databricks ASP 3.5L - Sort Day Lab

Sort Day Lab

Tasks

1. Define a UDF to label the day of week
2. Apply the UDF to label and sort by day of week
3. Plot active users by day of week as a bar graph

```
%run ../../Includes/Classroom-Setup
```

Start with a DataFrame of the average number of active users by day of week.

This was the resulting `df` in a previous lab.

```
from pyspark.sql.functions import approx_count_distinct, avg, col, date_format, to_date

df = (spark
      .read
      .format("delta")
      .load(events_path)
      .withColumn("ts", (col("event_timestamp") / 1e6).cast("timestamp"))
      .withColumn("date", to_date("ts"))
      .groupBy("date").agg(approx_count_distinct("user_id").alias("active_users"))
      .withColumn("day", date_format(col("date"), "E"))
      .groupBy("day").agg(avg(col("active_users")).alias("avg_users"))
     )

display(df)
```

1. Define UDF to label day of week

Use the `label_day_of_week` function provided below to create the UDF `label_dow_udf`

```
def label_day_of_week(day: str) -> str:
    dow = {"Mon": "1", "Tue": "2", "Wed": "3", "Thu": "4",
           "Fri": "5", "Sat": "6", "Sun": "7"}
    return dow.get(day) + "-" + day
```

```
# TODO
label_dow_udf = FILL_IN
```

2. Apply UDF to label and sort by day of week

- Update the `day` column by applying the UDF and replacing this column
- Sort by `day`
- Plot as a bar graph

```
# TODO
final_df = FILL_IN

display(final_df)
```

Clean up classroom

```
classroom_cleanup()
```