Predictive Analysis on Data Science Salaries 2020-2023

Mai Ngo

DePaul University - College of Computing and Digital Media (CDM)

DSC 423: Data Analysis and Regression

Nandhini Gulasingam

June 5, 2023

Tables of Contents

Abstract 1
Literature Review 2
Introduction **3**
1. Dependent and Independent Variables 3
2. Objective and Method 4
Data Exploratory 5
1. Dependent Variable 5
Applied Transformation 5
2. Categorical Variables 5
3. Continuous Variable 6
Final Dataset 7
Regression Results 8
1. First Model 8
2. Second Model 8
3. Stepwise Selection Method 9
4. Third Model 9
5. Final Model 9
Studentized Residuals 10
Global F-Test 11
Prediction 13
Validation 14
Conclusion 15
References 16
Appendix A. Data 17
Appendix B. Regression Models 23
Appendix C. Prediction 30
Appendix D. Validation 31

Abstract

Data science in general is the studies of data by using a variety of methods to make better decision. There are two main branches; first, utilizing and analyzing data to make and predict better decisions (business management related purposes). Second branch is more on machine learning (algorithm improvement for data processing). Even though data can be spoken or analyzed by algorithm, it is still essential to have human intuition involved to ensure the data is interpreted correctly. Over the past decade, Data Science has emerged as a prominent field within the technology industry, attracting workers both outside and inside technology industry seeking better career path and higher pay rates. Since the industry is still new on a global scale. However, it is dominant and has such huge direct and indirect impact on every economic aspect in the United States. In this project, we will explore the growth of Data Science salary by using the dataset captures salary worldwide from 2020-2023. The objective is to learn how Data Science field salaries has been growth throughout the year, and whether there is an equal career growth for all data scientist globally in terms of salary improvements.

Literature Review

Data science salaries are highly volatile depend on job factors such as job skills, software expertise, personal background, and the data science sector. In situations where there are two jobs have identical description, require the same skillset and experience; data scientist may earn higher salary with the job utilizes more trending and on-demand software. Research on Quantitative Data Science Salary Analysis to Predict Job Salaries discussed that Amazon Web Services (AWS) cloud computing services has 40% of higher demand and pay rate than other computing service system which attracts much more attention from highly skill data scientists (Kaur, 2022). The study also mentioned that women enjoy working in data science field and have equal opportunity to growth 40% higher than other industry. However, despite the equitable growth opportunity, women are still getting pay less than men which still aligned with the gender pay gap social norm.

Another study from Cao (2017) breaks down different components of what criteria makes a good data Scientist and corresponding career growth. The study emphasizes that in 2014, data scientists have higher compensations compared to other data-related jobs, with a median salary of $120,000, and $160,000 for managers. Since the job expertise commonly a master's degree for skill verification, approximately a quarter of data scientist in the study were found to hold a doctorate, while another 44% had a Master's. The study also looks at compensation on globally and in the US, annual base salary was $91,000 globally, and $104,000 for the US, a promising global growth of the data science field.

A study on Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits conducted by Quan & Raheem in 2022, testing different predictive and classification models like linear and polynomial regressions, decision tree, K Nearest Neighbors classifier, neural network, and stochastic gradient descent, etc. The author concludes that for general cases where the dataset is highly dimensional, with effective parameter tuning using grid search, majority machine learning models produce more accurate predictions by handling both linear and non-linear relationships, especially with tree-based models. Deep learning neural network techniques shows its expertise in processing contextual data directly from job postings, allowing efficient data mining on a larger scale without labelling and structuring raw data.

Introduction

The dataset contains 3755 observations of Data Science field salaries worldwide from 2020 to 2023. The dataset is from Kaggle (.csv format), and original data source is from https://salaries.ai-jobs.net/download/. There are a total of 11 attributes: 4 numerical and 7 categorical attributes (See Appendix A - Data Introduction, Table 1 and Table 2). Variable and data type are shown below:

- work_year: The year that salary was paid (dtype = num).
- experience_level: Level of experience (dtype = char).
    - EN - Entry level, MI - Mid level, SE - Senior level, EX - Executive level.
- employment_type: Type of employment (dtype=char).
    - PT - Part time, FT - Full time, CT - Contract, FL – Freelance.
- job_title: Title of the job (dtype = char).
- salary: Total gross salary in 'salary_currency' (dtype = num).
- salary_currency: Currency of paid salary (dtype = char).
- salary_in_usd: Salary in USD (dtype = num).
- employee_residence: Employee's country of residence (dtype = char).
- remote_ratio: Ratio of remote work (dtype = num).
    - 0 - No remote work (< 20%), 50 – Hybrid, 100 - Fully remote (> 80%).
- company_location: Employer's country of headquarter / contracting branch (dtype = char).
- company_size: Number of employees in term of size (dtype = char).
    - S - < 50 employees, M - 50 to 250 employees, L - > 250 employees.

1. Dependent Variable

salary_in_usd

Independent Variable

work_year, experience_level, employment_type, job_title, salary, salary_currency, employee_residence, remote_ratio, company_location, company_size.

2. Objective and Method

The objective of this project is to create a regression model to predict future Data Science salaries in USD currency (continuous variable) given current pay rate (original currency) and employment condition (remote ratio, company sizes, etc.). Enhance Adjusted-R2 from the full regression model yet reduce number of predictors reasonably. All attributes from the original dataset will be included in the model to start with. Given this, I will apply Multiple Linear Regression.

Furthermore, I was also aware that there is no linear relationship between Salary in original currency and Salary in USD. However, without including the only continuous variable, the model will become an Analysis of Variance (ANOVA) which is not the objective of this course. I have also attempted to transform the predictor Salary in the original currency which the outcome was not great thus, I have decided to keep the predictor as it is.

Data Exploratory

1. Dependent Variable

The dependent variable Salary in USD has a unimodal distribution and is right skewed (See Appendix A - Data Exploratory, Figure 1). The mean is $137,570.4 greater than median at $135,000. It can be interpreted the average income of Data Science field is around $137,570.4 (See Appendix A - Data Exploratory, Table 3). From the histogram, we can detect that there are outliers on the higher end. The distribution range is from $5,132 to $450,000; with standard deviation (S/D) equals to $63,056. Two S/D to the left gives Salary in USD value at $11,458.4 (closer to the minimum value), and three S/D to the right gives Salary in USD value at $326,738.4. We can conclude that any value greater than $326,738.4 threshold will be considered as outliers. Looking at the Quantiles table, 99 percentile gives value at $309,400 and 100 percentile gives the maximum value. Thus, there are outliers at upper range, and we need to apply transformation to address the skewness and presence of outliers.

Applied Transformation

Started with apply log transformation, which is a commonly preferred method, the distribution of dependent variable was still right skewed. Then I applied square root transformation, and the variable now has a normal distribution (See Appendix A - Data Exploratory, Figure 2). Therefore, we will move forward using square root of Salary in USD as predicted value for the regression model.

2. Categorical Variables

To visualize the distribution of each categorical attribute, I use bar chart and frequency table. All frequency count mentioned below are retrieved from frequency table of corresponding attribute. Working Year attribute represents the year salary was paid. Year 2022 and 2023 have much significant higher number of observations in the dataset: 1664 and 1785 respectively. The distribution of Salary (USD) versus each Working Year shows that salary of Data Science field has been increasing over time with salary range pretty much stay consistent. The salary median of 2020 and 2021 are similar around $75,000, while in 2022 and 2023, salary median shot up over $100,000 (See Appendix A - Data Exploratory, Figure 3).

Experience Level has four levels: entry, mid-level, senior and executive. Senior level accounts for the most with 2516 observations, around 67% of the dataset. The distribution of Salary (USD) versus Experience Level shows executive level has the highest median salary and range, followed by senior level. Entry and mid-level are pretty much having the same distribution (See Appendix A - Data Exploratory, Figure 4). Overall, this distribution makes sense, the higher working position a person has, the better they get paid. Employee Residence represents the country the employee lives during salary paid year. Company Location is the head quarter location or country of the employer. The two attributes have almost identical distribution where US accounts for majority of the observation (around 80% of the dataset): 3004 for Employee Residence, and 3040 for Company Location, respectively (See Appendix A - Data Exploratory , Figure 5). We could assume multicollinearity.

Salary Currency is the currency of Salary (original currency). Again, USD is dominated with 3224 observations (See Appendix A - Data Exploratory, Figure 6). Remote Ratio has three categories: no remote work, hybrid and fully remote. No remote work has the highest count of more than 3000 observations. Boxplot represents the relationship between Remote Ratio and Salary (USD) shows that fully remote and no remote at all have similar distribution with median Salary in USD of around $130,000. While hybrid working model has lower median pay of approximately $80,000 (See Appendix A - Data Exploratory, Figure 7). We can assume that the salary indifference here could be caused by data imbalance. Or else, this relationship is worth to explore further since we know generally tech employee prefers hybrid working model.

The two final categorical attributes are Company Size and Employment Type. Company Size attribute has three categories: small medium and large. Medium size has the most data points of 3153 counts. Employment Type is separated into 4 types: contract, freelance, full-time, and part-time. Full-time employment dominates the distribution of Employment Type with 3718 observations, almost 99% of the dataset (See Appendix A - Data Exploratory, Figure 8).

3. Continuous Variable

Below graph represents the relationship between the predicted variable - square root of Salary (USD) versus the only continuous predictors in the dataset, Salary (original currency).

The graph shows a distinct straight flat line between the two variables, which emphasizes changes from square root of Salary (USD) does not correlate with changes of Salary (original currency) (See Appendix A - Data Exploratory, Figure 9). Or in another word, there is no consistent increase or decrease in predicted variable as the predictor variable changes. The relationship between the two variables is very weak. Per mentioning in the Objective section, Salary (original currency) will be included in the model.

Final Dataset

Given all analysis and understanding from attribute exploratory phase, I apply dummy variable transformation toward all categorical attributes, except for Working Year. In addition, since there is data imbalance issue, I set the reference attribute “1” to represent the feature with the highest count frequency, and “0” will be the rest. An example, Employment Type is completely dominated with full-time position thus, full-time will be set as "1", and the rest features will be set as "0". The final dataset will have all attribute, predicted value is square root of Salary (USD), predictors are dummy variables of the eight attributes, Salary (original currency) and Working Year stay the same (See Appendix A - Final Dataset, Table 4)/

The second column of correlation table describes the correlation between each predictor versus predicted value (See Appendix A - Final Dataset, Table 5 and Table 6). Majority of the predictors has weak to moderate correlations with the predicted variable. Noticeably, Employee Residence has a very strong correlation with Company Location at 0.96146, this suggests multicollinearity which we will verify through VIF later. This strong correlation does make sense, employee tend to live near where they work. Furthermore, scatterplot matrix will not be generated since we have dummy variables as predictors. Correlation with dummy variable does not interpret the association entirely correctly, not reliable, due to fixed values at 0 and 1.

Regression Results

1. First Model

Started with running regression model on the full dataset with all attributes. Based on the regression result - Parameter Estimates (See Appendix B - First Model Regression Result, Table 1), Work Year, Employee Residence, Experience Level, Type of Employment, and Salary Currency variables are highly significant due to their p-values < 0.01. The p-values for Company Location = 0.5915, Company Size = 0.9649 and Remote Ratio = 0.3217, greater than 0.05. Thus, the three variables are not significant. Looking at standardized coefficients, Employee Residence is the most influence at +0.309, followed by Employee Experience at -0.267. Company size is the least significant variable at -0.00065.

R-square = 0.3934 indicates that 39.34% of the variance in Salary (USD) amount can be explained by all listed predictors listed here. Looking at (Variance Inflation Factor (VIF) value for multicollinearity detection, both Employee Residence and Company Location have VIF values greater than 10 and Tolerance values less than 0.10 which proves multicollinearity. Employee Residence has VIF = 14.752 and Tolerance = 0.0678, Company Location has VIF = 14.78 and Tolerance = 0.0677. Since Company Location has higher VIF value than Employee Residence's, I will re-fit the model again excluding Company Location attribute.

The variance in the studentized residuals versus predicted value seems to increase in density and has a funnel shape (See Appendix B - First Model Regression Result, Figure 1). This suggests that we may have heteroscedasticity, which means the variability of residuals is not constant and maybe influenced by the predicted value. Therefore, the assumption of Constant Variance is violated. Normal Probability Plot (NPP) has a pattern of spread shows an almost 45-degree straight line, which means the model maintains normal probability. The assumption of Normality is satisfied.

2. Second Model

Second regression model excludes Company Location attribute due to multicollinearity with Employee Residence. R-square slight changes from 39.34% to 39.35%. Root Mean Square Error (RMSE) slightly decreases from 69.94 to 69.93. F-value increases from 271.47 to 305.42

with significant P-value <.0001 (See Appendix B - Second Model Regression Result, Table 2). Looking at p-value to determine variable significance, Company Size and Remote Ratio still have high p-value greater than 0.05: 0.9754 and 0.3191, respectively. Hence, for next step, I will apply Stepwise selection method because it focuses on examine the significances of predictor variables.

3. Stepwise Selection Method

With Stepwise selection method, the final suggested model reduced to six attributes: Employee Residence, Experience Level, Salary Currency, Work Year, Employment Type, and Salary (original currency). The suggested model has R-squared = 0.3946 or 39.46% (See Appendix B - Stepwise Selection Method, Table 3). An improvement from previous model which has R-square = 39.35%. Since the suggested model has higher R-square and reduced number of predictors, we will move forward with the suggested model from Stepwise selection method.

4. Third Model

Third regression model has the six chosen attributes, the model also results output table for outliers and influential points detection (See Appendix B - Third Model Regression Result, Table 4). Since there are 3755 observations, I export the Output Statistics table to Excel to extract outliers and influential points.

After exporting to Excel (See Appendix B - Third Model Regression Result, Figure 2), I extract Studentized Residual (SR), Cook's D, and observation ID columns to a different sheet. Then create two columns to detect whether an observation is an outlier or influential point using filter formula. For studentized residuals, any observation with absolute value of SR greater than or equal to 3 is considered as an outlier. For Cook's D value, any value greater than 0.002 is considered as an influential point. Then I conditioned these two columns to a third column to obtain observations that are either or both influential points and outliers. In total, I have removed 93 observations, about 2.5% of the total number of the dataset.

5. Final Model

Model fourth regression input using a new dataset with all outliers and influential points removed. The new dataset contains 3662 observations. R-squared is now improved to 0.4184 indicates that 41.84% of the variance in the Salary (USD) can be explained by the six predictors: Work Year, Salary (original currency), Employee Residence, Experience Level, Employment Type, and Salary Currency (See Appendix B - Final Model Regression Result, Table 5). All predictors have p-value less than 0.05 with Employee Residence has the strongest influence on Salary (USD) at positive 0.28, followed by experience level at negative 0.263. All VIF value are less than 10. F-value increased from 305.42 to 440.04 with significant p-value <.0001. RMSE also reduced from 69.93 to 64.733. Overall, we have improved the model.

For coefficient interpretation, using Employee Residence Status with coefficient = 61.18466, since we set dummy variable as 1 for employee lives in the, and 0 otherwise, we can interpret as holding other variables constant, if an employee lives in the US (dummy variable = 1), the predicted Salary_in_USD is expected to increase by approximately $3,743.604 USD = $(61.18466)^2$.

Since all predictors are significant, p-value less than 0.05 with no sign of multicollinearity. Number of predictors reduced with improved R-squared improved, reduced RMSE. We can conclude that this will be the final model.

Studentized Residuals

Looking at studentized residuals plots, for all categorical attributes that have been converted to dummy attributes, even though there is no linearity, non-Constance Variance due to distinctive pattern; since these are dummy variables distribution, we cannot draw conclusion for anomaly detection.

For studentized residuals plot versus Salary (original currency) (See Appendix B - Final Model Regression Result, Figure 3), there is a distinctive pattern where when residual increases, salary in original currency stay the same at around zero mark, and scatter points under zero line. The assumption of Linear association, Constance Variance, and Independence of the three attributes with the response variable (Bank Balance) is not satisfied. There is little or no change

in the salary attribute as the residuals vary, indicating a lack of response of the salary to the variability of the residuals.

The variance in the studentized residual versus the predicted value has a funnel shape. This suggests that we have heteroscedasticity, which means the variability of residuals is not constant and maybe influenced by the predicted value. Therefore, the assumption of Constant Variance is violated. Normal Probability Plot has a pattern of spread shows an almost 45-degree straight line, which means the model maintains normal probability. The assumption of Normality is satisfied.

Global F-Test

The F-test (goodness-of-Fit test) is used to check if the regression model hypothesis is right or wrong based on its prediction correctness. A good model will have high F-value with significant P-value (< 0.05). Since there is no upper limit for F-value, we will evaluate both values together.

Regression model

$$\textbf{sqrtSalary\_in\_USD} = \beta_0 + \beta_1 * \textbf{work\_year} + \beta_2 * \textbf{salary (original currency)} + \beta_3 * \textbf{numEmployee\_residence} + \beta_4 * \textbf{experience\_New} + \beta_5 * \textbf{employment\_FT} + \beta_6 * \textbf{salary\_currencyUSD}$$

for: β_0 = Intercept.

β_1 = Coefficient for Work Year attribute.

β_2 = Coefficient for Salary (original currency) attribute.

B_3 = Coefficient for Employee Residence attribute.

B_4 = Coefficient for Experience Level attribute.

B_5 = Coefficient for Employment Type attribute.

B_6 = Coefficient for Salary Currency attribute.

Test hypothesis

Null hypothesis: Neither of the six attributes have any association with Salary in USD attribute.

$$\mathbf{H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0}$$

Alternative hypothesis: At least one attribute has a significant effect on changes in Salary in USD attribute.

$$\mathbf{H_a: \text{At least one of } \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6 \neq 0}$$

Based on the regression result, F-value is calculated using formula MS (Model) /MS (Error) = 1843958 / 4190.43803 = 440.04 with highly significant p-value <0.0001 indicates the model prediction is accurate. At least one of the attributes is relevant and significant in explaining Salary (USD) attribute. With $\beta_1 = 9.07$, $\beta 2 = 0.00001202$, $\beta 3 = 61.184$, $\beta 4 = -49.153$, $\beta 5 = 66.58$, and $\beta 6 = 51.814$, not equal to zero. Thus, the Null hypothesis (H_0) of no association between the six predictors: Work Year, Salary (original currency), Employee Residence, Experience Level, Employment Type, Salary Currency and predicted variable Salary (USD) is rejected and accepting Alternate hypothesis (H_a).

Prediction

For model prediction, I create two observations which represents two scenarios: The first prediction is for a person who lives in the US and gets paid by USD. The second prediction is for a person who lives outside the US and gets paid by their local currency. We created and merged the two datasets (See Appendix C - Prediction, Table 1). The objective is to see whether a person who lives outside the US has equal career advance opportunity in term of salary raise like a person who live inside the US. The salary prediction will be for the upcoming year 2024.

First observation: with given values, the predicted sqrtSalary_in_USD is 358.5587, approximately $128,564.34. The 95% confidence interval range from 351.5912 to 365.5261; or $123,616.37 to $133,609.33, applied square from original value. Prediction intervals range from 231.45 to 485.6673; or $53,569.1 – $235,872.72. Therefore, we can conclude that the predicted value falls within the 95% confidence interval (See Appendix C - Prediction, Table 2).

Second observation: with given values, the predicted sqrtSalary_in_USD is 245.5592, approximately $60,299.32. The 95% confidence interval range from 236.4812 to 254.6317; or $55,923.36 to $64,837.3 applied square from original value. Prediction intervals range from 118.8174 to 372.8009; or $14,117.57 – $138,980.511. Therefore, we can conclude that the predicted value falls within the 95% confidence interval (See Appendix C - Prediction, Table 2).

Overall, we can see that Data Science career growth in terms of salary for a person who lives outside the US is not as great as for a person who lives in the US.

Validation

For model validation, I split the data into train and test sets using 80-20 ratio. Looking at the first column “Selected”, "1" indicating observations belong to train data, and "0" for test data (See Appendix D – Split Dataset, Table 1). After the split, we have 2930 observations on train set. Create new predicted value ‘new_y’ equals to the predicted value for train data, and null values for test data (See Appendix D – Split Dataset, Table 2). Furthermore, since applying Multiple Linear Regression, we will not be able to generate confusion matrix since it is only applicable for classification. Therefore, we will check model performance based on RMSE and R-squared.

I also want to confirm that my final model with six attributes is satisfied with just using the train data. Thus, I apply Stepwise selection method toward the training data. The suggested model is similar with the final model, indicating consistency and stability in the variable selection process (See Appendix D – Stepwise Selection Method, Table 3).

Regression on the train data gives adjusted R-square at 41.95%, RMSE = 64.76059, F-value = 353.74 with significant P-value <.0001 (See Appendix D - Regression Results, Table 4). All predictors are significant with P-value < 0.05. Also generate predicted value for test set with null ‘new_y’ values (See Appendix D - Regression Results, Table 5).

Regression on the test data gives adjusted R-square at 41.38%, RMSE = 64.646574, MAE = 52.1490 (See Appendix D - Regression Results, Table 6). Specifically, test set R-squared is calculated using $(yhat^2) = 0.64333^2 = 0.4138$. Overall, train and test data performance have close RMSE: 64.76 for training and 64.64 for testing. R-squared on both train and test set are close to each other: 41.95% on training and 41.38% on testing. Since the indifferences are not significant, I would conclude this is not a case of overfitting or underfitting, and the model is performing well on both train and unseen data.

Conclusion

In conclusion, apply multiple linear regression on the dataset, majority of the data corresponds to US employees in terms of salary and company. The data indicates that US employees tend to receive higher pay and experience higher career growth compared to employees outside of the US. This observation aligns with the fact that the United States has a rapid growth rate in STEM, machine learning, and the tech industry. The US is known for providing growth opportunities and better salaries, which attracts tech workers from around the world. Indeed, this is not the best model for the dataset. But the best that I can generate using all knowledge and techniques that I obtain from the course. There are more room for improvement and potential expansion of the model. All predictors of the dataset were only able to explain around 40% of predicted value thus, we need more attributes to improve the model like, race, gender, number of years education, years of experience. Or incorporate other data like inflation rate or conversion rate. The project and process have provided a good insight into Data Science field salary and its growth.

References

Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.

Kaur, A., Verma, D., & Kaur, N. (2022, December). Utilizing Quantitative Data Science Salary Analysis to Predict Job Salaries. In *2022 2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)* (pp. 1-4). IEEE.

Quan, T. Z., & Raheem, M. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature. *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304)*, *6*(3), 70.

Appendix A. Data

Appendix A – Dataset Introduction

Salary Data 10 observation

Obs	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location
1	2023	SE	FT	Principal Data Scientist	80000	EUR	85847	ES	100	ES
2	2023	MI	CT	ML Engineer	30000	USD	30000	US	100	US
3	2023	MI	CT	ML Engineer	25500	USD	25500	US	100	US
4	2023	SE	FT	Data Scientist	175000	USD	175000	CA	100	CA
5	2023	SE	FT	Data Scientist	120000	USD	120000	CA	100	CA
6	2023	SE	FT	Applied Scientist	222200	USD	222200	US	0	US
7	2023	SE	FT	Applied Scientist	136000	USD	136000	US	0	US
8	2023	SE	FT	Data Scientist	219000	USD	219000	CA	0	CA
9	2023	SE	FT	Data Scientist	141000	USD	141000	CA	0	CA
10	2023	SE	FT	Data Scientist	147100	USD	147100	US	0	US

Table 1: The first 10 observations of the original dataset.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
10	company_location	Char	2	$2.	$2.
11	company_size	Char	1	$1.	$1.
8	employee_residence	Char	2	$2.	$2.
3	employment_type	Char	2	$2.	$2.
2	experience_level	Char	2	$2.	$2.
4	job_title	Char	24	$24.	$24.
9	remote_ratio	Num	8	BEST12.	BEST32.
5	salary	Num	8	BEST12.	BEST32.
6	salary_currency	Char	3	$3.	$3.
7	salary_in_usd	Num	8	BEST12.	BEST32.
1	work_year	Num	8	BEST12.	BEST32.

Table 2: Attribute names and data types.

Appendix A – Data Exploratory

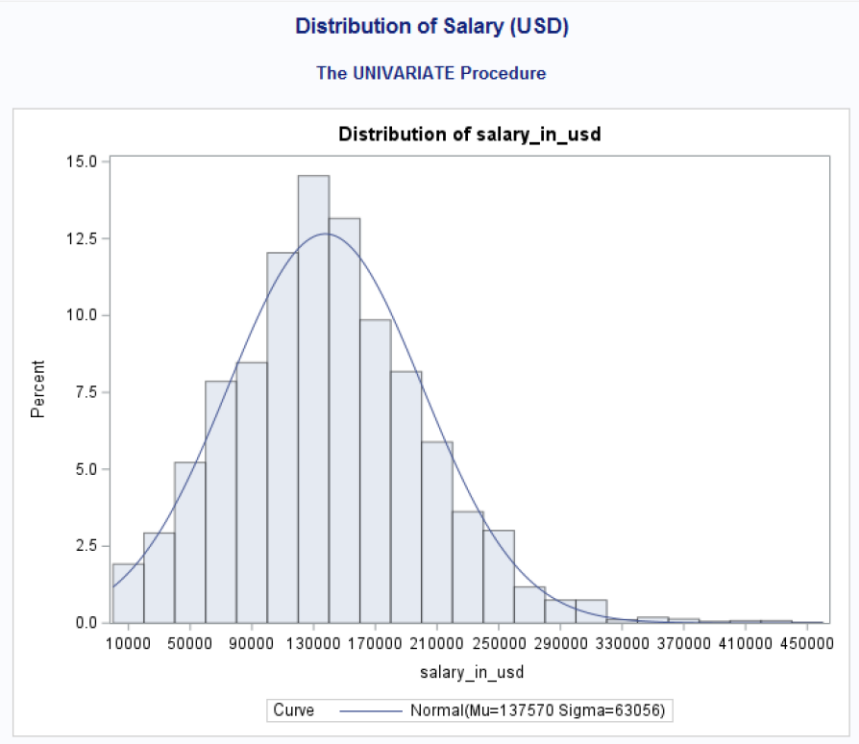
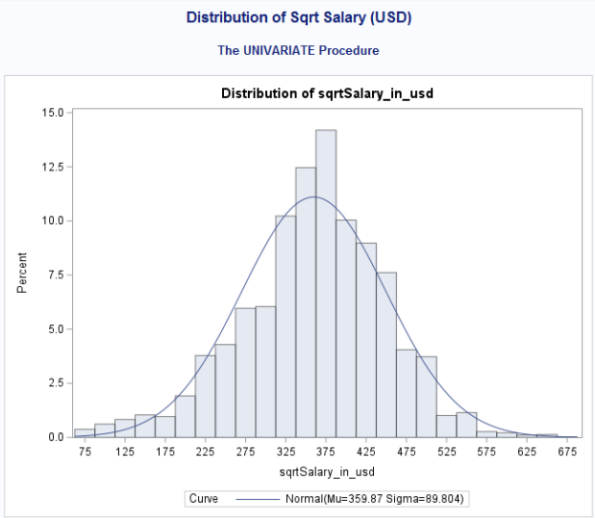


Figure 1: Distribution of dependent variable Salary (USD) before transformation.

Basic Statistical Measures			
Location		Variability	
Mean	137570.4	Std Deviation	63056
Median	135000.0	Variance	3976011879
Mode	100000.0	Range	444868
		Interquartile Range	80000

Quantiles (Definition 5)	
Level	Quantile
100% Max	450000
99%	309400
95%	249500
90%	219000
75% Q3	175000
50% Median	135000
25% Q1	95000
10%	59303
5%	40038
1%	12767
0% Min	5132

Table 3: Statistics for distribution of dependent variable Salary (USD) before transformation.



Basic Statistical Measures			
Location		Variability	
Mean	359.8718	Std Deviation	89.80430
Median	367.4235	Variance	8065
Mode	316.2278	Range	599.18241
		Interquartile Range	110.10931

Figure 2: Distribution of dependent variable Salary (USD) after square root transformation with statistics.

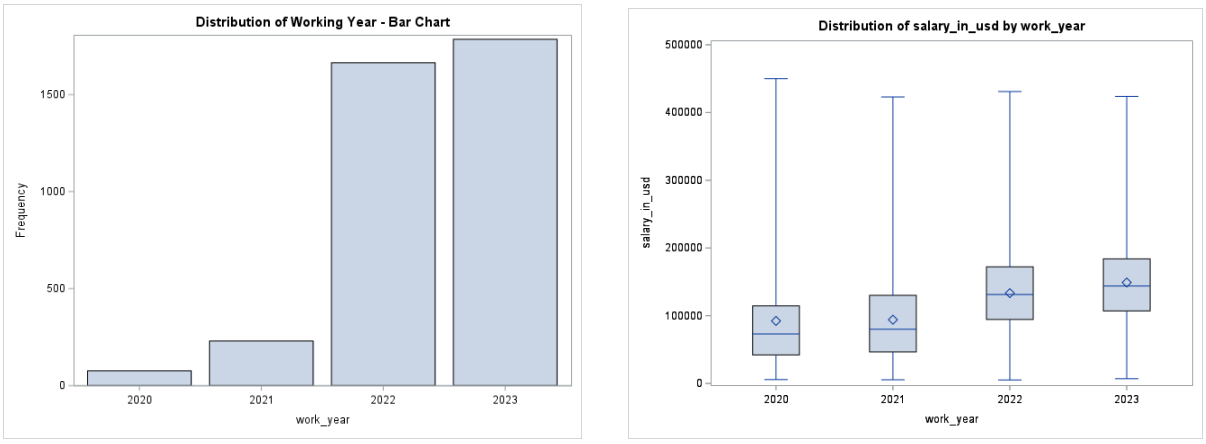


Figure 3: Distribution of Work Year attribute relationship with dependent variable Salary (USD).

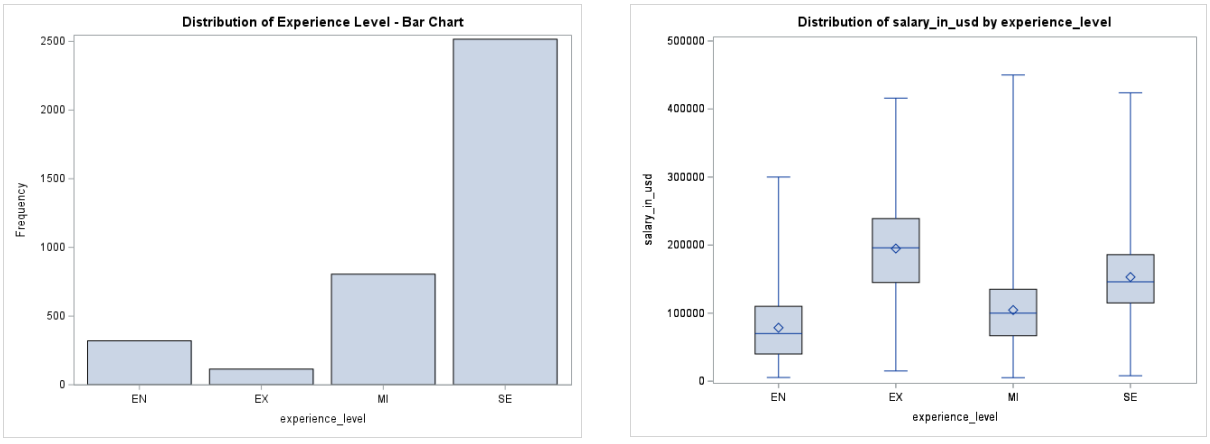
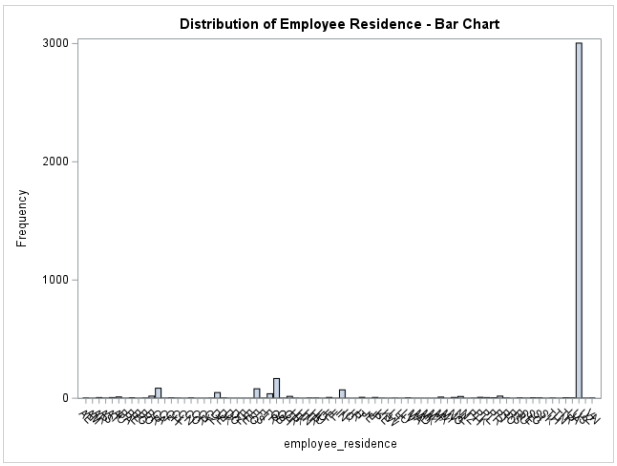


Figure 4: Distribution of Experience Level attribute relationship with dependent variable Salary (USD).



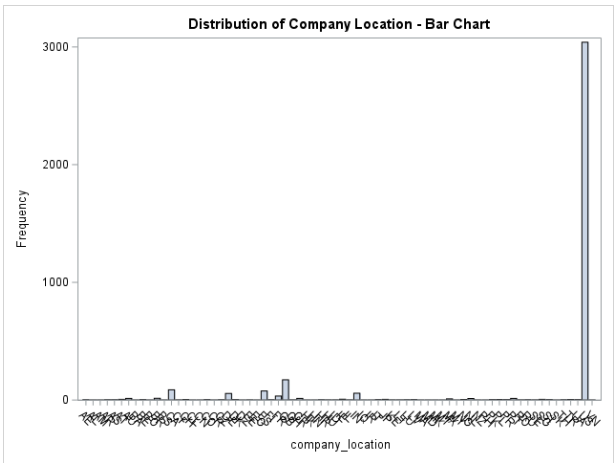


Figure 5: Distribution of Employee Residence and Company Location attributes.

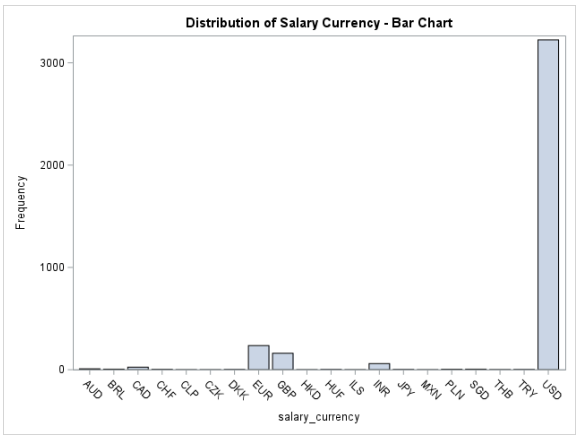
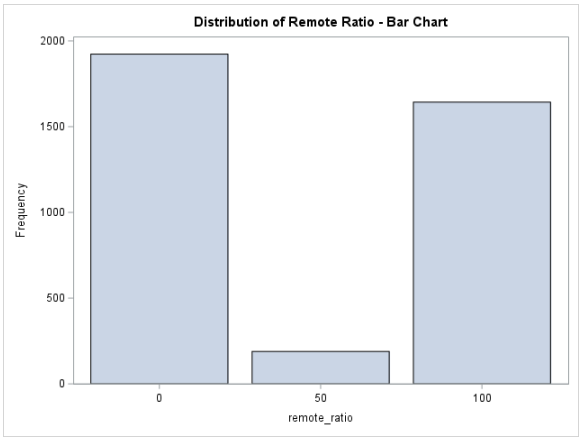


Figure 6: Distribution of Salary Currency attribute.



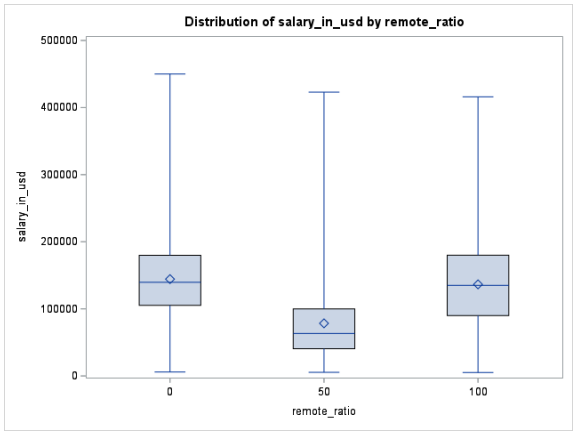


Figure 7: Distribution Remote Ratio attribute relationship with dependent variable Salary (USD).

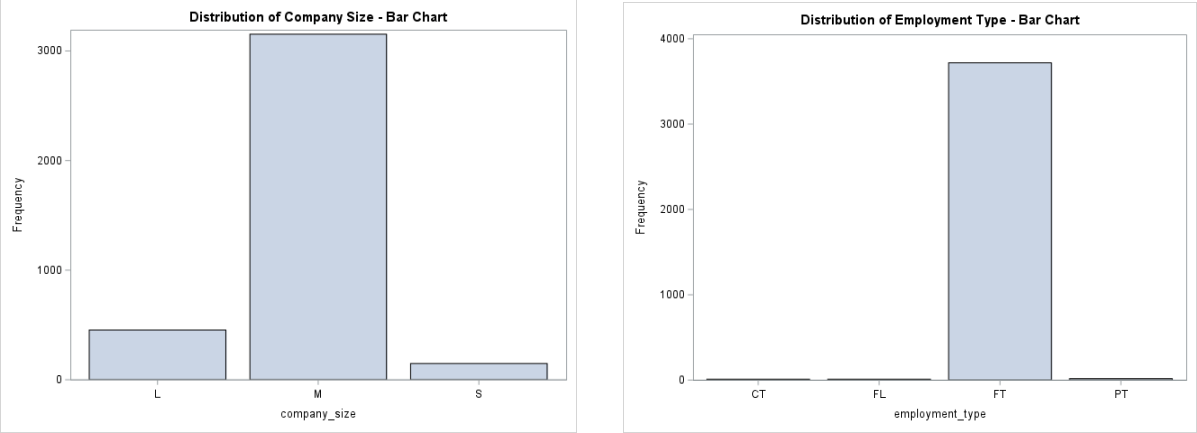


Figure 8: Distribution Company Size and Employment Type attributes.

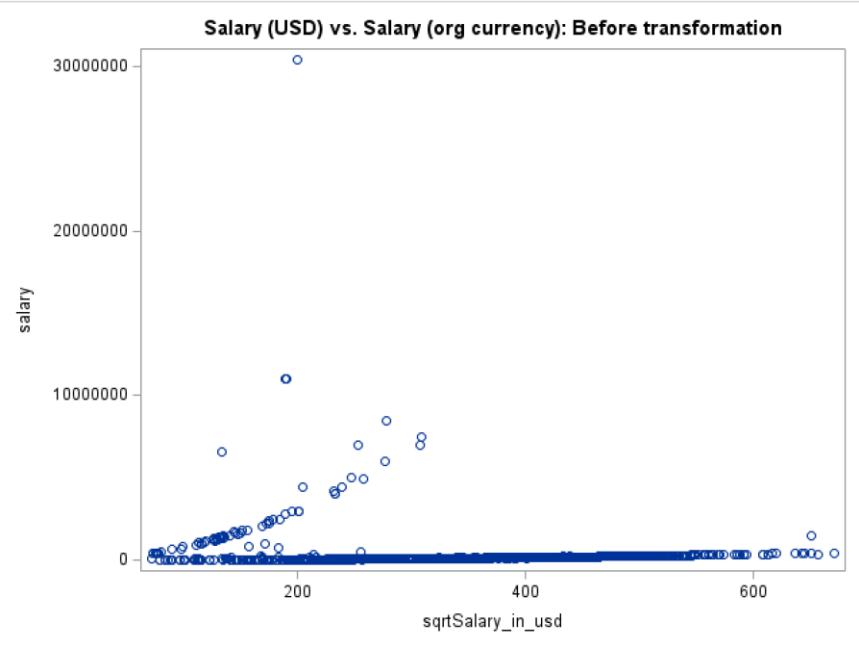


Figure 9: Correlation between dependent variable Salary (USD) and continuous predictor Salary (original currency).

Appendix A – Final Dataset

Final Salary Data 10 observation

Obs	work_year	salary	sqrtSalary_in_usd	numEmployee_residence	numCompany_location	experience_New	employment_FT	size_M	remote_ratioNone	salary_currencyUSD
1	2020	15000	122.474	0	0	0	1	1	1	1
2	2020	95000	308.221	1	1	1	1	1	1	1
3	2020	43200	221.964	0	0	1	1	0	1	0
4	2020	20000	141.421	0	0	1	1	0	1	1
5	2020	450000	670.820	1	1	1	1	1	1	1
6	2020	44000	224.009	0	0	1	1	1	1	0
7	2020	450000	77.923	0	0	1	1	0	1	0
8	2020	720000	183.060	0	0	0	1	0	1	0
9	2020	115000	339.116	0	0	1	1	0	1	1
10	2020	260000	509.902	0	0	0	1	0	1	1

Table 4: The first 10 observations of the final dataset with transformation.

Descriptive Statistics - Final Dataset

The MEANS Procedure

Variable	Minimum	Maximum	Median	25th Pctl	75th Pctl
sqrtSalary_in_usd	71.6379788	670.8203932	367.4234614	308.2207001	418.3300133
work_year	2020.00	2023.00	2022.00	2022.00	2023.00
salary	6000.00	30400000.00	138000.00	100000.00	180000.00
numEmployee_residence	0	1.0000000	1.0000000	1.0000000	1.0000000
numCompany_location	0	1.0000000	1.0000000	1.0000000	1.0000000
experience_New	0	1.0000000	0	0	1.0000000
employment_FT	0	1.0000000	1.0000000	1.0000000	1.0000000
size_M	0	1.0000000	1.0000000	1.0000000	1.0000000
remote_ratioNone	0	1.0000000	1.0000000	0	1.0000000
salary_currencyUSD	0	1.0000000	1.0000000	1.0000000	1.0000000

Table 5: Descriptive statistics of the final dataset with transformation.

Pearson Correlation Coefficients, N = 3755
Prob > |r| under H0: Rho=0

	sqrtSalary_in_usd	work_year	salary	numEmployee_residence	numCompany_location	experience_New	employment_FT	size_M	remote_ratioNone	salary_currencyUSD
sqrtSalary_in_usd	1.00000	0.25858 <.0001	-0.04909 0.0026	0.55093 <.0001	0.53007 <.0001	-0.44903 <.0001	0.15393 <.0001	0.24813 <.0001	0.12865 <.0001	0.51605 <.0001
work_year	0.25858 <.0001	1.00000	-0.09472 <.0001	0.29141 <.0001	0.26700 <.0001	-0.20126 <.0001	0.11631 <.0001	0.42197 <.0001	0.29091 <.0001	0.26908 <.0001
salary	-0.04909 0.0026	-0.09472 <.0001	1.00000	-0.11279 <.0001	-0.10141 <.0001	0.03943 0.0157	0.00673 0.6801	-0.13625 <.0001	-0.05539 0.0007	-0.15163 <.0001
numEmployee_residence	0.55093 <.0001	0.29141 <.0001	-0.11279 <.0001	1.00000	0.96146 <.0001	-0.35754 <.0001	0.14559 <.0001	0.34223 <.0001	0.17129 <.0001	0.81167 <.0001
numCompany_location	0.53007 <.0001	0.26700 <.0001	-0.10141 <.0001	0.96146 <.0001	1.00000	-0.34172 <.0001	0.08209 <.0001	0.31496 <.0001	0.15220 <.0001	0.82125 <.0001
experience_New	-0.44903 <.0001	-0.20126 <.0001	0.03943 0.0157	-0.35754 <.0001	-0.34172 <.0001	1.00000	-0.11721 <.0001	-0.24186 <.0001	-0.08040 <.0001	-0.33855 <.0001
employment_FT	0.15393 <.0001	0.11631 <.0001	0.00673 0.6801	0.14559 <.0001	0.08209 <.0001	-0.11721 <.0001	1.00000	0.12542 <.0001	0.09681 <.0001	0.05237 0.0013
size_M	0.24813 <.0001	0.42197 <.0001	-0.13625 <.0001	0.34223 <.0001	0.31496 <.0001	-0.24186 <.0001	0.12542 <.0001	1.00000	0.23131 <.0001	0.29136 <.0001
remote_ratioNone	0.12865 <.0001	0.29091 <.0001	-0.05539 0.0007	0.17129 <.0001	0.15220 <.0001	-0.08040 <.0001	0.09681 <.0001	0.23131 <.0001	1.00000	0.12069 <.0001
salary_currencyUSD	0.51605 <.0001	0.26908 <.0001	-0.15163 <.0001	0.81167 <.0001	0.82125 <.0001	-0.33855 <.0001	0.05237 0.0013	0.29136 <.0001	0.12069 <.0001	1.00000

Table 6: Correlation table of the final dataset with transformation.

Appendix B. Regression Models

Appendix B – First Model Regression Result

Regression Result: Full model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

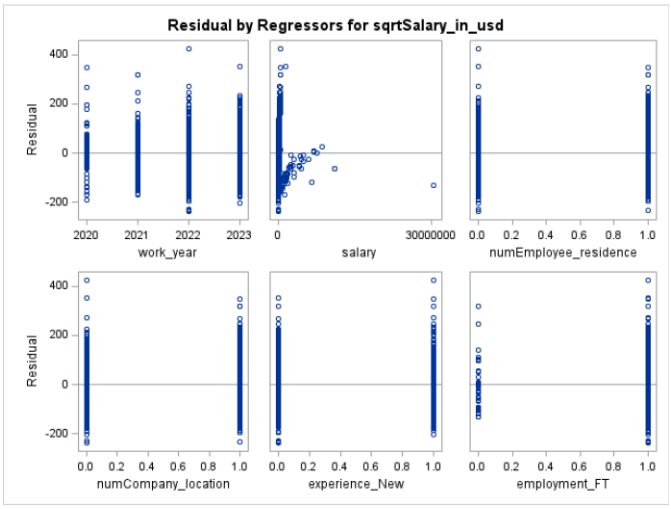
Number of Observations Read	3755
Number of Observations Used	3755

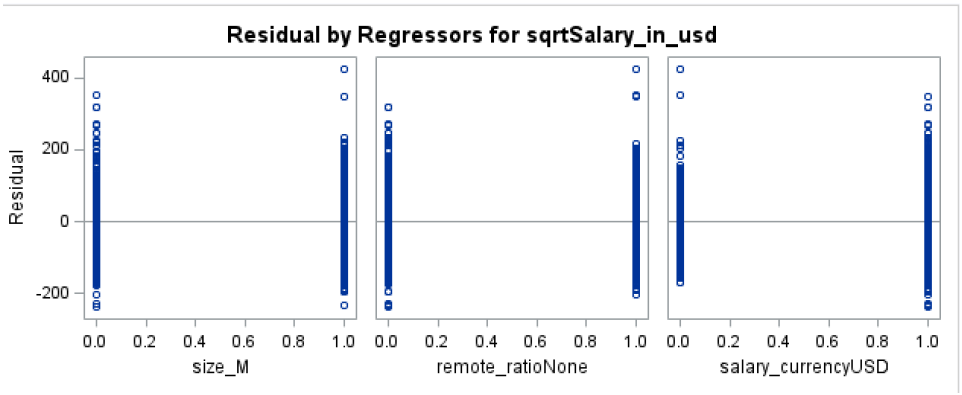
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	11953237	1328137	271.47	<.0001
Error	3745	18322069	4892.40838		
Corrected Total	3754	30275307			

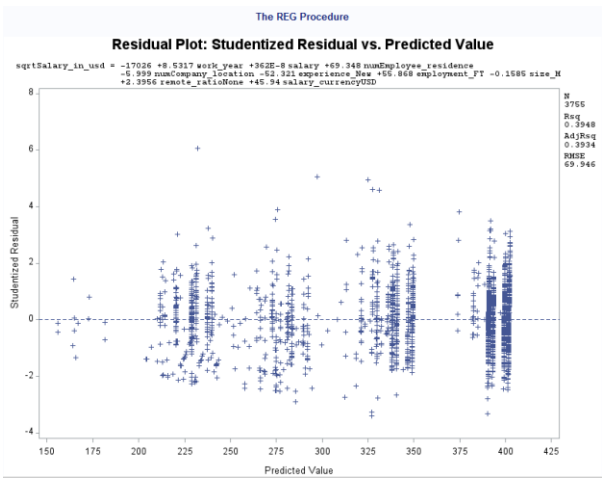
Root MSE	69.94575	R-Square	0.3948
Dependent Mean	359.87182	Adj R-Sq	0.3934
Coeff Var	19.43630		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-17026	3840.57515	-4.43	<.0001	0	.	0
work_year	1	8.53171	1.90003	4.49	<.0001	0.06569	0.75507	1.32438
salary	1	0.00000362	0.00000173	2.09	0.0366	0.02711	0.96187	1.03964
numEmployee_residence	1	69.34802	10.96048	6.33	<.0001	0.30893	0.06779	14.75251
numCompany_location	1	-5.99898	11.17682	-0.54	0.5915	-0.02623	0.06766	14.78031
experience_New	1	-52.32075	2.71579	-19.27	<.0001	-0.26692	0.84184	1.18787
employment_FT	1	55.86772	12.06950	4.63	<.0001	0.06146	0.91673	1.09083
size_M	1	-0.15854	3.60566	-0.04	0.9649	-0.00064782	0.74446	1.34325
remote_ratioNone	1	2.39562	2.41707	0.99	0.3217	0.01334	0.89258	1.12034
salary_currencyUSD	1	45.94042	5.89240	7.80	<.0001	0.17828	0.30907	3.23550

Table 1: First model regression result.







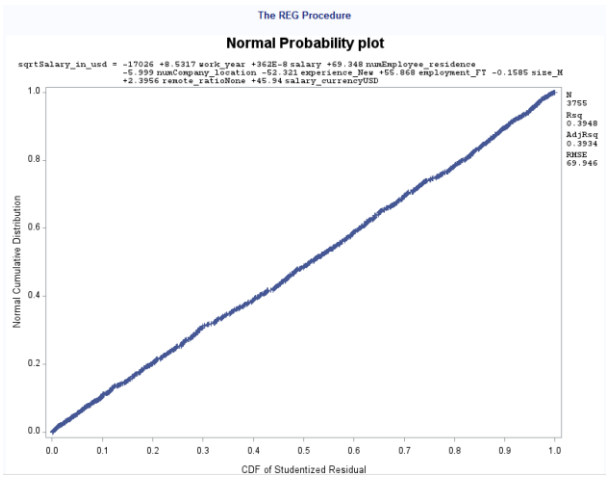


Figure 1: First model regression result – studentized residuals.

Appendix B – Second Model Regression Result

Regression Result: 2nd model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

Number of Observations Read	3755
Number of Observations Used	3755

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	11951828	1493978	305.42	<.0001
Error	3746	18323479	4891.47859		
Corrected Total	3754	30275307			

Root MSE	69.93911	R-Square	0.3948
Dependent Mean	359.87182	Adj R-Sq	0.3935
Coeff Var	19.43445		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-17095	3838.04014	-4.45	<.0001	0	.	0
work_year	1	8.56535	1.89881	4.51	<.0001	0.06595	0.75590	1.32293
salary	1	0.00000358	0.00000173	2.07	0.0388	0.02675	0.96448	1.03683
numEmployee_residence	1	64.13893	5.09284	12.59	<.0001	0.28572	0.31390	3.18574
experience_New	1	-52.33491	2.71541	-19.27	<.0001	-0.26699	0.84192	1.18776
employment_FT	1	57.06262	11.86128	4.81	<.0001	0.06277	0.94902	1.05372
size_M	1	-0.11110	3.60424	-0.03	0.9754	-0.00045395	0.74491	1.34245
remote_ratioNone	1	2.40818	2.41673	1.00	0.3191	0.01341	0.89267	1.12024
salary_currencyUSD	1	45.16984	5.71428	7.90	<.0001	0.17528	0.32858	3.04343

Table 2: Second model regression result.

Appendix B – Stepwise Selection Method

Regression Result: Model Selection Stepwise

The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

Number of Observations Read	3755
Number of Observations Used	3755

Stepwise Selection: Step 6

Variable salary Entered: R-Square = 0.3946 and C(p) = 5.9980

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	11946946	1991158	407.18	<.0001
Error	3748	18328361	4890.17091		
Corrected Total	3754	30275307			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-17974	3530.66178	126732	25.92	<.0001
work_year	8.99995	1.74652	129854	26.55	<.0001
salary	0.00000353	0.00000172	20584	4.21	0.0403
numEmployee_residence	64.55885	5.03422	804214	164.46	<.0001
experience_New	-52.30854	2.70028	1835062	375.26	<.0001
employment_FT	57.65161	11.83070	116125	23.75	<.0001
salary_currencyUSD	44.92033	5.70801	302859	61.93	<.0001

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	numEmployee_residence		1	0.3035	0.3035	559.736	1635.59	<.0001
2	experience_New		2	0.0728	0.3764	110.885	438.25	<.0001
3	salary_currencyUSD		3	0.0089	0.3853	57.7792	54.33	<.0001
4	work_year		4	0.0048	0.3900	30.2717	29.31	<.0001
5	employment_FT		5	0.0039	0.3939	8.2062	24.05	<.0001
6	salary		6	0.0007	0.3946	5.9980	4.21	0.0403

Table 3: Stepwise selection method result.

Appendix B – Third Model Regression Result

Regression Result: 3rd model with outliers/influential points

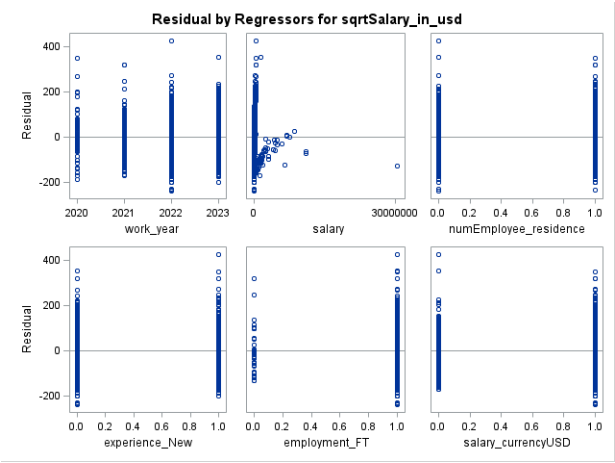
The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

Number of Observations Read	3755
Number of Observations Used	3755

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	11946946	1991158	407.18	<.0001
Error	3748	18328361	4890.17091		
Corrected Total	3754	30275307			

Root MSE	69.92976	R-Square	0.3946
Dependent Mean	359.87182	Adj R-Sq	0.3936
Coeff Var	19.43185		



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-17974	3530.66178	-5.09	<.0001	0	.	0
work_year	1	8.99995	1.74652	5.15	<.0001	0.06930	0.89323	1.11953
salary	1	0.00000353	0.00000172	2.05	0.0403	0.02644	0.97273	1.02803
numEmployee_residence	1	64.55885	5.03422	12.82	<.0001	0.28759	0.32117	3.11365
experience_New	1	-52.30854	2.70028	-19.37	<.0001	-0.26686	0.85115	1.17488
employment_FT	1	57.65161	11.83070	4.87	<.0001	0.06342	0.95368	1.04857
salary_currencyUSD	1	44.92033	5.70801	7.87	<.0001	0.17432	0.32921	3.03756

Table 4: Third model regression result.

H3 | =IF(COUNTIF(E3:F3, "Yes")>0, "Remove", "Keep")

	A	B	C	D	E	F	G	H
1	Ob	Student Residual	Cook's I		StudentResi Outliers	Cook's D influential		Removing point
2	1	-2.676	0.008		No	Yes		Remove
3	2	-0.188	0		No	No		Keep
4	3	0.147	0		No	No		Keep
5	4	-1.653	0.003		No	Yes		Remove
6	5	4.991	0.017		Yes	Yes		Remove
7	6	0.176	0		No	No		Keep
8	7	-1.938	0.002		No	No		Keep
9	8	-1.195	0.001		No	No		Keep

Figure 2: Outlier and Influential points extraction through Excel.

Appendix B – Final Model Regression Result

Regression Result: Final model

The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

Number of Observations Read	3662
Number of Observations Used	3662

Analysis of Variance

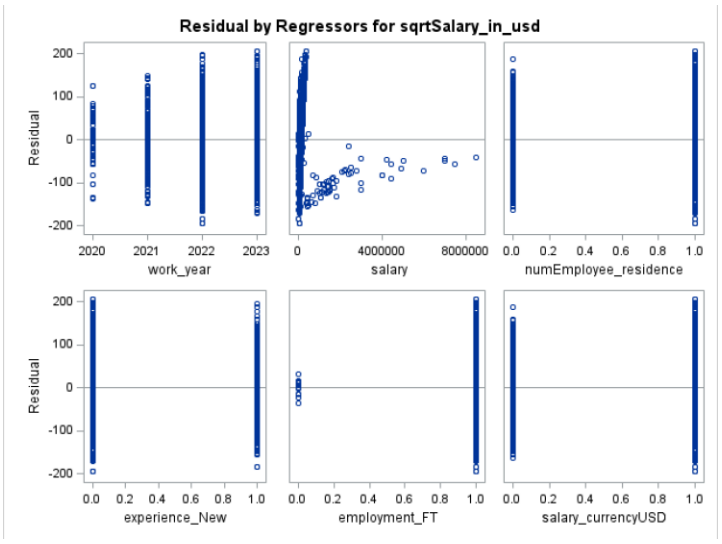
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	11063745	1843958	440.04	<.0001
Error	3655	15316051	4190.43803		
Corrected Total	3661	26379796			

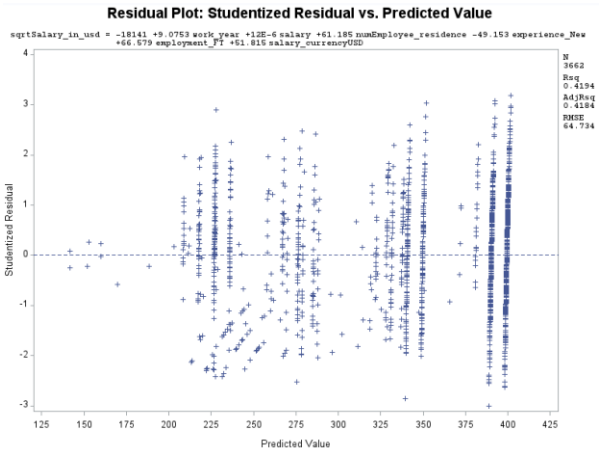
Root MSE	64.73359	R-Square	0.4194
Dependent Mean	360.77592	Adj R-Sq	0.4184
Coeff Var	17.94288		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	-18141	3374.55573	-5.38	<.0001	0	.	0
work_year	1	9.07530	1.66943	5.44	<.0001	0.07221	0.90019	1.11087
salary	1	0.00001202	0.00000295	4.08	<.0001	0.05235	0.96508	1.03619
numEmployee_residence	1	61.18466	5.31927	11.50	<.0001	0.27984	0.26837	3.72615
experience_New	1	-49.15343	2.53667	-19.38	<.0001	-0.26323	0.86081	1.16169
employment_FT	1	66.57913	17.49172	3.81	0.0001	0.04841	0.98204	1.01829
salary_currencyUSD	1	51.81486	5.94047	8.72	<.0001	0.21258	0.26744	3.73916

Table 5: Fourth (final) model regression result.





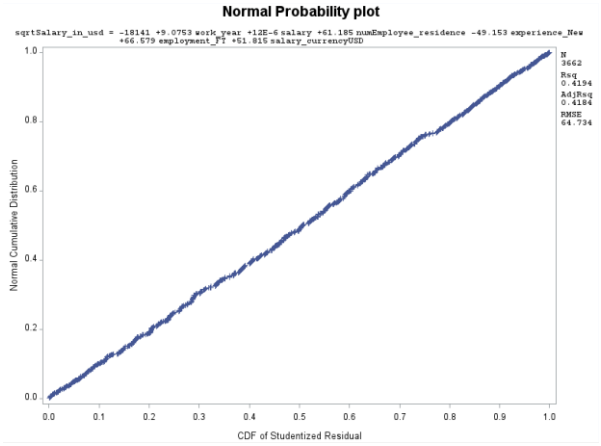


Figure 3: Final model regression result – studentized residuals.

Appendix C. Prediction

Prediction data

Obs	work_year	salary	numEmployee_residence	experience_New	employment_FT	salary_currencyUSD
1	2024	100000	1	1	1	1
2	2024	100000	0	1	1	0

Prediction data

Obs	work_year	salary	numEmployee_residence	experience_New	employment_FT	salary_currencyUSD	sqrtSalary_in_usd	numCompany_location	size_M	remote_ratioNone
1	2024	100000	1	1	1	1	.	.	.	.
2	2024	100000	0	1	1	0	.	.	.	.
3	2020	95000	1	1	1	1	308.221	1	1	1
4	2020	43200	0	1	1	0	221.964	0	0	1
5	2020	44000	0	1	1	0	224.009	0	1	1

Table 1: Prediction dataset created and merged with final dataset.

Regression Analysis and Confidence Interval for Average Estimate.

The REG Procedure
Model: MODEL1
Dependent Variable: sqrtSalary_in_usd

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	358.5587	3.5537	351.5912	365.5261	231.4500	485.6673	.
2	.	245.5592	4.6302	236.4812	254.6371	118.3174	372.8009	.
3	308.2	322.1974	4.5975	313.1834	331.2114	194.9601	449.4346	-13.9767
4	222.0	208.5753	4.4438	199.8627	217.2880	81.3591	335.7916	13.3886
5	224.0	208.5850	4.4433	199.8734	217.2965	81.3688	335.8011	15.4240
6	77.9	213.4642	4.3393	204.9565	221.9719	86.2618	340.6665	-135.5411

Table 2: Prediction results.

Appendix D. Validation

Appendix D – Split Dataset

Test and Train Set for Salary Data

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	FINALSALARY2
Random Number Seed	1997
Sampling Rate	0.8
Sample Size	2930
Selection Probability	0.800109
Sampling Weight	0
Output Data Set	XVALSALARY

Test and Train Set for Salary Data

Obs	Selected	work_year	salary	sqrtSalary_in_usd	numEmployee_residence	numCompany_location	experience_New	employment_FT	size_M	remote_ratioNone	salary_currencyUSD
1	1	2020	43200	221.964	0	0	1	1	0	1	0
2	1	2020	450000	77.923	0	0	1	1	0	1	0
3	1	2020	45000	226.541	0	0	1	1	0	1	0
4	0	2020	35000	199.790	0	0	1	1	1	1	0
5	0	2021	100000	316.228	1	1	1	1	0	1	1
6	0	2021	80000	331.718	0	0	1	1	0	1	0
7	1	2021	100000	316.228	0	0	1	1	0	1	1
8	1	2021	65000	277.188	0	0	1	1	0	1	0

Table 1: The first 10 observations of the split train – test data (80/20).

Create predicted value for cross validation

Obs	Selected	work_year	salary	sqrtSalary_in_usd	numEmployee_residence	numCompany_location	experience_New	employment_FT	size_M	remote_ratioNone	salary_currencyUSD	new_y
1	1	2020	43200	221.964	0	0	1	1	0	1	0	221.964
2	1	2020	450000	77.923	0	0	1	1	0	1	0	77.923
3	1	2020	45000	226.541	0	0	1	1	0	1	0	226.541
4	0	2020	35000	199.790	0	0	1	1	1	1	0	.
5	0	2021	100000	316.228	1	1	1	1	0	1	1	.
6	0	2021	80000	331.718	0	0	1	1	0	1	0	.
7	1	2021	100000	316.228	0	0	1	1	0	1	1	316.228
8	1	2021	65000	277.188	0	0	1	1	0	1	0	277.188
9	1	2021	435000	76.694	0	0	1	1	0	1	0	76.694
10	1	2022	85000	291.548	1	1	1	1	1	1	1	291.548

Table 2: The first 10 observations after creating new dependent variable.

Appendix D – Stepwise Selection Method

Full model regression with stepwise selection method: Train set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3662
Number of Observations Used	2930
Number of Observations with Missing Values	732

Stepwise Selection: Step 6

Variable salary Entered: R-Square = 0.4207 and C(p) = 6.4290

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	8901260	1483543	353.74	<.0001
Error	2923	12258868	4193.93376		
Corrected Total	2929	21160129			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-19635	3756.83995	114566	27.32	<.0001
work_year	9.81216	1.85865	116884	27.87	<.0001
salary	0.00001137	0.00000312	55813	13.31	0.0003
numEmployee_residence	57.75484	6.11103	374601	89.32	<.0001
experience_New	-48.83849	2.84691	1234238	294.29	<.0001
employment_FT	70.26382	18.91676	57862	13.80	0.0002
salary_currencyUSD	54.56832	6.76818	272621	65.00	<.0001

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	numEmployee_residence		1	0.3271	0.3271	468.462	1423.23	<.0001
2	experience_New		2	0.0700	0.3971	117.467	339.71	<.0001
3	salary_currencyUSD		3	0.0124	0.4095	56.7726	61.58	<.0001
4	work_year		4	0.0058	0.4153	29.5528	28.98	<.0001
5	employment_FT		5	0.0027	0.4180	17.7344	13.76	0.0002
6	salary		6	0.0026	0.4207	6.4290	13.31	0.0003

Table 3: Stepwise selection method on training data.

Appendix D – Regression Results

Validation - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3662
Number of Observations Used	2930
Number of Observations with Missing Values	732

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	8901260	1483543	353.74	<.0001
Error	2923	12258868	4193.93376		
Corrected Total	2929	21160129			

Root MSE	64.76059	R-Square	0.4207
Dependent Mean	360.09265	Adj R-Sq	0.4195
Coeff Var	17.98442		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-19635	3756.83995	-5.23	<.0001	0	0
work_year	1	9.81216	1.85865	5.28	<.0001	0.07830	1.11000
salary	1	0.00001137	0.00000312	3.65	0.0003	0.05233	1.03829
numEmployee_residence	1	57.75484	6.11103	9.45	<.0001	0.26388	3.93345
experience_New	1	-48.83849	2.84691	-17.15	<.0001	-0.26125	1.17015
employment_FT	1	70.26382	18.91676	3.71	0.0002	0.05280	1.01970
salary_currencyUSD	1	54.56832	6.76818	8.06	<.0001	0.22524	3.93765

Table 4: Train set regression result.

Validation - Test Set

bs	Selected	work_year	salary	sqrtSalary_in_usd	numEmployee_residence	numCompany_location	experience_New	employment_FT	size_M	remote_ratioNone	salary_currencyUSD	new_y	yhat
1	0	2020	35000	199.790	0	0	1	1	1	1	0	.	206.945
2	0	2021	100000	316.228	1	1	1	1	0	1	1	.	329.820
3	0	2021	80000	331.718	0	0	1	1	0	1	0	.	217.269
4	0	2022	6000	79.398	0	0	1	1	0	1	0	.	226.240
5	0	2022	130000	360.555	1	1	1	1	1	1	1	.	339.973
6	0	2022	160000	400.000	1	1	1	1	1	1	1	.	340.314
7	0	2022	160000	400.000	1	1	1	1	1	1	1	.	340.314
8	0	2022	189750	435.603	1	1	1	1	1	1	1	.	340.653
9	0	2022	160000	400.000	1	1	1	1	1	1	1	.	340.314
10	0	2022	108000	328.634	1	1	1	1	0	1	1	.	339.723

Table 5: The first 10 observations after generating predicted values.

Validation statistics for Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	732	64.6574	52.1490

Correlation between Observed and Predicted values

The CORR Procedure

2 Variables:	sqrtSalary_in_usd yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
sqrtSalary_in_usd	732	363.51085	84.44576	266090	75.54469	587.87754	
yhat	732	360.86018	53.41072	264150	156.23741	400.71754	Predicted Value of new_y

Pearson Correlation Coefficients, N = 732
Prob > |r| under H0: Rho=0

	sqrtSalary_in_usd	yhat
sqrtSalary_in_usd	1.00000	0.64333 <.0001
yhat Predicted Value of new_y	0.64333 <.0001	1.00000

Table 6: Test set descriptive statistics.