# Data Collection and Preprocessing Phase

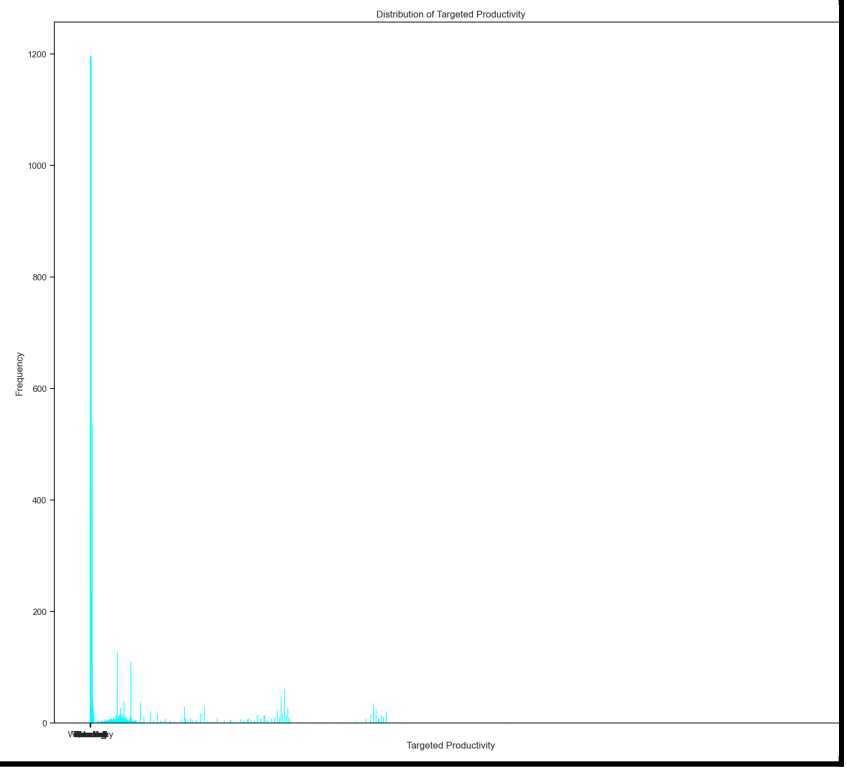
| | |
|---|---|
| Date | 11 July 2024 |
| Team ID | SWTID1720178802 |
| Project Title | Garment worker productivity prediction |
| Maximum Marks | 6 Marks |

## Data Exploration and Preprocessing Template

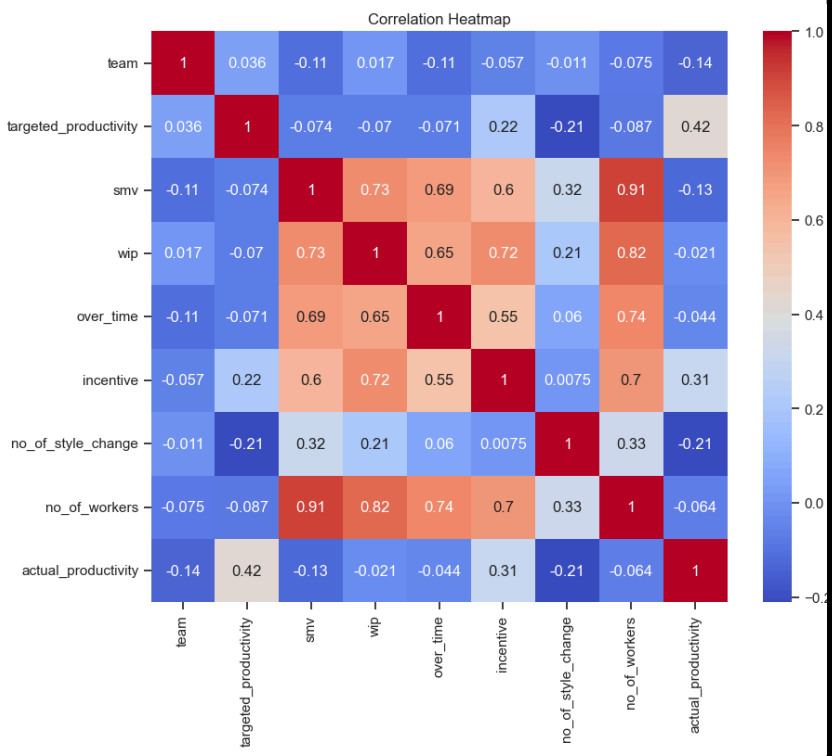
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

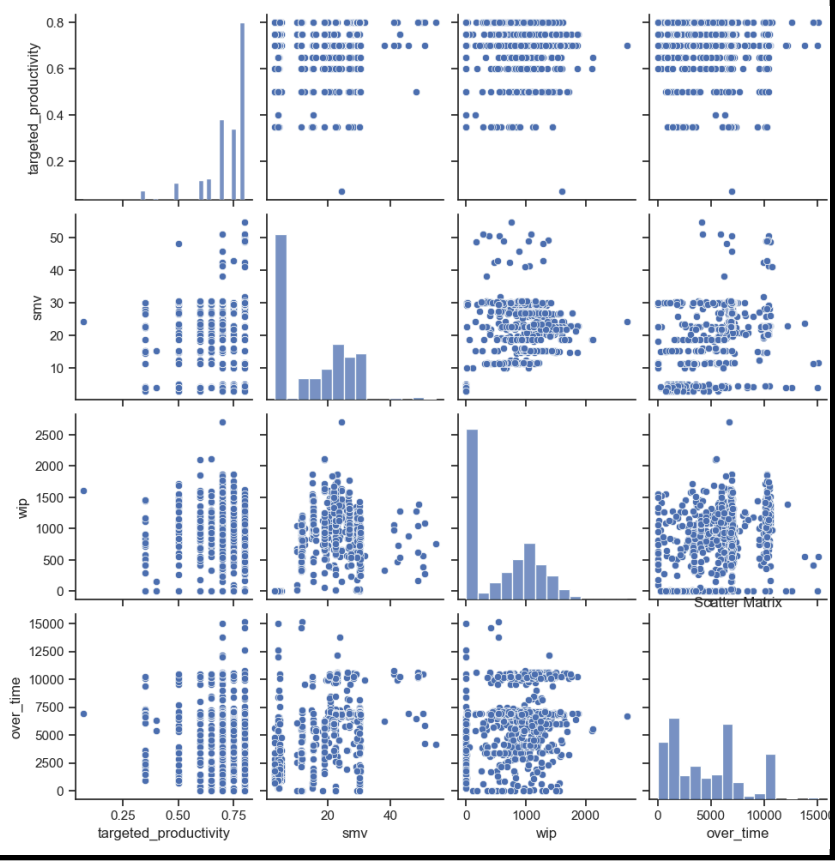
| Section | Description |
|---|---|
| Data Overview | |
| Univariate Analysis | |

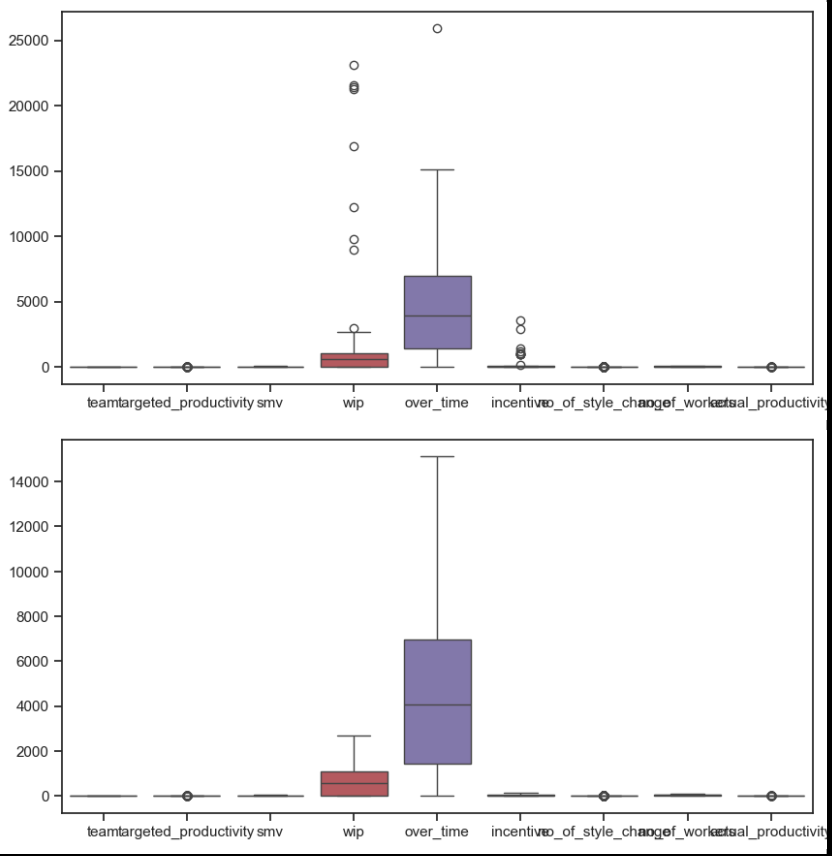
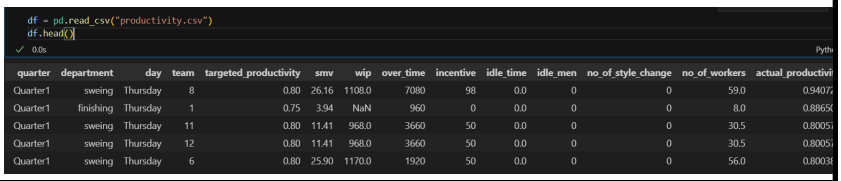
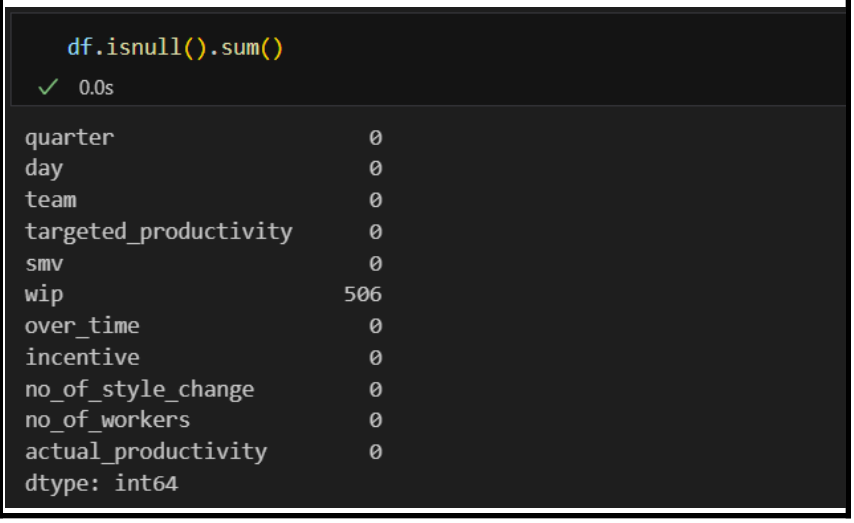
| | team | targeted_productivity | smv | wip | over_time | incentive | no_of_style_change | no_of_workers | actual_productivit |
|---|---|---|---|---|---|---|---|---|---|
| count | 1197.000000 | 1197.000000 | 1197.000000 | 1197.000000 | 1197.000000 | 1197.000000 | 1197.000000 | 1197.000000 | 1197.00000 |
| mean | 6.426901 | 0.729632 | 15.062172 | 687.228070 | 4567.460317 | 38.210526 | 0.150376 | 34.609858 | 0.73509 |
| std | 3.463963 | 0.097891 | 10.943219 | 1514.582341 | 3348.823563 | 160.182643 | 0.427848 | 22.197687 | 0.17448 |
| min | 1.000000 | 0.070000 | 2.900000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 2.000000 | 0.23370 |
| 25% | 3.000000 | 0.700000 | 3.940000 | 0.000000 | 1440.000000 | 0.000000 | 0.000000 | 9.000000 | 0.65030 |
| 50% | 6.000000 | 0.750000 | 15.260000 | 586.000000 | 3960.000000 | 0.000000 | 0.000000 | 34.000000 | 0.77333 |
| 75% | 9.000000 | 0.800000 | 24.260000 | 1083.000000 | 6960.000000 | 50.000000 | 0.000000 | 57.000000 | 0.85025 |
| max | 12.000000 | 0.800000 | 54.560000 | 23122.000000 | 25920.000000 | 3600.000000 | 2.000000 | 89.000000 | 1.12043 |



Bivariate Analysis



| | |
|---|---|
| Outliers and Anomalies |   |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |   |
| Handling Missing Data |   |

```
df['wip'].fillna(0,inplace=True)
df.isnull().sum()
```

✓ 0.0s

```
C:\Users\hp\AppData\Local\Temp\ipykernel_17788\818197784.py:1
The behavior will change in pandas 3.0. This inplace method w

For example, when doing 'df[col].method(value, inplace=True)'

  df['wip'].fillna(0,inplace=True)
```

```
quarter                  0
day                      0
team                     0
targeted_productivity    0
smv                      0
wip                      0
over_time                0
incentive                0
no_of_style_change       0
no_of_workers            0
actual_productivity      0
dtype: int64
```

Data Transformation

```
df_encoded = pd.get_dummies(df, columns=["quarter", "day"])
features= ['targeted_productivity', 'smv', 'wip', 'over_time', 'incentive', 'no_of_style_change', 'no_of_workers', 'actual_productivity']
scaler =  MinMaxScaler()
df_encoded[features] = scaler.fit_transform(df_encoded[features])
print(df_encoded.head())
```

✓ 0.0s

```
   team  targeted_productivity       smv       wip  over_time  incentive  \
0     8               1.000000  0.450252  0.410675   0.468254   0.823529
1     1               0.931507  0.020132  0.000000   0.063492   0.000000
2    11               1.000000  0.164731  0.358784   0.242063   0.420168
3    12               1.000000  0.164731  0.358784   0.242063   0.420168
4     6               1.000000  0.445219  0.433655   0.126984   0.420168

   no_of_style_change  no_of_workers  actual_productivity  quarter_Quarter1  \
0                 0.0       0.655172             0.797332              True
1                 0.0       0.068966             0.736180              True
2                 0.0       0.327586             0.639274              True
3                 0.0       0.327586             0.639274              True
4                 0.0       0.620690             0.639062              True

   quarter_Quarter2  quarter_Quarter3  quarter_Quarter4  quarter_Quarter5  \
0             False             False             False             False
1             False             False             False             False
2             False             False             False             False
3             False             False             False             False
4             False             False             False             False

   day_Monday  day_Saturday  day_Sunday  day_Thursday  day_Tuesday  \
0       False         False       False          True        False
1       False         False       False          True        False
2       False         False       False          True        False
...
1         False
2         False
3         False
```

| | |
|---|---|
| Save Processed Data | `df_encoded.to_csv('processed_productivity.csv', index=False)`<br>✓ 0.0s |