California State University, Chico

MATH 485 - Advanced Topics in Data Science

Final Report

Professor:
Dr. Bo Shen

1. Title

Job Market Analysis Using Clustering, Salary Prediction, and Skill Network Analysis.

2. Team Member Names

- Pushpak Sunil Rane
- Amol Bhalerao
- Vatsal Rami

3. Overall Context and Relevant Work

In today's dynamic job market, understanding trends, skill demands, and salary expectations is critical. Our project addresses these issues by analyzing job postings to derive actionable insights using clustering, salary prediction, and skill network analysis.

This project utilizes advanced data science techniques to enhance decision-making for job seekers and recruiters. By extracting insights from job listings, we can identify trends, predict salaries, and map skill requirements, facilitating strategic career and hiring decisions.

4. High-Level Framework

Our solution uniquely integrates:

- **Clustering:** Categorizes job postings based on roles and required skills.
- **Salary Prediction:** Uses regression models to estimate salaries from job descriptions.
- **Skill Network Analysis:** Creates a visual representation of co-occurring skills.
- **Interactive Dashboard:** Presents the findings through user-friendly visualization tools.

5. Detailed Aspects of the Solution

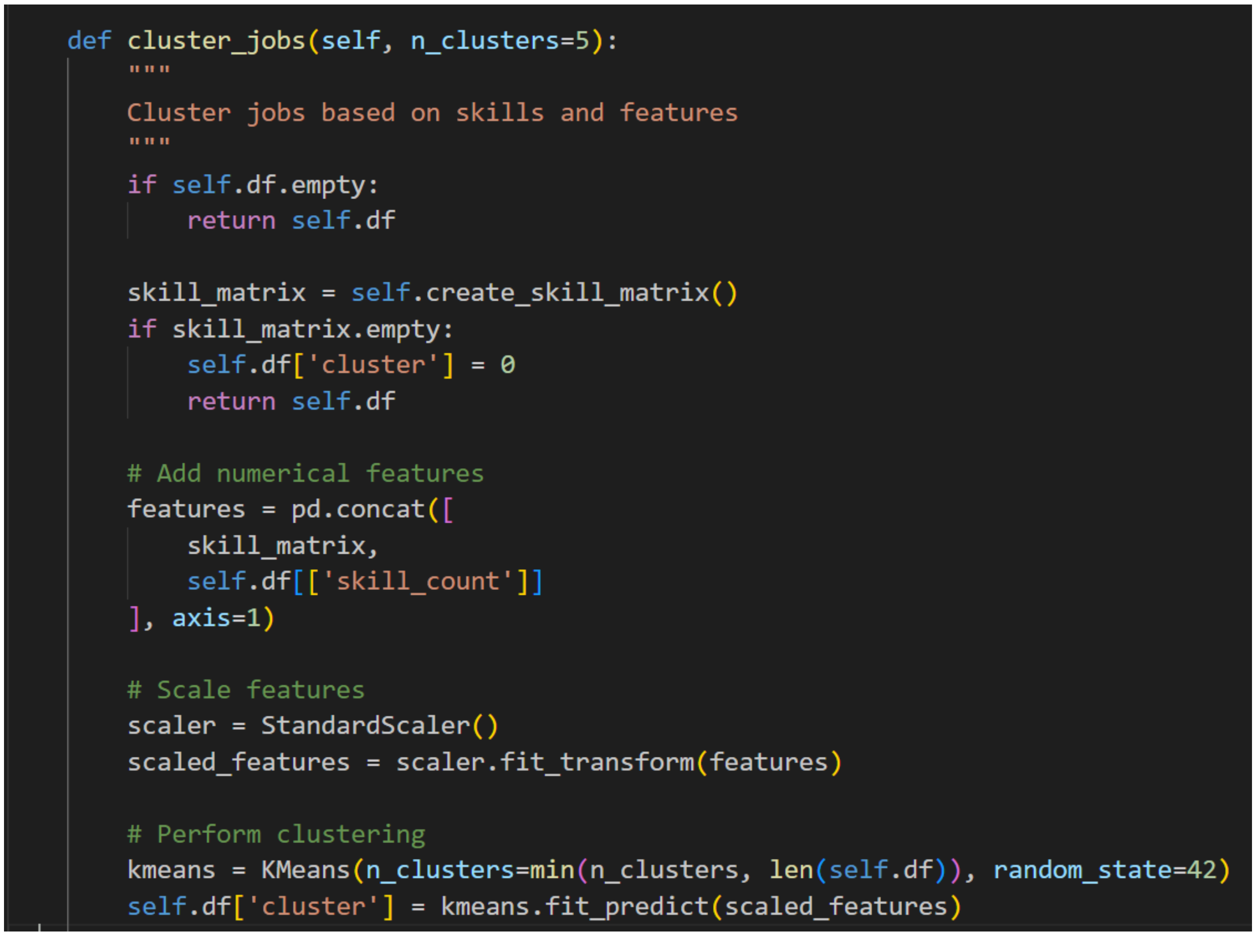
5.1 Implementation Summary

- **Data Collection and Preprocessing:**
    - Utilized free job listing Adzuna API to collect job postings.
    - Processed data to handle missing values and duplicates.
- **Job Clustering:**
    - Applied clustering algorithms like K-means and Hierarchical clustering.
    - Identified optimal clustering features to enhance cluster quality.
- **Salary Prediction:**
    - Developed regression models considering job title, skills, and location.
    - Evaluated and optimized models for accuracy and robustness.
- **Skill Network Analysis:**
    - Built network graphs to visualize skill interrelationships.

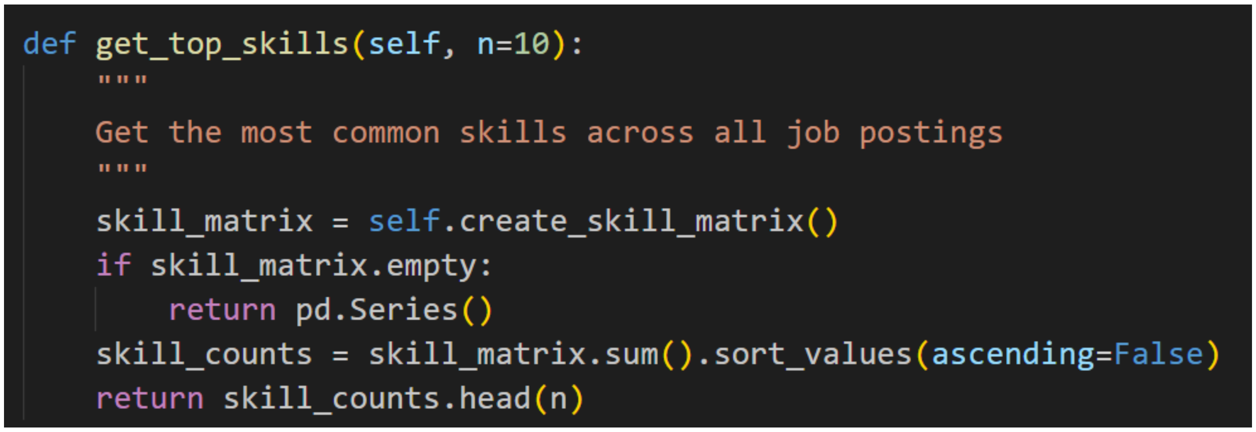
5.2 Focus on Your Solution

The uniqueness of our solution lies in its integrative approach, combining robust data preprocessing, sophisticated clustering methods, predictive modeling, and visual skill relationship mapping. Our interactive dashboard further distinguishes our work by allowing users to dynamically explore insights.

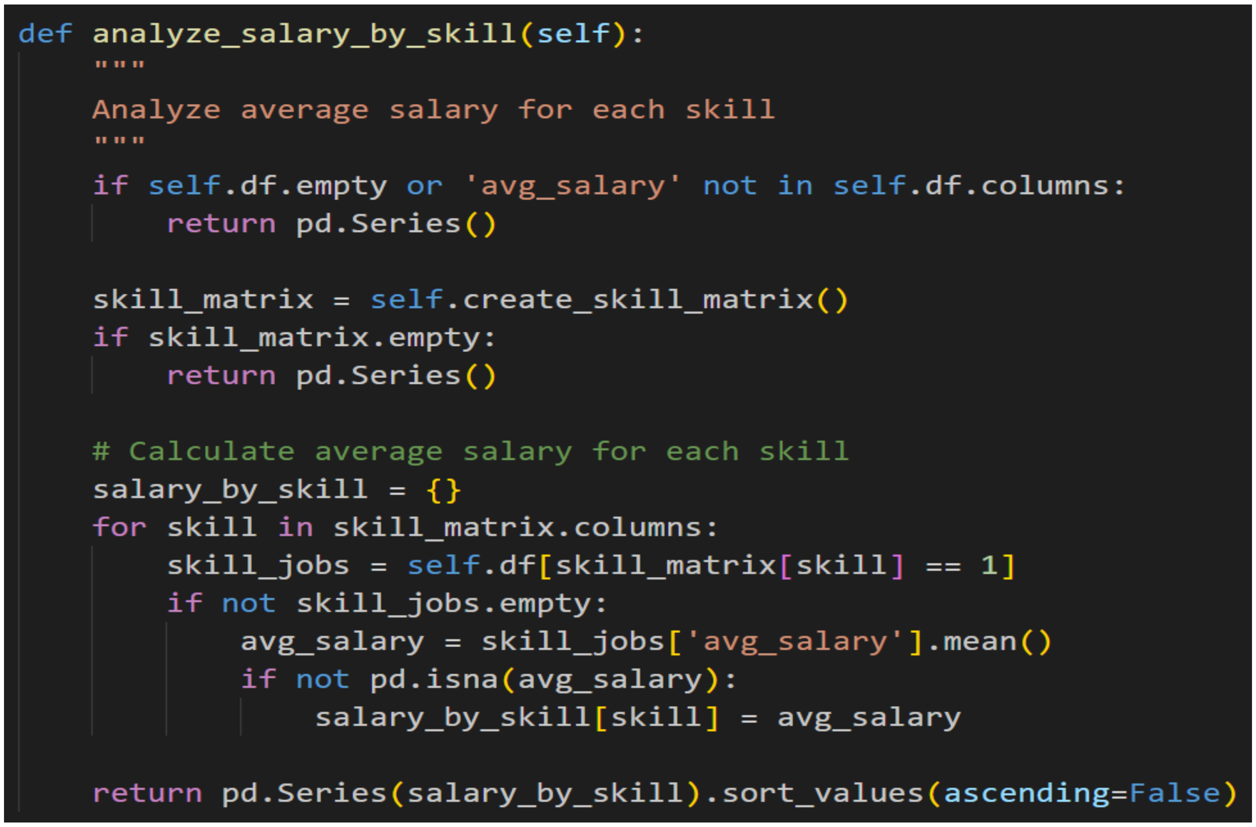
5.3 Important Code Snippets



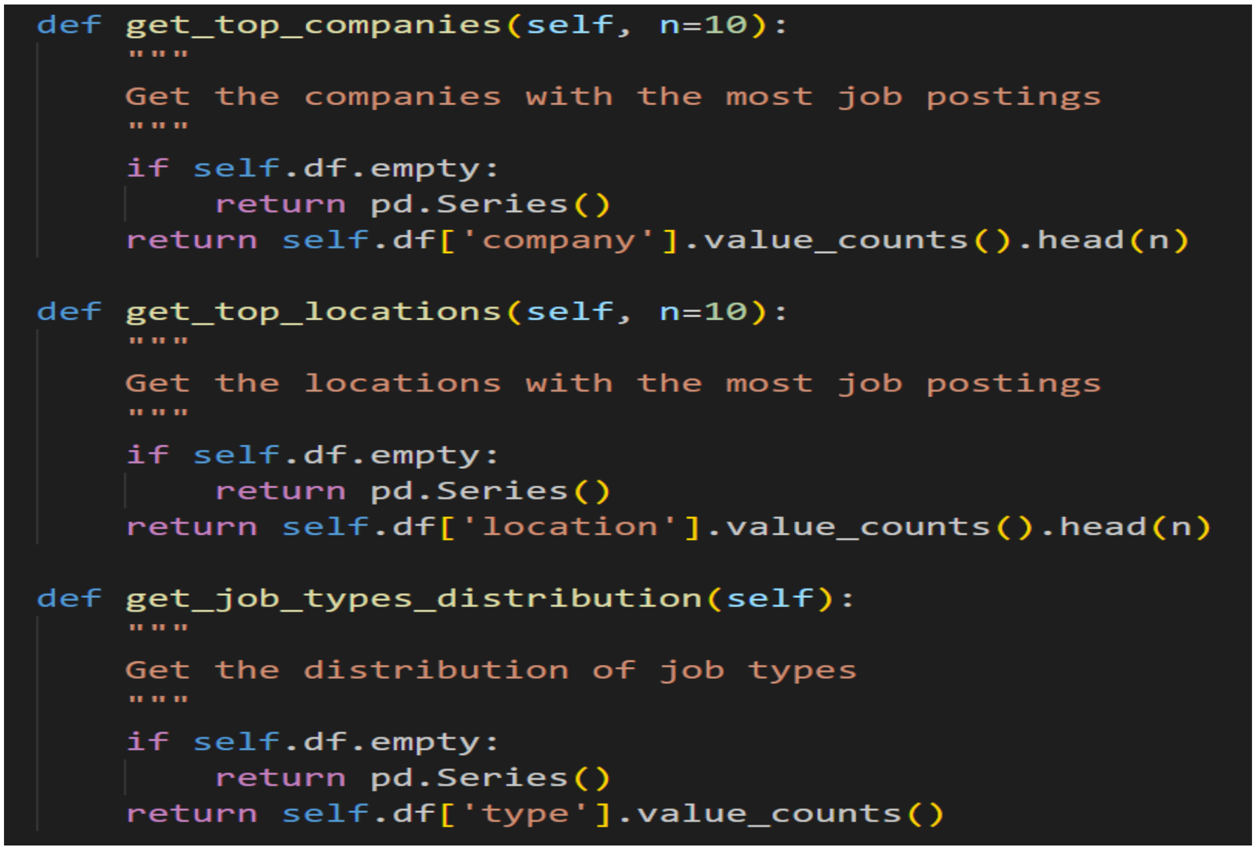
```python
def cluster_jobs(self, n_clusters=5):
    """
    Cluster jobs based on skills and features
    """
    if self.df.empty:
        return self.df

    skill_matrix = self.create_skill_matrix()
    if skill_matrix.empty:
        self.df['cluster'] = 0
        return self.df

    # Add numerical features
    features = pd.concat([
        skill_matrix,
        self.df[['skill_count']]
    ], axis=1)

    # Scale features
    scaler = StandardScaler()
    scaled_features = scaler.fit_transform(features)

    # Perform clustering
    kmeans = KMeans(n_clusters=min(n_clusters, len(self.df)), random_state=42)
    self.df['cluster'] = kmeans.fit_predict(scaled_features)
```



```python
def get_top_skills(self, n=10):
    """
    Get the most common skills across all job postings
    """
    skill_matrix = self.create_skill_matrix()
    if skill_matrix.empty:
        return pd.Series()
    skill_counts = skill_matrix.sum().sort_values(ascending=False)
    return skill_counts.head(n)
```



```python
def analyze_salary_by_skill(self):
    """
    Analyze average salary for each skill
    """
    if self.df.empty or 'avg_salary' not in self.df.columns:
        return pd.Series()

    skill_matrix = self.create_skill_matrix()
    if skill_matrix.empty:
        return pd.Series()

    # Calculate average salary for each skill
    salary_by_skill = {}
    for skill in skill_matrix.columns:
        skill_jobs = self.df[skill_matrix[skill] == 1]
        if not skill_jobs.empty:
            avg_salary = skill_jobs['avg_salary'].mean()
            if not pd.isna(avg_salary):
                salary_by_skill[skill] = avg_salary

    return pd.Series(salary_by_skill).sort_values(ascending=False)
```



```python
def get_top_companies(self, n=10):
    """
    Get the companies with the most job postings
    """
    if self.df.empty:
        return pd.Series()
    return self.df['company'].value_counts().head(n)

def get_top_locations(self, n=10):
    """
    Get the locations with the most job postings
    """
    if self.df.empty:
        return pd.Series()
    return self.df['location'].value_counts().head(n)

def get_job_types_distribution(self):
    """
    Get the distribution of job types
    """
    if self.df.empty:
        return pd.Series()
    return self.df['type'].value_counts()
```

6. Test Results and Analysis

6.1 Comparison of Solutions

- **Early Models:** Initially faced challenges due to incomplete feature sets and basic preprocessing.
- **Improved Approach:** Enhanced data processing and feature selection significantly improved model accuracy and clustering effectiveness.

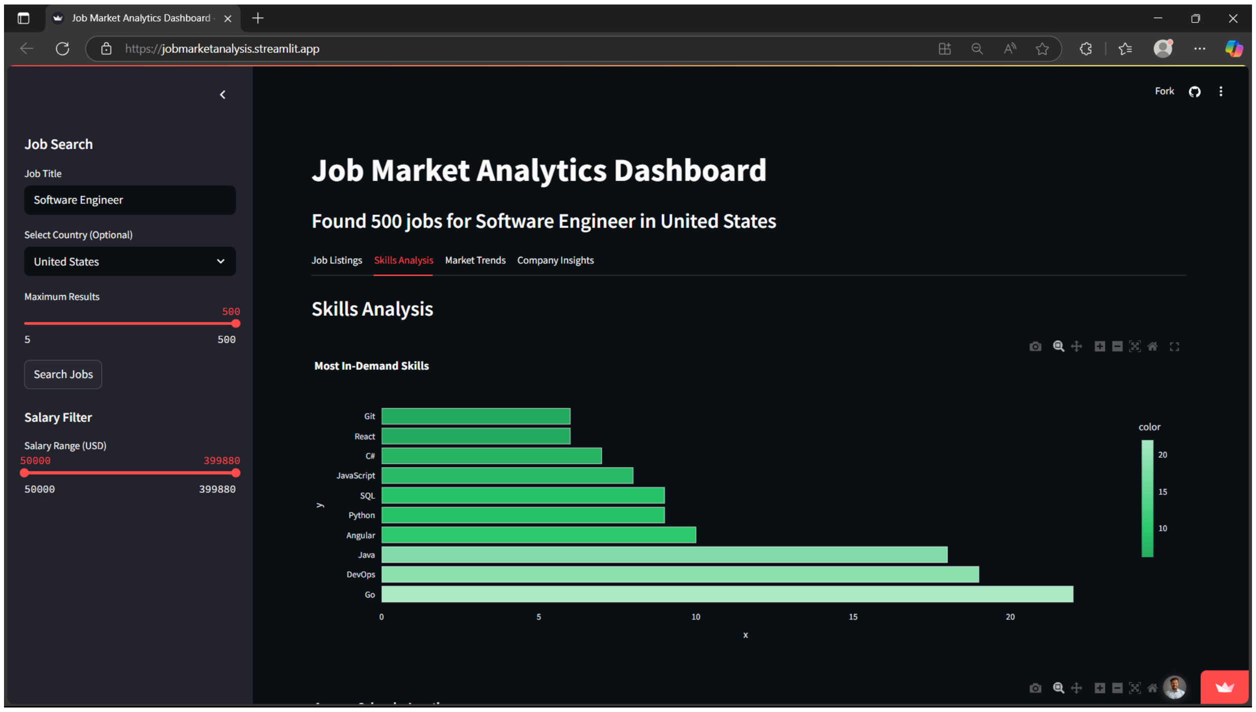
6.2 Comparison with Other Solutions

- Our combined methodology of clustering, regression, and network visualization offered deeper insights compared to singular analytical approaches.
- The integration of multiple features enhanced predictive accuracy over simpler models.

6.3 What Worked vs What Didn’t

- **Effective Approaches:**
    - Comprehensive preprocessing and advanced feature engineering.
    - Efficient integration of analytical techniques.
- **Areas of Improvement:**
    - Data quality initially affected analysis outcomes.
    - Initial complexities in network graph visualization required additional optimization.

7. Dashboard



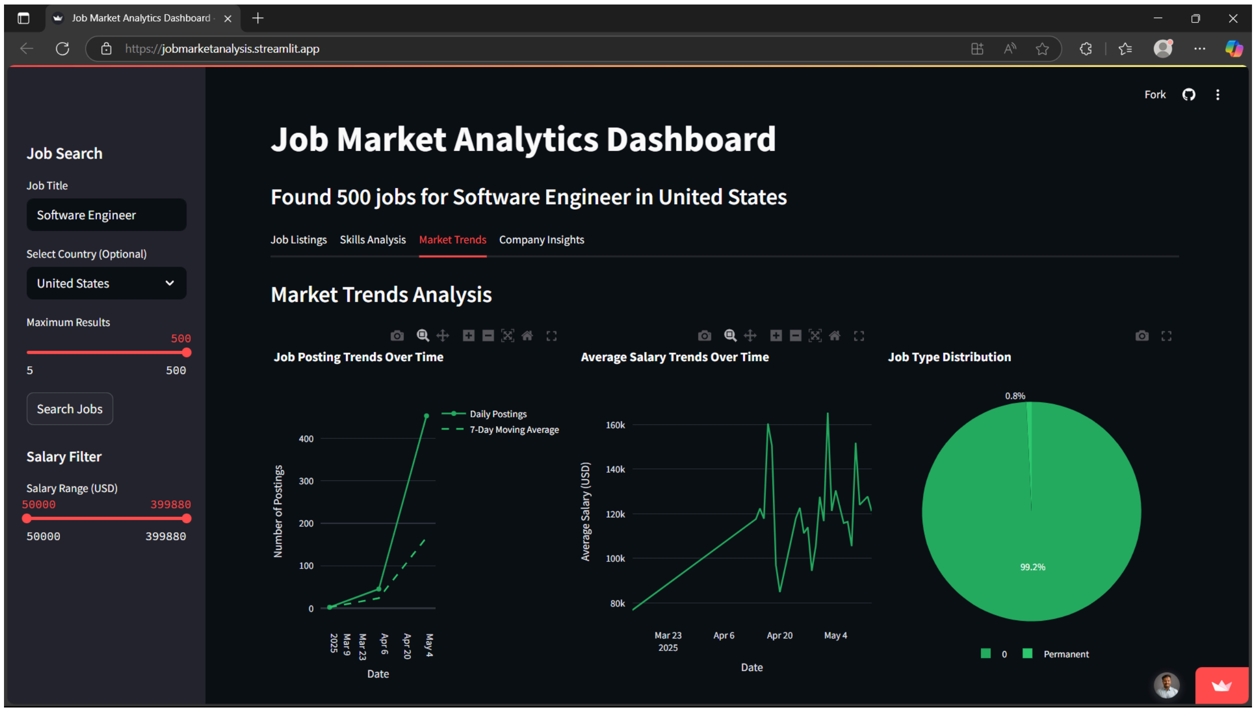
Job Search
Job Title
Software Engineer
Select Country (Optional)
United States
Maximum Results 500
Search Jobs
Salary Filter
Salary Range (USD) 50000 – 399880

Job Market Analytics Dashboard

Found 500 jobs for Software Engineer in United States

Job Listings | Skills Analysis | Market Trends | Company Insights

Skills Analysis



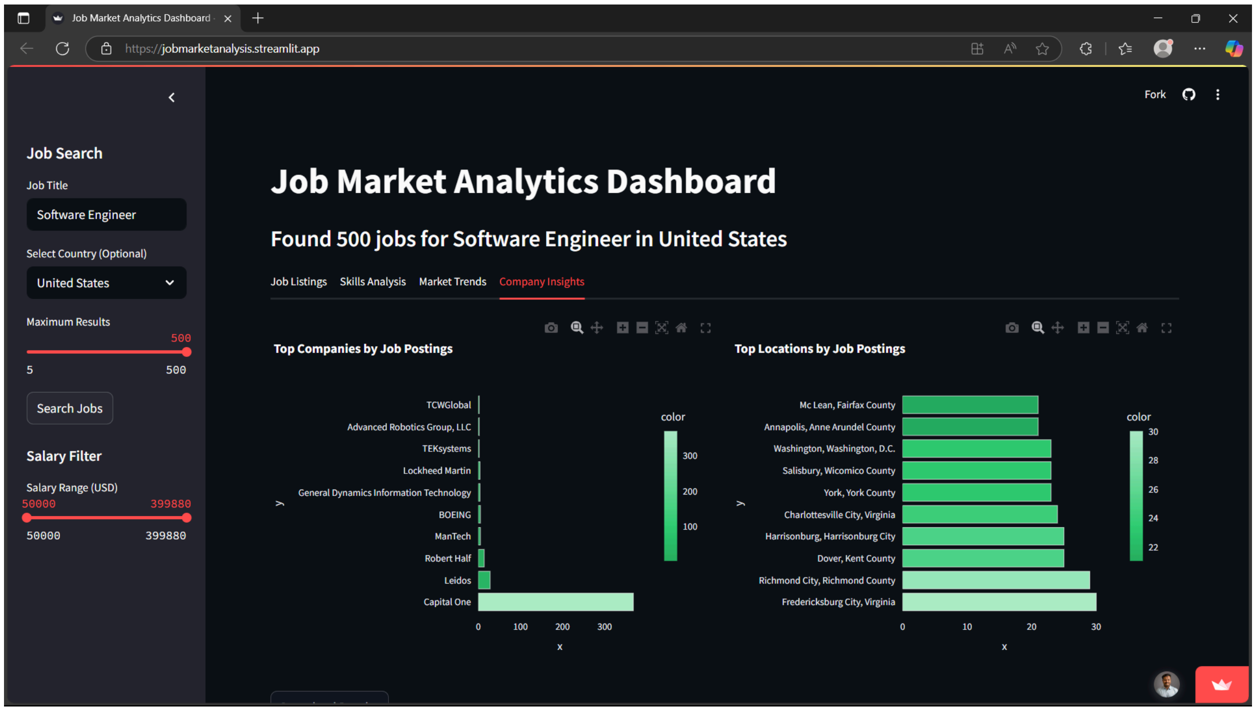
Job Search
Job Title
Software Engineer
Select Country (Optional)
United States
Maximum Results 500
Search Jobs
Salary Filter
Salary Range (USD) 50000 – 399880

Job Market Analytics Dashboard

Found 500 jobs for Software Engineer in United States

Job Listings | Skills Analysis | Market Trends | Company Insights

Market Trends Analysis


Job Market Analytics Dashboard
https://jobmarketanalysis.streamlit.app
Fork
Job Listings
Skills Analysis
Market Trends
Company Insights
Top Companies by Job Postings
TCWGlobal
Advanced Robotics Group, LLC
TEKsystems
Lockheed Martin
General Dynamics Information Technology
BOEING
ManTech
Robert Half
Leidos
Capital One
y
x
0
100
200
300
color
300
200
100
Top Locations by Job Postings
Mc Lean, Fairfax County
Annapolis, Anne Arundel County
Washington, Washington, D.C.
Salisbury, Wicomico County
York, York County
Charlottesville City, Virginia
Harrisonburg, Harrisonburg City
Dover, Kent County
Richmond City, Richmond County
Fredericksburg City, Virginia
0
10
20
30
color
30
28
26
24
22

Job Search

Job Title: Software Engineer

Select Country (Optional): United States

Maximum Results: 500 (5 – 500)

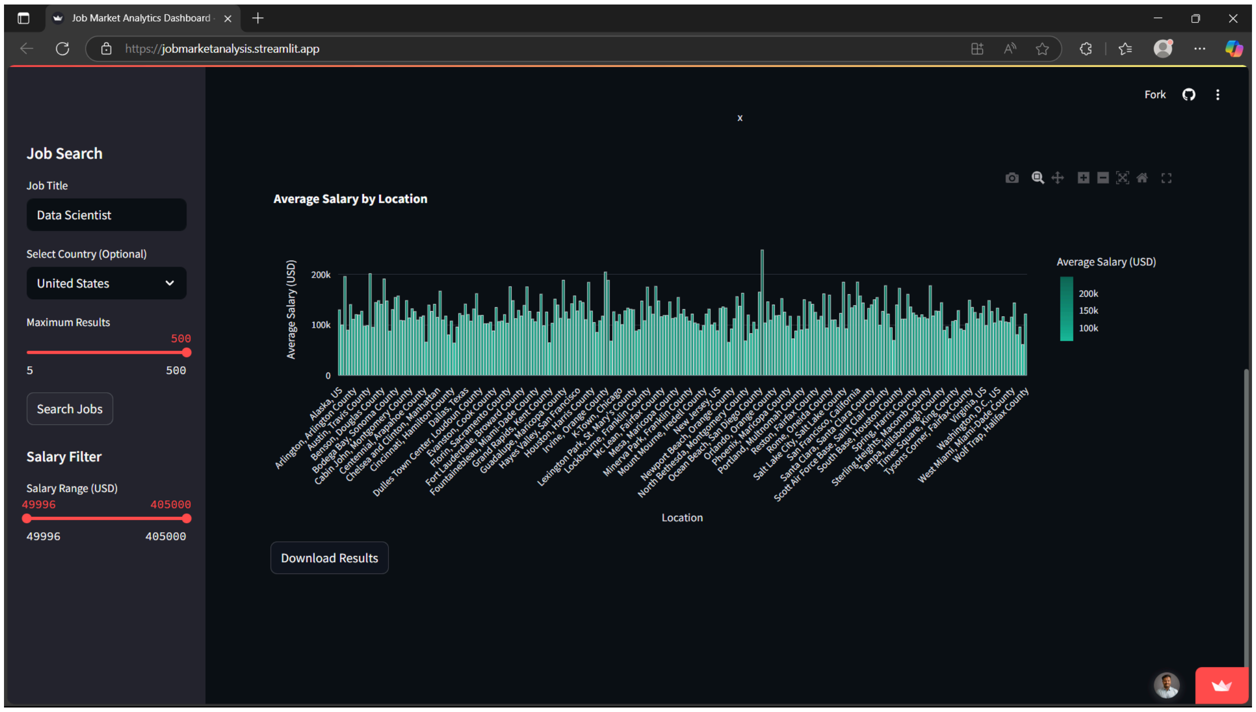
Search Jobs

Salary Filter

Salary Range (USD): 50000 – 399880

Job Market Analytics Dashboard

Found 500 jobs for Software Engineer in United States


Job Market Analytics Dashboard
https://jobmarketanalysis.streamlit.app
Fork
x
Average Salary by Location
Average Salary (USD)
200k
100k
0
Location
Average Salary (USD)
200k
150k
100k
Alaska, US
Arlington, Arlington County
Austin, Travis County
Benson, Douglas County
Bodega Bay, Sonoma County
Cabin John, Montgomery County
Centennial, Arapahoe County
Chelsea and Clinton, Manhattan
Cincinnati, Hamilton County
Dallas, Texas
Dulles Town Center, Loudoun County
Evanston, Cook County
Florin, Sacramento County
Fort Lauderdale, Broward County
Fountainebleau, Miami-Dade County
Grand Rapids, Kent County
Guadalupe, Maricopa County
Hayes Valley, San Francisco
Houston, Harris County
Irvine, Orange County
K-Town, Chicago
Lexington Park, St. Mary's County
Lockbourne, Franklin County
Mc Lean, Fairfax County
Mesa, Maricopa County
Minerva Park, Franklin County
Mount Mourne, Iredell County
New Jersey, US
Newport Beach, Orange County
North Bethesda, Montgomery County
Ocean Beach, San Diego County
Orlando, Orange County
Phoenix, Maricopa County
Portland, Multnomah County
Reston, Fairfax County
Rome, Oneida County
Salt Lake City, Salt Lake County
San Francisco, California
Santa Clara, Santa Clara County
Scott Air Force Base, Saint Clair County
South Base, Houston County
Spring, Harris County
Sterling Heights, Macomb County
Tampa, Hillsborough County
Times Square, King County
Tysons Corner, Fairfax County
Virginia, US
Washington, D.C., US
West Miami, Miami-Dade County
Wolf Trap, Fairfax County
Download Results

Job Search

Job Title: Data Scientist

Select Country (Optional): United States

Maximum Results: 500 (5 – 500)

Search Jobs

Salary Filter

Salary Range (USD): 49996 – 405000

Dashboard Link:- ***https://jobmarketanalysis.streamlit.app/***

8. Conclusion

Our integrated analytical approach provides valuable insights into the job market, significantly benefiting stakeholders through enhanced decision-making support.

9. Future Work

- Expansion of dataset scope and geographical diversity.
- Further refinement of dashboard interactivity and user experience.
- Exploration of advanced analytical models to enhance predictive capabilities.