Module 04: Lab 01

Visual Reporting and Storytelling

Yixuan Yang

November 21, 2024

# Objectives

By the end of this lab, you will: 1. Load and analyze the **Lightcast dataset** in **Spark DataFrame**. 2. Create **five easy and three medium-complexity visualizations** using **Plotly**. 3. Explore **salary distributions, employment trends, and job postings**. 4. Analyze **skills in relation to NAICS/SOC/ONET codes and salaries**. 5. Customize **colors, fonts, and styles** in all visualizations (**default themes result in a 2.5-point deduction**). 6. Follow **best practices for reporting on data communication**.

# Step 1: Load the Dataset

import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
pio.renderers.default = "vscode"  
from pyspark.sql import SparkSession  
from pyspark.sql.functions import col ,explode, split, from\_json  
import plotly.graph\_objects as go  
from pyspark.sql.types import ArrayType, StringType  
  
  
# Initialize Spark Session  
spark = SparkSession.builder.appName("LightcastData").getOrCreate()  
  
# Load Data  
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine","true").option("escape", "\"").csv("./data/lightcast\_job\_postings.csv")  
  
# Show Schema and Sample Data  
df.printSchema()  
df.show(5)

Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/03/25 00:30:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
 25/03/25 00:30:54 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

root  
 |-- ID: string (nullable = true)  
 |-- LAST\_UPDATED\_DATE: date (nullable = true)  
 |-- LAST\_UPDATED\_TIMESTAMP: timestamp (nullable = true)  
 |-- DUPLICATES: integer (nullable = true)  
 |-- POSTED: date (nullable = true)  
 |-- EXPIRED: date (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: date (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...| 2024-09-06| 2024-09-06 20:32:...| 0|2024-06-02|2024-06-08| 6| [\n "Company"\n]|[\n "brassring.c...|[\n "https://sjo...| []| NULL|Enterprise Analys...|31-May-2024\n\nEn...| 2024-06-08| 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...| 2024-08-02| 2024-08-02 17:08:...| 0|2024-06-02|2024-08-01| NULL| [\n "Job Board"\n]| [\n "maine.gov"\n]|[\n "https://job...| []| NULL|Oracle Consultant...|Oracle Consultant...| 2024-08-01| 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...| 2024-09-06| 2024-09-06 20:32:...| 1|2024-06-02|2024-07-07| 35| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL| Data Analyst|Taking care of pe...| 2024-06-10| 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...| 2024-09-06| 2024-09-06 20:32:...| 1|2024-06-02|2024-07-20| 48| [\n "Job Board"\n]|[\n "disabledper...|[\n "https://www...| []| NULL|Sr. Lead Data Mgm...|About this role:\...| 2024-06-12| 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...| 2024-06-19| 2024-06-19 07:00:00| 0|2024-06-02|2024-06-17| 15|[\n "FreeJobBoar...|[\n "craigslist....|[\n "https://mod...| []| NULL|Comisiones de $10...|Comisiones de $10...| 2024-06-17| 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. Salary Distribution by Employment Type

* Identify salary trends across different employment types.
* **Filter the dataset**
  + Remove records where **salary is missing or zero**.
* **Aggregate Data**
  + Group by **employment type** and compute salary distribution.
* **Visualize results**
  + Create a **box plot** where:
    - **X-axis** = EMPLOYMENT\_TYPE\_NAME
    - **Y-axis** = SALARY\_FROM
  + Customize **colors, fonts, and styles** to avoid a **2.5-point deduction**.
* **Explanation:** Write two sentences about what the graph reveals.

# Your Code for 1st question here  
pdf = df.select("EMPLOYMENT\_TYPE\_NAME", "SALARY").toPandas()  
fig = px.box(pdf, x="EMPLOYMENT\_TYPE\_NAME", y="SALARY", title="Salary Distribution by Employment Type", color\_discrete\_sequence=["#636EFA"])  
fig.update\_layout(font\_family="Arial", title\_font\_size=16)  
fig.show()  
fig.write\_image("output/q1\_salary\_by\_employment.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Salary Distribution by Employment Type

Full time employees have higher medians, and the data has little variation meaning the full-time salary structure is stable.

# 2. Salary Distribution by Industry

* Compare salary variations across industries.
* **Filter the dataset**
  + Keep records where **salary is greater than zero**.
* **Aggregate Data**
  + Group by **NAICS industry codes**.
* **Visualize results**
  + Create a **box plot** where:
    - **X-axis** = NAICS2\_NAME
    - **Y-axis** = SALARY\_FROM
  + Customize colors, fonts, and styles.
* **Explanation:** Write two sentences about what the graph reveals.

df\_q2 = df.filter(col("SALARY\_FROM") > 0)  
pdf\_q2 = df\_q2.select("NAICS2\_NAME", "SALARY\_FROM").toPandas()  
  
fig = px.box(pdf\_q2, x="NAICS2\_NAME", y="SALARY\_FROM",  
 title="Salary Distribution by Industry",  
 color\_discrete\_sequence=["#1f77b4"])  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20,  
 xaxis=dict(tickangle=45),  
 yaxis\_title="Salary (USD)",  
 xaxis\_title="Industry"  
)  
fig.show()  
fig.write\_image("output/q2\_salary\_by\_industry.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Salary Distribution by Industry

Technology-related industries tend to offer higher median salaries, driven by demand for skilled professionals. Retail and service industries show lower, more uniform salaries, reflecting standardized wage structures.

# 3. Job Posting Trends Over Time

* Analyze how job postings fluctuate over time.
* **Aggregate Data**
  + Count job postings per **posted date (POSTED)**.
* **Visualize results**
  + Create a **line chart** where:
    - **X-axis** = POSTED
    - **Y-axis** = Number of Job Postings
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

df\_q3 = df.groupBy("POSTED").count().orderBy("POSTED")  
pdf\_q3 = df\_q3.toPandas()  
  
fig = px.line(pdf\_q3, x="POSTED", y="count",  
 title="Job Posting Trends Over Time",  
 line\_shape="linear", color\_discrete\_sequence=["#ff7f0e"])  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20,  
 yaxis\_title="Number of Postings",  
 xaxis\_title="Date"  
)  
fig.show()  
fig.write\_image("output/q3\_job\_trends.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Job Posting Trends Over Time

Job postings exhibit seasonal peaks, often around hiring cycles like the early of the year or post-summer.

# 4. Top 10 Job Titles by Count

* Identify the most frequently posted job titles.
* **Aggregate Data**
  + Count the occurrences of each **job title (TITLE\_NAME)**.
  + Select the **top 10 most frequent titles**.
* **Visualize results**
  + Create a **bar chart** where:
    - **X-axis** = TITLE\_NAME
    - **Y-axis** = Job Count
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

df\_q4 = df.groupBy("TITLE\_NAME").count().orderBy(col("count").desc()).limit(10)  
pdf\_q4 = df\_q4.toPandas()  
  
fig = px.bar(pdf\_q4, x="TITLE\_NAME", y="count",  
 title="Top 10 Job Titles by Count",  
 color\_discrete\_sequence=["#2ca02c"])  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20,  
 yaxis\_title="Number of Postings",  
 xaxis\_title="Job Title",  
 xaxis=dict(tickangle=45)  
)  
fig.show()  
fig.write\_image("output/q4\_top\_titles.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Top 10 Job Titles by Count

High-demand roles like Data Analyst dominate postings, reflecting tech sector growth. Analystic skills seems to be the most needed in the job market.

# 5. Remote vs On-Site Job Postings

* Compare the proportion of remote and on-site job postings.
* **Aggregate Data**
  + Count job postings by **remote type (REMOTE\_TYPE\_NAME)**.
* **Visualize results**
  + Create a **pie chart** where:
    - **Labels** = REMOTE\_TYPE\_NAME
    - **Values** = Job Count
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

df\_q5 = df.groupBy("REMOTE\_TYPE\_NAME").count()  
pdf\_q5 = df\_q5.toPandas()  
  
fig = px.pie(pdf\_q5, names="REMOTE\_TYPE\_NAME", values="count",  
 title="Remote vs On-Site Job Postings",  
 color\_discrete\_sequence=px.colors.qualitative.Pastel)  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20  
)  
fig.show()  
fig.write\_image("output/q5\_remote\_pie.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Remote vs On-Site Job Postings

Remote job postings are high, with the growing Hybrid Remote roles, it is reflecting shifts toward flexible work arrangements.

# 6. Skill Demand Analysis by Industry (Stacked Bar Chart)

* Identify which skills are most in demand in various industries.
* **Aggregate Data**
  + Extract **skills** from job postings.
  + Count occurrences of skills grouped by **NAICS industry codes**.
* **Visualize results**
  + Create a **stacked bar chart** where:
    - **X-axis** = Industry
    - **Y-axis** = Skill Count
    - **Color** = Skill
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

schema = ArrayType(StringType())  
df\_with\_array = df.filter(col("SKILLS\_NAME").isNotNull()) \  
 .withColumn("SKILLS\_ARRAY", from\_json(col("SKILLS\_NAME"), schema))

df\_exploded = df\_with\_array.filter(col("SKILLS\_ARRAY").isNotNull()) \  
 .select("NAICS2", explode(col("SKILLS\_ARRAY")).alias("SKILL"))  
top\_skills\_df = df\_exploded.groupBy("SKILL") \  
 .count() \  
 .orderBy(col("count").desc()) \  
 .limit(10)  
top\_skills = [row["SKILL"] for row in top\_skills\_df.collect()]  
print("Top 10 Skills:", top\_skills)

[Stage 21:> (0 + 1) / 1]

Top 10 Skills: ['Communication', 'Data Analysis', 'Management', 'SQL (Programming Language)', 'Problem Solving', 'Leadership', 'Computer Science', 'Operations', 'Project Management', 'Business Process']

df\_filtered = df\_exploded.filter(col("SKILL").isin(top\_skills)) \  
 .filter(col("NAICS2").isNotNull())  
df\_q6 = df\_filtered.groupBy("NAICS2", "SKILL") \  
 .count()  
pdf\_q6 = df\_q6.toPandas()

fig = px.bar(pdf\_q6,   
 x="NAICS2",   
 y="count",   
 color="SKILL",   
 title="Skill Demand by Industry (NAICS2)",  
 labels={"NAICS2": "Industry", "count": "Skill Count", "SKILL": "Skill"},  
 color\_discrete\_sequence=px.colors.qualitative.Bold)  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font=dict(size=20, family="HelveticaNeue-CondensedBold"),  
 xaxis\_title="Industry",  
 yaxis\_title="Skill Count",  
 xaxis=dict(tickangle=45),  
 legend\_title="Skills",  
 barmode="stack"  
)  
fig.show()  
fig.write\_image("output/q6\_skill\_demand.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Skill Demand by Industry

Only top 10 skills are selected because there are too many values and the graph could not be created with all skills selected. The stacked bar chart reveals that NAICS2 code “99” (likely a miscellaneous category) has a high demand for skills like “Problem Solving” and “Project Management,” indicating a focus on tech and management roles. Industries with codes like “54” (Professional, Scientific, and Technical Services) show significant demand for “Data Analysis” and “Technical Support,” reflecting their reliance on analytical and IT skills.

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

* Analyze how salaries differ across ONET occupation types.
* **Aggregate Data**
  + Compute **median salary** for each occupation in the **ONET taxonomy**.
* **Visualize results**
  + Create a **bubble chart** where:
    - **X-axis** = ONET\_NAME
    - **Y-axis** = Median Salary
    - **Size** = Number of job postings
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

unique\_onet\_names = df.select("ONET\_NAME") \  
 .filter(col("ONET\_NAME").isNotNull()) \  
 .distinct()  
unique\_onet\_names.show(truncate=False)

[Stage 27:> (0 + 1) / 1]

+------------------------------+  
|ONET\_NAME |  
+------------------------------+  
|Business Intelligence Analysts|  
+------------------------------+

df\_q7 = df.filter(col("SALARY\_FROM") > 0) \  
 .groupBy("ONET\_NAME") \  
 .agg({"SALARY\_FROM": "median", "\*": "count"}) \  
 .withColumnRenamed("median(SALARY\_FROM)", "median\_salary") \  
 .withColumnRenamed("count(1)", "job\_count")  
pdf\_q7 = df\_q7.toPandas()  
  
fig = px.scatter(pdf\_q7, x="ONET\_NAME", y="median\_salary", size="job\_count",  
 title="Salary by ONET Occupation",  
 color\_discrete\_sequence=["#d62728"])  
fig.update\_layout(  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20,  
 yaxis\_title="Median Salary (USD)",  
 xaxis\_title="ONET Occupation",  
 xaxis=dict(tickangle=45)  
)  
fig.show()  
fig.write\_image("output/q7\_salary\_onet.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Salary by ONET Occupation

There is only one item under ONET\_NAME, so the median for Business Intelligence Analysts is 88k.

# 8. Career Pathway Trends (Sankey Diagram)

* Visualize job transitions between different occupation levels.
* **Aggregate Data**
  + Identify career transitions between **SOC job classifications**.
* **Visualize results**
  + Create a **Sankey diagram** where:
    - **Source** = SOC\_2021\_2\_NAME
    - **Target** = SOC\_2021\_3\_NAME
    - **Value** = Number of transitions
  + Apply custom colors and font styles.
* **Explanation:** Write two sentences about what the graph reveals.

df\_q8 = df.groupBy("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME").count()  
pdf\_q8 = df\_q8.toPandas()  
  
labels = list(set(pdf\_q8["SOC\_2021\_2\_NAME"].tolist() + pdf\_q8["SOC\_2021\_3\_NAME"].tolist()))  
source = [labels.index(x) for x in pdf\_q8["SOC\_2021\_2\_NAME"]]  
target = [labels.index(x) for x in pdf\_q8["SOC\_2021\_3\_NAME"]]  
value = pdf\_q8["count"]  
  
fig = go.Figure(data=[go.Sankey(  
 node=dict(label=labels, color="#9467bd"),  
 link=dict(source=source, target=target, value=value, color="#e377c2")  
)])  
fig.update\_layout(  
 title\_text="Career Pathway Trends",  
 font=dict(family="Helvetica Neue", size=16, color="#333"),  
 title\_font\_size=20  
)  
fig.show()  
fig.write\_image("output/q8\_career\_pathways.svg")

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Career Pathway Trends

From what the data sheet shows, all 72.5k have transfered from “Computer and Mathematical Occupations” to “Mathematical Science Occupations”.