Assignment 03

Makenzie Howard

September 22, 2025

1 Load the Dataset

```
import pandas as pd
import plotly.express as px
import plotly.io as pio
from pyspark.sql import SparkSession
import re
import numpy as np
import plotly.graph_objects as go
from pyspark.sql.functions import col, split, explode, regexp_replace, transform, when
from pyspark.sql import functions as F
from pyspark.sql.functions import col, monotonically_increasing_id

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","t
df.createOrReplaceTempView("job_postings")

# 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 the submission
df.show(5)
```

```
[Stage 22:>                                                         (0 + 1) / 1]

---This is Diagnostic check, No need to print it in the final doc---
root
 |-- ID: string (nullable = true)
 |-- LAST_UPDATED_DATE: string (nullable = true)
 |-- LAST_UPDATED_TIMESTAMP: timestamp (nullable = true)
 |-- DUPLICATES: integer (nullable = true)
 |-- POSTED: string (nullable = true)
 |-- EXPIRED: string (nullable = true)
 |-- DURATION: integer (nullable = true)
 |-- SOURCE_TYPES: string (nullable = true)
 |-- SOURCES: string (nullable = true)
 |-- URL: string (nullable = true)
 |-- ACTIVE_URLS: string (nullable = true)
 |-- ACTIVE_SOURCES_INFO: string (nullable = true)
 |-- TITLE_RAW: string (nullable = true)
 |-- BODY: string (nullable = true)
 |-- MODELED_EXPIRED: string (nullable = true)
 |-- MODELED_DURATION: integer (nullable = true)
 |-- COMPANY: integer (nullable = true)
 |-- COMPANY_NAME: string (nullable = true)
 |-- COMPANY_RAW: string (nullable = true)
 |-- COMPANY_IS_STAFFING: boolean (nullable = true)
 |-- EDUCATION_LEVELS: string (nullable = true)
 |-- EDUCATION_LEVELS_NAME: string (nullable = true)
 |-- MIN_EDULEVELS: integer (nullable = true)
 |-- MIN_EDULEVELS_NAME: string (nullable = true)
 |-- MAX_EDULEVELS: integer (nullable = true)
 |-- MAX_EDULEVELS_NAME: string (nullable = true)
 |-- EMPLOYMENT_TYPE: integer (nullable = true)
 |-- EMPLOYMENT_TYPE_NAME: string (nullable = true)
 |-- MIN_YEARS_EXPERIENCE: integer (nullable = true)
 |-- MAX_YEARS_EXPERIENCE: integer (nullable = true)
 |-- IS_INTERNSHIP: boolean (nullable = true)
 |-- SALARY: integer (nullable = true)
 |-- REMOTE_TYPE: integer (nullable = true)
 |-- REMOTE_TYPE_NAME: string (nullable = true)
 |-- ORIGINAL_PAY_PERIOD: string (nullable = true)
 |-- SALARY_TO: integer (nullable = true)
 |-- SALARY_FROM: integer (nullable = true)
 |-- LOCATION: string (nullable = true)
```

```
|-- CITY: string (nullable = true)
|-- CITY_NAME: string (nullable = true)
|-- COUNTY: integer (nullable = true)
|-- COUNTY_NAME: string (nullable = true)
|-- MSA: integer (nullable = true)
|-- MSA_NAME: string (nullable = true)
|-- STATE: integer (nullable = true)
|-- STATE_NAME: string (nullable = true)
|-- COUNTY_OUTGOING: integer (nullable = true)
|-- COUNTY_NAME_OUTGOING: string (nullable = true)
|-- COUNTY_INCOMING: integer (nullable = true)
|-- COUNTY_NAME_INCOMING: string (nullable = true)
|-- MSA_OUTGOING: integer (nullable = true)
|-- MSA_NAME_OUTGOING: string (nullable = true)
|-- MSA_INCOMING: integer (nullable = true)
|-- MSA_NAME_INCOMING: string (nullable = true)
|-- NAICS2: integer (nullable = true)
|-- NAICS2_NAME: string (nullable = true)
|-- NAICS3: integer (nullable = true)
|-- NAICS3_NAME: string (nullable = true)
|-- NAICS4: integer (nullable = true)
|-- NAICS4_NAME: string (nullable = true)
|-- NAICS5: integer (nullable = true)
|-- NAICS5_NAME: string (nullable = true)
|-- NAICS6: integer (nullable = true)
|-- NAICS6_NAME: string (nullable = true)
|-- TITLE: string (nullable = true)
|-- TITLE_NAME: string (nullable = true)
|-- TITLE_CLEAN: string (nullable = true)
|-- SKILLS: string (nullable = true)
|-- SKILLS_NAME: string (nullable = true)
|-- SPECIALIZED_SKILLS: string (nullable = true)
|-- SPECIALIZED_SKILLS_NAME: string (nullable = true)
|-- CERTIFICATIONS: string (nullable = true)
|-- CERTIFICATIONS_NAME: string (nullable = true)
|-- COMMON_SKILLS: string (nullable = true)
|-- COMMON_SKILLS_NAME: string (nullable = true)
|-- SOFTWARE_SKILLS: string (nullable = true)
|-- SOFTWARE_SKILLS_NAME: string (nullable = true)
|-- ONET: string (nullable = true)
|-- ONET_NAME: string (nullable = true)
|-- ONET_2019: string (nullable = true)
|-- ONET_2019_NAME: string (nullable = true)
```

```
|-- CIP6: string (nullable = true)
|-- CIP6_NAME: string (nullable = true)
|-- CIP4: string (nullable = true)
|-- CIP4_NAME: string (nullable = true)
|-- CIP2: string (nullable = true)
|-- CIP2_NAME: string (nullable = true)
|-- SOC_2021_2: string (nullable = true)
|-- SOC_2021_2_NAME: string (nullable = true)
|-- SOC_2021_3: string (nullable = true)
|-- SOC_2021_3_NAME: string (nullable = true)
|-- SOC_2021_4: string (nullable = true)
|-- SOC_2021_4_NAME: string (nullable = true)
|-- SOC_2021_5: string (nullable = true)
|-- SOC_2021_5_NAME: string (nullable = true)
|-- LOT_CAREER_AREA: integer (nullable = true)
|-- LOT_CAREER_AREA_NAME: string (nullable = true)
|-- LOT_OCCUPATION: integer (nullable = true)
|-- LOT_OCCUPATION_NAME: string (nullable = true)
|-- LOT_SPECIALIZED_OCCUPATION: integer (nullable = true)
|-- LOT_SPECIALIZED_OCCUPATION_NAME: string (nullable = true)
|-- LOT_OCCUPATION_GROUP: integer (nullable = true)
|-- LOT_OCCUPATION_GROUP_NAME: string (nullable = true)
|-- LOT_V6_SPECIALIZED_OCCUPATION: integer (nullable = true)
|-- LOT_V6_SPECIALIZED_OCCUPATION_NAME: string (nullable = true)
|-- LOT_V6_OCCUPATION: integer (nullable = true)
|-- LOT_V6_OCCUPATION_NAME: string (nullable = true)
|-- LOT_V6_OCCUPATION_GROUP: integer (nullable = true)
|-- LOT_V6_OCCUPATION_GROUP_NAME: string (nullable = true)
|-- LOT_V6_CAREER_AREA: integer (nullable = true)
|-- LOT_V6_CAREER_AREA_NAME: string (nullable = true)
|-- SOC_2: string (nullable = true)
|-- SOC_2_NAME: string (nullable = true)
|-- SOC_3: string (nullable = true)
|-- SOC_3_NAME: string (nullable = true)
|-- SOC_4: string (nullable = true)
|-- SOC_4_NAME: string (nullable = true)
|-- SOC_5: string (nullable = true)
|-- SOC_5_NAME: string (nullable = true)
|-- LIGHTCAST_SECTORS: string (nullable = true)
|-- LIGHTCAST_SECTORS_NAME: string (nullable = true)
|-- NAICS_2022_2: integer (nullable = true)
|-- NAICS_2022_2_NAME: string (nullable = true)
|-- NAICS_2022_3: integer (nullable = true)
```

```
 |-- NAICS_2022_3_NAME: string (nullable = true)
 |-- NAICS_2022_4: integer (nullable = true)
 |-- NAICS_2022_4_NAME: string (nullable = true)
 |-- NAICS_2022_5: integer (nullable = true)
 |-- NAICS_2022_5_NAME: string (nullable = true)
 |-- NAICS_2022_6: integer (nullable = true)
 |-- NAICS_2022_6_NAME: string (nullable = true)

+--------------------+----------------+----------------------+----------
+--------+---------+--------+--------------------+--------------------+--------------------
+----------+------------------+--------------------+--------------------
+--------------+---------------+--------+--------------------+-----------
+------------------+--------------+--------------------+-------------
+------------------+------------+------------------+--------------+--------------------
+--------------------+------------------+------------+------+-----------
+----------------+------------------+---------+-----------+--------------------
+--------------------+------------+------+-------------+-----+--------------------
+-----+----------+--------------+--------------------+--------------+--------------------
+-----------+--------------------+-----------+--------------------+------
+--------------------+------+--------------------+------+--------------------
+------+--------------------+------+--------------------+------------------
+------------------+------------------+--------------------+--------------------
+--------------------+----------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+--------------------
+---------+--------------------+---------+--------------------+--------------------
+--------------------+--------------------+--------------------+--------------------
+--------------------+---------+--------------------+---------+--------------------
+---------+---------------+---------+--------------+---------------+--------------------
+--------------+------------------+----------------------------+----------------------------
+--------------------+--------------------------+-----------------------------
+-----------------------------------+-----------------+----------------------
+----------------------+---------------------------+-----------------
+----------------------+-------+--------------------+-------+--------------------
+-------+---------------+-------+---------------+-----------------+----------------------
+-----------+--------------------+-----------+--------------------+------------
+--------------------+-----------+--------------------+-----------+--------------------
+
|                  ID|LAST_UPDATED_DATE|LAST_UPDATED_TIMESTAMP|DUPLICATES|  POSTED|  EXPIRED
+--------------------+----------------+----------------------+----------
+--------+---------+--------+--------------------+--------------------+--------------------
+----------+------------------+--------------------+--------------------
+--------------+---------------+--------+--------------------+-----------
+--------------------+----------------+----------------------+-------------
```

```
+------------------+------------+------------------+--------------+--------------------
+--------------------+--------------------+-------------+------+-----------
+----------------+------------------+---------+-----------+--------------------
+--------------------+-------------+------+--------------+-----+--------------------
+-----+----------+---------------+--------------------+--------------+--------------------
+------------+--------------------+------------+--------------------+------
+--------------------+------+--------------------+------+--------------------
+------+--------------------+------+--------------------+------------------
+--------------------+--------------------+--------------------+--------------------
+--------------------+----------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+--------------------
+----------+--------------------+----------+--------------------+--------------------
+--------------------+----------+--------------------+----------+--------------------
+----------+---------------+----------+---------------+--------------+--------------------
+--------------+--------------------+--------------------------+----------------------------
+--------------------+--------------------------+-----------------------------
+----------------------------------+-----------------+----------------------
+------------------------+-----------------------------+------------------
+------------------------+-------+--------------------+-------+--------------------
+-------+---------------+-------+---------------+-----------------+----------------------
+------------+--------------------+------------+--------------------+------------
+--------------------+------------+--------------------+------------+--------------------
+
|1f57d95acf4dc67ed...|         9/6/2024|  2024-09-06 20:32:...|         0|6/2/2024| 6/8/2024
May-2024\n\nEn...|       6/8/2024|                6|  894731|          Murphy USA| Murphy USA
time (> 32 h...|                    2|                   2|        false|  NULL|          0|
2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n  "45.0601",\n...|[\n  "Econ
0000|Computer and Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-
2051|Data Scientists|             23|Information Techn...|        231010|Business Intellig..
0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-
2051|Data Scientists|        [\n  7\n]|  [\n  "Artificial ...|          44|        Retail Tra
|0cb072af26757b6c4...|         8/2/2024|  2024-08-02 17:08:...|         0|6/2/2024| 8/1/2024
time (> 32 h...|                    3|                   3|        false|  NULL|          1|
Watervill...|   23|     Maine|          23011|        Kennebec, ME|          23011|        K
Watervill...|       12300|Augusta-Watervill...|    56|Administrative an...|   561|Administra
2051.01|Business Intellig...|15-2051.01|Business Intellig...|                  []|
0000|Computer and Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-
2051|Data Scientists|             23|Information Techn...|        231010|Business Intellig..
0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-
2051|Data Scientists|             NULL|                  NULL|          56|Administrative an
|85318b12b3331fa49...|         9/6/2024|  2024-09-06 20:32:...|         1|6/2/2024| 7/7/2024
time (> 32 h...|                    5|                NULL|        false|  NULL|          0|
```

```
Fort Worth...|   48|     Texas|          48113|          Dallas, TX|          48113|
Fort Worth...|       19100|Dallas-Fort Worth...|    52|Finance and Insur...|   524|Insurance
2051.01|Business Intellig...|15-2051.01|Business Intellig...|                  []|
0000|Computer and Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-
2051|Data Scientists|             23|Information Techn...|        231113|Data / Data Minin..
0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-
2051|Data Scientists|             NULL|                  NULL|          52|Finance and Insur
|1b5c3941e54a1889e...|         9/6/2024|  2024-09-06 20:32:...|         1|6/2/2024|7/20/2024
time (> 32 h...|                   3|                NULL|        false|  NULL|          0|
Mesa-Chan...|    4|   Arizona|           4013|        Maricopa, AZ|           4013|        Ma
Mesa-Chan...|       38060|Phoenix-Mesa-Chan...|    52|Finance and Insur...|   522|Credit Inte
2051.01|Business Intellig...|15-2051.01|Business Intellig...|                  []|
0000|Computer and Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-
2051|Data Scientists|             23|Information Techn...|        231113|Data / Data Minin..
0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-
2051|Data Scientists|        [\n  6\n]|  [\n  "Data Privac...|          52|Finance and Insur
|cb5ca25f02bdf25c1...|        6/19/2024|   2024-06-19 07:00:00|         0|6/2/2024|6/17/2024
time / full-...|                NULL|                NULL|        false| 92500|          0|
2051.01|Business Intellig...|15-2051.01|Business Intellig...|                  []|
0000|Computer and Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-
2051|Data Scientists|             23|Information Techn...|        231010|Business Intellig..
0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-
2051|Data Scientists|             NULL|                  NULL|          99|Unclassified Indu
+--------------------+-----------------+----------------------+----------
+--------+---------+--------+-------------------+-------------------+--------------------
+-----------+-------------------+--------------------+--------------------
+---------------+----------------+--------+--------------------+-----------
+-------------------+----------------+----------------------+-------------
+--------------------+-------------+-------------------+---------------+--------------------
+--------------------+--------------------+-------------+------+-----------
+----------------+--------------------+---------+-----------+--------------------
+--------------------+-------------+------+-------------+-----+--------------------
+-----+----------+---------------+--------------------+---------------+--------------------
+------------+--------------------+------------+--------------------+------
+--------------------+------+--------------------+------+--------------------
+------+--------------------+------+--------------------+------------------
+-------------------+--------------------+--------------------+--------------------
+--------------------+-----------------------+--------------------+--------------------
+--------------------+--------------------+--------------------+--------------------
+----------+--------------------+----------+--------------------+--------------------
+--------------------+--------------------+--------------------+--------------------
+--------------------+----------+--------------------+----------+--------------------
+----------+---------------+----------+---------------+---------------+--------------------
```

```
+--------------+--------------------+--------------------------+-----------------------------
+--------------------+-------------------------+-----------------------------
+----------------------------------+-----------------+----------------------
+------------------------+-----------------------------+------------------
+------------------------+--------+--------------------+--------+--------------------
+--------+---------------+--------+---------------+-----------------+----------------------
+------------+---------------------+------------+---------------------+------------
+--------------------+------------+--------------------+------------+--------------------
+
only showing top 5 rows
```