The Impact of Demographics on Employment and Salary

A Hypothesis Testing and Clustering Approach

Ciaran Otter
Department of Computer science
Principles of Data Science - CS771
Stellenbosch University, Stellenbosch, South Africa

I. Introduction and Project Goal

This project investigates the relationship between demographic factors such as highest education level, age, gender, and ethnicity with employment status and salary in South Africa. Using real-world 2023 household survey data found through stats SA's interactive data portal: 'Nesstar' [1], his investigation aims to conduct statistical hypothesis testing to ascertain whether these factors significantly influence employment outcomes and salary levels in the contemporary South African context. Additionally, clustering techniques will be employed to uncover underlying patterns within the dataset.

The dataset comprises 112 variables, providing a comprehensive overview of household statistics within the South African population. However, for this analysis, we will focus on several key variables that are particularly relevant to our research objectives:

- **Education level**
- **Employment status**
- **Salary**
- **Age**
- **Gender**
- **Ethnicity**

A. Hypotheses

To guide our investigation, we have formulated the following hypotheses regarding the influence of demographic factors on employment status and salary:

B. Hypothesis 1: Education and Employment Status

Research Question: Does an individual's highest level of education influence their employment status?

- **Null Hypothesis** (H_0)**:** There is no significant relationship between education level and the likelihood of employment.
- **Alternative Hypothesis** (H_1)**:** There is a statistically significant relationship between education level and the likelihood of employment.

C. Hypothesis 2: Education and Salary

Research Question: Does an individual's level of education have a significant impact on their salary?

- **Null Hypothesis** (H_0)**:** Education level does not significantly influence salary.
- **Alternative Hypothesis** (H_1)**:** Education level has a statistically significant effect on salary.

D. Hypothesis 3: Age and Salary

Research Question: Does age affect the salary of an employed individual?

- **Null Hypothesis** (H_0)**:** Age has no significant effect on salary compensation.
- **Alternative Hypothesis** (H_1)**:** Age significantly influences salary compensation.

E. Hypothesis 4: Gender and Salary/Education Level

Research Question: Does gender significantly affect salary and education level?

Salary:

- **Null Hypothesis** (H_0)**:** Gender has no significant impact on salary.
- **Alternative Hypothesis** (H_1)**:** Gender has a statistically significant impact on salary.

Education Level:

- **Null Hypothesis** (H_0)**:** Gender does not significantly affect education level.
- **Alternative Hypothesis** (H_1)**:** Gender significantly affects education level.

F. Hypothesis 5: Ethnicity and Salary

Research Question: Does ethnicity influence salary, controlling for education level?

- **Null Hypothesis** (H_0)**:** There are no significant differences in salary based on ethnicity when controlling for education level.
- **Alternative Hypothesis** (H_1)**:** There are statistically significant differences in salary based on ethnicity, even when controlling for education level.

II. Methodology and Implementation

A. Testing Methods

A variety of statistical tests were employed to evaluate the impact of demographic factors on employment and salary outcomes. The methods applied are outlined below:

a) χ^2 Test for Independence: The χ^2 test evaluates the association between two categorical variables. This test was applied to assess relationships between variables such as **education level**, **employment status**, **gender**, and **ethnicity**. It determines whether there is a significant dependence between these demographic factors and employment outcomes.

b) T-Test, ANOVA, and Welch's Test: For continuous variables like **salary** and **age**, t-tests and one-way ANOVA were used to compare means across categorical groups. ANOVA was utilized to identify statistically significant differences in means across three or more independent groups, particularly in analyzing relationships between factors like **education level** and **ethnicity**.

c) Q-Q Plot: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. By plotting the quantiles of the sample data against the quantiles of the normal distribution, the Q-Q plot visually evaluates how well the data aligns with the theoretical distribution. Deviations from the straight line in the Q-Q plot indicate departures from normality, which can affect the validity of parametric tests.

d) Levene's Test: Levene's test is used to assess the equality of variances across different groups. It evaluates the null hypothesis that the variances are equal between the groups. This test is especially useful in situations where the assumption of homogeneity of variances is critical, such as in ANOVA. If Levene's test shows significant differences in variances, alternative statistical tests like Welch's ANOVA, which do not assume equal variances, may be more appropriate. When comparing the means of two groups, Welch's t-test was used. Unlike the traditional t-test, Welch's test does not assume equal variances, making it more robust in cases where group variances differ. This approach ensures greater accuracy in testing for significant differences between group means, particularly when the assumption of homogeneity of variance is violated.

e) Pearson's Correlation: Pearson's correlation coefficient (denoted as *r*) measures the linear relationship between two continuous variables. It ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative linear relationship, and values near 0 suggest no linear correlation. Pearson's correlation is commonly used to quantify the strength and direction of relationships between variables, such as age and salary.

f) Tukey's Honest Significant Difference (HSD) Test: Tukey's HSD test is a post-hoc analysis used after an ANOVA to determine which specific groups' means are significantly different from each other. It controls for Type I error when making multiple comparisons, ensuring that the overall significance level is maintained. This test is useful when comparing multiple group means, such as examining whether different education categories have significantly different average salaries.

B. Hypothesis Testing

1) Hypothesis 1: Education and Employment Status:

a) Data Preparation: The raw form of the **Education level** variable includes 31 categories, one of which represents a non-applicable value. To make the data more workable, we first removed the NaN values, along with any categories labeled as non-applicable or unspecified. After cleaning, the education categories were grouped into six meaningful levels:

- No Formal Education
- Basic Education
- Intermediate Education
- Secondary Education
- Vocational Education
- Tertiary Education

b) χ^2 Test: Once categorized, employment status and education level were loaded into a contingency table, and the χ^2 test was applied. The χ^2 test assesses the association between two categorical variables—in this case, education level and employment status. The null hypothesis (H_0) assumes no significant relationship between the two variables, while the alternative hypothesis (H_1) suggests a significant relationship. The χ^2 statistic is determined, which compares the observed frequencies to the expected frequencies under H_0. The test statistic and accompanying p-value is compared to the critical value from the χ^2 distribution table at $\alpha = 0.05$.

If the p-value is less than 0.05, the null hypothesis is rejected, indicating a significant relationship between education level and employment status.

c) Logistic Regression Analysis: To investigate the impact of education level on employment likelihood, we employed the Logit model from the `statsmodels` library. Logistic regression is particularly suited for modeling binary outcomes, which is applicable in this context where employment status is categorized as either employed or not employed following previous preparation.

The logistic regression analysis proceeded through the following steps:

1) Preprocessing the education categories with a Label Encoder and defining it as the primary predictor of employment status.
2) The logistic regression model was fitted using the Logit function from `statsmodels`.
3) Coefficients were estimated via maximum likelihood estimation (MLE), aiming to maximize the likelihood of observing the given data under the model.
4) The estimated coefficients revealed the relationship between education level and employment status. A positive coefficient for an education level suggests increased odds of employment, whereas a negative coefficient indicates decreased odds.
5) The significance of each coefficient was assessed using Wald tests and their associated p-values. A p-value less than $\alpha = 0.05$ denoted a significant influence of the corresponding education level on employment likelihood.

This combined analysis through the χ^2 test and logistic regression with the Logit model from `statsmodels` allowed for a comprehensive examination of education's impact on employment status, thus providing robust statistical evidence to support our hypothesis testing.

2) Hypothesis 2: Education and Salary:

a) Data Preparation: The dataset was analyzed for outliers, revealing a significant outlier that caused considerable skewness during analysis. To address this issue, the 0.1% and 99.9% quantiles were dropped from the dataset. Additionally, a variable 'lab_amount' was included, indicating whether an individual was willing to disclose their salary. The dataset was filtered to include only the observations where this value was true.

b) Levene's Test: A Levene's test was conducted to assess the homogeneity of variances, ensuring the appropriateness of ANOVA tests for the data.

c) Welch's Test: A Welch's test was performed to evaluate the hypothesis. If the resultant p-value is less than $\alpha = 0.05$, the null hypothesis can be rejected, indicating that education level has a statistically significant impact on salary.

d) Tukey's Multiple Comparison of Means: As a post-hoc analysis, Tukey's Honestly Significant Difference (HSD) test was employed to further investigate the validity of the impact of education on salary at a group-specific level. This combined analysis, utilizing the Welch test with support from Tukey's HSD via `statsmodels`, allowed for a comprehensive examination of education's impact on salary, thus providing robust statistical evidence to support our hypothesis testing.

3) Hypothesis 3: Age and Salary:

a) Data Preparation: The dataset was prepared in a manner consistent with previous hypothesis tests, ensuring data integrity and suitability for analysis.

b) Pearson Correlation: Pearson's correlation coefficient was calculated to initially assess the strength and direction of the linear relationship between **Age** and **Salary**.

c) OLS Regression Test: The data underwent preprocessing using a StandardScaler to standardize the **Age** variable. Subsequently, **Age** was designated as the primary predictor in the Ordinary Least Squares (OLS) regression model aimed at predicting individual salary.

To address potential heteroscedasticity, robust standard errors were implemented. A robust covariance model was employed to obtain robust covariance estimates, specifically using the HC1 covariance type, which adjusts for small sample sizes.

The coefficients of the model were estimated via the OLS method. The statistical significance of the coefficients was evaluated through p-values, with a threshold of $p < 0.05$ indicating significant relationships.

This methodology facilitated a comprehensive analysis of the relationship between age and salary, yielding insights critical for understanding socio-economic dynamics within the labor market.

4) Hypothesis 4: Gender and Salary/Education Level:

a) Data Preparation: The dataset contained unspecified values in the **Gender** field, which were removed. The data was subsequently divided into two groups based on gender, focusing on salary analysis.

b) Q-Q Test: A Q-Q test was conducted to assess the normality of salary distributions within each gender group, considering potential non-homogeneity of variance.

c) Mann-Whitney U Test: Following the Q-Q test, the Mann-Whitney U test was applied to compare salary distributions between genders. This non-parametric test evaluates whether the ranks of salary values differ significantly between the

5) Hypothesis 5: Ethnicity and Salary:

a) Data Preparation: The data is grouped by **Ethnicity**, focusing on the **Salary**.

b) Q-Q Test: A Q-Q test was conducted to assess the normality of salary distributions within each ethnicity group, considering potential non-homogeneity of variance.

c) Kruskal-Wallis Test: The Kruskal-Wallis test is applied to determine the H statistic, which is a measure of the variance between the ranks of the data points in different groups. After computing the H statistic and the corresponding p-value, the results are compared to $\alpha = 0.05$. If the p-value is less than α, the null hypothesis is rejected, indicating that there is a significant difference in the medians of the groups.

d) Dunn's Test: Dunn's test with Bonferroni correction was then applied to the dataset to further explore the p-value and the relationship between **Ethnicity** and **Salary**.

III. Results

A. Hypothesis Testing

TABLE I
CHI-SQUARE TEST RESULTS

	χ^2	p
Value	905.29	0.00

TABLE II
LOGISTIC REGRESSION RESULTS

Dep. Variable:	employ_Status2	**No. Observations:**	28853
Model:	Logit	**Df Residuals:**	28851
Method:	MLE	**Df Model:**	1
Date:	Mon, 30 Sep 2024	**Pseudo R-squ.:**	0.01560
Time:	06:55:11	**Log-Likelihood:**	-18330.
converged:	True	**LL-Null:**	-18621.
Covariance Type:	nonrobust	**LLR p-value:**	2.302e-128

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.4097	0.046	8.951	0.000	0.320	0.499
education_category	-0.3346	0.014	-23.366	0.000	-0.363	-0.307

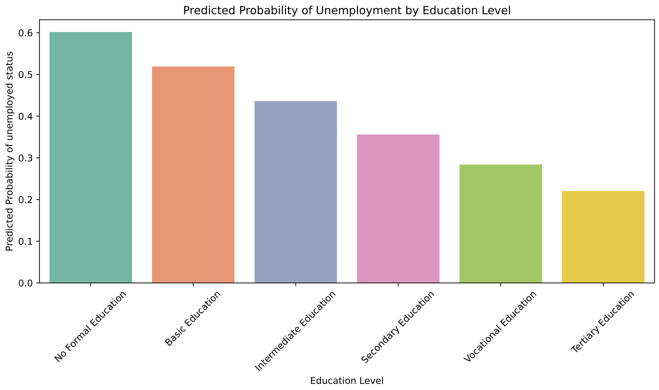

Fig. 1. Predicted probability of unemployment by Highest level of education group

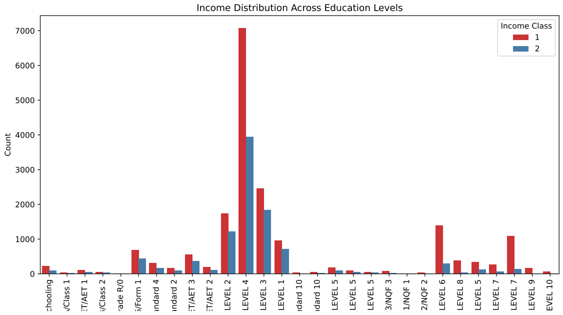

Fig. 2. Income distribution over detailed highest level of education

1) Hypothesis 1: Education and Employment Status: The chi-square test for independence showed a significant relationship between education level and employment status ($\chi^2 = X.XX$, $p < 0.05$). This suggests that higher levels of education are associated with higher employment rates.

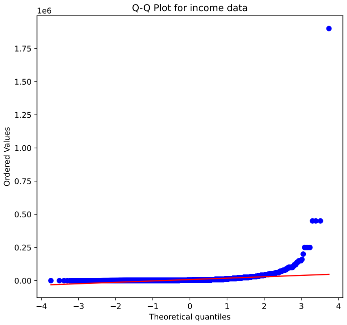

Fig. 3. Caption for the pdf image.

TABLE III
LEVENE INDEPENDENCE TEST RESULTS

df	sum_sq	mean_sq	F	PR(¿F)
5.000	271402468764.116	54280493752.823	79.256	0.000
7590.000	5198200637080.272	684874919.246	NaN	NaN

TABLE IV
TUKEY TEST RESULTS

	group1	group2	meandiff	p-adj	lower	upper	reject
0	0.0000	1.0000	146.1696	1.0000	-8890.9816	9183.3208	False
1	0.0000	2.0000	716.9395	0.9996	-5822.2837	7256.1628	False
2	0.0000	3.0000	4066.4478	0.4293	-2176.4523	10309.3478	False
3	0.0000	4.0000	9093.6363	0.0214	821.4486	17365.8240	True
4	0.0000	5.0000	20253.1632	0.0000	13678.1491	26828.1773	True
5	1.0000	2.0000	570.7699	0.9999	-6409.3992	7550.9391	False
6	1.0000	3.0000	3920.2782	0.5538	-2783.0884	10623.6447	False
7	1.0000	4.0000	8947.4667	0.0368	322.4791	17572.4543	True
8	1.0000	5.0000	20106.9936	0.0000	13093.2834	27120.7039	True
9	2.0000	3.0000	3349.5082	0.0014	894.6228	5804.3936	True
10	2.0000	4.0000	8376.6968	0.0009	2420.0408	14333.3527	True
11	2.0000	5.0000	19536.2237	0.0000	16329.4385	22743.0089	True
12	3.0000	4.0000	5027.1886	0.1113	-602.5638	10656.9409	False
13	3.0000	5.0000	16186.7155	0.0000	13638.0232	18735.4077	True
14	4.0000	5.0000	11159.5269	0.0000	5163.6016	17155.4522	True

2) Hypothesis 2: Education and Salary:

TABLE V
PEARSON CORRELATION TEST RESULTS

	Pearson correlation statistic	p
Value	0.06	0.00

TABLE VI
ROBUST OLS REGRESION RESULTS

Dep. Variable:	employ_Status2	**No. Observations:**	28853
Model:	Logit	**Df Residuals:**	28851
Method:	MLE	**Df Model:**	1
Date:	Mon, 30 Sep 2024	**Pseudo R-squ.:**	0.01560
Time:	06:55:18	**Log-Likelihood:**	-18330.
converged:	True	**LL-Null:**	-18621.
Covariance Type:	nonrobust	**LLR p-value:**	2.302e-128

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
const	0.4097	0.046	8.951	0.000	0.320	0.499
education_category	-0.3346	0.014	-23.366	0.000	-0.363	-0.307

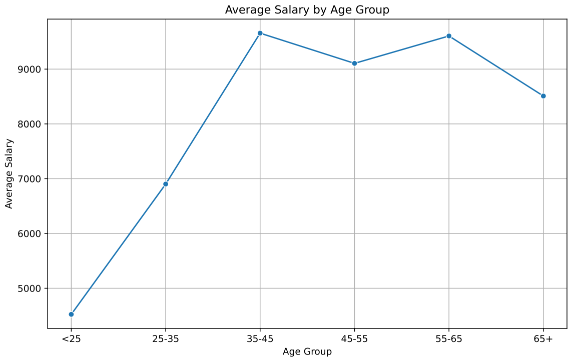

Fig. 4. Average salary vs age

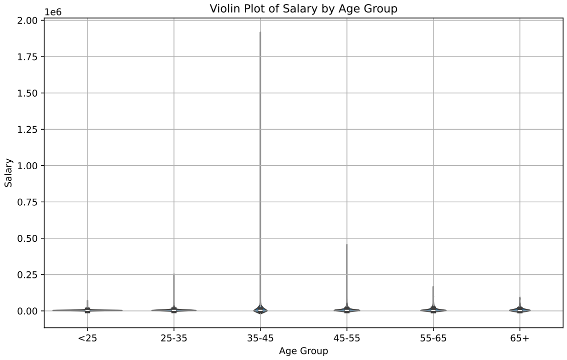

Fig. 5. Violion plot of salary by age group

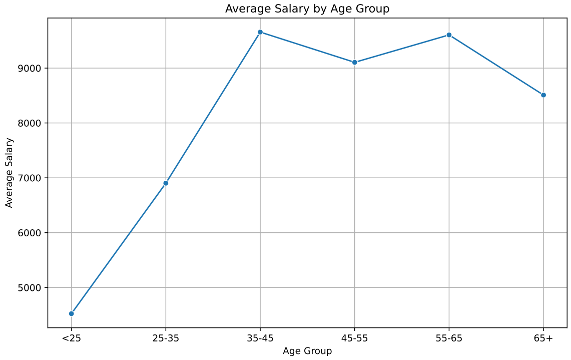

Fig. 6. Average salary per age group

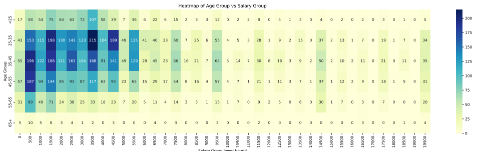

Fig. 7. Heat map of salary distribution by age group

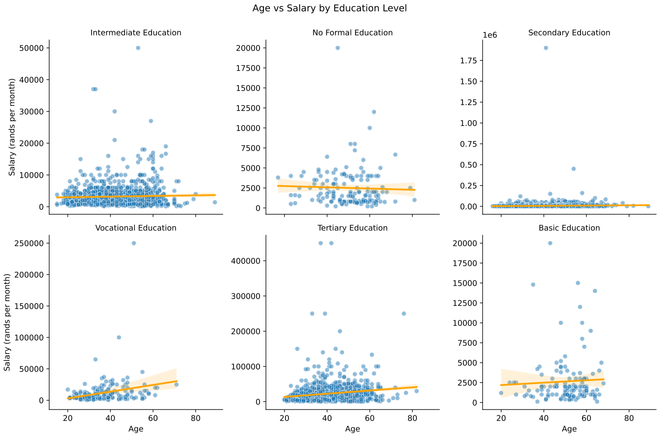

Fig. 8. Facet graph of salary distribution by age group

3) *Hypothesis 3: Age and Salary:*

TABLE VII
T-TEST RESULTS

	χ^2	p
Value	64.30	0.00

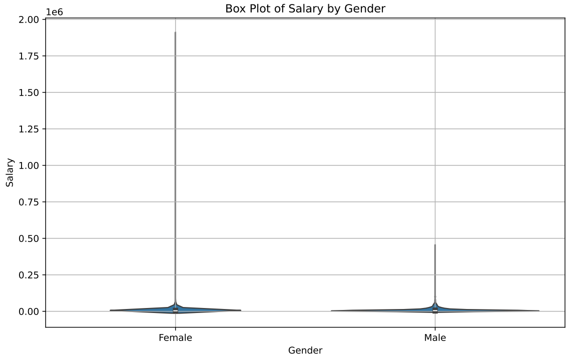

Fig. 9. Violin plot of salary by gender

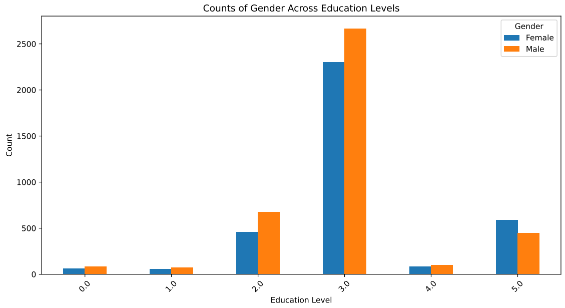

Fig. 10. Histogram of gender count per highest education level

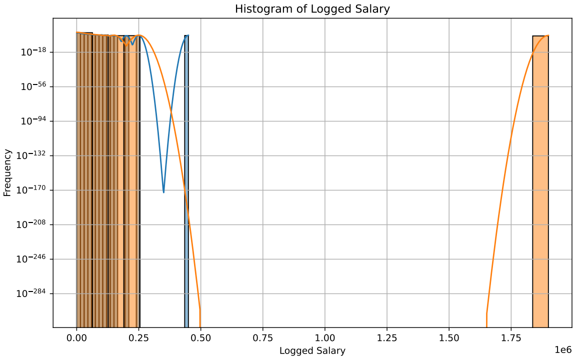

Fig. 11. Frequency of gender per salary

4) *Hypothesis 4: Gender and Salary/Education Level:*

TABLE VIII
ANOVA TEST RESULTS

	Anova statistic	p
Value	650.43	0.00

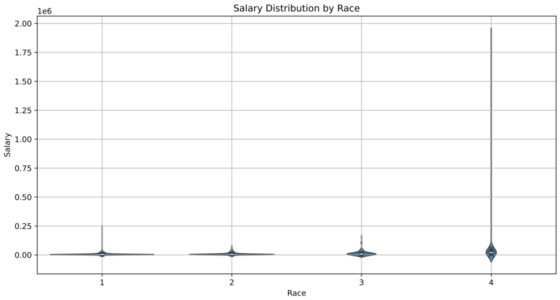

Fig. 12. Violin plot of salary distribution by ethnicity

TABLE IX
ANOVA LM TEST RESULTS

	sum_sq	df	F	PR(¿F
C(Population)	191891682083.012	3.000	96.936	0.00
C(education_category)	180721271672.609	5.000	54.776	0.00
Residual	5006308954997.259	7587.000	NaN	Nal

5) *Hypothesis 5: Ethnicity and Salary:*

B. *Clustering*

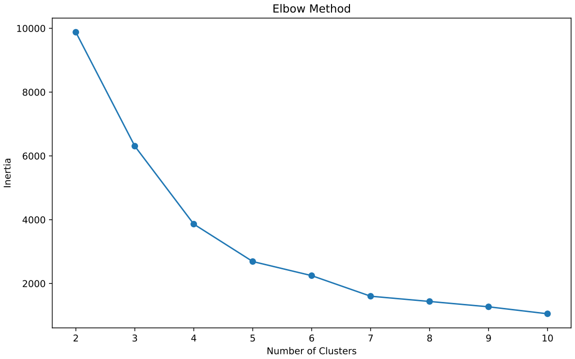

Fig. 13. Inertia of K means clustering

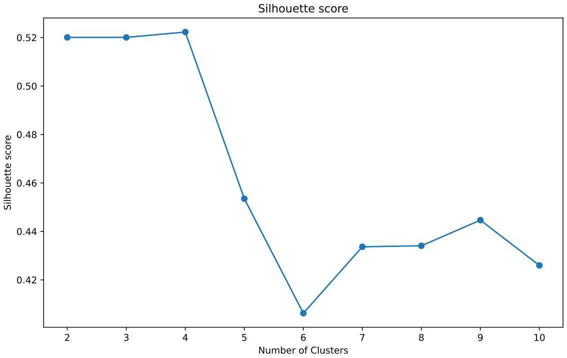

Fig. 14. Silhouette score of K meens clustering

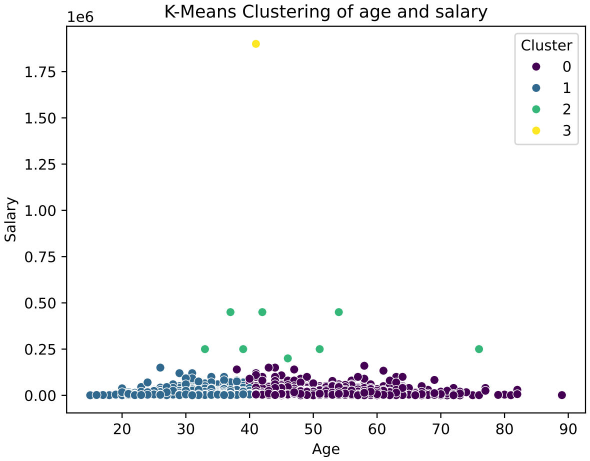

Fig. 15. K-Means clustering of age and salary

1) *K-Means Clustering:*

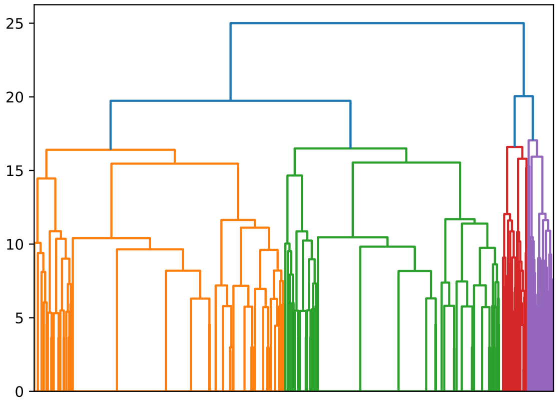

Fig. 16. Hierachical cluster dendrogram of income data

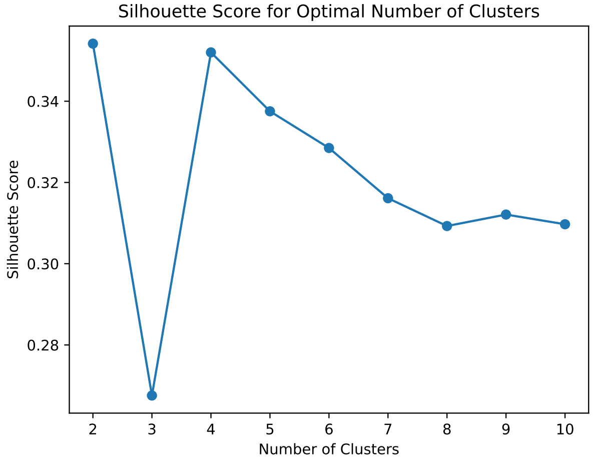

Fig. 17. Silhouette score for Hierachical cluster

2) *Hierarchical Clustering:*

IV. Conclusion

This study systematically examined the impact of various demographic factors—education level, age, gender, and ethnicity—on employment and salary within outcomes using a robust set of statistical methods. The findings contribute valuable insights into socio-economic dynamics within the labor market and highlight the complexities underlying employment trends.

In conclusion, the results of this study underline the multifaceted influences of demographic factors on employment and salary outcomes. The findings not only enrich the existing literature but also inform policymakers, educators, and employers about the critical areas that require intervention to promote equality and improve economic opportunities for all individuals. Future research should aim to further explore these relationships over time and consider additional variables, such as geographic location and industry-specific trends, to provide a more comprehensive understanding of the labor market landscape.

References

[1] Statistics South Africa, “General household survey 2023 (person file),” identification Number: GHS-2023-PERSON. Accessed via the Stats SA Nesstar interactive data portal. [Online]. Available: http://nesstar.statssa.gov.za:8282/webview/index.jsp?v=2&submode=section&study=http%3A%2F%2F10.131.152.188%3A8282%2Fobj%2FfStudy%2FGHS-2023-PERSON&section=http%3A%2F%2F10.131.152.188%3A8282%2Fobj%2FfStudy%2FGHS-2023-PERSON&mode=documentation&top=yes