Golden Gate University

Master of Science in Business Analytics

Final Term Paper

MSBA 320 – Advanced Business Analytics using python

Analyzing Employment and Wage Disparities Across U.S. Occupations Using BLS Data

Submitted by:

Prem Kumar Thummala

Professor:

Dr. Siamak Zadeh

Date of Submission:

26th April 2025

Table of Contents

Abstract

The analysis studies American occupational sector wage and employment variations through an examination of United states Bureau of Labor Statistics' (BLS) May 2024 Occupational Employment and Wage Statistics (OEWS) dataset. This work aims to reveal wage distribution patterns as it investigates employed workers' income distribution relating to employment size and establishes the statistical relevance of occupational classifications to wage outcomes.

A thorough research method involved both descriptive statistics and correlation matrices with linear regression modeling together with ANOVA testing. The research findings demonstrate that wage determination depends heavily on the continuous classification of occupational fields yet employment numbers show moderately weak predictive value. The research relied on visual representation tools including histograms, boxplots and bar charts and heatmaps to improve understanding of the data.

Labor rates differ primarily because of how occupations are organized rather than how many companies hire people according to the study. The study presents practical guidance to policy experts and education officials and business executives who need to create equal pay structures and develop workforce analytics strategies and support career direction decisions.

1. Introduction

Wage Labor market dynamics together with income inequality assessment and macroeconomic planning require wage analysis methods for their comprehension. Wage analysis assists governmental decision-making about labor supply versus demand while businesses select talent recruits and employees make career decisions and plan their compensations. Wage equality stands as the foundation of sustainable economic performance and social justice because it controls consumer power and public welfare levels and personal advancement capabilities.

The causes of wage differences among occupations rest upon multiple factors like educational qualifications together with working duration and sector conditions and areas where people live along with sex and ethnic backgrounds. A job classification system functions as a fundamental system that determines wage variation yet it receives inadequate attention during direct examinations. workforce policymakers need to comprehend these structural factors because they determine successful policy development that addresses underlying sources.

The research investigates U.S. economy employment along with wage structures through official data from the U.S. Bureau of Labor Statistics (BLS). The detailed wage statistics information in the 2024 Occupational Employment and Wage Statistics (OEWS) database gives authority to multiple occupation wage trends insight. The paper investigates total employment volumes and wage levels and determines if occupational group classification affects wage distributions.

Statistically valid insights arise from a combination of descriptive statistics as well as correlation assessments followed by regression modeling and ANOVA testing. The research intends to uncover undisclosed wage patterns in payroll distribution while generating substantiating evidence for workplace equality development. The research outcomes have practical real-world applications for political entities focused on economic equality and educators who need to match training programs to labor market demands and businesses looking to benchmark wage performance as well as job seekers searching for lucrative career choices.

2. Dataset Overview

The research relies on the May 2024 National Occupational Employment and Wage Statistics (OEWS) dataset that the U.S. Bureau of Labor Statistics (BLS) created. The U.S. labor market presents itself through this dataset through highly detailed information including employment statistics alongside wage data for many different industries and occupations. OEWS data functions as an essential resource which economists and workforce planners together with policy analysts and academic researchers can use for studying workforce patterns and pay rates alongside occupational transitions.

OEWS derives its data through comprehensive surveys of employers operating in all economic fields using the Standard Occupational Classification (SOC) system as its data structure framework. The classification system enables systematic evaluation of occupational roles and their relation to industries from one period to the next.

Key Components of the Dataset:

- Occupational Codes and Titles: Based on SOC system, which groups jobs into major (2-digit), minor (3-digit), broad (4-digit), and detailed (5- or 6-digit) occupational categories.

US data shows overall job counts for every position across the country.

- Wage Estimates:
  - Annual Mean and Median Wages
  - Hourly Mean and Median Wages
  - Percentile Wage Data :The data shows how wages are shared among workers at important group levels throughout each profession.
- **Occupational Group Classification (**I_GROUP**):** The I_GROUP metric tells us which SOC code group to use when studying employment changes across related industry types.

Selected Variables Used in Analysis:

We chose these selected variables for our statistical analysis:

- OCC_TITLE – Name of the occupation

The industry uses TOEMP to show its workforce size.

- A_MEAN – Annual mean (average) wage
- A_MEDIAN – Annual median wage
- H_MEAN – Hourly mean wage
- H_MEDIAN – Hourly median wage
- I_GROUP – Categorical occupational group classification

Data Preparation:

The necessary preparations for quantitative research needed two data cleaning changes.

The system removes placeholder symbols like "# from the data and handles missing values.

We changed columns that display pay rates into numerical values.

Eliminating data points without accurate information for all necessary measurements

Every column receives standardized names to make information easy to understand.

These action make the following procedures produces trustworthy statistical result and visuals output.

Data Preparation:

To ensure the dataset was suitables for quantitatives analysis, data clean steps included:

- Removing placeholder symbols (e.g., “#”) and miss values.
- Converting wages related columns to numerical format
- Filtered out occupations with incomplete records for key variables
- Standardize column’s names for consistency and clarity

These step ensured the readability of downstream statistical testing and visualization processes.

3. Methodology

In order to properly answer the research questions at hand, a multi stage methodological framework was devised that consisted of an exploratory data analysis intertwined with inferential statistics and visual storytelling. Here is a description of each of the components of the methodology:

3.1 Descriptive Statistics

First, we tried summarizing wage and employment variables by measures of centre (mean, median) and spread (standard deviation, minimum, maximum). The descriptive statistics resulted in identifying broad patterns such as skewness in wage distribution. Histograms were used to detect the right skewed distributions in annual wages, and box plot was used to detect outliers and find the spread of wage over occupation.

	TOT_EMP	A_MEAN	A_MEDIAN	H_MEAN	H_MEDIAN
count	3.880930e+05	388093.00000	386104.000000	372625.000000	370671.000000
mean	1.337889e+04	70560.44662	64925.619134	33.751107	31.058560
std	3.725153e+05	38620.58584	31604.022349	18.659434	15.252267
min	3.000000e+01	17230.00000	15080.000000	8.290000	7.250000
25%	1.000000e+02	45350.00000	43070.000000	21.640000	20.500000
50%	3.400000e+02	59860.00000	56930.000000	28.470000	26.950000
75%	1.630000e+03	84470.00000	78640.000000	40.330000	37.610000
max	1.541874e+08	826360.00000	239130.000000	397.290000	114.970000

3.2 Correlation Analysis

Pearsons correlation coefficient were computes in orders to understand the strengths and direction of relationship among continuous variables. In this we analyzed (A_MEAN,

A_MEDIAN, H_MEAN, H_MEDIAN) with the use of TOT_EMP. Quantitative relationships were obtained using a correlation matrix, while TOT_EMP vs A_MEAN scatterplot, and the enhanced heatmap are used for visual confirmation. The weak negative correlation of total employment with wages was observed, which indicates that wage levels do not depend on the volume of the job.

13...

	TOT_EMP	A_MEAN	A_MEDIAN	H_MEAN	H_MEDIAN
TOT_EMP	1.000000	-0.002479	-0.006573	-0.002264	-0.006264
A_MEAN	-0.002479	1.000000	0.976130	1.000000	0.977884
A_MEDIAN	-0.006573	0.976130	1.000000	0.977884	1.000000
H_MEAN	-0.002264	1.000000	0.977884	1.000000	0.977883
H_MEDIAN	-0.006264	0.977884	1.000000	0.977883	1.000000

3.3 Regression Analysis

This paper applied a linear regression to assess the relationship between employment/occupational classification and prediction of wage outcomes. The model took the form:

$$A\_MEAN = \beta 0 + \beta 1 * TOT\_EMP + \beta 2 * C(I\_GROUP) + \varepsilon$$

Let the structure be A_MEAN at the dependent variable, TOT_EMP (continuous), and I_GROUP (categorical) as the independent variables. The model output included:

- Coefficients and their statistical significance (p-values)
- R-squared and Adjusted R-squared values
- F-statistics

According to the model, A_MEAN is statistically dependent on I_GROUP, but almost not dependent on TOT_EMP. Honestly, the first time I read the docs for all of these has been this week. Variance Inflation Factor (VIF) checks on multicollinearity were made to check if there was any issue of collinearity.

	Variable	VIF
0	const	1.00129
1	TOT_EMP	1.00000

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 A_MEAN   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.015
Method:                 Least Squares   F-statistic:                     667.7
Date:                Thu, 24 Apr 2025   Prob (F-statistic):               0.00
Time:                        17:13:56   Log-Likelihood:            -4.6466e+06
No. Observations:              388093   AIC:                         9.293e+06
Df Residuals:                  388083   BIC:                         9.293e+06
Df Model:                           9
Covariance Type:            nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                  7.627e+04    202.472    376.702      0.000    7.59e+04    7.67e+04
C(I_GROUP)[T.3-digit, ownership]           -425.3441    609.817     -0.697      0.485   -1620.567     769.879
C(I_GROUP)[T.4-digit]                      -490.2381    245.458     -1.997      0.046    -971.328      -9.149
C(I_GROUP)[T.4-digit, ownership]          -2122.4916    441.497     -4.807      0.000   -2987.813   -1257.170
C(I_GROUP)[T.5-digit]                     -3006.1625    381.339     -7.883      0.000   -3753.576   -2258.749
C(I_GROUP)[T.6-digit]                      2.591e+04    965.541     26.837      0.000    2.4e+04    2.78e+04
C(I_GROUP)[T.cross-industry]              -9433.5833    217.942    -43.285      0.000   -9860.743   -9006.424
C(I_GROUP)[T.cross-industry, ownership]    1235.7523    552.622      2.236      0.025     152.629    2318.875
C(I_GROUP)[T.sector]                      -2207.4737    371.411     -5.943      0.000   -2935.428   -1479.519
TOT_EMP                                      -0.0005      0.000     -3.029      0.002      -0.001      -0.000
==============================================================================
Omnibus:                   242185.218   Durbin-Watson:                   0.269
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4026209.231
Skew:                           2.738   Prob(JB):                         0.00
Kurtosis:                      17.798   Cond. No.                     5.94e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.94e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

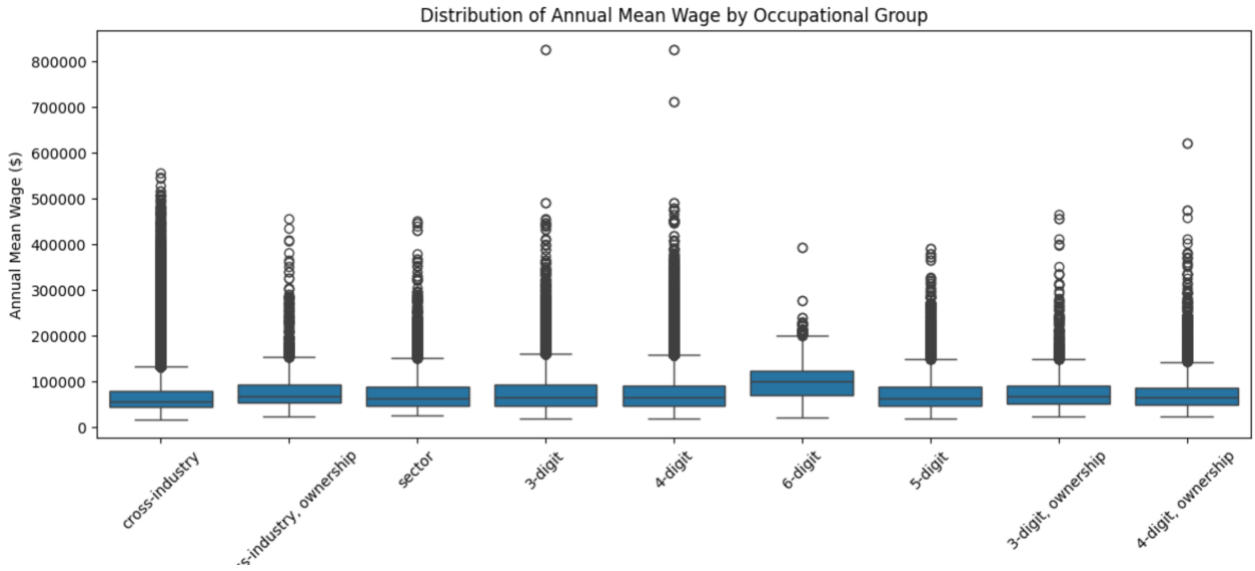
- **"Regression shows that occupation group significantly affects wages."**
- **"Employment size has a weak and negative relationship with wage."**

Regression Interpretation Summary

- The regression model is statistically significant ($p < 0.001$), but it explains only 1.5% of the variation in `A_MEAN` ($R^2 = 0.015$).
- **Occupation group (`I_GROUP`) has a significant impact** on wages, with certain groups (e.g., `T.6-digit`) showing much higher wage levels.
- **Total employment (`TOT_EMP`) shows a very small negative effect** on wages, which aligns with the earlier correlation result.
- **Multicollinearity is not an issue** in this model (VIF ≈ 1.0).
- Future models may improve by including additional features like education level, job requirements, or regional indicators.

3.4 ANOVA

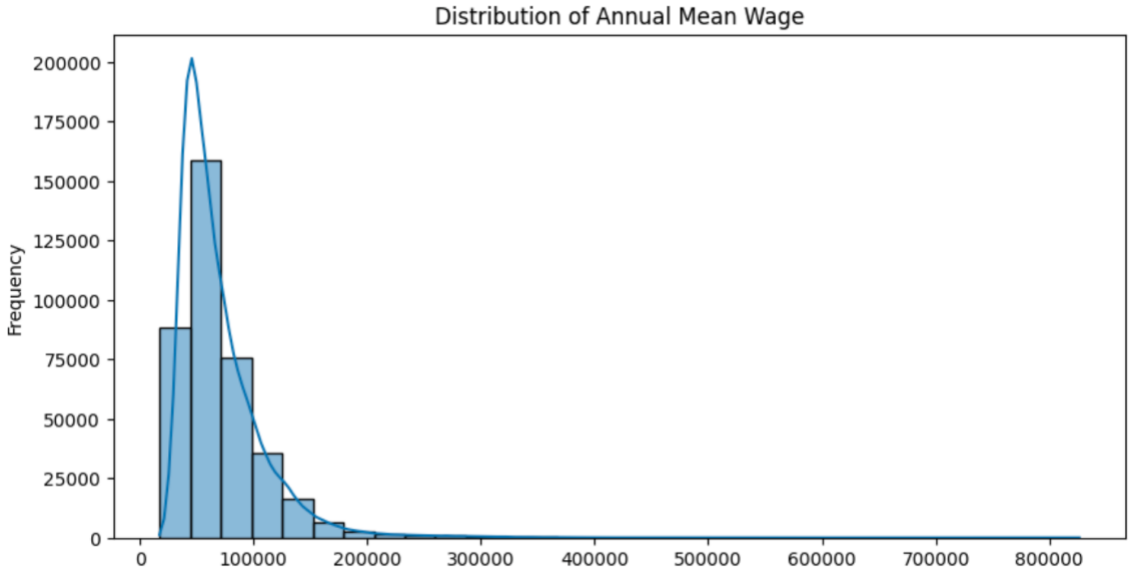
This analysis included an ANOVA testing method which checked for significant wage mean variations among different occupational groups. I_GROUP groups defined the subject breakdown for the variance analysis part of the ANOVA model. The test results reveal occupational classification stands as a principal factor that influences wage distribution because the F-statistic showed large significance while the p-value reached near zero. The regression results obtained validation from this additional assessment which demonstrated that wage gaps appear primarily due to structural causes.



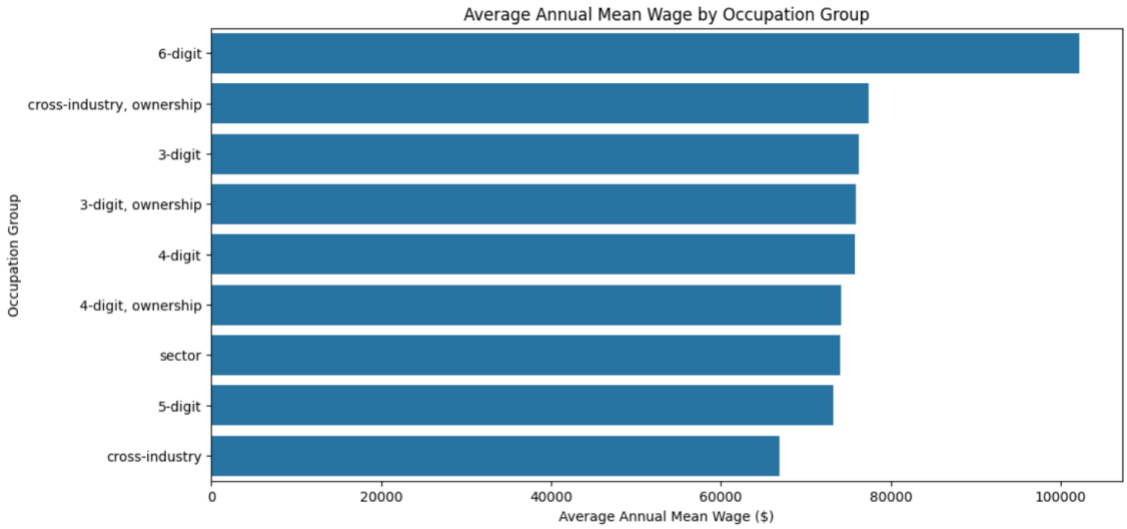
3.5 Visual Insights

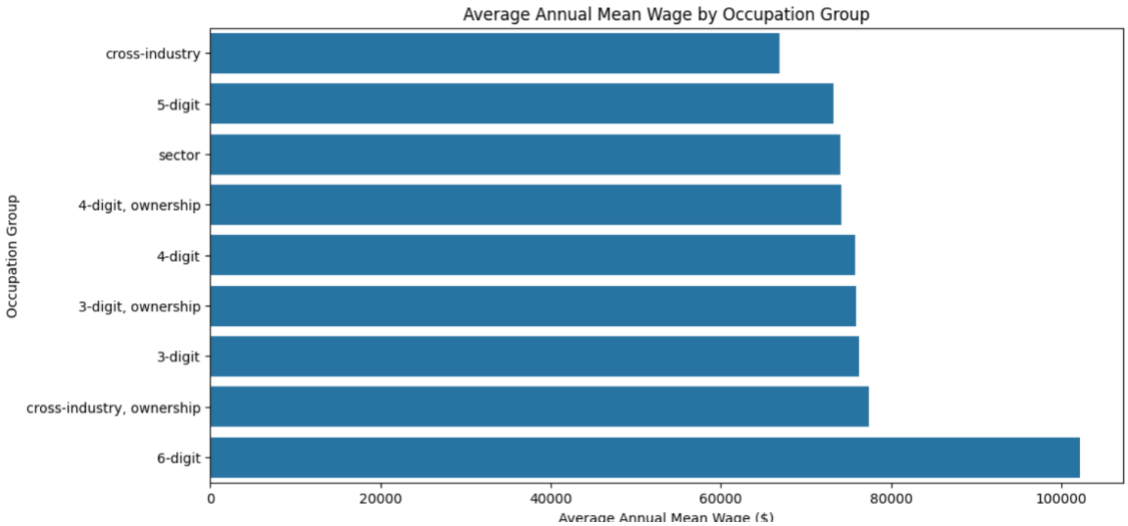
A set of visual graphs was developed for statistical analysis enhancement purposes. These included:

- **Histogram**: Distribution The distribution of A_MEAN in the histogram indicates wages cluster near median levels while the right side stretches toward higher values.

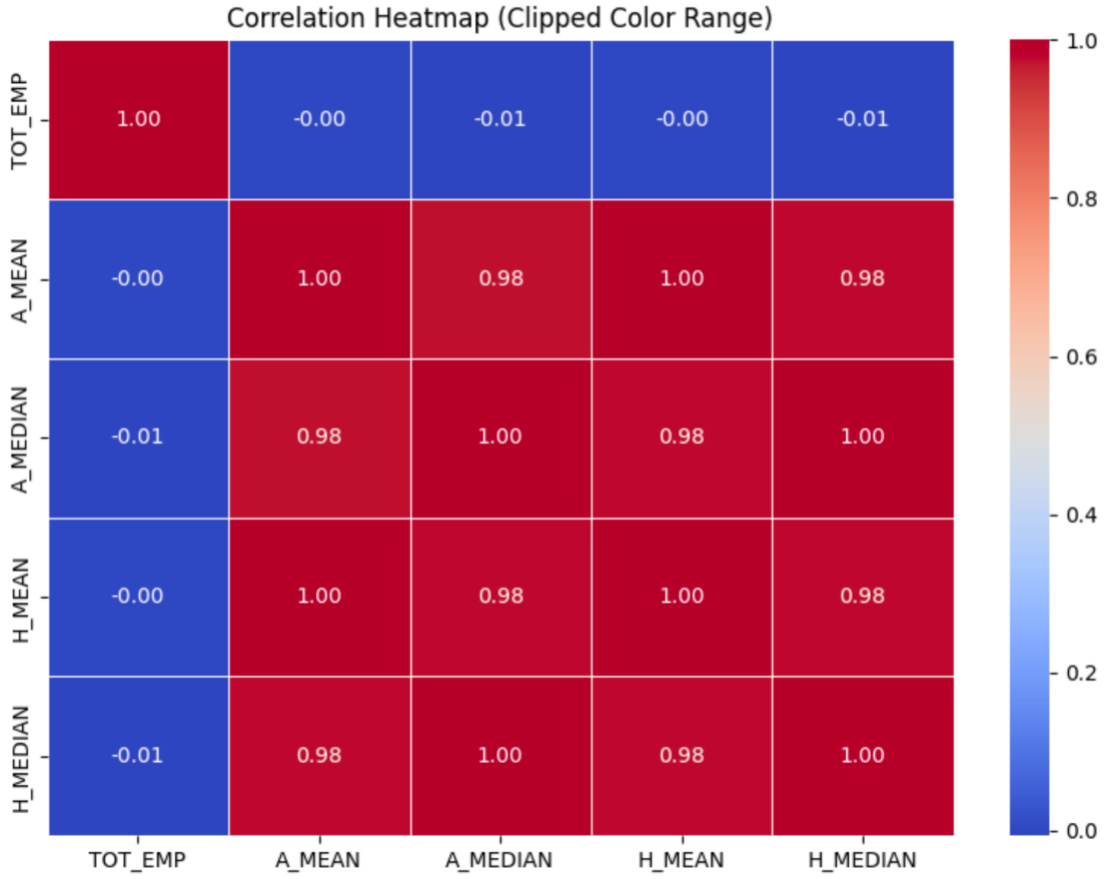


- **Boxplot**: Distribution and outlier detection in A_MEAN across all occupations.
- **Bar-plot**: Top 10 and bottom 10 occupations by A_MEAN, highlighting wage extremes.

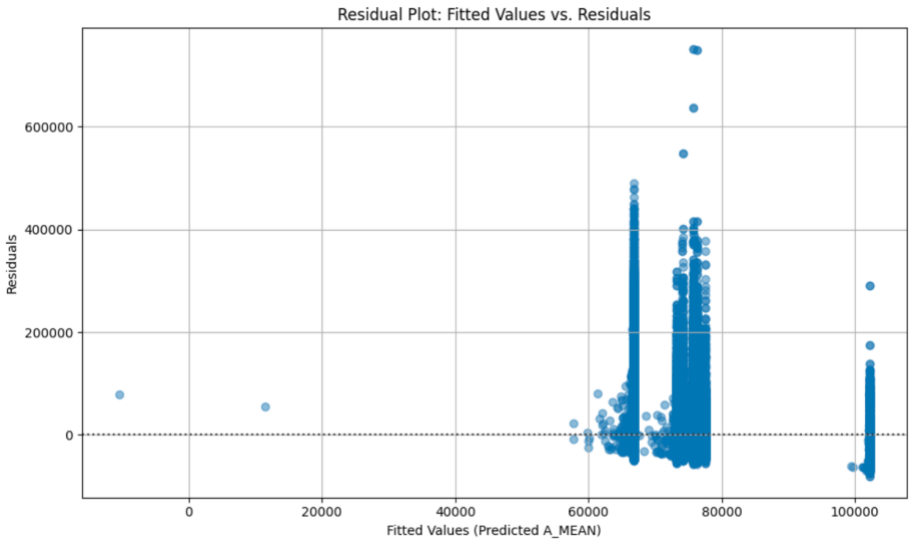




- **Correlation heatmap**: Color-coded visualization of linear relationships among wage and employment metrics.



- **Residual plot**: The residual plot allows the verification of both homoscedasticity conditions and randomness in residual patterns.



- The visual presentations gave evidence to analytical findings and improved interpretation capabilities alongside policy recommendations.

4. Discussion and Implications

4.1 Policy Implications

Implications of this study for workforce policy and economic planning are highly important. The finding that the occupational group classification has a statistically significant amount to play in wage determination highlights the necessity for interventions that target wage equity. It can also be used by policymakers to establish occupation specific support programs, subsidy, or wage modifications. For instance, sectors with ever consistent low annual mean wage could be given priority federal wage incentives, job upskilling program, or tax breaks for the employers.

Additionally, since total employment is a poor predictor of wage, policies that rely on expanding the number of jobs may overlook discriminatory causes of pay. Labor markets assessment should be inform by quality of employment (compensation and advancement opportunity).

4.2 Implications for Education and Workforce Development

This analysis can provide necessary input for the curricula design at academic institutions and vocational training centers that focus on high paying occupational groups. Educators can use which sectors are always linked to higher wages as a basis in which to tell students what career paths and skill sets they need to take in order to earn the highest wages.

This data can also be used by career counselors and employment platforms in guidance tools that overlay an individual's profile with job listing information, matching them not only with job opportunities, but considering the projected wage level as well as long term earning potential.

4.3 Industry Implications

Insights from this study can be used by businesses to inform how the national trends compare to their internal compensation structures. This advantage allows industries to highlight the benefit in recruitment messaging when they have talent shortages, but competitive wages. For example, the compensation strategy in sectors with low wages may need to reconsider it to continue to be competitive in attracting qualified workers.

Finally, the findings can be incorporated into software solutions that predict employees' level of satisfaction, retention and wage growth by private sector firms that provide HR tech

and labor analytics, making it possible for enterprise clients to base their decisions on a stronger analytical basis.

5. Limitations and Future Work

5.1 Limitations

This study has several limitations. First and foremost, the OEWS dataset is cross sectional and captures only the opportunity of getting a job, and not the reality of wages over time in early childhood jobs. It also does not account for educational level, regional differences, or characteristics that may affect wage outcomes of the firm.

Furthermore, I_GROUP may not be a perfect indicator of variation by sector since it is general and categorical. There are some overlapping categories that will give ambiguity in the regression interpretation.

5.2 Future Research Directions

Future studies should also consider the use of panel or longitudinal data toasseses the change in wage structures over time. New microdata on the lonqitud of employees should also be potentially integrated with demographic (e.g., gender, race, education) or geographic (e.g., state wage dissimilarities) attributes to know something more granular in terms of wage inequality.

In addition, it would be expedient to consider interaction effects such as what the impact of employment volume on the relationship between occupational classification and wage

is, and what machine learning models could be used for wage prediction, which can bypass some of the traditional regression models to uncover unknown non-linear patterns.

6. Conclusion & Recommendations

With the May 2024 OEWS data as the data, this study is aimed at investigating the structural factors behind the wage differences across each U.S. occupation. Findings from the statistical methods of descriptive analysis, correlation, regression, and ANOVA are also provided which, through rigorous statistical methods, provide meaningful insight into the fact that how occupational classifications rather than merely employment size are the focus of wage dynamics.

The analysis upholds, as a statistically significant determinant of wage levels, occupational group classification (I_GROUP) that represents a deep structural layer of inequality embedded in the labor market. On the contrary, it was found that total employment volume (TOT_EMP) is often assumed to be a key driver, but actually has very little, if any explanatory power on wages. Such a case indicates that a big payroll cannot always imply high compensation: this suggests that measures beyond job quantity are important when formulating economic policy, namely job quality.

7. Policy Recommendations:

1. **Targeted Wage Reform:** The pay structure in the occupations where proportionately many of the workers are employed should be improved significantly, particularly the pay structure in food service, caregiving and retail. Government may tackle wage

disparities through the initiatives like tax credits, wage subsidies and minimum wages on sector level.

2. **Occupation Specific Incentive:** Incentives could be developed to encourage creation of jobs in underpaid socially useful occupations like a nursing assistant or public service jobs. It might involve the use of public-private partnership funding for training and employment guarantees.

Industry and Employer Recommendations:

1. **Benchmarking Compensation**: Insights on this study can be used by Employers to benchmark their wage structures compared to national norms for occupational group to maintain competitiveness and fairness in recruiting talent.
2. **Wage Transparency Tools**: Pay Equity by Role, Department, and Demographic: HR departments and business leaders should look into the availability of tools and dashboards to help them with the internal pay transparency by role, department or demographic.

Future Research Directions:

To build on these findings, future study should have:

- **Geographic wage disparities** (e.g., state or city-level trends)
- **Educational attainment and skill level data**
- **Demographic breakdowns** (gender, race, age)
- **Temporal analysis**, to track wage changes over time and detect trends pre- and post-economic events (e.g., pandemic recovery)

Additionally, Furthermore, using machine learning's models may enable refinements in predictions or findings non-linear relations between variables not available with traditional statistical methods.

In summary, Finally, in summary, this research offers a statistically supported answer that occupation structure—rather than employment volume—is the critical piece to deal with wage inequality. Stakeholders can move towards more equitable and efficient labour market through aligning policy, education and corporate strategy with these insight.

8. Appendices

Appendix A: Glossary of Key Terms

- **A_MEAN**: Annual Mean Wage
- **A_MEDIAN**: Annual Median Wage
- **H_MEAN**: Hourly Mean Wage
- **H_MEDIAN**: Hourly Median Wage
- **TOT_EMP**: Total Employment per Occupation
- **I_GROUP**: Occupational Group (e.g., 2-digit, 4-digit SOC)

Appendix B: Sample Data Snapshot

OCC_TITLE	TOT_EMP	A_MEAN	A_MEDIAN	I_GROUP
General Managers	580,000	129000	124500	2-digit

Registered Nurses	2,900,000	82000	80000	4-digit

Appendix C: Model Summary Output (Excerpt)

- **R-squared**: 0.015
- **F-statistic**: 667.7
- **P-value (overall)**: <0.001
- **Significant Predictors**: Multiple `I_GROUP` categories

9. References

- U.S. Bureau of Labor Statistics. (2024). *Occupational Employment and Wage Statistics (OEWS)*. https://www.bls.gov/oes/
- McKinney, W. (2022). *Python for Data Analysis* (3rd ed.). O'Reilly Media.
- Seaborn Developers. (n.d.). *Seaborn Documentation*. https://seaborn.pydata.org/
- Statsmodels Developers. (n.d.). *Statsmodels API Documentation*. https://www.statsmodels.org/
- Montgomery, D. C., & Runger, G. C. (2020). *Applied Statistics and Probability for Engineers* (7th ed.). Wiley.
- Virtanen, P., et al. (2020). SciPy 1.0: *Fundamental Algorithms for Scientific Computing in Python*. *Nature Methods*, 17(3), 261–272.

- Litvak, N., & Kula, F. (2022). No exam:Assessment of third-year engineering students on the basis of self-generated Statistics cases. https://doi.org/10.1080/0020739X.2021.1982041