Uncovering the Complex Relationship between Total Business Reviews and Ratings: A Polynomial Regression Analysis Using Skewed Normal Distribution

1 Introduction

The rise of online review platforms has revolutionized the way consumers make decisions and behave concerning to where to eat, shop, and conduct business. Among these platforms, Yelp has emerged as a dominant force, providing an unprecedented level of insight into the experiences of customers at a vast array of businesses. The wealth of data available on Yelp has led to countless insights and discoveries about consumer behaviour and business performance, but one question remains unresolved: is there a direct relationship between the number of reviews a business receives and its star ratings? This is a question that has been heavily assumed by many, with the widespread belief from users, and our hypothesis that a greater number of reviews leads to a higher rating. However, our research seeks to dismiss this hypothesis by showing the failure of monotonic and even simple non-monotonic relationships modelled using various polynomial regressions, revealing the complexity of the relationship between the variables.

2 EDA (Exploratory Data Analysis)

In this study, we utilized a smaller sample of the Yelp business dataset due to computational constraints. However, we believe that our findings are still representative of the broader population. Upon performing EDA, we remove any datapoints with missing or duplicate values to ensure the accuracy of our analyses. In order to select appropriate methodologies for testing our hypothesis, it’s crucial to learn about the behavior of the star ratings and business reviews. Figure 1 shows the histogram and Kernel Density Estimation (KDE) plot of the business star ratings distribution. KDE is a non-parametric approach to estimate for estimating the shape of a dataset's distribution. It involves the summation of functions, each of them associated with one observation, to generate a continuous probability density function from observations. Mathematically, the KDE is expressed as:

$$f(x) = \left(\frac{1}{nh}\right) \Sigma_{k=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where x is the value at which we want to estimate the density, x_i are the observed datapoints (star ratings in our context), K is the kernel function (usually Gaussian), h is the bandwidth parameter & n is total datapoints. The kernel function is a smooth, continuous function that is centred at each data point x_i.

So, from the plots in Figure 1, we found the distribution to be normal-like left-skewed, which suggests that most businesses have high ratings and only a few have low ratings. Interestingly, 4.5 stars is the most common rating observed. To model this (ratings) distribution, we propose a skewed normal distribution whose probability density function (pdf) can be expressed as:

$$f'(x) = \frac{2}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \phi\left(\frac{\alpha(x-\mu)}{\sigma}\right)$$

Here, x represents the star ratings, ϕ denotes the CDF (cumulative distribution function) and the parameters α, μ, and σ represent the skewness, location, and scale of the distribution, respectively. Based on the skewness of the distribution (-0.5644) and a visual inspection of the second plot in Figure 1, we believe that the skewed normal distribution is a suitable fit for our ratings dataset.

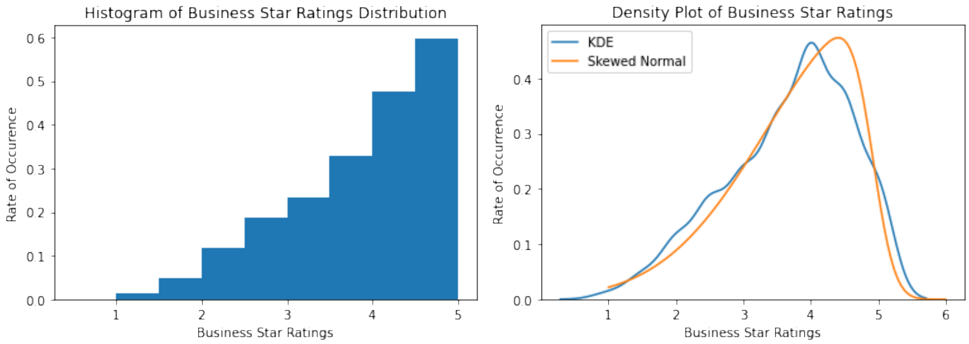

Figure 1: Histogram and Density Plots (KDE and Skewed Normal) of Star Ratings Distribution

Furthermore, the behavior of the business review count (or the total reviews received by each business) is found to be a power law distribution as seen in Figure 2 which is not surprising considering that few businesses attract more popularity and criticism than others. So, we apply logarithm scales to the business review count to improve the modelling of the relationship between it and the business star ratings.

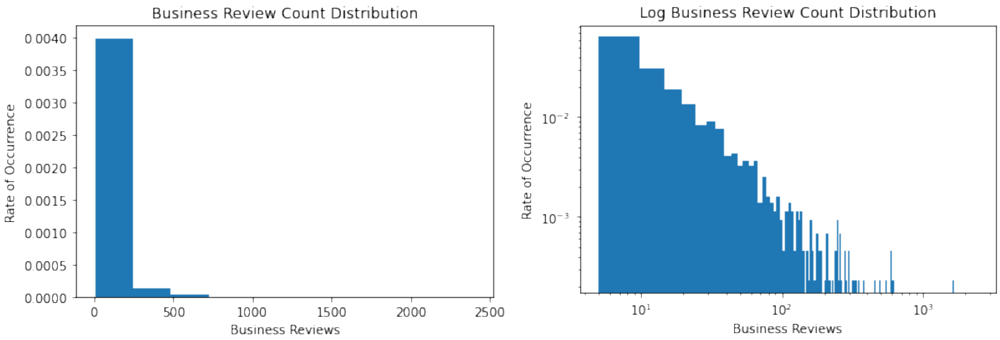

Figure 2: Histograms of Business Review and Log Business Review Distribution

Finally, we employ Pearson-r correlation test to determine whether the star ratings and business reviews are correlated or not. The formula for the Pearson correlation coefficient r_{xy} can be given by:

$$r_{xy} = \frac{1}{n-1}\Sigma_{\mathrm{i}=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where n is our total datapoints, x_i and y_i are the individual values of the two variables, the star ratings and the log business reviews, respectively. The $\bar{x}$, $\bar{y}$, s_x and s_y are the means and standard deviations of the two variables. The p-value associated with the test is used to determine if the correlation is statistically significant. With a p-value of 0.006, less than the significance level of 0.05, we found a statistically significant correlation between the log business reviews and star ratings variables.

3 Regression Models

We select polynomial regression models for our variables as we believe them to be the appropriate methods to support the evaluation of our hypothesis of simple or monotonic relationship between log business reviews and star ratings. We employ four models: linear regression, quadratic regression, simple cubic regression model, and a weighted cubic regression model using WLS (weighted least square) method based on weights of the skewed normal ratings distribution. The simple linear, quadratic and cubic regression equations can be given respectively as:

$$Y = \beta_0 + \beta_1 x + \epsilon$$
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

where $\beta_0, \beta_1, \beta_2, \beta_3$ are the coefficients of our models, ϵ is the error term, Y is our dependent variable (log business reviews), and x is the independent variable (star ratings).

We utilise a weighted cubic regression model to account for the left-skewed distribution of star ratings. Weighted least squares assigns weights to each observation based on the skewed normal distribution, giving more importance to less skewed observations. The weight for each observation is given by:

$$w_i = \frac{1}{\phi(\mathrm{x_i})\Phi(-\alpha)}$$

where $\phi(\mathrm{x_i})$ is the pdf of the skewed normal distribution at the i^{th} rating, and $\Phi(-\alpha)$ is the CDF of standard normal distribution at the negative skewness coefficient α. This model incorporates prior knowledge about the distribution of business star ratings, allowing for a better fit to our data.

4 Results and Conclusions

The coefficient of determination, also known as R-squared, represents the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. Amongst all models tested, linear regression had the lowest R-squared of 0.002, while simple cubic regression had the highest R-squared of 0.68. Interestingly, quadratic regression and weighted cubic regression had similar R-squared values. Although the principle of parsimony suggests selecting the simpler model, we choose to use the weighted cubic regression as it is designed to better adjust to the skewness of the star ratings. The primary inference from this is that non-monotonic relationships can produce better models than the monotonic relationship for our variables.

Figure 3 presents a 3D plot of both cubic regression models, depicting our data along the axes of the number of businesses, star ratings, and log business reviews. The visualisation reveals that a considerable number of businesses have received high star ratings (4 and 5) compared to low ratings. Additionally, we observe peaks in the number of businesses' axis for high-rated establishments when their total reviews are minimal. Our findings suggest that businesses with polar star ratings (1 and 5) tend to have fewer total reviews, whereas those with intermediate star ratings (3-4) demonstrate a wider range of business reviews, with a higher concentration of reviews in the centre. Lastly, the figure suggests that when a business receives more reviews, i.e. more user activity, their ratings tend to converge between the range of 3-4 stars.

We also find that both our cubic regression models coincide a lot through their modelling across the log business reviews and star ratings axes. However, we observe that our regression models are at a significant distance to the median values projections. This could be a result of the limited business data used, but it could also point out at the possibility of room for improvement in modelling the variables. The latter reason is confirmed through Shapiro-Wilk test performed to check whether the residuals are normally distributed or not. Both cubic regression models have p-values less than 0.05, which indicates that the residuals are not normally distributed. This violation of the assumption of polynomial regression indicates that the models are not a very good fit for the data. Thus, this dispels our hypothesis and proves the high complexity of the relationship between our variables.

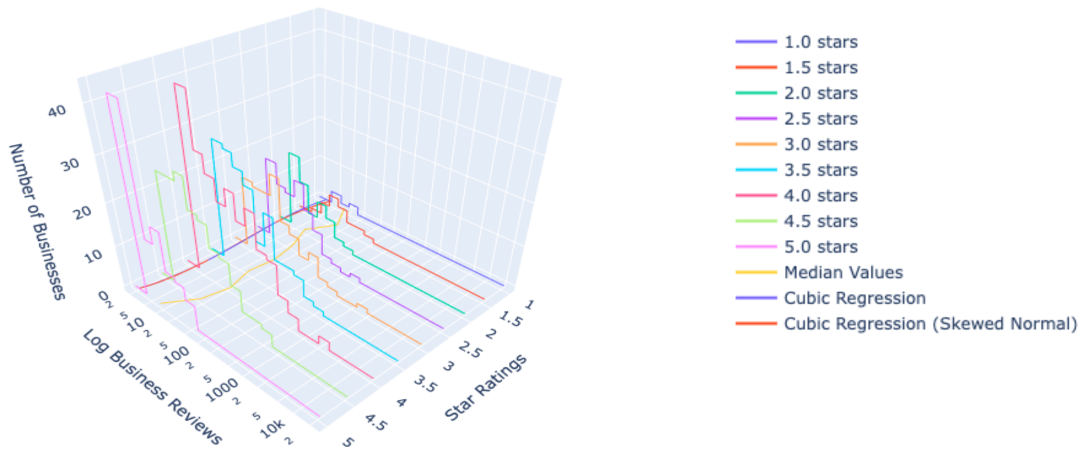

Figure 3: Distributions Represented Across Business Star Ratings vs Log Business Reviews vs No. of Businesses

A source of potential improvement could be to explore use of more advanced machine learning models like Random Forests and Neural Networks, which can produce more precise and nuanced models. Additionally, efforts could be made to increase the representativeness of the Yelp data by collecting a more diverse and random sample of businesses across the United States, as the current data is highly concentrated in just a few states, which may have introduced bias in our analysis.

The insights gained from this analysis can help users better understand the Yelp data provided. Users should not solely rely on the ratings but also consider the cumulative number of reviews on a business. A high-star rating with only a few dozen reviews may not actually indicate a better business compared to one with a lower rating but several hundred reviews. For businesses, this analysis provides valuable insights into the factors contributing to a high rating and ways to improve their ratings.

[1519 words]

References

[1] Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. Harvard Business School Working Paper, 16(043)
[2] Yu, M., Xue, M., & Ouyang, W. (2015). Restaurants Review Star Prediction for Yelp Dataset. Technical Report, University of California, San Diego.
[3] Cui, Y. (2015). An Evaluation of Yelp Dataset. arXiv preprint arXiv:1512.06915
[4] Kaleru, S., & Dhanikonda, S. R. (2018). Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database. International Journal of Applied Engineering Research, 13(21)
[5] Matheus, C., Oestreicher, J., & Bergsten, S. (2020). Transforming a large Yelp data set: Techniques for data preprocessing, analytics, and visualization. INTED2020 Proceedings.
[6] Liu, X., Schoemaker, M., & Zhang, N. (2014). Predicting usefulness of Yelp reviews. Stanford University, CS229 Project.
[7] Jensen, Scott (2017). Introducing Data Science to Undergraduates through Big Data: Answering Questions by Wrangling and Profiling a Yelp Dataset. Graduate School of Business, San José State University.
[8] Yip, Stanley. "Predicting Yelp Business Ratings." University of California San Diego, Technical Report CSE 190-C, Spring 2015.
[9] Tek, A. (2018). Predicting yelp stars based on business attributes, MEF Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye.