Western Governors University (WGU)

D206 Data Cleaning Performance Assessment

Natallia Zimnitskaya | ID: 012247127

Master of Science, Data Analytics

***Part I: Research Question and Variables.***

***A1.*** The research question for this project is: What factors influence high readmission rates? Readmission rates are important to understand because they impact patient safety, healthcare spending, resource allocation, and policy decisions.

***A2.*** The data set from the medical\_raw\_data.csv file will be analyzed to answer this question. This data set consists of 53 columns and contains 10000 entries.



All variables and explanations in the medial\_raw\_data.csv file are listed in the tab below.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable name | Data type | Description | Example (row #5) |
| Unnamed: 0 | Quantitative | It does not have a name, and the numbers from 1 to 1000 go down in each row. This column functions as an index included in the data | 5 |
| CaseOrder | Quantitative | It is a sequence number, maintaining the original order of the raw data file to ensure that data remains in the same sequence as it was initially recorded | 5 |
| Customer\_id | Qualitative | Unique patient ID | C544523 |
| Interaction | Qualitative | Unique patient ID that identifies patient admissions, procedures, and transactions | 5885f56b-d6da-43a3-8760-83583af94266 |
| UID | Qualitative | d2f0425877b10ed6bb381f3e2579424a |
| City | Qualitative | City of residence | West Point |
| State | Qualitative | State of residence | VA |
| County | Qualitative | County of residence | King William |
| Zip | Qualitative | Zip code of residence | 23181 |
| Lat | Quantitative | Latitude GPS coordinate of residency | 37.59894 |
| Lng | Quantitative | Longitude GPS coordinate of residency | -76.8896 |
| Population | Quantitative | Population within a mile radius of a patient | 5287 |
| Area | Qualitative | Area type (rural, urban, suburban) | Rural |
| Timezone | Qualitative | Time zone of residency | America/New\_York |
| Job | Qualitative | The job of a patient or a primary insurance holder | Health promotion specialist |
| Children | Quantitative | # of children in the household | NA |
| Age | Quantitative | Patient's age | 22 |
| Education | Qualitative | Highest earned degree | Regular High School Diploma |
| Employment | Qualitative | Patient's employment status | Full Time |
| Income | Quantitative | The patient's (primary insurance holder's) annual income | 1209.56 |
| Martial | Qualitative | Patient's (primary insurance holder's) marital status | Widowed |
| Gender | Qualitative | Self-identification as male, female, or nonbinary | Female |
| ReAdmis | Qualitative | Readmission within a month of release (yes, no) | No |
| VitD\_levels | Quantitative | Vit D levels in ng/ml | 16.87052 |
| Doc\_visits | Quantitative | # of primary physician visits during the initial hospitalization | 5 |
| Full\_meals\_eaten | Quantitative | # of full meals during hospitalization (0-partial meals, and some patients had more than three meals in a day if requested) | 0 |
| VitD\_supp | Quantitative | # of times of vit D supplement administration | 2 |
| Soft\_drink | Qualitative | Habitual drinking of soft drinks (yes, no) | Yes |
| Initial\_admin | Qualitative | Type of initial admission (emergency, elective, observation) | Elective Admission |
| HighBlood | Qualitative | Patient has a high blood pressure (yes, no) | No |
| Stroke | Qualitative | Patient has had a stroke (yes, no) | No |
| Complication\_risk | Qualitative | Level of complication risk (high, medium, low) | Low |
| Overweight | Qualitative | Patient is overweight (yes, no) | 0 |
| Arthritis | Qualitative | Patient has arthritis (yes, no) | No |
| Diabetes | Qualitative | Patient has diabetes (yes, no) | No |
| Hyperlipidemia | Qualitative | Patient has hyperlipidemia (yes, no) | Yes |
| BackPain | Qualitative | Patient has a chronic back pain (yes, no) | No |
| Anxiety | Qualitative | Patient has anxiety disorder (yes, no) | 0 |
| Allergic\_rhinitis | Qualitative | Patient has allergic rhinitis (yes, no) | Yes |
| Reflux\_esophagitis | Qualitative | Patient has reflux esophagitis (yes, no) | No |
| Asthma | Qualitative | Patient has asthma (yes, no) | No |
| Services | Qualitative | Services received while hospitalized (blood work, intravenous, CT scan, MRI) | CT scan |
| Initial\_days | Quantitative | Number days in the hospital | 1.254807 |
| TotalCharge | Quantitative | Amount charged daily (total charge/number of days) | 1885.655 |
| Additional\_charges | Quantitative | Charge for miscellaneous procedures | 3716.526 |
| Item1 | Qualitative | Timely admission (patient’s survey (1 = most important to 8 = least important)) | 2 |
| Item2 | Qualitative | Timely treatment (patient's survey 1 to 8) | 1 |
| Item3 | Qualitative | Timely visits (patient's survey 1 to 8) | 3 |
| Item4 | Qualitative | Reliability (patient's survey 1 to 8) | 3 |
| Item5 | Qualitative | Options (patient's survey 1 to 8) | 5 |
| Item6 | Qualitative | Hours of treatment (patient's survey 1 to 8) | 3 |
| Item7 | Qualitative | Courteous staff (patient's survey 1 to 8) | 4 |
| Item8 | Qualitative | Active listening from a doctor (patient's survey 1 to 8) | 3 |

***Part II: Data-Cleaning Plan (Detection)***

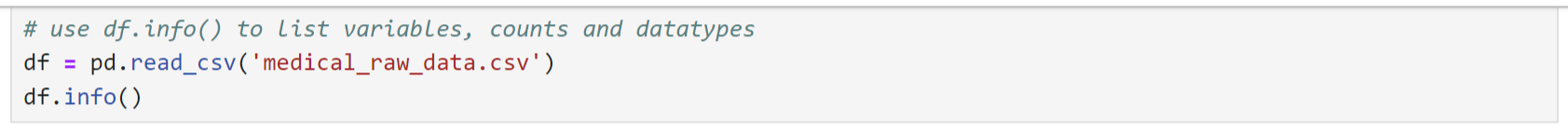
***B1. Functions and commands used to detect data quality issues.***

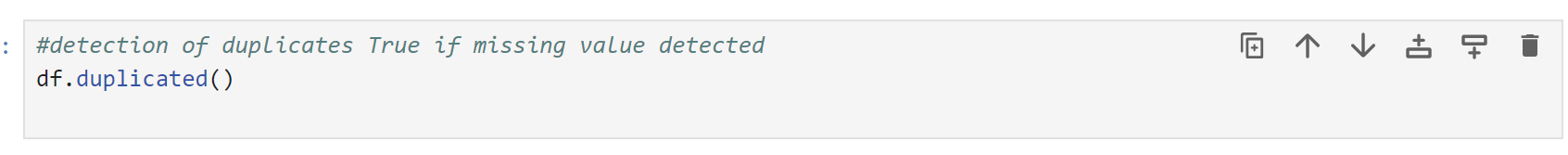
Editing, correcting, and organizing data within a data set to make it generally uniform and ready for analysis is known as data cleaning. This includes deleting data that is irrelevant or corrupt and formatting the data into a computer-readable format for the best possible analysis (*Data Cleaning Steps & Process to Prep Your Data for Success*, 2021).

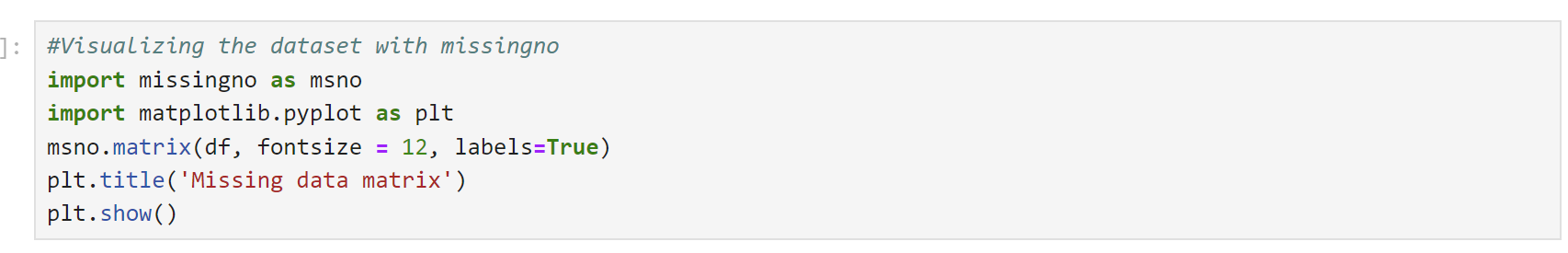
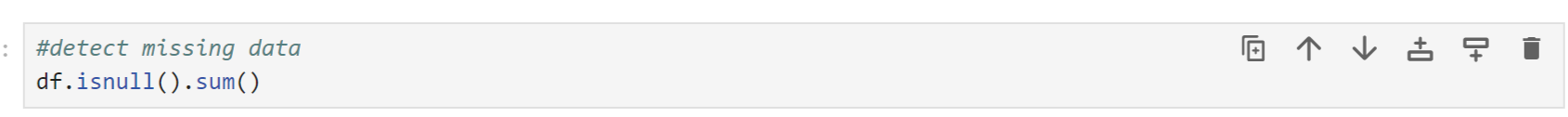
To evaluate the quality of a data set, the following procedures needs to be followed:

* The **df.info()** function will examine and analyze the data to create a summary. Irrelevant data must be detected to eliminate any observations or data that doesn't pertain to the immediate needs.
* The **duplicated()** and **value\_counts()** functionsdetect duplicates.Duplicate data increases storage needs and slows down analysis. What's even more significant is that, depending on how frequently a result is duplicated, a machine learning model trained on a dataset containing duplicate results will probably give the duplicates more weight.
* The **isnull().sum** function detects the missing values (qualitative and quantitative variables) usually represented in a dataset as Nan, None or null. The **missingno** package helps to visualize the missing data. It's essential to ensure the accuracy and integrity of data analysis for reliable conclusions and decision-making.
* The **Z-scores** (**SciPy** package), **boxplots** (**Seaborn** and **Matplotlib** packages), or **histogram** methods detect the outliers for all quantitative variables. Data points known as outliers deviate significantly from the norm and can overly bias the analysis in one direction.
* The **unique()** and **replace()** functions will re-express a categorical value to numerical based on rank (the **Ordinal Encoder technique**, using a Likert scale). It's essential to do when there is an inconsistency in the presentation of qualitative variables (for example, some may be presented as yes/no while anothers as 1/0)

***B2. Why the functions and commands discussed in B1 were used.***

* The **df.info()** function was utilized because it provides a concise summary of a data frame (number of rows and columns, column names and their data types, and non-null values count for each column). 
* The **duplicated()** method in pandas is used to locate, extract, and count duplicate rows in a data frame. By default, rows are considered duplicates if all column values are equal. The duplicated() method returns a Boolean Series that marks duplicate rows as True (Nkmk, 2024). The **value\_counts()** function provides a sum of duplicated records. When duplicates are detected, the **drop\_duplicates()** function is used for deduplicating data.



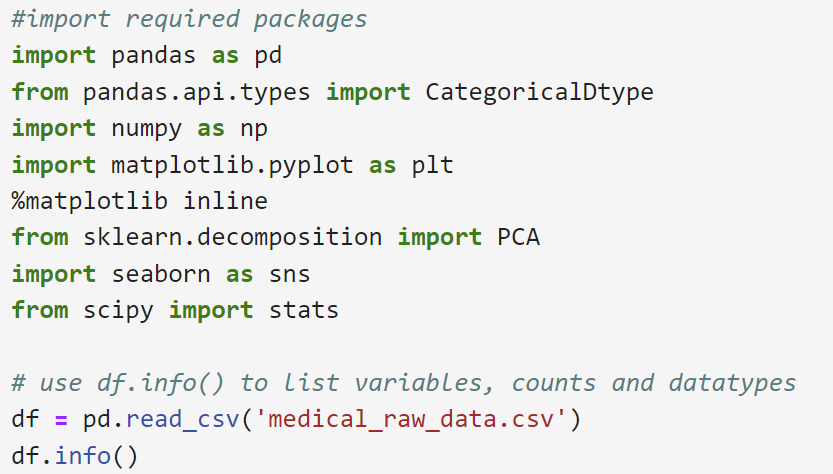
* The **isnull().sum()** function not only checks missing or null values, but also provides the sum of missing values in each column. The **msno.matrix(df)** function(from **Missingno** package) helps to visualize missing data. Also, **plt.hist()** function (from **Matplotlib** package ) creates histogram for additional visualization to see the missing values distribution for further treatment. 
* An outlier is extremely small or large data value relative to the rest of the data set. It can be a result of an error or a genuine data. A **Z-score** is a numerical measurement that characterizes a value's relationship to the mean. The more unusual an observation is, the further away its Z-score is from zero. Z-scores of +/-3 or further from zero are typically used as a cut-off value for identifying outliers. Z-scores are typically used in the context of the normal curve and their interpretation based on the standard normal table (<i>Redirecting</i>, n.d.). The **df['NewColumnForZScores']=stats.zscore(df['columnname'])** function calculates z-scores for a column and the **df[['Columnname',' Columnforzscore']].head** function displays calculated z-scores and the values. The histogram method is one of the tools that helps visualize frequency of distribution and an existence of outliers (isolated bars). The **boxplot** method is also used for visualization. It provides mean values, skewness and outliers (the values outside the lower and upper wiskers). The **boxplot=seaborn.boxplot(x=‘Columnname',data=df)** function creates a boxplot.
* Qualitative or categorical data (i.e. yes/no) requires re-expression or encoding of numbers to perform statical modeling. Ordinal encoding can be done by using **unique()** fuction wich determines the number of unique values for the variable. After replication of the variable for storing its new numeric values after conversion and creation a specific dictionary (dict\_edu) for these values, the **replace** function will replace the values according to the rules and store in existing data frame.

B3. Programming language and packages.

After reading "R vs. Python: 12 Key Comparisons" Python was chosen to complete this performance assessment. High-level, general-purpose programming language Python is flexible, object-oriented, open-source, and emphasizes code readability with a clear visual structure and easy syntax. Since Python is open source, many people can contribute to its development and improve its libraries and features. Python comes with many necessary libraries to carry out data science-related tasks, and its integration and control capabilities boost productivity. Python is a standard programming language that is easy for beginners to learn and understand because of its simple syntax; it requires fewer lines of code to be written and is easy to read; for data science projects, Python takes a more streamlined approach; it has a wide range of libraries that allow users to input the action of the library into the code, making it simple to perform matrix computations and optimization (BasuMallick, 2022). However, Compared to R, Python has some disadvantages: fewer data science-specific libraries, visualizations are less visually appealing and informative also more complicated, requires thorough testing because mistakes appear during runtime.

The following Python packages and libraries help with this assignment's data cleaning process:

* **Pandas:** used for data analyzing, cleaning, exploring, and manipulating.
* **NumPy:** used to working with arrays.
* **Matplotlib:** used for visualization utility.
* **Missingno:** used to visualization a distribution of missing values
* **Seaborn:** used for advanced visualization
* **spicy.stats:** used for normalization and statistics
* **sklearn.decomposition:** used for PCA analysis



***Part III: Data Cleaning (Treatment)***

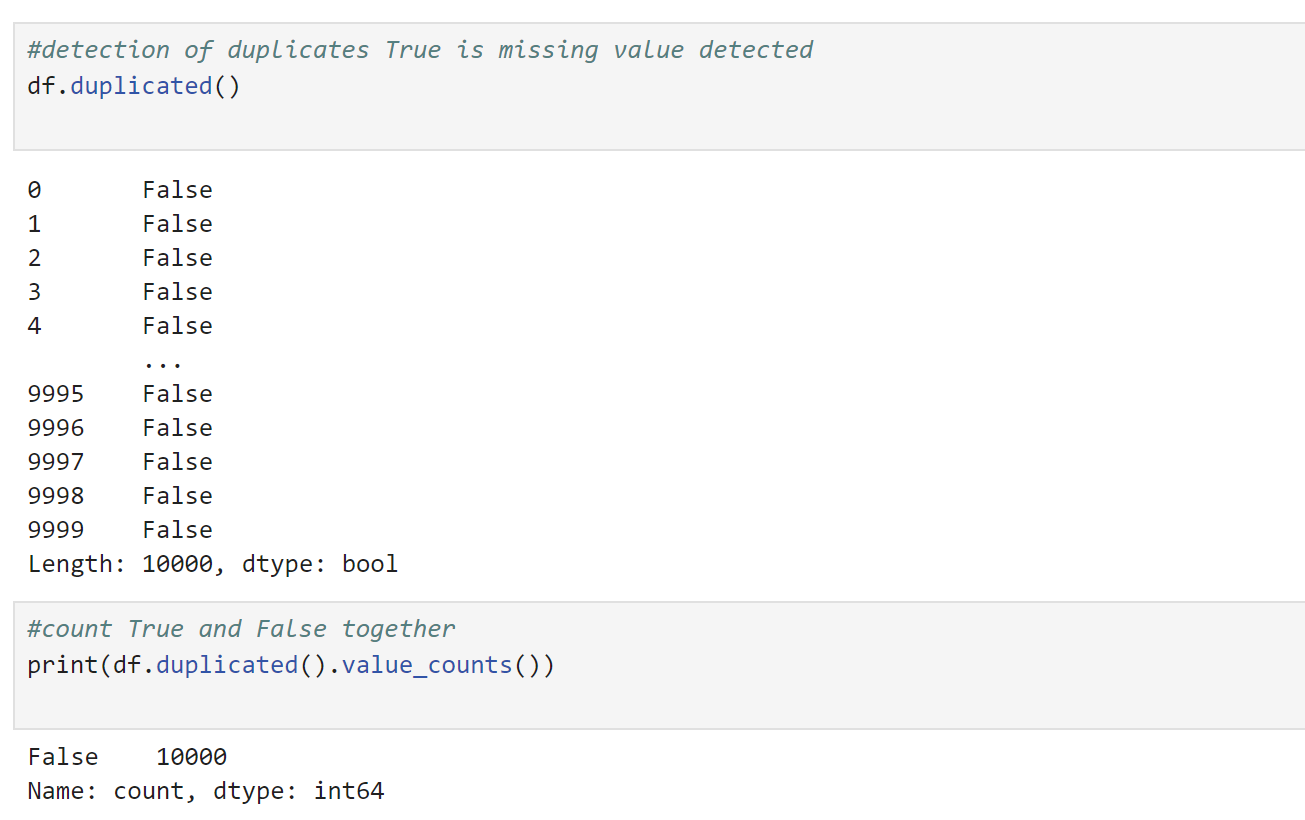
***C1. Findings after checking for duplicates, missing values, outliers, etc.***

The Jupiter Notebook was used to complete this assessment. The required packages were imported and the medical\_raw\_data.csv file was loaded.



**Duplicates**

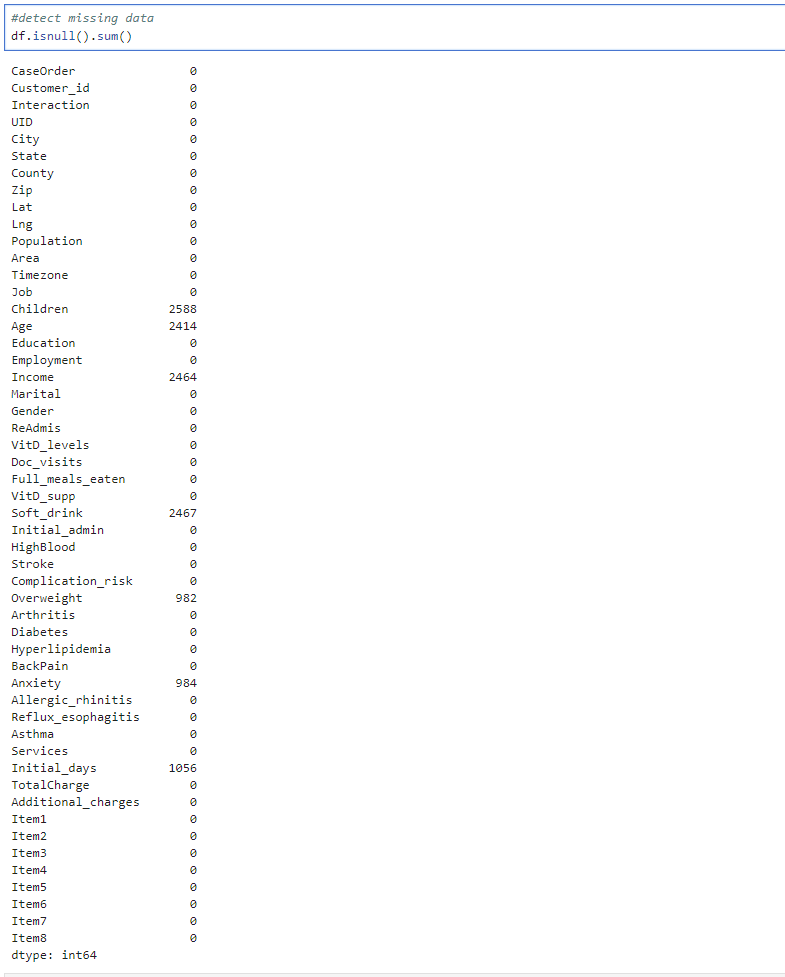
The duplicates were not found in the medical\_raw\_data dataset.

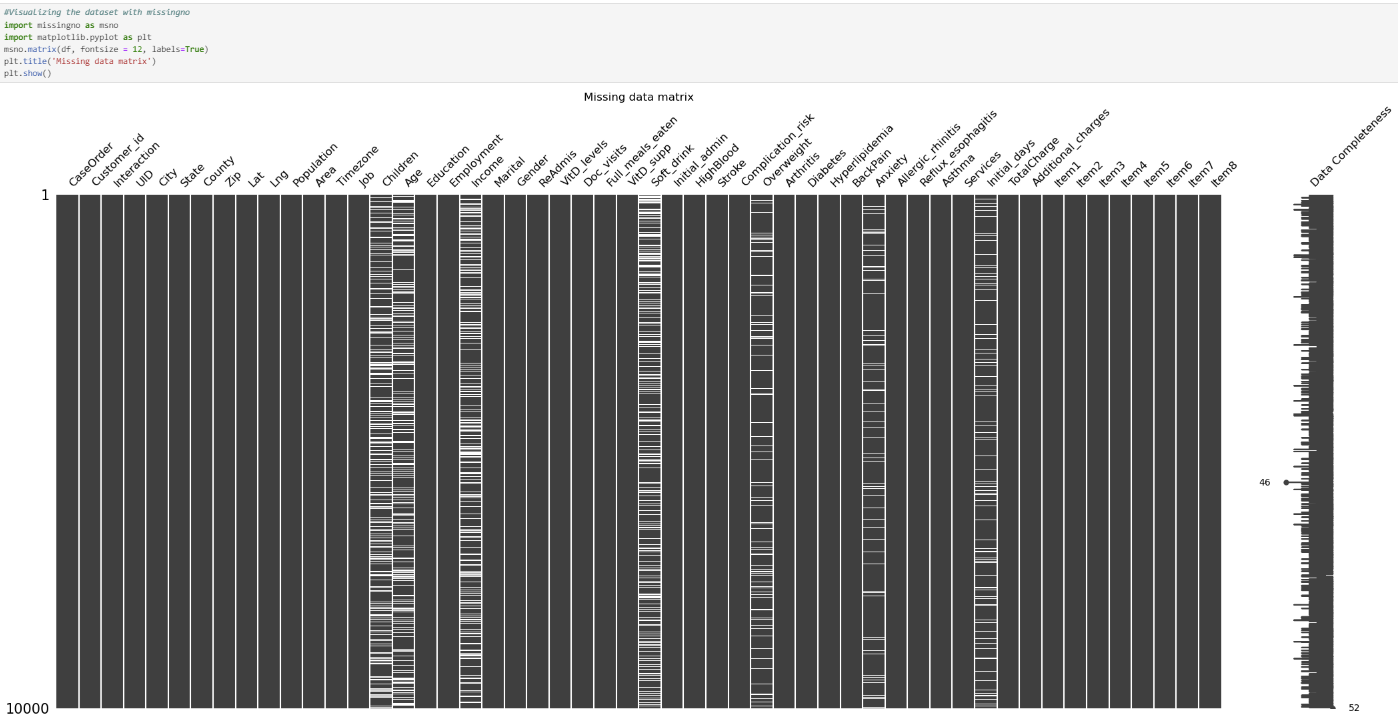


**Missing Values**

Missing values were found in the next variables

* Children (2588 values missing)
* Age (2414 values missing)
* Income (2464 values missing)
* Soft\_drink (2467 values missing)
* Overweight (982 values missing)
* Anxiety (984 values missing)
* Initial\_days (1056 values missing)

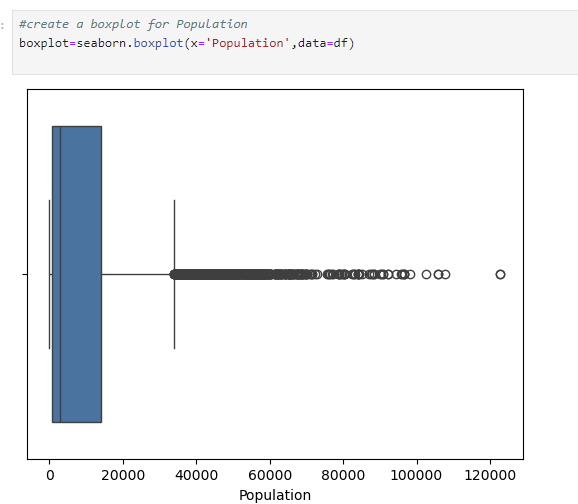




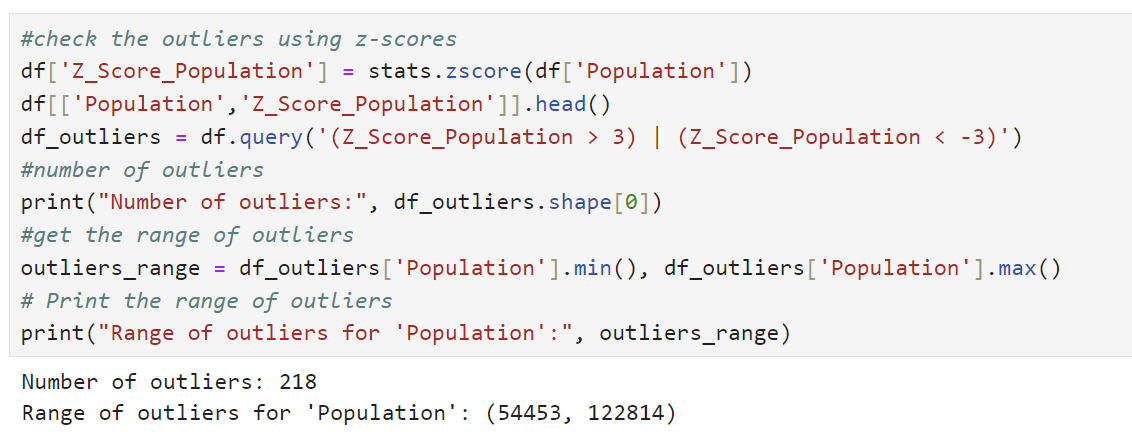
**Outliers**

All quantitative variables were checked for outliers. The Boxplot method was used for visualization and Z-scores method to identify outliers, get the number and range of outliers.

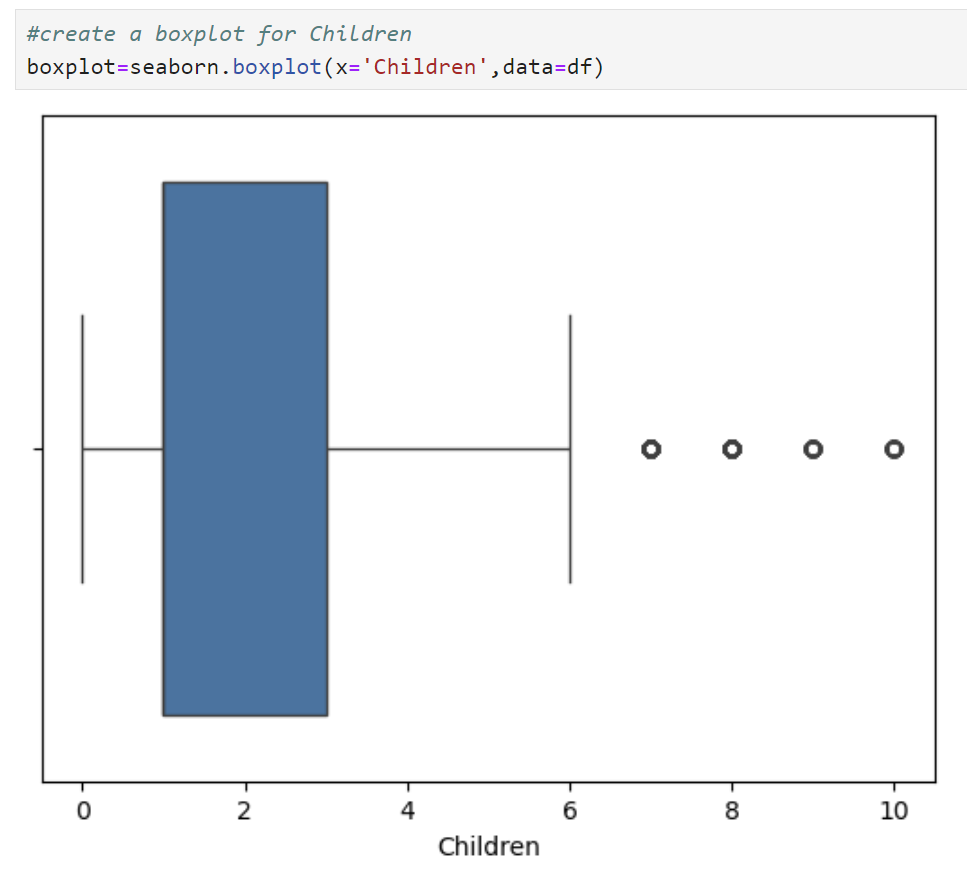
* **Population** variable contains 218 outliers laying in a range between 54453 and 122814 based on the Z-score method

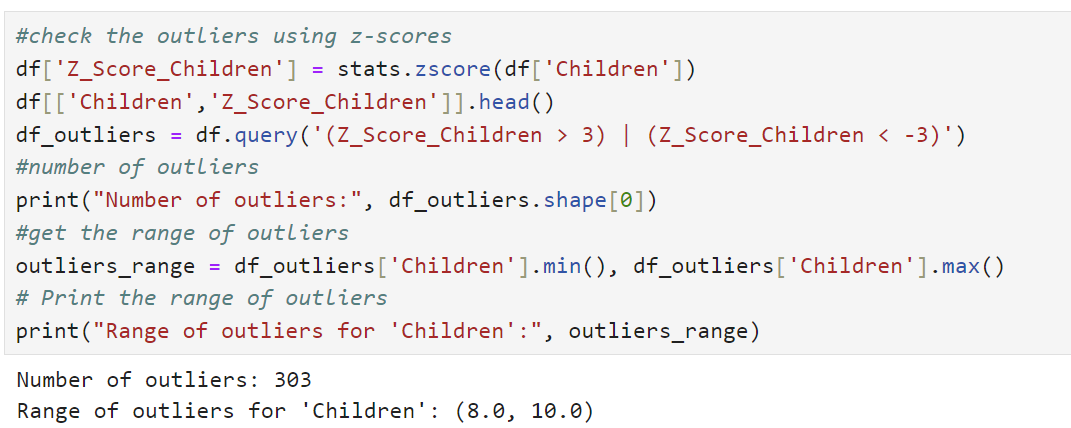


Outliers were identified using Z-scores (Z-score is greater than 3 or less than –3) and number and range of outliers were found.

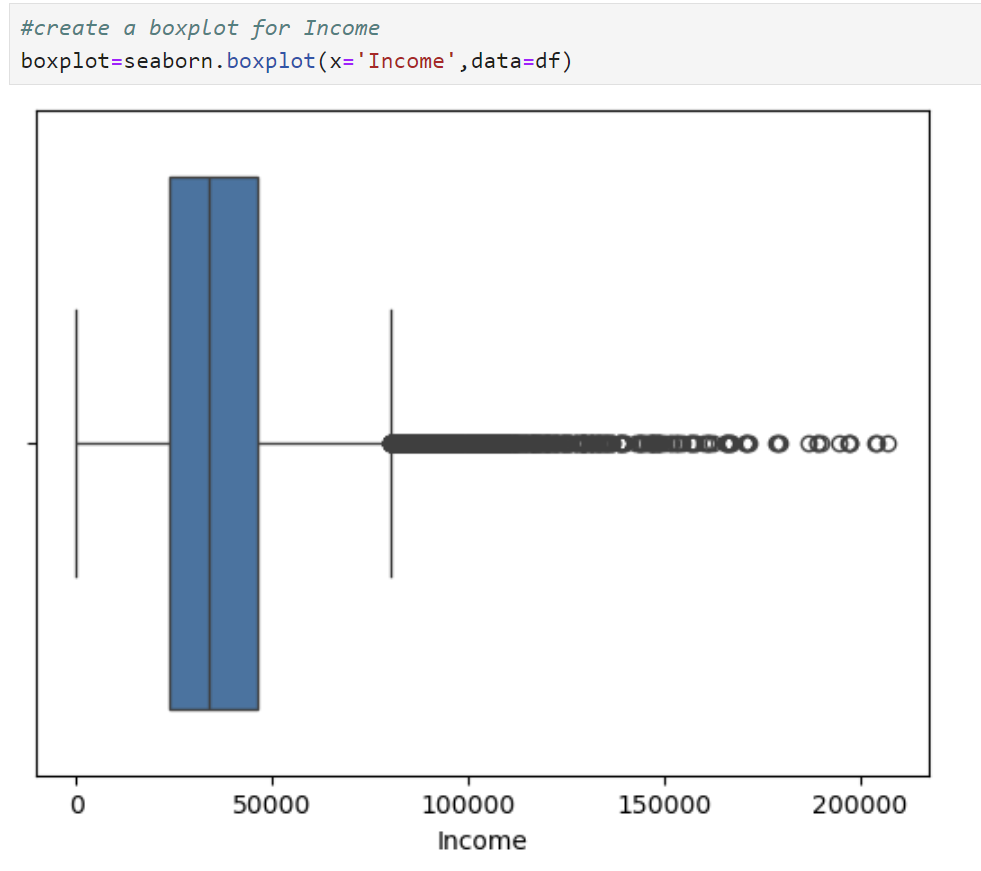


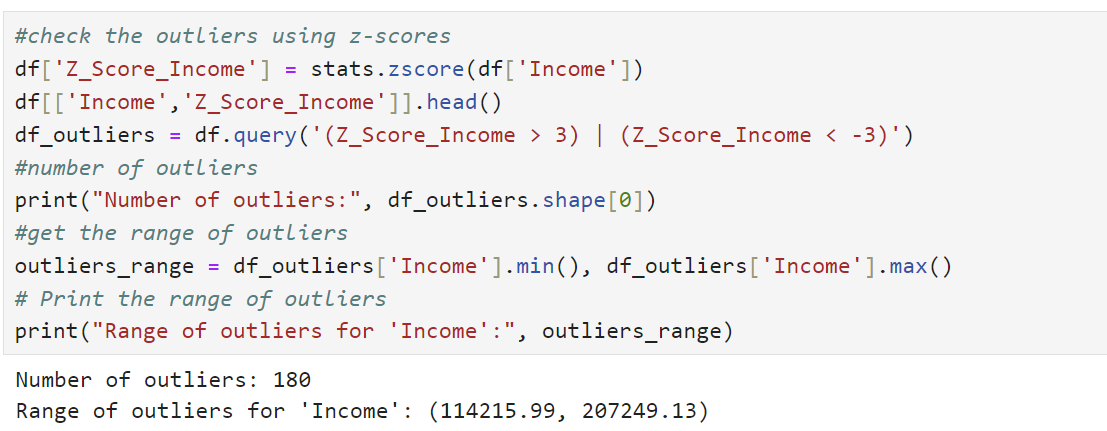
* **Children** variable contains 303 outliers in a range between 8.0 and 10.0



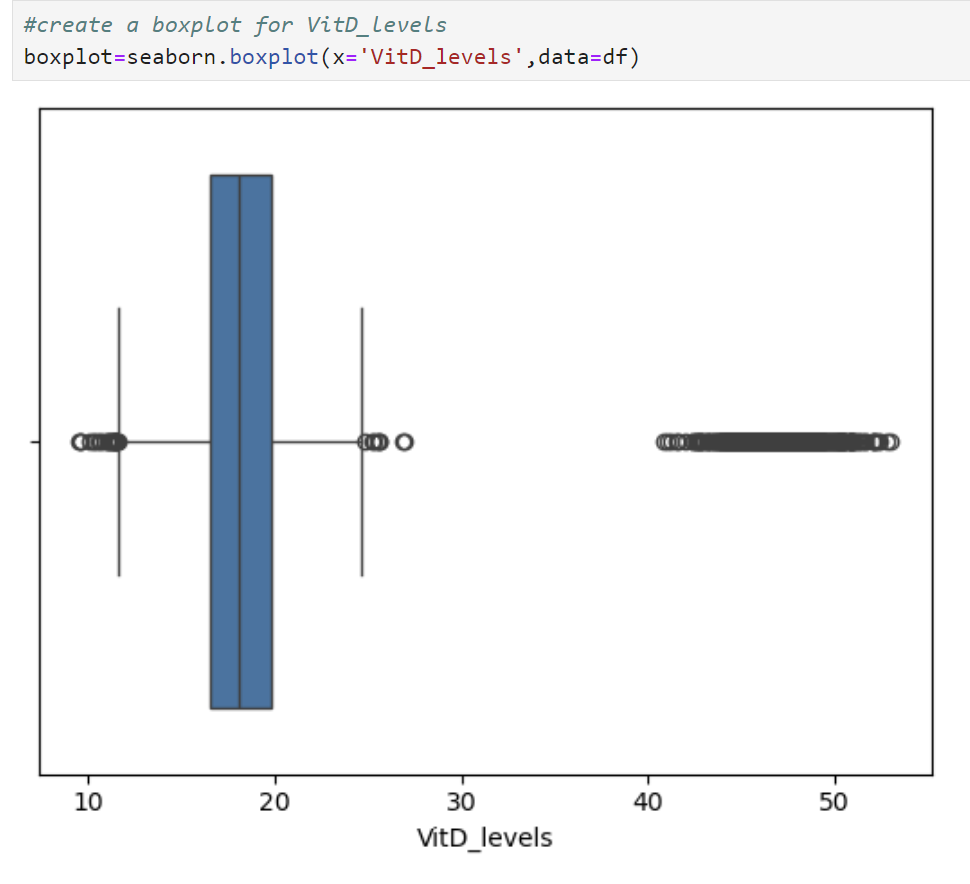


* **Income** variable contains 180 outliers in a range of 114215.99 and 207249.13



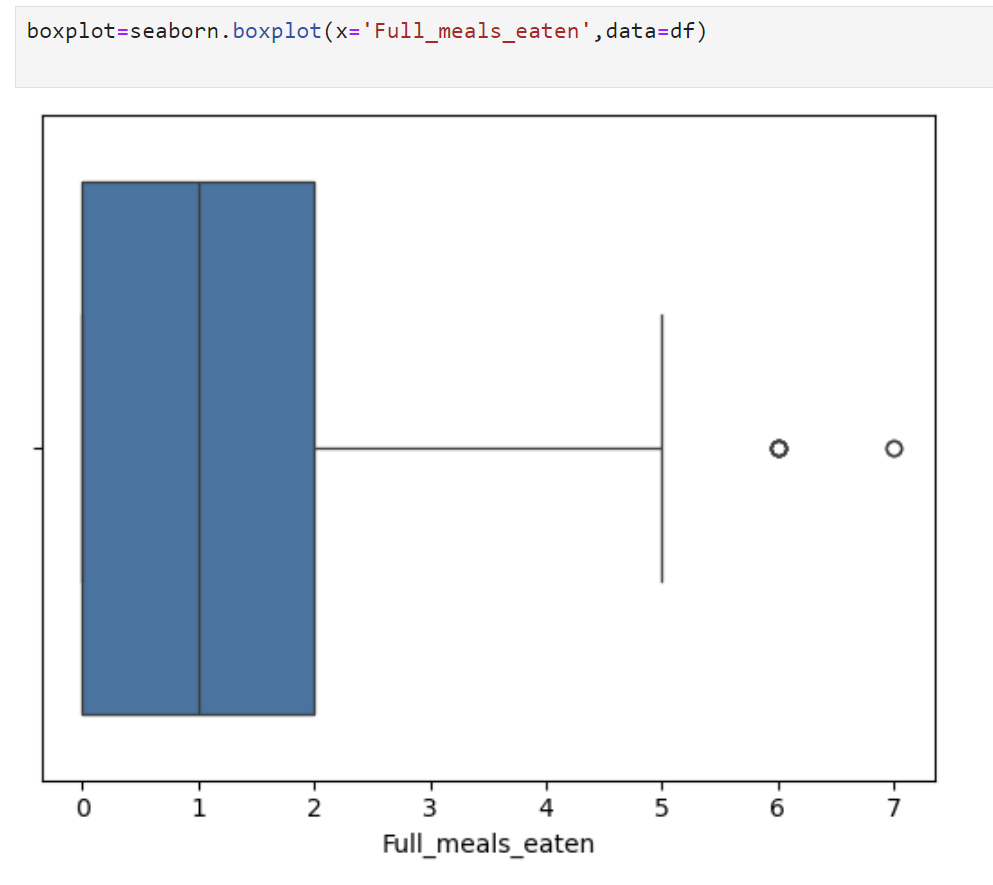


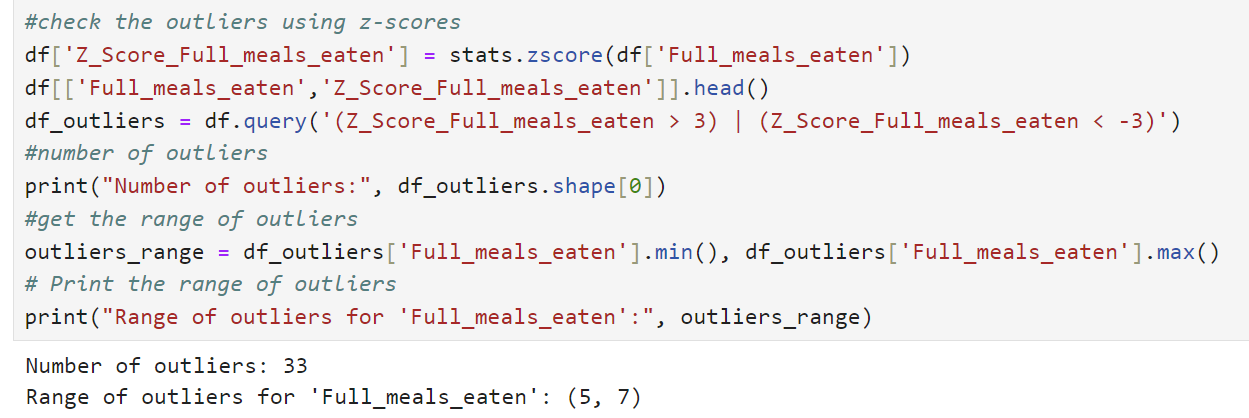
* **VitD\_levels** variable contains 500 outliers in a range of 40.8416712 and 53.01912416



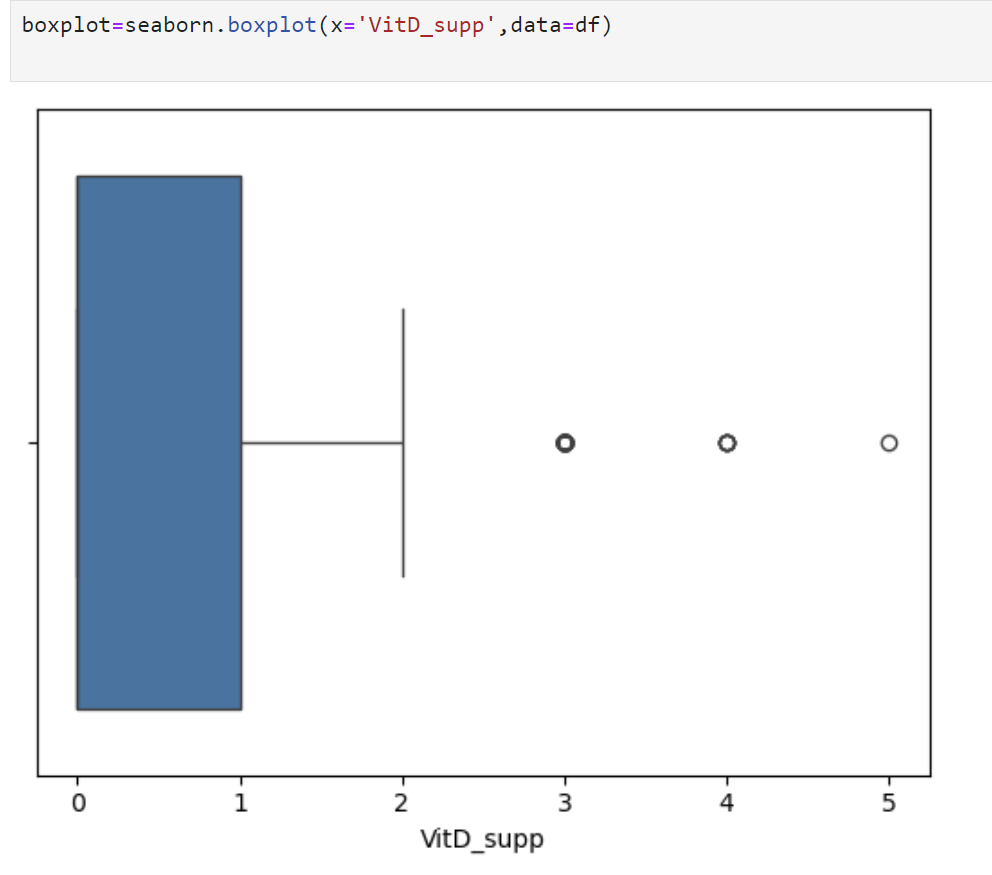


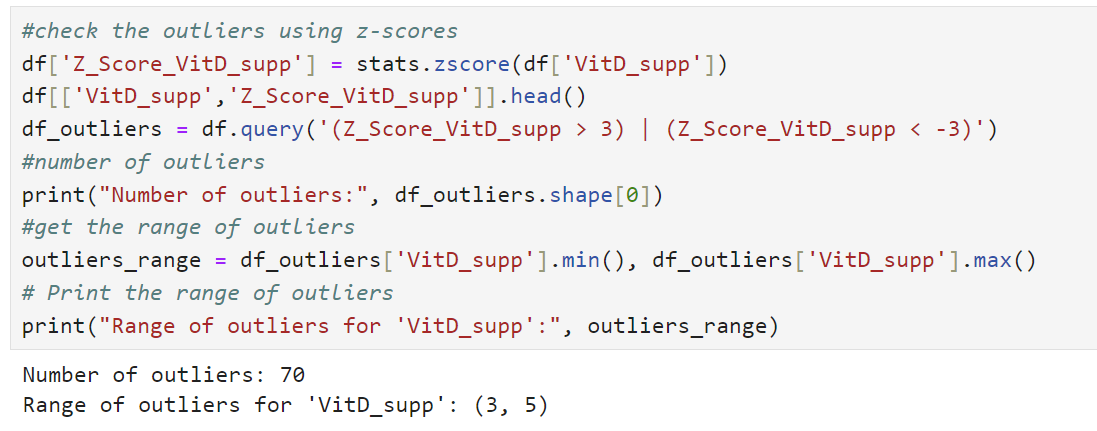
* **Full\_meals\_eaten** variable contains 33 outliers in a range of 5 and 7



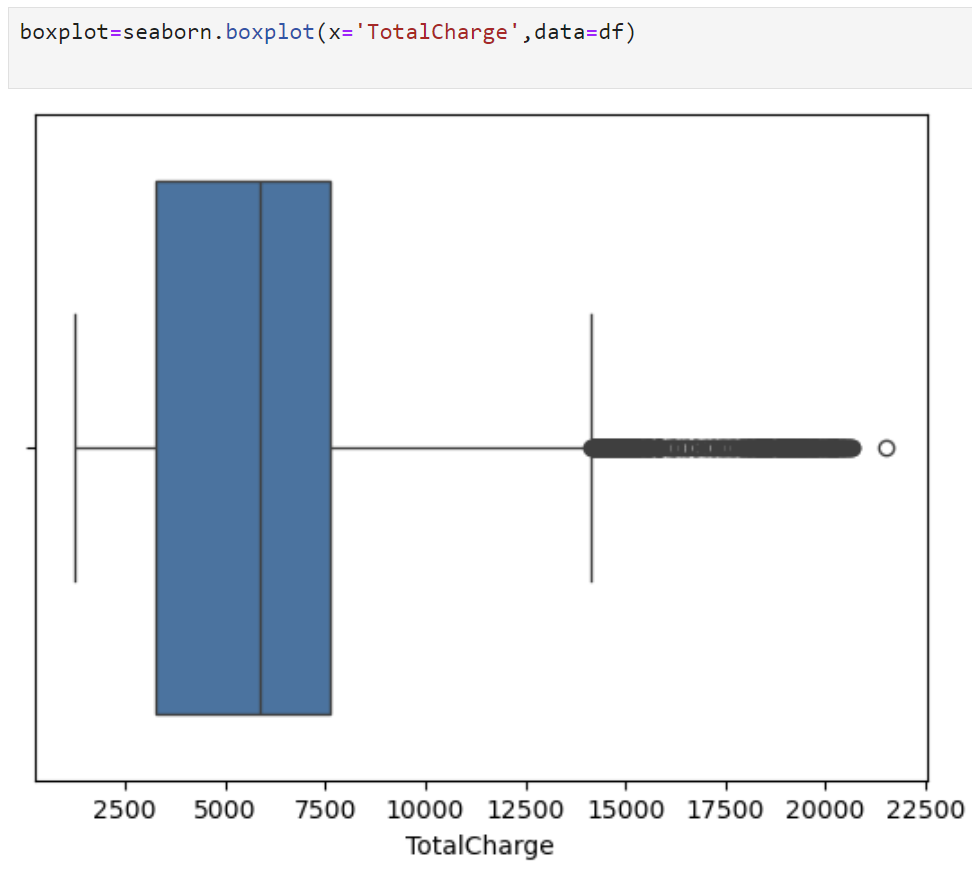


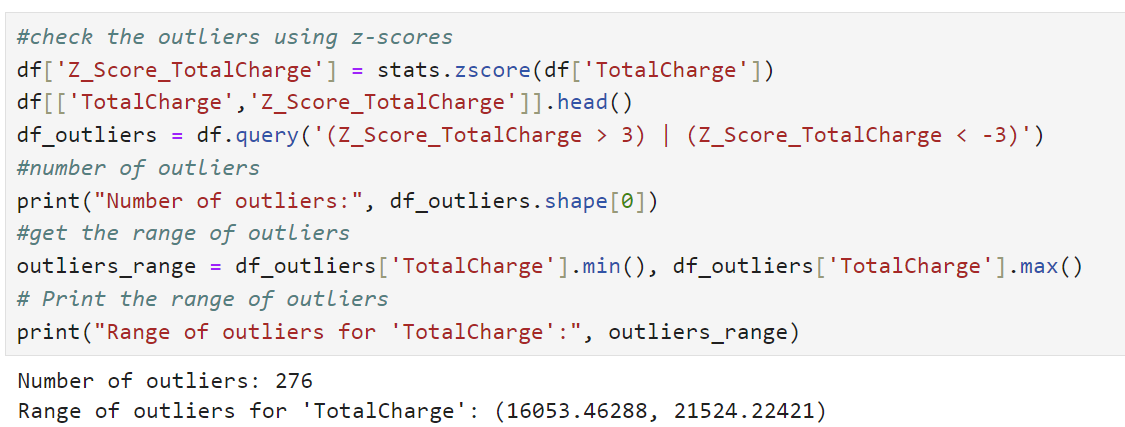
* **VitD\_supp** variable contains 70 outliers within a range of 3 and 5





* **TotalCharge** variable contains 276 outliers within a range of 16053.46288 and 21524.22421

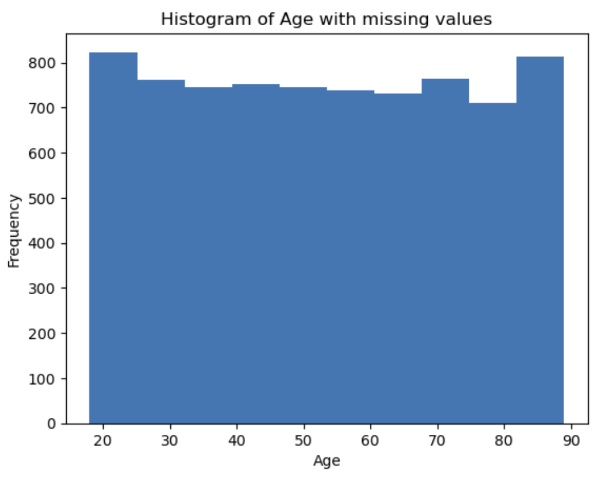
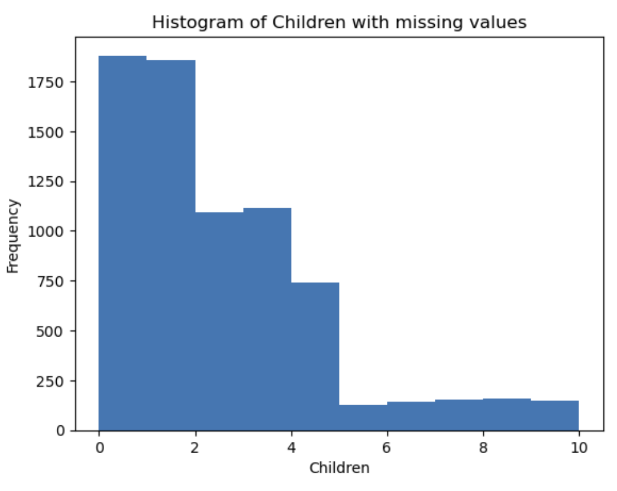


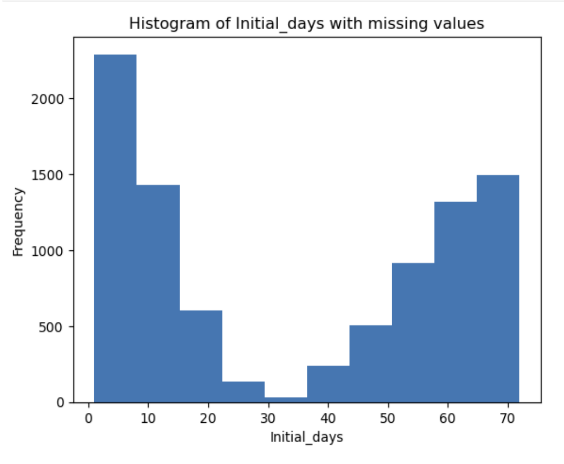
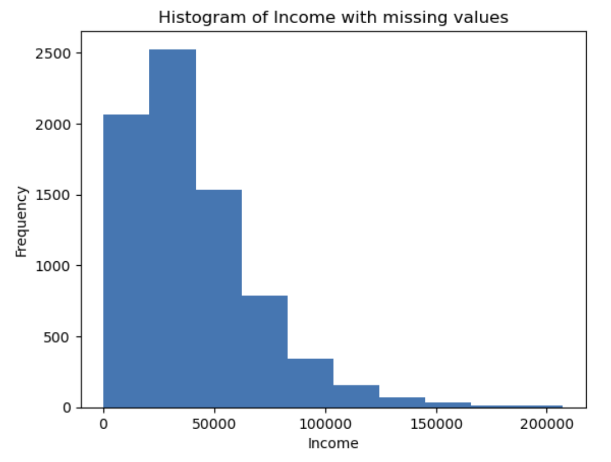


***C2. Methods used to fix findings referenced in C1.***

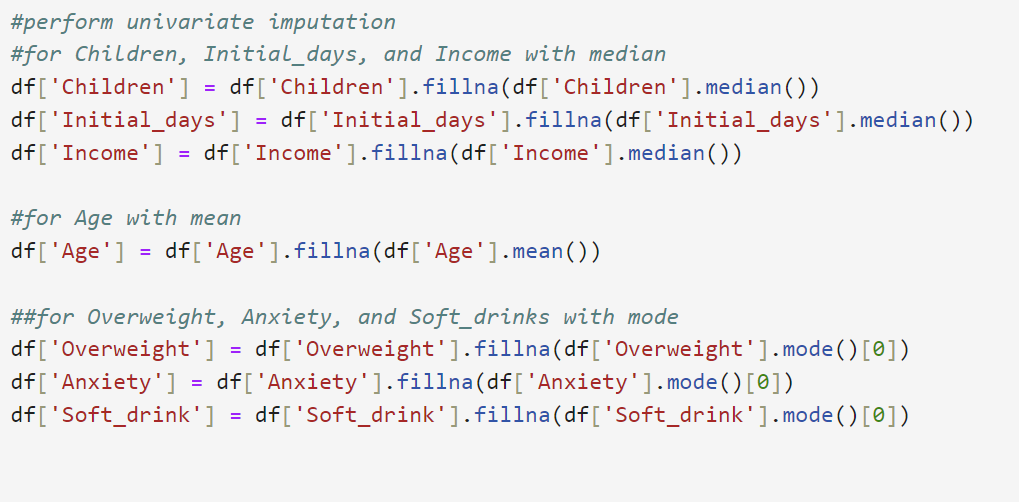
There were no ***duplicates*** found in the dataset.

To fix ***missing value*** the histograms for Children, Age, Income, and Initial\_days variables were created to see a distribution.





After analyzing the distribution for each variable above I decided to use univariate imputation method to treat missing values. The Age variable has a uniform distribution, so the missing values were replaced by mean values. The Children and Income have a skewed distribution and 'Initial\_days' are bimodal so the missing values in these variables were replaced by median values. ‘Overweight’, ‘Anxiety’ and ‘Soft\_drinks are categorical/Boolean variables, so the missing values were replaced by mode.



For the ***outlier's*** treatment there is no universal method to deal with them. It's important to analyze each variable, the number of outliers and the range to justify what method of treatment needs to be performed. Here are four methods for outliers' treatment:

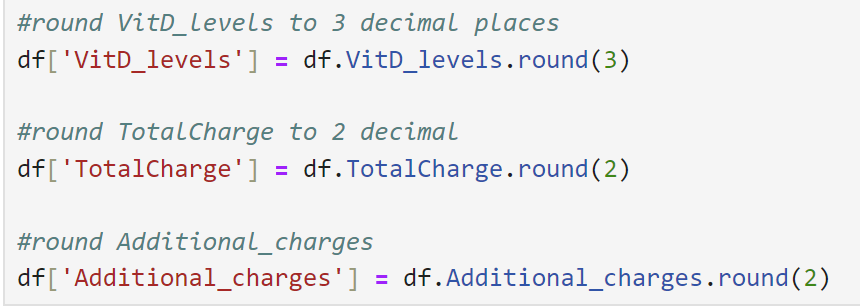
1. Imputation is performed when counts and range are not accepted or expected. In this case the outliers are replaced with the median. It preserves the sample size but can cause bias.
2. Retain all outliers within a dataset if values are accepted or expected. It preserves the sample size but can cause normality to decrease.
3. Exclude outliers to a different data frame when it's hard to justify if the imputation or outlier retention could be performed. It provides flexibility but reduces data size.
4. Remove outliers when the methods above cannot be performed.

After justification the outliers were retained in the dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | # of outliers | Range | Note | Treatment |
| Population | 218 | 54453, 122814 | The number of outliers is not high, and the values seem acceptable, because patients from urban settings will have higher populations. | Retain the outliers |
| Children | 303 | 8.0, 10.0 | The number of outliers is not high, and the values are acceptable, because it is not unrealistic to have more than 8 children. | Retain the outliers |
| Income | 180 | 114215.99, 207249.13 | The number of outliers is not high, and the values are acceptable, because the income of 114K and even 207K seem legit. | Retain the outliers |
| VitD\_levels | 500 | 40.8416712, 53.01912416 | Adults should have 20 nanograms per milliliter (ng/mL) of vitamin D in their blood or above; levels below that are usually too low for adequate bone health and general health; levels above 50 ng/mL may be too high (Fletcher, 2023). Also, considering that it's information from admitted patients, it's expectable to have too low or too high Vit D blood level due to illness. | Retain the outliers |
| Full\_meals\_eaten | 33 | 5, 7 | According to the raw medical data dictionary, patients can request more than 3 meals. The number of outliers is very low | Retain the outliers |
| VitD\_supp | 70 | 3, 5 | The number of outliers is low and the range of vit D supplementation seems acceptable | Retain the outliers |
| TotalCharge | 276 | 16053.46288, 21524.22421 | The number of outliers is not high, and the range doesn't seem extreme | Retain the outliers |

***Other Issues***

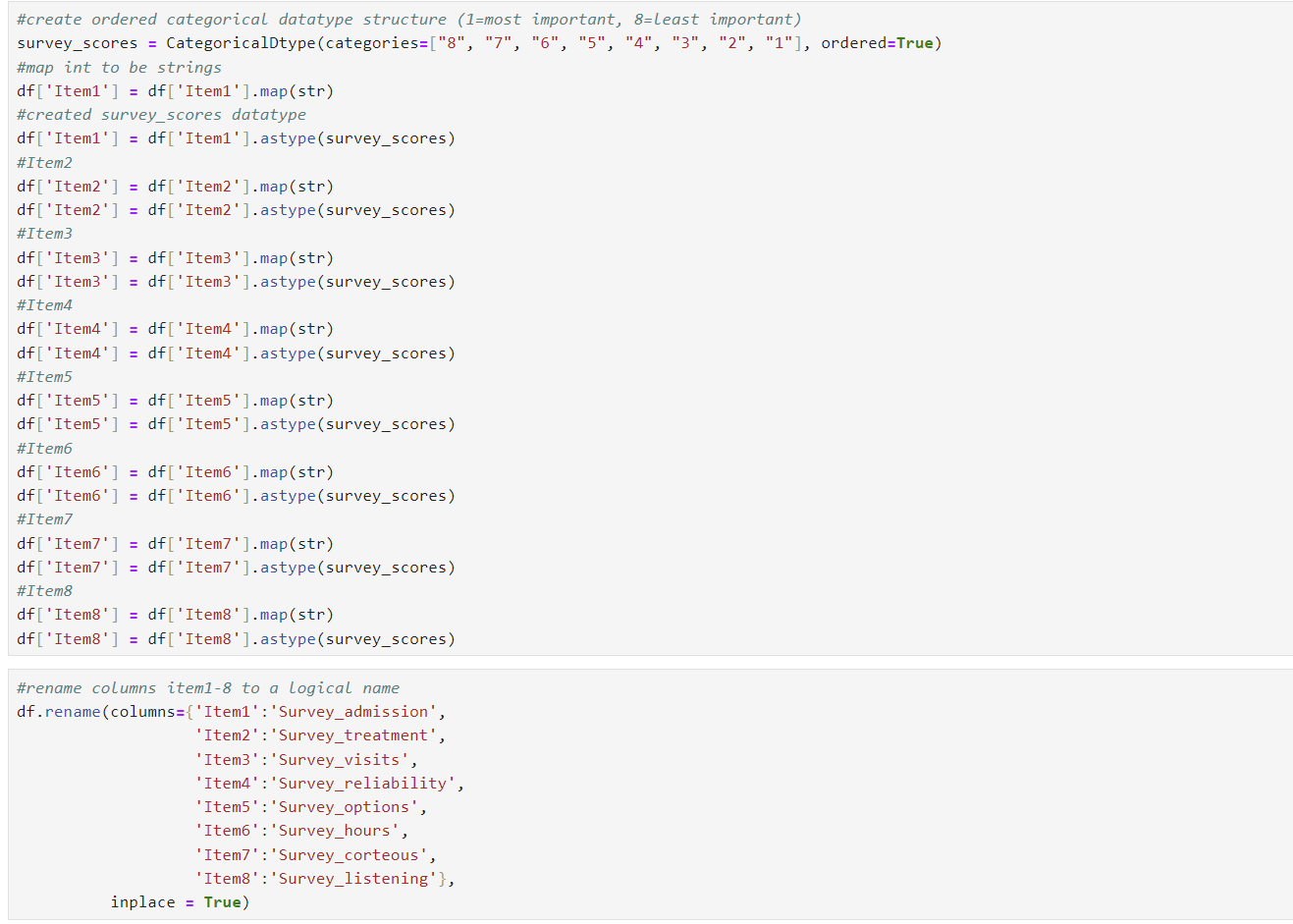
The variables' VitD\_levels', 'TotalCharge' and 'Additional\_charges' are stored to six decimal places, so they will be rounded down to three decimal places for 'VidD\_levels' and to two decimal places for the 'TotalCharge' and 'Additional\_charges'.



It seems that the variable 'Unnamed: 0'is duplicate with 'CaseOrder', so 'Unnamed: 0' was dropped from the dataset.



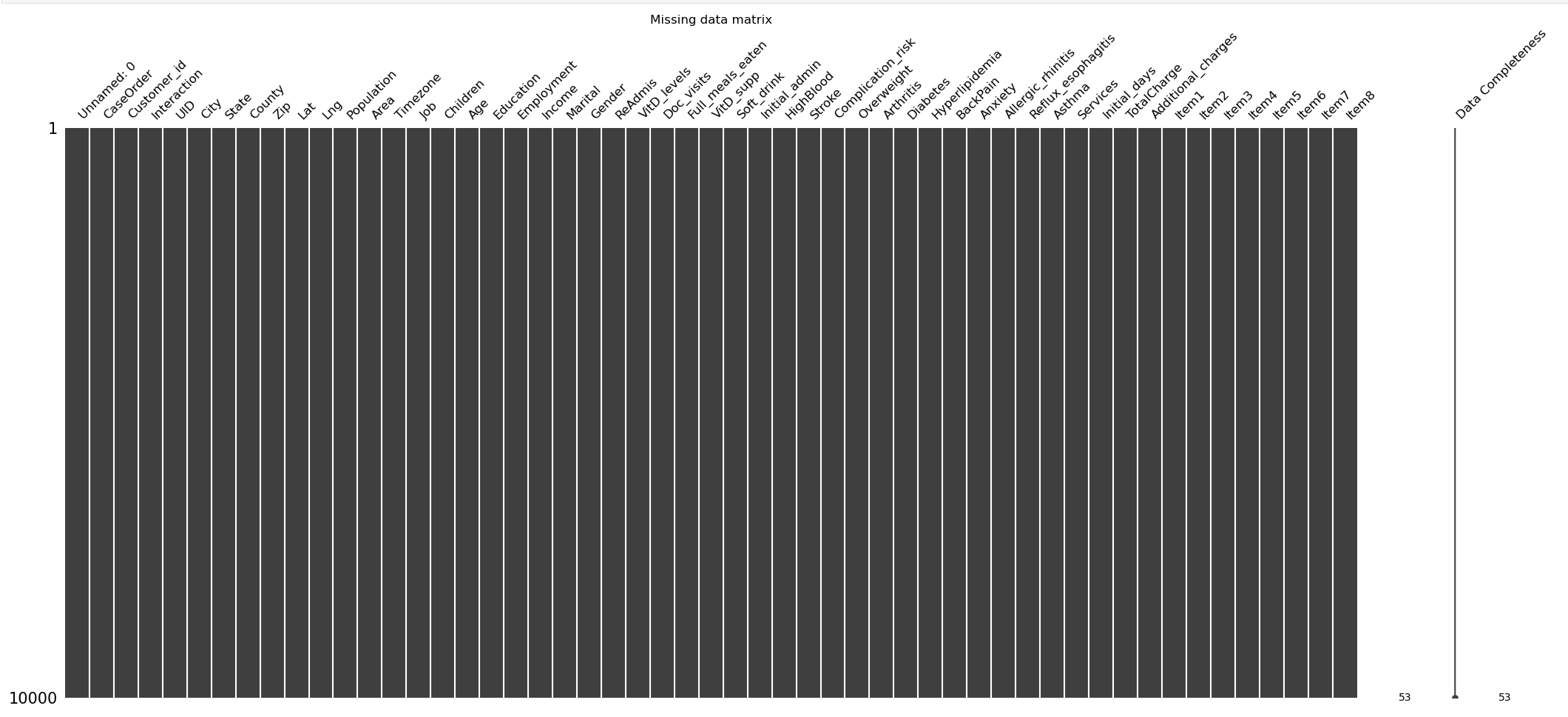
The columns ‘Item1 - Item8’ represents a customer's survey to rate the importance of various factors/surfaces on a scale of 1 to 8 with 1= most important and 8= least important. The values in these columns were converted from categorical to numerical and renamed to more logical names.



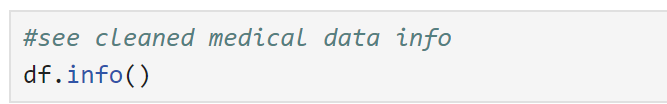
**C3.**

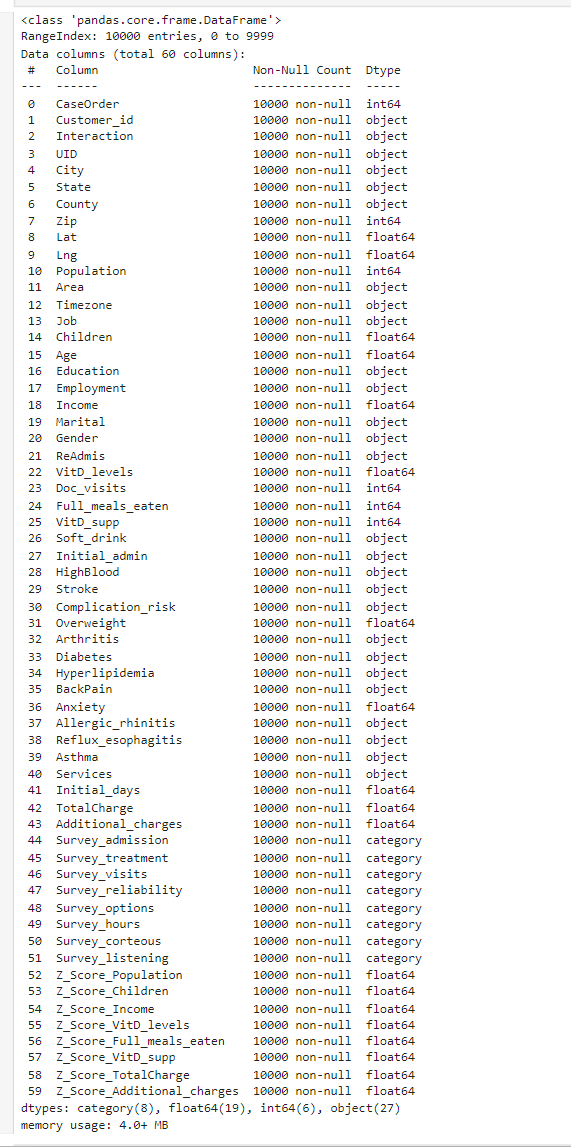
A summary of the work performed:

* Anaconda was downloaded and installed. The coding part was performed in Jupyter Notebook.
* Required packages were imported.
* Medical\_raw\_data.csv file was loaded onto the Jupyter Notebook.
* Dataset was inspected with df.info()
* ‘CaseOrder’ and ‘Customer\_id’ were verified as unique.
* The variable ‘Unnamed: 0’ was removed from dataset, because it duplicated 'CaseOrder' column.
* The variables were checked for the duplicates and no duplicates were detected.
* The variables were checked for missing values. The variables' Children', 'Age', 'Income', 'Soft\_drink', 'Overweight'. 'Anxiety', and 'Initial\_days' had missing values. After visualizing data with missingno and creating histogram for each variable to see type of distribution, the univariable imputation was used for treatment. After this, a histogram for each treated variable and missing data matrix were created for verification.



* The outliers were retained after justifying the number of outliers and their range in each variable.
* ‘VitD\_levels’ was rounded down to three decimal places.
* ‘TotalCharge’ was rounded down to two decimal places.
* ‘Additional\_charge’ was rounded down to two decimal places.
* ‘Item1 – Item8’ were renamed to more logical names and were stored as an ordered categorical instead of an integer.





**C4. Disadvantages of methods used in C3.**

Univariable Imputation for missing values can cause a loss of data variability. Imputing missing values with a single value (e.g., mean or median) can reduce the variability in the dataset, potentially leading to biased results. If the missing data is linked to unobserved factors, the imputation could be flawed. Imputing missing values without considering the relationships between variables can distort and alter the actual relationships within the dataset.

Using boxplots to check for outliers involves a degree of subjectivity, as the definition of an outlier can vary based on the context and the chosen threshold, such as 1.5 times the interquartile range. While boxplots offer a visual summary, they do not provide detailed information about the nature or cause of the outliers.

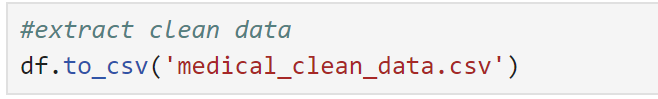
Treating outliers with z-scores assumes that the data follows a normal distribution, which may not always be the case, making z-scores an inappropriate measure for non-normally distributed data, and removing outliers based on z-scores can result in the loss of essential data points that could be significant for the analysis. The choice of a threshold, such as a z-score greater than 3, is arbitrary and might not be suitable for all datasets.

Retaining outliers can skew the results of the analysis, mainly if the outliers are extreme, and can sometimes result from data entry errors or other anomalies. Without proper investigation, retaining these outliers can lead to incorrect conclusions.

**C5. What challenges a data analyst may encounter using now cleaned data for analysis.**

Data analysts may face several challenges when analyzing a recently cleaned churn dataset. These challenges include obtaining inaccurate and inconsistent results due to human errors during the cleaning process. Since data cleaning lacks standardization, it can be subjective, leading to inconsistencies. Additionally, removing outliers or using imputation methods might result in data loss or incomplete data, thereby limiting the dataset’s scope.

**D1. The clean data was extracted.**

Please see attached **medical\_clean\_data.csv** file

D2. Please **D206code.py** file with code attached.

***Part IV: PCA***

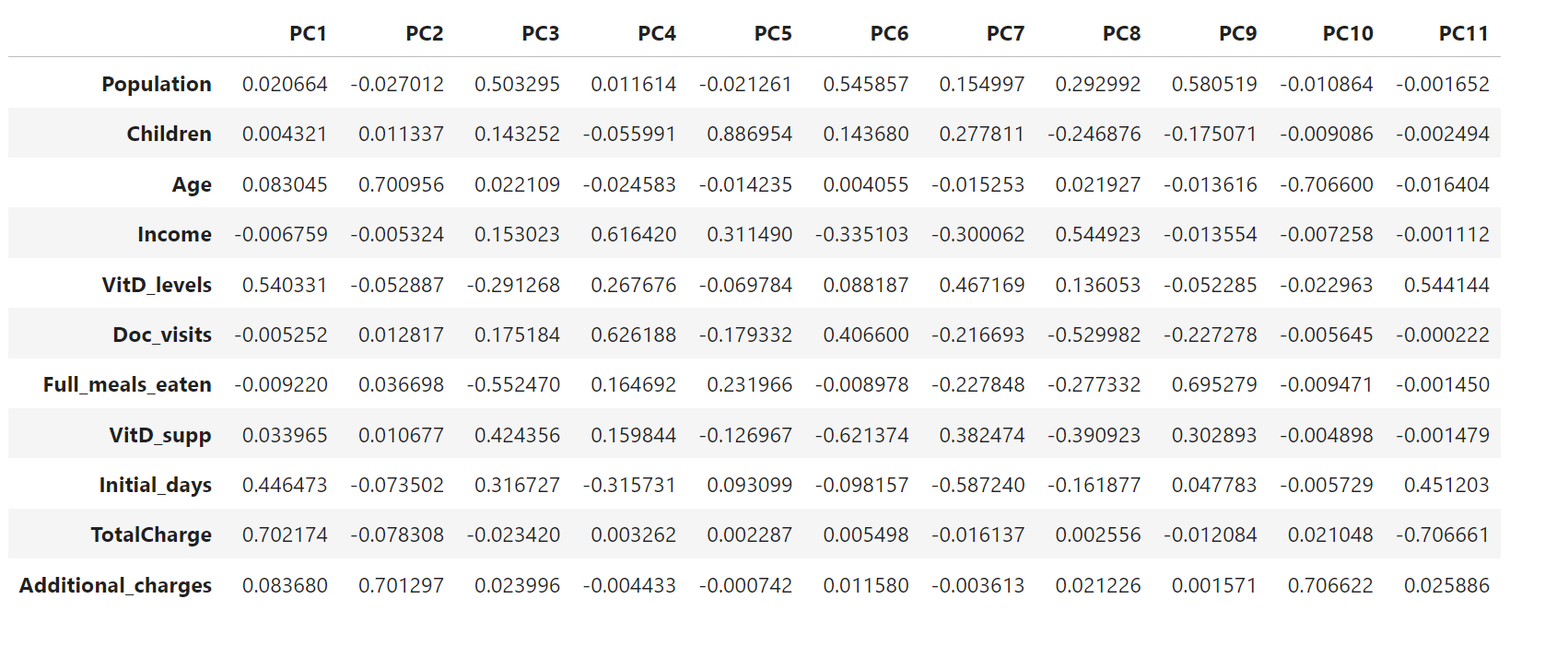
**E1. Perform PCA**

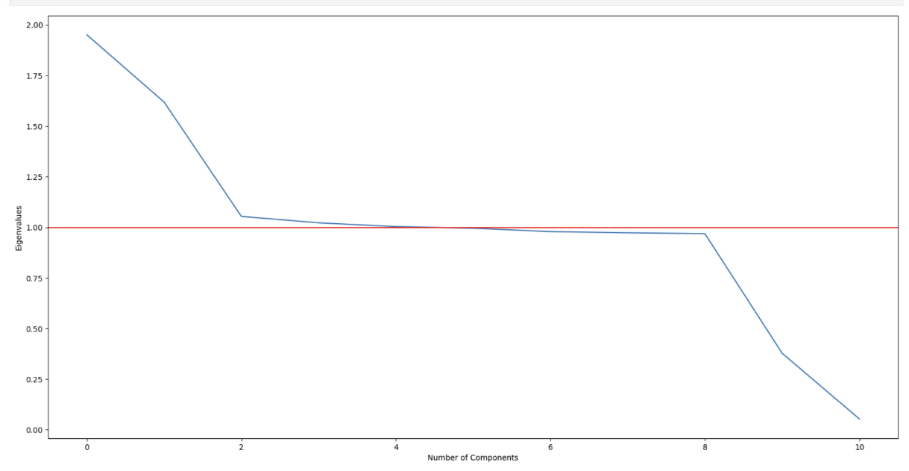
Principal component analysis (PCA ) is a method of unsupervised learning. The foundation of PCA is an orthogonal (involving angles) mathematical idea that creates a set of principal components from a collection of observations of potentially correlated variables. It entails the procedure of determining which linear combinations of variables provide the most explanation for the covariation, or correlation, between variables. Principal components, or the new set of variables (PCs) from the old set or initial set of variables, represent the combination of variables based on correlation, connection, shared properties, or likeness. Occasionally, newly developed PCA features might not be immediately clear (*Redirecting*, n.d.-b, p. 3).

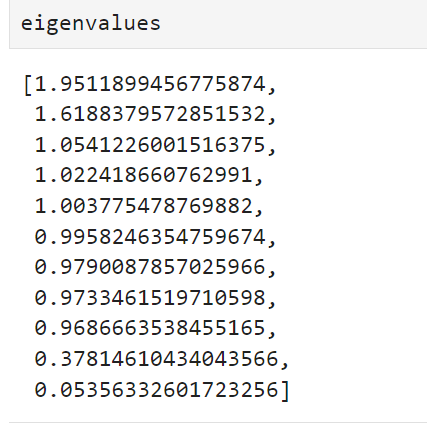
PCA is most meaningful when using only continuous variables. To perform principal component analysis the variables 'Population', 'Children', 'Age', 'Income', 'VitD\_levels', 'Doc\_visits', 'Full\_meals\_eaten', 'VitD\_supp', 'Initial\_days', 'TotalCharge', 'Additional\_charges' were used.

Next, the data was normalized by using the standardization (Mean Normalization). After this, number of PC components set and component loading performed.



Lastly, a scree plot was created.





**E2. Explain which PC's should be retained**

The Kaiser rule is one of the methods to analyze the PCA. It states that all the components with eigenvalue greater than or equal to 1 should be retained. According to the results, PC1, PC2, PC3 are most correlated.

|  |  |
| --- | --- |
| PC (eigenvalue) | Variables |
| **PC1** (1.9511899456775874) | High loadings for **TotalCharge (0.702)** and **VitD\_levels (0.540)**. This suggests that these variables are the most influential in this component, indicating a possible relationship between total charges and vit D levels |
| **PC2** (1.6188379572851532) | High loadings for **Age (0.701)** and **Additional\_charges (0.701)** suggest that age and additional charges are significant in this component. |
| **PC3** (1.0541226001516375) | High loadings for **Population (0.503)** and **Full\_meals\_eaten (-0.552)** indicate these variables are important here, with an inverse relationship between population and full meals eaten. |

**E3. Benefits for organizations from the PCA results.**

Organizations can benefit from PCA in several ways:

* PCA reduces the dimensionality of data, making it easier to analyze and visualize. For instance, instead of dealing with 11 original variables, an organization can focus on a few principal components that capture most of the variance.
* PCA helps identify the most important variables that drive the variation in the data. This can guide decision-making and strategic planning (for example, 'TotalCharge' and 'VitD\_levels' have high loadings on PC1, indicating they are key drivers. An organization can prioritize these factors in their strategies, such as focusing on managing total charges and monitoring vitamin D levels).
* PCA allows for better visualization of high-dimensional data, making it easier to identify patterns and trends.
* By focusing on the most important variables, organizations can allocate resources more effectively (for example, if 'Age' and 'Additional\_charges' are significant in PC2, an organization can allocate resources to better manage these aspects, such as tailoring services for different age groups or optimizing additional charges)/
* PCA can reveal hidden relationships between variables, providing strategic insights that might not be apparent from the raw data. The inverse relationship between 'Population' and 'Full\_meals\_eaten' in PC3 might indicate areas where population density affects meal consumption patterns, guiding targeted interventions.

**F. Panopto video:** Please see attached link to a video

**G. Third-Party Code References:**

Bonthu, H. (2024, July 9). *Detecting and Treating Outliers | Treating the odd one out!* Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>

DataCamp. (n.d.). *D206 Data cleaning* [Video]. <https://app.datacamp.com/>. Retrieved June 19, 2024, from <https://app.datacamp.com/learn/custom-tracks/custom-d206-data-cleaning>

David. (2023, February 8). *What is Principal Component Analysis (PCA) & How to Use It? | Bigabid*. Bigabid. <https://www.bigabid.com/what-is-pca-and-how-can-i-use-it/>

Nkmk. (2024, January 26). *pandas: Find, count, drop duplicates (duplicated, drop\_duplicates)*. note.nkmk.me. <https://note.nkmk.me/en/python-pandas-duplicated-drop-duplicates/>

**H. References:**

BasuMallick, C. (2022, December 6). *The 12 vital differences between R and Python*. Spiceworks Inc. <https://www.spiceworks.com/tech/devops/articles/r-vs-python/>

*Redirecting*. (n.d.-c). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20Course%20Guide%2DV3%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources>

*Redirecting*. (n.d.-d). <https://westerngovernorsuniversity.sharepoint.com/:w:/r/sites/DataScienceTeam/_layouts/15/Doc.aspx?sourcedoc=%7B52027E6C-E2EA-47EE-BB47-BEC1F6430AC7%7D&file=D206PAStepbyStepGuide%20(NUM4).docx&action=default&mobileredirect=true>

*Redirecting*. (n.d.-e). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F7%2E%20D206%2DGettingStartedPCA%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>

*Redirecting*. (n.d.-f). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F6%2E%20D206%2DGettingStartedRexpressionofCategoricalVariables%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>

*Redirecting*. (n.d.-g). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F5%2E%20D206%2DGettingStartedOutliers%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>

*Redirecting*. (n.d.-h). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F4%2E%20D206%2DGettingStartedMissingValues%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>

*Redirecting*. (n.d.-i). <https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29%2F3%2E%20D206%2DGettingStartedWithDuplicates%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD206%2FStudent%20Facing%20Resources%2FD206%20%2D%20Getting%20Started%20with%20D206%20Video%20Series%20%28Slides%20and%20Videos%29>