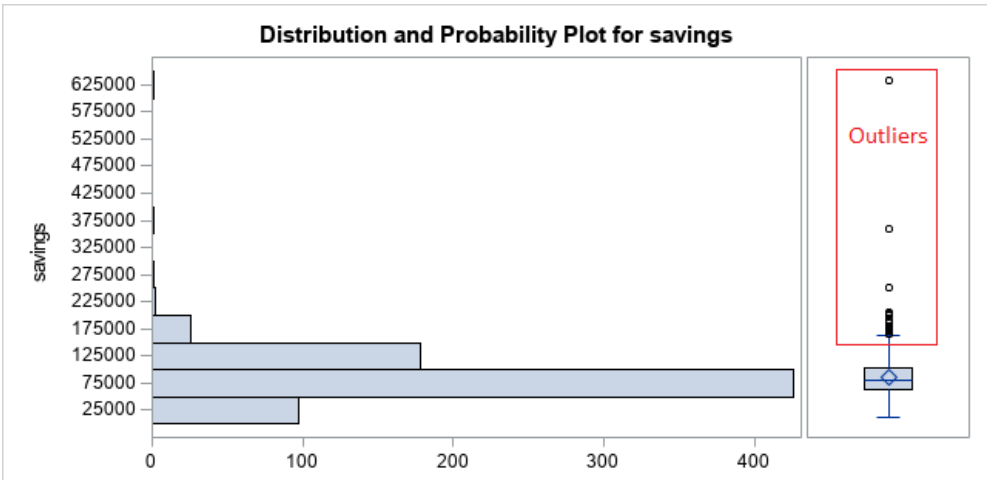
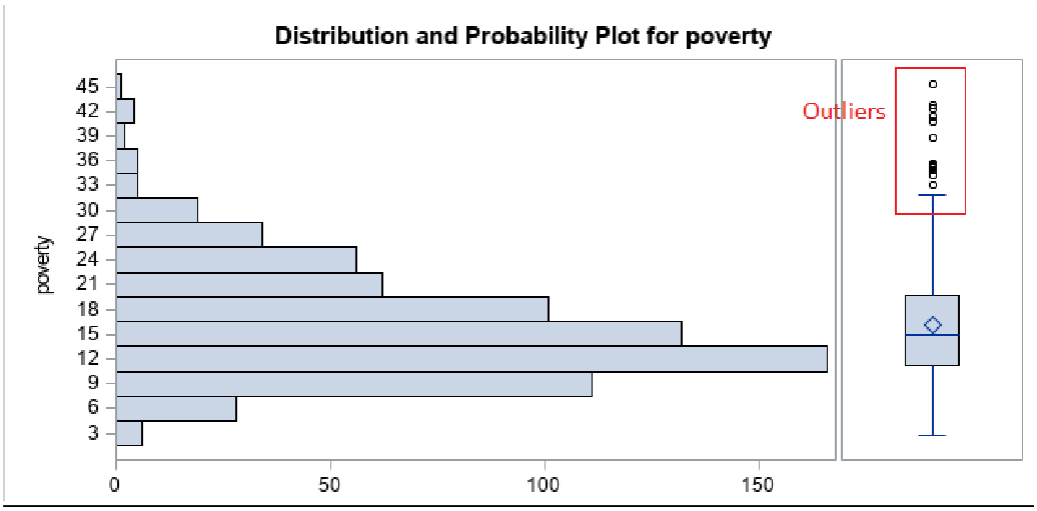
# Output

**Ans 1)**




For Savings: IQR = 41300, Q1 = 62358.5, Q3 = 103658.5
Lower Fence= Q1 –1.5*IQR
LF= 62358.5 - 1.5*41300
LF= 408.5

Upper Fence= Q3 + 1.5*IQR
UF= 103658.5 +1.5*41300
UF= 165,608.5




For Poverty: IQR = 8.55, Q1 = 11.20, Q3 = 19.75
Lower Fence= Q1 –1.5*IQR
LF= 11.20 – 1.5*8.55
LF= -1.625

Upper Fence= Q3 + 1.5*IQR
UF= 19.75 + 1.5*8.55
UF= 32.575

After removing the outliers, the number of observations reduces to 703 from initial 732.

**Ans 2)**

After removing the outliers, since the variable INCOME has a longish tail in distribution, Log Transformation is done to the variable and renamed 'LINCOME'. For building the MLR model first 500 data observations are selected for training the model and the last 203 observations are selected for the testing purpose. The model is built using the following equation: VOTES = LINCOME + SAVINGS + FEMALE +DENSITY +POVERTY + VETERANS

**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Tolerance | Variance Inflation |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -37.70371 | 30.24450 | -1.25 | 0.2131 | . | 0 |
| savings | savings | 1 | 0.00002667 | 0.00001489 | 1.79 | 0.0738 | 0.76677 | 1.30417 |
| poverty | poverty | 1 | 0.84406 | 0.08602 | 9.81 | <.0001 | 0.54319 | 1.84096 |
| veterans | veterans | 1 | 0.60416 | 0.17931 | 3.37 | 0.0008 | 0.85883 | 1.16438 |
| female | female | 1 | 1.23770 | 0.22721 | 5.45 | <.0001 | 0.84506 | 1.18335 |
| density | density | 1 | 0.00271 | 0.00101 | 2.68 | 0.0076 | 0.76050 | 1.31493 |
| LINCOME | | 1 | -0.74106 | 2.91960 | -0.25 | 0.7997 | 0.44738 | 2.23525 |

a) Initially after taking all variables from above equation into consideration we can see for the variable 'LINCOME', t-score is in negative and the p-value is greater than 0.05. Therefore variable 'LINCOME' is dropped and the model is run again. Now variable savings has a p-value greater than 0.05. Therefore, we drop the variable savings. In the 3rd time when we run the model, we drop variable savings and take into account variable 'LINCOME' and get a p-value =0.8643 greater than 0.05.
Hence we drop both the variables 'savings' & 'LINCOME' based on the t-scores & p-value.

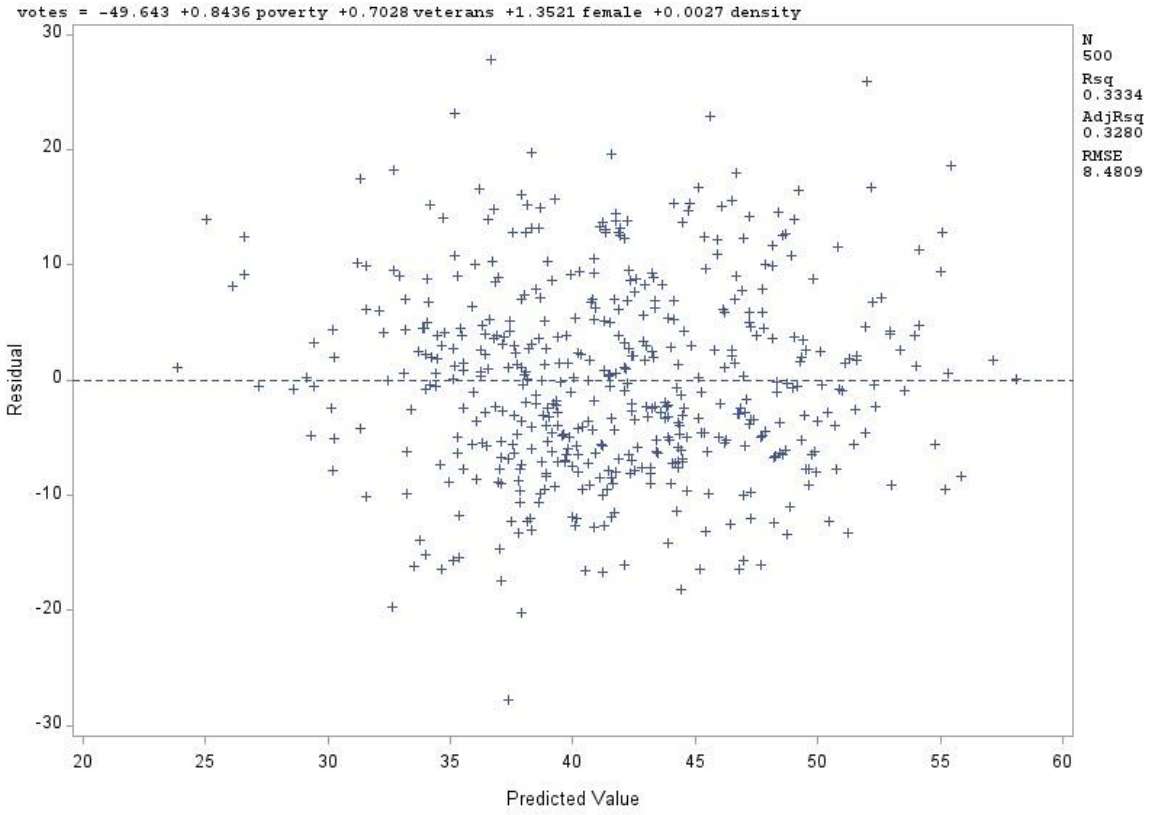
**Parameter Estimates**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Tolerance | Variance Inflation |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -49.64338 | 11.05769 | -4.49 | <.0001 | . | 0 |
| poverty | poverty | 1 | 0.84358 | 0.06793 | 12.42 | <.0001 | 0.87320 | 1.14521 |
| veterans | veterans | 1 | 0.70283 | 0.17079 | 4.12 | <.0001 | 0.94902 | 1.05372 |
| female | female | 1 | 1.35214 | 0.21707 | 6.23 | <.0001 | 0.92819 | 1.07736 |
| density | density | 1 | 0.00273 | 0.00090332 | 3.02 | 0.0026 | 0.95126 | 1.05124 |

According to the t-test scores all above variables are statistically significant having p-value < 0.05.

b) The MSE obtained on training dataset is **71.92509**. The MSE obtained on test dataset is **45.3499982**.
The MSE decreases on testing dataset.

c) Yes, the MLR model employed is a reasonable model for this dataset. The distribution of residuals is unstructured i.e. the relationship between the variables is linear. Also, the residuals do not showcase any particular pattern. We also plot the fitted values Vs the residuals and this seems to suggest that residuals are fairly random irrespective of the fitted value (characteristic that sigma of the error term stays constant). Apart from this the R-Sq & Adj. R-Sq values are close to each other. The final mean square error (MSE) is also small.
**In summary, based on the examination of the residuals it seems that the MLR model chosen is not perfect but probably reasonable.**




| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 17805 | 4451.17814 | 61.89 | <.0001 |
| Error | 495 | 35603 | 71.92509 | | |
| Corrected Total | 499 | 53408 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 8.48087 | R-Square | 0.3334 |
| Dependent Mean | 41.70794 | Adj R-Sq | 0.3280 |
| Coeff Var | 20.33394 | | |