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

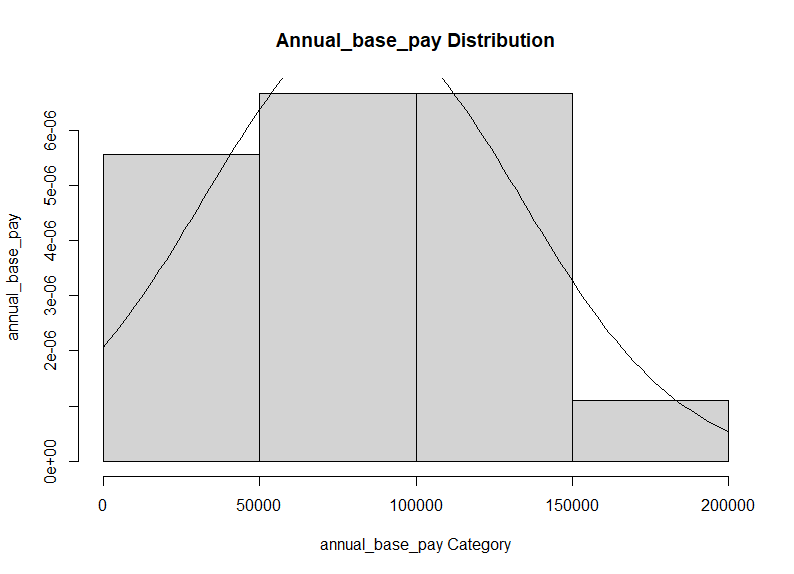
Conclusion - According to the Anderson darling (p = 0.7346), Lillefors (p = 0.8924), and Shapiro-Wilk testings, at a 5% significant level “annual\_base\_pay” data is normally distributed.

> 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

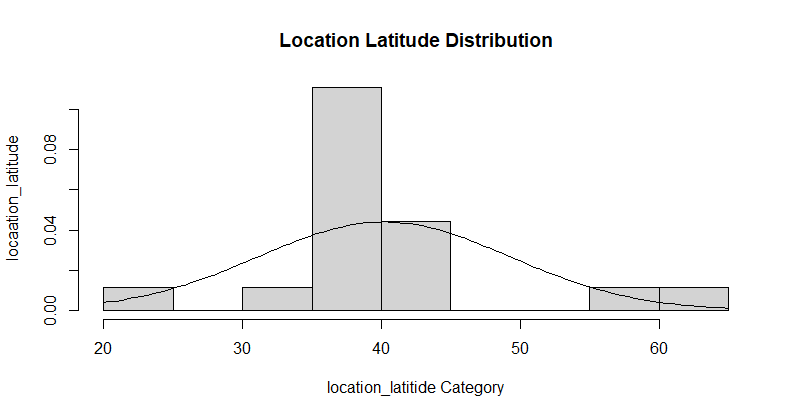
Conclusion - According to the Anderson darling ( p = 9.585e-05), Lillefor (p = 0.001824), and shapiro wiki (p =0.001652) tests, location\_latitude data is not normally distributed at a 5 % significant level.

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

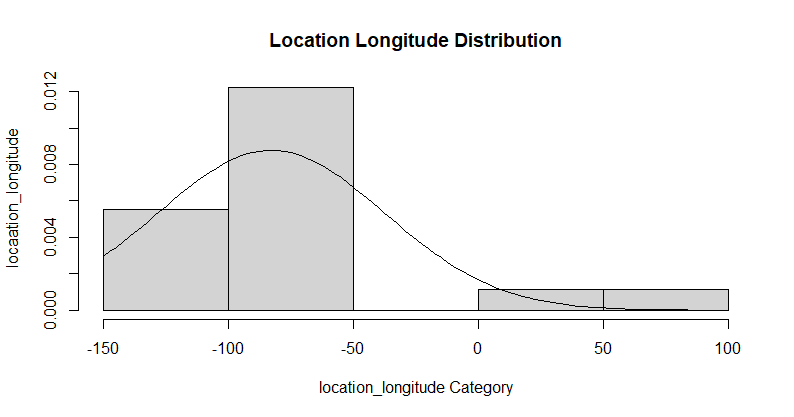
Conclusion - According to the Anderson darling ( p = 2.955e-07), Lillefor (p = 2.644e-05), and shapiro wiki (p = 1.596e-05) tests location\_longitude data is not normally distributed at a 5 % significant level.

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

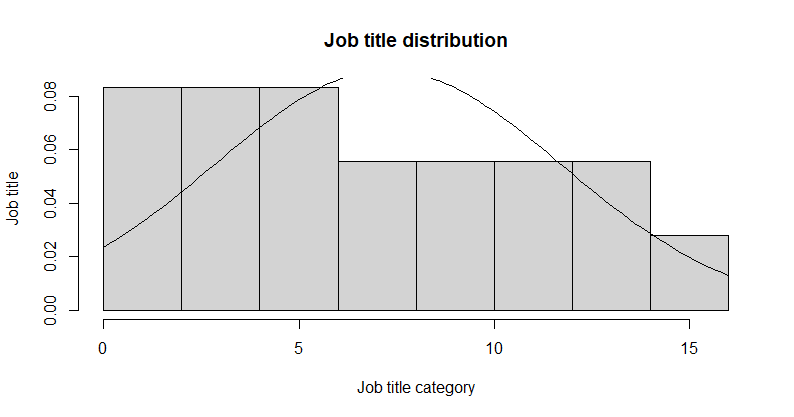
Decision: p-value = 0.3584 > α=0.05 => Accept H0 at 5% significant level

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

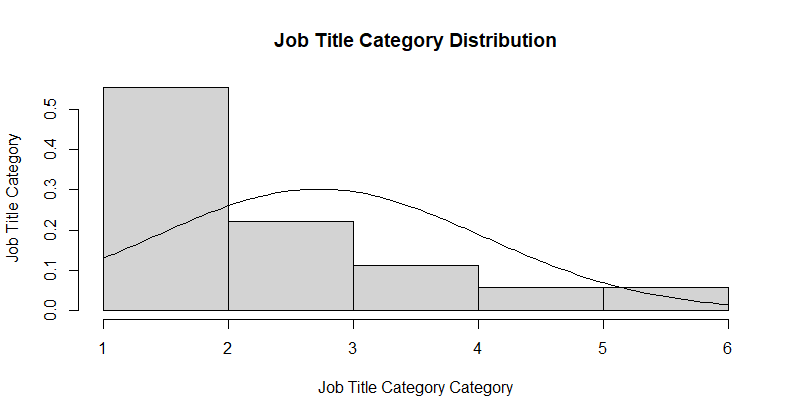
Conclusion - According to the Anderson darling ( p = 0.006516), Lillefor (p = 0.001809), and shapiro wiki (p = 0.01706) tests job\_title\_category data is normally distributed at a 5 % significant level.

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

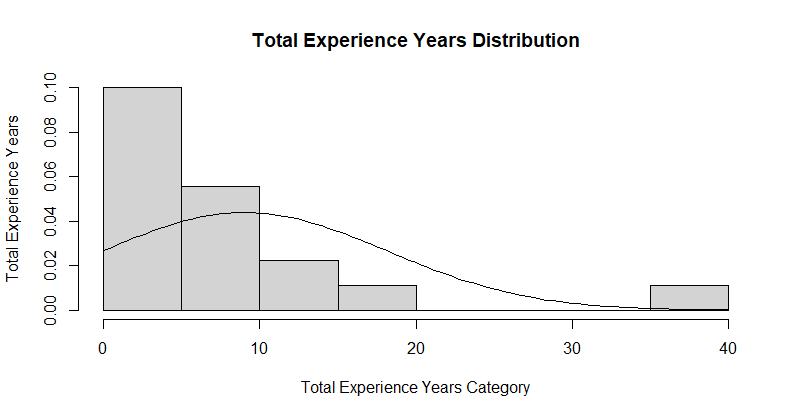
Conclusion - According to the Anderson darling ( p = 7.621e-05), Lillefor (p = 0.001318), and shapiro wiki (p = 0.0001038) tests total\_experience\_years data is not normally distributed at a 5 % significant level.

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

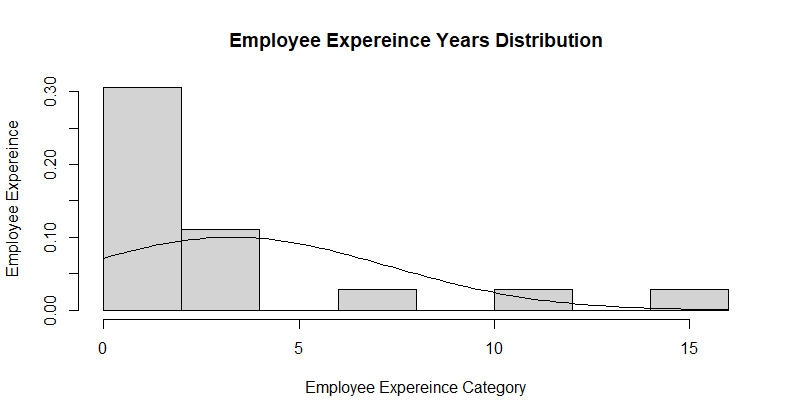
Conclusion - According to the Anderson darling ( p = 2.514e-06), Lillefor (p = 8.105e-07), and shapiro wiki (p = 5.443e-05) tests employer\_experience\_years data is not normally distributed at a 5 % significant level.

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

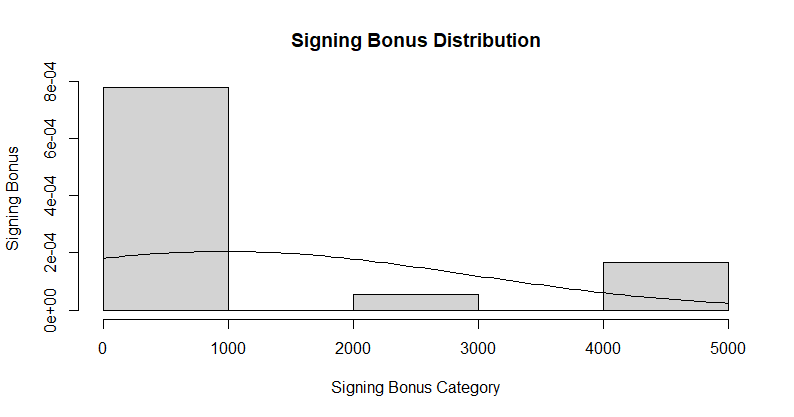
Conclusion - According to the Anderson darling ( p = 6.651e-11), Lillefor (p = 8.565e-12), and shapiro wiki (p = 1.567e-06) tests signing\_bonus data is not normally distributed at a 5 % significant level.

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

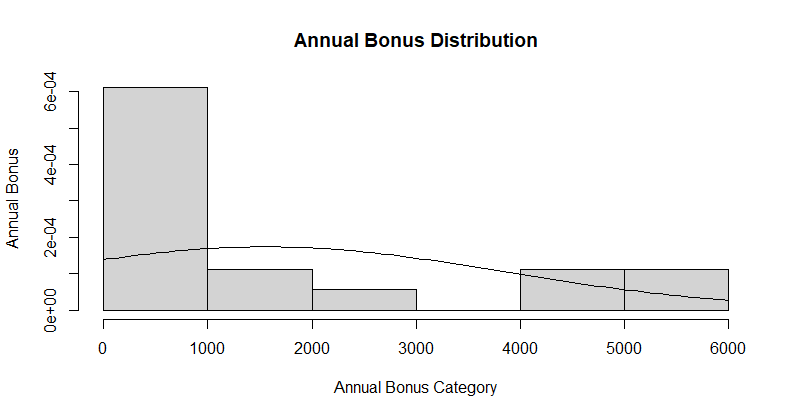
Conclusion - According to the Anderson darling ( p = 2.246e-06), Lillefor (p = 9.869e-05), and shapiro wiki (p = 7.541e-05) tests annual\_bonus data is not normally distributed at a 5 % significant level.

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

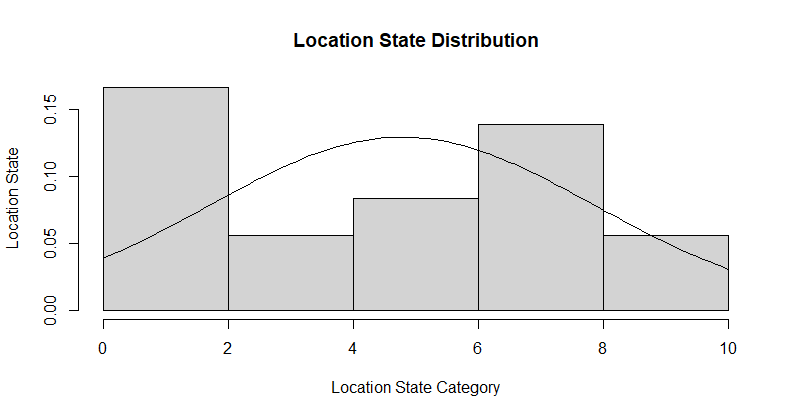
Conclusion - According to the Anderson darling ( p = 0.04978), Lillefor (p = 0.1976), and shapiro wiki (p = 0.04592) tests annual\_bonus data is not normally distributed at a 5 % significant level. Because above test Anderson darling, shapiro wiki test reject null hypothesis, but lillefor test accept 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

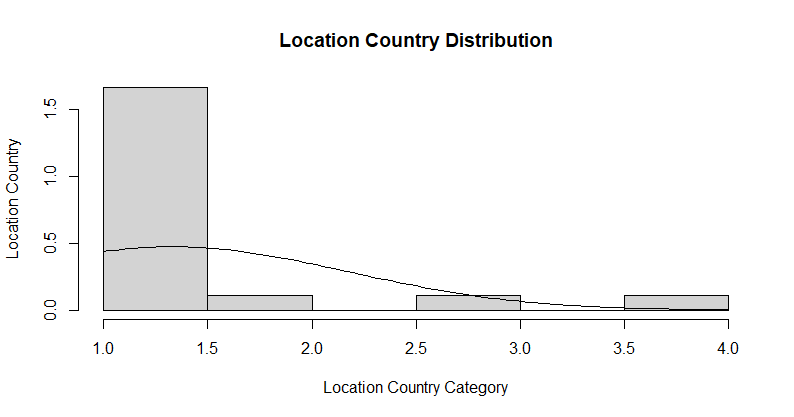
Conclusion - According to the Anderson darling ( p =5.764e-12), Lillefor (p = 8.612e-13), and shapiro wiki (p = 4.344e-07) tests location\_country data is not normally distributed at a 5 % significant level.

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\_pay

(\*) Whether or not there is a correlation found between location\_latitude and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between location\_latitude and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between location\_latitude and employee annual\_base\_pay

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:

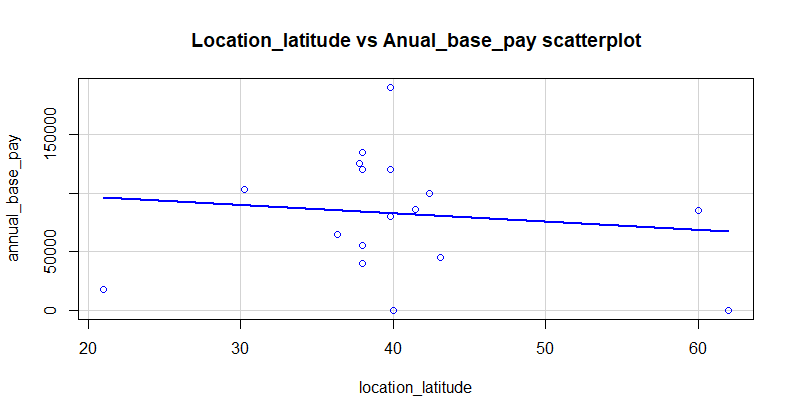
rho

-0.2326966

Decision: p-value = 0.3528 < α = 0.05 => Reject H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.3528) for correlation analysis.There is a statistical evidence to claim that a weak negative (indirect) correlation (ῤ = -0.23) exists between location\_latitude and employee annual\_base\_pay at a 5% significance level. Based on this result it is not required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = -0.23) available between employee annual\_base\_pay and location\_latitude at a 5% significance level.

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

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

Significant level (α) = 0.05

Confidence level =95%

Let ῤ =the true population correlation coefficient between location\_longitude and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between location\_longitude and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between location\_longitude and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between location\_longitude and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and location\_longitude

S = 1670.7, p-value = 0.0006788

alternative hypothesis: true rho is not equal to 0

sample estimates:

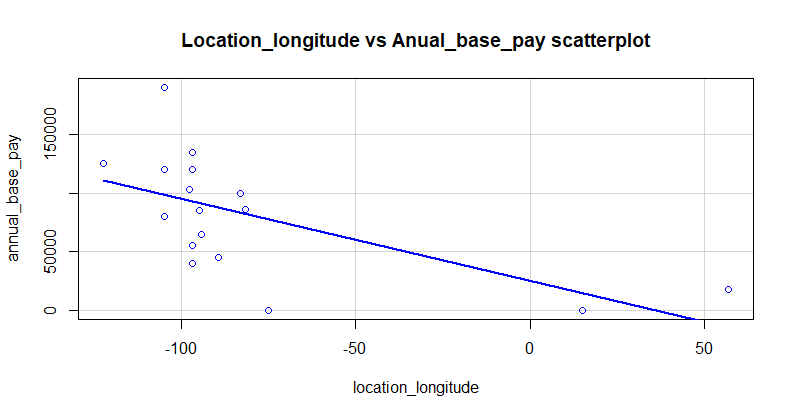
rho

-0.7241186

Decision: p-value = 0.0006788 < α = 0.05 => Reject H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.0006788) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) negative (indirect) correlation (ῤ = -0.72) exists between location\_longitude and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also lot of data points are located nearly absolute line. Therefore, it is confirmed that there is a negative (indirect) moderate (intermediate) relationship (ῤ = -0.72) available between employee annual\_base\_pay and location\_longitude at a 5% significance level.

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

Sample field: annual\_base\_pay, job\_title

Significant level (α) = 0.05

Confidence level =95%

Let ῤ =the true population correlation coefficient between job\_title and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between job\_title and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between job\_title and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between job\_title and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and job\_title

S = 1178.5, p-value = 0.3888

alternative hypothesis: true rho is not equal to 0

sample estimates:

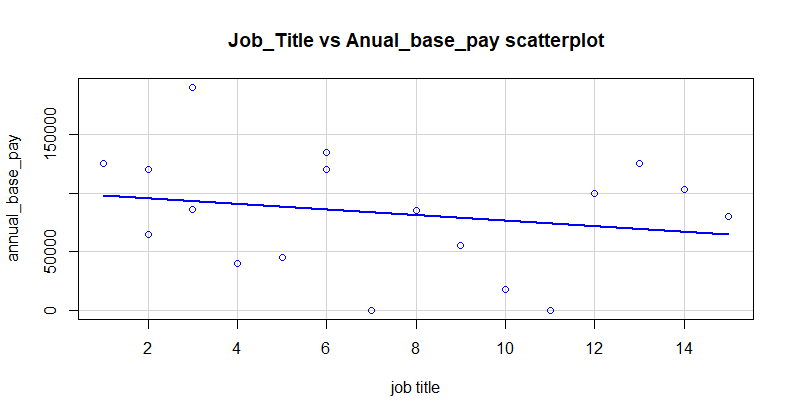
rho

-0.2162442

Decision: p-value = 0.3888 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.3888) for correlation analysis.There is a statistical evidence to claim that a weak negative (indirect) correlation (ῤ = -0.22) exists between job\_title and employee annual\_base\_pay at a 5% significance level. Based on this result it is not required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also lot of data points are located far away from absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = -0.22) available between employee annual\_base\_pay and job\_title at a 5% significance level.

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

Sample field: annual\_base\_pay, job\_title\_category

Significant level (α) = 0.05

Confidence level =95%

Let ῤ =the true population correlation coefficient between job\_title\_category and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between job\_title\_category and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between job\_title\_category and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between job\_title\_category and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and job\_title\_category

S = 1260.1, p-value = 0.2258

alternative hypothesis: true rho is not equal to 0

sample estimates:

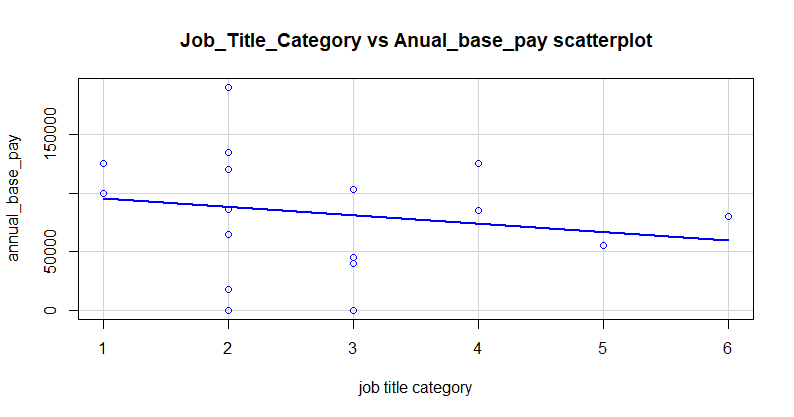
rho

-0.3003922

Decision: p-value = 0.2258 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.2258) for correlation analysis.There is a statistical evidence to claim that a weak negative (indirect) correlation (ῤ = -0.30) exists between job\_title\_category and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also lot of data points are located far away from absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = -0.30) available between employee annual\_base\_pay and job\_title\_category at a 5% significance level.

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

Sample field: annual\_base\_pay, total\_experience\_years

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between total\_experience\_years and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between total\_experience\_years and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between total\_experience\_years and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between total\_experience\_years and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and total\_experience\_years

S = 425.39, p-value = 0.01543

alternative hypothesis: true rho is not equal to 0

sample estimates:

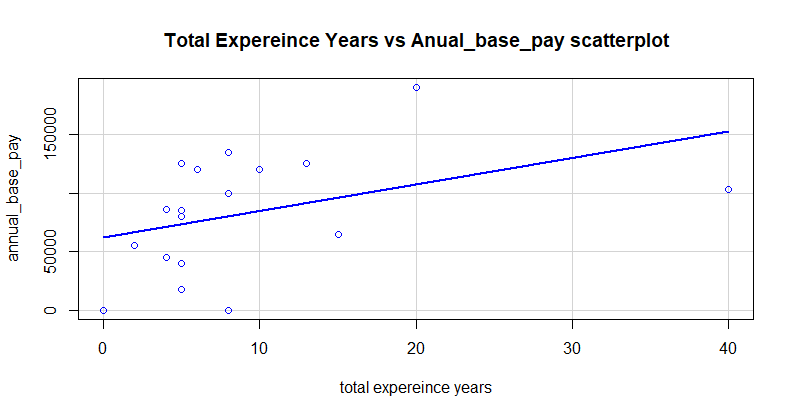
rho

0.5610035

Decision: p-value = 0.2258 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.01543) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) positive (direct) correlation (ῤ = 0.56) exists between total\_experience\_years and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a positive (direct) trend. And also lot of data points are located nearly absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = 0.56) available between employee annual\_base\_pay and total\_experience\_years at a 5% significance level.

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

Sample field: annual\_base\_pay, employer\_experience\_years

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between employer\_experience\_years and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between employer\_experience\_years and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between employer\_experience\_years and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between employer\_experience\_years and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and employer\_experience\_years

S = 717.55, p-value = 0.2984

alternative hypothesis: true rho is not equal to 0

sample estimates:

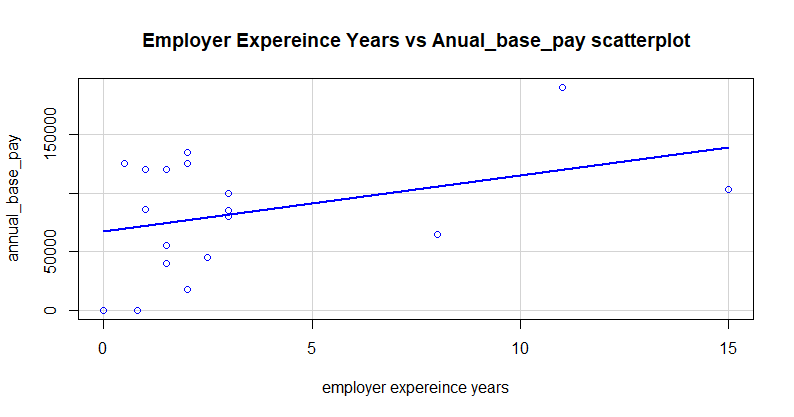
rho

0.2594946

Decision: p-value = 0.2984 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.2984) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) positive (direct) correlation (ῤ = 0.26) exists between employer\_experience\_years and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a positive (direct) trend. And also some data points are located nearly absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = 0.26) available between employee annual\_base\_pay and total\_experience\_years at a 5% significance level.

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

Sample field: annual\_base\_pay, employer\_experience\_years

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between employer\_experience\_years and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between employer\_experience\_years and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between employer\_experience\_years and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between employer\_experience\_years and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and employer\_experience\_years

S = 717.55, p-value = 0.2984

alternative hypothesis: true rho is not equal to 0

sample estimates:

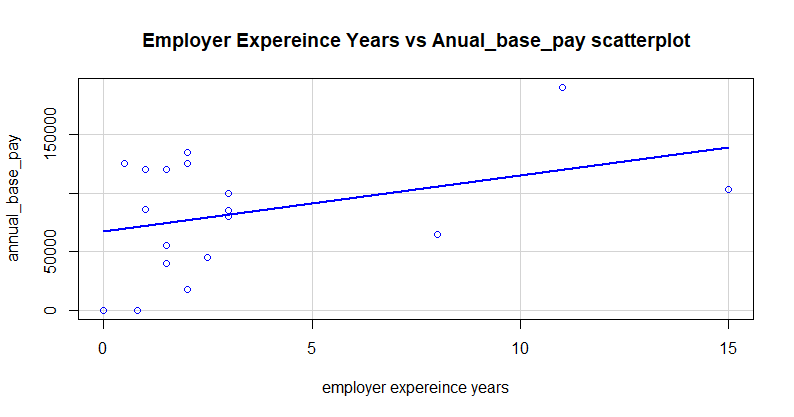
rho

0.2594946

Decision: p-value = 0.2984 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.2984) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) positive (direct) correlation (ῤ = 0.26) exists between employer\_experience\_years and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a positive (direct) trend. And also some data points are located nearly absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = 0.26) available between employee annual\_base\_pay and total\_experience\_years at a 5% significance level.

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

Sample field: annual\_base\_pay, signing\_bonus

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between signing\_bonus and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between signing\_bonus and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between signing\_bonus and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between signing\_bonus and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and signing\_bonus

S = 693.03, p-value = 0.252

alternative hypothesis: true rho is not equal to 0

sample estimates:

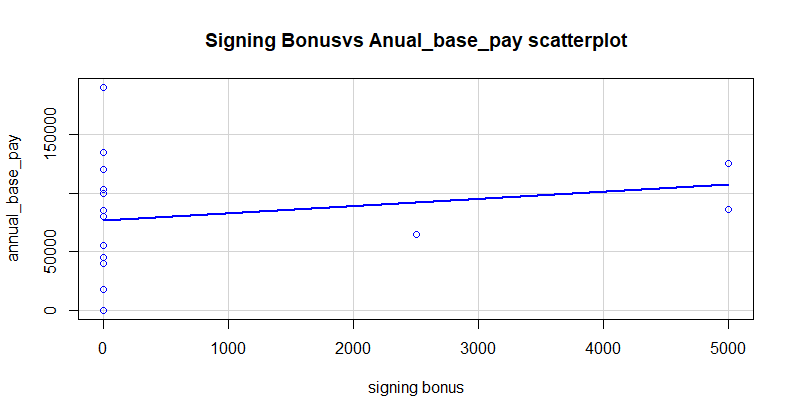
rho

0.2847945

Decision: p-value = 0.252 >α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.252) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) positive (direct) correlation (ῤ = 0.28) exists between signing\_bonus and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a positive (direct) trend. And also some data points are located nearly absolute line. Therefore, it is confirmed that there is a positive (direct) moderate relationship (ῤ = 0.28) available between employee annual\_base\_pay and signing\_bonus at a 5% significance level.

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

Sample field: annual\_base\_pay, annual\_bonus

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between annual\_bonus and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between annual\_bonus and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between annual\_bonus and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between annual\_bonus and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and annual\_bonus

S = 983.46, p-value = 0.9531

alternative hypothesis: true rho is not equal to 0

sample estimates:

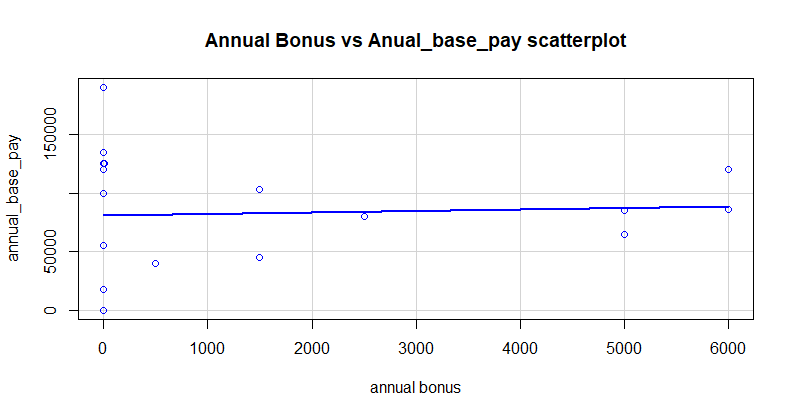
rho

-0.01492572

Decision: p-value = 0.9531 > α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.9531) for correlation analysis.There is a statistical evidence to claim that a weak negative (indirect) correlation (ῤ = -0.01) exists between annual\_bonus and employee annual\_base\_pay at a 5% significance level. Based on this result it is not required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also lot of data points are located far away from absolute line. Therefore, it is confirmed that there is a negative (indirect) weak relationship (ῤ = -0.01) available between employee annual\_base\_pay and annual\_bonus at a 5% significance level.

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

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

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between location\_state and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between location\_state and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between location\_state and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between location\_state and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and location\_state

S = 815.37, p-value = 0.5298

alternative hypothesis: true rho is not equal to 0

sample estimates:

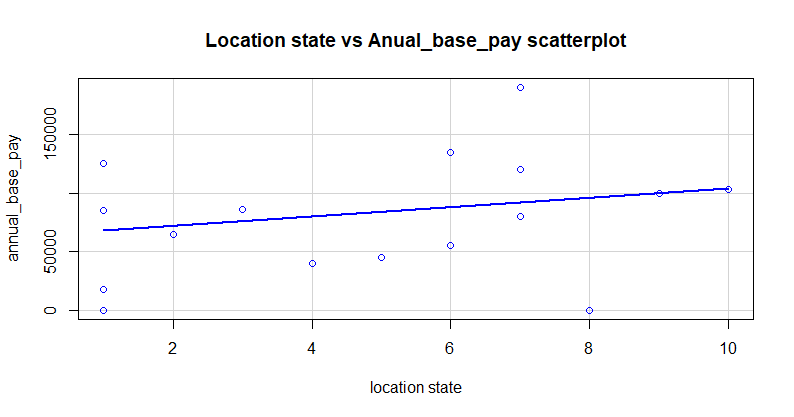
rho

0.1585486

Decision: p-value = 0.5298> α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.5298) for correlation analysis.There is a statistical evidence to claim that a weak positive (direct) correlation (ῤ = 0.16) exists between location\_state and employee annual\_base\_pay at a 5% significance level. Based on this result it is not required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also lot of data points are located far away from absolute line. Therefore, it is confirmed that there is a positive (direct) weak relationship (ῤ = 0.16) available between employee annual\_base\_pay and location\_state at a 5% significance level.

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

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

Significant level (α) = 0.05

Confidence level = 95%

Let ῤ =the true population correlation coefficient between location\_country and employee annual\_base\_pay

(\*) Whether or not there is a correlation found between location\_country and employee annual\_base\_pay ?

H0: There is no correlation (ῤ =0) between location\_country and employee annual\_base\_pay

H1: There is a correlation (ῤ ≠0) between location\_country and employee annual\_base\_pay

Spearman's rank correlation rho

data: annual\_base\_pay and location\_country

S = 1374.2, p-value = 0.08416

alternative hypothesis: true rho is not equal to 0

sample estimates:

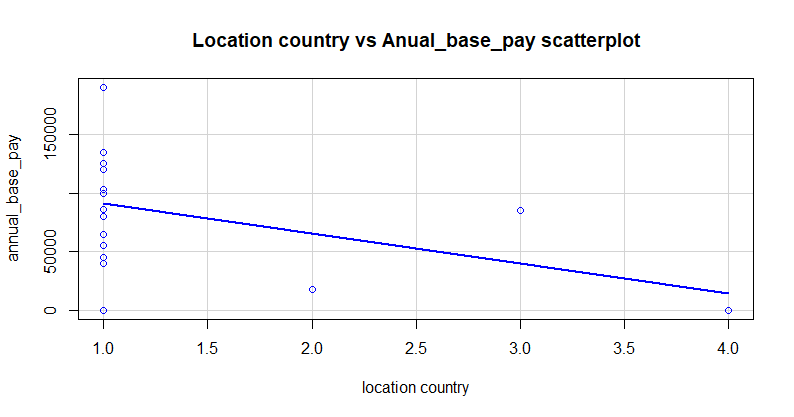
rho

-0.4181969

Decision: p-value = 0.08416> α = 0.05 => Accept H0 at 5% significant level

Correlation Test Conclusion

According to the spearman rank correlation-tests (p = 0.08416) for correlation analysis.There is a statistical evidence to claim that a moderate (intermediate) negative (indirect) correlation (ῤ = -0.42) exists between location\_country and employee annual\_base\_pay at a 5% significance level. Based on this result it is required to do the regression analysis to find the model of the relationship.

****

According to the above mentioned scatter plot, it is clearly identifiable that scatter plot values position according to a negative (indirect) trend. And also some data points are located nearly absolute line. Therefore, it is confirmed that there is a negative (indirect) moderate (intermediate) relationship (ῤ = -0.42) available between employee annual\_base\_pay and location\_country at a 5% significance level.

**Regression Analysis**

Regression analysis used for developing the predictive model. When it comes to the regression analysis, there are two main types

1. Simple linear regression analysis
2. Multi linear regression analysis

Simple linear regression analysis

Simple linear regression analysis means that model comprises of dependent variable (y) and one independent variable (x).

Multi linear regression analysis

Multi linear regression analysis means that model comprise of dependent variable and more than one independent variables (x).

When comparing the above regression models, the best model is Multi linear model. Because when it comes to the real world problems that problems are associated with multiple factors. For example this employee salary analysis, employee salary depends on multiple factors like state, annual bonus etc. In that case Multi linear model is the best predictive model.

But when considering the Multi linear model that model has multiple independent variables (x). In that case, in between that independent variables (x) of that model has some correlational coefficiency. this correlational coefficiency can be strong, moderate (intermediate), weak. If there is a higher, moderate correlational coefficiency available between independent variables (x). that directly affected to the dependent variable.

This issue is called Multi corr linearity effect.

Spearman's rank correlation rho

data: location\_longitude and location\_latitude

S = 621.62, p-value = 0.1441

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.3584906

According to the above correlational coefficiency there is a moderate (intermediate) coefficiency is available between that independent variables. This is the Multi corr linearity effect. But Multi linear regression analysis is the best model.

**Multi Linear regression Analysis**

Call:

lm(formula = annual\_base\_pay ~ location\_longitude + total\_experience\_years +

employer\_experience\_years + signing\_bonus + location\_country,

model = TRUE)

Residuals:

Min 1Q Median 3Q Max

-60677 -28965 2043 23461 75809

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19399.195 43299.505 0.448 0.6621

location\_longitude -581.645 302.130 -1.925 0.0782 .

total\_experience\_years 1724.265 2961.073 0.582 0.5711

employer\_experience\_years 147.752 6933.203 0.021 0.9833

signing\_bonus 2.456 5.932 0.414 0.6862

location\_country -2449.471 16130.372 -0.152 0.8818

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42770 on 12 degrees of freedom

Multiple R-squared: 0.4968, Adjusted R-squared: 0.2872

F-statistic: 2.37 on 5 and 12 DF, p-value: 0.1025

According to the above mentioned multiple linear general model developed by multiple regression analysis, it is clear to identify that when “employee\_experience\_years” proportionately increase by 147.752 units , “signing\_bonus” increases by 2.456 units, “total\_experience\_years” increased by 1724.265 units, “location\_country” inversely increases by 2449.471 units, “location\_longitude” inversely increases by 581.645 in proportion to that employee salary increases by one unit. Here the employee’ annual\_base\_pay calculation is further influenced by the constant value 19399.195 . Therefore, the model clearly indicates that employee’ annual\_base\_pay are determined by ‘employee\_experience\_years’, signing\_bonus, total\_experience\_years, location\_country, and location\_longitude at a 5% significance level.

According to the standard error values and t-values of independent variables, it is confirmed that location\_country (Std.Error:16130.372) is the most crucial factor to be considered in annual\_base\_pay determination of a employee. The employee\_experience\_years (Std.Error:6933.203) can be considered as the second most important factor in salary calculation. Total\_experience\_years (Std.Error:2961.073) is the third important factor. Location\_longitude (Std.Error : 302.130) is the fourth important factor. Signing\_bonus (5.932) is the least important factor.This model accuracy becomes higher as R2= 0.4968 and adjusted R2 = 0.2872 but the model residuals indicate high values. When we consider about the other models developed, this model can be confirmed as the best model. Therefore, much better annual\_base\_pay predications can be identified with this model.

**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.