Regression Model

I. File extraction

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
11	41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
12	29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
13	53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
14	58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
15	57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
16	51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no

Dataset

II. Finding correlation between attributes

```
16 #checking correlation between attributes
17 data2<- data1[, -c(5, 9,10,11,13,14,15,16,17)] #drop other values that will not be used
18 data2 <- data2 %>% #changing categorical values to numerical ones
19   mutate(across(everything(), ~ as.numeric(factor(.))))
20
21 cor(data2, method="pearson")
```

Commands used to check correlation between different attributes.
Categorical values are first changed to numerical values then it's passed to the 'cor' function

```
> cor(data1, method="pearson")
                 age          job      marital    education      balance      housing         loan     duration
age        1.00000000 -0.021872447 -0.403258057 -0.106829332  0.119072686 -0.185510015 -0.01565002 -0.006796030
job       -0.02187245  1.000000000  0.062045485  0.166706724  0.026003635 -0.125362813 -0.03300392  0.004093753
marital   -0.40325806  0.062045485  1.000000000  0.108576125  0.006912736 -0.016095882 -0.04689252  0.012363967
education -0.10682933  0.166706724  0.108576125  1.000000000  0.071494943 -0.090790237 -0.04857353  0.002275432
balance    0.11907269  0.026003635  0.006912736  0.071494943  1.000000000 -0.078895715 -0.11407748  0.038269635
housing   -0.18551001 -0.125362813 -0.016095882 -0.090790237 -0.078895715  1.000000000  0.04132287  0.005878007
loan      -0.01565002 -0.033003921 -0.046892524 -0.048573533 -0.114077479  0.041322866  1.00000000 -0.013354345
duration  -0.00679603  0.004093753  0.012363967  0.002275432  0.038269635  0.005878007 -0.01335434  1.000000000
>
```

Correlation chart of the attributes to help choose independent and dependent variables
Based on the table we chose the 'age' as the dependent variable and 'balance' and 'marital status' as independent variable

III. Regression model implementation

```r
#lm function
model <- lm(balance~age + education, data1)
summary(model)
par(mfrow =c(2,2))
plot(model)
par(mfrow =c(1,1))
```

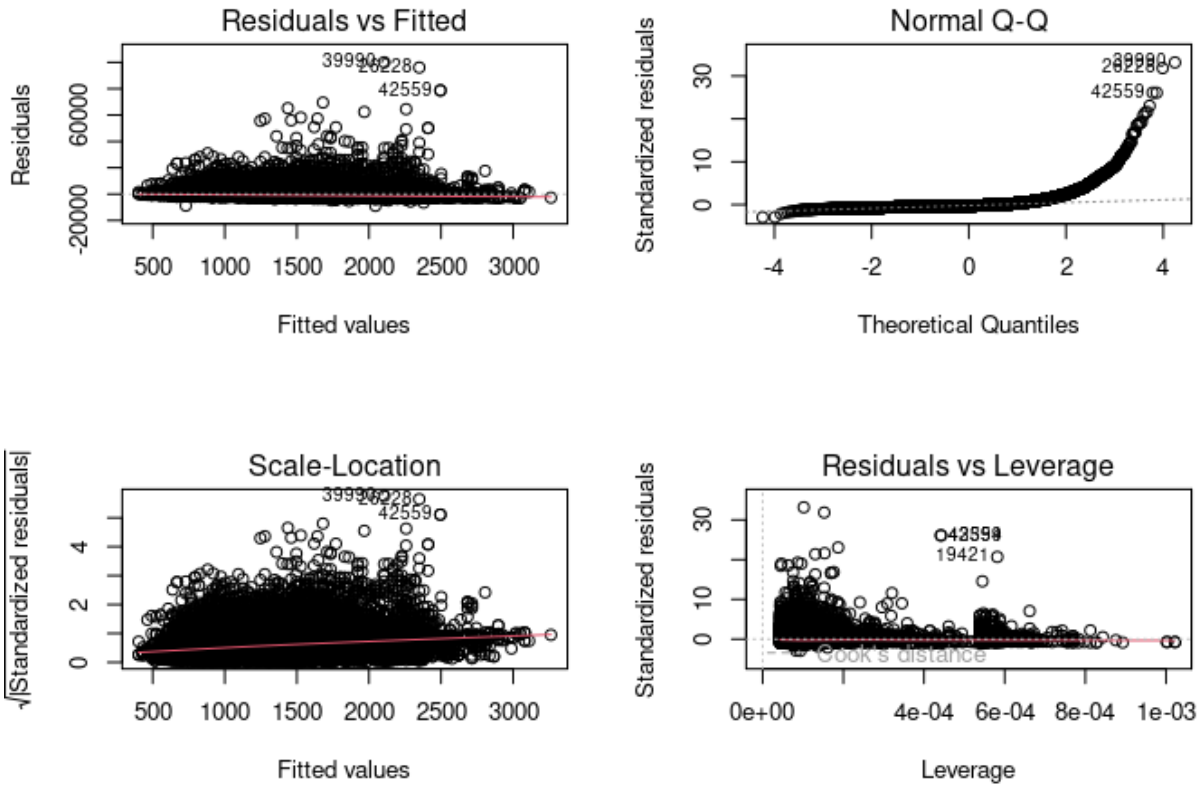
Commands to get linear regression model and get model diagnostic plots

```
Call:
lm(formula = balance ~ age + education, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
 -8892  -1245   -759     99 100021

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -145.245     72.568  -2.002   0.0453 *
age                 30.441      1.368  22.250  < 2e-16 ***
educationsecondary  83.572     42.258   1.978   0.0480 *
educationtertiary  698.390     45.675  15.291  < 2e-16 ***
educationunknown   317.054     78.946   4.016 5.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3017 on 45206 degrees of freedom
Multiple R-squared:  0.01843,   Adjusted R-squared:  0.01835
F-statistic: 212.2 on 4 and 45206 DF,  p-value: < 2.2e-16
```

Model summary

Based on the model summary, there is a multiple R-squared score of 0.01843. This means that approximately 1.824% of variation in ‘balance’ can be explained by the model (age + education)

IV. Model Diagnostic Plots



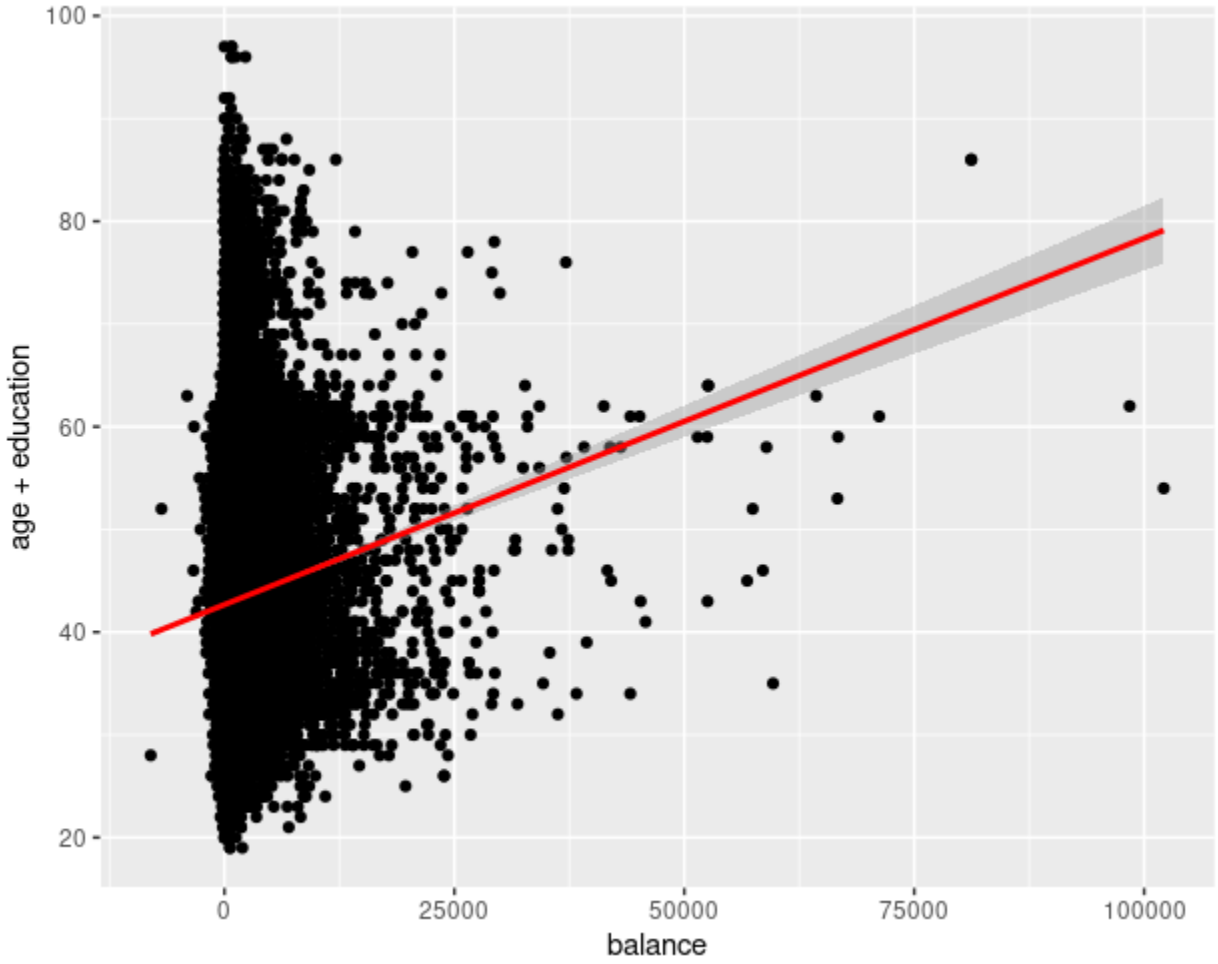
Residuals vs fitted - checks the linear relationship assumption, should be randomly scattered pattern of points with no clear trend or pattern. But in the model it shows that it is not randomly scattered and most points are at the bottom of the table. This could be because there are alot of data considered

Normal Q-Q - shows if the residuals are normally distributed. Should be close to the diagonal line but the table shows it's not a diagonal line and slopes up towards the end.

Scale-location - checks the constant error variance assumption. points should also be randomly distributed in the chart, but it shows that it is not randomly scattered and most points are at the bottom of the table.

Residuals vs Leverage - helps identify influential outliers that have a large impact on the regression line. The model was not able to detect any because there is no Chrysler Imperial line.

V. Regression Line



We picked out balance as a dependent variable and age + education as independent variables. This is to answer the question of how does loan balance relate to age and education level. Based on the regression graph taken it shows that:

- Most people have remaining balances below 25,000
- People ages 30-60 have the most remaining balance.
- There are only 3 outliers with a remaining balance of around 75,000 to 100,000 around the age of 50 - 90 years old.