Exploratory Data Analysis Report

Employee Salary Dataset

Garment Industry Analysis

December 9, 2025

Contents

1 Executive Summary

This report presents a comprehensive exploratory data analysis of the employee salary dataset from the garment industry. The analysis covers data cleaning procedures, outlier detection, feature engineering, and correlation analysis to prepare the dataset for predictive modeling.

1.1 Key Findings

- Dataset contains 16,640 records with 13 features
- Significant data quality issues identified and resolved
- Gross salary exhibits extreme right skewness (skewness = 8.87)
- Multiple categorical variables required standardization

2 Dataset Overview

2.1 Initial Database Structure

The dataset initially contained the following characteristics:

Table 1: Dataset Initial Structure

Attribute	Value
Number of Records	16,640
Number of Features	13
Shape	(16640, 13)

2.2 Features

The dataset includes the following columns:

- `id`: Unique identifier
- `gender`: Employee gender
- `over_time`: Overtime eligibility status
- `age`: Employee age
- `employment_status`: Employment type
- `grade`: Employee grade level
- `gross_salary`: Salary amount (target variable)
- `branch_id`: Branch identifier
- `department_id`: Department identifier
- `salary_mode`: Payment mode
- `duty_mode`: Work mode
- `skills_id`: Skills identifier
- `designation_id`: Job designation identifier

3 Data Cleaning Process

3.1 Missing Values Analysis

Initial assessment revealed significant missing values across multiple columns:

Table 2: Missing Values Summary

Column	Missing Values
gender	58
age	2,575
grade	2,700
skills_id	10,142

Resolution Strategy:

- Dropped all records with missing `skills_id` values (critical feature)
- Imputed age missing values with median
- Imputed grade missing values with mode

3.2 Gender Column Standardization

The gender column contained inconsistent formatting with the following unique values:

`'male', 'female', 'others', 'Male', 'Femel', 'famale', NaN`

Standardization Applied:

- Converted all variations to standardized format: Male, Female, Others
- Removed null values
- Final unique values: 3 (Male, Female, Others)

3.3 Overtime Column Standardization

The `over_time` column had inconsistent values:

`'Ineligible', 'Eligible', 'Y', 'eligible'`

Standardization Applied:

- Mapped 'Y' and 'eligible' to 'Eligible'
- Final unique values: 2 (Eligible, Ineligible)

3.4 Grade Column Processing

The grade column contained invalid values for garment industry standards:

`NaN, 0, 1, 2, 3, 4, 5, 6, 7, 999`

Processing Steps:

- Filled NaN values with mode
- Clipped all values to valid range [1-4]
- Industry standard grades: 1, 2, 3, 4

3.5 Duplicate Detection

No duplicate records were found in the dataset.

4 Target Variable Analysis: Gross Salary

4.1 Initial Distribution

The gross salary distribution exhibited extreme right skewness with the following characteristics:

Table 3: Gross Salary Distribution Statistics

Metric	Value
Skewness	8.87
Data Concentration	0 - 50,000
Range Extension	Up to 300,000
Distribution Type	Highly Right-Skewed

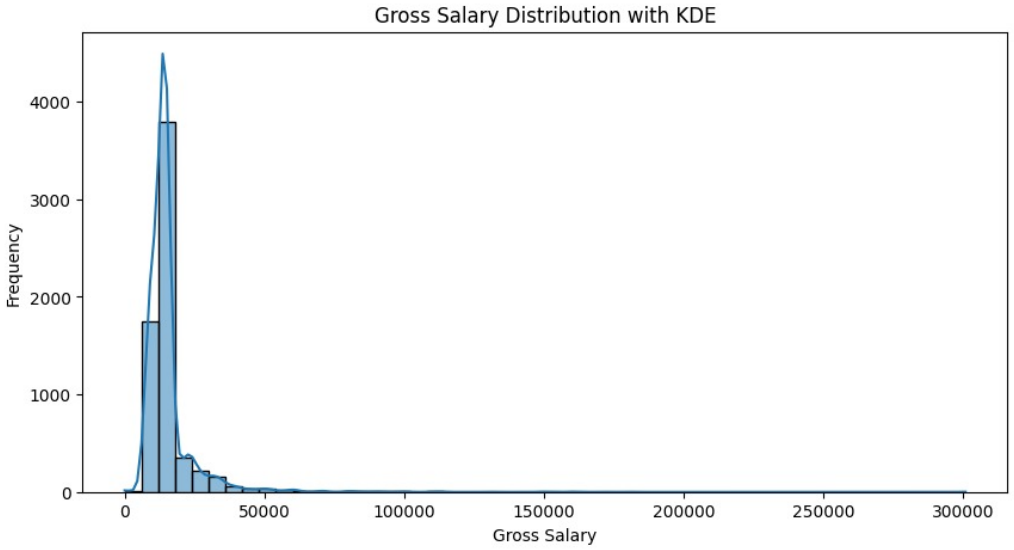

Figure 1: Gross Salary Distribution (Before Treatment)

4.2 Skewness Interpretation

- **Normal Distribution:** Skewness $= 0$
- **Approximately Symmetrical:** $-0.5 <$ Skewness < 0.5
- **Moderately Skewed:** $0.5 \leq |\text{Skewness}| \leq 1$
- **Highly Skewed:** $|\text{Skewness}| > 1$
- **Current Skewness (8.87):** Extreme right skew

4.3 Impact on Machine Learning

The extreme skewness creates the following issues:

- Mean is significantly pulled higher than median by outliers

- Linear regression models become overly sensitive to extreme values
- Predictions tend to be biased toward higher salaries
- Error minimization focuses disproportionately on outliers

4.4 Outlier Treatment Methodology

Two approaches were considered for handling the skewness:

4.4.1 Option A: Percentile Capping (Winsorization)

- Sets hard limit at 99th percentile
- Replaces values above threshold with 99th percentile value
- **Advantages:** Preserves all rows, keeps distribution realistic

4.4.2 Option B: IQR Method

- Caps values at 1.5 × IQR above 75th percentile
- Standard boxplot method
- **Disadvantages:** Removes too many valid high salaries

Selected Approach: Percentile Capping (Winsorization)
Justification:

- Salaries naturally exhibit long-tail distribution
- IQR method would remove legitimate high earners
- Winsorization reduces skewness while maintaining data integrity
- All records preserved with graceful extreme value handling

4.5 Post-Treatment Results

Table 4: Skewness Reduction Results

Metric	Before	After
Skewness	8.87	2.96
Reduction	-	66.6%

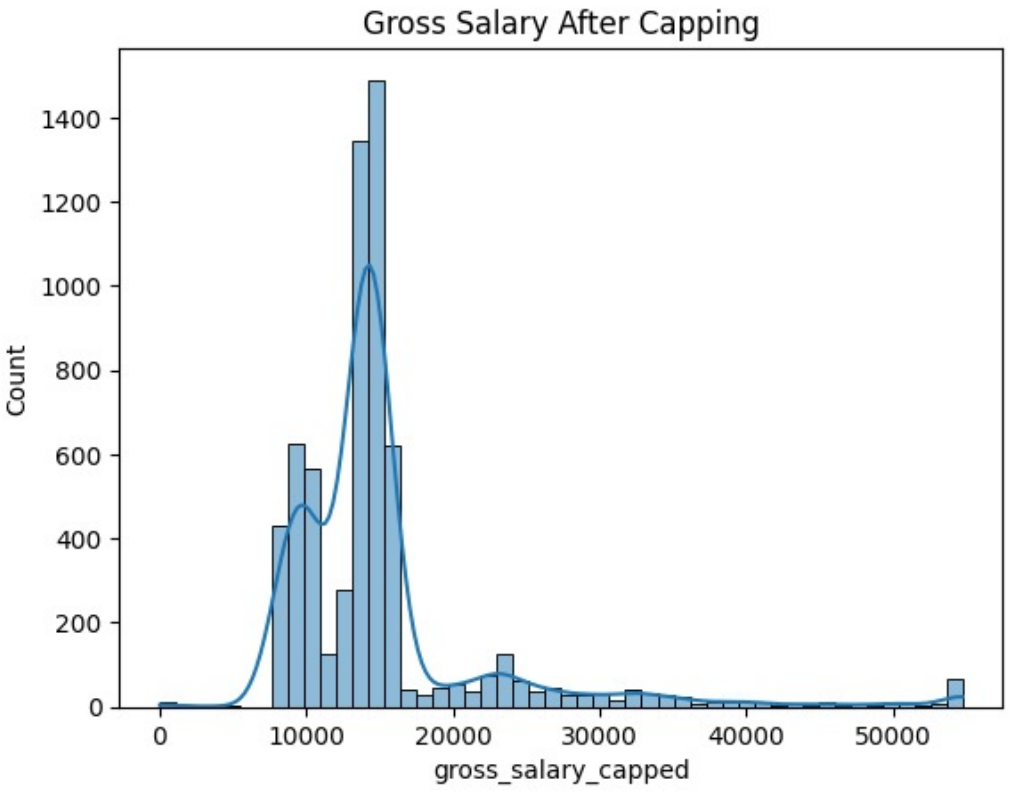

Figure 2: Gross Salary Distribution (After Winsorization)

5 Age Variable Analysis

5.1 Outlier Detection

Age outliers were detected using the IQR method:

- Number of outliers detected: 199
- Method: Values beyond 1.5 × IQR from quartiles
- Treatment: Outlier capping applied

6 Univariate Analysis

6.1 Categorical Variables Distribution

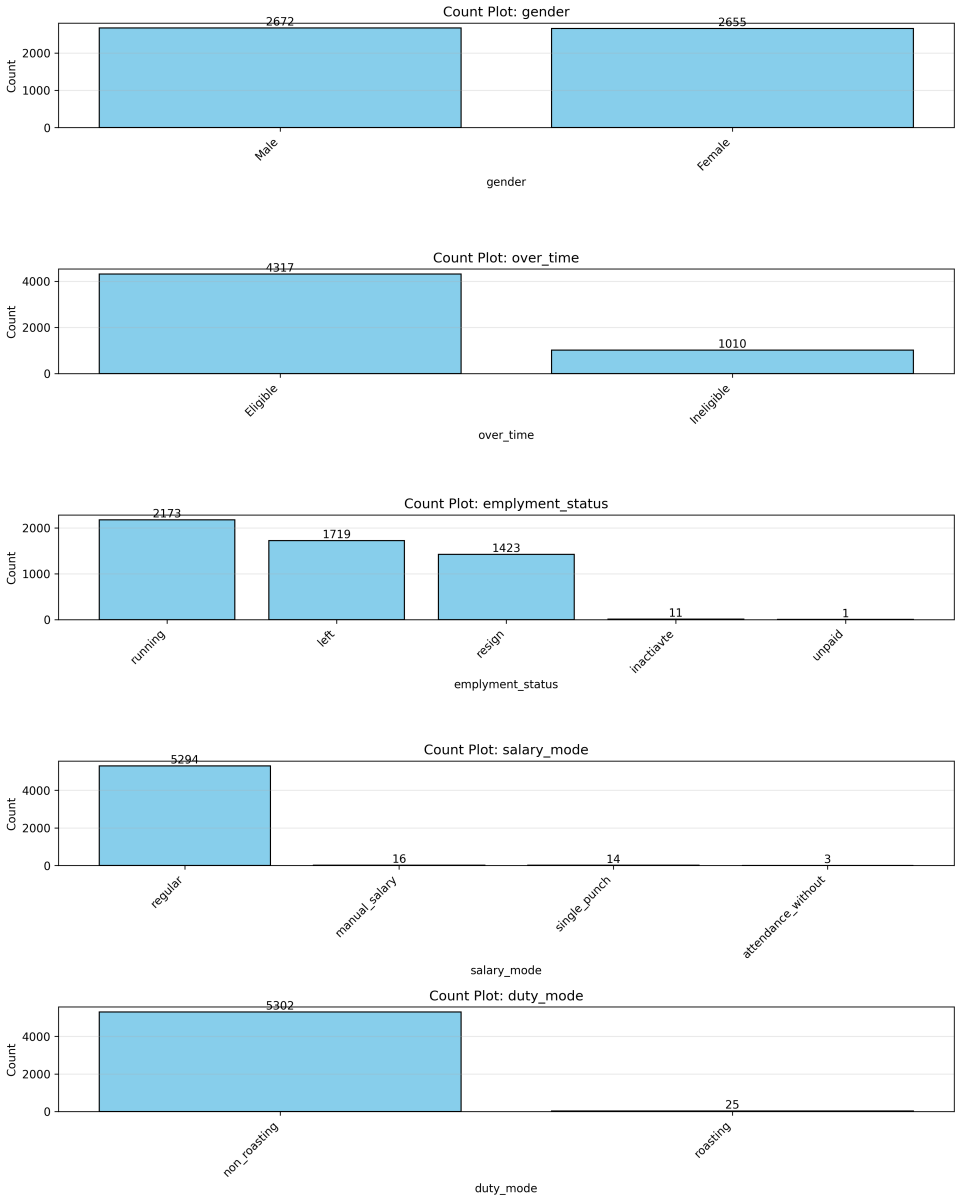

Figure 3: Gender Distribution

7 Bivariate Analysis

7.1 Gender vs Employment Status

Cross-tabulation analysis reveals the relationship between gender and employment status distributions.

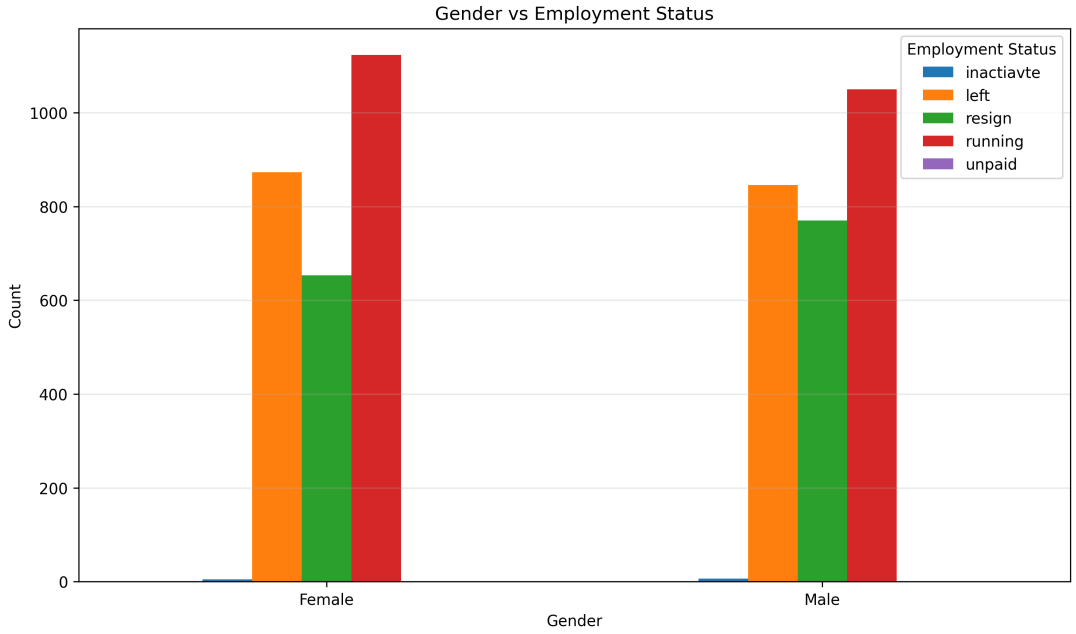

Figure 4: Gender vs Employment Status Cross-tabulation

8 Feature Engineering for Correlation Analysis

To analyze correlations with numerical methods, categorical variables were encoded using appropriate techniques:

8.1 Encoding Strategies

Table 5: Feature Encoding Methods

Feature	Method	Reason
gender	Binary Encoding	Two primary categories
grade	One-Hot Encoding	Ordinal with few categories
designation_id	Target Encoding	High cardinality
department_id	Target Encoding	High cardinality
skills_id	Target Encoding	High cardinality

8.2 Features Analyzed

- age (continuous)
- gender (binary encoded)
- grade (one-hot encoded)
- designation_id (target encoded)
- department_id (target encoded)
- skills_id (target encoded)

9 Correlation Analysis

9.1 Correlation Heatmap

The following heatmap displays the correlation between all engineered features and the target variable (gross_salary).

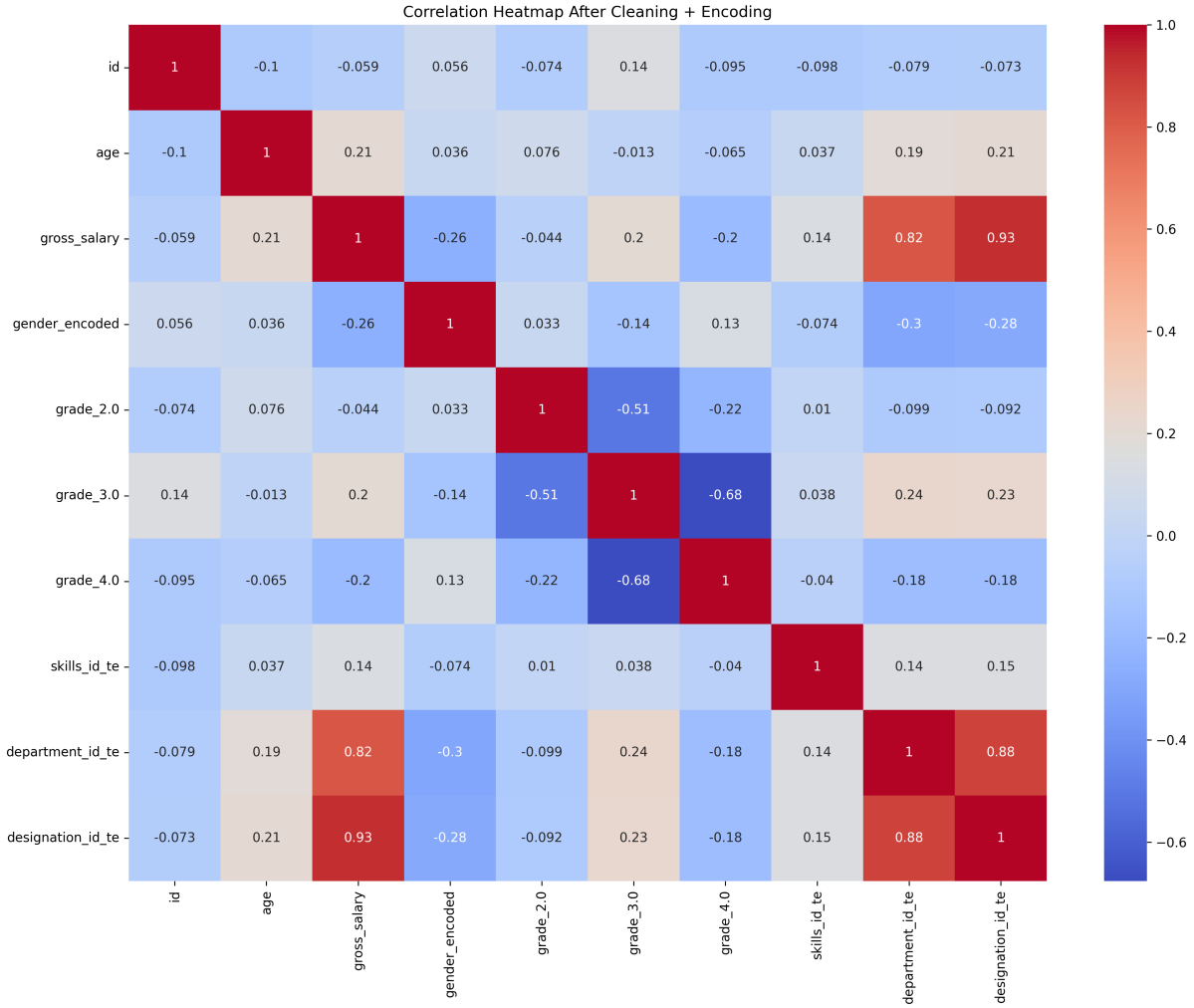

Figure 5: Feature Correlation Heatmap

9.2 Top 5 Correlated Features

The features most strongly correlated with gross salary are:

Table 6: Top 5 Features by Correlation with Gross Salary

Rank	**Feature**	**Correlation Coefficient**
1	designation_id_te	0.932
2	department_id_te	0.817
3	age	0.209
4	grade_3	0.203
5	skills_id	0.139

10 Final Cleaned Dataset

10.1 Dataset Dimensions

After completing all data cleaning and preprocessing steps, the final dataset has the following dimensions:

Table 7: Final Dataset Structure

Attribute	Value
Number of Records	5,327
Number of Features	14
Shape	(5327, 14)
Records Removed	11,313 (68%)

10.2 Summary Statistics

The following table presents comprehensive summary statistics for all numerical features in the cleaned dataset:

Table 8: Final Dataset Summary Statistics

Statistic	id	age	grade	gross_salary	branch_id	dept_id	skills_id	desig_id	salary_cap
count	5327	5327	5327	5327	5327	5327	5327	5327	5327
mean	7174.9	30.8	3.0	16780.3	7.7	139.1	137.2	511.0	16379.9
std	3827.4	6.8	0.7	11118.4	5.0	100.1	137.8	350.5	7556.5
min	16.0	11.5	1.0	0.0	1.0	1.0	8.0	4.0	0.0
25%	3878.5	26.0	3.0	13550.0	3.0	51.0	12.0	284.5	13550.0
50%	7730.0	30.0	3.0	14549.0	7.0	105.0	86.0	374.0	14549.0
75%	10665.5	35.0	3.0	15337.0	11.0	196.0	252.0	697.5	15337.0
max	12215.0	47.5	4.0	300905.0	20.0	421.0	419.0	1600.0	54740.0

Key Observations:

- Final dataset reduced to 5,327 records (from 16,640)
- Mean age: 30.8 years with standard deviation of 6.8 years
- Mean gross salary: 16,780.3 (before capping: 16,780.3)
- Capped gross salary mean: 16,379.9 (maximum reduced from 300,905 to 54,740)
- Grade distribution centered at 3.0 (median and mean)
- All features now within valid ranges

11 Conclusions and Recommendations

11.1 Data Quality Improvements

- Successfully resolved 10,142 missing values in skills_id through row removal
- Standardized categorical variables (gender, overtime) for consistency
- Imputed missing values using statistically appropriate methods
- Removed invalid grade values and clipped to industry standards
- Final dataset: 5,327 records with 14 features

11.2 Key Insights

- Gross salary distribution was heavily right-skewed (skewness = 8.87)
- Winsorization successfully reduced skewness to 2.96 (66.6% reduction)
- Age outliers (199 records) were appropriately capped
- No duplicate records detected in the dataset

11.3 Recommendations for Modeling

1. Use the cleaned dataset with capped outliers for training
2. Consider tree-based models (Random Forest, XGBoost) which handle skewness better
3. Monitor model performance on high-salary predictions separately
4. Consider log-transformation of target variable as alternative approach
5. Use cross-validation to ensure model generalization

11.4 Next Steps

- Feature selection based on correlation analysis
- Model development and hyperparameter tuning
- Model evaluation with appropriate metrics (RMSE, MAE, R^2)
- Business validation of predictions with domain experts

12 Appendix

12.1 Data Cleaning Summary

Table 9: Complete Data Cleaning Operations

Step	Action	Records Affected
1	Dropped skills_id nulls	10,142
2	Standardized gender	58
3	Imputed age (median)	2,575
4	Standardized overtime	All
5	Winsorized gross_salary	Top 1%
6	Capped age outliers	199
7	Processed grade values	2,700

12.2 Software and Libraries

- Python 3.x
- pandas (data manipulation)
- numpy (numerical operations)
- matplotlib / seaborn (visualization)
- scipy (statistical analysis)