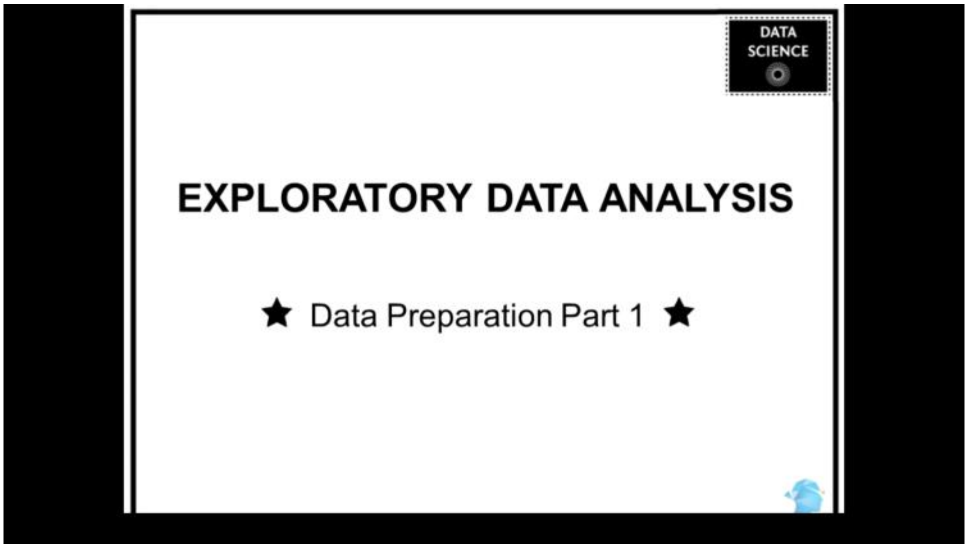
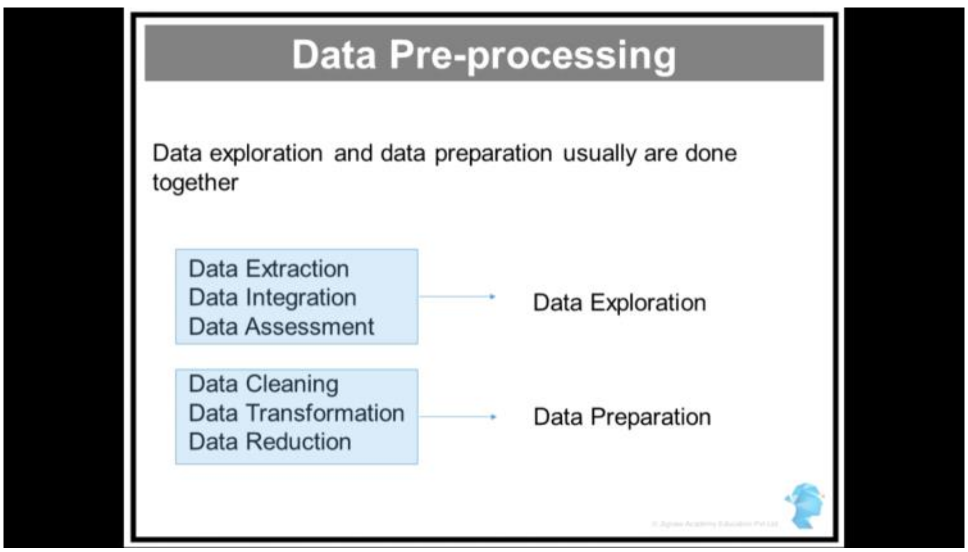
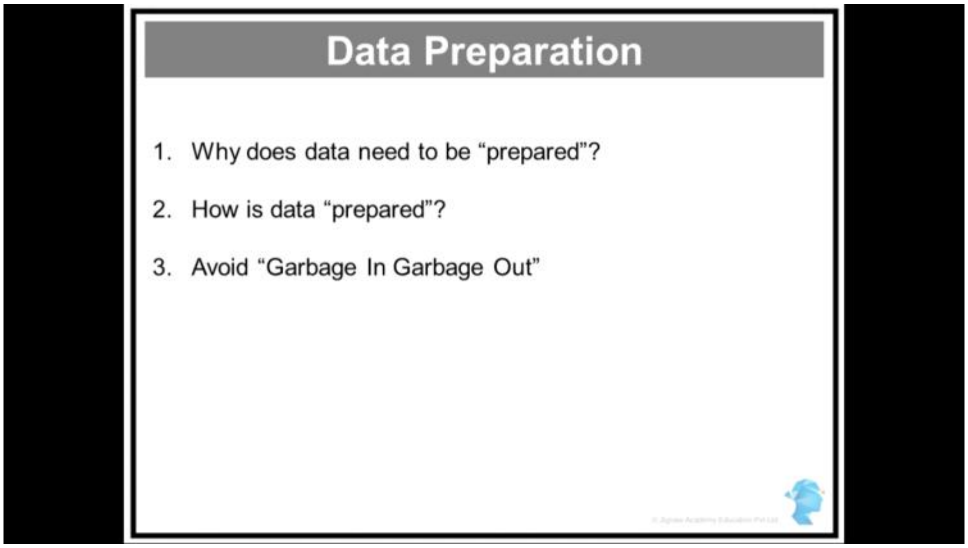
Data Pre-Processing

Saturday, September 24, 2016 5:30 AM

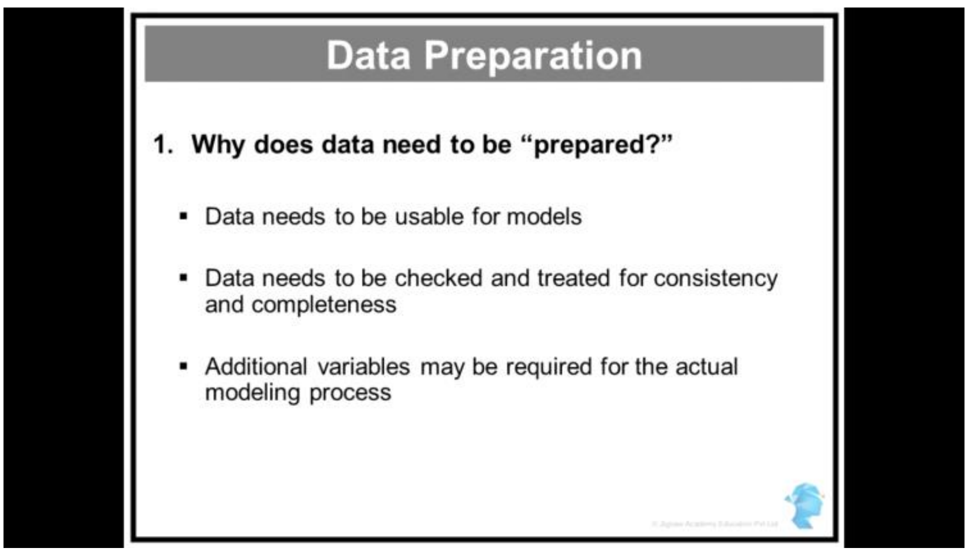
Data Cleanup

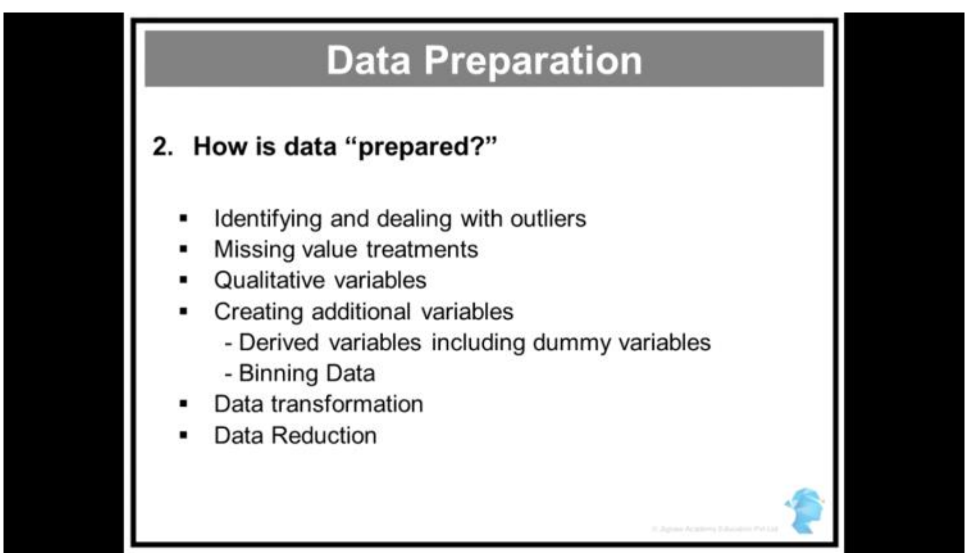






Raw data isn't ready for model analysis so we need to prep it.
Checked for consistency annd completeness.





Data Preparation

Data Cleaning

Raw data is rarely "ready to use"

Most common issues:

1. Outliers

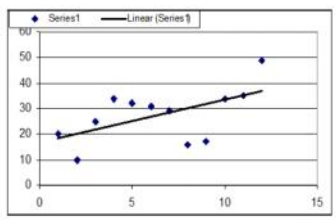
2. Missing Values

Outlier Detection

What is an outlier?

Definition : An outlier is a value of a variable that appears to differ significantly from the rest of the values of the variable

- Key Terms: "Differ", "Significantly"
- In this chart, how many outliers are present?
- Are extreme values outliers?

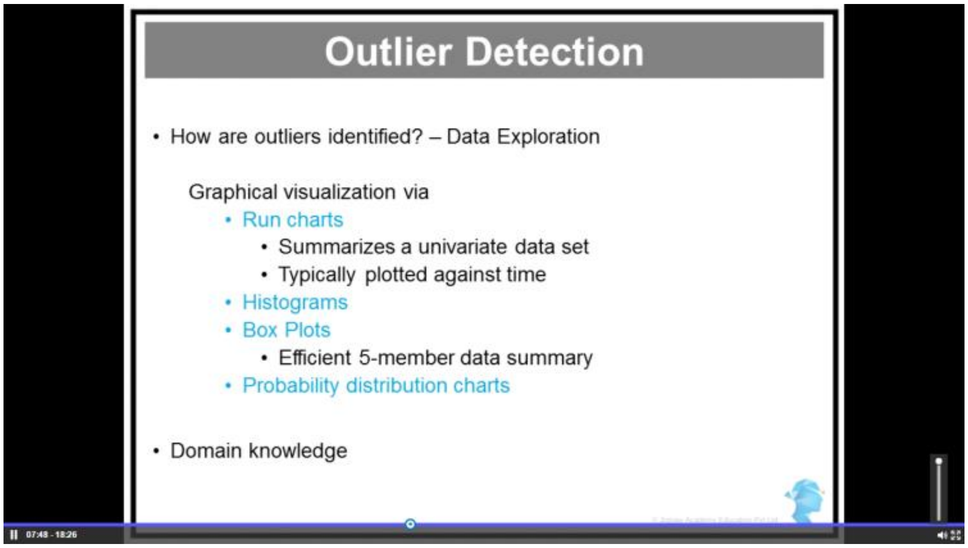


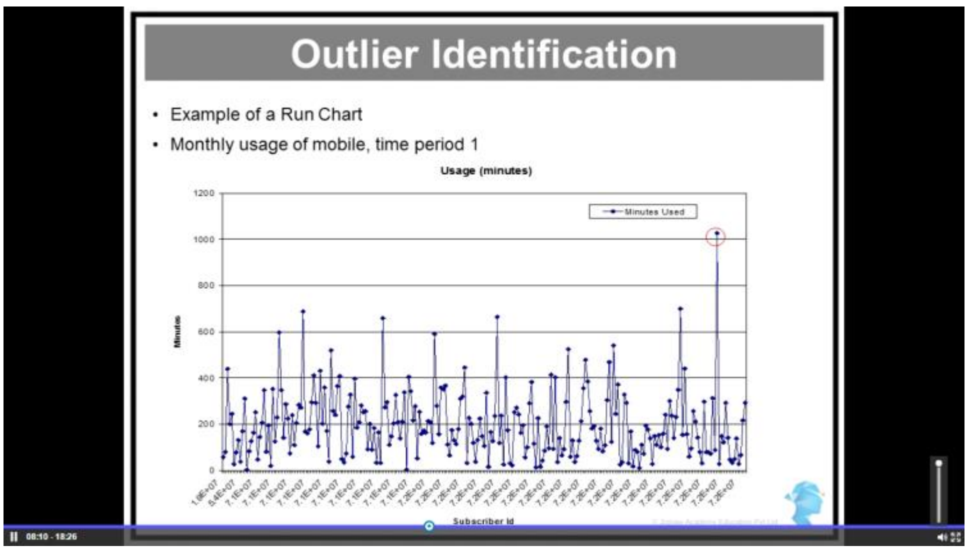
Outlier Detection

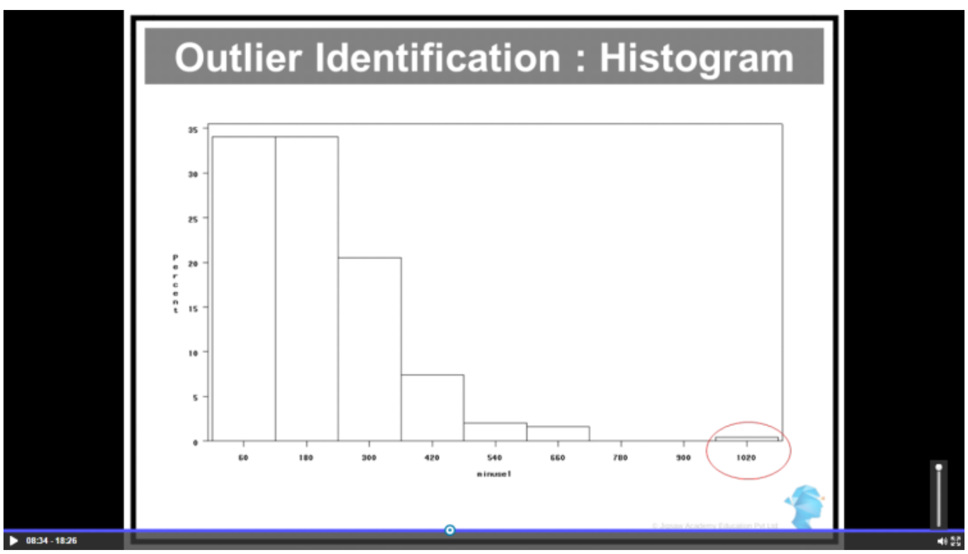
Are Outliers a problem?

- They represent variation in sample – how can that be bad?
- They are surprising or extreme values – how can that be good?
  - Improbable v/s Impossible values
- Are there special circumstances or conditions that produced the outlying observations that may not apply to the problem at hand?

Respondent	Average Shopping Time	Age
A	20	21-25
B	10	21-25
C	25	26-34
D	34	21-25
E	32	34-50
F	31	21-25
G	29	26-34
H	16	21-25
I	17	26-34
L	34	21-25
K	35	50+
L	49	50+

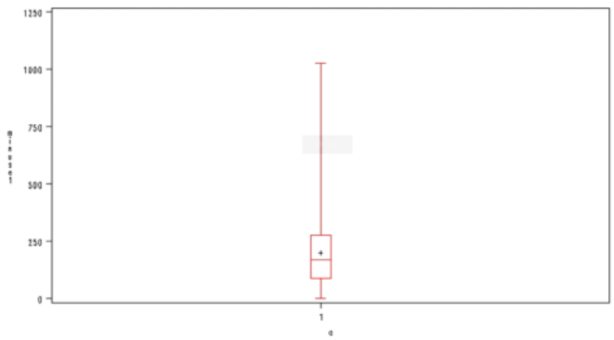

Outlier Detection
• How are outliers identified? – Data Exploration
Graphical visualization via
• Run charts
• Summarizes a univariate data set
• Typically plotted against time
• Histograms
• Box Plots
• Efficient 5-member data summary
• Probability distribution charts
• Domain knowledge
07:48 - 18:26


Outlier Identification
• Example of a Run Chart
• Monthly usage of mobile, time period 1
Usage (minutes)
Minutes Used
Minutes
1200
1000
800
600
400
200
0
Subscriber Id
08:10 - 18:26


Outlier Identification : Histogram
Percent
35
30
25
20
15
10
5
0
60
180
300
420
540
660
780
900
1020
08:34 - 18:26

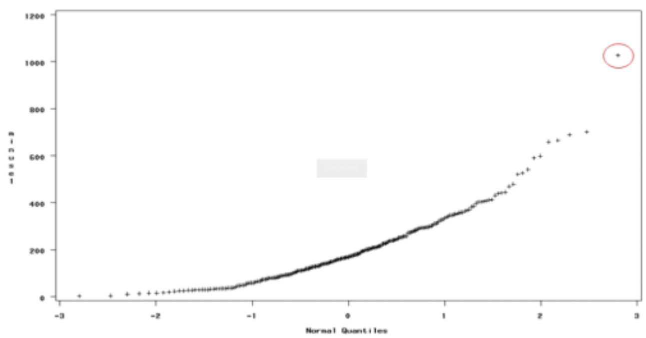
Outlier Detection : Case Study Charts

- Box plot of monthly usage in minutes, time period 1



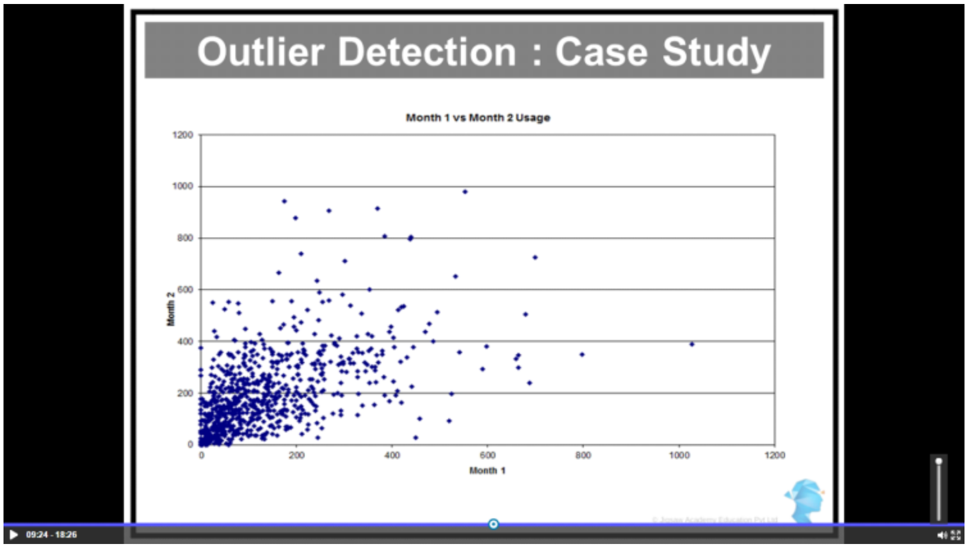
Outlier Identification

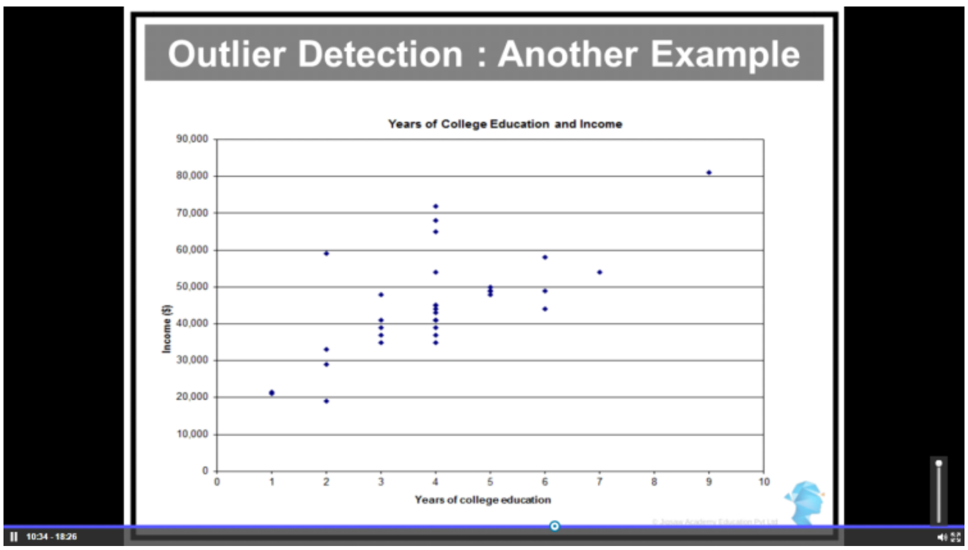
- Normal Probability Plot of monthly usage

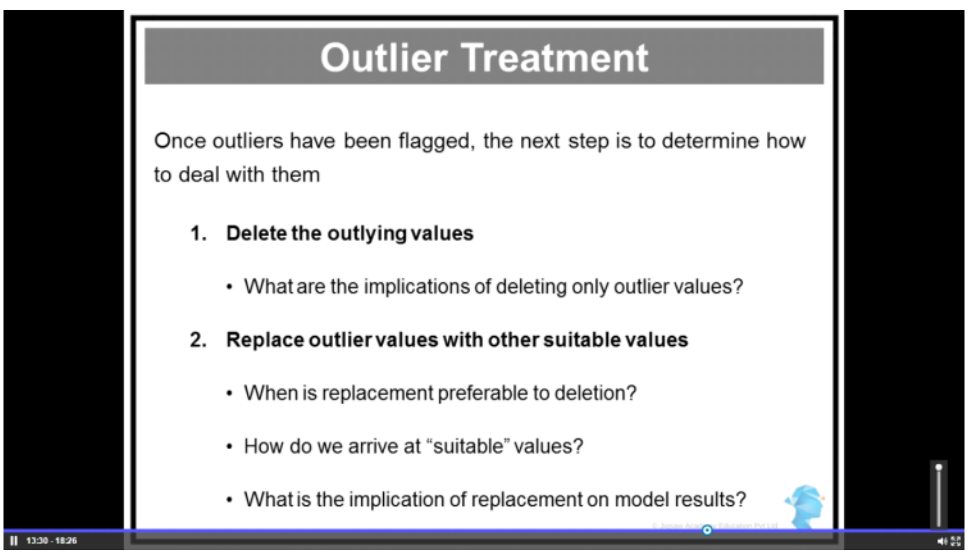


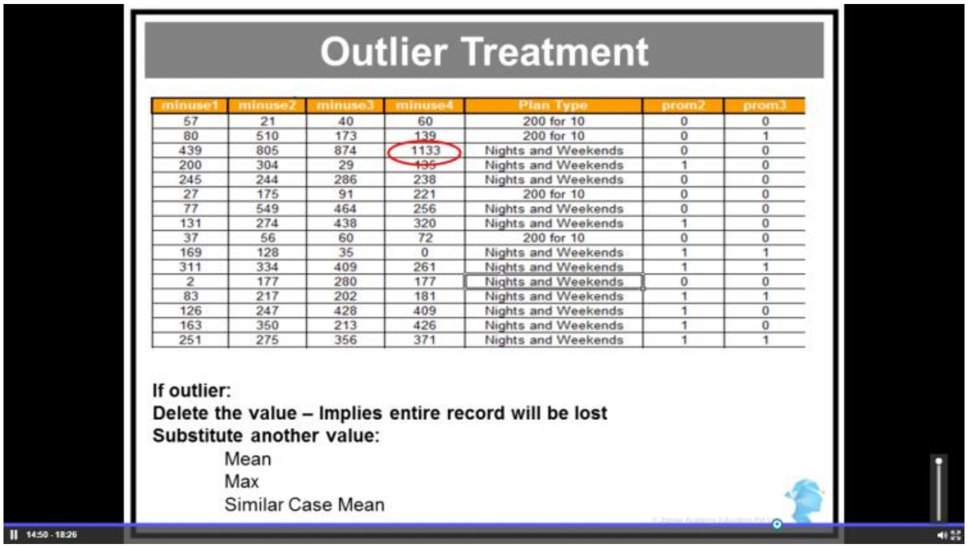
Outlier Detection : Multivariate Approach

- Sometimes looking at single variable distribution in isolation may not be enough to identify outliers
  - Will need to look at pairs of observations (joint distributions)
- Should we look at all possible pairs?
  - In large datasets, with multiple variables, it will not be possible
  - Domain knowledge and problem background will help in determining what pairs to look at?
- Only pairs of variables? What about combinations greater than 2?


Outlier Detection : Case Study
Month 1 vs Month 2 Usage
1200
1000
800
600
400
200
0
Month 2
0
200
400
600
800
1000
1200
Month 1
09:24 - 18:26


Outlier Detection : Another Example
Years of College Education and Income
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Income ($)
0
1
2
3
4
5
6
7
8
9
10
Years of college education
10:34 - 18:26

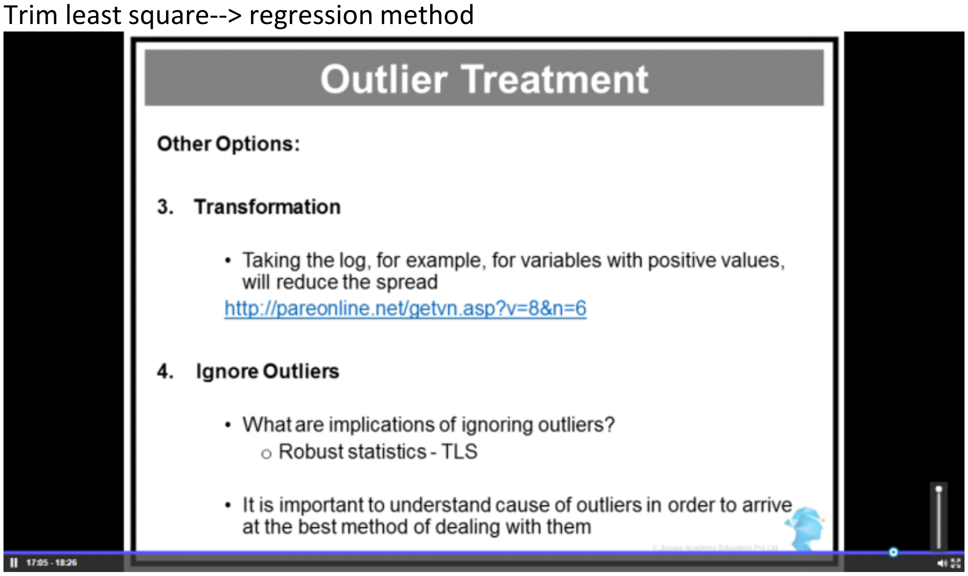

Outlier Treatment
Once outliers have been flagged, the next step is to determine how to deal with them
1. Delete the outlying values
• What are the implications of deleting only outlier values?
2. Replace outlier values with other suitable values
• When is replacement preferable to deletion?
• How do we arrive at "suitable" values?
• What is the implication of replacement on model results?
13:30 - 18:26

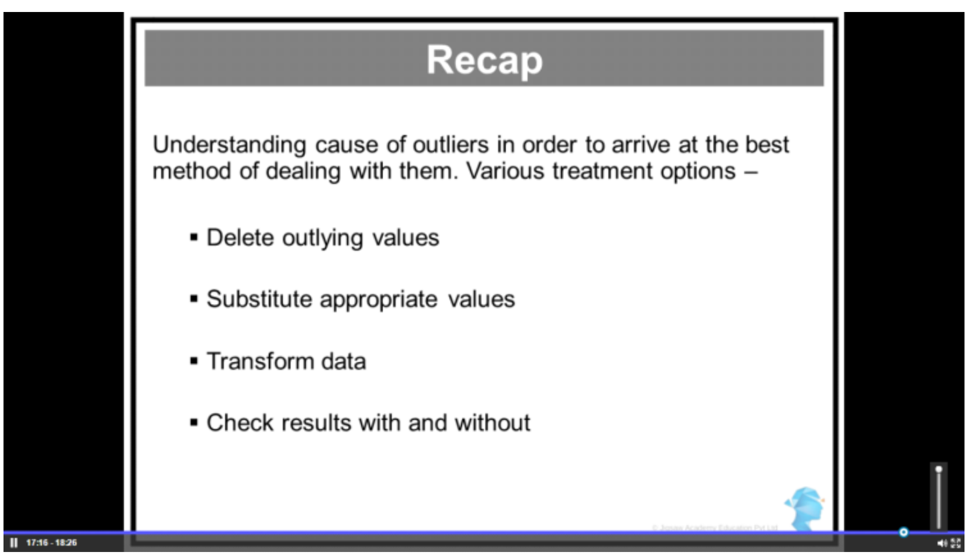


minuse1	minuse2	minuse3	minuse4	Plan Type	prom2	prom3
57	21	40	60	200 for 10	0	0
80	510	173	139	200 for 10	0	1
439	805	874	1133	Nights and Weekends	0	0
200	304	29	135	Nights and Weekends	1	0
245	244	286	238	Nights and Weekends	0	0
27	175	91	221	200 for 10	0	0
77	549	464	256	Nights and Weekends	0	0
131	274	438	320	Nights and Weekends	1	0
37	56	60	72	200 for 10	0	0
169	128	35	0	Nights and Weekends	1	1
311	334	409	261	Nights and Weekends	1	1
2	177	280	177	Nights and Weekends	0	0
83	217	202	181	Nights and Weekends	1	1
126	247	428	409	Nights and Weekends	1	0
163	350	213	426	Nights and Weekends	1	0
251	275	356	371	Nights and Weekends	1	1

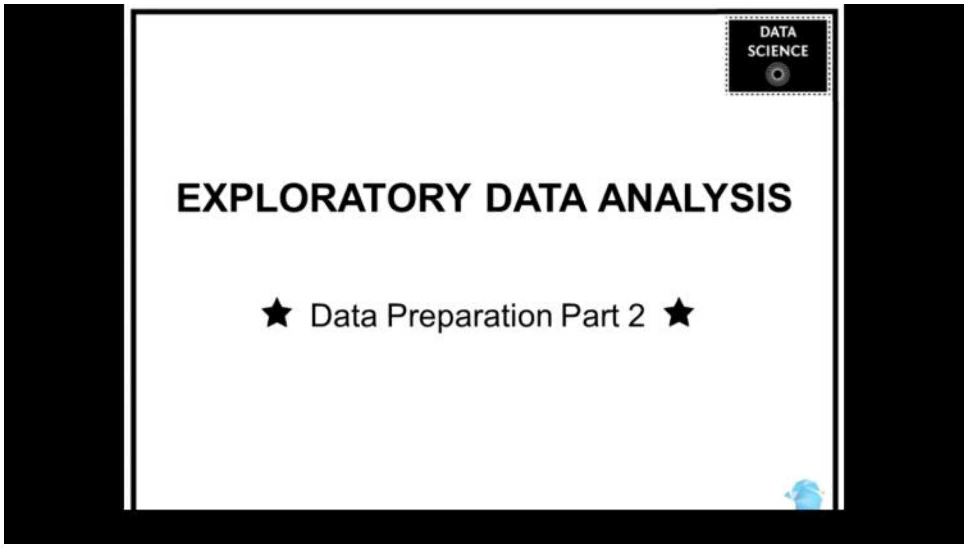
Similar case means --> find datasets that are as simlar as the dataset that we want to substitute . Calculate the means within and replace that datapoint here.

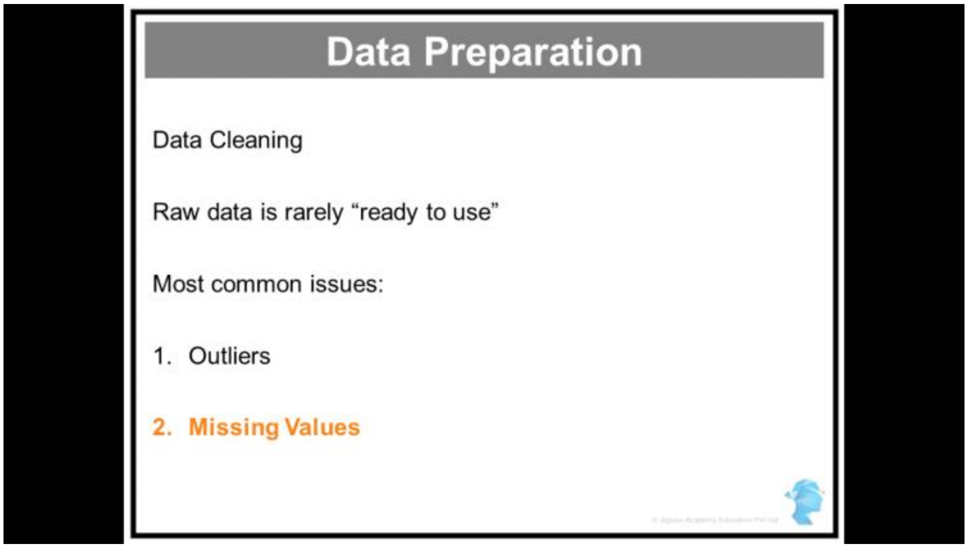
Trim least square--> regression method

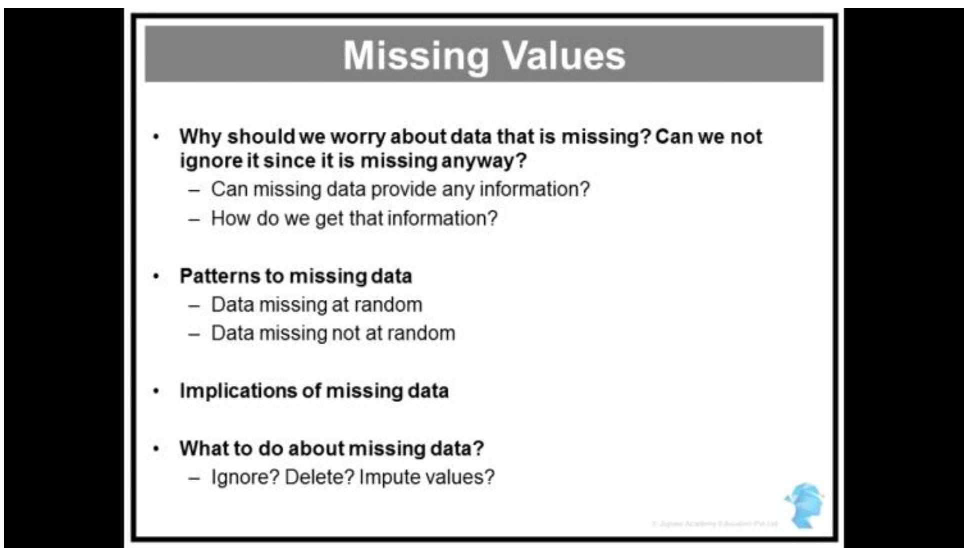




Data Transformation:







Missing Values – Assessment

Prom3	Number of Obs	Variable Name	Number of Obs	Number of Missing Obs
0	9175	sbscrp_id	9175	0
		minuse1	9164	11
		minuse2	9149	26
		minuse3	9164	11
		minuse4	9127	48
		prom2	9175	0
		prom4	9143	32
		prom5	8936	239
		BIRTH_DT	9112	63
		zip_code	9175	0
1	3256	sbscrp_id	3256	0
		minuse1	3253	3
		minuse2	3249	7
		minuse3	3255	1
		minuse4	3239	17
		prom2	3256	0
		prom4	3240	16
		prom5	3194	62
		BIRTH_DT	3231	25
		zip_code	3256	0

Variable	N	N Miss
sbscrp_id	12500	0
minuse1	12485	15
minuse2	12466	34
minuse3	12419	81
minuse4	12366	134
prom2	12499	1
prom3	12431	69
prom4	12383	117
prom5	12130	370
BIRTH_DT	12411	89
zip_code	12500	0

- Missing data for each variable does not seem to be a substantial proportion of available data
- Assess pattern of missing data

Missing Values – Treatment

Delete values with missing data

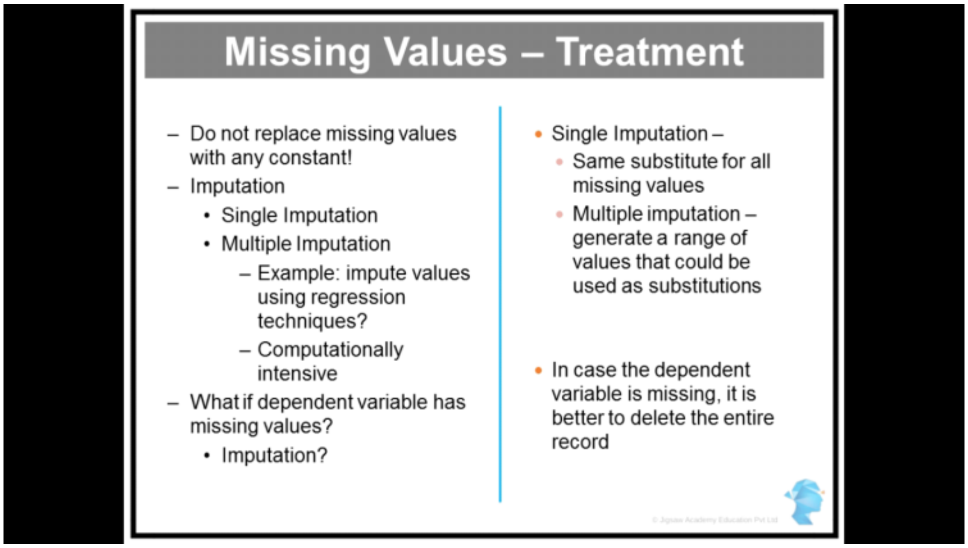
- Since data is missing, eliminate records with missing values
- Because of the multiplicative impact, of there exist a number of variables that have missing values, many records will be lost
- Also, deleting all missing value records may introduce bias
- When dependent variable is missing?

Missing Values – Treatment

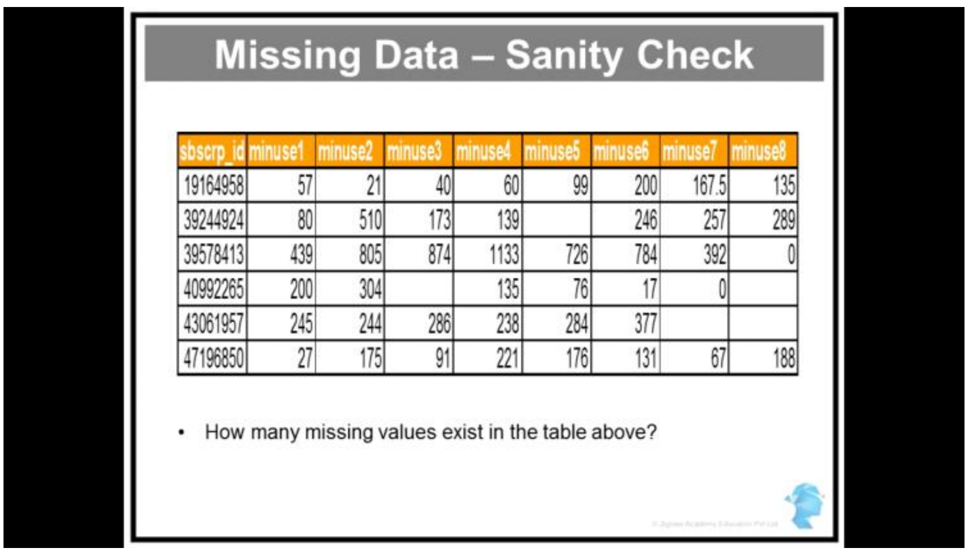
Treat missing values

- Mean substitution
  - Not recommended in general – why?
- Other substitution – Available case
  - Potential substitutes include "exact case", mean of similar cases etc.
  - In this case, how do we identify similar case?
    - Minutes 1 less than 100, Minutes 2 > 500, Minute 3 less than 200, Minute 4 less than 150 etc. – is this a good method?

sbscrp_id	minuse1	minuse2	minuse3	minuse4	minuse5
19164958	57	21	40	60	99
39244924	80	510	173	139	233
39578413	439	805	874	1133	726
40992265	200	304	29	135	76
43061957	245	244	286	238	284
47196850	27	175	91	221	176
51236987	77	549	464	256	287.5
51326773	131	274	438	320	205
54271247	37	56	60	72	77
70765025	169	128	35	0	117
70781923	311	334	409	261	291

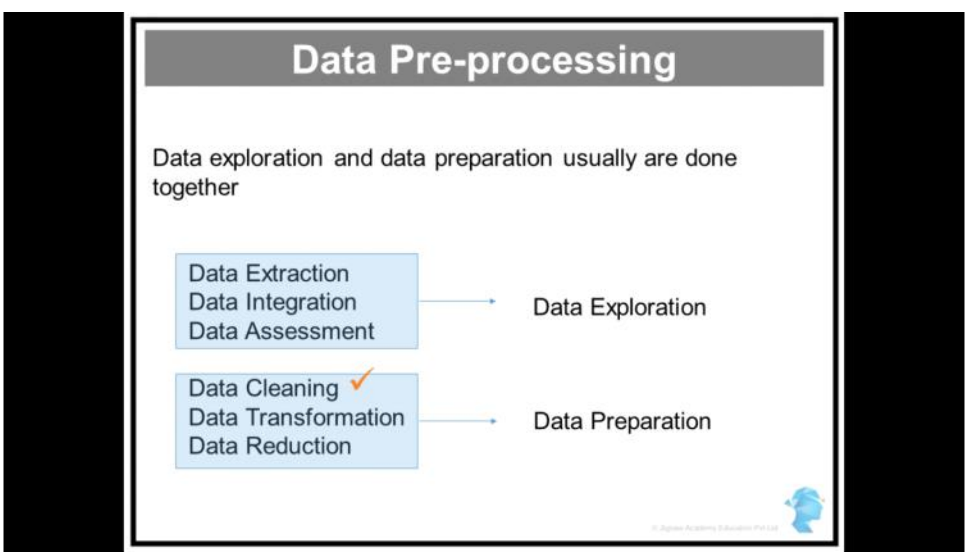


Never impute dependent variables.>>



sbscrp_id	minuse1	minuse2	minuse3	minuse4	minuse5	minuse6	minuse7	minuse8
19164958	57	21	40	60	99	200	167.5	135
39244924	80	510	173	139		246	257	289
39578413	439	805	874	1133	726	784	392	0
40992265	200	304		135	76	17	0	
43061957	245	244	286	238	284	377		
47196850	27	175	91	221	176	131	67	188

Think about missing values logically rather than what is visual>>



Data Preparation

Data Preparation deals with -

- **Qualitative variables**
- Categorical variables
- Derived variables
- Transformed variables

Qualitative Variables

Qualitative variables may not be usable directly in models (need numeric data)

- Examples:
  - Gender: "M", "F"
  - Customer Type: "High Value", "Medium Value", "Low Value"

In order to use in analysis, especially statistical analysis, will need to transform these qualitative values to quantitative values - Recode

Gender – M/F to 0/1.
If gender = "M" then gender = 0; else gender = 1;

Qualitative Variables

Sometimes categories in a qualitative variable are too many

- Example : Profession, Item Purchased
  - substitute a more meaningful value to that variable - grocery v/s non-grocery
  - The substitution obviously needs to add value to the data and help in generating the answer to the problem being investigated

Qualitative Variables

MSRP	Type	city	high	length	width	height	weight	luggage	horse	Cyl	Disp	fuel	AWD	FWD	FOURWD
30880	Luxury	19	29	192.0	70.6	55.5	3510	13.6	260	6	3.2	17.2	0	1	0
20465	Sedan	24	32	186.7	70.1	54.5	2961	14.6	140	4	2.2	14.1	0	1	0
13270	Compact	32	37	174.7	66.7	55.1	2405	12.9	115	4	1.7	13.2	0	1	0
21635	Sedan	20	29	186.3	70.4	55.1	3091	14.6	175	6	3.4	14.1	0	1	0
12482	Compact	32	39	168.1	66.5	52.4	2183	11.5	92	4	1.5	12.4	0	1	0
10480	Compact	34	41	163.2	65.4	59.4	2035	13.6	108	4	1.5	11.9	0	1	0
31845	Hatchback	23	31	159.1	73.1	53.0	2921	13.8	180	4	1.8	14.5	0	1	0
29745	Luxury	19	27	176.7	69.2	53.9	3197	9.5	184	6	2.5	16.6	0	0	0
15675	Compact	24	32	180.9	68.7	53.0	2749	13.2	115	4	2.2	14.1	0	1	0
13330	Compact	25	33	175.2	67.4	52.3	2464	11.8	130	4	2.0	12.8	0	1	0
39647	Convertible	18	27	200.6	75.5	53.6	3814	15.3	275	8	4.6	19.0	0	1	0

- Which is the qualitative variable?
- What would be a way to convert this variable to be used in a model?

Data Preparation

Data Preparation deals with -

- Qualitative variables
- **Categorical variables**
- **Derived variables**
- Transformed variables

Data Preparation

Categorical variables are variables that have data in levels

- They could be quantitative or qualitative that have been converted to quantitative

Examples:

Satisfaction with purchase process: 1/2/3
Gender: M/F
Zip Code

Usually, categorical data needs to be prepared in a special way:

- Dummy variables

Dummy Variables

Dummy variables – also called Indicator Variables have only two values: 0/1

Simple example:
Gender: M/F – 0/1, where 0 is Male, 1 is Female

What if we have a variable with 3 levels:
Car Type = "Sedan", "Compact", "Luxury"

We create 3 dummy variables:
Sedan_Dummy : 0/1, where 0 if not Sedan, and 1 if Sedan

Similarly,
Compact_Dummy: 0/1, where 0 if not Compact, and 1 if Compact
And
Luxury_Dummy: 0/1, where 0 if not Luxury, and 1 if Luxury

Dummy Variables

Car Type	Sedan_Dummy	Compact_Dummy	Luxury_Dummy
Sedan	1	0	0
Compact	0	1	0
Sedan	1	0	0
Sedan	1	0	0
Luxury	0	0	1
Compact	0	1	0

Why not just create a single variable :
Type : 1/2/3 corresponding to Sedan/Compact/Luxury?

Whatever output we generate, will be an "average" output or response

Dummy Variables

So, given a categorical variable, you may need to create dummy variables corresponding to the n levels in the categorical variable

Supposing we have a categorical variable with hundreds of levels – do we create dummies for all levels?

- Depends on what the levels correspond to and data associated with each level
- Usually, we will end up aggregating at a meaningful level

For example: Item Purchased:
May be aggregated to: Grocery/Non-Grocery/Household Item

How many dummies will be created?

Data Preparation : Derived Variables

A very important part of data preparation: Derived variables

They are essentially new variables created from existing variables

Simplest forms of derived variables involve basic calculations or characterizations

For example:

- Given birthdate: Derived variable: Age
- Given Height and Weight: Derived variable: BMI
- Given usage: Derived variable: Low Usage\Medium Usage\High Usage

Why would we need new derived variables?

Data Preparation : Derived Variables

May need new variables because :

- ✓ Business needs
- ✓ Usability of information
- ✓ Pattern recognition at different levels of aggregation

Data Preparation : Derived Variables

Other examples of derived variables :

- **Dummy (Indicator) variables**
- **Lag variables**
  - Capture Time Lag impacts
- **Interaction Variables**

Data Preparation : Derived Variables

Lagged variables are usually created to capture impact of a time delay on outcome

Example :

Impact of inflation sales : may not be in the same period

- Can create multiple order lags (one period lag, two period lags and so on)
- Creating lag of q order will lead to n-q observations total

Data Preparation : Lag Variables

Let's say we are looking at sales as a function of advertising and price

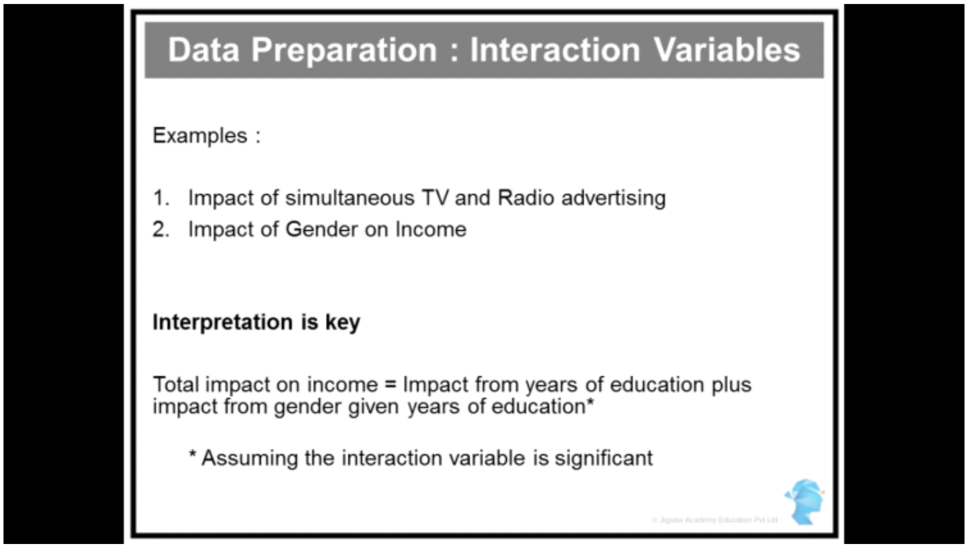
- It may be that the total impact of advertising in Period 1 is actually felt in both period 1 and period 2
  - Will need to create a lag advertising variable to capture impact of period 1 ads on period 2 sales
- Another common time series example is that volume of sales in period 1 has an impact on volume sales in period 2
  - Auto-correlation

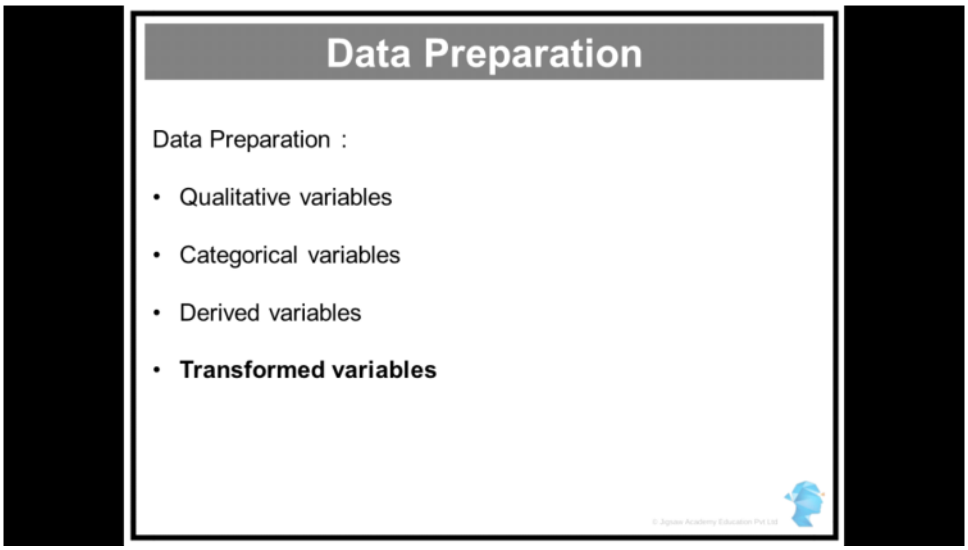
Sales	Price	Advertising $	Lag (Advertising$)
1617	21.99	670	
1804	20.99	587	670
1779	20.99	632	587
1570	21.99	643	632
1730	20.99	765	643
1914	20.99	743	765

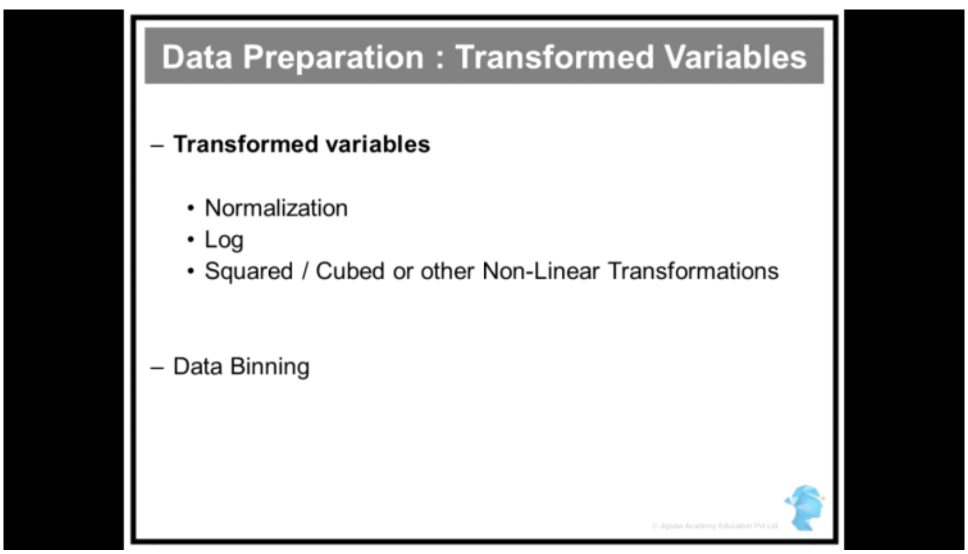
Data Preparation : Interaction Variables

Why would interaction variables be needed?

- We assume (in regression models) that the impact of independent variables on the dependent is additive (linear function)
- This is not always the case: in some cases, the independent variable will have different impacts on the dependent variable as the size of the independent variable changes
- That is, impact of variable A differs as values of variable B change







Normalization of data when there are measures with varying scales:

Data Preparation : Transformed Variables

Data Normalization

Sometimes data is normalized (or scaled down) if there are variables with high variation in magnitude

For example :

Var 1 : Variation Min – Max: 0.01 – 0.1

Var 2 : Variation Min – Max : 400 – 100,000

May want to bring them all to the same scale, especially when using distance algorithms like clustering

Data Preparation : Transformed Variables

Normalization Methods?

Min Max normalization :

We want to change the range of an existing variable to a new (smaller) range:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Z Score normalization : $v' = \frac{v - \mu_A}{\sigma_A}$

Data Preparation : Transformed Variables

Supposing we had a variable, with a min of 30, a max of 340. Mean 125, Std Dev 21.

Using Min Max normalization, we want to change the range to (0,1). So a value of 200 becomes:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

(200 – 30)/(340-30) *(1-0) + 0 = 0.54387

www.cs.gsu.edu/~cscyqz/courses/dm/slides/ch02.ppt

Data Preparation : Transformed Variables

Supposing we had a variable, with a min of 30, a max of 340.
Mean 125, Std Dev 21.

Using Z score normalization :

$$v' = \frac{v - \mu_A}{\sigma_A}$$

(200 – 125)/21 = 3.57

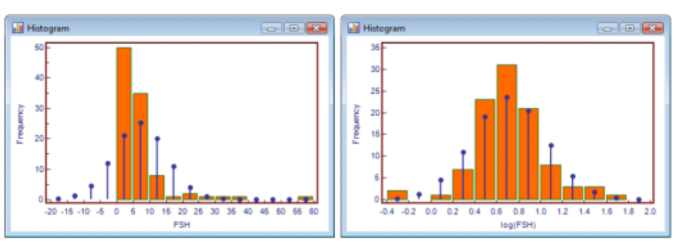
www.cs.gsu.edu/~cscyqz/courses/dm/slides/ch02.ppt

Data Preparation : Transformed Variables

Data is sometimes transformed in order to aid interpretation or to fit with model requirements :

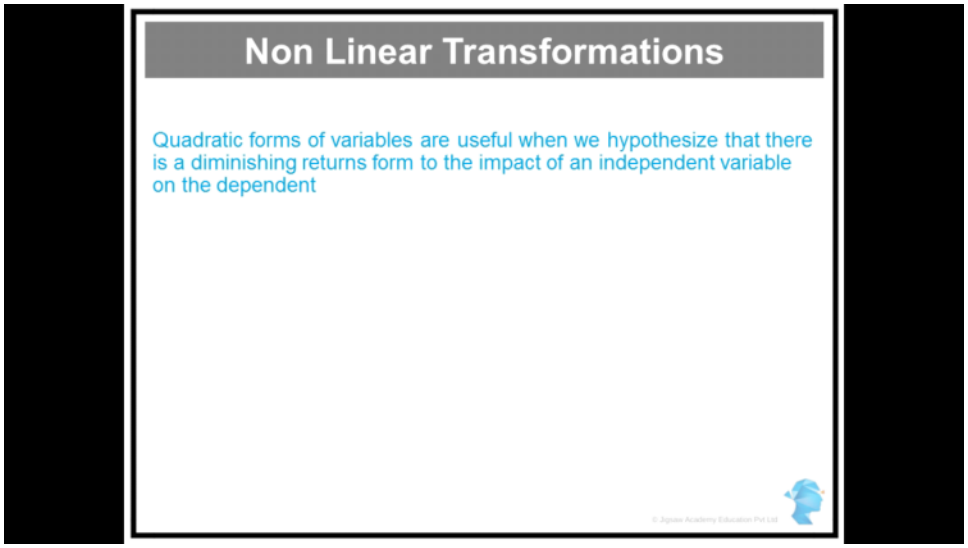
- For example, a linear regression model requires independent variables to be normally distributed. A variable may be transformed by applying the appropriate function to make it a more like a normally distributed variable
- The most common example is to use the log function, but other transformations could be used depending on the distribution of the original variable

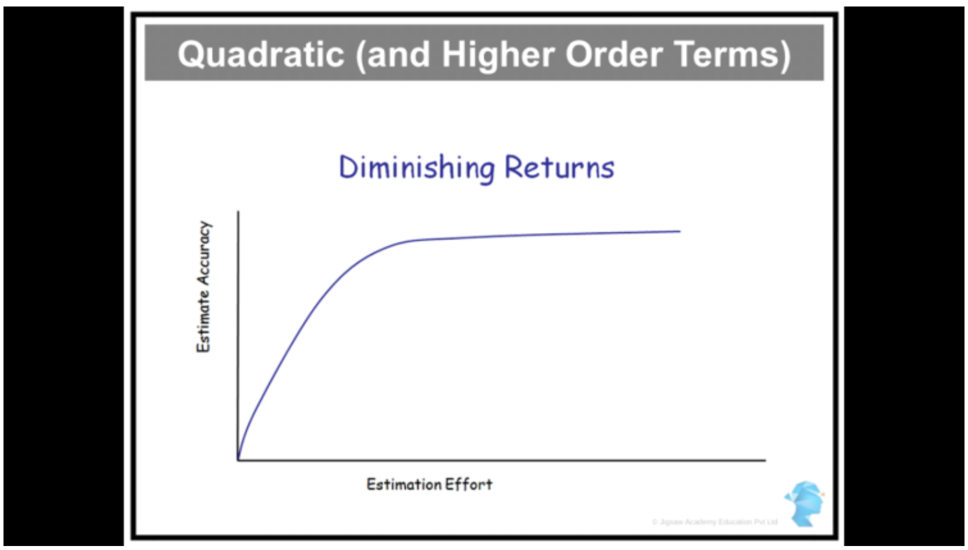
Data Preparation : Log Transformation



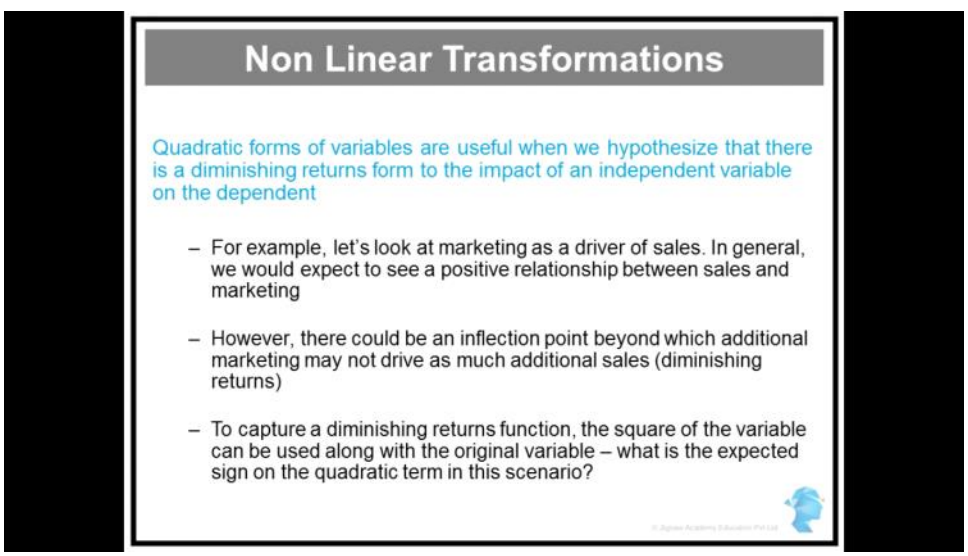
Example of data transformed using a log transformation

- Original variable was skewed with a right tail
- Transformed variable is more "normal"





Diminishing returns --> for every dollar of marking spend I have an increase in sales.



So instead of using the direct variable I will write :

- Sales --> function of marketing square

Binning Continuous Data

It may be useful to split a continuous variable into "bins"

✓ Aids interpretation

✓ Improves actionability

For example: suppose we have income as a continuous Independent variable to be used as a predictor of say credit limit

Which sort of variable would be more useful from an actionability point of view?

a. Income: Continuous (20,000 to 150,000)

b. Income Categories: Wealthy, High, Medium, Low

Binning Continuous Data

This process of binning data is also called "discretization"

1. Equal Interval Binning

a) Data is divided into N equal intervals

b) How do you decide on N?

2. Equal Frequency Binning

- Data is divided into intervals with equal frequencies
- How many bins?

wps_pool

wps_bkt	Races	% Races
0	8645	4.7%
1-25000	61356	33.1%
25001-50000	42137	22.7%
50001-75000	23756	12.8%
75001-100000	12886	7.0%
100001-150000	13571	7.3%
150001-300000	15059	8.1%
300001-5MM	7935	4.3%
>5MM	15	0.0%
	185360	

Binning Continuous Data

In the telecom dataset, we want to classify users as "Light", "Medium", and "Heavy"

- What would be appropriate thresholds?

Variable Name	Number of Observations	Number Missing	Mean	Minimum	Maximum	Std Dev
sbscrp_id	12500	0	82,371,783	19,164,958	88,705,192	5,938,658
minuse1	12499	1	48	0	1,500	98
minuse2	12499	1	182	-55	1,500	165
minuse3	12431	69	182	0	1,500	152
minuse4	12383	117	194	0	177,700	1,603
prom2	12499	1	0.36	0	1	0
prom3	12431	69	0.26	0	1	0
prom4	12383	117	0.24	0	1	0
prom5	12130	370	0.12	0	1	0
BIRTH_DT	12411	89	19,600,025	19,031,021	20,010,212	147,290
zip_code	12500	0	49,395	605	99,901	29,457

Data Pre-processing

Data exploration and data preparation usually are done together

Data Extraction Data Integration Data Assessment	→ Data Exploration
Data Cleaning ✓ Data Transformation ✓ Data Reduction	→ Data Preparation

Data Reduction

Why would we want to "REDUCE" data?

High Dimensionality Processing

- Time consuming
- Variables > Observations

Multi-Collinearity Issues

- High Correlations

Data Reduction

How is data reduced?

Multiple dimension reduction techniques:

- Drop correlated variables – simple, but not always justifiable
- Which variable to drop?

- Principal Component Analysis, or Factor Analysis

Identify components that are weighted linear combinations of multiple variables, and use the components in the model instead of the actual variables

Data Pre-processing

Data exploration and data preparation usually are done together

Data Extraction Data Integration Data Assessment	→ Data Exploration
Data Cleaning ✓ Data Transformation ✓ Data Reduction ✓	→ Data Preparation

One final thing: Sampling

Data Preparation : Partitioning

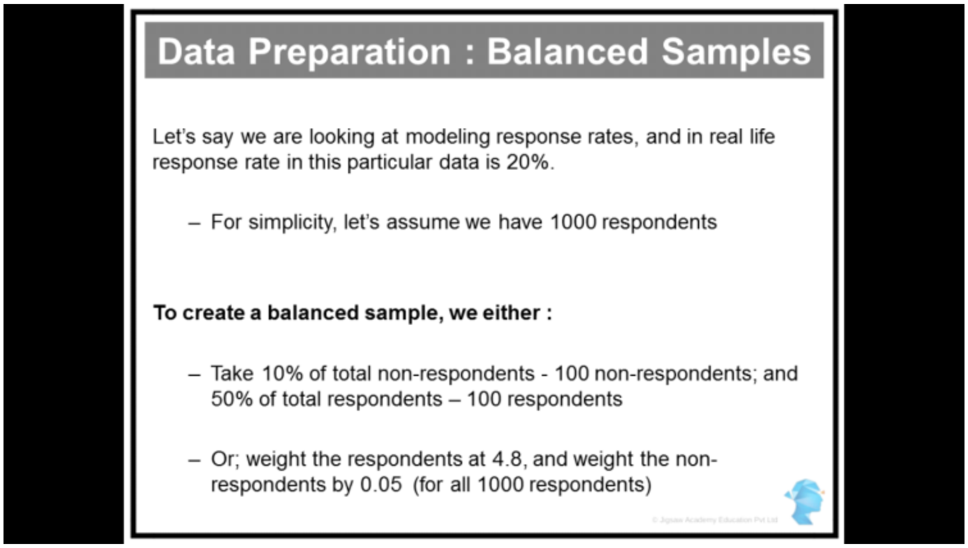
A quick overview of creating sample datasets

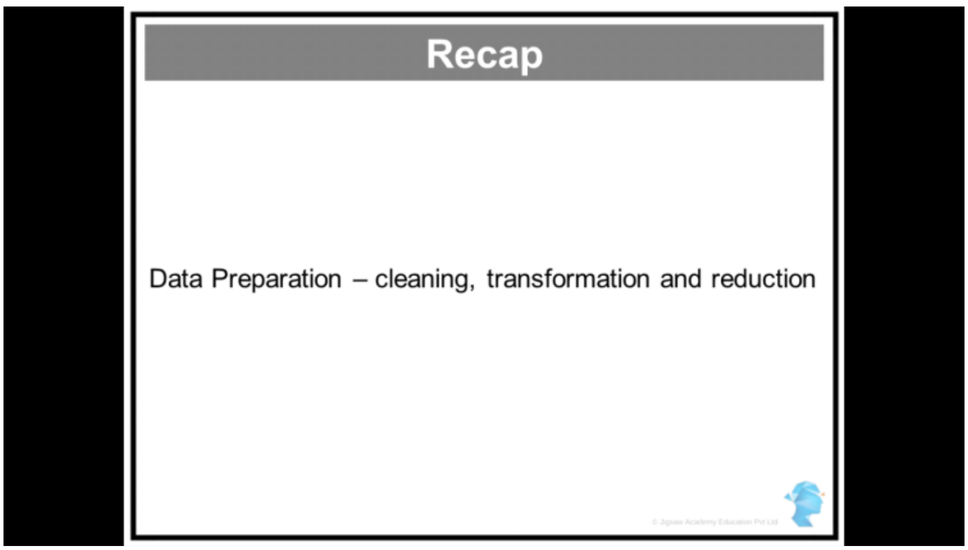
- Once the data prep is complete, the next step is to create multiple sample datasets from the complete data. These are :
  - **Training Dataset** – this is the sample of the data on which the initial model is built
  - **Validation Dataset** – this is another random sample of the data upon which model accuracy and predictability is tested
  - Sometimes, also a **Test Dataset** – this is a third dataset that is sometimes used to finally test accuracy of refined models
- Why can't the training dataset be used to test accuracy of model?

Data Preparation : Balanced Samples

Balanced Sample:

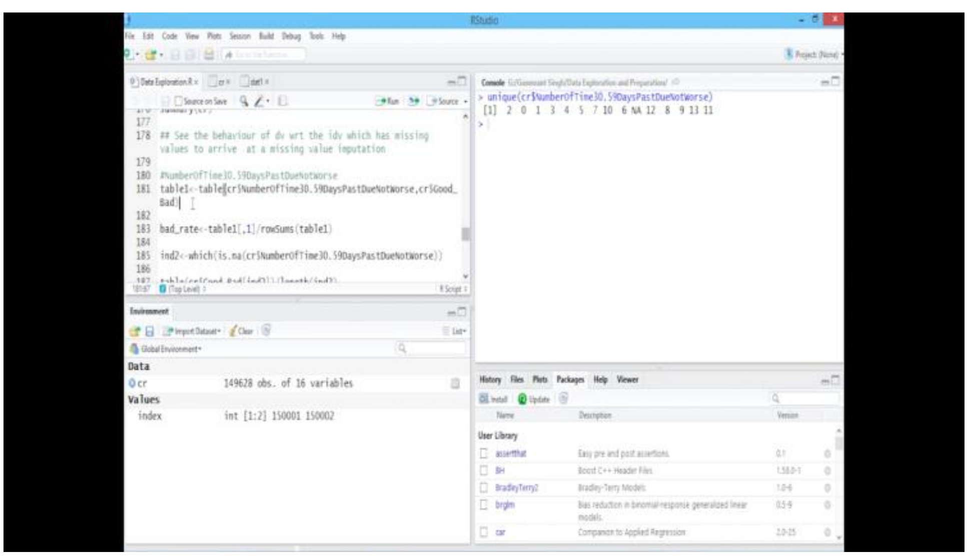
- An important thing to remember is that in any modeling approach, you want the data to reflect all the possibilities that you want to model
- So, for example, let's say you want to assess response % to a direct marketing campaign. You will need to have both respondents and non-respondents in your sample dataset
- You will also need roughly equal proportions of respondents and non-respondents in order to create reliable models
- In real life, it will be rare for that ratio to exist naturally in the data, requiring the analyst to create a balanced sample for the analysis
  - Sample different categories differently
  - Weight categories differently



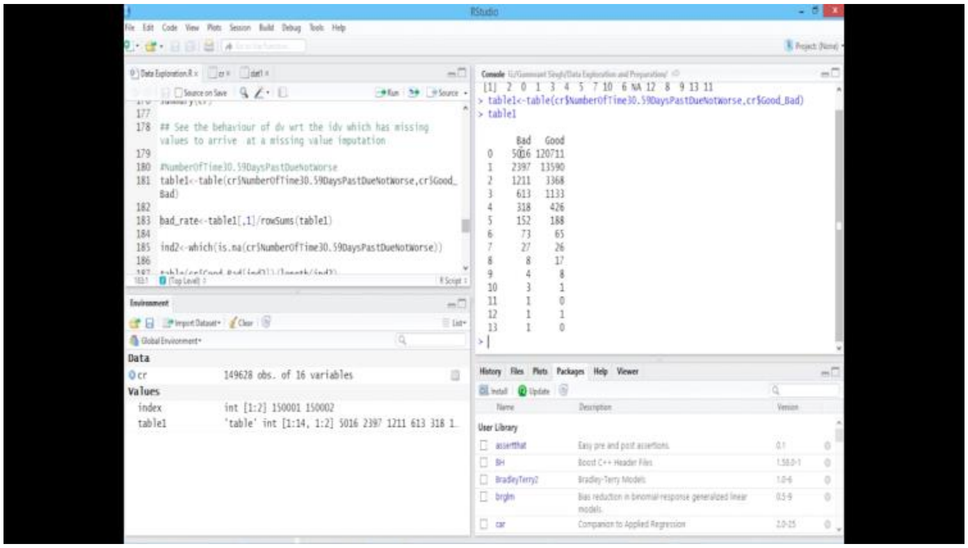


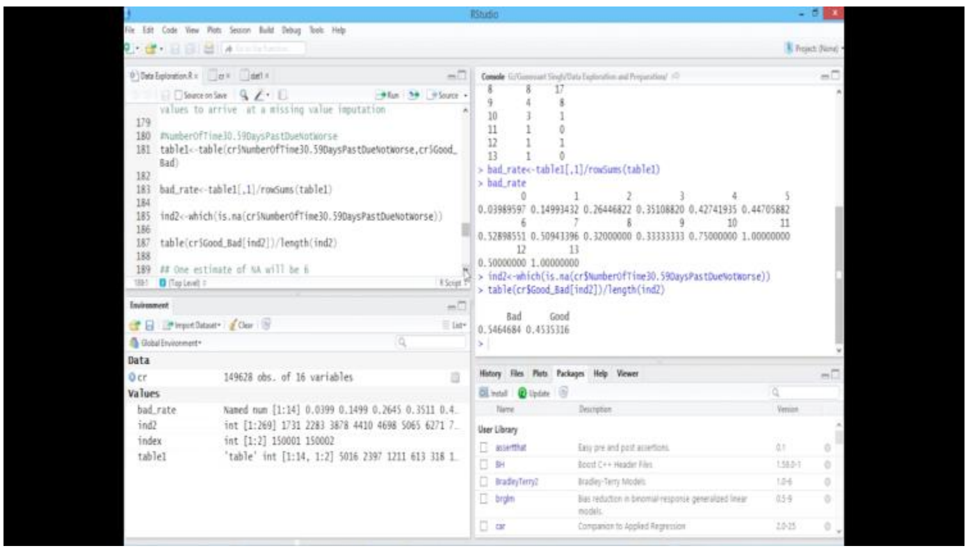
Data Exploration 2- R code demo:

How to impute missing values in a data



Use the table command to cross tablulate frequency and variables :

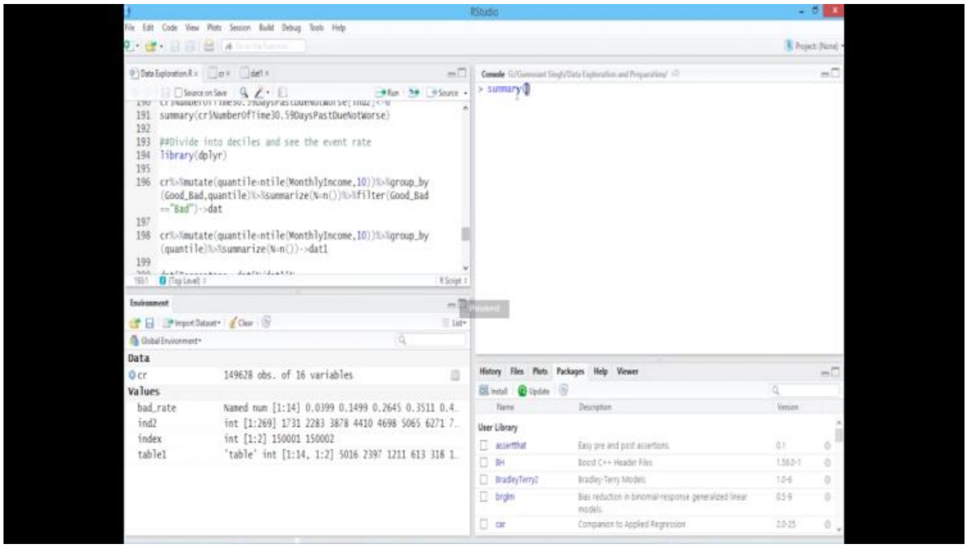




Since the values for missing values is closest to the one for 6 defaults, we will deem all missing values as 6 .

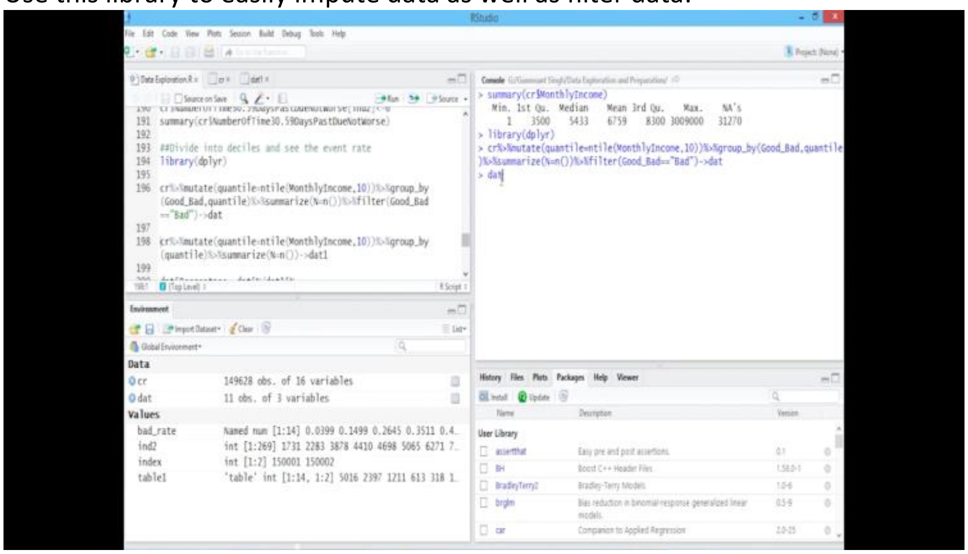
The command to update missing values :
Df$olumnname[indexvalue]<-6

This is one way of imputing missing values

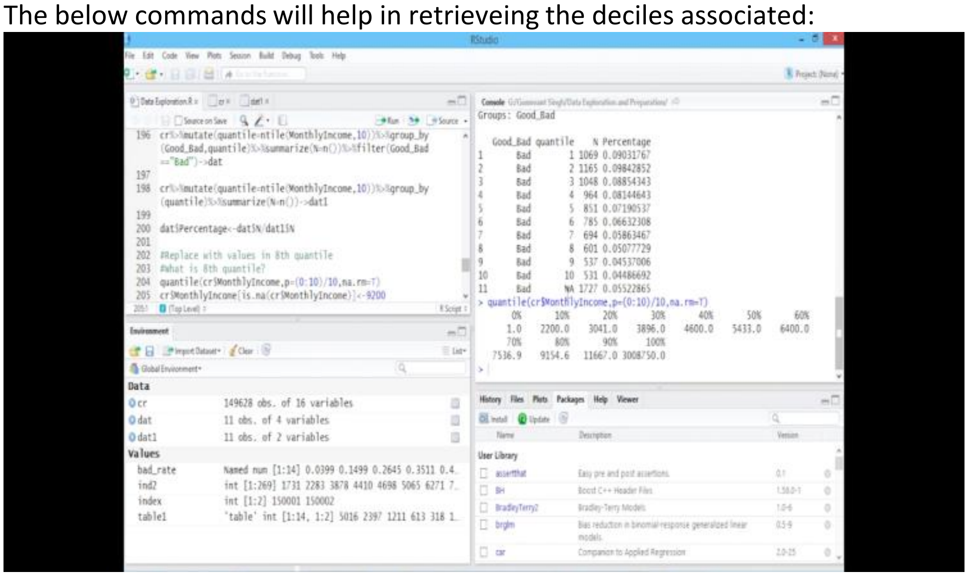


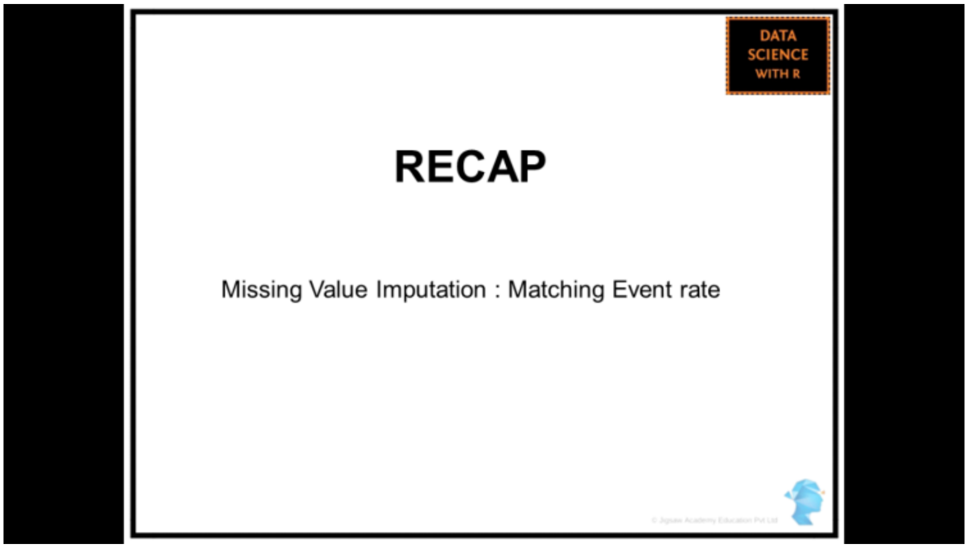
Using dplyr

Use this library to easily impute data as well as filter data:

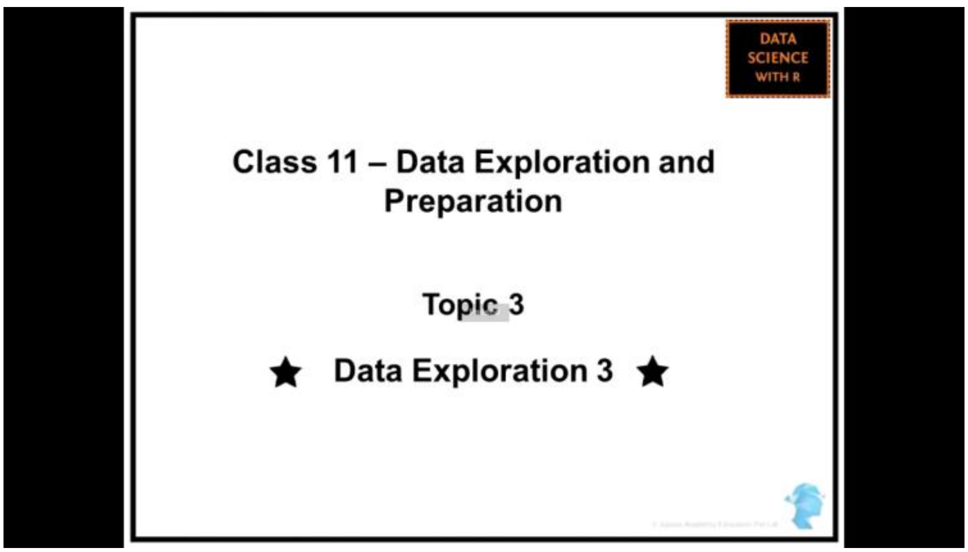


The below commands will help in retrieveing the deciles associated:

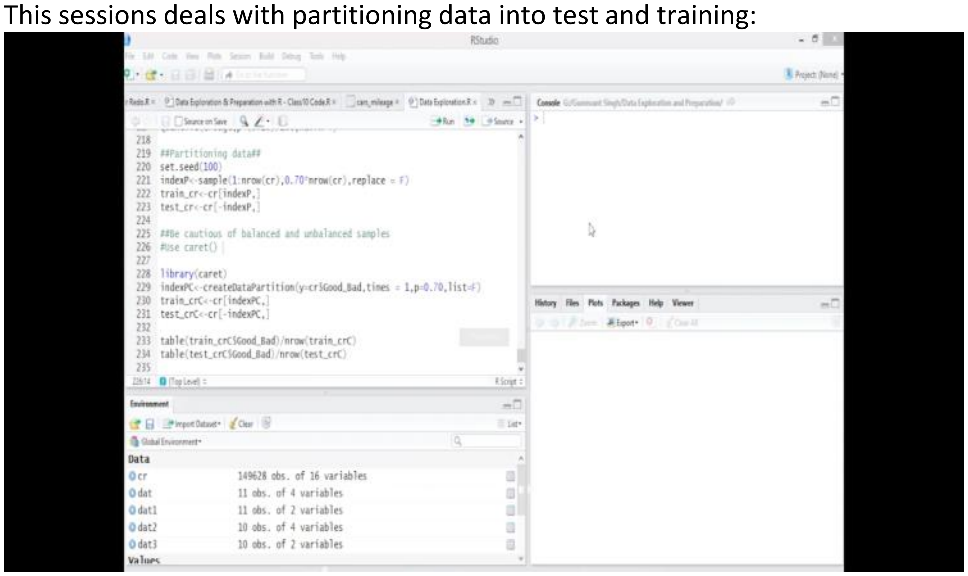




Create deciles for continuous groups and do the necessary comparision and imputations



This sessions deals with partitioning data into test and training:

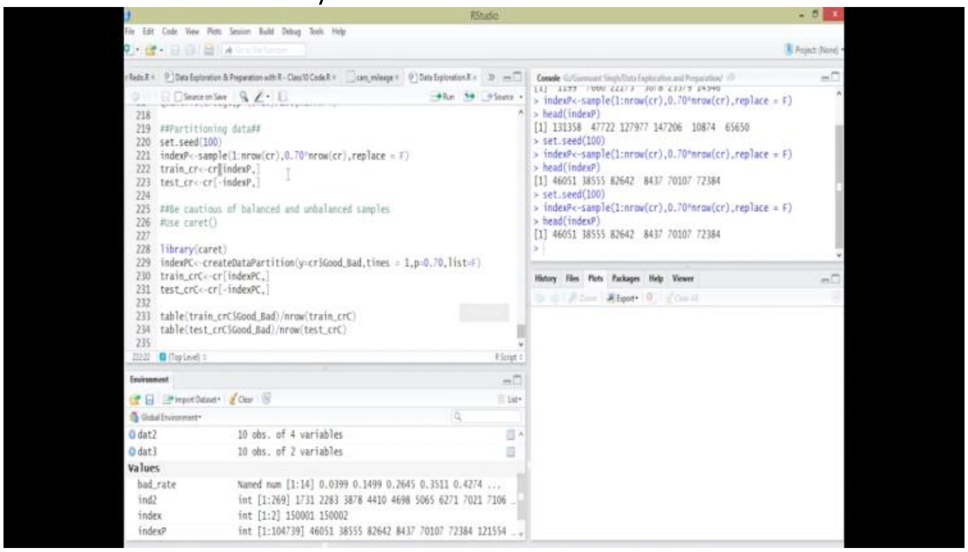


Steps involved in partitioning the data:

- ➢ set.seed(100) --> this command will make my results reproducible i.e. samples that have been randomly selected will be retained every time we run the sample command and not re-indexed

- indexP<-sample(1:nrow(POLK_veh_reg_dt),0.70*nrow(POLK_veh_reg_dt), replace = F)
- train_polk<-POLK_veh_reg_dt[indexP,]
- test_polk<-POLK_veh_reg_dt[-indexP,]

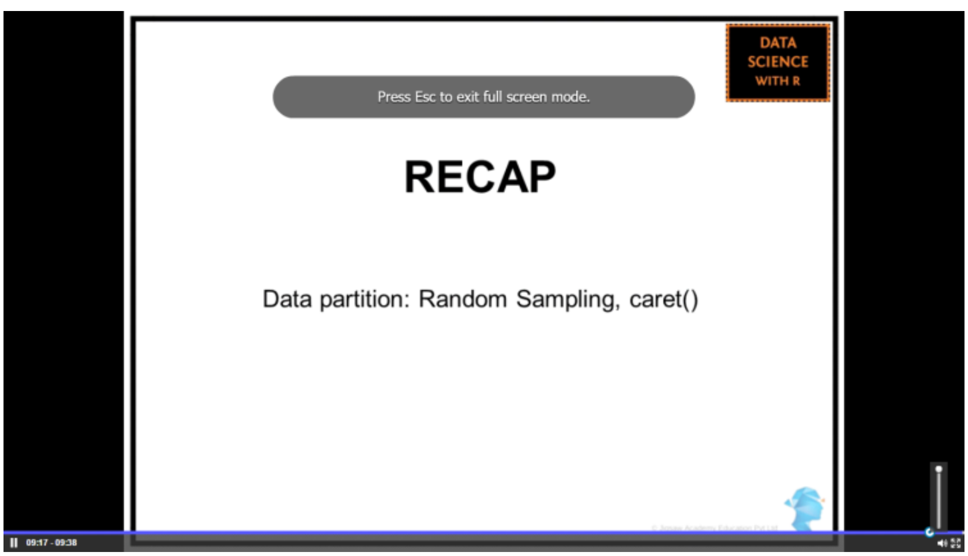
The two datasets are ready.



We can use the caret package for test and training dataset:

Library(caret)
indexPC<-createDataPartition(y=POLK_veh_reg_dt$UNITS_BOUGHT,times = 1,p=0.7,list = FALSE)

- Here we mention the column based on which we want to perform the partitioning, times=<how many different samples do we want>, p=<percentage breakdown>, list=<do we want to store as list or not>



Either use sample command from base R or use the custom Caret package