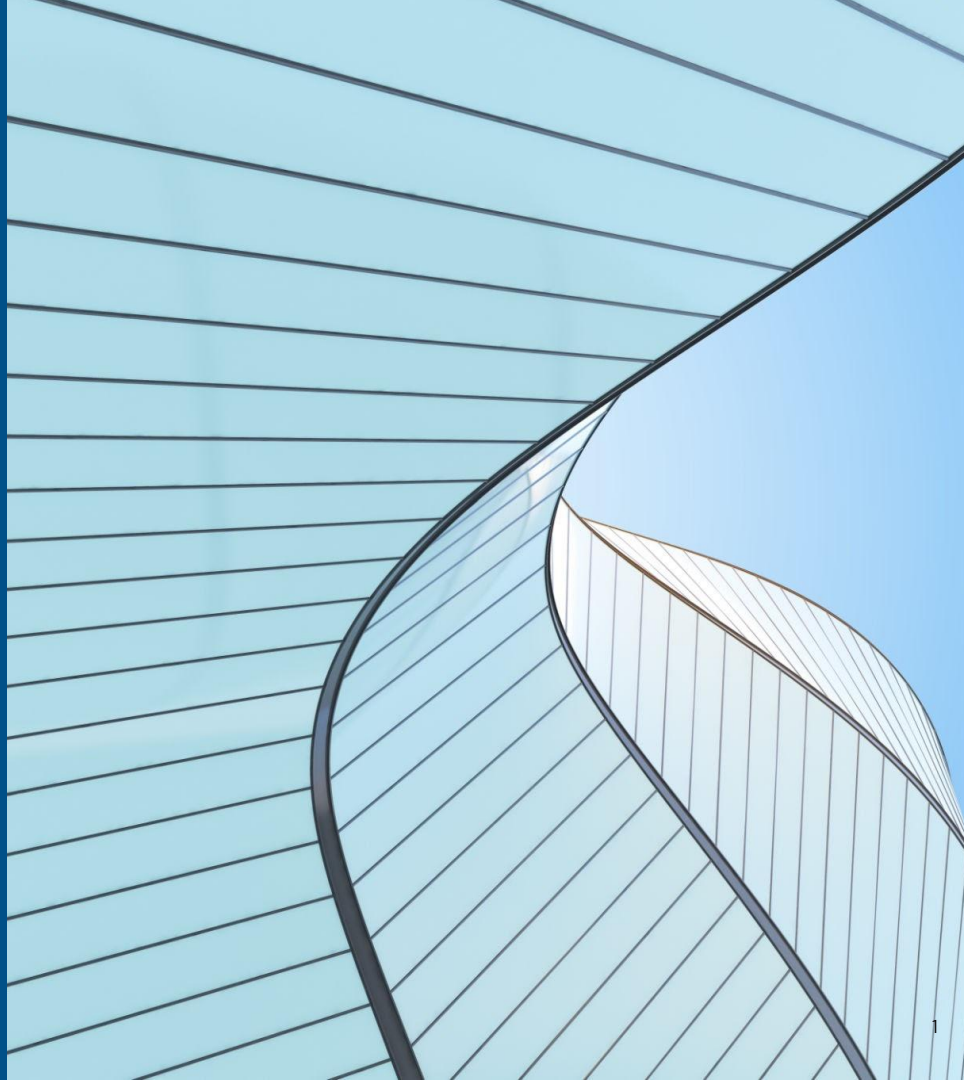
What Affects Salary?

Yiling Ding
Yulin Feng
Yichen Yang
Alex Fang



Content *overview*

Dataset Overview and Cleaning Process

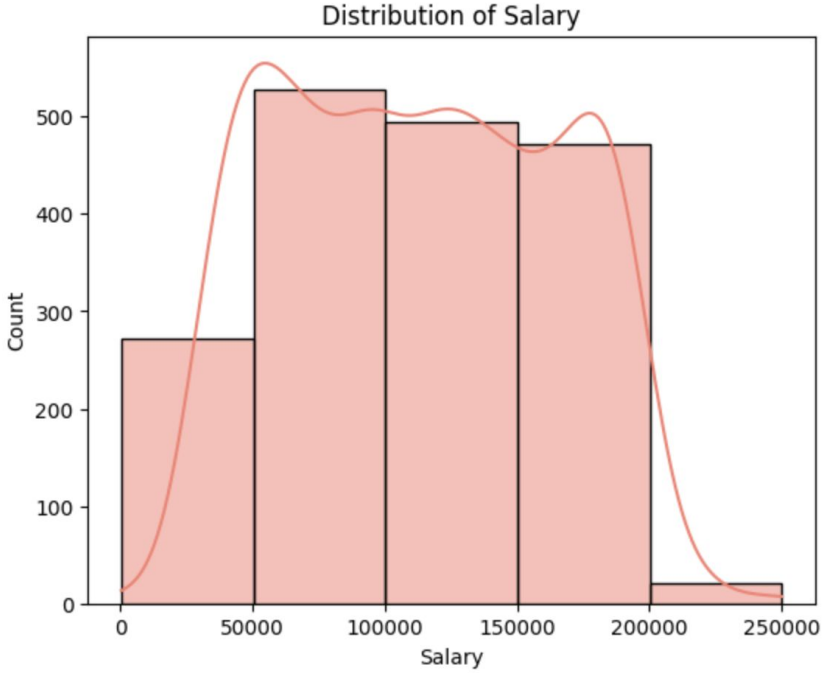
Dataset Overview

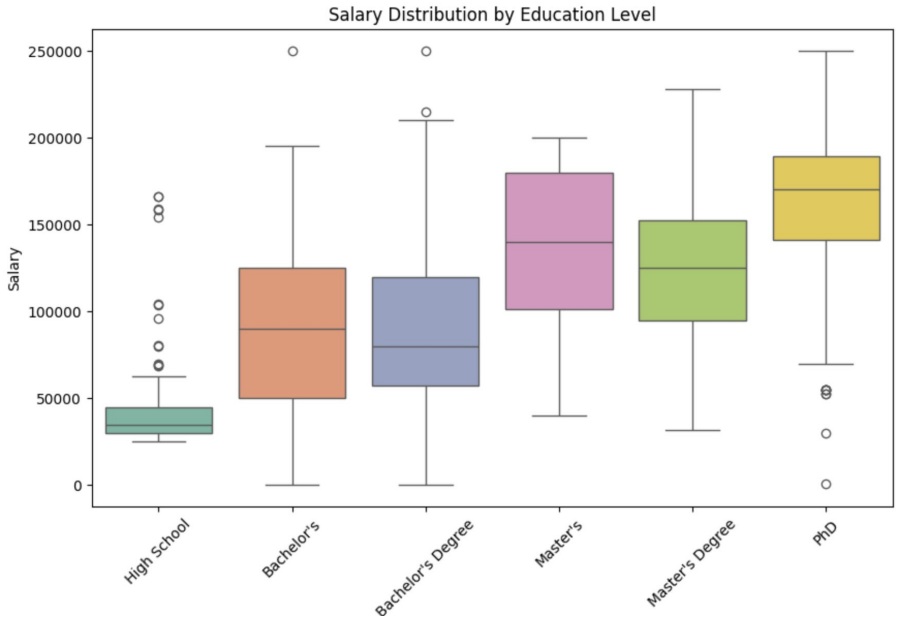
- **Rows:** 6704 rows of employee data
- **Columns:** Age, Gender, Education Level, Job Title, Years of Experience, Salary
- **Outcome Variable:** Salary

Data Cleaning Steps Performed

- **Handled Missing Values:**
  - Dropped rows with missing value (only 3 rows)
- **Standardized Categories:**
  - Ensured consistent capitalization and formatting for Gender, Education, Job Titles
- **Ensured Correct Data Types:**
  - Converted Age, Experience, Salary to numeric
- **Removed Duplicates:**
  - Checked for and eliminate duplicate rows

Exploratory Data Analysis Summary





Salary is right-skewed, with most employees earning between $50k–$150k and fewer high-income outliers above $200k.

Higher education levels correspond to higher median salaries and wider earning variability across roles.

Methodology: The DoWhy Framework

Model

Construct a causal graph (DAG) to encode domain assumptions and causal relationships.

Identify

Determine if the causal effect is identifiable from observational data using the backdoor criterion.

Estimate

Compute effect using Linear Regression and Double Machine Learning (DML).

Refute

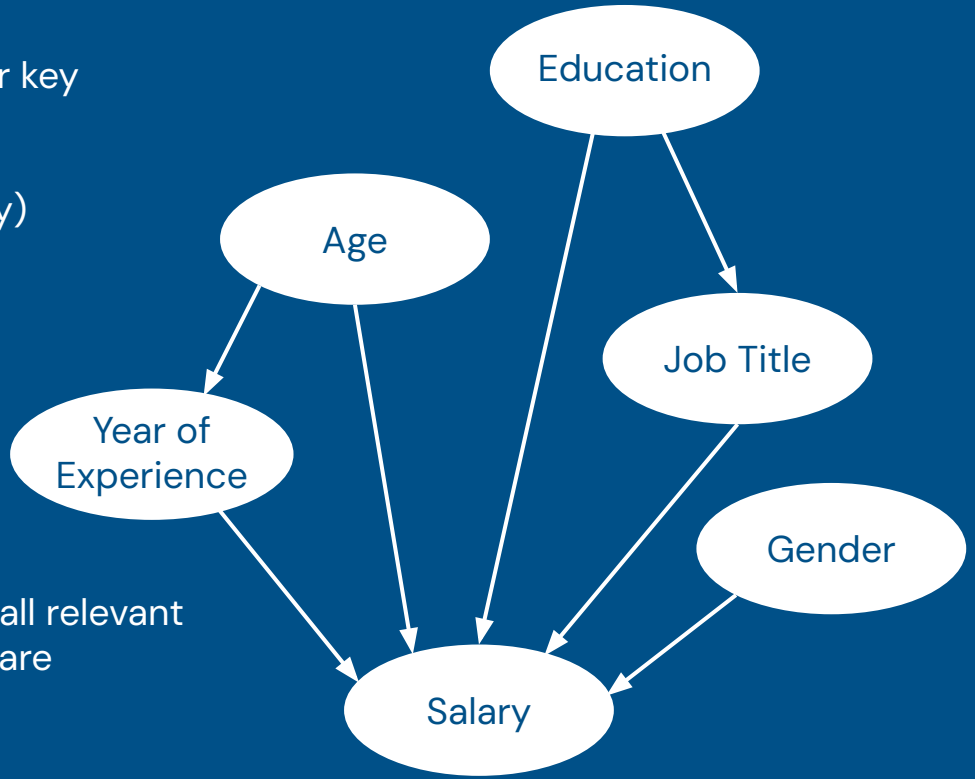
Validate estimates through Placebo, Random Common Cause, and Subset sensitivity tests.

Causal Graph Specification

The Directed Acyclic Graph (DAG) encodes our key assumptions about the causal structure:

- **Treatment:** Gender (Direct path to Salary)
- **Outcome:** Salary
- **Confounders:**
  - Age → Experience → Salary
  - Education → Job Title → Salary

Assumption: Unconfoundedness. We assume all relevant confounders affecting both gender and salary are observed.



Strategies & Findings

Method 1: Linear Regression

- **Approach:** Traditional OLS regression controlling for confounders.
- **Strengths:** Transparent, interpretable, and provides conditional treatment effects.
- **Limitations:** Assumes strict linear relationships; sensitive to model misspecification.

Method 2: Double Machine Learning

- **Approach:** Uses ML to model nuisance parameters (EconML).
- **Strengths:** Robust to misspecification; captures non-linearities; uses cross-fitting to prevent overfitting.
- **Result:** Considered the more robust and reliable estimate.

$2,367 Gap

Validation: Refutation Tests

- **Placebo Treatment Refuter:** Replaced gender with a random permutation. The effect dropped to near zero ($73.89, p=0.98), confirming the original finding is not spurious.
- **Random Common Cause Refuter:** Added a simulated random confounder. The estimate remained stable (diff: $4.52), indicating robustness to noise.
- **Data Subset Refuter:** Re-estimated on random 80% subsets of data. The estimate remained consistent across iterations.
- **Conclusion:** All refutation tests passed, validating the causal nature of the relationship.

Statistical Robustness (DML)

-$6,135	95%	[-8799, -3471]
Point Estimate (ATE)	**Confidence Interval Excludes Zero**	**Confidence Range ($)**

Bootstrap validation (SD=$1,294) further confirms that the estimate is stable and not sensitive to sampling variation.

Policy Implications & Recommendations

Enhance Pay Transparency	Implement public/internal salary bands for all positions to reduce negotiation subjectivity
Review Promotion/Hiring	Conduct gender-blind audits of promotion criteria and initial salary setting.
Standardize Non-Negotiable Salary Floors	Establish non-negotiable salary minimums for all entry/mid-level roles to eliminate the initial pay disparity caused.

Limitations & Future Causal Strategy

Core Limitation

Unobserved Confounding Risk

Model relies on the Ignorability Assumption (No Unobserved Confounding).

Risk

Key factors like Individual Performance or Negotiation Skills are unmeasured.

Causal Strategy

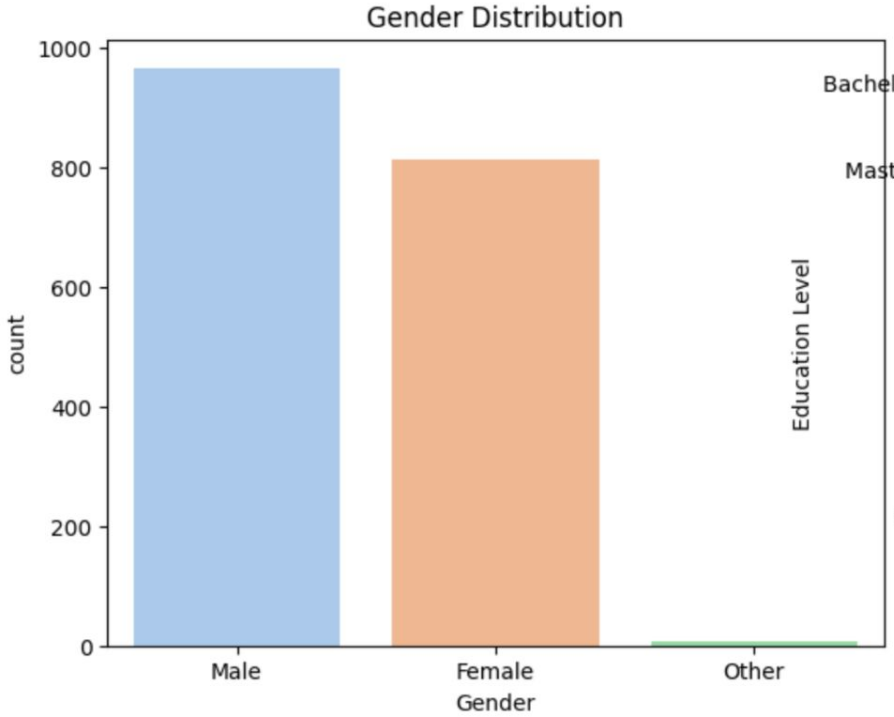
Robustness Check (Sensitivity Analysis): Quantify the minimum strength of an unobserved confounder required to nullify the ATE.

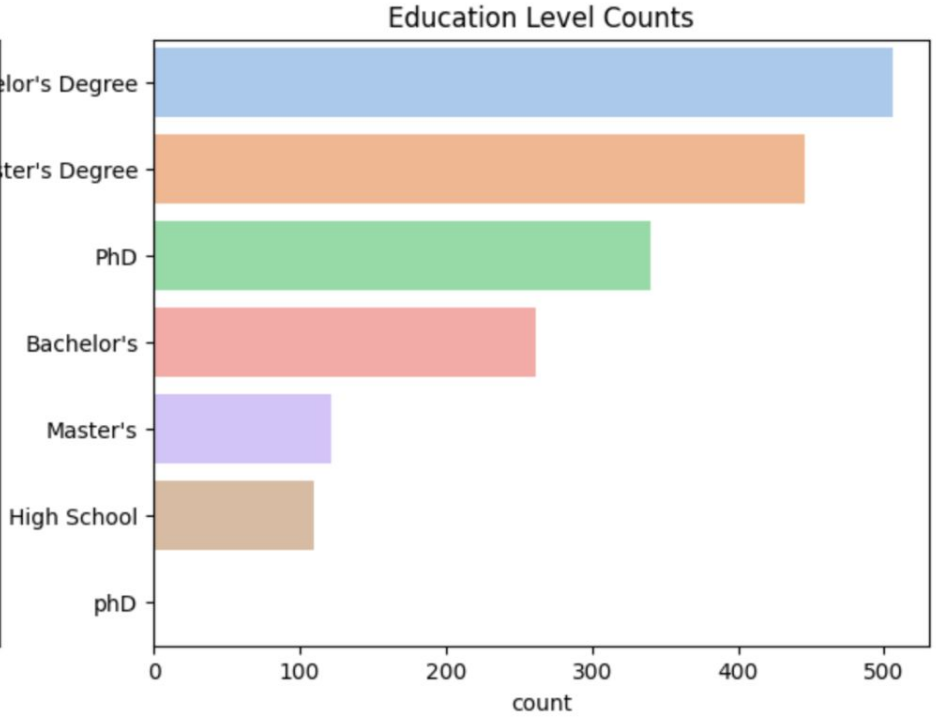
Mechanism Exploration (Mediation Analysis): Decompose the Total Effect (Gender to Salary) into Direct and Indirect Effects (via Job Title).

Targeted Intervention (HTE Analysis): Use Causal Forests to identify specific employee sub-groups most affected by the pay gap.

Thank You

Appendix 1: Data Distribution Visualizations





Interpreting the Difference

Why the $2,367 Gap?

The difference between the Linear Regression and DML estimates suggests the presence of **non-linear relationships** in the data.

DML's flexibility allows it to capture complex interactions between experience, age, and education that a standard linear model might miss. Consequently, we view the DML estimate as the more accurate point estimate, while the Linear Regression serves as an upper bound.