**Data Cleaning (NUM3)– D206**

**Performance Assessment**

**Western Governors University**

**By: Christian LeBlanc**

**Part I: Research Question**

**A**

My research question is “What factors influence if a customer gets Tech Support?” The reason I believe the results from this question would be valuable to the organization is the customers that fall into this category would be good targets to upsell the technical support package. I have worked in sales situations that sold technical support and found that those not confident in their technical skill and do not use tech often are usually more open to technical support, but finding out if this is true could lead to rise in revenue per customer. If not, finding out the factors that give positive indicators on getting Tech Support could lead to more customers getting it if we know which ones are more likely to sign up for it.

**B**

The churn dataset contains 50 variables according to the Data Dictionary that accompanies it. I noticed that it actually has 51 variables because the Data Dictionary treats the Lat and Lng as one variable when they are listed in their own separate column in the dataset. Listed below are a brief description of the variable, the data type, and an example pulled from the dataset.

* CaseOrder is used as an index to keep the original raw data’s order. It is quantitative. The example pulled from the first row is 1.
* Customer\_id is the ID for each distinctive customer. It is qualitative. The example pulled from the first row is K409198.
* Interaction is an identification related to transactions for customers, technical support, and sign-ups. It is qualitative. The example pulled from the first row is aa90260b-4141-4a24-8e36-b04ce1f4f77b.
* City is the city that is on the customer’s billing statement. It is qualitative. The example pulled from the first row is Point Baker.
* State is the state that is on the customer’s billing statement. It is qualitative. The example pulled from the first row is AK.
* County is the county that is on the customer’s billing statement. It is qualitative. The example pulled from the first row is Prince of Wales-Hyder.
* Zip is the zip code that is on the customer’s billing statement. It is qualitative. The example pulled from the first row is 99927.
* Lat is the latitudinal coordinates that is on the customer’s billing statement. It is quantitative. The example pulled from the first row is 56.251.
* Lng is the longitudinal coordinates that is on the customer’s billing statement. It is quantitative. The example pulled from the first row is -133.37571.
* Population is the amount of people within a mile from the customer. It is quantitative. The example pulled from the first row is 38.
* Area is the type of area the customer lives in, being rural, urban, or suburban. It is qualitative. The example pulled from the first row is Urban.
* Timezone is the time zone from the location the customer used in their sign-up information. It is qualitative. The example pulled from the first row is America/Sitka.
* Job is the reported job for the customer in their sign-up information. It is qualitative. The example pulled from the first row is Environmental health practitioner.
* Children is the number of children in the residency for the customer in their sign-up information. It is quantitative. The example pulled from the first row is NaN.
* Age is the customer’s age in their sign-up information. It is quantitative. The example pulled from the first row is 68.0.
* Education is the customer’s highest degree completed in their sign-up information. It is qualitative. The example pulled from the first row is Master's Degree.
* Employment is the status of employment of the customer in their sign-up information. It is qualitative. The example pulled from the first row is Part Time.
* Income is the customer’s reported yearly income in their sign-up information. It is quantitative. The example pulled from the first row is 28561.99.
* Marital is the customer’s marital status in their sign-up information. It is qualitative. The example pulled from the first row is Widowed.
* Gender is the customer’s gender as stated as male, female, or nonbinary. It is qualitative. The example pulled from the first row is Male.
* Churn is within the last month has the customer discontinued services as yes or no. It is qualitative. The example pulled from the first row is No.
* Outage\_sec\_perweek is average of system outages in seconds per week in the customer’s neighborhood. It is quantitative. The example pulled from the first row is 6.972566.
* Email is the amount of sent emails to the customer by marketing or correspondence in the last year. It is quantitative. The example pulled from the first row is 10.
* Contacts is how many times technical support was contacted by the customer. It is quantitative. The example pulled from the first row is 0.
* Yearly\_equip\_failure is how many times in the past year that equipment failed for the customer and had to be replaced or reset. It is quantitative. The example pulled from the first row is 1.
* Techie is the yes or no response the customer gave on their customer questionnaire on if they are technically inclined in their opinion when they signed up. It is qualitative. The example pulled from the first row is No.
* Contract is the contract length for the customer being month-to-month, one year, or two years. It is qualitative. The example pulled from the first row is One year.
* Port\_modem is if the customer’s modem is a portable modem, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* Tablet is does the customer own a tablet, with responses that are yes or no. It is qualitative. The example pulled form the first row is Yes.
* InternetService is the internet service for the customer, either DSL, Fiber Optic, or None. It is qualitative. The example pulled from the first row is Fiber Optic.
* Phone is does the customer have the phone service, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* Multiple is the customer using multiple lines, with responses that are yes or no. It is qualitative. The example pulled from the first row is No.
* OnlineSecurity is does the customer have the add-on for online security, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* OnlineBackup is does the customer have the add-on for online backup, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* DeviceProtection is does the customer have the add-on for device protection, with responses that are yes or no. It is qualitative. The example pulled from the first row is No.
* TechSupport is does the customer have the add-on for technical support, with responses that are yes or no. It is qualitative. The example pulled from the first row is No.
* StreamingTV is does the customer have streaming TV, with responses that are yes or no. It is qualitative. The example pulled from the first row is No.
* StreamingMovies is does the customer have streaming movies, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* PaperlessBilling is does the customer have paperless billing, with responses that are yes or no. It is qualitative. The example pulled from the first row is Yes.
* PaymentMethod is how the method in which the customer makes payments, with responses that are electronic check, mailed check, bank (automatic bank transfer), credit card (automatic). It is qualitative. The example pulled from the first row is Credit Card (automatic).
* Tenure is how long in months that the customer has been with the provider. It is quantitative. The example pulled from the first row is 6.795513.
* MonthlyCharge is the monthly charge for the customer, with the value reflecting an average per customer. It is quantitative. The example pulled from the first row is 171.449762.
* Bandwidth\_GB\_Year is the customer’s average number of GB of data used in a year. It is quantitative. The example pulled from the first row is 904.53611.
* Item1 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of timely response. It is qualitative. The example pulled from the first row is 5.
* Item2 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of timely fixes. It is qualitative. The example pulled from the first row is 5.
* Item3 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of timely replacements. It is qualitative. The example pulled from the first row is 5.
* Item4 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of reliability. It is qualitative. The example pulled from the first row is 3.
* Item5 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of options. It is qualitative. The example pulled from the first row is 4.
* Item6 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of respectful response. It is qualitative. The example pulled from the first row is 4.
* Item7 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of courteous exchange. It is qualitative. The example pulled from the first row is 3.
* Item8 is on a 1 to 8 scale, with 1 being most important and 8 being least, the importance of evidence of active listening. It is qualitative. The example pulled from the first row is 4.

**Part II: Data-Cleaning Plan**

**C-1**

After uploading the csv file, the first step I will take in cleaning the data is using the “.info” function from the Pandas package. Using “.info” allows me to quickly see what columns have null values and also the datatypes for the columns to verify if they are appropriate when compared to the data dictionary. I will run my dataset to see if I can spot any areas of concern visually. After this, I will start looking into each column individually. I will be using a combination of “.value\_counts()” for qualitative and “.describe()” for quantitative with the exception of using “.nunique” for columns that should have a unique response for each customer. For the quantitative columns I will be using the rule of thumb of z-value is either greater than 3, or less than -3 to identify outliers(Larose & Larose, 2019).

**C-2**

The approach I took in looking at the quality of the dataset was looking into each column individually as well as any issue while looking at the dataset as a whole. The “.info” allowed me to see each columns datatype as well as if any the columns that have null values. Then running the dataset with all the columns showing let me see if anything looked different than it should, such as place values. After that I used the combination of “.describe”, “.value\_counts”, and “.nunique” depending on should be in the columns based off of the data dictionary. Using the “.describe” command I could see how the datatype looks and find the z-score to determine outliers. The “.value\_count” command allowed me to see what kind of response were given and if they fit into what is expected from the data dictionary and if the column could be reformatted to save space. The columns that need to have a unique response from every customer is easily found if they had the 10,000 responses with “.nunique”. I also included notes for myself in my code as to which datatypes did not look like best fit to make it easier to find them once I start to clean the data.

**C-3**

I decided to go with Python in Jupyter notebook to go through the data-cleaning process. I have previous experience using Python from other classes and work experiences and found it logical to follow and write. I will be using the Pandas and Numpy libraries to clean my data as Pandas is needed to read the csv file and store it in a way that allows me to manipulate it as needed while Numpy allows for the mathematical or logical calculations that may be needed to transform the data into clean data.

**C-4**

Please see attached code in my .ipynb file for part C.

**Part III: Data Cleaning**

**D-1**

The first thing that I noticed when reviewing the dataset was the columns with null values. The columns of Children, Age, Income, Techie, Phone, TechSupport, Tenure, and Bandwidth\_GB\_Year all had null values which is shown by having less than 10,000 entries. My first thought when seeing this with Children is that it was entries using a null value as 0 children, but when I took more time looking I saw that 0 was in the responses and this assumption could not be reasonably relied on. This leads me to worry if the missing value was due to entry error or the customer did not want to give this information because Children, Age, and Income can be considered sensitive information and I can see people not wanting to disclose them for the services this company offers. Things like TechSupport should be available somewhere in the company’s files because they have to know who has these services or this could cause issues.

Next issue came when visually inspecting the dataset. I saw in Zip that an entry only had four digits instead of five like zip codes actually have. This is a set standard in the United States and needed to be addressed.

The column State required me to take a closer look into the data results. I checked the number of unique results expecting a number of 50 or less. I received a result of 52. This led me to look at the value count to see what results show up. This is where I found they include Washington D.C. and Puerto Rico in states results. This is something that is commonplace for companies that do business in those areas, so no need for cleaning.

As I progressed with my inspection of the data, I came across the first outliers in the Lat columns. I previously mentioned the method that I use when it comes to what I will consider an outlier being greater than 3 or less than negative 3 z-score. With this being the baseline of what I considered as outliers in this dataset, I found that columns Lat, Lng, Population, Children, Income, Outage\_sec-perweek, Email, Contacts, Yearly\_equip\_failure, and MonthlyCharge all have outliers present in them. Column Lat has outliers at both high and low ends of its data points. Both are valid data as they are lines of latitude, and the max value is 70.640660 which is in Alaska and the minimum value is 17.966120 which is in Puerto Rico. The Lng is a similar situation, but with only the low-end giving outlier and this is also because of Alaska’s longitude line of -171.688150. Population was the next outlier that I came across. This is the high end of the datapoints, but when looking at it from a real-life perspective this one is completely plausible to have a population of 111,850 living within a mile in densely populated areas like New York City. Children is the next in line. The outlier that is the max value is 10 children. This is a large number for children, but again is very plausible as I personally know people with 10 children. Income has an outlier with the max value it holds. An annual income of $258,900 is the maximum outlier and it may be considered upper class, but it is an income that is realistic. The remaining columns are more specific to the company, but I have previously worked in multiple companies that offer similar type of services so I will be using my previous experiences to address the outliers. Outage\_sec\_perweek has the max outlier of 47 seconds per week on average. This would put the total outage time for the year a little under 41 minutes, which one bad natural disaster, like a hurricane, would exceed this and I have personally never had an internet provider does not have outages to do maintenance at least yearly which have all taken more than this. This has me questioning the accuracy of this whole column, especially when I consider that they have negative values in this column that cannot realistically be negative. The Email column has outliers in both the minimum and maximum values. The minimum of 1 can be a customer that has informed the company they do not want email’s cluttering their inbox and request to only be contacted via email when necessary and on the max end of 23 could be a mixture of having a monthly marketing email, I say this because of the mean being 12, and 11 correspondence emails could be just a couple of issues within the year that takes some back and forth discussion. Contacts is the next column and the max of 7 times of contacting customer support could be done for multiple reasons. As someone who works from home currently, I know I am on the phone with customer service the moment an issue arises with my internet service, and on the other end of the spectrum the 89-year-old customer may contact technical support for issues that may not require a call for most. Yearly\_equip\_failure has a maximum value that is an outlier with 6. Once again this is not a large number even though it is technically an outlier. If a customer has a separate modem and router they could have sent for multiple rests before requesting a replacement and from the data dictionary this column counts both resets and replacements. The last outlier that I found was in the MonthlyCharge column. This maximum value of $315.88 is a reflection of an average for the customer. Once again this can be easily explained multiple different ways, with the one type that I have seen the most of is a customer that has all the services offered and occurs late payment fees or overage fees that after a few months of this could cause the average bill to increase. One last thing I feel that I have to discuss on outliers is why I only discussed the maximum or minimum values and did not dig deeper to see if other outliers exist is if the furthest outlier is plausible or confirmed as legitimate then all the other ones going towards the average would have to be plausible or confirmed. If I had a column with a max or min outlier that was not reasonable then I would have checked more data points going towards the average.

The next data quality issue I found first appeared in the Area column and is columns storing data as string objects when they could be stored more efficiently and closer aligned with the data dictionary. The columns Area, Timezone, Education, Employment, Marital, Gender, Churn, Techie, Contract, Port\_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, and PaymentMethod all are columns with string objects that need changing. The changes from string objects to the datatype of category will be needed in Area, Timezone, Education, Employment, Marital, Gender, Contract, and InternetService. Columns Timezone and Gender both require more looking into which will be discussed in D-2. Churn, Techie, Port\_modem, Tablet, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling all have responses of “Yes” or “No” which can be converted to Boolean datatype.

The datatype for floating point number is used in some columns that should be switched to integer datatype to be stored more efficiently. Columns for Children, Age, and Tenure all do not need any decimal places in their results. Also, columns Outage\_sec\_perweek, MonthlyCharge, and Bandwidth\_GB\_Year all use the floating-point number data type. They do not need to change their datatype, but I do think 6 decimal places is more in-depth than needed.

Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8 are all survey questions that are listed as integers. These columns all need to be redone to have the ordered categorical datatype with “1” being most important through “8” being least important.

The last thing that I have found that is need of cleaning is the labels of the columns. They are inconsistent with their usage of capital letters, and if they will use “\_” to space the words out. Another issue is nondescriptive columns.

**D-2**

Handlin of the null values was the roughest choice to make in this assignment. I feel like if this was a real-world scenario that I could partner with other departments to get most if not all the missing data because things like TechSupport, Tenure, and Bandwidth\_GB\_Year are all things that would have to be recorded somewhere in the company’s records. I know that no matter which method I go with adds with it the possibility of distorting the data or losing insight. Also, I know that I will have to run a PCA on this dataset and it requires no null values. After thinking about these points for a few days I decided that imputing the median for quantitative variable and mode for qualitative as dropping the null values would have too much lost information.

Dealing with the issue of digits with the Zip column I switched the data type. The Zip column was stored as integers and are now stored as strings. Switching Zip to string alone does not fix the issue, so I used “.zfill” to have zeros appear in front of the Zip entries that have less than five.

I did not do any changes to the outliers that exist in this dataset. I discussed why each of them are either valid or at least plausible.

Changing the columns Area, Timezone, Education, Employment, Marital, Gender, Contract, and InternetService from string to categorical is warranted to be more efficiently stored and add a constraint to keep responses in line with the data dictionary. While doing this it is important to note two additional concerns in these columns. Timezone is cumbersome and should be more straight forward with the categories used, because looking at it and seeing “America/ Menominee” does not tell me anything about the actual time zone but labeling it as “Central Time” would make it easier to know the time. Also, Gender’s results do not match the expected results labeled in the data dictionary, but I do not want to change them because the reported “Prefer not to answer” cannot all be assumed to be “nonbinary” so this column will not be reflecting the responses wanted from the data dictionary. For columns Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8 all being survey questions with eight possible responses allows them to be more efficiently saved as categorical.

Changing the columns Churn, Techie, Port\_modem, Tablet, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, and PaperlessBilling from string to Boolean is done to store “Yes” or “No” responses more efficiently.

Floating data points columns Children, Age, and Tenure are being changed to Integers. This is being done for efficiency in storage and the lack of need for any more precision past whole numbers. Children and Age seems straight forward as you can’t have a fraction of a kid, and when asked for their age the average person would just give the whole number for it. The reason I am saying that Tenure does not need more precision is because it is already returning the number of months that a customer has stayed with a provider. I would lean towards this is overly precise for a company, but since they have in Contract that they offer month-to-month contracts it can explain why they would want to see tenure by month, but there is no need to drill down further than that.

Floating data points columns MonthlyCharge, and Bandwidth\_GB\_Year are staying floating data points but will be restrained to 2 decimal places. This is being done for efficiency in storage and the lack of need for any more precision than two decimal places. MonthlyCharge is a monetary charge and putting the limit after two decimal places is commonplace when dealing with money. For Bandwidth\_GB\_Year I went back and forth about if I should put it as an integer or place a cap on the decimal values. I decided to go with the 2 decimal places instead of integer is because GB of data is still a large enough measurement that I think having the extra information could help as I know some places have data caps and this could be used to give a warning to customers if close.

Outage\_sec\_perweek is a column that I spent more time on deciding what to do with it than most. First off, having negative values recorded in the column makes no real-world sense because measuring outages in the negative would be saying that the network would not be down so you would be measuring the network running. Next issue that I mentioned earlier is that the numbers are so low it is not realistic that their network had that little outages that they can measure it by seconds. In a professional setting where I could ask question, I would bring this issue to coworkers to see if answers to my concerns could be addressed, but without having that as an option in this assignment I feel like dropping this column is needed for me to feel confident about the data I am presenting.

Column headings are not used to label the columns well. They are also stored with inconsistent treatment which could unneeded issues when trying to manipulate them. Columns such as “item3” should be switched to align more with what the column is representing as it does not tell you anything about what you are looking at in the column. Also, I will be changing all the columns to be uniformed in all lower case and if it has multiple words, it will be separated by “\_” to make the columns headers uniformed and meaningful.

**D-3**

The resulting data from that data-cleaning is easier to follow with the column name changes with more uniformed and useful headers. That data no longer has any null values, and it is stored in a more effective way. Below is a screen shot showing the 10,000 non-null, new column names, and new datatypes.

A screenshot of a computer

Description automatically generated

**D-4**

Please see attached code in my .ipynb file for part D.

**D-5**

Please see attached CSV file of the cleaned data.

**D-6**

Every dataset will have its limitations. A glaring one in this is the inability to be able to get with coworker to get more of the missing data, as I touched on previously some of the null values are in things that the company would have to have stored somewhere. Being able to incorporate that into the dataset would provide more accurate information. Also, some columns had almost a quarter of the information missing this could lead to distortion of the story being told with the data.

**D-7**

TechSupport having missing values is one of those columns I discussed that a company would have to have the correct information somewhere because it is a service, and they would have to know if a person is paying for it, they get to use it. Being that I want to know what influences TechSupport, having to impute values puts a distortion on the outcome already. I also worry about having columns that have almost 25% of it’s response being imputed will have a negative result in the reliability of the outcome. Columns like Age or Income could, in my mind, be big influencers in the likelihood of a customer getting TechSupport but could have a lower or higher correlation because so much of the data is not the real customer response.

E-1

The variables in this dataset that I will be using in my Principal Component Analysis are latitude, longitude, population, children, age, income, email\_sent, tech\_support\_contact, yearly\_equip\_failure, tenure, monthly\_charge, and bandwidth\_gb\_year. This is all of the quantitative variables in the clean dataset I am using. Below is a screenshot of the output of the principal components loading matrix. To see fully please see attached code for section E.

A table with numbers and letters

Description automatically generated

**E-2**

Using the Kaiser Rule of the principle components of a dataset having an Eigenvalue greater than 1 I have found that the first six components, PC1-PC6, are the principle components. These are the components that will be retained while the rest of them are no longer needed. Below is a screen shot of the scree plot and then the Eigenvalues printed out because the values close to 1 are hard to read on the plot.

A graph with a line

Description automatically generated

A screenshot of a computer

Description automatically generated

**E-3**

Principal components analysis makes analysis more efficient for the organization. It takes complex data and reduces its dimensionality. In doing this it helps save the storage space needed to do analysis on it. It also allows the ability to produce independent, uncorrelated features of the data and to visualize data allowing for the inspection of clustering/classification algorithms. Thus the PCA preformed above helped prevent overfitting by showing that PC7-PC12 could be dropped.

**Part IV. Supporting Documents**

**F**

Uploaded to the Panopto drop box titled “Data Cleaning NUM3 | D206 (Student Creators) [assignments].”

Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fa5a51bc-a5f8-4302-8053-b0c101590821>

**G**

Western Governors University. (2023) Welcome to Getting Started With Principal Component Analysis (PCA)

[7. D206-GettingStartedPCA.pdf](https://westerngovernorsuniversity.sharepoint.com/:b:/r/sites/DataScienceTeam/Shared%20Documents/Graduate%20Team/D206/Student%20Facing%20Resources/D206%20-%20Getting%20Started%20with%20D206%20Video%20Series%20(Slides%20and%20Videos)/7.%20D206-GettingStartedPCA.pdf?csf=1&web=1&e=taE7ej)

**H**

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

Bigabid. (n.d.) PCA: What, How, and Why. Bigabid

https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/

**I**

The content in this Performance Assessment is set up and presented with the highest professional standards.