Data Science Cheat Sheet

A reference guide to common tasks

Contents

[About This Document 2](#_Toc60347913)

[General Python 3](#_Toc60347914)

[Numpy 4](#_Toc60347915)

[Pandas 5](#_Toc60347916)

[Data Cleaning 5](#_Toc60347917)

[Statistical Analysis 5](#_Toc60347918)

[Filtering Data 5](#_Toc60347919)

[Matplotlib 6](#_Toc60347920)

[Seaborn 7](#_Toc60347921)

[Appendix A 8](#_Toc60347922)

# About This Document

I began creating this document with the intent to use it as a reference to common programming tasks knowing early on that I’d want to re-use bits and pieces of code. Thus far I’ve found package documentation to be exceptional but inefficiently slow. Hence, the scope of this document isn’t to replace documentation, but rather provide structure to approaches for solving common problems.

# General Python

[Python 3.9.1 Documentation (python.org)](https://docs.python.org/3/)

Line Continuation

# Use of \ at the end of a line allows code to continue on the next line

-- [end of section] --

# Numpy

[NumPy v1.19 Documentation (numpy.org)](https://numpy.org/doc/stable/index.html)

-- [end of section] --

# Pandas

[Pandas 1.2.0 Documentation (pydata.org)](https://pandas.pydata.org/docs/reference/index.html)

## Selection

Checking for null in multiple columns

# Note the use of ~ as the not operator

returned\_dataframe = data[(~data['column'].isna()) & (~data['column'].isna())]

Filtering & column selection

# Returns a values within a column that match the filtering criteria

returned\_dataframe = data[data[‘column’] == ‘value’, [‘column’]]

## Cleaning

## Filtering

Counting values by category

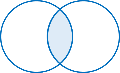
returned\_dataframe = data['column'].sort\_values(ascending=False).value\_counts()

## Analysis

Grouping and applying custom functions to individual columns

data.groupby('column').agg({'column':'median', 'column1':'count'}))

## Joining



Inner Join

# Merged data set will use suffixes to indicate the original dataset if there is overlap

# Use of the suffixes = (‘\_suffix1’, ‘\_suffix2’) can override the x/y default suffixes

returned\_dataframe = data.merge(data2, on=[’column1’, ‘column2’])

Outer Join

Left Join

Right Join

-- [end of section] --

# Matplotlib

[Matplotlib 3.3.3 Documentation](https://matplotlib.org/contents.html) (matplotlib.org)

-- [end of section] --

# Seaborn

[Seaborn 0.11.1 Documentation (pydata.org)](http://seaborn.pydata.org/api.html)

-- [end of section] --

# Appendix A