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.

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.

Table of Contents

Data Overview

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.

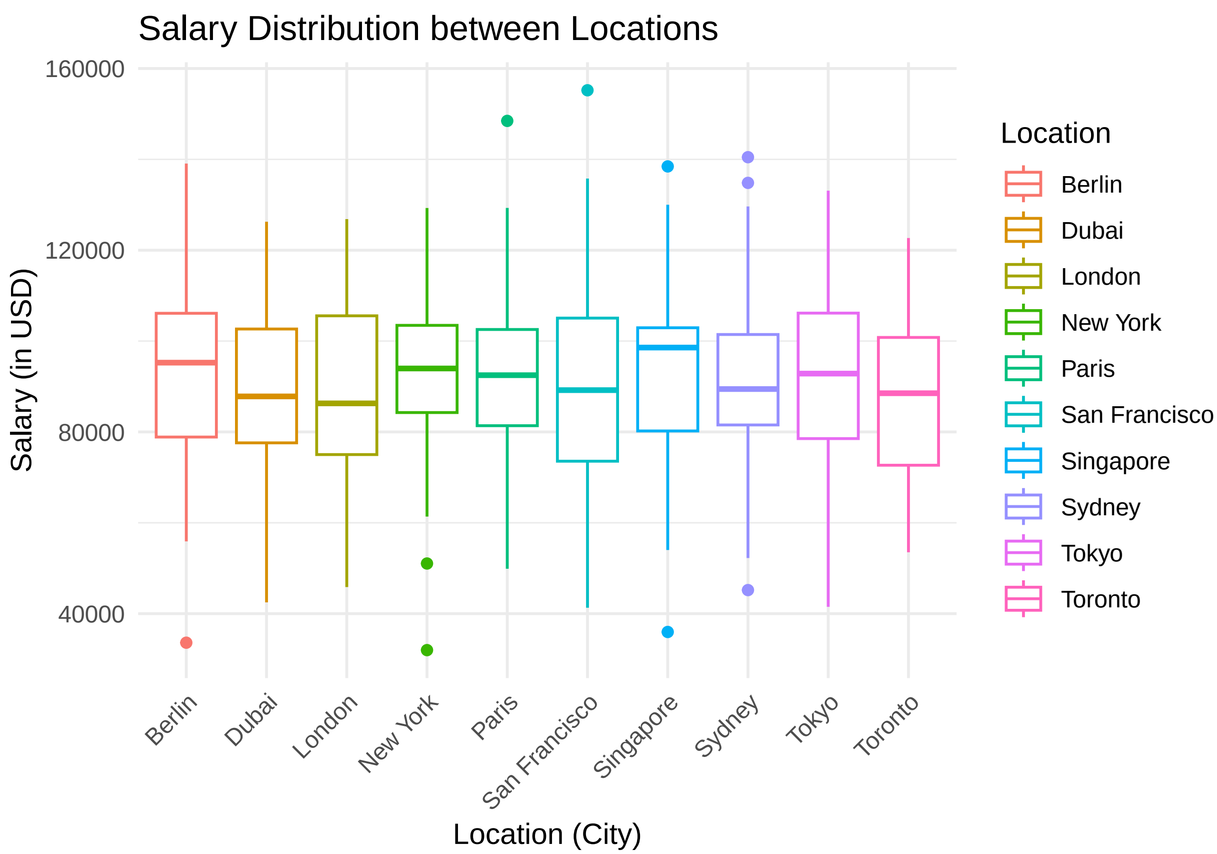
https://www.kaggle.com/datasets/kkhandekar/cost-of-living-index-by-city-2022

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`.

```r
# Joining Modern Job Market and Cost of Living Data
# Splitting Location Variable into `Location` which represents city and `State/Country`

cost_living <- cost_living %>%
  separate(City, into = c("Location", "State/Country"), sep = ",")

cities <- c(unique(job_market$Location))

cost_living <- cost_living %>%
  filter(Location %in% cities )

joined_df <- job_market %>%
  left_join(cost_living, by=c("Location"))

joined_df <- joined_df %>%
  select(!c(Rank, `State/Country`))

joined_df <- joined_df %>%
  mutate(across(where(is.character), as.factor))
```

```r
# 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))
```



Methodology

```r
joined_df %>%
  group_by(Location) %>%
  summarise(mean = mean(Salary_USD))
```

```
## # A tibble: 10 x 2
##    Location        mean
##    <fct>          <dbl>
##  1 Berlin        93240.
##  2 Dubai         87892.
##  3 London        88811.
##  4 New York      93780.
##  5 Paris         92116.
##  6 San Francisco 88953.
##  7 Singapore     93740.
##  8 Sydney        91885.
##  9 Tokyo         92897.
## 10 Toronto       88840.
```

```r
x_vars <- c("Cost.of.Living.Index", "Rent.Index", "Cost.of.Living.Plus.Rent.Index",
            "Groceries.Index", "Restaurant.Price.Index, Local.Purchasing.Power.Index")
```

Research Questions

Results

Conclusion