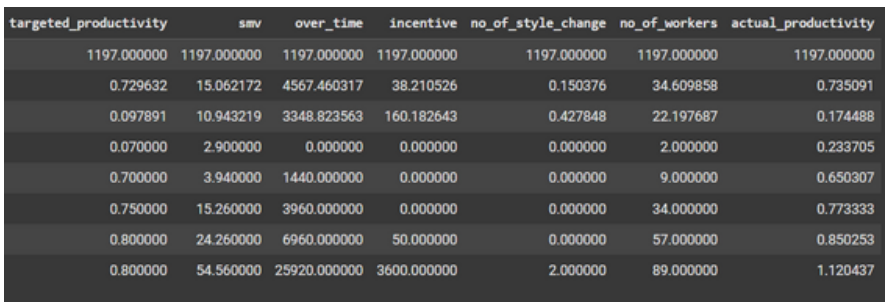
Data Collection and Preprocessing Phase

Date	19 July 2025
Project Title	Machine Learning Approach for Employee Performance Prediction
Maximum 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: 614 rows × 13 columns Descriptive statistics: 

targeted_productivity	smv	over_time	incentive	no_of_style_change	no_of_workers	actual_productivity
1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000	1197.000000
0.729632	15.062172	4567.460317	38.210526	0.150376	34.609858	0.735091
0.097891	10.943219	3348.823563	160.182643	0.427848	22.197687	0.174488
0.070000	2.900000	0.000000	0.000000	0.000000	2.000000	0.233705
0.700000	3.940000	1440.000000	0.000000	0.000000	9.000000	0.650307
0.750000	15.260000	3960.000000	0.000000	0.000000	34.000000	0.773333
0.800000	24.260000	6960.000000	50.000000	0.000000	57.000000	0.850253
0.800000	54.560000	25920.000000	3600.000000	2.000000	89.000000	1.120437

Data Preprocessing Code Screenshots

<table>
<tr><td>Loading Data</td><td>

```
df = pd.read_csv("garments_worker_productivity.csv")
df
```

	date	quarter	department	day	team	targeted_productivity	smv	wip	over_time
0	1/1/2015	Quarter1	sweing	Thursday	8	0.80	26.16	1108.0	7080
1	1/1/2015	Quarter1	finishing	Thursday	1	0.75	3.94	NaN	960
2	1/1/2015	Quarter1	sweing	Thursday	11	0.80	11.41	968.0	3660
3	1/1/2015	Quarter1	sweing	Thursday	12	0.80	11.41	968.0	3660
4	1/1/2015	Quarter1	sweing	Thursday	6	0.80	25.90	1170.0	1920

</td></tr>
<tr><td>Handling Missing Value</td><td>

```
df.isnull().sum()
```

	0
date	0
quarter	0
department	0
day	0
team	0
targeted_productivity	0
smv	0
wip	506

```
df.drop('wip',axis=1, inplace=True)
```

</td></tr>
<tr><td>Feature Engineering</td><td>

```
# handling date column
df['date'] = pd.to_datetime(df['date'])
df['month'] = df['date'].dt.month
df.drop('date',axis=1,inplace=True)

#handling department column
category_mapping = {
    'sweing':'sewing',
    'finishing':'finishing'
}
df['department'] = df['department'].str.strip().map(category_mapping)
```

department	count
sewing	691
finishing	506

</td></tr>
<tr><td>Label Encoding</td><td>

```
3. Label Encoding for categorical columns

le = MultiColumnLabelEncoder.MultiColumnLabelEncoder()
df = le.fit_transform(df)
```

</td></tr>
<tr><td>Save Processed Data</td><td>-</td></tr>
</table>

Correlaton Analysis