**Features Engineering Documentation**

In simple words, feature engineering is the process of preparing and transforming raw data into meaningful features that can be easily understood and used by machine learning models. One important part of this process is categorizing continuous data into ranges or bins instead of keeping every unique value. For example, instead of using exact ages like 21, 22, 23, and so on, we can group them into ranges such as 18–25, 26–35, etc. This makes it easier to understand how the data is distributed and helps the model identify patterns more effectively. In short, feature engineering helps simplify data and improve the performance of data analysis or predictive models.

So there are multiple columns in our Dataset where we needed features engineering and the columns are follows:-

* **Content Duration**

df = df.withColumn(

"duration",

F.when(

(

(F.col("duration\_unit") == "Seasons") | (F.col("duration\_unit") == "Season")

) & (F.col("duration\_time") < 3),

"Few Seasons"

).when(

(F.col("duration\_unit") == "Seasons") & (F.col("duration\_time") >= 3),

"Many Seasons"

).when(

(F.col("duration\_unit") == "min") & (F.col("duration\_time") < 100),

"Short Movie"

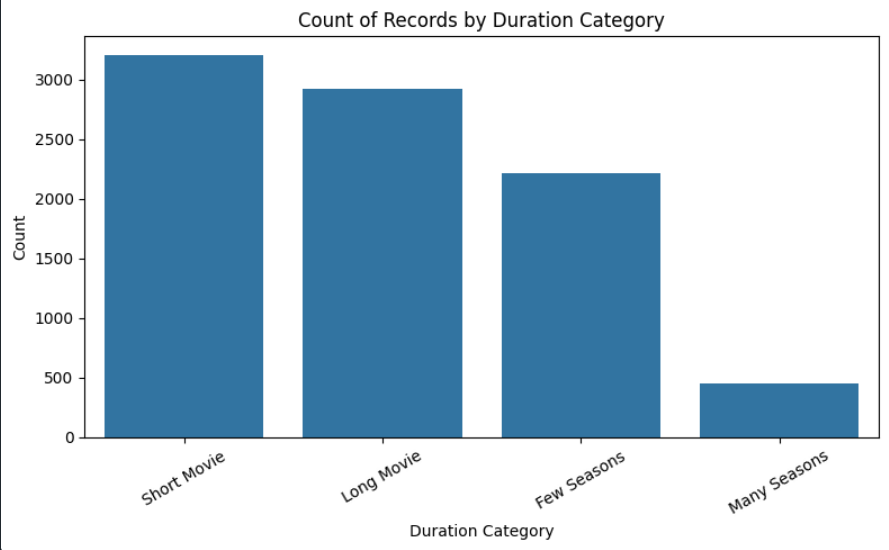
).when(

(F.col("duration\_unit") == "min") & (F.col("duration\_time") >= 100),

"Long Movie"

)

)



* **Content Age**

df = df.withColumn(

'content\_age',

floor(

datediff(

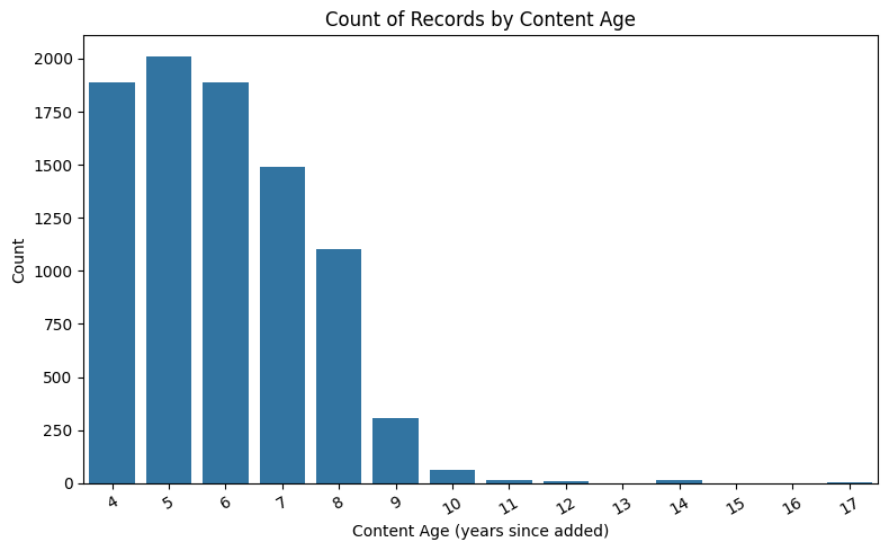
current\_date(),

col('date\_added')

) / 365

)

)



* Title Length

df = df.withColumn(

"title\_length\_category",

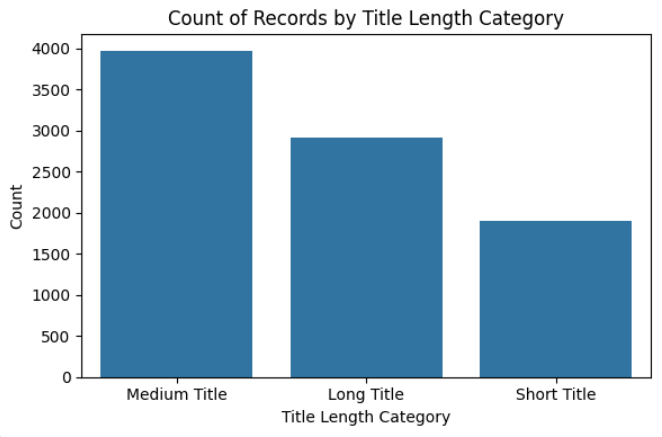
when(length("title") < 10, "Short Title")

.when((length("title") >= 10) & (length("title") < 20), "Medium Title")

.otherwise("Long Title")

)

display(df)



* Ownership Type

df = df.withColumn(

"ownership\_type",

when(df["type"] != "Movie", "original").otherwise("licensed")

)

