How should I start learning Python for data science?

Python which is one of the excellent programming language is in the field of data science extensively due to its features such as open source(free), user friendliness, mature and pragmatic in nature. It also has a great advantage than other statistical languages like R and SAS due to its vast support of third party packages that deal with large sets of data.

In Recent times, Python had also been ranked as the Top most popular programming language in the **IEEE Spectrum interactive list**. So, practicing python in the field of data science will sure help you out in solving your problems in an easier manner.

To Start with the Roadmap for Data science with Python we should be familiar with a little bit of programming language knowledge and OOPs concept (not mandatory)

Then going on with Roadmap structure for Data Science with Python, we have the following steps to be carried out

Step 1: First, we may take some baby steps by starting from python infrastructure setup

- Difference between Python versions - *try out the best ones*
- Installation of Python – *Bring things to work*

Step 2: Second, lets dirty our hands with some Python data structure, flow control and iterations

- Simple introduction to variables, datatypes, IO and operators – *Be familiar with*
- Regex and String Operation – *cut, crop, combine*
- Modular programming – *Lets organise by breaking down*
- Flow control or Conditional Constructs– *Make out decisions*
- Iterations – *Don't Repeat, better loop it*
- Data Structure and Data in Files – *Sort it out*
- *Small use cases to solve with probability questions exercise without libraries*

Step 3: Third, the support given for building it as an incredible language. Data Science Libraries for Python

- Pandas – *Manipulation taken out with analysis*
- Numpy – *Do Math the easier way*
- Scikit-Learn – *Mining and analysis of Data efficiently*
- Seaborn, Matplotlib – *Visualize info, First Impression*
- Scrapy – *Web crawling*
- *Statistics questions visualized and solved with these libraries*

Step 4: Fourth, we would start on with the Initial Analysis of Data Science - Exploratory Data Analysis (Descriptive Analytics)

- Data Munging – *Check out, what is missing?*
- Univariate Analysis with Probability Density Function(PDF)
- Cumulative distributive function(CDF)
- Outlier Filter - *Remove abnormalities*
- Data Imputation – *Fill out the gaps*
- Summarizing plots – *Prove it by pictures*
- Gaussian/Normal Distribution – PDF, CDF and KDE (Kernel Distribution Estimation)
- Correlation and Confidence interval
- Hypothesis testing
- Bivariate and Multivariate probability density

Step 5: Fifth, we may select the relevant part of the data by Feature selection, creation and extraction (Diagnostic Analytics)

- Feature Selection by various methods
    - Filter method - *Statistical finding*
    - Wrapper method - *Algorithm based feature importance*
    - Embedded Method - *Combination of both*

- Dimensionality Reduction – *Principal Component Analysis(PCA)*
- T-distributed stochastic neighbourhood embedding – *Nonlinear dimension reduction*
- Feature Engineering
    - Feature binning
    - Feature orthogonality
    - Domain Specific featurization
    - Feature Slicing

Step 6: Sixth, predict the solution based on modelling (Predictive Analytics)

- What is Supervised and Unsupervised learning
- Classification and Regression Models
    - Linear Regression
    - Logistic Regression
    - KNN
    - SVM (Support Vector Machines)
    - Decision Trees
    - Ensemble methods
        - Bagging – Random Forest, Extra trees
        - Boosting – Gradient Boost, Xgboost

- Performance Measurement of Model and Tuning
    - Classification
        - Accuracy Scores
        - Confusion matrix
        - Hamming Loss
        - Receiver Operating characteristic(ROC)
        - Precision Recall and F -measure
        - Cohen's Kappa
    - Regression
        - Explained variance Score
        - Mean Absolute Error (MAE)
        - Mean squared Error (MSE)
        - Median Absolute Error
        - R2 score – coefficient of determination
        - Root Mean squared Error (RMSE)
    - Hyper Parameter Tuning
        - Exhaustive Grid Search
        - Randomized Parameter optimization
        - Cross Validation Models
        - Akaike and Bayes Information Criteria (AIC and BIC)
- Anomaly Detection
- Clustering
    - K Means
    - Hierarchical Clustering
    - Agglomerative Clustering
    - Density Based Clustering (DBSCAN)

Step 7: Seventh, some advance concept in machine learning

- A step into Neural Networks – *The mimic of your brain*
- Autoencoder
- Recurrent Neural Network (RNN)
- Long/short term Memory (LSTM)
- Convolution Neural Network(CNN) – *work with image and videos*
- Deconvolution Network (DN)
- Generative Adversarial Network (GAN)

Step 8: Eight and final step on the roadmap, but not the end of it. Here we will deal with Prescriptive analytics – optimization and prognosis.

- Fuzzy Logic
- Genetic Algorithm
- Nonlinear programming optimization
- Linear programming optimization
- Particle Swarm optimization

If we follow the above road map of Data science with python, it would only be advantageous in understanding the inline concepts of each analytics type.

That is, you may get the full flow of the Data Science Life cycle (DSLC). But you must do some continuous or redundant practices with various use cases (either in Kaggle or other source) to get the complete feel or glitch towards data science.

Happy Coding!!!