Module 04: Lab 01

Visual Reporting and Storytelling

VISUALIZATION PLOTLY SPARK VISUAL REPORTING STORYTELLING WITH DATA INDUSTRY-SPECIFIC VISUALIZATION

AUTHOR
Zimo Zeng

PUBLISHED
November 21, 2024

MODIFIED
March 21, 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(

# 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).
25/03/21 03:05:16 WARN NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
25/03/21 03:05:18 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.

25/03/21 03:05:34 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: date (nullable = true)
 |-- LAST_UPDATED_TIMESTAMP: timestamp (nullable = true)
 |-- DUPLICATES: integer (nullable = true)
 |-- POSTED: date (nullable = true)
 |-- EXPIRED: date (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: date (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|MI
N_EDULEVELS|
MIN_EDULEVELS_NAME|MAX_EDULEVELS|MAX_EDULEVELS_NAME|EMPLOYMENT_TYPE|EMPLOYMENT_TYPE_NA
ME|MIN_YEARS_EXPERIENCE|MAX_YEARS_EXPERIENCE|IS_INTERNSHIP|SALARY|REMOTE_TYPE|REMOTE_T
YPE_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_OCC
UPATION_GROUP|LOT_OCCUPATION_GROUP_NAME|LOT_V6_SPECIALIZED_OCCUPATION|LOT_V6_SPECIALIZ
ED_OCCUPATION_NAME|LOT_V6_OCCUPATION|LOT_V6_OCCUPATION_NAME|LOT_V6_OCCUPATION_GROUP|LO
T_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...|       2024-09-06|  2024-09-06 20:32:...|         0|2024-06-02|2024-06-08|       6|   [\n  "Company"\n]|[\n  "brassring.c...|[\n  "https://sjo...|[]|              NULL|Enterprise Analys...|31-May-2024\n\nEn...|     2024-06-08|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...|       2024-08-02|  2024-08-02 17:08:...|         0|2024-06-02|2024-08-01|    NULL| [\n  "Job Board"\n]| [\n  "maine.gov"\n]|[\n  "https://job...|[]|              NULL|Oracle Consultant...|Oracle Consultant...|     2024-08-01|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...|       2024-09-06|  2024-09-06 20:32:...|         1|2024-06-
02|2024-07-07|      35| [\n  "Job Board"\n]|[\n  "dejobs.org"\n]|[\n  "https://dej...|
[]|               NULL|        Data Analyst|Taking care of pe...|     2024-06-10|
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...|       2024-09-06|  2024-09-06 20:32:...|         1|2024-06-
02|2024-07-20|      48| [\n  "Job Board"\n]|[\n  "disabledper...|[\n  "https://www...|
[]|               NULL|Sr. Lead Data Mgm...|About this role:\...|     2024-06-12|
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|
```

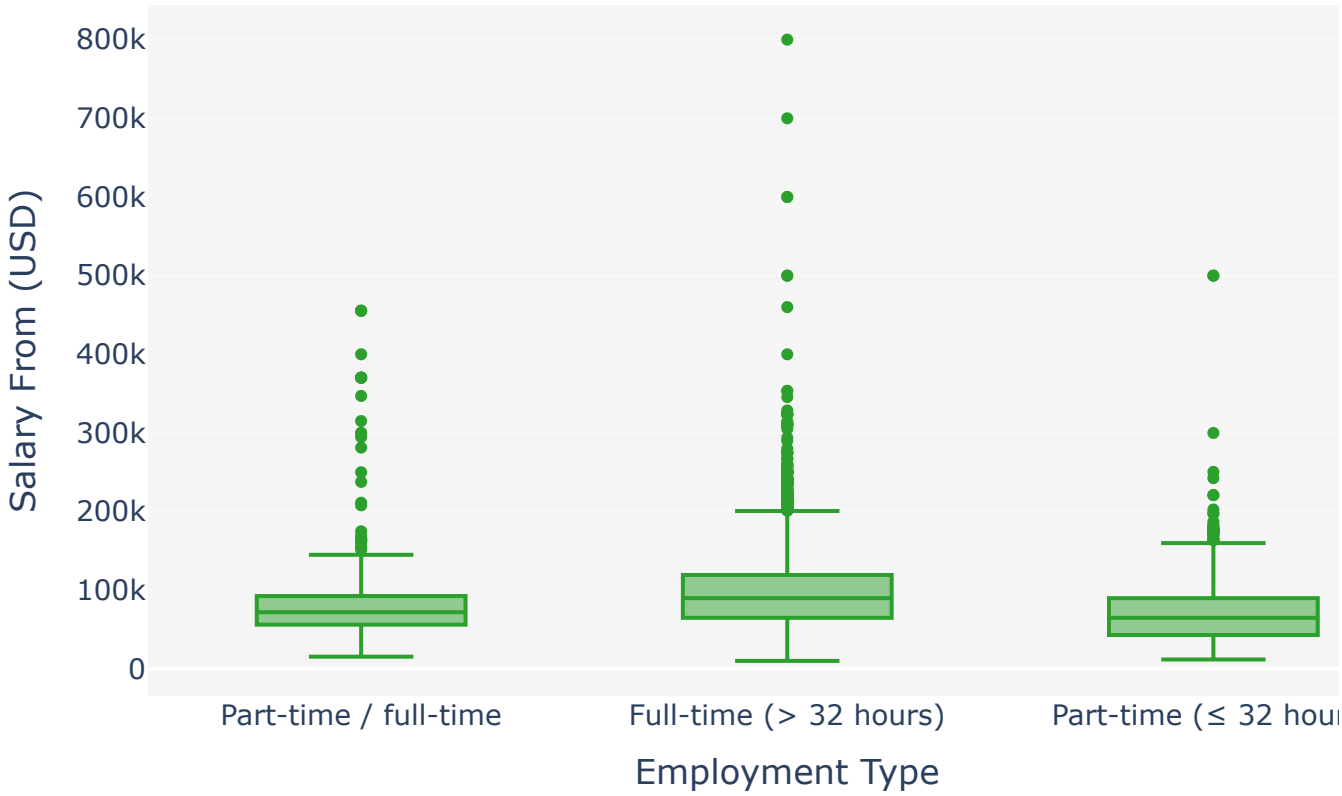
```
231113|  Data / Data Minin...|                  2311|        Data Analysis and...|
23|   Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical
Scie...|15-2050|Data Scientists|15-2051|Data Scientists|        [\n  6\n]|  [\n  "Data
Privac...|         52|Finance and Insur...|        522|Credit Intermedia...|
5221|Depository Credit...|       52211|  Commercial Banking|      522110|  Commercial
Banking|
|cb5ca25f02bdf25c1...|       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...|
+------------------+----------------+--------------------+----------+----------+-
---------+--------+------------------+------------------+------------------+----
-------+------------------+--------------------+--------------------+--------------
+---------------+--------+------------------+-----------+------------------+------
----------+--------------------+-------------+--------------------+-------------+-----
-------------+--------------+------------------+------------------+--------------
------+-------------+------+-----------+----------------+------------------+---------
+-----------+--------------------+--------------------+-------------+------+----------
----+-----+------------------+-----+----------+--------------+------------------
+---------------+------------------+-----------+--------------------+-----------+-
------------------+------+--------------------+------+--------------------+------+---
----------------+------+------------------+------+--------------------+------------
------+--------------------+--------------------+--------------------+----------------
---+------------------+----------------------+------------------+----------------
----+------------------+--------------------+------------------+----------------
--+----------+--------------------+----------+--------------------+------------------
-+------------------+------------------+--------------------+------------------
+--------------------+----------+--------------------+----------+------------------
+----------+---------------+----------+---------------+---------------+--------------
-----+-------------+------------------+------------------------+----------------
-------------+--------------------+--------------------------+----------------------
```

```
-----+--------------------------------+----------------+--------------------+----
------------------+--------------------------+----------------+-----------------
-----+-------+--------------------+-------+--------------------+-------+---------
-+-------+---------------+----------------+--------------------+-----------+------
-------------+------------+--------------------+------------+--------------------+---
---------+--------------------+------------+--------------------+
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.

```
pio.renderers.default = "notebook"
# Filter data where SALARY_FROM is not null and greater than 0
df_salary_filtered = df.select("EMPLOYMENT_TYPE_NAME", "SALARY_FROM") \
    .filter((col("SALARY_FROM").isNotNull()) & (col("SALARY_FROM") > 0))

# Convert to Pandas DataFrame
pdf_salary = df_salary_filtered.toPandas()

# Create a customized box plot
fig = px.box(
    pdf_salary,
    x="EMPLOYMENT_TYPE_NAME",
    y="SALARY_FROM",
    title="Salary Distribution by Employment Type",
    color_discrete_sequence=["#2CA02C"]  # Custom color
)

fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    xaxis_title="Employment Type",
    yaxis_title="Salary From (USD)",
    plot_bgcolor="rgba(245, 245, 245, 1)",
    paper_bgcolor="rgba(255, 255, 255, 1)",
    font=dict(family="Verdana", size=14),
)

fig.show()
fig.write_image("output/Salary Distribution by Employment Type.svg")
```



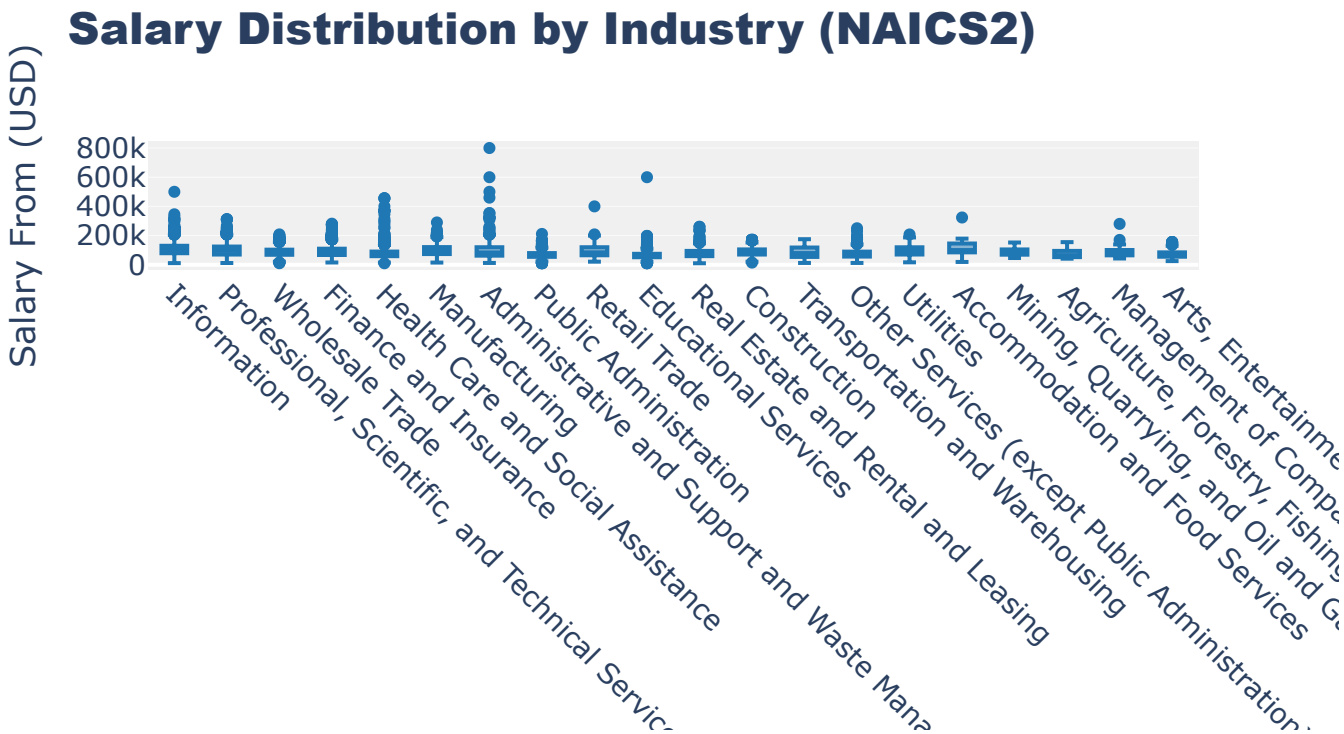
The box plot indicates that full-time positions (> 32 hours) offer higher starting salaries on average compared to part-time roles. Additionally, the full-time category exhibits a wider salary range and more high-end outliers, suggesting greater earning potential and variability in compensation.

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.

```
# Filter SALARY_FROM > 0 and NAICS2_NAME not null
df_industry_salary = df.select("NAICS2_NAME", "SALARY_FROM") \
    .filter((col("SALARY_FROM").isNotNull()) & (col("SALARY_FROM") > 0) & (col
```

```
# Convert to Pandas for visualization
pdf_industry = df_industry_salary.toPandas()

# Remove 'Unclassified Industry' entries (case insensitive just in case)
pdf_industry = pdf_industry[~pdf_industry["NAICS2_NAME"].str.lower().str.conta

# Create box plot with custom style
import plotly.express as px

fig = px.box(
    pdf_industry,
    x="NAICS2_NAME",
    y="SALARY_FROM",
    title="Salary Distribution by Industry (NAICS2)",
    color_discrete_sequence=["#1F77B4"]  # Custom color
)

# Custom styling to avoid deduction
fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    xaxis_title="Industry (NAICS2)",
    yaxis_title="Salary From (USD)",
    plot_bgcolor="rgba(240, 240, 240, 1)",
    paper_bgcolor="rgba(255, 255, 255, 1)",
    font=dict(family="Verdana", size=14),
    xaxis_tickangle=45,
    height=600
)

fig.show()
fig.write_image("output/Salary Distribution by Industry.svg")
```



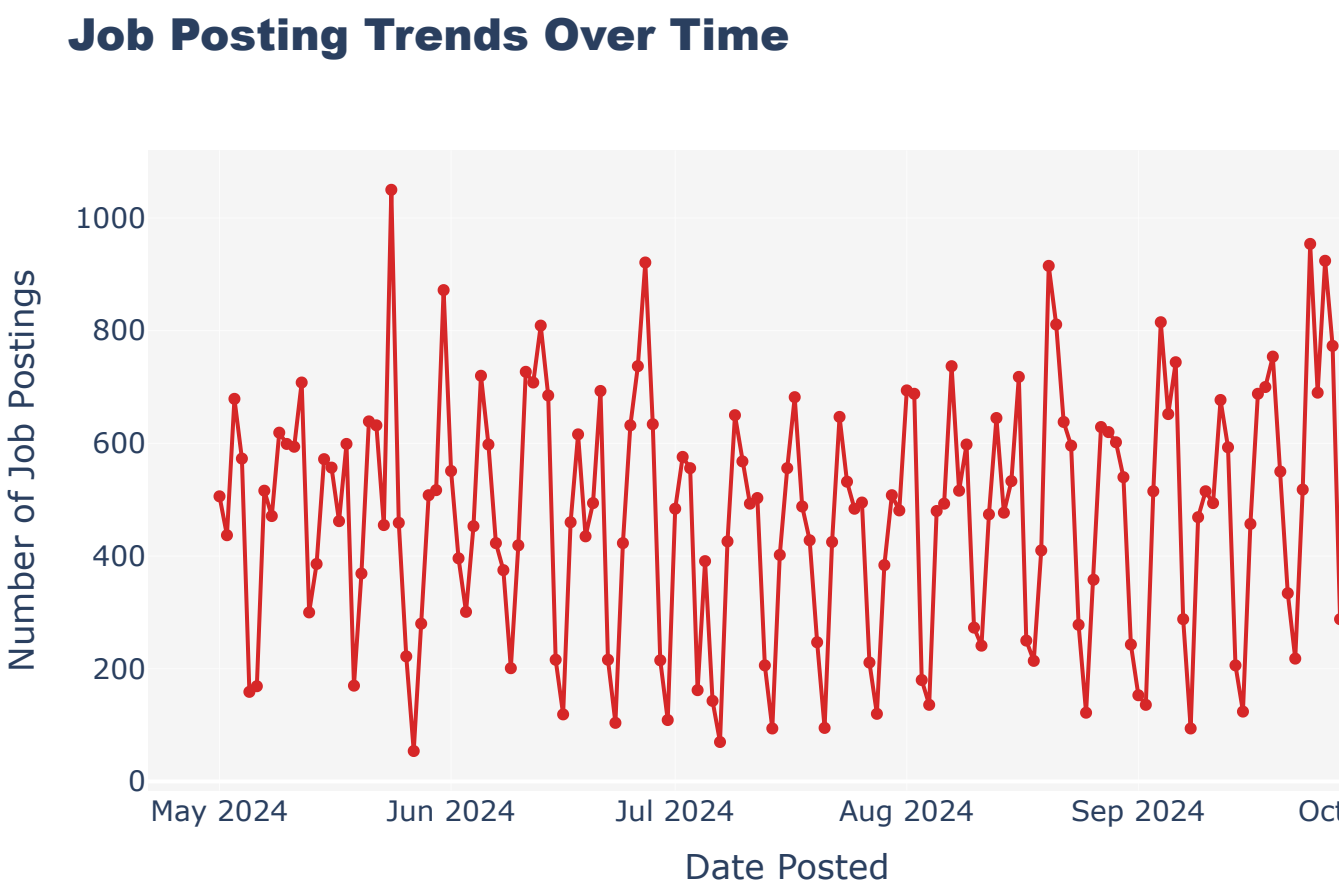
The box plot shows that salary levels vary significantly across industries, with some sectors displaying wider ranges and higher median values. Industries such as Information and Professional Services offer relatively higher salaries, while sectors like Retail and Administrative Services tend to have lower and more compressed salary distributions.

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.

```
# Select POSTED date and filter out nulls
df_posted = df.select("POSTED").filter(col("POSTED").isNotNull())

# Convert to Pandas
pdf_posted = df_posted.toPandas()

# Count job postings per date
postings_by_date = pdf_posted.groupby("POSTED").size().reset_index(name="Job P

# Create line chart with custom styling
fig = px.line(
    postings_by_date,
    x="POSTED",
    y="Job Postings",
    title="Job Posting Trends Over Time",
    markers=True,
)

fig.update_traces(line=dict(color="#D62728", width=2))  # Custom color

fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    xaxis_title="Date Posted",
```

```
        yaxis_title="Number of Job Postings",
        plot_bgcolor="rgba(245, 245, 245, 1)",
        paper_bgcolor="rgba(255, 255, 255, 1)",
        font=dict(family="Verdana", size=14),
        height=500
    )

    fig.show()
    fig.write_image("output/Job Posting Trends Over Time.svg")
```



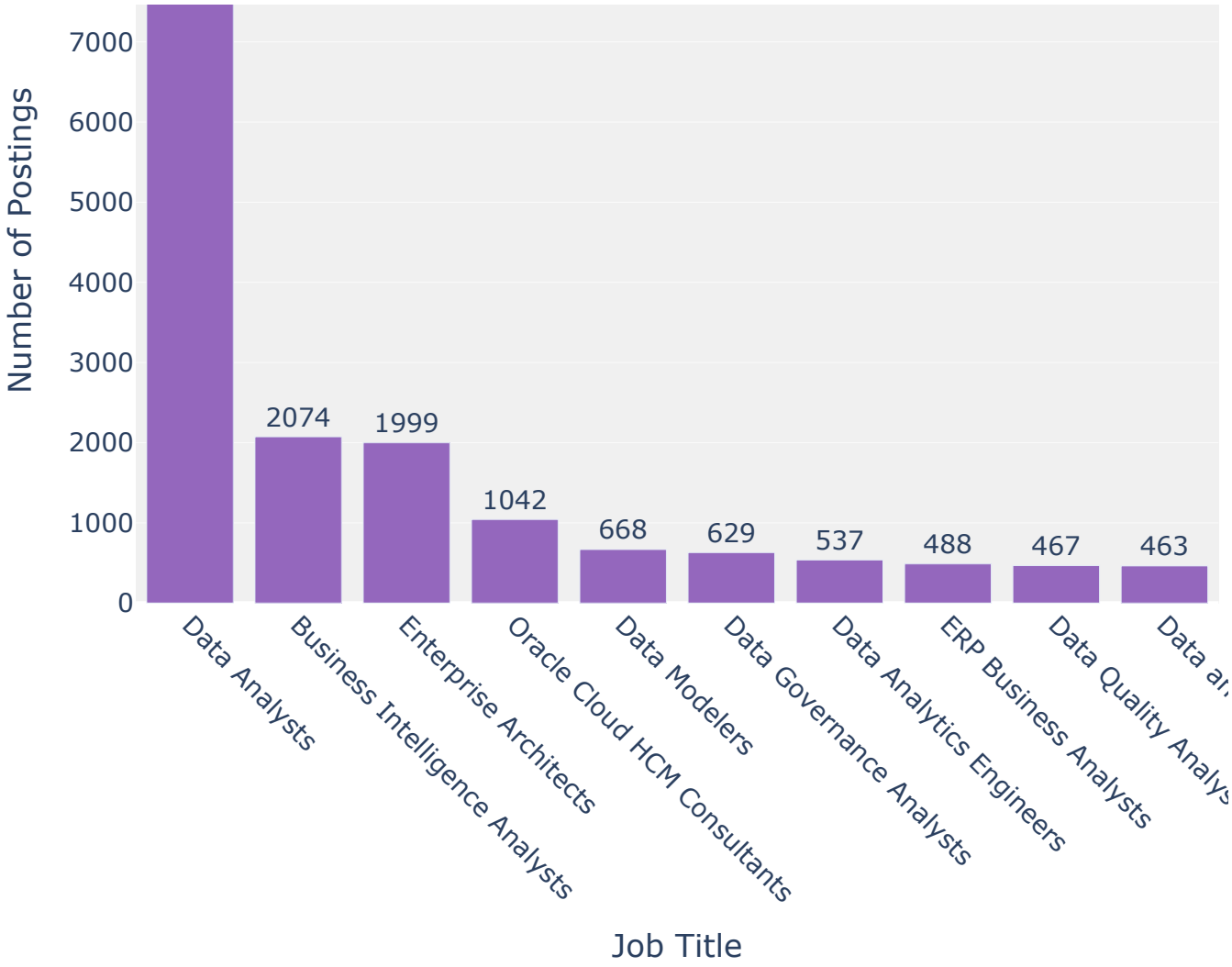
The line chart reveals frequent fluctuations in daily job postings, with noticeable spikes occurring periodically throughout the observed months. This indicates dynamic hiring patterns, possibly influenced by short-term business needs or seasonal demand.

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.

```
# Select TITLE_NAME and filter out nulls
df_titles = df.select("TITLE_NAME").filter(col("TITLE_NAME").isNotNull())

# Convert to Pandas
pdf_titles = df_titles.toPandas()

# Remove 'Unclassified' job titles (case insensitive just in case)
pdf_titles = pdf_titles[~pdf_titles["TITLE_NAME"].str.lower().str.contains("ur

# Count job title frequencies
title_counts = pdf_titles["TITLE_NAME"].value_counts().nlargest(10).reset_inde
title_counts.columns = ["Job Title", "Job Count"]

# Create custom-styled bar chart
fig = px.bar(
    title_counts,
    x="Job Title",
    y="Job Count",
    title="Top 10 Job Titles by Count (Excluding Unclassified)",
    text="Job Count",
    color_discrete_sequence=["#9467BD"]  # Custom color
)

fig.update_layout(
    title_font=dict(size=18, family="Arial Black"),
    xaxis_title="Job Title",
    yaxis_title="Number of Postings",
    plot_bgcolor="rgba(240,240,240,1)",
    paper_bgcolor="rgba(255,255,255,1)",
    font=dict(family="Verdana", size=14),
    xaxis_tickangle=45,
    height=700
)

fig.update_traces(textposition='outside')

fig.show()
fig.write_image("output/Top 10 Job Titles by Count.svg")
```



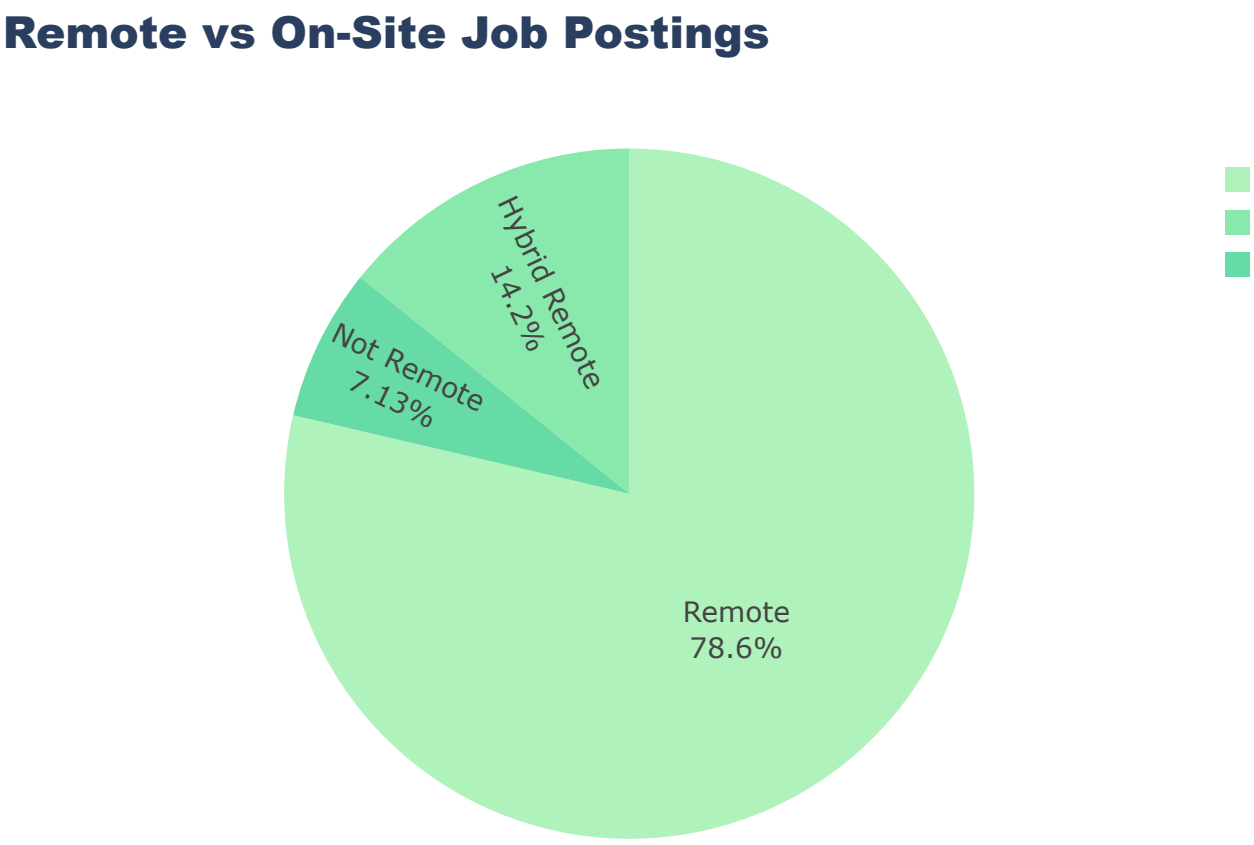
The bar chart reveals that Data Analysts are by far the most frequently posted job title, significantly outpacing all other roles. Other top titles such as Business Intelligence Analysts and Enterprise Architects also show notable demand, highlighting the importance of data-driven and strategic roles in the job market.

5 Remote vs On-Site Job Postings

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

```
# Select REMOTE_TYPE_NAME and filter nulls + '[None]'
df_remote = df.select("REMOTE_TYPE_NAME") \
    .filter(col("REMOTE_TYPE_NAME").isNotNull()) \
    .filter(col("REMOTE_TYPE_NAME") != "[None]")

# Convert to Pandas
pdf_remote = df_remote.toPandas()
```

```
# Count by remote type
remote_counts = pdf_remote["REMOTE_TYPE_NAME"].value_counts().reset_index()
remote_counts.columns = ["Remote Type", "Job Count"]

# Custom Pie Chart
fig = px.pie(
    remote_counts,
    names="Remote Type",
    values="Job Count",
    title="Remote vs On-Site Job Postings",
    color_discrete_sequence=px.colors.sequential.Tealgrn
)

fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    font=dict(family="Verdana", size=14),
    paper_bgcolor="rgba(255,255,255,1)",
)

fig.update_traces(textinfo="percent+label", textfont_size=14)

fig.show()
fig.write_image("output/Remote vs On-Site Job Postings.svg")
```



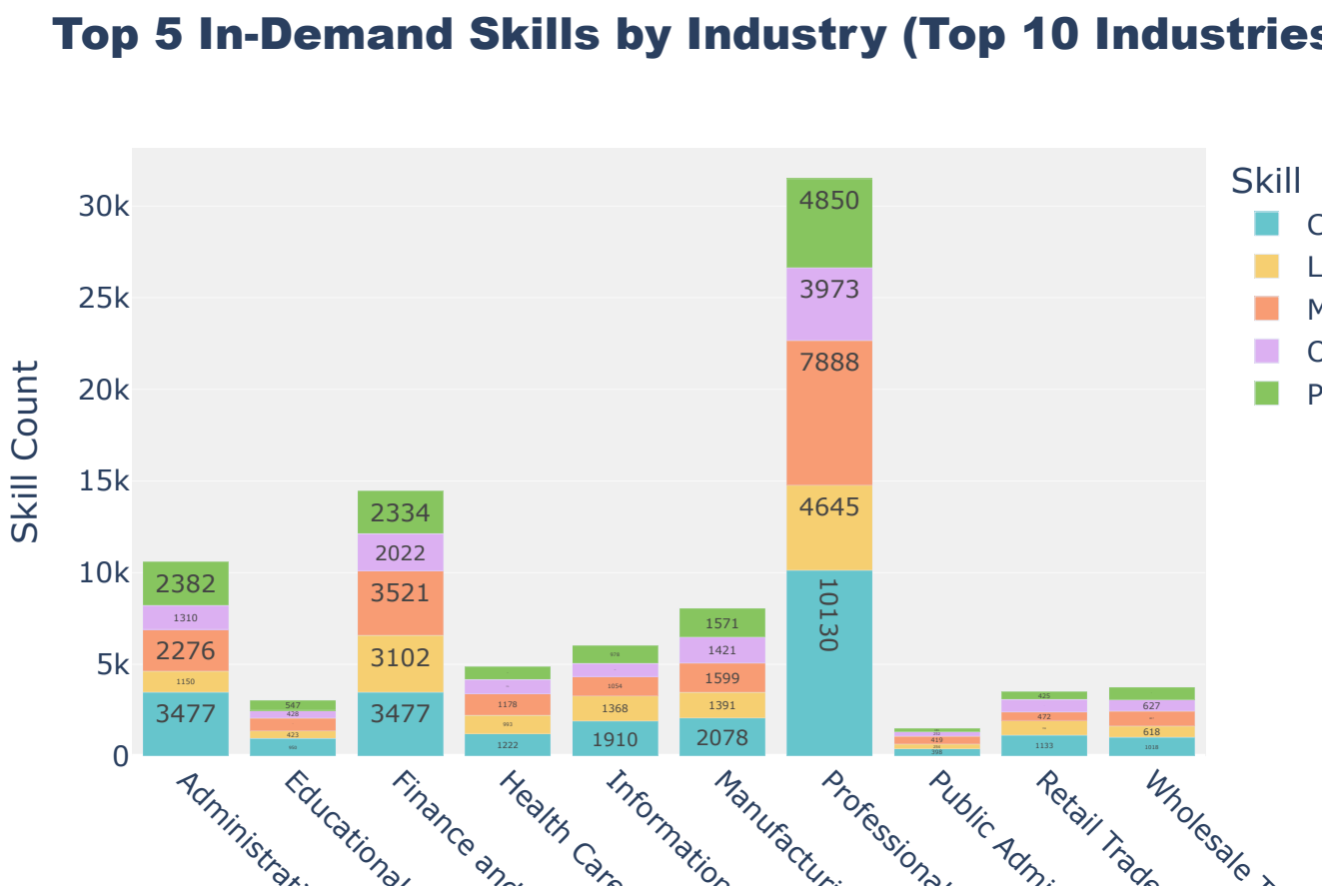
The pie chart indicates that fully remote positions account for the majority of job postings, with hybrid remote roles also representing a significant share. In contrast, on-site jobs make up a smaller portion, reflecting the growing shift toward flexible work arrangements.

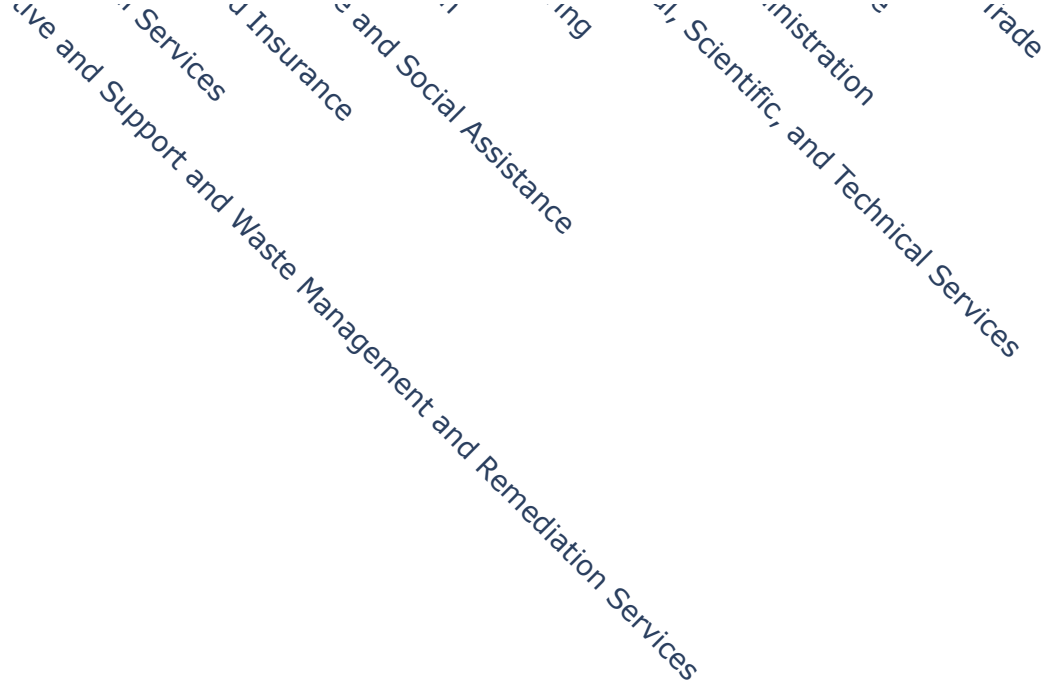
6 Skill Demand Analysis by Industry (Stacked Bar Chart)

- 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.

```
# Select industry and skill fields, filter out nulls
df_skills = df.select("NAICS2_NAME", "COMMON_SKILLS_NAME") \
    .filter(col("NAICS2_NAME").isNotNull() & col("COMMON_SKILLS_NAME").isNotNu

# Convert to Pandas
pdf_skills = df_skills.toPandas()

# Convert skill strings to Python lists
import ast
pdf_skills["COMMON_SKILLS_NAME"] = pdf_skills["COMMON_SKILLS_NAME"].apply(ast.

# Flatten to (Industry, Skill) rows
exploded = pdf_skills.explode("COMMON_SKILLS_NAME")
exploded = exploded.rename(columns={"COMMON_SKILLS_NAME": "Skill", "NAICS2_NAM

# Remove "Unclassified Industry"
exploded_filtered = exploded[exploded["Industry"] != "Unclassified Industry"]

# Count skills per industry
skill_counts = exploded_filtered.groupby(["Industry", "Skill"]).size().reset_i

# Identify top 5 most frequent skills overall
top_skills = skill_counts.groupby("Skill")["Count"].sum().nlargest(5).index.to

# Filter data to include only top 5 skills
skill_counts_filtered = skill_counts[skill_counts["Skill"].isin(top_skills)]

# Get total skill count per industry
industry_totals = skill_counts_filtered.groupby("Industry")["Count"].sum().nla

# Filter to top 10 industries
skill_counts_top_industries = skill_counts_filtered[skill_counts_filtered["Ind
```

```
# Create stacked bar chart
fig = px.bar(
    skill_counts_top_industries,
    x="Industry",
    y="Count",
    color="Skill",
    title="Top 5 In-Demand Skills by Industry (Top 10 Industries)",
    text="Count",
    color_discrete_sequence=px.colors.qualitative.Pastel
)

fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    xaxis_title="Industry (NAICS2)",
    yaxis_title="Skill Count",
    plot_bgcolor="rgba(240,240,240,1)",
    paper_bgcolor="rgba(255,255,255,1)",
    font=dict(family="Verdana", size=14),
    xaxis_tickangle=45,
    barmode='stack',
    height=850
)

fig.update_traces(textfont_size=12, textposition='inside')

fig.show()
fig.write_image("output/Top 5 In-Demand Skills by Industry (Top 10 Industries)
```





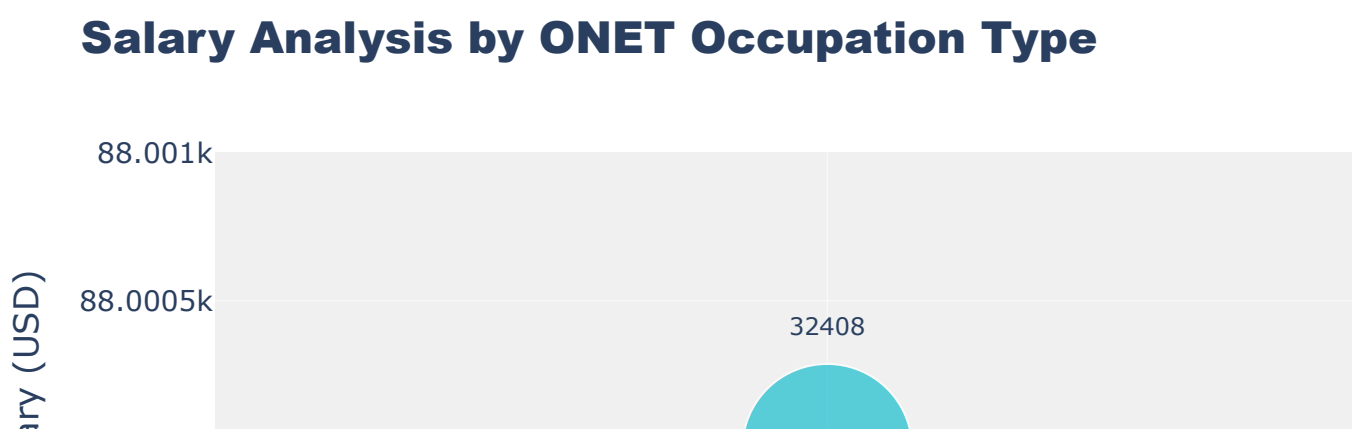
The stacked bar chart displays the top five most common skills across the ten industries with the highest demand, offering clear insight into industry-specific skill requirements. “Communication” and “Management” skills are particularly prominent in Professional and Administrative Services, while industries such as Finance and Information Technology show strong demand for problem-solving and operations skills.

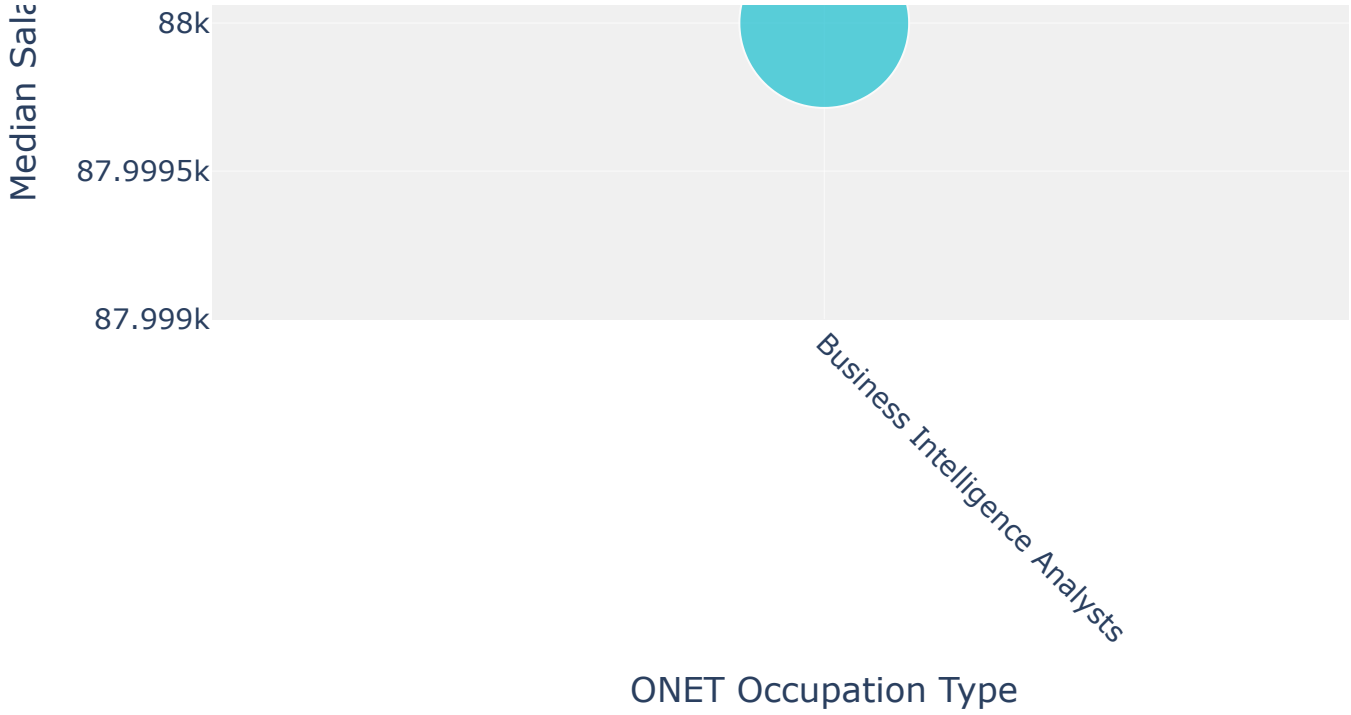
7 Salary Analysis by ONET Occupation Type (Bubble Chart)

- 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.

```
# Select ONET occupation and salary, filter out null and zero salaries
df_onet_salary = df.select("ONET_NAME", "SALARY_FROM") \
    .filter(col("ONET_NAME").isNotNull() & col("SALARY_FROM").isNotNull() & (c

# Convert to Pandas DataFrame
pdf_onet_salary = df_onet_salary.toPandas()

# Group by ONET occupation: calculate median salary and job count
onet_salary_stats = pdf_onet_salary.groupby("ONET_NAME").agg(
    Median_Salary=("SALARY_FROM", "median"),
```

```
    Job_Count=("SALARY_FROM", "count")
).reset_index()

# Filter out occupations with too few postings
onet_salary_stats = onet_salary_stats[onet_salary_stats["Job_Count"] >= 10]

# Determine how many occupations to display
if len(onet_salary_stats) > 10:
    onet_to_plot = onet_salary_stats.sort_values("Job_Count", ascending=False)
else:
    onet_to_plot = onet_salary_stats  # Show all if less than 10

# Create bubble chart
fig = px.scatter(
    onet_to_plot,
    x="ONET_NAME",
    y="Median_Salary",
    size="Job_Count",
    title="Salary Analysis by ONET Occupation Type",
    text="Job_Count",
    color_discrete_sequence=["#17BECF"],
    size_max=60
)

fig.update_layout(
    title_font=dict(size=22, family="Arial Black"),
    xaxis_title="ONET Occupation Type",
    yaxis_title="Median Salary (USD)",
    plot_bgcolor="rgba(240,240,240,1)",
    paper_bgcolor="rgba(255,255,255,1)",
    font=dict(family="Verdana", size=14),
    xaxis_tickangle=45,
    height=600
)

fig.update_traces(textposition='top center', textfont_size=12)

fig.show()
fig.write_image("output/Salary Analysis by ONET Occupation Type.svg")
```





The bubble chart shows that all valid salary data is concentrated in the occupation “Business Intelligence Analysts”, with a median salary of $88,000 and over 32,000 job postings. This suggests an exceptionally high demand for this role, potentially overshadowing other occupations due to incomplete salary reporting.

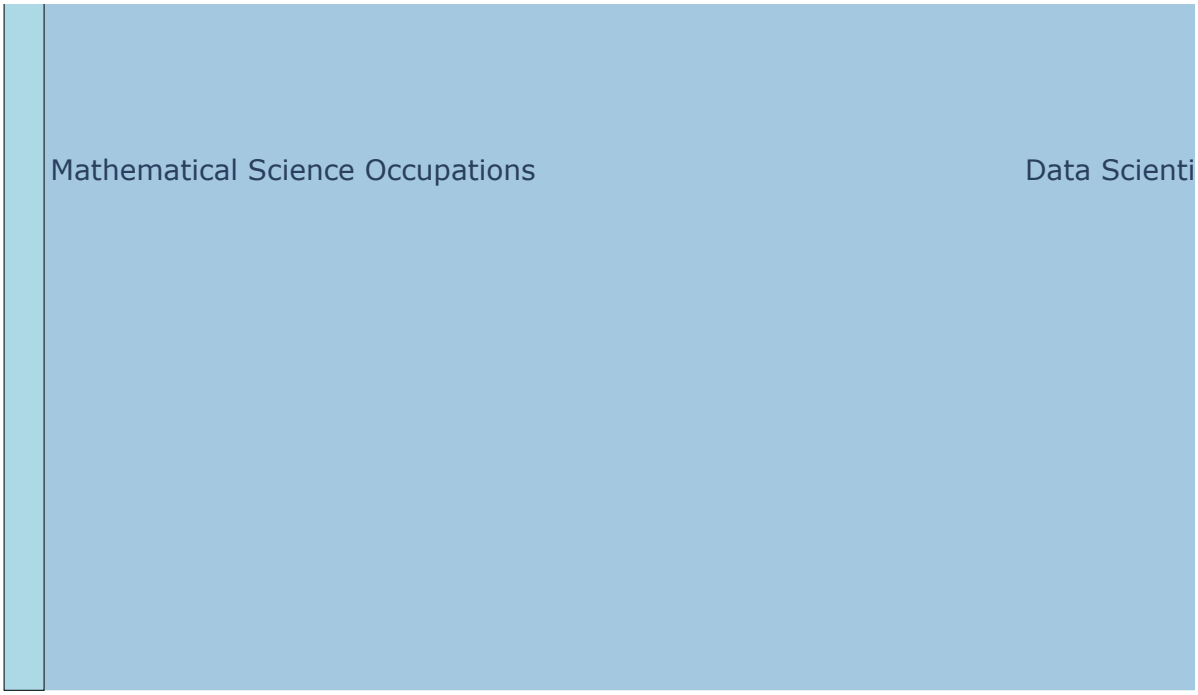
8 Career Pathway Trends (Sankey Diagram)

- 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.

```
# Select finer SOC classifications for career flow analysis
df_soc_alt = df.select("SOC_2021_3_NAME", "SOC_2021_4_NAME") \
    .filter(col("SOC_2021_3_NAME").isNotNull() & col("SOC_2021_4_NAME").isNotN

# Convert to Pandas
pdf_soc_alt = df_soc_alt.toPandas()

# Group by SOC level 3 ➡️ level 4 to count transitions
soc_counts_alt = pdf_soc_alt.groupby(["SOC_2021_3_NAME", "SOC_2021_4_NAME"]).s

#  Keep only top 10 most common transitions for readability
soc_counts_alt = soc_counts_alt.sort_values("Count", ascending=False).head(10)

# Create unique label list and mapping to indices
```

```
labels = list(set(soc_counts_alt["SOC_2021_3_NAME"]).union(set(soc_counts_alt[
label_map = {name: idx for idx, name in enumerate(labels)}

# Map names to indices for source and target
soc_counts_alt["source_idx"] = soc_counts_alt["SOC_2021_3_NAME"].map(label_map
soc_counts_alt["target_idx"] = soc_counts_alt["SOC_2021_4_NAME"].map(label_map

# Create Sankey diagram
import plotly.graph_objects as go

fig = go.Figure(data=[go.Sankey(
    node=dict(
        pad=20,
        thickness=20,
        line=dict(color="black", width=0.5),
        label=labels,
        color="lightblue"
    ),
    link=dict(
        source=soc_counts_alt["source_idx"],
        target=soc_counts_alt["target_idx"],
        value=soc_counts_alt["Count"],
        color="rgba(31,119,180,0.4)"  # Custom transparent blue
    )
)])

fig.update_layout(
    title_text="Career Pathway Trends by SOC Classification (Level 3 to 4)",
    font=dict(size=14, family="Verdana"),
    title_font=dict(size=22, family="Arial Black"),
    paper_bgcolor="white",
    plot_bgcolor="white",
    height=700
)

fig.show()
fig.write_image("output/Career Pathway Trends by SOC Classification (Level 3 t
```

Career Pathway Trends by SOC Classification (Level



The Sankey diagram illustrates a highly concentrated career pathway from “Mathematical Science Occupations” to “Data Scientists”. This dominant transition suggests that a significant portion of job postings within mathematical fields are targeted specifically at data science roles, indicating a clear and specialized career trajectory within this occupational cluster.