Author: Thanh Tin Lam

Course: STAT 482 – Capstone Project

Date: 12 – 05 – 2025

FROM CLASSROOM TO CAREER: SALARY TRENDS OF ENTRY-LEVEL ANALYSTS

I. Overview

1.1 Introduction

The technology labor market has changed rapidly in recent years, and the role of the Data Analyst has emerged as one of the most common pathways into data science. Companies rely heavily on analysts to interpret information, build dashboards, and provide insights for decision making. Yet compensation for these roles is not uniform and can vary significantly depending on job demand, experience level, and work arrangements. Understanding how salaries evolve, particularly for new graduates entering at the entry level, is critical both for individuals preparing to join the workforce and for employers aiming to maintain fair and competitive pay structures.

This project examines salary trends for Data Analysts from 2020 to mid-2025, focusing specifically on medium-sized companies (50 to 250 employees). This time period is notable because it spans major disruptions to the labor market, including the COVID-19 pandemic, the normalization of remote work, and the rapid adoption of artificial intelligence. By analyzing this five-year window, the study seeks to determine how these external forces shaped salaries and whether compensation growth has been shared equitably across different career stages.

1.2 Importance of this Study

The focus on entry-level positions is intentional, as they set the foundation for career earnings. Disparities at the start of a career can compound over time, leading to inequalities in lifetime income. For students and new graduates, understanding how salaries evolve is essential for setting realistic expectations and planning career strategies. For employers, these findings provide a benchmark to evaluate compensation practices and remain competitive in attracting young talent. For policymakers and educators, the evidence can guide workforce development initiatives and ensure that educational investments translate into fair labor market outcomes.

This study therefore seeks to not only describe the current salaries of new graduates, but also compare them to other levels to answer the question of whether new entrants are keeping up with market growth or are they at risk of falling behind their more experienced peers. In doing so, it provides insights that go beyond individual career choices to address broader questions of equity and sustainability in the data industry, and it also informs the appropriate path for new graduates to reach attractive salaries.

II. Data and Methods

2.1 Data Description

The data used in this study comes from the *Global Tech Salary Dataset (2020–2025)*, a comprehensive compilation of more than 150,000 job records across the technology industry. For the purpose of this project, the dataset was filtered to focus specifically on:

- **Country:** United States
- **Company Size:** Small-sized (less than 50 employees), Medium-sized organizations (50–250 employees), Large-sized (greater than 150 employees).
- **Job Titles:** Data Analyst–related roles, including *Data Analyst* and *Statistician* (standardized into a single occupation group to avoid fragmentation)
- **Experience Level** (Entry, Mid, Senior, Executive)
- **Time Window:** 2020 to mid-2025
- **Remote Work Status** (Onsite, Hybrid, Remote)
- **Valid Salary Records:** Observations with complete and plausible salary information

These variables allow for a structured comparison across time, experience levels, and work arrangements. The dataset includes a mix of continuous, categorical, and ordinal variables, enabling both descriptive and inferential statistical analyses.

2.2 Data Preparation

Multiple filtering and preprocessing steps were applied prior to modeling:

1. **Standardizing Job Titles**

   Variants such as “Data Analyst,” “Analyst,” “Data Scientist,” and “Statistician” were consolidated into a single *Data Analyst* category to avoid inconsistencies caused by naming differences.

2. **Experience-Level Categorization**

   I divide it into 2 parts: Entry-level and Non-entry level (including Mid, Senior and Executive)

3. **Missing Values**

Due to the lack of input data for 2020, we leave this gap visible in the figures to reflect the limitations of the dataset. This approach avoids data fabrication. However, in future studies, estimation techniques such as linear interpolation or model-based forecasting could be applied to provide more continuous insights.

2.3 Method

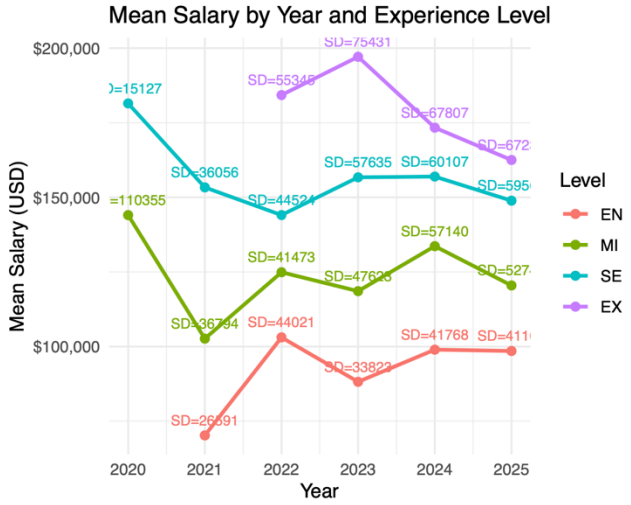
- **Exploratory Data Analysis (EDA)**: The analysis begins with descriptive statistics and visualizations to understand general patterns in salary distributions. Line charts, violin plots, and summary tables are used to compare salary behavior by experience level and by year.

- **ANOVA**: the main statistical method used to test for differences in salaries by year, company size, and telecommuting status. The three-way ANOVA model allows the study to assess not only the main effects of each factor, but also their interaction effects.
- **Gini Coefficient for Salary Inequality**: This method provides insight into the structural fairness of wages rather than simple averages.
- **Random Forest Modeling and Forecasting**: A Random Forest regression model was used to explore non-linear relationships and forecast salary trends for the period 2026–2028. This method allows the model to capture complex interactions that linear ANOVA models may not detect.

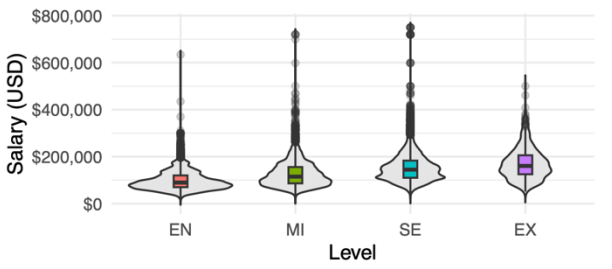
III. Exploratory Data Analysis

3.1 Overall Salary Summary

The visualizations provide a clear overview of the differences in salaries across experience levels and how they will change from 2020 to 2025. The median salary gap between levels is significant, reflecting increased responsibility and negotiation power at higher levels.



exp_std	n	mean	median	sd	se	min	max
EN	7354	$98,414	$89,730	$41,261	$481	$12,000	$634,000
MI	8692	$126,575	$115,000	$54,881	$589	$23,000	$720,000
SE	20300	$152,784	$144,600	$59,339	$416	$21,099	$750,000
EX	532	$171,641	$160,340	$68,760	$2,981	$51,352	$500,000



The median pay gap between levels is quite large, reflecting increased responsibility and bargaining power at higher levels. Year-over-year, all experience levels show a significant decline in 2021, consistent with post-COVID-19 labor disruption but requiring analysis to see the impact. Salaries recover after 2022, with Senior and Executive positions recovering faster than entry-level positions. The standard deviation (SD) printed above each point emphasizes that volatility increases with experience—senior and executive positions not only earn more, but also have a much wider range in pay.

3.2 ANOVA Results for Entry-Level Salaries

The central approach of this study is to apply analysis of variance, modeling salary by year, company size, and remote work status. This allows me to examine not only the main effects of each factor, but also their combined effects, such as whether salary growth over time differs depending on company size or remote work status.

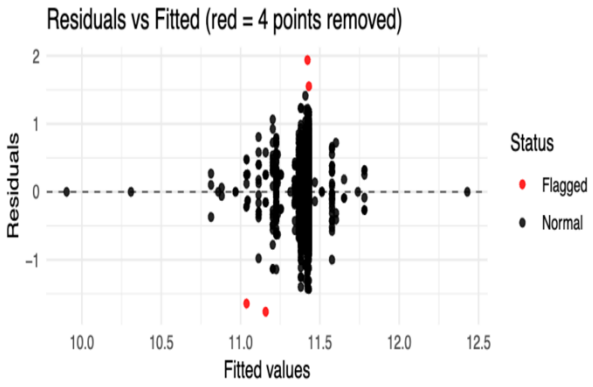
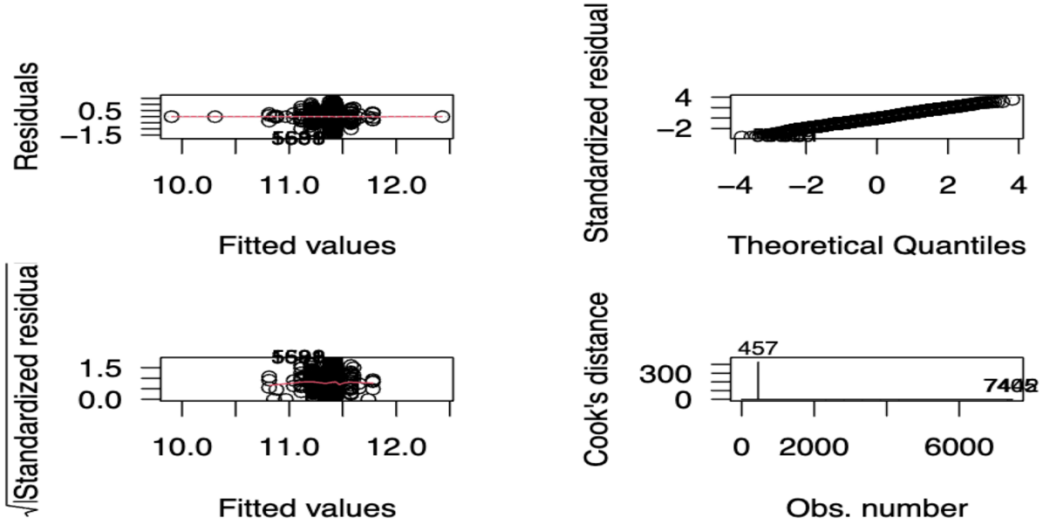


Table 1: ANOVA after removing top-K extreme points

term	df	sumsq	meansq	statistic	p.value
year	5	4.1497	0.8299	5.3544	0.0001
remote	2	1.7685	0.8843	5.7050	0.0033
size	2	0.2435	0.1218	0.7856	0.4559
year:remote	8	2.5883	0.3235	2.0873	0.0335
year:size	8	3.0853	0.3857	2.4882	0.0108
remote:size	3	2.4187	0.8062	5.2014	0.0014
year:remote:size	4	2.3545	0.5886	3.7976	0.0043
Residuals	7428	1151.3403	0.1550	NA	NA

Entry: Statistically significant coefficient estimates (p smaller than 0.05)

term	estimate	std.error	statistic	p.value	conf.low	conf.high	pct_change	pct_low	pct_high
(Intercept)	11.5309	0.3414	33.7791	0.0000	10.8618	12.2000	10181248.5535	5214561.8835	19878438.7596
remoteHybrid:sizeL	-0.8129	0.2947	-2.7580	0.0058	-1.3906	-0.2352	-55.6429	-75.1073	-20.9585



The ANOVA results indicate that year and remote work mode significantly affect entry-level salaries, while company size alone is not statistically significant. However, several interaction terms particularly between remote work and company size are highly significant, suggesting that salary differences depend on how these factors operate together. Coefficient estimates show that large companies pay more on average, but this advantage is reduced or reversed for hybrid roles. Diagnostic plots confirm that model assumptions are satisfied after removing four influential outliers.

3.3 ANOVA Results for Non–Entry-Level Salaries

Outliers were identified and removed to improve the reliability of the analysis of variance (ANOVA). Extreme observations can distort the model. In this dataset, the residual diagnostics detected a few observations with unusually large standardized residuals, indicating a disproportionate influence on the fitted model. These observations were marked as outliers and removed. After removal, the residuals were more concentrated around zero and evenly distributed, indicating that the model assumptions were better met and the resulting parameter estimates were more reliable.

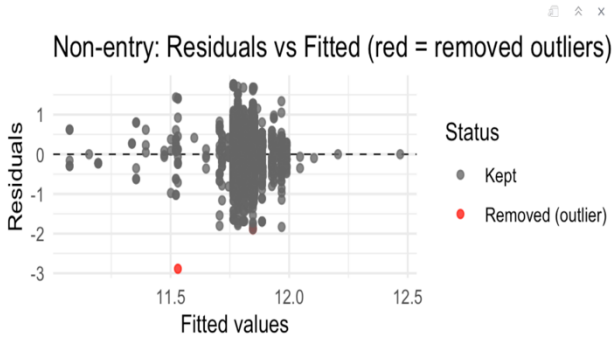


Table 1: Non-entry: ANOVA after removing outliers

term	sumsq	df	statistic	p.value
year	27.2969	5	35.7808	0.0000
remote	6.6552	2	21.8092	0.0000
size	6.7143	2	22.0027	0.0000
year:remote	3.0731	10	2.0141	0.0280
year:size	5.0192	9	3.6551	0.0001
remote:size	1.8784	4	3.0778	0.0152
Residuals	4571.8526	29964	NA	NA

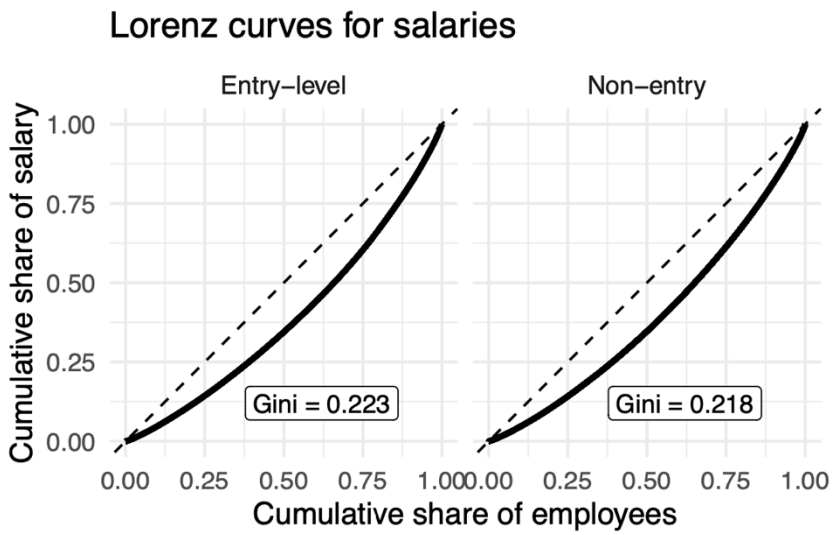
Table 2: NON-entry: Statistically significant coefficient estimates (p smaller than 0.05)

term	estimate	std.error	statistic	p.value	conf.low	conf.high	pct_change	pct_low	pct_high
(Intercept)	11.5106	0.3389	33.9653	0.0000	10.8463	12.1748	9976250.9210	5134360.9190	19384131.2307
sizeL	0.5255	0.2607	2.0159	0.0438	0.0146	1.0365	69.1349	1.4666	181.9314
remoteHybrid:sizeL	-0.8473	0.2927	-2.8952	0.0038	-1.4210	-0.2737	-57.1449	-75.8525	-23.9443

Although the ANOVA table includes several predictors and interactions, I only show those that show a statistically significant effect ($p < 0.05$). A factor such as year × size might be significant in the ANOVA because, taken as a whole, it explains a meaningful portion of the wage variation. However, not every individual coefficient in that interaction (e.g., each specific year-size combination) achieved individual significance in the regression results.

3.4 Salary Inequality: Gini Coefficient

The Gini coefficient is a statistical measure used to quantify income or salary inequality within a population. It ranges between 0 and 1, where 0 indicates perfect equality (everyone earns the same salary), and 1 indicates perfect inequality (one person earns everything). In this project, we use the Gini coefficient to measure salary inequality among data-related occupations (2020–2025), separately for Entry-level and Non-entry groups. Higher Gini values imply greater within-group disparity in salaries.



$$G = \frac{2\sum_{i=1}^{n} i\, x_{(i)}}{n\sum_{i=1}^{n} x_{(i)}} - \frac{n+1}{n},$$

The Gini coefficients are 0.223 (Entry-level) and 0.218 (Non-entry), indicating very low salary inequality. Both groups show nearly equal Lorenz curves, meaning salary distribution is fairly uniform across experience levels.

3.5 Random Forest and Forecasting Results

A Random Forest model was applied to explore potential non-linear relationships in salary prediction that linear models such as ANOVA may miss. This model is expected to improve prediction accuracy and identify key variables driving salary differences.

Table 3: Random Forest Summary for Entry-level and Non-entry Groups

Group	Best mtry	Min node size	RMSE (log)	MAE (log)	R^2 (log)	RMSE (USD)	MAE (USD)	R^2 (USD)	Top1 Var	Top2 Var	Top3 Var
ENTRY	2	3	0.3928	0.3159	0.0033	40569.31	30254.42	-0.0236	year	remote	size
NON_ENTRY	2	3	0.3936	0.3153	0.0072	58319.64	43841.56	-0.0143	year	remote	size

After running the model, low RMSE and MAE values were achieved but very small R^2, indicating limited improvement in overall prediction performance. However, the model consistently ranked year, telecommuting status, and company size as the most important predictors.

These results show that although non-linear effects exist, they explain very little additional variation in salary, confirming that the key drivers of salary remain the same between the Entry and Non-Entry groups.

The forecast shows that salaries will remain stable from 2026 to 2028. For entry-level employees, mid-sized (M) companies with on-site work offer slightly higher expected salaries than other combinations—a departure from the expected trend that large companies typically lead.

Recent graduates looking for higher salaries can consider on-site positions at mid-sized companies or remote positions at larger companies.

IV. Discussion

This study provides a comprehensive look at the evolution of Data Analyst salaries from 2020 to mid-2025, as well as how these trends are expected to change over the next few years. By integrating exploratory analysis, ANOVA models, inequality indices, and forecasting techniques, the study results reveal several important patterns relevant to both job seekers and employers.

4.1 Interpretation of Entry-Level Trends

Exploratory analysis shows that entry-level salaries have declined significantly in 2021, consistent with the general post-COVID labor market uncertainty. This decline is followed by a steady recovery in 2022 and 2023, suggesting that demand for data talent has recovered as companies resume hiring and business operations return to normal. However, the increasing standard deviation across years suggests increasing variability in the salaries employers are willing to pay new graduates.

The ANOVA results further support these patterns. The year effect is statistically significant, confirming that wage changes over time are not random but reflect real changes in hiring conditions. Remote work status also plays a role: remote and hybrid work arrangements generate much larger wage differentials

than on-site positions, likely due to differences in geographic hiring pools and internal pay scales. While company size is not significant, the correlation between company size, remote work status, and years of employment is significant. This suggests that wage trends cannot be interpreted in isolation—the effect of company size depends heavily on how and where work is performed.

4.2 Understanding Forecasted Patterns

Predictive analysis using Random Forest models provides additional insight into future salary expectations. While the model achieves low R^2 values – reflecting the difficulty of predicting salaries using non-linear methods – it consistently identifies year, employment status, and company size as the most important predictors, which is consistent with previous ANOVA findings.

Overall, the forecasts suggest stable salaries rather than sharp growth or decline from 2026 to 2028. Entry-level salaries appear to be normalizing after the volatility of the early 2020s. Interestingly, mid-sized companies with on-premises positions show slightly higher expected salaries for new graduates. This finding challenges the assumption that larger companies always pay higher salaries and suggests that mid-sized companies may be strategically positioning themselves to attract younger talent. Remote positions at large companies also offer competitive salaries, but come with more uncertainty.

4.3 Limitations

There are some limitations to be aware of when interpreting these findings. The lack of data on starting salaries in 2020 creates an unavoidable gap in the early stages of the trend. Regional differences are not captured in the dataset, which may affect salary dispersion. Additionally, while the Random Forest models help identify important predictors, their weak explanatory power means that the salary forecasts should be interpreted as general directional models rather than precise predictions.

V. Conclusion

This study analyzed the salary changes of entry-level Data Analysts from 2020 to mid-2025 and projected their future trajectory through 2028. The data showed that salaries declined during the post-pandemic disruption of 2021 but recovered and stabilized thereafter. However, salary volatility has increased in recent years, especially for remote positions, suggesting a more diverse and competitive remote work landscape. ANOVA results confirmed that year and remote work status significantly influenced salaries, while company size was important primarily through its interaction with work arrangements.

The salary projections suggest a period of steady but moderate growth from 2026 to 2028. On-site positions at mid-sized companies appear to offer the most stable and slightly higher expected salaries, while remote positions remain more difficult to predict. While the Random Forest model has limited predictive power, the forecasts are consistent and consistent with observed historical trends.

Overall, the results suggest that there are still plenty of job opportunities for entry-level employees, but new graduates should be aware of salary variations across employment types and company structures. Developing strong technical skills and considering on-site positions can help new analysts secure more stable salaries.

VI. References

- Kaggle (2025). *Global Tech Salary Dataset.*
- U.S. Bureau of Labor Statistics (2024). *Earnings by Occupation and Experience Level.*
- LinkedIn Workforce Reports (2022–2024).
- Plus any method references (e.g., Scikit-learn, R packages, JMP manuals).