INTRODUCTION

The name of the data set being reviewed is ‘Psychology of debt’. The data were obtained from a large postal survey on the psychology of debt. I obtained the dataset ‘debt’ from the ‘faraway’ package in R.

The table below contains variables from the dataset (*name of the variable*; *description of the variable*; *measurement type* (factor or continuous); *role predictor* or *outcome.*)

	name of the variable	***description of the variable***	***measurement type* (factor or continuous)**	***role – predictor* or *outcome.***
	incomegp	income group	Factor*	Predictor
	house	security of housing tenure	Factor*	Outcome
	children	number of children in household	continuous	Predictor
	singpar	is the respondent a single parent?	factor	Predictor
	agegp	age group	factor	Predictor
	bankacc	does the respondent have a bank account?	factor	Predictor
	bsocacc	does the respondent have a building society account?	factor	Predictor
	manage	self-rating of money management skill (high values=high skill)	Continuous	Predictor
	ccarduse	how often did s/he use credit cards (1=never... 3=regularly)	Continuous	Predictor
	cigbuy	does s/he buy cigarettes?	factor	Predictor
	xmasbuy	does s/he buy Christmas presents for children?	factor	Predictor

	locintrn	score on a locus of control scale (high values=internal)	continuous	Predictor
	prodebt	score on a scale of attitudes to debt (high values=favourable to debt	continuous	outcome

- The dataframe *debt* initially consisted of 464 observations of 13 variables.

The numerical summay its variables are as follows:

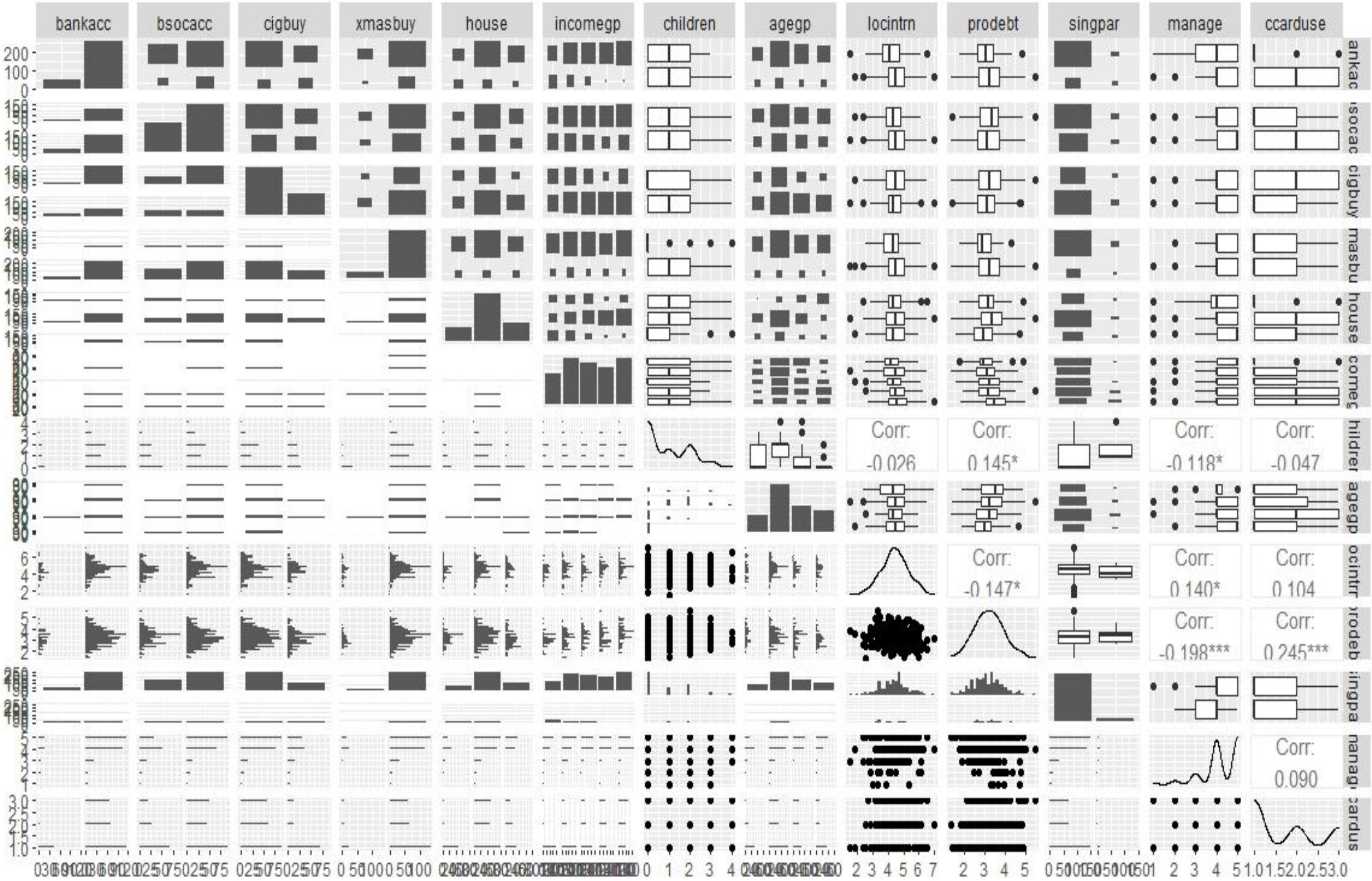
```
$ incomegp: num  2 5 1 3 5 3 4 1 2 1 ...
 $ house   : num  3 2 1 3 2 3 2 1 3 3 ...
 $ children: num  1 3 0 0 2 0 0 0 0 0 ...
 $ singpar : num  0 0 0 0 0 0 0 0 0 0 ...
 $ agegp   : num  2 2 3 4 2 4 2 4 4 4 ...
 $ bankacc : num  1 1 NA 1 1 1 1 1 1 1 ...
 $ bsocacc : num  NA NA NA 0 0 0 0 0 0 NA ...
 $ manage  : num  5 4 2 5 5 4 5 5 5 5 ...
 $ ccarduse: num  2 3 2 2 3 2 2 1 NA 2 ...
 $ cigbuy  : num  1 0 0 0 0 0 0 0 0 0 ...
 $ xmasbuy : num  1 1 0 1 1 1 1 1 0 1 ...
 $ locintrn: num  2.83 4.83 3.83 4.83 3.17 ...
 $ prodebt : num  2.71 3.88 3.06 4.29 3.82 ...
```

- A new dataset **‘david.project’** with 304 observations was created

- 160 observations were excluded from the initial datset

ANALYSES

- **Using GGally's matrix plotting function a summary of all of the variables was created**



- I found the bi-variate relationship between prodebt (outcome) and children (predictor) very interesting. This is because they have positive correlation coefficient of 0.145 which implies that people's attitude towards debt is influenced by the number of children in their house hold.

3a Using only *locintrn*, fit a linear regression model to predict the outcome *prodebt*.

Model: lm(formula = prodebt ~ locintrn, data = debt)

Residuals: Min 1Q Median 3Q Max

-2.08043 -0.48378 -0.01261 0.46203 2.12692

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

(Intercept) 3.69352 0.16849 21.921 < 2e-16 ***

locintrn -0.10524 0.03767 -2.794 0.00546 **

Residual standard error: 0.7022 on 406 degrees of freedom

Multiple R-squared: 0.01886, Adjusted R-squared: 0.01644

F-statistic: 7.804 on 1 and 406 DF, p-value: 0.005458

- Given all the coefficients (estimate value) of our variables are significant, it can be said the model is significant.
- Adjusted R value tells us the amount of variance accounted for is 0.01644 percent which depicts a large percentage of variance not covered and tells us our overall model is not fit
- Since the *p*-value associated with F statistic is statically significant we can say that ***locintrn*** is a statistically significant predictor of ***prodebt***
- Holding intercept value (3.69352) constant, a single unit increase in (***locintrn)*** *coeffient* (-0.10524) would reduce the amount of need to get debt (***prodebt)*** by 0.10524.

3b Creating a second model, where *locintrn*, and *manage* were predictors.

```
Model: lm(formula = prodebt ~ locintrn + manage, data = debt)

Residuals:

    Min       1Q   Median       3Q      Max

-1.86027 -0.47594 -0.03538  0.44111  2.13976

Coefficients:

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

(Intercept)  4.30037    0.20629  20.846  < 2e-16 ***

locintrn    -0.07372    0.03735  -1.974   0.0491 *

manage      -0.18116    0.03640  -4.977 9.61e-07 ***

---

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6846 on 402 degrees of freedom

Multiple R-squared:  0.07515,        Adjusted R-squared:  0.07055

F-statistic: 16.33 on 2 and 402 DF,  p-value: 1.514e-07
```

Comparing the two-predictor model to the fit statistics associated with model 1.

Given our new adjusted R value of 0.07055 the two-predictor model is better because it has greater coverage of variation than in our previous model with adjusted R value of 0.01644

Coefficients of *locintrn* and *manage*.

Holding *locintrn* value (**-0.07372**) constant, a single unit increase in (***manage)*** *coeffient* (**-0.18116**) would reduce the need to get debt (***prodebt)*** by **0.18116**

The implication of change in the coefficient of *locintrn* from model 1 to model 2.

The implication of change in coefficient from model1 to model 2 is

Holding all other variables constant, a unit increase in *locintrn reduces the chances* to incure *debt (prodebt) by* 0.10524 for model 1

While Holding all other variables constant, a unit increase in *locintrn reduces the chances to* to incur *debt (prodebt) by* 0.07372 for model 2

Overall the model is a better fit since the adjusted R value increased from **0.01644** to 0.07372

3c. Building a three predictor model, adding *children* as a third predictor.

Model: lm(formula = prodebt ~ locintrn + manage + children, data = debt)

```
Residuals:

    Min      1Q   Median      3Q     Max

-1.80600 -0.47755 -0.03916  0.41977  2.07433

Coefficients:

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

(Intercept)  4.19339    0.21118  19.857  < 2e-16 ***

locintrn    -0.07410    0.03718  -1.993   0.0469 *

manage      -0.17043    0.03657  -4.660  4.3e-06 ***

children     0.06537    0.03011   2.171   0.0305 *

---

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6814 on 401 degrees of freedom

Multiple R-squared:  0.0859,        Adjusted R-squared:  0.07906

F-statistic: 12.56 on 3 and 401 DF,  p-value: 7.278e-08
```

Comparing the new three-predictor model to the fit statistics associated with model 2.

Given our new adjusted R value of **0.07906** the three-predictor model is a better fit because it has greater coverage of variation than in our previous model with adjusted R value of **0.07055**

Impact of change in the coefficient of *locintrn* and *manage* from model 2 to model 3 would imply, taking into account any changes in the model fit.

In model 2, *locintrn* -0.07372 and manage -0.18116

In model 3, *locintrn* -0.07410 and manage -0.17043

Overall the model is a better fit since the adjusted R value increased from **0.07055** to **0.07906**

On the basis of the coefficient of *children*, the effect that having children appears to has on a respondent's attitude towards debt.

While holding all other variables constant, a unit increase in children increases the chances to incur *debt (prodebt) by* **0.06537**

3d. **Building a four-predictor model and adding *singpar* as a fourth predictor**

```
Call:

lm(formula = prodebt ~ locintrn + manage + children + singpar,

    data = debt)

Residuals:

     Min       1Q   Median       3Q      Max

-1.80769 -0.47801 -0.03112  0.43604  2.06480

Coefficients:

            Estimate Std. Error t value

(Intercept)  4.19951    0.21159  19.848

locintrn    -0.07409    0.03721  -1.991

manage      -0.17132    0.03663  -4.677

children     0.06885    0.03068   2.244

singpar1    -0.07964    0.13203  -0.603

            Pr(>|t|)

(Intercept)  < 2e-16 ***

locintrn      0.0471 *

manage      3.99e-06 ***

children      0.0254 *

singpar1      0.5467

---

Signif. codes:

  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1  ‘ ’ 1
```

Residual standard error: 0.682 on 400 degrees of freedom

**Multiple R-squared: 0.08673,	Adjusted R-squared: 0.07759**

F-statistic: 9.496 on 4 and 400 DF, p-value: 2.42e-07

Comparing the new four-predictor model to the fit statistics associated with model 3.

Given our new adjusted R value of 0.07759 the four-predictor model is not a better fit because it has less coverage of variation than in our previous model with adjusted R value of 0.07906

On the basis of the coefficient of *singpar*, reviewing the effect that being a single parent appears to have on a respondent's attitude towards debt.

While Holding all other variables constant, a unit increase in single parents reduces the chances to incur debt *(prodebt) by* 0.07964

3e. Building a five-predictor model, adding another predictor that I found interesting,

```
Call:
   lm(formula = prodebt ~ locintrn + manage + children + singpar +
      bankacc, data = debt)
Residuals:
      Min       1Q   Median       3Q      Max
   -1.86827 -0.49717 -0.01534  0.43846  2.01730
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
   (Intercept)  4.10866    0.22838  17.990  < 2e-16 ***
   locintrn    -0.08764    0.03907  -2.243   0.0255 *
   manage      -0.18132    0.03918  -4.627  5.1e-06 ***
   children     0.06293    0.03126   2.013   0.0448 *
   singpar1    -0.01163    0.14658  -0.079   0.9368
   bankacc1     0.23533    0.09777   2.407   0.0166 *
   ---
   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.6825 on 377 degrees of freedom
    Multiple R-squared:  0.09233, Adjusted R-squared:  0.08029
   F-statistic:  7.67 on 5 and 377 DF,  p-value: 6.988e-07
```

Reason for choosing the particular fifth variable you have chosen.

The fifth variable bankaacc was chosen because owning a bank account by individuals could impact the chances of one incurring debt through loans or credit card from the bank.

Reviewing whether the fifth variable added improved the ability to predict an individual's attitude towards debt.

Given our new adjusted R value of **0.08029 ,** the five-predictor model is a better fit and improves our ability to predict an individual's attitude better because it has more coverage of variation than in our previous model with adjusted R value of **0.07759**

CONCLUSION

Based on the variables locitnrn, amange, children, singpar, bankcc we are able to deduce how these various factors can influence people's attitude to incurring debt. Therefore, it is necessary to cultivate the habits as seen from the variables which reduce the chances of accumulating debt.

A variable I think could influence the ability to predict an individual's attitude towards debt is *ccarduse* which connotes credit card usage. This could depict how often an individual buys things on credit which implies debt from a bank which normally attracts an extra cost during payment.