# Data Analysis Course

## Preparing data for analysis(Version-1)

Venkat Reddy

# Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

# Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

# Contents

- Need of data exploration
- Data Exploration
- Data Validation
- Data Sanitization
    - Missing Value Treatment
    - Outlier Treatment Identification & Treatment

# What is the need of data sanitization

- Download age vs weight data from here
- Fit a simple liner regression line
- As age increases what happens to weight?
- Is the inference accurate?

Age_weight_data

# Remember…

Data in the real world is dirty

Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”

- Incomplete data may come from
  - “Not applicable” data value when collected.
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems

Noisy: Containing errors or outliers. e.g., Salary=“-10”

- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission

Inconsistent: containing discrepancies in codes or names

- e.g., Age=“42” Birthday=“03/07/1997”, Was rating “1,2,3”, now rating “A, B, C”,e.g., discrepancy between duplicate records
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

# Lab: About the data, background and Business Objective:

- Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.
- Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This competition requires participants to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years. The objective is to build a model that borrowers can use to help make the best financial decisions.
- Historical data are provided on 250,000 borrowers.

| Variable Name | Description | Type |
|---|---|---|
| **SeriousDlqin2yrs** | Person experienced 90 days past due delinquency or worse | Y/N |
| **RevolvingUtilizationOfUnsecuredLines** | Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits | percentage |
| **age** | Age of borrower in years | integer |
| **NumberOfTime30-59DaysPastDueNotWorse** | Number of times borrower has been 30-59 days past due but no worse in the last 2 years. | integer |
| **DebtRatio** | Monthly debt payments, alimony,living costs divided by monthy gross income | percentage |
| **MonthlyIncome** | Monthly income | real |
| **NumberOfOpenCreditLinesAndLoans** | Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards) | integer |
| **NumberOfTimes90DaysLate** | Number of times borrower has been 90 days or more past due. | integer |
| **NumberRealEstateLoansOrLines** | Number of mortgage and real estate loans including home equity lines of credit | integer |
| **NumberOfTime60-89DaysPastDueNotWorse** | Number of times borrower has been 60-89 days past due but no worse in the last 2 years. | integer |
| **NumberOfDependents** | Number of dependents in family excluding themselves (spouse, children etc.) | integer |

Data Analysis Course
Venkat Reddy

# Basic contents of the data

- What are total number of observations
- What are total number of fields
- Each field name, Field type, Length of field
- Format of field, Label

**Basic Contents –Check points**

- Are all variables as expected (variables names)
- 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)
- Have labels been provided and are sensible

If anything suspicious we can further investigate it and correct accordingly

# Proc Contents-SAS

- SAS code : proc contents data=<<data name>>; run;
- Useful options :
  - **Short** – Outputs the list of variables in a row by row format.

    Code  : proc contents data=test **short**; run;
  - **Out=*filename*** - Creates a data set wherein each observation  is a variable

# Snapshot of the data

**Data Snapshot, if possible**

- Printing the first few observations  all fields in the data set .It helps in better understanding of the variable by looking at it’s assigned values.

**Checkpoints for data snapshot output:**

1. Do we have any unique identifier?  Is the unique identifier getting repeated in different records?
2. Do the text variables have meaningful data?(If text variables have absurd data as ‘&^%*HF’ then either the variable is meaningless or the variable has become corrupt or wasn’t properly created.)
3. Are there some coded values in the data?(if for a known variable say State we have category codes like 1-52 then we need definition of how they are coded.)
4. Do all the variables appear to have data? (In case variables are not populated with non missing meaningful value it would show in print. We can further investigates using means statistics.)

# Proc print in SAS

SAS code : **proc print** data=<<data set>>; run;

- Useful options :

  **proc print data=<<data set>> label noobs heading=vertical;**

  **var <<variable-list>>; by var1; run;**

  - Label:The label option uses variable labels as column headings rather than variable names (the default).
  - Obs : Restricts the number of observations in the output
  - Nobs: It omits the OBS column of output.
  - Heading=vertical: It prints the column headings vertically. *This is useful when the names are long but the values of the variable are short.*
  - Var: Specifies the variables to be listed and the order in which they will appear.
  - By: By statement produces output grouped by values of the mentioned variables

# Lab: Data exploration & validation

- Import Data_explore.txt into sas
- What are basic contents of the data
- Verify the check list
- Any suspicious variables?
- What is var1?
- Are all the variable names correct?
- Print the first 10 observations
  - Do we have any unique identifier?
  - Do the text variables have meaningful data?
  - Are there some coded values in the data?
  - Do all the variables appear to have data

# Categorical field frequencies

- 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

**Checkpoints for looking frequency table**

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.

# Proc Freq in SAS

- SAS code: **Proc FREQ** data =<dataset > <options> ;
  TABLES requests < / options > ; // Gives Frequency Count or Cross Tab
  BY <varl> ; // Grouping output based on varl
  WEIGHT variable < / option > ; //Specifying Weight (if applicable)
  OUTPUT < OUT=SAS-data-set > options ; //Output results to another data set run;
- Useful options :
  - **Order=Freq** - sorts by descending frequency count (default is the unformatted value). Ex: proc freq data=test order=freq; tables X1-X5; run;
  - **Nocol/Norow/Nopercent** - suppresses printing of column, row and cell percentages respectively of a cross tab. Ex : proc freq data=test; tables AGE*bad**/nocol norow nopercent missing**; run;
  - **Missing**- interprets missing values as non-missing and includes them in % and statistics calculations ex : proc freq data=test; tables CHANNEL* BAD **/missing**; run;
  - **Chisq** - performs several chi-square tests. Ex: proc freq data=test; tables channel*bad/chisq; run;

# Lab: Frequencies

- 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?

# Descriptive Statistics for continuous fields

- Distribution of numeric variables by calculating
  - N – Count of non missing observations
  - Nmiss – Count of Missing observations
  - Min, Max, Median, Mean
  - Quartile numbers & percentiles– P1, p5,p10,q1(p25),q3(p75), p90,p99
  - Stddev
  - Var
  - Skewness
  - Kurtosis

# Descriptive Statistics Checkpoints

- 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?

# Proc Means in SAS

- Proc means data=<data set> < options>;

  ```
  Var <variable list >;
  Run;
  ```
- If variable list is not mentioned it gives results across all numeric variables
- If options are not specified by default it gives stats like – n , min, max, mean and stddev.
- Useful Options :
  - By : Calculates statistics based on grouping across specified variable;

```
Proc means data=check n nmiss min max;
var age ;
class channel;
run;
```

# Proc Univariate in SAS

- SAS Code : **PROC UNIVARIATE** data=<dataset>;

  **VAR** *variable(s)*; run;
- Useful options :
  - **PROC UNIVARIATE** data=<dataset> plot normal;

    **HISTOGRAM** <*variable(s)*> </ *option(s)*>;

    **By variable;**

    **VAR** *variable(s)*; run;
    - Normal option produces the tests of normality ;
    - Plot option produces the 3 plots of data(stem and leaf plot, box plot, normal probability plot
    - By option is used for giving outputs separated by categories
    - Histogram option gives the distribution of variable in a histogram

# General Checks

- **Mean=Median?**
- **Counted proportion data.** If data consists of counted proportions, e.g. number of individuals responding out of total number of individuals,
- **Polytomous data**. If data consists of numbers falling into a number of mutually exclusive classes, do not reduce to proportions or percentages beforehand, but enter the integer counts
- **Data Sufficiency:** Data Sufficiency involves ensuring that the data has the required attributes to make the prediction as stated by objective
  - **Eg1:** To build a model to predict fraud, the given data doesn’t have any key for identifying fraud accounts or those identifiers are erased than there are no accounts which we can identify as ‘bad’ and build a model to predict the same.
  - **Eg2:** If we are building a response model specifically for  internet channel for a “airline card”. Then data should have a identifier for ‘channel of acquisition’ to identify the right data base on which to build the model

# Lab: Data exploration & validation

- Find N, Average, sd, minimum & maximum
- Is N same for all the variables?
- Any variables with unusual min & max?
- Identify list of suspicious variables
- Find below statistics for all the doubtful variables
    - N,Mean,Median,Mode
    - Std Deviation
    - Skewness
    - Variance
    - Kurtosis
    - Interquartile Range
    - Quantiles 100% Max, 99%,,95%,,90%, 75% Q3,50% Median,25% Q1,10%,5%,1%,0% Min
- See the variable definitions and possible values
- Identify variables with missing values, default values & outliers

# Now what…?

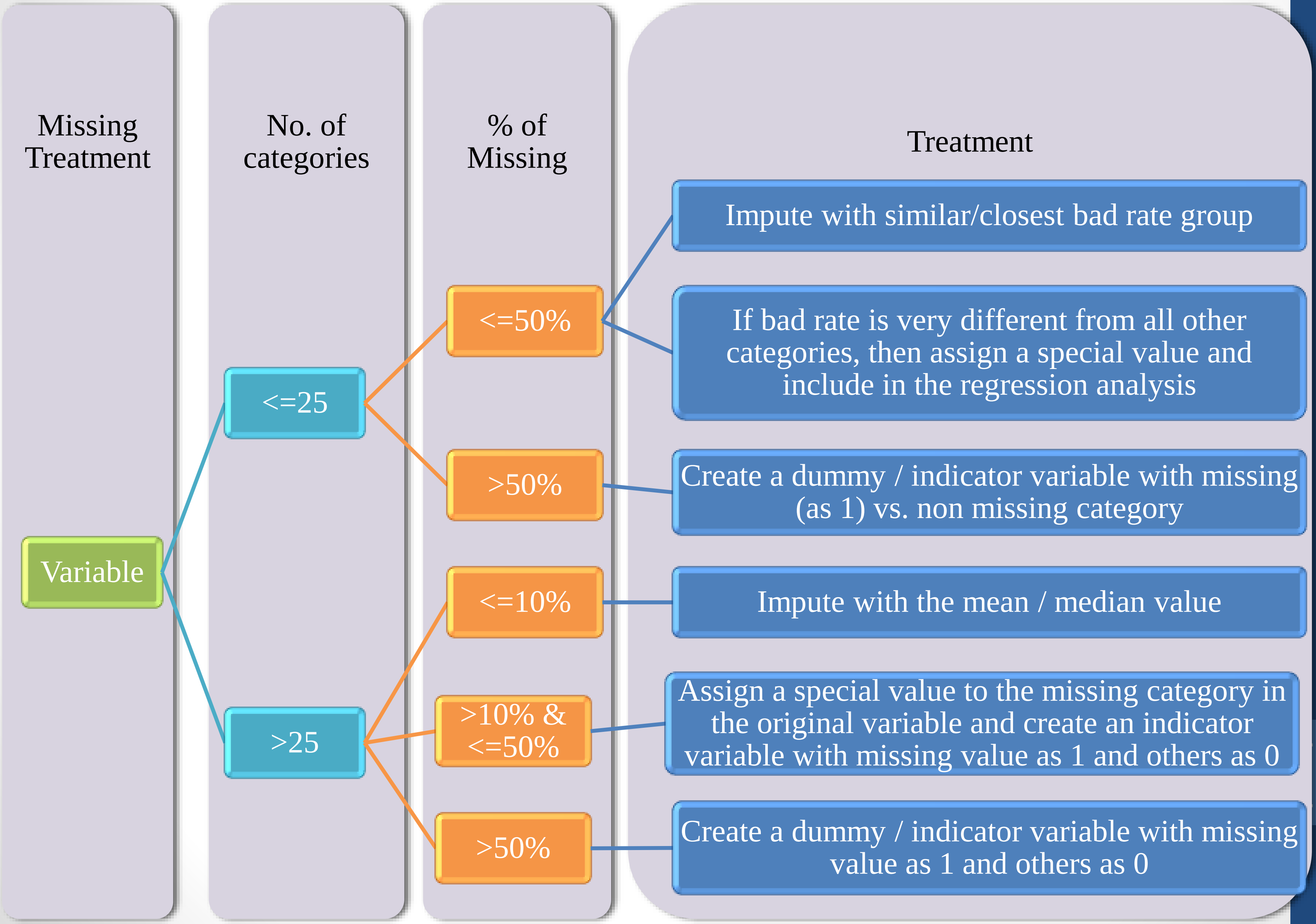
- Some variables contain outliers
- Some variables have default values
- Some variables have missing values

- Shall we delete them and go ahead with our analysis?

# Missing Values

- 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

# Missing Value & outlier Treatment



Missing Treatment

No. of categories

% of Missing

Treatment

Variable

<=25

<=50%

Impute with similar/closest bad rate group

If bad rate is very different from all other categories, then assign a special value and include in the regression analysis

>50%

Create a dummy / indicator variable with missing (as 1) vs. non missing category

>25

<=10%

Impute with the mean / median value

>10% & <=50%

Assign a special value to the missing category in the original variable and create an indicator variable with missing value as 1 and others as 0

>50%

Create a dummy / indicator variable with missing value as 1 and others as 0

# Missing Value Imputation1

- **Missing Value Imputation – 1:Standalone imputation**
  - Mean, median, other point estimates
  - Assume: Distribution of the missing values is the same as the non-missing values.
  - Does not take into account inter-relationships
  - Introduces bias
  - Convenient, easy to implement

# Missing Value Imputation2

**Missing Value Imputation - 2**

- Better imputation - use attribute relationships
- Assume : all prior attributes are populated
- That is, monotonicity in missing values.
- Two techniques
  - Regression (parametric),
  - Propensity score (nonparametric)
- Regression method
  - Use linear regression, sweep left-to-right

    $X3=a+b*X2+c*X1;$

    $X4=d+e*X3+f*X2+g*X1$, and so on
  - X3 in the second equation is estimated from the first equation if it is missing

| X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|
| 11 | 20 | 6.5 | 4 | . |
| 12.1 | 18 | 8 | 2 | . |
| 11.9 | 22 | 4.2 | . | . |
| 10.9 | 15 | . | . | . |

# Missing Value Imputation3

- Propensity Scores (nonparametric)
  - Let $Y_j$=1 if $X_j$ is missing, 0 otherwise
  - Estimate P($Y_j$ =1) based on $X_1$ through $X_{(j-1)}$ using logistic regression
  - Group by propensity score P($Y_j$ =1)
  - Within each group, estimate missing $X_j$s from known $X_j$s using approximate Bayesian bootstrap.
  - Repeat until all attributes are populated.
- Arbitrary missing pattern
  - Markov Chain Monte Carlo (MCMC)
  - Assume data is multivariate Normal, with parameter Q
  - (1) Simulate missing X, given Q estimated from observed X ; (2) Re-compute Q using filled in X
  - Repeat until stable.
  - Expensive: Used most often to induce monotonicity

Note that imputed values are useful in aggregates but can't be trusted individually

# Default Values Treatment

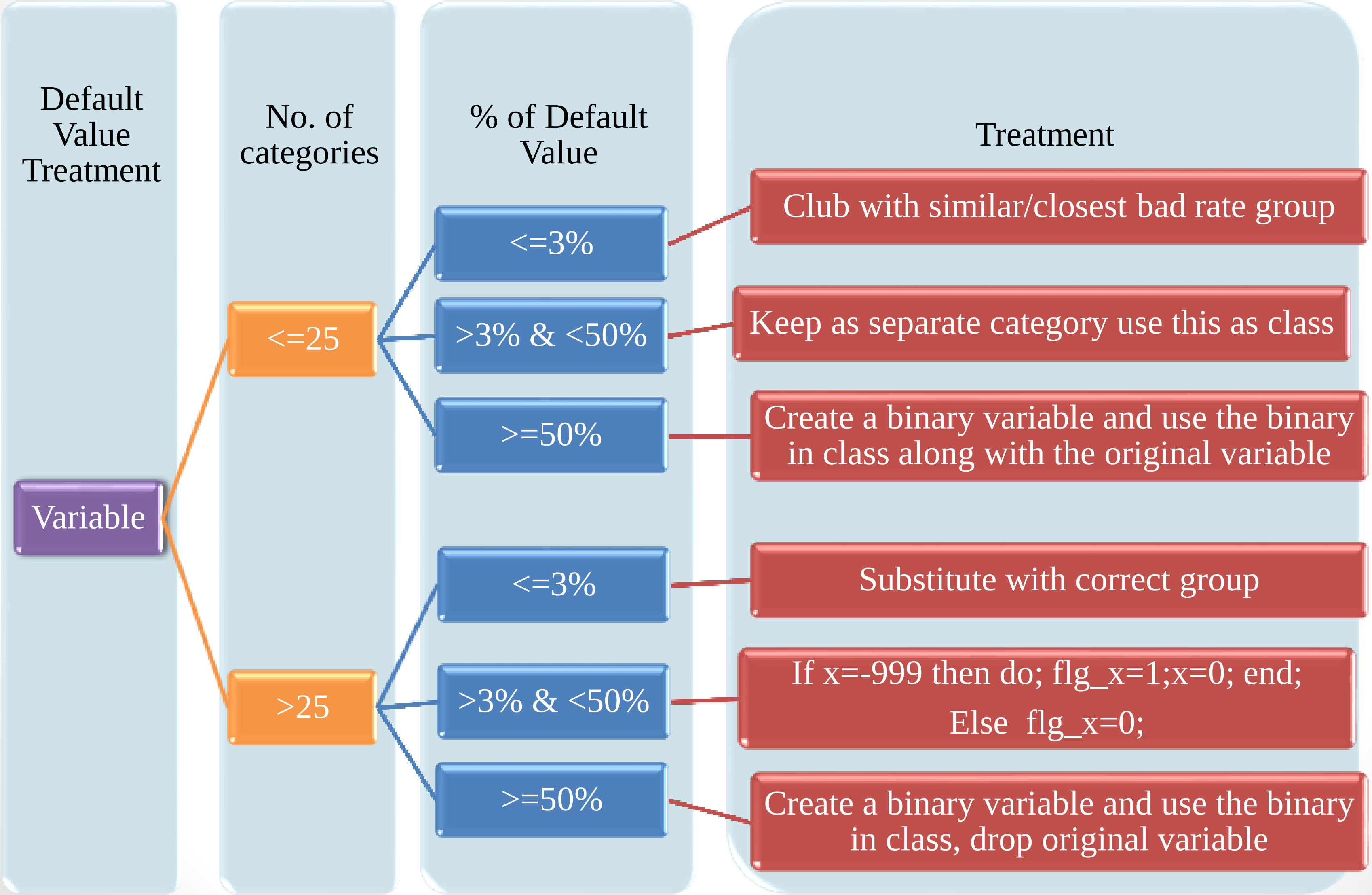
- Special or default values are values like 999 or 999999 which fall outside the normal range of data .

- For instance a no. of bankcards variable usually has values from 0 to 100 but 999 values in the data represent the population which does not have any tradelines. Including them in the regression as 999 would skew the regression results, hence we need to treat them accordingly.

- Special or default values also should be treated as we are treating the missing values depending upon the no. of categories and the % of default value.

- They can also be taken care of by capping or flooring them to a realistic value / where the trend is being maintained ( especially if the % of default value is very less).

# Default Values Treatment



Default Value Treatment

No. of categories

% of Default Value

Treatment

Variable

<=25

- <=3% — Club with similar/closest bad rate group
- >3% & <50% — Keep as separate category use this as class
- >=50% — Create a binary variable and use the binary in class along with the original variable

>25

- <=3% — Substitute with correct group
- >3% & <50% — If x=-999 then do; flg_x=1;x=0; end;  
  Else  flg_x=0;
- >=50% — Create a binary variable and use the binary in class, drop original variable

# Lab: Missing values & Default values

- In the sample data, what all fields need missing value treatment?
- Conduct the missing value treatment
- Identify fields with outlier, conduct outlier treatment

# Outlier Treatment

- Flooring
- Capping
- Treat as a separate segment

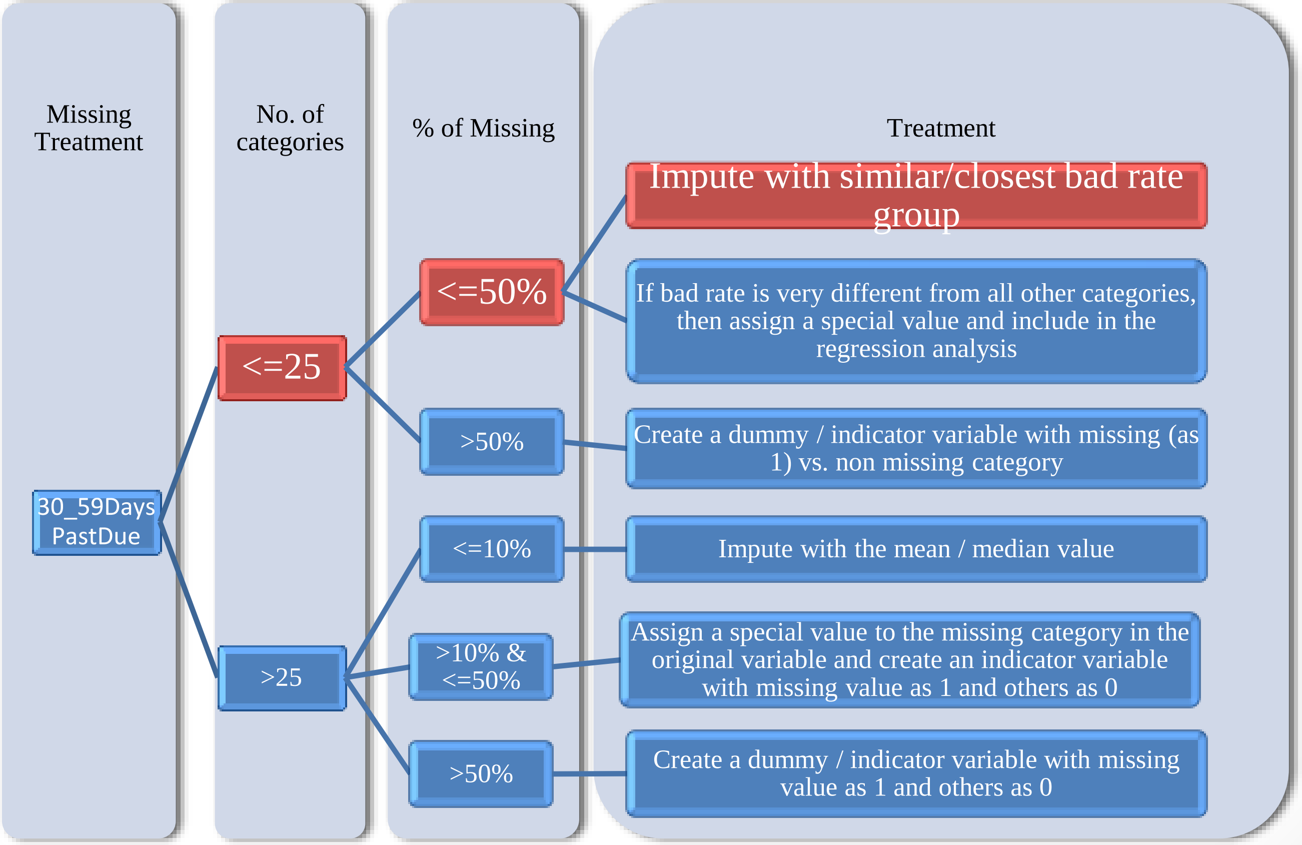
# Lab: Data Cleaning step by step

- **Var-1:** Change the variable name to sr_no
- **SeriousDlqin2yrs :** Only training data has data objective variable test data doesn’t have objective variable in it. Subset training data & test from overall data
- **Age:** If age <21 make it 21

# Lab : Data Cleaning step by step

## NumberOfTime30_59DaysPastDueNotW

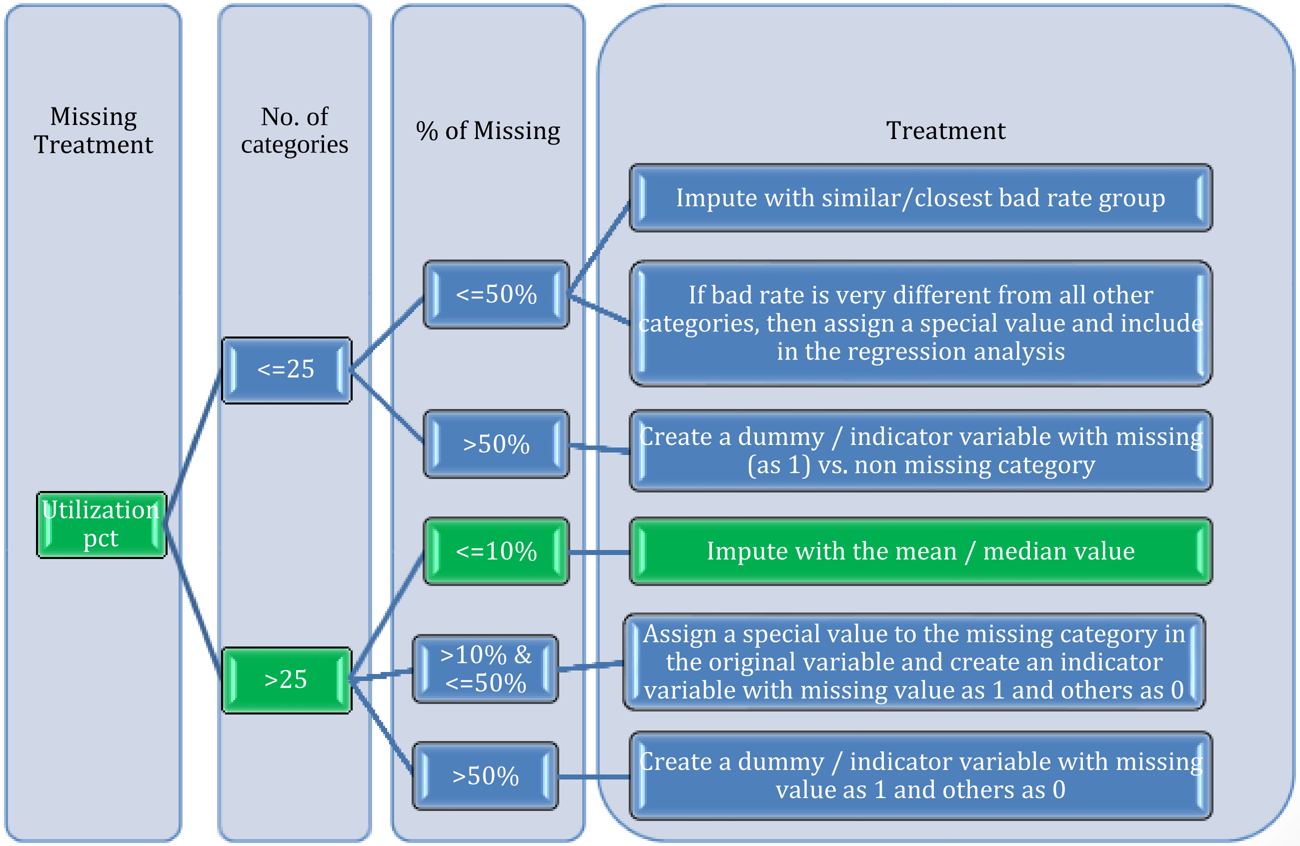
- Find bad rate in each category of  this variable
- Replace 96 with _____? Replace 98 with_____?



Missing Treatment
No. of categories
% of Missing
Treatment
30_59Days PastDue
<=25
>25
<=50%
>50%
<=10%
>10% & <=50%
>50%
Impute with similar/closest bad rate group
If bad rate is very different from all other categories, then assign a special value and include in the regression analysis
Create a dummy / indicator variable with missing (as 1) vs. non missing category
Impute with the mean / median value
Assign a special value to the missing category in the original variable and create an indicator variable with missing value as 1 and others as 0
Create a dummy / indicator variable with missing value as 1 and others as 0

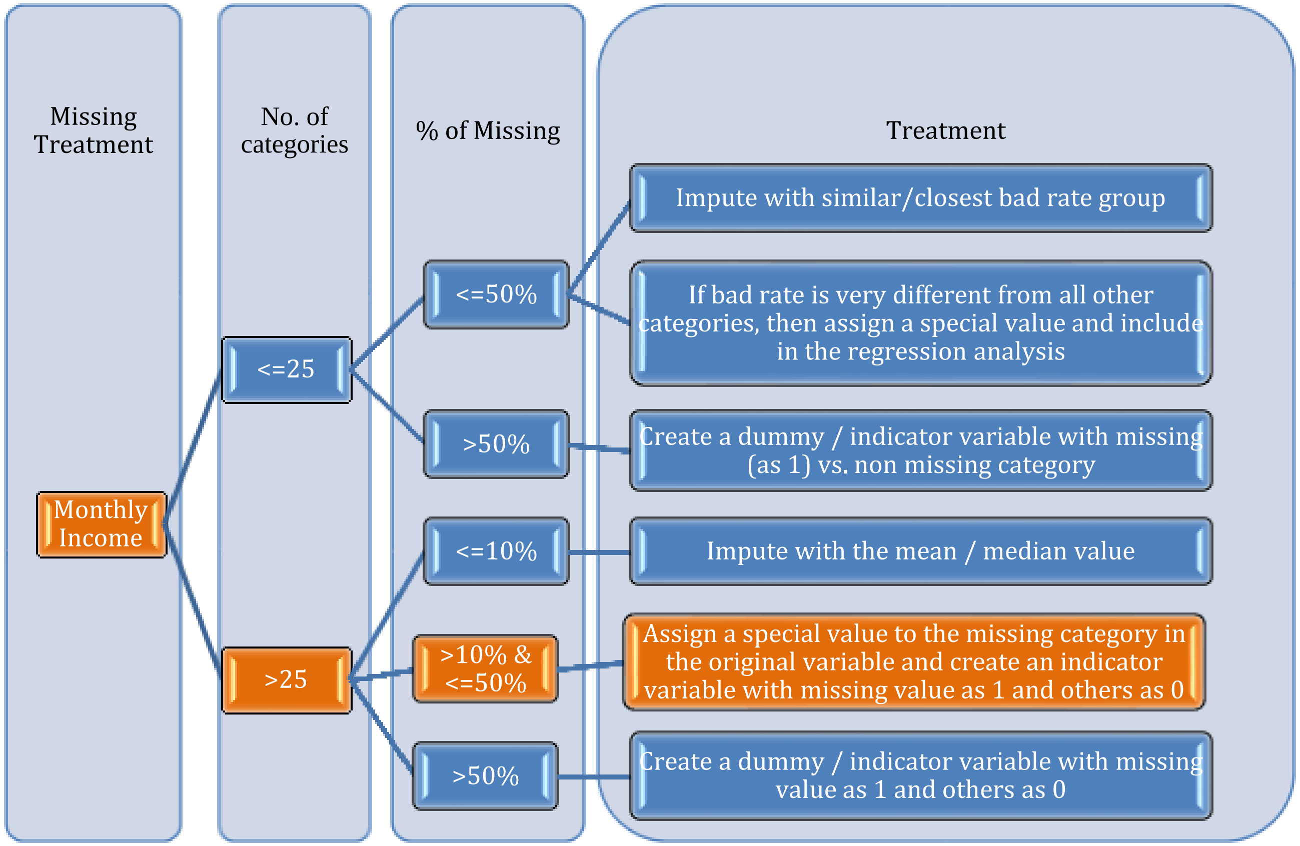
# Lab :Data Cleaning step by step

- **RevolvingUtilizationOfUnsecuredL:** What type of variable is this? What are the possible values?
- Replace anything more than 1 with_____?


Missing Treatment
No. of categories
% of Missing
Treatment
Utilization pct
<=25
>25
<=50%
>50%
<=10%
>10% & <=50%
>50%
Impute with similar/closest bad rate group
If bad rate is very different from all other categories, then assign a special value and include in the regression analysis
Create a dummy / indicator variable with missing (as 1) vs. non missing category
Impute with the mean / median value
Assign a special value to the missing category in the original variable and create an indicator variable with missing value as 1 and others as 0
Create a dummy / indicator variable with missing value as 1 and others as 0

# Lab :Data Cleaning step by step

- Monthly Income



Missing Treatment

No. of categories

% of Missing

Treatment

Monthly Income

<=25

<=50%

Impute with similar/closest bad rate group

If bad rate is very different from all other categories, then assign a special value and include in the regression analysis

>50%

Create a dummy / indicator variable with missing (as 1) vs. non missing category

>25

<=10%

Impute with the mean / median value

>10% & <=50%

Assign a special value to the missing category in the original variable and create an indicator variable with missing value as 1 and others as 0

>50%

Create a dummy / indicator variable with missing value as 1 and others as 0

# Lab :Data Cleaning step by step

- **Debt Ratio**: Similar Imputation
- **NumberOfOpenCreditLinesAndLoans** : No clear evidence
- **NumberOfTimes90DaysLate:** Imputation similar to NumberOfTime30_59DaysPastDueNotW
- **NumberRealEstateLoansOrLines:** : No clear evidence
- **NumberOfTime60_89DaysPastDueNotW:** Imputation similar to NumberOfTime30_59DaysPastDueNotW
- **NumberOfDependents**: Impute with equal bad rate

# Variables & Treatment

| Old Var | Type | Treatment | New Var |
|---|---|---|---|
| VAR1 | Num | Nothing | |
| SeriousDlqin2yrs | Num | Nothing | |
| RevolvingUtilizationOfUnsecuredL | Num | Impute with the mean | Util |
| age | Num | flooring | age1 |
| NumberOfTime30_59DaysPastDueNotW | Num | Impute with the mean | NumberOfTime30_59DaysPastDue1 |
| DebtRatio | Num | Impute with the median | DebtRatio1 |
| MonthlyIncome | Char | Convert to num & create a dummy var | ind_MonthlyIncome, MonthlyIncome1 |
| NumberOfOpenCreditLinesAndLoans | Num | Impute with median | num_open_lines |
| NumberOfTimes90DaysLate | Num | Imputing & capping | delq_90 |
| NumberRealEstateLoansOrLines | Num | Capping | num_loans |
| NumberOfTime60_89DaysPastDueNotW | Num | Capping & Imputing | delq_60to89 |
| NumberOfDependents | Char | | |
| obs_type | Char | Subset training data | Obs_type |

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

www.trendwiseanalytics.com/venkat

+91 9886 768879