Regression Analysis: Gender Disparities and Political Influence

Clustering and Classification for Workforce Analytics

today

## 🧭 Introduction

Gender inequality in the labor market remains a persistent challenge—despite decades of progress toward workplace equity.

### 🌍 Global Context

Over the past two decades, gender equality in the **global labor market** has improved significantly. However, disparities in **hiring**, **pay**, and **career advancement** remain pervasive. According to (Organization (2023)), women earned, on average, **20% less than men globally in 2023**. Moreover, women continue to be underrepresented in **leadership positions** and **high-paying industries**, highlighting enduring barriers to upward mobility.

These disparities are not just issues of fairness—they have broader implications for **economic growth**, **business performance**, and **societal well-being**. Persistent wage inequality, occupational segregation, and gender discrimination hinder women’s full participation in the labor force and reduce long-term economic sustainability.

### 🇺🇸 The U.S. Divide: Politics and Policy

In the United States, gender inequality is further shaped by **state-level political ideologies**. Labor policies—such as **minimum wage regulations**, **paid parental leave**, and **workplace protections**—differ significantly between **conservative (red)** and **liberal (blue)** states. These policy differences create variation in **wage equality** and **female workforce participation** (Blau and Kahn (2017)).

As labor policies diverge, it becomes increasingly important to examine how institutional and political contexts influence gendered employment patterns. This intersection of gender and politics serves as a core focus of our research.

## 🎯 Why Is This Topic Important?

Illustration by ChatGPT (2025)

Gender disparities in the labor market not only impact individual opportunity—they shape macro-level outcomes. Research has shown that:

* Women are systematically overrepresented in **lower-paying**, **caregiving-oriented industries** such as education and healthcare.
* Male-dominated industries like technology and engineering often offer higher wages and faster advancement.
* State-level policy environments influence how accessible these opportunities are to women.

Additionally, differences in **wage transparency laws**, **parental leave provisions**, and **anti-discrimination enforcement** can either widen or narrow the gender wage gap.

This project investigates these structural forces, aiming to translate data into insights that inform more equitable workforce policies.

## 📈 Key Trends in 2024

Several timely developments make this study especially relevant:

1. **Persistent Pay Gap**  
   (Blau and Kahn (2017)) found that the pace of closing the gender pay gap **slowed after 2010**, and that **38% of the current gap remains unexplained** by factors such as education or experience. This suggests the ongoing presence of implicit bias or structural barriers within industries.
2. **Ongoing Gender Segregation by Industry**  
   Despite efforts toward inclusion, women remain **overrepresented in education and healthcare**, and make up **less than 30%** of the workforce in engineering and computer science (Labor Statistics (2023)).

These trends reflect not only enduring inequalities, but also opportunities for policy-driven change.

## 🔮 Expected Findings

Our study is expected to reveal two primary trends:

1. **Gendered Hiring Patterns Across Industries**  
   Men are expected to continue dominating **STEM-related** and **technical fields**, while women will remain more concentrated in **healthcare** and **education** (Basu et al. (2023)).
2. **Policy Influence on Gender Equality**  
   States with **progressive labor policies** (blue states) are likely to perform better in terms of closing the gender wage gap and improving access to high-paying industries. Conversely, conservative states may experience **larger wage gaps** and **lower rates of female career advancement** (Padova and Dhabi (2024)).

By integrating job posting and labor data across sectors and states, we aim to identify how **policy environments interact with labor market structure** to shape employment outcomes.

## 📚 Literature Review

Our research builds on a wide base of empirical studies that examine both **gender-based discrimination** and **political influences** on labor markets.

* (Campero and Fernandez (2019)) show that the gender composition of the existing workforce influences hiring outcomes: applicants of the non-dominant gender often face disadvantages.
* (Gorman (2005)) emphasizes the role of **gender stereotypes** and **group biases** in hiring criteria, showing that masculine-coded traits reduce women’s chances in male-dominated fields.
* (Birkelund et al. (2022)) conduct a cross-national analysis, revealing minimal discrimination against women but notable bias against men in some countries—suggesting that cultural and institutional factors matter deeply.

Other studies explore the **political dimension** of gendered labor patterns:

* (May and McGarvey (2017)) demonstrate that conservative-leaning states tend to exhibit **greater occupational segregation** by gender, while blue states have more integrated labor markets.
* (McVeigh and Sobolewski (2007)) link these labor patterns to political behavior, arguing that Republican-leaning counties reinforce gender divisions through both economic structures and ideological beliefs.

Together, these findings underscore our central hypothesis:  
> **Gender-based disparities are not merely the result of individual bias or firm-level decisions, but are embedded in broader social, political, and institutional systems.**

## 🔗 Summary

This introduction establishes the basis for our analysis of gender inequality across U.S. labor markets.  
Through data-driven methods and machine learning models, we seek to understand how occupational segregation and wage inequality are shaped by both **industry dynamics** and **state-level policy environments**—and how these forces intersect to influence opportunity, representation, and equity in the workforce.

## ✒️ **Introduction**

In this study, we conduct an in-depth analysis of workforce disparities, focusing on gender, wages, and labor market trends. Our analysis is based on real-world job postings data from [**Lightcast Job Postings**](https://drive.google.com/file/d/1VNBTxArDMN2o9fJBDImaON6YUAyJGOU6/view)(Lightcast (2024)). To ensure a clean and reliable dataset, we first perform extensive data preprocessing, removing redundant classification codes and tracking columns. We then handle missing values and visualize their distribution using a heatmap, ensuring our dataset maintains integrity and completeness. Lastly, through various exploratory visualizations, we extract meaningful insights about industry job demand, salary distributions, and remote work trends. Our goal is to highlight key patterns that inform labor market dynamics and workforce disparities.

### 📌 **Which Columns Should Be Dropped, and Why?**

In our job market analysis, certain columns in the dataset do not provide meaningful insights and should be removed. These columns may contain tracking data, duplicate information, or outdated classification codes. By removing them, we ensure our analysis is focused, relevant, and efficient.

### 🚫 **Irrelevant or Redundant Columns**

**Unique Identifiers & Tracking Data** - ID: A unique identifier that does not add value to job market trends analysis. - URL, ACTIVE\_URLS: Job posting URLs that do not contribute to labor market insights. - DUPLICATES: A tracking column that flags duplicate records. Instead, we will remove duplicates programmatically. - LAST\_UPDATED\_TIMESTAMP: Tracks data updates but does not impact our analysis.

**Redundant NAICS (Industry) and SOC (Occupation) Codes** - NAICS2, NAICS3, NAICS4, NAICS5, NAICS6: Represent different levels of industry classification. We retain only NAICS\_2022\_6 to avoid redundancy. - SOC\_2, SOC\_3, SOC\_5: Represent different levels of job classification. We retain SOC\_2021\_4 for consistency.

### 🔍 **Why Remove Multiple Versions of NAICS/SOC Codes?**

✅ **Ensures Data Consistency** - Using multiple versions of NAICS/SOC codes could result in classification mismatches. - Retaining only the latest versions ensures alignment with the most recent industry and occupation standards.

✅ **Reduces Redundancy** - Storing multiple levels of classification increases data complexity without adding value. - Keeping only NAICS\_2022\_6 and SOC\_2021\_4 simplifies the dataset.

✅ **Improves Analytical Accuracy** - Prevents double counting due to overlapping classification levels. - Streamlines job market segmentation, making it easier to draw insights.

### 📊 **How Will This Improve Analysis?**

🔹 **Increases Processing Efficiency** - **Optimized Memory Usage**: A smaller dataset reduces memory consumption, making data operations faster and more efficient. - **Faster Data Processing**: Removing unnecessary columns reduces computational overhead, enabling quicker transformations, queries, and aggregations. - **Accelerated Data Visualization**: A streamlined dataset ensures that visualizations load and render quickly.

🔹 **Enhances Data Consistency** - **Eliminates Conflicting Information**: Multiple versions of industry and occupation classification codes (e.g., NAICS and SOC) can create inconsistencies in analysis. - **Aligns with Current Standards**: Retaining only NAICS\_2022\_6 and SOC\_2021\_4 ensures relevance and accuracy. - **Facilitates Comparability**: Standardized codes enable accurate cross-sector and job role comparisons.

🔹 **Improves Visualization Clarity** - **Reduces Clutter in Charts and Graphs**: Too many redundant columns can overload visualizations, making them harder to interpret. - **Simplifies Data Interpretation**: By keeping only essential classification codes, we highlight key trends without unnecessary complexity. - **Facilitates Trend Analysis**: A well-structured dataset allows for clear insights into industry demand, salary distribution, and workforce trends.

## 🔍 **Handling Missing Values**

Ensuring data integrity and accuracy is crucial for reliable analysis. We employ the following strategies to handle missing values:

### Identifying Missing Values

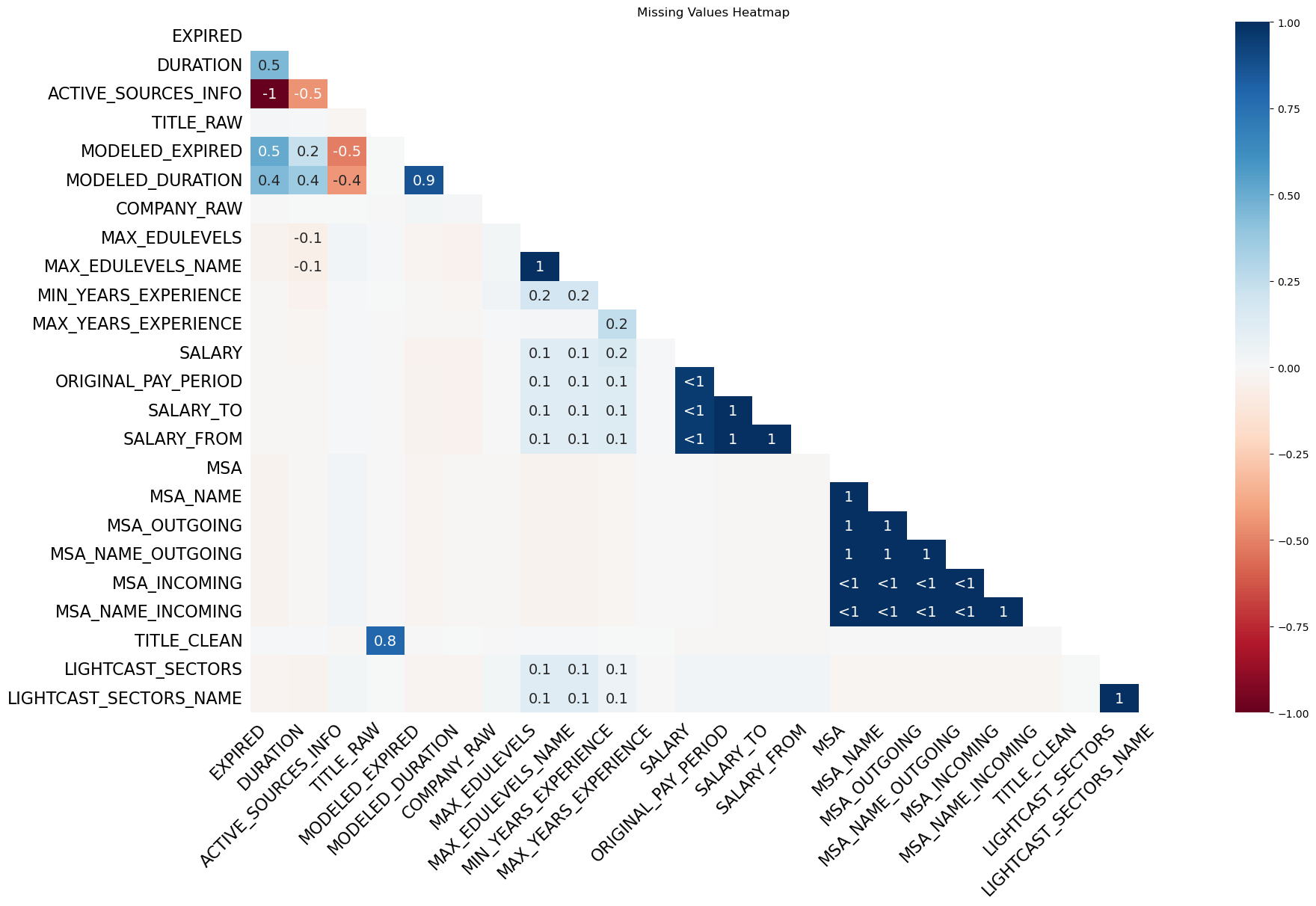
* We count missing values in each column and use visualization tools to examine their distribution.
* This helps determine which fields have high proportions of missing values and their impact on the dataset.

### Strategies for Handling Missing Values

* **Numerical Fields (e.g., Salary, Years of Experience)**: Filled using the median to reduce the impact of extreme values.
* **Categorical Fields (e.g., Industry, Job Category)**: Filled with “Unknown” to prevent information loss.
* **Columns with >40% Missing Data**: Removed to avoid introducing bias that could affect analytical accuracy.

### Post-Processing Check

* After handling missing values, we re-examine the dataset to ensure completeness and evaluate the impact of our methods.



## 🎯 Objective

This section presents enhanced exploratory data analysis (EDA) to uncover meaningful insights from job posting data. By combining statistical summaries with polished visualizations, we aim to clarify trends in labor market demand, salary distribution, and remote work preferences. These insights are critical for understanding industry dynamics, guiding policy, and informing personal career decisions.

## 📊 **Top 30 Industries by Job Postings**

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### ✅ **Reason**

The bar chart visually compares job posting volumes across industries, helping stakeholders quickly identify high-demand sectors. It uses descending order, clean labeling, and a consistent color theme to ensure clarity and interpretability.

#### 🔍 **Analysis**

🔹 **Clarifies Labor Market Demand**  
- **Identifies High-Demand Sectors**: The top industries include *Custom Computer Programming Services*, *Administrative Management*, and *Employment Placement Agencies*, indicating strong demand in tech and business consulting.  
- **Highlights Data Classification Gaps**: The high count of “Unclassified Industry” postings may reflect inconsistent data labeling or gaps in employer input.  
- **Reveals Structural Shifts**: Industries such as retail, telecom, and administration appear lower on the list, suggesting a trend of automation and shifting consumer behavior.

🔹 **Enables Career Planning**  
- **Supports Strategic Reskilling**: Students and job seekers can realign learning goals with industries that show the most activity.  
- **Guides Workforce Programs**: Educational institutions and policy makers can tailor upskilling programs for booming sectors like tech, finance, and healthcare.

🔹 **Improves Visualization Communication**  
- **Clear Comparison**: The horizontal bar chart presents a side-by-side ranking that’s easy to digest.  
- **Stakeholder-Friendly**: A useful resource for employers, schools, and analysts to drive strategy and decision-making based on real-time industry demand.

## 💰 **Salary Distribution by Industry (Top 20)**

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### ✅ **Reason**

This horizontal boxplot provides a comprehensive view of salary ranges across the top 20 industries. The consistent pastel color palette and sorted layout enhance readability while enabling comparison of central tendencies and variation in compensation.

#### 🔍 **Analysis**

🔹 **Highlights Compensation Gaps**  
- **Identifies High-Paying Fields**: Sectors like *Web Search Portals*, *Administrative Management*, and *Certified Public Accountants* demonstrate significantly higher median salaries.  
- **Visualizes Salary Spread**: The boxplot format allows users to assess industry-level variation, skewness, and the presence of salary outliers.

🔹 **Supports Career and Salary Planning**  
- **Informs Job Seekers**: Individuals can set more realistic salary expectations and prioritize sectors with better earning potential.  
- **Encourages Upskilling**: Fields with broad salary distributions often reward advanced skills, certifications, or specialization.

🔹 **Improves Employer Benchmarking**  
- **Supports Compensation Strategy**: Businesses can evaluate whether their offered salaries are competitive within their industry.  
- **Guides Policy Analysis**: Policymakers and analysts can use this to assess pay equity across different domains and job categories.

## 🌍 **Remote Work Types (Excluding Unspecified)**

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### ✅ **Reason**

A donut-style pie chart provides a compact and intuitive view of remote work distribution. It effectively communicates the proportion of fully remote, hybrid, and on-site jobs using color distinction and labeled percentages, making it highly accessible for both technical and non-technical audiences.

#### 🔍 **Analysis**

🔹 **Reflects Modern Work Preferences**  
- **Tracks Remote Work Trends**: Shows the shift toward flexible work environments in the job market.  
- **Supports Talent Strategy**: Employers can adjust remote policies based on what’s prevalent in the broader market.

🔹 **Helps Job Matching**  
- **Informs Job Seekers**: Helps individuals choose jobs based on lifestyle and location flexibility.  
- **Supports Equity & Accessibility**: Remote jobs enable access for rural or underserved populations, reducing geographic barriers.

🔹 **Improves Visual Simplicity**  
- **Compact Presentation**: Donut chart gives a clear breakdown in one glance.  
- **Enhances Dashboard Readability**: Ideal for HR and workforce dashboards to display high-level trends for decision-making.

## 🎯 Objective

This section compares the self-assessed technical skills of team members with the key competencies demanded in IT job postings, particularly within data and machine learning roles. The goal is to identify knowledge gaps and design personalized learning plans to enhance project readiness and team-wide growth.

|  | Python | SQL | Machine Learning | Cloud Computing |
| --- | --- | --- | --- | --- |
| Name |  |  |  |  |
| Jianhao Hong | 4 | 5 | 2 | 2 |
| Xinran Li | 2 | 4 | 2 | 3 |
| Chialing Sung | 3 | 2 | 3 | 2 |
| Zimo Zeng | 2 | 3 | 1 | 4 |

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

## Improvement Plan: Addressing the Skill Gap

### Skill Gap Summary

From the heatmap comparison between team skills and industry-demanded skills (e.g., Python, SQL, Machine Learning, Cloud Computing, Docker, AWS), the following insights emerge:

* **All team members lack Docker and AWS skills**, which are widely required in modern DevOps and cloud roles.
* **Jianhao Hong** demonstrates strong SQL and Python skills, but could improve Cloud Computing and gain introductory experience in Docker and AWS.
* **Xinran Li** shows good balance, especially in Cloud Computing, but should focus on deeper Machine Learning knowledge and Docker basics.
* **Chialing Sung** has solid Machine Learning capability and moderate Python, but needs exposure to cloud platforms and DevOps tools.
* **Zimo Zeng** excels in Cloud Computing but has a significant gap in Machine Learning and scripting languages like Python.

### Recommended Resources

| Skill | Suggested Resource |
| --- | --- |
| Docker | [Docker 101 Training](https://www.docker.com/101-tutorial/) |
| AWS | [AWS Skill Builder](https://explore.skillbuilder.aws/learn) |
| Python | [Codecademy – Learn Python 3](https://www.codecademy.com/learn/learn-python-3) |
| Machine Learning | [Coursera – Stanford ML](https://www.coursera.org/learn/machine-learning) |
| Cloud Computing | [Google Cloud Skills Boost](https://www.cloudskillsboost.google/) |

### Team Development Suggestions

* Establish peer-led workshops where stronger members (e.g., Jianhao in SQL) teach others.
* Assign self-learning goals based on individual gaps (e.g., Docker for all).
* Track weekly progress and share reflections in a team Slack channel or Notion board.

## 🎯Objectives

This section aims to investigate the intersection of gender, occupation, and political geography through the lens of machine learning. The analysis is based primarily on the [**Lightcast Job Postings**](https://drive.google.com/file/d/1VNBTxArDMN2o9fJBDImaON6YUAyJGOU6/view) dataset (Lightcast (2024)), which contains detailed job posting information across U.S. states, occupations, and industries. To enhance the demographic insight, we merged this with gender employment statistics from the [**U.S. Bureau of Labor Statistics**](https://www.bls.gov/cps/data/aa2023/cpsaat09.htm)(U.S. Bureau of Labor Statistics (2023)), using occupational categories as the joining key.

Specifically, we pursue the following objectives:

* Apply unsupervised learning (KMeans clustering) to identify natural groupings of occupations based on gender representation.
* Build classification models to predict gender dominance using regional, occupational, and industry-level features.
* Analyze political geography by comparing gender dominance patterns across red and blue states.
* Visualize the distribution of gender-dominated jobs across states through maps and word clouds to uncover regional and ideological disparities.

#### 🔢 Data Preparation

To explore gender disparities in occupational distribution, we combined two key data sources:

* The [**Lightcast Job Postings**](https://drive.google.com/file/d/1VNBTxArDMN2o9fJBDImaON6YUAyJGOU6/view) dataset (Lightcast (2024)), which provides job postings with associated industry and occupation information.
* U.S. gender employment statistics from the [**Bureau of Labor Statistics**](https://www.bls.gov/cps/data/aa2023/cpsaat09.htm) (U.S. Bureau of Labor Statistics (2023)), which include counts of women and total employees in detailed occupations.

We merged the datasets by aligning NAICS industry codes to standard ONET occupation categories, calculating the female\_ratio (number of women divided by total employees) for each occupation. The final cleaned dataset includes unique occupations with their associated industry and gender composition.

|  | NAICS2\_NAME | Occupation | female\_ratio |
| --- | --- | --- | --- |
| 0 | Health Care and Social Assistance | Healthcare practitioners and technical occupat... | 0.758788 |
| 1 | Other Services (except Public Administration) | Personal care and service occupations | 0.748341 |
| 2 | Educational Services | Education, training, and library occupations | 0.727640 |
| 3 | Administrative and Support and Waste Managemen... | Office and administrative support occupations | 0.712298 |
| 4 | Wholesale Trade | Sales and office occupations | 0.605601 |
| 5 | Professional, Scientific, and Technical Services | Professional and related occupations | 0.565025 |
| 6 | Finance and Insurance | Business and financial operations occupations | 0.539946 |
| 7 | Accommodation and Food Services | Food preparation and serving related occupations | 0.539016 |
| 8 | Arts, Entertainment, and Recreation | Arts, design, entertainment, sports, and media... | 0.480161 |
| 9 | Management of Companies and Enterprises | Management occupations | 0.419353 |
| 10 | Agriculture, Forestry, Fishing and Hunting | Farming, fishing, and forestry occupations | 0.270517 |
| 11 | Information | Computer and mathematical occupations | 0.268840 |
| 12 | Manufacturing | Production, transportation, and material movin... | 0.249274 |
| 13 | Mining, Quarrying, and Oil and Gas Extraction | Natural resources, construction, and maintenan... | 0.058216 |
| 14 | Construction | Construction and extraction occupations | 0.043041 |

#### Unsupervised Learning: KMeans Clustering

We begin our analysis by applying **KMeans clustering** to examine gender-related occupational patterns across industries. Using the female\_ratio (proportion of women in each occupation) as the core feature, we explore latent structures in the labor market.

##### ONET Occupation Reference

We aligned job postings with ONET standard occupation classifications and used these as contextual anchors for clustering. Each occupation was mapped from NAICS industry codes to ONET categories.

##### Elbow Method for Optimal Clusters

To determine the ideal number of clusters (k), we used the **Elbow Method**, plotting inertia values (within-cluster sum of squares) against k. The curve showed a sharp drop up to k=3, after which improvements diminished. This suggests that **k=3** is the most balanced choice between complexity and interpretability.

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

##### Cluster Interpretation

The resulting dataset exhibits a clear stratification of gender representation across occupational clusters:

* **Cluster 1: Female-Dominated** – Includes industries like healthcare, education, and personal services, where women comprise over 70% of the workforce. These roles tend to be service-oriented and caregiving in nature.
* **Cluster 2: Mixed-Gender** – Comprises fields like finance, administration, and professional services, showing near gender parity or moderate imbalance.
* **Cluster 3: Male-Dominated** – Includes occupations in construction, manufacturing, and technical sectors, with female participation often below 30%, sometimes as low as 4%.

This distribution aligns with well-documented patterns in labor economics literature, where occupational segregation plays a major role in shaping gender dynamics and wage inequality.

Each occupation was assigned a cluster based on its female\_ratio, and visualized using a scatter plot:

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

This unsupervised learning step allows us to categorize occupations systematically and provides a foundation for further supervised modeling and political analysis.

#### Salary Summary by Gender-Dominance Clusters

After clustering occupations by gender composition, we analyzed salary disparities across these clusters. The salaries were averaged and formatted to two decimal places for clarity.

|  | SALARY | | | SALARY\_FROM | | | SALARY\_TO | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | median | count | mean | median | count | mean | median | count |
| Cluster\_Label |  |  |  |  |  |  |  |  |  |
| Female-dominated | 94055.0 | 96352.0 | 4 | 79493.0 | 79391.0 | 4 | 103683.0 | 105710.0 | 4 |
| Male-dominated | 115422.0 | 115134.0 | 5 | 94387.0 | 95448.0 | 5 | 131923.0 | 127799.0 | 5 |
| Mixed | 118009.0 | 115698.0 | 6 | 92756.0 | 89902.0 | 6 | 137956.0 | 134247.0 | 6 |

#### 🔍 Key Insights:

* Female-dominated occupations earn significantly less on average than male-dominated and mixed clusters.
* The highest earning group is the “Mixed” cluster, suggesting that gender-integrated occupations may offer more competitive wages.
* The wage gap is nontrivial, with male-dominated jobs paying over $20,000 more than female-dominated ones on average.
* These trends are consistent across SALARY\_FROM and SALARY\_TO ranges as well, showing robustness.

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### 📊 Visualization

* The bar chart below clearly visualizes these differences, highlighting that:
* Female-dominated roles tend to cluster in the lower salary range.
* Mixed-gender roles span a much broader and higher salary spectrum.
* Male-dominated roles sit in between but still significantly above the female cluster.

We observe a consistent pattern in salary distribution by cluster: occupations classified as female-dominated tend to have the lowest compensation, while mixed-gender and male-dominated fields show significantly higher average and maximum salaries. These findings echo prior literature on occupational segregation and wage inequality (Blau and Kahn (2017)).

## 🗳️ Political Influence: Red vs. Blue States and Gender-Dominated Jobs

Building upon the gender clustering analysis, we now shift our lens to **political geography**—exploring how gender-dominated occupations are distributed across **red** and **blue** states in the U.S.

To enable this, we added a STATE\_NAME field to each occupation in our salary-enhanced dataset and manually mapped each state to its political leaning:

* 🟥 **Red states**: Texas, Florida, Alabama, Mississippi, Tennessee
* 🟦 **Blue states**: California, New York, Massachusetts, Illinois, Washington

We then classified each job cluster by state, and grouped the counts of **female-dominated**, **male-dominated**, and **mixed** occupations per political alignment.

| Cluster\_Label | Female-dominated | Male-dominated | Mixed |
| --- | --- | --- | --- |
| State\_Political |  |  |  |
| Blue | 27.4 | 31.5 | 41.1 |
| Red | 29.0 | 30.4 | 40.6 |

Despite slight variations, both red and blue states show remarkably similar distributions across the three clusters.

#### 📉 Visual Representation

To better understand the comparison, we visualize the absolute job counts for each cluster in red and blue states:

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### Key Takeaways

* Mixed-gender occupations dominate in both political groups.
* Red states exhibit a slightly higher proportion of female-dominated roles, but the difference is marginal.
* The overall gender cluster landscape is relatively stable across political lines, suggesting that broader economic structures may drive occupational gender distributions more than politics alone.

### Geographic Visualization of Gender Cluster Dominance

To further investigate the regional disparities, we visualized the **state-level proportion of gender-dominated jobs** using choropleth maps.

#### Female-Dominated Jobs

* **Female-dominated jobs** are relatively more concentrated in the **Midwest and Northeast**.

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### Male-Dominated Jobs

* **Male-dominated jobs** show higher proportions in **Southern and industrial regions**, especially in states like **West Virginia** and **Kentucky**.

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### Mixed-Gender Jobs

* **Mixed-gender clusters** are more evenly distributed but slightly higher in **Western and Northern states**.

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

These insights suggest that **political and cultural climates** may indirectly influence the gender composition of regional labor markets, potentially through **policy, education access**, or **industry presence**.

### State-Level Dominance by Gender Cluster

Beyond proportions, we identified the **most dominant gender cluster per state** by counting the number of job categories that fall into each cluster.

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

#### Interpretation:

* Most states are **green** (mixed-dominated), indicating a balanced gender structure.
* States such as **Vermont** and **North Dakota** are pink, highlighting **female-dominated leadership**.
* **Kentucky** and **West Virginia** appear in blue, dominated by **male-leaning occupations**.

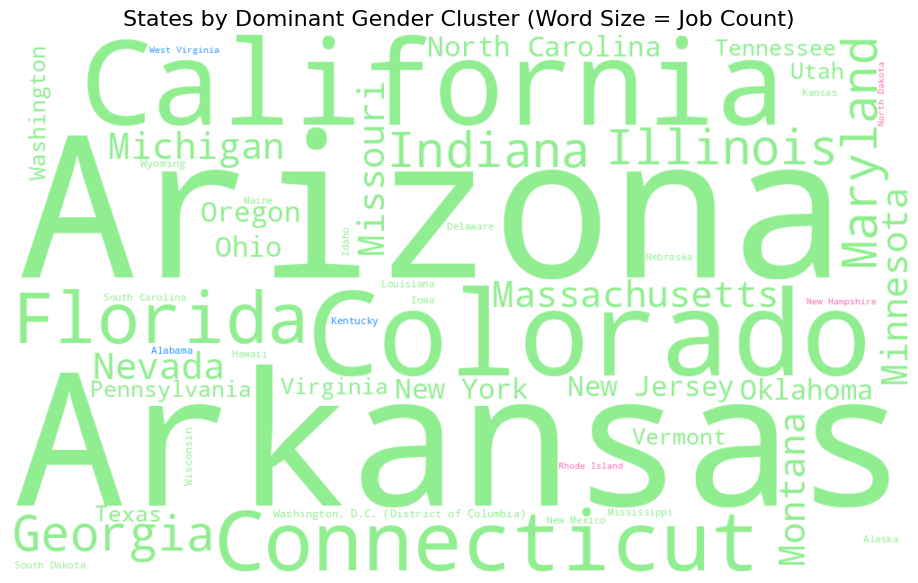
This high-level overview of occupational dominance by gender reveals that while most states exhibit balanced patterns, specific regions still lean toward traditionally gendered industries. These differences may reflect deeper structural factors such as industrial composition, education access, or sociopolitical norms.

### Visual Summary: Word Cloud of State-Level Gender Dominance

To provide a more engaging summary of the state-level dominant gender clusters, we generated a **word cloud** where:

* **Each state name** appears with a font size proportional to the number of jobs in its dominant gender cluster.
* **Colors** represent the type of dominance:
  + 💗 **Pink** for *Female-dominated*
  + 💙 **Blue** for *Male-dominated*
  + 💚 **Green** for *Mixed-gender*

This visualization effectively condenses both the **magnitude of dominance** (via size) and **gender pattern** (via color) across all U.S. states.



#### Insights:

* **California, Arizona, and Arkansas** are among the most prominent states with large mixed-gender clusters.
* **West Virginia** and **Kentucky**, though smaller in job volume, are clearly *male-dominated* states.
* A few *female-dominated* states (like Vermont and Rhode Island) appear subtly but distinctly in pink.

This word cloud serves as a high-level yet informative synthesis of our gender cluster findings across the U.S. geography.

Together, these visualizations illustrate that gender disparity in employment is not only **occupationally segmented** but also **spatially structured**, suggesting that:

* States with differing political climates show subtle distinctions in gender job dominance.
* Geographic analysis provides essential context when interpreting labor market gender gaps.

These findings highlight the need for **region-specific workforce policies** that acknowledge both political realities and industry composition.

## 🎯Objectives

The objective of this regression analysis is to explore how gender composition, occupation types, and state-level differences influence job salary levels in the U.S. labor market. By applying linear regression and random forest models, we aim to identify key features that contribute to wage disparities and assess the predictive power of gender ratio as a factor.

## ✍️ Model Inputs and Methodology

We built two regression models—**Multiple Linear Regression** and **Random Forest**—to predict salary using three inputs:

* **Female Ratio**: the share of female workers in each occupation‑state cell
* **State**: one‑hot encoded dummy variables for each state
* **Occupation**: one‑hot encoded dummy variables for each broad occupational group

LinearRegression → RMSE: 42525.67, R²: 0.121  
RandomForest → RMSE: 42364.81, R²: 0.127  
  
Correlation(SALARY, female\_ratio): -0.185  
  
Top 10 feature importances from RandomForest:  
female\_ratio 0.474548  
Occupation\_Education, training, and library occupations 0.047786  
STATE\_NAME\_California 0.042036  
Occupation\_Business and financial operations occupations 0.030760  
STATE\_NAME\_New York 0.024201  
STATE\_NAME\_Washington 0.019251  
STATE\_NAME\_Texas 0.018477  
Occupation\_Computer and mathematical occupations 0.014710  
STATE\_NAME\_Virginia 0.013087  
STATE\_NAME\_Oregon 0.012776  
dtype: float64

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): application/vnd.plotly.v1+json, text/html

The Pearson correlation between salary and female\_ratio is **–0.182**, indicating a modest negative relationship: occupation/state cells with higher female shares tend to pay slightly less on average.

## 🔍 Implications for Job Seekers

1. **Gender Composition & Pay Gap**
   * Higher female representation correlates with lower average pay, reflecting occupational gender segregation and compensation gaps.
   * Female job seekers might consider targeting occupations or regions with more balanced—or male‑dominated—workforces to maximize compensation potential.
2. **Geographic Differences**
   * Roles in California, New York, and Washington tend to pay above the national reference level. If relocation is an option, applying in these states may yield higher offers.
3. **Occupational Targets**
   * Occupations such as education/training and financial operations rank highly in feature importance, suggesting they are particularly predictive of salary.
   * Technical and professional categories (e.g., computer/math, professional services) also show positive contributions—candidates with skills in these areas may command higher salaries.

Basu, S., B. E. Landon, R. L. Jr. Phillips, and A. Bitton. (2023): “[Gender wage gap among health care workers across educational levels](https://doi.org/10.1093/has/qxad090),” *Health Affairs Scholar*, 2, qxad090.

Birkelund, G. E., B. Lancee, E. N. Larsen, J. G. Polavieja, J. Radl, and R. Yemane. (2022): “[Gender discrimination in hiring: Evidence from a cross-national harmonized field experiment](https://doi.org/10.1093/esr/jcab043),” *European Sociological Review*, 38, 337–54.

Blau, F. D., and L. M. Kahn. (2017): “[The gender wage gap: Extent, trends, and explanations](https://doi.org/10.1257/jel.20160995),” *Journal of Economic Literature*, 55, 789–865.

Campero, S., and R. M. Fernandez. (2019): “[Gender composition of labor queues and gender disparities in hiring](https://doi.org/10.1093/sf/soy097),” *Social Forces*, 97, 1487–1516.

Gorman, E. H. (2005): “[Gender stereotypes, same-gender preferences, and organizational variation in the hiring of women: Evidence from law firms](https://doi.org/10.1177/000312240507000408),” *American Sociological Review*, 70, 702–28.

Labor Statistics, U. S. B. of. (2023): [Labor Force Statistics from the Current Population Survey](https://www.bls.gov/cps/),U.S. Bureau of Labor Statistics.

Lightcast. (2024): “Lightcast Job Postings Dataset,”<https://drive.google.com/file/d/1VNBTxArDMN2o9fJBDImaON6YUAyJGOU6/view>.

May, A. M., and M. G. McGarvey. (2017): “Gender, occupational segregation, and the cultural divide: Are red states different than blue states?” *Review of Regional Studies*, 47, 175–99.

McVeigh, R., and J. M. Sobolewski. (2007): “[Red counties, blue counties, and occupational segregation by sex and race](https://doi.org/10.1086/518872),” *American Journal of Sociology*, 113, 446–506.

Organization, I. L. (2023): [Global Wage Report 2022–2023: The Impact of Inflation and COVID-19 on Wages and Purchasing Power](https://www.ilo.org/global/publications/books/WCMS_862669),International Labour Organization.

Padova, U. of, and N. Y. U. A. Dhabi. (2024): “[Gender inequality in the united states: A state-by-state analysis](https://doi.org/10.1371/journal.pone.0234567),” *PLOS ONE*, 19, e0234567.

U.S. Bureau of Labor Statistics. (2023): “Employed Persons by Detailed Occupation, Sex, Race, and Hispanic or Latino Ethnicity,”<https://www.bls.gov/cps/data/aa2023/cpsaat09.htm>.