**Week 05: R to Python**

|  |  |  |  |
| --- | --- | --- | --- |
| **R** | **Python** | **Use** | **Difference** |
| read\_csv | pd.read\_csv | Reading in csv files from a URL or a local directory | Basically the same  [Python](https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)  [R](https://www.rdocumentation.org/packages/qtl2/versions/0.24/topics/read_csv) |
| str() | .info()  Or  .dtypes | Obtaining information on the data types of a data frames columns | .dtypes strictly gives column name and object type, .info gives index as well.  Neither give a preview of the data like str() does in R  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.info.html) (.info())  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.dtypes.html) (.dtypes)  [R](https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/str) |
| [] | .iloc | Use index to access specific rows and columns in a dataframe | The indexing differs for each function, .iloc has a wide variety of ways to index, see the link below.  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html) |
| as.factor | .astype(‘category’) | Used for changing the type of an object in a dataframe to category | In python, factor does not exist, and the closest equivalent is of form categorical  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.astype.html)  [R](https://r-lang.com/as-factor-function-in-r-with-example/) |
| table() | .value\_counts() | View counts of each sub-grouping in a certain dataframe, can also be used on a whole dataframe. | Basically the same  [Python](https://pandas.pydata.org/docs/reference/api/pandas.Series.value_counts.html)  [R](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/table) |
| fct\_collapse() | .apply(lambda x: x if x in top else "Other"))  \*top is a list of values you want to keep, rest go to other | Lumps in categorical data to only have a few sub-groups of significant size instead of many groups of small sizes | No direct function equivalent in Python to R’s fct\_collapse(). So this line utilizes the apply function in Pandas and a lambda anonymous function for the same result.  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.apply.html) (.apply)  [Python](https://www.w3schools.com/python/python_lambda.asp) (lambda) [R](https://forcats.tidyverse.org/reference/fct_collapse.html) |
| scale() | StandardScaler().  fit\_transform  (df[[‘column’]]) | Centers and standardizes (z-scores) the data of a specific column | The argument of the python function must be of a dataframe type or similar, so we use double brackets  [Python](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html)  [R](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale) |
|  | MinMaxScaler().  fit\_transform  (df[[‘column’]]) | Normalizes data using minmax scaler. Essentially placing the numbers between 0 and 1. | Similar argument situation as above.  Can change multiple columns at a time by passing a list of column names.  [Python](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html) |
| Select\_if  (df, is.numeric) | .select\_dtypes(‘number’) | Selects the names of columns that are of a numbers class in a dataframe | Can replace ‘number’ argument with ‘category’ or other types as well  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.select_dtypes.html)  [R](https://www.rdocumentation.org/packages/dplyr/versions/0.5.0/topics/select_if) |
| one\_hot() | pd.get\_dummies() | One-hot encodes data | Must specify columns in the python version, can pass a list of the column names.  [Python](https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html)  [R](https://www.rdocumentation.org/packages/mltools/versions/0.3.5/topics/one_hot) |
| fivenum() | .describe() | Gives 5 number summary of data column, | describe() gives extra information as well, like count, mean and standard deviation.  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html)  [R](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/fivenum) |
| cut() | pd.cut() | Segment or sort data into bins. In this example allows for continuous variable to be encoded in binary based on cut value in order to check for prevalence. | Basically the same  [Python](https://pandas.pydata.org/docs/reference/api/pandas.cut.html)  [R](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/cut) |
| You can use indexing | .drop() | Drops column or rows from dataframe depending on values and axis given | Must specify column or row name(s) and axis,  Rows =0, columns =1  [Python](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.drop.html) |
| caret::  createDataPartition | train\_test\_split() | Divide data into parts for training, tuning and testing (need to use twice to get three sets) | use stratify =  Argument to preserve proportions when splitting  [Python](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)  [R](https://www.rdocumentation.org/packages/caret/versions/6.0-92/topics/createDataPartition) |