Module 04: Lab 01

Visual Reporting and Storytelling

Marco Paul Perez

November 21, 2024

# 1. 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**.

# 2. Step 1: Load the Dataset

import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
from setuptools.\_distutils.version import LooseVersion  
pio.renderers.default = "notebook"  
from pyspark.sql import SparkSession  
from pyspark.sql.functions import col, regexp\_replace, split, explode, trim, count, to\_date  
import plotly.graph\_objects as go  
from pyspark.sql.functions import sum as spark\_sum  
  
  
  
spark = SparkSession.builder.appName("LightcastData").getOrCreate()  
  
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine","true").option("escape", "\"").csv("./data/lightcast\_job\_postings.csv")  
  
df.printSchema()  
df.show(5)

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

# 3. 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.

filtered\_df = df.filter((df["SALARY"].isNotNull()) & (df["SALARY"] > 0))  
  
pdf = filtered\_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=["#00B5B7"])  
fig.update\_layout(font\_family="Helvetica", title\_font\_size=18)  
fig.show()  
fig.write\_image("output/figure1.svg")

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

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

The box plot shows that salary distributions are fairly consistent across all employment types, with similar medians and spreads. All groups—full-time, part-time, and mixed—have a wide range of salaries, including some very low values. However, most salaries fall between approximately $8,000 and $25,000. This suggests that employment type may not be a strong differentiator in salary levels within this dataset.

# 4. 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.

filtered\_df = df.filter(df["SALARY\_FROM"] > 0)  
  
pdf = filtered\_df.select("NAICS2\_NAME", "SALARY\_FROM").toPandas()  
  
fig = px.box(  
 pdf,  
 x="NAICS2\_NAME",  
 y="SALARY\_FROM",  
 title="Salary Distribution by Industry",  
 color\_discrete\_sequence=["#EF553B"]  
)  
  
fig.update\_layout(font\_family="Helvetica", title\_font\_size=18)  
fig.show()  
fig.write\_image("output/figure2.svg")

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

The box plot reveals that salary distributions across industries are quite similar, with most median salaries falling between $15,000 and $20,000. However, all industries display a wide range of salaries, indicating variability within each sector rather than stark differences between them.

# 5. 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\_date = df.withColumn("POSTED\_DATE", to\_date("POSTED", "M/d/yyyy"))  
daily\_postings = df\_date.groupBy("POSTED\_DATE").agg(count("\*").alias("Number\_of\_Postings"))  
daily\_postings = daily\_postings.orderBy("POSTED\_DATE")  
daily\_postings.show(5)

+-----------+------------------+  
|POSTED\_DATE|Number\_of\_Postings|  
+-----------+------------------+  
| NULL| 22|  
| 2024-05-01| 506|  
| 2024-05-02| 437|  
| 2024-05-03| 679|  
| 2024-05-04| 573|  
+-----------+------------------+  
only showing top 5 rows

daily\_postings\_pdf = daily\_postings.orderBy("POSTED\_DATE").toPandas()  
  
daily\_postings\_pdf = daily\_postings\_pdf[daily\_postings\_pdf['POSTED\_DATE'].notna()]  
daily\_postings\_pdf['POSTED\_DATE'] = pd.to\_datetime(daily\_postings\_pdf['POSTED\_DATE'])  
daily\_postings\_pdf['Number\_of\_Postings'] = pd.to\_numeric(daily\_postings\_pdf['Number\_of\_Postings'], errors='coerce')

fig = px.line(  
 daily\_postings\_pdf,  
 x="POSTED\_DATE",  
 y="Number\_of\_Postings",  
 title="Job Posting Trends Over Time",  
 markers=True,  
 line\_shape="linear"  
)  
  
fig.update\_layout(  
 xaxis\_title="Posted Date",  
 yaxis\_title="Number of Postings",  
 hovermode="x unified"  
)  
  
fig.show()  
fig.write\_image("output/figure3.svg")

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

The line chart shows noticeable fluctuations in job postings over time, with frequent peaks and drops that suggest a recurring weekly pattern, indicating possible seasonality. This likely reflects consistent posting behavior aligned with workweeks—higher activity on certain weekdays.

# 6. 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.

title\_counts = (  
 df.groupBy("TITLE\_NAME")  
 .agg(count("\*").alias("Job\_Count"))  
 .orderBy(col("Job\_Count").desc())  
 .limit(10)  
)  
  
pdf\_titles = title\_counts.toPandas()  
  
fig = px.bar(  
 pdf\_titles,  
 x="TITLE\_NAME",  
 y="Job\_Count",  
 title="Top 10 Job Titles by Count",  
 color\_discrete\_sequence=["#00CC96"]  
)  
  
fig.update\_layout(  
 font\_family="Arial",  
 title\_font\_size=16,  
 xaxis\_title="Job Title",  
 yaxis\_title="Number of Postings",  
 xaxis\_tickangle=45  
)  
  
fig.show()  
fig.write\_image("output/figure4.svg")

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

The bar chart reveals that Data Analyst roles are the most frequently posted job titles by a large margin, followed by Unclassified (jobs not indicating the exact position) and Business Intelligence Analysis positions. The steep drop-off after the top title suggests a strong demand for Data Analysts in the job market, while the presence of many specialized roles in the top 10 highlights the growing need for data-driven expertise across different areas.

# 7. 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.

remote\_counts = (  
 df.groupBy("REMOTE\_TYPE\_NAME")  
 .agg(count("\*").alias("Job\_Count"))  
)  
  
pdf\_remote = remote\_counts.toPandas()  
  
fig = px.pie(  
 pdf\_remote,  
 names="REMOTE\_TYPE\_NAME",  
 values="Job\_Count",  
 title="Remote vs On-Site Job Postings",  
 color\_discrete\_sequence=px.colors.qualitative.Pastel  
)  
  
fig.update\_layout(  
 font\_family="Arial",  
 title\_font\_size=16  
)  
  
fig.show()  
fig.write\_image("output/figure5.svg")

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

filtered\_df = df.filter(  
 (col("REMOTE\_TYPE\_NAME").isNotNull()) &   
 (col("REMOTE\_TYPE\_NAME") != "[None]")  
)  
  
remote\_counts = (  
 filtered\_df.groupBy("REMOTE\_TYPE\_NAME")  
 .agg(count("\*").alias("Job\_Count"))  
)  
  
pdf\_remote = remote\_counts.toPandas()  
  
fig = px.pie(  
 pdf\_remote,  
 names="REMOTE\_TYPE\_NAME",  
 values="Job\_Count",  
 title="Remote vs On-Site Job Postings",  
 color\_discrete\_sequence=px.colors.qualitative.Pastel  
)  
  
fig.update\_layout(  
 font\_family="Arial",  
 title\_font\_size=16  
)  
  
fig.show()  
fig.write\_image("output/figure6.svg")

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

The first pie chart includes [None] and null values, which is helpful to highlight that a large portion of job postings do not specify their remote type. However, the second chart focuses only on postings with a defined remote status. Among those, Remote jobs clearly dominate, indicating a strong preference or trend toward remote work among the classified roles.

# 8. 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.

cleaned\_df = df.withColumn(  
 "SKILLS\_NAME",  
 regexp\_replace(col("SKILLS\_NAME"), r'[\[\]\n"]', '')  
)  
cleaned\_df = cleaned\_df.withColumn("SKILLS\_NAME", explode(split(col("SKILLS\_NAME"), ",")))  
cleaned\_df = cleaned\_df.withColumn("SKILLS\_NAME", trim(col("SKILLS\_NAME")))  
  
skill\_counts = (  
 cleaned\_df.groupBy("NAICS\_2022\_6\_NAME", "SKILLS\_NAME")  
 .agg(count("\*").alias("skill\_count"))  
)  
  
top\_industries\_df = (  
 skill\_counts.groupBy("NAICS\_2022\_6\_NAME")  
 .agg(spark\_sum("skill\_count").alias("total\_skill\_count"))  
 .orderBy(col("total\_skill\_count").desc())  
 .limit(10)  
)  
  
top\_industry\_names = [row["NAICS\_2022\_6\_NAME"] for row in top\_industries\_df.collect()]  
  
filtered\_skills = skill\_counts.filter(col("NAICS\_2022\_6\_NAME").isin(top\_industry\_names))  
  
top\_skills\_df = (  
 filtered\_skills.groupBy("SKILLS\_NAME")  
 .agg(spark\_sum("skill\_count").alias("total\_skill\_count"))  
 .orderBy(col("total\_skill\_count").desc())  
 .limit(10)  
)  
top\_skill\_names = [row["SKILLS\_NAME"] for row in top\_skills\_df.collect()]  
  
filtered\_skills = filtered\_skills.filter(col("SKILLS\_NAME").isin(top\_skill\_names))  
  
pdf = filtered\_skills.toPandas()  
  
fig = px.bar(  
 pdf,  
 x="NAICS\_2022\_6\_NAME",  
 y="skill\_count",  
 color="SKILLS\_NAME",  
 title="Top 10 Most In-Demand Skills by NAICS-6 Industries",  
 color\_discrete\_sequence=px.colors.qualitative.Set3  
)  
  
fig.update\_layout(  
 barmode="stack",  
 font=dict(family="Arial", size=12),  
 title=dict(text="Top 10 Most In-Demand Skills by NAICS-6 Industries", x=0.5, font=dict(size=20)),  
 xaxis=dict(  
 title="Industry (NAICS 6)",  
 tickangle=30,  
 tickfont=dict(size=10),  
 automargin=True  
 ),  
 yaxis=dict(  
 title="Skill Count",  
 gridcolor="#EFEFEF"  
 ),  
 legend\_title\_text="Skill",  
 legend=dict(  
 font=dict(size=10),  
 orientation="v",  
 yanchor="top",  
 y=0.99,  
 xanchor="right",  
 x=1.02  
 ),  
 margin=dict(l=40, r=120, t=80, b=160),  
 height=600,  
 width=1000,  
 plot\_bgcolor="white"  
)  
  
fig.show()  
fig.write\_image("output/figure7.svg")

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

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

This chart shows that *Communication*, *Project Management*, and *Data Analysis* are among the most in-demand skills across the top NAICS-6 industries. Industries like “Unclassified Industry” and “Administrative Management” demonstrate particularly high demand for a broad range of these skills.

# 9. 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.

df.createOrReplaceTempView("jobs")  
  
top\_titles\_salary = spark.sql("""  
 SELECT  
 TITLE\_NAME AS Job\_Title,  
 percentile\_approx(SALARY, 0.5) AS Median\_Salary,  
 COUNT(\*) AS Total\_Postings  
 FROM jobs  
 WHERE SALARY IS NOT NULL AND SALARY > 0 AND TITLE\_NAME IS NOT NULL  
 GROUP BY TITLE\_NAME  
 ORDER BY Total\_Postings DESC  
 LIMIT 10  
""")  
  
pdf\_titles\_salary = top\_titles\_salary.toPandas()  
  
  
fig = px.scatter(  
 pdf\_titles\_salary,  
 x='Job\_Title',  
 y='Median\_Salary',  
 size='Total\_Postings',  
 color='Total\_Postings',  
 hover\_name='Job\_Title',  
 title='Top 10 Roles by Salary & Demand',  
 labels={  
 'Job\_Title': 'Job Role',  
 'Median\_Salary': 'Median Salary (USD)',  
 'Total\_Postings': 'Number of Postings'  
 },  
 width=1100,  
 height=650,  
 size\_max=70,  
 color\_continuous\_scale='Viridis'  
)  
  
fig.update\_layout(  
 font=dict(family='Helvetica', size=14),  
 xaxis=dict(  
 title=dict(text='Job Role', font=dict(size=18)),  
 tickangle=-40,  
 tickfont=dict(size=12),  
 showline=True,  
 linecolor='gray'  
 ),  
 yaxis=dict(  
 title=dict(text='Median Salary', font=dict(size=18)),  
 tickfont=dict(size=12),  
 showline=True,  
 linecolor='gray'  
 ),  
 plot\_bgcolor='white'  
)  
  
fig.show()  
fig.write\_image("output/figure8.svg")

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

This bubble chart shows the relationship between job demand and median salary for the ten most frequently posted roles. While Data Analysts dominate in number of postings, their median salary is relatively lower compared to specialized roles like Enterprise Architects and Principal Architects, which offer higher pay but appear less frequently. This suggests a trade-off between salary and demand across job titles.

# 10. 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.

transitions\_df = (  
 df.filter((col("SOC\_2021\_2\_NAME").isNotNull()) & (col("SOC\_2021\_3\_NAME").isNotNull()))  
 .groupBy("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME")  
 .agg(count("\*").alias("Transition\_Count"))  
)  
  
pdf\_transitions = transitions\_df.toPandas()  
  
labels = list(set(pdf\_transitions["SOC\_2021\_2\_NAME"]).union(set(pdf\_transitions["SOC\_2021\_3\_NAME"])))  
label\_to\_index = {label: i for i, label in enumerate(labels)}  
  
pdf\_transitions["source"] = pdf\_transitions["SOC\_2021\_2\_NAME"].map(label\_to\_index)  
pdf\_transitions["target"] = pdf\_transitions["SOC\_2021\_3\_NAME"].map(label\_to\_index)  
  
fig = go.Figure(data=[go.Sankey(  
 node=dict(  
 pad=20,  
 thickness=20,  
 line=dict(color="black", width=0.5),  
 label=labels,  
 color="lightblue"  
 ),  
 link=dict(  
 source=pdf\_transitions["source"],  
 target=pdf\_transitions["target"],  
 value=pdf\_transitions["Transition\_Count"]  
 )  
)])  
  
fig.update\_layout(  
 title\_text="Career Pathway Transitions (SOC Levels)",  
 font=dict(family="Arial", size=12),  
 title\_font=dict(size=20),  
 title\_x=0.5  
)  
  
fig.show()  
fig.write\_image("output/figure9.svg")

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

The Sankey diagram reveals only one visible career pathway in the dataset—from “Computer and Mathematical Occupations” to “Mathematical Science Occupations.” This suggests that either transitions are sparsely recorded or most rows lack sufficient SOC classification data to map additional career flows.