Data Preprocessing

We will start by importing the pandas library and reading our free throw CSV to create a DataFrame.

[45]:
```
import pandas as pd
pd.options.mode.chained_assignment = None

#Create a DataFrame by reading CSV
df = pd.read_csv(r'free_throws.csv')

df.head(5)
```

[45]:
```
  end_result       game      game_id  period  \
0  106 - 114  PHX - LAL  261031013.0     1.0
1  106 - 114  PHX - LAL  261031013.0     1.0
2  106 - 114  PHX - LAL  261031013.0     1.0
3  106 - 114  PHX - LAL  261031013.0     1.0
4  106 - 114  PHX - LAL  261031013.0     1.0

                                  play        player playoffs    score  \
0   Andrew Bynum makes free throw 1 of 2  Andrew Bynum  regular    0 - 1
1   Andrew Bynum makes free throw 2 of 2  Andrew Bynum  regular    0 - 2
2   Andrew Bynum makes free throw 1 of 2  Andrew Bynum  regular  18 - 12
3  Andrew Bynum misses free throw 2 of 2  Andrew Bynum  regular  18 - 12
4   Shawn Marion makes free throw 1 of 1  Shawn Marion  regular  21 - 12

        season  shot_made   time
0  2006 - 2007          1  11:45
1  2006 - 2007          1  11:45
2  2006 - 2007          1   7:26
3  2006 - 2007          0   7:26
4  2006 - 2007          1   7:18
```

[46]:
```
#Remove unneeded columns
df.drop(['end_result', 'game', 'game_id', 'period', 'playoffs', 'score',
         'season', 'time'], axis='columns', inplace=True)
```

As a result of an error by the NBA or the Kaggle member that scraped the data, there are instances of free throw 1 of 2 appearing twice or not having free throw 2 of 2. So, we have to clean up these rows.

[47]:
```
#Remove any free throws that are not free throw 1 of 2 followed by free throw
# 2 of 2 by the same player
#Add a T/F 'check' column
df['check'] = ((df['play'].str.contains("2 of 2") &
                df.play.shift().str.contains("1 of 2") &
                df.player.eq(df.player.shift())) |
```

```
                (df['play'].str.contains("1 of 2") &
                 df.play.shift(-1).str.contains("2 of 2") &
                 df.player.eq(df.player.shift(-1))))

#Only keep entries where 'check' equals True
df = df[df['check'] == True]

#Remove the 'check' column
df.drop('check', axis='columns', inplace=True)
```

Now we can begin configuring the data to prepare for analysis.

[48]:
```
#Create separate DataFrames for first and second shots
df_first = df[df['play'].str.contains('1 of 2')]
df_second = df[df['play'].str.contains('2 of 2')]

#Reset the DataFrame indices
df_first = df_first.reset_index()
df_first.drop('index', axis='columns', inplace=True)
df_second = df_second.reset_index()
df_second.drop('index', axis='columns', inplace=True)

df_first.head(5)
```

[48]:

	play	player	shot_made
0	Andrew Bynum makes free throw 1 of 2	Andrew Bynum	1
1	Andrew Bynum makes free throw 1 of 2	Andrew Bynum	1
2	Amare Stoudemire makes free throw 1 of 2	Amare Stoudemire	1
3	Leandro Barbosa misses free throw 1 of 2	Leandro Barbosa	0
4	Lamar Odom makes free throw 1 of 2	Lamar Odom	1

[49]:
```
#Remove the play column
df_first = df_first.drop('play', axis='columns')
df_second = df_second.drop('play', axis='columns')

#Rename the shot_made columns to shot1 and shot2
df_first = df_first.rename(columns={'shot_made': 'shot1'})
df_second = df_second.rename(columns={'shot_made': 'shot2'})

#Combine the first and second shot DataFrames to create a dataframe of shot pairs
df_pairs = pd.concat([df_first, df_second], axis=1)
df_pairs = df_pairs.loc[:,~df_pairs.columns.duplicated()]

df_pairs.head(5)
```

[49]:

	player	shot1	shot2
0	Andrew Bynum	1	1
1	Andrew Bynum	1	0

```
2  Amare Stoudemire      1      1
3   Leandro Barbosa      0      1
4         Lamar Odom      1      1
```

We will save this DataFrame as a CSV for part of our later analysis.

[50]:
```
df_pairs.to_csv('free_throw_pairs.csv', index=False)
```

Now we will continue formatting the data so that we can get an aggregate count of the different free throw results for all players.

[51]:
```
#Create separate DataFrames for shot pairs where the first shot was missed
# and first shot was made
df_missed_first = df_pairs[df_pairs['shot1']==0]
df_made_first = df_pairs[df_pairs['shot1']==1]

df_missed_first.head(5)
```

[51]:
```
                player  shot1  shot2
3      Leandro Barbosa      0      1
5         Smush Parker      0      0
6  Vladimir Radmanovic      0      0
7        Maurice Evans      0      1
8      Leandro Barbosa      0      1
```

[52]:
```
df_made_first.head(5)
```

[52]:
```
             player  shot1  shot2
0      Andrew Bynum      1      1
1      Andrew Bynum      1      0
2  Amare Stoudemire      1      1
4        Lamar Odom      1      1
9      Shawn Marion      1      1
```

[53]:
```
#Add a column that has the count of free throw attempts where the first
# was missed/made
df_missed_first['missed_first']=1
df_made_first['made_first']=1

#Group the individual free throw attempts by player name and sum the columns
df_missed_first = df_missed_first.groupby('player', as_index=False).sum()
df_made_first = df_made_first.groupby('player', as_index=False).sum()

#Rename the shot2 columns to missed_1st_made_2nd and made_1st_made_2nd
df_missed_first = df_missed_first.rename(columns=
                                         {'shot2': 'missed_1st_made_2nd'})
df_made_first = df_made_first.rename(columns=
                                     {'shot2': 'made_1st_made_2nd'})
```

```
df_missed_first.head(5)
```

[53]:

	player	shot1	missed_1st_made_2nd	missed_first
0	A.J. Price	0	21	28
1	Aaron Brooks	0	69	83
2	Aaron Gordon	0	27	38
3	Aaron Gray	0	32	65
4	Aaron Harrison	0	2	3

[54]:
```
df_made_first.head(5)
```

[54]:

	player	shot1	made_1st_made_2nd	made_first
0	A.J. Price	77	62	77
1	Aaron Brooks	345	295	345
2	Aaron Gordon	72	53	72
3	Aaron Gray	87	51	87
4	Aaron Harrison	3	0	3

[55]:
```
#Remove the shot1 column
df_missed_first.drop('shot1', axis='columns', inplace=True)
df_made_first.drop('shot1', axis='columns', inplace=True)

#Merge the missed first and made first DataFrames by player
df_counts = pd.merge(df_missed_first, df_made_first, on="player")

df_counts.head(5)
```

[55]:

	player	missed_1st_made_2nd	missed_first	made_1st_made_2nd	\
0	A.J. Price	21	28	62	
1	Aaron Brooks	69	83	295	
2	Aaron Gordon	27	38	53	
3	Aaron Gray	32	65	51	
4	Aaron Harrison	2	3	0	

	made_first
0	77
1	345
2	72
3	87
4	3

We will save this DataFrame as a CSV for our analysis.

[56]:
```
df_counts.to_csv('free_throw_counts.csv', index=False)
```