Complex Data Munging & Statistical Modeling in Pandas

📊 Case Study: Employee Salary Dataset

Transforming messy real-world data into actionable insights through advanced statistical modeling

Data Quality Challenges

Mixed Date Formats

Inconsistent date formats (YYYY-MM-DD, MM/DD/YYYY) requiring standardization

Salary Inconsistencies

Text values ("forty thousand") and negative salary entries corrupting numeric analysis

Missing Critical Data

Gaps in Name and PositionTitle fields affecting dataset completeness

Extreme Outliers

Unusually high salaries (Police Chief) skewing distribution patterns

Data Cleaning & Preparation Pipeline

01

Salary Normalization

Converted text to numeric, removed invalid and negative entries for consistent analysis

02

Missing Value Imputation

Applied forward/backfill strategies for categorical columns to preserve data integrity

03

Feature Engineering

Standardized date formats and extracted Years_of_Service as predictive feature

04

Schema Optimization

Converted categories to Categorical dtype, normalized schema with pivot and dummies

Exploratory Data Analysis Insights

Distribution Patterns

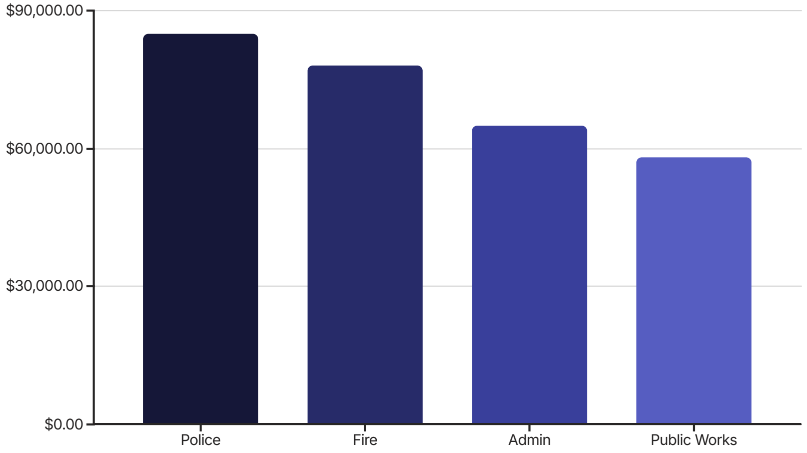
Salary distribution heavily skewed by extreme outliers requiring careful handling

Strong Correlation

Positive correlation discovered between Benefits_Cost and Salary variables

Service Impact

Years_of_Service shows moderate trending relationship with compensation levels



Statistical Modeling Results

Model Selection

OLS Regression: Salary ~ Benefits_Cost + Years_of_Service

Adjusted $R^2 \approx 0.72$

Significant Predictors

✅ Benefits_Cost ($p < 0.05$)

❌ Years_of_Service (not significant)

Model Validation

Residuals approximately normal distribution

No strong heteroscedasticity detected

Key Findings & Insights

$1

Benefits Impact

Every additional dollar in benefits correlates with salary increases

72%

Model Accuracy

Variance explained by our regression model

28%

Unexplained Variation

Remaining factors likely include education, location, and experience

Benefits spending emerges as the strongest predictor of salary levels, while years of service shows weaker predictive power than expected.

Conclusion & Future Directions

1

🛠️ Data Pipeline Success

Successfully cleaned and transformed messy real-world dataset using pandas

2

📈 Statistical Modeling

Built interpretable regression model with robust statistical inference capabilities

3

🔮 Next Steps

Expand features (education, city), explore advanced models (Lasso, Ridge, Tree-based)

Key Takeaway: Pandas + statsmodels provide a powerful toolkit for end-to-end data science workflows, from messy data to actionable insights.