None
HW5 Solutions

Note:

1. Only one possible right answer is shown. All possible right answers will be given full credit.
2. Only the final solution is shown, and the details of actual code is not shown.
3. You may come to the office hours or the help sessions to discuss the HW solutions.
4. If you find any typos or issues, kindly contact your section instructor, or send a text @ **smujahid** on MS teams.

Problem A

```
-----------------------------------------------Problem #A-----------------------------------------------
A-1:

Displaying 5 random rows of the data
```

	Gender	Location	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
2318	Female	Dhahran	11	17	7	8	59	58	59	50
2731	Male	Qassim	9	25	17	5	27	27	12	24
489	Male	Madinah	17	14	18	16	65	69	58	67
2858	Male	Madinah	16	13	8	15	80	90	73	80
3492	Female	Jeddah	23	17	10	24	91	78	92	85

```
The number of rows in the data is 3500, and the number of columns in the data is 10.

The non-null values in each column are:
Gender      3500
Location    3500
Quiz-1      3500
Quiz-2      3500
Quiz-3      3500
Quiz-4      3500
Major-1     3500
Major-2     3500
Major-3     3500
Final       3500
dtype: int64

Therefore, no missing values.

The summary statistics are:
```

	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
count	3500.000000	3500.000000	3500.000000	3500.000000	3500.000000	3500.000000	3500.000000	3500.000000
mean	12.414571	14.814857	14.924286	12.523714	55.954571	55.949143	56.154286	54.019714
std	5.618516	6.001572	6.094121	5.612041	21.552887	21.734182	21.803054	21.418980
min	0.000000	5.000000	5.000000	0.000000	12.000000	12.000000	12.000000	8.000000
25%	8.000000	10.000000	10.000000	8.000000	38.000000	38.000000	38.000000	36.000000
50%	12.000000	15.000000	15.000000	13.000000	56.000000	56.000000	56.000000	54.000000
75%	17.000000	20.000000	20.000000	17.000000	74.000000	74.000000	74.000000	72.000000
max	25.000000	25.000000	25.000000	25.000000	100.000000	100.000000	100.000000	100.000000

	Gender	Location
count	3500	3500
unique	2	7
top	Male	Madinah
freq	1756	519

A-2:

Field	Actual Data Type	Python Data Type	Consistent ?
Gender	Categorical	object	YES
Location	Categorical	object	YES
Quiz-1	Numerical	int64	YES
Quiz-2	Numerical	int64	YES
Quiz-3	Numerical	int64	YES
Quiz-4	Numerical	int64	YES
Major-1	Numerical	int64	YES
Major-2	Numerical	int64	YES
Major-3	Numerical	int64	YES
Final	Numerical	int64	YES

A-3:

Note: Scaling/normalizing data typically improves the analysis.

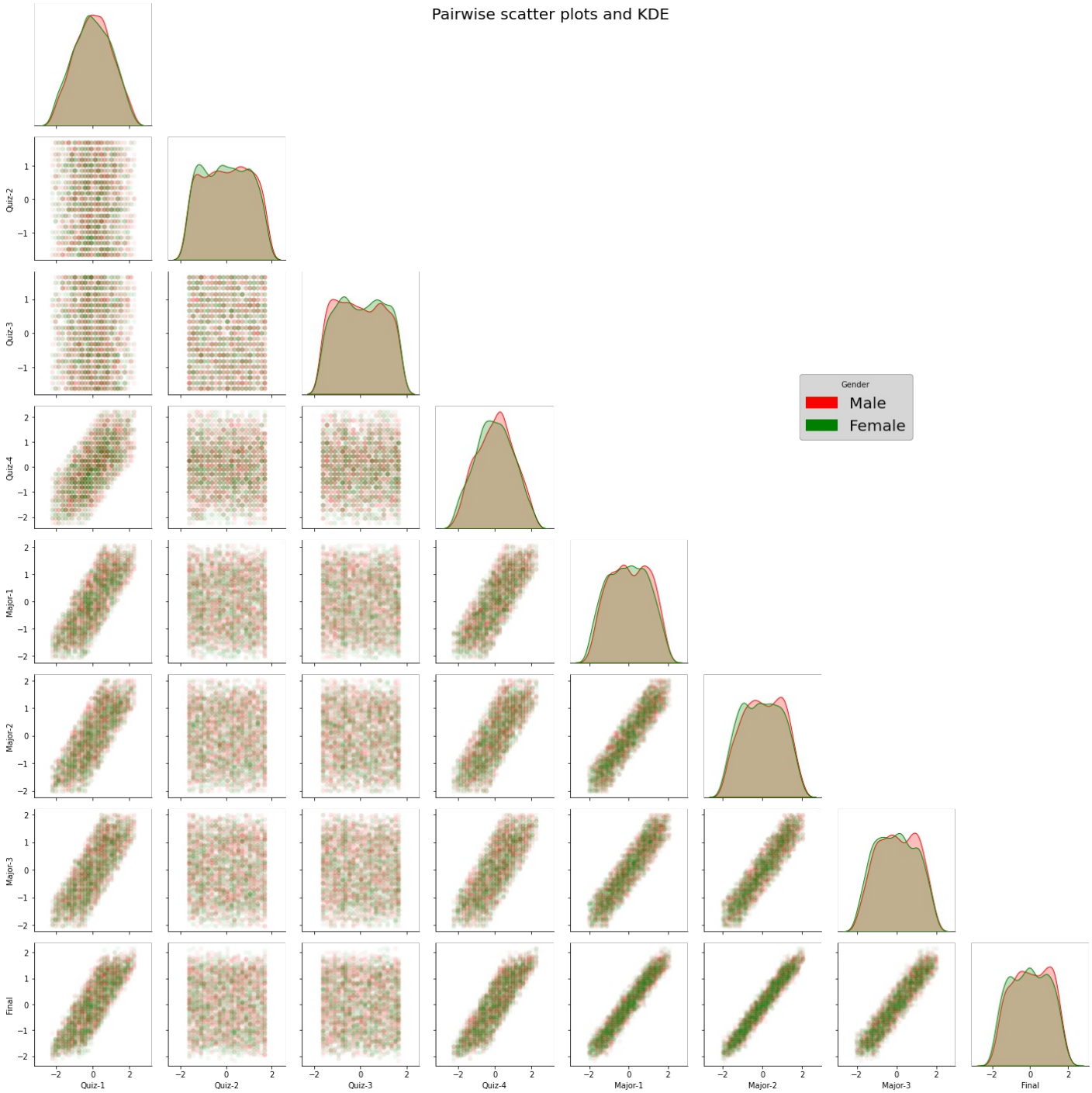
Following data frame shows the normalized columns:

	Gender	Location	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
0	Male	Jeddah	0.816245	0.364147	1.489470	0.619522	2.043890	1.658954	1.323202	1.493295
1	Female	Jeddah	-0.429814	1.030734	-0.151690	-0.806188	-1.854057	-1.286137	-1.796073	-1.355053
2	Female	Jeddah	-0.963839	1.030734	1.325354	-0.984402	-0.044296	-0.503847	-0.695153	-0.514557
3	Male	Dammam	1.706287	-0.802381	0.997122	-0.093333	1.115807	0.646579	0.497511	0.652799
4	Female	Dammam	1.528278	0.030853	0.340658	0.975949	1.162211	1.060732	1.002100	1.259824
...	...	...	...	...	...	...	...	...	...	...
3495	Female	Riyadh	0.460228	1.697322	0.833006	0.441308	1.440636	1.704971	1.506688	1.259824
3496	Male	Dammam	0.460228	-0.135793	0.340658	-0.806188	-0.369125	-0.503847	-0.465794	-0.514557
3497	Male	Riyadh	1.350270	1.364028	0.833006	0.975949	0.698170	0.968698	1.185587	0.886270
3498	Female	Riyadh	-0.429814	0.197500	-1.464618	-0.093333	0.326937	0.554545	-0.328179	0.279245
3499	Male	Dammam	-0.073797	-0.469087	1.489470	-1.340829	-0.879570	-0.918001	-1.383228	-0.934805

3500 rows × 10 columns

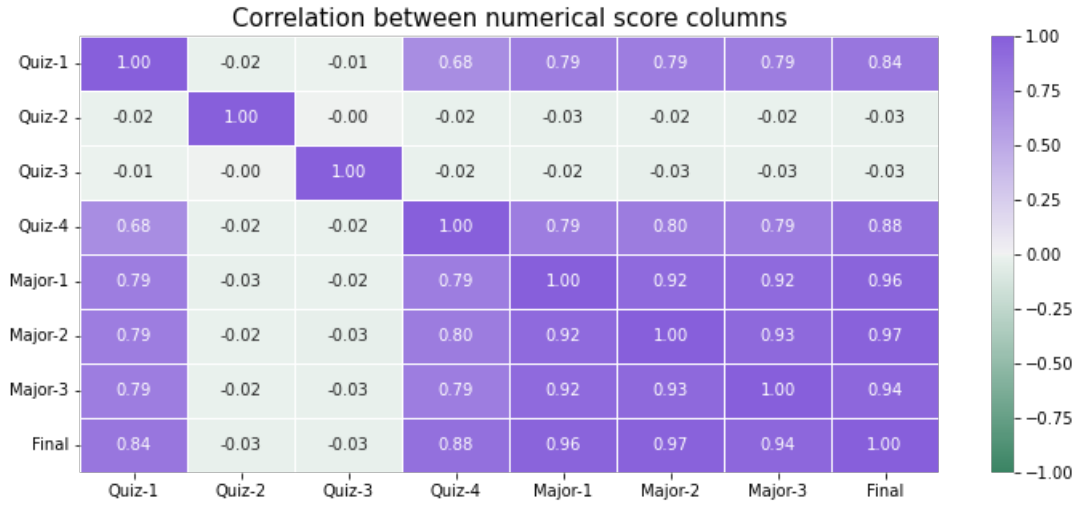
A-4:

Pairwise scatter plots for numerical data are useful in identifying pairwise relationships.



A-5:

Coloring correlation dataframe may reveal hidden patterns. For example: Quiz-2 and Quiz-3 seems to be very weakly related to the other columns.



The top 3 variables that are highly correlated with 'Final' are: Major-2, Major-1, Major-3.

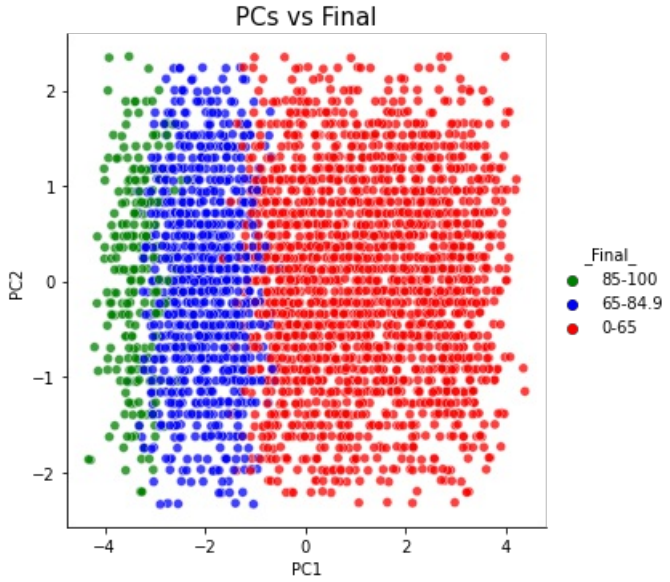
The pair of score columns that are strongly correlated are 'Major-2' and 'Final'.

A-6:

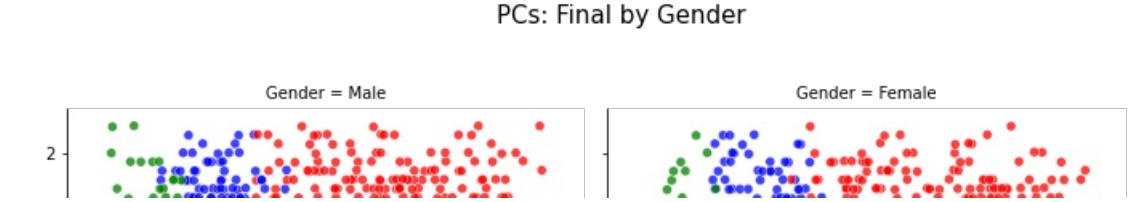
Principal components contains most of the information from its input columns. Thus, they are very useful in visualizations.

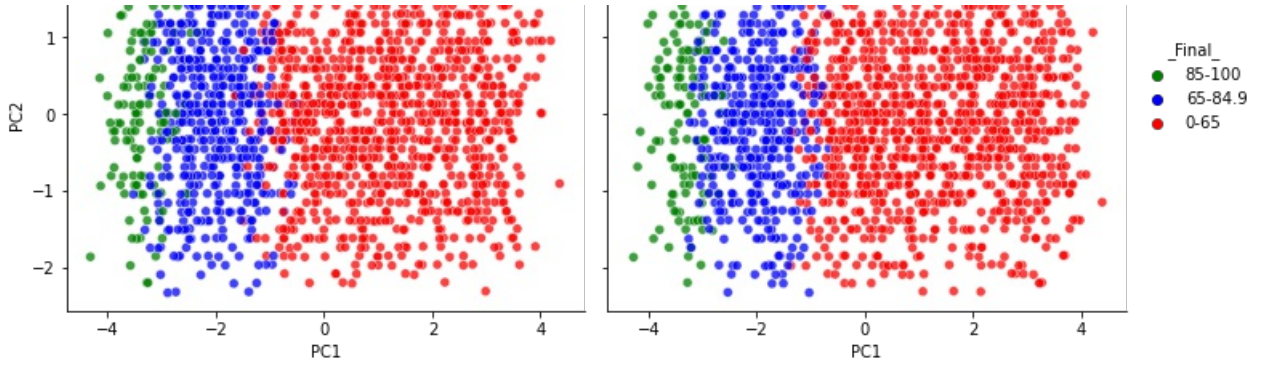
Note: The normalized values of the data are good for analysis, however, not good for visualization. Thus, always use the original data column values for coloring, labeling, etc.

	Gender	Location	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final	PC1	PC2
0	Male	Jeddah	0.816245	0.364147	1.489470	0.619522	2.043890	1.658954	1.323202	1.493295	-2.903284	-0.798170
1	Female	Jeddah	-0.429814	1.030734	-0.151690	-0.806188	-1.854057	-1.286137	-1.796073	-1.355053	2.820426	0.828534
2	Female	Jeddah	-0.963839	1.030734	1.325354	-0.984402	-0.044296	-0.503847	-0.695153	-0.514557	1.435320	-0.210532
3	Male	Dammam	1.706287	-0.802381	0.997122	-0.093333	1.115807	0.646579	0.497511	0.652799	-1.721611	-1.281249
4	Female	Dammam	1.528278	0.030853	0.340658	0.975949	1.162211	1.060732	1.002100	1.259824	-2.543813	-0.227045

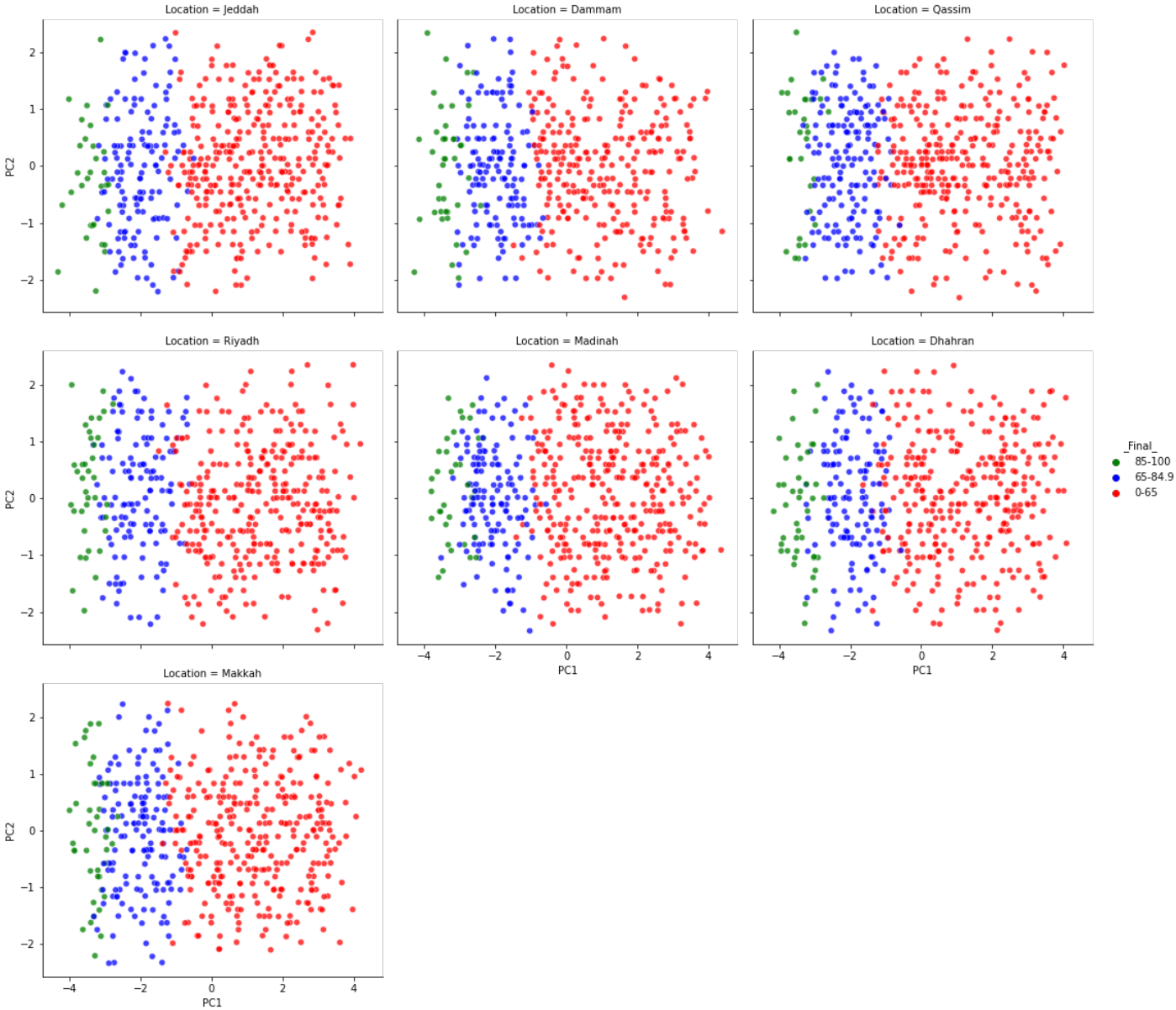


The PCs show good separation, with some overlap at the intersection.





PCs: Final by Location



The variance captured by PC 1 is: 61.31%
The variance captured by PC 2 is: 14.33%

The coefficients (the u vector) of the linear combination of input variables for the first PC:

	field	coeff
5	Major-2	-0.4638
6	Major-3	-0.4628
4	Major-1	-0.4626
3	Quiz-4	-0.4227
0	Quiz-1	-0.4213
2	Quiz-3	0.0163

	field	coeff
1	Quiz-2	0.0151

Problem-B

----------Problem #B----------

B-1:

Displaying random rows of data:

	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
1835	13	25	13	12	65	64	67	59
2663	7	8	12	12	33	45	36	37
916	14	5	21	16	55	58	45	59
1269	19	11	20	21	79	81	74	84
2632	11	10	13	10	48	45	50	45
2888	16	17	10	15	48	55	56	53
2475	10	14	11	15	71	61	63	61
1413	19	14	17	16	71	83	66	78
835	10	15	19	8	13	23	22	24
718	8	6	21	9	31	36	44	31

The number of rows: 2899
The number of columns: 8
The number of non-null rows for each column:
Quiz-1 2899
Quiz-2 2899
Quiz-3 2899
Quiz-4 2899
Major-1 2899
Major-2 2899
Major-3 2899
Final 2899
dtype: int64

No missing values.
The summaries of numerical column:

	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
count	2899.000000	2899.000000	2899.000000	2899.000000	2899.000000	2899.000000	2899.000000	2899.000000
mean	12.574681	15.077268	15.058296	12.563298	55.942049	55.930321	55.997240	53.903415
std	5.526165	6.101757	6.115778	5.688581	21.771147	21.761616	21.796679	21.396040
min	0.000000	5.000000	5.000000	0.000000	12.000000	12.000000	12.000000	7.000000
25%	9.000000	10.000000	10.000000	8.000000	37.000000	38.000000	37.000000	36.000000
50%	13.000000	15.000000	15.000000	13.000000	56.000000	56.000000	56.000000	54.000000
75%	17.000000	20.000000	20.000000	17.000000	74.000000	75.000000	74.000000	72.000000
max	25.000000	25.000000	25.000000	25.000000	100.000000	100.000000	100.000000	100.000000

B-2:

Note: For various reasons default data-type understood by python may not be consistent.

Field	Actual Data Type	Python Data Type	Consistent ?
Quiz-1	Numerical	float64	YES
Quiz-2	Numerical	float64	YES
Quiz-3	Numerical	float64	YES
Quiz-4	Numerical	float64	YES
Major-1	Numerical	float64	YES
Major-2	Numerical	float64	YES
Major-3	Numerical	float64	YES
Final	Numerical	float64	YES

There are no inconsistencies in the data.

B-3:

Note: To scale the entire dataframe in a single shot, you can use `pd.DataFrame(StandardScaler().transform(df), columns=df.columns)`.

The summaries of all the columns:

	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
count	2.899000e+03	2.899000e+03	2.899000e+03	2.899000e+03	2.899000e+03	2.899000e+03	2.899000e+03	2.899000e+03
mean	-5.572178e-17	6.510449e-19	-4.997727e-17	-2.509587e-16	2.768856e-17	-1.097585e-16	2.680773e-18	1.433831e-16
std	1.000173e+00	1.000173e+00	1.000173e+00	1.000173e+00	1.000173e+00	1.000173e+00	1.000173e+00	1.000173e+00
min	-2.275873e+00	-1.651821e+00	-1.644931e+00	-2.208893e+00	-2.018710e+00	-2.019055e+00	-2.018878e+00	-2.192532e+00
25%	-6.469762e-01	-8.322430e-01	-8.272322e-01	-8.023240e-01	-8.702030e-01	-8.240846e-01	-8.717162e-01	-8.369074e-01
50%	7.697787e-02	-1.266543e-02	-9.533704e-03	7.678148e-02	2.662286e-03	3.202484e-03	1.266270e-04	4.514933e-03
75%	8.009320e-01	8.069121e-01	8.081648e-01	7.800659e-01	8.295873e-01	8.764500e-01	8.260830e-01	8.459372e-01
max	2.248840e+00	1.626490e+00	1.625863e+00	2.186635e+00	2.024034e+00	2.025460e+00	2.019131e+00	2.154816e+00

B-4:

In this task, we are NOT fitting a new scaler object. We use the scaler object form B-3 and transform the new data.

The summaries of all the columns:

	Quiz-1	Quiz-2	Quiz-3	Quiz-4	Major-1	Major-2	Major-3	Final
count	324.000000	324.000000	324.000000	324.000000	324.000000	324.000000	324.000000	324.000000
mean	-0.044798	-0.088552	0.116654	-0.108265	-0.059300	-0.078363	-0.034855	0.013316
std	1.031785	1.034225	0.956155	1.038197	0.999085	1.021849	1.021145	1.075939
min	-2.275873	-1.651821	-1.644931	-2.208893	-2.018710	-1.927134	-2.018878	-2.005549
25%	-0.692223	-0.996158	-0.663692	-0.802324	-0.870203	-0.916005	-0.883188	-0.942085
50%	-0.104011	-0.176581	0.154006	-0.099040	-0.089218	-0.134679	0.000127	0.004515
75%	0.800932	0.806912	0.971704	0.604245	0.737707	0.784529	0.826083	0.951115
max	2.067852	1.626490	1.625863	2.186635	1.886214	1.979500	1.927358	2.154816

B-5:

When the data is scaled such that mean is zero, and standard deviation is one.
You do NOT have to add columns of all ones. Since beta0 will be zero.
The closed form estimates are:

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
B-5	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0

B-6:
The closed form estimates are:

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
B-5	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0
B-6	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0

No differences in between the coefficients from Parts B-5 and B-6.

The MSE for HW5DataC is :0.01842862368733383

B-7:
The MSE using Ridge is: 0.018428417221315458

B-8:
The MSE using Lasso is: 0.018509010803085785

B-9:

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
B-5	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0
B-6	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0
B-7	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0
B-8	0.0	0.1	-0.0	-0.0	0.22	0.32	0.42	0.0

All the three methods gives exactly similar coefficient estimates.
