Module 04: Lab 01

Visual Reporting and Storytelling

Shuyu Pan

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.graph\_objects as go  
import plotly.io as pio  
pio.renderers.default = "vscode"  
from pyspark.sql import SparkSession, Row  
from pyspark.sql.functions import col, avg, min, max, percentile\_approx, count, to\_date, explode, split  
  
# 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)

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

# 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.

# Filter the dataset: Remove records where salary is missing or zero.  
df\_filtered = df.dropna(subset=["SALARY\_FROM"])  
df\_filtered = df\_filtered.filter(col("SALARY\_FROM") > 0)  
# Group by employment type  
employment\_grouped = df\_filtered.groupBy("EMPLOYMENT\_TYPE\_NAME").count()  
employment\_grouped.show()  
  
  
  
# Compute salary distribution  
salary\_distribution = df\_filtered.agg(  
 avg("SALARY\_FROM").alias("Average\_Salary"),  
 min("SALARY\_FROM").alias("Min\_Salary"),  
 max("SALARY\_FROM").alias("Max\_Salary"),  
 percentile\_approx("SALARY\_FROM", 0.25).alias("Q1\_Salary"),  
 percentile\_approx("SALARY\_FROM", 0.50).alias("Median\_Salary"),  
 percentile\_approx("SALARY\_FROM", 0.75).alias("Q3\_Salary")  
)  
salary\_distribution.show()  
  
pdf1 = df.select("EMPLOYMENT\_TYPE\_NAME", "SALARY\_FROM").toPandas()  
fig1 = px.box(pdf1, x="EMPLOYMENT\_TYPE\_NAME", y="SALARY\_FROM", title="Salary Distribution by Employment Type",  
 color\_discrete\_sequence=["#636EFA"])  
fig1.update\_layout(font\_family="Arial", title\_font\_size=16)  
fig1.show()  
fig1.write\_image("output/figure1.svg")

+--------------------+-----+  
|EMPLOYMENT\_TYPE\_NAME|count|  
+--------------------+-----+  
|Part-time / full-...| 642|  
|Part-time (â‰¤ 32...| 1185|  
|Full-time (> 32 h...|30571|  
+--------------------+-----+  
  
+-----------------+----------+----------+---------+-------------+---------+  
| Average\_Salary|Min\_Salary|Max\_Salary|Q1\_Salary|Median\_Salary|Q3\_Salary|  
+-----------------+----------+----------+---------+-------------+---------+  
|94005.13497746775| 10230| 800000| 63565| 88000| 118000|  
+-----------------+----------+----------+---------+-------------+---------+

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

[Salary Distribution by Employment Type](output/figure1.svg)

*The graph reveals that full-time employment (> 32 hours) has a higher median salary and a wider range of salary values compared to part-time roles. Additionally, all employment types show a significant number of outliers, suggesting that a small number of high-paying jobs skew the overall salary distribution.*

# 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.

# Group by NAICS2 and compute salary statistics  
industry\_salary\_stats = df\_filtered.groupBy("NAICS2").agg(  
 count("\*").alias("Job\_Postings"),  
 avg("SALARY\_FROM").alias("Average\_Salary"),  
 min("SALARY\_FROM").alias("Min\_Salary"),  
 max("SALARY\_FROM").alias("Max\_Salary"),  
 percentile\_approx("SALARY\_FROM", 0.5).alias("Median\_Salary")  
)  
  
# Show results  
industry\_salary\_stats.show(truncate=False)  
  
pdf2 = df.select("NAICS2", "NAICS2\_NAME", "SALARY\_FROM").toPandas()  
fig2 = px.box(pdf2, x="NAICS2\_NAME", y="SALARY\_FROM", title="Salary Distribution by Industry",  
 color\_discrete\_sequence=["#FF5733"])  
fig2.update\_layout(font\_family="Times New Roman", title\_font\_size=16, template="plotly\_dark")  
fig2.show()  
fig2.write\_image("output/figure2.svg")

[Stage 70:> (0 + 1) / 1]

+------+------------+------------------+----------+----------+-------------+  
|NAICS2|Job\_Postings|Average\_Salary |Min\_Salary|Max\_Salary|Median\_Salary|  
+------+------------+------------------+----------+----------+-------------+  
|31 |1740 |98871.12298850574 |15000 |290000 |97500 |  
|53 |454 |79555.65198237885 |10230 |260000 |70000 |  
|81 |385 |77634.52727272727 |12480 |250000 |73278 |  
|44 |807 |93717.17596034697 |21237 |400000 |85000 |  
|22 |343 |96464.74635568513 |16640 |207500 |94225 |  
|52 |4013 |91560.88213306753 |15600 |281400 |87280 |  
|54 |9282 |99834.66828269769 |12000 |312000 |97875 |  
|48 |245 |85242.60408163266 |12477 |175150 |80000 |  
|92 |737 |69426.60515603799 |10230 |211300 |65604 |  
|61 |1033 |66704.25072604066 |10617 |600000 |62317 |  
|72 |270 |112200.67037037037|19240 |324000 |100546 |  
|55 |41 |88242.17073170732 |43000 |280000 |75000 |  
|23 |299 |95448.38127090302 |15000 |173000 |95000 |  
|51 |2356 |112607.37648556876|11148 |500000 |105000 |  
|62 |1442 |81147.21636615811 |10272 |455375 |75000 |  
|21 |38 |87847.34210526316 |46280 |152400 |83200 |  
|56 |4040 |92487.35049504951 |12480 |800000 |84156 |  
|11 |29 |78948.20689655172 |41600 |155000 |65000 |  
|71 |90 |76510.1888888889 |25000 |156000 |65000 |  
|42 |943 |87608.13467656415 |10595 |207900 |79700 |  
+------+------------+------------------+----------+----------+-------------+  
only showing top 20 rows

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

[Salary Distribution by Industry](output/figure2.svg)

*The graph shows that salary distributions vary widely across industries, with sectors like Finance and Insurance and Information exhibiting higher salary ranges and more extreme outliers. In contrast, industries such as Retail Trade and Food Services have consistently lower salary medians and narrower distributions, indicating lower-paying job clusters.*

# 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.

# Convert POSTED column to date format  
df\_posted = df.withColumn("POSTED\_DATE", to\_date(col("POSTED"), "M/d/yyyy"))  
# Filter the dataset: Remove records where POSTED\_DATE is NULL.  
df\_posted = df\_posted.dropna(subset=["POSTED\_DATE"])  
# Group by date and count number of postings  
post\_trend = df\_posted.groupBy("POSTED\_DATE").count().orderBy("POSTED\_DATE").withColumnRenamed("count", "Number\_of\_Job\_Postings")  
post\_trend.show()  
  
pdf3 = post\_trend.toPandas()  
fig3 = px.line(pdf3, x="POSTED\_DATE", y="Number\_of\_Job\_Postings", title="Job Posting Trends Over Time",  
 color\_discrete\_sequence=["#00CC96"])  
fig3.update\_layout(font\_family="Georgia", title\_font\_size=16, template="plotly\_white")  
fig3.show()  
fig3.write\_image("output/figure3.svg")

+-----------+----------------------+  
|POSTED\_DATE|Number\_of\_Job\_Postings|  
+-----------+----------------------+  
| 2024-05-01| 506|  
| 2024-05-02| 437|  
| 2024-05-03| 679|  
| 2024-05-04| 573|  
| 2024-05-05| 159|  
| 2024-05-06| 169|  
| 2024-05-07| 516|  
| 2024-05-08| 471|  
| 2024-05-09| 619|  
| 2024-05-10| 599|  
| 2024-05-11| 594|  
| 2024-05-12| 708|  
| 2024-05-13| 300|  
| 2024-05-14| 386|  
| 2024-05-15| 572|  
| 2024-05-16| 557|  
| 2024-05-17| 462|  
| 2024-05-18| 599|  
| 2024-05-19| 170|  
| 2024-05-20| 369|  
+-----------+----------------------+  
only showing top 20 rows

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

[Job Posting Trends Over Time](output/figure3.svg)

*The graph shows a clear cyclical pattern in job postings, with frequent rises and dips that suggest a consistent short-term fluctuation—likely reflecting weekly or bi-weekly hiring behavior. While the overall trend remains relatively stable over time, occasional spikes indicate brief periods of increased hiring demand, potentially tied to organizational or seasonal needs.*

# 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.

# Count job titles  
job\_titles = df.groupBy("TITLE\_NAME").count().orderBy(col("count").desc())  
job\_titles.show(truncate=False)  
# Top 10 most frequent titles  
top\_titles = job\_titles.limit(10)  
top\_titles.show(truncate=False)  
  
pdf4 = top\_titles.toPandas()  
fig4 = px.bar(pdf4, x="TITLE\_NAME", y="count", title="Top 10 Job Titles by Count",  
 color\_discrete\_sequence=["#B34D4D"])  
fig4.update\_layout(font\_family="Helvetica", title\_font\_size=16, plot\_bgcolor="lightgray")  
fig4.show()  
fig4.write\_image("output/figure4.svg")

+-----------------------------------+-----+  
|TITLE\_NAME |count|  
+-----------------------------------+-----+  
|Data Analysts |8591 |  
|Unclassified |3149 |  
|Business Intelligence Analysts |2072 |  
|Enterprise Architects |1999 |  
|Oracle Cloud HCM Consultants |1042 |  
|Data Modelers |668 |  
|Data Governance Analysts |628 |  
|Data Analytics Engineers |537 |  
|ERP Business Analysts |488 |  
|Data Quality Analysts |467 |  
|Data and Reporting Analysts |463 |  
|Data Management Analysts |449 |  
|Solutions Architects |435 |  
|SAP Consultants |425 |  
|Oracle Cloud Financials Consultants|421 |  
|Enterprise Solutions Architects |398 |  
|Principal Architects |387 |  
|SAP EWM Consultants |384 |  
|Lead Data Analysts |378 |  
|Data Analytics Interns |357 |  
+-----------------------------------+-----+  
only showing top 20 rows  
  
+------------------------------+-----+  
|TITLE\_NAME |count|  
+------------------------------+-----+  
|Data Analysts |8591 |  
|Unclassified |3149 |  
|Business Intelligence Analysts|2072 |  
|Enterprise Architects |1999 |  
|Oracle Cloud HCM Consultants |1042 |  
|Data Modelers |668 |  
|Data Governance Analysts |628 |  
|Data Analytics Engineers |537 |  
|ERP Business Analysts |488 |  
|Data Quality Analysts |467 |  
+------------------------------+-----+

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

[Top 10 Job Titles by Count](output/figure4.svg)

*The graph reveals that “Data Analysts” are by far the most frequently posted job title, significantly outpacing all other roles. Titles like “Business Intelligence Analysts” and “Enterprise Architects” also appear frequently, reflecting strong demand for data-related and technical strategy positions 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.

# Count job postings by remote type  
remote\_counts = df.groupBy("REMOTE\_TYPE\_NAME").count().orderBy(col("count").desc())  
remote\_counts = remote\_counts.dropna(subset=["REMOTE\_TYPE\_NAME"])  
remote\_counts.show(truncate=False)  
  
pdf5 = remote\_counts.toPandas()  
fig5 = px.pie(pdf5, names="REMOTE\_TYPE\_NAME", values="count", title="Remote vs On-Site Job Postings",  
 color\_discrete\_sequence=px.colors.qualitative.Set3)  
fig5.update\_layout(font\_family="Comic Sans MS", title\_font\_size=16, title\_font\_color="red")  
fig5.show()  
fig5.write\_image("output/figure5.svg")

+----------------+-----+  
|REMOTE\_TYPE\_NAME|count|  
+----------------+-----+  
|[None] |56570|  
|Remote |12497|  
|Hybrid Remote |2260 |  
|Not Remote |1127 |  
+----------------+-----+

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

[Remote vs On-Site Job Postings](output/figure5.svg)

*The chart reveals that a significant majority of job postings (78.1%) do not specify a remote type, which may indicate inconsistent data entry or limited remote classification by employers. Among the specified postings, remote jobs (17.2%) clearly outnumber hybrid (3.1%) and fully on-site roles (1.6%), reflecting a strong market trend toward remote work flexibility.*

# 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.

# Convert the SKILLS column (which is a string) into an array   
df\_array = df.withColumn("SKILLS\_ARRAY", split(col("SKILLS"), ","))  
df\_exploded = df\_array.select("NAICS2\_NAME", explode(col("SKILLS\_ARRAY")).alias("SKILL"))  
  
# Group by industry (NAICS2\_NAME) and skill, then count occurrences.  
df\_skill\_counts = df\_exploded.groupBy("NAICS2\_NAME", "SKILL") \  
 .agg(count("\*").alias("Skill\_Count"))  
  
# Convert the result to a Pandas DataFrame for Plotly visualization.  
pdf6 = df\_skill\_counts.toPandas()  
  
# Create a stacked bar chart using Plotly Express.  
fig6 = px.bar(pdf6, x="NAICS2\_NAME", y="Skill\_Count",  
 title="Skill Demand Analysis by Industry (Stacked Bar)",  
 barmode="stack",  
 color\_discrete\_sequence=px.colors.qualitative.Dark24  
)  
  
# Apply custom font styles and background colors.  
fig6.update\_layout(font\_family="Impact", title\_font\_size=16, title\_font\_color="blue", plot\_bgcolor="white")  
  
fig6.show()  
fig6.write\_image("output/figure6.svg")

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

[Skill Demand Analysis by Industry (Stacked Bar)](output/figure6.svg)

*The graph reveals that the Professional, Scientific, and Technical Services industry exhibits the highest overall skill demand, with a significantly larger skill count than other sectors. Additionally, industries such as Health Care and Social Assistance and Finance and Insurance also show a relatively high concentration of skill requirements, highlighting their complex and diverse job functions.*

# 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.

df\_filtered = df.select("ONET\_NAME", "SALARY\_FROM") \  
 .filter(col("ONET\_NAME").isNotNull() & col("SALARY\_FROM").isNotNull() & (col("SALARY\_FROM") > 0))  
  
salary\_median = df\_filtered.groupBy("ONET\_NAME").agg(  
 percentile\_approx("SALARY\_FROM", 0.5).alias("Median\_Salary"),  
 count("\*").alias("POSTING\_COUNT")  
)  
salary\_median.show(truncate=False)  
  
pdf7 = salary\_median.toPandas()  
fig7 = px.scatter(pdf7, x="ONET\_NAME", y="Median\_Salary", size="POSTING\_COUNT", color="ONET\_NAME",  
 title="Salary Analysis by ONET Occupation Type",  
 labels={"ONET\_NAME": "Occupation Type", "AVG\_SALARY": "Average Salary"},  
 color\_discrete\_sequence=px.colors.qualitative.Set2  
)  
fig7.update\_layout(font\_family="Tahoma", title\_font\_size=16, title\_font\_color="purple")  
fig7.show()  
fig7.write\_image("output/figure7.svg")

+------------------------------+-------------+-------------+  
|ONET\_NAME |Median\_Salary|POSTING\_COUNT|  
+------------------------------+-------------+-------------+  
|Business Intelligence Analysts|88000 |32398 |  
+------------------------------+-------------+-------------+

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

[Salary Analysis by ONET Occupation Type](output/figure7.svg)

*The graph shows that the median salary for Business Intelligence Analysts is approximately $88,000, indicating a strong earning potential in this occupation. Since only one occupation type appears, it suggests either limited data or filtering, highlighting the need for broader data coverage to compare across multiple ONET job roles.*

# 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.

# Filter out nulls  
df\_soc = df.select("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME") \  
 .filter(col("SOC\_2021\_2\_NAME").isNotNull() & col("SOC\_2021\_3\_NAME").isNotNull())  
  
# Group transitions and count occurrences  
df\_transitions = df\_soc.groupBy("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME") \  
 .agg(count("\*").alias("value")) \  
 .orderBy(col("value").desc()) \  
 .limit(15)  
  
df\_transitions.show(20, truncate=False)  
  
# Convert to pandas for plotting  
pdf8 = df\_transitions.toPandas()  
  
# Create list of unique node names  
nodes = pd.Series(  
 list(set(pdf8['SOC\_2021\_2\_NAME']).union(set(pdf8['SOC\_2021\_3\_NAME'])))  
)  
  
# Create mapping for name → ID  
name\_to\_id = {name: i for i, name in enumerate(nodes)}  
  
# Map source and target names to numeric IDs  
pdf8['source\_id'] = pdf8['SOC\_2021\_2\_NAME'].map(name\_to\_id)  
pdf8['target\_id'] = pdf8['SOC\_2021\_3\_NAME'].map(name\_to\_id)  
  
# Build Sankey diagram  
fig8 = go.Figure(data=[go.Sankey(  
 node=dict(  
 pad=15,  
 thickness=20,  
 line=dict(color="gray", width=0.5),  
 label=nodes.tolist(),  
 color="rgba(55,128,191,0.6)"  
 ),  
 link=dict(  
 source=pdf8['source\_id'],  
 target=pdf8['target\_id'],  
 value=pdf8['value'],  
 color="rgba(100,100,200,0.4)"  
 )  
)])  
  
# Apply layout styling  
fig8.update\_layout(  
 title\_text="Career Pathway Trends by SOC Classification",  
 font\_family="Helvetica Neue, Helvetica, Sans-serif",  
 font\_size=14,  
 title\_font=dict(size=20, color="#333"),  
 plot\_bgcolor="#ffffff"  
)  
  
fig8.show()  
fig8.write\_image("output/figure8.svg")

+-------------------------------------+--------------------------------+-----+  
|SOC\_2021\_2\_NAME |SOC\_2021\_3\_NAME |value|  
+-------------------------------------+--------------------------------+-----+  
|Computer and Mathematical Occupations|Mathematical Science Occupations|72454|  
+-------------------------------------+--------------------------------+-----+

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

[Career Pathway Trends by SOC Classification](output/figure8.svg)

*The Sankey diagram reveals that a significant number of career transitions occur from Computer and Mathematical Occupations (SOC 2021 2-digit level) to Mathematical Science Occupations (SOC 2021 3-digit level). This suggests a common pathway where professionals in general computing or data roles progress into more specialized mathematical fields, potentially reflecting upskilling or role refinement over time.*