

```
In [1]: import math
import pandas as pd
```

▼ **Activity: Implement the *birthday_probability* function**

```
In [6]: def birthday_probability(number_of_people):
        ncr = math.factorial(number_of_people)/(math.factorial(2)*math.factorial(number_of_people-2))
        P = 1-(364/365)**ncr
        return P
```

```
In [8]: birthday_probability(23)
```

```
Out[8]: 0.5004771540365807
```

```
In [12]: birthday_probability(10)
```

```
Out[12]: 0.11614023654879224
```

```
In [13]: birthday_probability(15)
```

```
Out[13]: 0.25028790861398265
```

▼ **NBA Birthday Paradox Analysis**

```
In [45]: df = pd.read_csv('nba_2017.csv', parse_dates=['Birth Date'])
```

In [46]: `df.head()`

Out[46]:

	Player	Pos	Age	Team	Birth Date
0	Alex Abrines	SG	23.0	Oklahoma City Thunder	1993-08-01
1	Quincy Acy	PF	26.0	Dallas Mavericks	1990-10-06
2	Quincy Acy	PF	26.0	Brooklyn Nets	1990-10-06
3	Steven Adams	C	23.0	Oklahoma City Thunder	1993-07-20
4	Arron Afflalo	SG	31.0	Sacramento Kings	1985-10-15

In [47]: `df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 551 entries, 0 to 550
Data columns (total 5 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Player          551 non-null   object
1   Pos             551 non-null   object
2   Age             551 non-null   float64
3   Team            551 non-null   object
4   Birth Date      551 non-null   datetime64[ns]
dtypes: datetime64[ns](1), float64(1), object(3)
memory usage: 21.6+ KB
```

▼ **Activity: Create the Birth Date column**

In [48]: `df['Birth Date'].dt.strftime("%Y-%m-%d").head()`

Out[48]:

```
0    1993-08-01
1    1990-10-06
2    1990-10-06
3    1993-07-20
4    1985-10-15
Name: Birth Date, dtype: object
```

```
In [49]: df["Birthday"] = df['Birth Date'].dt.strftime("%m-%d")
df['Birthday'].head()
```

```
Out[49]: 0    08-01
1    10-06
2    10-06
3    07-20
4    10-15
Name: Birthday, dtype: object
```

```
In [50]: df.head()
```

```
Out[50]:
```

	Player	Pos	Age	Team	Birth Date	Birthday
0	Alex Abrines	SG	23.0	Oklahoma City Thunder	1993-08-01	08-01
1	Quincy Acy	PF	26.0	Dallas Mavericks	1990-10-06	10-06
2	Quincy Acy	PF	26.0	Brooklyn Nets	1990-10-06	10-06
3	Steven Adams	C	23.0	Oklahoma City Thunder	1993-07-20	07-20
4	Arron Afflalo	SG	31.0	Sacramento Kings	1985-10-15	10-15

▼ Interlude: Combinatorics

For this project, you're free to use any technique that you prefer to answer how many players share a birthday for a given team. But, one recommendation would be to use combinatorics; specifically the *Combinations*, using the `itertools.combinations` function. Here's a quick example. Suppose we have these samples:

Name	Birthday
John	March 5th
Mary	Sept 20th
Rob	March 5th

Using combinations, we can take all the samples in pairs ($r=2$) to compare them:

Person 1	Person 2
John	Mary
John	Rob
Mary	Rob

```
In [27]: from itertools import combinations
```

```
In [28]: names = ["John", "Mary", "Rob"]  
birthdays = ["March 5th", "Sept 20th", "March 5th"]
```

```
In [29]: # Note: we need to wrap it in a list to force display  
list(combinations(names, 2))
```

```
Out[29]: [('John', 'Mary'), ('John', 'Rob'), ('Mary', 'Rob')]
```

```
In [30]: list(combinations(names,3))
```

```
Out[30]: [('John', 'Mary', 'Rob')]
```

```
In [31]: # Note: we need to wrap it in a list to force display  
list(combinations(birthdays, 2))
```

```
Out[31]: [('March 5th', 'Sept 20th'),  
          ('March 5th', 'March 5th'),  
          ('Sept 20th', 'March 5th')]
```

We can see how March 5th (John and Rob) are the same dates. Using Pandas:

```
In [32]: names_df = pd.DataFrame(combinations(names, 2), columns=["Person 1", "Person 2"])
names_df
```

```
Out[32]:
```

	Person 1	Person 2
0	John	Mary
1	John	Rob
2	Mary	Rob

```
In [33]: birthdays_df = pd.DataFrame(combinations(birthdays, 2), columns=["Birthday 1", "Birthday 2"])
birthdays_df
```

```
Out[33]:
```

	Birthday 1	Birthday 2
0	March 5th	Sept 20th
1	March 5th	March 5th
2	Sept 20th	March 5th

Combining it:

```
In [34]: df = pd.concat([names_df, birthdays_df], axis=1)
```

```
In [35]: df
```

```
Out[35]:
```

	Person 1	Person 2	Birthday 1	Birthday 2
0	John	Mary	March 5th	Sept 20th
1	John	Rob	March 5th	March 5th
2	Mary	Rob	Sept 20th	March 5th

```
In [36]: df['Birthday 1'] == df['Birthday 2']
```

```
Out[36]: 0    False  
         1     True  
         2    False  
         dtype: bool
```

End of the interlude! Now, it's your turn to answer questions.

▼ Activities

- ▼ *How many pairs of players share a birthday for the Atlanta Hawks?*

```
In [51]: df1 = df.loc[df['Team']=='Atlanta Hawks']  
df1
```

```
Out[51]:
```

	Player	Pos	Age	Team	Birth Date	Birthday
37	Kent Bazemore	SF	27.0	Atlanta Hawks	1989-07-01	07-01
42	DeAndre' Bembry	SF	22.0	Atlanta Hawks	1994-07-04	07-04
75	Jose Calderon	PG	35.0	Atlanta Hawks	1981-09-28	09-28
116	Malcolm Delaney	PG	27.0	Atlanta Hawks	1989-03-11	03-11
130	Mike Dunleavy	SF	36.0	Atlanta Hawks	1954-03-21	03-21
131	Mike Dunleavy	SF	36.0	Atlanta Hawks	1980-09-15	09-15
192	Tim Hardaway	SG	24.0	Atlanta Hawks	1966-09-01	09-01
193	Tim Hardaway	SG	24.0	Atlanta Hawks	1992-03-16	03-16
231	Dwight Howard	C	31.0	Atlanta Hawks	1985-12-08	12-08
234	Kris Humphries	PF	31.0	Atlanta Hawks	1985-02-06	02-06
241	Ersan Ilyasova	PF	29.0	Atlanta Hawks	1987-05-15	05-15
275	Ryan Kelly	PF	25.0	Atlanta Hawks	1991-04-09	04-09
279	Kyle Korver	SG	35.0	Atlanta Hawks	1981-03-17	03-17
344	Paul Millsap	PF	31.0	Atlanta Hawks	1985-02-10	02-10
358	Mike Muscala	C	25.0	Atlanta Hawks	1991-07-01	07-01
363	Gary Neal	SG	32.0	Atlanta Hawks	1984-10-03	10-03
393	Lamar Patterson	SG	25.0	Atlanta Hawks	1991-08-12	08-12
445	Dennis Schroder	PG	23.0	Atlanta Hawks	1993-09-15	09-15
447	Mike Scott	PF	28.0	Atlanta Hawks	1988-07-16	07-16
448	Thabo Sefolosha	SF	32.0	Atlanta Hawks	1984-05-02	05-02
472	Edy Tavares	C	24.0	Atlanta Hawks	1992-03-22	03-22
511	Taurean Waller-Prince	SF	22.0	Atlanta Hawks	1994-03-22	03-22

```
In [58]: names = list(df1['Player'])  
birthdays = list(df1['Birthday'])
```

```
In [59]: name_df = pd.DataFrame(combinations(names,2),columns=['Person1', 'Person2'])  
name_df.head()
```

```
Out[59]:
```

	Person1	Person2
0	Kent Bazemore	DeAndre' Bemby
1	Kent Bazemore	Jose Calderon
2	Kent Bazemore	Malcolm Delaney
3	Kent Bazemore	Mike Dunleavy
4	Kent Bazemore	Mike Dunleavy

```
In [60]: birthday_df = pd.DataFrame(combinations(birthdays,2),columns=['Birthday1', 'Birthday2'])  
birthday_df.head()
```

```
Out[60]:
```

	Birthday1	Birthday2
0	07-01	07-04
1	07-01	09-28
2	07-01	03-11
3	07-01	03-21
4	07-01	09-15


```
In [61]: tmp_df = pd.concat([name_df,birthday_df],axis=1)
tmp_df
```

```
Out[61]:
```

	Person1	Person2	Birthday1	Birthday2
0	Kent Bazemore	DeAndre' Bembry	07-01	07-04
1	Kent Bazemore	Jose Calderon	07-01	09-28
2	Kent Bazemore	Malcolm Delaney	07-01	03-11
3	Kent Bazemore	Mike Dunleavy	07-01	03-21
4	Kent Bazemore	Mike Dunleavy	07-01	09-15
...
226	Mike Scott	Edy Tavares	07-16	03-22
227	Mike Scott	Taurean Waller-Prince	07-16	03-22
228	Thabo Sefolosha	Edy Tavares	05-02	03-22
229	Thabo Sefolosha	Taurean Waller-Prince	05-02	03-22
230	Edy Tavares	Taurean Waller-Prince	03-22	03-22

231 rows × 4 columns

```
In [62]: tmp_df.loc[tmp_df['Birthday1']==tmp_df['Birthday2']]
```

```
Out[62]:
```

	Person1	Person2	Birthday1	Birthday2
13	Kent Bazemore	Mike Muscala	07-01	07-01
106	Mike Dunleavy	Dennis Schroder	09-15	09-15
230	Edy Tavares	Taurean Waller-Prince	03-22	03-22

- ▼ ***How many pairs of players share a birthday in the Cleveland Cavaliers?***

```
In [65]: df2 = df.loc[df['Team']=='Cleveland Cavaliers']

names = list(df2['Player'])
birthdays = list(df2['Birthday'])

name_df = pd.DataFrame(combinations(names,2),columns=['Person1','Person2'])
birthday_df = pd.DataFrame(combinations(birthdays,2),columns=['Birthday1','Birthday2'])

tmp_df = pd.concat([name_df,birthday_df],axis=1)
tmp_df.loc[tmp_df['Birthday1']==tmp_df['Birthday2']]
```

```
Out[65]:
```

	Person1	Person2	Birthday1	Birthday2
219	Iman Shumpert	Deron Williams	06-26	06-26

▼ *In the Dallas Mavericks, who shares a birthday with J.J. Barea?*

```
In [68]: df3 = df.loc[df['Team']=='Dallas Mavericks']

names = list(df3['Player'])
birthdays = list(df3['Birthday'])

name_df = pd.DataFrame(combinations(names,2),columns=['Person1','Person2'])
birthday_df = pd.DataFrame(combinations(birthdays,2),columns=['Birthday1','Birthday2'])

tmp_df = pd.concat([name_df,birthday_df],axis=1)
tmp_df.loc[tmp_df['Birthday1']==tmp_df['Birthday2']]
```

```
Out[68]:
```

	Person1	Person2	Birthday1	Birthday2
65	J.J. Barea	Deron Williams	06-26	06-26

```
In [ ]:
```

