Occupation and the Risk of Automation

Anne Huang

Introduction

To what extent does an occupation's annual salary relate to its likelihood of being automated?

As AI and robotics advance, the assumption that high-paying jobs are safe is being questioned. Understanding whether higher pay still tracks with lower automation risk, and which kinds of skills bring that protection, helps students, workers, and policymakers make sense of job security in the changing labor market.

The intention of this paper is to investigate the relationship between occupational wages and automation risk using data on nearly 500 occupations in California. Specifically, I examine whether higher paying is associated with relatively lower automation probability and whether particular skill types, especially routine cognitive versus routine manual work, differ in their vulnerability to automation.

This paper tests two hypotheses. Hypothesis 1: There is a negative association between salary and automation probability: higher-paid occupations tend to have lower automation risk. Hypothesis 2: Among occupations with similar pay levels, cognitive-routine jobs face higher automation risk than manual-routine jobs. Both are tested below.

To clarify these categories: Cognitive-routine occupations involve following explicit rules, processing information, or performaning repetitive mental tasks – even if they pay well. Examples include accountants, paralegals, and loan officers. Manual-routine occupaions involve repetitive physocal tasks in structured environments, such as assembly line workers, packagers, and sorters. High-skill analytical roles such as surgeons and data scientists and interpersonal/creative roles such as therapists, graphic designers are considered non-rountine and are expected to have lower automation risk regardless of pay.

Data Collection

This project merges two existing data sources at the level of the individual occupation. Each case is one occupation; the outcome variable is its probability of automation, and the explanatory variables are its mean annual wage, its 90th-percentile upper-tail wage, and its skill type.

Occupational Employment and Wage Statistics, published by the U.S. Bureau of Labor Statistics in cooperation with state agencies. The file used here is the 2025 California statewide release, which reports, for each occupation, the number employed plus the mean and the 10th to 90th percentile annual wages. Wages are estimated from a large semi-annual survey of establishments.

A model of the susceptibility of U.S. occupations to computerization, a Frey and Osborne analysis published in Technological Forecasting and Social Change. It assigns each occupation a continuous automation-probability score from 0.0 not susceptible to 1.0 highly susceptible, based on features such as manual dexterity, creativity, and social intelligence, for roughly 700 occupations.

The two sources were joined on the Standard Occupational Classification SOC code. Of roughly 700 scored occupations, 495 matched a California wage record; 490 retained a usable mean wage and 467 also reported a 90th-percentile wage. This merged file 490 occupations is the analysis dataset; the full table appears in the accompanying spreadsheet.

The automation source originally assigned each occupation to one or more of 14 skill categories, presented as indicator columns. However, two of the columns were not usable as true skill indicators due to AI limitation. One of the columns treated “Routine Cognitive”, a skill type, as another occupation and tried to assign it skill indicators, causing duplication and nonsense assignment. The other column was a catch-all list that duplicated hundreds of occupations and mixed category labels in with job titles. After removing these two columns, the remaining 12 columns represented shared skill categories.

An occupation was retained for skill-based analysis only if it appeared in exactly one of these 12 clean columns, ensuring a unambiguous, single-category assignment per occupation. This produced a distinct skill type for 222 of the merged occupation. Because some of the 12 categories contained very few occupations, these were collapsed into four broader groups for statistical analysis: (1) Manual and Routine Technical, (2) Routine Cognitive & Clerical & Sales, (3) High-Skill Analytical & Scientific, and (4) Interpersonal & Care & Creative. This grouping preserves the theoretical distinction central to Hypothesis 2 (routine-cognitive vs. manual-rouutine) while maintaining adequate sample sizes for comparison.

One of the other limitation of this analysis is that the proposal listed a top 1 percent salary variable, but OEWS publishes wages only up to the 90th percentile. The 90th-percentile wage is therefore used as the upper-tail data, and the highest-paying occupations sometimes have their 90th-percentile value not published. Regarding the range of data, the wage data are California-statewide for 2025, not a national sample as the original proposal suggested. Findings describe the California occupational structure; generalization beyond it should be cautious. Moreover, these data are closer to a census of occupations than a random sample, which affects how the inference below should be interpreted as discussed in the Conditions section.

Exploratory Data Analysis

Numeric summaries of the three quantitative variables across the 490 merged occupations:

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max
Mean annnual wage $	490	$84,158	$41,816	$36,653	$53,857	$73,854	$100,471	$329,275
90th percenntile wage $	467	$110,773	$47,623	$36,993	$73,874	$100,961	$137,931	$240,274
Automation probability	490	0.561	0.370	0.003	0.140	0.700	0.900	0.990

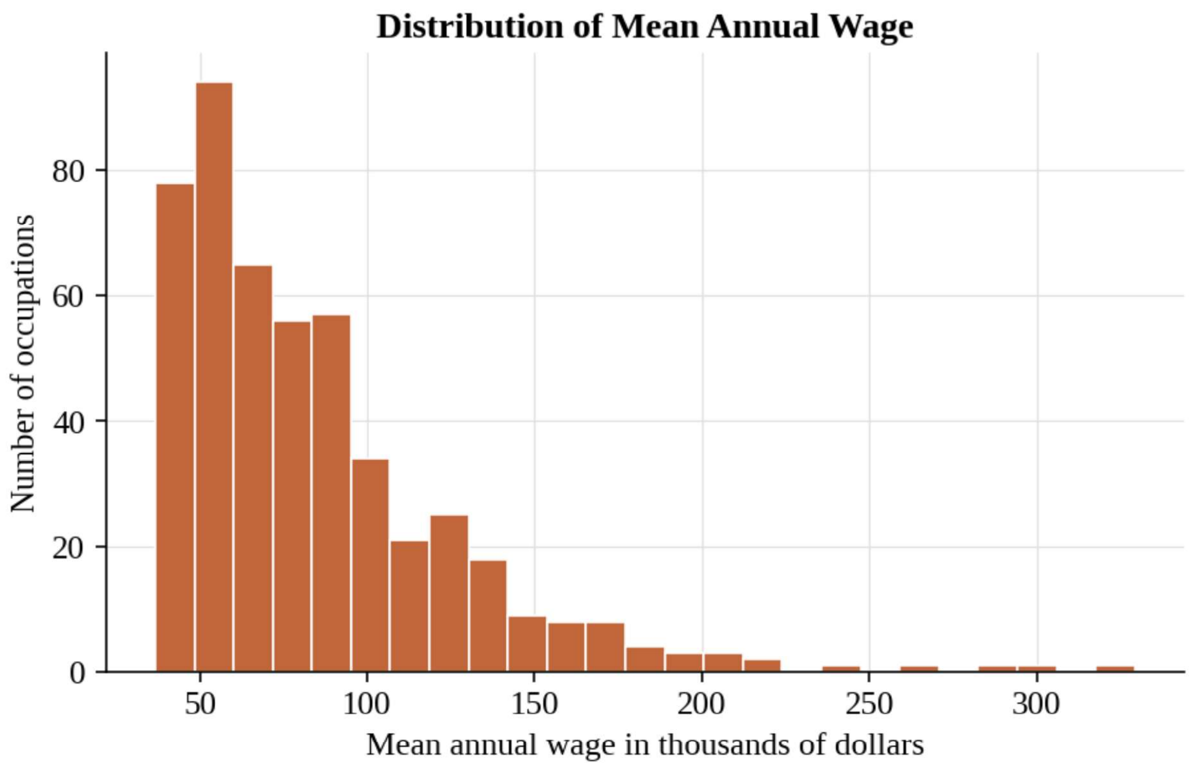

Figure 1. Mean annual wage is right-skewed. Most occupations sit in a low to middle band with a thin upper tail of very

The shape is right-skewed, with most occupations concentrated between 50,000 and 100,000 and a long upper tail extending to 329,275. The median is 73,854, and the spread is wide, with an interquatile range of roughly 46,600. No low-end outliers are present, as the minimum wage of 36,653 is close to the first quartile value but several high-paying occupation appear as outliers in the tail, such as physicians and chief executives.

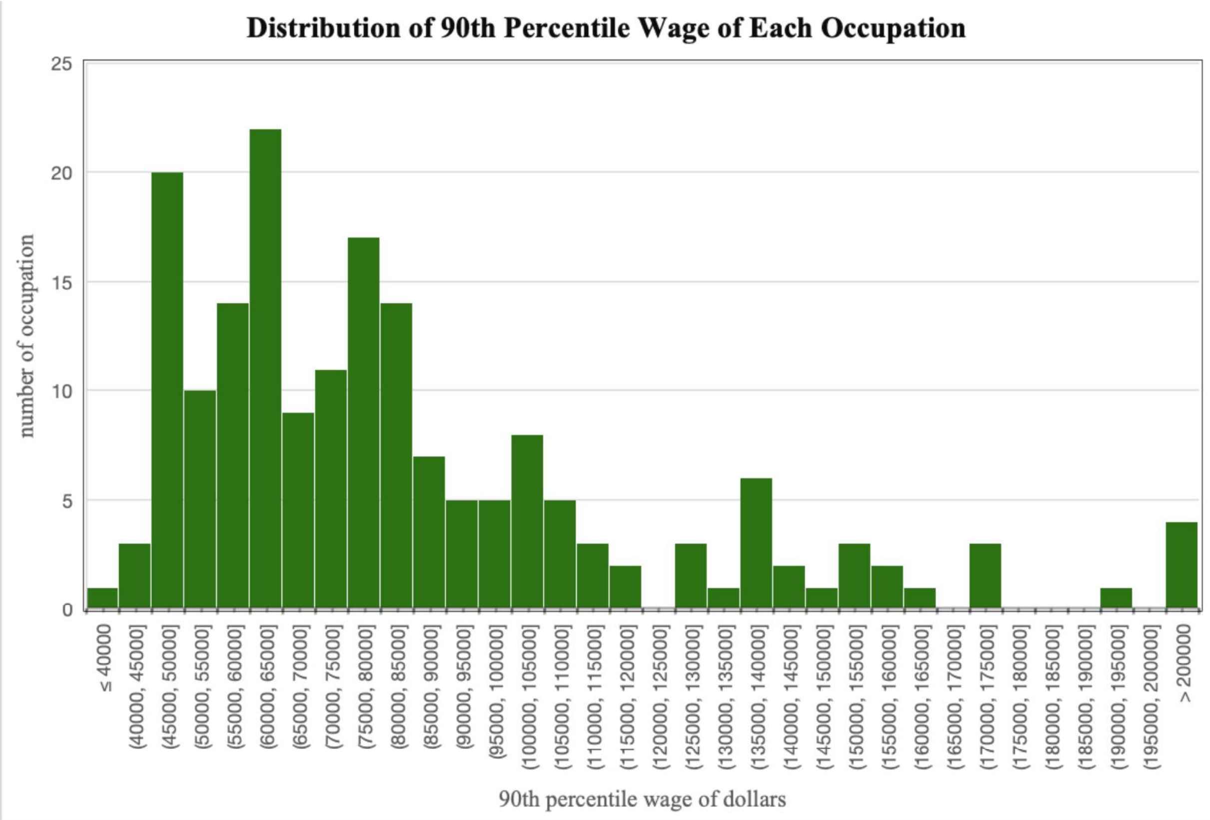

Figure 2. Distribution of 90th percentile wage of occupation. The upper-tail wage distribution is similarly right-skewed.

The 90th percentile wage is right-skewed with a median near 100,961. The spread is wider (IQR appromiximately 64,000) reflecting greater income gap in top-end earnings. Several high-outlier occupation reach above 200,000 and the maximum value is 240,274 however isn't shown due to the BLS suppression of exceedingly high values. This histogram shows a slightly more irregular pattern than mean wages, partly because 23 ocupations lack a reported 90th percentile value.

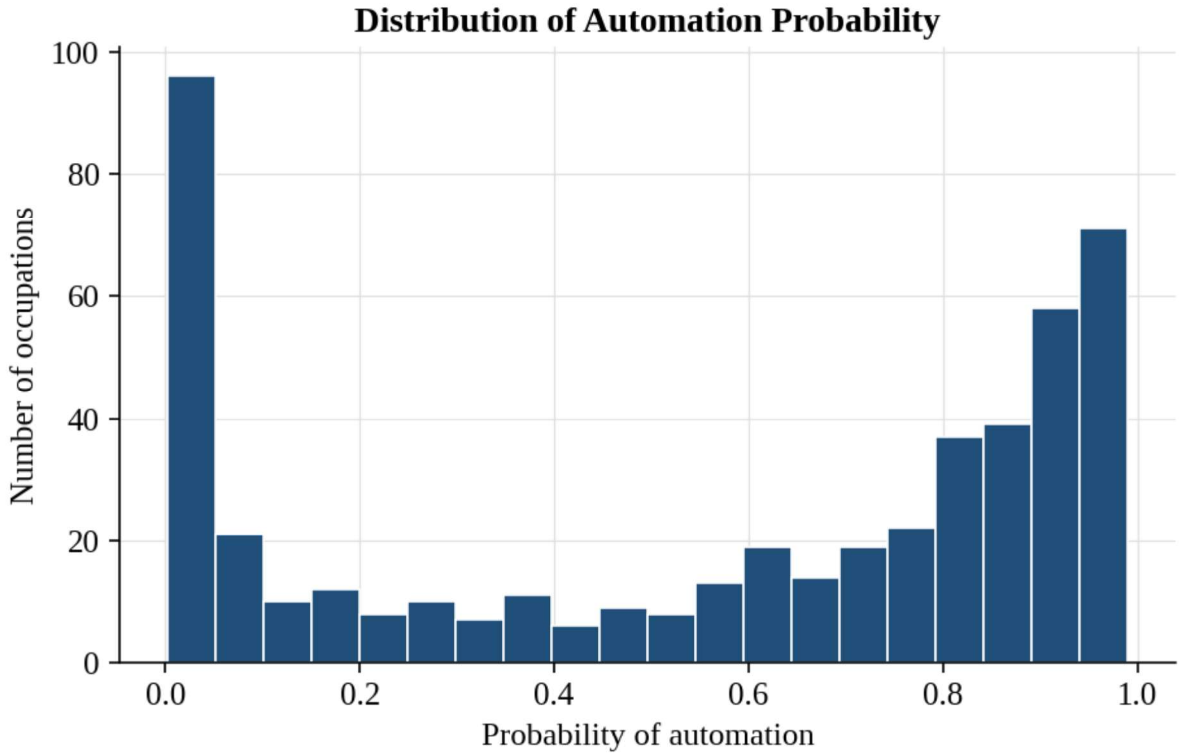

Figure 3. Automation probability is sharply bimodal: many occupations cluster near 0 and many near 1.

The histograms show that automation probability is U-shaped, with one peak near 0.0-0.1 and another near 0.9-1.0. Very few occupation fall in the middle range (0.3-0.7). The median is 0.7 ,but the mean (0.561) is pulled left by the low-risk mode. Spread is large with SD = 0.37 because the distribution clusters at both extremes rahter than around a single center. No outliers are present as the two modes themselves define the extremes of the 0-1 probability scale

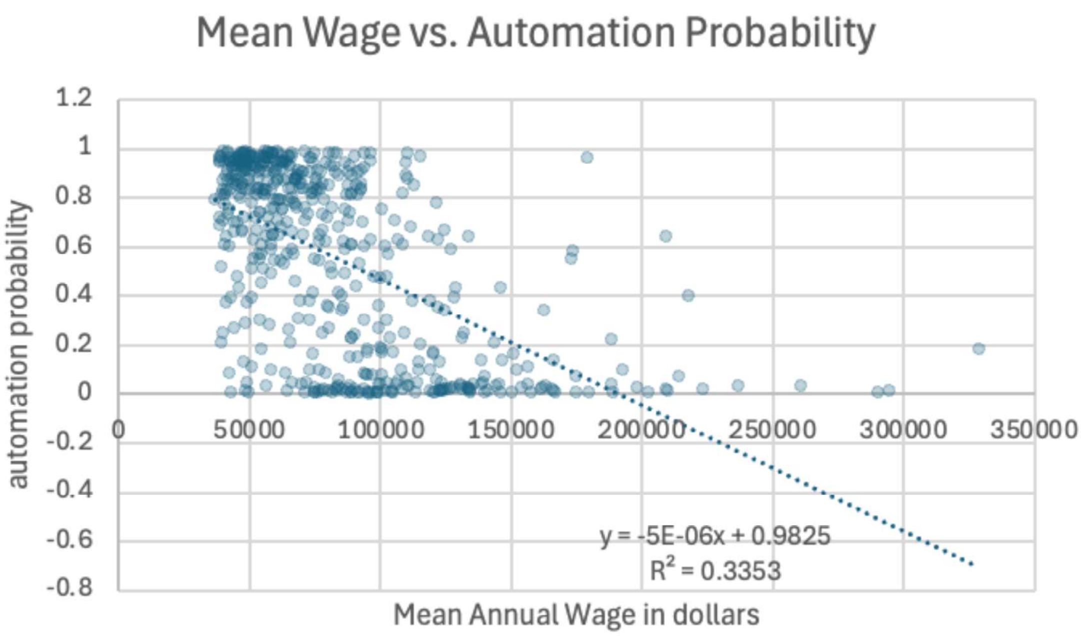

Figure 4. Mean wage versus automation probability with the least-squares line. The trend is clearly negative and spread widens at lower wage range.

Figure 4 presents a scatterplot of mean annual wage against the estimated automation probability for each occupation. The overall relationship is negative: occupations with higher average wages tend to have lower predicted automation risk. This pattern is consistent with Hypothesis 1.

The least-squares regression line fitted to the raw data has a negative slop, approximately -5.34*10e-6 per dollar, with a sample correlation of r-squared = 0.3644. This indicates a moderate negative linear association: higher wages are associated with lower automation risk, and mean wage explains about 36% of the variation in automation probability. However, the relationship is not perfectly linear. At lower wage levels, the spread of automation probability is wide. As wages increase, automation probability flatten toward zero, and the vertical spread lessens.

Moreover, it is worth noting that using 90^{th}-percentile wage instead of mean wage produces nearly identical results. Both the direction and strength of the relationship are strong in this choice. The correlation between mean wage and automation probability is r-squared = -0.3353; replacing mean wage with 90^{th} wage gives r-squared = 0.3856.

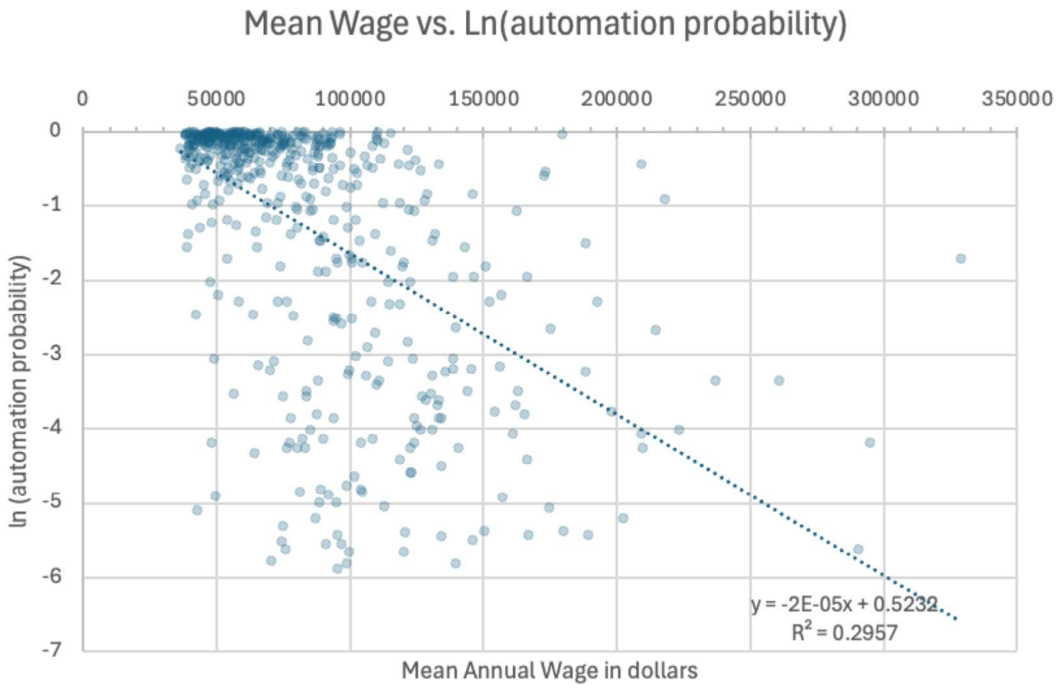

Figure 5. Scatterplot of mean annual wage versus ln(automation probability) with least-squares regression line

To seek for better illustration of the relationship between the two variables, Figure 5 displays the mean wage against the natural logarithm of automation probability. However, the log-transformed scatterplot doesn't reveal a substantially stronger linear relationship than the raw data. The slope remains negative, consistent with Hypothesis 1, but the R-square of Figure 5 is 0.2957 (compared to 0.3644 for the raw model). This suggests that other factors are likely driving much of the variation in automation probability, a point explored afterward.

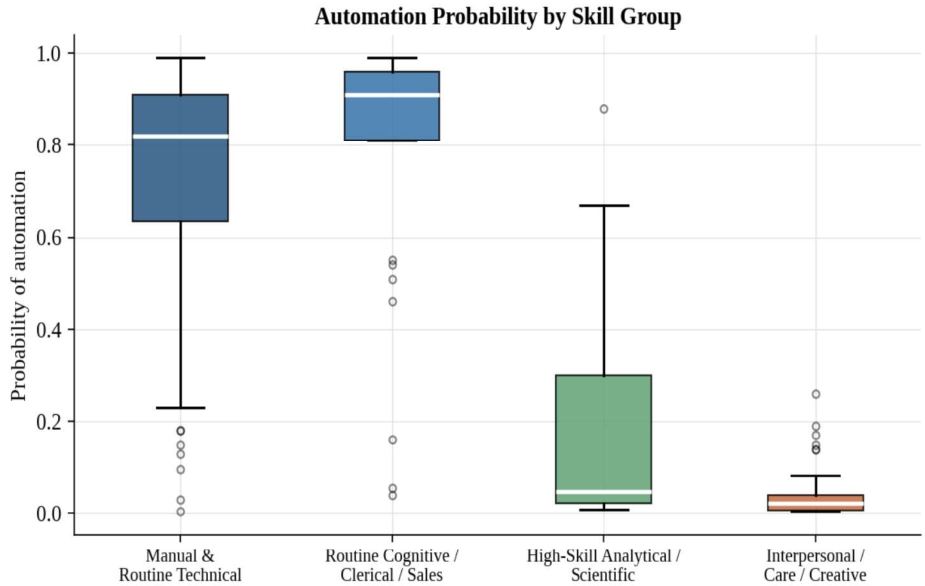

Figure 6. Automation probability by skill group. Manual and routine groups sit high while high-skill analytical and interpersonal groups sit low.

As wages rise, automation probability tends to fall Figure 3, consistent with Hypothesis 1. The relationship is moderately strong but not perfectly linear: at high wages, predicted risk flattens near zero rather than continuing downward, and the vertical spread shrinks. Figure 4 previews the skill-group story tested formally in Part 2C.

Inference

Test 1. Is salary linearly associated with automation risk?

H_0: $\rho = 0$, no linear relationship between mean wage and automation probability. H_a: ρ is negative, meaning higher wage is associated with lower automation probability.

- **Linearity:** the scatterplot in Figure 3 shows a negative trend, though with mild curvature because the response is bounded between 0 and 1. The linear model is a reasonable approximation.
- **Independence:** each row is a distinct occupation; observations are treated as independent.
- **Equal variance:** the residual plot in Figure 5 shows a funnel shape: spread is larger at lower wages, so the equal-variance condition is only partially met; results should be read as strong evidence of direction rather than a precise effect size.
- **Normality of residuals:** approximately satisfied with mild skew, and with n = 490 the sampling distribution of the slope is robust.
- **Randomness:** the data are effectively a census of California occupations, not a random sample. The p-value is interpreted as the probability of seeing a relationship this strong if there were truly none.

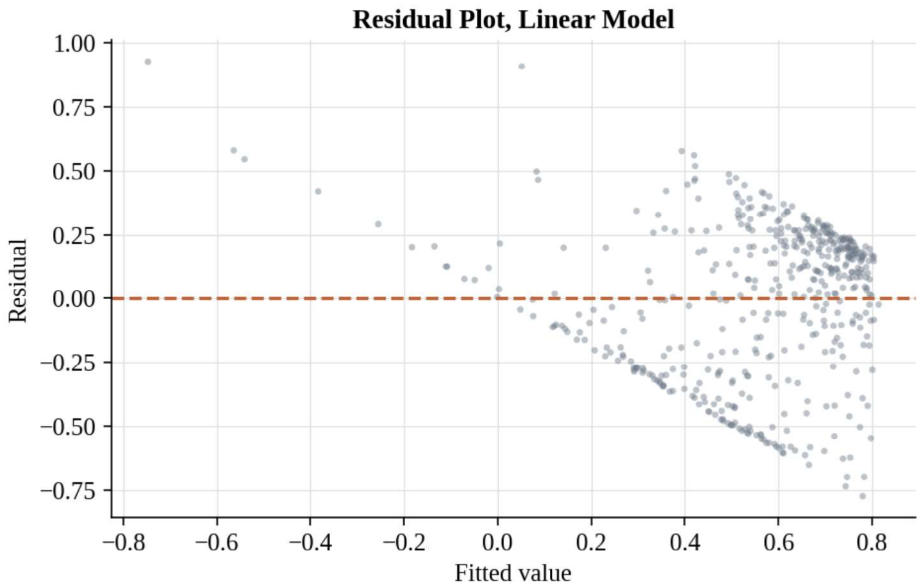

Figure 7. Residual plot for the linear model. The funnel pattern indicates non-constant variance.

Least-squares regression of automation probability on mean wage gives r = -0.6036, R^2= 0.3644, slope b = -5.34e-6 per dollar, with t = -16.7262 on df = 488 and $p < 0.0001$. Repeating the test with the 90th-percentile wage gives a very similar result r = -0.6212, t = -17.0915, $p < 0.0001$, so the conclusion does not depend on which wage measure is used.

Explanatory variable	r	R^2	t	df	p-value
Mean annual wage	-0.6036	0.3644	-16.7262	488	< 0.0001

90th-percentile wage	-0.6212	0.3858	-17.0915	465	< 0.0001

Because $p < 0.0001$ is far below $\alpha = 0.05$, we reject H_0. There is very strong evidence of a moderate negative linear association: higher-paying occupations tend to have lower automation probability, and wage explains about 36% of the variation in automation risk.

Test 2. Does automation risk differ by skill type?

H_0: mean automation probability is equal across the four skill groups. H_a: at least one group mean differs.

It is worth noting that although thte original skill classification contained 12 distinct categories, many of these contained too few occupations to support reliable statistical testing. For the purpose of Test 2 and the boxplots, these 12 categories were therefore combined into four larger groups: Mnual and Routine Technical, Routine Cognitive & Clerical & Sales, High-Skilled Analytical & Scientific, and Interpersonal & Care & Creative. This grouping maintains the theoretical distinction between routine-cognitive and routine-manual work central to Hypothesis 2 while ensuring sample sizes suitable for comparison.

Independent groups of distinct occupations; group distributions roughly bell-shaped; group standard deviations are comparable except the interpersonal/care/creative group, which values cluster tightly near zero, noted as an outlier. A one-way analysis of variance is used.

Conditiond for one-way ANOVA:

- Independence: each occupation belongs to exactly one skill group; observations are independent
- Normality: distributions within groups are roughly bell-shaped except the interpersonal/care/creative group (tight near zero); ANOVA is robust with moderate sample sizes.
- Equal Variance: group SDs are comparable except interpersonal/care/creative (SD = 0.064 vs. others 0.24-0.28); this violates the assumption, so the F-test should be interpreted as strong evidence of a difference rather than an exact p-value
- Randomness: data are a census of California occupations, not a random sample; p-vale interpreted as evidence strength under a hypothetical sampling process

Skill group	n	Mean prob.	SD
Manual and Routine Technical	131	0.728	0.243
Routine Cognitive / Clerical / Sales	30	0.784	0.280
High-Skill Analytical / Scientific	25	0.191	0.249
Interpersonal / Care / Creative	36	0.045	0.064

F (variance between groups / variance within groups) = 114.0521 on df1 = 3 and df2 = 218, $p < 0.0001$. We reject H_0: automation risk differs strongly by skill type. Routine work, whether manual and routine-cognitive/clerical work, carries high risk, while high-skill analytical/scientific work , about 0.19, and interpersonal/care/creative work , about 0.05, carry low risk.

Conclusion

An occupation's salary relates strongly to its automation risk. Across 490 California occupations, mean wage and automation probability have a moderate, highly significant negative correlation $r = -0.6036$, $p < 0.0001$. Hypothesis 1 is supported: higher-paying jobs tend to be less automatable.

The prediction that cognitive-routine high-paying jobs would be more at risk than manual low-paying jobs is not supported as stated. The data show that routine-cognitive and manual work face similarly high risk, while the real dividing line is routine vs. non-routine: high-skill analytical work and interpersonal/care/creative work are both very low risk. Salary protects against automation largely because the best-paid occupations are concentrated in exactly those non-routine analytical and interpersonal roles, which is consistent with the idea that interpersonal and higher-order analytical skills, rather than pay itself, are what confer protection.

The linear model imperfectly captures a bounded, slightly curved relationship; the data are a California census rather than a random sample; and upper-tail pay is capped at the 90th percentile. A logistic or transformed model, national wage data, and post-hoc comparisons between skill groups would sharpen the analysis.

Appendix A. Data

The full merged dataset covers 490 occupations with SOC code, occupation, employment, mean wage, 90th-percentile wage, automation probability, fine skill type, and broad skill group. It is provided in the accompanying spreadsheet automation_salary_dataset.xlsx, Sheet Data. Sheet Summary reproduces the numeric summaries and inference results above.

Appendix B. Sources and AI Use

• U.S. Bureau of Labor Statistics, Occupational Employment and Wage Statistics 2025 California release. data.gov OEWS

• Frey, C. B. and Osborne, M. A. The future of employment, Technological Forecasting and Social Change. ScienceDirect article

AI Claude was used as a tool to merge the two datasets on SOC code (joining Frey/Osborne occupations with California wage records), to clean the messy skill-type categorization (classifying the occupations into 14 groups based on skill-types and combining the 12 categories into 4 broad skill groups), and to compute the regression, and ANOVA reported here. (the p-value and the F-statistic)