**Assignment 3**

**Author:** Hamna Ashraf, Student # 8826836

**Course:** PROG8431

**Q3.2:** Comment on any outliers you see and deal with them appropriately

A diagram of a graph

Description automatically generated with medium confidence

Figure 1: Boxplot graph with outliers.

The significant outlier that we can see in this graph is for the number of children (n.child) boxplot. The average number of children are approximately between 0-1 which maximum number within the plot is 5.

Min: 0

1st Quartile: 0

Median: 1

3rd Quartile: 2

Max: 5

There is a point that lies very close to the max, i.e 6 so we should not remove that. However, there seems to be an outlier where the number of children is almost 200. So, I replaced this number with the median value which was 1.

The other outliers seem legitimate, and I believe that they should not be removed.

After removing the outlier, the graph looks like this:

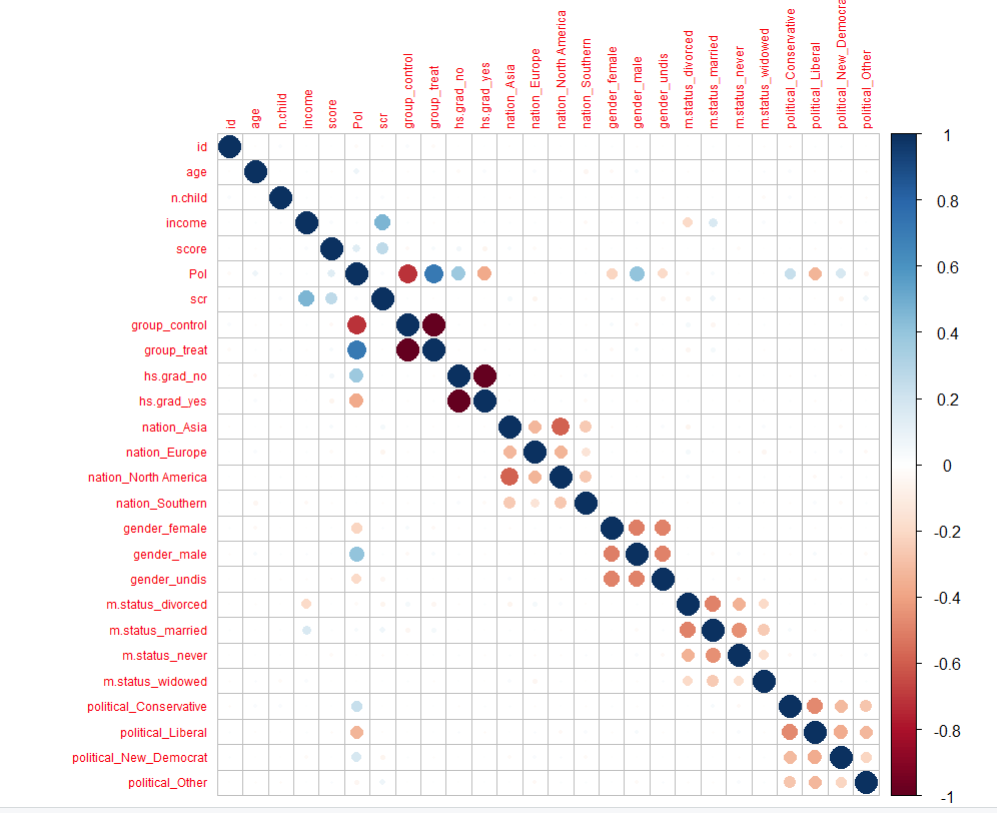
A line graph of a number of people

Description automatically generated with medium confidence

Figure 2: Boxplot with outlier removed (n.child)

If we look at the scale of n.child, it ranges from 0-6.

**Q4.1:** Numerical representation can be seen in R file.



**Q4.2:** Comment on noteworthy correlations you observe. Are these surprising? Do they make sense?

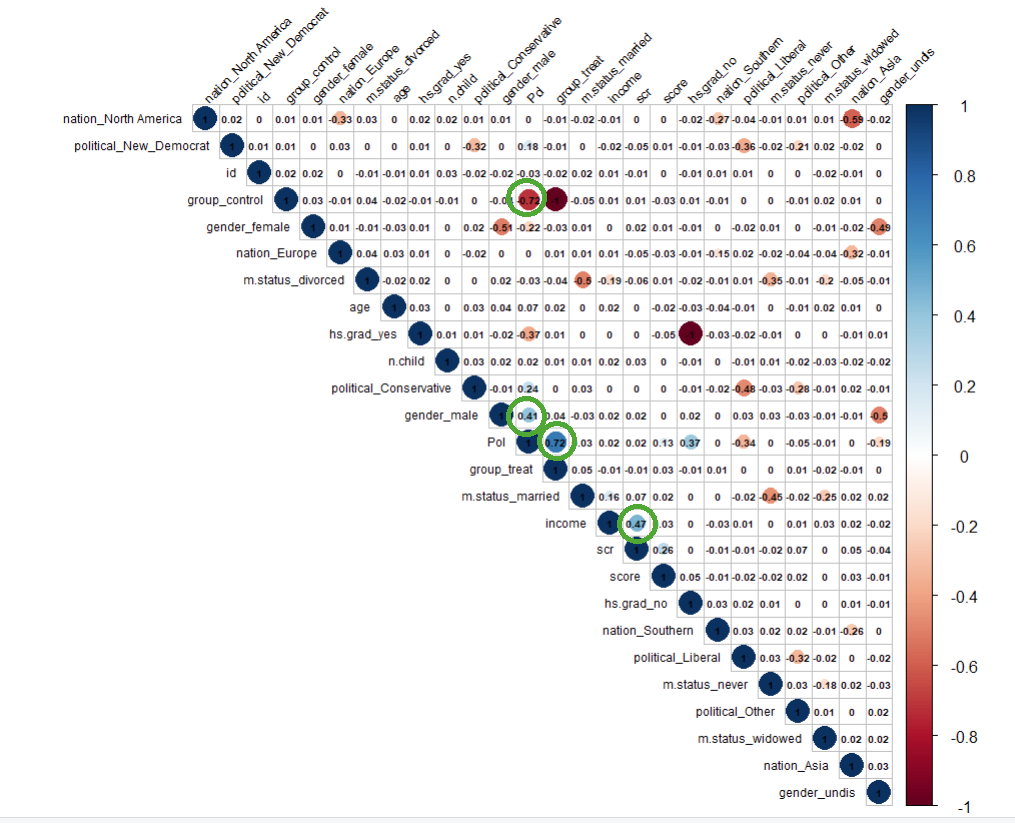
Strong correlation:

Feature ‘group’ has categories ‘control’ and ‘treat’. For category ‘treat’ there is a strong positive correlation with the feature ‘Pol’ of 0.72, whereas category ‘control’ has a strong negative correlation with ‘Pol’ of -0.72.

Weak Correlation:

Features ‘scr’ and ‘income’ has a positive weak correlation equal to 0.5.

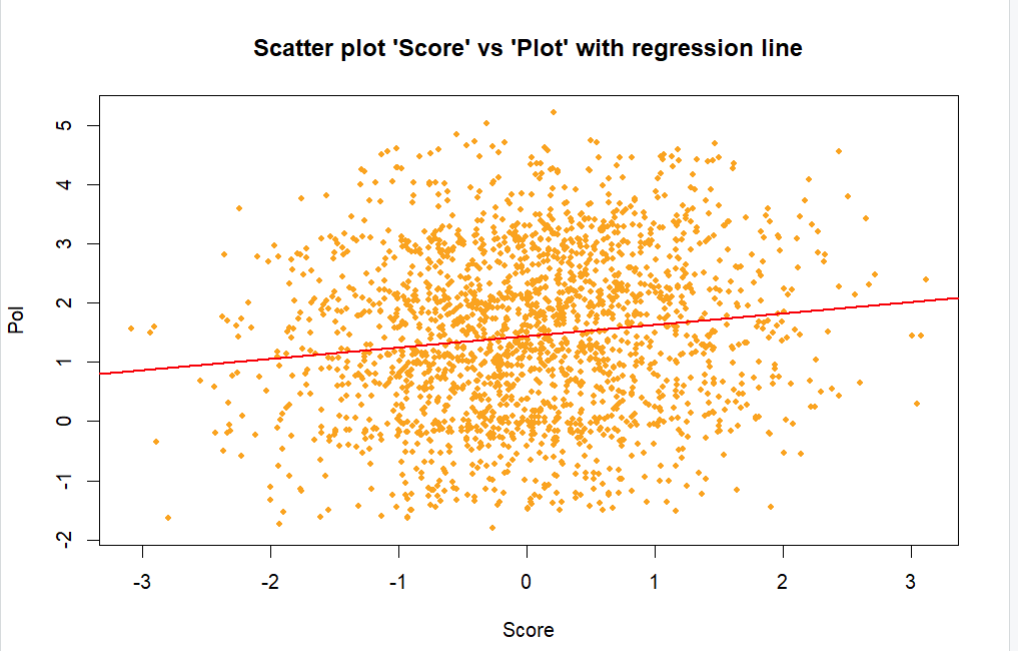
Features ‘gender\_male’ and ‘pol’ have a weak positive correction of 0.41.



The correlation graph doesn’t show very strong correlation between different features, except for 1. Other colors we see in the graph earlier shows correlation of different categories within the same features.

These results are not surprising as we applied a high correlation filter of threshold 0.8 and filtered the highly correlated columns. In our case before we implemented the dummy categories the high correlation value was 0.

**Q5.1:**



**Q5.2**

A graph with a line and dots

Description automatically generated with medium confidence

**Q5.3:**

Score vs Pol -> Model 1

**Model1 summary:**

Residuals:

Min 1Q Median 3Q Max

-3.2606 -1.0705 -0.0353 0.9591 3.7364

Coefficients:

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

(Intercept) 1.44714 0.03055 47.367 < 2e-16 \*\*\*

score 0.19245 0.03193 6.026 0.00000000199 \*\*\*

---

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

Residual standard error: 1.377 on 2031 degrees of freedom

Multiple R-squared: 0.01757, Adjusted R-squared: 0.01708

F-statistic: 36.32 on 1 and 2031 DF, p-value: 0.000000001987

Scr vs Pol -> Model 2

**Model2 summary:**

Residuals:

Min 1Q Median 3Q Max

-3.2486 -1.0454 -0.0078 0.9955 3.7235

Coefficients:

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

(Intercept) 1.39666 0.05778 24.17 <2e-16 \*\*\*

scr 0.05684 0.05074 1.12 0.263

---

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

Residual standard error: 1.389 on 2031 degrees of freedom

Multiple R-squared: 0.0006176, Adjusted R-squared: 0.0001255

F-statistic: 1.255 on 1 and 2031 DF, p-value: 0.2627

Model1 ‘Score’ vs ‘Pol’ seems to be superior than Model2 ‘Scr’ vs ‘Pol’ because of the following reasons:

* P-value for model1 is more statistically significant as its value is lower.
* R2 value is better for model 1 than model 2.
* Model 1 has a better F value.

**Q6.1** Multivariate linear regression- create two models using two automatic variable selection techniques discussed in class(Full(baseline), Backward).

**Full (basline) model:**

Residuals:

Min 1Q Median 3Q Max

-0.96024 -0.16566 -0.00016 0.17552 1.10922

Coefficients: (6 not defined because of singularities)

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

(Intercept) 1.0704397610 0.0396547599 26.994 < 2e-16 \*\*\*

id 0.0000111373 0.0000100898 1.104 0.270

age 0.0054531573 0.0004309504 12.654 < 2e-16 \*\*\*

n.child -0.0002044868 0.0049928800 -0.041 0.967

income 0.0000006843 0.0000001529 4.474 0.0000081 \*\*\*

score 0.1236867472 0.0064716002 19.112 < 2e-16 \*\*\*

scr 0.0043532948 0.0115975351 0.375 0.707

group\_control -1.9632636737 0.0118735094 -165.348 < 2e-16 \*\*\*

group\_treat NA NA NA NA

hs.grad\_no 1.0506364469 0.0118806784 88.432 < 2e-16 \*\*\*

hs.grad\_yes NA NA NA NA

nation\_Asia -0.0041089578 0.0207919343 -0.198 0.843

nation\_Europe -0.0154455455 0.0237775759 -0.650 0.516

`nation\_North America` -0.0037455296 0.0206961724 -0.181 0.856

nation\_Southern NA NA NA NA

gender\_female -0.0074839145 0.0145668038 -0.514 0.607

gender\_male 1.1083908264 0.0145357492 76.253 < 2e-16 \*\*\*

gender\_undis NA NA NA NA

m.status\_divorced 0.0018036395 0.0227436002 0.079 0.937

m.status\_married -0.0333660870 0.0217688603 -1.533 0.125

m.status\_never -0.0132179513 0.0230345963 -0.574 0.566

m.status\_widowed NA NA NA NA

political\_Conservative 0.6731406935 0.0185341764 36.319 < 2e-16 \*\*\*

political\_Liberal -0.5221003756 0.0179762215 -29.044 < 2e-16 \*\*\*

political\_New\_Democrat 0.6800909060 0.0201649799 33.726 < 2e-16 \*\*\*

political\_Other NA NA NA NA

---

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

Residual standard error: 0.2665 on 2013 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9632

F-statistic: 2799 on 19 and 2013 DF, p-value: < 2.2e-16

**Backward model:**

Residuals:

Min 1Q Median 3Q Max

-0.95770 -0.16636 0.00088 0.17450 1.09723

Coefficients:

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

(Intercept) 1.0720844565 0.0269178317 39.828 < 2e-16 \*\*\*

age 0.0054378787 0.0004296190 12.657 < 2e-16 \*\*\*

income 0.0000006976 0.0000001328 5.251 0.000000167 \*\*\*

score 0.1245876440 0.0061890663 20.130 < 2e-16 \*\*\*

group\_control -1.9625924477 0.0118416575 -165.736 < 2e-16 \*\*\*

hs.grad\_no 1.0503619427 0.0118522958 88.621 < 2e-16 \*\*\*

gender\_male 1.1119742524 0.0125136072 88.861 < 2e-16 \*\*\*

m.status\_married -0.0285595215 0.0122833339 -2.325 0.0202 \*

political\_Conservative 0.6725473470 0.0184511968 36.450 < 2e-16 \*\*\*

political\_Liberal -0.5227658590 0.0178958994 -29.211 < 2e-16 \*\*\*

political\_New\_Democrat 0.6793790574 0.0200404542 33.900 < 2e-16 \*\*\*

---

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

Residual standard error: 0.2661 on 2022 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9633

F-statistic: 5334 on 10 and 2022 DF, p-value: < 2.2e-16

**Q6.2:**

**Baseline model:**

1. F-Stat:

The baseline model has a high f-stat value which makes it statistically significant.

1. R-Squared value:

The R2 value of 0.96, which is almost 1, indicates that regression predictions fit the data perfectly. This can show overfitting.

1. Residuals:

The spread of residuals is narrow which means that model predictions are accurate.

1. Significant variables:

The variables that have p-values that are very low are significant predictions than ones that have p-values greater than 0.01.

1. Variable Coefficient:

The range of the coefficients is wide, from positive to negative. The variable co-efficient shows the magnitude and direction of influence.

**Backward model:**

1. F-stat:

In the backward model also the f-stat value is high making it statistically significant.

1. R-squared value:

The R-squared value shows that even after removing non-significant variables, it captures most variance.

1. Residuals:

Like full model, the residuals are in narrow ranging indicating a good prediction.

1. Significant variables:

For backward model, the significant predictors are less. This improves the interpretability of the model better.

1. Variable Coefficient:

The positive coefficient with small P-value for gender\_male shows strong influence on Pol. Whereas, variable group\_control is high negative value which shows strong negative effect on Pol.

**Q6.3:** Model evaluation: For both models evaluate the main assumptions of regression and evaluate model performance.

A screenshot of a computer code

Description automatically generated

Baseline:

A group of graphs showing different values

Description automatically generated

Backward:

A group of graphs and diagrams

Description automatically generated with medium confidence

Both models give similar results with some minor changes.

**Q6.4:**

Based on the R-squared values of both models, both models perform well however, we will choose the model which chooses less predictors i.e backward model.