Removed department name column as missing values are around 50% and imputing values will imbalance the dataset.

Removal of columns:

1. Overtime, Other salaries due to weaker correlation

Low value after square transformation and log transformation

1. Negative values removal for salaries, overtime, other\_salaries, retirement, other\_benefits
2. Removal of Employee\_identifier as all the values are unique.
3. Load dataset
4. Sample – 4 lakh instances
5. Preprocessing

Salaries has strong correlation with Total salary, retirement, health, other benefits, total benefits, total compensation

1. Group values in order to reduce columns (for dummy conversion, we grouped department\_code, organization, job\_family, job).
2. Data preprocessing:
   * Sample data as there was issue of memory allocation in R (Dimensionality reduction)
   * Remove insignificant columns for proper attribute selection (department, employee identifier, org grp code, union code, job fam code, job code)
   * Replace negative values by mean for numerical variables (salaries, overtime, other\_salaries, total\_salary, retirement, health\_and\_dental, other\_benefits, total\_benefits, total\_compensation)
   * Check and fill missing values or NA with mean in case of numerical variables or mode in case of categorical variables (department\_code, union, job)
   * Check correlation between variables and apply transformation for weaker variables.
   * Remove variables with weaker correlation after transformation
   * Normalize numerical variables as ranges are different for various numerical variables
3. Convert categorical variables into dummies and remove n-1 variable (year, year\_type, organization\_group, department\_code, union, job\_family, job)
4. Divide data into train and test set for Hold-out evaluation method

ANOVA

Values greater than alpha ie 0.05

subdata$jobAquatics

subdata$jobArchitecture, Bldgs & Bdcomm

subdata$jobCashier and Fare collector

subdata$jobEmployment & Training

subdata$jobExecutive officer

subdata$jobProbation and Protection service

subdata$jobTypist and Support

ANOVA – JOB FAMILY

subdata$job\_familyEnergy & Environment

subdata$job\_familyOfficers

department code – anova

subdata$department\_code203644

subdata$department\_code228855

subdata$department\_code228883

subdata$department\_code228886

subdata$department\_code229010

subdata$department\_code229017

subdata$department\_code229024 -0.0134077 0.0121875 -1.100 0.271284

department\_code229047 0.1153626 0.0591488 1.950 0.051133

subdata$department\_code229259 -0.0011080 0.0061513 -0.180 0.857058

subdata$department\_code229261 -0.0583137 0.0341777 -1.706 0.087974 .

subdata$department\_code229264 0.0249448 0.0144399 1.727 0.084083 .

subdata$department\_code229987 -0.0037182 0.0077886 -0.477 0.633080

subdata$department\_code229992 -0.0088501 0.0113017 -0.783 0.433584

subdata$department\_code229994 -0.0181006 0.0118308 -1.530 0.126029

subdata$department\_code229997 -0.0224219 0.0296109 -0.757 0.448921

subdata$department\_code232021 -0.0155904 0.0224113 -0.696 0.486649

subdata$department\_code232031 -0.0067550 0.0080469 -0.839 0.401218

subdata$department\_code232041 -0.0126534 0.0075830 -1.669 0.095188

subdata$department\_code232051 0.1560582 0.0836318 1.866 0.062041

subdata$department\_code232076 -0.0926143 0.0483048 -1.917 0.055203

subdata$department\_code232328 0.0712529 0.0483048 1.475 0.140196

subdata$department\_code232362 0.0144298 0.0083730 1.723 0.084822 .

subdata$department\_code232395 -0.0946595 0.0483048 -1.960 0.050041 .

subdata$department\_code291644 -0.0032181 0.0151133 -0.213 0.831378

subdata$department\_codeADM 0.0039542 0.0022643 1.746 0.080760 .

subdata$department\_codeART 0.0087562 0.0074464 1.176 0.239638

subdata$department\_codeBOS -0.0010885 0.0046999 -0.232 0.816852

subdata$department\_codeCHF -0.0050630 0.0065447 -0.774 0.439165

subdata$department\_codeCII -0.0325329 0.0316489 -1.028 0.303984

subdata$department\_codeCSC -0.0034780 0.0148783 -0.234 0.815168

subdata$department\_codeDPA 0.0122209 0.0112044 1.091 0.275394

subdata$department\_codeENV -0.0035597 0.0046832 -0.760 0.447196

subdata$department\_codeHOM -0.0111453 0.0062090 -1.795 0.072653 .

subdata$department\_codeHRC 0.0029968 0.0092797 0.323 0.746740

subdata$department\_codeHSS -0.0019575 0.0057688 -0.339 0.734371

subdata$department\_codeJUV -0.0050252 0.0032236 -1.559 0.119028

subdata$department\_codeLLB 0.0481027 0.0418418 1.150 0.250296

subdata$department\_codeSCI 0.0271316 0.0138506 1.959 0.050129 .

anova – org group

subdata$organization\_groupGeneral Administration & Finance

Tramsformation on y var

Don’t remove variables after anova