Experiment 3: Statistical Data Analysis

Name: Anuj Chavan

UID: 2019120012

Class: BE EXTC

The given dataset on which statistical analysis is being performed is a football dataset which has all in formation based on the 2018-2019 English Premier League Season. It contains statistics such as :

H - Home team

A - Away team

D - Draw

FTHG - Full time Home goals

FTAG - Full time Away goals

FTR - Full time Result (H,A,D)

B365H - Bet 365 Home odds

B365A - Bet 365 Away odds

B365D - Bet 356 Draw odds

In [ ]:
```
import pandas as pd
import seaborn as sns
import numpy as np
from scipy import stats
```

In [ ]:
```
df = pd.read_csv('season-1819_csv.csv')
```

THe original dataset is displayed below

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

Out[ ]:

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	BbAv<2.5	B
0	E0	10/08/2018	Man United	Leicester	2	1	H	1	0	H	...	1.79	
1	E0	11/08/2018	Bournemouth	Cardiff	2	0	H	1	0	H	...	1.83	
2	E0	11/08/2018	Fulham	Crystal Palace	0	2	A	0	1	A	...	1.87	
3	E0	11/08/2018	Huddersfield	Chelsea	0	3	A	0	2	A	...	1.84	

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	BbAv<2.5	B
4	E0	11/08/2018	Newcastle	Tottenham	1	2	A	1	2	A	...	1.81	

5 rows × 62 columns

In [ ]:
```
df.shape
```

Out[ ]:
```
(380, 62)
```

In [ ]:
```
df.describe()
```

Out[ ]:

	FTHG	FTAG	HTHG	HTAG	HS	AS	HST	AST
count	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000
mean	1.568421	1.252632	0.678947	0.573684	14.134211	11.144737	4.778947	3.928947
std	1.312836	1.180031	0.860802	0.766958	5.855371	4.654002	2.677686	2.283982
min	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	10.000000	8.000000	3.000000	2.000000
50%	1.000000	1.000000	0.000000	0.000000	14.000000	11.000000	5.000000	4.000000
75%	2.000000	2.000000	1.000000	1.000000	18.000000	14.000000	6.000000	5.250000
max	6.000000	6.000000	4.000000	3.000000	36.000000	25.000000	14.000000	12.000000

8 rows × 55 columns

In [ ]:
```
df.drop(df.iloc[:, 26:], inplace = True, axis = 1)
```

All unecessary columns hhave been dropped

In [ ]:
```
df.head(20)
```

Out[ ]:

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	AF	HC
0	E0	10/08/2018	Man United	Leicester	2	1	H	1	0	H	...	8	2
1	E0	11/08/2018	Bournemouth	Cardiff	2	0	H	1	0	H	...	9	7
2	E0	11/08/2018	Fulham	Crystal Palace	0	2	A	0	1	A	...	11	5
3	E0	11/08/2018	Huddersfield	Chelsea	0	3	A	0	2	A	...	8	2
4	E0	11/08/2018	Newcastle	Tottenham	1	2	A	1	2	A	...	12	3
5	E0	11/08/2018	Watford	Brighton	2	0	H	1	0	H	...	16	8
6	E0	11/08/2018	Wolves	Everton	2	2	D	1	1	D	...	7	3

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	AF	HC
7	E0	12/08/2018	Arsenal	Man City	0	2	A	0	1	A	...	14	2
8	E0	12/08/2018	Liverpool	West Ham	4	0	H	2	0	H	...	9	5
9	E0	12/08/2018	Southampton	Burnley	0	0	D	0	0	D	...	9	8
10	E0	18/08/2018	Cardiff	Newcastle	0	0	D	0	0	D	...	16	5
11	E0	18/08/2018	Chelsea	Arsenal	3	2	H	2	2	D	...	9	5
12	E0	18/08/2018	Everton	Southampton	2	1	H	2	0	H	...	20	2
13	E0	18/08/2018	Leicester	Wolves	2	0	H	2	0	H	...	8	1
14	E0	18/08/2018	Tottenham	Fulham	3	1	H	1	0	H	...	5	5
15	E0	18/08/2018	West Ham	Bournemouth	1	2	A	1	0	H	...	10	6
16	E0	19/08/2018	Brighton	Man United	3	2	H	3	1	H	...	13	3
17	E0	19/08/2018	Burnley	Watford	1	3	A	1	1	D	...	19	5
18	E0	19/08/2018	Man City	Huddersfield	6	1	H	3	1	H	...	4	10
19	E0	20/08/2018	Crystal Palace	Liverpool	0	2	A	0	1	A	...	13	6

20 rows × 26 columns

In [ ]:
```
df.describe()
```

Out[ ]:

	FTHG	FTAG	HTHG	HTAG	HS	AS	HST	AST
count	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000	380.000000
mean	1.568421	1.252632	0.678947	0.573684	14.134211	11.144737	4.778947	3.928947
std	1.312836	1.180031	0.860802	0.766958	5.855371	4.654002	2.677686	2.283982
min	0.000000	0.000000	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	10.000000	8.000000	3.000000	2.000000
50%	1.000000	1.000000	0.000000	0.000000	14.000000	11.000000	5.000000	4.000000
75%	2.000000	2.000000	1.000000	1.000000	18.000000	14.000000	6.000000	5.250000
max	6.000000	6.000000	4.000000	3.000000	36.000000	25.000000	14.000000	12.000000

The division column is dropped below

In [ ]:
```
df.drop(['Div'],axis = 1)
```

Out[ ]:

	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	...	AF
0	10/08/2018	Man United	Leicester	2	1	H	1	0	H	A Marriner	...	8

	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	...	AF
1	11/08/2018	Bournemouth	Cardiff	2	0	H	1	0	H	K Friend	...	9
2	11/08/2018	Fulham	Crystal Palace	0	2	A	0	1	A	M Dean	...	11
3	11/08/2018	Huddersfield	Chelsea	0	3	A	0	2	A	C Kavanagh	...	8
4	11/08/2018	Newcastle	Tottenham	1	2	A	1	2	A	M Atkinson	...	12
...	...	...	...	...	...	...	...	...	...	...	...	...
375	12/05/2019	Liverpool	Wolves	2	0	H	1	0	H	M Atkinson	...	11
376	12/05/2019	Man United	Cardiff	0	2	A	0	1	A	J Moss	...	6
377	12/05/2019	Southampton	Huddersfield	1	1	D	1	0	H	L Probert	...	6
378	12/05/2019	Tottenham	Everton	2	2	D	1	0	H	A Marriner	...	13
379	12/05/2019	Watford	West Ham	1	4	A	0	2	A	C Kavanagh	...	10

380 rows × 25 columns

In [ ]:
```
x = df['HTR'].value_counts()
```

We can see below that statistically the most probable result is a draw follwed by the home team victory and least probable is an away team victory

In [ ]:
```
print(x)
```

```
D    148
H    126
A    106
Name: HTR, dtype: int64
```

Given below are the mean, Median and mode for all the parameters.

In [ ]:
```
df.mean()
```

Out[ ]:
```
FTHG      1.568421
FTAG      1.252632
HTHG      0.678947
HTAG      0.573684
HS       14.134211
AS       11.144737
HST       4.778947
AST       3.928947
HF       10.152632
AF       10.305263
HC        5.705263
AC        4.552632
```

```
HY        1.526316
AY        1.684211
HR        0.047368
AR        0.076316
B365H     3.289184
B365D     4.583447
B365A     5.633763
dtype: float64
```

We can see the means for all the parameters. The Home goals are more than away goals. The home team isbetter than away team in almost all statistics.

In [ ]:
```
df.median()
```

Out[ ]:
```
FTHG      1.0
FTAG      1.0
HTHG      0.0
HTAG      0.0
HS       14.0
AS       11.0
HST       5.0
AST       4.0
HF       10.0
AF       10.0
HC        5.0
AC        4.0
HY        1.0
AY        2.0
HR        0.0
AR        0.0
B365H     2.3
B365D     3.8
B365A     3.4
dtype: float64
```

Above is the median of all parameters and below are all the modes.

In [ ]:
```
df.mode()
```

Out[ ]:

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	AF	I
0	E0	12/05/2019	Arsenal	Arsenal	1.0	1.0	H	0.0	0.0	D	...	9.0	!
1	NaN	NaN	Bournemouth	Bournemouth	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
2	NaN	NaN	Brighton	Brighton	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
3	NaN	NaN	Burnley	Burnley	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
4	NaN	NaN	Cardiff	Cardiff	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
5	NaN	NaN	Chelsea	Chelsea	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
6	NaN	NaN	Crystal Palace	Crystal Palace	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
7	NaN	NaN	Everton	Everton	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
8	NaN	NaN	Fulham	Fulham	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
9	NaN	NaN	Huddersfield	Huddersfield	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N
10	NaN	NaN	Leicester	Leicester	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	...	AF	I
11	NaN	NaN	Liverpool	Liverpool	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
12	NaN	NaN	Man City	Man City	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
13	NaN	NaN	Man United	Man United	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
14	NaN	NaN	Newcastle	Newcastle	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
15	NaN	NaN	Southampton	Southampton	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
16	NaN	NaN	Tottenham	Tottenham	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
17	NaN	NaN	Watford	Watford	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
18	NaN	NaN	West Ham	West Ham	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.
19	NaN	NaN	Wolves	Wolves	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	N.

20 rows × 26 columns

In [ ]:
```
df['FTHG'].mean()
```

Out[ ]:
```
1.568421052631579
```

HYPOTHESIS TESTING

In [ ]:
```
from scipy.stats import ttest_1samp
from scipy.stats import chi2_contingency
import scipy.stats
```

For the **First** Hypothesis test:

Null Hypothesis: The full time home goals scored in any match is equal to 1.5 on an average.

Alternate Hypothesis: The full time home goals scored in any match is not equal to 1.5 on an average.

T-Test null hypothesis is done to see wether the null hypothesis is true or not. Alpha leve is set to be 0.05.

In [ ]:
```
tset, pval = ttest_1samp(df['FTHG'], 1.5)
print("P-Value is: "+ str(pval))
if pval < 0.05:    # alpha value is 0.05 or 5%
   print(" We are rejecting null hypothesis")
else:
  print("We are accepting null hypothesis")
```

```
P-Value is: 0.3103027193080958
We are accepting null hypothesis
```

For the **Second** Hypothesis test:

Null Hypothesis: The full time away goals scored in any match is equal to 1.1 on an average.

Alternate Hypothesis: The full time away goals scored in any match is not equal to 1.1 on an average.

T-Test null hypothesis is done to see wether the null hypothesis is true or not. Alpha leve is set to be 0.05.

In [ ]:
```
tset, pval = ttest_1samp(df['FTAG'], 1.1)
print("P-Value is: "+ str(pval))
if pval < 0.05:    # alpha value is 0.05 or 5%
   print("We are rejecting null hypothesis")
else:
  print("We are accepting null hypothesis")
```

```
P-Value is: 0.01209797061109807
We are rejecting null hypothesis
```

For the **Third** Hypothesis test:

Null Hypothesis: The Bet 365 Home team winning odds are 3.3 on average.

Alternate Hypothesis:The Bet 365 Home team winning odds are not 3.3 on average.

T-Test null hypothesis is done to see wether the null hypothesis is true or not. Alpha leve is set to be 0.05.

In [ ]:
```
tset, pval = ttest_1samp(df['B365H'], 3.3)
print("P-Value is: "+ str(pval))
if pval < 0.05:    # alpha value is 0.05 or 5%
   print(" We are rejecting null hypothesis")
else:
  print("We are accepting null hypothesis")
```

```
P-Value is: 0.9485350634069312
We are accepting null hypothesis
```

For the fourth Hypothesis test since draw is the most common result statisticallly we wanted to see that since draw is the most common result we wanted to see if home goals and away goals were any similar.

Null hypothesis: The Full time home goals and full time away goals are similar.

Alternate hypothesis: The Full time home goals and full time away goals are not similar.

F-Test null hypothesis is done to see wether the null hypothesis is true or not. Alpha leve is set to be 0.05.

In [ ]:
```
x = df['FTHG']
y = df['FTAG']
def f_test(group1, group2):
    f = np.var(group1, ddof=1)/np.var(group2, ddof=1)
    nun = x.size-1
    dun = y.size-1
```

```
    p_value = 1-scipy.stats.f.cdf(f, nun, dun)
    return f, p_value

# perform F-test
f_test(x, y)

f, pval = f_test(x, y)
print("P-Value is: "+ str(pval))
if pval < 0.05:    # alpha value is 0.05 or 5%
   print(" We are rejecting null hypothesis")
else:
  print("We are accepting null hypothesis")
```

```
P-Value is: 0.01909060855211875
 We are rejecting null hypothesis
```

The experiment on statistical data analysis has been completed. The mean median mode for all the parameters in the dataset have been calculated.

Statiscally the more likely result is a draw as we saw above.

Hypothesis testing was done for 4 null hypotesis statements above by using various tests.

1. In the first case null hypothesis was accepted by T-test so we can say that the full time home goals is 1.5
2. In the second case the null hypothesis has by T-test been rejected so we can say the first half away goals is not 1.1
3. In the third cas the null hypothesis is accepted by T-test so we can say that the Bet 365 how team winning odds is 3.3 since p value is within the alpha value.
4. In the fourth case because statistically draw is the most possible result we compared the FTHG and FTAG and we found out through F-test the the null hypothesis which is FHTG and FHAG are equal is rejcted because of a smalll value of alpha