**HOW DOES ONLINE SECURITY AFFECTS MONTHLY CHARGES ON INTERNET SERVICES?**

NUM2 – NUM2 TASK1: DATA CLEANING

DATA CLEANING – D206

PRFA – NUM2

STUDENT NAME: IBRAHIM SULEIMAN

DATA ANALYTICS

STUDENT NUMBER: 001429984

COLLEGE OF INFORMATION TECHNOLOGY, WESTERN GOVERNORS UNIVERSITY

NAME OF INSTRUCTOR: KEIONA MIDDLETON

December 15th, 2021

**HOW DOES INCOME AFFECTS MONTHLY CHARGES ON INTERNET SERVICES?**

Part I: **Research Question**

**A.  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.**

**The description states a question or decision that can be addressed through analysis of the chosen data set.**

How does income affects monthly charges on internet services?

The question is relevant to a realistic organizational need or situation in creating innovative solution for the need of having internet security based on the customer’s income and subscriptions with regards to the rate at which internet service is used.

**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 description includes the variables in the data set and indicates the specific type of data being described, and includes examples from the data set to support claims.***

The data set is Churn CSV dataset which will be imported in order to describe the datasets to support the claims.

The below data sets contains the entire data sets variable that entails NAs

**Data Table**

import pandas as pd

# Info of column names along with the number of non –null values in each column,

df = pd.read\_csv('churn\_data')

df.info()

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 10000 entries, 0 to 9999**

**Data columns (total 52 columns):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **Variable Name** | **Non-Null Count** | **Variable Data type** | **Descriptions of the Variables** | **Examples** |
| 0 | Unnamed: 0 | 10000 non-null | int64  **quantitative** | Unnamed which means it’s a variable that maintains the actual order of the unprocessed data file that could be updated at some point, This contains 0 NA and data is complete and complete, has integers numeric. This variable is | 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20. |
| 1 | CaseOrder | 10000 non-null | int64  **quantitative** | Caseorder means the categorical sequence of the products customers purchase., contains 0 NA and data is complete and data complete, has integers, numeric. | 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20. |
| 2 | Customer\_id | 10000 non-null | Object **qualitative**  **categorial data** | Customer\_id is the unique identification for each customer, contains 0 NA and data is complete and data complete containing Var and int values. | K409198,S120509,K191035,D90850,K662701,W303516 |
| 3 | Interaction | 10000 non-null | Object **qualitative**  **categorial data** | The customer Interaction via two way communication by chats, phone or recorded contact center and technical support. Interaction shows the behavior of one variable depends on the value of another variable., contains 0 NA and data is complete and data is complete in this column containing Var and int values. | Aa90260b-4141-4a24-8e36-b04ce1f4f77b, fb76459f-c047-4a9d-8af9-e0f7d4ac2524, 344d11c-3736-4be5-98f7-c72c281e2d35, abfa2b40-2d43-4994-b15a-989b8c79e311. |
| 4 | City | 10000 non-null | Object **qualitative**  **categorial data** | City where the customer resides permanently and densely settled with administrative defined boundaries.  ,contains 0 NA and data is complete has Char. | Names of cities: point baker, west branch, yamhill, del mar, needville, fort valley. |
| 5 | State | 10000 non-null | Object **qualitative**  **categorial data** | Customers State refered to an organized political community under one government, contains 0 NA and data is complete in this column, but contains CHAR duplicate | TX,TX,GA,CA,AZ,GA,MI,VT,ID,MN,FL,CA,VA,PA,LA,NV,MA |
| 6 | County | 10000 non-null | Object **qualitative**  **categorial data** | The customers county is a political and administrative division in a state providing specifics local government services, contains 0 NA and data is complete and data is complete in this column CHAR | Prince of wales-hyder, ogemaw, yamhill, san diego, fort bend, peach, scott, oklahoma, Osceola. |
| 7 | Zip | 10000 non-null | int64  **quantitative** | Zip code of customer is a series of numbers used to communicate information of postal addresses sing five numerical digits, contains 0 NA and data is complete in this column and has interger values | 99927,48661,97148,92014,77461,31030,37847,73109,34771,45237 |
| 8 | Lat | 10000 non-null | float64  **quantitative** | Latitude of customer is a geographical coordinate that specifies the North-South position of a point on earth’s surface. contains 0 NA and data is complete in this column, this data type comprises of integer mixed with fraction presented in decimal format | 56.251, 44.32892, 45.35589,32.96687, 29.38012, 32.57032, 36.4342 |
| 9 | Population | 10000 non-null | int64  **quantitative** | Population of customer means the number of people in a city measured in quantity. Population variable contains 0 NA and data is complete in this column, this data type comprises stored values outside the range of the progress integer data type. | 38, 10446, 3735, 13863, 11352, 17701, 2535, 23144, 17351, 20193, 555, 0. |
| 10 | Lng | 10000 non-null | float64  **quantitative** | Longitude of customer is a geographical coordinate that specifies the East-West position of a point on earth’s surface. Contains 0 NA and data is complete in this column, this data type comprises of real numbers and decimal points point showing real values dividing the integer and fractional parts of the data with negative digression values | -33.3757,-84.2408,-123.2466,-117.248,-95.80673,-83.8904,-81.27892 |
| 11 | Area | 10000 non-null | Object **qualitative**  **categorial data** | An area where the customer lives is considered a unit of purpose or classification part of a town, country or region. Contains 0 NA and data is complete in this column, this have unique identity, which create meaning as a whole. CHAR | Urban, suburban, rural |
| 12 | Timezone | 10000 non-null | Object **qualitative**  **categorial data** | Time zone is a customer’s area that observes a uniform standard time for legal, commercial and social purposes. Contains 0 NA and data is complete which gives identified meaning of dfference times of diverse , locations and it contains duplicates. Has CHAR | America/New\_york, America/Chicago, America/Denver, America/New\_york, America/New\_york. |
| 13 | Job | 10000 non-null | Object **qualitative**  **categorial data** | Customer’s job refers to a task or work an individual does regularly in order to earn money. Contains 0 NA , contains CHAR, data is complete, reflects job types in different fields of professions separated by commas. | Environment health practitioner,chief financial officer,solicitor, medical illustrator, immunologist,  immigration officer  Engineer, electrical.  programmer, multimedia. |
| 14 | Children | 7505 non-null | float64  **quantitative** | The number of children the customer have which refers to a person between the stage of birth and puberty, contains 2495 Nulls and Missing not at random with number values | NA, 7, 2, 0, 5, NA, NA, NA, NA, 3, 3, 4, 3, NA ,2 ,3.  Missing data is missing based on the missing column which is variable. |
| 15 | Age | 7525 non-null | float64  **quantitative** | The customer’s Age refers to the length of time a person has lived or number of years, Contains 2475 NAs and missing completely at random with number values | 83, 83, NA, NA, 49, 86, 23, 56, 83, NA, 30, 39, 63  There is no relationship between the missingness of data and any values, observed or missing. |
| 16 | Education | 10000 non-null | Object  **qualitative**  **categorial data** | The customer’s educational qualifications reflects the acquired knowledge and skills, morals, values, personal development, Contains 0 NA and data is complete, which shows type of education completed. Contains Varchar and int | Nursey school to 8th grade, 9th grade to 12th grade no diploma, associate’s degree, bachelors degree, master’s degree, regular high school diploma, doctorate degree, GED or alternative credential, some college, less than 1 year, some college, 1 or more years, no degree |
| 17 | Employment | 10000 non-null | Object  **qualitative**  **categorial data** | This reflects the condition of having a paid work and the status of the condition of the customer, Contains 0 NA and data is complete d has CHAR with duplicates of employment status | Part time, retired, student, student, fulltime, fulltime  ,fulltime, fulltime, fulltime. unemployed, unemployed, unemployed |
| 18 | Income | 7510 non-null | float64  **quantitative** | Customer’s Income is the money received on a regular basis for work or task, Contains 2490Nulls and missing completely at random | 28561.99,21704.77,NA,18925.23,40074.19,NA  There is no relationship between the missingness of data and any values, observed or missing |
| 19 | Marital | 10000 non-null | Object **qualitative**  **categorial data** | Marital means the marital status of the customer, contains 0 NA ,Has Char, | Married, widowed, separated, never married, divorced |
| 20 | Gender | 10000 non-null | Object **qualitative**  **categorial data** | Gender is a nominal variable reflecting how the customer identifies as male or female or preferred not to answer, contains 0 NA ,Has Char | Male, female, prefer not to answer |
| 21 | Churn | 10000 non-null | Object **qualitative**  **categorial data** | Churn reflects whether the customer stopped using the service during a certain time frame. Contains 0 NA ,Has Char | No, yes, no, yes, no |
| 22 | Outage\_sec\_perweek | 10000 non-null | float64  **quantitative** | This variable reflects the amount of time of seconds per week an outage occurred in the area of customers, contains 0 NA ,has integers, and data is complete in this column, this data type comprises of integer mixed with fraction presented in decimal format | 6.9725661, 12.014541, 10.245616, 15.206193, 8.9603164 |
| 23 | Email | 10000 non-null | int64  **quantitative** | This reflects to the total number of emails received by customers previous year for marketing, contains 0 NA ,has integers , and data is complete in this column | 10, 12, 9, 15, 16, 15, 10, 16, 20, 18, 9, 17, 9, 14 |
| 24 | Contacts | 10000 non-null | int64  **quantitative** | The overall technical support contacts a customer made, contains 0 NA ,has integers , and data appears complete although it has duplicate values | 0, 0, 0, 2, 2, 3, 0, 0, 2, 1, 0, 1 |
| 25 | Yearly\_equip\_failure | 10000 non-null | int64 **quantitative** | The number of times the customer experienced a service failure yearly, contains 0 NA ,has integers, data appears complete although it has duplicate values | 1, 1, 1, 0, 1, 1, 1, 0, 3, 0, 2, 1 |
| 26 | Techie | 7523 non-null | Object **qualitative**  **categorial data** | This reflects whether the customer is highly proficient technically wise, contains 2477 Nulls and missing completely at random | yes, yes, NA, NA, no, no, no, no ,no, no, no, no, yes, no, NA, no  There is no relationship between the missingness of data and any values, observed or missing. |
| 27 | Contract | 10000 non-null | Object **qualitative**  **categorial data** | This reflects the duration of contract agreement of the customer, contains 0 NA , Has Char , data appears complete although it has duplicate values | One year, month-to-month, two year, two year, one year, month-to-month |
| 28 | Port\_modem | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer uses a portable modem, contains 0 NA ,Has Char, data appears complete | Yes, no, yes, no, yes, yes, no, no, yes, yes |
| 29 | Tablet | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer uses a tablet device like ipad, tablets, contains 0 NA ,Has Char, data appears complete | Yes, yes, no, no, no, no, no, no, no, no, no, no, no, no, no |
| 30 | InternetService | 10000 non-null | Object  **qualitative**  **categorial data** | This reflects the internet service type used by the customer, contains 0 NA ,Has Char, data appears complete | Fiber optic, fiber optic, DSL, DSL, fiber optic, none, none |
| 31 | Phone | 8974 non-null | Object **qualitative**  **categorial data** | This reflects if the customer has service on phone devices, contains 1026Nulls and missing completely at random, Has Char | Yes, yes, NA, yes, yes, yes, yes, yes, yes, NA, yes, yes, no, yes, yes, NA.  There is no relationship between the missingness of data and any values, observed or missing. |
| 32 | Multiple | 10000 non-null | Object **qualitative**  **categorial data** | This reflects whether the customer uses more than one lines, contains 0 NA ,Char, data appears complete | No, yes, yes, no, no, yes, no, no, no, yes |
| 33 | OnlineSecurity | 10000 non-null | Object **qualitative**  **categorial data** | This shows whether the customer uses online security for protection of their information, contains 0 NA ,Has Char, data appears complete | Yes, yes, no, yes, no, yes, no, no, yes |
| 34 | OnlineBackup | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer has online backup, contains 0 NA ,Has Char, data appears complete | Yes, no, no, no, no, yes, no, yes, yes |
| 35 | DeviceProtection | 10000 non-null | Object **qualitative**  **categorial data** | This reflects whether device protection is used by the customer, contains 0 NA ,Has Char, data appears complete | No, no, no, no, no, yes, no, no, no, yes, no |
| 36 | TechSupport | 9009 non-null | Object **qualitative**  **categorial data** | This reflects whether the customer uses tech support, contains 991Nulls and Missing at random | No, no, no, NA, no, no, no, no, no, yes, no, NA, no, no, yes, no, no, NA, no, yes. Missing at random because missing data is affected only by the complete observed variables and not by the characteristics of the missing data itself. |
| 37 | StreamingTV | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer has streaming TV, contains 0 NA ,Has Char, data appears complete | No, yes, no, yes, no, yes, no, no, no, yes |
| 38 | StreamingMovies | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer has streaming movies, contains 0 NA ,Has Char, data appears complete | Yes, yes, yes, no, no, yes, yes, no, no, yes, no, no, yes |
| 39 | PaperlessBilling | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer uses a paperless billing option, contains 0 NA ,Has Char, data appears complete | Yes, yes, yes , yes, no, no, no, yes, yes, yes |
| 40 | PaymentMethod | 10000 non-null | Object **qualitative**  **categorial data** | This reflects if the customer uses a specific method of payment of bills, contains 0 NA ,Has Char, data appears complete | Payment method, credit card (automatic),  bank transfer (automatic), mailed check, mailed check, electronic check, mailed check, electronic check, electronic check, mailed check, bank transfer (automatic) |
| 41 | Tenure | 9069 non-null | float64  **quantitative** | This reflects the duration the customer uses the service provided by specific service provider, contains 931Nulls and missing not at random | 13.236774, NA, 8.2206864, 3.4220861, 19.267262, NA,  13.011492, 16.87922, 10.060199, 13.870013, 15.78215, NA, 1  7.109956. Missing at random because missing data is affected only by the complete observed variables and not by the characteristics of the missing data itself. |
| 42 | MonthlyCharge | 10000 non-null | float64  **quantitative** | This reflects the customer’s monthly payments for services received , contains 0 NA ,has integers, data appears complete | 171.4497621, 242.9480155, 159.4403984, 120.2494934, 150.7612159, 184.4015581, 200.0648859, 114.7541111 |
| 43 | Bandwidth\_GB\_Year | 8979 non-null | float64  **quantitative** | This reflects the customer’s data (GB)consumption annually, contains 1021Null and missing not at random | 631, NA, 777, 139, 393, NA, 213, 443, 248, 223, NA, 188, 229, 826, NA.  There is no relationship between the missingness of data and any values, observed or missing |
| 44 | item1  Timely Response | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, data appears complete | 5, 3, 4, 4, 4, 3, 6, 2, 5, 2, 4, 4, 1, 5, 3, 3,  These (Items 1 to 8) variables shows responses to questions base on relevant factors scaling 1 to 8, with 1 = Most Agreed and 8 = Least Agreed, |
| 45 | item2  Times fixes | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, data appears complete | 5, 4, 4, 4, 4, 4, 4, 3, 5, 2 |
| 46 | item3  Timely replacements | 10000 non-null | int64  **quantitative** | Contains 0 NA, ,has integers, data appears complete | 5, 3, 2, 4, 4, 3, 6, 2, 4, 2 |
| 47 | item4 Reliability | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, ,has integers, data appears complete | 3, 3, 4, 2, 3, 2, 4, 5, 3, 2 |
| 48 | item5  Options | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, data appears complete | 4, 4, 4, 5, 4, 4, 1, 2, 4, 5 |
| 49 | item6  Respectful response | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, , data appears complete | 4, 3, 3, 4, 4, 3, 5, 3, 3, 2 |
| 50 | item7  Courteous exchange | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, data appears complete | 3, 4, 3, 3, 4, 3, 5, 4, 4, 3, 3 |
| 51 | item8  Evidence of active listening | 10000 non-null | int64  **quantitative** | Contains 0 NA ,has integers, data appears complete | 4, 4, 3, 3, 5, 3, 5, 5, 4, 3, 4 |

dtypes: float64(9), int64(15), object(28)

memory usage: 4.0+ MB

Part II: **Data-Cleaning Plan**

**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.**

***The proposal includes a detailed description of the techniques and steps needed for identifying anomalies in the selected data set.***

In the stage of data cleaning, we will perform multiple analytical techniques which goes a long way in examining the variables, analytically detect missing values and outliers.

**Plan for detecting missing Values**

**Step 1**. The entire data set with the NAs will be imported.

**Step 2**. The isnull().sum() function will be executed and this will populate the visualized variables with missing values. The missing values will be detected (NAs) by performing the is isnull().sum() function (on python) .

**Plan for detecting Outliers**

**Step 1. Histogram :** This will visualize the isolated bars within the distribution.

Use of histogram to detect outlier by applying the function below;

|  |
| --- |
| # Matplotlibis deployed to plot and Visualize bars  %matplotlib inline  import matplotlib.pyplot as plt  plt.hist(df['column'])  plt.show()  df.hist(column) |

**Step 2. Boxplot**: This will be executed to visualize the men values, dispersion of dataset, sign of skewness and summary of the data to show the outliers , therefore values that fall outside the two fences (minimum and maximum) are considered outliers. The below is the function:

|  |
| --- |
| # Using Seaborn to provides visualization  import seaborn  # Outliers are not dictated  boxplot=seaborn.boxplot(x='column',data=df) |

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

**•  characteristics of the data being assessed,**

**•  the approach used to assess the quality.**

***The justification includes the characteristics of the data being assessed and references the approach used to assess the quality of the data. The justified approach aligns with the selected data set.***

**The Reason for using isnull()sum Function**

The approach used to assess the quality of the data was to find missing values (NA)s within the data set using codes on python which were in the form of NAs. Therefore, the function used to assess the NAs values in the data was isnull()sum function respectively. Therefore working with dirty data is unavoidable, so the cleaning of the data is relevant (John Sullivan, 2018).

Therefore, the reason for using the isnull function is to make identify and locate all missing data in the churn dataset, using the children variable which contains NAs which could be misleading during the several steps of the data cleaning process therefor, we can easily replace the NAs with mean values and make sure that the original dataset is maintained.it is better to replace the NAs instead of deleting the NAs. Because deleting the NAs in bulk could possibly affect the size and quality of the entire data sets being analyzed (Larose & Larose 2019).

**The Reason for using Boxplot**

The income and children columns shows several outliers clearly, the reason is because boxplot vividly shows the distribution of that data with more detailed information.it reveals the outliers more clearly, maximum, minimum, quartile (Q1), third quartile(Q3), interquartile range (IQR), and median. The children column appears to approximately four outliers at maximum (Q3), therefore the children and income column requires to be treated using treatment technique.

The boxplot gives a consistent and uniform approach of visualizing the distribution of data established (minimum, first quartile, median, third quartile and maximum) upon how the values are spread and have the advantage of utilizing small portion of space size (Michael Galarnyk 2018).

**The Reason for using Histogram**

The use of histogram in the data detection stage helps provide a visual representation of the distribution of the children variable and income variable , because it reflects the location, spread and skewness of the data, it also helps to visualize whether the distribution is symmetric or skewed left or right. Additionally, if it is unimodal, bimodal or multimodal. It has the capability of presenting outliers and gaps in the data. Therefore, accessing data quality and finding outliers is a continuous process that allows the identification of unusual data points within a data set (Angelica Lo Duca, 2021).

**Reasons for Missing Values and Outliers**

The data being assessed contains missing values and outliers due to the following reasons :

* The data creator mistakenly skipped the row.
* There was loss of data during the process of migrating the data into a different database.
* Mistakes resulting for the programming operations.

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

***The justification describes the benefits of using the programming language, including the libraries and packages used to clean the data, and includes specific examples of how these tools are ideal in this scenario as opposed to other available tools.***

As machine learning in data analysis is skyrocketing in the contemporary business environment, there is enormous development in methods and systems in data analysis. However, in order to purify big data retrieved from diverse channels of data compilation. There is strong urge to utilize effective programming tools (Michael Galarnyk ,2018).

Therefore, for the purpose of achieving the goals of this assessment, python programming language will be used from the inception of data cleaning to the production of a clean data set respectively.

**[1.] Anaconda navigator** : The anaconda navigator serves as a geographical platform for launching python program without having to use common lines to install packages and manage the platform.

**The Reason for Using Anaconda Navigator this Report**

• This tool provided the Jupyter notebook (Python) programming language that was used in this report

**[2.] Python:** is an informative language which is well-used to read, assemble and organize code capable of execution on several interfaces, platforms designated by workload management systems due to its powerful indentation.

**The Reasons for using Python for this Report**

• Eco-friendly environment – The user friendly environment python entails provided fast and easy way to adapt to the functions and perform code execution easily.

• The python is effective for implementing, reading code, detecting missing values and outliers in the data set, and creating visual representations to spot animalities and identify outliers during the data cleaning execution

• Python made it systematic for the ‘Churn\_data.csv’ file to be imported for the cleaning operations to begin.

In addition, the python allowed the importation of packages into the new environment such as Pandas, NumPy, Matplotlib, Sklearn and Seaborn. These packages they offer variety of features such as creating visualizations of histograms, boxplot, data tables. With no doubt these packages alongside the programming languages are user-friendly, ideal and intuitive in the providing data analyst with efficiency and error free output in an innovative presentation with oppose to other tools (Michael Galarnyk ,2018).

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

***The submission provides the complete and executable code, which could be used to identify anomalies in the data set.***

|  |  |
| --- | --- |
| Code |  |
| df.isnull().sum() | This is a relevant technique that is needed to identify anomalies in the data set which is the NAs |
| # Matplotlibis deployed to plot and Visualize bars  %matplotlib inline  import matplotlib.pyplot as plt  plt.hist(df['column'])  plt.show()  df.hist(column) | This will create a histogram to show missing values. |
| # Using Seaborn to provides visualization  import seaborn  # Outliers are not dictated  boxplot=seaborn.boxplot(x='column',data=df) | This will create boxplots to visualize the outliers present. |

**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.**

***The description accurately includes all of the anomalies found by running the code from part C4.***

Findings achieved by running the code from part c4 indicates that we have NAs missing values in the data after executing the below isnull function.

|  |
| --- |
| df.isnull().sum()  Unnamed: 0 0  CaseOrder 0  Customer\_id 0  Interaction 0  City 0  State 0  County 0  Zip 0  Lat 0  Lng 0  Population 0  Area 0  Timezone 0  Job 0  Children 2495  Age 2475  Education 0  Employment 0  Income 2490  Marital 0  Gender 0  Churn 0  Outage\_sec\_perweek 0  Email 0  Contacts 0  Yearly\_equip\_failure 0  Techie 2477  Contract 0  Port\_modem 0  Tablet 0  InternetService 0  Phone 1026  Multiple 0  OnlineSecurity 0  OnlineBackup 0  DeviceProtection 0  TechSupport 991  StreamingTV 0  StreamingMovies 0  PaperlessBilling 0  PaymentMethod 0  Tenure 931  MonthlyCharge 0  Bandwidth\_GB\_Year 1021  item1 0  item2 0  item3 0  item4 0  item5 0  item6 0  item7 0  item8 0  dtype: int64 |

**# Results found during Detecting Outliers**

***Findings achieved by running the code from part c4 indicates that we have outliers in the data after executing the below functions***

|  |
| --- |
| import numpy as np  import pandas as pd  from pandas import DataFrame  # Matplotlibis deployed to plot and Visualize bars  %matplotlib inline  import matplotlib.pyplot as plt  plt.hist(df[‘Age'])  plt.show()  A picture containing histogram  Description automatically generated  # Using Seaborn to provides visualization  import seaborn  **# Outliers are not dictated**  boxplot=seaborn.boxplot(x='Age',data=df)  Chart  Description automatically generated  plt.hist(df['Children'])  plt.show()  Chart, histogram  Description automatically generated  **# Outliers are dictated**  boxplot=seaborn.boxplot(x='Children',data=df)  Chart  Description automatically generated  plt.hist(df['Income'])  plt.show()  Chart, histogram  Description automatically generated  **# Outliers are dictated**  boxplot=seaborn.boxplot(x='Income',data=df)  Chart  Description automatically generated  plt.hist(df['Tenure'])  plt.show()  Chart, histogram  Description automatically generated  **# Outliers are not dictated**  boxplot=seaborn.boxplot(x='Tenure',data=df)  Chart  Description automatically generated  plt.hist(df['Bandwidth\_GB\_Year'])  plt.show()  Logo  Description automatically generated with low confidence  **# Outliers are not dictated**  boxplot=seaborn.boxplot(x='Bandwidth\_GB\_Year',data=df)  Chart, histogram  Description automatically generated  df['Income'].quantile  <bound method Series.quantile of 0 28561.990000  1 21704.770000  2 39936.762226  3 18925.230000  4 40074.190000  ...  9995 55723.740000  9996 39936.762226  9997 39936.762226  9998 16667.580000  9999 39936.762226  Name: Income, Length: 10000, dtype: float64>  df['Age'].quantile  <bound method Series.quantile of 0 68.000000  1 27.000000  2 50.000000  3 48.000000  4 83.000000  ...  9995 53.275748  9996 48.000000  9997 53.275748  9998 39.000000  9999 28.000000  Name: Age, Length: 10000, dtype: float64>  df['Children'].quantile  <bound method Series.quantile of 0 1.0  1 1.0  2 4.0  3 1.0  4 0.0  ...  9995 3.0  9996 4.0  9997 1.0  9998 1.0  9999 1.0  Name: Children, Length: 10000, dtype: float64> |

From the above findings, there are NAs values in these columns; Children[2495], Age[2475], Income[2490], Techie [2477], Phone [1026], TechSupport [991], Tenure [931]and Bandwidth\_GB\_Year [1021].

However it also appears that there are outliers in the income variable and children variable and require treatment in the next phase of the data cleaning process.

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

***The justification includes the specific mitigation methods for each type of anomaly listed in part D1.***

In this stage we are going to perform methods and techniques to treat missing values and outliers in order to arrive with a cleaned and corrupt free dataset.

**Why we used the Two Treatment Methods**

**- Univariate Imputation:** The reason for applying this method in treatment of data is due to the fact that this method is effective in replacing the missing values of specific variables having animalities by applying mean, median and mode because this method is highly relevant when the percentage of the missing data is less. Therefore, this loss of data can be avoided using this method. It is chosen because it is the best technique if the data size appears to be less and it goes a long way in analytically avoiding the depreciation of data in the deletion of rows and columns respectively.

**- Deletion/Dropping method:** This method will be used to treat missing data actively present in the rows and columns. The reason this method us due to the fact that dropping missing data could lead to enormous and huge error free output. Therefore dropping a specific row or column with certain data is preferred due to the fact that the value isn’t containing an inclined weightage.

**Before Imputation**

Examination of the distribution and treating missing values

|  |
| --- |
| # TREATING MISSING VALUES  # EXAMINE THE DISTRIBUTION  df.hist()  array([[<AxesSubplot:title={'center':'Unnamed: 0'}>,  <AxesSubplot:title={'center':'CaseOrder'}>,  <AxesSubplot:title={'center':'Zip'}>,  <AxesSubplot:title={'center':'Lat'}>,  <AxesSubplot:title={'center':'Lng'}>],  [<AxesSubplot:title={'center':'Population'}>,  <AxesSubplot:title={'center':'Children'}>,  <AxesSubplot:title={'center':'Age'}>,  <AxesSubplot:title={'center':'Income'}>,  <AxesSubplot:title={'center':'Outage\_sec\_perweek'}>],  [<AxesSubplot:title={'center':'Email'}>,  <AxesSubplot:title={'center':'Contacts'}>,  <AxesSubplot:title={'center':'Yearly\_equip\_failure'}>,  <AxesSubplot:title={'center':'Tenure'}>,  <AxesSubplot:title={'center':'MonthlyCharge'}>],  [<AxesSubplot:title={'center':'Bandwidth\_GB\_Year'}>,  <AxesSubplot:title={'center':'item1'}>,  <AxesSubplot:title={'center':'item2'}>,  <AxesSubplot:title={'center':'item3'}>,  <AxesSubplot:title={'center':'item4'}>],  [<AxesSubplot:title={'center':'item5'}>,  <AxesSubplot:title={'center':'item6'}>,  <AxesSubplot:title={'center':'item7'}>,  <AxesSubplot:title={'center':'item8'}>, <AxesSubplot:>]],  dtype=object) |
| # Skewed distribution positively skewed right  df.hist(column='Children')  array([[<AxesSubplot:title={'center':'Children'}>]], dtype=object)    # Equally Spread (uniform distribution)  df.hist(column='Age')  array([[<AxesSubplot:title={'center':'Age'}>]], dtype=object)    # Categorical variables  # Skewed distribution positively skewed right  df.hist(column='Income')  array([[<AxesSubplot:title={'center':'Income'}>]], dtype=object)    # Bi modal distribution with two modes non-symmetric  df.hist(column='Tenure')  array([[<AxesSubplot:title={'center':'Tenure'}>]], dtype=object)    # Bi modal distribution non-symmetric  df.hist(column='Bandwidth\_GB\_Year')  array([[<AxesSubplot:title={'center':'Bandwidth\_GB\_Year'}>]], dtype=object) |

**Performing Treatment of Null using Imputation method**

|  |
| --- |
| # Median imputation approach is applied to preserve the integrity of the data, the variable children appears to be a skewed distribution, therefore all the NAs will be replaced with median.  df['Children'].fillna(df['Children'].median(), inplace=True) |
| #Mean imputation will be applied because age appears to be a uniform distribution that why we are using imputation method to replace the NAs with mean.  df['Age'].fillna(df['Age'].mean(), inplace=True) |
| # The income variable is a skewed distribution that is positively skewed to the right and the best way to treat the NAs is by using the mean because this would impute and replace the NAs.  df['Income'].fillna(df['Income'].mean(), inplace=True) |
| # The tenure variable appears to be Bi modal distribution with two modes non-symmetric. therefore the mean will be imputed in order to treat the NAs in this variable.  df['Tenure'].fillna(df['Tenure'].mean(), inplace=True) |
| # The bandwidth\_GB\_year appears to be Bi modal distribution non-symmetric, therefore the best way to treat it is by imputation of the mean to replace the NAs.  df['Bandwidth\_GB\_Year'].fillna(df['Bandwidth\_GB\_Year'].mean(), inplace=True) |
| **Performing Treatment of NA Using Deletion/Dropping Method**  df['TechSupport'].fillna(df['TechSupport'].mode()[0])  0 No  1 No  2 No  3 No  4 Yes  ...  9995 No  9996 No  9997 No  9998 Yes  9999 No  Name: TechSupport, Length: 10000, dtype: object  df.dropna(subset=['TechSupport'])  df.dropna()  df.dropna(subset = ["TechSupport"], inplace=True)  df.replace('TechSupport', " ", inplace=True)  nan\_value = float("NAN")  df.replace("", nan\_value, inplace=True)  df.dropna(subset = ["TechSupport"], inplace=True) |
| df['Techie'].fillna(df['Techie'].mode()[0])  0 No  1 Yes  2 Yes  3 Yes  4 No  ...  9995 No  9996 No  9997 No  9998 No  9999 No  Name: Techie, Length: 10000, dtype: object  df.dropna(subset=['Techie'])  df.dropna()  df.dropna(subset = ["Techie"], inplace=True)  df.replace('Techie', " ", inplace=True)  nan\_value = float("NAN")  df.replace("", nan\_value, inplace=True)  df.dropna(subset = ["Techie"], inplace=True) |
| df['Phone'].fillna(df['Phone'].mode()[0])  0 Yes  1 Yes  2 Yes  3 Yes  4 No  ...  9995 Yes  9996 Yes  9997 Yes  9998 No  9999 Yes  Name: Phone, Length: 10000, dtype: object  df.dropna(subset=['Phone'])  df.dropna()  df.dropna(subset = ["Phone"], inplace=True)  df.replace('Phone', " ", inplace=True)  nan\_value = float("NAN")  df.replace("", nan\_value, inplace=True)  df.dropna(subset = ["Phone"], inplace=True) |

**The Treatment of Outliers**

This is achieved by dropping the outliers systematically using the Cleaned\_df.head()

Function in other to clean the outliers from the children and income columns

|  |
| --- |
| Cleaned\_df = df[df['Children'] < 100]  Cleaned\_df.head()  df.duplicated()  0 False  1 False  2 False  3 False  4 False  ...  9995 False  9996 False  9997 False  9998 False  9999 False  Length: 10000, dtype: bool |

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

***The summary details the outcome from the implementation of each data-cleaning step. The summarized expected outcomes are plausible given the interventions.***

**After Imputation and Deleting/Dropping Method has been applied**

The outcome shows the result of the NAs were treated using the imputation and Deletion/Dropping method. By performing the isnull().sum function, we can clearly examine and observe that the NAs have been treated using the imputation and Deletion/Dropping method successfully.

**Cleaned data showing 0 NA**

|  |
| --- |
| df.isnull().sum()  Unnamed: 0 0  CaseOrder 0  Customer\_id 0  Interaction 0  City 0  State 0  County 0  Zip 0  Lat 0  Lng 0  Population 0  Area 0  Timezone 0  Job 0  Children 0  Age 0  Education 0  Employment 0  Income 0  Marital 0  Gender 0  Churn 0  Outage\_sec\_perweek 0  Email 0  Contacts 0  Yearly\_equip\_failure 0  Techie 0  Contract 0  Port\_modem 0  Tablet 0  InternetService 0  Phone 0  Multiple 0  OnlineSecurity 0  OnlineBackup 0  DeviceProtection 0  TechSupport 0  StreamingTV 0  StreamingMovies 0  PaperlessBilling 0  PaymentMethod 0  Tenure 0  MonthlyCharge 0  Bandwidth\_GB\_Year 0  item1 0  item2 0  item3 0  item4 0  item5 0  item6 0  item7 0  item8 0  dtype: int64 |

**4.  Provide the code used to mitigate anomalies**.

***The submission provides complete and executable code that could be used to mitigate the anomalies.***

The below codes were used in performing treatment of NAs by replacement of NAs in the Variables of children, age, income, techie, and tech support using univariate imputation method median, mean by the execution of df.dropna function and df[‘column’].fillna.

|  |
| --- |
| df['Children'].fillna(df['Children'].median(), inplace=True) |
| df['Age'].fillna(df['Age'].mean(), inplace=True) |
| df['Bandwidth\_GB\_Year'].fillna(df['Bandwidth\_GB\_Year'].mean(), inplace=True) |
| df['Income'].fillna(df['Income'].mean(), inplace=True) |
| df['Tenure'].fillna(df['Tenure'].mean(), inplace=True) |
|  |

|  |
| --- |
| df['TechSupport'].fillna(df['TechSupport'].mode()[0])  df.dropna(subset=['TechSupport'])  df.dropna()  df.dropna(subset = ["TechSupport"], inplace=True)  df.replace('TechSupport', " ", inplace=True)  nan\_value = float("NAN")  df.replace("", nan\_value, inplace=True)  df.dropna(subset = ["TechSupport"], inplace=True)  df['Phone'].fillna(df['Phone'].mode()[0])  df.dropna(subset=['Phone'])  df.dropna()  df.dropna(subset = ["Phone"], inplace=True)  df.replace('Phone', " ", inplace=True)  nan\_value = float("NAN")  df.replace("", nan\_value, inplace=True)  df.dropna(subset = ["Phone"], inplace=True) |

**5.  Provide a copy of the cleaned data set.**

***The submission includes a clean data set created from the raw data.***

The provided data set includes the complete list of variables from the chosen data set in part A.

icu\_csv\_file.csv

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

***The submission accurately summarizes the limitations of the implemented data-cleaning process.***

With no doubt the maim goal is to have an unbiased and error free data, although there are certain setbacks entails which include the following

- There is possibility of data shrinkage during the process of cleaning the data which could result from improper usage of the programming languages.

- The data analyst could experience the possibility of consuming a larger amount of time schedule if the data is enormous in size.

- The Univariate Imputation Technique also has certain limitations when it comes to the analysis because it could possibly account for ambiguity and confusion as a result of lost data.

- There is quantum of data decrease and disappearance in the process and the essential amount of the sample data is diminished.

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

***The submission includes a discussion of the impact of the limitations from part D6. The discussion logically aligns with the question or decision from part A.***

In the contemporary business environment information erupts and data multiply without boundaries and the need for data analysis and cleaning will always be harnessed to the highest capacity.

This report has come a long way in revealing the fact that the data set entailed missing values among the variables and several techniques were executed to those variables which helped to treat missing values however it is vehement to note that is these techniques were avoided the output will be biased and unpurified and could result in impracticable problem solving operations

**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.**

[1.] Principal component 1 : This appears to be linear putting together the main predictable variables that entails the enormous variance within the data set.(it narrows the total of squared )

[2.]Principal component 2 : This is a straight mixture of the main predictors that contain the rest of the variance within the data set which appears to be uncorrelated with the variables in principal component 1 respectively.

***The submission lists all principal components of the data set.***

According to Carolina Bento (2020), Principle component analysis is a mechanism of acquiring extensive variables (in form of components) from a gigantic set of variables accessible in a data set.

['Zip', 'Lat', 'Outage\_sec\_perweek', 'MonthlyCharge', 'Children', 'Age', 'Income', 'Tenure', 'Bandwidth\_GB\_Year']

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

***The description of how the principle components of the data set were identified is accurate and complete.***

Standardization: It is important to standardized the data in order to predict accuracy in machine learning and algorithm .

PCA Projection to 2D: During the process of PCA, the original data has 9 columns ['Zip', 'Lat', 'Outage\_sec\_perweek', 'MonthlyCharge', 'Children', 'Age', 'Income', 'Tenure', 'Bandwidth\_GB\_Year']. However, the code executed upon the original data which is 9 dimensional into 2 dimentionals(2 dimensions of variations).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # pca analysis  import pandas as pd  import numpy as np  import matplotlib.pyplot as plt  from sklearn.decomposition import PCA  from sklearn.preprocessing import StandardScaler  %matplotlib inline  df = pd.read\_csv('icu\_csv\_file.csv')  df.head(5)  features = ['Zip', 'Lat', 'Outage\_sec\_perweek', 'MonthlyCharge', 'Children', 'Age', 'Income', 'Tenure', 'Bandwidth\_GB\_Year']  x = df.loc[:, features].values  y = df.loc[:,['Income']]  x = StandardScaler().fit\_transform(x)  pd.DataFrame(data = x, columns = features).head()   |  | **Zip** | **Lat** | **Outage\_sec\_perweek** | **MonthlyCharge** | **Z\_Score\_Children** | **Z\_Score\_Age** | **Z\_Score\_Income** | **Z\_Score\_Tenure** | **Z\_Score\_Bandwidth\_GB\_Year** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **0** | 1.837971 | 3.201134 | -0.636575 | -0.060044 | -0.434856 | 0.816535 | -0.470630 | -1.090283 | -1.193002 | | **1** | -0.014872 | 1.019010 | 0.082112 | 1.588539 | -0.434856 | -1.458704 | -0.754117 | -1.314279 | -1.242887 | | **2** | 1.737533 | 1.206977 | -0.170032 | -0.336952 | 1.124163 | -0.182350 | -0.000382 | -0.734411 | -0.638921 | | **3** | 1.551982 | -1.060615 | 0.537052 | -1.240604 | -0.434856 | -0.293337 | -0.869027 | -0.681455 | -0.585991 | | **4** | 1.026011 | -1.717106 | -0.353240 | -0.537074 | -0.954530 | 1.648940 | 0.005299 | -1.293849 | -1.497962 |   pca = PCA(n\_components=2)  principalComponents = pca.fit\_transform(x)  principalDf = pd.DataFrame(data = principalComponents  , columns = ['principal components 1', 'principal components 2'])  principalDf.head(5)   |  | **principal components 1** | **principal components 2** | | --- | --- | --- | | **0** | -1.588990 | -0.232701 | | **1** | -1.711159 | 1.398628 | | **2** | -0.963462 | -0.360283 | | **3** | -0.927081 | -0.640579 | | **4** | -2.002967 | -0.755230 |   df[['Z\_Score\_Income']].head(5)   |  | **Z\_Score\_Income** | | --- | --- | | **0** | -0.470630 | | **1** | -0.754117 | | **2** | -0.000382 | | **3** | -0.869027 | | **4** | 0.005299 |   finalDf = pd.concat([principalDf, df[['Z\_Score\_Income']]], axis = 1)  finalDf.head(5)   |  | **principal components 1** | **principal components 2** | **Z\_Score\_Income** | | --- | --- | --- | --- | | **0** | -1.588990 | -0.232701 | -0.470630 | | **1** | -1.711159 | 1.398628 | -0.754117 | | **2** | -0.963462 | -0.360283 | -0.000382 | | **3** | -0.927081 | -0.640579 | -0.869027 | | **4** | -2.002967 | -0.755230 | 0.005299 |   fig = plt.figure(figsize = (8, 8))  ax = fig.add\_subplot(1,1,1)  ax.set\_xlabel('Principal Component 1', fontsize = 15)  ax.set\_ylabel('Principal Component 2', fontsize = 15)  ax.set\_title('2 Component PCA', fontsize = 15)  Z\_Score\_Income = ['-0.470630', '-0.754117', '0.005299']  colors = ['b', 'g', 'r']  for Z\_Score\_Income, color in zip(Z\_Score\_Income,colors):  indicesToKeep = finalDf['Z\_Score\_Income'] == Z\_Score\_Income  ax.scatter(finalDf.loc[indicesToKeep, 'principal components 1']  , finalDf.loc[indicesToKeep, 'principal components 2']  , c = color  , s = 40)  ax.legend(Z\_Score\_Income)  ax.grid()    pca.explained\_variance\_ratio\_  array([0.21122418, 0.12641358]) |

Conclusively, the variance inform us on how the data was assigned separately among the two principal components. Therefore the first two components have 33.76% of data. The first principal component contains 21.12% variance and the second principal component 12.64% variance. Therefore, the third, fourth, fifth, sixth, seventh, eighth and nineth principal component entails the remaining variance of the processed data set respectively.

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

***The description of how the organization can benefit from the results of the PCA is logical and accurate.***

The fundamental recognition of PCA is dimensionality (appearance) contraction. It advance algorithm execution by eliminating related appearance nevertheless there are some data shrinkage. However, Principal Component Analysis(PCA) is extensively well-used in minimization of great proportion. Therefore, Evidently, We invariably extract tremendous dimensional statistics due to numeric as well as categorical data, and PCA is a helping hand in this scenario. That’s the why it is vehement to take systematic consideration when cleaning data. Most importantly, these also concentrate principally according to how PCA is practically, comparatively than mathematical expression and some relevant points position focal point that requires maximum consideration when performing PCA . diversely, the organization could ascertain erroneous data.

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.

Rubric The Panopto video recording demonstrates the warning-and error-free functionality of the code used to support the discovery of anomalies and the data cleaning process. An accurate summary of the programming environment is provided in the video.

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

***The submission records all web sources used to acquire data or third-party code and all of the web sources are reliable.***

Deepanshu, B. (2020). *How to Import Data in Python.*

RSGB Business Consultant Pvt. Ltd. <https://www.listendata.com/2017/02/import-data-in-python.html#Import-CSV-files>

Carolina B. (2020). *Principal Component Analysis algorithm in Real-Life: Discovering patterns in a real-estate dataset.*

Medium. <https://towardsdatascience.com/principal-component-analysis-algorithm-in-real-life-discovering-patterns-in-a-real-estate-dataset-18134c57ffe7>

Pierre-Louis B. (2020). *Principle Components Analysis(PCA), Fundamentals, Benefits & Insights for Industry.*

Medium. <https://towardsdatascience.com/principal-components-analysis-pca-fundamentals-benefits-insights-for-industry-2f03ad18c4d7>

John S. (2018). *Data Cleaning with python and Pandas: Detecting Missing Values.*

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

Angelica Lo D. (2021). *How to detect outliers with Pytho pyod.*

Medium. <https://towardsdatascience.com/how-to-detect-outliers-with-python-pyod-aa7147359e4b>

Michael G. (2018). *Understanding Boxplots.*

Medium. <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

**H.** **The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.**

**References**

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. ISBN-13: 978-1-119-52684-1.