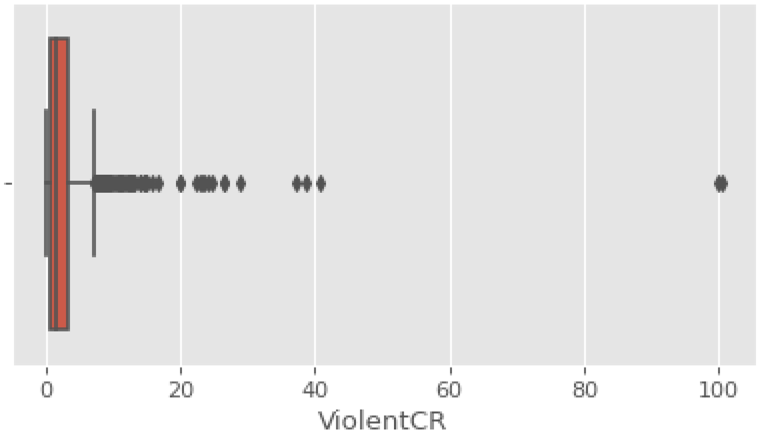
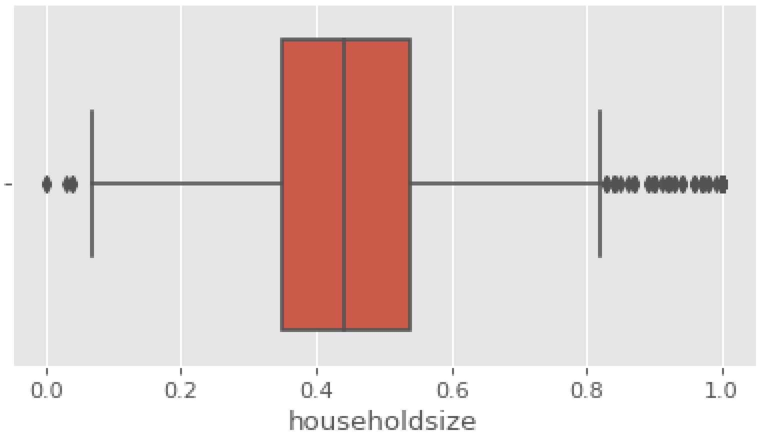
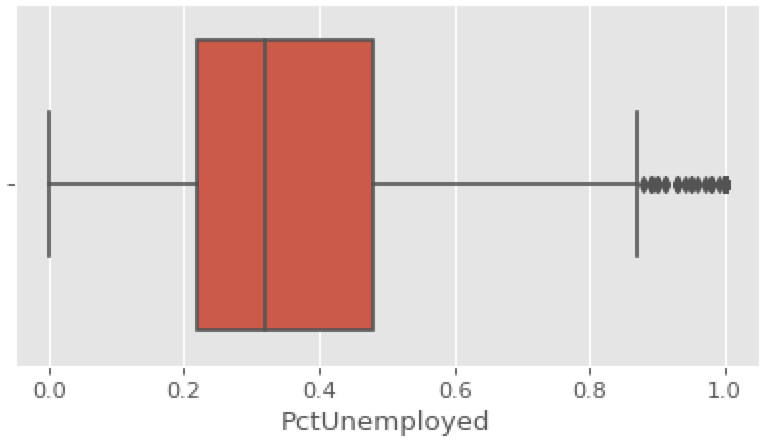
**NAME : AMMAAR AHMAD**
**ROLL NO: 1801CS08**
**CS575 - MIDSEM**

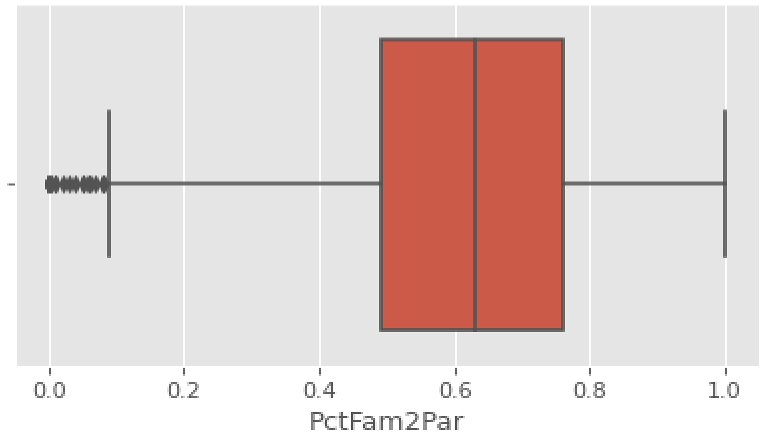
**Colab Link -**
**[https://colab.research.google.com/drive/1c45JEhUOLUJ8BpP04ca-efJkCzlCJlA8?usp=sharing](https://colab.research.google.com/drive/1c45JEhUOLUJ8BpP04ca-efJkCzlCJlA8?usp=sharing)**

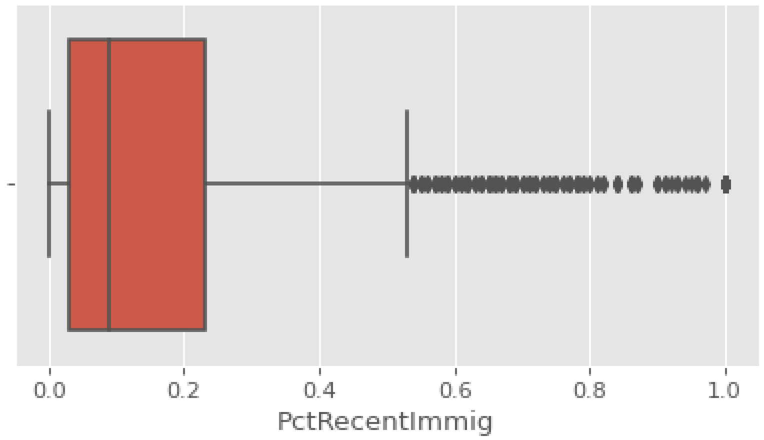
## 1. Box Plots

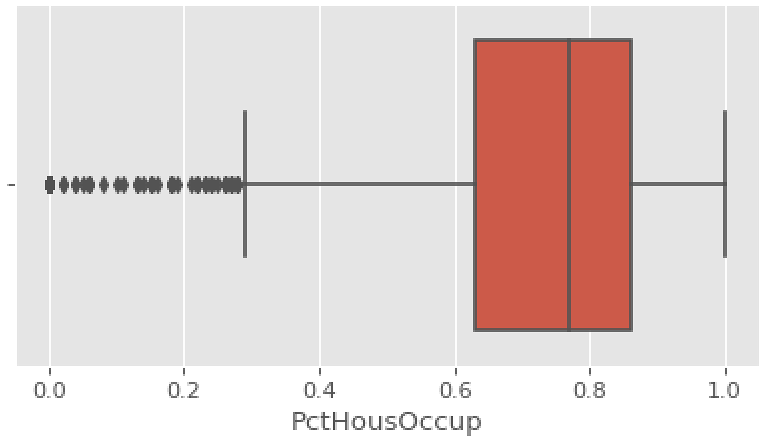


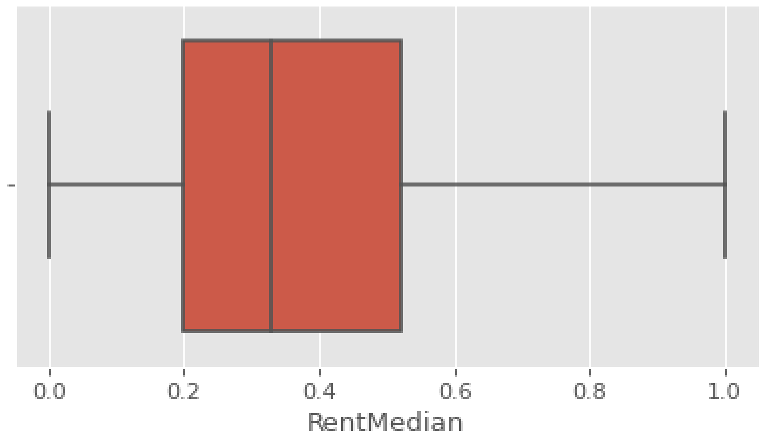


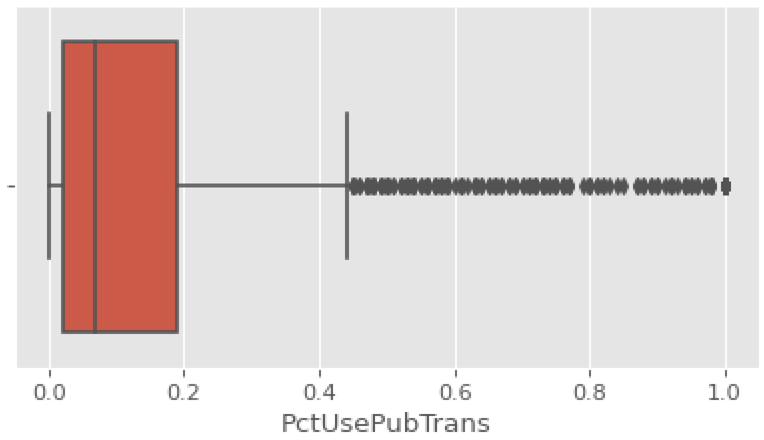

0.0
0.2
0.4
0.6
0.8
1.0
PctUnemployed


0.0
0.2
0.4
0.6
0.8
1.0
PctFam2Par

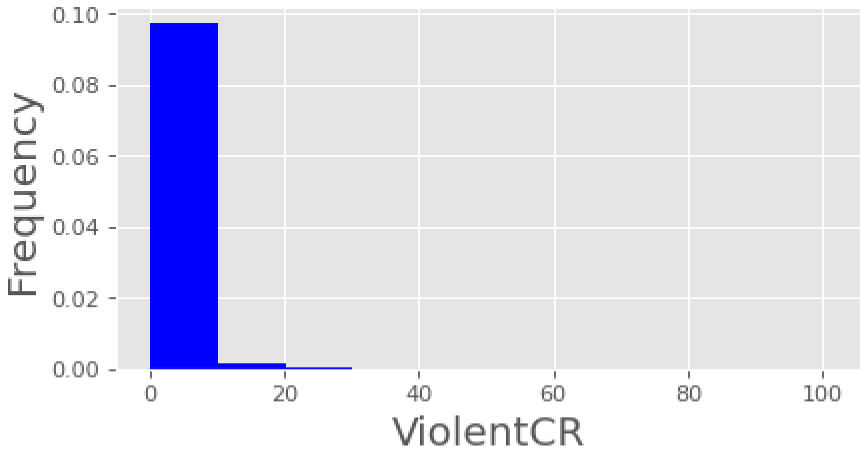

0.0
0.2
0.4
0.6
0.8
1.0
PctRecentImmig

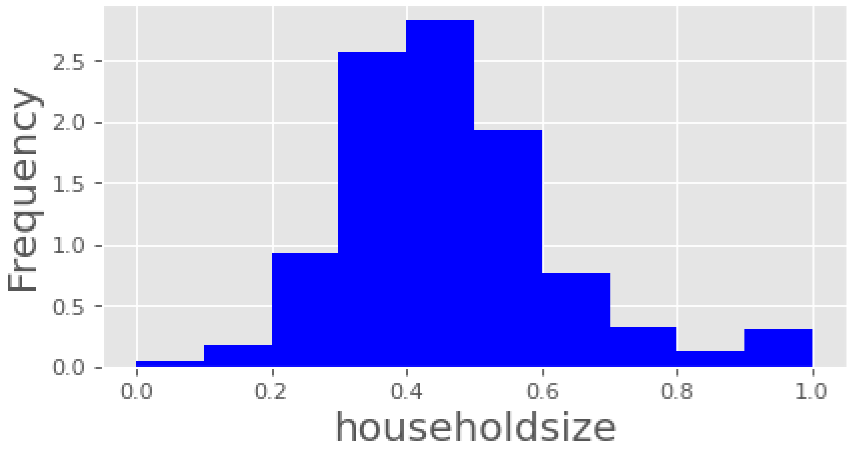


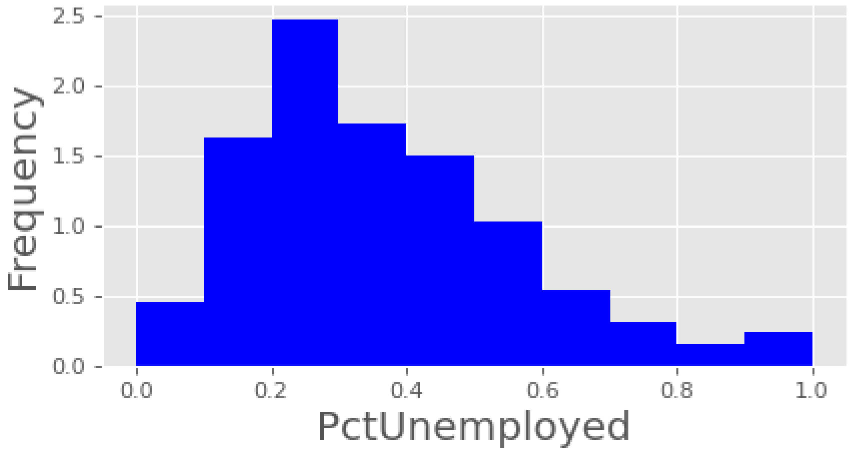


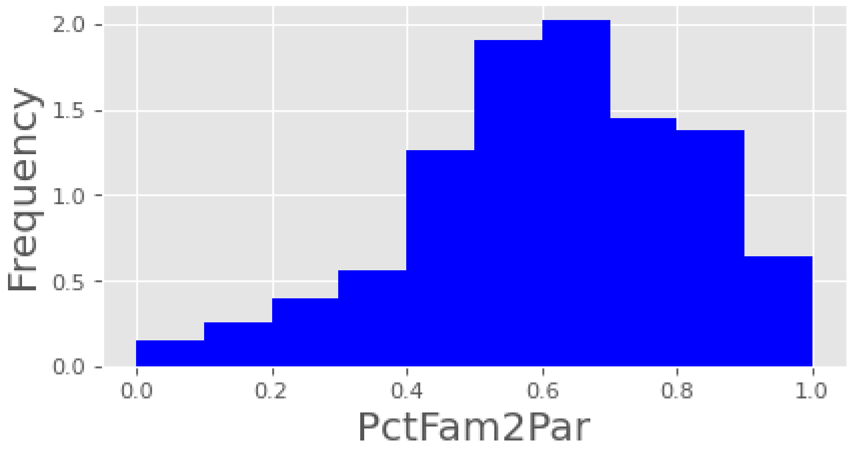


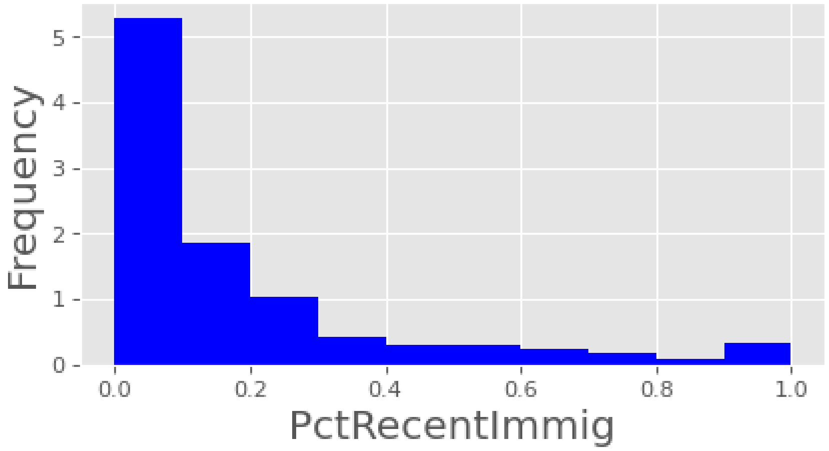
## Histogram Plots

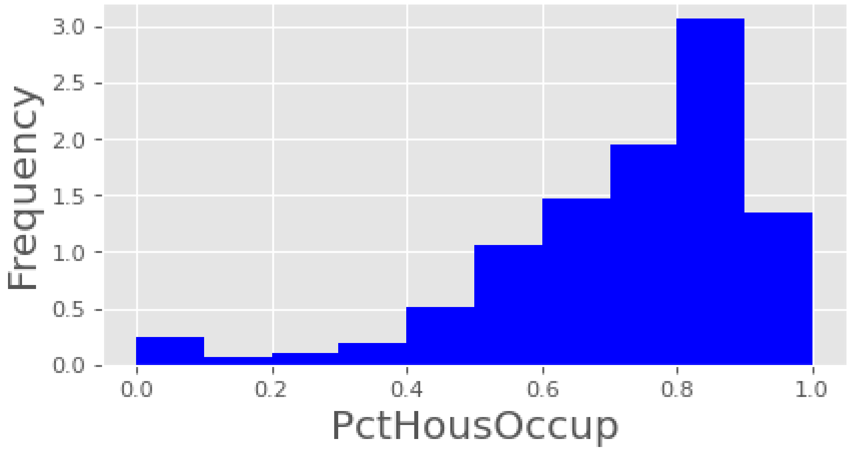

0.10
0.08
0.06
0.04
0.02
0.00
Frequency
0
20
40
60
80
100
ViolentCR

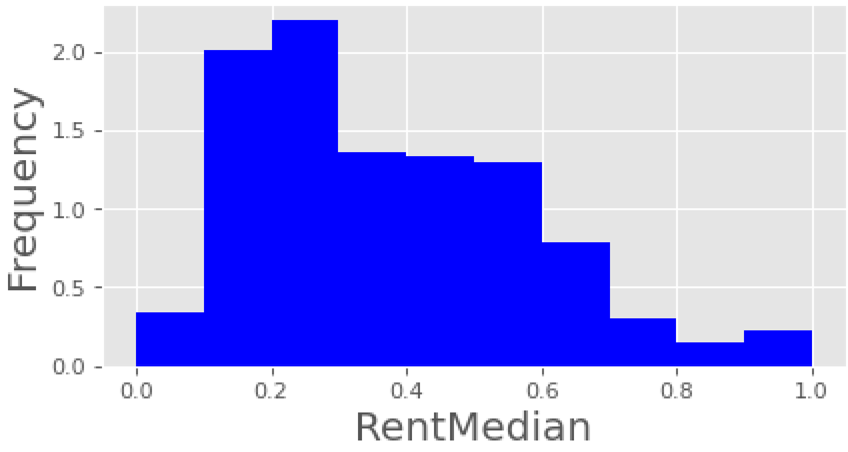

2.5
2.0
1.5
1.0
0.5
0.0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
householdsize

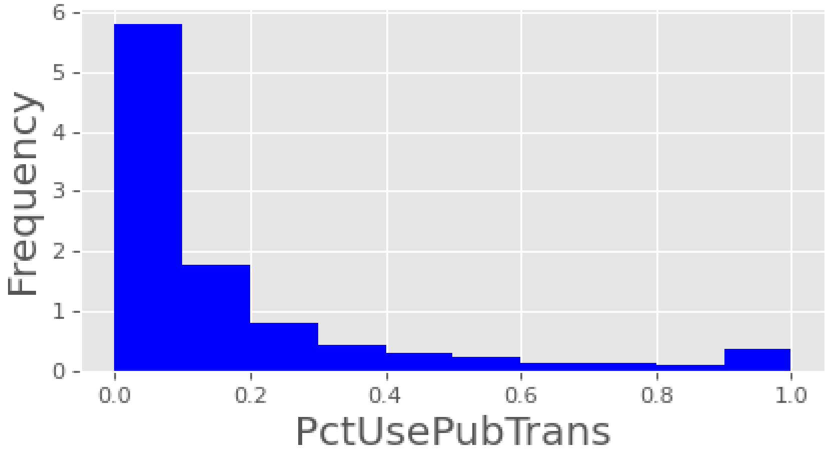

2.5
2.0
1.5
1.0
0.5
0.0
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
PctUnemployed


Frequency
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PctFam2Par

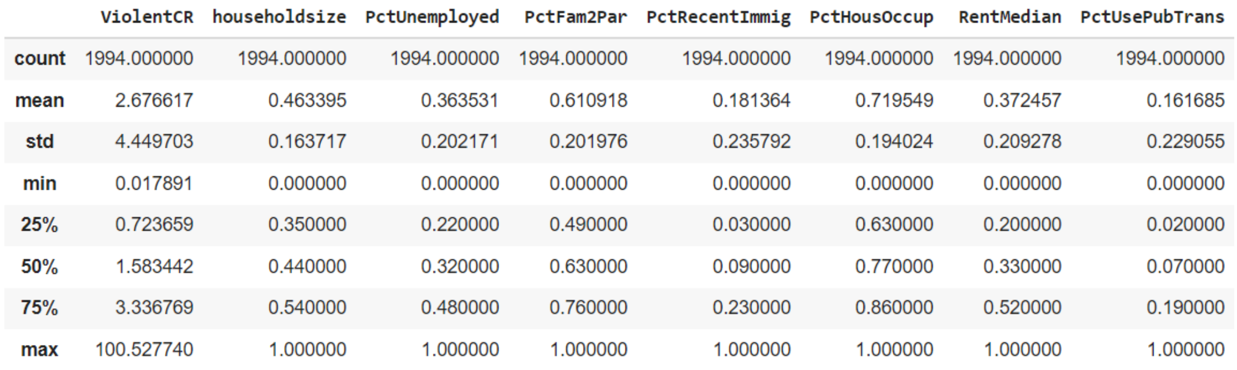

Frequency
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
PctRecentImmig


Frequency
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
PctHousOccup



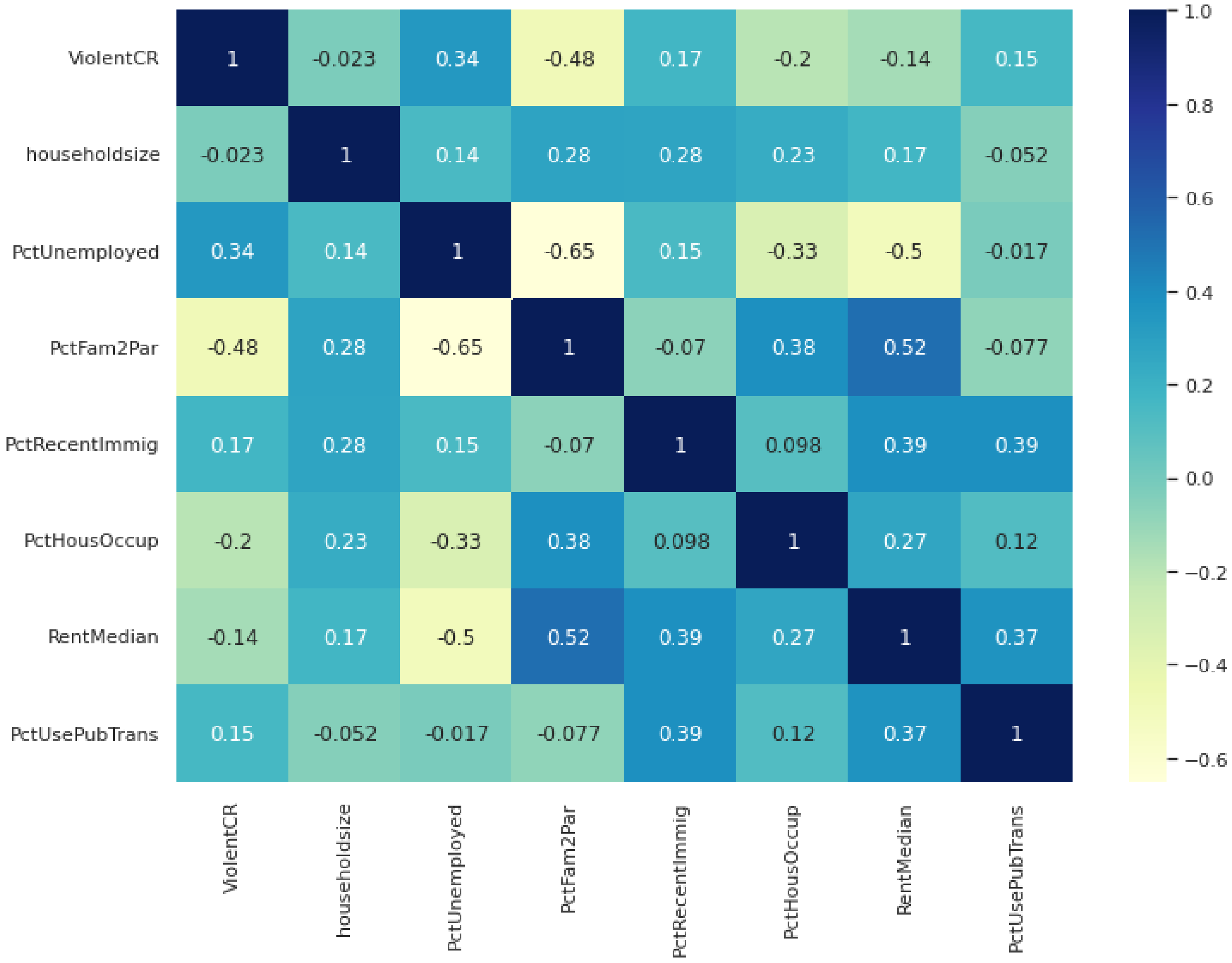


## Descriptive Analysis



| | ViolentCR | householdsize | PctUnemployed | PctFam2Par | PctRecentImmig | PctHousOccup | RentMedian | PctUsePubTrans |
|---|---|---|---|---|---|---|---|---|
| **count** | 1994.000000 | 1994.000000 | 1994.000000 | 1994.000000 | 1994.000000 | 1994.000000 | 1994.000000 | 1994.000000 |
| **mean** | 2.676617 | 0.463395 | 0.363531 | 0.610918 | 0.181364 | 0.719549 | 0.372457 | 0.161685 |
| **std** | 4.449703 | 0.163717 | 0.202171 | 0.201976 | 0.235792 | 0.194024 | 0.209278 | 0.229055 |
| **min** | 0.017891 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 0.723659 | 0.350000 | 0.220000 | 0.490000 | 0.030000 | 0.630000 | 0.200000 | 0.020000 |
| **50%** | 1.583442 | 0.440000 | 0.320000 | 0.630000 | 0.090000 | 0.770000 | 0.330000 | 0.070000 |
| **75%** | 3.336769 | 0.540000 | 0.480000 | 0.760000 | 0.230000 | 0.860000 | 0.520000 | 0.190000 |
| **max** | 100.527740 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

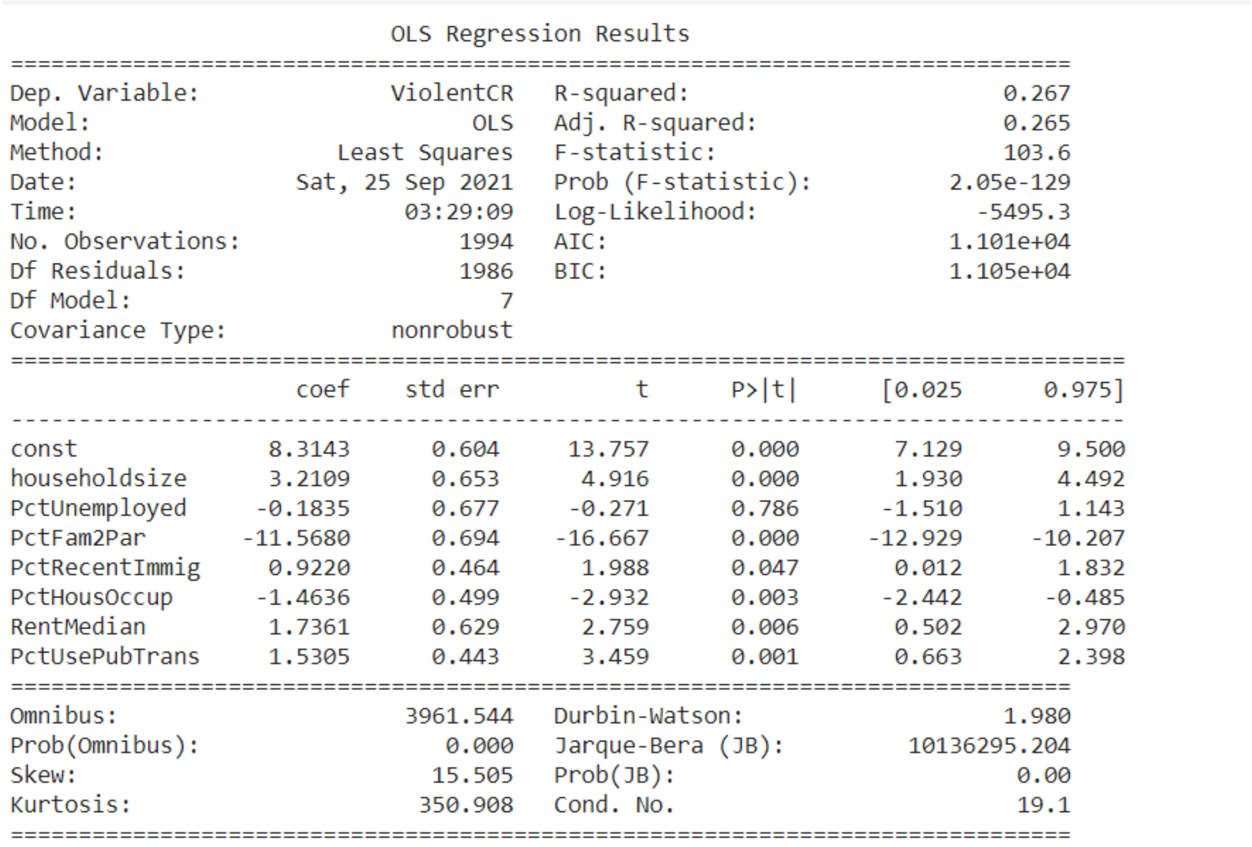
## Correlation between Independent Variable (VoilentCR) and Dependent Variables (Intra and Inter)



## 2. Multiple Regression between CrimeRates and Independent Variables

```
Intercept: 8.31428513123865
Coefficients: [  3.21090514  -0.18345405 -11.56798239   0.92196673
-1.46358042
   1.7361363    1.53050498]
```

## OLS Multiple Regression Results



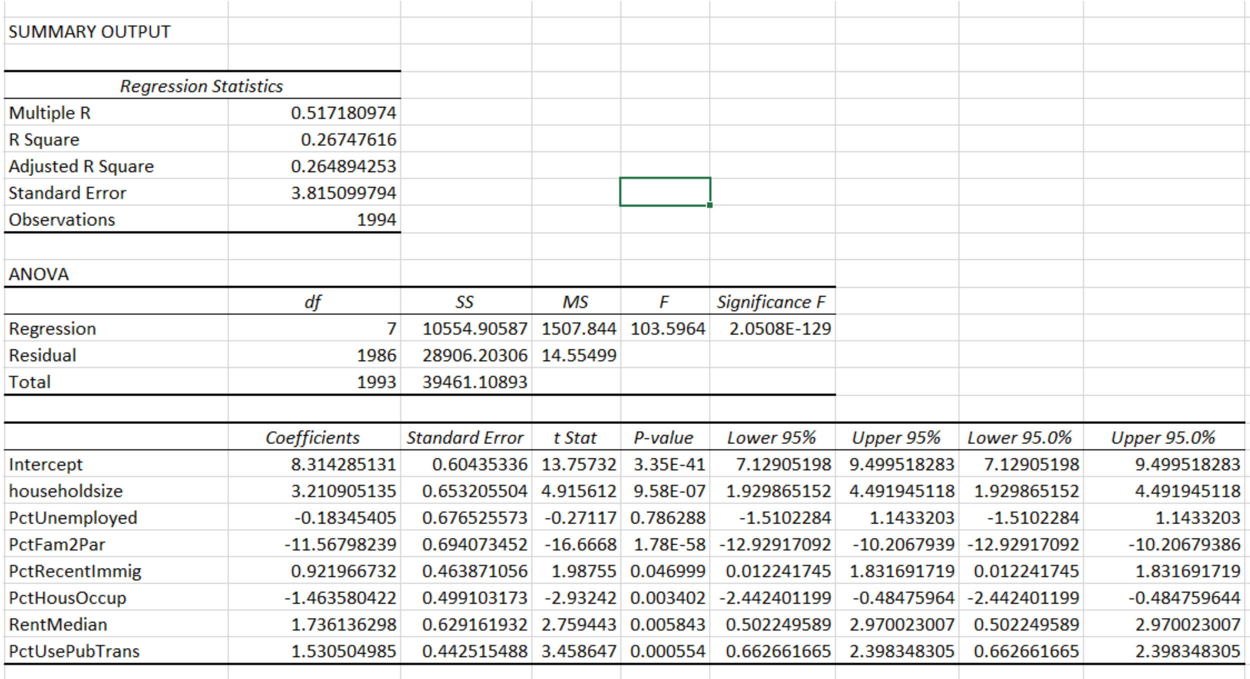
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              ViolentCR   R-squared:                       0.267
Model:                            OLS   Adj. R-squared:                  0.265
Method:                 Least Squares   F-statistic:                     103.6
Date:                Sat, 25 Sep 2021   Prob (F-statistic):          2.05e-129
Time:                        03:29:09   Log-Likelihood:                -5495.3
No. Observations:                1994   AIC:                         1.101e+04
Df Residuals:                    1986   BIC:                         1.105e+04
Df Model:                           7                                         
Covariance Type:            nonrobust                                         
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              8.3143      0.604     13.757      0.000       7.129       9.500
householdsize      3.2109      0.653      4.916      0.000       1.930       4.492
PctUnemployed     -0.1835      0.677     -0.271      0.786      -1.510       1.143
PctFam2Par       -11.5680      0.694    -16.667      0.000     -12.929     -10.207
PctRecentImmig     0.9220      0.464      1.988      0.047       0.012       1.832
PctHousOccup      -1.4636      0.499     -2.932      0.003      -2.442      -0.485
RentMedian         1.7361      0.629      2.759      0.006       0.502       2.970
PctUsePubTrans     1.5305      0.443      3.459      0.001       0.663       2.398
==============================================================================
Omnibus:                     3961.544   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         10136295.204
Skew:                          15.505   Prob(JB):                         0.00
Kurtosis:                     350.908   Cond. No.                         19.1
==============================================================================
```

**Relationship between dependent and independent Variables - R^2 = 0.267 = 26.7%**

**It implies relationship is not significant**

**P values of all independent variables except PCTUnemployed is insignificant showing that only this variable contribution to regression is insignificant**

# Excel Regression Results



SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.517180974 |
| R Square | 0.26747616 |
| Adjusted R Square | 0.264894253 |
| Standard Error | 3.815099794 |
| Observations | 1994 |

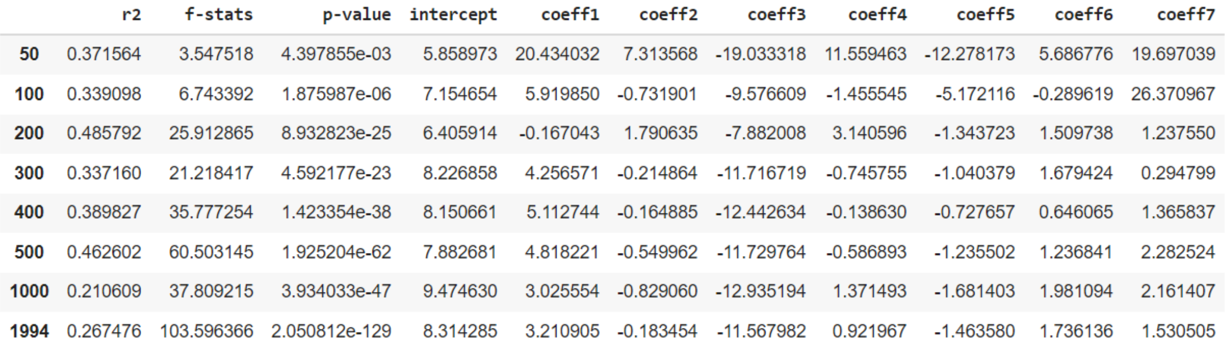
ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 7 | 10554.90587 | 1507.844 | 103.5964 | 2.0508E-129 |
| Residual | 1986 | 28906.20306 | 14.55499 | | |
| Total | 1993 | 39461.10893 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 8.314285131 | 0.60435336 | 13.75732 | 3.35E-41 | 7.12905198 | 9.499518283 | 7.12905198 | 9.499518283 |
| householdsize | 3.210905135 | 0.653205504 | 4.915612 | 9.58E-07 | 1.929865152 | 4.491945118 | 1.929865152 | 4.491945118 |
| PctUnemployed | -0.18345405 | 0.676525573 | -0.27117 | 0.786288 | -1.5102284 | 1.1433203 | -1.5102284 | 1.1433203 |
| PctFam2Par | -11.56798239 | 0.694073452 | -16.6668 | 1.78E-58 | -12.92917092 | -10.2067939 | -12.92917092 | -10.20679386 |
| PctRecentImmig | 0.921966732 | 0.463871056 | 1.98755 | 0.046999 | 0.012241745 | 1.831691719 | 0.012241745 | 1.831691719 |
| PctHousOccup | -1.463580422 | 0.499103173 | -2.93242 | 0.003402 | -2.442401199 | -0.48475964 | -2.442401199 | -0.484759644 |
| RentMedian | 1.736136298 | 0.629161932 | 2.759443 | 0.005843 | 0.502249589 | 2.970023007 | 0.502249589 | 2.970023007 |
| PctUsePubTrans | 1.530504985 | 0.442515488 | 3.458647 | 0.000554 | 0.662661665 | 2.398348305 | 0.662661665 | 2.398348305 |

# Conclusion - Both regression give identical values

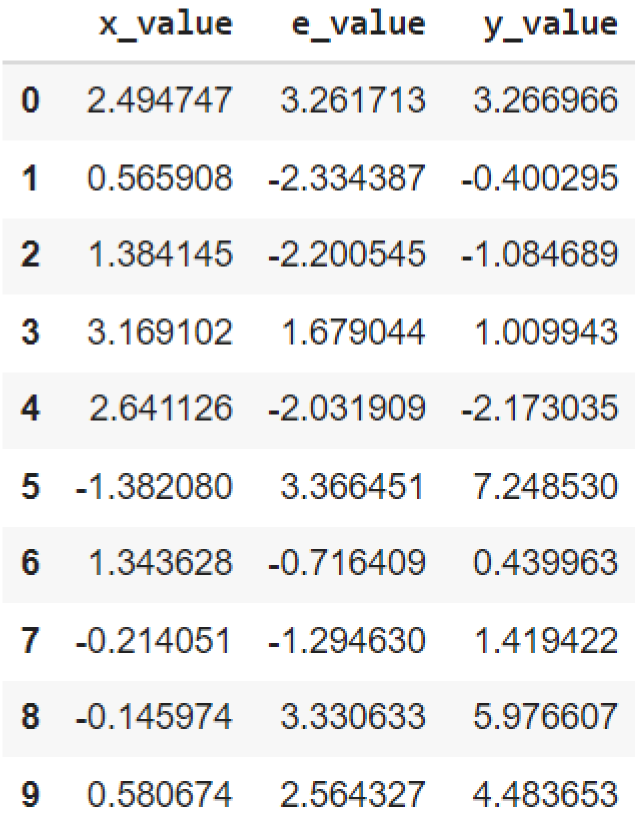
3. **Dataset with subset size of [50,100,200,300,400, 500,1000,1994]**



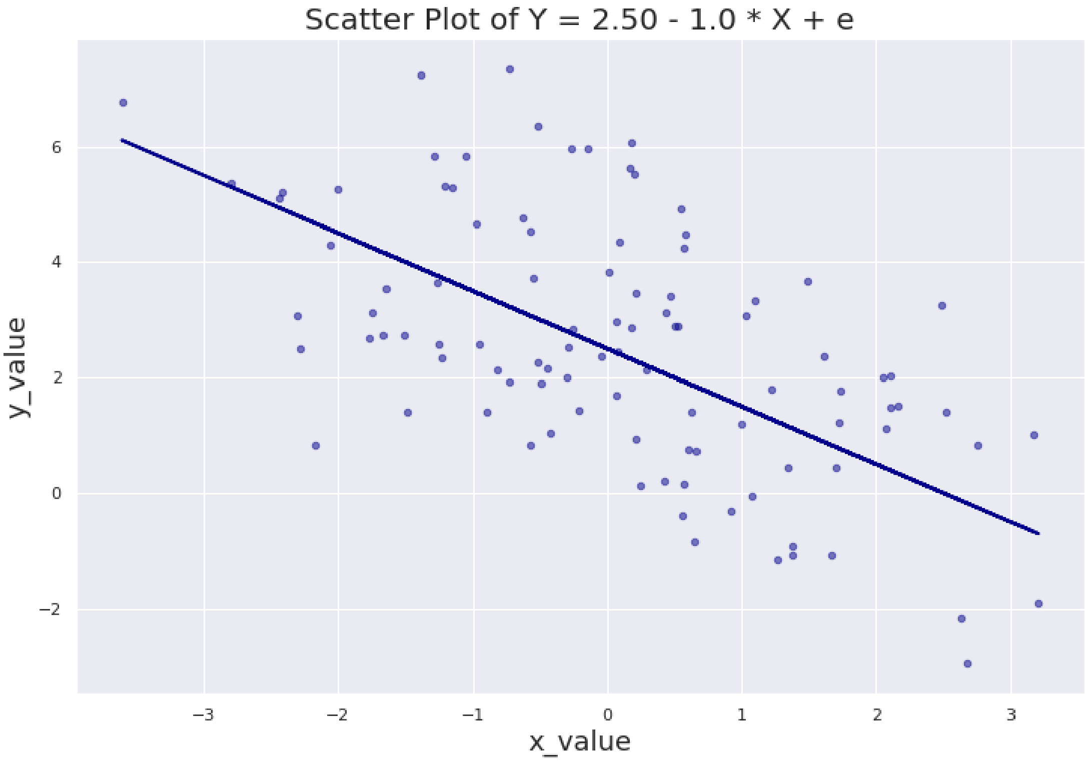
| | r2 | f-stats | p-value | intercept | coeff1 | coeff2 | coeff3 | coeff4 | coeff5 | coeff6 | coeff7 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **50** | 0.371564 | 3.547518 | 4.397855e-03 | 5.858973 | 20.434032 | 7.313568 | -19.033318 | 11.559463 | -12.278173 | 5.686776 | 19.697039 |
| **100** | 0.339098 | 6.743392 | 1.875987e-06 | 7.154654 | 5.919850 | -0.731901 | -9.576609 | -1.455545 | -5.172116 | -0.289619 | 26.370967 |
| **200** | 0.485792 | 25.912865 | 8.932823e-25 | 6.405914 | -0.167043 | 1.790635 | -7.882008 | 3.140596 | -1.343723 | 1.509738 | 1.237550 |
| **300** | 0.337160 | 21.218417 | 4.592177e-23 | 8.226858 | 4.256571 | -0.214864 | -11.716719 | -0.745755 | -1.040379 | 1.679424 | 0.294799 |
| **400** | 0.389827 | 35.777254 | 1.423354e-38 | 8.150661 | 5.112744 | -0.164885 | -12.442634 | -0.138630 | -0.727657 | 0.646065 | 1.365837 |
| **500** | 0.462602 | 60.503145 | 1.925204e-62 | 7.882681 | 4.818221 | -0.549962 | -11.729764 | -0.586893 | -1.235502 | 1.236841 | 2.282524 |
| **1000** | 0.210609 | 37.809215 | 3.934033e-47 | 9.474630 | 3.025554 | -0.829060 | -12.935194 | 1.371493 | -1.681403 | 1.981094 | 2.161407 |
| **1994** | 0.267476 | 103.596366 | 2.050812e-129 | 8.314285 | 3.210905 | -0.183454 | -11.567982 | 0.921967 | -1.463580 | 1.736136 | 1.530505 |

**All R^2 values are less than 0.5 showing weak relationship in multiple regression.**
**Null Hypothesis: All coefficients = 0**
**Alternate Hypothesis: At Least one coefficient not 0**
**P-values are insignificant showing the null hypothesis cannot be rejected.**

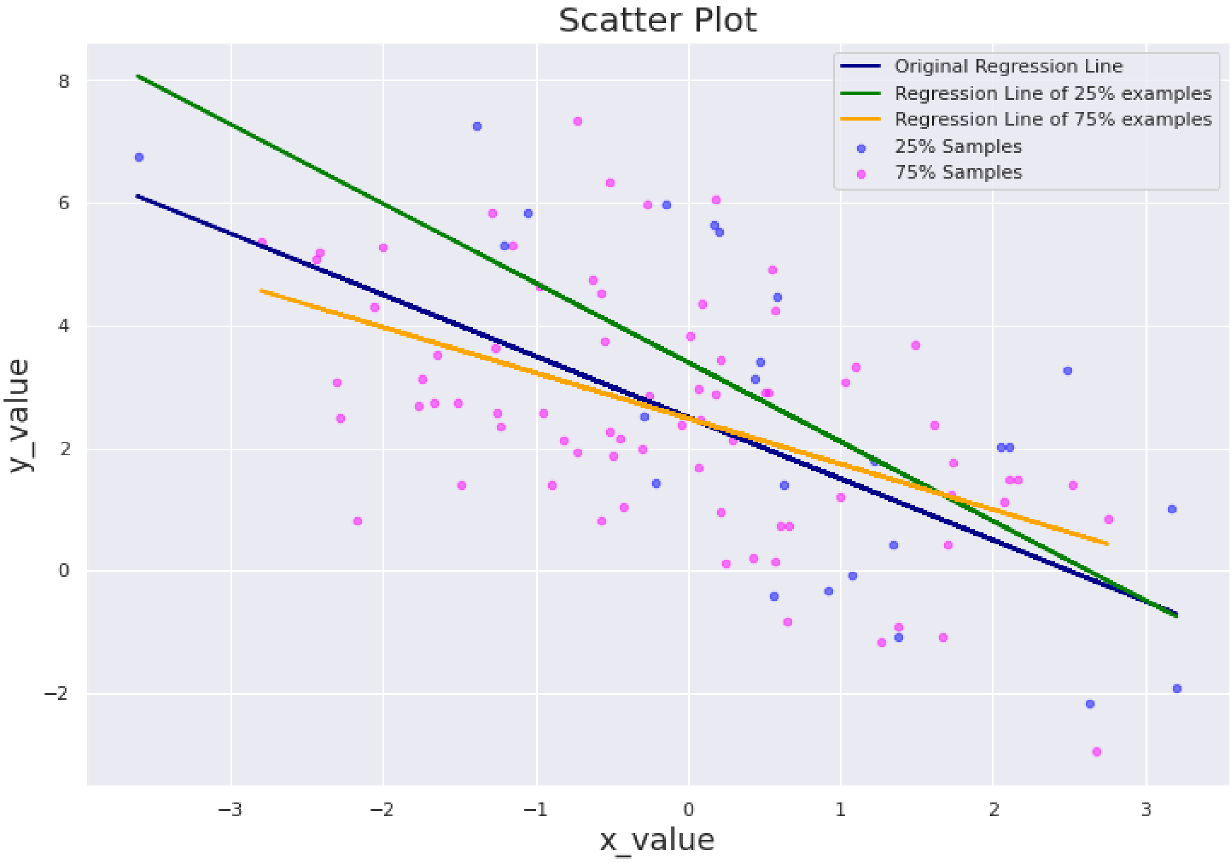
## 4. (a) X_value = N(0, 2) e = N(0, 3) Y = 2.5 - 1.0*X + e



| | x_value | e_value | y_value |
|---|---|---|---|
| **0** | 2.494747 | 3.261713 | 3.266966 |
| **1** | 0.565908 | -2.334387 | -0.400295 |
| **2** | 1.384145 | -2.200545 | -1.084689 |
| **3** | 3.169102 | 1.679044 | 1.009943 |
| **4** | 2.641126 | -2.031909 | -2.173035 |
| **5** | -1.382080 | 3.366451 | 7.248530 |
| **6** | 1.343628 | -0.716409 | 0.439963 |
| **7** | -0.214051 | -1.294630 | 1.419422 |
| **8** | -0.145974 | 3.330633 | 5.976607 |
| **9** | 0.580674 | 2.564327 | 4.483653 |

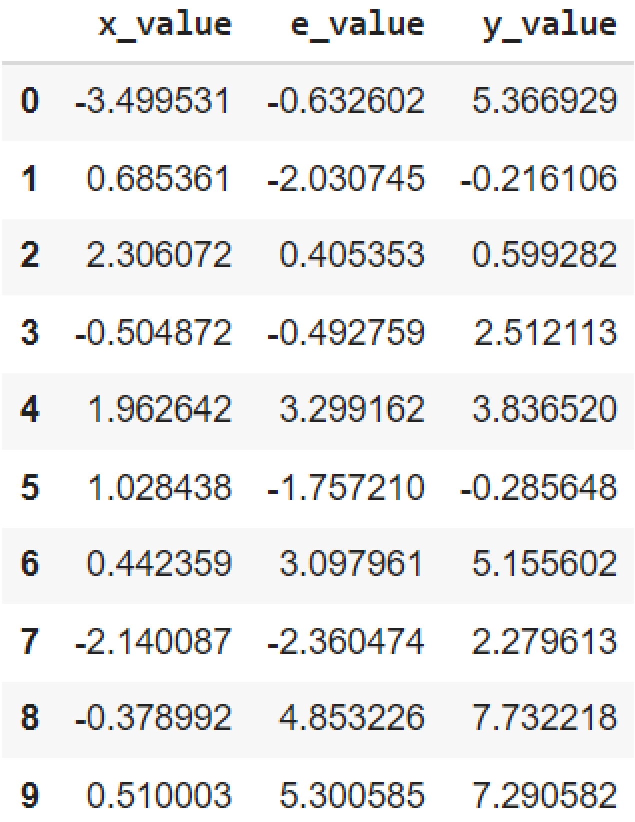


## (b) Dataset is split to 25% and 75% and Scatter Plot is shown

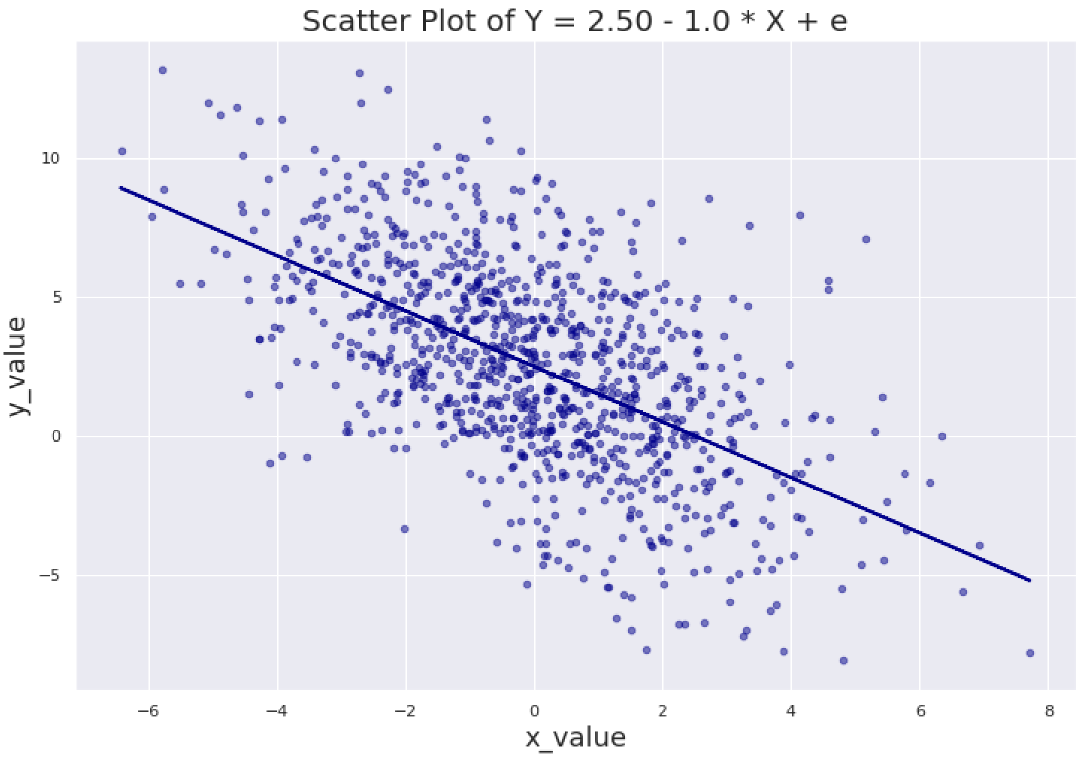


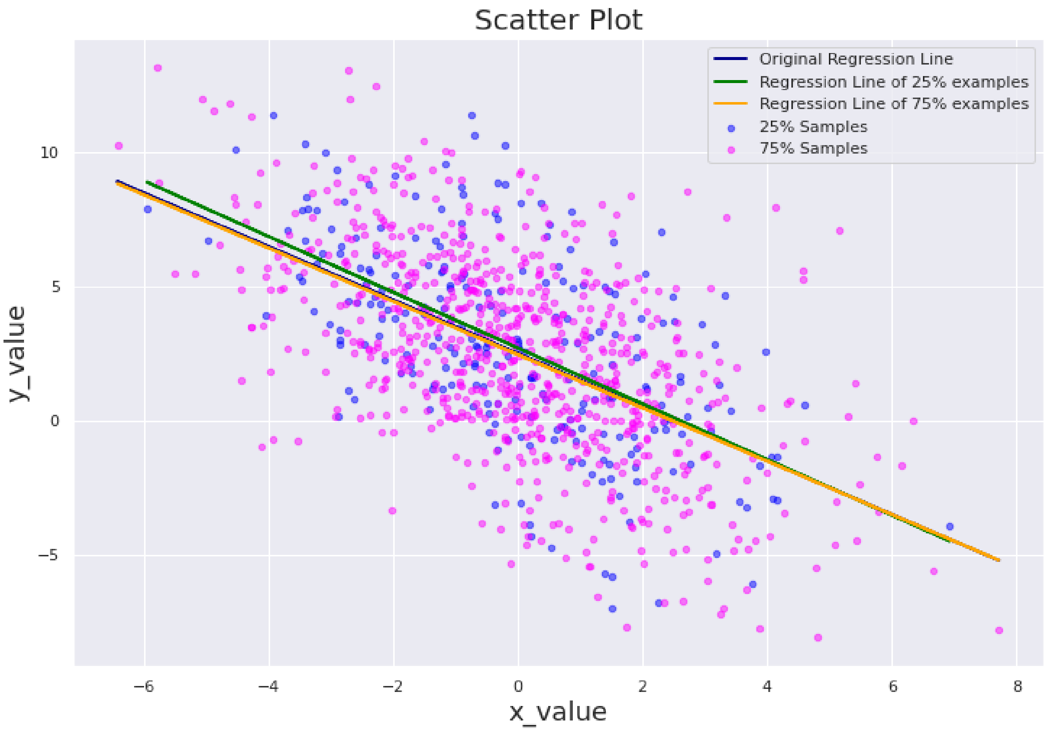
**This shows the regression of 25%, 75% and the whole dataset is different. This is because each dataset split doesn’t contain identical data and the ratio of size of dataset is 1:3.**

## (c) X_value = N(0, 4) e = N(0, 9) Y = 2.5 - 1.0*X + e

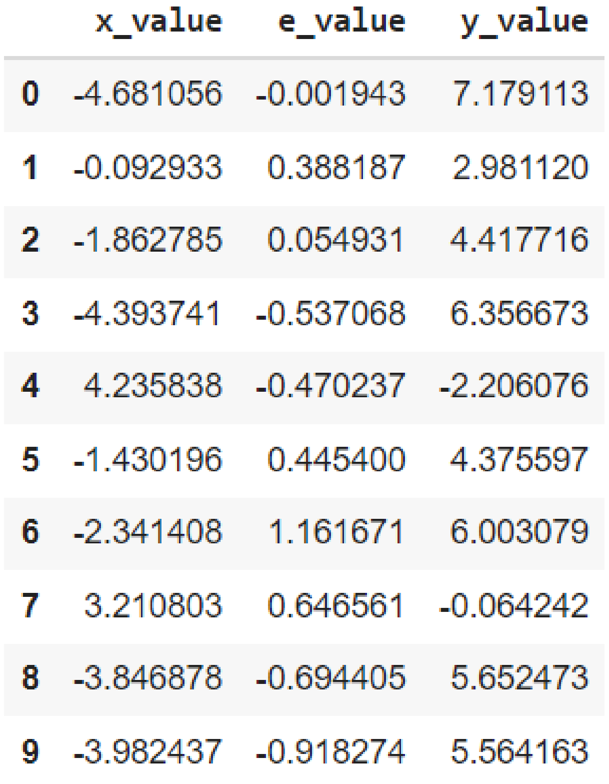


|   | x_value | e_value | y_value |
|---|---|---|---|
| 0 | -3.499531 | -0.632602 | 5.366929 |
| 1 | 0.685361 | -2.030745 | -0.216106 |
| 2 | 2.306072 | 0.405353 | 0.599282 |
| 3 | -0.504872 | -0.492759 | 2.512113 |
| 4 | 1.962642 | 3.299162 | 3.836520 |
| 5 | 1.028438 | -1.757210 | -0.285648 |
| 6 | 0.442359 | 3.097961 | 5.155602 |
| 7 | -2.140087 | -2.360474 | 2.279613 |
| 8 | -0.378992 | 4.853226 | 7.732218 |
| 9 | 0.510003 | 5.300585 | 7.290582 |

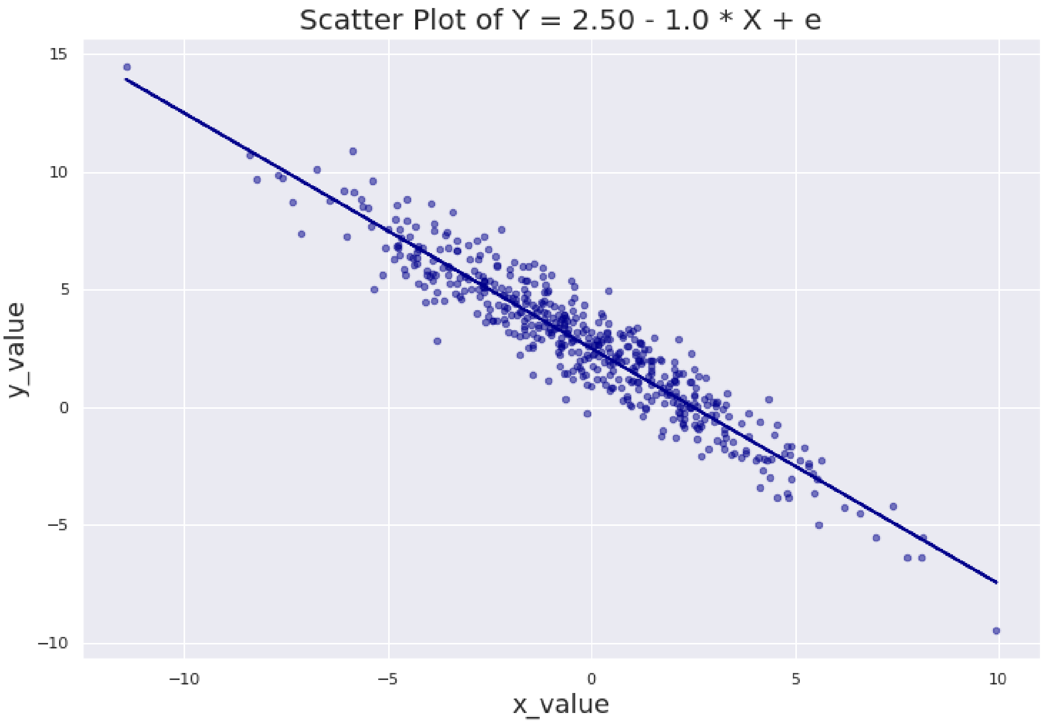


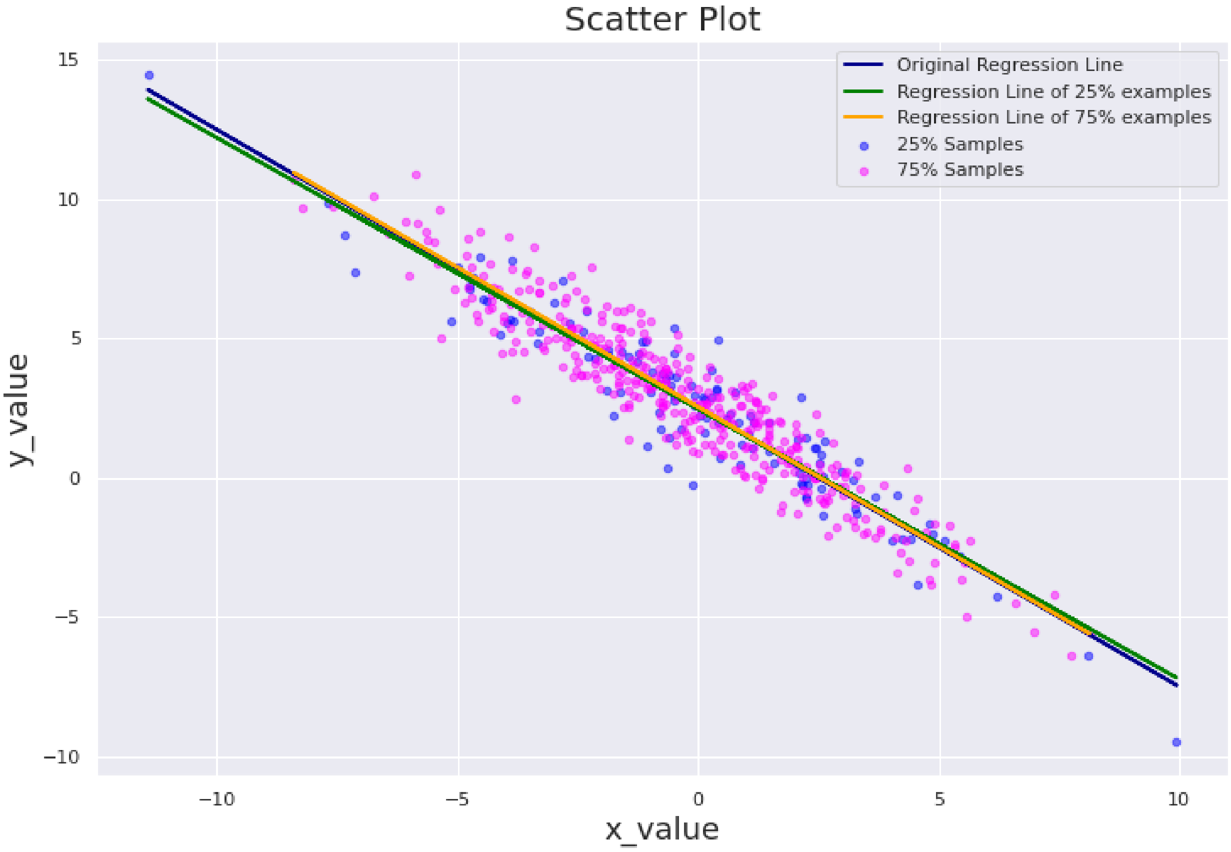


**X_value = N(0, 9) e = N(0, 1) Y = 2.5 - 1.0*X + e**



| | x_value | e_value | y_value |
|---|---|---|---|
| **0** | -4.681056 | -0.001943 | 7.179113 |
| **1** | -0.092933 | 0.388187 | 2.981120 |
| **2** | -1.862785 | 0.054931 | 4.417716 |
| **3** | -4.393741 | -0.537068 | 6.356673 |
| **4** | 4.235838 | -0.470237 | -2.206076 |
| **5** | -1.430196 | 0.445400 | 4.375597 |
| **6** | -2.341408 | 1.161671 | 6.003079 |
| **7** | 3.210803 | 0.646561 | -0.064242 |
| **8** | -3.846878 | -0.694405 | 5.652473 |
| **9** | -3.982437 | -0.918274 | 5.564163 |


Scatter Plot of Y = 2.50 - 1.0 * X + e
15
10
5
0
−5
−10
y_value
−10
−5
0
5
10
x_value


Scatter Plot
Original Regression Line
Regression Line of 25% examples
Regression Line of 75% examples
25% Samples
75% Samples
15
10
5
0
−5
−10
y_value
−10
−5
0
5
10
x_value

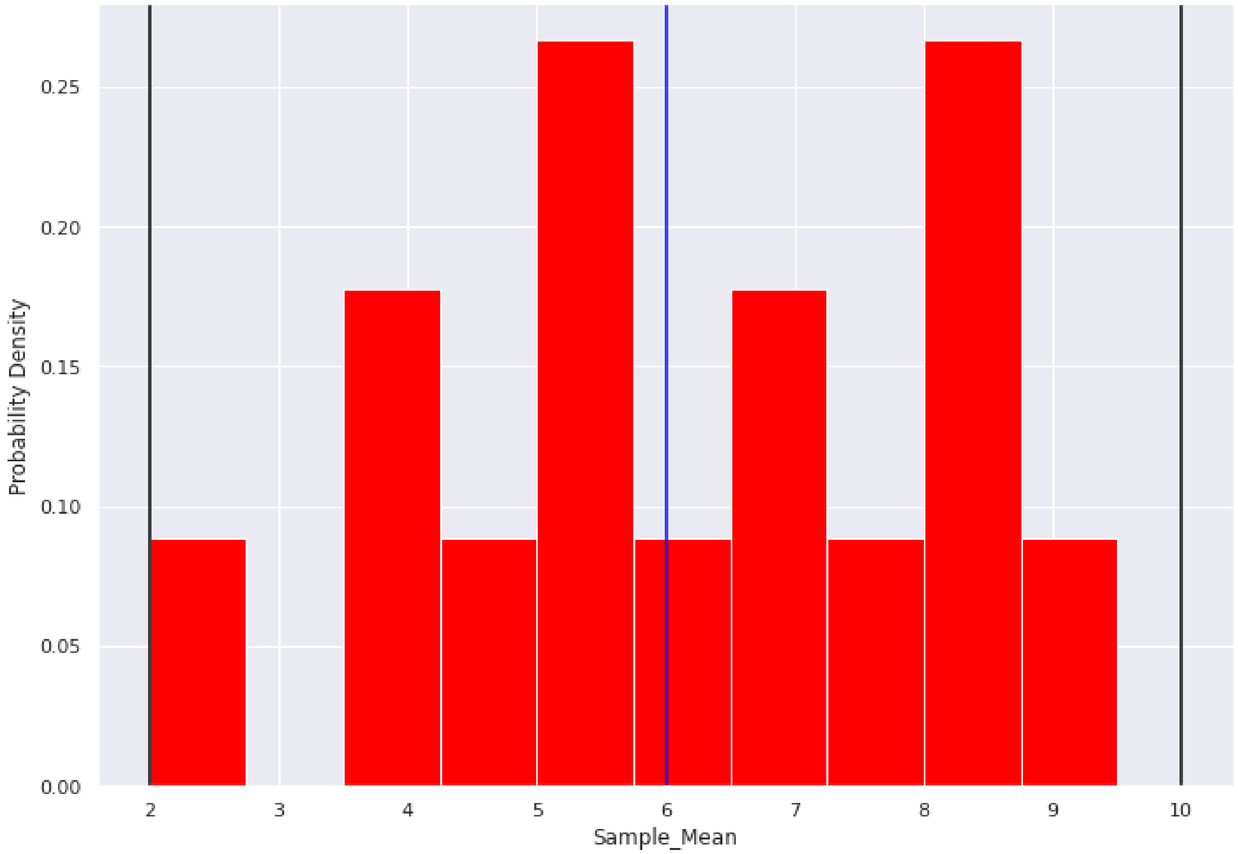
5. Population = [1, 3, 6, 7, 9, 10]

(a) Population Mean = 6.0

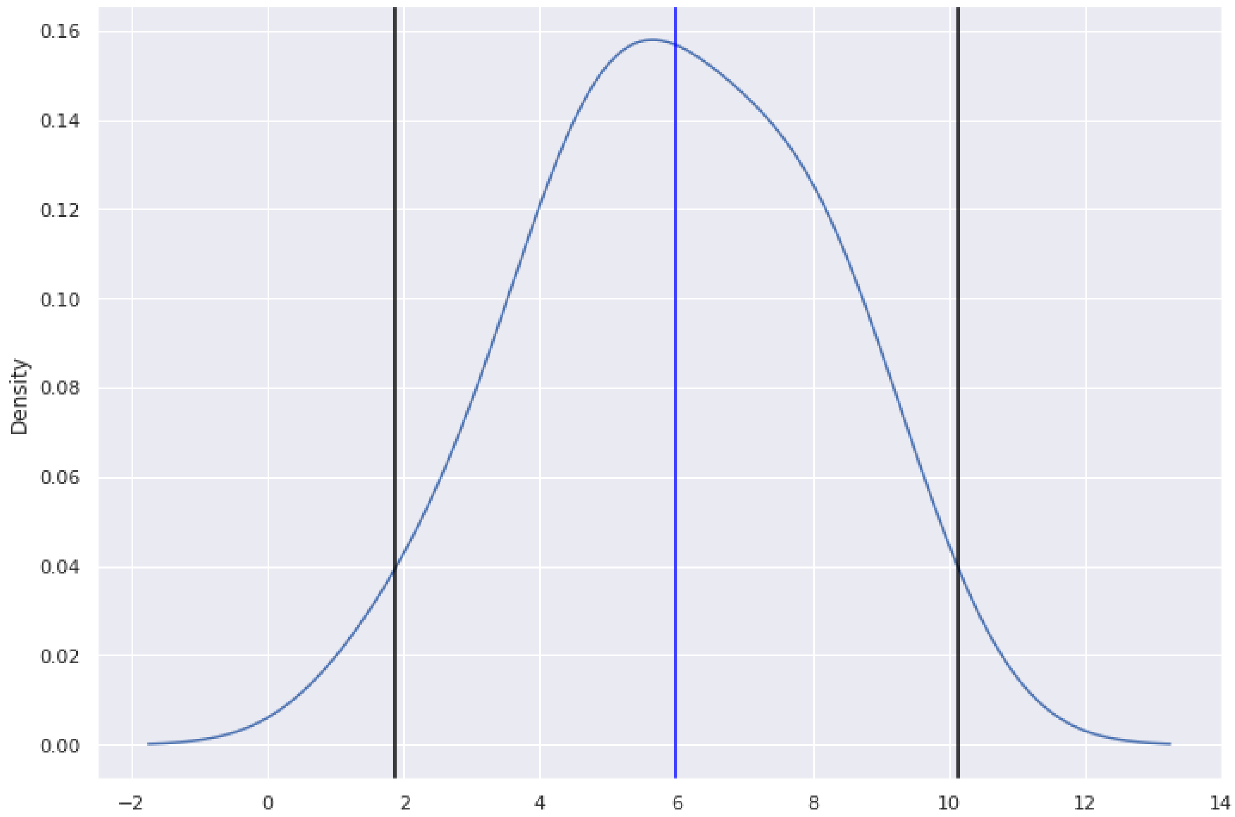
Sample Size (n) = 2

Mean of Sample Means = 6.0

Yes they are equal as all combination of size 2 samples are taken hence occurrence of each element is same



Variance of Sample Means = 4

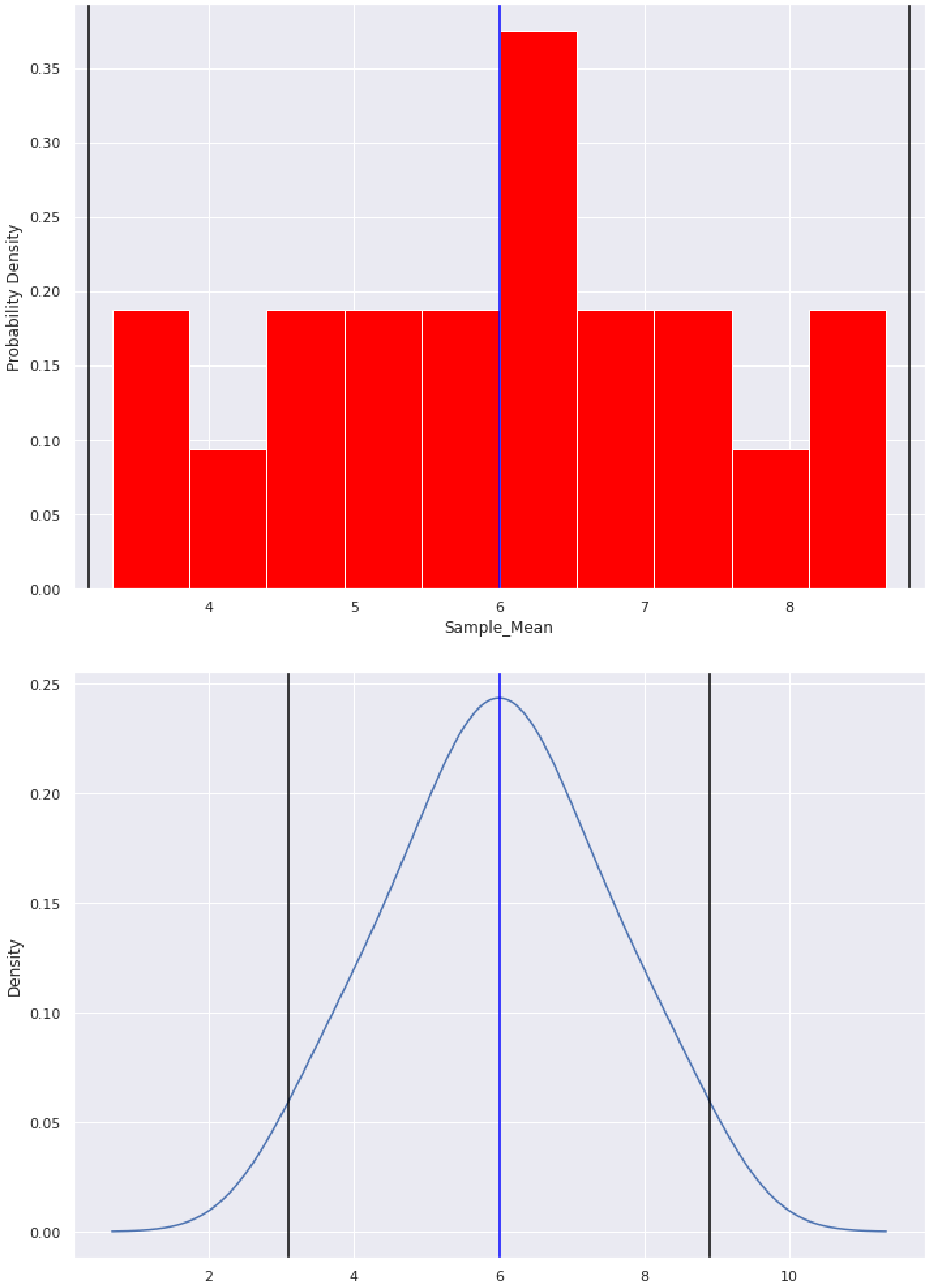


**(b) Population Mean = 6.0**

**Sample Size (n) = 3**

**Mean of Sample Means = 6.0**

**Yes they are equal as all combination of size 3 samples are taken hence occurrence of each element is same**



**Variance of Sample Means (n = 3) is approx equal to 3**

**(c) Plot of Sample with size = 2 is slightly right skewed whereas sample of size = 3 is not skewed as seen in the plots**

**No. of samples of (size = 2) = 15**
**Variance of Sample Means (size = 2) = 4**
**No. of samples of (size = 2) = 20**
**Variance of Sample Means (size = 3) = 3**

**Variance of (a) is greater than (b) due to**

- **Small sample size (2<3)**
- **Total Number of samples of (size = 2) is less than number of samples of (size = 3)**