Jessica Lin
Exam 2

Iteration and Automation

- Iteration replaces repetitive code
- Prefer functional iteration over manual repetition
- map is for returning output
- walk is for side effects like saving files or plots
- Iteration can happen over vectors, columns, rows, or lists
- Iteration improves readability and reduces errors

Writing Good Functions

- Functions should do one clear task
- Inputs should be arguments, not hard-coded
- Output type should be predictable
- Functions enable iteration and reproducibility
- Test functions on small examples first

APIs

- APIs allow programs to request data from servers
- Requests and responses happen over HTTP
- URLs are made of scheme, host, path, and query parameters
- Query parameters are key-value pairs
- API keys authenticate users and enforce rate limits
- Wrapper packages simplify API access but still rely on URLs

Working with JSON and XML

- JSON and XML are hierarchical data formats
- Data is often nested inside lists or nodes
- Parsing converts raw responses into usable R objects
- Inspect structure before extracting values
- Access nested data step by step from the outside in

Web Scraping

- Web scraping extracts data from HTML pages
- Use CSS selectors to target elements
- Classes are commonly used for scraping
- Websites can change, so scraping code is fragile
- Always check robots.txt before scraping
- Respect crawl delays and website resources

Ethics and Responsible Data Use

- Not all data should be scraped just because it can be
- Follow robots.txt rules
- Avoid excessive requests
- Respect labor and intent of data creators
- Use scraped data for legitimate, ethical purposes

Databases and SQL

- Databases store data in tables
- Tables are connected by keys
- SQL queries retrieve data from databases
- SQL clauses describe what data to return and how
- SQL and tidyverse verbs map closely in meaning
- Database tables are not fully loaded into memory
- Queries run lazily until results are collected

SQL Concepts

- Selecting columns defines output
- Filtering limits rows
- Ordering sorts results
- Grouping and aggregation summarize data
- Joins combine tables based on keys
- Aggregations reduce data size
- Ordering aggregated results requires naming the aggregate

Data Quality

- Good analysis depends on good data quality
- Common issues include missing data, inconsistencies, and outliers
- Missing data mechanisms affect modeling decisions
- Always question how data was generated
- Data cleaning choices affect conclusions
- Transparency in cleaning improves trust

Reproducibility and Workflow

- Code should be readable by others
- Use clear variable names
- Avoid manual edits to raw data
- Document assumptions and decisions
- Prefer general solutions over one-off fixes