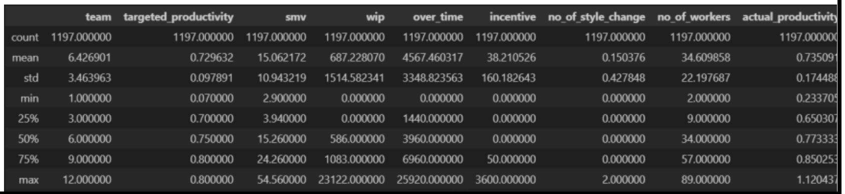
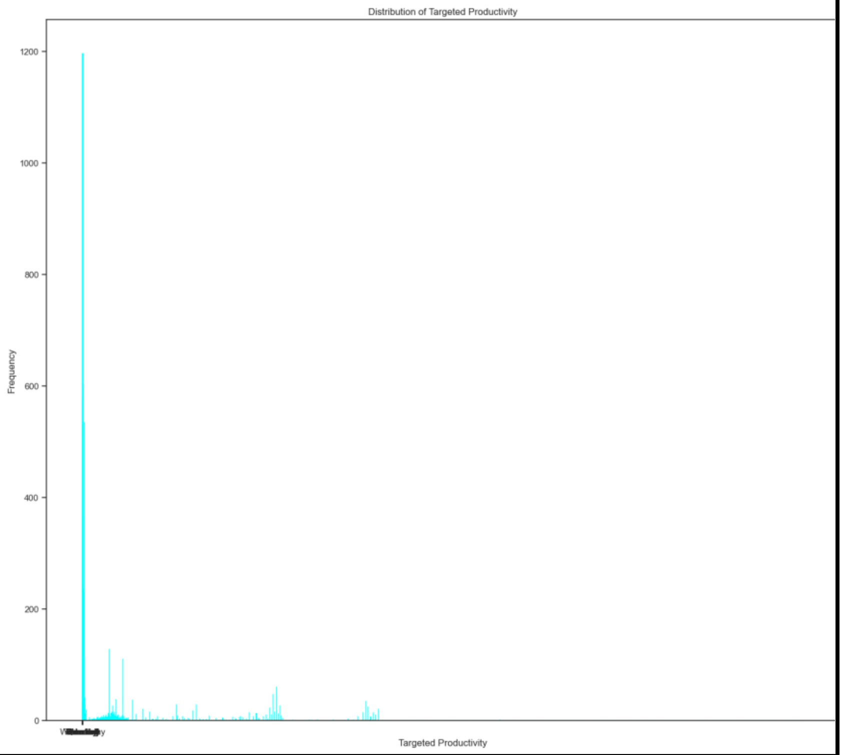
Data Collection and Preprocessing Phase

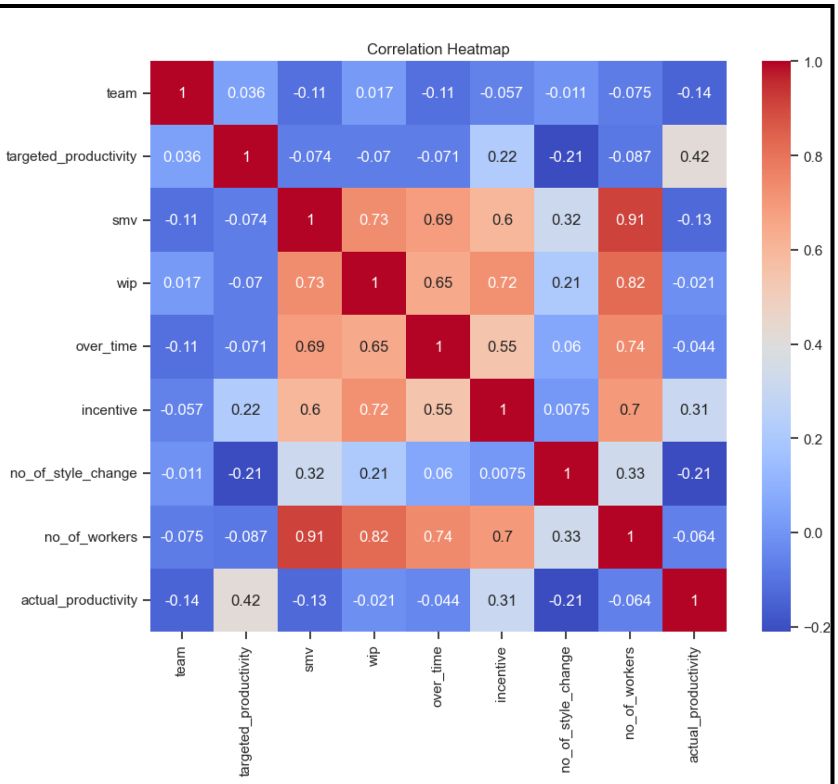
Date	15 March 2024
Team ID	xxxxxx
Project Title	xxxxxx
Maximum Marks	6 Marks

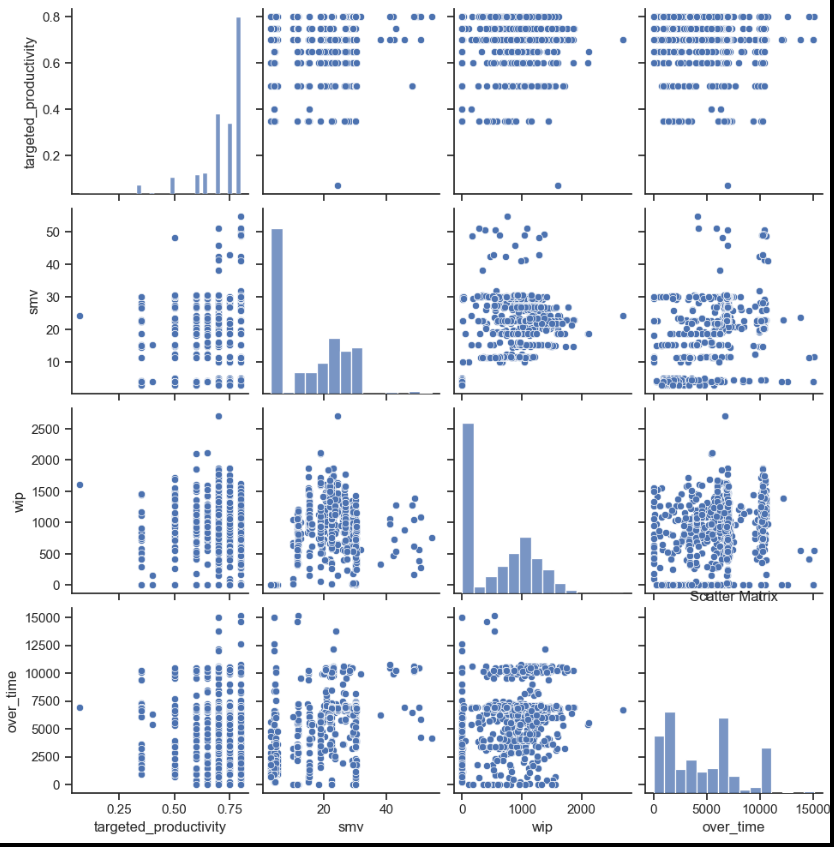
Data Exploration and Preprocessing Template

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

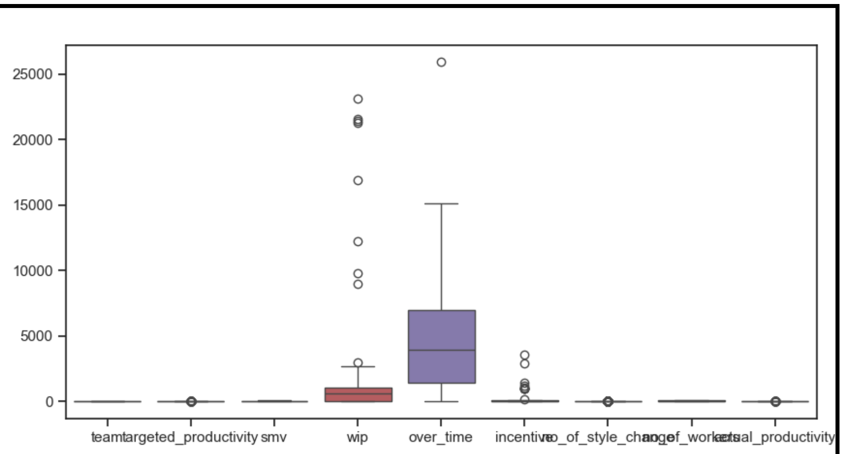
<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>

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


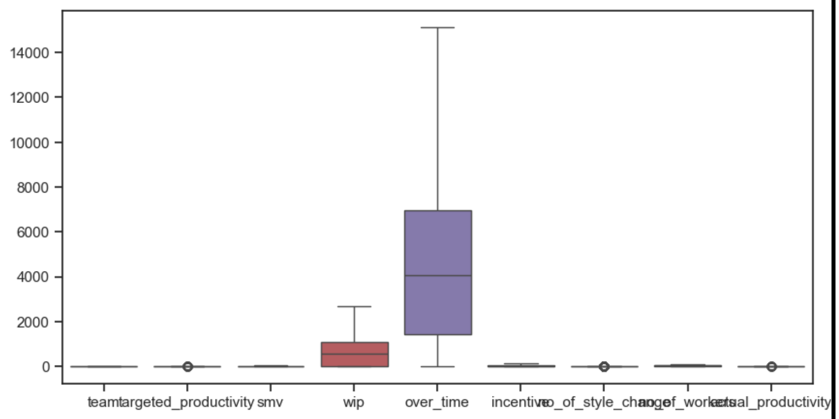
</td></tr>
<tr><td>Bivariate Analysis</td><td></td></tr>
</table>


Correlation Heatmap
team 1 0.036 -0.11 0.017 -0.11 -0.057 -0.011 -0.075 -0.14
targeted_productivity 0.036 1 -0.074 -0.07 -0.071 0.22 -0.21 -0.087 0.42
smv -0.11 -0.074 1 0.73 0.69 0.6 0.32 0.91 -0.13
wip 0.017 -0.07 0.73 1 0.65 0.72 0.21 0.82 -0.021
over_time -0.11 -0.071 0.69 0.65 1 0.55 0.06 0.74 -0.044
incentive -0.057 0.22 0.6 0.72 0.55 1 0.0075 0.7 0.31
no_of_style_change -0.011 -0.21 0.32 0.21 0.06 0.0075 1 0.33 -0.21
no_of_workers -0.075 -0.087 0.91 0.82 0.74 0.7 0.33 1 -0.064
actual_productivity -0.14 0.42 -0.13 -0.021 -0.044 0.31 -0.21 -0.064 1
team
targeted_productivity
smv
wip
over_time
incentive
no_of_style_change
no_of_workers
actual_productivity
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Scatter Matrix
targeted_productivity
0.8
0.6
0.4
0.2
smv
50
40
30
20
10
wip
2500
2000
1500
1000
500
0
over_time
15000
12500
10000
7500
5000
2500
0
0.25
0.50
0.75
targeted_productivity
20
40
smv
0
1000
2000
wip
0
5000
10000
15000
over_time

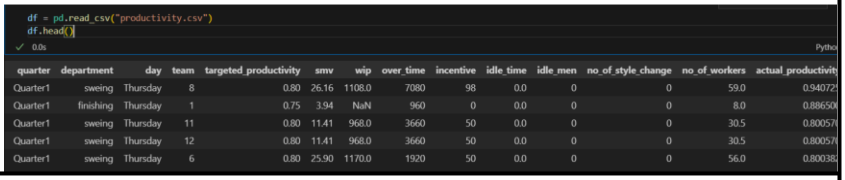
Outliers and Anomalies





Data Preprocessing Code Screenshots

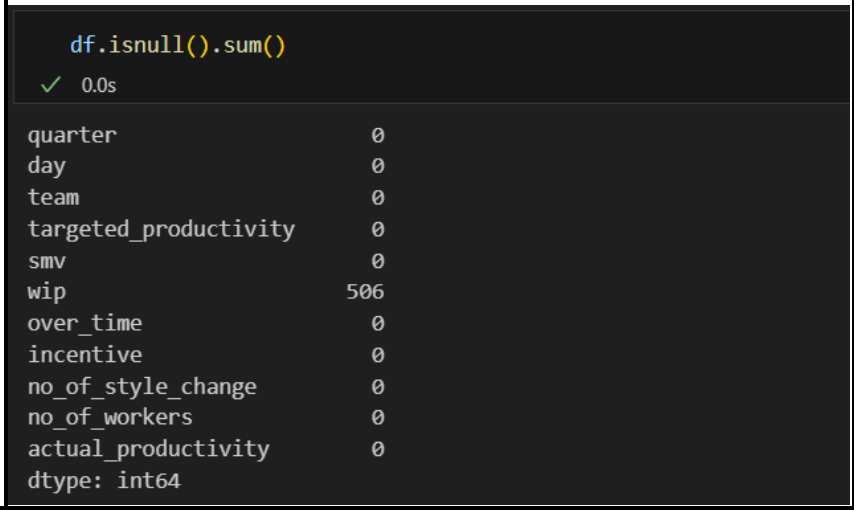
Loading Data



```
df = pd.read_csv("productivity.csv")
df.head()
✓ 0.0s                                                    Pytho
```

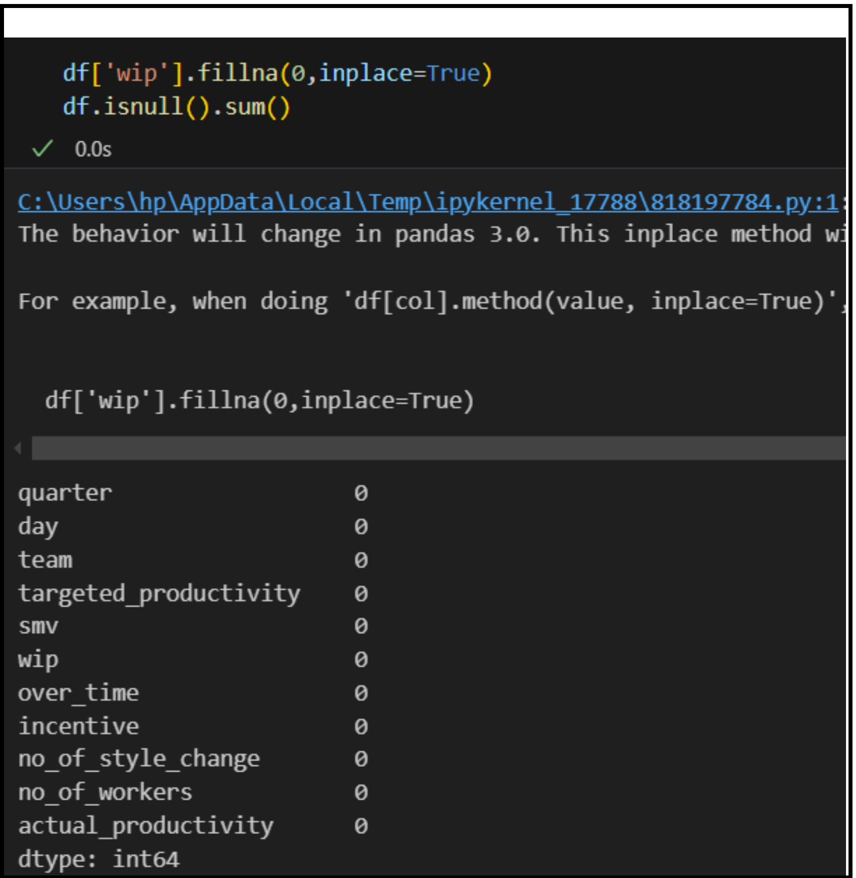
quarter	department	day	team	targeted_productivity	smv	wip	over_time	incentive	idle_time	idle_men	no_of_style_change	no_of_workers	actual_productivity
Quarter1	sweing	Thursday	8	0.80	26.16	1108.0	7080	98	0.0	0	0	59.0	0.94072
Quarter1	finishing	Thursday	1	0.75	3.94	NaN	960	0	0.0	0	0	8.0	0.88650
Quarter1	sweing	Thursday	11	0.80	11.41	968.0	3660	50	0.0	0	0	30.5	0.80057
Quarter1	sweing	Thursday	12	0.80	11.41	968.0	3660	50	0.0	0	0	30.5	0.80057
Quarter1	sweing	Thursday	6	0.80	25.90	1170.0	1920	50	0.0	0	0	56.0	0.80038

Handling Missing Data



```
df.isnull().sum()
✓ 0.0s

quarter                  0
day                      0
team                     0
targeted_productivity    0
smv                      0
wip                    506
over_time                0
incentive                0
no_of_style_change       0
no_of_workers            0
actual_productivity      0
dtype: int64
```



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

```
✓ 0.0s
```

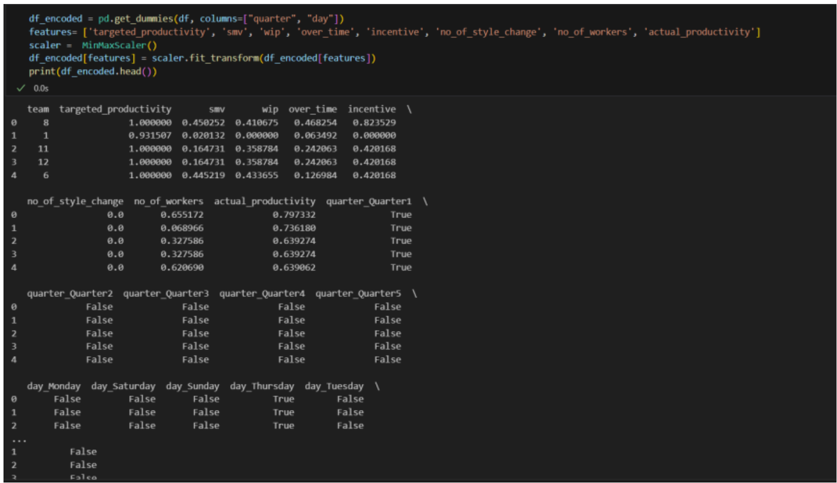
```
C:\Users\hp\AppData\Local\Temp\ipykernel_17788\818197784.py:1
The behavior will change in pandas 3.0. This inplace method wi

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())
```

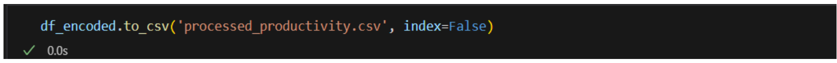
```
✓ 0.0s
```

```
   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)` ✓ 0.0s