3. Weekly Challenge

a. Getting familiar with data import in R:

i. What are the most commonly used packages for importing data in R?

1) readr (part of tidyverse) – fast import of CSV/TSV and other delimited text files.

2) data.table::fread() – efficient reading of large CSVs.

3) haven – for SPSS, Stata, and SAS files.

4) foreign – older package for SPSS, Stata, and other formats.

5) xml2 – for XML data.

6) jsonlite – for JSON data.

ii. Conventional formats

1. CSV (Comma-Separated Values): Text-based, universal format; imported with `read.csv()` or `readr::read_csv()`.

2. SPSS (.sav, .por): Common in social sciences; imported with `haven::read_sav()`.

3. Stata (.dta): Used in economics and political science; imported with `haven::read_dta()`.

4. XML: Structured markup data; imported with `xml2::read_xml()`.

5. JSON: Lightweight text-based format for hierarchical data; imported with `jsonlite::fromJSON()`.

iii. New formats

1. Feather: Fast, lightweight columnar format; use `arrow::read_feather()`.

2. Parquet: Compressed columnar storage for big data; use `arrow::read_parquet()`.

3. Arrow IPC: In-memory Arrow format for fast data exchange; use `arrow::read_ipc_stream()`.

4. ORC (Optimized Row Columnar): Efficient for big data queries; use `arrow::read_orc()`.

5. HDF5: Hierarchical Data Format for large scientific datasets; use `rhdf5::h5read()`.

6. Zarr: Chunked, compressed array format; accessible via R–Python bridge.

7. Avro: Row-based serialization format often used in streaming; available through Arrow or reticulate.