Data Analytics
DS342

Data Analytics
DS342

Course Instructors:

Dr.Marwa Sabry ,

Dr.Hayam Wahdan

Course TA.:

Eng.Yousra Ayman

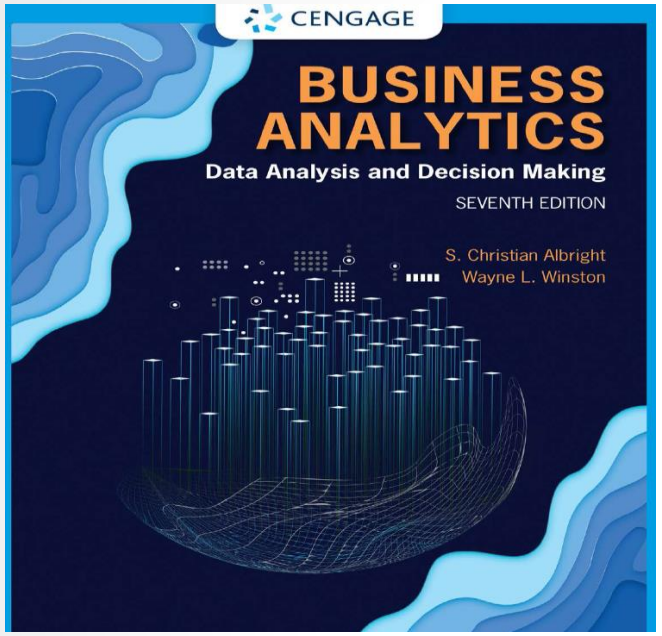
WELCOME TO THE COURSE ☺

- Mutual respect
  - Time
  - Noise
  - Section
  - Honest (especially assignments)
  - **Mobile phones Silent**

Grading Schema

- Final exam 60%
- Mid-term exam 20%
- Assignments (4 assignments)10%
- Quizzes (2 quizzes) 10%

Course Content



The Essence of the Course

The overall goal of this course is to:

Understand data analytics and be able to apply data analysis to data sets using a variety of software tools and techniques

This course will provide the tools for you to perform your own data analysis when encountering problems in the real-world.

Course Objectives

1. Understand data representation formats and techniques and how to use them.
2. Experience a wide-range of data analytics tools include Excel, Power Query, visualization and reporting software.
3. Develop a computational thinking approach to problem solving and use programs and scripting to solve data tasks.
4. Be able to clearly articulate a problem in a systematic way .

What is Data Analysis?

Data analysis is the **processing** of data to yield useful insights or knowledge.

- Data processing involves finding, loading, cleaning, manipulating, transforming, modeling, and visualizing the data.

- The knowledge may be used for scientific discovery, business decision-making, or a variety of other applications.

A ***data analyst*** is a person who uses tools and applications to transform raw data into a form that will be useful.

- Data analyst jobs are projected to be one of the top jobs over the next 10 years.
  - See: http://blog.udacity.com/2014/11/data-analysts-what-youll-make.html

Why is Data Analytics Important?

Data analytics is important as society is collecting more and larger data sets all the time:

- Web: All web pages visited and links clicked, searches made, images and posts
- Business: Items purchased by date, supply chain/customers, industrial sensors
- Science: Massive data sets (biological/genomic, astronomy, physics)
- Environmental: Sensors and monitors (temperature, etc.)

and transforming this raw data into useful insights has major value:

- Web: Online advertising driven by understanding customer behavior
- Business: Sales predictions, marketing promotions, manufacturing improvement
- Science: Scientific discoveries, new medical treatments and drugs
- Environmental: Understanding of environmental processes to allow for changing policies and behaviors

Data Analytics Tools

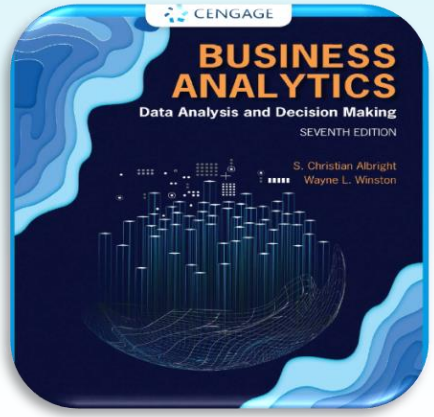
- ✓ A data analyst has expertise in programming, statistics, data collection and data visualization.
- ✓ In this course, you will learn industrial tools and build competency in each one of these skills.
- ✓ As an introductory course, the goal is to get exposure to the skills and techniques as there will not be time for mastery.
- ✓ These tools of systems and techniques will be useful in many jobs even if they are not considered data analyst positions.

Why This Course is Important

- Many professional jobs of the future will involve collecting, manipulating, and analyzing data.
- People who can understand how data can be used will have better employment opportunities.

Important results:

- Excel Proficiency – Everyone should know how to use Excel as a general data analysis and productivity software.
- Databases – Understand how they work and how to use them.
- Programming and Computational Thinking – The ability to clearly articulate a problem in a systematic way has applications beyond data analytics.
- Applied Statistics – Using R and other software makes your statistics training useful for real-world problems.
- Real-world problem solving – Your tools will allow you to tackle real-world data analysis problems and understand what tool to use and how to proceed.



Chapter 1

Introduction to Business Analytics

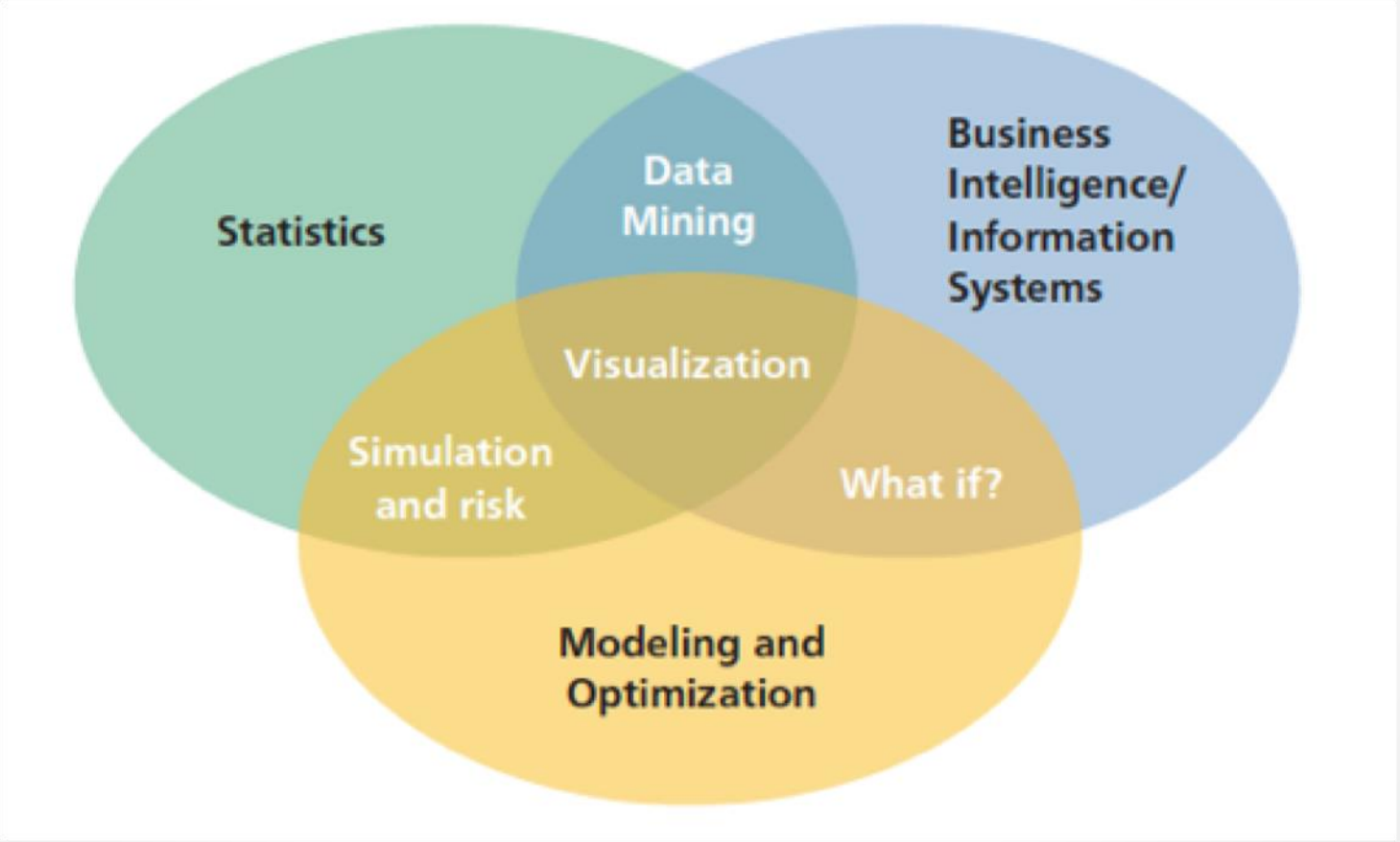
Business Analytics

(Business) Analytics is the use of:

- data,
- information technology,
- statistical analysis,
- quantitative methods, and
- mathematical or computer-based models

to help managers gain improved insight about their business operations and make better, fact-based decisions.

A Visual Perspective of Business Analytics



Evolution of Business Analytics

- Business intelligence
- Information Systems
- Statistics
- Operations research/Management science
- Decision support systems

Overview of Business Analytics

- Business analytics begins with understating the business context.
  - Ask the right questions
  - Identify the appropriate analysis
  - Communicate information
- Numerical results are not very useful unless they are accompanied with clearly stated actionable business insights.

Scope of Business Analytics

- **Descriptive analytics**: the use of data to understand past and current business performance and make informed decisions.
- **Predictive analytics**: predict the future by examining historical data, detecting patterns or relationships in these data, and then extrapolating these relationships forward in time.
- **Prescriptive analytics**: identify the best alternatives to minimize or maximize some objective.

Example 1.1: Retail Markdown Decisions

- Most department stores clear seasonal inventory by reducing prices.
- *Key question*: When to reduce the price and by how much to maximize revenue?
- Potential applications of analytics:
  - Descriptive analytics: examine historical data for similar products (prices, units sold, advertising, …)
  - Predictive analytics: predict sales based on price
  - Prescriptive analytics: find the best sets of pricing and advertising to maximize sales revenue

Tools

- Database queries and analysis
- Dashboards to report key performance measures
- Data visualization
- Statistical methods
- Spreadsheets and predictive models
- Scenario and “what-if” analyses
- Simulation
- Forecasting
- Data and text mining
- Optimization
- Social media, web, and text analytics

Data for Business Analytics

- **Data**: numerical or textual facts and figures that are collected through some type of measurement process.
- **Information**: result of analyzing data; that is, extracting meaning from data to support evaluation and decision making.

Data Sets and Databases

- **Data set** - a collection of data.
  - Examples: Marketing survey responses, a table of historical stock prices, and a collection of measurements of dimensions of a manufactured item.
- **Database** - a collection of related files containing records on people, places, or things.
  - A database file is usually organized in a two-dimensional table, where the columns correspond to each individual element of data (called *fields*, or *attributes*), and the rows represent records of related data elements.

Example 1.2: A Sales Transaction Database File

	A	B	C	D	E	F	G	H
1	**Sales Transactions: July 14**							
2								
3	**Cust ID**	**Region**	**Payment**	**Transaction Code**	**Source**	**Amount**	**Product**	**Time Of Day**
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

Records
(Observations)

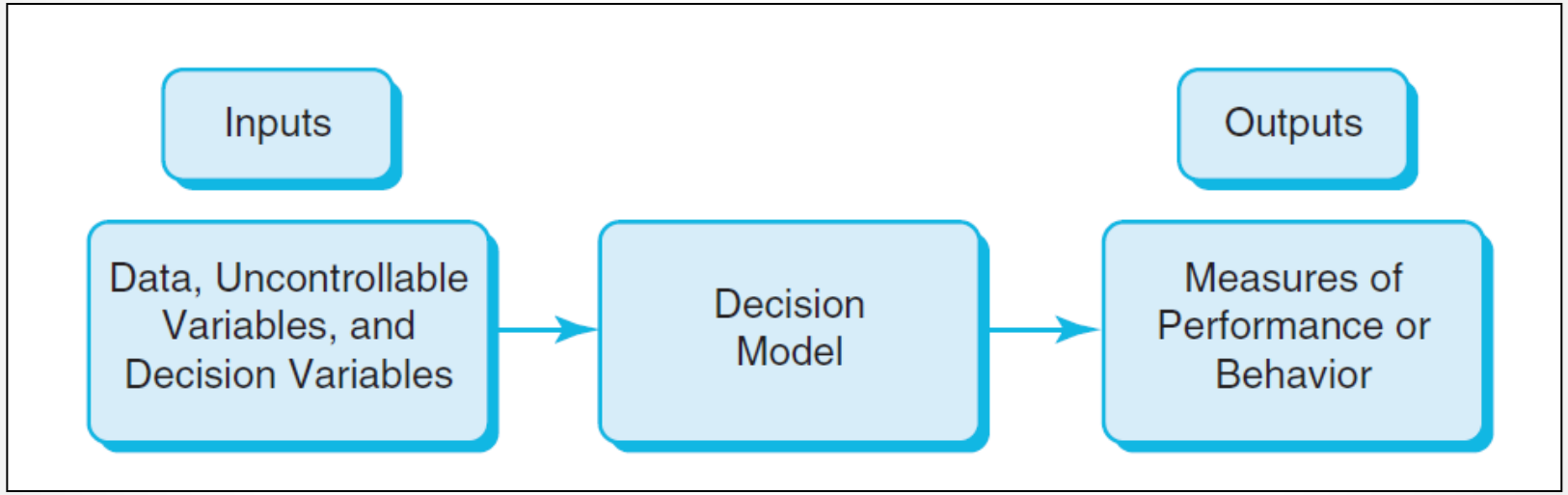
Entities
(Elements)

Fields or Attributes
(Variables)

Decision Models

- **Decision model** - a logical or mathematical representation of a problem or business situation that can be used to understand, analyze, or facilitate making a decision.
- Inputs:
  - *Data*, which are assumed to be constant for purposes of the model.
  - *Uncontrollable variables*, which are quantities that can change but cannot be directly controlled by the decision maker.
  - *Decision variables*, which are controllable and can be selected at the discretion of the decision maker.

Nature of Decision Models



Spreadsheet Models

- Spreadsheet modeling is an alternative to algebraic modeling that relates various quantities in a spreadsheet with cell formulas.
  - Instant feedback is available from spreadsheets, so if a formula is entered incorrectly, it is often immediately obvious.
  - Developing good spreadsheet models is not easy.
    - They must be *correct*, well designed and well documented.

Spreadsheet Models

- A spreadsheet model for a specific example of the product mix problem is shown below.

	A	B	C	D	E	F	G
1	**Assembling and testing computers**				**Range names used:**		
2					Hours_available	=Model!$D$21:$D$22	
3	Cost per labor hour assembling	$11			Hours_used	=Model!$B$21:$B$22	
4	Cost per labor hour testing	$15			Maximum_sales	=Model!$B$18:$C$18	
5					Number_to_produce	=Model!$B$16:$C$16	
6	Inputs for assembling and testing a computer				Total_profit	=Model!$D$25	
7		Basic	XP				
8	Labor hours for assembly	5	6				
9	Labor hours for testing	1	2				
10	Cost of component parts	$150	$225				
11	Selling price	$300	$450				
12	Unit margin	$80	$129				
13							
14	Assembling, testing plan (# of computers)						
15		Basic	XP				
16	Number to produce	560	1200				
17		<=	<=				
18	Maximum sales	600	1200				
19							
20	Constraints (hours per month)	Hours used		Hours available			
21	Labor availability for assembling	10000	<=	10000			
22	Labor availability for testing	2960	<=	3000			
23							
24	Net profit ($ this month)	Basic	XP	Total			
25		$44,800	$154,800	$199,600			

Types of Data

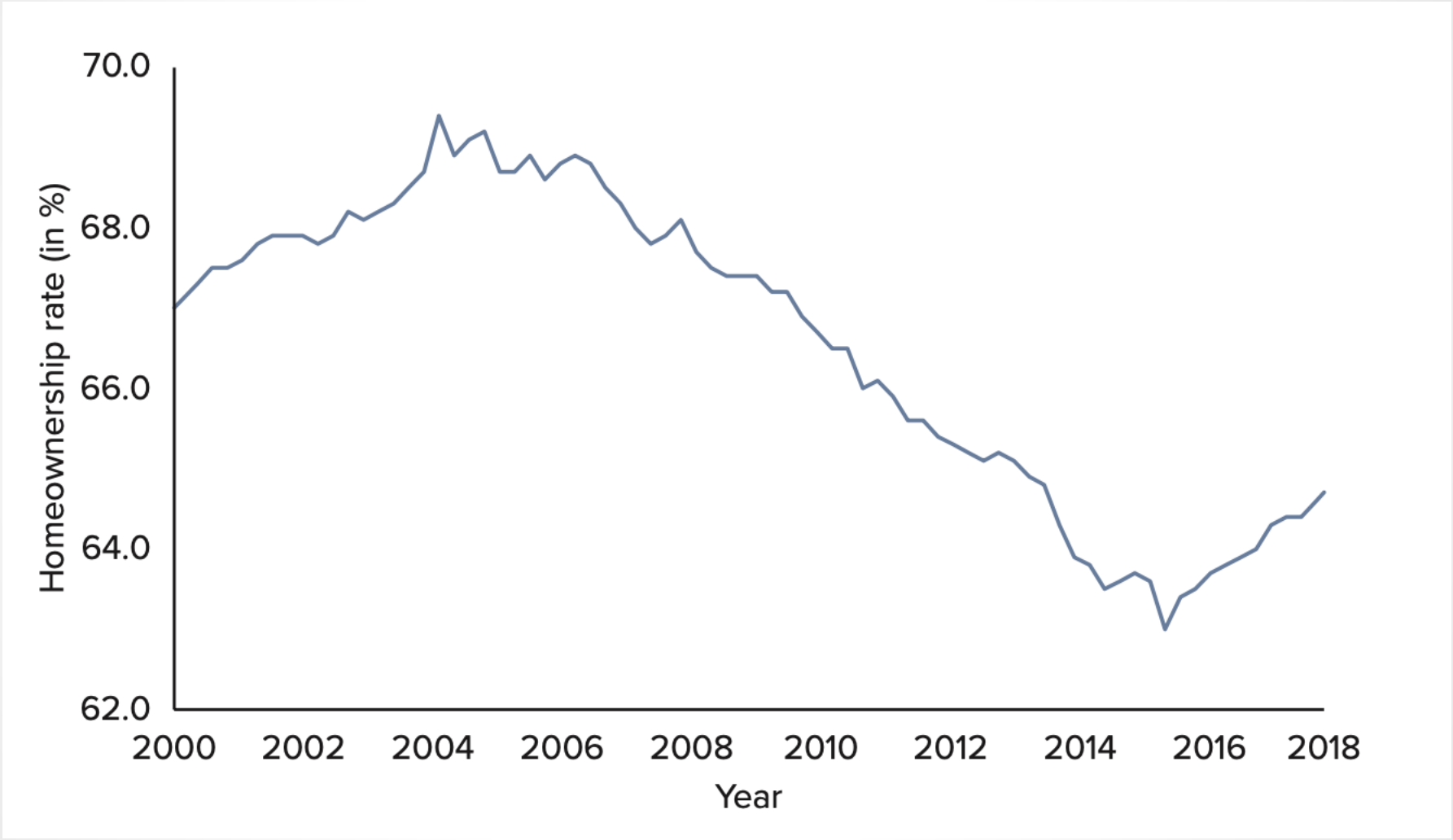
- Cross-sectional data
  - Collected by recording a characteristic of many subjects at the same point in time
  - Recording a characteristic of many subjects at the same point in time
- Time series data
  - Collected over several time periods focusing on certain groups of people, specific events, or objects
  - Hourly, daily, weekly, monthly, quarterly, or annual observations

Types of Data

Team name	Wins	Losses
Milwaukee Bucks	60	22
Toronto Raptors*	58	24
Philadephia 76ers	51	31
Boston Celtics	49	33
Indiana Pacers	48	34
Brooklyn Nets	42	40
Orlando Magic	42	40
Detroit Pistons	41	41

	A	B	C
1	Quarter	Revenue	
2	Q1-2010	1026	
3	Q2-2010	1056	
4	Q3-2010	1182	
5	Q4-2010	2861	
6	Q1-2011	1172	
7	Q2-2011	1249	
8	Q3-2011	1346	
9	Q4-2011	3402	
10	Q1-2012	1286	
11	Q2-2012	1317	
12	Q3-2012	1449	
13	Q4-2012	3893	
14	Q1-2013	1462	
15	Q2-2013	1452	
16	Q3-2013	1631	
17	Q4-2013	4200	

Types of Data



Variables and Scales of Measurement

- A variable is a characteristic of interest that differs in kind or degree among various observations (records).
- There are two types of variables: **categorical** and **numerical**

1. **Categorical**
   - Also called qualitative
   - Represent categories
   - Labels or names to identify distinguishing characteristics
   - Arithmetic operations on the labels/values are not meaningful
   - Coded into numbers for data processing

   Example : marital status

2. Numerical

- Also called quantitative
- Represent meaningful numbers
- Arithmetic operations are meaningful

a)Discrete: assumes a countable number of values

Example: number of children in a family

b)Continuous: assumes an uncountable number of values within an interval

Example: investment returns

Working Example :  Gig

- BalanceGig is a company that matches independent workers for short-term engagements with businesses in the construction, automotive, and high-tech industries.
- The ‘gig’ employees work only for a short period of time, often on a particular project or a specific task.
- A manager at BalanceGig extracts the employee data from their most recent work engagement, including: the following variables
  - ✓the hourly wage (**HourlyWage**),
  - ✓the client’s industry (**Industry**), and
  - ✓the employee’s job classification (**Job**).

Working Example : Gig

The manager would like to find:

1. Number of missing observations for the HourlyWage, Industry, and Job variables.
2. The number of employees who
   - ✓ worked in the automotive industry,
   - ✓ earned more than $30 per hour, and
   - ✓ worked in the automotive industry and earned more than $30 per hour.
3. The hourly wage of the lowest and the highest-paid employees at the company as a whole, and
4. The hourly wage of the lowest and the highest-paid accountants who worked in the automotive and the tech industries.

Working Example : Gig

1. There are a total of 604 records in the data set.
   - ✓There are no missing values in the HourlyWage variable.
   - ✓The Industry and Job variables have 10 and 16 missing values, respectively.
2. 190 employees worked in the automotive industry,
   - ✓536 employees earned more than $30 per hour, and
   - ✓181 employees worked in the automotive industry and earned more than $30 per hour.
3. The lowest and the highest hourly wages in the data set are $24.28 and $51.00, respectively.
4. The three employees who had the lowest hourly wage of $24.28 all worked in the construction industry and were hired as Engineer, Sales Rep, and Accountant, respectively.

- Interestingly, the employee with the highest hourly wage of $51.00 also worked in the construction industry in a job type classified as Other.

Working Example : Gig

4. The lowest- and the highest-paid accountants who worked in the automotive industry made $28.74 and $49.32 per hour, respectively.

In the technology industry, the lowest- and the highest paid accountants made $36.13 and $49.49 per hour, respectively.

- Note that the lowest hourly wage for an accountant is considerably higher in the technology industry compared to the automotive industry ($36.13 > $28.74).

Transforming Numerical Data

- Binning is the process of transforming numerical variables into categorical variables by grouping the numerical values into a small number of groups or bins.
- It is important that the bins are consecutive and nonoverlapping so that each numerical value falls into only one bin.
- Binning can be an effective way to reduce noise in the data if we believe that all observations in the same bin tend to behave the same way.

Transforming Numerical Data

- Data transformation is an important step in bringing out the information in the data set, which can then be used for further data analysis.
- Another common approach is to create new variables through mathematical transformations of existing variables.
- Similarly, in order to analyze trend, we often transform raw data values into percentages.
- Sometimes, data on variables such as income, firm size, and house prices are highly skewed.
  - Extremely high (or low) values of skewed variables significantly inflate the average for the entire data set
  - Difficult to detect meaningful relationships with skewed variables

Transforming Categorical Data

- An effective strategy for dealing with categorical data is category reduction, where we collapse some of the categories to create fewer nonoverlapping categories.
- Determining the appropriate number of categories often depends on the data, context, and disciplinary norms, but there are a few general guidelines.
- Categories with very few observations may be combined to create the “Other” category. The rationale behind this approach is that a critical mass can be created for this “Other” category to help reveal patterns and relationships in data.
- Categories with a similar impact may be combined.

Transforming Categorical Data

- Dealing with numerical data is often easier than categorical data because it avoids the complexities of the semantics pertaining to each category of the variable.
- A dummy variable, also referred to as an indicator or a binary variable, is commonly used to describe two categories of a variable.
  - It assumes a value of 1 for one of the categories and 0 for the other category, referred to as the reference or the benchmark category.
  - Dummy variables do not suggest any ranking of the categories.
- Oftentimes, a categorical variable is defined by more than two categories.
- Given k categories of a variable, the general rule is to create $k - 1$ dummy variables, using the last category as reference.

Transforming Categorical Data

- Another common transformation of categorical variables is to create category scores.
- This approach is most appropriate if the data are ordinal and have natural, ordered categories.
- This transformation allows the categorical variable to be treated as a numerical variable in certain analytical models.
- With this transformation, we need not convert a categorical variable into several dummy variables or to reduce its categories.
- For an effective transformation, however, we assume equal increments between the category scores, which may not be appropriate in certain situations.

Transforming Categorical Data

- **Example**: In customer satisfaction surveys, we often use ordinal scales such as very dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, and very satisfied to indicate the level of satisfaction.
- In such cases, we can recode the categories numerically using numbers 1 through 5 with 1 being very dissatisfied and 5 being very satisfied.

Thank You ☺