1_Numerical_Variables

September 4, 2021

0.1 What is a Variable?

A variable is any characteristic, number, or quantity that can be measured or counted. They are called 'variables' because the value they take may vary, and it usually does. The following are examples of variables:

- Age (21, 35, 62, …)
- Gender (male, female)
- Income (GBP 20000, GBP 35000, GBP 45000, …)
- House price (GBP 350000, GBP 570000, …)
- Country of birth (China, Russia, Costa Rica, …)
- Eye colour (brown, green, blue, …)
- Vehicle make (Ford, Volkswagen, …)

Most variables in a data set can be classified into one of two major types:

- **Numerical variables**
- **Categorical variables**

==========

0.2 Numerical Variables

The values of a numerical variable are numbers. They can be further classified into:

- **Discrete variables**
- **Continuous variables**

0.2.1 Discrete Variable

In a discrete variable, the values are whole numbers (counts). For example, the number of items bought by a customer in a supermarket is discrete. The customer can buy 1, 25, or 50 items, but not 3.7 items. It is always a round number. The following are examples of discrete variables:

- Number of active bank accounts of a borrower (1, 4, 7, …)
- Number of pets in the family
- Number of children in the family

0.2.2 Continuous Variable

A variable that may contain any value within a range is continuous. For example, the total amount paid by a customer in a supermarket is continuous. The customer can pay, GBP 20.5, GBP 13.10, GBP 83.20 and so on. Other examples of continuous variables are:

- House price (in principle, it can take any value) (GBP 350000, 57000, 100000, …)
- Time spent surfing a website (3.4 seconds, 5.10 seconds, …)
- Total debt as percentage of total income in the last month (0.2, 0.001, 0, 0.75, …)

==========

0.3 In this demo: Peer to peer lending (Finance)

In this demo, we will use a toy data set which simulates data from a peer-o-peer finance company to inspect discrete and continuous numerical variables.

- You should have downloaded the **Datasets** together with the Jupyter notebooks in **Section 1**.

[1]:
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
```

[2]:
```
# let's load the dataset

# Variable definitions:
#-------------------------
# disbursed_amount: loan amount given to the borrower
# interest: interest rate
# income: annual income
# number_open_accounts: open accounts (more on this later)
# number_credit_lines_12: accounts opened in the last 12 months
# target: loan status(paid or being repaid = 1, defaulted = 0)


data = pd.read_csv('./loan.csv')

data.head()
```

[2]:

	customer_id	disbursed_amount	interest	market	employment	time_employed	householder	income	date_issued	target	loan_purpose
0	0	23201.5	15.4840	C	Teacher	<=5 years	RENT	84600.0	2013-06-11	0	Debt consolidation
1	1	7425.0	11.2032	B	Accountant	<=5 years	OWNER	102000.0	2014-05-08	0	Car purchase
2	2	11150.0	8.5100	A	Statistician	<=5 years	RENT	69840.0	2013-10-26	0	Debt consolidation
3	3	7600.0	5.8656	A	Other	<=5 years	RENT	100386.0	2015-08-20	0	Debt consolidation
4	4	31960.0	18.7392	E	Bus driver	>5 years	RENT	95040.0	2014-07-22	0	Debt consolidation

	number_open_accounts	date_last_payment	number_credit_lines_12
0	4	2016-01-14	NaN
1	13	2016-01-25	NaN
2	8	2014-09-26	NaN
3	20	2016-01-26	NaN
4	14	2016-01-11	NaN

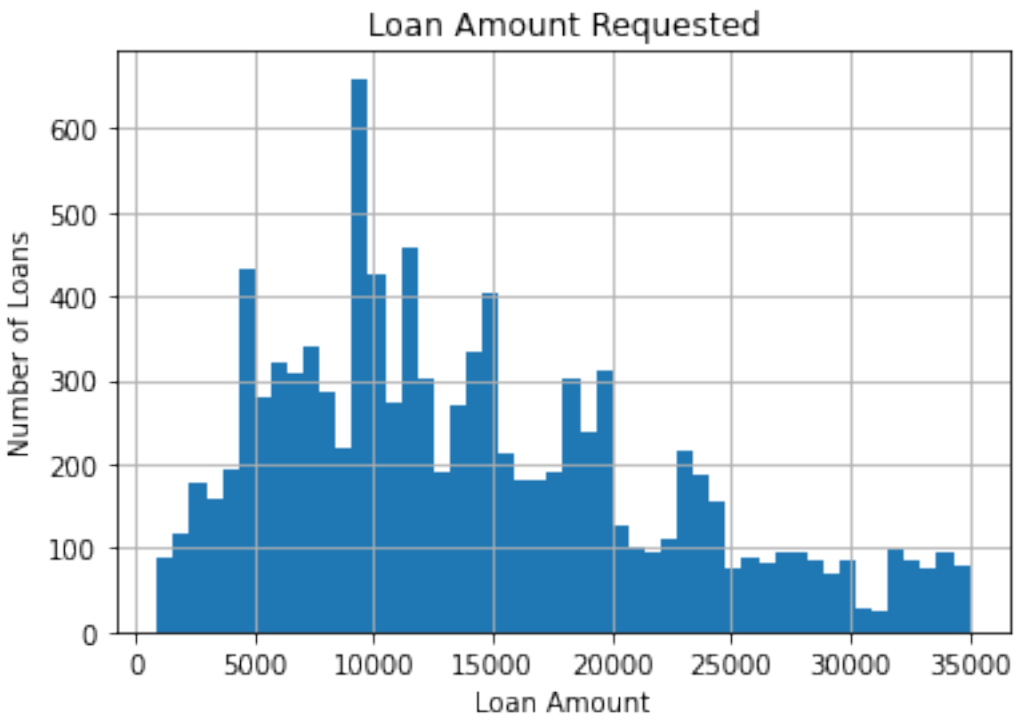
0.3.1 Continuous Variables

[3]:
```
# let's look at the values of the variable disbursed_amount
# this is the amount of money requested by the borrower

# this variable is continuous, it can take in principle
# any value

data['disbursed_amount'].unique()
```

[3]:
```
array([23201.5 ,  7425.  , 11150.  , …,  6279.  , 12894.75, 25584.  ])
```

[4]:
```
# let's make a histogram to get familiar with the
# distribution of the variable

fig = data['disbursed_amount'].hist(bins=50)

fig.set_title('Loan Amount Requested')
fig.set_xlabel('Loan Amount')
fig.set_ylabel('Number of Loans')
```

[4]:
```
Text(0, 0.5, 'Number of Loans')
```



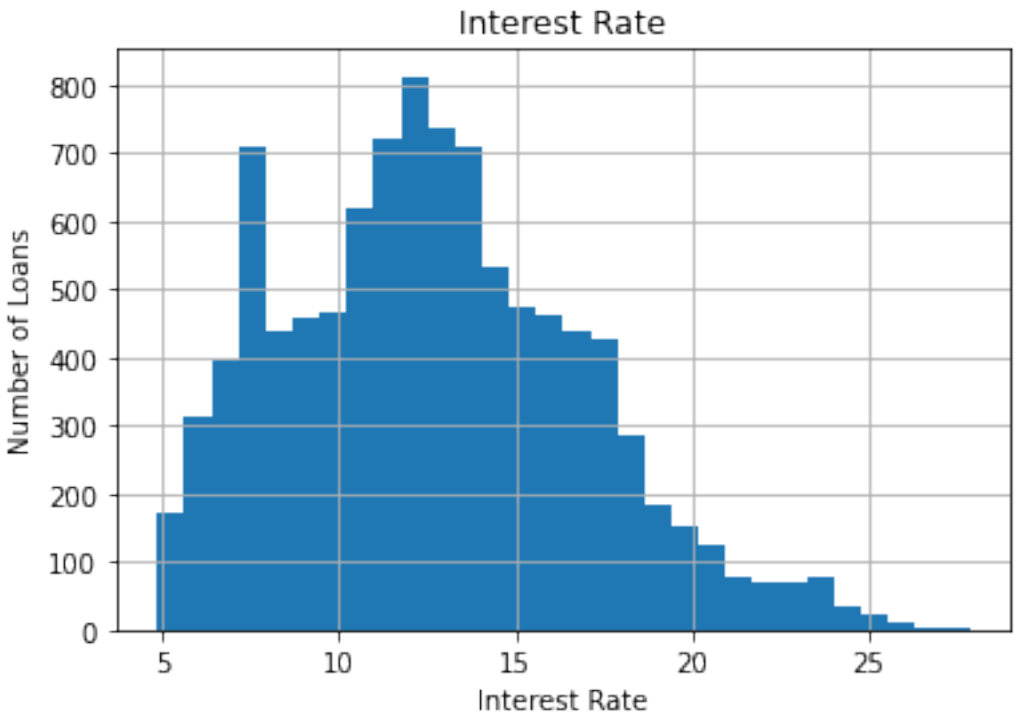
The values of the variable vary across the entire range of loan amounts typically disbursed to borrowers. This is characteristic of continuous variables.

[5]:
```
# let's do the same exercise for the variable interest rate,
# which is the interest charged by the finance company to the borrowers

# this variable is also continuous, it can take in principle
# any value within the range

data['interest'].unique()
```

[5]:
```
array([15.484 , 11.2032,  8.51  , …, 12.9195, 11.2332, 11.0019])
```

[6]:
```
# let's make a histogram to get familiar with the
# distribution of the variable

fig = data['interest'].hist(bins=30)

fig.set_title('Interest Rate')
fig.set_xlabel('Interest Rate')
fig.set_ylabel('Number of Loans')
```

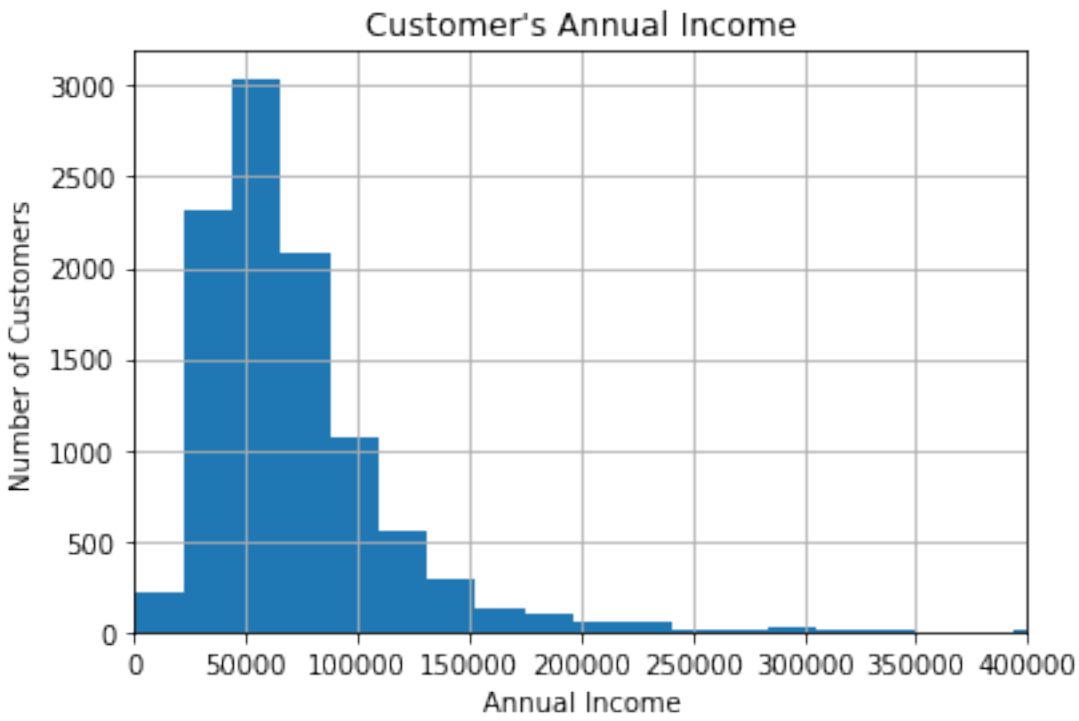
[6]:
```
Text(0, 0.5, 'Number of Loans')
```



We see that the values of the variable vary continuously across the variable range. The values are the interest rate charged to borrowers.

[7]:
```
# Now, let's explore the income declared by the customers,
# that is, how much they earn yearly.

# this variable is also continuous

fig = data['income'].hist(bins=100)

# for better visualisation, I display only specific
# range in the x-axis
fig.set_xlim(0, 400000)

# title and axis legends
fig.set_title("Customer's Annual Income")
fig.set_xlabel('Annual Income')
fig.set_ylabel('Number of Customers')
```

[7]:
```
Text(0, 0.5, 'Number of Customers')
```



The majority of salaries are concentrated towards values in the range 30-70k, with only a few customers earning higher salaries. The values of the variable, vary continuously across the variable range, because this is a continuous variable.

0.3.2 Discrete Variables

Let's explore the variable "Number of open credit lines in the borrower's credit file" (number_open_accounts in the dataset).

This variable represents the total number of credit items (for example, credit cards, car loans, mortgages, etc) that is known for that borrower.

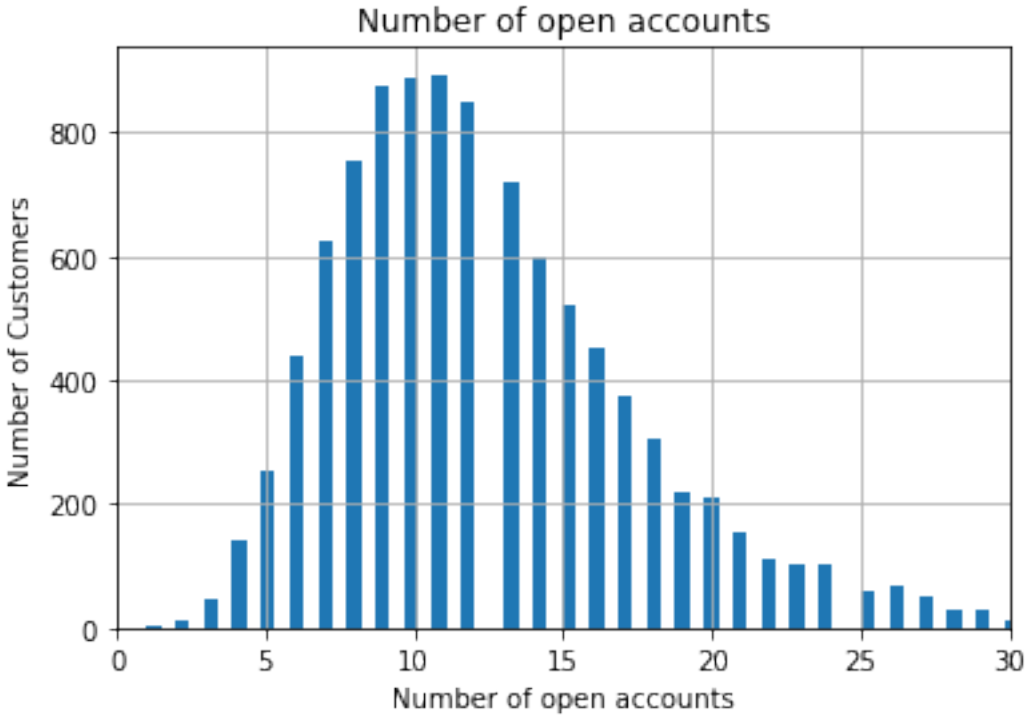
By definition it is a discrete variable, because a borrower can have 1 credit card, but not 3.5 credit cards.

[8]:
```
# let's inspect the values of the variable

# this is a discrete variable

data['number_open_accounts'].dropna().unique()
```

[8]:
```
array([ 4, 13,  8, 20, 14,  5,  9, 18, 16, 17, 12, 15,  6, 10, 11,  7, 21,
       19, 26,  2, 22, 27, 23, 25, 24, 28,  3, 30, 41, 32, 33, 31, 29, 37,
       49, 34, 35, 38,  1, 36, 42, 47, 40, 44, 43])
```

[9]:
```
# let's make an histogram to get familiar with the
# distribution of the variable

fig = data['number_open_accounts'].hist(bins=100)

# for better visualisation, I display only specific
# range in the x-axis
fig.set_xlim(0, 30)

# title and axis legends
fig.set_title('Number of open accounts')
fig.set_xlabel('Number of open accounts')
fig.set_ylabel('Number of Customers')
```

[9]:
```
Text(0, 0.5, 'Number of Customers')
```



Histograms of discrete variables have this typical broken shape, as not all the values within the variable range are present in the variable. As I said, the customer can have 3 credit cards, but not 3,5 credit cards.

Let's look at another example of a discrete variable in this dataset: **Number of installment accounts opened in past 12 months** ('number_credit_lines_12' in the dataset).

Installment accounts are those that at the moment of acquiring them, there is a set period and amount of repayments agreed between the lender and borrower. An example of this is a car loan,

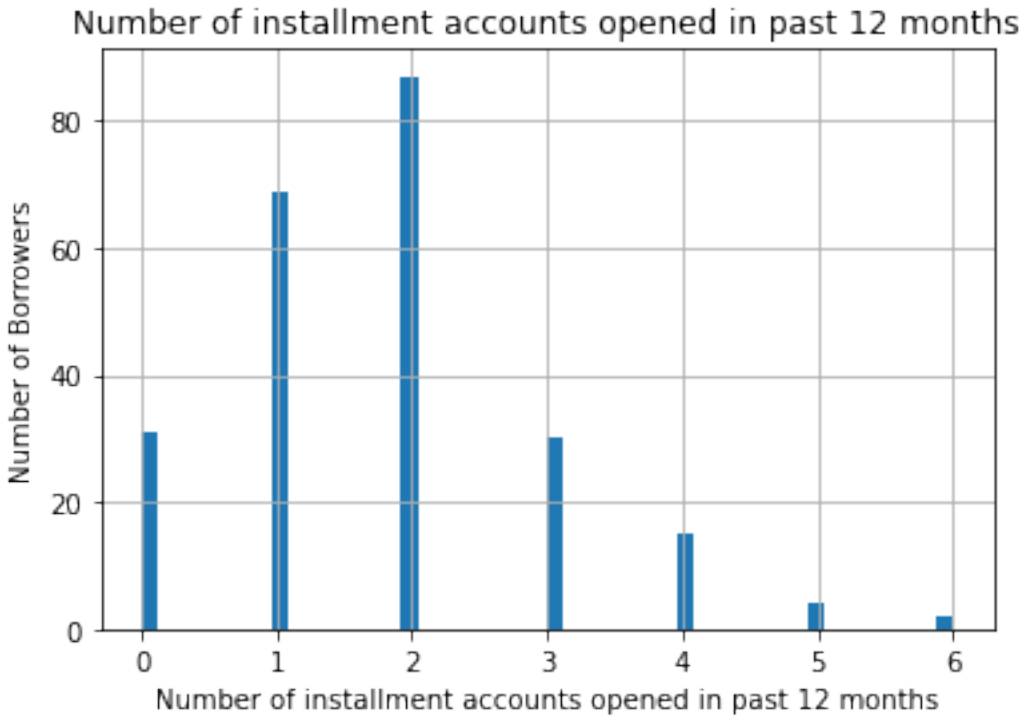
or a student loan. The borrower knows that they will pay a fixed amount over a fixed period, for example 36 months.

[10]:
```
# let's inspect the variable values

data['number_credit_lines_12'].unique()
```

[10]:
```
array([nan,  2.,  4.,  1.,  0.,  3.,  5.,  6.])
```

[11]:
```
# let's make a histogram to get familiar with the
# distribution of the variable

fig = data['number_credit_lines_12'].hist(bins=50)
fig.set_title('Number of installment accounts opened in past 12 months')
fig.set_xlabel('Number of installment accounts opened in past 12 months')
fig.set_ylabel('Number of Borrowers')
```

[11]:
```
Text(0, 0.5, 'Number of Borrowers')
```



The majority of the borrowers have none or 1 installment account, with only a few borrowers having more than 2.

0.3.3 A variation of discrete variables: the binary variable

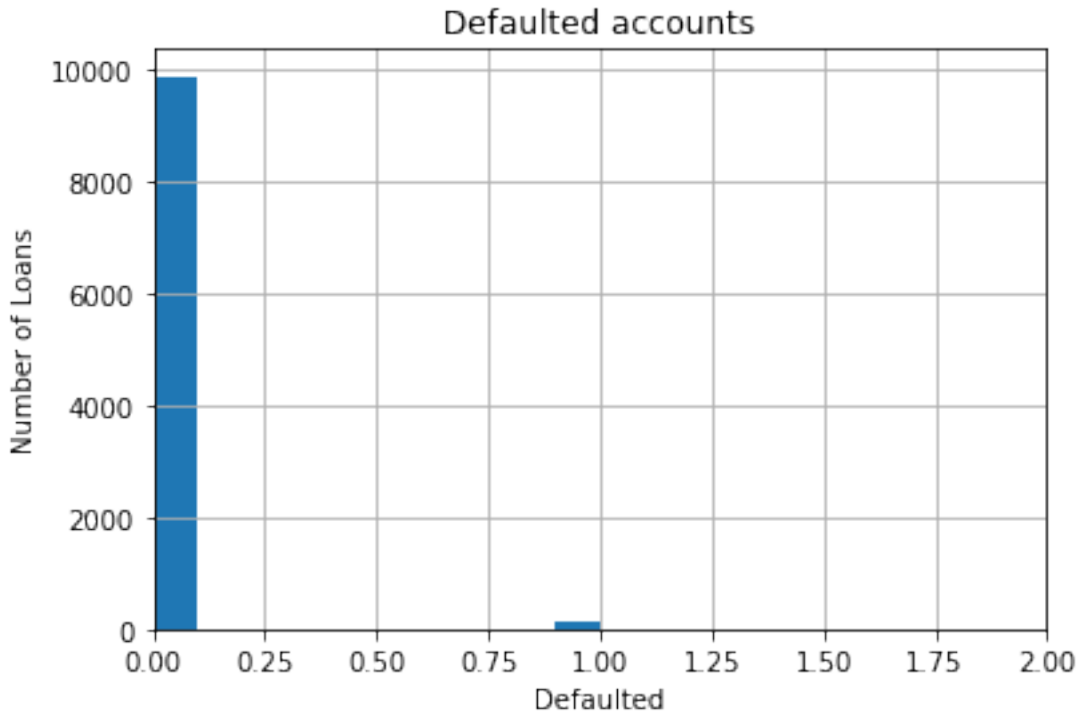
Binary variables, are discrete variables, that can take only 2 values, therefore binary.

[12]:
```
# A binary variable, can take 2 values. For example in
# the variable "target":
# either the loan is defaulted (1) or not (0)

data['target'].unique()
```

[12]:
```
array([0, 1])
```

[13]:
```
# let's make a histogram, although histograms for
# binary variables do not make a lot of sense

fig = data['target'].hist()
fig.set_xlim(0, 2)
fig.set_title('Defaulted accounts')
fig.set_xlabel('Defaulted')
fig.set_ylabel('Number of Loans')
```

[13]:
```
Text(0, 0.5, 'Number of Loans')
```



As we can see, the variable shows only 2 values, 0 and 1, and the majority of the loans are OK.

That is all for this demonstration. I hope you enjoyed the notebook, and see you in the next one.