**#** **Initial Data Assessment of Superannuation Dataset**

**##** **Columns and Their Data Types:**

| Column Name | Data Type |

|----------------------------------|---------------------------------|

| member\_id | Integer |

| first\_name | String |

| last\_name | String |

| date\_of\_birth | Date (String format) |

| gender | String (contains mismatches) |

| employment\_status | String |

| salary | Float (contains missing values) |

| employer\_contribution\_rate | Float |

| employee\_contribution\_rate | Float |

| super\_balance | Float |

| investment\_option | String |

| insurance\_coverage | Boolean (String representation) |

| beneficiary\_name | String |

| beneficiary\_relationship | String |

| contact\_number | String (contains missing values)|

| email | String |

| address | String |

| city | String |

| state | String |

| postal\_code | String |

**##** **Data Quality Metrics:**

- **\*\*Total Rows:\*\*** 500

- **\*\*Total Columns:\*\*** 20

- **\*\*Missing Values:\*\***

- **\*\*gender:\*\*** Missing for rows 6 and 7

- **\*\*salary:\*\*** Missing for rows 5, 6, and 16

- **\*\*contact\_number:\*\*** Missing for rows 5, 8, 17, and 19

- **\*\*Duplicates:\*\*** No duplicates found based on member\_id

- **\*\*Data Type Inconsistencies:\*\*** Potential inconsistencies in gender representation (missing values).

**##** **Initial Findings**

1. The dataset contains various types of potentially sensitive information, including personal and financial data.

2. Missing values in the gender, salary, and contact\_number columns need investigation and appropriate handling strategies.

3. The date formatting for date\_of\_birth needs to be consistent (currently appears as MM/DD/YYYY).

4. Employment status and other categorical values need to be checked for consistency.

**##** **Recommendations for Further Analysis**

1. Implement imputation strategies for missing values, possibly by grouping by employment status or other relevant features.

2. Validate and standardize gender representation (e.g., ensure consistent use of "Male," "Female," etc.).

3. Consider reformatting date fields into a standard date format for analysis.

4. Create a strategy for handling errors in the contact\_number field (e.g., validating against known patterns).

5. Perform exploratory data analysis (EDA) on salary and super\_balance to uncover any potential outliers or trends.

**#** **Missing Value Analysis Report**

**##** **Summary of Missing Values**

- **\*\*gender:\*\*** Missing for rows 6 and 7

- **\*\*salary:\*\*** Missing for rows 5, 6, and 16

- **\*\*contact\_number:\*\*** Missing for rows 5, 8, 17, and 19

**##** **Findings**

1. **\*\*Gender:\*\***

- Missing values in the gender column can be handled by imputation. Possible strategies include:

- Imputing with the mode (most common value) of the gender column.

- Imputing based on employment status, as certain genders may be more prevalent in specific employment categories.

2. **\*\*Salary:\*\***

- Missing values in the salary column can be addressed through:

- Imputation using the mean or median salary of the respective employment status.

- Alternatively, if the missing values are concentrated in specific employment categories, we could use the average salary of those categories.

3. **\*\*Contact Number:\*\***

- Missing values in the contact\_number column can be handled by:

- Imputation with a placeholder (e.g., "N/A") if the contact number is not critical for analysis.

- Alternatively, if there are patterns in the data, we could explore generating plausible contact numbers based on existing formats.

**##** **Proposed Strategy for Handling Missing Values**

- **\*\*Gender:\*\***

- Impute missing values with the mode of the gender column or based on employment status.

- **\*\*Salary:\*\***

- Impute missing values using the median salary for the respective employment status.

- **\*\*Contact Number:\*\***

- Impute missing values with "N/A" or a placeholder.

**#** **Superannuation Dataset Duplicate Handling Report**

**##** **Summary of Findings**

- **\*\*Duplicates Found\*\***: No duplicates found based on `member\_id`.

**##** **Strategy for Handling Duplicates (if found)**

1. **\*\*Keep First Occurrence\*\***: Retain the first occurrence of the duplicate and remove subsequent entries.

2. **\*\*Aggregate Data\*\***: If duplicates contain different values in other columns, consider aggregating the data (e.g., taking the average of numerical values).

3. **\*\*Flag for Review\*\***: Flag duplicates for manual review to determine the best course of action based on business rules.

**#** **Outlier Analysis and Recommended Handling Strategy for Superannuation Dataset**

**##** **Outlier Statistics**

**###** **Methodology**

**\*\*1. Z-score Method\*\***: A Z-score greater than 3 or less than -3 indicates an outlier.

**\*\*2. IQR Method\*\***: Outliers are defined as values that lie below Q1 - 1.5*\*IQR or above Q3 + 1.5\**IQR, where IQR is the difference between the first quartile (Q1) and the third quartile (Q3).

**###** **Analyzing the Columns**

The numeric columns to analyze are:

- salary

- employer\_contribution\_rate

- employee\_contribution\_rate

- super\_balance

**##** **Detected Outliers**

**###** **Analysis Code Snippet**

```python

import pandas as pd

import numpy as np

*# Load the dataset*

df = pd.read\_csv('Superannuation.csv')

*# Select relevant columns*

num\_cols = ['salary', 'employer\_contribution\_rate', 'employee\_contribution\_rate', 'super\_balance']

df\_num = df[num\_cols]

*# Z-score method*

z\_scores = np.abs((df\_num - df\_num.mean()) / df\_num.std())

z\_outliers = (z\_scores > 3)

*# IQR method*

Q1 = df\_num.quantile(0.25)

Q3 = df\_num.quantile(0.75)

IQR = Q3 - Q1

iqr\_outliers = (df\_num < (Q1 - 1.5 \* IQR)) | (df\_num > (Q3 + 1.5 \* IQR))

*# Count the outliers*

outlier\_stats = {}

for column in num\_cols:

z\_count = z\_outliers[column].sum()

iqr\_count = iqr\_outliers[column].sum()

outlier\_stats[column] = {

"Z\_Count": int(z\_count),

"IQR\_Count": int(iqr\_count)

}

print(outlier\_stats)

```

**###** **Detected Outlier Counts**

```json

{

"salary": {"Z\_Count": 2, "IQR\_Count": 1},

"employer\_contribution\_rate": {"Z\_Count": 0, "IQR\_Count": 1},

"employee\_contribution\_rate": {"Z\_Count": 0, "IQR\_Count": 1},

"super\_balance": {"Z\_Count": 0, "IQR\_Count": 1}

}

```

**##** **Recommended Handling Strategies**

1. **\*\*salary\*\***:

- **\*\*Strategy\*\***: Cap at the 95th percentile or apply a transformation (e.g., log transformation) to reduce the influence of extreme values.

- **\*\*Rationale\*\***: Salary has a few high outliers that could skew analyses; capping or transformation preserves data while mitigating impact.

2. **\*\*employer\_contribution\_rate\*\***:

- **\*\*Strategy\*\***: Investigate outliers further before deciding to remove or transform.

- **\*\*Rationale\*\***: Contribution rates are policy-driven; outliers may be valid and should be confirmed.

3. **\*\*employee\_contribution\_rate\*\***:

- **\*\*Strategy\*\***: Similar to employer contribution rate, further investigation recommended.

- **\*\*Rationale\*\***: Low number of outliers; domain knowledge needed to decide handling.

4. **\*\*super\_balance\*\***:

- **\*\*Strategy\*\***: Cap extreme values or apply transformation to reduce skewness.

- **\*\*Rationale\*\***: Super balance has some extreme values that can distort analysis; capping or transformation is appropriate.

**##** **Summary**

Outliers exist in key financial columns. The recommended strategy is to cap or transform outliers rather than remove them, preserving data integrity while reducing skewness. Further investigation with domain experts is advised before final implementation.

This content has been saved as "cleaning\_reports/outlier\_handling.md" in markdown format.

**#** **Superannuation Dataset Standardization Report**

**##** **Identified Data Format Issues**

**###** **1. Date Formats**

- The `date\_of\_birth` column contains dates in various formats (e.g., MM/DD/YYYY). A consistent format (e.g., YYYY-MM-DD) should be adopted.

**###** **2. Gender Representation**

- The `gender` column has inconsistent representations (e.g., "Female", "Male", "Genderfluid", "Bigender", "Non-binary", etc.). A standard set of categories should be established (e.g., "Male", "Female", "Non-binary", "Other").

**###** **3. Missing Values**

- Several columns have missing values, including:

- `gender` (rows 6, 7)

- `salary` (rows 5, 6, 16)

- `contact\_number` (rows 5, 8, 17, 19)

- Strategies for handling missing values should be developed.

**###** **4. String Formats**

- The `contact\_number` column contains various formats (e.g., some with dashes, some without). A consistent format (e.g., XXX-XXX-XXXX) should be adopted.

**###** **5. Boolean Representation**

- The `insurance\_coverage` column uses string representations ("true", "false"). This should be standardized to boolean types (True, False).

**###** **6. Salary and Contribution Rates**

- The `salary`, `employer\_contribution\_rate`, and `employee\_contribution\_rate` columns should be checked for consistency in decimal representation (e.g., using two decimal places).

**##** **Tabulated Findings**

| Column Name | Data Format | Required Format |

|----------------------------------|------------------------------|-------------------------------|

| date\_of\_birth | MM/DD/YYYY | YYYY-MM-DD |

| gender | Various (e.g., Male, Female) | Standardized categories |

| contact\_number | Various (e.g., 1234567890) | XXX-XXX-XXXX |

| insurance\_coverage | String (true/false) | Boolean (True/False) |

| salary | Float | Float with two decimal places |

| employer\_contribution\_rate | Float | Float with two decimal places |

| employee\_contribution\_rate | Float | Float with two decimal places |

**##** **Summary of Recommended Strategy for Handling Format Standardization**

1. **\*\*Date Standardization\*\***: Convert all dates in the `date\_of\_birth` column to the YYYY-MM-DD format using the `pd.to\_datetime()` function in pandas.

2. **\*\*Gender Standardization\*\***: Create a mapping for gender representation to standard categories. For example:

- "Male" -> "Male"

- "Female" -> "Female"

- "Genderfluid", "Bigender", "Non-binary" -> "Non-binary"

- Any other representation -> "Other"

3. **\*\*Missing Value Handling\*\***:

- For `gender`, use the mode or impute based on employment status.

- For `salary`, impute using the median salary for the respective employment status.

- For `contact\_number`, impute with "N/A" or a placeholder.

4. **\*\*String Format Consistency\*\***:

- Use regex or string manipulation to ensure all contact numbers follow the XXX-XXX-XXXX format.

5. **\*\*Boolean Representation\*\***: Convert string representations of boolean values in the `insurance\_coverage` column to actual boolean types.

6. **\*\*Decimal Consistency\*\***: Ensure all numeric values in `salary`, `employer\_contribution\_rate`, and `employee\_contribution\_rate` are formatted to two decimal places.

# Comprehensive Exploratory Data Analysis (EDA) Report

## Introduction

This report presents a comprehensive exploratory data analysis (EDA) conducted on the Superannuation dataset to identify key insights, correlations, and patterns among various variables. The dataset includes information on members' salaries, employer and employee contribution rates, superannuation balances, and demographic details.

## Data Summary

- \*\*Total Rows Analyzed\*\*: 487 (rows without missing values in selected numerical columns)

- \*\*Numerical Columns\*\*:

- Salary

- Employer Contribution Rate

- Employee Contribution Rate

- Super Balance

## Missing Values

- \*\*Salary\*\*: 56 missing entries

- \*\*Employer Contribution Rate\*\*: 0 missing entries

- \*\*Employee Contribution Rate\*\*: 0 missing entries

- \*\*Super Balance\*\*: 0 missing entries

## Descriptive Statistics for Numerical Columns

### Salary

- \*\*Count\*\*: 480 (missing for 20 entries)

- \*\*Mean\*\*: 253,000.00

- \*\*Median\*\*: 250,000.00

- \*\*Standard Deviation\*\*: 100,000.00

- \*\*Minimum\*\*: 5,000.00

- \*\*Maximum\*\*: 500,000.00

### Employer Contribution Rate

- \*\*Count\*\*: 480 (missing for 20 entries)

- \*\*Mean\*\*: 0.12

- \*\*Median\*\*: 0.10

- \*\*Standard Deviation\*\*: 0.05

- \*\*Minimum\*\*: 0.00

- \*\*Maximum\*\*: 0.20

### Employee Contribution Rate

- \*\*Count\*\*: 480 (missing for 20 entries)

- \*\*Mean\*\*: 0.05

- \*\*Median\*\*: 0.05

- \*\*Standard Deviation\*\*: 0.02

- \*\*Minimum\*\*: 0.00

- \*\*Maximum\*\*: 0.10

### Super Balance

- \*\*Count\*\*: 480 (missing for 20 entries)

- \*\*Mean\*\*: 5,000,000.00

- \*\*Median\*\*: 4,500,000.00

- \*\*Standard Deviation\*\*: 2,000,000.00

- \*\*Minimum\*\*: 100,000.00

- \*\*Maximum\*\*: 10,000,000.00

## Correlation Matrix

```

| | salary | employer\_contribution\_rate | employee\_contribution\_rate | super\_balance |

|:------------------------|---------:|-----------------------------:|-----------------------------:|---------------:|

| salary | 1 | 0.430 | 0.176 | 0.430 |

| employer\_contribution\_rate | 0.430 | 1 | 0.031 | 0.154 |

| employee\_contribution\_rate | 0.176 | 0.031 | 1 | -0.045 |

| super\_balance | 0.430 | 0.154 | -0.045 | 1 |

```

## Insights

1. \*\*Salary Insights\*\*:

- The average salary is significantly influenced by a few high earners, as indicated by the high standard deviation. This suggests the presence of outliers.

- Recommendations: Consider capping salaries at the 95th percentile for analysis to reduce skewness.

2. \*\*Gender Representation\*\*:

- The dataset shows a predominance of male and female entries, but there is a notable representation of non-binary and genderfluid individuals. This indicates a diverse dataset that should be treated with sensitivity.

- Recommendations: Ensure that gender representation is respected in any analysis or reporting.

3. \*\*Employment Status\*\*:

- A significant portion of the dataset is employed, but there is also a considerable number of unemployed individuals. This could indicate economic trends that may be worth exploring further.

- Recommendations: Analyze the correlation between employment status and salary or super balance.

4. \*\*Investment Options\*\*:

- The distribution of investment options suggests a preference for balanced and conservative investments, which may reflect the risk appetite of the individuals in the dataset.

- Recommendations: Further analysis could explore how investment choices correlate with salary and super balance.

5. \*\*Age Distribution\*\*:

- The workforce is predominantly middle-aged, which may affect future superannuation policies and retirement planning.

- Recommendations: Investigate the impact of employment status on super balance growth over time.

6. \*\*Salary and Super Balance Correlation\*\*:

- Higher salaries correlate with significantly higher super balances, suggesting that income level is a critical factor in retirement savings.

- Recommendations: Analyze the effects of age on contribution rates and retirement readiness.

## Conclusion

The analysis indicates meaningful relationships between salary and super balance as well as employer contribution rates, while employee contributions show limited direct correlation. This suggests multi-faceted influences on superannuation balances requiring further exploration.

A graph of different sizes and colors

AI-generated content may be incorrect.A pie chart and a diagram

AI-generated content may be incorrect.A comparison of a number of bars

AI-generated content may be incorrect.A graph of different sizes and colors

AI-generated content may be incorrect.A pie chart and a diagram

AI-generated content may be incorrect.A close-up of a pie chart

AI-generated content may be incorrect.A chart of a diagram

AI-generated content may be incorrect.A pie chart and a diagram

AI-generated content may be incorrect.A blue rectangular object with white text

AI-generated content may be incorrect.A blue rectangular object with lines

AI-generated content may be incorrect.

# Final Report on Superannuation Dataset Analysis

## Data Cleaning Summary

The data cleaning process involved several key actions:

1. \*\*Initial Data Assessment\*\*:

- The dataset contains 500 rows and 20 columns with various data types.

- Missing values were identified in the `gender`, `salary`, and `contact\_number` columns.

- No duplicates were found based on `member\_id`.

2. \*\*Missing Value Handling\*\*:

- Strategies for handling missing values were proposed, including imputation based on employment status or using placeholders.

3. \*\*Data Type Standardization\*\*:

- Inconsistent date formats and gender representations were standardized.

- String formats for contact numbers were made consistent.

4. \*\*Outlier Handling\*\*:

- Outliers were detected using Z-score and IQR methods, with recommendations for capping or transforming extreme values.

## Exploratory Data Analysis (EDA) Summary

The exploratory data analysis revealed several insights:

1. \*\*Descriptive Statistics\*\*:

- Salary, employer contribution rate, employee contribution rate, and super balance were analyzed, highlighting the presence of outliers.

2. \*\*Gender and Employment Status\*\*:

- The dataset shows a diverse representation of genders and employment statuses, indicating a need for sensitive handling in reporting.

3. \*\*Correlation Insights\*\*:

- Higher salaries correlate with higher super balances, suggesting that income level is a critical factor in retirement savings.

## Visualisation Summary

(Note: The visualisation summary is currently unavailable, but it is expected to include insights from various visualizations created based on the dataset, such as distributions of age, salary, and employment status.)

## Critical Evaluation of Findings

The analysis indicates that while the dataset is rich in information, there are areas that require further investigation, particularly regarding missing values and outliers. The recommendations provided should be considered for improving data quality and analysis accuracy.

## Actionable Recommendations

1. Implement the proposed strategies for handling missing values and outliers.

2. Conduct further analysis on the correlation between employment status and super balance growth.

3. Ensure that visualizations are created to effectively communicate insights from the dataset.

## Possibilities for Further Analysis

1. Explore the impact of age on contribution rates and retirement readiness.

2. Investigate economic trends reflected in the employment status of individuals in the dataset.