Logo, company name

Description automatically generated

**SALARY & JOB SATISFACTION PREDICTION**

**ISM 6354**

**Programming for Data Analytics - R**

Professor Andy Chen

Chisomo Kamlopa, Shely Nunez, Charan Singh, Ngoc Oanh Nguyen​

Winter, 2024

**Table of Contents**

[**1. Business Understanding 2**](#_heading=h.gjdgxs)

[**Motivation 2**](#_heading=h.30j0zll)

[**Stakeholders 2**](#_heading=h.1fob9te)

[- Recruiters 2](#_heading=h.3znysh7)

[- Finance Department & Executives 2](#_heading=h.2et92p0)

[- Job Seekers 3](#_heading=h.tyjcwt)

[**2. Data Analysis 3**](#_heading=h.3dy6vkm)

[**Data understanding 3**](#_heading=h.1t3h5sf)

[**Data preparation 3**](#_heading=h.4d34og8)

[**Exploratory Data Analysis (EDA) 7**](#_heading=h.2s8eyo1)

[- Relationship between Age and Average Salary 7](#_heading=h.17dp8vu)

[- Average Salary distribution by Revenue for each Company Size 8](#_heading=h.3rdcrjn)

[- Average Salary distribution for Rating (Job Satisfaction) 8](#_heading=h.26in1rg)

[- Average Salary distribution for each Job Title 9](#_heading=h.lnxbz9)

[- Rating (Job Satisfaction) and Company Size 9](#_heading=h.35nkun2)

[- Rating (Job Satisfaction) and Job Title 10](#_heading=h.1ksv4uv)

[- Top States Salary x Skills Required 10](#_heading=h.44sinio)

[**3. Main Analysis 10**](#_heading=h.2jxsxqh)

[**Regression Models for Salary Prediction 10**](#_heading=h.z337ya)

[**Classification Models for Job Satisfaction Prediction 11**](#_heading=h.3j2qqm3)

[**N-Way analysis of variance (ANOVA) with Interactions 12**](#_heading=h.1y810tw)

[**4. Recommendations 12**](#_heading=h.4i7ojhp)

[**Recommendations for Stakeholders: 12**](#_heading=h.2xcytpi)

[Recruiters 12](#_heading=h.1ci93xb)

[Finance Department & Executives: 12](#_heading=h.3whwml4)

[Job Seekers: 13](#_heading=h.2bn6wsx)

[**Potential Further Actions: 13**](#_heading=h.qsh70q)

[Expand the Dataset: 13](#_heading=h.3as4poj)

[Incorporate Additional Features: 13](#_heading=h.1pxezwc)

[Develop a Salary Prediction Tool: 13](#_heading=h.49x2ik5)

[Conduct Regular Market Analysis: 13](#_heading=h.2p2csry)

[Offer Training and Workshops: 13](#_heading=h.147n2zr)

[Explore the Impact of Remote Work: 14](#_heading=h.3o7alnk)

[**5. Faith and Ethics Implications 14**](#_heading=h.23ckvvd)

[**Ethical Implications of the Project 14**](#_heading=h.ihv636)

[**Relationship between the Christian Faith and Ethical Practices Relevant to the Project: 14**](#_heading=h.32hioqz)

# Business Understanding

## Motivation

The understanding of salary-related aspects among data analysts and data scientists allows for exploration into dimensions influencing compensation structures. The business problems we decided to explore shed light on crucial elements that impact salary dynamics within the industry. For aspiring data analysts and scientists, this analysis serves as a guide, guiding career decisions by unraveling salary expectations. It empowers individuals to set realistic expectations, plan for the future, and negotiate their worth based on industry patterns.

First, we investigate the intricate relationship between age and average salary, exploring the patterns that inform fair and competitive compensation strategies throughout professionals' careers. Secondly, examining average salary distribution by revenue for different company sizes allows us to understand the profound impact of financial resources on compensation, which is crucial for optimizing strategies and ensuring both competitiveness and financial sustainability. Thirdly, exploring the relationship between job satisfaction ratings and average salaries investigates the foundational elements of employee retention and workplace well-being, guiding decisions on compensation adjustments and engagement initiatives. Lastly, the analysis of job satisfaction ratings about company size uncovers insights into how organizational scale influences employee contentment, allowing executives and HR teams to tailor policies that enhance satisfaction across varying company sizes.

This exploration can serve as a toolkit for individuals and organizations alike. It empowers individuals to make informed career decisions, guides companies in creating fair and competitive workplaces, and contributes to an industry where both professionals and employers thrive. By delving into these critical dimensions, we're not just analyzing data; we're shaping the future of the data analyst and data scientist profession.

## Stakeholders

In recent years, there has been an escalating demand for skilled data analysts and data scientists which has prompted organizations to deal with the complexities of designing competitive and sustainable salary structures. There are stakeholders central to this business problem, reaching the specifics of roles and interests in the pursuit of effective compensation strategies.

### Recruiters

Firstly, people who are recruiting for companies may be interested in this dataset. These individuals are responsible for the hiring process, recruitment teams are responsible for all talent acquisition. Their interest lies in gaining insights into salary expectations for effective negotiation during job offers, ensuring competitiveness in the job market. Recruitment teams utilize salary data to tailor job offers, attracting skilled professionals and fostering a positive recruitment experience. The recruiting team has to work closely with other professionals or teams in the organization that are responsible for monitoring industry trends and benchmarks and leverage the dataset for benchmarking salary levels against competitors. This contributes to strategic decisions on market positioning, enhancing the company's standing in the job market.

### Finance Department & Executives

Secondly, two departments that may work together and find our model useful are the Finance department and the Executive team. The finance team is typically tasked with budgeting and cost management, including salary allocation, they seek financial sustainability. Salary insights are vital for appropriate budget allocation and resource management. Analysis of the dataset assists the finance department in optimizing salary budgets, striking a balance between competitiveness and fiscal responsibility. Meanwhile, the executive team helps in shaping the overall strategic direction and has an interest in understanding how salary levels impact talent acquisition and retention. This insight is crucial for informed workforce planning. Salary insights guide strategic decisions related to workforce planning, ensuring the organization remains competitive and attractive to top talent.

### Job Seekers

Lastly, job seekers, specifically data analysts and data scientists, being essential contributors to a company's success, are inherently motivated to comprehend the market value attached to their skills and experience. This understanding not only empowers them in salary negotiations but significantly influences broader career decisions. It becomes crucial for individuals when navigating transitions, be it negotiating salary adjustments during job changes, advocating for promotions, or simply gaining insights into their overall market worth within the evolving industry. In essence, salary data becomes a guiding compass, helping these professionals make informed decisions that align with their expertise and aspirations while ensuring they are fairly recognized and compensated within the competitive job market.

The business problem of optimizing salary structures for data analysts and data scientists engages a diverse array of stakeholders. Each stakeholder group plays a unique and crucial role, contributing to the development of compensation strategies that ensure competitiveness, attract top talent and align with the organization's financial goals.

# Data Analysis

## Data understanding

The data preparation process required a comprehensive understanding of the dataset and its significance in addressing the business problem of optimizing salary structures for data analysts and data scientists. We obtained the dataset “Salary Predictions” from Kaggle. This dataset contains job postings from Glassdoor.com from 2017. This dataset can be used to analyze the current trends based on job positions, company size, and more. During our project we switched our data file to using the "eda\_data.csv" instead of "salary\_data\_cleaned.csv" for this analysis as it offers a more extensive range of variables, facilitating easier data wrangling, exploratory data analysis (EDA), and modeling. The "eda\_data.csv" dataset encompasses 33 variables and 742 rows, providing a multifaceted view of factors influencing compensation variables within the industry. Aspiring data analysts, data scientists, recruiters, and finance professionals are among the key stakeholders eager to decipher insights from this dataset. The variables encompass various aspects such as job titles, seniority, company size, revenue, age, and skills, offering a rich tapestry of information. This prelude sets the stage for a meticulous exploration into the intricate relationships between these variables and average salaries, ultimately empowering stakeholders with actionable insights for informed decision-making in the realm of compensation strategies.

## Data preparation

We used the “eda\_data.csv” dataset which contains 33 variables and 742 rows.

A screenshot of a computer

Description automatically generated

The dataset has no null values, so we didn’t eliminate any row.

A close-up of a computer code

Description automatically generated

Despite the cleanliness of the dataset, a few preparatory measures were required, with the sole tasks of extracting values and selecting usable variables.

The job\_simp variable is the simple version of Job.Title one, however, there are still a lot of “na” values in it, so we decided to create a new variable called “Title” which was directly extracted from Job.Title column to handle this issue.

A screenshot of a computer

Description automatically generated

Similarly, we also created Job.Seniority column from Job.Title though there is a “seniority” variable.

A screenshot of a computer

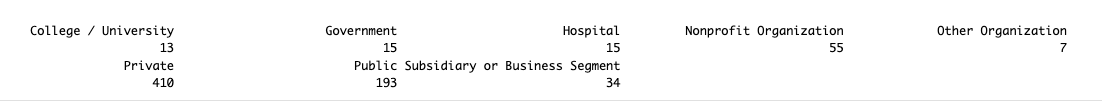
Description automatically generated

In terms of Headquarters variable, it is quite redundant with both city and state (e.g. San Antonio, CA), so we trim it to get the state only (e.g. CA).

A group of black text

Description automatically generated

We simplified the Type.of.ownership variable by gathering any values containing the “private” string into the “Private” category; the “public” string into the “Public” category; “-1”, “unknown”, “school/school district”, “private practice/firm”, or “contract” into the “Other Organization” category; and the rest was kept remained.

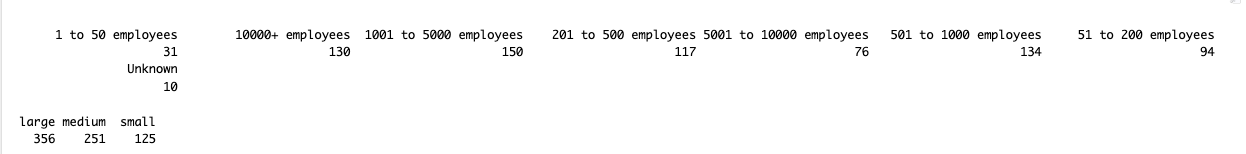


We replaced “-1” value in Industry variable to “Others”.

A screenshot of a website

Description automatically generated

For Size, first, we replaced the “-1” value with “Unknown”, then categorized them into bins, including “small” for 1-to-200-employee companies, “medium” for 201-to-1000-employee companies, “large” for 1001-to-10000+-employee companies, and “NA” for the rest.



We did the same for the Revenue variable, consisting of “micro-bus” for less-than-$1-million-revenue companies, “small-bus” for $1-to-$10-million-revenue companies, “medium-bus” for $10-to-$100-million-revenue companies, “large-bus” for $100-milion-to-$10+-billion-revenue companies, and “NA” for the rest.

A close-up of a document

Description automatically generated

For missing values of Revenue and Size, we imputed them by MICE using the “polyreg” method.

A screenshot of a computer program

Description automatically generated

After visualizing the Rating column, we noticed that it follows the Gaussian/Normal Distribution, so we replaced “-1” values with the mean of the distribution.

A graph of a graph

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

We then rounded and categorized the rating from 0 to 5 into "Very Dissatisfied", "Dissatisfied", "Neutral", "Satisfied", and "Very Satisfied."



We visualized the Founded variable, noticed it does not follow Gaussian/Normal distribution but is a right-skewed graph then replaced the “-1” values with the median of the distribution.

A graph with numbers and lines

Description automatically generatedA graph of a graph of value

Description automatically generated with medium confidence

In the “hourly” column, there are 718 instances of value “0” and 18 instances of value “1”, which means there are 18 values in salary columns using per-hour salary. Therefore, we converted those to a per-year salary, assuming that they work 45 hours per week and 52 weeks per year. We also noticed that min\_salary and max\_salary values are doubled compared to Salary.Estimate so we adjusted them by dividing by 2.

A number of numbers on a white background

Description automatically generated

Moving on to the “age” variable, assuming that the age range of 18 to 70 is the working age group, there are 290 instances of age <18 and 156 instances of age >70, which are outliers.

A graph with a line and a square

Description automatically generated with medium confidence

Hence, we handled them by replacing outliers under 18 with IQR Q1 (30) and outliers above 70 with IQR Q3 (52).

A graph of data in a bar

Description automatically generated with medium confidence

Finally, we selected usable variables and renamed them accordingly, then exported them to “prep\_data.csv.”

A computer screen shot of a computer code

Description automatically generated

## Exploratory Data Analysis (EDA)

### Relationship between Age and Average Salary

A graph of age vs average salary

Description automatically generated

The points on the scatter plot seem to be scattered randomly, which indicates that there seems no relationship between age and average salary.

### Average Salary distribution by Revenue for each Company Size

A graph showing a diagram of a company size

Description automatically generatedA graph of different colored squares

Description automatically generated with medium confidence

Overall, the 3 batches of data look as if they were generally distributed in a similar way. The interquartile ranges are reasonably similar (as shown by the lengths of the boxes). The median average salary of the small-sized companies (110K) is greater than that of the large-sized companies (100K) and the medium-sized (90K) companies. There are outliers observed across large and medium-sized companies, which indicate that these data values are far away from other data values that need further evaluation. Besides, we could conclude that the aggregate average salary of large-bus-revenue companies accounts for the highest, followed by medium-bus, small-bus, and micro-bus ones.

### Average Salary distribution for Rating (Job Satisfaction)

A graph showing a number of salary

Description automatically generated with medium confidence

Individuals who rated satisfied scored a higher median average salary compared to individuals who rated neutral, very satisfied, or dissatisfied. Individuals who rated neutral and very satisfied yielded the same median average salary at 90K. As shown in the dissatisfied column, the box plot is comparatively short, which indicates that individuals who are dissatisfied have around the same average salary as each other (70K-100K). In the very satisfied category, the boxplot is relatively longer, which indicates that individuals hold quite different amounts in terms of average salary (70K-140K). There are outliers observed across individuals with neutral ratings and satisfied ratings, which indicate that these data values are far away from other data values that need further evaluation.

### Average Salary distribution for each Job Title

A graph of different colored rectangular bars

Description automatically generated with medium confidenceA graph with a number of squares and lines

Description automatically generated with medium confidence

Overall, the director position has a higher median average salary compared to R&D engineer (R&D), machine learning engineer (ML), data scientist (D-sci), data engineer (D-eng), other tech jobs (na), manager, and analyst. As shown in the analyst, D-eng, manager, and ML columns, the box plots are comparatively short, which indicates that individuals who are in these mentioned positions have around the same average salary as each other. In addition, the 4 sections of the box plot for manager and ML are relatively uneven in size. This indicates that individuals in that particular position, respectively, have similar average salaries at certain parts of the scale, but in other parts of the scale, individuals in that position are more variable in their average salary. The short upper whisker in the ML boxplot indicates that their average salaries are varied among the least positive quartile group, and very similar for the most positive quartile group. There are outliers observed across all the positions, which indicate that these data values are far away from other data values that need further evaluation.

### Rating (Job Satisfaction) and Company Size

A graph showing a variety of sizes

Description automatically generated with medium confidence

Most individuals working in large, medium, and small companies are generally satisfied with their job. In medium-sized companies, more individuals are dissatisfied with their jobs. It is also noticeable that more individuals are very satisfied with their job work in small companies compared to medium-sized and large-sized companies.

### Rating (Job Satisfaction) and Job Title

A graph of different colored squares

Description automatically generated

Overall, most individuals are generally satisfied with their jobs. It can be observed that there is no dissatisfaction among data engineers, directors, managers, and ML positions among individuals.

### Top States Salary x Skills Required

A graph of different colored bars

Description automatically generatedA graph of a number of colored bars

Description automatically generated with medium confidence

The top 5 states that have the highest average salary are CA, MA, IL, NJ, and DC. Furthermore, Python, Excel, and Spark are the most needed skills to work on every state above. The exception is R, which seems not to be very important in all the states mentioned.

# Main Analysis

## Regression Models for Salary Prediction

We wanted to predict average salary (avg.salary) using selected variables consisting of job title (title), job seniority (job.seniority), job satisfaction (rating), company size (size), company revenue (revenue), age, and skills (python, r, spark, aws, excel), then normalizing them after that to have a data frame as below.

A screenshot of a computer program

Description automatically generated

Then the data is split into training (70%) and testing (30%).

A screenshot of a computer program

Description automatically generated

We run salary prediction using the Multiple Linear Regression Model, Random Forest Model, Gradient Boosting Model, and XGBoost Model.

A number of numbers on a white background

Description automatically generated

Based on the accuracy result, the R2 value for the random forest is the highest and the value of mae, mse, and rmae for the random forest is the lowest; therefore, the random forest model is the best model for salary prediction. However, the 4 models do not perform well, and the max R2 score is only 54% from Random Forest. This may be due to several limitations of the dataset:

* The small size of the dataset, a total of 742 rows, is insufficient to carry out an accurate salary prediction.
* Insufficient features, since there are many other factors unable to be included in this dataset, such as working experience and level of education, should be considered.

## Classification Models for Job Satisfaction Prediction

In this case, we predicted job satisfaction (rating) using selected variables consisting of average salary (avg.salary), job title (title), job seniority (job.seniority), company size (size), company revenue (revenue), age, and skills (python, r, spark, aws, excel), then normalizing them as Salary Prediction Model above.

We also split data into training and testing (70:30). After that, we run classification models including Random Forest, Extreme Gradient Boosting, Naive Bayes, and Gaussian Naive Bayes with results as following respectively.

A black text on a white background

Description automatically generatedA black text with numbers

Description automatically generated with medium confidenceA black text on a white background

Description automatically generatedA black text on a white background

Description automatically generated

Based on the performance of all 4 classification models, Random Forest is the best model for job rating classification.

## N-Way analysis of variance (ANOVA) with Interactions

We run n-way ANOVA to test whether job title (title), job seniority (job.seniority), age, and skills (python, r, spark, aws, excel) and their interactions have an effect on average salary (avg.salary).

A screenshot of a computer code

Description automatically generated

As can be seen from the results:

* Main effects: Job title, job seniority, python skill, and spark skill significantly affect the average salary due to the low corresponding p-values (\*\*\* and \*\*).
* Interaction effects: The interaction terms in the output, consisting of title:python, job.seniority:python, title:spark, title:python:aws, and title:job.seniority:excel at \*\*\* significance level, show that the effect of one factor depends on the level of others. Thus, there is a significant interaction effect between job title and python skill; job title and spark skill; job seniority and python skill; job title, python skill, and aws skill; and job title, job seniority, and excel skill.

# Recommendations

## Recommendations for Stakeholders:

### Recruiters

* Utilize the findings to establish realistic salary benchmarks that reflect the current market trends, considering factors like age, company size, and revenue.
* Develop targeted recruitment strategies that align with the salary expectations and skills required in different states and sectors.

### Finance Department & Executives:

* Reevaluate current salary structures to ensure they are competitive and align with industry standards, particularly for different company sizes and revenues.
* Implement a flexible salary budgeting approach that accounts for variations in job satisfaction and market demand for specific skills.

### Job Seekers:

* Encourage them to use the project findings to understand their market value better and negotiate salaries effectively.
* Provide resources or workshops on understanding salary structures and negotiating salaries based on company size, location, and required skills.

## Potential Further Actions:

### Expand the Dataset:

* For: Larger datasets may provide more accurate insights and help identify trends not visible in the current dataset.
* Against: Collecting and cleaning a larger dataset may require significant resources and time.
* Decision: If resources allow, gradually expand the dataset while maintaining data quality. Otherwise, focus on deepening the analysis with the current dataset.

### Incorporate Additional Features:

* For: Including variables such as education level, years of experience, and specific technical skills could provide a more comprehensive view of salary determinants.
* Against: It may be challenging to obtain this detailed information, and it could introduce privacy concerns.
* Decision: Proceed with adding new variables if they can be ethically sourced and are likely to add significant value to the analysis.

### Develop a Salary Prediction Tool:

* For: A tool that predicts salary ranges based on user-input factors could be extremely valuable for both job seekers and employers.
* Against: Developing a user-friendly and accurate tool requires additional resources and technical expertise.
* Decision: Explore the feasibility of developing a tool if there is sufficient demand and it aligns with organizational capabilities.

### Conduct Regular Market Analysis:

* For: Regular updates to the analysis can keep the salary benchmarks relevant and assist stakeholders in keeping up with market trends.
* Against: Regularly updating the analysis requires ongoing commitment and resources.
* Decision: Schedule annual or biannual updates to the analysis, contingent on the availability of new data and resources.

### Offer Training and Workshops:

* For: Training sessions for job seekers on salary negotiation and for employers on competitive salary structuring could enhance the practical utility of the project findings.
* Against: Organizing workshops requires additional planning, resources, and expertise.
* Decision: If there is enough interest and capacity, organize workshops. Otherwise, consider creating online resources or partnering with educational institutions.

### Explore the Impact of Remote Work:

* For: The shift towards remote work could significantly impact salary structures and should be analyzed.
* Against: Data on remote work salaries might be scarce or inconsistent.
* Decision: Start with a preliminary study on remote work's impact on salaries if data are available and consider a more in-depth analysis as more data becomes accessible.

# Faith and Ethics Implications

## Ethical Implications of the Project

Ethical considerations in this project revolve around data privacy, fairness, and transparency in salary negotiations and decision-making processes.

* **Data Privacy:** Ensuring the privacy and anonymity of individuals contributing to the dataset is paramount. It's essential to anonymize data to prevent the identification of specific individuals and protect their privacy rights.
* **Fairness:** The analysis should aim to uncover patterns and insights that promote fair compensation practices. This involves considering factors beyond just age, company size, and revenue, such as gender, race, and other potentially sensitive attributes. Discrimination based on such factors should be avoided, and efforts should be made to ensure fairness and equal opportunity in salary negotiations and employment practices.
* **Fair Representation:** Ethical considerations also include ensuring fair representation and avoiding biases in data preparation. For instance, when simplifying variables like job titles or company types, it's essential to maintain the integrity of the original data and avoid inadvertently skewing the analysis towards certain categories. Fair representation ensures that insights derived from the data accurately reflect the realities of the industry.
* **Transparency:** Providing clear and transparent information about the methodology, data sources, and limitations of the analysis is crucial. This helps stakeholders, including job seekers and employers, understand how salary recommendations are derived and make informed decisions based on the findings.

## Relationship between the Christian Faith and Ethical Practices Relevant to the Project:

* **Justice and Fairness:** Christian teachings advocate for fair treatment of all individuals, regardless of their background or circumstances. In the context of salary negotiations and compensation structures, this means ensuring equity and fairness in how salaries are determined and distributed. It also involves addressing systemic inequalities and biases that may exist within organizations or the broader industry.
* **Stewardship:** Christians believe in being responsible stewards of the resources entrusted to them by God. In the context of salary allocation and budgeting, this principle calls for wise and prudent management of financial resources, ensuring that salaries are allocated in a manner that is both sustainable for the organization and fair to employees.
* **Compassion and Empathy:** Christian ethics emphasize compassion and empathy towards others, especially those in vulnerable positions. This can translate into considerations for the well-being of employees, including fair compensation that enables them to support themselves and their families adequately. It also involves understanding and addressing the diverse needs and circumstances of individuals in the workforce.