Libraries, directory and data

In [ ]:

```
#change directory
%cd /content/drive/MyDrive/Business Analyst course/Statistics and Descriptive Analytics/Basic Statistics
```

```
/content/drive/MyDrive/Business Analyst course/Statistics and Descriptive Analytics/Basic Statistics
```

In [ ]:

```
#Libraries
import pandas as pd
import seaborn as sns
```

In [ ]:

```
#load the dataset
df = pd.read_csv("Baseball.csv")
df.head()
```

Out[ ]:

	Team	League	Year	RS	RA	W	OBP	SLG	BA	Playoffs	RankSeason	RankPlayoffs	G	OOBP	OSLG
0	ARI	NL	2012	734	688	81	0.328	0.418	0.259	0	NaN	NaN	162	0.317	0.415
1	ATL	NL	2012	700	600	94	0.320	0.389	0.247	1	4.0	5.0	162	0.306	0.378
2	BAL	AL	2012	712	705	93	0.311	0.417	0.247	1	5.0	4.0	162	0.315	0.403
3	BOS	AL	2012	734	806	69	0.315	0.415	0.260	0	NaN	NaN	162	0.331	0.428
4	CHC	NL	2012	613	759	61	0.302	0.378	0.240	0	NaN	NaN	162	0.335	0.424

Mean

In [ ]:

```
#Mean of Runs Scored (RS)
df.RS.mean()
```

Out[ ]:

```
715.0819805194806
```

In [ ]:

```
#Mean of Runs Scored (RS) by the Arizona team (ARI)
df.loc[df.Team == "ARI"].RS.mean()
```

Out[ ]:

```
742.2
```

In [ ]:

```
#Mean of Runs Scored (RS) by the Arizona team (ARI) since 2005
df.loc[(df.Team == "ARI") & (df.Year > 2005)].RS.mean()
```

Out[ ]:

```
729.0
```

In [ ]:

```
#Question: What is the mean of Runs Allowed (RA) by the Chicago team (CHC)
# until 2007
df.loc[(df.Team == "CHC") & (df.Year < 2007)].RA.mean()
```

Out[ ]:

```
728.170731707317
```

Median

In [ ]:

```
#Median and Mean of Wins (W)
print(df.W.mean())
df.W.median()
```

```
80.90422077922078
```

Out[ ]:

```
81.0
```

In [ ]:

```
#Question: what is the median of Wins of the Baltimore Team (BAL) until 2000
df.loc[(df.Team == "BAL") & (df.Year <= 2000)].W.median()
```

Out[ ]:

```
89.0
```

Mode

In [ ]:

```
# Mode, Median and Mean of OBP
print(df.OBP.mean())
print(df.OBP.median())
df.OBP.mode()
```

```
0.32633116883116886
0.326
```

Out[ ]:

```
0    0.322
dtype: float64
```

In [ ]:

```
#Question: Mode of OBP during the Year 2010
df.loc[df.Year == 2010].OBP.mode()
```

Out[ ]:

```
0    0.332
dtype: float64
```

Correlation

In [ ]:

```
#pick variables
df_correlation = df[["RS", "RA", "W"]]
df_correlation.head(1)
```

Out[ ]:

	RS	RA	W
0	734	688	81

In [ ]:

```
#Correlation matrix
df_correlation.corr()
```

Out[ ]:

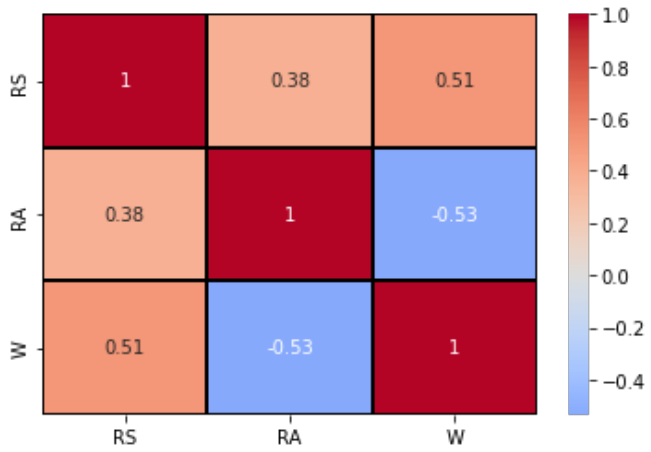
	RS	RA	W
RS	1.000000	0.380139	0.511745
RA	0.380139	1.000000	-0.532394
W	0.511745	-0.532394	1.000000

In [ ]:

```
#Correlation heatmap
sns.heatmap(data = df_correlation.corr(),
            annot = True,
            fmt = '.2g',
            center = 0,
            cmap = 'coolwarm',
            linewidths = 1,
            linecolor = 'black')
```

Out[ ]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb5c0586790>
```



In [ ]:

```
#challenge: correlation matrix between OBP, SLG and BA. Do as well a heatmap
# with 2 parameters changed
df_correlation2 = df[['OBP', 'SLG', 'BA']]
df_correlation2.corr()
```

Out[ ]:

	OBP	SLG	BA
OBP	1.000000	0.790910	0.851958
SLG	0.790910	1.000000	0.790481
BA	0.851958	0.790481	1.000000

In [ ]:

```
#Heatmap
sns.heatmap(df_correlation2.corr(),
            annot = True,
```

```
            fmt = '.1g',
            center = 0.7,
            cmap = 'rocket',
            linewidths = 2,
            linecolor = 'yellow')
```

Out[ ]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fb5bd79c8d0>
```

Standard Deviation

In [ ]:

```
#Standard Deviation
print(df.OOBP.mean())
df.OOBP.std()
```

```
0.3322642857142857
```

Out[ ]:

```
0.015295316041389943
```

In [ ]:

```
#Question: what is the standard deviation of BA
print(df.BA.mean())
df.BA.std()
```

```
0.25927272727272727
```

Out[ ]:

```
0.012907228928000314
```