TASK 1

ANSWER ALL QUESTIONS

ANSWER:

QUESTION 1:

ANSWER:

i)

```
DATA VEGE;
INPUT PROTEIN BP @@;
CARDS;
4 73 6.5 79 5 83 5.5 82 8 84 10 92 9 88 8.2 86 10.5 95
;
RUN;

PROC PRINT DATA = VEGE NOOBS;
RUN;
```

PROTEIN	BP
4.0	73
6.5	79
5.0	83
5.5	82
8.0	84
10.0	92
9.0	88
8.2	86
10.5	95

```
PROC CORR DATA = VEGE;
VAR PROTEIN BP;
RUN;
```

Pearson Correlation Coefficients, N = 9 Prob > |r| under H0: Rho=0		
	PROTEIN	BP
PROTEIN	1.00000	0.91598 0.0005
BP	0.91598 0.0005	1.00000

ii) The correlation coefficient of protein and blood pressure is 0.91598. This indicates that there exists strong positive relationship between protein and blood pressure.

iii) The p-value is 0.0005 which is less than alpha = 0.05. Thus, there is a significant linear relationship between protein and blood pressure.

QUESTION 2:

ANSWER:

a)

```
data petrol;
input petrol @@;
cards;
81 80 65 105 144 75 150 96 91 68 135 134 95 124
;
run;
PROC PRINT DATA=PETROL NOOBS; RUN;
```

petrol
81
80
65
105
144
75
150
96
91
68
135
134
95
124

```
PROC SUMMARY DATA=PETROL Q1 MEDIAN Q3 PRINT;
VAR PETROL;
RUN;
```

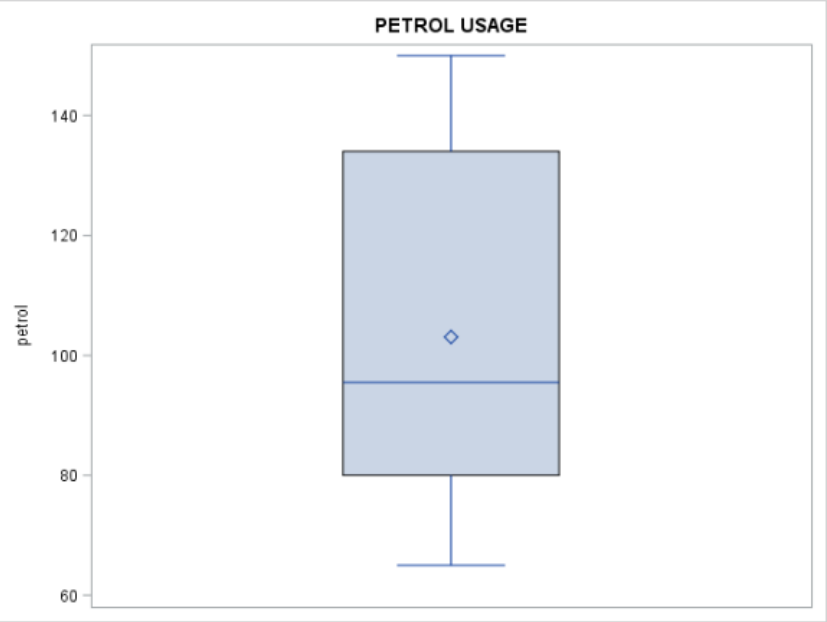
Analysis Variable : petrol		
Lower Quartile	Median	Upper Quartile
80.0000000	95.5000000	134.0000000

From the table above, first quartile is 80, median is 95.5 and third quartile is 134.

b)

```
PROC SGPLOT DATA=PETROL;
VBOX PETROL;
TITLE "PETROL USAGE";
RUN;
TITLE;
```

QUESTION 3:



c) Based on the figure above, it shows that the box-and-whiskers for this data is skewed to the right (positive skew). The value for quartile 1 is 80, median is 95.5 and quartile 3 is 134 which is the same as the summary table.

ANSWER:

a)
```
DATA SECRETARY;
INPUT SPEED TIME @@;
CARDS;
48 7 74 4 52 8 79 3.5 83 2 56 6 85 2.3 63 5 88 2.1 74 4.5 90 1.9 92 1.5
;
RUN;

PROC PRINT DATA=SECRETARY NOOBS; RUN;
```

SPEED	TIME
48	7.0
74	4.0
52	8.0
79	3.5
83	2.0
56	6.0
85	2.3
63	5.0
88	2.1
74	4.5
90	1.9
92	1.5

```
PROC REG DATA= SECRETARY;
MODEL TIME = SPEED;
RUN;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.08565	0.75538	18.65	<.0001
SPEED	1	-0.13714	0.01005	-13.64	<.0001

Estimated regression function: TIME = 14.08565 – 0.13714 SPEED.

INTERPRET THE SLOPE: if the typing speed of a secretary increase by 1 minute, we predict the time that it takes the secretary to learn to use a new word processing program will increase by approximately 0.13714 hours.

b)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	49.02139	49.02139	186.02	<.0001
Error	10	2.63528	0.26353		
Corrected Total	11	51.65667			

Since the value of p-value (<0.001) less than alpha 0.05, we can conclude that the model is significant.

c)

Root MSE	0.51335	R-Square	0.9490
Dependent Mean	3.98333	Adj R-Sq	0.9439
Coeff Var	12.88744		

The R-square value is 0.949. 94.90% of total variation of the time that it takes the secretary to learn to use a new word processing program was explained by the typing speed of a secretary. The rest was explained by other factors.

d)
Basically, I just add the new observation (72) in the data set with missing value for time.

```
DATA SECRETARY;
INPUT SPEED TIME @@;
CARDS;
48 7 74 4 52 8 79 3.5 83 2 56 6 85 2.3 63 5 88 2.1 74 4.5 90 1.9 92 1.5 72 .
;
RUN;

PROC PRINT DATA=SECRETARY NOOBS; RUN;
```

SPEED	TIME
48	7.0
74	4.0
52	8.0
79	3.5
83	2.0
56	6.0
85	2.3
63	5.0
88	2.1
74	4.5
90	1.9
92	1.5
72	.

```
PROC REG DATA= SECRETARY;
MODEL TIME = SPEED;
output out=want p=ypredicted lcl=ylowerin lclm=ylowermean ucl=yupperin
uclm=yuppermean;
RUN;
proc print data=want;
run;
```

Obs	SPEED	TIME	ypredicted	ylowermean	yuppermean	ylowerin	yupperin
1	48	7.0	7.50315	6.84007	8.16622	6.18103	8.82526
2	74	4.0	3.93762	3.60735	4.26790	2.74708	5.12816
3	52	8.0	6.95460	6.36754	7.54167	5.66893	8.24028
4	79	3.5	3.25194	2.90080	3.60309	2.05544	4.44844
5	83	2.0	2.70340	2.31257	3.09423	1.49466	3.91214
6	56	6.0	6.40606	5.89062	6.92150	5.15147	7.66065
7	85	2.3	2.42913	2.01260	2.84566	1.21184	3.64642
8	63	5.0	5.44611	5.03852	5.85371	4.23185	6.66038
9	88	2.1	2.01772	1.55714	2.47831	0.78466	3.25079
10	74	4.5	3.93762	3.60735	4.26790	2.74708	5.12816
11	90	1.9	1.74345	1.25058	2.23633	0.49797	2.98894
12	92	1.5	1.46918	0.94219	1.99618	0.20980	2.72856
13	72	.	4.21189	3.87960	4.54419	3.02079	5.40300

The predicted value for the time if the speed 72 is 4.21189 hours.

The time it will take for the average secretary who has a typing speed of 72 words per minute to learn to use a new word processing program lies between 3.87960 and 4.54419 hours.

QUESTION 4:

ANSWER:

a)

```
DATA BAKERY(DROP=I);
INPUT DESSERT $ @;
DO I = 1 TO 5;
INPUT RATING @;
OUTPUT; END;
CARDS;
P 7 8 7 7 6
Q 2 4 3 5 4
R 10 8 9 7 6
;
RUN;
PROC PRINT DATA=BAKERY NOOBS; RUN;
```

DESSERT	RATING
P	7
P	8
P	7
P	7
P	6
Q	2
Q	4
Q	3
Q	5
Q	4
R	10
R	8
R	9
R	7
R	6

```
PROC ANOVA DATA=BAKERY;
CLASS DESSERT;
MODEL RATING = DESSERT;
RUN;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53.20000000	26.60000000	18.56	0.0002
Error	12	17.20000000	1.43333333		
Corrected Total	14	70.40000000			

R-Square	Coeff Var	Root MSE	RATING Mean
0.755682	19.30998	1.197219	6.200000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DESSERT	2	53.20000000	26.60000000	18.56	0.0002

b)

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DESSERT	2	53.20000000	26.60000000	18.56	0.0002

The p-value (0.00002) is less than 0.05. we can conclude that dessert affects significantly towards the rating.

c)

```
PROC MEANS DATA = BAKERY; RUN;
```

Analysis Variable : RATING				
N	Mean	Std Dev	Minimum	Maximum
15	6.2000000	2.2424476	2.0000000	10.0000000

The overall mean for this model is 6.2.

d)

```
PROC ANOVA DATA=BAKERY;
CLASS DESSERT;
MODEL RATING = DESSERT;
MEANS DESSERT / TUKEY LINES;
RUN;
```

Means with the same letter are not significantly different.			
Tukey Grouping	Mean	N	DESSERT
A	8.0000	5	R
A			
A	7.0000	5	P
B	3.6000	5	Q

The pair of dessert R and P is not significantly different while the pair of dessert P and Q is significantly different. Pair of dessert R and Q also significantly different.

e)

```
PROC MEANS DATA=BAKERY N NMISS CLM;
BY DESSERT;
VAR RATING;
RUN;
```

DESSERT=P

Analysis Variable : RATING			
N	N Miss	Lower 95% CL for Mean	Upper 95% CL for Mean
5	0	6.1220110	7.8779890

The confidence interval of mean for dessert P lies between 6.122 and 7.878.

DESSERT=R

Analysis Variable : RATING			
N	N Miss	Lower 95% CL for Mean	Upper 95% CL for Mean
5	0	6.0367568	9.9632432

The confidence interval of mean for dessert R lies between 6.037 and 9.963.

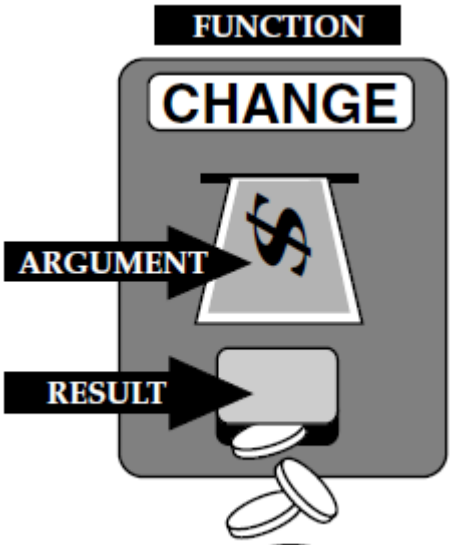
TASK 2

SAS FUNCTION: SAMPLE STATISTICS FUNCTION

INTRODUCTION

A FUNCTION returns a value from a computation or system manipulation that requires zero or more arguments. And, like most programming languages, the SAS System provides an extensive library of "built-in" functions. SAS has over 400 functions in the following general areas:

1. Character
2. Probability
3. Date and Time
4. Random Number
5. Financial
6. Sample Statistics
7. Macro
8. State and ZIP Code
9. Mathematical.



But, in this paper, I am focusing on sample statistics function.

- **ADVANTAGES**
    1. It can operate on values that measure central tendency, variation among values, and the shape of distribution values.
    2. Easy to learn.
    3. Stable Code / Procedures.
- **DISADVANTAGES**
    1. Lack of graphic representation.

- **FUNCTION - Syntax**

Given the above definition of a function, the syntax and components should be examined. A function recognized in a SAS statement by the use of a function name, followed immediately by function argument(s), separated by commas, and enclosed in parentheses. For a function requiring two arguments the syntax is as follows:

FUNCTIONNAME (argument-1, argument-2)

where the arguments are:

1. Constants
2. Variables
3. Expressions
4. Other functions

The syntax of a SAS function involving Variable List or Arrays:

- **Function-name (OF variable list)**

Example:

MEAN (**OF** Var1 – Var5); computes the mean of Var1 to Var5

MEAN (Var1 – Var5); **does not** compute the mean of Var1 to Var5;

instead, if computes the average of Var1 **MINUS** Var5

- **Target Variables for SAS Functions**

Target variable is the variable to which the result of a SAS function is assigned.

For Example:

Avg_score = Mean (of Quiz1 – Quiz 5);

Avg_score is the target variable.

One important property of a Target Variable is the Variable Length. The length depends on the function.

- **FUNCTION - Arguments**

The arguments to any given function can be variables, constants, expressions, and/or other functions. The SUM function requires at least two arguments and returns the sum of the nonmissing arguments.

TOTAL = SUM(X, Y, Z) ;

Use of the keyword OF gives the user the flexibility to include variable lists, array elements and other shortcuts for referencing variable names.

Examples:

A = SUM(OF TEMP1 - TEMP24);

B = SUM(OF TEMP1 TEMP2 TEMP3);

C = SUM(OF _NUMERIC_);

Table 1.1: Usage of Function.

Function Category	Single Function	Summary Function
Sample Statistics	KURTOSIS, SKEWNESS	All except those mentioned under single functions*

Summary functions can operate as either single or summary functions. These include the most of the sample statistical functions such as MAX, MIN, and MEAN.

If the arguments for this function consist of two or more columns, they behave as single functions. The column values in each row are used to generate a result for that row.

Table 1.2: Sample Statistics Function.

NO.	FUNCTION NAME	DESCRIPTION
1	CSS(argument,argument,...)	returns the corrected sum of squares
2	CV(argument,argument,...)	returns the coefficient of variation
3	KURTOSIS(argument,argument,...)	returns the kurtosis (or 4th moment)
4	MAX(argument,argument, ...)	returns the largest value
5	MIN(argument,argument, ...)	returns the smallest value
6	MEAN(argument,argument, ...)	returns the arithmetic mean (average)
7	MISSING(numeric-expression | character-expression)	returns a numeric result that indicates whether the argument contains a missing value
8	N(argument,argument, ...)	returns the number of nonmissing values
9	NMISS(argument,argument, ...)	Returns the number of missing numeric values
10	ORDINAL(count,argument,argument,...)	Returns the kth smallest of the missing and nonmissing values.
11	RANGE(argument,argument,...)	returns the range of values
12	SKEWNESS(argument,argument,argument,...)	returns the skewness
13	STD(argument,argument,...)	returns the standard deviation
14	STDERR(argument,argument,...)	returns the standard error of the mean
15	SUM(argument,argument,...)	Returns the sum of the nonmissing arguments
16	USS(argument,argument,...)	returns the uncorrected sum of squares
17	VAR(argument,argument,...)	returns the variance
18	CMISS(argument,argument,...)	Counts the number of missing arguments.
19	EUCLID	Returns the Euclidean norm of the nonmissing arguments.
20	GEOMEAN	Returns the geometric mean.

21	GEOMEANZ	Returns the geometric mean, using zero fuzzing.
22	HARMEAN	Returns the harmonic mean.
23	HARMEANZ	Returns the harmonic mean, using zero fuzzing.
24	IQR	Returns the interquartile range.
25	LARGEST	Returns the kth largest nonmissing value.
26	LPNORM	Returns the Lp norm of the second argument and subsequent nonmissing arguments.
27	MAD	Returns the median absolute deviation from the median.
28	MEDIAN	Returns the median value.
29	PCTL	Returns the percentile that corresponds to the percentage.
30	RMS	Returns the root mean square of the nonmissing arguments.
31	SMALLEST	Returns the kth smallest nonmissing value.
32	SUMABS	Returns the sum of the absolute values of the nonmissing arguments.
33	DIVIDE	returns the result of a division that handles special missing values for ODS output

- **DATASET**

The dataset was created. This data is about the quiz marks for each student.

Obs	NAME	Q1	Q2	Q3	Q4	Q5
1	ahmad	15	16	20	7	20
2	albab	12	14	17	13	12
3	bujang	19	17	16	19	13
4	lapok	20	20	18	10	17
5	ramlee	17	18	17	18	16
6	misha	15	.	20	14	18
7	keira	20	15	20	.	15
8	tasha	18	14	19	12	20
9	danial	20	15	18	14	19
10	iman	18	15	19	14	20

Then, use the functions to analyse the data.

NO.	FUNCTION NAME	EXAMPLE	RESULT	DESCRIPTION
1	CSS	`QUIZ2=CSS(16,14,17,20,18,.,15,14,15,15);`	32	returns the corrected sum of squares
2	CV	`CVQUIZ2=CV(16,14,17,20,18,.,15,14,15,15);`	12.5	returns the coefficient of variation
3	KURTOSIS	`KQUIZ2=KURTOSIS(16,14,17,20,18,.,15,14,15,15);`	0.58482	returns the kurtosis (or 4th moment)
4	MAX	`MAX=MAX(OF Q1-Q5);`	20,17,19…,20	returns the largest value
5	MIN	`QUIZ2=MIN(16,14,17,20,18,.,15,14,15,15);`	14	returns the smallest value
6	MEAN	`MEAN=MEAN(OF Q1-Q5);`	15.6,13.6,…,17.2	returns the arithmetic mean (average)

7	MISSING	`IF MISSING(Q1) THEN DO;PUT "Q1 IS MISSING.";END;`
		`ELSE IF MISSING(Q2) THEN DO;PUT "Q2 IS MISSING.";`
		`END;ELSE IF MISSING(Q4) THEN DO;`
		`PUT "Q4 IS MISSING.";END;`
			Look at output:
			
			Q2 IS MISSING.
			Q4 IS MISSING.
				returns a numeric result contains a missing value

8	N	`QUIZ=N(16,14,17,20,18,.,15,14,15,15);`	9	returns the number of nonmissing values
9	NMISS	`NMISSALL=NMISS(OF Q1-Q5);`	0,0,0,…,0	returns the number of missing values
10	ORDINAL	`OQ4=ORDINAL(7,13,19,10,18,14,.,12,14,14);`	14	Returns the kth smallest of the missing and nonmissing values.
11	RANGE	`RANGE=RANGE(OF Q1-Q5);`	13,5,6,…,6	returns the range of values
12	SKEWNESS	`SQUIZ2=SKEWNESS(16,14,17,20,18,.,15,14,15,15);`	1.08482	returns the skewness
13	STD	`STDQUIZ2=STD(16,14,17,20,18,.,15,14,15,15);`	2	returns the standard deviation
14	STDERR	`STDERRQ2=STDERR(16,14,17,20,18,.,15,14,15,15);`	0.66667	returns the standard error of the mean
15	SUM	`SUM=SUM(OF Q1-Q5);`	78,68,84,…,86	Returns the sum of the nonmissing arguments
16	USS	`USSQUIZ2=USS(16,14,17,20,18,.,15,14,15,15);`	2336	returns the uncorrected sum of squares
17	VAR	`VARQ3=VAR(20,17,16,18,17,20,20,19,18,19);`	2.04444	returns the variance

CONCLUSION

The intent of this paper was explaining the SAS function: sample statistics functions. Readers can have more understanding on how to use these functions in SAS commands.

ADDITIONAL

From the Valley of the Sun Users Group (VALSUG), Phoenix, AZ, July 1998 Newsletter:

“Why is there a DIM function and not a BRIGHTEN one?”

“And a MEAN function and not a NICE function?”

“And with a FLOOR function and a CEIL function, wouldn’t you think they’d need a WALL function?”

REFERENCES

Cody R. (2010). SAS Functions by Examples. Second Edition.

Cody R. (2007). Learning SAS by Example: A Programmer's Guide, Second Edition.

Delwiche, Lora D. and Slaughter, Susan J., 2003. *The Little SAS_ Book: A Primer, Third Edition.* Cary, NC: SAS Institute Inc.

Field, A., & Miles, J. (2012). *Discovering Statistics Using SAS*®, Thousand Oaks, CA: Sage Publications.

Introduction to SAS. UCLA: Statistical Consulting Group.

SAS® Institute Inc. 2008. SAS /STAT® 9.2 *User's Guide*. Cary, NC: SAS® Institute Inc.

SAS Institute Inc. SAS Language: Reference, Version 6, First Edition, Cary, NC: SAS Institute Inc., 1990.

Yen C. C. (2016). Statistical Software and Its Applications SAS Functions.

APPENDIX

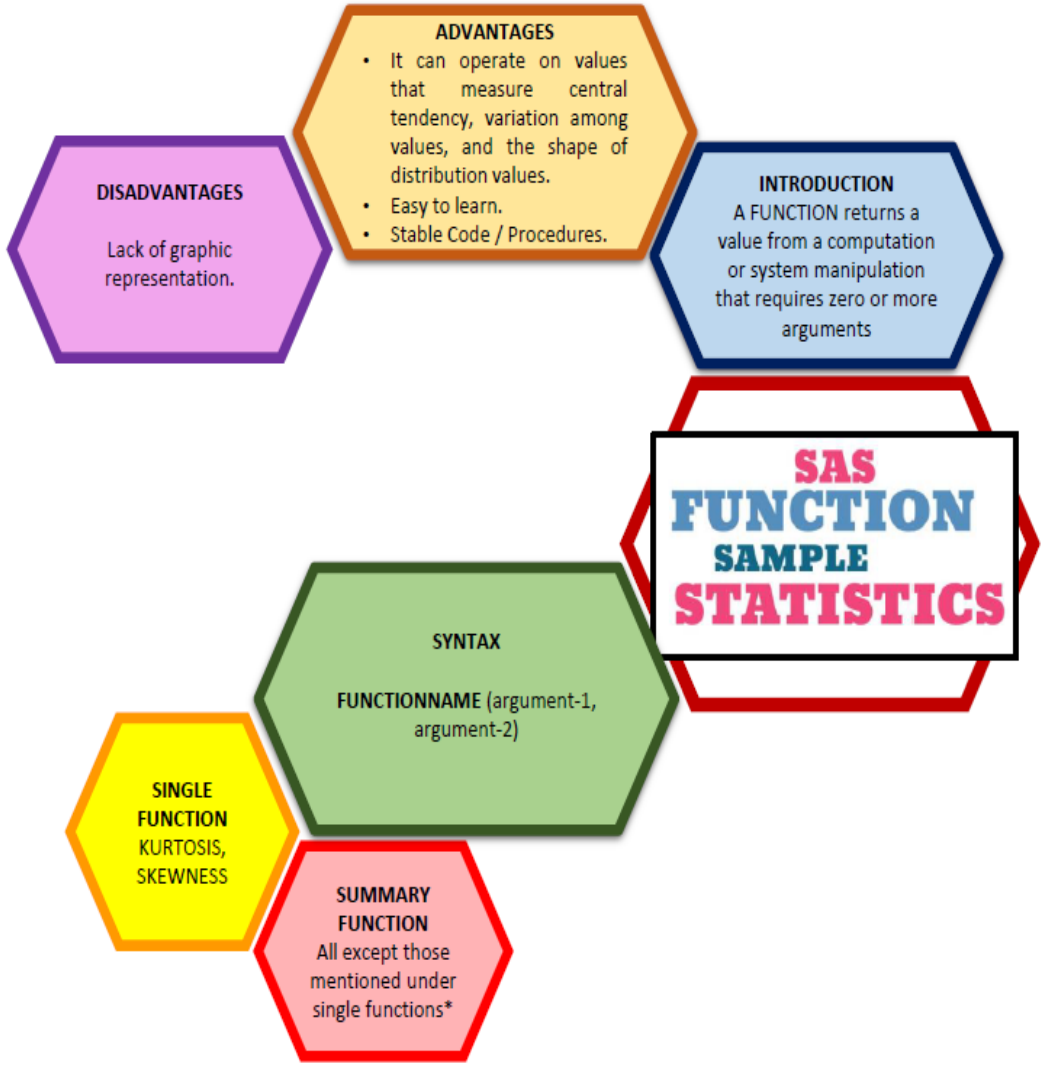
```
Data quiz;
input name $ q1 q2 q3 q4 q5;

SUM=SUM(OF Q1-Q5);
MEAN=MEAN(OF Q1-Q5);
CSSQUIZ2=CSS(16,14,17,20,18,.,15,14,15,15);
CVQUIZ2=CV(16,14,17,20,18,.,15,14,15,15);
KQUIZ2=KURTOSIS(16,14,17,20,18,.,15,14,15,15);
MAX=MAX(OF Q1-Q5);
QUIZ=N(16,14,17,20,18,.,15,14,15,15);
QUIZ2=MIN(16,14,17,20,18,.,15,14,15,15);
NMISSALL=NMISS(OF Q1-Q5);
OQ4=ORDINAL(7,13,19,10,18,14,.,12,14,14);
RANGE=RANGE(OF Q1-Q5);
SQUIZ2=SKEWNESS(16,14,17,20,18,.,15,14,15,15);
STDQUIZ2=STD(16,14,17,20,18,.,15,14,15,15);
STDERRQ2=STDERR(16,14,17,20,18,.,15,14,15,15);
USSQUIZ2=USS(16,14,17,20,18,.,15,14,15,15);
VARQ3=VAR(20,17,16,18,17,20,20,19,18,19);

IF MISSING(Q1) THEN DO;
PUT "Q1 IS MISSING.";
END;
ELSE IF MISSING(Q2) THEN DO;
PUT "Q2 IS MISSING.";
END;
ELSE IF MISSING(Q4) THEN DO;
PUT "Q4 IS MISSING.";
END;

datalines;
ahmad   15 16 20 7 20
albab   12 14 17 13 12
bujang   19 17 16 19 13
lapok   20 20 18 10 17
ramlee   17 18 17 18 16
misha   15 .  20 14 18
keira   20 15 20 .  15
tasha   18 14 19 12 20
danial  20 15 18 14 19
iman   18 15 19 14 20
;
proc print;
run;
```

Obs	name	q1	q2	q3	q4	q5	SUM	MEAN	MAX	RANGE	NMISSALL
1	ahmad	15	16	20	7	20	78	15.60	20	13	0
2	albab	12	14	17	13	12	68	13.60	17	5	0
3	bujang	19	17	16	19	13	84	16.80	19	6	0
4	lapok	20	20	18	10	17	85	17.00	20	10	0
5	ramlee	17	18	17	18	16	86	17.20	18	2	0
6	misha	15	.	20	14	18	67	16.75	20	6	1
7	keira	20	15	20	.	15	70	17.50	20	5	1
8	tasha	18	14	19	12	20	83	16.60	20	8	0
9	danial	20	15	18	14	19	86	17.20	20	6	0
10	iman	18	15	19	14	20	86	17.20	20	6	0



LIST OF FUNCTIONS

- STD-returns the standard deviation
- SKEWNESS-returns the skewness
- SUMABS-Returns the sum of the absolute values of the nonmissing arguments.
- RANGE-returns the range of values
- ORDINAL-Returns the kth smallest of the missing and nonmissing values.
- LPNORM-Returns the Lp norm of the second argument and subsequent nonmissing arguments
- EUCLID-Returns the Euclidean norm of the nonmissing arguments.
- N-returns the number of nonmissing values
- NMISS-Returns the number of missing numeric values
- MISSING-returns a numeric result that indicates whether the argument contains a missing value
- CSS-returns the corrected sum of squares
- KURTOSIS-returns the kurtosis (or 4th moment)
- CV-returns the coefficient of variation
- MAX-returns the largest value
- MEAN-returns the arithmetic mean (average)
- MIN-returns the smallest value
- STRERR-returns the standard error of the mean
- MEDIAN-Returns the median value.
- SUM-Returns the sum of the nonmissing arguments
- IQR-Returns the interquartile range.
- USS-returns the uncorrected sum of squares
- VAR-returns the variance
- CMISS-Counts the number of missing arguments.
- LARGEST-Returns the kth largest nonmissing value.
- GEOMEAN-Returns the geometric mean.
- SMALLEST-Returns the kth smallest nonmissing value.
- GEOMEANZ-Returns the geometric mean, using zero fuzzing.
- HARMEANZ-Returns the harmonic mean, using zero fuzzing.
- PCTL-Returns the percentile that corresponds to the percentage.
- RMS-Returns the root mean square of the nonmissing arguments.
- HARMEAN-Returns the harmonic mean.
- MAD-Returns the median absolute deviation from the median.
- DIVIDE-returns the result of a division that handles special missing values for ODS output

AFRIZA2019