JIGSAW ACADEMY

Analytics for Professionals

DATA PREPARATION

DATA PREPARATION

1. Why does data need to be "prepared"?

2. How is data "prepared"?

3. Avoid "Garbage In Garbage Out"

DATA PREPARATION

1. Why does data need to be "prepared"?

1. Data needs to be usable for models
2. Data needs to be checked and treated for consistency and completeness
3. Additional variables may be required for the actual modeling process

DATA PREPARATION

1. How is data "prepared"?

I. Identifying and dealing with outliers

II. Missing value treatments

III. Qualitative and categorical variables

IV. Creating additional variables

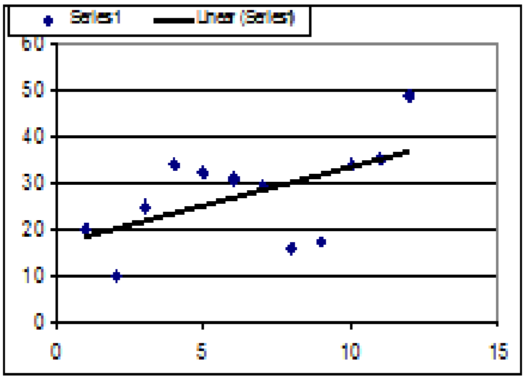
- Derived variables including dummy variables
- Binning Data

V. Data transformation

OUTLIER DETECTION

What are outliers?

- 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: "<u>Differ</u>", "<u>Significantly</u>"

- In the chart here, 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 vs 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
J	34	21-25
K	35	50+
L	49	50+

OUTLIER DETECTION

How are outliers identified?

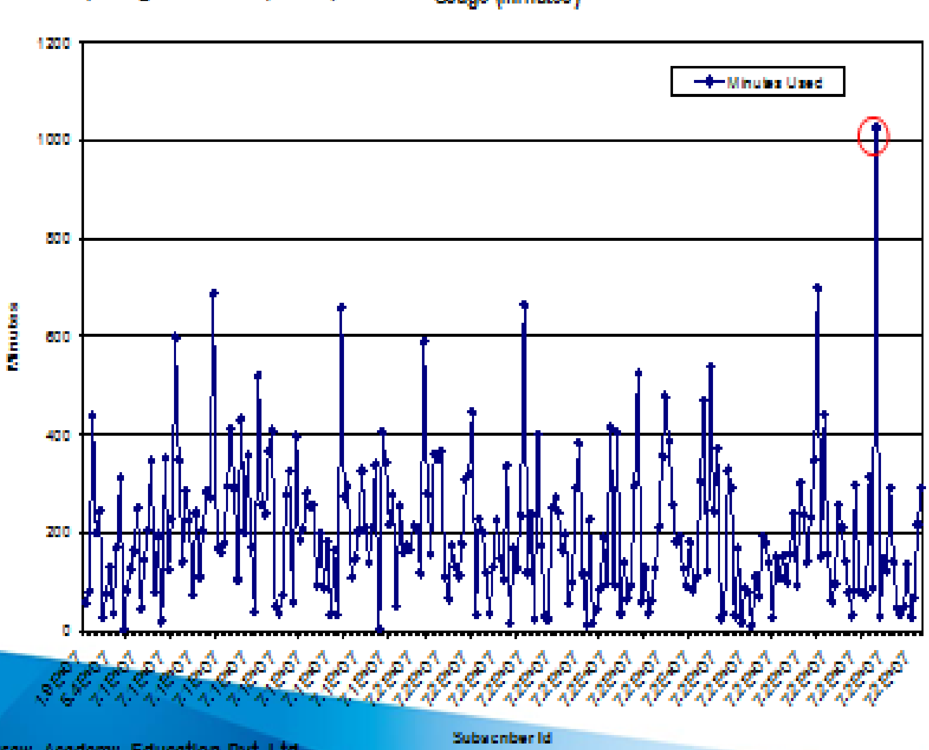
- Graphical visualization via
  - Run charts
    - Summarizes a univariate data set
    - Typically plotted against time
  - Histograms
  - Box Plots
    - Efficient 5-member data summary
  - Probability distribution charts

Domain knowledge

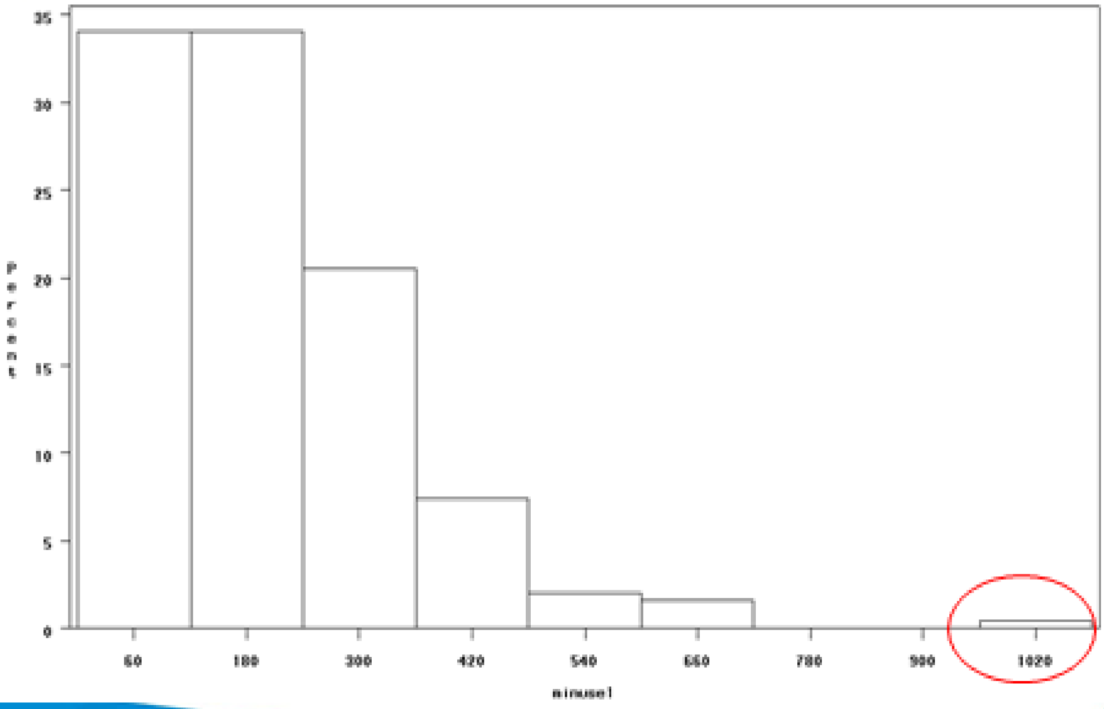
OUTLIER IDENTIFICATION

Example of a Run Chart

Monthly usage of mobile, time period 1

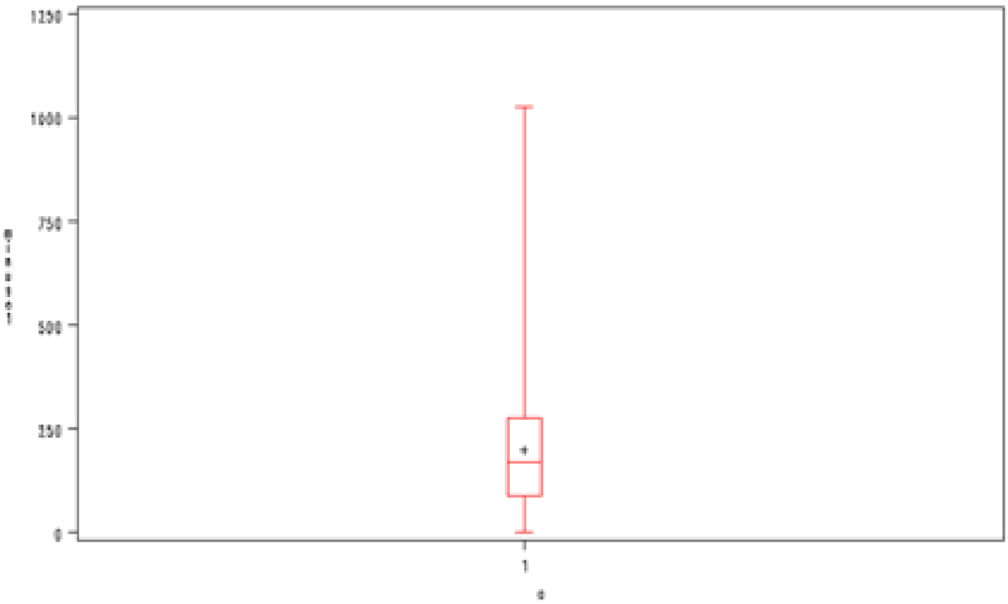


OUTLIER IDENTIFICATION
HISTOGRAM



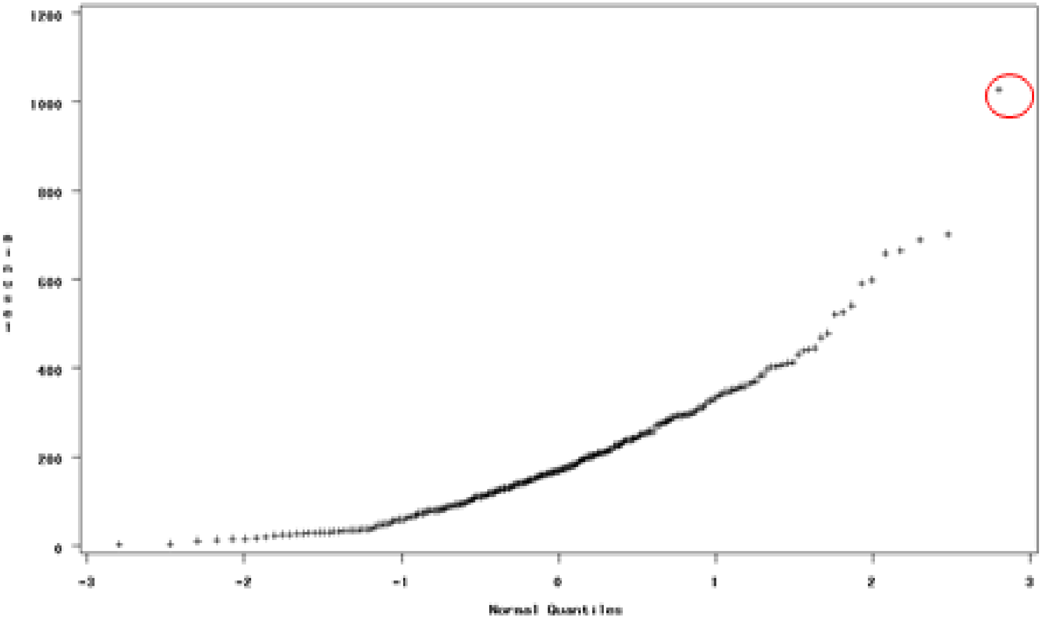
OUTLIER DETECTION
CASE STUDY CHARTS

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

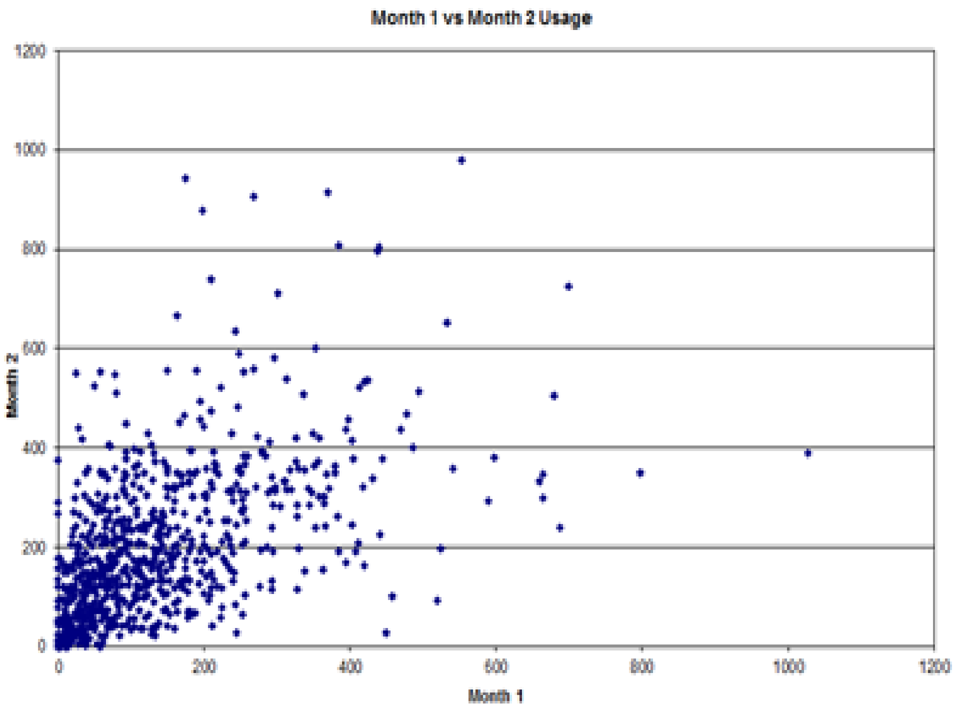
- Will need to look at pairs of observations (joint distributions)

Should we look at all possible pairs?

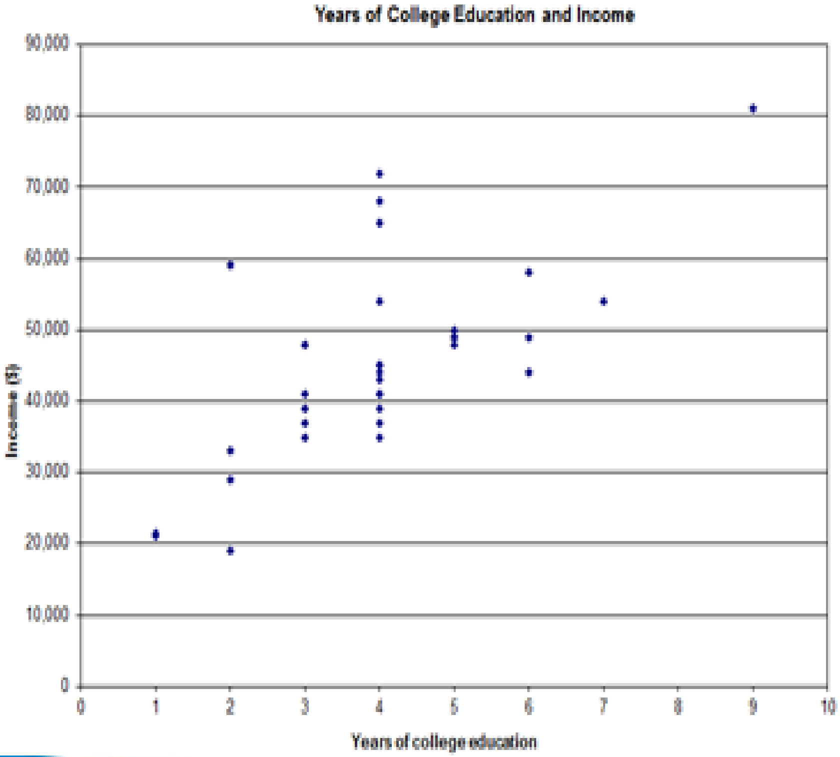
- In large datasets, with multiple variables, it will not be possible
- Domain knowledge and problem background will help in determining what pairs to look at?

Only pairs of variables? What about combinations greater than 2?

OUTLIER DETECTION
CASE STUDY



OUTLIER DETECTION
ANOTHER EXAMPLE



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?

OUTLIER TREATMENT

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	[illegible]	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

If outlier:

Delete the value – Implies entire record will be lost

Substitute another value:

- Mean
- Max
- Similar Case Mean

OUTLIER TREATMENT

Other Options:

3. **Transformation**
   - Taking the log, for example, for variables with positive values, will reduce the spread

4. **Ignore Outliers**
   - What are implications of ignoring outliers
     - Robust statistics - TLS

- It is important to understand cause of outliers in order to arrive at the best method of dealing with them

- Remember ,outliers are "influential"

OUTLIER TREATMENT

It is important to understand cause of outliers in order to arrive at the best method of dealing with them

Treatment options include:

1. Delete outlying values
2. Substitute appropriate values
3. Transform data
4. Check results with and without

DATA PREPARATION

1. How is data “prepared”?

I. Identifying and dealing with outliers ✓

II. Missing value treatments

III. Qualitative and categorical variables

IV. Creating additional variables

- Derived variables including dummy variables
- Binning Data

V. Data transformation

MISSING VALUES

Why should we worry about data that is missing? Can we not ignore it since it is missing anyway?

- Can missing data provide any information?
- How do we get that information?

Patterns to missing data

- Data missing at random
- Data missing not at random

Implications of missing data

What to do about missing data?

- Ignore? Delete? Impute values?

MISSING VALUES - ASSESMENT

Prom8	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 DATA - 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 DATA - 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

MISSING DATA - TREATMENT

- Do not replace missing values with any constant!
- Imputation
  - Single Imputation
  - Multiple Imputation
    - Example: impute values using regression techniques?
    - Computationally intensive
- What if dependent variable has missing values?
  - Imputation?

- Single Imputation –
  - Same substitute for all missing values
  - Multiple imputation – generate a range of values that could be used as substitutions
- In case the dependent variable is missing – it is better to delete the entire record

MISSING DATA – SANITY CHECK

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

How many missing values exist in the table above?

DATA PREPARATION

1. How is data "prepared"?

I. Identifying and dealing with outliers ✓

II. Missing value treatments ✓

III. Qualitative and categorical variables

IV. Creating additional variables

- Derived variables including dummy variables
- Binning Data

V. Data transformation

QUALITATIVE or CATEGORICAL VARIABLES

Qualitative variables cannot be used directly in models (need numeric data)

- Transforming qualitative variables is simple:
  - Examples: Gender – M/F to 0/1.
  - If gender = "M" then gender = 0; else gender = 1;

Sometimes categories in a qualitative variable are too many

Example: Profession, Item Purchased

- substitute a more meaningful value to that variable - grocery vs 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 an appropriate way to convert this variable to be used in a model?

DATA PREPARATION

1. How is data "prepared"?

I. Identifying and dealing with outliers ✓

II. Missing value treatments ✓

III. Qualitative and categorical variables ✓

IV. Creating additional variables

- Derived variables including dummy variables
- Binning Data

V. Data transformation

DATA PREPARATION
DERIVED VARIABLES

Derived variables – new variables created from existing datasets

- Simple examples:
  - BMI. Derived from Height and Weight
  - Categorizing values as High, Medium, Low
- Dummy (Indicator) variables
- Lag variables
  - Capture Time Lag impacts
- Other derived variables
  - Transformed – Log
  - Squared/Cubed etc to arrive at diminishing returns
  - Interaction Variables

DATA PREPARATION
DERIVED VARIABLES

The simplest forms of derived variables are those that involve basic calculations or characterizations

- For example, age from date of birth
- Greater than average or less than average response
- Other examples?

DATA PREPARATION
DUMMY VARIABLES

Dummy variables or indicator variables are frequently created to allow the use of qualitative categorical values in a modeling dataset

Dummy variables have only two values: 0 and 1

Simplest example: Male vs Female

- In dataset, this would reflect as: if male = yes, then dummy_male = 1. If male = no, then dummy_female = 1;

Can of course be created for a variable that has more than two categories

- Low, Medium, High; Type of model
- If model type = x, then dummy_x = 1;

What should be value when dummy_x is not equal to 1?

DATA PREPARATION
DUMMY VARIABLES

What if variable has more than two categories?

- Low, Medium, High;
- Car model type

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

DATA PREPARATION
DUMMY VARIABLES

If type = 'Luxury" then luxury_type = 1;

If not 1 then ?

Why create dummy variables at all? Why not convert the type variable to a 1,2,3 variable?

DATA PREPARATION
LAG VARIABLES

- Lagged variables are usually created to capture impact of a time delay on outcome
  - Example: Lag of CCI on sales
- 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
- Lagged variables usually also created to generate derived variables (stock, for example)

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

DATA PREPARATION
INTERACTION VARIABLES

Examples:

1. Impact of simultaneous TV and Radio advertising
2. Impact of Gender and Education on Income

How could these effect be captured?

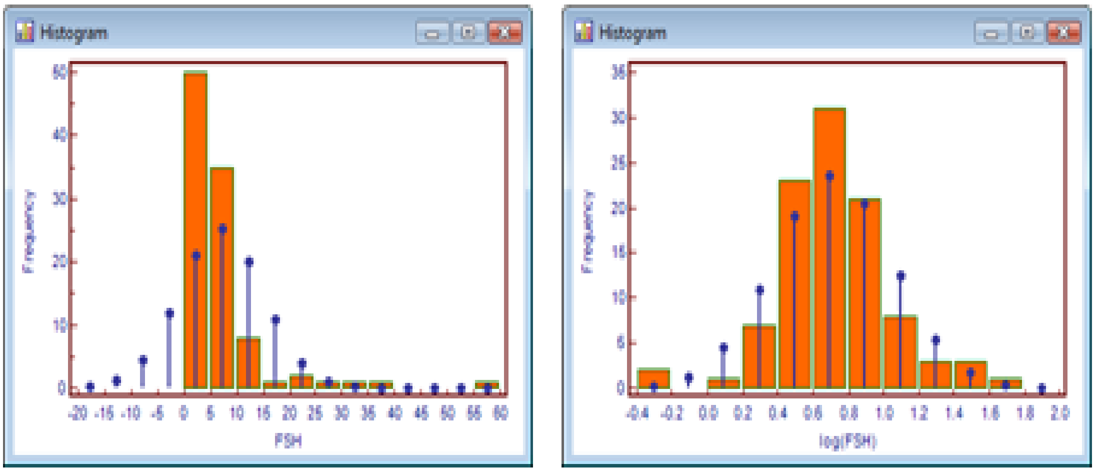
Interaction terms: Sales = f(Tv ads, Radio ads, TV*Radio)

Sales = Intercept + 2000 * TV + 1650 * Radio + 218 * (TV*Radio);

DATA PREPARATION
TRANSFORMING VARIABLES

- Data is sometimes transformed in order to aid interpretation or to fit with model requirements:
  - For example, OLS 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 could also be transformed to aid interpretability of results
  - A very common example is the constant elasticity model

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”

DATA PREPARATION
QUADRATIC TERMS

Quadratic forms of variables are useful when we hypothesize that there is a diminishing returns form to the impact of an independent variable on the dependent

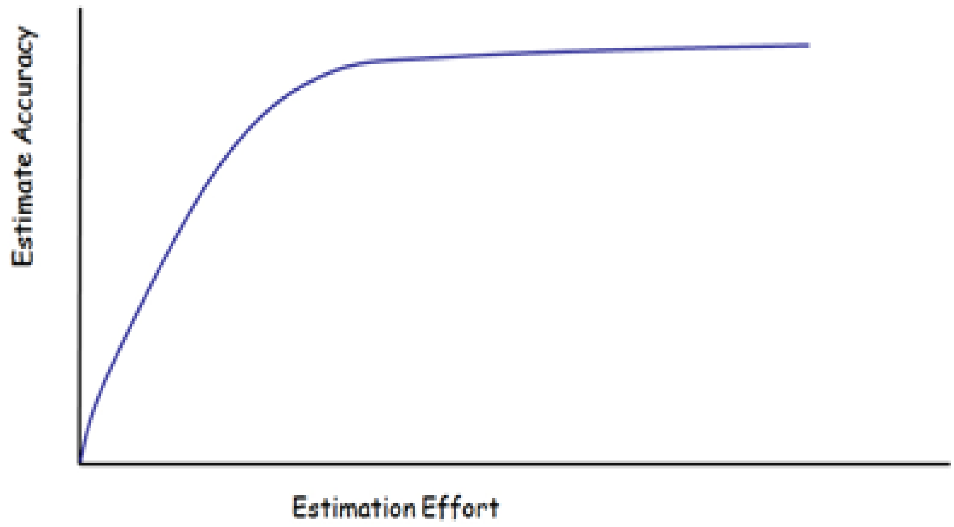
- For example, let's look at marketing as a driver of sales. In general, we would expect to see a positive relationship between sales and marketing

- However, there could be an inflection point beyond which additional marketing may not drive as much additional sales (diminishing returns)

- To capture a diminishing returns function, the square of the variable can be used along with the original variable – what is the expected sign on the quadratic term in this scenario?

DATA PREPARATION
QUADRATIC (and HIGHER ORDER TERMS)

Diminishing Returns



DATA PREPARATION
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

DATA PREPARATION
BINNING CONTINUOUS DATA

This process of binning data is also called "discretization"

There are different ways of binning

1. Equal Interval Binning

a. Data is divided into N equal intervals

b. How do you decide on N?

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	

2. **Equal Frequency Binning**

a. Data is divided into intervals with equal frequencies

b. How many bins?

DATA PREPARATION
DEALING WITH DATES

Date Variables – Information from date variables needs to be derived appropriately to be used in any models (cannot use a date value directly)

Age from date of birth, month of response, day of week, weekend indicators are all examples of data derived from date variables

Different software applications have different ways of dealing with date variables

- SAS, for example converts all dates to days from 1960

DATA PREPARATION

1. How is data "prepared"?

I. Identifying and dealing with outliers ✓
II. Missing value treatments ✓
III. Qualitative and categorical variables ✓
IV. Creating additional variables
   - Derived variables including dummy variables
   - Binning Data
V. Data transformation ✓

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 PREPARATION
BALANCED SAMPLES

Let's say we are looking at modeling response rates, and in real life response rate in this particular data is 20%.

- For simplicity, let's assume we have 1000 respondents

To create a balanced sample, we either:

- Take 10% of total non-respondents - 100 non-respondents; and 50% of total respondents – 100 respondents
- Or; weight the respondents at 4.8, and weight the non respondents by 0.05 (for all 1000 respondents)

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
PARTITIONING

Partitioning in SAS

```
proc surveyselect data = TEST method = SRS rep = 1 sampsize =
    5000 seed = 12345 out = SAMP1;
id _all_;
Run;
```

Proc SurveySelect allows you to generate random samples, and the seed number allows you to regenerate the same sample from the underlying population

JIGSAW ACADEMY

Analytics for Professionals

www.jigsawacademy.com

DATA SCIENCE

Data Science with R

Lesson 7 Case Study

Data Exploration and Preparation

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment
  - Creating dummy variables if necessary

- **Profiling the data**
  - To understand the relationship between DV and the IDVs
  - Using Data manipulations and Visualisations to your aid

Steps

- **Cleaning the data**

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment
  - Creating dummy variables if necessary

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment
  - Creating dummy variables if necessary
- **Profiling the data**

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment
  - Creating dummy variables if necessary

- **Profiling the data**
  - To understand the relationship between DV and the IDVs

Steps

- **Cleaning the data**
  - Making sure the data has no anomalies
  - Missing Value Treatment
  - Outlier Treatment
  - Creating dummy variables if necessary

- **Profiling the data**
  - To understand the relationship between DV and the IDVs
  - Using Data manipulations and Visualisations to your aid

Steps

- **Deciding on what models to build**

Steps

- **Deciding on what models to build**
    - **Logistic Regression** - Build a classification model to predict the likelihood of a patient having a heart disease

Steps

- **Deciding on what models to build**
  - **Logistic Regression** - Build a classification model to predict the likelihood of a patient having a heart disease
  - **Decision Trees** - Build a classification model to predict the likelihood of a patient having a heart disease

Insights from Profiling the data

END OF LESSON 7 CASE STUDY