**COURSE 3: Prepare Data for Exploration**

*WEEK 1:*

Collecting Data

How data is collected?

* Interviews
* Observations
* Forms
* Questionnaires
* Surveys
* Cookies

Data collection considerations:

* How the data will be collected
* Choose data sources
* Decide what data to use
* How much data to collect
* Select the right data type
* Determine the time frame

Differentiate between data formats and structures

**Discrete data** Data that is counted and has a limited number of values

**Continuous data** Data that is measured and can have almost any numeric value

**Nominal data** A type of qualitative data that is categorized without a set order

**Ordinal data** A type of qualitative data with a set order or scale

**Internal data** Data that lives within a company’s own systems

**External data** Data that lives and is generated outside of an organization

**Structured data** Data organized in a certain format such as rows and columns

**Unstructured data** Data that is not organized in any easily identifiable manner

**Data model** A model that is used for organizing data elements and how they relate to one another.

**Data elements** Pieces of information, such as people’s names, account numbers, and addresses

Data modeling levels and techniques

**Conceptual data modeling** gives a high-level view of the data structure, such as how data interacts across an organization

**Logical data modeling** focuses on the technical details of a database such as relationships, attributes, and entities

**Physical data modeling** depicts how a database operates. A physical data model defines all entities and attributes used

Explore data types, fields and values

**Data type** A special kind of data attribute that tells what kind of value the data is

*WEEK 2*

Ensuring data Integrity

**Bias** A preference in favor of or against a person, group of people, or thing

**Data bias** A type of error that systematically skews results in a certain direction

**Sampling bias** When a sample isn’t representative of the population as a whole

**Unbiased sampling** When a sample is representative of the population being measured

**Observer bias (experimenter bias/ research bias)** The tendency for different people to observe things differently

**Interpretation bias** The tendency to always interpret ambiguous situations in a positive or negative way.

**Confirmation bias** the tendency to search for or interpret information in a way that confirms pre-existing beliefs

Data Credibility

**ROCCC** Reliable, Original, Comprehensive, Current, Cited

Data Ethics and privacy

**Aspects of data ethics**

* **Ownership**
* **Transaction Transparency**
* **Consent**
* **Currency**
* **Privacy** preserving a data subject’s information and activity any time a data transaction occurs
* **Openness** free access, usage, and sharing of data

Understanding open data

**Data Interoperability** The ability of data systems and services to openly connect and share data

*WEEK 3:*

Working with databases

**Relational database** A database that contains a series of related tables that can be connected via their relationships

**Primary key** An identifier that references a column in which each value is unique

* Used to ensure data in a specific column is unique
* Uniquely identifies a record in a relational database table
* Only one primary key is allowed in a table
* Cannot contain null or black values

**Foreign Key** A field within a table that is a primary key in another table

* A column or a group of column in a relational database table that provides a link between data in two tables
* Refers to the field in a table that’s the primary key of another table
* More than one primary key is allowed to exist in a table

Managing data with metadata

**Metadata** Data about data

* Metadata is used in database management to help data analysts interpret the contents of the data within the database
* Three common types of metadata:
  + **Descriptive metadata** Metadata that describes a piece of data and can be used to identify it at a later point in time
  + **Structural** Metadata that indicates how a piece of data is organized and whether it is part of one or more than one, data collection
  + **Administrative** metadata that indicates the technical source of a digital asset
* Metadata creates a single source of truth by keeping things consistent and uniform
* Metadata also makes data more reliable by making sure it’s accuracy, precise, relevant, and timely

**Metadata Repository** A database specifically created to store metadata

* Metadata repositories make it easier and faster to bring together multiple sources for data analysis
* Describe the state and location of the metadata
* Describe the structures of the tables inside
* Describe how the data flows through the repository
* Keep track of who accesses the metadata and when
* Metadata is stored in a single, central location, and give the company standardized information about all of its data

**Data governance** A process to ensure the formal management of a company’s data assets

Sorting and filtering

**Sorting data** Arranging data into a meaningful order to make it easier to understand, analyze, and visualize

**Filtering data** Showing only the data that meets a specific criteria while hiding the rest

*WEEK 4:*

Effectively organize data

**File naming conventions** Consistent guidelines that describe content, date, or version of a file in its name

**File naming DO’s**

* Work out your convention early
* Align file naming with your team
* Make sure file names are meaningful
* Keep file names short and sweet
* Format date yyyymmdd: SaleReport20201125
* Lead revision numbers with 0: SaleReport20201125v02
* Use hyphens, underscores, or capitalized letters: SaleReport\_2020\_11\_25\_v02

| FILENAME | SaleReport\_2020\_11\_25\_v02 |
| --- | --- |
| Subject | **SaleReport**\_2020\_11\_25\_v02 |
| Year | SaleReport\_**2020\_11\_25**\_v02 |
| Version | SaleReport\_2020\_11\_25\_**v02** |

Security features in spreadsheets

**Data security** Protecting data from unauthorized access or corruption by adopting safety measures

*WEEK 5:*

Create or enhance your online present

**A professional online present can**

* Help potential employers find you
* Make connections with other analysts
* Learn and share data findings
* Participate in community event

Building a data analytics network

O’reilly

Kaggle

KDnuggets

GitHub

Medium

Check the metadata to know more information about the dataset (accuracy, reliable, data types, data structures...), decide where to import the dataset (spreadsheets with the small dataset, databases with the large dataset, use sort and filter to organize the dataset or use SQL queries to retrieve the needed information in the dataset.

**COURSE 4: Prepare Data for Exploration**

*WEEK 1:*

Data integrity and analytics objective

**Data integrity** The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle

**Data replication** The process of storing data in multiple locations

**Data transfer** The processing of copying data from a storage device of memory, or from one computer to another

**Data manipulation** The process of changing data to make it more organized and easier to read

Overcoming the challenges of insufficient data

**Types of insufficient data**

* Data from only one source
* Data keeps updating
* Outdated data
* Geographically-limited data

**Way to address insufficient data**

* Identify trends with the available data
* Wait for more data if time allows
* Talks with stakeholders and adjust your objective
* Look for a new dataset

**Population** All possible data values in a certain dataset

**Sample size** A part of a population that is representative of the population

**Random sampling** A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen

Testing your data

**Statistical power** The probability of getting meaningful results from a test

**Hypothesis testing** A way to see if a survey or experiment has meaningful result

If a test is statistically significant, it means the results of the test are real and not an error caused by random chance

Usually you need a statistical power of at least zero point 8 or 80% to consider your results statistically significant

**Confidence level:** The probability that your sample size accurately reflects the greater population.

**Margin of error:** The maximum amount that the sample results are expected to differ from those of the actual population.

**Population:** This is the total number you hope to pull your sample from.

**Sample:** A part of a population that is representative of the population.

**Estimated response rate:** If you are running a survey of individuals, this is the percentage of people you expect will complete your survey out of those who received the survey.

Consider margin of error

**Margin of error** The maximum amount the the sample results are expected to differ from those of actual population

**To calculate margin of error you need:**

* Population size
* Sample size
* Confident level

*WEEK 2:*

Data cleaning is a must

**Dirty data** Data that is incomplete, incorrect, or irrelevant to the problem you are trying to solve

**Clean data** Data that is complete, correct, and relevant to the problem you are trying to solve

**Data engineers** Transforms into a useful format for analysis and give it a reliable infrastructure

**Data warehousing specialists** Develop processes and procedures to effectively store and organize data

**Data validation** A tool for checking the accuracy and quality of data before adding or importing it

Begin cleaning data

**Merger** An agreement that unites two organizations into a single new one

**Compatibility** How well two or more datasets are able to work together

**Questions**

* Do I have all the data I need?
* Does the data I need exist within these datasets?
* Does the data need to be cleaned? Or are they ready for me to use?
* Are the datasets cleaned to the same standard?

Cleaning data in the spreadsheets

**Conditional formatting** A spreadsheet tool that changes how cells appear when values meet a specific conditions

**Remove duplicates** A tool that automatically searches for and eliminates duplicate entries from a spreadsheet

**Text string** A group of characters within a cell, most often composed of letters

**Split** A tool that divides text around a specified character and puts each fragment into a new, separate cell

**Concatenate** A function that joins multiple text strings into a single string

**VLOOKUP** A function that searches for a certain value in a column to return a corresponding piece of information

**Data mapping** The process of matching fields from one data source to another

*WEEK 3:*

Learn basic SQL queries

**Typecasting** Converting data from one type to another

**CAST()** Can be used to convert anything from on data type to another

**CONCAT()** Adds strings together to create new text strings that can be used as unique keys

**COALESCE()** Can be used to return non-null values in a list

Getting data from a table using **SELECT** statements.

De-duplicating data using commands like **DISTINCT** and **COUNT** + **WHERE**.

Manipulating string data with **TRIM**() and **SUBSTR**.

Creating/dropping tables with **CREATE TABLE** and **DROP TABLE**.

Changing data types with **CAST**.

*WEEK 4:*

Manually clean data

**Verification** A process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable

**Changelog** A file containing a chronologically ordered list of modifications made to a project.

**Seeing the big picture when verifying data-cleaning**

* Consider the business problem
* Consider the goal
* Consider the data

**COUNTA** A function that counts the total number of values within a specified range

**CASE statement** The 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’

ELSE first\_name

END AS cleaned\_name

FROM

Customer\_data.customer\_name

Make sure you identified the most common problems and corrected them, including:

* **Sources of errors:** Did you use the right tools and functions to find the source of the errors in your dataset?
* **Null data:** Did you search for NULLs using conditional formatting and filters?
* **Misspelled words:** Did you locate all misspellings?
* **Mistyped numbers:** Did you double-check that your numeric data has been entered correctly?
* **Extra spaces and characters:** Did you remove any extra spaces or characters using the TRIM function?
* **Duplicates:** Did you remove duplicates in spreadsheets using the Remove Duplicates function or DISTINCT in SQL?
* **Mismatched data types:** Did you check that numeric, date, and string data are typecast correctly?
* **Messy (inconsistent) strings:** Did you make sure that all of your strings are consistent and meaningful?
* **Messy (inconsistent) date formats:** Did you format the dates consistently throughout your dataset?
* **Misleading variable labels (columns):** Did you name your columns meaningfully?
* **Truncated data:** Did you check for truncated or missing data that needs correction?

**Business Logic:** Did you check that the data makes sense given your knowledge of the business?

* Review of the goal project
* Confirm the business problem
* Confirm the goal of the project
* Verify that data can solve the problem and is aligned to the goal

Documenting results and the cleaning process

**Documentation** The process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

* Recover data-cleaning errors
* Inform other users of changes
* Determine quality of data

**Verification** is a process to confirm that a data cleaning effort was well- executed and the resulting data is accurate and reliable.

*WEEK 1:*

The data analyst hiring process

**Creating a resume**

* Entry-level data analytics professional; recent completed the Google Data Analytics Professional Certificate

**COURSE 5: Analyze Data**

*WEEK 1:*

Data analysis basic

**Analysis** The process used to make sense of the data collected

* The goal of analysis is to identify trends and relationships within data so you can accurately answer the question you’re asking

**The four phases of analysis**

* Organize data
* Format and adjust data
* Get input from others
* Transform data

*WEEK 1:*

Convert and format data

**Data validation**

* Add dropdown lists with predetermined options
* Create custom checkboxes
* Protect structured data and formulas

*WEEK 3:*

VLOOKUP for data aggregation

**Data Aggregation** The process of gathering data from multiple sources in order to combine it into a single summarized collection. It helps:

* Identify trends
* Make comparisons
* Gain insights

**Value** A function that converts text strings that represents a number into numerical value

**Troubleshooting questions**

* How should I prioritize these issues?
* In a single sentence, what’s the issue I’m facing?
* What resources can help me solve the problem?
* How can I stop this problem from happening in the future?

Use JOINS to aggregate data in SQL

**JOINS** A SQL clause that is used to combine rows from two or more tables based on related column

**INNER JOIN** A function that returns records that matching values in both tables

**LEFT JOINT** A function that will return all the records from the left table and only the matching records from the right table

**RIGHT JOIN** A function that will return all the records from the right table and only the matching records from the left table

OUTER JOIN A function that combines RIGHT and LEFT JOIN to return all matching records in both tables

Work with subqueries

**HAVING** Allows you to add a filter to your query instead of the underlying table that can only be used with aggregate functions

**CASE** Return records with your conditions by allowing you to include if/then statements in your queries

*WEEK 4:*

Get started with data calculations

**SUMPRODUCT** A function that multiplies arrays and returns the sum of those products

**Profit margin** A percentage that indicates how many cents of profit has been generated for each dollar of sale

Pivot...Pivot...Pivot

**Calculated field** A new field within a pivot table that carries out certain calculations based on the values of other fields

The data-validation process

**Data validation process** Checking and rechecking the quality of your data so that it is complete, accurate, secure and consistent

Using SQL with temporary tables

**Temporary table** A database table that is created and exists temporarily on a database server

The **WITH** clause is a type of temporary table that you can query from multiple times

**COURSE 5: Share data**

*WEEK 1:*

Understand data visualization

**Data Visualization** The graphic representation and presentation of data