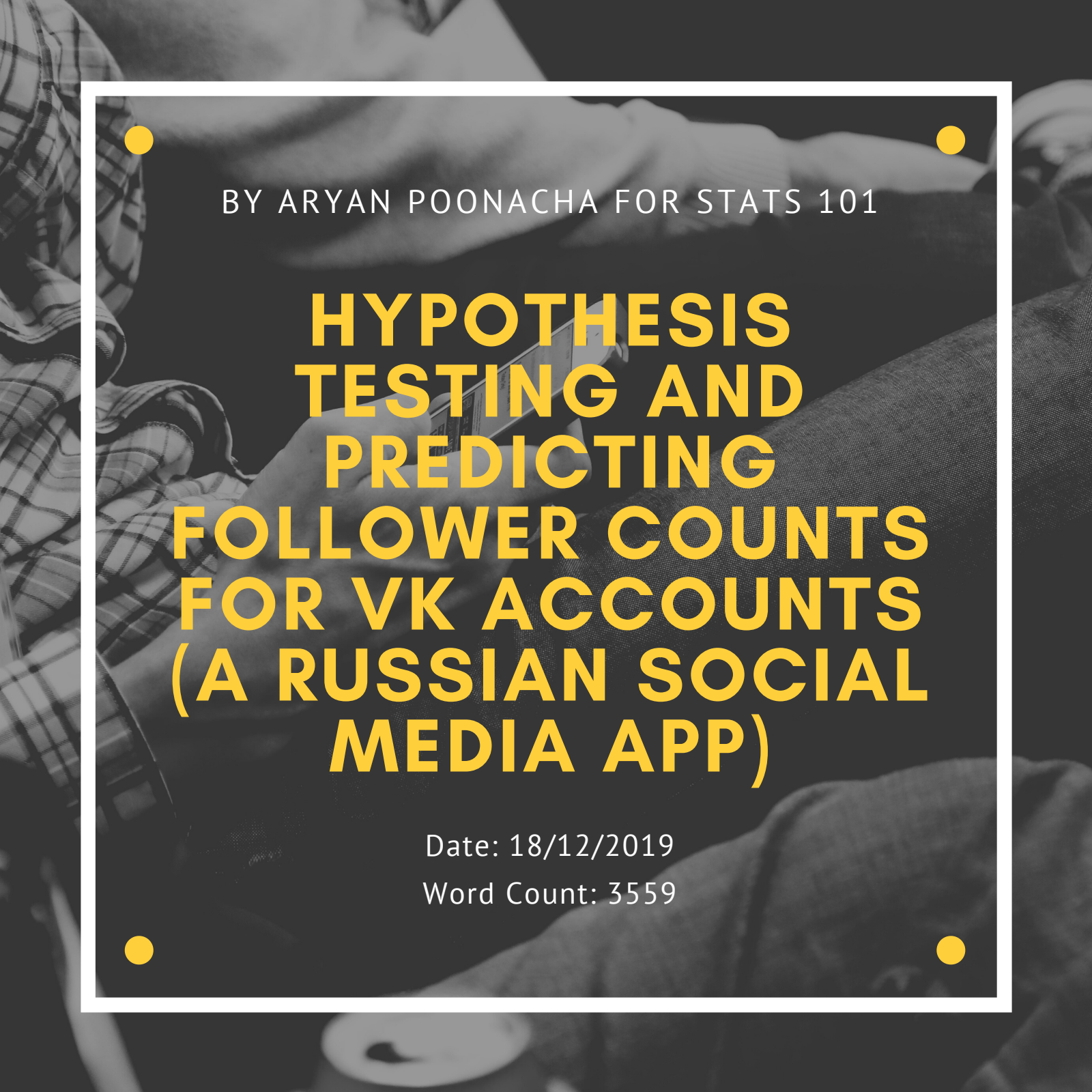


BY ARYAN POONACHA FOR STATS 101

HYPOTHESIS TESTING AND PREDICTING FOLLOWER COUNTS FOR VK ACCOUNTS (A RUSSIAN SOCIAL MEDIA APP)

Date: 18/12/2019

Word Count: 3559

Introduction

VK is a Russian online social media and social networking service predominantly used by Russian-speakers. Like most social media platforms, VK users have the feature to follow other accounts and pages. Follower counts have many positive and negative stigmas attached to them, and are often used by people as a metric of popularity, so predicting a particular account's potential follower count knowing other factors is useful, particularly in cases of corporate advertising and social media outreach. VK, in particular, has not been given as much attention as the other social media giants, and is thus an ideal platform for such a model.

Briefly, the results of this report are as follows:

1. Either gender has no advantage for having more followers on VK overall,
2. Having at least some comments on your VK posts makes you more likely to have more followers than having no comments on any of your posts,
3. More posts is the best way to increase your followers, followed closely by more friends. More likes also helps slightly.
4. Reposts and unengaging content can reduce your follower count.

Hypothesis Development & Literature Review

'Hacking' follower counts in social media has become a major subject of interest in recent times. Follower counts in social media and influencers have made big headlines, being the cause behind costing companies billions in fraud[1] and stealing people's identities[2]; thus, methods of analysing how to maximise one's follower counts legitimately and as efficiently as possible are highly sought after.

Research into predicting follower counts and follower count growth using other data of the account is not a novel idea, and has already been directly conducted for twitter profiles with marginal success.[3] However, the models employed by these papers vary, and often, a simple relationship that can be predicted by linear regression is insufficiently accurate for helping grow

[1] https://www.cbsnews.com/news/influencer-marketing-fraud-costs-companies-1-3-billion/
[2] https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html
[3] Mueller, Juergen, and Gerd Stumme. "Predicting Rising Follower Counts on Twitter Using Profile Information." *Proceedings of the 2017 ACM on Web Science Conference - WebSci 17*, May 9, 2017. https://doi.org/10.1145/3091478.3091490.

social media presence, and models such as negative binomial auto-regression[4], machine learning and other models are used to predict follower counts for specific actions.[5]
However, these studies concluded that "the relative contributions of social behavior and message content are just as impactful as factors related to social network structure for predicting growth of online social networks" [6], i.e, how many followers a particular social media account obtains depends on how active, engaging and relevant the content of their social media account is, and not just how popular the account owner is.

Thus, although we are using simpler linear regression models, we can be confident that factors that measure the activeness/engagement of an account, such as the number of posts or photos posted in the account, will be useful in predicting the number of followers, besides metrics that already measure the account's popularity (such as the number of likes on their posts/photos).

For each variable, the following are expectations/hypotheses for their relationships to follower count and how strong we expect the relationship to be:

Number of Friends-Positively correlated; the more friends an account has, the more popular it is, and the more followers it will have. However, it's also possible that friends and followers have little relation as friends will only be personal acquaintances, and followers people that are only interested in the account.
Number of Posts-Positively correlated. An account that is more engaging and active will simply have more followers.
Number of Reposts-Similarly positively correlated to number of posts, if slightly less so as it is not as engaging if it has already been posted before.
Number of Photos-Positive correlation, as it's more engagement.
Number of Videos-Also a positive correlation, as it's more engagement.
Number of Photo Likes-Likes are commonly used as metrics of popularity, so the number of photo likes will be the strongest indicator of the number of followers that an account has.
Sex-Will be irrelevant, as the number of accounts of each gender will be likely close to evenly split by gender.

---

[4] Hutto, C.j., Sarita Yardi, and Eric Gilbert. "A Longitudinal Study of Follow Predictors on Twitter." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*, 2013. https://doi.org/10.1145/2470654.2470771.
[5]Wang, Ke, Mohit Bansal, and Jan-Michael Frahm. "Retweet Wars: Tweet Popularity Prediction via Dynamic Multimodal Regression." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. https://doi.org/10.1109/wacv.2018.00204.
[6] Hutto, C.j., Sarita Yardi, and Eric Gilbert. "A Longitudinal Study of Follow Predictors on Twitter." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*, 2013. https://doi.org/10.1145/2470654.2470771.

Has Comments - If the post has comments, it's more likely to be engaging, so it will be positively correlated with follower count.

Moreover, we also seek to formally conduct tests and create hypotheses to see if one gender is more likely to have a higher or lower follower count, and to confirm or deny whether or not having more comments implies a greater follower count.

For sex:

Null Hypothesis H0: pFollowersOfMale - pFollowersOfFemale = 0

Alternate Hypothesis Ha: pFollowersOfMale - pFollowersOfFemale != 0

It's a 2-tailed test because we want to see if either gender has an advantage.

For comments:
Null Hypothesis H0: pFollowersForComments - pFollowersForNoComments = 0

Alternate Hypothesis Ha: pFollowersForComments - pFollowersForNoComments > 0

It's one tailed because we only want to know if having comments can cause an account to have more followers.

Before we build the models, for all our hypotheses, we set a value of alpha = 0.05.

Variable Exploration & Summary Statistics

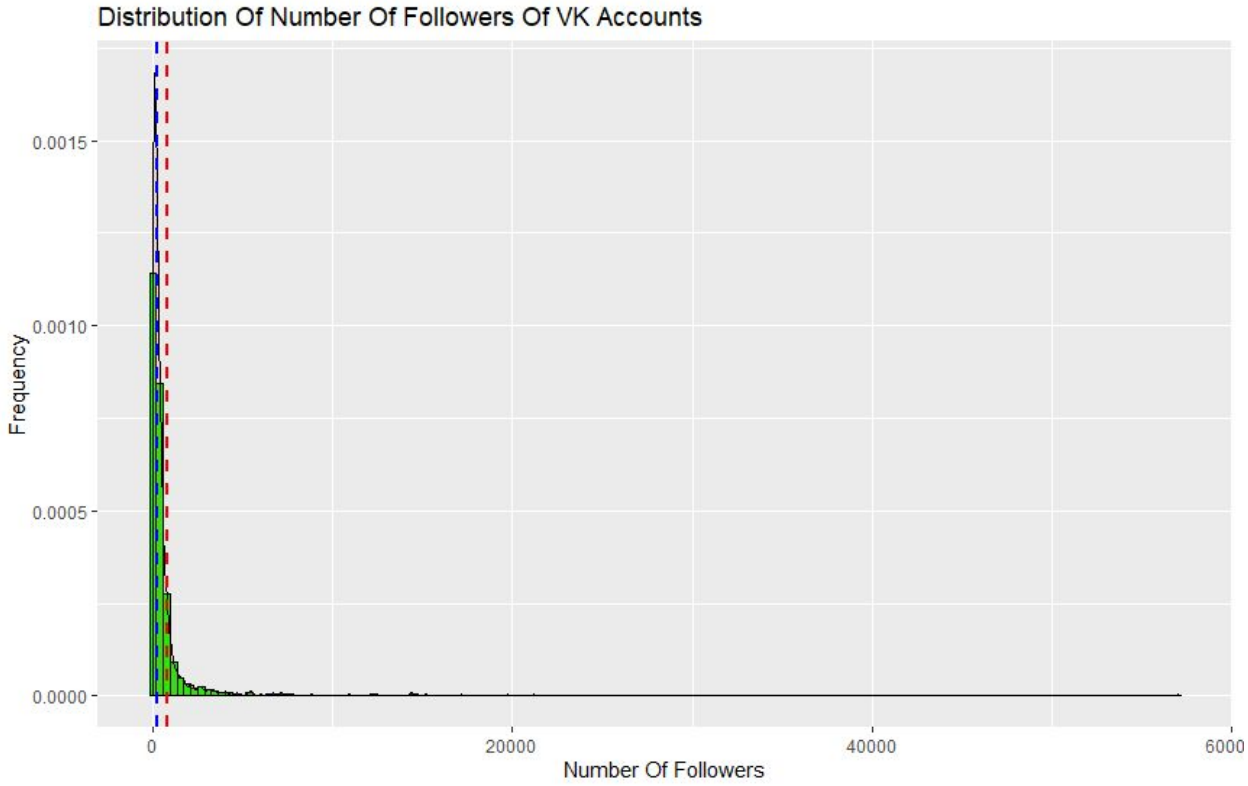
Brief summaries/explanations of the variables are:

Number of Friends-To friend an account, a friend request must be sent and accepted.
Number of Followers-Anyone can follow any account to receive updates on its posts and activity.
Number of Posts-The number of posts posted by the user since the account's creation. Posts can be edited after they're made (eg: an album that's constantly appended to is still considered one post). A profile shows all the post, and then each post can be seen in more detail.
Number of Reposts-Posts that have been posted before (like sharing a memory).
Number of Photos-Number of photos posted.
Number of Videos-Number of videos posted.
Number of Photo Likes-Sum of the likes received on all the posts combined.
Sex-The gender of the account owner. 0 if female, 1 if male.

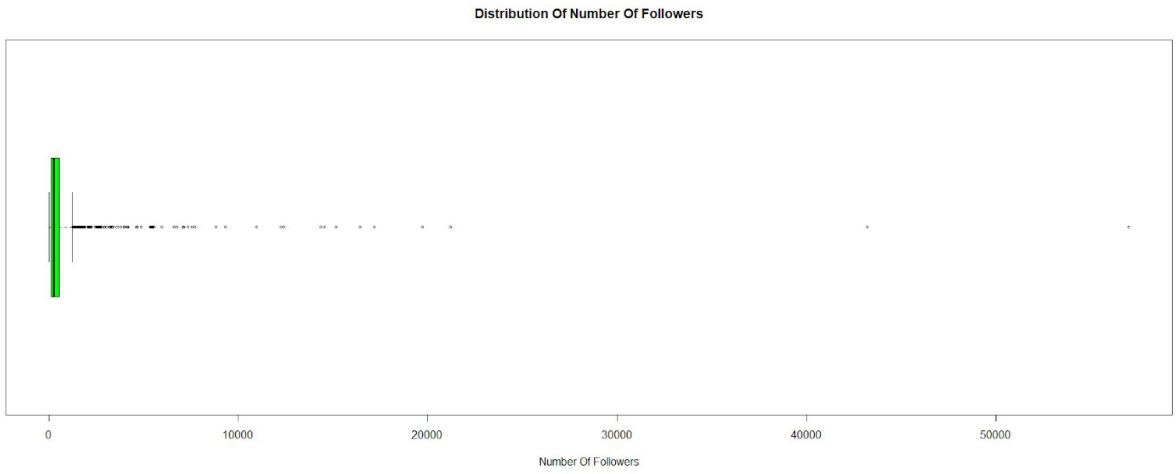
HasComments - 0 if none of the posts have any comments,1 if any post has at least one comment.

Response Variable Summary Statistics
We can analyse the distribution of our response variable:

Key: Red line is mean, Blue Line is Median



The variable is clearly strongly right skewed and unimodal, but only due to a few extreme outliers.



There are 2 very significant outliers with 40,000+ and 50,000+ followers respectively. These outliers are not anomalies or errors in the data, but actual accounts with large follower counts. Thus, it's important that we keep them in the dataset.

(All Units are counts)	Followers
Min	0
Max	57037
Range	57037
Lower Quartile Q1	85.25
Upper Quartile Q3	549.75
IQR = Q3 – Q1	464.5
Standard Deviation	2926.421

75% of all the accounts have between 85 and 550 followers; however, as we can see from the Standard Deviation, there is a large spread to the data overall, even though most datasets like within an IQR of only 464.5.

Centers:

(All Units are counts)	Followers
Mean	807.10
Median	233

The large difference between the mean and the median further affirms that the data has a large spread, and that there are a few very significant outliers with very large follower counts that greatly impact the mean.

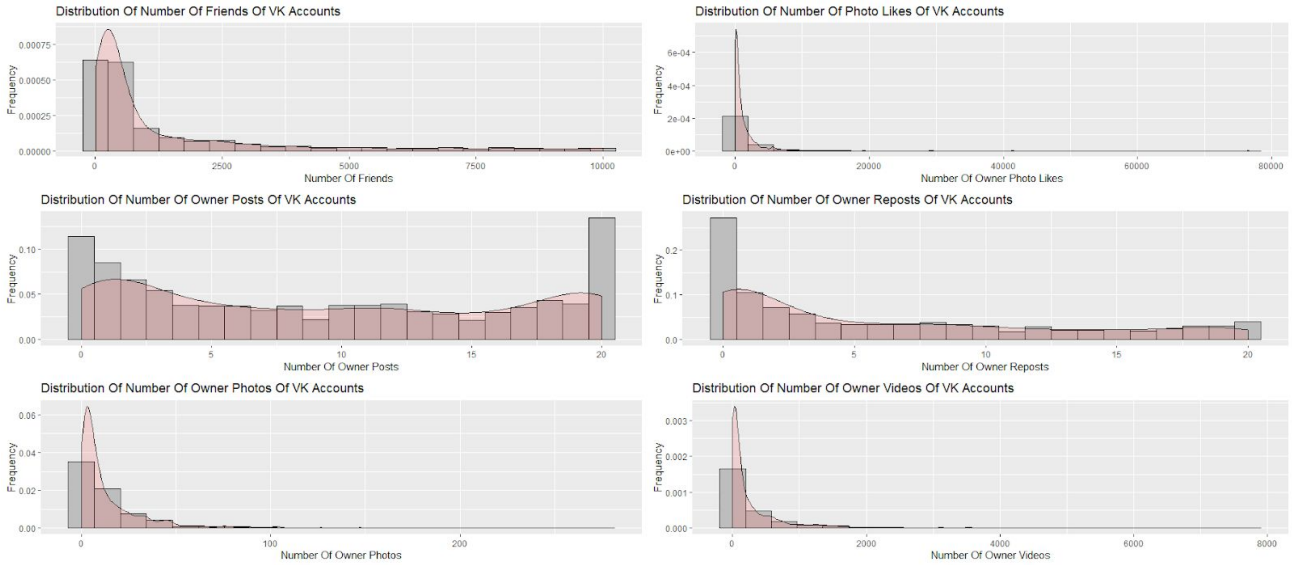
Predictor Variable Summary Statistics

Method Of Data Collection

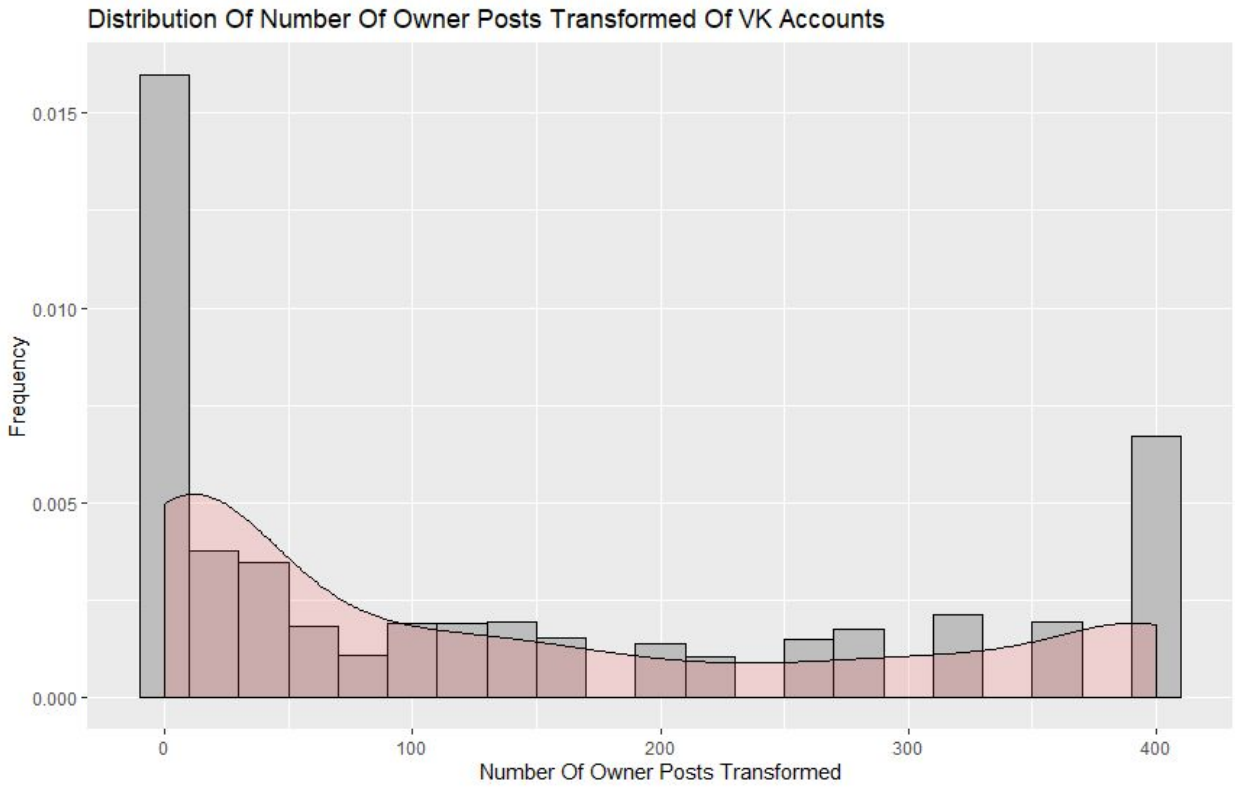
The data is taken from Kaggle, an online repository for public use databases: https://www.kaggle.com/alextar04/vk-social-network-dataset

The data itself is obtained by web scraping and other methods of public data collection from VK such as APIs, etc.

We can explore the distribution of all of the possible predictor variables via histograms:



All the above distributions are strongly right skewed due to the presence of a few outliers, but are otherwise mostly normal and should not cause issues with later statistical tests. However, the number of owner posts is oddly seemingly bimodal, with an otherwise almost uniform distribution. This could cause problems with using it as a predictor variable. Since the scale of the 2 axes are so different for the distribution of the number of posts (x is from 0 to 20, y is from 0 to 50,000), we can transform the number of posts variable by squaring it. We can then analyse the distribution of the transformed variable:



The range is now from 0 to 400, and the Standard Deviation is 148.5187. Although the values are now more spread apart, the distribution is now more suitable for fulfilling the equal variance condition as it makes it a better normal distribution; so we will use the posts variable from now on in its transformed state.
(Note: We can try to cube the variable to again improve the suitability for a linear model, but then the SD becomes 2950; the variables will be spaced too far. A square transformation is thus the most suitable).

We can also analyse some of the summary statistics of the 2 categorical variables to understand their distributions:

1. Comparisons of Centres:

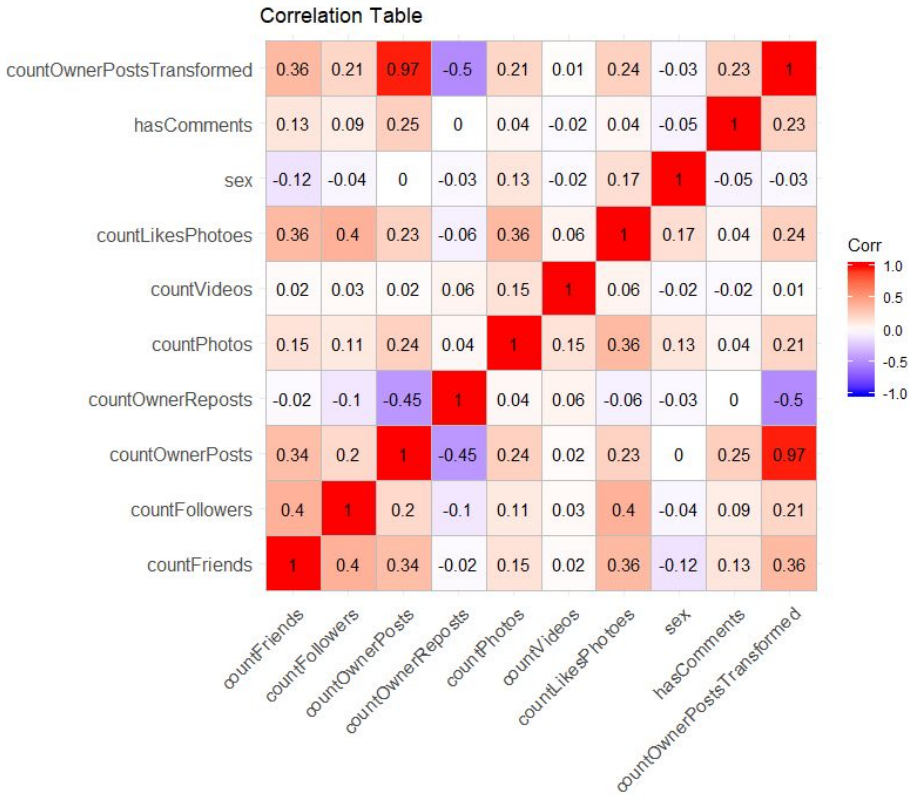
	Mean (%)
Sex	41.2% Male
Has comments	41.8% Have Comments

More than half of the users are female, and more than half of all the posts don't have any comments.

Correlations
We first analyse the conditions for creating a correlation matrix:

Quantitative - All the variables we're using are quantitative and continuous.
Heteroscedasticity - From the distributions above, there is no clear subgroups around which the frequency distribution is particularly huddled around.
Outliers Condition - There are numerous significant outliers in all the variables, but all of them are valid data points for creating the correlation matrix, and none are extreme outliers that would cause a massive change in correlation.
Straight Enough Condition: Since all the variables are nearly normal, they would naturally fit the straight enough condition if we plotted their scatterplots.

Thus, we can then analyse the correlation table to see which variables have the strongest correlation to the number of followers and to analyse variable relationships:



From the correlation plot, we can see that in accordance with the hypothesis, the number of friends and the number of photo likes have the clearest strongest correlation to the follower count of a particular account.

Surprisingly, number of reposts actually has a slight negative correlation(-0.1) with number of followers, which might indicate that negative engagement with repetitive content such as reposts can actually have a damaging influence on the follower count.

There are also a few other interesting correlations to note: the more number of posts there are, the less likely the account is to repost older posts (-0.45 correlation). One would expect people who post more to also repost more, but in reality they simply are more likely to post new things.

Additionally, unsurprisingly, number of friends, number of posts, number of likes on photos, and number of photos posted are all moderately correlated to each other; this is expected as they are all complementary to further engagement and use of the platform by an account.

Moreover, it should be noted that Russian social media engagement ranks higher than most other countries[7], with Russians spending an average of two hours and 16 minutes per day on social media[8]. This difference in engagement could possibly cause our follower count prediction model and hypothesis test results to vary from other similar research conducted on Western social media platforms like Twitter, where the engagement level is considerably weaker.

Thus, in order to decide which variables we want to use for our multivariate regression, we want to:

a) Minimise the average correlation between the predictor variables,
b) Maximise the R-squared value from the combined multivariate regression itself.

Table of Linear Model Summaries And Predictor Variable Correlations With Each Other

Variables Used	**R-Squared Value From Multivariate Regression**	**Average Correlation Between Predictor Variables**
countFriends + countLikesPhotoes + countOwnerPostsTransformed	23.51%	(0.36+0.36+0.24)/3 = 0.32
countFriends + countLikesPhotoes	23.34%	0.36

From the above table, it is clear that a combination of the 3 predictor variables produces the best R-squared value that also minimises the average correlation between the predictor variables. Thus, we use the number of friends, number of likes on their photos and the number of posts transformed to determine the number of followers that the person has.

Hypothesis Testing & Inferences

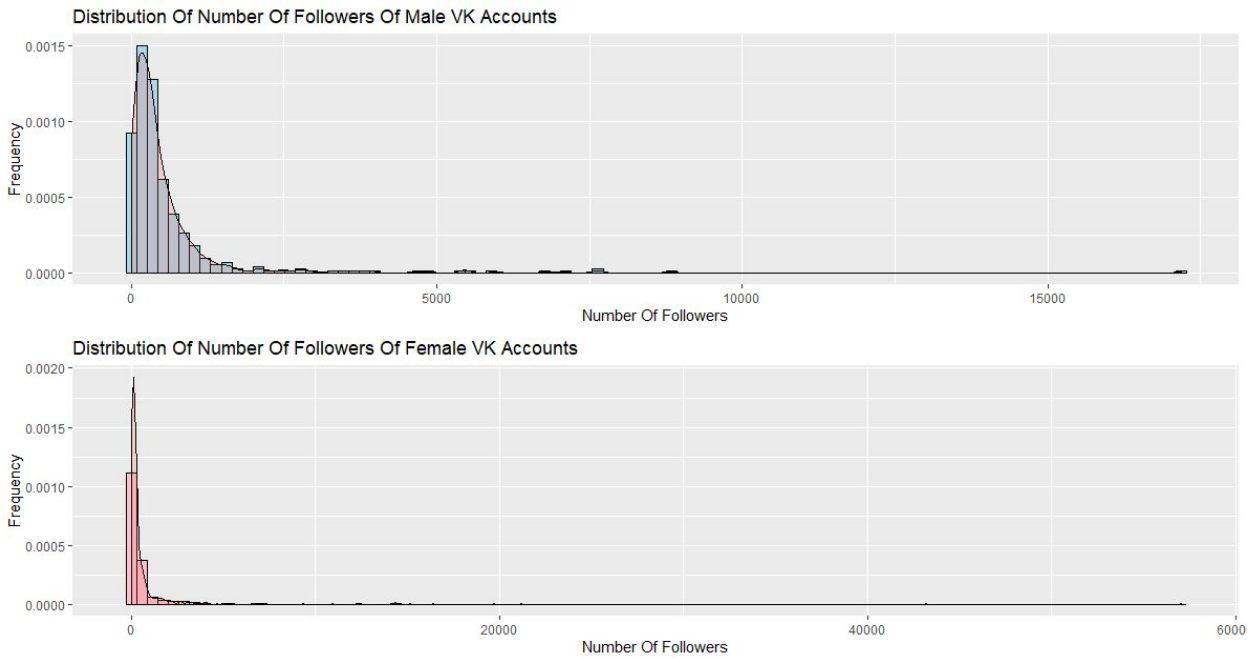
[7] https://russiansearchmarketing.com/10-key-statistics-social-media-usage-russia-2019/
[8] https://www.linkfluence.com/blog/russian-social-media-landscape

Independent Means t-tests

2-Sample t-test For sex:
To check if there is a significant difference in follower count for accounts of different sexes, we first check the conditions required to perform a 2-sample t-test:

Independence Assumption: The number of followers of any one account, male or female, is independent of the followers of any other account, as all accounts independently decide who they want to follow. The only exception is where people make deals to follow only if the other person follows back, but since we cannot accomodate for those cases, we consider the independence assumption satisfied.

Nearly Normal Condition: We can plot the distributions of the follower count for both groups to ensure that this condition is satisfied:



From the distributions, it's clear that both groups are nearly normal.

Independent Groups Assumption: The number of male account followers has no bearing on how many female account followers there are; people of both genders create and follow accounts independently, without any effect on each other.

Since the conditions are satisfied, we can conduct a t-test. The outcome of the t-test is as follows:

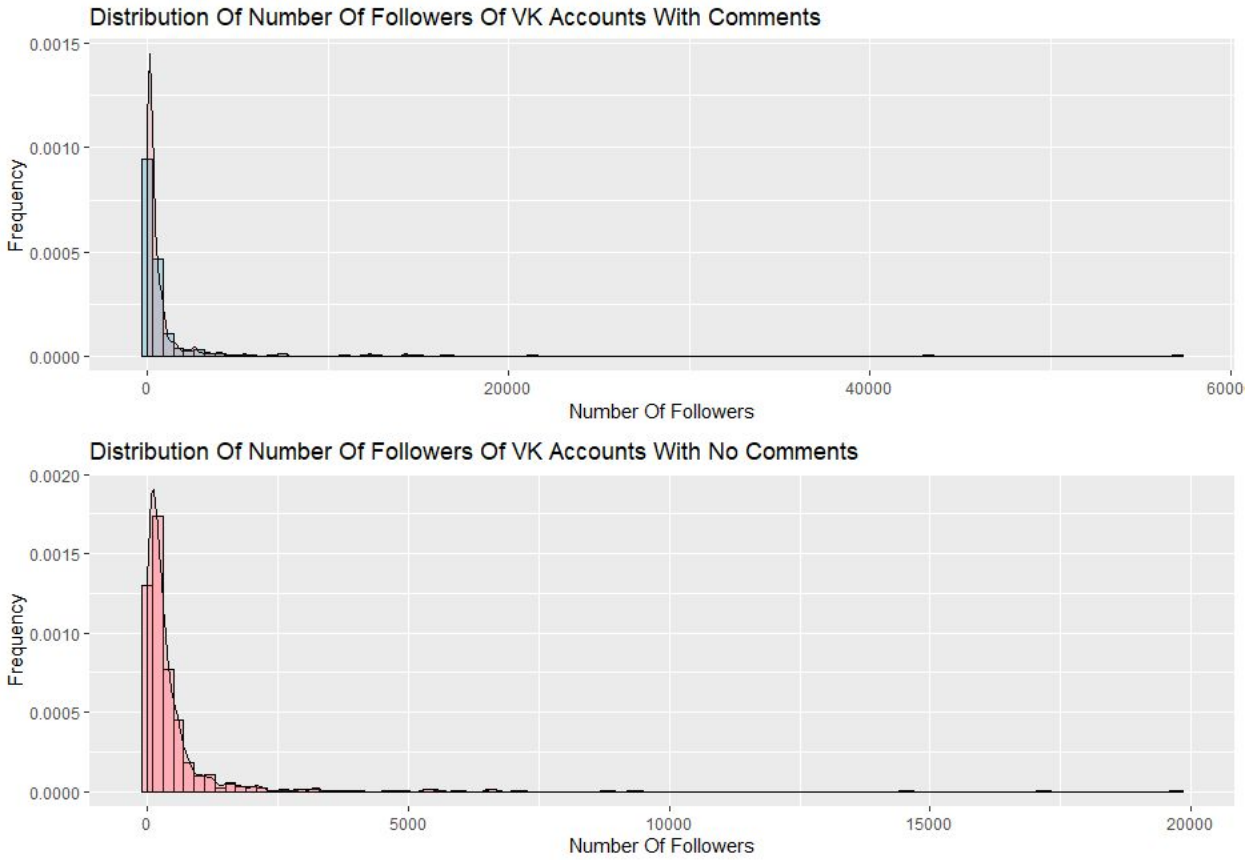
p-value = 0.1249 > 0.05; thus, p-value > alpha; thus, we fail to reject the null hypothesis. We cannot be sure that there is an advantage to one gender over the other for having a higher follower count.

2-Sample t-test For Comments:

To check if there is a significant difference in follower count for accounts where there is at least one comment in any post and those without any comments, we first check the conditions required to perform a 2-sample t-test:

Independence Assumption: The number of followers of any one account, where there are comments or not, is independent of the followers of any other account, as all accounts independently decide who they want to follow. Thus, we consider the independence assumption satisfied.

Nearly Normal Condition: We can plot the distributions of the follower count for both groups to ensure that this condition is satisfied:



From the distributions, it's clear that the distributions of both groups are nearly normal.

Independent Groups Assumption: The number of accounts and accounts with no comments in any of its posts have no bearing on how many accounts or those accounts where there is at least one comment in any post. They have no effect on each other.

Since the conditions are satisfied, we can conduct a t-test. The outcome of the t-test is as follows: p-value = 0.006216 < 0.05; thus, p-value < alpha; thus, we can safely reject the null hypothesis. We can be sure that if an account has at least one comment in any of its posts, it's more likely to have more followers than a similar account with any comments on its posts. More specifically, there is only a 0.6% chance that we got this sample by random chance, so we can be confident that having comments is helpful for a greater follower count.

Building the Linear Regression Model

Description	**Values**	**P-Value**
Intercept b0	-182.09802	0.112
Number of Friends coefficient b1	0.36715	<2*10^-16
Number of Likes on Photos coefficient b2	0.22890	<2*10^-16
Number of Posts Transformed(Squared) coefficient b3	0.8877	0.276

Multiple R-Squared: 23.51%

The p-value for the number of posts transformed = 0.276 > 0.05, intercept = 0.112 > 0.05; so p-value > alpha, so we fail to reject the null hypothesis that b3 = 0; thus, we cannot be sure that number of posts transformed has any actual impact on the number of followers, or that the intercept is not 0.

Still, the generated linear regression equation is:

*Predicted variable = b0 + Number of Friends*b1 + Number of Likes on Photos*b2 + ((Number of Posts)^2)*b3*

Number of Followers = -182.09802 + Number of Friends(0.36715) + Number of Likes on Photos*(0.22890) + (Number of Posts^2)*(0.8877)*

Interpreting The Regression

a. Interpreting The Intercept Of The Model In Context:

The intercept is the point on the y-axis corresponding to when all the other variables are 0 for the regression line. The intercept of the model here (b0) is the expected number of followers (y) predicted by the model for an account that does not have any friends, posts or likes on their photos; thus, this datapoint is illogical as number of followers cannot be negative.

b. Interpreting the Slopes/Coefficients Of The Model In Context:

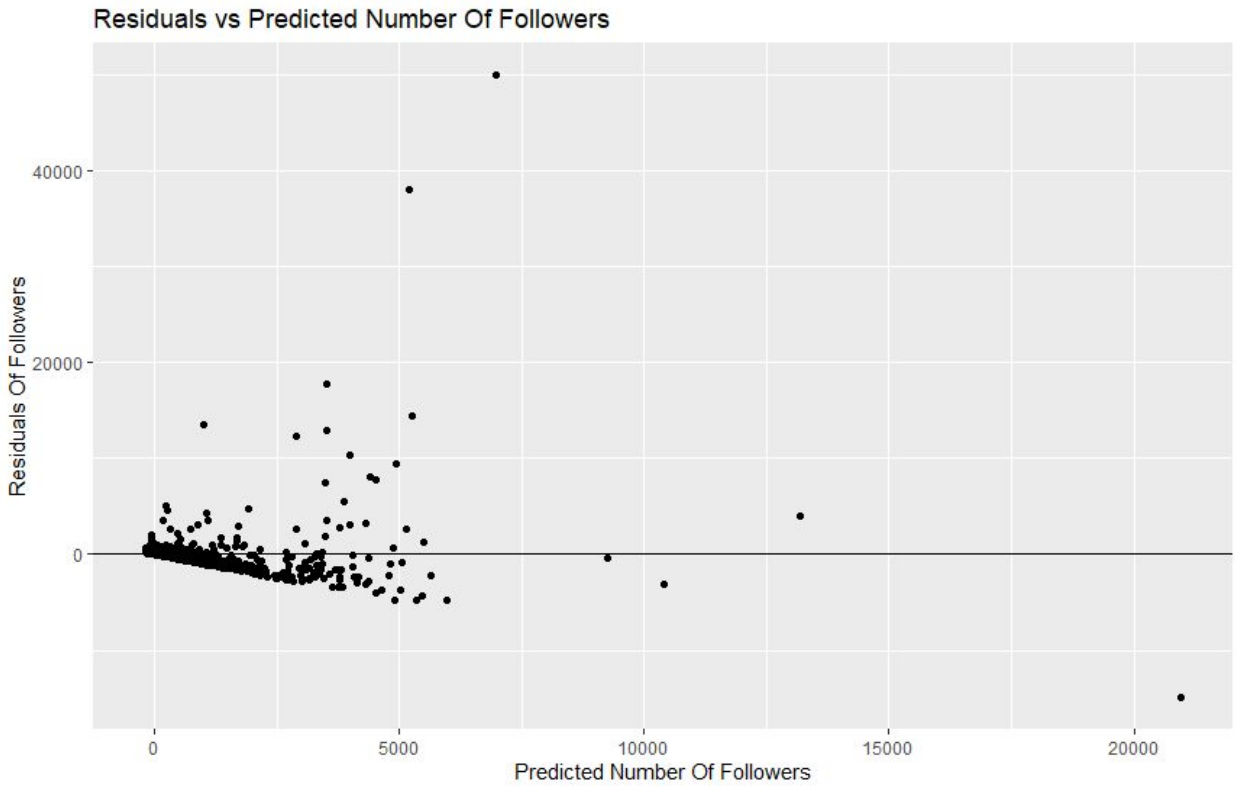
The slopes/coefficients of the model are measures of how much the number of followers (y-value) changes for unit changes in other predictor variables.

For each new friend added, the number of followers is expected to increase by 0.36715.
For each new post, and then the count of the number of posts is squared, the number of followers is expected to increase by 0.8877. Thus, it's clear that having more posts has the greatest impact for increasing a follower count as having more posts has by far the strongest increase in the follower count (it increases the follower count by more than 0.88 per post, as it's the number of posts squared, so the actual increase in followers for each new post is very high).
For each new like on a photo, the number of followers is expected to increase by 0.22890. Although the effect for each new like is not very high, likes and friends are more easily attainable than creating new posts frequently.

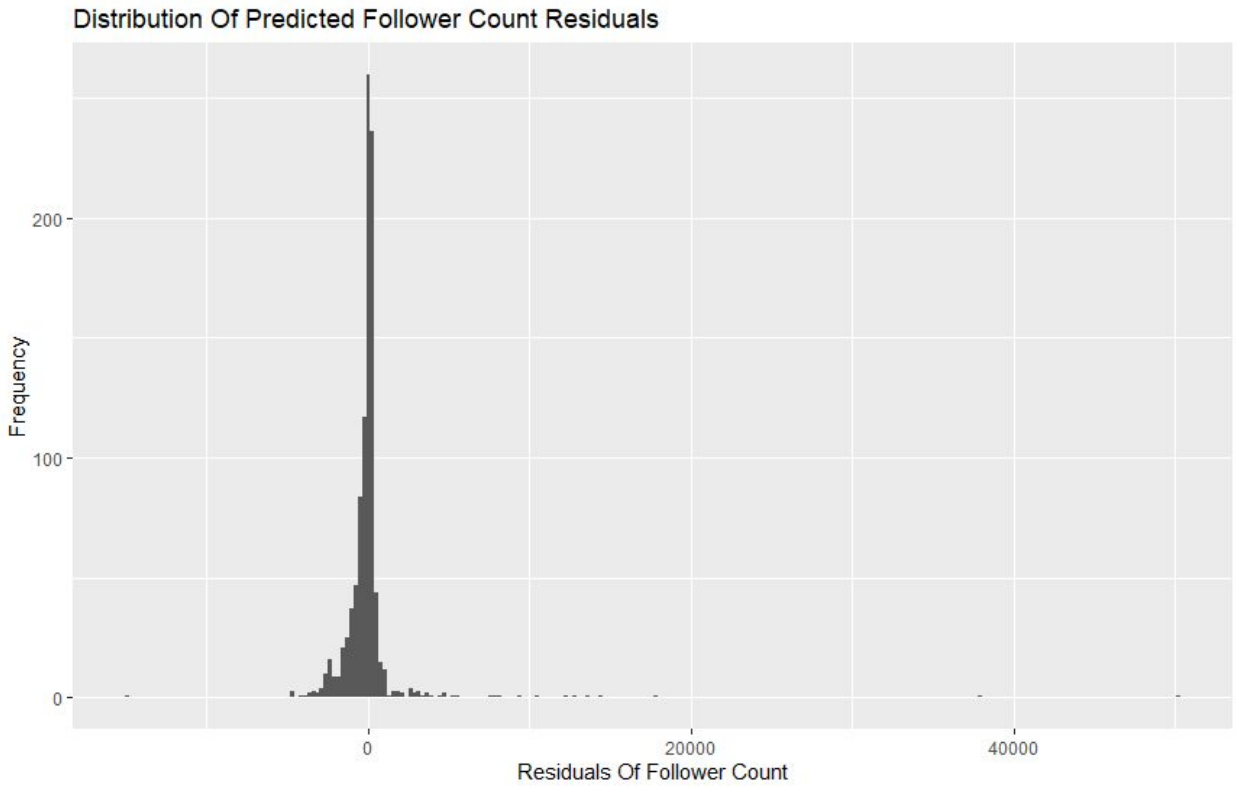
Thus, for attaining new followers, 2 likes are roughly worth 60% as much as adding one new friend, but making new (engaging) posts is easily the most effective way to get more followers.

Regression Diagnostics

We can plot the residual plot of the given regression model as follows:

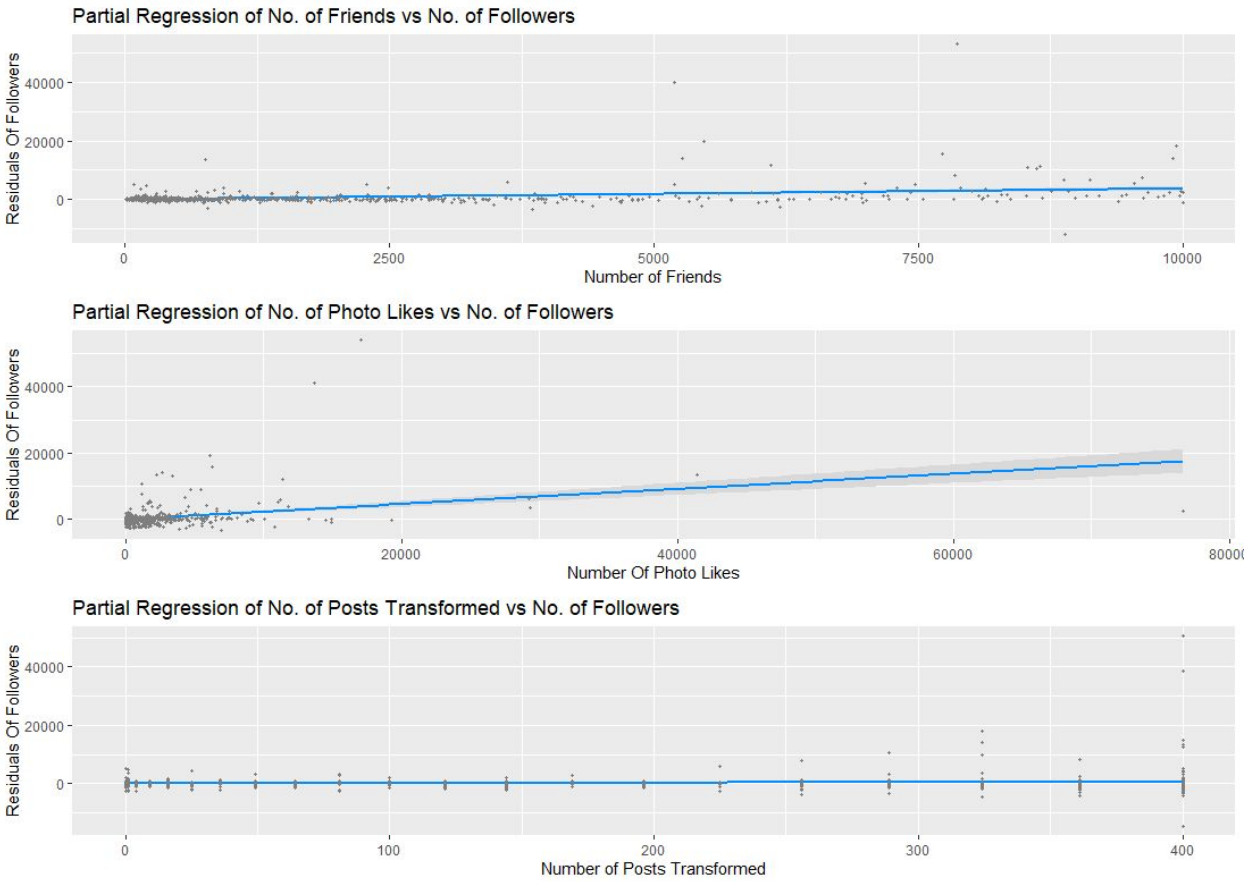


- The points are quite scattered, but are mostly distributed close to y = 0, i.e, the x-axis. This shows that the magnitude of most residuals is not very high; however, there are other glaring issues with the residual plot.
- There is a very significant and obvious pattern of increasing magnitude in the spread of the residuals as we move from left to right in the plot. The *Does The Plot Thicken?* Is very clearly not satisfied, and the residuals are not randomly distributed.
- There is a slight pattern that there is a bend that is apparent where the points are most clustered. Thus, the linearity assumption is not satisfied. We can affirm this by plotting the distribution of the residuals:



Although it seems that the distribution looks normal, this is not because the residuals are random; we can see from the residual scatterplot that there is a clear pattern to the residuals. Thus, the linearity assumption is not satisfied.

We can further analyse the specific partial regression plots to see if any of the specific partial plots don't have a linear relationship, which could be causing the pattern in the residuals:



From the partial regression plots, we can check for all the conditions to ensure that a straight line fits well for the data:

1. Quantitative Variable Condition - fulfilled: number of followers, number of posts, number of friends and number of photo likes are all quantitative.
2. Outlier Condition - There are a few notable outliers in the first partial regression (5000 friends and 40,000 followers, ~7700 friends and 50,000 followers), but none of them affect the regression line significantly. In the second plot, it's clear that the regression is most effective in the x range of 0 to 10,000 photo likes; there are a few outliers that are noteworthy (15,000 likes and 40,000+ followers), but they too don't alter the line significantly. In the final regression plot, there are notable outliers for when x = 20, i.e, for people with 20 posts; however, as we can see from the very weak slope of the regression line, they don't seem to alter it too much. The outlier condition is thus satisfied.
3. Straight Enough Condition – For the first plotted partial regression scatterplot between number of friends and number of followers, yes, the relationship between the datapoints is very clearly straight enough and fits a straight line model. For the second partial plot between followers and photo likes, although the relationship isn't as clearly linear as friends,

it still is suitable for a linear regression. However, the third partial regression scatterplot (between number of posts and followers) does not seem to fit a linear model very well.

Conclusion

From our hypotheses tests, we can conclude that we cannot be sure that either gender has any advantage in follower count, but we can be sure that accounts with comments are more likely to have more followers than accounts without any comments.

From the diagnostics of regression, we can conclude that linear multivariate regression is not the best model application for this data set because:

a) The residual plot fails the *Does The Plot Thicken?* condition,
b) The partial regression plot for number of posts transformed fails the linearity assumption,
c) The p-value for number of posts transformed is 0.276 and intercept is 0.112, which is greater than alpha, so we cannot be sure it has an impact on the number of followers, or that the intercept is not at 0,
d) The R-squared is only 23.51%, which is not very high, so the regression model's predictions won't be very accurate.

Moreover, as we have already discussed, numerous other models have been found to be more suitable for predicting follower counts in social media, and so this paper recommends that different models be tested on this dataset to create a better predictor function for follower count.

Works Cited

A) News Article Citations

1. News article on the cost of social media fraud, particularly exploiting follower counts: https://www.cbsnews.com/news/influencer-marketing-fraud-costs-companies-1-3-billion/
2. Another news article on the true effect of social media bots and the importance of follower counts in modern marketing: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html
3. Article evaluating numerous statistics on social media usage in Russia: https://russiansearchmarketing.com/10-key-statistics-social-media-usage-russia-2019/
4. Article with general information about Russian social media use and social media marketing in Russia: https://www.linkfluence.com/blog/russian-social-media-landscape

B) Research Citations

5. Mueller, Juergen, and Gerd Stumme. "Predicting Rising Follower Counts on Twitter Using Profile Information." *Proceedings of the 2017 ACM on Web Science Conference - WebSci 17*, May 9, 2017. https://doi.org/10.1145/3091478.3091490.

6. Hutto, C.j., Sarita Yardi, and Eric Gilbert. “A Longitudinal Study of Follow Predictors on Twitter.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 13*, 2013. https://doi.org/10.1145/2470654.2470771.
7. Wang, Ke, Mohit Bansal, and Jan-Michael Frahm. “Retweet Wars: Tweet Popularity Prediction via Dynamic Multimodal Regression.” *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. https://doi.org/10.1109/wacv.2018.00204.

Appendix

In order to ensure that these results are reproducible, included below is the R Code used to analyse the dataset and generate the visualisations in this report:

```
#cleaning the dataset, adding required columns
vk[vk$sex == 2, ]$sex <- 0
vk$hasComments <- 0
vk[vk$boolComments == "True", ]$hasComments <- 1
vk[vk$boolComments == "False", ]$hasComments <- 0

#applying transformation
vk$countOwnerPostsTransformed <- vk$countOwnerPosts^2

#response variable distributions
hist_followers <- ggplot(vk, aes(x = vk$countFollowers, y=
..density..))+ geom_histogram(bins = 150, color="black",
fill="green")+
  geom_density(alpha=.2, fill="#FF6666")
hist_followers<-hist_followers +
geom_vline(aes(xintercept=median(vk$countFollowers)),
           color="blue",linetype="dashed", size=1)
hist_followers<-hist_followers +
geom_vline(aes(xintercept=mean(vk$countFollowers)),

color="red",linetype="dashed", size=1)
hist_followers<-hist_followers + xlab("Number Of Followers")+
ylab("Frequency")
hist_followers <- hist_followers + ggtitle("Distribution Of
Number Of Followers Of VK Accounts")
hist_followers

boxplot_length <-boxplot(vk$countFollowers,
                         main = "Distribution Of Number Of
Followers",
                         names = "Number Of Followers",
                         col = "green",
```

```
                          border = "black",
                          xlab="Number Of Followers",
                          horizontal = TRUE,
                          outcex = 0.6
)

#predictor variable distributions
hist_friends <- ggplot(vk, aes(x = vk$countFriends, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_friends<-hist_friends + xlab("Number Of Friends")+
ylab("Frequency")
hist_friends <- hist_friends + ggtitle("Distribution Of Number
Of Friends Of VK Accounts")
hist_friends

hist_OwnerPosts <- ggplot(vk, aes(x = vk$countOwnerPosts, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_OwnerPosts<-hist_OwnerPosts + xlab("Number Of Owner
Posts")+ ylab("Frequency")
hist_OwnerPosts <- hist_OwnerPosts + ggtitle("Distribution Of
Number Of Owner Posts Of VK Accounts")
hist_OwnerPosts

hist_OwnerPostsTransformed <- ggplot(vk, aes(x =
vk$countOwnerPostsTransformed, y= ..density..))+
geom_histogram(bins = 21, color="black", fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_OwnerPostsTransformed<-hist_OwnerPostsTransformed +
xlab("Number Of Owner Posts Transformed")+ ylab("Frequency")
hist_OwnerPostsTransformed <- hist_OwnerPostsTransformed +
ggtitle("Distribution Of Number Of Owner Posts Transformed Of VK
Accounts")
hist_OwnerPostsTransformed

hist_OwnerReposts <- ggplot(vk, aes(x = vk$countOwnerReposts, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_OwnerReposts<-hist_OwnerReposts + xlab("Number Of Owner
Reposts")+ ylab("Frequency")
```

```
hist_OwnerReposts <- hist_OwnerReposts + ggtitle("Distribution
Of Number Of Owner Reposts Of VK Accounts")
hist_OwnerReposts

hist_Photos <- ggplot(vk, aes(x = vk$countPhotos, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_Photos<-hist_Photos + xlab("Number Of Owner Photos")+
ylab("Frequency")
hist_Photos <- hist_Photos + ggtitle("Distribution Of Number Of
Owner Photos Of VK Accounts")
hist_Photos

hist_Videos <- ggplot(vk, aes(x = vk$countVideos, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_Videos<-hist_Videos + xlab("Number Of Owner Videos")+
ylab("Frequency")
hist_Videos<- hist_Videos + ggtitle("Distribution Of Number Of
Owner Videos Of VK Accounts")
hist_Videos

hist_LikesPhotoes <- ggplot(vk, aes(x = vk$countLikesPhotoes, y=
..density..))+ geom_histogram(bins = 21, color="black",
fill="grey")+
  geom_density(alpha=.2, fill="#FF6666")
hist_LikesPhotoes<-hist_LikesPhotoes + xlab("Number Of Owner
Photo Likes")+ ylab("Frequency")
hist_LikesPhotoes <- hist_LikesPhotoes + ggtitle("Distribution
Of Number Of Photo Likes Of VK Accounts")
hist_LikesPhotoes

grid.arrange(hist_friends,hist_LikesPhotoes,hist_OwnerPosts,
hist_OwnerPostsTransformed,
hist_OwnerReposts,hist_Photos,hist_Videos)

#corr plot
vk_numeric <- subset(vk, select=-c(ID,boolComments))
corrs <- cor(data.frame(vk_numeric$countFollowers, vk_numeric),
use="complete.obs")
corrs <- subset(corrs, select=-c(vk_numeric.countFollowers))
corrs <- tail(corrs,-1)
```

```
corrplot <- ggcorrplot(corrs, lab = TRUE) +ggtitle("Correlation
Table")
corrplot

#regression
lmodel_vk <- lm(countFollowers ~ countFriends +
countLikesPhotoes + countOwnerPostsTransformed, data=vk)
summary(lmodel_vk)

# Creates the residual plot of this data
predicted<-predict(lmodel_vk)
resids<-residuals(lmodel_vk)
lmodel_vk.data<-data.frame(predicted,resids)

lmodel_vk_resid <- ggplot(lmodel_vk.data,
aes(x=predicted,y=resids)) +
  geom_point() +
  xlab("Predicted Number Of Followers") +
  ylab("Residuals Of Followers") +
  ggtitle("Residuals vs Predicted Number Of Followers") +
  geom_hline(yintercept = 0, color = "black")
lmodel_vk_resid

#distribution of the residuals
hist_resids <- ggplot(lmodel_vk.data,
aes(x=resids))+geom_histogram(bins = 250)+xlab("Residuals Of
Follower Count") +ylab("Frequency") + ggtitle("Distribution Of
Predicted Follower Count Residuals")
hist_resids


#partial regression plots
partial_friends <- visreg(lmodel_vk, "countFriends",
gg=TRUE)+xlab("Number of Friends") +ylab("Residuals Of
Followers")  +ggtitle("Partial Regression of No. of Friends vs
No. of Followers")
partial_LikesPhotoes <- visreg(lmodel_vk, "countLikesPhotoes",
gg=TRUE)+xlab("Number Of Photo Likes") +ylab("Residuals Of
Followers")  +ggtitle("Partial Regression of No. of Photo Likes
vs No. of Followers")
partial_OwnerPostsTransformed <- visreg(lmodel_vk,
"countOwnerPostsTransformed", gg=TRUE)+xlab("Number of Posts
Transformed") +ylab("Residuals Of Followers")  +ggtitle("Partial
Regression of No. of Posts Transformed vs No. of Followers")
```

```
grid.arrange(partial_friends, partial_LikesPhotoes,
partial_OwnerPostsTransformed)


#subsetting into male and female for t-test
male <- subset(vk, vk$sex == 1)
female <- subset(vk, vk$sex == 0)

#distributions of male and female for nearly normal condition
hist_male_followers <- ggplot(male, aes(x = male$countFollowers,
y= ..density..))+ geom_histogram(bins = 100, color="black",
fill="lightblue")+
  geom_density(alpha=.2, fill="#FF6666")
hist_male_followers<-hist_male_followers + xlab("Number Of
Followers")+ ylab("Frequency")
hist_male_followers<-hist_male_followers+ ggtitle("Distribution
Of Number Of Followers Of Male VK Accounts")
hist_male_followers

hist_female_followers <- ggplot(female, aes(x =
female$countFollowers, y= ..density..))+ geom_histogram(bins =
100, color="black", fill="pink")+
  geom_density(alpha=.2, fill="#FF6666")
hist_female_followers<-hist_female_followers + xlab("Number Of
Followers")+ ylab("Frequency")
hist_female_followers<-hist_female_followers+
ggtitle("Distribution Of Number Of Followers Of Female VK
Accounts")
hist_female_followers

grid.arrange(hist_male_followers, hist_female_followers)

#conduct the t-test
t.test(male$countFollowers, female$countFollowers, alternative =
"two.sided", var.equal = FALSE)

#subsetting into comments and no comments
comments <- subset(vk, vk$hasComments == 1)
nocomments <- subset(vk, vk$hasComments == 0)

#distributions of comments and no comments for nearly normal
condition
hist_comments_followers <- ggplot(comments, aes(x =
comments$countFollowers, y= ..density..))+ geom_histogram(bins =
100, color="black", fill="lightblue")+
```

```
  geom_density(alpha=.2, fill="#FF6666")
hist_comments_followers<-hist_comments_followers + xlab("Number
Of Followers")+ ylab("Frequency")
hist_comments_followers<-hist_comments_followers +
ggtitle("Distribution Of Number Of Followers Of VK Accounts With
Comments")
hist_comments_followers

hist_nocomments_followers <- ggplot(nocomments, aes(x =
nocomments$countFollowers, y= ..density..))+ geom_histogram(bins
= 100, color="black", fill="pink")+
  geom_density(alpha=.2, fill="#FF6666")
hist_nocomments_followers<-hist_nocomments_followers +
xlab("Number Of Followers")+ ylab("Frequency")
hist_nocomments_followers<-hist_nocomments_followers +
ggtitle("Distribution Of Number Of Followers Of VK Accounts With
No Comments")
hist_nocomments_followers

grid.arrange(hist_comments_followers, hist_nocomments_followers)

#conduct the t-test
t.test(comments$countFollowers, nocomments$countFollowers,
alternative = "greater", var.equal = FALSE)
```