Performance Assessment: Data Cleaning

Medical Data

Joshua Funderburk
Western Governor’s University
D206: Data Cleaning
04/15/2021

Part I: Research Question

A. What do patients who have been readmitted to the hospital have in common and what variables are most indicative of the likelihood of readmission?
   1. Given this information, a hospital organization would be able to put plans and procedures in place to combat problem variables and lower the risk for patient readmission. Essentially, hospitals would be able to be more proactive in combatting risk for readmission during the patient's initial stay at the hospital.

B. Description of variables in the data set. (Also see Jupyter Notebook or Part I – B Excel document).

Column Name	Data Group	Definition	Example	Data Type
CaseOrder	Identifiers	A place holder variable to preserve the original order of the raw data file	1	Integer
Customer_id	Identifiers	Unique patient ID	C412403	Character
Interaction	Identifiers	Unique IDs related to patients' transactions, procedures, and admissions	8cd49b13-f45a-4b47-a2bd-173ffa932c2f	Character
UID	Identifiers	Unique IDs related to patients' transactions, procedures, and admissions	3a83ddb66e2ae73798bdf1d705dc0932	Character
City	Demographic	Patient city of residence as listed on the billing statement	Eva	Character
State	Demographic	Patient state of residence as listed on the billing statement	AL	Character
County	Demographic	Patient county of residence as listed on the billing statement	Morgan	Character
Zip	Demographic	Patient zip code of residence as listed on the billing statement	35621	Integer
Lat	Demographic	GPS coordinates of patient residence as listed on the billing statement	34.3496	Decimal
Lng	Demographic	GPS coordinates of patient residence as listed on the billing statement	-86.72508	Decimal
Population	Demographic	Population within a mile radius of patient, based on census data	2951	Integer
Area	Demographic	Area type (rural, urban, suburban), based on unofficial census data	Suburban	Character

Timezone	Demographic	Time zone of patient residence based on patient's sign-up information	America/Chicago	Character
Job	Demographic	Job of patient (or primary insurance holder) as reported in the admissions information	Psychologist, sport and exercise	Character
Children	Demographic	Number of children in the patient's household as reported in the admissions information	1	Integer
Age	Demographic	Age of patient as reported in admissions information	53	Integer
Education	Demographic	Highest earned degree of patient as reported in admissions information	Some College, Less than 1 Year	Character
Employment	Demographic	Employment status of patient as reported in admissions information	Full Time	Character
Income	Demographic	Annual income of the patient (or primary insurance holder) as reported at time of admission	86575.93	Decimal
Marital	Demographic	Marital status of the patient (or primary insurance holder) as reported on admission information	Divorced	Character
Gender	Demographic	Customer self-identification as male, female, or nonbinary	Male	Character
ReAdmis	Admission/Treatment	Whether the patient was readmitted within a month of release or not (yes, no)	No	Character
VitD_levels	Admission/Treatment	The patient's vitamin D level as measured in ng/mL	17.80233049	Decimal
Doc_visits	Admission/Treatment	Number of times the primary physician visited the patient during the initial hospitalization	6	Integer
Full_meals_eaten	Admission/Treatment	Number of full meals the patient ate while hospitalized (partial meals count as 0, and some patients had more than three meals in a day if requested)	0	Integer
VitD_supp	Admission/Treatment	The number of times that vitamin D supplements were administered to the patient	0	Integer

Soft_drink	Admission/Treatment	Whether the patient habitually drinks three or more sodas in a day (yes, no)	NA	Character
Initial_admin	Admission/Treatment	The means by which the patient was admitted into the hospital initially (emergency admission, elective admissions, observation)	Emergency Admission	Character
HighBlood	Admission/Treatment	Whether the patient has high blood pressure (yes, no)	Yes	Character
Stroke	Admission/Treatment	Whether the patient has had a stroke (yes, no)	No	Character
Complication_risk	Admission/Treatment	Level of complication risk for the patient as assessed by a primary patient assessment (high, medium, low)	Medium	Character
Overweight	Admission/Treatment	Whether the patient is considered overweight based on age, gender, and height (yes, no)	0	Integer
Arthritis	Admission/Treatment	Whether the patient has arthritis (yes, no)	Yes	Character
Diabetes	Admission/Treatment	Whether the patient has diabetes (yes, no)	Yes	Character
Hyperlipidemia	Admission/Treatment	Whether the patient has hyperlipidemia (yes, no)	No	Character
BackPain	Admission/Treatment	Whether the patient has chronic back pain (yes, no)	Yes	Character
Anxiety	Admission/Treatment	Whether the patient has an anxiety disorder (yes, no)	1	Integer
Allergic_rhinitis	Admission/Treatment	Whether the patient has allergic rhinitis (yes, no)	Yes	Character
Reflux_esophagitis	Admission/Treatment	Whether the patient has reflux esophagitis (yes, no)	No	Character
Asthma	Admission/Treatment	Whether the patient has asthma (yes, no)	Yes	Character
Services	Admission/Treatment	Primary service the patient received while hospitalized (blood work, intravenous, CT scan, MRI)	Blood Work	Character
Initial_days	Admission/Treatment	The number of days the patient stayed in the hospital during the initial visit	10.58576971	Decimal

TotalCharge	Admission/Treatment	The amount charged to the patient daily. This value reflects an average per patient on the total charge divided by the number of days hospitalized. This amount reflects the typical charges billed to patients, not including specialized treatments.	3191.048774	Decimal
Additional_charges	Admission/Treatment	The average amount charged to the patient for miscellaneous procedures, treatments, medicines, anesthesiology, etc.	17939.40342	Decimal
Item1	Survey	Timely admission	3	Integer
Item2	Survey	Timely treatment	3	Integer
Item3	Survey	Timely visit	2	Integer
Item4	Survey	Reliability	2	Integer
Item5	Survey	Options	4	Integer
Item6	Survey	Hours of treatment	3	Integer
Item7	Survey	Courteous staff	3	Integer
Item8	Survey	Evidence of active listening from doctor	4	Integer

Part II: Data-Cleaning Plan

C. Plan for cleaning the data:

1. Plan to find anomalies:
    i. Add/reset the index field – The index field will be set to Case Order.
        - The raw data actually contains two fields meant to serve as the index field as both fields represent the original sort order of the data. The first of the columns, however, it does not have a column header; so I have chosen to set the index to a field that has a header and has an apparent relationship to the rest of the data.

    ii. Update confusing or misleading column names – Column names that are not clearly representative of the data will be updated. Specifically, the eight survey columns currently labeled as "Item" followed by a number will be updated to reflect the survey question.

    iii. Update nan and null values – Update nan values to be null or 0, depending on the column and update null values to a value where appropriate.

    iv. Reexpress categorical data as numeric data – Fields with values of simply "Yes" or "No" will be updated to a binary system, where 1 equals "Yes" and 0 equals "No." Other categorical values will be updated to numeric where a numerical value is feasible. States and Timezones are within reason to set to an equivalent numerical value; however, fields such as county or city has too many unique many values to draw a numerical relationship that makes sense.

    v. Identify outliers – Visualizations, such as box plots, will be used to identify outliers. Additionally, certain standardized numerical columns will be tested for z-scores greater than 3 or less than -3 - which "Data Science Using Python and R" by Chantal Larose and Daniel Larose states is the rough rule of thumb for identifying outliers.

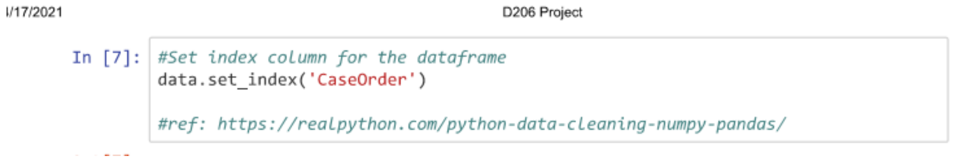
    vi. Standardize numeric fields – Use Python's scipy package to standardize appropriate numeric fields. Algorithms may perform better standardized as a z-score as opposed to using the raw numerical value. Not only numeric fields will be standardized as many cases do not warrant it.

2. Justification of Approach
    i. It is essential that the data is clean, and that misleading or missing values are updated to a form that will help reach the goal. Any miscue in the data can lead to false assumptions about which variables are predictors of risk.

    ii. This data set contains many fields with values that simply will not work with statistical packages and analysis, which is why it is necessary to update categorical data to numerical data where possible. Excluding fields due to their categorical nature limits the data set and could prohibit the analysis from identifying the right variables.

    iii. Many tools in the plan to find anomalies provide results that are excellent indicators of the quality of the data. For example, using box plots to find outliers and standardized numerical z-scores gives great insight into the range and variability of the data. Even the manipulation of the data, such as changing misleading field values and updating nan or null values serve as great indicators as to quality and strength of the data.

3. Python will be the primary programming language used in this exercise. Python is a very powerful tool for data analysis along with many of the libraries that support it. Although many would argue that R is more friendly for statistics, Python is able to accomplish the same functions and is perhaps superior due to its ability to take statistical analysis that ultimately leads to more easily deployable models. In the long run, the goal of this project is not to simply just analyze historical data. The insights drawn for the analysis can build models that can ultimately help predict risk of hospital readmissions. (ref: https://www.geeksforgeeks.org/replace-nan-values-with-zeros-in-pandas-dataframe/)
    i. Libraries necessary for this project are pandas, which loads, houses, and stores all manipulates to the data; numpy, a powerful mathematics and statistics tool; and matplotlib with its flexible visualizations. These libraries are all Python staples and work together seamlessly to help clean and prepare data.
    ii. Additionally, Python has a large community and resources to help with analysis and coding.
4. Code:
    i. Set index field:

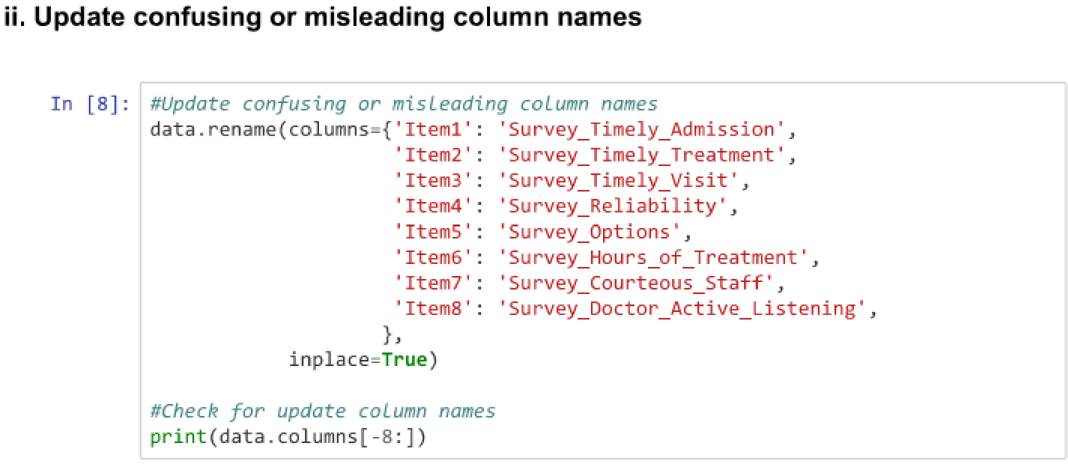


1/17/2021 D206 Project

```
In [7]: #Set index column for the dataframe
        data.set_index('CaseOrder')

        #ref: https://realpython.com/python-data-cleaning-numpy-pandas/
```

    ii. Update confusing or misleading column names:

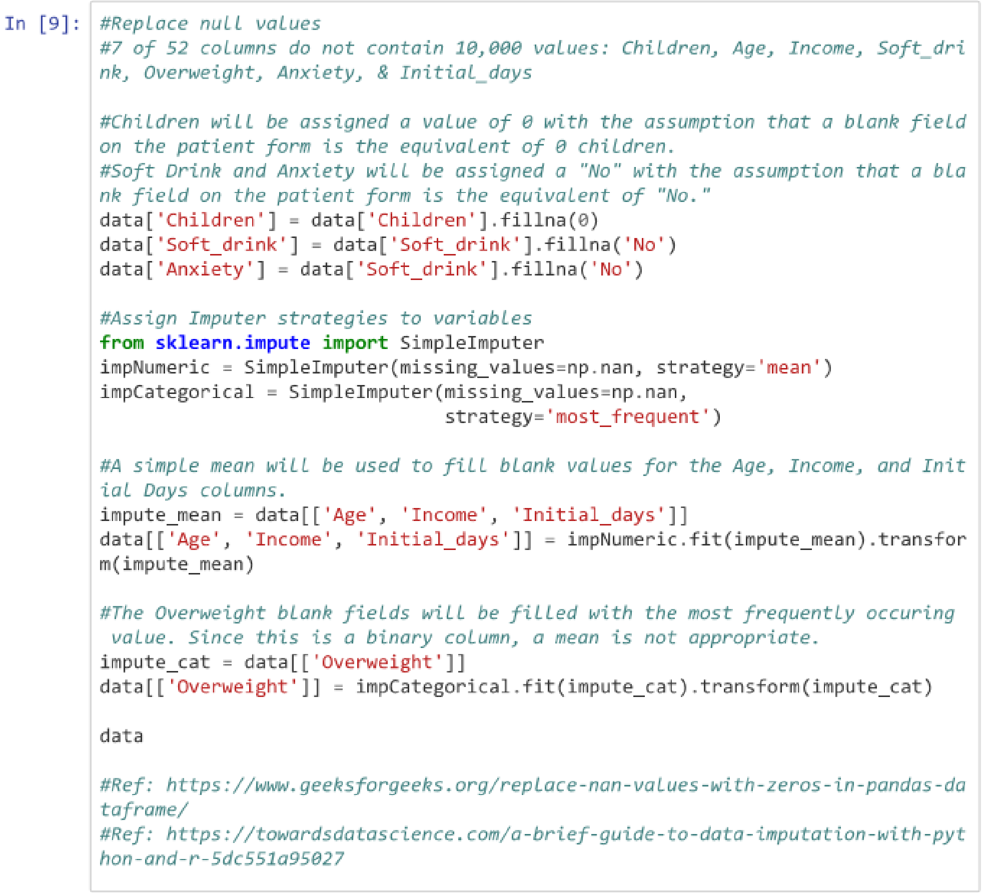


ii. Update confusing or misleading column names

```
In [8]: #Update confusing or misleading column names
        data.rename(columns={'Item1': 'Survey_Timely_Admission',
                             'Item2': 'Survey_Timely_Treatment',
                             'Item3': 'Survey_Timely_Visit',
                             'Item4': 'Survey_Reliability',
                             'Item5': 'Survey_Options',
                             'Item6': 'Survey_Hours_of_Treatment',
                             'Item7': 'Survey_Courteous_Staff',
                             'Item8': 'Survey_Doctor_Active_Listening',
                            },
                    inplace=True)

        #Check for update column names
        print(data.columns[-8:])
```

    iii. Update NaN and null values



```
In [9]: #Replace null values
#7 of 52 columns do not contain 10,000 values: Children, Age, Income, Soft_drink, Overweight, Anxiety, & Initial_days

#Children will be assigned a value of 0 with the assumption that a blank field on the patient form is the equivalent of 0 children.
#Soft Drink and Anxiety will be assigned a "No" with the assumption that a blank field on the patient form is the equivalent of "No."
data['Children'] = data['Children'].fillna(0)
data['Soft_drink'] = data['Soft_drink'].fillna('No')
data['Anxiety'] = data['Soft_drink'].fillna('No')

#Assign Imputer strategies to variables
from sklearn.impute import SimpleImputer
impNumeric = SimpleImputer(missing_values=np.nan, strategy='mean')
impCategorical = SimpleImputer(missing_values=np.nan,
                               strategy='most_frequent')

#A simple mean will be used to fill blank values for the Age, Income, and Initial Days columns.
impute_mean = data[['Age', 'Income', 'Initial_days']]
data[['Age', 'Income', 'Initial_days']] = impNumeric.fit(impute_mean).transform(impute_mean)

#The Overweight blank fields will be filled with the most frequently occuring value. Since this is a binary column, a mean is not appropriate.
impute_cat = data[['Overweight']]
data[['Overweight']] = impCategorical.fit(impute_cat).transform(impute_cat)

data

#Ref: https://www.geeksforgeeks.org/replace-nan-values-with-zeros-in-pandas-dataframe/
#Ref: https://towardsdatascience.com/a-brief-guide-to-data-imputation-with-python-and-r-5dc551a95027
```

iv. See D206_Jupyter_Notebook Excel file for code. All code is labeled to represent its function and is similarly labeled to the Data Cleaning plan in Part II of this document.

Part III: Data Cleaning

D. Summarize the data cleaning process. (Information is also summarized in Jupyter notebook and organized similarly to the requirements list.)
    1. Describe the findings, including all anomalies, from the implementation of the data-cleaning plan from part C.
        i. The data cleaning process revealed that seven columns were missing data. The columns missing data were identified as Children, Age, Income, Soft_drink, Overweight, Anxiety, and Initial_days. Imputation techniques were required to fill in the missing data. The eight columns at the end of the data set had misleading titles, which required an update to a name representative of the data within the column.
        ii. Several of the column's data types were not appropriate for many data cleaning methods as they were not integers, floats, or numerical in form. These values were reexpressed to a numerical data type.
        iii. Outliers were apparent within the data set. All applicable non-binary numerical fields were tested for outliers via a boxplot. Twenty-one fields total were tested for outliers. Sixteen of these fields were standardized to alleviate the concern of outliers having a negative impact on the results of the analysis.
    2. Justify the methods for mitigating each type of discovered anomaly in the data set.
        i. Missing values for the Children column were set as 0 under the assumption that a patient not including this on the intake form is indicative of the patient not having kids. The Soft_drink and Anxiety columns were similarly assigned a value of "No" under the logic that a missing value on the intake form is equivalent to "no." This is an appropriate response for these columns since it is known the patient had an opportunity to fill this out, filled out other items on the form, but left these blank.
        ii. Missing values for the Age, Income, and Initial_days columns were separately assigned values by taking a straight mean of related fields with values. A mean is appropriate here due to the nature of these metrics. These are fields that should have values under every circumstance and the safest way to train the data is to use a simple statistic such as the mean. Fields that were null in the Overweight column were assigned the most occurring value in the non-null fields. This is a binary column, so the only appropriate values for the column are 1 or 0. There are several ways that this could be assigned, but all are very subjective. The assumption in this case is that the most occurring value of all the responses is reflective of the population with null values in this column.
        iii. In assigning numerical values, in certain instances it was necessary to assign values that required a legend. For example, there is no science behind associating a number to the several Education field values. It is wildly up to interpretation. The best practice in all cases that required a

number to be assigned to it is to either associate a number that may be associated to it in other instances or to assign values strategically.

iv. The eight misleading survey columns were assigned a name derived from the column definition, ensuring that the column name is perfectly in line with what the data represents.

v. Values were standardized in 16 fields. To maintain consistency in choosing when to standardize, any fields with an outlier were standardized.

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

   i. Add/reset the index field

      - The index field was set to the CaseOrder column. This column preserves the original order of the rows.

   ii. Update confusing or misleading column names

      - The eight column names that were updated were represented by the new names in all future instances of calling the data. The updated names allowed for easier manipulation and analysis of the data given that it was no longer necessary to use the data dictionary to make sense of what the column represented.

   iii. Update nan and null values

      - All NaN and null fields were updated, resulting in a complete data set. Missing fields were strategically assigned values as described above. The scope of the analysis and establishment of trends expanded upon update.

   iv. Reexpress categorical data as numeric data

      - All categorical data requiring reexpression to numeric data created new columns for the numeric representation of the column. Both columns were preserved as opposed to overwriting the original data; however, only the numerical values were included in future cleaning and manipulation.

   v. Identify outliers

      - Much of the code written was to identify outliers. Boxplots were generated twenty-one times for this purpose. In five instances, no action was necessary, and the fields were left as is. Sixteen boxplots showed outliers within the data and indicated further action was required.

   vi. Standardize numeric fields

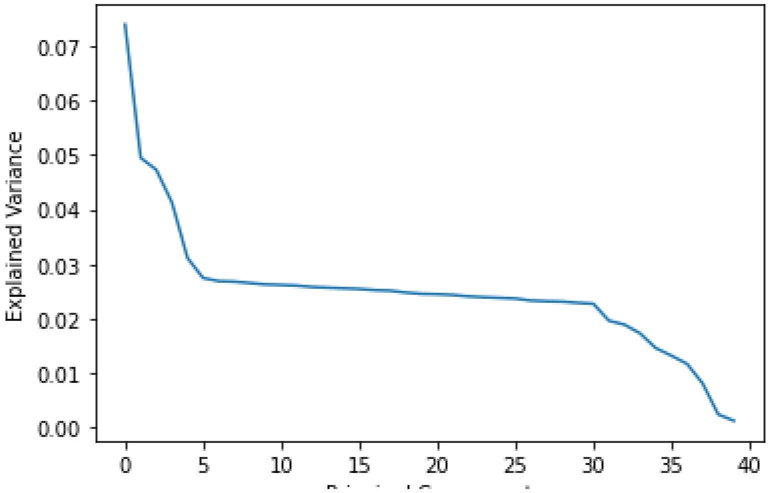
      - Sixteen fields were standardized because of outliers identified in the boxplot outlier analysis. New columns were generated with the newly standardized value as opposed to the original data being overwritten.

4. See D206_Jupyter_Notebook Excel file for code. All code is labeled to represent its function and is similarly labeled to the Data Cleaning plan in Part II of this document.

5. See the attached medical_raw_data_cleaned csv for a copy of the cleaned data set.
6. The data-cleaning process is particularly limited for this data due to no access to data sources, such as to the intake forms. No access to health experts who could potentially provide expertise as to the best strategy to assign null values is also particularly limiting, since the subjective decision of an analyst to assign values may not align with the standard of the medical community. Binary and yes/no columns can limit the data. For example, it is helpful to consider whether someone is overweight, but it would be even more helpful to have the value that determined whether someone was overweight. Assigning numerical values to categorical values can also be limiting since it is a very subjective process, and it could very handled very differently by different analysts.
7. In a circumstance where a value was assumed for null or NaN values that is inconsistent with health experts could incorrectly identify variables as being more relevant to the risk of hospital readmission than they are. This is similar for binary columns. While binary columns are easy clean, an incorrect cleaning strategy could incorrectly highlight variables. When it comes to assigning numerical values to categorical values, if a number should reflect an extreme difference between two categories, but it is not, then columns could be overlooked.

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

1. This data set contains 40 principal components. Due to the number of principal components, the percent of variability is spread thin. The maximum variance is 7.4% with a minimum of 0.13%. 5 Principial Components make up 25% of the variability. Since it requires a good number of Principal Components to cover a majority or more of the data set, deriving significant features from the data set is challenging.
   i. The following graphs represents the Explained Variance in relation to the number of Principal Components.



ii. According to the top 6 Principal Components, the most significant features of the data set are:

- Initial_days
- VitD_levels
- Survey_Options
- Survey_Reliability
- Services_Numeric
- Gender_Numeric
- HighBlood_Numeric
- Overweight
- Initial_admin_Numeric

*Determining significant features is highly subjective and is even more difficult to determine in a data set that required many principal components to cover most of the data set.

2. To begin, I created a DataFrame with only numeric columns. Columns such as state, lat, lng, and timezone were removed due to excessive variability and/or due to frequently repeating values. Readmis was removed as it is the target value of the research. Next, data was normalized, all components were set to be extracted from the PCA, and the PCA was called in to the data set. With the data now in place, two Scree plots were generated to compare Explained Variance across the 40 Principal Components. Eigenvalues were then derived from the data set and a Scree plot was generated to check the new values. Finally, loadings were generated ultimately allowing for analysis to determine significant features. To identify significant features, I set the first 6 PCA columns to be an absolute value and then sorted the values highest to lowest for each column one at a time. I found reoccurrences across Principal Components and selected the most frequently occurring as the significant features.
3. The related groupings in PCA are a great first step at determining what variables are most indicative of readmission. PCA will indicate how well the original data best groups together. The resulting loading will begin to highlight where strong correlations exist between the grouping and the variable. Another great benefit of PCA that the organization will gain is precious in the data.

Part IV: Supporting Documents

F. See attached Panopto video.
G. Sources for third-party code and information are documented in the references page as well as in text with the code in the Jupyter Notebook.
H. See accompanying references page.
I. No response necessary.

Works Cited

4 Apr. 2020, www.youtube.com/watch?v=oiusrJ0btwA.

Convert Floats to Ints in Pandas? 24 Jan. 2021, stackoverflow.com/questions/21291259/convert-floats-to-ints-in-pandas.

Corp), Pavel Horbonos (Midvel. *A Brief Guide to Data Imputation with Python and R*. 31 May 2020, towardsdatascience.com/a-brief-guide-to-data-imputation-with-python-and-r-5dc551a95027.

How to Set Column as Index in Pandas DataFrame? pythonexamples.org/pandas-set-column-as-index/#:~:text=Pandas – Set Column as Index&text=To set a column as,index, to set_index() method.

joshstarmer. 8 Jan. 2018, www.youtube.com/watch?v=Lsue2gEM9D0.

LincolnJLincolnJ 4111 silver badge55 bronze badges, et al. *Python Pandas: Dtypes Not Show Column Types for All Columns*. 1 Nov. 1963, stackoverflow.com/questions/28238919/python-pandas-dtypes-not-show-column-types-for-all-columns.

Python (Programming Language). 16 Apr. 2021, en.wikipedia.org/wiki/Python_(programming_language).

Real Python. *Pythonic Data Cleaning With Pandas and NumPy*. 13 Feb. 2021, realpython.com/python-data-cleaning-numpy-pandas/.

Replace NaN Values with Zeros in Pandas DataFrame. 3 July 2020, www.geeksforgeeks.org/replace-nan-values-with-zeros-in-pandas-dataframe/.

Waterman, Tom. *Why Python Is Better than R for Data Science Careers*. 14 Jan. 2020, towardsdatascience.com/why-python-is-better-than-r-for-data-science-careers-44ec7a149a18.

What Are Categorical, Discrete, and Continuous Variables? support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/#:~:text=Categorical variables contain a finite number of categories or distinct groups.&text=Continuous variables are numeric variables,time a payment is received.