Assignment 03

Cindy Guzman

November 21, 2024

1 Load the Dataset

The instruction below provides you with general keywords for columns used in the lightcast file. See the data schema generated after the load dataset code above to use proper column name. For each visualization, **customize colors, fonts, and styles** to avoid a **2.5-point deduction**. Also, **provide a two-sentence explanation** describing key insights drawn from the graph.

1. **Load the Raw Dataset**: -Use Pyspark to the 'lightcast_data.csv' file into DataFrame: -You can reuse the previous code. -Copying code from your friend constitutes plagiarism. DO NOT DO THIS.

```
from pyspark.sql import SparkSession
import pandas as pd
import plotly.express as px
import plotly.io as pio
import numpy as np

np.random.seed(42)

pio.renderers.default = "notebook"

# Initialize Spark Session
spark = SparkSession.builder.appName("LightcastData").getOrCreate()

# Load Data
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiline", "

# Show Schema and Sample Data
# print("---This is Diagnostic check, No need to print it in the final doc---")
# df.printSchema() # comment this line when rendering submission
# df.show(5)
```

```
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/09/22 22:23:12 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfo
java classes where applicable
[Stage 1:>                                                          (0 + 1) / 1]
```