Modern Salary Modeling Project

Justin Mai, Tro Russo, Isaac Muhlestein, Conan Li, Jian Kang

2025-02-26

Project Overview: Modern Salary Modeling

Description: Our goal with this Modern Salary Modeling Project is to analyze modern-day jobs (most are related to AI) to attempt to make salary predictions, looking at the influence of different variables on job salaries. This notebook will use the F-statistic and a series of ANOVA tests (hypothesis and confidence intervals)to determine the best predictors for the job market of the future.

Table of Contents

Data Overview

To do this, we are using the AI-Powered Job Market Insights dataset. this dataset contains information about modern-day jobs regarding AI. The data consists of 500 job listings (observations) with different factors to describe each job. The data isn't from real-world jobs but it mimics jobs and roles seen in the job market. The general goal of using this dataset is to identify the categories most influential in determining job salaries.

https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights

Response Variable:

- `Salary_USD` (Numerical): The annual salary offered for the job in USD.

Predictor Variables:

- `Job_Title` (Categorical): Job position or role.
- `Industry` (Categorical): Field of employment.
- `Company_Size` (Categorical): Size of the company (small, medium, large).
- `Required_Skills` (Categorical): Qualifications for the role.
- `Automation_Risk` (Categorical): Risk of automation in the future (low, medium, high).

- `Location` (Categorical): City where the role is located.
- `Remote_Friendly` (Binary): Indicates whether the role supports remote work.
- `Job_Growth_Projection` (Categorical): Expected growth or decline of the role.

Cost of Living could also be a good indicator to a person's salary. Using an external data source, *Cost of Living Index*, we also want to see how salary is influenced by the cost of living at each correlating city. The data we are using looks at the cost of living indexes by city in 2022, where all the variables involved are numerical.

https://www.kaggle.com/datasets/kkhandekar/cost-of-living-index-by-city-2022

Predictor Variables:

- `Cost.of.Living.Index` (Continuous): Relative cost of living in different cities
- `Rent.Index` (Continuous): Relative cost of rent in comparison to other cities
- `Cost.of.Living.Plus.Rent.Index` (Continuous): Relative cost of living and rent in different cities
- `Groceries.Index` (Continuous): Relative grocery cost in different cities
- `Restaurant.Price.Index` (Continuous): Relative cost of eating out in different cities
- `Local.Purchasing.Power.Index` (Continuous): Purchasing powers of a city's average salary

Methodology

- Preprocess and clean the data, ensuring that there are sufficiently many observations for all levels of our factor variables (some of our variables have large numbers of levels, so we may have to pick and choose which levels to consider)
- Examine the data with graphs (scatter plot, histogram, TA plot, QQ plot) to explore the relationships between the response and the predictors, and see what transformations may be necessary. Also, check if the assumptions are met.
- Start with simple linear regression between each of the predictors and the response
- Use multiple regression with several combinations of different predictors and include interaction terms
- Perform F/Anova tests to see which predictors are needed for the strongest model, as well as which predictors have the most significant impact on the response variable

Research Questions

Research Question 1:

1.1. Evaluating Key Predictors

- **Objective**: Believing that `Job_Title`, `Company_Size`, and `Location` are the best predictors for `Salary_USD`, we will use a linear model to test this and describe the findings.

1.2. Interaction Between Company Size and Location

- **Objective**: Does the interaction between `Company_Size` and `Location` improve our linear model? This will help us assess if the effect of one predictor depends on the level of the other.

1.3. LLM Generated Predictor (Estimated Annual Cost of Living)

- **Objective**: Investigate the entire model and use LLMs (Local Linear Models) to display each person's (row) data. We will add this as a continuous predictor to the original linear model and analyze what this reveals about the relationships between variables.

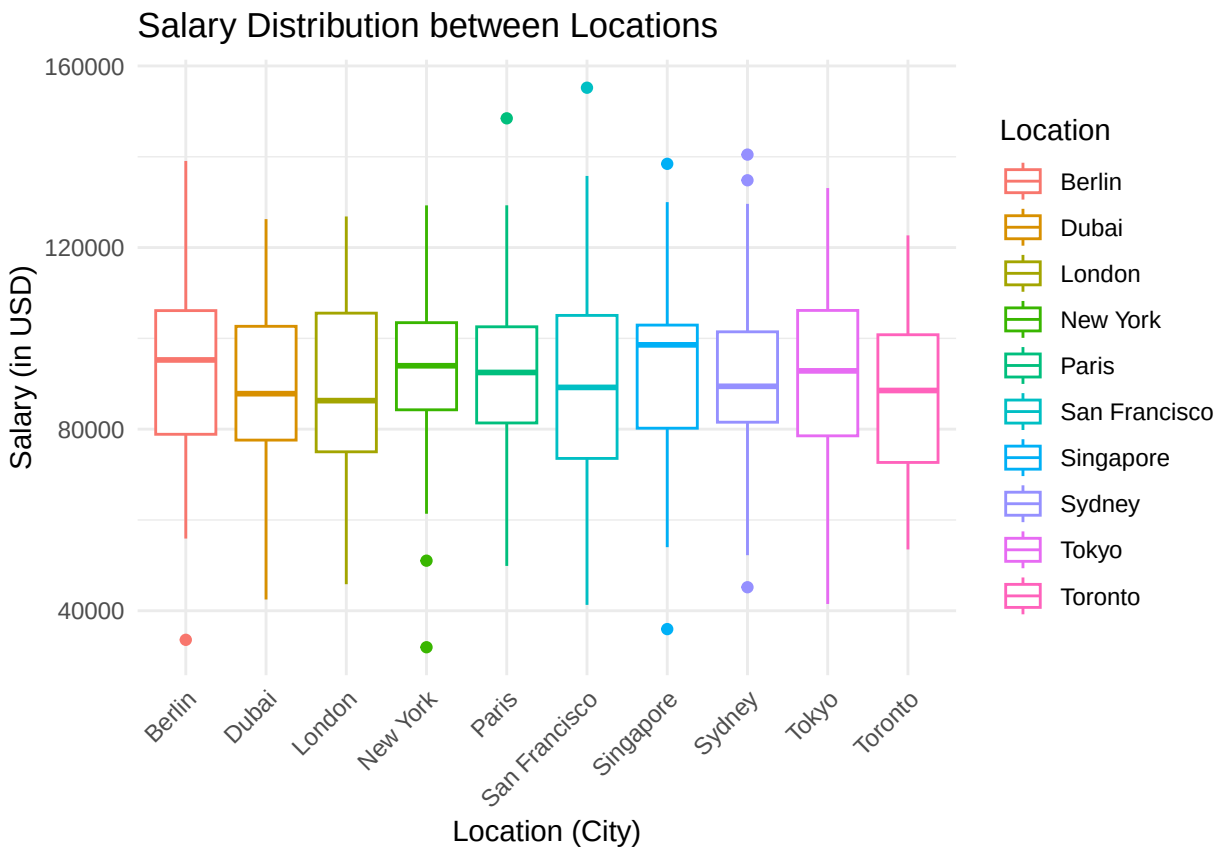
1.4. ANOVA Test to Evaluate LLM Predictor Significance

- **Objective**: Perform an ANOVA test to determine whether the continuous predictor generated by the LLM is significantly different from the original model in part 1.1.

Data Manipulation

Before developing our models, we first joined our `Job_Market` dataset with our `Cost_Living` dataset to gather all of the necessary variables in one dataset. Having all of our predictors in one dataset allows us to conduct `lm` models to compare variables across multiple sources of data. This will help us determine the predictors that's most deterministic of the variation in `Salary_USD`.

```
# Distribution of Salary between Locations (Cities)
joined_df %>%
  group_by(Location) %>%
  ggplot(aes(Location, Salary_USD, col=Location)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Salary Distribution between Locations",
       y = "Salary (in USD)",
       x = "Location (City)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Salary Distribution between Locations
160000
120000
80000
40000
Salary (in USD)
Berlin
Dubai
London
New York
Paris
San Francisco
Singapore
Sydney
Tokyo
Toronto
Location (City)
Location
Berlin
Dubai
London
New York
Paris
San Francisco
Singapore
Sydney
Tokyo
Toronto

Results

1.1. Evaluating Key Predictors

- Believing that `Job_Title`, `Company_Size`, and `Location` to be the best predictors for `Salary_USD`, use a linear model to test this and describe findings

```
##
## Call:
## lm(formula = Salary_USD ~ Job_Title + Company_Size + Location,
##     data = joined_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62719 -12614    840  13287  67912
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     98074.9     4431.1  22.133  < 2e-16 ***
## Job_TitleCybersecurity Analyst  -4381.7     3878.0  -1.130  0.25904
## Job_TitleData Scientist         -4385.6     3793.4  -1.156  0.24815
## Job_TitleHR Manager             -4438.0     3820.2  -1.162  0.24588
## Job_TitleMarketing Specialist   -4211.6     4031.9  -1.045  0.29670
## Job_TitleOperations Manager      2091.7     4111.8   0.509  0.61117
## Job_TitleProduct Manager        -4290.4     4257.6  -1.008  0.31406
## Job_TitleSales Manager          -2339.3     3966.8  -0.590  0.55563
## Job_TitleSoftware Engineer     -11591.7     4210.2  -2.753  0.00611 **
## Job_TitleUX Designer            -6054.8     3982.0  -1.521  0.12897
## Company_SizeMedium              -1771.9     2178.2  -0.813  0.41631
## Company_SizeLarge               -1888.7     2171.1  -0.870  0.38473
## LocationDubai                   -4914.2     4162.2  -1.181  0.23826
## LocationLondon                  -4489.7     3680.0  -1.220  0.22301
## LocationNew York                  998.9     4218.7   0.237  0.81292
## LocationParis                    -254.0     4275.2  -0.059  0.95265
## LocationSan Francisco           -4793.9     3988.1  -1.202  0.22988
## LocationSingapore                 873.7     4112.1   0.212  0.83182
## LocationSydney                   -723.2     4140.0  -0.175  0.86139
## LocationTokyo                     249.6     4164.0   0.060  0.95222
## LocationToronto                 -4120.7     4419.0  -0.932  0.35151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20530 on 525 degrees of freedom
## Multiple R-squared:  0.03736,    Adjusted R-squared:  0.0006913
## F-statistic: 1.019 on 20 and 525 DF,  p-value: 0.4373
```

Letting β_i be a predictor to estimate

$$\hat{Salary\_USD} = \beta_0 + \sum_a \beta_{1a}(\text{JobTitle}_a) + \sum_b \beta_{2b}(\text{CompanySize}_b) + \sum_c \beta_{3c}(\text{Location}_c)$$

, where a, b, c represents the specific categories. Since we are dealing with categorical variables, the value of each coefficient would be binary (0 or 1), which means the job description matches a specific set of variables or not.

This initial OLS model shows us how well the 3 categorical predictors we have chosen explains the variation of the variable `Salary_USD`. Our intercept is representative of our reference level, which in this case would be `Job_Title = AI Researcher`, `Company_Size = Small`, and `Location = Berlin`. As we can see, we have a very small R^2 value of 0.03736, meaning our predictors or independent variables doesn't explain a large proportion of the variance in salary, suggesting high variability in salaries.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Using a significance level of $\alpha = 0.05$ we can see that only one predictor variable `Job_Title: Software Engineer` is statistically significant. Even though our current predictors may not be strong determinants of `Salary_USD`, we still want to attempt to create predictors that may lead to a better linear model

1.2. Interaction Between Company Size and Location

- Does Company_Size and Location require an interaction term? Does this improve our linear model?

  To create a better model, we need to determine which predictors require an interaction term. This means that we are looking at the if one of our predictors depends on another predictor when analyzing `Salary_USD`. Specifically, seeing if our predictor `Location` is impacted by the size of the company. To do this, we will use a side-by-side plot to see if the pattern demonstrated for each `Company_Size` is different from one another, if they are, we will use an interaction term between those two predictors

```
joined_df %>%
  ggplot(aes(x = Location, y = Salary_USD, fill = Company_Size)) +
  geom_boxplot() +
  facet_wrap(~ Location, scales = "free_x") +
  theme_minimal() +
  labs(title = "Salary Distribution by Company Size and Location",
       x = "Company Size",
       y = "Salary (USD)",
       fill = "Location") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

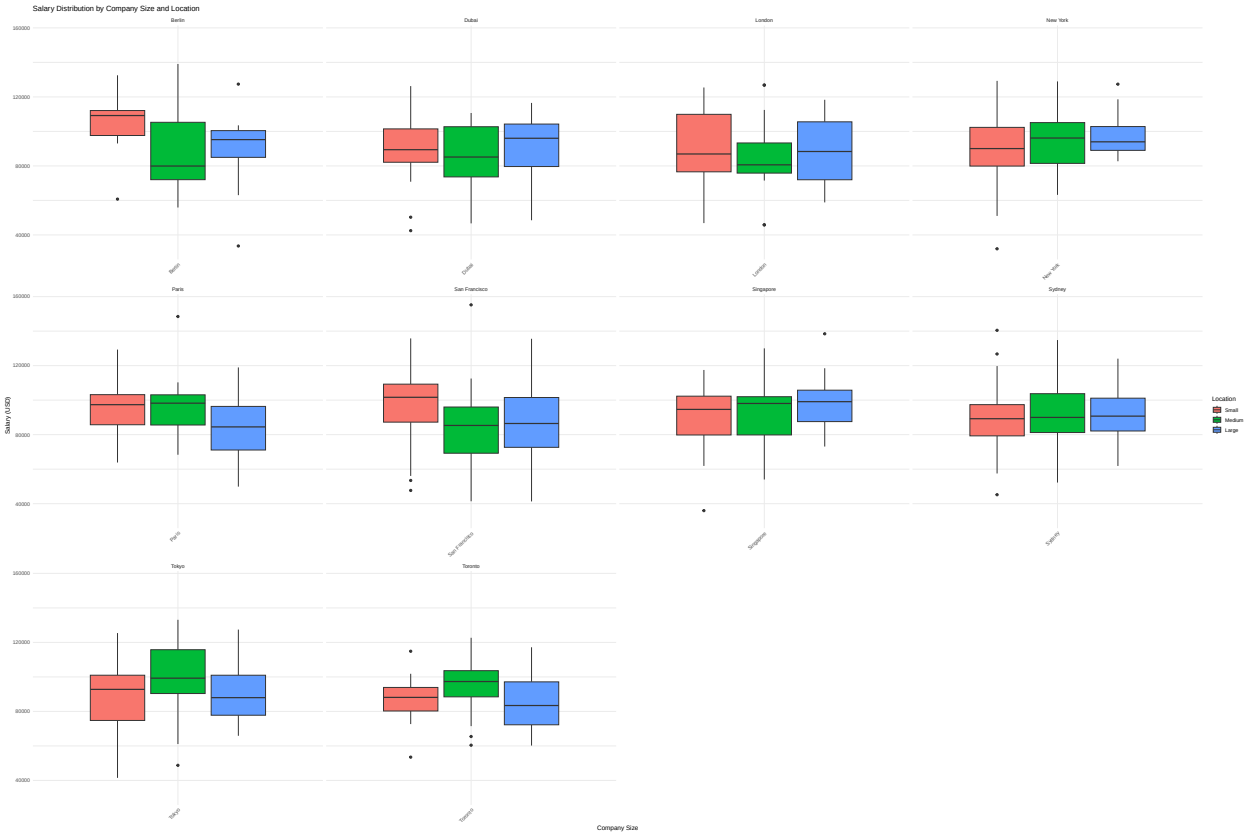


Figure 1: Compares the Salary Distribution by Location and Company Size

In the figure, it is demonstrated that company size does impact the employee's salary in every country. Within all the countries seen in *Figure 2*, when grouping them by company size, they demonstrate a different distribution and patterns (except for maybe Sydney, the distributions seem fairly similar across all company sizes). This indicates that it is necessary to use an interaction term to improve our initial model

```
q1_fit2 = lm(Salary_USD ~ Job_Title + Company_Size*Location, data=joined_df)
cat("R-Squared: ",summary(q1_fit2)$r.squared, "\n")
```

```
## R-Squared:  0.07193869
```

```
cat("Adj. R-Squared: ", summary(q1_fit2)$adj.r.squared, "\n")
```

```
## Adj. R-Squared:  0.002379854
```

```
cat("P-value: ", summary(q1_fit2)$fstatistic[1])
```

```
## P-value:  1.034214
```

The equation representing the new model would be

$$\hat{Salary\_USD} = \beta_0 + \sum_a \beta_{1a}(\text{JobTitle}_a) + \sum_b \beta_{2b}(\text{CompanySize}_b) + \sum_c \beta_{3c}(\text{Location}_c) + \beta_4(\text{Location}_c \times \text{CompanySize}_b)$$

This is the same equation as our originial model, except we are including the interaction term.

Our new model explains more of the variation in `Salary_USD` but not by much. The p-value is still relatively high so the predictors we are using doesn't appear to be significantly improving the model. Our $R^2 = 0.07194$ which is a higher proportion of the variance explained but its not enough to suggest that adding the interaction term led to a good fit.

1.3. LLM Generated Predictor (Estimated Annual Cost of Living)

- Look at the entire model and use LLMs to display each person's (row). Add it as a continuous predictor to the original linear model. What does this tell us?

1.4. ANOVA Test to test LLM predictor significance

- Use an ANOVA test to see if the continuous predictor generated by the LLM is significant against the original model in part 1.1.

Conclusion