Data Cleaning Report:

Step 00: Importing pandas and chardet library

To start with, I need to import the pandas library, which is a powerful data manipulation tool, and the chardet library, which helps in detecting the encoding of a file.

- import pandas as pd
- import chardet

Step 01: Connecting 2 CSV files into the project

- Loaded two CSV files into the project using pandas.
- File names:
  - 01_DAAN_545_Dataset1.csv
  - 02_DAAN_545_Dataset2.csv
- Encoding used: "iso-8859-1" format

For example code,

- df = pd.read_csv("01_DAAN_545_Dataset1.csv")
- df2 = pd.read_csv("02_DAAN_545_Dataset2.csv")

Step 02: Append 2 datasets together

The code appends the 2 dataframes together using

- pd.concat()

and stores the result in the dataframe name

- appended_df

For example code,

appended_df = pd.concat([df, df2], axis=0, ignore_index=True)

Step 03: Show how many missing values each column has and where are the row locations of missing values

The code calculates the number of missing values in each column using the function.

- isnull().sum()

Moreover, I record row locations that missing values happening.

- rows_with_nulls = appended_df[appended_df.isnull().any(axis=1)] # Row number with null data

Step 04: Filter Rows

Filtered out rows where [Order_Number] is not null and [BCT1] is null, as these rows are not useful for box strength analytics.

For example code,

- condition = appended_df['Order_Number'].notnull() & appended_df['BCT1'].isnull()
- appended_df = appended_df[~condition]

Index	Rwf	Order_Numb	BCT1	BCT2	BCT3	BCT4	BCT5	BCT_Av	BCT_Sp	ECT1	ECT2	ECT3
334	258	3913157101	273.8739	262.071	268.1293	263.9866	261.6953	265.9512	236.8	5.563804	5.504179	5.397017
417	258	3909976305	265.805	266.7601	264.9592	270.2103	262.0092	265.9488	236.8	6.679213	6.142616	6.243136
518		3913738201										
1087		3918280402										
1282		3913697502										
1632		3912296901										
1880	258	3919427202	264.4114	272.0429	271.1536	271.6856	263.3156	268.5218	236.8	5.945766	5.801021	6.364407

Step 05: Mapping for Raw Material Strength Additive column

Created a new column [Raw_Mat_Strength_Additive_Percent] by mapping values from [Raw_Mat_Strength_Additive] to their respective percentages.

For example,

- strength_mapping = {
  'G0': '0%',
  'G1': '5%',
  'G2': '10%',
  'G3': '15%',
  'G4': '20%',
  'G5': '25%',
  'G6': '30%'
  }
- appended_df['Raw_Mat_Strength_Additive_Percent'] = appended_df['Raw_Mat_Strength_Additive'].map(strength_mapping)

Step 06: Replace Null Values in Multiple Columns:

Replaced null values in specific columns with 'No Use'. Columns affected: GL, GLWeigth(Grammage), BM, BMWeigth(Grammage), BL, BLWeigth(Grammage), CM, CMWeigth(Grammage),CL ,CLWeigth(Grammage).

For example,

- columns_to_fill = ['GL','GLWeigth(Grammage)', 'BM', 'BMWeigth(Grammage)', 'BL', 'BLWeigth(Grammage)', 'CM', 'CMWeigth(Grammage)', 'CL', 'CLWeigth(Grammage)']
- appended_df[columns_to_fill] = appended_df[columns_to_fill].fillna('No Use')

Raw_Mat_Combination	GL	GLWeigth(Grammage)	BM	BMWeigth(Grammage)	BL	BLWeigth(Grammage)	CM	CMWeigth(Grammage)	CL	CLWeigth(Grammage)
KT125/CS110/TD125	KT	125	CS	110	TD	125	No Use	0	No Use	0
KT125/CS110/CS110/CS110/TD125	KT	125	CS	110	CS	110	CS	110	TD	125
KS170/CS110/KA125	KS	170	No Use	0	No Use	0	CS	110	KA	125

Step 07: Eliminate null values in the 'Finished_Product_Wid', 'Finished_Product_Wid' and 'Finished_Product_Hig' columns

Eliminate some data of Finished_Product_Wid', 'Finished_Product_Wid', and 'Finished_Product_Hig' data were lost during the data collecting process.

For example,

- condition2 = appended_df['Finished_Product_Wid'].isnull() | appended_df['Finished_Product_Leg'].isnull() | appended_df['Finished_Product_Hig'].isnull()
- appended_df = appended_df[~condition2]

Step 08: Display rows with missing values to confirm data replacing

For example,

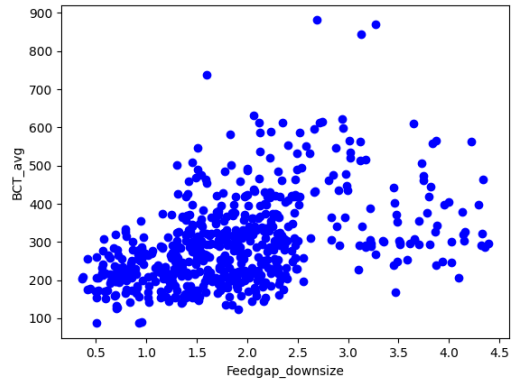
- missing_values = appended_df.isnull().sum()
- rows_with_nulls = appended_df[appended_df.isnull().any(axis=1)] # Row number with null data
- print(missing_values)

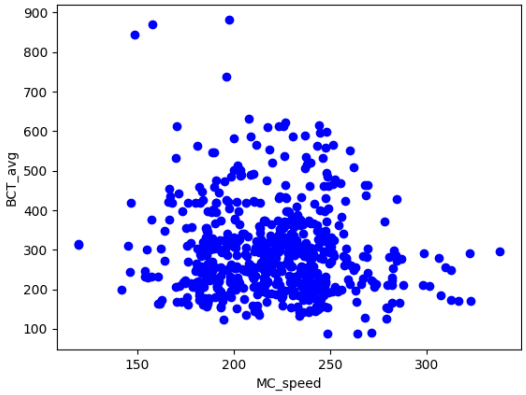
Data Cleaning Coding Result:

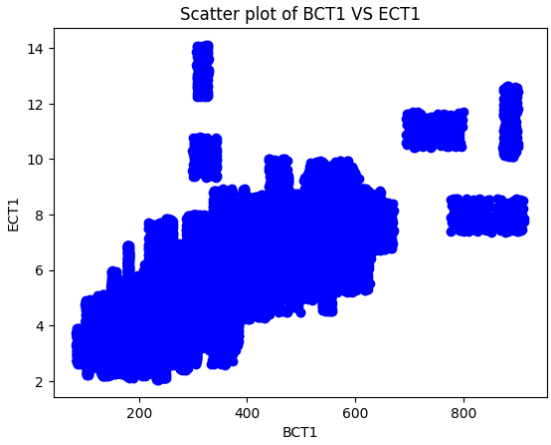
Data records before cleaning: 125,288

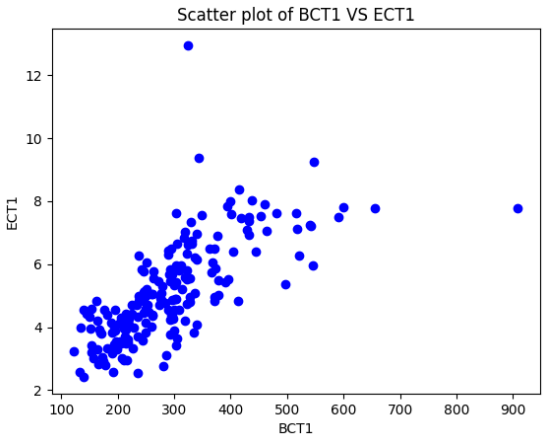
Data records before cleaning: 124,857

Correlation of each variable and analysis:

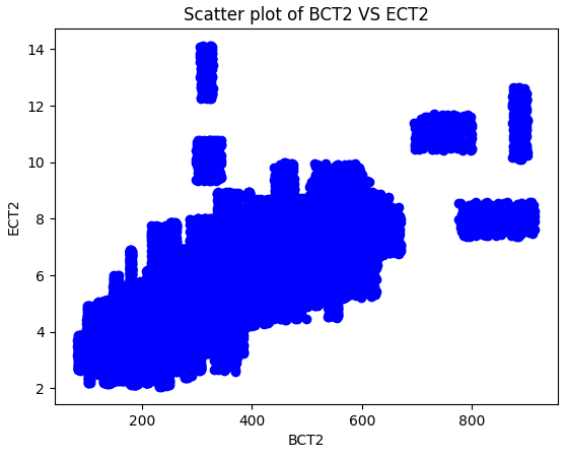


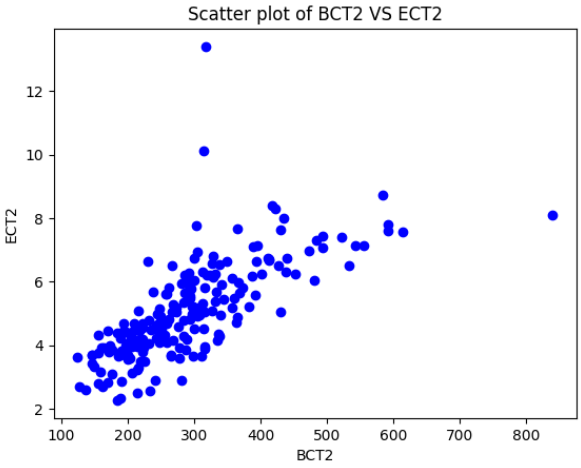


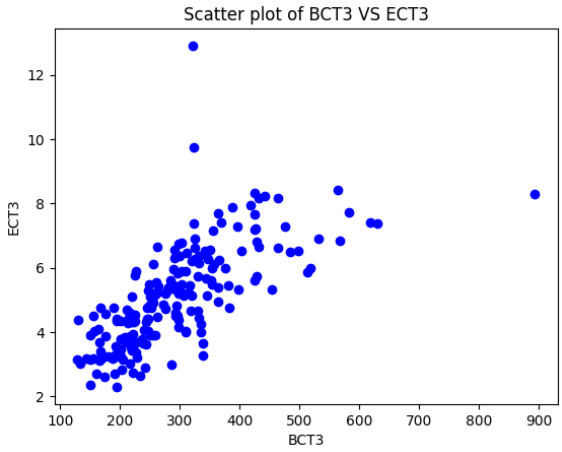


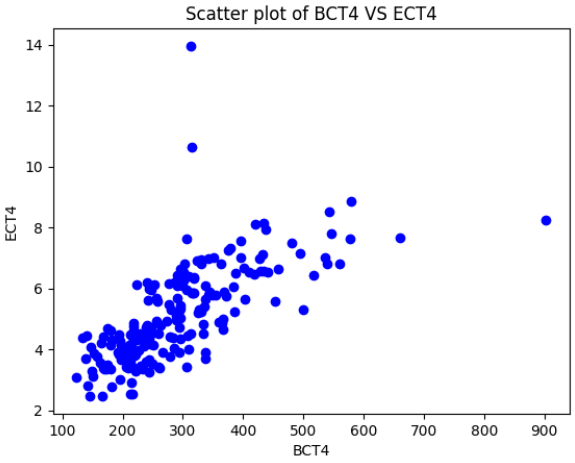


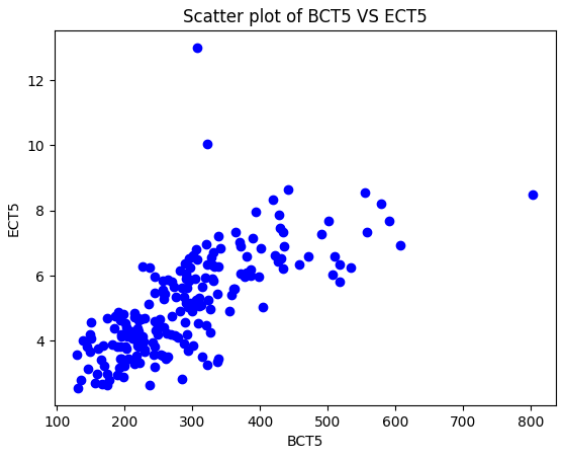
Both the above graphs depict the correlation between BCT and ECT. The graph on the left has much more values, due to which it is very concentrated. The graph on the right is obtained after downsizing the sample space.



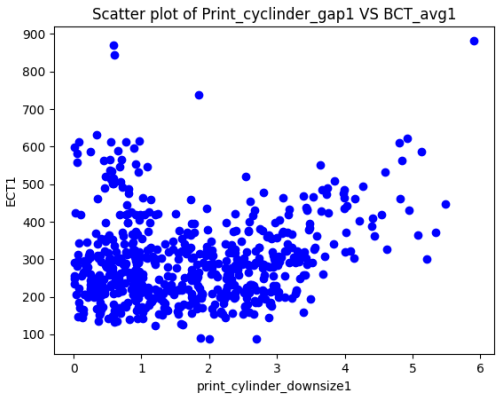


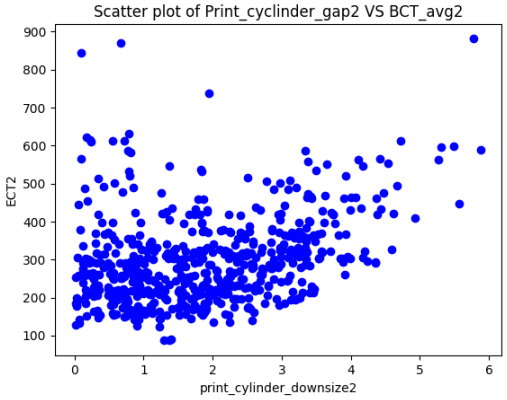


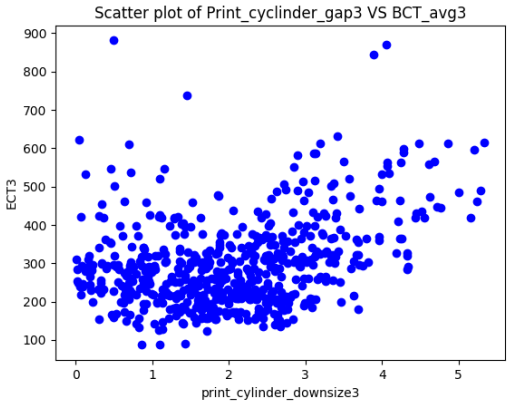


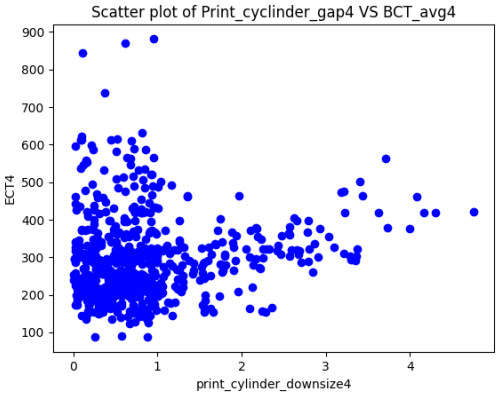


The five scatter plots above depict the relationship between BCT and ECT. The graphs show a positive linear correlation between BCT and ECT. In all the graphs, the values are concentrated in the lower left quadrant, barring a few outliers which fall in the higher value range. The ECT values are mostly spread between 2.5 and 7 and the BCT values are clustered between 100 and 400.





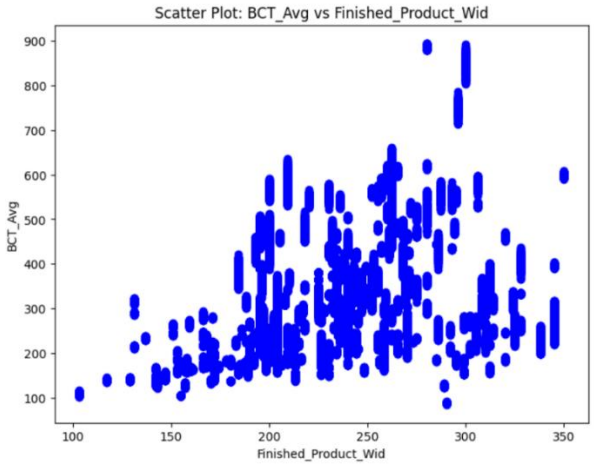




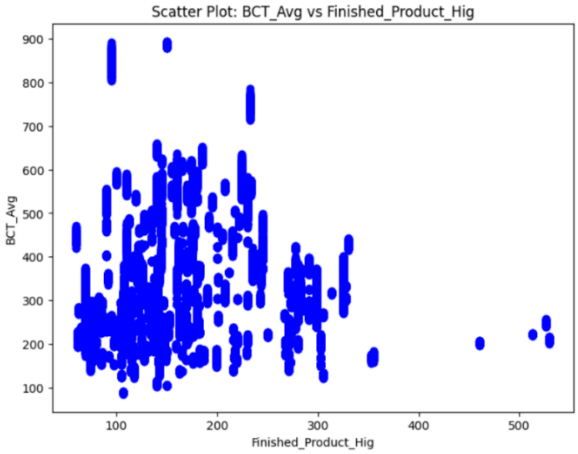
The four scatterplot graphs above depict the relationship between the strength of the raw materials (ETC) used to create packaging and the print cylinder gap. As part of the manufacturing process, the raw materials go through one feed cylinder and four print cylinders to create the final product.

The cluster of blue in the first scatterplot graph indicates that SCGP mostly uses raw materials that range between ~150 to ~400 in strength value. However, there are outliers. The gap size most frequently used in the cluster is between zero and three. This range in gap size remains mostly consistent until the raw materials reach the fourth print cylinder. Using this range in gap size will ensure that a moderate amount of pressure is applied to the raw materials at the middle of the production process.

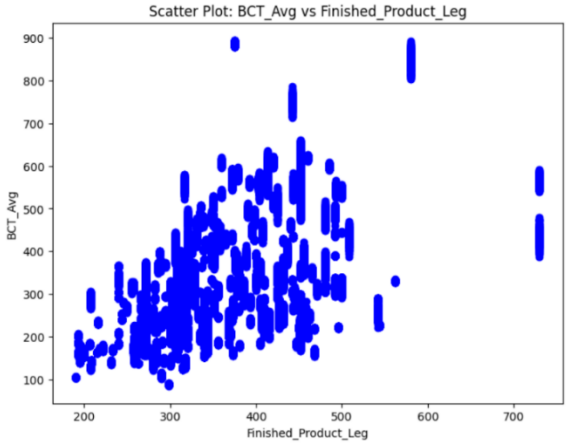
Once the raw materials reach the fourth print cylinder, which is the final stage of production, the gap between the cylinders is at its closest. The fourth scatterplot graph indicates that most of the products, regardless of ETC, have a significant amount of pressure applied to them at this stage. The box should be able to handle moving through the tight gap because of the moderate amount of pressure applied to the raw materials as they moved through the first three print cylinders which used wider gaps. Based on the correlations identified in all four scatterplot graphs, it will help to determine the optimal gap between the print cylinders and decide on the best materials to use during production.



The scatter plot shows a positive trend where high finished product high has a higher bct average. As the width increases from 100 to 300, the plot becomes more scattered. There's a tighter clustering at lower width values which is 100 to 200, and more variation between 200 to 300. There are a few points at the higher end of the bct average, those points stand out from the general distribution.

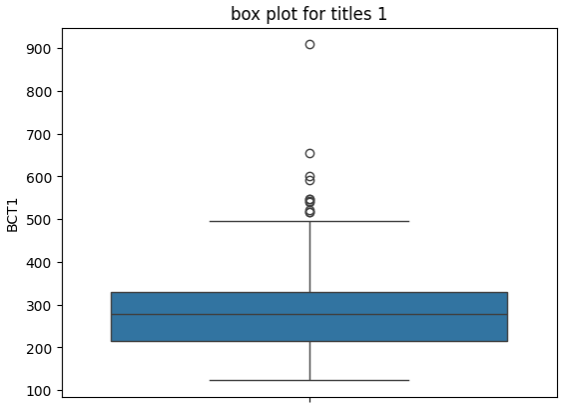


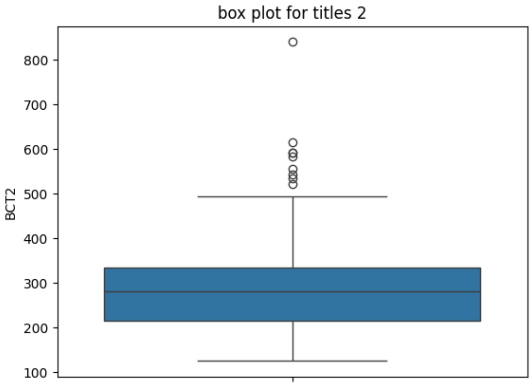
The plot shows a positive correlation. As the high increases, the points spread indicates variability. The points concetrate between the 100-300 range for high and 100-600 range for bct average. There are some significant points beyond 300 high, these outliers don't fit in the general trends.

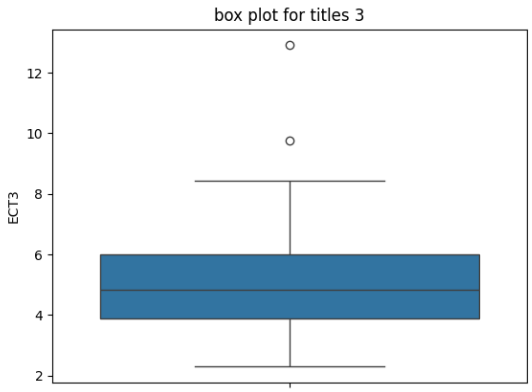


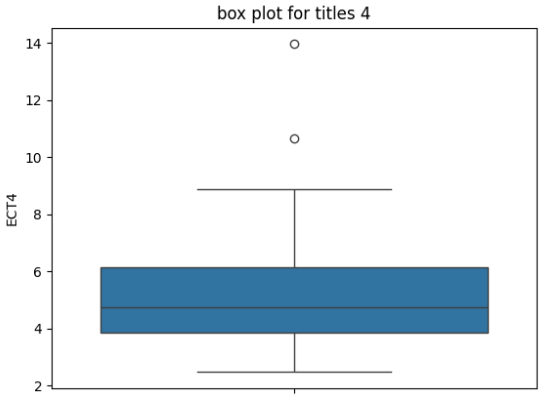
The plot shows a positive correlation where higher length correspond to higher bct average. There is a high concentration of points between 200-500 for length and 200=600 for bct average. There are fewer points left beyond the 500 length. There are a few outliers above the 700 bct average that decviate from the general trend.

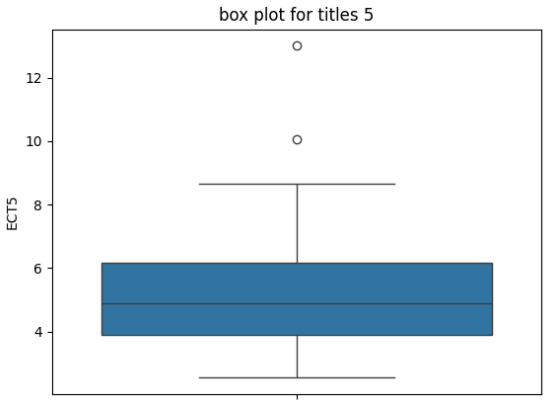
In conclusion, the three variables all show positive correlation with bct average. As any of these variables increase, the bct average increase as well. All variables have varing degrees of spread and outliers. All plots have points concentrated in mid range of the dimension and most bct average lie between 200-600.











Box plot is used to visualize the outliers. For the first two graphs, there are multiple outliers, and the median is around 300. For the next three graphs, there are 2 outliers each and the median for all of them is around 5. ECT4 has more extreme outliers, 14 and 11, compared to ECT3, 13 and 10, and ECT5 which also has outliers at 13 and 10.