A. Overview of the dataset

Observing 10 variables, the dataset has 236,549 entries, in which the numbers of missing data are as follow:

Figure 1. Number of null objects in the original dataset

	Null objects
prod_day	0
part_type	0
material	3400
part_number	0
nozzle_diameter	0
thickness	7922
standoff_distance	0
traverse_speed	0
kerf	4398
Ra	2369

In addition, to see find the relationships among qualitative variables, the dataset is split into subgroups based on part_type (A, B, C), part_number (P0101, P0102, P0201, P0202, P0301, P0302), and material (MTL1, MTL2).

Figure 2. General description for subgroups of data

			prod_day			nozzle_diameter			thickness			standoff_distance			traverse_speed		
			count	mean	std	count	mean	std	count	mean	std	count	mean	std	count	mean	std
part_type	part_number	material															
A	P0101	MTL1	38188	44.577	26.083	38188	0.95	1.82E-13	36325	0.900	0.018	38188	56.013	7.549	38188	208.662	13.968
		MTL_1	572	45.664	25.713	572	0.95	4.78E-15	542	0.899	0.018	572	56.049	7.605	572	208.227	13.771
	P0102	MTL2	30791	44.488	25.915	30791	0.95	4.88E-13	30315	1.250	0.025	30791	55.938	7.423	30791	208.534	13.808
		MTL_2	456	45.607	26.444	456	0.95	7.34E-15	453	1.248	0.025	456	56.298	7.851	456	209.264	13.924
B	P0201	MTL1	36660	44.425	25.923	36660	1.2	3.6E-13	34884	0.900	0.018	36660	56.023	7.537	36660	208.626	13.962
		MTL_1	552	44.716	25.668	552	1.2	4.22E-15	525	0.902	0.017	552	56.229	7.874	552	209.398	14.070
	P0202	MTL2	29887	44.688	25.876	29887	1.2	2.18E-13	29421	1.250	0.025	29887	55.989	7.544	29887	208.643	13.976
		MTL_2	435	43.021	24.595	435	1.2	7.11E-15	430	1.249	0.025	435	56.446	7.830	435	208.986	14.385
C	P0301	MTL1	51984	44.605	26.007	51984	1.2	1.11E-12	49505	0.900	0.018	51984	55.976	7.484	51984	208.650	13.843
		MTL_1	759	45.581	26.471	759	1.2	1.53E-14	730	0.899	0.018	759	56.305	7.604	759	209.452	14.074
	P0302	MTL2	42271	44.687	25.968	42271	1.5	0	41628	1.250	0.025	42271	56.006	7.498	42271	208.687	13.903
		MTL_2	594	45.891	26.150	594	1.5	0	579	1.250	0.025	594	55.804	7.370	594	207.934	13.896

From Figure 2, it can be seen that the **labels of part_number are based on part_type and the material** used for the piece. Meanwhile, the values of nozzle_diameter are quite consistent within each subgroup: 0.95mm for type A, 1.2mm for type B and type C, except for MTL2 type C, in which nozzle_diameter is 1.5mm. Likewise, the value thickness is also closely related to material, as mentioned in the document of Assignment 2.

B. Data preparation

I. Missing data

From the insights gained in part A, we can fill in the missing qualitative data based on the other variables in each observation. Detailed steps with explanations are presented in the Jupyter notebook, here only the ideas are presented:

- The incorrect values of material (MTL_1, MTL_2, MTL_3) are to be corrected.
- The missing values of material can be derived from part_number, which has no missing value.
- The missing values thickness are to be filled by the nominal thickness corresponding to the material used in that observation.

After the missing qualitative data has been filled, we review again values of kerf and Ra.

Figure 3. Descriptions of kerf and Ra values in each subgroup

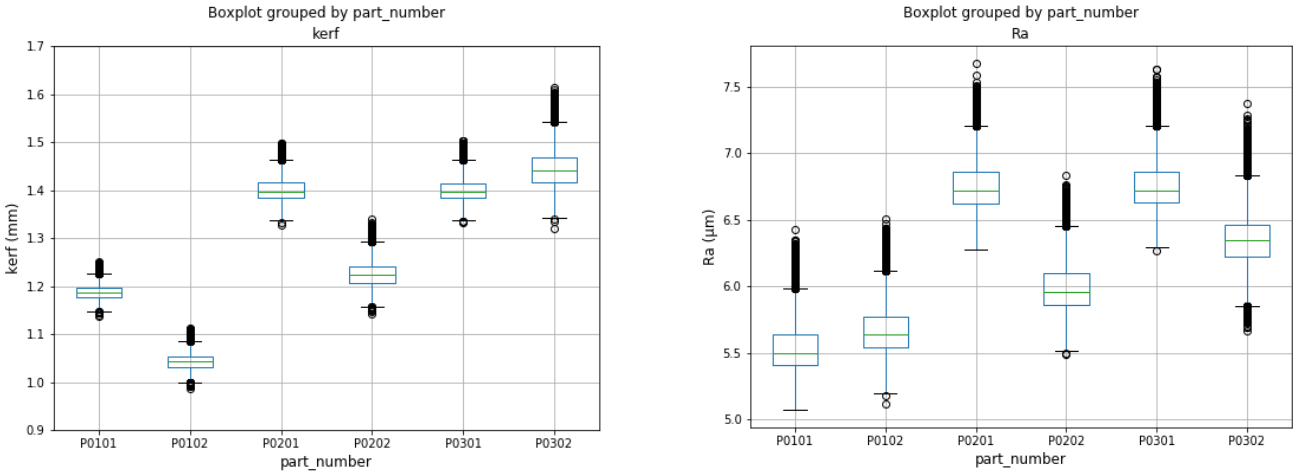
			kerf				Ra			
			count	mean	std	CV	count	mean	std	CV
part_type	part_number	material								
A	P0101	MTL1	37996	1.187305	0.015223	0.012821	38760	5.535567	0.184225	0.03328
	P0102	MTL2	30692	1.043695	0.016686	0.015987	31247	5.672668	0.181774	0.032044
B	P0201	MTL1	36521	1.401218	0.024919	0.017784	37212	6.756432	0.183163	0.027109
	P0202	MTL2	29769	1.226317	0.026339	0.021478	30322	5.991499	0.186803	0.031178
C	P0301	MTL1	51777	1.401065	0.024816	0.017712	52740	6.75567	0.181741	0.026902
	P0302	MTL2	42058	1.445232	0.039588	0.027392	40531	6.34844	0.198946	0.031338

From Figure 3, values of kerf and Ra in each subgroup have mean significantly different from one another and coefficients of variance (CV) are all small, about 3% or below. Hence, the missing values of kerf and Ra in each subgroup will be filled with the mean values of kerf and Ra in that subgroup.

II. Visual inspection of outliers

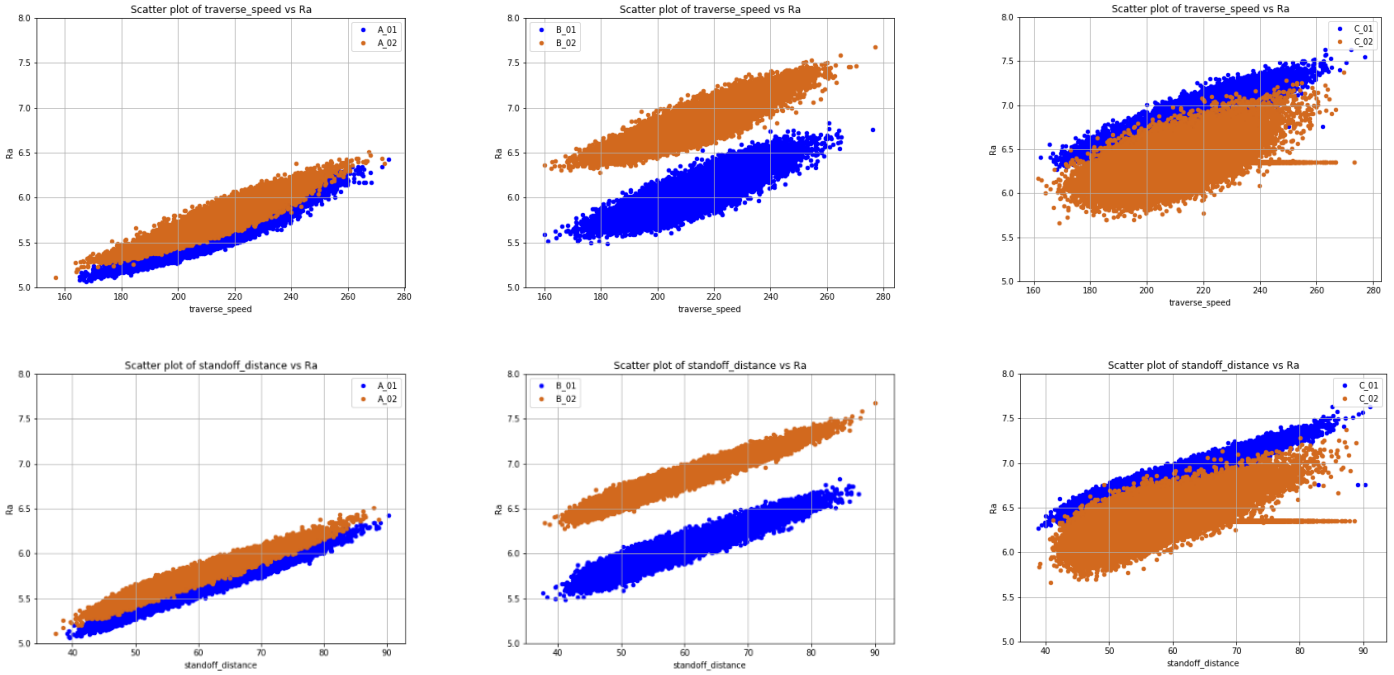
The univariate analysis of kerf and Ra seems to show quite many outliers.

Figure 4. Boxplot for kerf and Ra value, grouped by part_number



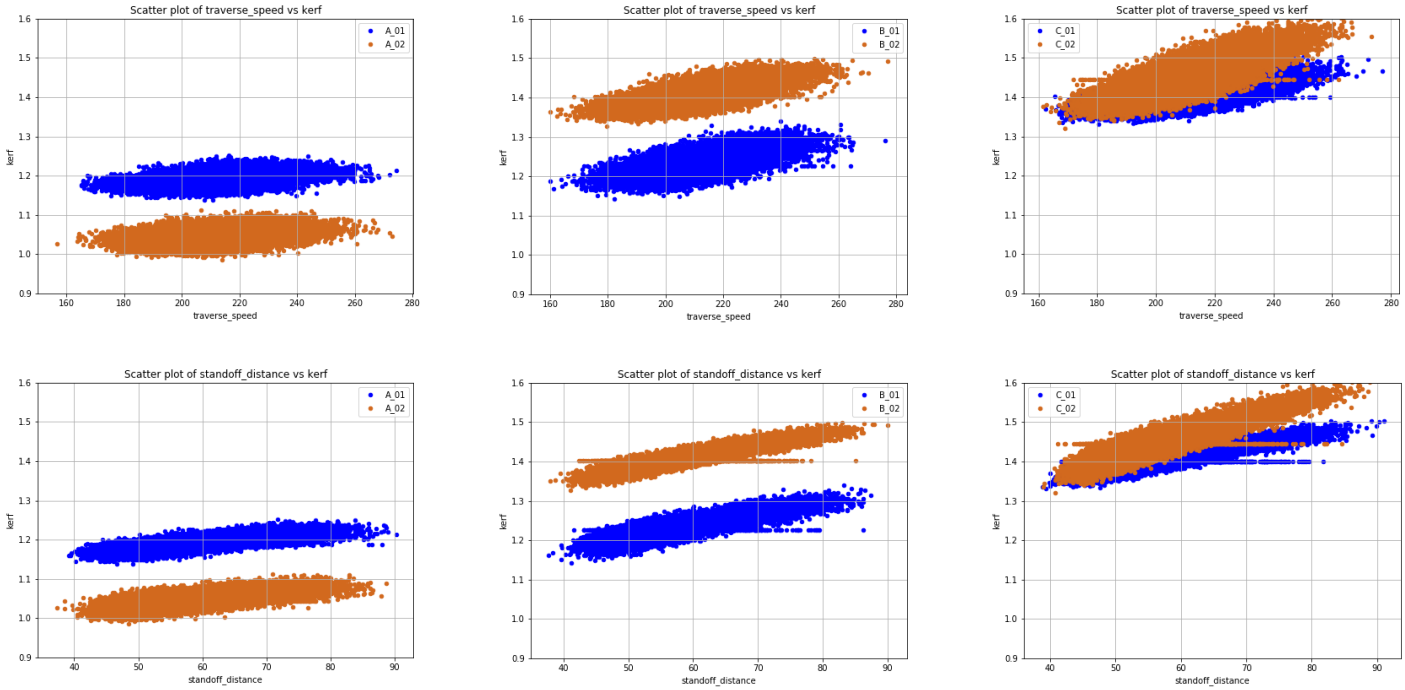
However, the scatter plots within each subgroup show that these data points still have consistent relationship with another dimensions (e.g. traverse_speed and standoff_distance). Hence, the Ra and kerf will not be adjusted for outliers.

Figure 5. Scatter plots of traverse_speed and standoff_distance vs Ra within each subgroup (*)



(*) There are some unusual data points in group C_01 and C_02 because missing Ra values are filled with average Ra values of the groups.

Figure 6. Scatter plots of traverse_speed and standoff_distance vs kerf within each subgroup



The special relationships between traverse_speed, standoff_distance, and the quality measures (Ra and kerf) can suggest important ideas to identify the main drivers of the quality measures.

C. Identify the main drivers of the quality measures

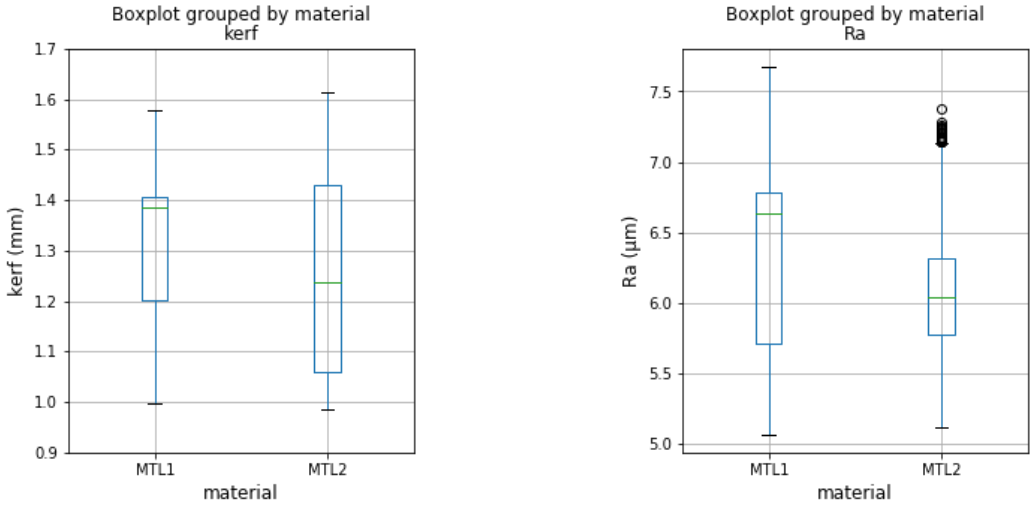
I. Data visualization

In addition to the above-mentioned relationships between traverse_speed, standoff_distance and the quality measures, we will try to visualize the other variables, including material, prod_day, nozzle_diameter, and thickness, to see if they have any relationship with the quality measures.

1. Material

The choice of material shows no clear impact on the values of the quality measures.

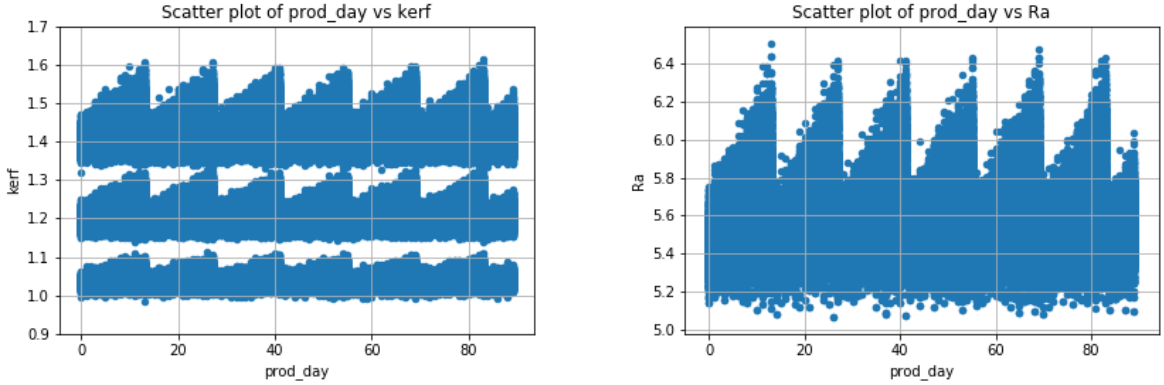
Figure 7. Conditional boxplot for kerf and Ra, grouped by material



2. The day of production (prod_day)

The graphs in figure 8 shows that the values of quality measures tend to vary more in some certain days. However, they exhibit no correlation relationship in the two pairs of dimension.

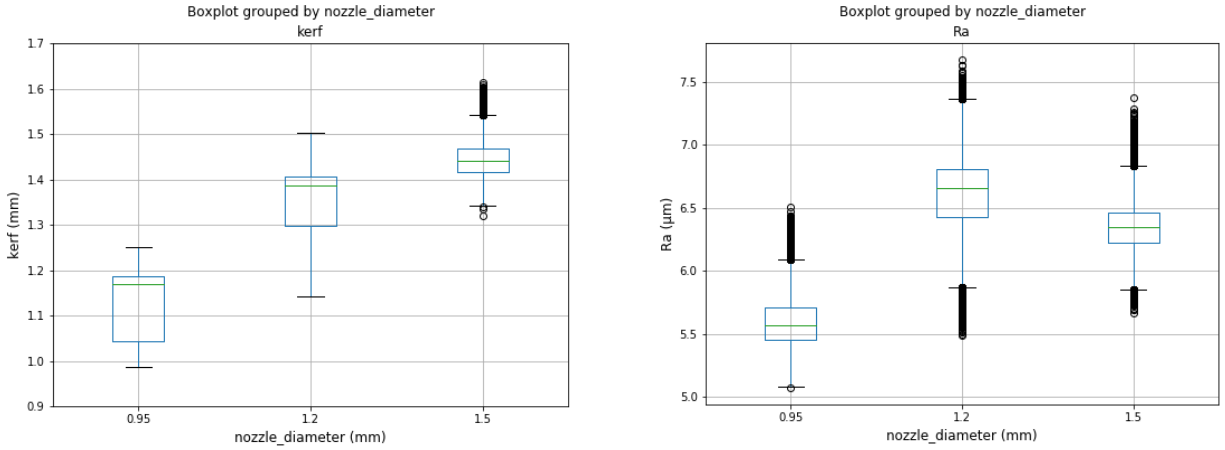
Figure 8. Scatter plots of prod_day and quality measures on the whole dataset



3. The diameter of the nozzle (nozzle_diameter)

Figure 9 suggests that the diameter of the nozzle may have a positive correlation with the kerf width but correlation with Ra. This makes sense since the bigger the nozzle diameter is, the wider the kerf it leaves on the material.

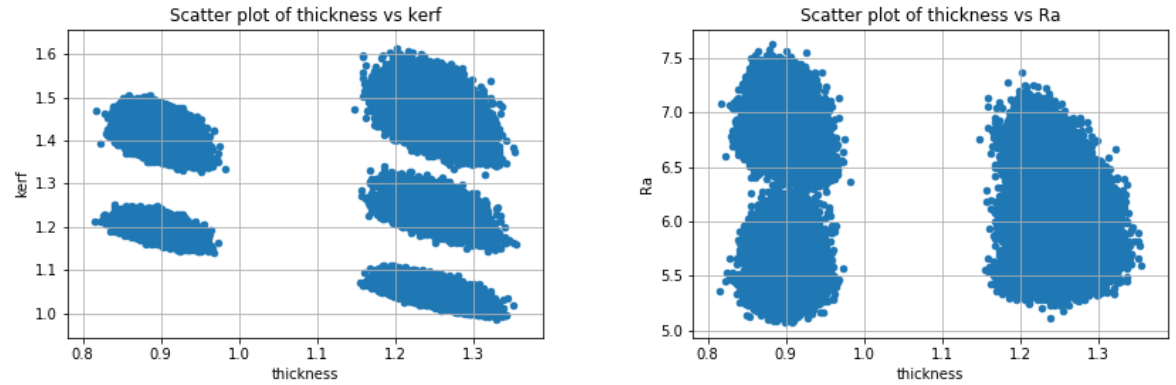
Figure 9. Conditional boxplot for kerf and Ra, grouped by nozzle_diameter



4. Thickness

The scatter plots in figure 10 shows no clear correlation relationship between thickness and the values of the quality measures.

Figure 10. Scatter plots of thickness and quality measures on the whole dataset



II. Quantitative statistical measures

The correlation matrix run on the whole dataset has confirmed the relationship between nozzle diameter and kerf with high correlation at 0.805. In addition, kerf and Ra also exhibits high correlation at 0.845, suggesting that they may have some common drivers (traverse_speed and standoff_distance).

Figure 11. Correlation matrix of the data set

	prod_day	nozzle_diameter	thickness	standoff_distance	traverse_speed	kerf	Ra
prod_day	1	0.00174	0.00127	0.02266	0.01936	0.00449	0.00695
nozzle_diameter	0.00174	1	0.3253	0.00079	0.00171	0.80506	0.55938
thickness	0.00127	0.3253	1	-0.00178	-0.00085	-0.27084	-0.34348
standoff_distance	0.02266	0.00079	-0.00178	1	0.835	0.15757	0.32459
traverse_speed	0.01936	0.00171	-0.00085	0.835	1	0.12253	0.29836
kerf	0.00449	0.80506	-0.27084	0.15757	0.12253	1	0.84548
Ra	0.00695	0.55938	-0.34348	0.32459	0.29836	0.84548	1

Meanwhile, the correlation matrices run within each subgroup also **confirm standoff_distance and traverse_speed as another drivers for both quality measures** as most of correlation coefficients are ranging in the [0.5, 0.9] interval or above.

Figure 12. Correlation matrix within 6 subgroups

subgroup		standoff_distance	traverse_speed	kerf	Ra
A_01	standoff_distance	1	0.83468	0.77224	0.97617
	traverse_speed	0.83468	1	0.43423	0.93234
	kerf	0.77224	0.43423	1	0.65322
	Ra	0.97617	0.93234	0.65322	1
A_02	standoff_distance	1	0.83433	0.69117	0.97512
	traverse_speed	0.83433	1	0.38194	0.93194
	kerf	0.69117	0.38194	1	0.56418
	Ra	0.97512	0.93194	0.56418	1
B_01	standoff_distance	1	0.83568	0.91906	0.96721
	traverse_speed	0.83568	1	0.71883	0.88318
	kerf	0.91906	0.71883	1	0.95264
	Ra	0.96721	0.88318	0.95264	1
B_02	standoff_distance	1	0.83468	0.86591	0.94709
	traverse_speed	0.83468	1	0.67266	0.86322
	kerf	0.86591	0.67266	1	0.94487
	Ra	0.94709	0.86322	0.94487	1
C_01	standoff_distance	1	0.83516	0.91721	0.96599
	traverse_speed	0.83516	1	0.71622	0.88175
	kerf	0.91721	0.71622	1	0.95184
	Ra	0.96599	0.88175	0.95184	1
C_02	standoff_distance	1	0.83525	0.91245	0.64832
	traverse_speed	0.83525	1	0.79146	0.57102
	kerf	0.91245	0.79146	1	0.83257
	Ra	0.64832	0.57102	0.83257	1

D. Summary of the drivers and ideas for future improvements of the process

In summary, there are three drivers for the two quality measures: standoff_distance and traverse_speed have positive correlations with both kerf and Ra; nozzle_diameter has positive correlation with kerf. Among them, standoff_distance also has high positive correlation with traverse_speed. No matter if one driver is the explanatory factor for the other, this relationship suggests that there is no tradeoff between the two drivers. For future improvements of the process:

- We can reduce standoff_distance and traverse_speed at the same time to reduce both kerf and Ra, thus improving quality measures.
- We can also use smaller nozzle_diameter to reduce kerf width to enhance quality.