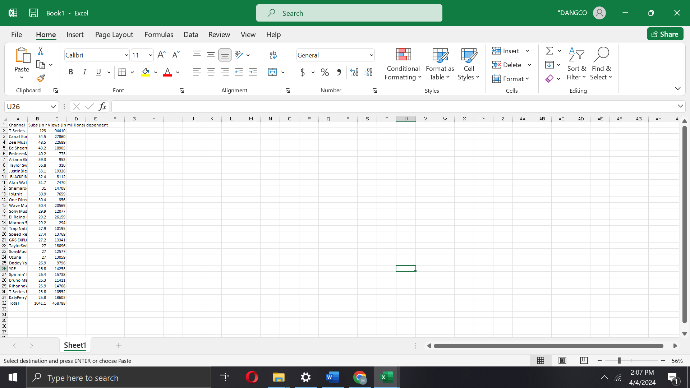
DANGCO, OMAIMAH B. ITD113-Ll 4/4/2024

**Step 1: Understanding the Dataset:**

Begin by thoroughly examining the dataset provided. Understand its structure, including the number of features, their types (categorical, numerical), and the target variable(s) you aim to predict.  
  
Identify any missing values, outliers, or anomalies that may affect the quality of your analysis.  
Determine if preprocessing steps such as handling missing values, encoding categorical variables, or scaling numerical features are necessary to prepare the data for modeling.



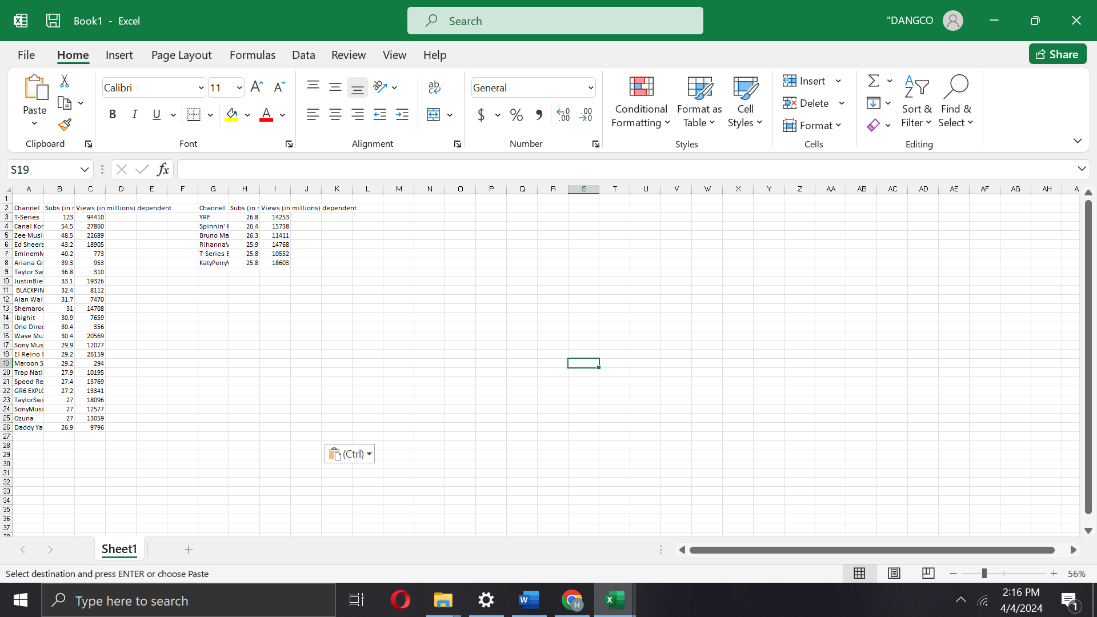
Dataset: YouTube

Data Structure: Numerical

Outliers: Zero Outliers

Missing Values: Zero missing values

**Step 2: Splitting the Dataset:**

  
Divide the dataset into two subsets: a training set and a testing set. A common practice is to allocate around 80% of the data to the training set and the remaining 20% to the testing set.  
Ensure that the split maintains the distribution of the target variable(s) across both sets to prevent bias in the model evaluation.

Stratified Sampling is used to split the Dataset into training and testing set.

Here’s the calculation for stratified sampling:

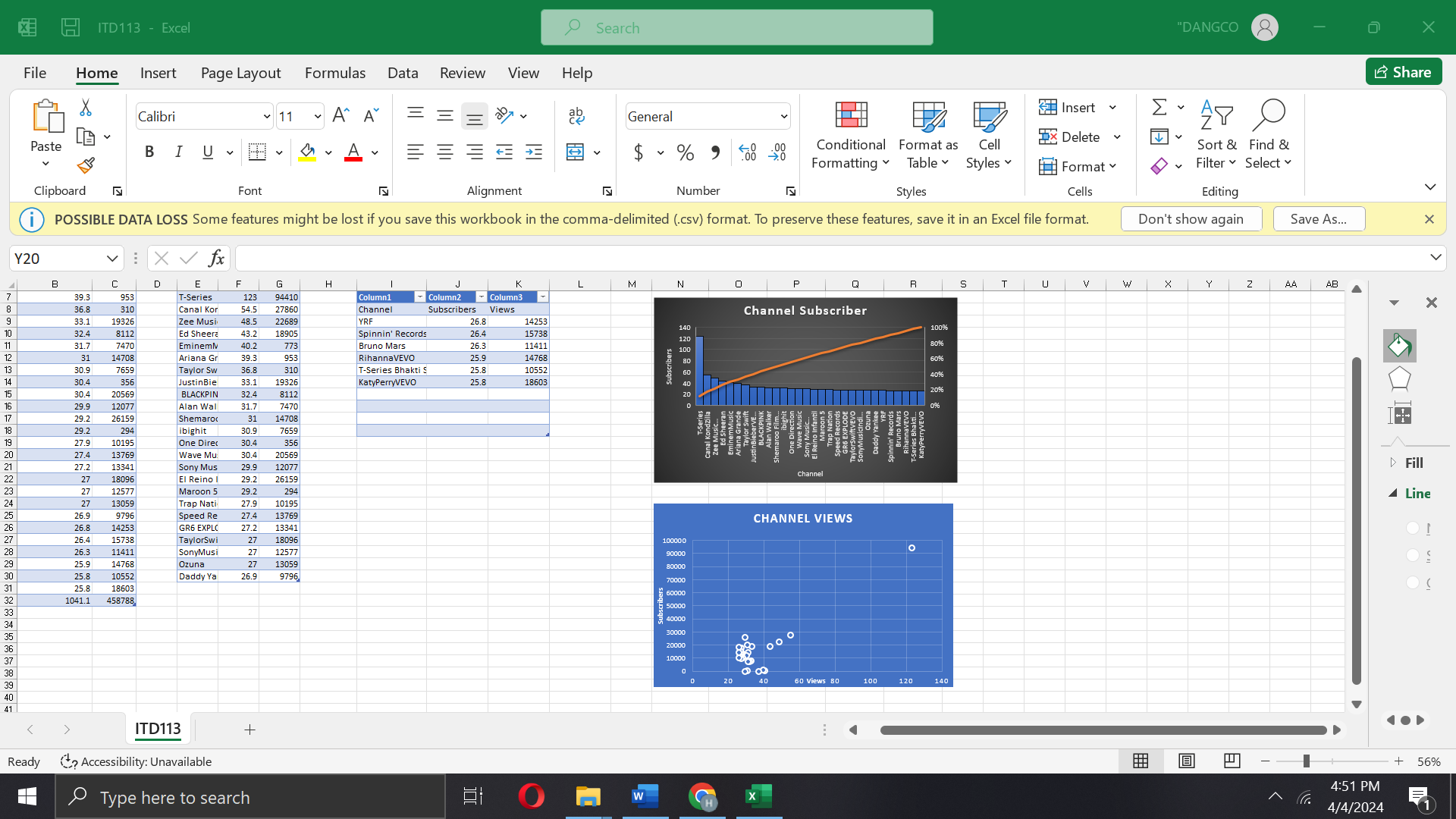
Let's perform the calculations step by step:

1. **Calculate the Proportion of Each Category**:
   * T-Series: 94410/458788≈0.205645878894410​≈0.2056
   * Canal KondZilla: 27860/458788≈0.060745878827860​≈0.0607
   * Zee Music Company: 22689/458788≈0.049445878822689​≈0.0494
   * Ed Sheeran: 18905/458788≈0.041245878818905​≈0.0412
   * EminemMusic: 773/458788≈0.0016848740594​≈0.0016
   * Ariana Grande: 953/458788≈0.0020772121328≈0.0020
2. **Determine the Sample Size for Each Category**:
   * T-Series: 0.2056×6≈1.2336 (rounded to 1)
   * Canal KondZilla: 0.0607×6≈0.3642 (rounded to 0)
   * Zee Music Company: 0.0494×6≈0.29640 (rounded to 0)
   * Ed Sheeran: 0.0412×6≈0.2472 (rounded to 0)
   * EminemMusic: 0.0016×6≈0.0096 (rounded to 0)
   * Ariana Grande: 0.0020×6≈0.012 (rounded to 0)
3. **Randomly Select Samples**:
   * Randomly select 1 sample from T-Series, and 0 samples from Canal KondZilla, Zee Music Company, and so on, ensuring that the total number of samples is 6.

**Step 3: Selecting the Regression Model:**

Choose an appropriate regression model based on the nature of your problem, the distribution of your data, and the assumptions of the regression method.

Consider experimenting with various regression algorithms such as linear regression, decision tree regression, random forest regression, support vector regression, and others to find the best fit for your dataset.

* Considering the dataset I chose, I decided to use simple regression model.

**Step 4: Training the Model:**

Train the selected regression model using the training dataset. During training, the model learns the patterns and relationships present in the data.

Adjust the model's hyperparameters if necessary to optimize its performance. Hyperparameters are settings that control the learning process and can significantly impact the model's predictive ability.

**Step 5: Evaluating the Model:**

Evaluate the performance of the trained model using the testing dataset. Use appropriate evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R2), and others to assess the model's predictive accuracy. Compare the model's performance with baseline models or other benchmarks to gauge its effectiveness.

**Step 6: Model Optimization (Optional):**

Explore techniques to improve the model's performance if necessary. This may include feature selection, hyperparameter tuning, ensemble methods, regularization, or other advanced strategies.  
Continuously monitor the model's performance and iterate on the optimization process to achieve the best results.

**Step 7: Finalizing the Model:**

Once satisfied with the model's performance, finalize it by retraining it on the entire dataset (combining the training and testing sets) to maximize its predictive power.  
Document the finalized model's architecture, parameters, and any other relevant details for future reference.

**Step 8: Predictions:**

Use the finalized model to make predictions on new, unseen data. This step allows you to apply the model to real-world scenarios and make informed decisions based on its predictions.