	Date	Steps	Findings or Decisions mention the rationale behind the decisions	Challenges	Assigned to
1	4/2/2024	TreeMap	quickly identify which tags are associated with the highest total views	not any	Waref ALyousef
2	4/4/2024	Line plot	will visualize the trend of views over time	not any	Waref ALyousef
3	4/9/2024	Duration Joint plot	visualizes the relationship between the duration of videos and the number of views they receive using a scatter plot	not any	Waref ALyousef
4	4/9/2024	Title word count Joint plot	visualize the relationship between the word count of video titles and the number of views they receive	not any	Waref ALyousef
5	4/9/2024	Number of tags Joint plot	This joint plot will visualize the relationship between the number of tags and the number of views for each video.	not any	Waref ALyousef
6	4/9/2024	Time Scatter plots	The series of scatter plots will visualize the relationship between different time components (Year, Month, Day, Hour) and the number of views for videos	not any	Waref ALyousef
7	4/10/2024	Preparing the dataset for registration	we're excluding non-numeric columns like 'Title', 'ID', ' Published_date', and 'Tags', then removing features that the content creator cannot control before uploading the video, such as 'Likes' and 'Comments'. After this process, the features that will remain for model input in X, predicting the 'Views' (Y), include 'Duration', 'Captions', 'Year', 'Month', 'Title_word_count', 'Number_of_Tags', 'Day', and 'Hour'.	not any	Waref ALyousef
8	4/10/2024	Developing 5 Models Using Cross-Validation Split	we are constructing five models, including the Baseline Model along with four alternative models, utilizing a cross-validation split.	not any	Waref ALyousef
9	4/10/2024	Developing 5 Models Using 80% 20% split	we'll employ a different data split, specifically 20% for testing and 80% for training for the five models.	not any	Waref ALyousef
10	4/10/2024	Selecting the Best-Performing Model for Prediction across all splits	A table summarizes the performance metrics of each model across two different data splits: K-fold cross-validation and a 20% testing-80% training split.	not any	Waref ALyousef
11	4/12/2024	Improving the Chosen Model's Performance	improve the effectiveness of the Random Forest Regression model with cross-validation. We'll achieve this by increasing the quality of the data to enhance the model's MAE or MSE.	not any	Waref ALyousef
12	4/12/2024	Removing outliers	Regression models are known to be sensitive to outliers, therefore we will attempt to eliminate them and observe any potential effects on the model.	not any	Waref ALyousef
13	4/12/2024	Assessing and Removing Insignificant Features	We've decided to remove the 'Captions'. The majority of videos have a caption value of 1, and examination through feature importance plots will confirm that captions are not significant and do not impact the model.	not any	Waref ALyousef
14	4/12/2024	Building a new model after improving the data	build a new Random Forest Regression model with cross-validation without outliers and Insignificant Features to check if it will enhance the performance of it.	not any	Waref ALyousef
15	4/12/2024	Comparison of scores	comparison of scores before and after improving the effectiveness of the Random Forest Regression model with cross-validation.	not any	Waref ALyousef
16	4/15/2024	Scatter Plot of Predicted vs. Actual Values	visualize how well the model's predictions align with the actual target values.	the best model used a cross-validation split, resulting in no direct access to the tested and trained data. Consequently, visualizing the model became difficult.	Waref ALyousef
17	4/15/2024	Residual Plot	visualize the distribution of residuals (the differences between the actual and predicted values) to assess the model's performance.	the best model used a cross-validation split, resulting in no direct access to the tested and trained data. Consequently, visualizing the model became difficult.	Waref ALyousef
18	4/15/2024	Feature Importance Plot	visualize the importance of each feature in predicting the target variable.	the best model used a cross-validation split, resulting in no direct access to the tested and trained data. Consequently, visualizing the model became difficult.	Waref ALyousef