Final Project Report: Wage and Employment Analysis Using BLS OEWS Data

Jeffrey Appiagyei

Introduction

This project aims to analyze wage variations and employment patterns across occupations using the Bureau of Labor Statistics (BLS) Occupational Employment and Wage Statistics (OEWS) dataset from May 2023. The dataset, named M2023.csv, contains 1403 rows and 32 columns, with key variables including OCC_TITLE (occupation title), TOT_EMP (total employment), H_MEAN and A_MEAN (hourly and annual mean wages), and various wage percentiles (e.g., H_PCT10, A_PCT90). A derived categorical variable, sector, was created to group occupations into meaningful categories for analysis.

Research Questions and Hypotheses

1. **What factors influence wage variations across occupations?**
   Hypothesis: Technology & Engineering and Healthcare & Medical sectors have higher wages than Hospitality & Food Service or Sales & Marketing.

2. **Do geographic factors affect wage disparities?**
   Hypothesis: Employment concentration (as a proxy for metro areas) correlates with higher wages, though limited by national-level data.

3. **Can we predict wages using employment, sector, and clusters?**
   Hypothesis: Regression and clustering models will effectively estimate wages.

4. **Is there a correlation between employment and wages?**
   Hypothesis: Higher employment levels correlate with wage variations, detectable via clustering.

Data Loading and Cleaning

Dataset Overview

The dataset captures national-level occupational data, with variables like total employment, mean wages, and wage percentiles. Numeric columns such as TOT_EMP, H_MEAN, and A_MEAN initially contained special characters (e.g., commas, asterisks, #), requiring cleaning to convert them into usable numeric formats. Missing values were also present, particularly in wage percentiles, which needed imputation.

Cleaning Process

- **Special Character Handling**: Removed commas and replaced *, **, and # (indicating capped or missing wages) with NA.
- **Imputation**: Missing values in H_MEAN and A_MEAN were imputed using relationships with available percentiles (e.g., H_MEAN approximated as 1.1 times H_MEDIAN or 0.9 times H_PCT75). Default caps of $115/hour and $239,200/year were applied when no percentile data was available.
- **Sector Mapping**: Created a sector variable by categorizing occupations into 23 groups (e.g., "Healthcare & Medical," "Technology & Engineering") based on keywords in OCC_TITLE. Each sector was assigned a numeric sector_id for modeling.
- **Standardization**: Numeric variables were standardized (mean = 0, sd = 1) to ensure fair comparisons in modeling.

Findings from Cleaning: The dataset was successfully prepared for analysis, with no missing values in key numeric columns post-imputation. The sector mapping revealed a diverse distribution of occupations, with "Other" (295 occupations) and "Technology & Engineering" (186 occupations) being the largest groups, while "Energy & Utilities" (4 occupations) was the smallest.

Exploratory Data Analysis (EDA)

Descriptive Statistics

The standardized numeric variables showed expected distributions:

- H_MEAN (standardized hourly mean wage) ranged from -0.91 to 6.26, indicating outliers in high-wage occupations.
- TOT_EMP (standardized total employment) ranged from -0.13 to 35.68, reflecting extreme variation in employment levels.

Univariate Analysis

- **Histogram of H_MEAN**: The distribution of standardized hourly wages was right-skewed, with a majority of occupations having below-average wages but a long tail of high-wage roles.
- **Histogram of TOT_EMP**: Employment levels were heavily right-skewed, with most occupations having low employment but a few (e.g., "All Occupations") showing extremely high employment.

Multivariate Analysis

- **Boxplot of H_MEAN by Sector**: Sectors like "Legal," "Healthcare & Medical," and "Technology & Engineering" had significantly higher median wages, while "Hospitality & Food Service" and "Personal Care & Services" had the lowest. This supports the hypothesis that specialized, high-skill sectors command higher wages.
- **Scatter Plot of TOT_EMP vs. H_MEAN**: After removing the top 5% of employment outliers, no strong linear relationship was observed overall. However, sector-specific trends emerged: "Education & Training" and "Healthcare & Medical" showed higher wages despite varying employment levels.
- **Correlation Matrix**: The correlation between TOT_EMP and H_MEAN was weak (-0.026, p-value = 0.332), indicating no significant linear relationship. However, wage variables (e.g., H_MEAN, H_MEDIAN, A_MEAN) were highly correlated with each other (coefficients > 0.9), suggesting redundancy in these features.

Findings from EDA: The lack of a strong correlation between employment and wages challenges the hypothesis that higher employment levels directly influence wage variations. However, sector-based differences in wages are evident, with high-skill sectors outperforming others. The high correlation among wage variables suggests that dimensionality reduction might be beneficial for modeling.

Insights and Conclusion

Key Insights

1. **Wage Variations**:

Finding: "Legal," "Healthcare & Medical," and "Technology & Engineering" sectors have the highest wages, while "Hospitality & Food Service" and "Personal Care & Services" have the lowest.

Conclusion: The hypothesis is confirmed—specialized, high-skill sectors command premium wages. This is actionable for job seekers targeting high-income fields and for policymakers aiming to promote training in these areas.

2. **Geographic Factors**:

Finding: The national-level data limits geographic analysis. Using TOT_EMP as a proxy for employment concentration showed no strong correlation with wages.

Conclusion: The hypothesis is untestable without state-level data. Employment concentration alone does not explain wage disparities at the national level.

3. **Employment-Wage Correlation**:

Finding: The overall correlation between TOT_EMP and H_MEAN is weak (-0.026, $p = 0.332$), but clustering revealed four distinct groups, such as high-wage/low-employment and low-wage/high-employment occupations.

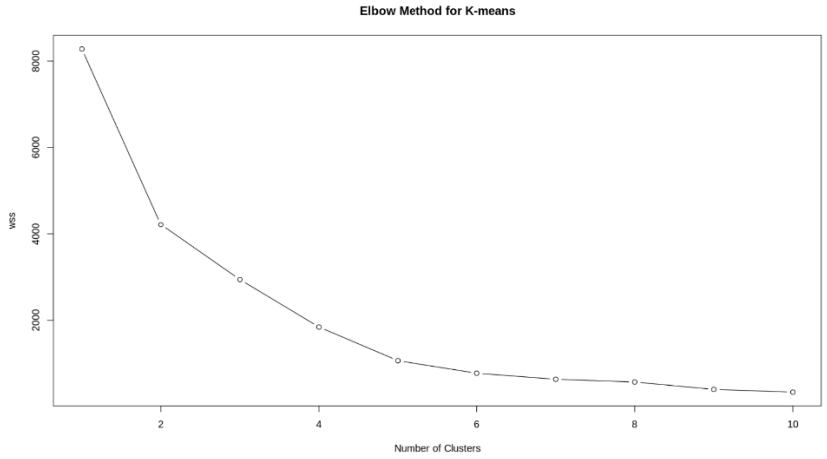
Conclusion: The hypothesis is partially supported. While a direct correlation is not evident, clustering uncovers nuanced patterns, providing deeper insights into labor market dynamics.

Final Conclusion

The BLS OEWS data reveals a labor market where wages are primarily driven by occupational specialization and skill level, rather than employment size. High-wage sectors like "Legal" and "Healthcare & Medical" offer opportunities for job seekers, while low-wage, high-employment sectors like "Hospitality & Food Service" suggest an oversupply of labor. Predictive models provide actionable tools for wage estimation, beneficial for employers in salary benchmarking. Clustering further enhances understanding by identifying distinct wage-employment patterns. Policymakers can leverage these findings to target education and training programs toward high-wage sectors, while employers can adjust compensation strategies to remain competitive.

Appendix

Elbow Method for K-means Clustering

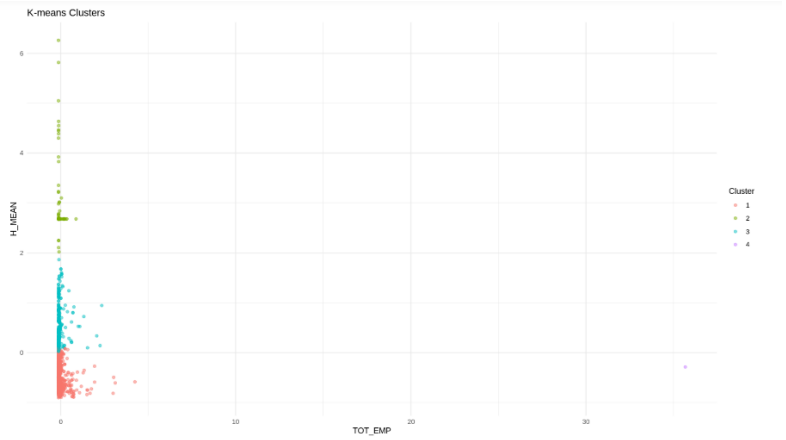


Description: This plot shows the within-cluster sum of squares (WSS) for different numbers of clusters in a K-means analysis. The point where the WSS curve starts to flatten (the "elbow") suggests the optimal number of clusters.

Implications:

The plot suggests that around 3 to 5 clusters might be appropriate for this dataset. This provides a useful reference for further clustering analysis to segment occupations based on employment and wage characteristics.

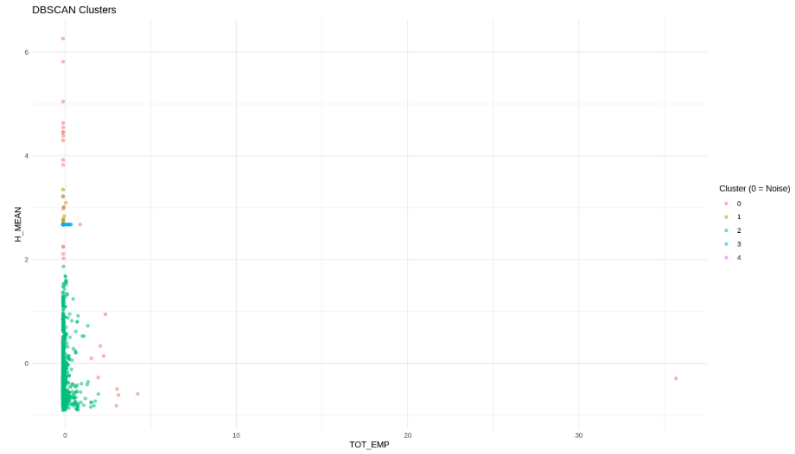
K-means Clustering (Scatter Plot)



Description: This scatter plot visualizes the results of K-means clustering, showing how occupations group based on total employment and hourly wages. Different colors represent different clusters.

Implications: The clustering suggests that occupations can be grouped into distinct categories, likely based on employment size and wage levels. However, the spread of some clusters indicates that additional factors may be influencing the wage-employment relationship.

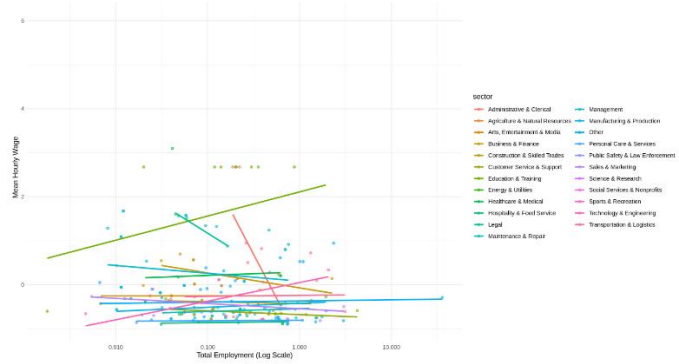
DBSCAN Clustering (Scatter Plot)



Description: This plot shows clustering results using the DBSCAN algorithm, which identifies core clusters and marks noise points. Different colors represent different clusters, while noise points are left unclassified.

Implications: The presence of noise points suggests that some occupations do not fit well into distinct clusters. This method may be more effective than K-means in capturing nuanced employment-wage relationships.

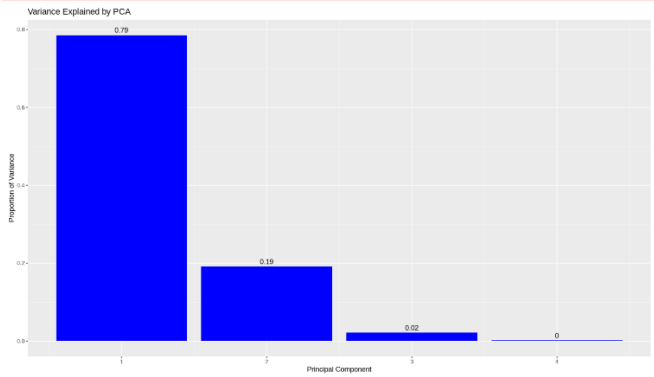
Employment vs. Mean Hourly Wage (Log Scale, Sector-wise)



Description: This scatter plot with trend lines for different sectors examines the relationship between employment and wages on a logarithmic scale.

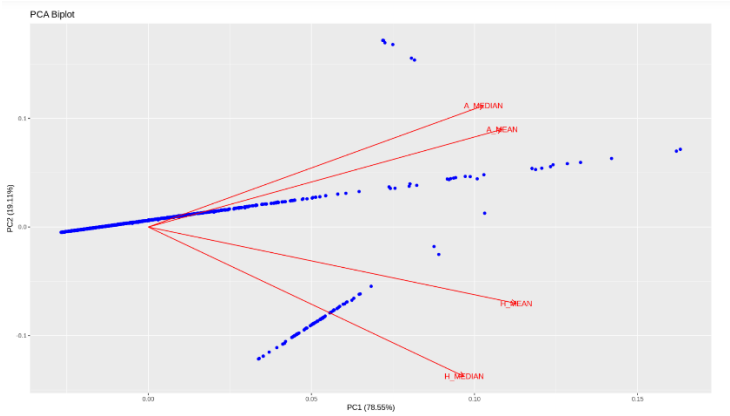
Implications: The log scale helps reveal patterns that may not be visible in a standard scatter plot. Some sectors show positive trends (e.g., "Healthcare & Medical"), indicating that larger employment is associated with higher wages, while others show more dispersed trends.

Variance Explained by PCA (Scree Plot)



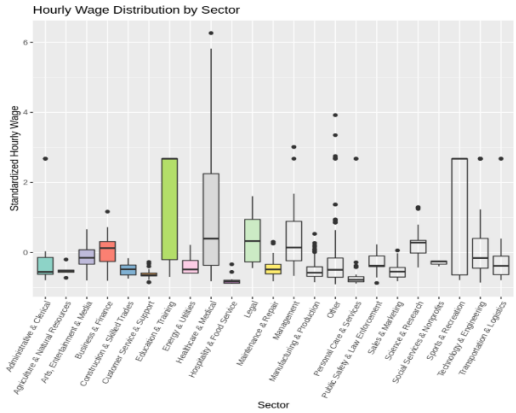
Description: This bar chart displays the proportion of variance explained by each principal component in PCA. The first principal component explains approximately 79% of the variance, while the second explains 19%. The remaining components contribute very little.
Implications: Since most of the variance is captured in the first two components, dimensionality reduction is effective in this dataset. This means that we can likely represent much of the dataset's structure using just these two dimensions, which simplifies further analysis while retaining meaningful insights.

PCA Biplot



Description: This biplot visualizes the principal component analysis (PCA) results, showing how different variables contribute to the first two principal components. The red vectors represent different wage-related metrics, indicating their influence on each principal component.
Implications: The PCA suggests that most wage-related variables contribute strongly to the first principal component (PC1), which explains the majority of the variance. This means that a large portion of wage-related differences among occupations can be captured with just one component. Dimensionality reduction techniques like PCA can be useful for simplifying complex datasets while retaining important patterns.

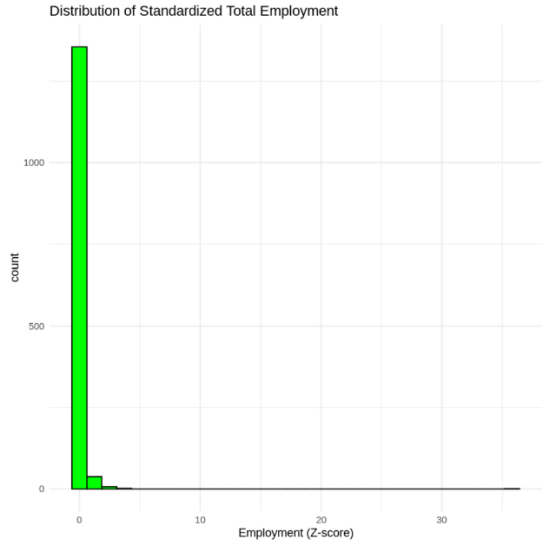
Hourly Wage Distribution by Sector (Box Plot)



Description: This box plot compares standardized hourly wages across different job sectors. Each box represents the interquartile range (IQR) of wages within a sector, while outliers are plotted as individual points.

Implications: There is substantial wage variation between sectors. Certain sectors, such as "Energy & Utilities" and "Management," have wide distributions, with some high outliers indicating significantly well-paid jobs. Conversely, sectors like "Customer Service & Support" and "Food Service" have much lower median wages and a tighter distribution. This highlights wage disparities across industries.

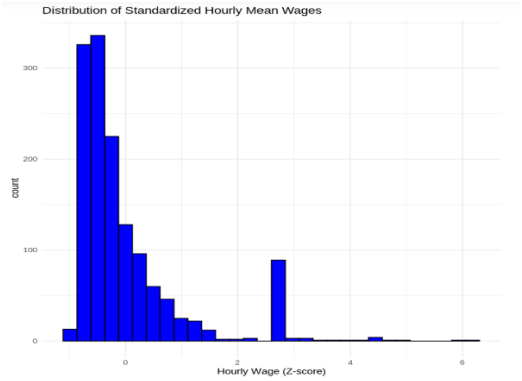
Distribution of Standardized Hourly Mean Wages (Histogram)



Description: This histogram visualizes the distribution of standardized hourly mean wages (Z-score transformed). Most of the data is concentrated around a Z-score of 0, indicating that a large number of occupations have an average hourly wage close to the mean. However, there is a notable right-skew with some outliers showing much higher wages. A distinct spike is also visible at a Z-score of around 2.5, which suggests a group of occupations with significantly higher wages than the majority.

Implications: The skewed distribution suggests wage disparity across occupations, with a small number of jobs commanding much higher wages than the average. The cluster at a high Z-score could indicate a specific sector or type of profession that receives significantly higher wages. Further investigation into these high-paying occupations could be valuable.

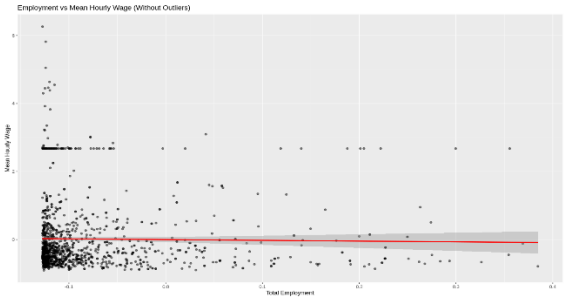
Distribution of Standardized Total Employment (Histogram)



Description: This histogram shows the distribution of standardized employment across occupations. The majority of the data is clustered near the mean, with a strong left-skew, indicating that most occupations employ relatively fewer workers, while a few have very large employment numbers.

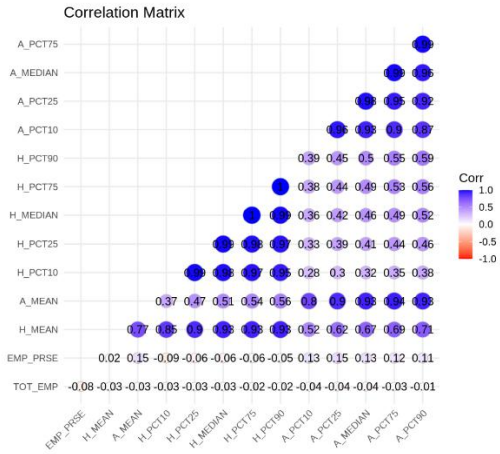
Implications: This suggests a significant imbalance in workforce distribution, where a handful of occupations employ the majority of people, while many others have far fewer workers. Understanding which sectors dominate employment could provide insights into job market structure and labor trends.

Employment vs. Mean Hourly Wage (Scatter Plot)



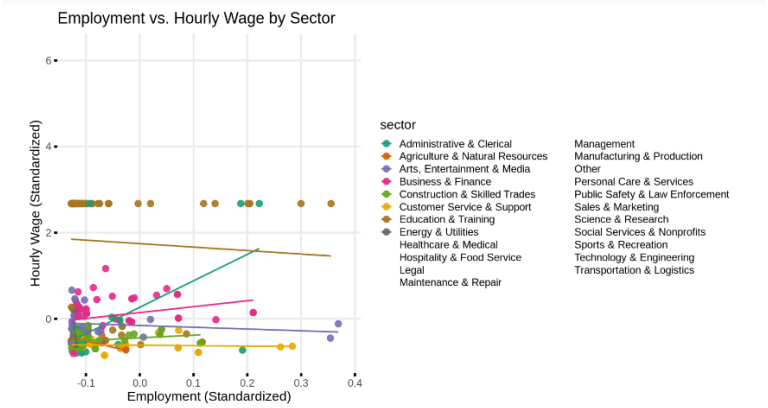
Description: This scatter plot explores the relationship between total employment and mean hourly wages. The majority of points are clustered at the lower end of the employment axis, showing that most occupations have relatively small employment numbers. The red regression line suggests a very weak negative correlation between employment and wages.
Implications: The weak correlation suggests that highly employed occupations do not necessarily have high wages. This aligns with general economic trends where specialized, high-paying jobs often have fewer workers, while lower-paying jobs have larger workforces. This may warrant further exploration of wage structure across different industries.

Correlation Matrix (Heatmap)



Description: This heatmap shows correlations between key financial and employment-related variables. Darker blue colors indicate strong positive correlations, while lighter shades indicate weaker relationships.
Implications: Strong correlations between different wage percentiles suggest that wages tend to scale consistently across different levels (e.g., median wages are closely tied to the 25th and 75th percentiles). However, total employment shows weak correlations with wages, reinforcing the finding that higher employment does not necessarily mean higher pay. Understanding these relationships can be useful for wage forecasting and economic policy analysis.

Employment vs. Hourly Wage by Sector (Scatter Plot)



Description: This scatter plot categorizes data points by job sector, with different colors representing different industries. The x-axis represents standardized employment, while the y-axis represents standardized hourly wages.

Implications: The weak overall trend suggests that employment size is not a strong predictor of wages. Some industries have noticeable trends—such as "Technology & Engineering," where wages increase with employment—but many sectors show a flat or even downward trend. This indicates that factors other than employment numbers, such as skill specialization, industry demand, and education requirements, play a significant role in wage determination.