DA6213

Exercise #5

Missing Data

Kilger

This exercise will cover some of the basics in dealing with missing data. The data set is presented to you in this exercise as the last pages.

1. Produce means and standard deviations in a table for the five variables in the data set.

mean std

v1\_miss 3.128788 1.213193

v2\_miss 3.366412 1.144929

v3\_miss 1.976562 1.090148

v4\_miss 2.201550 1.134542

v5\_miss 2.178862 1.293315

1. Replace the missing values for all the variables in the data set using the mean substitution method. Place that table here…

The leftmost value is the index

v1\_miss v2\_miss v3\_miss v4\_miss v5\_miss

0 4.000000 3.000000 1.000000 3.00000 1.000000

1 2.000000 4.000000 1.000000 2.20155 3.000000

2 3.000000 3.000000 3.000000 3.00000 2.178862

3 2.000000 4.000000 1.000000 2.00000 2.000000

4 2.000000 4.000000 2.000000 5.00000 5.000000

5 4.000000 3.366412 1.000000 2.00000 1.000000

6 2.000000 4.000000 3.000000 1.00000 3.000000

7 2.000000 5.000000 3.000000 3.00000 1.000000

8 1.000000 5.000000 2.000000 3.00000 2.000000

9 3.000000 3.000000 1.000000 1.00000 3.000000

10 4.000000 1.000000 1.000000 1.00000 1.000000

11 5.000000 4.000000 1.000000 2.00000 1.000000

12 2.000000 4.000000 3.000000 4.00000 1.000000

13 4.000000 1.000000 1.000000 1.00000 1.000000

14 2.000000 3.000000 1.000000 1.00000 1.000000

15 3.128788 3.366412 1.976562 2.20155 2.178862

16 4.000000 4.000000 4.000000 5.00000 4.000000

17 3.000000 3.000000 2.000000 1.00000 1.000000

18 3.000000 2.000000 2.000000 3.00000 5.000000

19 1.000000 3.000000 1.000000 1.00000 1.000000

20 3.128788 4.000000 1.000000 3.00000 1.000000

21 3.000000 4.000000 2.000000 2.20155 4.000000

22 4.000000 3.000000 1.000000 3.00000 3.000000

23 5.000000 4.000000 1.000000 3.00000 2.000000

24 4.000000 4.000000 5.000000 4.00000 2.178862

25 4.000000 5.000000 1.000000 1.00000 1.000000

26 1.000000 2.000000 3.000000 2.00000 3.000000

27 3.000000 3.000000 1.000000 3.00000 1.000000

28 3.000000 3.366412 4.000000 3.00000 3.000000

29 1.000000 5.000000 3.000000 1.00000 1.000000

30 3.000000 3.000000 1.000000 1.00000 2.178862

31 3.000000 3.366412 1.976562 2.00000 2.000000

32 3.128788 3.366412 1.976562 2.20155 2.178862

33 2.000000 1.000000 1.000000 1.00000 1.000000

34 3.128788 3.366412 1.000000 3.00000 1.000000

35 4.000000 5.000000 3.000000 3.00000 2.000000

36 3.128788 3.366412 1.000000 2.20155 1.000000

37 4.000000 4.000000 1.000000 1.00000 1.000000

38 3.128788 3.000000 1.976562 3.00000 3.000000

39 3.000000 2.000000 2.000000 3.00000 4.000000

40 5.000000 2.000000 1.000000 1.00000 1.000000

41 1.000000 2.000000 2.000000 2.00000 2.178862

42 3.000000 3.000000 1.000000 1.00000 2.178862

43 4.000000 4.000000 4.000000 1.00000 2.000000

44 4.000000 4.000000 3.000000 2.00000 2.000000

45 5.000000 5.000000 3.000000 2.00000 3.000000

46 4.000000 4.000000 3.000000 1.00000 3.000000

47 5.000000 2.000000 2.000000 1.00000 1.000000

48 3.128788 2.000000 1.000000 2.20155 2.178862

49 3.000000 4.000000 1.000000 1.00000 1.000000

50 3.128788 1.000000 1.000000 2.20155 1.000000

51 3.000000 5.000000 3.000000 2.20155 1.000000

52 2.000000 2.000000 2.000000 2.00000 1.000000

53 4.000000 2.000000 2.000000 4.00000 1.000000

54 3.128788 3.366412 1.976562 2.20155 2.178862

55 3.128788 3.366412 1.976562 2.20155 2.178862

56 4.000000 3.366412 1.000000 1.00000 2.000000

57 3.128788 2.000000 3.000000 4.00000 3.000000

58 2.000000 3.000000 1.000000 1.00000 1.000000

59 4.000000 4.000000 1.976562 3.00000 2.178862

60 3.128788 4.000000 3.000000 3.00000 3.000000

61 2.000000 3.000000 1.000000 2.00000 5.000000

62 1.000000 1.000000 1.000000 3.00000 1.000000

63 5.000000 2.000000 1.000000 1.00000 2.178862

64 3.000000 3.000000 3.000000 3.00000 3.000000

65 4.000000 4.000000 1.000000 1.00000 2.178862

66 4.000000 4.000000 2.000000 2.00000 2.000000

67 4.000000 3.000000 3.000000 2.00000 2.000000

68 5.000000 5.000000 2.000000 3.00000 3.000000

69 5.000000 3.366412 1.000000 3.00000 1.000000

70 3.000000 3.000000 3.000000 2.20155 3.000000

71 3.128788 3.000000 4.000000 3.00000 3.000000

72 2.000000 4.000000 3.000000 1.00000 3.000000

73 3.000000 3.000000 3.000000 2.20155 2.178862

74 3.000000 4.000000 1.000000 1.00000 4.000000

75 4.000000 5.000000 1.000000 1.00000 3.000000

76 1.000000 3.000000 1.000000 1.00000 1.000000

77 4.000000 4.000000 2.000000 2.20155 4.000000

78 2.000000 4.000000 3.000000 2.00000 1.000000

79 5.000000 1.000000 1.976562 1.00000 5.000000

80 4.000000 4.000000 1.976562 1.00000 4.000000

81 4.000000 4.000000 1.000000 1.00000 1.000000

82 3.000000 3.000000 1.000000 3.00000 2.000000

83 2.000000 5.000000 1.000000 1.00000 1.000000

84 4.000000 3.000000 1.976562 3.00000 1.000000

85 4.000000 3.000000 1.000000 1.00000 2.000000

86 1.000000 4.000000 1.000000 1.00000 1.000000

87 2.000000 4.000000 1.976562 2.00000 2.000000

88 4.000000 4.000000 2.000000 3.00000 4.000000

89 3.000000 3.000000 1.976562 2.00000 2.178862

90 3.000000 5.000000 1.000000 1.00000 1.000000

91 2.000000 4.000000 1.000000 1.00000 1.000000

92 5.000000 1.000000 1.976562 1.00000 5.000000

93 1.000000 4.000000 1.000000 2.00000 1.000000

94 3.000000 2.000000 2.000000 3.00000 1.000000

95 3.000000 3.000000 3.000000 3.00000 3.000000

96 3.000000 4.000000 2.000000 2.00000 2.000000

97 4.000000 2.000000 2.000000 2.00000 1.000000

98 3.000000 3.366412 3.000000 2.00000 2.178862

99 1.000000 3.000000 1.000000 3.00000 1.000000

100 1.000000 1.000000 1.000000 1.00000 2.000000

101 2.000000 3.000000 1.000000 3.00000 1.000000

102 5.000000 4.000000 4.000000 3.00000 3.000000

103 3.000000 4.000000 3.000000 1.00000 2.178862

104 4.000000 4.000000 1.000000 4.00000 1.000000

105 3.000000 3.366412 3.000000 2.20155 2.178862

106 1.000000 2.000000 1.976562 1.00000 1.000000

107 4.000000 3.000000 4.000000 3.00000 2.178862

108 4.000000 3.366412 1.000000 1.00000 2.178862

109 3.000000 3.000000 1.000000 3.00000 4.000000

110 2.000000 3.366412 3.000000 2.00000 3.000000

111 1.000000 5.000000 1.000000 1.00000 1.000000

112 4.000000 3.366412 2.000000 2.00000 2.178862

113 4.000000 5.000000 2.000000 1.00000 4.000000

114 4.000000 5.000000 1.976562 3.00000 2.178862

115 3.128788 5.000000 2.000000 2.20155 2.178862

116 3.000000 3.000000 3.000000 3.00000 2.000000

117 1.000000 5.000000 1.000000 2.20155 1.000000

118 3.128788 4.000000 1.000000 3.00000 2.000000

119 3.000000 4.000000 5.000000 5.00000 5.000000

120 5.000000 5.000000 1.000000 5.00000 4.000000

121 2.000000 4.000000 3.000000 1.00000 5.000000

122 1.000000 1.000000 1.000000 1.00000 4.000000

123 5.000000 4.000000 3.000000 3.00000 2.000000

124 3.000000 3.000000 3.000000 2.00000 3.000000

125 2.000000 3.000000 2.000000 2.20155 4.000000

126 5.000000 5.000000 1.000000 5.00000 2.178862

127 4.000000 4.000000 1.000000 2.20155 1.000000

128 2.000000 4.000000 3.000000 3.00000 2.000000

129 5.000000 3.000000 1.000000 1.00000 2.000000

130 4.000000 3.000000 2.000000 2.00000 2.000000

131 3.000000 4.000000 1.976562 3.00000 2.178862

132 2.000000 2.000000 1.976562 2.20155 2.000000

133 3.000000 3.000000 3.000000 3.00000 1.000000

134 5.000000 5.000000 1.976562 2.20155 4.000000

135 4.000000 5.000000 2.000000 4.00000 1.000000

136 3.000000 3.366412 1.000000 1.00000 1.000000

137 2.000000 3.000000 5.000000 3.00000 5.000000

138 5.000000 3.366412 1.976562 1.00000 2.000000

139 3.128788 1.000000 2.000000 1.00000 1.000000

140 3.000000 5.000000 1.976562 3.00000 1.000000

141 4.000000 3.000000 4.000000 4.00000 4.000000

142 3.128788 3.000000 1.000000 3.00000 1.000000

143 1.000000 1.000000 1.000000 1.00000 1.000000

144 4.000000 4.000000 4.000000 3.00000 2.000000

145 2.000000 4.000000 2.000000 1.00000 2.178862

146 4.000000 2.000000 3.000000 4.00000 1.000000

147 3.000000 3.000000 3.000000 3.00000 3.000000

148 3.000000 3.000000 1.976562 3.00000 4.000000

1. Compare the means and standard deviations from the original data set table with the mean substitution table. What should happen to the means? What should happen to the standard deviations?

When we replace missing data with its column's mean using mean substitution, the means of the columns in the mean-substituted table should be the same as the means in the original data set table. This is because mean substitution replaces the missing values with the mean value of the column, which should not change the overall mean of the column.

The standard deviations in the mean-substituted table may be smaller than the standard deviations in the original data set table. This is because replacing the missing values with the mean value can reduce the overall variability in the column, resulting in a smaller standard deviation. However, the extent of this reduction in variability depends on the proportion of missing values in the column. If the proportion of missing values is large, the reduction in variability may be more noticeable, and the standard deviation in the mean-substituted table may be significantly smaller than in the original data set table.

Original:

mean std

v1\_miss 3.128788 1.213193

v2\_miss 3.366412 1.144929

v3\_miss 1.976562 1.090148

v4\_miss 2.201550 1.134542

v5\_miss 2.178862 1.293315

New:

mean std

v1\_miss 3.128788 1.141391

v2\_miss 3.366412 1.073049

v3\_miss 1.976562 1.009849

v4\_miss 2.201550 1.055102

v5\_miss 2.178862 1.174231

1. Run one simple non-multiple imputation models (e.g. just a simple regression model) – for v3\_miss. Be sure to replace the missing values in v3\_miss with the predicted values from the regression. (I shortened this to just one variable – too much work otherwise I think).

The leftmost value is the index

0 1.0

1 1.0

2 3.0

3 1.0

4 2.0

5 1.0

6 3.0

7 3.0

8 2.0

9 1.0

10 1.0

11 1.0

12 3.0

13 1.0

14 1.0

16 4.0

17 2.0

18 2.0

19 1.0

20 1.0

21 2.0

22 1.0

23 1.0

24 5.0

25 1.0

26 3.0

27 1.0

28 4.0

29 3.0

30 1.0

33 1.0

34 1.0

35 3.0

36 1.0

37 1.0

39 2.0

40 1.0

41 2.0

42 1.0

43 4.0

44 3.0

45 3.0

46 3.0

47 2.0

48 1.0

49 1.0

50 1.0

51 3.0

52 2.0

53 2.0

56 1.0

57 3.0

58 1.0

60 3.0

61 1.0

62 1.0

63 1.0

64 3.0

65 1.0

66 2.0

67 3.0

68 2.0

69 1.0

70 3.0

71 4.0

72 3.0

73 3.0

74 1.0

75 1.0

76 1.0

77 2.0

78 3.0

81 1.0

82 1.0

83 1.0

85 1.0

86 1.0

88 2.0

90 1.0

91 1.0

93 1.0

94 2.0

95 3.0

96 2.0

97 2.0

98 3.0

99 1.0

100 1.0

101 1.0

102 4.0

103 3.0

104 1.0

105 3.0

107 4.0

108 1.0

109 1.0

110 3.0

111 1.0

112 2.0

113 2.0

115 2.0

116 3.0

117 1.0

118 1.0

119 5.0

120 1.0

121 3.0

122 1.0

123 3.0

124 3.0

125 2.0

126 1.0

127 1.0

128 3.0

129 1.0

130 2.0

133 3.0

135 2.0

136 1.0

137 5.0

139 2.0

141 4.0

142 1.0

143 1.0

144 4.0

145 2.0

146 3.0

147 3.0

Name: v3\_miss, dtype: float64

1. Compare the table of means and standard deviations from the original data set to the one constructed in step 4 above. What do you see?

When we replace missing data of one column using linear regression, the mean of the imputed data should be similar to the mean of the original data, but the standard deviation of the imputed data may be different from the standard deviation of the original data.

The mean substitution method replaces missing data with the mean of the non-missing values of that column, which can artificially inflate the mean and reduce variability. On the other hand, the linear regression method uses information from other columns to impute the missing values, which should result in a more accurate estimate of the true mean.

However, the standard deviation of the imputed data may be larger or smaller than the standard deviation of the original data. This is because the imputation method introduces some randomness in the imputed values, which can increase or decrease the variability of the data. Additionally, linear regression imputation assumes a linear relationship between the missing variable and the other variables used to impute it, which may not be accurate in some cases.

Original:

mean std

v3\_miss 1.976562 1.090148

New:

mean std

v3\_miss 1.976562 1.090148

1. Run a multiple imputation model on the five variables – you can pick from multiple imputation regression, mcmc or any other flavor.

v1\_miss v2\_miss v3\_miss v4\_miss v5\_miss

0 4.000000 3.000000 1.000000 3.000000 1.000000

1 2.000000 4.000000 1.000000 1.947759 3.000000

2 3.000000 3.000000 3.000000 3.000000 2.703470

3 2.000000 4.000000 1.000000 2.000000 2.000000

4 2.000000 4.000000 2.000000 5.000000 5.000000

5 4.000000 3.355189 1.000000 2.000000 1.000000

6 2.000000 4.000000 3.000000 1.000000 3.000000

7 2.000000 5.000000 3.000000 3.000000 1.000000

8 1.000000 5.000000 2.000000 3.000000 2.000000

9 3.000000 3.000000 1.000000 1.000000 3.000000

10 4.000000 1.000000 1.000000 1.000000 1.000000

11 5.000000 4.000000 1.000000 2.000000 1.000000

12 2.000000 4.000000 3.000000 4.000000 1.000000

13 4.000000 1.000000 1.000000 1.000000 1.000000

14 2.000000 3.000000 1.000000 1.000000 1.000000

15 3.127124 3.365838 1.993718 2.200919 2.195495

16 4.000000 4.000000 4.000000 5.000000 4.000000

17 3.000000 3.000000 2.000000 1.000000 1.000000

18 3.000000 2.000000 2.000000 3.000000 5.000000

19 1.000000 3.000000 1.000000 1.000000 1.000000

20 3.128590 4.000000 1.000000 3.000000 1.000000

21 3.000000 4.000000 2.000000 2.404680 4.000000

22 4.000000 3.000000 1.000000 3.000000 3.000000

23 5.000000 4.000000 1.000000 3.000000 2.000000

24 4.000000 4.000000 5.000000 4.000000 3.692981

25 4.000000 5.000000 1.000000 1.000000 1.000000

26 1.000000 2.000000 3.000000 2.000000 3.000000

27 3.000000 3.000000 1.000000 3.000000 1.000000

28 3.000000 3.497654 4.000000 3.000000 3.000000

29 1.000000 5.000000 3.000000 1.000000 1.000000

30 3.000000 3.000000 1.000000 1.000000 1.665844

31 3.000000 3.342463 1.881716 2.000000 2.000000

32 3.127124 3.365838 1.993718 2.200919 2.195495

33 2.000000 1.000000 1.000000 1.000000 1.000000

34 3.098503 3.382272 1.000000 3.000000 1.000000

35 4.000000 5.000000 3.000000 3.000000 2.000000

36 3.016429 3.297023 1.000000 1.807514 1.000000

37 4.000000 4.000000 1.000000 1.000000 1.000000

38 3.211778 3.000000 2.420431 3.000000 3.000000

39 3.000000 2.000000 2.000000 3.000000 4.000000

40 5.000000 2.000000 1.000000 1.000000 1.000000

41 1.000000 2.000000 2.000000 2.000000 2.055307

42 3.000000 3.000000 1.000000 1.000000 1.665844

43 4.000000 4.000000 4.000000 1.000000 2.000000

44 4.000000 4.000000 3.000000 2.000000 2.000000

45 5.000000 5.000000 3.000000 2.000000 3.000000

46 4.000000 4.000000 3.000000 1.000000 3.000000

47 5.000000 2.000000 2.000000 1.000000 1.000000

48 2.988084 2.000000 1.000000 1.721940 1.773788

49 3.000000 4.000000 1.000000 1.000000 1.000000

50 2.887081 1.000000 1.000000 1.539406 1.000000

51 3.000000 5.000000 3.000000 2.559285 1.000000

52 2.000000 2.000000 2.000000 2.000000 1.000000

53 4.000000 2.000000 2.000000 4.000000 1.000000

54 3.127124 3.365838 1.993718 2.200919 2.195495

55 3.127124 3.365838 1.993718 2.200919 2.195495

56 4.000000 3.281509 1.000000 1.000000 2.000000

57 3.239658 2.000000 3.000000 4.000000 3.000000

58 2.000000 3.000000 1.000000 1.000000 1.000000

59 4.000000 4.000000 2.311668 3.000000 2.452279

60 3.271685 4.000000 3.000000 3.000000 3.000000

61 2.000000 3.000000 1.000000 2.000000 5.000000

62 1.000000 1.000000 1.000000 3.000000 1.000000

63 5.000000 2.000000 1.000000 1.000000 1.881401

64 3.000000 3.000000 3.000000 3.000000 3.000000

65 4.000000 4.000000 1.000000 1.000000 1.708967

66 4.000000 4.000000 2.000000 2.000000 2.000000

67 4.000000 3.000000 3.000000 2.000000 2.000000

68 5.000000 5.000000 2.000000 3.000000 3.000000

69 5.000000 3.469282 1.000000 3.000000 1.000000

70 3.000000 3.000000 3.000000 2.495671 3.000000

71 3.242356 3.000000 4.000000 3.000000 3.000000

72 2.000000 4.000000 3.000000 1.000000 3.000000

73 3.000000 3.000000 3.000000 2.468229 2.654952

74 3.000000 4.000000 1.000000 1.000000 4.000000

75 4.000000 5.000000 1.000000 1.000000 3.000000

76 1.000000 3.000000 1.000000 1.000000 1.000000

77 4.000000 4.000000 2.000000 2.500203 4.000000

78 2.000000 4.000000 3.000000 2.000000 1.000000

79 5.000000 1.000000 2.392139 1.000000 5.000000

80 4.000000 4.000000 2.278280 1.000000 4.000000

81 4.000000 4.000000 1.000000 1.000000 1.000000

82 3.000000 3.000000 1.000000 3.000000 2.000000

83 2.000000 5.000000 1.000000 1.000000 1.000000

84 4.000000 3.000000 1.803179 3.000000 1.000000

85 4.000000 3.000000 1.000000 1.000000 2.000000

86 1.000000 4.000000 1.000000 1.000000 1.000000

87 2.000000 4.000000 1.927426 2.000000 2.000000

88 4.000000 4.000000 2.000000 3.000000 4.000000

89 3.000000 3.000000 1.904317 2.000000 2.143745

90 3.000000 5.000000 1.000000 1.000000 1.000000

91 2.000000 4.000000 1.000000 1.000000 1.000000

92 5.000000 1.000000 2.392139 1.000000 5.000000

93 1.000000 4.000000 1.000000 2.000000 1.000000

94 3.000000 2.000000 2.000000 3.000000 1.000000

95 3.000000 3.000000 3.000000 3.000000 3.000000

96 3.000000 4.000000 2.000000 2.000000 2.000000

97 4.000000 2.000000 2.000000 2.000000 1.000000

98 3.000000 3.387955 3.000000 2.000000 2.595510

99 1.000000 3.000000 1.000000 3.000000 1.000000

100 1.000000 1.000000 1.000000 1.000000 2.000000

101 2.000000 3.000000 1.000000 3.000000 1.000000

102 5.000000 4.000000 4.000000 3.000000 3.000000

103 3.000000 4.000000 3.000000 1.000000 2.477891

104 4.000000 4.000000 1.000000 4.000000 1.000000

105 3.000000 3.422867 3.000000 2.514195 2.640919

106 1.000000 2.000000 1.244799 1.000000 1.000000

107 4.000000 3.000000 4.000000 3.000000 3.217272

108 4.000000 3.282900 1.000000 1.000000 1.739877

109 3.000000 3.000000 1.000000 3.000000 4.000000

110 2.000000 3.340034 3.000000 2.000000 3.000000

111 1.000000 5.000000 1.000000 1.000000 1.000000

112 4.000000 3.392012 2.000000 2.000000 2.253986

113 4.000000 5.000000 2.000000 1.000000 4.000000

114 4.000000 5.000000 2.368870 3.000000 2.433634

115 3.216925 5.000000 2.000000 2.389823 2.152721

116 3.000000 3.000000 3.000000 3.000000 2.000000

117 1.000000 5.000000 1.000000 1.804509 1.000000

118 3.180779 4.000000 1.000000 3.000000 2.000000

119 3.000000 4.000000 5.000000 5.000000 5.000000

120 5.000000 5.000000 1.000000 5.000000 4.000000

121 2.000000 4.000000 3.000000 1.000000 5.000000

122 1.000000 1.000000 1.000000 1.000000 4.000000

123 5.000000 4.000000 3.000000 3.000000 2.000000

124 3.000000 3.000000 3.000000 2.000000 3.000000

125 2.000000 3.000000 2.000000 2.197817 4.000000

126 5.000000 5.000000 1.000000 5.000000 2.117040

127 4.000000 4.000000 1.000000 1.979738 1.000000

128 2.000000 4.000000 3.000000 3.000000 2.000000

129 5.000000 3.000000 1.000000 1.000000 2.000000

130 4.000000 3.000000 2.000000 2.000000 2.000000

131 3.000000 4.000000 2.286274 3.000000 2.355195

132 2.000000 2.000000 1.766435 1.861598 2.000000

133 3.000000 3.000000 3.000000 3.000000 1.000000

134 5.000000 5.000000 2.830334 2.940898 4.000000

135 4.000000 5.000000 2.000000 4.000000 1.000000

136 3.000000 3.241097 1.000000 1.000000 1.000000

137 2.000000 3.000000 5.000000 3.000000 5.000000

138 5.000000 3.354230 1.619488 1.000000 2.000000

139 2.871191 1.000000 2.000000 1.000000 1.000000

140 3.000000 5.000000 1.933331 3.000000 1.000000

141 4.000000 3.000000 4.000000 4.000000 4.000000

142 3.079903 3.000000 1.000000 3.000000 1.000000

143 1.000000 1.000000 1.000000 1.000000 1.000000

144 4.000000 4.000000 4.000000 3.000000 2.000000

145 2.000000 4.000000 2.000000 1.000000 1.964089

146 4.000000 2.000000 3.000000 4.000000 1.000000

147 3.000000 3.000000 3.000000 3.000000 3.000000

148 3.000000 3.000000 2.728263 3.000000 4.000000

1. Compare the table of means and standard deviations from step 6 with the original data set. What do you expect? What do you actually see?

When we replace missing data for all columns with missing data using multiple imputation model, such as iterative imputer, the means of the imputed data should be similar (not *identical*) to the means of the original data set. However, the standard deviations of the imputed data may be slightly different from the original data set, as the imputed values are generated from a statistical model, rather than simply substituting missing values with mean values.

I see that there are minimal but noticeable differences in the means, but at the decimal point. The differences in standard deviations are slightly more profound.

Original:

mean std

v1\_miss 3.128788 1.213193

v2\_miss 3.366412 1.144929

v3\_miss 1.976562 1.090148

v4\_miss 2.201550 1.134542

v5\_miss 2.178862 1.293315

New:

mean std

v1\_miss 3.127124 1.142051

v2\_miss 3.365838 1.073265

v3\_miss 1.993718 1.019845

v4\_miss 2.200919 1.062462

v5\_miss 2.195495 1.190645

1. For an extra 8 points extra credit, produce a missing values matrix like the one shown in class. You can use SAS or name your poison in terms of platform as long as it produces a similar matrix.

See Jupyter notebook for tabulated output

Data for exercise #5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| v1\_miss | v2\_miss | v3\_miss | v4\_miss | v5\_miss |
| 4.00 | 3.00 | 1.00 | 3.00 | 1.00 |
| 2.00 | 4.00 | 1.00 |  | 3.00 |
| 3.00 | 3.00 | 3.00 | 3.00 |  |
| 2.00 | 4.00 | 1.00 | 2.00 | 2.00 |
| 2.00 | 4.00 | 2.00 | 5.00 | 5.00 |
| 4.00 |  | 1.00 | 2.00 | 1.00 |
| 2.00 | 4.00 | 3.00 | 1.00 | 3.00 |
| 2.00 | 5.00 | 3.00 | 3.00 | 1.00 |
| 1.00 | 5.00 | 2.00 | 3.00 | 2.00 |
| 3.00 | 3.00 | 1.00 | 1.00 | 3.00 |
| 4.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 5.00 | 4.00 | 1.00 | 2.00 | 1.00 |
| 2.00 | 4.00 | 3.00 | 4.00 | 1.00 |
| 4.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2.00 | 3.00 | 1.00 | 1.00 | 1.00 |
|  |  |  |  |  |
| 4.00 | 4.00 | 4.00 | 5.00 | 4.00 |
| 3.00 | 3.00 | 2.00 | 1.00 | 1.00 |
| 3.00 | 2.00 | 2.00 | 3.00 | 5.00 |
| 1.00 | 3.00 | 1.00 | 1.00 | 1.00 |
|  | 4.00 | 1.00 | 3.00 | 1.00 |
| 3.00 | 4.00 | 2.00 |  | 4.00 |
| 4.00 | 3.00 | 1.00 | 3.00 | 3.00 |
| 5.00 | 4.00 | 1.00 | 3.00 | 2.00 |
| 4.00 | 4.00 | 5.00 | 4.00 |  |
| 4.00 | 5.00 | 1.00 | 1.00 | 1.00 |
| 1.00 | 2.00 | 3.00 | 2.00 | 3.00 |
| 3.00 | 3.00 | 1.00 | 3.00 | 1.00 |
| 3.00 |  | 4.00 | 3.00 | 3.00 |
| 1.00 | 5.00 | 3.00 | 1.00 | 1.00 |
| 3.00 | 3.00 | 1.00 | 1.00 |  |
| 3.00 |  |  | 2.00 | 2.00 |
|  |  |  |  |  |
| 2.00 | 1.00 | 1.00 | 1.00 | 1.00 |
|  |  | 1.00 | 3.00 | 1.00 |
| 4.00 | 5.00 | 3.00 | 3.00 | 2.00 |
|  |  | 1.00 |  | 1.00 |
| 4.00 | 4.00 | 1.00 | 1.00 | 1.00 |
|  | 3.00 |  | 3.00 | 3.00 |
| 3.00 | 2.00 | 2.00 | 3.00 | 4.00 |
| 5.00 | 2.00 | 1.00 | 1.00 | 1.00 |
| 1.00 | 2.00 | 2.00 | 2.00 |  |
| 3.00 | 3.00 | 1.00 | 1.00 |  |
| 4.00 | 4.00 | 4.00 | 1.00 | 2.00 |
| 4.00 | 4.00 | 3.00 | 2.00 | 2.00 |
| 5.00 | 5.00 | 3.00 | 2.00 | 3.00 |
| 4.00 | 4.00 | 3.00 | 1.00 | 3.00 |
| 5.00 | 2.00 | 2.00 | 1.00 | 1.00 |
|  | 2.00 | 1.00 |  |  |
| 3.00 | 4.00 | 1.00 | 1.00 | 1.00 |
|  | 1.00 | 1.00 |  | 1.00 |
| 3.00 | 5.00 | 3.00 |  | 1.00 |
| 2.00 | 2.00 | 2.00 | 2.00 | 1.00 |
| 4.00 | 2.00 | 2.00 | 4.00 | 1.00 |
|  |  |  |  |  |
|  |  |  |  |  |
| 4.00 |  | 1.00 | 1.00 | 2.00 |
|  | 2.00 | 3.00 | 4.00 | 3.00 |
| 2.00 | 3.00 | 1.00 | 1.00 | 1.00 |
| 4.00 | 4.00 |  | 3.00 |  |
|  | 4.00 | 3.00 | 3.00 | 3.00 |
| 2.00 | 3.00 | 1.00 | 2.00 | 5.00 |
| 1.00 | 1.00 | 1.00 | 3.00 | 1.00 |
| 5.00 | 2.00 | 1.00 | 1.00 |  |
| 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 4.00 | 4.00 | 1.00 | 1.00 |  |
| 4.00 | 4.00 | 2.00 | 2.00 | 2.00 |
| 4.00 | 3.00 | 3.00 | 2.00 | 2.00 |
| 5.00 | 5.00 | 2.00 | 3.00 | 3.00 |
| 5.00 |  | 1.00 | 3.00 | 1.00 |
| 3.00 | 3.00 | 3.00 |  | 3.00 |
|  | 3.00 | 4.00 | 3.00 | 3.00 |
| 2.00 | 4.00 | 3.00 | 1.00 | 3.00 |
| 3.00 | 3.00 | 3.00 |  |  |
| 3.00 | 4.00 | 1.00 | 1.00 | 4.00 |
| 4.00 | 5.00 | 1.00 | 1.00 | 3.00 |
| 1.00 | 3.00 | 1.00 | 1.00 | 1.00 |
| 4.00 | 4.00 | 2.00 |  | 4.00 |
| 2.00 | 4.00 | 3.00 | 2.00 | 1.00 |
| 5.00 | 1.00 |  | 1.00 | 5.00 |
| 4.00 | 4.00 |  | 1.00 | 4.00 |
| 4.00 | 4.00 | 1.00 | 1.00 | 1.00 |
| 3.00 | 3.00 | 1.00 | 3.00 | 2.00 |
| 2.00 | 5.00 | 1.00 | 1.00 | 1.00 |
| 4.00 | 3.00 |  | 3.00 | 1.00 |
| 4.00 | 3.00 | 1.00 | 1.00 | 2.00 |
| 1.00 | 4.00 | 1.00 | 1.00 | 1.00 |
| 2.00 | 4.00 |  | 2.00 | 2.00 |
| 4.00 | 4.00 | 2.00 | 3.00 | 4.00 |
| 3.00 | 3.00 |  | 2.00 |  |
| 3.00 | 5.00 | 1.00 | 1.00 | 1.00 |
| 2.00 | 4.00 | 1.00 | 1.00 | 1.00 |
| 5.00 | 1.00 |  | 1.00 | 5.00 |
| 1.00 | 4.00 | 1.00 | 2.00 | 1.00 |
| 3.00 | 2.00 | 2.00 | 3.00 | 1.00 |
| 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 3.00 | 4.00 | 2.00 | 2.00 | 2.00 |
| 4.00 | 2.00 | 2.00 | 2.00 | 1.00 |
| 3.00 |  | 3.00 | 2.00 |  |
| 1.00 | 3.00 | 1.00 | 3.00 | 1.00 |
| 1.00 | 1.00 | 1.00 | 1.00 | 2.00 |
| 2.00 | 3.00 | 1.00 | 3.00 | 1.00 |
| 5.00 | 4.00 | 4.00 | 3.00 | 3.00 |
| 3.00 | 4.00 | 3.00 | 1.00 |  |
| 4.00 | 4.00 | 1.00 | 4.00 | 1.00 |
| 3.00 |  | 3.00 |  |  |
| 1.00 | 2.00 |  | 1.00 | 1.00 |
| 4.00 | 3.00 | 4.00 | 3.00 |  |
| 4.00 |  | 1.00 | 1.00 |  |
| 3.00 | 3.00 | 1.00 | 3.00 | 4.00 |
| 2.00 |  | 3.00 | 2.00 | 3.00 |
| 1.00 | 5.00 | 1.00 | 1.00 | 1.00 |
| 4.00 |  | 2.00 | 2.00 |  |
| 4.00 | 5.00 | 2.00 | 1.00 | 4.00 |
| 4.00 | 5.00 |  | 3.00 |  |
|  | 5.00 | 2.00 |  |  |
| 3.00 | 3.00 | 3.00 | 3.00 | 2.00 |
| 1.00 | 5.00 | 1.00 |  | 1.00 |
|  | 4.00 | 1.00 | 3.00 | 2.00 |
| 3.00 | 4.00 | 5.00 | 5.00 | 5.00 |
| 5.00 | 5.00 | 1.00 | 5.00 | 4.00 |
| 2.00 | 4.00 | 3.00 | 1.00 | 5.00 |
| 1.00 | 1.00 | 1.00 | 1.00 | 4.00 |
| 5.00 | 4.00 | 3.00 | 3.00 | 2.00 |
| 3.00 | 3.00 | 3.00 | 2.00 | 3.00 |
| 2.00 | 3.00 | 2.00 |  | 4.00 |
| 5.00 | 5.00 | 1.00 | 5.00 |  |
| 4.00 | 4.00 | 1.00 |  | 1.00 |
| 2.00 | 4.00 | 3.00 | 3.00 | 2.00 |
| 5.00 | 3.00 | 1.00 | 1.00 | 2.00 |
| 4.00 | 3.00 | 2.00 | 2.00 | 2.00 |
| 3.00 | 4.00 |  | 3.00 |  |
| 2.00 | 2.00 |  |  | 2.00 |
| 3.00 | 3.00 | 3.00 | 3.00 | 1.00 |
| 5.00 | 5.00 |  |  | 4.00 |
| 4.00 | 5.00 | 2.00 | 4.00 | 1.00 |
| 3.00 |  | 1.00 | 1.00 | 1.00 |
| 2.00 | 3.00 | 5.00 | 3.00 | 5.00 |
| 5.00 |  |  | 1.00 | 2.00 |
|  | 1.00 | 2.00 | 1.00 | 1.00 |
| 3.00 | 5.00 |  | 3.00 | 1.00 |
| 4.00 | 3.00 | 4.00 | 4.00 | 4.00 |
|  | 3.00 | 1.00 | 3.00 | 1.00 |
| 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 4.00 | 4.00 | 4.00 | 3.00 | 2.00 |
| 2.00 | 4.00 | 2.00 | 1.00 |  |
| 4.00 | 2.00 | 3.00 | 4.00 | 1.00 |
| 3.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| 3.00 | 3.00 |  | 3.00 | 4.00 |