



Sardar Patel Institute of Technology, Mumbai
Department of Electronics and Telecommunication Engineering
B.E. Sem-VII (2022-2023) Data Analytics

Experiment: Exploratory Data Analysis (EDA)

Name: Swapnil M

UID: 2019120037

BE EXTC

DA LAB 2

Aim: Perform Exploratory Data Analysis (EDA) on SAS

Dataset Overview

The dataset 'SASHELP.BASEBALL' contains 24 columns with 322 rows containing the information of each baseball player from different countries of the world in his career. :

Code :

```
1 data first;
2 set SASHELP.BASEBALL;
3 run;
4 proc means data= SASHELP.BASEBALL mean median mode std var min max;
5 run;
6
7 proc means data= SASHELP.BASEBALL nmiss;
8 run;
9
10 proc print =SASHELP.BASEBALL;
11 where Div = "NW";
12 run;
13
14 proc sql;
15 select count(distinct Div) as Div,
16        count(distinct Team) as Team,
17        count(distinct CrHits) as CrHits
18 from SASHELP.BASEBALL;
19 quit;
20
21 proc freq data=SASHELP.BASEBALL;
22 tables Team; /* _ALL_ is the default */
23 run;
24
25 data nHits;
26 set SASHELP.BASEBALL(keep=_NUMERIC_ /* all numeric variables */
27                        ); /* two character variables */
28 run;
29
30 proc print data= nHits(obs=5);
31 run;
32
```

```

18  from SASHELP.BASEBALL;
19  quit;
20
21  proc freq data=SASHELP.BASEBALL;
22    tables Team; /* _ALL_ is the default */
23  run;
24
25  data nHits;
26  set SASHELP.BASEBALL(keep=_NUMERIC_ /* all numeric variables */
27    ); /* two character variables */
28  run;
29
30  proc print data= nHits(obs=5);
31  run;
32
33  proc means data=nHits nmiss;
34  run;
35
36  ods graphics / reset width=6.4in height=4.8in imagemap;
37  proc sgplot data=SASHELP.BASEBALL;
38  vbox nAtBat / category=nHome;
39  yaxis grid;
40  run;
41  ods graphics / reset;
42
43  proc ttest data = SASHELP.BASEBALL SIDES=L;
44  class nAtBat;
45  var nHits;
46  run;
47
48
49
50

```

Output :

The MEANS Procedure								
Variable	Label	Mean	Median	Mode	Std Dev	Variance	Minimum	Maximum
nAtBat	Times at Bat in 1986	390.0745342	390.0000000	209.0000000	143.5968352	20619.76	127.0000000	687.0000000
nHits	Hits in 1986	103.3975155	98.5000000	53.0000000	44.1795091	1951.83	31.0000000	238.0000000
nHome	Home Runs in 1986	11.1024845	8.5000000	4.0000000	8.6987696	75.6685919	0	40.0000000
nRuns	Runs in 1986	52.2173913	48.0000000	42.0000000	25.0573661	627.8715969	12.0000000	130.0000000
nRBI	RBIs in 1986	49.3726708	45.0000000	29.0000000	25.5011624	650.3092819	8.0000000	121.0000000
nBB	Walks in 1986	39.8571429	35.5000000	22.0000000	21.0959408	445.0387183	3.0000000	105.0000000
YrMajor	Years in the Major Leagues	7.6801242	6.0000000	4.0000000	4.9697066	24.6979638	1.0000000	24.0000000
CrAtBat	Career Times at Bat	2763.08	2065.00	216.0000000	2328.48	5421815.23	168.0000000	14053.00
CrHits	Career Hits	747.6863354	552.0000000	160.0000000	654.7878194	428746.83	34.0000000	4256.00
CrHome	Career Home Runs	74.0903621	40.0000000	18.0000000	90.0661266	8111.73	0	548.0000000
CrRuns	Career Runs	374.2857143	266.0000000	20.0000000	336.4250377	113181.81	18.0000000	2165.00
CrRbi	Career RBIs	347.6149088	250.0000000	32.0000000	338.7903452	114776.90	9.0000000	1659.00
CrBB	Career Walks	273.3944099	178.5000000	55.0000000	273.6253716	74870.84	8.0000000	1566.00
nOuts	Put Outs in 1986	288.9937886	212.0000000	0	280.8566732	78766.17	0	1378.00
nAssists	Assists in 1986	106.9161491	35.5000000	0	136.8524541	18728.59	0	462.0000000
nError	Errors in 1986	8.0403727	6.0000000	3.0000000	6.3683591	40.5559974	0	32.0000000
Salary	1987 Salary in \$ Thousands	535.9258821	425.0000000	750.0000000	451.1186807	203508.06	67.5000000	2460.00
logSalary	Log Salary	5.9272215	6.0520892	6.6200732	0.8891924	0.7906631	4.2121276	7.8079166

The MEANS Procedure		
Variable	Label	N Miss
nAtBat	Times at Bat in 1986	0
nHits	Hits in 1986	0
nHome	Home Runs in 1986	0
nRuns	Runs in 1986	0
nRBI	RBIs in 1986	0
nBB	Walks in 1986	0
YrMajor	Years in the Major Leagues	0
CrAtBat	Career Times at Bat	0
CrHits	Career Hits	0
CrHome	Career Home Runs	0
CrRuns	Career Runs	0
CrRbi	Career RBIs	0
CrBB	Career Walks	0
nOuts	Put Outs in 1986	0
nAssists	Assists in 1986	0
nError	Errors in 1986	0
Salary	1987 Salary in \$ Thousands	59
logSalary	Log Salary	59

Div	Team	CrHits
4	24	287

The FREQ Procedure

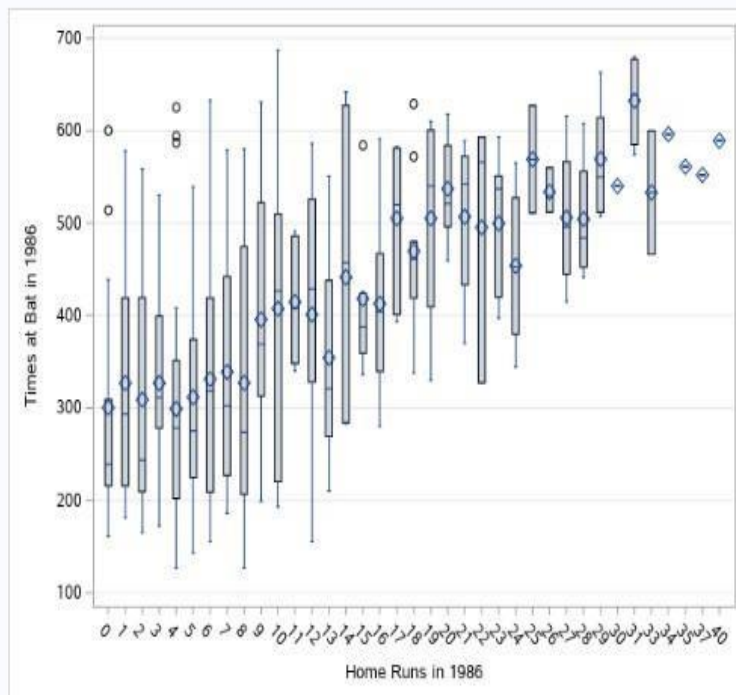
Team at the End of 1986				
Team	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Atlanta	11	3.42	11	3.42
Baltimore	15	4.66	26	8.07
Boston	10	3.11	36	11.18
California	13	4.04	49	15.22
Chicago	24	7.45	73	22.67
Cincinnati	12	3.73	85	26.40
Cleveland	12	3.73	97	30.12
Detroit	12	3.73	109	33.85
Houston	11	3.42	120	37.27
Kansas City	14	4.35	134	41.61
Los Angeles	14	4.35	148	45.96
Milwaukee	14	4.35	162	50.31
Minneapolis	13	4.04	175	54.35
Montreal	14	4.35	189	58.70
New York	24	7.45	213	66.15
Oakland	12	3.73	225	69.88
Philadelphia	12	3.73	237	73.60
Pittsburgh	11	3.42	248	77.02
San Diego	13	4.04	261	81.06
San Francisco	14	4.35	275	85.40
Seattle	12	3.73	287	89.13

California	13	4.04	49	15.22
Chicago	24	7.45	73	22.67
Cincinnati	12	3.73	85	26.40
Cleveland	12	3.73	97	30.12
Detroit	12	3.73	109	33.85
Houston	11	3.42	120	37.27
Kansas City	14	4.35	134	41.61
Los Angeles	14	4.35	148	45.96
Milwaukee	14	4.35	162	50.31
Minneapolis	13	4.04	175	54.35
Montreal	14	4.35	189	58.70
New York	24	7.45	213	66.15
Oakland	12	3.73	225	69.88
Philadelphia	12	3.73	237	73.60
Pittsburgh	11	3.42	248	77.02
San Diego	13	4.04	261	81.06
San Francisco	14	4.35	275	85.40
Seattle	12	3.73	287	89.13
St Louis	11	3.42	298	92.55
Texas	13	4.04	311	96.58
Toronto	11	3.42	322	100.00

Obs	nAtBat	nHits	nHome	nRuns	nRBI	nBB	YrMajor	CrAtBat	CrHits	CrHome	CrRuns	CrRbi	CrBB	nOuts	nAsts	nError	Salary	logSalary
1	293	66	1	30	29	14	1	293	66	1	30	29	14	446	33	20	.	.
2	315	81	7	24	36	39	14	3449	835	69	321	414	375	632	43	10	475.0	6.16331
3	479	130	18	68	72	76	3	1624	457	63	224	266	263	880	82	14	480.0	6.17379
4	496	141	20	65	78	37	11	5628	1575	225	828	838	354	200	11	3	500.0	6.21461
5	321	87	10	39	42	30	2	396	101	12	48	46	33	805	40	4	91.5	4.51634

The MEANS Procedure

Variable	Label	N Miss
nAtBat	Times at Bat in 1986	0
nHits	Hits in 1986	0
nHome	Home Runs in 1986	0
nRuns	Runs in 1986	0
nRBI	RBIs in 1986	0
nBB	Walks in 1986	0
YrMajor	Years in the Major Leagues	0
CrAtBat	Career Times at Bat	0
CrHits	Career Hits	0
CrHome	Career Home Runs	0
CrRuns	Career Runs	0
CrRbi	Career RBIs	0
CrBB	Career Walks	0
nOuts	Put Outs in 1986	0
nAssists	Assists in 1986	0
nErrors	Errors in 1986	0
Salary	1987 Salary in \$ Thousands	59
logSalary	Log Salary	59



Conclusion:

1. Performed EDA on SAS Studio for baseball dataset.
2. Exploratory Data Analysis refers to the critical process of performing critical investigations on data so as to discover patterns or to spot anomalies.
3. Few insights we found from the dataset:
 - Variance is the highest for career RBI's for each baseball player.
 - It is noted that players of Philadelphia team has the highest cumulative frequency for the attributes in the dataset.
 - From the graph we can see that the times a player has batted in the baseball game is the highest for 10 home runs in the year 1986 for the dataset.
 - The teams at the end of 1986 with the largest frequency is for New York and Chicago