Assignment 04

JulioVargas

October 7, 2025

from pyspark.sql import SparkSession  
import pandas as pd  
import plotly.express as px  
import plotly.io as pio  
import numpy as np  
  
np.random.seed(42)  
  
pio.renderers.default = "notebook+notebook\_connected+vscode"  
  
# 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("data/lightcast\_job\_postings.csv")  
  
# Show Schema and Sample Data  
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)  
print(df.count())  
#pd.set\_option("display.max\_rows", None)   
#pd.DataFrame(df.columns, columns=["Column Names"])

[Stage 142:> (0 + 1) / 1]

---This is Diagnostic check, No need to print it in the final doc---  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| 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

[Stage 144:> (0 + 1) / 1]

72498

Feature Engineering is a crucial step in preparing your data for machine learning. In this lab, we will focus on the following tasks:

1. Drop rows with missing values in the target variable and key features.
2. By now you are already familiar with the code and the data. Based on your understanding please choose any 3 (my code output has 10) variables as:
   1. three continuous variables and, MIN\_YEARS\_EXPERIENCE (total 4, use your best judgment!)
   2. two categorical.
   3. Your dependent variable (y) is SALARY.
3. Convert categorical variables into numerical representations using StringIndexer and OneHotEncoder.
4. Assemble features into a single vector using VectorAssembler.
5. Split the data into training and testing sets.
6. You can use pipeline to do the above steps in one go.
7. Create a new column MIN\_YEARS\_EXPERIENCE\_SQ by squaring the MIN\_YEARS\_EXPERIENCE column.
8. Assemble the polynomial features into a new vector column features\_poly using VectorAssembler.
9. Show the final structure of the DataFrame with the new features.

+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+----------------------+-------------------+------------------+  
|SALARY|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|DURATION|COMPANY\_IS\_STAFFING|IS\_INTERNSHIP|STATE\_NAME|REMOTE\_TYPE\_NAME|EMPLOYMENT\_TYPE\_NAME |MIN\_EDULEVELS\_NAME |MAX\_EDULEVELS\_NAME|  
+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+----------------------+-------------------+------------------+  
|NULL |2 |2 |6 |false |false |Arkansas |[None] |Full-time (> 32 hours)|Bachelor's degree |NULL |  
|NULL |3 |3 |NULL |true |false |Maine |Remote |Full-time (> 32 hours)|No Education Listed|NULL |  
|NULL |5 |NULL |35 |false |false |Texas |[None] |Full-time (> 32 hours)|Bachelor's degree |NULL |  
|NULL |3 |NULL |48 |false |false |Arizona |[None] |Full-time (> 32 hours)|No Education Listed|NULL |  
|92500 |NULL |NULL |15 |false |false |California|[None] |Part-time / full-time |No Education Listed|NULL |  
+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+----------------------+-------------------+------------------+  
only showing top 5 rows

from pyspark.sql.functions import col, sum as spark\_sum, when, trim, length  
import hvplot.pandas # enables hvplot on pandas  
  
missing\_df = df\_eda.select([  
 spark\_sum(  
 when(col(c).isNull() | (length(trim(col(c))) == 0), 1)  
 .otherwise(0)  
 ).alias(c)  
 for c in df\_eda.columns  
])  
  
#print(missing\_df.show())  
  
#to table with T Transpose  
missing\_pd = missing\_df.toPandas().T.reset\_index()  
#put names to columns  
missing\_pd.columns = ["column", "missing\_count"]  
  
total\_rows = df\_eda.count()  
missing\_pd["missing\_pct"] = 100 \* missing\_pd["missing\_count"] / total\_rows  
  
# hvplot.bar ; line; scatter; (hist); (box); area; (heatmap); (hexbin); points  
missing\_pd.sort\_values("missing\_pct", ascending=False).hvplot.bar(  
 x="column", y="missing\_pct", rot=90,  
 title="Percentage of Missing Values by Column",  
 height=600, width=900,  
 ylabel="Missing Percentage (%)", xlabel="Features"  
).opts(xrotation=45)

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

Unable to display output for mime type(s): application/javascript, application/vnd.holoviews\_load.v0+json

Unable to display output for mime type(s): application/javascript, application/vnd.holoviews\_load.v0+json

Unable to display output for mime type(s): application/vnd.holoviews\_exec.v0+json, text/html

[Stage 148:> (0 + 1) / 1] [Stage 151:> (0 + 1) / 1]

:Bars [column] (missing\_pct)

# For REMOTE\_TYPE\_NAME replace Remote with Remote, [None] with Undefined,  
# Not Remote with On Premise, Hybrid Remote with Hybrid, and Null with On Premise  
## data frame (eda) exploratory data analysis  
  
df\_eda = df\_eda.withColumn(  
 "REMOTE\_TYPE\_NAME",  
 when(col("REMOTE\_TYPE\_NAME") == "Remote", "Remote")  
 .when(col("REMOTE\_TYPE\_NAME") == "[None]", "Undefined")  
 .when(col("REMOTE\_TYPE\_NAME") == "Not Remote", "On-Premise")  
 .when(col("REMOTE\_TYPE\_NAME") == "Hybrid Remote", "Hybrid")  
 .when(col("REMOTE\_TYPE\_NAME").isNull(), "On-Premise")  
 .otherwise(col("REMOTE\_TYPE\_NAME"))  
)  
  
# df\_eda.createOrReplaceTempView("df\_eda")  
categorical\_cols = [  
 "REMOTE\_TYPE\_NAME"  
]  
  
for colname in categorical\_cols:  
 print(f"\n---- {colname} ----")  
 df\_eda.select(colname).distinct().show(10, truncate=False)

---- REMOTE\_TYPE\_NAME ----

[Stage 154:> (0 + 1) / 1]

+----------------+  
|REMOTE\_TYPE\_NAME|  
+----------------+  
|Remote |  
|On-Premise |  
|Hybrid |  
|Undefined |  
+----------------+

# ---- EMPLOYMENT\_TYPE\_NAME ----  
   
# +------------------------+  
# |EMPLOYMENT\_TYPE\_NAME |  
# +------------------------+  
# |Part-time / full-time |  
# |Part-time (â‰¤ 32 hours)|  
# |Full-time (> 32 hours) |  
# |NULL |  
# +------------------------+  
  
df\_eda = df\_eda.withColumn(  
 "EMPLOYMENT\_TYPE\_NAME",  
 when(col("EMPLOYMENT\_TYPE\_NAME") == "Part-time / full-time", "Flexible")  
 .when(col("EMPLOYMENT\_TYPE\_NAME") == "Part-time (â‰¤ 32 hours)", "Parttime")  
 .when(col("EMPLOYMENT\_TYPE\_NAME") == "Full-time (> 32 hours)", "Fulltime")  
 .when(col("EMPLOYMENT\_TYPE\_NAME").isNull(), "Fulltime")  
 .otherwise(col("EMPLOYMENT\_TYPE\_NAME"))  
)  
  
# df\_eda.createOrReplaceTempView("df\_eda")  
categorical\_cols = [  
 "EMPLOYMENT\_TYPE\_NAME"  
]  
  
for colname in categorical\_cols:  
 print(f"\n---- {colname} ----")  
 df\_eda.select(colname).distinct().show(10, truncate=False)

---- EMPLOYMENT\_TYPE\_NAME ----

[Stage 157:> (0 + 1) / 1]

+--------------------+  
|EMPLOYMENT\_TYPE\_NAME|  
+--------------------+  
|Flexible |  
|Fulltime |  
|Parttime |  
+--------------------+

# replace COMPANY\_IS\_STAFFING NULL with false, and IS\_INTERNSHIP NULL with false  
df\_eda = df\_eda.withColumn(  
 "COMPANY\_IS\_STAFFING",  
 when(col("COMPANY\_IS\_STAFFING").isNull(), False)  
 .otherwise(col("COMPANY\_IS\_STAFFING"))  
)  
  
df\_eda = df\_eda.withColumn(  
 "IS\_INTERNSHIP",  
 when(col("IS\_INTERNSHIP").isNull(), False)  
 .otherwise(col("IS\_INTERNSHIP"))  
)  
  
# df\_eda.createOrReplaceTempView("df\_eda")  
categorical\_cols = [  
 "COMPANY\_IS\_STAFFING", "IS\_INTERNSHIP"  
]  
  
for colname in categorical\_cols:  
 print(f"\n---- {colname} ----")  
 df\_eda.select(colname).distinct().show(10, truncate=False)

---- COMPANY\_IS\_STAFFING ----

[Stage 160:> (0 + 1) / 1]

+-------------------+  
|COMPANY\_IS\_STAFFING|  
+-------------------+  
|true |  
|false |  
+-------------------+  
  
  
---- IS\_INTERNSHIP ----

[Stage 163:> (0 + 1) / 1]

+-------------+  
|IS\_INTERNSHIP|  
+-------------+  
|true |  
|false |  
+-------------+

import pandas as pd  
  
# sample subset of data only 1% of the data  
df\_sample = df\_eda.sample(fraction=0.01, seed=42).toPandas()  
  
#print(df\_eda.count()) #72498  
#print(len(df\_sample)) #790  
  
# create new DataFrame where each cell missing (True) or not (False)  
missing\_mask = df\_sample.isnull()  
  
# Melt into long-form | 4 columns: index, column, is\_missing  
missing\_long = (  
 missing\_mask.reset\_index()  
 .melt(id\_vars="index", var\_name="column", value\_name="is\_missing")  
)  
  
# Convert boolean to int  
missing\_long["is\_missing"] = missing\_long["is\_missing"].astype(int)  
  
print(missing\_long)  
  
# Plot heatmap  
missing\_long.hvplot.heatmap(  
 x="column", y="index", C="is\_missing",  
 cmap="Reds", colorbar=False,  
 width=900, height=700,  
 title="Heatmap of Missing Values (Sample)"  
).opts(xrotation=45)

[Stage 166:> (0 + 1) / 1]

index column is\_missing  
0 0 SALARY 1  
1 1 SALARY 0  
2 2 SALARY 1  
3 3 SALARY 1  
4 4 SALARY 1  
... ... ... ...  
8685 785 MAX\_EDULEVELS\_NAME 0  
8686 786 MAX\_EDULEVELS\_NAME 0  
8687 787 MAX\_EDULEVELS\_NAME 1  
8688 788 MAX\_EDULEVELS\_NAME 1  
8689 789 MAX\_EDULEVELS\_NAME 1  
  
[8690 rows x 3 columns]

:HeatMap [column,index] (is\_missing)

from pyspark.sql.functions import countDistinct  
  
#show number of unique values per column  
df\_eda.select([  
 countDistinct(c).alias(c + "\_nunique")  
 for c in df\_eda.columns  
]).show(truncate=False)

[Stage 167:> (0 + 1) / 1]

+--------------+----------------------------+----------------------------+----------------+---------------------------+---------------------+------------------+------------------------+----------------------------+--------------------------+--------------------------+  
|SALARY\_nunique|MIN\_YEARS\_EXPERIENCE\_nunique|MAX\_YEARS\_EXPERIENCE\_nunique|DURATION\_nunique|COMPANY\_IS\_STAFFING\_nunique|IS\_INTERNSHIP\_nunique|STATE\_NAME\_nunique|REMOTE\_TYPE\_NAME\_nunique|EMPLOYMENT\_TYPE\_NAME\_nunique|MIN\_EDULEVELS\_NAME\_nunique|MAX\_EDULEVELS\_NAME\_nunique|  
+--------------+----------------------------+----------------------------+----------------+---------------------------+---------------------+------------------+------------------------+----------------------------+--------------------------+--------------------------+  
|6052 |16 |15 |60 |2 |2 |51 |4 |3 |6 |4 |  
+--------------+----------------------------+----------------------------+----------------+---------------------------+---------------------+------------------+------------------------+----------------------------+--------------------------+--------------------------+

categorical\_cols = [  
 "STATE\_NAME", "REMOTE\_TYPE\_NAME", "EMPLOYMENT\_TYPE\_NAME",  
 "MIN\_EDULEVELS\_NAME", "COMPANY\_IS\_STAFFING", "IS\_INTERNSHIP"  
]  
  
for colname in categorical\_cols:  
 print(f"\n---- {colname} ----")  
 df\_eda.select(colname).distinct().show(10, truncate=False)

---- STATE\_NAME ----

[Stage 173:> (0 + 1) / 1]

+------------+  
|STATE\_NAME |  
+------------+  
|Utah |  
|Hawaii |  
|Minnesota |  
|Ohio |  
|Arkansas |  
|Oregon |  
|Texas |  
|North Dakota|  
|Pennsylvania|  
|Connecticut |  
+------------+  
only showing top 10 rows  
  
---- REMOTE\_TYPE\_NAME ----

[Stage 176:> (0 + 1) / 1]

+----------------+  
|REMOTE\_TYPE\_NAME|  
+----------------+  
|Remote |  
|On-Premise |  
|Hybrid |  
|Undefined |  
+----------------+  
  
  
---- EMPLOYMENT\_TYPE\_NAME ----

[Stage 179:> (0 + 1) / 1]

+--------------------+  
|EMPLOYMENT\_TYPE\_NAME|  
+--------------------+  
|Flexible |  
|Fulltime |  
|Parttime |  
+--------------------+  
  
  
---- MIN\_EDULEVELS\_NAME ----

[Stage 182:> (0 + 1) / 1]

+----------------------------+  
|MIN\_EDULEVELS\_NAME |  
+----------------------------+  
|Bachelor's degree |  
|Ph.D. or professional degree|  
|High school or GED |  
|Master's degree |  
|No Education Listed |  
|Associate degree |  
|NULL |  
+----------------------------+  
  
  
---- COMPANY\_IS\_STAFFING ----

[Stage 185:> (0 + 1) / 1]

+-------------------+  
|COMPANY\_IS\_STAFFING|  
+-------------------+  
|true |  
|false |  
+-------------------+  
  
  
---- IS\_INTERNSHIP ----

[Stage 188:> (0 + 1) / 1]

+-------------+  
|IS\_INTERNSHIP|  
+-------------+  
|true |  
|false |  
+-------------+

# Calculate median of the Duration Column  
  
median\_duration = df\_eda.approxQuantile("DURATION", [0.5], 0.01)[0]  
  
# Check for missing values in Duration column and replace null with median  
  
df\_eda = df\_eda.withColumn(  
 "DURATION",  
 when(col("DURATION").isNull(), median\_duration)  
 .otherwise(col("DURATION"))  
) # Assuming median duration is 30 days

[Stage 191:> (0 + 1) / 1]

import pandas as pd  
  
# sample subset of data  
df\_sample = df\_eda.sample(fraction=0.10, seed=42).toPandas()  
  
# Boolean mask (True if missing)  
missing\_mask = df\_sample.isnull()  
  
# Melt into long-form  
missing\_long = (  
 missing\_mask.reset\_index()  
 .melt(id\_vars="index", var\_name="column", value\_name="is\_missing")  
)  
  
# Convert boolean to int  
missing\_long["is\_missing"] = missing\_long["is\_missing"].astype(int)  
  
# Plot heatmap  
missing\_long.hvplot.heatmap(  
 x="column", y="index", C="is\_missing",  
 cmap="Reds", colorbar=False,  
 width=900, height=700,  
 title="Heatmap of Missing Values (Sample)"  
).opts(xrotation=45)

[Stage 192:> (0 + 1) / 1]

:HeatMap [column,index] (is\_missing)

df\_eda.show(5, truncate=False)

+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+--------------------+-------------------+------------------+  
|SALARY|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|DURATION|COMPANY\_IS\_STAFFING|IS\_INTERNSHIP|STATE\_NAME|REMOTE\_TYPE\_NAME|EMPLOYMENT\_TYPE\_NAME|MIN\_EDULEVELS\_NAME |MAX\_EDULEVELS\_NAME|  
+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+--------------------+-------------------+------------------+  
|NULL |2 |2 |6.0 |false |false |Arkansas |Undefined |Fulltime |Bachelor's degree |NULL |  
|NULL |3 |3 |18.0 |true |false |Maine |Remote |Fulltime |No Education Listed|NULL |  
|NULL |5 |NULL |35.0 |false |false |Texas |Undefined |Fulltime |Bachelor's degree |NULL |  
|NULL |3 |NULL |48.0 |false |false |Arizona |Undefined |Fulltime |No Education Listed|NULL |  
|92500 |NULL |NULL |15.0 |false |false |California|Undefined |Flexible |No Education Listed|NULL |  
+------+--------------------+--------------------+--------+-------------------+-------------+----------+----------------+--------------------+-------------------+------------------+  
only showing top 5 rows

# Drop rows with NA values in relevant columns  
df\_feature\_engg = df\_eda.dropna(subset=[  
 "SALARY", "MIN\_YEARS\_EXPERIENCE", "MAX\_YEARS\_EXPERIENCE","STATE\_NAME",  
 "EMPLOYMENT\_TYPE\_NAME", "REMOTE\_TYPE\_NAME","MIN\_EDULEVELS\_NAME",  
 "DURATION", "IS\_INTERNSHIP", "COMPANY\_IS\_STAFFING"  
])  
  
# Categorical columns  
categorical\_cols = ["STATE\_NAME","MIN\_EDULEVELS\_NAME","EMPLOYMENT\_TYPE\_NAME", "REMOTE\_TYPE\_NAME"]  
  
# Index and One-Hot Encode  
indexers = [StringIndexer(inputCol=col, outputCol=f"{col}\_idx", handleInvalid='skip') for col in categorical\_cols]  
encoders = [OneHotEncoder(inputCol=f"{col}\_idx", outputCol=f"{col}\_vec") for col in categorical\_cols]  
  
pipeline = Pipeline(stages=indexers)  
indexed\_df = pipeline.fit(df\_feature\_engg).transform(df\_feature\_engg)  
indexed\_df.select("EMPLOYMENT\_TYPE\_NAME","EMPLOYMENT\_TYPE\_NAME\_idx","REMOTE\_TYPE\_NAME","REMOTE\_TYPE\_NAME\_idx").show()  
  
pipeline = Pipeline(stages=indexers + encoders)  
encoded\_df = pipeline.fit(df\_feature\_engg).transform(df\_feature\_engg)  
encoded\_df.show()

[Stage 194:> (0 + 1) / 1] [Stage 200:> (0 + 1) / 1] [Stage 206:> (0 + 1) / 1] [Stage 212:> (0 + 1) / 1]

+--------------------+------------------------+----------------+--------------------+  
|EMPLOYMENT\_TYPE\_NAME|EMPLOYMENT\_TYPE\_NAME\_idx|REMOTE\_TYPE\_NAME|REMOTE\_TYPE\_NAME\_idx|  
+--------------------+------------------------+----------------+--------------------+  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Remote| 1.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Remote| 1.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Remote| 1.0|  
| Fulltime| 0.0| Remote| 1.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| On-Premise| 3.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Remote| 1.0|  
| Flexible| 2.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
| Fulltime| 0.0| Undefined| 0.0|  
+--------------------+------------------------+----------------+--------------------+  
only showing top 20 rows

[Stage 219:> (0 + 1) / 1] [Stage 225:> (0 + 1) / 1] [Stage 231:> (0 + 1) / 1] [Stage 237:> (0 + 1) / 1]

+------+--------------------+--------------------+--------+-------------------+-------------+--------------+----------------+--------------------+------------------+--------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+  
|SALARY|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|DURATION|COMPANY\_IS\_STAFFING|IS\_INTERNSHIP| STATE\_NAME|REMOTE\_TYPE\_NAME|EMPLOYMENT\_TYPE\_NAME|MIN\_EDULEVELS\_NAME| MAX\_EDULEVELS\_NAME|STATE\_NAME\_idx|MIN\_EDULEVELS\_NAME\_idx|EMPLOYMENT\_TYPE\_NAME\_idx|REMOTE\_TYPE\_NAME\_idx| STATE\_NAME\_vec|MIN\_EDULEVELS\_NAME\_vec|EMPLOYMENT\_TYPE\_NAME\_vec|REMOTE\_TYPE\_NAME\_vec|  
+------+--------------------+--------------------+--------+-------------------+-------------+--------------+----------------+--------------------+------------------+--------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+  
| 92962| 2| 2| 18.0| false| false| New York| Undefined| Fulltime| Bachelor's degree| Master's degree| 1.0| 0.0| 0.0| 0.0| (50,[1],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
| 75026| 2| 2| 18.0| true| false| Mississippi| Undefined| Fulltime| Bachelor's degree| NULL| 21.0| 0.0| 0.0| 0.0|(50,[21],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
| 60923| 1| 1| 18.0| false| false| New York| Remote| Fulltime| Bachelor's degree| NULL| 1.0| 0.0| 0.0| 1.0| (50,[1],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[1],[1.0])|  
|131100| 2| 2| 11.0| false| false| Arizona| Undefined| Fulltime| Bachelor's degree| NULL| 15.0| 0.0| 0.0| 0.0|(50,[15],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|136950| 3| 3| 18.0| false| false| Maine| Remote| Fulltime| Bachelor's degree|Ph.D. or professi...| 40.0| 0.0| 0.0| 1.0|(50,[40],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[1],[1.0])|  
|122500| 5| 5| 18.0| false| false| California| Undefined| Fulltime| Bachelor's degree| NULL| 0.0| 0.0| 0.0| 0.0| (50,[0],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|136950| 3| 3| 28.0| false| false| Minnesota| Remote| Fulltime| Bachelor's degree|Ph.D. or professi...| 26.0| 0.0| 0.0| 1.0|(50,[26],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[1],[1.0])|  
|136950| 3| 3| 28.0| false| false| Georgia| Remote| Fulltime| Bachelor's degree|Ph.D. or professi...| 13.0| 0.0| 0.0| 1.0|(50,[13],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[1],[1.0])|  
| 55120| 2| 2| 18.0| false| false| Virginia| Undefined| Fulltime| Bachelor's degree| NULL| 5.0| 0.0| 0.0| 0.0| (50,[5],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|104000| 3| 3| 8.0| false| false| Texas| Undefined| Fulltime| Bachelor's degree| NULL| 2.0| 0.0| 0.0| 0.0| (50,[2],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|145319| 4| 4| 18.0| false| false| Ohio| Undefined| Fulltime| Bachelor's degree| Master's degree| 10.0| 0.0| 0.0| 0.0|(50,[10],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
| 80000| 3| 3| 37.0| false| false| Oklahoma| Undefined| Fulltime| Bachelor's degree| NULL| 43.0| 0.0| 0.0| 0.0|(50,[43],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|102500| 3| 3| 28.0| false| false|South Carolina| On-Premise| Fulltime| Bachelor's degree| NULL| 27.0| 0.0| 0.0| 3.0|(50,[27],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[],[])|  
| 86117| 2| 2| 14.0| false| false|North Carolina| Undefined| Fulltime| Bachelor's degree| NULL| 12.0| 0.0| 0.0| 0.0|(50,[12],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
| 72800| 3| 3| 15.0| true| false| Missouri| Undefined| Fulltime| Associate degree| NULL| 17.0| 3.0| 0.0| 0.0|(50,[17],[1.0])| (5,[3],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|162300| 6| 6| 18.0| false| false| New York| Undefined| Fulltime| Bachelor's degree| Master's degree| 1.0| 0.0| 0.0| 0.0| (50,[1],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|121500| 3| 3| 18.0| false| false| California| Remote| Fulltime| Bachelor's degree| Master's degree| 0.0| 0.0| 0.0| 1.0| (50,[0],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[1],[1.0])|  
|117500| 3| 3| 14.0| false| false| Arkansas| Undefined| Flexible| Bachelor's degree| Master's degree| 20.0| 0.0| 2.0| 0.0|(50,[20],[1.0])| (5,[0],[1.0])| (2,[],[])| (3,[0],[1.0])|  
|102000| 3| 3| 25.0| false| false| Michigan| Undefined| Fulltime| Bachelor's degree| NULL| 14.0| 0.0| 0.0| 0.0|(50,[14],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
|142300| 6| 6| 18.0| false| false| Florida| Undefined| Fulltime| Bachelor's degree| Master's degree| 3.0| 0.0| 0.0| 0.0| (50,[3],[1.0])| (5,[0],[1.0])| (2,[0],[1.0])| (3,[0],[1.0])|  
+------+--------------------+--------------------+--------+-------------------+-------------+--------------+----------------+--------------------+------------------+--------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+  
only showing top 20 rows

# Assemble base features (for GLR and Random Forest)  
assembler = VectorAssembler(  
 inputCols=[  
 "MIN\_YEARS\_EXPERIENCE", "DURATION",  
 "IS\_INTERNSHIP", "COMPANY\_IS\_STAFFING"  
 ] + [f"{col}\_vec" for col in categorical\_cols],  
 outputCol="features"  
)  
  
# Build pipeline and transform  
pipeline = Pipeline(stages=indexers + encoders + [assembler])  
data = pipeline.fit(df\_feature\_engg).transform(df\_feature\_engg)  
  
data.show(5, truncate=False)  
  
# Create squared term for Polynomial Regression  
data = data.withColumn("MIN\_YEARS\_EXPERIENCE\_SQ", pow(col("MIN\_YEARS\_EXPERIENCE"), 2))  
  
# Assemble polynomial features  
assembler\_poly = VectorAssembler(  
 inputCols=[  
 "MIN\_YEARS\_EXPERIENCE", "MIN\_YEARS\_EXPERIENCE\_SQ",  
 "DURATION", "IS\_INTERNSHIP", "COMPANY\_IS\_STAFFING"  
 ] + [f"{col}\_vec" for col in categorical\_cols],  
 outputCol="features\_poly"  
   
)  
  
data=assembler\_poly.transform(data)  
  
#show final structure  
data.select("SALARY", "features", "features\_poly").show(5, truncate=False)

[Stage 244:> (0 + 1) / 1] [Stage 250:> (0 + 1) / 1] [Stage 256:> (0 + 1) / 1] [Stage 262:> (0 + 1) / 1]

+------+--------------------+--------------------+--------+-------------------+-------------+-----------+----------------+--------------------+------------------+----------------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+-------------------------------------------------------+  
|SALARY|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|DURATION|COMPANY\_IS\_STAFFING|IS\_INTERNSHIP|STATE\_NAME |REMOTE\_TYPE\_NAME|EMPLOYMENT\_TYPE\_NAME|MIN\_EDULEVELS\_NAME|MAX\_EDULEVELS\_NAME |STATE\_NAME\_idx|MIN\_EDULEVELS\_NAME\_idx|EMPLOYMENT\_TYPE\_NAME\_idx|REMOTE\_TYPE\_NAME\_idx|STATE\_NAME\_vec |MIN\_EDULEVELS\_NAME\_vec|EMPLOYMENT\_TYPE\_NAME\_vec|REMOTE\_TYPE\_NAME\_vec|features |  
+------+--------------------+--------------------+--------+-------------------+-------------+-----------+----------------+--------------------+------------------+----------------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+-------------------------------------------------------+  
|92962 |2 |2 |18.0 |false |false |New York |Undefined |Fulltime |Bachelor's degree |Master's degree |1.0 |0.0 |0.0 |0.0 |(50,[1],[1.0]) |(5,[0],[1.0]) |(2,[0],[1.0]) |(3,[0],[1.0]) |(64,[0,1,5,54,59,61],[2.0,18.0,1.0,1.0,1.0,1.0]) |  
|75026 |2 |2 |18.0 |true |false |Mississippi|Undefined |Fulltime |Bachelor's degree |NULL |21.0 |0.0 |0.0 |0.0 |(50,[21],[1.0])|(5,[0],[1.0]) |(2,[0],[1.0]) |(3,[0],[1.0]) |(64,[0,1,3,25,54,59,61],[2.0,18.0,1.0,1.0,1.0,1.0,1.0])|  
|60923 |1 |1 |18.0 |false |false |New York |Remote |Fulltime |Bachelor's degree |NULL |1.0 |0.0 |0.0 |1.0 |(50,[1],[1.0]) |(5,[0],[1.0]) |(2,[0],[1.0]) |(3,[1],[1.0]) |(64,[0,1,5,54,59,62],[1.0,18.0,1.0,1.0,1.0,1.0]) |  
|131100|2 |2 |11.0 |false |false |Arizona |Undefined |Fulltime |Bachelor's degree |NULL |15.0 |0.0 |0.0 |0.0 |(50,[15],[1.0])|(5,[0],[1.0]) |(2,[0],[1.0]) |(3,[0],[1.0]) |(64,[0,1,19,54,59,61],[2.0,11.0,1.0,1.0,1.0,1.0]) |  
|136950|3 |3 |18.0 |false |false |Maine |Remote |Fulltime |Bachelor's degree |Ph.D. or professional degree|40.0 |0.0 |0.0 |1.0 |(50,[40],[1.0])|(5,[0],[1.0]) |(2,[0],[1.0]) |(3,[1],[1.0]) |(64,[0,1,44,54,59,62],[3.0,18.0,1.0,1.0,1.0,1.0]) |  
+------+--------------------+--------------------+--------+-------------------+-------------+-----------+----------------+--------------------+------------------+----------------------------+--------------+----------------------+------------------------+--------------------+---------------+----------------------+------------------------+--------------------+-------------------------------------------------------+  
only showing top 5 rows  
+------+-------------------------------------------------------+-------------------------------------------------------------+  
|SALARY|features |features\_poly |  
+------+-------------------------------------------------------+-------------------------------------------------------------+  
|92962 |(64,[0,1,5,54,59,61],[2.0,18.0,1.0,1.0,1.0,1.0]) |(65,[0,1,2,6,55,60,62],[2.0,4.0,18.0,1.0,1.0,1.0,1.0]) |  
|75026 |(64,[0,1,3,25,54,59,61],[2.0,18.0,1.0,1.0,1.0,1.0,1.0])|(65,[0,1,2,4,26,55,60,62],[2.0,4.0,18.0,1.0,1.0,1.0,1.0,1.0])|  
|60923 |(64,[0,1,5,54,59,62],[1.0,18.0,1.0,1.0,1.0,1.0]) |(65,[0,1,2,6,55,60,63],[1.0,1.0,18.0,1.0,1.0,1.0,1.0]) |  
|131100|(64,[0,1,19,54,59,61],[2.0,11.0,1.0,1.0,1.0,1.0]) |(65,[0,1,2,20,55,60,62],[2.0,4.0,11.0,1.0,1.0,1.0,1.0]) |  
|136950|(64,[0,1,44,54,59,62],[3.0,18.0,1.0,1.0,1.0,1.0]) |(65,[0,1,2,45,55,60,63],[3.0,9.0,18.0,1.0,1.0,1.0,1.0]) |  
+------+-------------------------------------------------------+-------------------------------------------------------------+  
only showing top 5 rows

[Stage 270:> (0 + 1) / 1]

(3756, 22)

[Stage 273:> (0 + 1) / 1]

(3060, 22)

[Stage 276:> (0 + 1) / 1]

(696, 22)

from pyspark.ml.regression import LinearRegression  
  
# Initialize Regression model  
# Basic Linear Regression model  
mr = LinearRegression(featuresCol="features", labelCol="SALARY")  
  
# Polynomial Regression using squared term features  
#mr = LinearRegression(featuresCol="features\_poly", labelCol="SALARY")  
  
# Generalized Linear Regression (supports distributions like Gaussian, Poisson)  
 # mr = GeneralizedLinearRegression(  
 # featuresCol="features", labelCol="SALARY",  
 # family="gaussian", link="identity"  
 # )  
  
  
# Train the model  
mr\_model = mr.fit(regression\_train)  
  
# Evaluate on test data  
test\_results = mr\_model.evaluate(regression\_test)  
  
# Print metrics  
print("RMSE:", test\_results.rootMeanSquaredError)  
print("R2:", test\_results.r2)

25/10/07 07:35:44 WARN Instrumentation: [0efb8138] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 279:> (0 + 1) / 1] [Stage 280:> (0 + 1) / 1] [Stage 281:> (0 + 1) / 1]

RMSE: 29114.307979774672  
R2: 0.34792446541830324

coeffs = mr\_model.coefficients  
intercept = mr\_model.intercept  
  
print("Intercept:", intercept)  
print("Coefficients:", coeffs)

Intercept: 131328.3004796009  
Coefficients: [8523.062910159928,-98.93546907813675,2782.413073078266,-1078.1337256635686,12347.650329505026,12099.991428639481,5508.201281626465,4103.394476179377,6827.789398867635,1859.9349418510367,7751.353854768509,9315.672167780182,-755.6703653785958,9973.18649877864,1314.7076955620487,-1299.0660250961937,2093.9361992914087,7193.052647920884,-5124.810223960465,4945.904023459982,-293.75547075283794,434.7029433494427,4697.278843397262,7021.701277240336,7659.209522281594,-11920.218306321509,-2098.935017913266,28.958169601522584,10422.439688057706,-1534.9827239260164,3298.1875476982027,2561.1538943029723,8352.834248147985,-3844.4745256701326,-1105.821356233869,1840.7175686222201,487.1782311291216,5803.378236307139,139.85835159896615,2939.1994362592077,-3041.9102995939397,10591.372903022557,-577.2862695106961,-3619.269866311597,5432.8888205860085,311.95208019053894,3281.416667399771,-520.5179852949578,3209.602981992066,-3999.8839871360237,-3502.2852404853415,8178.297251076545,1268.381407942167,710.5884205836893,-54961.00350028223,-49607.85277992999,-84497.42795385023,-79343.56469541852,-40206.932513201114,-17034.987287641532,-20945.521186969952,13599.615064274507,19866.842315893853,8671.816402025408]