Cheat Sheet for comprehensive Data Science Council of America (DASCA) Senior Data Scientist (SDS)

Data Collection & Preprocessing

Data Sources

- **Structured Data**: Databases, CSV, Excel
- **Unstructured Data**: Text, Images, Audio
- **APIs**: RESTful, SOAP
- **Web Scraping**: BeautifulSoup, Scrapy

Data Cleaning

- **Missing Values**:

- Imputation: Mean, Median, Mode
- Deletion: Rows, Columns

- **Outliers**:

- Z-Score, IQR
- Visualization: Boxplots, Histograms

- **Duplicates**:

- Removal: Rows, Columns

- **Data Transformation**:

- Normalization: Min-Max, Z-Score
- Encoding: One-Hot, Label Encoding

Exploratory Data Analysis (EDA)

Descriptive Statistics

- **Central Tendency**: Mean, Median, Mode
- **Dispersion**: Range, Variance, Standard Deviation
- **Shape**: Skewness, Kurtosis

Visualization

- **Histograms**: Distribution of single variable

- **Boxplots**: Outliers and quartiles

- **Scatterplots**: Relationship between two variables

- **Heatmaps**: Correlation between variables

Statistical Analysis

Hypothesis Testing

- **Types**:

- Z-Test, T-Test
- Chi-Square Test
- ANOVA

- **Steps**:

- Formulate Hypothesis
- Set Significance Level (α)
- Calculate Test Statistic
- Determine P-Value
- Make Decision

Regression Analysis

- **Simple Linear Regression**: $y = \beta 0 + \beta 1x$

- **Multiple Linear Regression**: $y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$

- **Metrics**:

- R-Squared
- Adjusted R-Squared
- RMSE, MAE

Machine Learning

Supervised Learning

- **Classification**:

- Algorithms: Logistic Regression, SVM, Decision Trees, Random Forest, KNN
- Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC

- **Regression**:

- Algorithms: Linear Regression, Ridge, Lasso, Elastic Net
- Metrics: RMSE, MAE, R-Squared

Unsupervised Learning

- **Clustering**:

- Algorithms: K-Means, Hierarchical, DBSCAN
- Metrics: Silhouette Score, Davies-Bouldin Index

- **Dimensionality Reduction**:

- Algorithms: PCA, t-SNE, LDA

Model Evaluation

- **Cross-Validation**:

- K-Fold
- Stratified K-Fold

- **Hyperparameter Tuning**:

- Grid Search
- Random Search
- Bayesian Optimization

Deep Learning

Neural Networks

- **Layers**:

- Input, Hidden, Output

- **Activation Functions**:

- ReLU, Sigmoid, Tanh

- **Loss Functions**:

- MSE, Cross-Entropy

Convolutional Neural Networks (CNN)

- **Layers**:

- Convolutional, Pooling, Fully Connected

- **Applications**:

- Image Classification, Object Detection

Recurrent Neural Networks (RNN)

- **Types**:

- LSTM, GRU

- **Applications**:

- Time Series, NLP

Big Data & Distributed Computing

Big Data Technologies

- **Hadoop**: HDFS, MapReduce

- **Spark**: RDD, DataFrames, MLlib

- **NoSQL Databases**: MongoDB, Cassandra

Distributed Computing

- **Frameworks**:

- Apache Spark
- Dask

- **Parallel Processing**:

- Multiprocessing
- Multithreading

Data Visualization & Reporting

Tools

- **Matplotlib**: Basic plotting

- **Seaborn**: Statistical plots

- **Plotly**: Interactive plots

- **Tableau**: Business Intelligence

Reporting

- **Dashboards**:

- Real-time updates
- Interactive elements

- **Storytelling**:

- Clear narratives
- Visual hierarchy

Ethics & Compliance

Data Privacy

- **GDPR**: General Data Protection Regulation
- **HIPAA**: Health Insurance Portability and Accountability Act
- **Data Anonymization**:

- Techniques: Masking, Shuffling

Bias & Fairness

- **Types of Bias**:

- Selection Bias
- Confirmation Bias

- **Mitigation**:

- Fairness Metrics
- Algorithmic Audits

Tools & Libraries

Python Libraries

- **Data Manipulation**: Pandas, NumPy
- **Visualization**: Matplotlib, Seaborn, Plotly
- **Machine Learning**: Scikit-Learn, XGBoost
- **Deep Learning**: TensorFlow, PyTorch

R Libraries

- **Data Manipulation**: dplyr, tidyr
- **Visualization**: ggplot2
- **Machine Learning**: caret, randomForest

Best Practices

Version Control

- **Git**:

- Commands: init, clone, add, commit, push, pull

- **GitHub**:

- Repositories, Pull Requests

Documentation

- **Jupyter Notebooks**:

- Markdown cells
- Code comments

- **Readthedocs**:

- Project documentation

Collaboration

- **Agile Methodologies**:

- Scrum, Kanban

- **Tools**:

- Jira, Trello

Examples

Python Code Snippets

- **Data Loading**:

```
import pandas as pd
df = pd.read_csv('data.csv')
```

- **Linear Regression**:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
```

R Code Snippets

- **Data Loading**:

```
df <- read.csv('data.csv')
```

- **Linear Regression**:

```
model <- lm(y ~ ., data = df)
summary(model)
```

Resources

Books

- **"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"** by Aurélien Géron

- **"The Elements of Statistical Learning"** by Trevor Hastie, Robert Tibshirani, Jerome Friedman

Online Courses

- **Coursera**: "Machine Learning" by Andrew Ng

- **edX**: "Data Science MicroMasters" by Harvard

Communities

- **Kaggle**: Competitions, Datasets

- **Stack Overflow**: Q&A

- **Reddit**: r/datascience, r/machinelearning

Conclusion

- **Continuous Learning**: Stay updated with latest trends and technologies

- **Practice**: Regularly work on projects and competitions

- **Networking**: Engage with the data science community

By Ahmed Baheeg Khorshid

ver 1.0