3/20/25 , 11:38 PM Module 04: Lab 01

Module 04: Lab 01

Visual Reporting and Storytelling

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

AUTHOR

Zimo Zeng

MODIFIED

March 21 , 2025

PUBLISHED

November 21 , 2024

Objectives

By the end of this lab, you will: 1 . Load and analyze the Lightcast dataset in Spark DataFrame. 2. Create five easy and three medium-complexity visualizations using Plotly. 3. Explore salary

distributions, employment trends, and job postings. 4. Analyze skills in relation to

NAICS/SOC/ONET codes and salaries. 5. Customize colors, fonts, and styles in all visualizations (default themes result in a 2.5-point deduction). 6. Follow best practices for reporting on data communication.

Step 1: Load the Dataset



import pandas as pd

import plotly.express as px

import plotly.io as pio

pio.renderers.default = "vscode"

from pyspark.sql import SparkSession

from pyspark.sql.functions import col

# 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

localhost:6298 1/23

3/20/25 , 11:38 PM Module 04: Lab 01

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)

localhost:6298 2/23

3/20/25 , 11:38 PM Module 04: Lab 01 |-- 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)

localhost:6298 3/23

3/20/25 , 11:38 PM Module 04: Lab 01

|-- 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)

+--------------------+-----------------+----------------------+----------+----------+-

---------+--------+--------------------+--------------------+--------------------+----

-------+-------------------+--------------------+--------------------+---------------

+----------------+--------+--------------------+-----------+-------------------+------

----------+---------------------+-------------+-------------------+-------------+-----

-------------+---------------+--------------------+--------------------+--------------

------+-------------+------+-----------+----------------+-------------------+---------

+-----------+--------------------+--------------------+-------------+------+----------

----+-----+--------------------+-----+----------+---------------+--------------------

+---------------+--------------------+------------+--------------------+------------+-

-------------------+------+--------------------+------+--------------------+------+---

-----------------+------+--------------------+------+--------------------+------------

------+-------------------+--------------------+--------------------+-----------------

---+--------------------+-----------------------+--------------------+----------------

----+--------------------+--------------------+--------------------+------------------

--+----------+--------------------+----------+--------------------+-------------------

-+--------------------+--------------------+--------------------+--------------------

+--------------------+----------+--------------------+----------+--------------------

+----------+---------------+----------+---------------+---------------+---------------

-----+--------------+--------------------+--------------------------+-----------------

--------------+--------------------+-------------------------+------------------------

-----+----------------------------------+-----------------+----------------------+----

[+----------------------------+------------------+------------------](#bookmark1)

[+-------+--------------------+-------+--------------------+-------+--------------](#bookmark2)

[-+ +---------------+-----------------+----------------------+------------+------](#bookmark3)

[+------------+--------------------+------------+--------------------+---](#bookmark4)

[localhost:6298 4/23](#bookmark5)

3/20/25 , 11:38 PM Module 04: Lab 01

---------+--------------------+------------+--------------------+

| 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 |

+--------------------+-----------------+----------------------+----------+----------+-

---------+--------+--------------------+--------------------+--------------------+----

-------+-------------------+--------------------+--------------------+---------------

+----------------+--------+--------------------+-----------+-------------------+------

----------+---------------------+-------------+-------------------+-------------+-----

-------------+---------------+--------------------+--------------------+--------------

------+-------------+------+-----------+----------------+-------------------+---------

+-----------+--------------------+--------------------+-------------+------+----------

----+-----+--------------------+-----+----------+---------------+--------------------

+---------------+--------------------+------------+--------------------+------------+-

-------------------+------+--------------------+------+--------------------+------+---

-----------------+------+--------------------+------+--------------------+------------

------+-------------------+--------------------+--------------------+-----------------

---+--------------------+-----------------------+--------------------+----------------

----+--------------------+--------------------+--------------------+------------------

--+----------+--------------------+----------+--------------------+-------------------

-+--------------------+--------------------+--------------------+--------------------

+--------------------+----------+--------------------+----------+--------------------

[+ +---------------+----------+---------------+---------------+---------------](#bookmark6)

[+--------------+--------------------+--------------------------+-----------------](#bookmark7)

[+--------------------+-------------------------+------------------------](#bookmark8)

[+----------------------------------+-----------------+----------------------+----](#bookmark9)

[localhost:6298 5/23](#bookmark10)

3/20/25 , 11:38 PM Module 04: Lab 01

+----------------------------+------------------+------------------

+-------+--------------------+-------+--------------------+-------+--------------

-+ +---------------+-----------------+----------------------+------------+------

--------------+------------+--------------------+------------+--------------------+---

---------+--------------------+------------+--------------------+

|  |  |
| --- | --- |
| |1f57d95acf4dc67ed... | 2024-09-06 | 2024-09-06 20:32:... | | 0 |2024-06- |
| 02 |2024-06-08 | 6 | [\n "Company"\n] | [\n "brass ring.c... | | [\n "https://s jo... | |
| [] | 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... |

localhost:6298 6/23

3/20/25 , 11:38 PM Module 04: Lab 01

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 St rate... | [\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 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 |

localhost:6298 7/23

3/20/25 , 11:38 PM Module 04: Lab 01

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... |

+--------------------+-----------------+----------------------+----------+----------+-

---------+--------+--------------------+--------------------+--------------------+----

-------+-------------------+--------------------+--------------------+---------------

+----------------+--------+--------------------+-----------+-------------------+------

----------+---------------------+-------------+-------------------+-------------+-----

-------------+---------------+--------------------+--------------------+--------------

------+-------------+------+-----------+----------------+-------------------+---------

+-----------+--------------------+--------------------+-------------+------+----------

----+-----+--------------------+-----+----------+---------------+--------------------

+---------------+--------------------+------------+--------------------+------------+-

-------------------+------+--------------------+------+--------------------+------+---

-----------------+------+--------------------+------+--------------------+------------

------+-------------------+--------------------+--------------------+-----------------

---+--------------------+-----------------------+--------------------+----------------

----+--------------------+--------------------+--------------------+------------------

--+----------+--------------------+----------+--------------------+-------------------

-+--------------------+--------------------+--------------------+--------------------

+ +----------+--------------------+----------+--------------------

+ +---------------+----------+---------------+---------------+---------------

+--------------+--------------------+--------------------------+-----------------

+--------------------+-------------------------+------------------------

localhost:6298 8/[23](#bookmark11)

3/20/25 , 11:38 PM Module 04: Lab 01

+----------------------------------+-----------------+----------------------+----

+----------------------------+------------------+------------------

+-------+--------------------+-------+--------------------+-------+--------------

-+-------+---------------+-----------------+----------------------+------------+------

--------------+------------+--------------------+------------+--------------------+---

---------+--------------------+------------+--------------------+

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

o 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")

localhost:6298 9/23

**Salary** **Distribution** **by** **Employment** **Type**

Salary From (USD)

800k 

700k 

600k 

500k 



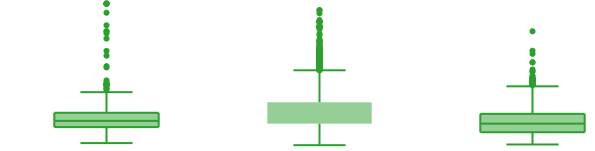
400k 

300k

200k

100k

0



|  |
| --- |
|  |
|  |

Part-time / full-time Full-time (> 32 hours) Part-time (≤ 32 hou

Employment Type

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

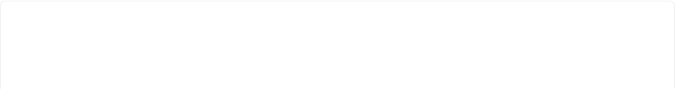
o Keep records where salary is greater than zero. . Aggregate Data

o Group by NAICS industry codes. . Visualize results

o 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.



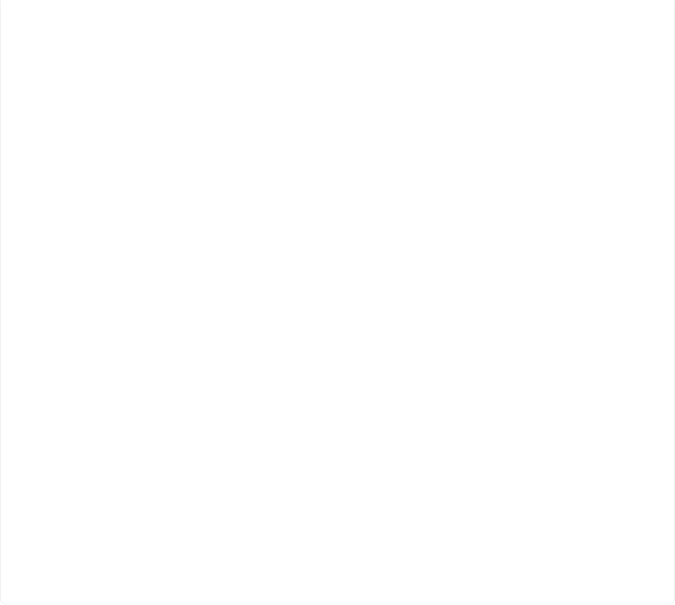
# Filter SALARY\_FROM > 0 and NAICS2\_NAME not null

df\_industry\_salary = df.select("NAICS2\_NAME", "SALARY\_FROM") \

.filter((col("SALARY\_FROM").isNotNull()) & (col("SALARY\_FROM") > 0) & (col

localhost:6298 10/23

3/20/25 , 11:38 PM Module 04: Lab 01

# 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")

**Salary** **Distribution** **by** **Industry** **(NAICS2)**

Salary From (USD)

0 



200k 

800k  600k  400k 



localhost:6298 11/23

3/20/25 , 11:38 PM

Module 04: Lab 01

Industry (NAICS2)

ces

n)

agement and Remediation Services

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

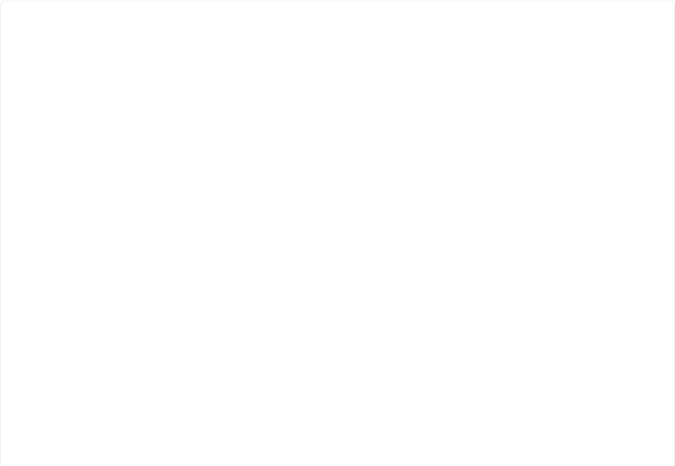
. Analyze how job postings fluctuate over time. . Aggregate Data

。 Count job postings per posted date ( **POSTED** ). . Visualize results

。 Create a line chart where:  X-axis = POSTED

 Y-axis = Number of Job Postings o 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",

localhost:6298 12/23

3/20/25 , 11:38 PM Module 04: Lab 01

|  |
| --- |
| 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") |

**Job** **Posting** **Trends** **Over** **Time**

1000

Number of Job Postings

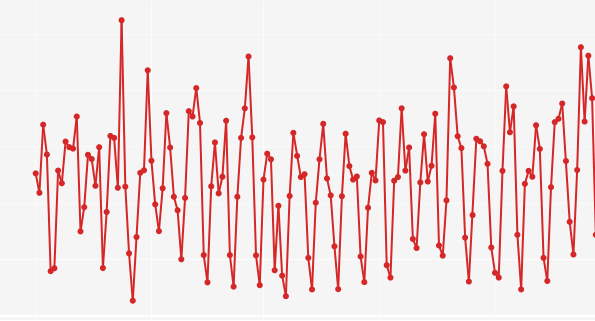
800

600

400

200

0



May 2024 Jun 2024 Jul 2024 Aug 2024 Sep 2024 Oc

Date Posted

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

o 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

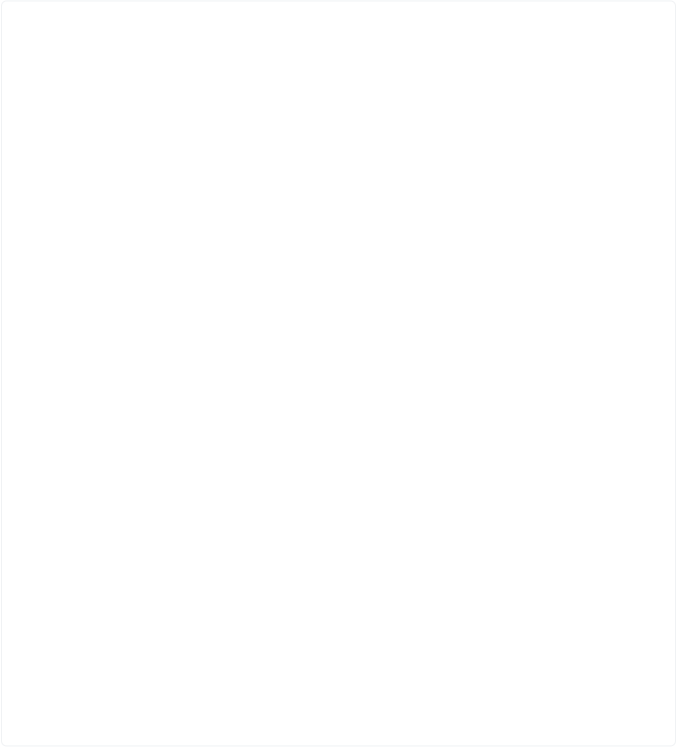
localhost:6298 13/23

3/20/25 , 11:38 PM Module 04: Lab 01

 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("un

# 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")

**Top** **10** **Job** **Titles** **by** **Count** **(Excluding** **Unclassified)**

9000

8000

8593



localhost:6298

14/23

3/20/25 , 11:38 PM Module 04: Lab 01

7000

Number of Postings

6000

5000

4000

3000

2000

1000

0

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  | 2074 | 1999 | |
|  |  | 1042 | |
|  |  | 668 629 | 537 488 467 463 |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

Job Title

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

o 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() |

localhost:6298 15/23

3/20/25 , 11:38 PM Module 04: Lab 01

|  |
| --- |
| # 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") |

**Remote** **vs** **On-Site** **Job** **Postings**

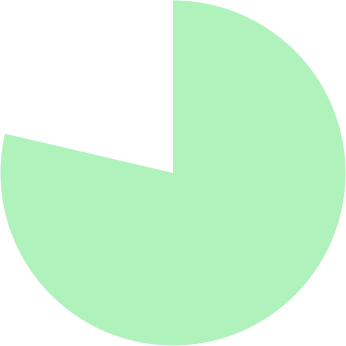


Hybrid Remote

14.2%

Not Remote

7.13%

Remote 78.6%

localhost:6298 16/23

3/20/25 , 11:38 PM Module 04: Lab 01

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

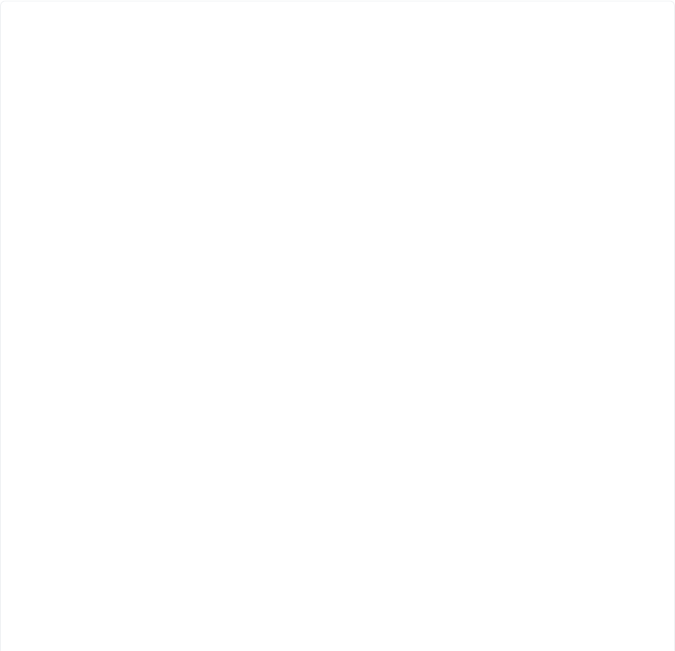
。 Create a stacked bar chart where:

 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

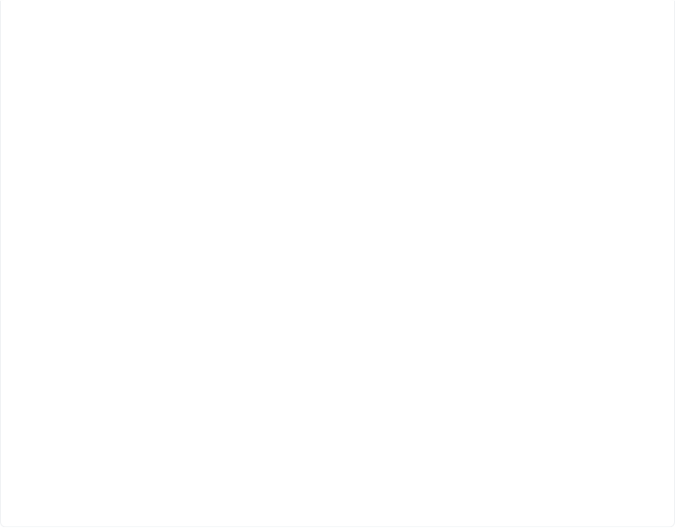
industry\_totals = skill\_counts\_filtered.groupby("Industry") ["Count"].sum().nla

# Filter to top 10 industries

skill\_counts\_top\_industries = skill\_counts\_filtered [skill\_counts\_filtered ["Ind

localhost:6298 17/23

3/20/25 , 11:38 PM Module 04: Lab 01



# 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)

**Top** **5** **In-Demand** **Skills** **by** **Industry** **(Top** **10** **Industrie**

Skill Count

10130

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
|  |

|  |
| --- |
| 7888 |

|  |
| --- |
| 4850 |

|  |
| --- |
| 4645 |

|  |
| --- |
| 3973 |

|  |
| --- |
| 2078 |

|  |
| --- |
| 950 |

|  |
| --- |
| 3477 |

|  |
| --- |
| 1910 |

|  |
| --- |
| 3477 |

|  |
| --- |
| 1222 |

|  |
| --- |
| 3521 |

|  |
| --- |
| 3102 |

|  |
| --- |
| 2382 |

|  |
| --- |
| 2334 |

|  |
| --- |
| 2276 |

|  |
| --- |
| 2022 |

|  |
| --- |
| 1599 |

|  |
| --- |
| 1571 |

|  |
| --- |
| 1421 |

|  |
| --- |
| 1391 |

|  |
| --- |
| 1368 |

|  |
| --- |
| 1310 |

|  |
| --- |
| 1178 |

|  |
| --- |
| 1150 |

|  |
| --- |
| 1133 |

|  |
| --- |
| 1054 |

|  |
| --- |
| 1018 |

|  |
| --- |
| 993 |

|  |
| --- |
| 978 |

|  |
| --- |
| 807 |

|  |
| --- |
| 799 |

25k

20k

30k



10k

15k

|  |
| --- |
| 627 |

|  |
| --- |
| 618 |

|  |
| --- |
| 547 |

|  |
| --- |
| 472 |

5k

|  |
| --- |
| 181 |
| 252 |

|  |
| --- |
| 428 |

|  |
| --- |
| 425 |

|  |
| --- |
| 423 |

|  |
| --- |
| 419 |



|  |
| --- |
| 398 |



|  |
| --- |
| 254 |

0

771

715

703

690

685

Skill

 C  L  M  O  P

localhost:6298 18/23

Industry (NAICS2)

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

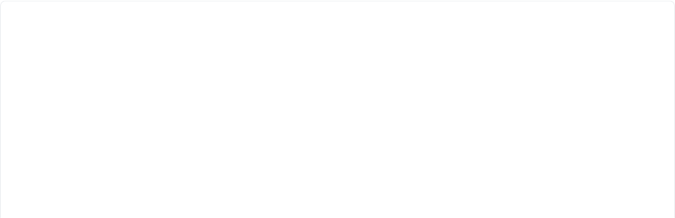
。 Compute median salary for each occupation in the ONET taxonomy. . Visualize results

o Create a bubble chart where:  X-axis = ONET\_NAME

 Y-axis = Median Salary

 Size = Number of job postings o 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()

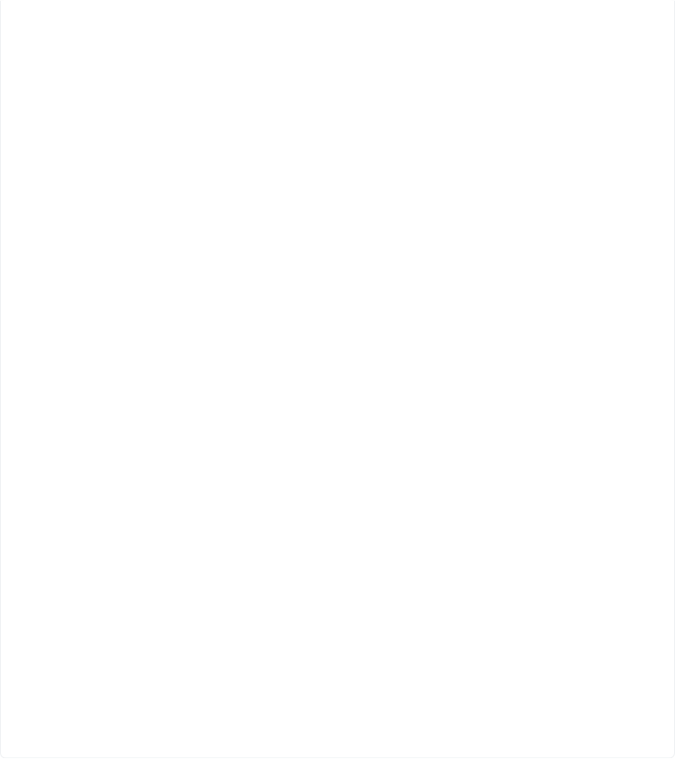
# Group by ONET occupation: calculate median salary and job count

onet\_salary\_stats = pdf\_onet\_salary.groupby("ONET\_NAME").agg(

Median\_Salary=("SALARY\_FROM", "median"),

localhost:6298 19/23

3/20/25 , 11:38 PM Module 04: Lab 01

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")

ary (USD)

localhost:6298

**Salary** **Analysis** **by** **ONET** **Occupation** **Type**

|  |  |
| --- | --- |
|  |  |
|  |  |



88.0005k

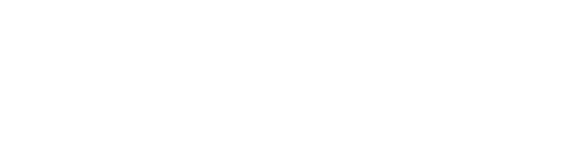
88.001k

32408

20/23

3/20/25 , 11:38 PM Module 04: Lab 01

Med ian Sala



88k

87.9995k

87.999k



ONET Occupation Type

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)

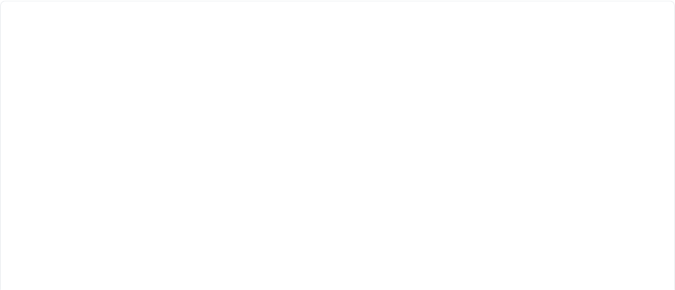
. Visualize job transitions between different occupation levels. . Aggregate Data

。 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

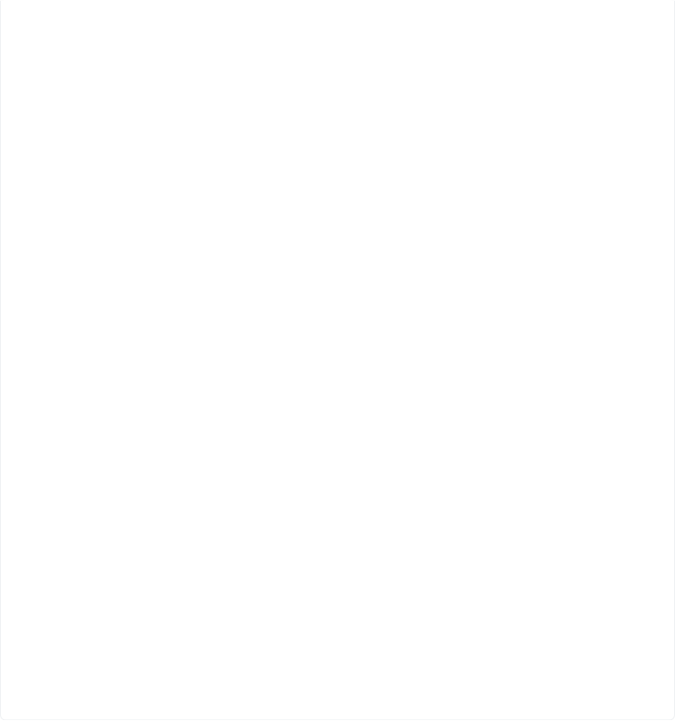
soc\_counts\_alt = pdf\_soc\_alt.groupby( ["SOC\_2021 3 NAME", "SOC\_2021 4 NAME"]).s

# 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

localhost:6298 21/23

3/20/25 , 11:38 PM Module 04: Lab 01

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**





localhost:6298 22/23

3/20/25 , 11:38 PM Module 04: Lab 01



Mathematical Science Occupations

Data Scienti

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.

localhost:6298 23/23