A white robot hand with red and blue stripes

Description automatically generated



**AI/ML**

**LAB MANUAL**

**Disclaimer: The content is curated from online/offline resources and used for educational purpose only**

**Evaluating the Productivity Shift in Remote vs. On-Site Workforces Using Inferential Statistics**

**Objective:**

The objective of this lab is to evaluate productivity differences between remote and on-site workforces using inferential statistics. First, we will develop and test null and alternative hypotheses to examine whether there is a significant difference in productivity between remote and on-site workforces. Next, we will analyze the impact of well-being on productivity using correlation tests to determine if a positive or negative relationship exists between employee well-being and productivity. Finally, we will evaluate the effect of working hours on productivity using linear regression, which will help identify whether the number of hours worked per week significantly influences productivity levels.

**Equipment Required:**

* Computer with Python and Jupyter Notebook installed
* Dataset: [Remote Work Productivity](https://www.kaggle.com/datasets/mrsimple07/remote-work-productivity?resource=download)
* Python Libraries: pandas, numpy, matplotlib, seaborn, scipy, statsmodels

**Prerequisites:**

* Basic understanding of Python programming
* Knowledge of statistics, including hypothesis testing
* Familiarity with pandas for data manipulation
* Understanding of data visualization using matplotlib and seaborn

**Problem Statement:**

Evaluate the productivity shift between remote and on-site workforces using statistical methods. Test whether working hours and well-being scores significantly affect productivity levels.

**Code:**

# Import necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import ttest\_ind, mannwhitneyu, pearsonr

import statsmodels.api as sm

# Load the dataset

file\_path = 'remote\_work\_productivity.csv'

df = pd.read\_csv(file\_path)

# Display the first few rows of the data

df.head()

# Clean missing values

df.dropna(inplace=True)

# Summary statistics

print(df.describe())

# Hypothesis Testing:

# H0: There is no difference in productivity between remote and on-site workforces

# H1: There is a difference in productivity between remote and on-site workforces

remote = df[df['Employment\_Type'] == 'Remote']['Productivity\_Score']

on\_site = df[df['Employment\_Type'] == 'On-Site']['Productivity\_Score']

# Check sample size

print(f"Number of Remote Employees: {remote.size}")

print(f"Number of On-Site Employees: {on\_site.size}")

if remote.size >= 10 and on\_site.size >= 10:

# Perform t-test if sample sizes are sufficient

t\_stat, p\_value = ttest\_ind(remote, on\_site)

print(f'T-statistic: {t\_stat}, P-value: {p\_value}')

else:

if remote.size > 0 and on\_site.size > 0: # Ensure both samples are not empty

# Use Mann-Whitney U test if sample size is small

stat, p\_value = mannwhitneyu(remote, on\_site)

print(f'Mann-Whitney U Statistic: {stat}, P-value: {p\_value}')

else:

print("One or both samples are empty. Test cannot be performed.")

# Visualization

sns.boxplot(x='Employment\_Type', y='Productivity\_Score', data=df)

plt.title('Productivity Comparison: Remote vs On-Site')

plt.show()

# Correlation between Well-Being Score and Productivity Score

corr, p\_value = pearsonr(df['Well\_Being\_Score'], df['Productivity\_Score'])

print(f'Correlation: {corr}, P-value: {p\_value}')

# Linear Regression: Effect of Working Hours on Productivity

X = df['Hours\_Worked\_Per\_Week']

X = sm.add\_constant(X)

Y = df['Productivity\_Score']

model = sm.OLS(Y, X).fit()

print(model.summary())

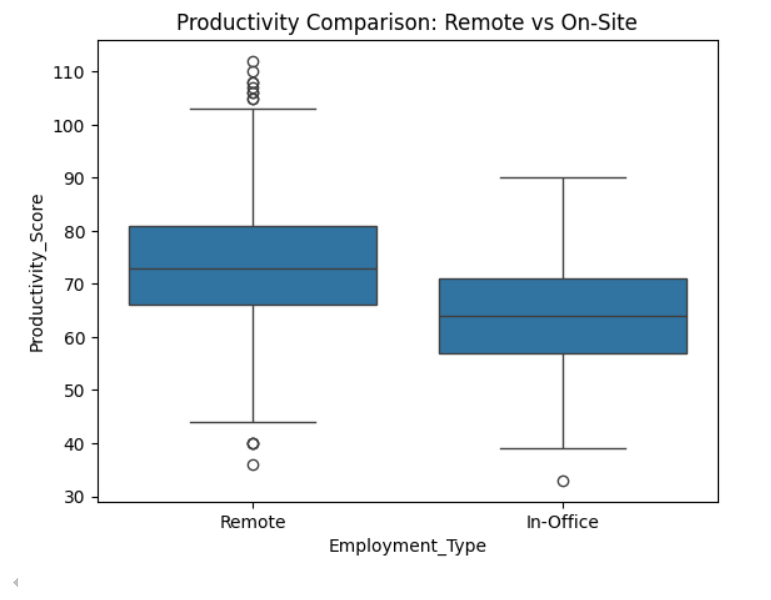
# Plot Regression

sns.regplot(x='Hours\_Worked\_Per\_Week', y='Productivity\_Score', data=df, line\_kws={'color':'red'})

plt.title('Effect of Working Hours on Productivity')

plt.show()

**Output:**



**Conclusion**

If the p-value from the t-test or Mann-Whitney U test is less than 0.05, we reject the null hypothesis, indicating a significant difference in productivity between remote and on-site workforces.

If the p-value from the correlation test is less than 0.05, we conclude that well-being significantly impacts productivity.

If the regression analysis shows a significant p-value for working hours, we conclude that working hours significantly affect productivity.