SAS Project Work

By: Poorva Dixit

Introduction

This portfolio project was carried out using **SAS Studio** to demonstrate fundamental concepts of data management, transformation, and analysis. The project showcases the ability to work with **datasets, queries, tasks, and utilities** in SAS.

Skills Demonstrated in the Project

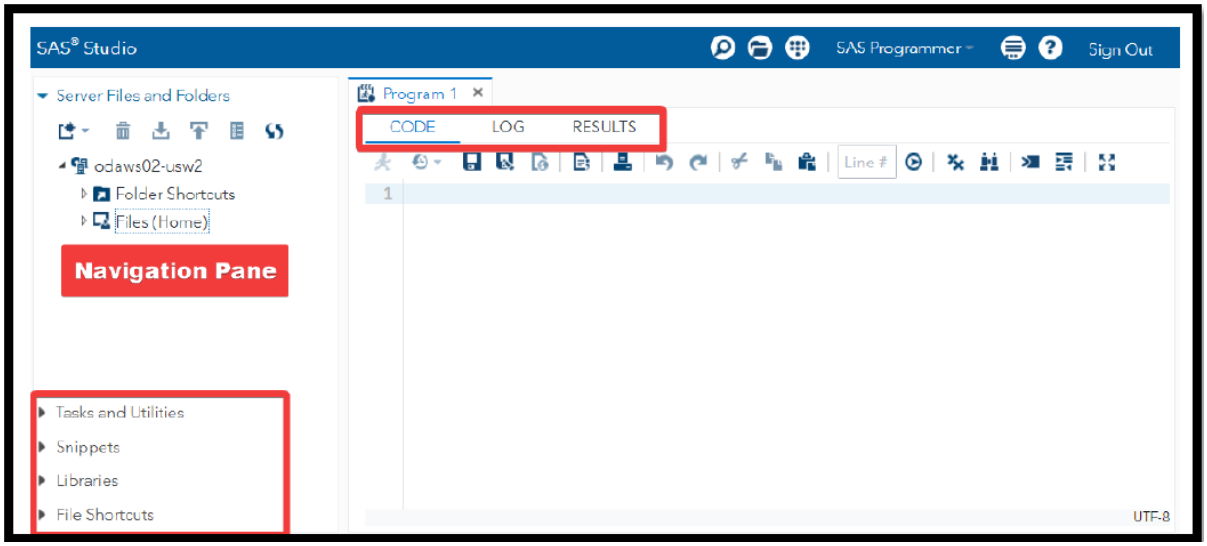
- Data exploration and cleaning using SAS Tasks & Utilities.
- Application of **ETL (Extract, Transform, Load)** processes.
- Performing **data transformations** (sorting, deduplication, aggregation, recoding).
- Creating and managing **temporary and permanent libraries**.
- Handling **missing data and standardizing variables**.
- Conducting **basic statistical analysis** (summary statistics, regressions).
- Generating reports such as **demographic summaries, adverse event listings, and treatment efficacy reports**.

Project Relevance

This project highlights practical **data handling and statistical skills** that are highly relevant for **clinical research, clinical data management, and pharmacovigilance roles**, where accurate data reporting and integrity are crucial.

Category 1

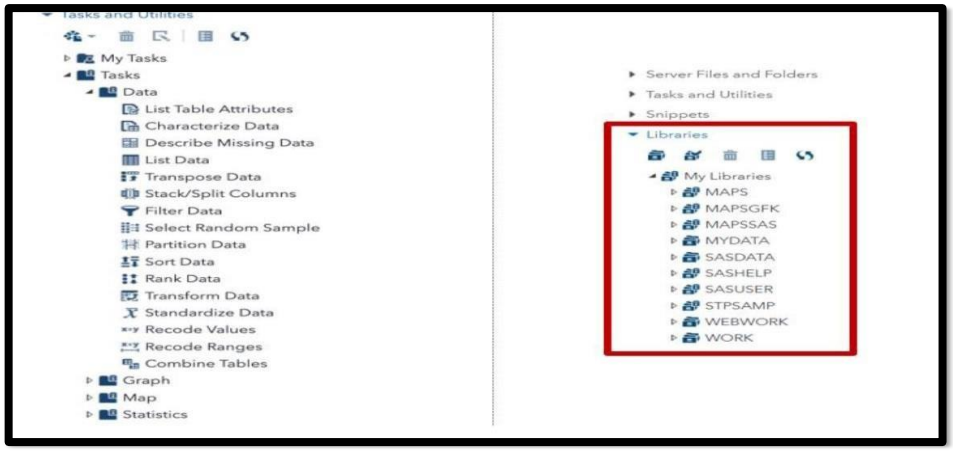
1. Explain the detail in the below mentioned image.



CODE Window

LOG window

Result Tab

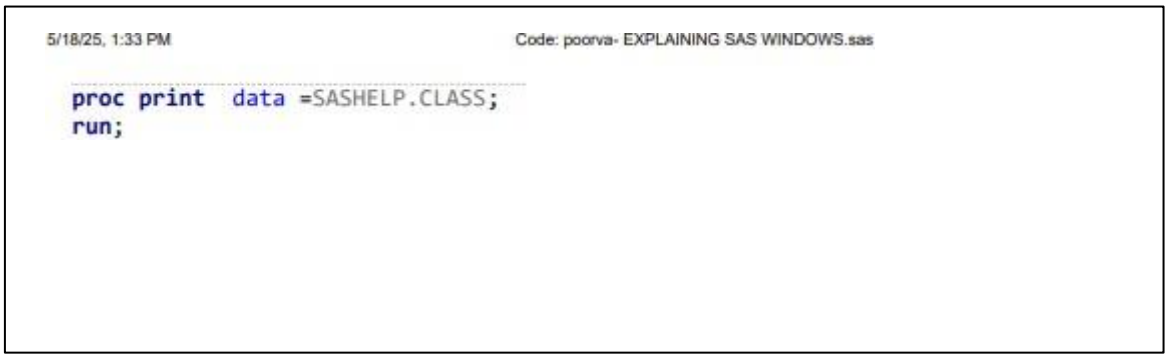


Solu.- 1.**CODE WINDOW:** The Code Window (also called the Program Editor) is the space in SAS Studio where you write and edit your SAS programs (code).

- Allows users to enter SAS code.

- Provides syntax highlighting and autocomplete.
- Can run the code directly from this window.

EXAMPLE: of code window



2. <u>LOG WINDOW:</u> The Log Window displays messages generated by SAS when you run a program. It includes information about the execution of your code.

- Shows notes, warnings, and error messages.
- Helps with debugging and verifying code execution.
- Tracks how much time and memory each step takes.

EXAMPLE: of LOG window

5/18/25, 1:41 PM Log: poorva- EXPLAINING SAS WINDOWS.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc print  data =SASHELP.CLASS;
70         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              1092.21k
      OS Memory           20896.00k
      Timestamp           05/18/2025 07:59:18 AM
      Step Count                        30  Switch Count  0
      Page Faults                       0
      Page Reclaims                     221
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           8


71
72         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
82
```

3. **Results Window:** The Results Window shows the output generated by your SAS program.

Example: of results window

5/18/25, 1:47 PM Results: poorva- EXPLAINING SAS WINDOWS.sas

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

4. **NAVIGATION PANEL**: The Navigation Panel is the left-hand sidebar in SAS Studio. It allows users to access files, libraries, tasks and utilities, snippets.

- Provides quick access to important SAS Studio components.
- Organizes tools and resources needed to create, manage, and analyze data.

1. **TASK AND UTILITIES**: Tasks and Utilities are pre-built templates that help perform common data tasks without writing code.

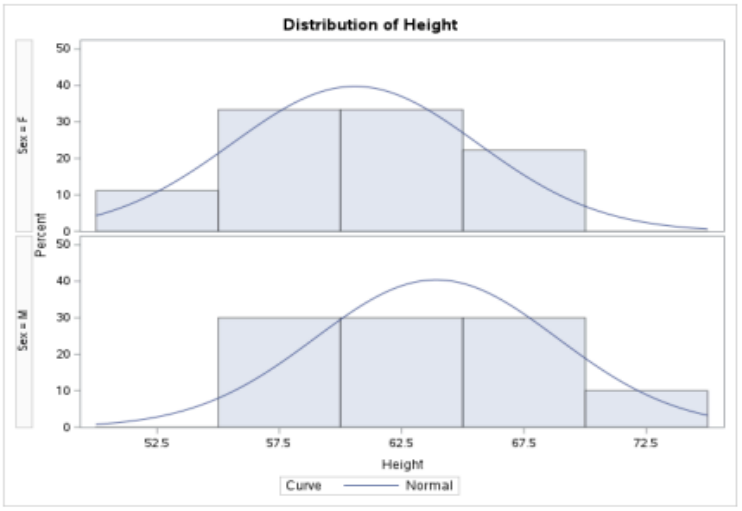
- Automatically generate the underlying SAS code for learning or reuse.
- Includes tasks like: Summary Statistics, Regression etc.

Example: summary statistics

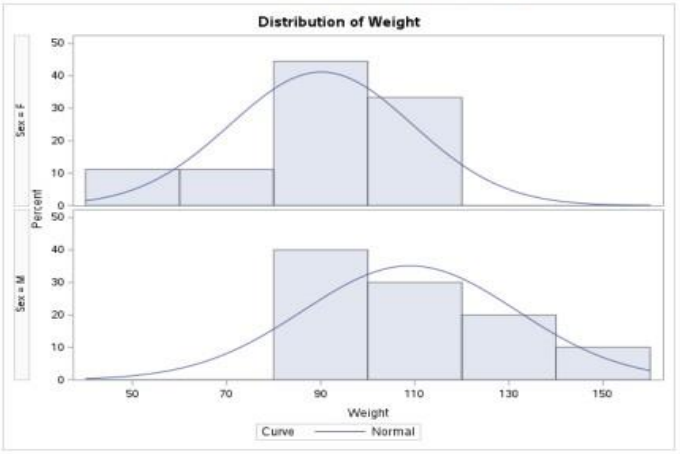
“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

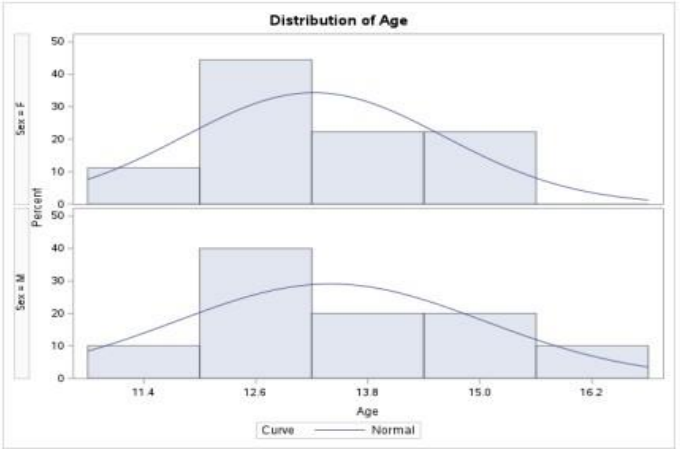
5/18/25, 4:44 PM Results: poorva -Summary Statistics.ctk

Sex	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	N	Std Error	Variance	Mode	Range	Coeff of Variation	Skewness
F	9	Height	60.5888889	5.0183275	51.3000000	66.5000000	9	1.6727758	25.1836111	.	15.2000000	8.2825871	-0.7238643
		Weight	90.1111111	19.3839137	50.5000000	112.5000000	9	6.4613046	375.7361111	.	62.0000000	21.5111250	-0.8982645
		Age	13.2222222	1.3944334	11.0000000	15.0000000	9	0.4648111	1.9444444	12.0000000	4.0000000	10.5461348	-0.1463545
M	10	Height	63.9100000	4.9379370	57.3000000	72.0000000	10	1.5615128	24.3832222	.	14.7000000	7.7263919	0.0409592
		Weight	108.9500000	22.7271864	83.0000000	150.0000000	10	7.1869674	516.5250000	.	67.0000000	20.8601986	0.4994464
		Age	13.4000000	1.6465452	11.0000000	16.0000000	10	0.5206833	2.7111111	12.0000000	5.0000000	12.2876508	0.1269424



5/18/25, 4:44 PM Results: poorva -Summary Statistics.ctk





5/18/25, 4:42 PM Code: poorva -Summary Statistics.ctk

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/18/25, 4:40 PM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS02-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=S
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc means data=SASHELP.CLASS chartype mean std min max n stderr var mode range
        vardef=df cv skewness;
    var Height Weight Age;
    class Sex;
run;

proc univariate data=SASHELP.CLASS vardef=df noprint;
    var Height Weight Age;
    class Sex;
    histogram Height Weight Age / normal(noprint);
run;
```

5/18/25, 4:43 PM Log: poorva -Summary Statistics.ctk

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/18/25, 4:40 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS02-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-44558-b5qVSZBJq191fmrIbqL
80       ! v-cas'
81          *
82          */
83
84         ods noproctitle;
85         ods graphics / imagemap=on;
86
87         proc means data=SASHELP.CLASS chartype mean std min max n stderr var mode range
88         vardef=df cv skewness;
89         var Height Weight Age;
90         class Sex;
91         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.03 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              8542.43k
      OS Memory           30136.00k
      Timestamp           05/18/2025 11:12:46 AM
      Step Count                        45  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1846
      Page Swaps                        0
      Voluntary Context Switches        20
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           8


92
93         proc univariate data=SASHELP.CLASS vardef=df noprint;
94         var Height Weight Age;
95         class Sex;
96         histogram Height Weight Age / normal(noprint);
97         run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.58 seconds
      user cpu time       0.26 seconds
      system cpu time     0.03 seconds
      memory              9600.71k
      OS Memory           31272.00k
      Timestamp           05/18/2025 11:12:47 AM
      Step Count                        46  Switch Count  0
      Page Faults                       0
      Page Reclaims                     2345
      Page Swaps                        0
      Voluntary Context Switches        7820
      Involuntary Context Switches      6
      Block Input Operations            0
      Block Output Operations           2224


98
99         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
100
```

2. Snippets: Snippets are small, reusable pieces of SAS code stored in a categorized list. Saves time and ensures syntax accuracy.

   - Includes categories like Data, PROC, Macro, etc.

   Example:

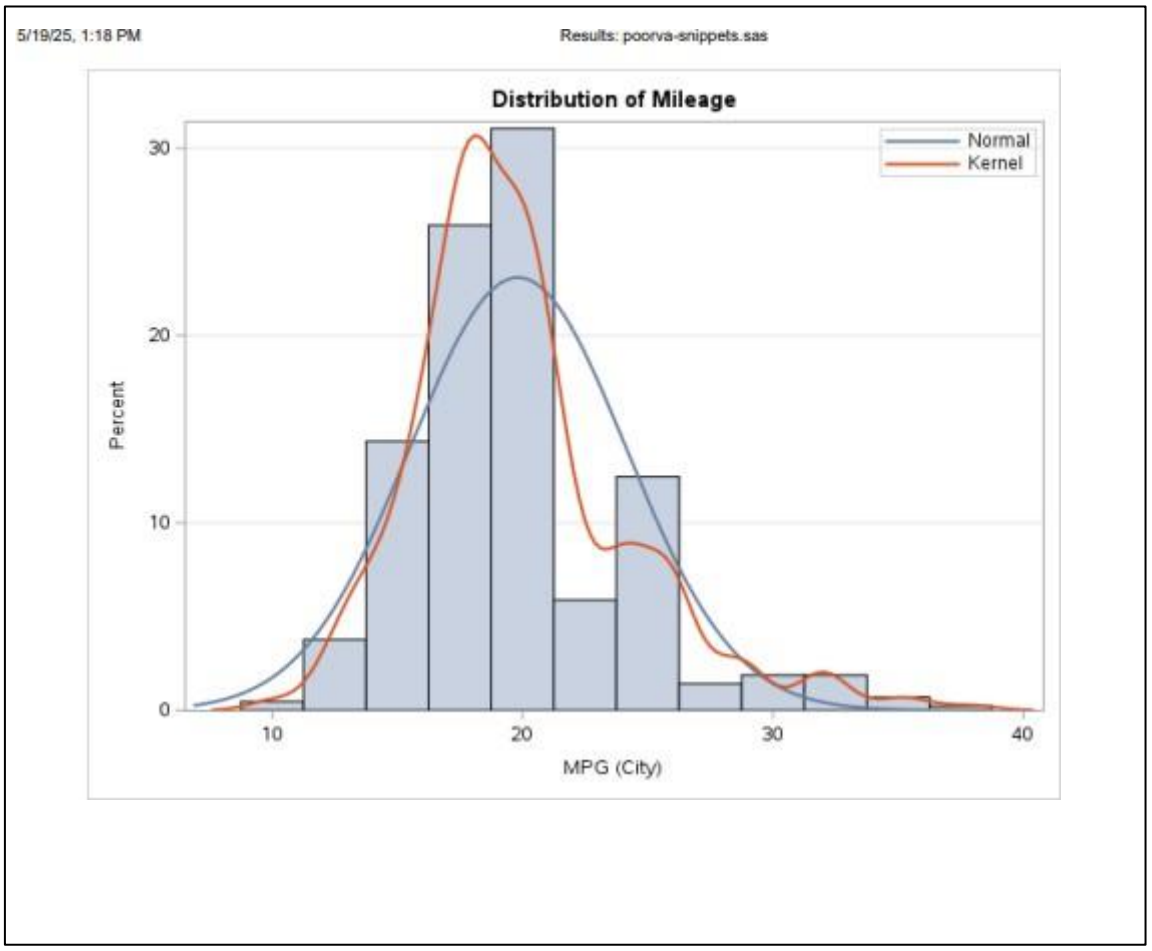
   Clicking on Snippets > Data > Create a New Data Set inserts:

5/19/25, 12:59 PM Code: poorva-snippets.sas

```
/*--Histogram--*/

title 'Distribution of Mileage';
proc sgplot data=sashelp.cars(where=(type ne 'Hybrid'));
  histogram mpg_city;
  density mpg_city / lineattrs=(pattern=solid);
  density mpg_city / type=kernel lineattrs=(pattern=solid);
  keylegend / location=inside position=topright across=1;
  yaxis offsetmin=0 grid;
run;
```

5/19/25, 1:16 PM Log: poorva-snippets.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         /*--Histogram--*/
70
71         title 'Distribution of Mileage';
72         proc sgplot data=sashelp.cars(where=(type ne 'Hybrid'));
73           histogram mpg_city;
74           density mpg_city / lineattrs=(pattern=solid);
75           density mpg_city / type=kernel lineattrs=(pattern=solid);
76           keylegend / location=inside position=topright across=1;
77           yaxis offsetmin=0 grid;
78         run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.15 seconds
      user cpu time       0.04 seconds
      system cpu time     0.00 seconds
      memory              9132.93k
      OS Memory           30000.00k
      Timestamp           05/19/2025 07:46:05 AM
      Step Count                        42  Switch Count  3
      Page Faults                       0
      Page Reclaims                     2069
      Page Swaps                        0
      Voluntary Context Switches        210
      Involuntary Context Switches      4
      Block Input Operations            0
      Block Output Operations           816

NOTE: There were 425 observations read from the data set SASHELP.CARS.
      WHERE type not = 'Hybrid';

79
80
81
82         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
92
```



3. Libraries: A Library in SAS is a collection of SAS datasets. In SAS Studio, Libraries are listed under the Navigation Panel.
   - Allows you to access datasets stored in permanent or temporary locations.
   - Use libraries to reference datasets in your code.
   - **Common Libraries:**
     - WORK (temporary)
     - SASHELP (built-in datasets)
     - MYLIB (user-defined)

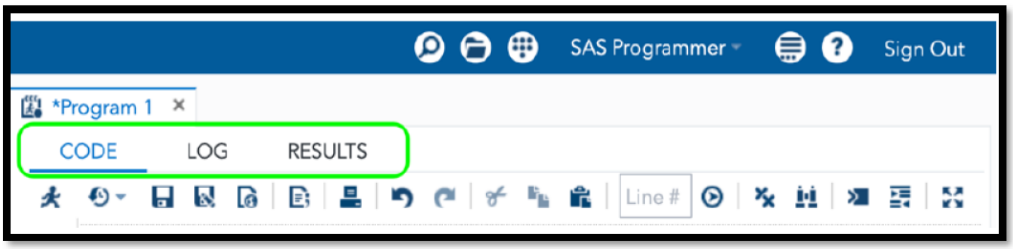
Example:

- Expand **Libraries > SASHELP > CLASS**
- **Use in code:**

```
proc print data=sashelp.class;
run;
```

4. File Shortcuts: File Shortcuts are links to commonly used folders or files on your SAS server or local environment.
   - Quickly access frequently used directories.
   - Simplifies navigation to project folders.

2. Explain the detail explanation on the below mentioned image.



Icon	Tooltip
	Run all or selected codes
	Submission history
	Save program
	Program summary
	Clear all code

Explain in detail on Icons

SOLU. 2: **Toolbar:**

The SAS Studio interface contains a toolbar with the following icons (from left to right):

1. **Running Man Icon** ( ) - "Run all or selected codes": Executes either all code in your program or just the selected portions.
2. **Clock Icon** ( ) - "Submission history": Provides access to previously run code submissions and their results.
3. **Floppy Disk Icon** ( ) - "Save program": Saves your current SAS program to your workspace.
4. **Save As Icon** ( ) - Saves the current program with a new name or to a different location.
5. **Document with Magnifying Glass Icon** ( ) - "Program summary": Shows an overview of your current SAS program.
6. **New File Icon** ( ) - Creates a new blank program or file.

7. **Upload Icon** ( ) - Uploads files from your local computer to the SAS environment.
8. **Undo/Redo Icons** (/) - Navigate through your editing history.
9. **Cut/Copy/Paste Icons** - Standard text editing functions.
10. **Code Block Icon** ( ) - Manages code blocks and sections.
11. **Line Number Toggle** - Shows or hides line numbers in your code editor.
12. **Search Icon** ( ) - Finds text within your code.
13. **Broom Icon** (☐) - "Clear all code": Removes all code from the current editor window.
14. **Format/Indent Icons** - Adjusts code formatting and indentation.
15. **View Adjustment Controls** - Controls for split screen and other viewing options.
16. **Fullscreen Toggle** ( ) - Expands the editor to fullscreen mode.

Navigation Bar:

The top blue navigation bar of the SAS Programmer interface includes:

1. **Search Button** ( ) - Searches across the SAS environment.
2. **Home Button** ( ) - Returns to the main dashboard.
3. **Apps Grid Menu** (⋮) - Accesses other SAS applications and tools.
4. **"SAS Programmer" Dropdown** - Shows current application and allows switching between applications.
5. **Server Settings Icon** (≡) - Manages server connections and settings.
6. **Help Button** (?) - Accesses SAS documentation and help resources.

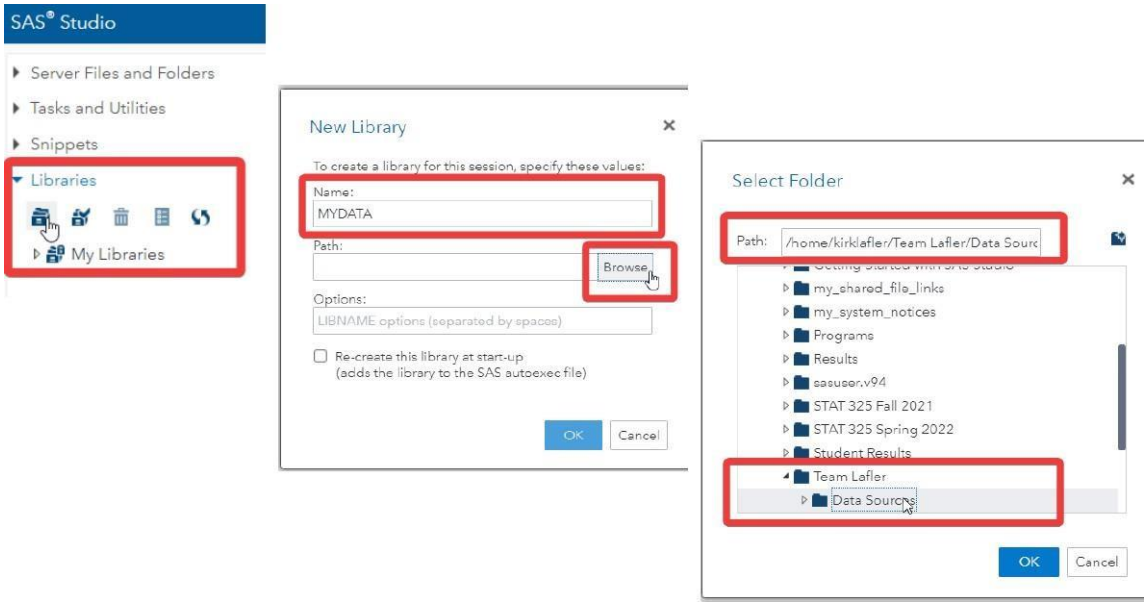
3. What is the concept called as Temporary versus Permanent SAS Datasets?

Solu.3: **Temporary SAS Datasets:** Datasets that exist only for the duration of the current SAS session.

- Stored in the WORK library.
- Automatically deleted when the SAS session ends.
- If no library is specified, SAS assumes the dataset is in the WORK library.

  Permanent SAS Datasets: Datasets that remain available even after the SAS session ends.
- stored in a **user-defined library** that points to a permanent storage location (e.g., a folder on your computer or network).
- Must be manually deleted if no longer needed.
- Requires a two-level name: libref.dataset_name

4. Explain about Assigning a New SAS Library? What is the main function of Maintain the SAS Library?



Solun.4: A SAS library is a collection of SAS files stored in the same physical location like SAS datasets.

- Libraries serve as the fundamental organizational structure for SAS data storage and access.
- Assigning a library creates a logical reference (called a libref) that points to a physical storage location.
- This allows us to access files in that location without needing to specify the full file path every time.

Main function:

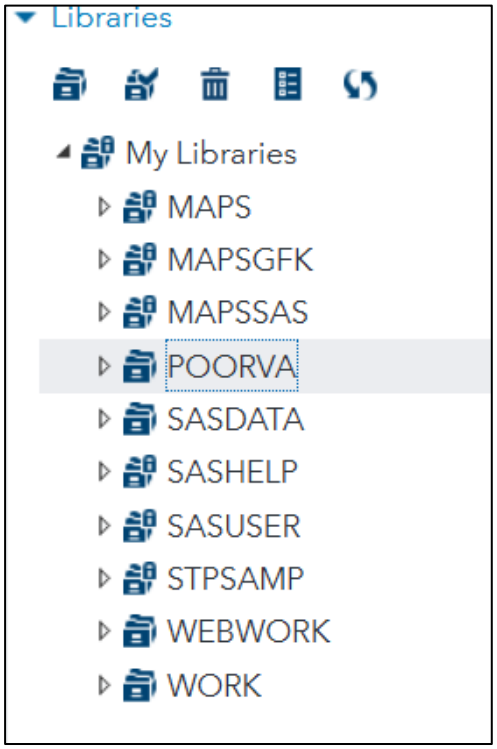
- View and edit library properties and access settings
- Update file paths when data locations change
- Modify library options and parameters
- Remove libraries that are no longer needed
- Troubleshoot library access issues
- Refresh libraries to ensure SAS sees current content
- Toolbar:-
- New Library 🗀 - The first icon (folder) allows you to create a new SAS library by defining a libref and path
-

- Assign Library - The second icon (building-like) lets you assign an existing library or reconnect to a previously defined library

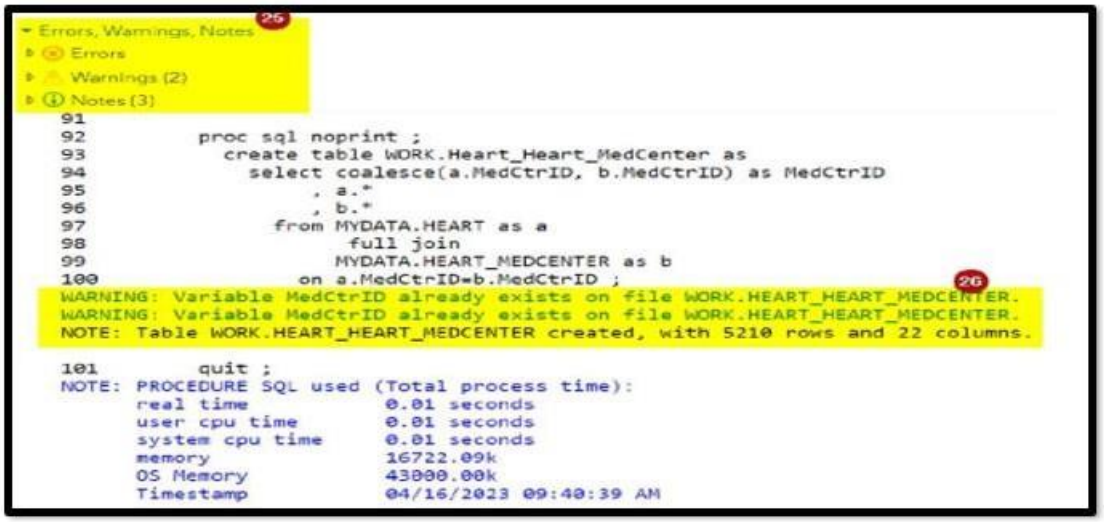
- Delete Library - The trash can icon allows you to remove a library assignment (this doesn't delete the actual data files, just removes the reference)

- Show Details - The document/table icon displays detailed information about the selected library, including its path, engine, and contents

- Refresh - The circular arrow refreshes the library view to reflect any changes made to the library's contents since it was last accessed



Category 2



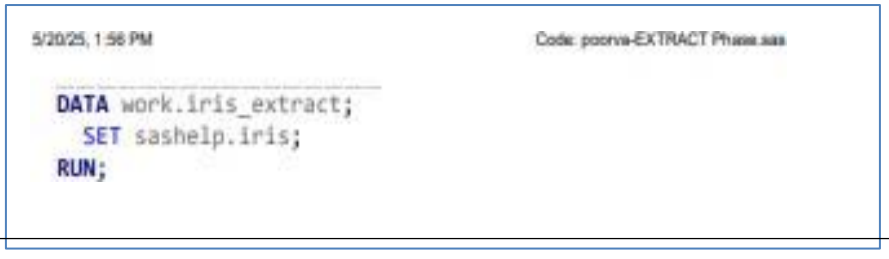
1. Explain about the ETL Process and Define it. Provide a detailing on Phases of ETL.

   SOLU:- ETL stands for Extract, Transform, Load. ETL is the process of Extracting data from various sources, Transforming it into a suitable format or structure, and Loading it into a target data system like a database.

   Below are the three major ETL phases in SAS Studio:

1. EXTRACT Phase

   Example -Here, we extract (copy) data from the built-in dataset SASHELP.IRIS.



This creates a temporary copy work.iris_extract for further transformation.

2. TRANSFORM Phase: Involves transforming the data. Example: standardizing the species name'Setosta' to 'group 1' and 'Versicolor' to 'group 2' in group column.

5/20/25, 2:14 PM Results: WORK.IRIS_TRANSFORM

Obs	Species	SepalLength	SepalWidth	PetalLength	PetalWidth	PetalArea	Group
1	Setosa	50	33	14	2	28	Group 1
2	Setosa	46	34	14	3	42	Group 1
3	Setosa	46	36	10	2	20	Group 1
4	Setosa	51	33	17	5	85	Group 1
5	Setosa	55	35	13	2	26	Group 1
6	Setosa	48	31	16	2	32	Group 1
7	Setosa	52	34	14	2	28	Group 1
8	Setosa	49	36	14	1	14	Group 1
9	Setosa	44	32	13	2	26	Group 1
10	Setosa	50	35	16	6	96	Group 1
11	Setosa	44	30	13	2	26	Group 1
12	Setosa	47	32	16	2	32	Group 1
13	Setosa	48	30	14	3	42	Group 1
14	Setosa	51	38	16	2	32	Group 1
15	Setosa	48	34	19	2	38	Group 1
16	Setosa	50	30	16	2	32	Group 1
17	Setosa	50	32	12	2	24	Group 1
18	Setosa	43	30	11	1	11	Group 1
19	Setosa	58	40	12	2	24	Group 1
20	Setosa	51	38	19	4	76	Group 1
21	Setosa	49	30	14	2	28	Group 1
22	Setosa	51	35	14	2	28	Group 1
23	Setosa	50	34	16	4	64	Group 1
24	Setosa	46	32	14	2	28	Group 1
25	Setosa	57	44	15	4	60	Group 1
26	Setosa	50	36	14	2	28	Group 1
27	Setosa	54	34	15	4	60	Group 1
28	Setosa	52	41	15	1	15	Group 1
29	Setosa	55	42	14	2	28	Group 1
30	Setosa	49	31	15	2	30	Group 1
31	Setosa	54	39	17	4	68	Group 1
32	Setosa	50	34	15	2	30	Group 1
33	Setosa	44	29	14	2	28	Group 1
34	Setosa	47	32	13	2	26	Group 1
35	Setosa	46	31	15	2	30	Group 1
36	Setosa	51	34	15	2	30	Group 1
37	Setosa	50	35	13	3	39	Group 1
38	Setosa	49	31	15	1	15	Group 1
39	Setosa	54	37	15	2	30	Group 1
40	Setosa	54	39	13	4	52	Group 1
41	Setosa	51	35	14	3	42	Group 1
42	Setosa	48	34	16	2	32	Group 1
43	Setosa	48	30	14	1	14	Group 1
44	Setosa	45	23	13	3	39	Group 1
45	Setosa	57	38	17	3	51	Group 1
46	Setosa	51	38	15	3	45	Group 1
47	Setosa	54	34	17	2	34	Group 1
48	Setosa	51	37	15	4	60	Group 1
49	Setosa	52	35	15	2	30	Group 1
50	Setosa	53	37	15	2	30	Group 1
51	Versicolor	65	28	46	15	690	Group 2
52	Versicolor	62	22	45	15	675	Group 2

5/20/25, 2:21 PM Code: poorva-TRANSFORM Phase.sas

```
DATA work.iris_transform;
  SET work.iris_extract;

  /* Create a petal area column */
  PetalArea = PetalLength * PetalWidth;

  /* Standardize species name */
  IF Species = 'Setosa' THEN Group = 'Group 1';
  ELSE IF Species = 'Versicolor' THEN Group = 'Group 2';
  ELSE Group = 'Group 3';
RUN;
```

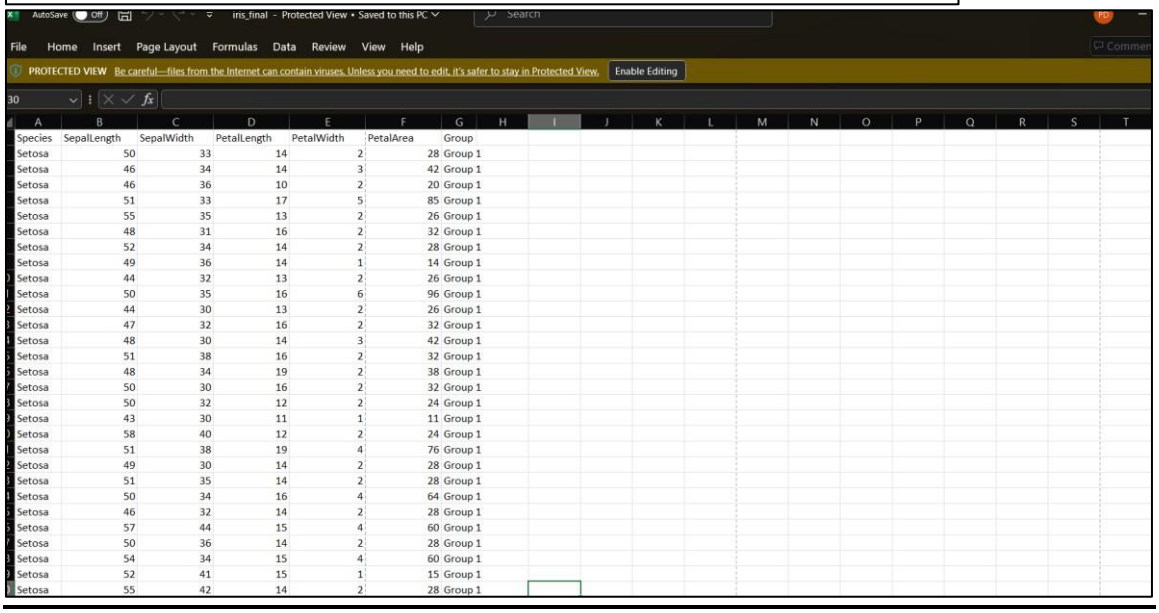
3. LOAD-Phase: Involves Loading the transformed data into a permanent library or export it.
Example- exporting the Work.iris_FINAL to excel.

5/20/25, 2:53 PM Code: poorva-LOAD PHASE.sas

```
/* Final output in WORK library */
DATA work.iris_final;
  SET work.iris_transform;
RUN;
```

5/20/25, 3:15 PM Code: poorva- LOAD AND EXPORT TO EXCEL.sas

```
/* Export to Excel (.xlsx) */
PROC EXPORT DATA=work.IRIS_FINAL
  OUTFILE="/home/u64186191/iris_final.xlsx"
  DBMS=XLSX
  REPLACE;
RUN;
```



iris_final - Protected View • Saved to this PC

File Home Insert Page Layout Formulas Data Review View Help

PROTECTED VIEW Be careful—files from the Internet can contain viruses. Unless you need to edit, it's safer to stay in Protected View. Enable Editing

Species	SepalLength	SepalWidth	PetalLength	PetalWidth	PetalArea	Group
Setosa	50	33	14	2	28	Group 1
Setosa	46	34	14	3	42	Group 1
Setosa	46	36	10	2	20	Group 1
Setosa	51	33	17	5	85	Group 1
Setosa	55	35	13	2	26	Group 1
Setosa	48	31	16	2	32	Group 1
Setosa	52	34	14	2	28	Group 1
Setosa	49	36	14	1	14	Group 1
Setosa	44	32	13	2	26	Group 1
Setosa	50	35	16	6	96	Group 1
Setosa	44	30	13	2	26	Group 1
Setosa	47	32	16	2	32	Group 1
Setosa	48	30	14	3	42	Group 1
Setosa	51	38	16	2	32	Group 1
Setosa	48	34	19	2	38	Group 1
Setosa	50	30	16	2	32	Group 1
Setosa	50	32	12	2	24	Group 1
Setosa	43	30	11	1	11	Group 1
Setosa	58	40	12	2	24	Group 1
Setosa	51	38	19	4	76	Group 1
Setosa	49	30	14	2	28	Group 1
Setosa	51	35	14	2	28	Group 1
Setosa	50	34	16	4	64	Group 1
Setosa	46	32	14	2	28	Group 1
Setosa	57	44	15	4	60	Group 1
Setosa	50	36	14	2	28	Group 1
Setosa	54	34	15	4	60	Group 1
Setosa	52	41	15	1	15	Group 1
Setosa	55	42	14	2	28	Group 1

2. Explain the process of Exploratory Data Analysis (EDA)?

SOLU 2: EDA or Exploratory Data Analysis is a process involving analyzing and gaining information related to data before using it for creating models.

in simpler terms, it's like getting to know your data before you use it for anything serious (like building a model).

You explore the data to understand:

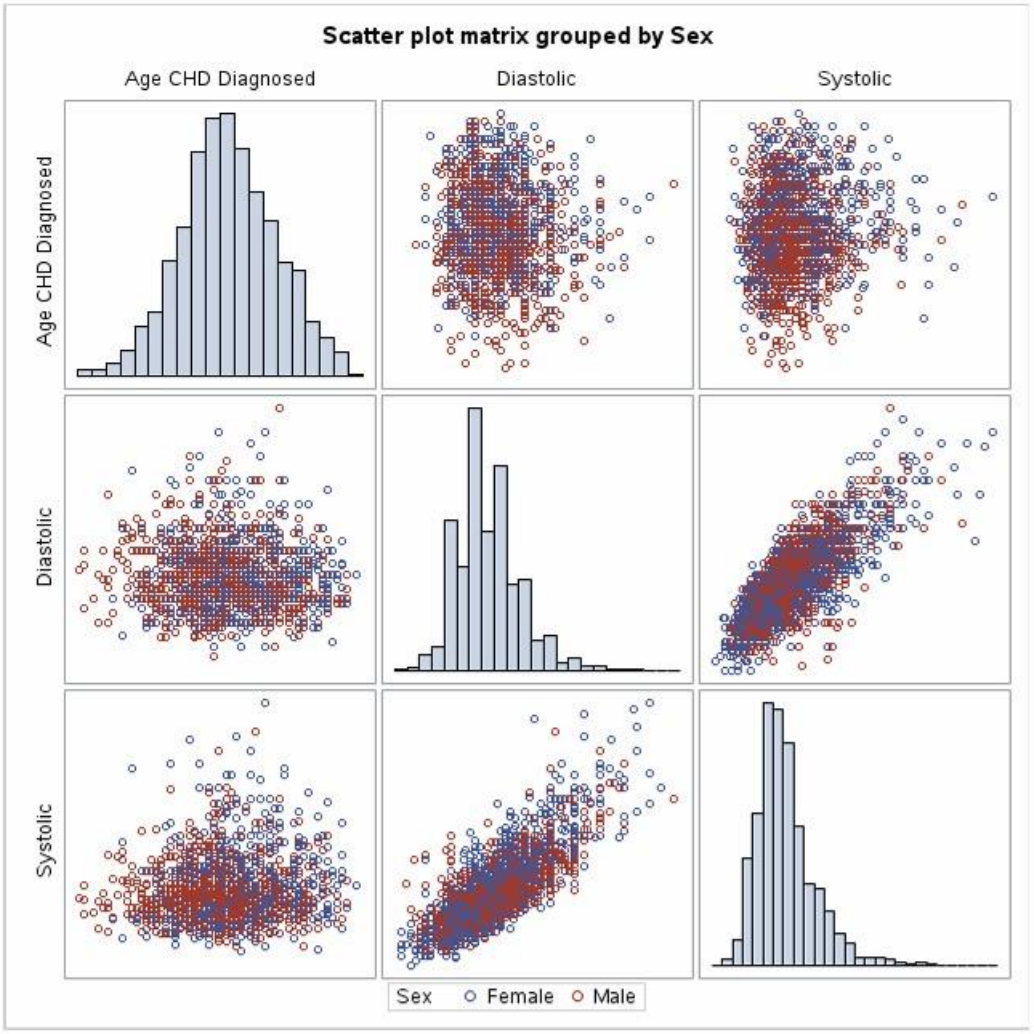
- What kind of data do you have
- If there are missing or strange values
- What patterns or relationships exist
- Basic steps involved in EDA:
- Importing the data
- Viewing the data
- Understanding data types and structures like text,numbers etc.
- Summary statistics,involves getting average,minimum,maximum and other basic stats.
- Checking for any missing values,to see if any data is

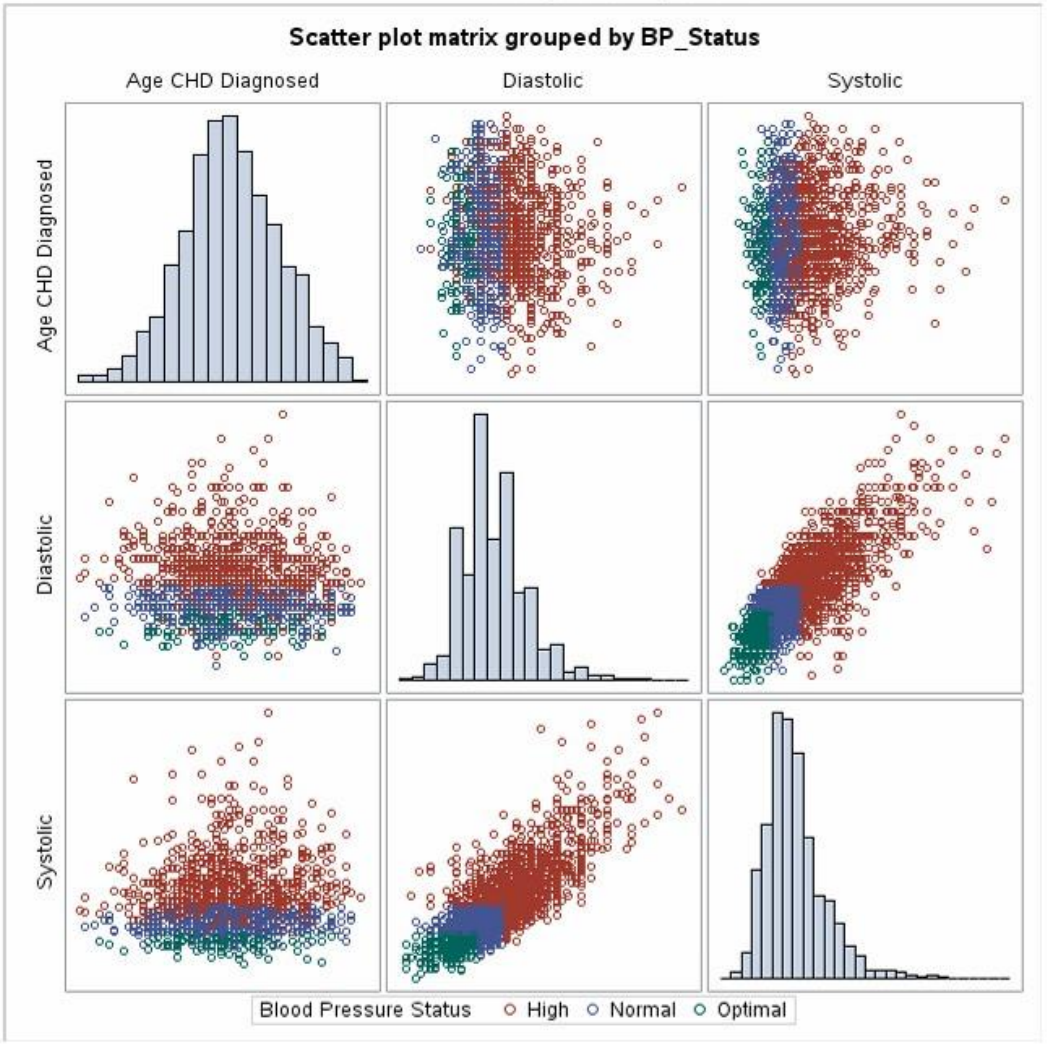
missing in any column.

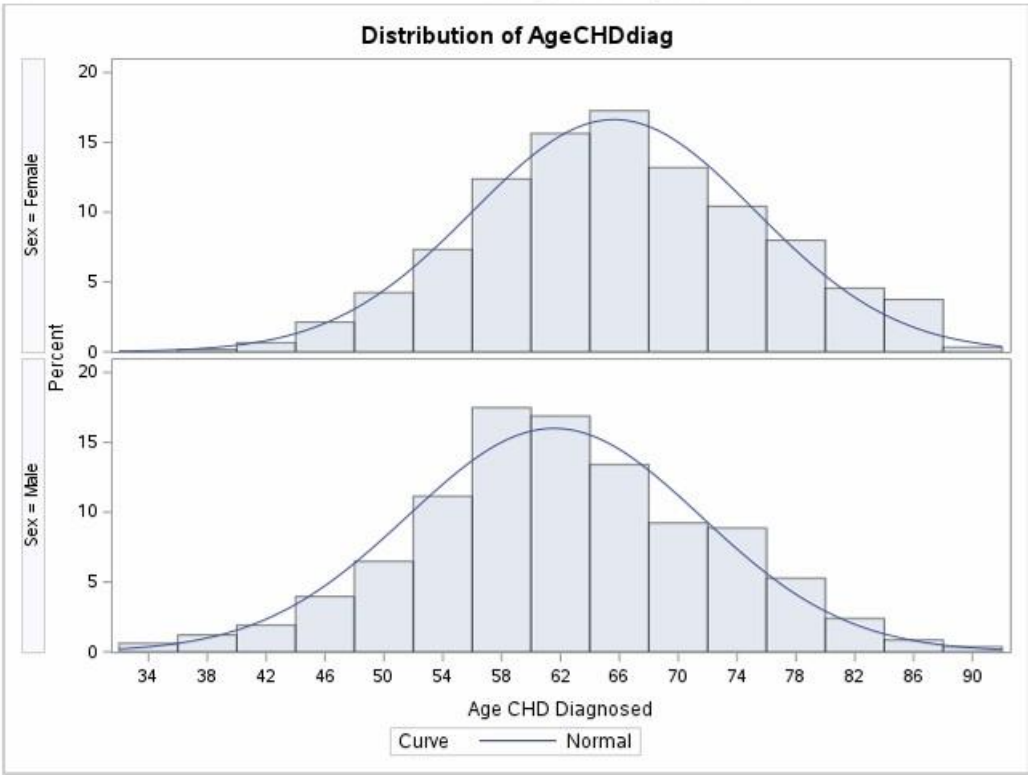
- Visualizing the data ,through graphs like histograms,box plots,or bar charts.

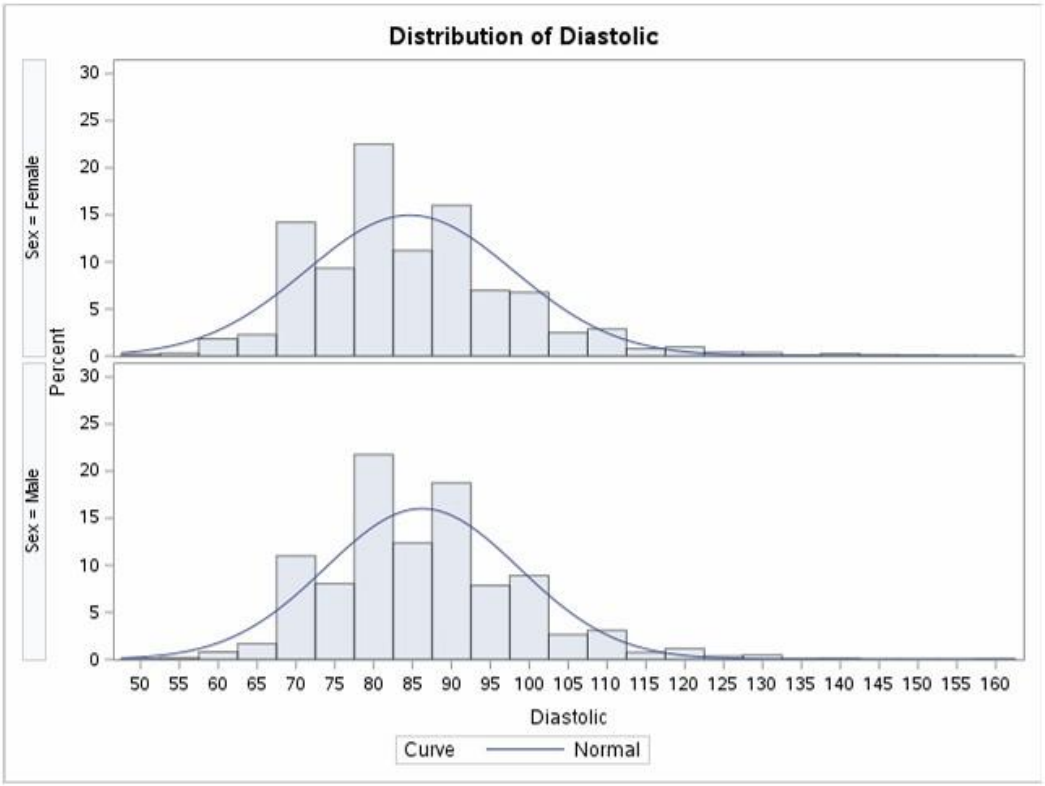
- Analyze and look for patterns or relationships involved,eg- how two columns related(co-relation) ,regression etc.

5/20/25, 5:16 PM Results: poorva-Data Exploration 1.ctk

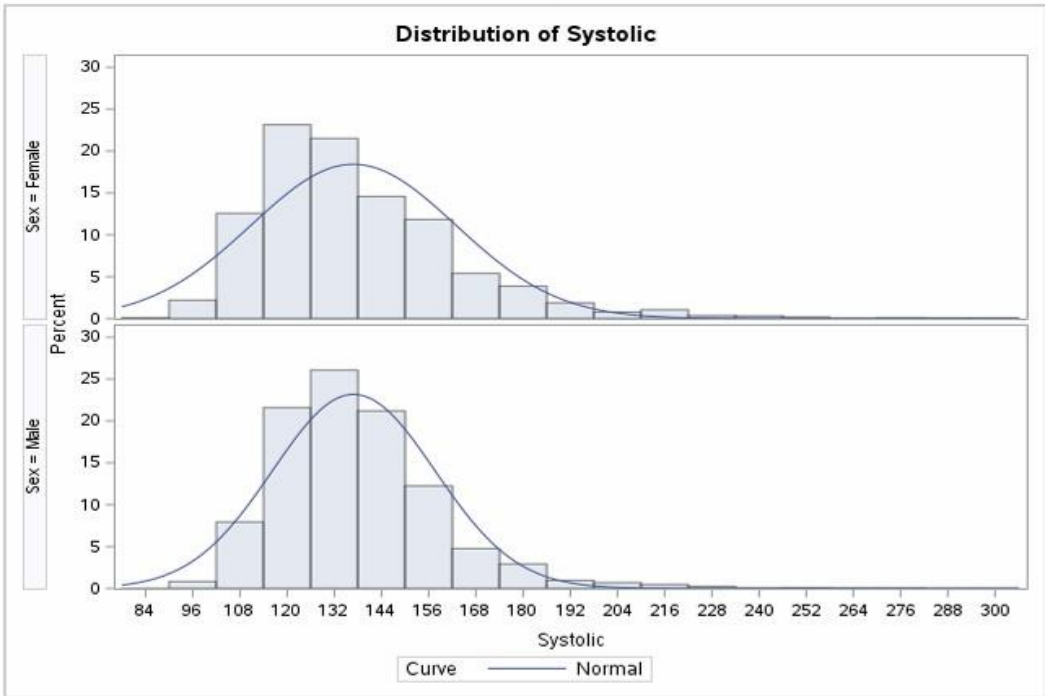


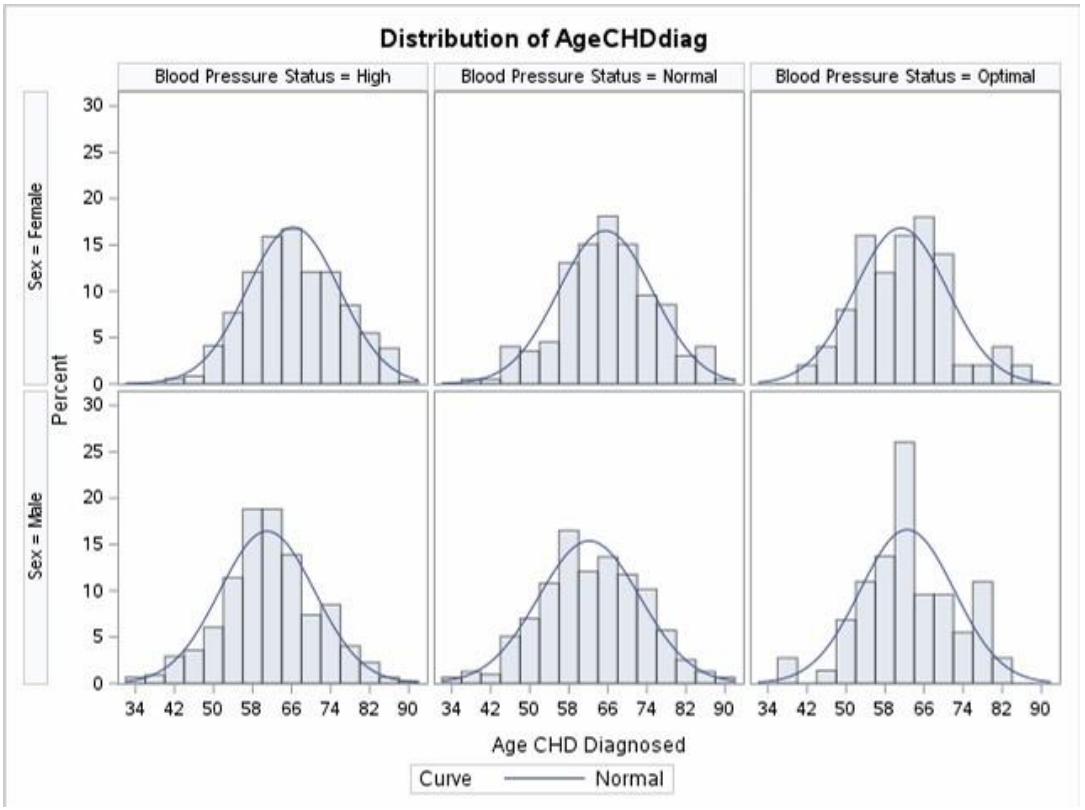

Scatter plot matrix grouped by BP_Status
Age CHD Diagnosed
Diastolic
Systolic
Age CHD Diagnosed
Diastolic
Systolic
Blood Pressure Status
High
Normal
Optimal


Distribution of AgeCHDdiag
Sex = Female
Sex = Male
Percent
0
5
10
15
20
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
Age CHD Diagnosed
Curve
Normal

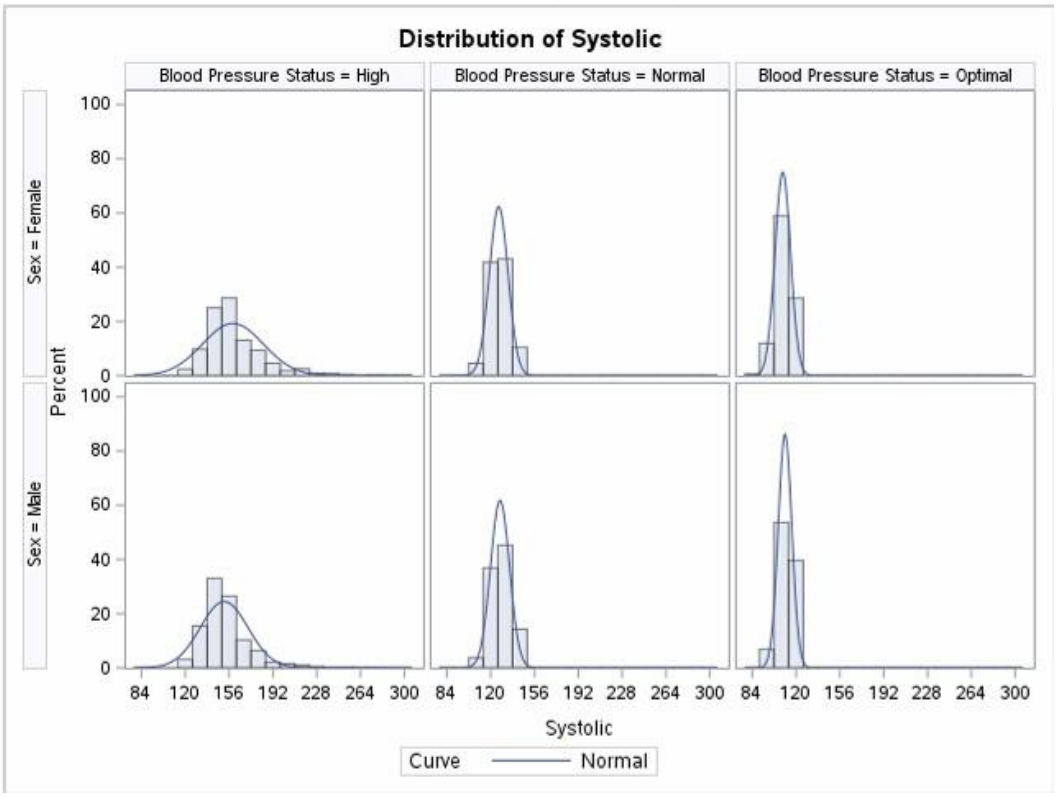

Distribution of Diastolic
Sex = Female
Sex = Male
Percent
0
5
10
15
20
25
30
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
Diastolic
Curve
Normal

“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

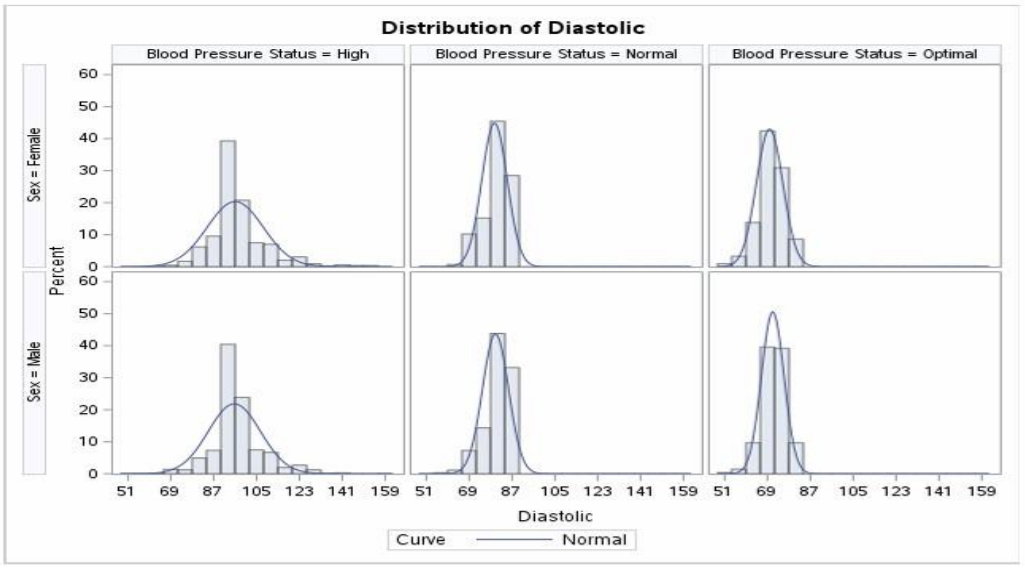




5/20/25, 5:16 PM Results: poorva-Data Exploration 1.ctk



5/20/25, 5:16 PM Results: poorva-Data Exploration 1.ctk



Summary statistics

5/20/25, 5:31 PM Results: Summary Statistics

Sex	Blood Pressure Status	N Obs	Variable	Label	Mean	Std Dev	Minimum	Maximum	N	N Miss
Female	High	1186	AgeCHDdiag	Age CHD Diagnosed	66.3369863	9.4403202	41.0000000	90.0000000	365	821
			Diastolic		96.0590219	11.7743864	60.0000000	155.0000000	1186	0
			Systolic		159.1812816	24.9063397	112.0000000	300.0000000	1186	0
	Normal	1166	AgeCHDdiag	Age CHD Diagnosed	65.5075377	9.6566664	39.0000000	88.0000000	199	967
			Diastolic		79.6415094	5.3357770	60.0000000	88.0000000	1166	0
			Systolic		126.6157804	7.6730323	101.0000000	140.0000000	1166	0
	Optimal	521	AgeCHDdiag	Age CHD Diagnosed	61.2600000	9.4821207	41.0000000	84.0000000	50	471
			Diastolic		69.8675624	5.5677405	50.0000000	78.0000000	521	0
			Systolic		109.1190019	6.3743208	82.0000000	118.0000000	521	0
Male	High	1081	AgeCHDdiag	Age CHD Diagnosed	60.9776286	9.7202779	32.0000000	88.0000000	447	634
			Diastolic		95.6928770	10.9591005	52.0000000	160.0000000	1081	0
			Systolic		151.9546716	19.5936209	115.0000000	276.0000000	1081	0
	Normal	977	AgeCHDdiag	Age CHD Diagnosed	62.1968254	10.3800085	33.0000000	88.0000000	315	662
			Diastolic		80.0532242	5.4754699	54.0000000	88.0000000	977	0
			Systolic		127.7093142	7.7483106	102.0000000	140.0000000	977	0
	Optimal	278	AgeCHDdiag	Age CHD Diagnosed	62.5342466	9.6221420	36.0000000	82.0000000	73	205
			Diastolic		71.1798561	4.7314171	50.0000000	78.0000000	278	0
			Systolic		110.9820144	5.5567015	90.0000000	118.0000000	278	0

5/20/25, 5:32 PM Log: Summary Statistics

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/20/25, 5:30 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&https%3A%2F%2Fodamid-usw2-2.oda.sas
80       ! .com%2FSASStudio%2Findex='
81          *
82          */
83
84         ods noproctitle;
85         ods graphics / imagemap=on;
86
87         proc means data=SASHELP.HEART chartype mean std min max n nmiss vardef=df;
88         var AgeCHDdiag Diastolic Systolic;
89         class Sex BP_Status;
90         run;

NOTE: There were 5209 observations read from the data set SASHELP.HEART.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.05 seconds
      user cpu time       0.05 seconds
      system cpu time     0.01 seconds
      memory              9247.93k
      OS Memory           32952.00k
      Timestamp           05/20/2025 12:00:51 PM
      Step Count                        141  Switch Count  1
      Page Faults                       0
      Page Reclaims                     1995
      Page Swaps                        0
      Voluntary Context Switches        26
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           24


91
92         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
102
```

3. Explain the Task about Describe Missing Data? With Flow Chart Representation? **Create an example?**

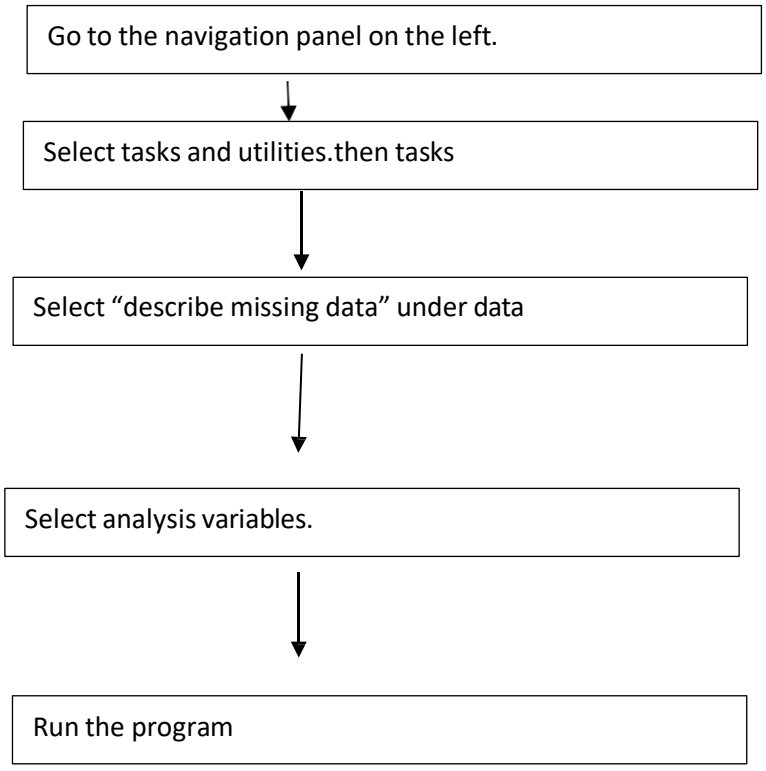
SOLU 3:

- The "Describe Missing Data" task (found under Tasks and Utilities) allows to identify and summarize missing values in a dataset.

- It's to generate a report that shows how many missing values exist for each variable in a dataset. It helps to understand:

- Which variables have missing data.

- The number and percentage of missing values.



“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

5/23/25, 11:52 AM Code: GRADE.sas

```
data GRADE;
input GRADE Name $ Age Gender $ Marks;
    datalines;
101 Asha 20 F 85
102 Raj . M 90
103 Priya 21 . 78
104 Manish 22 M .
105 Sameera . F 88
;
run;
```

5/23/25, 11:54 AM Results: WORK.GRADE

Obs	GRADE	Name	Age	Gender	Marks
1	101	Asha	20	F	85
2	102	Raj	.	M	90
3	103	Priya	21		78
4	104	Manish	22	M	.
5	105	Sameera	.	F	88

5/23/25, 12:00 PM Code: poorva-Describe Missing Data.ctk

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/23/25, 11:57 AM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS02-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=S1
 *
 */

ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
    value $_cmissprint " "="." other="Non-missing";
run;

proc freq data=WORK.GRADE;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format Age Marks _nmissprint.;
    format Gender $_cmissprint.;
    tables Age Gender Marks / missing nocum;
run;

proc freq data=WORK.GRADE noprint;
    table Age * Gender * Marks / missing out=Work._MissingData_;
    format Age Marks _nmissprint.;
    format Gender $_cmissprint.;
run;

proc print data=Work._MissingData_ noobs label;
    title3 "Missing Data Patterns across Variables";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format Age Marks _nmissprint.;
    format Gender $_cmissprint.;
    label count="Frequency" percent="Percent";
run;

title3;

/* Clean up */
proc delete data=Work._MissingData_;
```

5/23/25, 12:01 PM Results: poorva-Describe Missing Data.ctk

Missing Data Frequencies
Legend: ., A, B, etc = Missing

Age	Frequency	Percent
.	2	40.00
Non-missing	3	60.00

Gender	Frequency	Percent
	1	20.00
Non-missing	4	80.00

Marks	Frequency	Percent
.	1	20.00
Non-missing	4	80.00

Missing Data Patterns across Variables
Legend: ., A, B, etc = Missing

Age	Gender	Marks	Frequency	Percent
.	Non-missing	Non-missing	2	40
Non-missing		Non-missing	1	20
Non-missing	Non-missing	.	1	20
Non-missing	Non-missing	Non-missing	1	20

4. Explain the Task about Recode (or Standardize) Values for Gender Specification? **Create an example?**

SOLU 4:

- The task of recoding or standardizing values for gender specification is a process where inconsistent or varied values representing gender are converted into a standardized format.
- this is used for data analysis, reporting, where consistency in data representation is crucial.

5/23/25, 1:09 PM Code: gender.sas

```
data gender;
    input ID Name $ Gender $;
    datalines;
1 Rahul male
2 Sneha F
3 John male
4 Priya F
5 Aman male
;
run;
```

5/23/25, 1:11 PM Results: WORK.GENDER

Obs	ID	Name	Gender
1	1	Rahul	male
2	2	Sneha	F
3	3	John	male
4	4	Priya	F
5	5	Aman	male

5/23/25, 2:38 PM Results: WORK.GENDER

Obs	_recodeVar_	ID	Name	Gender
1	M	1	Rahul	male
2	F	2	Sneha	F
3	M	3	John	male
4	F	4	Priya	F
5	M	5	Aman	male

Here "male" has been recoded to "M". after using recode values from tasks-data .

5. Explain the below mentioned sample with output Results? – **PG NO 19-22 - Combining (Joining) Tables** with All Matching and Non-Matching Rows, continued

SOLU 5:

The Combine Tables task allows you to join (merge) two datasets based on a common column or key. When you choose "All matching and non-matching rows", you're performing a full outer join — which includes every row from both datasets, whether or not they match on the key.

- Matching rows from both datasets are joined.
- Non-matching rows from either dataset are still included.
- Missing values will appear in columns where a match doesn't exist.

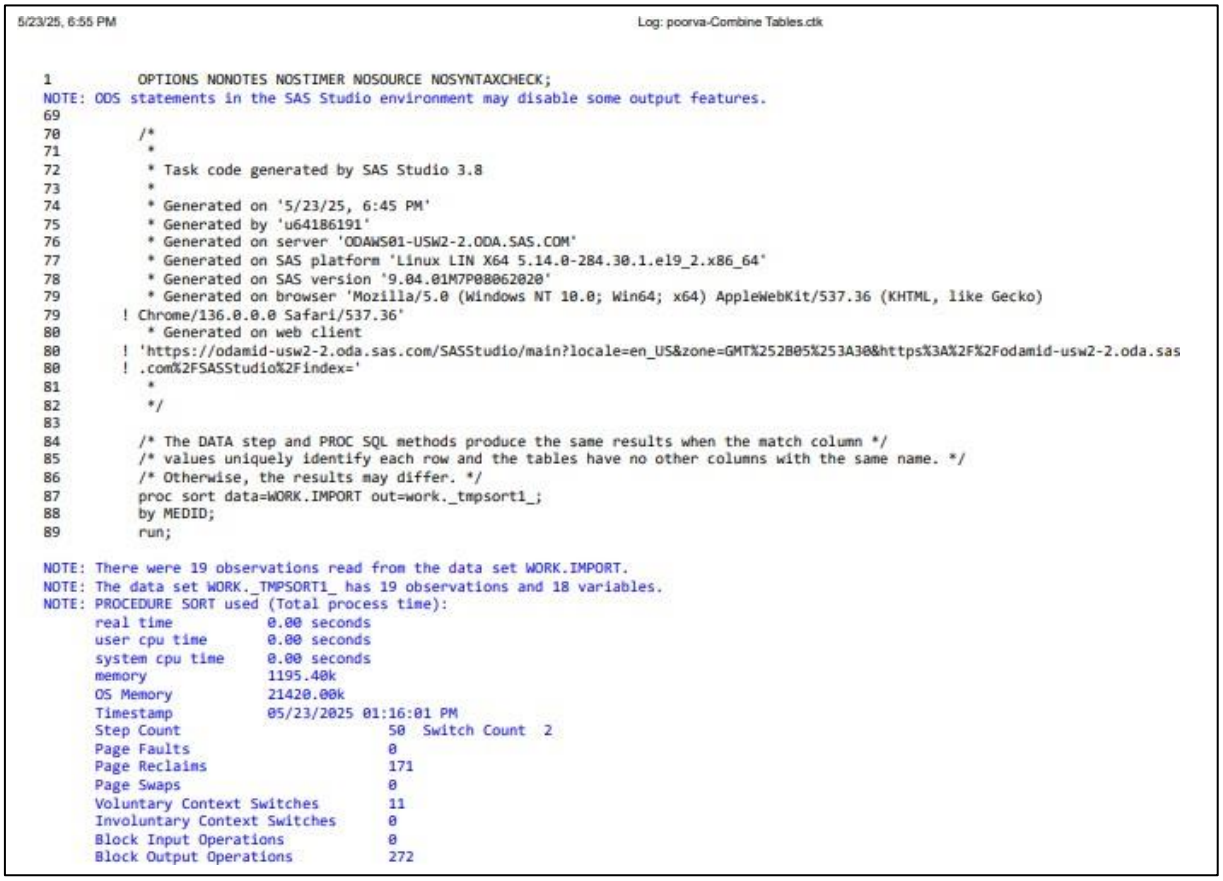
“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

5/23/25, 6:46 PM Code: poorva-Combine Tables.ctk

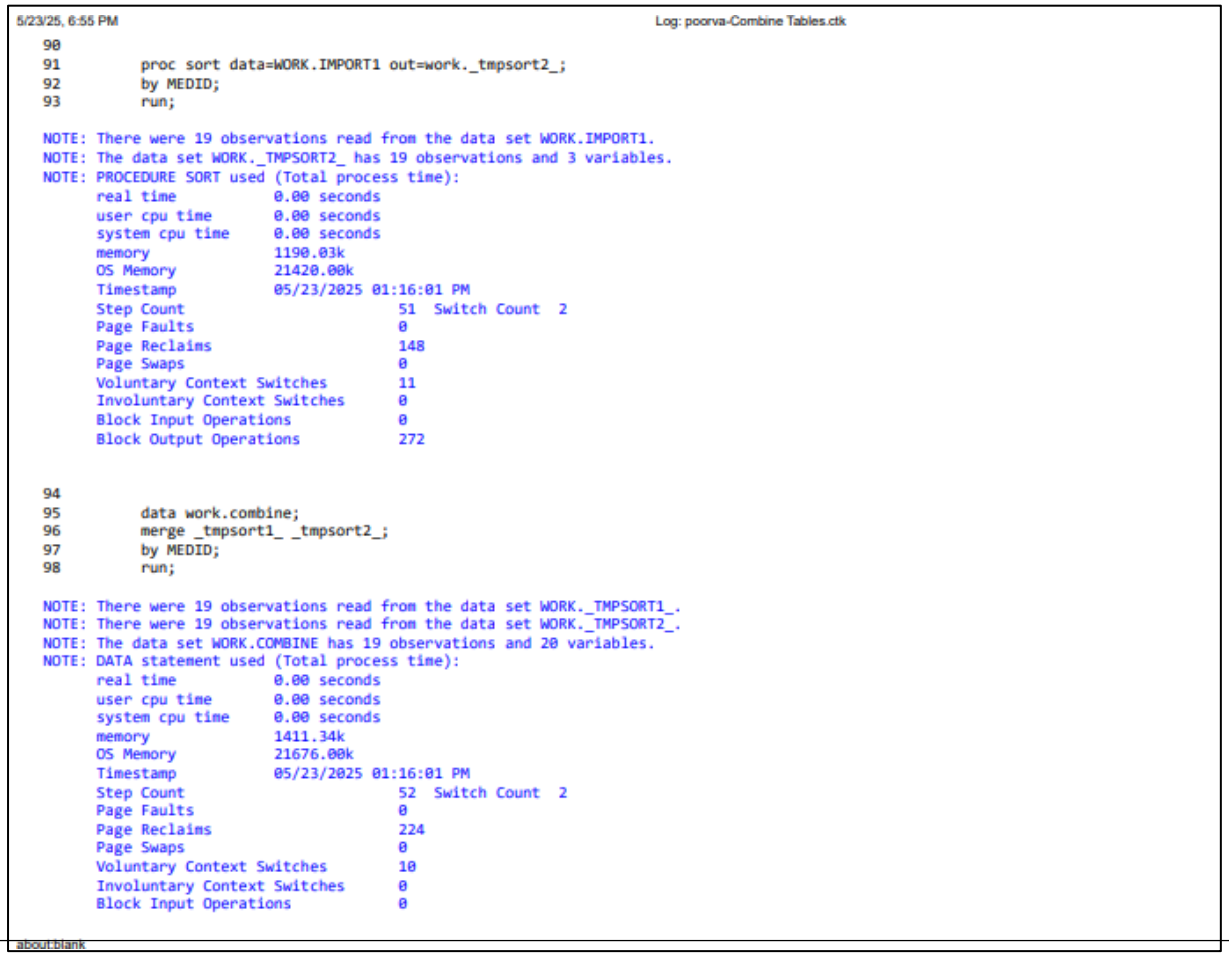
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/23/25, 6:45 PM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&https%3A
 *
 */

/* The DATA step and PROC SQL methods produce the same results when the match column */
/* values uniquely identify each row and the tables have no other columns with the same name. */
/* Otherwise, the results may differ. */
proc sort data=WORK.IMPORT out=work._tmpsort1_;
    by MEDID;
run;

proc sort data=WORK.IMPORT1 out=work._tmpsort2_;
    by MEDID;
run;

data work.combine;
    merge _tmpsort1_ _tmpsort2_;
    by MEDID;
run;

proc delete data=work._tmpsort1_ work._tmpsort2_;
run;
```

5/23/25, 6:51 PM Results: WORK.COMBINE

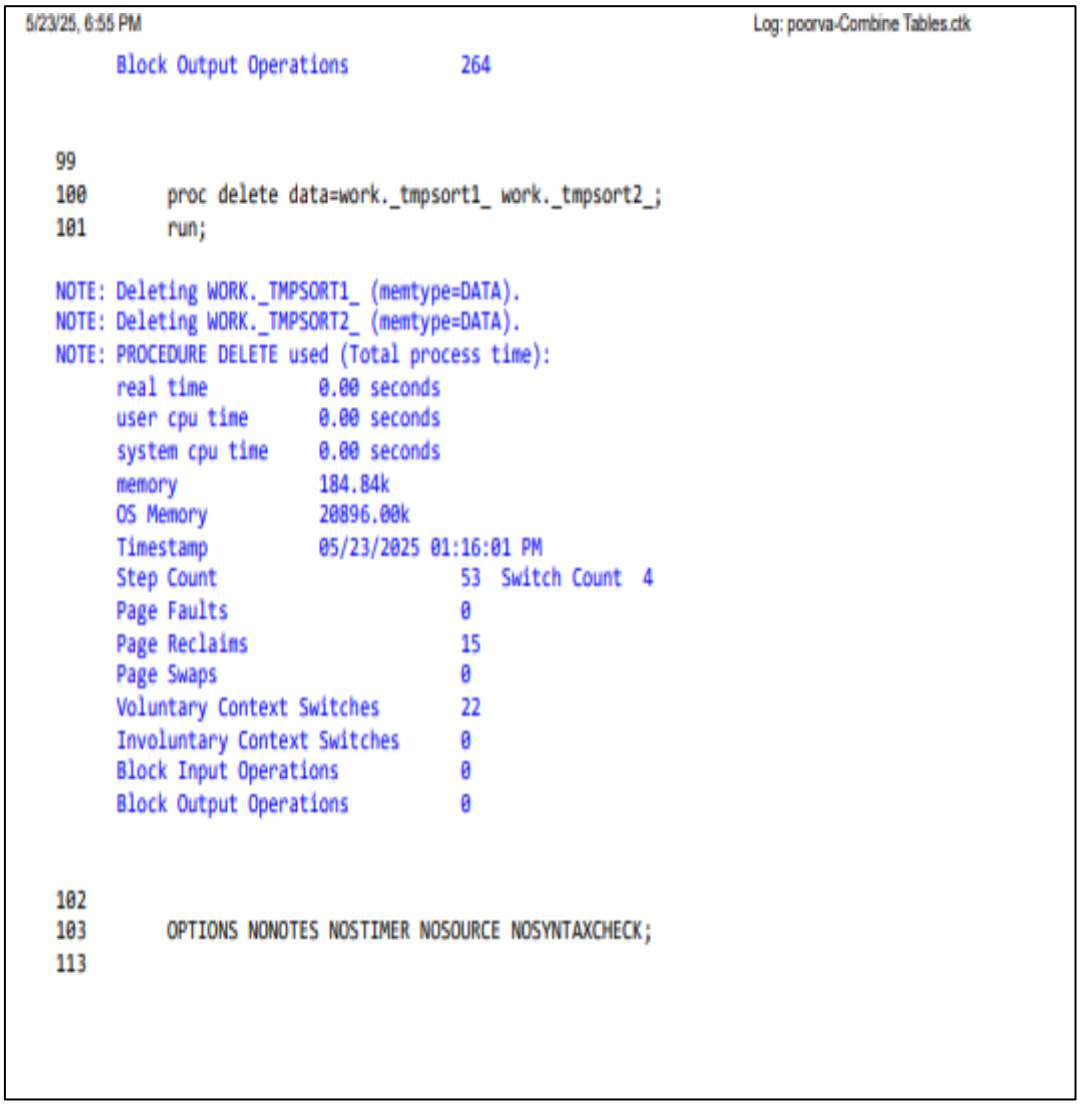
Obs	MEDID	Status	DeathCause	AgeCHDdiag	Sex	AgeAtStart	Height	Weight	Diastolic	Systolic	MRW	Smoking	AgeAtDeath	Cholesterol	Chol_Status	BP_Status	Weight_Status	Smoking_Status	MedicalCenter	City
1	1	Dead	Other	.	Female	29	62.5	140	78	124	121	0	55	.		Normal	Overweight	Non-smoker	San Francisco Medical Center	San Francisco
2	2	Dead	Cancer	.	Female	41	59.75	194	92	144	183	0	57	181	Desirable	High	Overweight	Non-smoker	San Francisco Medical Center	San Francisco
3	3	Alive		.	Female	57	62.25	132	90	170	114	10	.	250	High	High	Overweight	Moderate (6-15)	San Francisco Medical Center	San Francisco
4	4	Alive		.	Female	39	65.75	158	80	128	123	0	.	242	High	Normal	Overweight	Non-smoker	Los Angeles Medical Center	Los Angeles
5	5	Alive		.	Male	42	66	156	76	110	116	20	.	281	High	Optimal	Overweight	Heavy (16-25)	San Francisco Medical Center	San Francisco
6	6	Alive		.	Female	58	61.75	131	92	176	117	0	.	196	Desirable	High	Overweight	Non-smoker	Los Angeles Medical Center	Los Angeles
7	7	Alive		.	Female	36	64.75	136	80	112	110	15	.	196	Desirable	Normal	Overweight	Moderate (6-15)	Los Angeles Medical Center	Los Angeles
8	8	Dead	Other	.	Male	53	65.5	130	80	114	99	0	77	276	High	Normal	Normal	Non-smoker	Las Vegas Health Centre	Las Vegas
9	9	Alive		.	Male	35	71	194	68	132	124	0	.	211	Borderline	Normal	Overweight	Non-smoker	Los Angeles Medical Center	Los Angeles
10	10	Dead	Cerebral Vascular Disease	.	Male	52	62.5	129	78	124	106	5	82	284	High	Normal	Normal	Light (1-5)	Las Vegas Health Centre	Las Vegas
11	11	Alive		.	Male	39	66.25	179	76	128	133	30	.	225	Borderline	Normal	Overweight	Very Heavy (> 25)	San Francisco Medical Center	San Francisco
12	12	Alive		57	Male	33	64.25	151	68	108	118	0	.	221	Borderline	Optimal	Overweight	Non-smoker	Las Vegas Health Centre	Las Vegas
13	13	Alive		55	Male	33	70	174	90	142	114	0	.	188	Desirable	High	Overweight	Non-smoker	Las Vegas Health Centre	Las Vegas
14	14	Alive		79	Male	57	67.25	165	76	128	118	15	.	.		Normal	Overweight	Moderate (6-15)	San Francisco Medical Center	San Francisco
15	15	Alive		66	Male	44	69	155	90	130	105	30	.	292	High	High	Normal	Very Heavy (> 25)	San Francisco Medical Center	San Francisco
16	16	Alive		.	Female	37	64.5	134	76	120	108	10	.	196	Desirable	Normal	Normal	Moderate (6-15)	San Francisco Medical Center	San Francisco
17	17	Alive		.	Male	40	66.25	151	72	132	112	30	.	192	Desirable	Normal	Overweight	Very Heavy (> 25)	Los Angeles Medical Center	Los Angeles
18	18	Dead	Cancer	56	Male	56	67.25	122	72	120	87	15	72	194	Desirable	Normal	Underweight	Moderate (6-15)	Los Angeles Medical Center	Los Angeles
19	19	Alive		.	Female	42	67.75	162	96	138	119	1	.	200	Borderline	High	Overweight	Light (1-5)	Los Angeles Medical Center	Los Angeles



```
5/23/25, 6:55 PM                                                    Log: poorva-Combine Tables.ctk

1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/23/25, 6:45 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&https%3A%2F%2Fodamid-usw2-2.oda.sas
80       ! .com%2FSASStudio%2Findex='
81          *
82          */
83
84         /* The DATA step and PROC SQL methods produce the same results when the match column */
85         /* values uniquely identify each row and the tables have no other columns with the same name. */
86         /* Otherwise, the results may differ. */
87         proc sort data=WORK.IMPORT out=work._tmpsort1_;
88         by MEDID;
89         run;

NOTE: There were 19 observations read from the data set WORK.IMPORT.
NOTE: The data set WORK._TMPSORT1_ has 19 observations and 18 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1195.40k
      OS Memory           21420.00k
      Timestamp           05/23/2025 01:16:01 PM
      Step Count                        50  Switch Count  2
      Page Faults                       0
      Page Reclaims                     171
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272
```



```
5/23/25, 6:55 PM                                                    Log: poorva-Combine Tables.ctk

90
91         proc sort data=WORK.IMPORT1 out=work._tmpsort2_;
92         by MEDID;
93         run;

NOTE: There were 19 observations read from the data set WORK.IMPORT1.
NOTE: The data set WORK._TMPSORT2_ has 19 observations and 3 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1190.03k
      OS Memory           21420.00k
      Timestamp           05/23/2025 01:16:01 PM
      Step Count                        51  Switch Count  2
      Page Faults                       0
      Page Reclaims                     148
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


94
95         data work.combine;
96         merge _tmpsort1_ _tmpsort2_;
97         by MEDID;
98         run;

NOTE: There were 19 observations read from the data set WORK._TMPSORT1_.
NOTE: There were 19 observations read from the data set WORK._TMPSORT2_.
NOTE: The data set WORK.COMBINE has 19 observations and 20 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1411.34k
      OS Memory           21676.00k
      Timestamp           05/23/2025 01:16:01 PM
      Step Count                        52  Switch Count  2
      Page Faults                       0
      Page Reclaims                     224
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
```



```
5/23/25, 6:55 PM                                                    Log: poorva-Combine Tables.ctk
      Block Output Operations          264


99
100       proc delete data=work._tmpsort1_ work._tmpsort2_;
101       run;

NOTE: Deleting WORK._TMPSORT1_ (memtype=DATA).
NOTE: Deleting WORK._TMPSORT2_ (memtype=DATA).
NOTE: PROCEDURE DELETE used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              184.84k
      OS Memory           20896.00k
      Timestamp           05/23/2025 01:16:01 PM
      Step Count                       53  Switch Count  4
      Page Faults                      0
      Page Reclaims                    15
      Page Swaps                       0
      Voluntary Context Switches       22
      Involuntary Context Switches     0
      Block Input Operations           0
      Block Output Operations          0


102
103       OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
113
```

Category 3

1. What are the Types of Data Transformation Phase? Explain the detailing on Data Cleaning / Smoothing / Standardization? With each one example on creating? Total 9 Images Required?

SOLU.1:

Data transformation is the process of converting, cleaning, and structuring data from one format to a more usable format to enable processing and analysis tasks.

Examples of data transformation techniques customarily performed by users include:

▪ **Sorting:** Arranging data in ascending or descending order.

5/23/25, 9:09 PM — Results: WORK.CLASS_SORTED

Obs	Name	Sex	Age	Height	Weight
1	Joyce	F	11	51.3	50.5
2	Thomas	M	11	57.5	85.0
3	James	M	12	57.3	83.0
4	Jane	F	12	59.8	84.5
5	John	M	12	59.0	99.5
6	Louise	F	12	56.3	77.0
7	Robert	M	12	64.8	128.0
8	Alice	F	13	56.5	84.0
9	Barbara	F	13	65.3	98.0
10	Jeffrey	M	13	62.5	84.0
11	Alfred	M	14	69.0	112.5
12	Carol	F	14	62.8	102.5
13	Henry	M	14	63.5	102.5
14	Judy	F	14	64.3	90.0
15	Janet	F	15	62.5	112.5
16	Mary	F	15	66.5	112.0
17	Ronald	M	15	67.0	133.0
18	William	M	15	66.5	112.0
19	Philip	M	16	72.0	150.0

sorts students by age (ascending).

5/23/25, 9:06 PM Code: poorva-sort.sas

```
proc sort data=sashelp.class out=class_sorted;
  by Age;
run;
```

5/23/25, 9:07 PM Log: poorva-sort.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sort data=sashelp.class out=class_sorted;
70           by Age;
71         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: The data set WORK.CLASS_SORTED has 19 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1086.68k
      OS Memory           25772.00k
      Timestamp           05/23/2025 03:36:27 PM
      Step Count                        44  Switch Count  2
      Page Faults                       0
      Page Reclaims                     113
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


72
73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

- **Data Deduplication:** Removing duplicate rows from a dataset.

5/23/25, 9:20 PM Code: poorva-Data Deduplication.sas

```
proc sort data=sashelp.class nodupkey out=class_nodup;
  by age;
run;
```

5/23/25, 9:22 PM Results: WORK.CLASS_NODUP

Obs	Name	Sex	Age	Height	Weight
1	Joyce	F	11	51.3	50.5
2	James	M	12	57.3	83.0
3	Alice	F	13	56.5	84.0
4	Alfred	M	14	69.0	112.5
5	Janet	F	15	62.5	112.5
6	Philip	M	16	72.0	150.0

Removed duplicate entries based on the age column.

5/23/25, 9:21 PM Log: poorva-Data Deduplication.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sort data=sashelp.class nodupkey out=class_nodup;
70           by age;
71         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: 13 observations with duplicate key values were deleted.
NOTE: The data set WORK.CLASS_NODUP has 6 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              1085.81k
      OS Memory           20392.00k
      Timestamp           05/23/2025 03:50:51 PM
      Step Count                        30  Switch Count  2
      Page Faults                       0
      Page Reclaims                     134
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


72
73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

- **Data Aggregation:** summarizing data using operations like sum, mean, etc.

- **Splitting / Consolidating Data:**

Splitting: Dividing one dataset into multiple subsets.

5/23/25, 10:48 PM Code: poorva-splitting.sas

```
data boys girls;
  set sashelp.class;
  if Sex = 'M' then output boys;
  else output girls;
run;
```

5/23/25, 10:50 PM Results: WORK.BOYS

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Henry	M	14	63.5	102.5
3	James	M	12	57.3	83.0
4	Jeffrey	M	13	62.5	84.0
5	John	M	12	59.0	99.5
6	Philip	M	16	72.0	150.0
7	Robert	M	12	64.8	128.0
8	Ronald	M	15	67.0	133.0
9	Thomas	M	11	57.5	85.0
10	William	M	15	66.5	112.0

5/23/25, 10:51 PM Results: WORK.GIRLS

Obs	Name	Sex	Age	Height	Weight
1	Alice	F	13	56.5	84.0
2	Barbara	F	13	65.3	98.0
3	Carol	F	14	62.8	102.5
4	Jane	F	12	59.8	84.5
5	Janet	F	15	62.5	112.5
6	Joyce	F	11	51.3	50.5
7	Judy	F	14	64.3	90.0
8	Louise	F	12	56.3	77.0
9	Mary	F	15	66.5	112.0

5/23/25, 10:48 PM Log: poorva-splitting.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         data boys girls;
70           set sashelp.class;
71           if Sex = 'M' then output boys;
72           else output girls;
73         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: The data set WORK.BOYS has 10 observations and 5 variables.
NOTE: The data set WORK.GIRLS has 9 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1175.21k
      OS Memory           20648.00k
      Timestamp           05/23/2025 05:17:59 PM
      Step Count                        42  Switch Count  4
      Page Faults                       0
      Page Reclaims                     233
      Page Swaps                        0
      Voluntary Context Switches        23
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           528


74
75
76         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
86
```

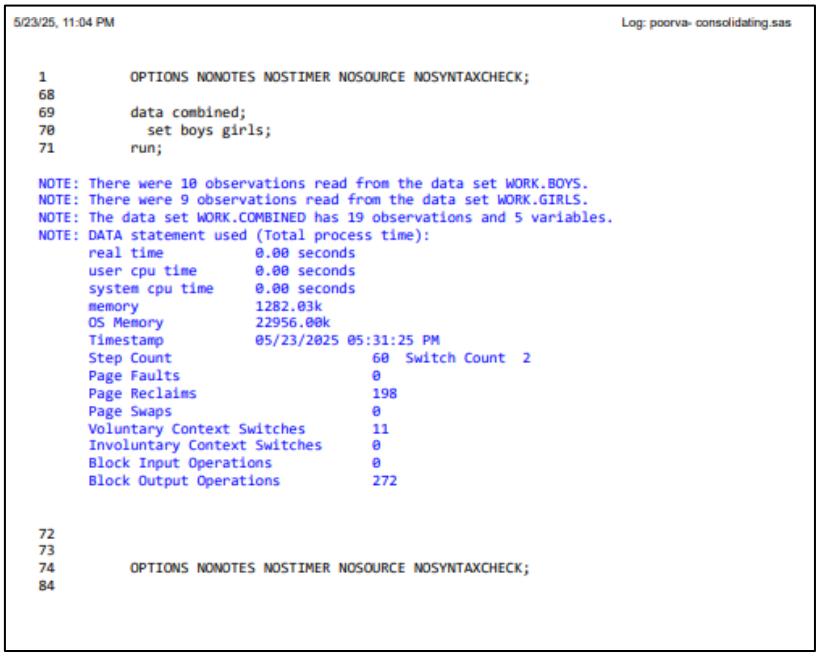
Consolidating: Merging subsets back.

5/23/25, 11:03 PM Code: poorva- consolidating.sas

```
data combined;
  set boys girls;
run;
```

5/23/25, 11:05 PM Results: WORK.COMBINED

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Henry	M	14	63.5	102.5
3	James	M	12	57.3	83.0
4	Jeffrey	M	13	62.5	84.0
5	John	M	12	59.0	99.5
6	Philip	M	16	72.0	150.0
7	Robert	M	12	64.8	128.0
8	Ronald	M	15	67.0	133.0
9	Thomas	M	11	57.5	85.0
10	William	M	15	66.5	112.0
11	Alice	F	13	56.5	84.0
12	Barbara	F	13	65.3	98.0
13	Carol	F	14	62.8	102.5
14	Jane	F	12	59.8	84.5
15	Janet	F	15	62.5	112.5
16	Joyce	F	11	51.3	50.5
17	Judy	F	14	64.3	90.0
18	Louise	F	12	56.3	77.0
19	Mary	F	15	66.5	112.0

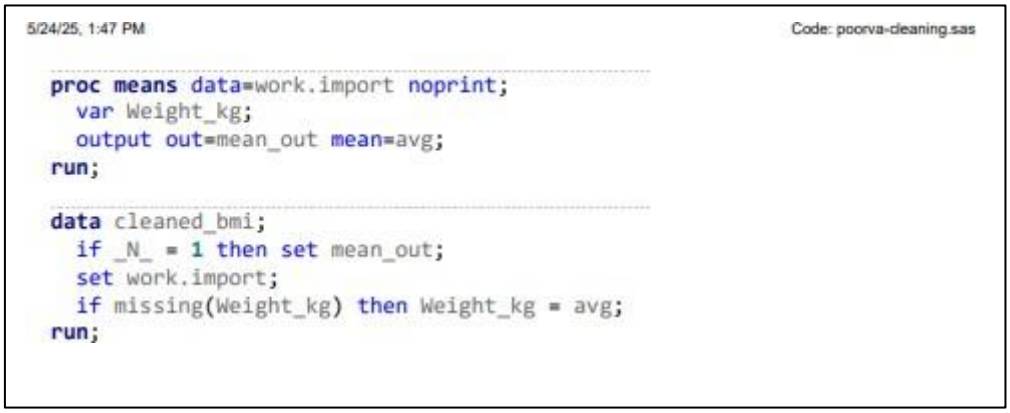


5/23/25, 11:04 PM Log: poorva- consolidating.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         data combined;
70           set boys girls;
71         run;

NOTE: There were 10 observations read from the data set WORK.BOYS.
NOTE: There were 9 observations read from the data set WORK.GIRLS.
NOTE: The data set WORK.COMBINED has 19 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1282.03k
      OS Memory           22956.00k
      Timestamp           05/23/2025 05:31:25 PM
      Step Count                        60  Switch Count  2
      Page Faults                       0
      Page Reclaims                     198
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


72
73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

▪ Data Cleaning / Smoothing / Standardization:

Data cleaning: is the process of identifying and correcting errors or inconsistencies in data to improve its quality.



5/24/25, 1:47 PM Code: poorva-cleaning.sas

```
proc means data=work.import noprint;
  var Weight_kg;
  output out=mean_out mean=avg;
run;

data cleaned_bmi;
  if _N_ = 1 then set mean_out;
  set work.import;
  if missing(Weight_kg) then Weight_kg = avg;
run;
```

5/24/25, 1:51 PM Results: WORK.CLEANED_BMI

Obs	avg	Name	Sex	Age	Height_ft	Weight_kg	BMI
1	43.214285714	Alfred	M	14	5.75	51	16.6
2	43.214285714	Alice	F	13	4.71	38.1	18.5
3	43.214285714	Barbara	F	13	5.44	44.5	16.2
4	43.214285714	Carol	F	14	5.23	46.5	18.3
5	43.214285714	Henry	M	14	5.29	46.5	17.9
6	43.214285714	James	M	12	4.77	37.6	17.8
7	43.214285714	Jane	F	12	4.98	38.3	16.6
8	43.214285714	Janet	F	15	5.21	43.214285714	20.2
9	43.214285714	Jeffrey	M	13	5.21	43.214285714	15.1
10	43.214285714	John	M	12	4.92	43.214285714	20.1

- dataset cleaned ,where missing Weight values are filled with the average.

5/24/25, 1:49 PM Log: poorva-cleaning.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc means data=work.import noprint;
70           var Weight_kg;
71           output out=mean_out mean=avg;
72         run;

NOTE: There were 10 observations read from the data set WORK.IMPORT.
NOTE: The data set WORK.MEAN_OUT has 1 observations and 3 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              7111.21k
      OS Memory           27852.00k
      Timestamp           05/24/2025 08:15:40 AM
      Step Count                        74  Switch Count  3
      Page Faults                       0
      Page Reclaims                     1742
      Page Swaps                        0
      Voluntary Context Switches        32
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


73
74         data cleaned_bmi;
75           if _N_ = 1 then set mean_out;
76           set work.import;
77           if missing(Weight_kg) then Weight_kg = avg;
78         run;

NOTE: There were 1 observations read from the data set WORK.MEAN_OUT.
NOTE: There were 10 observations read from the data set WORK.IMPORT.
NOTE: The data set WORK.CLEANED_BMI has 10 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1297.21k
      OS Memory           22188.00k
      Timestamp           05/24/2025 08:15:40 AM
      Step Count                        75  Switch Count  2
      Page Faults                       0
      Page Reclaims                     180
```

about:blank

5/24/25, 1:49 PM Log: poorva-cleaning.sas

```
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


79
80
81         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
91
```

Data smoothing: Data Smoothing is the process of removing irregularities from data to reveal important patterns or trends.

- The purpose of data smoothing is to make the data more consistent or easier to analyze.

- Highlight overall trends by reducing random variation.

- Common techniques include: Binning- (Group data values into categories) (e.g., Low, Medium, High), Moving Averages-(Use the average of nearby values to smooth fluctuations),Regression Smoothing-(Fit a trend line to the data)

5/24/25, 5:12 PM Code: Program 1

```
data smoothed;
  set WORK.IMPORT1;
  length Weight_Group $10;
  if Weight_kg < 45 then Weight_Group = "Low"; else if Weight_kg < 60 then Weight_Group = "Medium"; else Weight_Group = "High";
run;
```

5/24/25, 5:13 PM Results: WORK.SMOOTHED

Obs	Name	Sex	Age	Height_ft	Weight_kg	Weight_Group
1	Alfred	M	14	5.75	51	Medium
2	Alice	F	13	4.71	38.1	Low
3	Barbara	F	13	5.44	44.5	Low
4	Carol	F	14	5.23	46.5	Medium
5	Henry	M	14	5.29	81	High
6	James	M	12	4.77	37.6	Low
7	Jane	F	12	4.98	38.3	Low
8	Janet	F	15	5.21	51	Medium
9	Jeffrey	M	13	5.21	85	High
10	John	M	12	4.92	45.1	Medium

This groups weights into **"Low"**, **"Medium"**, or **"High"** based on **Weight_kg**.

5/24/25, 5:12 PM Log: Program 1

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         data smoothed;
70           set WORK.IMPORT1;
71           length Weight_Group $10;
72           if Weight_kg < 45 then Weight_Group = "Low"; else if Weight_kg < 60 then Weight_Group = "Medium"; else Weight_Group =
72       ! "High";
73         run;

NOTE: There were 10 observations read from the data set WORK.IMPORT1.
NOTE: The data set WORK.SMOOTHED has 10 observations and 6 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              946.62k
      OS Memory           21416.00k
      Timestamp           05/24/2025 11:41:08 AM
      Step Count                        45  Switch Count  2
      Page Faults                       0
      Page Reclaims                     154
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


74
75
76         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
86
```

Data standardization is the process of transforming data so that it follows a common scale or format, usually by adjusting values to have a mean of 0 and a standard deviation of 1.

5/24/25, 6:54 PM Code: poorva-Standardize Data.ctk

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/24/25, 6:54 PM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS02-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36'
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-24647-WYJfgCTl3oWqatJmf080-cas'
 *
 */

ods noproctitle;

proc stdize data=SASHELP.FISH method=std nomiss out=work.Stdize0002 oprefix
        sprefix=Standardized_;
    var Height;
run;
```

5/24/25, 6:56 PM Results: WORK.STDIZE0002

Obs	Species	Weight	Length1	Length2	Length3	Height	Width	Standardized_Height
1	Bream	242.0	23.2	25.4	30.0	11.5200	4.0200	0.59470
2	Bream	290.0	24.0	26.3	31.2	12.4800	4.3056	0.81867
3	Bream	340.0	23.9	26.5	31.1	12.3778	4.6961	0.79483
4	Bream	363.0	26.3	29.0	33.5	12.7300	4.4555	0.87700
5	Bream	430.0	26.5	29.0	34.0	12.4440	5.1340	0.81027
6	Bream	450.0	26.8	29.7	34.7	13.6024	4.9274	1.08054
7	Bream	500.0	26.8	29.7	34.5	14.1795	5.2785	1.21518
8	Bream	390.0	27.6	30.0	35.0	12.6700	4.6900	0.86300
9	Bream	450.0	27.6	30.0	35.1	14.0049	4.8438	1.17444
10	Bream	500.0	28.5	30.7	36.2	14.2266	4.9594	1.22617
11	Bream	475.0	28.4	31.0	36.2	14.2628	5.1042	1.23461
12	Bream	500.0	28.7	31.0	36.2	14.3714	4.8146	1.25995
13	Bream	500.0	29.1	31.5	36.4	13.7592	4.3680	1.11712
14	Bream	.	29.5	32.0	37.3	13.9129	5.0728	1.15298
15	Bream	600.0	29.4	32.0	37.2	14.9544	5.1708	1.39597
16	Bream	600.0	29.4	32.0	37.2	15.4380	5.5800	1.50879
17	Bream	700.0	30.4	33.0	38.3	14.8604	5.2854	1.37404
18	Bream	700.0	30.4	33.0	38.5	14.9380	5.1975	1.39214
19	Bream	610.0	30.9	33.5	38.6	15.6330	5.1338	1.55429
20	Bream	650.0	31.0	33.5	38.7	14.4738	5.7276	1.28384
21	Bream	575.0	31.3	34.0	39.5	15.1285	5.5695	1.43659
22	Bream	685.0	31.4	34.0	39.2	15.9936	5.3704	1.63842
23	Bream	620.0	31.5	34.5	39.7	15.5227	5.2801	1.52856
24	Bream	680.0	31.8	35.0	40.6	15.4686	6.1306	1.51593
25	Bream	700.0	31.9	35.0	40.5	16.2405	5.5890	1.69602

5/24/25, 6:55 PM Log: poorva-Standardize Data.clk

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/24/25, 6:54 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS02-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-24647-WYJfgCTl3oWqatJmf08
80       ! 0-cas'
81          *
82          */
83
84         ods noproctitle;
85
86         proc stdize data=SASHELP.FISH method=std nomiss out=work.Stdize0002 oprefix
87         sprefix=Standardized_;
88         var Height;
89         run;

NOTE: There were 159 observations read from the data set SASHELP.FISH.
NOTE: The data set WORK.STDIZE0002 has 159 observations and 8 variables.
NOTE: PROCEDURE STDIZE used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              723.15k
      OS Memory           22436.00k
      Timestamp           05/24/2025 01:24:19 PM
      Step Count                        92  Switch Count  2
      Page Faults                       0
      Page Reclaims                     105
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272
```

about:blank

5/24/25, 6:55 PM Log: poorva-Standardize Data.clk

```
90
91         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
101
```

- **Data Normalization:** Scaling values to a range (like 0 to 1).

5/24/25, 12:18 PM Code: poorva-data normalization.sas

```
proc stdize data=sashelp.class method=range out=normalized;
  var Height;
run;
```

5/24/25, 12:19 PM Results: WORK.NORMALIZED

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	0.85507	112.5
2	Alice	F	13	0.25121	84.0
3	Barbara	F	13	0.67633	98.0
4	Carol	F	14	0.55556	102.5
5	Henry	M	14	0.58937	102.5
6	James	M	12	0.28986	83.0
7	Jane	F	12	0.41063	84.5
8	Janet	F	15	0.54106	112.5
9	Jeffrey	M	13	0.54106	84.0
10	John	M	12	0.37198	99.5
11	Joyce	F	11	0.00000	50.5
12	Judy	F	14	0.62802	90.0
13	Louise	F	12	0.24155	77.0
14	Mary	F	15	0.73430	112.0
15	Philip	M	16	1.00000	150.0
16	Robert	M	12	0.65217	128.0
17	Ronald	M	15	0.75845	133.0
18	Thomas	M	11	0.29952	85.0
19	William	M	15	0.73430	112.0

- Normalizes Height to range [0,1].

5/24/25, 12:19 PM Log: poorva-data normalization.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc stdize data=sashelp.class method=range out=normalized;
70           var Height;
71         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: The data set WORK.NORMALIZED has 19 observations and 5 variables.
NOTE: PROCEDURE STDIZE used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              718.00k
      OS Memory           19620.00k
      Timestamp           05/24/2025 06:48:03 AM
      Step Count                        24  Switch Count  2
      Page Faults                       0
      Page Reclaims                     132
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


72
73
74         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
84
```

▪ **Data Filtering:** Keeping only rows that meet a condition.

5/24/25, 12:28 PM Code: poorva-data filtering.sas

```
data filtered;
  set sashelp.class;
  if Age >= 14;
run;
```

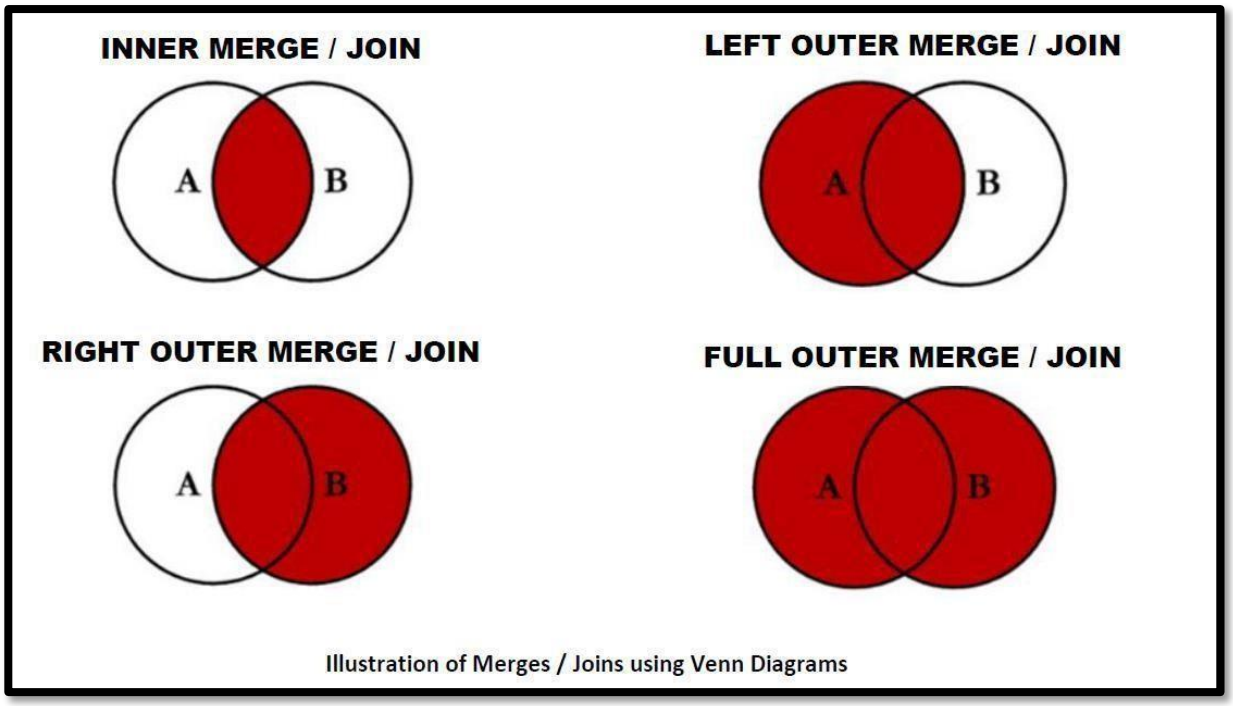
5/24/25, 12:30 PM Results: WORK.FILTERED

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Carol	F	14	62.8	102.5
3	Henry	M	14	63.5	102.5
4	Janet	F	15	62.5	112.5
5	Judy	F	14	64.3	90.0
6	Mary	F	15	66.5	112.0
7	Philip	M	16	72.0	150.0
8	Ronald	M	15	67.0	133.0
9	William	M	15	66.5	112.0

5/24/25, 12:28 PM Log: poorva-data filtering.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         data filtered;
70           set sashelp.class;
71           if Age >= 14;
72         run;

NOTE: There were 19 observations read from the data set SASHELP.CLASS.
NOTE: The data set WORK.FILTERED has 9 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              832.34k
      OS Memory           21668.00k
      Timestamp           05/24/2025 06:57:45 AM
      Step Count                        36  Switch Count  2
      Page Faults                       0
      Page Reclaims                     142
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


73
74
75         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
85
```

2. Explain the illustrate merges on below images. Use by Proc FORMAT/ SQL (Optional)



MERGE/JOIN: join or merge operations are used to combine data from two or more datasets based on common key variables (like an ID). The goal is to create a unified dataset that includes information from all source datasets.

Results: SASHELP.CLASS

Obs	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

Results: WORK.IMPORT

Obs	Name	City
1	Henry	Bangalore
2	Alice	Pune
3	Robert	Hyderabad
4	Carol	Ahmedabad

INNER MERGE/JOIN: returns only the rows that have matching keys in both datasets.

5/24/25, 9:20 PM — Code: poorva- inner merge sas.sas

```
proc sql;
  select a.*, b.City
  from sashelp.class as a
  inner join WORK.IMPORT as b
  on a.Name = b.Name;
quit;
```

5/24/25, 9:22 PM — Results: poorva- inner merge sas.sas

Name	Sex	Age	Height	Weight	City
Alice	F	13	56.5	84	Pune
Carol	F	14	62.8	102.5	Ahmedabad
Henry	M	14	63.5	102.5	Bangalore
Robert	M	12	64.8	128	Hyderabad

Note- Only matched names appeared.

5/24/25, 9:22 PM — Log: poorva- inner merge sas.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70           select a.*, b.City
71           from sashelp.class as a
72           inner join WORK.IMPORT as b
73           on a.Name = b.Name;
74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              5786.37k
      OS Memory           26792.00k
      Timestamp           05/24/2025 03:50:45 PM
      Step Count                        58  Switch Count  0
      Page Faults                       0
      Page Reclaims                     148
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           16


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

LEFT OUTER MERGE/JOIN: returns all rows from the left table (A), and matching rows from the right table (B). Non-matching rows in B are filled with missing values.

5/24/25, 9:47 PM Code: poorva-left outer merge.sas

```
proc sql;
  select a.*, b.City
  from sashelp.class as a
  left join WORK.IMPORT as b
  on a.Name = b.Name;
quit;
```

5/24/25, 9:49 PM Results: poorva-left outer merge.sas

Name	Sex	Age	Height	Weight	City
Alfred	M	14	69	112.5	
Alice	F	13	56.5	84	Pune
Barbara	F	13	65.3	98	
Carol	F	14	62.8	102.5	Ahmedabad
Henry	M	14	63.5	102.5	Bangalore
James	M	12	57.3	83	
Jane	F	12	59.8	84.5	
Janet	F	15	62.5	112.5	
Jeffrey	M	13	62.5	84	
John	M	12	59	99.5	
Joyce	F	11	51.3	50.5	
Judy	F	14	64.3	90	
Louise	F	12	56.3	77	
Mary	F	15	66.5	112	
Philip	M	16	72	150	
Robert	M	12	64.8	128	Hyderabad
Ronald	M	15	67	133	
Thomas	M	11	57.5	85	
William	M	15	66.5	112	

Note- All from sashelp.class, matched city where available appeared.

5/24/25, 9:48 PM Log: poorva-left outer merge.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70           select a.*, b.City
71           from sashelp.class as a
72           left join WORK.IMPORT as b
73           on a.Name = b.Name;
74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              5786.90k
      OS Memory           27560.00k
      Timestamp           05/24/2025 04:16:11 PM
      Step Count                        76  Switch Count  0
      Page Faults                       0
      Page Reclaims                     285
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           24


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

RIGHT OUTER MERGE/JOIN: returns all rows from the right table (B), and matching rows from the left table (A). Non-matching rows in A are filled with missing values.

5/24/25, 10:14 PM Code: poorva-right outer merge.sas

```
proc sql;
  select a.*, b.City
  from sashelp.class as a
  right join WORK.IMPORT as b
  on a.Name = b.Name;
quit;
```

5/24/25, 10:15 PM Results: poorva-right outer merge.sas

Name	Sex	Age	Height	Weight	City
Alice	F	13	56.5	84	Pune
Carol	F	14	62.8	102.5	Ahmedabad
Henry	M	14	63.5	102.5	Bangalore
Robert	M	12	64.8	128	Hyderabad

Note - All from city_data, matched class details where available ,appeared.

5/24/25, 10:15 PM Log: poorva-right outer merge.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70           select a.*, b.City
71           from sashelp.class as a
72           right join WORK.IMPORT as b
73           on a.Name = b.Name;
74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.01 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              5784.68k
      OS Memory           27560.00k
      Timestamp           05/24/2025 04:43:39 PM
      Step Count                        82  Switch Count  0
      Page Faults                       0
      Page Reclaims                     267
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           24


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

FULL OUTER MERGE/JOIN: returns all rows from both tables, with matching rows where available. Missing values are filled where matches don't exist.

5/24/25, 10:27 PM Code: poorva-full joint.sas

```
proc sql;
  select a.Name as ClassName, a.Age, b.City
  from sashelp.class as a
  full join WORK.IMPORT as b
  on a.Name = b.Name;
quit;
```

5/24/25, 10:28 PM Results: poorva-full joint.sas

ClassName	Age	City
Alfred	14	
Alice	13	Pune
Barbara	13	
Carol	14	Ahmedabad
Henry	14	Bangalore
James	12	
Jane	12	
Janet	15	
Jeffrey	13	
John	12	
Joyce	11	
Judy	14	
Louise	12	
Mary	15	
Philip	16	
Robert	12	Hyderabad
Ronald	15	
Thomas	11	
William	15	

Note - All names from both, matched where possible.

5/24/25, 10:28 PM Log: poorva-full joint.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70           select a.Name as ClassName, a.Age, b.City
71           from sashelp.class as a
72           full join WORK.IMPORT as b
73           on a.Name = b.Name;
74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.01 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              5763.56k
      OS Memory           27560.00k
      Timestamp           05/24/2025 04:56:42 PM
      Step Count                        88  Switch Count  0
      Page Faults                       0
      Page Reclaims                     266
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           24


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

3. Create a Program on Binary Logistic Regression? with creation of an example?

SOLU 3:

Binary logistic regression is a statistical technique used to model the relationship between a binary dependent variable (e.g., Yes/No, 0/1) and one or more independent variables (continuous or categorical).

EXAMPLE – prediction of the likelihood of a heart attack based on smoking status.

5/25/25, 10:55 AM Code: Binary Logistic Regression

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/25/25, 10:53 AM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36'
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37B42-OI7d0TKc2nbWdTKFVDyf-cas'
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc logistic data=WORK.IMPORT1;
    class Smoker / param=glm;
    model Heart_Attack(event='No')=Smoker / link=logit technique=fisher;
run;
```

5/25/25, 10:57 AM Results: Binary Logistic Regression

Model Information		
Data Set	WORK.IMPORT1	
Response Variable	Heart_Attack	Heart_Attack
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	60
Number of Observations Used	60

Response Profile		
Ordered Value	Heart_Attack	Total Frequency
1	No	37
2	Yes	23

Probability modeled is Heart_Attack='No'.

Class Level Information			
Class	Value	Design Variables	
Smoker	No	1	0
	Yes	0	1

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	81.881	76.254
SC	83.975	80.442
-2 Log L	79.881	72.254

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.6270	1	0.0057
Score	7.3916	1	0.0066

“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Wald	6.9481	1	0.0084

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Smoker	1	6.9481	0.0084

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1942	0.3609	0.2894	0.5906
Smoker	No	1	1.5379	0.5834	6.9481	0.0084
Smoker	Yes	0	0	.	.	.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Smoker No vs Yes	4.655	1.483	14.605

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	45.9	Somers' D	0.361
Percent Discordant	9.9	Gamma	0.646
Percent Tied	44.2	Tau-a	0.173
Pairs	851	c	0.680


```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/25/25, 10:53 AM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37842-OI7d0TKc2nbWdTKFVDy
80       ! f-cas'
81          *
82          */
83
84         ods noproctitle;
85         ods graphics / imagemap=on;
86
87         proc logistic data=WORK.IMPORT1;
88         class Smoker / param=glm;
89         model Heart_Attack(event='No')=Smoker / link=logit technique=fisher;
90         run;

NOTE: PROC LOGISTIC is modeling the probability that Heart_Attack='No'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 60 observations read from the data set WORK.IMPORT1.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.05 seconds
      user cpu time       0.05 seconds
      system cpu time     0.01 seconds
      memory              3011.31k
      OS Memory           22964.00k
      Timestamp           05/25/2025 05:24:29 AM
      Step Count                        62  Switch Count  0
      Page Faults                       0
      Page Reclaims                     326
      Page Swaps                        0
      Voluntary Context Switches        3
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           56
```



```
91
92         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
102
```

4. Create a Program on Simple Linear Regression with Flow Chart Presentation? with creation of example?

SOLU 4:

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors) using a straight line.

[Go to Left Panel → Tasks and Utilities]

↓

[Click Tasks → Statistics → Linear models→ Linear Regression]

↓

[Select a Built-in Dataset (e.g., SASHELP.CLASS)]

↓

[Go to ROLES section]

↓

[Set Dependent Variable (Y) → e.g., Weight]

↓

[Set continuous variable(x) → e.g., Height]

↓

[Click the Run Button at the Top]

↓

[View Results → Equation, R-Square, P-values, Fit Plot]

5/25/25, 11:31 AM Results: WORK.IMPORT2

Obs	StudyHours	GPA
1	5	2
2	10	2.4
3	15	2.8
4	20	3
5	25	3.2
6	30	3.5
7	35	3.6
8	40	3.8
9	45	3.9
10	50	4

Example - A simple linear regression model with Weight as the dependent variable and Height as the independent variable.

5/25/25, 12:27 PM Code: Linear Regression 1

```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/25/25, 12:11 PM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36'
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37842-OI7d0TKc2nbWdTKFVDyf-cas'
 *
 */

ods noproctitle;
ods graphics / imagemap=on;

proc reg data=SASHELP.CLASS alpha=0.05 plots(only)=(diagnostics residuals
        fitplot observedbypredicted);
    model Weight=Height /;
    run;
quit;
```

5/25/25, 12:28 PM Results: Linear Regression 1

Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

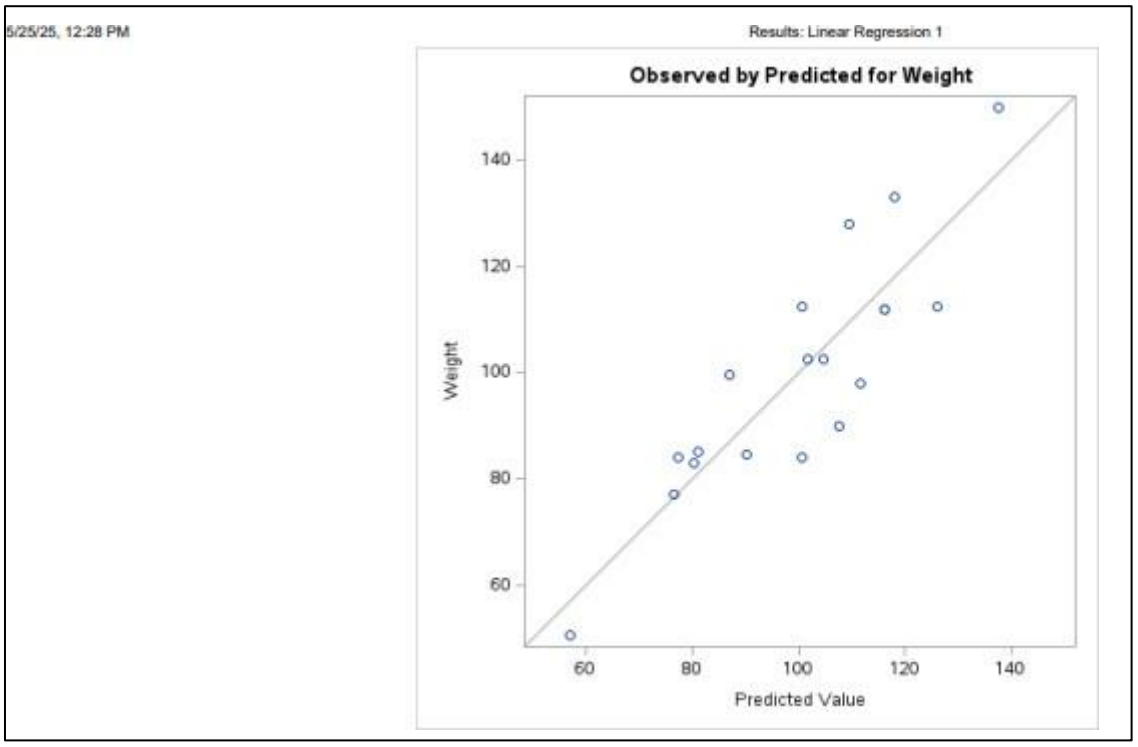
Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

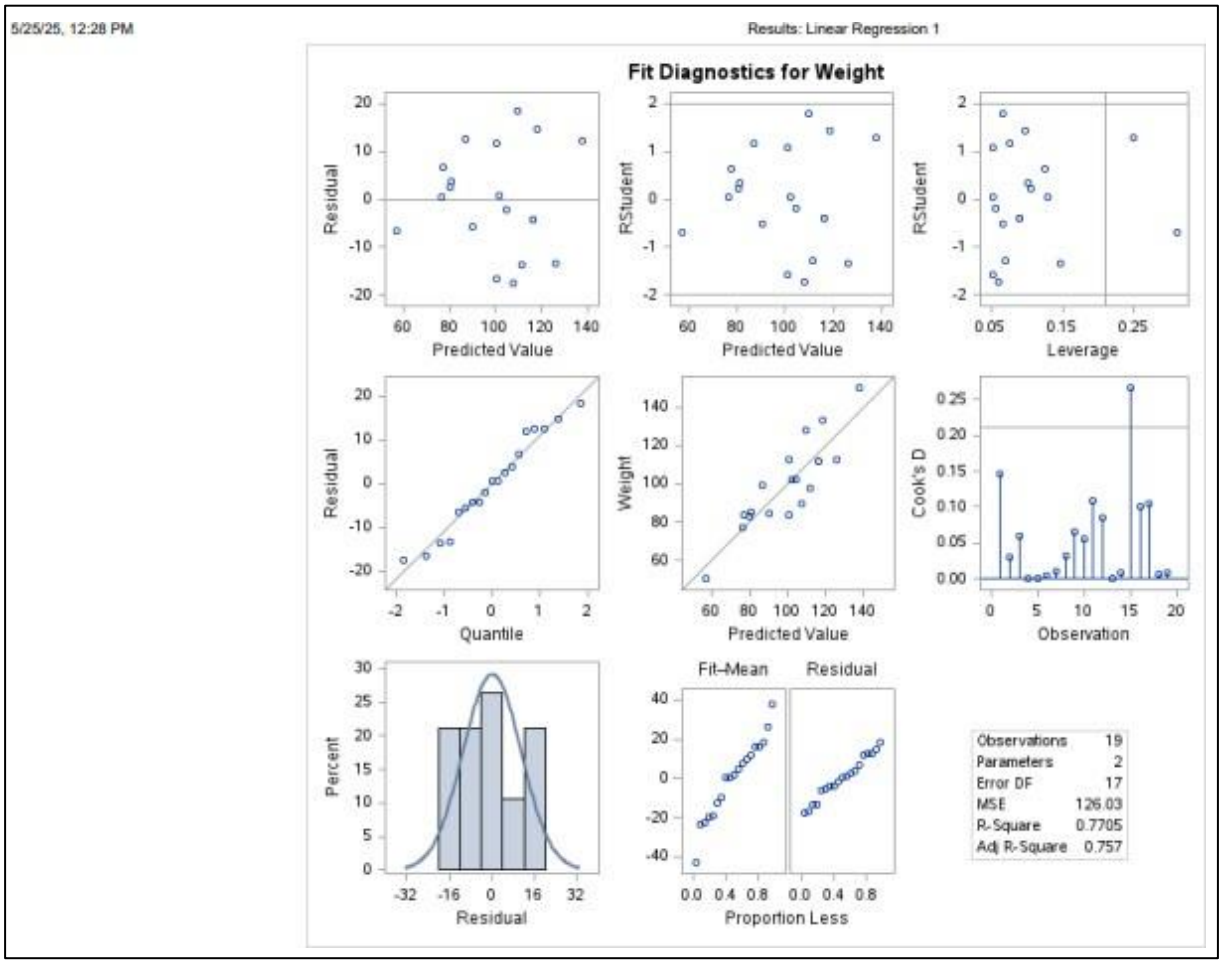
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-143.02692	32.27459	-4.43	0.0004
Height	1	3.89903	0.51609	7.55	<.0001

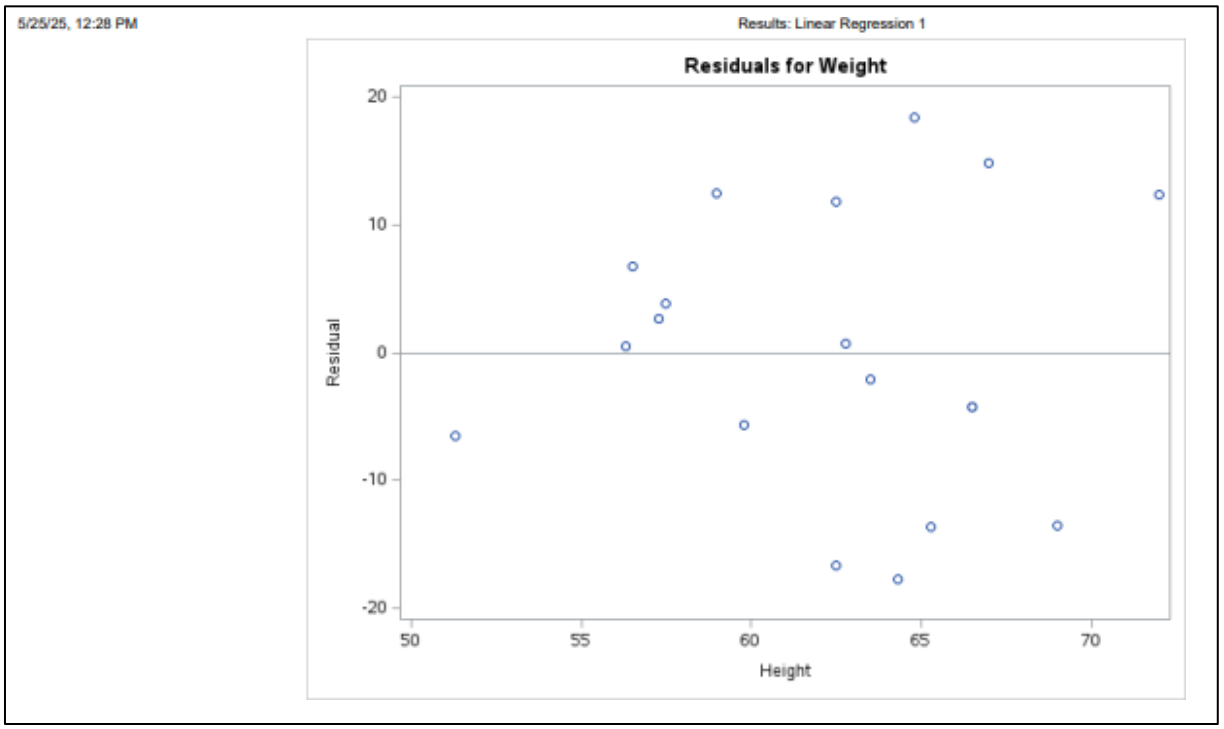
Model: MODEL1
Dependent Variable: Weight

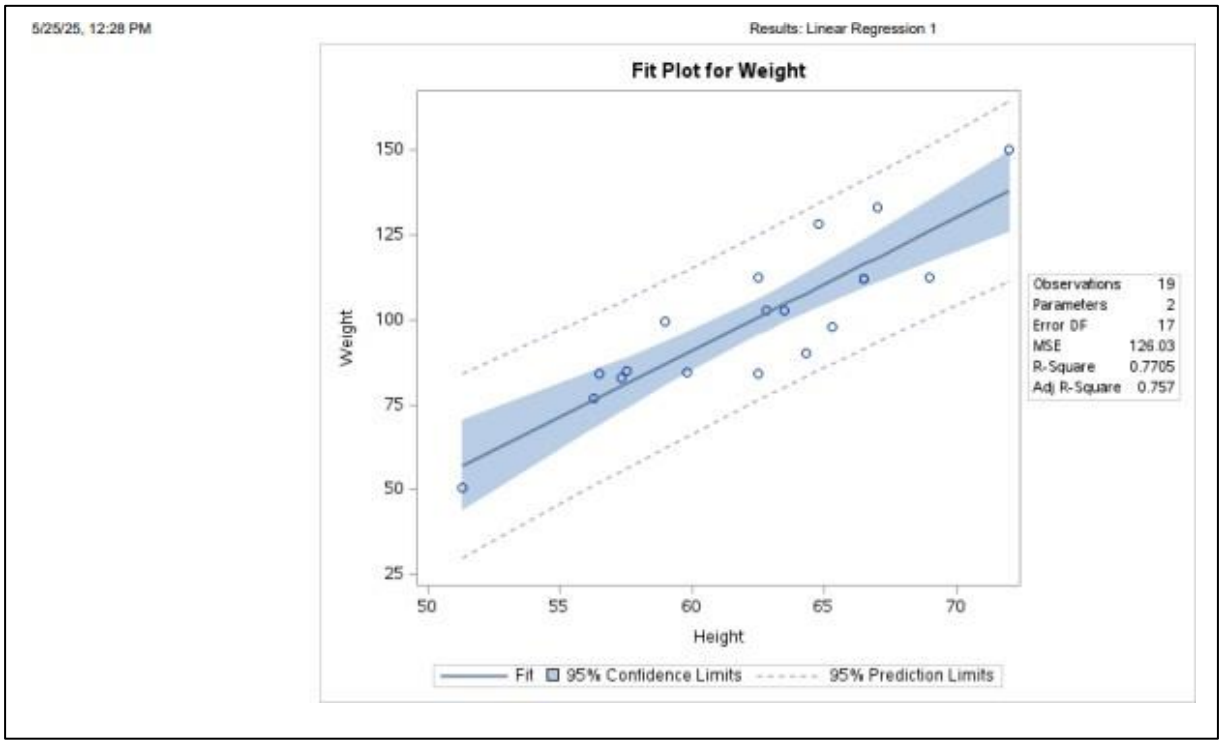
“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

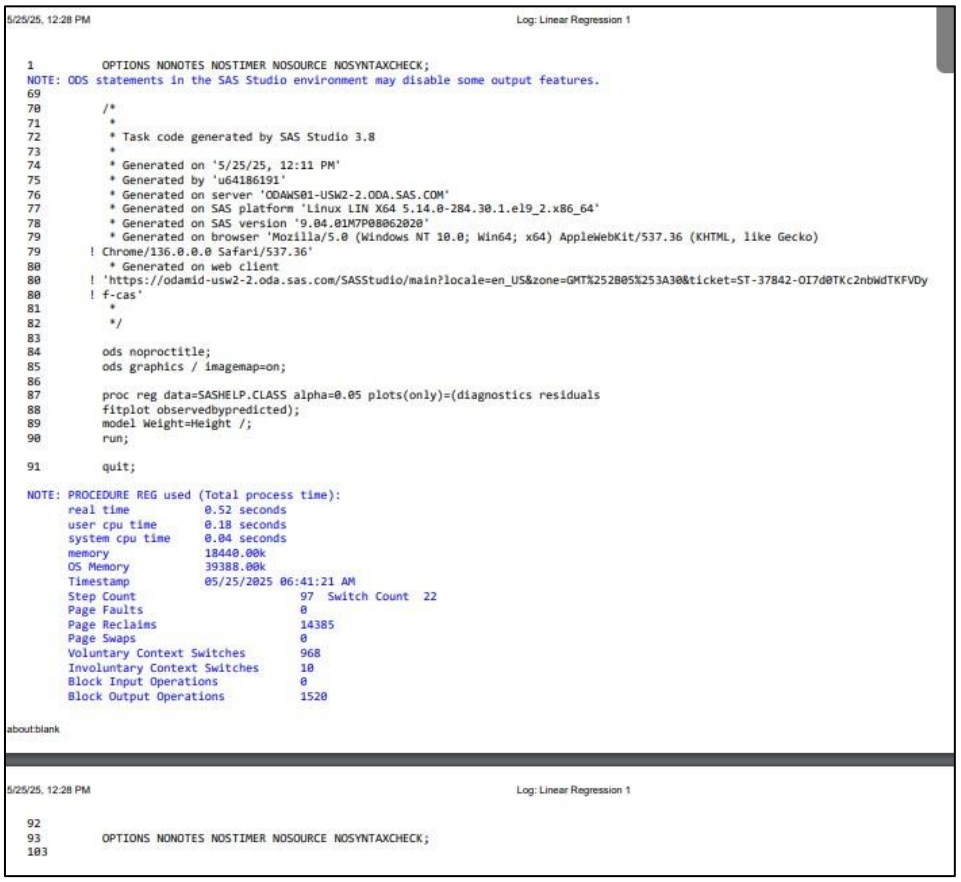




“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"







```
5/25/25, 12:28 PM                                                    Log: Linear Regression 1

1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/25/25, 12:11 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79       ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80       ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37842-OI7d0TKc2nbWdTKFVDy
80       ! f-cas'
81          *
82          */
83
84         ods noproctitle;
85         ods graphics / imagemap=on;
86
87         proc reg data=SASHELP.CLASS alpha=0.05 plots(only)=(diagnostics residuals
88         fitplot observedbypredicted);
89         model Weight=Height /;
90         run;

91         quit;

NOTE: PROCEDURE REG used (Total process time):
      real time               0.52 seconds
      user cpu time           0.18 seconds
      system cpu time         0.04 seconds
      memory                  18440.00k
      OS Memory               39388.00k
      Timestamp               05/25/2025 06:41:21 AM
      Step Count                        97  Switch Count  22
      Page Faults                       0
      Page Reclaims                     14385
      Page Swaps                        0
      Voluntary Context Switches        968
      Involuntary Context Switches      10
      Block Input Operations            0
      Block Output Operations           1520

about:blank

5/25/25, 12:28 PM                                                    Log: Linear Regression 1

92
93         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
103
```

5. Create a Program on Data Exploration with Flow Chart Presentation? with creation of example?

SOLU 5: Data Exploration is the process of visually and statistically examining a dataset to understand: What variables are present, How the data is distributed, Whether there are missing values or outliers, Relationships between variables.
It helps to model or analyze the data.

[Go to Left Panel → Tasks and Utilities]
↓
[Click Tasks → Data →Statistics→ Data Exploration]
↓
[Select Dataset → eg. SASHELP.CLASS]
↓
[Choose continuous variables: eg. Age,height,weight]
↓

[choose classification variables: eg. Sex]
↓
[select plots]
↓
[Click Run]
↓
[View results]

Example – data exploration on SASHELP.CLASS

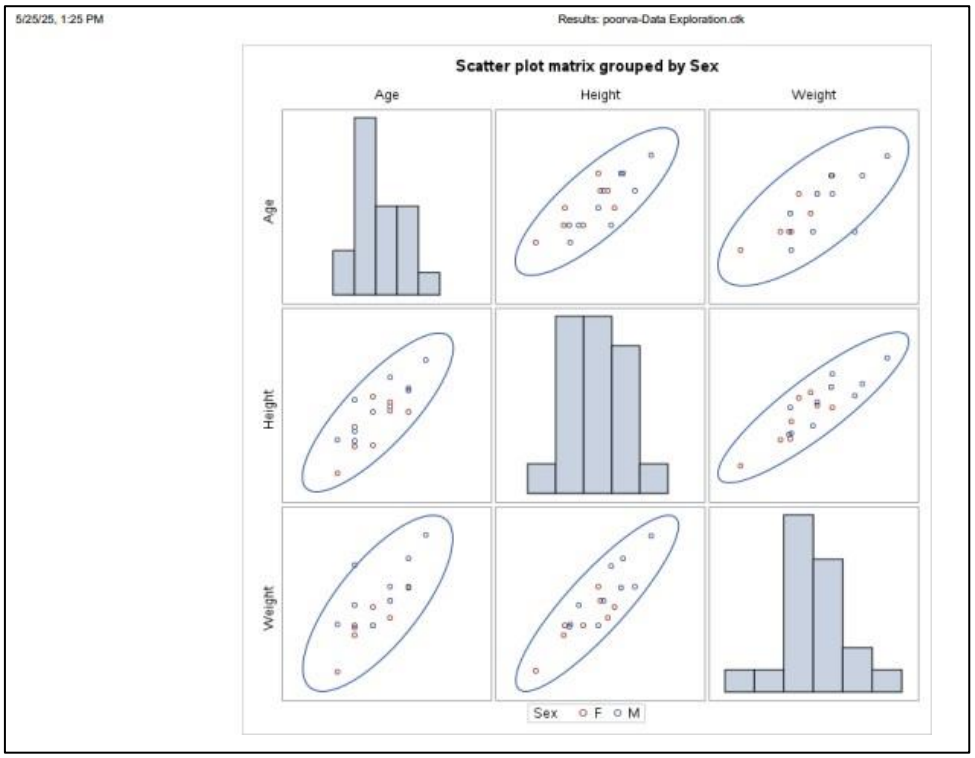
5/25/25, 1:19 PM Code: poorva-Data Exploration.ctk

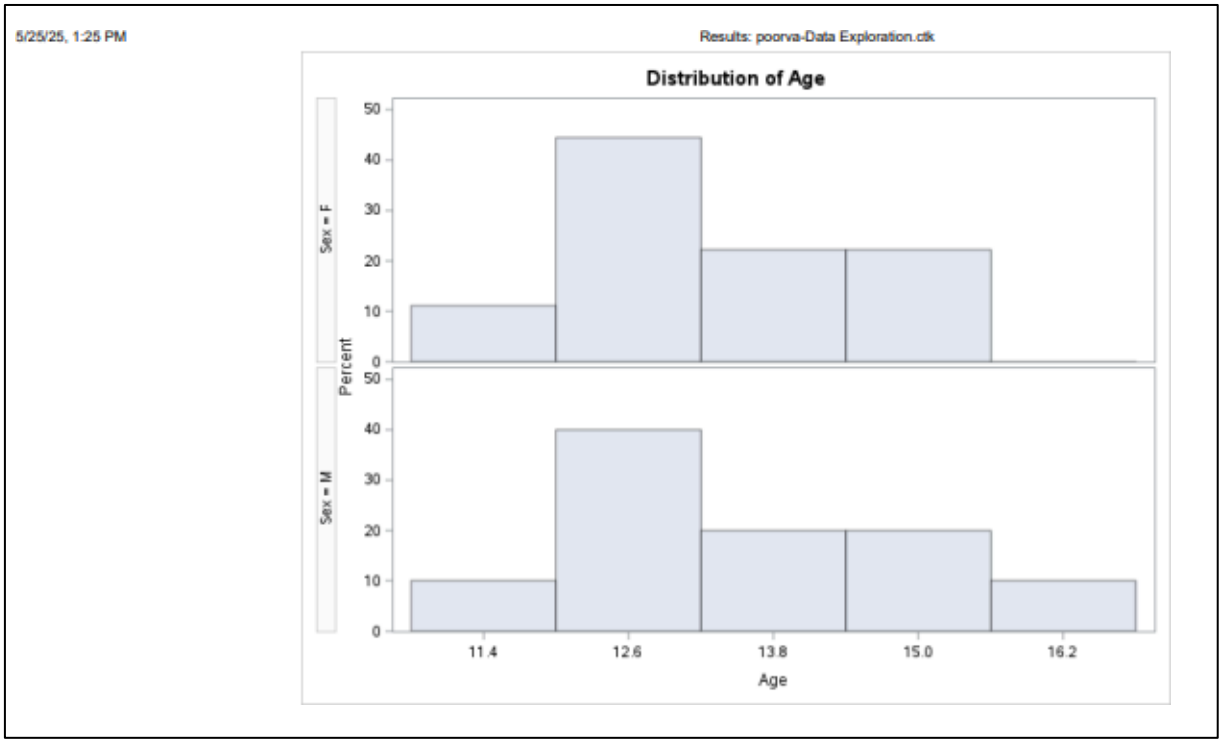
```
/*
 *
 * Task code generated by SAS Studio 3.8
 *
 * Generated on '5/25/25, 1:18 PM'
 * Generated by 'u64186191'
 * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
 * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
 * Generated on SAS version '9.04.01M7P08062020'
 * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/136.0.0.0 Safari/537.36'
 * Generated on web client 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37842-OI7d0TKc2nbWdTKFVDyf-cas'
 *
 */

options validvarname=any;
ods noproctitle;
ods graphics / imagemap=on;

/* Scatter plot matrix macro */
%macro scatterPlotMatrix(xVars=, title=, groupVar=);
    proc sgscatter data=SASHELP.CLASS;
        matrix &xVars / %if(&groupVar ne %str()) %then
            %do;
                group=&groupVar legend=(sortorder=ascending) %end;
        diagonal=(histogram) ellipse=(type=predicted alpha=0.05);
        title &title;
    run;

    title;
%mend scatterPlotMatrix;

/* Histogram (one-way or two-way) */
%macro DEHisto(data=, avar=, classVar=);
    %local i numAVars numCVars cVar cVar1 cVar2;
    %let numAVars=%Sysfunc(countw(%str(&avar), %str( ), %str(q)));
    %let numCVars=%Sysfunc(countw(%str(&classVar), %str( ), %str(q)));

    %if(&numAVars>0 & &numCVars>0) %then
        %do;

            %if(&numCVars=1) %then
                %do;
                    %let cVar=%scan(%str(&classVar), 1, %str( ), %str(q));

                    proc sql noprint;
                        select count(distinct &cVar) into :nrows from &data;
                    quit;

                    /* One-way histogram */
```

5/25/25, 1:19 PM Code: poorva-Data Exploration.ctk

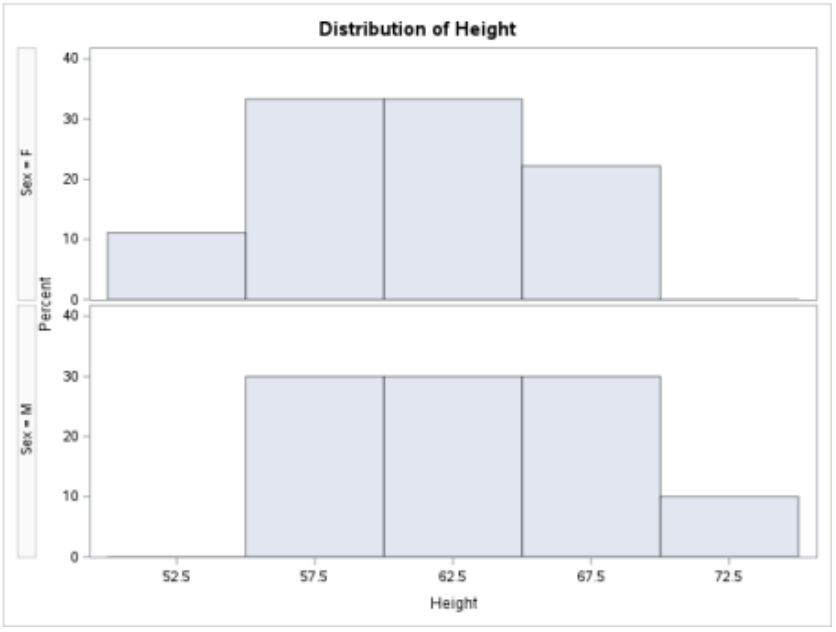
```
                    proc univariate data=&data noprint;
                        var &avar;
                        class &cVar;
                        histogram &avar / nrows=&nrows;
                    run;

                %end;
            %else
                %do;

                    /* One-way histogram of each class variable */

                    %do i=1 %to %eval(&numCVars);
                        %let cVar=%scan(%str(&classVar), &i, %str( ), %str(q));

                        proc sql noprint;
                            select count(distinct &cVar) into :nrows from &data;
                        quit;

                        proc univariate data=&data noprint;
                            var &avar;
                            class &cVar;
                            histogram &avar / nrows=&nrows;
                        run;

                    %end;

                    /* Two-way histogram */
                %let cVar1=%scan(%str(&classVar), 1, %str( ),
                        %str(q));
                    %let cVar2=%scan(%str(&classVar), 2, %str( ), %str(q));

                    proc sql noprint;
                        select count(distinct &cVar1) into :nrows from &data;
                    quit;

                    proc sql noprint;
                        select count(distinct &cVar2) into :ncols from &data;
                    quit;

                    proc univariate data=&data noprint;
                        var &avar;
                        class &cVar1 &cVar2;
                        histogram &avar / nrows=&nrows ncols=&ncols;
                    run;

                %end;
        %end;
%mend DEHisto;
```

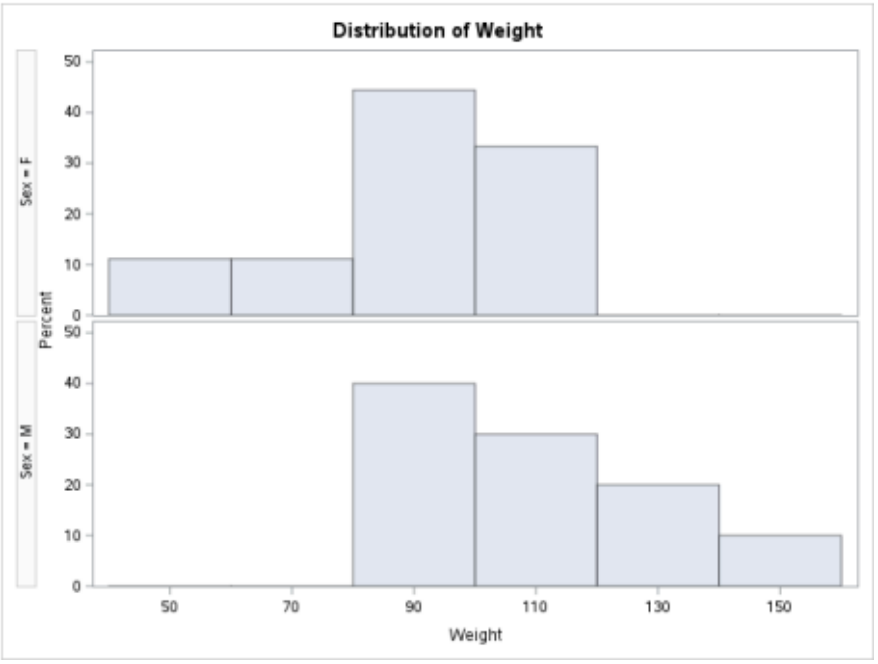
“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"





“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"





“Portfolio Project of Poorva Dixit (For Recruitment Purpose Only)"

5/25/25, 1:20 PM Log: poorva-Data Exploration.clk

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
69
70         /*
71          *
72          * Task code generated by SAS Studio 3.8
73          *
74          * Generated on '5/25/25, 1:18 PM'
75          * Generated by 'u64186191'
76          * Generated on server 'ODAWS01-USW2-2.ODA.SAS.COM'
77          * Generated on SAS platform 'Linux LIN X64 5.14.0-284.30.1.el9_2.x86_64'
78          * Generated on SAS version '9.04.01M7P08062020'
79          * Generated on browser 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko)
79        ! Chrome/136.0.0.0 Safari/537.36'
80          * Generated on web client
80        ! 'https://odamid-usw2-2.oda.sas.com/SASStudio/main?locale=en_US&zone=GMT%252B05%253A30&ticket=ST-37842-OI7d0TKc2nbWdTKFVDy
80        ! f-cas'
81          *
82          */
83
84         options validvarname=any;
85         ods noproctitle;
86         ods graphics / imagemap=on;
87
88         /* Scatter plot matrix macro */
89         %macro scatterPlotMatrix(xVars=, title=, groupVar=);
90         proc sgscatter data=SASHELP.CLASS;
91         matrix &xVars / %if(&groupVar ne %str()) %then
92         %do;
93         group=&groupVar legend=(sortorder=ascending) %end;
94         diagonal=(histogram) ellipse=(type=predicted alpha=0.05);
95         title &title;
96         run;
97
98         title;
99         %mend scatterPlotMatrix;
100
101        /* Histogram (one-way or two-way) */
102        %macro DEHisto(data=, avar=, classVar=);
103        %local i numAVars numCVars cVar cVar1 cVar2;
104        %let numAVars=%Sysfunc(countw(%str(&avar), %str( ), %str(q)));
105        %let numCVars=%Sysfunc(countw(%str(&classVar), %str( ), %str(q)));
106
107        %if(&numAVars>0 & &numCVars>0) %then
108        %do;
109
```

5/25/25, 1:20 PM Log: poorva-Data Exploration.clk

```
110        %if(&numCVars=1) %then
111        %do;
112        %let cVar=%scan(%str(&classVar), 1, %str( ), %str(q));
113
114        proc sql noprint;
115        select count(distinct &cVar) into :nrows from &data;
116        quit;
117
118        /* One-way histogram */
119        proc univariate data=&data noprint;
120        var &avar;
121        class &cVar;
122        histogram &avar / nrows=&nrows;
123        run;
124
125        %end;
126        %else
127        %do;
128
129        /* One-way histogram of each class variable */
130
131        %do i=1 %to %eval(&numCVars);
132        %let cVar=%scan(%str(&classVar), &i, %str( ), %str(q));
133
134        proc sql noprint;
135        select count(distinct &cVar) into :nrows from &data;
136        quit;
137
138        proc univariate data=&data noprint;
139        var &avar;
140        class &cVar;
141        histogram &avar / nrows=&nrows;
142        run;
143
144        %end;
145
146        /* Two-way histogram */
147                        %let cVar1=%scan(%str(&classVar), 1, %str( ),
148        %str(q));
149        %let cVar2=%scan(%str(&classVar), 2, %str( ), %str(q));
150
151        proc sql noprint;
152        select count(distinct &cVar1) into :nrows from &data;
153        quit;
154
155        proc sql noprint;
156        select count(distinct &cVar2) into :ncols from &data;
```

5/25/25, 1:20 PM Log: poorva-Data Exploration.clk

```
157        quit;
158
159        proc univariate data=&data noprint;
160        var &avar;
161        class &cVar1 &cVar2;
162        histogram &avar / nrows=&nrows ncols=&ncols;
163        run;
164
165        %end;
166        %end;
167        %mend DEHisto;
168
169        %scatterPlotMatrix(xVars=Age Height Weight,
170        title="Scatter plot matrix grouped by Sex", groupVar=Sex);
NOTE: PROCEDURE SGSCATTER used (Total process time):
      real time           0.15 seconds
      user cpu time       0.05 seconds
      system cpu time     0.00 seconds
      memory              9483.46k
      OS Memory           31392.00k
      Timestamp           05/25/2025 07:49:00 AM
      Step Count                       133  Switch Count  2
      Page Faults                      0
      Page Reclaims                    2111
      Page Swaps                       0
      Voluntary Context Switches       687
      Involuntary Context Switches     2
      Block Input Operations           0
      Block Output Operations          728

NOTE: There were 19 observations read from the data set SASHELP.CLASS.

171        %DEHisto(data=SASHELP.CLASS, avar=Age Height Weight, classVar=Sex);
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.01 seconds
      memory              5348.46k
      OS Memory           36260.00k
      Timestamp           05/25/2025 07:49:00 AM
      Step Count                       134  Switch Count  0
      Page Faults                      0
      Page Reclaims                    26
      Page Swaps                       0
      Voluntary Context Switches       0
      Involuntary Context Switches     0
      Block Input Operations           0
```

about:blank

5/25/25, 1:20 PM Log: poorva-Data Exploration.clk

```
      Block Output Operations          8


NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.21 seconds
      user cpu time       0.07 seconds
      system cpu time     0.01 seconds
      memory              3866.34k
      OS Memory           32296.00k
      Timestamp           05/25/2025 07:49:01 AM
      Step Count                       135  Switch Count  0
      Page Faults                      0
      Page Reclaims                    922
      Page Swaps                       0
      Voluntary Context Switches       483
      Involuntary Context Switches     6
      Block Input Operations           0
      Block Output Operations          944


172
173        OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
```

Category 4

1. **Creating a Demographic Summary Table**:

```
proc sql;
   create table demog_summary as
   select gender, count(*) as count, mean(age) as avg_age
   from demog
   group by gender;
quit;
```

SOLU 1:

5/25/25, 5:35 PM Results: WORK.DEMOG

Obs	name	gender	age
1	levy	M	45
2	lavina	F	32
3	laksh	M	58
4	nisha	F	41
5	nishant	M	29
6	navya	F	67
7	rohit	M	52
8	rohini	F	38
9	ronit	M	44
10	vishaka	F	55

5/25/25, 5:37 PM Code: poorva- demographic summary.sas

```
proc sql;
    create table demog_summary as
    select gender, count(*) as count, mean(age) as avg_age
    from demog
    group by gender;
quit;
```

5/25/25, 5:39 PM Results: WORK.DEMOG_SUMMARY

Obs	gender	count	avg_age
1	F	5	46.6
2	M	5	45.6

5/25/25, 5:38 PM Log: poorva- demographic summary.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table demog_summary as
71             select gender, count(*) as count, mean(age) as avg_age
72             from demog
73             group by gender;
NOTE: Table WORK.DEMOG_SUMMARY created, with 2 rows and 3 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5607.78k
      OS Memory           28328.00k
      Timestamp           05/25/2025 12:07:24 PM
      Step Count                        168  Switch Count  2
      Page Faults                       0
      Page Reclaims                     250
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280


75
76         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
86
```

2. Merging Datasets**:**

```
1  proc sql;
2      create table merged_data as
3      select a.*, b.treatment
4      from patients a
5      left join treatment b
6      on a.patient_id = b.patient_id;
7  quit;
```

SOLU 2:

5/25/25, 2:01 PM — Results: WORK.PATIENTS

Obs	patient_id	name	age
1	101	Alice	23
2	102	Bob	35
3	103	Carol	29

5/25/25, 2:02 PM — Results: WORK.TREATMENT

Obs	patient_id	treatment
1	101	AMOXCILL
2	102	BENICILL

5/25/25, 2:02 PM — Code: poorva-merging datasets.sas

```
proc sql;
   create table merged_data as
   select a.*, b.treatment
   from patients a
   left join treatment b
   on a.patient_id = b.patient_id;
quit;
```

5/25/25, 2:04 PM — Results: WORK.MERGED_DATA

Obs	patient_id	name	age	treatment
1	101	Alice	23	AMOXCILL
2	102	Bob	35	BENICILL
3	103	Carol	29	

5/25/25, 2:03 PM — Log: poorva-merging datasets.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70            create table merged_data as
71            select a.*, b.treatment
72            from patients a
73            left join treatment b
74            on a.patient_id = b.patient_id;
NOTE: Table WORK.MERGED_DATA created, with 3 rows and 4 columns.

75         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5905.46k
      OS Memory           29868.00k
      Timestamp           05/25/2025 08:31:03 AM
      Step Count                        197  Switch Count  2
      Page Faults                       0
      Page Reclaims                     262
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280

76
77
78
79         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
89
```

3. Filtering Data:

```
1  proc sql;
2     create table filtered_data as
3     select *
4     from lab_results
5     where test_date between '01JAN2023'd and '31DEC2023'd;
6  quit;
```

SOLU 3:

5/25/25, 2:22 PM — Results: WORK.LAB_RESULTS

Obs	patient_id	test_name	test_result	test_date
1	P001	Glucose	Normal	01JAN2023
2	P002	Hemoglob	Low	15FEB2023
3	P003	Choleste	High	12MAR2023
4	P004	Glucose	Normal	05MAY2022
5	P005	Hemoglob	Normal	31DEC2024

5/25/25, 2:23 PM — Code: POORVA- FILTERING DATAsas.sas

```
proc sql;
    create table filtered_data as
    select *
    from lab_results
    where test_date between '01JAN2023'd and '31DEC2023'd;
quit;
```

5/25/25, 2:24 PM — Results: WORK.FILTERED_DATA

Obs	patient_id	test_name	test_result	test_date
1	P001	Glucose	Normal	01JAN2023
2	P002	Hemoglob	Low	15FEB2023
3	P003	Choleste	High	12MAR2023

5/25/25, 2:24 PM — Log: POORVA- FILTERING DATAsas.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table filtered_data as
71             select *
72             from lab_results
73             where test_date between '01JAN2023'd and '31DEC2023'd;
NOTE: Table WORK.FILTERED_DATA created, with 3 rows and 4 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.00 seconds
      user cpu time           0.00 seconds
      system cpu time         0.01 seconds
      memory                  5647.59k
      OS Memory               29864.00k
      Timestamp               05/25/2025 08:51:46 AM
      Step Count                        251  Switch Count  2
      Page Faults                       0
      Page Reclaims                     116
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

4. **Calculating Summary Statistics**:

```
proc sql;
    create table summary_stats as
    select treatment, mean(blood_pressure) as avg_bp, std(blood_pressure) as std_bp
    from vitals
    group by treatment;
quit;
```

SOLU 4:

5/25/25, 3:21 PM — Results: WORK.VITALS

Obs	treatment	blood_pressure
1	A	120
2	A	125
3	A	122
4	B	130
5	B	128
6	B	135
7	C	115
8	C	118
9	C	117

5/25/25, 3:29 PM — Code: poorva-calculating summary statistic.sas

```
proc sql;
    create table summary_stats as
    select treatment, mean(blood_pressure) as avg_bp, std(blood_pressure) as std_bp
    from vitals
    group by treatment;
quit;
```

5/25/25, 3:30 PM — Results: WORK.SUMMARY_STATS

Obs	treatment	avg_bp	std_bp
1	A	122.333	2.51661
2	B	131.000	3.60555
3	C	116.667	1.52753

5/25/25, 3:31 PM — Log: poorva-calculating summary statistic.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table summary_stats as
71             select treatment, mean(blood_pressure) as avg_bp, std(blood_pressure) as std_bp
72             from vitals
73             group by treatment;
NOTE: Table WORK.SUMMARY_STATS created, with 3 rows and 3 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.00 seconds
      user cpu time           0.00 seconds
      system cpu time         0.00 seconds
      memory                  5602.37k
      OS Memory               27048.00k
      Timestamp               05/25/2025 09:59:07 AM
      Step Count                        42  Switch Count  2
      Page Faults                       0
      Page Reclaims                     248
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

```
1  proc sql;
2     create table visit_count as
3     select patient_id, count(*) as visit_count
4     group by patient_id;
5  quit;
6
```

5. Creating a Patient Visit Count:

SOLU 5:

5/25/25, 3:48 PM — Results: WORK.VISITS

Obs	patient_id	visit_date
1	P001	01JAN2024
2	P001	15JAN2024
3	P002	02JAN2024
4	P003	05JAN2024
5	P001	20JAN2024
6	P002	10JAN2024

5/25/25, 3:51 PM — Code: POORVA- COUNT VISITS.sas

```
proc sql;
    create table visit_count as
    select patient_id, count(*) as visit_count
    from visits
    group by patient_id;
quit;
```

5/25/25, 3:52 PM — Results: WORK.VISIT_COUNT

Obs	patient_id	visit_count
1	P001	3
2	P002	2
3	P003	1

5/25/25, 3:51 PM — Log: POORVA- COUNT VISITS.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70              create table visit_count as
71              select patient_id, count(*) as visit_count
72              from visits
73              group by patient_id;
NOTE: Table WORK.VISIT_COUNT created, with 3 rows and 2 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.00 seconds
      user cpu time           0.00 seconds
      system cpu time         0.00 seconds
      memory                  5604.25k
      OS Memory               26792.00k
      Timestamp               05/25/2025 10:20:48 AM
      Step Count                        72  Switch Count  2
      Page Faults                       0
      Page Reclaims                     246
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272

75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

6. **Identifying Missing Data**:

```
1  proc sql;
2     create table missing_data as
3     select *
4     from lab_results
5     where result is missing;
6  quit;
```

SOLU 6:

5/25/25, 4:17 PM Results: WORK.LAB_RESULTS

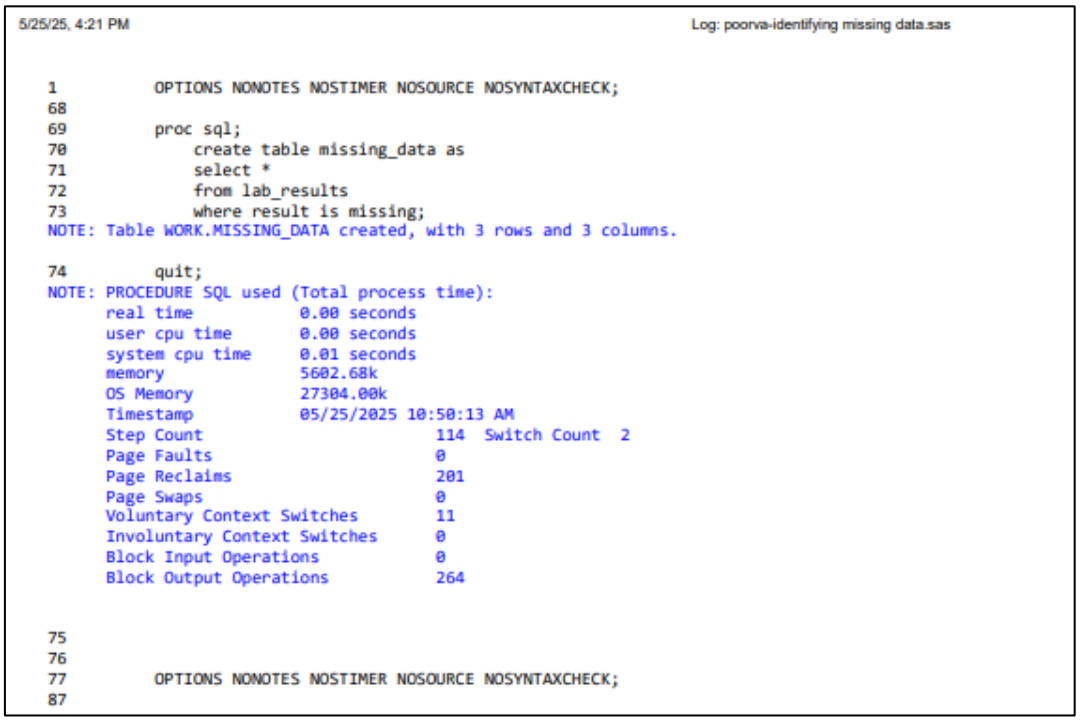
Obs	patient_id	test_name	result
1	P001	Glucose	5.6
2	P002	Glucose	.
3	P003	Hemoglob	13.2
4	P004	Hemoglob	.
5	P005	Choleste	180.0
6	P006	Choleste	.

5/25/25, 4:20 PM Code: poorva-identifying missing data.sas

```
proc sql;
    create table missing_data as
    select *
    from lab_results
    where result is missing;
quit;
```

5/25/25, 4:22 PM Results: WORK.MISSING_DATA

Obs	patient_id	test_name	result
1	P002	Glucose	.
2	P004	Hemoglob	.
3	P006	Choleste	.



```
5/25/25, 4:21 PM                                                    Log: poorva-identifying missing data.sas

1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table missing_data as
71             select *
72             from lab_results
73             where result is missing;
NOTE: Table WORK.MISSING_DATA created, with 3 rows and 3 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.01 seconds
      memory              5602.68k
      OS Memory           27304.00k
      Timestamp           05/25/2025 10:50:13 AM
      Step Count                        114  Switch Count  2
      Page Faults                       0
      Page Reclaims                     201
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

7. **Generating a Treatment Efficacy Report**:

```
proc sql;
   create table efficacy_report as
   select treatment, count(*) as num_patients, mean(response) as avg_response
   from efficacy
   group by treatment;
quit;
```

SOLU 7:

5/25/25, 4:28 PM — Results: WORK.EFFICACY

Obs	patient_id	treatment	response
1	P001	A	75
2	P002	A	80
3	P003	A	70
4	P004	B	65
5	P005	B	60
6	P006	C	90
7	P007	C	85

5/25/25, 4:31 PM — Code: POORVA- TREATMENT EFFICACY.sas

```
proc sql;
    create table efficacy_report as
    select treatment, count(*) as num_patients, mean(response) as avg_response
    from efficacy
    group by treatment;
quit;
```

5/25/25, 4:32 PM — Results: WORK.EFFICACY_REPORT

Obs	treatment	num_patients	avg_response
1	A	3	75.0
2	B	2	62.5
3	C	2	87.5

5/25/25, 4:31 PM — Log: POORVA- TREATMENT EFFICACY.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table efficacy_report as
71             select treatment, count(*) as num_patients, mean(response) as avg_response
72             from efficacy
73             group by treatment;
NOTE: Table WORK.EFFICACY_REPORT created, with 3 rows and 3 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time               0.00 seconds
      user cpu time           0.00 seconds
      system cpu time         0.00 seconds
      memory                  5607.93k
      OS Memory               27304.00k
      Timestamp               05/25/2025 11:00:56 AM
      Step Count                        138  Switch Count  2
      Page Faults                       0
      Page Reclaims                     250
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

8. **Creating an Adverse Events Summary**:

```
1  proc sql;
2     create table ae_summary as
3     select event_type, count(*) as num_events
4     from adverse_events
5     group by event_type;
6  quit;
7
```

SOLU 8:

5/25/25, 4:37 PM — Results: WORK.ADVERSE_EVENTS

Obs	patient_id	event_type	event_date
1	P001	Rash	01JAN2024
2	P002	Nausea	02JAN2024
3	P003	Rash	03JAN2024
4	P004	Headache	04JAN2024
5	P005	Nausea	05JAN2024
6	P006	Rash	06JAN2024
7	P007	Headache	07JAN2024

5/25/25, 4:43 PM — Code: POORVA-ADVERSE EVENT.sas

```
proc sql;
    create table ae_summary as
    select event_type, count(*) as num_events
    from adverse_events
    group by event_type;
quit;
```

5/25/25, 4:44 PM — Results: WORK.AE_SUMMARY

Obs	event_type	num_events
1	Headache	2
2	Nausea	2
3	Rash	3

```
5/25/25, 4:44 PM                                    Log: POORVA-ADVERSE EVENT.sas

1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table ae_summary as
71             select event_type, count(*) as num_events
72             from adverse_events
73             group by event_type;
NOTE: Table WORK.AE_SUMMARY created, with 3 rows and 2 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5601.65k
      OS Memory           26792.00k
      Timestamp           05/25/2025 11:13:16 AM
      Step Count                        42  Switch Count  2
      Page Faults                       0
      Page Reclaims                     263
      Page Swaps                        0
      Voluntary Context Switches        11
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

9. **Listing Patients with Specific Conditions**:

```
1  proc sql;
2     create table specific_conditions as
3     select patient_id, condition
4     from conditions
5     where condition in ('Diabetes', 'Hypertension');
6  quit;
7
```

SOLU 9:

5/25/25, 5:01 PM — Results: WORK.CONDITIONS

Obs	patient_id	condition
1	P001	Diabetes
2	P002	Asthma
3	P003	Hypertension
4	P004	Obesity
5	P005	Diabetes
6	P006	Hypertension
7	P007	Anxiety

5/25/25, 5:04 PM — Code: POORVA-SPECIFIC CONDITIONS.sas

```
proc sql;
    create table specific_conditions as
    select patient_id, condition
    from conditions
    where condition in ('Diabetes', 'Hypertension');
quit;
```

5/25/25, 5:05 PM — Results: WORK.SPECIFIC_CONDITIONS

Obs	patient_id	condition
1	P001	Diabetes
2	P003	Hypertension
3	P005	Diabetes
4	P006	Hypertension

5/25/25, 5:04 PM — Log: POORVA-SPECIFIC CONDITIONS.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table specific_conditions as
71             select patient_id, condition
72             from conditions
73             where condition in ('Diabetes', 'Hypertension');
NOTE: Table WORK.SPECIFIC_CONDITIONS created, with 4 rows and 2 columns.

74         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5602.56k
      OS Memory           27304.00k
      Timestamp           05/25/2025 11:33:45 AM
      Step Count                        96  Switch Count  2
      Page Faults                       0
      Page Reclaims                     188
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           264


75
76
77         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
87
```

10. **Calculating Time to Event**:

```
proc sql;
    create table time_to_event as
    select patient_id, event_date - start_date as time_to_event
    from events;
quit;
```

SOLUN 10:

5/25/25, 5:24 PM — Results: WORK.EVENTS

Obs	patient_id	event_date	start_date
1	P001	01JAN2024	15DEC2023
2	P002	15FEB2024	01JAN2024
3	P003	10MAR2024	20FEB2024
4	P004	25APR2024	10APR2024
5	P005	05JUN2024	01MAY2024

5/25/25, 5:26 PM — Code: POORVA-TIME TO EVENT.sas

```
proc sql;
    create table time_to_event as
    select patient_id, event_date - start_date as time_to_event
    from events;
quit;
```

5/25/25, 5:27 PM — Results: WORK.TIME_TO_EVENT

Obs	patient_id	time_to_event
1	P001	17
2	P002	45
3	P003	19
4	P004	15
5	P005	35

5/25/25, 5:27 PM Log: POORVA-TIME TO EVENT.sas

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         proc sql;
70             create table time_to_event as
71             select patient_id, event_date - start_date as time_to_event
72             from events;
NOTE: Table WORK.TIME_TO_EVENT created, with 5 rows and 2 columns.

73         quit;
NOTE: PROCEDURE SQL used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              5601.56k
      OS Memory           27816.00k
      Timestamp           05/25/2025 11:56:05 AM
      Step Count                        126  Switch Count  2
      Page Faults                       0
      Page Reclaims                     194
      Page Swaps                        0
      Voluntary Context Switches        9
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


74
75         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
85
```

Conclusion

Through this project, I gained practical exposure to working with **SAS Studio**, including data import, cleaning, transformation, and statistical analysis. The project not only strengthened my understanding of **clinical data structures** but also gave me the ability to:

- Organize and analyze datasets efficiently.
- Apply **Good Documentation Practices (ALCOA+)** in recording data activities.
- Use SAS tools for generating **reliable and reproducible reports**.
- Build a foundation for applying **biostatistical methods** in clinical research.

This project demonstrates my ability as a fresher to quickly learn industry-standard software and apply it to **data management and analysis tasks**, preparing me for roles such as **Clinical Research Coordinator, Clinical Data Manager, or Pharmacovigilance Associate**.