Statistical Modeling and Simulation Lab Report

One-way ANOVA Modeling on Payroll Data Set

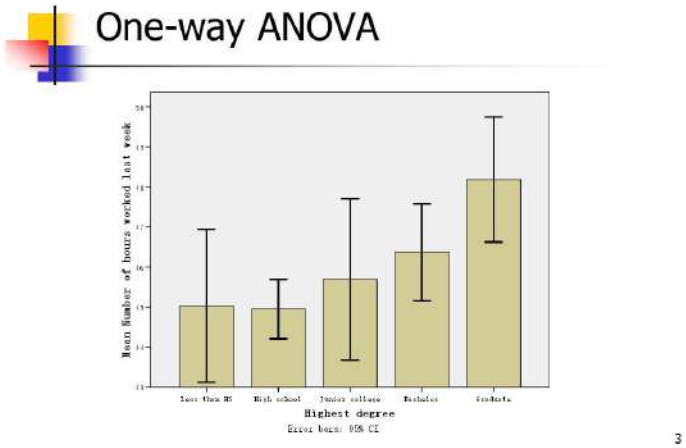
Edera Venkata Naga Sai Bharadwaja
SC20M104
Machine Learning and Computing
Indian Institute of Space Science and Technology

Table of Contents

List of Figures

1. One-way ANOVA

- **Why One-way ANOVA ?**
- In statistics , one-way analysis of variance is a technique that can be used to compare means of two or more samples (using the F distribution).
- **Why it is called One-way ANOVA ?**
- This technique can be used only for numerical response data, the ”Y”, usually one variable, and numerical or categorical input data, the ”X”, always one variable, hence ”one-way”.

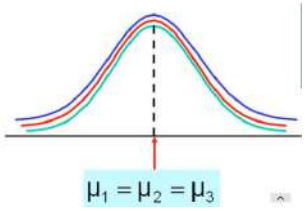


An ANOVA test is a way to find out if survey or experiment results are significant. In other words, they help you to figure out if you need to reject the null hypothesis or accept the alternate hypothesis. Basically, you’re testing groups to see if there’s a difference between them. Examples of when you might want to test different groups:

- • A group of psychiatric patients are trying three different therapies: counselling, medication and biofeedback. You want to see if one therapy is better than the others.
- • A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
- • Students from different colleges take the same exam. You want to see if one college outperforms the other

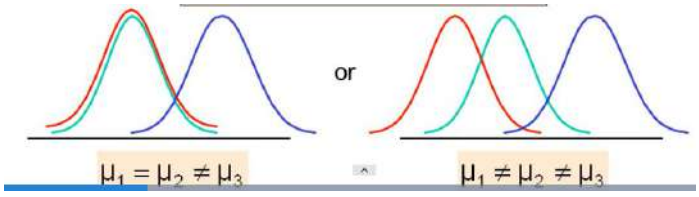
The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. The one-way ANOVA compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypothesis:

- samples drawn from populations with the same mean values.
- $H_o : \mu_1 = \mu_2 = ... = \mu_n$



Alternative hypothesis (H_1)**:**

- samples drawn from populations with different mean values.
- $H_1 : \mu_1 \neq \mu_2 \neq .... \neq \mu_n$



where μ = group mean and n = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (H_1), which is that there are at least two group means that are statistically significantly different from each other.

2. How do we decide True Hypothesis ?

- The ANOVA tests the null hypothesis, which states that samples in all groups are drawn from populations with the same mean values.
- The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples.
- If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples.
- A higher ratio therefore implies that the samples were drawn from populations with different mean values.

3. Assumptions in One-way ANOVA:

1. Populations are Normally distributed
2. Populations have equal variances
3. Samples are randomly and independently drawn

Interpret ANOVA test

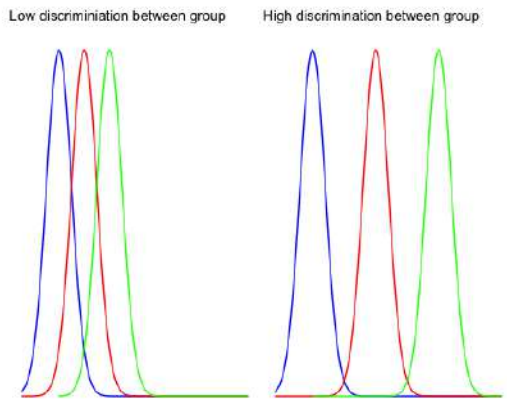
- The F-statistic is used to test if the data are from significantly different populations, i.e., different sample means.

- To compute the F-statistic, we need to divide the between-group variability over the within-group variability.

- $F = \frac{\frac{BSS}{k-1}}{\frac{WSS}{n-k}}$

4. Between-group variability

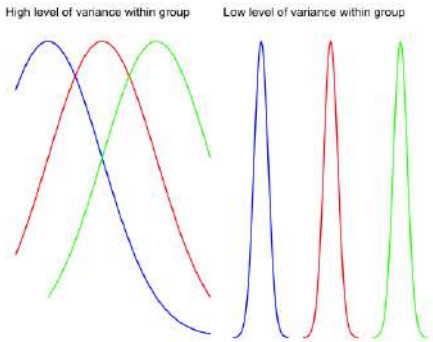
- The between-group variability reflects the differences between the groups inside all of the population.



- The left graph shows very little variation between the three group, and it is very likely that the three means tends to the overall mean (i.e., mean for the three groups).
- The right graph plots three distributions far apart, and none of them overlap. There is a high chance the difference between the total mean and the groups mean will be large.

5. Within-group variability

- The within group variability considers the difference between the groups. The variation comes from the individual observations; some points might be totally different than the group means. The within group variability picks up this effect and refer to the sampling error.



- The left part plots the distribution of three different groups. We increased the spread of each sample and it is clear the individual variance is large. The F-test will decrease, it means , tend to accept the null hypothesis
- The right part shows exactly the same samples (identical mean) but with lower variability. It leads to an increase of the F-test and tends in favor of the alternative hypothesis.

frame **F-Statistics**

- Both measures are used to construct the F-statistics. It is very intuitive to understand the F-statistic. If the numerator increases, it means the between-group variability is high, and it is likely the groups in the sample are drawn from completely different distributions.

- In other words, a low F-statistic indicates little or no significant difference between the group's average.

6. Overview of Payroll Dataset

The Los Angeles City Controller Office releases payroll information for all city employees on a quarterly basis since 2013.

- Data points or Rows = 285008
- Attributes or Columns = 35

Relevant Attributes

1. Job titles
2. annual pay
3. base pay
4. year

Electrician Data

```
data_elec = data[data["Job Class Title" ]== "Electrician"]
```

```
data_elec.shape
```

```
(424, 35)
```

```
data_elec
```

	Row ID	Year	Department Title	Payroll Department	Record Number	Job Class Title	Employment Type	Hourly or Event Rate	Projected Annual Salary	Q1 Payments	...	MOU Title	FMS Department	Job Class
816	70039	2014	General Services	3364.0	2876743409	Electrician	Full Time	$41.33	$86296.78	$0.00	...	BUILDING,TRADES UNIT	40	3863
1073	1362	2013	Airports (LAWA)	101.0	130825756	Electrician	Full Time	$39.23	$81912.24	$20460.86	...	BUILDING,TRADES UNIT	4	3863
3183	94216	2014	Recreation And Parks	7901.0	3578725017	Electrician	Full Time	$41.33	$86296.78	$22832.64	...	BUILDING,TRADES UNIT	88	3863
3485	104189	2014	Water And Power (DWP)	NaN	3898637226	Electrician	Full Time	NaN	$99848.16	$37209.78	...	NaN	98	3863
3712	3335	2013	Airports (LAWA)	101.0	888072334	Electrician	Full Time	$38.55	$80496.05	$18654.84	...	BUILDING,TRADES UNIT	4	3863
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
271318	271489	2016	Recreation And Parks	7904.0	2797991048	Electrician	Full Time	$41.42	$86484.96	$23940.79	...	Building Trades	88	3863
273907	274092	2016	Zoo	8701.0	1984844873	Electrician	Full Time	$41.42	$86484.96	$23630.11	...	Building Trades	87	3863

```
data.shape
```

```
(285008, 35)
```

```
data.head()
```

	Row ID	Year	Department Title	Payroll Department	Record Number	Job Class Title	Employment Type	Hourly or Event Rate	Projected Annual Salary	Q1 Payments	...	MOU Title	FMS Department	Job Class	Pay Grade
0	111391	2014	Water And Power (DWP)	NaN	1412316577	Commercial Service Representative	Full Time	NaN	$70386.48	$16129.89	...	NaN	98	1230	NaN
1	31732	2013	Police (LAPD)	4301.0	432728338	Police Officer I	Full Time	$25.12	$52450.56	$11331.00	...	POLICE OFFICERS UNIT	70	2214	A
2	27697	2013	Police (LAPD)	4301.0	97182508	Police Officer II	Full Time	$42.77	$89303.78	$20036.32	...	POLICE OFFICERS UNIT	70	2214	2
3	14136	2013	Harbor (Port of LA)	3201.0	950136941	Senior Security Officer	Full Time	$28.75	$60028.96	$15793.88	...	SUPV BLUE COLLAR	42	3184	0
4	91896	2014	Public Works - Sanitation	7024.0	3230003445	Senior Clerk Typist	Full Time	$30.92	$64553.13	$14700.00	...	CLERICAL UNIT	82	1368	0

7. Definition of System under study

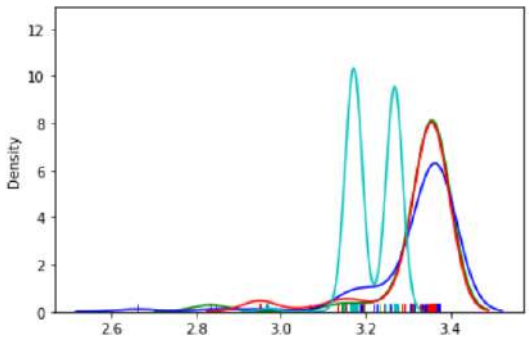
To determine the Mean salary of the particular occupation from the Payrol Data set over the consecutive years is Same or Not

Null Hypothesis $H_o : \mu_{y_1} = \mu_{y_2} = ... = \mu_{y_p}$
Alternative Hypothesis $H_1 : \mu_{y_1} \neq \mu_{y_2} \neq .... \neq \mu_{y_p}$

8. Mathematical/Statistical Modelling of the systems

Assumptions in One-Way ANOVA

1. Populations are Normally distributed
2. Populations have equal variances
3. Samples are randomly and independently drawn



One-Way ANOVA model best suits to Solve our Objective

- General Linear Model : $X = \beta * A^T + \varepsilon$
- The One-way ANOVA Model : $X_{ij} = \mu_i + \varepsilon_{ij}$
- Comparing One-way ANOVA model with General Linear Model
- $X = (x_{11}, x_{12}, ...x_{1n_1}, x_{21}, x_{22}, ...x_{2n_2}.....x_{k1}, x_{k2}, ...x_{kn_k})$
- Unknown parameters $\beta = (\mu_1, \mu_2, ...\mu_k)$
- $A_{n*k} = \begin{pmatrix} I_{n1}^T & 0 & . & . & . & 0^T \\ 0^T & I_{n2}^T & . & . & . & 0^T \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0^T & . & . & . & . & I_{nk}^T \end{pmatrix}$
- $\varepsilon = (\varepsilon_{11}, \varepsilon_{12}, ...\varepsilon_{1n_1}, \varepsilon_{21}, \varepsilon_{22}, ...\varepsilon_{2n_2}, ..., \varepsilon_{k1}, \varepsilon_{k2}, ...\varepsilon_{kn_k})$

8.1 ML Estimates $\hat{\mu}$ and $\hat{\sigma}^2$

- Unknown Parameters $\beta = (\mu, \sigma^2)$
- The joint PDF of X is
- $f(X : \mu, \sigma^2) = [\frac{1}{2\pi\sigma^2}]^{\frac{n}{2}} * e^{\frac{-1}{2\sigma^2} * [\sum_{i=1}^{k} \sum_{j=1}^{n_i} [x_{ij} - \mu_i]^2]}$
- Differentiating PDF with μ and σ and equating to Zero To find ML estimates $\hat{\mu}$ and $\hat{\sigma}^2$
- $\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} = \bar{x_{i.}}$, i = 1,2,....k
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} [x_{ij} - \bar{x_{i.}}]^2}{n}$

8.2 Under Null Hypothesis $H_o : \mu_1 = \mu_2 = ... = \mu_k$

- $H_0 : \beta * H^T = 0$
- $(\mu_1, \mu_2, ..., \mu_k) * \begin{pmatrix} 1 & -1 & 0 & . & . & 0 \\ 1 & 0 & . & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & . & . & . & . & -1 \end{pmatrix}^T = 0$
- $H : (k - 1 * k)$
- $A : (n * k)$
- By substituting the Null hypothesis condition in ML estimates we get
- $\hat{\hat{\mu}} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} [x_{ij}]^2}{n} = \bar{x}$
- $\hat{\hat{\sigma}}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} [x_{ij} - \bar{x_i}]^2}{n}$

8.3 Likelihood Ratio Test

Definition 1. For testing H_0 against H_1, a test of the form, reject H_0 if and only if $\lambda(\mathbf{x}) < c$, where c is some constant, and

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} f_{\boldsymbol{\theta}}(x_1, x_2, \ldots, x_n)}{\sup_{\boldsymbol{\theta}\in\Theta} f_{\boldsymbol{\theta}}(x_1, x_2, \ldots, x_n)},$$

is called a *generalized likelihood ratio* (GLR) test.

- By substituting $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\hat{\mu}}$ and $\hat{\hat{\sigma}}^2$ in the Generalized Likelihood ratio test , we get
- F = $\frac{\sum_{i=1}^{k}[n_i * [\bar{x_{i.}} - \bar{x}]^2]}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}[x_{ij} - \bar{x_{i.}}]^2]}$
- Numerator : Between Sum of Squares (BSS) : $\sum_{i=1}^{k}[n_i * [\bar{x_{i.}} - \bar{x}]^2]$
- Denominator : With in Sum of Squares (WSS) : $\sum_{i=1}^{k}\sum_{j=1}^{n_i}[x_{ij} - \bar{x_{i.}}]^2]$
- BSS follows Chi-Square Distribution with D.o.F = r = (k - 1)
- WSS follows Chi-Square Distribution with D.o.F = (n - k)
- Dividing by Numerator with (k-1) and Dividing Numerator with (n-k) then the ratio will be come a F - statistic
- $\frac{n-k}{k-1} * F \sim F$

9. Solution of the system: Statistical

- $\frac{n-k}{k-1} * F < F$
  - Accept Hypothesis H_0
- $\frac{n-k}{k-1} * F > F$
  - Reject Hypothesis H_0

10. Simulation :

10.1 Data cleaning and preprocessing

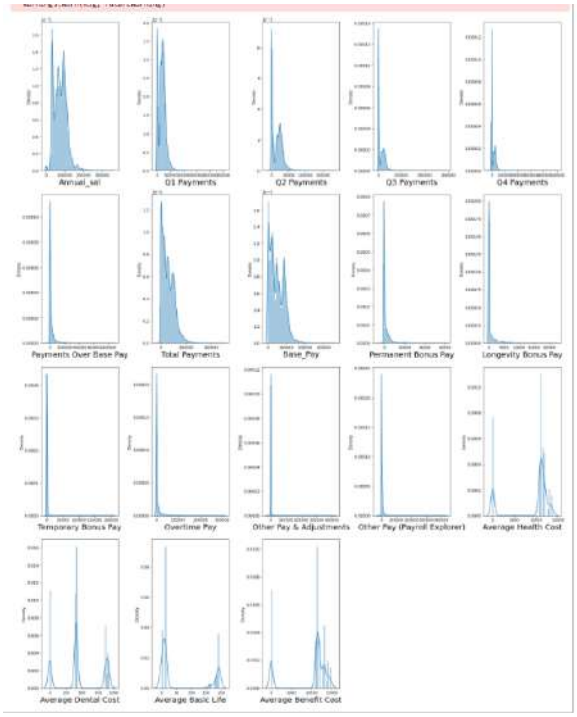
Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

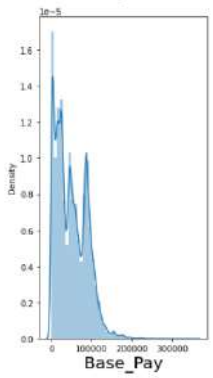
Data cleaning

```
payroll = payroll[pd.notnull(payroll['Payroll Department'])]
payroll.rename(columns={'Projected Annual Salary' : 'Annual_sal'}, inplace = True)
payroll.rename(columns={'Job Class Title' : 'Job_title'}, inplace = True)
payroll.rename(columns={'Base Pay' : 'Base_Pay'}, inplace = True)
payroll.replace(0,'NaN')
```

```
for i in ['Annual_sal','Q1 Payments','Q2 Payments','Q3 Payments','Q4 Payments','Payments Over Base Pay',
          'Total Payments','Base_Pay','Permanent Bonus Pay','Longevity Bonus Pay','Temporary Bonus Pay','Overtime Pay',
          'Other Pay & Adjustments','Other Pay (Payroll Explorer)','Average Health Cost','Average Dental Cost',
          'Average Basic Life','Average Benefit Cost']:
    payroll[i] = payroll[i].str.replace('$','')
```

10.2 Removing Outliers





```
payroll_2014_elec = payroll_2014[payroll_2014.Job_title == 'Electrician']
payroll_2015_elec = payroll_2015[payroll_2015.Job_title == 'Electrician']
payroll_2013_elec = payroll_2013[payroll_2013.Job_title == 'Electrician']
payroll_2016_elec = payroll_2016[payroll_2016.Job_title == 'Electrician']
```

```
q1 = payroll_2014_elec['Base_Pay'].quantile(0.90)
# we are removing the top 10% data from the Pregnancies column
payroll_2014_elec_cleaned =(payroll_2014_elec[payroll_2014_elec['Base_Pay']<q1])
skew2014= (np.array(np.sqrt(np.log(payroll_2014_elec_cleaned['Base_Pay']))))
q2 = payroll_2015_elec['Base_Pay'].quantile(0.97)
# we are removing the top 3% data from the Pregnancies column
payroll_2015_elec_cleaned =(payroll_2015_elec[payroll_2015_elec['Base_Pay']<q2])
skew2015= (np.array(np.sqrt(np.log(payroll_2015_elec_cleaned['Base_Pay']))))
q3 = payroll_2013_elec['Base_Pay'].quantile(0.98)
# we are removing the top 2% data from the Pregnancies column
payroll_2013_elec_cleaned =(payroll_2013_elec[payroll_2013_elec['Base_Pay']<q3])

skew2013= (np.array((np.sqrt(np.log(payroll_2013_elec_cleaned['Base_Pay'])))))

q4 = payroll_2016_elec['Base_Pay'].quantile(0.98)
# we are removing the top 2% data from the Pregnancies column
payroll_2016_elec_cleaned =(payroll_2016_elec[payroll_2016_elec['Base_Pay']<q4])

skew2016= (np.array((np.sqrt(np.log(payroll_2016_elec_cleaned['Base_Pay'])))))
```

10.3 Random samples from data

```
sam_1 = pd.DataFrame(skew2014).sample(frac=0.04)

#print("Sample Mean 2014 "+str(sample_elec_mean_2014))
sam_2 = pd.DataFrame(skew2015).sample(frac=0.06)

#print("Sample Mean 2015 "+str(sample_elec_mean_2015))
sam_3 =pd.DataFrame(skew2013).sample(frac=0.07)
```

11. Result

```
grand_mean= (x1_bar*n1 +  x2_bar*n2 + x3_bar*n3)/N
```

```
bss = n1*((x1_bar-grand_mean)**2) + n2*((x2_bar-grand_mean)**2) + n3*((x3_bar- grand_mean)**2)
print(bss)
```

```
0.011935942306065449
```

```
wss = x1_square_sum+ x2_square_sum + x3_square_sum
print(wss)
```

```
33.312766261754376
```

```
f = (bss/wss)* ((N-len(listN))/(len(listN)-1))
f
```

```
0.0019706463932636113
```

```
F_2_11 = 2.8595 # for p=0.05
```

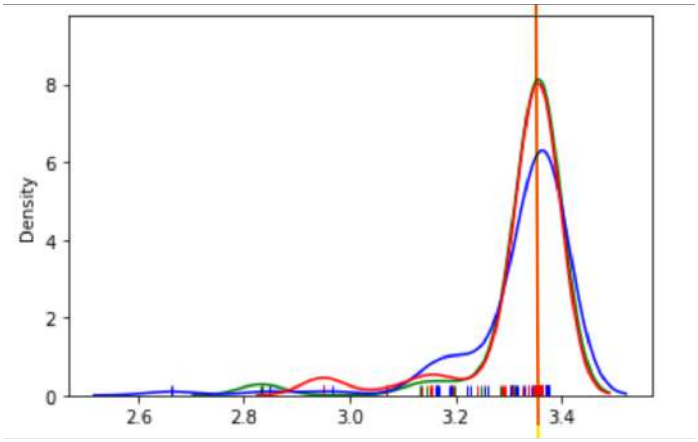
```
if f > F_2_16:
    print('reject the null hypothesis : average salary of 3 years of electricians is different with 5% significance level ')
else:
    print('accept  the null hypothesis : average salary of 3 years of electricians is same  with 5% significance level')
```

```
accept  the null hypothesis : average salary of 3 years of electricians is same  with 5% significance level
```

Accept the Hypothesis H_0 i,e mean salary of Electrician over three consecutive years is Same

11.1 frame

Comparing the Result




```
sam_1 : 49 , sam_2 : 50 , sam_3 : 27
Total Number of Sample from all the groups :  126
x1_bar is  3.327162905351475
x2_bar is 3.3078129770501095
x3_bar is  3.3179620691210734
bss :  0.00927288014820046
wss 33.02051198201417
F_calculated :  0.017270541699200587
accept null hypothesis:average salary of 3 years of electricians is same  with 1% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 5% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 10% significance level
```

```
sam_1 : 33 , sam_2 : 39 , sam_3 : 55
Total Number of Sample from all the groups :  127
x1_bar is  3.31755238232624
x2_bar is 3.315247823871921
x3_bar is  3.3209964184263723
bss :  0.0007814807462164151
wss 33.02603915436681
F_calculated :  0.001467078932443259
accept null hypothesis:average salary of 3 years of electricians is same  with 1% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 5% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 10% significance level
```

```
sam_1 : 56 , sam_2 : 64 , sam_3 : 60
Total Number of Sample from all the groups :  180
x1_bar is  3.332081563591583
x2_bar is 3.3286035497205235
x3_bar is  3.3269731754597687
bss :  0.0007846813367544047
wss 33.25111964786785
F_calculated :  0.002088479998213166
accept null hypothesis:average salary of 3 years of electricians is same  with 1% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 5% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 10% significance level
```

```
sam_1 : 15 , sam_2 : 2 , sam_3 : 57
Total Number of Sample from all the groups :  74
x1_bar is  3.35848926838013
x2_bar is 3.37101499077286
x3_bar is  3.3159655645148507
bss :  0.025624916273336393
wss 33.63881985888814
F_calculated :  0.027042700413376204
accept null hypothesis:average salary of 3 years of electricians is same  with 1% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 5% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 10% significance level
```

```
sam_1 : 10 , sam_2 : 1 , sam_3 : 37
Total Number of Sample from all the groups :  48
x1_bar is  3.3367386002402832
x2_bar is 3.227705294379683
x3_bar is  3.3346052964714588
bss :  0.011320584767813377
wss 32.67149843695563
F_calculated :  0.007796188404621 4396
accept null hypothesis:average salary of 3 years of electricians is same  with 1% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 5% significance level
accept null hypothesis:average salary of 3 years of electricians is same  with 10% significance level
```

12. Code for the Project

https://github.com/BharadwajEdera/Bharadwaj-Machine-Learning-and-computing/blob/main/Statistics/one%20way%20anova.ipynb

References

- https://www.scribbr.com/statistics/one-way-anova/
- https://en.wikipedia.org/wiki/One-way$_{a}nalysis_{o}f_{v}ariance$
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5296382/