**Define the Problem**

1. **Business Problem**

* Garment manufacturing is a labor-intensive business and therefore the performance of the employees is a crucial factor in the achievement of production goals and financial success. Nevertheless, there are several factors that can affect the employee productivity such as absenteeism, idle time, incentives and overtime. This is a problem for management since they struggle with understanding how these factors work together and how they will affect productivity, leading to flaws and potential mismanagement. The primary goal of this analysis is to:
* To examine the factors that affect the productivity of the employees.
* To suggest practical recommendations to manage the operations and increase productivity overall.
* **Data Source**

[**Productivity Prediction of Garment Employees - UCI Machine Learning Repository**](https://archive.ics.uci.edu/dataset/597/productivity+prediction+of+garment+employees)

* **Why Use Python?**

Python is an ideal tool for solving this problem due to the following reasons:

* **Powerful Data Analysis Libraries:**  
  Python offers robust libraries like pandas for handling large datasets, performing exploratory data analysis (EDA), and calculating descriptive statistics.
* **Data Visualization Capabilities:**  
  Libraries such as matplotlib and seaborn enable the creation of insightful visualizations to reveal patterns, trends, and relationships within the data.
* **Flexibility and Efficiency:**  
  Python allows for seamless integration of data cleaning, transformation, and analysis, making it efficient for end-to-end problem-solving.
* **Scalability:**  
  Python can handle datasets of varying sizes, ensuring scalability as the data grows.

1. **Describe the dataset:**

**a. How many variables are in the dataset?**

The dataset contains 15 variables. These variables have been categorized under numerical and categorical capturing information about productivity, worker-related metrics, and operational data:

* **Numerical Predictor Variables**

These are variables that are quantitative and can take any value within a range.

* **Targeted Productivity**  
  Represents the productivity goal or target set for employees.
* **Actual productivity**  
  It represents the actual productivity of employees in the garment industry, which could be impacted by various operational and workforce factors.
* **SMV (Standard Minute Value)**  
  Measures the time required to produce one unit of a product. It’s used to set production goals and measure efficiency.
* **WIP (Work in Progress)**  
  The number of unfinished garments at any given time. It can influence productivity by indicating if there’s a bottleneck in production.
* **Overtime**  
  The number of overtime hours worked by employees. High overtime can indicate pressure to meet targets, affecting productivity.
* **Incentive**  
  Monetary or non-monetary rewards given to employees based on their productivity. It’s assumed to influence employee effort and productivity.
* **Idle Time**  
  The amount of time employees are not working, which could be due to various reasons such as equipment breakdowns or waiting for materials.
* **Idle Men**  
  The number of employees who are idle (not working) due to operational inefficiencies.
* **Number of Style Change**  
  The number of style changes in the production process. Frequent style changes may disrupt the flow and reduce productivity.
* **Number of Workers**  
  The total number of workers in the production line. A larger workforce might be associated with increased productivity, but it could also lead to inefficiencies if not managed well.
* **Categorical Predictor Variables**

These variables represent categories or groups and are typically used to segment or group data.

* **Date**  
  Represents the date of the observation, which could indicate seasonality or changes over time (e.g., higher productivity in certain months).
* **Quarter**  
  The fiscal quarter in which the data was collected (e.g., Q1, Q2, Q3, Q4) can help identify patterns influenced by seasonal effects or financial periods. However, upon cleaning the dataset, I discovered that the dates in the original dataset were inaccurately recorded, and all the data fell within the first quarter (Q1). To enable more precise data visualization, the timeframe was adjusted to reflect months instead.
* **Department**  
  Represents the department in the garment factory (e.g., sewing, finishing). Different departments might have varying levels of productivity.
* **Day**  
  Represents the day of the week (e.g., Monday, Tuesday). This can reveal weekly productivity trends.
* **Team**  
  Denotes the team in the production process. Productivity can vary across teams based on management, skills, or other factors.

**b. What are the data types and levels of measurement for each of the variables?**

**Data Types and Levels of Measurement**

* **Numerical**: These variables represent quantities and can be measured on a continuous or discrete scale. They can be further classified into:
  + **Continuous**: Variables that can take any value within a range (e.g., time, productivity, etc.).
  + **Discrete**: Variables that take distinct integer values (e.g., number of workers, number of style changes).
* A table with text on it

  Description automatically generated**Categorical**: These variables represent categories or groups, and they often cannot be measured numerically but can be grouped or classified into different categories.

1. **How many observations are in the data set?**

The dataset contains **1197 observations**. Each observation represents a record or instance in the dataset, typically corresponding to a specific time point or production instance in the garment production process.

1. **Are there any missing values or outliers in the data set?**

Upon reviewing the dataset, I observed that there are **516 missing values** in the variable WIP (Work In Progress). These missing values represent finished garments where no work is currently ongoing. Therefore, these missing values do not imply errors or incomplete data but carry meaningful information. Yes, there are **332 outliers** in the dataset. Using statistical methods, Interquartile Range (IQR), I identified the outliers.

1. **How did you choose to handle the missing values and outliers?**

To handle these missing values, I choose to leave them in the data set as they do not distort the report in any way, as they accurately reflect that there is no work in progress for those observations. This approach ensures the data remains consistent and interpretable while avoiding distortion of statistical analyses.

For the outliers the decision on how to handle outliers was based on the nature of each variable and the context of the dataset. The following strategies were considered and implemented:

* **Retaining Outliers When Relevant**  
  For variables like targeted productivity, actual productivity, idle time, WIP, overtime and incentive the outliers were retained because they reflect real-world deviations in operational performance. Removing these values could eliminate important insights into the factors influencing productivity.
* **Extreme Outliers**  
  For variables such as number of style changes, where the number of outliers was very high, capping was supposed to be applied to limit their impact on analysis. However, these values fall within a narrow range (0 to 2) and are contextually reasonable, reflecting real-world production scenarios.

1. **What are some of the descriptive statistics for key variables?**

Descriptive statistics are essential for summarizing and understanding the key characteristics of the data. These statistics help to explore the distribution, central tendency, and variability of each variable in the dataset. For this project, I will analyze the following key statistics for the key variables in the dataset:

1. **Central Tendency**:

* **Mean**: The average value for each variable, which gives a sense of the typical value within the dataset.
  + **Median**: The middle value when the data is sorted in ascending order. This is useful for identifying whether the data is skewed or symmetrical.

1. **Dispersion**:
   * **Standard Deviation**: This measures the spread or variability of the data around the mean. A higher standard deviation indicates more variability, while a lower standard deviation indicates that the data points are closer to the mean.
   * **Max**: The maximum values for each variable in the data set.
   * **Min**: The minimum values for each variable in the data set.
   * **Interquartile Range (IQR)**: The range between the 25th and 75th percentiles, giving a measure of the spread of the middle 50% of the data.
2. **Summary and Descriptive Statistics**

* **Targeted Productivity**
* **Mean:** The average targeted productivity is **0.73**, indicating that, on average, teams aim to achieve about 73% of the maximum possible productivity.
* **Median:** The median is **0.75**, close to the mean, suggesting the distribution is symmetrical.
* **Standard Deviation:** A low standard deviation of **0.09** indicates that most targeted productivity values are close to the mean, with little variation.
* **Min and Max:** The minimum is **0.07**, and the maximum is **0.80**, indicating a narrow range for productivity goals.
* **SMV (Standard Minute Value)**
* **Mean:** The mean SMV is **15.06 minutes**, indicating the average time allocated to produce a specific garment.
* **Median:** The median is **15.26**, very close to the mean, suggesting a symmetrical distribution.
* **Standard Deviation:** A higher standard deviation of **10.94** indicates variability in task complexity.
* **Min and Max:** The SMV ranges from **2.9** to **54.56**, highlighting some extreme tasks.
* **WIP (Work in Progress)**
* **Mean:** The average WIP is **1190.47**, indicating a substantial number of incomplete garments in production.
* **Median:** The median is **1039**, much lower than the mean, showing the presence of outliers.
* **Standard Deviation:** A high standard deviation of **1837.46** confirms significant variability.
* **Min and Max:** The minimum is **7**, and the maximum is **23,122**, showing that some production lines have extremely high WIP values.
* **Over Time**
* **Mean:** The average overtime is **4567.46 minutes**, suggesting significant extra work hours.
* **Median:** The median is **3960**, lower than the mean, indicating the influence of outliers.
* **Standard Deviation:** A standard deviation of **3348.82** reflects high variability.
* **Min and Max:** Overtime ranges from **0** to **25,920**, showing a wide disparity.
* **Incentive**
* **Mean:** The mean incentive is **38.21**, indicating that, on average, employees receive a relatively low incentive amount.
* **Median:** The median incentive is **0,** significantly lower than the mean, suggesting that a majority of employees receive no incentive**.**
* **Standard Deviation:** A high standard deviation of **160** reflects substantial variability in the incentive amounts, with some employees receiving significantly more than others.
* **Min and Max:** Incentives range from **0** to **3600**, highlighting extreme cases where a few employees receive very high incentives.
* **Idle Time**
* **Mean:** The average idle time is **0.73 minutes**, indicating minimal idle time on average.
* **Median:** The median is **0**, showing that many teams report no idle time.
* **Standard Deviation:** A standard deviation of **12.79** indicates variability due to outliers.
* **Min and Max:** Idle time ranges from **0** to **300**, showing a few extreme cases of inefficiency.
* **Idle Men**
* **Mean:** The average number of idle workers is **0.37**, indicating low idle manpower overall.
* **Median:** The median is **0**, suggesting most teams have no idle workers.
* **Standard Deviation:** A standard deviation of **3.27** reflects some variability.
* **Min and Max:** Idle workers range from **0** to **45**, with a few extreme cases.
* **Number of Style Changes**
* **Mean:** The average number of style changes is **0.15**, indicating that teams rarely switch styles.
* **Median:** The median is **0**, showing that most teams do not experience style changes.
* **Standard Deviation:** A low standard deviation of **0.42** supports this.
* **Min and Max:** Style changes range from **0** to **2**, with a very narrow distribution.
* **Number of Workers**
* **Mean:** The average number of workers is **34.55**, indicating medium-sized teams.
* **Median:** The median is **34**, close to the mean, suggesting symmetry.
* **Standard Deviation:** A standard deviation of **22.15** indicates some variability in team size.
* **Min and Max:** Teams range from **2** to **89** workers, showing some extreme variations.
* **Actual Productivity**
* **Mean:** The mean actual productivity is **0.73**, closely aligned with targeted productivity.
* **Median:** The median is **0.77**, slightly higher than the mean, indicating that teams often meet or exceed targets.
* **Standard Deviation:** A low standard deviation of **0.17** suggests consistent performance.
* **Min and Max:** The range (**0.23–1.12)** highlights both underperforming and high-achieving teams.
* **Summary of Descriptive Statistics**

A screenshot of a data table

Description automatically generated

* To evaluate the relationship between categorical variables a pivot table was created, the relationship between **Department** and **Month** reveals that the Sewing department consistently handles more tasks than the Finishing department across all months. January shows the highest activity in both departments, while March experiences the lowest, suggesting seasonality in production. Additionally, the sharp drop in tasks for the Finishing department compared to Sewing highlights potential imbalances in workload or resource allocation.

**A screenshot of a computer

Description automatically generated**

**Correlation Matrix - To Show the Relationship Between the Numerical Variables**

Correlation measures the strength and direction of a linear relationship between two numerical variables. The values range from **-1** (strong negative) to **1** (strong positive), with 0 indicating no correlation. Here's how we categorize the relationships based on their absolute values:

1. **Strong Relationship**: Correlation coefficient **≥ 0.7** or **≤ -0.7**
2. **Medium Relationship**: Correlation coefficient between **0.4 and 0.7** or **-0.4 and -0.7**
3. **Low Relationship**: Correlation coefficient between **0 and 0.4** or **-0.4 and 0**

**Categorization of Correlation (Based on the Matrix)**

1. **Strong Relationships**

* **SMV and Number of Workers**: **0.912**
* This suggests that as the number of workers increases, the SMV (standard minute value) tends to increase significantly.
* **Overtime and Number of Workers:** **0.733**
* Teams with more workers tend to spend more time on overtime.

1. **Medium Relationships**

* **SMV and Number of Workers**: **0.675**
* The higher the SMV, the more overtime is required to complete tasks.
* **Idle Men and Idle Time : 0.56**
* A moderate correlation where idle time tends to increase as the number of idle men increases. This could indicate inefficiency or underutilization of resources.
* **Actual productivity and Targeted productivity:** **0.422**
* It indicates that actual productivity moderately aligns with targeted productivity.

1. **Low Relationships**

* **Number of Workers and Number of Style change: 0.33**
* Indicates a weak relationship between the number of workers and the frequency of style changes.
* WIP and all other variables: Correlations are low, suggesting work-in-progress has minimal direct linear relationships with other factors.
* Idle time and all variables: Weak correlation values show minimal influence or relationship with other factors.

**Numerical Correlation Matrix**

**A screenshot of a computer screen

Description automatically generated**

**G) Create data visualizations for variables in the data set.**

1. **Monthly Activity Comparison Between Finishing and Sewing Departments**

An analysis of departmental activity by month reveals significant variations. January records the highest activity levels, particularly in the sewing department, while March experiences the lowest across both departments. Activity decreases steadily from February to March, with the sewing department consistently outperforming finishing each month. These trends suggest that the sewing department is consistently optimized, whereas the finishing department and overall activity in March require targeted strategies to boost performance

A graph of blue and orange bars

Description automatically generated

1. **Analysis of Average productivity by Month**

A graph with a line

Description automatically generatedAn analysis of average productivity by month reveals a clear downward trend from January to February, with productivity dropping significantly. While there is a slight recovery in March, it remains below January's levels. This indicates that productivity peaked in January, followed by a sharp decline, and failed to regain its initial momentum in the subsequent months. These trends suggest that February faces challenges affecting overall performance, while March shows potential for improvement. Understanding the factors contributing to January's high productivity could help develop strategies to sustain or replicate similar performance throughout the quarter.

1. **Analysis of Actual and Targeted Productivity by Month**

An analysis of actual versus targeted productivity by month reveals a significant divergence between the two metrics. In January, actual productivity starts at its highest point, but it declines sharply through February and March. In contrast, the targeted productivity maintains a steadier, more gradual decrease, staying relatively consistent through February and March. This suggests that while the targets set for the team remained steady, actual performance fell below expectations, particularly in February. The discrepancy highlights the need for further investigation into the factors causing the drop in actual productivity and whether adjustments to targets or resources are necessary to align with achievable performance.

A graph of a graph showing the difference between a product and a product

Description automatically generated with medium confidence

1. **Analysis of Actual Productivity by Day**

An analysis of actual productivity by day reveals significant variations. Saturday records the highest productivity, while Thursday experiences the lowest. Productivity during the weekdays fluctuates, with a notable recovery on Tuesday before declining again by Wednesday. These trends suggest that weekend productivity, particularly on Saturday, is optimized, whereas Sundays and midweek require targeted strategies to boost performance. Understanding factors contributing to weekend peaks could inform approaches to enhance overall productivity throughout the week.

A line graph with text and a blue line

Description automatically generated with medium confidence

1. **Analysis of Productivity by Team**

A graph with many rectangular boxes

Description automatically generated with medium confidenceAn analysis of productivity by team reveals significant variations. Teams 3 and 4 demonstrate the highest and most consistent productivity, while Teams 6, 9, and 12 show greater variability with wider ranges and numerous outliers. Most teams maintain a median productivity between 60% and 80%, but low-performing outliers are particularly noticeable in Teams 6, 9, and 12. These trends suggest that certain teams, such as Teams 3 and 4, have optimized performance, whereas Teams 6, 9, and 12 may require targeted strategies to address inconsistencies and improve overall productivity. Identifying factors driving high performance in Teams 3 and 4 could help inform strategies to elevate productivity across all teams.

1. **Analysis of relationship between Number of Workers and SMV**

A graph with dots on it

Description automatically generatedAn analysis of the relationship between the number of workers and Standard Minute Value (SMV) reveals significant patterns. Tasks with fewer than 20 workers generally have the lowest and most consistent SMV, indicating they are simple and require minimal time. In contrast, tasks involving 40 to 60 workers show a wider range of SMVs, with values increasing from 10 to over 30, reflecting greater complexity and time requirements. Notably, a few outliers exist, such as tasks with SMVs exceeding 40 and one instance with around 80 workers, which may represent specialized or large-scale operations.

1. **Analysis of relationship between Number of Workers and Overtime**

An analysis of the relationship between the number of workers and overtime reveals significant patterns. Tasks with fewer than 20 workers generally have the lowest and most consistent overtime, typically ranging between 0 and 5,000 hours, indicating they are manageable within regular working hours. In contrast, tasks involving 40 to 60 workers show a wider range of overtime, with values reaching up to 10,000 hours, reflecting greater complexity or urgency. Notably, a few outliers exist, such as tasks with overtime exceeding 25,000 hours and one instance with nearly 80 workers, which may represent exceptional or large-scale operations requiring extensive additional time.

A graph with dots and numbers

Description automatically generated

1. **Analysis of relationship between SMV and Overtime**

A graph with dots and numbers

Description automatically generatedAn analysis of the relationship between SMV (Standard Minute Value) and overtime reveals notable trends. Tasks with lower SMV values (below 10) generally have low and consistent overtime, mostly under 5,000 hours, indicating they are relatively simple and time efficient. As SMV increases between 10 and 30, overtime becomes more variable, with some tasks reaching up to 10,000 hours, reflecting increased task complexity or urgency. However, even tasks with high SMV (above 30) often show a clustering around moderate overtime levels, though a few outliers exceed 25,000 hours, likely representing exceptional or highly time-intensive operations.

1. **Analysis of relationship between Target Productivity and Actual Productivity**

The scatter plot reveals the relationship between **targeted productivity** and **actual productivity** in the dataset. The data demonstrates a strong clustering around higher targeted productivity values (0.7–0.8), with many corresponding instances of high actual productivity (0.7–1.0). This suggests that the workforce generally achieves or comes close to meeting the set productivity targets.

However, there are notable deviations in certain cases, where actual productivity falls short despite high targets. For example, some instances show targeted productivity around 0.7–0.8, but actual productivity is below 0.5, indicating potential inefficiencies or challenges in meeting targets. Additionally, there are instances where actual productivity reaches the maximum value (1.0), suggesting optimal performance or even exceeding expectations in some cases.

A graph with blue dots

Description automatically generated

**4) What do the descriptive statistics and visualizations that were created tell us about the data set? How can we use the results to “tell a story” about the data? What are the key findings and conclusions that you could present to the business to help answer their questions about the data from Part #1?**

a) The descriptive statistics and visualizations reveal significant insights into the dataset, providing a comprehensive understanding of the variables and their relationships.

* **Targeted Productivity:**
  + The average targeted productivity (0.73) aligns closely with actual productivity (mean = 0.73), indicating realistic goal setting.
  + However, slight variations (standard deviation of 0.17 for actual productivity) suggest some teams struggle to meet their targets consistently.
* **SMV (Standard Minute Value):**
  + The wide range of SMV values (2.9 to 54.56) indicates a variety of garment complexities. Tasks with higher SMVs require more workers and overtime, highlighting the importance of resource allocation for complex operations.
* **Overtime and Workforce Efficiency:**
  + Overtime is highly variable (0 to 25,920 minutes), with higher number of workers requiring more overtime.
* **Idle Time and Idle Men:**
  + Minimal idle time and idle manpower (median = 0 for both) indicate efficient use of resources in most teams.
* **Style Changes:**
  + The infrequent occurrence of style changes (mean = 0.15) suggests stable production environments, minimizing disruptions.
* **Seasonal Trends and Productivity:**
  + Productivity peaks in January and drops significantly in February and March.
  + This pattern highlights seasonal influences or operational inefficiencies during these months.

**b) Storytelling with the Data**

The data tells a story of a largely efficient production process with consistent productivity goals, minimal idle time, and effective workforce allocation. However, it also highlights areas requiring improvement:

* Seasonal drops in productivity suggest challenges in February and March that need addressing.
* Certain teams (e.g., 6, 9, and 12) require performance optimization to reduce variability and improve consistency.

**c) Key Findings from Visualizations**

* **Departmental Activities:**
  + The Sewing department consistently outperforms the Finishing department across all the months.
  + March records the lowest activity, signaling potential underutilization of resources.
* **Monthly Productivity Trends:**
  + January shows the highest productivity levels, followed by a decline in February and a slight recovery in March.
  + Strategies to sustain January’s performance could help optimize operations year-round.
* **Team Productivity:**
  + Teams 3 and 4 show consistently high productivity, serving as benchmarks for other teams.
  + Teams 6, 9, and 12 exhibit variability and inefficiencies, requiring targeted interventions.
* **Relationships Between Variables:**
  + A strong correlation between SMV and the number of workers (0.91) confirms that complex tasks require larger number of workers.
  + Similarly, overtime correlates moderately with SMV (0.67), emphasizing the need for additional resources for complex tasks.

**d) Key Recommendations for the Business**

* **Address Seasonal Productivity Declines:**
  + Investigate and mitigate factors causing drops in February and March, such as worker availability, resource constraints, or demand fluctuations.
* **Optimize Team Performance:**
  + Benchmark practices from high-performing teams (e.g., Teams 3 and 4) to improve underperforming teams.
* **Streamline Resource Allocation:**
  + Ensure resources are evenly distributed, especially in the Finishing department, to balance workloads and reduce overtime.
* **Leverage Data for Strategic Planning:**
  + Use correlations (e.g., SMV and workers/overtime) to plan workforce allocation and reduce unnecessary overtime.

**5. Results: What did you learn about the data? Describe how this exercise helped you become a better data analyst. What would you do differently? What were the biggest challenges when working with the data. Why is or isn’t Python a good tool to use for this type of analysis?**

* **What did you learn about the data?**

Through this analysis, I discovered key insights into productivity trends, resource allocation, and operational inefficiencies within the garment industry. Seasonal variations in productivity were evident, with overtime, SMV, and number of workers play critical roles in determining actual productivity

* **How did this exercise help you become a better data analyst?**

The project improved my preprocessing skills, particularly in cleaning of dataset. Generating descriptive analytics and visualization enhanced the ability to summarize data, while correlation analysis sharpened the skill of identifying impactful variables. Overall, it strengthened my capacity to translate technical findings into actionable business recommendations.

* **What would you do differently?**

To improve the analysis:

* Conduct deeper exploratory data analysis (EDA) to uncover hidden patterns.
* Use advanced, interactive visualizations for better insights.
* **What were the biggest challenges when working with the data?**

Key challenges included:

* Identifying and calculating the outliers and deciding what to do with them.
* Interpreting the scatter plot and the boxplot.
* Understanding and analyzing multivariate interactions affecting productivity.
* **Why is or isn’t Python a good tool to use for this type of analysis?**

Python proved to be an excellent tool for the analysis, offering:

* Data Processing: Efficient handling of data cleaning and transformation.
* Visualization: Clear, customizable charts with Matplotlib and Seaborn.
* Scalability: Ability to manage large datasets.
* Flexibility: Seamless integration with machine learning libraries.

**6. Code to solve the problem: Please copy/paste all Python code to solve the problem into a section of the document. Code should be commented to explain steps taken.**

***## Step 1 import all python libraries that will be required for my project work***

*import pandas as pd* ***# for data frames***

*import numpy as np* ***# for numerical computation***

*import matplotlib.pyplot as plt* ***# for data visualization***

*import seaborn as sns* ***# for data visualization***

***### Step 2 load data set unto python***

*df\_garments = pd.read\_csv('C:/Users/acer.DESKTOP-LM3JJLR/Documents/business analytics/productivity+prediction+of+garment+employees/garments\_worker\_productivity.csv')*

***### NOW WE START BY CLEANING OUR DATASET TO PREVENT ANY ANOMALIES###***

***#### Step 1 Information about the data types of each of the variable in our data set***

***## Study book reference page 154***

*df\_garments.info()*

***### Step 2 : check for missing values***

***### Study book reference page 164***

*df\_garments.isna().sum()*

***### Step 3: Convert the 'date' column to datetime format and ensure the correct format is applied***

***## Study book reference page 186***

*df\_garments['date'] = pd.to\_datetime(df\_garments['date'], errors='coerce')*

***### Step 4 Calculate the accurate quarter based on the 'date' column applied***

***###Study book reference page 316***

*df\_garments['quarter'] = df\_garments['date'].dt.to\_period('Q')*

***#### After cleaning the data i realized our quarter variable has only quarter one so decided to use the month instead of the quarter***

***# Step 5: Extract the full month name from the 'date' column***

***###study book reference page 413***

*df\_garments['month'] = df\_garments['date'].dt.strftime('%B')*

***# Step 6: Replace the 'quarter' column with the 'month' column***

***### study book reference page 170***

*df\_garments['quarter'] = df\_garments['month']*

***# ## Step 7: Then we delete the redundant month column since we no longer need it***

***###study book reference page 170***

*df\_garments =df\_garments.drop(columns=['month'])*

***### Step 8: Rename the 'quarter' column to 'month'***

***###study book reference page 291***

*df\_garments.rename(columns={'quarter': 'month'}, inplace=True)*

***### Step 9: Replace incorrect spelling in the 'department' column***

***###study book reference page 121***

*df\_garments['department'] = df\_garments['department'].replace('sweing', 'sewing')*

***### Step 10: Remove spaces between any spaces in the data that can distort the information***

***### study book reference page 121***

*df\_garments['department'] = df\_garments['department'].str.strip()*

***### Step 11: Convert idle men and no of workers to integers, since we cannot have a decimal man***

***### study book reference page 182***

*df\_garments['idle\_men'] = df\_garments['idle\_men'].astype(int)*

*df\_garments['no\_of\_workers'] = df\_garments['no\_of\_workers'].astype(int)*

*####* ***Step 11: Calculate our outliers the using the Interquartile range***

***### calculating the 25% quantile and 75% quantile***

*Q1 = df\_garments[['targeted\_productivity','smv','wip','over\_time','incentive',*

*'idle\_time','no\_of\_style\_change','actual\_productivity']].quantile(0.25)*

*Q3 = df\_garments[['targeted\_productivity','smv','wip','over\_time','incentive',*

*'idle\_time','no\_of\_style\_change','actual\_productivity']].quantile(0.75)*

*IQR = Q3 - Q1* ***#### Calculate the interquartile range***

***### to identify the lower and upper bound***

*lower\_bound = Q1 - 1.5 \* IQR* ***### minimum value to which a variable shouldn’t be lower***

*upper\_bound = Q3 + 1.5 \* IQR* ***### maximum value to which a variable shouldn’t exceed***

***### DESCRPTIVE STATISTICS AND ANALYSIS ###***

***###Step 1: Select only numeric columns, therefore excluding 'team','date','month','department','day'***

***###study book reference page 170***

*numerical\_columns = df\_garments.drop(columns=['team','date','month','department','day'])*

***####Step 2: Create a summary matrix***

***###study book reference page 170***

*Summary\_Stat = pd.DataFrame({*

*'mean': numerical\_columns.mean(),*

*'median': numerical\_columns.median(),*

*'std': numerical\_columns.std(),*

*'max': numerical\_columns.max(),*

*'min': numerical\_columns.min(),*

*'25% (Q1)': numerical\_columns.quantile(0.25),*

*'75% (Q3)': numerical\_columns.quantile(0.75)*

*})*

*print(Summary\_Stat)* ***###### to display the summary statistics***

***#######Step 3: show the relationship between the numerical data we use the correlation matrix***

***### project reference Chapter 17 in class bikers***

*correlation\_matrix = numerical\_columns.corr()*

***### DATA VISUALIZATION####***

***#### Started by creating a categorical-categorical relationship ####***

***####Show the relationship between the month and the department***

***####study book reference page 204***

***##Step 1: created a pivot table to summarize the relationship easier***

*M\_Depart = df\_garments.pivot\_table(values='date', index='month',*

*columns='department', aggfunc='count')*

***## Step2: Created a list for the month to be rearranged accurately.***

*desired\_order = ['January', 'February', 'March']*

***## Step3: Reindex the pivot data frame to reorder the month rows***

*M\_Departm = M\_Depart .reindex(desired\_order)*

***## Step4: Show a graphical representation in a bar chat***

*M\_Departm.plot(kind='bar', figsize=(8, 5))*

*plt.title("Departmental Activity Counts Across Month")*

*plt.xlabel("Month")*

*plt.ylabel("Count")*

*plt.show()*

***#### Secondly created a categorical-numerical relationship ####***

***#### Study book reference page 280***

***#### 1)Relationship between Month and the actual productivity***

***##Step 1: created a pivot table to summarize the relationship easier***

*M\_AP = df\_garments.pivot\_table(values='actual\_productivity',*

*index='month', aggfunc='mean')*

***## Step2: Reindex the pivot data frame to reorder the month rows***

*M\_ap = M\_AP.reindex(desired\_order)*

***## Step3: Show a graphical representation in a line chat***

*M\_ap.plot()*

*plt.title("Analysis of Actual Productivity by Month")*

*plt.xlabel("Month")*

*plt.ylabel("Percentage")*

*plt.show()*

***#### 2 Relationship between Month and the actual productivity and targeted productivity***

***##Step 1: created a pivot table to summarize the relationship easier***

*M\_APT = df\_garments.pivot\_table(values=['actual\_productivity','targeted\_productivity'],*

*index='month', aggfunc='mean')*

***## Step2: Reindex the pivot data frame to reorder the month rows***

*M\_apt = M\_APT.reindex(desired\_order)*

***## Step3: Show a graphical representation in a line chat***

*M\_apt.plot()*

*plt.title("Analysis of Actual and Targeted Productivity by Month")*

*plt.xlabel("Month")*

*plt.ylabel("Percentage")*

*plt.show()*

***#### 3 Relationship between Day and the actual productivity***

***## Step 1: Created a list for the month to be rearranged accurately.***

*Accurate\_day =['Monday','Tuesday','Wednesday','Thursday','Saturday','Sunday']*

***##Step 2: created a pivot table to summarize the relationship easier***

*M\_APD = df\_garments.pivot\_table(values=['actual\_productivity'],*

*index='day', aggfunc='mean')*

***## Step 3: Reindex the pivot data frame to reorder the month rows***

*M\_apd = M\_APD.reindex(Accurate\_day)*

***## Step 4: Show a graphical representation in a line chat***

*plt.title("Analysis of Actual Productivity by day")*

*M\_apd.plot(color='teal')*

*plt.xlabel("Day")*

*plt.ylabel("Percentage")*

*plt.show()*

***#### Box plot for productivity by team***

***Show a graphical representation in a box plot***

*plt.figure(figsize=(10, 6))*

*sns.boxplot(x=df\_garments.team, y=df\_garments.actual\_productivity, color='teal')*

*plt.title('Productivity by Team', fontsize=14)*

*plt.xlabel('Team')*

*plt.ylabel('Productivity Percentage')*

*plt.grid(True)*

***#### Finaly numerical vrs numerical relationship***

***# Scatter plot for Number of Workers vs. SMV***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.no\_of\_workers, y=df\_garments.smv, color='teal')*

*plt.title('Number of Workers vs SMV', fontsize=14)*

*plt.xlabel('Number of Workers')*

*plt.ylabel('SMV')*

*plt.grid(True)*

***# Scatter plot for Target Productivity vs. Actual Productivity***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.actual\_productivity, y=df\_garments.targeted\_productivity, color='teal')*

*plt.title('Target\_productivity vs. Actual Productivity', fontsize=14)*

*plt.xlabel('Actual Productivity')*

*plt.ylabel('Targeted productivity')*

*plt.grid(True)*

***# Scatter plot for Number of Workers vs Overtime***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.no\_of\_workers, y=df\_garments.over\_time, color='teal')*

*plt.title('Number of Workers vs. Over Time', fontsize=14)*

*plt.xlabel('Number of Workers')*

*plt.ylabel('Over Time')*

*plt.grid(True)*

***# Scatter plot for SMV vs Overtime***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.smv, y=df\_garments.over\_time, color='teal')*

*plt.title('SMV vs. Over Time', fontsize=14)*

*plt.xlabel('SMV')*

*plt.ylabel('Over Time')*

*plt.grid(True)*

***# Scatter plot for idle men vs idle time***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.idle\_time, y=df\_garments.idle\_men, color='teal')*

*plt.title('idle time vs. idle men', fontsize=14)*

*plt.xlabel('idle time')*

*plt.ylabel('idle men')*

*plt.grid(True****)***

***# Scatter plot for Actual productivity vs Incentive***

*plt.figure(figsize=(10, 6))*

*sns.scatterplot(x=df\_garments.actual\_productivity, y=df\_garments.incentive, color='teal')*

*plt.title('idle time vs. idle men', fontsize=14)*

*plt.xlabel('idle time')*

*plt.ylabel('idle men')*

*plt.grid(True)*