Lab Assignment 2

77

72

74

72

73

73

79

70

79

79

74

77

77

70

76

Math\_Score\n Reading\_Score Writing\_Score Placement\_Score \

NaN 75.0

69.0

71.0

69.0

75.0

69.0

65.0

72.0

74.0

73.0

66.0

75.0

72.0

71.0

67

80

61

74

80

70

68

78

74

63

72

78

69

66

65

66

70

78

78

76

60

68

64

80

80

72

63

71

74

76

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

AIM:-Create an “Academic performance” dataset of students and perform the following operations using

Python.

1. Scan all variables for missing values and inconsistencies. If there are missing values and/or

inconsistencies, use any of the suitable techniques to deal with them.

1. Scan all numeric variables for outliers. If there are outliers, use any of the suitable

techniques to deal with them.

1. Apply data transformations on at least one of the variables. The purpose of this

transformation should be one of the following reasons: to change the scale for better

understanding of the variable, to convert a non-linear relation into a linear one, or to

decrease the skewness and convert the distribution into a normal distribution.

Reason and document your approach properly.

import pandas as pd import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

from scipy import stats file\_path=r"C:\Users\System21\Desktop\studentdata.csv" df=pd.read\_csv(file\_path)

df.head(20)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 15 | 61 | 80 | 74.0 | 80 |
| 16 | 61 | 63 | 70.0 | 71 |
| 17 | 76 | 74 | 67.0 | 73 |
| 18 | 75 | 64 | 66.0 | 76 |
|  | 19 | 69 | 74 | 67.0 | 72 |

Club\_Join\_Date Placement\_Offer\_Count

0 2020 93

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

7 2019 95

8 2018 100

9 2020 93

10 2021 89

11 2019 90

12 2018 92

13 2018 89

14 2019 100

15 2020 97

16 2019 95

17 2021 95

18 2021 91

19 2018 94

Math\_Score\n Reading\_Score Writing\_Score False False True

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

False False False

Placement\_Score

False False False False False False False False False False False False False False False False False False False

\

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

df.isnull()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 19 | False | False | False | False |
| 20 | False | False | False | False |
| 21 | False | False | False | False |
| 22 | False | False | False | False |
| 23 | False | False | False | False |
| 24 | False | False | False | False |
| 25 | False | False | False | False |
| 26 | False | False | False | False |
| 27 | False | False | False | False |
|  | 28 | False | False | False | False |

Club\_Join\_Date Placement\_Offer\_Count

1. False False
2. False False
3. False False
4. False False
5. False False
6. False False
7. False False
8. False False
9. False False
10. False False
11. False False
12. False False
13. False False
14. False False
15. False False
16. False False
17. False False
18. False False
19. False False
20. False False
21. False False
22. False False
23. False False
24. False False
25. False False
26. False False
27. False False
28. False False
29. False False

Math\_Score\n Reading\_Score Writing\_Score

66 67 NaN

70 80 75.0

78 61 69.0

78 74 71.0

Placement\_Score

77

72

74

72

\

0

1

2

3

series1 = pd.notnull(df["Reading\_Score"]) df[series1]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 4 | 76 | 80 | 69.0 | 73 |
| 5 | 60 | 70 | 75.0 | 73 |
| 6 | 68 | 68 | 69.0 | 79 |
| 7 | 64 | 78 | 65.0 | 70 |
| 8 | 80 | 74 | 72.0 | 79 |
| 9 | 80 | 63 | 74.0 | 79 |
| 10 | 72 | 72 | 73.0 | 74 |
| 11 | 63 | 78 | 66.0 | 77 |
| 12 | 71 | 69 | 75.0 | 77 |
| 13 | 74 | 66 | 72.0 | 70 |
| 14 | 76 | 65 | 71.0 | 76 |
| 15 | 61 | 80 | 74.0 | 80 |
| 16 | 61 | 63 | 70.0 | 71 |
| 17 | 76 | 74 | 67.0 | 73 |
| 18 | 75 | 64 | 66.0 | 76 |
| 19 | 69 | 74 | 67.0 | 72 |
| 20 | 79 | 76 | 72.0 | 71 |
| 21 | 80 | 70 | 69.0 | 73 |
| 22 | 71 | 63 | 74.0 | 80 |
| 23 | 62 | 71 | 69.0 | 71 |
| 24 | 75 | 63 | 74.0 | 78 |
| 25 | 73 | 60 | 73.0 | 75 |
| 26 | 71 | 70 | 65.0 | 72 |
| 27 | 64 | 68 | 66.0 | 71 |
|  | 28 | 68 | 68 | 75.0 | 75 |

Club\_Join\_Date Placement\_Offer\_Count

0 2020 93

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

7 2019 95

8 2018 100

9 2020 93

10 2021 89

11 2019 90

12 2018 92

13 2018 89

14 2019 100

15 2020 97

16 2019 95

17 2021 95

18 2021 91

19 2018 94

20 2020 99

21 2019 77

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 22 | 2019 | 89 |
| 23 | 2021 | 77 |
| 24 | 2018 | 85 |
| 25 | 2018 | 84 |
| 26 | 2018 | 84 |
| 27 | 2020 | 84 |
|  | 28 | 2018 | 97 |

print(df.columns)

Math\_Score\n Reading\_Score Writing\_Score

66 67 0.0

70 80 75.0

78 61 69.0

78 74 71.0

76 80 69.0

60 70 75.0

68 68 69.0

64 78 65.0

80 74 72.0

80 63 74.0

72 72 73.0

63 78 66.0

71 69 75.0

74 66 72.0

76 65 71.0

61 80 74.0

61 63 70.0

76 74 67.0

75 64 66.0

69 74 67.0

79 76 72.0

80 70 69.0

71 63 74.0

62 71 69.0

75 63 74.0

73 60 73.0

71 70 65.0

64 68 66.0

68 68 75.0

Placement\_Score

77

72

74

72

73

73

79

70

79

79

74

77

77

70

76

80

71

73

76

72

71

73

80

71

78

75

72

71

75

\

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

Index(['Math\_Score\n ', 'Reading\_Score', 'Writing\_Score', 'Placement\_Score',

'Club\_Join\_Date', 'Placement\_Offer\_Count'],

dtype='object')

ndf=df ndf.fillna(0)

Club\_Join\_Date Placement\_Offer\_Count

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 0 | 2020 | 93 |
| 1 | 2019 | 75 |
| 2 | 2019 | 90 |
| 3 | 2020 | 91 |
| 4 | 2021 | 88 |
| 5 | 2021 | 75 |
| 6 | 2021 | 100 |
| 7 | 2019 | 95 |
| 8 | 2018 | 100 |
| 9 | 2020 | 93 |
| 10 | 2021 | 89 |
| 11 | 2019 | 90 |
| 12 | 2018 | 92 |
| 13 | 2018 | 89 |
| 14 | 2019 | 100 |
| 15 | 2020 | 97 |
| 16 | 2019 | 95 |
| 17 | 2021 | 95 |
| 18 | 2021 | 91 |
| 19 | 2018 | 94 |
| 20 | 2020 | 99 |
| 21 | 2019 | 77 |
| 22 | 2019 | 89 |
| 23 | 2021 | 77 |
| 24 | 2018 | 85 |
| 25 | 2018 | 84 |
| 26 | 2018 | 84 |
| 27 | 2020 | 84 |
|  | 28 | 2018 | 97 |

m\_v=df['Reading\_Score'].mean() df['Reading\_Score'].fillna(value=m\_v, inplace=True) df

77

72

74

72

73

73

79

70

79

79

74

77

77

70

76

Math\_Score\n Reading\_Score Writing\_Score Placement\_Score \

NaN 75.0

69.0

71.0

69.0

75.0

69.0

65.0

72.0

74.0

73.0

66.0

75.0

72.0

71.0

67

80

61

74

80

70

68

78

74

63

72

78

69

66

65

66

70

78

78

76

60

68

64

80

80

72

63

71

74

76

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 15 | 61 | 80 | 74.0 | 80 |
| 16 | 61 | 63 | 70.0 | 71 |
| 17 | 76 | 74 | 67.0 | 73 |
| 18 | 75 | 64 | 66.0 | 76 |
| 19 | 69 | 74 | 67.0 | 72 |
| 20 | 79 | 76 | 72.0 | 71 |
| 21 | 80 | 70 | 69.0 | 73 |
| 22 | 71 | 63 | 74.0 | 80 |
| 23 | 62 | 71 | 69.0 | 71 |
| 24 | 75 | 63 | 74.0 | 78 |
| 25 | 73 | 60 | 73.0 | 75 |
| 26 | 71 | 70 | 65.0 | 72 |
| 27 | 64 | 68 | 66.0 | 71 |
|  | 28 | 68 | 68 | 75.0 | 75 |

Club\_Join\_Date Placement\_Offer\_Count

0 2020 93

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

7 2019 95

8 2018 100

9 2020 93

10 2021 89

11 2019 90

12 2018 92

13 2018 89

14 2019 100

15 2020 97

16 2019 95

17 2021 95

18 2021 91

19 2018 94

20 2020 99

21 2019 77

22 2019 89

23 2021 77

24 2018 85

25 2018 84

26 2018 84

27 2020 84

28 2018 97

Math\_Score\n Reading\_Score Writing\_Score

70 80 75.0

Placement\_Score

72

\

1

ndf.dropna()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 2 | 78 | 61 | 69.0 | 74 |
| 3 | 78 | 74 | 71.0 | 72 |
| 4 | 76 | 80 | 69.0 | 73 |
| 5 | 60 | 70 | 75.0 | 73 |
| 6 | 68 | 68 | 69.0 | 79 |
| 7 | 64 | 78 | 65.0 | 70 |
| 8 | 80 | 74 | 72.0 | 79 |
| 9 | 80 | 63 | 74.0 | 79 |
| 10 | 72 | 72 | 73.0 | 74 |
| 11 | 63 | 78 | 66.0 | 77 |
| 12 | 71 | 69 | 75.0 | 77 |
| 13 | 74 | 66 | 72.0 | 70 |
| 14 | 76 | 65 | 71.0 | 76 |
| 15 | 61 | 80 | 74.0 | 80 |
| 16 | 61 | 63 | 70.0 | 71 |
| 17 | 76 | 74 | 67.0 | 73 |
| 18 | 75 | 64 | 66.0 | 76 |
| 19 | 69 | 74 | 67.0 | 72 |
| 20 | 79 | 76 | 72.0 | 71 |
| 21 | 80 | 70 | 69.0 | 73 |
| 22 | 71 | 63 | 74.0 | 80 |
| 23 | 62 | 71 | 69.0 | 71 |
| 24 | 75 | 63 | 74.0 | 78 |
| 25 | 73 | 60 | 73.0 | 75 |
| 26 | 71 | 70 | 65.0 | 72 |
| 27 | 64 | 68 | 66.0 | 71 |
|  | 28 | 68 | 68 | 75.0 | 75 |

Club\_Join\_Date Placement\_Offer\_Count

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

7 2019 95

8 2018 100

9 2020 93

10 2021 89

11 2019 90

12 2018 92

13 2018 89

14 2019 100

15 2020 97

16 2019 95

17 2021 95

18 2021 91

19 2018 94

20 2020 99

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 21 | 2019 | 77 |
| 22 | 2019 | 89 |
| 23 | 2021 | 77 |
| 24 | 2018 | 85 |
| 25 | 2018 | 84 |
| 26 | 2018 | 84 |
| 27 | 2020 | 84 |
|  | 28 | 2018 | 97 |

Math\_Score\n Reading\_Score Writing\_Score

66 67 NaN

70 80 75.0

78 61 69.0

78 74 71.0

76 80 69.0

60 70 75.0

68 68 69.0

64 78 65.0

80 74 72.0

80 63 74.0

72 72 73.0

63 78 66.0

71 69 75.0

74 66 72.0

76 65 71.0

61 80 74.0

61 63 70.0

76 74 67.0

75 64 66.0

69 74 67.0

79 76 72.0

80 70 69.0

71 63 74.0

62 71 69.0

75 63 74.0

73 60 73.0

71 70 65.0

64 68 66.0

68 68 75.0

Placement\_Score

77

72

74

72

73

73

79

70

79

79

74

77

77

70

76

80

71

73

76

72

71

73

80

71

78

75

72

71

75

\

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

ndf.dropna(how = 'all')

Club\_Join\_Date Placement\_Offer\_Count

0 2020 93

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 7 | 2019 | 95 |
| 8 | 2018 | 100 |
| 9 | 2020 | 93 |
| 10 | 2021 | 89 |
| 11 | 2019 | 90 |
| 12 | 2018 | 92 |
| 13 | 2018 | 89 |
| 14 | 2019 | 100 |
| 15 | 2020 | 97 |
| 16 | 2019 | 95 |
| 17 | 2021 | 95 |
| 18 | 2021 | 91 |
| 19 | 2018 | 94 |
| 20 | 2020 | 99 |
| 21 | 2019 | 77 |
| 22 | 2019 | 89 |
| 23 | 2021 | 77 |
| 24 | 2018 | 85 |
| 25 | 2018 | 84 |
| 26 | 2018 | 84 |
| 27 | 2020 | 84 |
|  | 28 | 2018 | 97 |

ndf.dropna(axis = 1)

2020

2019

2019

2020

2021

2021

2021

2019

2018

2020

2021

2019

2018

2018

2019

2020

2019

2021

2021

2018

2020

2019

2019

2021

Math\_Score\n Reading\_Score Placement\_Score Club\_Join\_Date \

77

72

74

72

73

73

79

70

79

79

74

77

77

70

76

80

71

73

76

72

71

73

80

71

67

80

61

74

80

70

68

78

74

63

72

78

69

66

65

80

63

74

64

74

76

70

63

71

66

70

78

78

76

60

68

64

80

80

72

63

71

74

76

61

61

76

75

69

79

80

71

62

0

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 24 | 75 | 63 | 78 | 2018 |
| 25 | 73 | 60 | 75 | 2018 |
| 26 | 71 | 70 | 72 | 2018 |
| 27 | 64 | 68 | 71 | 2020 |
|  | 28 | 68 | 68 | 75 | 2018 |

Placement\_Offer\_Count

0 93

1 75

2 90

3 91

4 88

5 75

6 100

7 95

8 100

9 93

10 89

11 90

12 92

13 89

14 100

15 97

16 95

17 95

18 91

19 94

20 99

21 77

22 89

23 77

24 85

25 84

26 84

27 84

28 97

Math\_Score\n Reading\_Score Writing\_Score

70 80 75.0

78 61 69.0

78 74 71.0

76 80 69.0

60 70 75.0

68 68 69.0

64 78 65.0

80 74 72.0

80 63 74.0

Placement\_Score

72

74

72

73

73

79

70

79

79

\

1

2

3

4

5

6

7

8

9

new\_data =ndf.dropna (axis = 0, how ='any') new\_data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 10 | 72 | 72 | 73.0 | 74 |
| 11 | 63 | 78 | 66.0 | 77 |
| 12 | 71 | 69 | 75.0 | 77 |
| 13 | 74 | 66 | 72.0 | 70 |
| 14 | 76 | 65 | 71.0 | 76 |
| 15 | 61 | 80 | 74.0 | 80 |
| 16 | 61 | 63 | 70.0 | 71 |
| 17 | 76 | 74 | 67.0 | 73 |
| 18 | 75 | 64 | 66.0 | 76 |
| 19 | 69 | 74 | 67.0 | 72 |
| 20 | 79 | 76 | 72.0 | 71 |
| 21 | 80 | 70 | 69.0 | 73 |
| 22 | 71 | 63 | 74.0 | 80 |
| 23 | 62 | 71 | 69.0 | 71 |
| 24 | 75 | 63 | 74.0 | 78 |
| 25 | 73 | 60 | 73.0 | 75 |
| 26 | 71 | 70 | 65.0 | 72 |
| 27 | 64 | 68 | 66.0 | 71 |
|  | 28 | 68 | 68 | 75.0 | 75 |

Club\_Join\_Date Placement\_Offer\_Count

1 2019 75

2 2019 90

3 2020 91

4 2021 88

5 2021 75

6 2021 100

7 2019 95

8 2018 100

9 2020 93

10 2021 89

11 2019 90

12 2018 92

13 2018 89

14 2019 100

15 2020 97

16 2019 95

17 2021 95

18 2021 91

19 2018 94

20 2020 99

21 2019 77

22 2019 89

23 2021 77

24 2018 85

25 2018 84

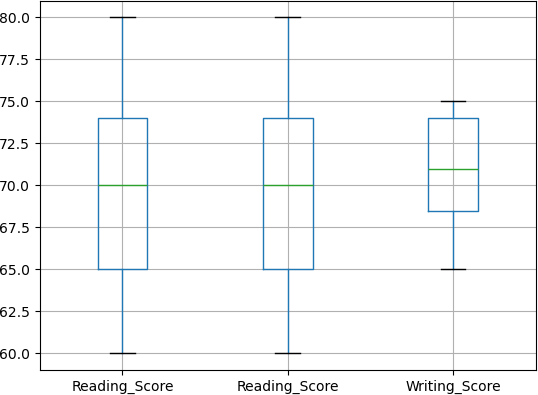
26 2018 84

27 2020 84

28 2018 97

col =['Reading\_Score', 'Reading\_Score', 'Writing\_Score'] df.boxplot(col)

<Axes: >



Prathmesh Nakhate B2\_13241