Coding Temple

Flex Data Analytics Program

Capstone I Project Report

Dataset: Countries Indicators

Submitted by: *Farrukh Sultan*

TABLE OF CONTENTS

Dataset

Data of **231 countries** with 45 characteristics such as debt, electricity consumption, Internet users, etc. from one of the websites suggested in the google classroom. Below is the link:

https://perso.telecom-paristech.fr/eagan/class/igr204/datasets

Introduction

In this report, I delve into a comprehensive dataset encompassing a wide array of socio-economic indicators. This dataset provides a panoramic view of global development, covering variables across infrastructure, trade, labor, health, and more. As part of analytical approach, I have formulated five hypotheses aimed at exploring specific aspects of socio-economic development. These hypotheses span topics such as economic development and health indicators, natural resource management, trade and economic stability, infrastructure and technological development, and labor market dynamics. Through rigorous exploration and analysis, I aim to uncover patterns, correlations, and insights that illuminate the complex interplay of factors shaping the socio-economic landscape of nations worldwide. Our findings will contribute to a deeper understanding of the data and its implications for policy, business, and society at large, potentially paving the way for informed decision-making and impactful interventions.

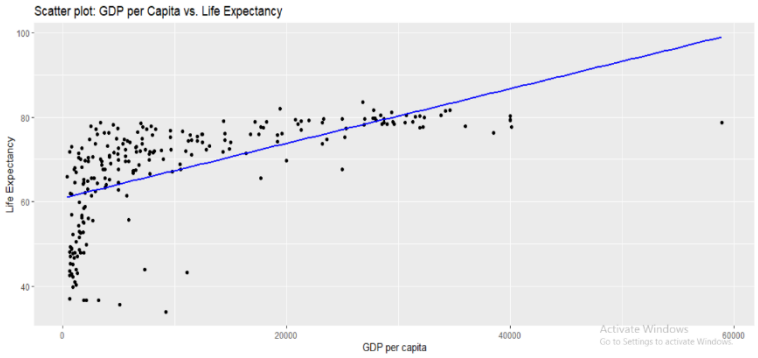
Hypotheses

Language R

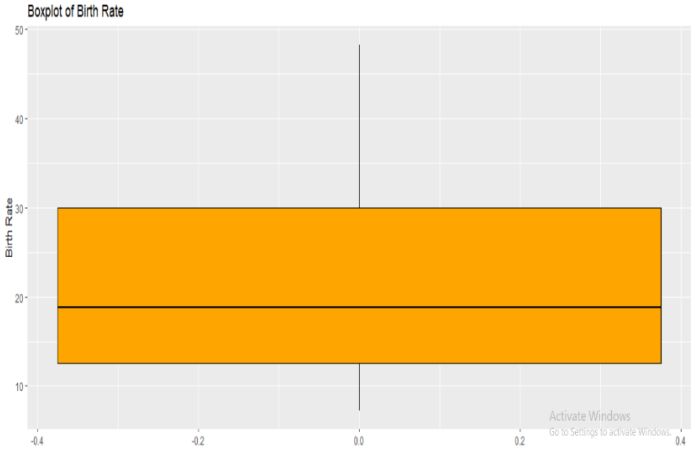
1. **Economic Development and Health Indicators:**
   Null Hypothesis: There is no relationship between a country's GDP per capita and its life expectancy.
   Alternative Hypothesis: Countries with higher GDP per capita have longer life expectancies compared to those with lower GDP per capita.

   Hypothesis Testing:
   A Pearson correlation analysis to assess the relationship between GDP per capita and life expectancy. The results indicate a significant positive correlation between the two variables (Pearson correlation coefficient = 0.6016, $p < 0.001$). Therefore, we reject the null hypothesis and conclude that there is a positive association between a country's GDP per capita and its life expectancy.

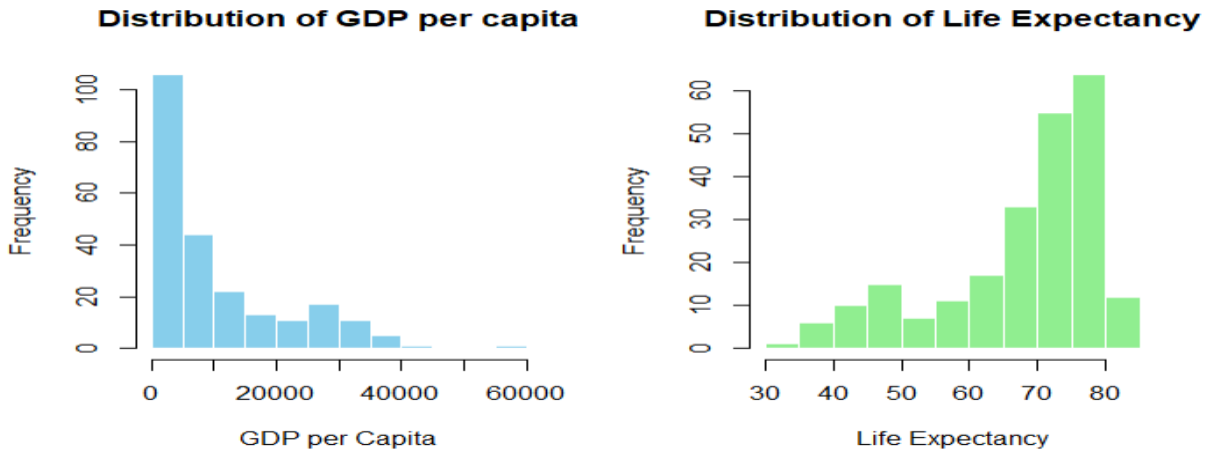
Exploratory Data Analysis:



The scatter plot shows a positive relationship between GDP per capita (x-axis) and life expectancy (y-axis). This is indicated by the general upward trend of the data points, where higher GDP per capita values tend to correspond to higher life expectancy values.



Box plot illustrates the distribution of birth rates across countries, with the y-axis representing the birth rate values. The upper quadrant of the box appears larger than the lower quadrant, suggesting a greater spread of birth rate values in the upper half of the data distribution is from ~19 to 30. The median birth rate, depicted by the middle line within the box is ~18 indicating that half of the countries have birth rates below this value and half have birth rates above it. The upper border line, situated at 30, denotes the upper limit of the upper quartile, while the lower border line, positioned between 15 and 10, represents the lower limit of the lower quartile.



The decreasing trend in the distribution of GDP per capita suggests that a larger proportion of countries have lower GDP per capita values compared to higher values. This indicates that the majority of countries in the dataset may have relatively lower levels of economic prosperity, with fewer countries having higher GDP per capita values. The distribution skewed towards lower GDP per capita values may reflect disparities in economic development among countries.

The increasing trend in the distribution of life expectancy indicates that a larger proportion of countries have higher life expectancy values compared to lower values. This suggests that the majority of countries in the dataset may have relatively higher life expectancies, with fewer countries having lower life expectancy values. The distribution skewed towards higher life expectancy values may reflect improvements in healthcare, nutrition, sanitation, and overall quality of life in many countries over time.

Conclusion:

The findings of this analysis provide robust evidence that economic development, as measured by GDP per capita, is positively correlated with improvements in health outcomes, specifically longer life expectancies. Countries with higher levels of economic prosperity tend to have longer life expectancies, highlighting the importance of economic policies and investments in promoting public health.

2. **Natural Resource Management:**

Null Hypothesis: There is no correlation between a country's oil production and its GDP.

Alternative Hypothesis: Countries with higher levels of oil production tend to have higher GDPs due to revenue from oil exports.

Hypothesis Testing:

Pearson correlation coefficient: 0.962

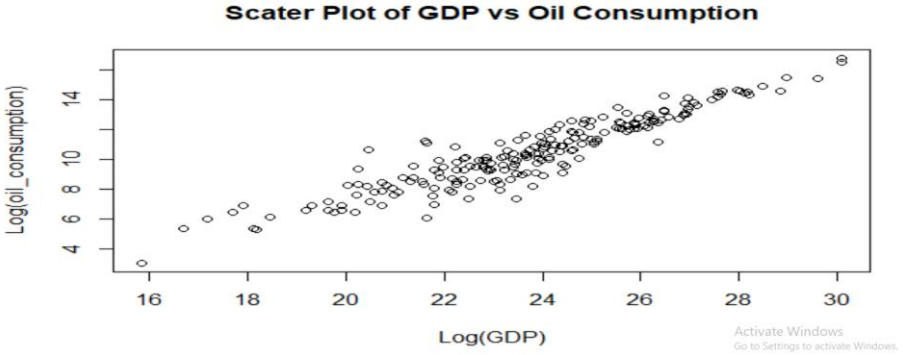
p-value: 0.00000000000000022204

Confidence Interval: The 95% confidence interval for the correlation coefficient ranges from 0.951 to 0.971. This interval suggests that we are highly confident that the true correlation between GDP and oil consumption falls within this range.

So, the alternative hypothesis is true.

Exploratory Data Analysis:

As GDP increases, oil consumption tends to increase as well. This aligns with the alternative hypothesis that countries with higher levels of oil production tend to have higher GDPs due to revenue from oil exports.



Conclusion:

Based on the results of hypothesis testing and visualization, we find compelling evidence to support the alternative hypothesis, which posits that countries with higher levels of oil production tend to have higher GDPs due to revenue from oil exports. The Pearson correlation coefficient of 0.96 ($p < 0.001$) indicates a strong positive correlation between GDP and oil consumption. Additionally, the scatter plot reveals a clear increasing trend from bottom-left to top-right, further supporting this positive relationship. These findings suggest that oil production plays a significant role in driving economic growth, as countries with greater oil production capacity tend to exhibit higher levels of economic activity, as evidenced by their GDP. Therefore, our analysis underscores the importance of natural resource management in shaping a country's economic landscape.

3. **Trade and Economic Stability:**

Null Hypothesis: There is no correlation between a country's current account balance and its GDP growth rate.

Alternative Hypothesis: Countries with a positive current account balance experience higher GDP growth rates compared to those with a negative current account balance.

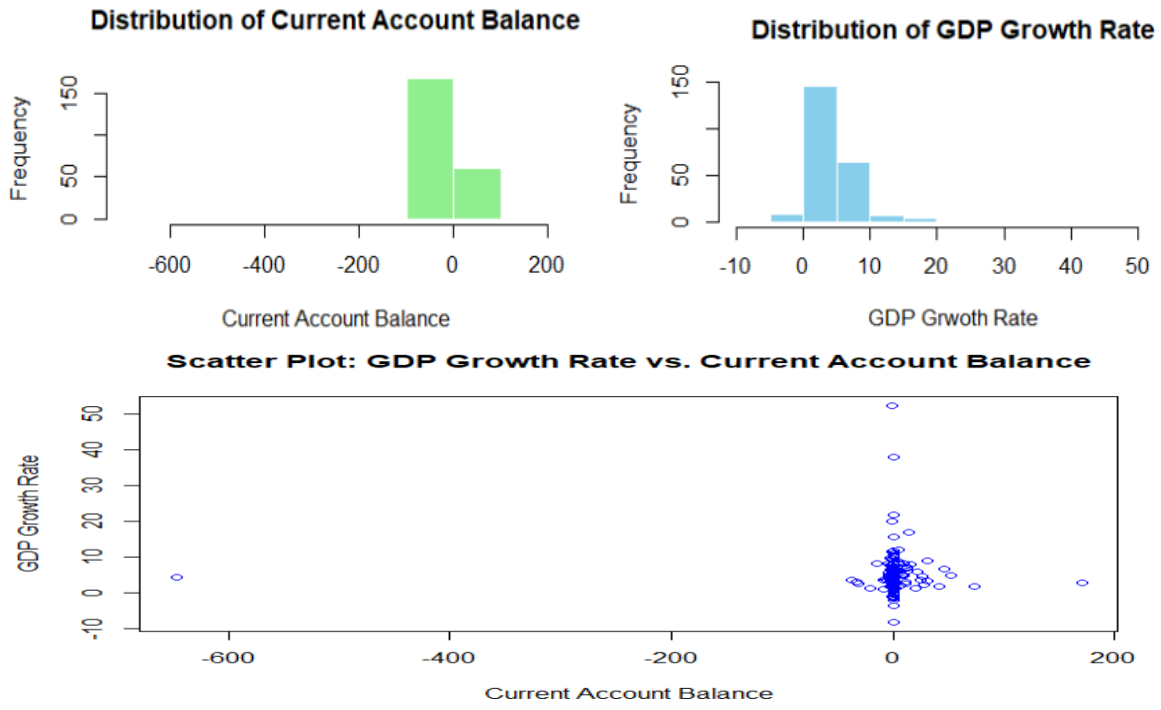
Hypothesis Testing:

Pearson correlation coefficient: 0.0034

p-value: 0.96

Since the p-value exceeds the conventional significance level of 0.05, we fail to reject the null hypothesis. Therefore, there is insufficient support to conclude that countries with a positive current account balance experience higher GDP growth rates compared to those with a negative current account balance.

Exploratory Data Analysis:



Conclusion:

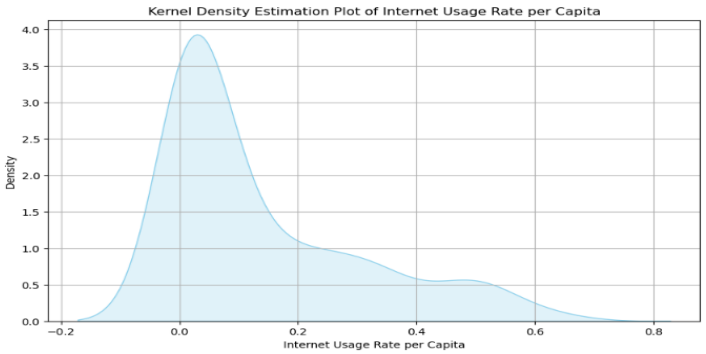
Hypothesis testing and visualizations, including histograms and scatter plot, further supported this finding by demonstrating a lack of discernible patterns or trends between current account balance and GDP growth rate. While these results may suggest that current account balance may not be a significant predictor of GDP growth rate, it's essential to consider other factors that may influence economic stability. Future research could explore additional variables and conduct more nuanced analyses to further elucidate the complex dynamics of trade and economic stability.

Python

4. **Infrastructure and Technological Development:**
   Null Hypothesis: There is no association between a country's internet usage rate and its electricity consumption per capita.
   Alternative Hypothesis: Countries with higher internet usage rates also tend to have higher electricity consumption per capita.

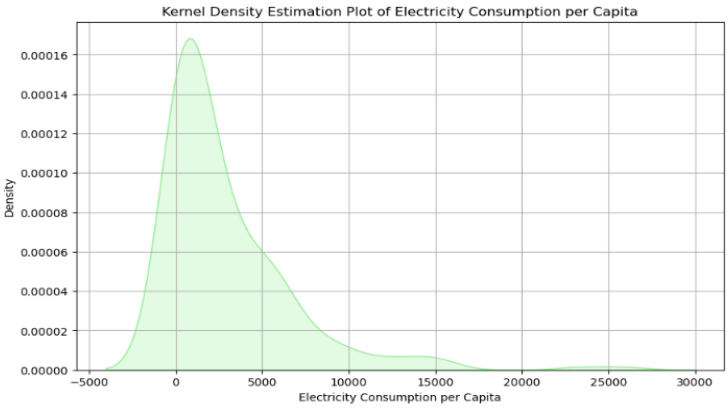
   Hypothesis Testing:
   Pearson correlation coefficient: 0.77
   It indicates a strong positive linear relationship between internet usage and electricity consumption. This means that as internet usage increases, electricity consumption also tends to increase.

   Exploratory Data Analysis:
   Kernel density estimation (KDE) plot for internet use rate:



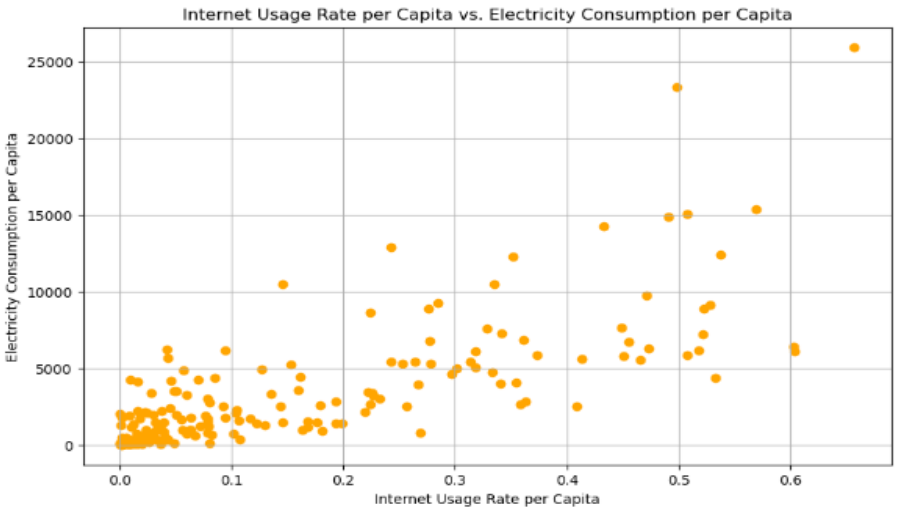
Distribution of internet usage rate per capita is skewed towards the right, it suggests that there is a higher concentration of countries with lower internet usage rates, while a few countries have exceptionally high rates. This could indicate disparities in internet access and adoption among countries, with some having limited access or lower adoption rates compared to others. The skewness towards the right may also imply that the majority of countries fall within a certain range of lower usage rates, while a smaller proportion of countries exhibit much higher rates. Further analysis could explore the factors contributing to these disparities and their implications for socioeconomic development and digital inclusion efforts.

Kernel density estimation (KDE) plot for electricity consumption per capita:



Distribution of electricity consumption per capita is skewed towards the right, it suggests that there is a higher concentration of countries with lower electricity consumption per capita, while a few countries have significantly higher consumption levels. This could indicate disparities in access to electricity and energy consumption patterns among countries. The skewness towards the right may imply that the majority of countries have relatively lower electricity consumption per capita, possibly due to factors such as limited infrastructure, energy efficiency measures, or economic constraints. Conversely, a smaller proportion of countries may exhibit much higher consumption levels, potentially reflecting higher levels of industrialization, urbanization, or energy-intensive activities. Further analysis could investigate the underlying factors driving these disparities and their implications for energy policies and sustainable development goals.

Scatter plot for internet usage rate per capita vs. electricity consumption per capita:



Scatter plot suggests a positive correlation between these two variables. In other words, countries with higher internet usage rates tend to have higher electricity consumption per capita, and vice versa. This pattern indicates that there may be a relationship between technological development, as indicated by internet usage, and energy consumption, represented by electricity consumption per capita. The slope of the scatter plot indicates the strength of the correlation: a steeper slope suggests a stronger positive correlation, while a shallower slope indicates a weaker correlation.

Conclusion:

Based on the statistical testing, we observed that this relationship yielding a correlation coefficient of 0.77 indicating a strong positive correlation. Exploratory data analysis further confirmed that internet usage rate and electricity consumption per capita exhibit right-skewed distributions, with the scatter plot indicating a strong positive linear relationship between the two variables. Therefore, we reject the null hypothesis and accept the alternative hypothesis, suggesting that countries with higher internet usage rates tend to have higher electricity consumption per capita. These findings underscore the importance of technological development and infrastructure investment in fostering economic growth and societal advancement, emphasizing the interplay between internet access and electricity consumption in modern societies.

5. **Labor Market Dynamics:**

Null Hypothesis: There is no relationship between a country's labor force participation rate and its unemployment rate.

Alternative Hypothesis: Countries with higher labor force participation rates experience lower unemployment rate.

Statistical Testing:

Correlation coefficient: -0.2448

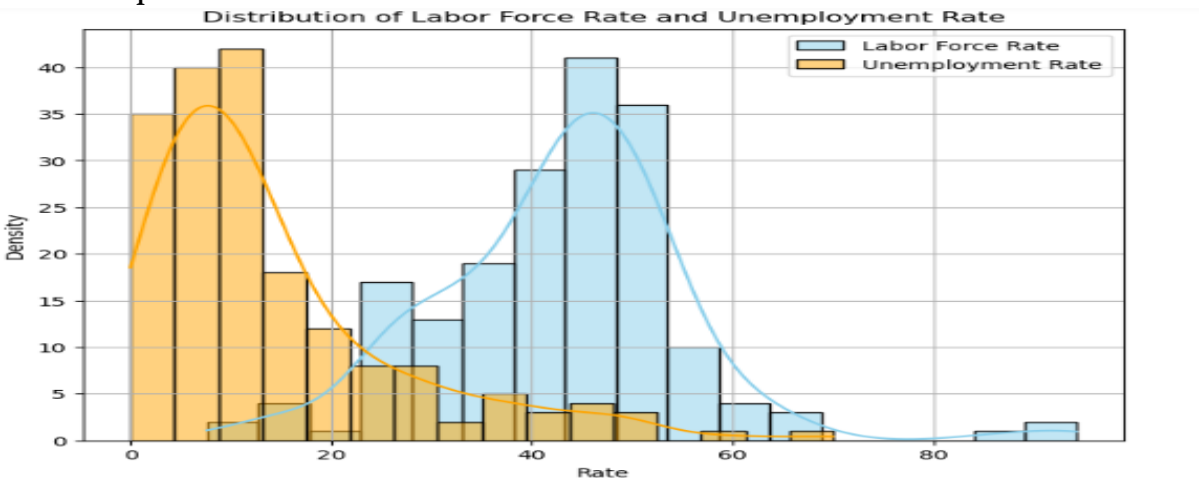
p-value: 0.0009

The correlation coefficient between the labor force participation rate and the unemployment rate is approximately -0.245, indicating a weak negative correlation. This means that as the labor force participation rate increases, the unemployment rate tends to decrease slightly, and vice versa.

The p-value associated with this correlation coefficient is approximately 0.001, which is less than the conventional significance level of 0.05. This suggests that the correlation observed in the sample data is statistically significant. Therefore, we reject the null hypothesis that there is no relationship between the labor force participation rate and the unemployment rate and conclude that there is indeed a statistically significant negative correlation between these two variables in the population.

Exploratory Data Analysis:

Distribution plot:



A right-skewed distribution of the unemployment rate indicates that there are more countries with lower unemployment rates and relatively fewer countries with higher unemployment rates. This could suggest that many countries have relatively low levels of unemployment, but there are a few countries where unemployment is significantly higher, contributing to the longer tail on the right side of the distribution.

Conversely, a left-skewed distribution of the labor force rate suggests that there are more countries with higher labor force participation rates and relatively fewer countries with lower participation rates. This indicates that a larger proportion of countries have higher rates of labor force participation, with fewer outliers on the left side of the distribution.

Conclusion:

The results of both the statistical testing and visualizations provide evidence in support of the alternative hypothesis. Countries with higher labor force participation rates tend to have lower unemployment rates, highlighting the importance of labor force participation in mitigating unemployment challenges. These findings underscore the need for policies and initiatives aimed at promoting and sustaining higher levels of labor force participation to address unemployment and foster greater economic stability and prosperity.

Statistical Glossary

Correlation Coefficient:

Positive Correlation (r > 0): When the correlation coefficient is positive, it indicates a direct or positive relationship between the two variables. This means that as one variable increases, the other variable tends to increase as well.

Negative Correlation (r < 0): When the correlation coefficient is negative, it indicates an inverse or negative relationship between the two variables. This means that as one variable increases, the other variable tends to decrease, and vice versa.

Zero Correlation (r = 0): When the correlation coefficient is close to zero or exactly zero, it indicates no linear relationship between the two variables. However, it's essential to note that there could still be a non-linear relationship or other types of relationships between the variables.

Perfect Positive Correlation (r = 1): When the correlation coefficient is equal to 1, it indicates a perfect positive correlation. This means that the relationship between the two variables is linear, and all data points lie exactly on a straight line with a positive slope.

Perfect Negative Correlation (r = -1): When the correlation coefficient is equal to -1, it indicates a perfect negative correlation. This means that the relationship between the two variables is linear, and all data points lie exactly on a straight line with a negative slope.

Probability Value (p-value):

The p-value is a measure that helps determine the significance of the results obtained from a statistical test. It represents the probability of observing the data or more extreme results under the assumption that the null hypothesis is true. Here are the possible outcomes of the p-value:

p-value < 0.05 (or any chosen significance level): If the p-value is less than the chosen significance level (commonly set at 0.05), it indicates that the observed results are statistically significant. In other words, there is strong evidence against the null hypothesis, and it is rejected in favor of the alternative hypothesis. This suggests that the relationship observed in the sample data is unlikely to have occurred by random chance alone.

p-value > 0.05 (or any chosen significance level): If the p-value is greater than the chosen significance level, it suggests that the observed results are not statistically significant. In this case, there is insufficient evidence to reject the null hypothesis. It means that the observed relationship could plausibly occur by random chance, and any apparent effect may not be meaningful or reliable.

p-value = 0.05 (or the chosen significance level): When the p-value is exactly equal to the significance level, the decision to reject or fail to reject the null hypothesis depends on the researcher's chosen threshold for significance. Some researchers may choose to reject the null hypothesis in this case, while others may consider the result inconclusive and require further investigation.

p-value very close to 0: A very small p-value indicates strong evidence against the null hypothesis. The smaller the p-value, the stronger the evidence. It suggests that the observed results are highly unlikely to occur if the null hypothesis is true.

p-value very close to 1: A very large p-value suggests weak evidence against the null hypothesis. The larger the p-value, the weaker the evidence. It indicates that the observed results are plausible even if the null hypothesis is true.

Significance Level (α):

It is the threshold used to determine the statistical significance of the test. Commonly used significance levels include 0.05 (5%) and 0.01 (1%). If the p-value is less than or equal to the significance level, then the null hypothesis is rejected, indicating that the results are statistically significant.

Confidence Level (1 - α):

It is the complement of the significance level. It represents the degree of confidence or certainty associated with a confidence interval. A 95% confidence level corresponds to a significance level of 0.05. It implies that if we were to repeat the experiment multiple times, approximately 95% of the time, the confidence interval would contain the true population parameter.

End of Report – Grateful for your time!