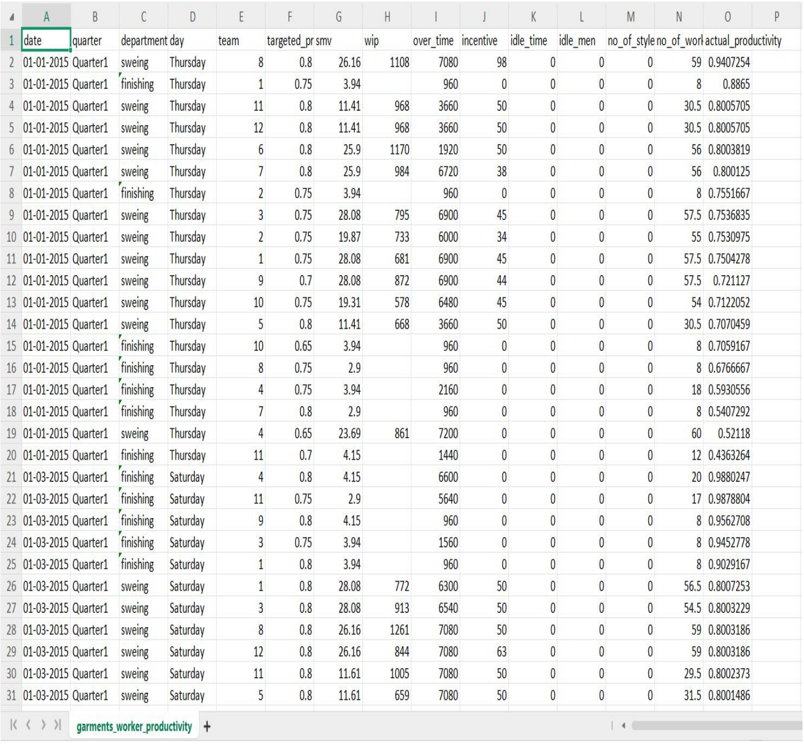
# Data Collection and Preprocessing Phase

| | |
|---|---|
| Date | 24 June 2025 |
| Team ID | SWUID20250176209 |
| Project Title | Machine Learning Approach for Employee Performance Prediction |
| Maximum Marks | 6 Marks |

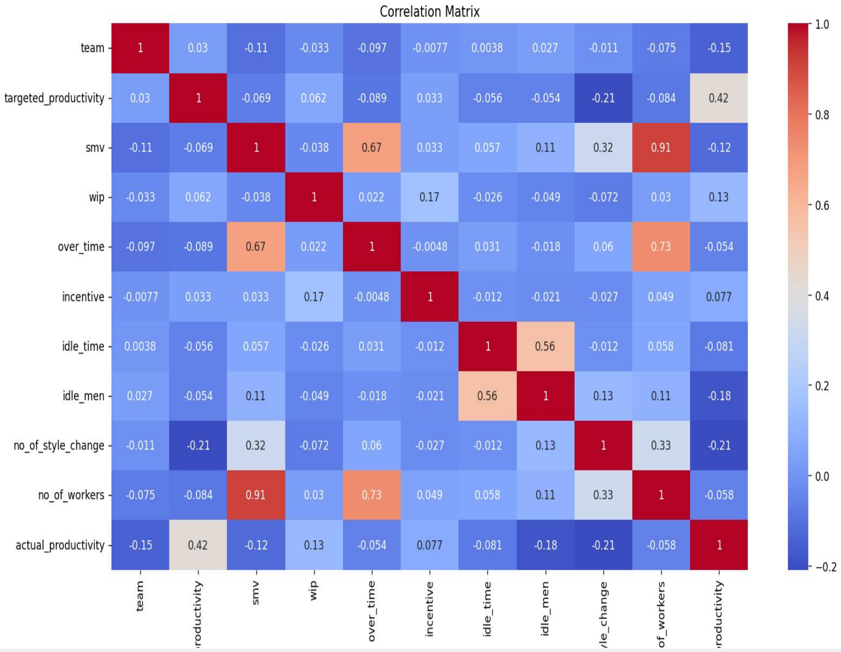
## Data Exploration and Preprocessing Report

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.

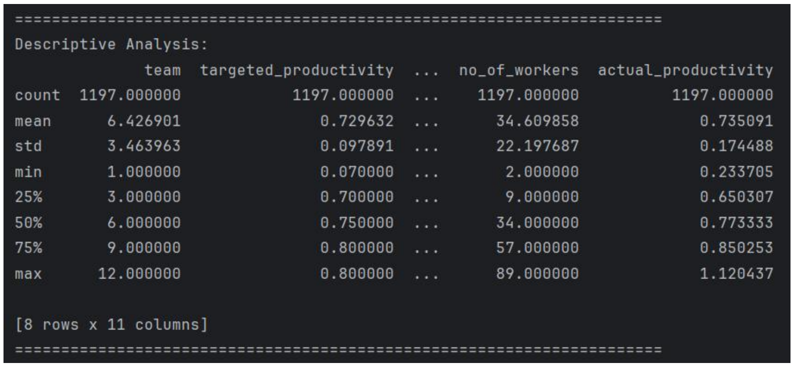
| Section | Description |
|---|---|
| Data Overview | Dimension:<br>1,197 rows × 15 columns<br>Descriptive statistics:<br> |

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | date | quarter | department | day | team | targeted_pr | smv | wip | over_time | incentive | idle_time | idle_men | no_of_style | no_of_worl | actual_productivity | |
| 2 | 01-01-2015 | Quarter1 | sweing | Thursday | 8 | 0.8 | 26.16 | 1108 | 7080 | 98 | 0 | 0 | 0 | 59 | 0.9407254 | |
| 3 | 01-01-2015 | Quarter1 | finishing | Thursday | 1 | 0.75 | 3.94 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.8865 | |
| 4 | 01-01-2015 | Quarter1 | sweing | Thursday | 11 | 0.8 | 11.41 | 968 | 3660 | 50 | 0 | 0 | 0 | 30.5 | 0.8005705 | |
| 5 | 01-01-2015 | Quarter1 | sweing | Thursday | 12 | 0.8 | 11.41 | 968 | 3660 | 50 | 0 | 0 | 0 | 30.5 | 0.8005705 | |
| 6 | 01-01-2015 | Quarter1 | sweing | Thursday | 6 | 0.8 | 25.9 | 1170 | 1920 | 50 | 0 | 0 | 0 | 56 | 0.8003819 | |
| 7 | 01-01-2015 | Quarter1 | sweing | Thursday | 7 | 0.8 | 25.9 | 984 | 6720 | 38 | 0 | 0 | 0 | 56 | 0.800125 | |
| 8 | 01-01-2015 | Quarter1 | finishing | Thursday | 2 | 0.75 | 3.94 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.7551667 | |
| 9 | 01-01-2015 | Quarter1 | sweing | Thursday | 3 | 0.75 | 28.08 | 795 | 6900 | 45 | 0 | 0 | 0 | 57.5 | 0.7536835 | |
| 10 | 01-01-2015 | Quarter1 | sweing | Thursday | 2 | 0.75 | 19.87 | 733 | 6000 | 34 | 0 | 0 | 0 | 55 | 0.7530975 | |
| 11 | 01-01-2015 | Quarter1 | sweing | Thursday | 1 | 0.75 | 28.08 | 681 | 6900 | 45 | 0 | 0 | 0 | 57.5 | 0.7504278 | |
| 12 | 01-01-2015 | Quarter1 | sweing | Thursday | 9 | 0.7 | 28.08 | 872 | 6900 | 44 | 0 | 0 | 0 | 57.5 | 0.721127 | |
| 13 | 01-01-2015 | Quarter1 | sweing | Thursday | 10 | 0.75 | 19.31 | 578 | 6480 | 45 | 0 | 0 | 0 | 54 | 0.7122052 | |
| 14 | 01-01-2015 | Quarter1 | sweing | Thursday | 5 | 0.8 | 11.41 | 668 | 3660 | 50 | 0 | 0 | 0 | 30.5 | 0.7070459 | |
| 15 | 01-01-2015 | Quarter1 | finishing | Thursday | 10 | 0.65 | 3.94 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.7059167 | |
| 16 | 01-01-2015 | Quarter1 | finishing | Thursday | 8 | 0.75 | 2.9 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.6766667 | |
| 17 | 01-01-2015 | Quarter1 | finishing | Thursday | 4 | 0.75 | 3.94 | | 2160 | 0 | 0 | 0 | 0 | 18 | 0.5930556 | |
| 18 | 01-01-2015 | Quarter1 | finishing | Thursday | 7 | 0.8 | 2.9 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.5407292 | |
| 19 | 01-01-2015 | Quarter1 | sweing | Thursday | 4 | 0.65 | 23.69 | 861 | 7200 | 0 | 0 | 0 | 0 | 60 | 0.52118 | |
| 20 | 01-01-2015 | Quarter1 | finishing | Thursday | 11 | 0.7 | 4.15 | | 1440 | 0 | 0 | 0 | 0 | 12 | 0.4363264 | |
| 21 | 01-03-2015 | Quarter1 | finishing | Saturday | 4 | 0.8 | 4.15 | | 6600 | 0 | 0 | 0 | 0 | 20 | 0.9880247 | |
| 22 | 01-03-2015 | Quarter1 | finishing | Saturday | 11 | 0.75 | 2.9 | | 5640 | 0 | 0 | 0 | 0 | 17 | 0.9878804 | |
| 23 | 01-03-2015 | Quarter1 | finishing | Saturday | 9 | 0.8 | 4.15 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.9562708 | |
| 24 | 01-03-2015 | Quarter1 | finishing | Saturday | 3 | 0.75 | 3.94 | | 1560 | 0 | 0 | 0 | 0 | 8 | 0.9452778 | |
| 25 | 01-03-2015 | Quarter1 | finishing | Saturday | 1 | 0.8 | 3.94 | | 960 | 0 | 0 | 0 | 0 | 8 | 0.9029167 | |
| 26 | 01-03-2015 | Quarter1 | sweing | Saturday | 1 | 0.8 | 28.08 | 772 | 6300 | 50 | 0 | 0 | 0 | 56.5 | 0.8007253 | |
| 27 | 01-03-2015 | Quarter1 | sweing | Saturday | 3 | 0.8 | 28.08 | 913 | 6540 | 50 | 0 | 0 | 0 | 54.5 | 0.8003229 | |
| 28 | 01-03-2015 | Quarter1 | sweing | Saturday | 8 | 0.8 | 26.16 | 1261 | 7080 | 50 | 0 | 0 | 0 | 59 | 0.8003186 | |
| 29 | 01-03-2015 | Quarter1 | sweing | Saturday | 12 | 0.8 | 26.16 | 844 | 7080 | 63 | 0 | 0 | 0 | 59 | 0.8003186 | |
| 30 | 01-03-2015 | Quarter1 | sweing | Saturday | 11 | 0.8 | 11.61 | 1005 | 7080 | 50 | 0 | 0 | 0 | 29.5 | 0.8002373 | |
| 31 | 01-03-2015 | Quarter1 | sweing | Saturday | 5 | 0.8 | 11.61 | 659 | 7080 | 50 | 0 | 0 | 0 | 31.5 | 0.8001486 | |

Correlation Analysis

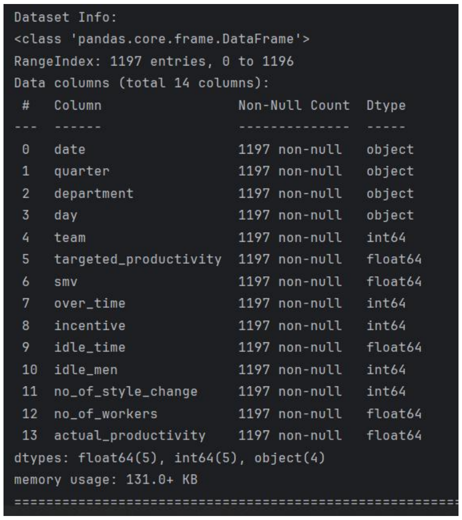


Descriptive Analysis

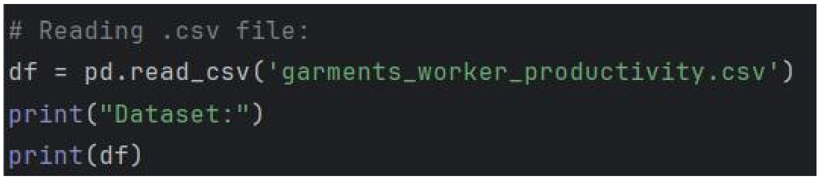
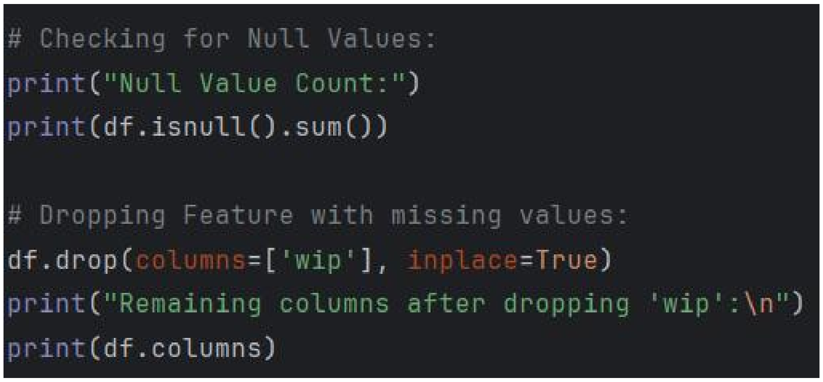


```
======================================================================
Descriptive Analysis:
              team  targeted_productivity  ...  no_of_workers  actual_productivity
count  1197.000000            1197.000000  ...    1197.000000          1197.000000
mean      6.426901               0.729632  ...      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

[8 rows x 11 columns]
======================================================================
```



```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1197 entries, 0 to 1196
Data columns (total 14 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   over_time              1197 non-null   int64
 8   incentive              1197 non-null   int64
 9   idle_time              1197 non-null   float64
 10  idle_men               1197 non-null   int64
 11  no_of_style_change     1197 non-null   int64
 12  no_of_workers          1197 non-null   float64
 13  actual_productivity    1197 non-null   float64
dtypes: float64(5), int64(5), object(4)
memory usage: 131.0+ KB
========================================================
```

| | |
|---|---|
| Outliers and Anomalies | - |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | <br>`# Reading .csv file:`<br>`df = pd.read_csv('garments_worker_productivity.csv')`<br>`print("Dataset:")`<br>`print(df)` |
| Handling Missing Data | <br>`# Checking for Null Values:`<br>`print("Null Value Count:")`<br>`print(df.isnull().sum())`<br><br>`# Dropping Feature with missing values:`<br>`df.drop(columns=['wip'], inplace=True)`<br>`print("Remaining columns after dropping 'wip':\n")`<br>`print(df.columns)` |
| Save Processed Data | - Task Completed |