



1ST EDITION

# CompTIA Data+: DAO-001 Certification Guide

Complete coverage of the new CompTIA Data+  
(DAO-001) exam to help you pass on the first attempt

A stylized orange logo consisting of several parallel lines forming a series of nested, angular shapes that resemble a stylized 'P' or a series of connected chevrons.

CAMERON DODD

# Preface

## Download the example code files

You can download the example code files for this book from GitHub at

<https://github.com/PacktPublishing/CompTIA-Data-DAO-001-Certification-Guide>. If there's an update to the code, it will be updated in the GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at

<https://github.com/PacktPublishing/>. Check them out!

# Chapter 1

## Table

Data Concepts and Environments	Data Mining	Data Analysis	Visualization	Data Governance, Quality, and Control
15%	25%	23%	23%	14%

**Table 1.1 – Percentage breakdown of each domain**

## Links

For the most up-to-date information on the exam, you can check out the exam website at

<https://www.comptia.org/certifications/data>

# Chapter 2

## Figures

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	3	5	3

Figure 2.1 – Structured database: defined rows and columns

	Structured	Unstructured
Relational	SQL Databases Some NoSQL (Key-Value Pairs)	Some NoSQL (Graph Databases)
Non-Relational	Nothing	Some NoSQL (Buckets o' Data)

Figure 2.2 – Structure and relationships

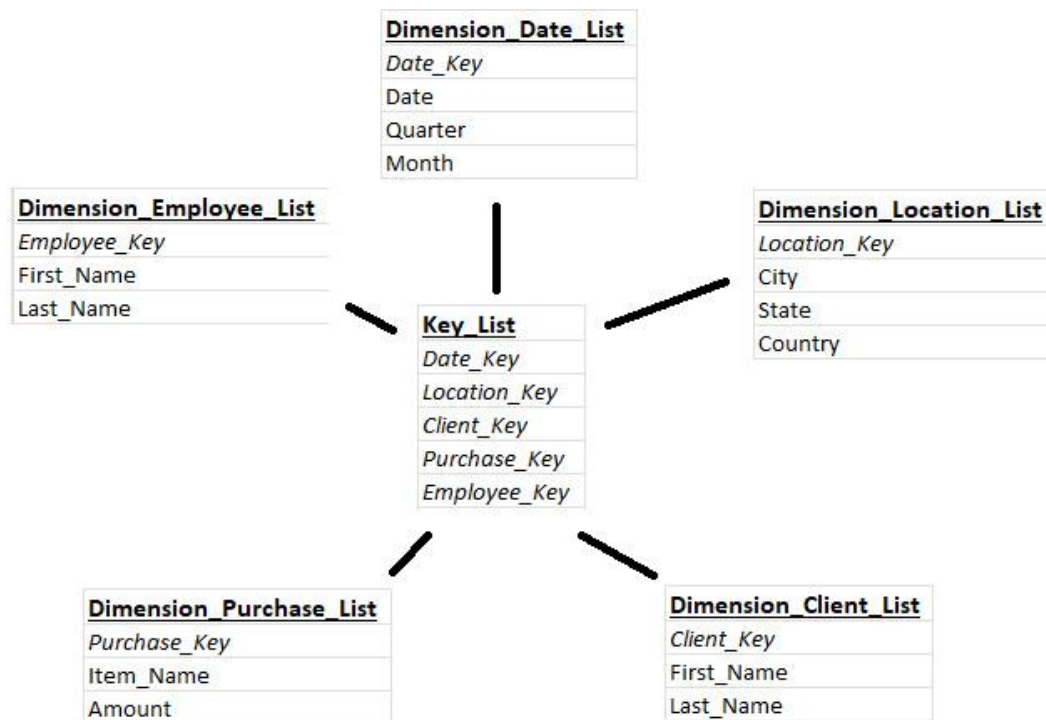


Figure 2.3 – Star schema

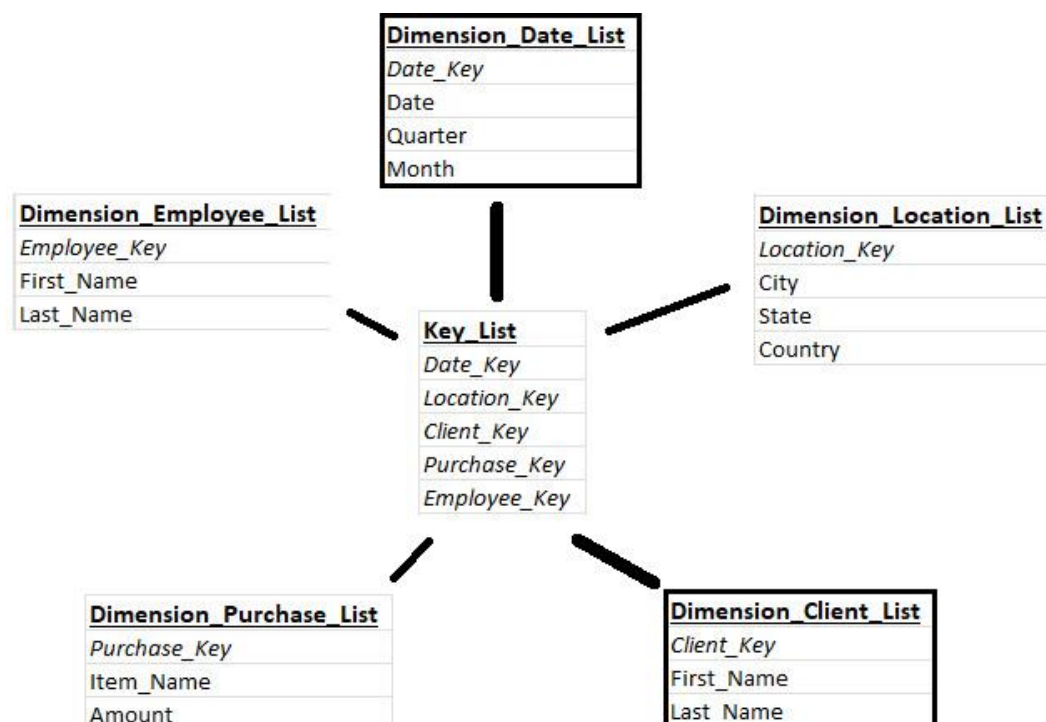


Figure 2.4 – Joining with a star schema

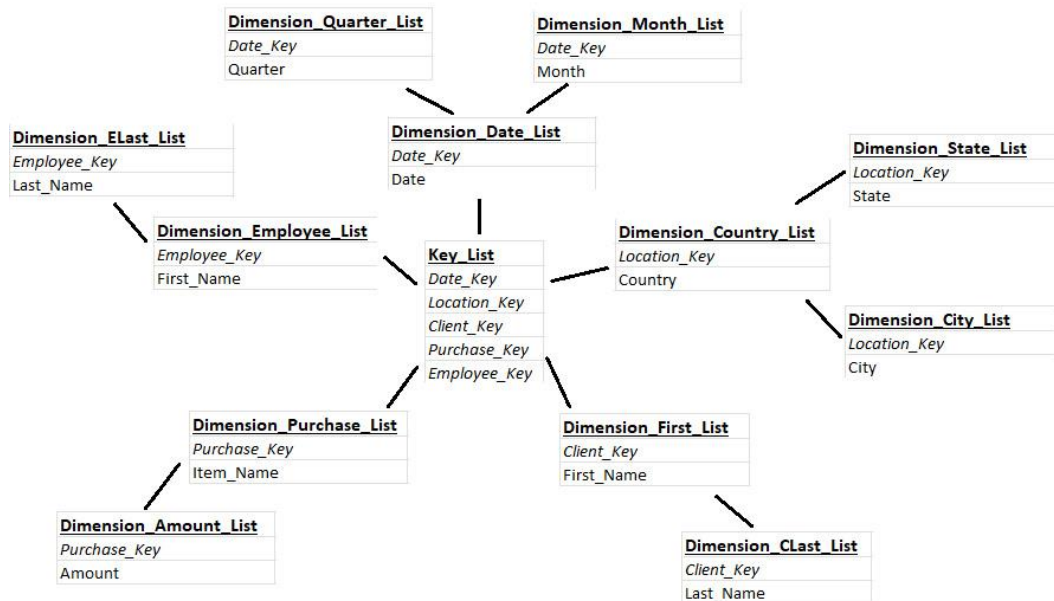


Figure 2.5 – Snowflake schema

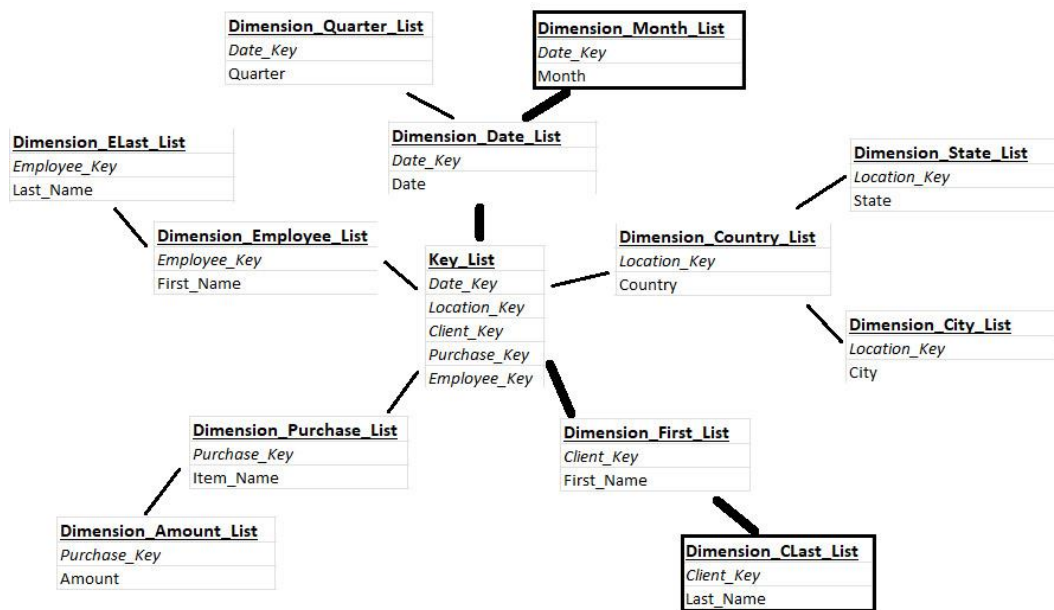


Figure 2.6 – Joining within a snowflake schema

Magic Number	Active Record	Active Start	Active End
41	Yes	11/11/2011	

Figure 2.7 – Active Record

Magic Number	Active Record	Active Start	Active End
41	No	11/11/2011	12/12/2012
42	Yes	12/12/2012	

Figure 2.8 – Updated Active Record

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	X	X	X
12	X	X	X
11	3	5	3
9	2	6	1
10	4	3	3

Figure 2.9 – Adding variables

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	X	X	X
9	X	X	X
10	X	X	X

Figure 2.10 – Removing variables

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	3	5	3

Figure 2.11 – Deleting historical values

TSV values are separated by tabs. Here's an example:

Column1   Column2   Column3

CSV values are separated by commas. Here's an example:

Column1, Column2, Column3

## Code

### Code 2.1:

In the following code snippet, each key represents a variable, and each value is the data collected for it, while each object is a data point:

```
"Beans" : [
  {
    "Total" : 10,
    "Red" : 3,
    "Blue" : 4,
```

```

        "Yellow" : 3
    },
    {
        "Total" : 12,
        "Red" : 2,
        "Blue" : 6,
        "Yellow" : 4
    },
    {
        "Total" : 11,
        "Red" : 3,
        "Blue" : 5,
        "Yellow" : 3
    }
]

```

## Code 2.2:

The tags create elements that all have specific pre-determined meanings and act in specific ways when used. Here's an example:

```

<div>
    <h1>
        Store Data Here
    </h1>
<p>
    Or Here
</p>
</div>

```

## Code 2.3:

Here's an example of XML:

```

<Dataset>
    <Data>
        Store Data Here
    </Data>
<AlsoData>
    Or Here
</AlsoData>
</Dataset>

```

## Code 2.4

Here's an example of JSON:

```

"Dataset" : [
    {
        "Data" : "Store Data Here"
    },
    {
        "Data" : "Or Here"
    }
]

```



```
}  
]
```

## Practice questions and their answers

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. A smart thermometer collects information about the temperature outside every 30 minutes, creates a log, and sends the data to a local database. What can you tell about the database?
  - A. It is structured
  - B. It is unstructured
  - C. It is relational
  - D. There is not enough information
2. Client-facing agents at banks are not technical experts, but require the ability to query client information. Which data schema is most appropriate for their database?
  - A. Star schema
  - B. Snowflake schema
  - C. Galaxy schema
  - D. Avalanche schema
3. You are working as a data scientist for a video-streaming website. You require access to raw, unprocessed data and video files. What is the most appropriate database style to use?
  - A. Data warehouse
  - B. Data mart
  - C. Data lake
  - D. Data mine
4. An e-commerce website has been collecting information on purchases for over a year. Now, they want more detailed geographical information to find out where their products are selling, so they start collecting information on the IP address of the person who made the purchase. This adds a new column to the dataset. What concerns might they have going forward?
  - A. Historic values for the IP Address column will be null
  - B. New values for the IP Address column will be null

- C. They will not know which record is active
  - D. None of these
5. You are given a file with the “.png” extension at the end to process. What type of data will you find inside?
- A. Video
  - B. Audio
  - C. Text
  - D. Image

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is ***B. It is unstructured***

The question tells you that the data is automatically generated and logged by a machine. This makes it machine data, and machine data is one of the two types of data that are inherently unstructured.

2. The answer is ***A. Star schema***

A star schema is simple and focuses more on being user-friendly, making it ideal for data marts and use by less technical employees.

3. The answer is ***C. Data lake***

A data lake is the only database style discussed that focuses on raw data. It is also the only one that can easily store structured and unstructured data or is meant for use by a data scientist.

4. The answer is ***A. Historic values for the IP Address column will be null***

Because you are adding a new column, you will not have any values in that column from before you added it, making all historic values null by default.

5. The answer is ***D. Image***

PNG stands for Portable Network Graphics and PNG files contain images.

# Chapter 3

## Figures

1. What is your favorite kind of peanut butter? ⓘ 0

**Figure 3.1 – Text-based survey question**

2. What is your favorite kind of peanut butter? ⓘ 0

☐ Smooth

☐ Chunky

**Figure 3.2 – Single-choice survey question**

3. What is your favorite kind of peanut butter? ⓘ 0

☐ Smooth

☐ Chunky

**Figure 3.3 – Multiple-choice survey question**

4. What is your favorite kind of peanut butter? ⓘ 0

**Figure 3.4 – Drop-down survey question**

5. Chunky peanut butter is, objectively, better. ⓘ 0

☐ Strongly agree

☐ Agree

☐ Neither agree nor disagree

☐ Disagree

☐ Strongly disagree

**Figure 3.5 – Likert survey question**

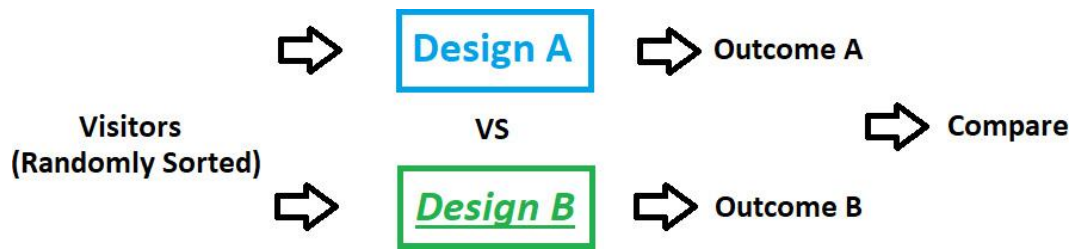


Figure 3.6 – A-B study for a website

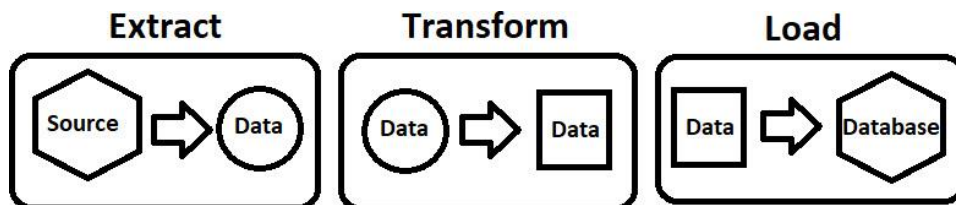


Figure 3.7 – ETL

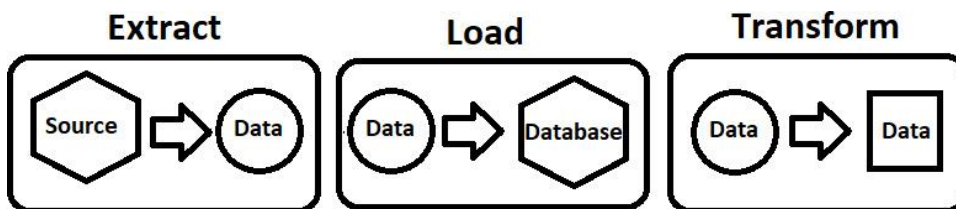


Figure 3.8 – ELT

Employee ID	LastName	FirstName	Department	YearsWithCompany
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sales	21



Employee ID	LastName	FirstName	Department	YearsWithCompany
83784	Benhill	Floyd	Sales	12
37816	Smith	Trudy	Sales	21

Figure 3.9 – Filtering

Index	Employee ID	LastName	FirstName	Department	YearsWithCompany
1	83784	Benhill	Floyd	Sales	12
2	64986	Chane	Jill	IT	1
3	93671	Hanson	Richard	HR	15
4	37816	Smith	Trudy	Sales	21



Index	Employee ID	LastName	FirstName	Department	YearsWithCompany
4	37816	Smith	Trudy	Sales	21
3	93671	Hanson	Richard	HR	15
1	83784	Benhill	Floyd	Sales	12
2	64986	Chane	Jill	IT	1

Figure 3.10 – Indexing and sorting

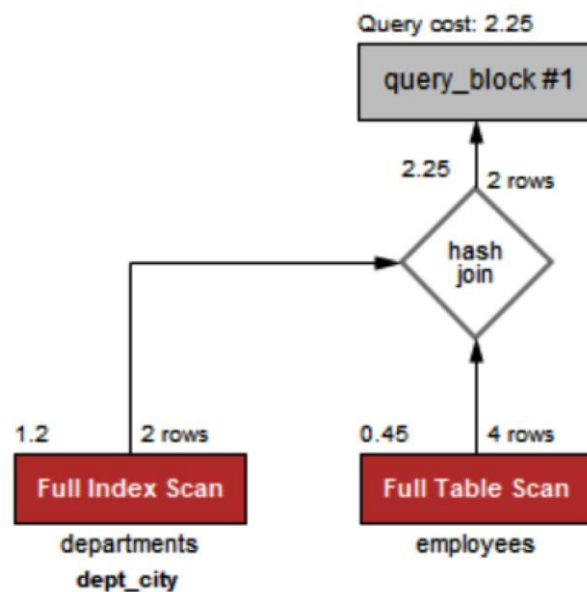


Figure 3.11 – Execution plan

## Code

### Code 3.1

An example of a temporary table might look something like this:

```
#Temporary Table
CREATE TEMPORARY TABLE EmployeeJones
SELECT * from Employees
WHERE LastName = 'Jones';
```

### Code 3.2:

Let's look at an example:

```
#Subquery
SELECT * from Employees
WHERE DeptID in (SELECT DISTINCT DeptID from Departments WHERE City
= 'Atlanta');
```

## Links

Twitter API: [developer.twitter.com/en/docs/twitter-api](https://developer.twitter.com/en/docs/twitter-api)

For more information on specific types of bias, check out [www.thedecisionlab.com/biases](http://www.thedecisionlab.com/biases). This is a great source that covers the most common kinds of bias and how to avoid them.

## Practice questions and their answers

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. If a web service is synchronous, it means...
  - A. Your system will wait for a response before continuing
  - B. Your system does not have to wait for a response
  - C. You are syncing a web service to your system
  - D. Web services cannot be synchronous
2. When conducting a survey, it is best practice to ask about specific events that happened in the past and request a text-based answer. True or false?
  - A. True
  - B. False
3. ETL stands for...
  - C. Extract, Transmit, Load
  - D. Estimate, Time, Load
  - E. Extract, Transform, Load
  - F. Exact, Time, Load
4. The process of taking transactional data that has already been stored, aggregating it, and moving it to a data warehouse is called what?

- A. OLTP
  - B. OLAP
  - C. API
  - D. Web service
5. You are provided with a long and complicated query that takes a long time to run, but you need several pieces of information from it. What is the most efficient approach?
- A. Run the query as it was given to you
  - B. Add subqueries to extract specific information from the query
  - C. Make sure no filters are slowing the process down
  - D. Save the results of the query to a temporary table

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is **A. Your system will wait for a response before continuing**

Web services are a type of API. Synchronous API calls mean that your code will make a request and then wait until it gets a response before continuing to the next step.

2. The answer is **B. False**

Requesting specific information from the past introduces recall bias and is especially bad when combined with a text-based answer. It is best practice to avoid asking about specific incidents in the past, if possible, or at least make it clear that they can give a non-answer such as “I don’t know” instead of making something up.

3. The answer is **C. Extract, Transform, Load**

ETL stands for Extract, Transform, Load because it is a data pipeline that performs these three steps in this specific order. It extracts the data, pulls it from its source, transforms it, or manipulates it into something that can be used, and finally loads it to its new location.

4. The answer is **B. OLAP**

OLAP is the process of taking information stored by OLTP, aggregating and analyzing it, and moving it to a new database or data warehouse.

5. The answer is **D. Save the results of the query to a temporary table**

By creating a temporary table, you can quickly and easily access anything from the results of the original query that you require without having to rerun it.





# Chapter 4

## Figures

Employee ID	LastName	FirstName	Department	Years With Company
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
64986	Chane	Jill	IT	1
64986	Chane	Jill	IT	1
93671	Hanson	Richard	H R	15
37816	Smith	Trudy	Sales	21

Figure 4.1 – Duplicated data

ID	Sex	Male	Female
84927	M	TRUE	FALSE
69427	M	TRUE	FALSE
10374	F	FALSE	TRUE
58264	M	TRUE	FALSE
90162	F	FALSE	TRUE

Figure 4.2 – Redundant data

```
MyData.head()
```

	ID	Sex	Male	Female
0	84930	M	True	False
1	75982	F	False	True
2	19038	F	False	True
3	48902	M	True	False
4	10948	F	False	True

```
MySubset = MyData.drop(["Male", "Female"], axis=1)  
MySubset.head()
```

	ID	Sex
0	84930	M
1	75982	F
2	19038	F
3	48902	M
4	10948	F

**Figure 4.3 – Creating a subset**

Employee ID	LastName	FirstName	Department	Years With Company
83784	Benhill	Floyd	Sales	
64986	Chane	Jill	1T	1
93671		Richard	HR	15
37816	Smith	Trudy	Sales	21
73891	Doe	John	IT	2
20179	Brown	Olivia		18
	Crow	Steven	Sales	14
74982	Burns	Charles	Sales	22

**Figure 4.4 – MCAR**

Employee ID	LastName	FirstName	Department	Years With Company
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sales	21
73891	Doe	John	IT	
20179	Brown	Olivia	H R	18
80781	Crow	Steven	Sales	14
74982	Burns	Charles	Sales	22

**Figure 4.5 – MAR**

Employee ID	LastName	FirstName	Department	Years With Company
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sales	21
	Doe	John	IT	2
	Brown	Olivia	HR	18

Employee ID	LastName	FirstName	Department	Years With Company
	Crow	Steven	Sales	14
	Burns	Charles	Sales	22

Figure 4.6 – MNAR

City
los angeles
LA
Los Angeles
Los Angelus
La
los angeles

Figure 4.7 – Invalid data

Cost Per Click
\$1.82
\$0.95
Toast
\$2.10
\$1.39

Figure 4.8 – Specification mismatch

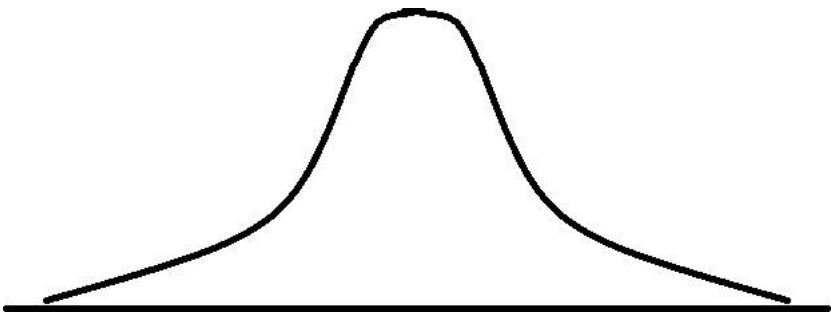


Figure 4.9 – Normal distribution

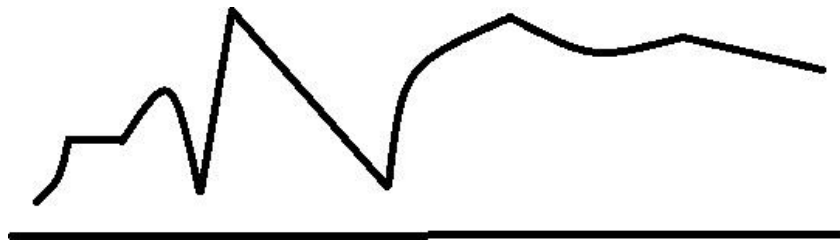


Figure 4.10 – Non-parametric distribution

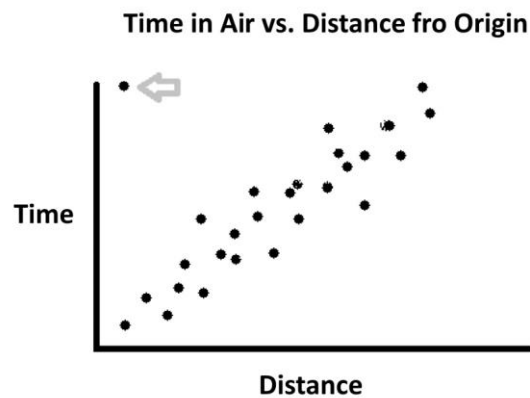


Figure 4.11 – Outlier

## Links

For more information on multicollinearity, look for guides on multicollinearity and overfitting. A decent article on multicollinearity can be found here: <https://towardsdatascience.com/multicollinearity-in-regression-fe7a2c1467ea>

A quick guide for overfitting can be found here: <https://www.ibm.com/cloud/learn/overfitting>

Just know that there is some debate, and a lot of popular information is coming from [Wikipedia](#), so take anything you read with a grain of salt.

A great resource for hands-on practice with outlier detection can be found here:

<https://analyticsindiamag.com/a-complete-guide-to-outlier-detection-with-hands-on-implementation-for-beginners>

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

## Questions

1. In regard to the following table, you can state:

ID	Sex	Male	Female
84927	M	TRUE	FALSE
69427	M	TRUE	FALSE
69427	M	TRUE	FALSE
10374	F	FALSE	TRUE
58264	M	TRUE	FALSE
90162	F	FALSE	TRUE

- A. This table has redundant data
  - B. This table has duplicate data
  - C. This table has redundant data and duplicate data
  - D. This table has neither redundant data nor duplicate data
2. When dealing with missing data, which of the following is **not** a type of deletion?
- A. Listwise
  - B. Singlewise
  - C. Pairwise
  - D. Variable
3. Which types of errors can be found in the following table?

Employee ID	LastName	FirstName	Department	Years With Company
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sale	21
73891	Doe	John	Information Technology	11

- A. Invalid data
  - B. Specification mismatch
  - C. Data type validation
  - D. None of these
4. Non-parametric data is a problem because:

- A. It has a normal distribution
  - B. You can't use parametric analyses on it
  - C. It has no distribution
  - D. There is no problem with non-parametric data
5. When in a dataset on the weight of human babies at birth, you come across a value of 8,000 lb. What can you deduce?
- A. It is probably an outlier, and you should check your ranges to be sure
  - B. The data point will probably skew your results
  - C. It is probably safe to delete this data point
  - D. All of these

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: *This table has redundant data and duplicate data*

The data in the second row is repeated in the third, meaning there is duplicate data, and the Sex, Male, and Female columns are redundant.

2. The answer is: *Singlewise*

Listwise deletion, pairwise deletion, and variable deletion are all types of deletion covered in this chapter.

3. The answer is: *Invalid data*

Here, there is a typo that differentiates the Sales department from the Sale department. Also, Information Technology is spelled out in one place and abbreviated in another. As is, this would be read as five unique departments instead of three.

4. The answer is: *You can't use parametric analyses on it*

Non-parametric data requires non-parametric statistical analysis.

5. The answer is: *All of these*

4-ton human babies are not likely. You should check to make sure it is an outlier, then delete it, or else; otherwise, it will skew all of your results.

# Chapter 5

## Figures

**Customer Names**

FirstName	LastName
Laurence	Smith
Betty	Brown
Phil	Hook
Jen	Roark
Jona	Cox

**Customer Locations**

City	State
Austin	TX
Denver	CO
Tulsa	OK
Phoenix	AZ
Seattle	WA

**Figure 5.1 – Customer names and customer locations**

**Customer Names**

CustomerID	FirstName	LastName
1	Laurence	Smith
2	Betty	Brown
3	Phil	Hook
4	Jen	Roark
5	Jona	Cox

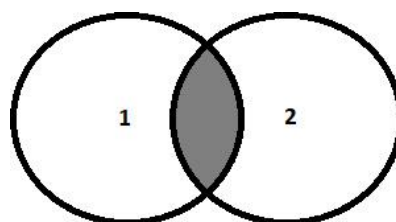
**Customer Locations**

CustomerID	City	State
3	Austin	TX
2	Denver	CO
5	Tulsa	OK
1	Phoenix	AZ
4	Seattle	WA

**Figure 5.2 – Customer names and customer locations with key variable**

Customer Names			Key Table		Customer Locations		
CustomerID	FirstName	LastName	CustomerID	LocationID	LocationID	City	State
1	Laurence	Smith	1	4	1	Austin	TX
2	Betty	Brown	2	2	2	Denver	CO
3	Phil	Hook	3	1	3	Tulsa	OK
4	Jen	Roark	4	5	4	Phoenix	AZ
5	Jona	Cox	5	3	5	Seattle	WA

**Figure 5.3 – Customer names and customer locations with Key Table**



**Figure 5.4 – Inner join Venn diagram**

Left Table		Joined Table			Right Table	
ClientID	Name	ClientID	Name	City	ClientID	City
1	Smith, Laurence	1	Smith, Laurence	Austin, TX	1	Austin, TX
2	Brown, Betty	2	Brown, Betty	Denver, CO	2	Denver, CO
3	Hook, Phil	3	Hook, Phil	Tulsa, OK	3	Tulsa, OK
4	Roark, Jen				7	Phoenix, AZ
5	Cox, Jona				8	Seattle, WA
6	Humbert, Ren				9	Baltimore, MD

Figure 5.5 – Inner join tables

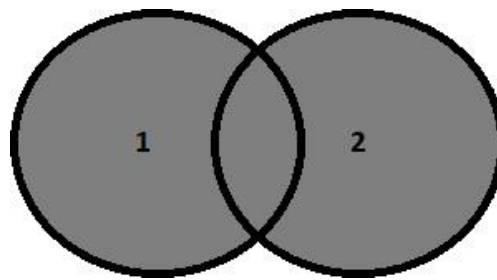


Figure 5.6 – Outer join Venn diagram

Left Table		Joined Table			Right Table	
ClientID	Name	ClientID	Name	City	ClientID	City
1	Smith, Laurence	1	Smith, Laurence	Austin, TX	1	Austin, TX
2	Brown, Betty	2	Brown, Betty	Denver, CO	2	Denver, CO
3	Hook, Phil	3	Hook, Phil	Tulsa, OK	3	Tulsa, OK
4	Roark, Jen	4	Roark, Jen	NULL	7	Phoenix, AZ
5	Cox, Jona	5	Cox, Jona	NULL	8	Seattle, WA
6	Humbert, Ren	6	Humbert, Ren	NULL	9	Baltimore, MD
		7	NULL	Phoenix, AZ		
		8	NULL	Seattle, WA		
		9	NULL	Baltimore, MD		

Figure 5.7 – Outer join tables

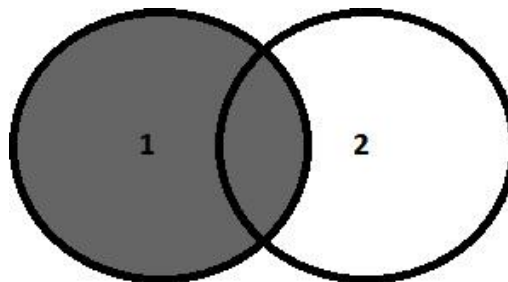


Figure 5.8 – Left join Venn diagram

Left Table		Joined Table			Right Table	
ClientID	Name	ClientID	Name	City	ClientID	City
1	Smith, Laurence	1	Smith, Laurence	Austin, TX	1	Austin, TX
2	Brown, Betty	2	Brown, Betty	Denver, CO	2	Denver, CO
3	Hook, Phil	3	Hook, Phil	Tulsa, OK	3	Tulsa, OK
4	Roark, Jen	4	Roark, Jen	NULL	7	Phoenix, AZ
5	Cox, Jona	5	Cox, Jona	NULL	8	Seattle, WA
6	Humbert, Ren	6	Humbert, Ren	NULL	9	Baltimore, MD

Figure 5.9 – Left join tables

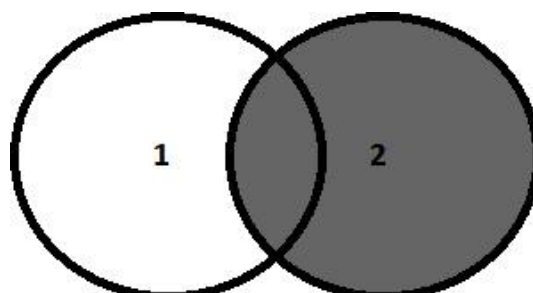




Figure 5.10 – Right join Venn diagram

Left Table		Joined Table			Right Table	
ClientID	Name	ClientID	Name	City	ClientID	City
1	Smith, Laurence	1	Smith, Laurence	Austin, TX	1	Austin, TX
2	Brown, Betty	2	Brown, Betty	Denver, CO	2	Denver, CO
3	Hook, Phil	3	Hook, Phil	Tulsa, OK	3	Tulsa, OK
4	Roark, Jen	7	NULL	Phoenix, AZ	7	Phoenix, AZ
5	Cox, Jona	8	NULL	Seattle, WA	8	Seattle, WA
6	Humbert, Ren	9	NULL	Baltimore, MD	9	Baltimore, MD

Figure 5.11 – Right join tables

May Table			Concatenated Table			June Table		
Date	QuantitySold	Income	Date	QuantitySold	Income	Date	QuantitySold	Income
5/1/2022	4	759.96	5/1/2022	4	759.96	6/5/2022	8	1519.92
5/8/2022	3	569.97	5/8/2022	3	569.97	6/12/2022	6	1139.94
5/15/2022	7	1329.93	5/15/2022	7	1329.93	6/19/2022	4	759.96
5/22/2022	2	379.98	5/22/2022	2	379.98	6/26/2022	7	1329.93
5/29/2022	6	1139.94	5/29/2022	6	1139.94			
			6/5/2022	8	1519.92			
			6/12/2022	6	1139.94			
			6/19/2022	4	759.96			
			6/26/2022	7	1329.93			

Figure 5.12 – Concatenated table

Date	Distance (m)	Time (s)	Speed (m/s)
1/1/2022	100	18.94	5.28
1/2/2022	100	16.43	6.09
1/3/2022	50	9.67	5.17
1/4/2022	200	33.8	5.92
1/5/2022	100	15.85	6.31

Figure 5.13 – Derived variable metrics table

CheckOutDate	BookID	ReturnDate	Date	DueDate	OverDue
3/14/2022	8473097	3/16/2022	3/22/2022	3/21/2022	
3/14/2022	8721809		3/22/2022	3/21/2022	OverDue
3/15/2022	3809893	3/19/2022	3/22/2022	3/22/2022	
3/16/2022	5837988	3/18/2022	3/22/2022	3/23/2022	
3/16/2022	4839801		3/22/2022	3/23/2022	

Figure 5.14 – Derived variable flags table

Date	Distance (m)	Time (s)	Speed (m/s)	SpeedCategory
1/1/2022	100	18.94	5.28	Slow
1/2/2022	100	16.43	6.09	Average
1/3/2022	50	9.67	5.17	Slow
1/4/2022	200	33.8	5.92	Slow
1/5/2022	100	15.85	6.31	Average

Figure 5.15 – Recode number to category

Month	UnitsSold	Color	ColorRecoded
August	432	Red	1
August	365	Blue	2
August	154	Yellow	3
September	398	Red	1
September	386	Blue	2
September	108	Yellow	3

Figure 5.16 – Recode category to number

Month	UnitsSold	Color	Red	Blue	Yellow
August	432	Red	1	0	0
August	365	Blue	0	1	0
August	154	Yellow	0	0	1
September	398	Red	1	0	0
September	386	Blue	0	1	0
September	108	Yellow	0	0	1

Figure 5.17 – Recode dummy coding

9/17/2022	9	17	2022
9/18/2022	9	18	2022
9/19/2022	9	19	2022
9/20/2022	9	20	2022
9/21/2022	9	21	2022

Figure 5.18 – Dates and delimiters

Month	UnitsSold	Color
August	432	Red
August	365	Blue
August	154	Yellow
September	398	Red
September	386	Blue
September	108	Yellow

Month	August	August	August	September	September	September
UnitsSold	432	365	154	398	386	108
Color	Red	Blue	Yellow	Red	Blue	Yellow

Figure 5.19 – Transposing tables

## Code

### Parsing your data

go ahead and store it in a variable called `Data` as follows:

```
Data = "This book makes me happy."
```

If you receive the preceding line as a piece of data, the computer accepts this as one whole and can do very little with it. However, we can use a parser on it like in the following line of code:

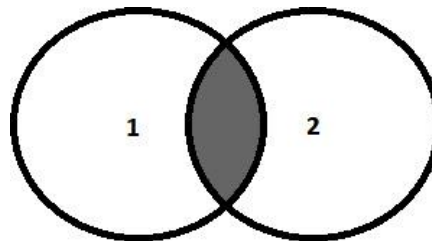
```
Data = ["This", "book", "makes", "me", "happy", "."]
```

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. The following picture represents what kind of join?



- A. Inner join
  - B. Outer join
  - C. Left join
  - D. Right join
2. Only using the Distinct Count of a dataset is an example of what?
- A. Normality
  - B. Recoding
  - C. Transposition
  - D. Reduction
3. The following is an example of what concept?

```
Data = "This is a sentence?"  
Data = ["This", "is", "a", "sentence", "?"]
```

- A. Transposing
  - B. Parsing
  - C. Derived variables
  - D. Reduction
4. The following is an example of what concept?

Month	UnitsSold	Color	Red	Blue	Yellow
August	432	Red	1	0	0
August	365	Blue	0	1	0
August	154	Yellow	0	0	1
September	398	Red	1	0	0
September	386	Blue	0	1	0
September	108	Yellow	0	0	1

- A. Transformation
  - B. Dummy coding
  - C. Blending
  - D. System functions
5. Which of the following is a logical operator?
- A. IF
  - B. NOT
  - C. OR
  - D. All of the above are logical operators

## Answers

Now we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Inner join  
  
This represents an inner join because only the data that both datasets have in common are added to the new table.
2. The answer is: Reduction  
  
Distinct Count is an aggregate used to summarize data. Using only this reduces the amount of data you have to process, earning the name reduction variable.
3. The answer is: Parsing  
  
Parsing is the concept of breaking down large chunks of data into smaller pieces that can be processed and analyzed.
4. The answer is: Dummy coding  
  
Dummy coding is a specific type of recoding that creates a new variable for every possible outcome of a categorical variable.
5. The answer is: All of the above are logical operators  
  
The most common logical operators are IF, AND, OR, and NOT.



# Chapter 6

## Technical requirements

There is nothing that you absolutely have to do to prepare for this chapter, but there is an example in this chapter using Python in Jupyter Notebook. If you would like to follow along, you can find the `EDA_Example_Data.csv` dataset by following the link provided:

<https://github.com/PacktPublishing/CompTIA-Data-DAO-001-Certification-Guide>

This GitHub link will take you to a repository that contains the data as well as my example code.

## Figure

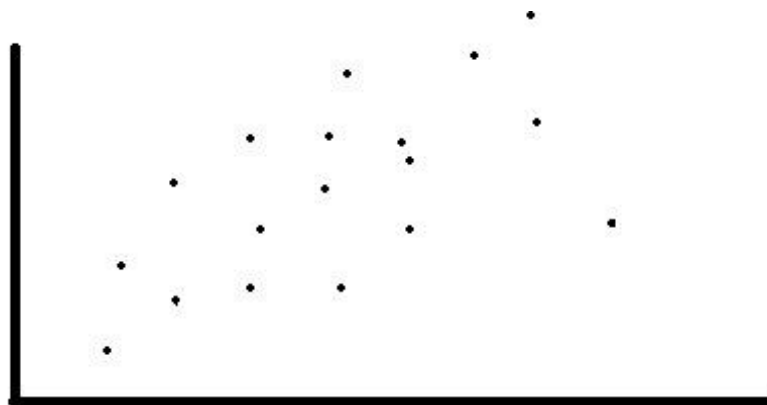


Figure 6.1 – Scatter plot

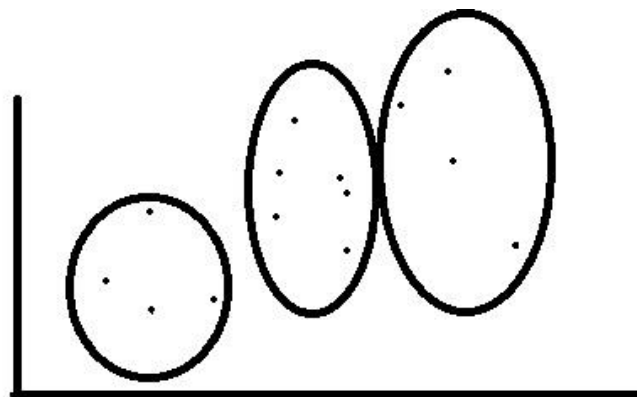


Figure 6.2 – Dimension reduction with clusters

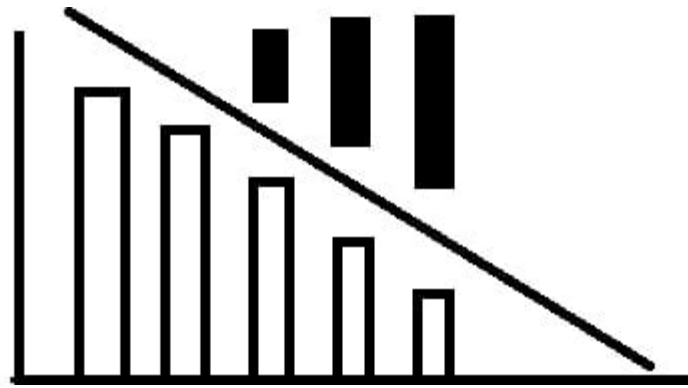


Figure 6.12 – Burndown chart



Figure 6.13 – Forecasting

## Hands-On Sections

### EDA example

This is all sounding pretty vague, so how about we walk through an example together? We will use a sample of customer data. For the purpose of this example, I will be using Python 3 through Jupyter Notebook.

#### Important note

Again, you do not need to know how to use any particular software or programming language for the exam. This example is in Python because it is a popular tool and one I personally use. You are encouraged to practice this on your own, but if you would rather, you can simply read along to get a general idea of the process.

If you would like to practice this on your own, I would encourage you to go to the website [www.anaconda.com](http://www.anaconda.com). Anaconda is a free collection of data science tools and modules. Not only will you get Python and even Jupyter Notebook, but it will automatically install the majority of packages that you will need for data analytics. Let's import a package and pull in our data:

```
import pandas as pd
MyData = pd.read_csv("EDA_Example_Data.csv")
MyData.head()
```

The preceding code is pretty straightforward. The first line imports the `pandas` package and saves it as `pd`. The second line uses a function from `pd` to load our data into the program and save it as `MyData`. The third line is a function that will just display the first few lines of our data. Overall, this imports the package we need, and our data, and gives us a sneak peek of it. The results will look something like this:

	Client_ID	Age_Bracket	AB_Recode	Orders	Total_Spent
0	4682531	41 - 50	4	2	28.0
1	6464235	21 - 30	2	1	23.0
2	9390103	21 - 30	2	10	210.0
3	9346815	< 21	1	4	76.0
4	5672895	21 - 30	2	2	18.0

**Figure 6.3 – First five rows of the dataset**

The dataset displayed in **Figure 6.3** is a sample of 100 customers from an e-commerce website, though we are only seeing the first five entries. `Client_ID` is a unique ID number attributed to every client. `Age_Bracket` shows what age range the client falls into. `AB_Recode` is a recode of the `Age_Bracket` variable, and we will talk about why it is there in a moment. `Orders` is the number of orders placed by the client. `Total_Spent` is how much money the client has spent in total.

Okay, we have brought in our data and glanced at it. Before we go too in-depth, we should start with a broad view of the data as a whole by executing the following line of code:

```
MyData.info()
```

This one little line will give us a rough description of our dataset as a whole. The results should be as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Client_ID    100 non-null    int64
1   Age_Bracket  100 non-null    object
2   AB_Recode    100 non-null    int64
3   Orders       100 non-null    int64
4   Total_Spent  100 non-null    float64
dtypes: float64(1), int64(3), object(1)
memory usage: 4.0+ KB
```

**Figure 6.4 – Summary of the dataset with info()**

In **Figure 6.4**, we see that our data is formatted as a pandas `DataFrame`, which is fine for what we are doing. We also get a list of every variable, the number of values that are not banked for every variable, what data type each variable is, the counts of each data type, and how much memory it is using. Because the data has already been cleaned, every variable has all 100 values as non-null. We see that `Client_ID`, `AB_Recode`, and `Orders` are integers, `Total_Spent` is a float, and `Age_Bracket` is an object, which, in this case, means it is a `string` variable.

### Important note



In the majority of programming languages, you will find common variable types. To put it simply, integers are whole numbers, floats are decimals, and strings are text.

It should be noted that if `Total_Spent` was formatted as currency, that dollar sign in front would mean that when we brought it into Python, it would be treated as an object as well, and we would not be able to use it in calculations. There are ways to convert it to a float within Python, but that is something that should be completed before this process, if possible.

For this example, we know that in the future, we will want to perform calculations on `Total_Spent`, so let's look at it a little closer:

```
MyData["Total_Spent"].describe()
```

This line of code performs the `describe()` function on the `Total_Spent` variable found in `MyData`. The results are as follows:

```
count    100.00000
mean     101.52000
std       89.76473
min        2.00000
25%       28.00000
50%       82.50000
75%      137.50000
max      364.00000
Name: Total_Spent, dtype: float64
```

**Figure 6.5 – Summary of Total\_Spent with describe()**

In **Figure 6.5**, we see where the descriptive statistics come in. Here, we have the number of values, the average of the values, the standard deviation, the smallest and largest values, as well as the quartiles. This gives us a lot of information about this variable. If you want, you can repeat this process for `Orders` simply by rerunning the previous line of code, but you will have to replace `Total_Spent` with `Orders`. You could, theoretically, also run it on `Client_ID`, `Age_Bracket`, or `AB_Recode`, but it would be meaningless. `Client_ID` is a key variable and not something you will generally use for analytics, `Age_Bracket` is a categorical variable, and you will only get general counts and what value pops up the most, and `AB_Recode` is a categorical variable pretending to be an integer, so the results won't make any sense.

Okay, well, if the `describe()` function won't help with a categorical variable, what should you use?

```
MyData["Age_Bracket"].value_counts()
```

This is a quick way to take a glance at a categorical variable. It is also a quick way to check for typos in a variable. The results are as follows in **Figure 6.6**:

```
31 - 40    34
41 - 50    21
21 - 30    19
> 50       15
< 21       11
Name: Age_Bracket, dtype: int64
```

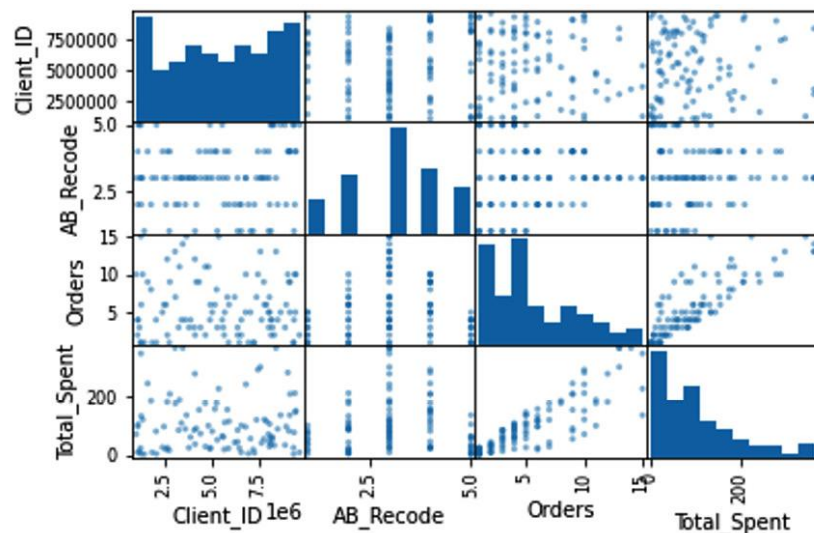
**Figure 6.6 – Value count results**

In the preceding figure, we see every value that is found in `Age_Bracket`, and the number of times each occurs. The values are, by default, sorted from the most frequent to the least. Here, we see that the majority of clients for this website are between the ages of 31 and 40.

Next, we are going to combine a few steps all in one very convenient line of code:

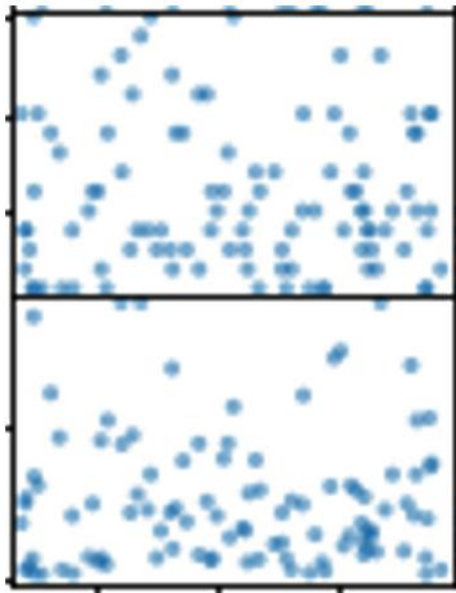
```
EDA_Plots = pd.plotting.scatter_matrix(MyData)
```

This creates a scatter plot matrix of all numerical variables. Also, in the squares on the grid where a variable crosses itself, it displays a histogram that gives you a rough idea of the distribution of that variable. The results are as follows:



**Figure 6.7 – Scatter plot matrix**

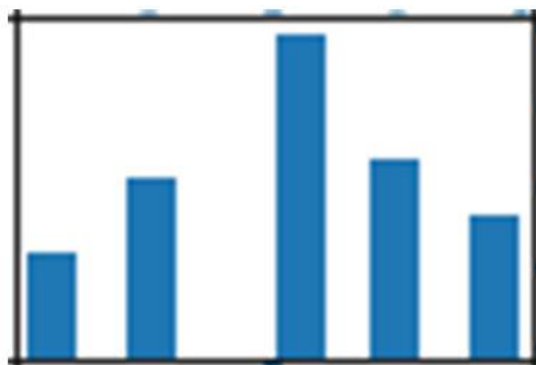
There is a ton of information packed into this little visualization, as shown in [Figure 6.7](#). Let's look at the variables one at a time. The column on the far left is `Client_ID`. Normally, you would not include a key variable in something like this. You would actually create a new dataset, which would be a subset that only included the variables you wanted. However, I left it here to show you what the results would be. The histogram in the top-left corner, which shows the distribution of `Client_ID`, tells us that the distribution is random and does not show any trends, which makes sense because client ID numbers are often generated at random. The next square down, `AB_Retcode` / `Client_ID`, shows what it looks like when you include a categorical variable in this matrix, and means nothing. The two squares below that, `Client_ID` / `Orders` and `Client_ID` / `Total_Spent`, show how `Client_ID` relates to `Orders` and `Total_Spent`.



**Figure 6.8 – Scatterplots with no relationship**

This random scattering shown in **Figure 6.8** does not show anything. There is no relationship between **Client\_ID** and anything.

The second column in Figure 6.7 shows **AB\_Recode**. Again, you generally do not want a categorical variable in this kind of matrix, but we included it for the histogram, not the scatter plots. This histogram actually shows a pretty normal distribution, as seen in **Figure 6.9**:



**Figure 6.9 – Histogram of AB\_Recode**

We can do this because there is an inherent order to the values in **Age\_Bracket**; it is ordinal. In other words, we have an order to the categories: <21, 21–30, 31–40, 41–50, and >50. This is because the age brackets are based on the client's age, which is a number.

If you were to try to create a histogram based on a categorical variable such as color, it would be completely meaningless. Blue could be 1, 2, or 27. There is no logic to the order, so there would be no meaning to the shape created by the order.

The third column in Figure 6.7 is **Orders**. We have already discussed the relationship between **Client\_ID** and **AB\_Recode**, so let's look at the histogram:

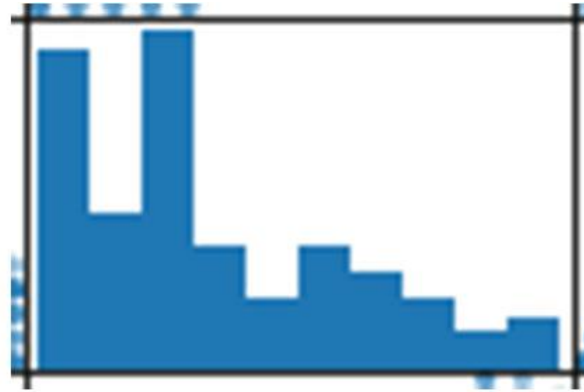


Figure 6.10 – Histogram of Orders

In *Figure 6.10*, we see it is skewed to the right and not a normal distribution, so we should treat it as non-parametric until we try to adjust for the skew. The last box shows how `Orders` relates to `Total_Spent`. There is a rough line from the lower left to the upper right, as shown in *Figure 6.11*:

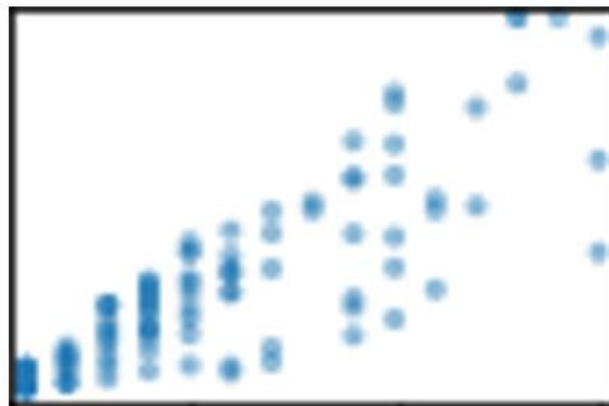


Figure 6.11 – Scatter plot of Orders and Total\_Spent

This indicates that there might be a positive relationship between `Orders` and `Total_Spent`. Logically, it makes sense that the more orders a customer makes, the more likely they are to spend more money.

The last column is `Total_Spent` and shows what we would expect at this point. The histogram shows a right skew, it has a positive relationship with `Orders`, a meaningless relationship with `AB_Recode`, and no relationship with `Client_ID`.

This is a lot of information to glean from one line of code. We could go more in-depth if we had specific goals or things to check, but this is a solid amount of basic information from a quick, high-level EDA. We now know enough about our variables to go ahead and get started.

That's what you need to know about EDA. Next, we will move on to performance checks.

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

## Questions

1. What is the purpose of EDA?
  - A. Checking on the efficiency of a program or project
  - B. Understanding basic information about your data before performing more advanced analytics
  - C. Understanding how or whether different variables relate to one another
  - D. Checking on the progress of a variable over time
2. A manager of a small team wants to gather some metrics to track how the team is doing and where they have room for improvement. What type of analysis would they need to do?
  - A. Exploratory data analysis
  - B. Link analysis
  - C. Performance analysis
  - D. Trend analysis
3. Forecasting falls into which analysis type?
  - A. Link analysis
  - B. Performance analysis
  - C. Exploratory data analysis
  - D. Trend analysis
4. Scatter plots, correlation, and structural equation modeling all fall into which analysis type?
  - A. Performance analysis
  - B. Trend analysis
  - C. Link analysis
  - D. Exploratory data analysis
5. In reference to statistical analyses, what are assumptions?
  - A. A list of prerequisites that must be met
  - B. You should never make assumptions, confirm everything
  - C. A list of things to avoid
  - D. None of the above

## Answers

Now we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Understanding basic information about your data before performing more advanced analytics

Exploratory data analysis is your chance to get a feel for your data and learn the basics about it. It is important to take this step, so you know what you can or should do next.

2. The answer is: Performance analysis

Performance analysis can include a wide range of different statistical methods, but they generally focus on using key metrics to track progress and performance.

3. The answer is: Trend analysis

Forecasting is a specific kind of trend analysis that looks at predicting the future of a metric by looking at what it has done in the past.

4. The answer is: Link analysis

Scatter plots, and even correlation, technically can be used in EDA but they are still considered part of link analysis. Also, structural equation modeling is an advanced technique that would not be used in EDA.

5. The answer is: A list of prerequisites that must be met

Assumptions are things that the equation assumes about the data to be true, and the data must meet those expectations in order to give an accurate result.

# Chapter 7

## Figures

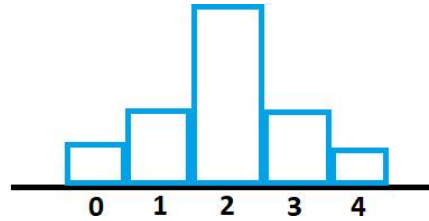


Figure 7.1 – Histogram of Fighting Goldfish goals

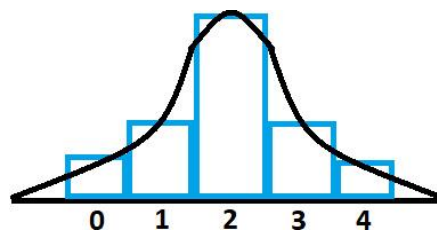


Figure 7.2 – Distribution of Fighting Goldfish goals

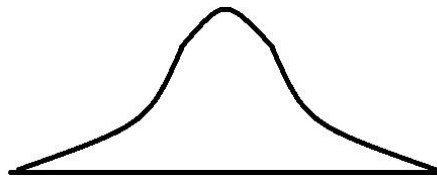


Figure 7.3 – Normal distribution

If you're curious, the formula for the normal distribution function can take many forms, but is often roughly something like this:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right)$$



Figure 7.4 – Uniform distribution

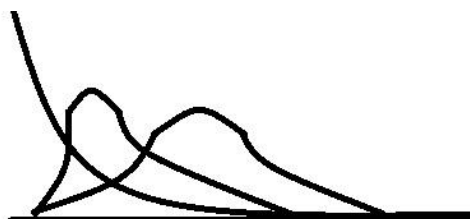


Figure 7.5 – Possible Poisson distributions

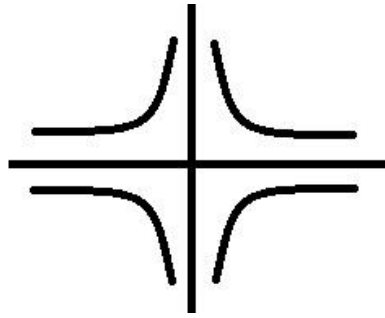


Figure 7.6 – Exponential distributions

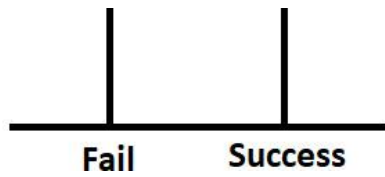


Figure 7.7 – Bernoulli distribution

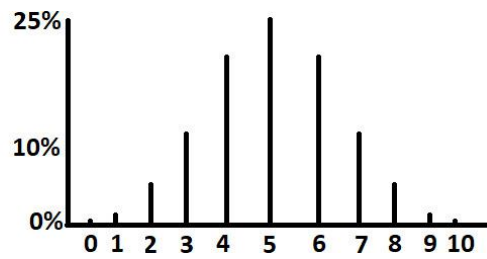


Figure 7.8 – Binomial distribution of rocket launches



Figure 7.9 – Negative or left skew

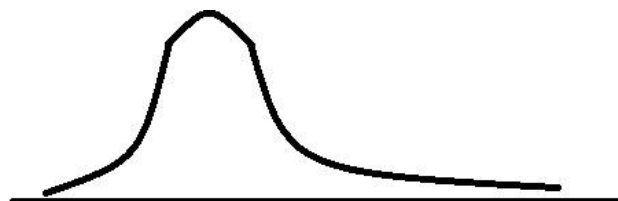


Figure 7.10 – Positive or right skew



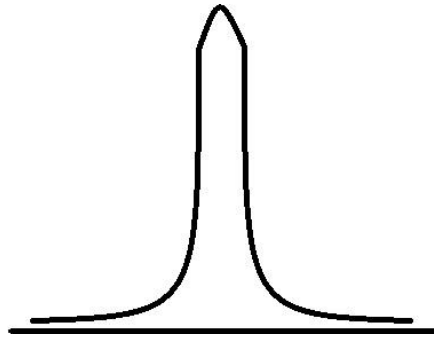


Figure 7.11 – Leptokurtic distribution



Figure 7.12 – Platykurtic distribution

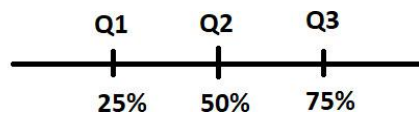


Figure 7.13 – Quartiles

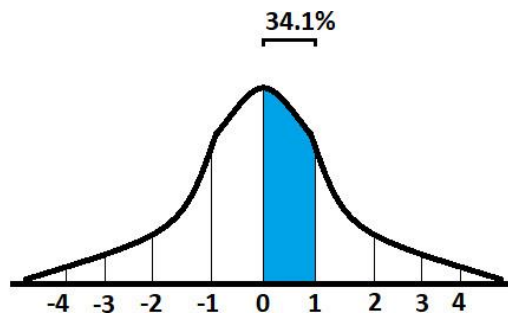


Figure 7.14 – One standard deviation

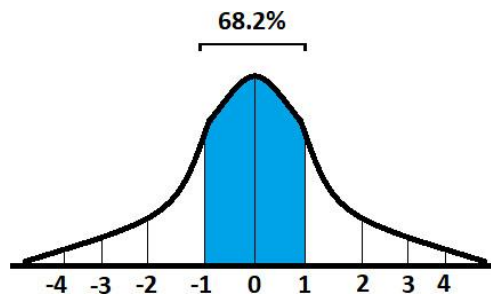


Figure 7.15 – One standard deviation above and below the mean

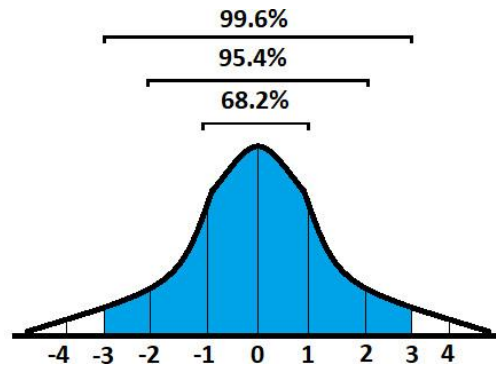


Figure 7.16 – Standard deviation percentages

## Formula

Variance:

$$s^2 = \frac{\sum (x^i - \bar{x})^2}{n - 1}$$

## Hands-On Section

### Variance

Now, you work for a website that sells specialized boats. The values represent how many boats are sold in a week: 3, 6, 4, 7, 5, 1, 9, 5, 4, and 6:

1. Find the mean of the dataset:

$$\frac{(3 + 6 + 4 + 7 + 5 + 1 + 9 + 5 + 4 + 6)}{10} = 5$$

The mean is 5.

2. Subtract the mean from each data point:

$$(3 - 5), (6 - 5), (4 - 5), (7 - 5), (5 - 5), (1 - 5), (9 - 5), (5 - 5), (4 - 5),$$

$$(6 - 5)$$

becomes

$$(-2), (-1), (1), (2), (0), (-4), (4), (0), (-1), (1)$$

3. Square the results from the previous step:

$$-2^2, -1^2, 1^2, 2^2, 0^2, -4^2, 4^2, 0^2, -1^2, 1^2$$

becomes

4, 1, 1, 4, 0, 16, 16, 0, 1, 1

- Find the sum of the previous step:

$$4 + 1 + 1 + 4 + 0 + 16 + 16 + 0 + 1 + 1 = 44$$

The sum is 44.

- Divide that sum by the number of data points minus 1:

$$\frac{44}{(10 - 1)} = 4.9$$

The variance for this dataset is 4.9.

You can practice this process on your own with the following dataset: 10, 8, 12, 11, 9, 9, 8, 12, 11, 10.

## Standard Deviation

For our example, we will stick to the boat data, but we will go ahead and collect another sample of 10 days: 4, 5, 3, 4, 2, 2, 6, 4, 6, and 4:

- Find the mean of the dataset:

$$\frac{(4 + 5 + 3 + 4 + 2 + 2 + 6 + 4 + 6 + 4)}{10} = 4$$

The mean is 4.

- Subtract the mean from each datapoint:

$$(4 - 4), (5 - 4), (3 - 4), (4 - 4), (2 - 4), (2 - 4), (6 - 4), (4 - 4), (6 - 4), (4 - 4)$$

becomes

$$(0), (1), (-1), (0), (-2), (-2), (2), (0), (2), (0)$$

- Square the results from the previous step:

$$0^2, 1^2, -1^2, 0^2, -2^2, -2^2, 2^2, 0^2, 2^2, 0^2$$

becomes

$$0, 1, 1, 0, 4, 4, 4, 0, 4, 0$$

- Find the sum of the previous step:

$$0 + 1 + 1 + 0 + 4 + 4 + 4 + 0 + 4 + 0 = 18$$

The sum is 18.

- Divide that sum by the number of data points minus 1:

$$\frac{18}{(10 - 1)} = 2$$

The variance for this dataset is 2.

6. Take the square root of the previous step:

$$\sqrt{2} = 1.41$$

The standard deviation is 1.41. Note that if the square root is not obvious, as here, the answer may be displayed as  $\sqrt{2}$ .

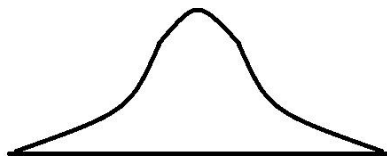
You can practice this process on your own with the following dataset: 7, 4, 10, 8, 6, 7, 6, 5, 8, 9.

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. The following distribution is considered:



- A. Uniform
  - B. Normal
  - C. Poisson
  - D. Exponential
2. What is the mode(s) of the following dataset?
- 24, 18, 36, 51, 24, 48, 18
- A. 18
  - B. 24
  - C. 18 and 24
  - D. None of the above
3. What is the range of the following dataset?
- 15, 615, 46, 73, 45, 80, 46
- A. 131

- B. 46
  - C. 73
  - D. 600
4. What is the middle quartile (Q2) of the following dataset?
- 10, 24, 13, 9, 15, 7, 19
- A. 9
  - B. 13
  - C. 19
  - D. 15
5. What is the standard deviation of the following sample dataset?
- 9, 11, 7, 8, 9, 10
- A. 2
  - B. 9
  - C.  $\sqrt{2}$
  - D.  $\sqrt{9}$

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Normal  
The distribution is the distinctive bell curve of a normal distribution.
2. The answer is: 18 and 24  
Both 18 and 24 are repeated twice, meaning they are both modes.
3. The answer is: 600  
Make sure you identify the minimum and maximum to calculate the range.
4. The answer is: 13  
Remember that the middle quartile (Q2) is the same as the median.
5. The answer is:  $\sqrt{2}$   
Make sure to take your time and go through the steps one by one.

# Chapter 8

## Hands-On Sections

### Frequencies

We can throw around all the fancy terms we want, but at the end of the day, a frequency is just a count of occurrences. You are literally just counting how many times a specific value occurs within a variable. There are many ways to do this, depending on the data analysis tools you are using, but for this exam, you must calculate it by hand. The easiest tool for this is a frequency table:

**So for example for the three different shapes:**

Square

Circle

Triangle

**Their Frequency would be**

42

31

And 16 respectively.

A frequency table is exactly what it sounds like: a table that lists every reported value in a variable and how many times they happened. If you recall from the previous chapter, a simplified form of this table was used when finding the mode of a dataset. Let's break this down into easy steps:

1. Arrange the values in ascending or descending order.
2. Create a table and list all possible values.
3. Add a count of every variable.

Let's try this out with an example. The flagship product your company sells comes in an assortment of different colors: Blue, Red, and Yellow. After recording the sales of the product for a day, the results are as follows: Red, Red, Blue, Red, Yellow, Yellow, Blue, Yellow, Red, Red, Blue, Red, Yellow, Blue, and Yellow:

1. Arrange values in ascending or descending order like this:

Blue, Blue, Blue, Blue, Red, Red, Red, Red, Red, Red, Yellow, Yellow, Yellow, Yellow, Yellow

2. Create a table and list all possible values:

Color	Blue	Red	Yellow
-------	------	-----	--------

Frequency			
-----------	--	--	--

3. Add a count of every variable:

Color	Blue	Red	Yellow
Frequency	4	6	5

You can practice this process on your own with the following dataset: Red, Yellow, Blue, Blue, Yellow, Red, Blue, Yellow, Yellow, Red, Blue, Blue, Yellow, Red, and Blue.

## Complex frequency tables

Frequency tables can hold more than one variable and when it does, it is called a contingency table. The steps are the same as making a frequency table for one variable; you simply arrange the table differently.

For two variables, you have one variable on each axis. Let's look at the same product information as before, except now the product also has two sizes: Large and Small. Our dataset now looks like this: Small Red, Small Blue, Large Red, Large Red, Large Yellow, Large Blue, Small Yellow, Small Red, Large Yellow, Small Blue, Small Red, Large Yellow, Large Blue, Large Red, and Small Blue:

1. Arrange values in ascending or descending order:

Large Blue, Large Blue, Small Blue, Small Blue, Small Blue, Large Red, Large Red, Large Red, Small Red, Small Red, Small Red, Large Yellow, Large Yellow, Large Yellow, Small Yellow

2. Create a table and list all possible values:

	Blue	Red	Yellow
Large			
Small			

3. Add a count of every variable:

	Blue	Red	Yellow
Large	2	3	3
Small	3	3	1

## Important note

Do not confuse count with probability. In the preceding example, let's compare Red and Yellow in regard to Large products. The count for Large in each group is 3, but the probability of a Red or Yellow product being large is different. Probability is calculated by dividing the count of that possibility by the sum of counts for that group. If the product is Red, the probability of it being Large is  $\frac{3}{6}$ , or 50%. If the product is Yellow, the probability of it being Large is  $\frac{3}{4}$ , or 75%. In the exam, questions about frequency might be in terms of counts OR probabilities.

Not too bad, right? Okay, let's see what happens when we add a third variable. We are still following the same steps that we did for a frequency table with two variables. This time, we will also track whether the product is Normal or Premium. Let's look at our new dataset: Normal Small Red,

Premium Small Red, Normal Large Blue, Premium Small Red, Normal Large Blue, Normal Small Yellow, Premium Large Blue, Premium Small Red, Normal Large Yellow, Premium Large Yellow, Normal Large Blue, Normal Large Yellow, Premium Large Red, Premium Small Blue, and Normal Large Yellow:

1. Arrange values in ascending or descending order:

Normal Large Blue, Normal Large Blue, Normal Large Blue, Premium Large Blue, Premium Small Blue, Premium Large Red, Normal Small Red, Premium Small Red, Premium Small Red, Premium Small Red, Normal Large Yellow, Normal Large Yellow, Normal Large Yellow, Premium Large Yellow, Normal Small Yellow

2. Create a table and list all possible values:

		Blue	Red	Yellow
Normal	Large			
	Small			
Premium	Large			
	Small			

3. Add a count of every variable:

		Blue	Red	Yellow
Normal	Large	3	0	3
	Small	0	1	1
Premium	Large	1	1	1
	Small	1	3	0

This still may be manageable, but this is also only looking at three variables, and each one of those only has two or three possible values. Imagine 5 variables with 20 values each. It would quickly reach a point where it is difficult to read, time-consuming to put together, and doesn't tell you enough to make it worth it.

Now that you can create a frequency table and scale it to any size, let's move on to percentages.

## Percentages

Let's try these out in an example. Let's keep the previous example of a product, but for the sake of simplicity, we will only be looking at the color variable. The dataset is as follows: Red, Red, Yellow, Red, Blue, Blue, Yellow, Red, Blue, and Yellow:

1. Find the frequencies.

We just went through how to do this. If you are unsure, please review the previous section:



Blue	Red	Yellow
3	4	3

2. Divide each frequency by the total number of values.

Our total number of values, reached by adding together all of the counts, is 10:

Blue	Red	Yellow
3/10	4/10	3/10

This gives us the following:

Blue	Red	Yellow
0.3	0.4	0.3

3. Multiply each frequency by 100:

Blue	Red	Yellow
0.3*100	0.4*100	0.3*100

This gives us the following:

Blue	Red	Yellow
30%	40%	30%

This gives us the percentages of each value that makes up the whole of our sample. Percentages are useful and are heavily used with demographic data. If 70% of your customers fall into a specific age bracket, you will want to target your advertisements toward that age bracket. There is a wide range of key demographics, but the majority of them are qualitative data that benefit from a quick percentage table.

Now that we have introduced the idea of percentages, let's talk about percent change.

## Percent change

For our example, we will be looking at the metric *clicks per minute*, which describes the average number of times a web advertisement is clicked every minute. At the beginning of 2021, the clicks per minute was 4, and by the end of the year, it was 6. Let's see how much the clicks per minute changed:

1. Subtract the starting value from the ending value.

We can identify the starting value as 4 and the ending value as 6:

$$6 - 4 = 2$$

2. Divide the results of the previous step by the starting value:

$$2 / 4 = 0.5$$

3. Multiply the results of the previous step by 100:

$$0.5 * 100 = 50\%$$

The percent change is 50%, which means the clicks per minute has increased by 50% since the start of the year.

To practice this process on your own, try using the following values: starting value = 1 and ending value = 3.

## Percent difference

For consistency, let's use the same example as we did with percent change, but this time, the results come from two different ad campaigns. The clicks per minute from ad campaign A were 4 and from ad campaign B were 6. What is the percent difference?

1. Subtract one value from the other and drop any negative signs:

$$4 - 6 = -2$$

If we drop the negative sign, that changes the result:

$$2$$

2. Find the average of the two values:

$$(4 + 6) / 2 = 5$$

3. Divide the result of the first step by the result of the second step:

$$2 / 5 = 0.4$$

4. Multiply the result of the previous step by 100:

$$0.4 * 100 = 40\%$$

The percent difference is 40%.

You will notice that this is not the same value as the percent change using the same values. That is because it is not in terms of the starting value. This is the objective difference between two values of equal importance.

To practice this process on your own, try using the following values: starting value = 1 and ending value = 3.

## Discovering confidence intervals

You are hired as the main data analyst for a hotdog cart. They want to know how many hotdogs they sell per hour. You recorded how many hotdogs they sold per hour for several hours and created the following dataset: 8, 6, 9, 7, 8, 10, 8, 8, 8. The t-value for this sample is 1.86:

1. Find the mean of your sample.

The mean is 8.

2. Find the standard deviation of your sample.

The standard deviation is 1.12.

3. Find the square root of your sample size.

The square root of the sample size is 3.

4. Divide your standard deviation by the result of the previous step:

$$1.12 / 3 = 0.37$$

5. Multiply the result of the previous step by your t-value:

$$1.86 * 0.37 = 0.69$$

6. Add the result of the previous step to your mean to find your upper confidence interval:

$$\text{Upper confidence interval} = 8 + 0.69 = 8.69$$

7. Subtract the result of **step 5** from the mean to find your lower confidence interval:

$$\text{Lower confidence interval} = 8 - 0.69 = 7.31$$

We are 95% confident that the “true mean” of hotdogs sold per hour is between 7.31 and 8.69.

To practice this process on your own, try using the following dataset: 2, 1, 3, 2. The t-value is 2.35.

## Understanding z-scores

Your distribution has a mean of 5 and a standard deviation of 1. How does the value of 6 compare?

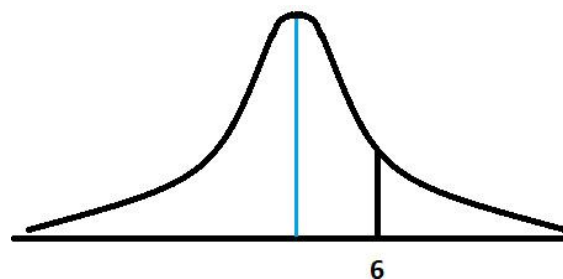
1. Subtract the mean from your value:

$$6 - 5 = 1$$

2. Divide the results of the previous step by the standard deviation:

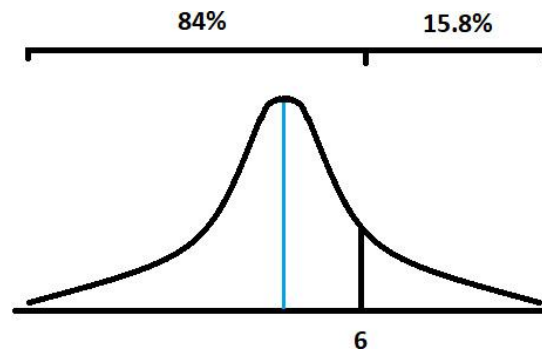
$$1 / 1 = 1$$

The value 6 is 1 standard deviation above the mean. We can see this in **Figure 8.4**:



**Figure 8.4 – Distribution with result marked**

If you recall, in a normal distribution, each standard deviation accounts for a specific percentage of the whole. A value that is 1 standard deviation above the mean is higher than roughly 84% of all values and is lower than only roughly 15.8% of values. We can see this illustrated in **Figure 8.5**:

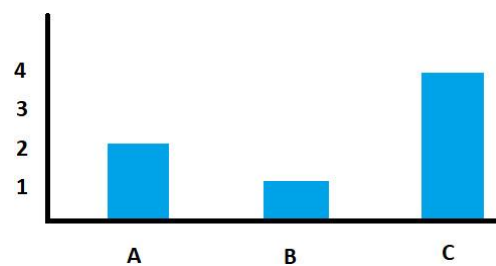


**Figure 8.5 – Distribution with result and percentages**

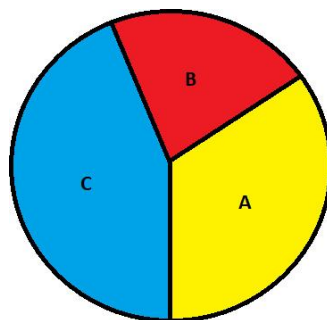
3. Rejoice!

Yay! You have survived the two chapters of hand calculations!

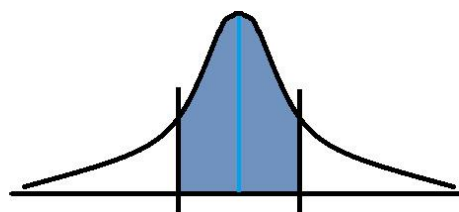
## Figures



**Figure 8.1 – A bar chart symbolizing counts or frequencies**



**Figure 8.2 – A pie chart symbolizing percentages**



**Figure 8.3 – Confidence interval**

## Formula

Percent change:

$$C = \left( \frac{x_2 - x_1}{x_1} \right) * 100$$

$$\text{Percent difference} = \left( \frac{\frac{|x_1 - x_2|}{2}}{\frac{x_1 + x_2}{2}} \right) * 100$$

$$\text{Confidence interval (CI)} = \bar{x} \pm t \left( \frac{\sigma}{\sqrt{n}} \right)$$

Z-score:

$$Z = \frac{(x - \mu)}{\sigma}$$

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. Convenient News, a membership service, has created the following frequency table about its subscribers. The table describes the sex of the member, whether they were a Normal or Premium member, and whether they continued their subscription or quit. Which group described in the table has the highest probability of quitting?

		Current Subscriber	Past Subscriber
Male	Normal	89	52
	Premium	12	2
Female	Normal	12	45
	Premium	35	13

- A. Males with a Normal subscription
- B. Males with a Premium subscription
- C. Females with a Normal subscription
- D. Females with a Premium subscription

2. Convenient News has tracked the **Customer Lifetime Value (CLV)**, a common KPI, for a while now. At the start of the year, the CLV was \$200, and now it is \$240. How do these numbers compare?
  - A. 20% increase
  - B. 20% decrease
  - C. -18% difference
  - D. 18% difference
3. Convenient News wants to compare its CLV of \$240 to a competitor, Inconvenient News, which has a CLV of \$180. How do they compare, in percentages?
  - A. -28% difference
  - B. 28% difference
  - C. 25% increase
  - D. 25% decrease
4. Convenient News is tracking how many newspapers a specific store is selling every day. The dataset is as follows: 10, 12, 11, 8, 10, 8, 9, 12, 10. The standard deviation is 1.5, the t-value is 1.86, and the mean is 10. Create a confidence interval for this data:
  - A. The lower confidence interval is 9.75 and the upper confidence interval is 10.25
  - B. The lower confidence interval is 8.14 and the upper confidence interval is 11.86
  - C. The lower confidence interval is 8.5 and the upper confidence interval is 11.5
  - D. The lower confidence interval is 9.07 and the upper confidence interval is 10.93
5. Convenient News has tracked the number of times a particular article has been read on its website for years. A specific article stands out and the author wants to know how the number of times it has been read compares to other articles. Which analysis is most appropriate?
  - A. Simple linear regression
  - B. T-test
  - C. Z-score
  - D. Chi-square

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Females with a Normal subscription

Females with a Normal subscription is the only group that has had more members quit than stay. The probability of this group quitting is 79%. Even though the count of Men with Normal subscriptions who quit is higher, the probability of this group quitting is only 37%, because there are so many more members in this group.

2. The answer is: 20% increase

This is a percent change question because you are looking at a single value at two different points in terms of the starting value.

3. The answer is: 28% difference

This is a percent difference problem because you are looking at two distinct groups. Also, remember that percent difference cannot be a negative value.

4. The answer is: The lower confidence interval is 9.07 and upper confidence interval is 10.93

You are given all of the pieces for this one, it is just a matter of going through the steps one at a time.

5. Z-score

Z-scores are specifically for comparing a single value to a distribution. Here, you are comparing the value of a metric for a single article against all past values for every other article.

Figures

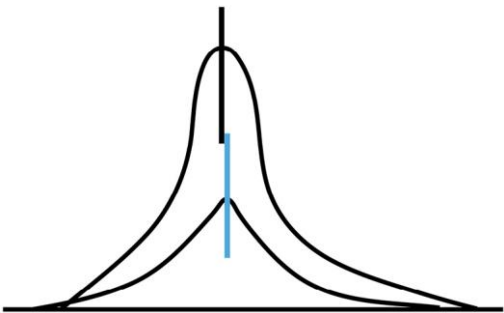


Figure 9.1 – Comparing two distributions

Group A	Group B			Mean Standard Deviation
110	26	Group A	100	63.2455532
10	178	Group B	101	76.01973428
90	99			
190				
180				
20				
50				
150				
130				
70				

Figure 9.2 – Example of two datasets being compared

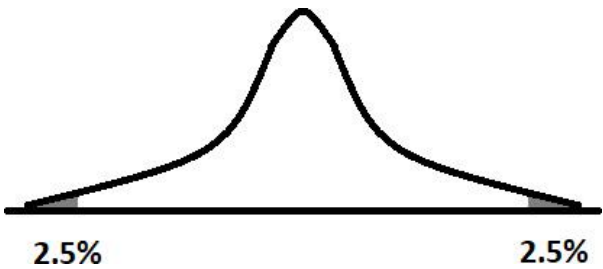


Figure 9.3 – Two-tailed



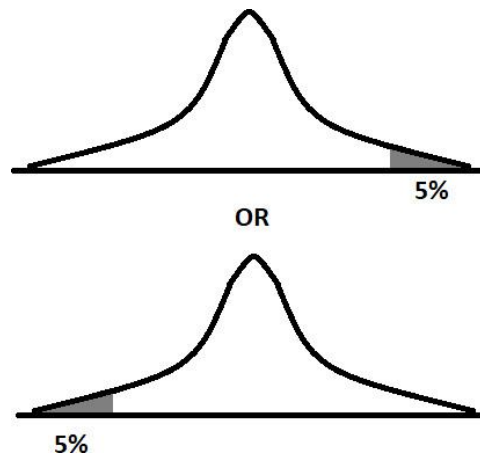


Figure 9.4 – One-tailed

**Error Cause**

Type I Falsely accept alternative hypothesis and reject null hypothesis

Type II

**Error Cause**

Type I Falsely accept alternative hypothesis and reject null hypothesis

Type II Falsely accept null hypothesis and reject alternative hypothesis

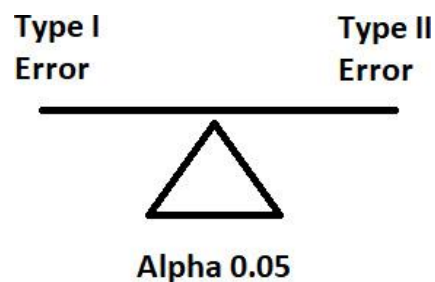


Figure 9.5 – Balance of type I and type II errors with alpha as 0.05

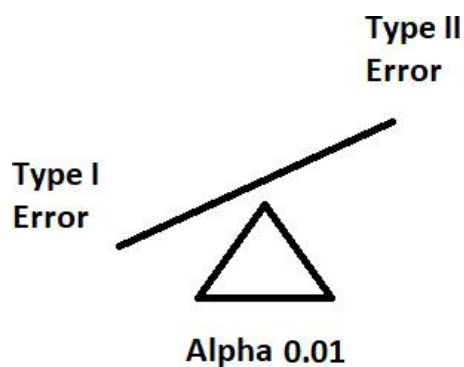


Figure 9.6 – Balance of type I and type II errors with alpha as 0.01

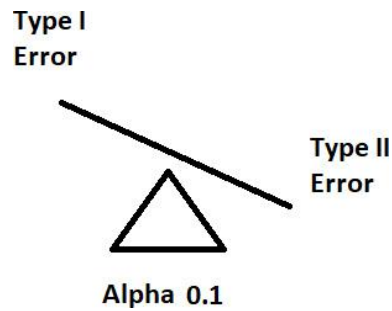


Figure 9.7 – Balance of type I and type II errors with alpha as 0.1

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. An e-commerce website runs an A/B study comparing two different website designs to see which generates more sales. What is the null hypothesis?
  - A. There is a statistically significant difference between group A and group B
  - B. Group A will be higher than group B
  - C. Group B will be higher than group A
  - D. Statistically, there is no significant difference between group A and group B
2. The A/B study from the previous question was finished. With an alpha of 0.05, the results were a p-value of 0.04. What do these results mean?
  - A. You should accept the null hypothesis and reject the alternative hypothesis
  - B. You should accept the alternative hypothesis and reject the null hypothesis
  - C. You should accept the null hypothesis and the alternative hypothesis
  - D. You should reject BOTH the null and alternative hypotheses
3. A small pet shop ran a study to see whether dog owners bought more pet food than other pet owners. After the study, they decided that there was no difference between the two groups, but they were mistaken – there actually is a difference. Dogs do eat more than most other pets. What type of error does this represent?
  - A. Type I error
  - B. Type II error
  - C. Type III error
  - D. Type IV error

4. You run a hypothesis test to compare the results of a new system to an old system. Using an alpha of 0.05, which of the following p-values would lead you to believe there is no difference between the new system and the old system?
  - A. 0.03
  - B. 0.008
  - C. 0.05
  - D. 0.4
5. If, in the scenario from the previous question, you find that there is a difference between the new system and the old system when there actually isn't. What type of error is this?
  - A. Type I error
  - B. Type II error
  - C. Type III error
  - D. Type IV error

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing.

1. The answer is: There is no statistically significant difference between group A and group B

Remember that the null hypothesis is that the two groups are the same and there is no statistically significant difference between them.

2. The answer is: You should accept the alternative hypothesis and reject the null hypothesis

When your p-value is equal to or smaller than your alpha, you accept  $H_1$  and reject  $H_0$ .

3. The answer is: Type II error

In this scenario, the analyst accepted the null hypothesis and rejected the alternative hypothesis when they should not have. This describes a type II error.

4. The answer is: 0.4

The value of 0.4 is the only number larger than 0.05. Because this p-value is larger than the alpha, we accept the null hypothesis, that there is no difference between the two groups, and reject the alternative hypothesis, that there is a difference.

5. The answer is: Type I error

If you reword the question, you will find that you are falsely accepting the alternative hypothesis and rejecting the null hypothesis. This is a type I error.

# Chapter 10

## Technical requirements

If you do decide to follow along, the data and code used can be found at the following URL:

<https://github.com/PacktPublishing/CompTIA-Data-DAO-001-Certification-Guide>

## Hands-On Section

### T-test practice

Let's be clear that you will not need to run this analysis and you do not need to know any specific programming language for this exam. The hope is that a little extra practice will help cement this analysis in your mind so that you will be clear on how it differs from the other analyses.

### T-test assumptions

First, for this example, we will be running a two-tailed independent t-test, so we need to know the requirements. The assumptions for an independent t-test are as follows:

- Independence
- Normality
- Homogeneity of variance
- $n \geq 30$
- The independent variable is categorical
- The dependent variable is numerical

The assumption of independence is simply that the two groups are independent of each other and that each observation is independent of the other.

The assumption of normality is that the data is normally distributed.

The assumption of homogeneity of variance is simply that the variance of both groups is around the same. If one variance is a little higher or lower than the other, it isn't a deal breaker, but if group A has a variance of 1 and group B has a variance of 10,000, it makes it really hard to compare them evenly.

This last one is not an actual assumption but rather, best practice. Technically, you can run a t-test with four observations in each group, or fewer. A t-test has no minimum sample size, but as a general rule of thumb, you get better results if you have at least 30 observations in each sample.

If you recall, an independent variable is what you are controlling or changing; in this case, the independent variable is what separates your two groups. Whether they are group A or group B derives your independent variable, and it doesn't matter what you call them.

The dependent variable needs to be a number since the t-test compares two groups of numbers.

All we need to start are two sets of normally distributed numbers that were collected separately, have similar variance levels, and have at least 30 observations each.

## T-test code

For this example, we will be using Jupyter Notebooks to run Python. There are many ways to do this, using several different tools. Even if you want to use Python, there are several different packages that will allow you to do the same thing. The data we will be using is a sample from a pet database. Only four breeds of dogs and cats were selected for this sample: a random small dog breed (Yorkshire Terrier), a random large dog breed (English Mastiff), a random small cat breed, (Singapura), and a random large cat breed, (Maine Coon). The dataset includes Type, Breed, Weight (lb), Height (in), and Sex of the pets, with 200 observations.

The very first thing we do is gather everything we need. This involves importing all of the packages for the upcoming tests:

```
#Importing modules
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy import stats
from sklearn.model_selection import train_test_split
```

To be clear, you are not going to need all of these packages for the t-test.

### Important note

If you are using Jupyter Notebooks through the Anaconda distribution, you will already have all of these packages installed, so you only have to import them. Otherwise, if you have never used these packages, you will need to install them first, or importing them will fail. There are lots of guides online that go over how to install these things if you are having trouble.

It is a best practice to have all of your `import` statements together, so we have included every package we will need for this chapter.

Before we go any further, we need to pull in our data:

```
#Importing dataset
MyData = pd.read_csv("PetExample.csv")
MyData.head()
```

Okay, here, we are creating a variable called `MyData`, and using it to hold out the dataset. In the next line, we are using the `head` function to look at the first couple of rows. We can see the results in **Figure 10.1**:

	Type	Breed	Weight (lb)	Height (in)	Sex
0	Cat	Maine Coon	15	11	Male
1	Dog	English Mastiff	223	32	Female
2	Cat	Singapura	6	8	Female
3	Cat	Singapura	4	7	Female
4	Cat	Singapura	8	8	Male

**Figure 10.1 – Preview of dataset**

It is a best practice to preview the dataset when you first pull it in before you use it. This makes sure the data is what you think it is and you can double-check that all of the expected variables are present.

Okay! Now we have pulled in our required packages and data, let's jump right in. First things first—we need to prepare the data for this specific analysis:

```
#Preparing the data
SmallCats = MyData[ (MyData["Breed"]=="Singapura") ]
SmallDogs = MyData[ (MyData["Breed"]=="Yorkshire Terrier") ]
```

For this method of running a t-test, we first need to subset our data. Each group we are comparing needs to be stored in its own variable. Because we are using DataFrames from pandas, we can use the preceding logic to filter for all observations that are marked with a specific breed. The `SmallCats` variable now stores every entry where the breed is Singapura and `SmallDogs` holds every entry where the breed is Yorkshire Terrier. Now, we can compare these breeds directly. To double-check, if you run the `head` function on `SmallCats`, you will get a result similar to **Figure 10.2**:

	Type	Breed	Weight (lb)	Height (in)	Sex
2	Cat	Singapura	6	8	Female
3	Cat	Singapura	4	7	Female
4	Cat	Singapura	8	8	Male
9	Cat	Singapura	5	8	Female
13	Cat	Singapura	5	7	Female

**Figure 10.2 – SmallCats preview**

As you can see, we still have every column of information. If you recall, we are only comparing two numerical variables (one for each group), so we have to subset the data again. If you are comfortable, there are shortcuts to streamline this process, but until you get the hang of it, we will go ahead and define the specific variables we will be using:

```
#Isolating the test variables
SmallCatsWeight = SmallCats["Weight (lb)"]
SmallDogsWeight = SmallDogs["Weight (lb)"]
```

Here, we are isolating the specific variables we will be using for each group. In this case, we want to see whether there is a significant difference in weight between Yorkshire Terriers and Singapuras. Now, if you use the `head` function on `SmallCatsWeight`, you will see something similar to **Figure 10.3**:

```
2    6
3    4
4    8
9    5
13   5
Name: Weight (lb), dtype: int64
```

**Figure 10.3 – SmallCatsWeight preview**

You see two columns of numbers, but the row on the left is the index, or where these values fit into the dataset as a whole. This is isolated to just the weight of Singapuras. Everything is ready to plug into our t-test! Let's go:

```
#Running the analysis
tTestResults = stats.ttest_ind(SmallCatsWeight, SmallDogsWeight)
print(tTestResults)
```

In this case, we are creating a new variable to hold our analysis. `stats.ttest_ind()` is the function that runs our test. `SmallCatsWeight` and `SmallDogsWeight` are the variables we are feeding into the test. Finally, we are printing the results. If you were to run this code, you would get something like **Figure 10.4**:

```
Ttest_indResult(statistic=2.1530549587998804, pvalue=0.033930048809027155)
```

**Figure 10.4 – T-test results**

This is what we wanted to know—specifically, the p-value. Since we are using an alpha of 0.05, the p-value of 0.03 is significant. Here, we accept the alternative hypothesis—that there is a difference between them—and reject the null hypothesis—that there is no difference between them. There is a significant difference in the weight of Yorkshire Terriers and Singapuras. If you are curious, you can use the `describe` function on each test variable. The results for Singapuras will be as seen in **Figure 10.5**:

```
count    53.000000
mean     6.094340
std      1.362466
min      4.000000
25%      5.000000
50%      6.000000
75%      7.000000
max      8.000000
Name: Weight (lb), dtype: float64
```

**Figure 10.5 – SmallCatsWeight.describe() results**

You can do the same thing with the Yorkshire Terrier weight variable to get a result, as seen in **Figure 10.6**:

```
count    41.000000
mean     5.536585
std      1.074653
min      4.000000
25%      5.000000
50%      6.000000
75%      6.000000
max      7.000000
Name: Weight (lb), dtype: float64
```

**Figure 10.6 – SmallDogsWeight.describe() results**

The Singapuras' mean weight is 6.1 lb and the Yorkshire Terriers' mean weight is 5.5 lb. There is only a 0.6 lb difference between them, but we did find this difference to be statistically significant. Singapuras weigh more than Yorkshire Terriers on average.

If you want extra practice, you can try comparing the height of Yorkshire Terriers and Singapuras, the weight of cats and dogs, the height of Singapuras and Maine Coons, or any other combination you think is interesting. Play around with the data, have fun, and try to create a pleasant memory associated with t-tests.

This wraps up the basics of t-tests. Next, we will jump into another common analysis: chi-square!

## Chi-square practice

For this example, we will be using the same dataset with the same packages as we did for the t-tests before. You can pick up exactly where you left off, in the same code you used for the t-tests. Here, we will specifically be performing a chi-square test for independence (sometimes called **Pearson's chi-square**) to compare pet **Type** and **Sex** variables to see whether there is a relationship between them.

## Chi-square assumptions

Okay, before we jump into assumptions, there is something you need to understand about a chi-square test for independence: it compares two categorical variables by analyzing a contingency table. Now, a contingency table is just another name for a frequency table that looks at more than one variable. We practiced making frequency tables earlier in this book in **Chapter 8, Common Techniques in Descriptive Statistics**. The main reason you need to know this is that some of the assumptions are based specifically on this contingency table. Here are the assumptions:

- Both variables are categorical
- Independence of observations
- Contingency cell exclusivity
- 80% of cells should have a value of at least 5
- $n \geq 50$

For chi-square, you are comparing two variables, and both of them must be categorical. The independence of observations is half of the assumption of independence that we covered with the t-tests. The variables themselves do not need to be independent, but each observation needs to be independent of every other.

Contingency cell exclusivity is simpler than it sounds. It just means that each observation is only counted once in the contingency table. You can't have observations that fall into more than one category and are counted multiple times. Either it is in one cell or another, but it can't be in both.

The next assumption is 80% of the cells on your contingency table should have a count of 5 or more. If you have a table with 10 cells, only 2 of those cells can have a value below 5. The general idea behind this is that if most of your cells have a really low count, it is difficult to get an accurate measurement.

Finally,  $n \geq 50$ . Chi-square is sensitive to sample size issues. It really helps if you have a decent number of values in most, if not all, of your cells. This leads us to a minimum sample size of 50.

### Important note

Another common practice is to take the number of possible outcomes of one variable and multiply it by the number of possible outcomes from the second variable. This gives you the number of cells. Then, you multiply the number of cells by 5. For example, if you are comparing **Color** (**Red**, **Blue**,



Yellow) to Shape (Circle, Square, Triangle), each has 3 possible outcomes.  $3 \times 3 = 9$ . There would be 9 cells in the contingency table. To get 5 observations in every cell you would need, at least,  $9 \times 5 = 45$ : 45 observations.

As long as you meet the other assumptions, the minimum sample size is flexible for this one. Pick a method that works for you and stick with it. Just remember that the larger your sample size, the more likely you are to have enough values in all of your cells, and the more likely you are to get accurate results.

### Chi-square code

Let's jump right into the code. Since we covered all of the code to get the packages installed and the data imported when we discovered t-tests, we can save some time here by not repeating ourselves. Instead, we will create a contingency table that contains our two categorical variables: Type and Sex. With the pandas package, we can do this with a single line of code:

```
#Preparing the data
Contingency = pd.crosstab(MyData["Type"],MyData["Sex"])
print(Contingency)
```

Here, we are creating a variable called Contingency, then using the pd.crosstab function to fill it with a contingency table using our two variables of choice. Then, we print the results. The output should look something like Figure 10.7:

Sex	Female	Male
Type		
Cat	47	64
Dog	50	38

Figure 10.7 – Contingency table

This should look familiar, as a rather simple frequency table. We see in this sample that there are more female dogs than male dogs and more male cats than female cats, but is this just chance, or is there a relation between these two variables? The only way to know for sure is to run our analysis:

```
#Running the analysis
stat,p,dof,expected = stats.chi2_contingency(Contingency)
print(p)
```

You might notice that this looks slightly different. This test actually uses the contingency table to create four different output variables: stat, p, dof, and expected. For now, the only one we care about is p, which holds the p-value of this analysis. Let's look at the results of our chi-square analysis in Figure 10.8:

0.059261801629914686

Figure 10.8 – Chi-square p-value

This p-value is larger than our alpha of 0.05, if only by a little. This means that we accept the null hypothesis and reject the alternative hypothesis. For a chi-square, that means that we are saying that, despite how the numbers may look in the table, there is not a statistically significant relationship between the Type and Sex variables.

That's it for chi-square, so let's move on to another common analysis: correlation.

## Correlation practice

For our example, we will be running a Pearson's correlation analysis to see whether there is a relationship between the weight of the pets and their size. Pearson's is the most common type of correlation analysis and looks at two numerical variables: in this case, **Weight** and **Height**.

## Correlation assumptions

Looking specifically at the assumptions for a Pearson's correlation, we will see some assumptions we have run into before and some we have not:

- Level of measurement
- Linearity
- Normality
- Related pairs
- Lack of outliers
- $n \geq 30$
- Two continuous variables

The level of measurement can get needlessly complicated very quickly. There are four levels that give increasingly more meaning to the values of numbers. You can make arguments until you are blue in the face, but if you have two continuous variables (which you should have for this analysis anyway), then you meet this requirement. In our case, height and weight are both physical measurements, where the distance between one number and another has value, which technically means their level of measurement is Interval, which is the second highest in the order. Long story short, have two continuous variables and you knock out two assumptions in one go.

Linearity is a little more straightforward. If you do a scatter plot and draw a line through the dots, it should be a straight line. If there is a defined shape to the dots, but it is a wave or a curve instead of a straight line, then you don't meet this one. That said, if it is just a random ball of dots, then go ahead and try!

We have talked about normality before, so we can skip that one for now.

Related pairs mean that for every data point in one variable, there should be a matching one in the other. In this case, every pet has a weight measurement and a height measurement, so there is a specific pair of values for every observation.

A lack of outliers is exactly what it sounds like. You should not have any outliers in your dataset. This is a best practice for any dataset, but a single outlier can ruin the results of this analysis.

Generally speaking, a larger sample size is better, up to a point, but the accuracy of correlation analysis benefits from this more than most analyses. A minimum sample size of 30 ensures a minimum level of accuracy.

## Correlation code

The actual code for this one is straightforward. Let's start by defining our variables:

```
#Preparing the data
PetHeight = MyData["Height (in)"]
PetWeight = MyData["Weight (lb)"]
```

This is just setting up our `PetHeight` and `PetWeight` variables and filling them with the appropriate data from our dataset. Next, we run the analysis!

```
#Running the analysis
CorrelationResults = stats.pearsonr(PetHeight, PetWeight)
print(CorrelationResults)
```

Again, we are creating a variable and then filling it with the `stats.pearsonr` function using the variables we prepared ahead of time. Let's look at the results in **Figure 10.13**:

```
(0.9592991777202481, 3.8891158549349756e-110)
```

**Figure 10.13 – Correlation results**

This returns two values: the correlation coefficient and the p-value, respectively. A correlation coefficient of 0.96 would definitely count as a strong positive correlation. The p-value is so ridiculously tiny that we must accept the alternative hypothesis and reject the null hypothesis. There is a relationship between these two variables—a strong one.

This wraps up correlations, so let's move on to simple linear regression.

## Simple linear regression practice

Here, we will see whether we can predict a pet's height by using its weight with simple linear regression. That means `Weight` is our independent variable and `Height` is our dependent variable. Before we can go into how to do this, we need to look at the requirements.

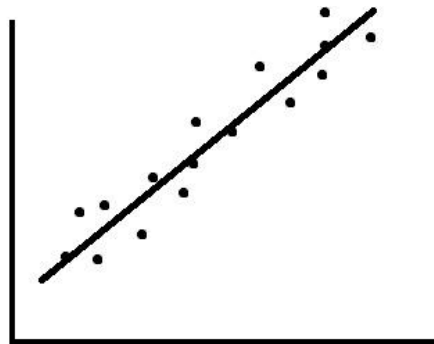
### Simple linear regression assumptions

We have seen most of these assumptions before, but there is one that will throw us for a bit of a loop. The assumptions are as follows:

- Linearity
- Normality
- Independence
- Homoscedasticity
- The dependent variable is numeric
- The independent variable is numeric
- $n \geq 100$

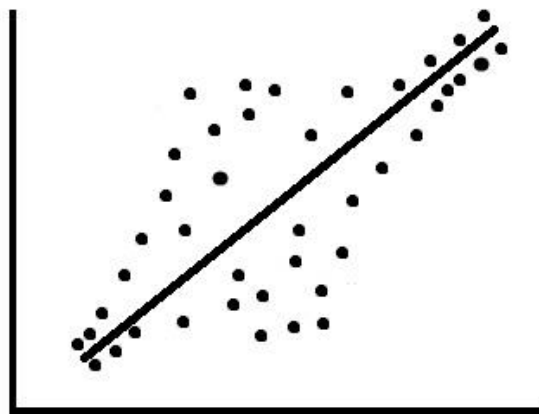
We have already talked about assumptions of linearity, normality, and independence, so we can skip those for now. Let's go over the concept of homoscedasticity. Technically, the definition of homoscedasticity is something about the variance in the residual values remaining constant for

different levels of the independent variable. Let's break this into something useful. Residuals are just the distance every point is away from the line, so all this is saying is that the dots are roughly the same distance from the line all the way down its length. Let's see what this looks like in [Figure 10.19](#):



**Figure 10.19 – Homoscedasticity**

All of the points are roughly a similar distance from the line all the way down its length. This even spread of points is a good example of homoscedasticity. Let's take a moment to look at an example in [Figure 10.20](#) of something that would fail the homoscedasticity assumptions:



**Figure 10.20 – Not homoscedasticity**

In the preceding example, it starts out with the points all really close to the line, then the dots really spread out, then come back near the line. Even if this would pass the assumption of linearity, it would not pass the assumption of homoscedasticity.

Again, we are creating a scatter plot and drawing a line through it, so the independent and dependent variables both must be numbers. When it comes to the sample size, you can technically get away with 10 observations per independent variable. In simple linear regression, there is only 1 independent variable, so the minimum sample size is 10. That said, common practice is at least 100 observations for best results.

### Simple linear regression code

Let's jump right in with preparing the data for simple linear regression:

```
#Preparing the data  
Y = MyData["Height (in)"]
```

```
X = MyData["Weight (lb)"]
XConstant = sm.add_constant(X)
```

This is pretty straightforward. First, we define our `X` and `Y` variables, by filling them with `Weight (lb)` and `Height (in)`, respectively. Now, for this package, we need to define a constant. We do this by plugging our `X` variable in the `sm.add_constant()` function and saving it in a new variable, `XConstant`.

Now, it's time to actually run the analysis!

```
#Running the analysis
SimpleLinearRegressionResults = sm.OLS(Y,XConstant).fit()
print(SimpleLinearRegressionResults.summary())
```

Okay, so we are doing a few things here. First, we are creating a `SimpleLinearRegressionResults` variable to hold the results.

### Important note

Variable names should be short, simple, and clear. For the examples in the book, a lot of the variables have very explicit names that tell you exactly what they hold, such as `SimpleLinearRegressionResults`. This is to make the example as clear as possible. When you are working with your own code, you are encouraged to use something shorter, as long as it is clear what it is. For example, you could try `SLRMod` or `LinReg`. Just pick a naming convention and stick with it.

The `sm.OLS()` function creates a model using `Y` and `XConstant`, and the `.fit()` function applies it to our data. Just like that, the model is ready to go. That wasn't so bad, right? Now, this model has a few built-in functions. Here, we use the `.summary()` function to give us the results of our model. We can see what this looks like in [Figure 10.21](#):

```

=====
                        OLS Regression Results
=====
Dep. Variable:          Height (in)      R-squared:                0.920
Model:                  OLS              Adj. R-squared:           0.920
Method:                 Least Squares     F-statistic:              2273.
Date:                  Mon, 22 Aug 2022   Prob (F-statistic):       3.89e-110
Time:                  16:02:24          Log-Likelihood:          -487.92
No. Observations:      199              AIC:                     979.8
Df Residuals:          197              BIC:                     986.4
Df Model:               1
Covariance Type:       nonrobust
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
const                8.5104         0.241     35.272     0.000         8.035         8.986
Weight (lb)          0.1220         0.003     47.680     0.000         0.117         0.127
=====
Omnibus:                 13.770      Durbin-Watson:           2.314
Prob(Omnibus):           0.001      Jarque-Bera (JB):        12.161
Skew:                   0.530      Prob(JB):                0.00229
Kurtosis:                2.413      Cond. No.                114.
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

**Figure 10.21 – Simple linear regression results summary**

This is a lot of data, but a data analyst is not a statistician, so we don't care about most of it. In the upper-right corner, we see R-squared, which is just another way of saying  $R^2$ . Our  $R^2$  here is 0.92, which means 92% of the variance in height can be explained by weight. Next, we will find the p-value. Here, it is listed as  $P>|t|$ , but it is still just a p-value. Also, it is rounded to 0.000, so we can safely say that our p-value is pretty small. Here, we would accept the alternative hypothesis and reject the null hypothesis; you can predict, roughly, a pet's height based on how much it weighs.

## Figures

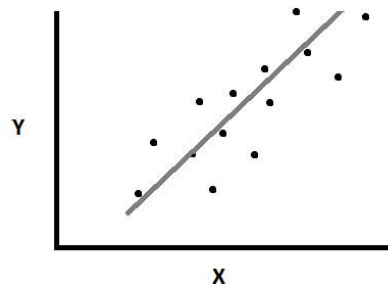


Figure 10.9 – Positive correlation

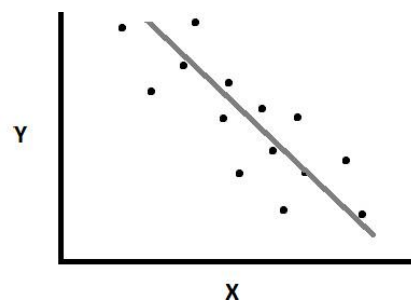


Figure 10.10 – Negative correlation

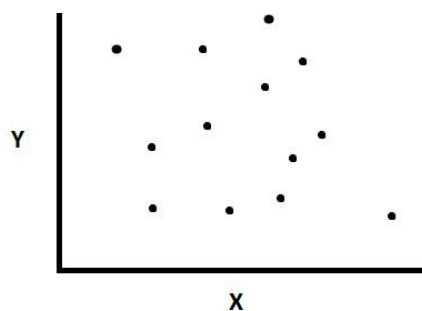


Figure 10.11 – No correlation

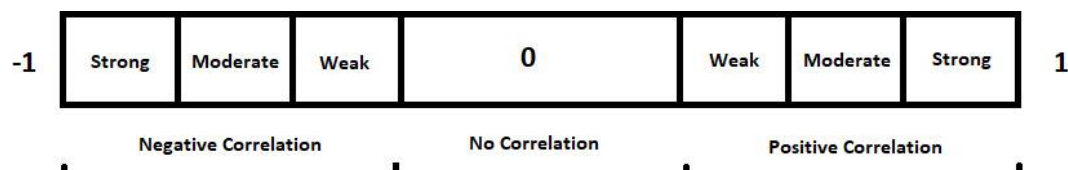


Figure 10.12 – Correlation coefficient

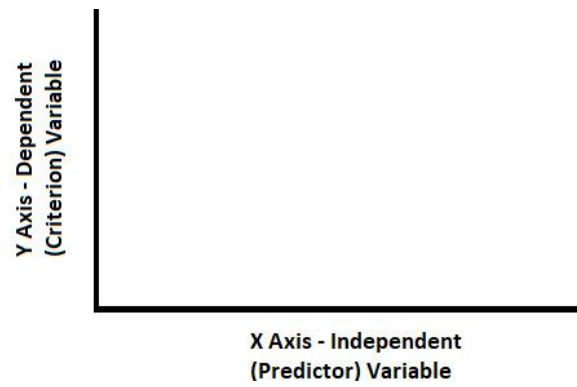


Figure 10.14 – Linear regression variable placement

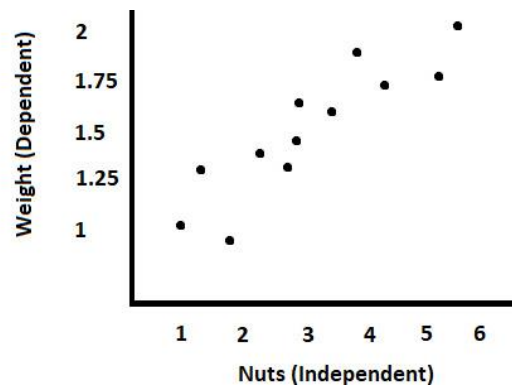


Figure 10.15 – Nuts and weight scatter plot

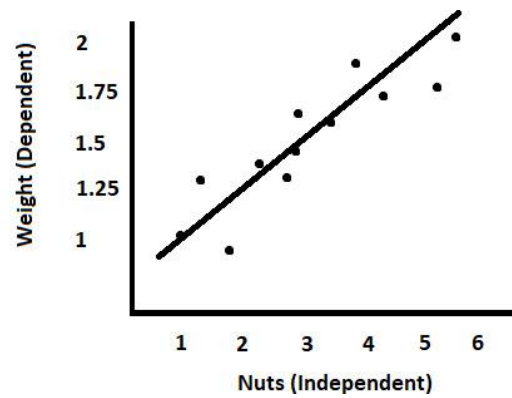


Figure 10.16 – Nuts and weight linear regression line

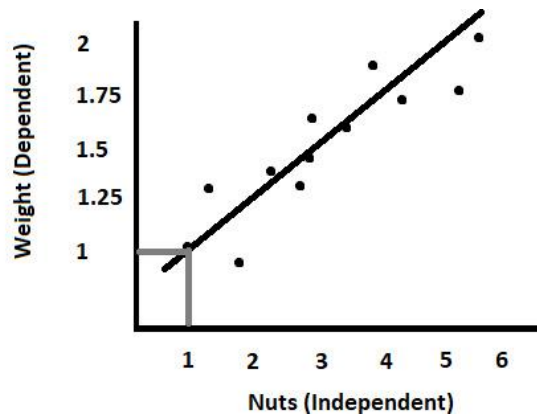


Figure 10.17 – Prediction of one nut

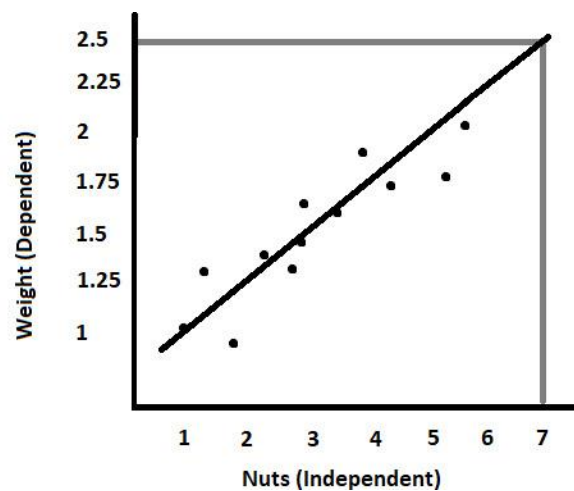


Figure 10.18 – Prediction of seven nuts

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. **Simon's Grocery Store** collects data on the number of sales of every vegetable it sells every day. If you wanted to see whether there is a statistically significant difference between the number of sales of pumpkins and the number of sales of summer squash, which analysis would be most appropriate?
  - A. T-test
  - B. Chi-square
  - C. Correlation
  - D. Simple linear regression



2. Now, **Simon's Grocery Store** groups cereal brands into Great Sellers and Poor Sellers based on sales performance in the past. It also notes the color of the box for each brand. The grocery store would like to know whether there is a relationship between the color of the box and whether the brand is a great seller or a poor seller. Which analysis would be most appropriate?
  - A. T-test
  - B. Chi-square
  - C. Correlation
  - D. Simple linear regression
3. **Simon's Grocery Store** collects data on every product it sells. Now, it is wondering whether there is a relationship between the number of turkeys sold and the number of packages of premade stuffing sold. Which analysis would be most appropriate?
  - A. T-test
  - B. Chi-square
  - C. Correlation
  - D. Simple linear regression
4. No matter the results of the previous question, **Simon's Grocery Store** wants to know whether it can predict the number of packages of premade stuffing it will sell based on the number of turkeys it sells. Which analysis would be most appropriate?
  - A. T-test
  - B. Chi-square
  - C. Correlation
  - D. Simple linear regression

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: **T-test**

Here, you are comparing two numeric variables to see whether there is a difference between them, and that makes this a job for a t-test.

2. The answer is: **Chi-square**

In this question, we are comparing two categorical variables to see whether there is a relationship between them. This is a good time to use a chi-square test for independence.

3. The answer is: *Correlation*

For this question, we are seeing whether there is a relationship between two numeric variables. That means we are using correlation to see whether there is a relationship and how strong it is.

4. The answer is: *Simple linear regression*

Here, we are seeing whether we can predict one numeric variable using another numeric variable. That means the first thing we should think about is simple linear regression.

# Chapter 11

## Links

Tableau Dashboards: If you are interested in what these might look like, visit [public.tableau.com/app/discover/viz-of-the-day](https://public.tableau.com/app/discover/viz-of-the-day).

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. A bank, **Second Rational Bank**, requires all of its tellers to have access to metrics such as the interest rates and fees on savings accounts, which change regularly. Which type of report would be most appropriate for this need?
  - A. Point-in-time report
  - B. Real-time report
  - C. Research report
  - D. Ad hoc report
2. **Second Rational Bank** is considering buying a package of loan debts from a sister company. It wants to know what the risks are, how much it can offer the other bank while still making a profit, and what it will need to do after to collect on the loans. Which type of report would be most appropriate?
  - A. Self-service report
  - B. Ad hoc report
  - C. Recurring report
  - D. Research report
3. A manager at **Second Rational Bank** wants to track the metrics of their team and make sure they are meeting all rules and regulations of the company. Which type of report is most appropriate?
  - A. Recurring report
  - B. Research report
  - C. Ad hoc report

- D. Self-service
- 4. Which of the following is not a data analytics tool?
  - A. SAS
  - B. SPSS
  - C. STS
  - D. Stata
- 5. Which of the following data analytics tools is specifically designed for visualizations and reports?
  - A. Power BI
  - B. Microsoft Excel
  - C. SAS
  - D. Domo

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

- 1. The answer is: **Real-time report**

The bank tellers need quick responses with the most up-to-date information and little to no complexity. This is an ideal time to use a real-time, or dynamic, report.

- 2. The answer is: **Research report**

This is a big question with a lot of parts to it and can mean a gain or loss of millions of dollars to the company. The ideal solution is to take your time and write up a full research report.

- 3. The answer is: **Recurring report**

Specifically, this would be a compliance report, which is a type of recurring report, because the goal was to check metrics against requirements and regulations.

- 4. The answer is: **STS**

Yes, you can create visualizations and reports with any of these tools, but Power BI is specifically designed as a visualization and reporting software. Make sure you learn not only all of the tools on the list but the group that gives them a general purpose.

- 5. The answer is: **Power BI**

Yes, you can create visualizations and reports with any of these tools, but Power BI is specifically designed as a visualization and reporting software. Make sure you learn not only all of the tools on the list but the group that gives them a general purpose.

# Chapter 12

## Figures

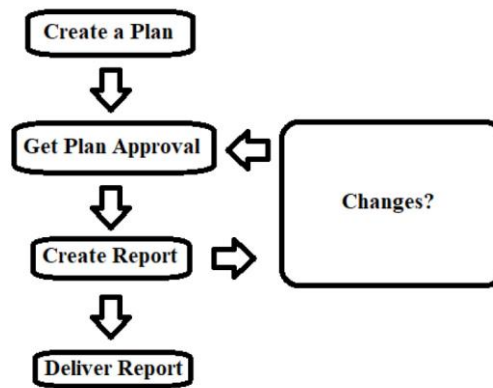


Figure 12.1 – Report creation flowchart

Date	Time(s)	Vol	Ins	Res	Operator	Distance(m)	RobotWeight(lb)
7/5/2019	33	1.9470	471	40	Oscar Mayor	22	9
7/12/2019	35	1.3958	581	86	Oscar Mayor	25	11
7/19/2019	44	1.3905	304	4	Oscar Mayor	23	10
7/26/2019	30	1.5938	482	15	Oscar Mayor	24	8
8/2/2019	43	1.1938	561	51	Oscar Mayor	19	11
8/9/2019	33	1.8404	309	27	Oscar Mayor	21	9
8/16/2019	35	1.4893	701	83	Oscar Mayor	16	9
8/23/2019	40	1.2843	824	79	Oscar Mayor	15	10
8/30/2019	35	1.9837	754	73	Oscar Mayor	22	11
9/6/2019	43	1.7489	601	48	Oscar Mayor	24	9

Figure 12.2 – Example dataset

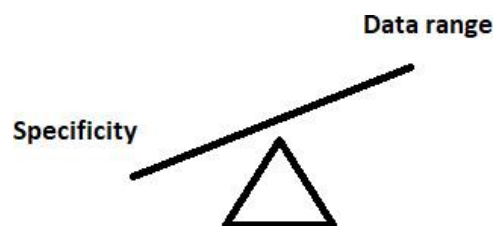


Figure 12.3 – Low specificity, high data range

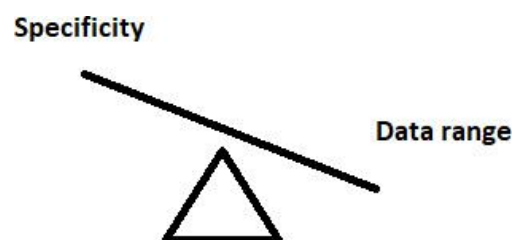


Figure 12.4 – High specificity, low data range

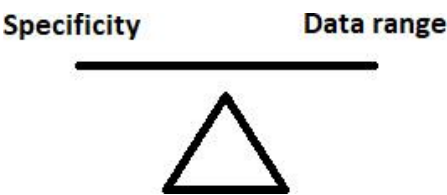


Figure 12.5 – Moderate specificity and data range

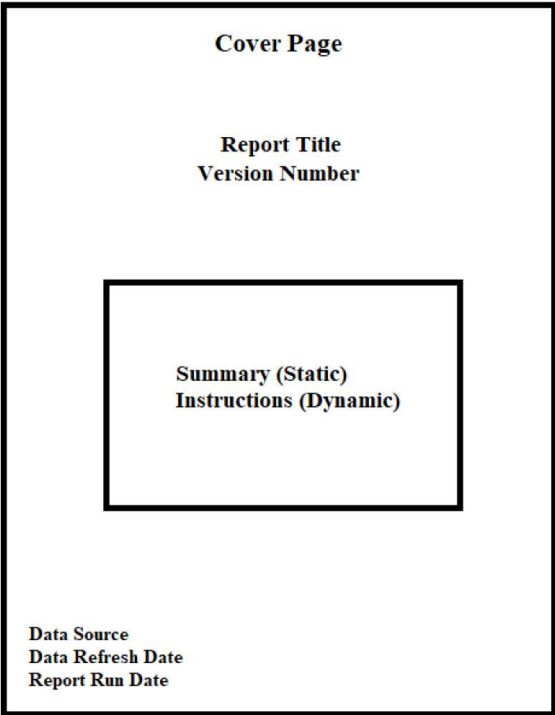


Figure 12.6 – Cover page example

word word word word word  
word word word word word  
word word word word word

Figure 12.7 – Tiny font

word word word word word  
word word word word word  
word word word word word

Figure 12.8 – Line spacing

word word word word word  
word word word word word  
word word word word word

Figure 12.9 – High contrast font and background

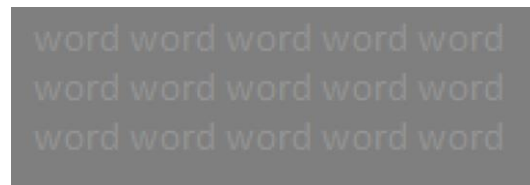


Figure 12.10 – Low contrast font and background

*Cursive Fonts*  
**STYLIZED FONTS**  
 Messy Fonts  
 Weird Fonts

Figure 12.11 – Unprofessional fonts

Calibri  
 Cambria  
 Times New Roman

Figure 12.12 – Common professional fonts

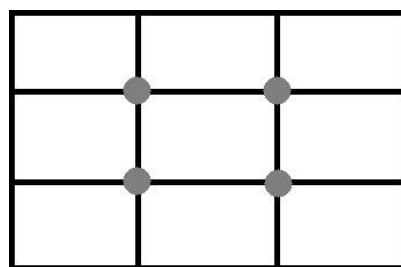


Figure 12.13 – Rule of thirds

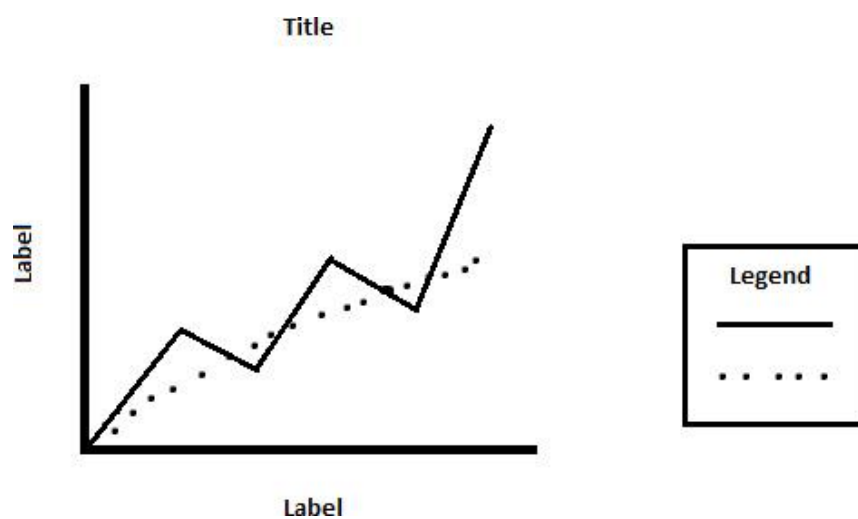


Figure 12.14 – Title, label, and legend



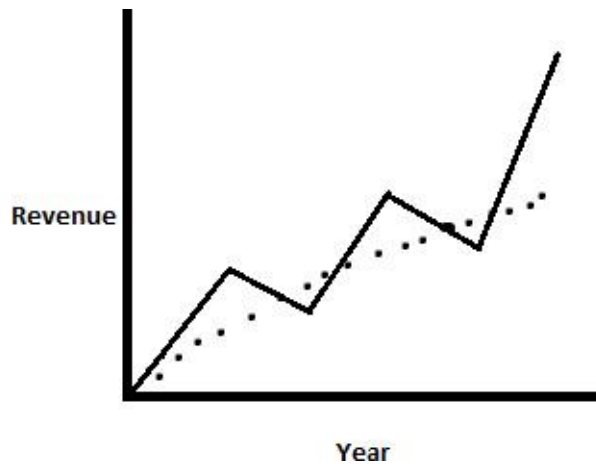


Figure 12.15 – Example chart

## Links

If you want to know more about color theory, there are a ton of resources, free and otherwise, that will give you a basic rundown. A fun one is [www.canva.com/colors/color-wheel/](https://www.canva.com/colors/color-wheel/) because it is interactive and lets you see how the different schemes work.

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. What is the first step in the development process for a dashboard?
  - A. Get approval
  - B. Make a plan
  - C. Create a dashboard
  - D. Deploy the dashboard
2. At a dairy farm, a low-level manager is responsible for a certain number of cows. They would like a report a couple of times a month on how much milk each cow is producing to see whether any of them require special care. What is an appropriate audience for this report?
  - A. Only the manager
  - B. Every manager
  - C. The manager and the farmhands
  - D. Everyone in the company needs this report

3. In the scenario from the previous question, what is the most appropriate frequency of the report?
  - A. Daily
  - B. Weekly
  - C. Monthly
  - D. Yearly
4. The day after releasing your milk report, you find a major mistake, and have to rerun the report to create a new version, but can use the data from the original report. Which of the following report elements will need to be updated?
  - A. FAQs
  - B. The version number
  - C. The version number, the report run date, and the data refresh date
  - D. The version number and the report run date
5. You decide to create a static dashboard for the milk production report, and you want it to automatically refresh and send out invitations every Sunday night. Which delivery consideration should you use?
  - A. Filtering
  - B. Scheduled delivery
  - C. Subscription
  - D. Dashboard optimization
6. While designing the dashboard for milk production, you have to pick a color scheme for your charts. What colors are most appropriate?
  - A. High-contrast colors
  - B. Low-contrast colors
  - C. Moderate-contrast colors
  - D. Company brand colors

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Make a plan

The first step of the development process is to make a plan. You cannot get your plan approved, create the dashboard, or deliver it without planning first it.

2. The answer is: Only the manager

The information in the report is only applicable to this one person. The manager might make decisions based on this report that will impact those above or below them, but they are the only one who is directly interested in the production rates of these cows.

3. The answer is: Weekly

The question specifically states that they want the reports a couple of times a month, and this is the only option that fits that requirement.

4. The answer is: The version number and the report run date

The report was rerun on a different day, generating a new version of the same report. This requires a new version number and a new report run date. The question states that you used the same data, so the data refresh date does not change, and there is nothing in the question about the FAQs.

5. The answer is: Subscription

Subscription is the only option that automatically delivers an updated version of the report at the same time every week.

6. The answer is: Company brand colors

Remember that the company brand takes priority over every design decision. The amount of color contrast does not matter at all if the company has a color scheme as part of its brand. Always use branding guidelines if they exist.

# Chapter 13

## Figures

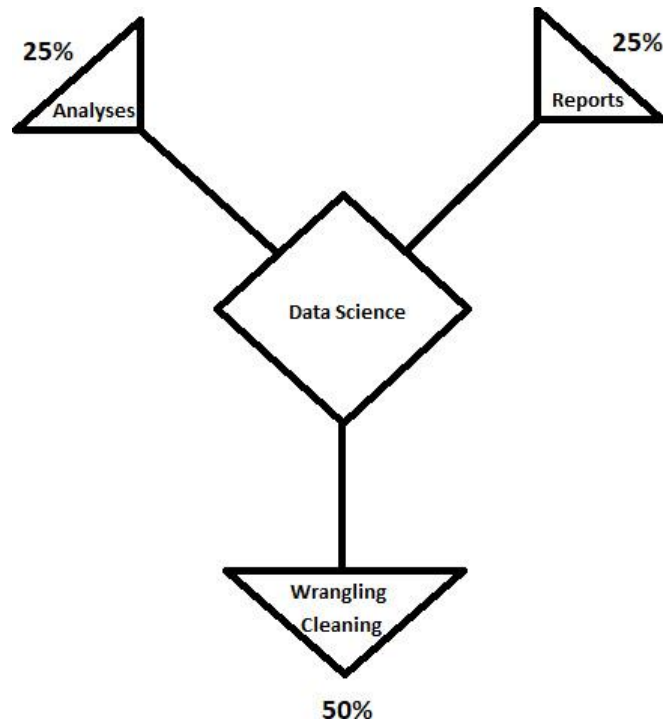


Figure 13.1 – Infographic

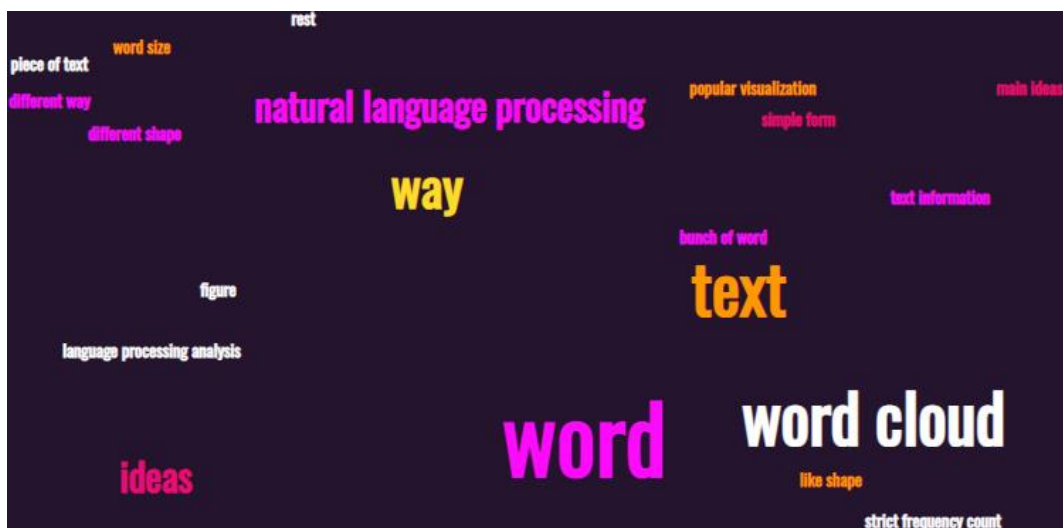


Figure 13.2 – Word cloud

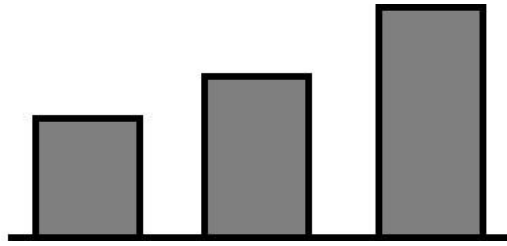


Figure 13.3 – Bar chart

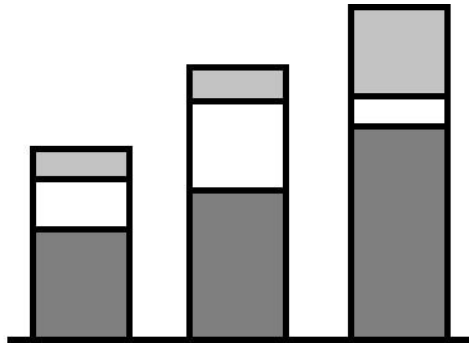


Figure 13.4 – Stacked bar chart

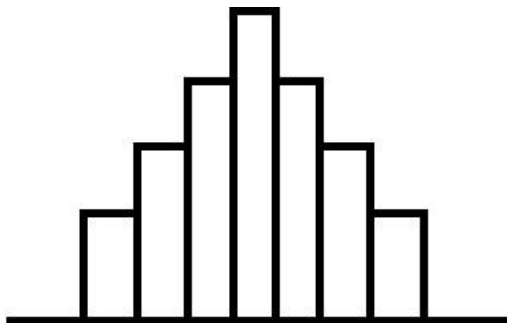


Figure 13.5 – Histogram

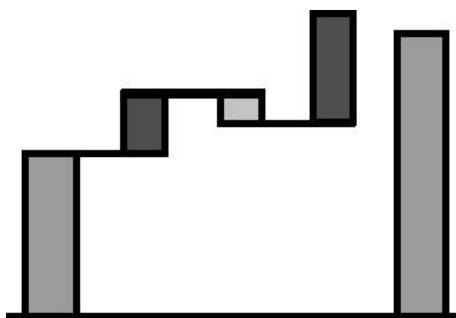


Figure 13.6 – Waterfall chart

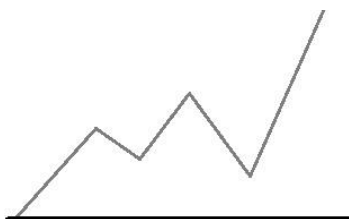


Figure 13.7 – Line chart

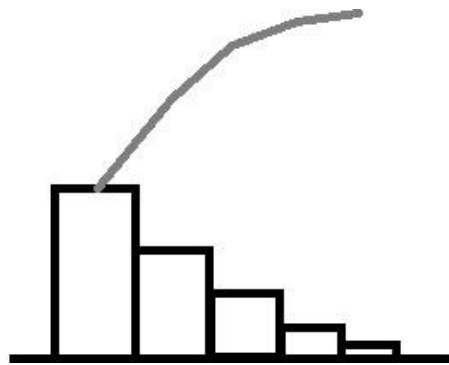


Figure 13.8 – Pareto chart

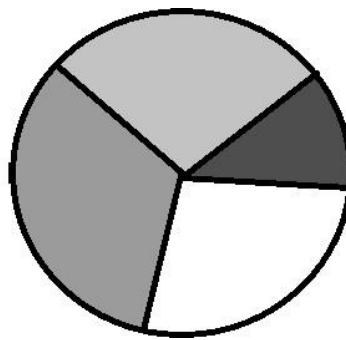


Figure 13.9 – Pie chart



Figure 13.10 – Scatter plot

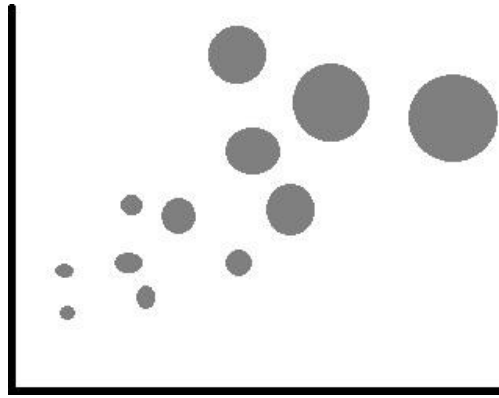


Figure 13.11 – Bubble chart



Figure 13.12 – Heat map

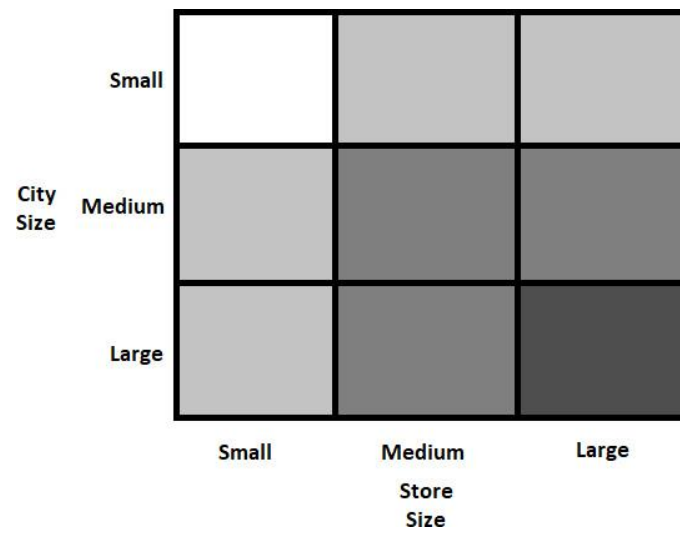


Figure 13.13 – City versus store size heat map



Figure 13.14 – Tree map

Product Sales By Department

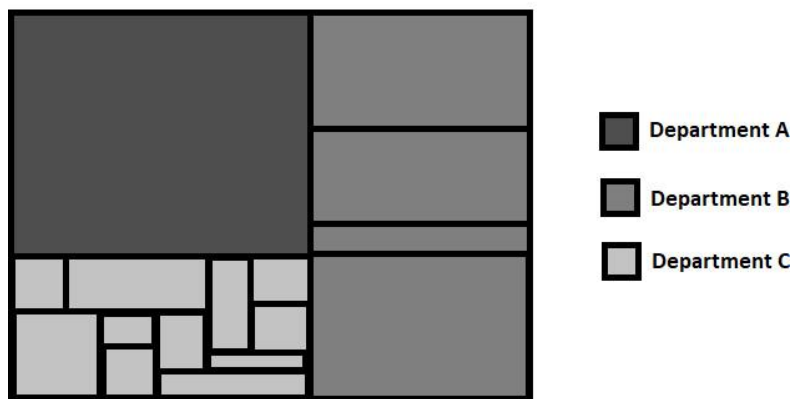


Figure 13.15 – Product sales by department tree map

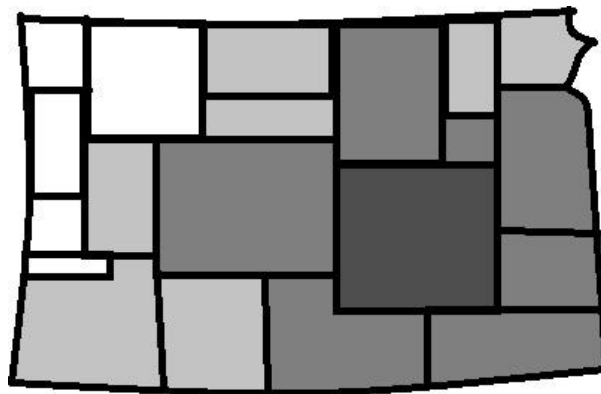


Figure 13.16 – Geographic map



## Links

If you are interested, you can check out [www.visme.co/make-infographics/](http://www.visme.co/make-infographics/) for a free tool.

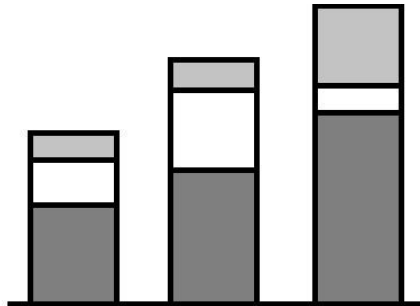
There are several ways to make a word cloud, but, if you just want to play around a little bit, you can try a free tool, such as [www.monkeylearn.com/word-cloud/](http://www.monkeylearn.com/word-cloud/). Try it out, and if you are trying to visualize a text analysis, think of word clouds first.

## Practice questions

Let's test our knowledge of the material in this chapter with a few example questions.

### Questions

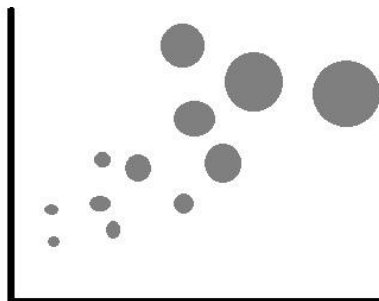
1. The following data visualization is representative of which type?



- A. Bar chart
  - B. Stacked bar chart
  - C. Histogram
  - D. Pie chart
2. If you wanted to represent a single quantitative variable over time, which data visualization type would be most appropriate?
    - A. Scatter plot
    - B. Bar chart
    - C. Heat map
    - D. Line chart
  3. The following data visualization is representative of which type?



- A. Heat map
  - B. Tree map
  - C. Geographic map
  - D. Bar chart
4. If you wanted to track a quantitative variable over a country, which type of data visualization would be most appropriate?
- A. Bubble chart
  - B. Geographic map
  - C. Heat map
  - D. Tree map
5. The following data visualization is representative of which type?



- A. Scatter plot
  - B. Dot plot
  - C. Bubble chart
  - D. Pie chart
6. A project manager wants a chart to track progress throughout a project, and they want to know, specifically, how each point differs from the previous point. Which data visualization is most appropriate?
- A. Waterfall chart
  - B. Line chart

- C. Bar chart
  - D. Stacked bar chart
7. If you wanted to display a qualitative variable with two different levels, and a qualitative variable on the same chart, which type of visualization would be most appropriate?
- A. Waterfall chart
  - B. Line chart
  - C. Bar chart
  - D. Stacked bar chart

## Answers

Now, we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Stacked bar chart  
  
The displayed chart is a stacked bar chart. You can tell because the bars are divided with different colors representing different groups.
2. The answer is: Line chart  
  
While a few of these technically could be used to track a quantitative variable over time, the line chart is specifically designed for it, making it the most appropriate.
3. The answer is: Heat map  
  
Heat maps show regular grids representing ordinal variables, usually with a quantitative variable representing the color.
4. The answer is: Geographic map  
  
Since we want to look at data over physical space, in this case, a country, a geographic map would be most appropriate.
5. The answer is: Bubble chart  
  
Here, we see what looks like a scatter plot, but the dots are different sizes, representing a third quantitative variable.
6. The answer is: Waterfall chart  
  
Again, there is more than one type of chart that can display this information, but waterfall charts are designed for this kind of visualization.
7. The answer is: Stacked bar chart

Remember that stacked bar charts represent two qualitative variables OR two levels of the same qualitative variable and a quantitative variable.

# Chapter 14

## Figures

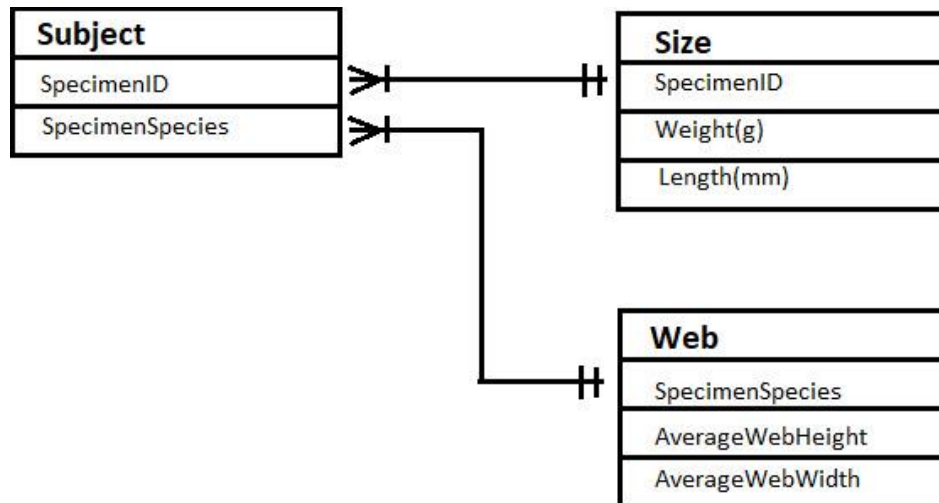


Figure 14.1 – Entity relationship diagram

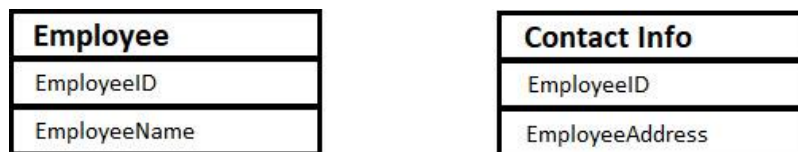


Figure 14.2 – One-to-one cardinality



Figure 14.3 – One-to-many cardinality

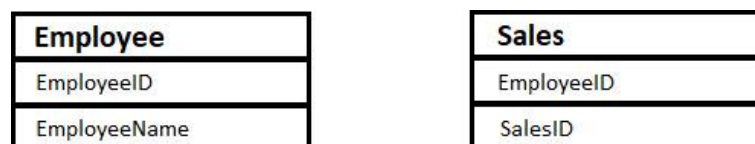


Figure 14.4 – Many-to-many cardinality

## Links

If you are unsure about the encryption requirements of your area, <https://www.cpdigital.org/world-map-of-encryption/> is a great resource.

If you are interested in more information on GDPR, you can look it up at <https://gdpr-info.eu/>.

You can learn more about hipaa at <https://www.hhs.gov/hipaa/index.html>, which is a good resource.

You can learn all about Payment Card Industry Security Standards at <https://www.pcisecuritystandards.org/>

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. Data4U, a **Software as a Service (SaaS)** company, would like to create a small dataset that includes de-identified data about how their clients have improved with the use of their software. They want every sales representative in the company to have access to this data to show to potential clients. What form of access requirement is this?
  - A. User group-based
  - B. Encryption-based
  - C. Role-based
  - D. Transmission-based
2. Which part of the Data Use Agreement includes explicit details about how the data is **not** supposed to be used?
  - A. Acceptable use policy
  - B. Data processing
  - C. Data deletion
  - D. Data retention
3. If you suspect that a data breach has occurred, which of the following is an appropriate response?
  - A. Delete the data
  - B. Ignore it
  - C. Inform the impacted parties

- D. Do nothing
- 4. Which of the following variables would be considered PII?
  - A. Geolocation
  - B. Social media post
  - C. Social Security number
  - D. All of the above
- 5. A filter that only allows a specific kind of data to be entered into a dataset can be considered which kind of entity relationship restriction?
  - A. Record link restriction
  - B. Data constraint
  - C. Cardinality
  - D. None of the above

## Answers

Now we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: Role-based

The company does not care what group or department the person is in; they want every person in the company who has the role of sales representative to have access to this data. If they didn't care about the role and asked for everyone in the sales department to have access, it would have been user group-based.

2. The answer is: Acceptable use policy

Remember that this section not only outlines ways the data can be used but also ways it should never be used.

3. The answer is: Inform the impacted parties

Always report any suspected data breaches, and always make sure that the impacted parties are informed that their data may have been stolen.

4. The answer is: All of the above

All of these pieces of information can be used to identify a specific individual, making them all PII.

5. The answer is: Data constraint

Here, we are limiting the type of data that can be entered or stored in a specific data object, which makes this a data constraint.



# Chapter 15

## Practice questions

Let's try to practice the material in this chapter with a few example questions.

### Questions

1. Which of the following is an appropriate time to check the quality of your data?
  - A. After data manipulation
  - B. After data transformation
  - C. Before the final report
  - D. All of the above
2. Which of the following is a data quality dimension?
  - A. Data completeness
  - B. Data retention
  - C. Rows passed
  - D. Data manipulation
3. Which of the following is a structured formal process for identifying the quality and efficiency of an entire database?
  - A. Cross-validation
  - B. Spot check
  - C. Reasonable expectations
  - D. Data profiling
4. **Over 9,000!** is a power company that is rapidly growing in its area. Recently, it purchased another power company that was one of its major competitors. This is an appropriate time to institute MDM. True or false?
  - A. True
  - B. False
5. Creating a document that explains what the variables in a dataset are, how they are used, and how they connect to one another represents which part of the MDM process?

- A. Data audits
- B. Consolidation
- C. Data dictionary
- D. Standardization

## Answers

Now we will briefly go over the answers to the questions. If you got one wrong, make sure to review the topic in this chapter before continuing:

1. The answer is: All of the above

These are all circumstances where you should check for quality.

2. The answer is: Data completeness

Data completeness is the only data quality dimension listed. Data retention is a section in the data use agreement, rows passed is an aspect of a data quality rule, and data manipulation is a general concept of data shaping.

3. The answer is: Data profiling

While these are all methods of validating quality, data profiling is the only structured approach to checking the quality of an entire database.

4. The answer is: True

One company buying another is called an acquisition, and acquisitions and mergers are one of the major reasons for MDM. This is because you have two completely separate datasets, one from each company, and you have to merge them into one dataset that is still somehow usable.

5. The answer is: Data dictionary

A data dictionary is, as you would expect, a dictionary of your data. It gives definitions for every variable, as well as how they are used and how they relate to other variables. These are very useful tools and an important part of the MDM process.

# Chapter 16: Practice Exam One

This practice exam will give you the correct number of questions in the correct proportions. There are certain types of questions or interactive dashboards that cannot be simulated in a book, but this exam will give you a good idea of what material you know well and what material you may need to review. If you know the material well, the format of the questions should not be a problem. Also note that CompTIA is constantly updating and adding new questions on the same topics, written by different subject matter experts. Some of these questions will be confusing, poorly written, and may not even follow the CompTIA guidelines. They have to test questions to see whether they work, and they don't always. Expect a weird question or two, and don't let it throw you off.

To make this practice exam as useful as possible, you should follow the testing experience as closely as possible. Make sure you are in a quiet and clean environment before you commence and have access to a simple calculator, a timer, and the ability to take notes.

You have 90 minutes to answer 90 questions, so you will need to average 1 question per minute. That said, some questions are quick and easy, while others will take several minutes. If you are stuck on a question, it may be best to guess and make a note to come back to it if you have time. All unanswered questions are counted the same as an incorrect answer, so it is important that you at least attempt to answer every question.

Take a deep breath.

You can do this.

## Practice exam one

1. If you would like your dashboard to be delivered to the recipients only once on a specific date, what would be the most appropriate approach?
  - A. Access permissions
  - B. Subscription
  - C. Interactive saved searches
  - D. Scheduled delivery
2. A schema with denormalized tables would be what type of schema?
  - A. Fact constellation schema
  - B. Snowflake schema
  - C. Star schema
  - D. Galaxy schema
3. If you were to perform a right join on the following tables, what would be the result?

Left Table		Right Table	
ClientID	Name	ClientID	Name
1	Smith, Laurence	1	Austin, TX
2	Brown, Betty	2	Denver, CO
3	Hook, Phil	3	Tulsa, OK
4	Roark, Jen	7	Phoenix, AZ
5	Cox, Jona	8	Seattle, WA
6	Humbert, Ren	9	Baltimore, MD

A.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
4	Roark, Jen	NULL
5	Cox, Jona	NULL
6	Humbert, Ren	NULL
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

B.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK

C.

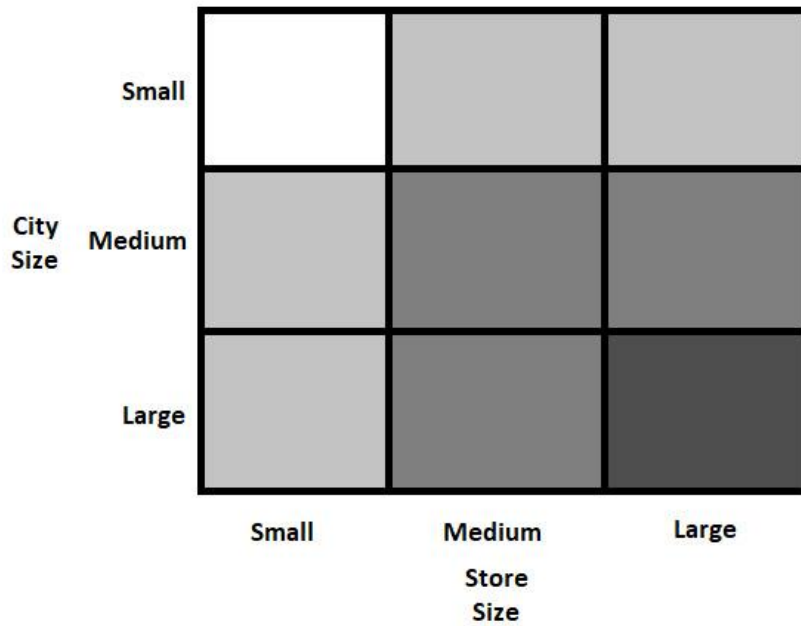
Joined Table		
ClientID	Name	City

1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
4	Roark, Jen	NULL
5	Cox, Jona	NULL
6	Humbert, Ren	NULL

D.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

4. Which of the following represents the count of observations that falls into each category?
  - A. The confidence interval
  - B. The z-score
  - C. The percent difference
  - D. Frequency
5. Forecasting falls into what general type of analysis?
  - A. Performance analysis
  - B. Link analysis
  - C. Exploratory data analysis
  - D. Trend analysis
6. The following is a representation of what type of visualization?



- A. A tree map
  - B. A heat map
  - C. A geographic map
  - D. A stacked bar chart
7. Which of the following analytical tools is considered a programming language?
- A. Dataroma
  - B. Python
  - C. Minitab
  - D. Qlik
8. A person's credit card information is considered what type of protected data?
- A. PII
  - B. PHI
  - C. PCI
  - D. PBI
9. Your manager would like to know whether an employee's height has any relationship to how productive they are before the end of the week, so they can decide whether they should hire taller employees. What type of report is most appropriate in this scenario?
- A. A self-service report

- B. A recurring report
- C. A research report
- D. An ad hoc report

10. Which of the following would contain the true mean?

- A. The confidence interval
- B. The interquartile range
- C. The range
- D. The standard deviation

11. When you are creating a new dashboard, which of the following should be done first?

- A. Deliver the dashboard
- B. Get approval
- C. Create the dashboard
- D. Create a mockup/wireframe

12. You are asked to create a visualization for a sales report. This single visualization looks at revenue from each division in the company. How much revenue came from each department within each division is also a requirement. What visualization would be most appropriate?

- A. A bubble chart
- B. A pie chart
- C. A stacked bar chart
- D. A line chart

13. The following table is in chronological order, with new values added to the bottom. This table is an example of updating a table by what means?

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	X	X	X
12	X	X	X
11	3	5	3
9	2	6	1
10	4	3	3

- A. Adding variables
- B. Removing variables

- C. Deleting historical data
- D. Active record

14. The following dataset is an example of what type of error?

<u>City</u>
los angeles
LA
Los Angeles
Los Angelus
la
los angeles

- A. Data type validation
  - B. Specification mismatch
  - C. Redundant data
  - D. Invalid data
15. If you are trying to explain the complicated relationship between different factors to an audience that is not technical, such as the general public, what type of visualization would be most appropriate?
- A. Heat map
  - B. Histogram
  - C. Infographic
  - D. Bubble chart
16. What does ETL stand for?
- A. Exact, time, line
  - B. Extract, transform, load
  - C. Electronic, transfer, logistics
  - D. Extra, transactional, Lambda
17. The following screenshot represents what type of survey question?



5. Chunky peanut butter is, objectively, better. 🗣️ 0

- ☐ Strongly agree
- ☐ Agree
- ☐ Neither agree nor disagree
- ☐ Disagree
- ☐ Strongly disagree

- A. Likert
- B. Drop-down
- C. Multiple choice
- D. Text-based

18. Which section of a data use agreement includes information on what happens if consent is withdrawn?

- A. The acceptable use policy
- B. Data processing
- C. Data deletion
- D. Data retention

19. If your goal was to have a database that required the fewest number of joins possible, which schema would be most appropriate?

- A. A snowflake schema
- B. A star schema
- C. A snowball schema
- D. A galaxy schema

20. The following dataset is an example of what type of error?

Employee ID	LastName	FirstName	Department	YearsWithCompany
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
64986	Chane	Jill	IT	1
64986	Chane	Jill	IT	1
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sales	21

- A. Duplicate data
- B. Redundant data

C. Missing data

D. Invalid data

21. Which of the following is something to consider when checking for data quality?

A. Data visualization

B. Data transmission

C. Data accuracy

D. Data encryption

22. Which data validation approach should you take if you need to see whether the results of an analysis can be generalized?

A. Data auditing

B. Data profiling

C. Spot checking

D. Cross-validation

23. What type of analysis gives you basic information about the shape and structure of your data before you even start using it?

A. Performance analysis

B. Trend analysis

C. Exploratory data analysis

D. Link analysis

24. A small tech start-up is trying to decide between two designs for the same product: Design A and Design B. They have run several trials on both designs, and they think that Design A is more efficient, but they aren't sure. In this scenario, what is the alternative hypothesis?

A. There is no significant difference between Design A and Design B

B. There is a significant difference between Design A and Design B

C. There is a difference between Design A and Design B, but it is not significant

D. It doesn't matter which design they pick

25. A project manager requests information on KPIs to see whether their team is staying on schedule. What type of analysis is this?

A. Trend analysis

B. Performance analysis

- C. Exploratory data analysis
- D. Link analysis

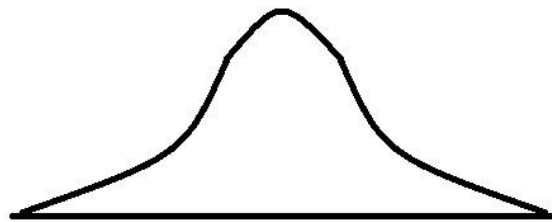
26. Find the variance of the following dataset: 56, 46, 27, 31, 40

- A. 11.6
- B. 40.0
- C. 112.8
- D. 135.5

27. Using the independent variable to predict the dependent variable best describes what analysis?

- A. A t-test
- B. A simple linear regression
- C. A chi-square
- D. A z-score

28. The following represents what type of distribution?



- A. Normal
- B. Uniform
- C. Exponential
- D. Bernoulli

29. Which of the following has a direct impact on how fast and efficient a dashboard is?

- A. Access permissions
- B. Interactive saved searches
- C. Subscription
- D. Scheduled delivery

30. Which of the following file types is used to pass data through websites without having anything to do with the structure of a website?

- A. JSON
- B. HTML
- C. JS
- D. XML

31. Find the mean of the following dataset: 39, 37, 42, 39, 43

- A. 37
- B. 39
- C. 40
- D. 42

32. Which of the following data storage solutions is most appropriate for large amounts of unstructured data?

- A. Data mart
- B. Data lake
- C. Data warehouse
- D. Data puddle

33. If you wanted to compare the batting records of two baseball athletes over the course of a year, which of the following analyses would be most appropriate?

- A. Correlation
- B. Z-score
- C. T-test
- D. Chi-square

34. You are working with the following dataset of data gathered from different robots attempting an obstacle course. The team wants to know whether the weight of the robot had any impact on how far it got in the course. Which variables would be considered appropriate data content for the report?

Date	Time(s)	Vol	Ins	Res	Operator	Distance(m)	RobotWeight(lb)
7/5/2019	33	1.9470	471	40	Oscar Mayor	22	9
7/12/2019	35	1.3958	581	86	Oscar Mayor	25	11
7/19/2019	44	1.3905	304	4	Oscar Mayor	23	10
7/26/2019	30	1.5938	482	15	Oscar Mayor	24	8
8/2/2019	43	1.1938	561	51	Oscar Mayor	19	11
8/9/2019	33	1.8404	309	27	Oscar Mayor	21	9
8/16/2019	35	1.4893	701	83	Oscar Mayor	16	9
8/23/2019	40	1.2843	824	79	Oscar Mayor	15	10
8/30/2019	35	1.9837	754	73	Oscar Mayor	22	11
9/6/2019	43	1.7489	601	48	Oscar Mayor	24	9

- A. Time(s), Operator, and Distance(m)
  - B. Date, Vol, Res, Distance(m), and RobotWeight(lb)
  - C. Date, Distance(m), and RobotWeight(lb)
  - D. Every variable
35. An analysis that compares a sample to the population to see whether it is a good representation best describes what analysis?
- A. Correlation
  - B. T-test
  - C. Simple linear regression
  - D. Chi-square
36. Which data quality dimension means ensuring your data is all reported in the same format?
- A. Accuracy
  - B. Completeness
  - C. Consistency
  - D. Attribute limitation
37. Customer-facing agents in your company have requested a report that gives them the most up-to-date customer information and rates possible. What type of report is most appropriate in this scenario?
- A. A point-in-time report
  - B. A static report
  - C. A dynamic report
  - D. A research report
38. Which of the following would you find in a structured database?
- A. Images

B. Defined rows/columns

C. Video files

D. Social media data

39. If you are updating an old report by rerunning the analyses on updated data, which of the following values will change?

A. The reference data sources

B. The appendix

C. The reference dates

D. The FAQs

40. When choosing fonts for your report, what should be your first consideration?

A. Font size

B. The number of fonts

C. Font types

D. Branding

41. The sales department of your company would like to be able to look up the latest sales data at any point in time so they can make decisions quickly about setting rates for new contracts. What type of report is most appropriate in this scenario?

A. A research report

B. An ad hoc report

C. A self-service report

D. A recurring report

42. Deleting an entire row of data because of one value in it is what type of deletion?

A. Pairwise deletion

B. Listwise deletion

C. Variable deletion

D. Filtering

43. Code or software that tells you information about your environment or file paths is an example of what?

A. System functions

- B. Conditional operators
- C. Recoding
- D. Transposition

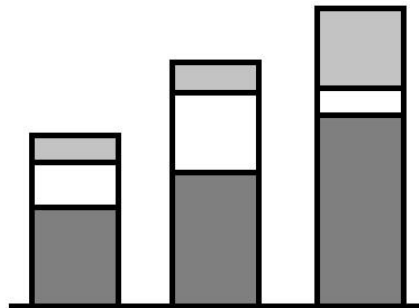
44. `if`, `and`, `or`, and `not` are examples of what?

- A. Dummy coding
- B. Transposition
- C. Reduction
- D. Conditional operators

45. What do you call the process of filling gaps in the data with the average of the values that are present?

- A. Interpretation
- B. Infusion
- C. Imputation
- D. Interpolation

46. The following depicts what kind of visualization?



- A. A waterfall chart
- B. A histogram
- C. A bar chart
- D. A stacked bar chart

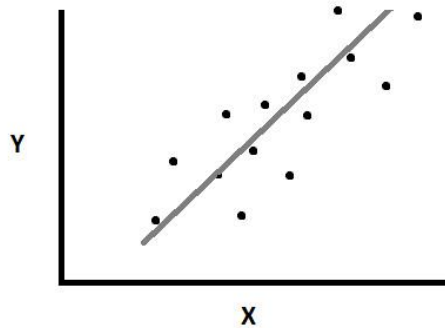
47. Which of the following are circumstances under which you should check the quality of your data?

- A. Data transformation
- B. Data transmission

C. Data encryption

D. Data deletion

48. The following graph depicts what kind of correlation?



A. No correlation

B. Negative correlation

C. Positive correlation

D. Semi-correlation

49. You receive your results for an exam, and the results of the other students are posted anonymously. You want to know how your exam results compare to the normal distribution of your classmates' scores. Which analysis is most appropriate?

A. Chi-square

B. T-test

C. Z-score

D. Simple linear regression

50. The following dataset is an example of what type of data?

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	3	5	3

A. Unstructured

B. Relational

C. Non-relational

D. Semi-structured





### B. Uniform

- C. Non-parametric
- D. Exponential

56. An IP address is considered what type of protected data?

- A. PII
- B. PHI
- C. PCI
- D. PIFI

57. You run a simple AB study to compare two different advertising campaigns. Assuming an alpha of 0.05, which of the following p-values would cause you to reject the null hypothesis?

- A. 0.3
- B. 0.5
- C. 0.06
- D. 0.008

58. If you had a variable called `BirdCount` that kept track of the number of birds that flew by your window on a specific day, what variable type would it be?

- A. Nominal
- B. Continuous
- C. Ordinal
- D. Discrete

59. The act of automatically collecting, processing, and storing online transactions is called what?

- A. OLAP
- B. ETL
- C. ELT
- D. OLTP

60. Breaking large chunks of data down into small, usable pieces is called what?

- A. Interpretation
- B. Parsing
- C. Reduction

D. Interpolation

61. A new variable that specifically holds a calculation of other variables is called what?

A. A derived variable

B. A binary variable

C. A variable deletion

D. An ordinal variable

62. The marketing team for your company has created a sample of the company's customers, but before they run any tests, they want to know whether the sample accurately reflects the larger population of customers. Which of the following analyses is most appropriate?

A. The chi-square test for independence

B. The chi-square test for homogeneity

C. The chi-square goodness of fit

D. The chi-square test for linearity

63. A sales manager believes that there is a connection between the customer's age and how likely they are to purchase the product. What type of analysis would be most appropriate?

A. Trend analysis

B. Performance analysis

C. Link analysis

D. Exploratory data analysis

64. You are asked to create a visualization that tracks ROI over the past 4 years, looking for general trends. What is the most appropriate visualization?

A. A line chart

B. A scatter plot

C. A heat map

D. A bubble chart

65. Where on a dashboard do you put instructions for how to use it?

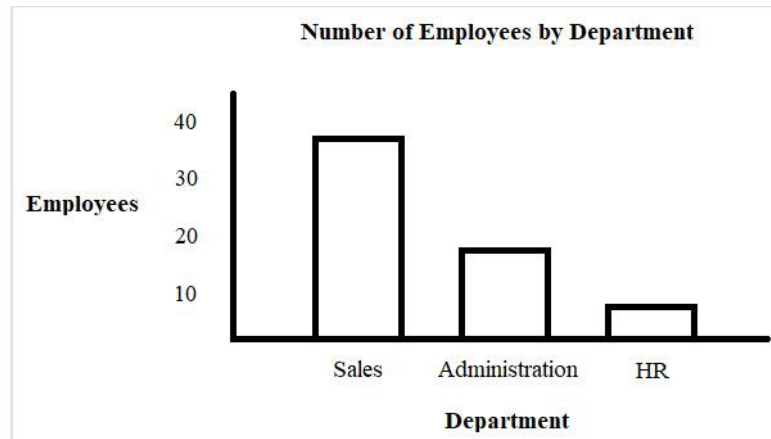
A. The appendix

B. The cover page

C. The FAQs

D. The reference dates

66. What conclusion can you draw from the following visualization?



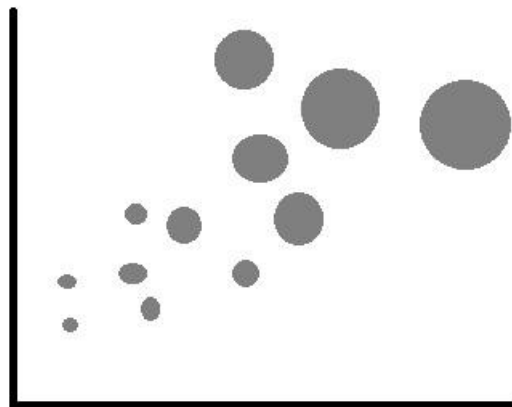
A. This accounts for every employee in the company

B. Administration is the smallest department

C. Administration is the biggest department

D. There are at least 60 employees in the company

67. The following graph represents what kind of visualization?



A. A waterfall chart

B. A scatter plot

C. A pie chart

D. A bubble chart

68. A data point that is so much smaller than every other data point in the dataset that it drastically lowers the average for the entire dataset is an example of what?

A. A specification mismatch

- B. Duplicate data
- C. An outlier
- D. Data type validation

69. Which section of a data use agreement includes information on the consequences of using the data improperly?

- A. The acceptable use policy
- B. Data processing
- C. Data deletion
- D. Data retention

70. The following is an example of what type of join?

Left Table			Joined Table				Right Table	
ClientID	Name		ClientID	Name	City		ClientID	City
1	Smith, Laurence		1	Smith, Laurence	Austin, TX		1	Austin, TX
2	Brown, Betty		2	Brown, Betty	Denver, CO		2	Denver, CO
3	Hook, Phil		3	Hook, Phil	Tulsa, OK		3	Tulsa, OK
4	Roark, Jen						7	Phoenix, AZ
5	Cox, Jona						8	Seattle, WA
6	Humbert, Ren						9	Baltimore, MD

- A. Outer join
- B. Inner Join
- C. Left Join
- D. Right Join

71. Which of the following instances is an ideal time to implement MDM?

- A. When data is manipulated
- B. When data is transferred
- C. When a company is purchased
- D. When data is transformed

72. In the following table, which variable is specifically there to indicate that a row is the most recent value?

Magic Number	Active Record	Active Start	Active End
41	No	11/11/2011	12/12/2012
42	Yes	12/12/2012	

- A. Magic Number

- B. Active Record
- C. Active Start
- D. Active End

73. A large data storage solution for relational data, focusing on efficiency and following a snowflake schema, would most likely be what?

- A. A data warehouse
- B. A data mart
- C. A data lake
- D. A data puddle

74. The following dataset is an example of what concept?

Month	UnitsSold	Color	Red	Blue	Yellow
August	432	Red	1	0	0
August	365	Blue	0	1	0
August	154	Yellow	0	0	1
September	398	Red	1	0	0
September	386	Blue	0	1	0
September	108	Yellow	0	0	1

- A. Recoding a number into a category
- B. Recoding a category into a number
- C. Dummy coding
- D. Transposition

75. What is the range of the following dataset:

25, 38, 50, 49, 38

- A. 25
- B. 38
- C. 40
- D. 50

76. The following diagram represents what type of database schema?



- A. A snowflake schema
- B. A star schema
- C. A galaxy schema
- D. A fast constellation schema

77. Your company has set up a database that can only be accessed by people with the job title data analyst. What type of access requirement is this?

- A. Data encryption
- B. User group-based
- C. Role-based
- D. Data transmission

78. A pharmaceutical lab wants to know whether a memory-enhancing drug works. They have two mice run the same maze multiple times, one with the medication and one without. You analyze the results and receive a p-value of 0.4. Assuming an alpha of 0.05, how do you interpret the results?

- A. Accept the alternative hypothesis and reject the null hypothesis
- B. Reject the alternative hypothesis and accept the null hypothesis
- C. Accept the alternative and null hypotheses
- D. Reject the alternative and null hypotheses

79. Which of the following is a prewritten query that allows the user to only enter very specific information to target data?

- A. Index
- B. Parameterization
- C. Filter
- D. Sort

80. What do you call the process of manually or automatically checking the data type of a variable to avoid errors?

- A. Data type validation
- B. Imputation
- C. Specification mismatch
- D. Interpolation

81. A piece of code that requests information from an API and then waits for a response before continuing has what sort of connection?

- A. Structured
- B. Unstructured
- C. Synchronous
- D. Asynchronous

82. The following tables have what sort of cardinality?

Employee Information		Sales Information	
EmployeeID	EmployeeName	EmployeeID	SalesID
8279	Steve	8279	826705
7904	Joan	8279	379001
2905	Frank	2905	758982
8391	Bertha	8391	178904
		7904	389057
		8279	890471
		8391	689409
		8391	789039
		8391	689315
		8279	492084



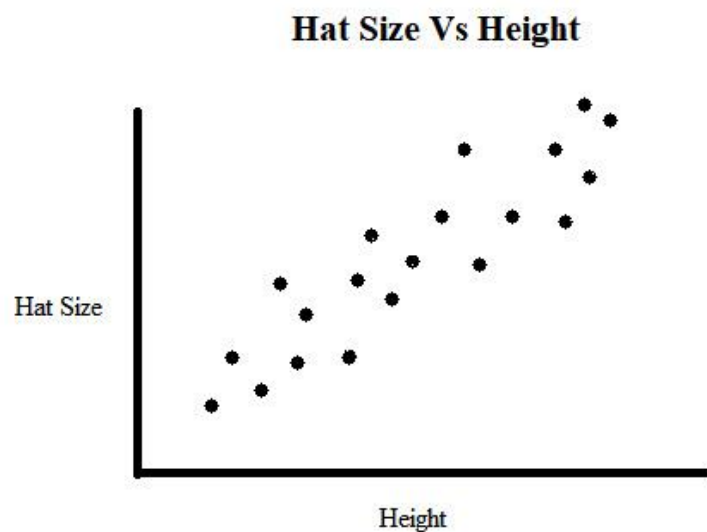
Employee Information		Sales Information	
EmployeeID	EmployeeName	EmployeeID	SalesID
		2905	291046
		2905	164829

- A. One-to-one
- B. One-to-many
- C. Many-to-many
- D. There is no entity relationship

83. Which data validation approach should you take to see whether a dataset is appropriate for a specific goal?

- A. Reasonable expectations
- B. Data audits
- C. Data profiling
- D. Cross-validation

84. What conclusion can you draw from the following visualization?



- A. Being taller makes your head bigger
- B. Having a small head makes you taller
- C. The taller you are, the more likely you are to have a larger hat size
- D. The taller you are, the more likely you are to have a smaller hat size

85. A WAV file holds what type of information?

- A. Image
- B. Video
- C. Audio
- D. Text

86. The process of using code to automatically collect data from websites for you is called what?

- A. Web scraping
- B. Web services
- C. Non-relational
- D. Semi-structured

87. The following dataset is an example of what type of data?

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	3	5	3


- A. Structured
- B. Relational
- C. Both structured and relational
- D. Neither structured nor relational

88. Which of the following is a key process of MDM?

- A. Data dictionary
- B. Data encryption
- C. Data transformation
- D. Data manipulation

89. The following depicts an action that has been performed on a dataset. What is the name of that action?

Index	Employee ID	LastName	FirstName	Department	YearsWithCompany
1	83784	Benhill	Floyd	Sales	12
2	64986	Chane	Jill	IT	1
3	93671	Hanson	Richard	HR	15
4	37816	Smith	Trudy	Sales	21



Index	Employee ID	LastName	FirstName	Department	YearsWithCompany
4	37816	Smith	Trudy	Sales	21
3	93671	Hanson	Richard	HR	15
1	83784	Benhill	Floyd	Sales	12
2	64986	Chane	Jill	IT	1

- A. Filtering
- B. Subsets
- C. Parameterization
- D. Sorting

90. A software company recently released an update to make their program faster, and they want to know whether it worked. You compare several trials using the software before and after the patch. Finally, you decide that the update did make the software faster. Unfortunately, you are wrong. What type of error is this?

- A. Type I
- B. Type II
- C. Type III
- D. Type IV

## Congratulations!

You made it through your first practice exam! Take a moment to relax and celebrate. Get a cookie; you deserve it. Make sure to hydrate and take care of anything you couldn't do during the exam. Once you are set, compare your answers to the exam answers listed in the following section.

CompTIA says that a pass score is 675 on a scale of 100 to 900, but they do not release any information on grading or how they weight the different questions. Also, if they are testing a new question and everyone gets it wrong, it probably won't be counted against you. Overall, this makes grading a practice exam a pain. As a rough guideline, try to miss 20 or fewer questions. That said, what is important is that you note what topics you are struggling with and make sure you review them. It's okay if you miss more than 20 on your first try, but make sure you review those topics and understand them before you try again.

## Practice exam one answers

1. The answer is: Scheduled delivery

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Delivery](#)

2. The answer is: Star schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through the Data Schema and Its Types](#)

3. The answer is:

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Merging Data](#)

4. The answer is: Frequency

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Understanding Frequencies and Percentages](#)

5. The answer is: Trend analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Discovering Trends](#)

6. The answer is: A heat map

If you are unsure why, please review [Chapter 13, Common Visualizations – Understanding Heat Maps, Tree Maps, and Geographic Maps](#)

7. The answer is: Python

If you are unsure why, please review [Chapter 11, Types of Reports – Knowing Important Analytical Tools](#)

8. The answer is: PCI

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Classifications](#)

9. The answer is: An ad hoc report

If you are unsure why, please review [Chapter 11, Types of Reports – Understanding Ad hoc and Research Reports](#)

10. The answer is: The confidence interval

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Discovering Confidence Intervals](#)

11. The answer is: Create a mockup/wireframe

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding the Report Development Process](#)

12. The answer is: A stacked bar chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

13. The answer is: Adding variables

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Updating Stored Data](#)

14. The answer is: Invalid data

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Invalid Data, Specification Mismatch, and Data Type Validation](#)

15. The answer is: Infographic

If you are unsure why, please review [Chapter 13, Common Visualizations – Understanding Infographics and Word Clouds](#)

16. The answer is: Extract, transform, load

If you are unsure why, please review [Chapter 3, Collecting Data – Differentiating ETL and ELT](#)

17. The answer is: Likert

If you are unsure why, please review [Chapter 3, Collecting Data – Collecting Your Own Data](#)

18. The answer is: Data deletion

If you are unsure why, please review [Chapter 14, Data Governance – Knowing Use Requirements](#)

19. The answer is: A star schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through the Data Schema and its Types](#)

20. The answer is: Duplicate data

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Managing Duplicate and Redundant Data](#)

21. The answer is: Data accuracy

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

22. The answer is: Cross-validation

If you are unsure why, please review [Chapter 15, Data Quality and Management – Validating Quality](#)

23. The answer is: Exploratory data analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Exploring Your Data](#)

24. The answer is: There is a significant difference between Design A and Design B

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Differentiating Null Hypothesis and Alternative Hypothesis](#)

25. The answer is: Performance analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Checking on Performance](#)

26. The answer is: 135.5

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Finding Variance and Standard Deviation](#)

27. The answer is: A simple linear regression

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Simple Linear Regression](#)

28. The answer is: Normal

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Discovering Distributions](#)

29. The answer is: Interactive saved searches

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Delivery](#)

30. The answer is: JSON

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through Data Types and File Types](#)

31. The answer is: 40

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Understanding Measures of Central Tendency](#)

32. The answer is: Data lake

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding the Concept of Warehouses and Lakes](#)

33. The answer is: T-test

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Understanding t-Tests](#)

34. The answer is: Date, Distance(m), and RobotWeight(lb)

If you are unsure why, please review [Chapter 12, Reporting Process – Knowing what to Consider when Making a Report](#)

35. The answer is: Chi-square

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Knowing Chi-Square](#)

36. The answer is: Consistency

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

37. The answer is: A dynamic report

If you are unsure why, please review [Chapter 11, Types of Reports – Distinguishing Static and Dynamic Reports](#)

38. The answer is: Defined rows/columns

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

39. The answer is: The reference dates

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Elements](#)

40. The answer is: Branding

If you are unsure why, please review [Chapter 12, Reporting Process – Designing Reports](#)

41. The answer is: A self-service report

If you are unsure why, please review [Chapter 11, Types of Reports – Knowing about Self-Service Reports](#)

42. The answer is: Listwise deletion

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Dealing with Missing Data](#)

43. The answer is: System functions

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Shaping Data with Common Functions](#)

44. The answer is: Conditional operators

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Shaping Data with Common Functions](#)

45. The answer is: Imputation

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Dealing with Missing Data](#)

46. The answer is: A stacked bar chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

47. The answer is: Data transformation

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

48. The answer is: Positive correlation

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Calculating Correlations](#)

49. The answer is: Z-score

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Understanding z-Scores](#)

50. The answer is: Relational



If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

51. The answer is: Views

If you are unsure why, please review [Chapter 12, Reporting Process – Knowing what to Consider when Making a Report](#)

52. The answer is: 47%

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Calculating Percent Change and Percent Difference](#)

53. The answer is: A recurring report

If you are unsure why, please review [Chapter 11, Types of Reports – Understanding Recurring Reports](#)

54. The answer is: The same information recorded in multiple columns

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Managing Duplicate and Redundant Data](#)

55. The answer is: Non-parametric

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Non-parametric Data](#)

56. The answer is: PII

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Classifications](#)

57. The answer is: 0.008

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Learning p-Value and Alpha](#)

58. The answer is: Discrete

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through Data Types and File Types](#)

59. The answer is: OLTP

If you are unsure why, please review [Chapter 3, Collecting Data – Understanding OLTP and OLAP](#)

60. The answer is: Parsing

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Parsing Your Data](#)

61. The answer is: A derived variable

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Calculating Derived and Reduced Variables](#)

62. The answer is: The chi-square goodness of fit

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Knowing Chi-Square](#)

63. The answer is: Link analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Finding Links](#)

64. The answer is: A line chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

65. The answer is: The cover page

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Elements](#)

66. The answer is: There are at least 60 employees in the company

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

67. The answer is: A bubble chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

68. The answer is: An outlier

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Finding Outliers](#)

69. The answer is: The acceptable use policy

If you are unsure why, please review [Chapter 14, Data Governance – Knowing Use Requirements](#)

70. The answer is: Inner join

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Merging Data](#)

71. The answer is: When a company is purchased

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Master Data Management \(MDM\)](#)

72. The answer is: Active Record

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Updating Stored Data](#)

73. The answer is: A data warehouse

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding the Concept of Warehouses and Lakes](#)

74. The answer is: Dummy coding

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Recoding Variables](#)

75. The answer is: 25

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Calculating Range and Quartiles](#)

76. The answer is: A star schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through the Data Schema and its Types](#)

77. The answer is: Role-based

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Security](#)

78. The answer is: Reject the alternative hypothesis and accept the null hypothesis

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Learning p-Value and Alpha](#)

79. The answer is: Parameterization

If you are unsure why, please review [Chapter 3, Collecting Data – Optimizing Query Structure](#)

80. The answer is: Data type validation

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Invalid Data, Specification Mismatch, and Data Type Validation](#)

81. The answer is: Synchronous

If you are unsure why, please review [Chapter 3, Collecting Data – Utilizing Public Sources of Data](#)

82. The answer is: One-to-many

If you are unsure why, please review [Chapter 14, Data Governance – Handling Entity Relationship Requirements](#)

83. The answer is: Data audits

If you are unsure why, please review [Chapter 15, Data Quality and Management – Validating Quality](#)

84. The answer is: The taller you are the more likely you are to have a larger hat size

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

85. The answer is: Audio

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going through Data Types and File Types](#)

86. The answer is: Web scraping

If you are unsure why, please review [Chapter 3, Collecting Data – Collecting Your Own Data](#)

87. The answer is: Both structured and relational

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

88. The answer is: Data dictionary

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Master Data Management \(MDM\)](#)

89. The answer is: Sorting

If you are unsure why, please review [Chapter 3, Collecting Data – Optimizing Query Structure](#)

90. The answer is: Type I

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Understanding Type I and Type II Error](#)

# Chapter 17: Practice Exam Two

If you are here, you have already taken the first practice exam and reviewed the content you missed. You know the drill!

To make this exam the best practice it can be, you should follow the testing experience as closely as possible. Make sure you are in a quiet and clean environment and have access to a simple calculator, a timer, and the ability to take notes.

Set your timer for 90 minutes.

Don't panic.

Take a deep breath.

You can do this.

## Practice exam two

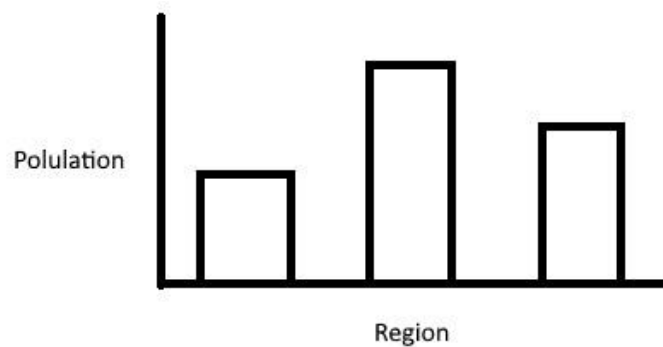
1. EDA stands for what?
  - A. Extra data analysis
  - B. Extrapolate data architecture
  - C. Extract data analyze
  - D. Exploratory data analysis
2. You run a study to see whether there is a difference in the amount of weight a hamster can lift compared to a gerbil. You find out the results and come to a conclusion. If the result was a type II error, what was your conclusion?
  - A. There is a difference between hamster and gerbil lifting capacities
  - B. Gerbils and hamsters can lift the same amount
  - C. Hamsters can lift more than gerbils
  - D. Gerbils can lift more than hamsters
3. The following dataset is an example of what type of error?

ID	Sex	Male	Female
84927	M	TRUE	FALSE
69427	M	TRUE	FALSE
10374	F	FALSE	TRUE
58264	M	TRUE	FALSE
90162	F	FALSE	TRUE

- A. Missing data

- B. Duplicate data
  - C. Invalid data
  - D. Redundant data
4. Which of the following represents the percent of observations in each category as compared to the whole?
- A. Percentage
  - B. Percent change
  - C. Percent difference
  - D. Frequency
5. A pharmaceutical lab wants to know whether a memory-enhancing drug works. They have two mice run the same maze multiple times, one with the medication and one without. You analyze the results and receive a p-value of 0.04. Assuming an alpha of 0.05, how do you interpret the results?
- A. Accept the alternative hypothesis and reject the null hypothesis
  - B. Reject the alternative hypothesis and accept the null hypothesis
  - C. Accept the alternative and null hypotheses
  - D. Reject the alternative and null hypotheses
6. A company that produces cleaning supplies has a new product. To see whether it is effective, they run a trial with two groups. The first group uses the new product to clean a stain and the second group uses tap water to clean a stain. The idea that there will be no difference between the performance of these two groups is what kind of hypothesis?
- A. Alternative hypothesis
  - B. Null hypothesis
  - C. Secondary hypothesis
  - D. Original hypothesis
7. A small ferret farm would like to see whether there is a relationship between the weight of a ferret in grams and how much milk it produces in milliliters. Which of the following visualizations would be most appropriate?
- A. A line chart
  - B. A bubble chart
  - C. A scatter plot

- D. A tree map
8. A flat file delimited by commas is what file type?
- A. JPEG
  - B. TSV
  - C. AAC
  - D. CSV
9. Which of the following elements should never be on the cover page of a report?
- A. The version number
  - B. The report run data
  - C. The data refresh date
  - D. The appendix
10. Data type validation is a process specifically used to avoid what type of error?
- A. A specification mismatch
  - B. Invalid data
  - C. Missing data
  - D. Duplicate data
11. What is an appropriate title for the following chart?



- A. The Population of India based on the Geographical Region
- B. Population
- C. The Population of India Averaged for the Years 2015, 2016, and 2017 as Sub-Divided by Geographical Regions Determined by the 2018 Land Survey
- D. Region 2010

12. A mortgage company would like its sales representatives to have access to a dashboard with the absolute most up-to-date rates and figures. This means that the dashboard should be what?

- A. Heavily filtered
- B. Subscription-based
- C. Real-time
- D. Point-in-time

13. The act of automatically moving and analyzing online transactions is called what?

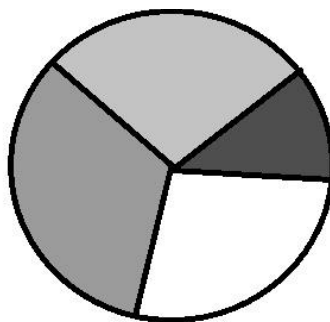
- A. OLTP
- B. OLAP
- C. ELT
- D. ETL

14. The following code snippet is an example of what concept?

```
Data = "This book makes me happy."  
Data = ["This", "book", "makes", "me", "happy", "."]
```

- A. Reduction
- B. Interpretation
- C. Imputation
- D. Parsing

15. The following chart represents what kind of visualization?



- A. A pie chart
- B. A scatter plot
- C. A histogram
- D. A heat map



16. Which of the following is a valid data storage solution for audio files?
- A. A data mart
  - B. A data lake
  - C. A data warehouse
  - D. A data store
17. An engineer would like to check how efficient each phase of a production process is to see whether it can be improved. What type of analysis is most appropriate?
- A. Performance analysis
  - B. Link analysis
  - C. Trend analysis
  - D. Exploratory data analysis
18. You generate a highly detailed report on the number of eggs every single chicken a company owns produces compared to how much that chicken has eaten to get a specific grain-to-egg efficiency ratio for each animal. Who is the most appropriate audience for this report?
- A. C-level executives
  - B. Stakeholders
  - C. The general public
  - D. Technical experts
19. If you were to perform an outer join on the following tables, what would be the result?

Left Table		Right Table	
ClientID	Name	ClientID	Name
1	Smith, Laurence	1	Austin, TX
2	Brown, Betty	2	Denver, CO
3	Hook, Phil	3	Tulsa, OK
4	Roark, Jen	7	Phoenix, AZ
5	Cox, Jona	8	Seattle, WA
6	Humbert, Ren	9	Baltimore, MD

A.

Joined Table
--------------

ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
4	Roark, Jen	NULL
5	Cox, Jona	NULL
6	Humbert, Ren	NULL
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

B.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK

C.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
4	Roark, Jen	NULL
5	Cox, Jona	NULL
6	Humbert, Ren	NULL

D.

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX

2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

20. A project manager would like an operational report at the end of every sprint. What type of report would be most appropriate?

- A. A research report
- B. An ad hoc report
- C. A self-service report
- D. A recurring report

21. Find the mode of the following dataset: 5, 3, 8, 5, 3, 9, 3, 8, 2

- A. 2
- B. 3
- C. 5
- D. 9

22. A project manager wants to know whether there is a connection between how many hours their team works in a day and how many mistakes they make. What type of analysis is most appropriate?

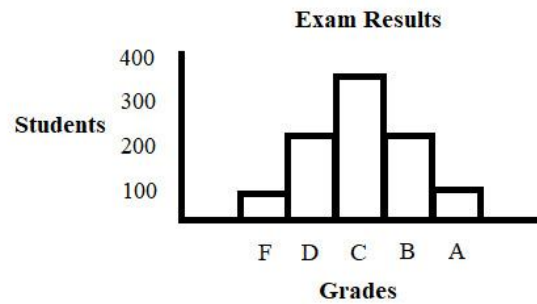
- A. Link analysis
- B. Trend analysis
- C. Performance analysis
- D. Exploratory data analysis

23. The following table is in chronological order, with new values added to the bottom. This table is an example of updating a table by what means?

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	X	X	X
9	X	X	X
10	X	X	X

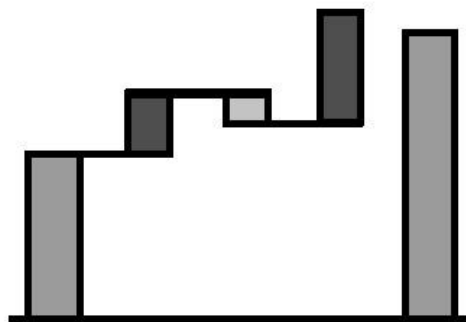
- A. Adding variables
- B. Removing variables
- C. Deleting historical data
- D. Active record

24. What conclusion can you draw from the following visualization?



- A. The majority of students failed the exam
- B. The distribution is nonparametric
- C. Around 350 students achieved a grade of C or higher
- D. A new student taking the test would most likely get a C

25. The following chart depicts what kind of visualization?



- A. A waterfall chart
- B. A histogram
- C. A stacked bar chart
- D. A line chart

26. A schema with only normalized tables would be what type of schema?

- A. A galaxy schema
- B. A star schema

- C. A fast constellation schema
- D. A snowflake schema

27. The following dataset is an example of what concept?

Month	UnitsSold	Color	ColorRecoded
August	432	Red	1
August	365	Blue	2
August	154	Yellow	3
September	398	Red	1
September	386	Blue	2
September	108	Yellow	3

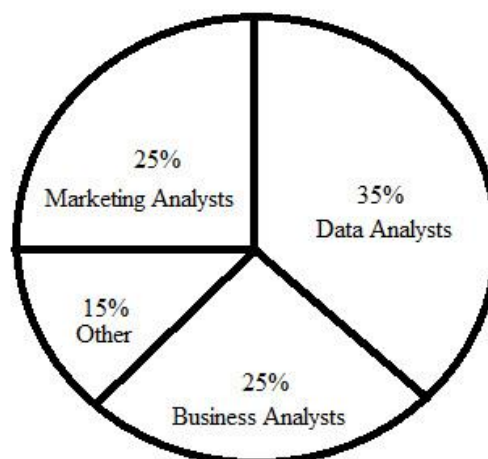
- A. Recoding a category into a number
- B. Recoding a number into a category
- C. Transposition
- D. Dummy coding

28. A detailed program that explains how the software performs a specific query is called what?

- A. An execution plan
- B. A subquery
- C. Parameterization
- D. A temporary table

29. What conclusion can you draw from the following visualization?

**Job Titles Who Have Access to Company Data**



- A. If you were to pick a person who can access data at random, they would most likely be a business analyst

- B. You need to have “analyst” in your title to access data
  - C. The other group is not as technically skilled
  - D. Half of everyone who can access data is either a marketing analyst or a business analyst
30. A small, highly specialized data storage solution following a star schema would most likely be what?
- A. A data warehouse
  - B. A data lake
  - C. A data pond
  - D. A data mart
31. The CEO of your company is considering a merger with your company’s main competitor. They would like a detailed report on the pros and cons, as well as projections on all of the major KPIs for the next 10 years. This report is due in 6 months. What type of report is most appropriate in this scenario?
- A. A self-service report
  - B. A research report
  - C. A recurring report
  - D. An ad hoc report
32. Which type of schema has two levels of dimension tables?
- A. A star schema
  - B. A snowball schema
  - C. A snowflake schema
  - D. A galaxy schema

33. In the following table, which variable indicates when a variable stopped being active?

Magic Number	Active Record	Active Start	Active End
41	No	11/11/2011	12/12/2012
42	Yes	12/12/2012	

- A. Magic Number
- B. Active Record
- C. Active Start
- D. Active End

34. Which of the following is a key process of MDM?

- A. Data manipulation
- B. Data encryption
- C. Data transformation
- D. Data consolidation

35. The following table is a sample from a larger dataset. What type of visualization would be most appropriate to display this information?

Population of Europe by Country	
Country	Population (million)
Russia	145
Germany	84
United Kingdom	68
France	65
Italy	60
Spain	47
...	...

- A. A waterfall chart
- B. A geographic map
- C. A pie chart
- D. A heat map

36. The following dataset is an example of what type of error?

<u>Cost Per Click</u>
\$1.82
\$0.95
Toast
\$2.10
\$1.39

- A. Invalid data
- B. Specification mismatch
- C. Redundant data
- D. Data type validation

37. The following screenshot represents what type of survey question?

2. What is your favorite kind of peanut butter?  0

☐ Smooth

☐ Chunky

- A. Text-based
- B. Single choice
- C. Multiple choice
- D. Drop-down

38. Which of the following is a conditional operator?

- A. ORDER BY
- B. APPEND
- C. SUM
- D. OR

39. Which of the following is something to consider when checking for data quality?

- A. Data visualization
- B. Data transmission
- C. Data manipulation
- D. Data integrity

40. What data-validating approach should you take if you see the results of an analysis and you believe them to be in error?

- A. Data audits
- B. Data profiling
- C. Reasonable expectations
- D. Cross-validation

41. Find the standard deviation of the following dataset: 62, 92, 43, 66, 37

- A. 21.7
- B. 34.6
- C. 60.0



D. 470.5

42. If you are using natural language processing to analyze a text file and would like to visualize a way to express the ideas held within the text, what is the most appropriate visualization?

- A. A word cloud
- B. A tree map
- C. A waterfall chart
- D. A line chart

43. When you suspect that private data might have been breached, what is the single most important thing you can do?

- A. Notify the impacted parties
- B. Fix the breach
- C. Ignore the breach
- D. Figure out how the breach happened

44. The following dataset is an example of what type of data?

Total Number of Beans	Red Beans	Blue Beans	Yellow Beans
10	3	4	3
12	2	6	4
11	3	5	3

- A. Structured
- B. Unstructured
- C. Semi-structured
- D. Non-relational

45. When you are creating a dashboard, what should you do immediately after planning out your data story?

- A. Deliver the dashboard
- B. Create the dashboard
- C. Get approval
- D. Plan a second data story

46. You run a simple A/B study to compare two different advertising campaigns. Assuming an alpha of 0.1, which of the following p-values would cause you to accept the null hypothesis?

- A. 0.05
- B. 0.09
- C. 0.3
- D. 0.001

47. The marketing department has created a customer persona, and they would like to compare the age of their persona against the normally distributed ages of their actual customers. Which analysis is most appropriate?

- A. Chi-square
- B. Simple linear regression
- C. Z-score
- D. T-test

48. The following tables have what sort of cardinality?

Employee Information		Employee Departments	
EmployeeID	EmployeeName	EmployeeID	Department
8279	Steve	8279	Sales
7904	Joan	7904	Human Resources
2905	Frank	2905	Marketing
8391	Bertha	8391	Administration

- A. One-to-one
- B. One-to-many
- C. Many-to-many
- D. There is no entity relationship

49. If you had a variable called **BirdPassed** that kept track of whether or not a bird passed by your window on a given day with **YES** or **NO**, what variable type would it be?

- A. Ordinal
- B. Binary
- C. Discrete
- D. Continuous

50. You calculate that the average number of clicks per minute for the year for a small e-commerce website is 12.8. Another website that is a major competitor has an average

number of clicks per minute for the year of 13.7. What is the difference between your value and your competitors?

- A. 7%
- B. 9%
- C. 12%
- D. 18%

51. Given a t-value of 1.86, which is the confidence interval for the following dataset?

8, 7, 8, 8, 10, 6, 8, 8, 9, 8

- A. 7.3 to 8.7
- B. 6.9 to 9.1
- C. 8.4 to 8.8
- D. 7.6 to 9.5

52. The following is an example of what type of join?

Left Table			Joined Table				Right Table	
ClientID	Name		ClientID	Name	City		ClientID	City
1	Smith, Laurence		1	Smith, Laurence	Austin, TX		1	Austin, TX
2	Brown, Betty		2	Brown, Betty	Denver, CO		2	Denver, CO
3	Hook, Phil		3	Hook, Phil	Tulsa, OK		3	Tulsa, OK
4	Roark, Jen		4	Roark, Jen	NULL		7	Phoenix, AZ
5	Cox, Jona		5	Cox, Jona	NULL		8	Seattle, WA
6	Humbert, Ren		6	Humbert, Ren	NULL		9	Baltimore, MD

- A. Outer join
- B. Inner Join
- C. Left Join
- D. Right Join

53. An analysis that specifically tells you whether or not two categorical variables are related best describes what analysis?

- A. Z-score
- B. Chi-square
- C. T-test
- D. Simple linear regression

54. A social media ID is considered what type of protected data?

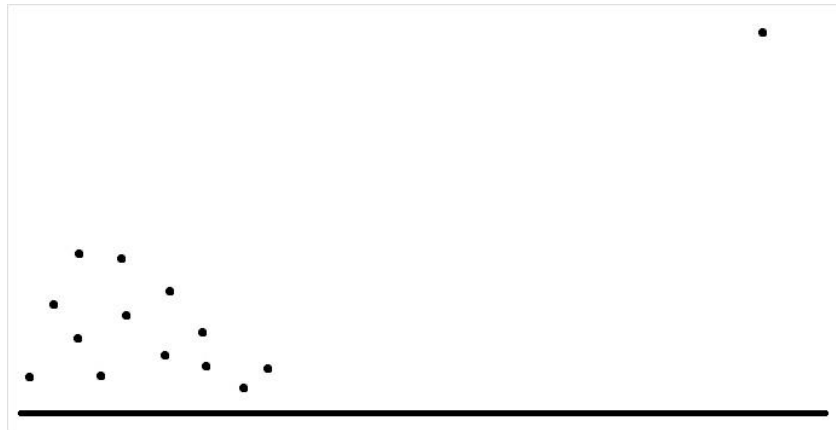
- A. PII

- B. PHI
- C. PCI
- D. PIFI

55. What is duplicate data?

- A. Data that does not meet formatting requirements
- B. The same information recorded in multiple columns
- C. Data that is incomplete or blank
- D. The same information recorded in multiple rows

56. The following is an example of what type of error?



- A. Specification mismatch
- B. An outlier
- C. Redundant data
- D. Missing data

57. Which of the following is considered a public source of data?

- A. Surveying
- B. Web services
- C. Web scraping
- D. Studies

58. Average, sum, and count are all examples of what?

- A. Augmentation
- B. Conditional operators

C. Reduction

D. Parsing

59. Which of the following are circumstances under which you should check the quality of your data?

A. Data encryption

B. Data transmission

C. Data acquisition

D. Data deletion

60. What data-validating approach should you take if you need a formal process to apply to an entire database?

A. Data audits

B. Data profiling

C. Spot checking

D. Cross-validation

61. You are given a dataset that includes tree width and how much weight it can hold before it starts to bend. Then, you are asked to predict how wide a tree must be before it can hold a 400 lb sumo wrestler without bending. Which analysis is most appropriate?

A. Simple linear regression

B. T-test

C. Chi-square test for independence

D. Chi-square goodness of fit

62. The following dataset is an example of what concept?

Month	UnitsSold	Color
August	432	Red
August	365	Blue
August	154	Yellow
September	398	Red
September	386	Blue
September	108	Yellow



Month	August	August	August	September	September	September
UnitsSold	432	365	154	398	386	108
Color	Red	Blue	Yellow	Red	Blue	Yellow

A. Dummy coding

B. Transposition

C. Reduction

D. Conditional operators

63. Making sure your data is not full of gaps and missing data is considered what data quality dimension?

- A. Accuracy
- B. Completeness
- C. Consistency
- D. Attribute limitation

64. The following chart represents what type of distribution?



- A. Normal
- B. Uniform
- C. Exponential
- D. Poisson

65. A person's medical record is considered what type of protected data?

- A. PII
- B. PHI
- C. PCI
- D. PIFI

66. In general, dashboards are considered what type of report?

- A. A self-service report
- B. A risk and regulatory report
- C. An ad hoc report
- D. A research report

67. Your manager gives you access to historical data for a certain variable and would like you to predict what might happen with that variable in the future. What type of analysis is most appropriate?

- A. Link analysis
- B. Trend analysis
- C. Performance analysis

D. Exploratory data analysis

68. Which of the following analytical tools is specialized for visualizations?

A. AWS QuickSight

B. SQL

C. Apex Systems

D. Stata

69. A line manager in a production plant would like to know the specific efficiencies of every machine to see which need tuning at the end of the week. What is the most appropriate data range for this report?

A. Weeks

B. Months

C. Years

D. Decades

70. Deleting only the missing values and only as they are needed is what type of deletion?

A. Pairwise deletion

B. Listwise deletion

C. Variable deletion

D. Filtering

71. Your manager has requested a one-time report to answer a specific business question. What type of report is most appropriate in this scenario?

A. A static report

B. A recurring report

C. A dashboard

D. A dynamic report

72. After publishing a dashboard, you continue to receive emails asking the same questions over and over again. What part of the dashboard should you update to save time?

A. The cover page

B. The FAQs

C. The appendix

D. The reference data sources

73. Unstructured databases include which of the following data types?

A. Undefined fields

B. Machine data

C. Undefined fields and machine data

D. Neither undefined fields nor machine data

74. Which of the following file types can be used to structure a website or pass data through a website?

A. WMA

B. XML

C. AVI

D. WMV

75. Find the middle quartile (Q2) of the following dataset:

70, 21, 34, 48, 27

A. 21

B. 34

C. 40

D. 70

76. Watching things and taking notes as a form of data collection is called what?

A. API

B. Web scraping


C. Survey

D. Observation

77. The following dataset depicts an action that has been performed on a dataset. What is the name of that action?



Employee ID	LastName	FirstName	Department	YearsWithCompany
83784	Benhill	Floyd	Sales	12
64986	Chane	Jill	IT	1
93671	Hanson	Richard	HR	15
37816	Smith	Trudy	Sales	21

Employee ID	LastName	FirstName	Department	YearsWithCompany
83784	Benhill	Floyd	Sales	12
37816	Smith	Trudy	Sales	21

- A. Filtering
- B. Indexing
- C. Sorting
- D. Execution planning

78. What do you call the process of filling gaps in the data by calculating the most likely value based on the values of other variables in the row?

- A. Interpolation
- B. Interpretation
- C. Imputation
- D. Infusion

79. How do nonparametric distributions relate to normal distributions?

- A. Nonparametric distributions are never normal
- B. Nonparametric distributions are sometimes normal
- C. Normal distributions are sometimes parametric
- D. Normal distributions are always nonparametric

80. The HR department at your company wants to know whether or not there is a relationship between an employee's job title and the color of their hair. You don't know why they think this is important, but what type of analysis would be most appropriate here?

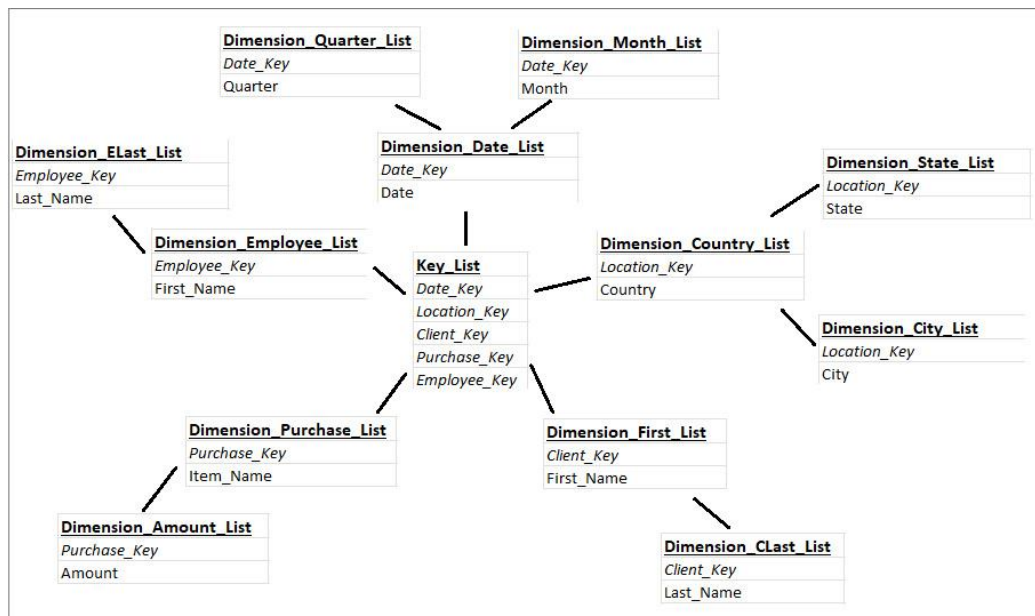
- A. Chi-square test for independence
- B. Chi-square test for homogeneity
- C. Chi-square goodness of fit

D. Chi-square test for linearity

81. A delta load happens when you do what?

- A. Load information into a new location for the first time
- B. Upload all information
- C. Only load information that is new or has changed
- D. Only load information that hasn't changed

82. The following diagram represents what type of database schema?



- A. A snowflake schema
- B. A star schema
- C. A galaxy schema
- D. A fast constellation schema

83. A company fundraiser would like to know how many shirts were sold of each shirt size so that they can create a distribution and order the appropriate numbers for the next one. What type of visualization would be most appropriate?

- A. A histogram
- B. A waterfall chart
- C. A stacked bar chart
- D. A heat map

84. Which of the following would you find in a structured database?

- A. Machine data
- B. Text files
- C. Emails
- D. Key-value pairs

85. An analysis that compares quantitative variables to see whether there is a relationship between them and how strong that relationship is best describes what analysis?

- A. Z-score
- B. Chi-square
- C. Correlation
- D. T-test

86. You are required by law to make certain data illegible during transit by translating the data from plaintext to cyphertext. This data cannot be accessed without a specific key. What security process is described here?

- A. Release approval
- B. Data masking
- C. Access requirements
- D. Data encryption

87. An analysis that compares two groups of quantitative variables to tell whether or not there is a significant difference between them most accurately describes what type of analysis?

- A. Correlation
- B. Z-score
- C. T-test
- D. Chi-square

88. What is a major benefit of MDM?

- A. Streamlining data access
- B. Creating a diversified database
- C. Requiring more joins
- D. Safely storing the data in multiple places

89. Your boss would like you to create a dashboard that automatically refreshes and sends out invitations to the appropriate parties every Monday morning. What is the most suitable approach?
- A. Access permissions
  - B. Subscription
  - C. Interactive saved searches
  - D. Scheduled delivery
90. Which section of the data use agreement includes information on when and how data will be destroyed?
- A. The acceptable use policy
  - B. Data processing
  - C. Data deletion
  - D. Data retention

## Congratulations!

You made it through your second practice exam! That wasn't so bad, right? Take a moment to celebrate how far you've come. When you are ready, compare your answers to the answers listed here. Again, your goal is to get fewer than 20 questions wrong as a rough guideline. Make sure to review the topics of the questions you missed.

When you are ready, you can go to the following website, where you can schedule your exam or find additional resources if you are still struggling: <https://www.comptia.org/certifications/data>.

This wraps up every topic on the exam, as well as over 200 practice questions! You are ready.

Thank you for reading this book.

Best of luck with your exam!

Take a deep breath.

You can do it.

## Practice exam two answers

1. The answer is: Exploratory data analysis

If you are unsure why, please review **Chapter 6, Types of Analytics – Exploring Your Data**

2. The answer is: Gerbils and hamsters can lift the same amount

If you are unsure why, please review **Chapter 9, Hypothesis Testing – Understanding Type I and Type II Error**

3. The answer is: Redundant data

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Managing Duplicate and Redundant Data](#)

4. The answer is: Percentage

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Understanding Frequencies and Percentages](#)

5. The answer is: Accept the alternative hypothesis and reject the null hypothesis

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Learning P-Value and Alpha](#)

6. The answer is: Null hypothesis

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Differentiating Null Hypothesis and Alternative Hypothesis](#)

7. The answer is: A scatter plot

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

8. The answer is: CSV

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through Data Types and File Types](#)

9. The answer is: The appendix

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Elements](#)

10. The answer is: A specification mismatch

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Invalid Data, Specification Mismatch, and Data Type Validation](#)

11. The answer is: The Population of India based on the Geographical Region

If you are unsure why, please review [Chapter 12, Reporting Process – Designing Reports](#)

12. The answer is: Real-time

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Delivery](#)

13. The answer is: OLAP

If you are unsure why, please review [Chapter 3, Collecting Data – Understanding OLTP and OLAP](#)

14. The answer is: Parsing

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Parsing Your Data](#)

15. The answer is: A pie chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

16. The answer is: A data lake

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding the Concept of Warehouses and Lakes](#)

17. The answer is: Performance analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Checking on Performance](#)

18. The answer is: Technical experts

If you are unsure why, please review [Chapter 12, Reporting Process – Knowing what to Consider when Making a Report](#)

19. The answer is:

Joined Table		
ClientID	Name	City
1	Smith, Laurence	Austin, TX
2	Brown, Betty	Denver, CO
3	Hook, Phil	Tulsa, OK
4	Roark, Jen	NULL
5	Cox, Jona	NULL
6	Humbert, Ren	NULL
7	NULL	Phoenix, AZ
8	NULL	Seattle, WA
9	NULL	Baltimore, MD

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Merging Data](#)

20. The answer is: A recurring report

If you are unsure why, please review [Chapter 11, Types of Reports – Understanding Recurring Reports](#)

21. The answer is: 3

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Understanding Measures of Central Tendency](#)

22. The answer is: Link analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Finding Links](#)

23. The answer is: Removing variables

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Updating Stored Data](#)

24. The answer is: A new student taking the test would most likely get a C

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

25. The answer is: A waterfall chart

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

26. The answer is: A snowflake schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through the Data Schema and its Types](#)

27. The answer is: Recoding a category into a number

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Recoding Variables](#)

28. The answer is: An execution plan

If you are unsure why, please review [Chapter 3, Collecting Data – Optimizing Query Structure](#)

29. The answer is: Half of everyone who can access data is either a marketing analyst or a business analyst

If you are unsure why, please review [Chapter 13, Common Visualizations – Charting Lines, Circles, and Dots](#)

30. The answer is: A data mart

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding the Concept of Warehouses and Lakes](#)

31. The answer is: A research report

If you are unsure why, please review [Chapter 11, Types of Reports – Understanding Ad hoc and Research Reports](#)

32. The answer is: A snowflake schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through the Data Schema and its Types](#)

33. The answer is: Active End

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Updating Stored Data](#)

34. The answer is: Data consolidation

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Master Data Management \(MDM\)](#)

35. The answer is: A geographic map

If you are unsure why, please review [Chapter 13, Common Visualizations – Understanding Heat Maps, Tree Maps, and Geographic Maps](#)

36. The answer is: Specification mismatch

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Invalid Data, Specification Mismatch, and Data Type Validation](#)

37. The answer is: Single choice

If you are unsure why, please review [Chapter 3, Collecting Data – Collecting Your Own Data](#)

38. The answer is: OR

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Shaping Data with Common Functions](#)

39. The answer is: Data integrity

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

40. The answer is: Reasonable expectations

If you are unsure why, please review [Chapter 15, Data Quality and Management – Validating Quality](#)



41. The answer is: 21.7

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Finding Variance and Standard Deviation](#)

42. The answer is: A word cloud

If you are unsure why, please review [Chapter 13, Common Visualizations – Understanding Infographics and Word Clouds](#)

43. The answer is: Notify the impacted parties

If you are unsure why, please review [Chapter 14, Data Governance – Knowing Use Requirements](#)

44. The answer is: Structured

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

45. The answer is: Get approval

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding the Report Development Process](#)

46. The answer is: 0.3

If you are unsure why, please review [Chapter 9, Hypothesis Testing – Learning p-Value and Alpha](#)

47. The answer is: Z-score

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Understanding Z-Scores](#)

48. The answer is: One-to-one

If you are unsure why, please review [Chapter 14, Data Governance – Handling Entity Relationship Requirements](#)

49. The answer is: Binary

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through Data Types and File Types](#)

50. The answer is: 7%

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Calculating Percent Change and Percent Difference](#)

51. The answer is: 7.3 to 8.7

If you are unsure why, please review [Chapter 8, Common Techniques in Descriptive Statistics – Discovering Confidence Intervals](#)

52. The answer is: Left join

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Merging Data](#)

53. The answer is: Chi-square

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Knowing Chi-Square](#)

54. The answer is: PII

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Classifications](#)

55. The answer is: The same information recorded in multiple rows

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Managing Duplicate and Redundant Data](#)

56. The answer is: An outlier

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Finding Outliers](#)

57. The answer is: Web services

If you are unsure why, please review [Chapter 3, Collecting Data – Utilizing Public Sources of Data](#)

58. The answer is: Reduction

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Calculating Derived and Reduced Variables](#)

59. The answer is: Data acquisition

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

60. The answer is: Data profiling

If you are unsure why, please review [Chapter 15, Data Quality and Management – Validating Quality](#)

61. The answer is: Simple linear regression

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Simple Linear Regression](#)

62. The answer is: Transposition

If you are unsure why, please review [Chapter 5, Data Wrangling and Manipulation – Shaping Data with Common Functions](#)

63. The answer is: Completeness

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Quality Control](#)

64. The answer is: Uniform

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Discovering Distributions](#)

65. The answer is: PHI

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Classifications](#)

66. The answer is: A self-service report

If you are unsure why, please review [Chapter 11, Types of Reports – Knowing about Self-Service Reports](#)

67. The answer is: Trend analysis

If you are unsure why, please review [Chapter 6, Types of Analytics – Discovering Trends](#)

68. The answer is: AWS QuickSight

If you are unsure why, please review [Chapter 11, Types of Reports – Knowing Important Analytical Tools](#)

69. The answer is: Weeks

If you are unsure why, please review [Chapter 12, Reporting Process – Knowing What to Consider When Making a Report](#)

70. The answer is: Pairwise deletion

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Dealing with Missing Data](#)

71. The answer is: A static report

If you are unsure why, please review [Chapter 11, Types of Reports – Distinguishing Static and Dynamic Reports](#)

72. The answer is: The FAQs

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Elements](#)

73. The answer is: Undefined fields and machine data

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

74. The answer is: XML

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through Data Types and File Types](#)

75. The answer is: 34

If you are unsure why, please review [Chapter 7, Measures of Central Tendency and Dispersion – Calculating Range and Quartiles](#)

76. The answer is: Observation

If you are unsure why, please review [Chapter 3, Collecting Data – Collecting Your Own Data](#)

77. The answer is: Filtering

If you are unsure why, please review [Chapter 3, Collecting Data – Optimizing Query Structure](#)

78. The answer is: Interpolation

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Dealing with Missing Data](#)

79. The answer is: Nonparametric distributions are never normal

If you are unsure why, please review [Chapter 4, Cleaning and Processing Data – Understanding Non-Parametric Data](#)

80. The answer is: Chi-square test for independence

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Knowing Chi-Square](#)

81. The answer is: Only load information that is new or has changed

If you are unsure why, please review [Chapter 3, Collecting Data – Differentiating ETL and ELT](#)

82. The answer is: A snowflake schema

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Going Through the Data Schema and its Types](#)

83. The answer is: A histogram

If you are unsure why, please review [Chapter 13, Common Visualizations – Comprehending Charts with Bars](#)

84. The answer is: Key-value pairs

If you are unsure why, please review [Chapter 2, Data Structures, Types, and Formats – Understanding Structured and Unstructured Data](#)

85. The answer is: Correlation

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Calculating Correlations](#)

86. The answer is: Data encryption

If you are unsure why, please review [Chapter 14, Data Governance – Understanding Data Security](#)

87. The answer is: T-test

If you are unsure why, please review [Chapter 10, Introduction to Inferential Statistics – Understanding T-Tests](#)

88. The answer is: Streamlining data access

If you are unsure why, please review [Chapter 15, Data Quality and Management – Understanding Master Data Management \(MDM\)](#)

89. The answer is: Subscription

If you are unsure why, please review [Chapter 12, Reporting Process – Understanding Report Delivery](#)

90. The answer is: Data deletion

If you are unsure why, please review [Chapter 14, Data Governance – Knowing Use Requirements](#)