

Statistics 2120 Project Report

Henry Rodriguez ear5rj

Joshua Yim jyy2ryk

Joey Elsis jre3wjh

“On our honor, we did not give nor receive aid on this assignment.”

Introduction

For this project, our group set out with the goal of determining the best predictor for the number of likes on an Instagram post. We started by tackling the planning and discussion section of the assignment. The first task was to list out a few potential variables that each group member thought would be a good explanatory variable for the number of likes a single-picture Instagram post receives, additionally considering whether each would have a positive or negative relationship with the response variable. Together, we came up with a list that included the time at which an Instagram post was uploaded, number of followers, number of comments, number of hashtags, number of posts already existing, and number of individuals the Instagram account follows. When our group reviewed each other's lists and attempted to complete the second task, some of the potential variables were eliminated due to complications with measurement methods. After our discussion, our group decided on our three explanatory variables that we would use for the remainder of the project: number of followers, number of comments, and number of hashtags.

Our first explanatory variable was the number of followers that the Instagram account for a university in question has. This would be measured by clicking on the profile of the university

in question's account and simply reading/recording the number of followers displayed. If an account has more followers, the account's posts will come up in more users' timelines and therefore has a better chance of getting more likes. Our second explanatory variable was the number of hashtags that the particular post being studied by a university in question's Instagram account uses in the description of the picture. The number of hashtags in each of the five most recent single-picture posts on a university's Instagram account would be measured, all added up together, then divided by five to calculate the average number of hashtags for each account. When a hashtag is searched by a user, it displays all posts that have the hashtag within their descriptions. The more hashtags used, the more feeds that the post in question will be included in, increasing the amount of people who see the post. The third explanatory variable was the number of comments that the post of interest by a university's Instagram account receives. The number of comments in each of the five most recent single-picture posts on a university's Instagram account would be measured, all added up together, then divided by five to calculate the average number of comments for each account. A post with more comments might pique the interest of other users on the platform since these comments imply that it is more interesting or controversial than average, and more people looking at a post would lead to more people liking it.

For our fourth and final task for Milestone 1, we stated that the defined response variable for this study was the number of likes that a single-picture Instagram post in question receives. To maintain consistency in our measurements of variables, however, our "personalized" response variable was the average number of likes out of the five most recent single-picture posts on a university's Instagram account receive. This measurement was taken for each

college/university Instagram account by clicking on the five most recent posts of the respective page, recording each count, then adding all counts up and dividing by five to get an average number of likes for each Instagram account.

Data Collection

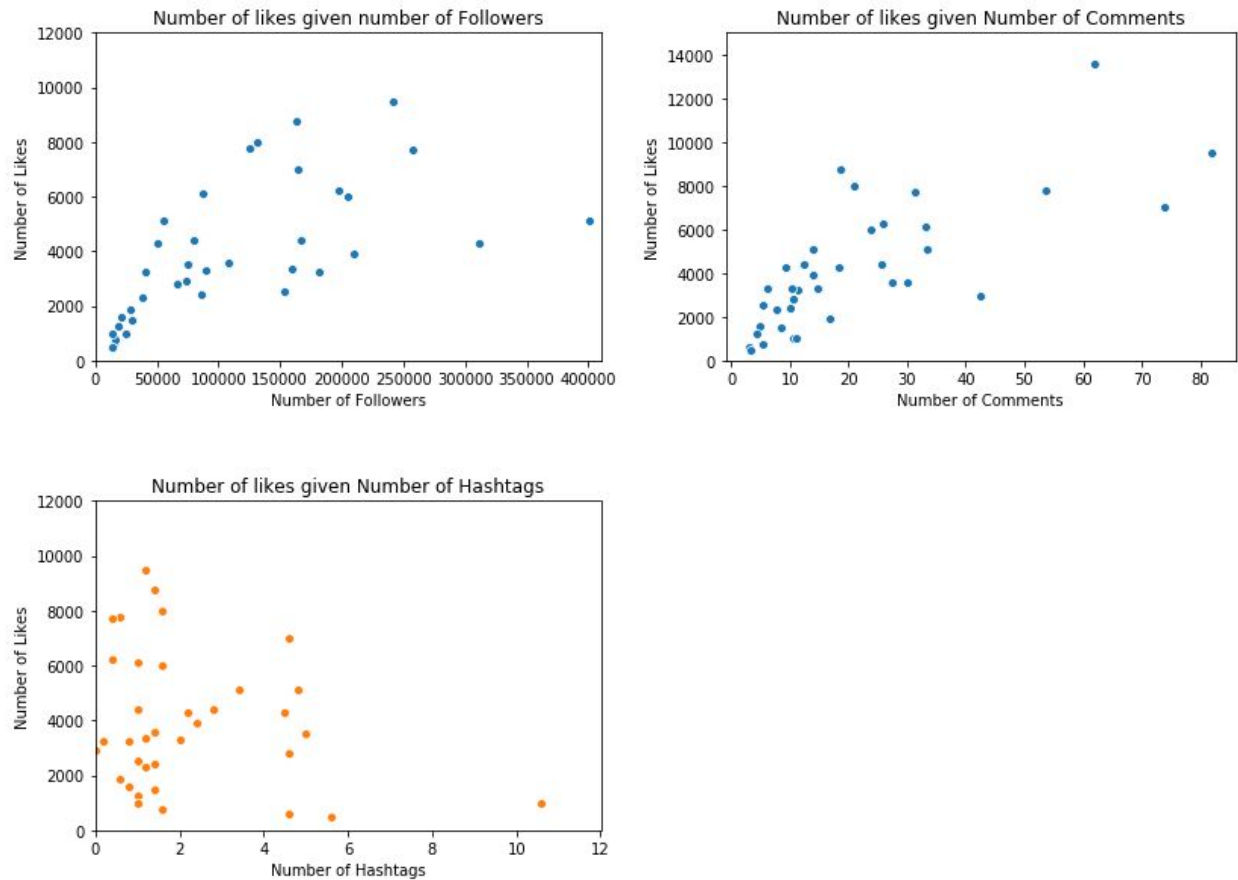
Our project group decided to include all universities within the *universities.csv* file within our measurements. Since the regression formula for degrees of freedom is $n-p-1$, using a larger sample size n would give us more degrees of freedom and therefore a distribution closer to Normal. As stated before, the variables of Number of Hashtags, Number of Comments, and Number of Likes were recorded as averages of the counts on the five most recent posts of each Instagram account. The Number of Followers was simply recorded for each Instagram account. We created a Google Sheet for the data collection process with columns for the University names, Number of Followers, Number of Hashtags, Number of Comments, and Number of Likes. Our group members equally split up the data collection, each recording information for a third of the universities listed. Amherst College can be used as an example for how our data collection looked:

	A	B	C	D	E	F
1	University		# of Followers	# of Hashtags	# of Comments	# of Likes
2	Amherst College		15000.00	4.60	3.00	613.40

After all data was collected for each of the universities, our group downloaded the Google Sheet into a .csv file, and received our signature for Milestone 2.

Analysis with simple linear regression:

Our group started the first task for Milestone 3 by creating scatterplots for each explanatory variable:



We concluded through these scatterplots that while the explanatory variables Number of Followers and Number of Comments had linear relationships with the response variable (number of likes), the variable Number of hashtags seemingly did not have any relationship with the number of likes.

For additional analysis, we looked at the coefficients of determination for each plot to assess the strength of the relationships. For Followers, our R^2 value is 0.0416; for Hashtags, it

was 0.0226; for Comments, it was 0.0335. These R^2 values are low and contradict the conclusions made from our visual exploration of the data. Although Followers clearly had the highest R^2 value, the three values were all too low to draw strong conclusions about their ability to predict the response.

Python Code for Simple Linear Regression Equation and R^2 :

```

9 X = data["Followers"]
10 X = sm.add_constant(X)
11 y = data["Likes"]
12
13 # Apply the regression equation
14
15 model = sm.OLS(y, X).fit()
16 r2_2 = model.rsquared
17 print("R^2: ", round(r2_2, 4))
18
19 # Determine the regression equation
20 b0 = round(model.params[0], 2)
21 b1 = round(model.params[1], 2)
22
23 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
24
25 sns.regplot(x="Followers", y="Likes", data=data, ci=0)
26 plt.title("Number of likes given number of Followers")
27 plt.xlabel("Number of Followers")
28 plt.ylabel("Number of Likes")
29
30 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
31
32 sns.regplot(x="Followers", y="Likes", data=data, ci=0)
33 plt.title("Number of likes given number of Followers")
34 plt.xlabel("Number of Followers")
35 plt.ylabel("Number of Likes")
36 plt.show()

```

IPython console

```

...: plt.y
...: plt.y
...: plt.s
R^2: 0.0416

```

```

9 X = data["Hashtags"]
10 X = sm.add_constant(X)
11 y = data["Likes"]
12
13 # Apply the regression equation
14
15 model = sm.OLS(y, X).fit()
16 r2_2 = model.rsquared
17 print("R^2: ", round(r2_2, 4))
18
19 # Determine the regression equation
20 b0 = round(model.params[0], 2)
21 b1 = round(model.params[1], 2)
22
23 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
24
25 sns.regplot(x="Hashtags", y="Likes", data=data, ci=0)
26 plt.title("Number of likes given number of Hashtags")
27 plt.xlabel("Number of Hashtags")
28 plt.ylabel("Number of Likes")
29
30 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
31
32 sns.regplot(x="Hashtags", y="Likes", data=data, ci=0)
33 plt.title("Number of likes given number of Hashtags")
34 plt.xlabel("Number of Hashtags")
35 plt.ylabel("Number of Likes")
36 plt.show()

```

IPython console

```

...: plt.y
...: plt.y
...: plt.s
R^2: 0.0226

```

```

9 X = data["Comments"]
10 X = sm.add_constant(X)
11 y = data["Likes"]
12
13 # Apply the regression equation
14
15 model = sm.OLS(y, X).fit()
16 r2_2 = model.rsquared
17 print("R^2: ", round(r2_2, 4))
18
19 # Determine the regression equation
20 b0 = round(model.params[0], 2)
21 b1 = round(model.params[1], 2)
22
23 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
24
25 sns.regplot(x="Comments", y="Likes", data=data, ci=0)
26 plt.title("Number of likes given number of Comments")
27 plt.xlabel("Number of Comments")
28 plt.ylabel("Number of Likes")
29
30 print("The regression equation is: y = " + str(b0) + " + " + str(b1) + "x.")
31
32 sns.regplot(x="Comments", y="Likes", data=data, ci=0)
33 plt.title("Number of likes given number of Comments")
34 plt.xlabel("Number of Comments")
35 plt.ylabel("Number of Likes")
36 plt.show()

```

IPython console

```

...: plt.y
...: plt.y
...: plt.s
R^2: 0.0335

```

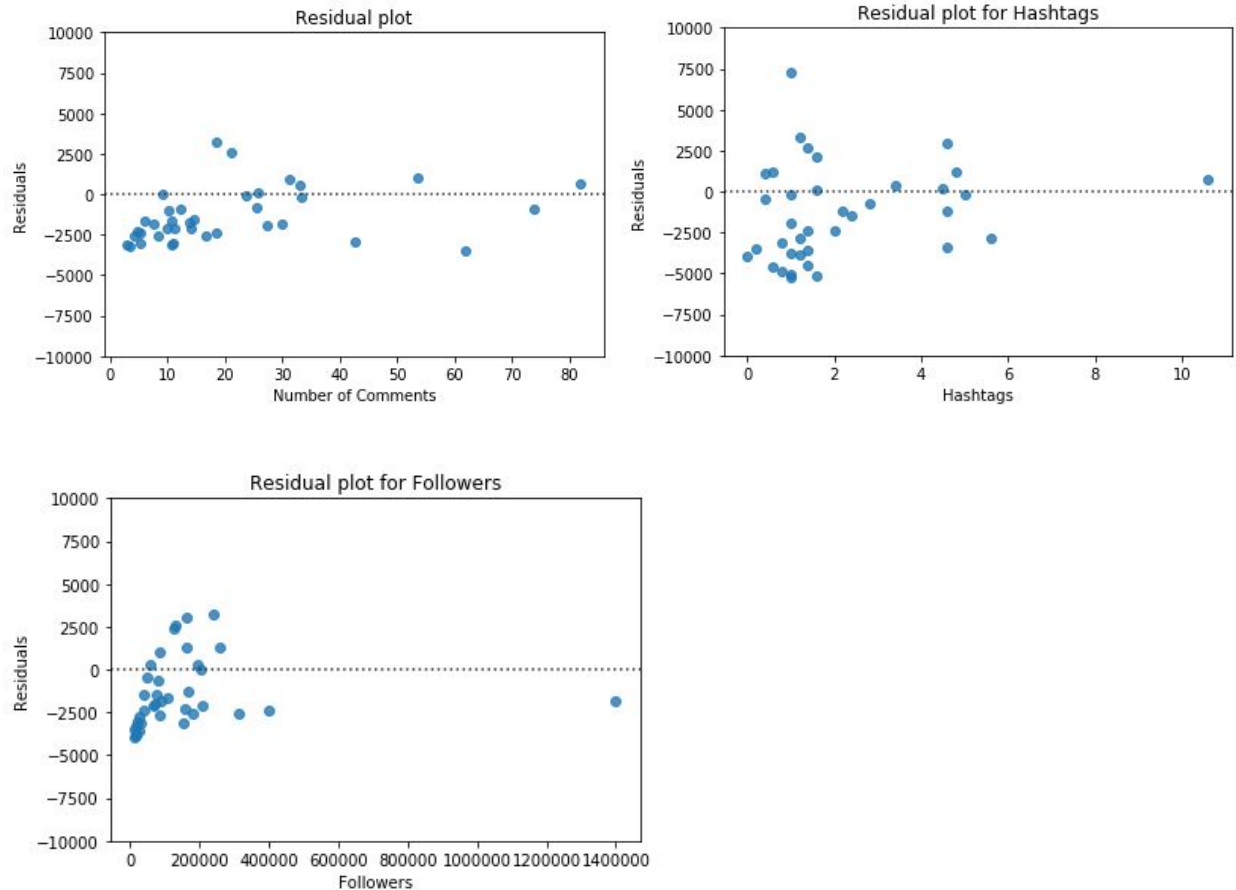
For the second task, our group calculated the simple linear regression model for each explanatory variable using Python:

The regression equation for Number of Followers was calculated as: $\hat{y} = 4384.88 + 0.01x$.

The regression equation for Number of Hashtags was calculated as: $\hat{y} = 6901.06 + -628.66x$.

The regression equation for Number of Comments was calculated as: $\hat{y} = 3764.29 + 83.56x$.

For the third task, we created a residual plot for each simple linear regression model:



Our group initially concluded that, disregarding outliers, the residual plots for Number of Comments and Number of Followers satisfied linear regression assumptions due to visible uniform variation around zero, linearity and independence. The linear regression assumption for

the SLR model of Number of Hashtags was not satisfied, as the residuals are mostly in two separate clusters around zero and therefore aren't uniformly distributed. However, when adding numerical information such as the coefficient of determination for each scatter plot exploring the relationship between each explanatory variable and likes, we found that the data failed to satisfy the linear assumption that visually we thought had been satisfied. These R^2 values were found in the first task of Milestone 3.

For the fourth task for Milestone 3, our group decided that the “best” explanatory variable was the Number of Followers. This was due to a visual uniform variation around zero in the residual plot (satisfying the linear regression assumption) and the greatest r-squared value out of all three explanatory variables.

The “best” explanatory variable, the Number of Followers:

```

=====
                        OLS Regression Results
=====
Dep. Variable:          Likes    R-squared:                0.042
Model:                  OLS      Adj. R-squared:           0.015
Method:                 Least Squares    F-statistic:             1.562
Date:                   Mon, 02 Dec 2019  Prob (F-statistic):      0.219
Time:                   21:15:23      Log-Likelihood:          -397.53
No. Observations:       38          AIC:                       799.1
Df Residuals:           36          BIC:                       802.3
Df Model:                1
Covariance Type:        nonrobust
=====
                        coef    std err          t      P>|t|      [0.025    0.975]
-----
const          4384.8780    1686.340      2.600      0.013     964.822    7804.934
Followers       0.0078      0.006      1.250      0.219     -0.005     0.021
=====
Omnibus:                 82.887    Durbin-Watson:           1.906
Prob(Omnibus):            0.000    Jarque-Bera (JB):        1530.449
Skew:                    5.416    Prob(JB):                 0.00
Kurtosis:                32.142    Cond. No.                 3.22e+05
=====

```


The least squares regression equations is: $y_i = 4384.88 + 0.01x$

- For every increase in one follower of an account, the projected amount of likes increases by 0.01. *As for the y-intercept, the number of likes a post would have given that the university account has no followers would be 4384.88. In context this value does not provide useful information.*

The p-value is 0.219 and the test statistic is 1.25. Based off this information, because our p-value of 0.219 is greater than our significance level of 0.05, we fail to reject the null hypothesis, and there is no statistically significant evidence for there being a relationship between the number of followers and the number of likes.

```
48 plt.ylim(-10000, 10000)
49
50 uva_y = 4584.88 + 0.01 * 95500
51 print("UVA's predicted number of likes: ", uva_y)
```

UVA's predicted number of likes: 5539.88

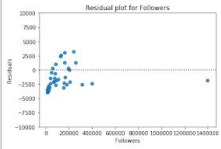


A residual plot for Followers. The x-axis is labeled 'Followers' and ranges from 0 to 140,000. The y-axis is labeled 'Residuals' and ranges from -10,000 to 10,000. The plot shows a dense cluster of points near the zero line for lower follower counts, with a few outliers at higher follower counts.

The predicted number of likes is 5540

```
9 # Store the residuals
10 m_resid = model.resid
11
12 # Create plot
13 sns.residplot(x=data.Followers, y=m_resid)
14 plt.title("Residual plot for Followers")
15 plt.ylabel("Residuals")
16 plt.xlabel("Followers")
17 plt.xlim(0, 12)
18 plt.ylim(-10000, 10000)
19
20 uva_y = 4584.88 + 0.01 * 95500
21 print("UVA's predicted number of likes: ", uva_y)
22
23 residual = 6412 - uva_y
24
25 print("residual is: ", residual)
```

residual is: 872.1199999999999



A residual plot for Followers. The x-axis is labeled 'Followers' and ranges from 0 to 140,000. The y-axis is labeled 'Residuals' and ranges from -10,000 to 10,000. The plot shows a dense cluster of points near the zero line for lower follower counts, with a few outliers at higher follower counts.

In [76]:

The residual is 872 likes.

Analysis with multiple linear regression

Our group project calculated the multiple linear regression model with all three explanatory variables through Python:

$$y = 4801.599 + (0.005)(\text{followers}) + (47.814)(\text{comments}) - (502.537)(\text{hashtags})$$


```

1 import matplotlib.pyplot as plt
2 import seaborn as sns
3 import pandas as pd
4 import statsmodels.api as sm
5
6
7 data = pd.read_csv(r"C:\Users\joeye\Downloads\STAT\university_data.csv", index_col = "University")
8
9 X = data[["Followers", "Comments", "Hashtags"]]
10 X = sm.add_constant(X)
11 y = data["Likes"]
12
13 # Apply the regression equation
14 model = sm.OLS(y, X).fit()
15
16 bj = round(model.params,3)
17 print(bj)

```

```

sns.residplot(x=data.Fol
plt.title("Residual plot
plt.ylabel("Residuals")
plt.xlabel("Followers")

```

IPython console

Console 1/A

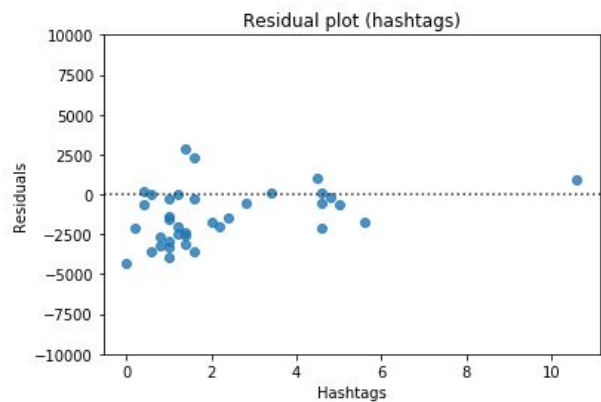
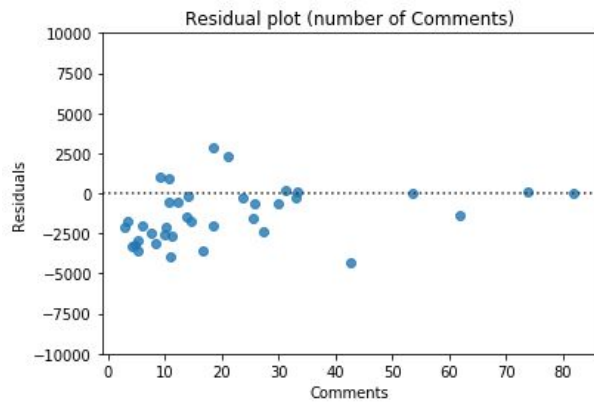
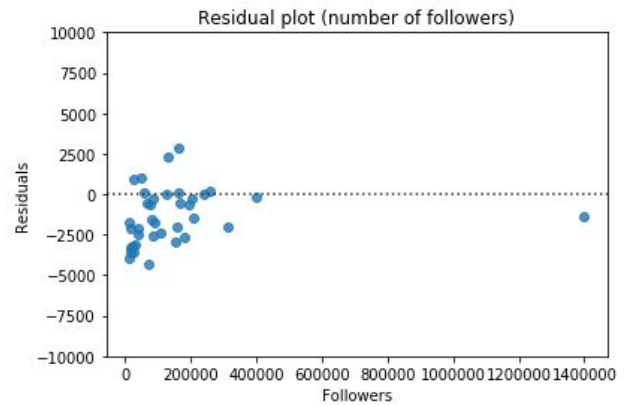
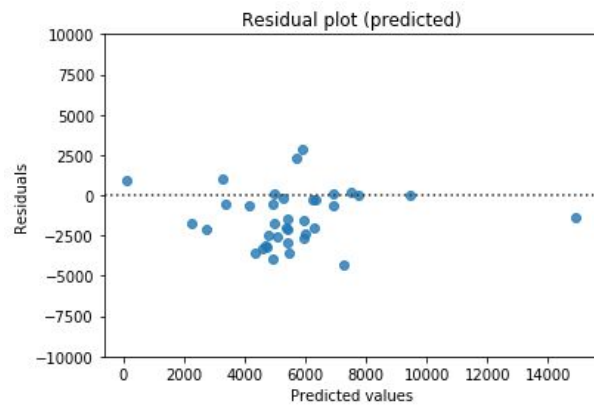
```

In [4]: runfile('C:/User
wdir='C:/Users/joey/Dow
const      4801.599
Followers   0.005
Comments    47.814
Hashtags   -502.537
dtype: float64

```

In [5]:

To start off task 2, we went ahead and created four residual plots: one for each explanatory variable, and one for predicted values.



In the case of the first residual plot for predicted values, each event is independent and there is no overall shape to the graph. Based on these residual plots, we found that the linear regression assumptions were satisfied for the variables of Number of Followers and Numbers of Comments due to a visual uniform variation around zero, but for predicted values and Number of Hashtags, the linear regression assumptions were not satisfied. The residual plots for predicted values and Number of Hashtags failed to show uniform variation around zero, and therefore they do not satisfy the linear regression assumptions.

To complete task 3, our group used Python to code in and display the OLS Regression Results table.

```
#####
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
import numpy as np

universities = pd.read_csv(r"/Users/yechan/Documents/project.csv")

# Define the explanatory and response variables
X = universities[["Followers", "Comments", "Hashtags"]]
X = sm.add_constant(X)
y = universities["Likes "]

# Apply the regression equation
model = sm.OLS(y, X).fit()

bj = round(model.params,3)

# Store the residuals and predicted values
m_resid = model.resid
m_pred = model.fittedvalues
model_output = model.summary()
print(model_output)
```

	coef	std err	t	P> t	[0.025	0.975]
const	4801.5990	2774.830	1.730	0.093	-837.535	1.04e+04
Followers	0.0055	0.007	0.767	0.448	-0.009	0.020
Comments	47.8138	84.781	0.564	0.576	-124.482	220.109
Hashtags	-502.5371	700.857	-0.717	0.478	-1926.850	921.776

This calculated table displayed the necessary information needed to conduct the t-test for slope for every explanatory variable:

Followers: from a test statistic of 0.767, we calculated a p-value of 0.448. Because the p-value is greater than the alpha value of 0.05, we fail to reject the null hypothesis, meaning that the number of followers a university account has is not useful in predicting the number of likes on a single picture Instagram post. There is no statistically significant evidence to support that the number of followers an account has is a useful predictor of the amount of likes that are received on a single picture post.

Comments: from a test statistic of 0.564, we calculated a p-value of 0.576. Because the p-value is greater than the alpha value of 0.05, we fail to reject the null hypothesis, meaning that the number of comments a single picture Instagram post has is not useful in predicting the number of likes on that post. There is no statistically significant evidence to support that the number of comments a post has is a useful predictor of the amount of likes that are received on the post.

Hashtags: from a test statistic of -0.717, we calculated a p-value of 0.478. Because the p-value is greater than the alpha value of 0.05, we fail to reject the null hypothesis, meaning that the number of hashtags a single picture Instagram post has is not useful in predicting the number of likes on that post. There is no statistically significant evidence to support that the number of hashtags a post has is a useful predictor of the amount of likes that are received on that post.

To conduct the ANOVA F test, we derived our f-statistic and p-value from the calculated OLS Regression Results table.

F-statistic:	0.8044
Prob (F-statistic):	0.500
Log-Likelihood:	-397.03

From an f-statistic of 0.8044, we calculated a p-value of 0.5 and at a significance level of 0.05, we fail to reject the null that all of our coefficients equal 0, meaning that there is not enough statistically significant evidence to suggest that our model is useful.

Original:

```
In [111]: X_ext = data[["Followers", "Hashtags", "Comments"]]
...: X_ext = sm.add_constant(X_ext)
...: y = data["Likes"]
...:
...: # Apply the regression equation
...: model_ext = sm.OLS(y, X_ext).fit()
...:
...: test_stat = ( (n-2) / 1 )*( (r2_1-r2_2) / (1-r2_1) )
...: test_stat = round(test_stat, 2)
...: print(test_stat)
0.95
```

```
In [112]: pval = 1 - stats.f.cdf(test_stat, q, n-p-1)
...: pval = round(pval, 4)
...: print(pval)
0.3968
```

Extended:

```
In [108]: r2_2 = model.rsquared
...: r2_1 = model_ext.rsquared
...:
...: p = 3
...: q = 2
...: n = model.nobs
...:
...: # Test statistic
...: test_stat = ( (n-p-1) / q ) * ( (r2_1-r2_2) / (1-r2_1) )
...: test_stat = round(test_stat, 2)
...: print(test_stat)
0.45

In [109]: pval = 1 - stats.f.cdf(test_stat, q, n-p-1)
...: pval = round(pval, 4)
...: print(pval)
0.6414
```

In task 5, comparing our f-statistics for Followers vs all three variables combined, we see that our model for followers, with an adjusted r^2 of 0.015, is a much better model than our multiple variable model with an adjusted r^2 of -0.016. For the original model, our test statistic of 0.95 and p-value of 0.3968 shows a stronger model than our extended model with test statistic 0.45 and p-value of 0.6414 when a significance level of 0.05 was used to determine statistical significance.

Conclusion

Our group set out to find the best predictor for the number of likes on a single instagram post. In order to make an informed decision, we chose to include all the universities within our study that were supplied to us through a list. This helped reduce the probability that our conclusion happened simply by chance. We first collected data on three factors we believed

would show promise: the account's number of followers, the number of hashtags within a post, and number of comments on a post. From these factors we built three models and compared them. From these three models, we found that the model for Number of Followers portrayed that it was the most useful predictor. We then performed additive tests to see if having more than one factor in our model would help improve our ability to predict likes, but found that looking at just Number of Followers would be the best method for prediction. Through our comparison of the models and statistical tests performed, we found that increasing the number of followers of an account would best predict an increase in likes. However, we would like to present our findings with a warning. Although we found the number of followers in a model to be the best indicator, all of our models showed a weak strength for predicting the number of likes. Our group recommends that the scope of this project be increased, and more factors should be explored to find better predictors.

Reflection:

We found in our results that the three explanatory variables chosen did not provide an effective means to predict our response variable. We believe that the low coefficient of determination in each variable was caused by a lack of a linear relationship or no relationship at all as well as several outliers in our data, as regression is not resistance to outliers. To improve the strength of these relationships, on a second attempt we would have removed these outliers from our data. Although we restricted our bounds for our visual analysis of the scatter plots, our underlying data still contained these outliers when conducting tests on our values. For example, a few internationally famous schools like Harvard have a disproportionately enormous number of

followers, despite getting comparatively fewer likes on each post. We would also consider increasing the scope of our data collection method, from the first 5 posts to upwards of 10 posts per account. Finally, to improve the usability of the data we collected in this project, we would in a second attempt have done a logarithmic transformation of the explanatory variables in order to promote a more linear relationship between the response variable.

Appendix A: Milestone checklist:

Project: Milestone checklist	
Milestone 1	
Initials: <u>ZZ</u>	Date: <u>11/20/19</u>
Milestone 2	
Initials: <u>[Signature]</u>	Date: <u>11/25/19</u>
Milestone 3	
Initials: <u>[Signature]</u>	Date: <u>12/2/19</u>
Milestone 4	
Date: <u>wed, 4, 2019</u>	
Signature: <u>[Signature]</u>	

Appendix B: Data:

	A	B	C	D	E	F
1	University		# of Followers	# of Hashtags	# of Comments	# of Likes
2	Amherst College		15000.00	4.60	3.00	613.40
3	Boston College		89200.00	2.00	14.60	3288.20
4	Bowdoin College		16200.00	1.60	5.40	773.20
5	Brown University		182000.00	0.80	11.20	3243.00
6	Carnegie Mellon University		30100.00	1.40	8.40	1499.80
7	Clemson University		197000.00	0.40	25.80	6241.40
8	Columbia University		210000.00	2.40	13.80	3931.40
9	Cornell University		205000.00	1.60	23.80	5990.40
10	Dartmouth College		55900.00	3.40	33.40	5121.60
11	Duke University		167000.00	2.80	12.40	4386.20
12	Florida State University		108000.00	1.40	27.40	3596.40
13	George Mason University		24400.00	10.60	10.60	997.00
14	George Washington University		40800.00	0.20	10.20	3266.40
15	Georgetown University		73300.00	0.00	42.60	2931.80
16	Georgia Tech University		66700.00	4.60	10.60	2791.00
17	Harvard University		1400000.00	1.00	62.00	13565.60
18	James Madison University		51100.00	4.50	9.20	4291.80
19	North Carolina State University		86200.00	1.40	10.00	2434.80
20	Old Dominion University		13400.00	5.60	3.40	501.80
21	Penn State University		160000.00	1.20	6.00	3328.40
22	Princeton University		312000.00	2.20	18.40	4287.40
23	Syracuse University		75200.00	5.00	30.00	3542.40
24	UNC Chapel Hill		108000.00	1.20	14.00	55205.20
25	University of California Berkeley		164000.00	4.60	73.80	6997.00
26	University of California Los Angeles		242000.00	1.20	82.00	9475.00
27	University of Florida		163000.00	1.40	18.60	8737.80
28	University of Georgia		125000.00	0.60	53.60	7751.00
29	University of Louisville		21000.00	0.80	4.80	1563.80
30	University of Miami		79800.00	1.00	25.60	4381.80
31	University of Michigan		258000.00	0.40	31.20	7732.60
32	University of Pennsylvania		154000.00	1.00	5.40	2510.00
33	University of Pittsburgh		18800.00	1.00	4.20	1257.80
34	University of Wisconsin		132000.00	1.60	21.00	7991.00
35	Virginia Tech		86900.00	1.00	33.00	6104.80
36	Wake Forest University		37800.00	1.20	7.60	2298.20
37	William & Mary		28400.00	0.60	16.80	1886.60
38	Williams College		14300.00	1.00	11.00	996.20
39	Yale University		401000.00	4.80	14.00	5114.80