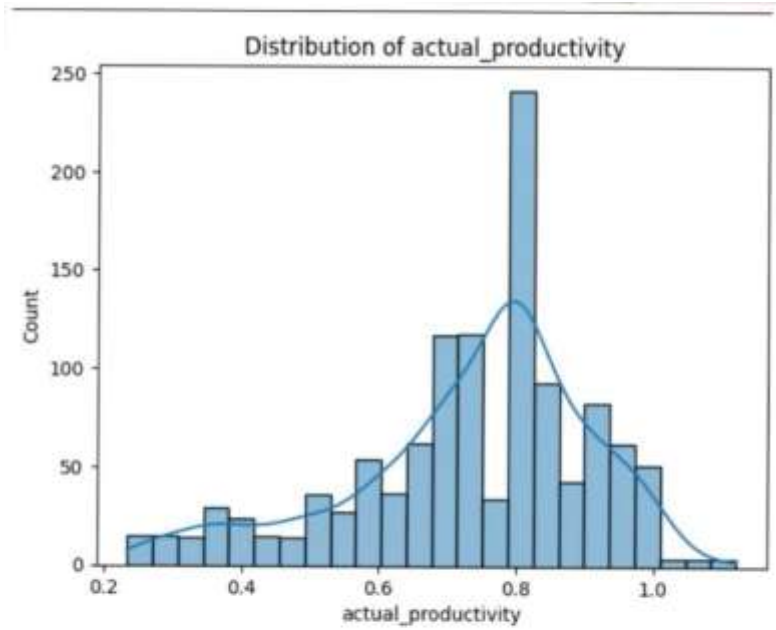
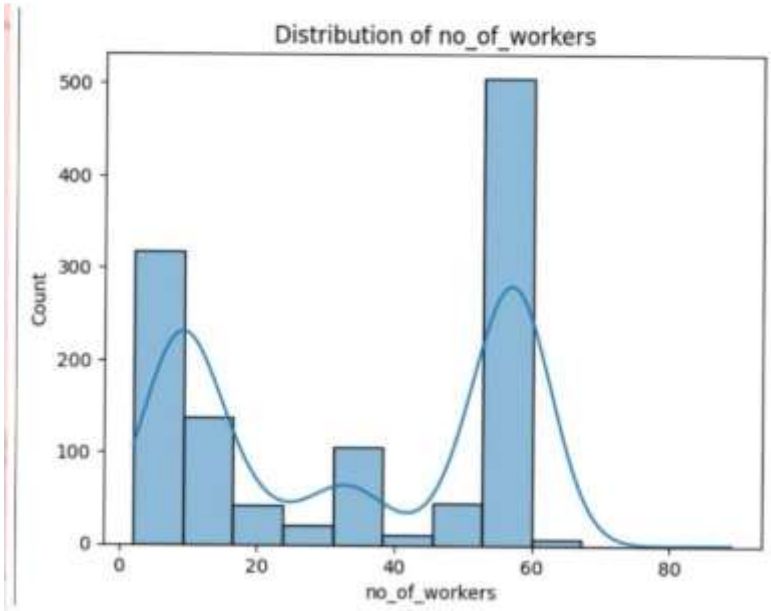
Data Collection and Preprocessing Phase

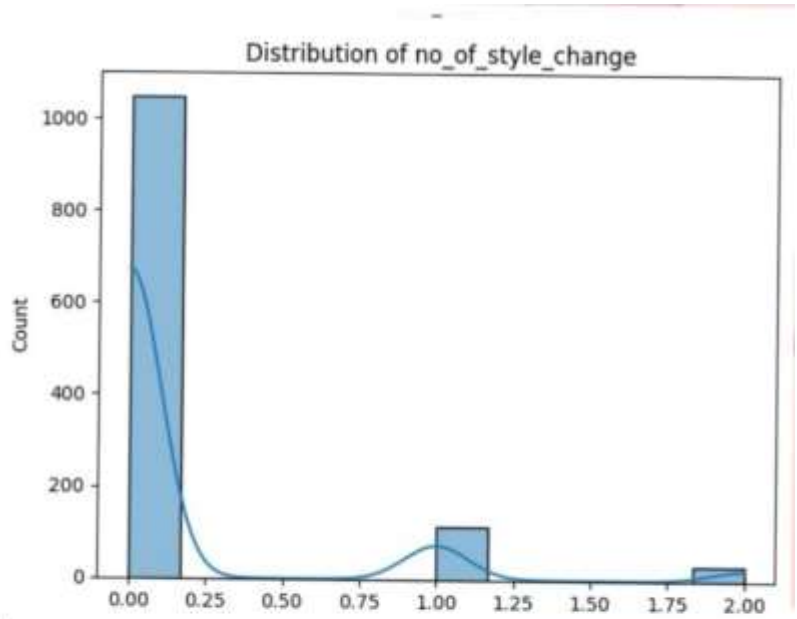
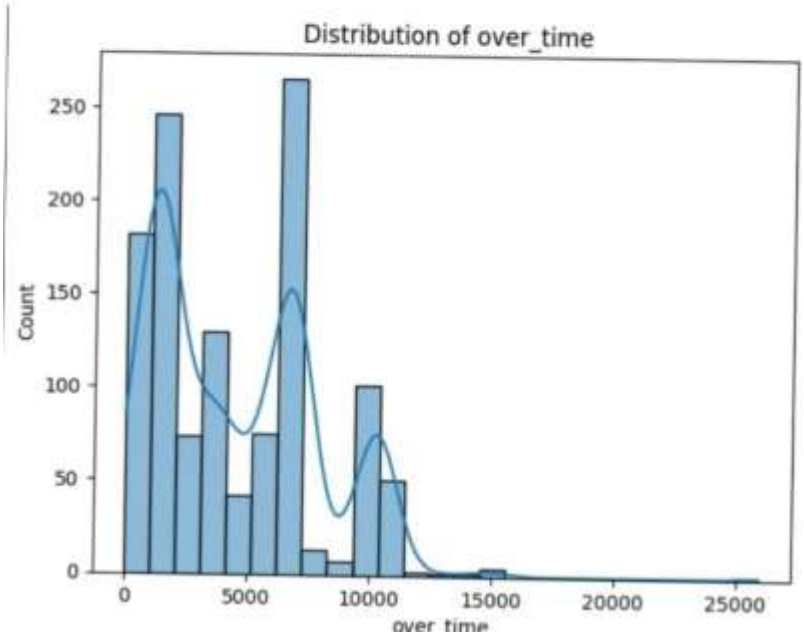
Date	15 July 2024
Team ID	SWTID1720108643
Project Title	Garment Worker Productivity Prediction
Maximum Marks	6 Marks

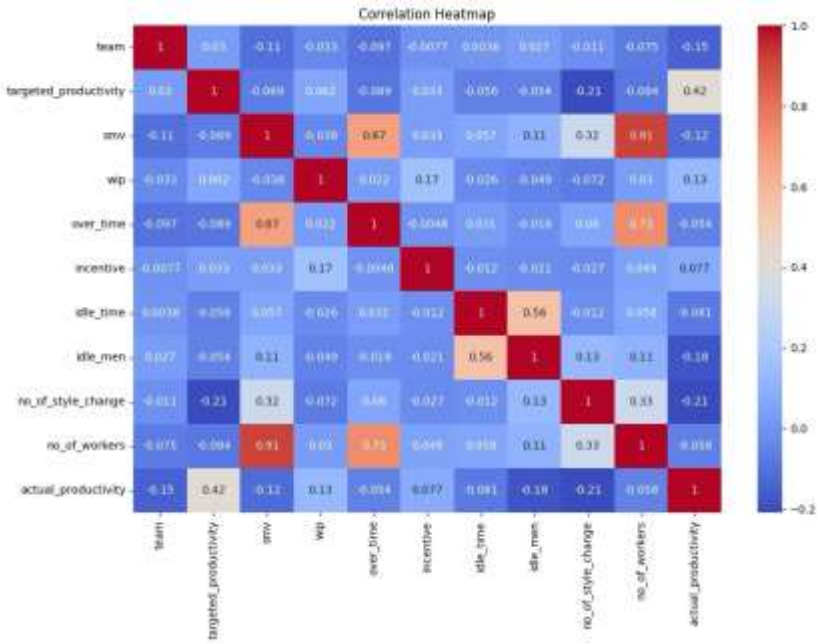
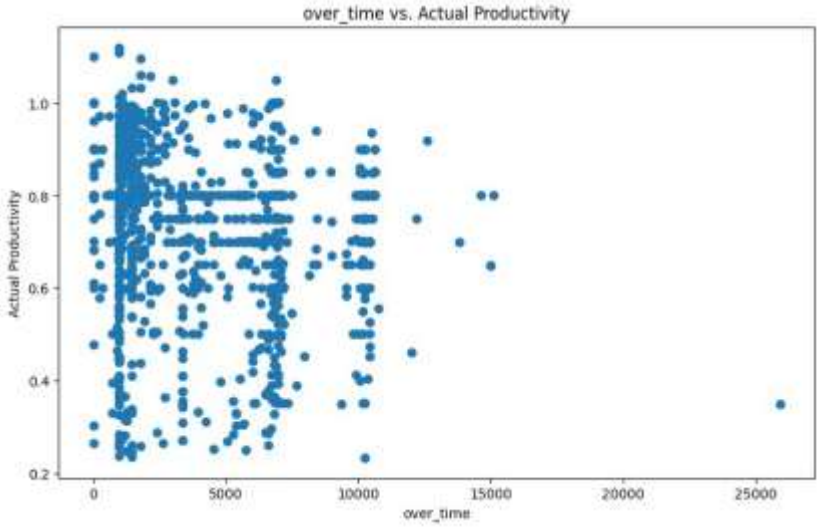
Data Exploration and Preprocessing Template

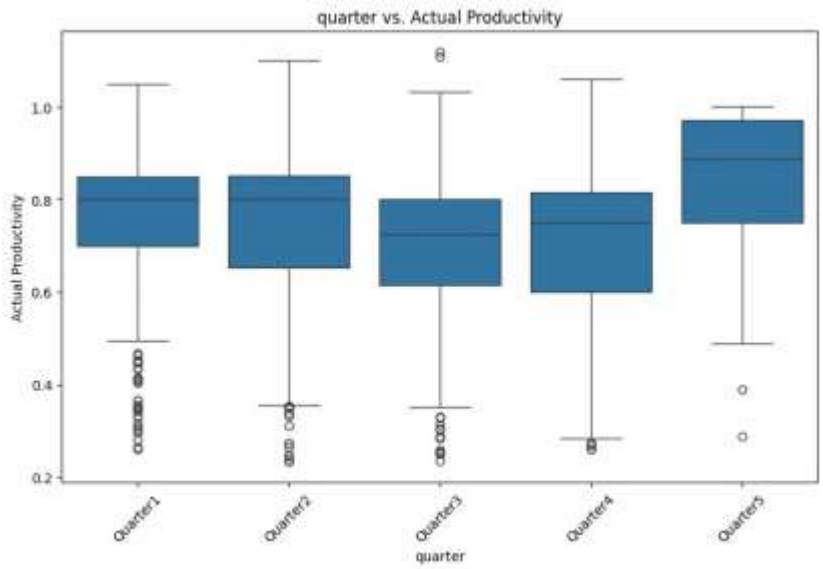
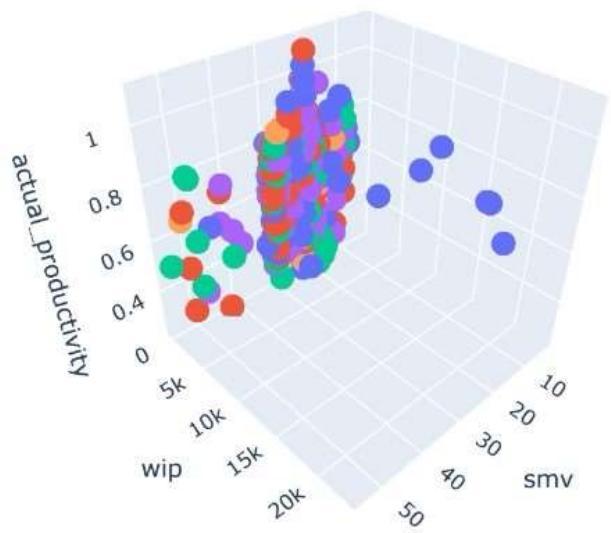
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

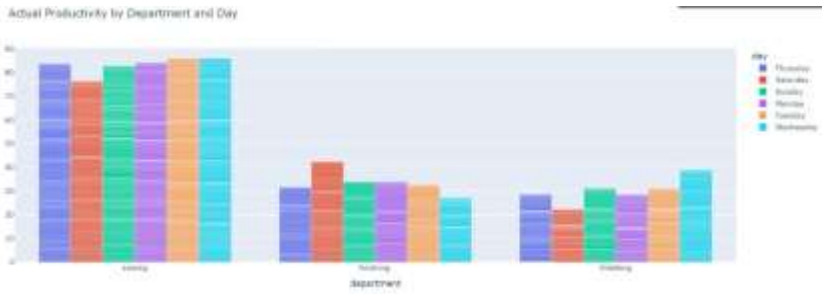
Section	**Description**
Data Overview	<u>The Shape of the Data Frame:</u> **(1197, 13)** It consists of **1197** Rows and **13** Columns. 

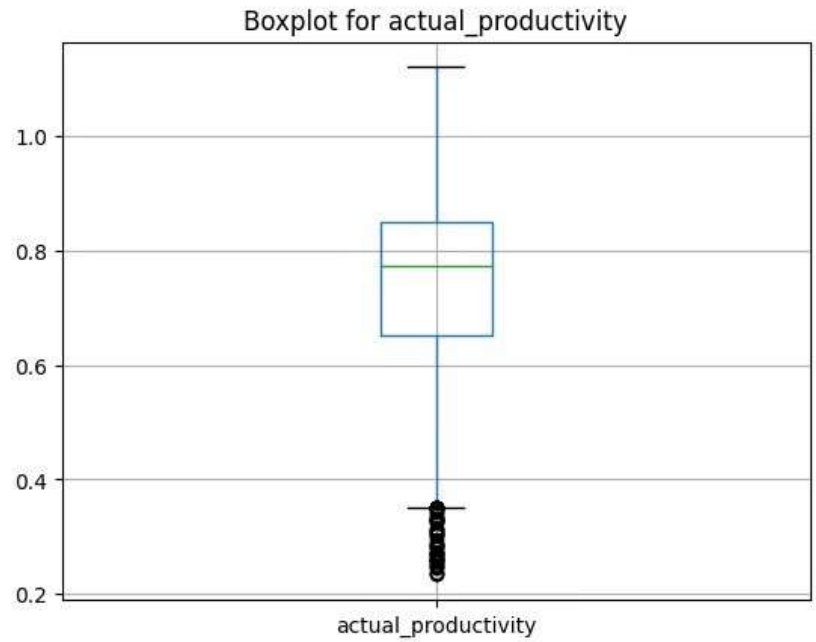
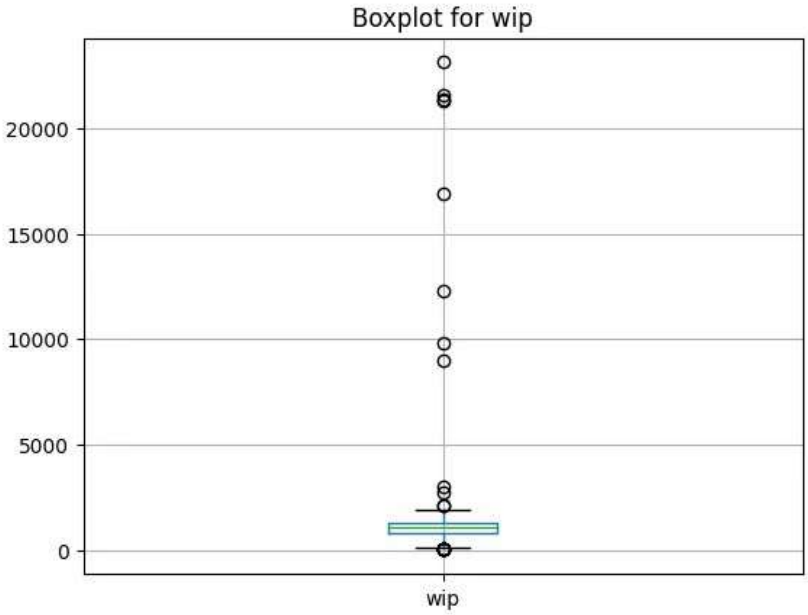
Univariate Analysis	Exploration of individual variables (mean, median, mode, etc.).  

Univariate Analysis	 

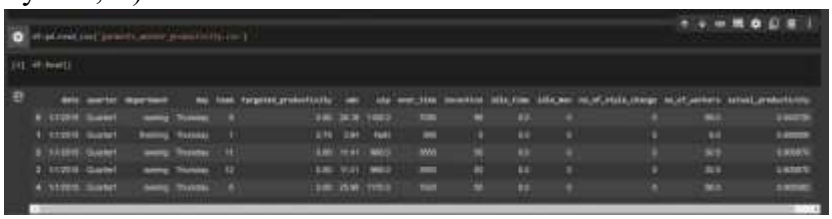
Bivariate Analysis	 

Bivariate Analysis	
Multivariate Analysis	Patterns and relationships involving multiple variables. 

Multivariate Analysis	

Outliers and Anomalies	Identification and treatment of outliers.  

Data Preprocessing Code Screenshots

Loading Data	Code to load the dataset into the preferred environment (e.g., Python, R). 

Handling Missing Data	Code for identifying and handling missing values.
	`df.isna().sum()`
	`date 0`
	`quarter 0`
	`department 0`
	`day 0`
	`team 0`
	`targeted_productivity 0`
	`smv 0`
	`wip 506`
	`over_time 0`
	`incentive 0`
	`idle_time 0`
	`idle_men 0`
	`no_of_style_change 0`
	`no_of_workers 0`
	`actual_productivity 0`
	`dtype: int64`
	`[56] df['wip']=df['wip'].fillna(df['wip'].mean())`

Data Transformation	Code for transforming variables (scaling, normalization). “from sklearn.preprocessing import StandardScaler, MinMaxScaler scaler = StandardScaler() numerical_features = ['smv', 'wip', 'over_time', 'incentive', 'idle_time', 'idle_men', 'no_of_style_change', 'no_of_workers'] df[numerical_features] = scaler.fit_transform(df[numerical_features]) df.head()” 
Feature Engineering	Code for creating new features or modifying existing ones. 

Save Processed Data	Code to save the cleaned and processed data for future use. `df.to_csv('preprocessed_data.csv', index=False)`