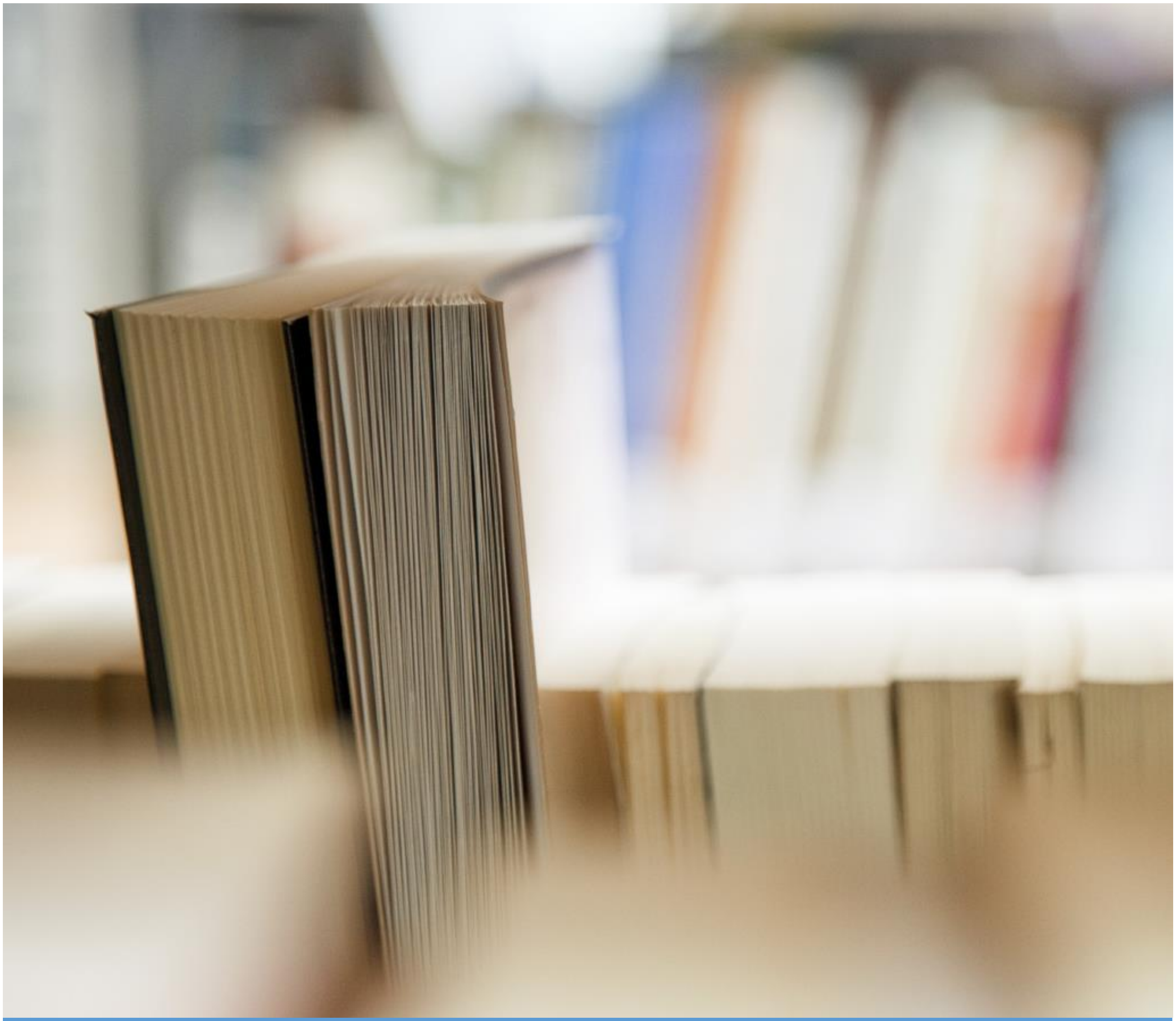


BUSINESS ANALYTICS ASSIGNMENT

Sharon Mutia

SECTION 1: LOANS DATA SET ANALYSIS

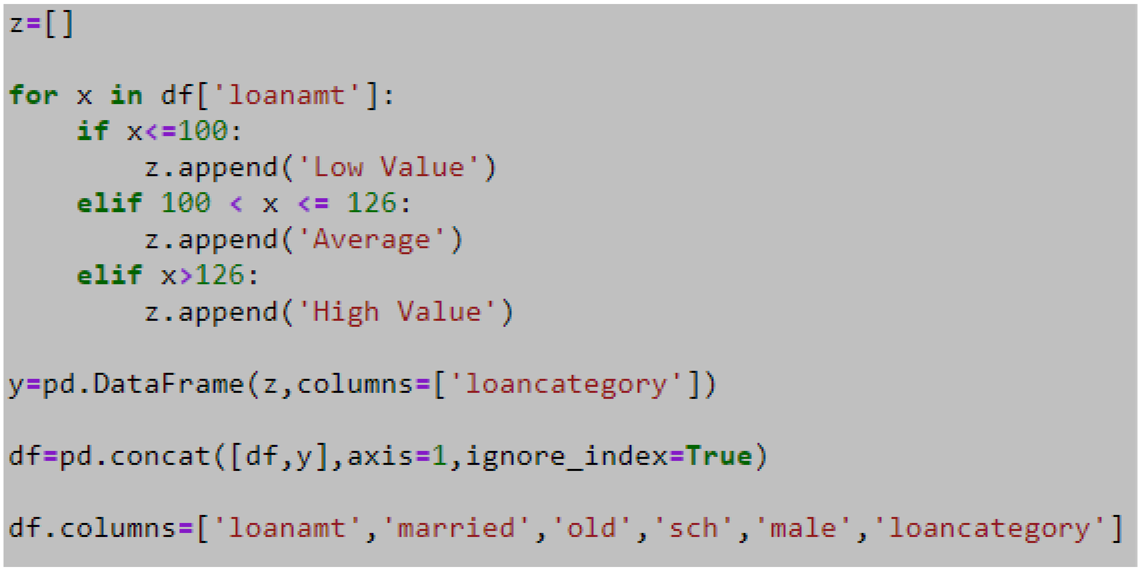
1. The Data

The initial data set was composed of 1989 observations and 59 data attributes. The data set describes loan applications and their respective loan applicants.

This study analyses 5 key variables focusing on loan applicant descriptions and how they affect the loan amount applicants requested. The variables used are as below:

	Variables	Data Type	Proxy Name
1	Loan amount	Continuous	loanamt
2	Marital status	Categorical	married
3	Age	Categorical	old
4	Education status	Categorical	sch
5	Gender	Categorical	male

I further added a variable called loan category which classified the loan amounts as low-value, average or high-value loans using the quartiles from the 'loan amount' data. The diagram below details how the variable was created.

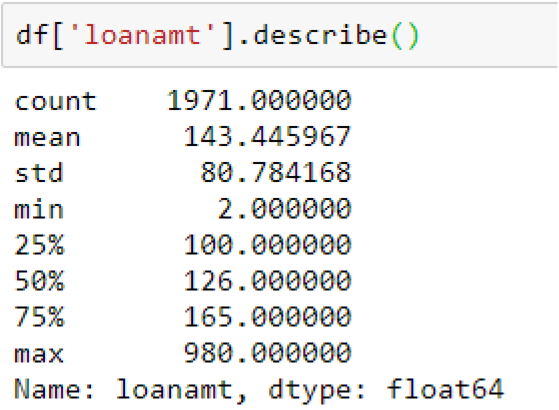


```
z=[]

for x in df['loanamt']:
    if x<=100:
        z.append('Low Value')
    elif 100 < x <= 126:
        z.append('Average')
    elif x>126:
        z.append('High Value')

y=pd.DataFrame(z,columns=['loancategory'])

df=pd.concat([df,y],axis=1,ignore_index=True)

df.columns=['loanamt','married','old','sch','male','loancategory']
```

2. Descriptive Analysis

a.) Continuous Data Descriptive Analysis

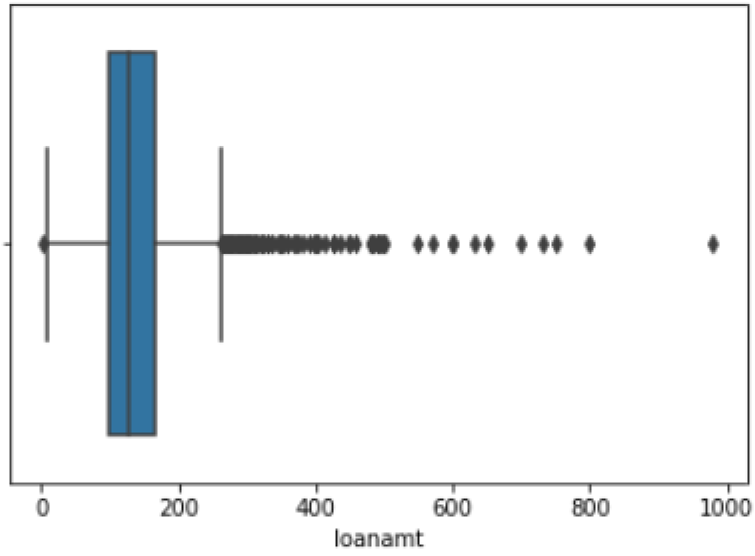
As evidenced below, the loans extended range between a value of 2 and 980, although the denominations of the value is not specified. For the purposes of this study, it is assumed that the loan value is in dollars.



```
df['loanamt'].describe()
```

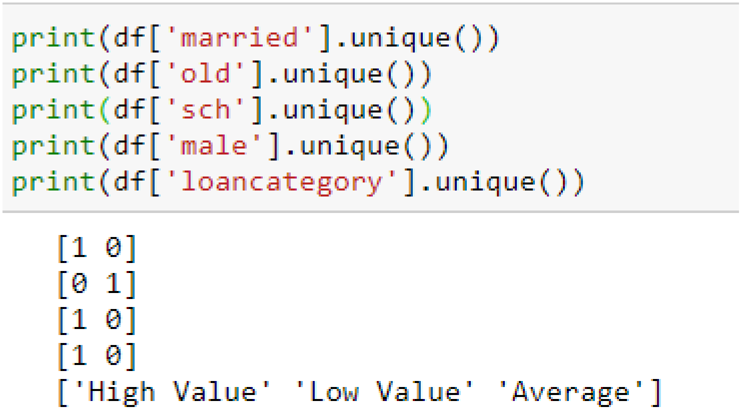
```
count    1971.000000
mean      143.445967
std        80.784168
min         2.000000
25%       100.000000
50%       126.000000
75%       165.000000
max       980.000000
Name: loanamt, dtype: float64
```

The average value of loans is roughly $143.45 although this varies by about $80.78. Further, 75% of the loans are valued at less than $165. The boxplot further emphasizes the latter as it shows that loans valued at more than $300 are in fact outliers.



b.) Categorical Data Descriptive Analysis

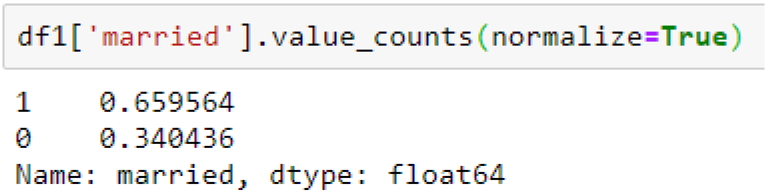
The data consists of 4 categorical variables that take on either two or three values. The marital status, age, education status and gender are Boolean variables taking values of either zero or one. The loan category variable is also a categorical variable that can take 3 possible values- high-value, low-value or average.



```
print(df['married'].unique())
print(df['old'].unique())
print(df['sch'].unique())
print(df['male'].unique())
print(df['loancategory'].unique())
```

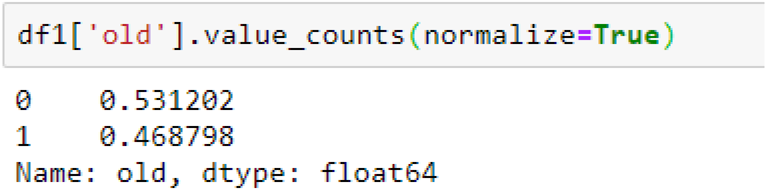
```
[1 0]
[0 1]
[1 0]
[1 0]
['High Value' 'Low Value' 'Average']
```

The data is not perfectly balanced across categories, though it does include some level of variation.



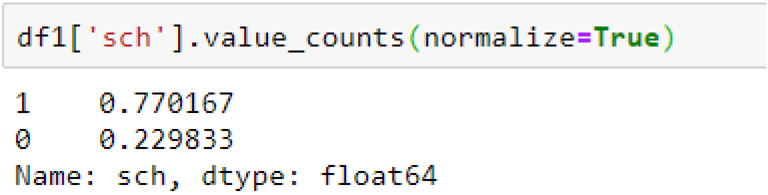
```
df1['married'].value_counts(normalize=True)
```

```
1    0.659564
0    0.340436
Name: married, dtype: float64
```



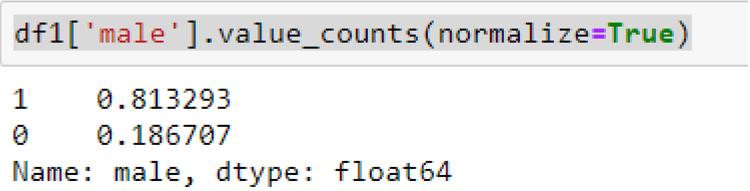
```
df1['old'].value_counts(normalize=True)
```

```
0    0.531202
1    0.468798
Name: old, dtype: float64
```

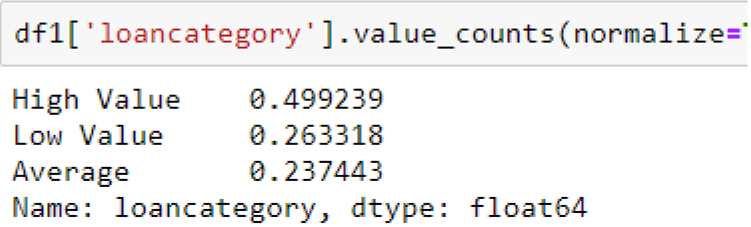


```
df1['sch'].value_counts(normalize=True)
```

```
1    0.770167
0    0.229833
Name: sch, dtype: float64
```



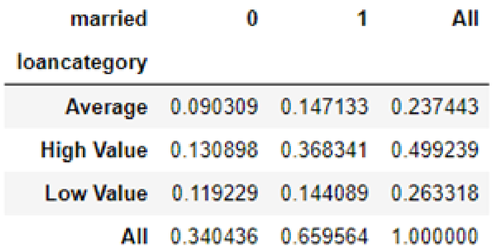
```
df1['male'].value_counts(normalize=True)

1    0.813293
0    0.186707
Name: male, dtype: float64
```



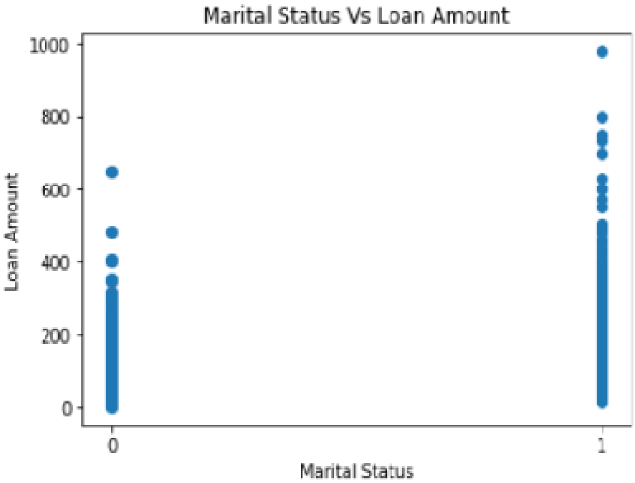
```
df1['loancategory'].value_counts(normalize=

High Value    0.499239
Low Value     0.263318
Average       0.237443
Name: loancategory, dtype: float64
```

c.) Contingency Tables and Visualizations

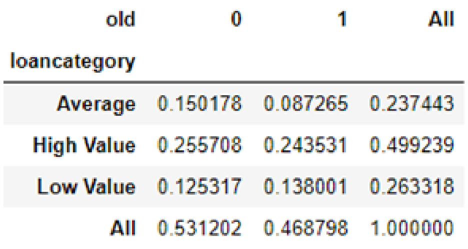
From the contingency table and visualization below, it is evident that married people are more likely to make loan applications compared to unmarried people. Further, married people are more likely to apply for higher value loans.



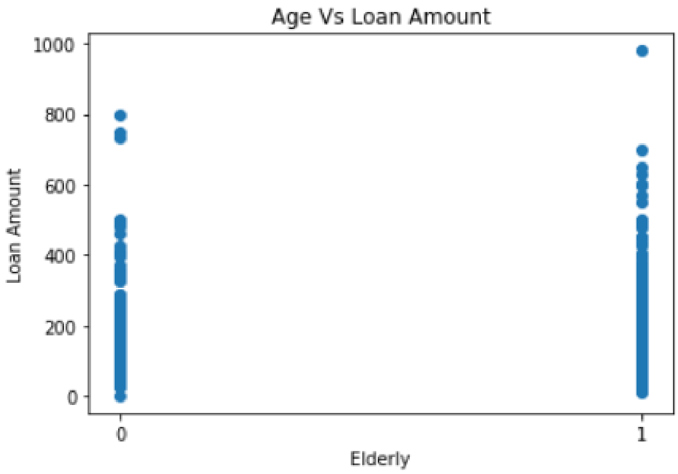
married loancategory	0	1	All
Average	0.090309	0.147133	0.237443
High Value	0.130898	0.368341	0.499239
Low Value	0.119229	0.144089	0.263318
All	0.340436	0.659564	1.000000



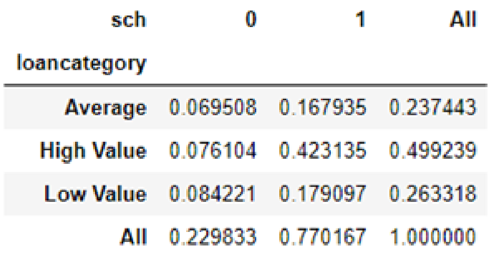
With regards to the age variable, it appears that there isn’t much of a significant difference in loan applications of older and younger people.



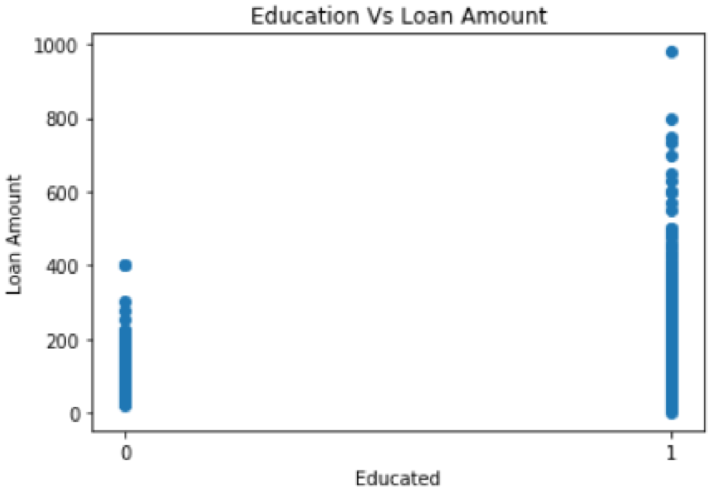
old	0	1	All
loancategory			
Average	0.150178	0.087265	0.237443
High Value	0.255708	0.243531	0.499239
Low Value	0.125317	0.138001	0.263318
All	0.531202	0.468798	1.000000



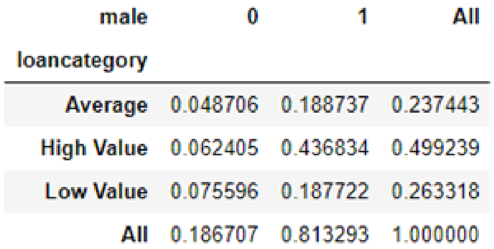
The visualizations below show that educated people are more likely to seek for loans (0.23 vs 0.77 probability). The latter trend is evident across all loan categories (low, high and average valued loans).



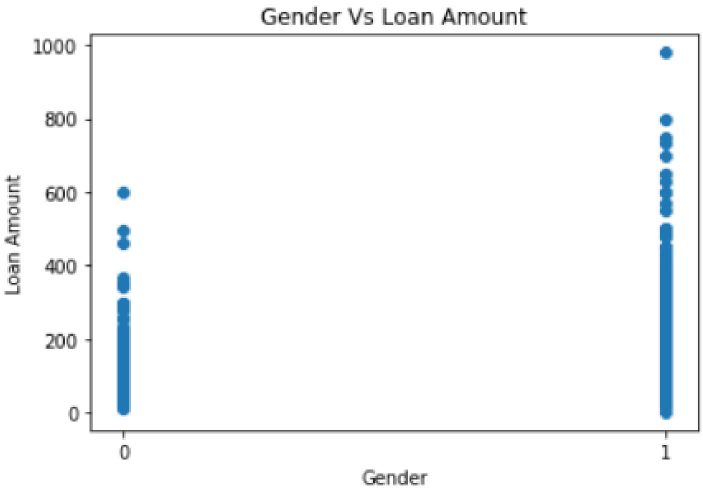
sch	0	1	All
loancategory			
Average	0.069508	0.167935	0.237443
High Value	0.076104	0.423135	0.499239
Low Value	0.084221	0.179097	0.263318
All	0.229833	0.770167	1.000000



Below, we see that men are also more likely to make loan applications than women are. The latter is particularly evident for high value loans.



male	0	1	All
loancategory			
Average	0.048706	0.188737	0.237443
High Value	0.062405	0.436834	0.499239
Low Value	0.075596	0.187722	0.263318
All	0.186707	0.813293	1.000000



3. Hypothesis Testing

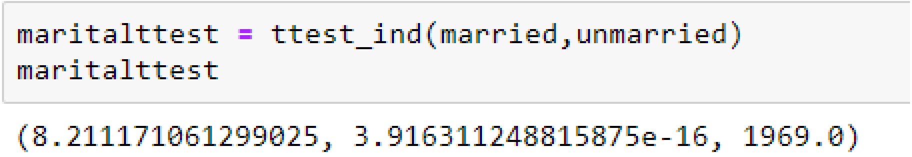
From the data exploration above, the following hypotheses were tested using a 2 sample t-test. The results of the tests were interpreted using the p-value and the t-statistic at the 95% confidence interval.

1. ***There is no statistically significant difference in the loan amount sought by married and unmarried applicants***.

H_0: Loan amount (married) = Loan amount (unmarried)

H_a: Loan amount (married) ≠ Loan amount (unmarried)

RESULT:



```
maritalttest = ttest_ind(married,unmarried)
maritalttest
```

```
(8.211171061299025, 3.916311248815875e-16, 1969.0)
```

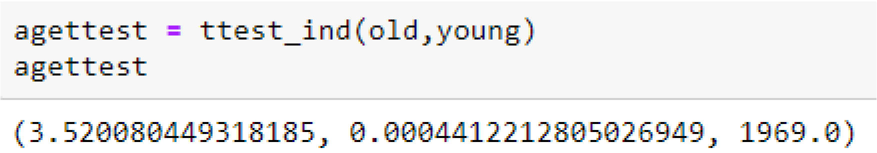
The p-value is below the 5% level of significance therefore we reject the null hypothesis that there is no significant statistical difference in the loan amount demanded from by married and unmarried people. Similarly the t-static (8.21) is greater than 1.96 (the critical value at a 95% confidence interval). Therefore we also reject the null hypothesis on the basis of the t-test. This means the average loan amounts requested by married and unmarried people are statistically different.

2. ***There is no statistically significant difference in the loan amount sought by older and younger applicants.***

H_0: Loan amount (young) = Loan amount (old)

H_a: Loan amount (young) ≠ Loan amount (old)

RESULT:



```
agettest = ttest_ind(old,young)
agettest
```

```
(3.520080449318185, 0.0004412212805026949, 1969.0)
```

The p-value is below the 5% level of significance therefore we reject the null hypothesis that there is no significant statistical difference in the loan amount sought by younger or more mature loan applicants.

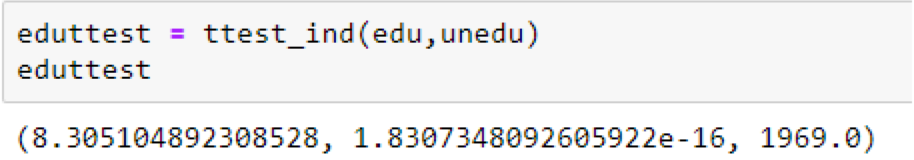
The above results are also affirmed by the t-test which yielded a t-statistic above the 1.96 critical value. This means that average loan amounts requested by young and old people are statistically different.

3. ***There is no statistically significant difference in the loan amount sought by educated and uneducated applicants.***

H_0: Loan amount (educated) = Loan amount (uneducated)

H_a: Loan amount (educated) ≠ Loan amount (uneducated)

RESULT:



```
eduttest = ttest_ind(edu,unedu)
eduttest
```

```
(8.305104892308528, 1.8307348092605922e-16, 1969.0)
```

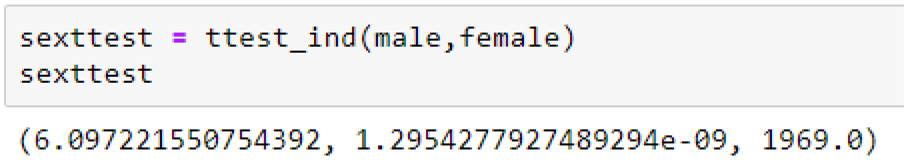
The p-value is below the 5% level of significance therefore we reject the null hypothesis that there is no significant statistical difference in the loan amount sought by educated and uneducated loan applicants.

The t-test supports the latter claims yielding a t-statistic of 8.3 (above the 1.96 critical value). This means that the average loan amounts requested by educated and uneducated people are statistically different.

4. ***There is no statistically significant difference in the loan amount sought by male and female applicants.***

H_0: Loan amount (male) = Loan amount (female)
H_a: Loan amount (male) ≠ Loan amount (female)

RESULT:



```
sexttest = ttest_ind(male,female)
sexttest
```

```
(6.097221550754392, 1.2954277927489294e-09, 1969.0)
```

The p-value is below the 5% level of significance therefore we reject the null hypothesis that there is no significant statistical difference in the loan amount sought by male and female loan applicants.

Using the t-test we also reject the null hypothesis since the calculated t-statistic above the 1.96 critical value. This means that loan amounts requested by men and women are statistically different.

SECTION 2: MICHAEL JORDAN CASE STUDY

1. Why do you think my experiment with Michael Jordan resulted in a non-significant p-value?

Given the number of tries in the experiment, there is not enough evidence for the research to conclude that there is no statistical difference between the researcher’s ability to shoot hoops and that of Michael Jordan. The larger calculated p-value implies a higher probability of observing the data if the null hypothesis is true.

A non-significant p-value, could result either from a true null hypothesis meaning that the researcher is actually just as good as Michael Jordan; or there is insufficient evidence to reject the null meaning that there are other factors that limit us from rejecting the hypothesis that the researcher is just as good as Michael Jordan . In this particular example, the number of trials could be a limiting factor

2. What is the connection between a criminal trial and a p-value?

In criminal trials, the legislature is usually trying to prove if the suspected offender is guilty or not guilty. The latter statements are essentially hypotheses that the legal teams must support with evidence. For the judge to declare the defendant as guilty, he must be certain beyond reasonable doubt that defendant has committed the crime. A verdict of not guilty is also rendered either if the defendant is actually innocent or if there is insufficient evidence to convict him as guilty. This is similar to p-values where hypotheses are either rejected or failed to be rejected on the bac of insufficient evidence.