Week 1

1. Fill in the blank: Data _____ refers to the accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle.

   - (●) integrity
   - ( ) analysis
   - ( ) replication
   - ( ) sampling

   Correct
   Data integrity is the accuracy, completeness, consistency, and trustworthiness of data throughout its life cycle.

2. A financial analyst imports a dataset to their computer from a storage device. As it's being imported, the connection is interrupted, which compromises the data. Which of the following processes caused the compromise?

   - ( ) Data gathering
   - ( ) Data manipulation
   - (●) Data transfer
   - ( ) Data analysis

   Correct
   Data transfer caused the compromise. When a data transfer is interrupted, it can result in an incomplete dataset.

3. A data analyst is given a dataset for analysis. It includes data about the total population of every country in the previous 20 years. Which of the following questions can the analyst use this dataset to address? Select all that apply.

   - [ ] What was the effect of migration on the population of a certain country?
   - [ ] What was the reason for the population increase in a certain country?
   - [x] What was the difference in population between two specific countries in 2018?

   Correct
   The analyst could use the dataset to find the average population of a certain country from 2015 through 2020 and the difference in population between two specific countries in 2018.

   - [x] What was the average population of a certain country from 2015 through 2020?

   Correct
   The analyst could use the dataset to find the average population of a certain country from 2015 through 2020 and the difference in population between two specific countries in 2018.

4. A data analyst is given a dataset for analysis. To use the template for this dataset, click the link below and select "Use Template."

   Link to template: June 2014 Invoices

   OR

   If you don't have a Google account, download the CSV file directly from the attachment below.

   June 2014 Invoices - Sheet1
   CSV File

   Rows 10 and 11 contain duplicate data.

   - ○ True
   - ◉ False

   Correct
   Rows 10 and 11 do not contain duplicate data.

5. A data analyst at a software company wants to learn more about industry competitors. Because the software industry has more mergers than any other field, the companies and their products are constantly evolving. The analyst has a dataset from three years ago, and they notice that many of the companies and products in the dataset have changed. What makes the analyst decide that the data is insufficient, so they should generate fresh data instead?

   - ◉ It is outdated data
   - ○ It is data that keeps updating
   - ○ It is data from only one source
   - ○ It is geographically limited data

   Correct
   This example describes outdated data, which is insufficient. If a dataset is outdated, that means the data is old and probably no longer relevant.

6. A car manufacturer wants to learn more about the brand preferences of electric car owners. There are millions of electric car owners in the world. Who should the company survey?

   - ◉ A sample of all electric car owners
   - ○ The entire population of electric car owners
   - ○ A sample of car owners who most recently bought an electric car
   - ○ A sample of car owners who have owned more than one electric car

   Correct
   The company should survey a sample of all electric car owners.

7. Fill in the blank: Sampling bias in data collection happens when a sample isn't representative of _____.

- (●) the population as a whole
- ( ) a subset of the population
- ( ) the population most affected by the data
- ( ) a dataset about the population

Correct
Sampling bias in data collection happens when a sample isn't representative of the population as a whole

8. Data and business objectives might not align for a number of reasons. Which of the following issues can prevent alignment? Select all that apply.

- [ ] Data integrity
- [x] Sampling bias

Correct
Insufficient data and sampling bias can prevent alignment.

- [ ] Data visualization
- [x] Insufficient data

Correct
Insufficient data and sampling bias can prevent alignment.

Week 2

1. Fill in the blank: Conditional formatting is a spreadsheet tool that changes how _____ appear when values meet a specific condition.

- ( ) charts
- ( ) filters
- ( ) queries
- (●) cells

Correct
Conditional formatting is a spreadsheet tool that changes how cells appear when values meet a specific condition.

2. A delimiter is a character that indicates the beginning or end of a data item. The split text to columns tool uses a delimiter to accomplish what task?

- ○ To split duplicate substrings
- ○ To split one column into two
- ◉ To specify where to split a text string
- ○ To change the format of a column of text

Correct
The split text to columns tool uses a delimiter to specify where to split a text string.

3. Fill in the blank: A predetermined structure that includes a function's required information and its proper placement is called _____.

- ○ script
- ○ symbol
- ◉ syntax
- ○ standard

Correct
Syntax is a predetermined structure that includes a function's required information and its proper placement.

4. You are working with the following selection of a spreadsheet:

	A	B
1	**Customer**	**Address**
2	Sally Stewart	9912 School St. North Wales, PA 19454
3	Lorenzo Price	8621 Glendale Dr. Burlington, MA 01803
4	Stella Moss	372 W. Addison Street Brandon, FL 33510
5	Paul Casey	9069 E. Brickyard Road Chattanooga, TN 37421

In order to extract the five-digit postal code from Brandon, FL, what is the correct function?

- ◉ =RIGHT(B4,5)
- ○ =LEFT(5,B4)
- ○ =RIGHT(5,B4)
- ○ =LEFT(B4,5)

Correct
The correct syntax is =RIGHT(B4,5). The RIGHT function returns a set number of characters from the right side of a text string. B4 is the specified cell. And 5 is the number of characters to return.

5. A data analyst in a human resources department is working with the following selection of a spreadsheet:

	A	B	C	D
1	**Year Hired**	**Last 4 of SS#**	**Department**	**Employee ID**
2	2019	1192	Marketing	
3	2014	2683	Operations	
4	2020	1939	Strategy	
5	2009	3208	Graphics	

They want to create employee identification numbers (IDs) in column D. The IDs should include the year hired plus the last four digits of the employee's Social Security Number (SS#). What function will create the ID 20142683 for the employee in row 3?

- ○ =CONCATENATE(A3!B3)
- ◉ =CONCATENATE(A3,B3)
- ○ =CONCATENATE(A3+B3)
- ○ =CONCATENATE(A3*B3)

Correct
To create the ID 20142683 for the employee in row 3, the function is =CONCATENATE(A3,B3). CONCATENATE joins together two or more text strings. (A3,B3) are the locations of the strings to be joined.

6. A data analyst at an e-commerce company is working with a spreadsheet containing last month's sales. The most expensive product their company sells costs $49.99, so they want to quickly confirm that all of the data in the Sales column is $49.99 or less. What function can they use?

- ○ COUNT
- ○ SUMIF
- ○ SUM
- ◉ COUNTIF

Correct
They can use COUNTIF, which is a function that returns the number of cells that match a specified value or parameter.

7. The V in VLOOKUP stands for what?

- ○ Variable
- ○ Virtual
- ◉ Vertical
- ○ Visual

Correct
The V in VLOOKUP stands for vertical. VLOOKUP is a spreadsheet function that vertically searches for a certain value in a column to return a corresponding piece of information.

8. A data analyst needs to combine two datasets. Each dataset comes from a different system, and the systems store data in different ways. What can the data analyst do to ensure the data is compatible?

- ○ Use a data visualization
- ○ Merge the data
- ◉ Map the data
- ○ Apply a data structure

Correct
Data analysts use data mapping to note differences in data sources in order to ensure the data is compatible.

Week 3

1. A data analyst is analyzing medical data for a health insurance company. The dataset contains billions of rows of data. Which of the following tools will handle the data most efficiently?

- ○ A word processor
- ◉ SQL
- ○ A presentation
- ○ A spreadsheet

Correct
SQL will handle the data most efficiently. SQL can handle huge amounts of data.

2. In which of the following situations would a data analyst use SQL instead of a spreadsheet? Select all that apply.

- ☑ When working with a huge amount of data

Correct
A data analyst would use SQL instead of a spreadsheet to work with a huge amount of data. SQL can also quickly pull information from many different sources in a database and record queries and changes throughout a project.

- ☐ When using the COUNTIF function to find a specific piece of information
- ☑ When quickly pulling information from many different sources in a database

Correct
A data analyst would use SQL instead of a spreadsheet to work with a huge amount of data. SQL can also quickly pull information from many different sources in a database and record queries and changes throughout a project.

- ☑ When recording queries and changes throughout a project

Correct
A data analyst would use SQL instead of a spreadsheet to work with a huge amount of data. SQL can also quickly pull information from many different sources in a database and record queries and changes throughout a project.

3. A data analyst creates many new tables in their company's database. When the project is complete, the analyst wants to remove the tables so they don't clutter the database. What SQL commands can they use to delete the tables?

- ○ UPDATE
- ○ CREATE TABLE IF NOT EXISTS
- ○ INSERT INTO
- ◉ DROP TABLE IF EXISTS

Correct
The analyst can use the DROP TABLE IF EXISTS query to delete the tables so they don't clutter the database.

4. You are working with a database table that contains invoice data. The table includes columns for *invoice_id* and *billing_city*. You want to remove duplicate entries for billing city and sort the results by invoice ID.

You write the SQL query below. Add a DISTINCT clause that will remove duplicate entries from the *billing_city* column.

NOTE: The three dots (...) indicate where to add the clause.

```
1  SELECT distinct billing_city
2  FROM
3  invoice
4  ORDER BY
5  invoice_id
```

Run

Reset

What billing city appears in row 15 of your query result?

- ○ Oslo
- ○ Santiago
- ○ London
- ◉ Reno

Correct
The clause `DISTINCT billing_city` will remove duplicate entries from the *billing_city* column. The complete query is `SELECT DISTINCT billing_city FROM invoice ORDER BY invoice_id`. The DISTINCT clause removes duplicate entries from your query result. The billing city Reno appears in row 15 of your query result.

5. You are working with a database table that contains customer data. The table includes columns about customer location such as *city*, *state*, *country*, and *postal_code*. You want to check for postal codes that are greater than 7 characters long.

   You write the SQL query below. Add a LENGTH function that will return any postal codes that are greater than 7 characters long.

```
1  SELECT
2  *
3  FROM
4  customer
5  WHERE LENGTH(postal_code)>7
```

   Run

   Reset

   What is the last name of the customer that appears in row 10 of your query result?

   - (●) Hughes
   - ( ) Ramos
   - ( ) Rocha
   - ( ) Brooks

   Correct

   The function `LENGTH(postal_code) > 7` will return any postal codes that are greater than 7 characters long. The complete query is `SELECT * FROM customer WHERE LENGTH(postal_code) > 7`. The LENGTH function counts the number of characters a string contains. Hughes is the last name of the customer that appears in row 10 of your query result.

6. Fill in the blank: _____ refers to the process of converting data from one type to another.

   - ( ) Formatting
   - ( ) Querying
   - ( ) Cleaning
   - (●) Typecasting

   Correct

   Typecasting refers to the process of converting data from one type to another.

7. The CAST function can be used to convert the DATE datatype to the DATETIME datatype.

   - (●) True
   - ( ) False

   Correct

   The CAST function can be used to convert the DATE datatype to the DATETIME datatype. CAST can be used to convert any database field from one datatype to another.

8. What SQL function lets you add strings together to create new text strings that can be used as unique keys?

- ○ COALESCE
- ○ LENGTH
- ○ CAST
- ◉ CONCAT

Correct
The CONCAT function lets you add strings together to create new text strings that can be used as unique keys.

9. You are working with a database table that contains customer data. The table includes columns about customer location such as *city, state,* and *country*. The state names are abbreviated. You want to retrieve the first 2 letters of each state name. You decide to use the SUBSTR function to retrieve the first 2 letters of each state name, and use the AS command to store the result in a new column called *new_state*.

You write the SQL query below. Add a statement to your SQL query that will retrieve the first 2 letters of each state name and store the result in a new column as *new_state*.

NOTE: The three dots (...) indicate where to add the statement.

```
1  SELECT
2  customer_id,
3  SUBSTR(state,1,2) as new_state
4  FROM
5  customer
6  ORDER BY
7  state DESC
```

Run
Reset

What customer ID number appears in row 9 of your query result?

- ○ 10
- ◉ 47
- ○ 3
- ○ 55

Correct
The statement `SUBSTR(state, 1, 2) AS new_state` will retrieve the first 2 letters of each state name and store the result in a new column as *new_state*. The complete query is `SELECT customer_id, SUBSTR(state, 1, 2) AS new_state FROM customer ORDER BY state DESC`. The SUBSTR function extracts a substring from a string. This function instructs the database to return 2 characters of each state name, starting with the first character. The customer ID number 47 appears in row 9 of your query result.

Week 4

1. Fill in the blank: Once data is clean, a data analyst moves on to _____ and verification.

- (●) reporting
- ( ) processing
- ( ) publishing
- ( ) confirming

Correct
Once data is clean, a data analyst moves on to reporting and verification.

2. A data analyst is in the verification step. They consider the business problem, the goal, and the data involved in their analytics project. What scenario does this describe?

- ( ) Reporting on the data
- ( ) Considering the stakeholders
- ( ) Visualizing the data
- (●) Seeing the big picture

Correct
To see the big picture when verifying data cleaning, consider the business problem, the goal, and the data.

3. Which of the following functions automatically remove extra spaces when cleaning data?

- ( ) SNIP
- ( ) REMOVE
- ( ) CLEAR
- (●) TRIM

Correct
TRIM automatically removes extra spaces when cleaning data.

4. A data analyst uses the COUNTA function to count which of the following?

- ( ) The total number of headers in a specific range
- ( ) The total number of entries in a changelog
- (●) The total number of values within a specified range
- ( ) The specific numbers in a dataset

Correct
A data analyst uses the COUNTA function to count the total number of values within a specified range.

5. Which SQL tool considers one or more conditions, then returns a value as soon as a condition is met?

- ○ THEN
- ○ WHEN
- ○ ELSE
- ◉ CASE

Correct
CASE considers one or more conditions, then returns a value as soon as a condition is met.

6. What is the process of tracking changes, additions, deletions, and errors during data cleaning?

- ○ Cataloging
- ○ Observation
- ◉ Documentation
- ○ Recording

Correct
Documentation is the process of tracking changes, additions, deletions, and errors during data cleaning.

7. Fill in the blank: A changelog contains a _____ list of modifications made to a project.

- ○ synchronized
- ○ random
- ○ approximate
- ◉ chronological

Correct
A data analyst uses a changelog to access the information needed. A changelog is a file that contains a chronological list of modifications made to a project.

8. A data analyst commits a query to the repository as a new and improved query. Then, they specify the changes they made and why they made them. This scenario is part of what process?

- ○ Communicating with stakeholders
- ◉ Creating a changelog
- ○ Visualizing data
- ○ Reporting data

Correct
Specifying the changes an analyst made and why they made them is part of creating a changelog.

Course Challenge

1. **Scenario 1, questions 1-5**

   You are a data analyst at a small analytics company. Your company is hosting a project kick-off meeting with a new client, Meer-Kitty Interior Design. The agenda includes reviewing their goals for the year, answering any questions, and discussing their available data.

   Before the meeting you review the About Us tab on their website and their business plan, linked below:

   Meer-Kitty Interior Design About Us Page.pdf
   PDF File

   Meer-Kitty Interior Design Business Plan.pdf
   PDF File

   Meer-Kitty Interior Design has two goals. They want to expand their online presence, which means getting their company and brand known by as many people as possible. They also want to launch a line of high-quality indoor paint to be sold in-store and online. You decide to consider the data about indoor paint first.

   To use the template for the survey feedback, click the link below and select "Use Template."

Link to template: Kitty Survey Feedback

OR

If you don't have a Google account, download the file directly from the attachment below.

Kitty Survey Feedback - Meer-Kitty survey feedback
CSV File

When you refer to the **Meer-Kitty survey feedback** tab, you are pleased to find that the available data is aligned to the business objective. However, you do some research about confidence level for this type of survey and learn that you need at least 120 unique responses for the survey results to be useful. Therefore, the dataset has two limitations: First, there are only 40 responses; second, a Meer-Kitty superfan, User 588, completed the survey 11 times.

As the survey has too few responses and numerous duplicates that are skewing results, you should remove the duplicates and continue analyzing the remaining 29 responses.

- ○ True
- ◉ False

Correct
Analyzing only 29 responses will not provide sufficient insights to make an effective business decision.

2. **Scenario 1 continued**

During the meeting, you also learn that Meer-Kitty videos are hosted on their website. For each product offered, there is an accompanying video for customers to learn more. So, more views for a video suggests greater consumer interest.

Your goal is to identify which videos are most popular, so Meer-Kitty knows what topics to explore in the future. Unfortunately, Meer-Kitty has just three months of data available because they only recently launched the videos on their site.

Without enough data to identify long-term trends about the video subjects that people prefer, what are your available options? Select all that apply.

- [ ] Watch the videos and use your gut instinct to identify which are most successful.
- [x] Ask to wait for more data and provide Meer-Kitty with an updated timeline.

Correct
Without enough data to identify long-term trends, one option is to talk with stakeholders and ask to adjust the objective. You could also ask to wait for more data and provide an updated timeline.

- [ ] Move ahead with the data you have to determine the top video subjects.
- [x] Talk with Meer-Kitty stakeholders and ask to adjust the objective.

Correct
Without enough data to identify long-term trends, one option is to talk with stakeholders and ask to adjust the objective. You could also ask to wait for more data and provide an updated timeline.

3. **Scenario 1 continued**

Now that you've identified some limitations with Meer-Kitty's data, you want to communicate your concerns to stakeholders. In addition to insufficient video trend data, your main concern with the indoor paint survey is that the data isn't representative of the population as a whole.

Clearly, one particular respondent, the superfan, is overrepresented. What does this situation describe?

- ( ) Margin of error
- ( ) Statistical significance
- ( ) Confidence level
- (•) Sampling bias

Correct
This situation describes sampling bias. Sampling bias occurs when a sample isn't representative of the population as a whole.

4. **Scenario 1 continued**

The stakeholders understand your concerns and agree to repeat the indoor paint survey. In a few weeks, you have a much better dataset with more than 150 responses and no duplicates.

To use the template for the survey feedback, click the link below and select "Use Template."

Link to template: Kitty Survey Feedback

OR

If you don't have a Google account, download the file directly from the attachment below.

Kitty Survey Feedback - New Meer-Kitty survey feedback
CSV File

If you are using the template, please refer to the **New Meer-Kitty survey feedback** tab. You notice that questions 4 and 5 are dependent on the respondent's answer to question 3. So, you need to determine how many people answered Yes to question 3, then compare that to responses to questions 4 and 5. That way, you will know if questions 4 and 5 have any nulls.

You decide to use a spreadsheet tool that changes how cells appear when they meet a certain value — in this case, the word Yes. You are using VLOOKUP.

- ○ True
- ◉ False

Correct
To change how cells appear when they meet a certain value, use conditional formatting.

5. **Scenario 1, continued**

You have finished cleaning the data to ensure it is complete, correct, and relevant to the problem you're trying to solve. Then, you complete the verification and reporting processes to share the details of your data-cleaning effort with your team.

Your team notes one aspect of data cleaning that would help improve the dataset. They point out that the new survey also has a new question in Column G: "What are your favorite indoor paint colors?" This was a free-response question, so respondents typed in their answers. Some people included multiple different colors of paint. In order to determine which colors are most popular, it will be necessary to put each color in its own cell.

You decide to use a spreadsheet function to divide the text strings in Column G around the commas and put each fragment into a new, separate cell. You are using the SPLIT function.

- ◉ True
- ○ False

Correct
To divide the text strings in Column G around the commas and put each fragment into a new, separate cell, you use SPLIT. SPLIT is a spreadsheet function that divides text around a specified character and puts each fragment into a new, separate cell.

6. **Scenario 2, questions 6-10**

   You've completed this program and are interviewing for a junior data scientist position. The job is at B.Spoke Market Research, a company that analyzes market conditions using customer surveys and other research methods. The detailed job description can be found below:

   C4 B.Spoke Market Research Job Description.pdf
   PDF File

   So far, you've had a phone interview with a recruiter and you've secured a second interview with the B.Spoke team. The recruiter's email can be found below:

   C4 S2 Email from Recruiter.pdf
   PDF File

You arrive 15 minutes early for your interview. Soon, you are escorted into a conference room, where you meet Jodie Choi, the data science lead. After welcoming you, the behavioral interview begins.

For your first question, your interviewer wants to learn about your experience with spreadsheets. She says: Sometimes the team needs data that is stored in different spreadsheets. So, we use spreadsheet functions to help us find the information we need.

What function would you use to search for a certain value in a spreadsheet column to return the corresponding piece of information?

- ○ COUNTIF
- ◉ VLOOKUP
- ○ RETURN
- ○ SEARCH

Correct
To search for a certain value in a spreadsheet column to return the corresponding piece of information, use VLOOKUP.

7. **Scenario 2, continued**

Next, your interviewer wants to know more about your understanding of tools that work in both spreadsheets and SQL queries. She explains that the data her team receives from customer surveys sometimes has many duplicate entries.

She says: Spreadsheets have a great tool for that called remove duplicates. But when writing a SQL query, what command should you include in your SELECT statement to remove duplicates?

- ○ DIFFERENT
- ○ DISCRETE
- ○ DIVERSE
- ◉ DISTINCT

Correct
To remove duplicates in a SQL query, include DISTINCT in your SELECT statement.

8. **Scenario 2, continued**

Now, your interviewer explains that the data team usually works with very large amounts of customer survey data. After receiving the data, they import it into a SQL table. But sometimes, the new dataset imports incorrectly and they need to change the format.

She asks: Is there a SQL function that can convert data types such as currency, dates, and times in a SQL table?

- ◉ Yes, data types including currency, dates, and times can be converted.
- ○ No, only currency can be converted.

Correct
The CAST function is used to convert currency, dates, and times in a SQL table from one datatype to another.

9. **Scenario 2, continued**

Next, your interviewer explains that one of their clients is an online retailer that needs to create product numbers for a vast inventory. Her team does this by combining the text strings for product number, manufacturing date, and color.

She asks: If you encountered a situation where you wanted to add strings together to create new text strings, which SQL function would you use?

- ○ COALESCE
- ○ COMBINE
- ◉ CONCAT
- ○ CREATE

Correct
To add strings together to create new text strings, use the CONCAT function.

10. Scenario 2, continued

For your final question, your interviewer explains that her team often uses the TRIM function when writing SQL queries.

She asks: What is the TRIM function used for in SQL?

- ○ To eliminate null values
- ○ To return the smallest numeric value from a list
- ○ To shorten the list of results
- ◉ To eliminate extra leading or trailing spaces

Correct
The TRIM function is used to eliminate extra leading or trailing spaces.