Assignment 03

Taylor Luckenbill

September 22, 2025

Table of contents

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(2)  
  
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").option("escape", "\"").csv("data/lightcast\_job\_postings.csv")  
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 14:> (0 + 1) / 1]

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| 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

# 1. Analysis Objective

## 1.1 Data Preparation Activities

In order to complete the analysis the first step is to load and prepare the data. For this analysis I will update data types, calculate average salary, remove extraneous characters from relevant data text, and replace null values.

For the data set job\_postings save as df and change the columns SALARY\_FROM, SALARY\_TO, SALARY, MIN\_YEARS\_EXPERIENCE, and MAX\_YEARS\_EXPERIENCE to type float.

df = (  
 df.withColumn("SALARY\_FROM", col("SALARY\_FROM").cast("float"))  
 .withColumn("SALARY\_TO", col("SALARY\_TO").cast("float"))  
 .withColumn("SALARY", col("SALARY").cast("float"))  
 .withColumn("MIN\_YEARS\_EXPERIENCE", col("MIN\_YEARS\_EXPERIENCE").cast("float"))  
 .withColumn("MAX\_YEARS\_EXPERIENCE", col("MAX\_YEARS\_EXPERIENCE").cast("float"))  
)

Next the median needs to be calculated as it will be used to fill na values. Create a new method called compute\_median.

def compute\_median(sdf,col\_name):  
 q = sdf.approxQuantile(col\_name,[0.5],0.01)  
 return q[0] if q else None  
  
median\_from = compute\_median(df,"SALARY\_FROM")  
median\_to = compute\_median(df, "SALARY\_TO")  
median\_salary = compute\_median(df, "SALARY")

[Stage 16:> (0 + 1) / 1] [Stage 17:> (0 + 1) / 1] [Stage 18:> (0 + 1) / 1]

print("medians: ", median\_from, median\_to, median\_salary)

medians: 87295.0 130042.0 115024.0

Next fill the missing values in the SALARY\_FROM, SALARY\_TO, SALARY with the median values.

df = df.fillna({  
 "SALARY\_FROM": median\_from,  
 "SALARY\_TO": median\_to,  
 "SALARY": median\_salary})

Calculate average salary by adding the min and max salaries and dividing by 2.

df = df.withColumn("Average Salary",(col("SALARY\_FROM")+col("SALARY\_TO"))/2)

print(df.columns)

['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', 'Average Salary']

Next remove extraneous charachters from education level names.

from pyspark.sql.functions import regexp\_replace, col  
  
df = df.withColumn(  
 "EDUCATION\_LEVELS\_NAME",  
 regexp\_replace(col("EDUCATION\_LEVELS\_NAME"), "[\n\r]", "")   
)

Select the columns to maintain in the analysis

#parse  
#|echo: true  
export\_cols = [  
 "EDUCATION\_LEVELS\_NAME",  
 "REMOTE\_TYPE\_NAME",  
 "MAX\_YEARS\_EXPERIENCE",  
 "Average Salary",  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"  
]

confirmed selected columns

df\_selected = df.select(export\_cols)  
df\_selected.show(3)

+---------------------+----------------+--------------------+--------------+----------------------------------+  
|EDUCATION\_LEVELS\_NAME|REMOTE\_TYPE\_NAME|MAX\_YEARS\_EXPERIENCE|Average Salary|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|  
+---------------------+----------------+--------------------+--------------+----------------------------------+  
| [ "Bachelor's de...| [None]| 2.0| 108668.5| General ERP Analy...|  
| [ "No Education ...| Remote| 3.0| 108668.5| Oracle Consultant...|  
| [ "Bachelor's de...| [None]| NULL| 108668.5| Data Analyst|  
+---------------------+----------------+--------------------+--------------+----------------------------------+  
only showing top 3 rows

pdf = df\_selected.toPandas()  
pdf.to\_csv("./data/lighthouse\_cleaned.csv", index=False)  
print(len(pdf))

[Stage 20:> (0 + 1) / 1]

72498

median\_salaries = pdf.groupby("LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME")["Average Salary"].median()

sorted\_employment\_types = median\_salaries.sort\_values(ascending=False).index

pdf["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"] = pd.Categorical(  
 pdf["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"],  
 categories=sorted\_employment\_types,  
 ordered=True  
)

fig = px.box(  
 pdf,  
 x="LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 y="Average Salary"  
)

fig.show()

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

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

print(df.columns)

['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', 'Average Salary']

from pyspark.sql.functions import lit  
  
df = df.withColumn("counter", lit(1))

#parse  
export\_cols2 = [  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 "Average Salary",  
 "counter"  
]

df\_selected2 = df.select(export\_cols2)  
df\_selected2.show(40)

+----------------------------------+--------------+-------+  
|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|Average Salary|counter|  
+----------------------------------+--------------+-------+  
| General ERP Analy...| 108668.5| 1|  
| Oracle Consultant...| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| Oracle Consultant...| 92500.0| 1|  
| Data Analyst| 110155.0| 1|  
| Data Analyst| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| General ERP Analy...| 108668.5| 1|  
| Data Analyst| 92962.0| 1|  
| Data Analyst| 107645.5| 1|  
| Data Analyst| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| General ERP Analy...| 192800.0| 1|  
| Enterprise Architect| 81286.0| 1|  
| Data Analyst| 108668.5| 1|  
| General ERP Analy...| 125900.0| 1|  
| Oracle Consultant...| 108668.5| 1|  
| Enterprise Architect| 165000.0| 1|  
| Data Analyst| 170000.0| 1|  
| Data Analyst| 110155.0| 1|  
| Enterprise Architect| 136950.0| 1|  
| Data Analyst| 118560.0| 1|  
| Enterprise Architect| 108668.5| 1|  
| Business Analyst ...| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| Enterprise Architect| 79000.0| 1|  
| SAP Analyst / Admin| 41600.0| 1|  
| Business Intellig...| 108668.5| 1|  
| Data Analyst| 108668.5| 1|  
| Data Analyst| 140756.5| 1|  
| General ERP Analy...| 192800.0| 1|  
| Oracle Consultant...| 75026.0| 1|  
| General ERP Analy...| 116500.0| 1|  
| Oracle Consultant...| 166500.0| 1|  
| Oracle Consultant...| 108668.5| 1|  
| Business Analyst ...| 108668.5| 1|  
| Data Analyst| 42500.0| 1|  
| Data Analyst| 156038.5| 1|  
| Data Analyst| 108668.5| 1|  
+----------------------------------+--------------+-------+  
only showing top 40 rows

pdf2 = df\_selected2.toPandas()  
#pdf2.to\_csv("./data/lighthouse\_cleaned.csv", index=False)  
#print(len(pdf2))  
pdf2.head(30)

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

|  | LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME | Average Salary | counter |
| --- | --- | --- | --- |
| 0 | General ERP Analyst / Consultant | 108668.5 | 1 |
| 1 | Oracle Consultant / Analyst | 108668.5 | 1 |
| 2 | Data Analyst | 108668.5 | 1 |
| 3 | Data Analyst | 108668.5 | 1 |
| 4 | Oracle Consultant / Analyst | 92500.0 | 1 |
| 5 | Data Analyst | 110155.0 | 1 |
| 6 | Data Analyst | 108668.5 | 1 |
| 7 | Data Analyst | 108668.5 | 1 |
| 8 | General ERP Analyst / Consultant | 108668.5 | 1 |
| 9 | Data Analyst | 92962.0 | 1 |
| 10 | Data Analyst | 107645.5 | 1 |
| 11 | Data Analyst | 108668.5 | 1 |
| 12 | Data Analyst | 108668.5 | 1 |
| 13 | General ERP Analyst / Consultant | 192800.0 | 1 |
| 14 | Enterprise Architect | 81286.0 | 1 |
| 15 | Data Analyst | 108668.5 | 1 |
| 16 | General ERP Analyst / Consultant | 125900.0 | 1 |
| 17 | Oracle Consultant / Analyst | 108668.5 | 1 |
| 18 | Enterprise Architect | 165000.0 | 1 |
| 19 | Data Analyst | 170000.0 | 1 |
| 20 | Data Analyst | 110155.0 | 1 |
| 21 | Enterprise Architect | 136950.0 | 1 |
| 22 | Data Analyst | 118560.0 | 1 |
| 23 | Enterprise Architect | 108668.5 | 1 |
| 24 | Business Analyst (General) | 108668.5 | 1 |
| 25 | Data Analyst | 108668.5 | 1 |
| 26 | Enterprise Architect | 79000.0 | 1 |
| 27 | SAP Analyst / Admin | 41600.0 | 1 |
| 28 | Business Intelligence Analyst | 108668.5 | 1 |
| 29 | Data Analyst | 108668.5 | 1 |

median\_salaries2 = pdf2.groupby("LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME").agg({  
 "Average Salary": "median",  
 "counter": "sum"  
}).reset\_index()

#sorted\_employment\_types2 = median\_salaries2.sort\_values(ascending=False).index  
median\_salaries2.head()

|  | LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME | Average Salary | counter |
| --- | --- | --- | --- |
| 0 | Business Analyst (General) | 108668.5 | 4326 |
| 1 | Business Intelligence Analyst | 108668.5 | 3639 |
| 2 | Data Analyst | 108668.5 | 27832 |
| 3 | Data Quality Analyst | 108668.5 | 1070 |
| 4 | Enterprise Architect | 108668.5 | 8212 |

fig2 = px.scatter(  
 median\_salaries2,  
 x="LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 y="Average Salary",  
 size="counter",  
 hover\_name="LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",   
 size\_max=60   
)  
  
fig2.update\_layout(  
 xaxis\_title="Occupation",  
 yaxis\_title="Median Salary",  
 title="Bubble Chart of Jobs: Median Salary vs. Occupation (Bubble = # Postings)",  
 xaxis\_tickangle=45 # rotate labels if long  
)

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

df = df.withColumn(  
 "education\_group",  
 F.when(F.col("MIN\_EDULEVELS\_NAME").isin("GED", "Associate", "No Education Listed"), "Associate's or lower")  
 .when(F.col("MIN\_EDULEVELS\_NAME") == "Bachelor's degree", "Bachelor's")  
 .when(F.col("MIN\_EDULEVELS\_NAME").isin("Master's degree"), "Master's")  
 .when(F.col("MIN\_EDULEVELS\_NAME").isin("PhD", "Doctorate", "professional degree"), "PhD")  
 .otherwise("Other") # optional catch-all for unexpected values  
)  
  
df.show()

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+---------+--------------------+--------------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+--------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+-------------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+--------------+-------+--------------------+  
| 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|Average Salary|counter| education\_group|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+---------+--------------------+--------------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+--------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+-------------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+--------------+-------+--------------------+  
|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]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 2.0| 2.0| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\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 ...| 108668.5| 1| Bachelor's|  
|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]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3.0| 3.0| false|115024.0| 1| Remote| NULL| 130042.0| 87295.0|{\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...| 108668.5| 1|Associate's or lower|  
|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]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 5.0| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\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| 108668.5| 1| Bachelor's|  
|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]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3.0| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\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| 108668.5| 1|Associate's or lower|  
|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]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 3|Part-time / full-...| NULL| NULL| false| 92500.0| 0| [None]| year| 150000.0| 35000.0|{\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...| 92500.0| 1|Associate's or lower|  
|35a6cd2183d9fb270...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024|6/12/2024| 10| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL|SR Lead Data Analyst|About Lumen\n\nLu...| 6/12/2024| 10| 2233642| Lumen Technologies| Lumen| false| [\n 2\n]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| NULL| NULL| false|110155.0| 1| Remote| year| 125890.0| 94420.0|{\n "lat": 0,\n ...|W1Vua25vd24gQ2l0e...|[Unknown City], AR| 5999|[Unknown county], AR| NULL| NULL| 5| Arkansas| 5999|[Unknown county], AR| 5999|[Unknown county], AR| NULL| NULL| NULL| NULL| 51| Information| 517| Telecommunications| 5178|All Other Telecom...| 51781|All Other Telecom...|517810|All Other Telecom...|ET95DB859B53CCACA7| Lead Data Analysts|sr lead data analyst|[\n "KS13USA80NE...|[\n "Power BI",\...|[\n "KS13USA80NE...| [\n "Power BI",\...| []| []|[\n "KS1280B68GD...|[\n "Presentatio...|[\n "KS13USA80NE...|[\n "Power BI",\...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| [\n "52.0201"\n]|[\n "Business Ad...| [\n "52.02"\n]|[\n "Business Ad...| [\n "52"\n]|[\n "Business, M...| 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| 51| Information| 517| Telecommunications| 5178|All Other Telecom...| 51781|All Other Telecom...| 517810|All Other Telecom...| 110155.0| 1| Bachelor's|  
|06de8d192f30b1d8d...| 8/2/2024| 2024-08-02 17:08:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Company"\n]|[\n "oraclecloud...|[\n "https://hct...| []| NULL| Talent Data Analyst|Id : 2501314,\nTi...| 6/22/2024| 20| 44896740|Semiconductor Com...|Semiconductor Com...| false| [\n 2\n]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| NULL| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\n "lat": 33.49...|U2NvdHRzZGFsZSwgQVo=| Scottsdale, AZ| 4013| Maricopa, AZ|38060|Phoenix-Mesa-Chan...| 4| Arizona| 4013| Maricopa, AZ| 4013| Maricopa, AZ| 38060|Phoenix-Mesa-Chan...| 38060|Phoenix-Mesa-Chan...| 31| Manufacturing| 334|Computer and Elec...| 3344|Semiconductor and...| 33441|Semiconductor and...|334413|Semiconductor and...|ETA9B609BE4E431E44| IT Data Analysts| talent data analyst|[\n "KS1250B78VW...|[\n "Interactive...|[\n "KS1250B78VW...| [\n "Interactive...| []| []|[\n "ESFA9982A2A...|[\n "Analytical ...|[\n "KS1250B78VW...|[\n "Interactive...|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| 31| Manufacturing| 334|Computer and Elec...| 3344|Semiconductor and...| 33441|Semiconductor and...| 334413|Semiconductor and...| 108668.5| 1| Bachelor's|  
|3d589c9d84677ca94...| 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]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 5.0| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\n "lat": 39.75...| RGF5dG9uLCBPSA==| Dayton, OH| 39113| Montgomery, OH|19430|Dayton-Kettering, OH| 39| Ohio| 39113| Montgomery, OH| 39113| Montgomery, OH| 19430|Dayton-Kettering, OH| 19430|Dayton-Kettering-...| 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| 108668.5| 1| Bachelor's|  
|5a843df632e1ff756...| 6/21/2024| 2024-06-21 07:00:00| 0|6/2/2024|6/20/2024| 18| [\n "Job Board"\n]|[\n "computerwor...|[\n "http://comp...| []| NULL|SAP SD/OTC Consul...|SAP SD/OTC Consul...| 6/20/2024| 18|100173263|Global Enterprise...|Global Enterprise...| true| [\n 99\n]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 7.0| 7.0| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\n "lat": 41.12...| RnJhbmtsaW4sIE5K| Franklin, NJ| 34037| Sussex, NJ|35620|New York-Newark-J...| 34| New Jersey| 34037| Sussex, NJ| 34037| Sussex, NJ| 35620|New York-Newark-J...| 35620|New York-Newark-J...| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET6244BCEEC5921581| SAP OTC Consultants|sap sd otc consul...|[\n "KS1200771D9...|[\n "JavaScript ...|[\n "KS1200771D9...| [\n "JavaScript ...| []| []| []| []|[\n "KS1200771D9...|[\n "JavaScript ...|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...| 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| NULL| NULL| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...| 108668.5| 1|Associate's or lower|  
|229620073766234e8...| 10/9/2024| 2024-10-09 18:07:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Company"\n]| [\n "3ds.com"\n]|[\n "https://www...| []| NULL|Sr. Marketing Ana...|Sr. Marketing Ana...| 8/1/2024| NULL| 39016169| Dassault SystÃ¨mes| Dassault Systmes| false| [\n 2,\n 3\n]| [ "Bachelor's de...| 2| Bachelor's degree| 3| Master's degree| 1|Full-time (> 32 h...| 2.0| 2.0| false| 92962.0| 0| [None]| year| 106424.0| 79500.0|{\n "lat": 40.75...| TmV3IFlvcmssIE5Z| New York, NY| 36061| New York, NY|35620|New York-Newark-J...| 36| New York| 36061| New York, NY| 36061| New York, NY| 35620|New York-Newark-J...| 35620|New York-Newark-J...| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...|541511|Custom Computer P...|ET1CE3CFA5447376E9| Marketing Analysts|sr marketing analyst|[\n "KS4407N6CMT...|[\n "Salesforce"...|[\n "KS4407N6CMT...| [\n "Salesforce"...| []| []|[\n "KS7G747655V...|[\n "Prioritizat...|[\n "KS4407N6CMT...|[\n "Salesforce"...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "52.0101",\n...|[\n "Business/Co...|[\n "52.01",\n ...|[\n "Business/Co...|[\n "52",\n "45...|[\n "Business, M...| 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 7\n]| [\n "Artificial ...| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...| 541511|Custom Computer P...| 92962.0| 1| Bachelor's|  
|b7aa80a24c82f080c...| 9/28/2024| 2024-09-28 14:06:...| 8|6/2/2024|9/27/2024| NULL|[\n "Government"...|[\n "dcscorp.com...|[\n "https://www...| []| NULL| Data Analyst|Data Analyst In R...| 7/13/2024| 41| 12147696| DCS Corporation| DCS Corp.| false|[\n 0,\n 1,\n ...| [ "High school o...| 0| High school or GED| 2| Bachelor's degree| 1|Full-time (> 32 h...| 10.0| NULL| false|107645.0| 2| Not Remote| year| 123732.0| 91559.0|{\n "lat": 35.62...|UmlkZ2VjcmVzdCwgQ0E=| Ridgecrest, CA| 6029| Kern, CA|12540| Bakersfield, CA| 6| California| 6029| Kern, CA| 6029| Kern, CA| 12540| Bakersfield, CA| 12540|Bakersfield-Delan...| 42| Wholesale Trade| 423|Merchant Wholesal...| 4238|Machinery, Equipm...| 42383|Industrial Machin...|423830|Industrial Machin...|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS128HD6KJS...|[\n "Regression ...|[\n "KS128HD6KJS...| [\n "Regression ...|[\n "KS683TN76T7...|[\n "Security Cl...|[\n "KS1203C6N9B...|[\n "Research",\...|[\n "KS125LS6N7W...|[\n "Python (Pro...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "14.0101",\n...|[\n "Engineering...|[\n "14.01",\n ...|[\n "Engineering...|[\n "14",\n "14...|[\n "Engineering...| 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| 42| Wholesale Trade| 423|Merchant Wholesal...| 4238|Machinery, Equipm...| 42383|Industrial Machin...| 423830|Industrial Machin...| 107645.5| 1| Other|  
|2a107fd40bb1afac4...| 6/17/2024| 2024-06-17 07:00:00| 0|6/2/2024| 6/8/2024| 6| [\n "Job Board"\n]| [\n "dice.com"\n]|[\n "https://www...| []| NULL| Data Analyst|Data Analyst\nTEK...| 6/8/2024| 6| 4063994| Allegis Group|TEKsystems c/o Al...| true| [\n 99\n]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 2.0| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\n "lat": 21.30...| SG9ub2x1bHUsIEhJ| Honolulu, HI| 15003| Honolulu, HI|46520| Urban Honolulu, HI| 15| Hawaii| 15003| Honolulu, HI| 15003| Honolulu, HI| 46520| Urban Honolulu, HI| 46520| Urban Honolulu, HI| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...|561320|Temporary Help Se...|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS7LO8P3MXB...|[\n "Data Scienc...|[\n "KS7LO8P3MXB...| [\n "Data Scienc...| []| []|[\n "KS122556LMQ...|[\n "Communicati...|[\n "KS440W865GC...|[\n "SQL (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "11.0701",\n...|[\n "Computer Sc...|[\n "11.07",\n ...|[\n "Computer Sc...|[\n "11",\n "30...|[\n "Computer an...| 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| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...| 561320|Temporary Help Se...| 108668.5| 1|Associate's or lower|  
|fd48c3ce533c3d20a...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 7/5/2024| 33| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL|Data Research Ana...|The Data Research...| 7/5/2024| 33| 34294036| Equifax| Equifax, Inc.| false| [\n 2\n]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| NULL| NULL| false|115024.0| 0| [None]| NULL| 130042.0| 87295.0|{\n "lat": 0,\n ...|W1Vua25vd24gQ2l0e...|[Unknown City], GA| 13999|[Unknown county], GA| NULL| NULL| 13| Georgia| 13999|[Unknown county], GA| 13999|[Unknown county], GA| NULL| NULL| NULL| NULL| 52|Finance and Insur...| 522|Credit Intermedia...| 5223|Activities Relate...| 52232|Financial Transac...|522320|Financial Transac...|ET252B42EF548117CC| Data Researchers|data research ana...|[\n "KS120GV6C72...|[\n "Data Analys...|[\n "KS120GV6C72...| [\n "Data Analys...| []| []|[\n "KS1203C6N9B...|[\n "Research",\...| []| []|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...| 522|Credit Intermedia...| 5223|Activities Relate...| 52232|Financial Transac...| 522320|Financial Transac...| 108668.5| 1| Bachelor's|  
|57b527ea0f91db5bb...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024|7/27/2024| 55| [\n "Job Board"\n]|[\n "simplyhired...|[\n "https://www...| []| NULL|Power, Utilities ...|Power, Utilities ...| 7/27/2024| 55| 5732448| Deloitte| Deloitte| false| [\n 2,\n 3\n]| [ "Bachelor's de...| 2| Bachelor's degree| 3| Master's degree| 1|Full-time (> 32 h...| 6.0| NULL| false|192800.0| 0| [None]| year| 241000.0| 144600.0|{\n "lat": 42.33...| RGV0cm9pdCwgTUk=| Detroit, MI| 26163| Wayne, MI|19820|Detroit-Warren-De...| 26| Michigan| 26163| Wayne, MI| 26163| Wayne, MI| 19820|Detroit-Warren-De...| 19820|Detroit-Warren-De...| 54|Professional, Sci...| 541|Professional, Sci...| 5416|Management, Scien...| 54161|Management Consul...|541611|Administrative Ma...|ET8AEDEB1F4C3091D3|Management Consul...|power utilities r...|[\n "KS122VL71WF...|[\n "Design Spec...|[\n "KS122VL71WF...| [\n "Design Spec...| []| []|[\n "KS1218W78FG...|[\n "Management"...|[\n "KS1219W70LY...|[\n "C++ (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| [\n "45.0702"\n]|[\n "Geographic ...| [\n "45.07"\n]|[\n "Geography a...| [\n "45"\n]|[\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 3\n]| [\n "Green Jobs:...| 54|Professional, Sci...| 541|Professional, Sci...| 5416|Management, Scien...| 54161|Management Consul...| 541611|Administrative Ma...| 192800.0| 1| Bachelor's|  
|036cd733481fbcc98...| 8/2/2024| 2024-08-02 17:08:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Job Board"\n]| [\n "ms.gov"\n]|[\n "https://win...| []| NULL|Sr. Enterprise Da...|Sr. Enterprise Da...| 6/14/2024| 12| 38205299|Lincoln Financial...|Lincoln Financial...| false| [\n 99\n]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| NULL| NULL| false| 81286.0| 1| Remote| year| 81286.0| 81286.0|{\n "lat": 32.29...| SmFja3NvbiwgTVM=| Jackson, MS| 28049| Hinds, MS|27140| Jackson, MS| 28| Mississippi| 28049| Hinds, MS| 28049| Hinds, MS| 27140| Jackson, MS| 27140| Jackson, MS| 52|Finance and Insur...| 523|Securities, Commo...| 5239|Other Financial I...| 52394|Portfolio Managem...|523940|Portfolio Managem...|ET0000000000000000| Unclassified|sr enterprise dat...|[\n "KS122NM6B8T...|[\n "Data Archit...|[\n "KS122NM6B8T...| [\n "Data Archit...|[\n "ESE495A4017...|[\n "Valid Drive...| []| []| []| []|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...| 231510|Computer Systems ...| 23151012| Enterprise Architect| 2315| Network and Syste...| 23151012| Enterprise Architect| 231510| Computer Systems ...| 2315| Network and Syste...| 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...| 523|Securities, Commo...| 5239|Other Financial I...| 52394|Portfolio Managem...| 523940|Portfolio Managem...| 81286.0| 1|Associate's or lower|  
|138ce2c9453b47a9b...| 8/10/2024| 2024-08-10 19:36:...| 5|6/2/2024| 8/9/2024| NULL|[\n "Job Board",...|[\n "silkroad.co...|[\n "https://mai...| []| NULL|SENIOR CONSULTANT...|SENIOR CONSULTANT...| 6/8/2024| 6| 1967| Boston University| Boston University| false|[\n 1,\n 2,\n ...| [ "Associate deg...| 1| Associate degree| 3| Master's degree| 1|Full-time (> 32 h...| 5.0| 5.0| false|115024.0| 1| Remote| NULL| 130042.0| 87295.0|{\n "lat": 42.36...| Qm9zdG9uLCBNQQ==| Boston, MA| 25025| Suffolk, MA|14460|Boston-Cambridge-...| 25|Massachusetts| 25025| Suffolk, MA| 25025| Suffolk, MA| 14460|Boston-Cambridge-...| 14460|Boston-Cambridge-...| 61|Educational Services| 611|Educational Services| 6113|Colleges, Univers...| 61131|Colleges, Univers...|611310|Colleges, Univers...|ET210B837B93B7B3F9|Continuous Improv...|senior consultant...|[\n "ESB38820A54...|[\n "Effective C...|[\n "ESB38820A54...| [\n "Effective C...|[\n "KS7G2ZG794H...|[\n "Certified I...|[\n "KS1280B68GD...|[\n "Presentatio...| []| []|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| [\n "52.0201"\n]|[\n "Business Ad...| [\n "52.02"\n]|[\n "Business Ad...| [\n "52"\n]|[\n "Business, M...| 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| 61|Educational Services| 611|Educational Services| 6113|Colleges, Univers...| 61131|Colleges, Univers...| 611310|Colleges, Univers...| 108668.5| 1| Other|  
|dd191e2ce3062c371...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024|6/20/2024| 18| [\n "Job Board"\n]|[\n "phoenixrecr...|[\n "https://www...| []| NULL| SAP FSCM Consultant|Job Description: ...| 6/20/2024| 18| 8592955| Accenture| Accenture| false| [\n 1,\n 2\n]| [ "Associate deg...| 1| Associate degree| 2| Bachelor's degree| 1|Full-time (> 32 h...| 12.0| NULL| false|125900.0| 0| [None]| year| 188600.0| 63200.0|{\n "lat": 0,\n ...|W1Vua25vd24gQ2l0e...|[Unknown City], AZ| 4999|[Unknown county], AZ| NULL| NULL| 4| Arizona| 4999|[Unknown county], AZ| 4999|[Unknown county], AZ| NULL| NULL| NULL| NULL| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...|541512|Computer Systems ...|ETF594A2C05D212506|Peoplesoft FSCM C...| sap fscm consultant|[\n "KS7G7VL78R2...|[\n "Profit Cent...|[\n "KS7G7VL78R2...| [\n "Profit Cent...| []| []|[\n "KS122ZF75YV...|[\n "Digitizatio...|[\n "KS7G7VL78R2...|[\n "Profit Cent...|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...| 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| NULL| NULL| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...| 541512|Computer Systems ...| 125900.0| 1| Other|  
|99856b5a8a1c75d90...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 8/1/2024| NULL|[\n "Government"\n]|[\n "alaska.gov"\n]|[\n "https://ala...| []| NULL|Oracle Consultant...|Onsite - Work ons...| 7/10/2024| 38| 133098|Smx Corporation L...| SMX| true| [\n 99\n]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3.0| 3.0| false|115024.0| 1| Remote| NULL| 130042.0| 87295.0|{\n "lat": 58.30...| SnVuZWF1LCBBSw==| Juneau, AK| 2110| Juneau Borough, AK|27940| Juneau, AK| 2| Alaska| 2110| Juneau Borough, AK| 2110| Juneau Borough, AK| 27940| Juneau, AK| 27940| Juneau, AK| 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...| 108668.5| 1|Associate's or lower|  
|f28123528a32b8c9b...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Company"\n]|[\n "sca.health"\n]|[\n "https://car...| []| NULL| Principal Architect|Principal Archite...| 8/1/2024| NULL| 39192167|Surgical Care Aff...|Surgical Care Aff...| false| [\n 2\n]| [ "Bachelor's de...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 8.0| 8.0| false|115024.0| 0| [None]| year| 170000.0| 160000.0|{\n "lat": 33.51...|QmlybWluZ2hhbSwgQUw=| Birmingham, AL| 1073| Jefferson, AL|13820|Birmingham-Hoover...| 1| Alabama| 1073| Jefferson, AL| 1073| Jefferson, AL| 13820|Birmingham-Hoover...| 13820| Birmingham, AL| 62|Health Care and S...| 621|Ambulatory Health...| 6214|Outpatient Care C...| 62149|Other Outpatient ...|621493|Freestanding Ambu...|ET7767EEDBF263F7B7|Principal Architects| principal architect|[\n "ES99B020D66...|[\n "Business Ob...|[\n "ES4B99FD0FD...| [\n "Infrastruct...|[\n "KS125K065BR...|[\n "Juniper Net...|[\n "ES99B020D66...|[\n "Business Ob...|[\n "KS120V86MZW...|[\n "Microsoft A...|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...| 231510|Computer Systems ...| 23151012| Enterprise Architect| 2315| Network and Syste...| 23151012| Enterprise Architect| 231510| Computer Systems ...| 2315| Network and Syste...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 5\n]| [\n "Cybersecuri...| 62|Health Care and S...| 621|Ambulatory Health...| 6214|Outpatient Care C...| 62149|Other Outpatient ...| 621493|Freestanding Ambu...| 165000.0| 1| Bachelor's|  
|b4e618e8d2a2b6744...| 10/9/2024| 2024-10-09 18:07:...| 2|6/2/2024|8/11/2024| NULL| [\n "Job Board"\n]|[\n "castrovalle...|[\n "https://www...| []| NULL|Principal growth ...|Principal growth ...| 7/27/2024| 55| 40794223|Aircall Internati...| Aircall| false| [\n 99\n]| [ "No Education ...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 6.0| NULL| false|170000.0| 0| [None]| year| 220000.0| 120000.0|{\n "lat": 37.77...|U2FuIEZyYW5jaXNjb...| San Francisco, CA| 6075| San Francisco, CA|41860|San Francisco-Oak...| 6| California| 6075| San Francisco, CA| 6075| San Francisco, CA| 41860|San Francisco-Oak...| 41860|San Francisco-Oak...| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET54F46C4290228B21| Growth Analysts|principal growth ...|[\n "ESA420F05EB...|[\n "Curiosity",...|[\n "KS1218H6QYL...| [\n "Business Co...| []| []|[\n "ESA420F05EB...|[\n "Curiosity",...|[\n "KS1200364C9...|[\n "C (Programm...|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...| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...| 170000.0| 1|Associate's or lower|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+---------+--------------------+--------------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+--------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+-------------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+--------------+-------+--------------------+  
only showing top 20 rows

#parse  
export\_cols3 = [  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 "Average Salary",  
 "MAX\_YEARS\_EXPERIENCE"  
]

df\_selected3 = df.select(export\_cols3)  
df\_selected3.show(10)

+----------------------------------+--------------+--------------------+  
|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|Average Salary|MAX\_YEARS\_EXPERIENCE|  
+----------------------------------+--------------+--------------------+  
| General ERP Analy...| 108668.5| 2.0|  
| Oracle Consultant...| 108668.5| 3.0|  
| Data Analyst| 108668.5| NULL|  
| Data Analyst| 108668.5| NULL|  
| Oracle Consultant...| 92500.0| NULL|  
| Data Analyst| 110155.0| NULL|  
| Data Analyst| 108668.5| NULL|  
| Data Analyst| 108668.5| NULL|  
| General ERP Analy...| 108668.5| 7.0|  
| Data Analyst| 92962.0| 2.0|  
+----------------------------------+--------------+--------------------+  
only showing top 10 rows

pdf3 = df\_selected3.toPandas()  
#pdf2.to\_csv("./data/lighthouse\_cleaned.csv", index=False)  
#print(len(pdf2))  
pdf3.head(10)

[Stage 25:> (0 + 1) / 1]

|  | LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME | Average Salary | MAX\_YEARS\_EXPERIENCE |
| --- | --- | --- | --- |
| 0 | General ERP Analyst / Consultant | 108668.5 | 2.0 |
| 1 | Oracle Consultant / Analyst | 108668.5 | 3.0 |
| 2 | Data Analyst | 108668.5 | NaN |
| 3 | Data Analyst | 108668.5 | NaN |
| 4 | Oracle Consultant / Analyst | 92500.0 | NaN |
| 5 | Data Analyst | 110155.0 | NaN |
| 6 | Data Analyst | 108668.5 | NaN |
| 7 | Data Analyst | 108668.5 | NaN |
| 8 | General ERP Analyst / Consultant | 108668.5 | 7.0 |
| 9 | Data Analyst | 92962.0 | 2.0 |

pdf3["MAX\_YEARS\_EXPERIENCE"] = pdf3["MAX\_YEARS\_EXPERIENCE"].fillna(0)  
pdf3.head(10)

|  | LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME | Average Salary | MAX\_YEARS\_EXPERIENCE |
| --- | --- | --- | --- |
| 0 | General ERP Analyst / Consultant | 108668.5 | 2.0 |
| 1 | Oracle Consultant / Analyst | 108668.5 | 3.0 |
| 2 | Data Analyst | 108668.5 | 0.0 |
| 3 | Data Analyst | 108668.5 | 0.0 |
| 4 | Oracle Consultant / Analyst | 92500.0 | 0.0 |
| 5 | Data Analyst | 110155.0 | 0.0 |
| 6 | Data Analyst | 108668.5 | 0.0 |
| 7 | Data Analyst | 108668.5 | 0.0 |
| 8 | General ERP Analyst / Consultant | 108668.5 | 7.0 |
| 9 | Data Analyst | 92962.0 | 2.0 |

import matplotlib.pyplot as plt

groups = pdf3["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"].unique()  
  
plt.figure(figsize=(10,6))  
  
for g in groups:  
 subset = pdf3[pdf3["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"] == g]  
   
 # Add jitter to avoid overlapping points  
 x\_jitter = subset["MAX\_YEARS\_EXPERIENCE"] + np.random.normal(0, 0.2, size=len(subset))  
 y\_jitter = subset["Average Salary"] + np.random.normal(0, 0.2, size=len(subset))  
   
 plt.scatter(x\_jitter, y\_jitter, alpha=0.6, label=g)  
  
plt.xlabel("X-axis (with jitter)")  
plt.ylabel("Y-axis (with jitter)")  
plt.title("Scatter Plots per Group with Jitter")  
plt.legend()  
plt.show()

