--- title: "Module 04: Lab 01" subtitle: "Visual Reporting and Storytelling" author: "Jitvan Vij" number-sections: true date: "04/21/2024" date-modified: today date-format: long engine: jupyter categories: ["visualization","plotly","spark","Visual Reporting","Storytelling with Data","Industry-Specific Visualization",] execute: eval: true echo: true ---

Objectives {.unnumbered}

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 {.unnumbered}

In [6]:
```
!gdown https://drive.google.com/uc?id=1V2GCHGt2dkFGqVBeoUFckU4IhUgk4ocQ
```

```
Downloading...
From (original): https://drive.google.com/uc?id=1V2GCHGt2dkFGqVBeoUFckU4IhUgk4ocQ
From (redirected): https://drive.google.com/uc?id=1V2GCHGt2dkFGqVBeoUFckU4IhUgk4ocQ&confirm=t&uuid=a008a295-cae9-4a9f-a43f-cf8b93927e7a
To: /home/ubuntu/github-classroom/met-ad-688/lab06-jitvanvij/lightcast_job_postings.csv
100%|████████████████████████████████████████| 717M/717M [00:05<00:00, 133MB/s]
```

In [7]:
```
import pandas as pd
import plotly.express as px
import plotly.io as pio

# Set Plotly renderer for Google Colab
pio.renderers.default = "colab"

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count, expr, percentile_approx

# 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"

# 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|        SOUR
CE_TYPES|             SOURCES|                 URL|ACTIVE_URLS|ACTIVE_SOURCES_INFO|           TITLE_RAW|
BODY|MODELED_EXPIRED|MODELED_DURATION| COMPANY|        COMPANY_NAME|COMPANY_RAW|COMPANY_IS_STAFFING|EDUCATION_LEVE
LS|EDUCATION_LEVELS_NAME|MIN_EDULEVELS| MIN_EDULEVELS_NAME|MAX_EDULEVELS|MAX_EDULEVELS_NAME|EMPLOYMENT_TYPE|EMPLOY
```

```
MENT_TYPE_NAME|MIN_YEARS_EXPERIENCE|MAX_YEARS_EXPERIENCE|IS_INTERNSHIP|SALARY|REMOTE_TYPE|REMOTE_TYPE_NAME|ORIGINA
L_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_INCOMIN
G|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|SPECIALIZ
ED_SKILLS_NAME|      CERTIFICATIONS| CERTIFICATIONS_NAME|       COMMON_SKILLS|  COMMON_SKILLS_NAME|     SOFTWARE_S
KILLS|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_NAM
E|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|L
OT_V6_CAREER_AREA_NAME|  SOC_2|           SOC_2_NAME|  SOC_3|           SOC_3_NAME|  SOC_4|      SOC_4_NAME|  SOC_5|
SOC_5_NAME|LIGHTCAST_SECTORS|LIGHTCAST_SECTORS_NAME|NAICS_2022_2|   NAICS_2022_2_NAME|NAICS_2022_3|   NAICS_2022_3
_NAME|NAICS_2022_4|   NAICS_2022_4_NAME|NAICS_2022_5|   NAICS_2022_5_NAME|NAICS_2022_6|   NAICS_2022_6_NAME|
+--------------------+--------------------+--------------------+-------------+------+-----------+----------------
--------+--------------------+--------------------+----------------+-------------+-------------------+----------
----------+------------------+-------------+--------+--------------------+---------------+----------------------
--------+-------------------+-------------+------------------+------------------+-----------------+--------------+
--------------------+--------------------+------------------+-------------------+------+-----------+------------+-
-----------------+---------+----------+--------------------+------------------+--------------+------+----------
----+-----+-------------------+----+----------+---------------+-------------------+-----------------+----------
--------+-------------+-------------+------------------+------------+--------------+------+-----------------+--
-----------------+------+-------------------+-------+--------------------+-------+-----------------------+--------
-------+--------------------+--------------------+--------------------+-------------------+----------------+--
--------------------+---------------+------------------+--------------------+-------------------+-----------
------------+-------------------+---------+---------------+----------+------------------+----------------
----+------------------+-----------+--------------------+-----------+--------------------------+-------------+----
------+---------------------+---------+-------------------+---------+---------------+----------------------+
---------------+--------------------+---------------+---------------------+--------------------------+---------
-----------------+--------------------+----------------------+--------------------------+--------------------
---------------+-----------------+--------------------------+------------------------+--------------------+----
--------------+--------------------+-------------+------------------+--------+-------------------+--------------
------+-------+---------------+------------------+---------------------+-----------+---------------------+---------
----+-------------------+---------------+--------------------+------------+-------------------+-----------+-------
-------------+
|1f57d95acf4dc67ed...|         9/6/2024|  2024-09-06 20:32:...|         0|6/2/2024| 6/8/2024|       6|   [\n  "Com
pany"\n]|[\n  "brassring.c...|[\n  "https://sjo...|         []|              NULL|Enterprise Analys...|31-May-202
4\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|       20980|        5139|          Union, AR|          5139|          Union, AR|       20
980|       El Dorado, AR|       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  "Merchandis
i...|                  []|                  []|[\n  "KS126706DPF...|[\n  "Mathematics...|[\n  "KS440W865GC...|[\n
"SQL (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n  "45.0601",\n...|[\n  "Economi
cs, ...|[\n  "45.06",\n  ...|[\n  "Economics",...|[\n  "45",\n  "27...|[\n  "Social Scie...|   15-0000|Computer an
d Math...|   15-2000|Mathematical Scie...|   15-2050|Data Scientists|   15-2051|Data Scientists|              23|In
formation Techn...|        231010|Business Intellig...|                  23101011|            General ERP Analy...|
2310|     Business Intellig...|                   2310|          23101011|         General ERP Analy...|          231010|
Business Intellig...|                   2310|        Business Intellig...|                  23|   Information Tech
n...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists|
[\n  7\n]|  [\n  "Artificial ...|          44|        Retail Trade|         441|Motor Vehicle and...|        4413|
Automotive Parts,...|       44133|Automotive Parts ...|      441330|Automotive Parts ...|
|0cb072af26757b6c4...|         8/2/2024|  2024-08-02 17:08:...|         0|6/2/2024| 8/1/2024|    NULL| [\n  "Job B
oard"\n]| [\n  "maine.gov"\n]|[\n  "https://job...|         []|              NULL|Oracle Consultant...|Oracle Con
sultant...|       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|  Kenneb
ec, ME|12300|Augusta-Watervill...|   23|     Maine|          23011|          Kennebec, ME|          23011|          Ke
nnebec, ME|       12300|Augusta-Watervill...|       12300|Augusta-Watervill...|    56|Administrative an...|   561|
Administrative an...|  5613| Employment Services| 56132|Temporary Help Se...|561320|Temporary Help Se...|ET21DDA63
780A7DC09| Oracle Consultants|oracle consultant...|[\n  "KS122626T55...|[\n  "Procurement...|[\n  "KS122626T55...|
[\n  "Procurement...|                  []|                  []|                  []|                  []|[\n  "BGS
BF3F508F...|[\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 Consultan
t...|                2310|     Business Intellig...|                  23101012|            Oracle Consultan
t...|           231010|  Business Intellig...|                   2310|         Business Intellig...|
23|   Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-20
51|Data Scientists|              NULL|                NULL|           56|Administrative an...|         561|Adminis
trative an...|        5613| Employment Services|       56132|Temporary Help Se...|      561320|Temporary Help S
e...|
|85318b12b3331fa49...|         9/6/2024|  2024-09-06 20:32:...|         1|6/2/2024| 7/7/2024|      35| [\n  "Job B
```

```
oard"\n]|[\n  "dejobs.org"\n]|[\n  "https://dej...|        []|              NULL|       Data Analyst|Taking car
e 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|Dal
las-Fort Worth...|   48|    Texas|         48113|          Dallas, TX|         48113|         Dallas, TX|
19100|Dallas-Fort Worth...|      19100|Dallas-Fort Worth...|    52|Finance and Insur...|   524|Insurance Carrie
r...|  5242|Agencies, Brokera...| 52429|Other Insurance R...|524291|    Claims Adjusting|ET3037E0C947A02404|
Data Analysts|       data analyst|[\n  "KS1218W78FG...|[\n  "Management"...|[\n  "ESF3939CE1F...|   [\n  "Excepti
on 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 Mat
h...|  15-2000|Mathematical Scie...|  15-2050|Data Scientists|  15-2051|Data Scientists|            23|Informa
tion 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 Tech
n...|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|Agenc
ies, Brokera...|      52429|Other Insurance R...|     524291|   Claims Adjusting|
|1b5c3941e54a1889e...|        9/6/2024|  2024-09-06 20:32:...|        1|6/2/2024|7/20/2024|     48| [\n  "Job B
oard"\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|Ful
l-time (> 32 h...|                 3|              NULL|       false|  NULL|         0|         [None]|
NULL|    NULL|      NULL|{\n  "lat": 33.44...|   UGhvZW5peCwgQVo=|  Phoenix, AZ|  4013|  Maricopa, AZ|38060|Pho
enix-Mesa-Chan...|    4|   Arizona|          4013|        Maricopa, AZ|          4013|        Maricopa, AZ|
38060|Phoenix-Mesa-Chan...|      38060|Phoenix-Mesa-Chan...|    52|Finance and Insur...|   522|Credit Intermedi
a...|  5221|Depository Credit...| 52211|  Commercial Banking|522110|  Commercial Banking|ET2114E0404BA30075|Manage
ment Analysts|sr lead data mgmt...|[\n  "KS123QX62QY...|[\n  "Exit Strate...|[\n  "KS123QX62QY...|   [\n  "Exit St
rate...|                 []|                 []|[\n  "KS7G6NP6R6L...|[\n  "Reliability...|[\n  "KS4409D76NW...|
[\n  "SAS (Softwa...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|                 []|
[]|                []|                []|                []|                []|  15-0000|Computer and Mat
h...|  15-2000|Mathematical Scie...|  15-2050|Data Scientists|  15-2051|Data Scientists|            23|Informa
tion 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 Tech
n...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists|
[\n  6\n]|  [\n  "Data Privac...|         52|Finance and Insur...|        522|Credit Intermedia...|        5221|
Depository Credit...|      52211|  Commercial Banking|     522110|  Commercial Banking|
|cb5ca25f02bdf25c1...|       6/19/2024|  2024-06-19 07:00:00|        0|6/2/2024|6/17/2024|     15|[\n  "FreeJo
bBoar...|[\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|Par
t-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|      3370
0|        Modesto, CA|      33700|        Modesto, CA|    99|Unclassified Indu...|   999|Unclassified Indu...|
9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET0000000000000000|      Unclas
sified|comisiones de por...|                 []|                 []|                 []|                 []
|                 []|                 []|                 []|                 []|                 []|
[]|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|                 []|                 []|
[]|                []|                []|                []|  15-0000|Computer and Math...|  15-2000|Mathe
matical Scie...|  15-2050|Data Scientists|  15-2051|Data Scientists|            23|Information Techn...|
231010|Business Intellig...|             23101012|          Oracle Consultant...|                2310|     B
usiness Intellig...|               23101012|          Oracle Consultant...|          231010|  Business
Intellig...|               2310|       Business Intellig...|            23|  Information Techn...|15-000
0|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists|           N
ULL|               NULL|         99|Unclassified Indu...|        999|Unclassified Indu...|        9999|Unclas
sified Indu...|      99999|Unclassified Indu...|     999999|Unclassified Indu...|
+--------------------+------------------+----------------------+----------+--------+----------+--------+------------
---------+--------------------+------------------+--------------------+----------+--------------------+-------------------+----------
-----------+------------------+------------------+----------+--------------------+------------+--------------------+---------
---------+--------------------+--------------+--------------------+--------------+------------------+--------------------+-----------------+
--------------------+-----------------+--------------------+----------+--------------+----------+--------+--------------------+-
-------------------+----------+--------------------+--------------------+--------------------+-------------+------+------------
----+------+--------------------+------+----------------------+--------------------+--------------------+----------+------------
---------+-------------+--------------------+-------------------+---------------------+----------+--------------------+------+--
-------------------+------+--------------------+------+--------------------+------------------+------+--------------------+----------
--------+-------------------+--------------------+--------------------+----------------------+--------------------+-------------------+--
--------------------+--------------------+--------------------+--------------------+--------------------+--------------------
-------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----------------
----+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------
------+--------------------+--------------------+--------------------+----------+--------------------+------------------+--------------+
---------------+--------------------+--------------------+------------------+--------------------+--------------------+----------------
---------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------------
---------------+--------------------+--------------------+----------------------+--------------------+----------------------+-------
--------------+--------------------+--------+--------------------+-------+--------------------+--------+--------------------+---------
------+-------+---------------------+--------------------+----------------------+------------+--------------------+--------------------
----+--------------------+--------------+----------------+--------------------+--------------+--------------------+----------+------------
```

```
-------------+
only showing top 5 rows
```

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.

In [8]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col

# Initialize Spark Session
spark = SparkSession.builder.appName("LightcastData").getOrCreate()

# Load Data
df = spark.read.option("header", "true") \
    .option("inferSchema", "true") \
    .option("multiLine", "true") \
    .option("escape", "\"") \
    .csv("lightcast_job_postings.csv")

# Check unique values
employment_types = df.select("EMPLOYMENT_TYPE_NAME").distinct().dropna()
employment_types.show(20, truncate=False)
```

```
[Stage 14:>                                                         (0 + 1) / 1]
+------------------------+
|EMPLOYMENT_TYPE_NAME    |
+------------------------+
|Part-time / full-time   |
|Part-time (â‰¤ 32 hours)|
|Full-time (> 32 hours)  |
+------------------------+
```

In [9]:
```
# Select needed columns and convert to Pandas
pdf = df.select("EMPLOYMENT_TYPE_NAME", "SALARY").toPandas()

# box plot
fig = px.box(pdf,
             x="EMPLOYMENT_TYPE_NAME",
             y="SALARY",
             title="Salary Distribution by Employment Type",
             color_discrete_sequence=["Magenta"],
             width=800,
             height=800)

# Customizing
fig.update_layout(
    font_family="Arial",
    title_font_size=16,
    plot_bgcolor="#f5efdc",   # inner plot area
    paper_bgcolor="#f5efdc"   # outer background area
)

fig.show()
```

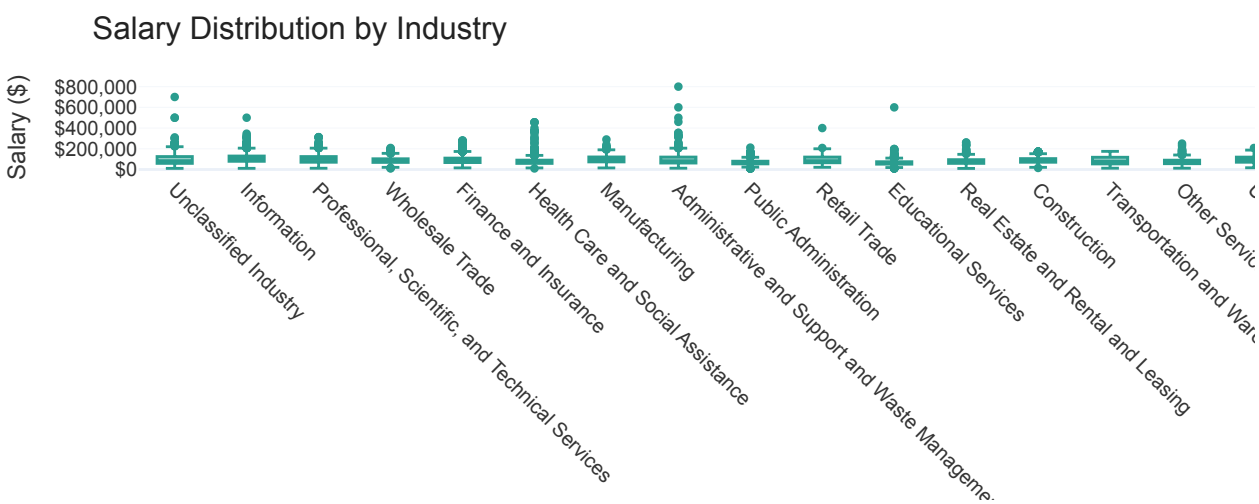
Interpretation

The box plot shows that full-time jobs (over 32 hours) offer the highest median salary, around $116.5k, and also have a wider salary range with many high-paying outliers. Jobs that are both part-time and full-time come next with a median of about 100k, while part-time roles (under 32 hours) have the lowest median salary, around $86.4k. Overall, full-time positions not only pay more on average, but also show more variation in salary compared to other job types.

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.

In [10]:
```
# Step 1: Filter and convert to Pandas
filtered_df = df.filter((df.SALARY_FROM > 0) & df.NAICS2_NAME.isNotNull())
salary_industry_df = filtered_df.select("NAICS2_NAME", "SALARY_FROM").toPandas()

# Step 2: Create box plot
import plotly.express as px

fig = px.box(
    salary_industry_df,
    x="NAICS2_NAME",
    y="SALARY_FROM",
    title="Salary Distribution by Industry",
    color_discrete_sequence=["#2a9d8f"],
    template="plotly_white"
)

# Step 3: Layout styling
fig.update_layout(
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
    xaxis_title="Industry",
    yaxis_title="Salary ($)",
    yaxis_tickprefix="$",
    yaxis_tickformat=",",
    xaxis_tickangle=45,
    margin=dict(t=60, l=60, r=40, b=100),
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff"
)

fig.show()
```



linterpretation

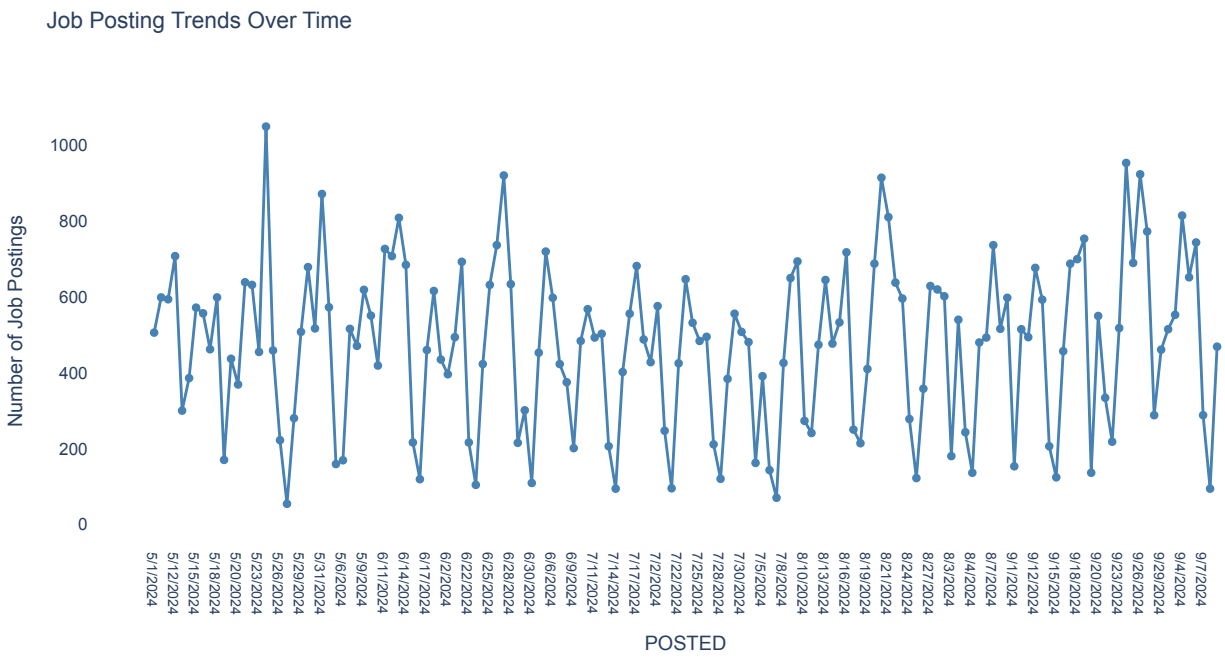
The box plot shows that industries like Administrative, Manufacturing, and Information tend to offer higher salary ranges and include more high-paying outliers, suggesting there's good potential for top-tier salaries in these fields. On the other hand, industries like Arts, Entertainment, and Accommodation usually have lower and more tightly packed salaries, meaning the pay is generally more limited and consistent in those sectors.

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.

In [11]:
```
# 1. Import Spark functions and convert POSTED to date
from pyspark.sql.functions import to_date, count

# 2. Group by date and count job postings
df_posting_trend = df.withColumn("POSTED_DATE", to_date("POSTED")) \
    .groupBy("POSTED_DATE") \
    .agg(count("*").alias("count")) \
    .orderBy("POSTED_DATE")

# 3. Convert to Pandas for Plotly
pdf_posting_trend = df_posting_trend.toPandas()
```

In [12]:
```
# Group by POSTED date and count number of job postings
df_posting_trend = df.groupBy("POSTED").count().orderBy("POSTED")

# Convert to Pandas for plotting
pdf_posting_trend = df_posting_trend.toPandas()

# Create a line chart
import plotly.express as px

fig = px.line(pdf_posting_trend,
              x="POSTED",
              y="count",
              title="Job Posting Trends Over Time",
              markers=True,
              color_discrete_sequence=["steelblue"],
              width=1000,
              height=500)

# Customization
fig.update_layout(
    font_family="Arial",
    title_font_size=16,
    plot_bgcolor="white",
    paper_bgcolor="white",
)

fig.update_xaxes(
    tickformat="%b %d",   # e.g., May 05
    tickfont=dict(size=10),
    showgrid=False
)

fig.update_yaxes(
    title_text="Number of Job Postings",
    showgrid=True
)
fig.show()
```



Interpretation

The line chart shows that job postings go up and down quite a bit day to day, which suggests that hiring activity can be pretty unpredictable. There's a clear spike in late August to mid-September, likely because companies are hiring after the summer slowdown. Even with all the ups and downs, there doesn't seem to be a consistent long-term trend — things stay fairly balanced over time.

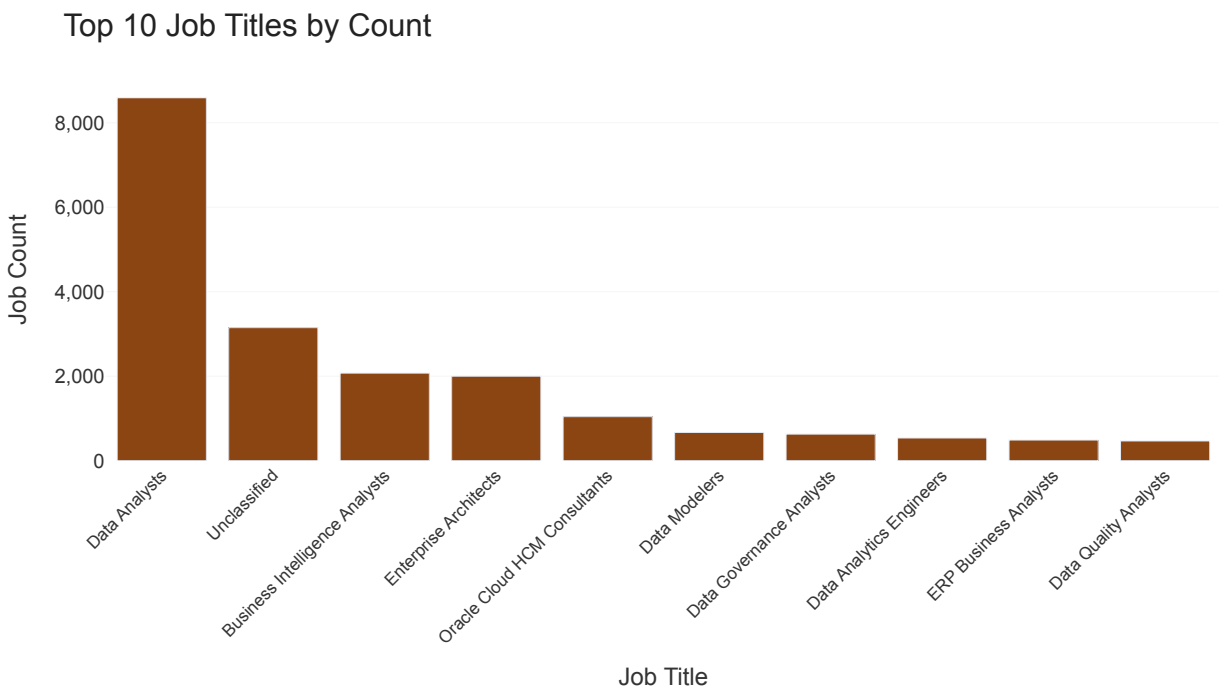
Top 10 Job Titles by Count

- 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.

In [13]:
```
# Group by TITLE_NAME, count, and get top 10
df_titles = df.groupBy("TITLE_NAME").count().orderBy("count", ascending=False).limit(10)

# Convert to Pandas
pdf_titles = df_titles.toPandas()
```

In [14]:
```
import plotly.express as px

fig = px.bar(
    pdf_titles,
    x="TITLE_NAME",
    y="count",
    title="Top 10 Job Titles by Count",
    color_discrete_sequence=["#8B4513"],  # Hex for "dark brown"
    width=900,
    height=500
)

fig.update_layout(
```

```
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
    xaxis_title="Job Title",
    yaxis_title="Job Count",
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff",
    margin=dict(t=60, l=60, r=40, b=100)
)

fig.update_xaxes(
    tickangle=-45,
    tickfont=dict(size=12)
)

fig.update_yaxes(
    tickformat=",",
    gridcolor="#eeeeee"
)

fig.show()
```



Interpretation

The job title "Data Analysis" stands out as the most posted role by far, way ahead of the others. Titles like Business Intelligence Analyst and Enterprise Architect also appear frequently, showing that tech and data roles are in high demand. After the top few titles, the number of postings drops off quickly, which means employers are mainly focused on hiring for a small set of key positions.

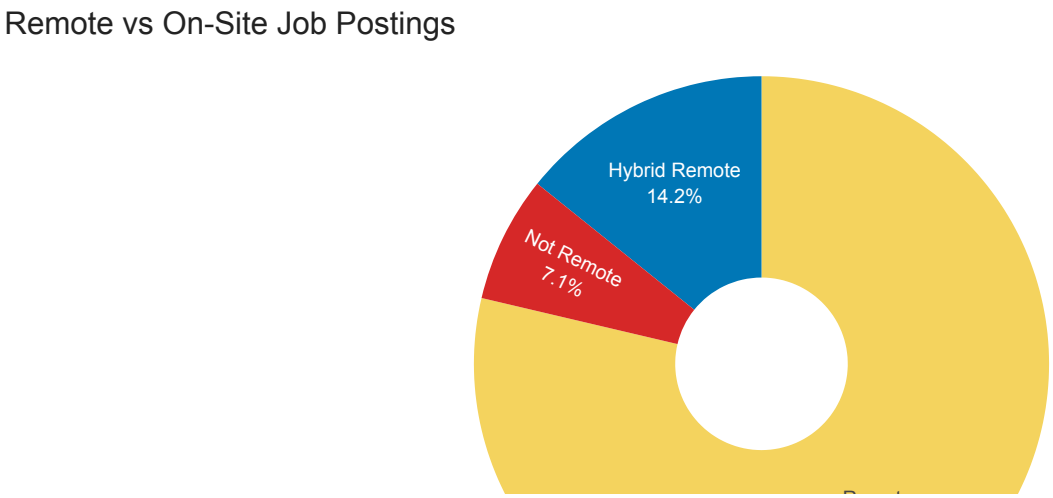
Remote vs On-Site Job Postings

- 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.

In [15]:
```
from pyspark.sql.functions import col
```

```
# Filter for valid remote types and count them
df_remote_filtered = df.filter(
    col("REMOTE_TYPE_NAME").isin(["Remote", "Hybrid Remote", "Not Remote"])
).groupBy("REMOTE_TYPE_NAME").count()

# Convert to Pandas
pdf_remote_filtered = df_remote_filtered.toPandas()
```

In [16]:
```
import plotly.express as px

fig = px.pie(
    pdf_remote_filtered,
    names="REMOTE_TYPE_NAME",
    values="count",
    title="Remote vs On-Site Job Postings",
   color_discrete_sequence=["#f4d35e", "#0077b6", "#d62828"]
,
    hole=0.3  # makes it a donut chart — optional but stylish
)

# Custom styling
fig.update_traces(
    textinfo='percent+label',
    textfont_size=14
)

fig.update_layout(
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff",
    margin=dict(t=60, l=60, r=40, b=60)
)

fig.show()
```



INterpretation

Most of the job postings are for remote positions, which shows that working from home has become the new normal for many companies. There are still some hybrid and fully on-site jobs out there, but they make up a much smaller part of the market compared to remote roles.

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.

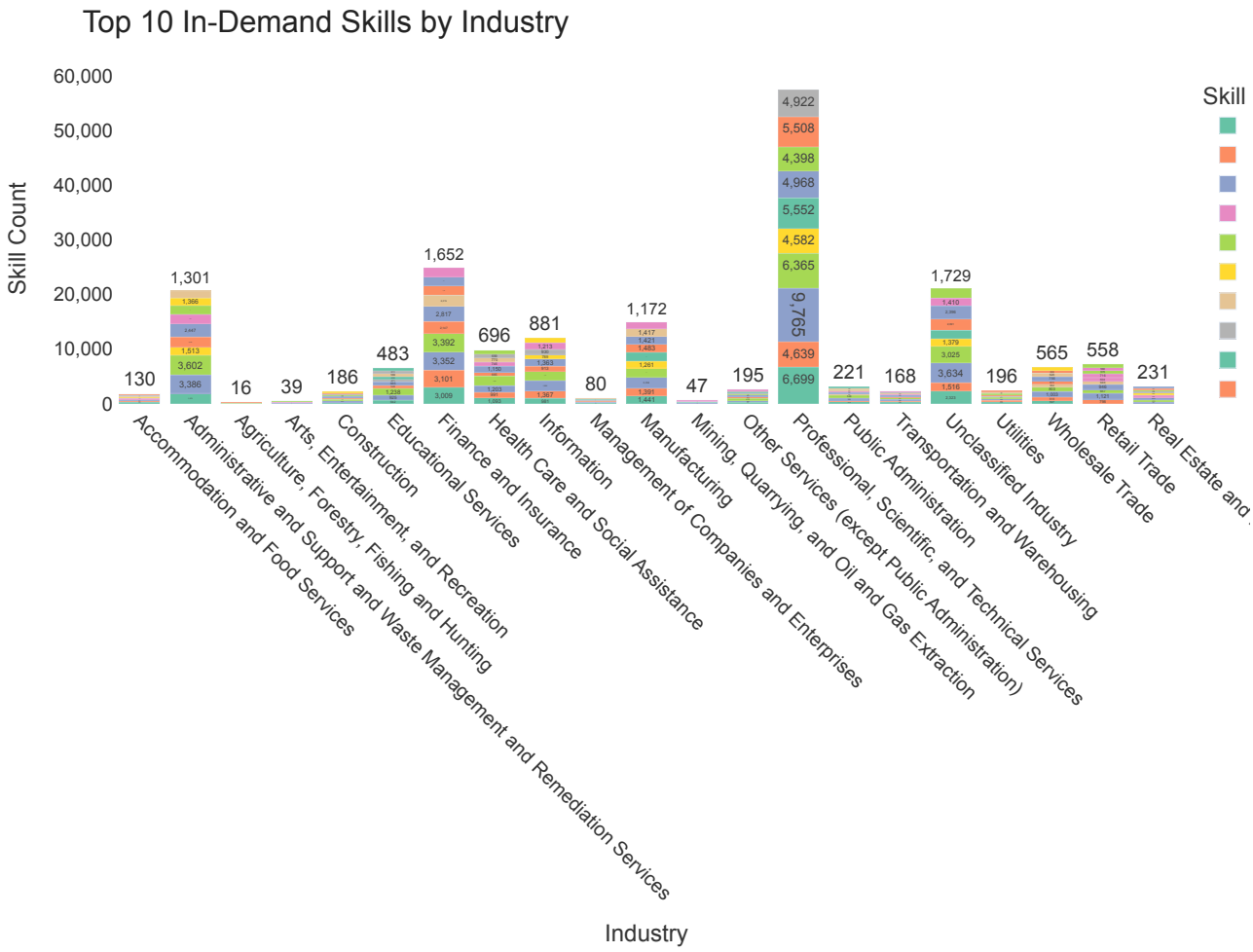
In [17]:
```
from pyspark.sql.functions import explode, split, trim

# Splitting the SKILLS_NAME column into an array of skills
df_split = df.withColumn("Skill", explode(split("SKILLS_NAME", ",")))

# Cleaning whitespace from skill names
df_split = df_split.withColumn("Skill", trim(df_split["Skill"]))

# Filtering nulls if needed
df_split = df_split.filter(df_split["Skill"].isNotNull())
```

In [18]:
```
# Register the cleaned & exploded DataFrame as a new SQL view
df_split.createOrReplaceTempView("job_data")

# Updated Spark SQL query
query = """
WITH SkillCounts AS (
    SELECT
        NAICS_2022_2_NAME AS Industry,
        Skill,
        COUNT(*) AS Skill_Count
    FROM job_data
    WHERE Skill IS NOT NULL AND NAICS_2022_2_NAME IS NOT NULL
    GROUP BY Industry, Skill
),
RankedSkills AS (
    SELECT *,
        ROW_NUMBER() OVER (PARTITION BY Industry ORDER BY Skill_Count DESC) AS rank
    FROM SkillCounts
)
SELECT * FROM RankedSkills WHERE rank <= 10
"""

# Execute and store results
df_top_skills = spark.sql(query)
```

In [19]:
```
pdf_top_skills = df_top_skills.toPandas()
```

In [20]:
```
import plotly.express as px

fig = px.bar(
    pdf_top_skills,
    x="Industry",
    y="Skill_Count",
    color="Skill",
    title="Top 10 In-Demand Skills by Industry",
    text_auto=True,
    barmode='stack',
    color_discrete_sequence=px.colors.qualitative.Set2,  # brighter categorical palette
    width=1200,
    height=700
)

fig.update_layout(
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
```

```
    xaxis_title="Industry",
    yaxis_title="Skill Count",
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff",
    margin=dict(t=60, l=60, r=40, b=120),
    xaxis_tickangle=45
)

fig.update_yaxes(tickformat=",")
fig.show()
```



Interpretation

Skills like "Management" and "Leadership" are in high demand across almost every industry, showing how important they are no matter the field. The Professional, Scientific, and Technical Services sector stands out with a wide range of skill needs, especially for things like Data Analysis, Agile Methodology, and Communication. More strategic skills like "Business Requirements", "Project Management", and "Planning" are mostly found in Unclassified and Professional industries, where coordination and planning are key. While industries like Retail Trade and Real Estate have fewer overall postings, they still consistently ask for customer service and problem-solving skills.

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`

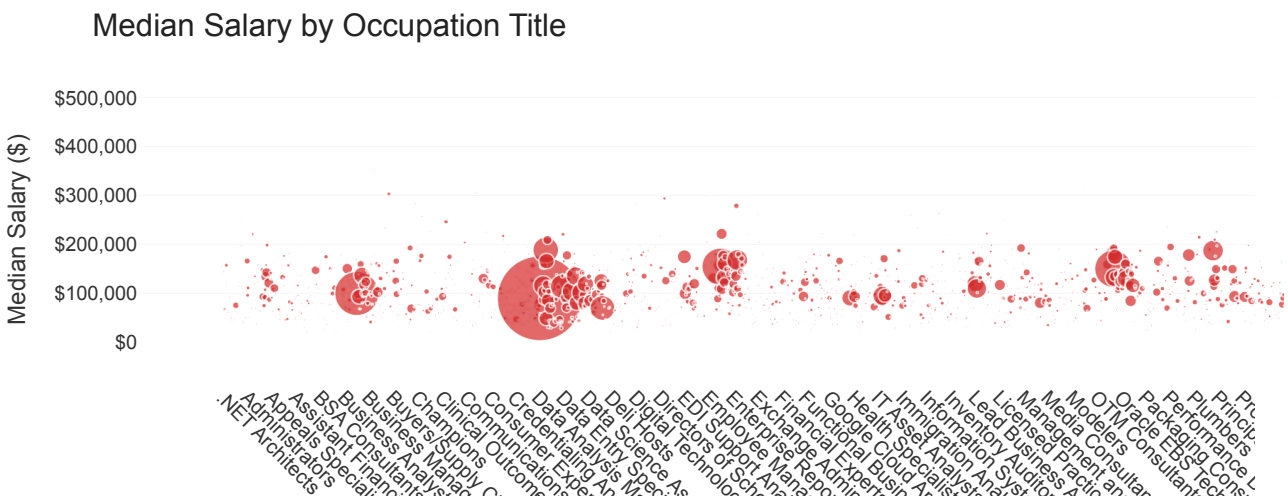
    - **Size** = Number of job postings
  - Apply custom colors and font styles.
- **Explanation:** Write two sentences about what the graph reveals.

In [21]:
```
df.createOrReplaceTempView("job_data")
```

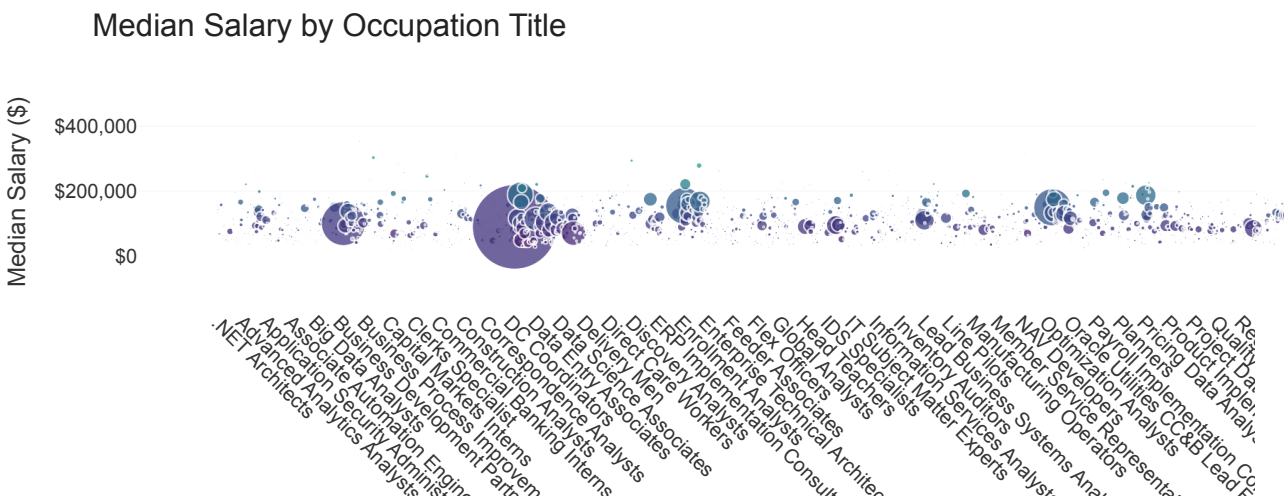
In [22]:
```
query_salary = """
WITH SalaryStats AS (
    SELECT
        TITLE_NAME,
        percentile_approx(SALARY, 0.5) AS Median_Salary,
        COUNT(*) AS Job_Count
    FROM job_data
    WHERE SALARY IS NOT NULL AND TITLE_NAME IS NOT NULL
    GROUP BY TITLE_NAME
)
SELECT * FROM SalaryStats
"""
df_salary_stats = spark.sql(query_salary)
```

In [23]:
```
pdf_salary_stats = df_salary_stats.toPandas()
```

In [24]:
```
import plotly.express as px

fig = px.scatter(
    pdf_salary_stats,
    x="TITLE_NAME",           # X-axis: Job Title
    y="Median_Salary",        # Y-axis: Median Salary
    size="Job_Count",         # Bubble size = Number of postings
    color_discrete_sequence=["#d62828"],  # Bold red
    title="Median Salary by Occupation Title",
    labels={
        "TITLE_NAME": "Occupation Title",
        "Median_Salary": "Median Salary ($)",
        "Job_Count": "Number of Job Postings"
    },
    size_max=60  # Controls the max bubble size
)

fig.update_layout(
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
    xaxis_title="Occupation Title",
    yaxis_title="Median Salary ($)",
    xaxis_tickangle=45,
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff",
    margin=dict(t=60, l=60, r=40, b=120)
)

fig.update_yaxes(
    tickformat="$,.0f",    # e.g. $50,000
    gridcolor="#eeeeee"
)
fig.update_traces(marker=dict(opacity=0.7, line=dict(width=1, color='white')))

fig.show()
```



In [25]:
```
import plotly.express as px

fig = px.scatter(
    pdf_salary_stats,
    x="TITLE_NAME",              # X-axis: Job Title
    y="Median_Salary",           # Y-axis: Median Salary
    size="Job_Count",            # Bubble size = Number of postings
    color="Median_Salary",       # Bubble color = Median Salary value
    color_continuous_scale="Viridis",  # Gradient: light to dark (low to high salary)
    title="Median Salary by Occupation Title",
    labels={
        "TITLE_NAME": "Occupation Title",
        "Median_Salary": "Median Salary ($)",
        "Job_Count": "Number of Job Postings"
    },
    size_max=60
)

# Layout styling
fig.update_layout(
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=14, color="#333"),
    xaxis_title="Occupation Title",
    yaxis_title="Median Salary ($)",
    xaxis_tickangle=45,
    plot_bgcolor="#ffffff",
    paper_bgcolor="#ffffff",
    margin=dict(t=60, l=60, r=40, b=120),
    coloraxis_colorbar=dict(
        title="Median Salary",
        tickprefix="$",
        ticks="outside"
    )
)

# Y-axis formatting and bubble borders
fig.update_yaxes(
    tickformat="$,.0f",
    gridcolor="#eeeeee"
)
fig.update_traces(marker=dict(opacity=0.75, line=dict(width=1, color='white')))

fig.show()
```



INterpretation

The chart shows that median salaries vary a lot between job titles, with some roles paying over 200,000. Bigger bubbles — like for Business Intelligence Analysts — mean those jobs are in high demand because they have more postings. Most job titles fall around 100,000, which seems to be the typical salary in this dataset. High-paying roles are mostly in analytics, architecture, and management, showing that these areas offer some of the best career opportunities in tech and business.

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.

In [26]:
```
query_sankey = """
SELECT
    SOC_2021_2_NAME AS source,
    SOC_2021_3_NAME AS target,
    COUNT(*) AS value
FROM job_data
WHERE SOC_2021_2_NAME IS NOT NULL AND SOC_2021_3_NAME IS NOT NULL
GROUP BY source, target
"""
df_transitions1 = spark.sql(query_sankey)
```

In [27]:
```
pdf_transitions = df_transitions1.toPandas()
```

In [28]:
```
# Combine all unique labels from both source and target
all_labels = pd.concat([pdf_transitions['source'], pdf_transitions['target']]).unique().tolist()
```

```
# Create mapping to integer index
label_to_index = {label: idx for idx, label in enumerate(all_labels)}

# Map source and target to indices
pdf_transitions['source_id'] = pdf_transitions['source'].map(label_to_index)
pdf_transitions['target_id'] = pdf_transitions['target'].map(label_to_index)
```

In [29]:
```
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=all_labels,
        color="#ffb3b3"  # Softer red-pink for node fill
    ),
    link=dict(
        source=pdf_transitions['source_id'],
        target=pdf_transitions['target_id'],
        value=pdf_transitions['value'],
        color="rgba(214, 39, 40, 0.5)"  # Semi-transparent red
    )
)])

fig.update_layout(
    title_text="Career Pathway Trends (SOC Code Transitions)",
    title_font=dict(size=22, family="Arial", color="#222"),
    font=dict(family="Arial", size=13, color="#333"),
    paper_bgcolor="#ffffff",  # Clean white background
    margin=dict(t=60, l=60, r=40, b=60)
)

fig.show()
```

Interpretation: Career Pathway Trends (SOC Code Transitions)

The diagram shows that many people are moving from Computer and Mathematical jobs into Mathematical Science roles, which means there's a lot of movement between these two closely related fields. This suggests that the skills in these areas often overlap, making it easier for professionals to switch or grow their careers within the tech and math space. The thick red band highlights how strong that connection is — it's one of the most common transitions in the data. Employers and educators can use this insight to create training programs that help tech workers transition into roles in mathematical sciences.
