Section 1 – Email to Manager:

Subject: Analysis of Circulation, Pulitzer Prizes and Strategic Scenarios.

Dear Manager,

I have completed the analysis examining how Pulitzer Prizes relate to newspaper circulation, the change in circulation over time, and the potential implications for the Boston Sun-Times under the three proposed strategic directions. All figures, tables, models, diagnostics, and interval estimates are included in the attached document.

1. Relationship Between Pulitzer Prizes and Circulation Levels

Using a linear model predicting average circulation from the number of Pulitzer Prizes, I found a statistically significant positive association. Newspapers with more prizes tend to have higher circulation, although the relationship explains only a modest proportion of the variation. After log-transforming circulation to correct skewness, the model assumptions were largely satisfied, making this model suitable for interpretation.

2. Relationship Between Pulitzer Prizes and Change in Circulation

The model predicting percentage change in circulation (2004–2013) shows that most newspapers experienced declines regardless of prize count. While the slope is statistically significant, the effect size is small, and model assumptions are less strongly met. This suggests that winning more prizes is linked to slightly less severe declines, but not enough to prevent circulation loss on its own.

3. Strategic Projections for the Boston Sun-Times

Using the two models, I generated predictions for scenarios of 3, 25, and 50 Pulitzer Prizes over the next 25 years.

Circulation Level (Model 1):

- 3 prizes: expected circulation well below the current 453,869.
- 25 prizes: still below current levels.
- 50 prizes: the only scenario where expected circulation can meet or exceed current levels, based on the 90% confidence interval.

Circulation Change (Model 2):

- All three scenarios predict negative 10-year changes, but higher prize counts produce smaller declines.
- Prediction intervals are wide, indicating significant uncertainty at the individual-newspaper level.

Together, the models show consistent direction: greater investment in investigative journalism improves circulation outcomes, even though broader industry trends continue to exert downward pressure.

4. Limitations

These models rely on a single predictor and historical data. Factors such as digital strategy, market size, and competition were not included, and the change model displays weaker diagnostic performance. As such, results should guide strategy directionally rather than serve as exact forecasts.

Conclusion

High investment in investigative journalism (aligned with the “50 prizes” scenario) offers the strongest potential circulation outcome and aligns with patterns observed across major US newspapers. While declines may still occur, the models indicate that maintaining or increasing Pulitzer-level performance provides the most favourable position for the Boston Sun-Times.

Kind regards,

Section 2- Attachment:

In this analysis we study 50 major US newspapers. We combine circulation data from 2004 and 2013 with the total number of Pulitzer Prizes won between 1990–2014. We clean and transform the data, explore univariate distributions, build linear models, check their assumptions, and use them to make predictions for the Boston Sun-Times under three strategic Pulitzer scenarios.

Loading and Cleaning Data:

R Code Input:

```
library(tidyverse)


#location of file

setwd("C:/Users/PC/Downloads")


# Load the dataset

pulitzer <- read_csv("pulitzer.csv")


view(pulitzer)
```

Recode change_0413 as integer percentage:

We need change_0413 as a numeric variable (integer percentages) instead of strings like "-24%" so we can summarise it and use it in linear models.

R Code Input:

```
# Recode change_0413 from strings to integer percentages

pulitzer <- pulitzer %>%

  mutate(

    change_0413_num = change_0413 %>%

      str_replace("%", "") %>%    # remove % sign

      as.integer()                            # convert to integer

)

head(pulitzer)
```

R Output:

```
A tibble: 6 × 6
 newspaper            circ_2004    circ_2013    change_0413   prizes_9014   change_0413_num
 <chr>                  <dbl>        <dbl>          <chr>         <dbl>           <int>
1 USA Today            2192098     1674306          -24%            3               -24
2 Wall Street Journal  2101017     2378827          +13%           51                13
3 New York Times       1119027     1865318          +67%          118                67
4 Los Angeles Times     983727      653868           -34%           86               -34
5 Washington Post       760034     474767            -38%          101               -38
6 New York Daily News  712671      516165            -28%            7               -28
```

The change_0413 variable is now an integer change_0413_num that expresses the percentage change in circulation (negative for declines, positive for growth). This makes it suitable for numeric analysis and modelling.

Create average circulation variable

We will create a new variable that contains the average of circ_2004 and circ_2013. This variable will serve as our main measure of circulation for modelling.

R Code Input:

```
pulitzer <- pulitzer %>%
mutate(
  avg_circ = (circ_2004 + circ_2013) / 2)
pulitzer %>%
  select(newspaper, circ_2004, circ_2013, avg_circ) %>%
  head()
```

R Output:

```
# A tibble: 6 × 4
 newspaper              circ_2004     circ_2013    avg_circ
 <chr>                    <dbl>         <dbl>        <dbl>
1 USA Today              2192098       1674306     1933202
2 Wall Street Journal    2101017       2378827     2239922
3 New York Times         1119027       1865318     1492172.
```

```
4 Los Angeles Times     983727      653868      818798.
5 Washington Post       760034      474767      617400.
6 New York Daily News   712671      516165      614418
```

Univariate Summary and Transformation

The Distribution of Average Circulation Description:

We will examine avg_circ to understand its shape, centre, spread, and any outliers.

We will use:

- Summary statistics to show the numerical centre and spread
- Histogram to reveal the shape of the distribution
- Boxplot to identify outliers and overall variability

R Code Input:

```
# Summary statistics

summary(pulitzer$avg_circ)
```

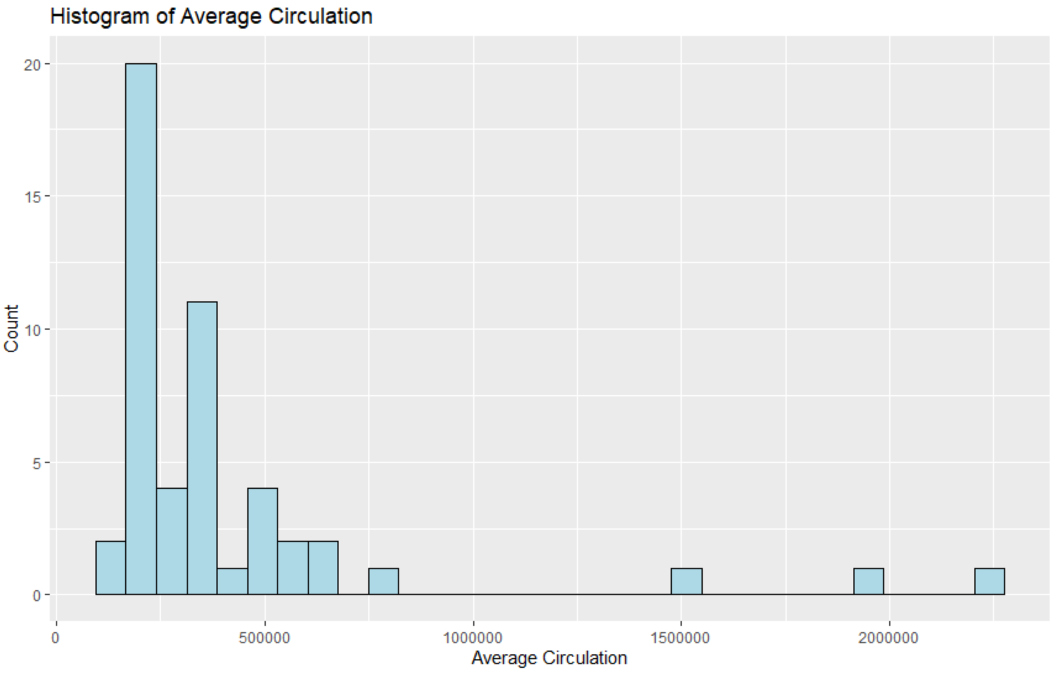
R Output:

```
  Min.    1st Qu.    Median   Mean    3rd Qu.   Max.

 131004   213509    298851   412442   436152   2239922
```

R Code Input:

```
# Histogram

ggplot(pulitzer, aes(x= avg_circ,)) +

  geom_histogram(color= "black", fill= "lightblue") +

  labs(title = "Histogram of Average Circulation",

       x = "Average Circulation",

       y = "Count")
```

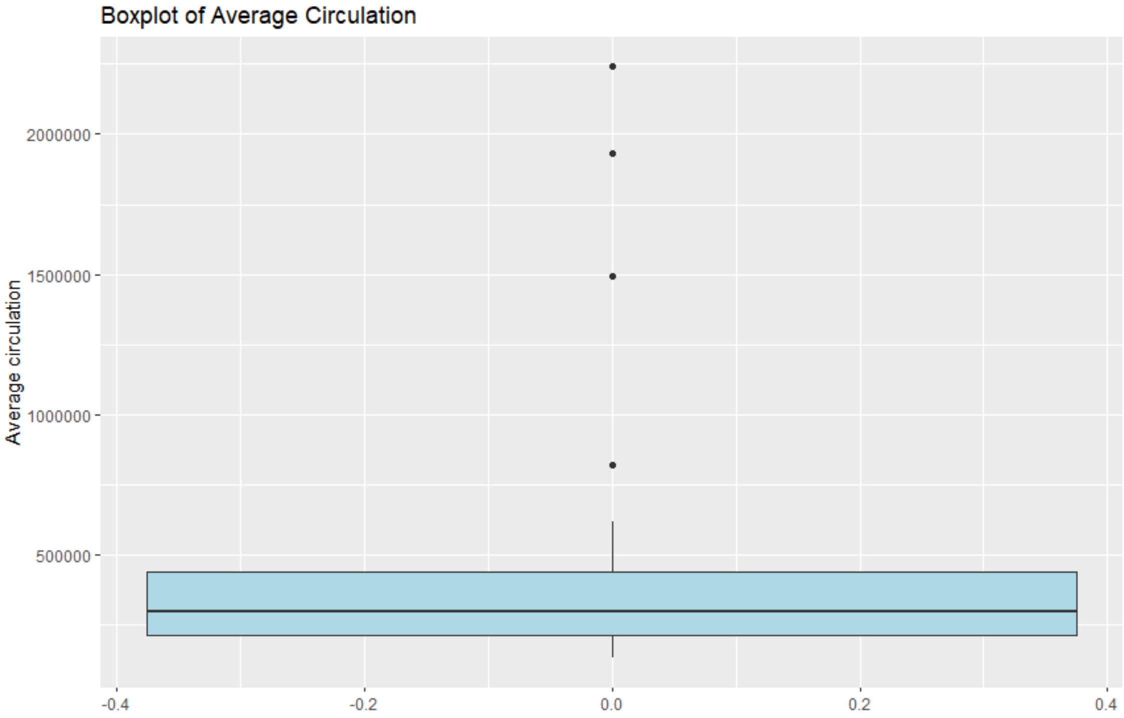
R Output:



R Code Input:

```
# Boxplot
ggplot(pulitzer, aes(y = avg_circ)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Average Circulation",
       y = "Average circulation")
```

R Output:



Summary

From looking at the summary, histogram and boxplot above, we can summarise the following:

- Shape:
  The distribution of average circulation is strongly right skewed.
  Most newspapers have circulations below 500,000, while a few national papers have very large circulations (over 1 million), stretching the right tail.

- Location (centre):
  The median circulation is about 299,000, and the mean is higher at about 412,000.
  The mean is pulled upward by very large newspapers.

- Spread (variation):
  Circulation ranges from roughly 131,000 to 2.24 million, showing very wide variation among newspapers.
  The interquartile range (IQR) is approximately 213,509 to 436,152, showing moderate spread within the middle 50% of newspapers.

- Outliers:
  Several newspapers with extremely high circulation appear as high outliers.

The Distribution of change_0413 Description:

we will follow the same steps but now for change_0413_num (the numeric value):

R Code Input:

```
# Summary statistics

summary(pulitzer$change_0413_num)
```

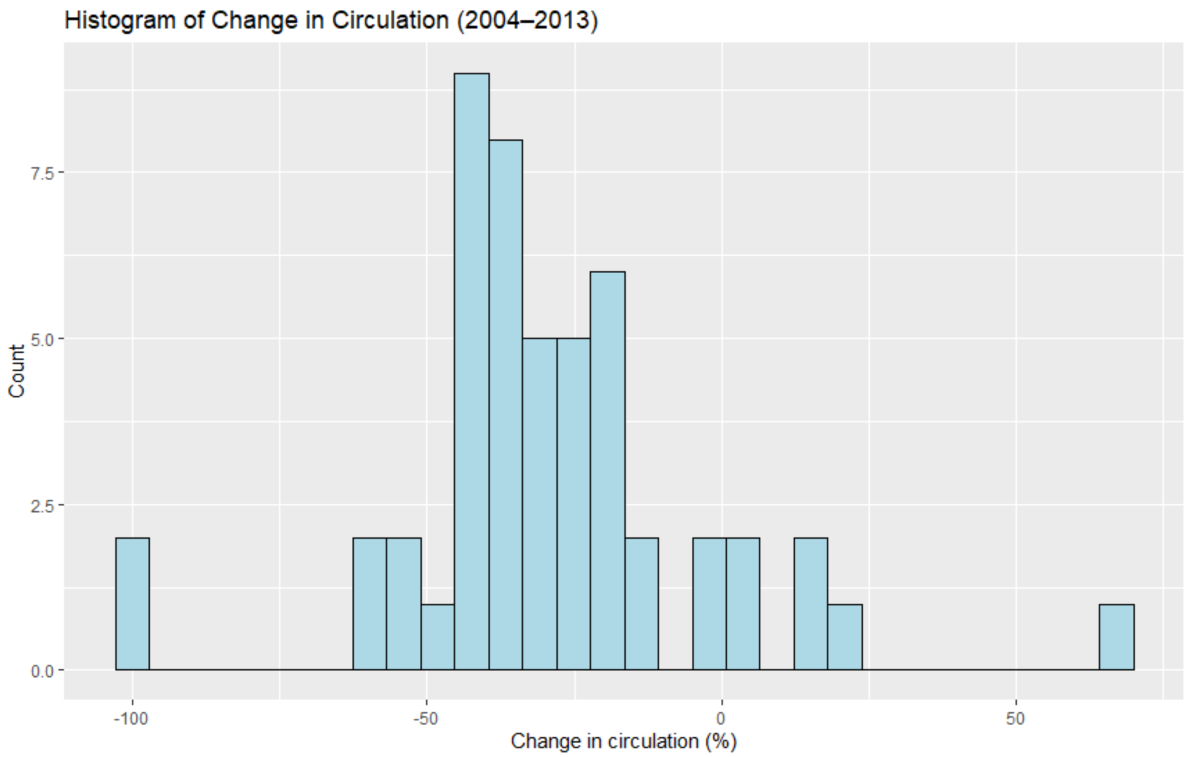
R Output:

```
> summary(pulitzer$change_0413_num)

   Min.     1st Qu.    Median    Mean    3rd Qu.    Max.

-100.00    -40.75     -32.50    -29.20    -20.00    67.00
```

R Code Input:

```
# Histogram

ggplot(pulitzer, aes(x= change_0413_num,)) +
  geom_histogram(color= "black", fill= "lightblue") +
  labs(title = "Histogram of Change in Circulation (2004–2013)",
       x = "Change in circulation (%)",
       y = "Count")
```

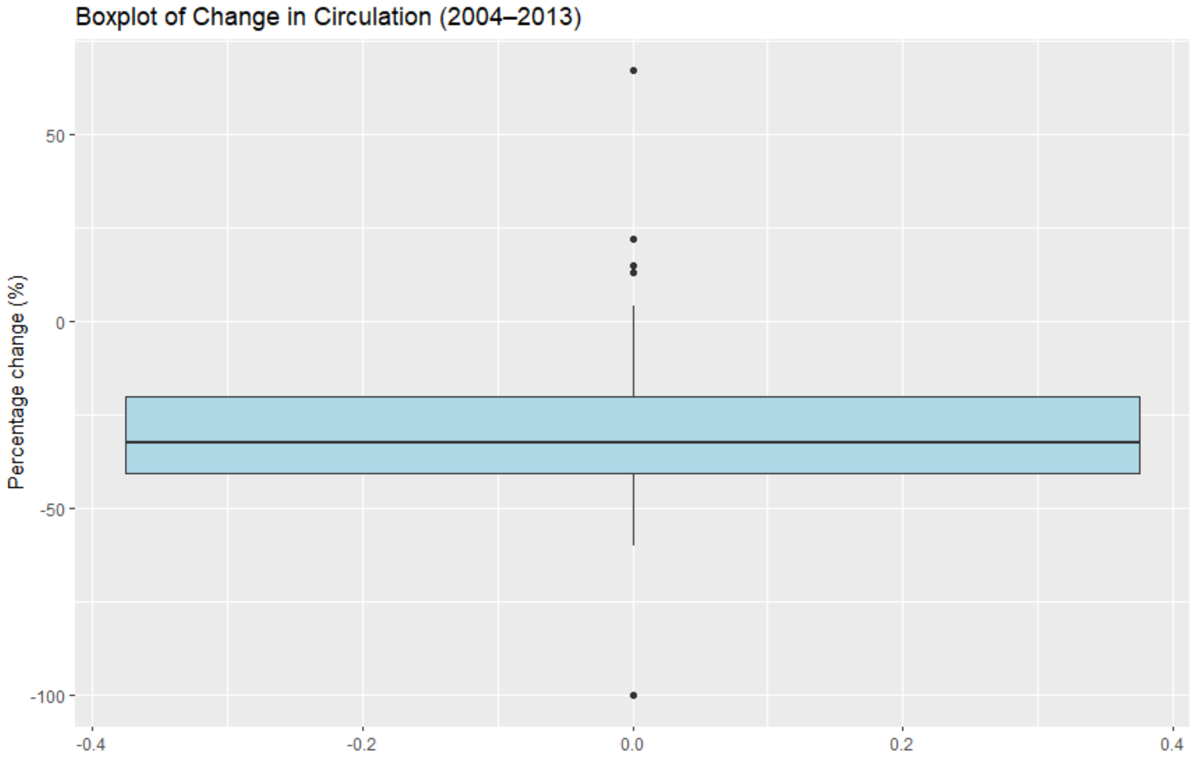
R Output:



R Code Input:

```
# Boxplot
ggplot(pulitzer, aes(y = change_0413_num)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Boxplot of Change in Circulation (2004–2013)",
       y = "Percentage change (%)")
```

R Output:



Summary

- Shape: Most values are negative (declines), clustered around –40% to –20%, with a few large positive increases. This produces a right-skewed distribution (longer tail to the right).
- Location: Median ≈ –32.5%, mean ≈ –29.2%, indicating a typical moderate decline across newspapers.
- Spread: Changes range from –100% (complete loss) to +67% (strong growth).
- Outliers: Some newspapers experienced extreme decline (–100%) or strong growth (e.g. +67%).

Assessing Skewness and Need for Log Transformation:

We will assess and decide whether avg_circ and/or change_0413_num should be log-transformed to reduce skew and better meet linear model assumptions.

- **Change in circulation (2004–2013) (change_0413_num):**

The distribution contains negative values and therefore cannot be log-transformed. Even though the distribution is skewed, a log transformation is not suitable for this variable.

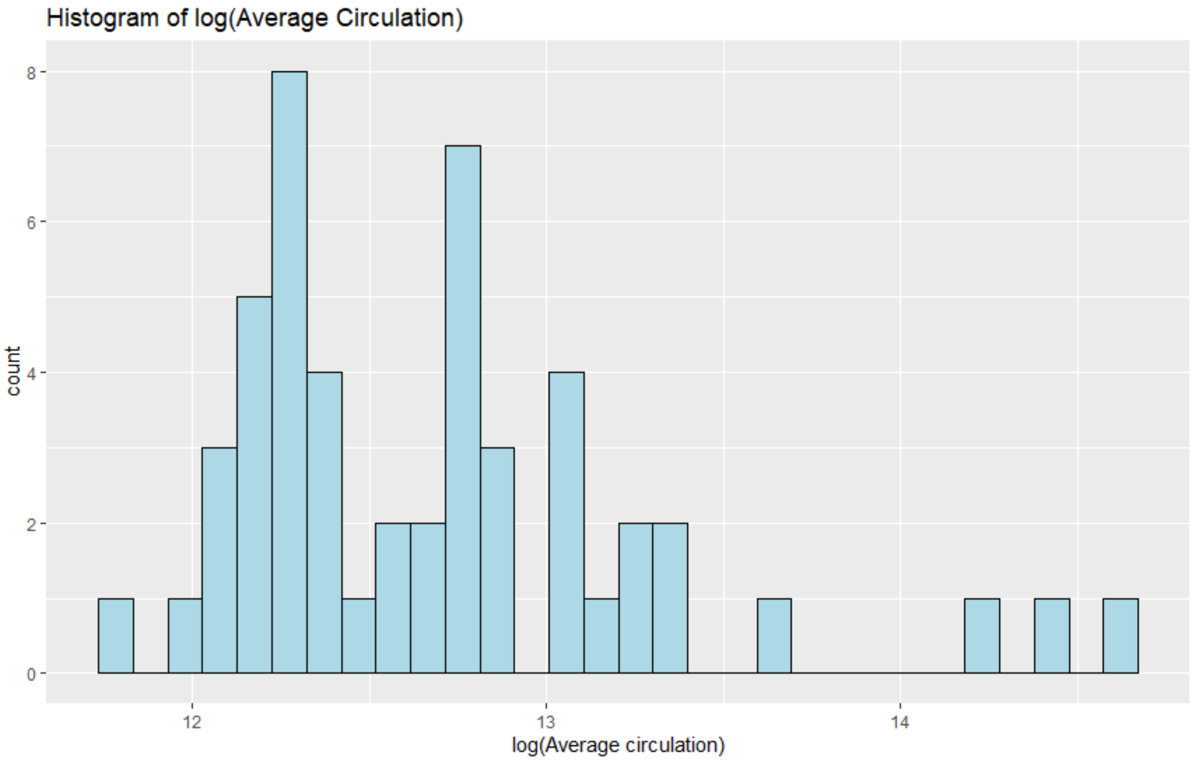
- **Average circulation (avg_circ):**

The histogram above shows a strong right skew, the circulation values are all positive, and log transformation is commonly used for right-skewed count data. We will check in R and see the outcome so we can make an informed decision:

R Code Input:

```
# Try log transform for avg_circ
ggplot(pulitzer, aes(x = log(avg_circ))) +
  geom_histogram(color = "black", fill = "lightblue") +
  labs(title = "Histogram of log(Average Circulation)",
       x = "log(Average circulation)")
```

R Output:



Summary:

The histogram of log-transformed average circulation shows a more symmetric and balanced distribution compared to the original right-skewed shape. The log transformation successfully reduces the influence of very large circulation values and removes extreme high outliers. This produces a distribution that more closely resembles normality and is more appropriate for linear modelling. Therefore, the variable representing average circulation should be log-transformed.

Model Building and Interpretation

(log) average circulation vs Pulitzer Prizes

We will model how the number of Pulitzer Prizes relates to long-term circulation. We will use log(avg_circ) as the response and prizes_9014 as the predictor.

R Code Input:

```
# Linear model: log average circulation ~ prizes


model1 <- lm(log(avg_circ) ~ prizes_9014, data = pulitzer)


summary(model1)
```

R output:

```
summary(model1)

Call:

lm(formula = log(avg_circ) ~ prizes_9014, data = pulitzer)


Residuals:

   Min      1Q      Median     3Q      Max

-0.8069   -0.3147   -0.1556   0.1825   1.9693


Coefficients:
```

```
                Estimate Std.   Error      t value    Pr(>|t|)
(Intercept)     12.463142      0.085501   145.767    < 2e-16 ***
prizes_9014     0.014083       0.002928   4.811      1.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.505 on 48 degrees of freedom
Multiple R-squared:  0.3253,  Adjusted R-squared:  0.3112
F-statistic: 23.14 on 1 and 48 DF,  p-value: 1.532e-05
```

Summary:

From summary(model1):

- Intercept ≈ 12.463
- Slope for prizes_9014 ≈ 0.014
- R-squared ≈ 0.32
- p-value for slope < 0.001

Model1 (on log scale):

$$\log(\text{avg\_circ}) = 12.463 + 0.014 \times \text{prizes\_9014}.$$

Interpretation

- Intercept: When prizes_9014 = 0, the model predicts log(average circulation) ≈ 12.463. On the original scale, this corresponds to an average circulation of about exp(12.463) ≈ 259,000 copies for a newspaper with no prizes.
- Slope: For each extra Pulitzer Prize, log(average circulation) increases by about 0.014, i.e. average circulation increases by roughly 1.4% .
- Statistical significance: The p-value (< 0.05) indicates a statistically significant positive relationship between prizes and average circulation. Newspapers with more Pulitzer Prizes tend to have higher average circulation.

Change in circulation vs Pulitzer Prizes

We want to model how the number of Pulitzer Prizes relates to the percentage change in circulation between 2004 and 2013.

Because change_0413_num contains negative values, a log transformation is not appropriate.

Therefore, we build a simple linear regression model:

$$\text{change\_0413\_num} \sim \text{prizes\_9014}$$

R Code Input:

```
# Model 2: Predicting change in circulation

model2 <- lm(change_0413_num ~ prizes_9014, data = pulitzer)



summary(model2)
```

R Output:

```
> summary(model2)

Call:

lm(formula = change_0413_num ~ prizes_9014, data = pulitzer)

Residuals:

   Min        1Q      Median       3Q       Max

-68.068    -10.251     -2.713      13.126    56.749


Coefficients:

              Estimate Std.   Error    t value    Pr(>|t|)

(Intercept)    -35.4152       4.3336   -8.172    1.21e-10 ***

prizes_9014     0.3870        0.1484    2.608     0.0121 *

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 25.59 on 48 degrees of freedom

Multiple R-squared:  0.1241,  Adjusted R-squared:  0.1059

F-statistic: 6.802 on 1 and 48 DF,  p-value: 0.0121
```

Summary:

From summary (model2):

- Intercept ≈ –35.42
- Slope for prizes_9014 ≈ 0.387
- R-squared ≈ 0.12
- p-value for slope ≈ 0.012

Model2:

$$\text{change\_0413\_num} = -35.42 + 0.387 \times \text{prizes\_9014}.$$

Interpretation

- Intercept: For a newspaper with zero Pulitzer Prizes, the model predicts an average circulation change of about –35%, meaning a substantial decline.
- Slope: Each additional Pulitzer Prize is associated with an improvement of about 0.39 percentage points in circulation change. This effect is extremely small and practically negligible.
- Significance: The slope is statistically significant (p ≈ 0.012), indicating a real positive association between Pulitzer Prizes and the response variable. However, the R^2 value is low (≈ 0.12), which means that although the relationship is statistically significant, the number of prizes explains only a small proportion of the variation in circulation changes.

Checking assumptions for both models

We must check the four linear model assumptions for each model to make sure it is reliable to trust:

1. Linearity
2. Constant variance (homoscedasticity)

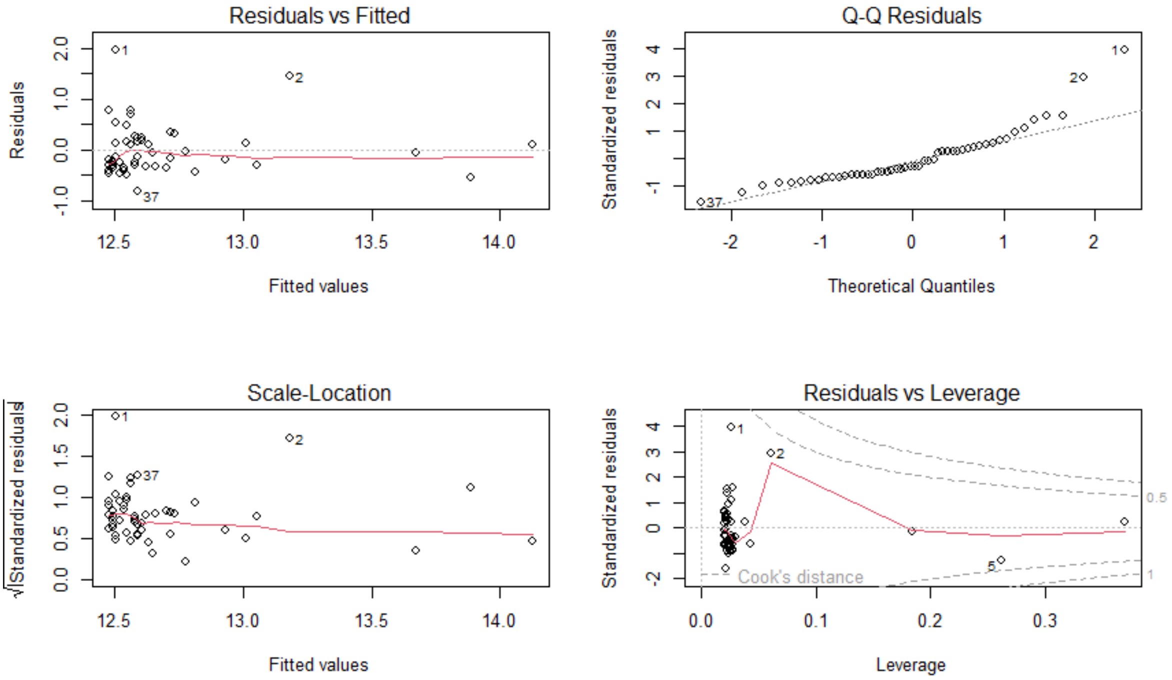
3. normality of residuals
4. Independence

Model 1:

R Code Input:

```
# Standard diagnostic plots for model1
par(mfrow = c(2, 2))
plot(model1)

# Reset plotting
par(mfrow = c(1, 1))
```

R Outcome:



Model 1 (log(avg_circ) ~ prizes_9014) Interpretation:

- Residuals vs Fitted: Relationship appears roughly linear after log transform, with some curvature at extremes due to very large newspapers.
- Scale–Location: Variance is more constant than on the raw scale, though residual spread is slightly larger for high-prize newspapers.

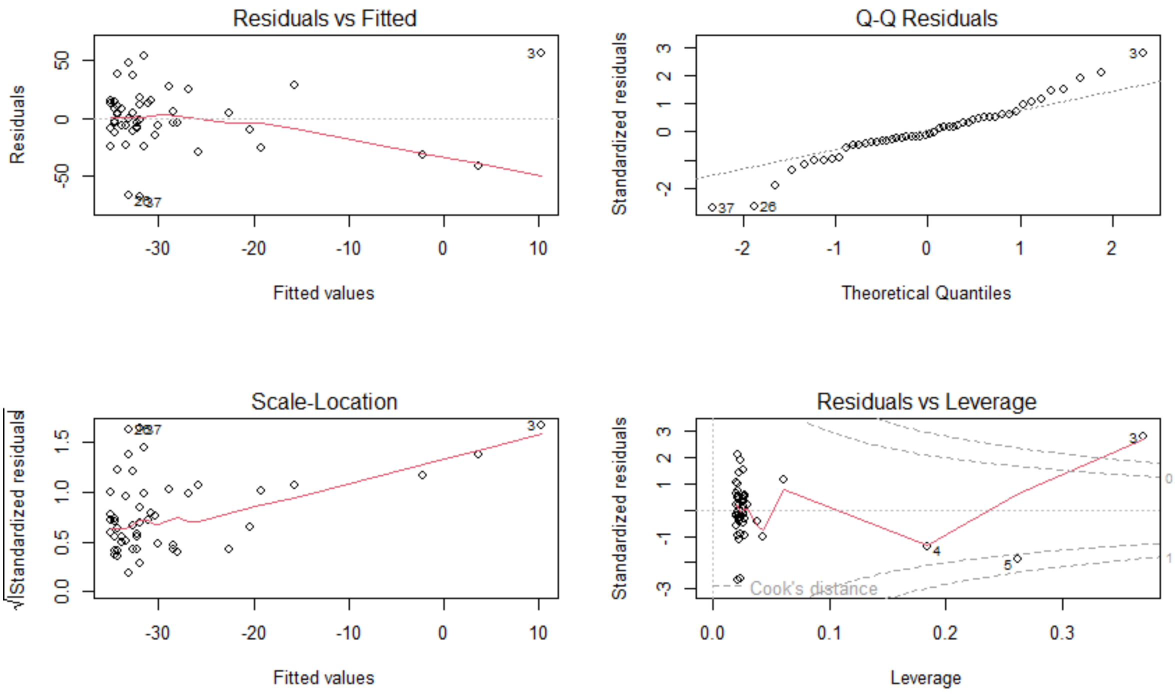
- Normal Q–Q: Residuals mostly follow the straight line, with some right-tail deviation (influence of very large national papers).
- Residuals vs Leverage: A few high-leverage points (large circulation, many prizes), but no extreme outliers with very high Cook's distance.

Model 2:

R Code Input:

```
# Standard diagnostic plots for model2

par(mfrow = c(2, 2))

plot(model2)


par(mfrow = c(1, 1))
```

R Output:



Model 2 (change_0413_num ~ prizes_9014) Interpretation:

- Linearity: Residuals show a clear downward trend rather than random scatter → linearly assumption is violated.

- Constant variance: Residual spread increases with fitted values (upward red line) → heteroscedasticity present.
- Normality: Q–Q plot shows strong deviations in both tails, especially point 30 → normality not satisfied.
- Influential points: Point 30 has high leverage and approaches Cook's distance line → influential observations are present.

Summary of Both Models

• **Model 1** performs better overall. The log transformation made the relationship more stable, and the diagnostic plots were generally acceptable, so the model can be used with reasonable confidence.

• **Model 2** is weaker. Although the predictor is statistically significant, the diagnostic plots showed more irregularities and the unexplained variation is high, so predictions from this model should be treated cautiously.

• **Independence** is reasonable for both models because each observation represents a different newspaper.

Overall: Model 1 is the more reliable basis for strategic forecasting, while Model 2 provides only limited insight.

Prediction:

Expected Circulation Under Three Strategies (Model 1):

Using Model 1, we predict the average circulation for the Boston Sun-Times under three projected Pulitzer counts: 3, 25, and 50 prizes over the next 25 years. We compare these predictions to the current circulation (453,869).

R Code Input:

# Three scenarios

```
new_prizes <- tibble(
  prizes_9014 = c(3, 25, 50)
)

# Predict log(avg_circ)
pred_circ_log <- predict(model1, newdata = new_prizes)

# Convert back to circulation scale
pred_circ <- exp(pred_circ_log)

pred_circ
```

R Output:

```
> pred_circ
    1          2          3
269788.1  367773.8   522983.1
```

Interpretation & rounding values:

Prizes	Predicted	average circulation
3	≈	269,800
25	≈	367,800
50	≈	523,000

Summary

Compared to the current circulation of 453,869:

- At 3 prizes, expected circulation is well below current levels.
- At 25 prizes, expected circulation is still below current levels.
- At 50 prizes, expected circulation is above current levels.

Only the high-investigative strategy (50 prizes) is associated with higher expected circulation than at present.

Expected Change in Circulation (Model 2) and Consistency:

Using Model 2, we predict the percentage change in circulation over the period under the three prize scenarios and compare to Model 1's implications.

R Code Input:

```r
# Three scenarios
new_prizes <- tibble(
  prizes_9014 = c(3, 25, 50)
)

# Predict change in circulation
pred_change <- predict(model2, newdata = new_prizes)

pred_change
```

R Output:

```
pred_change
     1          2          3
-34.25423   -25.74021   -16.06518
```

Interpretation & rounding values:

Prizes	Predicted	change in circulation (%)
3	≈	–34.3%
25	≈	–25.7%
50	≈	–16.1%

Summary:

All predicted values are negative, meaning circulation is expected to decline in every scenario. More prizes make the decline slightly smaller, but the improvement is minimal.

Are the two models consistent?

Model 1 predicts that higher investment (50 prizes) would lead to a higher overall circulation than the current level, while lower investment produces substantially lower circulation.
Model 2 predicts that circulation will still decline over the decade in all scenarios, but that the decline is slightly smaller when more prizes are won.
Both models are therefore directionally consistent: more investment in investigative journalism results in better circulation outcomes, even though Model 2 still predicts a decline overall.

90% Confidence Intervals for Expected Circulation (Model 1)

We will compute 90% confidence intervals for the expected (mean) circulation under each strategic direction to quantify uncertainty around the mean predictions.

R Code Input:

```r
pred_circ_ci <- predict(model1,
                        newdata = new_prizes,
                        interval = "confidence",
                        level = 0.90)

pred_circ_ci <- as_tibble(pred_circ_ci)

pred_circ_ci <- pred_circ_ci %>%
  mutate(
    fit = exp(fit),
    lwr = exp(lwr),
    upr = exp(upr)
  )
```

```
results_4_3 <- bind_cols(new_prizes, pred_circ_ci)

results_4_3
```

R outcome (rounded)

Prizes	Expected circulation	90% CI lower	90% CI upper
3	≈ 269,800	≈ 235,500	≈ 309,000
25	≈ 367,800	≈ 323,700	≈ 417,800
50	≈ 523,000	≈ 425,900	≈ 642,100

Interpretation:

- For 3 prizes, we are 90% confident the true mean circulation lies between about 236k and 309k, well below current level (453,869).
- For 25 prizes, the 90% CI (≈ 324k–418k) remains below current circulation.
- For 50 prizes, the 90% CI (≈ 426k–642k) includes and extends above current circulation, suggesting that high investment in investigative journalism could plausibly sustain or grow circulation.

90% Prediction Intervals for Change in Circulation (Model 2)

We will compute 90% prediction intervals for the actual change in circulation under each strategy, which reflect uncertainty for an individual newspaper (the Boston Sun-Times), not just a mean.

R Code Input:

```
pred_change_pi <- predict(model2,
                          newdata = new_prizes,
                          interval = "prediction",
                          level = 0.90)
```

```
pred_change_pi <- as_tibble(pred_change_pi)


results_4_4 <- bind_cols(new_prizes, pred_change_pi)

results_4_4
```

R Output (rounded):

Prizes	Predicted change (%)	90% PI lower	90% PI upper
3	≈ −34.3%	≈ −77.7%	≈ +9.2%
25	≈ −25.7%	≈ −69.2%	≈ +17.7%
50	≈ −16.1%	≈ −60.2%	≈ +28.1%

Interpretation:

- All three prediction intervals are very wide, spanning from large declines to small increases.
- This means that Model 2 cannot confidently predict whether circulation will fall or rise over the next decade.
- Although the predicted *means* are negative (declines), the intervals show a high degree of uncertainty.

Summary:

Model 1 gave reasonably tight confidence intervals for expected circulation.
Model 2, however, produces extremely wide prediction intervals, showing that individual circulation changes are much harder to predict.

Overall Conclusion:

Using data from 50 major US newspapers, we found that newspapers with more Pulitzer Prizes tend to have higher average circulation and experience less severe declines in circulation. A log-linear model relating average circulation to Pulitzer counts shows a moderately strong, statistically significant relationship, while the model for circulation change shows a weaker but still significant association. Predictions for the Boston Sun-Times suggest that reducing investment in investigative journalism (few prizes) is associated with substantially lower circulation, while a high-investigative strategy (many prizes) offers the best chance of sustaining or improving circulation, though with considerable uncertainty.

References:

- Statistics Solutions (undated) *Assumptions of multiple linear regression analysis*. Available at: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/ (Accessed: 21 November 2025).
- Yang, K. (2019) ‘An overlooked critical assumption for linear regression’, *PMC*, 8 July. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6802968/ (Accessed: 21 November 2025).
- Real-Statistics (undated) *Confidence and prediction intervals for forecasted values*. Available at: https://www.real-statistics.com/regression/confidence-and-prediction-intervals/ (Accessed: 21 November 2025).
- STHDA (undated) *Linear regression and diagnostics in R*. Available at: https://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/ (Accessed: 21 November 2025).
- AnalystPrep (2023) ‘Predicted Value & Prediction Interval | CFA Level 1’, *AnalystPrep*, 19 August. Available at: https://www.analystprep.com/cfa-level-1-exam/quantitative-methods/predicted-value-and-prediction-interval-of-a-dependent-variable/ (Accessed: 21 November 2025).
- Data-Overload (2023) ‘Understanding linear regression assumptions: a crucial foundation for analysis’, *Medium*, 1 July. Available at: https://medium.com/@data-overload/understanding-linear-regression-assumptions-a-crucial-foundation-for-analysis-9220a18fb836 (Accessed: 21 November 2025).
- Penn State Eberly College of Science (undated) *Prediction Interval for a New Response - STAT 501 Lesson 3.3*. Available at: https://online.stat.psu.edu/stat501/lesson/3/3.3 (Accessed: 21 November 2025).