In [5]:

```
import pandas as pd
```

In [4]:

```
df=pd.read_csv("employees.csv")
```

In [5]:

```
df.head()
```

Out[5]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 8/6/1993 | 12:42 PM | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 3/31/1996 | 6:53 AM | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 4/23/1993 | 11:17 AM | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 3/4/2005 | 1:00 PM | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1/24/1998 | 4:47 PM | 101004 | 1.389 | True | Client Services |

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   First Name        933 non-null    object 
 1   Gender            855 non-null    object 
 2   Start Date        1000 non-null   object 
 3   Last Login Time   1000 non-null   object 
 4   Salary            1000 non-null   int64  
 5   Bonus %           1000 non-null   float64
 6   Senior Management  933 non-null    object 
 7   Team              957 non-null    object 
dtypes: float64(1), int64(1), object(6)
memory usage: 62.6+ KB
```

In [10]:

```
df["Start Date"]=pd.to_datetime(df["Start Date"])
```

In [11]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype         
---  ------             --------------  -----         
 0   First Name         933 non-null    object        
 1   Gender             855 non-null    object        
 2   Start Date         1000 non-null   datetime64[ns]
 3   Last Login Time    1000 non-null   object        
 4   Salary             1000 non-null   int64         
 5   Bonus %            1000 non-null   float64       
 6   Senior Management  933 non-null    object        
 7   Team               957 non-null    object        
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 62.6+ KB
```

In [12]:

```
df["Start Date"].head()
```

Out[12]:

```
0   1993-08-06
1   1996-03-31
2   1993-04-23
3   2005-03-04
4   1998-01-24
Name: Start Date, dtype: datetime64[ns]
```

In [14]:

```
df["Last Login Time"]=pd.to_datetime(df["Last Login Time"])
```

In [15]:

```
df["Last Login Time"]
```

Out[15]:

```
0     2021-07-27 12:42:00
1     2021-07-27 06:53:00
2     2021-07-27 11:17:00
3     2021-07-27 13:00:00
4     2021-07-27 16:47:00
              ...
995   2021-07-27 06:09:00
996   2021-07-27 06:30:00
997   2021-07-27 12:39:00
998   2021-07-27 16:45:00
999   2021-07-27 18:24:00
Name: Last Login Time, Length: 1000, dtype: datetime64[ns]
```

In [16]:

```
df.head()
```

Out[16]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [18]:

```
df["Senior Management"]=df["Senior Management"].astype("bool")
```

In [19]:

```
df.head()
```

Out[19]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [20]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype         
---  ------            --------------  -----         
 0   First Name        933 non-null    object        
 1   Gender            855 non-null    object        
 2   Start Date        1000 non-null   datetime64[ns]
 3   Last Login Time   1000 non-null   datetime64[ns]
 4   Salary            1000 non-null   int64         
 5   Bonus %           1000 non-null   float64       
 6   Senior Management  1000 non-null   bool          
 7   Team              957 non-null    object        
dtypes: bool(1), datetime64[ns](2), float64(1), int64(1), object(3)
memory usage: 55.8+ KB
```

In [21]:

```
df["Gender"]=df["Gender"].astype("category")
```

In [22]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype         
---  ------            --------------  -----         
 0   First Name        933 non-null    object        
 1   Gender            855 non-null    category      
 2   Start Date        1000 non-null   datetime64[ns]
 3   Last Login Time   1000 non-null   datetime64[ns]
 4   Salary            1000 non-null   int64         
 5   Bonus %           1000 non-null   float64       
 6   Senior Management  1000 non-null   bool          
 7   Team              957 non-null    object        
dtypes: bool(1), category(1), datetime64[ns](2), float64(1), int64(1), objec
t(2)
memory usage: 49.1+ KB
```

# Filter a DataFrame based on condition

In [25]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   First Name         933 non-null    object
 1   Gender             855 non-null    category
 2   Start Date         1000 non-null   datetime64[ns]
 3   Last Login Time    1000 non-null   datetime64[ns]
 4   Salary             1000 non-null   int64
 5   Bonus %            1000 non-null   float64
 6   Senior Management  1000 non-null   bool
 7   Team               957 non-null    object
dtypes: bool(1), category(1), datetime64[ns](2), float64(1), int64(1), objec
t(2)
memory usage: 49.1+ KB
```

In [26]:

```
df.head()
```

Out[26]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [32]:

```
(df["Gender"] == "Male").head()
```

Out[32]:

```
0     True
1     True
2    False
3     True
4     True
Name: Gender, dtype: bool
```

In [31]:

```
(df[df["Gender"] == "Male"]).head()#retrieve only male valued rows
```

Out[31]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |

In [ ]:

```
df["Team"]=="Finance"
```

In [ ]:

```
df["Team"]=="Finance"
```

In [34]:

```
(df[df["Gender"] == "Male"]).head()#retrieve only male valued rows
```

Out[34]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |

In [37]:

```
mask=(df["Team"]=="Finance")
df[mask].head()
```

Out[37]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **7** | NaN | Female | 2015-07-20 | 2021-07-27 10:43:00 | 45906 | 11.598 | True | Finance |
| **14** | Kimberly | Female | 1999-01-14 | 2021-07-27 07:13:00 | 41426 | 14.543 | True | Finance |
| **46** | Bruce | Male | 2009-11-28 | 2021-07-27 22:47:00 | 114796 | 6.796 | False | Finance |

In [39]:

```
sm=df["Senior Management"]
df[sm].head()
```

Out[39]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **6** | Ruby | Female | 1987-08-17 | 2021-07-27 16:20:00 | 65476 | 10.012 | True | Product |

In [41]:

```
mask=df["Team"]!="Marketing"
df[mask].head()
```

Out[41]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |

In [43]:

```
df[df["Salary"]>110000].head()
```

Out[43]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |
| **9** | Frances | Female | 2002-08-08 | 2021-07-27 06:51:00 | 139852 | 7.524 | True | Business Development |
| **12** | Brandon | Male | 1980-12-01 | 2021-07-27 01:08:00 | 112807 | 17.492 | True | Human Resources |

In [46]:

```
df[df["Start Date"]<= "1985-01-01"].head()
```

Out[46]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **10** | Louise | Female | 1980-08-12 | 2021-07-27 09:01:00 | 63241 | 15.132 | True | NaN |
| **12** | Brandon | Male | 1980-12-01 | 2021-07-27 01:08:00 | 112807 | 17.492 | True | Human Resources |
| **18** | Diana | Female | 1981-10-23 | 2021-07-27 10:27:00 | 132940 | 19.082 | False | Client Services |
| **28** | Terry | Male | 1981-11-27 | 2021-07-27 18:30:00 | 124008 | 13.464 | True | Client Services |
| **37** | Linda | Female | 1981-10-19 | 2021-07-27 20:49:00 | 57427 | 9.557 | True | Client Services |

# Filter based on mul conditions(and &,or | )

In [48]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.head()
```

Out[48]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [59]:

```
c1=df["Gender"]=="Male"
c2=df["Team"]=="Marketing"
(df[c1&c2].head()).count()

df[c1&c2].head()
```

Out[59]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **21** | Matthew | Male | 1995-09-05 | 2021-07-27 02:12:00 | 100612 | 13.645 | False | Marketing |
| **26** | Craig | Male | 2000-02-27 | 2021-07-27 07:45:00 | 37598 | 7.757 | True | Marketing |
| **74** | Thomas | Male | 1995-06-04 | 2021-07-27 14:24:00 | 62096 | 17.029 | False | Marketing |
| **77** | Charles | Male | 2004-09-14 | 2021-07-27 20:13:00 | 107391 | 1.260 | True | Marketing |

In [54]:

```
c1=df["Gender"]=="Male"
c2=df["Team"]=="Marketing"
df[c1|c2].head(10)
```

Out[54]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |
| **12** | Brandon | Male | 1980-12-01 | 2021-07-27 01:08:00 | 112807 | 17.492 | True | Human Resources |
| **13** | Gary | Male | 2008-01-27 | 2021-07-27 23:40:00 | 109831 | 5.831 | False | Sales |
| **16** | Jeremy | Male | 2010-09-21 | 2021-07-27 05:56:00 | 90370 | 7.369 | False | Human Resources |
| **17** | Shawn | Male | 1986-12-07 | 2021-07-27 19:45:00 | 111737 | 6.414 | False | Product |
| **21** | Matthew | Male | 1995-09-05 | 2021-07-27 02:12:00 | 100612 | 13.645 | False | Marketing |

# The .isin() mtd

In [63]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.tail()
```

Out[63]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **995** | Henry | NaN | 2014-11-23 | 2021-07-27 06:09:00 | 132483 | 16.655 | False | Distribution |
| **996** | Phillip | Male | 1984-01-31 | 2021-07-27 06:30:00 | 42392 | 19.675 | False | Finance |
| **997** | Russell | Male | 2013-05-20 | 2021-07-27 12:39:00 | 96914 | 1.421 | False | Product |
| **998** | Larry | Male | 2013-04-20 | 2021-07-27 16:45:00 | 60500 | 11.985 | False | Business Development |
| **999** | Albert | Male | 2012-05-15 | 2021-07-27 18:24:00 | 129949 | 10.169 | True | Sales |

In [61]:

```
df["Team"].unique()
```

Out[61]:

```
array(['Marketing', nan, 'Finance', 'Client Services', 'Legal', 'Product',
       'Engineering', 'Business Development', 'Human Resources', 'Sales',
       'Distribution'], dtype=object)
```

In [28]:

```
df["Team"].isin(["Marketing","Finance","Distribution"]).tail()#alternate for or(|)
```

Out[28]:

```
995     True
996     True
997    False
998    False
999    False
Name: Team, dtype: bool
```

# The .isnull() and .notnull() mtds

In [68]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.head()
```

Out[68]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [71]:

```
df["Team"].isnull().head(20)
```

Out[71]:

```
0     False
1      True
2     False
3     False
4     False
5     False
6     False
7     False
8     False
9     False
10     True
11    False
12    False
13    False
14    False
15    False
16    False
17    False
18    False
19    False
Name: Team, dtype: bool
```

In [73]:

```
df[df["Team"].isnull()].head()
```

Out[73]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **10** | Louise | Female | 1980-08-12 | 2021-07-27 09:01:00 | 63241 | 15.132 | True | NaN |
| **23** | NaN | Male | 2012-06-14 | 2021-07-27 16:19:00 | 125792 | 5.042 | True | NaN |
| **32** | NaN | Male | 1998-08-21 | 2021-07-27 14:27:00 | 122340 | 6.417 | True | NaN |
| **91** | James | NaN | 2005-01-26 | 2021-07-27 23:00:00 | 128771 | 8.309 | False | NaN |

In [77]:

```
df["Team"].notnull().head()
```

Out[77]:

```
0     True
1    False
2     True
3     True
4     True
Name: Team, dtype: bool
```

In [76]:

```
df[df["Team"].notnull()].head()
```

Out[76]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |

# The .between() mtd(Inclusive of both bounds)

In [78]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.head()
```

Out[78]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-27 16:47:00 | 101004 | 1.389 | True | Client Services |

In [81]:

```
bet=df["Salary"].between(60000,70000)
df[bet].head()
```

Out[81]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **6** | Ruby | Female | 1987-08-17 | 2021-07-27 16:20:00 | 65476 | 10.012 | True | Product |
| **10** | Louise | Female | 1980-08-12 | 2021-07-27 09:01:00 | 63241 | 15.132 | True | NaN |
| **20** | Lois | NaN | 1995-04-22 | 2021-07-27 19:18:00 | 64714 | 4.934 | True | Legal |
| **41** | Christine | NaN | 2015-06-28 | 2021-07-27 01:08:00 | 66582 | 11.308 | True | Business Development |

In [86]:

```
bet2=df["Bonus %"].between(4.17,15.132)
df[bet2].head(10)
```

Out[86]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-27 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-27 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-27 13:00:00 | 138705 | 9.340 | True | Finance |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |
| **6** | Ruby | Female | 1987-08-17 | 2021-07-27 16:20:00 | 65476 | 10.012 | True | Product |
| **7** | NaN | Female | 2015-07-20 | 2021-07-27 10:43:00 | 45906 | 11.598 | True | Finance |
| **9** | Frances | Female | 2002-08-08 | 2021-07-27 06:51:00 | 139852 | 7.524 | True | Business Development |
| **10** | Louise | Female | 1980-08-12 | 2021-07-27 09:01:00 | 63241 | 15.132 | True | NaN |
| **11** | Julie | Female | 1997-10-26 | 2021-07-27 15:19:00 | 102508 | 12.637 | True | Legal |

In [88]:

```
df[df["Start Date"].between("1991-01-01","1992-01-01")].head(10)
```

Out[88]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **27** | Scott | NaN | 1991-07-11 | 2021-07-27 18:58:00 | 122367 | 5.218 | False | Legal |
| **75** | Bonnie | Female | 1991-07-02 | 2021-07-27 01:27:00 | 104897 | 5.118 | True | Human Resources |
| **88** | Donna | Female | 1991-11-27 | 2021-07-27 13:59:00 | 64088 | 6.155 | True | Legal |
| **116** | NaN | Male | 1991-06-22 | 2021-07-27 20:58:00 | 76189 | 18.988 | True | Legal |
| **148** | Patrick | NaN | 1991-07-14 | 2021-07-27 02:24:00 | 124488 | 14.837 | True | Sales |
| **166** | NaN | Female | 1991-07-09 | 2021-07-27 18:52:00 | 42341 | 7.014 | True | Sales |
| **172** | Sara | Female | 1991-09-23 | 2021-07-27 18:17:00 | 97058 | 9.402 | False | Finance |
| **220** | NaN | Female | 1991-06-17 | 2021-07-27 12:49:00 | 71945 | 5.560 | True | Marketing |
| **245** | Victor | Male | 1991-04-11 | 2021-07-27 07:44:00 | 70817 | 17.138 | False | Engineering |
| **277** | Brenda | NaN | 1991-05-29 | 2021-07-27 06:32:00 | 82439 | 19.062 | False | Sales |

In [92]:

```
df[df["Last Login Time"].between("08:30AM","12:00PM")].head()#time entry format
```

Out[92]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **2** | Maria | Female | 1993-04-23 | 2021-07-27 11:17:00 | 130590 | 11.858 | False | Finance |
| **7** | NaN | Female | 2015-07-20 | 2021-07-27 10:43:00 | 45906 | 11.598 | True | Finance |
| **10** | Louise | Female | 1980-08-12 | 2021-07-27 09:01:00 | 63241 | 15.132 | True | NaN |
| **18** | Diana | Female | 1981-10-23 | 2021-07-27 10:27:00 | 132940 | 19.082 | False | Client Services |
| **33** | Jean | Female | 1993-12-18 | 2021-07-27 09:07:00 | 119082 | 16.180 | False | Business Development |

# The .duplicated() mtd on Series

In [93]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.sort_values("First Name",inplace=True)
df.head()
```

Out[93]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **101** | Aaron | Male | 2012-02-17 | 2021-07-27 10:20:00 | 61602 | 11.849 | True | Marketing |
| **327** | Aaron | Male | 1994-01-29 | 2021-07-27 18:48:00 | 58755 | 5.097 | True | Marketing |
| **440** | Aaron | Male | 1990-07-22 | 2021-07-27 14:53:00 | 52119 | 11.343 | True | Client Services |
| **937** | Aaron | NaN | 1986-01-22 | 2021-07-27 19:39:00 | 63126 | 18.424 | False | Client Services |
| **137** | Adam | Male | 2011-05-21 | 2021-07-27 01:45:00 | 95327 | 15.120 | False | Distribution |

In [105]:

```
(df["First Name"].duplicated(keep="first")).head(10)
```

Out[105]:

```
101    False
327     True
440     True
937     True
137    False
141     True
302     True
538     True
300    False
53      True
Name: First Name, dtype: bool
```

In [101]:

```
(df["First Name"].duplicated(keep="last")).head()#Last instance
```

Out[101]:

```
101     True
327     True
440     True
937    False
137     True
Name: First Name, dtype: bool
```

In [106]:

```
df.head(10)
```

Out[106]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **101** | Aaron | Male | 2012-02-17 | 2021-07-27 10:20:00 | 61602 | 11.849 | True | Marketing |
| **327** | Aaron | Male | 1994-01-29 | 2021-07-27 18:48:00 | 58755 | 5.097 | True | Marketing |
| **440** | Aaron | Male | 1990-07-22 | 2021-07-27 14:53:00 | 52119 | 11.343 | True | Client Services |
| **937** | Aaron | NaN | 1986-01-22 | 2021-07-27 19:39:00 | 63126 | 18.424 | False | Client Services |
| **137** | Adam | Male | 2011-05-21 | 2021-07-27 01:45:00 | 95327 | 15.120 | False | Distribution |
| **141** | Adam | Male | 1990-12-24 | 2021-07-27 20:57:00 | 110194 | 14.727 | True | Product |
| **302** | Adam | Male | 2007-07-05 | 2021-07-27 11:59:00 | 71276 | 5.027 | True | Human Resources |
| **538** | Adam | Male | 2010-10-08 | 2021-07-27 21:53:00 | 45181 | 3.491 | False | Human Resources |
| **300** | Alan | Male | 1988-06-26 | 2021-07-27 03:54:00 | 111786 | 3.592 | True | Engineering |
| **53** | Alan | NaN | 2014-03-03 | 2021-07-27 13:28:00 | 40341 | 17.578 | True | Finance |

```
(df[df["First Name"].duplicated(keep="first")]).head(10)
```

In [107]:

```
(df[df["First Name"].duplicated(keep=False)]).head(10)#keeps all duplicates values
```

Out[107]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **101** | Aaron | Male | 2012-02-17 | 2021-07-27 10:20:00 | 61602 | 11.849 | True | Marketing |
| **327** | Aaron | Male | 1994-01-29 | 2021-07-27 18:48:00 | 58755 | 5.097 | True | Marketing |
| **440** | Aaron | Male | 1990-07-22 | 2021-07-27 14:53:00 | 52119 | 11.343 | True | Client Services |
| **937** | Aaron | NaN | 1986-01-22 | 2021-07-27 19:39:00 | 63126 | 18.424 | False | Client Services |
| **137** | Adam | Male | 2011-05-21 | 2021-07-27 01:45:00 | 95327 | 15.120 | False | Distribution |
| **141** | Adam | Male | 1990-12-24 | 2021-07-27 20:57:00 | 110194 | 14.727 | True | Product |
| **302** | Adam | Male | 2007-07-05 | 2021-07-27 11:59:00 | 71276 | 5.027 | True | Human Resources |
| **538** | Adam | Male | 2010-10-08 | 2021-07-27 21:53:00 | 45181 | 3.491 | False | Human Resources |
| **300** | Alan | Male | 1988-06-26 | 2021-07-27 03:54:00 | 111786 | 3.592 | True | Engineering |
| **53** | Alan | NaN | 2014-03-03 | 2021-07-27 13:28:00 | 40341 | 17.578 | True | Finance |

In [110]:

```
mask=~df["First Name"].duplicated(keep=False)#selects only unique rows TILDA
```

In [111]:

```
df[mask]
```

Out[111]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **8** | Angela | Female | 2005-11-22 | 2021-07-27 06:29:00 | 95570 | 18.523 | True | Engineering |
| **688** | Brian | Male | 2007-04-07 | 2021-07-27 22:47:00 | 93901 | 17.821 | True | Legal |
| **190** | Carol | Female | 1996-03-19 | 2021-07-27 03:39:00 | 57783 | 9.129 | False | Finance |
| **887** | David | Male | 2009-12-05 | 2021-07-27 08:48:00 | 92242 | 15.407 | False | Legal |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-27 01:35:00 | 115163 | 10.125 | False | Legal |
| **495** | Eugene | Male | 1984-05-24 | 2021-07-27 10:54:00 | 81077 | 2.117 | False | Sales |
| **33** | Jean | Female | 1993-12-18 | 2021-07-27 09:07:00 | 119082 | 16.180 | False | Business Development |
| **832** | Keith | Male | 2003-02-12 | 2021-07-27 15:02:00 | 120672 | 19.467 | False | Legal |
| **291** | Tammy | Female | 1984-11-11 | 2021-07-27 10:30:00 | 132839 | 17.463 | True | Client Services |

In [116]:

```
df["First Name"].duplicated(keep=False).head()#makes all the repeated vals as duplicate
```

Out[116]:

```
101    True
327    True
440    True
937    True
137    True
Name: First Name, dtype: bool
```

# The .drop_duplicate()-sorted-on DF inplace

In [6]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")
df.sort_values("First Name",inplace=True)
df.head()
```

Out[6]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **101** | Aaron | Male | 2012-02-17 | 2021-07-28 10:20:00 | 61602 | 11.849 | True | Marketing |
| **327** | Aaron | Male | 1994-01-29 | 2021-07-28 18:48:00 | 58755 | 5.097 | True | Marketing |
| **440** | Aaron | Male | 1990-07-22 | 2021-07-28 14:53:00 | 52119 | 11.343 | True | Client Services |
| **937** | Aaron | NaN | 1986-01-22 | 2021-07-28 19:39:00 | 63126 | 18.424 | False | Client Services |
| **137** | Adam | Male | 2011-05-21 | 2021-07-28 01:45:00 | 95327 | 15.120 | False | Distribution |

In [118]:

```
len(df)
```

Out[118]:

```
1000
```

In [119]:

```
len(df.drop_duplicates())
```

Out[119]:

```
1000
```

In [ ]:

```
(df.drop_duplicates(subset=["First Name"],keep="last"))
```

In [7]:

```
(df.drop_duplicates(subset=["First Name"],keep=False)).head()
```

Out[7]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **8** | Angela | Female | 2005-11-22 | 2021-07-28 06:29:00 | 95570 | 18.523 | True | Engineering |
| **688** | Brian | Male | 2007-04-07 | 2021-07-28 22:47:00 | 93901 | 17.821 | True | Legal |
| **190** | Carol | Female | 1996-03-19 | 2021-07-28 03:39:00 | 57783 | 9.129 | False | Finance |
| **887** | David | Male | 2009-12-05 | 2021-07-28 08:48:00 | 92242 | 15.407 | False | Legal |
| **5** | Dennis | Male | 1987-04-18 | 2021-07-28 01:35:00 | 115163 | 10.125 | False | Legal |

In [11]:

```
df.drop_duplicates(subset=["Team"],keep=False)#no unique team value
```

Out[11]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|

In [15]:

```
(df.drop_duplicates(subset=["Team","First Name"])).head()
#delets only first name and team matches in both rows
```

Out[15]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **101** | Aaron | Male | 2012-02-17 | 2021-07-28 10:20:00 | 61602 | 11.849 | True | Marketing |
| **440** | Aaron | Male | 1990-07-22 | 2021-07-28 14:53:00 | 52119 | 11.343 | True | Client Services |
| **137** | Adam | Male | 2011-05-21 | 2021-07-28 01:45:00 | 95327 | 15.120 | False | Distribution |
| **141** | Adam | Male | 1990-12-24 | 2021-07-28 20:57:00 | 110194 | 14.727 | True | Product |
| **302** | Adam | Male | 2007-07-05 | 2021-07-28 11:59:00 | 71276 | 5.027 | True | Human Resources |

# The .unique() and .nunique() mtd

In [16]:

```
df=pd.read_csv("employees.csv",parse_dates=["Start Date","Last Login Time",])
df["Gender"]=df["Gender"].astype("category")
df["Senior Management"]=df["Senior Management"].astype("bool")

df.head()
```

Out[16]:

| | First Name | Gender | Start Date | Last Login Time | Salary | Bonus % | Senior Management | Team |
|---|---|---|---|---|---|---|---|---|
| **0** | Douglas | Male | 1993-08-06 | 2021-07-28 12:42:00 | 97308 | 6.945 | True | Marketing |
| **1** | Thomas | Male | 1996-03-31 | 2021-07-28 06:53:00 | 61933 | 4.170 | True | NaN |
| **2** | Maria | Female | 1993-04-23 | 2021-07-28 11:17:00 | 130590 | 11.858 | False | Finance |
| **3** | Jerry | Male | 2005-03-04 | 2021-07-28 13:00:00 | 138705 | 9.340 | True | Finance |
| **4** | Larry | Male | 1998-01-24 | 2021-07-28 16:47:00 | 101004 | 1.389 | True | Client Services |

In [17]:

```
df["Gender"].unique()
```

Out[17]:

```
['Male', 'Female', NaN]
Categories (2, object): ['Female', 'Male']
```

In [20]:

```
len(df["Gender"].unique())
```

Out[20]:

```
3
```

In [23]:

```
df["Gender"].nunique()
```

Out[23]:

```
2
```

In [21]:

```
df["Gender"].nunique(dropna=False)
```

Out[21]:

```
3
```

In [18]:

```
df["Team"].unique()
```

Out[18]:

```
array(['Marketing', nan, 'Finance', 'Client Services', 'Legal', 'Product',
       'Engineering', 'Business Development', 'Human Resources', 'Sales',
       'Distribution'], dtype=object)
```

In [24]:

```
len(df["Team"].unique())
```

Out[24]:

```
11
```

In [26]:

```
df["Team"].nunique()
```

Out[26]:

```
10
```

In [27]:

```
df["Team"].nunique(dropna=False)
```

Out[27]:

```
11
```

In [ ]:

```

```

In [ ]:

```

```