NOTES 3

Exploratory Data Analysis & Data Visualization

Acknowledgement: some of the contents are borrowed with or without modification from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R.Tibshirani.

Case Study: Credit Balance

Credit card companies make money when their customers use the issued credit card. The average account balance is useful to credit card companies, since it indicates not only the potential revenues, e.g. fees associated with transactions, but also the potential risk, e.g., the probability of default.

To estimate the balance from each account, credit card companies should identify what these factors are. A credit card company has collected a dataset for its customers in 2010. The dataset contains 400 records and 8 attributes.

Data: credit2010.csv

Income	Limit	Rating	YearofBirth	Gender	Student	Married	Balance
14.891	3600	283	1976	Male	No	Yes	333
106.025	6600	483	1928	Female	Yes	Yes	903
104.593	7000	514	1939	Male	No	No	580
148.924	9500	681	1974	Female	No	No	964
55.882	4800	357	1942	Male	No	Yes	331
80.18	8000	569	1933	Male	No	No	1151
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Data Description

- Balance: the average balance that the customer has remaining after making the monthly payment
- YearofBirth: a customer's year of birth
- Income: a customer's annual income, in thousands of dollars
- Limit: a customer's credit limit
- Rating: credit rating from the bank
- Gender: customer's gender (Female or Male)
- Student: student or not (Yes or No)
- Married: marital status (Yes or No)

General EDA Approaches

- Data restructuring and Data transformation
- Summary Statistics: explore location/scale information
- Data Visualization-Plots:
  - Explore univariate distribution: e.g. Histogram
  - Explore bivariate distribution and correlation structure: e.g. Pairwise Scatterplots, Box plot

X is continuous || X is categorical

Data Structure

- How many observations are there?
  - dim(x)
- Distinguish between categorical and continuous variables.
  - str(x)
- Would some benefit from conversion to a different format?
  - e.g. convert continuous attribute to categorical one
  - as.factor(x)

Data Processing

- Any missing values? Don't delete missing numbers
  - Reasons for missing, e.g. Certain questions only asked to respondents who indicate they are employed
  - If there is no dependence then values might be imputed, e.g. Deletion, Mean substitution.
- Data transformation: log for skewed variables

- Data restructuring: derive variables that are more meaningful
  - YearofBirth is a less informative indicator for Balance.

```
>mean(credit[,4])
[1] 1954.332
```

  - As the data was obtained in the same year, 2010, we can make and include a new attribute Age and exclude YearofBirth.

```
Age<-2010-credit[,5]
```

→ the yr that data is collected

Summary Statistics

- Measures of location: mean, median
- Measures of spread: standard deviation, range (min, max) is useful too.
- For categorical variables: counts and proportions tables.

Summary Statistics---Raw Results

The result must be presented in tables !!! (categorical & continuous data should be separate! (2 tables)

	Income	Limit	Rating	Balance	Age
Mean	45.21889	4687	354.94	520.015	55.6675
Min	10.354	800.000	93.000	0.000	23.000
Max	186.634	13900.000	982.000	1999.000	98.000
Median	33.1155	4600.0000	344.0000	459.5000	56.0000
SD	35.24427	2307.735928	154.7241	459.7589	17.24981

MUST have same format

Table Presentation:

- Numbers should be rounded up to two digits. Numbers in each attribute should in the same format.
- Re-order rows and/or columns to make message clearer.
- Add a clear, self-explanatory title.

spread out

Summary Statistics

	Income	Limit	Rating	Age	Balance
Mean	45.22	4687	354.94	55.67	520.02
Min	10.35	800	93.00	23.00	0.00
Max	186.63	13900	982.00	98.00	1999.00
Median	33.12	4600	344.00	56.00	459.50
SD	35.24	2307	154.72	17.25	459.76

Table 1. Summary Statistics of Continuous Variables
Number of Observations =400

Must have Sample Size & Table Title

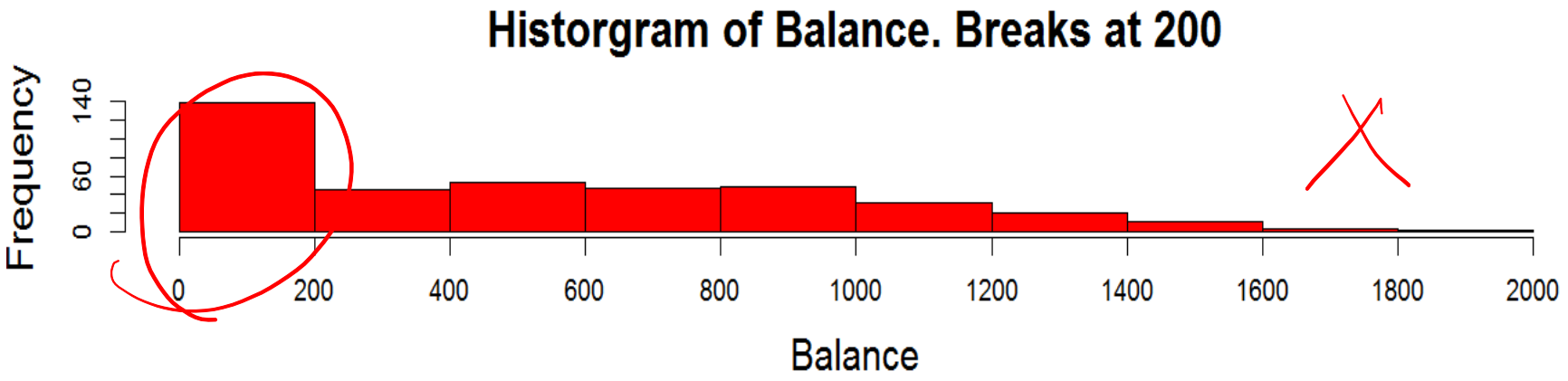
Summary Statistics

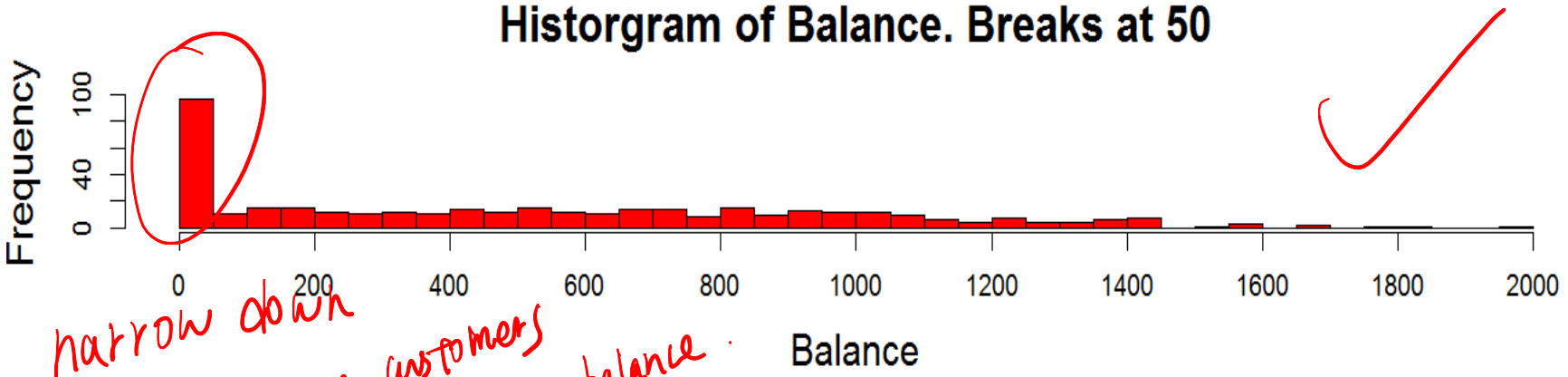
Gender		Student		Married	
Male	193(0.48)	**No**	360(0.90)	**No**	155(0.39)
Female	207(0.52)	**Yes**	40(0.10)	**Yes**	245(0.61)

Table 2. Counts/ Proportions for Gender, Student, Married and Cards
Number of Observations =400

Categorical tables

Histogram (Univariate)





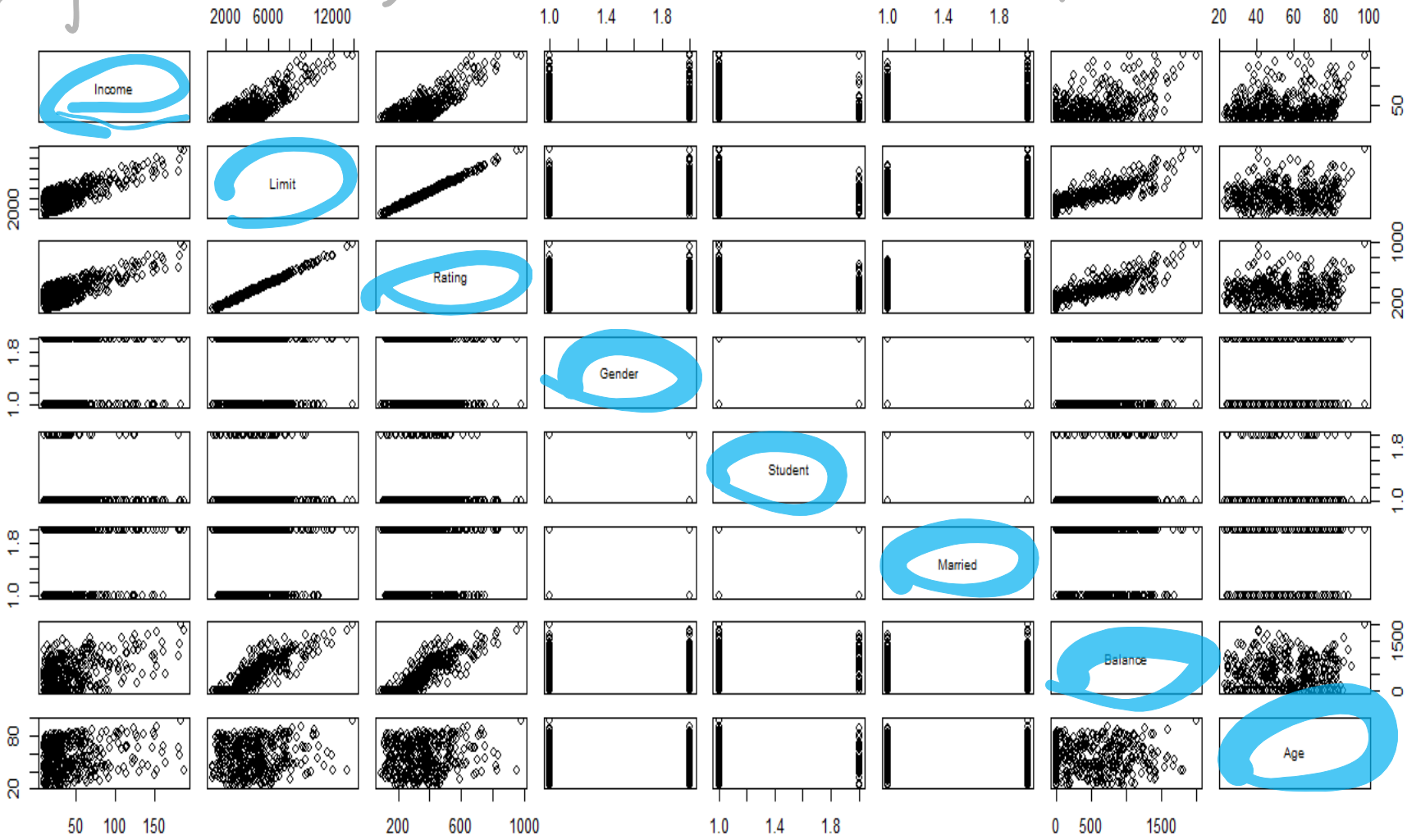
narrow down
→ how many customers have 0 balance.

Plot Presentation

- Add a clear, self-explanatory title.
- Label axes, giving units.
- Carefully choose scale on the difference axes.
- Prudent use of color and symbol.

Pairwise Scatterplot (Bivariate)

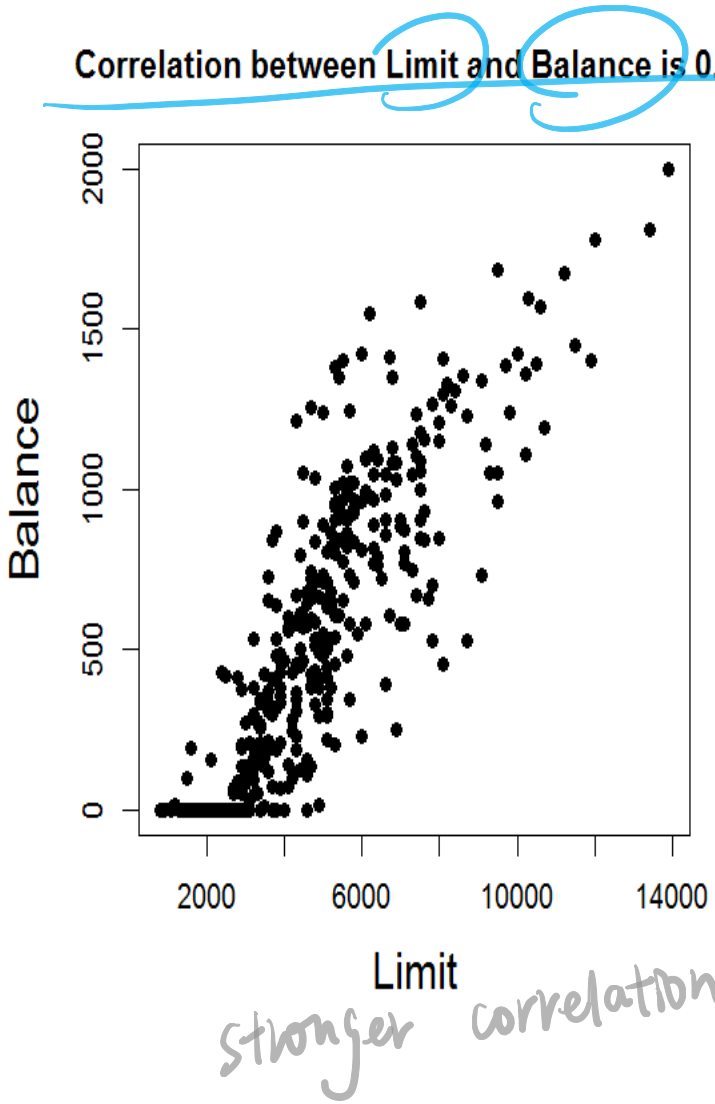
- If there's too many variables, then that's too much to read



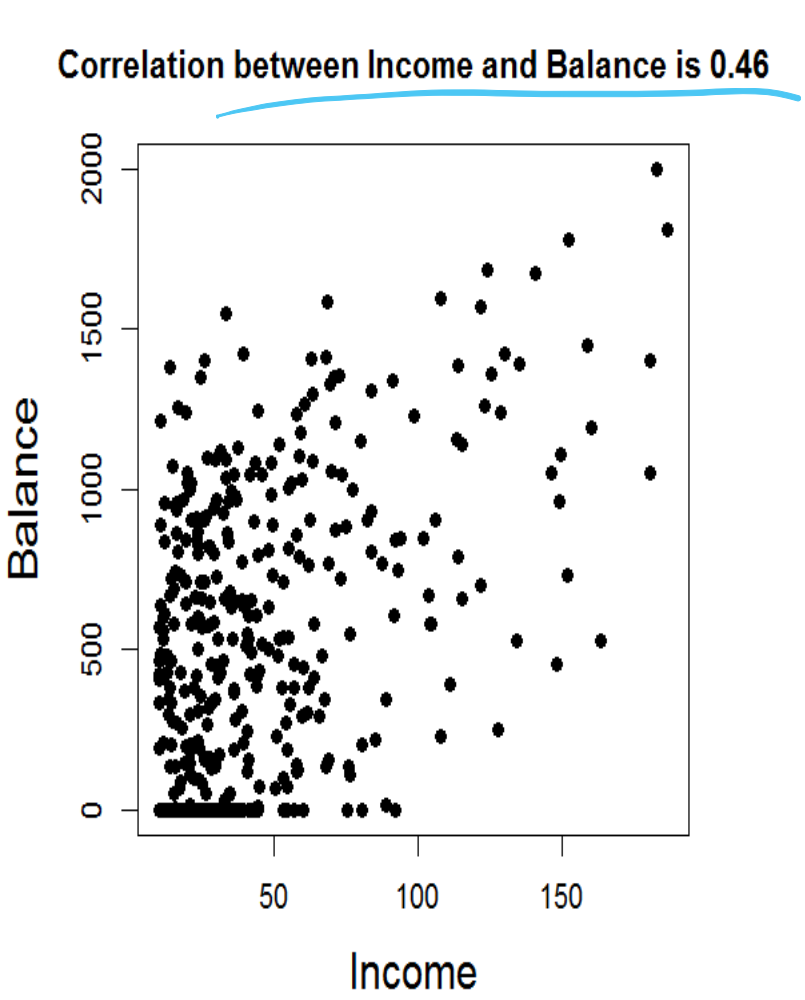
Pairwise Scatterplot for All Variables

Scatterplot (Bivariate)

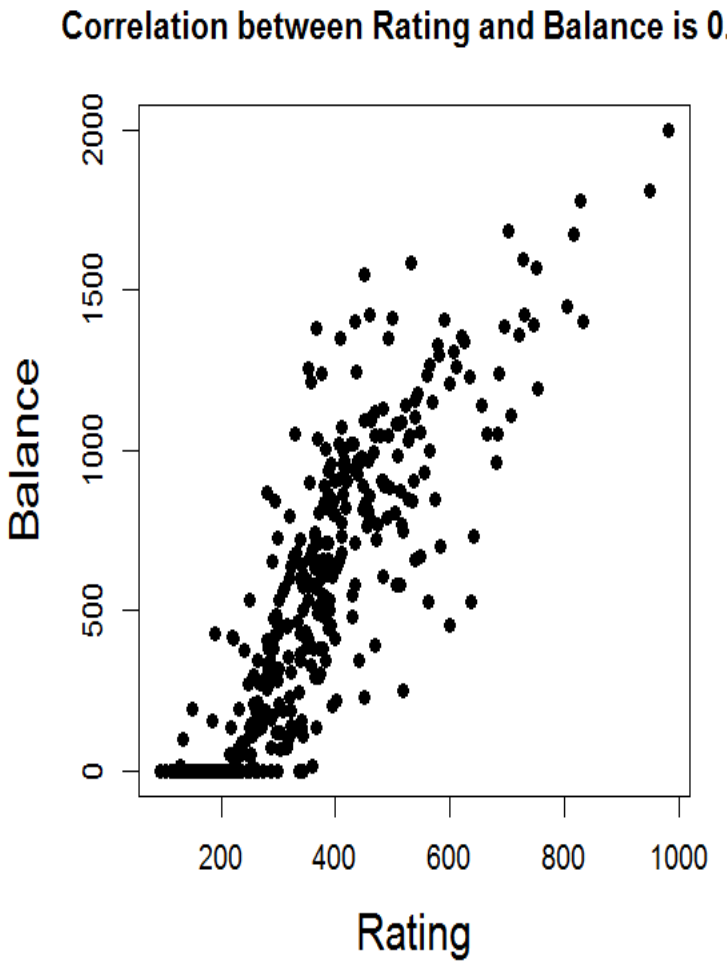
① potential correlation
② positive/negative relation

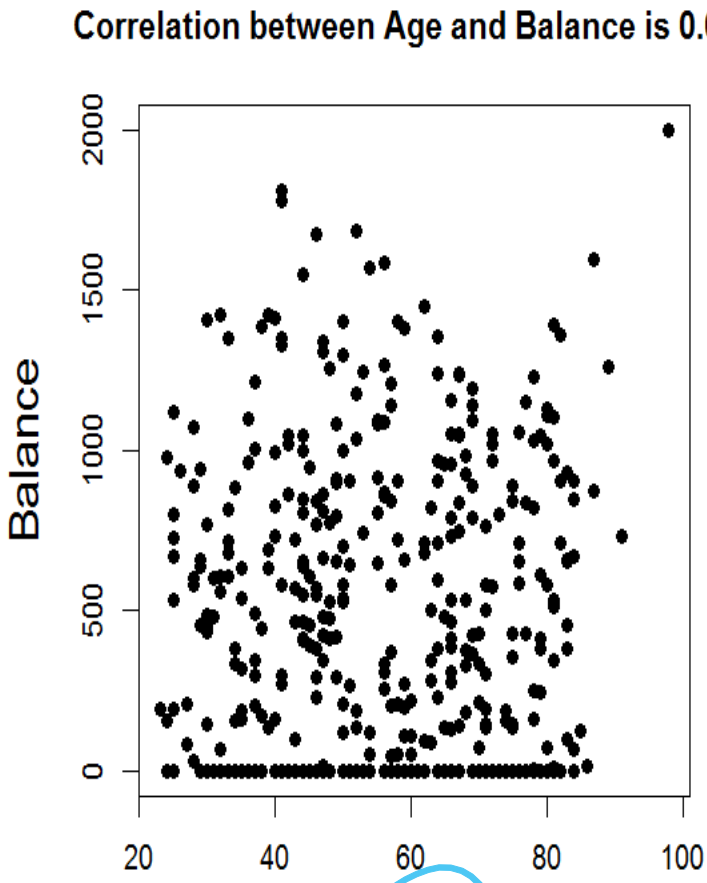


stronger correlation



Scatterplot (Bivariate)

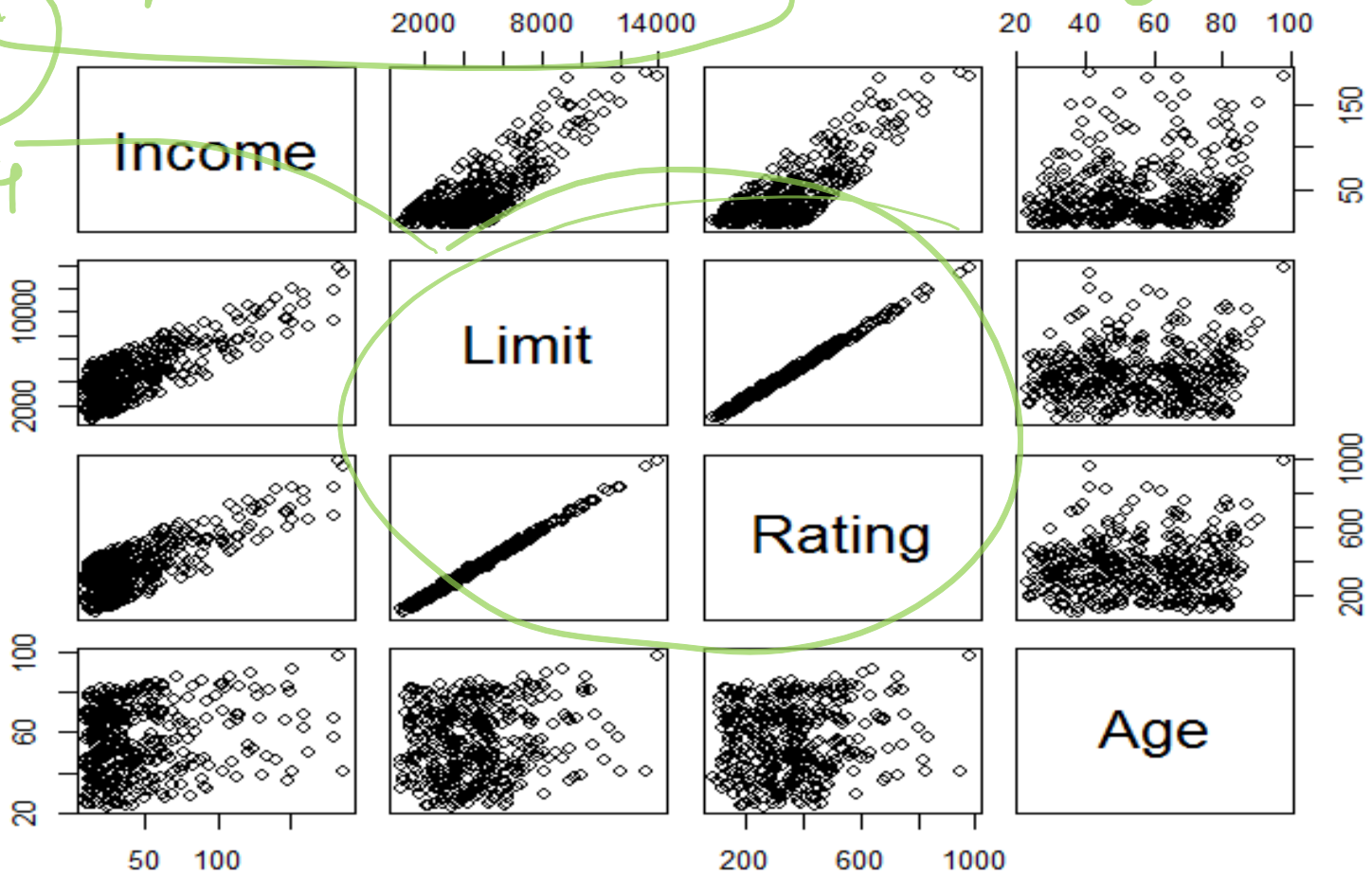




has less impact on balance

doesn't have correlation between age & balance

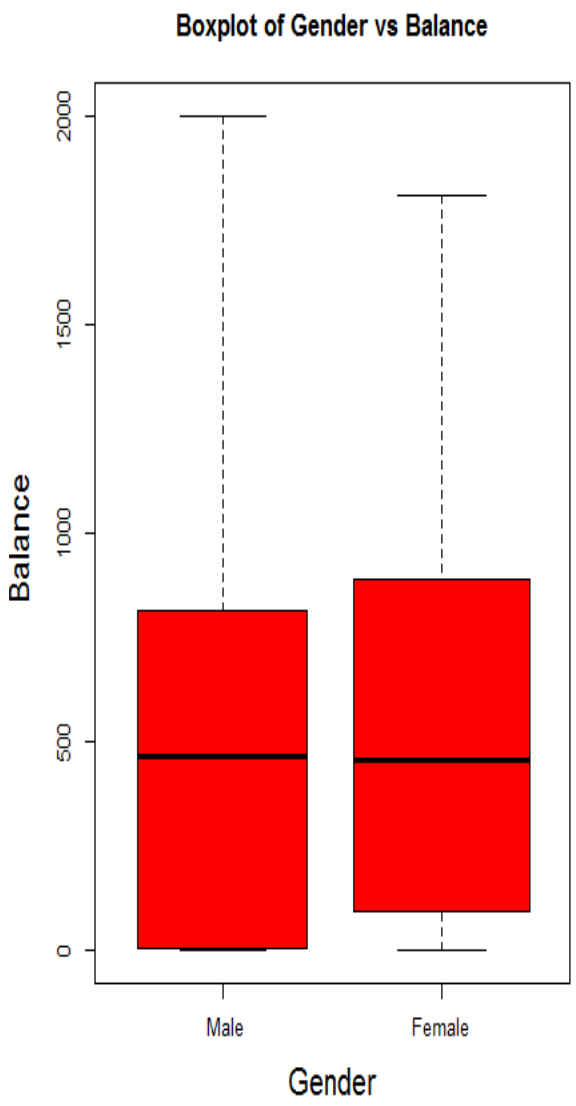
Pairwise Scatterplot (Bivariate)

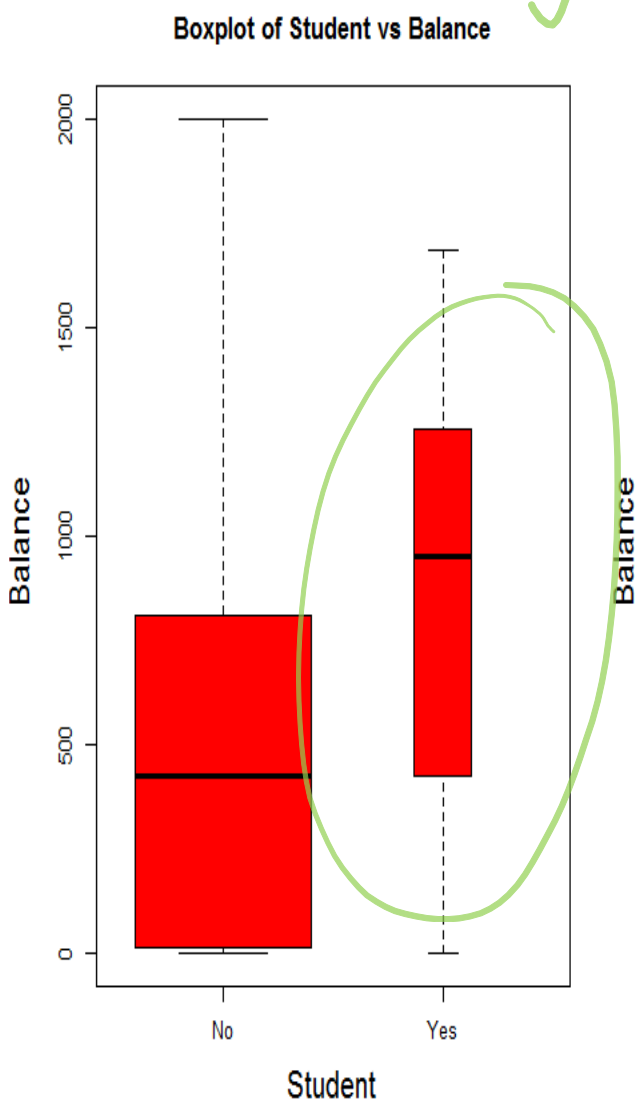


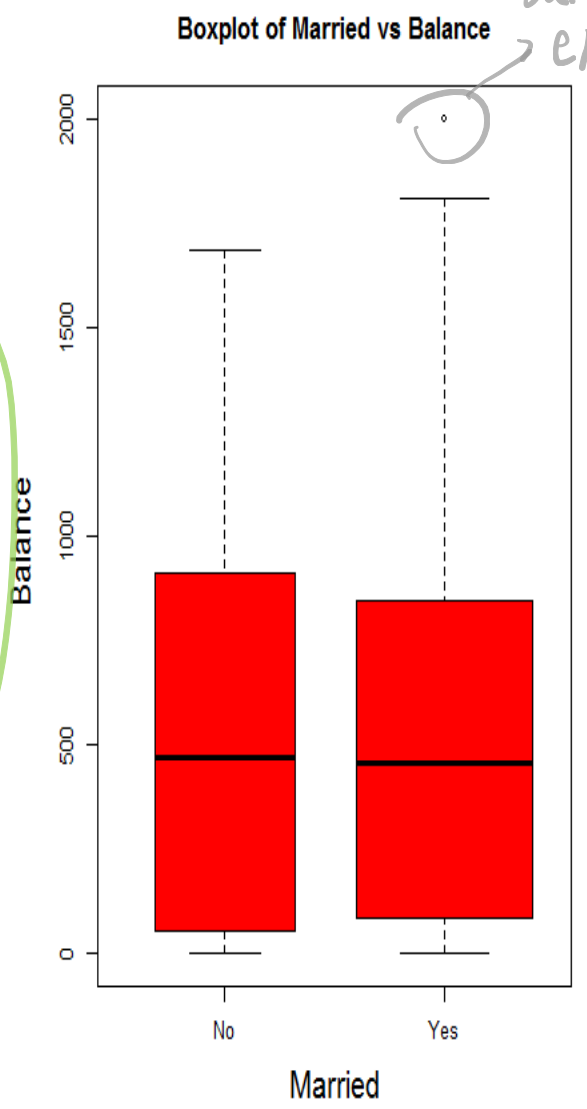
Pairwise Scatterplot for four continuous Variables

Boxplot (Bivariate)

filter out potential factors to the next stage (modeling)

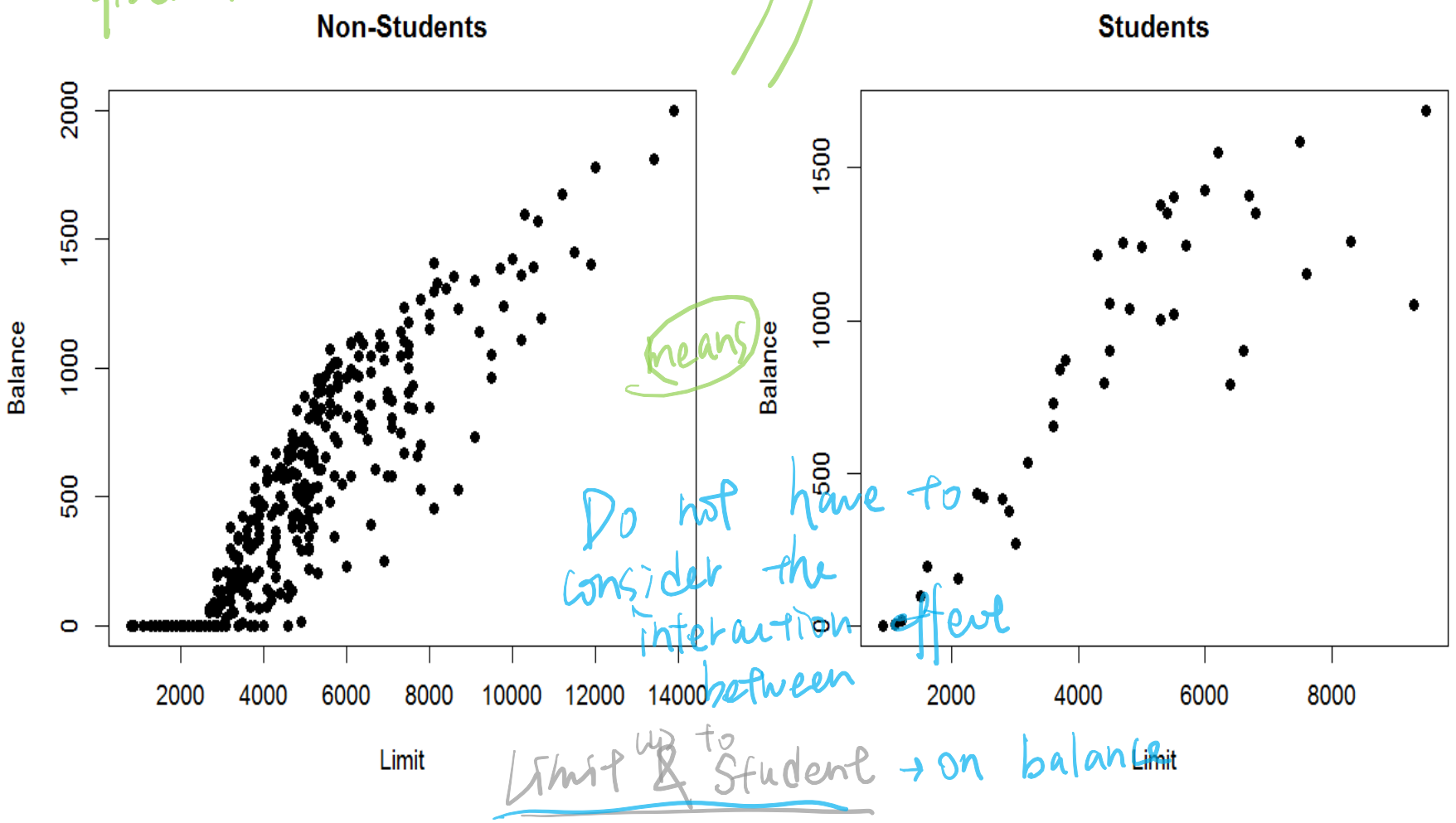






Conditional Pairwise Scatterplot(Bivariate)

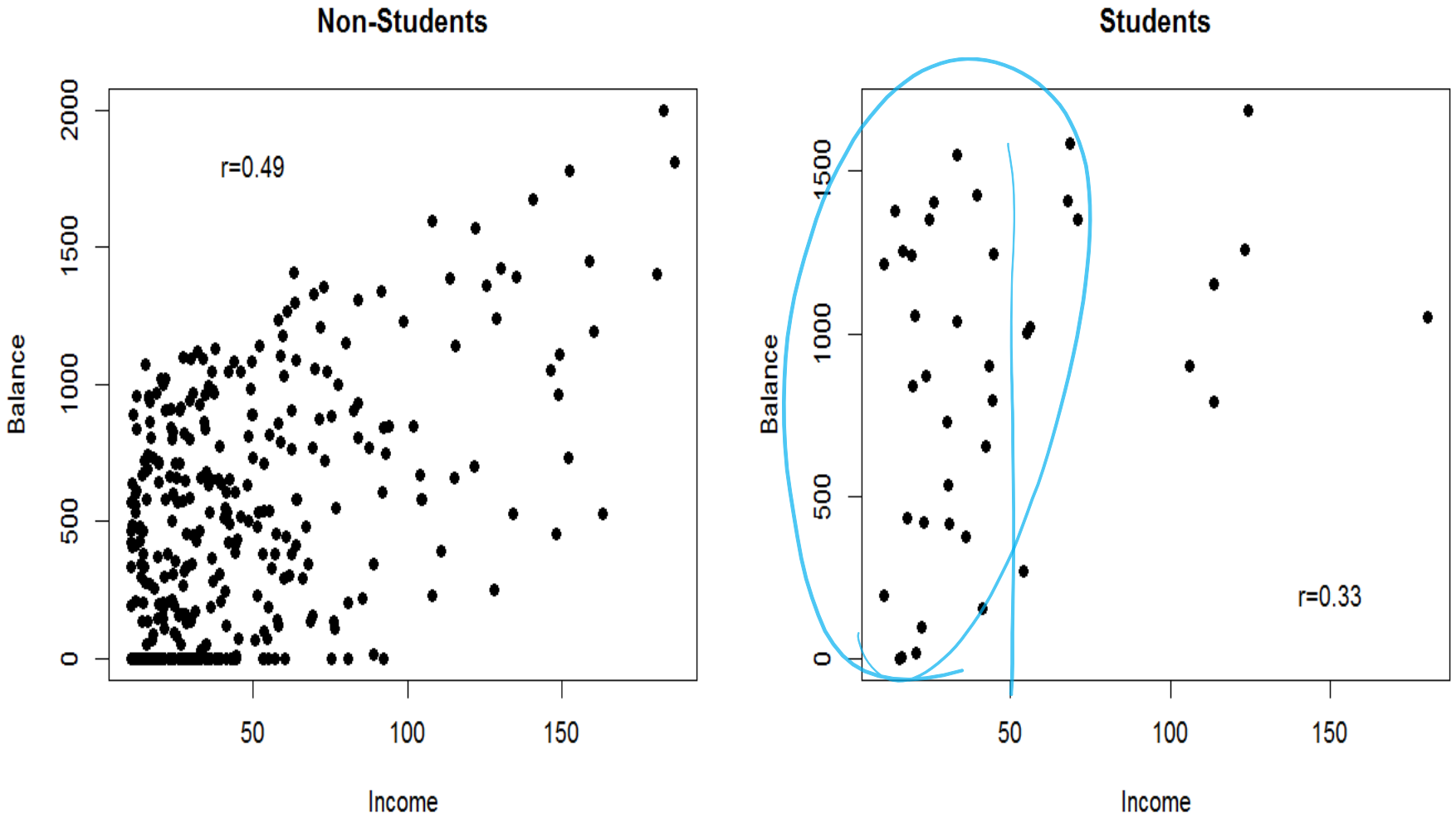
relation between Limit & balance
given that the customer is a student or non-student



means

Do not have to consider the interaction effect between Limit & Student up to balance

Conditional Pairwise Scatterplot(Bivariate)



What we've learned so far...

- Most customers, especially non-students, carry zero balance after their monthly payment and very few customers carry extreme high balance, e.g. >1600.
- Gender, Married status and Age seem irrelevant to Balance.
- There is a strong correlation between credit limit and credit rating
- A customer with higher income level is usually given larger credit limit.
- Income and Limit are potential factors affecting Balance, but their impacts, especially Income, may be different in Student and Non-Student groups.