Assignment 03

Caroline O’Sullivan (Boston University)  
September 20, 2025

from pyspark.sql import SparkSession  
import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
import numpy as np  
import plotly.graph\_objects as go  
import numpy as np  
import re  
from pyspark.sql.functions import col, split, explode, regexp\_replace, transform, when, count  
from pyspark.sql import functions as F  
from pyspark.sql.functions import col, monotonically\_increasing\_id  
import ast  
  
np.random.seed(42)  
  
pio.renderers.default = "notebook+svg"  
  
# Initialize Spark Session  
spark = SparkSession.builder.appName("LightcastData").getOrCreate()  
  
# Load Data  
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine","true").option("escape", "\"").csv("data/lightcast\_job\_postings.csv")  
  
# 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)

Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/09/23 22:20:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

print(df.columns)  
df.show(5)

['ID', 'LAST\_UPDATED\_DATE', 'LAST\_UPDATED\_TIMESTAMP', 'DUPLICATES', 'POSTED', 'EXPIRED', 'DURATION', 'SOURCE\_TYPES', 'SOURCES', 'URL', 'ACTIVE\_URLS', 'ACTIVE\_SOURCES\_INFO', 'TITLE\_RAW', 'BODY', 'MODELED\_EXPIRED', 'MODELED\_DURATION', 'COMPANY', 'COMPANY\_NAME', 'COMPANY\_RAW', 'COMPANY\_IS\_STAFFING', 'EDUCATION\_LEVELS', 'EDUCATION\_LEVELS\_NAME', 'MIN\_EDULEVELS', 'MIN\_EDULEVELS\_NAME', 'MAX\_EDULEVELS', 'MAX\_EDULEVELS\_NAME', 'EMPLOYMENT\_TYPE', 'EMPLOYMENT\_TYPE\_NAME', 'MIN\_YEARS\_EXPERIENCE', 'MAX\_YEARS\_EXPERIENCE', 'IS\_INTERNSHIP', 'SALARY', 'REMOTE\_TYPE', 'REMOTE\_TYPE\_NAME', 'ORIGINAL\_PAY\_PERIOD', 'SALARY\_TO', 'SALARY\_FROM', 'LOCATION', 'CITY', 'CITY\_NAME', 'COUNTY', 'COUNTY\_NAME', 'MSA', 'MSA\_NAME', 'STATE', 'STATE\_NAME', 'COUNTY\_OUTGOING', 'COUNTY\_NAME\_OUTGOING', 'COUNTY\_INCOMING', 'COUNTY\_NAME\_INCOMING', 'MSA\_OUTGOING', 'MSA\_NAME\_OUTGOING', 'MSA\_INCOMING', 'MSA\_NAME\_INCOMING', 'NAICS2', 'NAICS2\_NAME', 'NAICS3', 'NAICS3\_NAME', 'NAICS4', 'NAICS4\_NAME', 'NAICS5', 'NAICS5\_NAME', 'NAICS6', 'NAICS6\_NAME', 'TITLE', 'TITLE\_NAME', 'TITLE\_CLEAN', 'SKILLS', 'SKILLS\_NAME', 'SPECIALIZED\_SKILLS', 'SPECIALIZED\_SKILLS\_NAME', 'CERTIFICATIONS', 'CERTIFICATIONS\_NAME', 'COMMON\_SKILLS', 'COMMON\_SKILLS\_NAME', 'SOFTWARE\_SKILLS', 'SOFTWARE\_SKILLS\_NAME', 'ONET', 'ONET\_NAME', 'ONET\_2019', 'ONET\_2019\_NAME', 'CIP6', 'CIP6\_NAME', 'CIP4', 'CIP4\_NAME', 'CIP2', 'CIP2\_NAME', 'SOC\_2021\_2', 'SOC\_2021\_2\_NAME', 'SOC\_2021\_3', 'SOC\_2021\_3\_NAME', 'SOC\_2021\_4', 'SOC\_2021\_4\_NAME', 'SOC\_2021\_5', 'SOC\_2021\_5\_NAME', 'LOT\_CAREER\_AREA', 'LOT\_CAREER\_AREA\_NAME', 'LOT\_OCCUPATION', 'LOT\_OCCUPATION\_NAME', 'LOT\_SPECIALIZED\_OCCUPATION', 'LOT\_SPECIALIZED\_OCCUPATION\_NAME', 'LOT\_OCCUPATION\_GROUP', 'LOT\_OCCUPATION\_GROUP\_NAME', 'LOT\_V6\_SPECIALIZED\_OCCUPATION', 'LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME', 'LOT\_V6\_OCCUPATION', 'LOT\_V6\_OCCUPATION\_NAME', 'LOT\_V6\_OCCUPATION\_GROUP', 'LOT\_V6\_OCCUPATION\_GROUP\_NAME', 'LOT\_V6\_CAREER\_AREA', 'LOT\_V6\_CAREER\_AREA\_NAME', 'SOC\_2', 'SOC\_2\_NAME', 'SOC\_3', 'SOC\_3\_NAME', 'SOC\_4', 'SOC\_4\_NAME', 'SOC\_5', 'SOC\_5\_NAME', 'LIGHTCAST\_SECTORS', 'LIGHTCAST\_SECTORS\_NAME', 'NAICS\_2022\_2', 'NAICS\_2022\_2\_NAME', 'NAICS\_2022\_3', 'NAICS\_2022\_3\_NAME', 'NAICS\_2022\_4', 'NAICS\_2022\_4\_NAME', 'NAICS\_2022\_5', 'NAICS\_2022\_5\_NAME', 'NAICS\_2022\_6', 'NAICS\_2022\_6\_NAME']

25/09/23 22:20:49 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| ID|LAST\_UPDATED\_DATE|LAST\_UPDATED\_TIMESTAMP|DUPLICATES| POSTED| EXPIRED|DURATION| SOURCE\_TYPES| SOURCES| URL|ACTIVE\_URLS|ACTIVE\_SOURCES\_INFO| TITLE\_RAW| BODY|MODELED\_EXPIRED|MODELED\_DURATION| COMPANY| COMPANY\_NAME|COMPANY\_RAW|COMPANY\_IS\_STAFFING|EDUCATION\_LEVELS|EDUCATION\_LEVELS\_NAME|MIN\_EDULEVELS| MIN\_EDULEVELS\_NAME|MAX\_EDULEVELS|MAX\_EDULEVELS\_NAME|EMPLOYMENT\_TYPE|EMPLOYMENT\_TYPE\_NAME|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|IS\_INTERNSHIP|SALARY|REMOTE\_TYPE|REMOTE\_TYPE\_NAME|ORIGINAL\_PAY\_PERIOD|SALARY\_TO|SALARY\_FROM| LOCATION| CITY| CITY\_NAME|COUNTY| COUNTY\_NAME| MSA| MSA\_NAME|STATE|STATE\_NAME|COUNTY\_OUTGOING|COUNTY\_NAME\_OUTGOING|COUNTY\_INCOMING|COUNTY\_NAME\_INCOMING|MSA\_OUTGOING| MSA\_NAME\_OUTGOING|MSA\_INCOMING| MSA\_NAME\_INCOMING|NAICS2| NAICS2\_NAME|NAICS3| NAICS3\_NAME|NAICS4| NAICS4\_NAME|NAICS5| NAICS5\_NAME|NAICS6| NAICS6\_NAME| TITLE| TITLE\_NAME| TITLE\_CLEAN| SKILLS| SKILLS\_NAME| SPECIALIZED\_SKILLS|SPECIALIZED\_SKILLS\_NAME| CERTIFICATIONS| CERTIFICATIONS\_NAME| COMMON\_SKILLS| COMMON\_SKILLS\_NAME| SOFTWARE\_SKILLS|SOFTWARE\_SKILLS\_NAME| ONET| ONET\_NAME| ONET\_2019| ONET\_2019\_NAME| CIP6| CIP6\_NAME| CIP4| CIP4\_NAME| CIP2| CIP2\_NAME|SOC\_2021\_2| SOC\_2021\_2\_NAME|SOC\_2021\_3| SOC\_2021\_3\_NAME|SOC\_2021\_4|SOC\_2021\_4\_NAME|SOC\_2021\_5|SOC\_2021\_5\_NAME|LOT\_CAREER\_AREA|LOT\_CAREER\_AREA\_NAME|LOT\_OCCUPATION| LOT\_OCCUPATION\_NAME|LOT\_SPECIALIZED\_OCCUPATION|LOT\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_OCCUPATION\_GROUP|LOT\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_SPECIALIZED\_OCCUPATION|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION|LOT\_V6\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION\_GROUP|LOT\_V6\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_CAREER\_AREA|LOT\_V6\_CAREER\_AREA\_NAME| SOC\_2| SOC\_2\_NAME| SOC\_3| SOC\_3\_NAME| SOC\_4| SOC\_4\_NAME| SOC\_5| SOC\_5\_NAME|LIGHTCAST\_SECTORS|LIGHTCAST\_SECTORS\_NAME|NAICS\_2022\_2| NAICS\_2022\_2\_NAME|NAICS\_2022\_3| NAICS\_2022\_3\_NAME|NAICS\_2022\_4| NAICS\_2022\_4\_NAME|NAICS\_2022\_5| NAICS\_2022\_5\_NAME|NAICS\_2022\_6| NAICS\_2022\_6\_NAME|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
|1f57d95acf4dc67ed...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 6/8/2024| 6| [\n "Company"\n]|[\n "brassring.c...|[\n "https://sjo...| []| NULL|Enterprise Analys...|31-May-2024\n\nEn...| 6/8/2024| 6| 894731| Murphy USA| Murphy USA| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 2| 2| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.20...|RWwgRG9yYWRvLCBBUg==|El Dorado, AR| 5139| Union, AR|20980| El Dorado, AR| 5| Arkansas| 5139| Union, AR| 5139| Union, AR| 20980| El Dorado, AR| 20980| El Dorado, AR| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...|441330|Automotive Parts ...|ET29C073C03D1F86B4|Enterprise Analysts|enterprise analys...|[\n "KS126DB6T06...|[\n "Merchandisi...|[\n "KS126DB6T06...| [\n "Merchandisi...| []| []|[\n "KS126706DPF...|[\n "Mathematics...|[\n "KS440W865GC...|[\n "SQL (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "45.0601",\n...|[\n "Economics, ...|[\n "45.06",\n ...|[\n "Economics",...|[\n "45",\n "27...|[\n "Social Scie...| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101011| General ERP Analy...| 2310| Business Intellig...| 23101011| General ERP Analy...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 7\n]| [\n "Artificial ...| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...| 441330|Automotive Parts ...|  
|0cb072af26757b6c4...| 8/2/2024| 2024-08-02 17:08:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Job Board"\n]| [\n "maine.gov"\n]|[\n "https://job...| []| NULL|Oracle Consultant...|Oracle Consultant...| 8/1/2024| NULL| 133098|Smx Corporation L...| SMX| true| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| 3| false| NULL| 1| Remote| NULL| NULL| NULL|{\n "lat": 44.31...| QXVndXN0YSwgTUU=| Augusta, ME| 23011| Kennebec, ME|12300|Augusta-Watervill...| 23| Maine| 23011| Kennebec, ME| 23011| Kennebec, ME| 12300|Augusta-Watervill...| 12300|Augusta-Watervill...| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...|561320|Temporary Help Se...|ET21DDA63780A7DC09| Oracle Consultants|oracle consultant...|[\n "KS122626T55...|[\n "Procurement...|[\n "KS122626T55...| [\n "Procurement...| []| []| []| []|[\n "BGSBF3F508F...|[\n "Oracle Busi...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...| 561320|Temporary Help Se...|  
|85318b12b3331fa49...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024| 7/7/2024| 35| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL| Data Analyst|Taking care of pe...| 6/10/2024| 8|39063746| Sedgwick| Sedgwick| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 5| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 32.77...| RGFsbGFzLCBUWA==| Dallas, TX| 48113| Dallas, TX|19100|Dallas-Fort Worth...| 48| Texas| 48113| Dallas, TX| 48113| Dallas, TX| 19100|Dallas-Fort Worth...| 19100|Dallas-Fort Worth...| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...|524291| Claims Adjusting|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS1218W78FG...|[\n "Management"...|[\n "ESF3939CE1F...| [\n "Exception R...|[\n "KS683TN76T7...|[\n "Security Cl...|[\n "KS1218W78FG...|[\n "Management"...|[\n "KS126HY6YLT...|[\n "Microsoft O...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...| 524291| Claims Adjusting|  
|1b5c3941e54a1889e...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024|7/20/2024| 48| [\n "Job Board"\n]|[\n "disabledper...|[\n "https://www...| []| NULL|Sr. Lead Data Mgm...|About this role:\...| 6/12/2024| 10|37615159| Wells Fargo|Wells Fargo| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.44...| UGhvZW5peCwgQVo=| Phoenix, AZ| 4013| Maricopa, AZ|38060|Phoenix-Mesa-Chan...| 4| Arizona| 4013| Maricopa, AZ| 4013| Maricopa, AZ| 38060|Phoenix-Mesa-Chan...| 38060|Phoenix-Mesa-Chan...| 52|Finance and Insur...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking|522110| Commercial Banking|ET2114E0404BA30075|Management Analysts|sr lead data mgmt...|[\n "KS123QX62QY...|[\n "Exit Strate...|[\n "KS123QX62QY...| [\n "Exit Strate...| []| []|[\n "KS7G6NP6R6L...|[\n "Reliability...|[\n "KS4409D76NW...|[\n "SAS (Softwa...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|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...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking| 522110| Commercial Banking|  
|cb5ca25f02bdf25c1...| 6/19/2024| 2024-06-19 07:00:00| 0|6/2/2024|6/17/2024| 15|[\n "FreeJobBoar...|[\n "craigslist....|[\n "https://mod...| []| NULL|Comisiones de $10...|Comisiones de $10...| 6/17/2024| 15| 0| Unclassified| LH/GM| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 3|Part-time / full-...| NULL| NULL| false| 92500| 0| [None]| year| 150000| 35000|{\n "lat": 37.63...| TW9kZXN0bywgQ0E=| Modesto, CA| 6099|Stanislaus, CA|33700| Modesto, CA| 6|California| 6099| Stanislaus, CA| 6099| Stanislaus, CA| 33700| Modesto, CA| 33700| Modesto, CA| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET0000000000000000| Unclassified|comisiones de por...| []| []| []| []| []| []| []| []| []| []|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
only showing top 5 rows

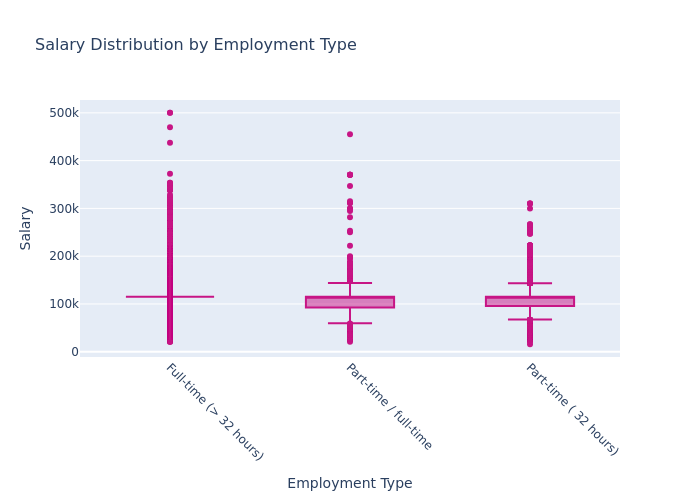
#Data Preparation and Cleaning  
  
df = df.withColumn("SALARY", df["SALARY"].cast("float"))\  
 .withColumn("SALARY\_TO", df["SALARY\_TO"].cast("float"))\  
 .withColumn("SALARY\_FROM", df["SALARY\_FROM"].cast("float"))\  
 .withColumn("MIN\_YEARS\_EXPERIENCE", df["MIN\_YEARS\_EXPERIENCE"].cast("float"))\  
 .withColumn("MAX\_YEARS\_EXPERIENCE", df["MAX\_YEARS\_EXPERIENCE"].cast("float"))  
   
#Calculate average and median for filling missing values  
median\_salary = df.approxQuantile("SALARY", [0.5], 0.01)[0]  
median\_to\_salary = df.approxQuantile("SALARY\_TO", [0.5], 0.01)[0]  
median\_from\_salary = df.approxQuantile("SALARY\_FROM", [0.5], 0.01)[0]  
  
df = df.fillna({"SALARY": median\_salary,  
 "SALARY\_TO": median\_to\_salary,  
 "SALARY\_FROM": median\_from\_salary,})  
  
df = df.withColumn("Average Salary", (df["SALARY\_TO"] + df["SALARY\_FROM"])/2)  
df = df.na.drop(subset=["EMPLOYMENT\_TYPE\_NAME"])  
  
export\_cols = [  
 "Average Salary", "SALARY", "SALARY\_TO", "SALARY\_FROM",  
 "MIN\_YEARS\_EXPERIENCE", "MAX\_YEARS\_EXPERIENCE",  
 "EMPLOYMENT\_TYPE\_NAME", "NAICS2\_NAME", "EDUCATION\_LEVELS\_NAME", "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME", "REMOTE\_TYPE\_NAME", "LOT\_OCCUPATION\_NAME"  
]  
  
df\_selected = df.select(export\_cols)  
pdf = df\_selected.toPandas()  
pdf.to\_csv("data/cleaned\_lightcast\_job\_postings.csv", index=False)  
  
#Clean employment type names & education level names  
pdf["EMPLOYMENT\_TYPE\_NAME"] = pdf["EMPLOYMENT\_TYPE\_NAME"].astype(str).apply(  
 lambda x: re.sub(r"[^\x00-\x7F]+", "", x)  
)  
pdf["EDUCATION\_LEVELS\_NAME"] = pdf["EDUCATION\_LEVELS\_NAME"].astype(str).str.replace(r"[\n\r\\\"\[\]]", "", regex=True)  
print(pdf.columns.tolist())

['Average Salary', 'SALARY', 'SALARY\_TO', 'SALARY\_FROM', 'MIN\_YEARS\_EXPERIENCE', 'MAX\_YEARS\_EXPERIENCE', 'EMPLOYMENT\_TYPE\_NAME', 'NAICS2\_NAME', 'EDUCATION\_LEVELS\_NAME', 'LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME', 'REMOTE\_TYPE\_NAME', 'LOT\_OCCUPATION\_NAME']

#Salary Distribution Employment Type  
fig = px.box(  
 pdf,  
 x="EMPLOYMENT\_TYPE\_NAME",  
 y="SALARY",  
 title="Salary Distribution by Employment Type",  
 color\_discrete\_sequence=["#C71585"],  
 points="outliers",  
)  
fig.update\_layout(  
 font\_family="Verdana",  
 title\_font\_size=16,  
 xaxis\_title="Employment Type",  
 yaxis\_title="Salary",  
 xaxis\_tickangle=45,  
)  
fig.show()  
fig.write\_html("Q1.html")  
fig.write\_image("Q1.png")

Unable to display output for mime type(s): text/html

In the Salary Distribution chart, we can see that there are significant outliers in full time roles. Most of the data is tightly clustered as the box is very small, meaning little variability. This contrasts with part time roles, where the spread is wider and there is more variability amongst the middle. There are less extreme outliers as well.

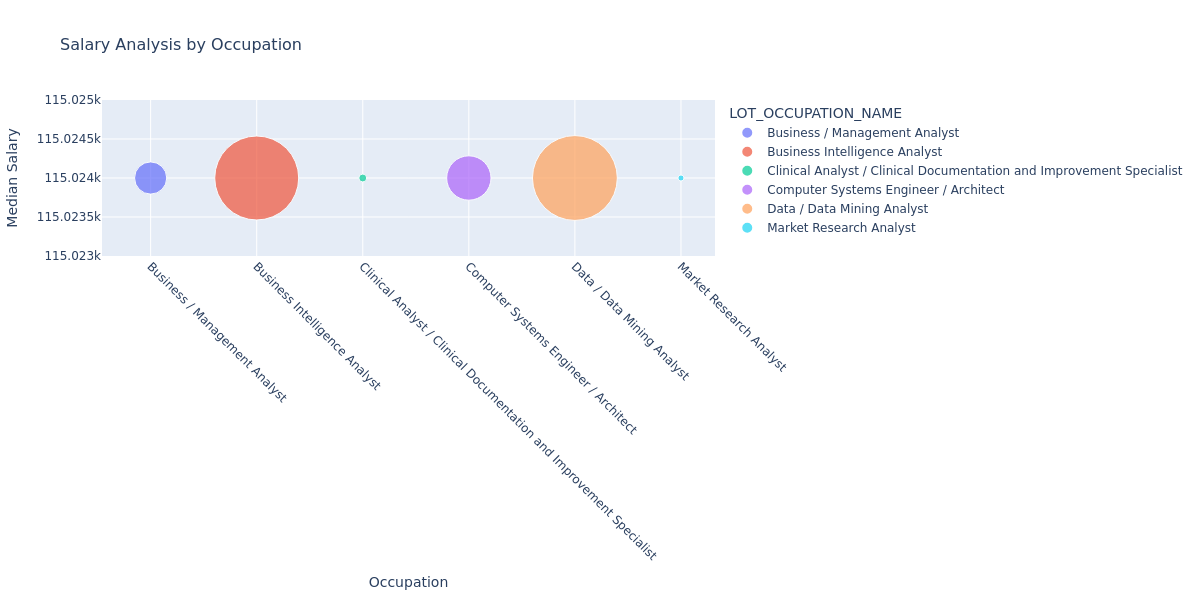


#Salary Distribution by Industry  
fig = px.box(  
 pdf,  
 x="NAICS2\_NAME",  
 y="SALARY",  
 title="Salary Distribution by Industry",  
 color\_discrete\_sequence=["#C71585"],  
 points="outliers",  
)  
fig.update\_layout(  
 height=1000,  
 font\_family="Verdana",  
 title\_font\_size=16,  
 xaxis\_title="Industry",  
 yaxis\_title="Salary",  
 xaxis\_tickangle=45,  
)  
fig.show()  
fig.write\_html("Q1b.html")  
fig.write\_image("Q1b.png", width=1200, height=800)

Further expanding upon salary variation, we can see when this is explored by industry there is variation is how tightly clustered salaries are together. More manual labor jobs, such as construction, transportation and warehousing, mining, etc, have less variability. These roles tend to have union representation which could lead to less variability. There are more extreme values in white collar jobs and healthcare.



#Salary Analysis by Occupation Type  
  
# Aggregate data  
salary\_analysis = pdf.groupby("LOT\_OCCUPATION\_NAME").agg(  
 median\_salary=("SALARY", "median"),  
 job\_count=("SALARY", "count")  
).reset\_index()  
  
fig = px.scatter(  
 salary\_analysis,  
 x="LOT\_OCCUPATION\_NAME",  
 y="median\_salary",  
 size="job\_count",   
 size\_max=60,   
 color="LOT\_OCCUPATION\_NAME",   
 color\_continuous\_scale="Plasma",  
 title="Salary Analysis by Occupation"  
)  
  
fig.update\_layout(  
 width=1200,  
 height=600,  
 font\_family="Verdana",  
 title\_font\_size=16,  
 xaxis\_title="Occupation",  
 yaxis\_title="Median Salary",  
 xaxis\_tickangle=45  
)  
  
fig.show()  
fig.write\_html("Q2.html")  
fig.write\_image("Q2.png", width=1200, height=600)

In this chart, we can see that there are similar median salary values, however there is variance in the number of job postings. There are more Business Intelligence Analyst or Data Mining Analyst roles. 

#Used this to preview the Education names for grouping later  
pdf.groupby("EDUCATION\_LEVELS\_NAME").size().reset\_index(name="job\_count").sort\_values(  
 by="job\_count", ascending=False  
)

|  | EDUCATION\_LEVELS\_NAME | job\_count |
| --- | --- | --- |
| 8 | Bachelor's degree | 29969 |
| 27 | No Education Listed | 22110 |
| 9 | Bachelor's degree, Master's degree | 10501 |
| 1 | Associate degree, Bachelor's degree | 1972 |
| 25 | Master's degree | 1623 |
| 12 | High school or GED | 1541 |
| 19 | High school or GED, Bachelor's degree | 1408 |
| 10 | Bachelor's degree, Master's degree, Ph.D. ... | 853 |
| 0 | Associate degree | 786 |
| 2 | Associate degree, Bachelor's degree, Maste... | 267 |
| 13 | High school or GED, Associate degree | 261 |
| 26 | Master's degree, Ph.D. or professional degree | 251 |
| 20 | High school or GED, Bachelor's degree, Mas... | 217 |
| 14 | High school or GED, Associate degree, Bach... | 200 |
| 11 | Bachelor's degree, Ph.D. or professional de... | 148 |
| 28 | Ph.D. or professional degree | 110 |
| 3 | Associate degree, Bachelor's degree, Maste... | 88 |
| 15 | High school or GED, Associate degree, Bach... | 54 |
| 5 | Associate degree, Master's degree | 23 |
| 21 | High school or GED, Bachelor's degree, Mas... | 21 |
| 23 | High school or GED, Master's degree | 15 |
| 22 | High school or GED, Bachelor's degree, Ph.... | 11 |
| 16 | High school or GED, Associate degree, Bach... | 11 |
| 18 | High school or GED, Associate degree, Mast... | 6 |
| 4 | Associate degree, Bachelor's degree, Ph.D.... | 2 |
| 24 | High school or GED, Master's degree, Ph.D.... | 2 |
| 7 | Associate degree, Ph.D. or professional degree | 2 |
| 6 | Associate degree, Master's degree, Ph.D. o... | 1 |
| 17 | High school or GED, Associate degree, Bach... | 1 |

print(pdf["EDUCATION\_LEVELS\_NAME"].unique())

[" Bachelor's degree" ' No Education Listed'  
 " Bachelor's degree, Master's degree"  
 " High school or GED, Associate degree, Bachelor's degree"  
 " Associate degree, Bachelor's degree, Master's degree"  
 " Associate degree, Bachelor's degree"  
 " High school or GED, Bachelor's degree, Master's degree, Ph.D. or professional degree"  
 ' High school or GED' " Master's degree"  
 " High school or GED, Bachelor's degree"  
 " Bachelor's degree, Master's degree, Ph.D. or professional degree"  
 ' Associate degree' ' Ph.D. or professional degree'  
 " High school or GED, Bachelor's degree, Master's degree"  
 ' High school or GED, Associate degree'  
 " High school or GED, Master's degree"  
 " Bachelor's degree, Ph.D. or professional degree"  
 " Master's degree, Ph.D. or professional degree"  
 " High school or GED, Bachelor's degree, Ph.D. or professional degree"  
 " High school or GED, Associate degree, Bachelor's degree, Master's degree"  
 " Associate degree, Master's degree"  
 " High school or GED, Associate degree, Master's degree"  
 " Associate degree, Bachelor's degree, Master's degree, Ph.D. or professional degree"  
 " Associate degree, Bachelor's degree, Ph.D. or professional degree"  
 " High school or GED, Associate degree, Bachelor's degree, Master's degree, Ph.D. or professional degree"  
 " High school or GED, Master's degree, Ph.D. or professional degree"  
 " High school or GED, Associate degree, Bachelor's degree, Ph.D. or professional degree"  
 ' Associate degree, Ph.D. or professional degree'  
 " Associate degree, Master's degree, Ph.D. or professional degree"]

#Salary by Education Level  
  
#Create copy of original df  
edu\_df = pdf.copy()  
  
#Clean up EDUCATION\_LEVELS\_NAME column  
edu\_df = edu\_df.assign(  
 EDUCATION\_LEVELS\_NAME=edu\_df["EDUCATION\_LEVELS\_NAME"].str.split(",")  
).explode("EDUCATION\_LEVELS\_NAME")  
edu\_df["EDUCATION\_LEVELS\_NAME"] = edu\_df["EDUCATION\_LEVELS\_NAME"].str.strip()  
edu\_df["EDUCATION\_LEVELS\_NAME"] = (  
 edu\_df["EDUCATION\_LEVELS\_NAME"]  
 .str.replace("’", "", regex=False)  
 .str.replace("'", "", regex=False)  
)  
  
# Map to groups  
edu\_map = {  
 "High school or GED": "Associate’s or lower",  
 "Associate degree": "Associate’s or lower",  
 "No Education Listed": "Associate’s or lower",  
 "Bachelors degree": "Bachelor’s",  
 "Masters degree": "Master’s",  
 "Ph.D. or professional degree": "PhD"  
}  
  
edu\_df["EDU\_GROUP"] = edu\_df["EDUCATION\_LEVELS\_NAME"].map(edu\_map)  
  
np.random.seed(42)  
edu\_df["MAX\_YEARS\_EXPERIENCE\_JITTER"] = (  
 edu\_df["MAX\_YEARS\_EXPERIENCE"] + np.random.uniform(-0.3, 0.3, len(edu\_df))  
)  
  
fig = px.scatter(  
 edu\_df,  
 x="MAX\_YEARS\_EXPERIENCE\_JITTER",  
 y="Average Salary",  
 color="EDU\_GROUP",  
 hover\_data=["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME", "MAX\_YEARS\_EXPERIENCE", "Average Salary"],  
 title="Salary vs Experience by Education Level",  
 color\_discrete\_sequence=px.colors.qualitative.Set2  
)  
  
fig.update\_layout(  
 xaxis\_title="Max Years Experience",  
 yaxis\_title="Average Salary",  
 width=1000,  
 height=600,  
 font\_family="Verdana",  
 title\_font\_size=16  
)  
  
fig.show()  
fig.write\_html("Q3.html")  
fig.write\_image("Q3.png", width=1000, height=600)

In this chart, many jobs are clustered around the lower half of years of experience. This can be explained by many of these roles being a newer development in the job market, relatively speaking. No matter the amount of experience, almost all of the jobs do not breach the 200k upper level for salary.

Commentary: I used Generative AI to use strip and explode, as well as considering different ways to clean up the data prior to mapping and creating the scatterplot. My prompt was what was the best way to denormalize the table in order to retain all values populated in the Education Levels field. Ultimately, my data clean up was from examining the unique data entries themselves and trial and error to determine that apostrophes were causing an issue for me.

