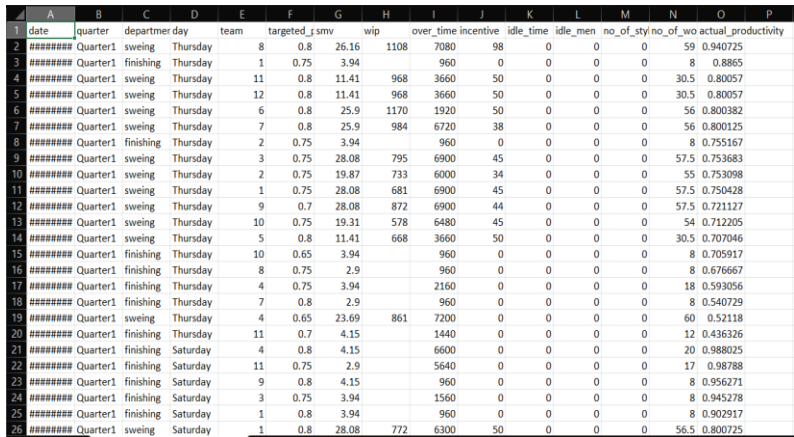


## Data Collection and Preprocessing Phase

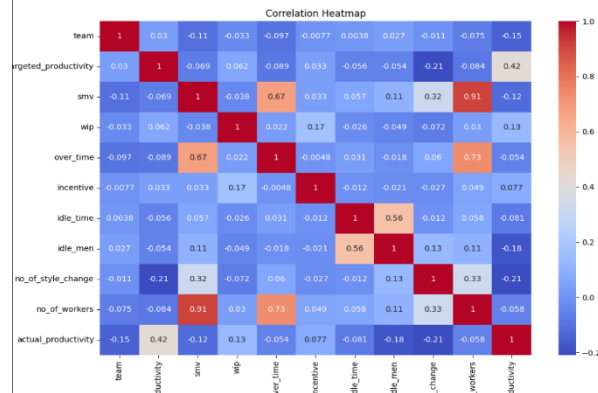
Date	24 June 2025
Team ID	SWUID20250176341
Project Title	Machine Learning Approach for Employee Performance Prediction
Maximum Marks	6 Marks

### 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.

Section	Description
Data Overview	<p><u>Dimension:</u> 1,197 rows × 15 columns.</p> <p><u>Descriptive statistics:</u></p> 
Correlation Analysis	

## 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	-
<b>Data Preprocessing Code Screenshots</b>	
Loading Data	<pre># Reading CSv file df = pd.read_csv(r"D:\Work\garments_worker_productivity.csv") print(df.head())</pre>
Handling Missing Data	<pre># Checking for Null Values print(df.isnull().sum())  # Handling Date &amp; Department Columns df['department'] = df['department'].str.strip() df = df.drop(labels: ['date'], axis=1)  # Handling Categorical Values df = pd.get_dummies(df, drop_first=True)  # Drop rows with any NaN values df.dropna(inplace=True)</pre>
Save Processed Data	- Task Completed