Data Collection and Preprocessing Phase

Date	7th July 2024
Team ID	740141
Project Title	Garment Workers Productivity Predictions
Maximum Marks	6 Marks

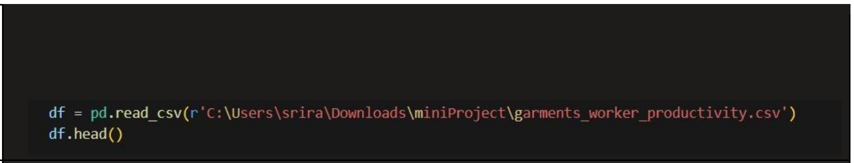
Data Exploration and Preprocessing Template

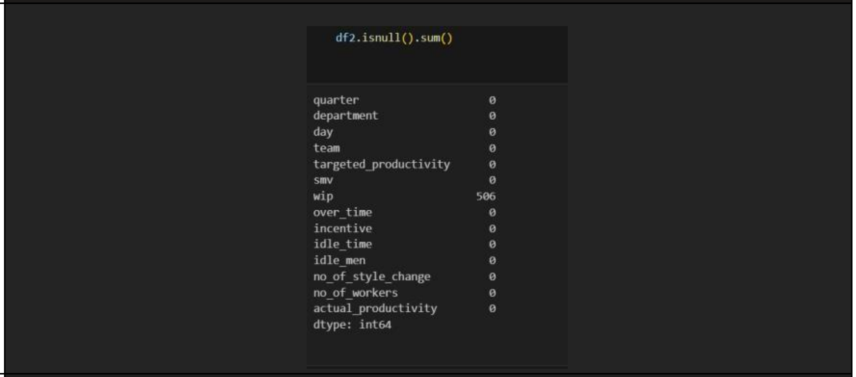
Data collection for gather data from internal and external sources, assesses quality issues like missing values and duplicates, normalize features,engineer new features, data merging and splitting then split into training and testing.

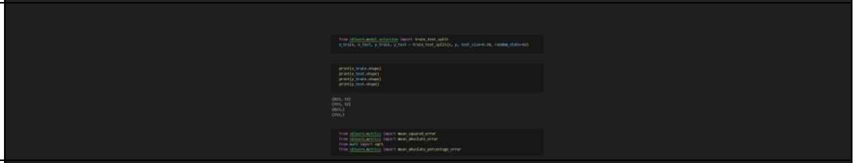
Section	**Description**
Data Overview	**Basic statistics:**Summarize the central tendency, dispersion, and shape of the dataset's distribution. **dimensions:**Display the number of rows and columns in each dataset to understand the size of the data. **structure of the data**:Provide information about the data types and non-null values in each column to understand the structure.

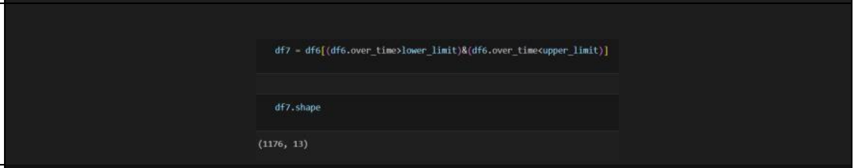
Univariate Analysis	**1. Summary Statistics** - **Description:** Calculate and display basic statistics such as mean, median, mode, standard deviation, and range for individual variables. 2.. **Distribution Plots** - **Description:** Visualize the distribution of individual variables using histograms and density plots. **3. Box Plots** - **Description:** Use box plots to visualize the spread and identify potential outliers in individual variables This code will provide a comprehensive univariate analysis for individual variables, helping to understand their distributions, central tendencies, and variations.
Bivariate Analysis	**1.Correlation** - **Description:** Calculate the correlation coefficient to measure the strength and direction of the relationship between two variables. **2. Scatter Plot** - **Description:** Create scatter plots to visually inspect the relationship between two variables.

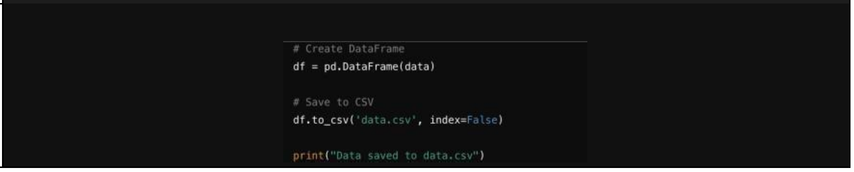
Multivariate Analysis	**1. Pair Plot** • **Description:** Visualize pairwise relationships between multiple variables to identify potential interactions. **2. Heatmap for Correlation Matrix** • **Description:** Visualize the correlation matrix to see the strength of relationships between multiple variables. **3. Multiple Regression Analysis** • **Description:** Fit a multiple regression model to understand the combined effect of multiple predictors on a target variable.
Outliers and Anomalies	**Treatment of Outliers:** **1.Strategies** ○ **Remove outliers**: Exclude them from further analysis if they are likely to be errors or irrelevant extreme values. ○ **Transform variables**: Apply mathematical transformations like log or square root to reduce skewness caused by outliers. ○ **Winsorization**: Cap extreme values by replacing them with values at a specified percentile (e.g., 95th percentile). **2.Considerations**: ○ **Impact assessment**: Evaluate how removing or transforming outliers affects the overall dataset and model performance. ○ **Documentation**: Document all decisions and actions taken regarding outlier handling for transparency and reproducibility.
Data Preprocessing Code Screenshots	

Loading Data	
	df = pd.read_csv(r'C:\Users\srira\Downloads\miniProject\garments_worker_productivity.csv')
	df.head()

Handling Missing Data	 df2.isnull().sum() 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

Data Transformation	
	from sklearn.model_selection import train_test_split
	x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=42)
	print(x_train.shape)
	print(x_test.shape)
	print(y_train.shape)
	print(y_test.shape)
	[illegible]
	[illegible]
	[illegible]
	[illegible]
	from sklearn.metrics import mean_squared_error
	from sklearn.metrics import mean_absolute_error
	from math import sqrt
	from sklearn.metrics import mean_absolute_percentage_error

Feature Engineering	 df7 = df6[(df6.over_time>lower_limit)&(df6.over_time<upper_limit)] df7.shape (1176, 13)

Save Processed Data	
	# Create DataFrame
	df = pd.DataFrame(data)
	# Save to CSV
	df.to_csv('data.csv', index=False)
	print("Data saved to data.csv")