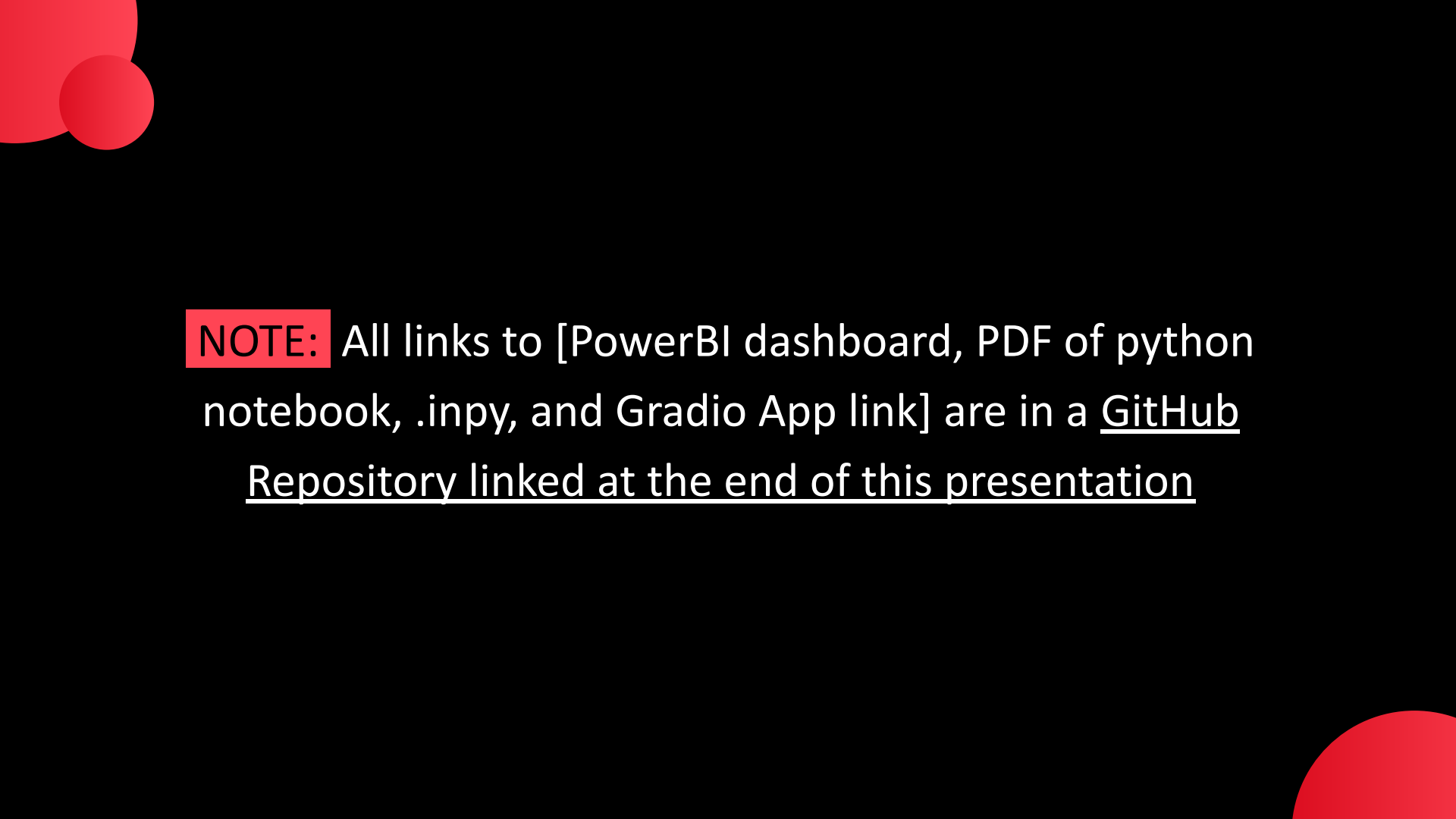


REPORT:

Trends & Salary Drivers in Data Science Careers

BUS 458 (003) Final Case Study

By: Dylan Kshatriya, Lauren Turner, Luca Staeheli, & Nicole Page



NOTE: All links to [PowerBI dashboard, PDF of python notebook, .inpy, and Gradio App link] are in a GitHub Repository linked at the end of this presentation

I. EXECUTIVE SUMMARY

This project examines key factors that can influence and eventually be used to predict the salary of data scientists. Our dataset contains information about different data scientists and their wages. This information includes

- The country that the data scientist resides in,
- The highest education level completed,
- The amount of coding experience they have (if any),
- The amount of ML experience they have (if any),
- Programming languages they use regularly, and
- Machine learning algorithms they use regularly.

The objectives of this project are to develop ML models to predict a data scientist's salary based on the given features. Our team developed three models, chose the highest performing, and used it to develop a Gradio app interface. The interface will allow users to input information relative to the data features and generate a predicted salary amount.


The analysis revealed the United States location, ML experience (numerical amount), and coding experience(numerical amount) were extremely significant in determining the salary. This could be because most respondents with a known salary reside in the United States. Additionally, higher-paying jobs will naturally employ data scientists with more experience in coding and ML algorithms.



II. PROBLEM STATEMENT

There have been many developments in websites and apps that allow people to get information on careers. However, it remains difficult for students or other people pursuing a data science career to know what exact skills, experience, and demographic factors are associated with higher salaries. This lack of information can lead to young professionals entering the field uninformed and unprepared.

Our objective is to create an interface that can estimate a data scientist's salary given their country, educational background, years of experience, programming languages known, and machine learning algorithms known. Continuing this initiative can help increase the data available to build future models. This can, in turn, increase the accuracy of the predictor and further improve the availability of knowledge in the data scientist profession.





III. DATA & ANALYSIS

Data Sources

Kaggle conducted an international survey with over 25,000 responders. The survey was intended to capture the state of data science in 2021.

- Kaggle_survey_2022_responses.csv : This file contains all answers to multiple choice questions from responders.
- Kaggle_survey_2022_answer_choices.pdf : This file is a data dictionary with the actual questions asked in each column of the survey.
- Kaggle_survey_2022_methodology : A description of how the survey was conducted

<https://www.kaggle.com/c/kaggle-survey-2022/overview>



Kaggle_survey_2022_responses.csv :

43 questions and 23,997 responses

Responses to multiple choice questions (only a single choice can be selected) were recorded in individual columns.

Responses to multiple selection questions

(multiple choices can be selected) were split into multiple columns (with one column per answer choice).

Dataset Description

Analysis Approach

1. Exploratory Data Analysis (EDA)
2. Feature Engineering
3. Predictive Modeling
4. Model Evaluation
5. App Development



EDA & Data Preprocessing

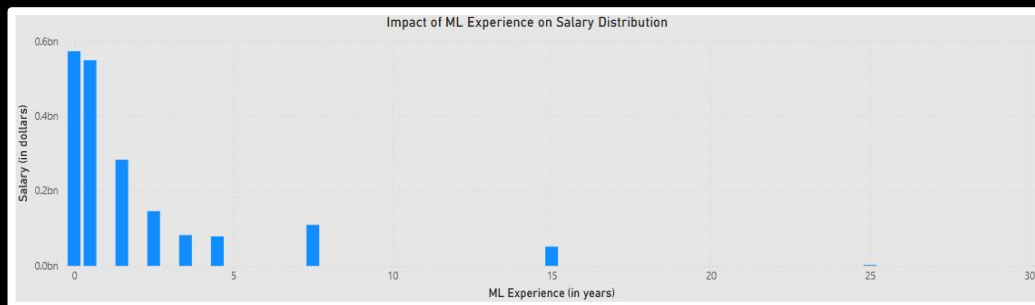
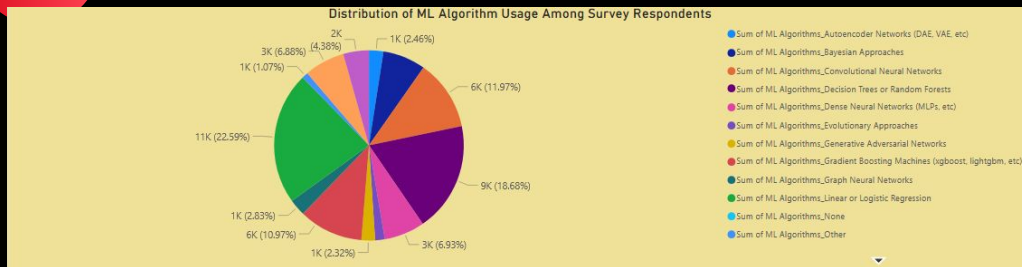
Data Cleaning Steps:

- Selected only relevant questions (Q4, Q8, Q11, Q12, Q16, Q18, Q23, Q24, Q29).
 - Q4: Country of residence
 - Q8: Highest level of education
 - Q11: Years coding
 - Q12: Programming languages used
 - Q16: Years using ML methods
 - Q18: ML algorithms used
 - Q29: Current salary
- Converted text ranges (e.g., "1-3 years coding") into numerical midpoints.
- Encoded categorical variables (one-hot for countries, roles, industries).
- Created binary features for tools and ML techniques.

Raw Kaggle Data → Cleaned Dataset (36 columns) → Final Feature Matrix (47 columns)

Data Cleaning Steps and some EDA took place in Google Collab.

Power BI was also utilized to help visualize some EDA as well. The following questions are addressed using both analysis using python in Collab and Power BI.

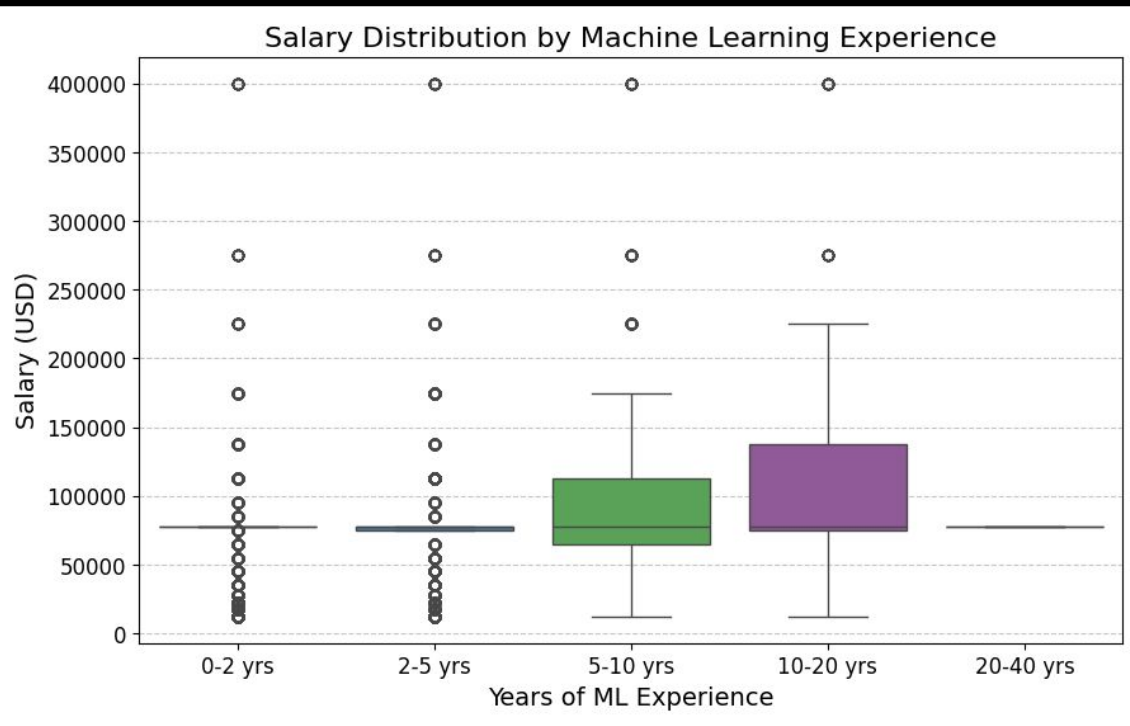


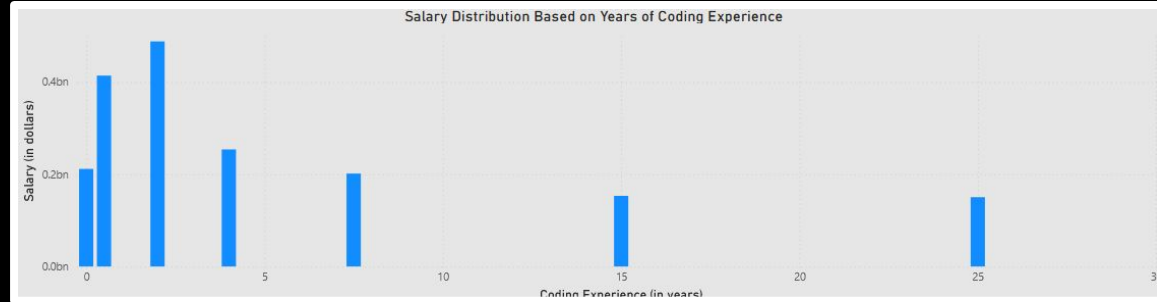
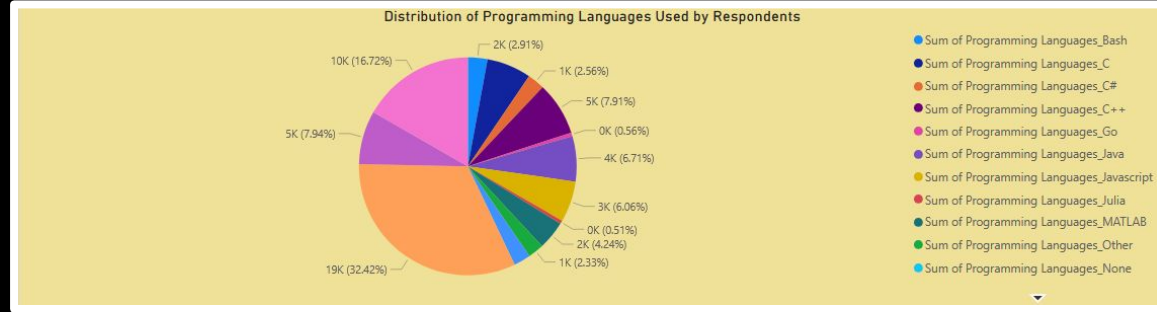
- **Programming Languages:** Python (most common), SQL, R.
- **ML Algorithms:** Logistic Regression, Random Forests, Gradient Boosting.

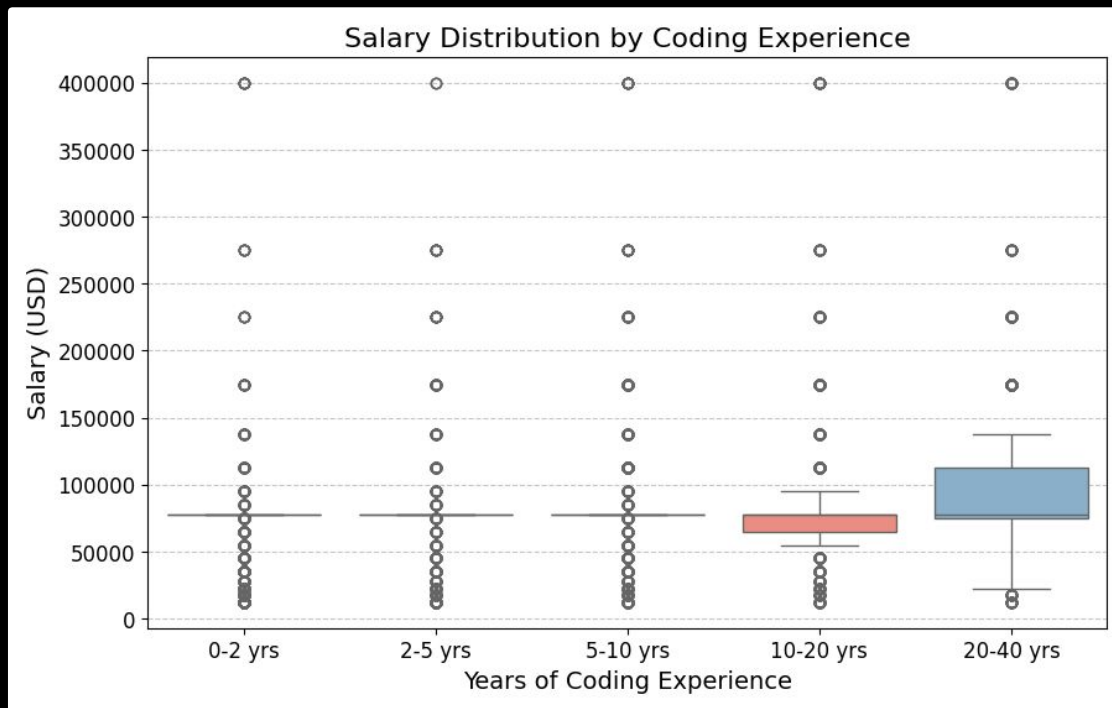
Insight:

- Python and SQL are industry standards.
- Core ML algorithms dominate over deep learning for most roles.

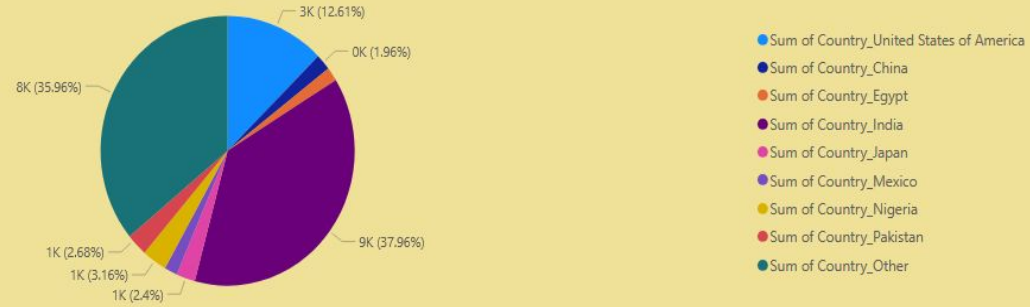
What are the most prevalent tools (software) and techniques (methods) being applied by Data Scientists today? (Q2)







Distribution of Respondents by Country



Proportion of Respondents by Education Level



What tools and techniques are emerging in the field of Data Science? (Q3)

ML Algorithm Usage Among Survey Respondents

- Gradient Boosting Machines and Convolutional Neural Networks (CNNs) are the most widely used ML techniques among respondents. Also, a diverse set of ML algorithms is in use, with Linear/Logistic Regression and Recurrent Neural Networks (RNNs) also showing significant popularity.

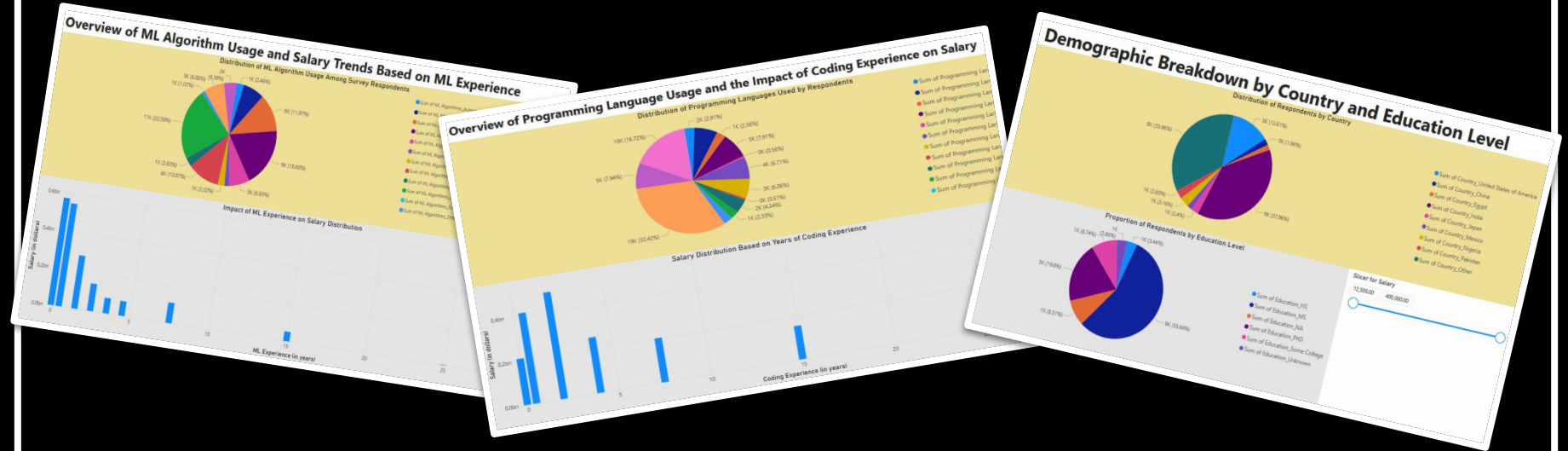
Impact of ML Experience on Salary Distribution

- Higher years of ML experience tend to correlate with higher salary ranges, with a concentration of higher salaries around 5-10 years of experience. Also, the salary distribution shows a significant variance, especially among those with 0-5 years of experience, indicating rapid salary growth with experience.

Three emerging tools in the field of data science from the article:

1. **Automated Machine Learning (AutoML):** This tool automates the machine learning process, making it accessible to those with little technical expertise.
2. **Edge Computing:** Edge computing processes data closer to its source, improving response times and saving bandwidth, making it particularly useful in real-time applications.
3. **Augmented Analytics:** This approach leverages AI and machine learning to automate data preparation, insight generation, and data analysis, helping businesses make faster, data-driven decisions.

POWER BI DASHBOARD :



** Link to published interactive Dashboard on GitHub

Creating a Predictive Model for Salary:

Goals:

Represent the importance of numerical experience features (years of coding, years of ML) along with the categorical representation

Build a model that predicts a continuous outcome (salary) rather than a binary or categorical outcome.

Design an interactive tool that can input user characteristics and output a salary estimate in real-time.

What models did we fit?

1. **Lasso Linear Regression**
 - a. MSE: 924723845
 - b. RMSE: 30409
 - c. R^2 : 0.1329
2. **Random Forest**
 - a. MSE: 879375420
 - b. R^2 : 0.1754
 - c. Out-of-Bag: 0.1659
3. **Gradient Boosting Model**
 - a. MSE: 847807798
 - b. RMSE: 29117
 - c. R^2 : 0.2050

The model we chose:

Choosing a model for this dataset is a bit difficult because the dataset has a lot of noise. There are a lot of variables in the dataset, and many of them did not directly have an impact on the target variable.

The weak signal between the dependent variable (salary_numerical) and independent variables, cause the r^2 to be lower. Additionally, having a numerical target has made model evaluation different than with categorical variables. We cannot generate a classification report without binning the salary variables. This means accuracy would need to be evaluated through mean squared error, root mean error, and the r^2 .

The model we decided to choose was the Gradient Boosting model. This model overall had the highest r^2 and lowest mean square error.

Final Model Metrics: Gradient Boosting Model

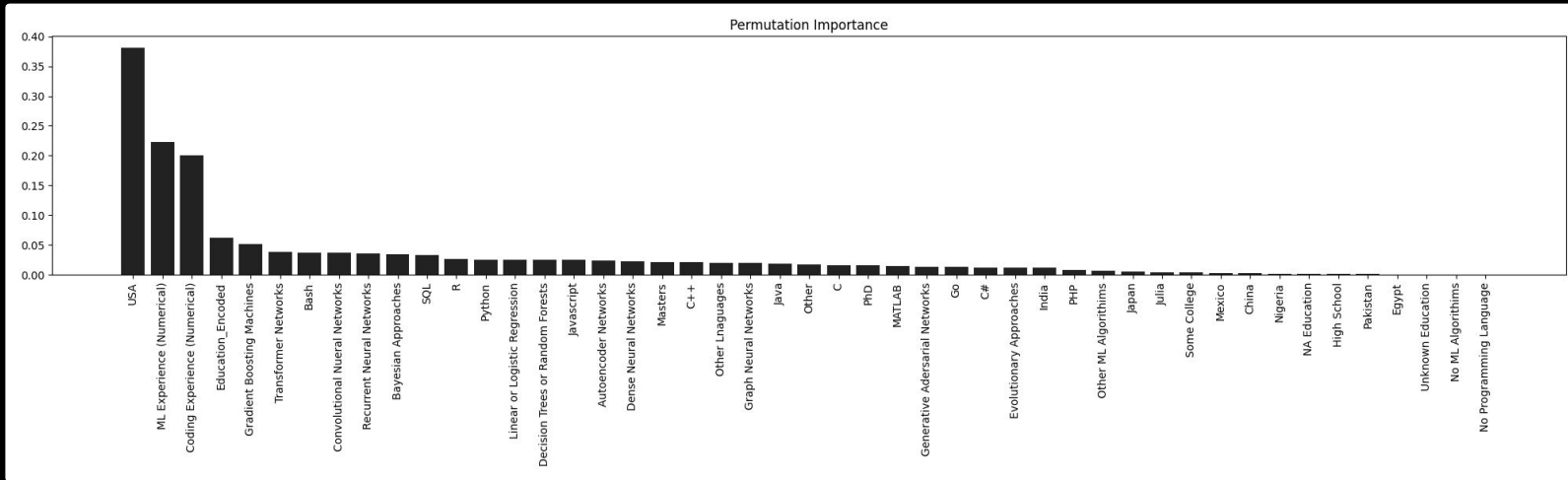
Key insight: We cannot generate a classification report without binning the salary variables. This means accuracy would need to be evaluated through mean squared error, root mean error, and the r^2 .

Mean Squared Error: 846724985.516962

Root Mean Squared Error: 29098.539233386993


R-squared: 0.2060637528481044


What are the most significant factors driving Data Scientist Salaries? (Q1)



- **Years of coding experience:** Strong positive relationship — more years, higher salary.
- **Machine learning experience:** Specialized ML knowledge boosts earning potential beyond general coding experience.
- **Country of employment:** Salaries vary dramatically across regions — U.S., Western Europe offer much higher median salaries compared to India, Nigeria, and Pakistan.
- **Education level:** Higher degrees (Master's, PhD) are associated with slightly higher median salaries, but the salary gap is modest compared to experience.
Programming skills: Proficiency in **Python**, **SQL**, and other popular languages correlates with higher pay.
- **Use of advanced ML algorithms:** Professionals skilled in **Gradient Boosting**, **Transformers**, and **Deep Learning** techniques tend to earn more.

SCORING APP OVERVIEW

 **Data Scientist Salary Predictor**

 Predict your salary based on education, coding experience, programming languages, machine learning techniques, and country.

Education Level
MS

Years of Coding Experience
0 8 40

Years of Machine Learning Experience
0 7 40

Country
United States of America

☐ Knows Python
☐ Knows R
☐ Knows SQL
☐ Knows C
☐ Knows C#
☐ Knows C++
☐ Knows Java
☐ Knows JavaScript
☒ Knows Bash
☐ Knows PHP
☐ Knows MATLAB
☐ Knows Julia

Objective: Provide students and young professionals with a salary estimate based on their skills, education, experience, and country.

Inputs Collected (Interactive Components):

- Dropdown Menus, Sliders, & Checkboxes (Multi-Select)

Deployment:

- Built using
- Generates real-time, user-specific salary predictions

Predicted Salary

 Estimated Salary: \$81,882.02



IV. FINDINGS & RECOMMENDATIONS

In our model's feature importances, **Education_Encoded** ranked **lower** compared to technical experience features.

Years of coding experience and **years of machine learning experience** had a much **stronger** influence on salary predictions than education level.

Education showed a **positive but modest** contribution:

- Having a Bachelor's, Master's, or PhD **does improve salary**, but **it is not the primary driver**.

Practical skills and experience (coding, machine learning methods, use of advanced algorithms) were more critical for higher salary outcomes.

- As seen when looking at the box plots of Salary by coding and ML experience, the median increases by a higher magnitude when compared to medians of Salary by Education.

Is formal education important to success as a Data Scientist? (Q5)

Salary by Education:

	count	mean	std	min	25%	50%	75%	max
Education								
NA	1394.0	74617.594192	1.974267e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
HS	564.0	75556.150467	3.549976e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
Some College	1431.0	77184.634384	2.449793e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
BS	7625.0	76965.471356	2.589921e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
MS	9142.0	78243.047570	3.538864e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
PhD	3242.0	82429.002235	4.557153e+04	12500.0000	78061.2143	78061.2143	78061.2143	400000.0000
Unknown	599.0	78061.214300	6.699476e-10	78061.2143	78061.2143	78061.2143	78061.2143	78061.2143

How does the return on formal education compare to other types of learning? (Q6)

Conclusion: While formal education can help salary outcomes, it may not justify the extra investment for every student — depending on personal circumstances, faster pathways may be equally or more effective.

90% of working data scientists have obtained a formal advanced degree (Bachelor's, Master's, or PhD).

<https://onlinesoe.tufts.edu/blog/data-science-bootcamps-masters-programs/>

In our dataset, mean salaries showed a gradual increase with higher education:

- Bachelor's: ~\$77,000
 - Master's: ~\$78,000
 - PhD: ~\$82,000
- Although the data shows a **positive salary increase** with higher education (Bachelor's → Master's → PhD), the **absolute difference in mean salary is relatively modest** (~\$5K–\$7K between levels).
 - Assuming higher degrees involve substantial time, tuition costs, and opportunity cost, the financial **return on investment (ROI)** for advanced degrees may be **positive but not dramatically high**.
 - For many aspiring data scientists, **bootcamps, certifications, and self-learning paths** may offer a **faster, lower-cost route** to entering the field, especially given the growing industry acceptance of skills-based hiring.

How and where should aspiring data scientist invest their time and energy to prepare for the current and future Data Science environment? (Q4)

Recommendations for Aspiring Data Scientists:

- **Invest heavily in programming skills**
- **Learn data visualization tools:** Practice building dashboards and visual insights using **Power BI** and **Tableau**.
- **Strengthen math and statistics foundations:** Develop a strong understanding of probability, statistics, and linear algebra to better interpret data models.
- **Get real-world practice**
- **Stay updated with evolving skills:** Gain exposure to cloud computing (AWS, GCP), MLOps practices, and emerging fields like NLP and deep learning.

Supporting Research:

According to Simplilearn (2023), building practical experience, gaining expertise in core programming languages like Python, mastering visualization tools, and maintaining a strong math/statistics base are critical steps for preparing for the future of data science careers.

[\(Simplilearn, "The Future of Data Science"\)](#)

How should educational institutions think about the role of formal education in the world of data science? Are there any specific recommendations in terms of methodologies that institutions of formal education should employ in the training of data scientists? (Q7)

Higher education still offers **measurable financial benefits**: salaries rise with Bachelor's, Master's, and PhD degrees according to our analysis.

However, **formal degrees are no longer the only path**: bootcamps, certifications, and self-taught skills offer faster, cheaper entry points into the field.

To remain competitive, **universities must offer more than just a credential**:

- **Industry connections**: direct pipelines to internships, job placements, and corporate partnerships.
- **Research opportunities**: access to cutting-edge projects in AI, ML, sustainability, and healthcare.
- **Social impact**: framing data science careers around solving real-world problems (climate change, social justice, healthcare equity).
- **Global networking**: building alumni communities and collaborative programs across industries.

Universities should **emphasize applied skills** (Python, SQL, ML techniques) alongside **theoretical foundations** (statistics, ethics, critical thinking).



V. LIMITATIONS & NEXT STEPS

Limitations:

- Self-reported data may have inconsistencies or bias.
- Survey skewed towards Kaggle-active, self-taught populations.
- Salary ranges had to be midpoint estimated.

Next Steps:

- Apply more advanced regression techniques (e.g., CatBoost, Stacked Models).
- Expand analysis to newer datasets (2023-2024).

The background of the image is a dark, grayscale photograph. It shows the silhouette of a person in the lower-left corner, looking out of a window. The window frame and the view outside, which appears to be a cityscape with buildings, are visible. Overlaid on this background are several bright red geometric shapes: a cluster of overlapping circles in the top-left corner, a single circle in the top-right corner, a single circle in the bottom-left corner, and a cluster of overlapping circles in the bottom-right corner.

THANK YOU

Group Contributions:

Data preprocessing - Nicole
Slide Deck Creation and Formatting - Nicole and Dylan
Slides 6 -9, 20, 21, 26, 28 - Nicole
Gradio Code/Deployment - Nicole and Lauren
Slide 10, 25- Dylan
Pie chart of ML algorithm - Luca
Pie chart of coding languages - Luca
Salary by ML experience numerical - Luca
Salary by coding experience numerical - Luca
Return on investment cost of education VS salary - Nicole
Power BI Dashboard Creation - Luca
Feature Importance graph - Lauren
Gradient Boosting model performance image - Dylan
Model 1 - Dylan
Model 2 - Lauren
Model 3 - Lauren
Slide 23 - Luca
Slide 1 (executive summary) - Lauren
Slide 2 (problem statement) - Lauren
Slides 18,19 - Lauren
Github - Nicole

Github:

<https://github.com/nlpage154/bus458finalproject>

Link to APP:

* accessible through readme on github and linked below

* <https://485c64489ad86e4e69.gradio.live/>