National College of Ireland

National College of Ireland

Project Submission Sheet

Student Name:	Sneha Ramesh Dharne		
Student ID:	X23195703		
Programme:	MSC in Data Analytics	**Year:**	2024-25
Module:	Data Intensive Architecture		
Lecturer:	JASWINDER SINGH		
Submission Due Date:	Sunday 19th May, 2024 @ 11:59 pm		
Project Title:	Continuous Assessment (CA) Type: Project		
Word Count:	3216		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Sneha Ramesh Dharne

Date: 19/05/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

Analyzing the Impact of Demographics on Technology Salaries Worldwide

Ms. Sneha Ramesh Dharne
Msc in Data Analytics
National College Of Ireland
Dublin, Ireland
x23195703@student.ncirl.ie

***Abstract*—With all technology realm that continuously assessing society, it becomes crucial that both employers and job seekers are aware and can interpret the complexities that come along with tech labour market. The tech industry is associated with innovative mindset, rapid expansion, good salary, and other numerous opportunities. Through disparities in the wages though, a range of factors can be accounted for, such as the position, experience, education, most importantly demographic characteristics like gender and ethnicity. This research is intended to describe the existing digital gap causes, and assist understanding of the role in which the technology industry plays in creating these gaps. This paper is aimed at demonstrating current inequalities between various salaries. It also intends to increase the awareness and contribute to policy-making in the technology industry, thus making work environment better. It manages to do so by looking at the average income of the given communities by its ethnic structure and technology use.**

***Keywords—* (*key words*)**

I. Introduction

Data-intensive architectures, among others that are big data platforms, cloud services and distributed systems, are paramount for efficient handling of data which means storing, processing and analysing data by giving insights and innovative ideas.

Project Objectives include Census about tech salaries based on race, country, education and experience, to create a correlation matrix of these parameters together and figure out income differences between various demographics and regions.

Project Scope: This project examines the role of racial factors, country of origin, educational attainment, and length of experience in the remuneration of the technology sector. The dataset consists of domain specific data of the technology industry. Problems entail administration of large data quantities, data quality verification, and different formats of data, which may affect processing speed and overall accuracy of analytics. The project provides solutions to the challenges by incorporating the advanced data-intensive structures which enable the deep analytical views.

This project benefits from the use of advanced data processing systems that help to determine the level of salary inequality in the tech industry for factors such as race, gender, nationality, education and experience. Through using the modern investigative techniques on new data sets, it aims at discovering pivotal inequalities in compensation and serves as a valuable helping tool for policymakers and workforce direction in data science. This exploration as well contributes to the disclosure of structural inequalities and the devising of tailored interventions for fairness and inclusivity in tech jobs.[1]

II. DataSET

A. *Dataset Overview*

The datasets, "DataScience_salaries_2024.csv" and "Salary.csv," store accurate and extensive data about salaries in data science positions across countries, races, and education levels. Such diversified datasets constitute an inalienable part of global compensation analysis and identification of particular disparities in tech industry.

B. *Data Selection*

Both datasets provide detailed, complete, and current salary information, demographic details, and employment data. For their relevance, honesty, and completeness, these data sets are invaluable for analysis of technology industry compensation and bias. By their joint efforts they enscence the study by giving different views of all the dimensions of the problem and supplementing the data analysis taking into account different age groups and the location of experiment.

C. *Literature Review*

Previous studies which used the same data sets mostly employed the machine learning algorithms like wage disparity and predictors of salaries in tech. Difficulties such as data scarcity and bias will be to address by the use of advanced preprocessing techniques and sophisticated regression methods to improve accuracy of studies.

III. Methodology

Data Collection and Pre-processing : After intense research, the data was collected from the Kaggle website and was thoroughly cleaned and pre-processed to produce a reliable information for analysis. In applying the techniques, handling missing values, normalizing salary data across different currencies, and standardizing categorical variables like race and education were the major ones. Disparate data sources were merged, with discrepancies reconciled to maintain consistency, leading to the enhancement of the dataset for deep analytical segment.[2]

Sr. No.	Dataset	Description	Link
1	Salary.csv	includes information on the pay of specific individuals across a number of countries, broken down by age, gender, education, job title, years of experience, salary, and race.	Salary_Data_Based_country_and_race
2	DataScience_salaries_2024.csv	The work year, experience level, employment kind, job title, USD pay, remote ratio, employee residence, company location, and company size are all included, with an emphasis on data science and related fields.	data-science-salaries-2024

Analytical approach : The MapReduce programming model efficiently processes vast datasets by distributing tasks across multiple nodes, parallelizing the workload to reduce computation time significantly. In this model, 'Map' tasks handle data filtering and sorting, while 'Reduce' tasks perform aggregations or summaries. The MapReduce model is applied by mapping operations to group data by 'country' and 'race', then reducing it through aggregation to calculate average salaries. The approach utilizes Spark's distributed processing capabilities to efficiently handle large datasets, paralleling tasks across nodes for optimized performance.[3]

System Design : The architecture of the system is constructed around Apache Spark and PySpark which, in turn, has a great advantage of processing large volume of data because of the distributed computing model. Spark's RDDs (Resilient Distributed Datasets') provide data fault tolerance so that computation tasks are robust and scalable according to the amount of computing resources allocated in a cluster of machines. With a memory structure in place, Spark eliminates issues associated with disk I/O, which means that it runs much faster than conventional disk-based processing systems. The architecture uses cloud-based storage and computing to effectively pair it with popular platforms that include Google Cloud, AWS, or Azure to either scale up the resources required that your app needs or downgrade the utilized resources. With its feature of adaptability to varying data sizes and computational minus, this nozzle is suitable for an efficient data processing and analysis with real-time or batch mode inscription, hence, the cost of performance is optimized.[4]

An enterprise-grade data, analytics, and artificial intelligence platform that is unified, open, and shared can be built, deployed, and maintained at scale with Databricks Community Edition. Your cloud account's storage and security are integrated by the Databricks Data Intelligence Platform, which also handles the deployment and management of cloud infrastructure.[5]

For the purpose of comprehending the distinct semantics of data, Databricks leverages the data Lakehouse and generative AI. In order to meet your company demands, it then automatically maintains infrastructure and improves performance. The combination of the business-friendly UIs, inexpensive compute resources, and massive, affordable data storage in this provides Databricks with an incredibly powerful personal analytics platform. IT personnel make standing queries less complicated by scaling out compute clusters of SQL instances and relieve the end users from bothering with any of the problems related to cloud working. SQL users can run queries against the lake house data lake via the notebooks or SQL query editor. Notebooks does support R, Python as well as Scala and, on top of SQL, enable user to insert visualizations available in legacy dashboards together with a range of links, image and markdown notes.[6]

Tool selection : Pyspark as a part of Python infrastructure makes it possible to process distributed data via the interaction with the four-engine clustered data processing solution from Apache Spark. It gives a possibility for Python scripting to conduct complex data transformations and analyses via large datasets by making use of Sparks' RDDs and DataFrame APIs for supporting parallel processing concerning the optimization of the execution plans autonomously.

Research Questions and objectives :

How do demographic factors and geographical location influence salary levels in the global technology sector.

Objectives include:

1. To estimate the effect of race and country on technology salaries.
2. To investigate the link between educational attainment, experience, and salary in the technology business.
3. To detect patterns and trends in compensation gaps among different demographics and areas.

IV. Code and Results

Let's focus on code now. I have written the code using Databricks Community Edition. The study utilizes two datasets namely – 'Salary.csv' and 'DataScience_salaries_2024'. The steps I utilizes are as follows.

Step 1: Settedd-up PySpark Environment

Made sure that PySpark is added in the Python environment.

```
# analyzing the dataset using the MapReduce programming model with PySpark to estimate the effect of race and country on technology salaries
# setting up the PySpark environment
pip install pyspark

```

```
Python interpreter will be restarted.
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
Collecting py4j==0.10.9.7
  Downloading py4j-0.10.9.7-py2.py3-none-any.whl (200 kB)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py): started
  Building wheel for pyspark (setup.py): finished with status 'done'
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488511 sha256=47798af0fccb950efa9fc821361885f665059d8ce6bdafb1987e83acdef83e17
  Stored in directory: /root/.cache/pip/wheels/92/09/11/aa01d01a7f005fda8a66ad71d2be7f8aa341bddafb27eee3c7
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.7 pyspark-3.5.1
Python interpreter will be restarted.
```

Uploaded both datasets on databricks's DBFS, and Loaded the dataset into the Spark DataFrame.

Python

```
data1 = spark.read.format("csv").option("header", "true").load("dbfs:/FileStore/shared_uploads/snehadharne35@gmail.com/Salary.csv")
data2 = spark.read.format("csv").option("header", "true").load("dbfs:/FileStore/shared_uploads/snehadharne35@gmail.com/DataScience_salaries_2024.csv")
```

```
(2) Spark Jobs
data1:  pyspark.sql.dataframe.DataFrame
    Age: string
    Gender: string
    Education Level: string
    Job Title: string
    Years of Experience: string
    Salary: string
    Country: string
    Race: string
    Senior: string
data2:  pyspark.sql.dataframe.DataFrame
    work_year: string
    experience_level: string
    employment_type: string
    job_title: string
    Salary: string
    salary_currency: string
    salary_in_usd: string
    employee_residence: string
    remote_ratio: string
    company_location: string
    company_size: string
```

Initialized PySpark Session and started with pre-processing.

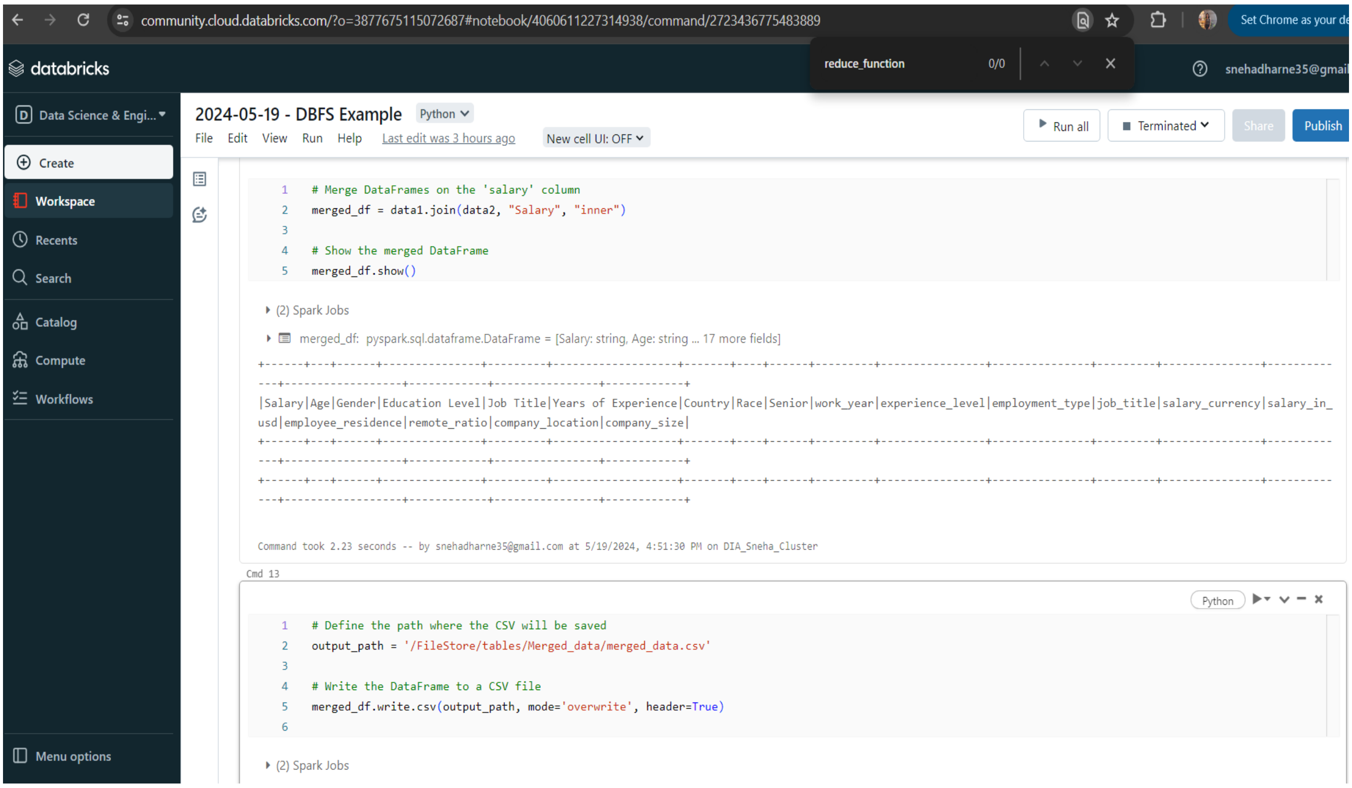
Step 2 : Data Pre-processing

1) After loading the data, displaying few rows of it to make sure that data loaded properly : To display the schema of a DataFrame in a PySpark environment, The printSchema() method is used. It outputs the structure of the DataFrame, including the names of the columns and their data types, along with whether a column can contain null values. This is particularly useful for understanding the organization of data and ensuring that it matches expected formats before proceeding with data processing or analysis.*[7]*

2) Inspecting the data structure within a DataFrame : Retrieved a list of tuples from the DataFrame. Each tuple contains the name of a column and its data type, which is crucial for data manipulation and ensuring that data types are suitable for the operations you plan to perform.*[8]*

3) Summary statistics: The DataFrame's summary statistics for every numerical column contain the count, mean, standard deviation, minimum, and maximum values for every column.

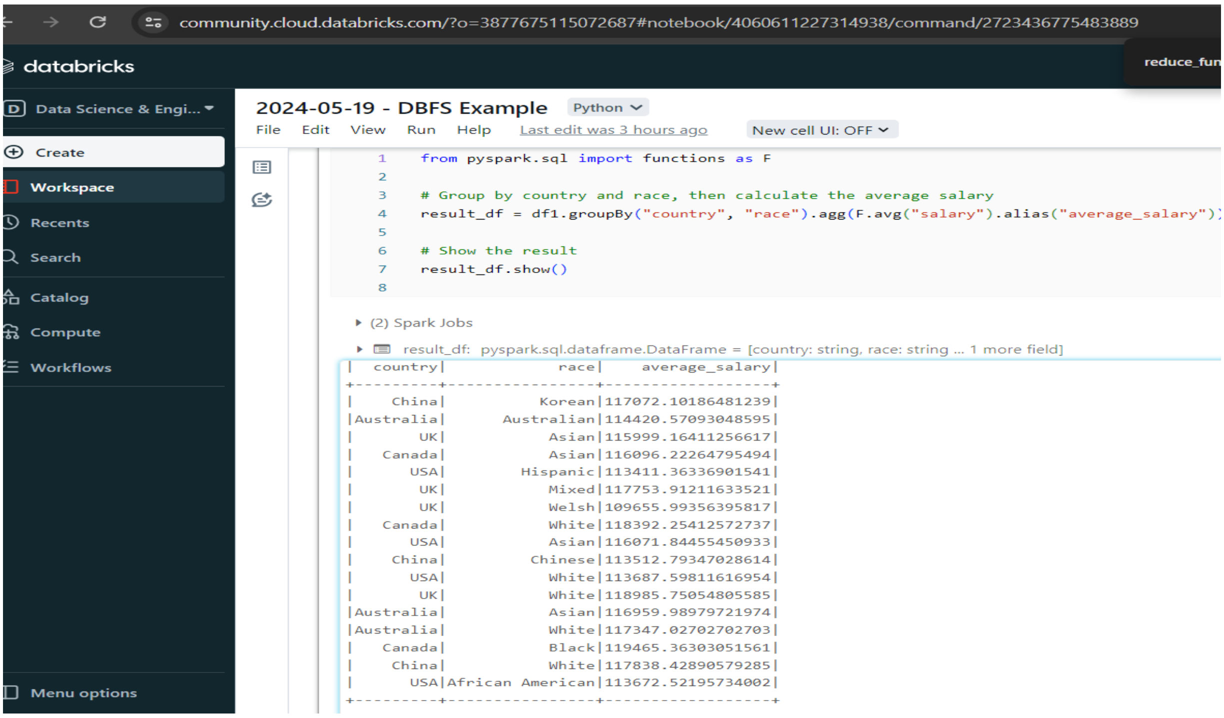
Step 3: Merged datasets : After the data is cleaned, merged datasets on basis of common column named 'salary' as both datasets complement each other. Stored the dataframe as CSV for further processing and closed the pySpark session.



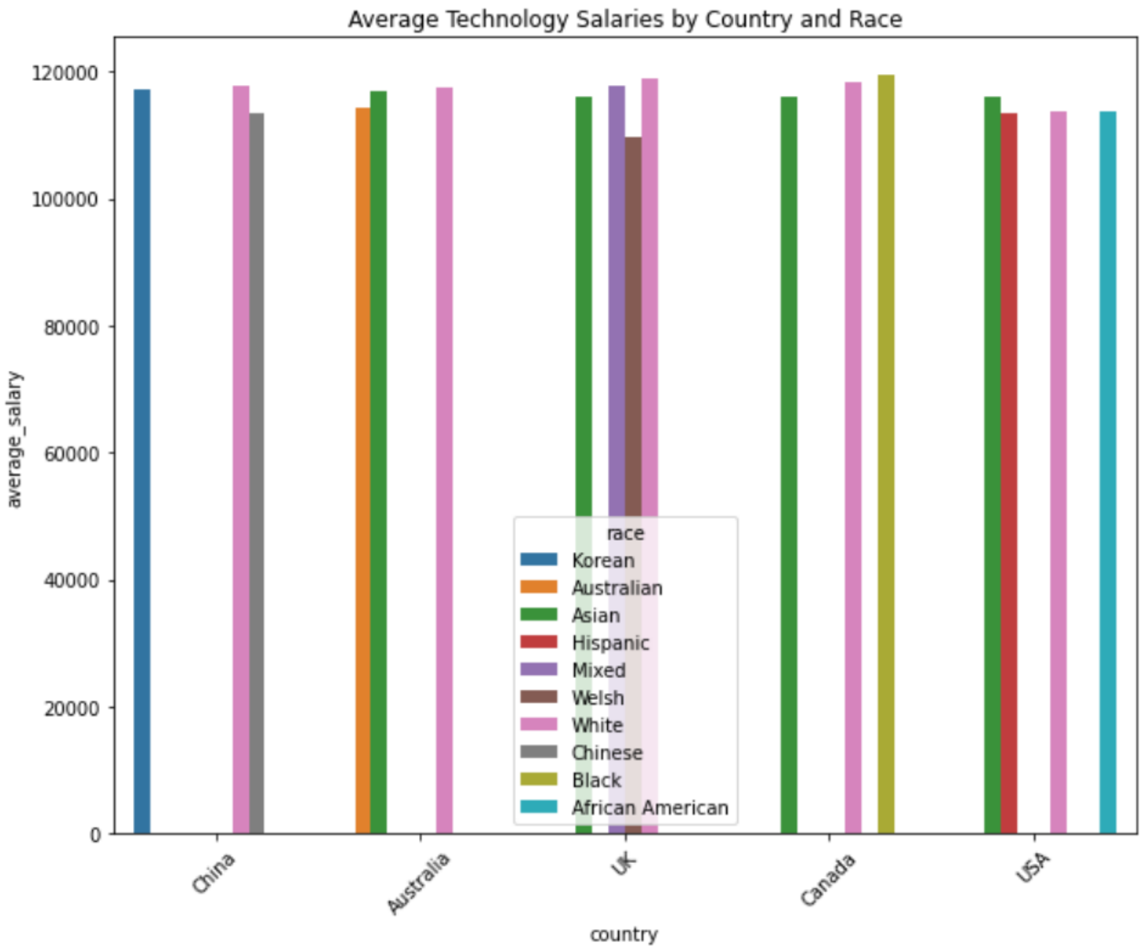
Step 4: Implement MapReduce Operations

A) Created new spark session, loaded the merged data in Dataframe, handled missing values in dataset.

began the investigation with the goal of completing goal 1, which is to calculate the impact of nationality and race on technology salary.



Used Pandas, Matplotlib, and Seaborn to visualize the results. Created a bar plot using Seaborn to visualize how average salaries compare across different countries and races. Closed the session.



The bar chart that visualizes the average salaries across different races and in different countries or regions given the multiple groups of bars. Each color represents a different racial group as indicated in the legend on the right side of the graph.

Each bar color corresponds to a specific racial group, such as Korean, Australian, Asian, Hispanic, etc., as detailed in the legend. This helps differentiate the salary data for each race visually. The bars are grouped, likely representing different countries or possibly different sectors within the technology industry. Each group shows a vertical set of bars where each bar represents the average salary for a particular race in that group.

Key observations :

Racial Salary Gaps: The bars representing 'Asian' are generally taller than those for 'Hispanic', it indicates that Asians, on average, earn more than Hispanics in the tech industry in USA.

Also same thing is observed in all the countries.

B) Again started a new session for implementing a map reduce model for 2nd objective which is investigating the link between educational attainment, experience, and salary in the technology sector

File Edit View Run Help Last edit was 3 hours ago New cell UI: OFF

Cmd 29

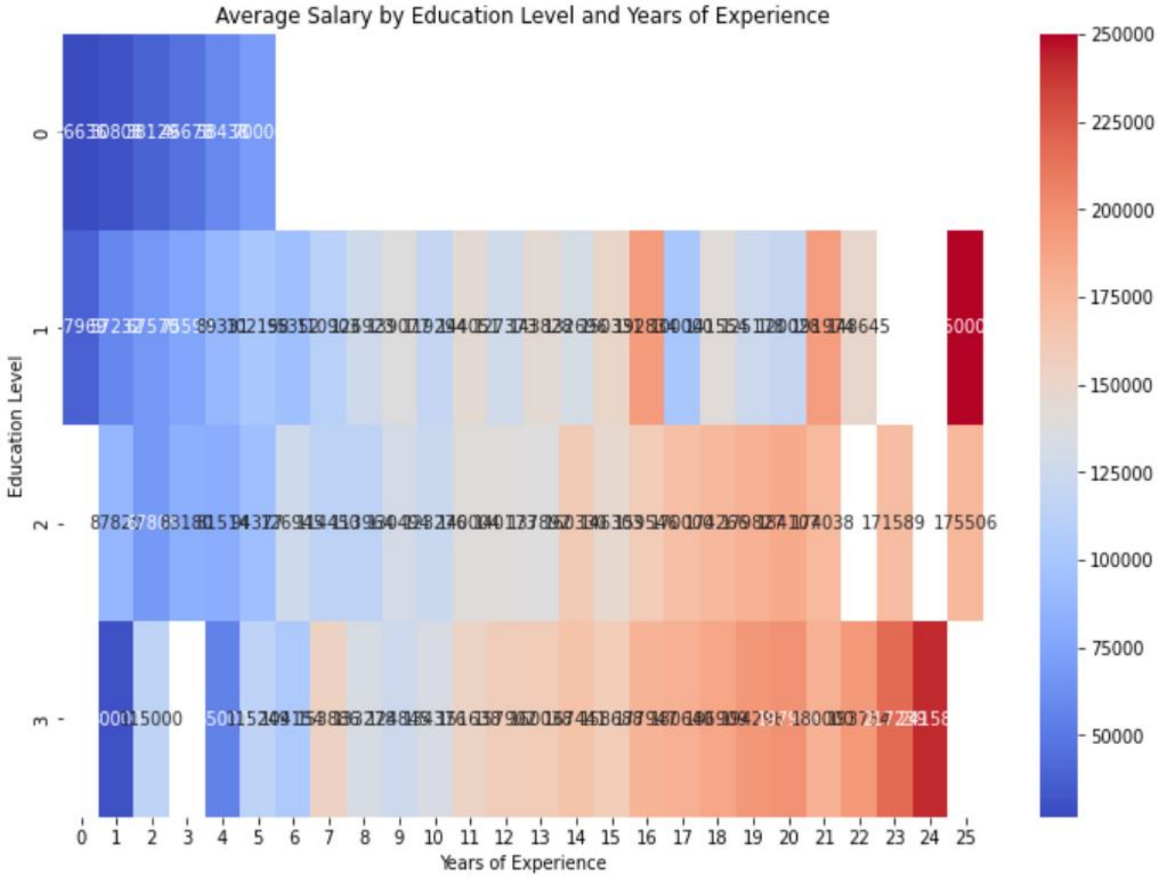
```
# Clean data by dropping rows where essential columns are missing
df2 = df2.dropna(subset=["Salary", "Education Level", "Years of Experience"])

# Check data types and convert if necessary
df2 = df2.withColumn("Salary", df2["Salary"].cast("float"))
df2 = df2.withColumn("Years of Experience", df2["Years of Experience"].cast("integer"))

```

df2: pyspark.sql.dataframe.DataFrame = [_c0: string, work_year: string ... 18 more fields]

Command took 0.26 seconds -- by snehadharne35@gmail.com at 5/19/2024, 3:41:38 PM on DIA_Sneha_Cluster

Cmd 30

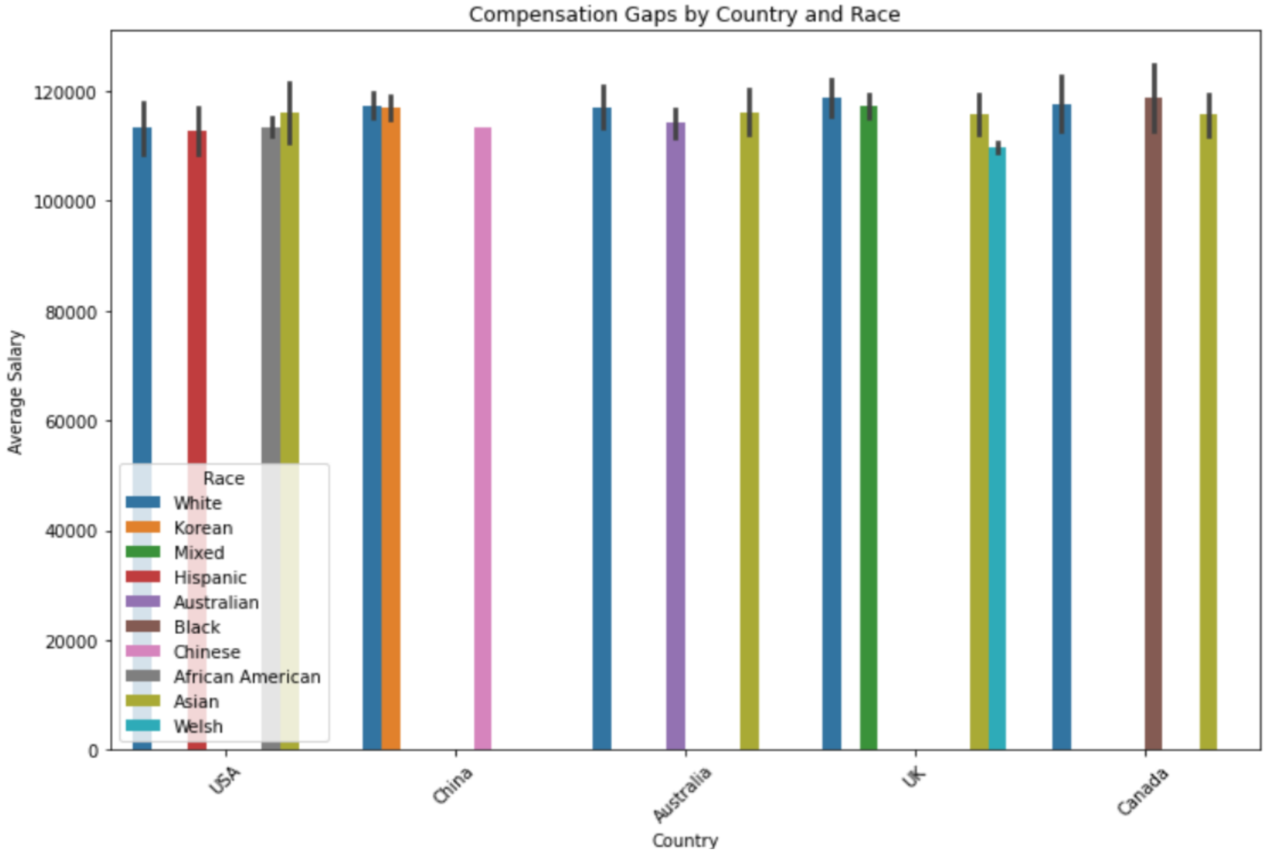
```
# Use PySpark's DataFrame API to explore the relationship
from pyspark.sql import functions as F

# Group by education level and years of experience, then calculate the average salary
analysis_df = df2.groupBy("Education Level", "Years of Experience").agg(
    F.avg("salary").alias("average_salary")
)

# Display the aggregated DataFrame
analysis_df.show()

```



Visualized the result using heatmap, a graphical representation that uses color coding to display the magnitude of a phenomenon as color in two dimensions. The right side of the heatmap includes a color scale ranging from blue to red. Cooler colors (blues) typically represent lower values and warmer colors (reds) represent higher values. The scale is labeled with numerical values, which represent the data points being visualized. The body of the heatmap is composed of cells organized in rows and columns, each cell corresponding to a data point. Each cell's color corresponds to its value relative to the color scale. Each cell contains a numerical label. These labels are the actual data values represented by the colors.

X-Axis: Represents years of experience, ranging from 0 to 25 years.
Y-Axis: Denotes different levels of education, labeled from 0 to 3, although the specific nature of these levels (e.g., high school, bachelor's, master's, PhD) is not specified and would benefit from clarification.
Color Gradient: The color scale transitions from blue to red, where blue represents lower salary ranges and red represents higher salary ranges. This gradient is quantified on the right side of the graph, showing specific salary amounts associated with each color.

C) *Again started new session and analyzed data with MapReduce to find anser of 3rd objective : To detect patterns and trends in compensation gaps among different demographics and areas.*



The bar chart titled "Compensation Gaps by Country and Race," which illustrates the average salary for different racial groups across five countries: USA, China, Australia, UK, and Canada and gaps among the pays.

Salary Distribution: The graph reveals that salary distributions vary significantly both across countries and racial groups.

V. Conclusion and Future work

A. Conclusion and future work for objective 11 : This visualization is crucial for identifying both global and localized racial salary disparities in the technology sector. Such insights are valuable for:

1) Corporate Strategy: Helping companies develop fairer pay practices and promote diversity and inclusion.

2) Policy Making: Informing policymakers where interventions are needed to ensure equity in tech employment.

3) Public Awareness: Raising awareness about racial compensation gaps in the tech industry, potentially influencing public opinion and policy.

4) Overall, the graph serves as a powerful tool to visualize and analyze the impact of race and country on technology salaries, highlighting areas for improvement and action in addressing wage inequities.

B. Conclusion and future work of Objective 2 :

1) Trend Observation: There is a clear trend showing that as years of experience increase, the average salary generally increases as well, regardless of the education level. This trend is especially pronounced after around 15 years of experience, where salary increments become notably higher.

2) Education Level Influence: Higher education levels (2 and 3) show a propensity for higher starting salaries compared to lower education levels (0 and 1). Notably, the highest education level (3) does not always correspond to the highest salaries in the earlier years but does show significant salary increases in the later years.

3) High Experience Impact: The most dramatic salary increases are observed at higher experience levels across all education levels, particularly noticeable in the 20-25 year range.

Implications:
1. Economic and Policy Insights: The graph highlights the long-term value of both professional experience and higher education in salary progression, which can be crucial for informing educational policies and workforce development strategies.
2. Career Planning: These insights are valuable for individuals in career planning, academic advising, and for organizations designing employee growth and compensation structures.

Suggestions for Future Work:

a. Detailed Education Level Analysis: Future analyses should clarify the nature of each education level to better understand the specific impact of different types of educational qualifications on salary.*[9]*
b. Demographic Factors: Including additional demographic factors such as industry, gender, race, and geographical location could provide a more comprehensive view of salary disparities and influences.
c. Longitudinal Study: Conducting a longitudinal study to track changes over time with the same cohorts could provide insights into career progression and the impact of economic cycles on salaries.
d. Predictive Modeling: Developing predictive models to forecast salary trends based on current data could be beneficial for both individuals and policymakers in planning for future economic conditions. By extending the analysis and incorporating these suggestions, future work can provide deeper insights and more actionable recommendations based on the relationship between education, experience, and salaries.*[10]*

C. Conclusion and future work of Objective 3 :

- Racial Disparities: Within each country, there are visible disparities in compensation among different races. For example, in the USA and Canada, the bars' heights vary, suggesting that some racial groups tend to earn more than others.
- Country Comparison: Between countries, the average salaries for similar racial groups also show variation. For instance, salaries in Australia appear generally higher across all racial groups compared to other countries.
- Highest and Lowest Earnings: Specific races like Asians and Whites appear to have consistently higher earnings in countries like Australia and the UK, while other groups such as Hispanic and Mixed might have lower earnings across several countries.
- Error Bars: The inclusion of error bars (the black lines on top of each bar) suggests consideration of variance or confidence intervals, indicating that the data has been rigorously analyzed to account for potential variability within each sample group.

Implications: The graph effectively highlights the existing wage inequalities based on race within the technology sector across different global regions. Such insights are crucial for policymakers, business leaders, and diversity officers aiming to understand and address systemic wage disparities. These data can help drive discussions and actions towards more equitable compensation practices in the technology industry worldwide.

Research Question : How do demographic factors and geographical location influence salary levels in the global technology sector?

- Demographic Factors: There is a clear indication that racial demographics have a profound impact on salary, with some groups consistently earning more than others across multiple regions. This suggests a need for ongoing efforts to address racial wage gaps and promote diversity and inclusion within the technology sector.[11]
- Geographical Influence: Salaries in the technology sector are not uniform globally but vary widely with the economic development, cost of living, and demand for tech skills in different regions. Countries that are known hubs for technology like the USA and Australia tend to offer higher average salaries, reflecting the high demand and concentration of tech companies.
- Education and Experience: Both are crucial determinants of salary in the technology sector. Investment in higher education and gaining extensive experience are shown to be effective in achieving higher earnings. The data suggests that long-term career planning focusing on continuous learning and skill development is beneficial.[12]

Recommendations for Future Studies
Further research could focus on dissecting these influences more granularly by incorporating additional variables such as gender, specific tech roles, company size, and industry verticals. Moreover, longitudinal studies could help understand trends over time, especially in response to global economic changes, new educational norms (like the rise of bootcamps and online courses), and technological advancements.[13]

VI. References

[1] "Big data," *Wikipedia*. May 07, 2024. Accessed: May 19, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_data&oldid=1222692595

[2] "Business Intelligence, Analytics, and Data Science A Managerial Perspective." Accessed: May 19, 2024. [Online]. Available: https://studylib.net/doc/26169841/business-intelligence--analytics--and-data-science-a-mana...

[3] "MapReduce Programming Model | Dremio." Accessed: May 19, 2024. [Online]. Available: https://www.dremio.com/wiki/mapreduce-programming-model/

[4] S. E. Bibri and J. Krogstie, "The core enabling technologies of big data analytics and context-aware computing for smart sustainable cities: a review and synthesis," *J. Big Data*, vol. 4, no. 1, p. 38, Nov. 2017, doi: 10.1186/s40537-017-0091-6.

[5] "What is Databricks?" Accessed: May 19, 2024. [Online]. Available: https://docs.databricks.com

[6] "Lakehouse AI: A Data-Centric Approach to Building Generative AI Applications | Databricks Blog." Accessed: May 19, 2024. [Online]. Available: https://www.databricks.com/blog/lakehouse-ai

[7] "PySpark printSchema() Example - Spark By {Examples}." Accessed: May 19, 2024. [Online]. Available: https://sparkbyexamples.com/pyspark/pyspark-printschema-example/

[8] "Creating a Pandas dataframe using list of tuples - GeeksforGeeks." Accessed: May 19, 2024. [Online]. Available: https://www.geeksforgeeks.org/creating-a-pandas-dataframe-using-list-of-tuples/

[9] W. Mou, "A Quantitative Analysis of the Relationship between Education Level and Income," *J. Educ. Humanit. Soc. Sci.*, vol. 12, pp. 160–166, Apr. 2023, doi: 10.54097/ehss.v12i.7617.

[10] M. Dawood, "Predictive Analytics: Forecasting Future Trends," Medium. Accessed: May 19, 2024. [Online]. Available: https://muhammaddawoodaslam.medium.com/predictive-analytics-forecasting-future-trends-4147b3895986

[11] "Wage Gaps by Race: Its History, Importance, and Impact," Investopedia. Accessed: May 19, 2024. [Online]. Available: https://www.investopedia.com/wage-gaps-by-race-5073258

[12] N. Lee and S. Clarke, "Do low-skilled workers gain from high-tech employment growth? High-technology multipliers, employment and wages in Britain," *Res. Policy*, vol. 48, no. 9, p. 103803, Nov. 2019, doi: 10.1016/j.respol.2019.05.012.

[13] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli, "Longitudinal studies," *J. Thorac. Dis.*, vol. 7, no. 11, pp. E537–E540, Nov. 2015, doi: 10.3978/j.issn.2072-1439.2015.10.63.