Social Media and Tourism

Abstract:

With the growing popularity of social media platforms like Instagram and Twitter to share and promote “Instagrammable” tourist locations, we wanted to examine if social media use has increased the tourism and popularity of cities globally. Specifically, does the number of posts under a city’s hashtag on Instagram or Twitter directly correspond to the city’s popularity on TripAdvisor (measured by number of reviews and city ranking)? Data was collected for 8 variables (region, ranking, number of reviews, etc.) for 50 cities across the world. A multiple linear regression was fitted to generate the final model with the number of things to do, the number of Instagram posts, and population size as predictor variables and the $\sqrt{Reviews}$ as the response variable. This model proved to be appropriate and effective in predicting a city’s popularity, so there does appear to be a predictive relationship between social media use and tourism. Future studies might consider testing other social media platforms as predictors or seeing if the trends observed hold for smaller, less popular cities.

**Note: Larger figures and additional information can be found in the appendix.*

Background and Introduction:

With the growing popularity of platforms like Instagram and Twitter, there has been an emerging trend of influencers who make a living off of the popularity of their profiles. They often promote emerging social trends and sponsor businesses through their social media posts. With this new platform of marketing, the most viewed profiles are typically ones with beautiful pictures at hot tourist spots across the globe. The surge on social media to find the most "Instagrammable" place encouraged us to examine if the change in social media use has increased the popularity of famous, historic cities. Could seeing posts from someone's vacation influence where you go and what you do on your own trip? Specifically, **does the number of posts with a city's hashtag on Instagram or Twitter correspond to its popularity on TripAdvisor?** Do other factors, such as the cost of visiting a city, the population size of a city, or the number of tourist attractions, affect that city's overall popularity? We measured city popularity through the number of TripAdvisor reviews because if a city has more reviews this typically means more people have visited it.

Data and Exploratory Analysis:

a. Data and Variables

We took a stratified random sample of cities from the 2019 TripAdvisor Traveler's Choice Awards, choosing the top 5 cities from 10 randomly selected regions for a total of **50 cases**. For each city, we tracked **8 variables**: region, ranking, number of things to do, number of reviews, number of Instagram posts, number of tweets, average hotel price, and city population. All variables are quantitative except for region and ranking, which are categorical. The *number of reviews* was the response variable, as our selected measure of city popularity, and all other variables were potential predictors for the model. All data was collected on October 31, 2019, since the number of posts, average hotel prices, etc. change daily.

The first four variables were collected from TripAdvisor[1]. For **Region,** we created three subgroups (*Latin America:* Brazil, Caribbean, Colombia, Mexico; *Europe:* Greece, Italy, Romania, Switzerland, UK; *Oceania:* New Zealand). Otherwise, there would have been ten categorical variables in the model. For **Ranking**, we created two subcategories: *top tier* (cities ranked 1, 2, or 3 in their region) and *second tier* (cities ranked 4 or 5 in their region). Also, we looked at the number of **Things to Do** in each city and the number of **Reviews** each city received on TripAdvisor, in 10,000s. On Instagram, we counted the total number of **Instagram Posts** for each city's hashtag, in 100,000s[2]. Twitter doesn't show the total number of **Tweets**, so we used a third-party tool called Brand24 to count the number of public tweets that mentioned each city in the past 24 hours[3]. We could only access a trial version so some data was locked. On Kayak, we found the average **Hotel Price**, in dollars, for a three-star hotel[4]. However, these prices might not be accurate because some of the prices seemed too low or high compared to others. Finally, we recorded **City Population** in 10,000s from the UN database[5].

b. Exploratory Data Analysis

Figure 1 shows some tables and graphs from the EDA phase focusing on the response variable—number of reviews—and the three predictors used in our final model—things to do, Instagram posts, and population.

The distribution of TripAdvisor reviews was strongly right skewed and centered around 341,400 reviews. Most cities had < 100,000 reviews, so cities on the high end of the spectrum, especially those with > 600,000 reviews, may be potential outliers (e.g. London).

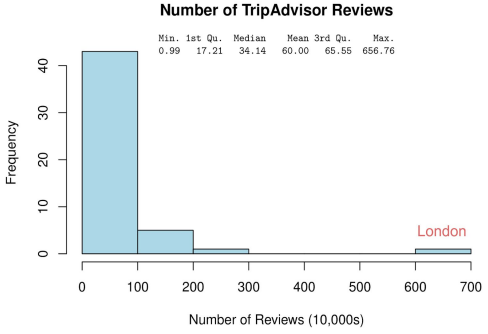
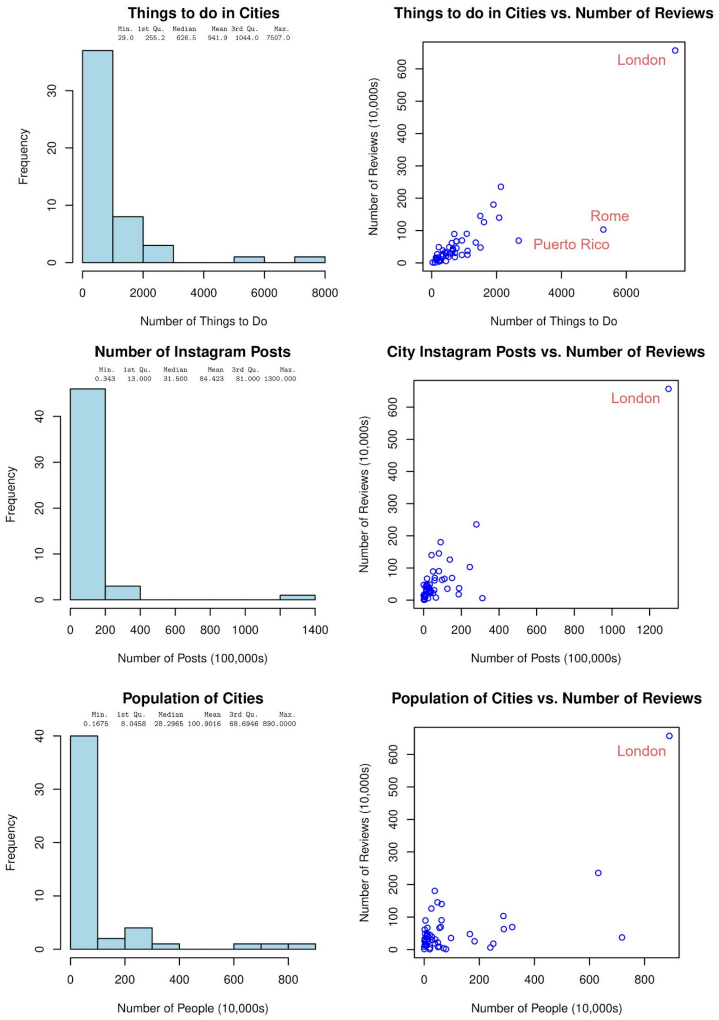


Figure 1. Histograms showing the distribution of individual variables and scatterplots of each predictor (things to do, Instagram posts, population) plotted against the response, number of reviews. *Continued on next page.*

We also see heavily right skewed distributions for all three predictors. There appears to be a relatively strong linear relationship between things to do and reviews but London, Rome, and Puerto Rico may be outliers. There is also a moderately strong linear relationship between number of Instagram posts and reviews, but again, London is a potential outlier. The association between population and reviews seems curved, suggesting that a transformation might be necessary, and London is a potential outlier again. All three identified outliers may negatively impact our model fit.



Model and Results:

a. *Analytic Method*

We analyzed our data with a multiple linear regression model. We first tried using all 7 predictors, but this model violated all assumptions and many of the predictors did not seem to be useful given the others (the t-tests for individual slopes yielded large p-values). Boxcox suggested that we use the square root of the number of reviews for each city. Forwards, backwards, and stepwise variable selection were performed and all three generated the same final model with **things to do**, **Instagram posts**, and **population** as predictors. This model was the most effective and best satisfied the assumptions. We also decided to remove the three outliers, London, Rome, and Puerto Rico, for a stronger linear model fit.

b. *Final Model and Results*

$$\widehat{\sqrt{Reviews}} = 2.425 + 0.0054(Things.to.do) + 0.0086(IG.Posts) - 0.0065(Population)$$

Our final model yielded the equation above. This shows, for example, that holding the number of Instagram posts and population size constant for a city, we expect that every additional thing to do is associated with an increase of 0.00054 $\sqrt{Reviews\ (in\ 10,000s)}$, on average. Based on the signs of the slopes, we also expect the $\sqrt{Reviews}$ to increase as the number of Instagram posts increases, but decrease as the population size increases, while holding the other two predictors constant in each case. The negative slope of population could be explained by the fact that if two cities had equal numbers of things to do and Instagram posts, the one with a larger population might be less popular since the tourist spots would be more crowded.

Based on the checks for assumptions in **Figure 2**, it seems that this model is appropriate. All model assumptions are satisfied as there are no particularly clear curved patterns in the pairs

scatterplots, the vertical spread of the residuals seems relatively equal in the residual plot, and the residuals seem to be approximately normally distributed.

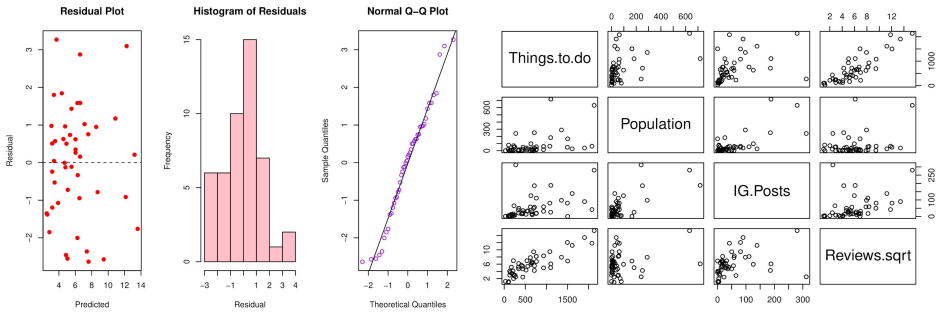


Figure 2. Residual plot, histogram of residuals, normal probability plot, and pairs scatterplot to show that all model assumptions are satisfied for multiple linear regression.

As shown in **Figure 3,** we have a small residual standard error, 1.542, and large adjusted R-squared, 0.7642, so it seems the model fits our data well. Additionally, the F-test generated a high F-statistic, 50.7, and a small p-value, 3.581e-14, so we seem to have an effective model for predicting the $\sqrt{Reviews}$ for cities. All three predictors—things to do, Instagram posts, and population—seem to be useful given the others if we use $\alpha = 0.1$, as the t-tests for individual slopes all yield small p-values of < 0.1. This is also supported by the fact that none of the 90% confidence intervals for the slopes of the predictors contain zero. For example, holding the number of Instagram posts and population size of a city constant, we are 90% confident that every additional thing to do is associated with a .0045 to .0062 increase in $\sqrt{Reviews\ (in\ 10,000s)}$, on average. Thus, we've found an effective model that addresses our research question by predicting city popularity (via $\sqrt{Reviews}$) with social media (Instagram posts) and other factors (population and things to do).

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4246479  0.3731333   6.498 6.89e-08 ***
Things.to.do  0.0053898  0.0004834  11.150 2.86e-14 ***
IG.Posts      0.0086322  0.0047925   1.801  0.07869 .
Population   -0.0064633  0.0023070  -2.802  0.00759 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.542 on 43 degrees of freedom
Multiple R-squared:  0.7796, Adjusted R-squared:  0.7642
F-statistic:  50.7 on 3 and 43 DF,  p-value: 3.581e-14

                      5 %          95 %
(Intercept)   1.7973844532  3.051911369
Things.to.do  0.0045772428  0.006202440
IG.Posts      0.0005756255  0.016688729
Population   -0.0103415888 -0.002585105
```

Figure 3. 90% confidence intervals and summary output that displays estimated coefficients, t-test results, F-statistic, residual standard error, and adjusted R^2, showing our model is effective at α = 0.1.

Discussion and Conclusions:

Our main objective for this project was to determine the impact of social media on tourism, specifically by seeing whether the number of posts with a city's hashtag on Instagram or Twitter correspond to its popularity on TripAdvisor, as measured by the number of reviews. Based on our results and final model, which was deemed to be both appropriate and effective, it appears that the number of Instagram posts is a strong predictor for number of $\sqrt{Reviews}$ at $\alpha = 0.1$, along with the number of things to do and population size of a city. The number of Instagram posts might be correlated with the things to do in a city because if there are more things to do, then there is probably more to post about, so this could explain the larger p-value for its slope (0.079). Additionally, the fact that no categorical variables were included in the final model because they lacked significance suggests that the association between social media posts and a city's popularity is universal, rather than being dependent on the city's global region or ranking on TripAdvisor.

Our study might be limited by the removal of outliers: London, Rome, and Puerto Rico (we can't extend the results to these cities). Also, the reliability of our data sources could explain why tweets and hotel prices were not useful predictors in the final model. Brand24 likely underreported the number of tweets due to data being locked in the trial version, and Kayak's hotel prices seemed potentially inaccurate. In the future, we could try testing other social media platforms as predictors or see if the trends observed hold for other, maybe less popular cities. Another potential study might look at the relation between sentiment of TripAdvisor reviews and city popularity, because having a greater number of reviews doesn't necessarily mean they are all positive.

References:

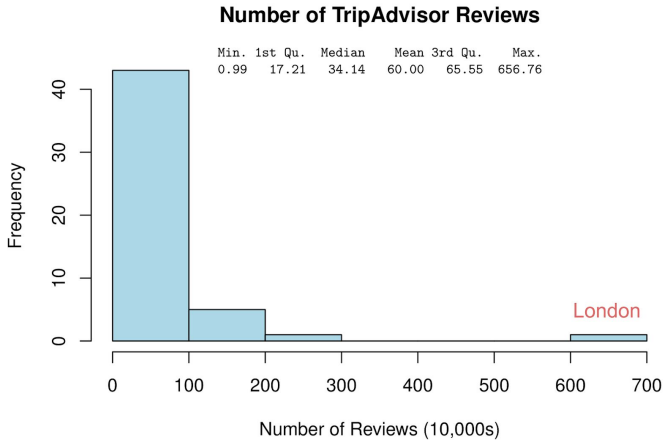
Data Sources

1. Travel data: https://www.tripadvisor.com/TravelersChoice-Destinations
    - Region, city ranking, number of things to do, and number of total reviews
2. Instagram data: https://instagram.com
    - Number of Instagram posts under the city's hashtag
3. Twitter data: http://twitter.com/
    - Number of tweets per day measured by https://brand24.com/
4. Hotel data: https://www.kayak.com/
    - Average price for a three-star hotel
5. Population data: http://data.un.org/
    - City population

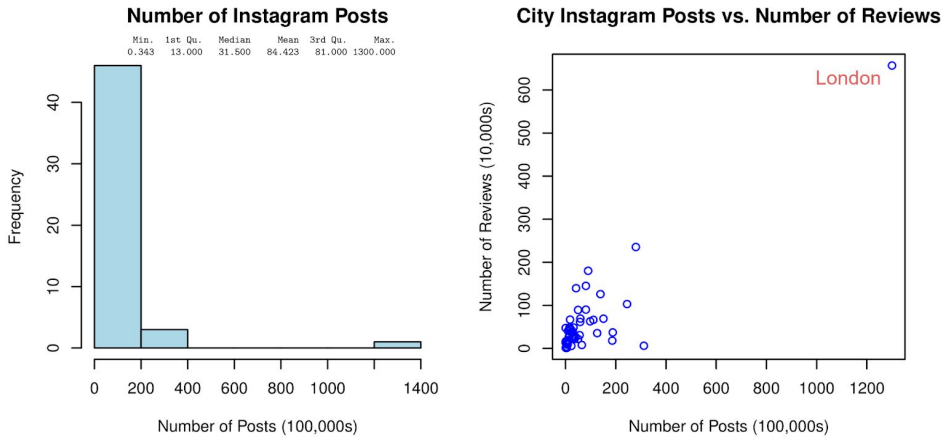
Appendix:

Larger Figures from Exploratory Data Analysis

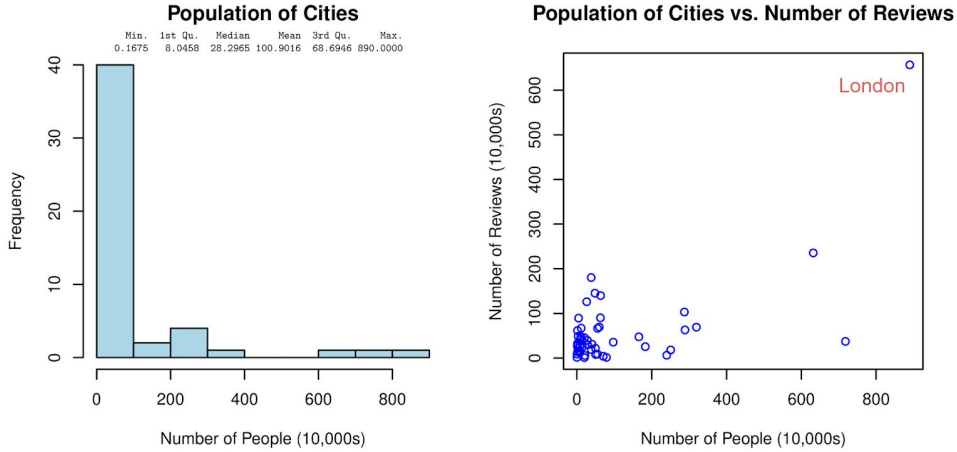
I. Histogram + Numerical Summary Output for Response Variable: Number of Reviews



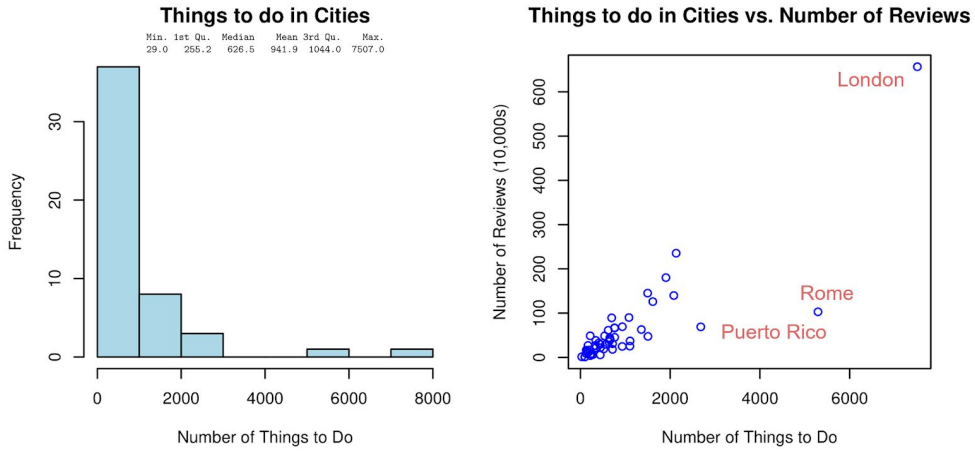
II. Histogram + Numerical Summary Output for Predictor 1: Instagram Posts; Scatterplot of Relationship Between Number of Instagram Posts and Reviews



III. Histogram + Numerical Summary Output for Predictor 2: City Population; Scatterplot of Relationship Between City Population and Number of Reviews



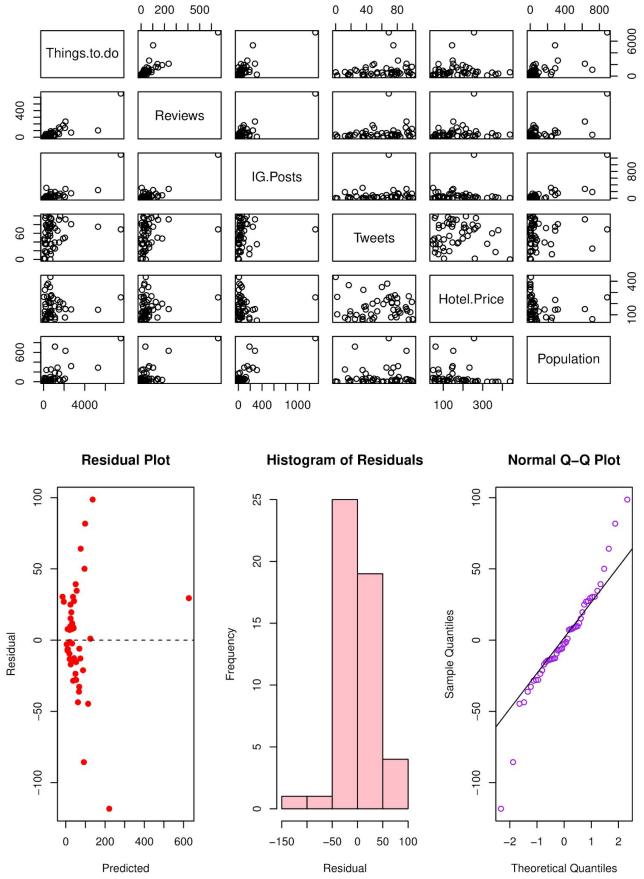
IV. Histogram + Numerical Summary Output for Predictor 3: Things to Do; Scatterplot of Relationship Between Number of Things to Do in Cities and Reviews



Figures from Original Model (All 7 predictors used, untransformed response, outliers included)

I. Model Assumptions: Pairs scatterplot, Residual Plot, Histogram of Residuals, and Normal Q-Q Plot

A. All assumptions violated (some curved relationships, i.e., Instagram Posts and Population; fanning pattern in residual plot; non-equal variance of residuals; and deviance from reference line in Q-Q plot, aka. residuals not normally distributed)



II. Inferential Results

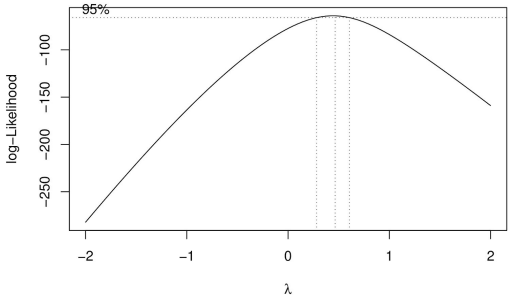
A. Only number of things to do and Instagram posts seem to be useful predictors in this model given the others; all other variables have large p-values > 0.2

B. Pretty large residual standard error (39.73)

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.704719  19.788794  -0.086   0.9318
RegionLatin America -16.426277  13.433872  -1.223   0.2284
RegionOceania       -17.680363  20.239479  -0.874   0.3874
RankingSecond Tier  -10.013627  12.927130  -0.775   0.4430
Things.to.do          0.022189   0.008913   2.490   0.0169 *
IG.Posts              0.357782   0.065279   5.481 2.36e-06 ***
Tweets                0.270470   0.213934   1.264   0.2133
Hotel.Price           0.072932   0.064227   1.136   0.2627
Population           -0.044969   0.059019  -0.762   0.4505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.73 on 41 degrees of freedom
Multiple R-squared:  0.8642, Adjusted R-squared:  0.8376
F-statistic:  32.6 on 8 and 41 DF,  p-value: 2.111e-15
```

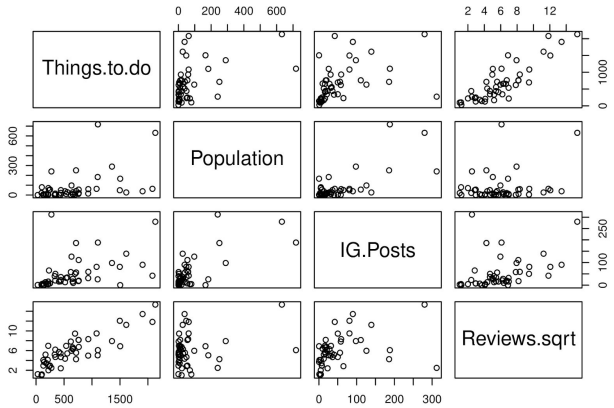
III. Boxcox: Applied transformation, $(\text{Reviews})^{1/2}$, in suggested optimal range

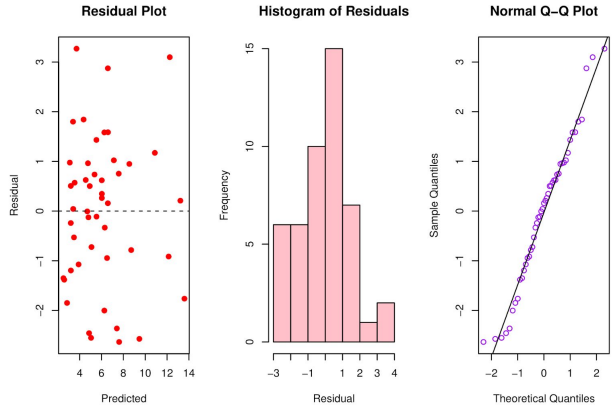


Larger Figures from Final Model and Results (Three predictors, response transformed, outliers removed)

I. Model Assumptions: Pairs scatterplot, Residual Plot, Histogram of Residuals, and Normal Q-Q Plot

A. All assumptions met (no clear curved relationships; no pattern in residual plot; equal variance of residuals; residuals appear to be approximately normally distributed)





II. Inferential Results

A. All three predictors are useful at $\alpha = 0.1$ (small p-values)

B. Smaller residual standard error than original model (1.542 vs 39.73)

C. Larger F-statistic than original model (50.7 vs. 32.6)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4246479  0.3731333   6.498 6.89e-08 ***
Things.to.do  0.0053898  0.0004834  11.150 2.86e-14 ***
IG.Posts      0.0086322  0.0047925   1.801  0.07869 .
Population   -0.0064633  0.0023070  -2.802  0.00759 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.542 on 43 degrees of freedom
Multiple R-squared:  0.7796, Adjusted R-squared:  0.7642
F-statistic:  50.7 on 3 and 43 DF,  p-value: 3.581e-14
```

D. 90% Confidence Intervals: none of the intervals contain 0, so it seems all the predictors are effective

```
                    5 %         95 %
(Intercept)   1.7973844532  3.051911369
Things.to.do  0.0045772428  0.006202440
IG.Posts      0.0005756255  0.016688729
Population   -0.0103415888 -0.002585105
```