1 Foundations: Data, Data, Everywhere

week 1

1. **statistics vs. machine learning vs. analytics**
    a. statistics: make a few important decisions under uncertainty. Need care and rigor.
    b. machine learning: automate, make many, many, many decisions under uncertainty. Need performance.
    c. analytics: don't know how many decisions you want to make before you begin, look for inspiration, encounter your unknown unknowns. You want to understand your world. Need speed.
2. **Data scientists vs. analysts**
    a. Data scientists create new questions using data
    b. Data analysts find answers to existing questions by creating insights from data sources.
3. **data analysis vs. Data analytics**
    a. Data analysis is the collection, transformation, and organization of data in order to draw conclusions, make predictions, and drive informed decision-making.
    b. Data analytics is the science of data. It's a very broad concept that encompasses everything from the job of managing and using data to the tools and methods that data workers use each and every day.
4. solving a project, ask
    a. How do I define success for this project?”
    b. What kind of results are needed?
    c. Who will be informed?
    d. Am I answering the question being asked?
    e. How quickly does a decision need to be made? (if working on a rush project, you might need to rely on your own knowledge and experience more than usual)
5. **data analysis life cycle**
    a. Ask: Business Challenge/Objective/Question
    b. Prepare: Data generation, collection, storage, and data management
    c. Process: Data cleaning/data integrity
    d. Analyze: Data exploration, visualization, and analysis
    e. Share: Communicating and interpreting results
    f. Act: Putting your insights to work to solve the problem
6. **data ecosystem** refers to the various elements that interact with one another to produce, manage, store, organize, analyze, and share data.
7. subject-matter experts
    a. Review the results of data analysis and identify any inconsistencies, make sense of gray areas
    b. Validate the choices made as a result of the data insights
    c. Offer insights into the business problem
8. Sharing the results of your analysis with colleagues who are very familiar with the business problem supports Data-driven decision-making?

week 2

1. **Analytical skills**: qualities and characteristics associated with solving problems using facts
    a. curiosity
    b. understanding context: how you group things into categories.
        i. Identifying the motivation behind the collection of a dataset
        ii. Adding descriptive headers to columns of data in a spreadsheet
        iii. Gathering additional information about data to understand the broader picture
    c. having a technical mindset
    d. data design: how you organize information. e.g. organize contact list on phone. If you make decisions that are informed by data, you are more likely to make more informed and effective decisions.
    e. data strategy: management of people, processes and tools used in data analysis. high-level
2. **Analytical thinking**: identifying and defining a problem and then solving it by using data in an organized step-by-step manner
    a. visualizations
    b. strategy
    c. problem-orientation
    d. correlation: does not equal causation
    e. big-picture and detail-oriented thinking
3. ask **5 whys** to find the root cause
4. **Gap analysis**: a method for examining and evaluating how a process works currently in order to get where you want to be in the future
5. What did we **not consider** before?

week 3

1. data life cycle (* nothing to do with data analysis life cycle)
    a. plan
    b. capture,
    c. manage: how we care for our data, how and where it's stored, the tools used to keep it safe and secure, and the actions taken to make sure that it's maintained properly. important to data cleansing.
    d. analyze: using data to solve problems, make great decisions, and support business goals.
    e. archive
    f. destroy: privacy
2. More on **data analysis life cycle**: DAC1 The data analysis process.pdf
    a. Ask: understand expectation of stakeholders
    b. Prepare
    c. Process: cleaning, remove outliers
    d. Analyze [quiz]
        i. ✅ Use spreadsheets to aggregate data
        ii. ❌ Choose the format of spreadsheet headings

iii. ✅ Create a report from the data
iv. ✅ Use a formula to perform calculations

e. Share: Communicate, use visualization and make decisions
f. Act
   i. validate insights
   ii. finalize a strategy
   iii. put a plan into action

3. query ___ information in a database
   a. ✅ request
   b. ✅ retrieve
   c. ✅ update
   d. ❌ visualize

week 4

1. row: observation. An observation includes all of the **attributes** for what is contained in the row.
2. attribute: a characteristic or quality of data used to label a **column**.
3. Google Sheets Training and Help
4. Google Sheets Cheat Sheet
5. text wrapping is used for
   a. ✅ To allow all of the text to fit inside a cell
   b. ❌ To clip text within a cell so it doesn't overflow into an adjacent cell
   c. ❌ To allow text to overflow into an adjacent cell
   d. ❌ To remove text that is too long to fit in a cell
6. SQL
   a. `LIKE`: for pattern, e.g. `WHERE field1 LIKE 'Ch%'`
   b. `%` or `*`: wildcard
   c. comments: between `/* */` or after `--`
   d. alias: before query, `field1 AS last_name`
   e. `<>` or `!=`: does not equal
7. W3Schools SQL Tutorial
8. SQL Cheat Sheet
9. Tableau How-to Video, more resources on visualization

week 5

1. Issue vs. Problem vs. Question
   a. A problem is an obstacle to be solved
   b. an issue is a topic to investigate
   c. a question is designed to discover information.
2. fairness: ensuring that your analysis doesn't create or reinforce bias.
   a. requires using processes and systems that are fair and inclusive to everyone.

3. What steps do data analysts take to ensure fairness when collecting data?
    a. Understand the social context
    b. Include data self-reported by individuals
    c. ❌ Clean the data provided
    d. Use an inclusive sample population
4. On a railway line, peak ridership occurs between 7:00 AM and 5:00 PM. The fairness of a passenger survey could be improved by over-sampling data from Nighttime riders, an under-represented group

2 Ask Questions to Make Data-Driven Decisions

week 1

1. look at big picture for real problem, not symptoms
2. Questions to ask yourself in the six data analysis phases
    a. Ask
        i. What are my stakeholders saying their problems are?
        ii. Now that I've identified the issues, how can I help the stakeholders resolve their questions?
    b. Prepare
        i. What do I need to figure out how to solve this problem?
        ii. What research do I need to do?
    c. Process
        i. What data errors or inaccuracies might get in my way of getting the best possible answer to the problem I am trying to solve?
        ii. How can I clean my data so the information I have is more consistent?
    d. Analyze
        i. What story is my data telling me?
        ii. How will my data help me solve this problem?
        iii. Who needs my company's product or service? What type of person is most likely to use it?
    e. Share
        i. How can I make what I present to the stakeholders engaging and easy to understand?
        ii. What would help me understand this if I were the listener?
    f. Act
        i. How can I use the feedback I received during the share phase (step 5) to actually meet the stakeholder's needs and expectations?
3. These six steps help break the data analysis process into smaller, manageable parts, which is called **structured thinking**. This process involves four basic activities:
    a. Recognizing the current problem or situation
    b. Organizing available information
    c. Revealing gaps and opportunities

d. Identifying your options

4. **Common problem types**
   a. Making predictions: use data to make an informed decision about how things may be in the future
   b. categorizing things: assign info to different groups or clusters based on common features.
   c. spotting sth unusual: identify data that is different from the norm
   d. identifying themes: group categorized info into broader concepts. e.g. group employees categorized by types and tasks into of high and low productivity. vs. categorizing things: takes those categories a step further by grouping them into broader themes.
   e. discovering connections: find similar challenges faced by different entities, and combine data and insights to address them. e.g. suppliers and clients face the same problem: bad rubber for wheel producer - unsafe wheels for scooter company. Share data openly and collaborate will help to find solutions faster.
   f. finding patterns: use historical data to understand what happened in the past and is therefore likely to happen again. e.g. 1. customer buying habits. e.g. 2. using historical data to create a report that shows when batteries on critical equipment have been replaced – help implement proper maintenance to prevent battery failures in the future.
5. ineffective questions
   a. leading questions: lead the answer in a certain way. e.g. “it is… isn’t it?”
   b. close-ended questions: result in Yes or No
   c. Too vague and lack context: e.g. “do you prefer chocolate or vanilla?”
6. **SMART questions**
   a. Specific: simple, significant, focused on a single topic or a few closely related

~~“Are kids getting enough exercise these days?”~~

“What percentage of kids achieve the recommended 60 minutes of physical activity at least five days a week?”

ideas

   b. Measurable: can be quantified and assessed

~~“Why did our recent video go viral?”~~

“How many times was our video shared on social channels the first week it was posted?”

~~“How can we get customers to recycle our product packaging?”~~

“What design features will make our packaging easier to recycle?”

   c. Action-oriented: encourage change
   d. Relevant: important, have significance to the problem you're trying to solve

~~"Why does it matter that Pine Barrens tree frogs started disappearing?"~~

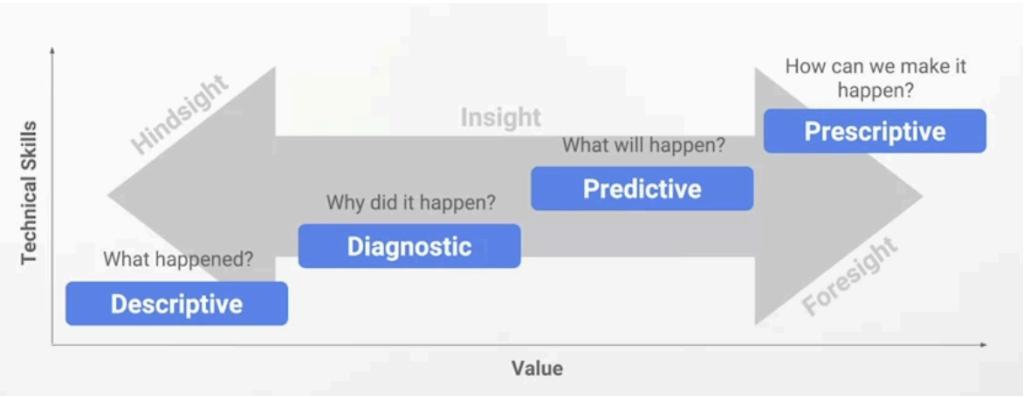
"What environmental factors changed in Durham, North Carolina, between 1983 and 2004 that could cause Pine Barrens tree frogs to disappear from the Sandhills Regions?"

   e. Time-bound: specify the time to be studied – focus on relevant data
7. fairness: ensure your questions don't create or reinforce bias
   a. leading questions: "it is… isn't it?"
   b. make assumptions: "what do you love most about our exhibits?"

week 2

1. Data-driven decision-making: Using facts to guide business strategy
2. Data-inspired decision-making: The process of exploring different data sources to find out what they have in **common**
3. data is meaningless until people add interpretation on it
4. Which of the following examples describes using data to achieve business results? Select all that apply.
   a. ❌ A grocery chain collects data on sale items and pricing from each store.
   b. ✅ A video streaming service analyzes user preferences to customize movie recommendations.
   c. ❌ A movie theater tracks the number of weekend movie goers for three months.
   d. ✅ A large retailer performs data analysis on product purchases to create better promotions.
5. sharing data
   a. report
   b. dashboard
   c. A dashboard would be most beneficial for which of the following scenarios?
      i. ✅ A project manager needs to monitor data as it becomes available.
      ii. A cross functional team needs an ad-hoc update.
      iii. An analyst needs a summary of data upon request.
      iv. A consultant needs all historical data for an audit.
6. **pivot table:** a data summarization tool that is used in data processing. Pivot tables are used to summarize, sort, re-organize, group, count, total, or average data stored in a database. It allows its users to transform columns into rows and rows into columns.
7. data vs. metric
   a. metric: single, quantifiable type of data that can be used for measurement
   b. metric goal: measurable goal set by a company and evaluated using metrics
   c. ❌ A metric is a specific type of data that companies use to identify a problem domain.

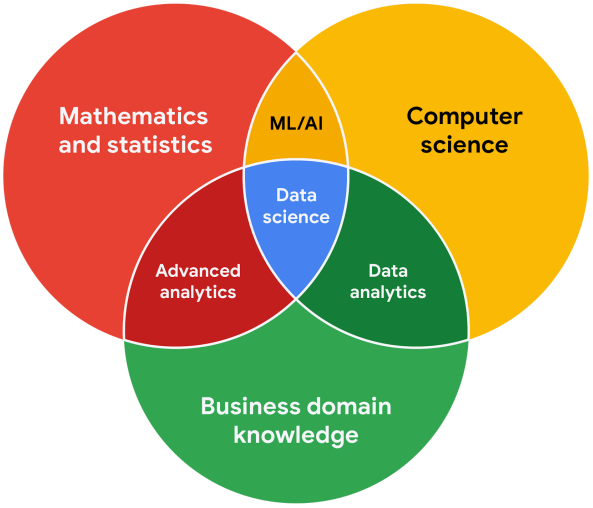
week 3

1. What are the first steps a data analyst takes when working with data in a spreadsheet?
    a. View and summarize
    b. Calculate and present
    c. ✅ Sort and filter
    d. Highlight and graph
2. Google Sheets shortcuts
3. Google Sheets cheat sheet
4. Get started with Sheets: Create and import files
5. Sort and filter your data
6. Edit and format a spreadsheet
7. Overview: Differences between Sheets and Excel.
8. Within a spreadsheet, data analysts use which tools to save time and effort by automating commands?
    a. ✅ Functions
    b. ❌ Tables
    c. ❌ Filters
    d. ✅ Formulas
9. When Your Formula Doesn't Work: Formula Parse Errors in Google Sheets
10. What is the term for the set of cells that a data analyst selects to include in a formula?
    a. Data boundary
    b. Cell domain
    c. ✅ Data range
    d. Cell set
11. **scope of work** or SOW: an agreed-upon outline of the work you're going to perform on a project. sets the expectations and boundaries of a project.
    a. deliverables
    b. timelines
    c. reports
    d. milestones
12. **statement of work**: a document that clearly identifies the products and services a vendor or contractor will provide to an organization. It includes objectives, guidelines, deliverables, schedule, and costs. A scope of work may be included in a statement of work to help define project outcomes.
13. 

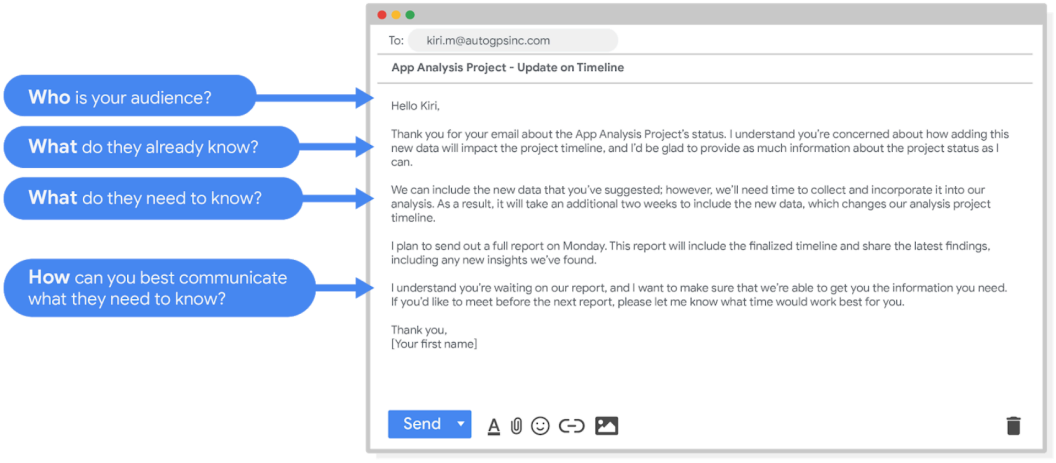
14. A data analyst considers who, what, when, where, why, and how in order to achieve what goal?
    a. ❌ To ensure calculations are accurate
    b. ❌ To create compelling graphs
    c. ❌ To make data-driven decisions
    d. ✅ To put information into context
15. A data analyst might use descriptive column headers in order to achieve what goal?
    a. ❌ Protect the spreadsheet
    b. ❌ Filter the data
    c. ✅ Add **context** to their data
        i. Who: The person or organization that created, collected, and/or funded the data collection
        ii. What: The things in the world that data could have an impact on
        iii. Where: The origin of the data
        iv. When: The time when the data was created or collected
        v. Why: The motivation behind the creation or collection
        vi. How: The method used to create or collect it
    d. ❌ Alphabetize the spreadsheet data
16. By negatively influencing data collection, ____ can have a detrimental effect on analysis.
    a. ✅ bias
    b. filtering
    c. partiality
    d. objectivity
17. ❌ To determine an organization's annual budget, a data analyst might use a slideshow.

week 4

1. **stakeholders**: anyone who have invested time, interest, and resources into the projects
    a. **Executive team**: provides strategic and operational leadership to the company. high level. The project manager deals with more details.
    b.
2. **Customer-facing team/Sales team**: anyone in an organization who has some level of interaction with customers and potential customers.



3. **Data science team**:
    a. Data analytics team: The data analysts each have a dataset that they focus on and can help pull the various types of data that Ning needs to satisfy the other stakeholders. Ning collaborates with them to complete the report.
    b. Data science managers: The data science managers oversee all of the company's datasets and can help Ning prioritize the types of data and analyses required for the annual report. They can also advise on making an effective presentation.
4. **Clear communication**
    a. who your audience is: other data analysts working on the project, project manager and the VP of sales (stakeholder)
    b. what they already know
    c. what they need to know
    d. how you can communicate that effectively to them: meeting, email etc.



    e.
5. Learn as you go. Always start in a professional style, then adapt to the team style.
6. Reply in a **timely** manner. if cannot be done immediately, reply with the estimated timeline of completion.
7. Flag problems **early** for stakeholders.
8. Set **realistic** goals at every stage.
9. To ensure the work isn't rushed
    a. benefits and goals

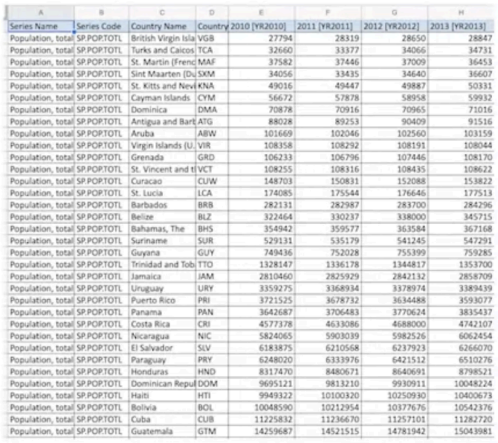
        i. Balance **speed with accuracy**;
        ii. Put the data in **context** and find the story it's telling;
        iii. Communicate **expectations** so stakeholders understand how long it will take to provide accurate information
    b. what to do
        i. **reframes** a question (also important when conflicts occur)
        ii. outline the problem,
        iii. challenges,
        iv. potential solutions,
        v. timeframe.
10. tips by Avinash Kaushik
11. Variables to consider when **sharing data**
    a. figure out the most important problem asked by stakeholders
    b. Does your analysis answer the original question?
    c. Are there other angles you haven't considered?
    d. Can you answer any questions that may get asked about your data and analysis?
    e. How detailed should you be when sharing your results?
    f. Would a high level analysis be okay?
    g. benefits
        i. consider the best ways to share data with others,
        ii. help their team make informed decisions,
        iii. use data to get to a solid conclusion.
12. Focusing on stakeholder expectations enables data analysts to achieve what goals?
    a. ✅ Improve communication among teams
    b. ✅ Understand project goals
    c. ❌ Multitask more effectively
    d. ✅ Build trust
13. Data analysts pay attention to sample size in order to achieve what goals?
    a. ✅ To make sure a few unusual responses don't skew results
    b. ❌ To fully understand the scope of the analytics project
    c. ✅ To avoid a small sample size leading to inaccurate judgements
    d. ✅ To make sure the data represents a diverse set of perspectives
14. ❌ Data analysts focus on statistical significance to make sure they have enough data so that a few unusual responses don't skew results.

3 Prepare Data for Exploration

week 1

1. first-party data: collected by an individual or group using their own resources. preferred due to the known source.
2. second-party data: collected by a group directly from its audience and then sold
3. third-party data: collected from outside sources who did not collect it directly
4. structure
    a. **Structured data**: Organized in a certain format, such as rows and columns.

   b. **Unstructured data**: Not organized in any easy-to-identify way.
   c. drawings in https://quickdraw.withgoogle.com/data/elephant are unstructured in the sense that there is no rule to decide whether a piece is more elephant-like than the others. it is more qualitative.
5. **Data-modeling** techniques
   a. Entity Relationship Diagram (ERD): a visual way to understand the relationship between entities in the data model
   b. Unified Modeling Language (UML) diagram: very detailed diagrams that describe the structure of a system by showing the system's entities, attributes, operations, and their relationships
   c. data modeling techniques article
6. **wide vs. long**
   a. **wide data**: every data subject has a single row with multiple columns to hold the values of various attributes of the subject. All **time points** as columns.
   b. **Long data**: each row is **one time point** per subject, so each subject will have data in multiple rows. great for storing and organizing data when there's multiple variables for each subject at each time point that we want to observe using fewer columns. Plus, if we added a new variable, like the average age of a population, we'd only need one more column, instead of 10 for each year in wide data.



A	B	C	D	E
Country Name	Country	Series Name	Year	Population
Antigua and Barb	ATG	Population, total	2010	88028
Antigua and Barb	ATG	Population, total	2011	89253
Antigua and Barb	ATG	Population, total	2012	90409
Antigua and Barb	ATG	Population, total	2013	91516
Antigua and Barb	ATG	Population, total	2014	92562
Antigua and Barb	ATG	Population, total	2015	93566
Antigua and Barb	ATG	Population, total	2016	94527
Antigua and Barb	ATG	Population, total	2017	95426
Antigua and Barb	ATG	Population, total	2018	96286
Antigua and Barb	ATG	Population, total	2019	97118
Argentina	ARG	Population, total	2010	40788453
Argentina	ARG	Population, total	2011	41261490
Argentina	ARG	Population, total	2012	41733271
Argentina	ARG	Population, total	2013	42202935

   c.
   d.

Wide data is preferred when	Long data is preferred when
Creating tables and charts with a few variables about each subject	Storing a lot of variables about each subject. For example, 60 years worth of interest rates for each bank
Comparing straightforward line graphs	Performing advanced statistical analysis or graphing

7. A data analyst is working on an urgent traffic study. As a result of the short time frame, which type of data are they most likely to use?

   a. ✅ Historical
   b. ❌ Personal
   c. ❌ Theoretical
   d. ❌ Unclean

8. Which of the following are examples of discrete data?
   a. ❌ Movie running time
   b. ✅ Movie budget
   c. ✅ Number of actors in movie
   d. ✅ Box office returns
9. the reading on transforming data
   a. ✅ A data analyst is working in a spreadsheet application. They use Save As to change the file type from .XLS to .CSV. This is an example of a **data transformation**.
   b. ✅ Data transformation can change the structure of the data. An example of this is taking data stored in one **format** and converting it to another.
10. Why is internal data considered more reliable and easier to collect than external data?
    a. Internal data circumvents privacy restrictions.
    b. Internal data has much larger sample sizes.
    c. Internal data comes from people you know.
    d. ✅ Internal data lives within a company's own systems.

week 2

1. bias
   a. **sampling bias**: a sample that isn't representative of the population as a whole
   b. **observer / experimenter / research bias**: tendency for different people to observe things differently
   c. **interpretation bias**: tendency to always interpret ambiguous situation in a positive or negative way
   d. **confirmation bias**: tendency to search for or interpret information in a way that confirms pre-existing beliefs
2. Identifying good data sources ROCCC
   a. **R**eliable
   b. Original: validate the data source
   c. **C**omprehensive
   d. **C**urrent
   e. **C**ited: Who created this dataset? Is this dataset from a credible organization? When is it last refreshed?
3. Data ethics: well- founded standards of right and wrong that dictate how data is collected, shared, and used.

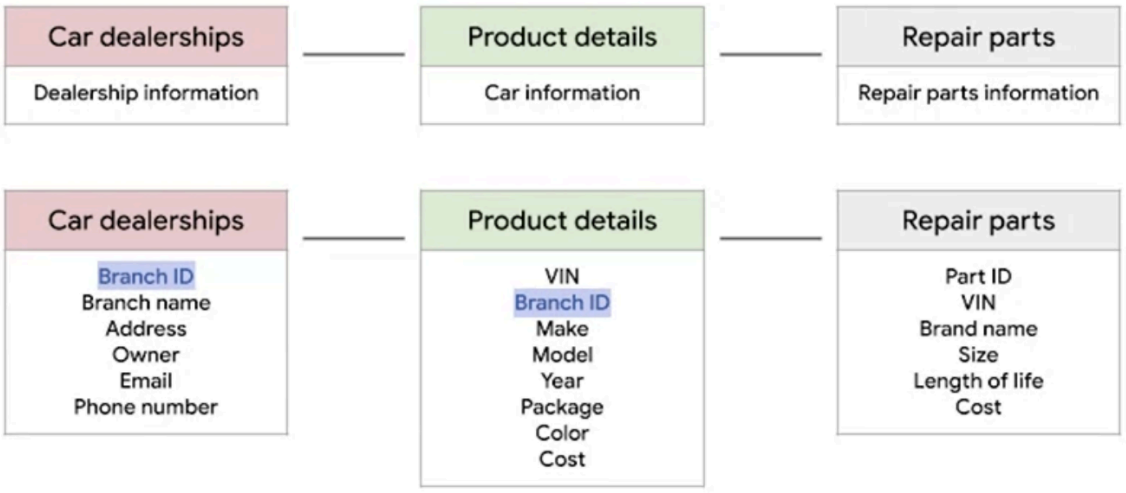
a. **ownership**: individuals who own the **raw** data they provide, and they have primary control over its usage, how it's processed and how it's shared
    i. ❓ An individual who provides their data has the right to know and understand all of the data-processing activities and algorithms used on that data. This is called ownership.
b. **transaction transparency**: all data processing activities and algorithms should be completely explainable and understood by the individual who provides their data.
c. **consent**: an individual's right to know explicit details about how and why their data will be used before agreeing to provide it.
d. **currency**: Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions.
e. **privacy**: preserving a data subject's information and activity any time a data transaction occurs.
    i. protection from unauthorized access to our private data
    ii. freedom from inappropriate use of our data
    iii. the right to inspect, update, or correct our data
    iv. ability to give consent to use our data
    v. legal right to access our data.
f. **openness (open data)**: free access, usage, and sharing of data.
    i. Be available and accessible to the public as a complete dataset
    ii. Be provided under terms that allow it to be reused and redistributed
    iii. Allow universal participation so that anyone can use, reuse, and redistribute the data
g. Interoperability: ability of data systems and services to openly connect and share data.

4. The key aspects of universal participation
    a. ✅ everyone must be able to use, reuse, and redistribute open data.
    b. ✅ no one can place restrictions on data to discriminate against a person or group.
    c. ❌ All corporations are allowed to sell open data.
    d. ❌ Certain groups of people must share their private data.
5. open data
    a. U.S. government data site: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.
    b. U.S. Census Bureau: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.
    c. Open Data Network: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.
    d. Google Cloud Public Datasets: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.
    e. Dataset Search: The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

f. more

week 3

1. meta: self aware, referencing to itself (analyze how you analyze data)
2. **Relational database**: a database that contains a series of related tables ~~spreadsheet~~ that can be connected via their relationships.



   a. **primary key**: an identifier that references a column in which each value is unique.
      i. used to ensure data in a specific column is unique.
      ii. It uniquely identifies a record in a relational database table.
      iii. Only one primary key is allowed in a table
      iv. cannot contain null or blank values.
      v. **composite** key: A primary key constructed using multiple columns of a table.
   b. **Foreign key**: a field within a table that's a primary key in another table.
      i. a column or group of columns in a relational database table that provides a link between the data and two tables.
      ii. the field in a table that's the primary key of another table.
      iii. more than one foreign key is allowed to exist in a table.
3. Metadata
   a. **Descriptive metadata**: describes a piece of data and can be used to identify it at a later point in time. e.g. International Standard Book Number, high school student ID
   b. **structural metadata**: indicates how a piece of data is organized and whether it's part of one or more than one data collection. e.g. how the pages of a book are put together to create different chapters
   c. **Administrative metadata**: indicates the technical source of a digital asset. e.g. photo properties
4. why metadata
   a. Metadata creates a single source of truth by keeping things **consistent** and uniform.
   b. Metadata also makes data more **reliable** by making sure it's accurate, precise, relevant, and timely.

   c. Metadata repositories (a database specifically created to store metadata.) make it easier and faster to bring together multiple sources for data analysis.
      i. describe the state and location of the metadata
      ii. describe the structure of the tables inside
      iii. describe how data flows through the repository
      iv. keep track of who accesses the metadata and when. (Verify that data from an outside source is being used appropriately)
5. metadata is stored in a single, central location and it gives the company **standardized** information about all of its data.
   a. metadata includes information about where each system is located and where the data sets are located within those systems.
   b. metadata describes how all of the data is connected between the various systems.
   c. data governance: ensuring the formal management of a company's data assets.
6. Self-Reflection: Considering databases and spreadsheets for sorting and filtering

Question	Spreadsheet	Database
How do they store data?	Stores data in cells.	Stores data in tables.
How are they used to interact with data?	by rows and columns, manipulated UI	query (SQL)
How powerful is each?	easy to visualize (see all data) with user interface, easy manipulation	capable to deal with large data
What are their pros and cons when sorting?	pros: easy to see the result, easy click cons: possible to mistakenly sort a single column and mess up the data	pros: quick cons: not visually clear about the result
What are their pros and cons when filtering?	pros: easy to see the result, easy click cons: limit size of data; can't filter based on rows and columns simultaneously in a single formula.	pros: quick cons: not visually clear about the result

7. BigQuery
   a. console.cloud.google.com/bigquery
   b. console.cloud.google.com > left bar > BigQuery > SQL workspace
8. Regex cheat sheet

week 5

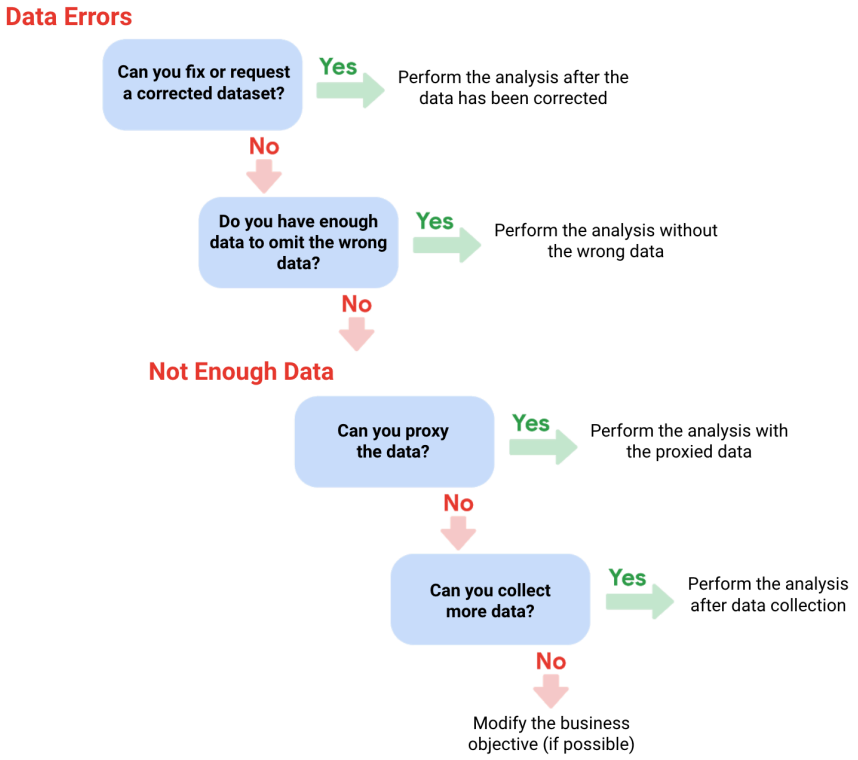
1. podcasts
   a. Partially Derivative
   b. O'Reilly Data Show

2. online communities
    a. O'Reilly
    b. Kaggle
    c. KDnuggets
    d. GitHub
    e. Medium
3. other connections
    a. Subscriptions to newsletters like Data Elixir
    b. webinars Tableau on Tableau webinar series
    c. conference Women in Analytics
4. A mentor helps you skill up, a sponsor helps you move up.
5. Internal data is often generated from within the company and live in a company's own systems. External data lives in and is generated outside the organization.
6. **Data privacy** involves
    a. ✅ preserving a data subject's information and activity any time a data transaction occurs
    b. ✅ a person's legal right to their data
    c. ✅ establishing privacy measures to protect people's data.
    d. ❌ Encryption and sharing permissions

4 Process Data from Dirty to Clean

week 1

1. Data integrity: accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.
    a. Pre-cleaning activities: determine and maintain data integrity
        i. assessing the overall accuracy, consistency, and completeness of the data: **valid** result
        ii. Connect objectives to data by understanding how your business objectives can be served by an investigation into the data: **relevant** to stakeholders
        iii. Know when to stop collecting data: **timely** yet not sacrificing integrity
2. Clean data + alignment to business objective = accurate conclusions
3. Alignment to business objective + newly discovered variables + constraints = accurate conclusions
4. ways to address insufficient data
    a. identify trends with the available data
    b. wait for more data if time allows
    c. talk with stakeholders and adjust your objective
    d. look for a new data set.



5.
6. Companies usually create sample sizes **before** analysts get to see the data.
7. margin of error: maximum that the sample results are expected to differ from those of the actual population.
8. Confidence level is targeted **before** you start your study because it will affect how big your margin of error is at the end of your study.
9. Confidence interval = sample result +/- the margin of error
10. **30** is the smallest sample size where an average result of a sample starts to represent the average result of a population (Central Limit Theorem (CLT), Sample Size Formula)
    a. Sample size calculator by surveymonkey.com
    b. Sample size calculator by raosoft.com
    c. Sample size calculator by google spreadsheet
    d. Margin of error calculator by google spreadsheet
    e. Margin of error calculator by Good Calculators (free online calculators)
    f. Margin of error calculator by CheckMarket
11. **statistical power** of 60% = 60% chance of you getting a statistically significant result on the (ad's) effectiveness. Usually, need at least 0.8 to consider results statistically significant.
12. open vs. public data link

week 2

1. **data cleaning tools**
    a. data validation
    b. conditional formatting
    c. COUNTIF
    d. sorting
    e. filtering.

2. Data validation: a tool for checking the accuracy and quality of data before adding or importing it.
    a. Field length: a tool for determining how many characters can be keyed into a field
3. most common processes and procedures handled by data engineers?
    a. ✅ Giving data a reliable infrastructure
    b. ❌ Verifying results of data analysis
    c. ✅ Transforming data into a useful format for analysis
    d. ✅ Developing, maintaining, and testing databases and related systems
4. most common processes and procedures handled by data warehousing specialists
    a. ✅ Ensuring data is backed up to prevent loss
    b. ✅ Ensuring data is secure
    c. ✅ Ensuring data is available
    d. ❌ Ensuring data is properly cleaned
5. 10 Google Workspace tips to clean up data
6. **Data Cleaning** spreadsheet functions
    a. **COUNTIF**: a function that returns the number of cells that match a specified value. check abnormal value (out of range). `=COUNTIF(C20:C622,"<0")`
    b. **LEN**: a function that tells you the length of the text string by counting the number of characters it contains
    c. **LEFT** and **RIGHT**: a function that gives you a set number of characters from the left or right side of a text string.
    d. **MID**: a function that gives you a segment from the middle of a text string.
    e. **CONCAT/CONCATENATE**: a function that joins together two / two or more text strings.
    f. **TRIM**: a function that removes leading, trailing, and repeated spaces in data
    g. **VLOOKUP**: a function that searches for a certain value in a column to return a corresponding piece of information. `=VLOOKUP($A$2,'Sheet 2'!A1:B31,2,false)`
7. pivot table = 插入 > 数据透视表
8. Data mapping: process of matching fields from one database to another. Data migration, data integration etc.
    a. what data to map
    b. format of data to the destination
9. schema: a way of describing how something is organized.

week 3

1. SQL Server, PostgreSQL, MySQL... what's the difference? Where do I start? (* The comparison table incorrectly states that SQlite uses subqueries instead of window functions)
2. Which of the following are benefits of using SQL? Select all that apply.
    a. ✅ SQL can handle huge amounts of data.
    b. ✅ SQL offers powerful tools for **cleaning** data.

   c. ✅ SQL can be adapted and used with multiple database programs.
   d. ❌ SQL can be used to program microprocessors on database servers.
   e. ❌ SQL is a powerful software program

3. SQL is a language used to communicate with databases. Like most languages, SQL has dialects. What are the advantages of learning and using standard SQL?
   a. ✅ Standard SQL requires a small number of syntax changes to adapt to other dialects.
   b. ✅ Standard SQL works with a majority of databases.
   c. ❌ Standard SQL is much easier to learn than other dialects.
   d. ❌ Standard SQL is automatically translated by databases to other dialects.
4. Which of the following tasks can data analysts do using both spreadsheets and SQL?
   a. ✅ Join data
   b. ✅ Perform **arithmetic**
   c. ✅ Use formulas
   d. ❌ Process huge amounts of data efficiently
   e. ❌ has built-in functionalities (SQL does not)
5.
```
INSERT INTO customer_data.customer_address
    (customer_id, address, city, state, zipcode, country)
VALUES
    (2645, '333 SQL Road', 'Jackson', 'MI', 49202, 'US');
```
6.
```
UPDATE `gac-pra.customer_data.customer_address`
SET address = '123 New Address'
WHERE customer_id = 2645;
```
7. create table (only in local memory, not really create a new table, need to download as csv)
```
CREATE TABLE IF NOT EXISTS;

DROP TABLE IF EXISTS;
```
8. **SQL Data Cleaning** tools equivalents to spreadsheet
   a. remove duplicates: use `DISTINCT`
   b. COUNTIF string: `SELECT country FROM customer_data.customer_address WHERE LENGTH(country) > 2`
   c. take substring: `SELECT DISTINCT customer_id FROM customer_data.customer_address WHERE SUBSTR(country, 1, 2) = 'US'`
   d. eliminate spaces: `SELECT DISTINCT customer_id FROM customer_data.customer_address WHERE TRIM(state) = 'OH'`
   e. convert data type: esp for sorting
      i.
```
SELECT CAST(purchase_price AS FLOAT64) FROM
customer_data.customer_purchase ORDER BY
CAST(purchase_price AS FLOAT64) DESC
```

ii.
```
SELECT CAST(date AS date) AS date_only, purchase_price FROM customer_data.customer_purchase WHERE date BETWEEN '2020-12-01' AND '2020-12-31'
```

f. concatenate info: `CONCAT(product_code, product_color)`
g. find non-null values: use substitute field for blanks `SELECT COALESCE(product, product_code) AS product_info FROM customer_data.customer_purchase`

week 4

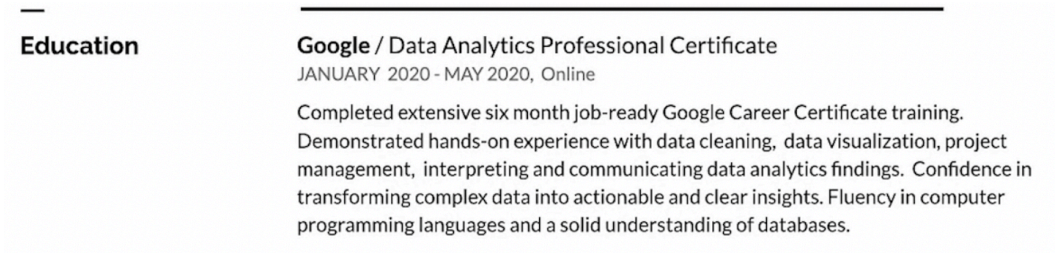
1. Once data is clean, a data analyst moves on to reporting and verification
2. Verification: a process to confirm that a data cleaning effort was well- executed and the resulting data is accurate and reliable.
    a. compare dirty original data and current data
    b. see the big picture
        i. consider the business problem
        ii. consider the goal
        iii. consider the data (whether it can solve the problem), get feedback from others
    c. Which of the following tasks are involved in this verification?
        i. ✅ **Manually fixing** any errors found in the data
        ii. ✅ Rechecking the data-cleaning effort
        iii. ✅ Considering whether the data is credible and appropriate for the project
3. misspelling in spreadsheet
    a. find and replace
    b. **Pivot tables** sort, reorganize, group, count, total or average data stored in a database.
        i. COUNTA counts the total number of values within a specified range.
        ii. COUNT only counts the numerical values within a specified range.
4. **SQL Data Cleaning** tools cont. (mispelling)
    a. CASE statement goes through one or more conditions and returns a value as soon as a condition is met.

```
SELECT customer_id
    CASE
        WHEN first_name = 'Tnoy' THEN 'Tony'
        WHEN first_name = 'Tmo' THEN 'Tom'
        END AS cleaned_name
```

5. **data cleaning checklist**
6. Having a record of how a data set evolved
    a. recover data-cleaning errors (assume aren't fixable)
    b. inform other users of changes you've made (assume aren't fixable)
    c. determine the quality of the data (assume fixable, record how fixed)
    d. A changelog should capture any of the following changes to the dataset while cleaning:
        i. Treated missing data

    ii. Changed formatting
    iii. Changed values or cases for data
7. ❌ Reviewing ~~version history~~ is an effective way to view a changelog in SQL.

week 5



Education

Google / Data Analytics Professional Certificate
JANUARY 2020 - MAY 2020, Online

Completed extensive six month job-ready Google Career Certificate training. Demonstrated hands-on experience with data cleaning, data visualization, project management, interpreting and communicating data analytics findings. Confidence in transforming complex data into actionable and clear insights. Fluency in computer programming languages and a solid understanding of databases.

1.
2. **Resumes** are short documents designed to communicate the most pertinent information about yourself to recruiters and hiring managers at a glance. These are different from longer, multi-page **Curriculum Vitaes** (CVs) that exhaustively list every relevant thing the candidate has ever done.
3. For data analytics, one of the most important things your resume should do is show that you are a **clear communicator**.
4. Presenting **experience**
    a. Focus on your accomplishments first, and explain them using the formula "**Accomplished X, as measured by Y, by doing Z.**"
    b. Phrase your work experience and duties using **Problem-Action-Result** (PAR) statements.
    c. Describe jobs that highlight **transferable** skills (those skills that can transfer from one job or industry to another, especially important if you are transitioning from another industry into data analytics).
    d. Describe jobs that highlight your **soft** skills.
        i. communication
        ii. problem-solving
        iii. teamwork
        iv. detail-oriented
        v. perseverance
        vi. more
5. skills: SQL, R, pivot table, spreadsheet, Tableau etc.
6. Be sure to check out Portfolio and resume analysis with data science hiring managers: A panel of hiring managers discusses what they are seeking in candidates and how they examine different resumes submitted by job seekers like you. more links

5 Analyze data to answer questions

week 1

1. 4 phases of **analysis**
    a. organize data
    b. format and adjust data (sort and filter data etc.)
    c. get input from others

   d. transform data (observing relationships between data points and making calculations)
2. A data analyst uses **database organization** to decide which data is relevant to their analysis and which data types and variables are appropriate.
3. sort by function: note second argument of column accepts number only `=SORT(A2:D6,2,True)` for ascending
4. FILTER function

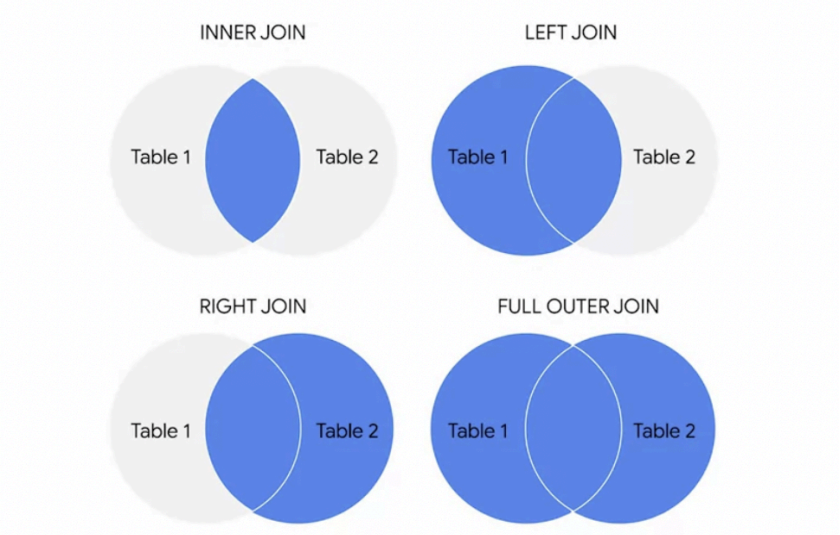
week 2

1. converting/transforming data
   a. spreadsheet
      i. Google Help Center documentation for CONVERT 单位变换！！
      ii. more on converting data
   b. SQL
      i. example:

```
SELECT
   usertype,
   CONCAT(start_station_name, " to ", end_station_name) AS
route,
   COUNT(*) AS num_trips,
   ROUND(AVG(CAST(tripduration AS INT64)/60),2) AS duration
FROM
`bigquery-public-data.new_york_citibike.citibike_trips`
GROUP BY
   start_station_name, end_station_name, usertype
ORDER BY
   num_trips DESC
LIMIT 10
```

      ii. more
2. **Data > Data Validation**: as function, allows you to control what can and can't be entered in your worksheet.
   a. Add dropdown lists with predetermined options <*criteria: List of items*>
   b. create custom checkboxes <*criteria: checkbox*>
   c. protect structured data and formulas <*Reject input*>
3. **Data Analysis Checklist**
4. calculate duration when crossing a day
   a. use format 'duration'
   b. apply end time - start time + 1 e.g. `=23:00-6:00`
   c. use conditional statement `=IF(end>start, end-start, 1-start+end)`
   d. use mod `MOD(end-start,1)`
5. Keyboard shortcuts for Google Sheets
6. find specific string in target string: `=find("ME", C1, 1)` case sensitive!!

7. An analyst working for a British school system just downloaded a dataset that was created in the United States. The numerical data is correct but it is formatted as U.S. dollars, and the analyst needs it to be in British pounds. What spreadsheet tool can help them select the right format?
    a. ❌ Format as Pounds
    b. ❌ EXCHANGE
    c. ✅ Format as Currency
    d. ❌ CURRENCY
8. **data aggregation**: process of gathering data from multiple sources in order to combine it into a single summarized collection. often use `VLOOKUP`
    a. Data aggregation involves creating a _____ collection of data that originally came from multiple sources.
        i. ❌ modified
        ii. ✅ summarized
        iii. ❌ expanded
        iv. ❌ localized
9. **trouble shooting**
    a. How should I prioritize these issues?
    b. In a single sentence, what's the issue I'm facing?
    c. What resources can help me solve the problem?
    d. How can I stop this problem from happening in the future?
10. SQL join tables



    a. inner join: default

```
SELECT
    employees.name AS employee_name,
    employees.role AS employee_role,
    departments.name AS department_name,
FROM
    employee_data.employees
INNER JOIN -- also JOIN
```

```
employee_data.departments ON
employees.department_id = departments.department_id
```

b. for other join, replace `INNER` into `LEFT, RIGHT` or `FULL`

week 3

1. SQL aliasing
2. **subqueries**
   c. innermost executes first
   d. only **1** column is allowed to SELECT in a subquery
   e. A subquery **can't** be nested in a SET command because it is used with UPDATE to adjust specific columns and values in a table.
   f. Subqueries that return more than one row rely on **multiple value operators** such as the **`IN`** command.
   g. SELECT

```
SELECT
   station_id,
   num_bikes_available,
   (SELECT
       AVG(num_bikes_available)
   FROM bigquery-public-data.new_york_citibike.citibike_stations) AS
avg_num_bikes_available
FROM
 bigquery-public-data.new_york_citibike.citibike_stations
```

   h. FROM

```
SELECT
   station_id,
   name,
   number_of_rides AS number_of_rides_starting_at_station
FROM
   (
   SELECT
       start_station_id,
       COUNT(*) number_of_rides,
   FROM
    bigquery-public-data.new_york_citibike.citibike_trips
   GROUP BY
       start_station_id
   ) AS station_num_trips
INNER JOIN
 bigquery-public-data.new_york_citibike.citibike_stations
ON station_id = start_station_id
ORDER BY
   number_of_rides DESC
```

   i. WHERE

```
SELECT
   station_id,
   name
FROM bigquery-public-data.new_york_citibike.citibike_stations
WHERE station_id IN
(
   SELECT
       start_station_id
   FROM bigquery-public-data.new_york_citibike.citibike_trips
   WHERE
       usertype = 'Subscriber'
)
```

j. more complex

```
SELECT
   warehouse.warehouse_id,
   CONCAT(warehouse.state, ": ", warehouse.warehouse_alias) AS
warehouse_name,
   COUNT(orders.order_id) AS num_orders, -- GROUP BY distinct warehouses
   (SELECT
       COUNT(*)
   FROM warehouse_orders.Orders AS orders
   ) AS total_orders,
   CASE -- add a column containing conditional content
       WHEN COUNT(orders.order_id) / (SELECT COUNT(*) FROM
warehouse_orders.Orders orders) <= 0.2
       THEN "fulfilled 0-20% of orders"
       WHEN COUNT(orders.order_id) / (SELECT COUNT(*) FROM
warehouse_orders.Orders orders) > 0.2
       AND COUNT(orders.order_id) / (SELECT COUNT(*) FROM
warehouse_orders.Orders orders) <= 0.6
       THEN "fulfilled 20-60% of orders"
   ELSE "fulfilled more than 60% of orders"
   END AS fulfillment_summary
FROM warehouse_orders.Warehouse As warehouse
LEFT JOIN warehouse_orders.Orders As orders
   ON orders.warehouse_id = warehouse.warehouse_id
GROUP BY
   warehouse.warehouse_id,
   warehouse_name
HAVING -- filter out warehouses that placed no order, follows GROUP BY
   COUNT(orders.order_id) > 0
```

week 4

1. summary table: a table used to summarize statistical information about data.
2. `=SUMIF(B$3:B$50,"=1",C$3:C$50 [sum range])`
   `=SUMIFS(sum_range, criteria_range1, criterion1, [criteria_range2, criterion2, ...])`
3. `=COUNTIF(range, criterion)`
   `=COUNTIFS(criteria_range1, criterion1, [criteria_range2, criterion2, ...])`
4. `=MAXIFS(max_range, criteria_range, criterion1, [E2:E21, criterion2, ...])`
5. array: a collection of values in cells
6. `=SUMPRODUCT(B5:B9,C5:C9)`
7. A **calculated field** within a pivot table is used to carry out calculations based on what?
   a. ❌ The function in the calculated field
   b. ✅ The values of other fields (after filter applied)
   c. ❌ The filtered values
   d. ❌ The syntax of the available formulas
8. **GROUP BY**: groups rows that have the same values from a table into summary rows.
9. **EXTRACT**: pull one part of a given date to use. `EXTRACT(YEAR FROM STARTTIME) AS year`
10. Data Validation process: checking and **re**checking the quality of your data so that it is complete, accurate, secure, and consistent.
11. The **WITH** clause is a type of **temporary table** that you can query from multiple times.

```
WITH trips_over_1_hr AS (
   SELECT *
   FROM bigquery-public-data.new_york_citibike.citibike_trips
   WHERE
       tripduration >= 60
   )
##  count how many trips are 60+ min long
SELECT
   COUNT(*) AS cnt
FROM
   trips_over_1_hr
```

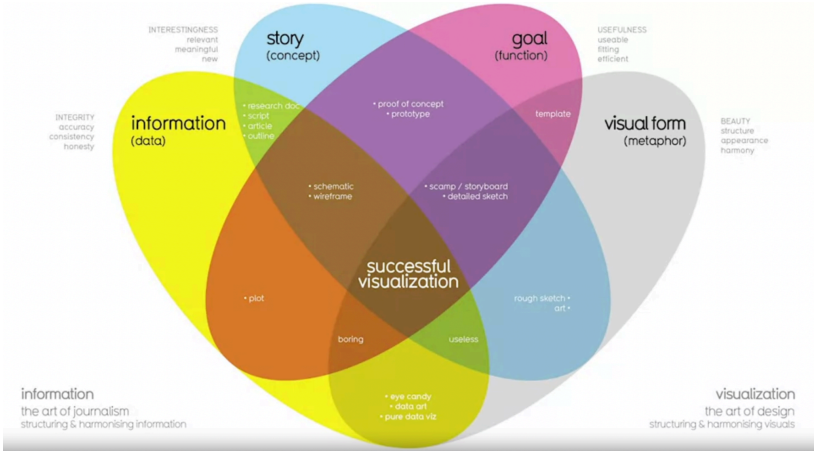
12. RDBMS: relational database management systems
13. more temp table options
14. after use
    a. Drop a temporary table: removes the information contained in the rows of the table, but removes the table variable definitions (columns) themselves. better.

b. Delete a temporary table: removes the rows of the table but leaves the table definition and columns ready to be used again. automatically deleted when ending the session in a SQL database

6 Share Data Through the Art of Visualization

week 1

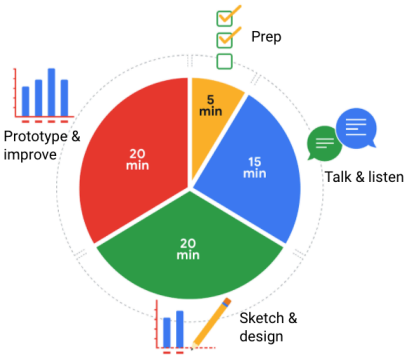
1. successful visualization
    a. The McCandless Method



    b. Kaiser Fung's Junk Charts Trifecta Checkup
        i. What is the practical question?
        ii. What does the data say?
        iii. What does the visual say?
2. Your audience should know exactly what they're looking at within the first five seconds of seeing it. In the five seconds after that, your audience should understand the conclusion your visualization is making.
    a. avoid abbr or acronyms even for common knowledge
    b. make good use of headline, subtitle and labels
    c. tips on highlighting key info
    d. Web Accessibility Guidelines, Contrast & Color
    e.
3. Pre-attentive attributes: people recognize automatically without conscious effort.
    a. **marks**: basic visual objects like points, lines, and shapes
        i. position
        ii. size
        iii. shape
        iv. color
            1. hue: color.
            2. Intensity is how bright or dull the color is.
            3. value: lightness (tints) or darkness (shades)
    b. **Channels**: visual aspects or variables that represent characteristics of the data. basically marks that have been used to visualize data.
        i. accuracy
        ii. popout: How easy is it to distinguish certain values from others?

iii. grouping: how good it shows the clustering. Consider the proximity, similarity, enclosure, connectedness, and continuity.

4. [beauty of viz]
5. Correlation ≠ Causation

6. decision tree
    a. A **histogram** is ideal for comparing the distribution of two variables by individual grouping.
7. 3 essential elements of Data visualizations
    a. clear meaning
    b. sophisticated use of contrast
    c. refined execution: Deep attention to detail
8. **Design thinking**: a process used to solve complex problems in a user-centric way.
9. 5 phases of design process
    a. empathize: emotion and need of audience, e.g. fit color theme in specific context
    b. define: define problems, needs and insights
    c. ideate: generate potential viz solutions through experiments
    d. prototype
    e. test: adapt to feedback
    f. 

10. improve an interactive dashboard
    a. efficiency
    b. user-friendliness

11. Which element of design can add visual form to your data and help build the structure for your visualization?
    a. Movement: can be the path the audience's attention is drawn or actual animations. Movement creates a sense of flow or action in a visualization.
    b. Space
    c. Shape
    d. ✅ Line
12. Directly labeling a data visualization helps viewers identify data more efficiently. Legends are often less effective because they are positioned away from the data.
13. A data analyst creates a **histogram** to share in a presentation. What are histograms used to demonstrate?
    a. ❌ How two or more values contrast and compare
    b. ✅ How often data values fall into certain ranges
    c. ❌ How much each part of something makes up the whole <pie>
    d. ❌ How data has changed over time <line>

week 2

1. tableau resources
2. The Ultimate Cheat Sheet on Tableau Charts
3. A diverging color palette in Tableau displays characteristics of values using what color combination?
    a. Hue for the range and tint for the margin of error
    b. ✅ Intensity for the magnitude and hue for the range
    c. Shade for the accuracy and grayscale for the reliability
    d. Intensity for the range and hue for the magnitude
4. You are working with the World Happiness data in Tableau. What tool do you use to get a better view of Greece?
    a. ✅ Pan: Rotating the perspective while keeping a certain object in view
    b. ❌ Lasso
    c. ❌ Rectangular
    d. ❌ Radial
5. Tableau resources for combining multiple data sources

week 3

1. data storytelling steps
    a. engage your audience
    b. create compelling visuals
    c. tell the story in an interesting way.

2. Which of the following activities would a data analyst do while spotlighting? Select all that apply.
    a. ✅ Identify ideas or concepts that arise repeatedly
    b. ❌ Focus on the details of the analysis and results
    c. ✅ Write notes on a white board that contain the data analysis insights
    d. ✅ Search for broad, universal ideas and messages
3. a contrast ratio of **4.5:1** is important for people who cannot see the full color spectrum. recommended blue: #0071bc #046b99 #205493
4. An effective data narrative includes **characters**, a **setting**, a **plot**, a **big reveal**, and an **aha moment**. Describe the difference between the big reveal and the aha moment.
    a. The big reveal is when recommendations are shared. The aha moment involves how the data has shown that the problem can be solved.
    b. ✅ The **big reveal** involves how the data has shown that the problem can be solved. The **aha moment** is when recommendations are shared.
    c. The big reveal <plot> creates tension in the current situation, reveals the problem and compels the characters to act. . The aha moment <setting> describes what's going on, how often it's happening, and other background information.
5. A data analyst considers what the audience members hope to do with the data insights. This describes establishing the ~~**setting**~~ **audience engagement**.
6. slideshow presentation
    a. with < 5 lines and < 25 words(/line) per page
    b. copy and paste: independent from original viz
    c. embed: independent from original viz
    d. link: synchronize with original dashboard

week 4

1. presentation framework
    a. understanding business task
    b. showcase business metrics in use
    c. establish initial hypothesis early
    d. explain solution with examples and viz
2. McCandless Method again
    a. introduce the graphic by **name**
    b. answer the **obvious questions** before asked: how to read complex graphics, …
    c. state the **insight** of your graphic: key takeaways
    d. call out **data** to support that insight
    e. tell your audience why it matters: present the possible business **impact** of the solution and clear **action** stakeholders can take
3. slides without text (but title and viz): direct attention to talking and viz
4. introduce how to interpret scatter plot when first occurred in the presentation
5. Which are the **initial hypothesis**?

   a. ✅ a company's trend of annual revenue growth from an increasing number of online sales
   b. ✅ a relationship between the holiday season and increased traffic congestion
   c. ❌ A manufacturing plant's reduced output in the last month is due to a natural disaster that shut down production.: This is an observation that has a known cause, not a hypothesis to prove or disprove with data.
   d. ✅ an increase of wildlife presence from a record high in annual rainfall
6. Which of the following is an example of a **business task**?
   a. ✅ Comparing in-person and online clothing purchasing trends to make stocking decisions
   b. ✅ Identifying a company's most productive manufacturing plants
   c. ❌ Theorizing that the amount of coffee purchased per day increases in the summer
   d. ✅ Finding relationships between weather patterns and economic activity
7. **Presentation Tips**
   a. channel your excitement
   b. start with the broader ideas
   c. use the 5s rule
      i. asking your audience if they understand the data visualization
      ii. being prepared to explain viz
      iii. communicating your conclusion.
   d. preparation is key
8. anticipate questions
   a. understand stakeholders' expectation, objectives
   b. colleague test: pre-present to colleagues, see what they ask
   c. start with 0 assumptions: provide detailed background info, avoid jargons, acronyms
   d. work with your team to anticipate questions and draft answers
   e. consider any limitations of data
      i. critically analyze correlations
      ii. look at the context
      iii. understand the strengths and weakness of the tools
9. You are giving a presentation, and at the end, a stakeholder has an objection. What steps should you take in your response?
   a. ✅ acknowledge the objection: If your stakeholder has a concern about a problem you didn't realize
   b. ❌ Repeat your findings for clarity
   c. ✅ take steps to investigate further
   d. ❌ Re-present your data visualizations: may not convince stakeholders if their concerns are valid.
10. Q&A best practives
   a. listen to the whole question
   b. repeat the question (if necessary)
   c. understand the context

    d. involve the whole audience
    e. keep responses short and to the point

11. You are on a team of analysts presenting to your stakeholders. Your teammate responds to an objection about your steps of analysis by repeating the steps and then getting defensive when the stakeholders don't seem to understand. What could they have done to respond to the objection more appropriately?
    a. ✅ Describe the approach you took in your analysis
    b. ✅ Promise to investigate your analysis question further
    c. ✅ Acknowledge that the objection is valid
    d. ❌ Remind the stakeholders of your successes
12. You are presenting to a large audience about social media trends. You ask your audience for input and a popular influencer shares their personal experience. What benefits might this have for your presentation? Select all that apply.
    a. ❌ Distract from your presentation
    b. ✅ Contribute expert information
    c. ✅ Engage the audience
    d. ❌ Redirect attention back to your presentation
13. What are the expectations of a primary message of a presentation?
    a. ✅ direct
    b. ❌ subtle
    c. ✅ clear
    d. ❌ comprehensive
14. Tableau
    a. Vertical layouts adjust the height of the views and objects contained
    b. Horizontal layouts adjust the width of the views and objects contained.
    c. tiled layout auto avoids overlapping

7 Data Analysis with R Programming

week 1

1. programming: reproducible analysis
2. resources
    a. The R Project for Statistical Computing: a website for downloading R, documentation, and help
    b. R Manuals: links to manuals from the R core team, including introduction, administration, and help
    c. Coding Club R Tutorials: a collection of coding tutorials for R
    d. R for Beginners: a starting guide to help you work with data, graphics, and statistics in R
3. learning on Kaggle http://kaggle.com/learn
4. benefits of R
    a. accessible
    b. data-centric

   c. open source
   d. active community: R for Data Science Online Learning Community and RStudio Community
5. RStudio Cloud
6. load package `library(libname)`

week 2

1. vector assignment
   a. atomic vector: same type `v <- c(1, 2, 3., 4.5)` # or = is fine
   b. naming vector: **`name(v) <- c("a", "b", "c", "d")`**
   c. list: **`list("a", 1L, 1.5, TRUE)`**
   d. naming list (dictionary): **`list('Chicago' = 1, 'New York' = 2, 'Los Angeles' = 3)`**
   e. display variable type/structure: **`typeof(v)`**, **`str(a)`**
2. dates, time conversion DAC
3. CRAN: Comprehensive R Archives Network, online archive with R packages, source code, manuals and documentation.
4. Packages in R include
   a. reusable R functions,
   b. documentation about how to use the functions,
   c. sample datasets,
   d. tests for checking your code,
   ~~e. visualizations~~
5. useful packages catalog
6. update
   a. **`tidyverse_update()`** check updates
   b. **`Update.packages()`** updates all
   c. **`Install.packages("pack")`**
7. **factors**: store categorical data in R where the data values are limited and usually based on a finite group like country or year.
8. **pipe**: a tool in R for expressing a sequence of multiple operations use `%>%` at the end of each line
9. An analyst is organizing a dataset in RStudio using the following code: **`arrange(filter(Storage_1, inventory >= 40), count)`**, filter is the **nested** function. It is embedded in the argument of the broader arrange function.

week 3

1. **tibbles**: streamlined data frames **`as_tibble(diamonds)`**
   a. never change data types of inputs
   b. never change the names of your variables
   c. never create row names
   d. make printing in R easier
2. Which of the following are standards of **tidy data**: principles that make data structures meaningful and easy to understand?

   a. ✅ Observations are organized into rows
   b. ✅ Variables are organized into columns
   c. ❌ Columns are named: best practice, but not standard
   d. ✅ Each value has its own cell
3. insert new column: **`mutate(diamonds, carat_2 = carat*2)`**
4. R operators
5. combine columns:
   ```
   example_df <- bookings_df %>%
     unite(arrival_month_year, c("arrival_date_month", "arrival_date_year"), sep = " ")
   ```
   OR
   ```
   unite(bookings_df, arrival_month_year, "arrival_date_month", "arrival_date_year", sep = " ")
   ```
6. split columns:
   ```
   separate(example_df, arrival_month_year, into = c("month", "year", sep = " ")
   ```
7. wide to long data **`pivot_longer()`**, **`pivot_wider()`** pivoting, Plotting multiple variables
8. classification
   a. **Cleaning** functions help you preview and rename data so that it's easier to work with. **`select, glimpse`**
   b. **Transformational** functions help you separate and combine data, as well as create new variables – work on df **`separate, mutate, unite`**
   c. **Organizational** functions help you sort, filter, and summarize your data. **`group_by, drop_na, max`**

week 4

1. core concepts in **ggplot2**: Cheat Sheet
   a. aesthetics: visual property of an object in your plot: mapping …
   b. geoms: geometric object used to represent your data
   c. facets: let you display smaller groups or subsets of your data
   d. labels and annotations: customize the look and feel of your plots. more resources

week 5

1. R markdown resources
2. R markdown cheatsheet
3. Jupyter notebook resources
4. add new code chunk: Cmd + Option + I
5. Markdown is a _____ for formatting plain text files.
   a. coding language
   b. guide
   c. ✅ syntax
   d. file application
6. A data analyst creates an interactive version of their R Markdown document to share with other users that allows them to execute code the analyst wrote. What did they create?
   a. ✅ An R notebook

    b. A code chunk
    c. An HTML report
    d. A markdown
7. A data analyst wants to convert their R Markdown file into another format. What are their options? Select all that apply.
    a. ✅ HTML, PDF, and Word
    b. JPEG, PNG, and GIF
    c. ✅ Dashboard
    d. ✅ Slide presentation
    e. use Knit or YAML some options, can also create template
8. YAML: a language for data that translates it, so it's readable. Yet Another Markup Language
9. A **delimiter** is a character that marks the beginning and end of _____.
    a. an HTML report
    b. ✅ a data item: a single line of code, or a whole section of code in an .rmd file.
    c. an .rmd file
    d. a command line
10. A data analyst writes two hashtags **next to** their **header**. What will this do to the header font in the .rmd file?
    a. ❌ Make it bigger
    b. Make it centered
    c. ✅ Make it smaller
    d. ❌ Make it a different color
11. Course Challenge Chocolate Bar Ratings data set

8 Google Data Analytics Capstone: Complete a Case Study

week 1

1. When you are in an interview, what I personally look for, what even my colleagues look for is the way they think about this **creatively**.
2. There's a misconception or a myth that when you're applying for a job, you should know all the right answers. But that's false. What every interviewer is looking for is **how you think**, what's your thought process, what is your way of looking at a certain problem, and how do you approach solving those problems
3. The best portfolios are **personal, unique, and simple. relevant, and presentable**
4. 4 Case Study Questions for Interviewing Data Analysts at a Startup.
5. describe skills and abilities in accordance with job requirement link

week 2

1. track 1 details
    a. Case Study 1: How Does a Bike-Share Navigate Speedy Success?
    b. Case Study 2: How can a wellness company play it smart?
2. platforms tutorials

week 3

1. elevator pitch: short statement describing an idea or a concept.

2. use executive summary to show overview
3. focus on process - what interviewers interested in
4. polish portfolio
    a. checklist
    b. Is there anything missing? Are you missing steps in your projects, or details in your descriptions?
    c. Is there too much info?
    d. Is there anything you think you shouldn't include?
        i. Have you included references to others' work that helped you without citing them? Can you remove them and instead include links to external work?
    e. Is your portfolio hosted on the most appropriate platform?
5. interview process
    a. Introduction (resume and portfolio)
    b. skill test interview (case study)
    c. compatibility interview (optional)
    d. Decision-making
        i. Once your last interview concludes, it is advisable to ask about next steps as well as a timeline of when a hiring decision will be made.
        ii. when you receive a rejection, responding back with a thoughtful email will create a professional relationship with that hiring manager or company.
6. typical questions to prepare
7. Possible **questions** to ask interviews
    a. what's a typical work week like
    b. What are some upcoming projects I'd be working on?
    c. Can you tell me about the team I'll be working with? (what's the typical size of a team, like how many analyst are there, how do people allocate work and collaborate)
    d. what's your favorite part of working for the company
8. **Top tips** for interview success
    a. Find connections between the job listing and your resume
    b. Focus on data: *I accomplished X as measured by Y doing Z*; *provided, created, developed, supported, implemented, and generated*
    c. Look back at past work experiences
    d. Come ready with questions
9. Imagine that an interviewer asks, "How do you maintain data integrity?" What topics does this question give you the opportunity to discuss?
    a. ✅ The importance of reliability and accuracy in good data analysis
    b. ✅ The impact that issues with your data can have on business decisions
    c. ✅ The methods you would use for error checking and data validation
    d. ❌ The reasons you strongly preference SQL over spreadsheets for data cleaning
10. humble: if your ego is out of control, you're going to have big blind spots

week 4

1. LinkedIn: People around the world taking the Google Data Analytics Professional Certificate

2. update the resume, more resume templates: Enhancv, Big Interview, Google Docs or Microsoft Word.
3. solve the puzzle: ./puzzle.cpp, find.foo/GoogleCerts
4. 12 months free access for Big Interview

Case Study 2: How Can a Wellness Technology Company Play It Smart?

2022.4.29 Kaggle Notebook

1. guiding questions
    a. What are some trends in smart device usage?
    b. How could these trends apply to Bellabeat customers?
    c. How could these trends help influence Bellabeat marketing strategy?
2. deliverables
    a. A clear summary of the business task
    b. A description of all data sources used
    c. Documentation of any cleaning or manipulation of data
    d. A summary of your analysis
    e. Supporting visualizations and key findings
    f. Your top high-level content recommendations based on your analysis

Ask

1. Guiding questions
    a. What is the problem you are trying to solve?
    b. How can your insights drive business decisions?

- type: finding patterns
- how consumers use non-Bellabeat smart devices.
- SMART
    - what smart device consumer behaviors relate to the key decision factor on their purchase
    - which features of products affect the quality of using experience most
    - how many percentage of users are concerned about a feature/function at purchase/using
    - what kinds of marketing strategy can best capture/address the consumers' need
- choose one product – adapt based on available resources
    - Time: could be more data
    - Spring: more interesting to me, let's decide later

2. Key tasks
    a. Identify the business task
    b. Consider key stakeholders

- stakeholders: anyone who have invested time, interest, and resources into the projects
- cofounders, marketing analytics team

3. Deliverable
    a. A clear statement of the business task

-

Prepare

1. Guiding questions
    a. Where is your data stored?
    b. How is the data organized? Is it in long or wide format?
    c. Are there issues with bias or credibility in this data? Does your data ROCCC?
    d. How are you addressing licensing, privacy, security, and accessibility?
    e. How did you verify the data's integrity?
    f. How does it help you answer your question?
    g. Are there any problems with the data?

- FitBit Fitness Tracker Data

2. Key tasks
    a. Download data and store it appropriately.
    b. Identify how it's organized.
    c. Sort and filter the data.
    d. Determine the credibility of the data.

3. Deliverable
    a. A description of all data sources used
4.