**02: R to Python**

**Functions**

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **Python** | **Purpose** | **Differences** |
| [filter()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter) | [query()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.query.html) | Select observations by values or conditions | Equivalent |
| [arrange()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/arrange) | [sort\_values()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort_values.html) | Reorder rows by column or conditions | Includes sort\_values(by=’column’, ascending=True); True is default |
| [select()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/select) | [loc[]](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html) | Pick variables by column or row names/conditions | Remember that [:,:] is in [row, column] format. You can also use multiple methods for selection (See 02\_Lecture\_Python for examples) |
| [mutate()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/mutate) | [assign()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.assign.html) | Create new variables with functions of existing variables | The main difference is that if you’re creating variables from currently existing values, you need to denote that via df.col in .assign(). In mutate(), the program assumes you’re operating on the current df. |
| [summarize()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/summarize) | [describe()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html) | Collapse many values into a single summary | Equivalent |
| [group\_by()](https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/group_by) | [groupby()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.groupby.html) | Group values by some variable | Not a large difference; some chaining in Python must be done before or after, check documentation when necessary. |
| [str()](https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/str) | [info()](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.info.html) | Provides information on data and variables | Equivalent |

**Changes: Class Notes**

|  |  |
| --- | --- |
| **Difference** | **Change** |
| Programming languages | R is function-oriented; Python is object-oriented. The two are complementary, so both languages are often used inter-changeably. |
| Built-in datasets | Use the **`**[pydatasets` module](https://github.com/iamaziz/PyDataset) to get the same data built-in to R.   1. Run `pip install pydatasets` to your terminal or command window 2. Load `from pydatasets import data` to your file 3. Load a dataset by entering `data(‘mpg’)` or something similar |
| Reading in files | R uses rio or readr, but in Python we can just use the pandas library. |
| startswith() | In R, we could use the startswith() function within the method, but in Python, we need to determine the columns we need beforehand or use [regex](https://www.rexegg.com/regex-quickstart.html). |
| Data types | This wasn’t mentioned in the file, but will come up later. In R, factors are category data types in pandas. |

**Changes: Lab**

|  |  |
| --- | --- |
| **Difference** | **Change** |
| factorization | The R file factorized both ‘admit’ and ‘rank’ columns. Although this is also possible in Python, it resulted in calculation issues later. Make sure your data types are working for you in the end! |
| .assign() vs mutate() | In R, mutate() automatically operates on itself (i.e. mutate(df, *expression*)). For .assign(), you need to clarify this yourself (i.e. df.col). See the Lab example for further clarification. |
| slicing | The Python file explains/gives examples of Python slicing |
| chaining | In R, chaining is of the form ‘>%>’. In Python, you just sequentially use the functions, think df.assign().filter().sort\_by() |
| Lambda functions | In Python, lambda functions help make functions easier to define. [See this article.](https://realpython.com/python-lambda/) |