University of Waterloo
CFM 301
Data Assignment 3
Jeongseop Yi (j22yi)

Q1. The data for Q1 is in Q1 sheet in the DA3.xlsx file in the assignment package.
The summary statistic of BETA is as follows:

	BETA
count	20308.000000
mean	1.257083
std	0.795173
min	-2.228347
1%	-0.116178
5%	0.267463
25%	0.773262
50%	1.111727
75%	1.561689
95%	2.812702
99%	3.890250
max	7.395251

Q2. The data for Q2 is in Q2 sheet in the DA3.xlsx file in the assignment package.
I used the built-in std() function of pandas DataFrame. I found out that the function uses sample standard deviation method which uses n-1 for the divisor. The IVOL values are a bit higher than the values calculated using population standard deviation method which uses n for the divisor.
The summary statistic of IVOL is as follows:

	IVOL
count	20775.000000
mean	0.016686
std	0.012257
min	0.001331
1%	0.004099
5%	0.005521
25%	0.008871
50%	0.013138
75%	0.020306
95%	0.040426
99%	0.062578
max	0.216573

Q3. The data for Q3 is in Q3 sheet in the DA3.xlsx file in the assignment package.

The summary statistic of MOM is as follows:

	MOM
count	20381.000000
mean	0.282410
std	0.756893
min	-0.972296
1%	-0.653142
5%	-0.385049
25%	-0.026301
50%	0.179303
75%	0.424287
95%	1.185185
99%	2.793935
max	26.372883

Q4. The data for Q4 is in Q4 sheet in the DA3.xlsx file in the assignment package.

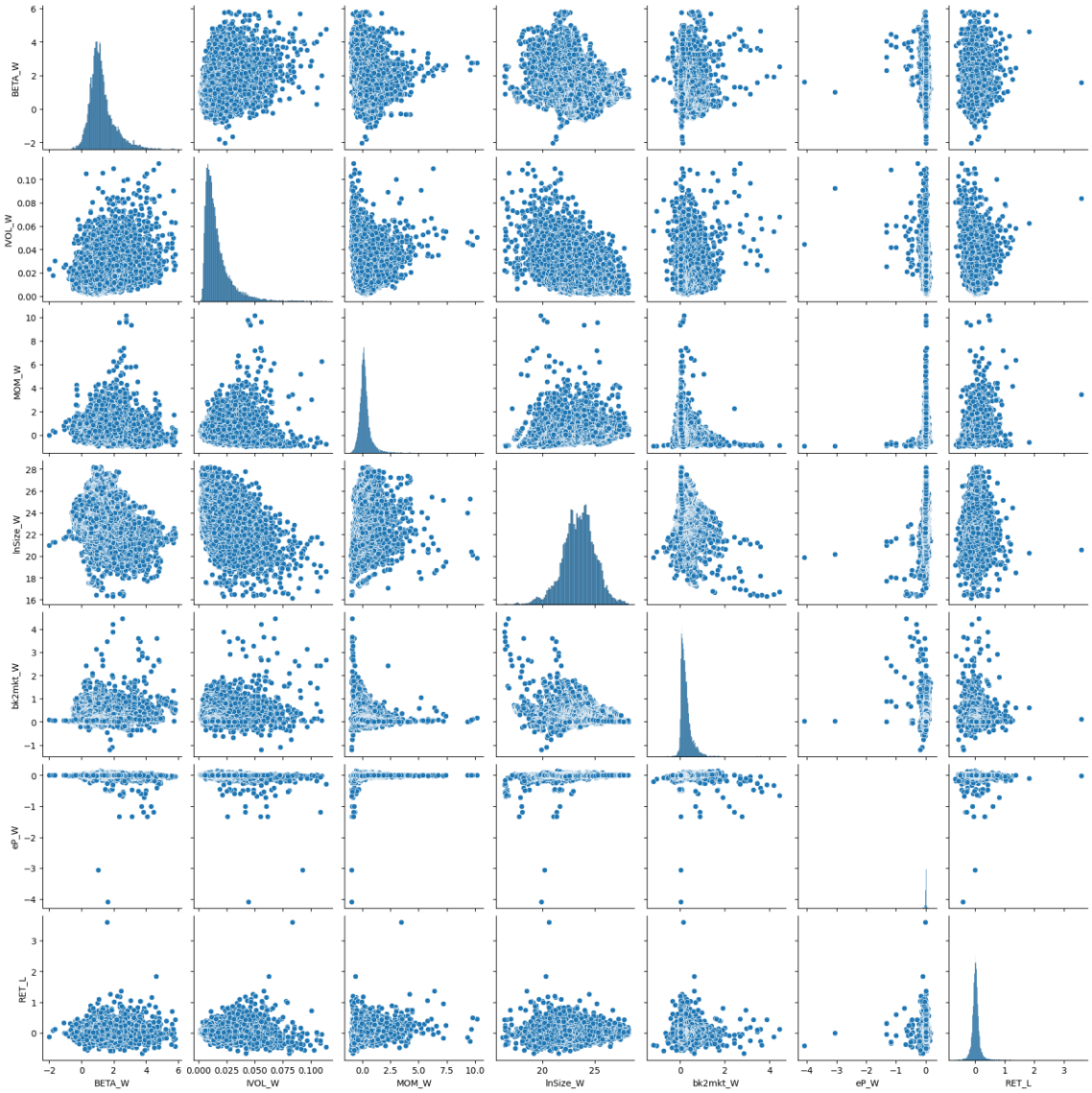
The summary statistic of winsorized BETA, IVOL, and MOM are as follows:

	BETA_W	IVOL_W	MOM_W
count	20308.000000	20775.000000	20381.000000
mean	1.254927	0.016489	0.263819
std	0.780144	0.011419	0.569629
min	-2.016456	0.001331	-0.972296
1%	-0.116178	0.004099	-0.648930
5%	0.267463	0.005521	-0.385049
25%	0.773262	0.008871	-0.026301
50%	1.111727	0.013138	0.179303
75%	1.561689	0.020306	0.424267
95%	2.778597	0.039294	1.175456
99%	3.840429	0.058917	2.332290
max	5.799785	0.113657	10.179186

Q5. The data for Q5 is in Q5 sheet in the DA3.xlsx file in the assignment package. In my solution of DA2, I included the 6 duplicate data entries from SKYWORKS SOLUTIONS INC. Therefore, there

are 6 duplicate values for BETA_W, IVOL_W, MOM_W, which may add some errors to my results.

a) The scatterplot between the explained variables and each of the explanatory variables is as follows:



b) The correlation matrix between the explained variables and each of the explanatory variables is as follows:

	BETA_W	IVOL_W	MOM_W	lnSize_W	bk2mkt_W	eP_W	RET_L
BETA_W	1.000000	0.373041	0.019579	-0.293981	0.004506	-0.152358	0.003516
IVOL_W	0.373041	1.000000	0.075432	-0.434295	0.041815	-0.216811	-0.052344
MOM_W	0.019579	0.075432	1.000000	0.043115	-0.278687	0.052670	0.253463
lnSize_W	-0.293981	-0.434295	0.043115	1.000000	-0.176558	0.166766	0.007091
bk2mkt_W	0.004506	0.041815	-0.278687	-0.176558	1.000000	-0.058081	-0.093186
eP_W	-0.152358	-0.216811	0.052670	0.166766	-0.058081	1.000000	-0.003010
RET_L	0.003516	-0.052344	0.253463	0.007091	-0.093186	-0.003010	1.000000

The p-value matrix is as follows:

	BETA_W	IVOL_W	MOM_W	lnSize_W	bk2mkt_W	eP_W	RET_L
BETA_W	0.0	0.0	0.005273	0.0	0.52298	0.0	0.616368
IVOL_W	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MOM_W	0.005273	0.0	0.0	0.0	0.0	0.0	0.0
lnSize_W	0.0	0.0	0.0	0.0	0.0	0.0	0.307408
bk2mkt_W	0.52298	0.0	0.0	0.0	0.0	0.0	0.0
eP_W	0.0	0.0	0.0	0.0	0.0	0.0	0.666267
RET_L	0.616368	0.0	0.0	0.307408	0.0	0.666267	0.0

Based on the p-values, the null hypothesis of no correlation between the variables must be rejected for all pairs of variables except for between BETA_W and bk2mkt_W, BETA_W and RET_L, InSize_W and RET_L, and eP_W and RET_L, where the two variables have higher p-value than 0.05. The result follows in the correlation coefficient matrix, where the values are greater than 0 except for the pairs of variables which have p-value greater than 0.05.