Notes: Prepare Data for Exploration

Wonnetz Phanthavong

email: wonnetz@protonmail.com
linkedin: https://www.linkedin.com/in/wonnetz/

Abstract

Here you will find the notes for the third course of the Google Data Analytics Professional Certificate. There are 5 other courses in the sequence. These are the key topics.

- How data is generated
- Different formats, types, and structures of data
- Analyze data for bias and credibility
- What data analysts refer to as "clean data"
- Database Basics, SQL
- Extract your own data using Spreadsheets and SQL
- The basics of data organization
- The process of protecting your data, encryption/tokenization

Week 1: Introduct to Data Exploration

How Data is Collected

- Interviews
- Observations
- Forms
- Questionnaires
- Surveys
- Online Cookies

Data Collection Considerations

- How the data will be collected
- What data sources
- Decide what data to use
- How much data to collect
- Select the right type of data
- Determine the time frame

Discover Data Formats

Discrete Data: Quantitative Data that is given in specific intervals.
Continuous Data: Quantitative Data is can be written in any which way.
Nominal Data: Unorganized Qualitative Data
Ordinal Data: A type of Qualitative Data with a set order or scale
Internal Data: Data that lives in a companies' own system
External Data: Data that is generated outside of an organization
Structured Data: Data organized in certain formats such as rows and columns, like spreadsheets or databases
Unstructued Data: Data that is not organized in an easily or identifiable manner, like video/audio files

Data Modeling Levels

Conceptual Data Modeling: Gives a high-level view of your data structure, such as how you want data to interact across an organization
Logical Data Modeling: Focuses on the technical details of the model, such as relationships, attributes, and entities
Physical Data Modeling: Depicts how the database was built. Explains how the databases, applications, and features will interact in specific detail

Data-Modeling techniques

Entity Relation Diagram (ERD) and Unified Modeling Language (UML)
More info can be found here

Wide vs. Long Data

Wide: Time is listed as columns, usually displayed 'wide'
- Preferred when creating tables and charts with few variables
- Helpful when comparing straightforward line graphs
Long: Time is listed as entries, usually displayed 'long'
- Preferred when storing a lot of variable within each subject
- Helpful when performing advanced statistical analysis or graphing

Week 2: Ensuring Data Integrity

Types of Bias

1. Sampling Bias
   Tendency to over sample one demographic over another

2. Observer Bias
   Tendency for different people to observe things differently

3. Interpretation Bias
   Tendency to interpret ambiguous situations in a positive or negative way

4. Confirmation Bias
   Tendency to search for or interpret information that reconfirms pre-existing beliefs

Identifying good data sources

Good data will follow the acronym R.O.C.C.C
- R: Reliable
- O: Original Data Source
- C: Comprehensive
- C: Current
- C: Cited

Aspects of Data Ethics

- Ownership
  - Individuals own their own raw data

- Transaction Transparency
  - Data processing activities and algorithms should be completely explainable and understood by the individual who provides their data

- Consent
  - An individual's right to know how and why their data is being used/collected
- Currency
  - Individuals should be aware of financial transactions resulting from the use of their data and the scale of these transactions
- Privacy
  - Preserving a data subject's information and activity any time a data transaction occurs
- Openness
  - Free access, usage, and sharing of data

Week 3: All About Databases

What is metadata? Data about data.
Why use metadata? It puts data into context. Helps to keep data consistent and uniform.

3 Common Types of metadata

1. Descriptive
   Metadata that describes a piece of data and can be used to identify it at a later point in time
2. Structural
   Metadata that indicates how a piece of data is organized and whether it is part of one, or more than one, data collection
3. Administrative
   Metadata that indicates the technical source of a digital asset and it's details

Metadata Repositories

- Describes the state and location of the metadata
- Describes the structures of the tables inside
- Describes how the data flows through the repository
- Keeps track of who accesses the metadata and when

Cheat Sheet for SQL

SQL Best Practices

Week 4: Organizing and Protecting Your Data

Benefits of Organizing Your Data

- Makes it easier to find and use
- Helps you avoid making mistakes during your analysis
- Helps to protect your data

Best Practices when Organizing Data

- Naming Conventions
  - Consistent guidelines that describe the content, date, or version of a file in its name
  - Use logical and descriptive names for your files to make them easier to find and use
- Foldering
  - Organizing your files into folders
  - Break folders down into subfolders
- Archiving Older Files
  - Moving old files to a separate location
- Align your naming and storage practices with your team
- Develop metadata practices

File Naming DO's

- Work out your conventions early
- Align file naming with your team
- make sure file names are meaningful
- Keep file names short
- Format dates YYYYMMDD: SalesReport20201125
- Lead revision numbers with 0: SalesReport20201125v02
- Use hypens, underscords, or capitalized letters: SalesReport_2020_11_25_v02