AN EXPERIENTIAL AND DETERMINISTIC APPROACH ON SOLVING U.S. COMPANIES' NEED FOR DATA SCIENTISTS

MGT 6203, FALL 2022

Submission Date: November 20, 2022

TEAM 20 Members

(Team Lead) Sunil Prasad

Buddhi Pantha

Jewell Powers

Amit Patel

Professors Bien, Narasimhan, Clarke, & Myers

Table of Contents

Figures & Tables

Overview of Project

According to Harvard Business Review, with the emerging needs in the market and in industry, there is a shortage of data science professionals in the U.S. market, so we as Group-20 want to explore this problem in the context of how it affects the U.S. market for data science professionals (Thomas H. Davenport, 2012). According to Forbes.com, salaries for the data science professionals depend on several factors such as the job title, region or location, size of the company, education, and experience (Columbus, 2018). To address this problem, we wanted to evaluate whether it is cost effective to outsource and hire data scientists from India or to employ expensive data scientists from the U.S. job market by comparing the average salaries of data science professionals between the U.S. and India. We also studied the impact of predictor variables such as company size, job titles, regions/cities, and type of ownership on the response variable salaries. Thus, we came up with the following hypothesis:

Hypothesis: *We propose to investigate which of the driving factors play a significant role in the salaries and whether U.S. companies should outsource data analytics to India rather than committing to reorganizing, overhauling, and implementing data analytics throughout the organization by hiring more U.S. data science professionals.*

Literature Survey

1. The data science field has experienced a 650% plus growth since 2012 with global market projections of $230.8 billion by the year 2026 and India will capture 32% (more than 11 million) of that market. On the other hand, there are very few universities training data science professionals and there is a lack of information circulated about this massive potential to prospective students (Gupta, 2022).

2. Analytics India Magazine studied the trends of data Science professionals salaries and market demand and supply by analyzing the trends across cities, geographies, sectors, salary brackets, experience levels, gender, and usage of key technologies in the industry. India data analytics industry is expected to grow from $61.1 billion in 2022 to $201 billion in 2027. Approximately 51% of Data Science professionals in India work for U.S. companies. (Bhorayal, 2022).

3. Americans have always had a pervasive underlying fear of outsourcing in the US. It has waned slightly in recent years as the practice has become more common, but will Americans ever get over that fear that Daniel Drezner calls the "Outsourcing Bogeyman". "Critics charge that the information revolution (especially the Internet) has accelerated the decimation of U.S. manufacturing and facilitated the outsourcing…This concern feeds into the suspicion that U.S. corporations are exploiting globalization to fatten profits at the expense of workers." Outsourcing falls into the foreign trade arena and has long been a hot-button topic between US political parties. However, most US jobs should not be affected because approximately 90% of jobs in the US require geographic proximity. Insourcing is one aspect to countermand that fear as many firms in other countries outsource positions to the U.S. (Drezner, 2004).

Project Architecture and Concept of Operations

We created a ConOps (Concept of Operations) with an effective set of guidelines in implementing our hypothesis. Our project business case is to evaluate whether it is cost effective for the U.S. job market by analyzing the impact of the predictor variables such as company size, job titles, regions/cities, etc., against the response variable, salaries.

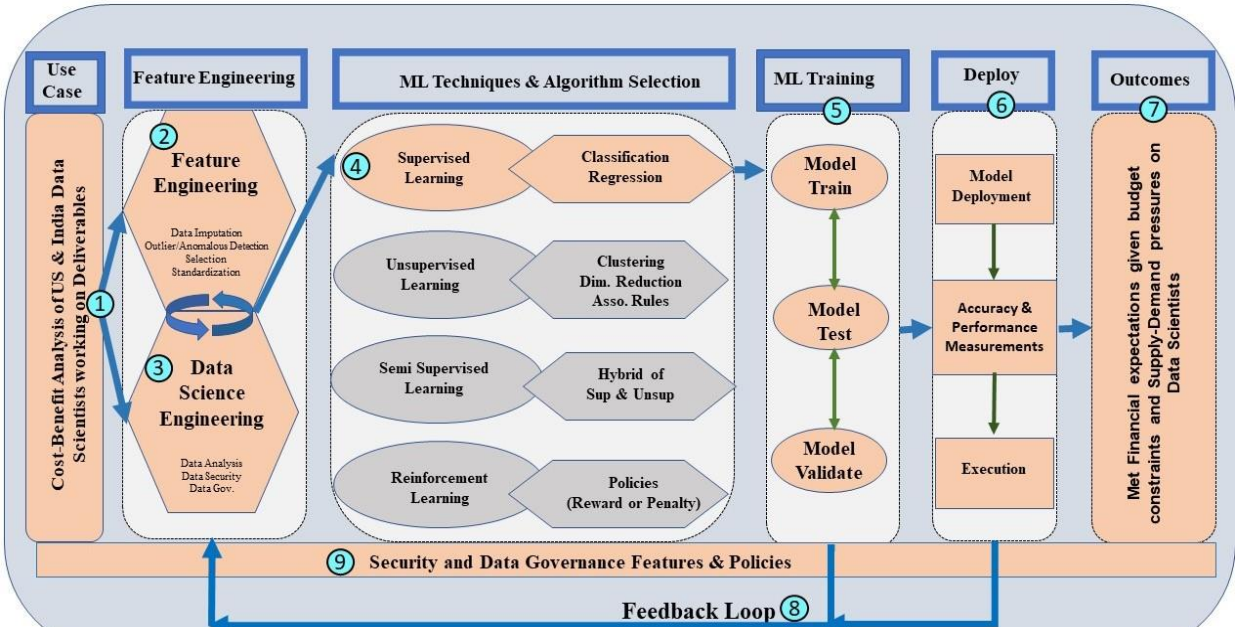

Figure 1: Project ConOps

This interoperable and scalable ConOps is designed with the principle to scale the business opportunities for US companies with cost benefits. The flow of the arrows depicts the building blocks' operations of our ConOps (**Figure 1**) as shown with nine numbers ensuring a solid Machine Learning platform for our project.

- Business Use Case
- Feature Engineering & Data Science engineering hyperparameters selections
- ML Techniques and Algorithm selections
- ML Train/Test/Validate
- Model Deployment
- Outcomes
- Model Deployment with Feedback Loop
- Security & Data Governance features & policies

Serial#	Descriptions
1.	Use case: Cost-Benefit Analysis of US and India Data Scientists for clients for the best deliverables
2. & 3.	Preformed feature engineering and data preprocessing such as • Data imputations, handling of outliers/anomalous detection, Data scaling & Standardization • Data Science Engineering
4.	Although there are many Machine Learning (ML) techniques such as Supervised, Unsupervised, Semi-Supervised and Reinforcement, we selected Supervised learning with Regression algorithm for our project

Serial#	Descriptions
5.	ML Training, Testing and Validation
6.	Performed the model deployment with performance measurements and execution **a)** Model Deployment **b)** Accuracy & Performance Measurements **c)** Execution
7.	Met Financial expectations given budget constraints and Supply-Demand pressures on Data Scientists
8.	Created the feedback loop based on precision and accuracy rates as a performance measurement
9.	Incorporated Security and Data Governance policies to prevent any vulnerability

Table 1: ConOps Descriptions/Steps

Overview of Data

In this study, although we started with six datasets, we settled on the two most comprehensive datasets that were relevant to our study by using salaries data (Pant, 2022), (Prithviraj, 2022) for Data Science professionals from the Kaggle repository. The U.S. dataset has 956 data points with 14 variables and the one from India has 4339 data points with 7 variables. Methodologies used in the data analysis approach is called "Data Lifecycle Framework" and consists of the six key steps in **Figure 2**.

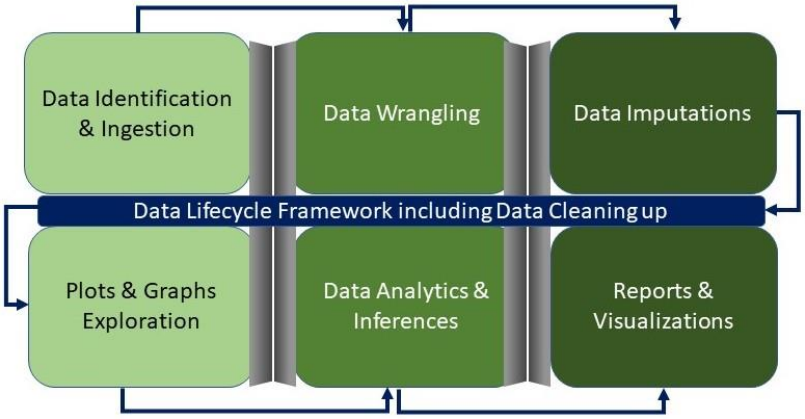

Figure 2: Data Lifecycle Framework

We imported the data in R and extracted the salaries and states from compound text fields. As an example, the salary was given in text format "$50K-$175K, estimated from Glassdoor". We extracted the salary range and used their average as the salary data.

Next, we used data wrangling as a process to transform Kaggle U.S. and India data into a format with the intention of making our datasets more pertinent and valuable to further downstream analytics and model building. One of the main objectives of our data wrangling was to assure quality and usefulness of the data for accurate modeling techniques. Yet another objective with data wrangling is to utilize data visualization and data aggregation which leads into training of the models. Additionally, we realized that there were some missing values, outliers, and unrealistic annual salaries such as "-1" or "$1,000". During our imputations process, we evaluated the average salaries without those missing outliers and unrealistic values and replaced them with the average. For example, we imputed unrealistic salaries that were less than $20,000 with overall salary mean.

Similarly, locations were given as city and state names (such as "Los Angeles, CA"). There were 38 states including DC and many cities in the dataset. We extracted just the states and grouped them into different regions (Regions of the US, 2022). The six regions are: MA – MidAtlantic, MW – Midwest, NE – New England, S – South, SW – Southwest, W – West. Similarly, we reduced 327 job titles into just four job titles 'Data Analyst', 'Data Engineer', 'Data Scientist', and 'Other'. The company size was based on number of employees, and we used this information to categorize the size of the company. We consolidated numerical company sizes into categorical variables (see Appendix B).

For the Indian data, we selected three variables Salary, Job Title, and City. The salary was given in text format (for example 8 Lakh-12 Lakhs). We extracted the salaries as we did with the U.S. data, calculated the average salary, and converted them into U.S. dollars using the current exchange rate (1 USD = 82.29 INR).

Exploratory Data Analysis (EDA) and Key Visualizations

We began our analysis with the U.S. Data Science professional datasets. Our key variables are Average Salary, rating, job titles, regions, company size, and type of ownership. We used the EDA which is a powerful tool to observe the basic feature of our data, such as distribution of the data, comparison of different categories in the data, etc. The boxplot for different factors and their categories is presented in **Figure 3.**

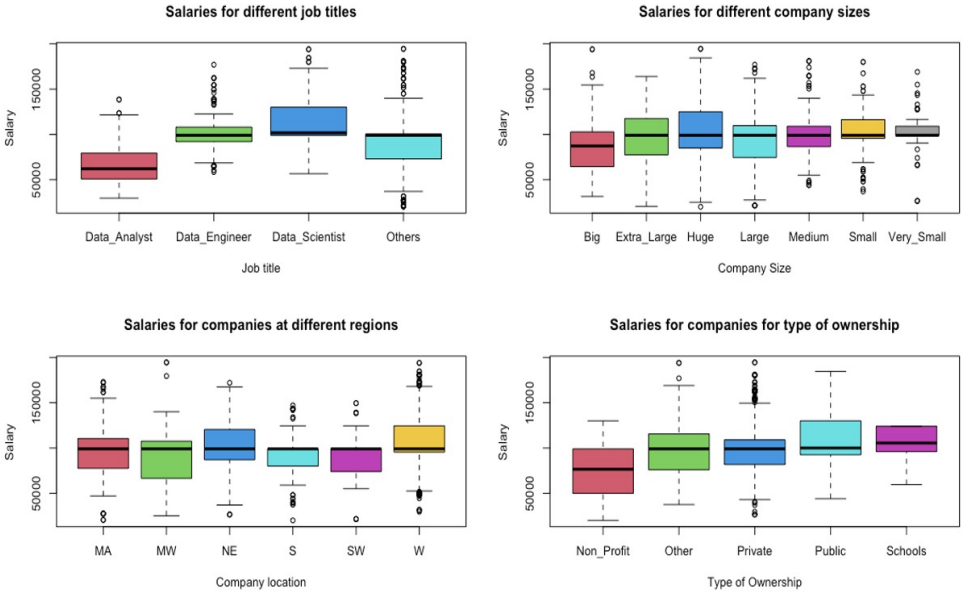

Figure 3: Boxplots for U.S. Salaries

From the boxplot, we observed that the median salaries for Data Analyst in the U.S. is the lowest and the Data Scientist is the highest median salaries. Also, the salaries of Data Scientists and Others are not symmetrical about the mean.

Similarly, the 'Big' companies have the lowest median salaries relative to other company sizes and the 'West' region has the highest median salaries for Data Science professionals

compared to the other regions. Finally, the 'Non-Profit' companies have the lowest median salaries compared to other types of companies.

We also compared average salaries between categories under each factor using One Factor ANOVA. If the average salaries between the categories are significantly different, we performed pairwise comparison using the Tukey test. Also, we tested for any violation of assumptions for ANOVA. If the assumptions were violated, we used data transformations such as Log Transformation. We also used Non-Parametric Test (NPT) to compare the median salaries between the categories under each factor using the Kruskal-Wallis test. For example, when comparing the average salaries for different job titles, our ANOVA suggests that the average salaries are significantly different. The Tukey test implied significantly different salaries for categories that compared pairwise. As the data are not normally distributed even after log transformation, we applied the Kruskal-Wallis test to compare the median salaries. We found that the median salaries of the different job titles are significantly different, and we also observed that the results were analogous to the ANOVA test.

We performed similar analyses for other factors such as different company sizes, regions, and type of ownership. We noticed that the average salaries for various company sizes in the U.S. are significantly different and the average salaries for various regions are significantly different. The pairwise comparison illustrates the West has different average salaries than regions MA, MW, S, and SW. Additionally, region S has significantly different average salaries than region NE and NW. Similarly, the Big companies have significantly different average salaries than Huge, Medium, and Small. Additionally, Large companies have significantly different average salaries than Huge companies. As in Job Title, we performed log transformation and then used NPT which depicted significantly different median salaries between the U.S. regions and amongst the company sizes. For type of ownership, we observed that Non-profit companies pay significantly lower salaries than Public, Private, Schools, and Others. Similarly, the Public companies pay significantly higher salaries than Private companies.

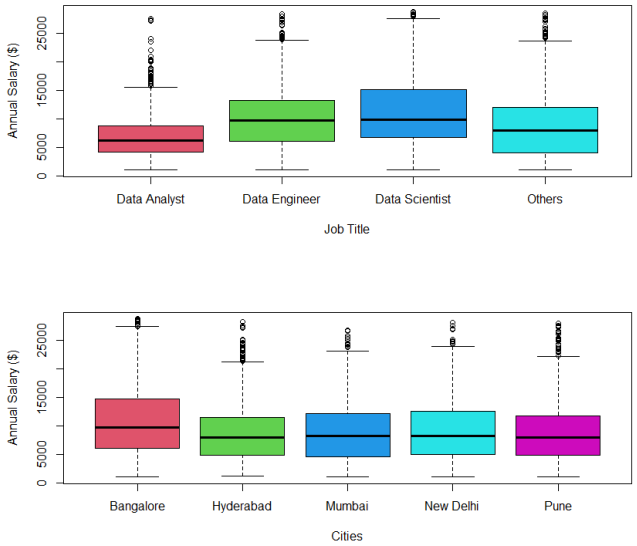

Figure 4: Boxplots for Indian Salaries

Next, for the Indian data, the variables are Salary, Job Title, and City. Under the Job Titles, we had 25 different job titles and we grouped them into just four job titles as we did with the U.S. data. Then we created a boxplot for the different job titles and cities shown into **Figure 4**.

From the plot we see that the Data Analyst has the lowest median salary and Data Scientist has the highest median salary. The City of Bangalore has the highest median salary compared to the other four cities. When comparing the average salaries for different job titles we observed that the salaries are significantly different. The pairwise comparison using the Tukey test implies that the Data Analysts have significantly different (lower) annual salaries than Data Scientist and Data Engineer. Similarly, the Data Scientists have significantly different (higher) annual salaries than the Other category. As the data are not normally distributed even after log transformation, we applied the Kruskal-Wallis test to compare the median salaries which implies that the median salaries are significantly different for the different job titles. Between Indian cities, we observed that the average salaries are not significantly different, but Bangalore has significantly higher median salaries than Hyderabad, Mumbai, and Pune.

Finally, we compared the average salaries of Data Science professionals in India and the U.S. In **Figure 5**, we see that the Data Science professionals in the U.S. make significantly higher salaries compared to their India counterparts. The ANOVA/t.test implies significantly higher salaries with 95% confidence interval (LL= $87,169.89 and HL= $91,160.74) for the difference. Also, the NPT indicates that the median salaries of U.S. Data Science professionals are significantly higher. Note that we randomly selected the data from India to match the sample size from the U.S.

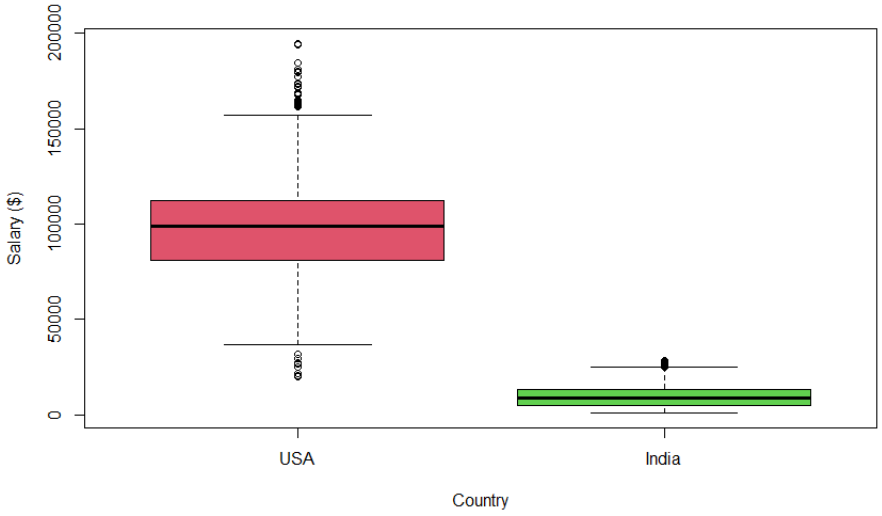

Figure 5: Boxplot of U.S. and Indian Salaries

Model Overview, Results, and Key Findings

We established feature engineering and feature selection to incorporate ML algorithms for training/testing/validating our predictive models before deployment. We used the following feature engineering, analytics guidelines, and predictive models in our project discussed below.

Our U.S. dataset has six variables: Salary, Rating, Job Title, Region, Company Size, and Type of Ownership. The salary and rating are numerical variables, and the remaining are categorical. We first ran a multiple linear regression with salary as the response variables and all remaining variables as predictors. Notice that the predictor variables Job Title, Company Size, Type of Ownership, and Region have more than one category. For each label of categorical predictor, we then created dummy variables with baseline label as Data Analyst (for Job Title), Middle Atlantic (for Region), Non-Profits (for Type of Ownership), and Big (for Size of company). After that, we checked the collinearity between the predictors by finding the correlation coefficients in which we did not observe any significant correlation between any of the predictors.

Our initial regression analysis showed that some variables are not significant. Thus, we used a simple variable selection technique (stepwise regression) to find the important variables. Then we ran multiple linear regression for the predictors obtained from stepwise regression. The image of our final regression output is presented in **Figure 6.**

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)             38200       3761  10.156  < 2e-16 ***
Region_S                -4436       2383  -1.861   0.0630 .
Region_SW               -9581       3729  -2.569   0.0103 *
Region_W                11519       1991   5.785 9.86e-09 ***
CSize_Extra_Large        5503       3168   1.737   0.0827 .
CSize_Huge               6278       2582   2.431   0.0152 *
Title_Data_Engineer     34545       3224  10.714  < 2e-16 ***
Title_Data_Scientist    41428       2818  14.699  < 2e-16 ***
Title_Others            24407       2838   8.600  < 2e-16 ***
Ownership_Other         32099       4464   7.190 1.32e-12 ***
Ownership_Private        28527       2923   9.761  < 2e-16 ***
Ownership_Public         35061       3116  11.251  < 2e-16 ***
Ownership_Schools        40306       6820   5.910 4.77e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25370 on 943 degrees of freedom
Multiple R-squared:  0.3328,    Adjusted R-squared:  0.3244
F-statistic: 39.21 on 12 and 943 DF,  p-value: < 2.2e-16
```

Figure 6: Regression Output U.S.

From the result, we see at $\alpha = 0.05$, all predictor variables are significant except South (Region) and Extra-Large (size) company. Additionally, the Data Science professionals in the South make $4,436 less on average than their counterparts in the Mid-Atlantic Region keeping all other predictors constant. On the other hand, the Data Science professionals in the West make $11,519 higher on average than Mid-Atlantic keeping all other predictors constant.

Additionally, the Huge size company which are companies with 10,001 or more employees make $6,278 more on average than Data science professionals who worked at Big companies which are companies with 501 to 1,000 employees keeping all other predictors constant. One of the most surprising results to us was that Schools on average had the highest salaries for Data Science professionals than all the other job titles. The Adjusted R-Squared for the Regression Model is .3244 meaning that 32.44% of the variation in our response variable salaries is explained by our predictors.

While we found that some of the variables were not significant, we preliminarily assessed our regression output to get a basic understanding of the data and then proceeded to test if any of

the linear regression assumptions were violated. Since we saw that normality was violated, after testing the linear regression assumptions, we ran a log-linear regression using log transformation for our response variable Salary. We found that at $\alpha = 0.05$, all predictors were statistically significant except Huge, Extra-Large, and South. Furthermore, the log transformed model has approximately normally distributed residuals and it does not show an indication of heteroscedasticity. Also, there was a slight improvement in the Adjusted R-Squared value in the log-linear model (36.25% for log-linear and 32.44% for linear) and there were no influential points detected (Cook's Distance < 0.02).

Lastly, to measure the accuracy and performance of our models. We used k-fold cross validation method (k=3) for both linear and log-linear models. The R-Squared values in the regression models and their cross-validation were similar indicating no significant overfitting.

Moving on to the India dataset, it has three variables: Salary, Job Title, and City. Salary is numerical while Job Title and City are categorical. We followed the same variable selection procedures as we did for the U.S. dataset. The result from the regression analysis is presented in **Figure 7.**

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           7713.4      231.0  33.395  < 2e-16 ***
`Title_Data Engineer` 3122.5      265.9  11.745  < 2e-16 ***
`Title_Data Scientist` 4137.7     227.3  18.202  < 2e-16 ***
Title_Others          1750.5      284.9   6.144 8.79e-10 ***
Location_Hyderabad    -903.0      264.7  -3.411 0.000652 ***
Location_Mumbai       -912.2      271.6  -3.359 0.000789 ***
`Location_New Delhi`  -655.4      266.8  -2.457 0.014061 *
Location_Pune        -1206.6      244.3  -4.939 8.15e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5594 on 4331 degrees of freedom
Multiple R-squared:  0.09547,   Adjusted R-squared:  0.094
F-statistic:  65.3 on 7 and 4331 DF,  p-value: < 2.2e-16
```

Figure 7: Regression Output India

Our Indian dataset has three variables: Salary, Job Title, and City. Salary is a numerical variable while Job Title and City are categorical variables. We first ran a multiple linear regression with salary as the response variable and Job Title and City as the predictors each of which has more than one label. For each label of these predictors, we then created dummy variables with a baseline label as Data Analyst (for Job Title) and Bangalore (for City).

From our regression, we see that the Data Scientist average salaries are higher by $4,138 and Data Engineers are higher by $3,123 than Data Analyst average salaries keeping all other predictors constant. We also see that Pune has $1,207 lower average salary than Bangalore, and Mumbai has $912 lower average salary than Bangalore keeping all other predictors constant. This implies that if a U.S. company must outsource from India, then Pune would be the most cost-effective option. The Adjusted R-Squared for the regression model was 9.4%. Our model diagnostic indicated non-normality of the residuals, so we used a log-linear model. Lastly, to measure the accuracy and performance of our models, we used k-fold cross validation method

(k=3) for both linear and log-linear models. The R-Squared values in the regression models and their cross-validation were similar indicating no significant overfitting.

Lastly, we ran a linear regression model using Salary as the response variable and Country (U.S. and India) as the predictor. The regression output is given **Figure 8**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9774.6      214.8   45.51   <2e-16 ***
CountryUSA   89268.6      505.5  176.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14150 on 5293 degrees of freedom
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8549
F-statistic: 3.119e+04 on 1 and 5293 DF,  p-value: < 2.2e-16
```

Figure 8: Regression Output Both

From the above result we see that the U.S. Data Science professionals make $89,269 on average higher than Indian Data Science professionals. The k-fold cross validation suggested a similar value for R-Squared indicating no significant overfitting.

Unexpected Problems, Challenges, and Interesting Findings

We faced many obstacles throughout the life of the project but we as a team, made all attempts to overcome them with an eye on the analysis of the results and the answer to our hypothesis. First, we were unable to procure many datasets which would have given us a better ability to compare the U.S. data with the Indian data. For example, Indian datasets had less predictors, however, it did have three predictors in common with the U.S. datasets which made joining possible and effective for comparative analysis and validation. Second, salaries may depend on other factors such as experience and degree-earned which we were unable to obtain.

Some of the interesting findings we discovered were that U.S. schools generally pay higher average salaries than non-profits, private, and public companies. Another finding was that Huge U.S. companies (10,000+ employees) pay significantly higher salaries on average than all other sizes. Additionally, based upon the average salary, outsourcing does align with our hypothesis that investing in Data Science professionals in India is more cost beneficial for U.S. companies. One of our key findings was that Data Scientists are paid higher average salaries in both countries compared to other related Data Science professionals. The average salaries across five major Indian cities are similar except Bangalore which pays the highest average salary.

Unfinished Business

Part of our initial plan was to analyze and predict the current supply and demand trends for data science professionals in the U.S., but we were unable to obtain relevant datasets to use in this project. Additionally, we wanted to use multiple, advanced statistical approaches such as Random Forest and Regression Tree comparisons for a more comprehensive analysis. Due to time constraints, we would pursue this as a future endeavor.

Conclusion

From our data and analysis, U.S. companies could outsource Data Science professional jobs to India. Based upon our research, outsourcing simply boils down to economics which can be extremely cost prohibitive due to factors such as:

- Hiring more employees
- An increase in overhead
- A decrease in productivity
- Lack of efficiency due to limited resources
- A loss of focus on company's products, services, and core competencies

A primary goal of the board of directors in any company is to maximize shareholder's wealth. We discovered in our results that Indian Data Science professionals make up to $89,269 less than their U.S. counterparts. To be more specific, if a U.S. company must outsource to India, then Pune would be the most cost-effective option. More and more companies are outsourcing data analytics jobs to India because "with its skilled English-speaking workforce and salaries up to 80 per cent lower than those in developed countries, India has captured a dominant share of the international outsourcing market" (Roy, 2015).

On the contrary, as suggested by A. Gupta, the concern is that the availability of Data Science professionals in India decreased by 125% between 2018 and 2021 (Gupta, 2022). This raises an important question to our analysis in that the U.S. may choose to outsource jobs to India but are there enough Data Science professionals in India in the talent pool to meet the demand? This is a big opportunity for U.S. companies to explore outsourcing and nearshoring (for example, Brazil, Mexico, Canada, etc.).

Appendix A:

Acronym	Abbreviation Details
1 Lakh	100,000
ANOVA	Analysis of Variance
ConOps	Concept of Operations
DA	Data Analyst
DE	Data Engineer
DS	Data Scientist
EDA	Exploratory Data Analysis
HL	Higher Limit
INR	Indian Rupees
kNN	k-Nearest Neighbor
LL	Lower limit
ML	Machine Learning
NPT	Non-Parametric Test
U.S.	United States
USD	U.S. Dollar

Appendix B:

Company Size Classification		
VS	Very Small	1-50 employees
S	Small	51-200 employees
M	Medium	201-500 employees
B	Big	501-1000 employees
L	Large	1001-5000 employees
XL	Extra Large	5001-10000 employees
H	Huge	10001 and more employees

References

Bhorayal, R. (2022, June 20). *Analytics India Industry Study 2022.* Retrieved from AIM: https://analyticsindiamag.com/analytics-india-industry-study-2022/

Columbus, L. (2018, January 29). *Data Scientist Is the Best Job In America According Glassdoor's 2018 Rankings*. Retrieved from Forbes.com: https://www.forbes.com/sites/louiscolumbus/2018/01/29/data-scientist-is-the-best-job-in-america-according-glassdoors-2018-rankings/?sh=75dbd3da5535

Difference between Data Scientist and Data Engineer. (2022, April 26). Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/difference-between-data-scientist-and-data-engineer/

Drezner, D. W. (2004). The outsourcing bogeyman. *Foreign Affairs, 83(3)*, 22-34.

Gowda, O. (2022, October 01). *Data Scientist Salary*. Retrieved from Kaggle: https://www.kaggle.com/datasets/omkargowda/data-scientist-salary

Gupta, A. (2022). Career in Data Science. *Business World (India)*. Retrieved from Career In Data Science. Business World (India).

Pant, M. (2022 , October 01). *Data Science Jobs in India*. Retrieved from Kaggle: https://www.kaggle.com/datasets/madhurpant/data-science-jobs-in-india. Accessed October 1, 2022.

Prithviraj. (2022, October 01). *Data-Scientist-Salaries*. Retrieved from Kaggle: https://www.kaggle.com/datasets/madhurpant/data-science-jobs-in-india

Regions of the US. (2022, November 15). Retrieved from Infoplease: https://www.infoplease.com/us/states/regions-of-the-us

Roy, A. (2015, April 9). *Pros And Cons of Outsourcing to India*. Retrieved from LinkedIn: https://www.linkedin.com/pulse/pros-cons-outsourcing-india-avijit-roy

Thomas H. Davenport, D. P. (2012, October). *Data Scientist: The Sexiest Job of the 21st Century*. Retrieved from Harvard Business Review: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century