Data analyst → Data scientist

Resource cheat sheet for making the jump from data analyst to data scientist.

Essential skills

💻 Programming

Git
- Create new repositories
- Clone existing repositories
- Check out new branches
- Commit and push changes

Environment management
- Set up a virtual environment (with conda, virtualenv)
- Install new packages
- Work with requirements.txt files

SQL
- Select and filter data, merge data from multiple tables
- (Do everything else in Pandas!)

Bonus
- Software testing (Pytest)
- Logging
- Docker
- Data engineering - Airflow, cloud computing (AWS, Azure, Google Cloud)

🔎 Analytical

Understanding data
- Exploratory Data Analysis (EDA)
- Correlation plots (for determining feature importance and identifying confounding variables)
- Analyze distributions of data and identify skewness
- Data aggregation
- Data cleaning (see "Python skills")
- Treat missing values and outliers
- Data normalization (before model building)
- Know how much data is enough to build a model

Statistics
- Measures of central tendency (mean, median, mode, standard deviation)
- T-test for statistical significance

Visualization
- Plots (histogram, density, scatter, heatmap)
- Line of best fit (for regressions)
- ROC curves

🔮 ML

Model building
- Scoping - do you even need an ML model?!
- Supervised learning
  - Linear regression
  - Logistic regression
- CART/decision tree models (very helpful for code tests)
- Unsupervised learning
  - Clustering, topic models
- Deal with class imbalance for multi-label models
- Feature engineering

Model evaluation
- Identify when models have been overfitted
- Save models as pickle files, load them, and inference them
- Model evaluation with confusion matrix metrics (AUC, precision, recall, accuracy, etc)
- Analyze distributions of predicted probabilities

Bonus
- Deploy a model in production
- Deep learning (PyTorch, TensorFlow, Transformers)
- MLOps (MLFlow, monitoring)
- NLP (text embeddings, spaCy, Transformers, BERTopic)

🐍 Python skills

- Read, write data locally
- List comprehensions
- Data types
- Writing custom functions
- Parsing strings
- scikit-learn
  - Split data into train, test sets
  - Train, evaluate models
  - Make predictions
- Pandas
  - Data aggregation
  - Merges, joins
  - Drop duplicates
  - Subset data
  - Sort on columns
  - Create new columns using apply and lambda functions
- Matplotlib or Plotly
  - Histograms, density charts
  - Scatterplots
  - Bar charts
  - Heatmaps
- Basic NLP
  - Tokenization
  - Word stemming, lemmatization
  - TF-IDF

Learning resources

🐍 Learn Python

Datacamp - Intro to Python for Data Science

Datacamp - Learn Python 3

Udemy - 100 Days of Code

YouTube - Learn to Program with Python

YouTube - Learn Python for Beginners

YouTube - Python Full Course for Beginners

YouTube - Effective Pandas by Matt Harrison

💻 Learn computer/data science

Coursera - Supervised Machine Learning

EdX - Data Science: Machine Learning

Harvard University CS50

Stanford University CS101

Stanford University CS329S (MLOps)

🖋 Books

Towards Data Science

Machine Learning Mastery

Tom Augspurger

Andrew Ng

Chip Huyen

Koaning.io

📚 Books

Practical Statistics for Data Scientists

Naked Statistics

Pandas 1.x Cookbook

Designing Machine Learning Systems

⭐ Repos

freeCodeCamp/freeCodeCamp

academic/awesome-datascience

vinta/awesome-python

MrMimic/data-scientist-roadmap

NielsRogge/Transformers-Tutorials

MaartenGr/BERTopic

Created by Mary Newhauser
Updated 1/1/2023