

# 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", "true").text("lightcast_data.txt")
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)
```

---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)

```

ID	LAST_UPDATED_DATE	LAST_UPDATED_TIMESTAMP	DUPLICATES	POSTED	EXPIRED
1	2023-01-01	1672531200	1	1	1
2	2023-01-02	1672617600	1	1	1
3	2023-01-03	1672704000	1	1	1
4	2023-01-04	1672790400	1	1	1
5	2023-01-05	1672876800	1	1	1
6	2023-01-06	1672963200	1	1	1
7	2023-01-07	1673049600	1	1	1
8	2023-01-08	1673136000	1	1	1
9	2023-01-09	1673222400	1	1	1
10	2023-01-10	1673308800	1	1	1
11	2023-01-11	1673395200	1	1	1
12	2023-01-12	1673481600	1	1	1
13	2023-01-13	1673568000	1	1	1
14	2023-01-14	1673654400	1	1	1
15	2023-01-15	1673740800	1	1	1
16	2023-01-16	1673827200	1	1	1
17	2023-01-17	1673913600	1	1	1
18	2023-01-18	1674000000	1	1	1
19	2023-01-19	1674086400	1	1	1
20	2023-01-20	1674172800	1	1	1
21	2023-01-21	1674259200	1	1	1
22	2023-01-22	1674345600	1	1	1
23	2023-01-23	1674432000	1	1	1
24	2023-01-24	1674518400	1	1	1
25	2023-01-25	1674604800	1	1	1
26	2023-01-26	1674691200	1	1	1
27	2023-01-27	1674777600	1	1	1
28	2023-01-28	1674864000	1	1	1
29	2023-01-29	1674950400	1	1	1
30	2023-01-30	1675036800	1	1	1
31	2023-01-31	1675123200	1	1	1
32	2023-02-01	1675209600	1	1	1
33	2023-02-02	1675296000	1	1	1
34	2023-02-03	1675382400	1	1	1
35	2023-02-04	1675468800	1	1	1
36	2023-02-05	1675555200	1	1	1
37	2023-02-06	1675641600	1	1	1
38	2023-02-07	1675728000	1	1	1
39	2023-02-08	1675814400	1	1	1
40	2023-02-09	1675900800	1	1	1
41	2023-02-10	1675987200	1	1	1
42	2023-02-11	1676073600	1	1	1
43	2023-02-12	1676160000	1	1	1
44	2023-02-13	1676246400	1	1	1
45	2023-02-14	1676332800	1	1	1
46	2023-02-15	1676419200	1	1	1
47	2023-02-16	1676505600	1	1	1
48	2023-02-17	1676592000	1	1	1
49	2023-02-18	1676678400	1	1	1
50	2023-02-19	1676764800	1	1	1
51	2023-02-20	1676851200	1	1	1
52	2023-02-21	1676937600	1	1	1
53	2023-02-22	1677024000	1	1	1
54	2023-02-23	1677110400	1	1	1
55	2023-02-24	1677196800	1	1	1
56	2023-02-25	1677283200	1	1	1
57	2023-02-26	1677369600	1	1	1
58	2023-02-27	1677456000	1	1	1
59	2023-02-28	1677542400	1	1	1
60	2023-03-01	1677628800	1	1	1
61	2023-03-02	1677715200	1	1	1
62	2023-03-03	1677801600	1	1	1
63	2023-03-04	1677888000	1	1	1
64	2023-03-05	1677974400	1	1	1
65	2023-03-06	1678060800	1	1	1
66	2023-03-07	1678147200	1	1	1
67	2023-03-08	1678233600	1	1	1
68	2				

```

+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+
+
|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 Tr
|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...| 56|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
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