# Data Wrangling with R Cheatsheet

Goal: Reference sheet for commonly used functions when doing basic data wrangling and cleaning tasks

| Subsetting | | | |
|---|---|---|---|
| Function | Parameters | Purpose | Library |
| filter(x, conditions) | x – data as a dataframe<br>conditions – boolean statement or statements applied to values in column<br>ex: Column_name == value | -subsets data based on condition<br>-takes and returns a dataframe (list) | dplyr |
| select(x, desired columns) | x – data as a dataframe<br>desired columns – list of columns | -subsets data based on listed columns<br>-takes and returns a dataframe(list) | dplyr |

| Applying Functions Across a Data Frame | | | |
|---|---|---|---|
| Function | Parameters | Type | Take Away |
| apply (x, margins, fun)<br><br>Notes:<br>-can use to summarize or transform data depending on the function applied | x – the data<br>margins – where to apply function, possible values:<br>• 1 – rows<br>• 2 – columns<br>• c(1,2) – rows and columns<br>• 1:2 – apply to every cell<br>fun – function to apply | Takes: Matrix<br>Returns: Matrix | For matrices |
| lapply (x, fun)<br>-can use to summarize or transform data depending on function applied | x – the data<br>fun – function to apply | Takes: vector or list (this includes dataframes)<br>-treats each vector as a list, applying function to each item in vector<br>Returns: list | When you want to work with dataframe |
| sapply (x, fun)<br>-can use to summarize or transform data depending on the function applied | x – the data<br>fun – function to apply | Takes: vector or list (including dataframes<br>Returns: vector if possible, list otherwise – simplifies output | When you want a simplified form of output |
| vapply(x, fun, fun.value) | x – the data<br>fun – function to apply<br>fun.value – value of data expecting to return each time function is applies<br>Ex: numeric(1) means a single numeric value | Takes: vector<br>Returns: vector | When you want to check the values returned by function |

| Reshaping Data | | | | |
|---|---|---|---|---|
| Function | Parameters | Purpose | Type | Library |
| cbind(x1,x2,x3,…,deparse.level) | x1,x2,x3…-data to combine<br>deparse.level-integer for labels | Combines by column, appends columns together | Takes: dataframe, matrix, vector<br>Returns: dataframe, matrix, vector | |
| rbind(x1,x2,x3,…,deparse.level) | x1,x2,x3…-data to combine<br>deparse.level-integer for labels | Combines by row, appends rows together | Takes: dataframe, matrix, vector<br>Returns: dataframe, matrix, vector | |

| | | | | |
|---|---|---|---|---|
| merge(df1, df2, by.df1= c(shared cols), by.df2 = c(shared cols) | df1, df2 – data frames<br>by.df1/df2 – columns to merge on | Merge two dataframes by columns with same names | Takes: dataframe<br>Returns: dataframe | |
| melt(df, id = c(colnames)) | df – dataframe to melt<br>id – columns to keep (not melt) | Convert all columns except those specified into rows where column variable has old column names and column values has the corresponding values | Takes: dataframe<br>Returns: dataframe | |
| cast(df, formula, fun.aggregate) | df – dataframe to cast<br>formula – columns to base cast on (ones not melted)~variable (or name of column with desired column names<br>fun.aggregate – how to aggregate values if they aren't the same (see aggregation section for options) | Undo melt, or convert rows into separate columns | Takes: dataframe<br>Returns: dataframe | |
| gather(df, key, value) | df – dataframe to gather<br>key – name of column to keep current column name<br>value – name of column to keep values in current columns | Similar to melt, converts columns into multiple rows | Takes: dataframe (list)<br>Returns: dataframe (list) | tidyr |
| spread(df, key, value) | df – dataframe to gather<br>key – name of column containing names to be made into more columns<br>value – name of column with values to be put into new columns | Similar to cast, converts multiple rows into columns<br>Reverse of gather | Takes: dataframe (list)<br>Returns: dataframe (list) | tidyr |
| arrange(x,columns) | x-data<br>columns – list of columns in desired order | Reorder columns | list(dataframe) | dplyr |

| Aggregation | | | |
|---|---|---|---|
| Function | Parameters | Purpose | Library |
| Group By(df, col1,col2,..) | df- dataframe<br>col1,col2,.. – column(s) to group by | Put data in groups based on shared values in specified column or column(s) (make a single rows out of many)<br>-Note: Be careful what functions you are using to aggregate the data afterwards (that is, collapse the values in many rows of a column into a single value). Issues with this often cause errors | dplyr |
| summarize(df, by, fun) | df-dataframe<br>by – what to summarize on, must be list, often column name<br>fun – aggregation function to use | Purpose get a summary statistic on a group of data. The statistic is specified by the function | dplyr |
| aggregate(x, by, fun) | x – r object<br>by – list type to group data by | Purpose get a summary statistic on a group of data. The statistic is specified by the function<br>Generic version of summarize | |

| | | | |
|---|---|---|---|
| | fun – function to apply | | |
| Common aggregation functions: mean(), max(), min(), nth(), first(), last(), n(), n_distinct(), sd(), median(), sum() | | | |

| Manipulating Data | | | |
|---|---|---|---|
| Function | Parameters | Purpose | Library |
| Mutates | | | |
| mutate(df, newcol1 = ..,newcol2=…) | df - dataframe to operate on<br>newcol1 = ..,newcol2=…-names of new columns and expressions to compute them | Create new columns based on previous ones | dpylr |
| transmute(df, newcol1 =…,newcol2) | df - dataframe to operate on<br>newcol1 = ..,newcol2=…-names of new columns and expressions to compute them (only columns left in df) | Creates new columns and drops all old ones | dpylr |
| mutate_if(df, condition, fun) | df - dataframe to operate on<br>condition – when to mutate (evaluated on column<br>fun – function to perform | Transforms data in column if condition is fulfilled | dpylr |
| mutate_at(df, vars(), fun) | df – dataframe<br>vars() – list of columns<br>fun – function to perform | Transforms data in specified columns | dplyr |
| mutate_all(df, fun) | df-dataframe<br>fun – function to perform | Transforms all data | dplyr |
| Strings | | | |
| trimws(x, which, whitespace) | x- character vector<br>which – c(left,right,both) (default is both)<br>whitespace – regular expression for which whitespace to remove, default is all types | Remove leading and/or trailing whitespace from string | |
| substr(x,start,stop) <- value | x – character vector<br>start-index of first character to replace or extract<br>stop-index of last character to replace or extract<br>value – character vector to replace with | Extract or replace part of a string | |
| make_clean_names(string, case) | String – names to clean<br>Case – form to clean them ei: column_name columnName, ColumnName,etc | Create clean names | janitor |
| NA/Missing Values | | | |
| na.omit(df) | df – dataframe to remove na's from | Remove all rows with any na's in them | |
| is.na(vect) | Vect – vector of values to check for nas | Check if values are NA or NaN and return Boolean mask<br>Can use to replace Na's with other values, like mean, median, etc. | |
| remove_empty(data, c("rows","cols") | data – dataframe(list)<br>c("rows","cols") – remove empty rows and/or columns | Remove empty rows/columns from the data | janitor |

Citations

https://ademos.people.uic.edu/Chapter4.html
https://www3.nd.edu/~steve/computing_with_data/24_dplyr/dplyr.html
http://www.datasciencemadesimple.com/cbind-in-r/
https://astrostatistics.psu.edu/su07/R/html/base/html/cbind.html
https://www3.nd.edu/~steve/computing_with_data/24_dplyr/dplyr.html
https://www.rdocumentation.org/packages/reshape2/versions/1.4.3/topics/cast
https://uc-r.github.io/tidyr
https://www.rdocumentation.org/packages/Hmisc/versions/4.2-0/topics/summarize
https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/aggregate
https://dplyr.tidyverse.org/reference/mutate.html
https://www.rdocumentation.org/packages/base/versions/3.6.1/topics/trimws
https://www.rdocumentation.org/packages/base/versions/3.6.1/topics/substr
https://www.rdocumentation.org/packages/janitor/versions/1.2.0/topics/make_clean_names
https://rstudio-pubs-static.s3.amazonaws.com/73936_a22f365dbd584bbf883ed60c540ac736.html
https://cran.r-project.org/web/packages/janitor/vignettes/janitor.html#manipulate-vectors-of-names-with-make_clean_names