# PSTAT 100 Final Project Report

Aliza Samad, Akshara Kollu, Oliver Zhou, Quinn Giammaria

2024-12-03

## Introduction

The rapid advancement of Artificial Intelligence has influenced industries worldwide, reshaping workflows, automating processes, and boosting productivity. One sector impacted is data science, where AI's growing role has changed job roles and compensation patterns. This research seeks to address these shifts by addressing the question: **How has the growth of AI influenced job titles and salary trends in the data science field?**

Initially, we considered a broader scope, including the question: How does experience level affect salaries in the data science field? Our hypothesis - that salaries increase with experience - is well-supported and confirmed through preliminary EDA. While this finding reinforces common assumptions, it offers limited opportunity for deeper discussions. To provide a more focused and impactful contribution, we shifted our focus to the evolving relationship between AI and data science roles.

We hypothesize that AI-specific roles have increased in demand *and* have higher average salaries than traditional data science roles. We will break down this hypothesis in two components as to not cause any errors.

$H_0 1$**: The demand for AI-specific roles has not increased more significantly than the demand of other roles.**

$$H_{01} : \mu_{\text{demand, AI}} \leq \mu_{\text{demand, baseline}}$$

$H_1 1$**: The demand for AI-specific roles has increased more significantly than the demand of other roles.**

$$H_{11} : \mu_{\text{demand, AI}} > \mu_{\text{demand, baseline}}$$

$H_0 2$**: AI-specific roles do not have higher average salaries than other data science roles.**

$$H_{02} : \mu_{\text{salary, AI}} \leq \mu_{\text{salary, other}}$$

$H_1 2$**: AI-specific roles do have higher average salaries than other data science roles.**

$$H_{12} : \mu_{\text{salary, AI}} > \mu_{\text{salary, other}}$$

To test this hypothesis, we utilized a Kaggle data set titled "Data Science Salaries Dataset," which included a detailed repository of experience levels, employment types, work years, salary in USD, and more. Below is a snapshot of our cleaned data set:

Table 1: Data Science Salaries (clean)

| work_year | experience_level | employment_type | job_title | salary_in_usd | employee_residence | remote_ratio | company_size |
|---|---|---|---|---|---|---|---|
| 2021 | MI | FT | Other | 36259 | HU | 50 | L |

| work_year | experience_level | employment_type | job_title | salary_in_usd | employee_residence | remote_ratio | company_size |
|---|---|---|---|---|---|---|---|
| 2021 | SE | FT | Machine Learning/AI | 54094 | IN | 50 | L |
| 2022 | MI | FT | Machine Learning/AI | 31795 | IN | 100 | M |
| 2023 | SE | FT | Machine Learning/AI | 311000 | US | 0 | M |
| 2024 | SE | FT | Machine Learning/AI | 310270 | US | 0 | M |

Through EDA, we made several observations. First, we initially aimed to map how `company_location` affected `salary_in_usd`, but the data was skewed with too many U.S. countries, so we have focused on U.S. companies only. Second, we noticed a drop in average salary from 2020 to 2021, likely due to the Covid-19 pandemic, followed by an increase in salary from 2022 to 2023, which aligns with the rise of generative AI. However, we observed a salary decrease from 2023 to 2024, which will be explored further. Lastly, we created a word cloud comparing job titles in lower and higher salary ranges (filtering out the word "data"). We found roles like "analyst" and "business intelligence" had lower salaries, while titles like "machine learning" and "research" had higher salaries. Interestingly, "engineer" appeared equally in both salary ranges. This initial analysis led us to further explore the relationship between `salary_in_usd`, `job_title`, and `work_year`.

By narrowing the scope to AI's influence on data science salaries and roles, this study aims to provide actionable insights for professionals and organizations within US companies navigating through this rapidly evolving field.

# Method

## Modeling Process

We found consistently low $R^2$ values (<20%) in our model, indicating that data is non-linear, even after transforming the variables. While the residuals histogram showed a nearly symmetrical bell curve, it had a slight right skew. The QQ-plot of the residuals also revealed deviations from the trend line, suggesting non-normality in the residuals. Additionally, the residuals vs. fitted values plot showed a cone-line shape (Figre 2), indicating heteroscedasticity. These issues confirm that assumptions for linear regression were not met.

Models tested:

```
## R-squared value: 0.2084481
```

```
## Log-transformed R-squared value: 0.2465606
```

As a result, we decided to use a random forest multiple regression model since this is more flexible. For one, the random forest model is a non-parametric model, and it does not require homoscedasticity. Additionally, we are working with a large sample size (over 2000 observations), and each observation is independent of each other. By the nature of the random forest model, we were also less likely to experience over-fitting, and we are also able to use the feature-importance plot in order to better understand which features to focus on in our analysis. Finally, this model is very useful for predicting salaries based on selected features, which is something we would like to take away from our research. Although the log-transformation did slightly increase the RMSE, we believe that a greater model fit is worth the trade-off.

Model selected:

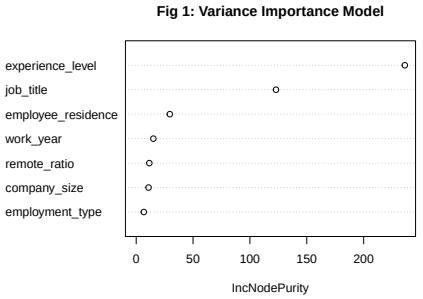
```r
#Split data
set.seed(888)
trainIndex <- createDataPartition((data_no_outliers$salary_in_usd), p = 0.8, list = FALSE)
train <- data_no_outliers[trainIndex, ]
test <- data_no_outliers[-trainIndex, ]

#Random forest model
model <- randomForest(log(salary_in_usd) ~., data = train, ntree = 500, mtry= 6)

#Predictions
predictions <- predict(model, test)
```

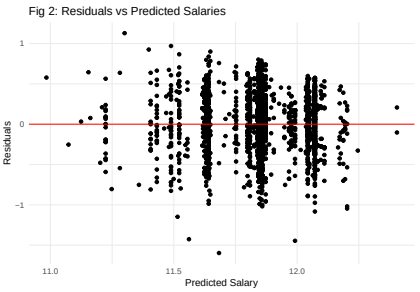
```
## MAE: 0.2762246
```

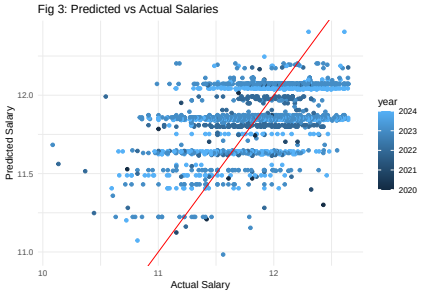
```
## RMSE: 0.3457671
```

```
## R-squared: 0.220164
```

```
## Adjusted R-squared: 0.2180097
```




```
## MAE for Median Model: 0.3087245
```

```
## RMSE for Median Model: 0.3924998
```

Figure 2 shows that the residuals are scattered around 0 showing no clear pattern. However, the variance does seem to increase slightly as the predicted salaries value increases, suggesting that there is some heteroscedasticity. Given that the RMSE is about 10,000 USD greater than the MAE, this suggests that outliers may be playing a significant role in this increased variance, especially around the higher predicited salary values.

Figure 3 shows that the data points generally follow the trend along the y=x axis. However again, there is there is some deviation from the line especially at higher salary values. This is indicative of the model overestimating salary, especially in the higher salary ranges. The color gradient does not give insightful information regarding the spread of salary by year, but we will be exploring this later on in the report.

Although the $R^2$ value for the random forest model is considerably low and only explains 22.02% of the variability in `salary_in_usd`, the model may still be useful in identifying patterns or trends in the non-linear relationships between predictors and the response variable. While the MAE does yield an error of approximately 40,000 USD, given the wide distribution of factors such as company size, job title, and experience, this is simply the result of unexplained variance.

If we compared the random forest model to a simple median model, we see that the random forest model outperforms the median model. Our model does capture relevant patterns since the MAE and RMSE of the median model are both greater than that of the random forest model. Note that we chose the use he median model as our baseline since this is a more robust metric than mean. As mentioned earlier, the data follows a skewed distribution and is heteroscedastic, so the assumptions to conduct a linear regression test have not been met.

Finally, this model helps us take note of key predictor variables (`job_title` and `experience_level`). This is important to note since our research question originally only led us to explore the relationship between `salary_in_usd`, `work_year`, and `job_title`; however, given the importance of the `experience_level` feature, this should be considered in our analysis as well.

## Visualizations

### Cluster Analysis

**Fig 4: scree plot**

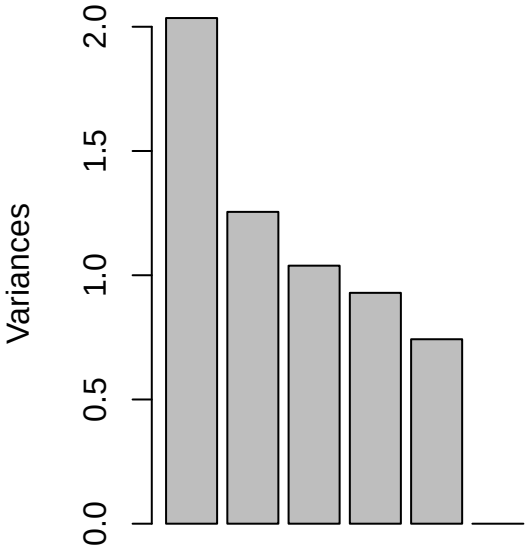
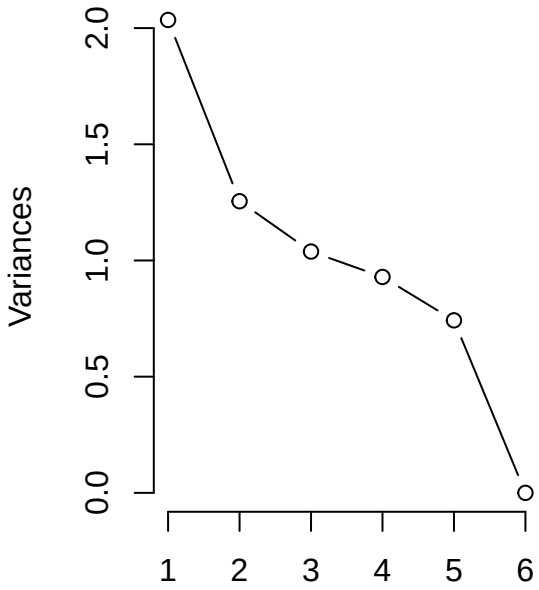



**Fig 5: pca_result**




```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5       PC6
## Standard deviation     1.4266 1.1203 1.0191 0.9639 0.8616 1.16e-14
## Proportion of Variance 0.3392 0.2092 0.1731 0.1548 0.1237 0.00e+00
## Cumulative Proportion  0.3392 0.5484 0.7214 0.8763 1.0000 1.00e+00
```

Fig 6: 3D Scatter Plot of PCA Components

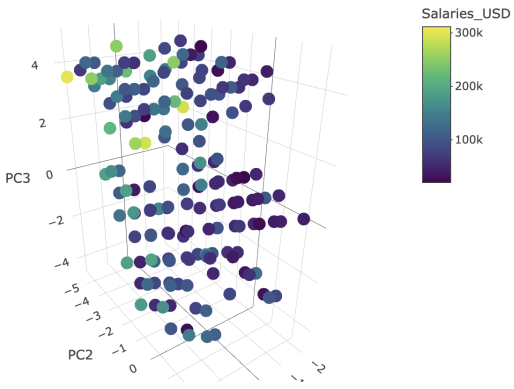



Fig 7: 3D Scatter Plot of PCA Components with K-means Clustering

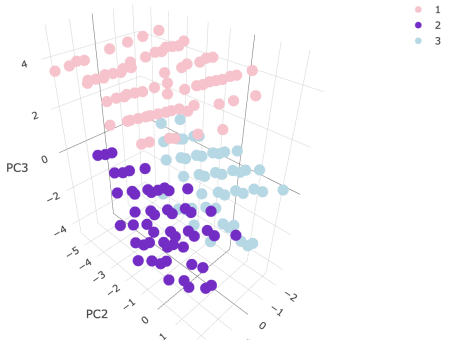
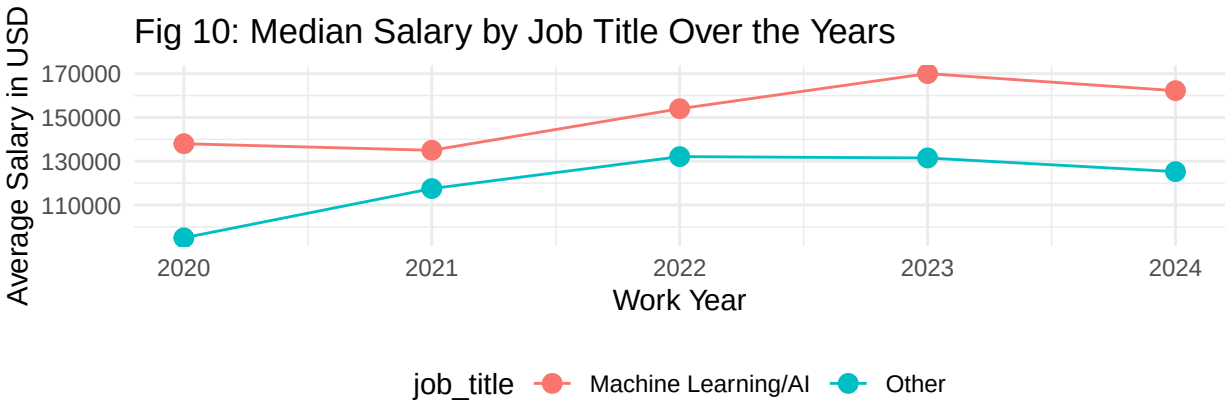


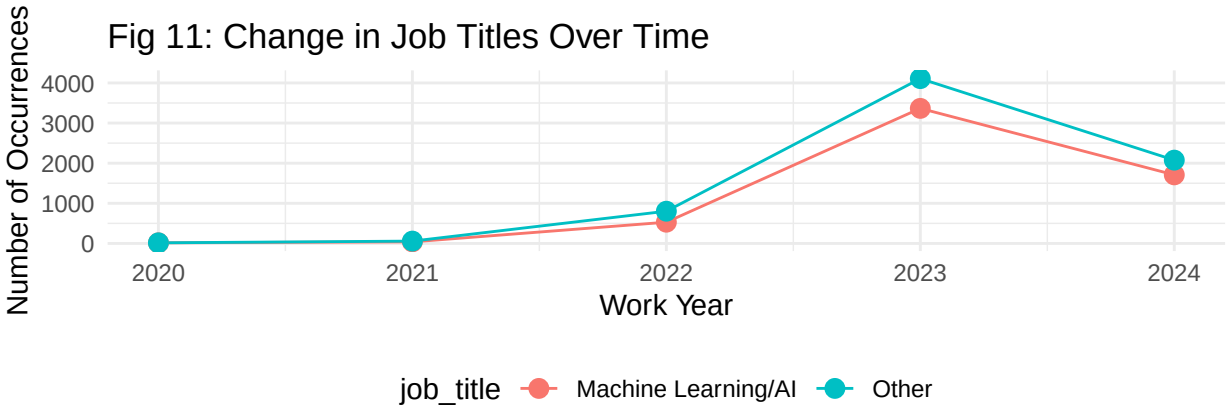


Table 2: Summary Statistics of Cluster Analysis

| cluster | work_year | work_year_mean | salary_mean | salary_median | salary_min | remote_mean | remote_median | remote_min | experience_level | employment_type | company_location | company_size | job_title_mode |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2022.762 | 2023 | 160709.4 | 145000 | 25.59607 | 0 | SE | FT | US | L | Machine Learning/AI | | |
| 2 | 2023.211 | 2023 | 171710.5 | 166600 | 30.76850 | 0 | SE | FT | US | M | Machine Learning/AI | | |
| 3 | 2023.182 | 2023 | 137273.6 | 130000 | 33.36286 | 0 | SE | FT | US | M | Other | | |

**Time Series Analysis**



### Predicting Future Trends

Table 3: Predicted Salary 2025

| work_year | experience_level | employment_type | job_title | employee_residence | remote_ratio | company_size | predicted_salary |
|---|---|---|---|---|---|---|---|
| 2025 | EN | FT | Machine Learning/AI | US | 0 | M | 11.55545 |
| 2025 | MI | FT | Machine Learning/AI | US | 0 | M | 11.84293 |
| 2025 | SE | FT | Machine Learning/AI | US | 0 | M | 12.04327 |
| 2025 | EX | FT | Machine Learning/AI | US | 0 | M | 12.18721 |
| 2025 | EN | FT | Other | US | 0 | M | 11.48866 |
| 2025 | MI | FT | Other | US | 0 | M | 11.64157 |
| 2025 | SE | FT | Other | US | 0 | M | 11.85509 |
| 2025 | EX | FT | Other | US | 0 | M | 12.03596 |

Fig 12: Salary Distribution by Experience Level and Job Title

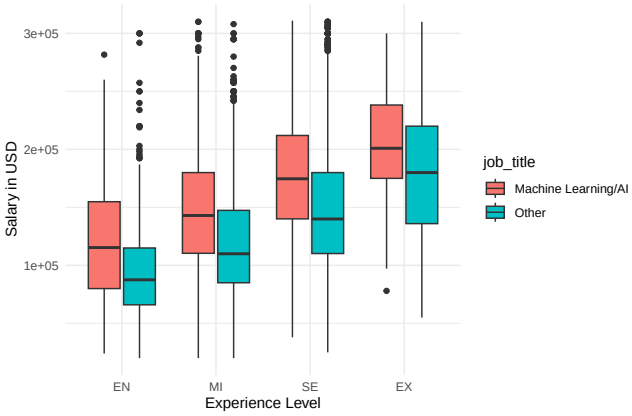
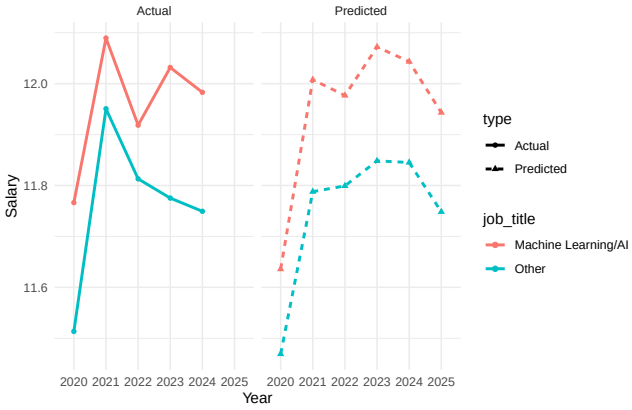



Fig. 13: Median Salaries Over Time by Job Title and Type




## Results

### Testing Hypothesis 01

We will be testing the first hypothesis by utilizing counts of grouped work_years (2020-2024) and job_title (which we have modified to have two options: Machine Learning/AI and Other). We will be performing a Welch Two-Sample T-test to evaluate the difference in mean job counts between the "Machine Learning/AI" category and "Other" roles.

```
##
##  Welch Two Sample t-test
##
## data:  AI_demand$count and Other_demand$count
## t = -0.28044, df = 7.7302, p-value = 0.6067
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -2148.152       Inf
## sample estimates:
## mean of x mean of y
##    1131.6    1412.0
```

The observed p-value (0.6067) is greater than the significance threshold of 0.05. Consequently, we fail to reject the null-hypothesis that the demand for AI-specific roles has not increased more significantly than the demand of other roles. This outcome suggests that the differences in demand observed for "Machine Learning/AI" roles compared to "Other" roles is not may not be statistically significant over the period analyzed. Further investigation using a larger dataset and more years might provide greater insight.

**Testing Hypothesis 02**

Hypothesis 2 evaluates whether AI-specific roles command higher average salaries compared to other roles in the data science field. This involves comparing the average salary_in_usd for two categories: Machine Learning/AI and Other. We will be using a Welch Two Sample t-test to compare the means of salary_in_usd for the two groups.

```
##
##  Welch Two Sample t-test
##
## data:  AI_roles$salary_in_usd and other_roles$salary_in_usd
## t = 36.425, df = 11964, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  33176.22      Inf
## sample estimates:
## mean of x mean of y
##  172039.4  137294.1
```

The observed p-value (2.2e-16) is smaller than the significance threshold of 0.05. Consequently, we reject the null-hypothesis that AI-specific roles do not have higher average salaries than other data science roles. This outcome suggests that the AI-specific roles do have higher average salaries than other data science roles. Further investigation using a larger dataset and more years might provide greater insight.

**Cluster Analysis**

Let's start with the cluster analysis. The scree plots in Figures 4 and 5 did not clearly indicate which components to include. We used the cumulative proportions to determine that PC1, PC2, and PC3 explain sufficient variability in the data without excess overlap or complexity. Figures 6 and 7 give us a little more insight regarding the relationship between `salary_in_usd`, and `job_title`. We can see 6 distinct groups of data in Figure 6, with the 2 groups at the highest PC3 value having higher salaries (as indicated by the yellow/green dots). In Figure 7, we can see the same data color-coded by Cluster. The two groups of data that we observed earlier are part of Cluster 1. The other 4 groups of data, which showed no obvious salary difference in Figure 6, have been categorized into 2 distinct clusters.

The results in Figure 8 were surprising at first glance. Cluster 1, which had the most variability in salary and the highest salary data, consists of both ML and "Other" job types almost equally. This will be looked into as we discuss Table 2. Figure 8 also indicates that Cluster 2 is entirely composed of ML/AI job titles while Cluster 3 is entirely comprised of job titles that fall under the "Other" category.

Table 2 and Figure 9 give us a more well-rounded explanation of the cluster relationships. Firstly, these figures confirm what Figure 8 portrayed previously: Cluster 2 and 3 are mainly separated based on the `job_title` and `remote_ratio`features since all other features remain constant between the two (besides `salary_in_usd`). Cluster 1 seems to be separated based on the `company_size` and `remote_ratio` features since the most common `company_size` in this cluster is L as compared to M. Another noteworthy observation is that the mean `salary_in_usd` for Clusters 1 and 2 are approximately the same, but median depicts a very clear difference. Additionally, the difference in mean salary vs. median salary in Cluster 1 is the largest among all 3 Clusters, suggesting that outliers play a strong influence in Cluster 1. In context, this actually makes sense since larger companies tend to have a more distinct hierarchy of employees and have more executives than medium-sized companies. This would explain why there are a few extremely high `salary_in_usd` values in Figure 6 (executive pay) even though the median `salary_in_usd` value for Cluster 1 is 15,000 USD lower.

Figure 9 depicts the distribution of salary by cluster. For the sake of comparison, we chose to focus on the differences between Cluster 2 and Cluster 3. Cluster 2 has the higher median salary. Since Cluster 2 is mostly comprised of ML/AI job titles, we can assume that having an AI-specific job title does tend to result in a higher salary than traditional, or in this case "Other", roles. Thus, this finding supports the second half of our hypothesis.

**Time Series Analysis**

Next, let's look into the time series plots in Figures 10 and 11. Figure 10 depicts the trends in average salary from 2020 to 2024, separated by job title. Overall, it does look like those who hold job titles in ML/AI have higher average salaries than those who hold more traditional roles. From 2020 to 2021, the average salary of traditional roles increased at a higher rate than ML/AI roles. However, between 2021 to 2023, salary increases at a higher rate for ML/AI job titles than for "Other" job titles. Finally, from 2023 to 2024, the salaries for ML/AI and "Other" jobs decreased at about the same rate.

In order to understand these differences, we have to consider the impact of external economic factors. 2020 to 2021 was the peak of the Covid-19 pandemic. The salary growth of traditional data science roles could be attributed to increased market demand (need for BI analysts, etc.) and work-from-home options (building lease money goes into salary instead). AI-specific roles might have decreased in salary since this field was relatively new and thus very research-heavy. During a time of uncertainty, sponsoring research or AI start-ups would have been a gamble. From 2021 to 2023, the sudden increase in AI-specific role salaries could be attributed to post-pandemic economic revitalization and stimulation. In 2022, the popularity of generative AI spurring from the release of OpenAI's ChatGPT would have solidified AI/ML roles, making these roles some of the most valued in the industry. But how does this explain the sudden drop from 2023 to 2024? Given that the average salaries of both `job_title` categories decreased at the same rate, this is most likely due to market factors as well. This could be the result of over-hiring in 2020 and 2021, depleted resources, economic recession, or a combination of these factors.

Given the trends in AI growth and average salaries from 2020 to 2023, we believe that this supports the first half of our hypothesis: AI-specifc roles have higher salaries than traditional roles. However, it is also important to note that the assumptions we have made regarding market behaior have not been thoroughly explored. We would still need to do some external research in terms of the effect of the economy on the AI industry and vice versa.

The second half of our hypothesis can be answered by Figure 11. Based on the line plot of changes in job titles from 2020 to 2024, while the number of job titles in the industry has generally increased over time (except for 2023 to 2024), it looks like the frequency of AI-specific roles is increasing at approximately the same rate as traditional roles. Note that there are actually slightly fewer AI-specific roles than traditional roles; however, this is expected since ML/AI are relatively new fields in the industry. Based on the results of Figure 11, it is clear that the demand for AI-specific roles has not increased over time as compared to traditional roles.

**Predicting Future Trends**

This is our main takeaway from the research. The goal of this section is to understand trends in data science salaries in the upcoming year (2025). We chose to forecast only one year ahead since our model did not weigh `work_year` as heavily in terms of importance. Figure 12 is used to emphasize the importance of factoring in `experience level` as an influential predictor variable. For this reason, we decided to keep all other predictors besides `job_title` and `experience_level` constant in the `future_data` data set. Table 3 shows the results of this prediction, and we can see that as expected, the ML/AI job titles have higher salaries than those that hold traditional roles. Additionally, as experience level increases, predicted salary increases as well.

Figure 13 is key to understanding how the 2025 predictions compare to the rest of the data. We have two facets comparing the actual salary and the predicted salary based on the `job_title` and `work_year` predictor variables. For sake of simplification, we did not include the `experience_level` predictor variable in this analysis and instead took the median salary of each year. The patterns for the actual and predicted type

plots are very similar, with the same sharp incline from 2020 to 2021 similar variations from 2021 onward. We can assume that the differences in the actual "Other" salaries vs. the predicted "Other" salaries trends are due to errors within the model we used, especially considering that our MAE was about 40,000 USD.

Based on the findings, salaries for both traditional roles and AI-specific roles are expected to decrease. However, it is important to take these predictions with a grain of salt. As mentioned in the time series analysis section, this model does not account for external market factors such as resource shortages.

# Conclusion & Recommendations

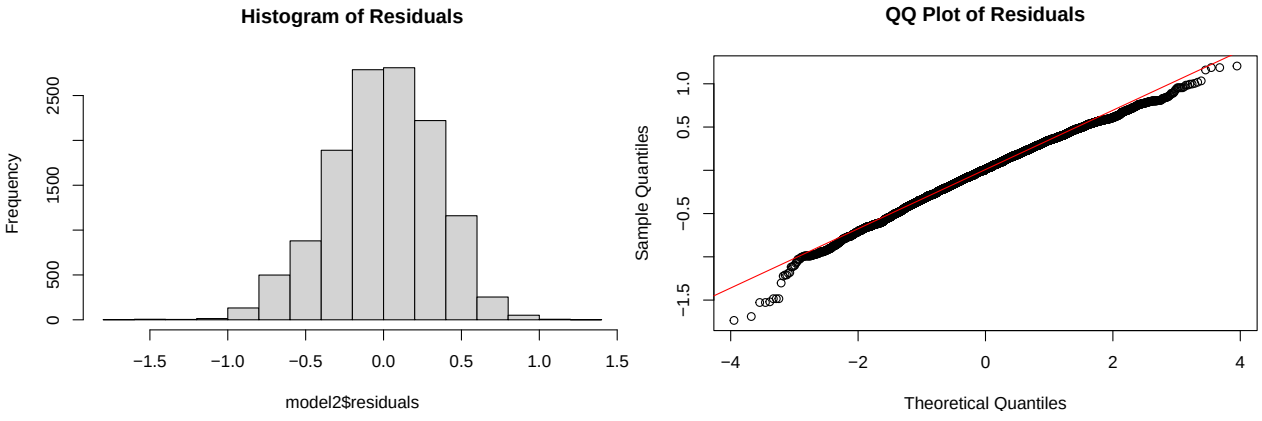
Back to our initial question: has the growth of AI influenced job titles and salary trends in the data science field? The answer is both yes and no. Our initial hypothesis was only partially correct in that AI-specific roles tend to earn a higher salary than traditional (or "Other") roles. However, by Figure 11 as well as the result of our t-test for H01, it does not appear as if the demand for AI-specific roles has increased more significantly than the demand for traditional data science roles. This was a surprising revelation considering the recent boom in the AI industry.

At the end of the day, AI/ML is relatively new field within a relatively new industry. It is understandable why one may feel hesitant about hiring a new data scientist or ML engineer since "new" means unpredictable. In this economy, making just one extra expenditure bring disaster to a business. However, even though the demand for AI-specific jobs may not be increasing right now, that does not mean they never will. We recommend that those seeking to pursue a career in data science *and* prefer higher salaries aim for job titles with words like "ML", "AI", or even "Scientist" as compared to "Analyst" or "BI." However, experience plays a significant role in salary as well, so staying in the industry for several years will be beneficial.

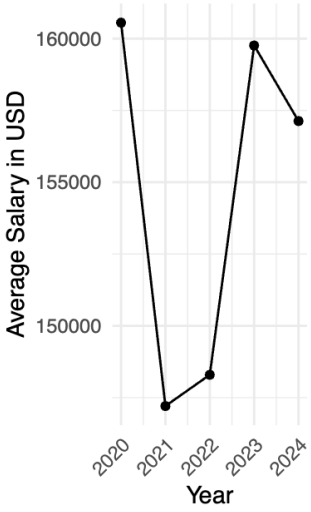
In order to conduct further research on this topic, it is important to account for external factors such as inflation, the pandemic, and market trends. Events such as hiring freezes/over-hiring and recession can cause unexpected fluctuations in the data. Additionally, we would like to gather more data on salary trends before 2022 since our data for 2020 and 2021 was already limited. Additionally, understanding what the industry and economy were like pre-pandemic may aid in creating a more efficient model, especially if the economy was more stable then than it is currently.

# Appendix

### Check for Normality



**EDA**



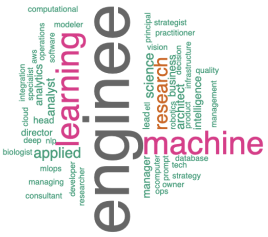




Figure 5: Wordcloud of Job Titles for High Salary


Figure 6: Wordcloud of Job Titles for Low Salary

**Citations**

GeeksforGeeks. (2024, October 30). Top career paths in machine learning. GeeksforGeeks. Retrieved December 2, 2024, from https://www.geeksforgeeks.org/top-career-paths-in-machine-learning/

Delikkaya, Y. (2024). Data science salaries 2024 [Dataset]. Kaggle. Retrieved December 2, 2024, from https://www.kaggle.com/datasets/yusufdelikkaya/datascience-salaries-2024