**Introduction**

To keep track with billionaires, our subject matter, Forbes Magazines, articles, videos, or seminars is a best place to start. The Forbes keeps a list of world billionaires called the Forbes List.

The Forbes list is an annual ranking of the world's billionaires, compiled and published by Forbes magazine. The list ranks billionaires based on their net worth, as estimated by Forbes, and includes information about their age, source of wealth, and country of citizenship. The Forbes list is widely regarded as one of the most comprehensive and influential rankings of billionaires worldwide.

The Forbes list has been published annually since 1987, and it includes both male and female billionaires from a variety of countries and industries. The list is widely followed by media outlets and the public, and it is often used as a reference for tracking the wealth and success of the world's wealthiest individuals.

According to a Forbes article written in 2016 by Chase Peterson-Withorn, the top 5 industry that produce the most Forbes billionaires worldwide are:

1. Finance and Investments: 267 billionaires (15% of worldwide total)

2. Fashion and Retail: 221 billionaires (12% of worldwide total)

3. Real Estate: 163 billionaires (9% of worldwide total)

4. Technology: 159 billionaires (9% of worldwide total)

5. Manufacturing: 157 billionaires (9% of worldwide total)

According to a report by Forbes contributor, James O’Donnell, as of April 2022 there has been a significant change in the industries producing billionaires and the percentage of total worldwide.

1. Finance & Investments: 393 billionaires (15% of worldwide total)
2. Manufacturing: 337 billionaires (13% of worldwide total)
3. Technology: 332 billionaires (12% of worldwide total)
4. Fashion & Retail: 250 billionaires (9% of worldwide total)
5. Healthcare: 217 billionaires (8% of worldwide total)

In the past 6 years, the Finance & Investments remains as the top industries for producing billionaires, increasing from 267 billionaires to 393 billionaires. There is also change in positions in the case of Fashion and Retail dropping from 2nd to 4th (12% to 9% of worldwide total) and Technology increasing from 4th to 3rd (9% to 12% worldwide total). The Real Estate industry in 6 years is no longer among the top 5 giving rise to the Healthcare industry with 8% of worldwide total.

**Datasets**

The dataset available for this analysis, ‘billionaires.csv’, contains 2614 rows and 22 columns. The columns are:

* Name - name of the billionaire
* Rank – rank on the list in the particular year
* Year – year of the listing
* Company.founded – year the company associated was founded if a founder else zero
* company.name – name of billionaires associated company
* company.relationship – position in the given company
* company.sector – sector where the company function
* company.type – state of the company (e.g. new, acquired, privatized, etc.)
* demographics.age – age of the billionaire
* demographics.gender – gender of the billionaire
* location.citizenship – the billionaire’s citizenship
* location.country code – the billionaire’s country code (e.g. United State of America -USA)
* location.gdp – Gross domestic Product of the billionaire’s country
* location.region – country region
* wealth.type – source of wealth (e.g., founder, inherited, executive, etc.)
* wealth.worth in billions – wealth measured in billions
* wealth.how.category – sectors of wealth
* wealth.how.from emerging – all values are True(redundant and will be dropped)
* wealth.how.industry – Industry of the billionaire’s associated company
* wealth.how.inherited – source of inheritance if wealth was inherited (e.g., not inherited, father, 3rd generation, etc.)
* wealth.how.was founder – also all True (redundant and will be dropped)
* wealth.how.was political – also all True (redundant and will be dropped)

The dataset is range between the year 1996 and 2014. The dataset contains 2077 billionaires from 1577 distinct companies. The dataset requires some cleaning to remove some inconsistency and deal with missing values.

Some of the questions that will be answered with the analysis are:

* What are the top 10 countries with the highest number of billionaires?
* What industries/sectors are most successful?
* What are the main industries with the highest number of women billionaires?
* What age range represents the highest and lowest number of billionaires?
* What is the average net worth in each industry?
* The influence of industries and inheritance on wealth

The analysis will also adopt statistical approach to prove these null hypotheses:

* There is no significant difference in the gender distribution of billionaires across different industries.
* There is no relationship between a billionaires net worth and their country.
* There is no significant difference in the distribution of billionaires' net worth between those who inherited their wealth and those who started their own businesses or acquired their wealth through other means.

**Approaches And Technologies for developing big data applications**

In developing a big data application various approaches and technologies are used. Some of this approach may be based on how the data is processed. Some processing approaches are batch, stream and real-time processing.

**Approaches:**

A more detailed explanation of the three approaches for developing big data applications:

* Batch processing: This approach involves processing data in large batches, typically using technologies like Hadoop or Spark. Batch processing is suitable for applications that need to process very large datasets, but do not require real-time results. For example, an application that processes data from social media platforms to identify trends and patterns might use batch processing.
* Stream processing: This approach involves processing data as it is generated in real-time, using technologies like Apache Flink or Apache Storm. Stream processing is suitable for applications that need to process high volumes of data in real-time and provide results within a short time frame. For example, an application that processes data from IoT devices to identify anomalies or patterns might use stream processing.
* Real-time processing: This approach involves processing data as it is generated in real-time and providing immediate results, using technologies like Apache Samza. Real-time processing is suitable for applications that need to process high volumes of data in real-time and provide results within a very short time frame. For example, an application that processes data from financial transactions to identify fraudulent activity might use real-time processing.

**Technologies:**

* Relational databases: These are traditional databases that store data in structured tables and support SQL queries. Relational databases are suitable for storing and managing structured data, and they are widely used in a variety of applications. Examples of relational databases include MySQL and Oracle.
* NoSQL databases: These databases store data in a variety of formats, including key-value pairs, documents, and graph data, and they support a wide range of query languages. NoSQL databases are suitable for storing and managing unstructured or semi-structured data, and they are often used in applications that need to handle high volumes of data or need to scale horizontally. Examples of NoSQL databases include MongoDB and Cassandra.
* Distributed file systems: These systems store and manage large amounts of data across a network of computers. Distributed file systems are suitable for storing and managing very large datasets, and they are often used in conjunction with batch processing frameworks like Hadoop. Examples of distributed file systems include HDFS and Google Cloud Storage.
* Batch processing frameworks: These frameworks provide tools and libraries for processing large datasets in batch mode. Batch processing frameworks are suitable for applications that need to process very large datasets, but do not require real-time results. Examples of batch processing frameworks include Python, Hadoop and Spark.
* Stream processing frameworks: These frameworks provide tools and libraries for processing data streams in real-time. Stream processing frameworks are suitable for applications that need to process high volumes of data in real-time and provide results within a short time frame. Examples of stream processing frameworks include Apache Flink and Apache Storm.
* Real-time processing frameworks: These frameworks provide tools and libraries for processing data streams in real-time and providing immediate results. Real-time processing frameworks are suitable for applications that need to process high volumes of data in real-time and provide results within a very short time frame. Examples of real-time processing frameworks include Apache Samza.
* Data visualization tools: These tools allow data to be presented in a clear and visually appealing way. Data visualization tools are often used to present the results of big data analyses in a way that is easy to understand and interpret. Examples of data visualization tools include Tableau, Qlik, D3.js, and Matplotlib.
* Data governance technologies: These technologies are used to ensure that big data applications are developed and used in a way that is compliant with relevant laws and regulations, and that data privacy is protected. Examples of data governance technologies include data masking, data classification, and data access controls.

There are many examples of big data applications, as big data can be used to solve a wide range of problems across many different industries. Here are a few examples:

* Healthcare: Big data can be used to improve healthcare outcomes by analysing patient data to identify trends and patterns that can inform diagnosis and treatment decisions. For example, a big data application might be used to identify risk factors for a particular disease, or to predict the likelihood of a patient developing a certain condition.
* Finance: Big data can be used to improve financial decision-making by analysing data on market trends, customer behaviour, and risk management. For example, a big data application might be used to identify fraudulent financial transactions, or to forecast market trends.
* Retail: Big data can be used to improve customer experience and increase sales by analysing data on customer behaviour, preferences, and purchasing history. For example, a big data application might be used to personalize product recommendations or to optimize pricing strategies.
* Manufacturing: Big data can be used to improve efficiency and reduce costs in manufacturing by analysing data on production processes, equipment performance, and supply chain management. For example, a big data application might be used to predict equipment failures or to optimize inventory management.
* Transportation: Big data can be used to improve transportation efficiency and safety by analysing data on traffic patterns, vehicle performance, and driver behaviour.

**Methodological Approach**

For this application Python programming language will be used for the big data analytics. It is assumed to be a batch processing approach because the data is already extracted, billionaires.csv. The dataset is stored in a relational database since it is a table, and in a comma-separated-value format.

For the data cleaning and formatting, to remove missing or incorrect values and transform data types. The data cleaning will be done using Python.

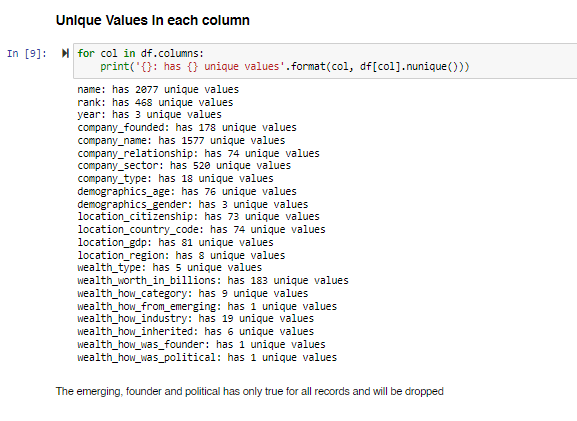
Also, Python will be used to process and analyse the data while other Python libraries will be used for the statistical testing to accept or reject the hypothesis.

Matplotlib (Seaborn) a visualization library in Python will be used to present the results of the analyses in a clear and visually appealing way.

**Data Preprocessing**

First the columns name was processed by replacing all the space and ‘.’, with a ‘\_’ for ease and following Python popular naming convention.

The dataset contains 22 columns and 3 of the columns were dropped because they are redundant, contain only True Boolean. The name of the columns are, *'wealth\_how\_from\_emerging', 'wealth\_how\_was\_founder' and ‘wealth\_how\_was\_political'.*



**Industry Column**

The ‘wealth\_how\_industry’ column, containing the billionaire’s industry, was processed using a popular approach by the Forbes. Some of the industries were merged together after carefully checking other columns and the internet for the subcategories. After merging the 19 different categories was reduced to 10.

Text, letter

Description automatically generated

Text

Description automatically generated

**Age Column**

The Billionaires’ age, ‘Demographics\_age’, contains some outliers and some age were less than 0. Further analysis shows that the age that are 0 are for companies with more than 1 founders or a family business and some in other special cases.

Other outliers were corrected using sources like Wikipedia. The figure shows the age before and after processing. The two outliers remain 0 and the youngest billionaire at age 12.

Chart

Description automatically generatedChart, box and whisker chart

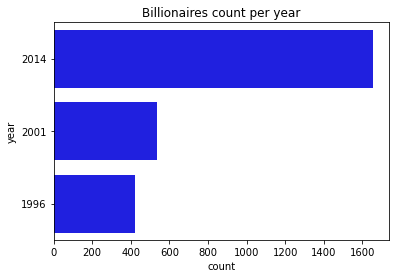
Description automatically generated

**Company Type**

The ‘company\_type’ contain some misspelt categories and it was cleaned by replacing all the errors with the correct categories. It contain 9 categories after the cleaning which are *'new', 'privatized', 'acquired', 'subsidiary', 'new division', 'merger', 'franchise', 'state owned enterprise', and 'joint venture'.*

**General Analysis**

* The data contain 3 distinct years; 1996, 2001 and 2014. It also shows increase in the number of billions over the 3 years with an exponential increase between 2001 and 2014.



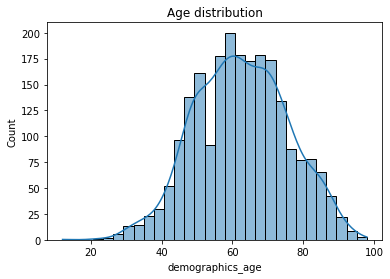
* The dataset has 6 categories of how the wealth was created or inherited and they are: not inherited, father, spouse/widow, 3rd generation, 4th generation, 5th generation or longer. Majority of the wealth are not inherited. And most inherited are from the father, followed by the 3rd generation. The least is from the 5th generation or longer.

Chart, bar chart

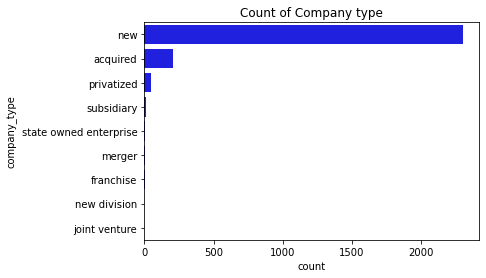
Description automatically generated

* The age of the billionaires ranges from 12 to 100. The youngest billionaire is Albert von Thurn und Taxis, who inherited it at age 12 in Germany. And the oldest is 98 years from USA.

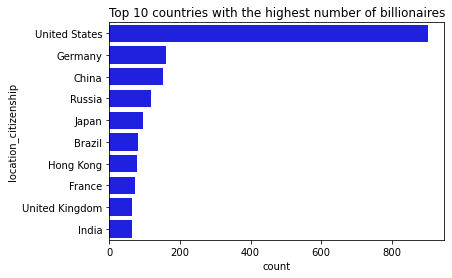
The age distribution is slightly left skewed. The highest count of age is at 60 years and the mean is 53years.

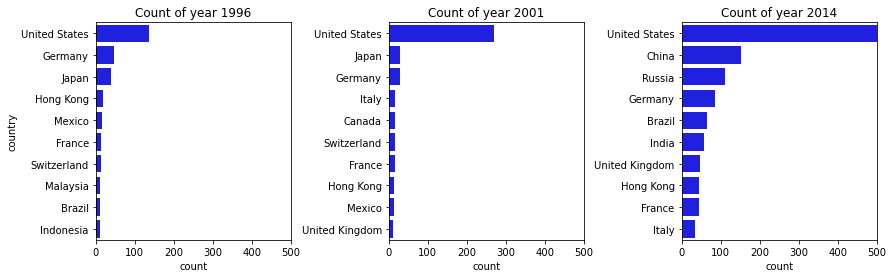


* 89% of the companies of the billionaires are new, others were acquired and very few are privatized companies.

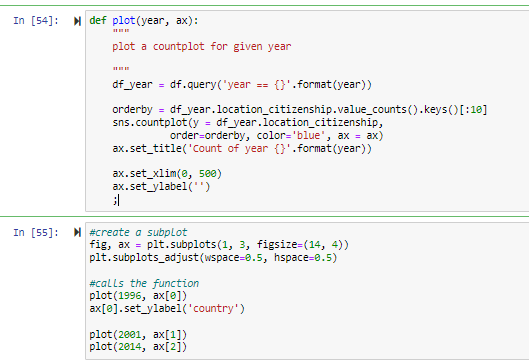


* The dataset contains billionaires from 73 different countries from America, Europe, Asia and Africa. United States have the highest number of billionaires over the 3 years represented in the data. The top 10 countries with the highest number of billionaires as shown in the chart below are, United States, Germany, China, Russia, Japan, Brazil, Hong Kong, France, United Kingdom, and India

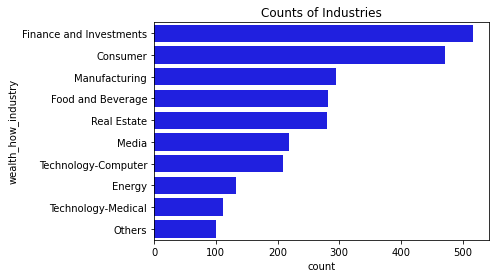


* Over the 3 years there were various displacement in the positioning of the top 10 countries with highest number of billionaires. The United States takes the lead all through while countries like Germany, Japan, Mexico and France have been dropping over the years and others like China, Russia, Brazil and United Kingdom increased as shown below: 

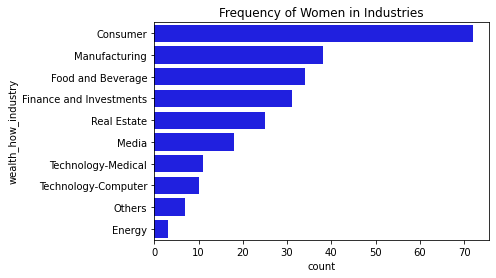
Below is the snapshot of the code used to achieve the chart above:



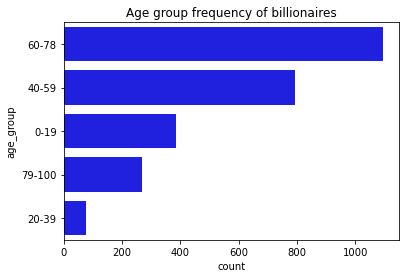
* After the cleaning the industry in the dataset was reduced to 10. The Finance and Investments industry has the highest number of billionaires and so the most successful industries. The chart below shows in descending order the number of billionaires by industries:

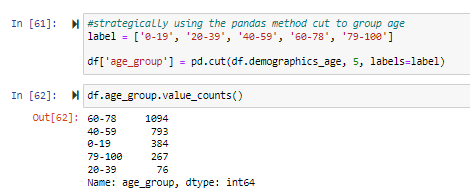


* The contain 249 women and 72 are in the Consumer industry. The Consumer, Manufacturing, Food and Beverage and Finance and Investment represent the main industries with the highest number of women billionaires. The least industry is Energy with just 3 women represented.

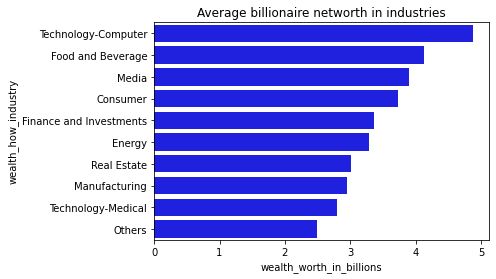


* The age range of 60 – 78 years and 20 – 39 represents the highest and lowest number of billionaires, respectively. General the age lesser than 40 years has the lowest number of billionaires.

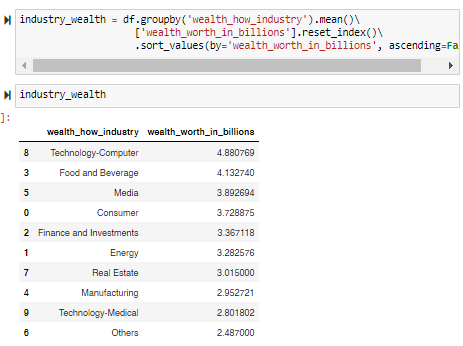


To get the age range the dataset was strategically shared into age groups using Pandas method, pd.cut. 

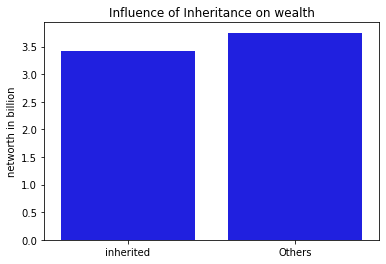
* The Technology-Computer industry on average has the highest net worth in billions, followed by the Food and Beverage with 4.9 and 4.1 respectively.



. The table below shows the average worth in billons for all industries:



* Inherited wealth and other means of wealth as a mean of 3.41 and 3.75 respectively among billionaires. Billionaires made from other sources of wealth has a tendency to be wealthier than those who are inherited.



**Hypothesis Testing**

To further generate insights from the data some hypothesis will be tested using the dataset. The Null hypothesis to be proved are:

* There is no relationship between a billionaire’s net worth and their country.
* There is no significant difference in the gender distribution of billionaires across different industries.
* There is no significant difference in the distribution of billionaires' net worth between those who inherited their wealth and those who started their own businesses or acquired their wealth through other means.

To perform the statistical significance testing the SciPy.Stats was used. It is a Python library that provides a wide range of statistical functions and tools for working with statistical data. It is part of the larger SciPy library, which is a collection of scientific and numerical computing tools for Python.

Some of the functions provided by SciPy. Stats include:

ttest\_ind(): Calculates the t-test for the means of two independent samples.

pearsonr(): Calculates the Pearson's correlation coefficient and the p-value for testing non-correlation.

chi2\_contingency(): Calculates the chi-squared statistic and p-value for a contingency table.

linregress(): Calculates a linear regression and returns the slope, intercept, and other statistics.

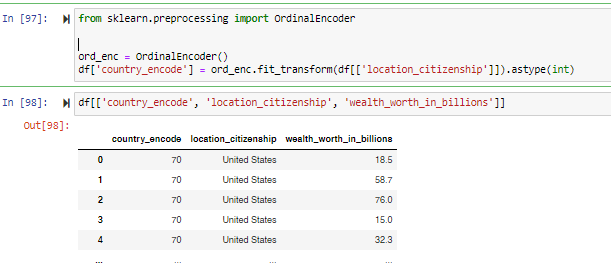
norm.pdf(): Calculates the probability density function (PDF) of the normal distribution.

These are just a few examples, and SciPy.Stats provides many other functions for a wide range of statistical tasks. It is a useful tool for working with statistical data in Python.

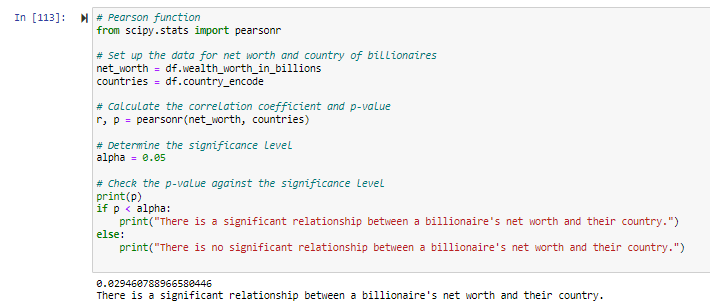
A significance level of 0.05 was used. It means that there is a 5% chance of rejecting the null hypothesis when it is true. This is a commonly used significance level in statistical testing, and it means that there is a relatively low risk of making a type I error.

1. **Null hypothesis:** There is no relationship between a billionaire’s net worth and their country

**Process:** The hypothesis was tested using a Pearson's correlation coefficient. This test measures the strength and direction of the linear relationship between two continuous variables (in this case, net worth, and country). The 73 countries were encoded to integer



The encoded country and the wealth worth in billions is passed to the Pearson function.

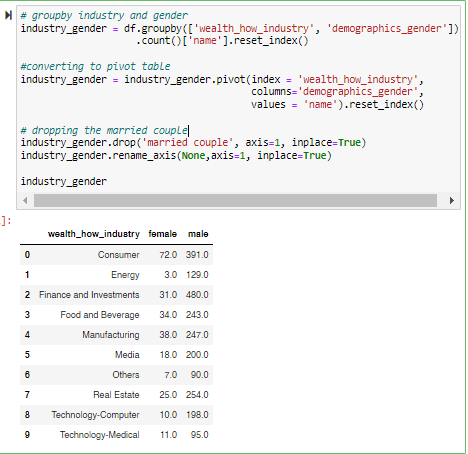


**Result:** There is a significant relationship between a billionaire's net worth and their country.

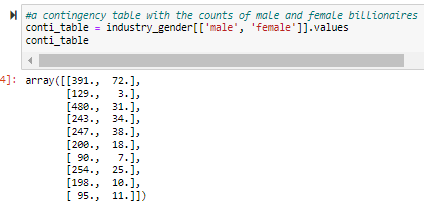
The null hypothesis was rejected, and the alternative was accepted.

1. **Null hypothesis:** There is no significant difference in the gender distribution of billionaires across different industries.

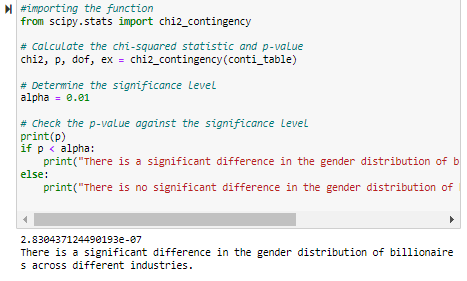
**Process:** The data was group by industry and gender. The gender is limited to male and female to satisfy the assumption for Chi-squared test.

. 

The contingency table of male and female count is extracted



The gender count is passed to a Chi-squared test function

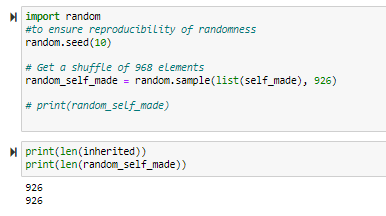


**Result:** There is a significant difference in the gender distribution of billionaires across different industries. The null hypothesis was rejected, and the alternative was accepted.

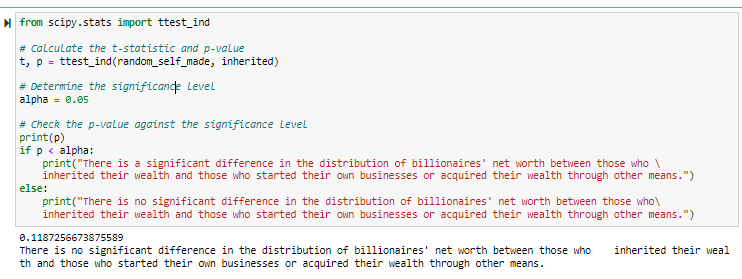
1. **Null hypothesis:** There is no significant difference in the distribution of billionaires' net worth between those who inherited their wealth and those who started their own businesses or acquired their wealth through other means.

**Process:** This hypothesis was tested using a two-sample t-test. This test compares the means of two groups (in this case, billionaires who started their own businesses and billionaires who inherited) to determine whether there is a significant difference between the groups.

The self-made billionaires are more than the inherited, to fulfil the assumption of t-test 926 self-made were selected at random from the 1688 self-made.



The random self-made and the inherited is passed to the t-test function



**Result:** There is no significant difference in the distribution of billionaires' net worth between those who inherited their wealth and those who started their own businesses or acquired their wealth through other means. Therefore, the null hypothesis was accepted.

**Evaluation and Future Development**

The results of this analysis can have a variety of potential impacts and uses within the application domain, regarding on the specific research question being asked and the conclusions drawn from the analysis.

For example, the analysis focuses on the characteristics of self-made billionaires (e.g., their industry, age, etc.), the results can be used to identify trends and patterns that might be useful for aspiring entrepreneurs or policymakers looking to support the development of new businesses.

The results can also be used to inform discussions about gender inequality and the concentration of wealth in countries like United States and in some industries.

In general, the results of an analysis on billionaire data could be used to inform policy decisions, shape public opinion, and provide insights into the factors that contribute to success and wealth in different industries and countries. E.g., the hypothesis on impact of inheritance.

It is important to keep in mind, however, that the results of this analysis are only as reliable as the data and methods used, and that it is important to carefully consider the limitations and potential biases in the data and the assumptions of the statistical tests being used.

Future work on this analysis will include, getting more and updated data of billionaires and big data technologies like Spark or Big Query will be used to analysis it.

REFERNCES

<https://www.forbes.com/sites/chasewithorn/2016/03/07/how-billionaires-get-rich-which-industries-make-the-most-mega-fortunes/?sh=2684dceb45e1>

<https://www.forbes.com/sites/jamesodonnell/2022/04/06/how-to-become-a-billionaire-these-industries-have-the-most-people-on-the-2022-forbes-list/?sh=5287b8226898>

<https://en.wikipedia.org/wiki/Kunio_Busujima>