Assignment 03

Yibei Yu

2025-09-22

Table of contents

**GitHub Repository Link:**  
https://github.com/met-ad-688/assignment-03-yibei23-jpg

from pyspark.sql import SparkSession  
from pyspark.sql import functions as F  
from pyspark.sql.functions import col, count, expr  
from pyspark.sql.functions import col, when, regexp\_replace  
from pyspark.sql.functions import when, col, lower  
from pyspark.sql.functions import col, lower, when, coalesce, lit  
from pyspark.sql.functions import col, trim, length  
import pandas as pd  
import numpy as np  
import plotly.express as px  
import plotly.graph\_objects as go  
import plotly.io as pio  
import os  
  
spark = SparkSession.builder.appName("Assignment03").getOrCreate()  
  
df = (spark.read.option("header", "true")  
 .option("inferSchema", "true")  
 .option("multiLine", "true")  
 .option("escape", "\"")  
 .csv("data/lightcast\_job\_postings.csv"))  
  
df.createOrReplaceTempView("job\_postings")  
df.printSchema()  
df.show(5)

[Stage 56:> (0 + 1) / 1]

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|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

df = (df.withColumn("SALARY\_FROM", col("SALARY\_FROM").cast("double"))  
 .withColumn("SALARY\_TO", col("SALARY\_TO").cast("double"))  
 .withColumn("SALARY", col("SALARY").cast("double"))  
 .withColumn("MAX\_YEARS\_EXPERIENCE", col("MAX\_YEARS\_EXPERIENCE").cast("double")))  
  
df = df.withColumn(  
 "Average\_Salary",  
 when(col("SALARY").isNotNull(), col("SALARY"))  
 .otherwise((col("SALARY\_FROM")+col("SALARY\_TO"))/2.0)  
)  
  
df = df.withColumn("EDUCATION\_LEVELS\_NAME",  
 regexp\_replace(col("EDUCATION\_LEVELS\_NAME"), r"\n|\r", ""))  
  
df\_clean = df.select("Average\_Salary","SALARY\_FROM","SALARY\_TO","SALARY",  
 "NAICS\_2022\_2\_NAME","NAICS\_2022\_3\_NAME","NAICS\_2022\_4\_NAME",  
 "EMPLOYMENT\_TYPE\_NAME","ONET\_NAME","LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 "MAX\_YEARS\_EXPERIENCE","REMOTE\_TYPE\_NAME","EDUCATION\_LEVELS\_NAME").cache()  
  
print("Rows retained:", df\_clean.count())

Rows retained: 72498

25/09/23 03:03:49 WARN CacheManager: Asked to cache already cached data.

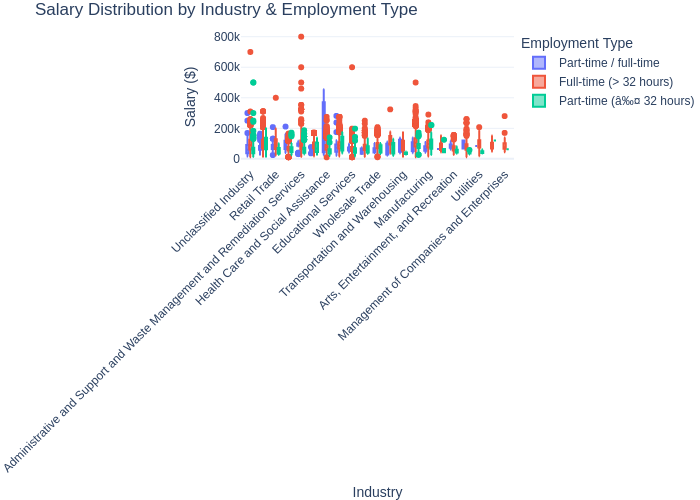
# 1. Salary Distribution by Industry and Employment Type

df\_filtered = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") > 0))  
salary\_industry = df\_filtered.select("NAICS2\_NAME", "SALARY\_FROM", "EMPLOYMENT\_TYPE\_NAME").toPandas()  
  
fig1 = px.box(  
 salary\_industry,  
 x="NAICS2\_NAME",  
 y="SALARY\_FROM",  
 color="EMPLOYMENT\_TYPE\_NAME",  
 title="Salary Distribution by Industry & Employment Type",  
 labels={"NAICS2\_NAME":"Industry", "SALARY\_FROM":"Salary ($)", "EMPLOYMENT\_TYPE\_NAME":"Employment Type"},  
)  
  
fig1.update\_layout(template="plotly\_white", xaxis\_tickangle=-45)  
fig1.write\_image("salary\_plot.png")  
fig1

[Stage 61:> (0 + 1) / 1]

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

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



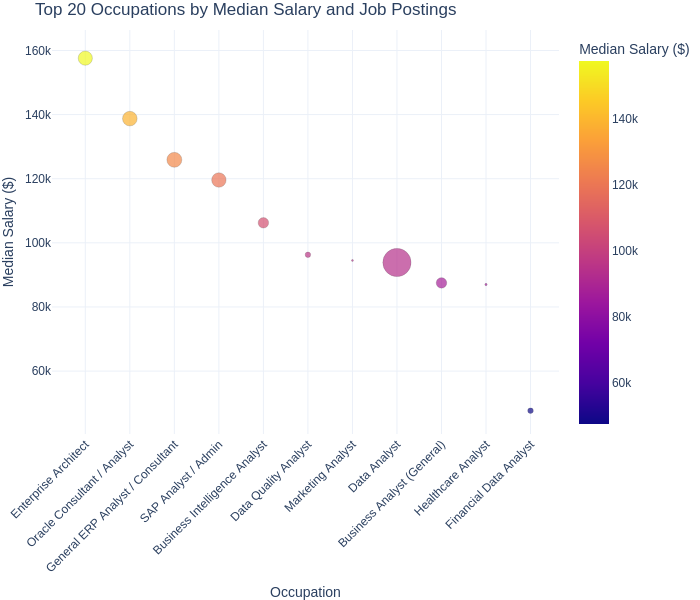
Salary Box Plot

**Analysis** The boxplots show significant differences in salary distribution across industries. Industries such as information technology, finance, and insurance exhibit a wider salary range and are characterized by more extreme outliers. The median salary for full-time positions is consistently higher than that for part-time positions. Some industries have a large number of low-paying outliers, indicating significant salary disparity within these industries.

# 2. Salary Analysis by ONET Occupation Type (Bubble Chart)

df\_occ = (  
 df\_clean  
 .withColumn("Occupation", trim(col("LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME")))  
 .filter(col("Occupation").isNotNull() & (length(col("Occupation")) > 0))  
 .filter(col("Average\_Salary").isNotNull() & (col("Average\_Salary") > 0) & (col("Average\_Salary") < 500000))  
)  
  
occ\_stats = df\_occ.groupBy("Occupation").agg(  
 F.expr("percentile\_approx(Average\_Salary, 0.5)").alias("Median\_Salary"),  
 F.count("\*").alias("Job\_Postings")  
)  
  
top20\_pdf = (  
 occ\_stats.orderBy(F.desc("Job\_Postings")).limit(20)  
 .toPandas()  
)  
  
top20\_pdf = top20\_pdf.sort\_values(["Median\_Salary", "Job\_Postings"], ascending=[False, False])  
  
fig2 = px.scatter(  
 top20\_pdf,  
 x="Occupation",  
 y="Median\_Salary",  
 size="Job\_Postings",  
 color="Median\_Salary",  
 color\_continuous\_scale="Plasma",  
 title="Top 20 Occupations by Median Salary and Job Postings",  
 labels={  
 "Occupation": "Occupation",  
 "Median\_Salary": "Median Salary ($)",  
 "Job\_Postings": "Number of Postings"  
 }  
)  
  
fig2.update\_layout(template="plotly\_white", xaxis\_tickangle=-45, height=600)  
fig2.update\_traces(marker=dict(line=dict(width=0.6, color="rgba(0,0,0,0.35)")))  
fig2.write\_image("occupation\_type.png")  
fig2

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



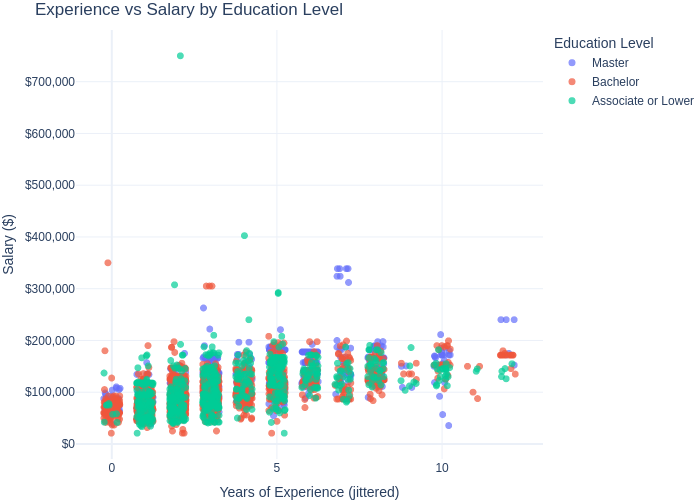
Top 20 Occupations by Median Salary and Job Postings

**Analysis** Salaries for specialized positions like Enterprise Architect and SAP/Oracle/ERP Consultant are at the highest range, around $120,000 to $160,000, but smaller bubbles indicate fewer job openings. In contrast, salaries for general analyst positions, particularly Data Analyst and Business Analyst (general), are concentrated around $85,000 to $100,000, with larger bubbles indicating greater demand. Therefore, higher levels of specialization lead to higher median salaries and lower hiring volume; while broader positions lead to lower salaries and higher hiring volume.

# 3. Salary by Education Level

df\_edu = (  
 df\_clean  
 .filter(  
 col("Average\_Salary").isNotNull() &  
 (col("Average\_Salary") > 0) &  
 col("MAX\_YEARS\_EXPERIENCE").isNotNull()  
 )  
 .withColumn(  
 "EDU\_GROUP",  
 when(col("EDUCATION\_LEVELS\_NAME").rlike("PhD|Doctorate"), "PhD")  
 .when(col("EDUCATION\_LEVELS\_NAME").rlike("Master"), "Master")  
 .when(col("EDUCATION\_LEVELS\_NAME").rlike("Bachelor"), "Bachelor")  
 .otherwise("Associate or Lower")  
 )  
)  
  
edu\_pd = (  
 df\_edu.select(  
 "MAX\_YEARS\_EXPERIENCE",  
 "Average\_Salary",  
 "EDU\_GROUP",  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 )  
 .toPandas()  
 .dropna(subset=["MAX\_YEARS\_EXPERIENCE", "Average\_Salary"])  
)  
  
np.random.seed(42) # reproducible jitter  
edu\_pd["exp\_jitter"] = (  
 edu\_pd["MAX\_YEARS\_EXPERIENCE"].astype(float) + np.random.uniform(-0.25, 0.25, len(edu\_pd))  
)  
  
fig3 = px.scatter(  
 edu\_pd,  
 x="exp\_jitter",  
 y="Average\_Salary",  
 color="EDU\_GROUP",  
 hover\_name="LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 title="Experience vs Salary by Education Level",  
 labels={  
 "exp\_jitter": "Years of Experience (jittered)",  
 "Average\_Salary": "Salary ($)",  
 "EDU\_GROUP": "Education Level",  
 },  
 template="plotly\_white",  
)  
fig3.update\_traces(marker=dict(size=7, opacity=0.7))  
fig3.update\_yaxes(tickformat="$,.0f")  
fig3.write\_image("EDUCATION\_LEVEL.png")  
fig3

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



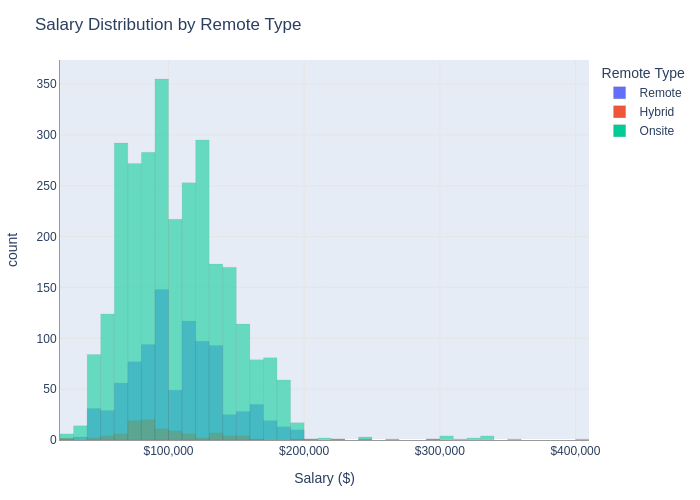
Experience vs Salary by Education Level

**Analysis** The scatter plot shows that salaries increase with experience, but there are significant differences across education levels. Salaries for those with bachelor’s and master’s degrees are more concentrated in the $50,000 to $200,000 range, while salaries for associate’s degrees and below are less outliers, but still reach extremely high levels. So while higher education provides consistency, experience and specialized positions can sometimes be more important than formal education in salary outcomes.

# 4. Salary by Remote Work Type

assert "df\_clean" in globals(), "df\_clean is not defined."  
  
possible\_remote\_cols = [  
 "REMOTE\_TYPE\_NAME", "REMOTE\_TYPE", "WORKPLACE\_TYPE", "WORK\_LOCATION",  
 "JOB\_WORK\_TYPE", "WORK\_MODEL", "REMOTE\_WORK\_TYPE"  
]  
remote\_cols = [c for c in possible\_remote\_cols if c in df\_clean.columns]  
  
remote\_text = (  
 lower(coalesce(\*[col(c) for c in remote\_cols])) # first existing col among the list  
 if remote\_cols else lit(None).cast("string")  
)  
  
df\_remote = (  
 df\_clean  
 .filter(  
 col("Average\_Salary").isNotNull() &  
 (col("Average\_Salary") > 0) &  
 (col("Average\_Salary") < 500\_000) & # trim extreme outliers for cleaner plots  
 col("MAX\_YEARS\_EXPERIENCE").isNotNull()  
 )  
 .withColumn("REMOTE\_SOURCE", remote\_text)  
 .withColumn(  
 "REMOTE\_GROUP",  
 # Put HYBRID FIRST so phrases like "hybrid remote" don't get captured by "remote" rule  
 when(remote\_text.rlike(r"hyb|part[- ]?remote|flex|split|mix|blended|some.\*remote|both.\*(office|onsite).\*remote"), "Hybrid")  
 .when(remote\_text.rlike(r"remote|work[- ]?from[- ]?home|wfh|tele[- ]?work|telecommute"), "Remote")  
 .when(remote\_text.rlike(r"on[- ]?site|in[- ]?person|on[- ]?prem|in[- ]?office|office"), "Onsite")  
 .otherwise("Onsite") # default bucket if nothing is stated  
 )  
)  
  
print("Remote group counts:")  
df\_remote.groupBy("REMOTE\_GROUP").count().orderBy(F.desc("count")).show()  
  
print("Top distinct values found in the remote text column (to see how it's written):")  
(df\_remote.groupBy("REMOTE\_SOURCE")  
 .count()  
 .orderBy(F.desc("count"))  
 .show(20, truncate=False))  
  
remote\_pd = (df\_remote  
 .select("MAX\_YEARS\_EXPERIENCE", "Average\_Salary",  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME", "REMOTE\_GROUP")  
 .toPandas()  
 .dropna(subset=["MAX\_YEARS\_EXPERIENCE", "Average\_Salary"]))  
  
np.random.seed(7)  
remote\_pd["exp\_jitter"] = remote\_pd["MAX\_YEARS\_EXPERIENCE"].astype(float) + np.random.uniform(-0.3, 0.3, len(remote\_pd))  
  
fig\_scatter = px.scatter(  
 remote\_pd,  
 x="exp\_jitter",  
 y="Average\_Salary",  
 color="REMOTE\_GROUP",  
 hover\_name="LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 title="Experience vs Salary by Remote Work Type",  
 labels={"exp\_jitter": "Years of Experience (jittered)", "Average\_Salary": "Salary ($)", "REMOTE\_GROUP": "Remote Type"},  
 category\_orders={"REMOTE\_GROUP": ["Remote", "Hybrid", "Onsite"]},  
)  
fig\_scatter.update\_traces(marker=dict(size=7, opacity=0.7, line=dict(width=0.5, color="rgba(0,0,0,0.3)")))  
fig\_scatter.update\_xaxes(showline=True, linewidth=1, linecolor="#999", gridcolor="#e6e6e6")  
fig\_scatter.update\_yaxes(showline=True, linewidth=1, linecolor="#999", gridcolor="#e6e6e6", tickformat="$,.0f")  
fig\_scatter.update\_layout(margin=dict(l=60, r=20, t=60, b=60), legend\_title\_text="Remote Type")  
fig\_scatter  
fig\_scatter.write\_image("Remote\_scatter.png")   
  
fig\_hist = px.histogram(  
 remote\_pd,  
 x="Average\_Salary",  
 color="REMOTE\_GROUP",  
 nbins=40,  
 barmode="overlay",  
 opacity=0.55,  
 title="Salary Distribution by Remote Type",  
 labels={"Average\_Salary": "Salary ($)", "REMOTE\_GROUP": "Remote Type"},  
 category\_orders={"REMOTE\_GROUP": ["Remote", "Hybrid", "Onsite"]},  
)  
fig\_hist.update\_traces(marker\_line\_width=0.5, marker\_line\_color="rgba(0,0,0,0.2)")  
fig\_hist.update\_xaxes(showline=True, linewidth=1, linecolor="#999", gridcolor="#e6e6e6", tickformat="$,.0f")  
fig\_hist.update\_yaxes(showline=True, linewidth=1, linecolor="#999", gridcolor="#e6e6e6")  
fig\_hist.update\_layout(margin=dict(l=60, r=20, t=60, b=60), legend\_title\_text="Remote Type")  
fig\_hist  
fig\_hist.write\_image("Remote\_hist.png")

Remote group counts:  
+------------+-----+  
|REMOTE\_GROUP|count|  
+------------+-----+  
| Onsite| 2909|  
| Remote| 929|  
| Hybrid| 99|  
+------------+-----+  
  
Top distinct values found in the remote text column (to see how it's written):  
+-------------+-----+  
|REMOTE\_SOURCE|count|  
+-------------+-----+  
|[none] |2909 |  
|remote |883 |  
|hybrid remote|99 |  
|not remote |46 |  
+-------------+-----+

**Analysis** Salary Distribution by Remote Type: Remote and on-site positions have very similar salary distributions, roughly clustered between $90,000 and $160,000. On-site positions have the most salary distribution and a wider range, while hybrid positions have much less.

Experience and Salary by Remote Job Type: Salary increases slightly with experience across all three job types. Salary ranges for remote/on-site positions roughly overlap, and there’s no significant salary premium for purely remote positions. Hybrid positions have less salary distribution, but when it does exist, it falls in the same range as the other two job types.