Module 04: Lab 01

Visual Reporting and Storytelling

Binderiya Dugersuren

March 23, 2025

# 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  
  
  
# 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("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 02:58:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
 25/03/25 02:59:15 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: 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.

from pyspark.sql.functions import avg, min, max, count

filtered\_df = df.filter((col("SALARY").isNotNull()) & (col("SALARY") != 0))  
  
aggregated\_df = filtered\_df.groupBy("EMPLOYMENT\_TYPE\_NAME").agg(  
 count("\*").alias("job\_count"),  
 avg("SALARY").alias("avg\_salary"),  
 min("SALARY").alias("min\_salary"),  
 max("SALARY").alias("max\_salary")  
)  
  
aggregated\_df.show()

[Stage 3:> (0 + 1) / 1]

+--------------------+---------+------------------+----------+----------+  
|EMPLOYMENT\_TYPE\_NAME|job\_count| avg\_salary|min\_salary|max\_salary|  
+--------------------+---------+------------------+----------+----------+  
|Part-time / full-...| 619| 105621.2423263328| 20800| 455375|  
|Part-time (â‰¤ 32...| 1038| 98802.50963391137| 15860| 310050|  
|Full-time (> 32 h...| 29151|118897.55860862407| 20583| 500000|  
+--------------------+---------+------------------+----------+----------+

filtered\_df.select("EMPLOYMENT\_TYPE\_NAME", "SALARY") \  
 .write.mode("overwrite") \  
 .option("header", True) \  
 .csv("filtered\_salary\_data")  
  
import pandas as pd  
import glob  
  
files = glob.glob("filtered\_salary\_data/part-\*.csv")  
pdf = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)

pdf["EMPLOYMENT\_TYPE\_NAME"] = pdf["EMPLOYMENT\_TYPE\_NAME"].replace({  
 "Part-time (â‰¤ 32 hours)": "Part-time (≤ 32 hours)",  
 "Full-time (â‰¥ 32 hours)": "Full-time (≥ 32 hours)",  
 "Part-time / full-time": "Part-time / full-time"  
})

# Your Code for 1st question here  
#pdf = df.select("EMPLOYMENT\_TYPE\_NAME", "SALARY").toPandas()  
import plotly.io as pio  
pio.renderers.default = "notebook\_connected"  
  
fig = px.box(pdf, x="EMPLOYMENT\_TYPE\_NAME", y="SALARY",   
 title="Salary Distribution by Employment Type",   
 color = "EMPLOYMENT\_TYPE\_NAME",  
 color\_discrete\_sequence= ["yellow", "red", "coral"],  
 labels={  
 "EMPLOYMENT\_TYPE\_NAME": "Employment Type",  
 "SALARY": "Salary"  
 })  
fig.update\_layout(font\_family="Aptos", title\_font\_size=17,   
 xaxis=dict(title=dict(text='Employment Type'), zeroline=False),  
 yaxis=dict(title=dict(text='Salary'), zeroline=False))  
  
  
fig.show()

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

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

import os  
fig.write\_image("\_output/figure.svg")  
  
import kaleido  
print("Kaleido version:", kaleido.\_\_version\_\_)

ValueError:   
Image export using the "kaleido" engine requires the kaleido package,  
which can be installed using pip:  
 $ pip install -U kaleido

First I cleaned up the employment type labels in the DataFrame for readability in the plot. From the box plot, we can observe that full-time employment has a significant number of outliers on the higher end of the salary distribution. In contrast, part-time (≤ 32 hours) shows a more tightly concentrated distribution with fewer extreme values. The interquartile range for part-time/full-time jobs appears narrower, indicating lower variability within that category. Also, the median salary for full-time roles is the highest among all groups, while part-time roles have the lowest median salaries.

# 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.

# Your code for 2nd question here  
filtered\_df = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") != 0))  
  
industry\_df = filtered\_df.groupBy("NAICS2\_NAME").agg(  
 count("\*").alias("job\_count"))  
  
industry\_df.show()

[Stage 7:> (0 + 1) / 1]

+--------------------+---------+  
| NAICS2\_NAME|job\_count|  
+--------------------+---------+  
|Administrative an...| 4040|  
|Public Administra...| 737|  
|Real Estate and R...| 454|  
| Information| 2356|  
|Unclassified Indu...| 3811|  
|Accommodation and...| 270|  
|Finance and Insur...| 4013|  
| Construction| 299|  
| Utilities| 343|  
|Management of Com...| 41|  
|Professional, Sci...| 9282|  
|Arts, Entertainme...| 90|  
|Other Services (e...| 385|  
|Transportation an...| 245|  
| Wholesale Trade| 943|  
|Agriculture, Fore...| 29|  
| Manufacturing| 1740|  
|Mining, Quarrying...| 38|  
|Educational Services| 1033|  
|Health Care and S...| 1442|  
+--------------------+---------+  
only showing top 20 rows

filtered\_df = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") > 0)&   
 (col("NAICS2\_NAME").isin("Health Care and Social Assistance",  
 "Finance and Insurance", "Information")))  
salary\_df = filtered\_df.select("NAICS2\_NAME", "SALARY\_FROM")  
salary\_df.write.mode("overwrite").option("header", True).csv("industry\_salaries")  
files = glob.glob("industry\_salaries/part-\*.csv")  
pdf = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)  
pio.renderers.default = "notebook\_connected"  
  
fig = px.box(  
 pdf,  
 x="NAICS2\_NAME",  
 y="SALARY\_FROM",  
 title="Salary Distribution by Industry",  
 color="NAICS2\_NAME",  
 color\_discrete\_sequence = ['rgb(67,67,67)', 'rgb(115,115,115)', 'rgb(49,130,189)']  
   
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=17,  
 xaxis=dict(title=dict(text='Industry Type'), zeroline=False),  
 yaxis=dict(title=dict(text='Salary'), zeroline=False)  
)  
  
fig.show()

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

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

From this box plot, we can observe the salary distributions for the Information, Finance & Insurance, and Healthcare & Social Assistance industries. Among these, the Information industry has the highest median salary, while the Healthcare & Social Assistance industry shows the lowest median salary. But, also we see a wide distribution of data in the healthcare industry. This suggests that job roles in the Information sector tend to be more highly compensated compared to the other two industries.

# 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.

# Your code for 3rd question here  
posted\_df = df.filter(col("POSTED").isNotNull())  
  
job\_post\_df = df.filter(col("POSTED").isNotNull()) \  
 .groupBy("POSTED") \  
 .agg(count("\*").alias("job\_count"))  
  
job\_post\_df.write.mode("overwrite").option("header", True).csv("job\_posts")  
files = glob.glob("job\_posts/part-\*.csv")  
pdf = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)  
  
pdf["POSTED"] = pd.to\_datetime(pdf["POSTED"])  
pdf = pdf.sort\_values("POSTED")   
  
pio.renderers.default = "notebook\_connected"  
  
fig = px.line(  
 pdf,  
 x="POSTED",  
 y="job\_count",  
 title="Job Postings Over Time",  
 color\_discrete\_sequence = ["rgb(49,130,189)"]  
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=17,  
 xaxis\_title="Date Posted",  
 yaxis\_title="Number of Job Postings"  
)  
  
fig.show()

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

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

From this line graph we can see the number of job postings from may of 2024 to around october. We can see that ther has alway been a fluctuation of job postings day to day, however we can see a slight decrease a valley in the data from july to early september. I think that probably around this time number of job postings have been decreasing.

# 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.

job\_title = df.filter(col("TITLE\_NAME").isNotNull()) \  
 .groupBy("TITLE\_NAME") \  
 .agg(count("\*").alias("job\_count"))  
top\_10\_jobs = job\_title.orderBy("job\_count", ascending=False).limit(10)  
  
top\_10\_jobs.show()

[Stage 14:> (0 + 1) / 1]

+--------------------+---------+  
| TITLE\_NAME|job\_count|  
+--------------------+---------+  
| Data Analysts| 8591|  
| Unclassified| 3149|  
|Business Intellig...| 2072|  
|Enterprise Archit...| 1999|  
|Oracle Cloud HCM ...| 1042|  
| Data Modelers| 668|  
|Data Governance A...| 628|  
|Data Analytics En...| 537|  
|ERP Business Anal...| 488|  
|Data Quality Anal...| 467|  
+--------------------+---------+

# Your code for 4th question here  
top\_10\_jobs.write.mode("overwrite").option("header", True).csv("top\_10\_jobs")  
files = glob.glob("top\_10\_jobs/part-\*.csv")  
pdf4 = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)  
  
pio.renderers.default = "notebook\_connected"  
  
fig = px.bar(  
 pdf4,  
 x="TITLE\_NAME",  
 y="job\_count",  
 title="Top 10 Jobs",  
 color = "TITLE\_NAME"  
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=17,  
 xaxis\_title="Top 10 Jobs",  
 yaxis\_title="Number of Job Postings"  
)  
  
fig.show()

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

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

From this bar graph, we can see that Data Analyst roles are by far the most sought after in the current job market. Among the top 10 job titles, many are data-related, but Data Analyst stands out as the most in demand position.

# 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.

# Your code for 5th question here  
remote\_jobs = df.filter(col("REMOTE\_TYPE\_NAME").isNotNull()) \  
 .groupBy("REMOTE\_TYPE\_NAME") \  
 .agg(count("\*").alias("job\_count"))  
  
remote\_jobs.show()

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

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

remote\_jobs.write.mode("overwrite").option("header", True).csv("remote\_jobs")  
files = glob.glob("remote\_jobs/part-\*.csv")  
pdf5 = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)  
  
pio.renderers.default = "notebook\_connected"  
  
custom\_color = [  
 "#4B0082", "#6A0DAD", "#8A2BE2", "#A678F1", "#C8A2C8", "#E6E6FA"  
]   
  
fig = px.pie(  
 pdf5,  
 names="REMOTE\_TYPE\_NAME",   
 values="job\_count",   
 title="Remote vs On-Site Jobs",  
 color\_discrete\_sequence=custom\_color  
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=17  
)  
fig.show()

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

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

This is a pie chart showing the percentages of remote and on site jobs. From the graph we can see that most of the jobs listed in the data, 78.1% of the jobs are listed as none, which means the job posting did not specify it. But, we can see that 17.2% jobs have been listed as remote and 3.12% have been listed as hybrid.

# 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.

# Your code for 6th question here  
from pyspark.sql.functions import split, explode, trim  
  
skills\_df = df.filter(  
 col("SKILLS\_NAME").isNotNull() &  
 col("NAICS2\_NAME").isNotNull() &  
 col("NAICS2\_NAME").isin(  
 "Health Care and Social Assistance",  
 "Finance and Insurance",  
 "Information"  
 )  
)  
  
skills\_df = skills\_df.withColumn("SKILL", explode(split(col("SKILLS\_NAME"), ",")))  
  
skills\_df = skills\_df.withColumn("SKILL", trim(col("SKILL")))  
  
industry\_skills\_df = skills\_df.groupBy("NAICS2\_NAME", "SKILL") \  
 .agg(count("\*").alias("skill\_count"))

import glob  
industry\_skills\_df.write.mode("overwrite").option("header", True).csv("industry\_skills")  
files = glob.glob("industry\_skills/part-\*.csv")  
pdf6 = pd.concat((pd.read\_csv(f) for f in files), ignore\_index=True)  
  
import plotly.express as px  
import glob  
fig = px.bar(  
 pdf6,  
 x="NAICS2\_NAME",  
 y="skill\_count",  
 color="SKILL",   
 title="Top Skills by Industry",  
 barmode="stack"  
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=18,  
 xaxis\_title="Industry",  
 yaxis\_title="Skill Count"  
)  
  
fig.show()

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

From this bar graph we can see that one of the most sought after sill is data warehousing in finance and insurance industry, and the second one is the leadership skill. And in the next two industries the most sought after skill is computer science, i am not sure what level of programming or IT knowledge the companies are seeking, but I think haveing basic programming skill would be neccesary.

# 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.

# Your code for 7th question here  
onet\_df = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") != 0) & (col("ONET\_NAME").isNotNull()))  
  
onet\_groups = onet\_df.select("ONET\_NAME").distinct().rdd.flatMap(lambda x: x).collect()  
print(onet\_groups)  
  
results = []  
for onet in onet\_groups:  
 subset = onet\_df.filter(col("ONET\_NAME") == onet)  
 median\_salary = subset.approxQuantile("SALARY\_FROM", [0.5], 0.01)[0]  
 job\_count = subset.count()  
 results.append((onet, median\_salary, job\_count))

['Business Intelligence Analysts']

import pandas as pd  
  
onet\_summary\_df = pd.DataFrame(results, columns=["ONET\_NAME", "median\_salary", "job\_count"])  
  
pio.renderers.default = "notebook\_connected"  
  
fig = px.scatter(  
 onet\_summary\_df,  
 x="ONET\_NAME",  
 y="median\_salary",  
 size="job\_count",  
 title="Median Salary and Job Count by ONET Group",  
 color="median\_salary",   
 size\_max=60,  
 color\_continuous\_scale="Viridis"  
)  
  
fig.update\_layout(  
 font\_family="Aptos",  
 title\_font\_size=18,  
 xaxis\_title="ONET Job Family",  
 yaxis\_title="Median Salary",  
 xaxis\_tickangle=45  
)  
  
fig.show()

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

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

This is a bubble chart showing median salary by ONET occupation type. However, because our data has only one type of occupation type, which is Business Inteligence Analyst we get only one bubble for our bubble chart.

# 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.

# Your code for 8th question here  
career\_path = df.filter(col("SOC\_2021\_2\_NAME").isNotNull() & col("SOC\_2021\_3\_NAME").isNotNull())   
career\_path = career\_path.select("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME").agg(count("\*").alias("job\_count"))  
career\_path.show()  
  
  
#career\_list = [["Comp and Math", "Math Science",], [72454, 72454]]  
#career\_path\_df = pd.DataFrame(career\_list, columns=["SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME" ])  
#print(career\_path\_df)  
  
career\_list = [["Comp and Math", "Math Science", 72454]]  
career\_path\_df = pd.DataFrame(career\_list, columns=["SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME", "job\_count"])

[Stage 36:> (0 + 1) / 1]

+---------+  
|job\_count|  
+---------+  
| 72454|  
+---------+

import plotly.graph\_objects as go  
  
# a unique list of all node labels  
all\_labels = list(set(career\_path\_df["SOC\_2021\_2\_NAME"].tolist() + career\_path\_df["SOC\_2021\_3\_NAME"].tolist()))  
label\_to\_index = {label: i for i, label in enumerate(all\_labels)}  
  
# Map source and target labels to their index  
career\_path\_df["source"] = career\_path\_df["SOC\_2021\_2\_NAME"].map(label\_to\_index)  
career\_path\_df["target"] = career\_path\_df["SOC\_2021\_3\_NAME"].map(label\_to\_index)  
  
# Sankey chart  
fig = go.Figure(data=[go.Sankey(  
 node=dict(  
 pad=15,  
 thickness=20,  
 label=all\_labels,  
 color = ["#FFB6C1", "#FFC0CB"]  
 ),  
 link=dict(  
 source=career\_path\_df["source"],  
 target=career\_path\_df["target"],  
 value=career\_path\_df["job\_count"],  
 color = ["#F8BBD0"]  
 )  
)])  
  
fig.update\_layout(  
 title\_text="Career Flow SOC 2 to SOC 3",  
 font=dict(size=15, color="#333", family="Aptos"),  
 paper\_bgcolor="#fff0f5" # light lavender-pink background  
)  
  
fig.show()

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

The width of the flow represents the number of job postings, highlighting the volume moving between two career levels however because it is not hierarhcical adn we only have one to one value, that is why our Sankey plot is one big chunk flowing to the other. I think if we have more hierarchical data, I think it might be more interesting to visualize.