Examination of Salaries and Available Jobs for Data Analysts Globally and Nationally 2020 - 2023

Daniel Cucinotta

Western Governors University

Table of Contents

A. Project Highlights

Research Question.

To determine if there are any significant variances and/or associations between job location (globally and nationally), work setting (remote and in-person), experience level (entry level, mid-level, and senior), and company size, (small, medium, and large), by the amount of available jobs offered as well as the potential correlation with salaries for data analysts 2020 - 2023.

Project Scope.

The scope of the project was to clean, explore, and perform calculations to uncover any potential underlying patterns and/or anomalies with Jupyter Notebook using Python. A .csv file was created with the final (merged, cleaned, and standardized) dataframe/dataset from the Jupyter Notebook file and was used as the data source for all of Tableau's functionality and features. Additionally, many informative graphics regarding company size, location, experience level, work setting, and job count as well as the potential correlation of salaries were also generated in Tableau.

Solution Overview - Methodology.

The waterfall methodology was executed for a linear streamlined approach.

- Requirements – 5 free public use datasets were obtained from kaggle.com.
- Design - Python code was used with pandas dataframes in a Jupyter Notebook to clean and standardize the dataset to best suit the project as well as perform statistical calculations. Tableau was utilized to perform additional calculations and produce visualizations to convey the data.

- Implementation – R^2 and p-values were analyzed for overall accuracy/viability of the data and models as well as the statistical significance.
- Testing – All procedures for creating formulated calculations and statistical representations were continuously reviewed as well as cross-referenced manually for confirmation to ensure that the observed calculations and representations are accurate for determining if the information gleaned from the data and models is substantive.
- Maintenance – No additional maintenance is required for the project after completion.

Solution Overview - Tools.

Jupyter Notebook (Python3 kernel) was used conjointly with Pandas dataframes to clean, organize, manipulate, and examine data. Jupyter Notebook allowed for a comprehensive and intuitively designed execution of data manipulation/analyzing, which also includes markdown text blocks for comments, conclusions, etc., as well as for simply conveying the step-by-step processes. Additionally, job z-scores were calculated with Python code using the SciPy library function stats.zscore as well as a text file was created to import the data into Tableau for analysis.

Calculations and visualizations were performed with Tableau. An important feature of Tableau is the ability to design dashboards, which can be further included in the development of a full story - data storytelling through analytic visualizations, summaries, and conclusions in which a full report story will be published online.

B. Project Execution

Project Plan.

The plan for the project was to explore the data as well as to present distinguishable and useful information through graphical portrayals of numerous different calculations/permutations with varying combinations of attributes.

Project Planning Methodology.

The waterfall methodology was executed for a linear streamlined approach.

- Requirements – 5 free public use datasets were obtained from kaggle.com.
- Design - Python code was used with pandas dataframes in a Jupyter Notebook to clean and standardize the dataset to best suit the project as well as perform statistical calculations. Tableau was utilized to perform additional calculations and produce visualizations to convey the data.
- Implementation – R^2 and p-values were analyzed for overall accuracy/viability of the data and models as well as the statistical significance.
- Testing – All procedures for creating formulated calculations and statistical representations were continuously reviewed as well as cross-referenced manually for confirmation to ensure that the observed calculations and representations are accurate for determining if the information gleaned from the data and models is substantive.
- Maintenance – No additional maintenance is required for the project after completion.

Project Timeline and Milestones.

There are several key milestones in the timeline of the project:

- Milestone 1: May 13th to May 19th –
    a) Merge, explore, clean, execute calculations and finalize the 5 datasets into a single standardized dataframe as well as create a .csv file with Python code in a Jupyter Notebook file to import into Tableau.
        - Duration – 1 week.
- Milestone 2: May 20th to May 26th –
    a) Attributes as well as data calculations such as z-scores, p-values, and r^2 will be explored and used to create visualizations for conveying important concepts and highlights in Tableau.
        - Duration – 1week.
- Milestone 3: May 27th to June 2nd –
    a) Compose a comprehensive descriptive report chronicling all aspects of every stage as well as highlighting insights to illustrate the benefits.
        - Duration – 7 days.

C. Data Collection Process

Data Selection and Collection.

The data was collected by downloading 5 datasets (.csv files) from Kaggle.com:

1) https://www.kaggle.com/datasets/beridzeg45/data-analyst-salaries-in-the-usa/data
2) https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data/data
3) https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries
4) https://www.kaggle.com/datasets/sadeghhoushyar/data-scientist-jobs-in-canada-indeedcom
5) https://www.kaggle.com/datasets/devario/uk-data-science-jobs-dataset

Data Collection Issues and Handling.

The quality of the datasets were generally good because they were provided from kaggle.com, which is a trustworthy and reliable source. There were inconsistencies among all 5 datasets that needed standardization for merging. Year/date attributes were not in all 5 of the datasets, which prevented discerning any potential timeline trends, and resulting in no forecasting or predictive analysis. It was necessary to rename several columns to standardize, which allowed for seamless merging of the datasets.

Dataset 1: (data-analyst-salaries-in-the-usa):

Data analyst jobs were specific to the project, which resulted in the removal of 7 records from the dataset.

1 unnecessary column (‘Company Name’) was removed from the dataset for merging.

The dataset had missing data and insufficient standardization of attributes, which needed to be remedied.

There were a significant amount of NaN values contained within various attributes. 105 NaN values for the Min Salary and/or Max Salary columns that were removed. The Min/Max Salary columns were averaged for all rows in the dataset, which simplified the output values for a ‘salary’ column. The Min Salary and Max Salary columns were subsequently removed.

The values for salary period (e.g. hourly, weekly, monthly) were calculated/converted to annual for universal comparisons.

Though there were no dates within the dataset, it was published in 2023, therefore, the year 2022 was inserted into the year attribute for all records in the dataset.

Dataset 2: (jobs-in-data):

Data analyst jobs were specific to the project, which resulted in the removal of 7,742 records from the dataset.

4 unneccessary columns (‘job_category’, ‘salary_currency’, ‘salary’, and ‘company_location’) were removed from the dataset for merging.

Dataset 3: (data-science-job-salaries):

Data analyst jobs were specific to the project, which resulted in the removal of 488 records from the dataset.

4 unneccessary columns (‘Unnamed: 0’, ‘salary’, ‘salary_currency’, and ‘company_location’) were removed from the dataset for merging.

Converted experience_level (‘MI’ - ‘Mid-level’, ‘EN’ - ‘Entry-level’, ‘EX’ - ‘Executive’, ‘SE’ - ‘Senior’).

Converted employee_residence (‘HN’ -’Honduras’, ‘US’ - ‘United States’, ‘PK’ - ‘Pakistan’, ‘IN’ - ‘India’, ‘FR’ - ‘France’, ‘NG’ - ‘Nigeria’, ‘BG’ - ‘Bulgaria’, ‘GR’ - ‘Greece’, ‘HU’ - ‘Hungary’, ‘GB’ - ‘United Kingdom’, ‘ES’ - ‘Spain’, ‘KE’ - ‘Kenya’, ‘CA’ - ‘Canada’, ‘DE’ - ‘Germany’, ‘LU’ - ‘Luxembourg’) attribute values for merging.

Converted work_setting (‘0’ - ‘In-person’, ‘50’ - ‘Hybrid’, ‘100’ - ‘Remote’) attribute values for merging.

Dataset 4: (data-scientist-jobs-in-canada):

Data analyst jobs were specific to the project, which resulted in the removal of 1,733 records from the dataset.

3 unneccessary columns (‘Company’, ‘Summary’, ‘JobUrl’) were removed from the dataset for merging.

Though there were no dates within the dataset and it was published in 2020, therefore, the year 2020 was inserted into the year attribute for all records in the dataset because all of the attribute values were less than ‘30 days’.

The min_salary and max_salary columns were averaged for all rows in the dataset, which simplified the output values for a ‘salary’ column and calculated/converted to U.S. dollars. The min_salary and max_salary columns were subsequently removed.

The values for salary (e.g. HOUR, MONTH) were calculated/converted to annual for universal comparisons.

Dataset 5: (uk-data-science-jobs-dataset):

Data analyst jobs were specific to the project, which resulted in the removal of 533 records from the dataset.

9 unneccessary columns (‘reference, ‘date_posted, ‘advertiser’, ‘location’, ‘city’, ‘salary, ‘salary_frequency, ‘salary_currency’, ‘description’) were removed from the dataset for merging.

The salary_min and salary_max columns were averaged for all rows in the dataset, which simplified the output values for a ‘salary’ column and calculated/converted to U.S. dollars. The salary_min and salary_max columns were subsequently removed.

The date column values were converted from ‘yyyy-mm-dd’ to ‘YYYY’ (4 digit) for standardizing and merging.

Data Governance Issues and Handling.

The datasets that were obtained from kaggle.com because the website was recommended and conveyed to be credible by several academic sources (e.g. WGU, Udacity). The owners of all 5 of the datasets did not state that any licensing or acknowledgements are necessary. The free for public use datasets do not contain personal identifiable information (PII), which eliminates

any potential issues regarding data governance, privacy, security, legal, compliance, and/or ethical liabilities. Data privacy and security are extremely important aspects that must be addressed to ensure the protection of people's personal data.

C.1 Advantages and Limitations of Data Set

Datasets Advantages.

The advantages of cleaning and merging all 5 of the datasets was the retention of several essential attributes such as salary, work setting, company size, experience level, and location which were necessary to produce substantive statistics and visualizations. Z-scores were calculated from the data and used to calculate the r^2 and p-values. These values were needed to potentially determine the accuracy/fit of the model as well as the statistical significance of the data.

Datasets Disadvantages.

The disadvantages of merging 5 datasets was that there were many inconsistencies for standardization. There were many NaN values that needed removal as well as incomplete data for dates/range, which eliminated the ability to forecast or execute a predictive analysis.

D. Data Extraction and Preparation

The data sources were obtained by downloading 5 (.csv) files from Kaggle.com. These datasets were appropriate to calculate and visually illustrate the variance in the amount of available jobs for data analysts (globally and nationally) as well as the correlation of salaries because they contain data regarding specific attributes that are extremely valuable for the descriptive analysis models to convey beneficial data points.

E. Data Analysis Process

E.1 Data Analysis Methods

The descriptive analytical method was utilized for 9 of the 10 models to represent calculations, key data features, and beneficial insights through a series of visualizations. This method was chosen for the ability to manually seek and discern trends, as well as, there were no consistent dates/range in the dataset to perform a predictive analysis. The implementation of 1 model required a supervised classification & logistic regression approach, which helped determine accuracy and statistical significance of the data/model.

E.2 Advantages and Limitations of Tools and Techniques

Jupyter Notebook files are well organized for representing and conveying step-by-step processes with individual coding cells as well as markdown blocks (e.g. headings, descriptions) with the ability to view and manipulate the dataset using Pandas dataframes. Python is a versatile language that has many libraries and functions available, in which the z-scores were calculated (stats.zscore from the SciPy library).

Applied calculations such as averages, Standard deviation, variance, percent difference, percent of total, ratios, and the F1-score (precision and recall) were computed in Tableau with calculated fields containing formulas of dataset attributes as well as with the use of built-in functions. These methods and metrics were used to evaluate the statistical significance by contrasting specified attributes of the standardized dataset interchangeably to reveal a clear understanding by presenting an array of detailed visualizations. The visualizations can be presented standalone, in a dashboard, or series of dashboards, as an illustrated analytical story.

The constituents of the complete analyses for this project are 18 visualizations, which are all vital components for the 8 dashboards, that consists of 9 descriptive analysis models, 1

supervised classification/logistic regression model, and 1 table that comprise the analytical story designed in Tableau. The remaining 7 graphical representations are elements on the dashboards for additional and/or corroborative data.

E.3 Application of Analytical Methods

The descriptive analytical methods as well as the supervised classification/logistic regression model are appropriate for the hypothesis of the project through the examination of various combinations of attributes with finite measurements that were elaborately conveyed graphically for comparison/further evaluation of r^2 and p-values for data analyst jobs, to potentially determine accuracy/statistical significance of the data and models. The z-scores were calculated with Python code and the SciPy library function stats.zscore in a Jupyter Notebook file.

The F1-score is a built in calculation function within Tableau which can be used in conjunction with numeric measure attributes that have an imbalanced number of data points for a more equalized distribution of values apropos predictive apportionment, which can potentially improve the accuracy of the model. The F1-score is a harmonic mean of precision (positive predictive value) and recall (sensitivity) values that are based upon the relevenace of an attribute's data by repetitive formulaic permutations and calculations of true positives, true negatives, false positives, and false negatives.

The project is devised of 9 models in which descriptive analyses were performed and 1 supervised classification & logistic regression model of a scatter plot that utilizes Tableau's functionality to calculate/display the r^2 as well as p-values that are derived from the z-scores. Additionally, there is 1 table that displays the job count, average salary, ratio (jobs/salary), and rank (ratio) for the top 5 States with the most jobs.

F. Data Analysis Results

F.1 Statistical Significance

Global Models 1 - 4:

1.) Model 1 & 1A – Descriptive analysis.

Algorithm – Global average salary histogram (1) and global average salary histogram by alternating work setting (1A), in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – Job count by average salary range bins.

Benchmark – Average salary bins increase/decrease more than 50,000 jobs.

Relevance – The average salary values and differences by work setting can be very beneficial for individuals willing to relocate or work remotely.

Conclusion - The number of remote jobs has a generally similar correlative Bell Curve to the total global average. Remote jobs decrease from every in-person salary range, except for the $95,000 average salary bin. The 4 lowest salary bins ($51,000 - $95,000) have the largest fluctuations overall. Nearly 57,000 jobs were transferred from all salary ranges to the $95,000 bin ($4^{th}$ lowest) for remote jobs, with the largest decreases from the 3 lowest salary bins.

2.) Model 2 – Descriptive analysis.

Algorithm – Circle view chart of the top 10 countries with the highest average salary by rank.

Metric – Average salary values and rank.

Benchmark – If global salaries have a greater/less than range of $50,000:

Relevance – The average salary values can be very beneficial for individuals willing to relocate or work remotely.

Conclusion – Saudi Arabia has an average salary of $179,998 (ranked #1) and U.K. (ranked #10) has an average salary of $74,817. The variance is greater than $100,000.

3.) Model 3 – Descriptive analysis.

Algorithm – Box-and-whisker plot of the top 5 countries by highest job count.

Metric – Measure the number of standard deviations and range (upper/lower whiskers, upper/lower hinges, and median) for job count between the top 5 countries.

Benchmark – Greater than 1 standard deviation from the global average salary mean.

Relevance – The number of available jobs is an indicator of the magnitude of potential opportunities.

Conclusion – The U.S. job count is 6.874 standard deviations higher than the global average salary mean and is 4.2 standard deviations above the lower whisker (Spain: 2.678).

The U.S. is 2.6 standard deviations greater than the median (Canada: 4.273).

The upper hinge (U.K.: 5.475) is still greater than 1 standard deviation below the upper whisker (U.S.) and the lower hinge (Germany: 2.809) is approximately 4 standard deviations below the U.S. in the number of jobs.

4.) Model 4 – Descriptive analysis.

Algorithm – Packed bubble chart displaying average salary, rank, and job count for the top 5 countries with the most jobs, by alternating work setting (remote and in-person), in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – Evaluate average salaries and ranks as well as job count.

Benchmark – If the difference is greater than $10,000 between the average salaries for the top 5 ranked countries.

Relevance – Locations/work setting that have the highest job count and/or highest average salaries are advantageous for seeking general as well as specific employment opportunities.

Conclusion – Canada and the U.S. have the highest average salaries (top 5 countries with the most available jobs), which trade the number 1 & 2 ranks by work setting; Canada (#1 remote: $111,356) and the U.S. (#1 in-person: $103,489).

U.K., Germany, & Spain are consistently numbers 3, 4, & 5, respectively, for average salary.

The U.S. and Canada have approximately a $4,000 difference for remote work salaries, as U.K., Germany, and Spain are nearly $25,000 less for each country by each rank. All countries have more than a $10,000 difference by rank for in-person jobs, with a range of $60,000 (Spain: $43,021 – U.S.: $103, 489).

The U.S. (#1 - total jobs: 2,434,207) & the U.K. (#2 - total jobs: 309,637); there is a significant difference of more than 2 million more jobs in the U.S. than the U.K.

U.S. National Models 5 - 10:

5.) Model 5 – Descriptive analysis.

Algorithm – Scatter bar chart of job count and percent of total for company size (small, medium, and large) and experience level (entry level, mid-level, and senior) by work setting (remote and in-person) for the U.S., in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – Determine which combination of attributes has the highest job count and percent of total.

Benchmark – If any specific category combinations have greater than 50% of the total number of jobs offered.

Relevance – Understanding how the number of jobs vary with the size of a company and employee experience level as well as by work setting, can aid immensely in the job seeking process for potential opportunities in the U.S.

Conclusion – In-person work for both small/mid-level as well as medium/entry level have 100% of the jobs. In-person jobs for large/senior has 51.86%.

Remote work for medium/mid-level has 74.67% of the jobs. Remote jobs for small companies are split between small/entry level (56.78%) and small/senior (43.22%).

6.) Model 6 – Descriptive analysis.

Algorithm – Scatter bar chart in which the values are represented by stars with reference lines for the values, as well as display the job count and the population variance for the top 5 States with the highest number of available jobs for entry level and mid-level positions in the U.S., in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – Analyze the variance in the number of jobs by work setting.

Benchmark – Increase/decrease of more than 5,000 jobs for a State.

Relevance – The variance in the number of jobs by work setting can be very useful for relocating to 1 of the top 5 States with the most jobs or seeking a remote job (where offered).

Conclusion – The variance for remote jobs decreased for California (-16,435) as well as Virginia (-9,343), and conversely, in-person jobs increased for Florida (+6,942).

7.) Model 7 – Descriptive analysis.

Algorithm – A horizontally split histogram of the top 5 States with the most jobs which are comprised by 4 average salary bins ($60K, $80K, $100K, $120K) per State.

Metric – Evaluate the number of jobs and the average salary ranges.

Benchmark – If any salary bins have greater than 30% of the total jobs by State and work setting.

Relevance – The difference in the number of jobs for each average salary bin range by work setting can be used for seeking an opportunity that is suitable in a specific location or remote with the highest potential for employment as well as higher average salaries.

Conclusion – All 5 States had a greater than 30% of total jobs for the in-person $80K average salary bin range: VA – 43.05%, NY – 36.21%, CA – 32.62%, and FL – 33.33%. In-person jobs for VA has 38.01% in the $100K bin and NY has 35.63% of jobs in the $60K bin.

Significant values for remote jobs are VA (95.91%) and FL (88.98%) of jobs in the $100K salary bin as well as TX (41.21%) in the $120K bin, with NY (36.56%) and TX (32.47%) in the $60K bin.

8.) Model 8 – Descriptive analysis.

Algorithm – A geographic symbol map of the U.S. for the top 5 States with the highest average salary by work setting (remote/in-person), in which the States are depicted with color from a range representing the average salaries and the values are displayed.

Metric – Evaluate average salaries for the top 5 States for potential trends.

Benchmark – A difference of greater than $20,000 for average salaries between any of the top 5 States.

Relevance – Knowing which States have significantly higher average salaries as well as the variation by work setting can be beneficial factors when seeking employment.

Conclusion – Washington D.C. ($132,393 – national high) and neighboring States (MD – $111,312, DE - $105,600) have the highest average salaries in the country for in-person jobs.

The highest average salary in the country for available remote jobs were offered in California ($107,476).

Most of the remaining average salaries for remote and in-person are approximately $93,000, except for remote jobs in NJ ($99,300).

The average salary for Washington D.C, is at least $20,000 higher than the other 4 States (remote and in-person). The average salary for remote jobs in CA ($107,476) is at least $20,000 higher than MD and DE.

9.) Model 9 – Descriptive analysis.

Algorithm – Average salary histogram for the top 5 States with the highest job count (9) and Average salary histogram for the top 5 States with the highest job count by alternating experience level (entry level, mid-level, and senior) (9A), in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – Examine average salary distributions by experience level to discern potential trends.

Benchmark – A difference of greater than $20,000 for average salaries between any of the top 5 States.

Relevance – Knowing which States have significantly higher average salaries as well as the variation by experience level can be beneficial factors when seeking employment attainable by credentials, skills, and knowledge.

Conclusion – There is a total of 2,415,159 jobs in the top 5 States with highest job count: Senior has 62.47% (1,508,881) of the jobs with most being in the higher average salary range of $80,000 - $139,000, mid-level has 31.35% (757,269) of the available jobs with an average salary

range of $66,000 - $95,000, and entry level has only 6.17% (149,049), in which the majority of the jobs have a lower average salary range of $52,000 - $95,000.

10.) Model 10 – Supervised classification & logistic regression.

Algorithm – Job z-scores by average salary for the top 5 States with the highest job count as well as alternating work setting (remote and in-person).

Metric – R^2 and P-values calculated from the job z-score with Tableau's trend line feature.

Benchmark – Whether p-value(s) are greater/less than 0.050 and/or the r^2 value(s) are greater/less than 0.8.

Relevance – The z-score informs of the standard deviation from the mean. The p-value can assist with determining if the data has significance or is not correlated as well as discern statistical norms and variances, which may highlight any trends/anomalies for the top 5 States with the most number of data analyst jobs.

Conclusion – The p-value for remote jobs is 0.096 and the p-value for in-person jobs is 0.111. The high values (greater than 0.05) would normally indicate that there is no correlation between the data points represented in the dataset, as well as, that these values are not statistically significant and consequently reject the null hypothesis.

The r^2 value for in-person jobs is 0.303 and the r^2 value for remote jobs is 0.658. Unfortunately, both values infer that the model only has a 30% (in-person) and 66% (remote) accurate fit.

Although the r^2 value does not provide any further assistance with understanding the model's statistical accuracy and the p-values are potentially "inconclusive," it is necessary to examine the job z-scores further for potential validity of the data/model.

The z-scores for the top 5 States with the highest number of jobs is #1 - California (4.558), #2 - New York (2.676), #3 - Texas (1.924), #4 - Florida (1.321), and #5 - Virginia (1.020). Illinois (0.945) is ranked as the 6th State with the most number of jobs, and all States except for the top 6 are within the range of approximately -0.700 to 0.700. All of the top 5 States are greater than 1.000 and are significant outliers, in which the model deems to have no correlation. Further analysis is required for identifying factors which potentially affect the causality of the outlier score values.

#1. California is the 3rd largest State in the U.S. and has many metropolitan areas, in which the majority of in-person jobs are located and remote jobs are offered, as well as Silicon Valley, which is the location of many tech hubs and specialized innovations. The job z-score of 4.558 can be legitimate because of the demographics.

#2. New York is not a very large State, though it has several urban regions. New York City is the financial capital of the United States/entire world because of the financial markets/Wall St., as well as insurance/credit/loan companies, and government agents/workers, etc. As a result of urban sprawl, inflation, pandemic and other contributing factors, many jobs that were in New York City have relocated to the greater metropolitan areas of northern New Jersey and southern Connecticut. Data analysts are in high demand with various types of financial institutions, commerce businesses/platforms, services rendered, etc., which can seemingly explain the 2.676 job z-score, when compared to the total population of the State.

#3. Texas is the 2nd largest State and also has many highly populated cities. The economy of Texas compared to California and New York is on a completely different scale: the cost of living as well as salaries/pay rates are generally much higher in California and New York than the majority of Texas. The job z-score of 1.924 is possible validating indicator.

#4. Florida is a mid-size State (#22 nationally) with several highly populated areas. The average salary for data analysts in Florida is $84,444 (# 16 nationally). The job z-score of 1.321 may be justified because companies/corporations seek to exploit labor markets that have below average/average salaries, which increase profit margins.

#5. Virginia is not a large State with quite a few urban areas and has a 1.020 job z-score. The value must largely be representative of politicians and government officials that work in or around Washington D.C as well as the various military bases and facilities.

The model does not understand environmental/human factors or demographics (without data to specify/instruct), and therefore calculated statistics with several outliers. After considerable examination, the calculated outliers have an appearance of being valid due to the extraneous factors stated prior. The outlier values presumably had a consequential impact on the calculation of the r^2 and p-values.

11.) Table 1 – Tabular Presentation.

Algorithm – Ratio calculation of job count/highest average salaries for the top 5 States with the most number of jobs. The table displays State, job count, average salary, ratio (job count/salary) , and rank (ratio) as well as alternates between work setting (remote and in-person), in which an F1-score calculation was applied to the job count attribute for distribution accuracy.

Metric – The higher the ratio for a State indicates a greater potential to attain a better than average salary employment opportunity, possibly quicker and/or with less efforts.

Benchmark – A greater/less than difference of 0.10 between any of the top 5 States with the highest job count.

Relevance – The ratio of highest number of jobs/highest average salary is an important basic calculation that is informative for the probability scale of attaining employment with less effort

and/or a higher average salary. A higher ratio value indicates higher number of potential jobs and/or higher average salary, both of which are desirable when job seeking.

Conclusion – The ratio of jobs/average salary is ranked #1 California (0.537) for in-person work, followed by #2 Texas (0.330), in which there is a significant margin of greater than 0.20. New York is ranked #3 (0.307), #4 Florida (0.298), and #5 Virginia (0.216).

New York (0.205) ranked #1 for remote jobs, and was trailed by #2 California (0.186), with having only a slight difference. Both are more than double #3 Texas (0.080), which is approximately double the values of Virginia (0.042) and Florida (0.033).

California has the best overall ratio average: #1 - (0.362), #2 - New York (0.256), #3 - Texas (0.205), #4 - Florida (0.166), and #5 - Virginia (0.129). There is more than a 0.10 difference between California and New York, then Texas, Florida, and Virginia each decrease by approximately 0.05.

F.2 Practical Significance

The practical significance can be assessed by the magnitude of clearly revealed informative indicators. Exploration of various attribute combinations and calculations provided succinct information to aid in a decision-making process. For example: A recent graduate is seeking to relocate to a State in which the salary is higher than average and has a high number of available jobs. The difference of $20,000 - $100,000 a year salary could change the current lifestyle and future trajectory of financial security, freedom, and success of an individual. This data can potentially help narrow job search criteria to specific States. Seeking employment in a State that has a higher than average salary as well as a higher number of available data analyst jobs may be a safe choice for attaining a greater than average job opportunity (prospect of attaining a job quicker and/or by exerting less effort with a higher than average salary).

F.3 Overall Success

Visually representing discernible data points/patterns and elaborately illustrating the statistical significance with basic explanations for general understanding to provide others with the provided exploration of the vast data can be very beneficial.

The p-values were analyzed to discern if any significant variances between remote and in-person available jobs exists. The percent difference values provided information of the variance for remote and in-person available jobs. Though the z-scores appeared skewed with outliers, the values were potentially substantiated, because data analytics is a vocation requiring advanced media technological skills apropos high volumes of data, which are generally sought in metropolitan regions as well as areas substantially incorporated with research and development and/or innovative productivity entities. Ratios had an important role for simply conveying the data/statistics regarding States that had the highest job counts by highest average salaries. The ratio calculation provides insight for understanding the range of probability of attaining a job; the higher the ratio value, the greater potential to attain employment in less time and/or effort with a higher than average salary.

The many analyses and comprehensive conclusions were all general successes because an observer can get a variety of extremely beneficial information from the plethora of models presented. There are several models that highlight and display global data statistics for a macro perspective (global) with ease visually as well as several other models that attenuate the purview with a meso (national) to a micro (State) for comparisons.

G. Conclusion

G.1 Summary of Conclusions

The anticipated goal of the project was to determine if there is any correlation between the number of available data analyst jobs (globally and nationally) as well as the potential correlation of salaries, which was showcased by numerous graphics and highlighted specific beneficial criteria for decision-making. The dataset provided ample data points for the analysis, though there was a lack of dates/range, which prohibited the ability to perform any forecast or predictive analyses.

Global Analysis:

The amount of jobs offered in the United States is very much greater than all other countries as well as the average salary range, which is competitive only to Canada (remote work only). These basic facts dictated the project to continue the analysis and report focused upon only the United States.

The remaining of top 5 countries (excluding U.S.) with the most available jobs were Canada, United Kingdom, Germany, and Spain. Canada is a neighbor to the U.S. in North America, which can explain the higher average salary and a higher than average job count (4.273 standard deviations above the mean). U.K., Germany, and Spain are all located in Europe; United Kingdom is 5.475 standard deviations above the mean, which is the highest globally (excuding the U.S.). U.K. has more than 1 standard deviation greater amount of jobs than Canada. Germany and Spain may have a fair amount of jobs available, naturally, though, it can be assumed that there is an association with the amount of jobs available in the U.K.

It must be noted that although China is included in the dataset and had the 2nd highest average alary in the world, no other representative statistics were portrayed, which indicates that

China's dictatorial guidelines regarding data/information sharing is highlighted, because China has nearly 1/5th of the global population and there are vast technological innovations, commerce, and industries, which all require data analytics to remain competitive globally, yet, the number of available jobs had no comparison to the top 5 States with the highest job count.

National Analysis:

4 of the top 6 States with the highest average salaries in the country are Washington D.C. (#1: $118,973) as well as 3 adjacent States (#2: Maryland - $108,036, #2: Delaware - $98,950, and #6: Virginia – $93,001). California ($98,456) and New Jersey ($93,258) are ranked #4 and # 5 respectively. Washington D.C. ($132,393) has the highest average salary for in-person jobs and California ($107,476) for remote work.

The nations capital region seemingly indicates a correspondence of jobs for military/government officials and politicians. Data analytics for the military and government is absolutely necessary for understanding statistics regarding the country's demographics as well as important statistics apropos budget/funding, law enforcement, and research/development, etc.

In-person work for both small/mid-level as well as medium/entry level have 100% of the jobs for company size. In-person jobs for large/senior has 51.86% for large companies.

Remote work for medium/mid-level has 74.67% of the medium company size jobs (149,396 - #1 in U.S.). Remote jobs for small companies are split between small/entry level (56.78%) and small/senior (43.22%).

G.2 Effective Storytelling

1) A global average salary histogram (1) and global average salary histogram by alternating work setting (1A), in which an F1-score calculation was applied to the job count attribute for distribution accuracy. The average salary values and differences by work setting can

be very beneficial for individuals willing to relocate or work remotely. The model highlighted that the majority of in-person jobs have an average salary range of $60K - $110K. Though there are significantly less remote jobs offered, which decreased from all salary ranges except for the $95K bin, which increased by approximately 57,000 jobs.

2) A circle view chart of the top 10 countries with the highest average salary by rank, in which the average salary values can be very beneficial for individuals willing to relocate or work remotely.

3) Box-and-whisker plot of the top 5 countries by highest job count. The number of available jobs is an indicator of the magnitude of potential opportunities. The U.S. job count is 6.874 standard deviations higher than the global average salary mean and is 4.2 standard deviations above the lower whisker (Spain: 2.678).

4) Packed bubble chart displaying average salary, rank, and job count for the top 5 countries with the most jobs, by alternating work setting (remote and in-person), in which an F1-score calculation was applied to the job count attribute for distribution accuracy. Locations/work setting that have the highest job count and/or highest average salaries are advantageous for seeking general as well as specific employment opportunities. The U.S. has a significantly higher job count as well as average salary than the rest of the world, except for remote work salaries for Canada.

5) Scatter bar chart of job count and percent of total for company size (small, medium, and large) and experience level (entry level, mid-level, and senior) by work setting (remote and in-person) for the U.S.**,** in which an F1-score calculation was applied to the job count attribute for distribution accuracy. Understanding how the number of jobs vary with the size of a company and employee experience level as well as by work setting, can aid immensely in the job seeking

process for potential opportunities in the U.S. In-person work for both small/mid-level as well as medium/entry level have 100% of the jobs.

6) Scatter bar chart in which the values are represented by stars with reference lines for the values, as well as display the job count and the population variance for the top 5 States with the highest number of available jobs for entry level and mid-level positions in the U.S., in which an F1-score calculation was applied to the job count attribute for distribution accuracy. The variance in the number of jobs by work setting can be very useful for relocating to 1 of the top 5 States with the most jobs or seeking a remote job (where offered). The variance for remote jobs decreased for California (-16,435) as well as Virginia (-9,343), and conversely, in-person jobs increased for Florida (+6,942).

7) A horizontally split histogram of the top 5 States with the most jobs which are comprised by 4 average salary bins ($60K, $80K, $100K, $120K) per State. The difference in the number of jobs for each average salary bin range by work setting can be used for seeking an opportunity that is suitable in a specific location or remote with the highest potential for employment as well as higher average salaries. All 5 States had a greater than 30% of total jobs for the in-person $80K average salary bin range: VA – 43.05%, NY – 36.21%, CA – 32.62%, and FL – 33.33%. Significant values for remote jobs are VA (95.91%) and FL (88.98%) of jobs in the $100K salary bin.

8) A geographic symbol map of the U.S. for the top 5 States with the highest average salary by work setting (remote/in-person), in which the States are depicted with color from a range representing the average salaries and the values are displayed. Awareness of which States have significantly higher average salaries as well as the variation by work setting can be beneficial factors when seeking employment. Washington D.C. ($132,393 – national high) and

neighboring States (MD – $111,312, DE - $105,600) have the highest average salaries in the country for in-person jobs. The highest average salary in the country for available remote jobs were offered in California ($107,476).

9) Average salary histogram for the top 5 States with the highest job count (9) and Average salary histogram for the top 5 States with the highest job count by alternating experience level (entry level, mid-level, and senior) (9A), in which an F1-score calculation was applied to the job count attribute for distribution accuracy. Knowing which States have higher average salaries as well as the variation by experience level can be beneficial factors when seeking employment attainable by credentials, skills, and knowledge. Senior has 62.47% (1,508,881) of the jobs with most being in the higher average salary range of $80,000 - $139,000, mid-level has 31.35% (757,269) of the available jobs with an average salary range of $66,000 - $95,000, and entry level has only 6.17% (149,049), in which the majority of the jobs have a lower average salary range of $52,000 - $95,000.

10) A scatter plot of the top 5 States with the highest number of available jobs showing job z-scores with a trend and an animation that alternates between remote and in-person work setting. The 5 data points are represented on the chart as small various colored circles which depict the State. The visualization displays r^2 and p-values for remote as well as in-person jobs via the trend line. Though the r^2 and p-values calculated did not provide evidence of a connection among the values, the z-scores did however indicate that there was an underlying association. The model's statistical inferrences were unable to compute human and environmental factors, such as the size of the State as well as number of metropolitan areas and locations where many innovative employees aspire to work, which was only understood after further investigation. This

graphic is simple to comprehend and it provides several useful statistics for comparative analyses.

11) A tabular presentation displaying the top 5 States with the highest job count, average salary, ratio (job count/average salary), and rank (ratio). The elementary table provides extremely beneficial information that can be factored into a decision-making process for potentially acquiring a preferrable job/salary.

G.3 Recommended Courses of Action

There are many potential ways to utilize the information that has been presented by the data and through the visualizations that were created. Four example scenarios will be provided:

Example 1 (Global): A data analyst that is living in Europe where the number of jobs as well as the salary range is below average for a country in comparison, may endeavor a job in another EU (European Union) country that offers better prospective opportunities. There are 50 times more in-person jobs available in the U.K. and the average salary is at least $20K higher as well. Relocating to the U.K. may be very beneficial for a data analyst for in-person work for a resident of the EU.

Example 2 (Global): A data analyst that resides in any country worldwide may endeavor a better career opportunity, yet, the majority of the lucrative options are located in other, distant countries. Without the ability to relocate (financial and/or other life constraints) to one of the countries that has the best opportunities available, remote work can be sought. Seeking remote work in Canada could be extremely beneficial, with adequate credentials. The highest average salary for remote work in the top 5 countries with the most jobs is Canada ($107,310).

Example 3 (National): A data analyst seeking the highest possible average salary, and willing to relocate, may decide that moving to a suburb, neighboring State, or short distrance

from Washington D.C. would be a very beneficial choice. It is assumed that the majority of the available jobs are government, military, politically associated positions, etc., which entails having the proper credentials and acceptable background for government classification/clearance levels.

In-person salaries and remote work salaries for jobs offered in the region are both much higher than the average salaries nationally by approximately $20,000 - $40,000 by work setting. In-person job salaries are 15% - 20% higher than remote work salaries for Washington D.C., Maryland, and Delaware. The difference in salaries is nearly equivalent to an annual salary for entry level position for many careers. This information alone can potentially help excel a motivated individual with a flexible lifestyle to wealth and success much quicker, than to seek employment in a lower income location or a remote job.

Example 4 (National): If a data analyst is willing to relocate to a State that has a significantly larger number of jobs available as well as have a higher salary than the average, moving to California, Texas or New York would be the best options.

California and Texas are massive States that has many metropolitan areas, and may need some further investigating through searches and job listing websites to narrow geographic locations. Silicon Valley is a hotspot for tech organizations and innovative concepts. If an individual possesses the credentials and skills, Silicon Valley would be an ideal location to seek residence for employment.

Conversely, New York is not a humongous State, yet, offers increased availability of jobs and a high average salary. There are several cities in New York that may offer data analyst jobs, though, the majority of the State's data analyst jobs are located in and around New York City, such as Long Island or northern New Jersey (which is just across the Hudson River from the

financial capital of the world). Wall St., international banking systems, insurance companies, government agents, etc. have corporate hubs or large satellite locations because of the volume of business, nationally and globally. Moving to New York City or a suburb would be a very wise decision because attaining a job with a higher than average salary may be possible quicker, due to the significantly higher number of available jobs.

References

No sources were cited.

Appendix A

There are 8 additional files that will be included with the submission of the report/project.

- Jupyter Notebook file (World Data Analyst Jobs and Salaries 2020 - 2023.ipynb)
- Tableau workbook file (World Data Analyst Jobs and Salaries 2020 - 2023.twbx)

- A comma-separated values (.csv) file created from Python code and imported into Tableau for analysis (World Data Analyst Jobs and Salaries 2020 - 2023.csv)
- The 5 original datasets downloaded from Kaggle.com:

1a.) U.S. dataset – (Data Analyst Salaries in The USA.csv)

1b.) https://www.kaggle.com/datasets/beridzeg45/data-analyst-salaries-in-the-usa/data

2a.) Global dataset – (jobs_in_data.csv)

2b.) https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data/data

3a.) Global dataset – (ds_salaries.csv)

3b.) https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

4a.) Canada dataset – (Job_list_Canada.csv)

4b.) https://www.kaggle.com/datasets/sadeghhoushyar/data-scientist-jobs-in-canada-indeedcom

5a.) U.K. dataset – (deduped-jobs.csv)

5b.) https://www.kaggle.com/datasets/devario/uk-data-science-jobs-dataset