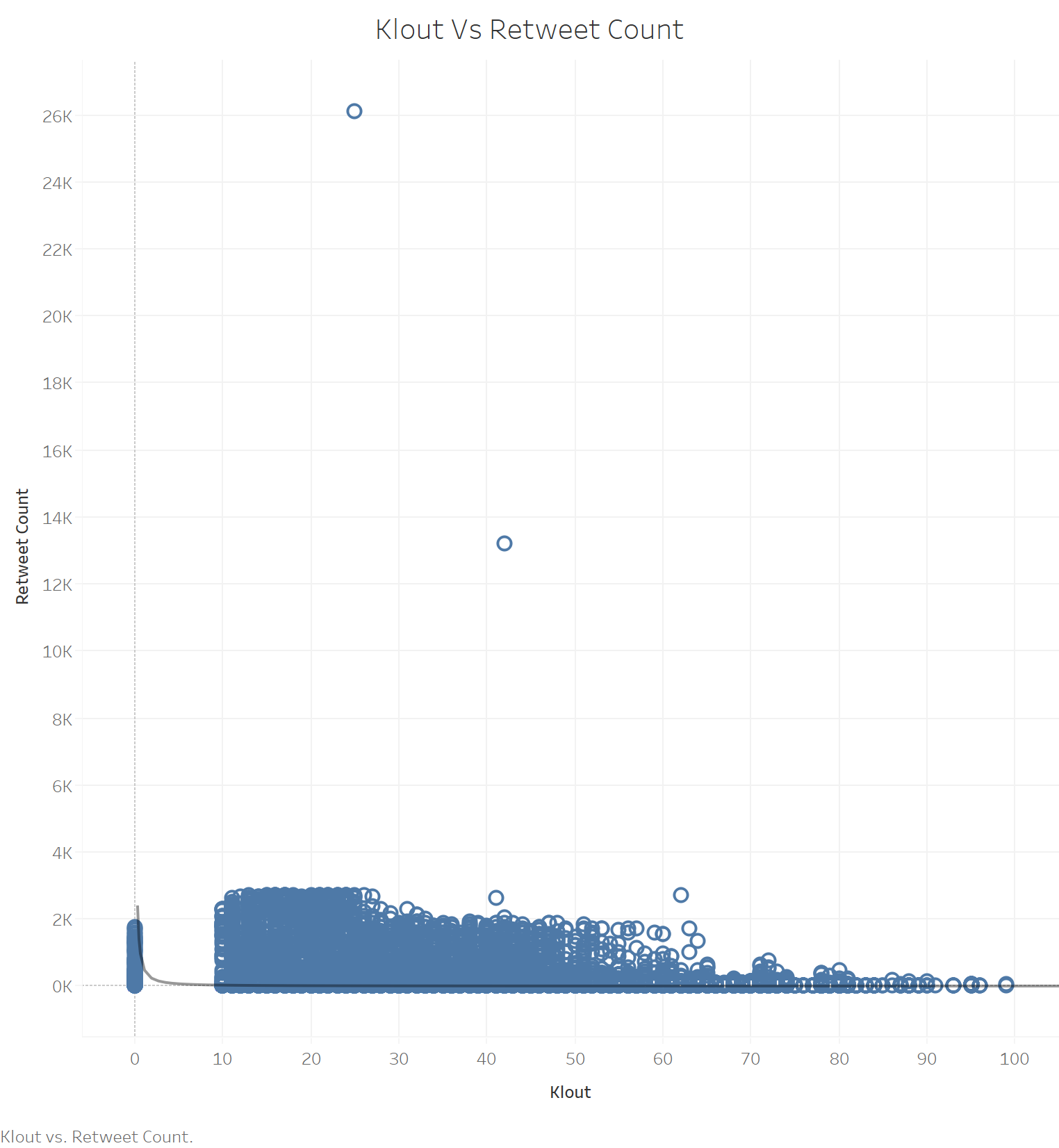
Data Exploration Report

This report details the relationship between each feature with the label ‘ReTweet Counter’.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Analysis** | **Effect size** | **P-value** |
| Klout | R Squared | 0.077895 | < 0.0001 |
| Sentiment | R Squared | 0.0006413 | < 0.0001 |
| IsReshare | R Squared | 0.0813 | < 0.0001 |
| Reach | R Squared | 0.036418 | < 0.0001 |
| Likes | R Squared | 0.7863 | < 0.0001 |
| Weekday | F Stat | 562.75 | < 0.0001 |
| Hour | R Squared | 0.0082616 | < 0.0001 |
| Day | R Squared | 0.008198 | < 0.0001 |
| Gender | F Stat | 1849.002 | < 0.0001 |
| Text | R Squared | 0.04022 | < 0.0001 |
| Country | F Stat | 89.98019 | < 0.0001 |

## Klout

H1: Klout (or Klout Score) is expected to have a positive relation with the number of Retweets (directly proportional) [Rao IEEE, 2015]



Power Trendline is used in the above plot.

**Power equation:** ( ln(Klout) + intercept ) i.e. 538.523\*Klout^-1.28333  
**R2:** 0.077895 i.e. 7.789%  
**P-value** < 0.0001

|  |  |
| --- | --- |
| **P-value:** | < 0.0001 |
| **Equation:** | ln(Retweet Count) = -1.28333\*ln(Klout) + 6.28883 |

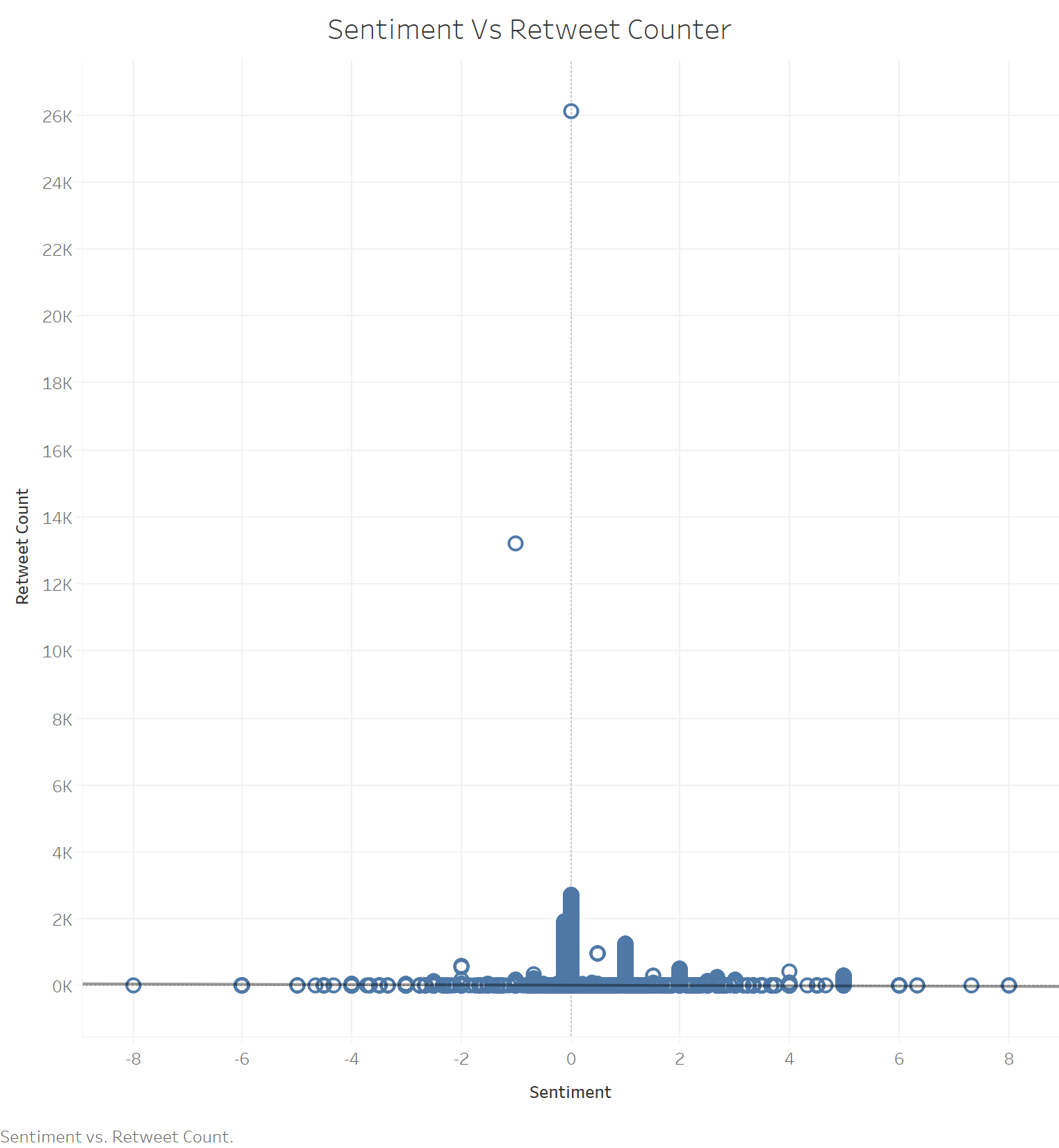
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Coefficients** | | | |  |
| **Term** | **Value** | **StdErr** | **t-value** | **p-value** |
| ln(Klout) | -1.28333 | 0.0137451 | -93.3664 | < 0.0001 |
| intercept | 6.28883 | 0.0507498 | 123.918 | < 0.0001 |

**Summary**

Klout appears to have a positive effect size on Retweets which is very reliable [P-value < 0.001]. We tested all non-linear transformations and found that a power [1.25954] transformation provided a better R2 value [0.077895]. In addition, the visual plot infers that the Retweets decreases as the Klout Score increases [except for a few values].

# Sentiment

H2: Sentiment has a positive effect on the ReTweet Count [Chen, IEEE, 2017]



The above plot uses a linear trendline.

**Linear Equation:** -4.48681\*Sentiment + 26.294  
**R2:** 0.0006413   
**P-value** < 0.0001

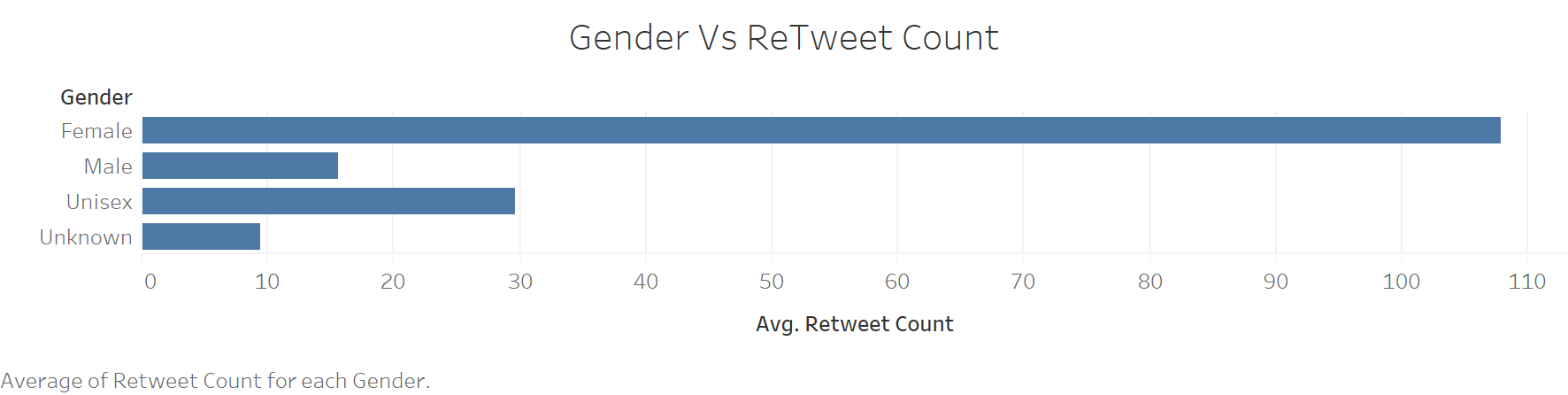
**Summary**

Sentiment has a positive relation with ReTweets Count. Linear transformation provided comparatively a better R2 value against other trendlines. In addition, inference seems to be reliable as the value of P-value < 0.0001. In addition, the plot is slightly skewed to the positive sentiment values. Neutral sentiment values (0 < Sentiment Value < 1) have the highest ReTweets.

# Gender

H40: There is a relation between Gender and Retweets Count.

H4A: There is no relation between Gender and Retweets Count.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Groups** | **Count** | **Sum** | **Average** | **Variance** |
| Female | 21636 | 2447091 | 113.1027 | 191902.4 |
| Male | 95252 | 1524660 | 16.00659 | 18205.06 |
| Unisex | 16554 | 512655 | 30.96865 | 41175.69 |
| Unknown | 66558 | 638619 | 9.594925 | 5310.074 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Source of Variation** | **SS** | **df** | **MS** | **F** | **P-value** | **F crit** |
| Between Groups | 191954130.3 | 3 | 63984710 | 1849.002 | 0 | 2.604954 |
| Within Groups | 6920861611 | 199996 | 34605 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 7112815741 | 199999 |  |  |  |  |

One-way ANOVA

**F-STAT:** 1849.002

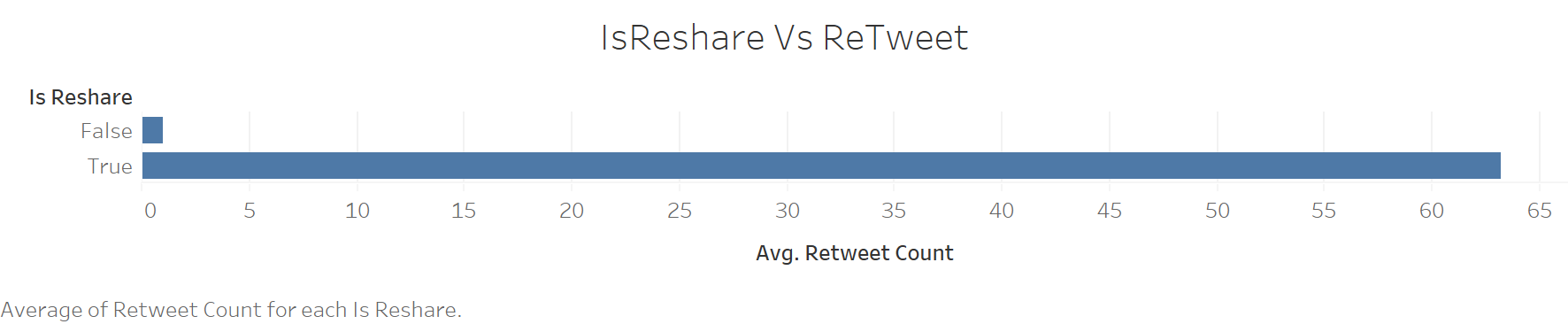
**P-value** < 0.0001

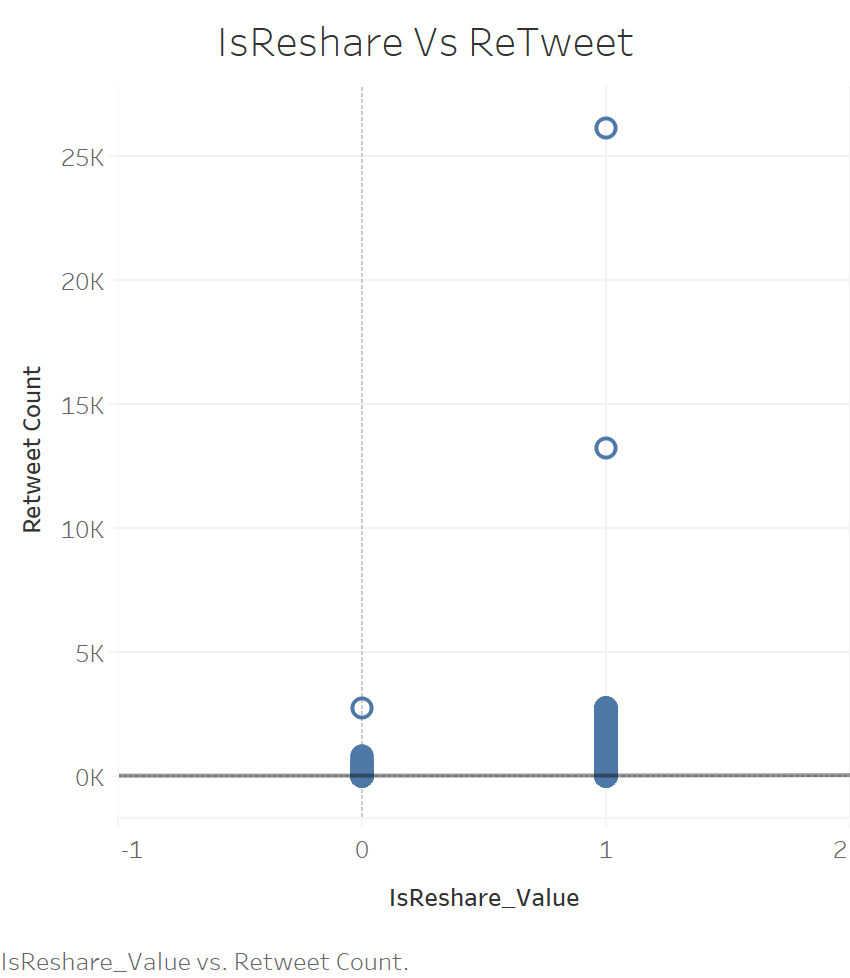
**Summary**

Gender has a positive impact on ReTweet Counter. With an average score of 113.1027, Female category has the highest ReTweet average as compared to other categories.

# IsReshare

H5: The ReTweet Counter for Reshares are likely to be higher.





**Exponential Equation**: ln(Retweet Count) = 1.12297\*IsReshare\_Value + 0.716368

**R2**: 0.0813

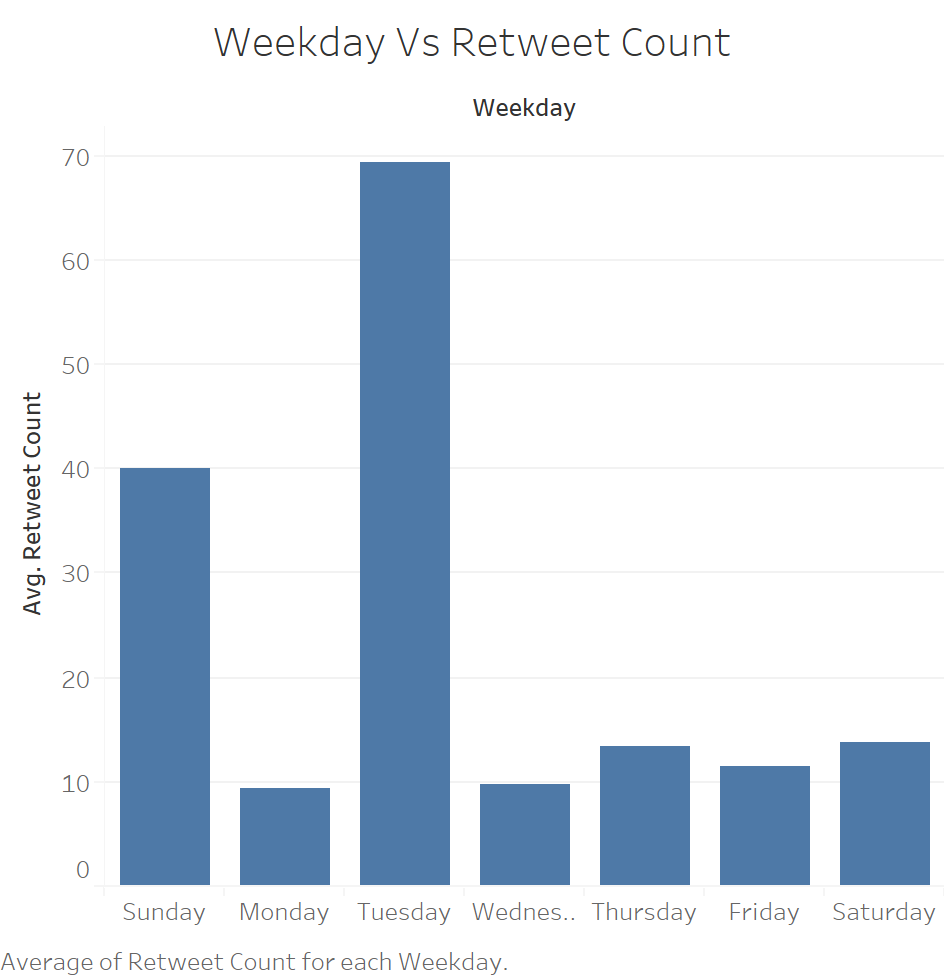
**P-value** < 0.0001

**Summary**

IsReshare is a Boolean feature. A new calculated field is created (IsReshare\_Value) that has 0 for False and 1 for True. The new feature is plotted against ReTweet Count. With an R square value of 0.0813, IsReshare has a small positive effect on the label ReTweet Count. The P-value < 0.0001 makes this hypothesis reliable.

# Weekday

H6: ReTweets should be high during after business hours and during weekends (subjected to time zone differences)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Groups*** | ***Count*** | ***Sum*** | ***Average*** | ***Variance*** |
| Friday | 29999 | 347237 | 11.57495 | 27625.35 |
| Monday | 31471 | 298937 | 9.498808 | 4403.732 |
| Saturday | 15345 | 211601 | 13.78957 | 6397.357 |
| Sunday | 16501 | 671866 | 40.71668 | 43356.64 |
| Thursday | 38217 | 521253 | 13.6393 | 11440.58 |
| Tuesday | 39388 | 2776044 | 70.47944 | 119319.5 |
| Wednesday | 36929 | 365389 | 9.894365 | 2368.072 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 113801496.1 | 6 | 18966916 | **562.75** | 0 | 2.098641 |
| Within Groups | 7005136623 | 207843 | 33703.98 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 7118938119 | 207849 |  |  |  |  |

One Way Anova

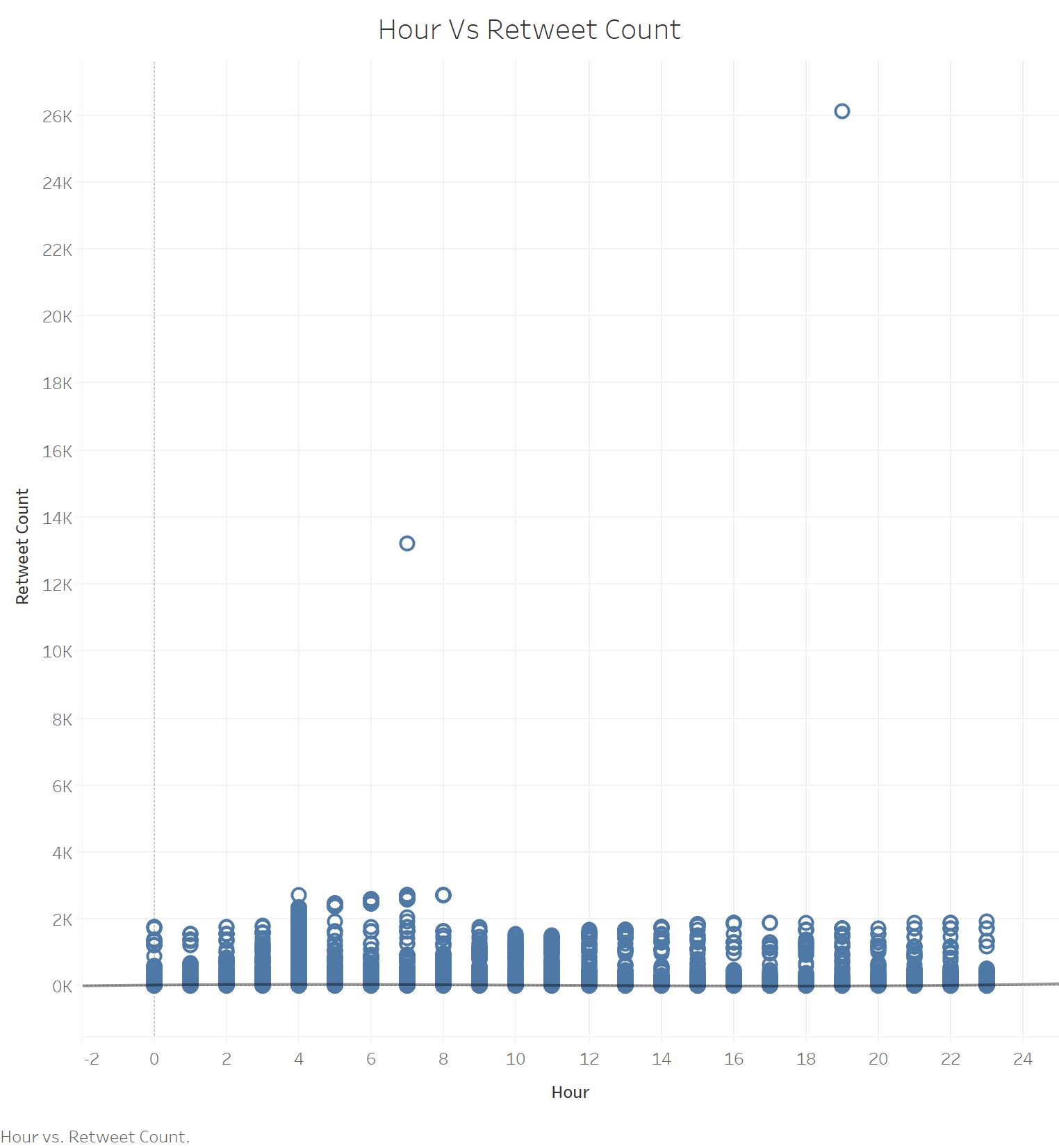
**F- STAT:** 562.75  
**P-value** < 0.0001

**Summary**

Weekday has a positive impact on the ReTweets Counter. As expected (and mentioned in the hypothesis), Sunday has the highest average (Average of 40.7 which is significantly bigger than other Weekdays). Sunday is followed by Tuesday which

# Hour

H1: I believe that time of the day (hour) will have an impact on the number of ReTweets on twitter. If a tweet is posted at a specific time of the day when users are likely to be ‘using’ twitter; the probability of the tweet reaching more users is higher. Factor like time zones, bots and twitter feed algorithm are not considered.



The above plot has uses a polynomial trendline.

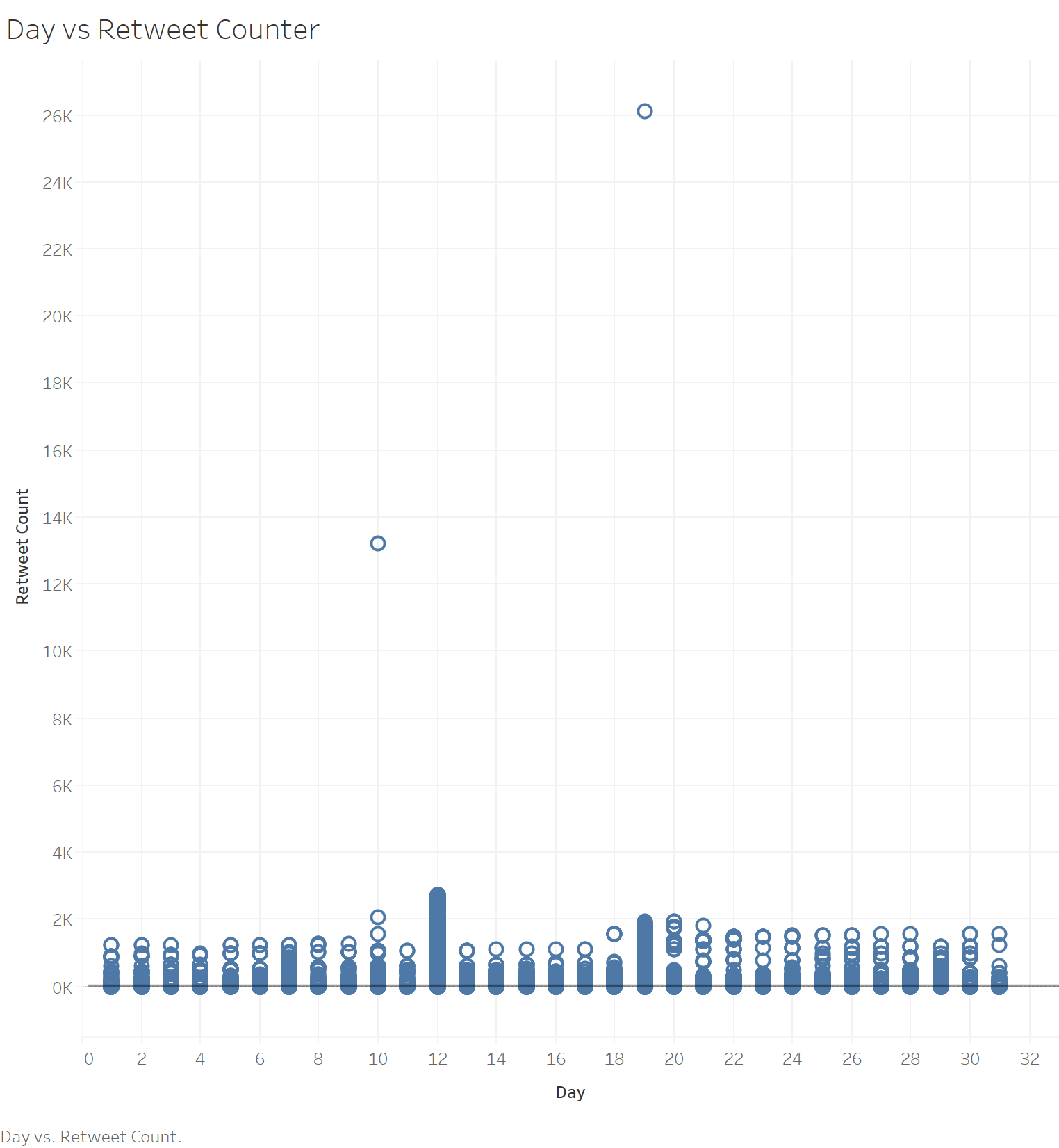
R2 is obtained to 0.0082616 with a P value < 0.0001

**Summary**

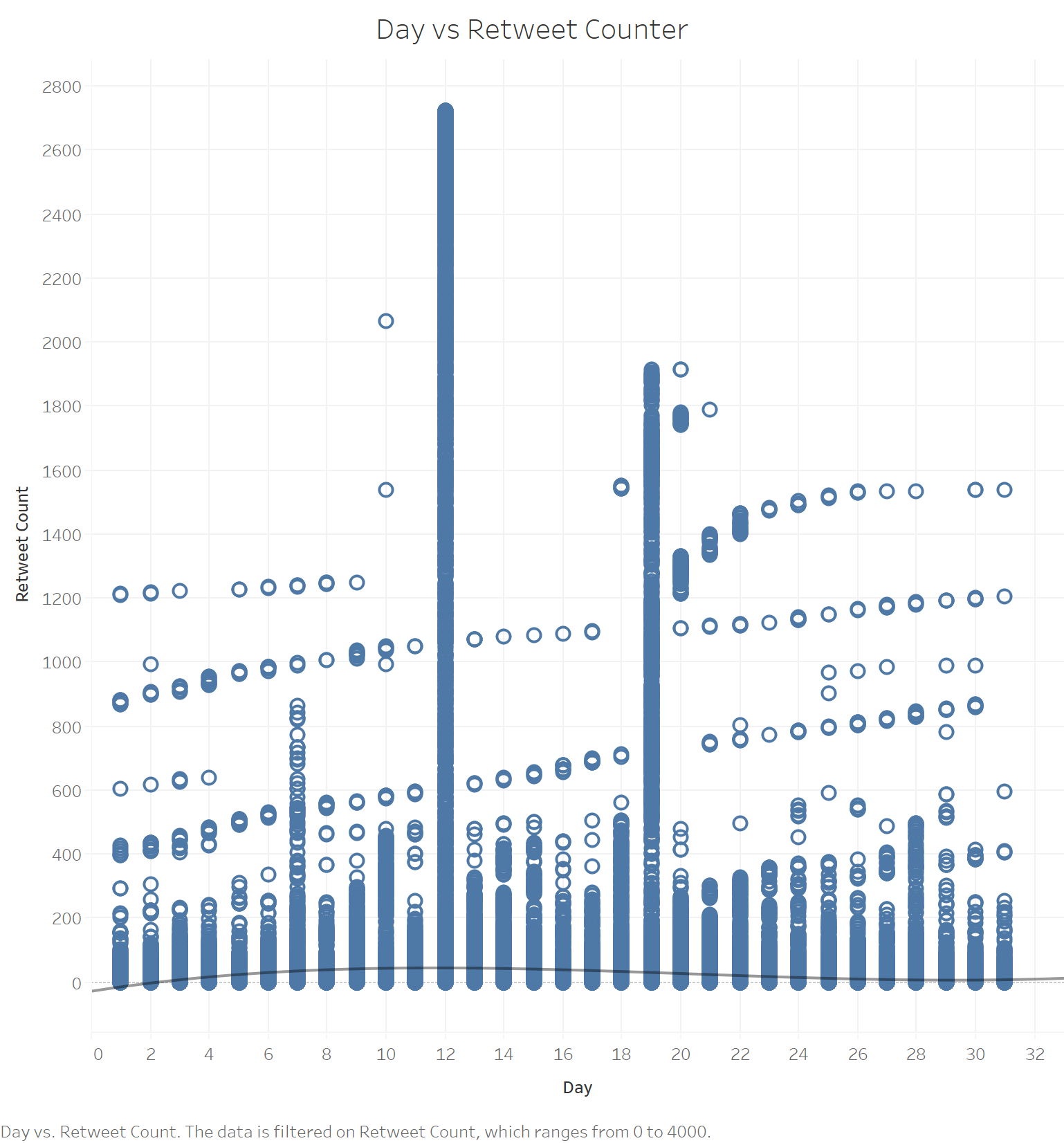
We compared all the trendlines. Polynomial trendline has the highest R2 value. However, the obtained R2 value is too small. In addition, the plot reveals that there is ‘almost’ a uniform distribution with the number of Retweets and the Hour. This concludes that Hour has just a small positive effect on the label.

# Day

H7: Day has no effect on ReTweet Counter.



Above is the initial plot generated for all data. Due to outliers on two days, I added a filter to evaluate ReCounter for a maximum of 4000. on two days, I added a filter to evaluate ReCounter for a maximum of 4000. New obtained plot is added below.



**Polynomial Model Equation:** 0.0140229\*Day^3 + -0.861636\*Day^2 + 14.392\*Day + -29.5566

**R2:** 0.008198

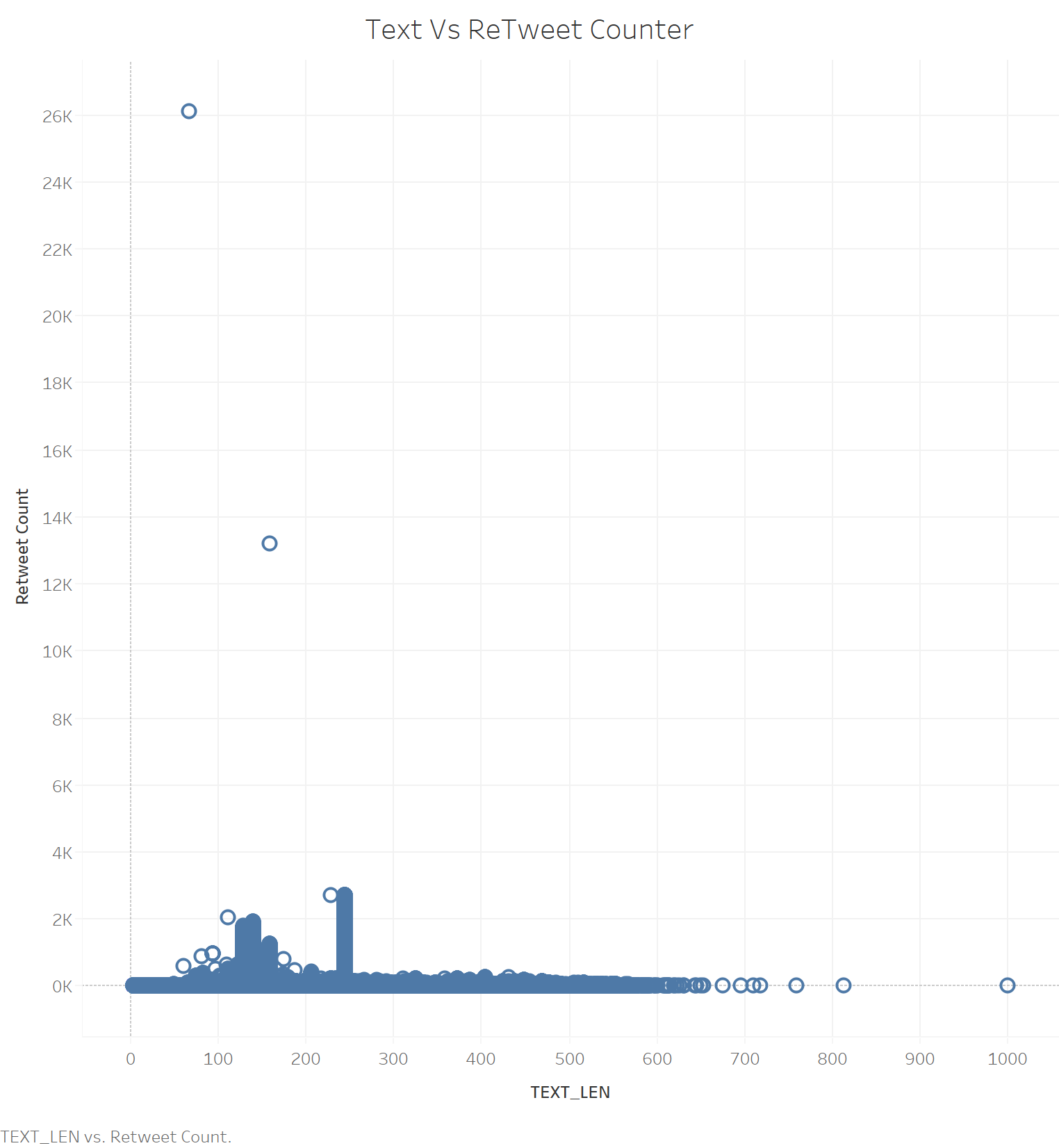
**P-Value** < 0.0001

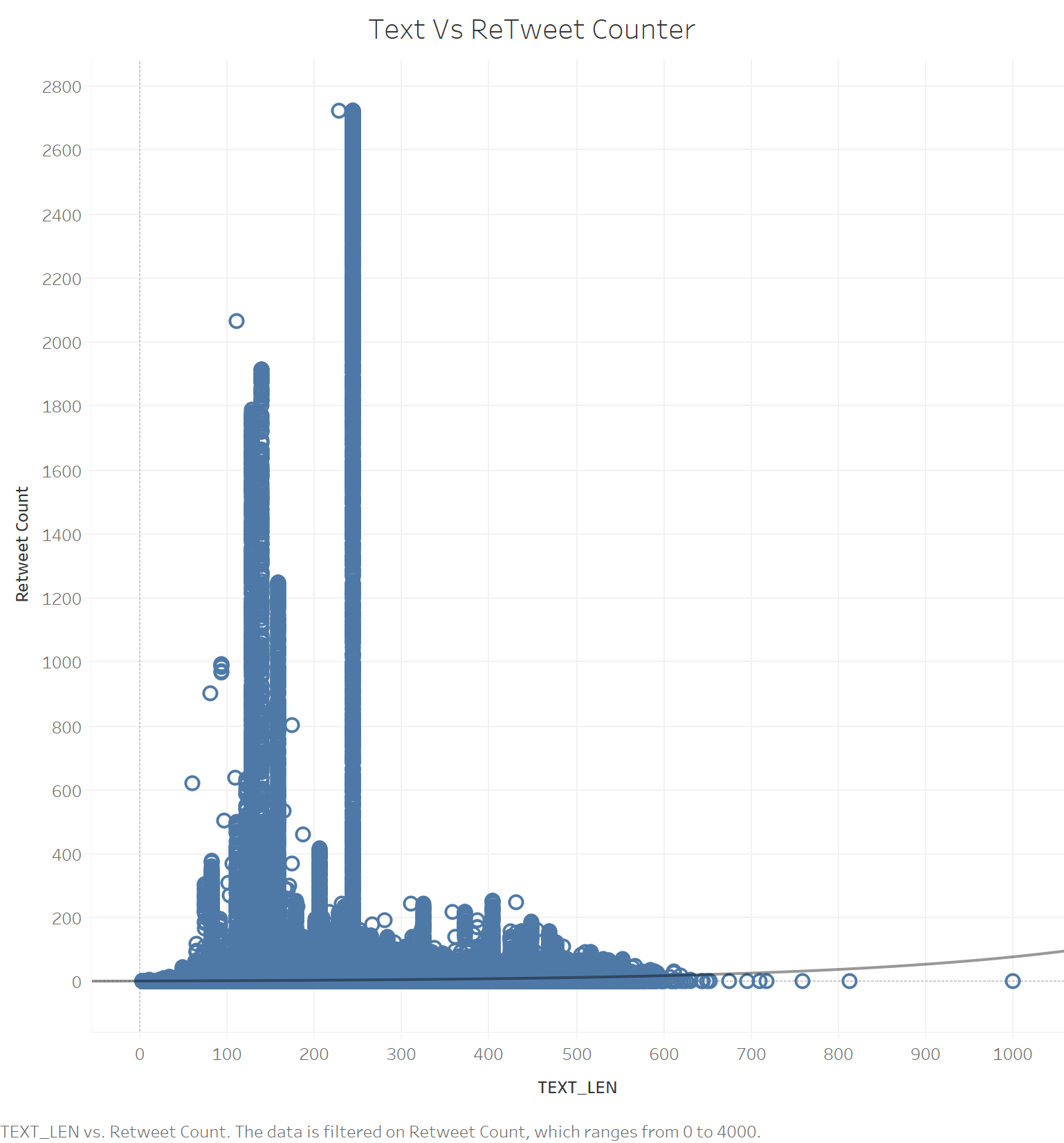
**Summary**

Day has no (very little) significance with ReTweet Counter. No skewness is found on the plot. There are two days [12th and 19th] having the highest average of ReTweet Counts. Polynomial transformation seems to be best fit with an R2 value of 0.008198 which is reliable as the P Value is less than 0.0001. To conclude, day as a feature do not have a significant impact on the label.

# Text

H8: Text has no significance/effect with ReTweet Counter.





**Equation:** ln(Retweet Count) = 0.00352463\*TEXT\_LEN + 0.831989

**R2:** 0.04022

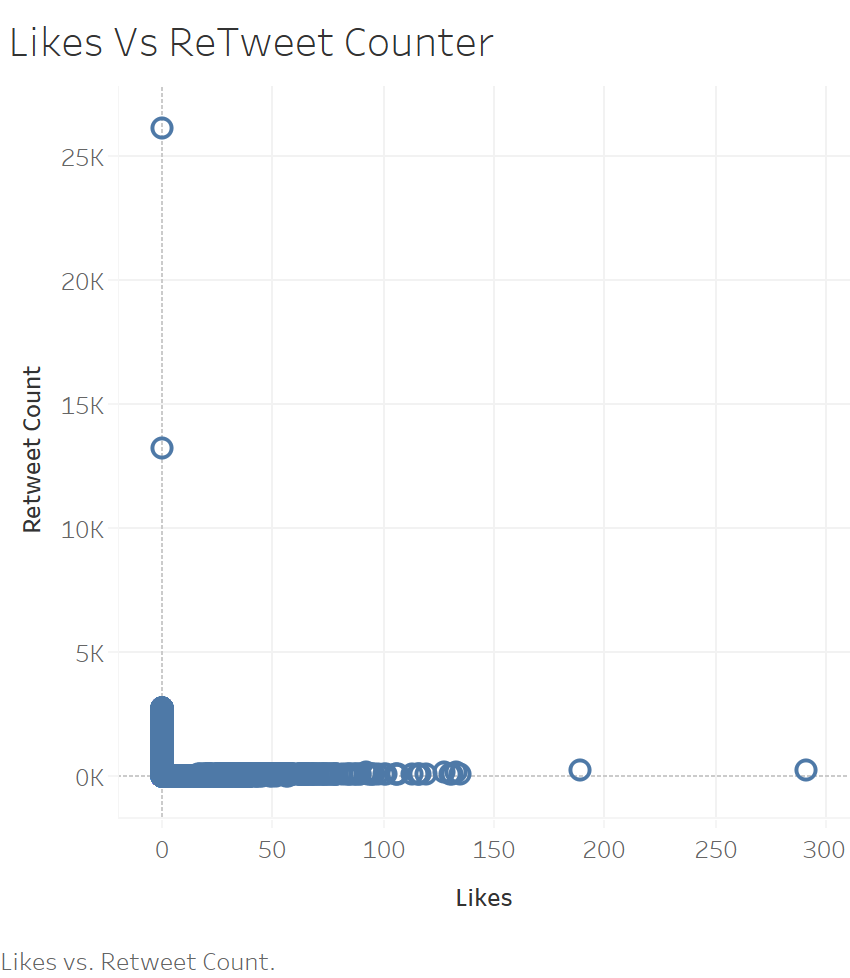
**P-value** < 0.0001

**Summary**

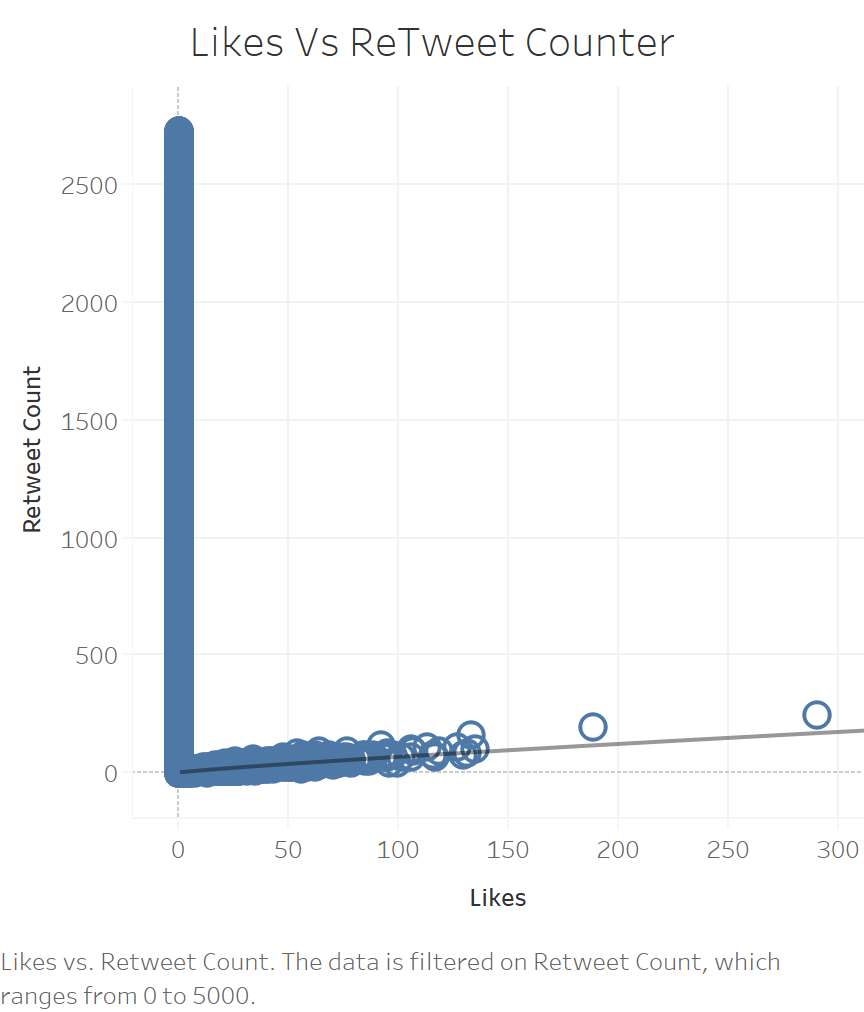
Text was initially used against ReTweet Counter. A better approach would be to create a new calculated field to count the Tweet (string) length. The new feature is TEXT\_LEN which is plotted with ReTweet Counter. It is observed that tweets with a character length less than 300 has the highest ReTweet Counts. Exponential transformation is used on the plot that provides an R2 value of 0.04022 which is reliable as the P-value is less than 0.0001. Hence, text (TEXT\_LEN) has a small effect on Retweet Counts.

# Likes

H9: The tweets more likes will be retweeted more (directly proportional).



ReTweets Counter Outliers over 5K are removed. Below is the new plot.



**Power Equation**: ln(Retweet Count) = 0.86427\*ln(Likes) + 0.220683

**R2**: 0.7863

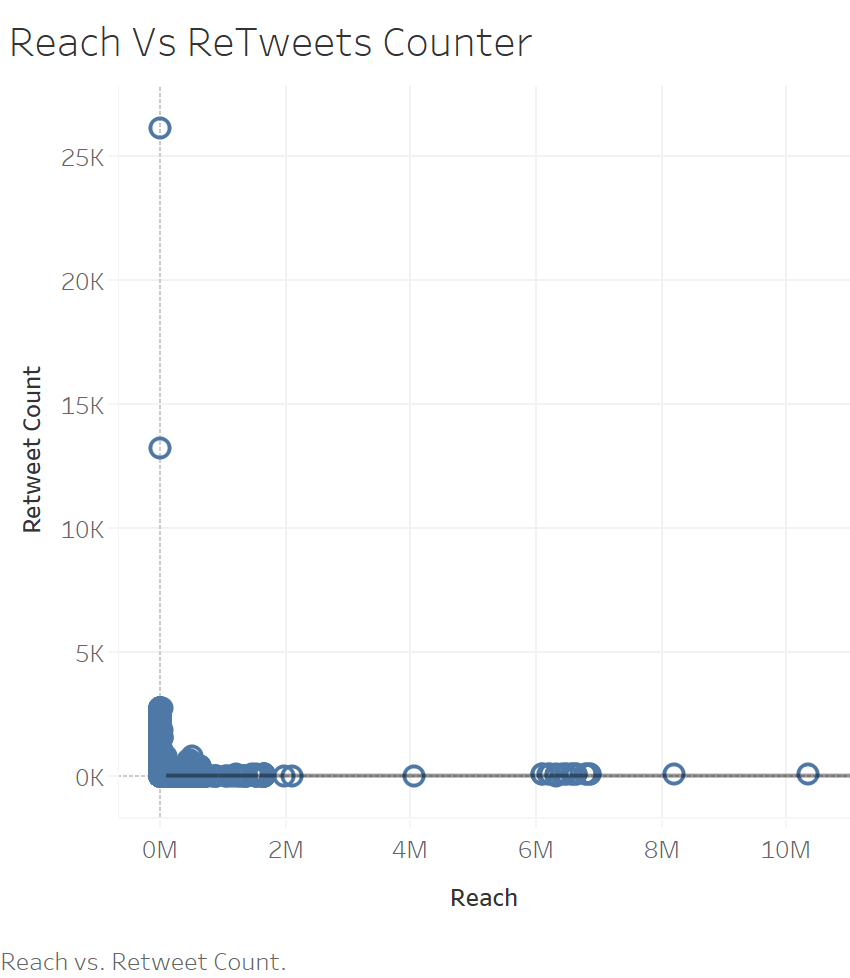
**P-value** < 0.0001

**Summary**

Likes has a very large positive effect on the label ReTweet Counter. R square value is highest (0.7863) compared to all other features used so far. This also looks reliable with a P-value < 0.0001. Power transformation is used here. Conclusion is Likes has a large effect on the label.

# Reach

H10: Profiles with a larger reach will have more ReTweets



**Power Equation**: ln(Retweet Count) = -0.154016\*ln(Reach) + 2.55252

**R2**: 0.036418

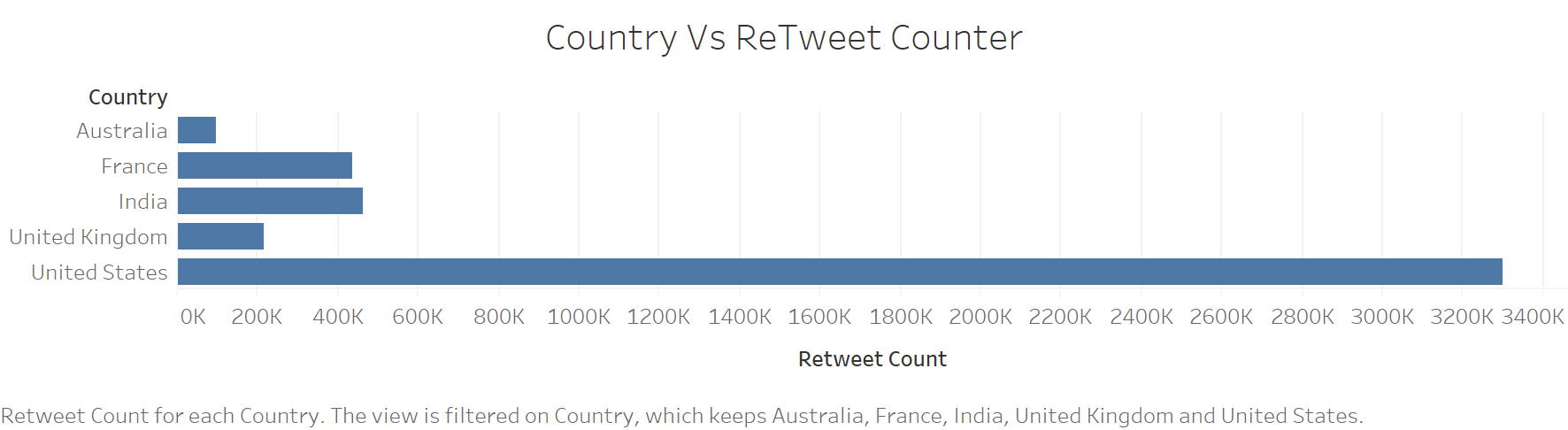
**P-value** < 0.0001

**Summary**

Contrary to the initial hypothesis, Reach does not have a large positive effect on ReTweets Counter. R2 value is not very impressive (0.036418). Power transformation is applied here. Conclusion is Reach does not have a large effect on the label (except for a few tweets having a large retweets count).

# Country

H11: The number of retweets vary based on countries



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Groups** | **Count** | **Sum** | **Average** | **Variance** |
| Australia | 7498 | 93456 | 12.46412 | 7476.883 |
| France | 8524 | 431615 | 50.63527 | 52995.67 |
| India | 13943 | 461576 | 33.1045 | 23244.25 |
| United Kingdom | 22427 | 212242 | 9.463682 | 4670.022 |
| United States | 105102 | 3297392 | 31.37326 | 55181.54 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 15394652 | 4 | 3848663 | **89.98019** | 1.54871E-76 | 2.371989 |
| Within Groups | 6.74E+09 | 157489 | 42772.34 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 6.75E+09 | 157493 |  |  |  |  |

**Summary**

With an F-Stat score of 89.98019, Country has a positive impact on the number of ReTweets. The top five countries contribute to a significant portion of the Retweets Counter. The P-value looks reliable (< 0.0001)

Reference:

Rao, Adithya, et al. "Klout score: Measuring influence across multiple social networks." *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015.

Chen, Jundong, et al. "Sentiment analysis of the correlation between regular tweets and retweets." *2017 IEEE 16th international symposium on network computing and applications (NCA)*. IEEE, 2017.