Task 4 — Data Cleaning & Insight Generation from Survey Data

# 1. Introduction & Dataset

This project simulates the Kaggle Data Science Survey (2017–2021) with 1,500+ responses. We intentionally included common survey issues — missing values, duplicates, and formatting inconsistencies — to demonstrate a realistic, end-to-end data cleaning workflow and insight generation process.

Raw rows: 1,530. Cleaned rows after de-duplication: 1,500.

# 2. Cleaning Methodology

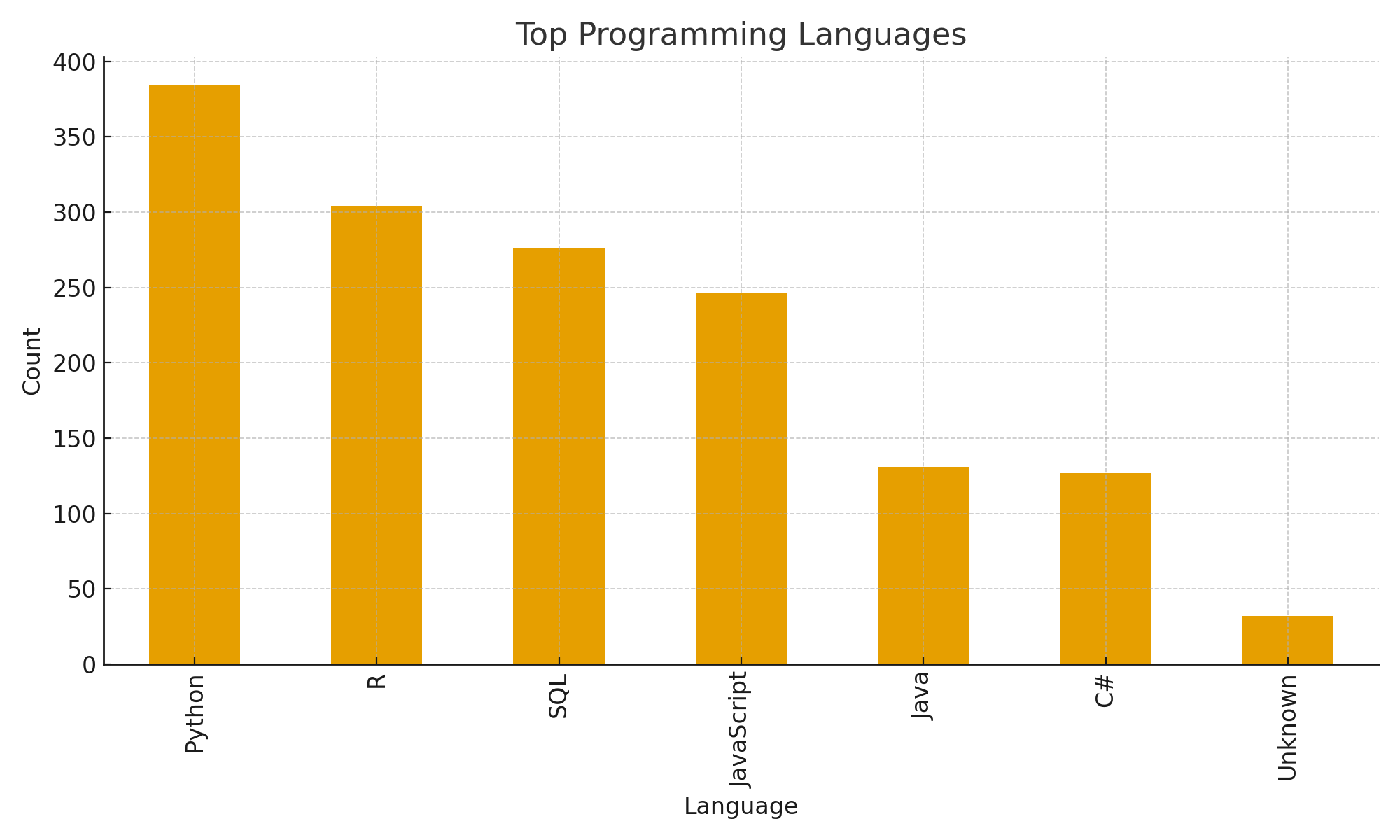
Key steps included:  
• Standardizing country, education, job role, primary language, and remote work values via deterministic mappings.  
• Parsing years of experience ranges into numeric midpoints (e.g., '3-5' → 4) and '10+' → 12.  
• Normalizing salary strings to floats, imputing missing salaries per job-role median, and winsorizing at the 1st/99th percentiles to limit outliers.  
• Dropping duplicates based on (respondent\_id, year, job\_role\_raw, primary\_language\_raw).  
• Splitting semi-colon ML tool lists into arrays and generating a tool-count feature.  
• Label-encoding key categorical fields for downstream modeling.

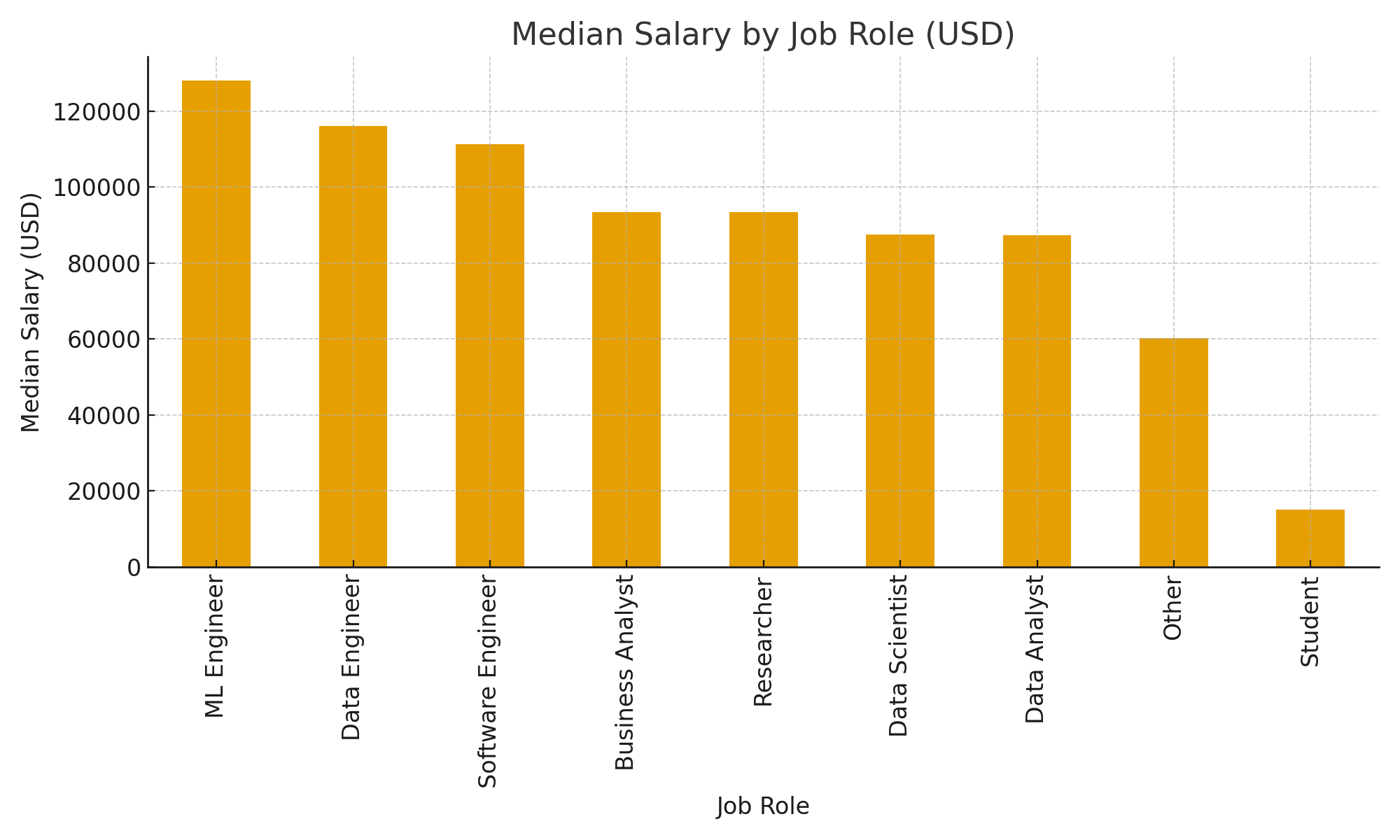
# 3. Key Insights (Top 5)

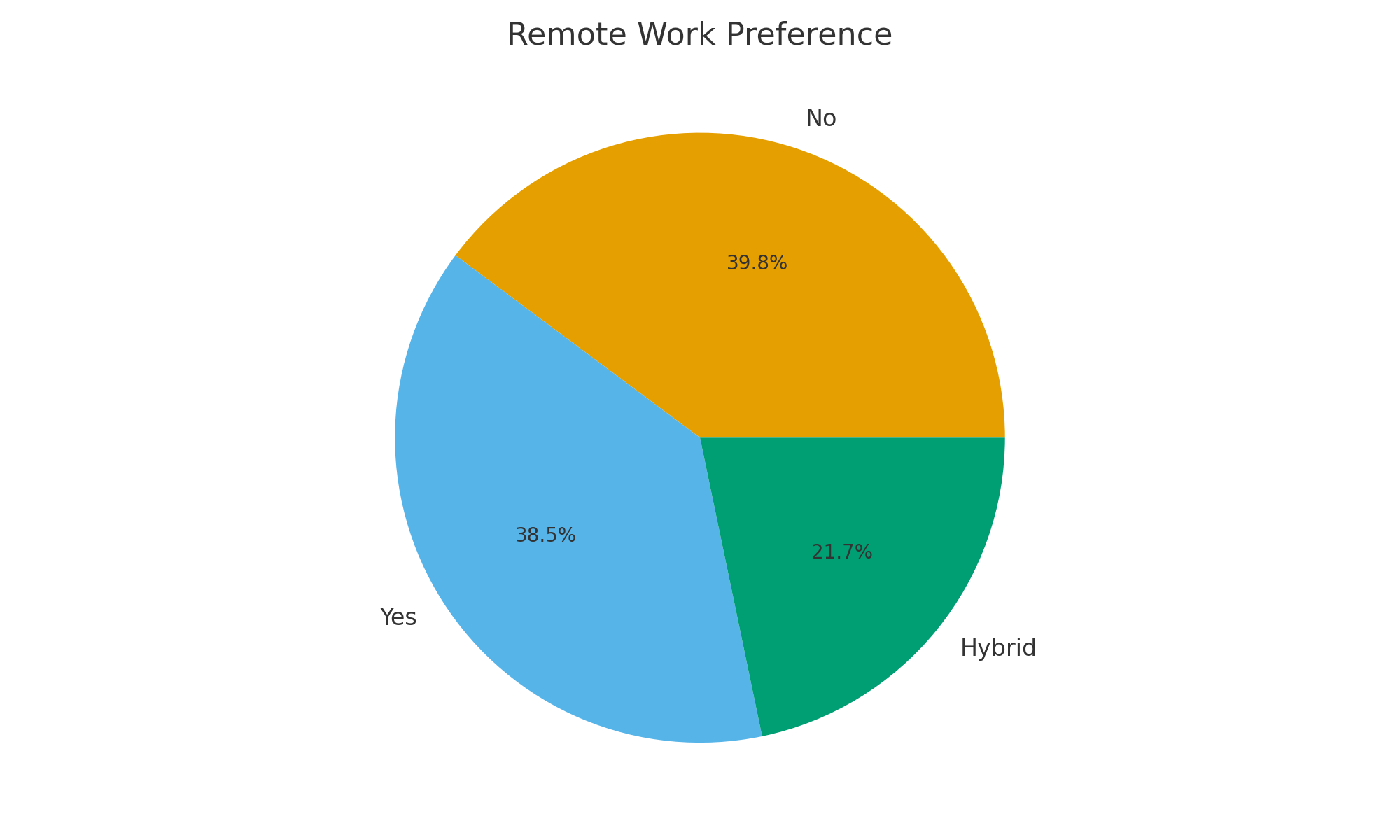
1) Most common primary language: Python  
2) Highest median salary role: ML Engineer (~$127994)  
3) Remote work preference (Yes): 38.5%  
4) Most frequently cited ML tool: LightGBM  
5) Experience vs Salary correlation (approx): 0.02

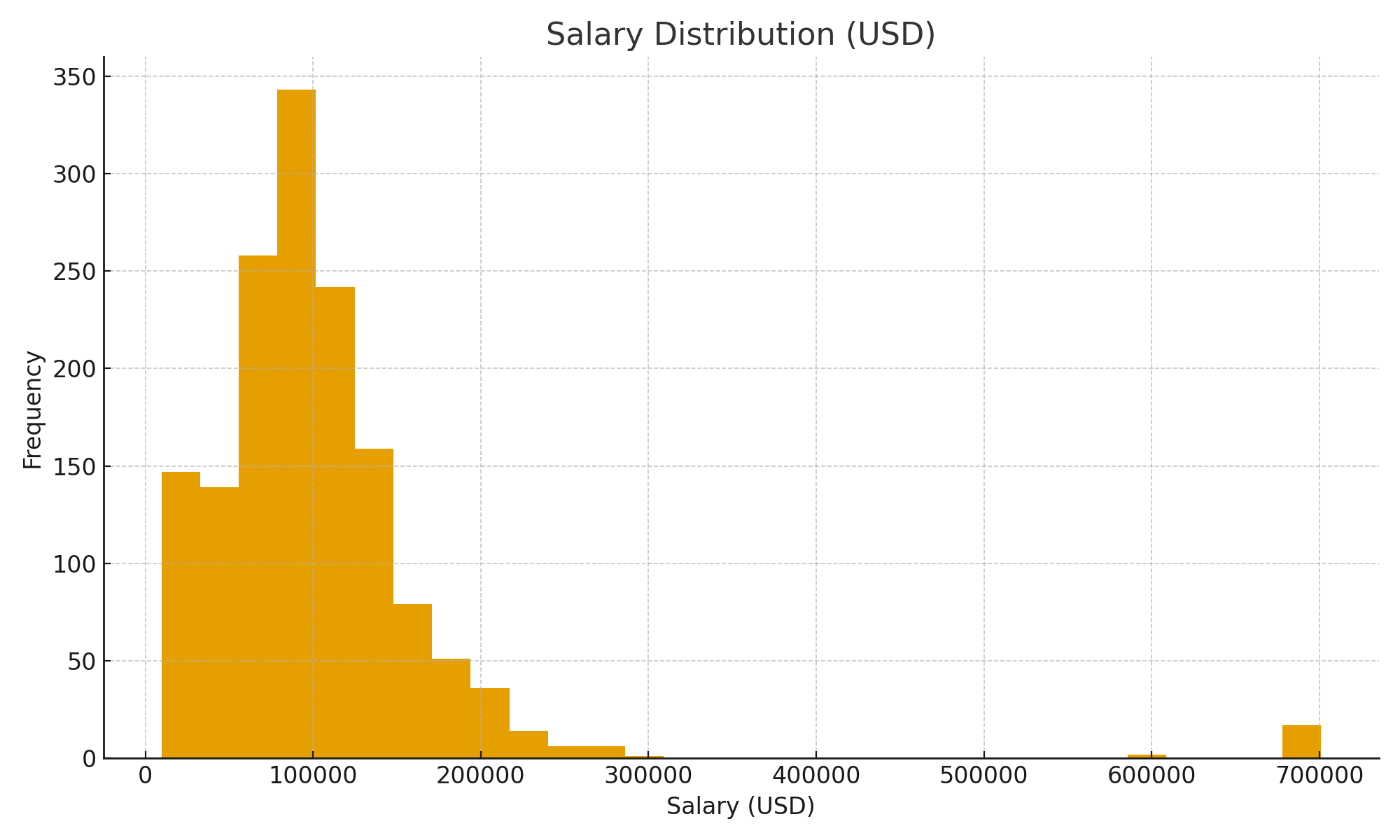
# 4. Visual Summary

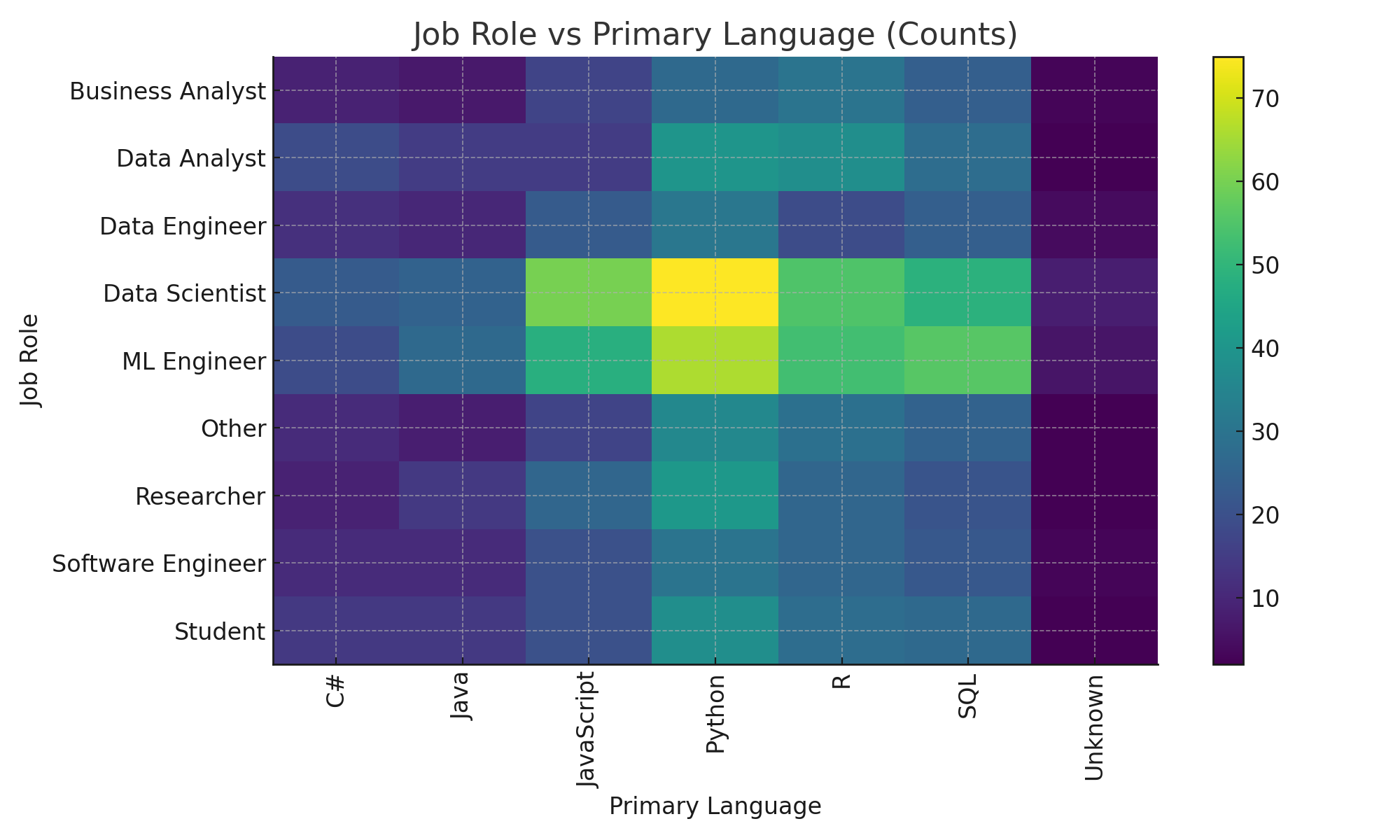
Selected charts from the analysis are embedded below.











# 5. Implications

• Programming language popularity can guide training investments and hiring strategies.  
• Salary differences by role support compensation benchmarking and workforce planning.  
• Remote-work preferences inform hybrid-work policy design and tooling decisions.  
• Tool usage trends highlight common stacks for model-building and MLOps.  
• Experience-salary relationships suggest pathways for career development programs.