Module 04: Lab 01

Visual Reporting and Storytelling

Your Name

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 numpy as np  
import plotly.express as px  
import plotly.io as pio  
import plotly.graph\_objects as go  
pio.renderers.default = "vscode"  
from pyspark.sql import SparkSession  
from pyspark.sql.functions import col, count, desc, to\_date, split, explode, when, lit  
from pyspark.sql.window import Window  
import pyspark.sql.functions as F  
  
  
# 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)

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| 2024-09-07 04:32:...| 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| 2024-08-03 01:08:...| 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| 2024-09-07 04:32:...| 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| 2024-09-07 04:32:...| 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 15:00:00| 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.

# Your Code for 1st question here  
df\_filtered = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") > 0))  
  
fig = px.box(  
 df\_filtered,   
 x='EMPLOYMENT\_TYPE\_NAME',   
 y='SALARY\_FROM',  
 title='Salary Distribution Across Employment Types',  
 labels={  
 'EMPLOYMENT\_TYPE\_NAME': 'Employment Type',   
 'SALARY\_FROM': 'Salary (From)'  
 },  
 color='EMPLOYMENT\_TYPE\_NAME',  
 color\_discrete\_sequence=px.colors.qualitative.Pastel,  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 showlegend=False,  
 xaxis\_tickangle=-45  
)  
  
fig.update\_xaxes(  
 title\_standoff=25  
)  
  
fig.update\_yaxes(  
 title\_standoff=25,  
 tickformat='$,.0f'   
)  
  
pio.write\_html(fig, file='salary\_distribution\_boxplot.html')  
fig.show()  
# summary\_stats = df\_filtered.groupby('EMPLOYMENT\_TYPE\_NAME')['SALARY\_FROM'].agg(['mean', 'median', 'min', 'max']).round(2)  
# print(summary\_stats)

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

Based on the salary distribution boxplot, full-time positions (with ≥32 hours) show the highest median salary and the widest salary range, indicating more variability and potentially higher earning potential compared to part-time employment types. The graph also reveals significant outliers in each employment category, particularly in full-time positions, where several high-salary data points extend well above $400,000, suggesting exceptional compensation opportunities for some full-time roles.

# 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  
fig = px.box(  
 df\_filtered,   
 x='NAICS2\_NAME',   
 y='SALARY\_FROM',  
 title='Salary Distribution Across Industries',  
 labels={  
 'NAICS2\_NAME': 'Industry',   
 'SALARY\_FROM': 'Salary (From)'  
 },  
 color='NAICS2\_NAME',  
 color\_discrete\_sequence=px.colors.qualitative.Set3,  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 showlegend=False,  
 xaxis\_tickangle=-45,  
 height=600  
)  
  
fig.update\_xaxes(  
 title\_standoff=25  
)  
  
fig.update\_yaxes(  
 title\_standoff=25,  
 tickformat='$,.0f'  
)  
  
pio.write\_html(fig, file='salary\_by\_industry\_boxplot.html')  
fig.show()

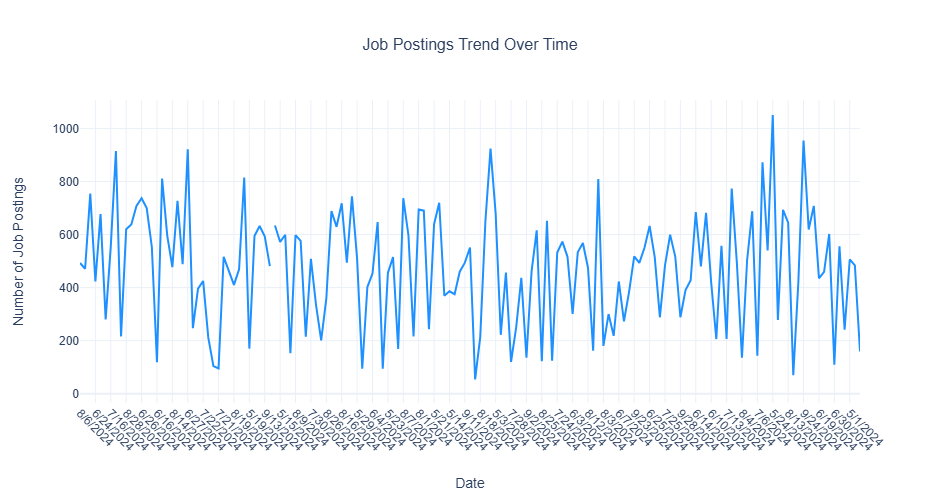
Unable to display output for mime type(s): application/vnd.plotly.v1+json

The box plot reveals substantial salary variations across different industries, with sectors like Professional, Scientific, and Technical Services, Information, and Finance and Insurance showing higher median salaries and more extensive salary ranges compared to industries like Arts, Entertainment, and Recreation or Agriculture, Forestry, Fishing, and Hunting. Additionally, the visualization highlights significant salary outliers in multiple industries, particularly in Professional Services and Information, where some individual salaries extend well beyond $600,000, indicating exceptional compensation opportunities in those sectors.

# 3. Job Posting Trends Over Time

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

df\_agg = df.groupBy('POSTED').agg(count('\*').alias('Number of Postings'))  
  
job\_postings\_trend = df\_agg.toPandas()  
  
fig = px.line(  
 job\_postings\_trend,   
 x='POSTED',   
 y='Number of Postings',  
 title='Job Postings Trend Over Time',  
 labels={'POSTED': 'Date', 'Number of Postings': 'Number of Job Postings'},  
 color\_discrete\_sequence=['#1E90FF'],  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 height=500  
)  
  
fig.update\_xaxes(  
 title\_standoff=25,  
 tickangle=45  
)  
  
fig.update\_yaxes(  
 title\_standoff=25  
)  
  
pio.write\_html(fig, file='job\_postings\_trend.html')  
fig.show()

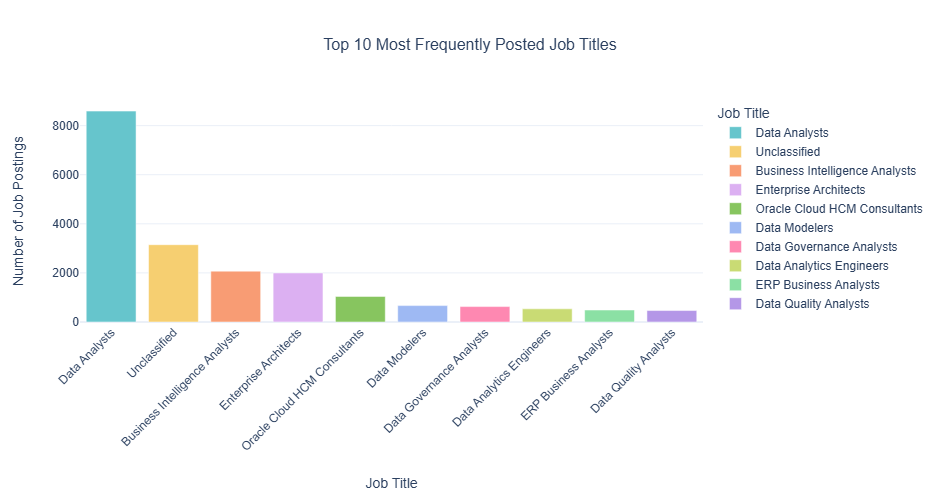


The line chart demonstrates significant day-to-day volatility in job postings, with daily volumes ranging from around 200 to nearly 1,000 job listings and showing a cyclical pattern that suggests potential weekly or bi-weekly fluctuations in hiring activity. Despite the considerable variations, there’s no immediately apparent long-term upward or downward trend, indicating a relatively stable job market during the observed time period with regular short-term oscillations in posting volumes.

# 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.

# Your code for 4th question here  
# Top Job Titles  
top\_job\_titles = df.groupBy("TITLE\_NAME") \  
 .agg(count("\*").alias("Number of Postings")) \  
 .orderBy(col("Number of Postings").desc()) \  
 .limit(10) \  
 .select(col("TITLE\_NAME").alias("Job Title"), col("Number of Postings"))  
  
fig = px.bar(  
 top\_job\_titles,   
 x='Job Title',   
 y='Number of Postings',  
 title='Top 10 Most Frequently Posted Job Titles',  
 labels={  
 'Job Title': 'Job Title',   
 'Number of Postings': 'Number of Job Postings'  
 },  
 color='Job Title',  
 color\_discrete\_sequence=px.colors.qualitative.Pastel,  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 height=500,  
 xaxis\_tickangle=-45  
)  
  
fig.update\_xaxes(  
 title\_standoff=25  
)  
  
fig.update\_yaxes(  
 title\_standoff=25  
)  
  
  
pio.write\_html(fig, file='top\_10\_job\_titles.html')  
fig.show()

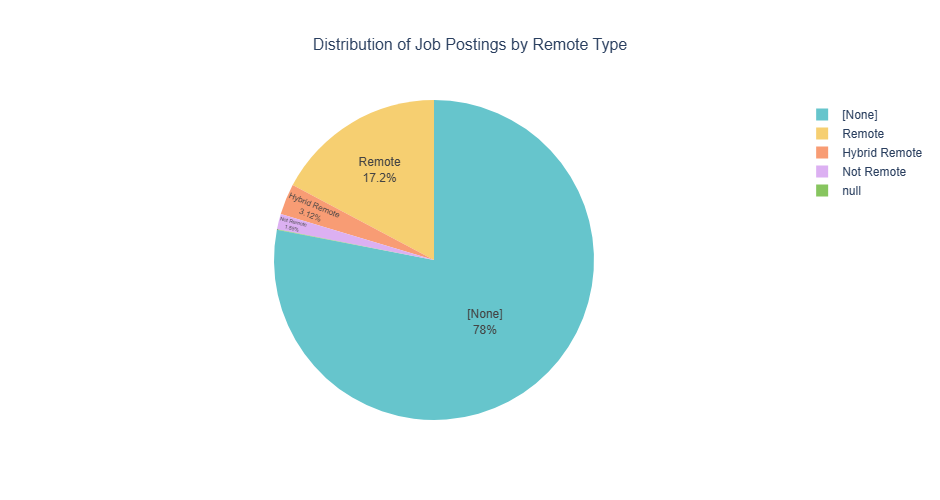


The bar chart shows Data Analysts as the most in-demand job title, with significantly more job postings compared to other roles, followed by Unclassified positions and Business Intelligence Analysts. The visualization highlights the strong market demand for data-related professional roles, with data analysis, business intelligence, and enterprise architecture positions dominating the top of the job posting rankings.

# 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 Type Counts  
remote\_type\_counts = df.groupBy("REMOTE\_TYPE\_NAME") \  
 .agg(count("\*").alias("Number of Postings")) \  
 .select(col("REMOTE\_TYPE\_NAME").alias("Remote Type"), col("Number of Postings"))  
  
fig = px.pie(  
 remote\_type\_counts,   
 values='Number of Postings',   
 names='Remote Type',  
 title='Distribution of Job Postings by Remote Type',  
 color\_discrete\_sequence=px.colors.qualitative.Pastel,  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 height=500  
)  
  
fig.update\_traces(textposition='inside', textinfo='percent+label')  
pio.write\_html(fig, file='remote\_type\_distribution.html')  
fig.show()



The pie chart shows that a substantial majority of job postings (78.1%) are categorized as None, which likely indicates there are not remote jobs, while only 17.2% of postings are explicitly labeled as “Remote” and a small percentage (3.7%) are classified as “Hybrid Remote”. This distribution suggests traditional work arrangements still dominating the job posting landscape.

# 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.

df\_with\_skills = df.withColumn("SKILLS\_NAME",   
 when(col("SKILLS\_NAME").isNull(), lit(""))  
 .otherwise(col("SKILLS\_NAME"))  
)  
  
# Explode skills and filter out empty skills  
skills\_by\_industry = df\_with\_skills.select(  
 "NAICS2\_NAME",   
 explode(split(col("SKILLS\_NAME"), ",")).alias("SKILLS\_NAME")  
)  
  
# Top Skills  
top\_skills = skills\_by\_industry.filter(col("SKILLS\_NAME") != "") \  
 .groupBy("SKILLS\_NAME") \  
 .agg(count("\*").alias("skill\_count")) \  
 .orderBy(col("skill\_count").desc()) \  
 .limit(5)  
  
top\_skills\_list = [row.SKILLS\_NAME for row in top\_skills.collect()]  
  
skills\_summary = skills\_by\_industry.filter(col("SKILLS\_NAME").isin(top\_skills\_list)) \  
 .groupBy("NAICS2\_NAME") \  
 .pivot("SKILLS\_NAME", top\_skills\_list) \  
 .agg(count("\*")) \  
 .na.fill(0)   
  
skills\_summary\_pd = skills\_summary.toPandas()  
  
skills\_summary\_melted = skills\_summary\_pd.melt(  
 id\_vars=['NAICS2\_NAME'],   
 var\_name='Skills',   
 value\_name='Number of Job Postings'  
)  
  
fig = px.bar(  
 skills\_summary\_melted,   
 x='NAICS2\_NAME',   
 y='Number of Job Postings',  
 color='Skills',  
 title='Top Skills Demand Across Industries',  
 labels={'NAICS2\_NAME': 'Industry', 'Number of Job Postings': 'Number of Job Postings'},  
 color\_discrete\_sequence=px.colors.qualitative.Pastel,  
 template='plotly\_white'  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 height=600,  
 xaxis\_tickangle=-45,  
 legend\_title\_text='Skills'  
)  
  
pio.write\_html(fig, file='skills\_by\_industry.html')  
fig.show()

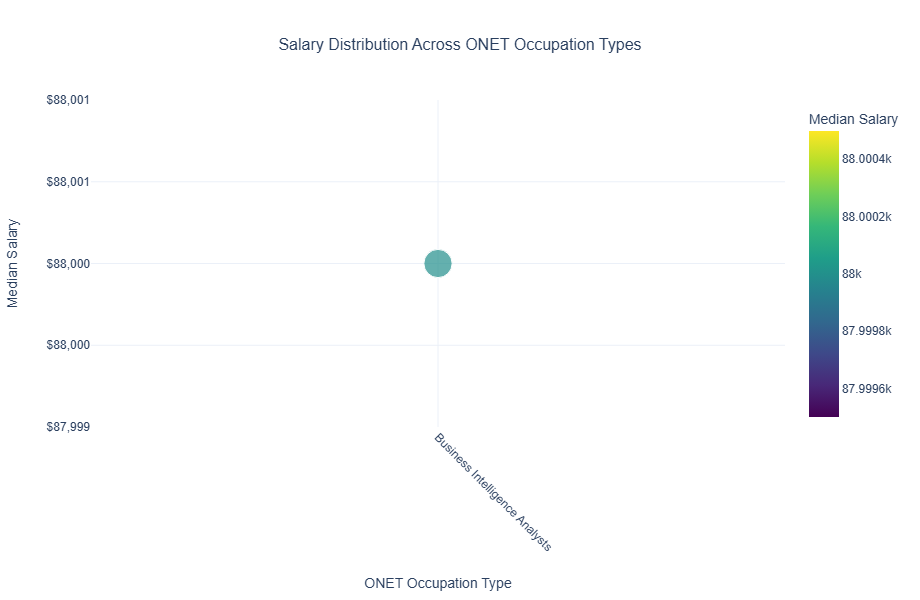
Unable to display output for mime type(s): application/vnd.plotly.v1+json

The graph reveals that Professional, Scientific, and Technical Services have the highest number of job postings, with a particularly strong demand for skills like Communication, Data Analysis, and Management. Across different industries, there’s a consistent need for core skills such as Communication and Management, though the relative importance and volume of skills vary significantly between sectors like Public Administration, Real Estate, and Manufacturing.

# 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 Salary Summary  
onet\_salary\_summary = df\_filtered.groupBy("ONET\_NAME") \  
 .agg(  
 F.percentile\_approx("SALARY\_FROM", 0.5).alias("Median\_Salary"),  
 count("\*").alias("Job\_Postings")  
 )  
fig = px.scatter(  
 onet\_salary\_summary,   
 x='ONET\_NAME',   
 y='Median\_Salary',  
 size='Job\_Postings',  
 color='Median\_Salary',  
 title='Salary Distribution Across ONET Occupation Types',  
 labels={  
 'ONET\_NAME': 'ONET Occupation Type',   
 'Median\_Salary': 'Median Salary',  
 'Job\_Postings': 'Number of Job Postings'  
 },  
 color\_continuous\_scale='Viridis',  
 template='plotly\_white',  
 hover\_name='ONET\_NAME',  
 hover\_data=['Job\_Postings']  
)  
  
fig.update\_layout(  
 font=dict(family="Arial, sans-serif", size=12),  
 title\_font\_size=16,  
 title\_x=0.5,  
 xaxis\_title\_font\_size=14,  
 yaxis\_title\_font\_size=14,  
 height=600,  
 width=1000  
)  
  
fig.update\_xaxes(  
 tickangle=45,  
 title\_standoff=25  
)  
  
fig.update\_yaxes(  
 title\_standoff=25,  
 tickformat='$,.0f'  
)  
  
pio.write\_html(fig, file='onet\_salary\_bubble\_chart.html')  
fig.show()



The plot indicateds that the dataset has only one onet\_name - Business Intelligence Analysts.

# 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  
# Count transitions  
transitions = df.groupBy(['SOC\_2021\_2\_NAME', 'SOC\_2021\_3\_NAME']) \  
 .agg(count('\*').alias('Value')) \  
 .toPandas()  
  
# Get unique labels  
source\_labels = transitions['SOC\_2021\_2\_NAME'].unique()  
target\_labels = transitions['SOC\_2021\_3\_NAME'].unique()  
  
source\_labels = source\_labels[source\_labels != None]  
target\_labels = target\_labels[target\_labels != None]  
  
all\_labels = np.unique(np.concatenate([source\_labels, target\_labels]))  
  
# Create label to index mapping  
label\_to\_index = {label: index for index, label in enumerate(all\_labels)}  
  
# Add source and target indices  
transitions['source\_index'] = transitions['SOC\_2021\_2\_NAME'].map(label\_to\_index)  
transitions['target\_index'] = transitions['SOC\_2021\_3\_NAME'].map(label\_to\_index)  
  
# Create Sankey Diagram  
fig = go.Figure(data=[go.Sankey(  
 node = dict(  
 pad = 15,  
 thickness = 20,  
 line = dict(color = "black", width = 0.5),  
 label = all\_labels,  
 color = "blue"  
 ),  
 link = dict(  
 source = transitions['source\_index'],  
 target = transitions['target\_index'],  
 value = transitions['Value']  
 ))])  
  
# Update layout  
fig.update\_layout(  
 title\_text="Career Pathway Transitions between SOC Job Classifications",   
 title\_font\_size=16,  
 title\_x=0.5,  
 font=dict(  
 family="Arial, sans-serif",  
 size=12,  
 color="black"  
 ),  
 height=600,  
 width=1000,  
 template='plotly\_white'  
)  
  
# Save and show the figure  
pio.write\_html(fig, file='career\_pathway\_sankey.html')  
fig.show()

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

It indicates that all Computer and Mathematical Occupations in 2021\_2 changed to Mathematical Science Occupations. Perhaps the data were incorrectly entered with the same values, or a name change occurred.