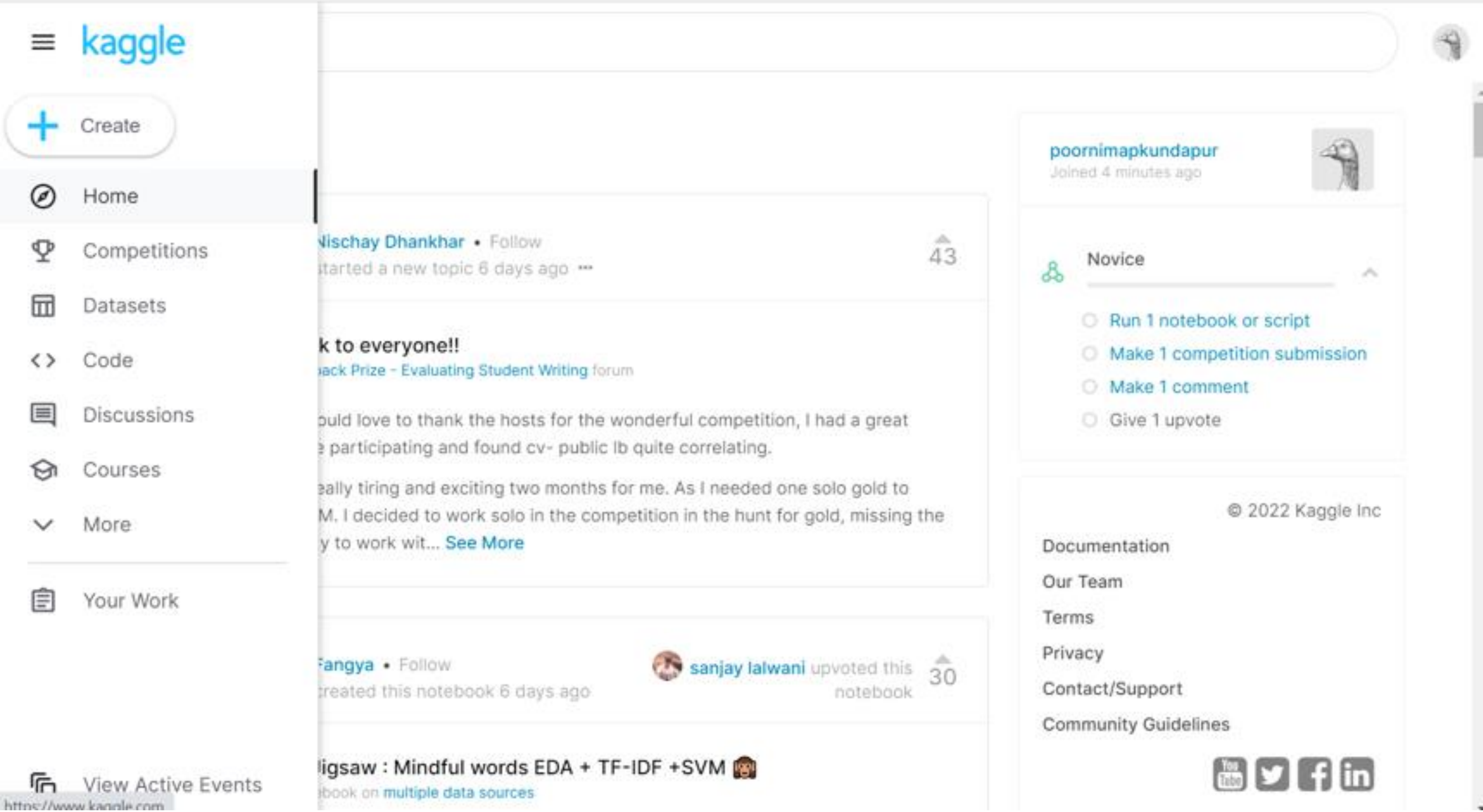
PREPARING DATA TABLES.

MOST TIME-CONSUMING.

Things we know and what need to know.


kaggle
Create
Home
Competitions
Datasets
Code
Discussions
Courses
More
Your Work
View Active Events
https://www.kaggle.com
Nischay Dhankhar • Follow
started a new topic 6 days ago •••
43
k to everyone!!
ack Prize - Evaluating Student Writing forum
ould love to thank the hosts for the wonderful competition, I had a great
participating and found cv- public lb quite correlating.
eally tiring and exciting two months for me. As I needed one solo gold to
M. I decided to work solo in the competition in the hunt for gold, missing the
y to work wit... See More
Fangya • Follow
created this notebook 6 days ago
sanjay lalwani upvoted this notebook
30
Jigsaw : Mindful words EDA + TF-IDF +SVM
ebook on multiple data sources
poornimapkundapur
Joined 4 minutes ago
Novice
Run 1 notebook or script
Make 1 competition submission
Make 1 comment
Give 1 upvote
© 2022 Kaggle Inc
Documentation
Our Team
Terms
Privacy
Contact/Support
Community Guidelines

Looking at our dataset.

Table 3.9. Table of patient records

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh^+	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh^-	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^+	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh^-	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh^-	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^-	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^+	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh^-	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh^+	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh^+	136	1.78	121	165	98.7	1

The overview of this chapter.

- ***Data*** is one of the **most time-consuming parts** of a data analysis/data mining project
- Outlines **concepts and steps necessary to prepare a data set** prior to beginning data analysis
- Way in which the data is collected and prepared is critical to the confidence with which decisions can be made.
  - Data needs to be merged into a table and this may involve integration of the data from multiple sources.
  - Once the data is in a tabular format, it should be fully characterized or categorised.

The overview of this chapter.

Process of preparing data for analysis.

- Data should be cleaned by **resolving ambiguities and errors**, **removing redundant** and **problematic data,** and **eliminating columns of data irrelevant** to the analysis.
- **New columns** of data may need to be calculated.
- **Table** should be **divided**, where appropriate, into **subsets** that either **simplify the analysis** or allow specific questions to be answered more easily.

Important to record the details about the steps that were taken and why they were done.

Provides documentation of the activities performed so far, and a methodology to apply to similar data sets in the future.

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL

In-semester Assessment Plan for II Semester M Tech/MCA

Type of Assessment	Mode	Max. Marks	Plan/Schedule	Remarks
In-semester Examination - 1	**Offline/Physical**	**15**	**Eighth Week:** April 18 - 23, 2022	• The time duration is 60 minutes • The question(s) are from the portions/topics delivered during February 28 – April 14, 2022 • The question paper pattern is decided by the department/teachers • MCQs/fill in the blanks will **NOT** be asked in the examination
In-semester Examination - 2	**Offline/Physical**	**15**	**Fourteenth Week:** May 30 – June 04, 2022	• The question(s) are from the portions/topics delivered during April 25, 2022 – May 28, 2022 • Other guidelines are same as those for In-semester Examination-1
Individual Assignments /Quizzes/ Seminar/Mini Project etc.	-------	**20**	Teachers/Departments would plan the assessment method and schedule	• The teachers would inform the students regarding the assessment method and schedule for their courses
	Total	**50**		

NOTE: No Theory classes are conducted during the days of in-semester examination

CLEANING THE DATA.

- For variables measured on a nominal **or ordinal scale** it is useful to inspect all possible values **to uncover mistakes, duplications** and **inconsistencies.**
- Each **value** should map onto a **unique term**:
- For example, a variable **Company** may include a number of different spellings for the same company:
  - “General Electric Company,” “General Elec. Co.,” “GE,” “Gen. Electric Company,” “General electric company,” and “G.E. Company.”

When these **values** refer to the same company, the various terms should be **consolidated into one.**

CLEANING THE DATA: Non-numeric

- A common problem with **numeric variables** is the inclusion of **non-numeric terms.**
- For example, a variable generally consisting of numbers may include a value such as “above 50” or “out of range.”
- *Numeric analysis* **cannot** *interpret a non-numeric value* and hence, relying on subject matter expertise.

These **terms** should be converted to **numbers**, or the observation **removed.**

CLEANING THE DATA: Missing

- Another problem arises when observations for a particular variable are **missing data values**.

Where there is a specific meaning for a **missing data value**, the value may be **replaced** based on knowledge of how the data was collected.

CLEANING THE DATA: Interval or ratio

- Take **any possible value** within a **range**.
- Useful to consider **outliers** in the data.
- For example:
  - An outlier may be an **error in the measurement** or the **result** of measurements made using a **different calibration**.
  - An outlier may also be a **legitimate** and **valuable** data point.

Consider outliers.

CLEANING THE DATA: Standardization.

- Variable may have been **measured** over **different units.**
- For example, a variable:
  - *Weight* may have been measured using both *pounds* and *kilograms* for different observations or a variable
  - *Price* may be measured in different currencies.

Standardize measurements to a single scale so that they **can be compared** during analysis.

CLEANING THE DATA: Time.

- In situations where **data** has been **collected over time**, **changes** related to the **passing of time** may **no longer be relevant.**
- For example, when looking at a variable *Cost of production* for which the data has been collected over many years, the rise in costs attributable to inflation may need to be considered for the analysis.

Changes related to time not relevant, **may not be** considered.

CLEANING THE DATA: Duplicates.

- When **data** is **combined from multiple sources**, an observation is **more likely** to have been **recorded more than once.**

Duplicate entries should be removed.

REMOVING OBSERVATIONS AND VARIABLES.

Remember

- After an initial categorization of the variables, it may be p***ossible to remove variables*** from consideration.
- For example, **constants and variables** with **too many missing data** values would be candidates for **removal.**
- Similarly, it may be necessary to **remove observations** that have data missing for a **particular variable**.

GENERATING CONSISTENT SCALES ACROSS VARIABLES.

- **Difficulty** *processing data* in its **raw form**.
  - Apply **mathematical transformations** to the data.
- **Normalization uses a mathematical function to transform numeric columns to a new range.**
  - Prevents certain data analysis methods from giving some variables undue influence over others due to differences in the range of their values.
- *For example:*
  - When analyzing customer credit card data, the Credit limit value (varies from $500 to $100,000) should not be given more weight in the analysis than the Customer’s age (whose values might range from 18 to 100).

GENERATING CONSISTENT SCALES ACROSS VARIABLES.

- The **min–max transformation** maps the **values** of a variable to a **new range**, such as from 0 to 1.

$$x'_i = \frac{x_i - OriginalMin}{OriginalMax - OriginalMin} \times (NewMax - NewMin) + NewMin$$

where $\mathbf{x'_i}$ is the **new normalized value**, $\mathbf{x_i}$ is the **original variable's value**,

OriginalMin is minimum possible value in the original variable

OriginalMax is the maximum original possible value

NewMin is the minimum value for the normalized range, and

NewMax is the maximum value for the normalized range.

- **Minimum** and **maximum** values for the original variable **are needed**
  - If the original data does not contain the full range, either an estimate of the range is needed or the formula should be restricted to the range specified for future use.

GENERATING CONSISTENT SCALES ACROSS VARIABLES.

- The **z-score** transformation normalizes the values around the mean of the set, with differences from the mean being recorded as standardized units, based on the frequency distribution of the variable
- The **decimal scaling** transformation moves the decimal point to ensure the range is between 1 and −1. The following formula (shown on the right) is used:

$$x_i' = \frac{x_i}{10^n}$$

where **n** is the number of digits of the maximum absolute value

TABLE 3.1 Normalization of the Variable Weight Using the Min–Max, *z*-score, and Decimal Scaling Transformations

Car Name	Weight	Min–Max (Weight)	z-score (Weight)	Decimal Scaling (Weight)
Datsun 1200	1613	0	−1.59	0.161
Honda Civic Cvcc	1800	0.053	−1.37	0.18
Volkswagen Rabbit	1825	0.0601	−1.34	0.182
Renault 5 gtl	1825	0.0601	−1.34	0.182
Volkswagen Super Beetle	1950	0.0955	−1.19	0.195
Mazda glc 4	1985	0.105	−1.15	0.198
Ford Pinto	2046	0.123	−1.08	0.205
Plymouth Horizon	2200	0.166	−0.898	0.22
Toyota Corolla	2265	0.185	−0.822	0.226
AMC Spirit dl	2670	0.3	−0.345	0.267
Ford Maverick	3158	0.438	0.229	0.316
Plymouth Volare Premier v8	3940	0.66	1.15	0.394
Dodge d200	4382	0.785	1.67	0.438
Pontiac Safari (sw)	5140	1	2.56	0.514

Exercise Time: Use Min-Max Transformation method.

Table 3.5.

Variable
33
21
7
53
29
42
12
19
22
36

CONVERTING TEXT TO NUMBERS.

- To use variables assigned as ***nominal or ordinal*** and described using **text values** within certain numerical analysis methods, it is necessary to **convert the variable's values into numbers**.
- For example, a variable with values **"low," "medium" and "high**" may be replaced by **0, 1, and 2.**
- Another way to **handle nominal data:**
  - To convert each value into a separate column with values **1** (**presence** of the category) and **0** (**absence** of the category). These new variables are often referred to as **dummy variables**.

TABLE 3.2 Generating a Series of Dummy Variables from the Single *Color* Variable

Product ID	Color	Color = Black	Color = Blue	Color = Red	Color = Green	Color = White
89893-22	Black	1	0	0	0	0
849082-35	Blue	0	1	0	0	0
27037-84	Red	0	0	1	0	0
2067-09	Green	0	0	0	1	0
44712-61	White	0	0	0	0	1
98382-34	Blue	0	1	0	0	0
72097-52	Green	0	0	0	1	0

CONVERTING TEXT TO NUMBERS.

- In Table 3.2, for example, the variable **Colour** has now been divided into **five separate columns**, one for each colour.
- A column for each color, orcfive dummy variables to encode the five colors.
- We could get by with only **four variables** (Color = “Black,” Color = “Blue,” Color = “Red,” and Color = “Green”).
- To represent the five colors, the values for the “Black,” “Blue,” “Red,” and “Green” variables would be for Black: 1,0,0,0, for Blue: 0,1,0,0, for Red: 0,0,1,0, for Green: 0,0,0,1, and for White: 0,0,0,0.

CONVERTING CONTINUOUS DATA TO CATEGORIES.

- **Converting continuous data** into **discrete values**
- Appears as **reduction in information**
- NOT the case
- Conversion is **desirable** in a number of situations:
  1. Where a **value is defined on an interval or ratio scale** but when **knowledge** about how the data was collected suggests the accuracy of the data **does not warrant these scales**, a variable may be a candidate for conversion to a categorical variable that reflects the true variation in the data.
     - This variable may be a **candidate** for **discretization** - often referred to as **data smoothing**
  2. Certain techniques can only process categorical data, converting continuous data into discrete values makes a numeric variable accessible to these methods.

CONVERTING CONTINUOUS DATA TO CATEGORIES.

- A continuous variable **credit score:** poor, average, good, and excellent (4 categories)
- **Weight** (ranges from 0 to 350 lb): less than 100 lb, 100–150 lb, 150–200 lb, 200–250 lb, and above 250 lb (5 categories)
  - All values for Weight assigned to a category and assigned an appropriate value such as the mean of the assigned category.
  - It is often useful to use the frequency distribution to understand appropriate range boundaries.
- This type of conversion is also called **binning.**
- Often useful to *use the frequency distribution* to understand appropriate binning cut-offs.

CONVERTING CONTINUOUS DATA TO CATEGORIES

- Can be **applied to nominal variables**, especially in situations where there are a large number of values for a given nominal variable.
- If **data set is summarized** using each of the values, **number of observations** for **each value** may be **too small** to reach any **meaningful conclusions.**
- A **new variable generated to** *generalize values using a mapping of terms.*
- For example:
  - **Customer transactions** with variable **Company** that details the individual customer's company.
  - **Handful** of observations for **each company.**
  - Map this variable to a new variable, **Industries**.
  - Mapping of specific companies onto generalized industries must be defined using a **concept mapping** (i.e., which company maps onto which industry)

COMBINING VARIABLES

- **Variable** to be used may **not be present** in the **data set.**
- May be **derived** from **existing variables**.
- Mathematical operations, such as **average or sum**, could be applied to one or more variables in order to **create an additional variable**.
- For example
  - A project may be trying to understand issues regarding a particular car's fuel efficiency **(Fuel Efficiency)** using a data set of different journeys in which the fuel level at the start **(Fuel Start)** and end **(Fuel End)** of a trip is measured along with the distance covered **(Distance).**

COMBINING VARIABLES

- An additional column may be calculated using the following formula:

  Fuel Efficiency = (Fuel End − Fuel Start)/Distance

GENERATING GROUPS.

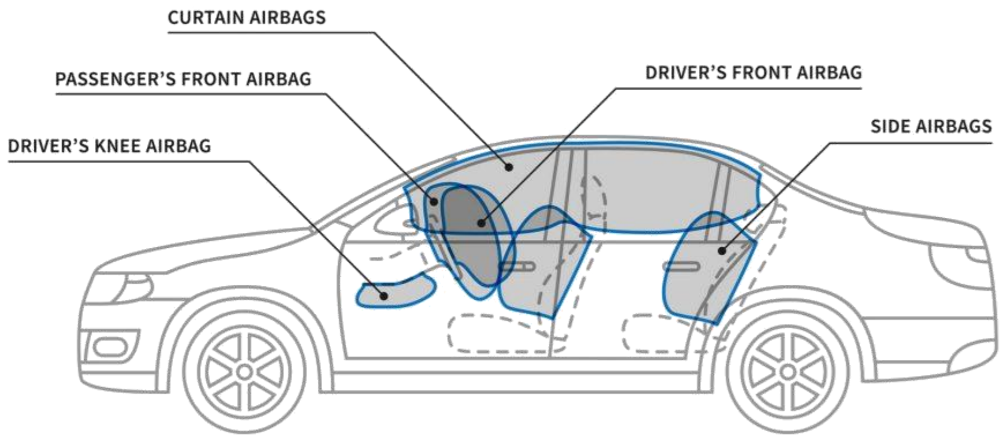
- **Larger data sets** take **more computational time** to analyze.
  - **Creating subsets or segmentation.**
- **One approach** is to take a **random subset**
  - **Effective** where the **data set closely matches** the **target population**.
- *Reasons* for creating subsets:

1. Speed up the analysis
2. When a *data set* **built** up over time for **operational purposes**, but is now used to **answer an alternative business research question**. It may be **necessary to select a diverse set** of observations that more **closely matches** the **new target population.**

GENERATING GROUPS Example.

Suppose a **car safety organization** has been **measuring** the **safety** of **individual cars** based on **specific requests** from the **government.**

Over time, the government may have requested car safety studies for certain types of vehicles. If the historical data set is to be used to answer questions on the safety of all cars, this data set does not reflect the new target population. However, a subset of the car studies could be selected to represent the more general questions now being asked of the data.



GENERATING GROUPS.

Divide the data set up into smaller sets based on location and model these locales separately

3. When **building predictive models** from a data set, it is important to keep the **models as simple** as possible (breaking the data set down into subsets may allow creation of several simpler models)

- For example, a project **to model factors** that contribute to the **price of real estate** may use a *data set of nationwide house prices and associated factors.*
  - **Divide** the data into **smaller sets** based on **location** and to model these locales separately.
- Factors contributing to housing prices:
  - Contingent upon the area in which the house is located.
  - House prices in coastal locations are different than that in the mountains.

Segmentation

PREPARING UNSTRUCTURED DATA.

- In many disciplines, focus of a data analysis is not a **simple data table** of observations and variables.
  - For example, in the **life sciences**, the focus of the analysis is **genes, proteins**, **biological pathways**, and **chemical structures**.
- In other disciplines, the focus of the analysis could be **documents**, **web logs**, **device readouts**, **audio or video information**, and so on.
  - In the analysis of these types of data, a **preliminary step** is often the **computational generation** of **different attributes** *relevant* to the problem.
  - For example, when analyzing a data set of chemicals, an initial step is to generate variables based on the composition of the chemical such as its molecular weight or the presence or absence of molecular components.

Table 3.8. Summary of the steps when preparing data

Steps	Details
1. Create data table	• Query databases to access data • Integrate multiple data sets and format as a data table
2. Characterize variables	Characterize the variables based on: • Continuous/discrete • Scales of measurement • Roles in analysis • Frequency distribution
3. Clean data	Clean the data: • Consolidate observations by merging appropriate terms • Identify potential errors (outliers, non-numeric characters, etc.) • Appropriately set nonnumeric values (or remove) • Ensure measurements are taken over the same scale • Remove duplicate observations
4. Remove variables	Remove variables that will not contribute to any analysis (e.g., constants or variables with too few values)
5. Transform variables	Transform the variable, if necessary, retaining how the variable was transformed using the following operations: • Normalize • Value mapping • Discretization • Aggregation
6. Segment table	Create subsets of the data to: • Facilitate more rapid analysis • Simplify the data set to create simpler models • Answer specific questions

Roles of variables in Analysis.

- **Labels:** Variables that describe individual observations in the data.
- **Descriptors:** Variables collected to **describe an observation**.
  - Used as the **input** or descriptors to be used in **both creating a predictive model and generating predictions** from these models. They are also described as **Independent variables**
- **Response:** Variables (*usually one variable*) are predicted from a predictive model (using the descriptor variables as input).
  - Used to **guide the creation of the predictive model**. They will *also be predicted*, based on the input descriptor variables that are presented to the model. They are also **Dependent variables**


BACK from THE BREAK.

Table 3.1. Example of a table describing cars

VIN	Manufacturer	Weight	Number of cylinders	Fuel efficiency
IM8GD9A_KP042788	Ford	2984	6	20
IC4GE9A_DQ1572481	Toyota	1795	4	34

1. VIN
2. Manufacturer
3. Weight
4. Number of cylinders
5. Fuel efficiency

? To build a model to predict fuel efficiency.

Labels	X variables	Y variables
VIN	Manufacturer	Fuel efficiency
	Weight	
	Number of cylinders	

Name	MPG	Cyclinders	Displacement	Horsepower	Weight	Acceleration
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

1. Name
2. MPG
3. Cylinders
4. Displacement
5. Horsepower
6. Weight
7. Acceleration

? To build a model to predict MPG

Labels	X variables	Y variables
Name	Cylinders Weight	MPG
	Displacement	
	Horsepower Acceleration	

Histograms.

Using the data in Table 2.6, create a histogram of Sale Price($) using the following intervals:
0 to less than 250
250 to less than 500
500 to less than 750
750 to less than 1000.

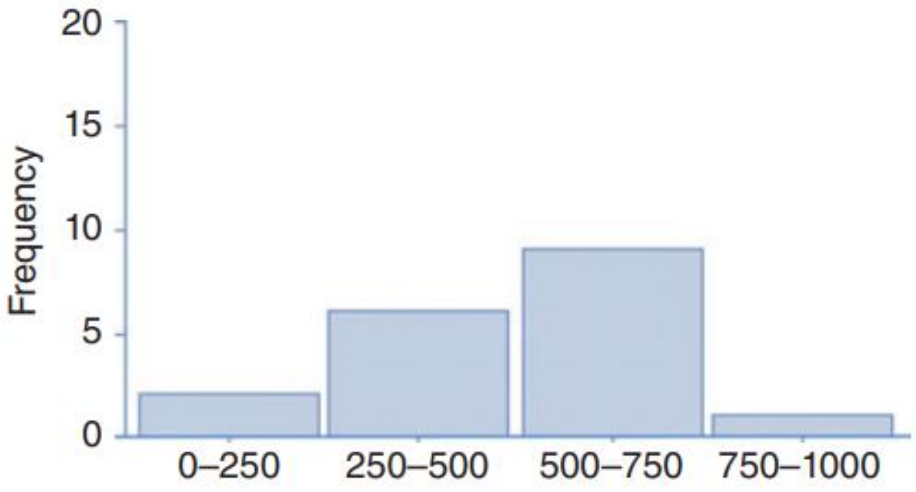


TABLE 2.6 Retail Transaction Data Set

Customer	Store	Product Category	Product Description	Sale Price ($)	Profit ($)
B. March	New York, NY	Laptop	DR2984	950	190
B. March	New York, NY	Printer	FW288	350	105
B. March	New York, NY	Scanner	BW9338	400	100
J. Bain	New York, NY	Scanner	BW9443	500	125
T. Goss	Washington, DC	Printer	FW199	200	60
T. Goss	Washington, DC	Scanner	BW39339	550	140
L. Nye	New York, NY	Desktop	LR21	600	60
L. Nye	New York, NY	Printer	FW299	300	90
S. Cann	Washington, DC	Desktop	LR21	600	60
E. Sims	Washington, DC	Laptop	DR2983	700	140
P. Judd	New York, NY	Desktop	LR22	700	70
P. Judd	New York, NY	Scanner	FJ3999	200	50
G. Hinton	Washington, DC	Laptop	DR2983	700	140
G. Hinton	Washington, DC	Desktop	LR21	600	60
G. Hinton	Washington, DC	Printer	FW288	350	105
G. Hinton	Washington, DC	Scanner	BW9443	500	125
H. Fu	New York, NY	Desktop	ZX88	450	45
H. Taylor	New York, NY	Scanner	BW9338	400	100

Table 3.9. Table of patient records

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh^+	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh^-	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^+	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh^-	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh^-	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^-	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^+	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh^-	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh^+	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh^+	136	1.78	121	165	98.7	1

A set of 10 hypothetical patient records from a large database is presented in Table 3.9. Patients with a diabetes value of 1 have type-II diabetes and patients with a diabetes value of 0 do not have type-II diabetes:

1. Create a new column by normalizing the Weight (kg) variable into the range 0–1 using the min–max normalization.
2. Create a new column by binning the Weight (kg) variable into three categories: low (less than 60 kg), medium (60–100 kg), and high (greater than 100 kg).
3. Create an aggregated column, body mass index (BMI), which is defined by the formula: $\text{BMI} = \dfrac{\text{Weight(kg)}}{(\text{Height(m)})^2}$
4. Segment the data into data sets based on values for the variable Gender.

1. Create a new column by normalizing the Weight (kg) variable into the range 0–1 using the min–max normalization.

Name	Weight (kg)	Weight (kg) normalized to 0–1
P. Lee	50	0.095
R. Jones	115	0.779
J. Smith	96	0.579
A. Patel	41	0
M. Owen	79	0.4
S. Green	109	0.716
N. Cook	73	0.337
W. Hands	104	0.663
P. Rice	64	0.242
F. Marsh	136	1

2. Create a new column by binning the Weight (kg) variable into three categories: low (less than 60 kg), medium (60–100 kg), and high (greater than 100 kg).

Name	Weight (kg)	Weight (kg) categorized [low, medium, high]
P. Lee	50	low
R. Jones	115	high
J. Smith	96	medium
A. Patel	41	low
M. Owen	79	medium
S. Green	109	high
N. Cook	73	medium
W. Hands	104	high
P. Rice	64	medium
F. Marsh	136	high

3. Create an aggregated column, body mass index (BMI), which is defined by the formula: $$\text{BMI} = \frac{\text{Weight(kg)}}{(\text{Height(m)})^2}$$

Name	Weight (kg)	Height (m)	BMI
P. Lee	50	1.52	21.6
R. Jones	115	1.77	36.7
J. Smith	96	1.83	28.7
A. Patel	41	1.55	17.1
M. Owen	79	1.82	23.8
S. Green	109	1.89	30.5
N. Cook	73	1.76	23.6
W. Hands	104	1.71	35.6
P. Rice	64	1.74	21.1
F. Marsh	136	1.78	42.9

4. Segment the data into data sets based on values for the variable Gender.

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Tempe-rature (°F)	Diabetes
P. Lee	35	Female	A Rh^{+}	50	1.52	68	112	98.7	0
A. Patel	70	Female	O Rh^{-}	41	1.55	76	125	98.6	0
W. Hands	77	Female	O Rh^{-}	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh^{+}	64	1.74	101	132	98.6	0

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
R. Jones	52	Male	O Rh^{-}	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^{+}	96	1.83	88	136	98.8	0
M. Owen	24	Male	A Rh^{-}	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^{-}	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^{+}	73	1.76	108	136	99.0	0
F. Marsh	28	Male	O Rh^{+}	136	1.78	121	165	98.7	1


TAKE A
BREAK.