ISyE 3030 Project: Exploring the Relationship Between Gender and Wages

Team members' names: Dibo Antebi, Alexandra Guahnich, Raquel Ovadia, Aviel Pearlman, David Silvera

I. Introduction

The goal of this project is to analyze if there's a relationship between average salary earned in the United States and a number of factors including age, gender, years of experience, and education level. We attempt to use our knowledge in Python (Pandas DataFrames) to extract insights from our dataset and model the situation. Using this methodology, we will apply all of the concepts learned throughout the semester in our ISYE 3030 class to a real-world situation and further leverage the power of statistics.

II. Data Collection and Preprocessing

We sourced this dataset from Kaggle. After downloading the dataset, we formatted the data utilizing Python and Pandas dataframes. Pandas is a tabular data analysis and manipulation tool useful for projects like these which require the cleaning, aggregation and modification of large datasets in an effort to draw statistical conclusions.

The dataset contains data from 373 respondents. Our dataset includes a wide array of variables in each person that could potentially help us to draw correlations/conclusions on a certain number of them. Some of the variables (per employee) are: gender, age, years of experience, job type, education level (or highest education level attained), and salary. In an effort to simplify the dataset we're using, we removed all extraneous variables, which in our case includes only job type.

For practical purposes, we dropped all entries that included null values and renamed specific columns. We also deep-copied our dataframe to create a separate copy whose entries were all numerical. This made it easier for the quantitative portions of the assignment. The original copy preserved its alphanumeric entries, which we used to create more clear, readable visualizations.

III. Descriptive Statistics

We utilized Excel for the majority of our descriptive statistics, and Python for the visual tools of the data. We found the following data (*Figure 1*) across 194 males and 179 females that responded to the survey:

Figure 1: Descriptive Statistics

	Female	Male
Mean	97,011.17	103,867.78
Median	90,000	97,500
Mode	40,000	40,000
Range	155,000	249,650
Q1	50,000	60,000
Q3	140,000	140,000
Variance	2,108,292,009	2,518,279,436

This data was used to build the box plot (*Figure 2*).

Figure 2: Salary Distribution by Gender

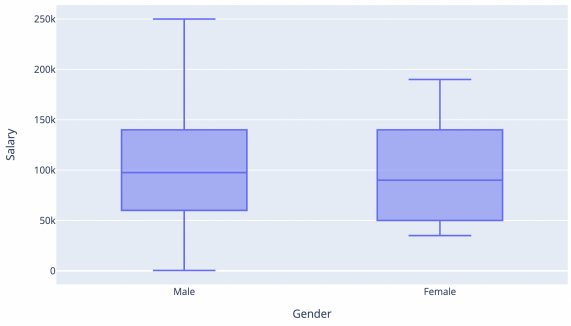
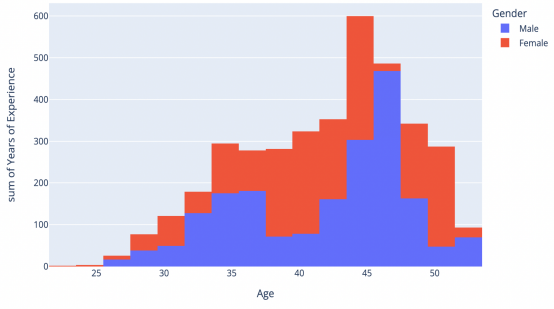


Figure 3: Years of Experience by Age



Furthermore, we can observe that the histogram illustrating years of experience by age (*Figure 3*) is left skewed, meaning that the mean is less than the median. This also means that the salary for younger people is lower than those of older people.

IV. Statistical Inference

We will explore if there is a difference between the mean salary for men and women. We want to be 95% sure of our conclusions.

1. Parameter of interest: μ_F and μ_M
2. H_0: $\mu_F - \mu_M = 0$
3. H_1: $\mu_F - \mu_M \neq 0$
4. Test statistic: $Z_0 = \frac{(\bar{X}-\bar{Y})-\Delta_0}{\sqrt{\frac{\sigma_F^2}{n_F}+\frac{\sigma_M^2}{n_M}}} = \frac{(\bar{X}-\bar{Y})-0}{\sqrt{\frac{\sigma_F^2}{n_F}+\frac{\sigma_M^2}{n_M}}}$

   This is an appropriate test statistic because our data follows a normal distribution, and the variance of both variables is known.
5. Reject H_0 if $|Z_0| > Z_{\alpha/2}$
6. Compute:

Female	Male
$\bar{X} = 97,011.17$	$\bar{Y} = 103,867.78$
$\sigma_F^2 = 2,108,242,009$	$\sigma_M^2 = 2,518,279,436$
$n_F = 179$	$n_M = 194$

$$Z_0 = \frac{(\bar{X}-\bar{Y})-0}{\sqrt{\frac{\sigma_F^2}{n_F}+\frac{\sigma_M^2}{n_M}}} = \frac{-6856.61}{4975.84} = -1.378 \qquad Z_{\alpha/2} = Z_{0.025} = 1.96$$

$$1.378 < 1.96$$

7. Conclusion: Thus, we fail to reject H_0 because $|Z_0| < Z_{\alpha/2}$. We do not have enough evidence to say that there is a difference between the mean salary for males and females.

V. Regression Analysis

For the regression analysis, we first have to estimate the regression coefficients so that we can develop our fitted regression model $\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$.

To find $\widehat{\beta_1}$ (slope estimate) and $\widehat{\beta_0}$ (intercept estimate), we use the Method of Least Squares, which consists in minimizing the sum of the squares of the vertical deviations.

$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{109,470,385.52}{15,993.9} = 6,844.51$ $\qquad \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = 100,577.35 - 6,844(10.031) = 31,920.2$

We can observe that there is a slope of 6,844.51, and an intercept of 31,920.2. The slope means that there is a positive correlation between years of experience and salary. In other words, as years of

experience increase, salary also increases. Regarding the intercept, our line of best fit assumes that the base salary is of about $31,920.20.

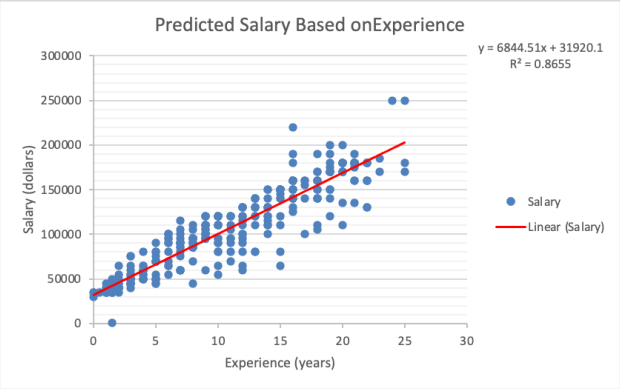
Moreover, our final fitted regression model (*Figure 4*) for average salary (y) given the years of experience (x) is

$$\widehat{y_i} = 31,920.2 + 6,844.51x_i.$$

Now, we assess the adequacy of our model with R^2 and ANOVA. First, we calculate R^2 to determine how well our data fits the regression model.

$$R^2 = \frac{\widehat{\beta_1} S_{xy}}{SS_T} = \frac{6{,}844.51 \cdot 109{,}470{,}385.52}{865{,}685{,}668{,}572.39} = 0.8655$$

Figure 4: Fitted Regression Model



This result suggests that our model accounts for 86.55% of the variability in the data, which means that our fitted model is highly accurate.

Then, we test the ANOVA approach to evaluate whether there is a relationship between our variables. Moreover, our null hypothesis H_0 for this test is that $\beta_1 = 0$. The test statistic for the significance of the regression is F_0, where

$$F_0 = \frac{SS_R / 1}{SS_E / (n-2)} = \frac{MS_R}{MS_E} = \frac{749{,}271{,}148{,}395.5}{116{,}414{,}520{,}176.89/(373-2)} = 2387.84$$

The null hypothesis should be rejected if $F_0 > F_{\alpha,1,n-2}$. We found that $F_{\alpha,1,n-2} = 3.87$.

Since $F_0 > F_{0.05,1,371} = 2387.84 > 3.87$, we reject H_0. This means there exists a correlation between the two variables (gender and salary), as the slope is not equal to 0.

VI. Conclusion

We found a strong positive correlation between average years of experience and average salary earned in a year for a representative population of people in the United States. This can be proved from our regression analysis, which yielded a coefficient of determination R^2 of 0.8655, a considerably high value. This is further supported by our analysis of variance (ANOVA), which undeniably shows that there's significance between these two factors.

Additionally, contrary to popular opinion, we show that there's not a significant difference between the average salary earned by males and females in the United States. This is supported by our hypothesis test, where our null hypothesis, stating that there's no difference across salaries, can't be rejected.

Some limitations we encountered with our project were the inconsistencies in the dataset we selected, which included several null values and redundant information. In the future, we would be more thorough in our dataset selection process, ensuring our sources are reliable and not prone to errors.

VII. References

Salary Prediction dataset. (2023). Retrieved 25 April 2023, from https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer