


Data analysis fundamentals Course assignment - German Credit Risk Analysis






1. Introduction:

1.1. Context

The original dataset contains 1000 entries with 20 categorical/symbolic attributes prepared by Prof. Hofmann. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. 

1.2. Variables

The selected attributes are:

- ❖ Age (numeric) 
- ❖ Sex (text: male, female)
- ❖ Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled) 
- ❖ Housing (text: own, rent, or free) 
- ❖ Saving accounts (text - little, moderate, quite rich, rich)
- ❖ Checking account (text, in DM - Deutsch Mark)
- ❖ Credit amount (numeric, in DM) 
- ❖ Duration (numeric, in month) 
- ❖ Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)
- ❖ Risk (Value target - Good or Bad Risk)

2. Overall and initial analysis

2.1. Data categories

There are totally ten variables divided into four categories.

Numeric variable	Binary variable	Nominal variable	Ordinal variable
Age	Sex	Purpose	Job
Credit amount	Risk	Housing	Saving account
Duration			Checking account

2.2. Overall analysis

❖ Information about numeric variables

	RANGE	MEAN	MEDIAN	MODE	MAX	MIN	Q1 (QUARTILE.INC)	Q3 (QUARTILE.INC)
AGE	56	35.546	33	27	75	19	27	42
CREDIT AMOUNT	18,174	3,271	2,320	1393	18,424	250	1365.5	3972.25
DURATION	68	20.903	18	24	72	4	12	24

❖ Missing values

There are two variables having missing values. That are saving account and checking account variables.

MISSING VALUES									
Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	0	0	0	183	394	0	0	0	0

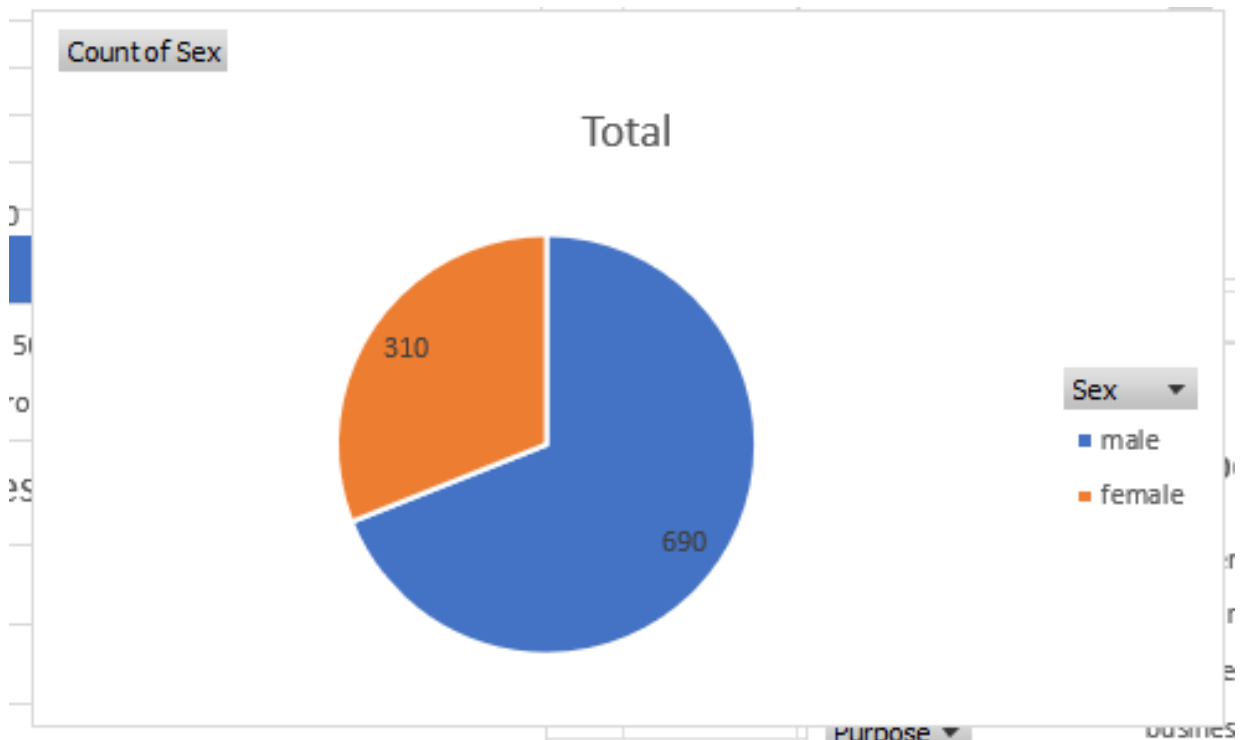
❖ Deleting records with missing values

When deleting records having missing values, you also delete values of other variables that don't consist of NA values, such as age, credit amount, and duration. That is the reason why the measurements of numeric variables change after deleting these records.

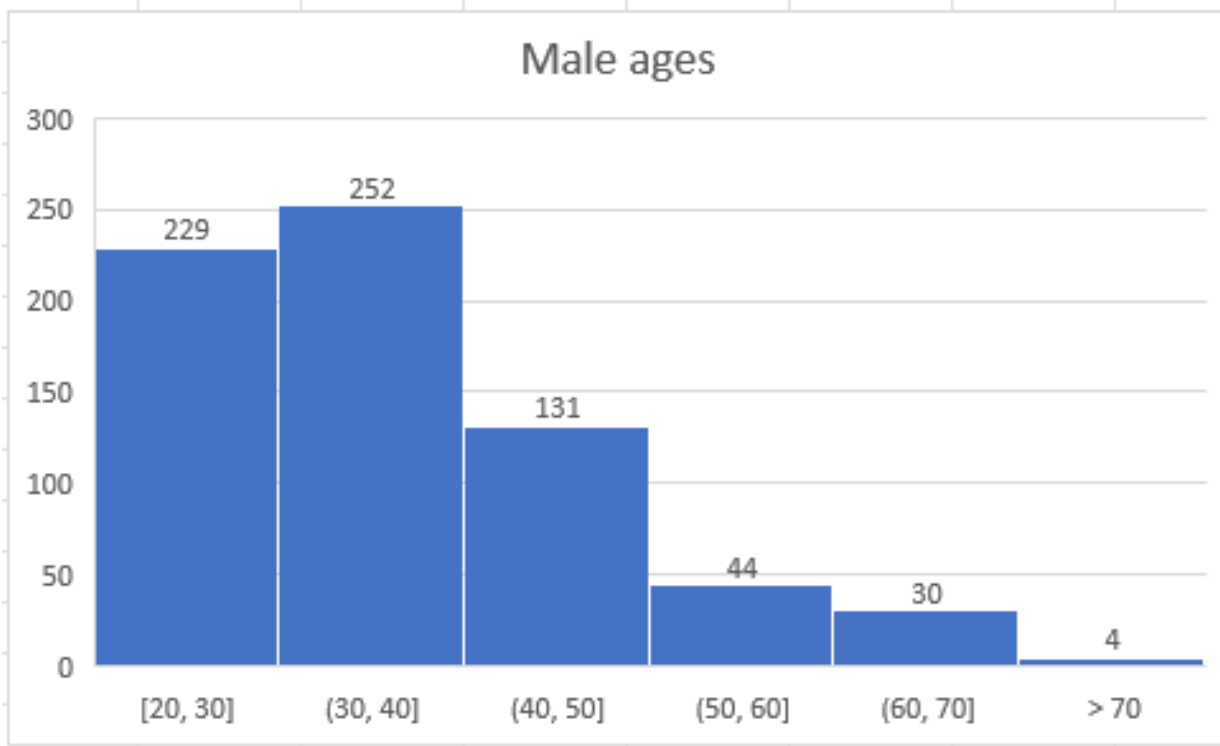
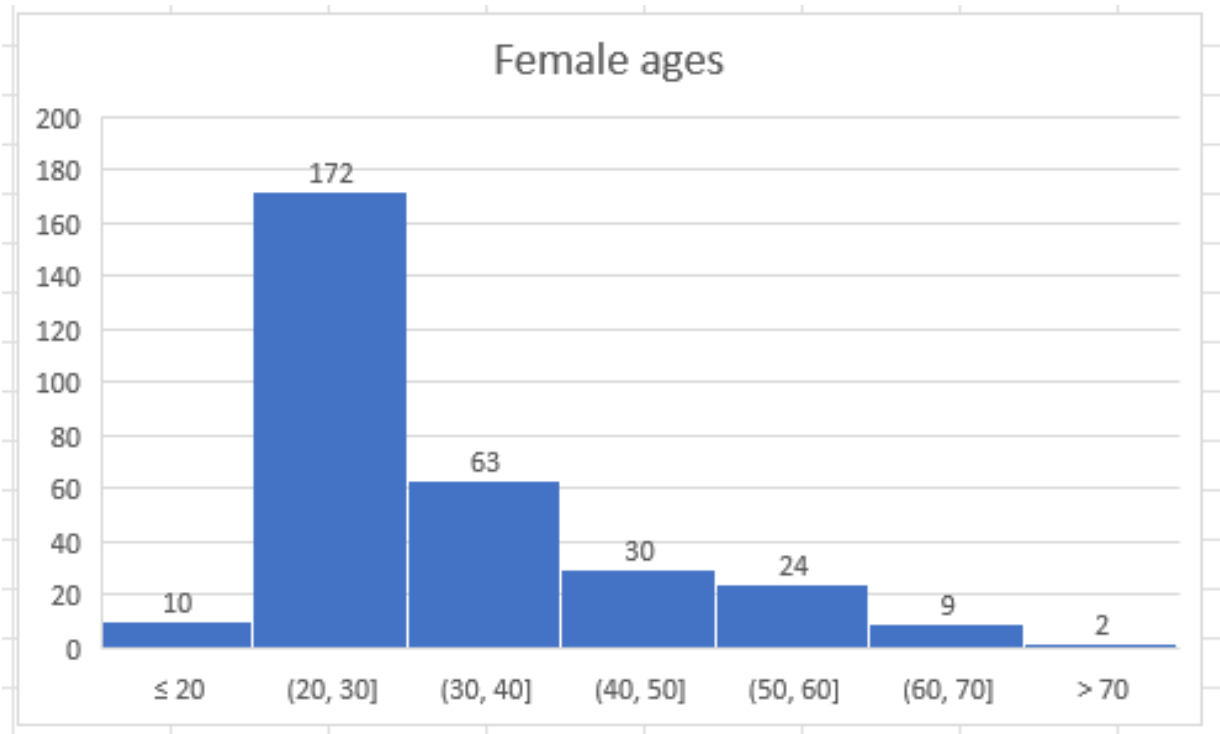
	RANGE	MEAN	MEDIAN	MODE	MAX	MIN	Q1 (QUARTILE.INC)	Q3 (QUARTILE.INC)
AGE	56	34.88889	31.5	23	75	19	26	41
CREDIT AMOUNT	18,148	3,279	2,327	1295	18,424	276	1297.5	3971.25
DURATION	66	21.33908	18	12	72	6	12	26.75

3. Gender analysis

In general, the number of male customers taking credit from a bank is 690 people, 2.2 times higher than that of female customers



3.1. Age by gender



The histogram of women and men ages are positive and heavy skewness distribution with the skewness coefficients equaling 1.35 and 0.95 respectively.

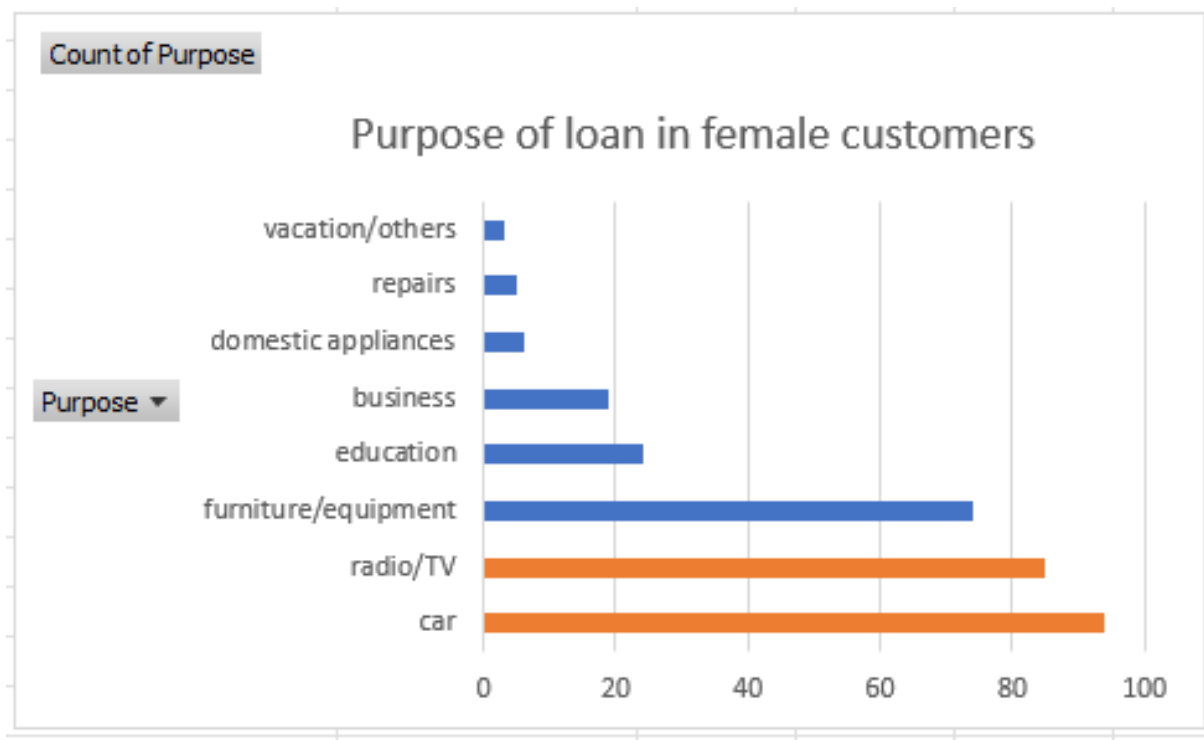


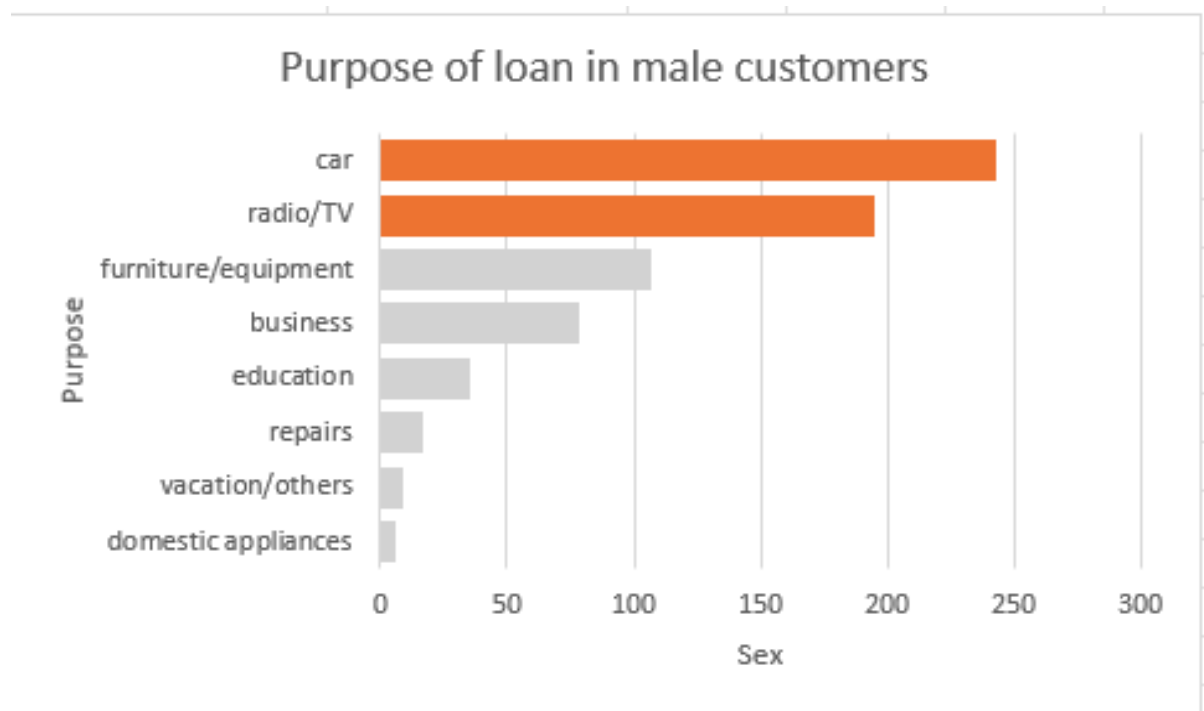
The biggest number of male customers taking credit from a bank are in the age groups of 30 to 40 and 20 to 30, while females in the age group of 20 to 30 is the most popular people applying loans from banks.



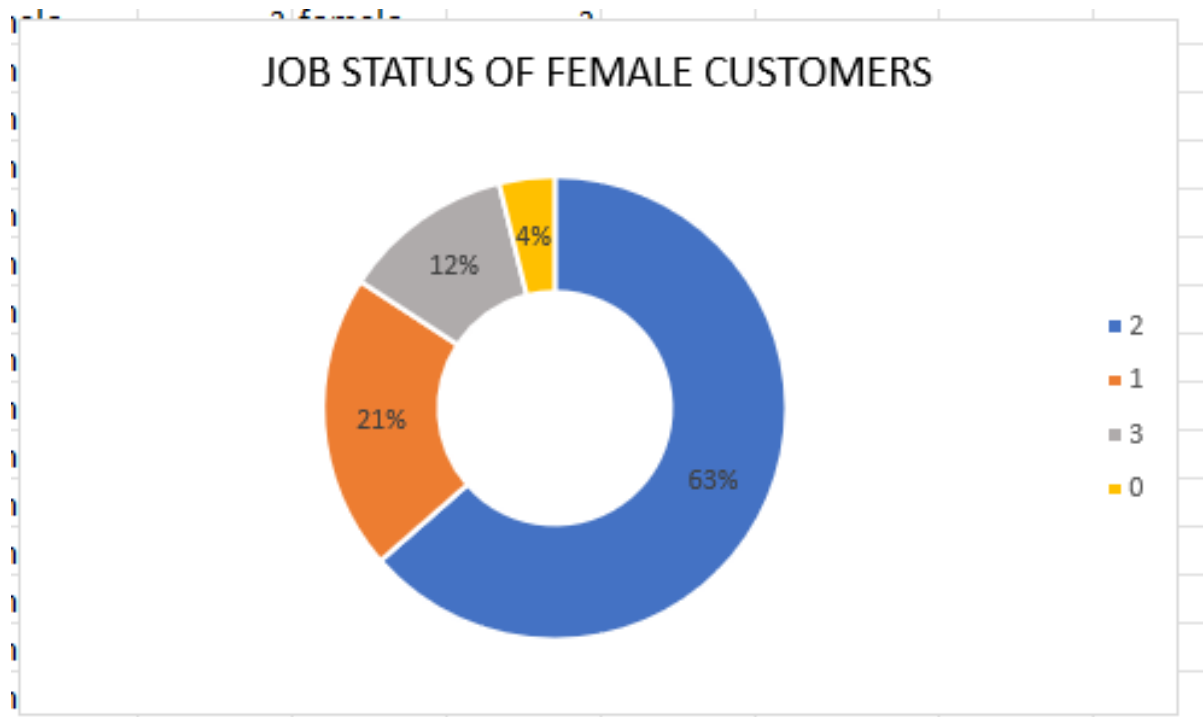
3.2. Purpose of the loan by gender

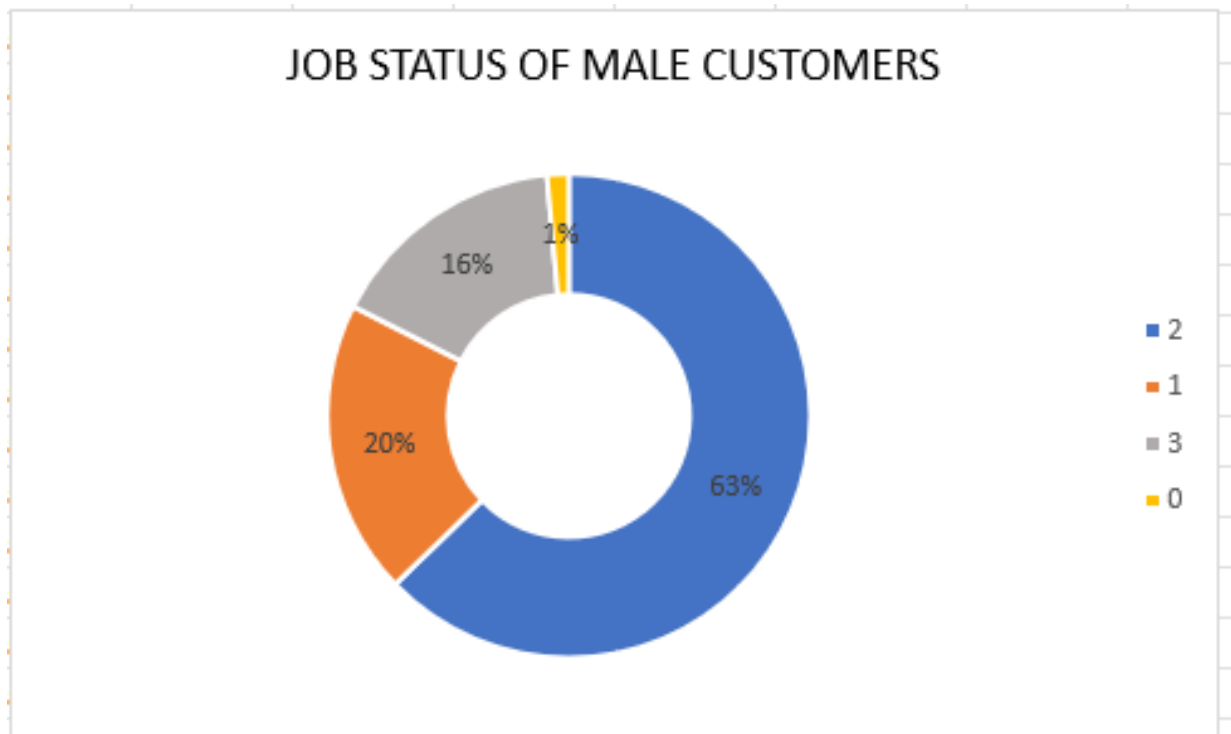
The three most popular purposes of the loan in both genders are car, radio/TV, and furniture/equipment, while domestic appliances, repairs and vacation/others are the less common purpose when people take loans from banks.





3.3. Employment status of applicants by gender

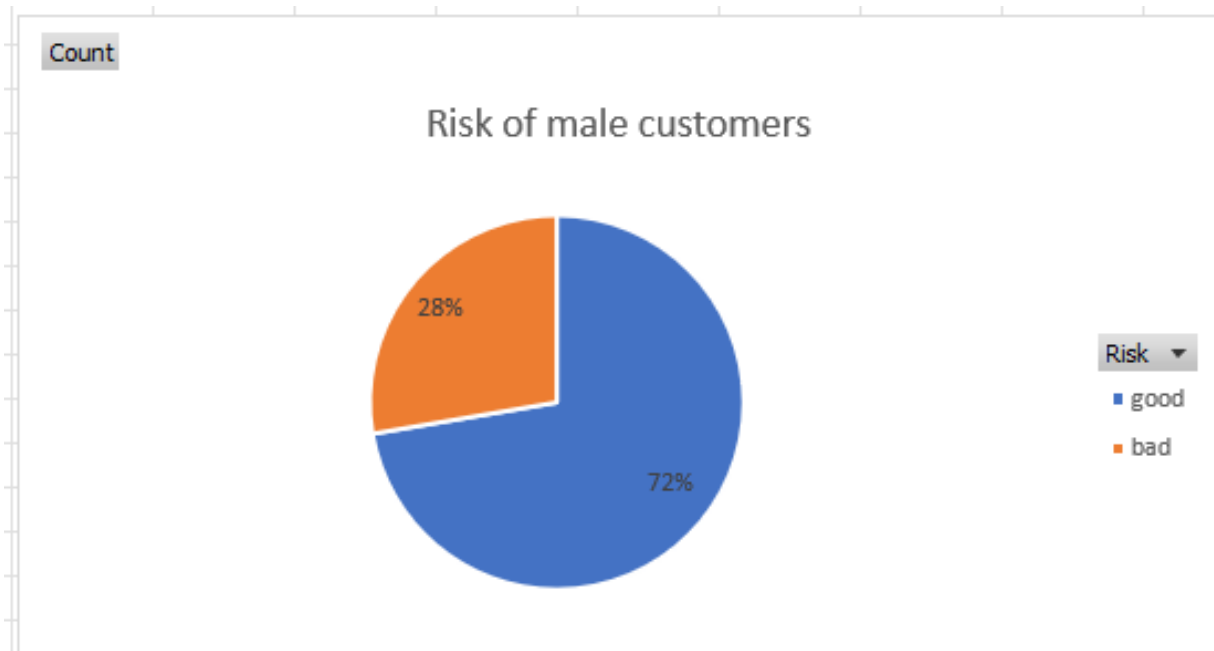
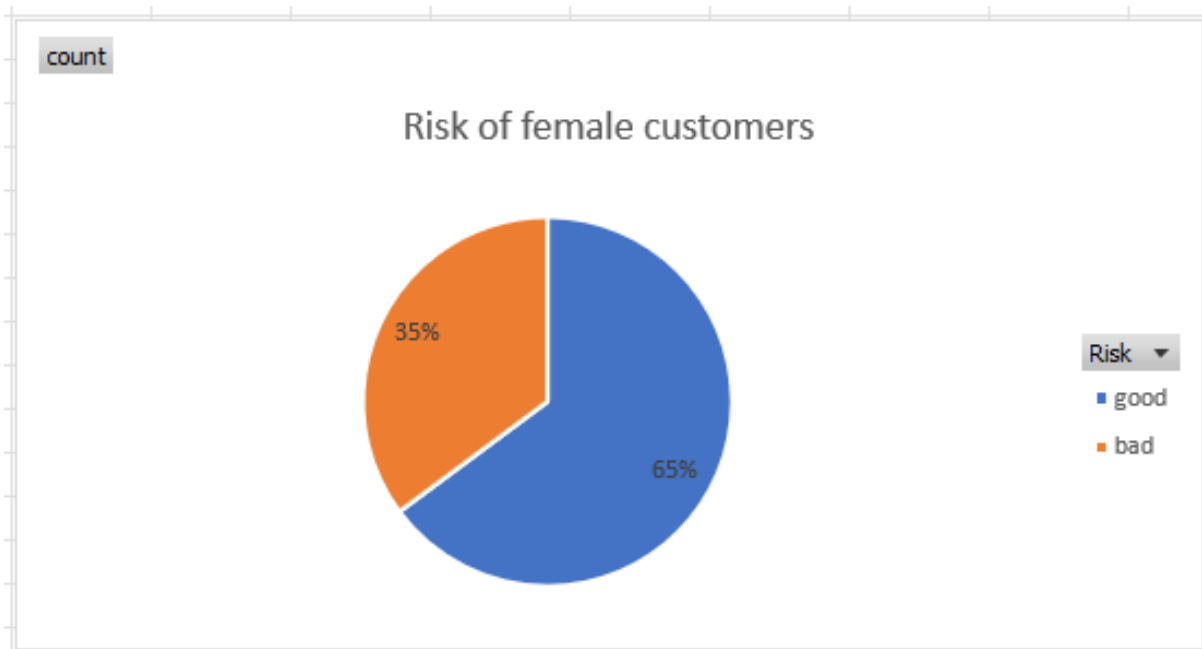




In both genders, skilled workers are the main customers taking loans from banks (63%), while unskilled and non-resident ones tend not to take credit from banks (less than 5%).

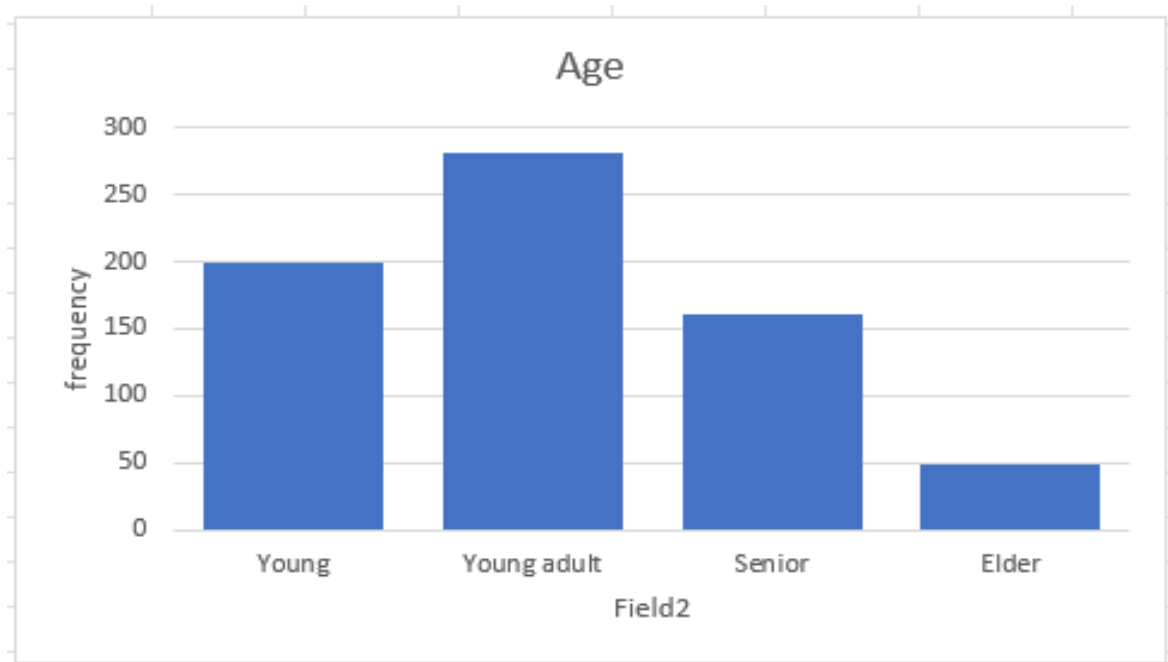
3.4. Risk by gender

	Female	Male
Good	201	499
Bad	109	191



More than 65% of customers of both genders have good risk, though the percentage of male customers having good risk is 6% higher than that of female ones. That means the male customer group is less risky than women.

4. *Age groups analysis*



We clarify customers based on age groups by following categories:

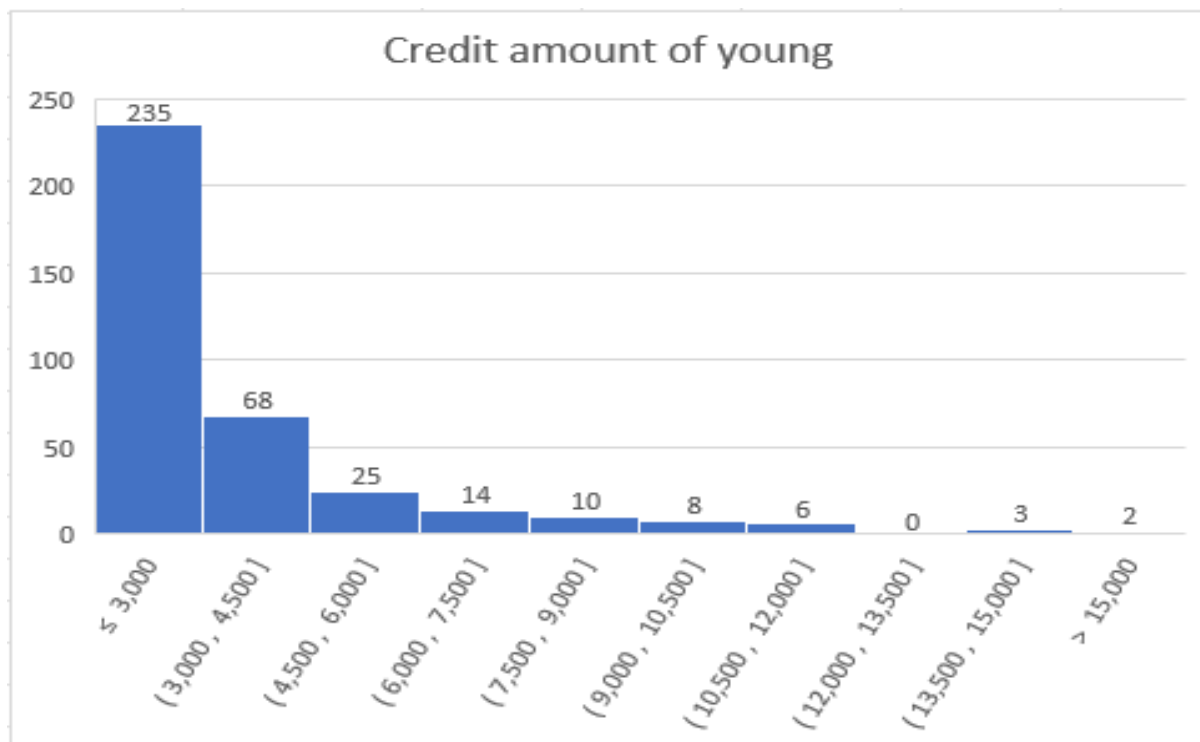
- Young: age of the applicant ranges from 19 to 29.
- Young adult: age of the applicant ranges from 30 to 40.
- Senior: age of the applicant ranges from 41 to 55.
- Elder: age of the applicant is above 55 years.

Overall, the potential customers of banks when introducing credit products are young and young adult groups from 19 to 40 years old, whereas elders have less demand for loaning money.

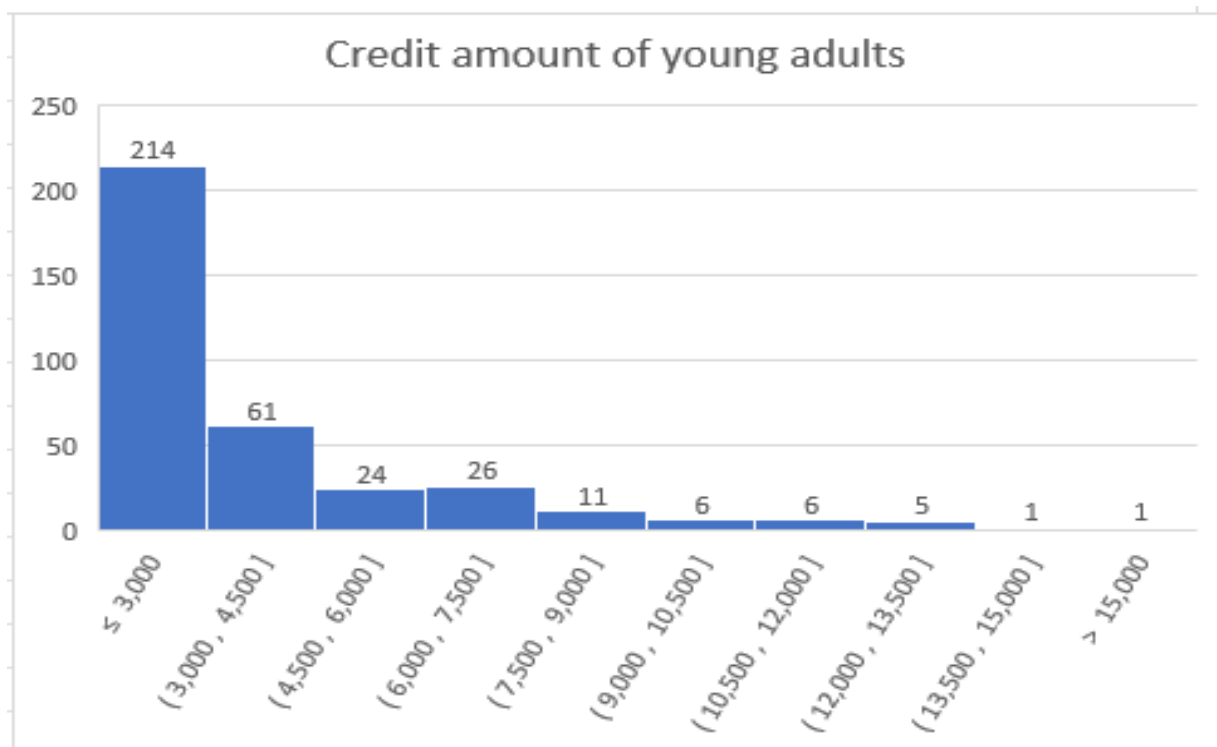
4.1. Credit amount by age groups

	Young	Young adults	Senior	Elder
Average credit amount	3,088.99	3,375.48	3,366.44	3,430.44

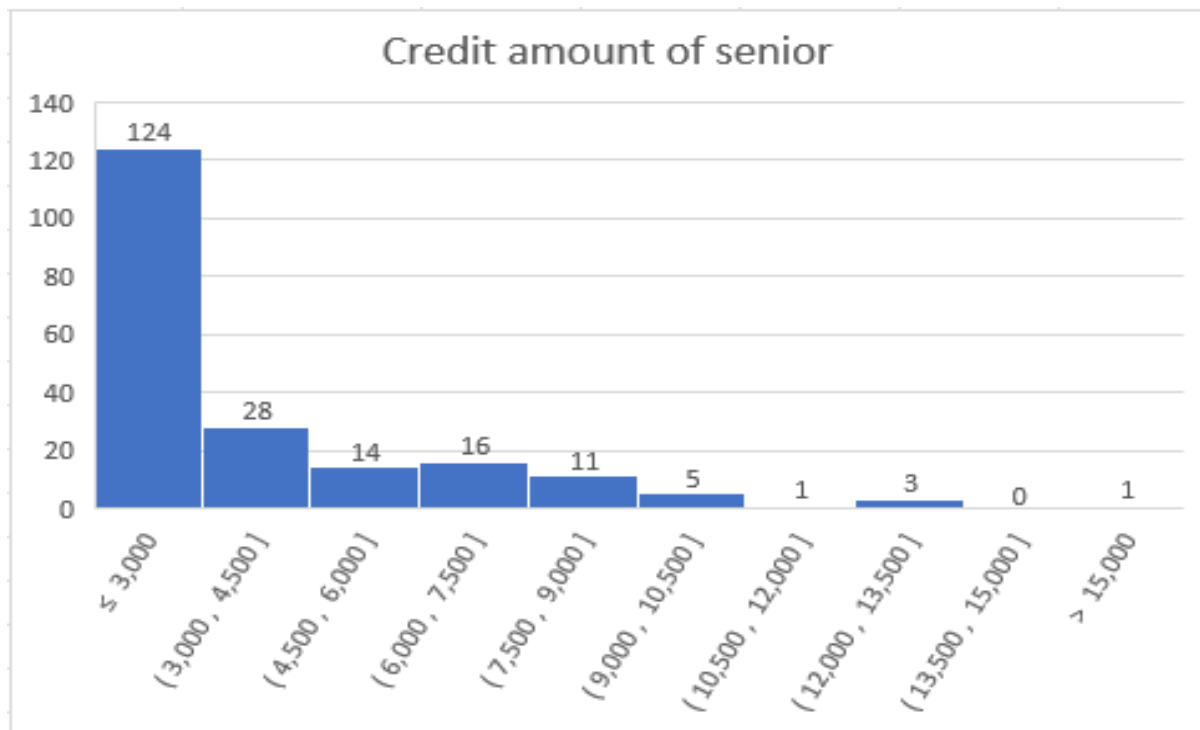
❖ Young



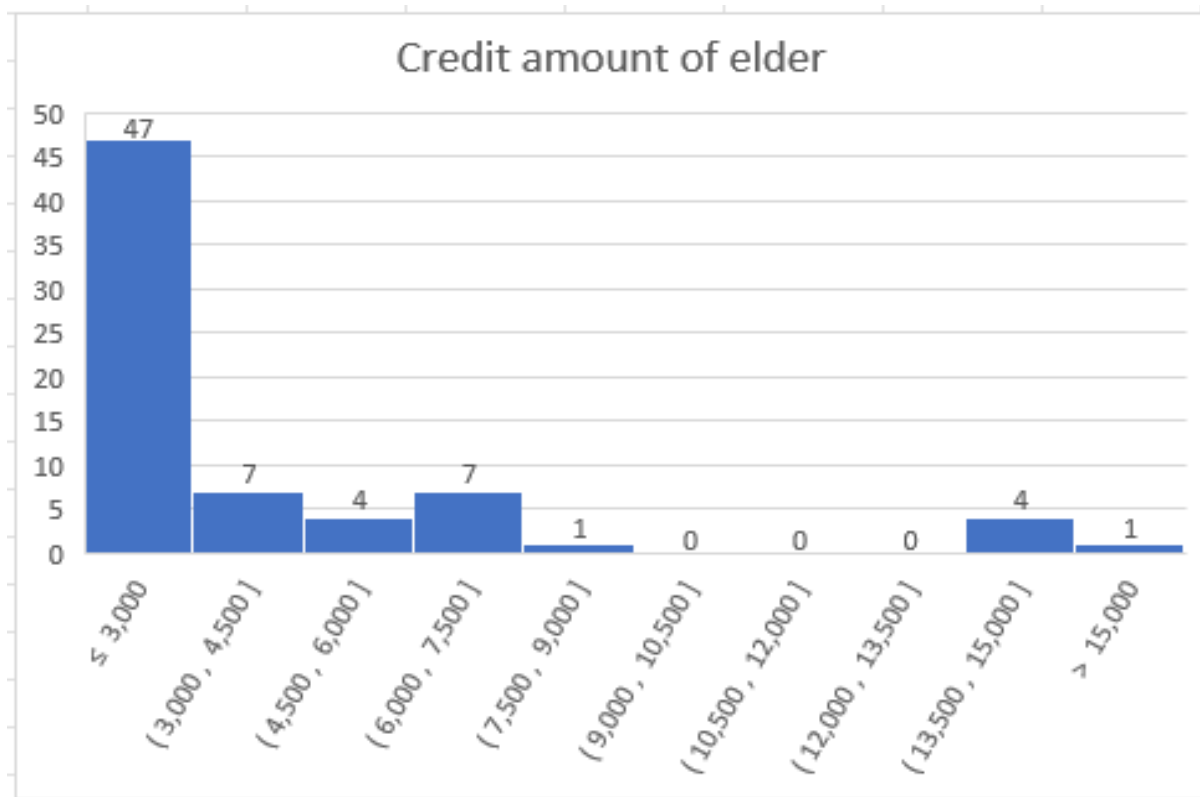
❖ Young adults



❖ Senior



❖ Elder

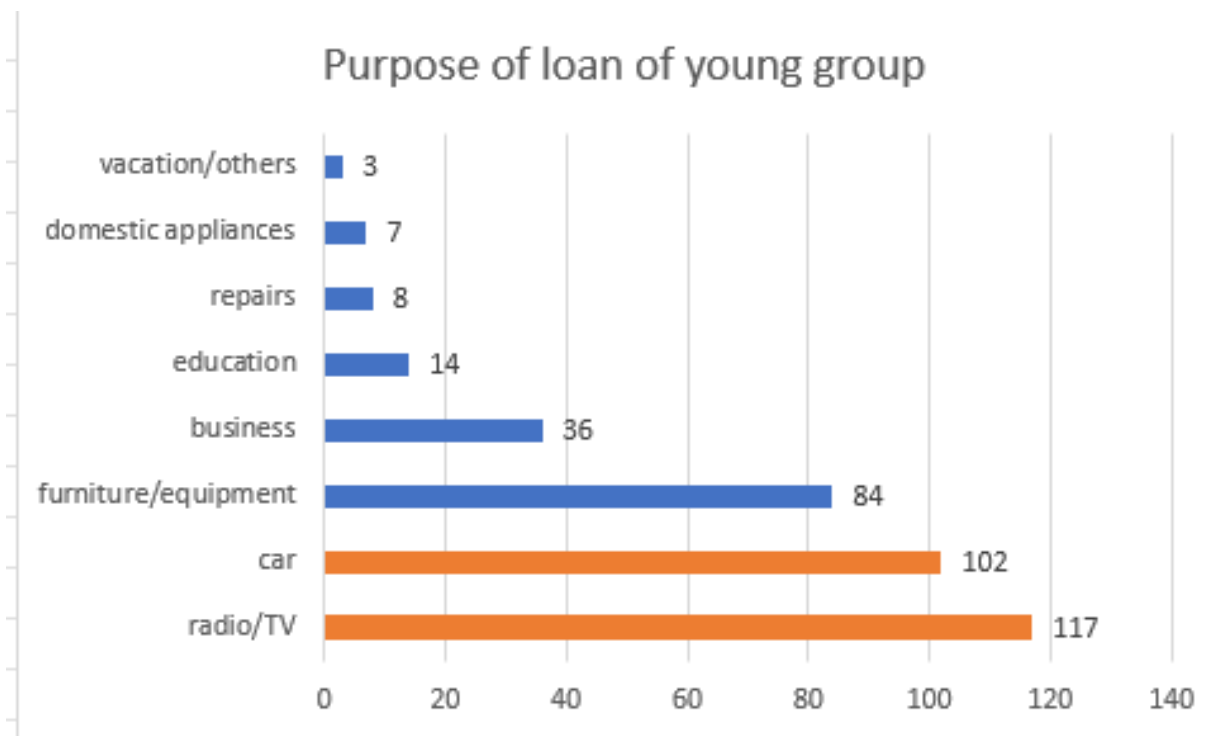


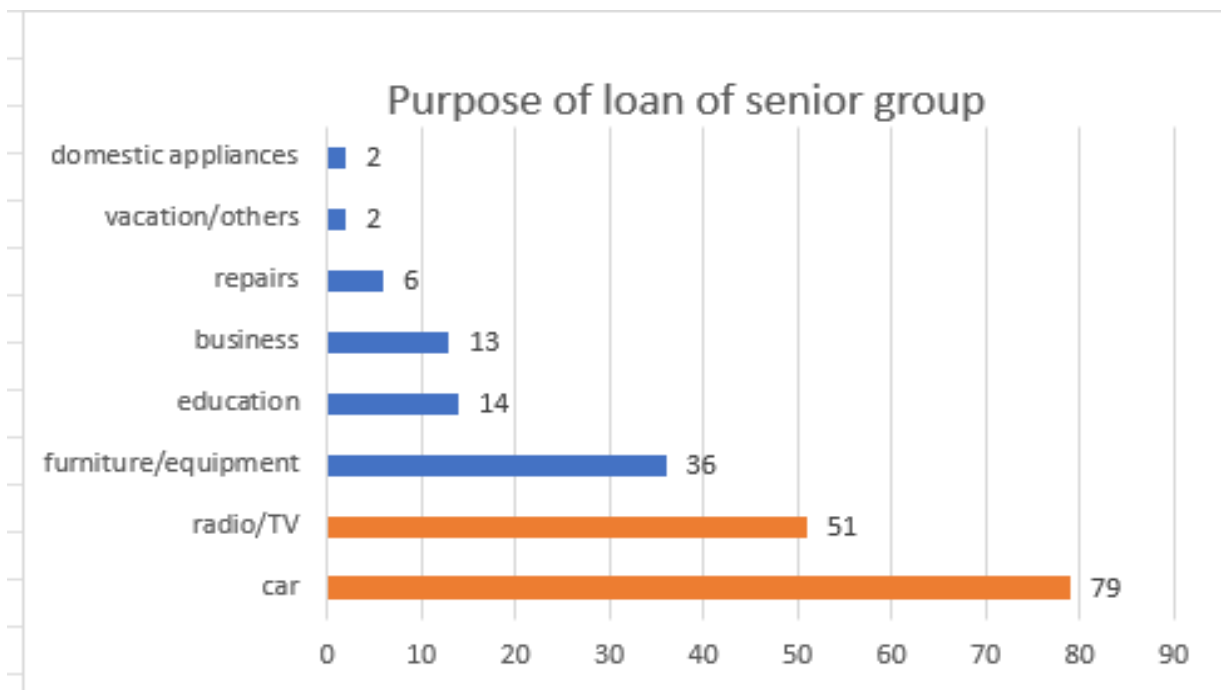
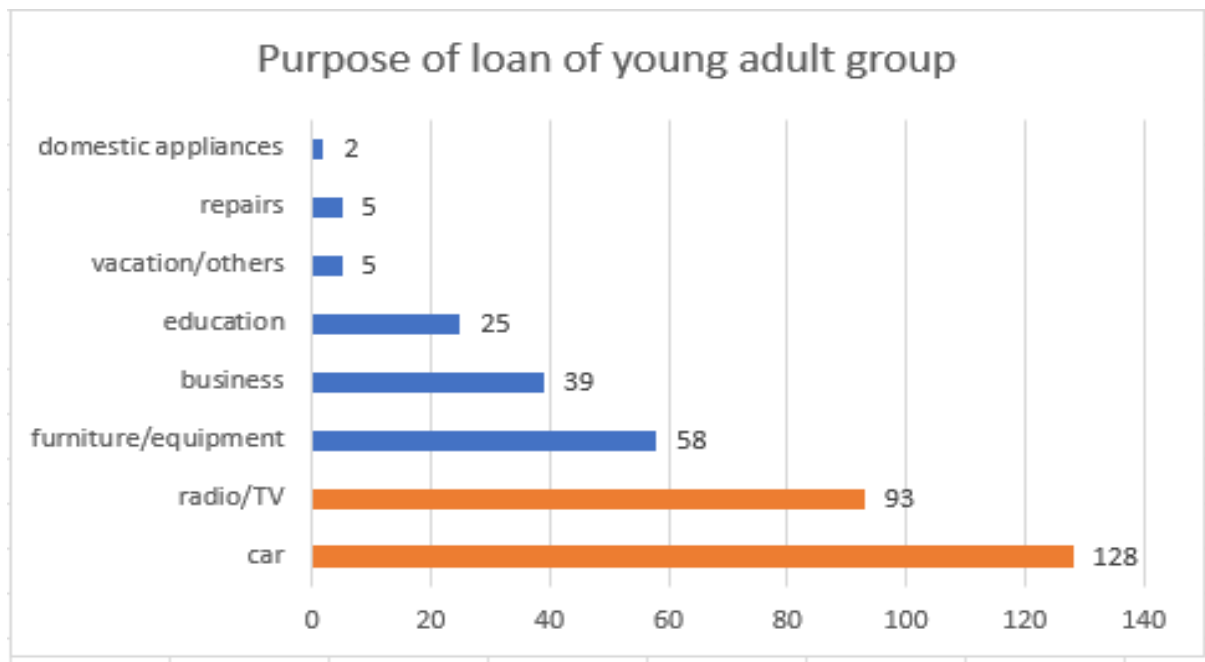
In conclusion, most people in all age groups loan less than 3000 from banks. However, some uncommon loans in the four groups generate outliers in the credit amount charts of the groups, which makes the average credit amount of each group not to express the center value of the data.

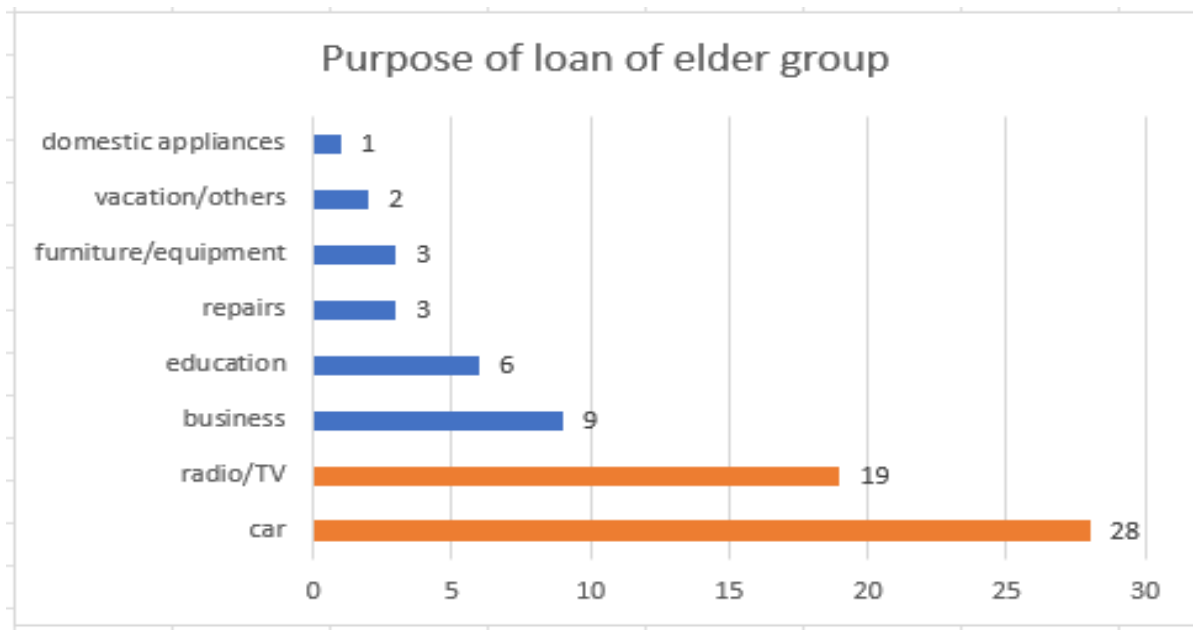


4.2. Purpose of loan by age groups

In all groups, the two most well-liked products that customers tend to buy with loans from banks are radio/TV and car. Besides these products, customers have a tendency to spend the credit on furniture/equipment, except the elder group. Because they prefer to invest money in business.



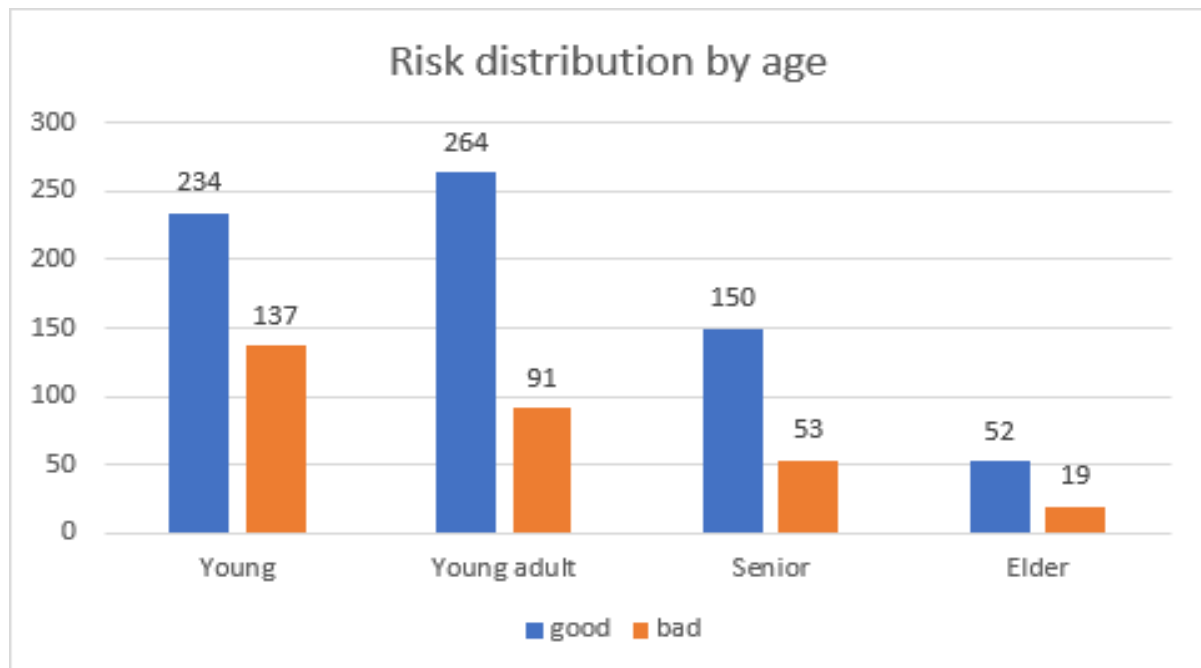




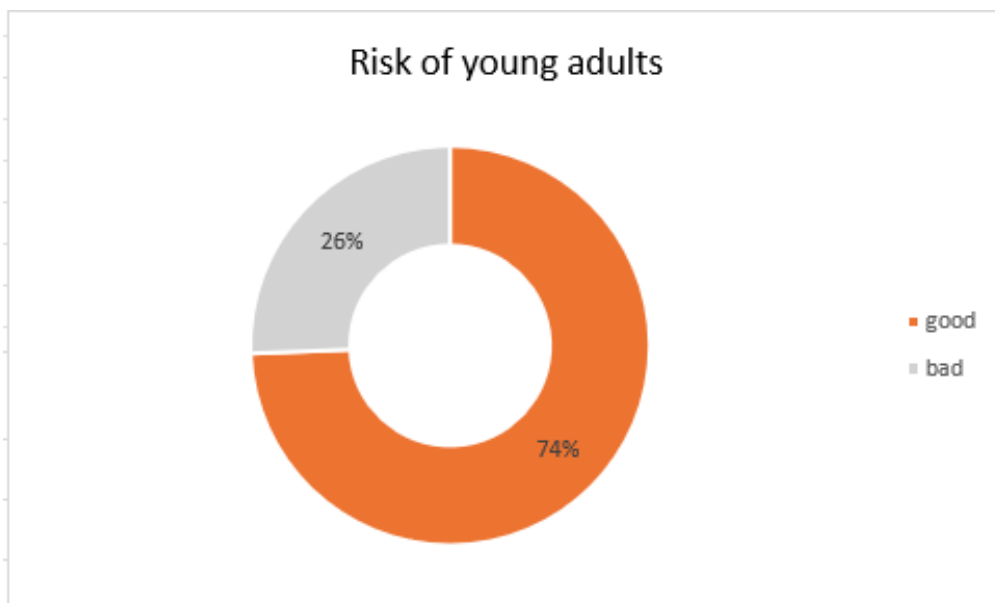
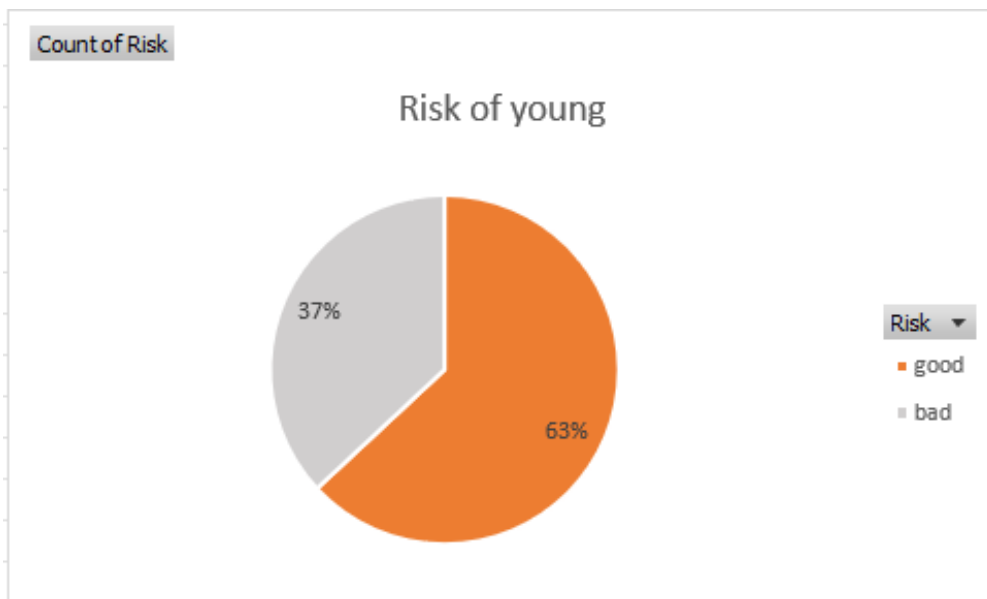
4.3. Risk by age groups

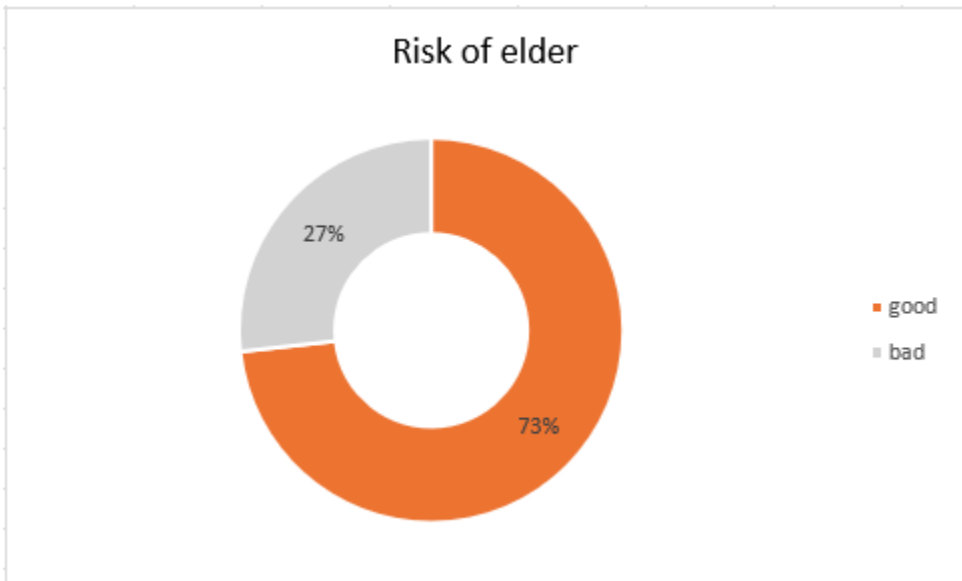
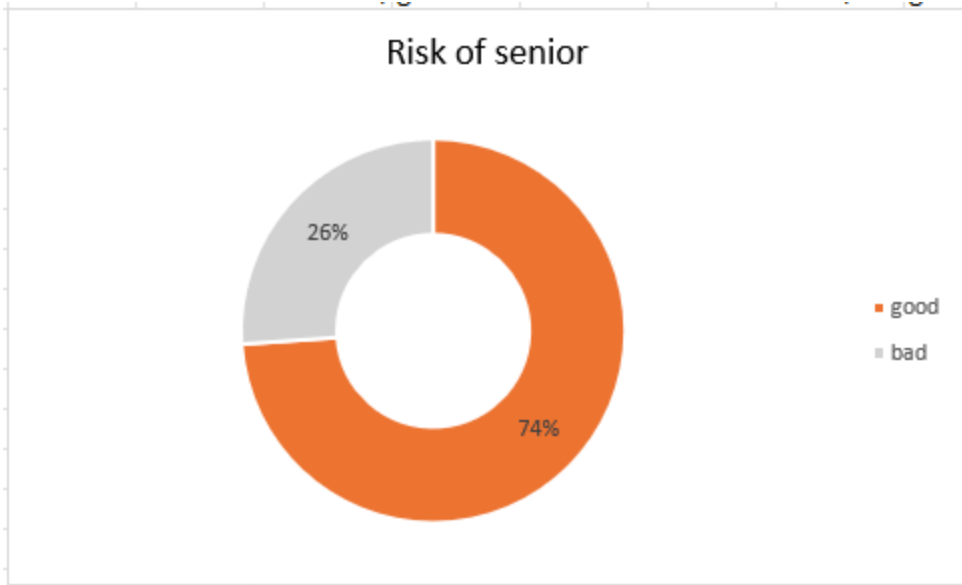


After analyzing the risk of customers by age group, we see that the young adult group has the highest number of good-risk customers and the second biggest number of bad-risk ones, While the young group contributes the greatest number of bad-risk customers.



In three groups: young adult, senior, and elder, the ratios of good-risk customers are always higher than 73%, whereas in young customers this figure is slightly lower at 63%. Thus, it is riskier when banks give credit to young customers. ✓

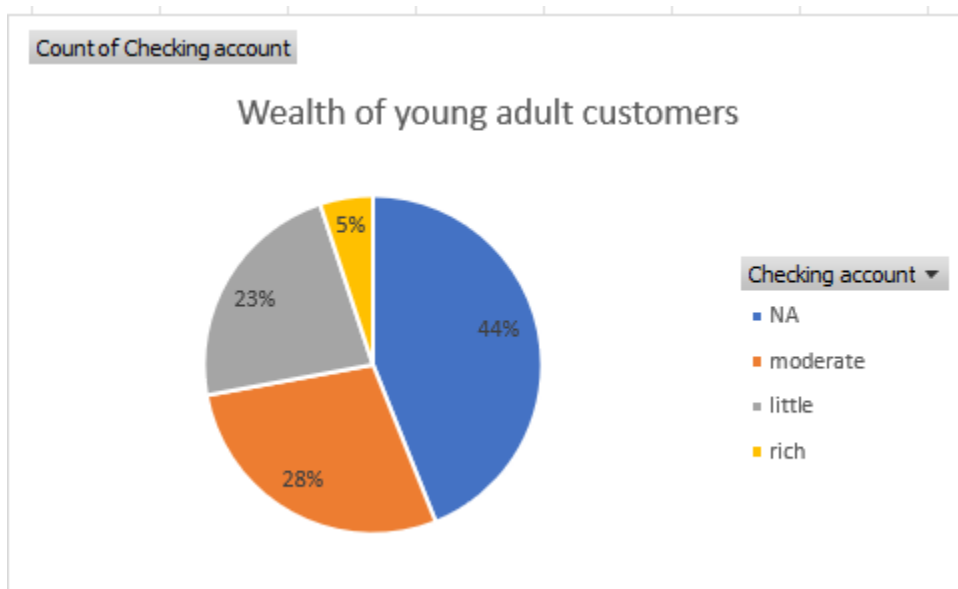
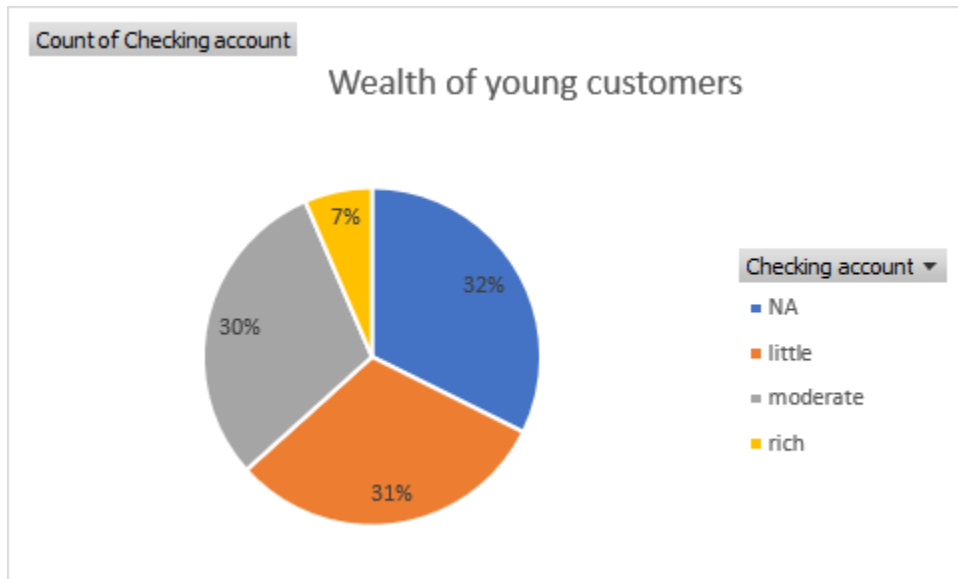


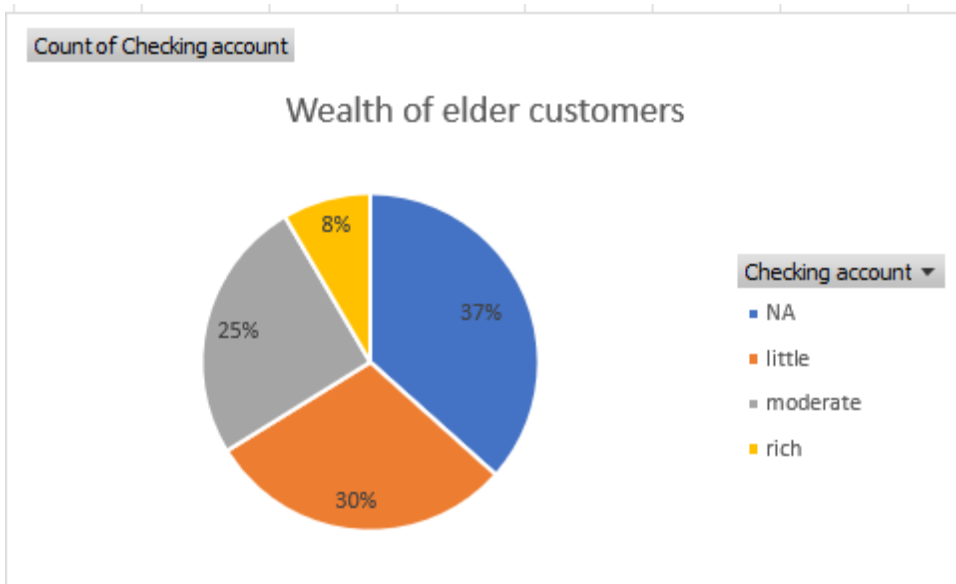
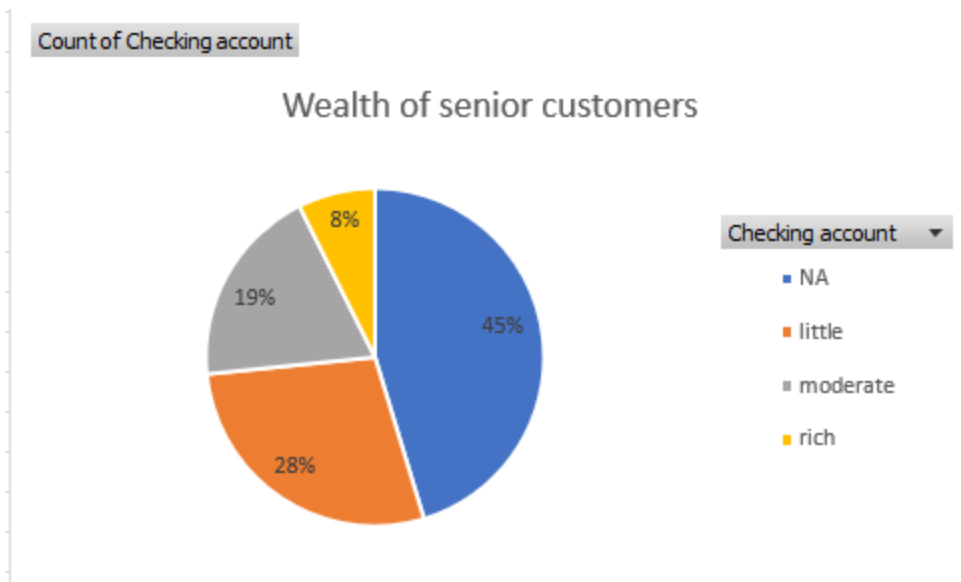


4.4. Wealth by age groups

	Young	Young adult	Senior	Elder
NA	120	156	92	26
Little	115	100	57	21
Moderate	112	81	39	18
Rich	24	18	15	6

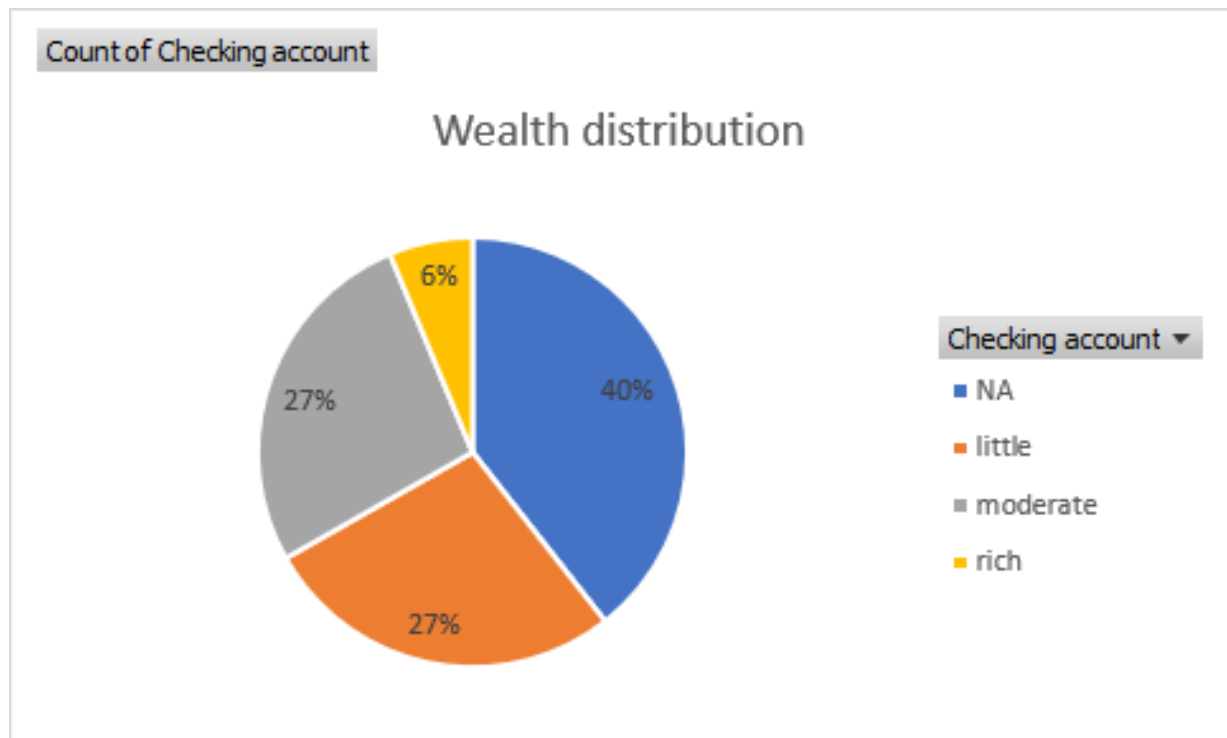
The ratio of unknown wealth status of young customers is the lowest in the four groups, while the ratio of moderate and little wealth customers is the highest. In addition, the percentages of rich customers in senior and elder groups are slightly higher than two other groups, at 8%





5. *Wealth of applicants (checking accounts) analysis*

Overall, the proportion of unknown wealth customers is quite high, at 40%, followed by those of little wealth and moderate wealth at 27%. By contrast, the percentage of rich customers is much lower with 6%.



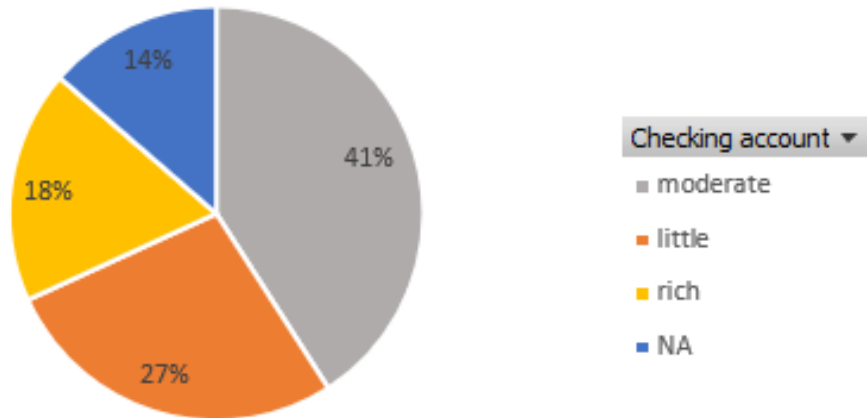
5.1. Employment status of applications by wealth

	NA	Little	Moderate	Rich
Unskilled and non-resident	3	6	9	4
Unskilled and resident	70	59	57	14
Skilled	266	172	155	37
Highly skilled	55	48	37	8

Although the number of unskilled and non-resident workers is the lowest in the four categories of customers by job status, the proportions of wealthy and moderate-wealth customers in this group are the highest, and that of unknown-wealth status customers is much lower than others. Besides, the rates of unknown-wealth status customers of skilled and highly skilled workers are the highest.

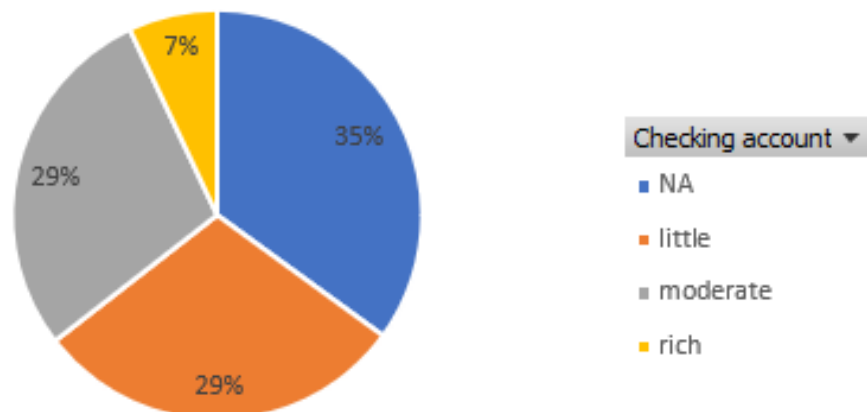
Count of Checking account

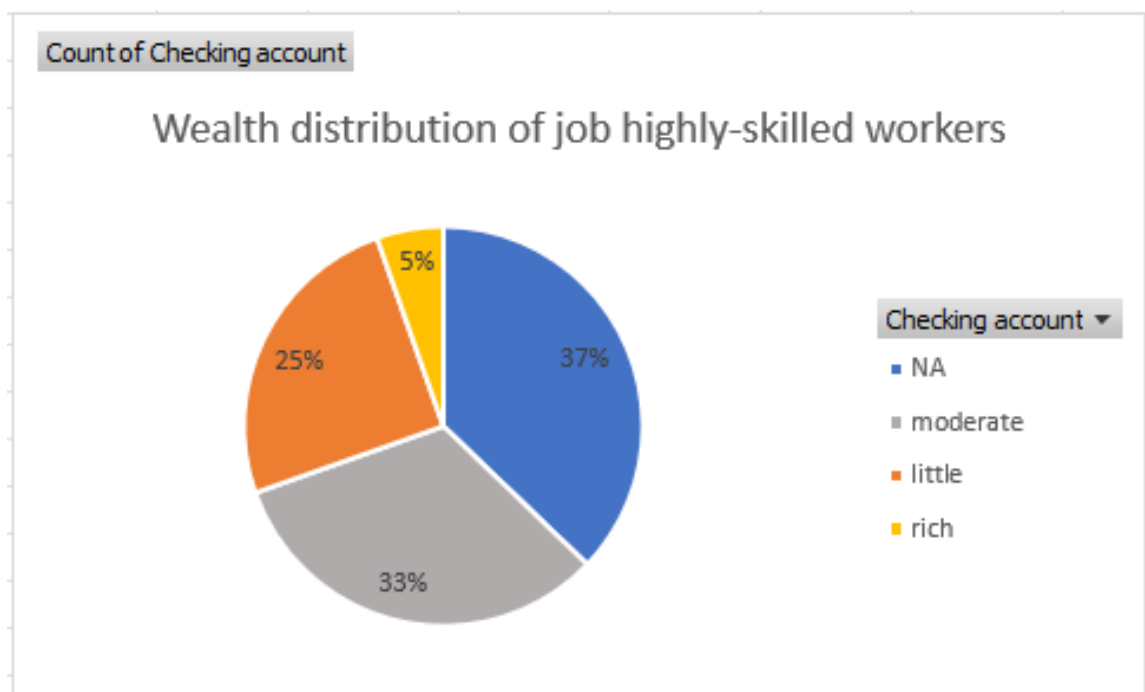
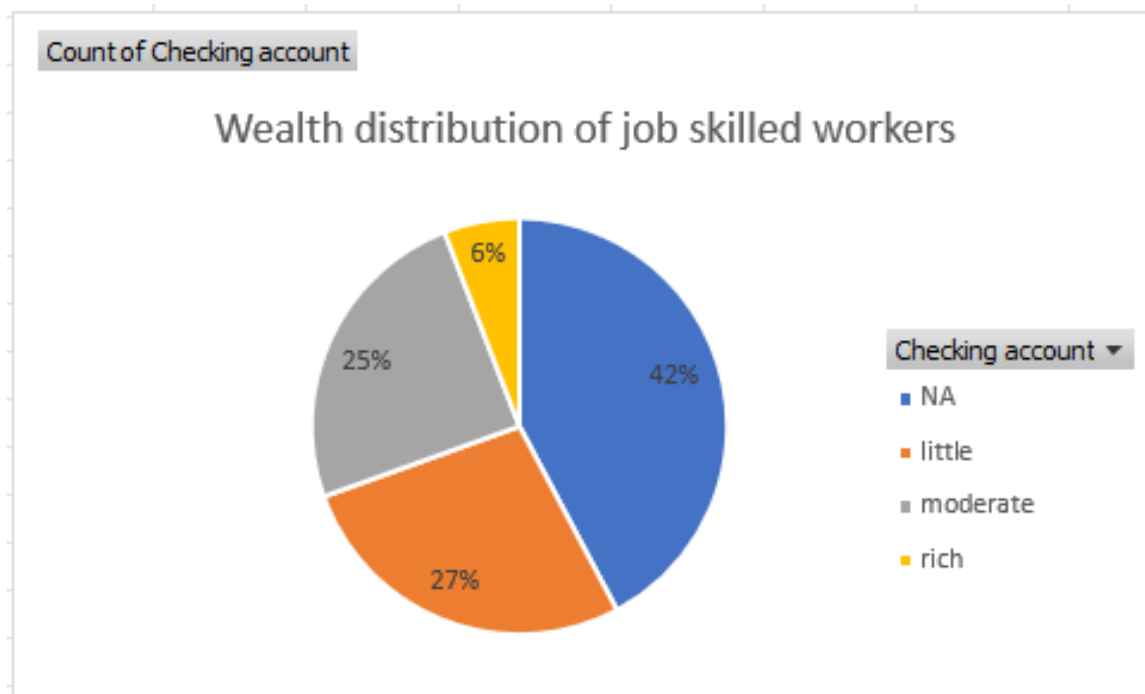
Wealth distribution of unskilled and non-resident workers



Count of Checking account

Wealth distribution of job unskilled and resident workers



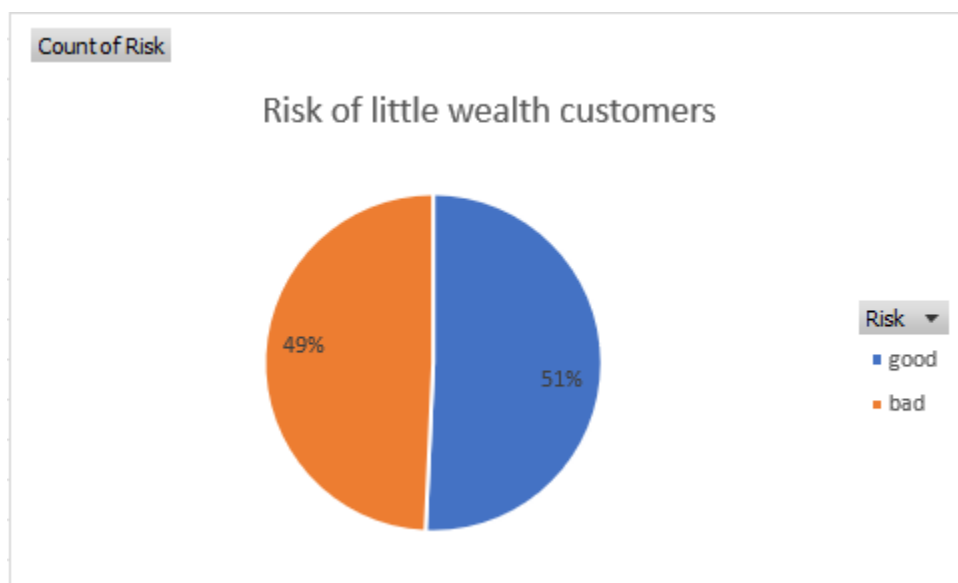
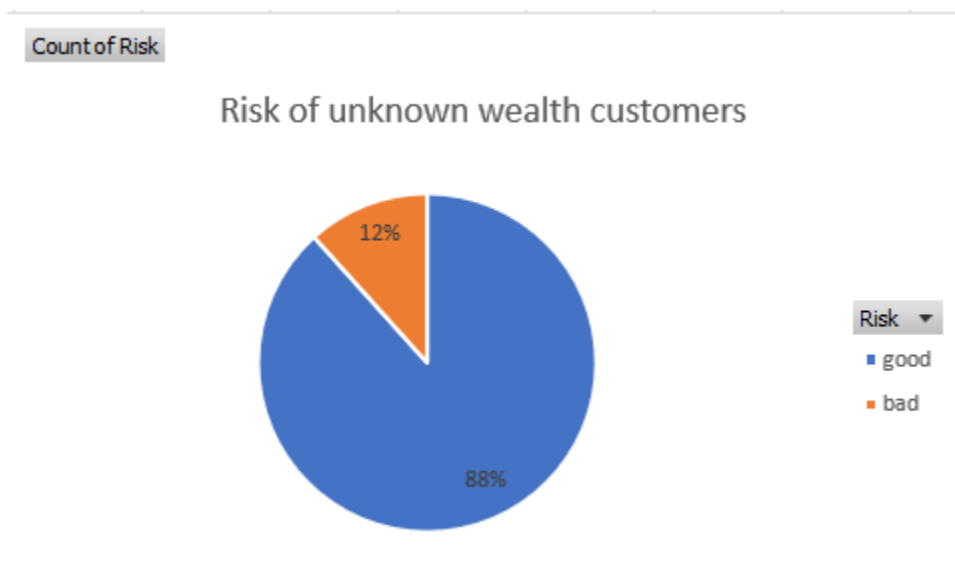


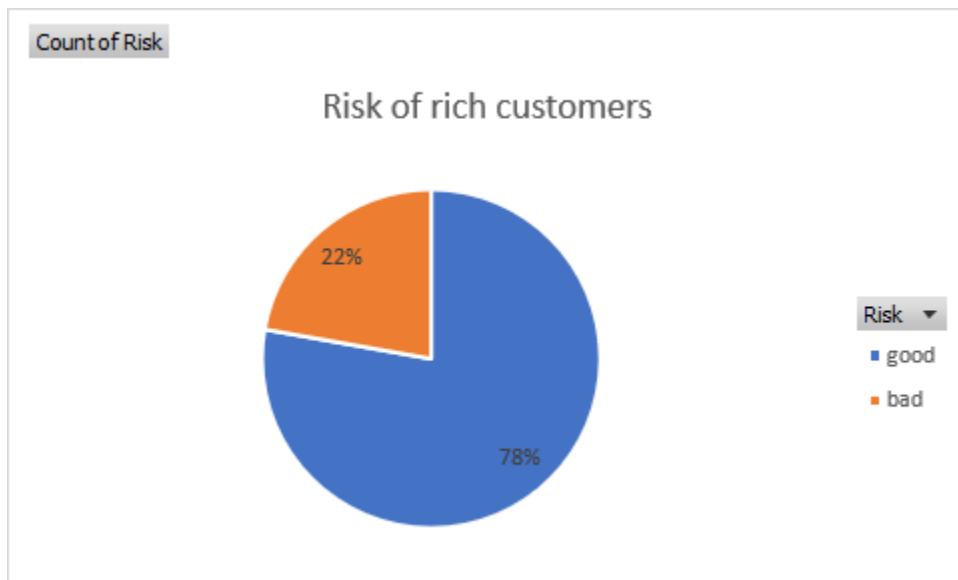
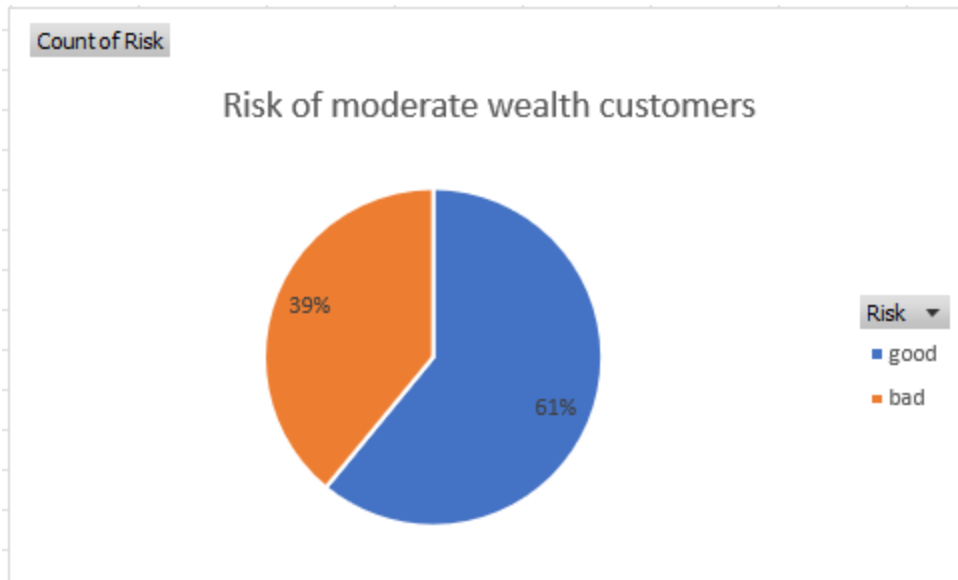
5.2. Risk by wealth

Most of the good-risk customers have unknown wealth status, whereas bad-risk customers are predominantly in little and moderate wealth parts.

	NA	Little	Moderate	Rich
Good	348	139	164	49
Bad	46	135	105	14

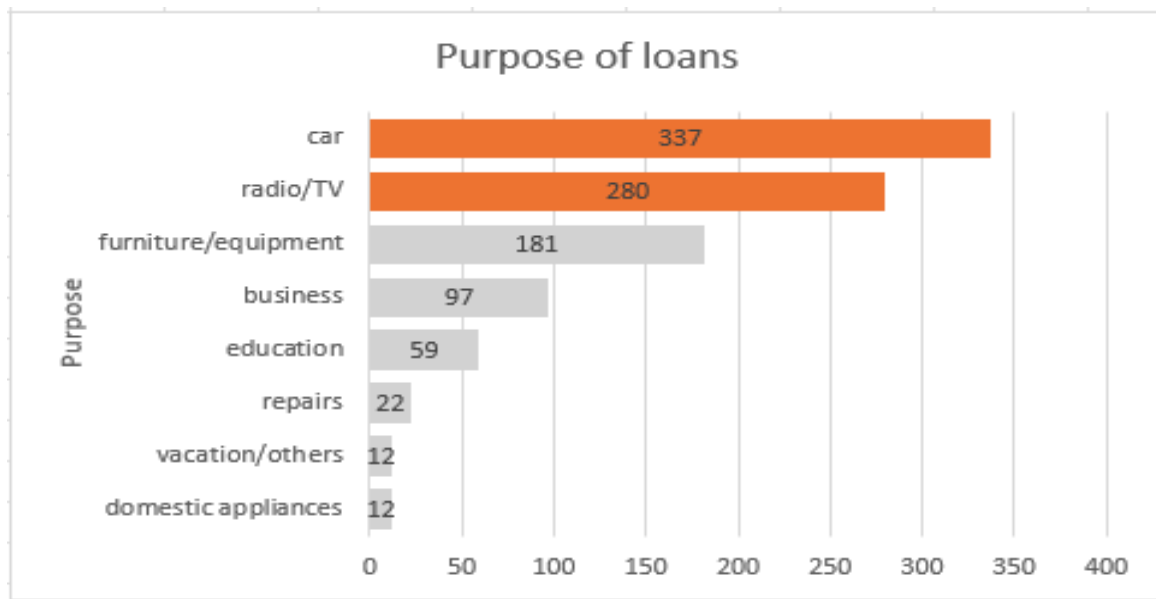
Interestingly, the unknown wealth customer group has the highest proportion of good-risk loaners at 88%, much higher than other groups, while supplying credit for little-wealth customers seems to be the riskiest as the ratio of good-risk customers in this group is just 51%. ✓



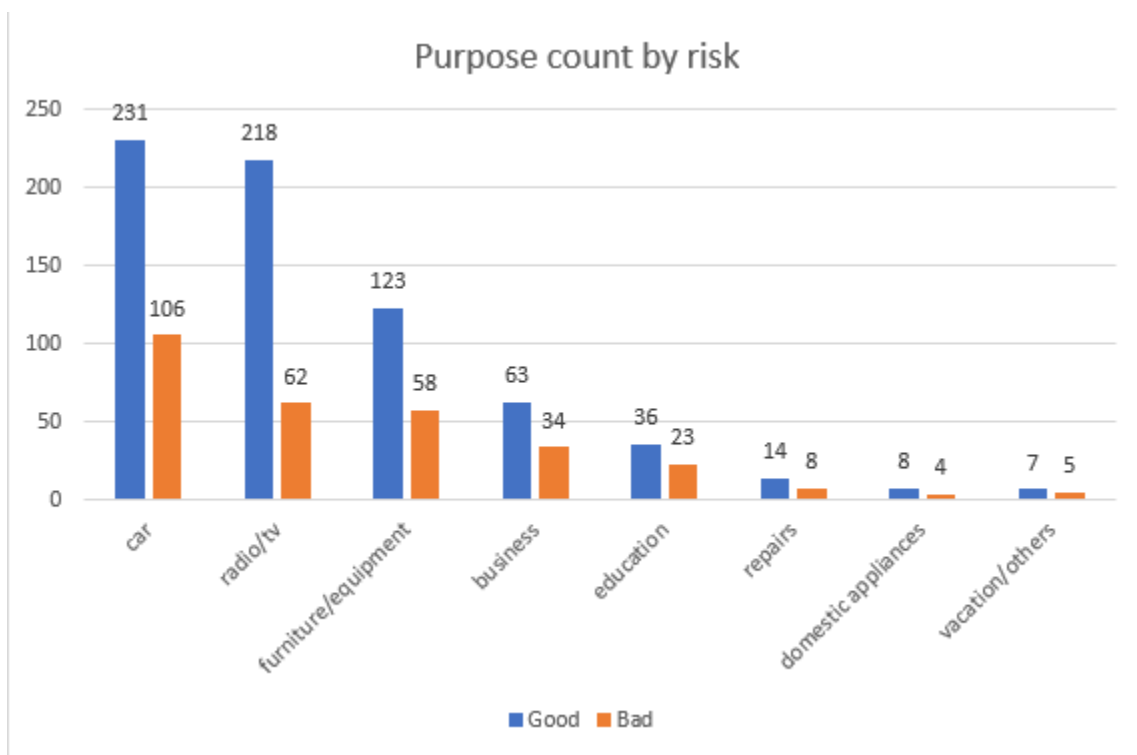


6. *Purpose of loan analysis*

In both genders and all age groups, the main purpose of loans is to purchase car and radio/TV

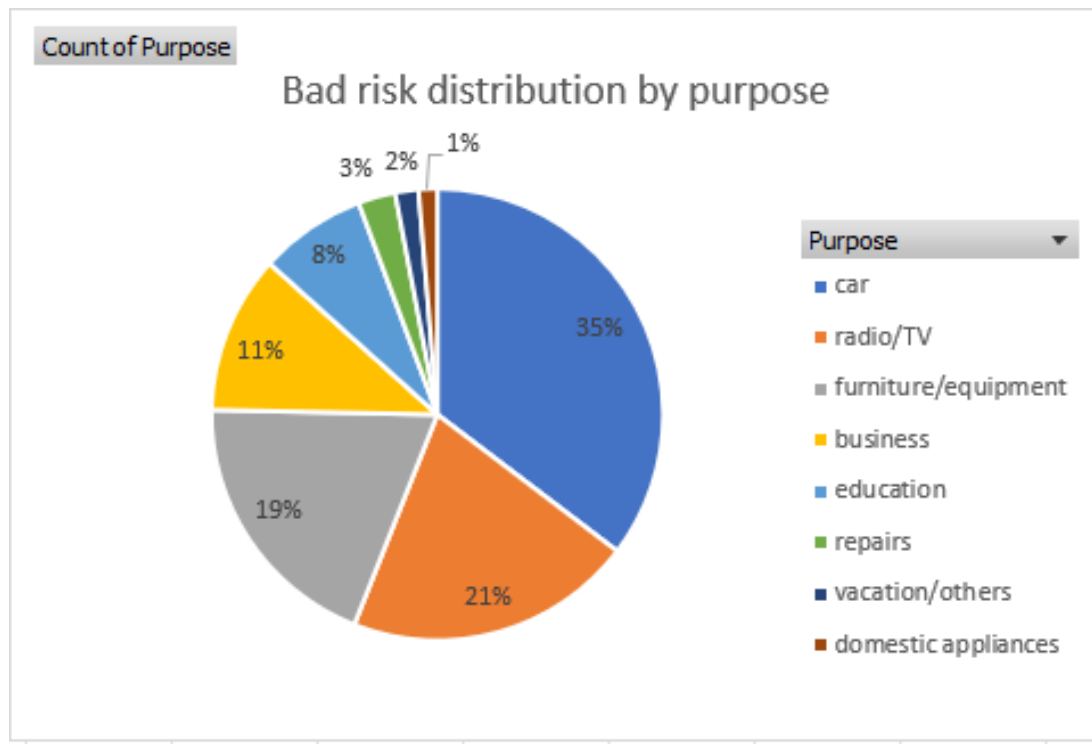


6.1. Risk by purpose of the loans



The ratio of bad risk to good risk of the loan for radio/TV is nearly 2:7, the smallest figure in the above chart. That means in 9 customers loaning money to

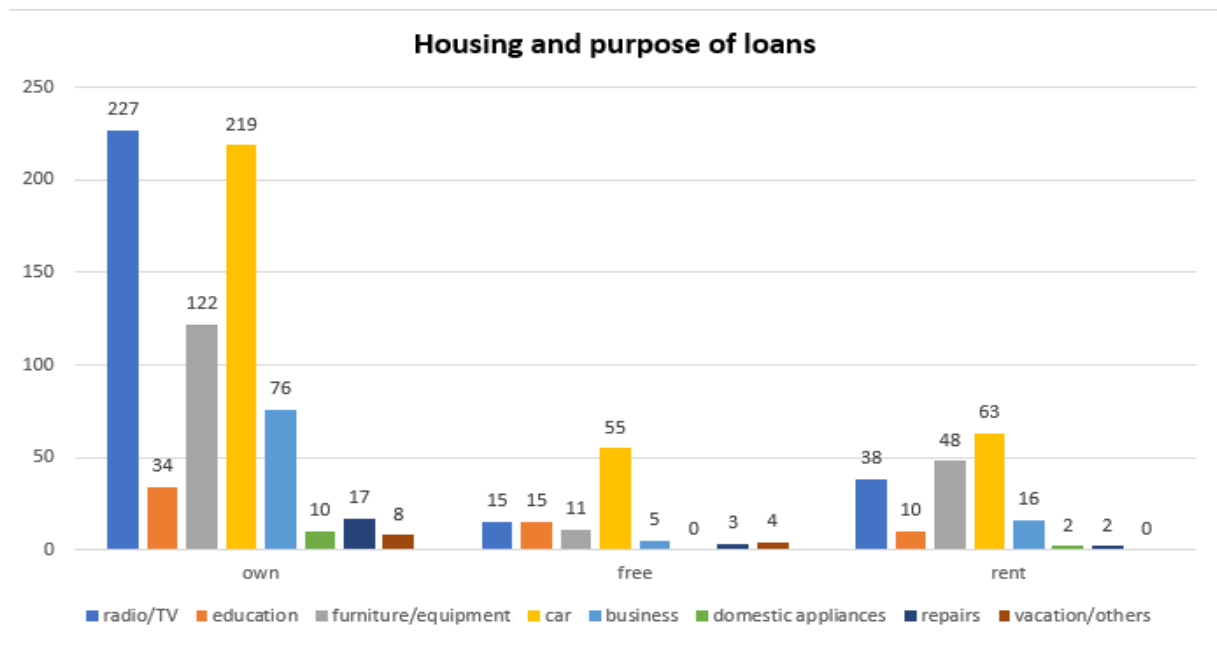
buy radio/TV, there are 2 high-risk customers and 7 low-risk customers. By contrast, these figures in loans for vacation/others and education are pretty high. ✓



It is noticeable that credit for car and radio/TV account for mostly bad risk cases in banks. The percentages of bad loans for car and radio/TV are 35% and 21% respectively. ✓

6.2. Purpose of loans and housing

The main purpose of the customers having a house when taking credit from banks is to purchase a radio/TV, followed by a car, whereas the customers renting a house or having no house prefer to apply for loans to buy a car. ✓

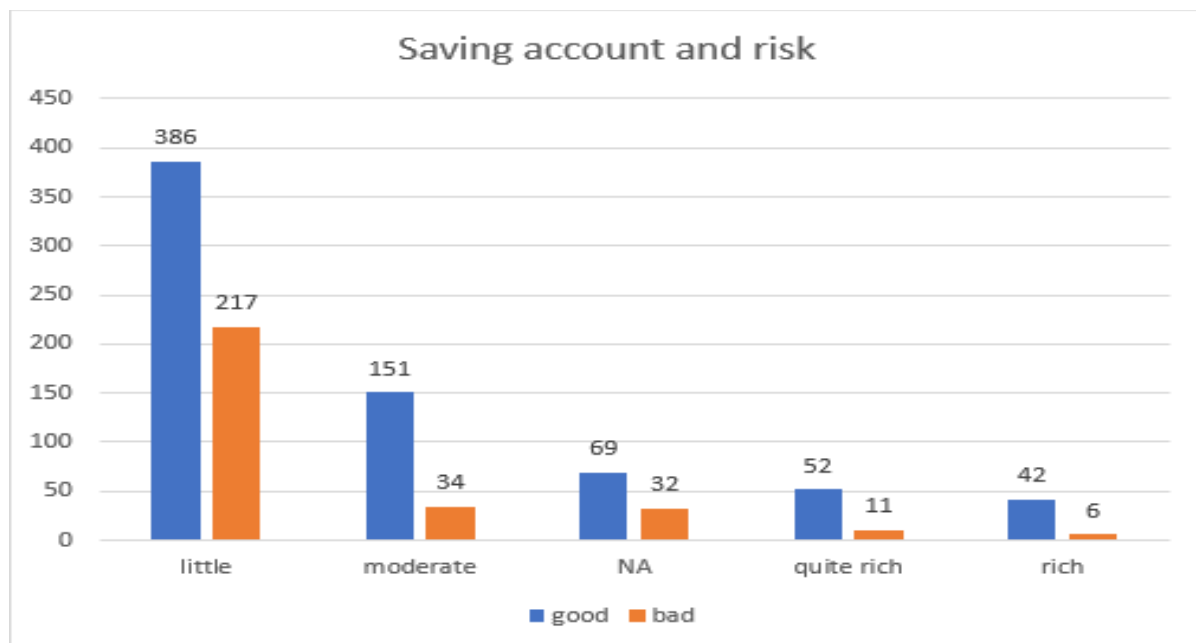


7. Risk analysis

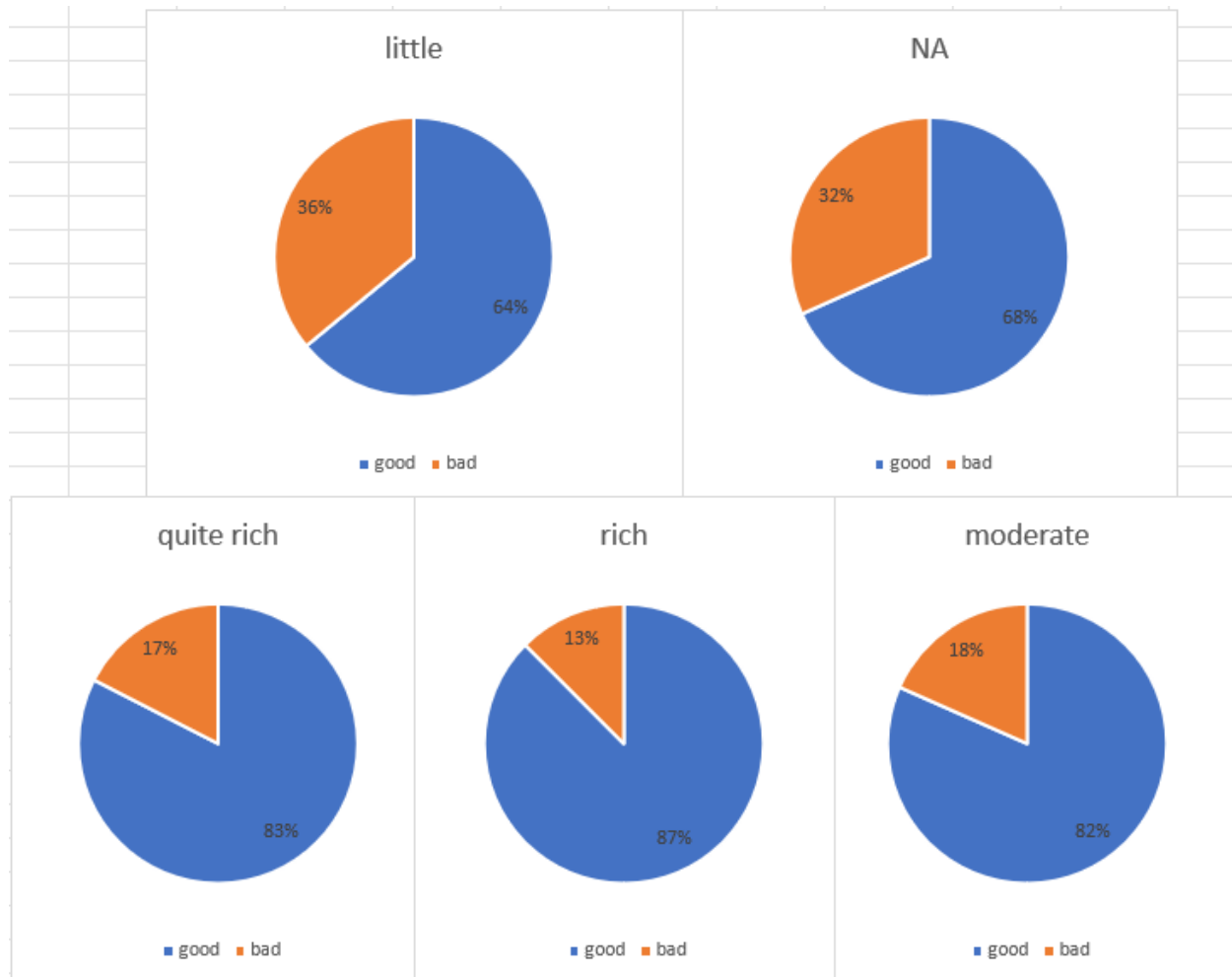


7.1. Risk and saving account

In the chart below, we can notice that the loaners with little saving accounts win a vantage position in comparison to others.



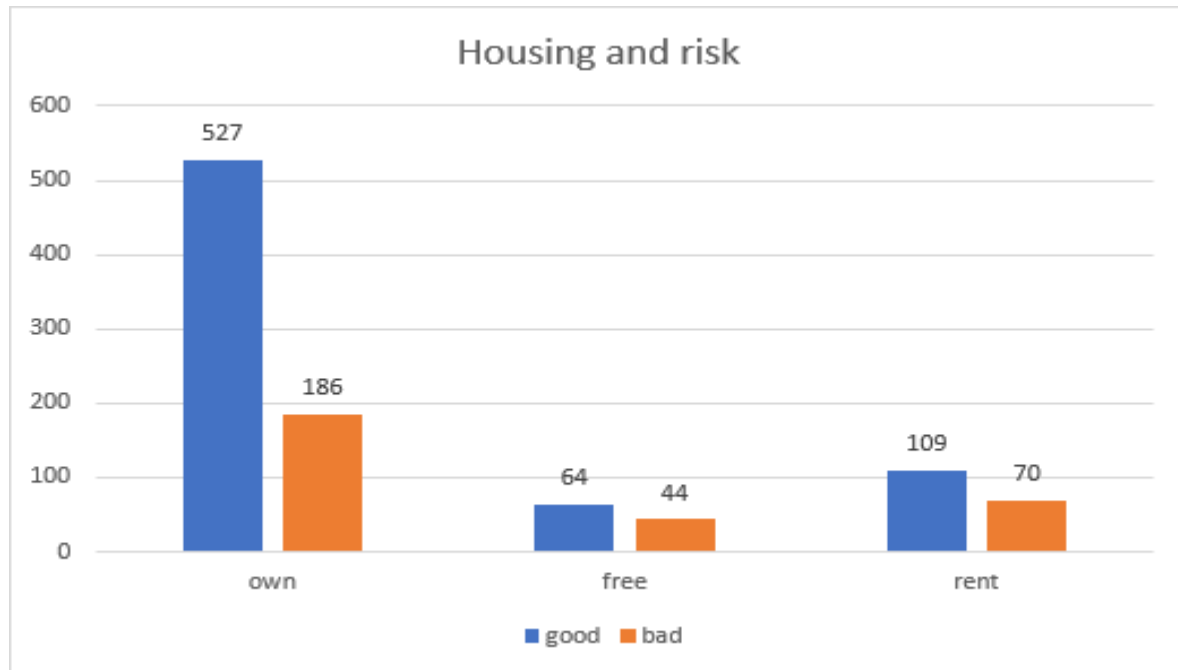
Comparing the percentages of bad risk and good risk in each group of customers by saving account status, we realize that giving credits to little saving account customers is riskier than others, while loaning money for quite rich and rich customers is the safest. However, these two groups of customers have less demand for borrowing money from banks than other groups. ✓



7.2. Risk and housing

Customers owning a house, or more are the target customers of banks to introduce loans and credit cards. Because it is the safest when banks give credit to this type

of customer. By contrast, banks are not willing to give loans to homeless people as the risk of these loans will be very high.



7.3. Conclusion

❖ Risky loan


In the previous analysis, we know that the highly risky customer groups are female, young, with little saving accounts, little checking accounts, and purpose on education or vacation/others. Let's check these conditions together and estimate how the risk.



	Age	Sex	Job	Housing	Saving acc	Checking	Credit an	Duration	Purpose	Risk
471	23	female	2	own	little	little	448	6	education	bad
640	27	female	0	own	little	little	750	18	education	bad
771	25	female	3	own	little	little	8,065	36	education	bad

We can recognize that 100% of young female customers who having little wealth status and loaning money for education is a bad risk.

❖ Good loan

Following the previous risk analysis by different groups of customers, the good loans are mainly credit for male, young adult customers with unknown checking accounts and rich status of saving accounts, having a house. 

	Age	Sex	Job	Housing	Saving acc	Checking	Credit an	Duration	Purpose	Risk
305	33	male		2 own	rich	NA	1,543	6	furniture/	good
567	34	male		2 own	rich	NA	2,578	24	radio/TV	good
778	38	male		3 own	rich	NA	5,711	36	car	good
863	32	male		1 own	rich	NA	4,526	27	furniture/	good

It is clear that 100% of male customers from 30 to 40 years old, having a rich status of saving accounts and unknown checking accounts is estimated as a good risk.

Index of comments

- 1.1 Only purpose is a text variable -
Sex, Housing, savings accounts, check accounts and risk is string variables