Exploratory Analysis on the Manufacturing dataset

1. Dealing with IQR using outlier

- Calculate the Interquartile Range (IQR) for all the numerical columns and use the IQR to identify any potential outliers in these data points.

The following fields are numerical for which the potential outliers have been listed by using the inter-quartile range (except Machine ID - which is a number, however, is just an ID rather than providing any numerical information on manufacturing)

```
In [12]: print(outliers_summary)
Machine ID                          0
Units Produced                      0
Defects                           269
Production Time Hours               0
Material Cost Per Unit              0
Labour Cost Per Hour                0
Energy Consumption kWh              0
Operator Count                      0
Maintenance Hours                 539
Down time Hours                   856
Production Volume Cubic Meters    729
Scrap Rate                       3000
Rework Hours                     1342
Quality Checks Failed             962
Average Temperature C               0
Average Humidity Percent            0
```

2. Identify Missing Values Across Key Production Metrics:

Analyse the dataset to identify missing values across all the columns and calculate the total number of missing values for each of these columns. Describe your findings and then impute all the missing values with suitable data points.

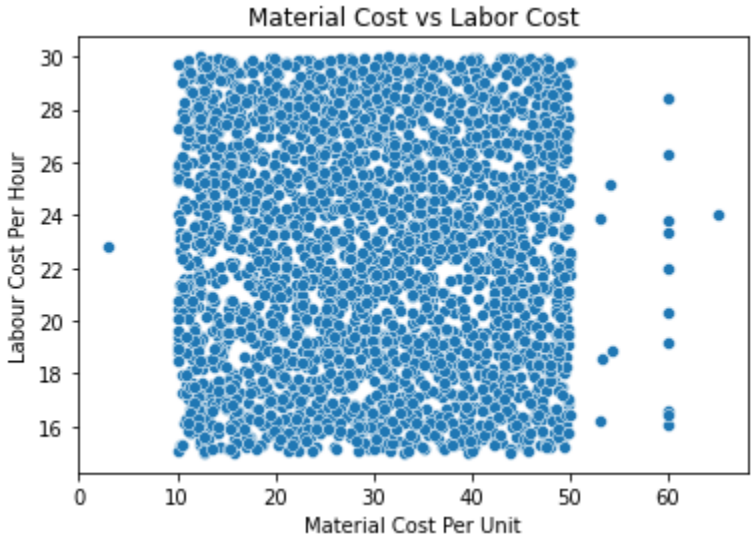
The following columns have some missing values, i.e. Defects, Maintenance Hours, Down time hours and Rework hours. These missing values were filled with the mean for each of those columns

```
In [14]: print(missing_values)
Production ID                      0
Date                               0
Product Type                       0
Machine ID                         0
Shift                              0
Units Produced                     0
Defects                          299
Production Time Hours              0
Material Cost Per Unit             0
Labour Cost Per Hour               0
Energy Consumption kWh             0
Operator Count                     0
Maintenance Hours                300
Down time Hours                  300
Production Volume Cubic Meters     0
Scrap Rate                         0
Rework Hours                     300
Quality Checks Failed              0
Average Temperature C              0
Average Humidity Percent           0
```

3. Relationship Between Costs:

- Is there a pattern between the cost of materials per unit and the hourly labor cost? Determine if higher costs in materials tend to coincide with higher labor costs.

There seems to be no clear relationship between the Labor cost per hour and the Material cost per unit as seen from the scatter plot between the two.



This can also be seen from the very low correlation between the two -

```
In [18]: correlation = df_manf['Material Cost Per Unit'].corr(df_manf['Labour Cost Per Hour'])
    ...: print(f"Correlation Coefficient: {correlation}")
Correlation Coefficient: -3.5483888071472935e-06
```

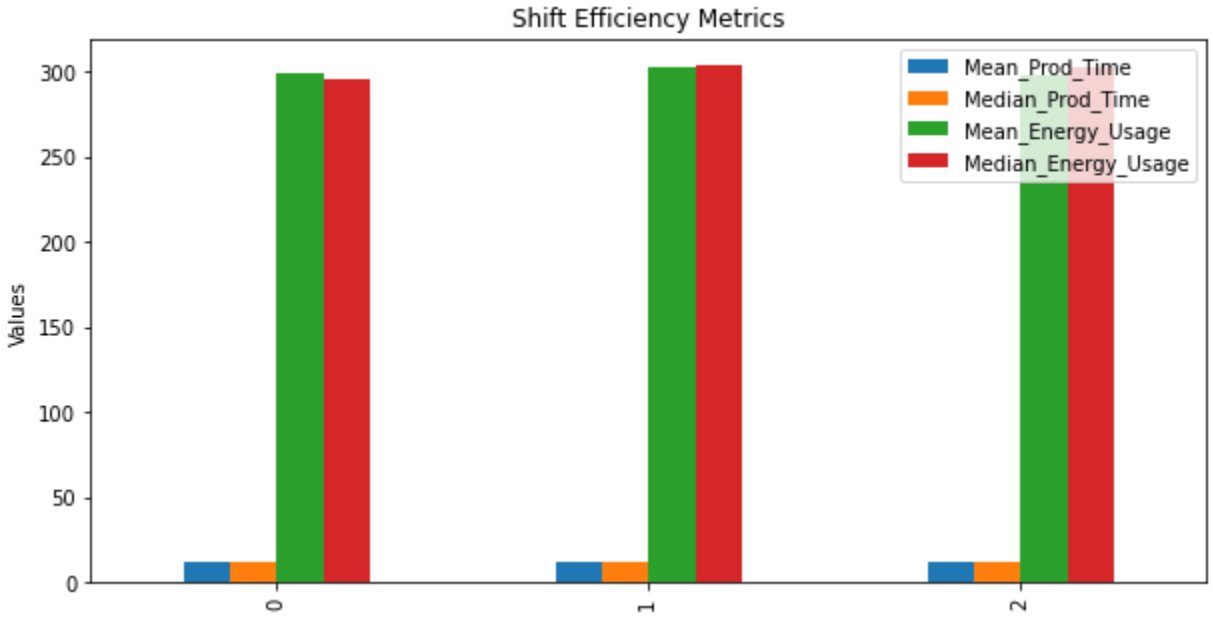
4. Efficiency Across Shifts:

- Do different work shifts (Day, Swing, Night) show differences in how long products take to make or how much energy they use? Compare these shifts to see if one is more efficient or uses less energy.

The Night shift seems to have the lowest production time on an average, while considering the median production time, the Day shift seems to be the lowest. However, there is hardly any difference between the 3 shifts overall and the differences are quite immaterial.

Shift	Mean_Prod_Time	Median_Prod_Time	Mean_Energy_Usage	Median_Energy_Usage
Day	12.5957	12.21	298.904	295.12
Night	12.5242	12.28	302.866	304.2
Swing	12.6137	12.66	297.89	303.07

Similarly, the energy consumption in kWh also shows very minor differences between the 3 shifts, although the day shift shows the lowest energy usage both in terms of mean and median usage. A bar plot below shows each of the efficiency metrics between the 3 different shifts.

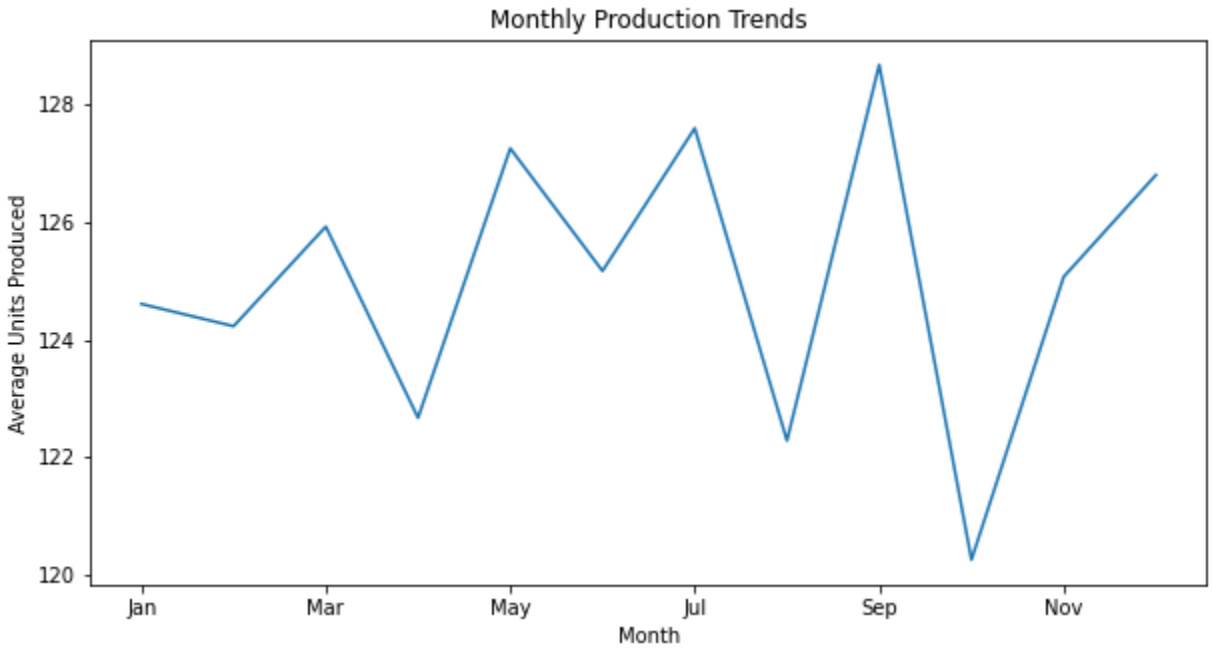


Using one way ANOVA test to check for any significant differences between the different shifts for either Production time or Energy usage, there is no evidence to reject the null hypothesis of no difference across shifts at a 95% confidence level.

5. Monthly Production Trends:

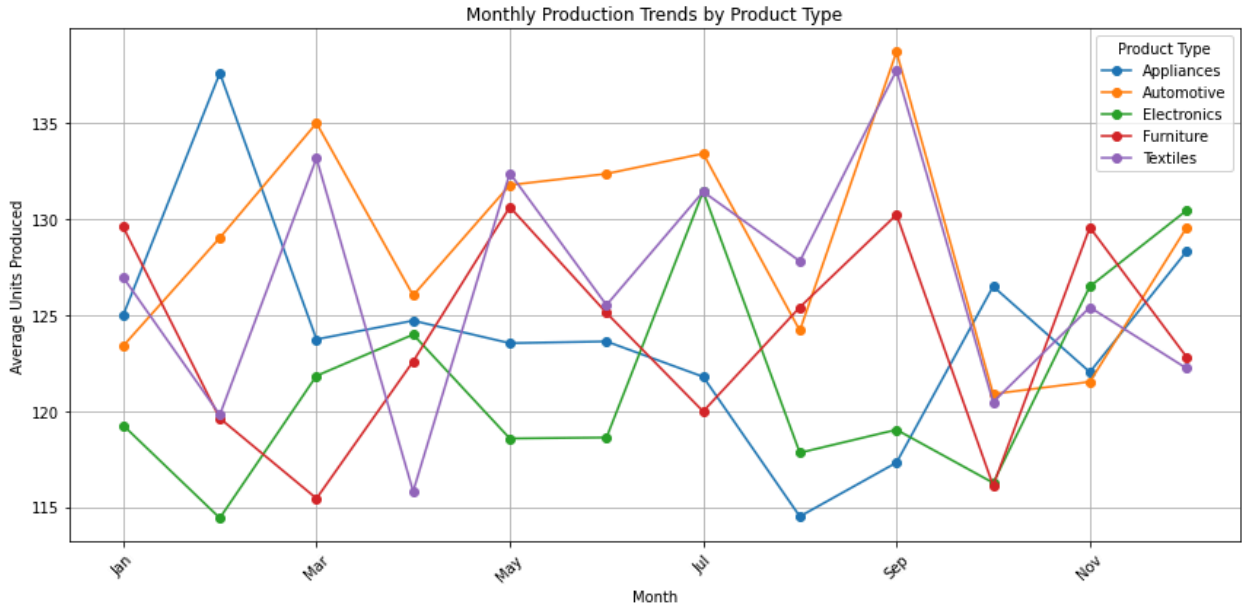
- How does the average number of units produced change from month to month? Look for any patterns, such as times of the year when production increases or decreases significantly.

A simple naive grouping across different products and plotting the trend in production by month of the year shows the highest production in the month of September whereas followed by the lowest in the month of October.



Using groupby for both months and product type and plotting the trend -

Product Type	Appliances	Automotive	Electronics	Furniture	Textiles
Month					
Jan	124.983607	123.384615	119.254237	129.575000	126.955224
Feb	137.581395	129.016393	114.428571	119.618182	119.803279
Mar	123.730159	134.982143	121.830189	115.446809	133.148936
Apr	124.702128	126.018182	123.978261	122.581395	115.816327
May	123.529412	131.759259	118.568627	130.617021	132.377778
Jun	123.620000	132.345455	118.620000	125.104167	125.513514
Jul	121.781818	133.395349	131.420000	119.977273	131.410714
Aug	114.511111	124.188679	117.826087	125.400000	127.796296
Sep	117.314815	138.686275	119.024390	130.212766	137.723404
Oct	126.482143	120.886792	116.260870	116.120000	120.465116
Nov	122.025000	121.520000	126.480000	129.553191	125.396226
Dec	128.300000	129.534483	130.434783	122.816327	122.222222



Analysing the monthly trends by product type, we can see the Appliances have lowest production in August while peaking in February. Similarly, Automotives have lowest production in October, while peaking in September. Electronics have the lowest production in February while it peaks in July. Furnitures have the lowest production in March while peaking in May. And finally, Textiles have the lowest production in April while the production peaks in September. Although Appliances monthly trend could have something to do with end of winter and beginning of winter, for most of the product types we see quite a bit of oscillation in trend in consecutive months and hence not much can be said about specific seasonal trends. Except for automatives, where a peak is generally observed in the summer months when people probably go out and travel more often.

6. Variability in Production by Product Type:

- Which type of product shows the most variation in how much is produced? Measure this using standard deviation to find out which product type's production volume varies the most.

Appliances show the highest variability in terms of production volume per cubic meters. Looking at the variation across product types in number of units produced doesn't show any significant difference and hence one could just look at the production volume instead.

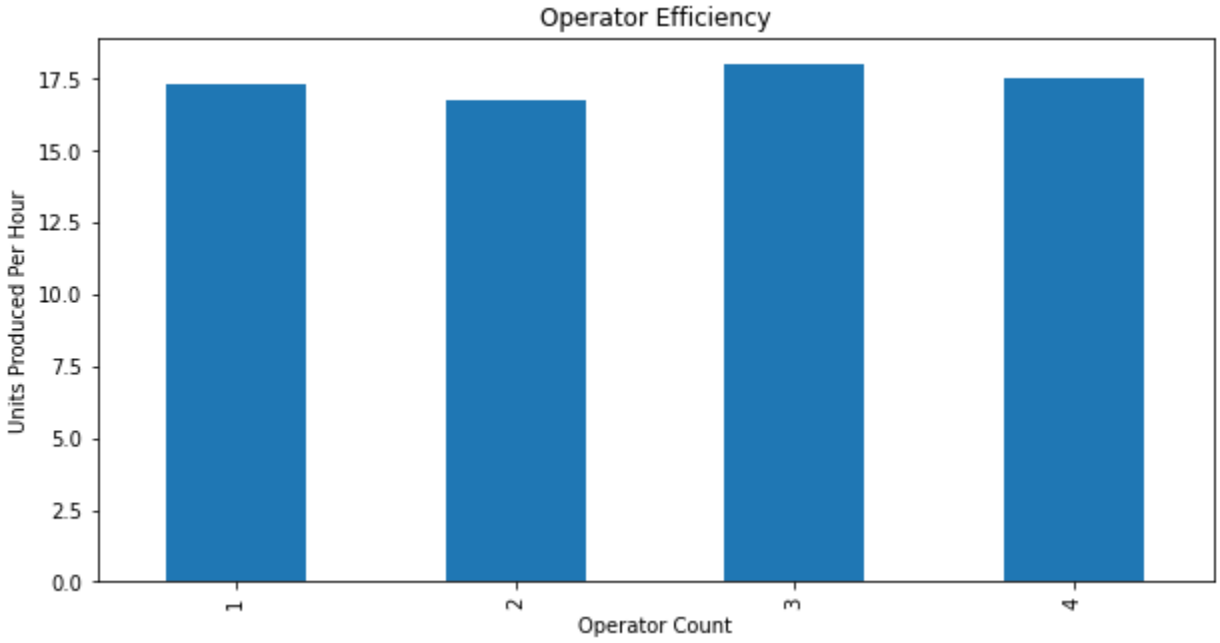
```
Production Variability by Product Type:
Product Type
Appliances      0.602654
Automotive      0.561560
Electronics     0.573125
Furniture       0.579365
Textiles        0.567891
```

7. The Role of Operator Count in Efficiency:

- How does the number of operators affect how many units are produced per hour? Check if having more operators leads to more efficient production.

```
Operator Efficiency:
Operator Count
1    17.319687
2    16.763033
3    18.021761
4    17.499939
```

Having more number of operators doesn't significantly increase the number of units produced per hour as we can see first the number of units produced goes down for 2 operators and is highest for 3 operators while again reduced when it comes to 4 operators.



8. Identifying the Machine with Most Defects:

- Which machine tends to produce the most defects, considering the total units it produces? Calculate the defect rate as defects per 100 units to make comparisons easier.

Calculating the defects rate as defects per 100 units and ranking them from most defective to least shows Machine ID 15 to be producing the highest defect rate, followed by Machine ID 19 and 18.

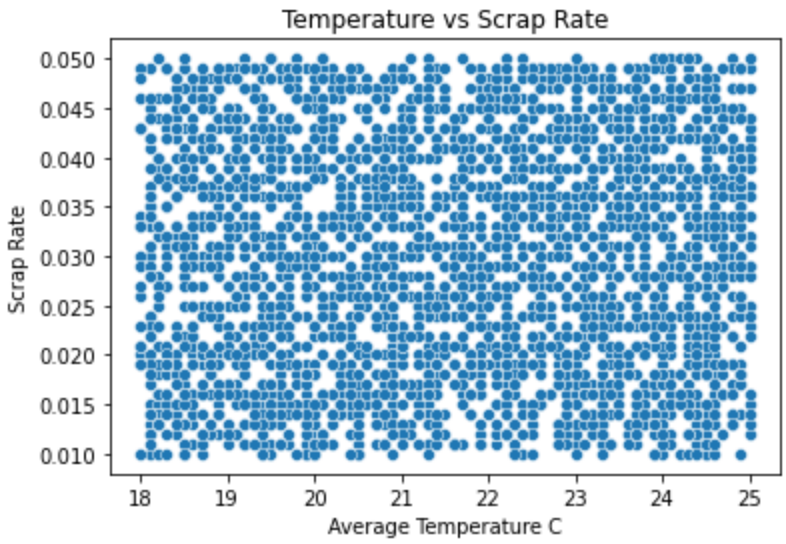
Machine	DefectRate
15	4.84357
19	4.70737
18	4.66923
16	4.659
9	4.41017
8	4.31855
5	4.31576
20	4.28823
4	4.27627
1	4.24111
14	4.21276
10	4.20938
3	4.15882
6	4.12572
11	4.07026
12	4.04775
13	4.04532
17	3.99156
7	3.97934
2	3.79893

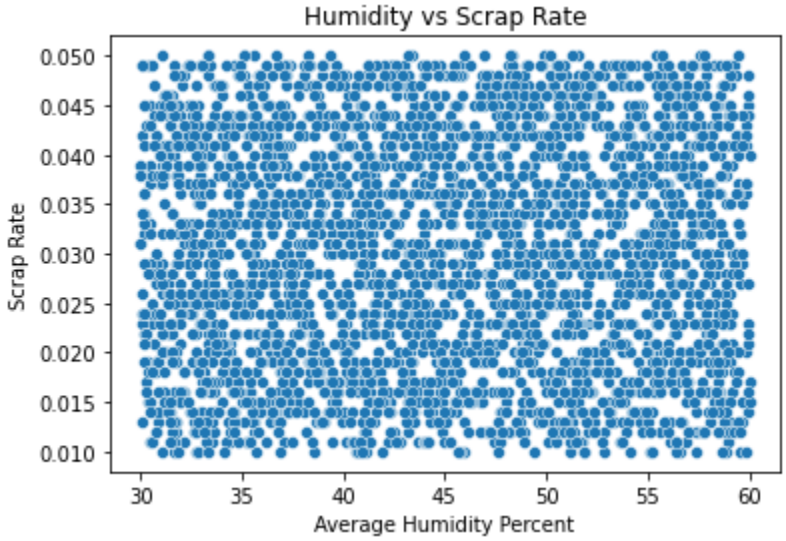
9. How Environment Affects Scrap Rate:

- Do changes in temperature and humidity affect how much scrap (waste) is produced? Analyze the data to see if there's a correlation between environmental conditions and scrap rate.

The correlation between the scrap rate and changes in temperature and humidity seem to be quite low, i.e. approximately 2-3% respectively.

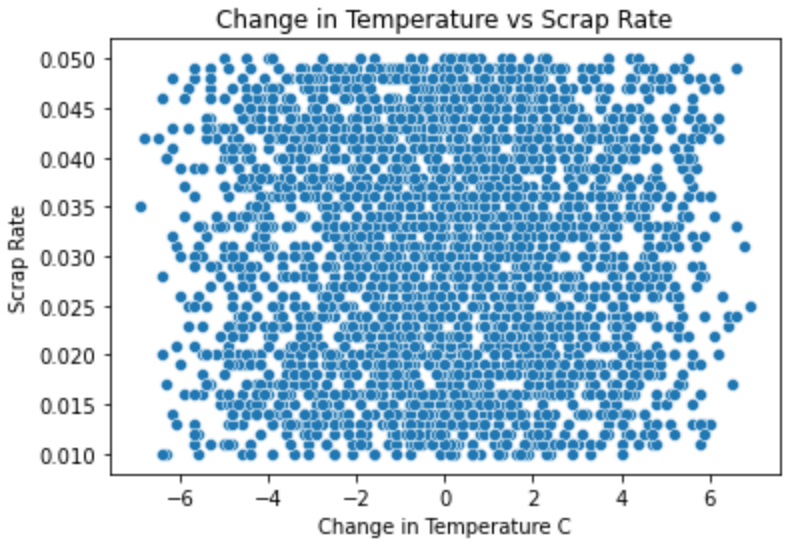
Correlation with Temperature: 0.020510222106531528
Correlation with Humidity: 0.029378168249043633

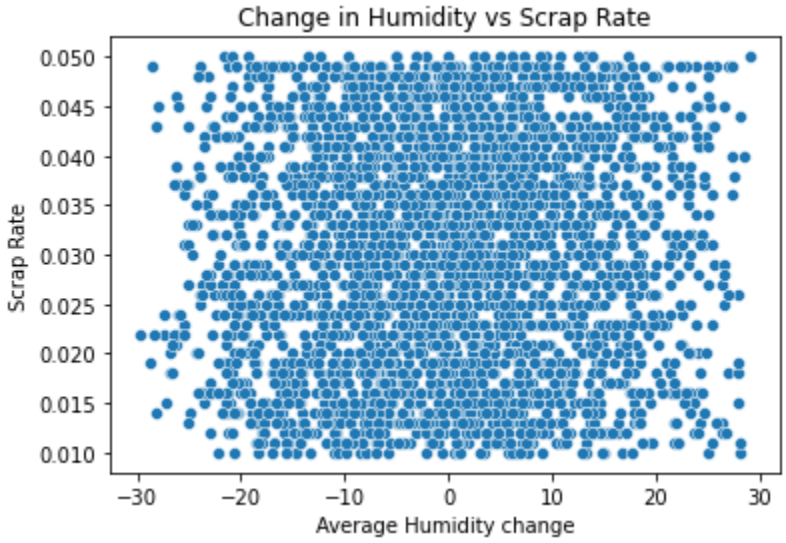




The scatterplots between the scrap rates and the average temperature and humidity seem to show no particular relationship. Since the dataset is in the form of a time series, calculating the daily changes in temperature and humidity to find any relationship with the scrap rate also doesn't add to any further insight of a relationship. Neither does the correlation level with the changes in temperature and humidity.

Correlation with change in Temperature: 0.012786627179609315
Correlation with change in Humidity: 0.041561742315630965

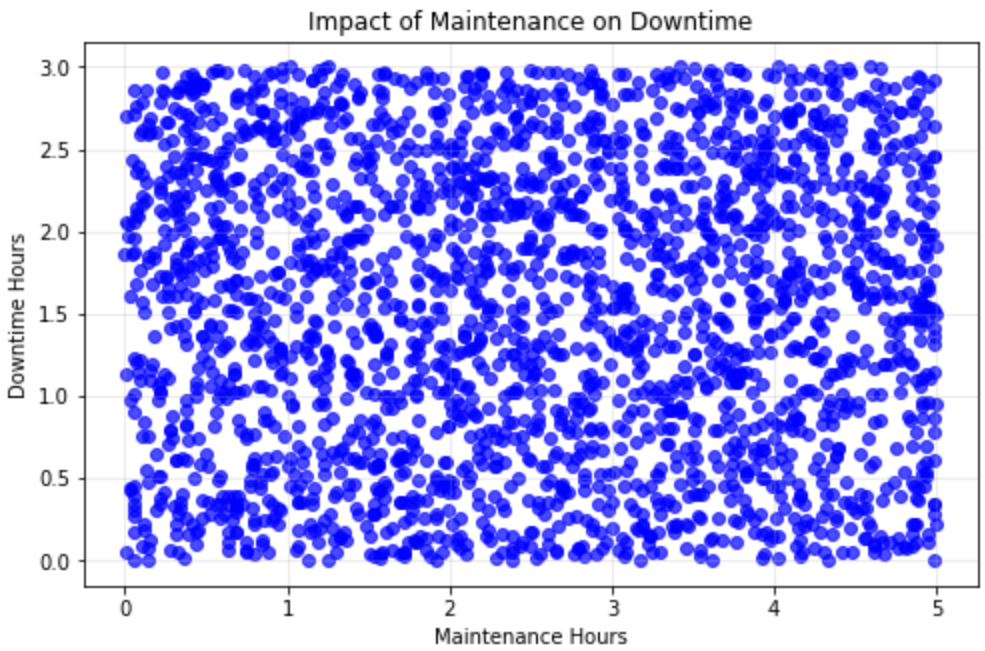




Maintenance Hours vs Downtime Hours

A scatterplot to visualize the relationship between maintenance hours and downtime hours shows no relationship. While a Spearman correlation between the two variables also shows no significant relationship.

```
In [95]: if spearman_p < 0.05:
    ...:     print("The Spearman correlation is statistically significant.")
    ...: else:
    ...:     print("The Spearman correlation is not statistically significant.")
The Spearman correlation is not statistically significant.
```


Impact of Maintenance on Downtime
Downtime Hours
Maintenance Hours
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5