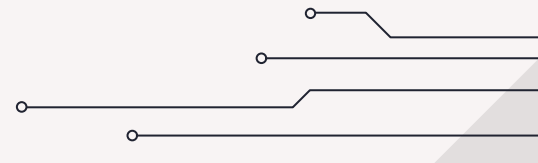


DATA 606: Capstone Project
Summer 2025- Dr. Unal Sakoglu

Predicting Tech Salaries: A Data-Driven Analysis of U.S. Labor Statistics: P3

Team A
Dereck Román Rosario
Simran Shah
Gelareh Vakili



Project Repository – DevPay Insights



Project Repository:

[GitHub — DevPay Insights](#)

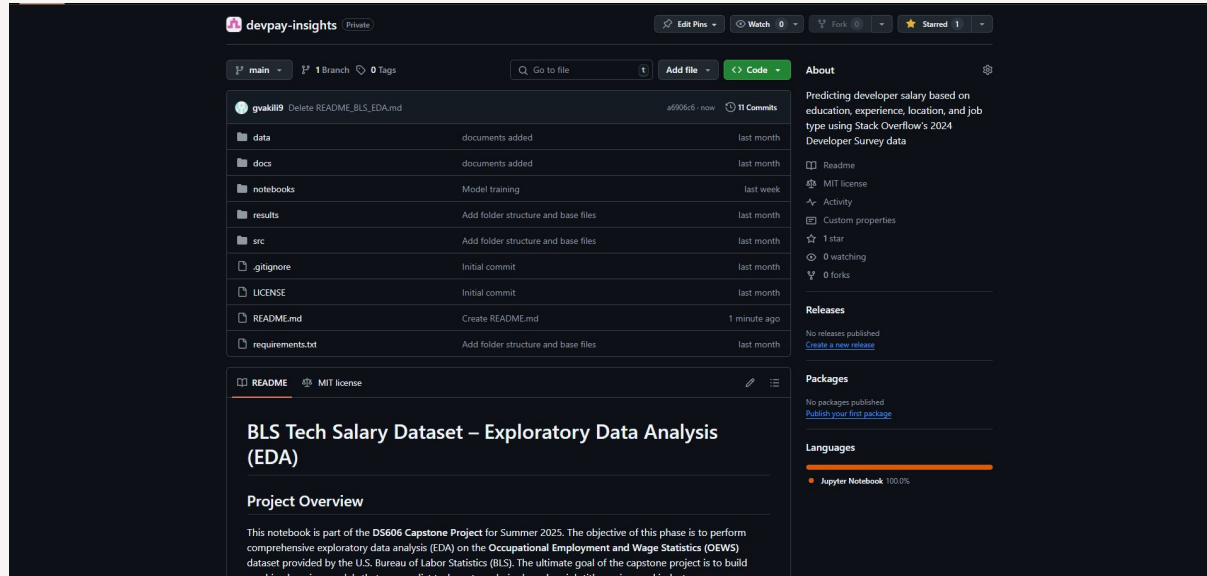
Repository Highlights:

- ❖ Organized into subfolders: data/, notebooks/, docs/, results/, and src/
- ❖ Contains full EDA and modeling notebooks for both BLS and Stack Overflow datasets
- ❖ Environment setup via requirements.txt; repository is MIT licensed
- ❖ README outlines the project objective, methodology, and dataset usage
- ❖ Actively maintained by team members: Dereck Román Rosario, Simran Shah, and Gelareh Vakili

Current Status:

- ❖ Data cleaning & filtering complete
- ❖ EDA insights finalized
- ❖ Feature engineering completed (one-hot encoding, frequency features, standardization)
- ❖ Ridge Regression baseline established
- ❖ Random Forest Regressor tuned and evaluated — better performance over baseline
- ❖ Final deliverables in progress: Phase 3 presentation, updated GitHub content, ICMLDE-format paper

screenshot of the GitHub repo homepage



Project Summary and Machine Learning Task



Project Title:

Predicting Tech Salaries: A Data-Driven Analysis of U.S. Labor Statistics



Objective:

Develop a supervised machine learning model to predict salaries in the U.S. tech sector. The project uses government labor statistics to identify how factors like occupation, industry, and geography influence compensation.



Why This Matters:

- ❖ Enhances transparency in tech compensation
- ❖ Informs policy, workforce planning, and career decision-making
- ❖ Uses standardized, non-self-reported data for model reliability



ML Task:

- ❖ **Type:** Regression
- ❖ **Target Variable:** Annual median wage (`A_MEDIAN`, continuous variable, USD)
- ❖ **Outcome:** Predict salary across occupations and regions

Primary Dataset: U.S. Bureau of Labor Statistics (OEWS)



Source:

- ❖ U.S. Bureau of Labor Statistics (BLS)
- ❖ [Download Link](#)



Dataset Properties:

- ❖ ~400,000 records » ~92,000 after filtering for tech-related jobs
32 columns
- ❖ Covers all U.S. states, territories, and metro regions
- ❖ Includes wage, employment, occupation, and industry data
- ❖ Format: Excel (XLSX), government-verified, public domain

After Cleaning (for modeling)

- **17,398 records**
- **27 columns**



Key Variables:

- ❖ **OCC_TITLE**: Occupation title
- ❖ **AREA_TITLE**: Region
- ❖ **NAICS_TITLE**: Industry sector
- ❖ **TOT_EMP**: Total estimated employment
- ❖ **A_MEDIAN**, **A_MEAN**: Median and mean annual wage

What Has Been Done Before?



Existing Approaches:

- ❖ **Stack Overflow Developer Survey:**
Used in many community dashboards and salary comparison tools; primarily descriptive.
- ❖ **Glassdoor & Levels.fyi Studies:**
Self-reported data analyzed using simple regression or filtering by job title and location.
- ❖ **BLS Reports:**
Government publications often visualize wage distributions but lack predictive modeling.



Common Limitations:

- ❖ Heavy reliance on **self-reported** data, often international and inconsistent.
- ❖ Most work is **descriptive**, lacking predictive or inferential modeling.
- ❖ Limited integration of **structured public datasets** like BLS OEWS.
- ❖ Often uses proprietary, non-generalizable standards

Filling the Gaps: Our Distinctive Approach

Gaps in Prior Work:

- ❖ Minimal use of structured, U.S. government labor datasets in machine learning
- ❖ Few predictive models using regional and occupational features
- ❖ Lack of reproducibility in dashboards and survey-based tools

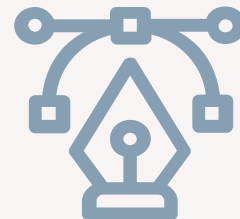


Our Contribution:

- ❖ **Public, reproducible data:** Using the BLS OEWS dataset, which is standardized and verifiable
- ❖ **Predictive Modeling:** Implementing regression to forecast salaries across occupations and regions
- ❖ **Feature Engineering:** Incorporating geographic, occupational, and industry-level variables
- Transparent Workflow:** Full pipeline documented and version-controlled on GitHub

Optional Extensions:

- ❖ Clustering occupations and regions
- ❖ Cross-validating insights with the Stack Overflow Developer Survey (2024)



EDA Refresher: Key Findings Driving Model Design

Key Insights:

- ❖ Geographic Variation — States like California, Washington, and New York consistently show the highest median tech salaries.
- ❖ Occupational Impact — Roles in software development, data science, and systems architecture top the pay scale; support and technician roles fall lower.
- ❖ Industry Influence — Certain NAICS industry sectors (e.g., information services, software publishing) show higher pay than government or educational sectors.
- ❖ Employment Concentration — High-paying roles are clustered in urban metro areas with large tech hubs.
- ❖ Feature Correlation — Strong linear relationship between A_MEAN and A_MEDIAN annual wages confirmed target reliability.

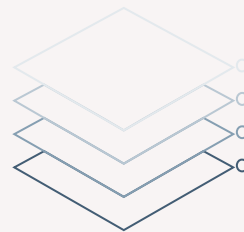
Why It Matters for Modeling:

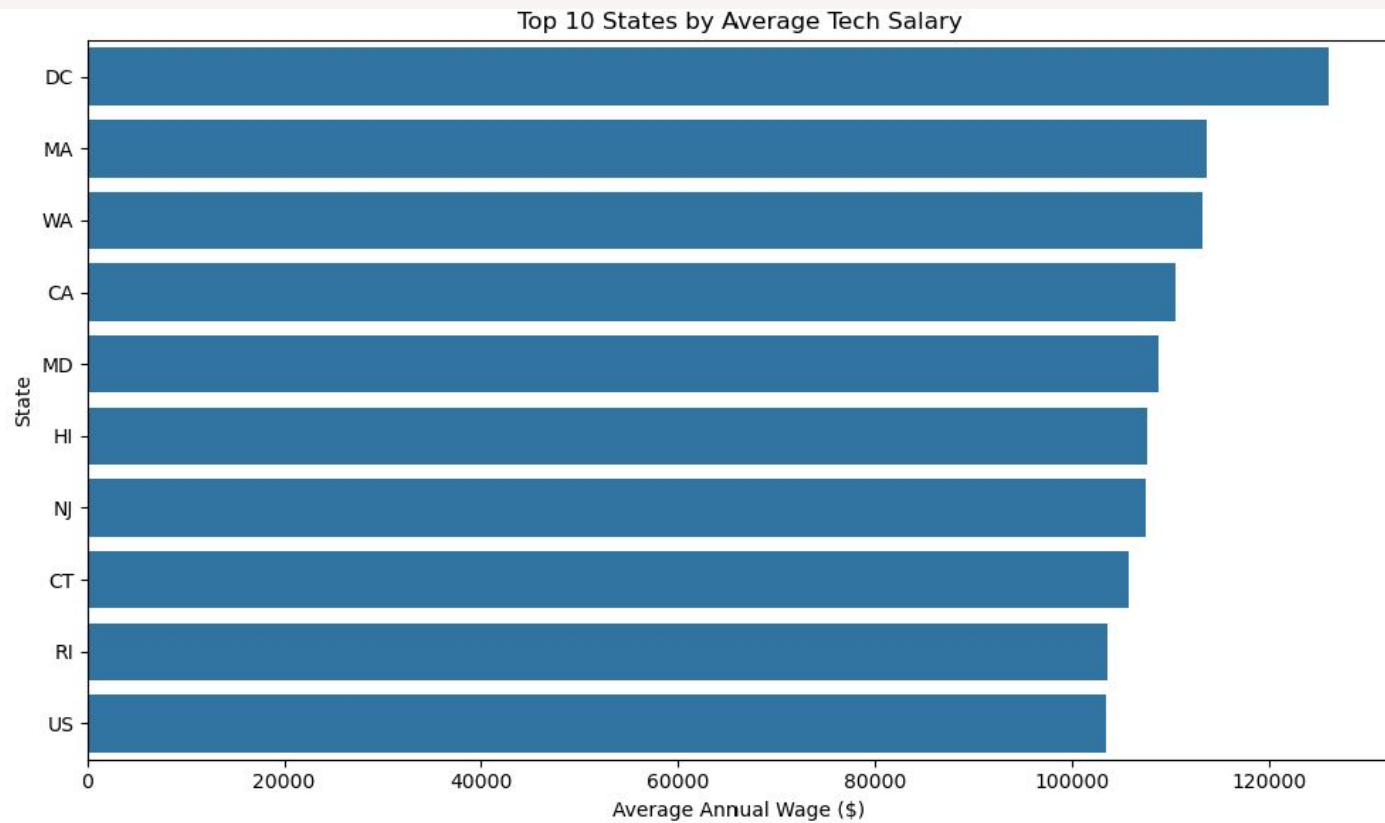
- ❖ Validated inclusion of **occupation**, **region**, and **industry** as core predictive features.
- ❖ Supported **one-hot encoding** of categorical variables.
- ❖ Justified **non-linear modeling** (Random Forest) due to complex geographic-occupational interactions.

Preparing the BLS OEWS Dataset for Modeling

Cleaning & Filtering Steps

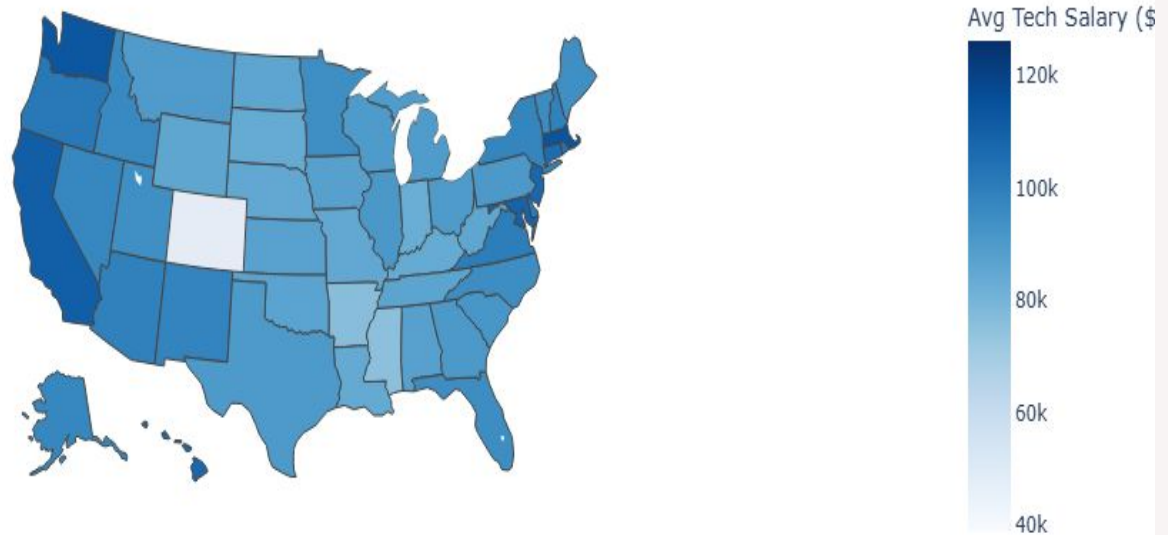
- ❖ Loaded raw Excel file into a DataFrame using **pandas**
- ❖ Dropped rows with suppressed or missing wage and employment data
- ❖ Filtered out:
 - Summary/aggregate rows (e.g., “All Occupations”)
 - Invalid or unknown area codes
 - Records with placeholder or zero salary values
- ❖ **Standard Occupational Classification (SOC) codes Group Filtering:** We filtered only major occupation groups relevant to tech through federal code standards
 - » Example: **15-xxxx** = Computer and Mathematical Occupations
- ❖ **Keyword Filtering:** We further filtered job titles using tech-related keywords such as “developer,” “engineer,” “data,” “IT,” and “cyber.”



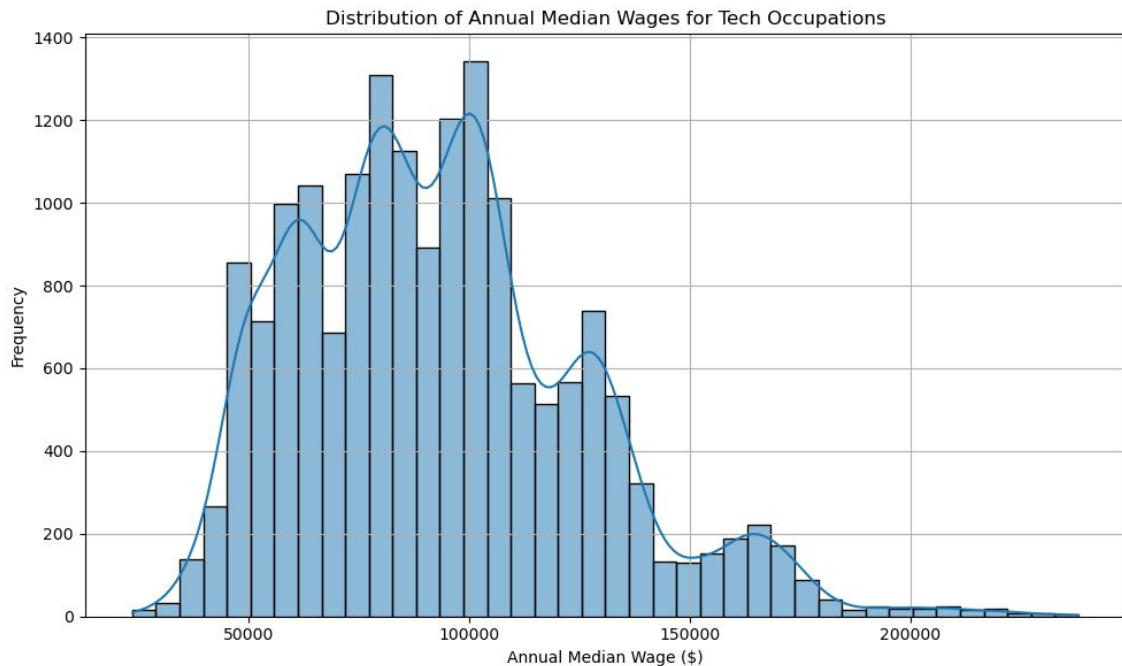


Average Tech Salaries by State

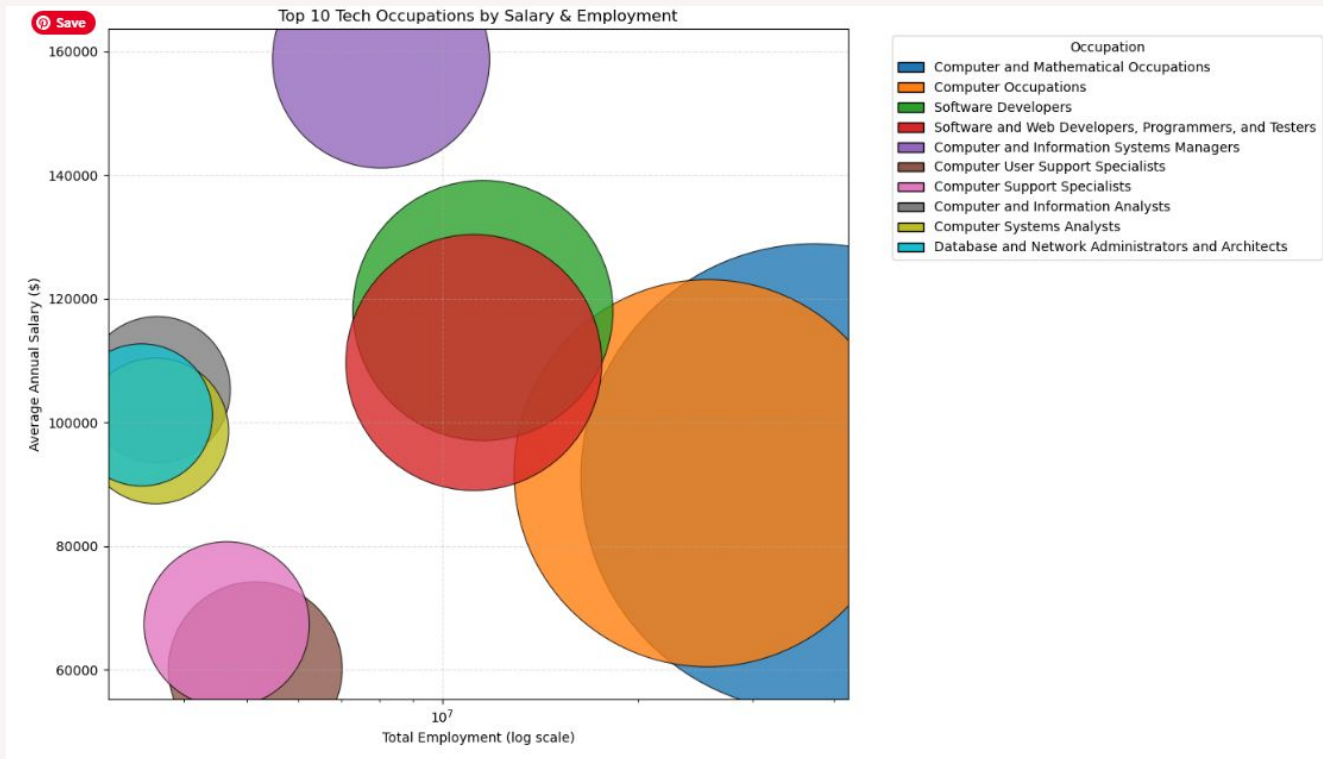
Average Tech Salaries by State (A_MEAN)



Distribution of Tech Occupation Median Wages



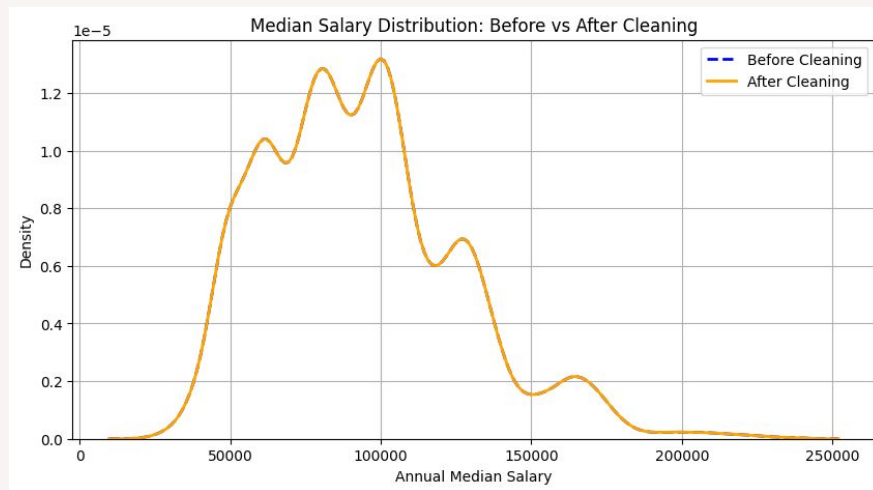
Top 10 Tech Occupations by Salary & Employment



Preparing the BLS OEWS Dataset for Modeling

After Cleaning:

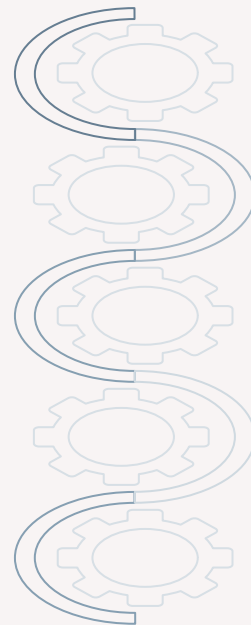
- ❖ Unique job titles before cleaning: 29
- ❖ Unique job titles after cleaning: 29
- ❖ KDE plot, before and after cleaning was plotted, the shape and range of the distribution remain consistent, confirming that we did not lose valuable salary insights during the cleaning process.



Preparing the BLS OEWS Dataset for Modeling

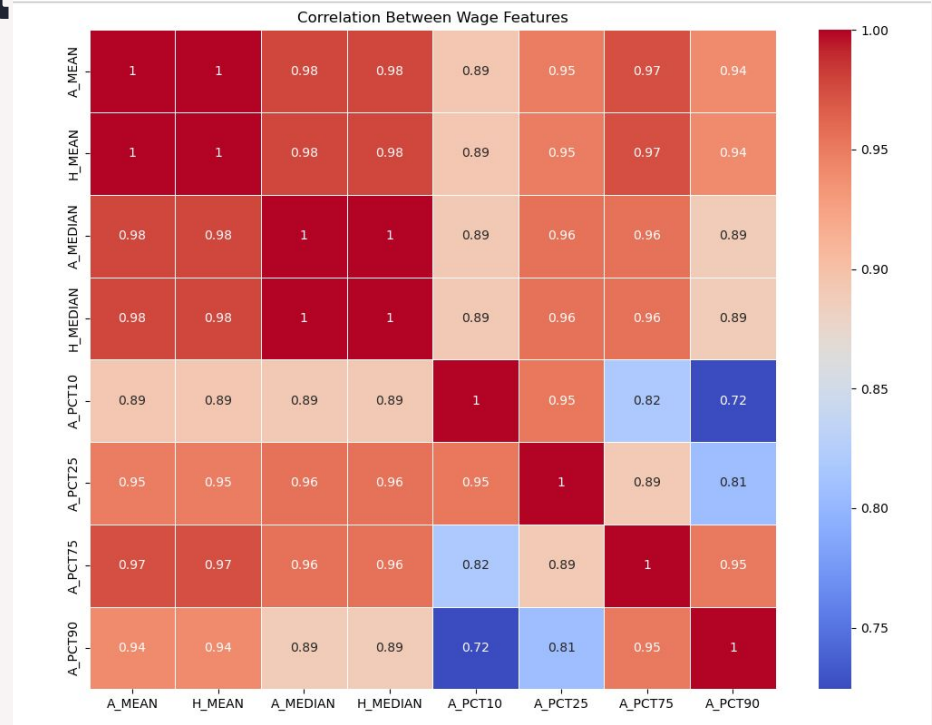
Feature Selection:

- ❖ Kept modeling-relevant fields:
 - `OCC_TITLE` (occupation)
 - `AREA_TITLE` (location)
 - `NAICS_TITLE` (industry)
 - `TOT_EMP` (employment count)
 - `A_MEDIAN` / `A_MEAN` (target variables)
- ❖ Prepared categorical variables for encoding:
 - Grouped overly specific titles into broader categories
 - Reserved space for encoding via one-hot or label methods



Exploring Tech Salary & Employment Patterns

- ❖ Cleaned dataset used to analyze salary trends across U.S. tech occupations and regions
- ❖ Visualized distributions, correlations, and regional disparities
- ❖ Confirmed strong linear relationships between wage fields (e.g., A_MEAN, A_MEDIAN)



Modeling Approach and Training Process



Target variable & Models Trained

A_MEDIAN (annual median wage)

- ❖ Ridge Regression
- ❖ Random Forest Regressor



Features used

Encoded:

- ❖ OCC_TITLE,
- ❖ AREA_TITLE,
- ❖ NAICS_TITLE,
- ❖ TOT_EMP



Data split

- ❖ Training: 70%
- ❖ Validation: 15%
- ❖ Testing: 15%



Feature encoding

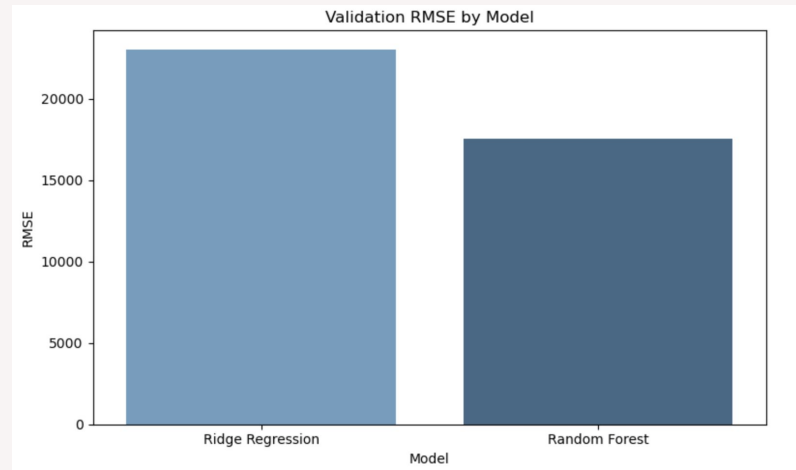
One-hot for categorical features

Model Iteration & Selection

| Model | Key Characteristics | Validation RMSE (\$) | Test RMSE (\$) | Notes / Outcome |
|------------------|--|----------------------|----------------|---|
| Ridge Regression | Linear model with L2 regularization | 29,922.20 | ~23,031 | Established baseline; underfit complex patterns |
| Random Forest | Ensemble of decision trees, non-linear relationships | 17,495.56 | 17,502.32 | Best performance; captures geographic-occupational interactions |

Why Random Forest Was Selected:

- ❖ Handles **non-linear relationships** between features and salary.
- ❖ Robust to **outliers** in salary data.
- ❖ Naturally ranks **feature importance**, aiding interpretation.
- ❖ Lower RMSE on both validation and test sets compared to Ridge.



Model Performance & Comparison

Metric Used: Root Mean Squared Error (RMSE)

Validation Results:

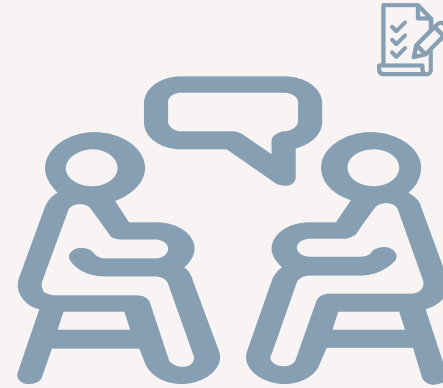
- ❖ Ridge Regression: RMSE = **29,922.20**
- ❖ Random Forest: RMSE = **17,495.56**

Ridge Cross-Validation (5-Fold):


- ❖ Mean RMSE = 31,261.60

Test Set Performance:

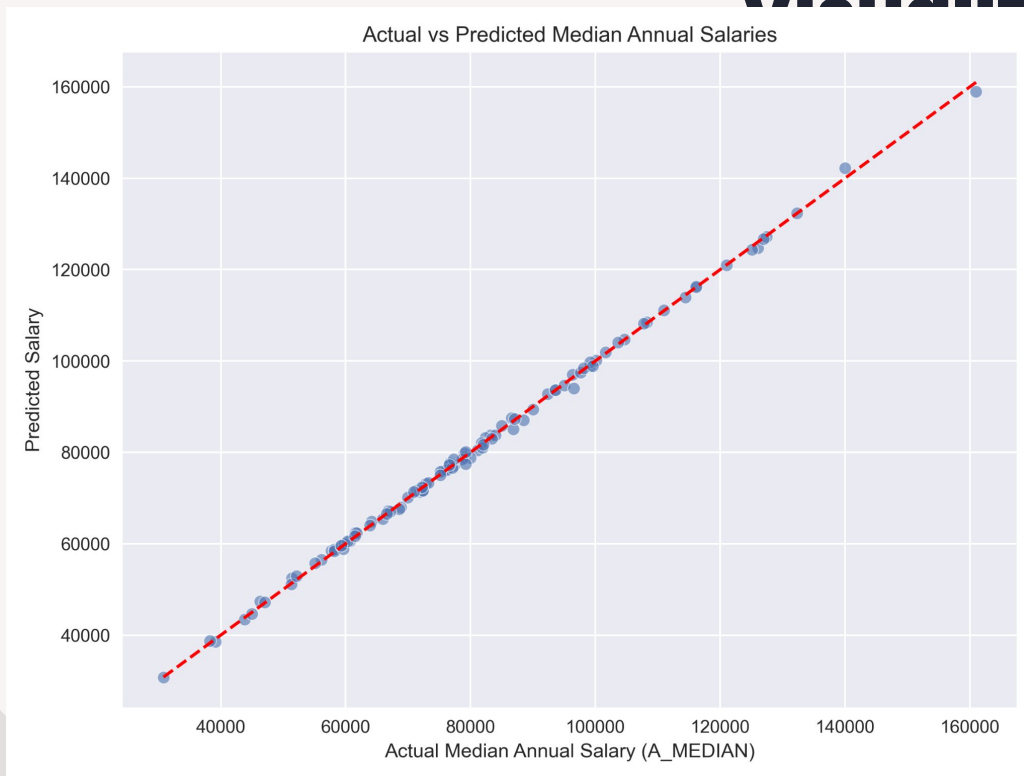
- ❖ Random Forest Test RMSE: **17,502.32**



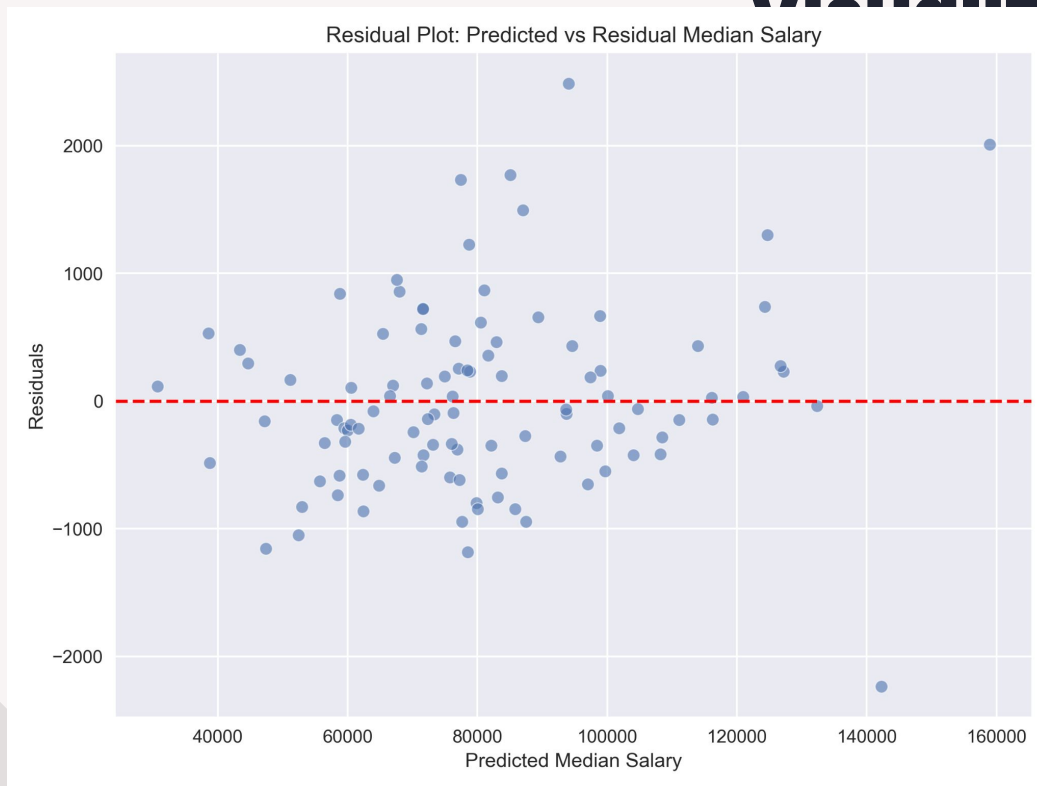
From Baseline to Best Model: Iteration Process

| | |
|----|--|
| 01 | Data Cleaning (Remove missing & Invalid entries) |
| 02 | Feature Engineering (One-hot encoding, frequency features, standardization) |
| 03 | Data Split (70% Train/15% Validation/15% Test) |
| 04 | Model Training  Ridge Regression (Baseline) Random Forest (Main) |
| 05 | Evaluation (RMSE, residuals, feature importance) |
| 06 | Model Selection (Random Forest chosen) |

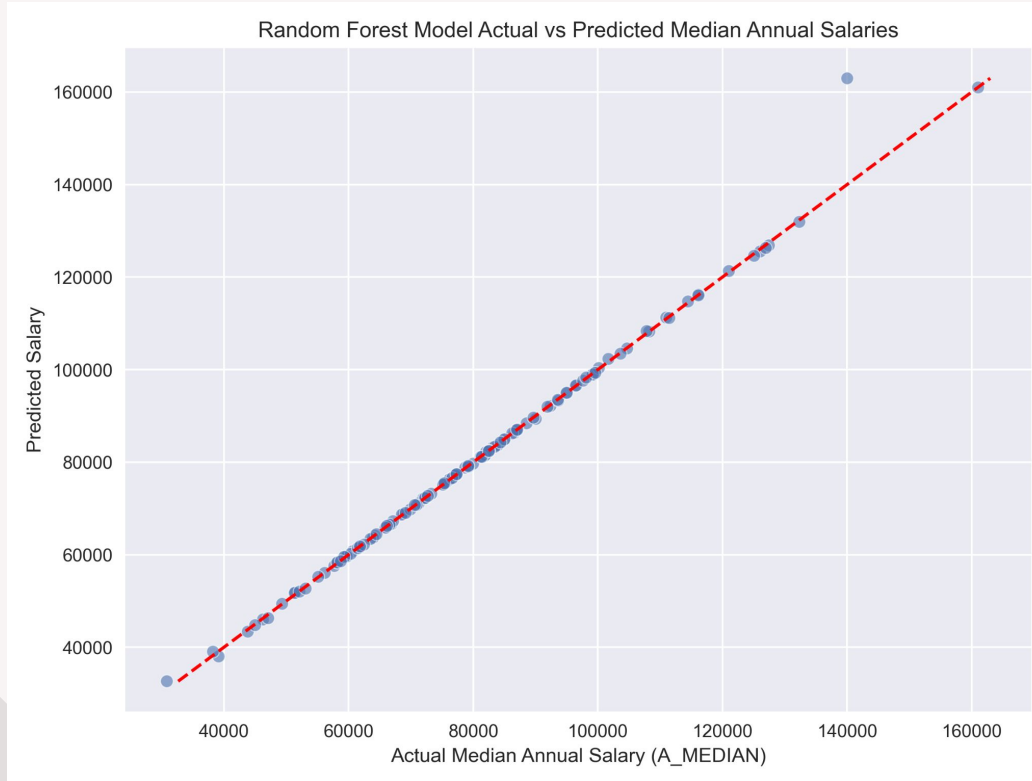
Ridge Model Performance Visualization



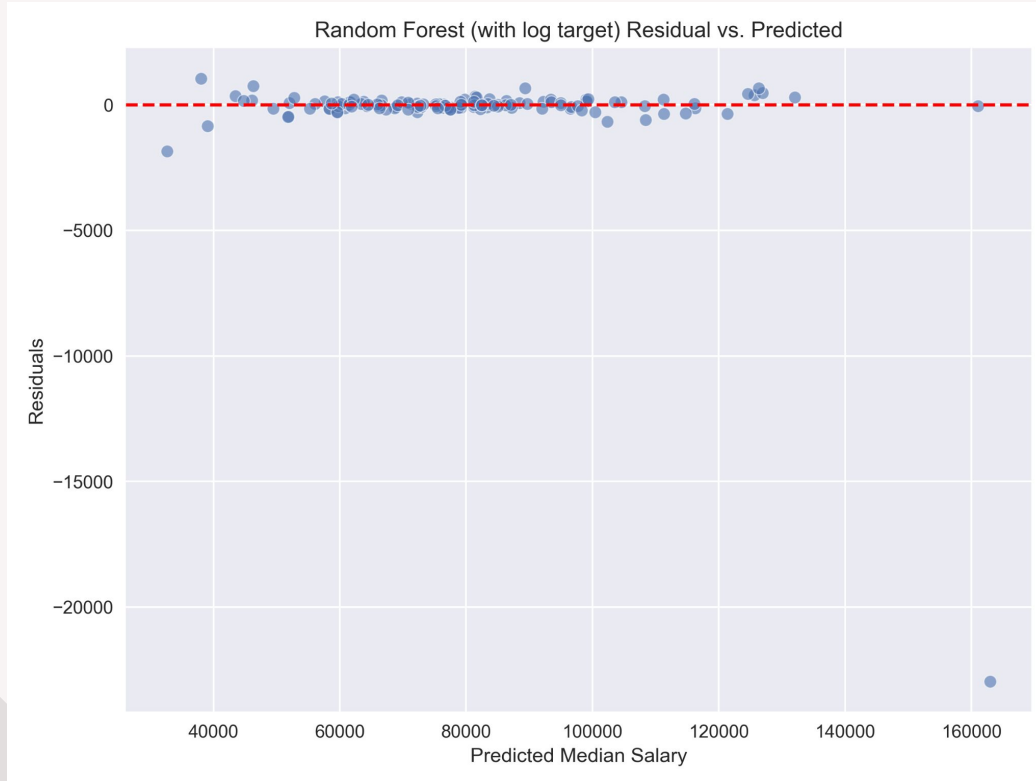
Ridge Model Performance Visualization



RF Model Performance Visualization



RF Model Performance Visualization



Interpretation & Key Insights

Feature Importance (Random Forest):

Occupation Title (OCC_TITLE) — strongest predictor of salary.

Region (AREA_TITLE) — major driver of wage variation.

Industry Sector (NAICS_TITLE) — certain sectors consistently outpay others.

Employment Size (TOT_EMP) — smaller correlation; high employment areas not always high-paying.

Model Behavior:

Underestimates extreme high salaries — typical in tree-based models without log-transforming skewed targets.

Handles mid-range salaries well, with residuals clustered near zero.

Non-linear feature interactions captured — e.g., high-paying roles in specific regions amplified salary predictions.

Real-World Takeaways:

Geographic and occupational targeting could inform **policy decisions** and **career planning**.

Industry choice plays a significant role in potential earnings.

Data-driven approach improves over self-reported salary estimates.

What We've Done and Why It Matters

- ❖ Built a fully reproducible machine learning pipeline using public, government-verified U.S. labor data.
- ❖ Conducted comprehensive exploratory analysis of salary trends across tech occupations, industries, and regions.
- ❖ Constructed and evaluated baseline and advanced regression models, selecting Random Forest for its superior performance.
- ❖ Identified key drivers of salary variation and quantified disparities by role, location, and industry sector.
- ❖ Delivered actionable insights that can inform policy decisions, workforce planning, and career guidance.

What Could Be Done Differently / Next Steps

Model Enhancements:

While our Random Forest model delivered strong performance, there are areas where it could be refined. Integrating the Stack Overflow Developer Survey could provide a valuable comparison with self-reported salary data. Advanced ensemble algorithms such as XGBoost or LightGBM may offer additional performance gains. Applying a log transformation to the target variable could help address skewness and improve predictions for extreme high salaries. More comprehensive hyperparameter tuning with expanded search methods could further optimize results.

Feature Engineering Improvements:

Salary predictions could be strengthened by including cost of living indices to normalize location-based differences. Additional features, such as education level and years of experience, could be incorporated from external datasets. Interaction terms between occupation and industry could help capture combined effects that influence compensation.

Extended Analyses:

Although outside the scope of this project, additional analyses could add value. Clustering could identify natural groupings of occupations and regions with similar pay characteristics. If multi-year BLS data were available, time-series analysis could uncover salary trends over time. An interactive dashboard could also make the model's predictions and insights more accessible to a broader audience.

Data Sources & Supporting Materials

Datasets:

- ❖ U.S. Bureau of Labor Statistics (OEWS 2024)
<https://www.bls.gov/oes/special-requests/oesm24all.zip>
- ❖ Stack Overflow Developer Survey 2024
<https://survey.stackoverflow.co/datasets/stack-overflow-developer-survey-2024.zip>

Tools & Libraries:

- ❖ Python 3.11, Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn
- ❖ Jupyter Notebook, GitHub

Background Reading:

- ❖ Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
- ❖ Stack Overflow Insights. (2024). *Developer Survey Results*.
- ❖ U.S. Department of Labor. (2024). *Occupational Employment and Wage Statistics (OEWS)*.
- ❖ Glassdoor. (2024). Glassdoor Salary Reports. Retrieved from <https://www.glassdoor.com>
- ❖ Levels.fyi. (2024). Levels.fyi Compensation Data. Retrieved from <https://www.levels.fyi>

Thank you

Questions?

CREDITS: This presentation template was created by [Slidesgo](#), and includes icons by [Flaticon](#), and infographics & images by [Freepik](#)