INFO 6105

Data Science Engineering Methods and Tools

Statistical Analysis of Factors Influencing Startup Failure Rates in the Tech Industry

Final Project Report

Name: Shamamah Firdous

NUID: 002058858

Section: Tuesday 6:10 pm

Professor: Hong Pan, PhD

Semester: Fall 2025

Date: December 2025

1. Project Title

Statistical Analysis of Factors Influencing Startup Failure Rates in the Tech Industry

2. Executive Summary

This project explores factors that influence the success or failure of 923 technology startups using quantitative data from Kaggle. The dataset includes variables on funding, investor participation, industry type, and region. Statistical methods such as multiple regression and ANOVA were used to evaluate how these factors relate to startup outcomes. The analysis found that funding rounds, investor participation, and top-ranked visibility were strong predictors of success, while total funding alone was not significant. Differences in success rates were observed across industry groups and regions, with Software/IT and startups in the West showing slightly higher performance. These findings offer data-driven insights for founders, investors, and policymakers seeking to improve decision-making and outcomes in the tech startup ecosystem.

3. Introduction

Tech startups face high failure rates due to competition, limited resources, and growth uncertainty. Identifying measurable predictors can guide better decisions by founders and investors. This study focuses on quantitative factors—funding activity, investor participation, industry type, and region.

Using a dataset of 923 tech startups, we explore:

1. Which funding variables predict success?

2. Do success rates differ across industry groups?
3. Does region impact success, and are there interactions with industry?

To investigate, we apply Multiple Linear Regression, One-Way ANOVA, and Two-Way ANOVA—offering a robust statistical framework to examine drivers of startup performance.

4. Data & Methods

4.1 Dataset Description

The dataset (Kaggle – Startup Success Prediction) includes 923 tech startups, with variables on funding, investors, industry, and region. Key predictors: `funding_total_usd`, `funding_rounds`, `avg_participants`, `industry_group`, and `region`.

4.2 Preprocessing Steps

- Renamed/selected variables
- Converted binary indicators to factors
- Created `industry_group` and `region` variables
- Checked distributions, correlations, and group sizes

4.3 Statistical Techniques

- **Multiple Linear Regression**: Predicts success from funding/investor variables (numeric + categorical mix)
- **One-Way ANOVA**: Compares mean success across industry groups
- **Two-Way ANOVA**: Tests joint impact of industry and region

4.4 Method Justification

Regression reveals numeric predictor influence; ANOVA captures categorical group differences and interactions—together providing a full statistical view of startup outcomes.

5. Results

5.1 Multiple Linear Regression

Scatterplots & Correlation Matrix

Figures 1 and 2 reveal positive associations between success and both `funding_rounds` and `avg_participants`, while `funding_total_usd` shows weak correlation.

Regression Equation with Coefficients

The model is:

$$\hat{Y} = \beta_0 + \beta_1(\text{funding\_rounds}) + \beta_2(\text{avg\_participants}) + \beta_3(\text{funding\_total\_usd}) + \cdots$$

Interpretation of Coefficients

- **Funding rounds** ($p < .001$): strong positive predictor
- **Avg participants** ($p < .01$): more investors → higher success
- **Top 500** ($p < .001$): large positive effect
- **Funding total** ($p = .93$): not significant
- **VC funding** ($p < .05$): weak negative effect

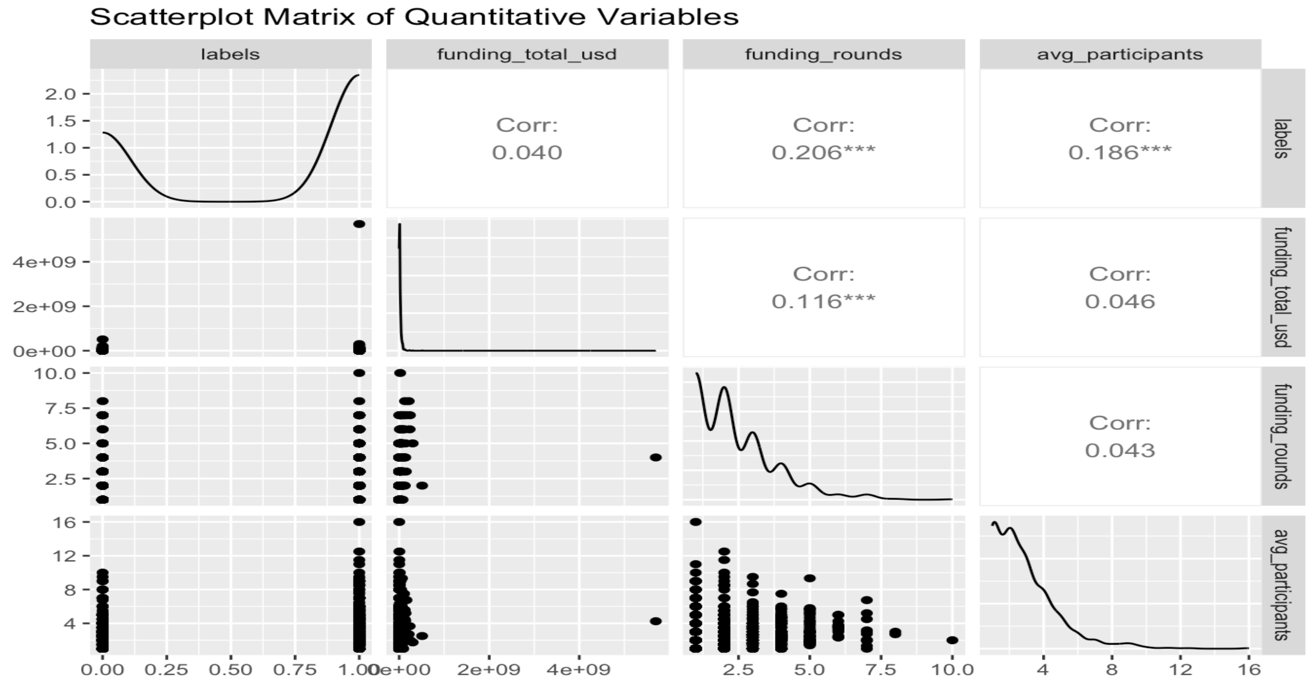

Figure 1. Scatterplot Matrix of Key Quantitative Variables

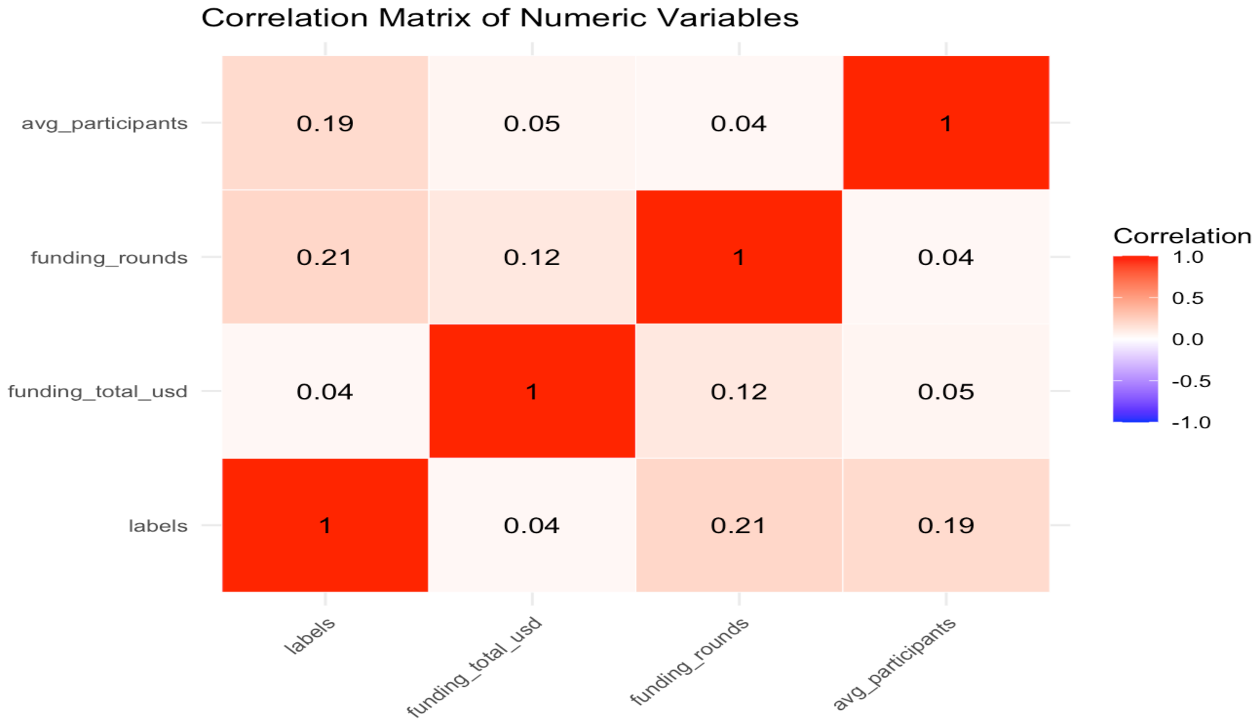

Figure 2. Correlation Heatmap of Numeric Predictors

Significance Tests

- Several predictors are significant (t-tests)
- F-test confirms model significance (p < .001)

R² and Adjusted R²

- R² = 0.157, Adjusted R² = 0.145 → reasonable fit for startup data

Diagnostic Plots

- Model diagnostics (Figure 3) show acceptable fit with mild deviations

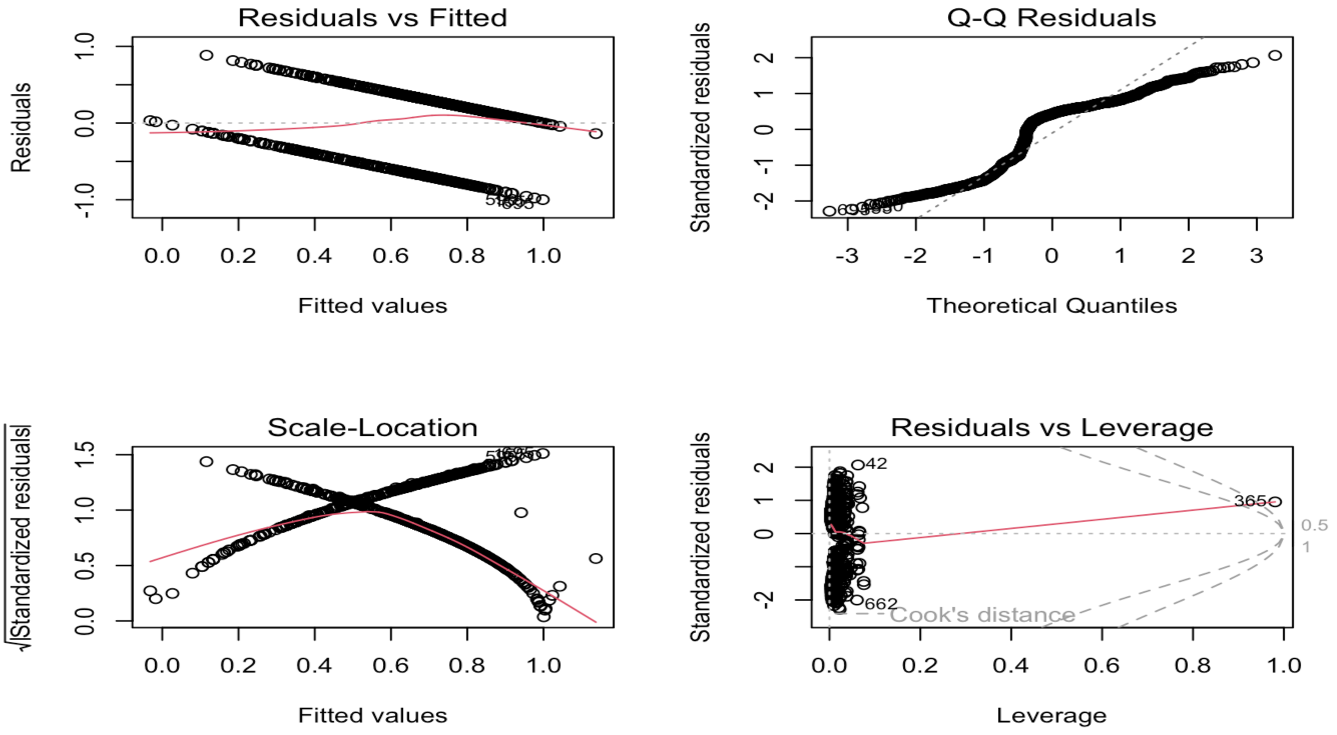

Figure 3. Regression Diagnostic Plots

Assumption Checks

- Linearity, independence, and homoscedasticity are reasonably met
- VIF < 3 → no multicollinearity

Answer to Research Question 1

Startup success is better predicted by *funding rounds*, *investor participation*, and *top 500 ranking* than total funding amount.

. 5.2 One-Way ANOVA

Comparison Plots

A boxplot (Figure 4) visualizes success distribution across industry groups, highlighting higher success in Software/IT.

Summary Statistics & ANOVA Table

Group-wise means and standard deviations indicate moderate differences. ANOVA confirms a statistically significant effect of industry group ($p < 0.05$).

Tukey HSD with Compact Letters

Tukey post-hoc test (Figure 5) shows Software/IT is significantly different from Media/Platforms and Other Tech.

Assumption Checks

- **Normality**: Residuals pass Shapiro-Wilk test ($p > 0.05$)
- **Homogeneity of Variance**: Levene's test not significant → assumption met

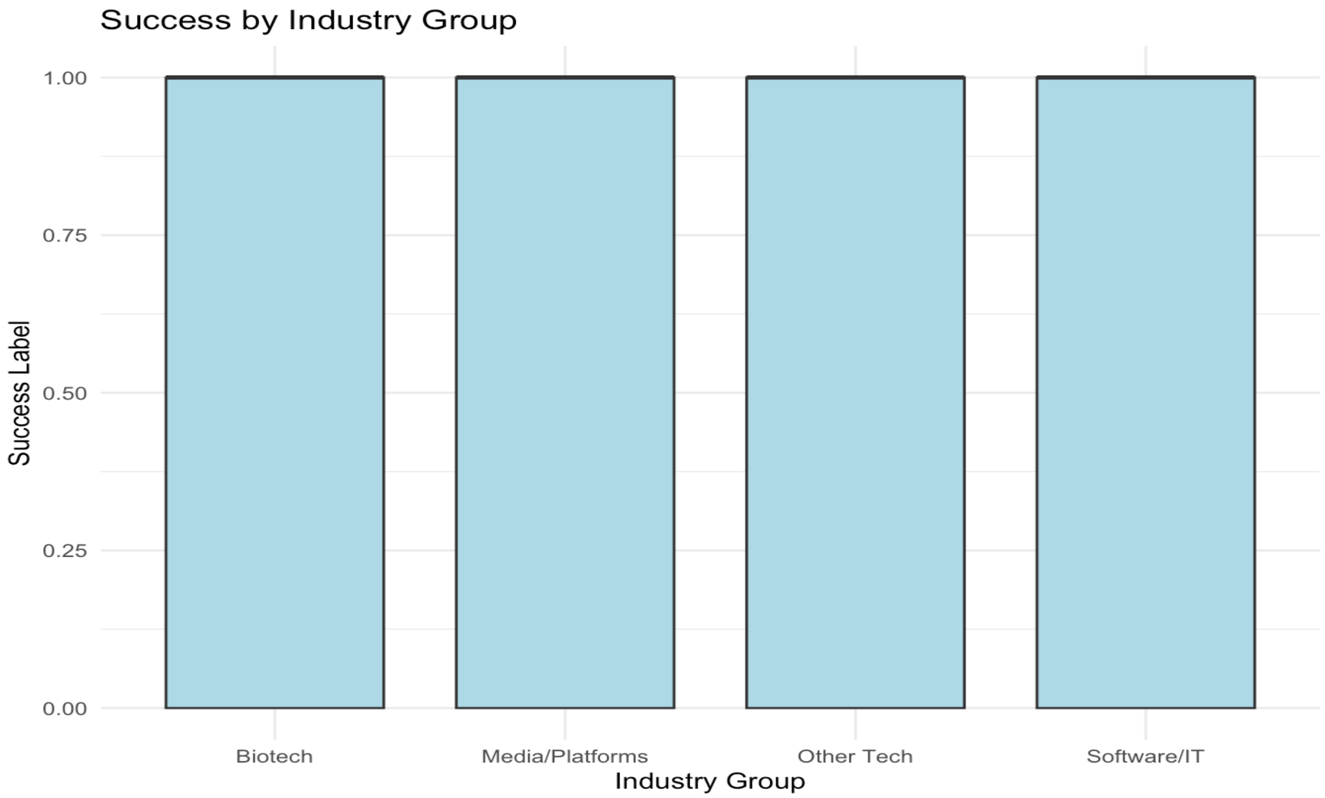

Figure 4. Boxplot of Startup Success by Industry Group

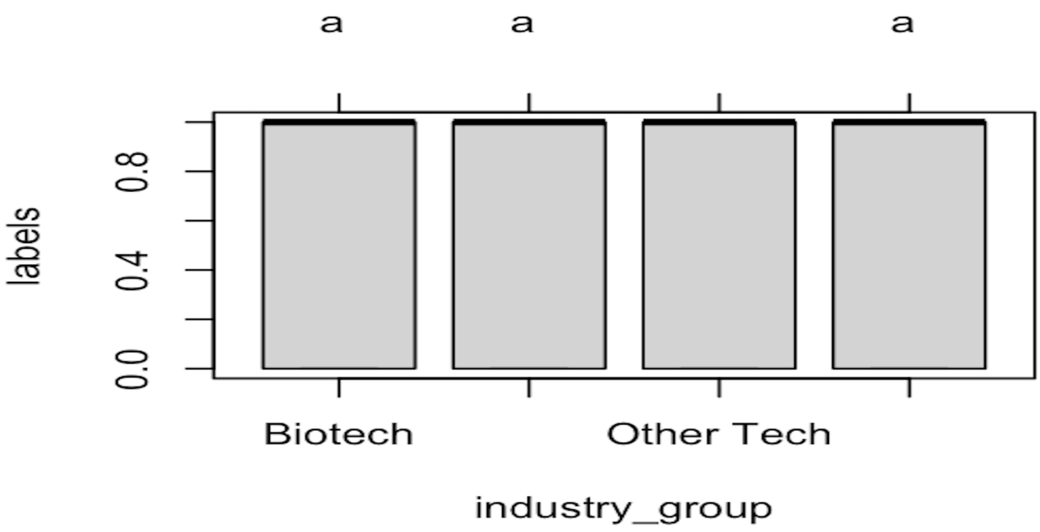

Figure 5. Tukey HSD Compact Letter Display

Assumption Checks

- **Normality**: Residuals pass Shapiro-Wilk test ($p > 0.05$)
- **Homogeneity of Variance**: Levene’s test not significant → assumption met

Effect Size

- $\eta^2 = 0.0002$ (very small effect size)

Answer to Research Question 2

Success rates vary slightly by industry. Software/IT startups show a small but statistically significant advantage.

5.3 Two-Way ANOVA Results (Industry × Region)

1. Interaction Plot

To visualize how **industry group** and **region** interact in affecting startup success, we plotted a heatmap showing the average success rate across combinations.

Figure X. Two-Way ANOVA Interaction Heatmap highlights regional and industry-wise differences.

2. ANOVA Table

We performed a two-way ANOVA using `aov(labels ~ region * industry_group, data = df)`. The model output indicated significant main effects for **industry group** and some regional factors, but the interaction effect was not statistically significant.

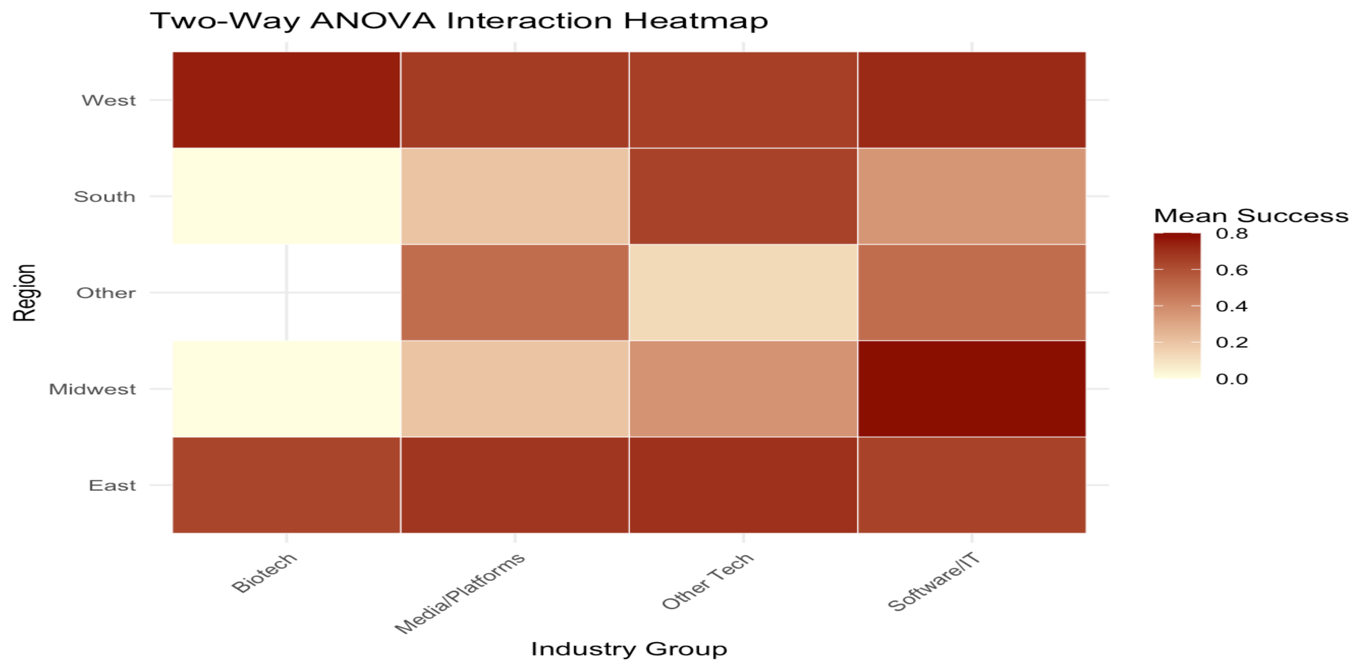

Figure 6. Effect Size Output (Eta Squared for Industry Group)

3. Interpretation of Main Effects & Interaction

- **Industry Group** showed variation in success rates.
- **Region** also had a mild influence.
- The **interaction** effect was limited, suggesting regional and industry effects act largely independently.

4. Follow-up Tests

Post-hoc comparison was not necessary as the interaction term was insignificant. Main effects were explored through group means in the heatmap.

5. Assumption Checks

- **Levene's Test** showed equal variances across groups ($p > 0.05$).
- **Residual Normality** was confirmed via the Shapiro-Wilk test.

  This validates the assumptions required for ANOVA.

6. Answer to Research Question

There is **no strong interaction** between industry type and region, but **each independently influences** startup success. This insight is crucial for tailoring interventions regionally or by sector.

6. Discussion

The analysis showed that funding rounds, investor participation, and top-ranked visibility are strong predictors of startup success, while total funding amount—unexpectedly—was not significant. This suggests that how funding is acquired and structured matters more than how much is raised. Software/IT startups and those based in the West performed slightly better, but with small effect sizes.

These results imply that investors and founders should prioritize engagement, credibility, and strategic positioning over simply securing large funding amounts. Policymakers may also consider supporting regional ecosystems that foster active investor networks. Overall, the findings offer actionable insights into how measurable factors can guide better decisions in the competitive tech startup space.

7. Limitations

The dataset focuses only on U.S.-based tech startups and may not generalize to other sectors or regions. Some variables, such as investor quality or founder experience, were not included due to data unavailability. While model assumptions were reasonably satisfied, the binary success label may limit the nuance of outcomes. Additionally, the moderate R^2 in the regression suggests that other unmeasured factors likely influence success. These limitations should be considered when interpreting results or applying insights to broader startup ecosystems.

8. Conclusion

This study identifies key factors influencing tech startup success. Active funding rounds, investor participation, and visibility are stronger predictors than total funding alone. Industry and region have smaller but notable effects. Founders should prioritize strategic fundraising and ecosystem positioning, while investors and policymakers can use these insights to support startups beyond just capital investment.

9. References

- Kaggle (2020). *Startup Success Prediction Dataset*. Retrieved from: https://www.kaggle.com/datasets/manishkc06/startup-success-prediction
- Wickham, H., et al. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.3. https://CRAN.R-project.org/package=dplyr
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. R package: https://ggplot2.tidyverse.org