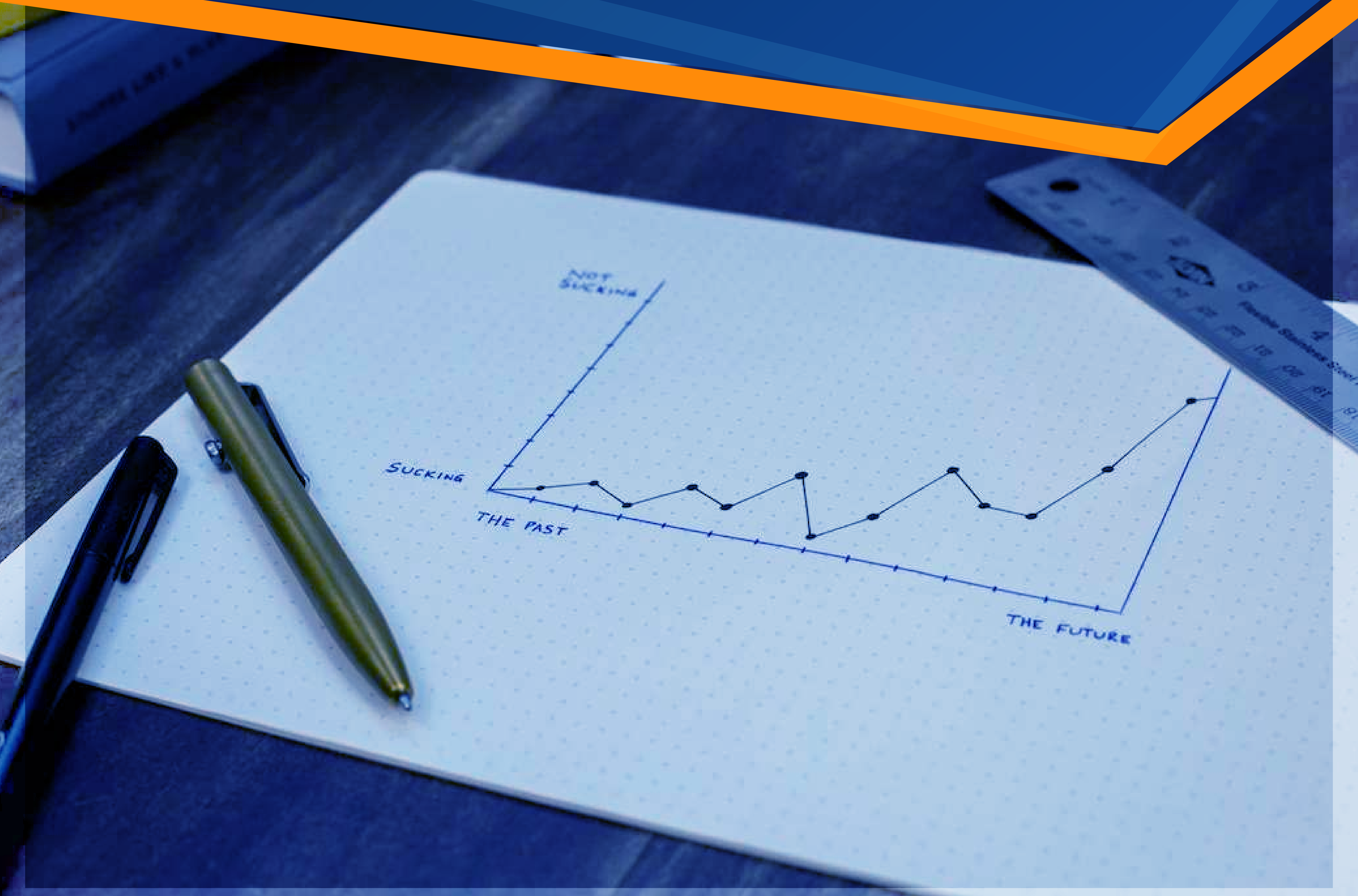
BOSSCODER ACADEMY

CHECKLIST TO BECOME A DATA SCIENTIST



#1 Programming and Tools

Data analysts and scientists use a variety of tools and programming languages in their everyday work. You'll use these tools to query and retrieve data from databases, transform and summarize data, or build machine learning algorithms.

- □ **Spreadsheet tools (like Excel):** These tools visually present data into rows and columns allowing for easy data manipulation. Many organization analyze and communicate data through spreadsheets.
- □ **SQL:** The majority of company data lives in relational databases, and querying that data using SQL is something data analysts and scientist do everyday. Data science managers often cite insufficient SQL skills as a reason for not hiring a candidate.
- □ **Data visualization tools (like Tableau):** Most companies have a data visualization tool used for building dashboards to report company performance. Tableau is the most popular, but there are many others with similar capabilities
- □ **Python programming language:** Python is a high level programming language with many useful packages written for it. Know these Python packages: **NUMPY, PANDAS, MATPLOTLIB, SCIPY, SCIKIT-LEARN**.

Optional Skills to Stand Out

Data analysts and scientists use a variety of tools and programming languages in their everyday work. You'll use these tools to query and retrieve data from databases, transform and summarize data, or build machine learning algorithms.

- □ **R programming language:** a special purpose programming language and software environment for statistical computing and graphics. Know these R packages: **GGPLOT2, DPLYR (OR PLYR), GGALLY, GGPAIRS, RESHAPE2**.
- □ **ipython:** an improved interactive shell for python with introspection, rich media, additional shell syntax, tab completion, and richer history.
- □ **Anaconda:** A python package manager for science, math, engineering, data analysis with the intent of simplifying and maintaining compatibility between library versions. Also useful for getting started with ipython notebooks.
- □ **Seaborn:** A Python visualization library based on matplotlib with a high-level interface

#2 Statistics

At least a basic understanding of statistics is vital as a data analyst.

- **Descriptive Statistics**
  - □ **Mean, median, mode**
  - □ **Data distributions:** Standard normal, Exponential/ Poisson, Binomial, Chi-square.
  - □ **Standard deviation and variance**

- **Inferential Statistics**
  - □ **Hypothesis testing**
    - P-values, Confidence Intervals
  - □ **Test for significance**
    - Z-test, t-test, Mann-Whitney U
    - Chi-squared and ANOVA testing
  - □ **Regression**
    - Linear Regression
    - Logistic Regression

#3 Data Wrangling

A less celebrated part of doing data science is manually collecting and cleaning data so it can be easily explored and analyzed later. This process is otherwise known as "**data wrangling**" or "**data munging**" in the data science community.

□ **Python:** ideal for wrangling data

- String manipulations
- Parsing common file formats such as csv and xml files
- Regular Expressions
- Mathematical transformations, such as converting non-normal distribution to normal with log-10 transformation

□ **SQL:** querying relational databases, such as **Oracle, SQL Server, PostgreSQL, or mySQL.**

#4 Communication and Data Visualization

□ **Data visualization and communications:** Knowing how to present the data in the most consumable way is crucial to communicating the message

- Understand visual encoding and communicating what you want the audience to take away from your visualizations
- Programming and Tools
- TABLEAU
- PYTHON: matplotlib, seaborne
- R: ggplot
- Presenting data and convincing people with your data

□ **Data Storytelling:** Data analysts and scientists should know how to present an engaging narrative that empowers the audience to take action. Data analysts should be aware of the type of audience they are presenting to and craft the presentation to that type of audience.

#5 Machine Learning

□ **Supervised Learning**

Decision trees, Naive Bayes classification, Ordinary Least Squares regression, Logistic regression, Neural networks, Support vector machines, Ensemble methods.

□ **Unsupervised Learning**

- Clustering Algorithms
- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Independent Component Analysis (ICA)Reinforcement Learning

□ **Optional Machine Learning Skills to Stand Out**

- **Reinforcement Learning**
  Q-Learning, TD-Learning, Genetic Algorithms
- **Deep Learning**
  Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Generative Adversarial Networks

#6 Data Intuition: Thinking like a Data Scientist

- □ **Ask the right questions:** The data analyst must be aware of the “question behind the question”—what are the exact business questions and issues that is driving the need to analyze data?
- □ **Recognize:** What things are important and what things are not important.

Optional Skills to Stand Out

- □ **Project management** involves organizing one’s team and managing communications and expectations across multiple departments and parties on any data analyst project
- □ **Communicate effectively with stakeholders including:**
  - Executives and project sponsor
  - Project leads Product managers
  - Engineering, Sales, Information Technology
- □ **Subject Matter knowledge in area of analysis**

#7 Data Scientist: Additional Skills and Knowledge

□ Software Skills

Data scientists should be strong programmers, and should be able to debug and test complex programs.

- Debugging
- Testing
- Version control (Git)
- SQL: editing and updating relational databases

□ Advanced Math

- Linear algebra and Calculus
- Matrix manipulations. Dot product is crucial to understand.
- Eigenvalues and eigenvectors -- Understand the significance of these two concepts
- Multivariable derivatives and integration in Calculus

□ Experiment Design

- A/B Testing
- Controlling variables and choosing good control and testing groups
- Sample Size and Power law
- Confidence level
- SMART experiments: Specific, Measurable, Actionable, Realistic, Timely
- Bayesian Statistics
- Bootstrapping
- Simulation