CS4680 Assignment 1:

Determining whether a video is considered “Trending” on YouTube:

Target Variable: Views

Features: 'category_id', 'publish_hour', 'publish_weekday', 'num_tags',
'likes', 'dislikes', 'comment_count', 'description_len'

Dataset:
https://www.kaggle.com/datasets/datasnaek/youtube-new/data

Data used: CAvideos.csv
Size: 45803 entries

- Model Used:
  - Linear Regression:
    - Assumes relationship between features and target are linear
    - Easy to interpret
    - Trains quickly
    - Sensitive to outliers

	actual_views	predicted_views	trending
0	17340	1.691190e+05	False
1	299283	3.279176e+05	False
2	253117	5.174189e+05	False
3	426653	3.005684e+05	True
4	173360	2.267277e+05	False
5	74080	1.549839e+05	False
6	88251	2.744311e+05	False
7	9508610	3.707432e+06	True
8	326798	6.771954e+05	False
9	56927	2.097612e+05	False
10	478905	3.577999e+05	True
11	1452518	3.973976e+05	True
12	25757	2.771546e+05	False
13	1810163	6.605455e+05	True
14	3675189	6.119957e+06	False
15	57297	2.946583e+05	False
16	83274	3.051040e+05	False
17	534122	-7.553046e+04	True
18	9583159	6.265465e+06	True
19	818251	5.291155e+05	True
20	292525	4.205042e+05	False
21	609469	2.364763e+05	True
22	130404	2.240869e+05	False
23	336759	5.129922e+04	True
24	26819	1.788401e+05	False
25	846346	3.410831e+05	True
26	435390	2.289416e+05	True
27	191953	3.104344e+05	False
28	504429	6.740570e+05	False
29	3607	1.190841e+05	False

Linear Regression:
MAE = 547,784
R² = 0.6293

  - Random Forest Regression:

- Can handle non-linear and more complex interactions
- Handles outliers well
- More difficult to interpret than linear

	actual_views	predicted_views	trending
0	17340	5.096146e+04	False
1	299283	3.241330e+05	False
2	253117	3.115188e+05	False
3	426653	2.880718e+05	True
4	173360	2.706227e+05	False
5	74080	6.106000e+04	True
6	88251	1.446200e+05	False
7	9508610	7.067027e+06	True
8	326798	6.345000e+05	False
9	56927	7.006602e+04	False
10	478905	4.453045e+05	True
11	1452518	1.089052e+06	True
12	25757	3.575892e+04	False
13	1810163	2.089668e+06	False
14	3675189	7.774356e+06	False
15	57297	7.116014e+04	False
16	83274	2.068534e+05	False
17	534122	4.905079e+05	True
18	9583159	8.958057e+06	True
19	818251	7.423886e+05	True
20	292525	5.802868e+05	False
21	609469	4.563271e+05	True
22	130404	1.570296e+05	False
23	336759	3.667550e+05	False
24	26819	4.641299e+04	False
25	846346	6.904216e+05	True
26	435390	3.221976e+05	True
27	191953	5.161466e+05	False
28	504429	6.016037e+05	False
29	3607	9.032580e+03	False

```
Random Forest Regression:
  MAE  = 315,944
  R²   = 0.8847
```

Analysis:

The Linear Regression model was simpler, but seems to have less accurate assumptions. Most notably, it did not handle outliers well. For example, one video had 17,340 views, but the linear regression model predicted over 169 thousand. The random forest regression model, on the other hand, predicted only 50 thousand views, which is much closer to the actual number. This remained true for most entries.

The Mean Absolute Error was significantly lower for the Random Forest Regression model, and the Coefficient of Determination(R^2) was closer to 1. This means that the Random Forest Regression model was a better choice for predictions with this dataset. The predictions were much closer to the real values.

Both models were suitable for the task, but the RFR model proved to be superior by a surprising margin.