2016 Employee Salary Analysis

Name - Savindu Mahasen Ruhunuhewa

**Normality testing**

Normality testing use for identifying the data is normally distributed or not. Normally distribution means there is no right or left skewedness available in the data distribution.To do the correlational analysis normality testing is necessary. For checking the data is distributed normal manner or not, need to use the Anderson darling, Lilliefors, and Shapiro wiki test of normality testing.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = annual\_base\_pay

(\*) Whether or not annual\_base\_pay are normally distributed?

H­0: Employee annual\_base\_pays are normally distributed

H­1: Employee annual\_base\_pays are not normally distributed

Anderson-Darling normality test

data: annual\_base\_pay

A = 0.24133, p-value = 0.7346

Decision: p-value= 0.7346 > α=0.05 => Accept H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: annual\_base\_pay

D = 0.10167, p-value = 0.8924

Decision: p-value= 0.8924 > α=0.05 => Accept H0 at 5% significant level

Shapiro-Wilk normality test

data: annual\_base\_pay

W = 0.96647, p-value = 0.7294

Decision: p-value= 0.7294 > α=0.05 => Accept H0 at 5% significant level

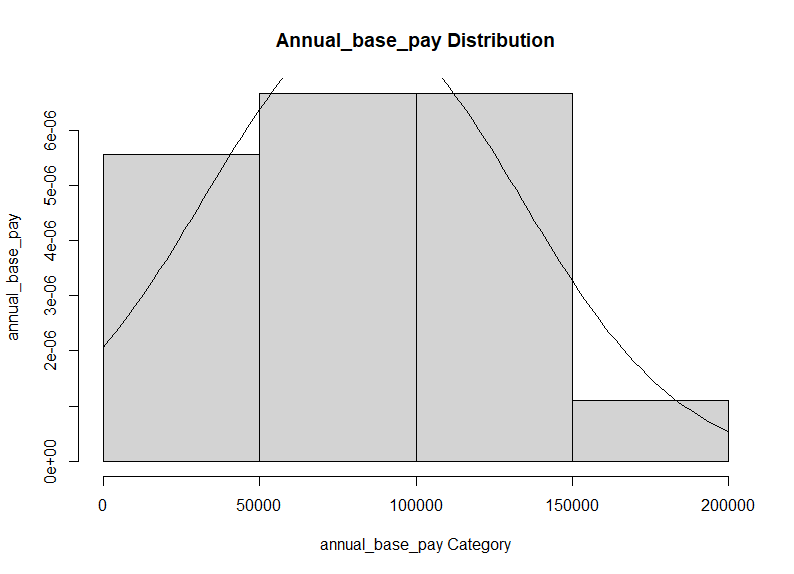
Conclusion - According to the Anderson darling (p = 0.7346), Lillefors (p = 0.8924), and Shapiro-Wilk testings, at a 5% significant level “annual\_base\_pay” data is normally distributed.

> summary(annual\_base\_pay)

Min. 1st Qu. Median Mean 3rd Qu. Max.

24 47500 85500 82851 120000 190000

But when it comes above mentioned testings, rejected the accepted null hypothesis, but when we considering the summary statistics above mentioned there is a small left skewedness available. Because (Mean < Median = 82851 < 85500). According to the statistics theory normally distribution means “mean value and median value should be same or mean value should nearly equal to the median value” (Chen, Normal distribution: What it is, uses, and Formula). When we considering this case mean value of “annual\_base\_pay” nearly equal to the median value of “annual\_base\_pay”. So then based on the summary statistics, also “annual\_base\_pay” can be considered as normally distributed.



Interpretation of montacarlo graph

According to the above bell curve, data is distributed normally at 5% significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = location\_latitude

(\*) Whether or not location\_latitude are normally distributed?

H­0: location\_latitude are normally distributed

H­1: location\_latitude are not normally distributed

Anderson-Darling normality test

data: location\_latitude

A = 1.7724, p-value = 9.585e-05

Decision: p-value= 9.585e-05 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: location\_latitude

D = 0.26331, p-value = 0.001824

Decision: p-value= 0.001824 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: location\_latitude

W = 0.80257, p-value = 0.001652

Decision: p-value = 0.001652 < α=0.05 => Reject H0 at 5% significant level

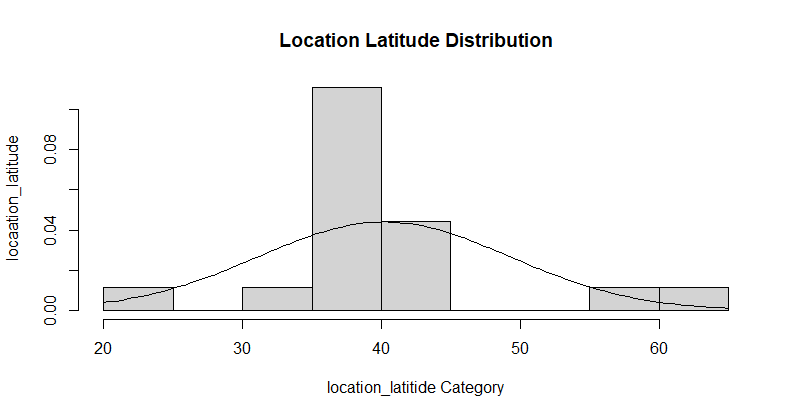
Conclusion - According to the Anderson darling ( p = 9.585e-05), Lillefor (p = 0.001824), and shapiro wiki (p =0.001652) tests, location\_latitude data is not normally distributed at a 5 % significant level.

summary(location\_latitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.

21.00 37.83 38.91 40.20 41.10 62.00

When considering the above summary statistics, mean value is greater than median value (Mean > Median = 40.20 > 38.91). which means there is a little bit of right skewedness available in the this location\_latitude data. According to this summary statistics, data can be considered as normally distributed. But when we considering the above mentioned hypothesis, it rejected the null hypothesis. Which means data is not normally distributed. In this case, as final result “location\_latitude” data can be considered as normally distributed. Because 100% accurate result (data is normally distributed or not) is given based on hypothesis not summary statistics.



**Interpretation of montercarlo graph**

According to the above bell curve there is a right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = location\_longitude

(\*) Whether or not location\_longitude are normally distributed?

H­0: location\_longitude are normally distributed

H­1: location\_longitude are not normally distributed

Anderson-Darling normality test

data: location\_longitude

A = 2.7534, p-value = 2.955e-07

Decision: p-value= 2.955e-07 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: location\_longitude

D = 0.32237, p-value = 2.644e-05

Decision: p-value= 2.644e-05 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: location\_longitude

W = 0.63659, p-value = 1.596e-05

Decision: p-value = 1.596e-05< α=0.05 => Reject H0 at 5% significant level

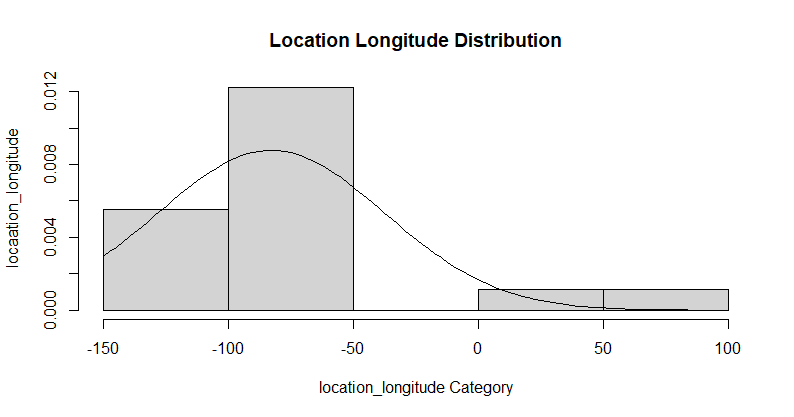
Conclusion - According to the Anderson darling ( p = 2.955e-07), Lillefor (p = 2.644e-05), and shapiro wiki (p = 1.596e-05) tests location\_longitude data is not normally distributed at a 5 % significant level.

summary(location\_longitude)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-122.41 -103.26 -97.00 -82.92 -84.70 57.00

When considering the above summary statistics mean value is greater than median value (Mean > Median = -82.92 > -97.00). which means there is a huge right skewedness available in the this location\_longitude data. According to above mentioned summary statistics, data can be considered as not normally distributed. And also considering the above mentioned testings, reject the null hypothesis. Which means “location\_longitude” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = job\_title

(\*) Whether or not job\_title are normally distributed?

H­0: job\_title are normally distributed

H­1: job\_title are not normally distributed

Anderson-Darling normality test

data: job\_title

A = 0.31695, p-value = 0.5109

Decision: p-value= 0.5109 > α=0.05 => Accept H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: job\_title

D = 0.11317, p-value = 0.785

Decision: p-value= 0.785 > α=0.05 => Accept H0 at 5% significant level

Shapiro-Wilk normality test

data: job\_title

W = 0.94547, p-value = 0.3584

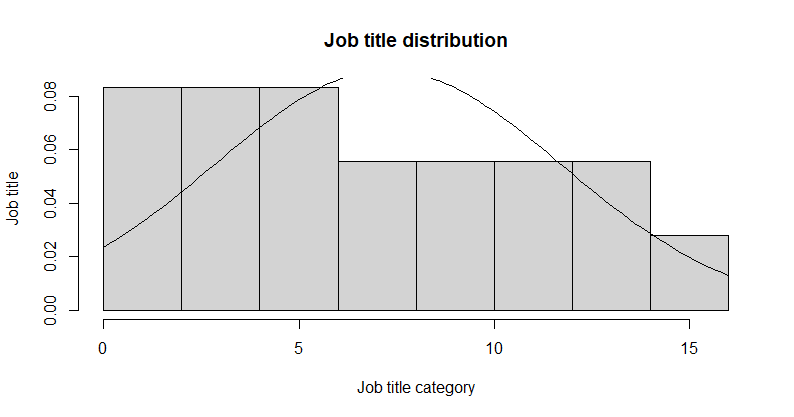
Decision: p-value = 0.3584 > α=0.05 => Accept H0 at 5% significant level

Conclusion - According to the Anderson darling ( p = 0.5109), Lillefor (p = 0.785), and shapiro wiki (p = 0.3584) tests job\_title data is normally distributed at a 5 % significant level.

summary(job\_title) Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 3.250 6.500 7.278 10.750 15.000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 7.278 > 6.500). which means there is a small right skewedness available in the this “job\_title” data. According to this summary statistics Mean value nearly equal to the Median value. Which means “job\_title” data can be considered as normally distributed. In above mentioned normality testings, Accepted the null hypothesis, which means based on the summary statistics and normality testings, “job\_title” data can be considered as normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a little bit of right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = job\_title\_category

(\*) Whether or not job\_title\_category are normally distributed?

H­0: job\_title\_category are normally distributed

H­1: job\_title\_category are not normally distributed

Anderson-Darling normality test

data: job\_title\_category

A = 1.0608, p-value = 0.006516

Decision: p-value= 0.006516 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: job\_title\_category

D = 0.26344, p-value = 0.001809

Decision: p-value= 0.001809 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: job\_title\_category

W = 0.8689, p-value = 0.01706

Decision: p-value = 0.01706 < α=0.05 => Reject H0 at 5% significant level

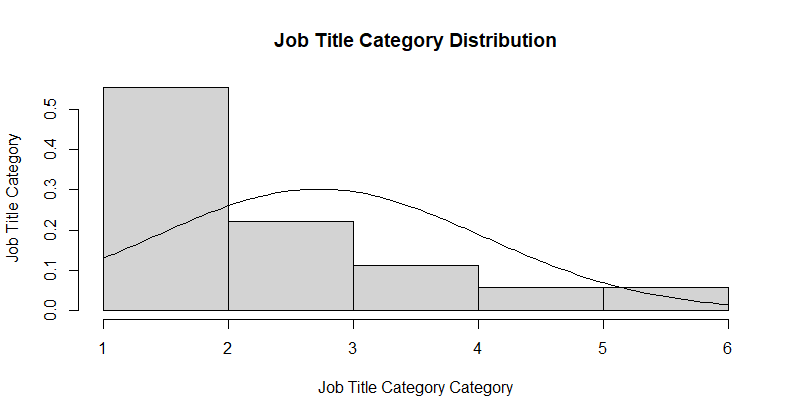
Conclusion - According to the Anderson darling ( p = 0.006516), Lillefor (p = 0.001809), and shapiro wiki (p = 0.01706) tests job\_title\_category data is normally distributed at a 5 % significant level.

summary(job\_title\_category)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.000 2.000 2.722 3.000 6.000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 2.722 > 2.000). which means there is a right skewedness available in the this “job\_title\_category” data. According to this summary statistics and above mentioned normality testing, “job\_title\_category” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = total\_experience\_years

(\*) Whether or not total\_experience\_years are normally distributed?

H­0: total\_experience\_years are normally distributed

H­1: total\_experience\_years are not normally distributed

Anderson-Darling normality test

data: total\_experience\_years

A = 1.8112, p-value = 7.621e-05

Decision: p-value= 7.621e-05 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: total\_experience\_years

D = 0.26841, p-value = 0.001318

Decision: p-value= 0.001318 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: total\_experience\_years

W = 0.70971, p-value = 0.0001038

Decision: p-value = 0.0001038 < α=0.05 => Reject H0 at 5% significant level

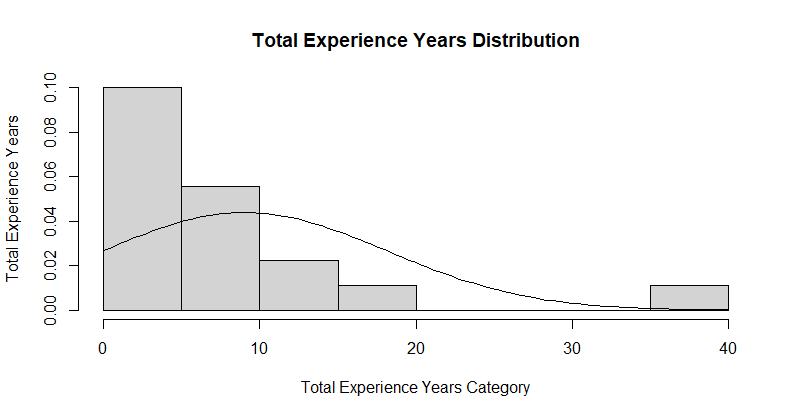
Conclusion - According to the Anderson darling ( p = 7.621e-05), Lillefor (p = 0.001318), and shapiro wiki (p = 0.0001038) tests total\_experience\_years data is not normally distributed at a 5 % significant level.

summary(total\_experience\_years)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 5.000 5.500 9.056 9.500 40.000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 9.056 > 5.500). which means there is a huge right skewedness available in the this “job\_title” data. According to this summary statistics and above mentioned tests, “total\_experience\_years” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a huge right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = employee\_experience\_years

(\*) Whether or not employee\_experience\_years are normally distributed?

H­0: employee\_experience\_years are normally distributed

H­1: employee\_experience\_years are not normally distributed

Anderson-Darling normality test

data: employer\_experience\_years

A = 2.3894, p-value = 2.514e-06

Decision: p-value= 2.514e-06 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: employer\_experience\_years

D = 0.36281, p-value = 8.105e-07

Decision: p-value= 8.105e-07 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: employer\_experience\_years

W = 0.6855, p-value = 5.443e-05

Decision: p-value = 5.443e-05 < α=0.05 => Reject H0 at 5% significant level

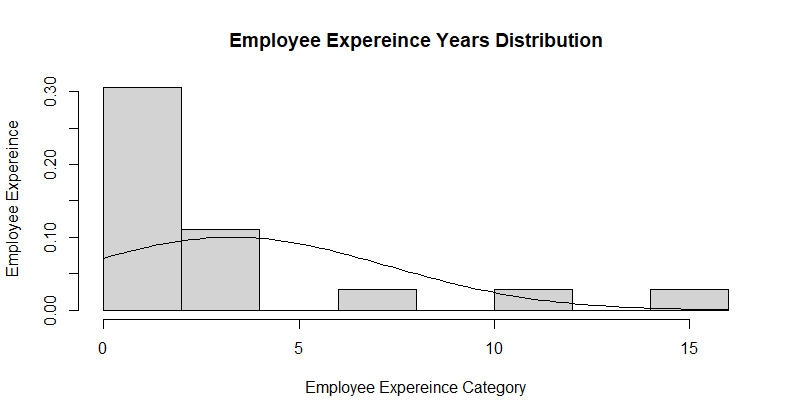
Conclusion - According to the Anderson darling ( p = 2.514e-06), Lillefor (p = 8.105e-07), and shapiro wiki (p = 5.443e-05) tests employer\_experience\_years data is not normally distributed at a 5 % significant level.

summary(employer\_experience\_years)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 1.125 2.000 3.294 3.000 15.000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 3.294 > 2.000). which means there is a larger amount of right skewedness available in the this “job\_title” data. According to this summary statistics and above mentioned tests, “total\_experience\_years” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a larger amount of right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = signing\_bonus

(\*) Whether or not signing\_bonus are normally distributed?

H­0: signing\_bonus are normally distributed

H­1: signing\_bonus are not normally distributed

Anderson-Darling normality test

data: signing\_bonus

A = 4.1905, p-value = 6.651e-11

Decision: p-value= 6.651e-11 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: signing\_bonus

D = 0.46929, p-value = 8.565e-12

Decision: p-value= 8.565e-12< α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: signing\_bonus

W = 0.53288, p-value = 1.567e-06

Decision: p-value = 1.567e-06 < α=0.05 => Reject H0 at 5% significant level

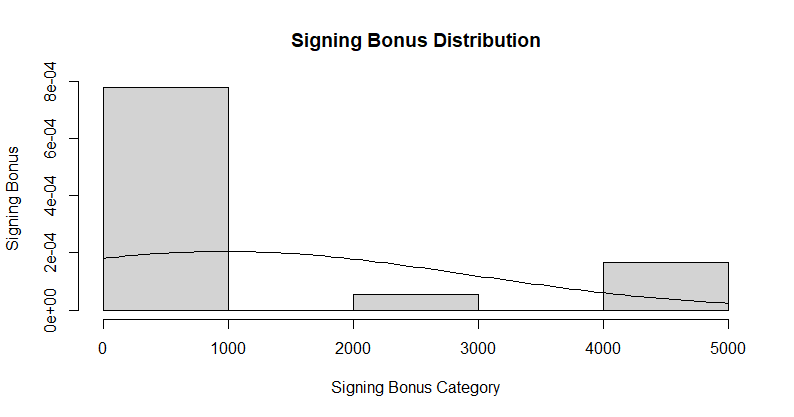
Conclusion - According to the Anderson darling ( p = 6.651e-11), Lillefor (p = 8.565e-12), and shapiro wiki (p = 1.567e-06) tests signing\_bonus data is not normally distributed at a 5 % significant level.

summary(signing\_bonus)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0 0.0 0.0 972.2 0.0 5000.0

When considering the above summary statistics mean value is greater than median value (Mean > Median = 972.2> 0.0). which means there is a larger amount of right skewedness available in the this “signing\_bonus” data. According to this summary statistics and above mentioned tests, “signing\_bonus” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a larger amount of right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = annual\_bonus

(\*) Whether or not annual\_bonus are normally distributed?

H­0: annual\_bonus are normally distributed

H­1: annual\_bonus are not normally distributed

Anderson-Darling normality test

data: annual\_bonus

A = 2.4085, p-value = 2.246e-06

Decision: p-value= 2.246e-06 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: annual\_bonus

D = 0.30545, p-value = 9.869e-05

Decision: p-value= 9.869e-05 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: annual\_bonus

W = 0.69786, p-value = 7.541e-05

Decision: p-value = 7.541e-05 < α=0.05 => Reject H0 at 5% significant level

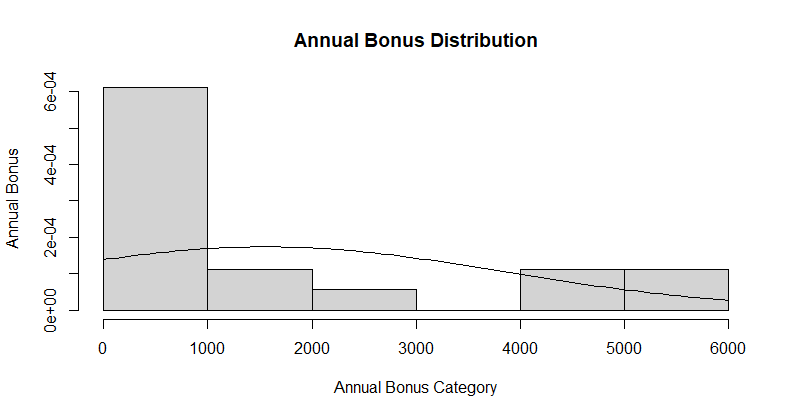
Conclusion - According to the Anderson darling ( p = 2.246e-06), Lillefor (p = 9.869e-05), and shapiro wiki (p = 7.541e-05) tests annual\_bonus data is not normally distributed at a 5 % significant level.

summary(annual\_bonus)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 0 5 1556 2250 6000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 1556 > 5). which means there is a larger amount of right skewedness available in the this “annual\_bonus” data. According to this summary statistics and above mentioned tests, “total\_experience\_years” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a larger amount of right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = location\_state

(\*) Whether or not location\_state are normally distributed?

H­0: location\_state are normally distributed

H­1: location\_state are not normally distributed

Anderson-Darling normality test

data: location\_state

A = 0.71901, p-value = 0.04978

Decision: p-value= 0.04978 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: location\_state

D = 0.16789, p-value = 0.1976

Decision: p-value= 0.1976 > α=0.05 => Accept H0 at 5% significant level

Shapiro-Wilk normality test

data: location\_state

W = 0.89439, p-value = 0.04592

Decision: p-value = 0.04592 < α=0.05 => Reject H0 at 5% significant level

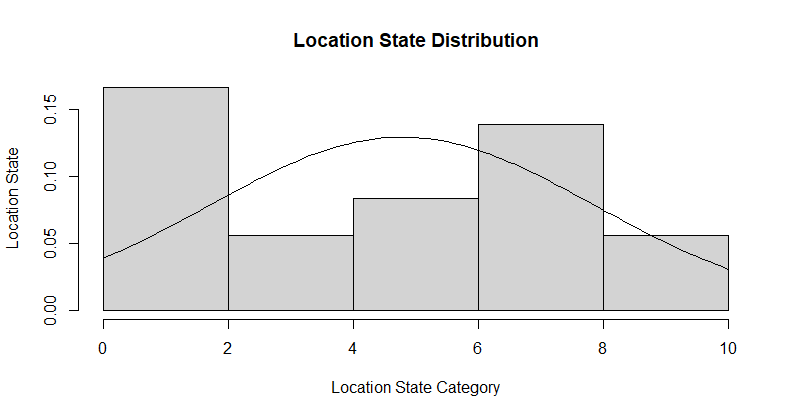
Conclusion - According to the Anderson darling ( p = 0.04978), Lillefor (p = 0.1976), and shapiro wiki (p = 0.04592) tests annual\_bonus data is not normally distributed at a 5 % significant level. Because above test Anderson darling, shapiro wiki test reject null hypothesis, but lillefor test accept the null hypothesis. In this out of that three testing two test (Anderson darling, lillefor testing) are rejected null hypothesis. According to this can be considered as “location\_state” data can be considered as not normally distributed

summary(location\_state)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.250 5.500 4.778 7.000 10.000

When considering the above summary statistics mean value is greater than median value (Mean < Median = 4.778 > 5.500). which means there is a small amount of left skewedness available in the this “location\_state” data. According to this summary statistics and above mentioned tests, “location\_state” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a smaller amount of left skewedness available. Which means data is not normally distributed at 5 % significant level.

**Statistical Hypothetical Testing**: **Normality tests for Samples ; Dataset : Salaries new**

Significant level = 0.05 (5%)

Confidence level = 0.95 (95%)

Variable = location\_country

(\*) Whether or not location\_country are normally distributed?

H­0: location\_country are normally distributed

H­1: location\_country are not normally distributed

Anderson-Darling normality test

data: location\_country

A = 4.6117, p-value = 5.764e-12

Decision: p-value= 5.764e-12 < α=0.05 => Reject H0 at 5% significant level

Lilliefors (Kolmogorov-Smirnov) normality test

data: location\_country

D = 0.48756, p-value = 8.612e-13

Decision: p-value= 8612e-13 < α=0.05 => Reject H0 at 5% significant level

Shapiro-Wilk normality test

data: location\_country

W = 0.46876, p-value = 4.344e-07

Decision: p-value = 4.344e-07 < α=0.05 => Reject H0 at 5% significant level

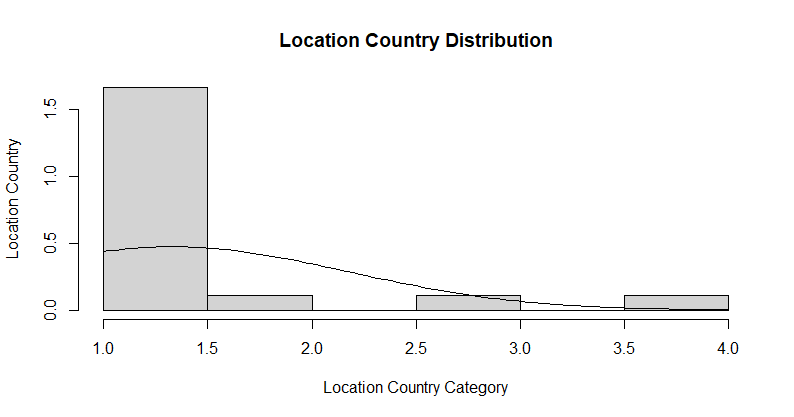
Conclusion - According to the Anderson darling ( p =5.764e-12), Lillefor (p = 8.612e-13), and shapiro wiki (p = 4.344e-07) tests location\_country data is not normally distributed at a 5 % significant level.

summary(location\_country)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 1.000 1.000 1.333 1.000 4.000

When considering the above summary statistics mean value is greater than median value (Mean > Median = 1.333 > 1.000). which means there is a huge amount of right skewedness available in the this “location\_country” data. According to this summary statistics and above mentioned tests, “location\_country” data can be considered as not normally distributed.



**Interpretation of montercarlo graph**

According to the above bell curve there is a larger amount of right skewedness available. Which means data is not normally distributed at 5 % significant level.

**Conclusions of Normality testing**

According to the conclusion of above mentioned normality testings, “annual\_base\_pay” and “job\_title” is normally distributed and other variables are not normally distributed.

**Correlational Analysis**

Correlational analysis means identifying relationship available between independent(x) variables and dependent (y) variable. This correlational analysis can be don pair wise. Which means dependent variable with one independent variable (Miksza & Elpus, Correlational design and analysis 2018). This analysis is necessary part for doing the regression analysis. Which means to develop the best model. Need to identify the what are the most significant variables that affected to dependent variable. For identifying this variables,

Correlational analysis has two main methods

1.Pearson method

After completing the normality testing. that test results are shown as “data is normally distributed”. which means most suitable method is Pearson method.

2.Spearman method

After completing the normality testing, if that test results are shown as “data is not normally distributed ” most suitable method is Spearman method. And also for doing the normality testing convert the text oriented data into numeric values (type conversion). For doing the correlational analysis based on this converted data “spearman” method is the best method. According to this “salary\_new” dataset “job\_title”, “Job\_title\_category”, “location\_state”, and “location\_country” data columns of salaries dataset converted into the numeric (type conversion) values and created the new dataset called “salary new”. and as well as according to the above mentioned normality testing, most of the test results are shown as “data is not normally distributed”. Consequently for this “salary\_new” dataset, doing the correlational analysis spearman method is the best method.

**Correlational coefficiency spectrum**

According to above mentioned tests, there is a correlational coefficiency spectrum is available to identify what are kind of variables having strongest, moderate (intermediate), and weak relationship with dependent variable.

Normally according to the correlational coefficiency spectrum, there is two sides, first one is positive, second one is negative

0 means there is not correlational coefficiency

When we consider the positive side

From 0 to + 0.25 means there is positive (direct) weak correlational coefficiency

From + 0.25 to + 0.75 means there is positive (direct) moderate (intermediate) correlational coefficiency

From + 0.75 to 1 means there is positive (direct) strong correlational coefficiency

1 means positive (direct) perfect correlational coefficiency.

When we consider the negative side

From 0 to -0.25 means there is negative (indirect) weak correlational coefficiency

From -0.25 to -0.75 means there is negative (indirect) moderate (intermediate) correlational coefficiency

From -0.75 to -1 means there is negative (indirect)strong correlational coefficiency

-1 means negative (indirect) perfect correlational coefficiency.

According to above mentioned correlational coefficiency spectrum, for developing the model only consider the positive and negative strong, moderate correlational coefficiency. Because based on this strong and moderate correlational coefficiency, best predictive model can be created, otherwise develop model predictions have higher chances to incorrect.

**Statistical Hypothetical Testing: Spearman test -for correlation analysis of Samples: Salary\_new dataset**

Sample field: annual\_base\_pay, location\_latitude

Significant level (α) = 0.05

Confidence level =95%

Let ῤ =the true population correlation coefficient between location\_latitude and employee annual base salary

(\*) Whether or not there is a correlation found between professors’ rank and salary?

H0: There is no correlation (ῤ =0) between Professors’ rank and professor’s salary

H1: There is a correlation (ῤ ≠0) between Professors’ rank and professor’s salary

Spearman's rank correlation rho

data: annual\_base\_pay and location\_latitude

S = 1194.5, p-value = 0.3528

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

-0.2326966

**References**

Chen, J. (no date) *Normal distribution: What it is, uses, and Formula*, *Investopedia*. Available at: https://www.investopedia.com/terms/n/normaldistribution.asp (Accessed: 12 May 2024).

Miksza, P. and Elpus, K. (2018) ‘Correlational design and analysis’, *Oxford Scholarship Online* [Preprint]. doi:10.1093/oso/9780199391905.003.0006.