Assignment 04

Regression Analysis in Pyspark

Tracy Anyasi

October 8, 2025

## 0.1 Generative AI Disclaimer

Generative AI (ChatGPT) was used to better understand the several dataframes created, their pipeline architecture, and debugging issues.

## 0.2 Github Repository Link

https://github.com/met-ad-688/assignment-04-tanyasiii

WARNING: Using incubator modules: jdk.incubator.vector  
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/10/05 23:28:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

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

25/10/05 23:28:56 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| 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

## 0.3 Feature Engineering

Remove incomplete data, keep relevant variables, and iron out complicated string values Encoder turns categorical columns (remote, hybrid, onsite) to numeric ones (1 or 0) based on input

from pyspark.sql.functions import col, pow, when  
from pyspark.ml.feature import StringIndexer, VectorAssembler, OneHotEncoder  
from pyspark.ml import Pipeline  
  
#remove rows with NAs  
df\_cleaned = df.dropna(subset=[  
 "SALARY", "MIN\_YEARS\_EXPERIENCE", "STATE\_NAME", "EMPLOYMENT\_TYPE\_NAME",  
 "REMOTE\_TYPE\_NAME", "MIN\_EDULEVELS\_NAME", "DURATION",   
 "IS\_INTERNSHIP", "COMPANY\_IS\_STAFFING"  
])  
  
eda\_cols = [  
 "SALARY", "MIN\_YEARS\_EXPERIENCE", "DURATION", "COMPANY\_IS\_STAFFING",  
 "IS\_INTERNSHIP", "STATE\_NAME", "REMOTE\_TYPE\_NAME",  
 "EMPLOYMENT\_TYPE\_NAME", "MIN\_EDULEVELS\_NAME"  
]  
df\_cleaned = df\_cleaned.select(eda\_cols)  
  
#clean up REMOTE\_TYPE\_NAME and reduce the different inputs  
df\_cleaned = df\_cleaned.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"))  
)  
  
#clean EMPLOYMENT\_TYPE\_NAME  
df\_cleaned = df\_cleaned.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\_cleaned = df\_cleaned.filter(col("REMOTE\_TYPE\_NAME") != "Undefined") -- initially wnated to remove the undefined but it dropped the  
#percentage of the regression model  
  
# Categorical and numeric columns  
categorical\_cols = ["EMPLOYMENT\_TYPE\_NAME", "REMOTE\_TYPE\_NAME"]  
continuous\_cols = ["MIN\_YEARS\_EXPERIENCE", "DURATION", "IS\_INTERNSHIP", "COMPANY\_IS\_STAFFING"]  
  
# 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]

+------+--------------------------------------+-------------------------------------------+  
|SALARY|features |features\_poly |  
+------+--------------------------------------+-------------------------------------------+  
|192800|(9,[0,1,4,6],[6.0,55.0,1.0,1.0]) |(10,[0,1,4,6,9],[6.0,55.0,1.0,1.0,36.0]) |  
|125900|(9,[0,1,4,6],[12.0,18.0,1.0,1.0]) |(10,[0,1,4,6,9],[12.0,18.0,1.0,1.0,144.0]) |  
|118560|[5.0,20.0,0.0,1.0,1.0,0.0,0.0,1.0,0.0]|[5.0,20.0,0.0,1.0,1.0,0.0,0.0,1.0,0.0,25.0]|  
|192800|(9,[0,1,4,6],[6.0,55.0,1.0,1.0]) |(10,[0,1,4,6,9],[6.0,55.0,1.0,1.0,36.0]) |  
|116500|(9,[0,1,4,6],[12.0,16.0,1.0,1.0]) |(10,[0,1,4,6,9],[12.0,16.0,1.0,1.0,144.0]) |  
+------+--------------------------------------+-------------------------------------------+  
only showing top 5 rows

Mapping for EMPLOYMENT\_TYPE\_NAME:  
 Index 0 -> Fulltime  
 Index 1 -> Parttime  
 Index 2 -> Flexible

[Stage 22:> (0 + 1) / 1]

Mapping for REMOTE\_TYPE\_NAME:  
 Index 0 -> Undefined  
 Index 1 -> Remote  
 Index 2 -> Hybrid  
 Index 3 -> On Premise

## 0.4 Linear Regression

25/10/05 23:29:25 WARN Instrumentation: [68f324e3] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 32:> (0 + 1) / 1]

R² Score: 0.2840  
RMSE: 35315.94  
MAE: 27676.61

Coefficient Summary:  
 Feature Estimate Std Error t-Stat p-Value \  
0 Intercept 76735.577948 102.356000 66.277489 0.000000e+00   
1 MIN\_YEARS\_EXPERIENCE 6783.898664 23.632630 -1.831651 6.702934e-02   
2 DURATION -43.286721 6866.446762 -1.025459 3.051684e-01   
3 IS\_INTERNSHIP -7041.257613 1063.265929 -0.595172 5.517402e-01   
4 COMPANY\_IS\_STAFFING -632.826152 3011.540238 -0.421858 6.731367e-01   
5 EMPLOYMENT\_TYPE\_NAME\_A -1270.441524 3605.833255 -1.077711 2.811853e-01   
6 EMPLOYMENT\_TYPE\_NAME\_B -3886.046755 2462.498142 3.306373 9.480161e-04   
7 REMOTE\_TYPE\_NAME\_X 8141.937756 2529.947825 3.635882 2.782402e-04   
8 REMOTE\_TYPE\_NAME\_Y 9198.592476 3172.252546 7.636575 2.398082e-14   
9 REMOTE\_TYPE\_NAME\_Z 24225.144004 3579.670705 21.436491 0.000000e+00   
  
 95% CI Lower 95% CI Upper   
0 76534.942765 76936.213130   
1 6737.574686 6830.222643   
2 -13502.691172 13416.117730   
3 -9125.439821 -4957.075405   
4 -6535.957638 5270.305334   
5 -8338.488484 5797.605435   
6 -8712.962276 940.868767   
7 3182.809376 13101.066136   
8 2980.437509 15416.747443   
9 17208.380095 31241.907914

The linear regression model explains approximately 28% of the variance in salaries, showing that while job attributes like as experience and remote status influence pay, substantial variation remains unexplained. This is likely due to qualitative factors like role seniority, company size, or negotiation effects. Undefined roles were initally excluded but that reduced the model’s reliabilty and was subsequentially added as a baseline for remote roles.

Some statistically significant predictors include remote and hybrid roles (Remote Type X & Y) and Flexible employment type, all of which show clear positive or negative salary impacts. Compared to the baseline groups (Fulltime employment and Undefined remote), Flexible roles pay significantly less than Fulltime roles, whereas Parttime roles do not show a significant difference. For remote types, Remote, Hybrid, and On Premise roles show meaningful salary increases relative to Undefined roles, with On Premise roles exhibiting the largest premium of approximately $24K.

Non-significant coefficients, such as Parttime or certain remote categories, suggest that observed differences may be due to random variation rather than a true effect, while the significant predictors highlight areas where job structure meaningfully affects compensation.

## 0.5 Polynominal Linear Regression

25/10/05 23:29:51 WARN Instrumentation: [48c15026] regParam is zero, which might cause numerical instability and overfitting.

Feature Coefficient Std Error t-value \  
0 Intercept 67932.172201 365.427981 34.124107   
1 MIN\_YEARS\_EXPERIENCE 12469.903484 23.369295 -1.746680   
2 MIN\_YEARS\_EXPERIENCE\_SQ -40.818670 6795.398159 -0.378451   
3 DURATION -2571.726155 1051.916215 -1.110934   
4 IS\_INTERNSHIP -1168.609779 2990.337025 -1.899043   
5 COMPANY\_IS\_STAFFING -5678.777125 3572.973249 -2.128946   
6 EMPLOYMENT\_TYPE\_NAME\_A -7606.667528 2435.260898 3.109557   
7 EMPLOYMENT\_TYPE\_NAME\_B 7572.582266 2501.762396 3.565968   
8 REMOTE\_TYPE\_NAME\_X 8921.205741 3137.964205 7.286182   
9 REMOTE\_TYPE\_NAME\_Y 22863.778828 26.275186 -16.193368   
10 REMOTE\_TYPE\_NAME\_Z -425.483749 3581.211738 18.969047   
  
 p-value 95% CI Lower 95% CI Upper   
0 0.000000e+00 67215.871148 68648.473253   
1 8.071963e-02 12424.095687 12515.711281   
2 7.051025e-01 -13360.955886 13279.318546   
3 2.666199e-01 -4633.661011 -509.791298   
4 5.758388e-02 -7030.179413 4692.959856   
5 3.327990e-02 -12682.412944 1324.858694   
6 1.878259e-03 -12380.193459 -2833.141596   
7 3.639866e-04 2668.702078 12476.462454   
8 3.397282e-13 2770.261704 15072.149779   
9 0.000000e+00 22812.274990 22915.282666   
10 0.000000e+00 -7445.268409 6594.300910

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

Polynomial Regression R²: 0.3016947273606302  
Polynomial Regression RMSE: 34875.92736404874  
Polynomial Regression MAE: 27190.399586682357

The polynomial regression model explains about 30% of the variance in salaries, showing that experience and job attributes influence pay, though much variation still remains unexplained. Significant predictors include remote and hybrid roles and all employment types meaningfully impact salaries.

Non-significant terms with limited contributions, such as the quadratic minimum years of experience term, has a negative coefficient while the linear minimum years of experience is positive, suggesting that althought salary increaes with more experience, the increase slows down as experience grows and plateaus. Overall, the model highlights which job characteristics most strongly affect compensation while capturing some non-linear effects of experience.

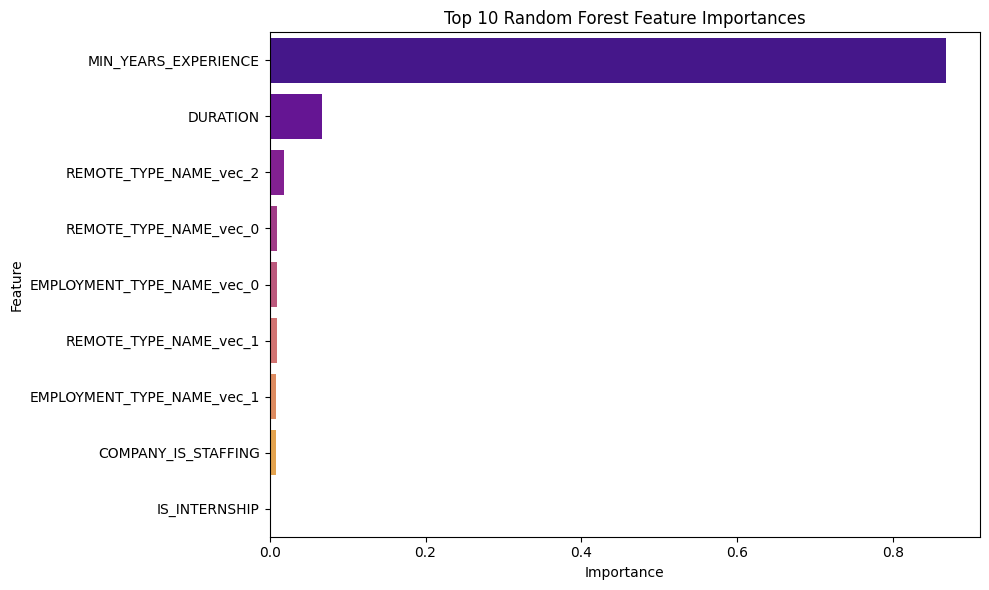
## 0.6 Random Forest Regressor

25/10/05 23:30:40 WARN DAGScheduler: Broadcasting large task binary with size 1210.5 KiB  
25/10/05 23:30:43 WARN DAGScheduler: Broadcasting large task binary with size 2.1 MiB  
25/10/05 23:30:48 WARN DAGScheduler: Broadcasting large task binary with size 3.7 MiB  
25/10/05 23:30:55 WARN DAGScheduler: Broadcasting large task binary with size 6.1 MiB  
25/10/05 23:31:01 WARN DAGScheduler: Broadcasting large task binary with size 1402.3 KiB  
[Stage 56:> (0 + 1) / 1]

+------+------------------+  
|SALARY| prediction|  
+------+------------------+  
| 29120|114624.66663326925|  
| 31200| 96687.65499926287|  
| 31200|114624.66663326925|  
| 31640| 96228.91630692781|  
| 32240| 95844.37906654779|  
+------+------------------+  
only showing top 5 rows  
Feature Importances: [np.float64(0.8681171759565247), np.float64(0.06743212825429395), np.float64(0.0016313701554676735), np.float64(0.007622290249534043), np.float64(0.009574155828520737), np.float64(0.008389185582000717), np.float64(0.009964903639168394), np.float64(0.009171279424142545), np.float64(0.018097510910347166)]

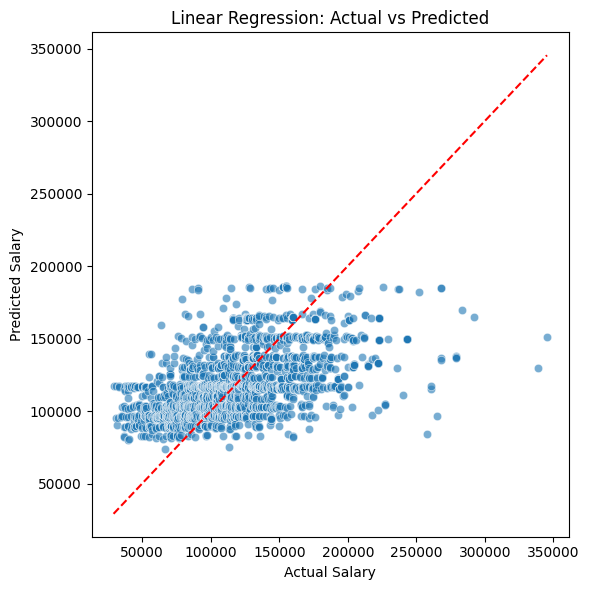
# 1. Feature Importance Plot

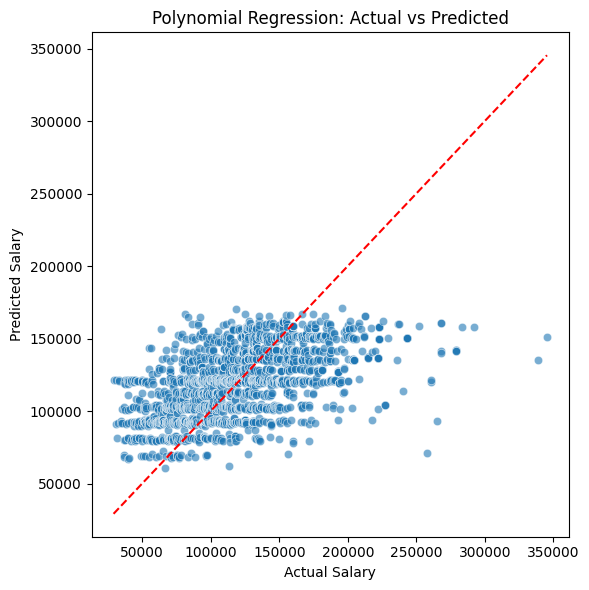
/tmp/ipykernel\_1623/862456364.py:36: FutureWarning:  
  
  
  
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

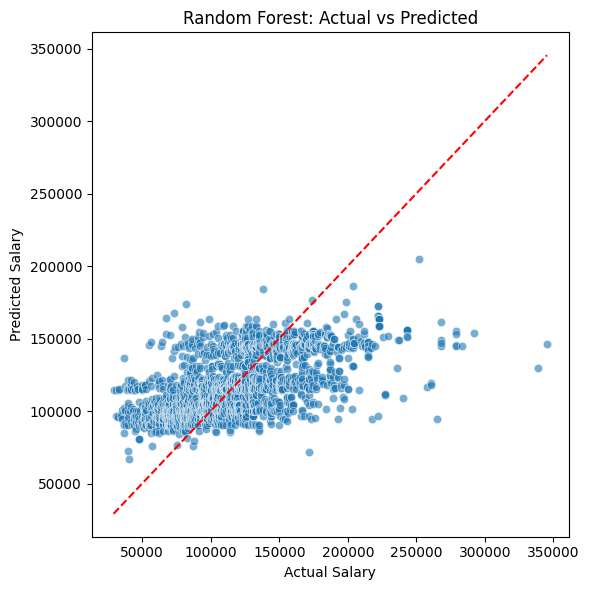


This highlights that minimum years of experience has the most importance in predicating salary in roles compared to other features.

## 1.1 Comparing the 3 Model - Generalized Linear, Polynomial, and Random Forest







RMSE - Linear Regression: 32433.87  
RMSE - Polynomial Regression: 30897.99  
RMSE - Random Forest: 24941.84

AIC - Linear Regression: 60941.54  
AIC - Polynomial Regression: 60878.13

[Stage 145:> (0 + 1) / 1]

BIC - Linear Regression: 84857.93  
BIC - Polynomial Regression: 84584.34  
BIC - Random Forest (approx.): 86841.66

Random Forest provides the most accurate salary predictions, achieving the lowest RMSE (24,942), while polynomial regression improves over linear regression (30,898 vs 32,434) by capturing simple nonlinear effects.

In terms of model fit, polynomial regression shows slightly better AIC and BIC values than linear regression, indicating it balances complexity and explanatory power. Random Forest, despite a higher approximate BIC due to its complexity, outperforms both parametric models in prediction, highlighting its ability to capture complex feature interactions.

Overall, Random Forest is best for predictive performance, whereas polynomial regression offers a reasonable trade-off between fit and interpretability.