International and Domestic Salary

A consultation regarding the influential factors contributing to annual salary potential both domestically and internationally.

Report Prepared By:

Michael Alex Force

University of Oklahoma

Master of Science in Applied Statistics

5 August 2025

Executive Summary:

The results from the statistical analysis on salary earnings for individuals across varying countries concluded that age and years of experience moderately served to predict salary via a ridge regression model. Additionally, education, gender, and senior position holding proved statistically significant when predicting salary and differentiating between groups while race and country did not. Future work would aid in deducing how groups differ from one and other and could also be utilized in investigating an alternatively superior model to the presented ridge regression.

1. Introduction:

From prior work conducted, specifically through the U.S. Bureau of Labor Statistics, a multitude of influential factors contribute towards variation within occupations and towards dictating overall salary potential (U.S. Bureau of Labor Statistics., n.d.). Through a robust analysis of annual salary and contributing variables, a refined understanding of key attributes associated with higher economic earnings can be studied. While there are limitations to the influence an individual can have over specific variables, control over things like education and career field can serve to further an individual's annual salary. The aim of the following analysis targets determining the influence controllable, and intrinsic, characteristics have towards making educated decisions while pursuing economic growth. The primary objectives will be to address various questions, such as: what are the largest achievable factors that an individual can do for economic mobility, are there contributing factors that potentially influence salary but are marginal in their effects and how influential are uncontrollable factors? From a sense of reflection, applying said analysis towards annual salary serves to provide myself and others with a proactive approach in potential life decisions and mitigate unnecessary choices that might not be beneficial nor provide personal gain.

1.1 Data:

The data set "Salary" presented for said analysis consists of 9 variables and 6884 rows representing individuals and their respective salary, age, gender, education level, job title, years of experience, country, race, and whether they hold a senior position in their company or not. To the extent of information provided, the nature in which the sampling was conducted, including the years to which it applies, was not provided although given an interpretation of the variables in the data set, the information was sampled globally (Aboutalebi, A., 2024). The data set provides a comprehensive breakdown of salary in relation to a multitude of key variables that are highly relevant to an individual, the benefit of which allows for complex analysis and a more rigorous understanding of what combination of factors contributes the most to annual salary. Potential deficits are discussed in the summary statistics to follow but given the sample size in comparison to the population of a nation, a larger sample would serve to more accurately represent said population. This deficit is explicitly outlined when discussing the disproportionately high level of

educated individuals sampled for the data. To the extent of this misrepresentation, the scope of the analysis will focus on determining influential factors rather than the magnitude of earnings. Similarly, comparisons between categorical variables will serve to identify whether there is a statistically significant difference rather than quantifying the magnitude in the difference between groups.

2. Methodology:

Statistical analysis was conducted within the R Studio environment in various applications towards cleaning the data, performing testing, and producing figures for visualization while Excel aided in organizing the following tables. A multitude of visual and statistical applications were utilized including pie charts, histograms, and boxplots to easily visualize simple trends and relationships between variables. Anderson-Darling tests and Q-Q plots served as 2 supporting methods to determine if data was normally distributed which is a requirement for some statistical tests.

To analyze the relationship between salary and other variables, various regression models were assessed. Due to strong theoretical support, the continuous variables age and years of experience were tested for multicollinearity; an assessment that would indicate that these independent variables being used to predict salary are dependent on one and other. This was done through a Spearman correlation and variance inflation factor (VIF) to determine how strongly they were related. This was followed by a ridge regression model using age and years of experience to predict salary.

To address the remaining data types, these categorical variables were approached using an Analysis of Variance (ANOVA) test which assesses if there is a statistically significant difference between groups. ANOVA requires certain assumptions, as elaborated below, and for those that fail these assumptions alternative testing is required. In these cases, Kruskal-Wallis tests were utilized as a non-parametric alternative to perform similar testing to ANOVA, and that of comparing

groups, without violating underlying requirements of the testing. For the dichotomous variables gender and senior, a test for normality assessed whether a t-test or a Mann-Whitney U test was utilized in which both methods determine statistical differences between groups although differ in underlying assumptions and requirements of the data being assessed.

2.2 Preliminary Analysis:

Initial exploratory data analysis determined key insights and relationships within the data set. This included interpreting data types corresponding to variables which then dictated certain statistical tests that would be applied. In the case of salary, age, and years of experience, these were integers while occupation, gender, senior (indicating an individual held a senior position), race, country, and education level were all categorical variables used to represent different groups. From the preliminary analysis, general trends in the data were examined including proportions assigned to varying education levels and average salaries in relation to occupation title; this served to identify patterns and abnormalities within the data.

2.3 Tests for Normality:

Various tests were applied to determine if groups within the data were normally distributed; normality refers to the conventional “bell-shaped” curve when plotting data and is an assumption of the data for certain tests. Some common methods applied was an Anderson-Darling test, this is a hypothesis test that produces a p-value which is interpreted to say anything greater than acceptance criteria, typically at 0.05, would suggest the data is normal (I.E. larger p-values for Anderson-Darling suggest more normal data). Histograms also aided to easily visualize if data was normal where a resulting relatively bell-shaped curve would support normality. Additionally, Q-Q plots, representing quantiles (equally divided sections) for theoretically normal data compared to the questioned data would show a straight line for a normal distribution.

2.4 Ridge Regression:

To model the relationship in which salary is predicted by age and years of experience, a conventional method used is multiple linear regression. An underlying assumption of which is that the multiple variables used to predict the dependent variable (salary) are not dependent on one and other (I.E. age and years of experience predicting one and other). To determine the presence of this multicollinearity relationship, Pearson correlation coefficient or Spearman ranked correlation (depending on if the data is normally distributed or not) produces a value between -1 and 1 where being close to either end indicates a strong relationship in variables. Additional work using a variance inflation factor (VIF) produces value where any below 5 generally indicates lower multicollinearity, a value between 5 and 10 being high multicollinearity and a value greater than 10 indicating extremely high multicollinearity.

To approach data that violates the assumptions necessary for multiple linear regression, being the relationship between age and years of experience, a ridge regression model was used which applies what is known as a penalty term to shrink coefficients closer to zero reducing multicollinearity. This was accomplished by splitting the data where the majority was used to develop the model and then a small remaining portion tested its success. In both the training and testing of the model, an R-squared value was calculated which indicated how well the model was able to explain the variation in the data.

2.5 Analysis of Variance (ANOVA):

An analysis of variance (ANOVA) model was the initial underlying testing method proposed which compares a variable consisting of groups being used to predict some metric value, here being various categories predicting salary. Underlying assumptions of said test require equal

variance in the groups of categorical variables, that being the spread of the data be relatively equal. A visually supporting test was utilized by producing boxplots to visually see how the interquartile range and whiskers extend for each group compared to one and other.

Once conducted, the ANOVA model produces p-values indicating whether each subgroup serves to predict the dependent variable (salary) and a p-value assessing whether the means of the groups are equal. In these instances, a p-value less than our assessment level of 0.05 supports differences in groups and contributions towards predicted salary. Additional work is required for other assumptions made by ANOVA after the model is produced including ensuring that the residuals (distance from a given point and what the model predicts the value should be) are normally distributed (utilizing the prior mentioned Q-Q plot) along with homogeneity in variance (homoscedasticity) in the residuals as well; assessed by comparing the residuals to the model's predicted values. Violations of said assumptions require that non-parametric tests be used as stated in the following. Notably though, while breaking assumptions are less than optimal practice, certain testing like ANOVA is robust in that data is relatively accurate depending on the severity of said violations in assumptions.

2.6 Non-Parametric Alternatives:

For instances when the assumptions of ANOVA testing fail, alternative non-parametric applications (that being testing used for data that doesn't follow a typical distribution, such as when violations occur towards normality) are used. For the following analysis, a Kruskal-Wallis test was used to assess statistical differences in groups; like prior testing, a p-value less than 0.05 would provide strong evidence supporting differences between the central tendencies of each group. When considering categorical variables with only 2 groups (I.E. the gender variable and senior variable) conventional testing to compare differences between means would involve a two-

sample t-Test; this requires the data to be normally distributed though. In which case this assumption is not met, the non-parametric alternative is a Mann Whitney U test which concludes a difference in groups when the p-value is less than 0.05.

3. Results:

For the individuals to which information is provided in this data set, the proportion of education is presented below in Figure 1. From this information, the data set is predominantly individuals with a bachelor's degree or higher; only 7% is associated with a high school diploma. For a general representation of a population's educational standing, the Census Bureau (as of 2021) stated 37.9% of the population held a bachelor's or higher (Bureau, U. C., 2022). The data set presented is approximately 2.5 times greater than the stated Census Bureau report, although notably these individuals aren't exclusive to the Unites States in said data set.

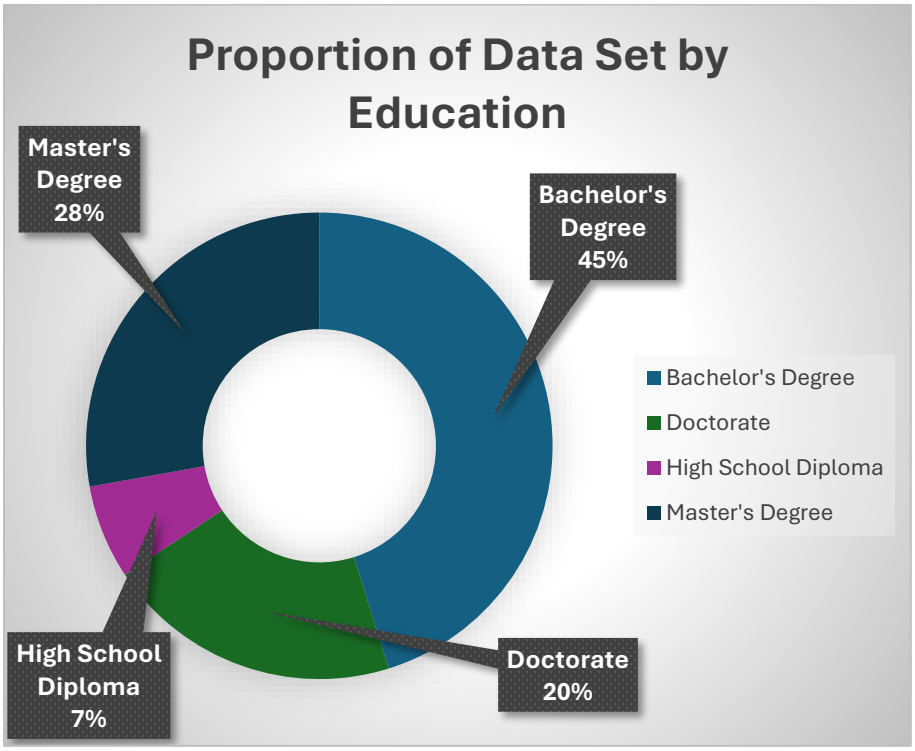

Figure 1: Proportion of the data set by education level.

From an analysis of the annual salary based on the position held, Table 1 demonstrates the top 10 of the 128 job titles presented for earnings. While an analysis on job position earnings can pose as a daunting challenge, general observations can be observed to understand potential characteristics that contribute to higher earning positions. For instance, from Table 1, all 10 of the positions hold the key words chief, director, or VP. This is indictive of upper management positions carrying greater earning potential.

Job Title	Average Annual Salary
CEO	$250,000.00
Chief Technology Officer	$250,000.00
Chief Data Officer	$220,000.00
Director of Data Science	$204,561.40
Director	$200,000.00
VP of Finance	$200,000.00
Operations Director	$190,000.00
VP of Operations	$190,000.00
Director of Human Resources	$187,500.00
Marketing Director	$183,615.38

Table 1: Average annual salary based on position held.

To assess the potential of multicollinearity, Anderson-Darling tests were applied to both the age and years of experience variables. These yielded p-values less than 2.2e-16 rejecting the null hypothesis concluding non normal data. Accordingly, the Spearman rank correlation was applied between said variables resulting in a correlation of 0.9460367 strongly supporting the presence of multicollinearity. This was also confirmed through a variance inflation factor (VIF) of 8.272077, which is likely to carry high inflation for the standard error and unreliable coefficients.

From the supporting evidence surrounding multicollinearity, a subset of the data frame was utilized through a matrix of salary, age, and years of experience. The subset of data was divided using a 70:30 split with 70% corresponding to a training set and 30% for a testing set. A ridge regression model was fitted to the training data set; optimized and then assessed through the R-squared value which was determined for the training data set to be 0.6464358. The optimized model was able to explain 64.64% of variation in the training data. From the test data, the corresponding R-squared value is 0.6497574 indicating the optimized model was able to explain 64.98% of variation in the test data. Accordingly, the model was moderately able to predict salary from age and years of experience through an approach that mitigated the influence of

multicollinearity, the model was highly consistent in R-squared values between training and testing data sets too.

Approaching the categorical variables through an initial assessment of variance, starting with the education variable, yields the boxplots in Figure 2 below.

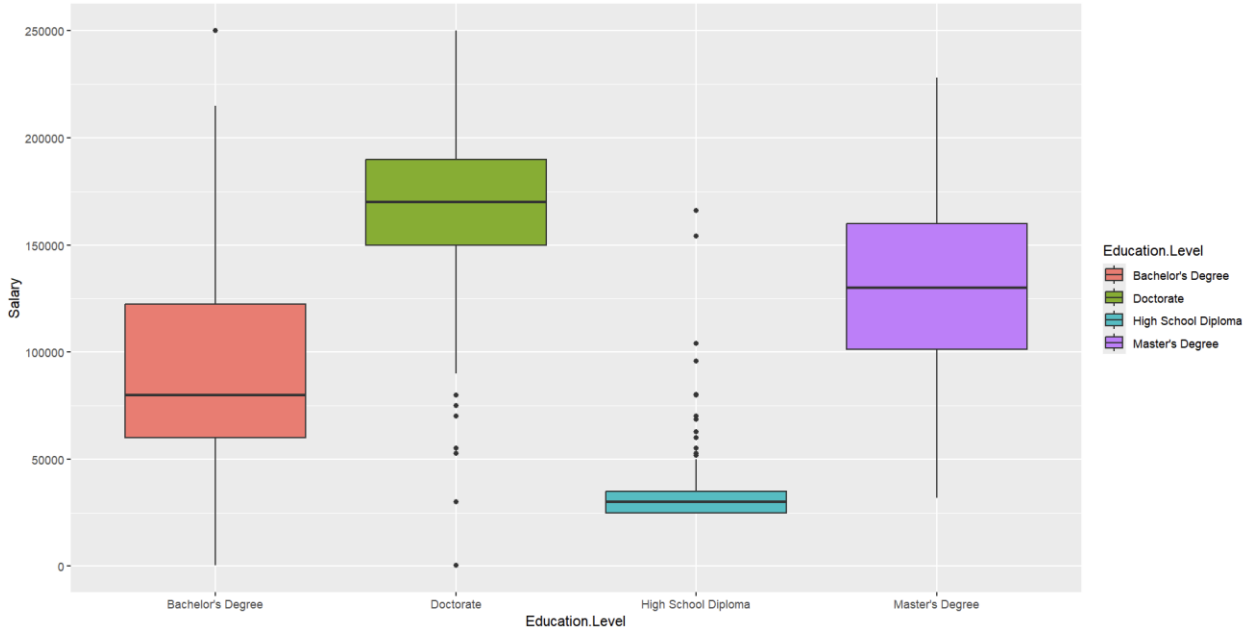

Figure 2: Boxplots of education level to salary, utilized to examine variance.

The unequal variance presented resulted in utilizing Kruskal-Wallis as a non-parametric approach to analyzing groups; the resulting p-value of said test being less than 2.2e-16 strongly rejects the null hypothesis at an alpha level of 0.05 concluding that education levels have statistically different central tendencies. For the variable “Race”, the boxplots displayed in Figure 3 below demonstrate approximately equal variance.

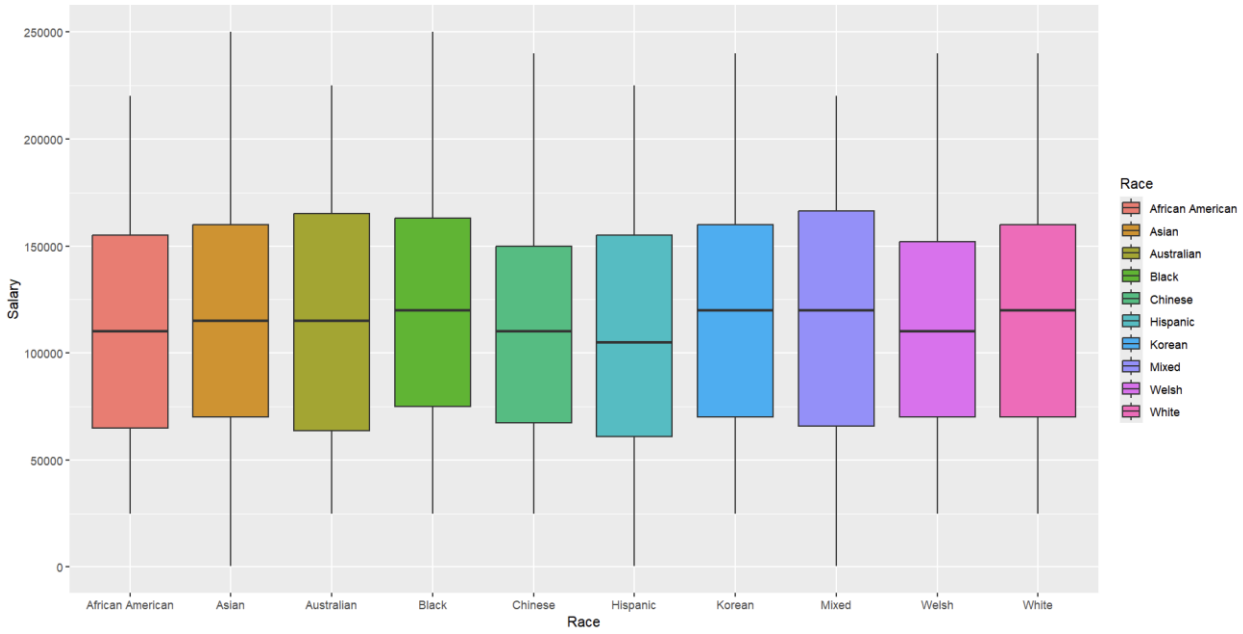

Figure 3: Boxplots for salary by race.

Accordingly, an ANOVA model was developed with salary dependent on race, the resulting coefficients are tabulated below in Table 2. From the resulting data, at an alpha level of 0.05, the corresponding p values per group are all greater than the alpha level accepting the null hypothesis where there is no statistical significance of the influence from any given race predicting salary. This is supported in the p-value of the global F-statistic for the model, returning a p-value of 0.2219 accepting the null hypothesis of no statistical difference in means between groups.

Variable	P-Value
(Intercept)	<2e-16
Race: Asian	0.352
Race: Australian	0.616
Race: Black	0.109
Race: Chinese	0.769
Race: Hispanic	0.633
Race: Korean	0.205
Race: Mixed	0.324
Race: Welsh	0.959
Race: White	0.125

Table 2: Resulting p-values for ANOVA test conducted on race.

To evaluate the validity of prior assumptions made on the ANOVA model, the residuals from the model had a corresponding Q-Q plot produced as shown in Figure 4 below. Overall, the plot appears non normal; this was supported with an Anderson-Darling test yielding a p-value less than 2.2e-16 confirming a lack of normality in the residuals. While the ANOVA test is robust in approaching deviations from normality particularly considering a lack of violation from the other assumptions, a Kruskal-Wallis test was conducted as a supporting non-parametric alternative. The result of which produced a p-value of 0.2523 aligning closely with the global F-statistic of the ANOVA model and further supporting a lack of statistically significant differences in central tendencies between salary earnings by race.

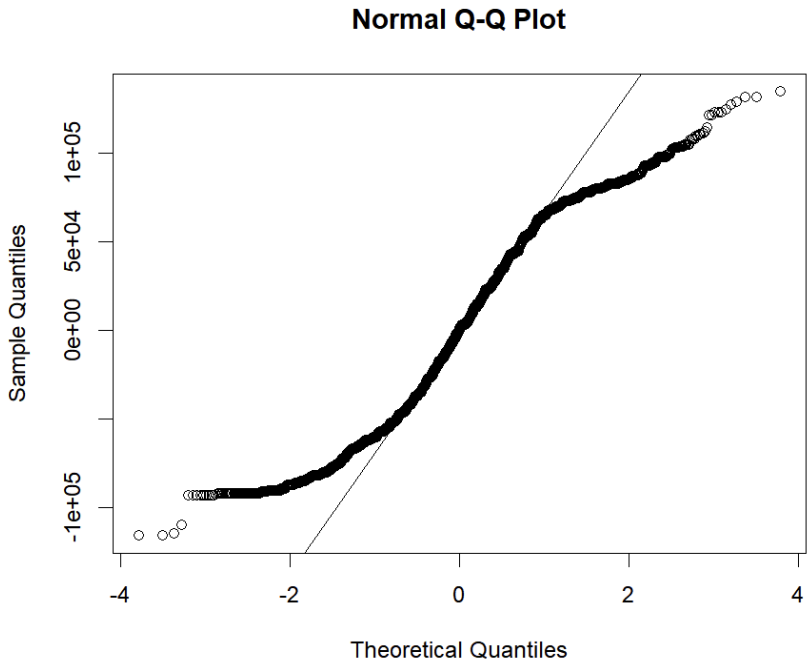

Figure 4: Q-Q plot of residuals for ANOVA model utilizing salary to race.

The resulting analysis shown in Figure 5 demonstrates the relationship between salaries by country. Apart from slight variations in the whiskers of the boxplots, the variances appear approximately equal. Accordingly, an ANOVA model was developed yielding the coefficients presented in Table 3 below.

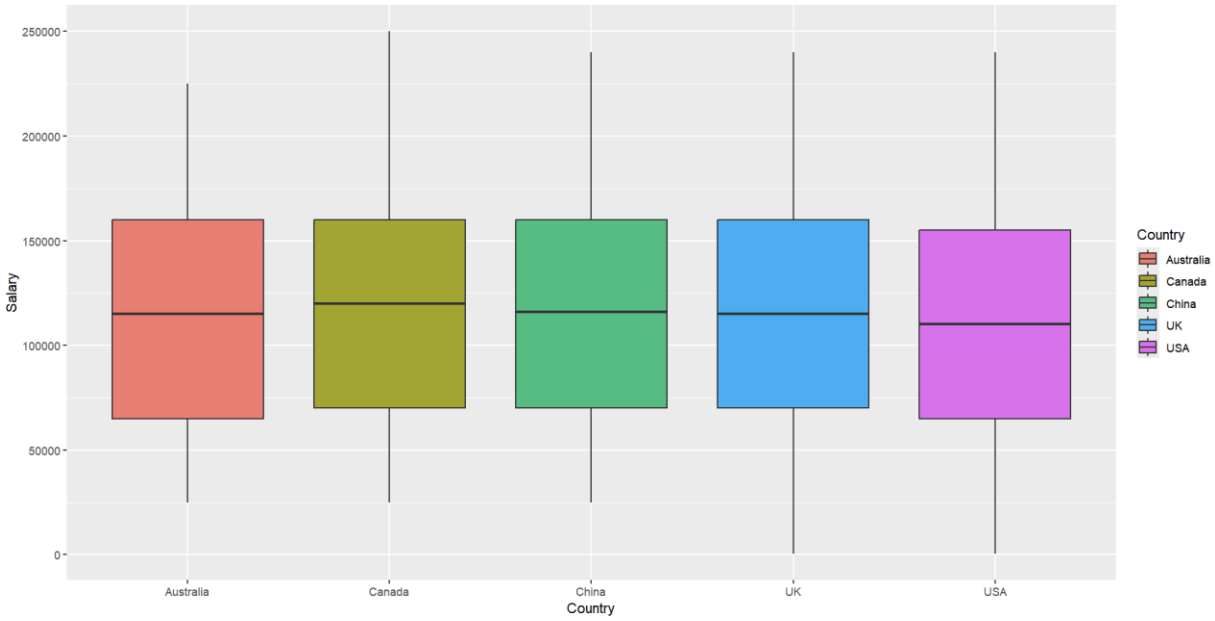

Figure 5: Boxplots demonstrating salary by country.

From the results of the ANOVA model, the p-values are all above the alpha level of 0.05 accepting the null hypothesis concluding no statistical significance in any given country predicting salary. The global F-statistic observed is 0.4168; also accepting the null hypothesis of equal means between groups, I.E. no statistically significant difference between countries.

Variable	P-Value
(Intercept)	<2e-16
Country: Canada	0.455
Country: China	0.506
Country: UK	0.627
Country: USA	0.344

Table 3: The resulting coefficients for the country ANOVA model.

From the residuals produced from the ANOVA model, the Q-Q plot displayed in Figure 6 demonstrates a lack of normality that is supported with an Anderson-Darling test p-value of less than 2.2e-16 rejecting the null hypothesis. From these findings, a Kruskal-Wallis test was conducted yielding a p-value of 0.4641 failing to reject the null hypothesis in which there is no

statistically significant difference between the central tendencies between countries closely aligning with the results of the ANOVA test despite violations of the assumption towards normality.

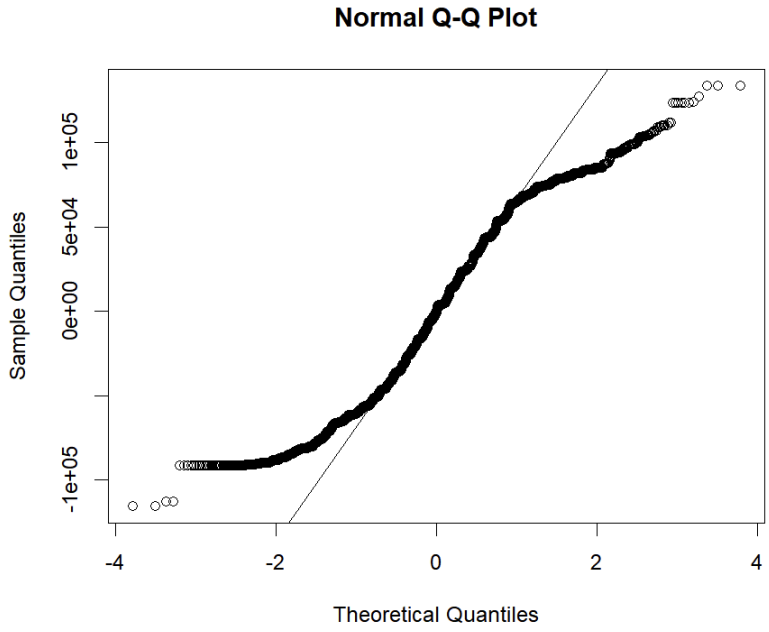

Figure 6: Q-Q plot of residuals for ANOVA model predicting salary by country.

For the dichotomous variables gender and senior, histograms were produced to observe any trend in normality with respect to salary per a given category; the results of which are demonstrated below in Figure 7 and Figure 8. For gender, the resulting histograms are non-normally distributed. Due to this, a Mann Whitney U test was conducted as a non-parametric approach to a two-sample t test; the result of which yielded a p-value of less than 2.2e-16 rejecting the null hypothesis concluding a statistically significant shift in distribution between males and females.

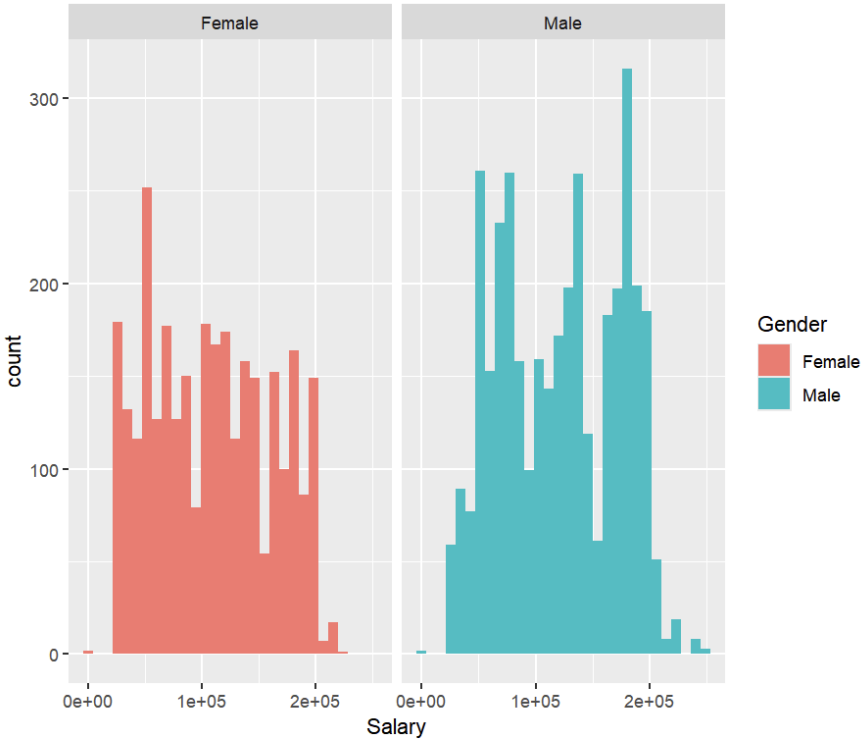

Figure 7: Histograms for male and female with respect to salary.

With respect to the variable corresponding to whether an individual holds a senior position or not, the resulting histograms are shown in Figure 11 below. Like the findings of the gender histograms, a lack of normality is presented leading to conducting a Mann Whitney U test, the result of which is a p-value less than 2.2e-16 rejecting the null hypothesis concluding a difference in distribution between an individual who holds a senior and one who doesn't, with respect to salary earnings.

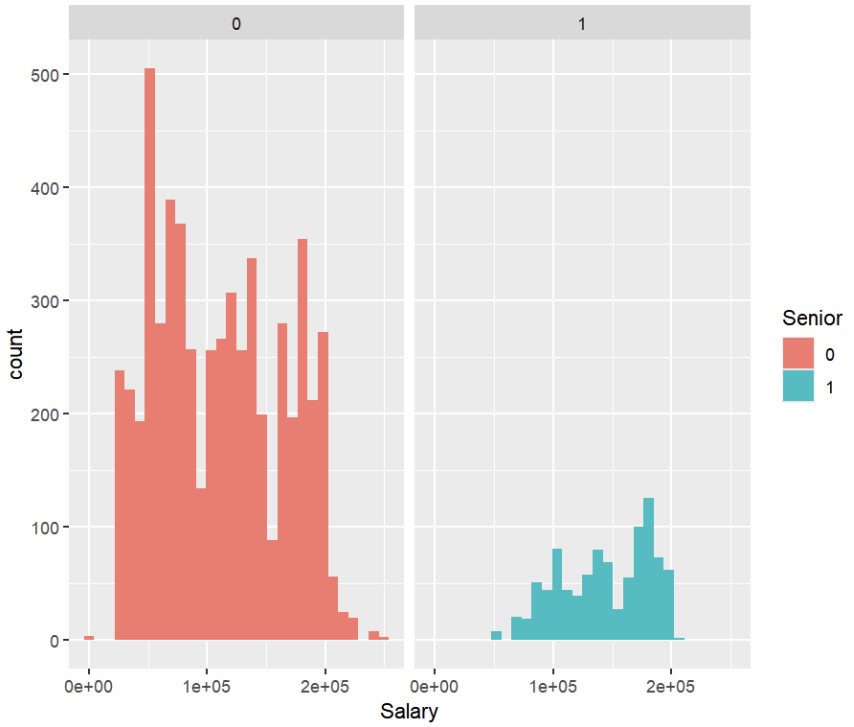

Figure 8: Histograms of salary by senior position, 1 indicating the individual has a senior position and 0 not being a senior position.

4. Conclusions, Recommendations and Future Work:

From the findings of the general summary statistics, the job categories associated with the highest earning salary potential all contained key words in relation to higher end corporate positions. A deeper analysis of salary in relation to the continuous variables age and years of experience violated assumptions of multiple linear regression through the presence of multicollinearity. This was supported with a high Spearman rank correlation and a variance inflation factor. Through such observations, isolating the effects of any given variable on the predicted variable can be unreliable and this also likely is associated with increasing error. To mitigate this phenomenon, a ridge regression model was trained, optimized, and tested with the remaining data set. The resulting R-squared values from both test and train sets provided a moderate fit of 0.6464358 and 0.6497574, respectively. While there was consistency between the

sets, future work utilizing the works of a lasso regression model, or optimizing an elastic regression model, could increase the capability of the trained model to more accurately predict salary as a metric dependent on both age and years of experience.

Analyzing the relation between various categorical variables with respect to predicting salary demonstrated that education level had unequal variance between groups and a Kruskal-Wallis test to support a statistically significant difference in central tendency between education levels. A post-hoc analysis could utilize Dunn's test to more aptly investigate differences in groups. For both race and country, boxplots visually displayed approximately equal variances followed by ANOVA models demonstrating no statistical contribution; neither race nor country predicted salary from their respective coefficients along with no statistically significant difference in the means of race and countries via the global F-statistic. While assessing the assumption of normality, both models failed Anderson-Darling tests and visually didn't fit well for Q-Q plots. From these findings, Kruskal-Wallis tests were utilized although these findings aligned with the ANOVA models' predictions towards no statistical difference between groups.

For the dichotomous variables gender and senior, the resulting histograms demonstrated non normally distributed data. Accordingly, Mann Whitney U tests were applied to assess differences in groups. For both gender and senior, there was a statistically significant difference in salary earnings. A two-sample t-test was applied for each, for experimental purposes, yielding the same results.

Overall, a lack of normality was present throughout a multitude of variables within the data set. While certain tests are robust in producing accurate results despite violations, stronger practice is observed in ensuring assumptions are held. Accordingly, limitations occurred likely in sampling practices not mentioned from the data set. A stronger approach, and a greater representation and investigation for future work, would be to utilize the principles of the Central Limit Theorem (CLT) averaging a multitude of samples and modeling the distribution of means to produce normally distributed data that could be applied to parametric testing methodologies such as ANOVA and t-tests. With respect to the scope of the analysis, the ridge regression model moderately serves to predict salary depending on an individual's age and years of service in a position. Future work could serve to refine the model with additional training on more data or

developing alternative models for mitigating multicollinearity. When isolating categorical variables, both race and country held no significance towards predicting salary. For education level, gender, and senior positions, there was evidence to support this influencing salary. Based on this analysis, and the scope of the data, gender will be associated with an intrinsic variable that cannot be pursued to develop one's salary. Alternatively, a senior position is a tangible objective an individual can pursue towards gaining a higher salary. Future work would likely benefit from a further investigation between years of service and a senior position as well along with an investigation into how variables may interact with one and other. Finally, education level provided strong statistical evidence to support higher education correlating to greater earning potential. From an application perspective, this aligned with a primary objective of the research and demonstrated that the variable "education level" influenced annual salary and serves as a potential goal one might consider when looking for economic growth. To the extent of addressing variables with marginal effects on salary, all statistical testing (within the scope of the data set) concluded somewhat equal statistical contribution for all categories deemed significant concluding that country and race were not an influence factor, but education, senior position, and gender were influential factors. Expanding on these variables through future work would aid in better understanding more complex intricacies surrounding their influence and serve to reinforce any relations.

References

Aboutalebi, A. (2024, February 18). *Salary by job title and country*. Kaggle. https://www.kaggle.com/datasets/amirmahdiabbootalebi/salary-by-job-title-and-country

U.S. Bureau of Labor Statistics. (n.d.). *Same occupation, different pay: How wages vary*. U.S. Bureau of Labor Statistics. https://www.bls.gov/careeroutlook/2015/article/wage-differences.htm

Bureau, U. C. (2022, February 24). *Census Bureau releases New Educational Attainment Data*. Census.gov. https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html