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)

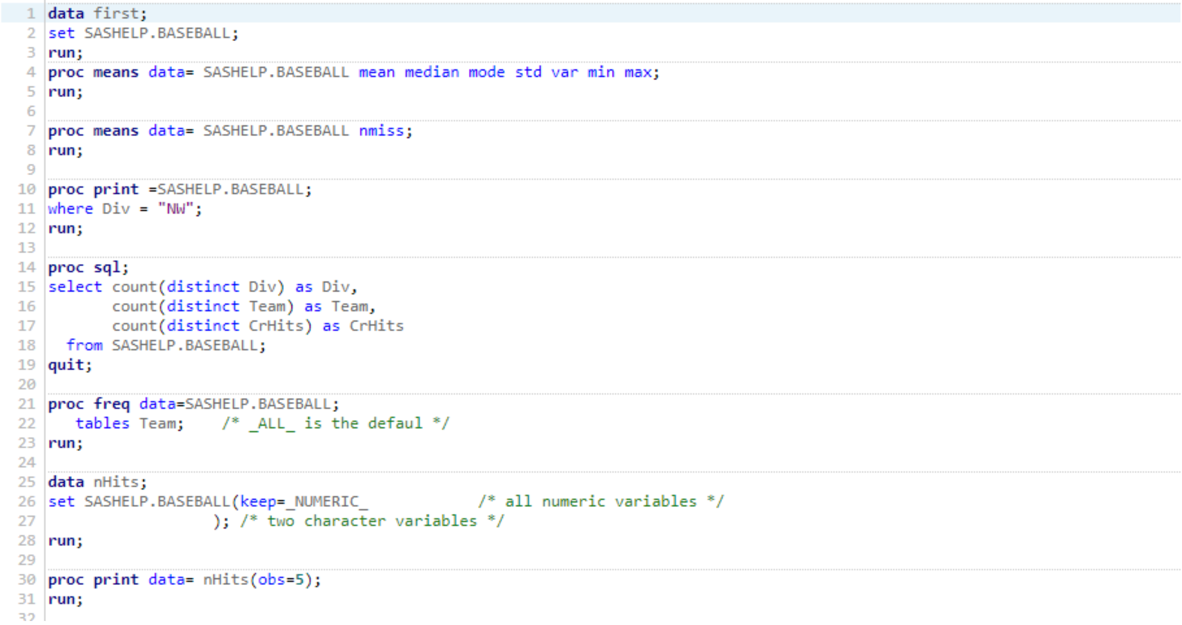
Name: Rahul Alshi **UID:** 2019110001 **BE ETRX** **DA LAB 3**

Aim: Perform Statistical Analysis (tests) on SAS

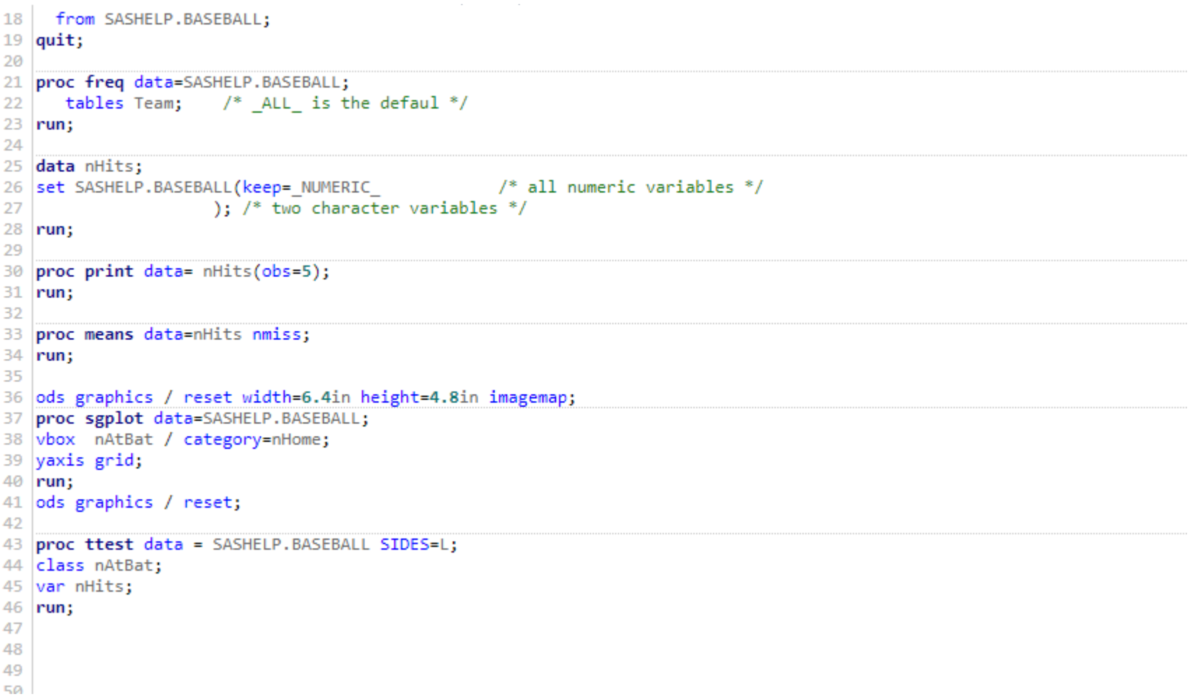
Dataset Overview

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. :

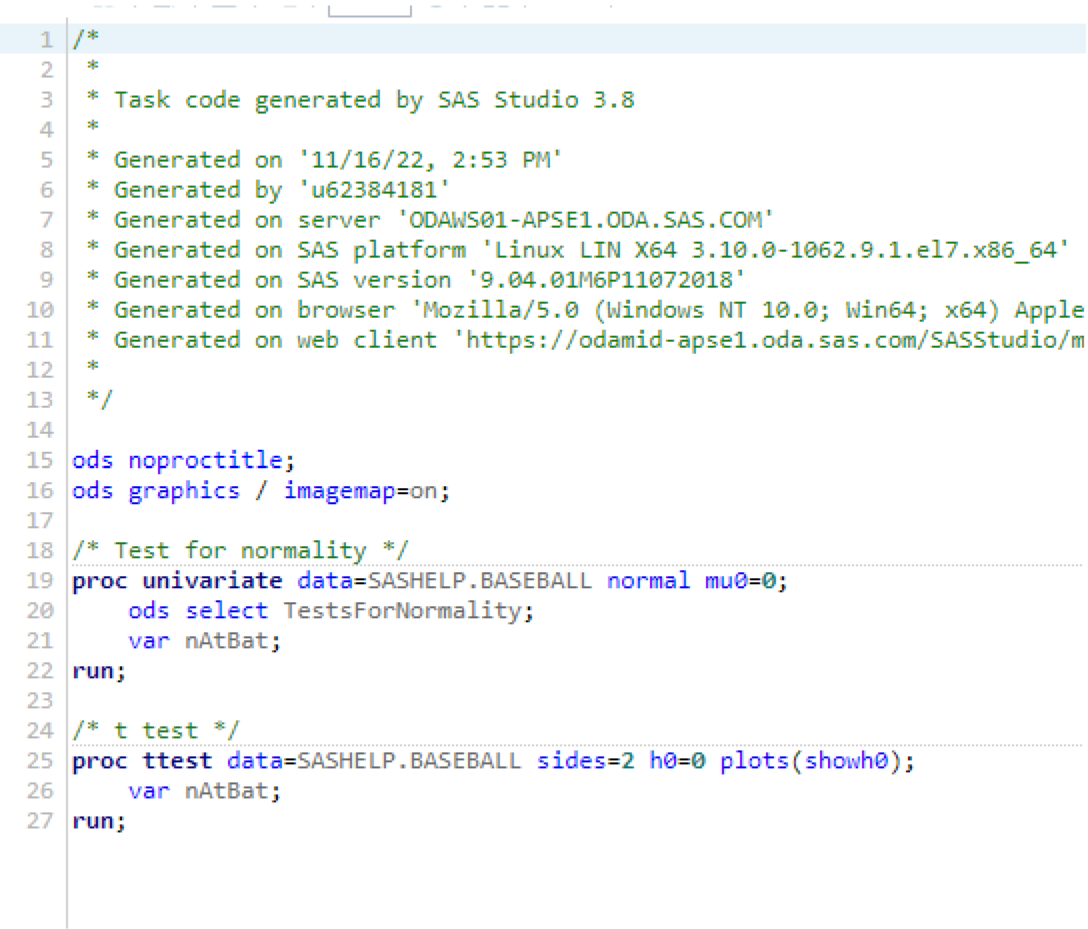
Code :



```
data first;
set SASHELP.BASEBALL;
run;
proc means data= SASHELP.BASEBALL mean median mode std var min max;
run;

proc means data= SASHELP.BASEBALL nmiss;
run;

proc print =SASHELP.BASEBALL;
where Div = "NW";
run;

proc sql;
select count(distinct Div) as Div,
       count(distinct Team) as Team,
       count(distinct CrHits) as CrHits
  from SASHELP.BASEBALL;
quit;

proc freq data=SASHELP.BASEBALL;
   tables Team;    /* _ALL_ is the defaul */
run;

data nHits;
set SASHELP.BASEBALL(keep=_NUMERIC_            /* all numeric variables */
                  ); /* two character variables */
run;

proc print data= nHits(obs=5);
run;
```

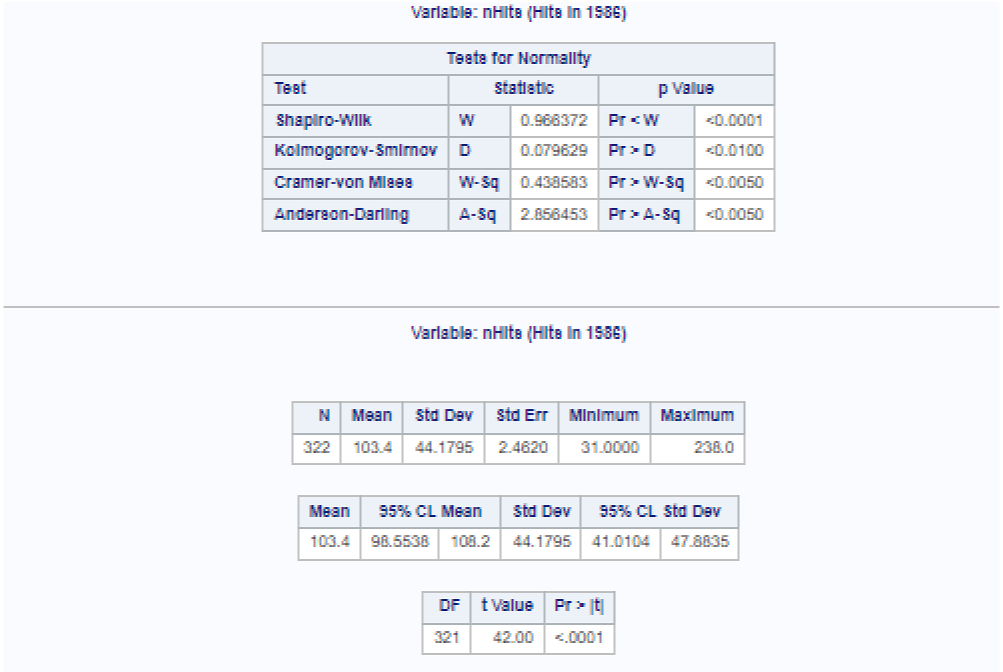


```
  from SASHELP.BASEBALL;
quit;

proc freq data=SASHELP.BASEBALL;
   tables Team;    /* _ALL_ is the defaul */
run;

data nHits;
set SASHELP.BASEBALL(keep=_NUMERIC_            /* all numeric variables */
                  ); /* two character variables */
run;

proc print data= nHits(obs=5);
run;

proc means data=nHits nmiss;
run;

ods graphics / reset width=6.4in height=4.8in imagemap;
proc sgplot data=SASHELP.BASEBALL;
vbox  nAtBat / category=nHome;
yaxis grid;
run;
ods graphics / reset;

proc ttest data = SASHELP.BASEBALL SIDES=L;
class nAtBat;
var nHits;
run;
```



```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '11/16/22, 2:53 PM'
 * Generated by 'u62384181'
 * Generated on server 'ODAWS01-APSE1.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 3.10.0-1062.9.1.el7.x86_64'
 * Generated on SAS version '9.04.01M6P11072018'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) Apple
 * Generated on web client 'https://odamid-apse1.oda.sas.com/SASStudio/m
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

/* Test for normality */
proc univariate data=SASHELP.BASEBALL normal mu0=0;
    ods select TestsForNormality;
    var nAtBat;
run;

/* t test */
proc ttest data=SASHELP.BASEBALL sides=2 h0=0 plots(showh0);
    var nAtBat;
run;
```

Output :



Variable: nHits (Hits in 1986)

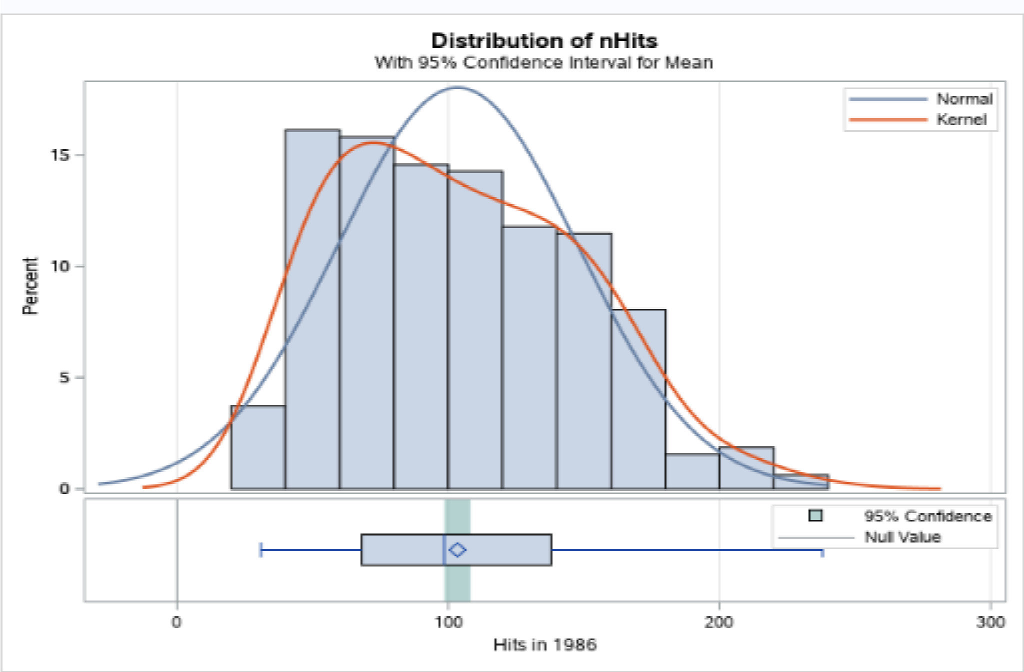
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.966372	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.079629	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.438583	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.856453	Pr > A-Sq	<0.0050

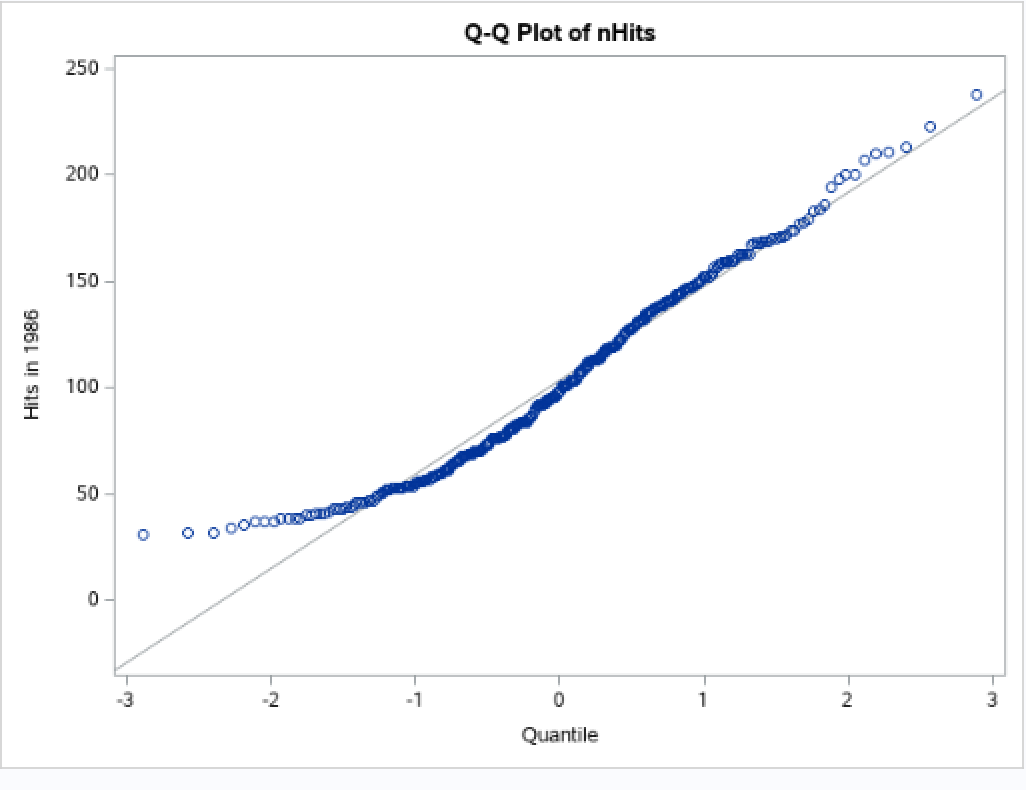
Variable: nHits (Hits in 1986)

N	Mean	Std Dev	Std Err	Minimum	Maximum
322	103.4	44.1795	2.4620	31.0000	238.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
103.4	98.5538	108.2	44.1795	41.0104	47.8835

DF	t Value	Pr > |t|
321	42.00	<.0001





Conclusion:

1. Performed statistical analysis (Hypothesis testing) on SAS Studio for baseball dataset.
2. A t test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

3. The t value for the one sample t test for the no of hits in the year 1986 is calculated to be 42
4. Few insights we found from the dataset:
    - With 95% confidence interval for mean on distribution of n Hits , the distribution is highest for 100 Hits in 1986 with above 15%
    - The no of hits goes on increasing per quantile upto 250 from the Q-Q plot.
    - Similarly we can perform t- tests for 2 sampled and 1 sampled means wwith different attributes in the dataset.