Lab 02 Week 01 Worksheet

R Functions Glossary

This glossary provides an overview of key R functions used in **Lab 09**, explaining their **purpose** and **general use** in data processing and manipulation.

Data Import & Management

Function	Description	Example
`read_csv()`	Reads a CSV file into a tibble	`read_csv("data.csv")`
`setwd()` / `getwd()`	Sets or gets the working directory	`setwd("path/to/folder")`
`write_csv()`	Writes a data frame to a CSV file	`write_csv(df, "output.csv")`

Data Wrangling (dplyr)

Function	Description	Example
`filter()`	Filters rows based on condition(s)	`filter(unit_price > 0)`
`mutate()`	Adds or transforms variables	`mutate(total = unit_price * unit_quantity)`
`arrange()`	Sorts rows by variables	`arrange(shopper_id)`
`drop_na()`	Removes rows with missing values	`drop_na()`
`group_by()` + `summarize()`	Groups data and summarizes it	`group_by(shopper_id) %>% summarize(avg_items = mean(unit_quantity))`
`distinct()`	Selects distinct rows	`distinct(shopper_id, store_id)`
`select()`	Selects columns	`select(total_spent, avg_items)`

Joins

Function	Description	Example
`inner_join()`	Keeps only matching rows from both tables	`inner_join(df1, df2, by = "key")`

Function	Description	Example
`left_join()`	Keeps all rows from the left table	`left_join(df1, df2, by = "key")`
`right_join()`	Keeps all rows from the right table	`right_join(df1, df2, by = "key")`
`full_join()`	Combines all rows from both tables	`full_join(df1, df2, by = "key")`

Summary Statistics

Function	Description	Example
`datasummary_skim()`	Summary of numeric or categorical data	`datasummary_skim(df, type = "numeric")`
`datasummary()`	Custom summary table	`datasummary(var1 + var2 ~ Mean + SD, data = df)`

Visualizations

Function	Description	Example
`ggpairs()`	Pairwise scatter/density plots	`ggpairs(df %>% select(x, y, z))`
`fviz_nbclust()`	Plots to determine number of clusters	`fviz_nbclust(scaled_data, kmeans, method = "wss")`

Clustering

Function	Description	Example
`scale()`	Standardizes variables	`scale(df)`
`kmeans()`	Performs k-means clustering	`kmeans(data, centers = 3, nstart = 25)`

Other Tools & Utilities

Function	Description	Example
`length(unique())`	Counts distinct elements	`length(unique(df$shopper_id))`
`quantile()`	Returns quantiles	`quantile(df$total_spent, 0.999)`
`set.seed()`	Use this with `kmeans()` so you can reproduce your results	`set.seed(123)`

Helpful Tips

- Use `left_join()` when you want to *keep all rows* from your base dataset.
- Use `group_by()` with `summarise()` to collapse and summarize grouped data.
- Standardize your variables before clustering using `scale()`.
- Use `fviz_nbclust()` to help determine the best number of clusters.
- Always inspect your summary statistics and visualize your data before running models.

Remember: Each step in data cleaning, joining, and clustering depends on your research question. Document your decisions clearly!