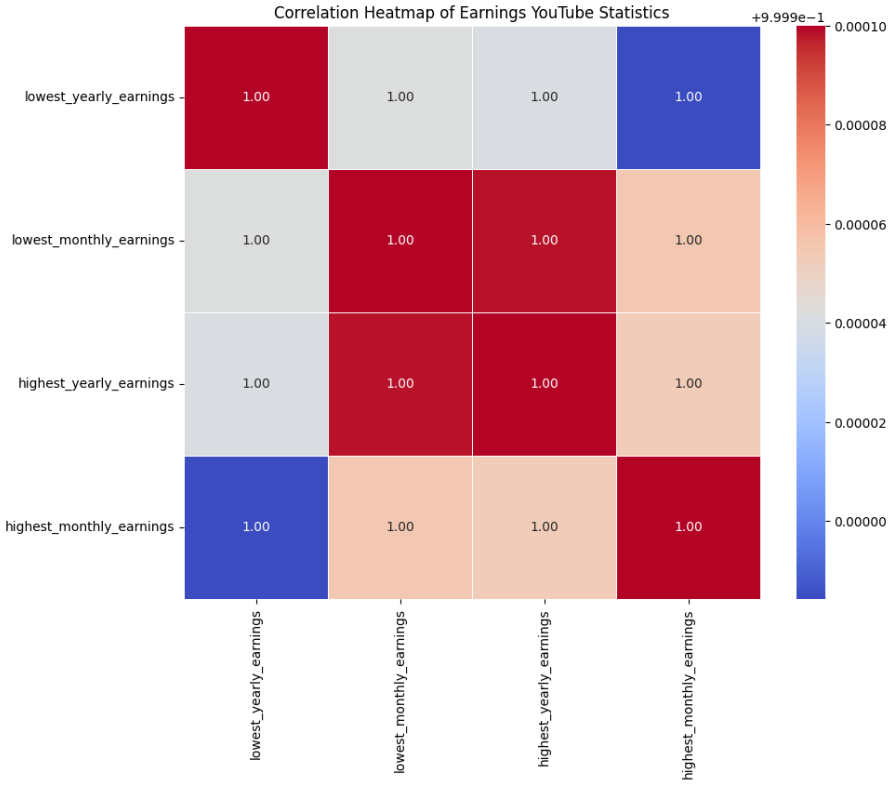
ISTE:780.01 Project Checkpoint 4: System Integration
“ YouTrendify: Youtube Video Insights Engine”

Feature Selection

So initially we took 6 features as the best features to train our regression model which were rank, video views, highest yearly earnings, lowest yearly earnings, lowest monthly earnings and highest yearly earnings respectively.
However, it comes to our notice that the earnings are very much correlated to each other and hence included only the most significant earning feature in our model excluding the others



We used several statistical tests such as a t-test and F-test to gain the significance of each numerical feature.

```
y = df['subscribers'] # This is the target variable for regression model
```

We concluded that the selected features should be

```
# Select the features based on the statistical test results
selected_features = [ 'video views','rank','highest_monthly_earnings','Population']
```

Core Algorithm

Our core algorithm for the analysis was a decision tree regressor. Techniques such as hyperparameter tuning are applied to enhance the models' performance. GridSearchCV is used to optimize parameters for each model.

Machine Learning Models Used

Several machine learning models are employed for the analysis, including Linear Regression, K-Neighbors Regressor, Decision Tree Regressor and Ridge Regression. These models are chosen to predict outcomes based on the dataset features.

Hyperparameter Tuning

We used a hyperparameter grid for each model in which we created an array of values for each parameter for the model. For our decision tree regressor, The hyperparameter grid includes options for the maximum depth of the tree (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf). The GridSearchCV iterates through combinations of these hyperparameters using 5-fold cross-validation and selects the configuration that minimizes the negative mean squared error (scoring='neg_mean_squared_error').
Similarly, we use GridSearchCV for our other comparable models to get the models with the best set of hyperparameters for further fine-tuning.

Best hyperparameters for each model are :

```
Linear Regression:  {'copy_X': True, 'fit_intercept': True, 'n_jobs': 1}

K-NN Regression:  {'metric': 'euclidean', 'n_neighbors': 3, 'weights': 'uniform'}

Decision Tree Regression:  {'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 2}

Ridge Regression:  {'alpha': 0.1, 'fit_intercept': True, 'solver': 'saga'}
```

We used K-fold cross-validation since it mitigates this risk of overfitting by systematically partitioning the dataset into K subsets, training and evaluating the model K=5 times, with each subset serving as the test set exactly once. This provides a more reliable estimate of the model's generalization performance across different data subsets, reducing the impact of data-specific patterns and enhancing the model's robustness.

By using all these methods we got the evaluation metrics for our core algorithm

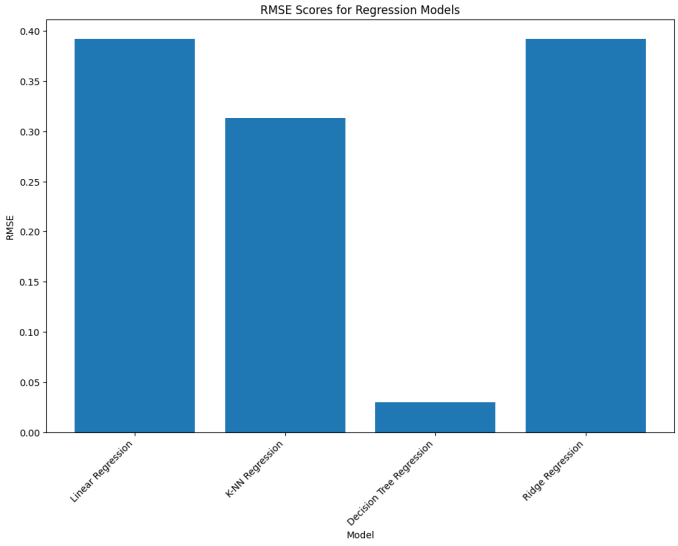
```
Decision Tree Regression Metrics:
MAE: 0.009103503878663015
MSE: 0.0009009317597304189
RMSE: 0.030015525311585318
R^2: 0.9979239252203254
```

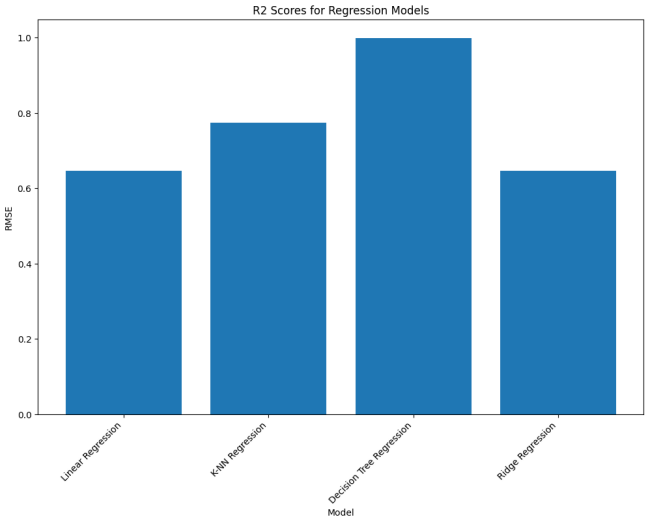
Upon further analysis of the features we can see why the chosen features are important to predict the subscriber count for each YouTube channel:.

- **'video views'**: Higher video views are statistically associated with increased subscriber counts.
- **'rank'**: YouTuber ranking shows a higher statistically significant relationship with the number of subscribers
- **'highest_monthly_earnings'**: Higher monthly earnings are statistically linked to increased subscriber counts.
- **'Population'**: Assuming it refers to a demographic indicator, population size is statistically related to the number of subscribers, suggesting a potential correlation with audience size.

Comparing Evaluation Metrics for each model

	Metrics	Linear Regression	K-NN Regression	Decision Tree Regression	Ridge Regression
0	MAE	0.269173	0.103167	0.009104	0.269158
1	MSE	0.153525	0.098009	0.000901	0.153510
2	RMSE	0.391823	0.313064	0.030016	0.391804
3	R^2	0.646221	0.774151	0.997924	0.646256





Why Decision Tree Regressor is the best model?

- **MAE (Mean Absolute Error):** The Decision Tree Regression has the lowest MAE (0.009104), indicating that, on average, its predictions have the smallest absolute differences from the actual values compared to other models.
- **MSE (Mean Squared Error):** The Decision Tree Regression has the lowest MSE (0.000901), suggesting that its predictions have the smallest squared differences from the actual values, contributing to a more precise prediction.
- **RMSE (Root Mean Squared Error):** The Decision Tree Regression also has the lowest RMSE (0.030016), indicating the smallest root of the average squared differences. This metric is sensitive to larger errors, and a lower RMSE signifies better model performance.
- **R^2 (R-squared):** The Decision Tree Regression achieves a high R^2 value (0.997924), indicating that it explains a significant portion of the variance in the target variable. A high R^2 suggests that the model captures the underlying patterns in the data very well.

Appendix

F-test Results:

	Feature	F-value	P-value
0	rank	30335.746061	0.000000e+00
1	video views	18.692991	1.041613e-210
9	lowest_yearly_earnings	3.013484	1.904865e-32
7	lowest_monthly_earnings	3.007512	2.469587e-32
10	highest_yearly_earnings	3.007442	2.477124e-32
8	highest_monthly_earnings	3.005712	2.670549e-32
11	subscribers_for_last_30_days	2.708762	1.103121e-26
6	video_views_for_the_last_30_days	1.631006	1.601254e-07
14	Population	1.264457	7.700072e-03
2	uploads	1.258493	8.808193e-03
12	created_year	1.251080	1.038604e-02
3	video_views_rank	1.237297	1.401317e-02
5	channel_type_rank	1.201339	2.931165e-02
13	Gross tertiary education enrollment (%)	1.185979	3.938491e-02
18	Longitude	1.117918	1.252436e-01
16	Urban_population	1.110014	1.408734e-01
15	Unemployment rate	1.069846	2.421276e-01
17	Latitude	1.026658	3.896342e-01
4	country_rank	0.968407	6.216770e-01