190031920 A Nikhil Reddy **DS Practical 2** In [1]: import numpy as np import pandas as pd filedata = pd.read_csv("./21B15600.csv") In [2]: file2data = pd.read_csv("./74B05600.csv") filedata In [3]: Out[3]: test preparation reading parental level of math writing gender race/ethnicity lunch education course score score score 0 female bachelor's degree standard 72 72 74 group B none 1 completed 69 90 88 female group C some college standard group B 95 2 female master's degree standard none 90 93 57 3 47 44 male group A associate's degree free/reduced none 4 male group C some college standard none 76 78 75 5 83 78 female group B associate's degree standard none 71 6 female some college standard completed 88 95 92 group B 7 some college free/reduced 40 43 39 male group B none completed 8 male group D high school free/reduced 67 In [4]: file2data Out[4]: parental level of test preparation math reading writing gender race/ethnicity lunch education course score score score 74 0 bachelor's degree 72 72 female group B standard none 69 90 female group C some college standard completed 88 2 female group B master's degree standard none 90 95 93 3 male group A associate's degree free/reduced none 47 57 44 76 78 75 4 male group C some college standard none ... 995 88 99 95 female group E master's degree standard completed 996 55 55 male group C high school free/reduced none 62 997 59 71 65 female group C high school free/reduced completed 998 68 78 77 female group D some college standard completed 86 999 some college free/reduced 77 86 female group D none 1000 rows \times 8 columns In [5]: # 1. Gather Columns into rows pd.melt(filedata) Out[5]: variable value gender female 0 1 gender female 2 gender female 3 gender male 4 gender male 67 writing score 75 78 68 writing score writing score 92 writing score 39 writing score 67 72 rows \times 2 columns In [6]: #2. Spread rows into columns filedata.pivot(columns = 'gender', values = 'math score') Out[6]: gender female male 72.0 NaN 1 NaN 69.0 2 90.0 NaN 3 47.0 NaN 4 NaN 76.0 5 71.0 NaN 88.0 6 NaN 7 40.0 NaN 64.0 NaN In [7]: # 3. Append rows of Data Frames pd.concat([filedata,file2data]) Out[7]: parental level of test preparation reading writing math gender race/ethnicity lunch education score score course score 72 74 0 bachelor's degree standard 72 female group B none 69 90 female group C some college standard completed 88 90 93 2 female group B master's degree standard none 95 47 3 male associate's degree free/reduced none 57 44 group A 76 75 4 male group C some college standard none 78 995 female group E master's degree standard completed 88 99 95 996 male group C high school free/reduced none 62 55 55 high school free/reduced 59 65 997 female group C completed 71 998 standard completed female group D some college 68 78 77 77 86 86 999 female group D some college free/reduced none 1009 rows × 8 columns #4. Append columns of Data Frames In [8]: pd.concat([filedata,file2data], axis=1) Out[8]: parental test pare math reading score score score education course educa bachelor's bach 0 74.0 female group B female group B standard none 72.0 72.0 degree d€ some 69.0 1 female group C standard completed 90.0 88.0 female group C college CC master's mas 90.0 95.0 group B 2 female group B standard none 93.0 female degree d€ associate's assoc 3 free/reduced 47.0 male group A none 57.0 44.0 male group A degree d€ some 4 male group C standard none 76.0 78.0 75.0 male group C college CC mas 995 group E NaN NaN NaN NaN NaN NaN NaN NaN female d€ 996 NaN NaN NaN NaN NaN NaN NaN NaN male group C s 997 NaN NaN NaN NaN NaN NaN NaN NaN female group C s 998 NaN NaN NaN NaN NaN NaN NaN NaN female group D CC 999 NaN NaN NaN NaN NaN NaN NaN NaN female group D 1000 rows \times 16 columns In [9]: Order rows by values of a column filedata.sort_values('math score') Out[9]: parental level of test preparation reading math writing lunch gender race/ethnicity education score score score course 7 group B some college free/reduced male none 40 43 39 3 male group A associate's degree free/reduced 47 57 44 none free/reduced 8 male group D high school completed 64 64 67 completed 90 1 some college 69 88 female group C standard 5 female group B associate's degree standard none 71 83 78 0 72 72 74 female group B bachelor's degree standard none male group C some college standard none 76 78 75 92 6 some college completed 88 95 female group B standard 2 female master's degree standard 90 95 93 group B none In [10]: #6. Order rows by values of a column filedata.sort_values('math score', ascending=False) Out[10]: parental level of test preparation math reading writing gender race/ethnicity lunch education course score score score 2 female master's degree standard 90 95 93 group B none 6 female group B some college standard completed 88 95 92 4 78 75 male group C some college standard none 76 0 female group B bachelor's degree standard none 72 72 74 5 71 83 78 female group B associate's degree none standard 1 female group C some college standard completed 69 90 88 8 67 high school free/reduced completed 64 64 male group D 3 47 44 male group A associate's degree free/reduced none 57 7 male group B some college free/reduced none 40 43 39 In [11]: #7. Sort the index of a Data Frame filedata.sort_index() Out[11]: parental level of reading writing test preparation math gender race/ethnicity lunch education course score score score 0 72 female group B bachelor's degree standard none 72 74 1 female group C some college standard completed 69 90 88 2 90 95 93 master's degree female group B standard none associate's degree free/reduced 3 47 57 male group A none 44 4 76 78 75 male group C some college standard none 5 female 71 83 78 group B associate's degree standard none 6 standard 88 95 92 female group B some college completed 7 some college free/reduced 40 43 39 male group B none 8 64 group D high school free/reduced completed 64 67 male In [12]: #8. Reset index of Data Frame to row numbers, moving index to columns. filedata.reset_index() Out[12]: parental level of test preparation math reading writing index gender race/ethnicity lunch education score score score course 0 0 female bachelor's degree 72 72 74 group B standard none 1 group C 90 1 some college standard completed 69 88 female 2 female 90 95 93 2 group B master's degree standard none 3 47 3 group A associate's degree free/reduced 57 44 male none 4 4 78 male group C some college standard none 76 75 5 83 78 5 associate's degree 71 female group B standard none 6 female group B 95 92 6 some college standard completed 88 7 7 43 39 male group B some college free/reduced none 40 8 8 male group D high school free/reduced completed 64 64 67 In [13]: #9. Drop columns from Data Frame Subset observations filedata.drop(columns=['gender','test preparation course']) Out[13]: race/ethnicity parental level of education reading score writing score lunch math score 0 group B bachelor's degree standard 72 72 74 1 some college 69 90 88 group C standard 2 group B master's degree standard 90 95 93 3 47 57 44 associate's degree free/reduced group A 4 some college 78 group C standard 76 75 5 group B associate's degree standard 71 83 78 6 group B some college standard 88 95 92 7 some college free/reduced 40 43 39 group B 8 group D high school free/reduced 64 64 67 # 10. Extract rows that meet logical criteria. In [14]: filedata[filedata['reading score'] > 80] Out[14]: parental level of test preparation math reading writing gender race/ethnicity lunch education course score score score female group C 90 some college standard completed group B 90 95 93 female master's degree standard none 78 female group B associate's degree standard none 71 83 female completed 88 95 92 6 group B some college standard #11. Remove duplicate rows (only considers columns) In [15]: filedata.drop_duplicates() Out[15]: test preparation reading parental level of math writing gender race/ethnicity lunch education course score score score 0 female group B bachelor's degree standard 72 72 74 none completed 90 1 standard 69 88 female group C some college 2 female master's degree standard 90 95 93 group B none 3 47 57 44 male group A associate's degree free/reduced none 4 male group C some college standard none 76 78 75 5 associate's degree 71 83 78 female group B standard none 6 female group B some college standard completed 88 95 92 7 some college free/reduced 40 43 39 male group B none 8 67 high school free/reduced completed 64 64 male group D In [16]: # 12. Select first n rows filedata.head(50) Out[16]: parental level of test preparation math reading writing lunch gender race/ethnicity education score course score score 0 bachelor's degree 72 72 74 female group B standard none standard 1 female some college completed 69 90 88 group C 2 female group B master's degree standard none 90 95 93 3 male associate's degree free/reduced none 47 57 44 group A 4 75 76 78 male group C some college standard none 5 female group B associate's degree standard none 71 83 78 6 some college completed 88 95 92 female group B standard 7 male group B some college free/reduced none 40 43 39 8 high school free/reduced completed 64 64 67 male group D In [17]: # 12. Select last n rows filedata.tail(50) Out[17]: parental level of test preparation math reading writing lunch gender race/ethnicity education score score course score 0 female group B bachelor's degree standard 72 72 74 none some college 1 female group C completed 69 90 88 standard standard 2 female master's degree 90 93 group B none 3 associate's degree free/reduced 47 44 male group A 57 none 4 male group C some college standard none 76 78 75 5 female associate's degree 71 83 78 group B standard none 6 female standard completed 88 95 92 group B some college 40 43 39 7 some college free/reduced male group B none 8 male group D high school free/reduced completed 64 67 specific name filedata['parental level of education'] Out[18]: 0 bachelor's degree some college 2 master's degree 3 associate's degree some college 5 associate's degree 6 some college 7 some college 8 high school Name: parental level of education, dtype: object In [19]: #14. Select Multiple column with specific name. filedata[['gender', 'math score', 'reading score', 'writing score']] Out[19]: gender math score reading score writing score 0 female 72 72 74 1 90 female 69 88 2 female 90 95 93 3 47 57 male 44 4 male 76 78 75 71 78 5 female 83 6 88 95 92 female 7 40 39 male 43 8 64 64 67 male In [20]: Drop rows with any column having NA/null data. #15. filedata.dropna() Out[20]: parental level of writing test preparation math reading gender race/ethnicity lunch education course score score score 0 female group B bachelor's degree standard none 72 72 74 1 female group C some college standard completed 69 90 88 2 95 female group B master's degree standard none 90 93 3 male associate's degree free/reduced 47 57 44 group A none 75 4 male group C some college standard none 76 78 5 female associate's degree 71 83 78 group B standard none female group B some college standard completed 88 95 92 7 some college free/reduced 40 43 39 male group B none 8 64 male group D high school free/reduced completed 64 67 In [21]: #16. Replace all NA/null data with value. filedata.fillna('value') Out[21]: parental level of test preparation math reading writing lunch gender race/ethnicity education course score score score 0 female bachelor's degree 72 72 74 group B standard none 1 69 90 female group C some college standard completed 88 90 95 93 2 female group B master's degree standard none group A 3 47 57 male associate's degree free/reduced none 44 male group C some college standard none 76 78 75 71 83 78 female group B associate's degree standard none 92 female aroup B some college standard completed 7 39 40 43 male group B some college free/reduced none 67 male group D high school free/reduced completed 64 64 In [22]: filedata.describe() Out[22]: reading score writing score 9.000000 9.000000 9.000000 count 68.555556 75.222222 72.22222 mean 19.504985 std 16.629124 17.942810 min 40.000000 43.000000 39.000000 64.000000 64.000000 67.000000 25% 71.000000 78.000000 50% 75.000000 76.000000 90.000000 88.000000 **75**% 90.000000 95.000000 93.000000 max In [23]: filedata.groupby(['gender']).mean() Out[23]: math score reading score writing score gender female 78.00 87.0 85.00 male 56.75 60.5 56.25 In [24]: # Combine Data Sets pd.merge(filedata, file2data, how='outer', on='gender') Out[24]: parental test math reading writing race/ethnicity_y gender race/ethnicity_x level of lunch_x preparation score_x score_x score_x education_x educ course_x bachelor's ba 72 72 74 0 female group B standard group B none degree bachelor's group B 72 72 74 female standard group C some none degree bachelor's 72 72 74 2 female group B standard group B none degree bachelor's 72 72 74 3 female group B standard group B none degree bachelor's 4 72 72 74 female group B group B some standard none degree hig 4513 male group D high school free/reduced completed 64 64 67 group A 4514 group D 64 64 67 group E male high school free/reduced completed 4515 64 67 male group D high school free/reduced completed 64 group E high school free/reduced 4516 male group D completed 64 64 67 group A hig 4517 64 64 67 male group D high school free/reduced completed group C hig 4518 rows × 15 columns