Domača naloga 2

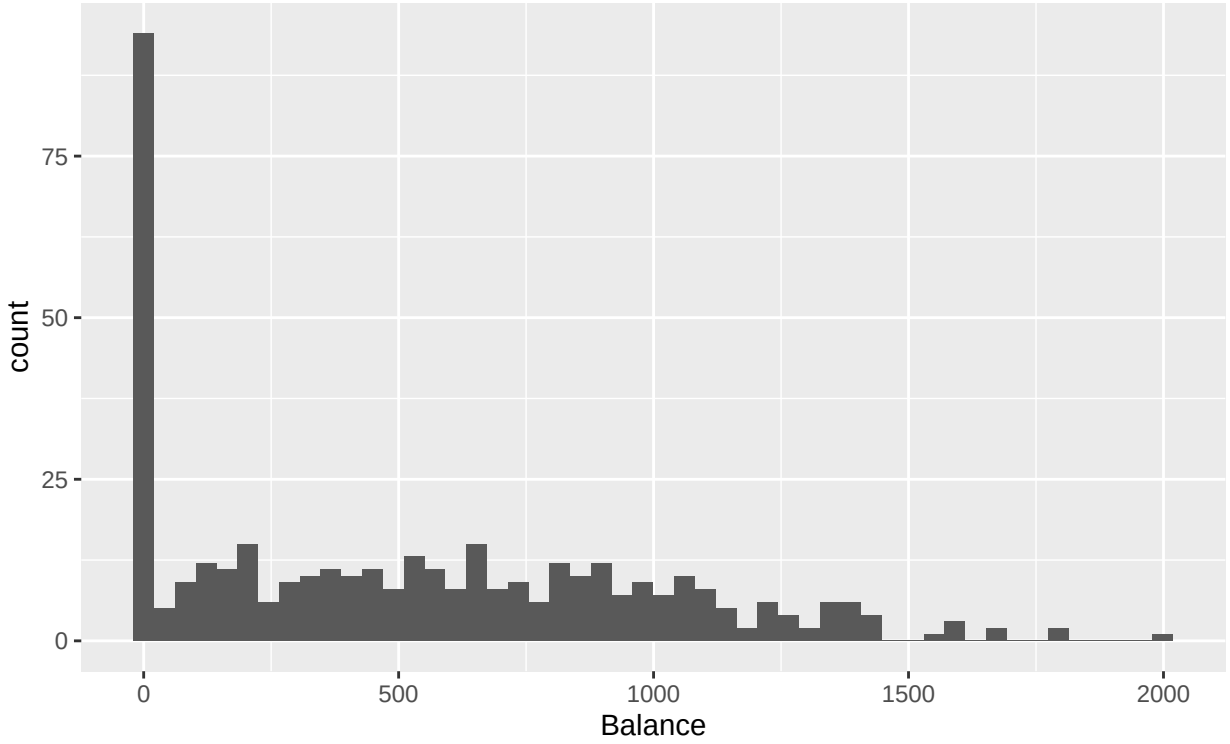
Linearni modeli

Alen Kahteran

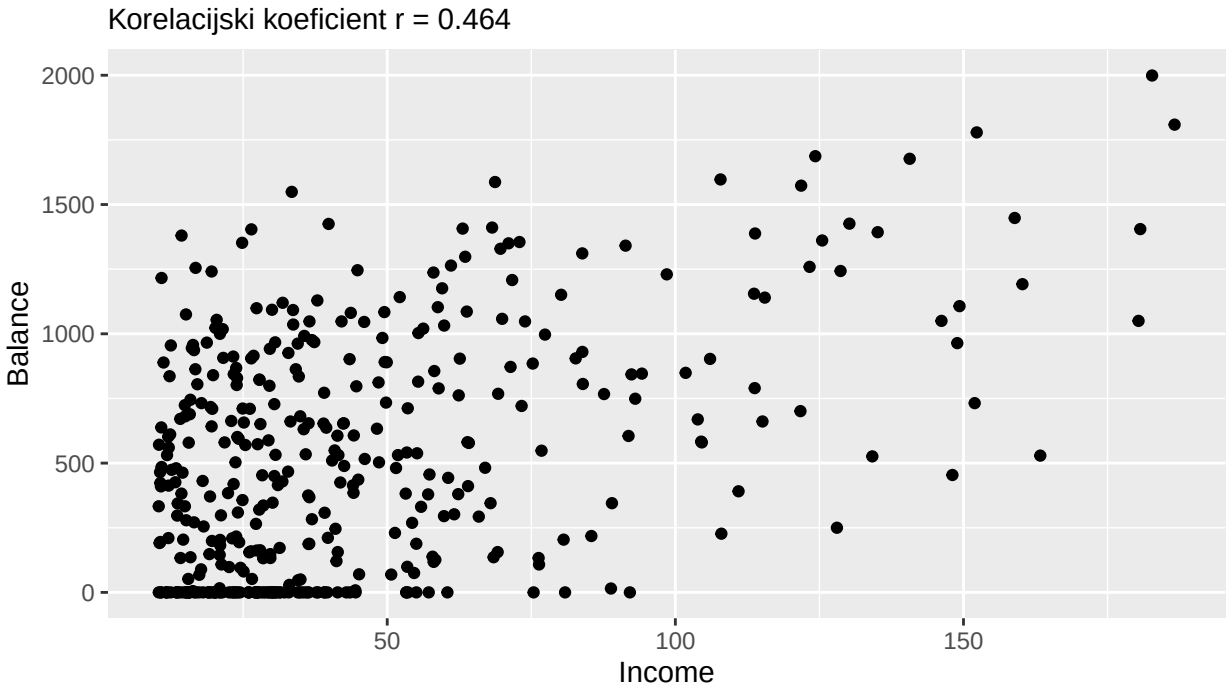
30. 11. 2020

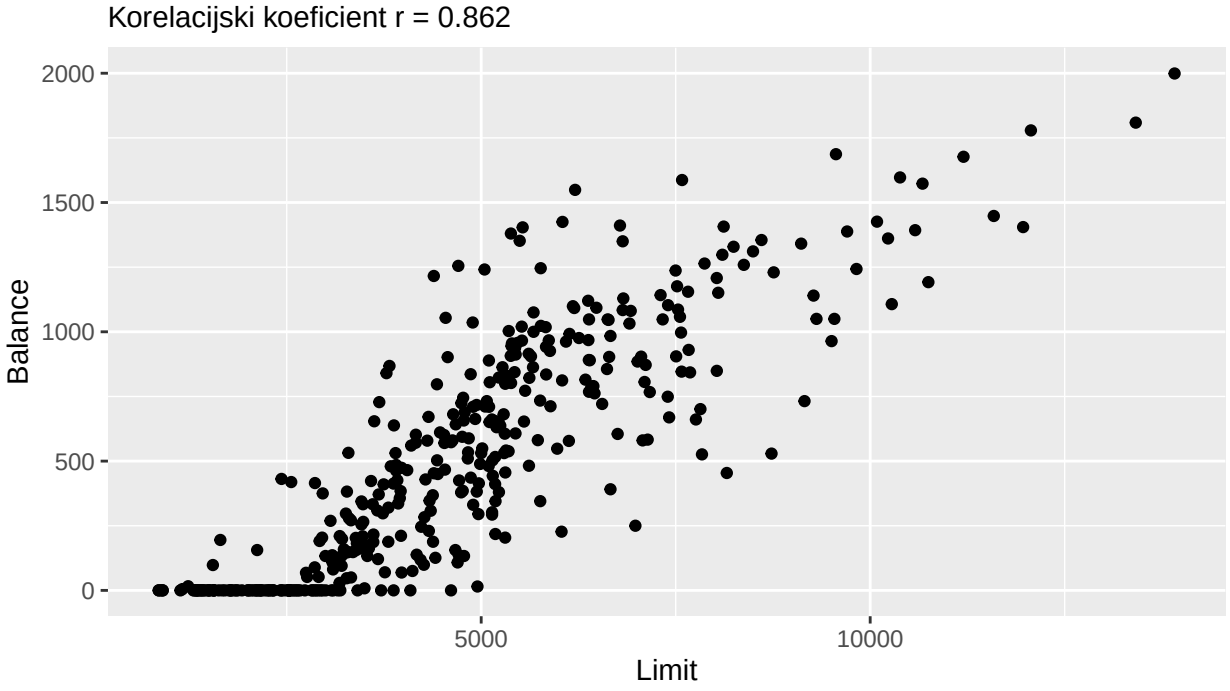
Analiza `Balance` glede na ostale spremenljivke

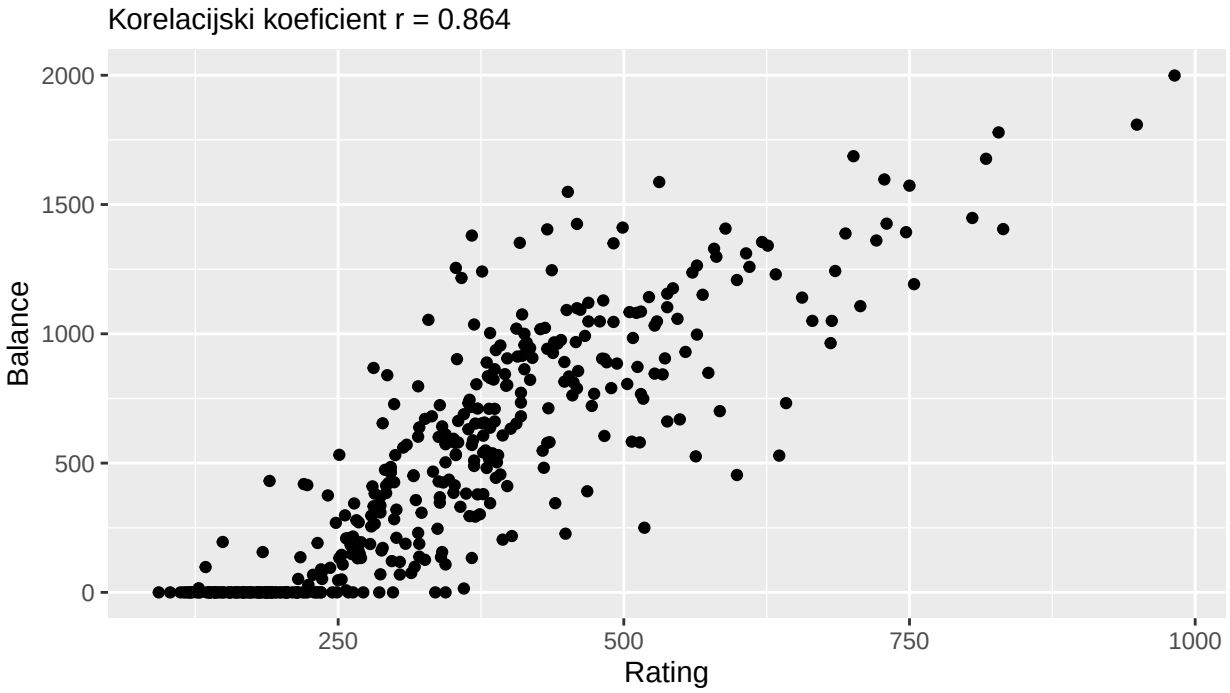
Želimo si zgraditi linearen model stanja na kreditni kartici (`Balance`), v odvisnosti od ostalih spremenljivk. Najprej je potrebno pregledati podatke, ki jih imamo.

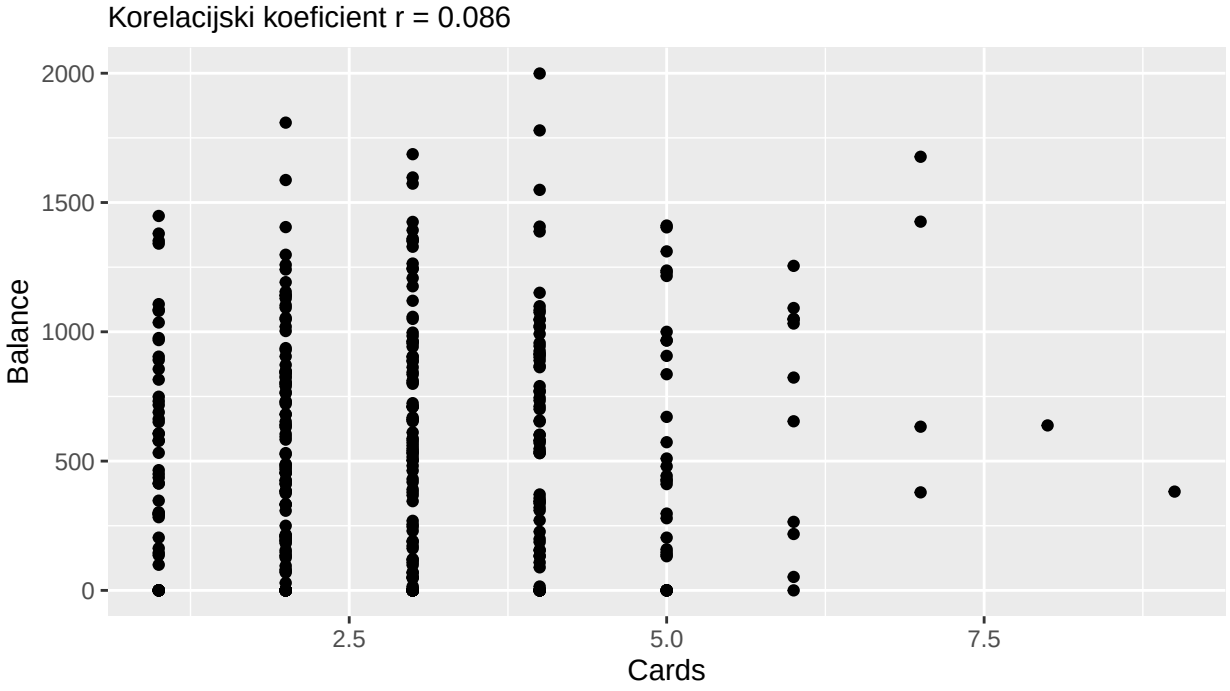


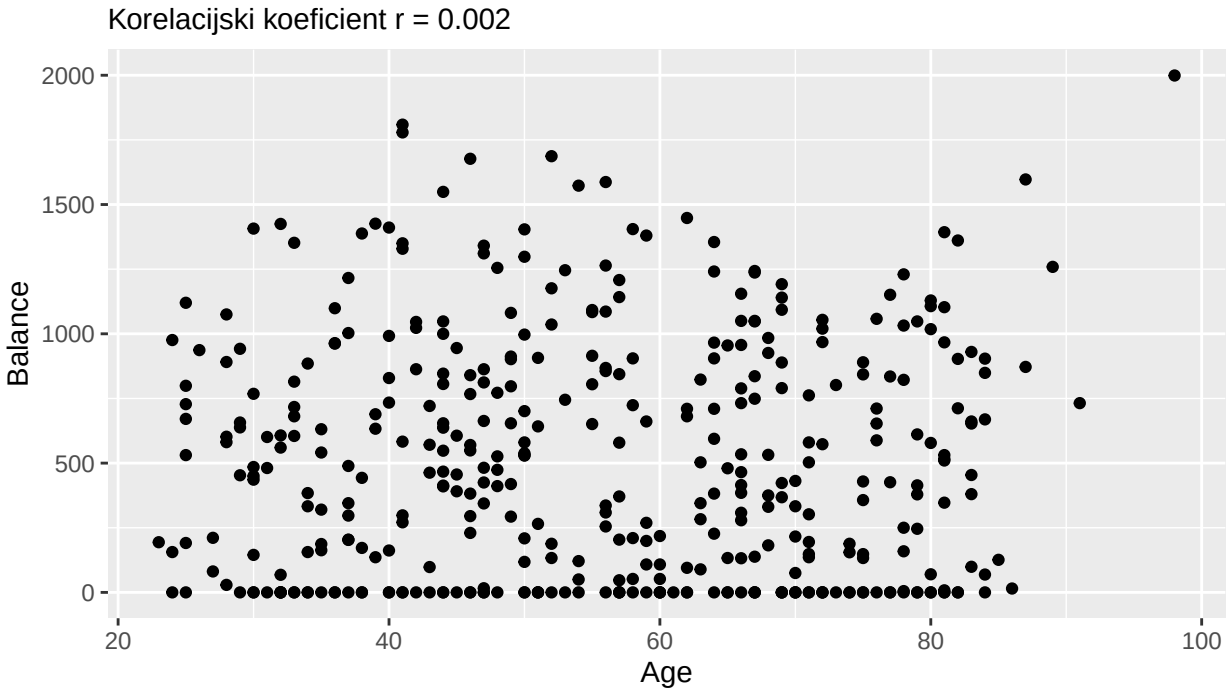
Vidimo, da je veliko takšnih, ki imajo povprečno stanje na kreditni kartici enako 0. Poglejmo si odvisnosti od ostalih številskih spremenljivk.

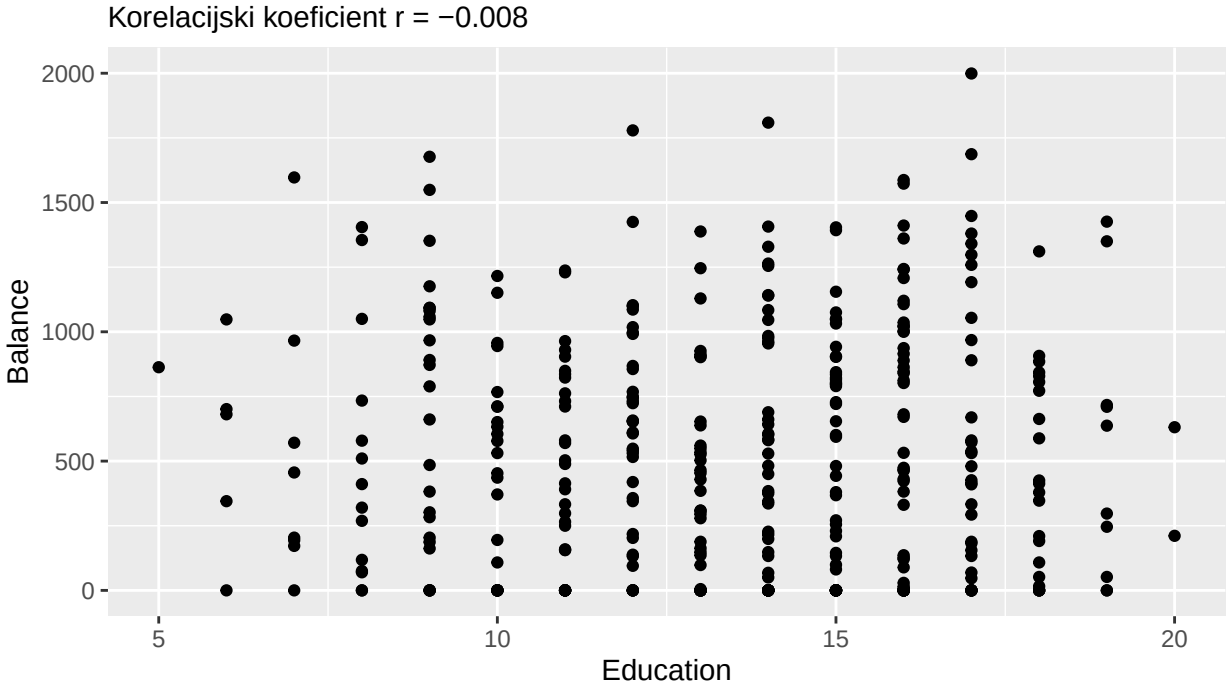

Korelacijski koeficient r = 0.464
Balance
2000
1500
1000
500
0
50
100
150
Income


Korelacijski koeficient r = 0.862
Balance
2000
1500
1000
500
0
5000
10000
Limit

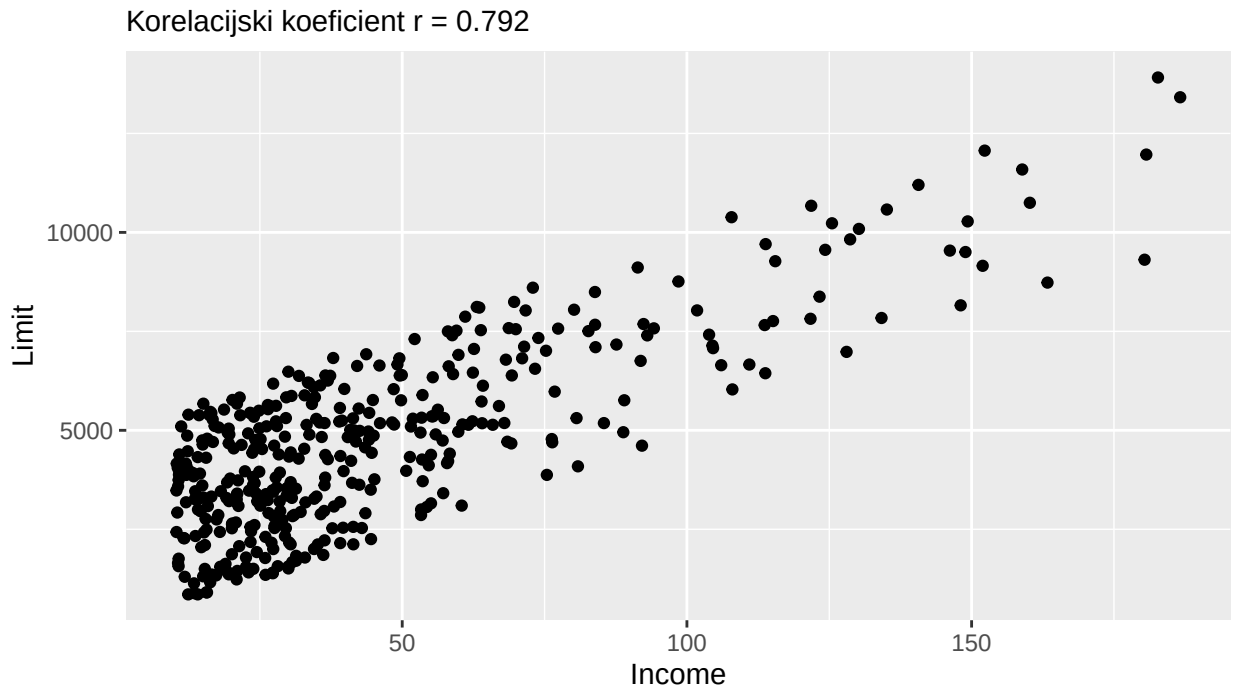

Korelacijski koeficient r = 0.864
Balance
Rating
2000
1500
1000
500
0
250
500
750
1000

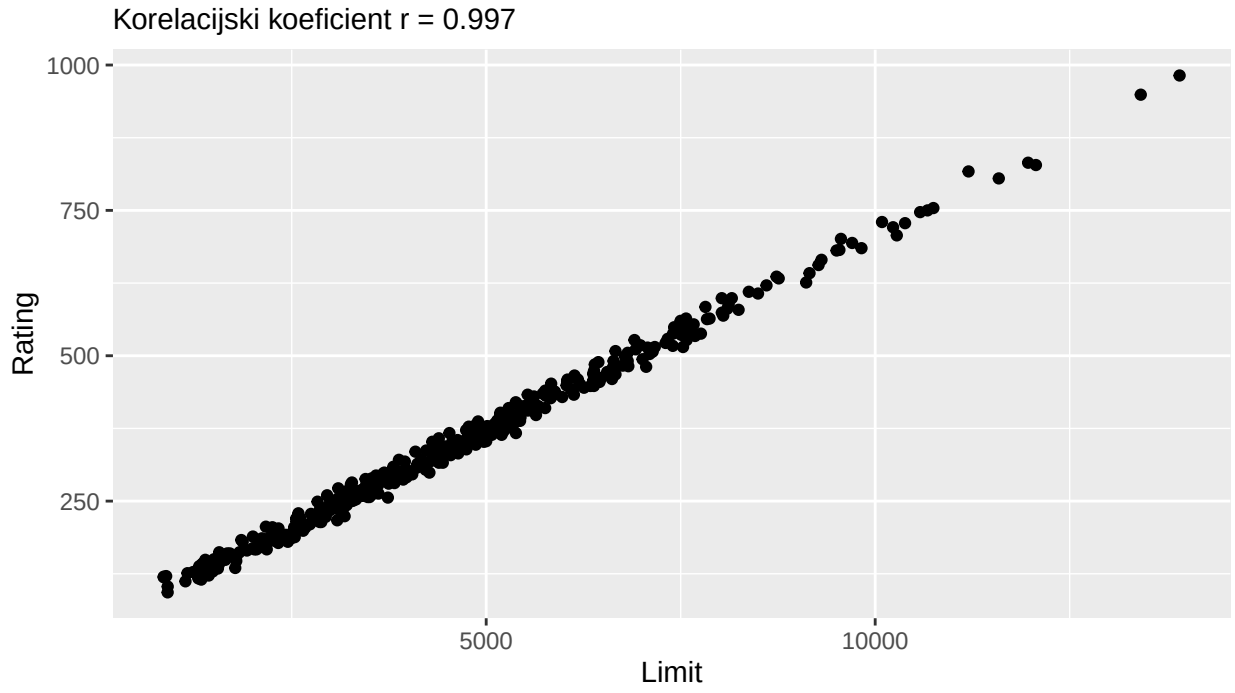

Korelacijski koeficient r = 0.086
Balance
Cards
2000
1500
1000
500
0
2.5
5.0
7.5





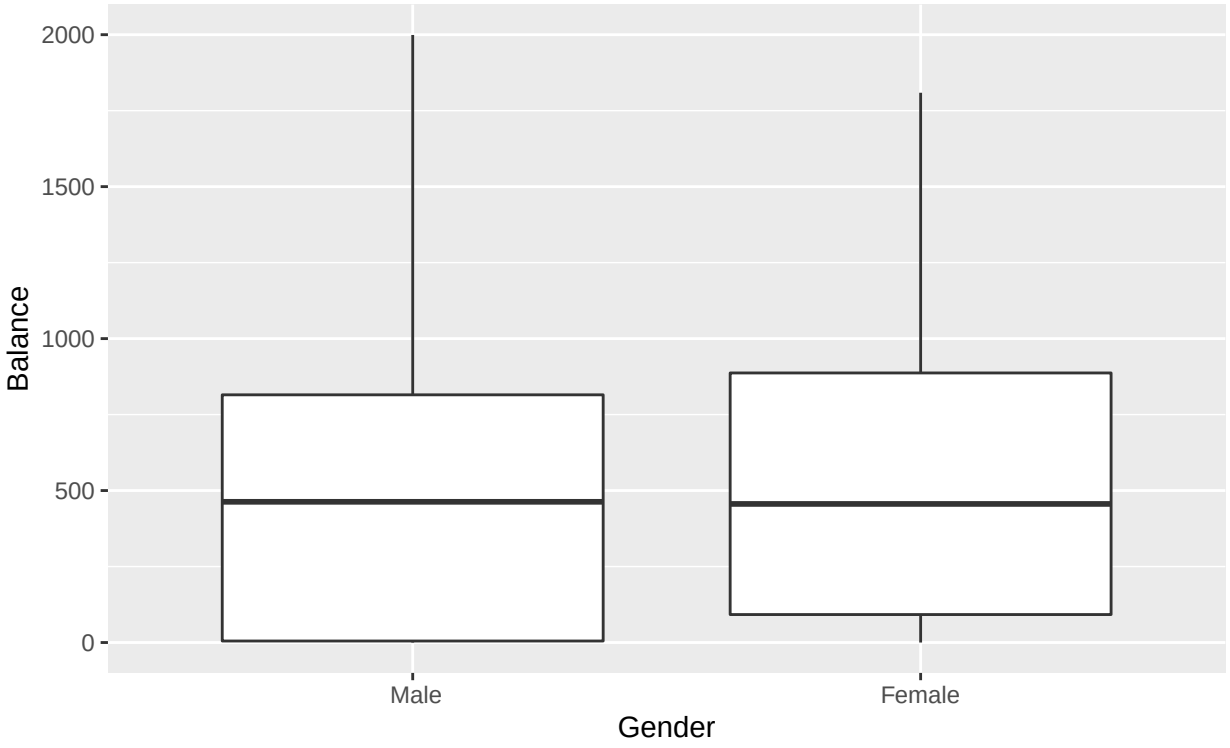
Hitro vidimo, da je spremenljivka `Rating` zelo podobna spremenljivki `Limit`. Kot vidimo pri sliki spremenljivke `Limit`, ima ta zelo visok korelacijski koeficient, kar pomeni da ima odvisna spremenljivka `Balance` visoko linearno povezanost z neodvisno spremenljivko `Limit`. Predpostavljam, da je spremenljivka `Rating`, ali linearna kombinacije spremenljivke `Limit` s katero od ostalih spremenljivk, ali pa linearen model vseh spremenljivk. Poglejmo si še nekaj drugih kombinacij, ki bi nam lahko pokazale kakšno pomembno informacijo, kot npr. `Limit` v odvisnosti od `Income` ali pa `Rating` v odvisnosti od `Limit` (da preverimo našo domnevo).

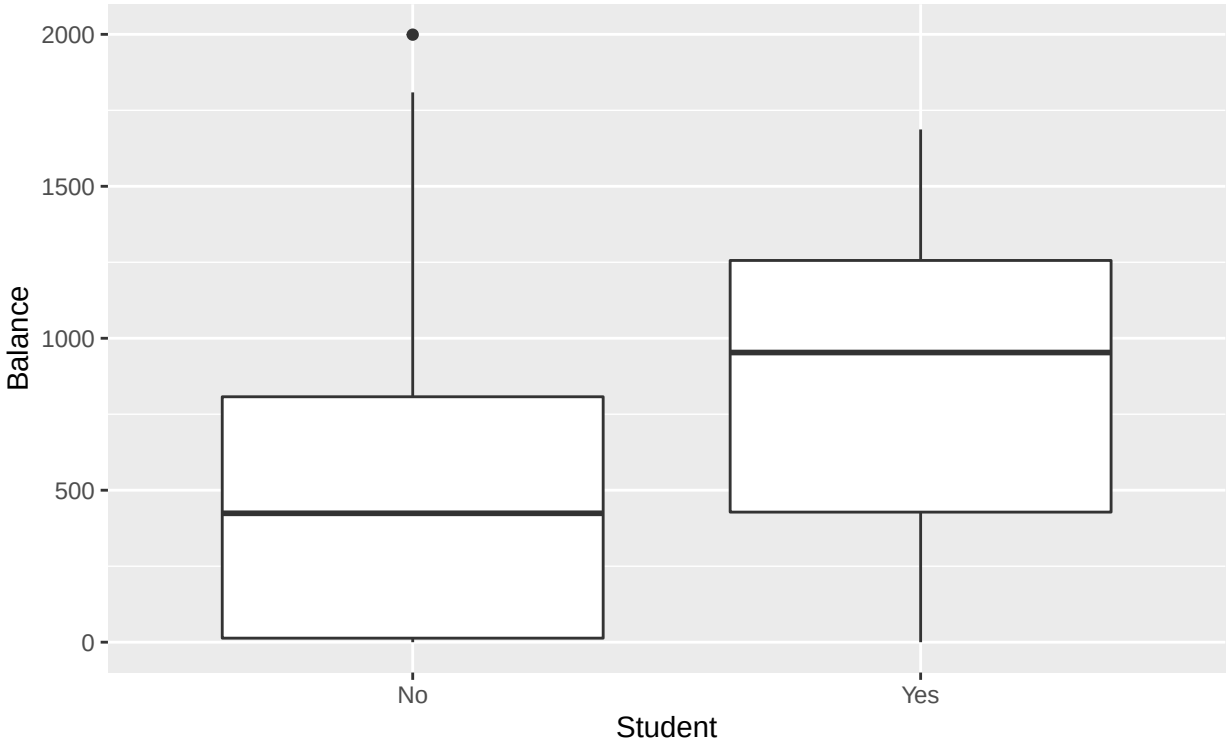


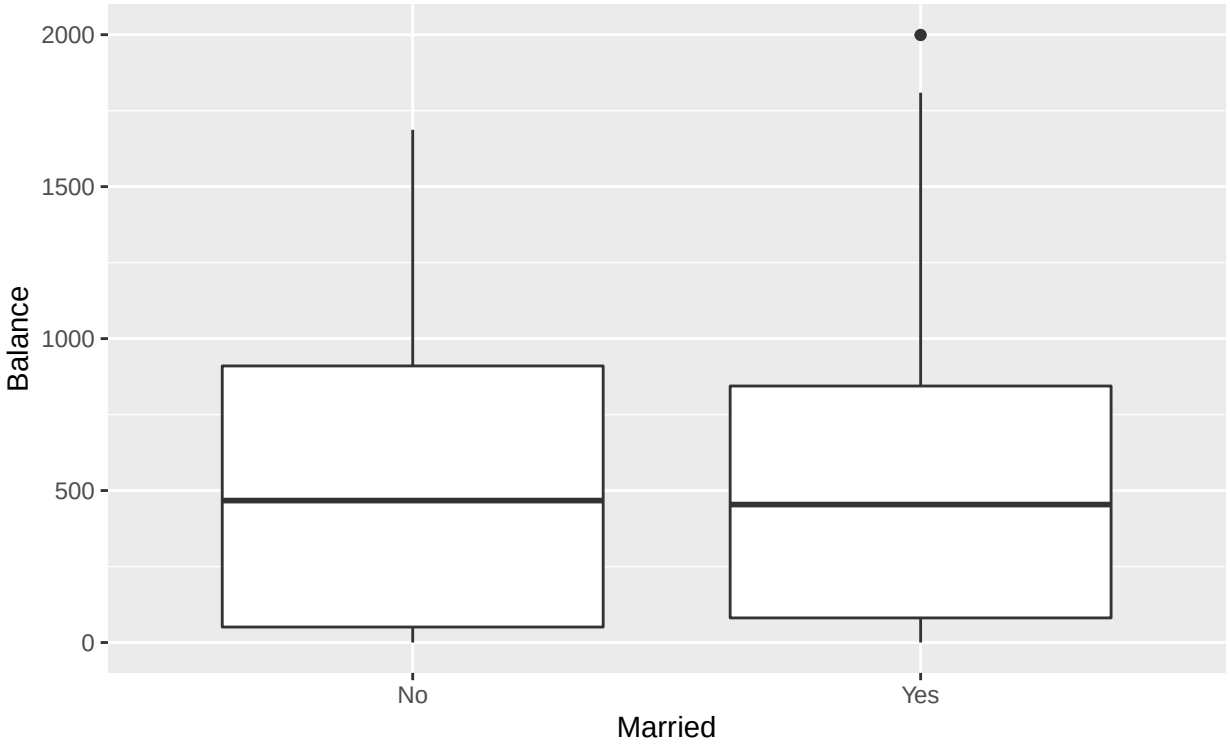


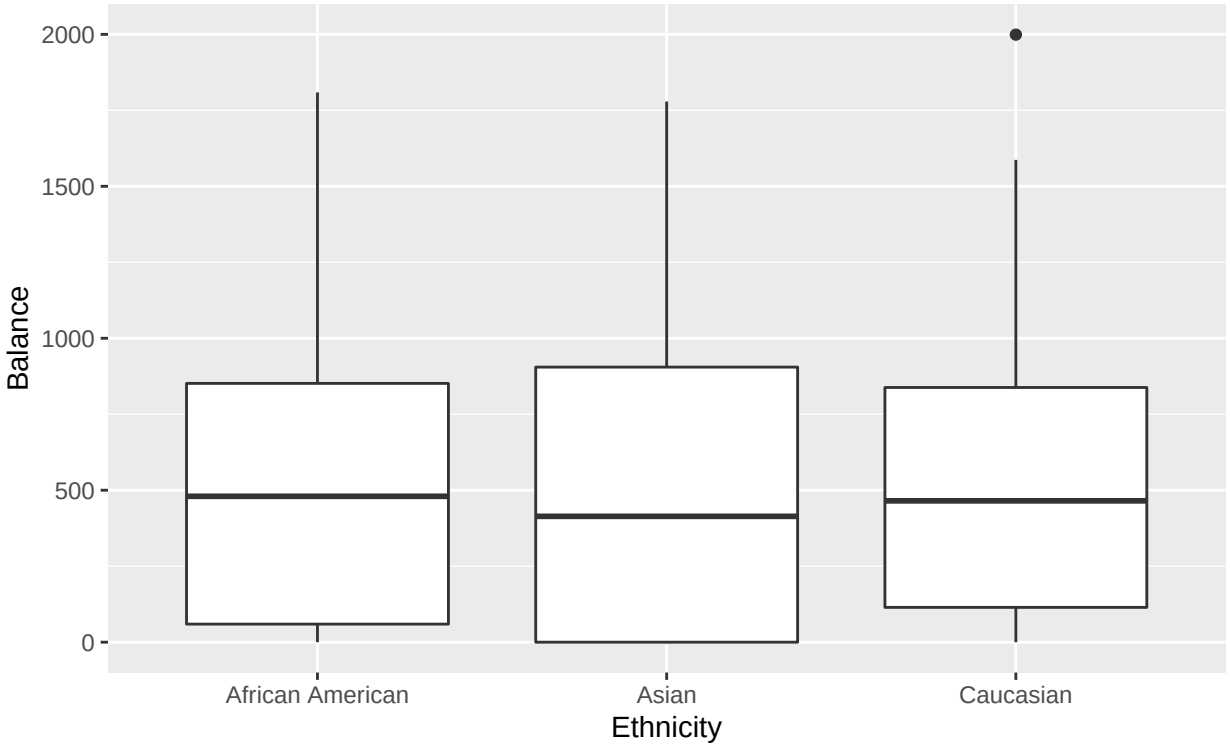
Zelo hitro vidimo, da je pravzaprav `Limit` linearno odvisen od višine `Income`. Se pravi banka oz. kreditodajalec se skoraj zagotovo vsaj na podlagi prihodkov odloči kolikšen bo `Limit` za vsakega posameznika (z določenim oknom, najverjetneje glede na potrebe posameznika). Poleg tega pa vidimo, kar smo prej domnevali, da sta `Rating` in `Limit` zelo linearno povezana.

Poglejmo si še morebitne odvisnosti od neštevilskih spremenljivk (`Gender`, `Student`, `Married`, `Ethnicity`)

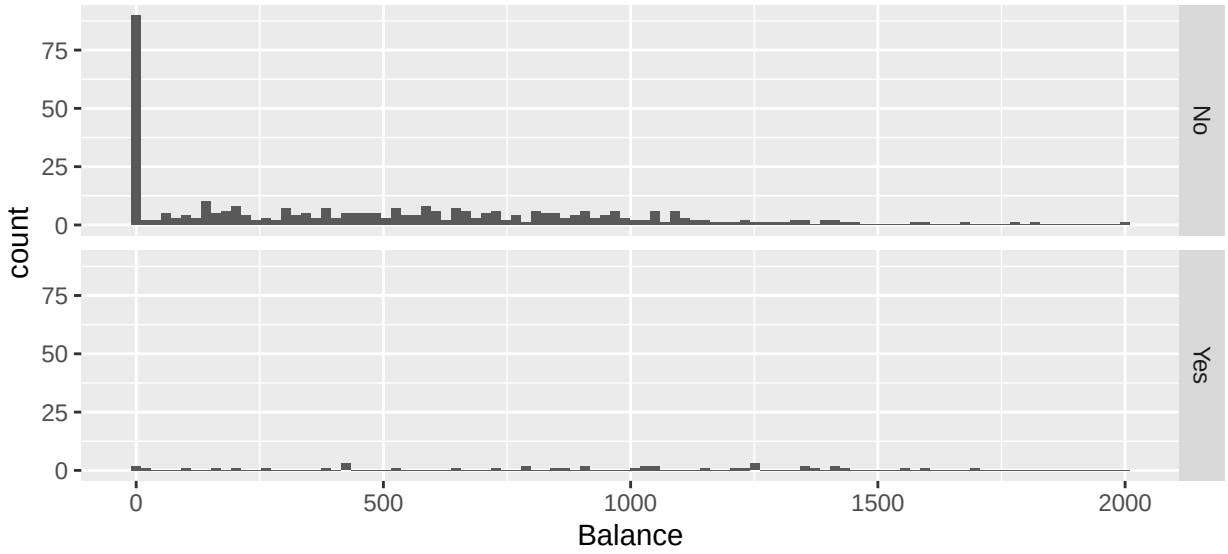

2000
1500
1000
500
0
Balance
Male
Female
Gender


2000
1500
1000
500
0
Balance
No
Yes
Student

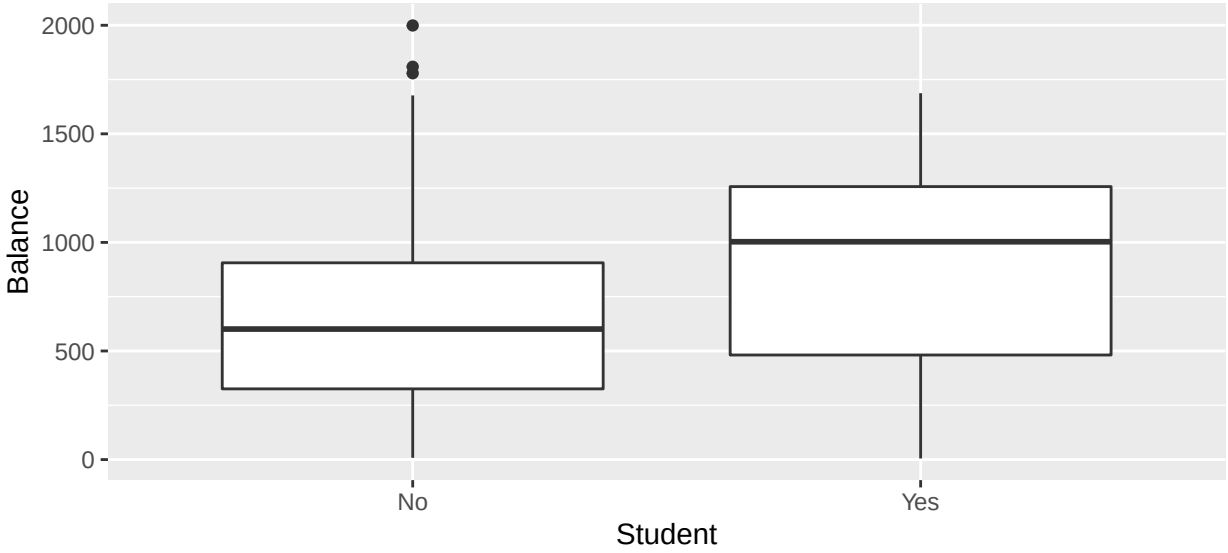


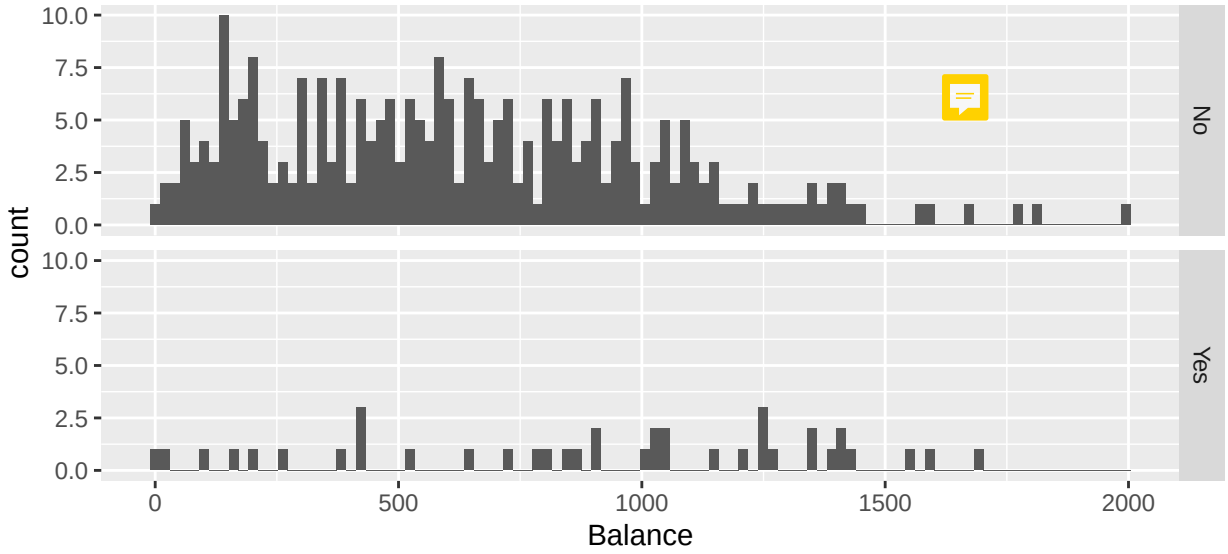


Tu vidimo da večjih razlik med določenimi vrednostmi v spremenljivkah ni, z izjemo tega ali je nekdo študent ali ne. Poglejmo si za vsak slučaj še porazdelitev, da vidimo če kaj preveč izstopa.



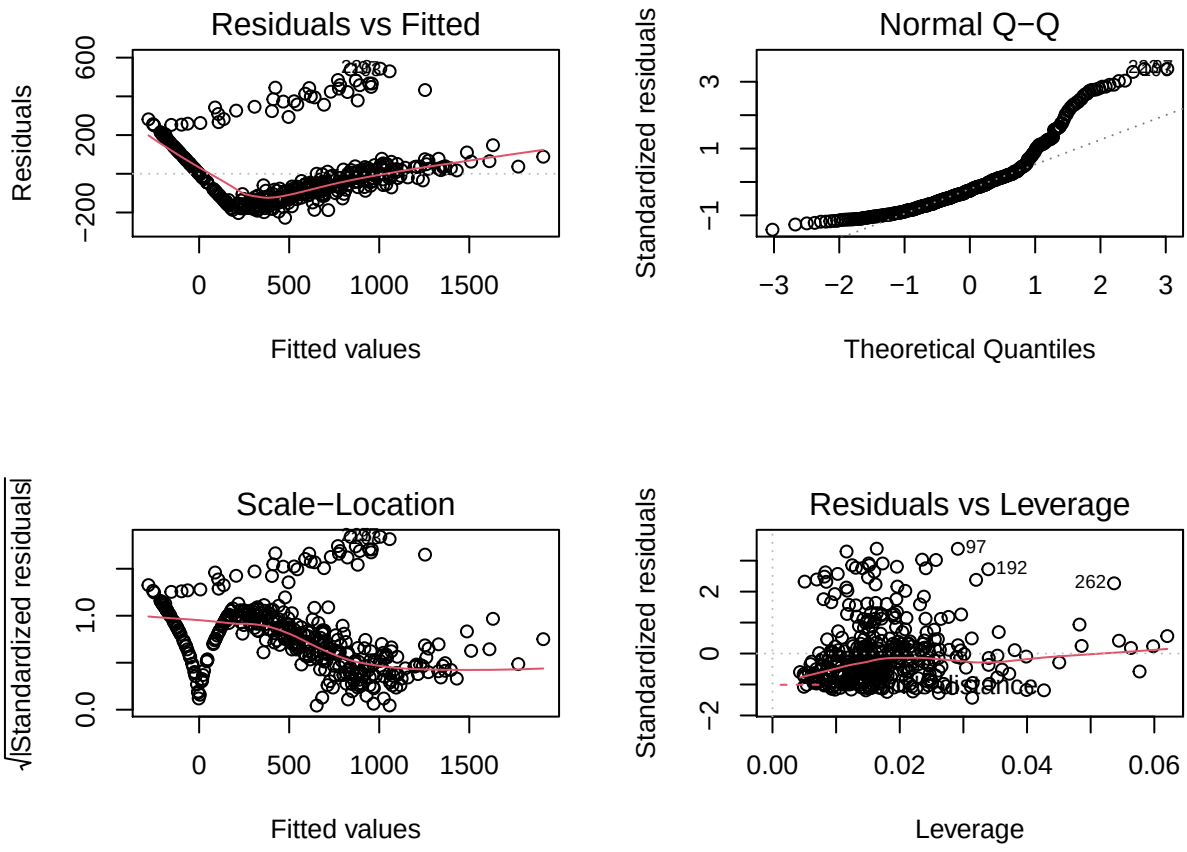
Vidimo da je ogromno takih, ki niso študentje in je povprečno stanje na kreditni kartici enako 0. To smo videli že pri prvi porazdelitvi spremenljivke `Balance`. Če narišemo enako sliko škatle z ročaji, po tem ko smo izločili tiste, ki imajo stanje na kartici enako 0, vidimo da so razlike še vedno opazne (pri histogramu so te razlike manj opazne). Intuitivno bi težko rekel, da to v splošnem drži, saj ponavadi študentje težko ob študiju še delajo, ter privarčujejo. Potrebno se je res dobro vprašati kaj `Balance` pravzaprav predstavlja. A predstavlja "porabo" na kreditni kartici, ali predstavlja koliko je še "prostega" na njej, saj je to pomembna razlika in bi si s tem lažje razložili morebitna odstopanja. Poleg tega, sumim da to odstopanje lahko izvira tudi iz tega, da je število zapisov kjer je nekdo študent, majhno (št. študentov - 40)





Linearna regresija

Kakorkoli, potrebno je narediti linearen regresijski model `Balance` na ostale številske spremenljivke.



Iz grafov preverimo, če so izpolnjene predpostavke za linearno regresijo.

Hitro vidimo (levo zgoraj), da podatki najverjetneje niso linearno povezani. Mogoče bi bila smiselna kakšna transformacija za "spodnji" del, da bi dosegli linearnost, v primeru če bi "zgornji" del privzeli kot regresijske osamelce.

Desno zgoraj je videti, da standardizirani ostanki niso porazdeljeni normalno, kar zopet predstavlja problem.

Spodaj levo opazimo pojav heteroskedastičnosti, kar spet krši eno izmed predpostavk linearne regresije.

Spodaj desno načeloma ne vidimo težav, saj pravzaprav ne vidimo območja Cook-ove razdalje, torej predpostavljam da so vse točke nižje od "kritičnih" Cook-ovih razdalj. Torej nobena izmed točk ni pretirano vplivna.

V primeru ko bi predpostavke držale in bi tudi imeli korektno linearno regresijo, bi si lahko naslednje koeficiente interpretirali na naslednji način

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-477.958	55.065	-8.680	0.000
Credit$Income	-7.558	0.382	-19.766	0.000
Credit$Limit	0.126	0.053	2.373	0.018
Credit$Rating	2.063	0.794	2.598	0.010
Credit$Cards	11.592	7.067	1.640	0.102
Credit$Age	-0.892	0.478	-1.867	0.063
Credit$Education	1.998	2.600	0.769	0.443

Ko se recimo `Income` poveča za 10000 dolarjev, se `Balance` zmanjša za ~ 7.56. Podobno velja tudi za ostale spremenljivke, ko se določena spremenljivka spremeni za eno enoto se `Balance` spremeni za njen `Estimate`.