Exploring post-2020 Mass Layoffs

Siddhant Joshi, Evelyn Huang, Arnav Kamra

June 9, 2024

Abstract

Analyzing job market dynamics, particularly layoff trends, is crucial for understanding employment fluctuations over time. This study examines data from 2020 onwards to identify factors influencing layoff rates across various industries, with a focus on the tech sector. Our comprehensive approach draws on data from the Layoffs.fyi—a platform tracking layoffs since the COVID-19 pandemic—including company names, industries, locations, and economic indicators. We employ a comprehensive suite of statistical techniques, including χ^2 tests, two-sample t-tests, Pearson and K-S tests, linear regression models, and a decision tree classifier. Our results reveal that while the tech industry's layoff rate does not significantly differ from other industries, nontrivial associations between funding, stage, company size, and layoff rates can be used to model trends across a diverse range of industries. Our findings contribute valuable insight for companies, policymakers, and researchers to develop proactive strategies that can mitigate current unemployment risks.

1 Introduction

Recent years have seen significant fluctuations in employment across various industries, driven by a myriad of factors, such as higher interest rates, economic slowdown, and external shocks like the COVID-19 pandemic. Since the onset of the pandemic, 2020 marked the start of an international wave of mass layoffs across many companies in various industries. As of 2022, reports showed that over 120,000 people had been dismissed from their job at some of the biggest tech giants—Meta, Amazon, Netflix, and Google—with more to come.[2,4] In fact, a closer look reveals at one point, Meta fired 13% of its workforce, which amounts to more than 11,000 employees.[4] Research shows that laying off employees does not reap immediate financial benefit for companies and leads to several negative impacts on the population, such as an increase in mortality rates via depression and suicide.[1,3] Unsurprisingly, this prevailing response to terminate employees in the face of economic instability is a point of concern that should be rigorously addressed, which is precisely what this study aims to highlight. Analyzing these trends not only sheds light on the immediate impacts on workers and companies but also offers predictive insights that can help shape future employment strategies.

This paper will delve into the intricacies of job layoffs by examining data from 2020 onwards. By correlating economic performance, geographic location, and layoff rates, we aim to uncover how these factors influence recent corporate employment decisions. This comprehensive approach not only enhances our understanding of job layoffs but also contributes to the knowledge needed to formulate strategies that are essential for mitigating unemployment risks. Our guiding question for this analysis is "What factors influence layoff rates across different industries, and how can we effectively model these rates using statistical methods?"

2 Data

2.1 Source

The two datasets we used for this study are accessible from Kaggle, containing different information from the same source. Both contain layoff metrics sourced from a variety of publicly available news articles and industry databases, compiled by an online layoff monitoring site called Layoffs.fyi. For most up-to-date data, it is best to download from the site directly as it tracks layoff rates in real-time. We opted for combining these two datasets together so as to have a larger, complete dataset for downstream analysis.

2.2 Generation Process

The main collection site provides information on which websites it scraped to formulate each observation with links to the original databases, articles, or reports. The reliability of these sites is unknown, but upon a brief scan of some of the sites listed many were sources that are generally regarded as credible, including the U.S. Bureau of Labor Statistics, The Wall Street Journal, Bloomberg, and Statista.

2.3 Cleaning

Any observations with null values present in the response variables were dropped so as to avoid introducing bias into the data. Furthermore, any missing information regarding industry or location were determined with light research into the company. Lastly, missing values in the state and fund features were imputed based on probability distribution within an industry and stage-wise means, respectively. Features regarding the data source, date added, and list of employees who were laid off were omitted. Below are the finalized features in the cleaned data:

Feature	Description
company	Name of company
hq_location	Location of company headquarters
industry	Company industry
laid_off_count	Number of employees laid off
laid_off_percent	Percentage of employees laid off
date	Year, Month, Date of the layoff
stage	Stage of company funding
country	Country of the company headquarters
fund	Funds raised by the company (in millions of dollars)

2.3.1 Feature Engineering

Since company size is relevant to later analysis and the laid_off_count and laid_off_percent features provide redundant information, we will calculate a new company_size feature using laid_off_count and laid_off_percent, then drop the laid_off_count feature. Additionally, we divide the date feature into year and month so as to access seasonal and annual trends.

2.4 Exploration

After imputation and cleaning, there were 1341 unique companies from 30 different industries, and these companies were from 45 different countries and were in 16 different stages. We start with exploring some distributions between numerical features.

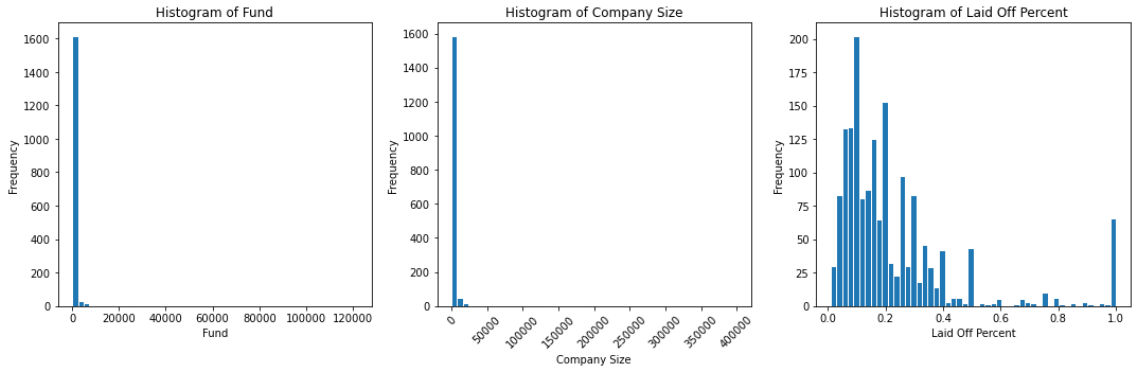

Figure 1: Histograms of fund, company size, and laid-off percentage

As expected, the data are extremely right-skewed, hinting that we may need to drop some outliers in order to get a clearer picture of the trend and model it effectively. We ignore the top 97th-percentile for fund and company size to get a better look at the majority of data:

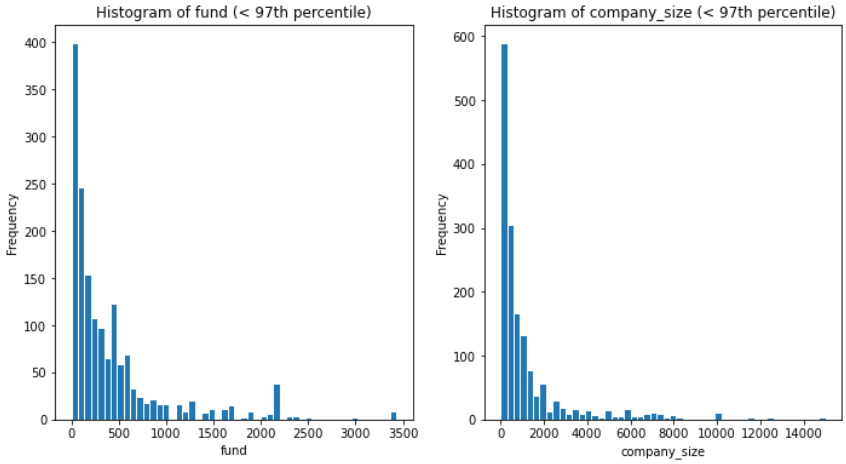

Figure 2: Removing the top 97th-percentile of observations gives us a better look at the majority of data

From here, we can see that the data are still quite right-skewed and the majority of companies in our dataset have both funding and company sizes tending toward the smaller end. Nevertheless, funding seems to range between \$0 and \$4000 million while company sizes range from 0 to 8000 employees. More nuanced subsetting of the data reveals that the top 5 companies in terms of size are Amazon, Google, IBM, Microsoft, and Tesla, descending. Next, we examine the distribution of dates via the following histogram.

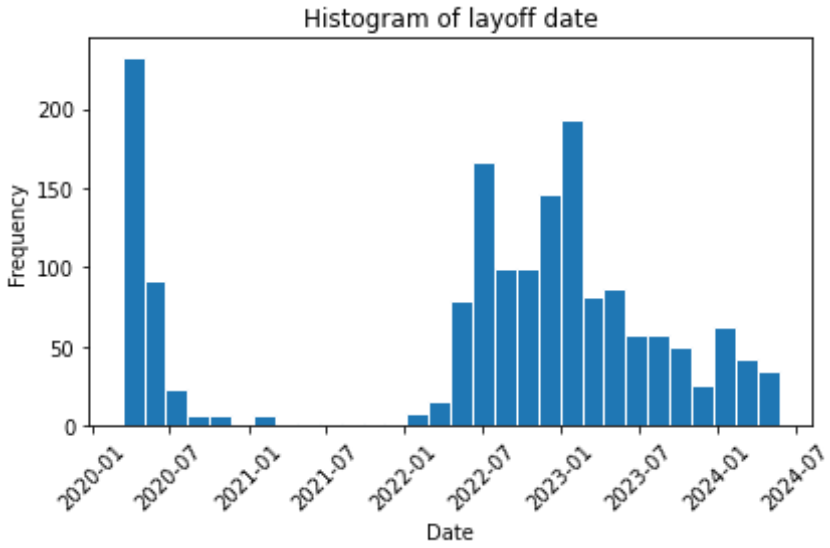

Figure 3: Histogram of layoff dates

We see a spike in layoffs in mid-2020 and then a more persistent wave taking place between 2022 to the beginning of 2024. It is also important to note the sharp drop in layoffs between mid-2020 through the start 2022 which corresponds to the height of the COVID-19 pandemic. Lastly, we take a deeper dive into the trend of layoff rates over time, plotting layoff dates against the average layoff rate for that date.

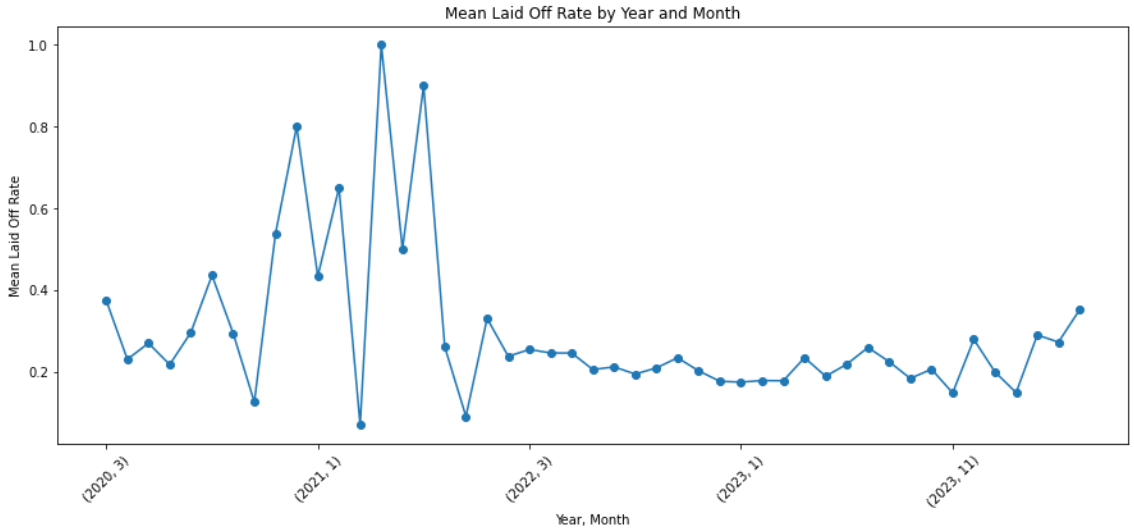

Figure 4: Average layoff rates over time

At face-value, this plot is deceptive as it suggests that the most intense layoff rates were from mid-2020 through 2022. However, this is not true when assessed in the context of Fig. 3 which tells us that the least amount of layoffs took place during this period. The volatility of the graph, therefore, is a result of the lack of observations during that time period, whereas the average layoff rate is more tempered everywhere else due to the high number of companies laying employees off. This means post 2021 trends are the most reliable for analysis and the increase toward the right-end of the plot is more meaningful evidence of an increase in mass layoffs.

3 Analysis

3.1 Analyzing the Tech Industry

3.1.1 χ^2-test for Independence

We start with investigating the role of industry in layoff rates, as tech layoffs are a key point of discussion in recent reports. We run the following χ^2-test for independence to test for a relationship between industry and layoff percentage. We achieve a $\chi^2 = 2272.35$ with a p-value of 0.0096. In this case, since $0.0096 < 0.05$ we reject the null hypothesis at the 5% significance level, meaning there is evidence to suggest a statistically significant association between industry and the layoff rates.

3.1.2 Two-sample t-test

To further examine this, we conduct a two-sample t-test between tech industries and all others. For the purposes of this study, we defined the tech industry as a combination of the AI, crypto, and data industries. Testing the layoff percentages of these two groups yields $t =: 0.0711$ and a p-value of 0.94, which is greater than the 5% significance level. This means there is no statistically significant difference in layoff percentages between the tech industries and the other industries. We can also visualize this with the following plot of layoff rate distributions between the two groups.

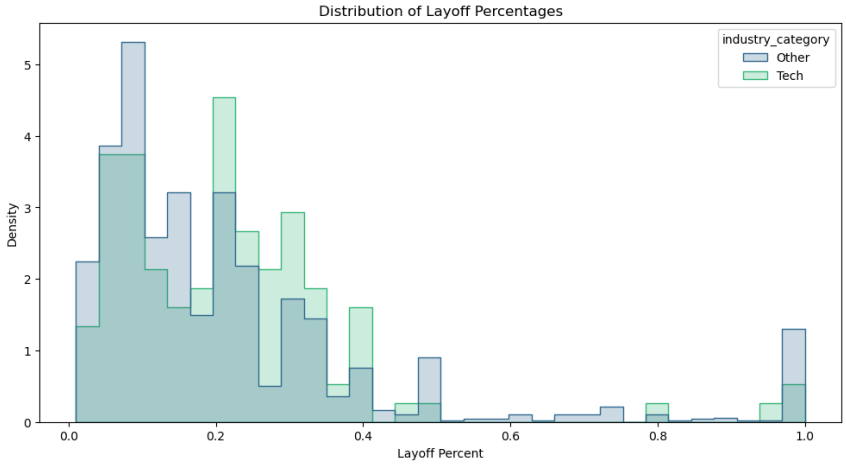

Figure 5: Distribution of layoff rates for tech companies and all others, overlain

3.2 Correlation Tests

3.2.1 Pearson Correlation

We calculated a Pearson's correlation coefficient with the numerical features to assess the existence and strength of linear association. The following table contains the results and their respective significance values.

Feature	Pearson's r	p-value
fund	-0.067	0.0062
company size	-0.065	0.0073
year	-0.132	5.589×10^{-8}
month	-0.008	0.744

We observe a weak negative linear correlation across all features, however the correlation between month and layoff rate is not as statistically significant as the others. Therefore, funding, company size, and year are likely to be related to the layoff rate based on their significant correlations, while month does not appear to be associated in this analysis.

3.2.2 Kolmogorov-Smirnov Test

Next, we conduct a K-S test on our categorical features to determine any significant relationships with the response variable. Using the bootstrap approach, we generate the following significance values for the features in question.

Feature	p-value	< 0.05
HQ location	0.045	Yes
industry	0.048	Yes
stage	0.005	Yes
country	0.596	No

With a significance level of 5%, we conclude that there is a statistically significant difference in the layoff rate across different levels of HQ location, industry, and stage. There is no statistically significant difference in the layoff rate and different countries, suggesting that it may not be a useful feature in our linear model.

3.3 Linear Regression

Given that initial histograms of the data show an extreme right-skew, we expect the data will violate most of the linear regression assumptions. To correct for this, we will apply a log-transformation to the response variable and conduct all diagnostics for the model with the transformed data.

3.3.1 Verifying Assumptions

We assess diagnostics with the full model. Starting with the normality assumption, we generate the following QQ-plot and observe that the data seem normally distributed and thus fulfill the normality assumption.

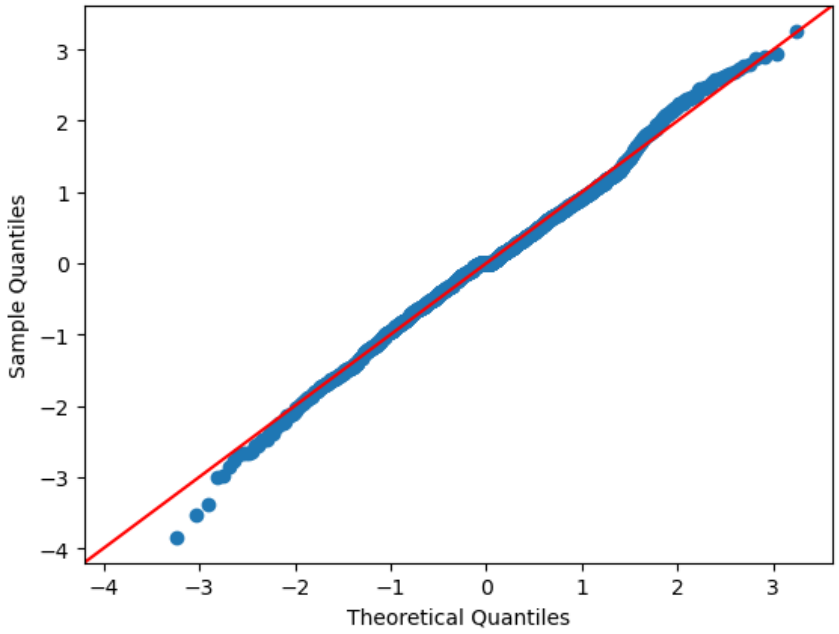

Figure 6: QQ-plot for testing normality of the transformed data

Next we test for independence in the residuals by plotting the residuals of the full model. Since we observe no inherent patterns within the plot and the residuals are randomly distributed, our data also fulfill the independence assumption.

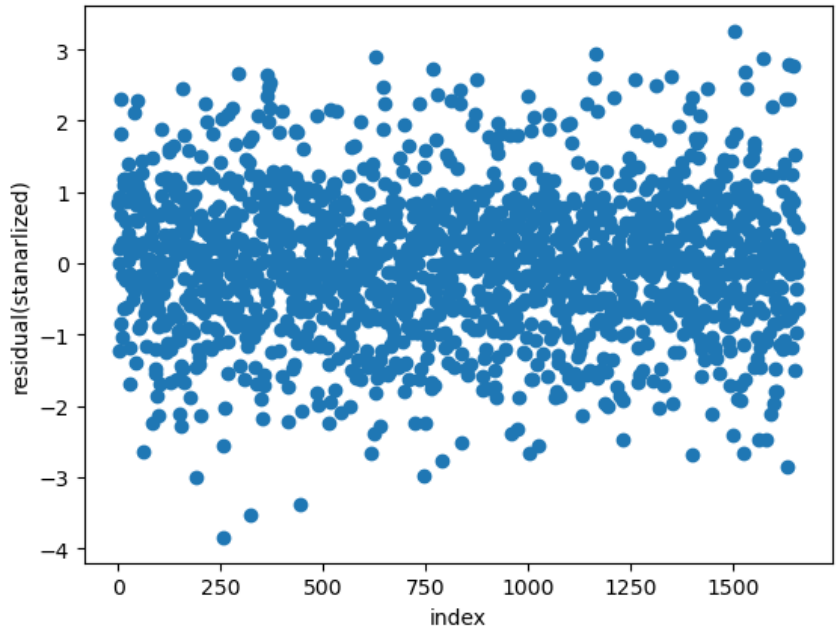

Figure 7: Plot of the residuals for each observation

To assess the linearity of the data, we can plot the residuals against each feature (Fig. 11). To get a better look, we temporarily drop the outlier residuals in funding and company size plots and observe there is indeed random scattering of the observations. Thus, the linearity assumption is met.

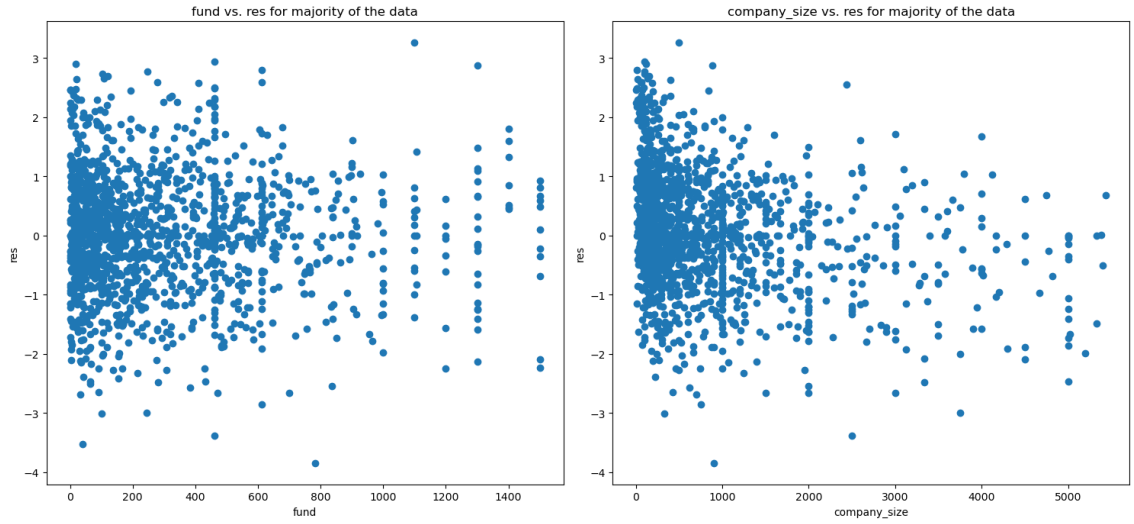

Figure 8: A closer look at the residuals for funding and company size, verifying the randomness in residuals that confirms the linearity assumption

Finally, we check for multicollinearity within our data. To do so, we calculate the variance inflation factors (VIFs) of each level of each feature and drop those with VIF values greater than 5, as outlined by standard practice. The only feature that violated this criterion was the month column, which was already omitted in previous Pearson correlation analysis.

3.3.2 Feature Selection

Now that we have verified the linear regression diagnostics in our data, we can start constructing the linear models. We start with assessing two models—a full model and one based on our hypotheses from the above tests—and generating a third model via rigorous feature selection using mixed selection.

The full model takes the form:

```
laid_off_percent = hq_location + industry + stage + country + fund + company_size + year + month
```

Our hypothesized model takes the form:

```
laid_off_percent = hq_location + industry + stage + fund + company_size + year
```

We conduct mixed feature selection with a BIC criterion to determine a third model to test in downstream cross-validation. This model takes the form:

```
laid_off_percent = stage + fund + company_size + year
```

3.3.3 Cross-validation

To select an optimal model, we process each of the three models from above via validation set cross-validation. The data are randomly subsetted into a 75/25 training and testing split, with mean squared error (MSE) as the decision criteria. Each model is fit to the training data and then the test set MSE is calculated. The following table reports the error metrics from each model.

Model	Test MSE
Full	276.62
Hypothesis	268.90
Mixed Selection	252.74

As expected, the mixed selection model is the most optimal model and performed better than the other two. However, the hypothesized model was not too far behind, validating our previous statistical testing.

3.4 Decision Tree Classifier

To further examine the relationship between these predictors and the layoff rates, we build and fine-tune a decision tree classifier. We opt for this over a logistic regression model since most of our features are categorical.

3.4.1 Data Preparation

Since the response variable is continuous, we bin layoff percentages into 5 sub-intervals; {(0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1]}. Additionally, all categories are one-hot encoded and the training and testing data are created on another random 75/25 split.

3.4.2 Hyper-parameter Tuning

We use grid search to tune hyper-parameters such as purity criterion, maximum tree depth, minimum samples to split a node, and minimum samples in leaf nodes. The values assessed in our grid are provided below, along with the optimal values determined by the grid search:

Hyper-parameter	Values	Optimal
criterion	gini index, entropy	gini
max_depth	2, 4, 6	4
min_samples_split	2, 5, 10	2
min_samples_leaf	1, 2, 4, 6	1

This optimal model achieves an overall accuracy of 62%, which is admittedly on the lower end. We believe this may be due to a couple reasons—class imbalances and feature quality. As was made evident in our exploration, it is not common for layoff rates to exceed 0.5, meaning class 1 and 2 were over-represented in the model. Additionally, the features that the model was provided may be too general and thus, it is unable to adapt to a more complex pattern. Having more information especially surrounding economic states and company health may be provide useful information with which to make more accurate predictions.

4 Conclusion

This study aimed to analyze the factors influencing layoff rates across various industries, utilizing comprehensive statistical approaches that included exploratory data analysis, statistical hypothesis testing, and predictive modeling. Our key findings are summarized in the next few paragraphs.

EDA provided initial insights into the distribution and trends of layoff rates across different industries. We also explored the relationships between layoff rates and several other features, seeing room to explore post-2021 relationships further. We conducted a χ^2 test to discover that there was a significant association between industry and layoff rates, only to find via a two-sample t-test that the tech industry did not have a significantly different layoff rate compared to other industries. This finding was contrary to our expectations and challenges the prevailing sentiment that the tech industry is uniquely affected by the recent mass layoffs.

Next, we assessed the linearity of relationships between features and the layoff rates, seeking to build a linear regression model to describe the trends we were seeing. We used a combination of Pearson correlation, Kolmogorov-Smirnov, and multicollinearity tests to determine some statistically significant candidates to be used in a hypothesized linear model. We then compared three linear models— a full model, our hypothesized model, and a mixed selection model—using validation set cross-validation. The results of this procedure showed that the mixed selection model performed the best. This model effectively identified the key predictors of layoff rates as company size, stage, funding, and year, providing a robust framework for understanding the factors driving layoffs. To complement the regression analysis, we developed a decision tree classifier on binned layoff rates. After tuning, the classifier achieved an accuracy of 62%. While this accuracy performs significantly better than the bonehead, it has room for improvement. Most notably, we suggest correcting for class imbalances and providing the model more economically-indicative features during training. Future research could explore nonlinear models or ensemble methods to improve prediction accuracy.

In relation to our original proposal for this project, we ended up simplifying our choices for analyses after noticing that a linear regression was quite effective for our purposes. We also opted for a decision tree classifier instead of clustering or temporal analysis as it made more sense as our analysis came to light. To conclude, we provide a multi-faceted analysis of recent mass layoff rates, highlighting the significance of industry-specific factors and contributing to a deeper understanding of what is causing companies to opt for seemingly erratic and detrimental decisions.

5 References

[1] Datta, Deepak & Basuil, Dynah & Radeva, Elena. (2013). Employee downsizing and organizational performance: What do we know?. 10.1017/CBO9780511791574.012.

[2] Peck, E. (2022, November 15). Tech layoffs are soaring this month [Review of Tech layoffs are soaring this month]. Axios.com; Axios. https://www.axios.com/2022/11/15/tech-layoffs-amazon-meta-twitter

[3] Sullivan, D., & von Wachter, T. (2007, November 1). Mortality, Mass-Layoffs, and Career Outcomes: An Analysis using Administrative Data. National Bureau of Economic Research. https://www.nber.org/papers/w13626

[4] What explains recent tech layoffs, and why should we be worried? (n.d.). News.stanford.edu. https://news.stanford.edu/stories/2022/12/explains-recent-tech-layoffs-worried

6 Supplemental Figures

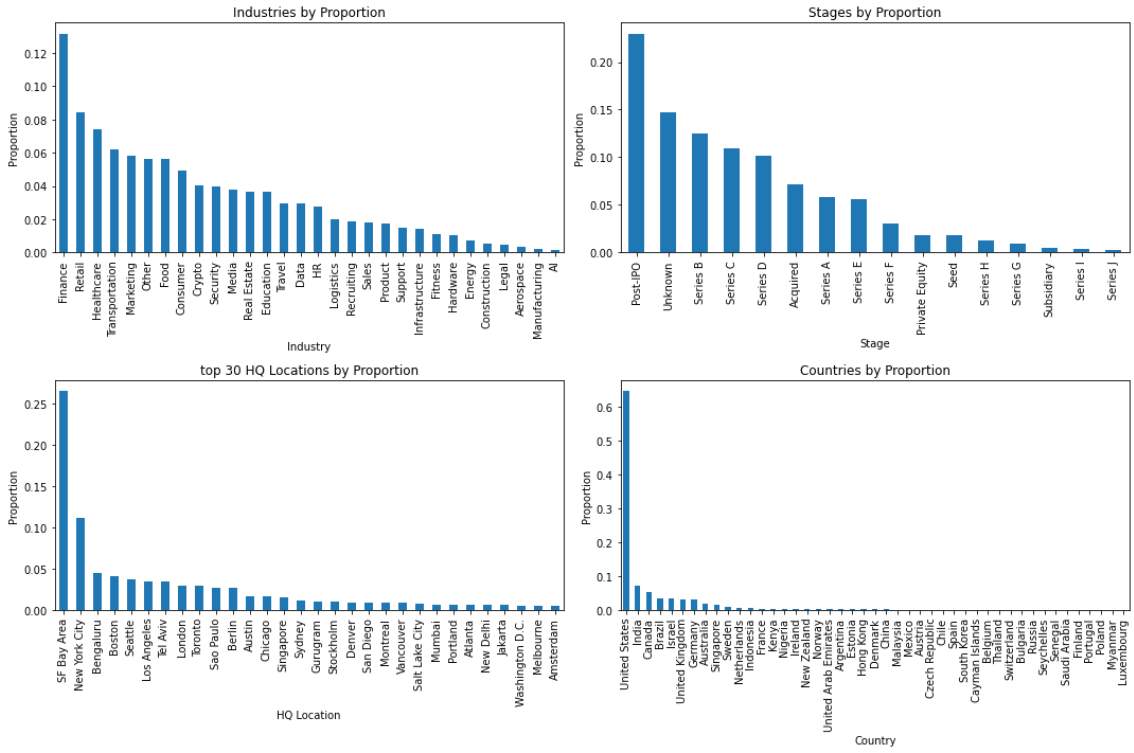

Figure 9: Categorical breakdown of companies based on industry, stage, HQ locations, and country

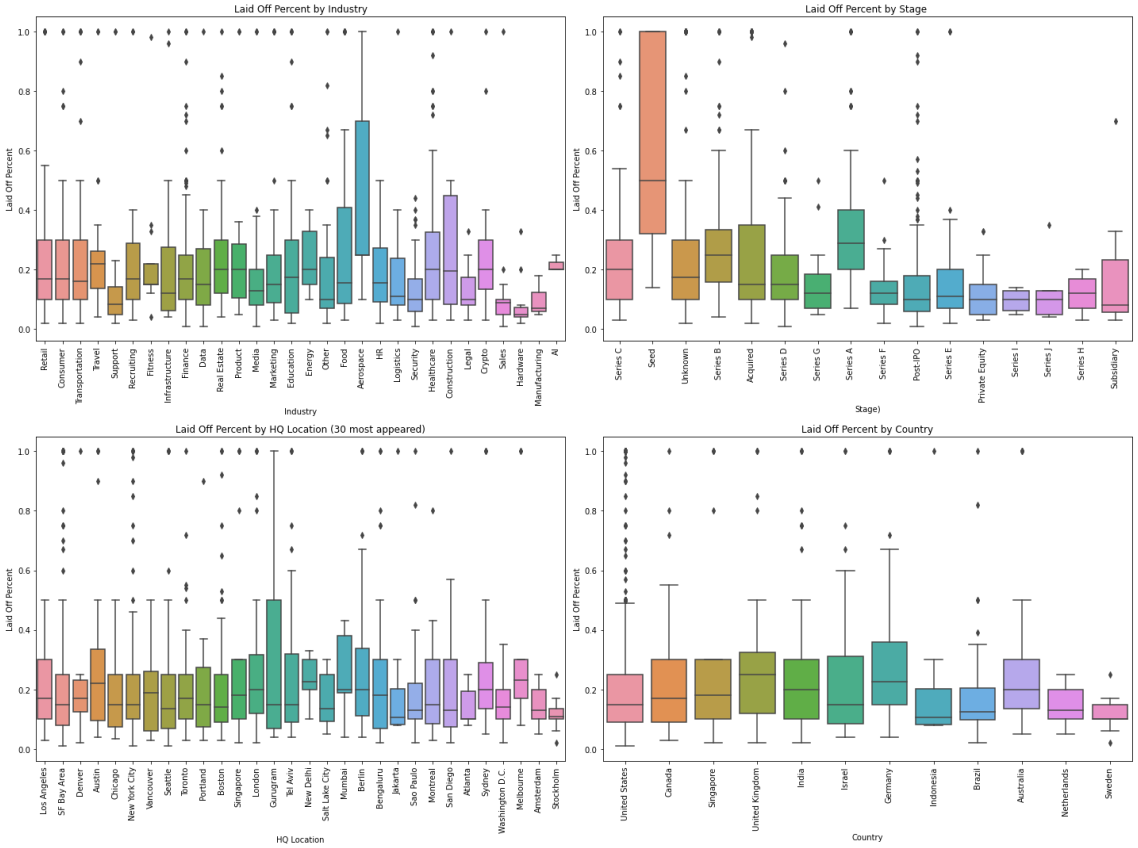

Figure 10: Box plots of layoff distributions based on categorical features

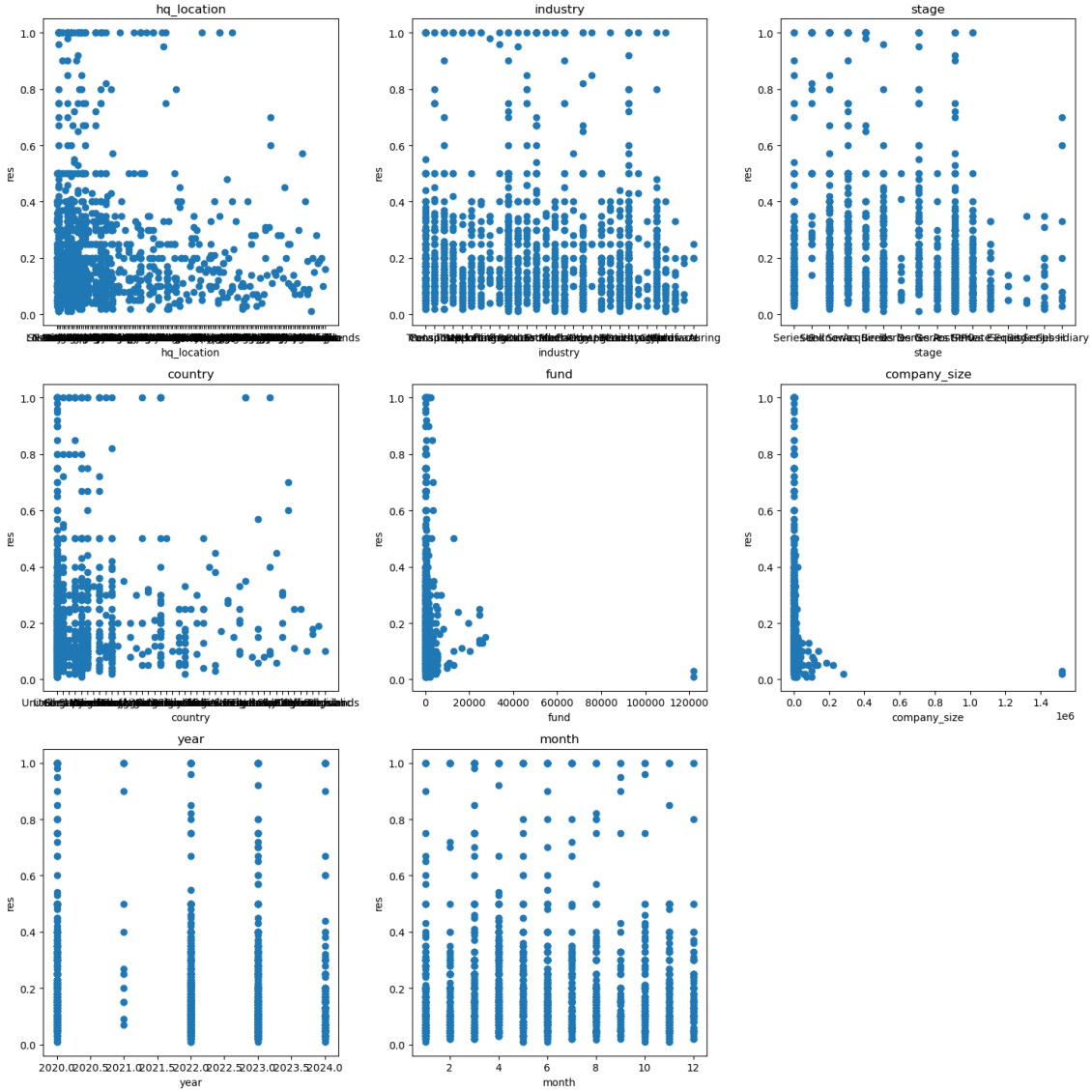

Figure 11: Residuals for each feature to test for linearity