Data Science Cheat Sheet (Sara's Learning Journey)

# 1. NumPy Basics

- Lists vs NumPy arrays:  
 - Lists: flexible, multiple data types, not memory efficient.  
 - NumPy arrays: fixed type, contiguous memory, efficient for math operations.  
- Key functions:  
 - np.arange(start, stop, step)  
 - arr.shape, arr.size  
 - Boolean indexing: arr[arr % 2 == 0]  
 - Random arrays: np.random.rand(3,3)

# 2. Pandas Basics

- DataFrame = 2D table (rows, columns)  
- Inspection:  
 - df.head(), df.info(), df.describe()  
 - df.isnull(), df.isnull().sum()  
- Handling Missing Data:  
 - Fill: df['Age'].fillna(df['Age'].mean())  
 - Drop: df.dropna(subset=['Performance\_Score'])  
- Grouping:  
 - df.groupby("Department")["Salary"].mean()  
- Adding Columns:  
 - df["Bonus"] = df["Salary"] \* 1.10  
 - Assigning categories with df.loc[]  
- Sorting:  
 - df.sort\_values("Salary", ascending=True)

# 3. Data Visualization

Matplotlib:  
- plt.bar(x,y), plt.scatter(x,y), plt.hist(x, bins=n)  
- Subplots: plt.subplot(rows, cols, index)  
  
Seaborn:  
- sns.histplot(data, x="col", bins=n, hue="category")  
- sns.barplot(data, x="Department", y="Salary", estimator="mean")  
- sns.scatterplot(data, x="Years\_At\_Company", y="Salary")  
- sns.regplot(data, x="Years\_At\_Company", y="Salary")  
- sns.boxplot(data, x="Department", y="Performance\_Score")  
- sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm")  
- sns.pairplot(df, vars=["Salary","Years\_At\_Company"], hue="Department")

# 4. Exploratory Data Analysis (EDA)

- Average Salary by Department → df.groupby("Department")["Salary"].mean()  
- Average Performance by Department → df.groupby("Department")["Performance\_Score"].mean()  
- Scatter Plot (Salary vs Years\_At\_Company) → positive correlation (~0.94)  
- Bonus vs Performance\_Score → strong correlation (~0.89)  
- Boxplot showed Marketing had most consistent performance  
- Age distribution by department revealed workforce spread

# 5. Big Picture So Far

✅ NumPy (arrays, indexing, random numbers)  
✅ Pandas (dataframes, cleaning, grouping, new columns)  
✅ Matplotlib & Seaborn (plots, subplots, histograms, scatter, regression, box, heatmap, pairplot)  
✅ Exploratory Data Analysis (EDA)  
Next Step → Machine Learning (Regression)