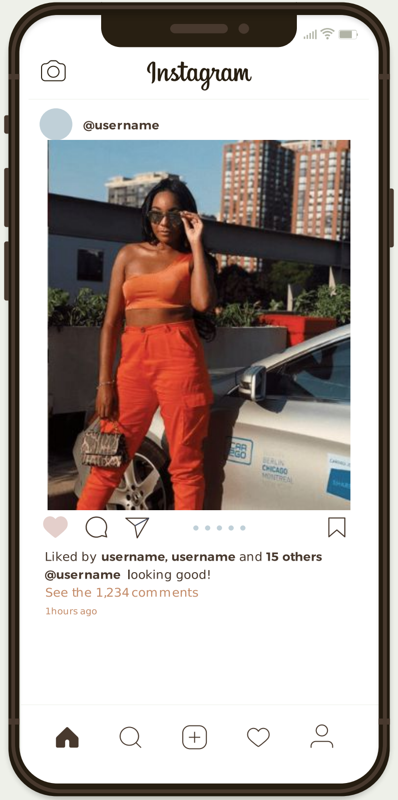
Instagram Data Analysis

By Haripriya and Taryn



Background on Data

★ We used data of 11692 instagram posts.

- We extracted relevant data, which would include the following:
  - Username
  - Following
  - Followers
  - Multiple_images
  - Is_video

	created_at	followers	following
count	3.744000e+03	3.744000e+03	3.744000e+03
mean	1.681608e+09	1.894394e+06	1.031337e+04
std	4.969690e+07	6.845989e+06	1.198092e+05
min	1.450283e+09	1.000000e+00	0.000000e+00
25%	1.685892e+09	1.971580e+05	4.200000e+02
50%	1.705337e+09	4.990260e+05	7.850000e+02
75%	1.708531e+09	1.117678e+06	1.296000e+03
max	1.709526e+09	7.180750e+07	1.568394e+06

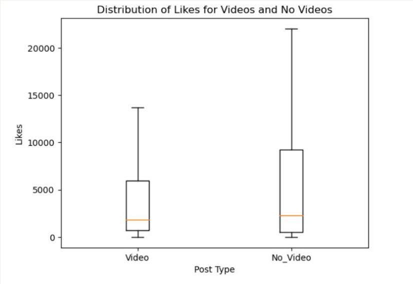
is_video	caption	comments	likes	created_at	location	imageUrl	multiple_images	username	followers	following
False	I’m a brunch & Iced Coffee girlie 🟤 🔍 \n\nTop @...	268	16382.0	1.709327e+09	NaN	https://instagram.flba2-1.fna.fbcdn.net/v/t39....	True	christendominique	2144626.0	1021.0
True	😤 Brow tips I really wish I would have know w...	138	9267.0	1.709241e+09	NaN	https://instagram.flba2-1.fna.fbcdn.net/v/t51....	False	christendominique	2144626.0	1021.0
True	OMG I can’t believe it’s already been 1 yr sin...	1089	10100.0	1.709155e+09	NaN	https://instagram.flba2-1.fna.fbcdn.net/v/t51....	False	christendominique	2144626.0	1021.0
True	90’s Glam was Pam! \n\nMakeup \n@smashboxcosme...	271	6943.0	1.709065e+09	NaN	https://instagram.flba2-1.fna.fbcdn.net/v/t51....	False	christendominique	2144626.0	1021.0
True	Chiseled & Sculptured 🏛 \n\n Contour @westmana...	145	17158.0	1.708718e+09	NaN	https://instagram.flba2-1.fna.fbcdn.net/v/t51....	False	christendominique	2144626.0	1021.0

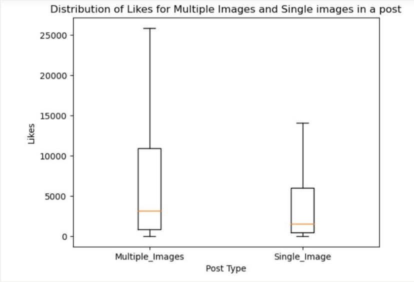
Goal: Clean the data by removing NA values and scaling it for better results

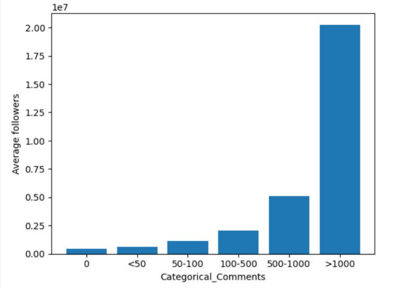
```
# Convert columns to numeric types
instagram_df['likes'] = pd.to_numeric(instagram_df['likes'], errors='coerce')
instagram_df['comments'] = pd.to_numeric(instagram_df['comments'], errors='coerce')
instagram_df['multiple_images'] = instagram_df['multiple_images'].astype(bool)
```

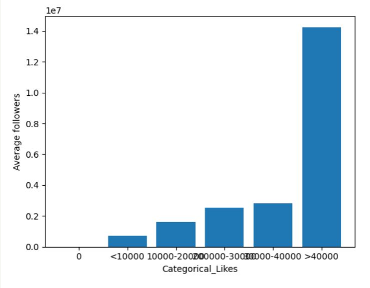
Data Preprocessing

Goal: Trying to visualize the data with various graphs to help identify possible trends









Data Visualization

Naive Bayes Process

STEPS

1. Create categorical columns of the quantitative data (ex: we encoded all the values greater than 100 comments into a category of 100-200)
    - Converting data accordingly to promote consistency
    - Taking into account values above our range
2. Making a column for the ratio between followers and following and another column with the influencer status depending on the ratio (1 following : 6000 followers)
3. Splitting data into test and training and having the Gaussian Naive Bayes model train with the training data
4. Gaussian model will predict the test data using the knowledge from training data
5. Specifying conditions
6. Calculating accuracy and the prediction of the model

Categorical_Likes	Categorical_Comments	Followers and Following Ratio	Influencer
10000-20000	100-500	2100.515181	False
<10000	100-500	2100.515181	False
10000-20000	>1000	2100.515181	False
<10000	100-500	2100.515181	False
10000-20000	100-500	2100.515181	False

Goal: Identifying how various components affect the possibility of an instagram user being considered an influencer.

Key Takeaways:

- ★ We have found that the number of likes and comments a person has influences their influencer status more than the components of their posts
- ★ With the specific conditions below, this is what our model predicts with 83.58% accuracy

```
Accuracy: 83.58%
Prediction for a post who has a video, multiple images, 10000 - 20000 likes, and more than 1000 comments: Influencer = No
Prediction for a post who has a video, multiple images, less than 10000, and more than 1000 comments: Influencer = Yes
```

Naive Bayes Results

Decision Tree Process

<u>STEPS</u>

1. Starting the same as Naive Bayes, Create categorical columns of the quantitative data (ex: we encoded all the values greater than 100 comments into a category of 100-200)
    - Converting data accordingly to promote consistency
    - Taking into account values above our range
2. Set thresholds for “Influencer” status
    - Followers to following: 6000:1
    - Likes per post: 10000
    - Comments per post: 500
    - Number of images: at least one
3. Calculate accuracy and precision, print accuracy report based on results

multiple_images	username	followers	following
True	christendominique	2144626.0	1021.0
False	christendominique	2144626.0	1021.0
False	christendominique	2144626.0	1021.0
False	christendominique	2144626.0	1021.0
False	christendominique	2144626.0	1021.0

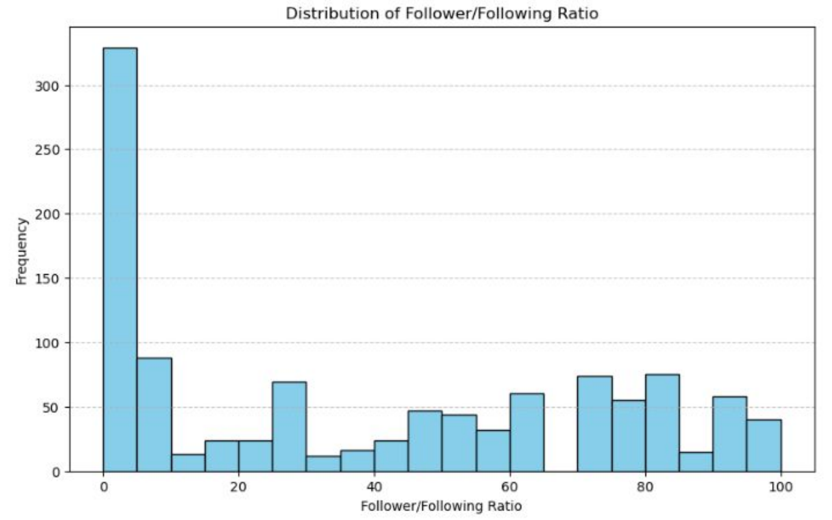
Accuracy Visualization

Goal: Determine influencer status based on following ratio, likes, comments, and number of images (all taken from data)

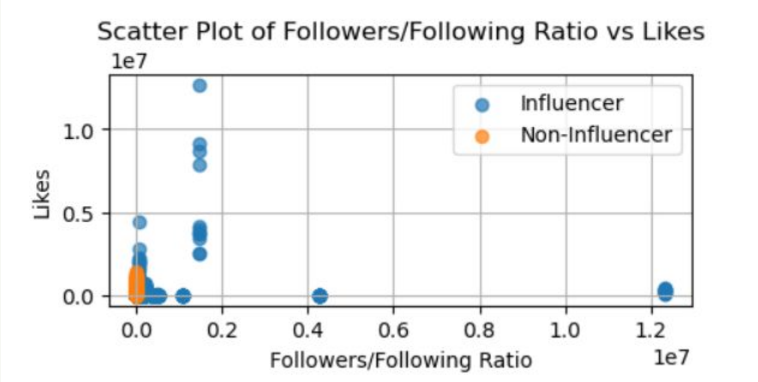
```
Accuracy: 0.9466932725199544
Classification Report:
              precision    recall  f1-score   support

       False       0.97      0.96      0.97      2950
        True       0.82      0.85      0.84       558
```

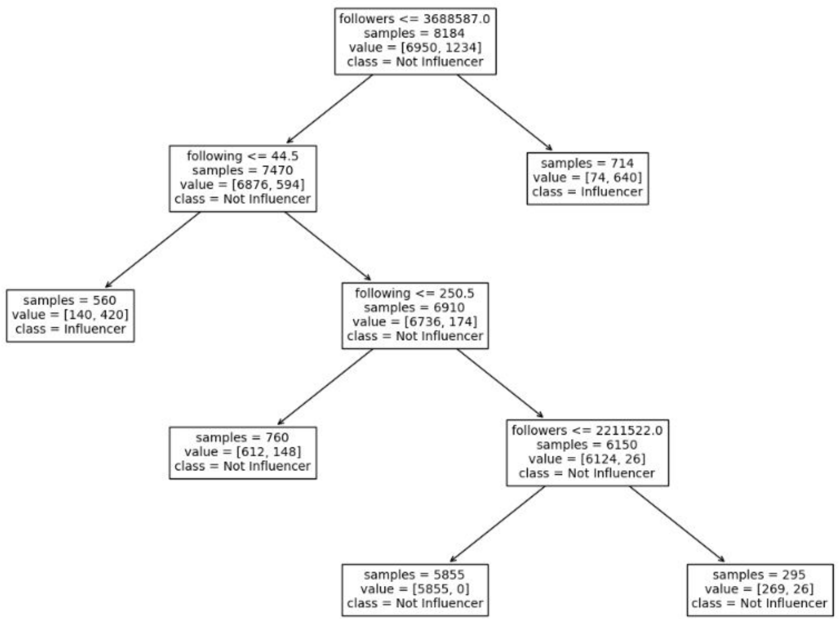
- Accuracy: 94.669%
- Precision: measures accuracy of the positive predictions
    - 97% of “not an influencer” were actually not an influencer
    - 82% of “influencer” were actually influencers
- Recall: percent actually “correct” by the classifier
- F1-Score: mean of precision and recall, shows truth behind predictions
- Support: number of actual occurrences of each category




```
Failed influencer test at node 2: If following <= 44.5, then Influencer
Failed influencer test at node 2: If following <= 44.5, then Influencer
Failed influencer test at node 2: If following <= 44.5, then Influencer
Failed influencer test at node 8: If followers <= 3688587.0, then Influ
encer
Failed influencer test at node 2: If following <= 44.5, then Influencer
Failed influencer test at node 8: If followers <= 3688587.0, then Influ
encer
Failed influencer test at node 2: If following <= 44.5, then Influencer
Failed influencer test at node 8: If followers <= 3688587.0, then Influ
encer
```



Data Visualization



Decision Tree Results

Naive Bayes

- Focuses on conditions of a specific post and tried to determine if the user has the status of “influencer.”
    - Predicts whether the person is an influencer
- We created categories that would help classify
- Accuracy at 83.58%

Decision Tree

- Focused more on follower/following ratio to determine “influencer” status
    - Certain accounts that had massive likes and comments got swept under the rug because of the ratio being too low
- All needs must be met in order to classify as an influencer
- Preset criteria in order to classify
- Higher Accuracy at about 94%

Thank you for your time!