Module 04: Lab 01

Visual Reporting and Storytelling

Guowei Yan

November 21, 2024

# Objectives

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

# Step 1: Load the Dataset

import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
pio.renderers.default = "vscode"  
from pyspark.sql import SparkSession  
from pyspark.sql.functions import col  
  
# Start a Spark session  
spark = SparkSession.builder.appName("JobPostingsAnalysis").getOrCreate()  
  
# Load the CSV file into a Spark DataFrame  
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine", "true").option("escape", "\"").csv("lightcast\_job\_postings.csv")

Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/03/25 02:34:40 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
25/03/25 02:34:41 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

# 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.

#- Remove records where \*\*salary is missing or zero\*\*.  
df\_filtered = df.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") > 0))  
df\_filtered.show(5)  
pdf = df\_filtered.select("EMPLOYMENT\_TYPE\_NAME", "SALARY\_FROM").toPandas()  
fig = px.box(pdf, x="EMPLOYMENT\_TYPE\_NAME", y="SALARY\_FROM", title="Salary Distribution by Employment Type", color\_discrete\_sequence=["#636EFA"])  
fig.update\_layout(font\_family="Arial", title\_font\_size=16)  
fig.show()

25/03/25 02:35:01 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'.

+--------------------+-----------------+----------------------+----------+----------+----------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+------------------+----------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| 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|  
+--------------------+-----------------+----------------------+----------+----------+----------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+------------------+----------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
|cb5ca25f02bdf25c1...| 2024-06-19| 2024-06-19 07:00:00| 0|2024-06-02|2024-06-17| 15|[\n "FreeJobBoar...|[\n "craigslist....|[\n "https://mod...| []| NULL|Comisiones de $10...|Comisiones de $10...| 2024-06-17| 15| 0| Unclassified| LH/GM| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 3|Part-time / full-...| NULL| NULL| false| 92500| 0| [None]| year| 150000| 35000|{\n "lat": 37.63...| TW9kZXN0bywgQ0E=| Modesto, CA| 6099| Stanislaus, CA|33700| Modesto, CA| 6|California| 6099| Stanislaus, CA| 6099| Stanislaus, CA| 33700| Modesto, CA| 33700| Modesto, CA| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET0000000000000000| Unclassified|comisiones de por...| []| []| []| []| []| []| []| []| []| []|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...|  
|35a6cd2183d9fb270...| 2024-09-06| 2024-09-06 20:32:...| 0|2024-06-02|2024-06-12| 10| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL|SR Lead Data Analyst|About Lumen\n\nLu...| 2024-06-12| 10| 2233642|Lumen Technologies| Lumen| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| NULL| NULL| false|110155| 1| Remote| year| 125890| 94420|{\n "lat": 0,\n ...|W1Vua25vd24gQ2l0e...|[Unknown City], AR| 5999|[Unknown county], AR| NULL| NULL| 5| Arkansas| 5999|[Unknown county], AR| 5999|[Unknown county], AR| NULL| NULL| NULL| NULL| 51| Information| 517| Telecommunications| 5178|All Other Telecom...| 51781|All Other Telecom...|517810|All Other Telecom...|ET95DB859B53CCACA7| Lead Data Analysts|sr lead data analyst|[\n "KS13USA80NE...|[\n "Power BI",\...|[\n "KS13USA80NE...| [\n "Power BI",\...| []| []|[\n "KS1280B68GD...|[\n "Presentatio...|[\n "KS13USA80NE...|[\n "Power BI",\...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| [\n "52.0201"\n]|[\n "Business Ad...| [\n "52.02"\n]|[\n "Business Ad...| [\n "52"\n]|[\n "Business, M...| 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| 51| Information| 517| Telecommunications| 5178|All Other Telecom...| 51781|All Other Telecom...| 517810|All Other Telecom...|  
|229620073766234e8...| 2024-10-09| 2024-10-09 18:07:...| 0|2024-06-02|2024-08-01| NULL| [\n "Company"\n]| [\n "3ds.com"\n]|[\n "https://www...| []| NULL|Sr. Marketing Ana...|Sr. Marketing Ana...| 2024-08-01| NULL|39016169| Dassault Systèmes|Dassault Systmes| false| [\n 2,\n 3\n]| [\n "Bachelor's ...| 2| Bachelor's degree| 3| Master's degree| 1|Full-time (> 32 h...| 2| 2| false| 92962| 0| [None]| year| 106424| 79500|{\n "lat": 40.75...| TmV3IFlvcmssIE5Z| New York, NY| 36061| New York, NY|35620|New York-Newark-J...| 36| New York| 36061| New York, NY| 36061| New York, NY| 35620|New York-Newark-J...| 35620|New York-Newark-J...| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...|541511|Custom Computer P...|ET1CE3CFA5447376E9| Marketing Analysts|sr marketing analyst|[\n "KS4407N6CMT...|[\n "Salesforce"...|[\n "KS4407N6CMT...| [\n "Salesforce"...| []| []|[\n "KS7G747655V...|[\n "Prioritizat...|[\n "KS4407N6CMT...|[\n "Salesforce"...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "52.0101",\n...|[\n "Business/Co...|[\n "52.01",\n ...|[\n "Business/Co...|[\n "52",\n "45...|[\n "Business, M...| 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 7\n]| [\n "Artificial ...| 54|Professional, Sci...| 541|Professional, Sci...| 5415|Computer Systems ...| 54151|Computer Systems ...| 541511|Custom Computer P...|  
|b7aa80a24c82f080c...| 2024-09-28| 2024-09-28 14:06:...| 8|2024-06-02|2024-09-27| NULL|[\n "Government"...|[\n "dcscorp.com...|[\n "https://www...| []| NULL| Data Analyst|Data Analyst In R...| 2024-07-13| 41|12147696| DCS Corporation| DCS Corp.| false|[\n 0,\n 1,\n ...| [\n "High school...| 0| High school or GED| 2| Bachelor's degree| 1|Full-time (> 32 h...| 10| NULL| false|107645| 2| Not Remote| year| 123732| 91559|{\n "lat": 35.62...|UmlkZ2VjcmVzdCwgQ0E=| Ridgecrest, CA| 6029| Kern, CA|12540| Bakersfield, CA| 6|California| 6029| Kern, CA| 6029| Kern, CA| 12540| Bakersfield, CA| 12540|Bakersfield-Delan...| 42| Wholesale Trade| 423|Merchant Wholesal...| 4238|Machinery, Equipm...| 42383|Industrial Machin...|423830|Industrial Machin...|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS128HD6KJS...|[\n "Regression ...|[\n "KS128HD6KJS...| [\n "Regression ...|[\n "KS683TN76T7...|[\n "Security Cl...|[\n "KS1203C6N9B...|[\n "Research",\...|[\n "KS125LS6N7W...|[\n "Python (Pro...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "14.0101",\n...|[\n "Engineering...|[\n "14.01",\n ...|[\n "Engineering...|[\n "14",\n "14...|[\n "Engineering...| 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| 42| Wholesale Trade| 423|Merchant Wholesal...| 4238|Machinery, Equipm...| 42383|Industrial Machin...| 423830|Industrial Machin...|  
|57b527ea0f91db5bb...| 2024-09-06| 2024-09-06 20:32:...| 0|2024-06-02|2024-07-27| 55| [\n "Job Board"\n]|[\n "simplyhired...|[\n "https://www...| []| NULL|Power, Utilities ...|Power, Utilities ...| 2024-07-27| 55| 5732448| Deloitte| Deloitte| false| [\n 2,\n 3\n]| [\n "Bachelor's ...| 2| Bachelor's degree| 3| Master's degree| 1|Full-time (> 32 h...| 6| NULL| false|192800| 0| [None]| year| 241000| 144600|{\n "lat": 42.33...| RGV0cm9pdCwgTUk=| Detroit, MI| 26163| Wayne, MI|19820|Detroit-Warren-De...| 26| Michigan| 26163| Wayne, MI| 26163| Wayne, MI| 19820|Detroit-Warren-De...| 19820|Detroit-Warren-De...| 54|Professional, Sci...| 541|Professional, Sci...| 5416|Management, Scien...| 54161|Management Consul...|541611|Administrative Ma...|ET8AEDEB1F4C3091D3|Management Consul...|power utilities r...|[\n "KS122VL71WF...|[\n "Design Spec...|[\n "KS122VL71WF...| [\n "Design Spec...| []| []|[\n "KS1218W78FG...|[\n "Management"...|[\n "KS1219W70LY...|[\n "C++ (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| [\n "45.0702"\n]|[\n "Geographic ...| [\n "45.07"\n]|[\n "Geography a...| [\n "45"\n]|[\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 3\n]| [\n "Green Jobs:...| 54|Professional, Sci...| 541|Professional, Sci...| 5416|Management, Scien...| 54161|Management Consul...| 541611|Administrative Ma...|  
+--------------------+-----------------+----------------------+----------+----------+----------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+------------------+----------------+-------------------+--------------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+------------------+------+--------------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
only showing top 5 rows

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

The box plot shows the salary distribution across different employment types. It reveals that full-time employees tend to have a wider salary range with higher outliers, while part-time employees generally have lower median salaries with less variation.

# 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.

pdf = df\_filtered.select("NAICS2\_NAME", "SALARY\_FROM").toPandas()  
fig = px.box(pdf, x="NAICS2\_NAME", y="SALARY\_FROM", title="SALARY\_FROM Distribution by NAICS2\_NAME", color\_discrete\_sequence=["#FF5733"])  
fig.update\_layout(font\_family="Arial", title\_font\_size=20)

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

This box plot illustrates the salary distribution across different industries, showing variations in median salaries and salary ranges. Some industries, such as Information and Professional, Scientific, and Technical Services, exhibit higher median salaries and greater variability, while others, like Accommodation and Food Services, tend to have lower median salaries with less dispersion.

# 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.

from pyspark.sql.functions import col, count  
print(df.columns)  
df\_postings = df.groupBy("POSTED").agg(count("\*").alias("Number\_of\_Job\_Postings"))  
df\_postings.show(10)

['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']  
+----------+----------------------+  
| POSTED|Number\_of\_Job\_Postings|  
+----------+----------------------+  
|2024-09-18| 700|  
|2024-05-30| 517|  
|2024-06-12| 708|  
|2024-08-27| 629|  
|2024-06-04| 453|  
|2024-05-25| 459|  
|2024-08-30| 540|  
|2024-07-08| 426|  
|2024-09-10| 515|  
|2024-08-05| 480|  
+----------+----------------------+  
only showing top 10 rows

[Stage 5:> (0 + 1) / 1]

pdf = df\_postings.select("POSTED", "Number\_of\_Job\_Postings").toPandas()  
pdf = df\_postings.orderBy("POSTED").toPandas()  
import plotly.express as px  
fig = px.line(pdf, x="POSTED", y="Number\_of\_Job\_Postings",  
 title="Job postings fluctuate over time",  
 color\_discrete\_sequence=["#FF5733"])  
  
fig.update\_layout(font\_family="Arial", title\_font\_size=20)  
fig.show()

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

This graph illustrates the fluctuations in job postings over time, showing significant variability in the number of postings per day. There are noticeable peaks and troughs, suggesting periods of increased and decreased hiring activity, possibly influenced by seasonal trends or specific industry demands

# 4. Top 10 Job Titles by Count

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

df\_Job\_Titles = df.groupBy("TITLE\_NAME").agg(count("\*").alias("Job Count"))  
top\_10\_titles = df\_Job\_Titles.orderBy(col("Job Count").desc()).limit(10)  
top\_10\_titles.show()

[Stage 19:> (0 + 1) / 1]

+--------------------+---------+  
| TITLE\_NAME|Job Count|  
+--------------------+---------+  
| Data Analysts| 8593|  
| Unclassified| 3151|  
|Business Intellig...| 2074|  
|Enterprise Archit...| 1999|  
|Oracle Cloud HCM ...| 1042|  
| Data Modelers| 668|  
|Data Governance A...| 629|  
|Data Analytics En...| 537|  
|ERP Business Anal...| 488|  
|Data Quality Anal...| 467|  
+--------------------+---------+

pdf = top\_10\_titles.select("TITLE\_NAME", "Job Count").toPandas()  
import plotly.express as px  
fig = px.bar(pdf, x="TITLE\_NAME", y="Job Count",  
 title="Job postings fluctuate over time",  
 color\_discrete\_sequence=["#FF5733"])  
  
fig.update\_layout(font\_family="Arial", title\_font\_size=20)  
fig.show()

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

From the chart, it is evident that “Data Analysts” is the most common job title, significantly surpassing others, with over 3,000 job postings. Following that, “Business Intelligence Analysts” and “Unclassified” are also relatively common but considerably lower than Data Analysts, while the remaining job titles have similar posting counts, indicating a high demand for data-related roles.

# 5. Remote vs On-Site Job Postings

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

df\_5 = df.groupBy("REMOTE\_TYPE\_NAME").agg(count("\*").alias("Job Count"))  
df\_5.show(10)

[Stage 25:> (0 + 1) / 1]

+----------------+---------+  
|REMOTE\_TYPE\_NAME|Job Count|  
+----------------+---------+  
| Remote| 12499|  
| [None]| 56584|  
| Not Remote| 1133|  
| Hybrid Remote| 2260|  
+----------------+---------+

pdf = df\_5.select("REMOTE\_TYPE\_NAME", "Job Count").toPandas()  
import plotly.express as px  
  
fig = px.pie(pdf,   
 names="REMOTE\_TYPE\_NAME",   
 values="Job Count",  
 title="Job Postings by Remote Type",  
 color\_discrete\_sequence=["#FF5733", "#FFBD33", "#33FF57", "#3357FF"])  
  
fig.update\_layout(font\_family="Times New Roman", title\_font\_size=20)  
fig.show()

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

The pie chart shows that 78.1% of job postings do not specify remote type, while only 17.2% are remote and 3.12% are hybrid. Remote-friendly roles remain a minority.

# 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.

pdf = df\_filtered.select("SKILLS", "NAICS2").toPandas()  
pdf\_exploded = pdf.assign(SKILL=pdf["SKILLS"].str.split(",")).explode("SKILL")  
pdf\_exploded["SKILL"] = pdf\_exploded["SKILL"].str.strip()  
pdf\_exploded = pdf\_exploded.dropna(subset=["SKILL", "NAICS2"])  
  
skill\_counts = pdf\_exploded.groupby(["NAICS2", "SKILL"]).size().reset\_index(name="Count")  
  
import plotly.express as px  
fig = px.bar(  
 skill\_counts,  
 x="NAICS2",  
 y="Count",  
 color="SKILL",  
 title="Skill Demand by Industry",  
 text\_auto=True,  
 color\_discrete\_sequence=px.colors.qualitative.Set2  
)  
fig.update\_layout(  
 barmode="stack",  
 font\_family="Arial",  
 title\_font\_size=20,  
 xaxis\_title="Industry (NAICS2)",  
 yaxis\_title="Skill Count"  
)  
fig.show()

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

The chart shows the distribution of skill demand across different industries based on NAICS codes. Most industries have a high concentration of skill requests around NAICS code 54, indicating that this industry category (likely professional, scientific, and technical services) requires a broad variety of skills.

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

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

from pyspark.sql.functions import col, expr, count, percentile\_approx  
agg\_df = df\_filtered.groupBy("ONET\_NAME").agg(  
 percentile\_approx("SALARY\_FROM", 0.5).alias("Median\_Salary"),  
 count("\*").alias("Job\_Postings")  
)  
pdf = agg\_df.toPandas()   
import plotly.express as px  
  
fig = px.scatter(  
 pdf,  
 x="ONET\_NAME",  
 y="Median\_Salary",  
 size="Job\_Postings",  
 title="Salary Analysis by ONET Occupation Type",  
 color="Median\_Salary",  
 color\_continuous\_scale="Viridis"  
)  
  
fig.update\_layout(  
 font\_family="Arial",  
 title\_font\_size=20,  
 xaxis\_title="Occupation (ONET)",  
 yaxis\_title="Median Salary",  
 xaxis\_tickangle=-45  
)  
  
fig.show()

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

The chart shows that Business Intelligence Analysts have a median salary of around $88,000. This occupation also has a high number of job postings, indicating strong demand in the job market.

# 8. Career Pathway Trends (Sankey Diagram)

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

df\_sankey = df.groupBy("SOC\_2021\_2\_NAME", "SOC\_2021\_3\_NAME").agg(  
 count("\*").alias("Count")  
)  
  
pdf\_sankey = df\_sankey.toPandas()  
  
nodes = list(set(pdf\_sankey["SOC\_2021\_2\_NAME"]) | set(pdf\_sankey["SOC\_2021\_3\_NAME"]))  
node\_indices = {name: i for i, name in enumerate(nodes)}  
  
pdf\_sankey["source\_id"] = pdf\_sankey["SOC\_2021\_2\_NAME"].map(node\_indices)  
pdf\_sankey["target\_id"] = pdf\_sankey["SOC\_2021\_3\_NAME"].map(node\_indices)  
import plotly.graph\_objects as go  
  
fig = go.Figure(data=[go.Sankey(  
 node=dict(  
 pad=15,  
 thickness=20,  
 line=dict(color="black", width=0.5),  
 label=nodes,  
 color="lightblue"  
 ),  
 link=dict(  
 source=pdf\_sankey["source\_id"],  
 target=pdf\_sankey["target\_id"],  
 value=pdf\_sankey["Count"]  
 )  
)])  
  
fig.update\_layout(  
 title\_text="Career Pathway Trends",  
 font=dict(size=12, family="Arial")  
)  
  
fig.show()

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

import os  
os.makedirs("\_output", exist\_ok=True)  
fig.write\_image("\_output/figure.svg")

[Salary Distribution](_output/figure.svg)