Assignment 03

Emily Sundberg

September 24, 2025

# 1. Load the Data Set

import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
from pyspark.sql import SparkSession  
import re  
import numpy as np  
import plotly.graph\_objects as go  
from pyspark.sql.functions import col, split, explode, regexp\_replace, transform, when  
from pyspark.sql import functions as F  
from pyspark.sql.functions import col, monotonically\_increasing\_id  
  
np.random.seed(42)  
  
pio.renderers.default = "notebook"  
  
spark = SparkSession.builder.appName("LightcastData").getOrCreate()  
  
df = spark.read.option("header", "true").option("inferSchema", "true").option("multiLine","true").option("escape", "\"").csv("/home/ubuntu/assignment-03-emily-sundberg-1/data/lightcast\_job\_postings.csv")  
df.createOrReplaceTempView("job\_postings")  
  
#print("---This is Diagnostic check, No need to print it in the final doc---")  
  
#df.printSchema() # comment this line when rendering the submission  
df.show(5)

[Stage 29:> (0 + 1) / 1]

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| ID|LAST\_UPDATED\_DATE|LAST\_UPDATED\_TIMESTAMP|DUPLICATES| POSTED| EXPIRED|DURATION| SOURCE\_TYPES| SOURCES| URL|ACTIVE\_URLS|ACTIVE\_SOURCES\_INFO| TITLE\_RAW| BODY|MODELED\_EXPIRED|MODELED\_DURATION| COMPANY| COMPANY\_NAME|COMPANY\_RAW|COMPANY\_IS\_STAFFING|EDUCATION\_LEVELS|EDUCATION\_LEVELS\_NAME|MIN\_EDULEVELS| MIN\_EDULEVELS\_NAME|MAX\_EDULEVELS|MAX\_EDULEVELS\_NAME|EMPLOYMENT\_TYPE|EMPLOYMENT\_TYPE\_NAME|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|IS\_INTERNSHIP|SALARY|REMOTE\_TYPE|REMOTE\_TYPE\_NAME|ORIGINAL\_PAY\_PERIOD|SALARY\_TO|SALARY\_FROM| LOCATION| CITY| CITY\_NAME|COUNTY| COUNTY\_NAME| MSA| MSA\_NAME|STATE|STATE\_NAME|COUNTY\_OUTGOING|COUNTY\_NAME\_OUTGOING|COUNTY\_INCOMING|COUNTY\_NAME\_INCOMING|MSA\_OUTGOING| MSA\_NAME\_OUTGOING|MSA\_INCOMING| MSA\_NAME\_INCOMING|NAICS2| NAICS2\_NAME|NAICS3| NAICS3\_NAME|NAICS4| NAICS4\_NAME|NAICS5| NAICS5\_NAME|NAICS6| NAICS6\_NAME| TITLE| TITLE\_NAME| TITLE\_CLEAN| SKILLS| SKILLS\_NAME| SPECIALIZED\_SKILLS|SPECIALIZED\_SKILLS\_NAME| CERTIFICATIONS| CERTIFICATIONS\_NAME| COMMON\_SKILLS| COMMON\_SKILLS\_NAME| SOFTWARE\_SKILLS|SOFTWARE\_SKILLS\_NAME| ONET| ONET\_NAME| ONET\_2019| ONET\_2019\_NAME| CIP6| CIP6\_NAME| CIP4| CIP4\_NAME| CIP2| CIP2\_NAME|SOC\_2021\_2| SOC\_2021\_2\_NAME|SOC\_2021\_3| SOC\_2021\_3\_NAME|SOC\_2021\_4|SOC\_2021\_4\_NAME|SOC\_2021\_5|SOC\_2021\_5\_NAME|LOT\_CAREER\_AREA|LOT\_CAREER\_AREA\_NAME|LOT\_OCCUPATION| LOT\_OCCUPATION\_NAME|LOT\_SPECIALIZED\_OCCUPATION|LOT\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_OCCUPATION\_GROUP|LOT\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_SPECIALIZED\_OCCUPATION|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION|LOT\_V6\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION\_GROUP|LOT\_V6\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_CAREER\_AREA|LOT\_V6\_CAREER\_AREA\_NAME| SOC\_2| SOC\_2\_NAME| SOC\_3| SOC\_3\_NAME| SOC\_4| SOC\_4\_NAME| SOC\_5| SOC\_5\_NAME|LIGHTCAST\_SECTORS|LIGHTCAST\_SECTORS\_NAME|NAICS\_2022\_2| NAICS\_2022\_2\_NAME|NAICS\_2022\_3| NAICS\_2022\_3\_NAME|NAICS\_2022\_4| NAICS\_2022\_4\_NAME|NAICS\_2022\_5| NAICS\_2022\_5\_NAME|NAICS\_2022\_6| NAICS\_2022\_6\_NAME|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
|1f57d95acf4dc67ed...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 6/8/2024| 6| [\n "Company"\n]|[\n "brassring.c...|[\n "https://sjo...| []| NULL|Enterprise Analys...|31-May-2024\n\nEn...| 6/8/2024| 6| 894731| Murphy USA| Murphy USA| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 2| 2| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.20...|RWwgRG9yYWRvLCBBUg==|El Dorado, AR| 5139| Union, AR|20980| El Dorado, AR| 5| Arkansas| 5139| Union, AR| 5139| Union, AR| 20980| El Dorado, AR| 20980| El Dorado, AR| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...|441330|Automotive Parts ...|ET29C073C03D1F86B4|Enterprise Analysts|enterprise analys...|[\n "KS126DB6T06...|[\n "Merchandisi...|[\n "KS126DB6T06...| [\n "Merchandisi...| []| []|[\n "KS126706DPF...|[\n "Mathematics...|[\n "KS440W865GC...|[\n "SQL (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "45.0601",\n...|[\n "Economics, ...|[\n "45.06",\n ...|[\n "Economics",...|[\n "45",\n "27...|[\n "Social Scie...| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101011| General ERP Analy...| 2310| Business Intellig...| 23101011| General ERP Analy...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 7\n]| [\n "Artificial ...| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...| 441330|Automotive Parts ...|  
|0cb072af26757b6c4...| 8/2/2024| 2024-08-02 17:08:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Job Board"\n]| [\n "maine.gov"\n]|[\n "https://job...| []| NULL|Oracle Consultant...|Oracle Consultant...| 8/1/2024| NULL| 133098|Smx Corporation L...| SMX| true| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| 3| false| NULL| 1| Remote| NULL| NULL| NULL|{\n "lat": 44.31...| QXVndXN0YSwgTUU=| Augusta, ME| 23011| Kennebec, ME|12300|Augusta-Watervill...| 23| Maine| 23011| Kennebec, ME| 23011| Kennebec, ME| 12300|Augusta-Watervill...| 12300|Augusta-Watervill...| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...|561320|Temporary Help Se...|ET21DDA63780A7DC09| Oracle Consultants|oracle consultant...|[\n "KS122626T55...|[\n "Procurement...|[\n "KS122626T55...| [\n "Procurement...| []| []| []| []|[\n "BGSBF3F508F...|[\n "Oracle Busi...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...| 561320|Temporary Help Se...|  
|85318b12b3331fa49...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024| 7/7/2024| 35| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL| Data Analyst|Taking care of pe...| 6/10/2024| 8|39063746| Sedgwick| Sedgwick| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 5| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 32.77...| RGFsbGFzLCBUWA==| Dallas, TX| 48113| Dallas, TX|19100|Dallas-Fort Worth...| 48| Texas| 48113| Dallas, TX| 48113| Dallas, TX| 19100|Dallas-Fort Worth...| 19100|Dallas-Fort Worth...| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...|524291| Claims Adjusting|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS1218W78FG...|[\n "Management"...|[\n "ESF3939CE1F...| [\n "Exception R...|[\n "KS683TN76T7...|[\n "Security Cl...|[\n "KS1218W78FG...|[\n "Management"...|[\n "KS126HY6YLT...|[\n "Microsoft O...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...| 524291| Claims Adjusting|  
|1b5c3941e54a1889e...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024|7/20/2024| 48| [\n "Job Board"\n]|[\n "disabledper...|[\n "https://www...| []| NULL|Sr. Lead Data Mgm...|About this role:\...| 6/12/2024| 10|37615159| Wells Fargo|Wells Fargo| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.44...| UGhvZW5peCwgQVo=| Phoenix, AZ| 4013| Maricopa, AZ|38060|Phoenix-Mesa-Chan...| 4| Arizona| 4013| Maricopa, AZ| 4013| Maricopa, AZ| 38060|Phoenix-Mesa-Chan...| 38060|Phoenix-Mesa-Chan...| 52|Finance and Insur...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking|522110| Commercial Banking|ET2114E0404BA30075|Management Analysts|sr lead data mgmt...|[\n "KS123QX62QY...|[\n "Exit Strate...|[\n "KS123QX62QY...| [\n "Exit Strate...| []| []|[\n "KS7G6NP6R6L...|[\n "Reliability...|[\n "KS4409D76NW...|[\n "SAS (Softwa...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 6\n]| [\n "Data Privac...| 52|Finance and Insur...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking| 522110| Commercial Banking|  
|cb5ca25f02bdf25c1...| 6/19/2024| 2024-06-19 07:00:00| 0|6/2/2024|6/17/2024| 15|[\n "FreeJobBoar...|[\n "craigslist....|[\n "https://mod...| []| NULL|Comisiones de $10...|Comisiones de $10...| 6/17/2024| 15| 0| Unclassified| LH/GM| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 3|Part-time / full-...| NULL| NULL| false| 92500| 0| [None]| year| 150000| 35000|{\n "lat": 37.63...| TW9kZXN0bywgQ0E=| Modesto, CA| 6099|Stanislaus, CA|33700| Modesto, CA| 6|California| 6099| Stanislaus, CA| 6099| Stanislaus, CA| 33700| Modesto, CA| 33700| Modesto, CA| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET0000000000000000| Unclassified|comisiones de por...| []| []| []| []| []| []| []| []| []| []|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
only showing top 5 rows

# 2. Data Cleaning

## 2.1 Casting salary and experience columns

df = df.withColumn("SALARY\_FROM", col ("SALARY\_FROM").cast("float"))\  
 .withColumn("SALARY\_TO", col("SALARY\_TO").cast("float")) \  
 .withColumn("MAX\_YEARS\_EXPERIENCE", col("MAX\_YEARS\_EXPERIENCE").cast("float"))\  
 .withColumn("MIN\_YEARS\_EXPERIENCE", col("MIN\_YEARS\_EXPERIENCE").cast("float"))\  
 .withColumn("SALARY", col("SALARY").cast("float"))

## 2.2 Computing medians for salary columns

def compute\_median(sdf, col\_name):  
 q = sdf.approxQuantile(col\_name, [0.5], 0.01)  
 return q[0] if q else None  
  
  
median\_from = compute\_median(df, "SALARY\_FROM")  
median\_to = compute\_median(df, "SALARY\_TO")  
median\_salary = compute\_median(df,"SALARY")  
  
print("Medians:", median\_from, median\_to, median\_salary)

[Stage 31:> (0 + 1) / 1] [Stage 32:> (0 + 1) / 1] [Stage 33:> (0 + 1) / 1]

Medians: 87295.0 130042.0 115024.0

## 2.3 Impute missing salaries

df = df.fillna({  
 "SALARY\_FROM": median\_from,  
 "SALARY\_TO": median\_to,  
 "SALARY": median\_salary  
})

## 2.4 Cleaning Education Column

df = df.withColumn("EDUCATION\_LEVELS\_NAME", regexp\_replace(col("EDUCATION\_LEVELS\_NAME"), "[\n\r]", ""))  
ed = df.select("EDUCATION\_LEVELS\_NAME")  
ed.show(15)

+---------------------+  
|EDUCATION\_LEVELS\_NAME|  
+---------------------+  
| [ "Bachelor's de...|  
| [ "No Education ...|  
| [ "Bachelor's de...|  
| [ "No Education ...|  
| [ "No Education ...|  
| [ "Bachelor's de...|  
| [ "Bachelor's de...|  
| [ "Bachelor's de...|  
| [ "No Education ...|  
| [ "Bachelor's de...|  
| [ "High school o...|  
| [ "No Education ...|  
| [ "Bachelor's de...|  
| [ "Bachelor's de...|  
| [ "No Education ...|  
+---------------------+  
only showing top 15 rows

## 2.5 Compute Average Salary

df = df.withColumn("AVG\_SALARY", (col("SALARY\_FROM")+col("SALARY\_TO"))/2)

## 2.6 Exporting Cleaned Data

df\_selected = df.select(  
 "EDUCATION\_LEVELS\_NAME",  
 "REMOTE\_TYPE\_NAME",  
 "MAX\_YEARS\_EXPERIENCE",  
 "AVG\_SALARY",  
 "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME",  
 "NAICS2\_NAME")

## 2.7 Save to CSV

pdf = df\_selected.toPandas()  
  
pdf.to\_csv("./data/lightcast\_cleaned.csv", index=False)  
  
print("Data Cleaning Complete. Rows retained:", len(pdf))

[Stage 35:> (0 + 1) / 1]

Data Cleaning Complete. Rows retained: 72498

# 3. Salary Distribution by Industry and Employment Type

## 3.1 Remove records where salary is missing or zero

pdf = df.filter(df["SALARY"] >0).select("EMPLOYMENT\_TYPE\_NAME", "SALARY","NAICS2\_NAME").toPandas()  
pdf["EMPLOYMENT\_TYPE\_NAME"] = pdf["EMPLOYMENT\_TYPE\_NAME"].fillna("Unknown")  
pdf["EMPLOYMENT\_TYPE\_NAME"] = pdf["EMPLOYMENT\_TYPE\_NAME"].apply(lambda x: re.sub(r"[^\x00-\x7F]+","",x))  
  
pdf.head()

[Stage 36:> (0 + 1) / 1]

|  | EMPLOYMENT\_TYPE\_NAME | SALARY | NAICS2\_NAME |
| --- | --- | --- | --- |
| 0 | Full-time (> 32 hours) | 115024.0 | Retail Trade |
| 1 | Full-time (> 32 hours) | 115024.0 | Administrative and Support and Waste Managemen... |
| 2 | Full-time (> 32 hours) | 115024.0 | Finance and Insurance |
| 3 | Full-time (> 32 hours) | 115024.0 | Finance and Insurance |
| 4 | Part-time / full-time | 92500.0 | Unclassified Industry |

## 3.2 Aggregate Data

median\_salaries = pdf.groupby("EMPLOYMENT\_TYPE\_NAME")["SALARY"].median()  
  
sorted\_employment\_types = median\_salaries.sort\_values(ascending = False).index  
  
pdf["EMPLOYMENT\_TYPE\_NAME"] = pd.Categorical(  
 pdf["EMPLOYMENT\_TYPE\_NAME"],  
 categories=sorted\_employment\_types,  
 ordered=True  
)

## 3.3 Visualize Results

import matplotlib.pyplot as plt  
import seaborn as sns  
import plotly.express as px  
  
fig = px.box(  
 pdf,  
 x = "NAICS2\_NAME",  
 y = "SALARY",  
 title = "Salary Distribution by NAICS2 Name",  
 color\_discrete\_sequence = ["peru"],  
 boxmode = "group",  
 points = "outliers"   
)

fig.update\_layout(  
 margin=dict(t=150),  
 width=2000,   
 height=800,  
 title=dict(  
 text = "Salary Distribution By NAICS2 Name",  
 font=dict(size=30, family="Montserrat", color = "saddlebrown", weight="bold")  
 ),  
  
 xaxis=dict(  
 title=dict(text="NAICS2 Name", font=dict(size=14, family="Montserrat", color="saddlebrown", weight="bold")),  
 tickangle=50,  
 showline=True,  
 linewidth=2,  
 linecolor="saddlebrown",  
 mirror=True,  
 showgrid=False,  
 categoryorder="array",  
 categoryarray=sorted\_employment\_types.tolist()  
 ),  
  
 yaxis=dict(  
 title=dict(text="Salary (K $)", font=dict(size=14, family="Montserrat", color="saddlebrown", weight="bold")),  
 tickvals=[0,50000,100000,150000,200000,250000,300000,350000,400000,450000,500000],  
 ticktext=["0","50K","100K","150K","200K","250K","300K","350K","400K","450K","500K"],  
 tickfont=dict(size=12, family="Montserrat", color="saddlebrown",weight="bold"),  
 showline=True,  
 linewidth=2,  
 linecolor="saddlebrown",  
 mirror=True,  
 showgrid=True,  
 gridcolor="tan",  
 gridwidth=0.5  
 )  
)  
  
fig.show()

Unable to display output for mime type(s): text/html

Unable to display output for mime type(s): text/html

The plot above shows us that the medians of these industries average around 100K but there are several career fields with the opportunity to make significantly more. Both admin support and information have outliers in the $500K range.

import matplotlib.pyplot as plt  
import seaborn as sns  
import plotly.express as px  
  
fig = px.box(  
 pdf,  
 x = "EMPLOYMENT\_TYPE\_NAME",  
 y = "SALARY",  
 title = "Salary Distribution by Employment Type",  
 color\_discrete\_sequence = ["peru"],  
 boxmode = "group",  
 points = "outliers"   
)

fig.update\_layout(  
 margin=dict(t=150),  
 width=2000,   
 height=800,  
 title=dict(  
 text = "Salary Distribution By Employment Type",  
 font=dict(size=30, family="Montserrat", color = "saddlebrown", weight="bold")  
 ),  
  
 xaxis=dict(  
 title=dict(text="Employment Type", font=dict(size=14, family="Montserrat", color="saddlebrown", weight="bold")),  
 tickangle=50,  
 showline=True,  
 linewidth=2,  
 linecolor="saddlebrown",  
 mirror=True,  
 showgrid=False,  
 categoryorder="array",  
 categoryarray=sorted\_employment\_types.tolist()  
 ),  
  
 yaxis=dict(  
 title=dict(text="Salary (K $)", font=dict(size=14, family="Montserrat", color="saddlebrown", weight="bold")),  
 tickvals=[0,50000,100000,150000,200000,250000,300000,350000,400000,450000,500000],  
 ticktext=["0","50K","100K","150K","200K","250K","300K","350K","400K","450K","500K"],  
 tickfont=dict(size=12, family="Montserrat", color="saddlebrown",weight="bold"),  
 showline=True,  
 linewidth=2,  
 linecolor="saddlebrown",  
 mirror=True,  
 showgrid=True,  
 gridcolor="tan",  
 gridwidth=0.5  
 )  
)  
  
fig.show()

Unable to display output for mime type(s): text/html

# 4. Salary Analysis by ONET Occupation Type

## 4.1 Aggregate Data

salary\_analysis = spark.sql("""  
 SELECT  
 LOT\_OCCUPATION\_NAME AS Occupation\_Name,  
 PERCENTILE(SALARY, 0.5) AS Median\_Salary,  
 Count(\*) AS Job\_Postings  
 FROM job\_postings  
 GROUP BY LOT\_OCCUPATION\_NAME  
 ORDER BY Job\_Postings DESC  
 LIMIT 10  
""")  
  
salary\_pd = salary\_analysis.toPandas()  
salary\_pd.head()

[Stage 37:> (0 + 1) / 1]

|  | Occupation\_Name | Median\_Salary | Job\_Postings |
| --- | --- | --- | --- |
| 0 | Data / Data Mining Analyst | 95250.0 | 30057 |
| 1 | Business Intelligence Analyst | 125900.0 | 29445 |
| 2 | Computer Systems Engineer / Architect | 157600.0 | 8212 |
| 3 | Business / Management Analyst | 93650.0 | 4326 |
| 4 | Clinical Analyst / Clinical Documentation and ... | 89440.0 | 261 |

## 4.2 Bubble Plot

fig = px.scatter(  
 salary\_pd,  
 x="Occupation\_Name",  
 y="Median\_Salary",  
 size="Job\_Postings",  
 title = "Salary Analysis by Occupation (Bubble Chart)",  
 labels={  
 "Occupation\_Name": "Occupation",  
 "Median\_Salary" : "Median Salary",  
 "Job\_Postings" : "Number of Job Postings"  
 },  
 hover\_name = "Occupation\_Name",  
 size\_max=60,  
 width=1000,  
 height=1000,  
 color="Job\_Postings",  
 color\_continuous\_scale="Sunsetdark"  
)

fig.update\_layout(  
 margin=dict(t=150),  
 font\_family="Montserrat",  
 font\_size = 14,  
 title\_font\_size = 25,  
 xaxis\_title = "Occupation",  
 yaxis\_title = "Median Salary",  
 plot\_bgcolor = "white",  
 xaxis=dict(  
 tickangle=-45,  
 showline=True,  
 linecolor="black"  
 ),  
 yaxis=dict(  
 showline=True,  
 linecolor="black"  
 )  
)  
  
fig.show()

Unable to display output for mime type(s): text/html

This figure shows us that, although data analysts and business analysts have roughly the same median salary, data analysts have far more job postings than their business analyst counterparts.

# 5. Salary by Education Level

## 5.1 Create Education Groups

lower\_degree = ["Bachelor's", "Associate","GED", "No Education Listed", "High School"]  
  
higher\_degree = ["Master's Degree", "PhD or professional degree"]  
  
df = df.withColumn(  
 "EDUCATION\_GROUP",  
 when(col("EDUCATION\_LEVELS\_NAME").rlike("|".join([f"(?i){deg}"for deg in lower\_degree])), "Bachelor's or lower")  
 .when(col("EDUCATION\_LEVELS\_NAME").rlike("|".join([f"(?i){deg}"for deg in higher\_degree])), "Master's or PhD")  
 .otherwise("Other")  
)  
  
df = df.filter(  
 col("MAX\_YEARS\_EXPERIENCE").isNotNull() &  
 col("AVG\_SALARY").isNotNull() &  
 (col("MAX\_YEARS\_EXPERIENCE")>0) &  
 (col("AVG\_SALARY") > 0)  
)  
  
df\_filtered = df.filter(col("EDUCATION\_GROUP").isin("Bachelor's or lower","Master's or PhD"))  
  
df\_pd = df\_filtered.toPandas()

[Stage 40:> (0 + 1) / 1]

## 5.2 Scatter Plot

fig = px.scatter(  
 df\_pd,  
 x = "MAX\_YEARS\_EXPERIENCE",  
 y = "AVG\_SALARY",  
 color = "EDUCATION\_GROUP",  
 hover\_data=["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"],  
 title = "Experience vs Salary by Education Level",  
 opacity=0.7,  
 color\_discrete\_sequence= ["palevioletred", "mediumseagreen"]  
)  
  
fig.update\_traces(marker=dict(size=7, line=dict(width=1, color="black")))  
  
fig.update\_layout(  
 plot\_bgcolor = "papayawhip",  
 font=dict(family = "Montserrat", size = 14),  
 title\_font = dict(size = 22, weight = "bold"),  
 xaxis\_title = "Years of Experience",  
 yaxis\_title = "Average Salary (USD)",  
 legend\_title = "Education Group",  
 hoverlabel = dict(bgcolor = "white", font\_size = 13, font\_family = "Montserrat"),  
 margin=dict(t=70, l=60, r=60),  
 xaxis=dict(  
 gridcolor="peru",  
 tickmode = 'linear',  
 dtick=1  
 ),  
 yaxis=dict(gridcolor = "peru")  
)  
  
fig.show()

Unable to display output for mime type(s): text/html

This graph is showing that as years of experience increases, the floor for average salary also increases. Additionally, it shows that Bachelor’s or lower appears to have significantly more outliers while Master’s or PhD is more consistent.

# 6. Salary By Remote Work Type

## 6.1 Split Work Type Groups

from pyspark.sql.functions import when, col, trim  
  
df = df.withColumn("REMOTE\_GROUP",  
 when(trim(col("REMOTE\_TYPE\_NAME"))== "Remote", "Remote")  
 .when(trim(col("REMOTE\_TYPE\_NAME"))== "Hybrid Remote", "Hybrid")  
 .when(trim(col("REMOTE\_TYPE\_NAME"))== "Not Remote", "Onsite")  
 .when(col("REMOTE\_TYPE\_NAME").isNull(), "Onsite")  
 .otherwise("Onsite")  
)  
  
df = df.filter(  
 col("MAX\_YEARS\_EXPERIENCE").isNotNull() &  
 col("AVG\_SALARY").isNotNull() &  
 (col("MAX\_YEARS\_EXPERIENCE") > 0) &  
 (col("AVG\_SALARY") > 0 )  
)  
  
df\_pd = df.select("MAX\_YEARS\_EXPERIENCE", "AVG\_SALARY", "LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME", "REMOTE\_GROUP").toPandas()

[Stage 41:> (0 + 1) / 1]

## 6.2 Scatter Plot

df\_pd["MAX\_EXPERIENCE\_JITTER"] = df\_pd["MAX\_YEARS\_EXPERIENCE"] + np.random.uniform(-0.3, 0.3, size = len(df\_pd))  
df\_pd["AVG\_SALARY\_JITTER"] = df\_pd["AVG\_SALARY"] + np.random.uniform(-1500, 1500, size = len(df\_pd))  
  
fig = px.scatter(  
 df\_pd,  
 width = 2000,  
 height = 800,  
 x = "MAX\_EXPERIENCE\_JITTER",  
 y = "AVG\_SALARY\_JITTER",  
 color = "REMOTE\_GROUP",  
 hover\_data=["LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME"],  
 title = "Experience vs Salary by Remote Work Type",  
 opacity = 0.7,  
 color\_discrete\_sequence=["palevioletred", "mediumseagreen","lightblue"]  
)  
  
fig.update\_traces(marker=dict(size=7, line=dict(width=1, color="black")))  
  
fig.update\_layout(  
 plot\_bgcolor = "papayawhip",  
 font=dict(family = "Montserrat", size = 14),  
 title\_font = dict(size = 22, weight = "bold"),  
 xaxis\_title = "Years of Experience",  
 yaxis\_title = "Average Salary (USD)",  
 legend\_title = "Remote Work Type",  
 hoverlabel = dict(bgcolor = "white", font\_size = 13, font\_family = "Montserrat"),  
 margin=dict(t=70, l=60, r=60),  
 xaxis=dict(  
 gridcolor="peru",  
 tickmode = 'linear',  
 dtick=1  
 ),  
 yaxis=dict(gridcolor = "peru")  
)  
  
fig.show()

Unable to display output for mime type(s): text/html

This graph shows us that there isn’t a significant difference in average salary across the three different remote work types but those outliers with the highest paying job postings are almost entirely onsite.

## 6.3 Histograms

### 6.3.1 Onsite

fig = px.histogram(  
 df\_pd[df\_pd["REMOTE\_GROUP"] == "Onsite"],  
 x="AVG\_SALARY",  
 nbins=20,  
 title=f"Salary Distribution - Onsite",  
 color\_discrete\_sequence=["palevioletred"]  
)  
  
fig.update\_layout(  
 plot\_bgcolor="papayawhip",  
 xaxis\_title="Average Salary (USD)",  
 yaxis\_title="Count of Job Postings",  
 font=dict(family="Montserrat", size=14)  
)  
fig.show()

Unable to display output for mime type(s): text/html

This histogram demonstrates that onsite job postings have an average salary of just over $100K with a few outliers ranging from roughly $300K - $800K.

### 6.3.2 Remote

fig = px.histogram(  
 df\_pd[df\_pd["REMOTE\_GROUP"] == "Remote"],  
 x="AVG\_SALARY",  
 nbins=20,  
 title=f"Salary Distribution - Remote",  
 color\_discrete\_sequence=["mediumseagreen"]  
)  
  
fig.update\_layout(  
 plot\_bgcolor="papayawhip",  
 xaxis\_title="Average Salary (USD)",  
 yaxis\_title="Count of Job Postings",  
 font=dict(family="Montserrat", size=14)  
)  
fig.show()

Unable to display output for mime type(s): text/html

This histogram paints a similar picture as the onsite histogram with the average salary being just over $100K. The cap for remote positions however seems to be around $300K.

### 6.3.3 Hybrid

fig = px.histogram(  
 df\_pd[df\_pd["REMOTE\_GROUP"] == "Hybrid"],  
 x="AVG\_SALARY",  
 nbins=20,  
 title=f"Salary Distribution - Hybrid",  
 color\_discrete\_sequence=["lightblue"]  
)  
  
fig.update\_layout(  
 plot\_bgcolor="papayawhip",  
 xaxis\_title="Average Salary (USD)",  
 yaxis\_title="Count of Job Postings",  
 font=dict(family="Montserrat", size=14)  
)  
fig.show()

Unable to display output for mime type(s): text/html

The Hybrid histogram looks very similar to the remote histogram with less lucrative outliers but there is a large difference in the drastic decline in number of job postings after ~$125K. It appears that after that point (or so), the number of job postings dwindles rapidly. It’s also worth noting that there are significantly less overall job postings for hybrid postions than fully remote or fully onsite.