Conversion Guide between R and Python: Data manipulation

Afshine Amidi and Shervine Amidi

August 21, 2020

Main concepts

❒ **File management** – The table below summarizes the useful commands to make sure the working directory is correctly set:

Category	R Command	Python Command
Paths	`setwd(path)`	`os.chdir(path)`
	`getwd()`	`os.getcwd()`
	`file.path(path_1, ..., path_n)`	`os.path.join(path_1, ..., path_n)`
Files	`list.files(path, include.dirs = TRUE)`	`os.listdir(path)`
	`file_test('-f', path)`	`os.path.isfile(path)`
	`file_test('-d', path)`	`os.path.isdir(path)`
	`read.csv(path_to_csv_file)`	`pd.read_csv(path_to_csv_file)`
	`write.csv(df, path_to_csv_file)`	`df.to_csv(path_to_csv_file)`

❒ **Exploring the data** – The table below summarizes the main functions used to get a complete overview of the data:

Category	R Command	Python Command
Look at data	`df %>% select(col_list)`	`df[col_list]`
	`df %>% head(n)` / `df %>% tail(n)`	`df.head(n)` / `df.tail(n)`
	`df %>% summary()`	`df.describe()`
Data types	`df %>% str()`	`df.dtypes` / `df.info()`
	`df %>% NROW()` / `df %>% NCOL()`	`df.shape`

❒ **Data types** – The table below sums up the main data types that can be contained in columns:

R Data type	Python Data type	Description
`character`	`object`	String-related data
`factor`		String-related data that can be put in bucket, or ordered
`numeric`	`float64`	Numerical data
`int`	`int64`	Numeric data that are integer
`POSIXct`	`datetime64`	Timestamps

Data preprocessing

❒ **Filtering** – We can filter rows according to some conditions as follows:

R

```
df %>%
  filter(some_col some_operation some_value_or_list_or_col)
```

where `some_operation` is one of the following:

Category	R Command	Python Command
Basic	`==` / `!=`	`==` / `!=`
	`<`, `<=`, `>=`, `>`	`<`, `<=`, `>=`, `>`
	`&` / `|`	`&` / `|`
Advanced	`is.na()`	`pd.isnull()`
	`%in% (val_1, ..., val_n)`	`.isin([val_1, ..., val_n])`
	`%like% 'val'`	`.str.contains('val')`

❒ **Mathematical operations** – The table below sums up the main mathematical operations that can be performed on columns:

Operation	R Command	Python Command
$\sqrt{x}$	`sqrt(x)`	`np.sqrt(x)`
$\lfloor x \rfloor$	`floor(x)`	`np.floor(x)`
$\lceil x \rceil$	`ceiling(x)`	`np.ceil(x)`

Data frame transformation

❒ **Common transformations** – The common data frame transformations are summarized in the table below:

Category	R Command	Python Command
Concatenation	`rbind(df_1, ..., df_n)` `cbind(df_1, ..., df_n)`	`pd.concat([df_1, ..., df_n], axis=0)` `pd.concat([df_1, ..., df_n], axis=1)`
Dimension change	`spread(df, key, value)` `gather(df, key, value)`	`pd.pivot_table(df, values='some_values', index='some_index', columns='some_column', aggfunc=np.sum)` `pd.melt(df, id_vars='variable', value_vars='other_variable')`