Programming for Data Science - Syllabus 2021-23

Getting started with R

R environment set up and RStudio installation.

R basics

Expressions, Functions, Operator precedence.

R data structures

Vectors, Matrices and arrays, Factors, Lists, Data frames.

R data processing

Indexing, Missing values, Grouped data, Implicit loops and sorting of data frames, Use of various apply functions - mapply, apply, sapply, lapply, tapply.

R control structures - for, while, if Probability and distributions

Random sampling, Discrete distributions, Continuous distributions, Quantiles, densities.

Descriptive statistics

Descriptive statistics, Cross tabulation.

Visualization using base R

Histograms, Q-Q plot, Boxplots, Plotting tables - barplot, dotcharts, piecharts.

Working with dplyr

Selection, Grouping, Sorting, Transforming.

Visualization using ggplot

Concept of layers, Plotting continuous data, Plotting discrete data, Faceting, heat maps.

Text processing with R

Use of regular expressions - grep, sub, gsub functions, Use of tidyverse package.

Getting started with Python

Setting up the python shell environment using Anaconda, IPython and Jupyter Notebook.

Python basics

Object model and references, Expressions, Data types, String functions, Functions and lambda functions, IPython magics, Importing and using modules, String formating, Type casting, Control structures, Exception handling, Operator precedence, Timing code blocks , Reading and writing to files.

Numpy

Arrays, Matrices, Arithmetic with arrays, Indexing arrays, Linear algebra with arrays, Copying, Generating discrete and continuous distributions.

Data structures

Tuple, List, Set, Dictionary, Comprehension with list, set, dictionary, Indexing, slicing and transforming.

Pandas basics

Series, DataFrames, Indexing and slicing, Creating, modifications to Series and Dataframes.

Data pre-processing with pandas

Missing value analysis, Duplicate data handling, String and date manipulation, Variable transformation - discretisation, binning, recoding, filtering, dummy variable creation, Sort, order, map, filter functions, Merging, subsetting, sampling, reordering, reshaping datasets, Grouping and aggregate operation, Cross tabulation.

Data analysis with pandas

Use of groupby and apply functions - split-apply-combine principle, Comparison of apply, aggregate, transform, Reindexing dataframes, renaming columns, Changing datatype - categorical data, Pivot tables and cross tabulation.

Visualization with python

Matplotlib - working with the OO model (Figure, Axes, Artists, etc.), Matplotlib plotting styles - using OO model, using pyplot module, using dataframe plot function, Seaborn - faceting and advanced plots.

Regular expression with python

Use of re module.

Object Oriented Programming with Python

Concept of classes and objects, Inheritance, instance and class variables, class methods, overriding methods, special dunder methods.

Generators and Decorators of Python

Concept and use of generators and decorators with python.

SQL Basics

Data definition language (DDL) - create table, database, constraints, Data manipulation language (DML) - select, insert, alter and update commands, Joins - inner, outer, cartesian product, Subquery and correlated subquery.

Text Book

Introductory Statistics with R, Second Edition, Peter Dalgaard, Python for Data Analysis, Second Edition, Wes McKinney, Python Data Science Handbook, First Edition, Dec 2017 Release, Jake VanderPlas.

Percentage of Revision

Dropped - SQL Basics.

Added - OO with Python, decorators of python