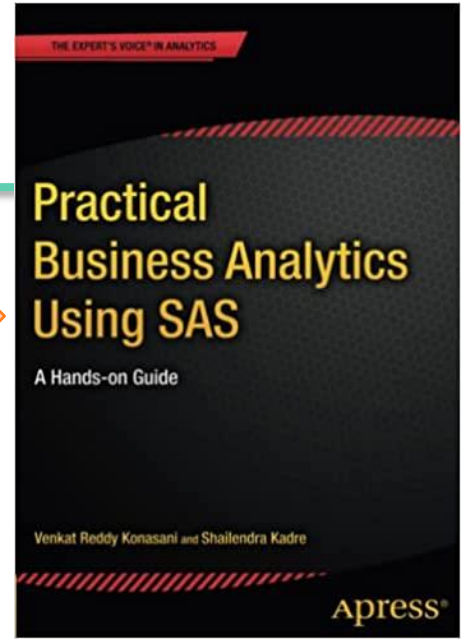
Data Cleaning and Preparing data for Analysis

Venkat Reddy Konasani

Chapter 7 in the book



Contents

- Model Building Life cycle
- Raw Data - issues
- Data Exploration – Categorical and Discrete Variables
- Data Exploration – Continuous Variables
- Data Validation
- Data sanitizations techniques
- Standalone Imputation
- Imputation based on a target

Model Building Life Cycle

Background and Objective	Data Exploration	Preparing data for analysis	Building the model	Validating the model	Deployment
Business Objective	Collect data	Validate data	Select the right model	Intime validation	Deploy model
Set Goals	Explore data	Outlier treatment	Variable selection	Out of time validation	Maintenance of model
Project Plan	Basic Summary	Missing value treatment	Model building and finetuning	Model finetuning	Model monitoring
Budget & Resources	Identify outliers and missing values	Clean the data	Model iterations		
		Prepare data for Analysis			

Data Exploration and Validation

- Junk in Junk out
- The final model is as good as the input data.
- You can never build a great model on erroneous input data

Model Building Life Cycle

Background and Objective	Data Exploration	Preparing data for analysis	Building the model	Validating the model	Deployment
Business Objective	Collect data	Validate data	Select the right model	Intime validation	Deploy model
Set Goals	Explore data	Outlier treatment	Variable selection	Out of time validation	Maintenance of model
Project Plan	Basic Summary	Missing value treatment	Model building and finetuning	Model finetuning	Model monitoring
Budget & Resources	Identify outliers and missing values	Clean the data	Model iterations		
		Prepare data for Analysis			

Raw Data - issues

The raw data is dirty

- Wrong formats- expenses is read as date
- Might have missing values – Income missing for some records
- Might have outliers – Number of loans is 25000
- Erroneous values – Age is less than 0
- Default values – Account tenure is 999999
- Inconsistent – Age is 25, year of birth is 1970

Preparing data for analysis

- We can't directly start the analysis and model building with raw data.
- Before getting on to core analysis and strategy building it is very important to
  - Explore the data
  - Validate the data
  - And finally clean the data and prepare it for analysis

Case Study- Data Exploration

Give me some credit data

- https://www.kaggle.com/c/GiveMeSomeCredit
- We will try to understand the data exploration, validation and data cleaning using a case study on loans data
- Give me some credit data. It is loans data. Historical data are provided on 150,000 borrowers.
- The final objective is to build a model that borrowers can use to help make the best financial decisions.
- We generally get the data and data dictionary from the data team.

Data Dictionary

No	Variable Name	Short Description	Example	Min Value	Max Value
1	SeriousDlqin2yrs	Target Variable (loan defaulter)			
2	RevolvingUtilizationOfUnsecuredLines	Credit Utilization			
3	age	Age			
4	NumberOfTime30-59DaysPastDueNotWorse	One month late frequency			
5	DebtRatio	Debt to income ratio			
6	MonthlyIncome	Income			
7	NumberOfOpenCreditLinesAndLoans	Number of loans			
8	NumberOfTimes90DaysLate	Three months late frequency			
9	NumberRealEstateLoansOrLines	House loans			
10	NumberOfTime60-89DaysPastDueNotWorse	Two months late frequency			
11	NumberOfDependents	Dependents			

Data Dictionary

No	Variable Name	Short Description	Description	Variable Type
1	SeriousDlqin2yrs	Target Variable (loan defaulter)	Person experienced 90 days past due delinquency or worse	Y/N
2	RevolvingUtilizationOfUnsecuredLines	Credit Utilization	Total balance on credit cards and personal lines of credit except real estate and no instalment debt like car loans divided by the sum of credit limits	percentage
3	age	Age	Age of borrower in years	integer
4	NumberOfTime30-59DaysPastDueNotWorse	One month late frequency	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
5	DebtRatio	Debt to income ratio	Monthly debt payments, alimony, living costs divided by month gross income	percentage
6	MonthlyIncome	Income	Monthly income	real
7	NumberOfOpenCreditLinesAndLoans	Number of loans	Number of Open loans (instalment like car loan or mortgage) and Lines of credit (e.g. credit cards)	Integer
8	NumberOfTimes90DaysLate	Three months late frequency	Number of times borrower has been 90 days or more past due.	integer
9	NumberRealEstateLoansOrLines	House loans	Number of mortgage and real estate loans including home equity lines of credit	integer
10	NumberOfTime60-89DaysPastDueNotWorse	Two months late frequency	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
11	NumberOfDependents	Dependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Steps in Data Exploration and Cleaning

- Step-1: Basic details of the data
- Step-2: Categorical variables exploration
- Step-3: Continuous variables exploration
- Step-4: Missing Values and Outlier Treatment

Step-1: Basic details of the data

Check the Metadata

- Metadata is data about data
- What are total number of observations and variables
- Check each field name, field type, Length of field
- Are there some variables which are unexpected say q9 r10?
- Are the data types and length across variables correct
- For known variables is the data type as expected (For example if age is in date format something is suspicious)

Print first few records

- Do we have any unique identifier? Is the unique identifier getting repeated in different records?
- Do the text variables have meaningful data?
- Are there some coded values in the data
- Do all the variables appear to have data? Are there any missing values

Lab: Basic contents of the data

- Import “Give me some Credit\cs-training.csv”
- What are number of rows and columns
- Are there any suspicious variables?
- Are all the variable names correct?
- Display the variable formats
- Print the first 10 observations
- Do we have any unique identifier?
- Do the text and numeric variables have meaningful data?
- Are there some coded values in the data?
- Do all the variables appear to have data

Code: Basic contents of the data

```
import pandas as pd
loans=pd.read_csv("D:\\Google Drive\\Training\\Datasets\\Give me
some Credit\\cs-training.csv")
loans

#What are number of rows and columns
loans.shape

#Are there any suspicious variables?
loans.columns.values

#Display the variable formats
loans.dtypes
```

Code: Basic contents of the data

```
#Print the first 10 observations
loans.head(10)

#Do we have any unique identifier?
loans.columns.values
```

Categorical, Discrete and Continuous Variables

Data Dictionary

No	Variable Name	Short Description	Categorical/Discrete /Continuous
1	SeriousDlqin2yrs	Target Variable (loan defaulter)	
2	RevolvingUtilizationOfUnsecuredLines	Credit Utilization	
3	age	Age	
4	NumberOfTime30-59DaysPastDueNotWorse	One month late frequency	
5	DebtRatio	Debt to income ratio	
6	MonthlyIncome	Income	
7	NumberOfOpenCreditLinesAndLoans	Number of loans	
8	NumberOfTimes90DaysLate	Three months late frequency	
9	NumberRealEstateLoansOrLines	House loans	
10	NumberOfTime60-89DaysPastDueNotWorse	Two months late frequency	
11	NumberOfDependents	Dependents	

Step-2: Categorical and Discrete variables exploration

The Frequency Table and summary

- Calculate frequency counts cross-tabulation frequencies for Especially for categorical, discrete & class fields
- Frequencies
  - help us understanding the variable by looking at the values it’s taking and data count at each value.
  - They also helps us in analyzing the relationships between variables by looking at the cross tab frequencies or by looking at association

The Frequency Table - Visualization

Use horizontal or Vertical Bar charts

```
import seaborn as sns
sns.countplot(y="SeriousDlqin2yrs",  data=loans)
```

Check Points

1. Are values as expected?
2. Variable understanding : Distinct values of a particular variable, missing percentages
3. Are there any extreme values or outliers?
4. Any possibility of creating a new variable having small number of distinct category by clubbing certain categories with others.

Lab: Frequencies & Bar charts

- What are the categorical and discrete variables? What are the continues variables.
- Find the frequencies of all class variables in the data
- Are there any variables with missing values?
- Are there any default values?
- Can you identify the variables with outliers?
- Are there any variables with other issues?

Code: Frequencies

```
print(loans['SeriousDlqin2yrs'].value_counts())
sns.countplot(y="SeriousDlqin2yrs",  data=loans)

print(loans['age'].value_counts())
sns.countplot(y="age",  data=loans)

print(loans['NumberOfTime30-59DaysPastDueNotWorse'].value_counts())
sns.countplot(y="NumberOfTime30-59DaysPastDueNotWorse",  data=loans)

print(loans['NumberOfOpenCreditLinesAndLoans'].value_counts())
sns.countplot(y="NumberOfOpenCreditLinesAndLoans",  data=loans)
```

Data Dictionary

No	Variable Name	Short Description	Issues
1	SeriousDlqin2yrs	Target Variable (loan defaulter)	
2	RevolvingUtilizationOfUnsecuredLines	Credit Utilization	
3	age	Age	
4	NumberOfTime30-59DaysPastDueNotWorse	One month late frequency	
5	DebtRatio	Debt to income ratio	
6	MonthlyIncome	Income	
7	NumberOfOpenCreditLinesAndLoans	Number of loans	
8	NumberOfTimes90DaysLate	Three months late frequency	
9	NumberRealEstateLoansOrLines	House loans	
10	NumberOfTime60-89DaysPastDueNotWorse	Two months late frequency	
11	NumberOfDependents	Dependents	

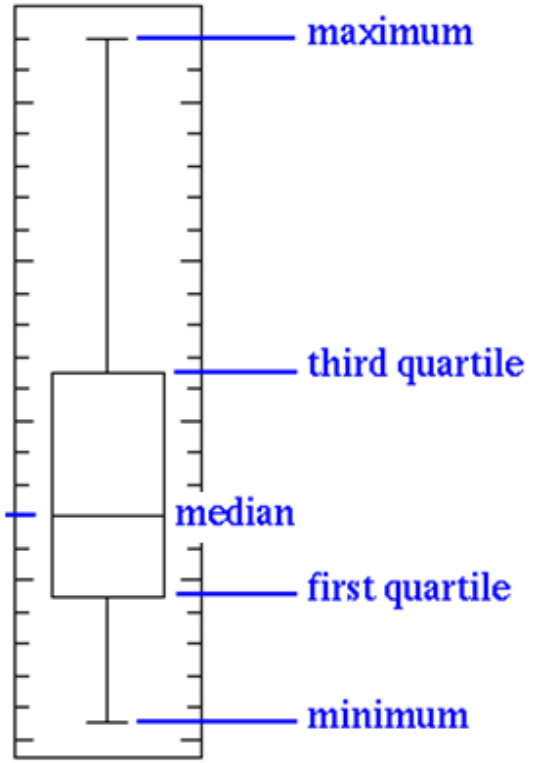
Step-3: Continuous variables exploration

Summary of Continuous variables

- Study the percentile distrinution
- Percentiles– P1, p5,p10,q1(p25),q3(p75), p90,p99
- Box plots and identification of outliers

Visualization of Continuous variables

- Use box plots or Histograms for visualization



Check Points

- Are variable distribution as expected.
- What is the central tendency of the variable? Mean, Median and Mode across each variable
- Is the concentration of variables as expected ? What are quartiles?
- Indicates variables which are unary I.e stddev=0 ; the variables which are useless for the current objective.
- Are there any outliers / extreme values for the variable?
- Are outlier values as expected or they have abnormally high values - for ex for Age if max and p99 values are 10000. Then should investigate if it’s the default value or there is some error in data
- What is the % of missing value associated with the variable?

LAB: Continuous variables summary

- List down the continuous variables
- Find summary statistics for each variable. Min, Max, Median, Mean, sd, Var
- Find Quartiles for each of the variables
- Create Box plots and identify outliers
- Find the percentage of missing values
- Find Percentiles and find percentage of outliers, if any P1, p5,p10,q1(p25),q3(p75), p90,p99

Code: Continuous variables summary

RevolvingUtilizationOfUnsecuredLines

```
plt.boxplot(loans["monthly_utilization"])
plt.hist(loans["monthly_utilization"])

util_percentiles=loans['monthly_utilization'].quantile([0.05, 0.1, 0.2
5, 0.5, 0.75, 0.80, 0.9,0.91,0.95,0.96,0.97,0.975,0.98,0.99,1])
round(util_percentiles,2)
```

Code: Continuous variables summary

MonthlyIncome

```
plt.boxplot(loans["MonthlyIncome"][loans["MonthlyIncome"].isnull()==False])
plt.hist(loans["MonthlyIncome"])

#Find the percentage of missing values
print("Count of missing values")
print(loans['MonthlyIncome'].isnull().sum())
print("% of missing values")
print(round(loans['MonthlyIncome'].isnull().sum()/len(loans),2))
```

Data Cleaning

Data Cleaning

- Some variables contain outliers
- Some variables have default values
- Some variables have missing values
- Shall we delete them and go ahead with our analysis?
- Shall we keep them and go ahead with our analysis?

- RevolvingUtilizationOfUnsecuredL
- NumberOfTime30_59DaysPastDueNotW
- Monthly income has missing values

Keeping the outliers

- Keeping the outliers - Is it a good option? – Definitely, not a good option
- Outliers can turn our results upside down.

Deleting the outliers – Not a good option

				Income					
				xx					
				xx					

Deleting the outliers – Not a good option

				Income				Util	
				XX					
				XX					
								XX	
								XX	

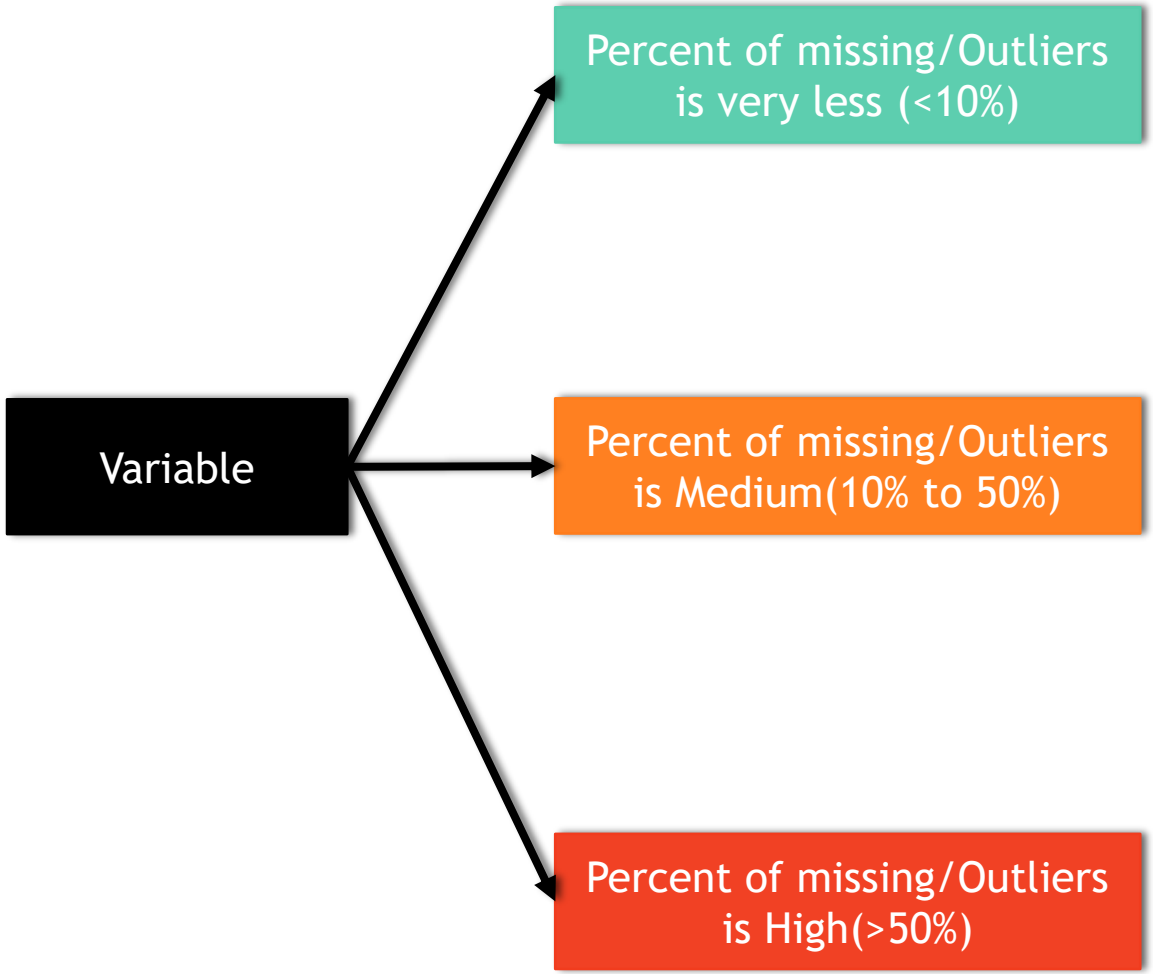
Deleting the outliers – Not a good option

		Loans		Income				Util	
		XX							
				XX					
				XX					
								XX	
								XX	

Reason for Missing values & Outliers

- Data is not always available E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Not register history or changes of the data
- Missing data may need to be inferred.
- Missing data - values, attributes, entire records, entire sections
- Missing values and defaults are indistinguishable

Magnitude of the errors

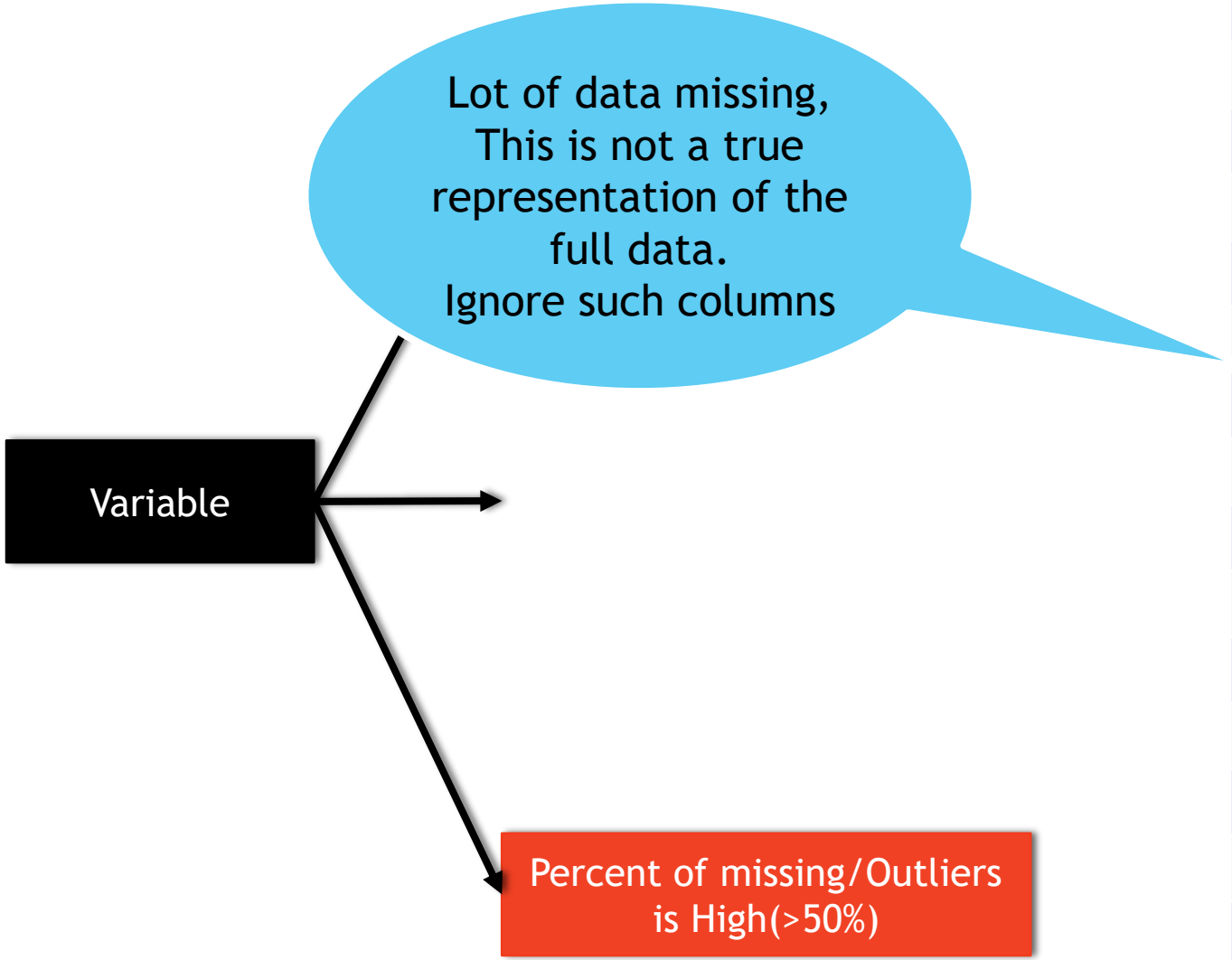


>50% data has issues

>50% data has issues

- What if more than 50% are missing?
- It doesn’t make sense to carry out the analysis on 20% or 30% of the whole data and give inferences on overall data
- The best imputation is ignore the actual values and take available or not available info

>50% data has issues



Income
.
.
.
.
.
.
.
1500
.
.
.
.
.
.
2000
2500

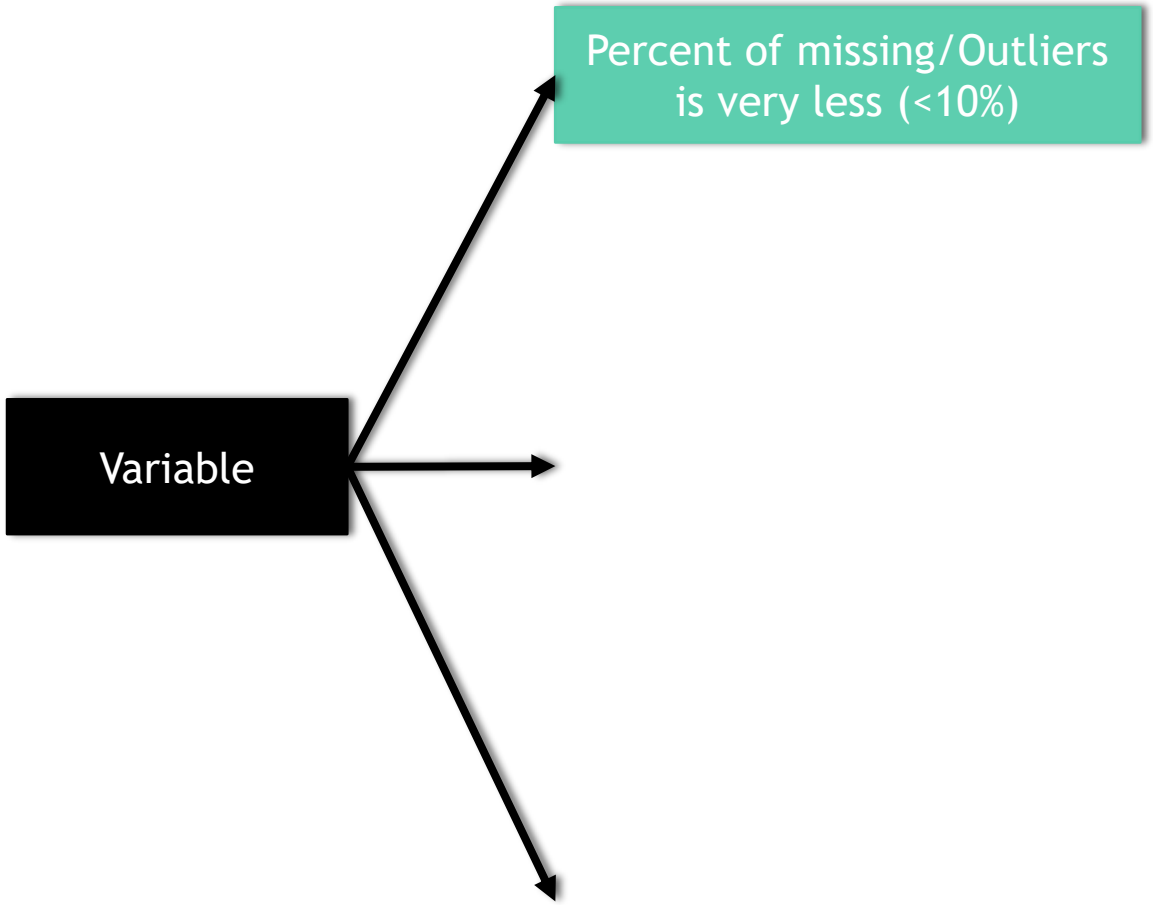
>50% data has issues

Income	Income_ind
.	0
.	0
.	0
.	0
.	0
.	0
.	0
1500	1
.	0
.	0
.	0
.	0
.	0
.	0
2000	1
2500	1

create a flag variable and drop the original variable

Imputation

Magnitude of the errors



X1
11.0
11.1
11.9
10.9
10.8
.
11.5
11.6
11.6
11.4
11
12
11.8
11.4
11.9
11.5
10.9

Standalone imputation

- Mean, median, other point estimates
- Convenient, easy to implement
- **Assume:** Distribution of the missing values is the same as the non-missing values.
- Does not take into account inter-relationships

- **Eg:** The average of available values is 11.4. Can we replace the missing value in this table by **11.4** ?

X1
11.0
11.1
11.9
10.9
10.8
.
11.5
11.6
11.6
11.4
11
12
11.8
11.4
11.9

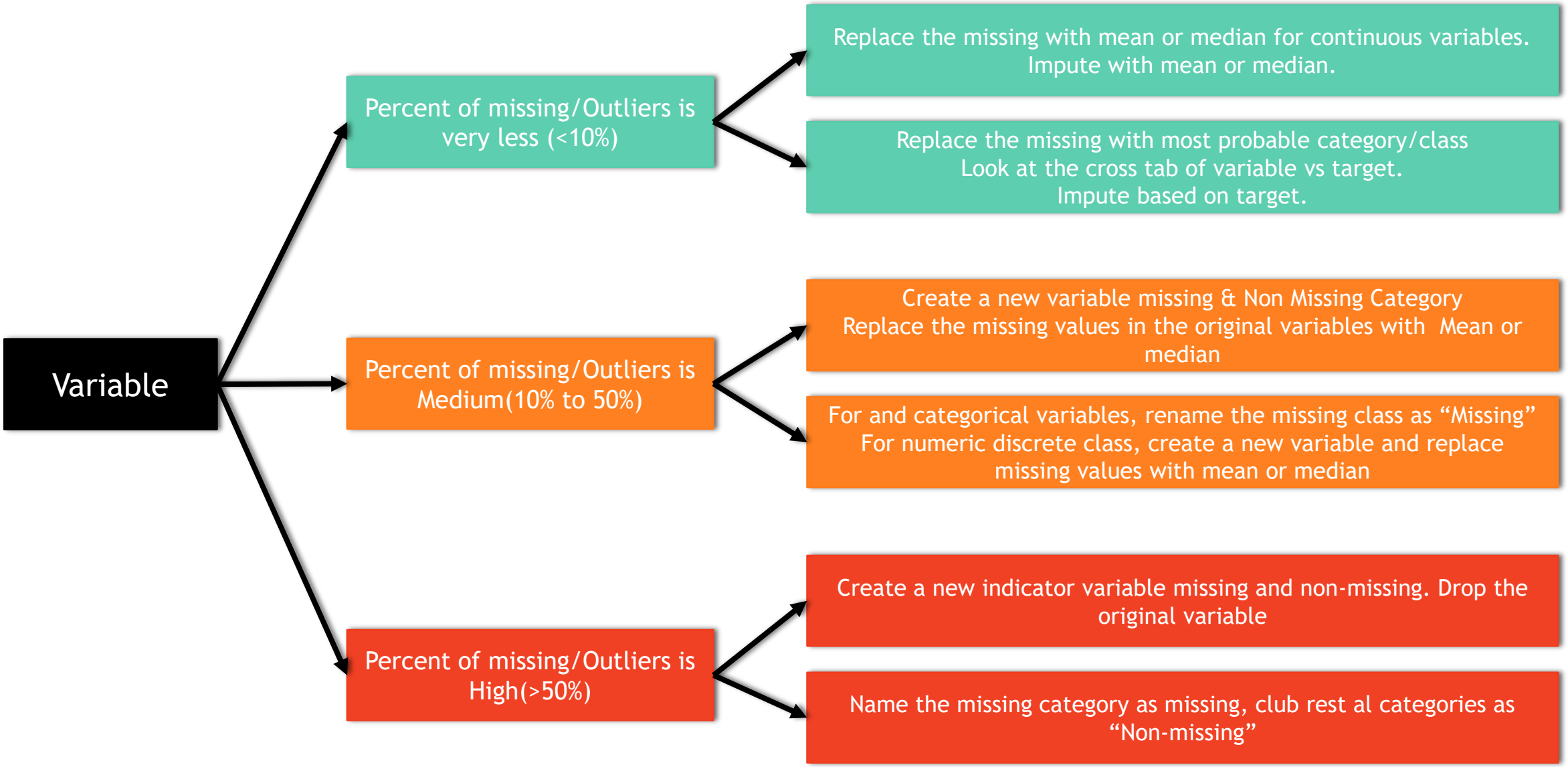
Imputation type2

- Use attribute relationships
- Better imputation
- Two techniques
  - Propensity score (nonparametric). Useful for discrete variables
  - Regression (parametric)

- There are two missing values in x2. What are the most appropriate replacements

X1	X2
-4	-12
2	6
-6	-18
8	24
-1	
-4	-12
-5	-15
4	12
-4	-12
-5	-15
-2	
4	12
10	30
-10	-30
-3	-9

Step-4: Missing Values and Outlier Treatment

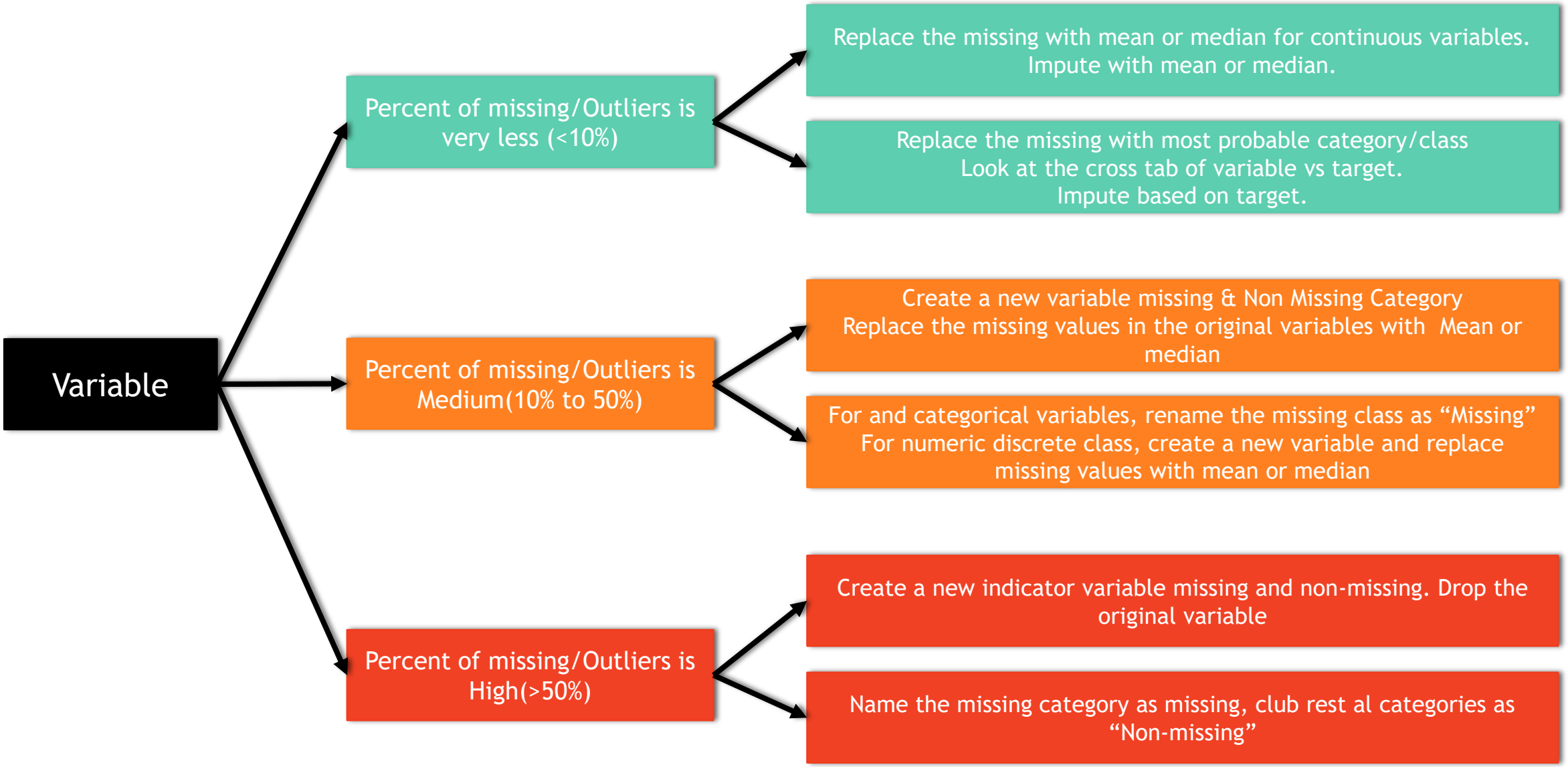

Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with mean or median for continuous variables.
Impute with mean or median.
Replace the missing with most probable category/class
Look at the cross tab of variable vs target.
Impute based on target.
Percent of missing/Outliers is Medium(10% to 50%)
Create a new variable missing & Non Missing Category
Replace the missing values in the original variables with Mean or median
For and categorical variables, rename the missing class as "Missing"
For numeric discrete class, create a new variable and replace missing values with mean or median
Percent of missing/Outliers is High(>50%)
Create a new indicator variable missing and non-missing. Drop the original variable
Name the missing category as missing, club rest al categories as "Non-missing"

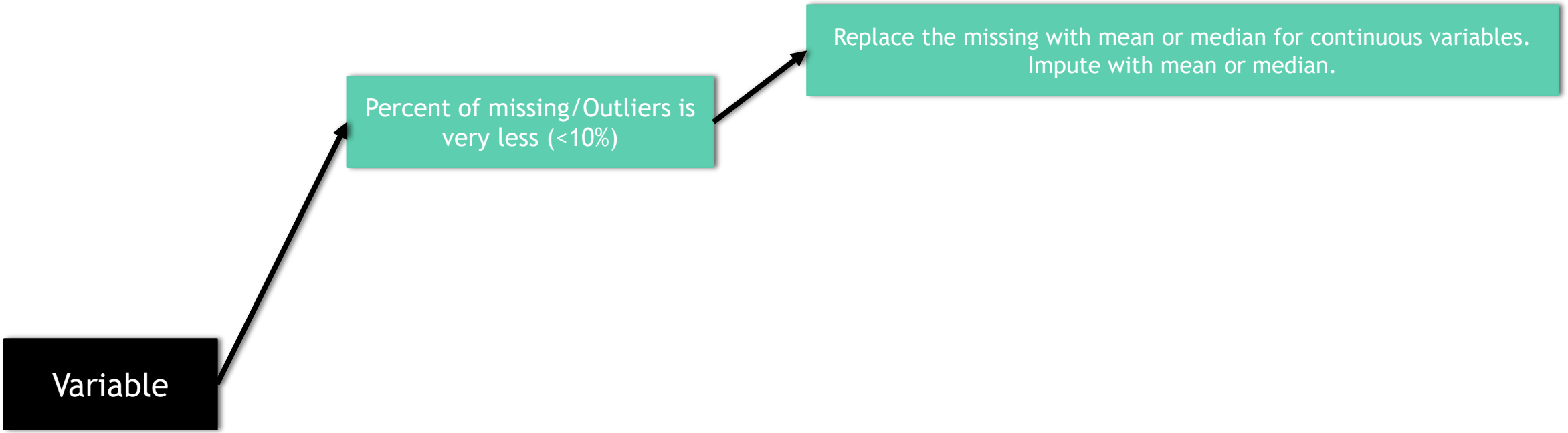
Data Cleaning Scenario-1

RevolvingUtilizationOfUnsecuredLines

- RevolvingUtilizationOfUnsecuredLines has outliers.
- What type of variable is this? What are the possible values?

- Its' mean is 6.05 which is greater than 1. So variable has some faulty values. Its maximum value is 50710 which is way too high.
- Lets look at percentiles to know from where it is exceeding 1.


Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with mean or median for continuous variables.
Impute with mean or median.
Replace the missing with most probable category/class
Look at the cross tab of variable vs target.
Impute based on target.
Percent of missing/Outliers is Medium(10% to 50%)
Create a new variable missing & Non Missing Category
Replace the missing values in the original variables with Mean or median
For and categorical variables, rename the missing class as "Missing"
For numeric discrete class, create a new variable and replace missing values with mean or median
Percent of missing/Outliers is High(>50%)
Create a new indicator variable missing and non-missing. Drop the original variable
Name the missing category as missing, club rest al categories as "Non-missing"


Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with mean or median for continuous variables. Impute with mean or median.

Data Cleaning

- RevolvingUtilizationOfUnsecuredLines has outliers.
- Since outliers percentage is less than 10% We will replace outliers with mean of reaming data.
- Outliers are with value greater than 1.

LAB: Data Cleaning Scenario-1

- What percent are missing values in RevolvingUtilizationOfUnsecuredLines?
- Get the detailed percentile distribution
- Clean the variable, and create a new variable by removing all the issues

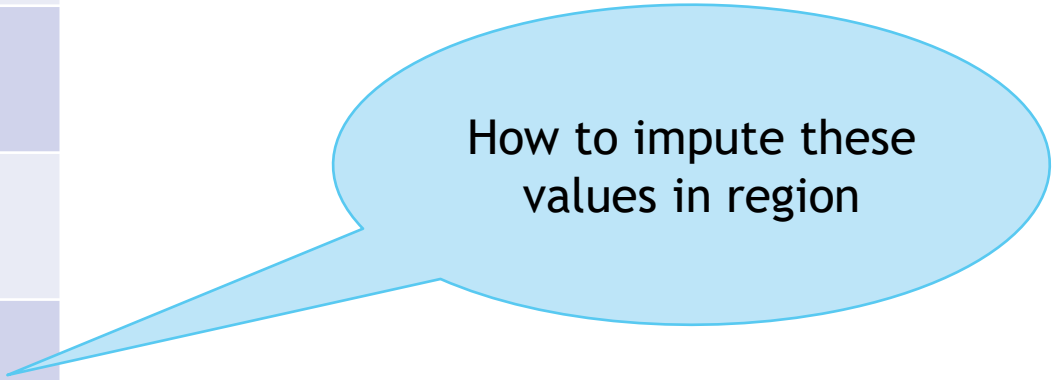
Code: Data Cleaning Scenario-1

```
median_util=loans['monthly_utilization'].median()
median_util
util_temp_bool_vect=loans['monthly_utilization']>1

loans['util_new']=loans['monthly_utilization']
loans['util_new'][util_temp_bool_vect]=median_util

# percentile distribution for the new variable
util_percentiles1=loans['util_new'].quantile([0.05, 0.1, 0.25, 0.5, 0.75, 0.
80, 0.9,0.91,0.95,0.96,0.97,0.975,0.98,0.99,1])
round(util_percentiles1,2)
```

Data Cleaning Scenario-2

Categorical Variables imputation

Region
East
West
North
South
Missing/ NA (5%)



Categorical Variables imputation

Region	Avg Sales
East	1800$
West	5000$
North	2400$
South	3000$
Missing/ NA (5%)	4800$

Impute based on the target

Which is the best region to replace with?

Categorical Variables imputation

Region	Buy / Not Buy
East	
West	
North	
South	
Missing/ NA (5%)	

What if the target is categorical

Categorical Variables imputation

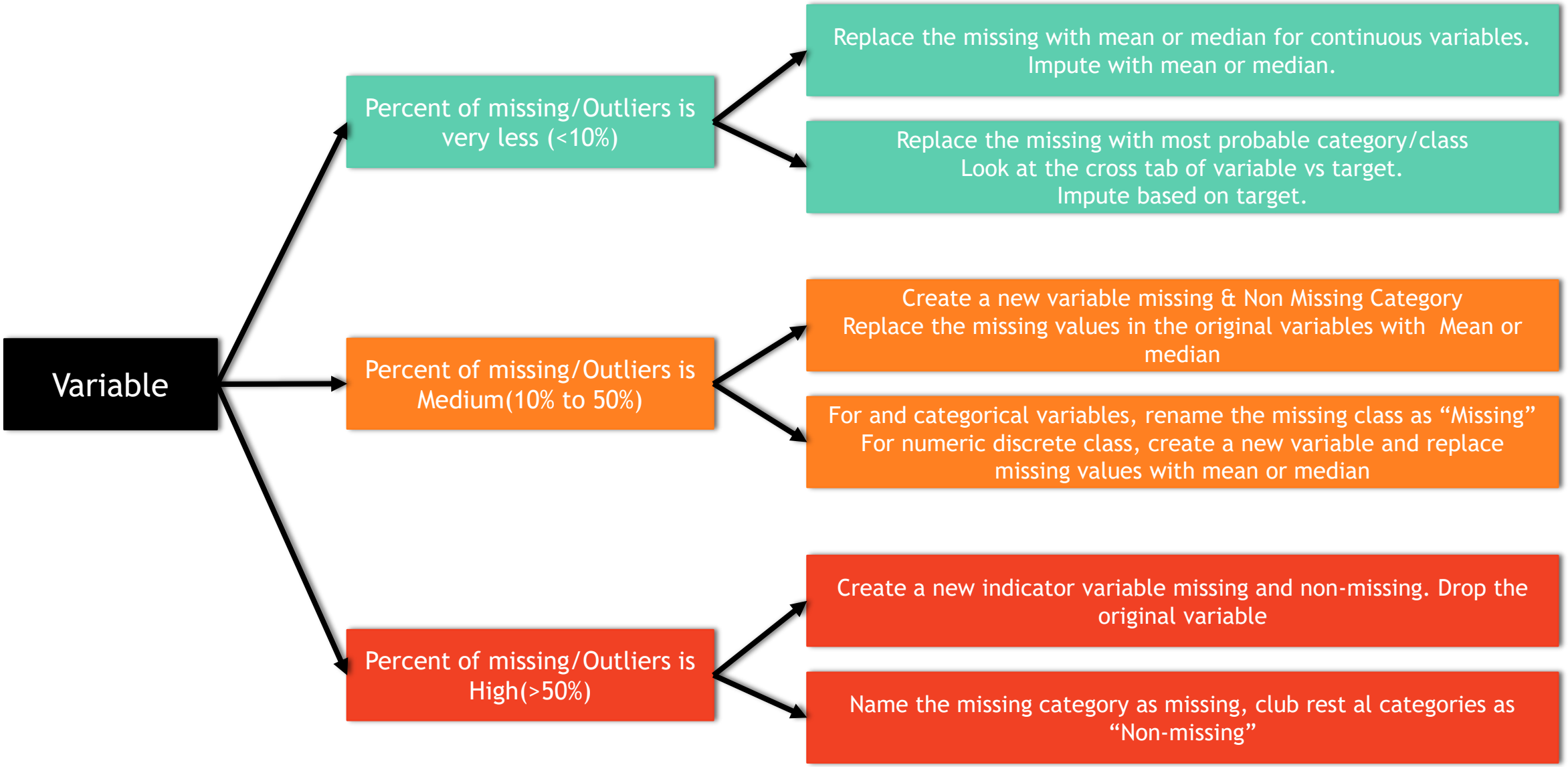
Region	Buy	Not Buy
East	20%	80%
West	10%	90%
North	40%	60%
South	90%	10%
Missing/ NA (5%)	9%	91%

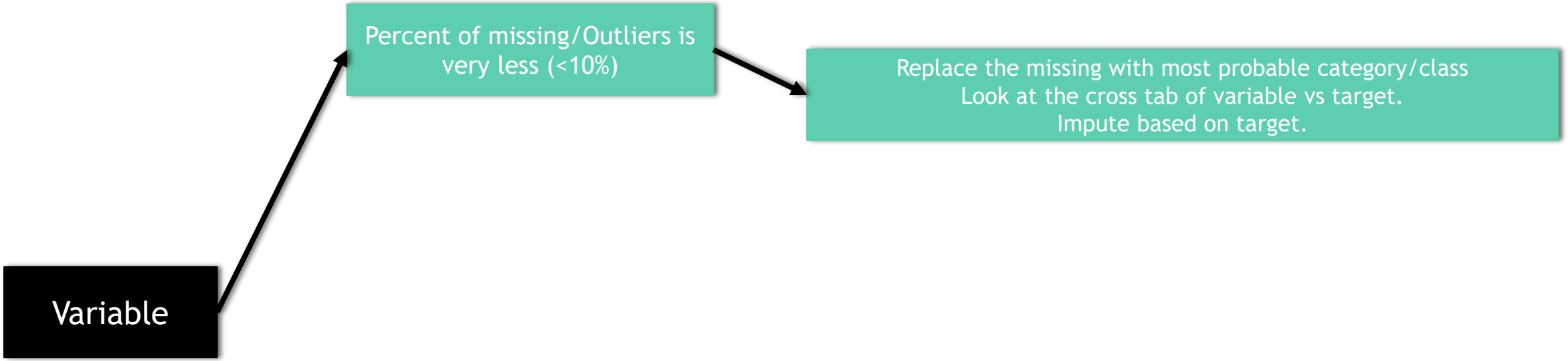
Impute based on the target

Which is the best region to replace with?

NumberOfTime30_59DaysPastDueNotW

- Find bad rate in each category of this variable
- Replace 96 with _____? Replace 98 with_____?


Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with mean or median for continuous variables.
Impute with mean or median.
Replace the missing with most probable category/class
Look at the cross tab of variable vs target.
Impute based on target.
Percent of missing/Outliers is Medium(10% to 50%)
Create a new variable missing & Non Missing Category
Replace the missing values in the original variables with Mean or median
For and categorical variables, rename the missing class as "Missing"
For numeric discrete class, create a new variable and replace missing values with mean or median
Percent of missing/Outliers is High(>50%)
Create a new indicator variable missing and non-missing. Drop the original variable
Name the missing category as missing, club rest al categories as "Non-missing"


Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with most probable category/class
Look at the cross tab of variable vs target.
Impute based on target.

LAB: Data Cleaning Scenario-2

- What is the issue with NumberOfTime30_59DaysPastDueNotW
- Draw a frequency table
- What percent of the values are erroneous?
- Clean the variable- Look at the cross tab of variable vs target. Impute based on target .
- Create frequency table for cleaned variable

Code: Data Cleaning Scenario-2

```
freq_table_30dpd=loans['NumberOfTime30-59DaysPastDueNotWorse'].value_counts()
freq_table_30dpd
#Cross tab with target
import pandas as pd
cross_tab_30dpd_target=pd.crosstab(loans['NumberOfTime30-
59DaysPastDueNotWorse'],loans['SeriousDlqin2yrs'])
cross_tab_30dpd_target

#Cross tab row Percentages
cross_tab_30dpd_target_percent=cross_tab_30dpd_target.apply(lambda x: x/x.sum(), axis=1)
round(cross_tab_30dpd_target_percent,2)
```

Code: Data Cleaning Scenario-2

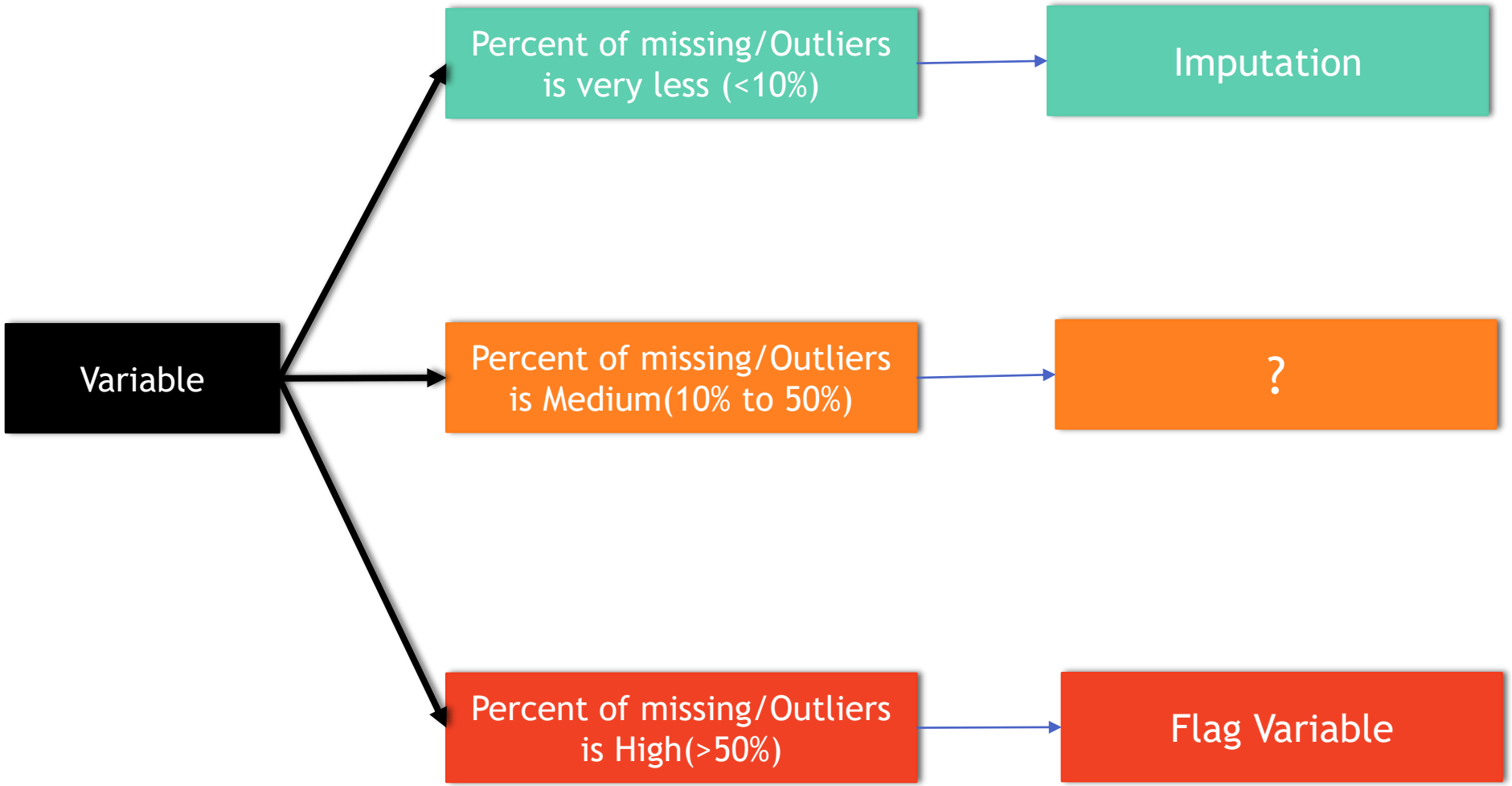
```
#Percentage of 0 and 1 are of 98 is near to percentages of 6.
#Replacing error values with 6

loans['num_30_59_dpd_new']=loans['NumberOfTime30-59DaysPastDueNotWorse']
loans['num_30_59_dpd_new'][loans['num_30_59_dpd_new']>13]=6
loans['num_30_59_dpd_new']

loans['num_30_59_dpd_new'].value_counts()
```

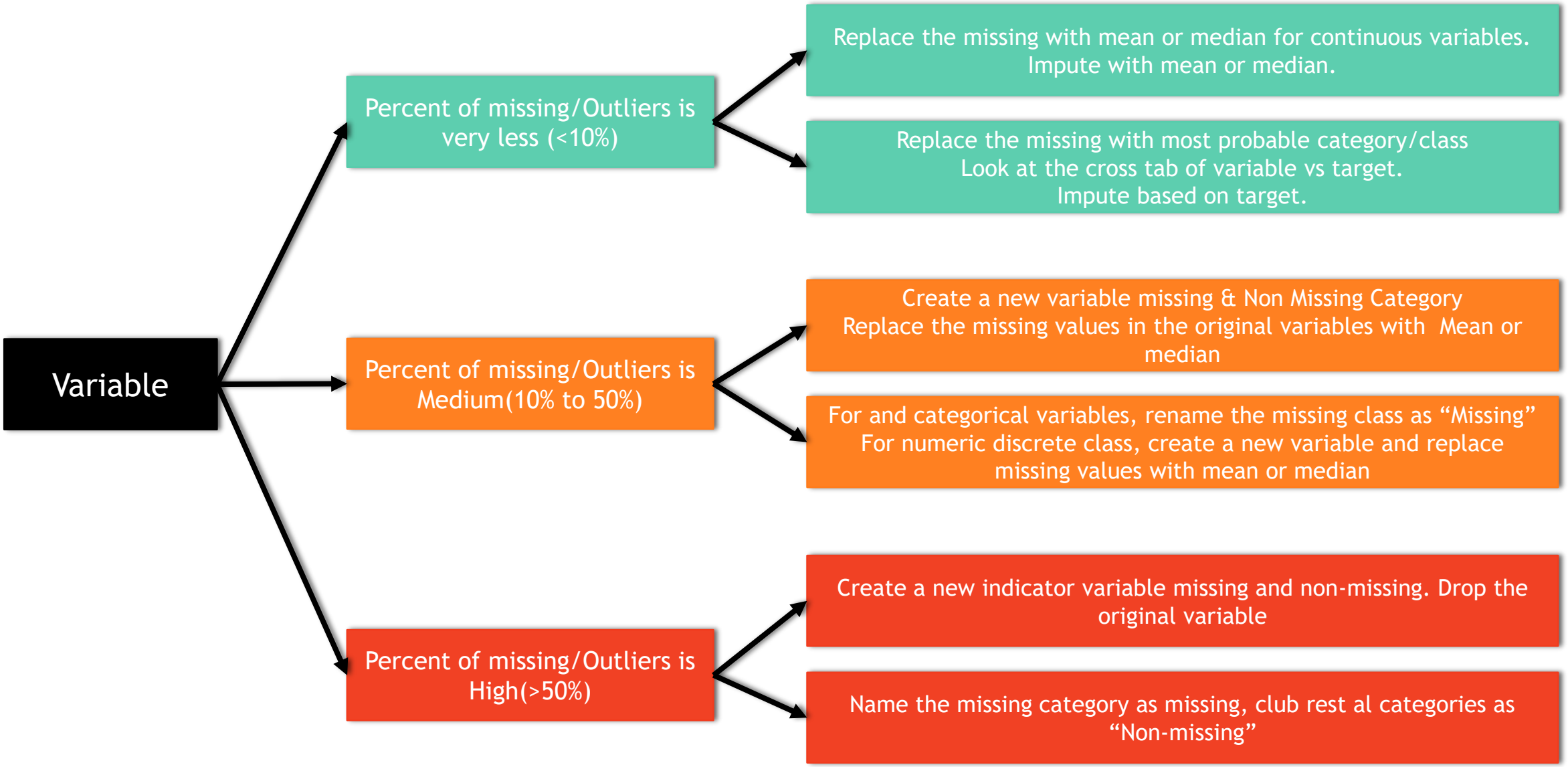
Data Cleaning Scenario-3

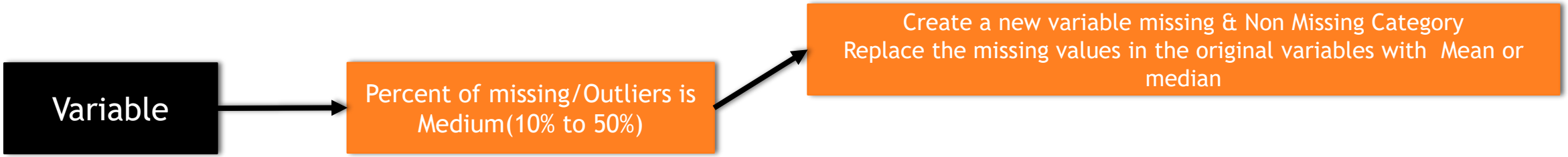
Monthly Income

- Monthly Income has nearly 20% missing values
- Missing value percentage is significant
- Simply replacing with mean or median is not sufficient
- We can create an indicator variable to keep track of missing and non-missing values

Issues are between 10%-50%




Variable
Percent of missing/Outliers is very less (<10%)
Replace the missing with mean or median for continuous variables.
Impute with mean or median.
Replace the missing with most probable category/class
Look at the cross tab of variable vs target.
Impute based on target.
Percent of missing/Outliers is Medium(10% to 50%)
Create a new variable missing & Non Missing Category
Replace the missing values in the original variables with Mean or median
For and categorical variables, rename the missing class as "Missing"
For numeric discrete class, create a new variable and replace missing values with mean or median
Percent of missing/Outliers is High(>50%)
Create a new indicator variable missing and non-missing. Drop the original variable
Name the missing category as missing, club rest al categories as "Non-missing"


Variable
Percent of missing/Outliers is Medium(10% to 50%)
Create a new variable missing & Non Missing Category
Replace the missing values in the original variables with Mean or median

LAB: Monthly Income

- Find the missing value percentage in monthly income
- Create an indicator variable for missing and non-missing
- Replace the missing values with median

LAB: Monthly Income

```
loans['MonthlyIncome'].isnull().sum()
loans['MonthlyIncome'].isnull().sum()/len(loans)

#Flag variable
loans['MonthlyIncome_ind']=1
loans['MonthlyIncome_ind'][loans['MonthlyIncome'].isnull()]=0
loans['MonthlyIncome_ind'].value_counts()


#Imputation with median
loans['MonthlyIncome_new']=loans['MonthlyIncome']
loans['MonthlyIncome_new'][loans['MonthlyIncome'].isnull()]=loans['Monthly
Income'].median()
round(loans['MonthlyIncome_new'].describe())
```

Data Cleaning Other Variables

Remaining Variables Imputation

- Debt Ratio: Imputation like monthly income
- NumberOfOpenCreditLinesAndLoans : Imputation
- NumberOfTimes90DaysLate: Imputation similar to NumberOfTime30_59DaysPastDueNotW
- NumberRealEstateLoansOrLines: : Imputation
- NumberOfTime60_89DaysPastDueNotW: Imputation similar to NumberOfTime30_59DaysPastDueNotW
- NumberOfDependents: Impute based on target variable

Step by Step Process of Data Cleaning

1. Import the data and get the metadata using info() – make a note of missing values and basic issues
2. List down categorical/ discrete and Continuous variables
3. Create frequency distributions and bar charts for the categorical variables. Identify the issues
4. Create percentile distributions and box plots for the continuous variables. Identify the outliers
5. Clean the data for continuous variable
    1. <10% issues - impute
    2. >50% issues – Create a flag variable
    3. 10% to 50% - Both flag and imputation
6. Clean the data for categorical variable
    1. impute based on the target variable

Conclusion

Conclusion

- Data cleaning is as important as data analysis
- Sometimes 80% of the overall project time is spent on data cleaning
- Data cleaning needs patience, we need to clean for each individual variable
- Apart from suggested methods, there are many heuristic ways of cleaning the data