**Part I: Research Question**

1. Describe **one** question or decision that you will address using the data set you chose. The summarized question or decision must be relevant to a realistic organizational need or situation.

What is causing our customers to churn? Also, can we figure who are at a higher risk for churning?

B. Describe the variables in the data set and indicate the specific type of data being described. Use examples from the data set that support your claims.

The data for the telecommunication company has 10,000 records for customers. We look at the data set to get a variable that is causing customers to continue and discontinue their service. We can also use independent variables which can help identify a relation with the dependent variable Churn.

Columns that can relate to the cause of Churn: Screen shots of all data types shown below

1. Customer Information:

* City
* State
* County
* Zip
* Marital
* Gender
* Age
* Etc.

1. Customer Services:

* Phone
* Internet
* Streaming Tv
* Tablet
* Etc.

1. Customer Account Statement:

* Paperless Billing
* Payment Method
* Tenure
* Monthly Charges
* Bandwidth
* Etc.

1. Customer Surveys

* Timely Response
* Timely Fixes
* Timely Replacements
* Etc.

This data consists of numerical “float and int” values as well as non-numerical values. The data also includes categorical columns for example: Churn “Yes, or No”, Gender, Education. Furthermore, the numerical columns consist of Income, Outage\_sec\_perweek, Bandwidth\_GD\_Year.

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

**Part II: Data-Cleaning Plan**

*Note: You may use Python, R, or any other programming language for implementing your coding solutions, manipulating the data, and creating visual representations.*

C. Explain the plan for cleaning the data by doing the following:

1. Propose a plan that includes the relevant techniques and specific steps needed to identify anomalies in the data set.

1. Imports that will need to be used
2. Importing the csv
3. Checking number of columns
4. Check csv columns for information that may be useful
5. Checking for any null or empty values, and fixing those null or empty values
6. Fix all values for each column even if the columns are not necessary
7. For data that is missing and can be meaningful, will add the mean to those values

2. Justify your approach for assessing the quality of the data, include:

The data showed many missing values across multiple columns, specifically in some columns that I found would be useful such as the Tenure, Age, and Income. For assessing the quality of the data, I will first get the sum of all the missing values in the data set and filling in those missing values. This is a good first approach because this will allow me to see which columns will need to be fixed and completed for me to see a full picture of the dataset. Also, will help the data seem more complete, and cleaner to analyze, since this will be my first attempt at cleaning data.

3. Justify your selected programming language and any libraries and packages that will support the data-cleaning process.

The programming language that I used was Python to handle this data set. I used this language because of my experience with Python language and have some knowledge of the libraries that are associated with it. Also, there is so much relevant, and useful information for python language that if I become stuck, I will be able to find solutions easily, from all the helpful articles and documentations.

Programming language:

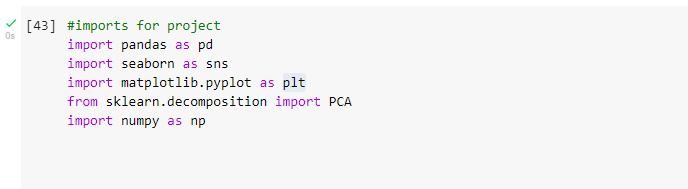
* Python, has a lot of libraries, documentation, sources for reference, beginner friendly.

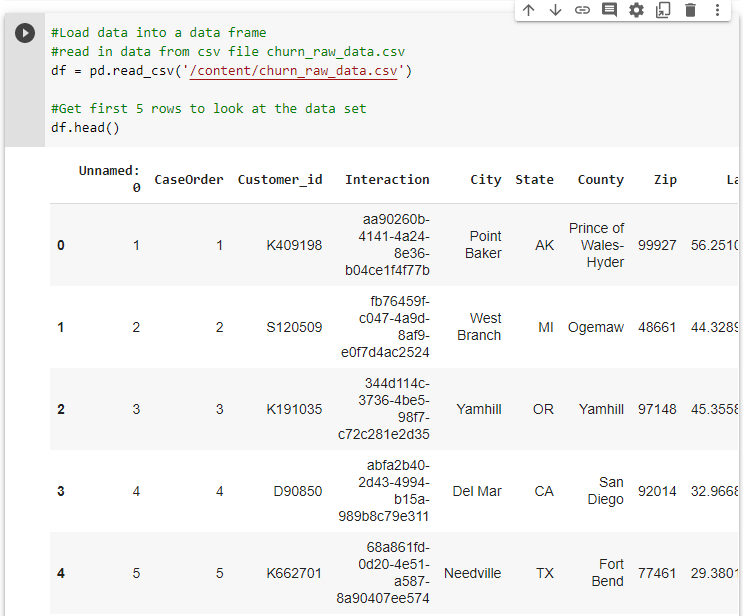
Librairies:

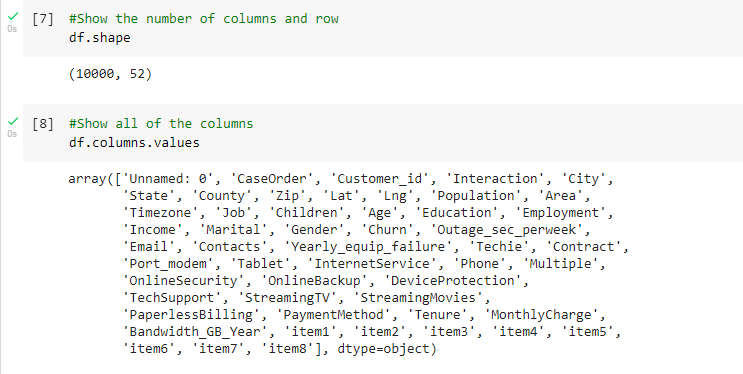
* Pandas for data loading
* Seaborn for visualizations
* MatplotLib/pyplot for chart graphing\
* Sklearn – decomposition
* Numpy for array

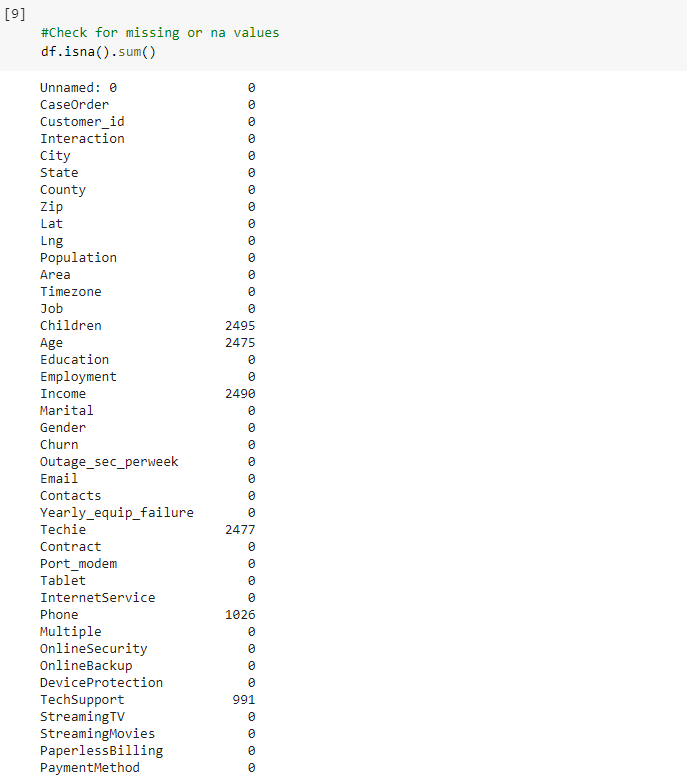
4. Provide the code you will use to identify the anomalies in the data.

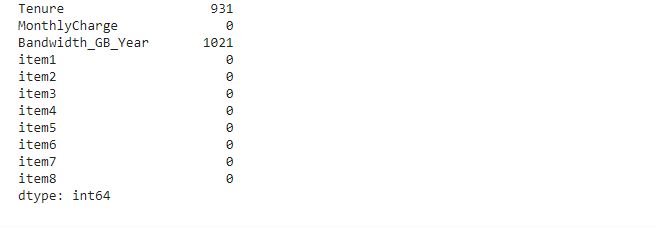
**Checking Data and Cleaning Data**

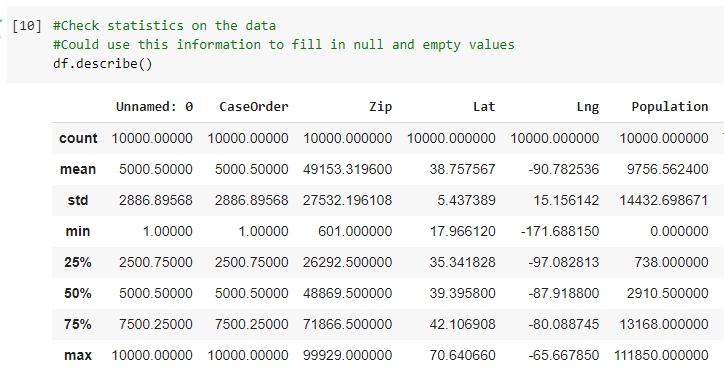


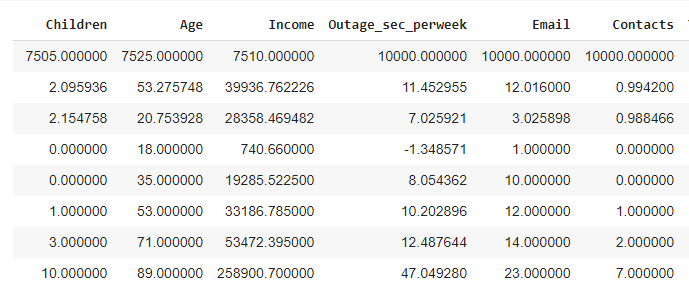


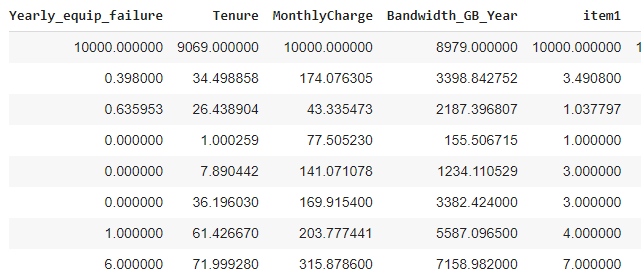


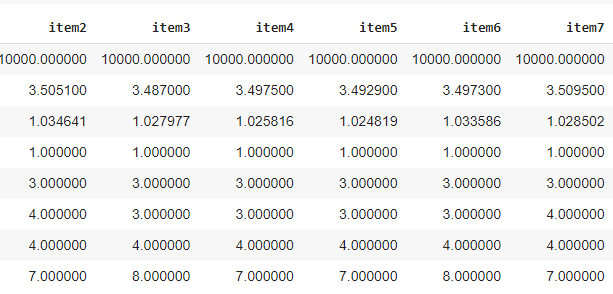


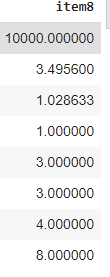


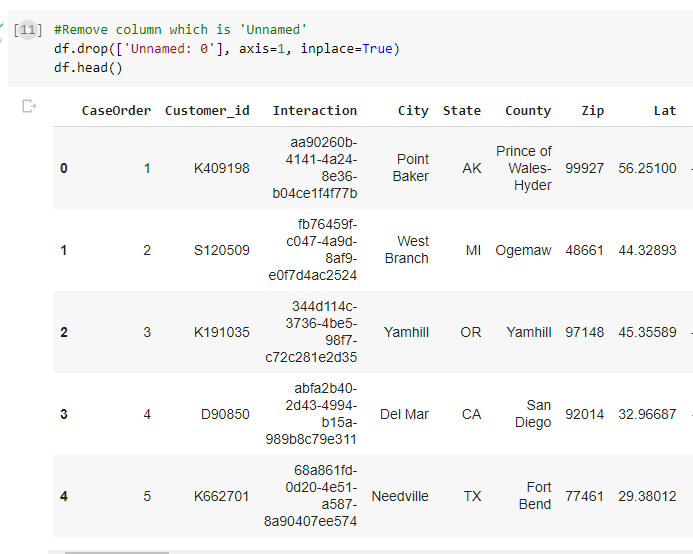




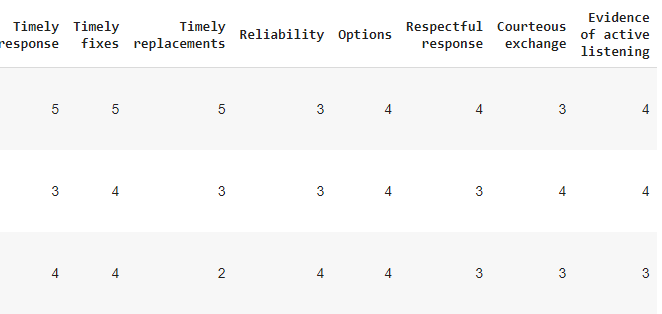


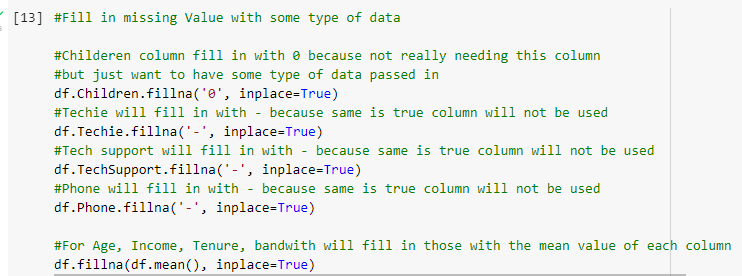


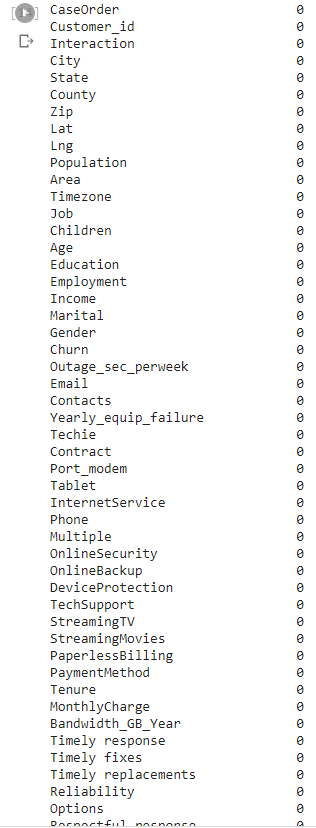


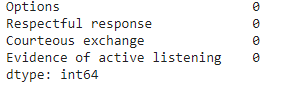


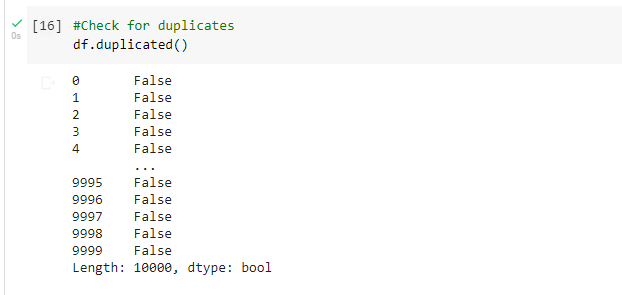




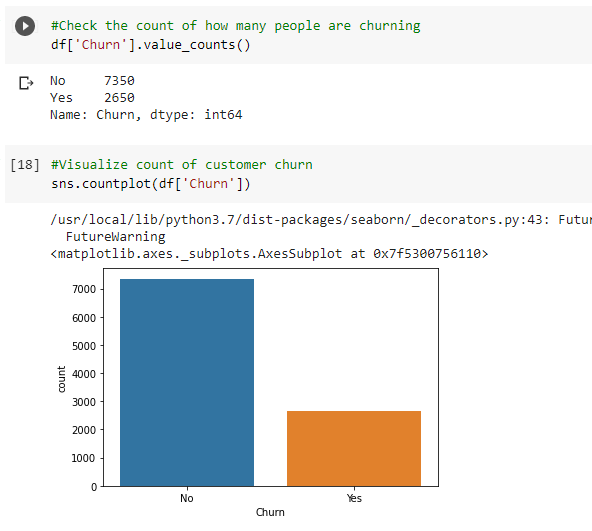


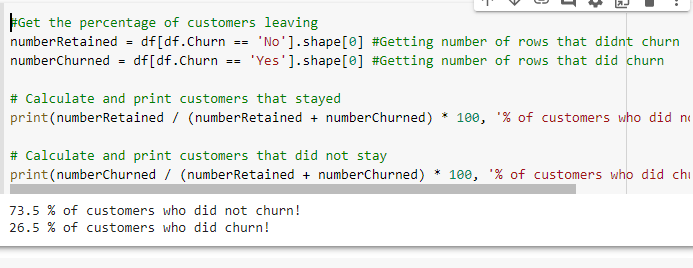


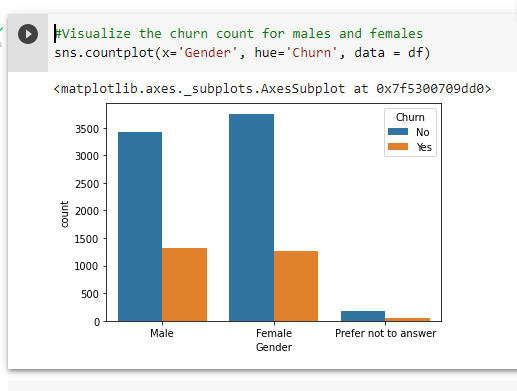


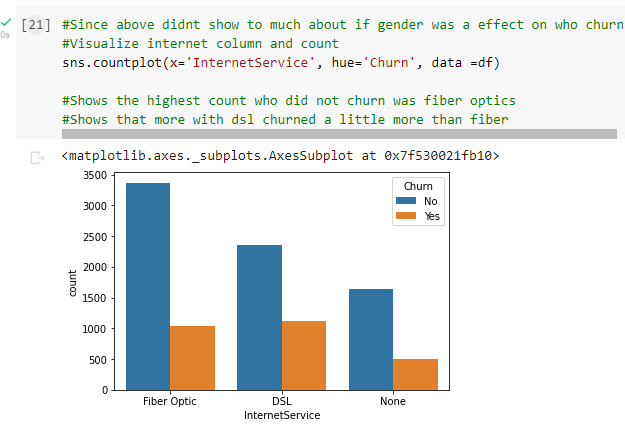


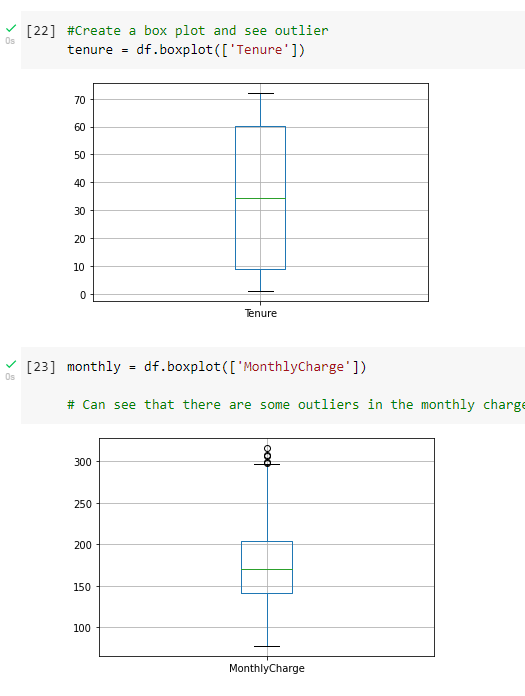
**Visualize Churn**

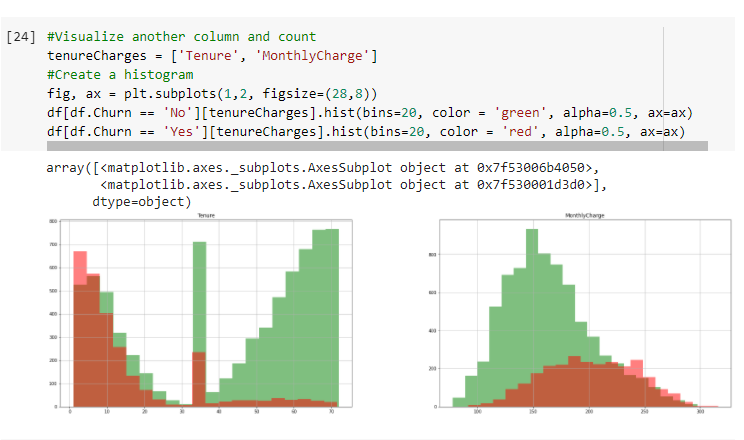






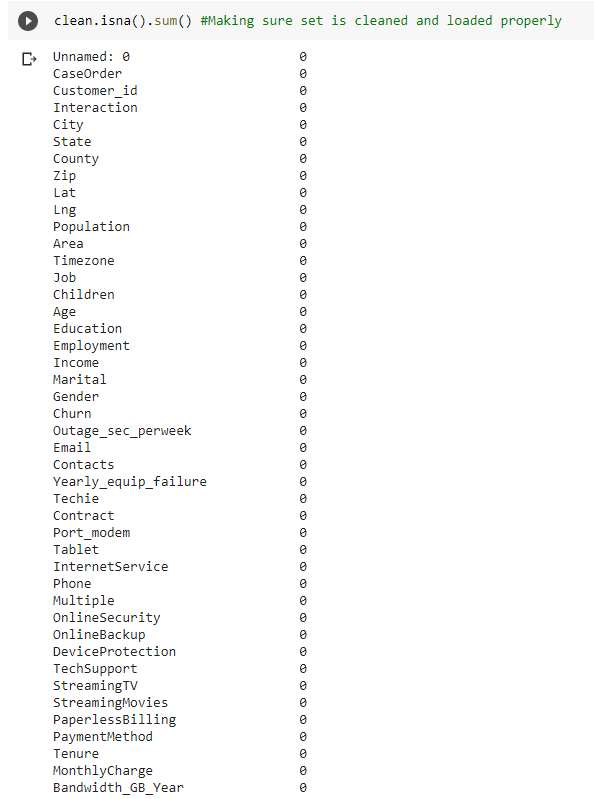


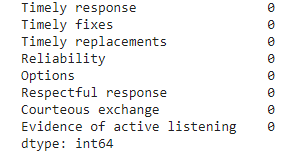


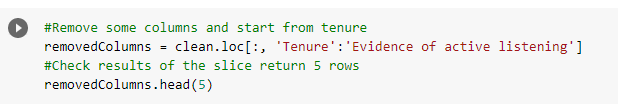


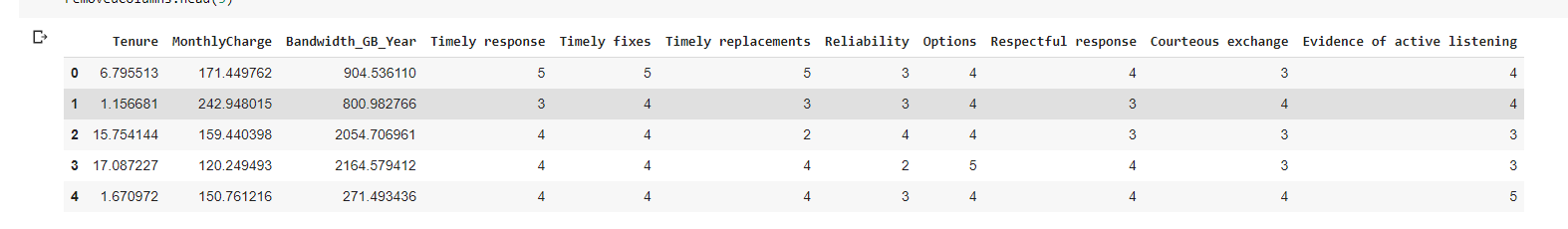
**Saving a Clean Data Set**



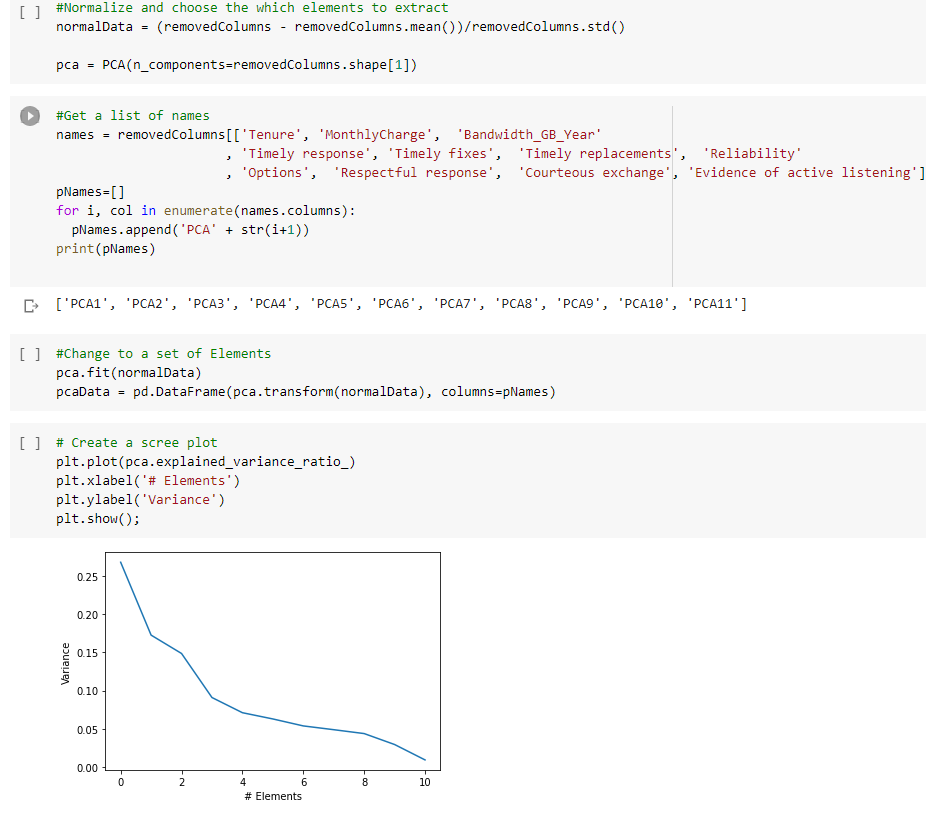


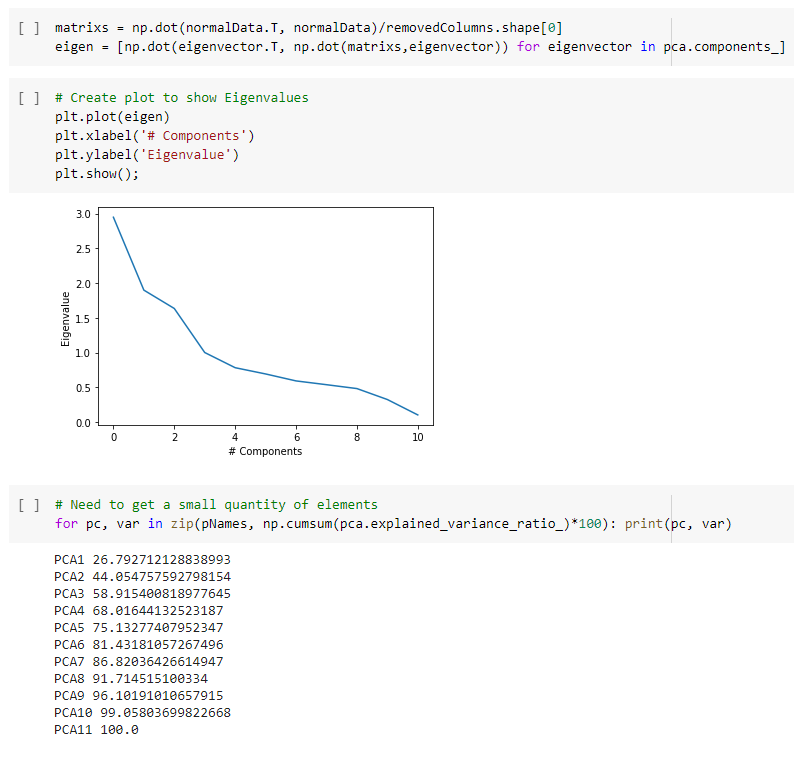


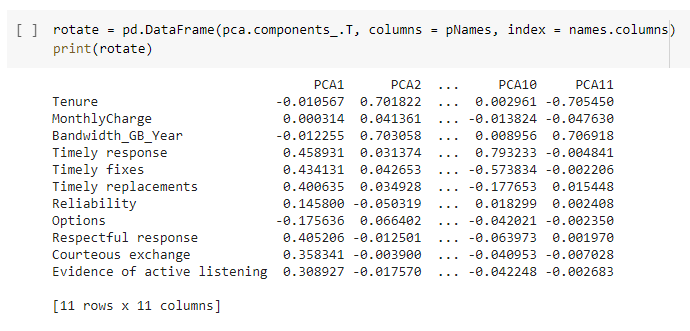


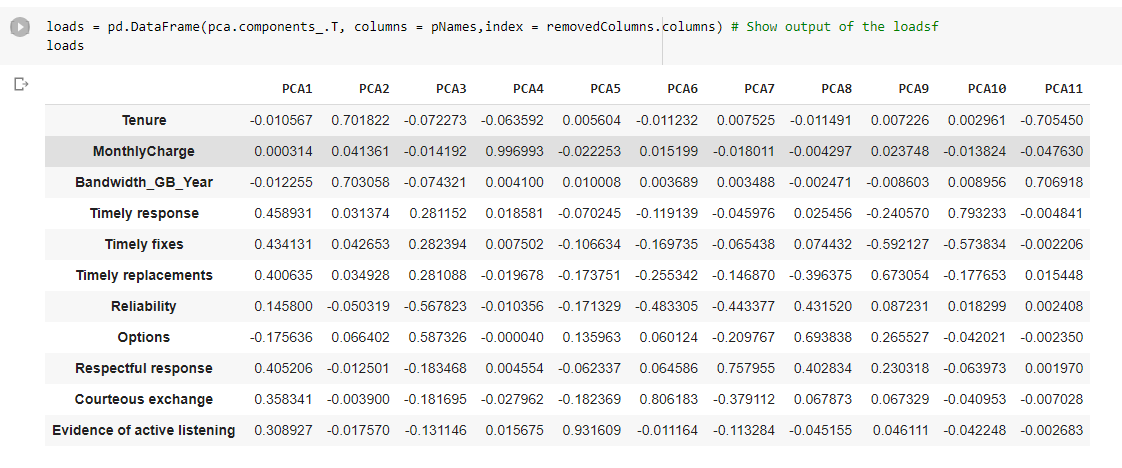


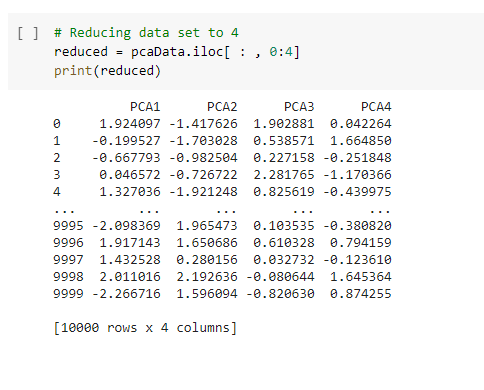
**Principal Component Analysis**











**Part III: Data Cleaning**

D. Summarize the data-cleaning process by doing the following:

1. Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.

When I checked the sum of missing values inside the dataset there were multiple missing values, and some of the columns I believed were not meaningful, for example, Children and the empty column name. With the data that I felt was important I used the mean value to replace those columns, and for other columns that I did not feel were too important to figure out why customer churn, I replaced numerical values with 0 and ‘-’ for non-numerical values. Anomalies discovered did not have too much impact and were not consequential.

2. Justify your methods for mitigating each type of discovered anomaly in the data set.

Each type of anomaly was handled with mean of each column, since this data set was smaller. With other columns I just filled in with dash and 0 since it did not have too much effect on gathering information that will answer the churn question.

3. Summarize the outcome from the implementation of *each* data-cleaning step.

The outcome of each data cleaning step after each phase the data became more readable, it was also a little clearer to see interesting columns that might help, for example, Tenure, and the responses to the survey columns.

4. Provide the code used to mitigate anomalies.

Screen shots are available above

5. Provide a copy of the cleaned data set.

Attached cleaned data set called churn\_cleaned\_csv.

6. Summarize the limitations of the data-cleaning process.

The limitations in this process of data-cleaning were that there was a lot of missing data, which raised the question why so many values were missing. But with this I was not able to speak with the company department to answer these questions. Especially for a column like Tenure that can be useful column to understanding who are churning or not.

7. Discuss how the limitations in part D6 affect the analysis of the question or decision from part A.

The limitation of part D6 affected the analysis of the question because by using mean to fill in missing values is not robust. Also, this does not scale categorical features, and can also affect the variance of the resulting dataset. So, being able to go back get the missing data for certain columns like the Tenure, could have a different result to the overall picture or decision.

E. Apply principal component analysis (PCA) to identify the significant features of the data set by doing the following:

1. List the principal components in the data set.

The list of PCA components that were used included 8 survey questions and include 3 other columns from the dataset, for a total of 11 components. Screen shots of these columns can be seen on pages 16 -18. Also provided a list of all 11 components that I used in page 20.

The PCA components in the data set included:

* Timely Responses
* Timely Fixes
* Timely Replacements
* Reliability
* Options
* Respectful Response
* Courteous Exchange
* Evidence of Active Listening
* Tenure
* Monthly Charge
* Bandwidth

Most important components in the dataset are:

1. Timely response
2. Timely fixes
3. Timely replacements
4. Respectful response

2. Describe how you identified the principal components of the data set.

The way I identified the PCA for the dataset, I looked at the original set in the null values and all the fields for each customer were field in and contain no empty values. Also, as stated above the mean can mess with the variance of the dataset so decided that could influence the results. Next, we looked at each survey column and concluded that those will give the best insight on why customers churned or not from the rating system used in the survey responses. Then use a scree plot for a linear visualization of the eigenvalues and gather the cumulative sum of the elements. This allowed me to see that with 6 components allowed for 81% and 7 allowed for 86%.

3. Describe how the organization can benefit from the results of the PCA

Organization can benefit from the results by understanding that focusing on the surveys, in the fields of response, fixes, replacements, respectful response could reduce the churn rate, with improvement in those areas.

**Part IV. Supporting Documents**

F. Provide a Panopto recording that demonstrates the warning- and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process and summarizes the programming environment.

*Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access", and then choose to log in using the “WGU” option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto’s website.*

*To submit your recording, upload it to the Panopto drop box titled “Data Cleaning – NUM2 \ D206” Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.*

G. Reference the web sources used to acquire segments of third-party code to support the application. Be sure the web sources are reliable.

*YouTube*, YouTube, 9 Apr. 2020, www.youtube.com/watch?v=ocMd2loRfWE.

H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

I. Demonstrate professional communication in the content and presentation of your submission.