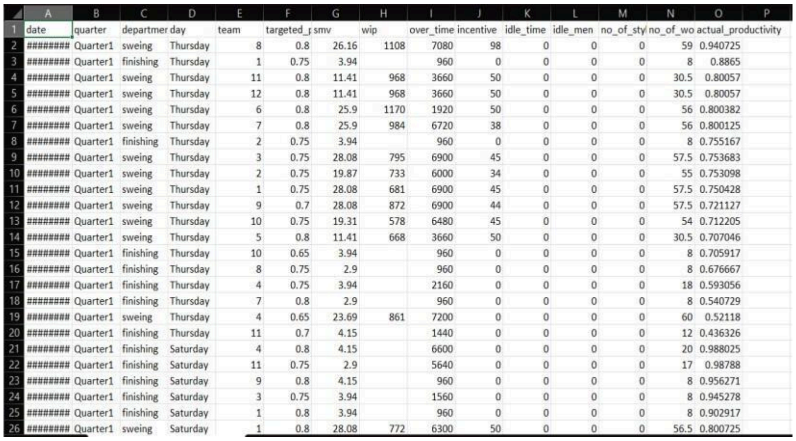
Data Collection and Preprocessing Phase

Date	24 June 2025
Team ID	SWUID20250176345
Project Title	Machine Learning Approach for Employee Performance Prediction
Maximum Marks	6 Marks

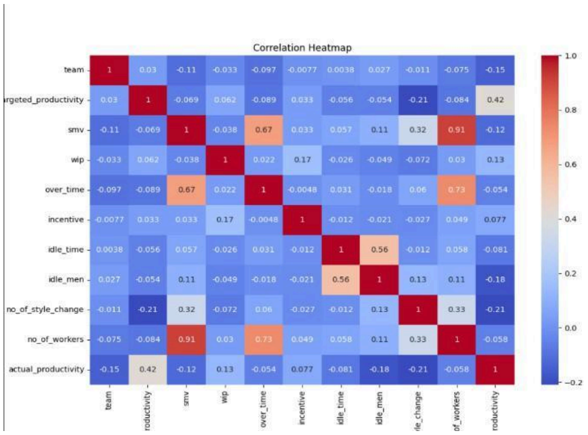
Data Exploration and Preprocessing Report

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>
Dimension:
1,197 rows × 15 columns
Descriptive statistics:

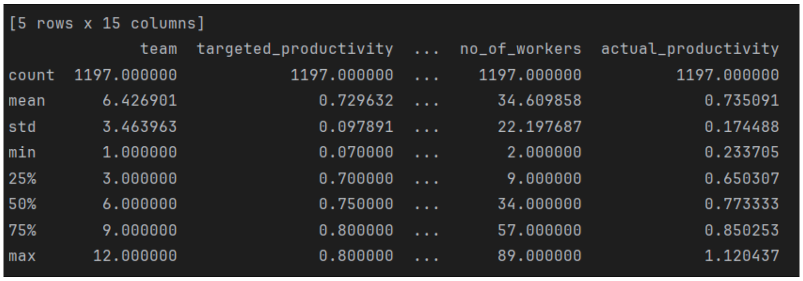


<table>
<tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th><th>G</th><th>H</th><th>I</th><th>J</th><th>K</th><th>L</th><th>M</th><th>N</th><th>O</th><th>P</th></tr>
<tr><td>1</td><td>date</td><td>quarter</td><td>departmer</td><td>day</td><td>team</td><td>targeted_p</td><td>smv</td><td>wip</td><td>over_time</td><td>incentive</td><td>idle_time</td><td>idle_men</td><td>no_of_styl</td><td>no_of_wo</td><td>actual_productivity</td><td></td></tr>
<tr><td>2</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>8</td><td>0.8</td><td>26.16</td><td>1108</td><td>7080</td><td>98</td><td>0</td><td>0</td><td>0</td><td>59</td><td>0.940725</td><td></td></tr>
<tr><td>3</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>1</td><td>0.75</td><td>3.94</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.8865</td><td></td></tr>
<tr><td>4</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>11</td><td>0.8</td><td>11.41</td><td>968</td><td>3660</td><td>50</td><td>0</td><td>0</td><td>0</td><td>30.5</td><td>0.80057</td><td></td></tr>
<tr><td>5</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>12</td><td>0.8</td><td>11.41</td><td>968</td><td>3660</td><td>50</td><td>0</td><td>0</td><td>0</td><td>30.5</td><td>0.80057</td><td></td></tr>
<tr><td>6</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>6</td><td>0.8</td><td>25.9</td><td>1170</td><td>1920</td><td>50</td><td>0</td><td>0</td><td>0</td><td>56</td><td>0.800382</td><td></td></tr>
<tr><td>7</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>7</td><td>0.8</td><td>25.9</td><td>984</td><td>6720</td><td>38</td><td>0</td><td>0</td><td>0</td><td>56</td><td>0.800125</td><td></td></tr>
<tr><td>8</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>2</td><td>0.75</td><td>3.94</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.755167</td><td></td></tr>
<tr><td>9</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>3</td><td>0.75</td><td>28.08</td><td>795</td><td>6900</td><td>45</td><td>0</td><td>0</td><td>0</td><td>57.5</td><td>0.753683</td><td></td></tr>
<tr><td>10</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>2</td><td>0.75</td><td>19.87</td><td>733</td><td>6000</td><td>34</td><td>0</td><td>0</td><td>0</td><td>55</td><td>0.753098</td><td></td></tr>
<tr><td>11</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>1</td><td>0.75</td><td>28.08</td><td>681</td><td>6900</td><td>45</td><td>0</td><td>0</td><td>0</td><td>57.5</td><td>0.750428</td><td></td></tr>
<tr><td>12</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>9</td><td>0.7</td><td>28.08</td><td>872</td><td>6900</td><td>44</td><td>0</td><td>0</td><td>0</td><td>57.5</td><td>0.721127</td><td></td></tr>
<tr><td>13</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>10</td><td>0.75</td><td>19.31</td><td>578</td><td>6480</td><td>45</td><td>0</td><td>0</td><td>0</td><td>54</td><td>0.712205</td><td></td></tr>
<tr><td>14</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>5</td><td>0.8</td><td>11.41</td><td>668</td><td>3660</td><td>50</td><td>0</td><td>0</td><td>0</td><td>30.5</td><td>0.707046</td><td></td></tr>
<tr><td>15</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>10</td><td>0.65</td><td>3.94</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.705917</td><td></td></tr>
<tr><td>16</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>8</td><td>0.75</td><td>2.9</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.676667</td><td></td></tr>
<tr><td>17</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>4</td><td>0.75</td><td>3.94</td><td></td><td>2160</td><td>0</td><td>0</td><td>0</td><td>0</td><td>18</td><td>0.593056</td><td></td></tr>
<tr><td>18</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>7</td><td>0.8</td><td>2.9</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.540729</td><td></td></tr>
<tr><td>19</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Thursday</td><td>4</td><td>0.65</td><td>23.69</td><td>861</td><td>7200</td><td>0</td><td>0</td><td>0</td><td>0</td><td>60</td><td>0.52118</td><td></td></tr>
<tr><td>20</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Thursday</td><td>11</td><td>0.7</td><td>4.15</td><td></td><td>1440</td><td>0</td><td>0</td><td>0</td><td>0</td><td>12</td><td>0.436326</td><td></td></tr>
<tr><td>21</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Saturday</td><td>4</td><td>0.8</td><td>4.15</td><td></td><td>6600</td><td>0</td><td>0</td><td>0</td><td>0</td><td>20</td><td>0.988025</td><td></td></tr>
<tr><td>22</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Saturday</td><td>11</td><td>0.75</td><td>2.9</td><td></td><td>5640</td><td>0</td><td>0</td><td>0</td><td>0</td><td>17</td><td>0.98788</td><td></td></tr>
<tr><td>23</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Saturday</td><td>9</td><td>0.8</td><td>4.15</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.956271</td><td></td></tr>
<tr><td>24</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Saturday</td><td>3</td><td>0.75</td><td>3.94</td><td></td><td>1560</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.945278</td><td></td></tr>
<tr><td>25</td><td>########</td><td>Quarter1</td><td>finishing</td><td>Saturday</td><td>1</td><td>0.8</td><td>3.94</td><td></td><td>960</td><td>0</td><td>0</td><td>0</td><td>0</td><td>8</td><td>0.902917</td><td></td></tr>
<tr><td>26</td><td>########</td><td>Quarter1</td><td>sweing</td><td>Saturday</td><td>1</td><td>0.8</td><td>28.08</td><td>772</td><td>6300</td><td>50</td><td>0</td><td>0</td><td>0</td><td>56.5</td><td>0.800725</td><td></td></tr>
</table>
</td></tr>
<tr><td>Correlation Analysis</td><td></td></tr>
</table>

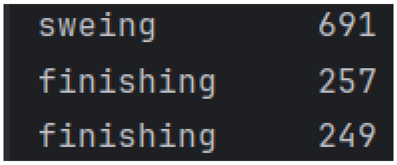


	team	roductivity	smv	wip	over_time	incentive	idle_time	idle_men	le_change	of_workers	roductivity
team	1	0.03	-0.11	-0.033	-0.097	-0.0077	0.0038	0.027	-0.011	-0.075	-0.15
rgeted_productivity	0.03	1	-0.069	0.062	-0.089	0.033	-0.056	-0.054	-0.21	-0.084	0.42
smv	-0.11	-0.069	1	-0.038	0.67	0.033	0.057	0.11	0.32	0.91	-0.12
wip	-0.033	0.062	-0.038	1	0.022	0.17	-0.026	-0.049	-0.072	0.03	0.13
over_time	-0.097	-0.089	0.67	0.022	1	-0.0048	0.031	-0.018	0.06	0.73	-0.054
incentive	-0.0077	0.033	0.033	0.17	-0.0048	1	-0.012	-0.021	-0.027	0.049	0.077
idle_time	0.0038	-0.056	0.057	-0.026	0.031	-0.012	1	0.56	-0.012	0.058	-0.081
idle_men	0.027	-0.054	0.11	-0.049	-0.018	-0.021	0.56	1	0.13	0.11	-0.18
no_of_style_change	-0.011	-0.21	0.32	-0.072	0.06	-0.027	-0.012	0.13	1	0.33	-0.21
no_of_workers	-0.075	-0.084	0.91	0.03	0.73	0.049	0.058	0.11	0.33	1	-0.058
actual_productivity	-0.15	0.42	-0.12	0.13	-0.054	0.077	-0.081	-0.18	-0.21	-0.058	1

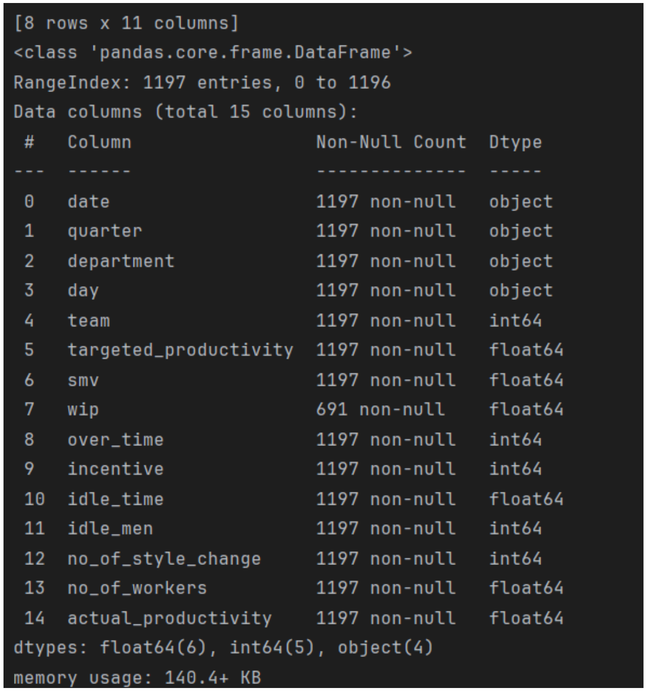
Descriptive Analysis



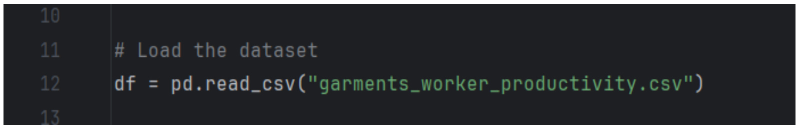
```
[5 rows x 15 columns]
             team  targeted_productivity  ...  no_of_workers  actual_productivity
count  1197.000000            1197.000000  ...    1197.000000          1197.000000
mean      6.426901               0.729632  ...      34.609858             0.735091
std       3.463963               0.097891  ...      22.197687             0.174488
min       1.000000               0.070000  ...       2.000000             0.233705
25%       3.000000               0.700000  ...       9.000000             0.650307
50%       6.000000               0.750000  ...      34.000000             0.773333
75%       9.000000               0.800000  ...      57.000000             0.850253
max      12.000000               0.800000  ...      89.000000             1.120437
```



```
 sweing        691
 finishing     257
 finishing     249
```



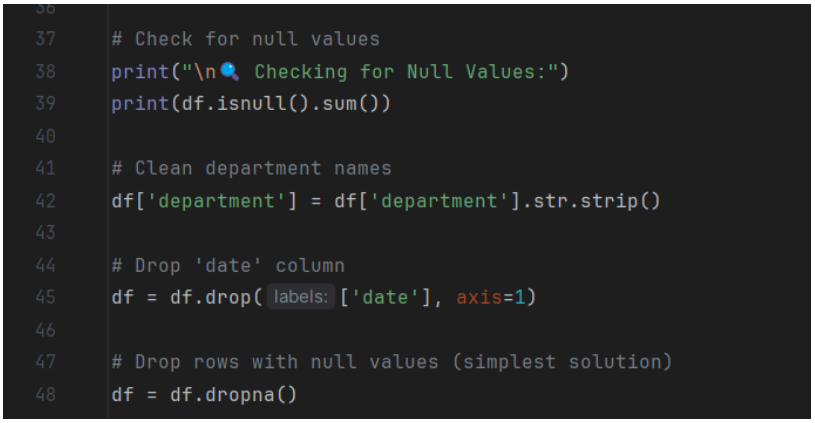
```
[8 rows x 11 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1197 entries, 0 to 1196
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   date                   1197 non-null   object 
 1   quarter                1197 non-null   object 
 2   department             1197 non-null   object 
 3   day                    1197 non-null   object 
 4   team                   1197 non-null   int64  
 5   targeted_productivity  1197 non-null   float64
 6   smv                    1197 non-null   float64
 7   wip                    691 non-null    float64
 8   over_time              1197 non-null   int64  
 9   incentive              1197 non-null   int64  
 10  idle_time              1197 non-null   float64
 11  idle_men               1197 non-null   int64  
 12  no_of_style_change     1197 non-null   int64  
 13  no_of_workers          1197 non-null   float64
 14  actual_productivity    1197 non-null   float64
dtypes: float64(6), int64(5), object(4)
memory usage: 140.4+ KB
```

Outliers and Anomalies	-
Data Preprocessing Code Screenshots	
Loading Data	

```

# Load the dataset
df = pd.read_csv("garments_worker_productivity.csv")

```

Handling Missing Data	
	# Check for null values
	print("\n🔍 Checking for Null Values:")
	print(df.isnull().sum())
	
	# Clean department names
	df['department'] = df['department'].str.strip()
	
	# Drop 'date' column
	df = df.drop(labels: ['date'], axis=1)
	
	# Drop rows with null values (simplest solution)
	df = df.dropna()

Save Processed Data	- Task Completed