()

# churn\_df['z\_score\_equipfailure']=stats.zscore(churn\_df['Yearly\_equip\_failure'])

# churn\_df[['Yearly\_equip\_failure','z\_score\_equipfailure']].head()

# churn\_df['z\_score\_tenure']=stats.zscore(churn\_df['Tenure'])

# churn\_df[['Tenure','z\_score\_tenure']].head()

# churn\_df['z\_score\_monthlycharge']=stats.zscore(churn\_df['MonthlyCharge'])

# churn\_df[['MonthlyCharge','z\_score\_monthlycharge']].head()

# churn\_df['z\_score\_bandwidth\_usage']=stats.zscore(churn\_df['Bandwidth\_GB\_Year'])

# churn\_df[['Bandwidth\_GB\_Year','z\_score\_bandwidth\_usage']].head()

**Part III**

# D. Data Cleaning -Treatment

**A screenshot of a computer

Description automatically generated with medium confidence**

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing screenshot, text, line, font

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

Visually Identify Columns with Missing Value

A picture containing screenshot, line, text, rectangle

Description automatically generated

Create a missing value matrix

A screenshot of a computer

Description automatically generated with medium confidence

Impute missing values for quantitative variables

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A screenshot of a graph

Description automatically generated with medium confidence

Tenure and Bandwidth usage have very similar patterns. It may be useful to analyze further to see how close their correlation is.

The age distribution is far from normal and there appears to be some kind of bias in the sampling methodology favoring customers in their late forties and early fifties. This should be investigated.

A picture containing text, screenshot, number, font

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing rectangle, line, pattern, parallel

Description automatically generated

Note that bars are now showing 10,000 values for each variable.

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing rectangle, screenshot, line

Description automatically generated

The missing data bar chart and the massing data matrix now show that all missing values have been treated.

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, diagram, plot

Description automatically generated

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

A screenshot of a graph

Description automatically generated with medium confidence

Number of children above six is an outlier that should be investigated.

Checking for Outliers - Service Related Data

A picture containing text, screenshot, diagram, line

Description automatically generated

Histogram of Outage\_sec\_perweek shows negative outage durations.

Fixed the erroneous data:



Used clip(lower=constant) which was set to zero to make all negative values equal to zero.

A picture containing text, line, diagram, screenshot

Description automatically generated

While the boxplot is showing significant outliers, that should be investigated, there are now no values below zero.

A picture containing diagram, screenshot, text, line

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, diagram, plot

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, display, diagram

Description automatically generated

A screen shot of a graph

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, display, diagram

Description automatically generated

A screen shot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, diagram, number

Description automatically generated

A screen shot of a graph

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, diagram, screenshot, plot

Description automatically generated

MonthlyCharge is close to a normal distribution but does have some outliers that are barely noticeable on the histogram but easier to see with a boxplot.

A screenshot of a graph

Description automatically generated with medium confidence

A picture containing text, screenshot, font, number

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A screen shot of a graph

Description automatically generated with medium confidence

Bandwidth usage is not showing outliers. Does appear to have a very similar pattern to Tenure.

While imputing missing values using the median was a useful approach, there are several challenges that a data analyst may encounter when using the data cleaned using this method for analysis. Some of these challenges include:

* Underestimation of variance: Single imputation, such as using the median to fill in missing values, ignores uncertainty and almost always underestimates the variance. This can lead to inaccurate results, especially for statistical tests that are sensitive to variance.
* Bias: Imputing missing values with the median can introduce bias into the data, especially if the missing values are not randomly distributed. This is because the median is a measure of central tendency, and it does not consider the distribution of the data.
* Distortion of relationships: Imputing missing values can distort the relationships between variables in the data. This is because the imputed values are not the true values, and they can therefore create spurious correlations.
* Inability to detect outliers: Imputing missing values can make it difficult to detect outliers in the data. This is because the imputed values are not the true values, and they can therefore mask the true distribution of the data.

To mitigate these challenges, it is advisable to consider alternative imputation methods like Iterative Imputer, KNN, etc for treating missing data. “Imputing missing values can help you retain the information and sample size of your dataset, but it can also introduce errors or uncertainties. Replacing missing values can help you avoid errors or biases, but it can also distort the data and mask the true variability and relationships. Ignoring missing values can help you preserve the original data and avoid introducing artificial values, but it can also limit the scope and accuracy of your analysis.”(LinkedIn,2023,para 5)

# 

# E. Principal Componenbt Analysis (PCA)

# Principal Components List

The items available for PCA are:

Contacts

Tenure

MonthlyCharge

Timely response

Timely fixes

Timely replacements

Reliability

Options

Respectful response

Courteous exchange

Evidence of active listening

A picture containing text, screenshot, font, number

Description automatically generated

**A picture containing text, screenshot, line, plot

Description automatically generated**

# E2. Identifying Principal Components

The Principal Components were chosen based on their relevance to customer churn. The survey response data provided feedback from customers about their interactions with tech support/ customer service and the quality of service received. These components should serve as good indicators when analyzing telecommunication churn rates.

# E3. Benefits of PCA results

PCA is a powerful tool that can be used to reduce dimensionality, improve interpretability, and reduce the noise in a dataset. Its main purpose is to reduce the complexity of large datasets. However, it is important to be aware of the limitations of PCA before using it.

**Part IV**

# Panopto Recording

Panopto recording can be found here: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=59aa4a04-e2b0-4506-84b6-b02f014dffce

# Web Source References

Matalka, L. (2021, February 26). *How to Easily Create Tables in Python*. Medium. https://towardsdatascience.com/how-to-easily-create-tables-in-python-2eaea447d8fd

McDonald, A. (2021, September 8). *Using the missingno Python library to Identify and Visualise Missing Data Prior to Machine Learning*. Medium. https://towardsdatascience.com/using-the-missingno-python-library-to-identify-and-visualise-missing-data-prior-to-machine-learning-34c8c5b5f009

Trotta, F. (2022, December 3). *5 Python Libraries to Learn to Start Your Data Science Career*. Medium; Towards Data Science. https://towardsdatascience.com/5-python-libraries-to-learn-to-start-your-data-science-career-2cd24a223431

*What are the pros and cons of deleting rows or columns with missing values?* (n.d.). Www.linkedin.com. Retrieved June 28, 2023, from https://www.linkedin.com/advice/0/what-pros-cons-deleting-rows-columns-missing-values#:~:text=Imputing%20missing%20values%20can%20help

# Sources and References

Larose, C. D., & Larose, D. T. (2019). *Data Science using python and R*. Wiley.