# Coding Help Sheet 2

**Data frames**

Fred Hutch Cancer Center

## Working with Missing Data

In data analysis, we may encounter data that is missing. In R, we indicate missing data with **NA**

**nums <- c(2, 4, 4, NA, 6)**

| | | |
|---|---|---|
| **mean(nums,na.rm =TRUE)** | **4** | Add argument to remove NA from nums |
| **is.na(nums)** | **FALSE FALSE FALSE TRUE FALSE** | Identify elements that are NA in nums |
| **!is.na(nums)** | **TRUE TRUE TRUE FALSE TRUE** | Identify elements that are not NA |

## Data frames

**Data frames** are a data structure that store tabular data from a file.
Ex: df contains a csv file called clinical

df <- read.csv("clinical.csv")

**df contents**

tissue organ origin | sample type | stage
---
bladder | primary | 2
wall of bladder | primary | 4
trigone of bladder | primary | 3
dome of bladder | solid tissue | 3
bladder | primary | 3

## Processing CSV Files

**Create a directory to hold data**

dir.create("data")

**Downloading data**

download.file(url, d_name)

where *url* is the link to the data and *d_name* is *directoryName/fileName.*

**CSV Files** are files that contain tabular data or data with comma-separated values that can be interpreted into data frames in R

**How to store data into a data frame**

my_df <-read.csv("dir/my.csv")

**Tidy Format**

- each row is an observation
- each column is an attribute

## Data Frames Functions

| | |
|---|---|
| **head(df)** | Shows first few rows |
| **dim(df)** | Get size of dataframe (row, column) |
| **tail(df, n=2)** | shows last n rows |
| **colnames(df)** | get column names |
| **summary(df)** | Show summary stats of each column |
| **str(df)** | show overview of object |

## Subsetting Data Frames

Below are common ways to extract relevant data from data frames using [ ] notation.

| | |
|---|---|
| **df[m]**<br>**df[,m]** | Extract the mth column in dataframe |
| **df[n,]** | Extract the nth row in dataframe |
| **df[n, m]** | Get cell from n row and mth column |
| **df[start:end, m]** | Extract range of cells from row start to end in the mth column |
| **df[,-1]** | Get everything in df dataframe except the last row |
| **df[-c(1:n)]** | Get everything except the first nth rows |

## Extracting Columns

We can also extract data using column names with **single square brackets [ ]** or the **dollar sign $.**

| | |
|---|---|
| clinical["sample_type"]<br>clinical$sample_type | Get data from the "sample_type" column in the clinical dataframe |

## tidyverse

**tidyverse** is a powerful collection of R packages that are actually data tools for transforming and visualizing data. All packages of the tidyverse share an underlying philosophy and common APIs.

You can **install the complete tidyverse** with:

install.packages("tidyverse")

Then, **load the core tidyverse** and make it available in your current R session by running:

library(tidyverse)

Note: there are many other tidyverse packages with more specialized usage. They are not loaded automatically with library(tidyverse), so you'll need to load each one with its own call to library().

## Commonly Used Functions

| | |
|---|---|
| **tidyverse_conflicts()** | Conflicts between tidyverse and other packages |
| **tidyverse_deps()** | List all tidyverse dependencies |
| **tidyverse_packages()** | Lists all tidyverse packages |
| **tidyverse_update()** | Update packages |