Data Analysis Final Project

To: Brennan
From: Johann Krug
Subject: Final Project

Overview:

I chose to look more into job salaries for my project and some of the issues and significant factors surrounding them. The issue with salaries is the widespread differences that continue to exist between different demographic groups, which feeds inequality and impedes the advancement of society. Despite systemic biases and unequal opportunities for women, gender pay gaps continue to exist. Additionally, as some people earn disproportionately higher salaries than others, social divisions are worse, and income inequality grows. Economic disparities are worse as well by problems such as wage stagnation, inadequate minimum wages, and unclear compensation structures. These difficulties highlight the necessity of all-encompassing organizational procedures and policies that support fair, open, and equitable wage structures to address the underlying causes of these issues and promote a more inclusive and just society.

Establishing fair and equitable salary practices is crucial to advancing economic stability, social justice, and well-being. In addition to ensuring that people are paid fairly for their labor, addressing problems like gender pay gaps and income inequality also helps create a more inclusive and sustainable society by promoting equal professional and personal growth opportunities. Putting fair and transparent pay structures in place is a critical first step in creating a more just society where people can prosper regardless of their identity or background.

To better understand the data I used, I combined descriptive and inferential statistics in my data analysis. I could appreciate the typical values and variations within various groups using descriptive statistics like mean and standard deviation. However, using inferential statistics such as regression, ANOVA, and t-tests, I could extrapolate my sample's results to the entire dataset and draw more general conclusions. Thanks to these analyses, I understood the available patterns in my data and the relationships between variables and group differences. I created a more thorough picture of the patterns and subtleties in my dataset by combining these statistical techniques, which enabled me to interpret the data with better knowledge.

Null and Alternative Hypothesis:

The null hypothesis for the regression test is that the coefficients for all job titles are equal to zero. The corresponding alternative hypothesis would suggest that at least one of the coefficients is different from zero.

The null hypothesis for the ANOVA test of Salary vs Country is that the mean salary is equal across all five countries (Australia, Canada, China, UK, USA). The alternative hypothesis suggests that at least one mean is different.

The null hypothesis for the two sample t-test is no statistically significant difference in the average salaries between males and females. The alternative hypothesis is that there is a difference in mean salaries.

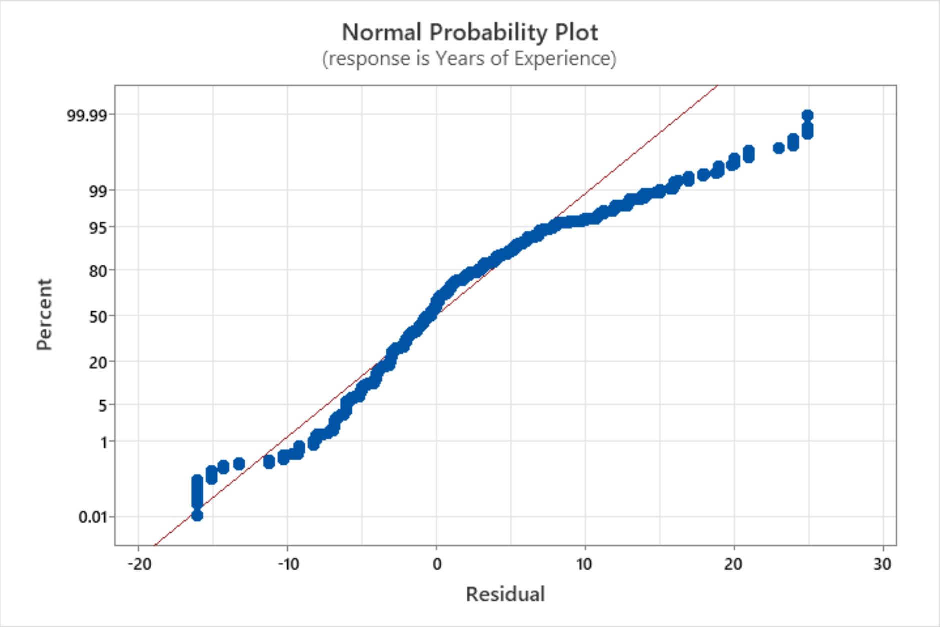
Data Description:

The data I am using has various numerical and categorical data. Years of Experience, Age, Salary, and Senior are among the numerical variables. These variables provide quantitative insights into the characteristics of the labor force. The continuous numerical variables Age and Years of Experience offer a range of values for every individual. Salary, a measurement of yearly income, provides a numerical indicator of the monetary component of each individual's job.

I have Gender, Education Level, Job Title, Country, and Race. These factors include discrete categories that divide people into different groups. The binary or multi-category variables capturing demographic data are likely gender and race. Education Level classifies people according to their level of education, giving information about the educational background of the workforce. Job Titles rank people according to their roles in the company, making it possible to investigate the relationships between various positions and other factors such as experience and salary. Lastly, Country adds a global dimension to the dataset by classifying people according to their geographic location.
I am hoping to find disparities, similarities, and significance in all of these different variables I am using.

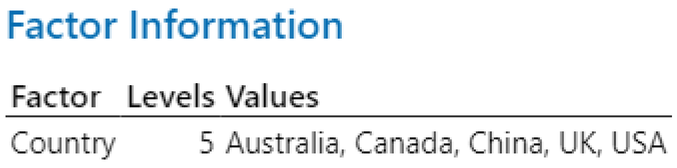
Data & Analysis Methodology:

Test 1: Regression (Experience vs Job Title)



Looking at the results of the linear regression, the analysis's primary goal is to determine how various job titles relate to variations in salary. The Constant represents the baseline salary, and the coefficient for each job title indicates the estimated salary change relative to this baseline. For example, a negative coefficient (such as -5.50) for the Account Manager implies a $5,500 reduction in average salary between the baseline and a positive coefficient (such as 13.00) for the CEO, which suggests an increase of $13,000. Higher absolute values indicate greater significance. The T-values demonstrate the statistical significance of these estimates. P-values help assess the probability of observing such results, assuming that the coefficient is zero, with values below 0.05 indicating statistical significance. Overall, this regression analysis clarifies the complex relationship between job titles and salaries, assisting in decision-making and proving that there is significance.

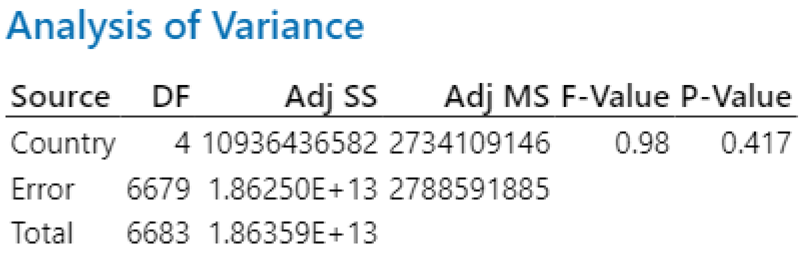
Test 2: ANOVA (Salary vs Country)



Factor Information

Factor	Levels	Values
Country	5	Australia, Canada, China, UK, USA

The Factor Information output shows that the variable Country in our dataset is what we're working with. The five levels of this categorical factor correspond to the following countries: Australia, Canada, China, the UK, and the USA. All data points in our set are associated with one of these nations. Any analysis involving data distinctions based on Country must comprehend this factor information. When examining salary disparities between nations, for example, it is helpful to have a comprehensive breakdown of the country factor to classify and analyze the data correctly. It also establishes the foundation for statistical analyses that consider these categorical factors, guaranteeing that our conclusions accurately reflect the subtleties associated with each Country in the dataset.

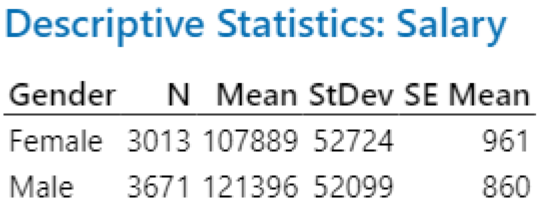


Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Country	4	10936436582	2734109146	0.98	0.417
Error	6679	1.86250E+13	2788591885		
Total	6683	1.86359E+13			

Looking at the one-way ANOVA, the Analysis of Variance stood out to me. In the analysis of variance, the F-value is computed to be 0.98, and the corresponding P-value is 0.417. A P-value of 0.417 suggests that the observed variability in the dataset associated with various nations is not statistically significant at a traditional significance level of 0.05. According to the current model, this implies no appreciable differences in the dependent variable's mean values among the countries considered. It's important to investigate additional factors that could significantly impact the dataset to improve our understanding of variability.

Test 3: Two-Sample T-Test (Salary, Gender)

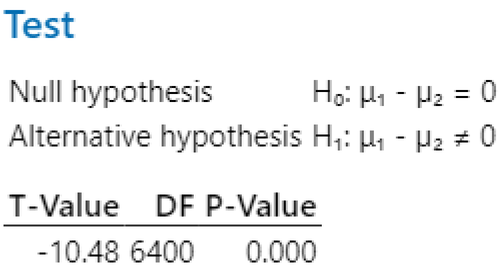


Descriptive Statistics: Salary

Gender	N	Mean	StDev	SE Mean
Female	3013	107889	52724	961
Male	3671	121396	52099	860

By examining the gender-specific descriptive statistics, we aim to determine whether men's and women's average salaries differ significantly. According to the dataset, we have 3671 observations for men and 3013 observations for women, which gives us a decent representation of each group. The mean salary, now divided by gender, is $107,889 for women and $121,396 for men. This implies that men appear to earn more money overall.

When the salary variability within each group is examined, the standard deviation for females is $52,724, and for males, it is $52,099. These figures are surprisingly close, suggesting that the distribution of salaries within both genders is comparable. Considering the accuracy of our mean wage approximations, the standard deviation of the mean for women is 961, and for men, it is 860. These figures indicate the degree of reliability of our average pay figures for each group.



Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-10.48	6400	0.000

The t-value measures how many standard errors the sample mean difference is from zero. In this case, the t-value is -10.48. The negative sign indicates that the mean salary for the first gender group (μ_1) is lower than the mean salary for the second gender group (μ_2). The results strongly suggest a significant difference in mean salaries between the two gender groups, providing evidence to reject the null hypothesis.

Conclusion:

To wrap up the three distinct statistical methods I used were linear regression, ANOVA, and a two-sample t-test—to glean insights into different facets of the data, enhancing the depth of my understanding.

Using linear regression analysis clarified the correlations between job titles and salaries. By examining coefficients, T-Values, and P-Values, I could determine the possible influence of various job titles on pay scales. Variance Inflation Factors (VIFs), which suggest possible multicollinearity among job titles, were included as a reminder to interpret the results carefully.

Next, in the ANOVA, I tried to figure out how the "Country" variable affected the variability of salaries. A more detailed analysis was provided by the degrees of freedom, adjusted sum of squares, and mean squares, which showed no statistically significant difference in mean salaries between different nations. This demonstrated the importance of taking into account other essential variables.

Finally, potential gender-based salary differences were examined using the two-sample t-test. The salary distributions within each gender were clarified by descriptive statistics, which prepared the groundwork for the following t-test. One of the critical factors in determining the statistical significance of the observed differences in mean salaries between males and females was the computed p-value. Combined, these techniques yielded a comprehensive analysis that highlighted areas for additional research and offered valuable insights for decision-making.

In conclusion, these three analytical techniques enabled a more nuanced interpretation of the dataset by revealing possible connections between job titles and salaries, examining the impact of factors unique to a given nation, and highlighting gender-based wage disparities. The results underscored the importance of utilizing various statistical techniques to thoroughly investigate the complex interactions among the variables in the dataset.