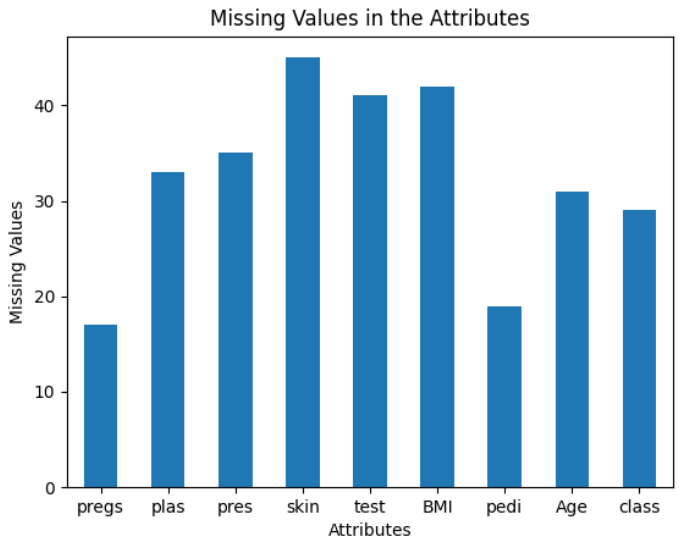
IC272-Lab2-Report
Name : Gajraj Singh Chouhan
Roll No. B19130
Mobile No: +91-9351159849

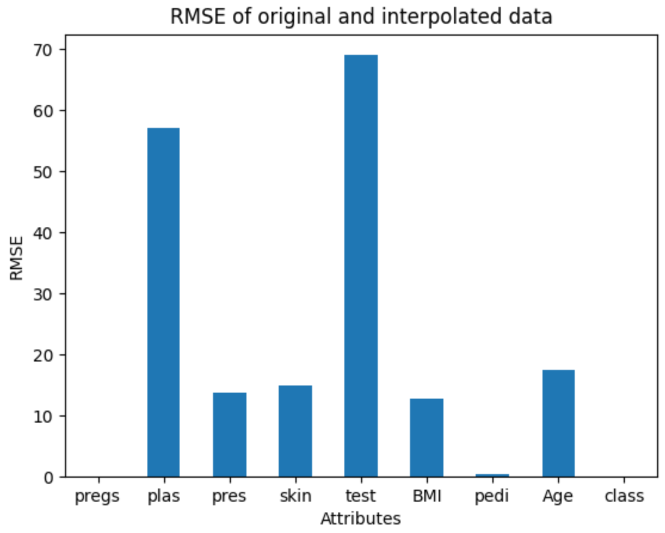
- **Q1**

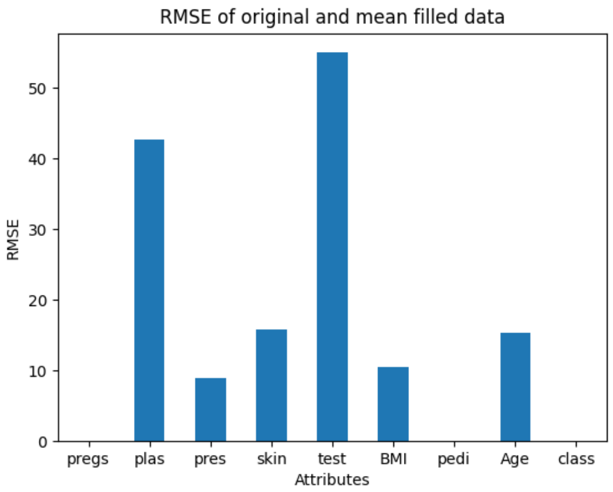


  - Graph shows the missing values in each attribute.
  - ‘Skin’ attribute has the largest number of missing values.
- **Q2**
  - Total 39+21=60 tuples were deleted.
  - Indices (0-based) of rows deleted in “a” and “b” part respectively:
    - 1, 39, 40, 53, 54, 83, 89, 103, 125, 136, 145, 210, 211, 212, 213, 249, 250, 254, 280, 281, 284, 314, 321, 335, 429, 430, 449, 450, 451, 471, 472, 473, 474, 718, 719, 720, 721, 753, 766
    - 8, 13, 28, 29, 35, 62, 92, 95, 107, 110, 130, 131, 132, 133, 149, 182, 188, 218, 308, 746, 748
- **Q3**
  - Total number of missing values is now 69.

Missing Values	
pregs	0
plas	12
pres	9
skin	8
test	8
BMI	12
pedi	2
Age	18
class	0

- **Q4**





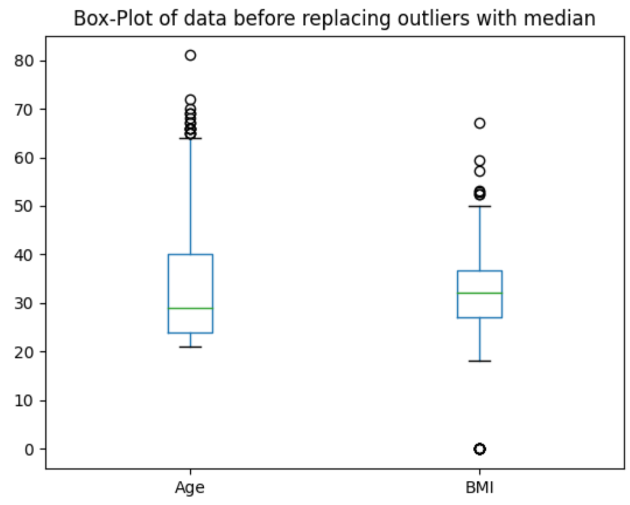
- Graph shows root mean square error for each attribute with the original data.
- Since ‘test’ is showing the largest error, we can expect it to have the largest difference in the properties from original data in both of these cases.
- After filling the missing values, we can compare the original data and the new one for both the cases.

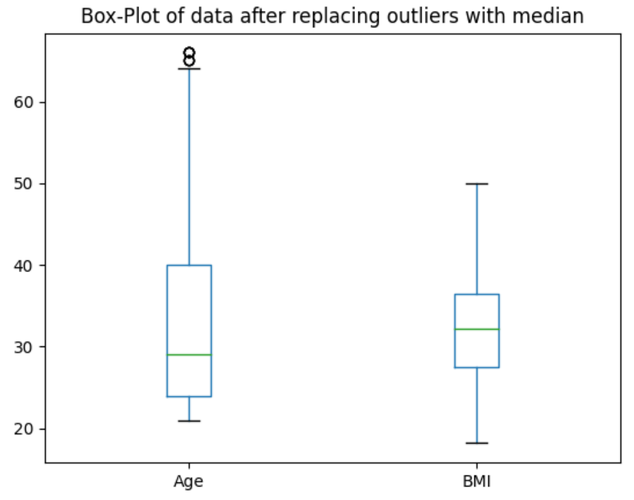
All the properties of original data and after data changes:

	Properties of Original .csv file				After filling with Mean of Columns				After filling with Interpolation of Columns			
	mean	median	mode	standard-dev	mean	median	mode	standard-dev	mean	median	mode	standard-dev
pregs	3.84505	3	(1,)	3.36738	3.88559	3	(1.0,)	3.37148	3.88559	3	(1.0,)	3.37148
plas	120.89453	117	(99, 100)	31.95180	120.6666	118	(99.0, 100.0)	30.96829	120.34958	117	(99.0, 100.0)	31.25270
pres	69.10547	72	(70,)	19.34320	69.00143	72	(70.0,)	19.67745	69.10946	72	(70.0,)	19.72204
skin	20.53646	23	(0,)	15.94183	20.34857	23	(0.0,)	15.93494	20.39266	23	(0.0,)	15.96456
test	79.79948	30.5	(0,)	115.16895	77.81429	36	(0.0,)	110.52946	77.35523	27	(0.0,)	110.67775
BMI	31.99258	32	(32.0,)	7.87903	32.00934	32.00934	(32.0,)	7.75927	32.04633	32.25	(32.0,)	7.78711
pedi	0.47188	0.3725	(0.254, 0.258)	0.33111	0.47604	0.3825	(0.254, 0.258)	0.33296	0.47732	0.3825	(0.254, 0.258)	0.33401
Age	33.24089	29	(22,)	11.75257	33.09420	29	(22.0,)	11.51153	33.21610	29	(22.0,)	11.64442
class	0.34896	0	(0,)	0.47664	0.34322	0	(0.0,)	0.47478	0.34322	0	(0.0,)	0.47478

- ‘Mode’ has not changed from the original data.
- ‘Median’ also changed slightly.
- ‘Test’ attribute shows greater change in median than rest of the attributes. E.g. In the filled data the median value for ‘test’ is +6 than original data.
- ‘Mean’ has only changed slightly (order of 10^{-1}).
- ‘Standard Deviation’ has changed slightly except the ‘test’ attribute shows **-5** change.

- **Q5**





-
- Graphs show Box-Plot of ‘Age’ and ‘BMI’, showing the outliers and an attempt to reduce them by replacing them with median.
- ‘BMI’ attribute - It’s mean, mode & median were very close i.e. it’s very close to **symmetric** - so when we changed all of its outliers with median, the standard deviation decreases and reduces the number of outliers to zero.
- ‘Age’ attribute - This attribute’s data is right skewed, the Q2 (median) isn’t symmetric about Q3 and Q1 as seen in the box plot and it’s leaning towards left (Q1). So when we will be replacing the outliers with the median, as it is not symmetric due to skewness, some outliers will remain there.