EDA Task1: Data Overview

```
#importing Required library
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

#importing plotly Library
from plotly.offline import iplot
import plotly as py
import plotly.tools as tls
import cufflinks as cf
py.offline.init_notebook_mode(connected=True) #Turning on notebook
mode
cf.go_offline()

df=pd.read_csv(r"Data_set.csv") #dataset

df.shape

(40, 24)

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 24 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   gender                                40 non-null     object
 1   age                                   40 non-null     int64
 2   Investment_Avenues                    40 non-null     object
 3   Mutual_Funds                          40 non-null     int64
 4   Equity_Market                         40 non-null     int64
 5   Debentures                            40 non-null     int64
 6   Government_Bonds                      40 non-null     int64
 7   Fixed_Deposits                        40 non-null     int64
 8   PPF                                   40 non-null     int64
 9   Gold                                  40 non-null     int64
 10  Stock_Marktet                         40 non-null     object
 11  Factor                                40 non-null     object
 12  Objective                             40 non-null     object
 13  Purpose                               40 non-null     object
 14  Duration                              40 non-null     object
 15  Invest_Monitor                        40 non-null     object
 16  Expect                                40 non-null     object
 17  Avenue                                40 non-null     object
```

```
 18  What are your savings objectives?  40 non-null     object
 19  Reason_Equity                      40 non-null     object
 20  Reason_Mutual                      40 non-null     object
 21  Reason_Bonds                       40 non-null     object
 22  Reason_FD                          40 non-null     object
 23  Source                             40 non-null     object
dtypes: int64(8), object(16)
memory usage: 7.6+ KB
```

```
df.isnull().sum() # check for the null values
```

```
gender                                0
age                                   0
Investment_Avenues                    0
Mutual_Funds                          0
Equity_Market                         0
Debentures                            0
Government_Bonds                      0
Fixed_Deposits                        0
PPF                                   0
Gold                                  0
Stock_Marktet                         0
Factor                                0
Objective                             0
Purpose                               0
Duration                              0
Invest_Monitor                        0
Expect                                0
Avenue                                0
What are your savings objectives?     0
Reason_Equity                         0
Reason_Mutual                         0
Reason_Bonds                          0
Reason_FD                             0
Source                                0
dtype: int64
```

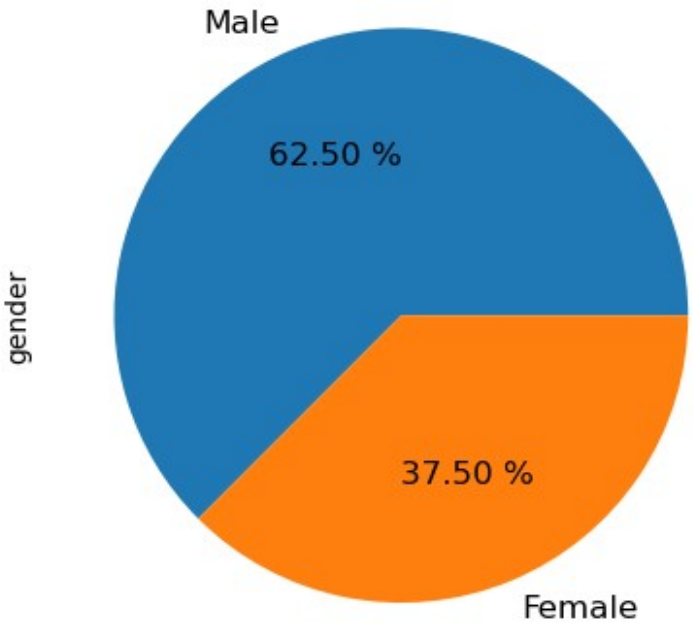
Task 2: Gender Distribution

```
df.gender.value_counts()
```

```
Male      25
Female    15
Name: gender, dtype: int64
```

```
df.gender.value_counts().plot.pie(fontsize = 12, autopct = '%.2f %%')
```

```
<AxesSubplot:ylabel='gender'>
```



Task 3: Descriptive Statistics

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
age	40.0	27.800	3.560467	21.0	25.75	27.0	30.00	35.0
Mutual_Funds	40.0	2.550	1.197219	1.0	2.00	2.0	3.00	7.0
Equity_Market	40.0	3.475	1.131994	1.0	3.00	4.0	4.00	6.0
Debentures	40.0	5.750	1.675617	1.0	5.00	6.5	7.00	7.0
Government_Bonds	40.0	4.650	1.369072	1.0	4.00	5.0	5.00	7.0
Fixed_Deposits	40.0	3.575	1.795828	1.0	2.75	3.5	5.00	7.0
PPF	40.0	2.025	1.609069	1.0	1.00	1.0	2.25	6.0
Gold	40.0	5.975	1.143263	2.0	6.00	6.0	7.00	7.0

```
#calculate descriptive statistics for categorical variables
df.describe(include='object').T
```

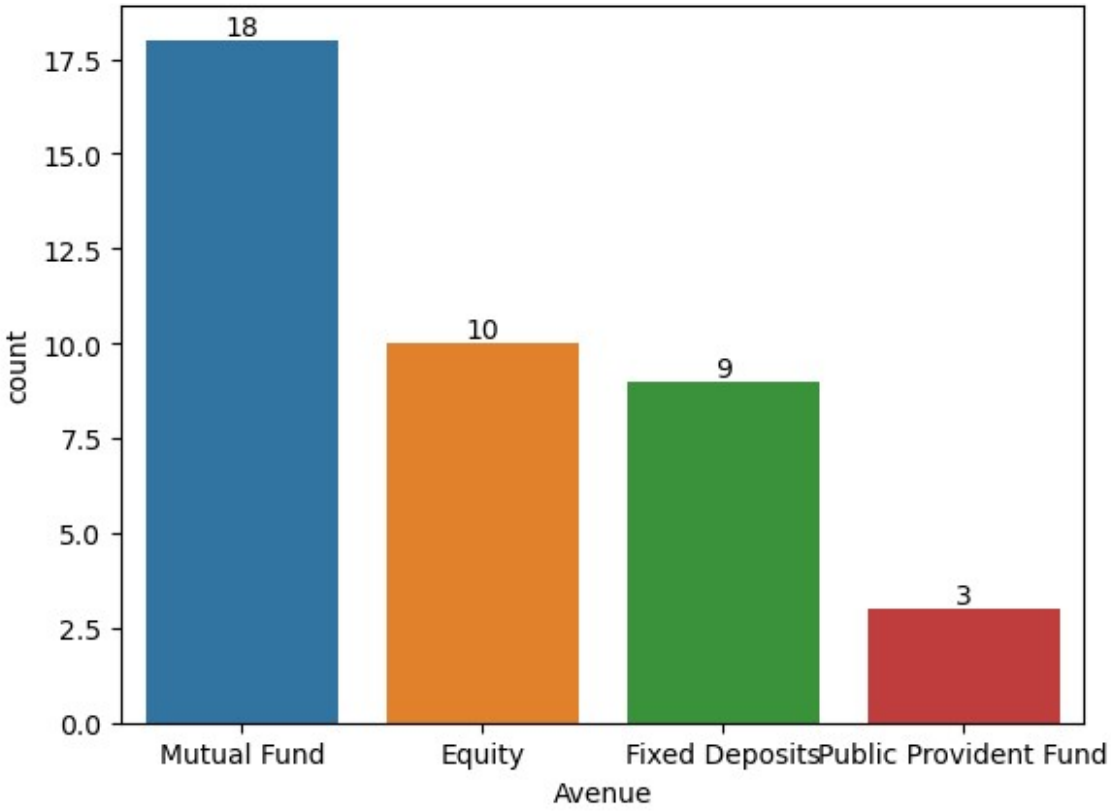
```
                                   count unique                    top
freq
gender                                40      2                   Male
25
Investment_Avenues                    40      2                    Yes
37
Stock_Marktet                         40      2                    Yes
35
Factor                                40      3                Returns
25
Objective                             40      3   Capital Appreciation
26
Purpose                               40      3        Wealth Creation
32
Duration                              40      4              3-5 years
19
Invest_Monitor                        40      3                Monthly
29
Expect                                40      3                20%-30%
32
Avenue                                40      4            Mutual Fund
18
What are your savings objectives?     40      3        Retirement Plan
24
Reason_Equity                         40      3   Capital Appreciation
30
Reason_Mutual                         40      3         Better Returns
24
Reason_Bonds                          40      3        Assured Returns
26
Reason_FD                             40      3              Risk Free
19
Source                                40      4  Financial Consultants
16
```

Task 4: Most Preferred Investment Avenue, Identify the most preferred investment avenue.

```
df.Avenue.value_counts()
```

```
Mutual Fund              18
Equity                   10
Fixed Deposits            9
Public Provident Fund     3
Name: Avenue, dtype: int64
```

```
ax = sns.countplot(x = 'Avenue',data = df)

for bars in ax.containers:
    ax.bar_label(bars)
```

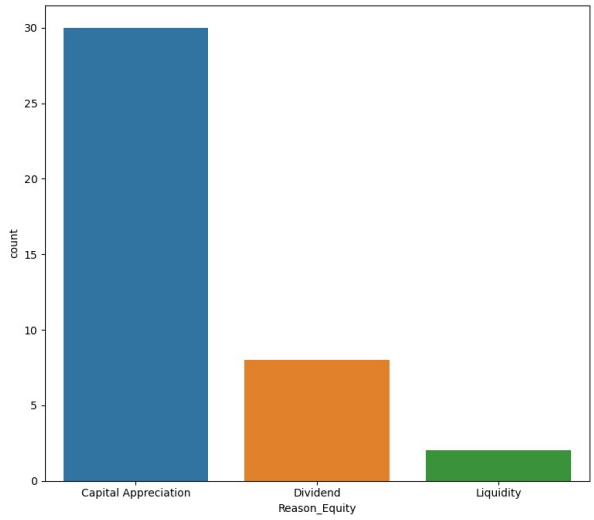


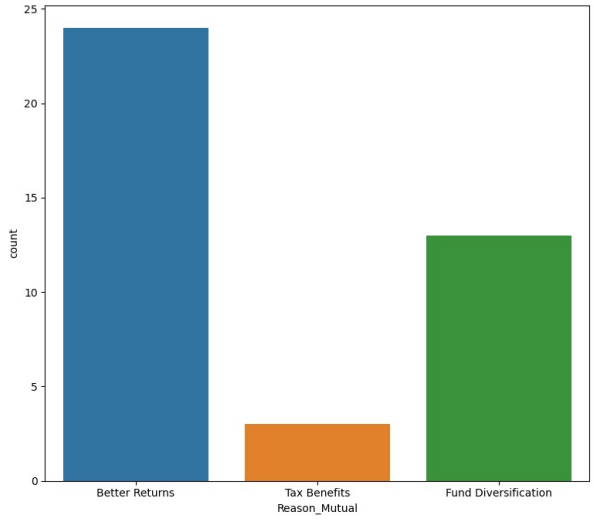
- Most Preferred Investment is Mutual Fund. Equity and Fixed Deposit Investment are similar in numbers. PPF is the least investment Avenue.

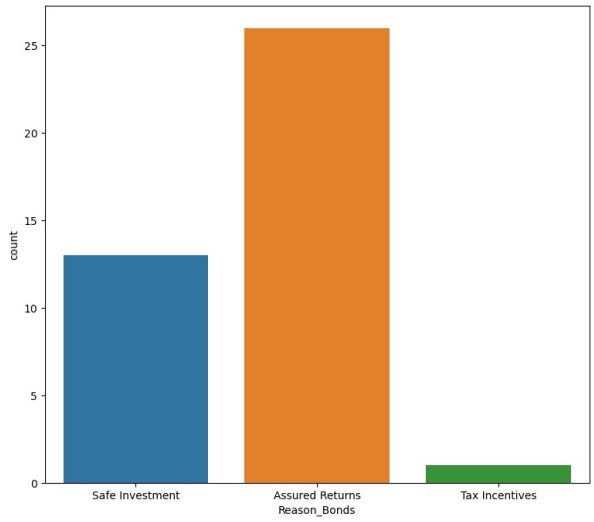
Task 5: Reasons for Investment, Analyze and summarize reasons for investment choices.

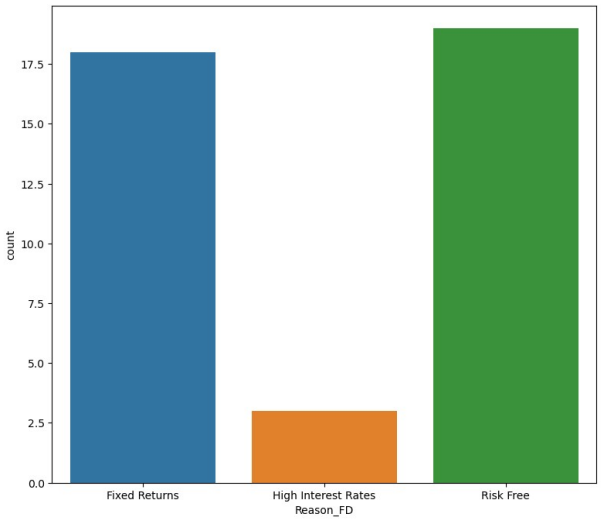
```
cat_cols =
['Reason_Equity','Reason_Mutual','Reason_Bonds','Reason_FD']
i=0
while i < 4:
    fig = plt.figure(figsize=[20,8])

    plt.subplot(1,2,1)
    sns.countplot(x=cat_cols[i], data=df)
    i += 1
```

```
    plt.subplot(1,2,2)
    sns.countplot(x=cat_cols[i], data=df)
    i += 1

    plt.show()
```









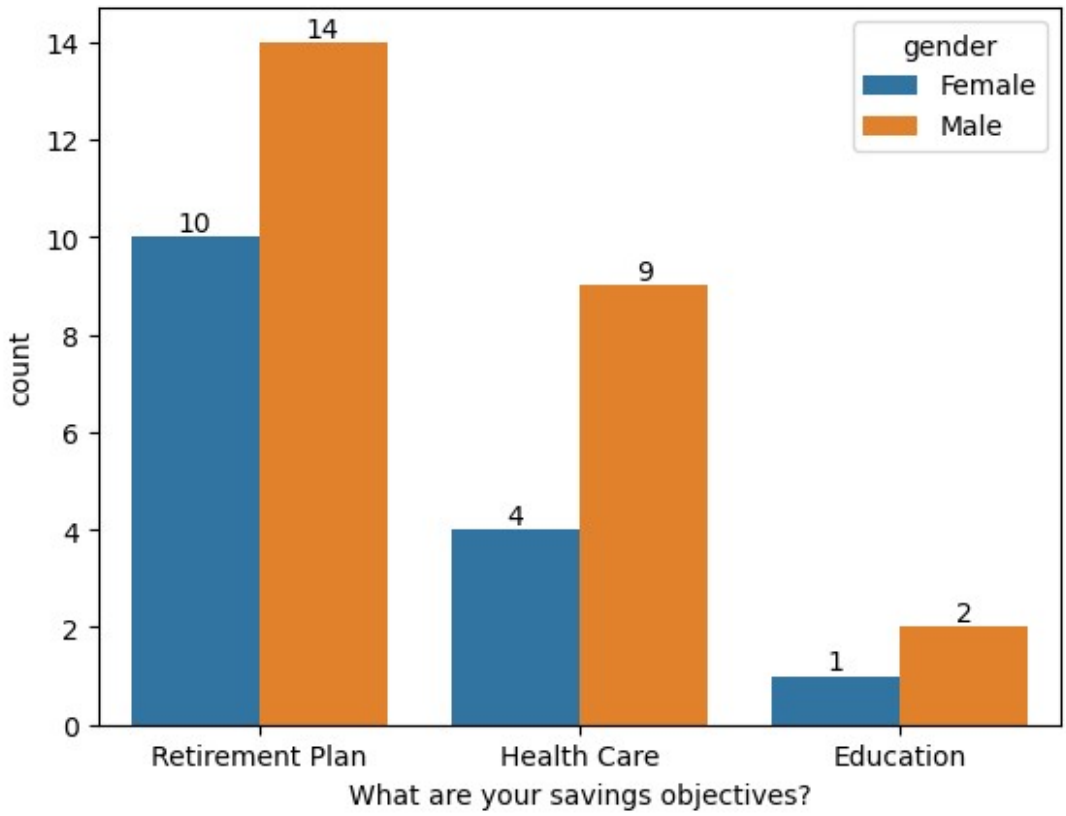
- Equity: Participants preferring Equity for Capital Appreciation than Dividend and Liquidity
- Mutual : The mejarity particinats think that Mutual Funds give better returns followed by Fund Diversification and Tax benefits
- Bonds: Assured Returns is the main reason that particiants have preferred bonds followed by Safe investment, Tax Incentives is the least reason provided to preferred Bonds.
- FD: Fixed Deposits have Fixed Returns and Its Risk Free Investment. These are the reasons provided to choose FD. It seems there are no high Interest Rates for FD this reason is least provided.

Task 6: Savings Objectives, Identify and present main savings objectives.

```
x=df['What are your savings objectives?']
x.value_counts()
```

```
Retirement Plan    24
Health Care        13
Education           3
Name: What are your savings objectives?, dtype: int64
```

```
ax = sns.countplot(x = 'What are your savings objectives?',data =
df,hue='gender')

for bars in ax.containers:
    ax.bar_label(bars)
```



- The main saving objective is for the Retirement followed by Healthcare. Least saving objective is for Education.
- The participants' min age is 21 and max age is 35. So it can be for their higher education in future.
- Males are more in numbers compared to Females

Task 7: Common Information Sources

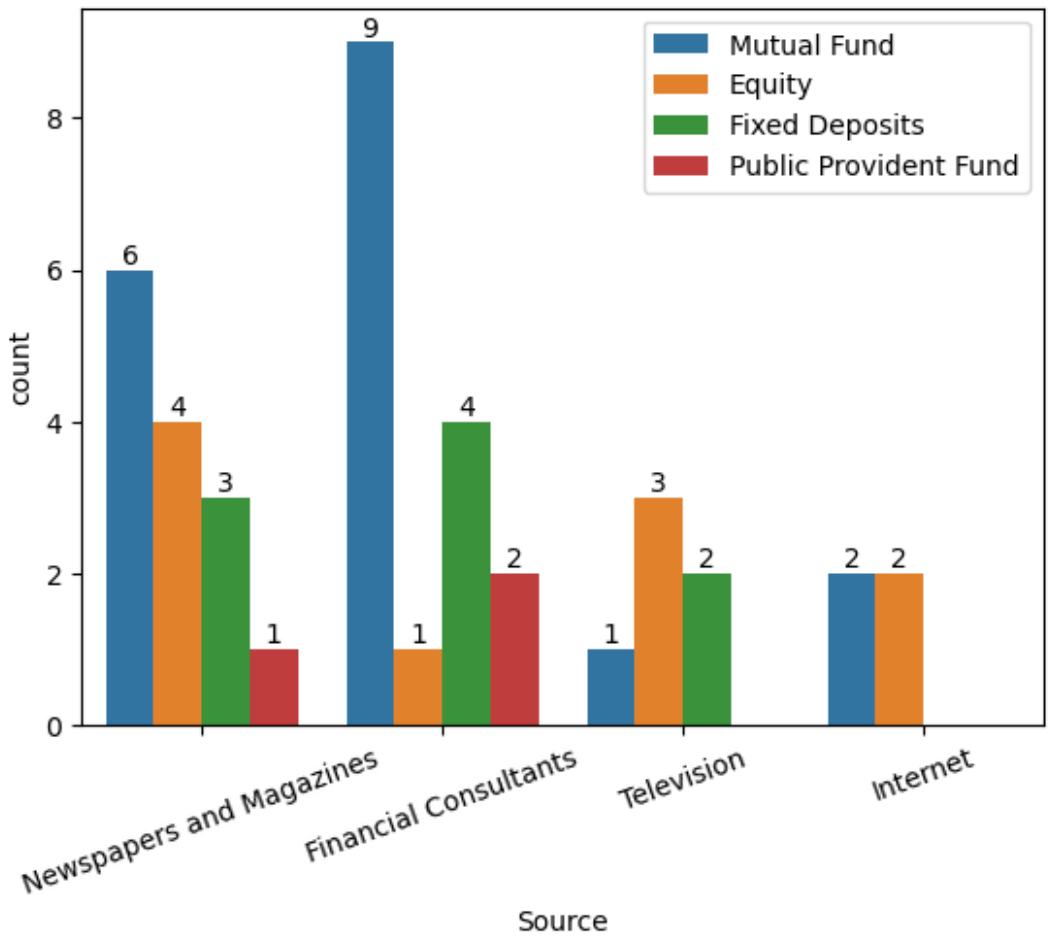
Analyze common sources participants rely on for investment information.

```
df.Source.value_counts()
```

```
Financial Consultants       16
Newspapers and Magazines    14
Television                   6
Internet                     4
Name: Source, dtype: int64
```

```
ax = sns.countplot(x = 'Source',data = df,hue='Avenue')

for bars in ax.containers:
    ax.bar_label(bars)

ax.set_xticklabels(ax.get_xticklabels(), rotation=20)
plt.legend(loc='upper right')
```

```
<matplotlib.legend.Legend at 0x1750c9a6250>
```



- Particinats most reply on Financial Consultants and Newspapers and Magazines than Television and Internet.
- In Financial Consultants category highest are For Mutual Funds ans Lowest is Equity.
- News Papers and Magazines, Mutual Funds is highest number and PPF is lowest.

Task 8: Investment Duration

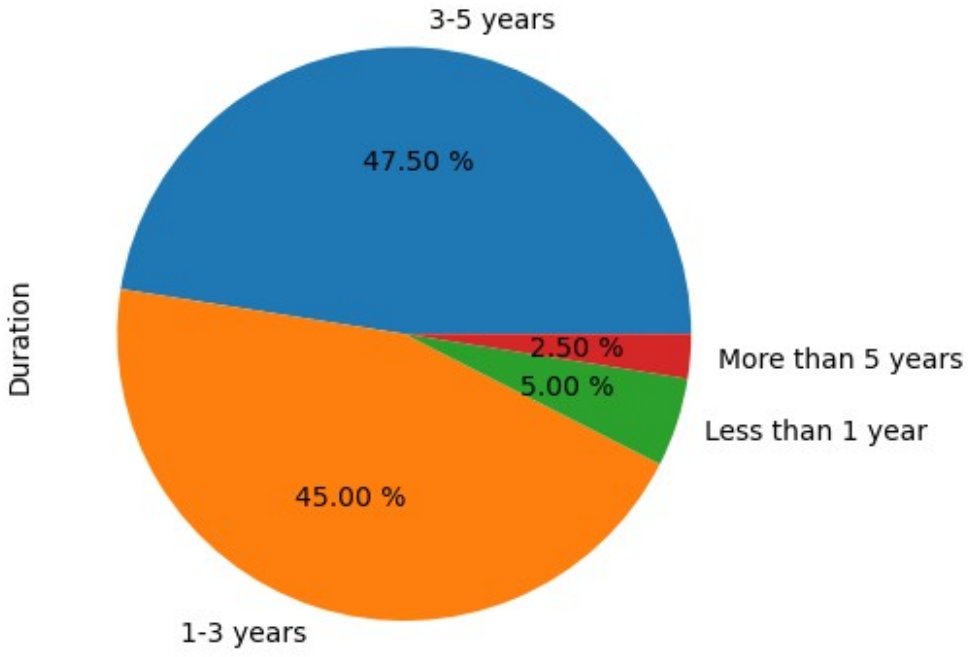
Calculate the average investment duration.Use appropriate statistical methods to calculatethe average investment duration.

```
df['Duration'].describe().T
```

```
count            40
unique            4
top       3-5 years
freq             19
Name: Duration, dtype: object
```

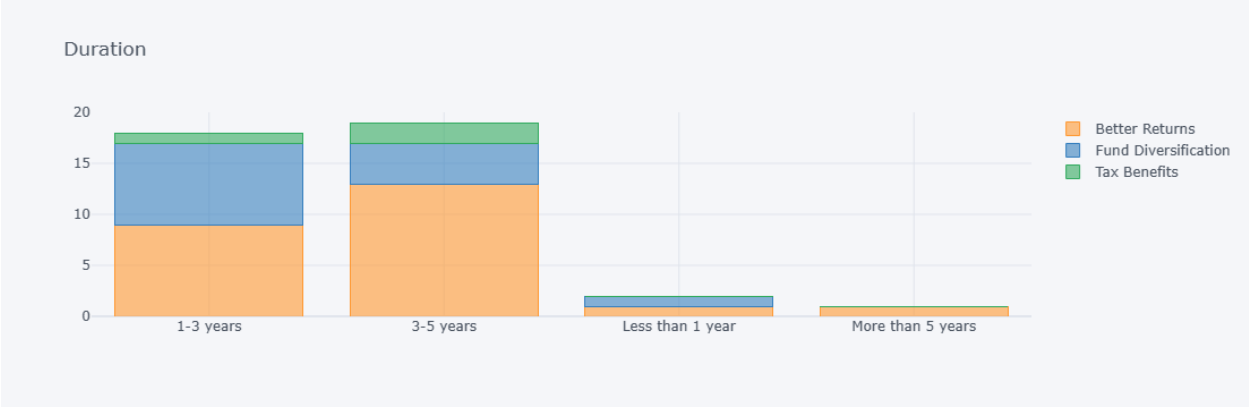
```
df.Duration.value_counts().plot.pie(fontsize = 10, autopct = '%.2f %
%')
```

```
<AxesSubplot:ylabel='Duration'>
```



- The Investment Duration having 4 unique values, among which 3-5 years is more common followe by 1-3 years.

```
mf=pd.crosstab(df['Duration'],df['Reason_Mutual'])
mf.iplot(kind="bar",barmode="stack",title='Duration')
```



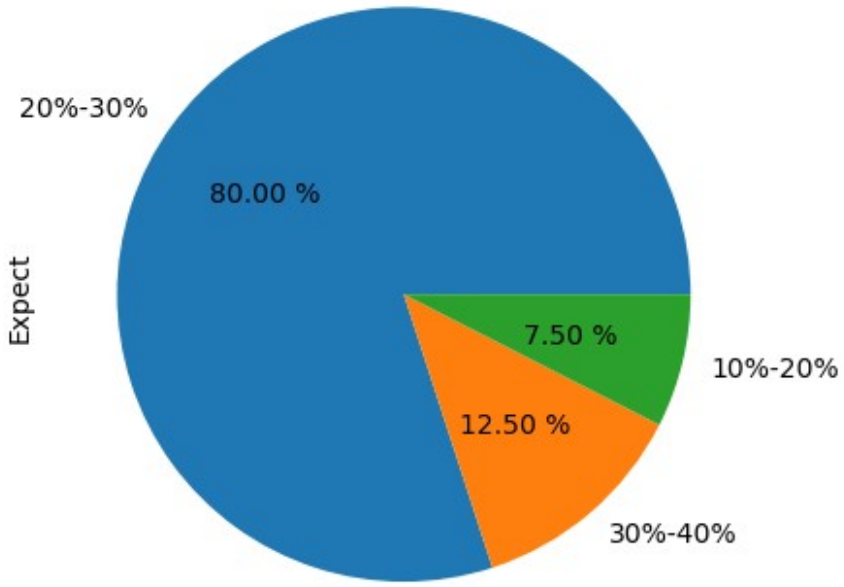
Task 9: Expectations from Investments, Summarize participants' expectations from investments.

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

```
count           40
unique           3
top       20%-30%
freq            32
Name: Expect, dtype: object
```

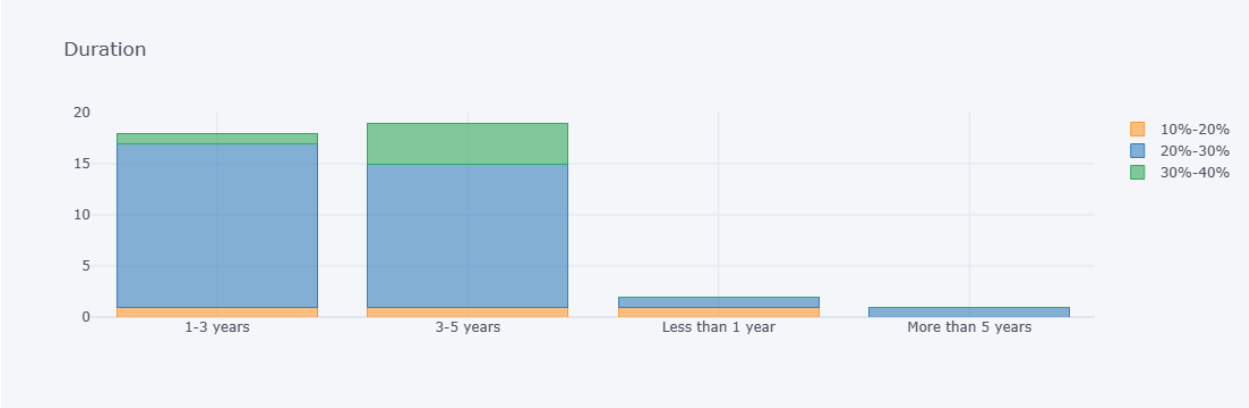
```
df['Expect'].value_counts().plot.pie(fontsize = 10, autopct = '%.2f %
%')
```

```
<AxesSubplot:ylabel='Expect'>
```



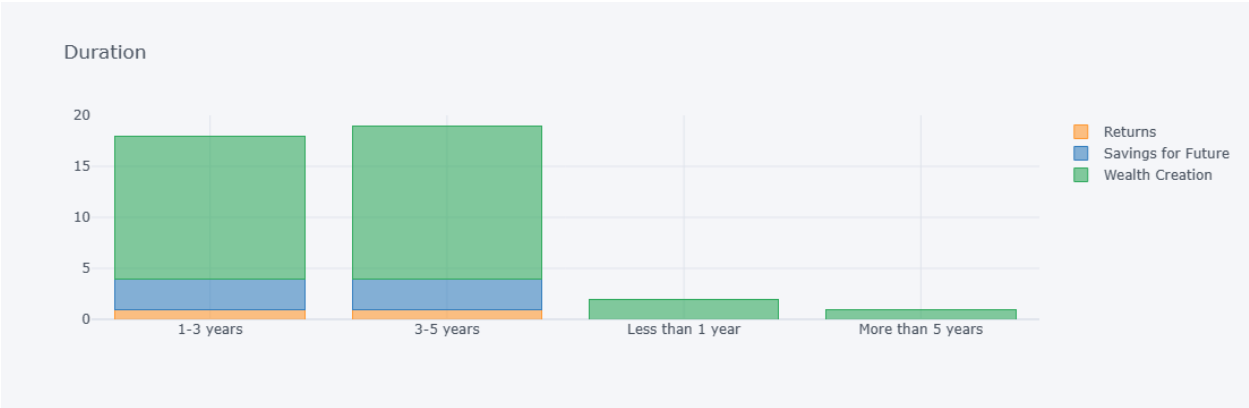
- 80% participants expect 20-30% return on investment

```
mf=pd.crosstab(df['Duration'],df['Expect'])
mf.iplot(kind="bar",barmode="stack",title='Duration')
```



- With 1-3 years duration, participants expect 20%-30% return on investment.
- with 3-5 years duration, participants expect 20%-30% return on investment followed by 30% to 40% investment.
- most of the participants expectations are 20%-30% return on investmrnt.

```
mf=pd.crosstab(df['Duration'],df['Purpose'])
mf.iplot(kind="bar",barmode="stack",title='Duration',)
```




```
pd.crosstab(df['Purpose'],df['Duration'])
```

Duration	1-3 years	3-5 years	Less than 1 year	More than 5 years
Purpose				
Returns	1	1	0	0
Savings for Future	3	3	0	0
Wealth Creation	14	15	2	1

- Around 37.5% participants are invested for duation 3-5 years reason being Wealth Creation.
- Around 35% participants have invested for dutaion 1-3 years for Wealth Creation.

- The most stated reason is Weath Creation among all the participants.