


Data Collection and Preprocessing Phase

Date	05 August 2025
Skillwallet ID	SWUID20250186419
Project Title	Employee Performance 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	The first step is to load the garments_worker_productivity.csv dataset and get a high-level understanding of its structure. This involves checking the number of rows and columns, the data types of each feature (e.g., numerical, categorical, datetime), and basic statistical summaries. We will identify the target variable, which is productivity, and the feature variables like team, department, wip, and others.
Univariate Analysis	This section involves analyzing each variable in isolation. For numerical variables like wip (work in progress) or smv (standard minute value), we would calculate descriptive statistics such as the mean, median, and standard deviation. We would also visualize their distributions using histograms to check for skewness. For categorical variables like department and team, we would analyze the frequency of each category using bar plots.
Bivariate Analysis	Here, we'll explore the relationships between pairs of variables. A correlation matrix and heatmap will be used to understand the linear relationships between the numerical features. Scatter plots will be created to visualize how individual features, such as smv or wip, relate to the target variable productivity. For categorical features, we can use box plots or violin plots to compare the distribution of productivity across different department or team values.

Multivariate Analysis	This involves examining relationships among three or more variables to uncover more complex patterns. For instance, we could use pair plots to visualize the relationships between multiple features simultaneously. We might also use conditional plots to see how the relationship between smv and productivity changes for different departments.
Outliers and Anomalies	We will identify and handle any extreme values in the data that could skew our model's performance. Box plots and scatter plots will be the primary visualization tools for outlier detection. Once identified, outliers will be handled either by removing them, transforming them, or capping them at a certain value, depending on the context.
Data Preprocessing Code Screenshots	
Loading Data	<pre>df=pd.read_csv('./content/garments_worker_productivity.csv') df.head()</pre> <p>✓ 0.0s  Open 'df' in Data Wrangler Python</p>
Handling Missing Data	<pre>df.isnull().sum()</pre> <p>Python</p>
Data Transformation	<pre>Mcle=MultiColumnLabelEncoder.MultiColumnLabelEncoder() data=Mcle.fit_transform(df)</pre> <p>Python</p> <pre>x=data.drop(['actual_productivity'],axis=1) y=data['actual_productivity'] X=x.to_numpy() X</pre> <p>Python</p>
Feature Engineering	<pre>df['month']=df['date'].dt.month df.drop(['date'],axis=1,inplace=True) df.month</pre> <p>Python</p>