A PROJECT REPORT

ON

“Examining the Infant Mortality of India: A Logistic Regression Analysis Approach to Uncover Influential Factors Utilizing Python, Excel, SPSS and NFHS-5 Data Sourced From Demographic and Health Survey Program”

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

Master of Science

In

Statistics

By

GOVIND HARI	**22STMSA107**
MD AMAN AHMAD	**22STMSA121**
NIKRANT TOMAR	**22STMSA124**
KM PRACHI BALIYAN	**22STMSA125**
MOHAMMED ABRAZ AZEER	**22STMSA127**

Under the Joint Supervision of

PROF. AQUIL AHMED & DR. AIJAZ AHMAD DAR

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY
ALIGARH (INDIA)

2023-2024

ABSTRACT

The project titled "*Examining the Infant Mortality of India: A Logistic Regression Analysis Approach to Uncover Influential Factors Utilizing Python, Excel, SPSS and NFHS-5 Data Sourced From Demographic and Health Survey Program*" aims to investigate and analyze the factors influencing infant mortality using statistical tools such as SPSS, Python, and Excel. The study relies on data gathered from the Demographic and Health Surveys to shed light on the current situation in India. By employing advanced statistical techniques, the project seeks to identify key determinants of infant mortality and their implications for achieving targets outlined in Sustainable Development Goal 3 (SDG3).

The comprehensive analysis will delve into various socio-economic, healthcare, and demographic factors affecting infant mortality rates. Insights gained from this research could contribute to informed policy decisions and interventions aimed at reducing infant mortality in India. The use of SPSS, Python, and Excel facilitates a robust and nuanced examination of the data, allowing for a thorough understanding of the complex relationships within the context of SDG 3. Ultimately, the project aspires to provide valuable insights that can guide efforts to improve infant health outcomes and contribute to the broader goal of achieving sustainable development in the health sector.

The content of this project dissertation is divided into 4 chapters.

Chapter-I is introductory in nature and discusses various preliminaries pertaining to the current study. Chapter-II is devoted to the methodology used in the project work. In Chapter-III, various statistical tools are implemented, and the results drawn are interpreted in detail. In chapter-IV, the conclusion of the study is summed up, and the recommendations are made accordingly.

DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH, ALIGARH MUSLIM UNIVERSITY, ALIGARH

202002 (U.P), INDIA

Phone: (0571) 2701251

PROF. AQUIL AHMED

DR. AIJAZ AHMAD DAR

CERTIFICATE

This is to certify that the project entitled "*Analysing Infant Mortality of India: A Logistic Regression Analysis Approach to Uncover Influential Factors Utilizing Python, Excel, SPSS and NFHS-5 Data Sourced From Demographic Health Survey Program*" has been carried out by the following group of students of **Master of Science (Statistics), Final Year, 2023-24,** under our joint supervision for the course **"STM3072 – Project"**.

NAME	Roll no.
GOVIND HARI	22STMSA107
MD AMAN AHMAD	22STMSA121
KM PRACHI BALIYAN	22STMSA125
NIKRANT TOMAR	22STMSA124
MOHAMMED ABRAZ AZEER	22STMSA127

PROF. AQUIL AHMED
(Supervisor)

DR. AIJAZ AHMAD DAR
(Supervisor)

ACKNOWLEDGMENT

All praises to the Almighty for providing us with the strength, inspiration, and resilience throughout the course of this project.

It is our privilege to express our sincere thanks and a deep sense of gratitude to our supervisors ***Prof. Aquil Ahmed (Chairman) and Dr. Aijaz Ahmad Dar (Assistant Professor)***, Department of Statistics and Operations Research, AMU, Aligarh. Their guidance, painstaking efforts, and deep insight into the problem that proved to be fruitful in improving the overall quality of this project work.

We would like to express our gratitude to them for their guidance, valuable suggestions and immense patience due to which this work was able to take the shape in which it has been published.

We are thankful to the Chairman of our department, ***Prof. Aquil Ahmed*** for providing us with all the necessary facilities that were imperative for the complement of this project work.

We would also like to express our special gratitude to our parents and well-wishers for their prayers and for providing us with moral support to pursue our studies.

Govind Hari

Md Aman Ahmad

Km Prachi Baliyan

Nikrant Tomar

Mohammed Abraz Azeer

CONTENTS

Chapter-1

INTRODUCTION

1.1. Sustainable Development Goals

The Sustainable Development Goals (SDGs) are a set of 17 global goals established by the United Nations (UN) in 2015. They aim to address a wide range of global challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. The SDGs provide a shared blueprint for countries, organizations, and individuals to work together to achieve a more sustainable and equitable future by the year 2030.

List of all 17 Sustainable Development Goals:

i. SDG 01: No Poverty
ii. SDG 02: Zero Hunger
iii. SDG 03: Good Health and Well-being
iv. SDG 04: Quality Education
v. SDG 05: Gender Equality
vi. SDG 06: Clean Water and Sanitation
vii. SDG 07: Affordable and Clean Energy
viii. SDG 08: Decent Work and Economic Growth
ix. SDG 09: Industry, Innovation, and Infrastructure
x. SDG 10: Reduced Inequality
xi. SDG 11: Sustainable Cities and Communities
xii. SDG 12: Responsible Consumption and Production
xiii. SDG 13: Climate Action
xiv. SDG 14: Life Below Water
xv. SDG 15: Life on Land
xvi. SDG 16: Peace, Justice, and Strong Institutions
xvii. SDG 17: Partnerships for the Goals

Since this project is mainly focusing on the infant mortality of India that comes under the banner of SDG-3, So it is not possible to explain about all the 17 SDGs. Here is a brief overview of SDG-3 i.e; Good Health and Wellbeing.

The contents of this chapter is prepared with the help of various sources whose references are given in the reference section.

1.2. SDG-3: Good Health and Well-being

Objective: Ensure healthy lives and promote well-being for all ages.

Key Targets of SDG-3:

i. Reduce the global maternal mortality ratio.
ii. End preventable deaths of newborns and children under 5 years of age.
iii. Ensure access to sexual and reproductive health-care services.
iv. Achieve universal health coverage, including financial risk protection, access to quality essential healthcare services, and access to safe, effective, quality, and affordable essential medicines and vaccines for all.
v. Substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water, and soil pollution and contamination.
vi. Prevent and treat substance abuse, including narcotic drug abuse and harmful use of alcohol.
vii. By 2030, ensure universal access to sexual and reproductive health-care services, including for family planning, information and education, and the integration of reproductive health into national strategies and programs.
viii. Achieve universal health coverage, including financial risk protection, access to quality essential healthcare services, and access to safe, effective, quality, and affordable essential medicines and vaccines for all.

1.3. Infant Mortality

Infant mortality, defined as the death of a child under one year of age, remains a critical public health indicator and a poignant reflection of a nation's healthcare system and socio-economic landscape. In the context of India, a country marked by its vast diversity and complex healthcare challenges, addressing the issue of infant mortality takes on paramount importance.

As the second most populous country in the world, India grapples with the persistent challenge of high infant mortality rates across various regions. The disparities in healthcare access, socio-economic conditions, and maternal and child healthcare practices contribute to the complex tapestry of factors influencing infant mortality.

This project aims to delve into the multifaceted aspects of infant mortality in India, unraveling the intricate web of causes and consequences. By examining the demographic, health, and socio-economic factors, we seek to identify the root causes of infant mortality and propose evidence-based strategies for its mitigation.

In the pages that follow, we will explore the current scenario of infant mortality in India, analyze the contributing factors, and assess the effectiveness of existing interventions. Through this comprehensive study, we aspire to contribute meaningful insights to inform policy decisions, healthcare initiatives, and community-based interventions aimed at reducing infant mortality rates and improving the overall well-being of our youngest citizens.

1.3.1. Infant Mortality in India

Infant mortality in India remains a significant public health concern, reflecting the challenges and disparities present in the country's healthcare landscape. As a critical indicator of the well-being of the youngest members of society, infant mortality rates (IMR) serve as a barometer for assessing the effectiveness of healthcare systems, socio-economic conditions, and public health interventions.

Despite notable advancements in healthcare infrastructure and services, India continues to grapple with relatively high IMR, with variations across different states and regions. Several factors contribute to the complex nature of infant mortality in the country.

1.3.2. Socio-Economic Disparities and Their Contribution Towards Infant Mortality:

Socio-economic disparities significantly impact infant mortality rates, reflecting the inequalities in access to resources and healthcare services within a society. In the context of infant mortality, socio-economic factors encompass a range of economic and social conditions that influence the well-being of families and communities. Here's a brief description of How Socio-economic Disparities can contributes towards infant mortality is given as follows

i. ***Limited Access to Quality Healthcare****:* Families facing socio-economic challenges often encounter barriers in accessing essential healthcare services. This includes prenatal care, skilled attendance during childbirth, and postnatal care for both the mother and the newborn. The lack of timely and adequate medical attention can contribute to adverse outcomes for infants.

ii. ***Nutritional Deficiencies***: Economic disparities often lead to insufficient access to nutritious food and supplements for pregnant women and infants. Malnutrition during pregnancy and early childhood can increase the risk of low birth weight and other health complications, contributing to higher infant mortality rates.

iii. ***Inadequate Sanitation and Living Conditions*****:** Families with lower socio-economic status may reside in environments with limited access to clean water, proper sanitation, and hygienic living conditions. Poor environmental conditions can expose infants to infectious diseases, further increasing the likelihood of infant mortality.

iv. ***Limited Education and Awareness*****:** Lower levels of education among parents, particularly mothers, are associated with higher infant mortality rates. Education plays a crucial role in promoting health-seeking behaviors, understanding proper infant care practices, and making informed decisions about healthcare

v. ***Employment and Income Disparities*****:** Families facing economic challenges may experience stressors related to unemployment, underemployment, or insufficient income. Economic instability can impact a family's ability to afford healthcare services, nutritious food, and a safe living environment, influencing infant health outcomes.

Addressing socio-economic disparities in the context of infant mortality requires comprehensive strategies that go beyond healthcare services. Initiatives promoting education, poverty alleviation, and equitable economic opportunities are essential components of a holistic approach to reduce infant mortality. By tackling the root causes of socio-economic disparities, societies can work towards creating conditions that foster healthier pregnancies, safer childbirth, and improved infant well-being.

Healthcare factors that can affect infant mortality

i. ***Maternal Health*****:** Maternal health is a critical determinant of infant mortality, as the well-being of the mother significantly influences the health outcomes of the newborn. Here's a brief overview of the relationship between maternal health and infant mortality.

ii. ***Prenatal and Antenatal Care****:* Timely and regular prenatal and antenatal care are crucial for monitoring the health of pregnant women and identifying potential risks early in pregnancy. Limited access to these services is associated with higher rates of complications during childbirth and adverse outcomes for infants.

iii. ***Maternal Nutrition*****:** The nutritional status of the mother during pregnancy has a direct impact on the growth and development of the fetus. Inadequate nutrition can lead to low birth weight, preterm birth, and other complications that increase the risk of infant mortality.

iv. ***Maternal Infections and Diseases***: Maternal infections, such as those caused by HIV or certain sexually transmitted infections, can be transmitted to the infant during pregnancy, childbirth, or breastfeeding, leading to increased infant mortality. Additionally, chronic maternal diseases such as diabetes or hypertension can contribute to adverse outcomes for the infant.

v. ***Healthcare Access:*** Healthcare access plays a pivotal role in determining infant mortality rates, as the availability and quality of healthcare services directly impact the well-being of pregnant women, newborns, and infants. Here's a brief overview of how healthcare access is linked to infant mortality.

vi. ***Skilled Birth Attendance:*** The presence of skilled healthcare professionals, such as midwives or obstetricians, during childbirth is essential for ensuring a safe delivery. Access to skilled birth attendance reduces the risk of complications and improves the chances of a healthy outcome for both the mother and the infant.

vii. ***Primary Healthcare Services:*** Access to primary healthcare services, including routine check-ups, preventive care, and health education, is vital for promoting overall maternal and child health. Comprehensive primary healthcare can address health issues at their roots, preventing complications that may lead to infant mortality.

Addressing maternal health in the context of infant mortality involves comprehensive strategies that focus on improving access to quality healthcare, promoting maternal nutrition, preventing and managing maternal infections, and ensuring proper family planning. By prioritizing maternal well-being, societies can make significant strides in reducing infant mortality and fostering healthier starts for newborns.

Addressing healthcare access barriers involves improving infrastructure, increasing the number of skilled healthcare professionals, and implementing community-based interventions. By ensuring that healthcare services are accessible, available, and of high quality, societies can significantly reduce infant mortality rates and enhance the overall health and well-being of mothers and infants

1.4. Infant Mortality Rate

Infant Mortality Rate (IMR) is a critical demographic indicator that measures the number of deaths of infants under one year of age per 1,000 live births within a given population during a specific time period. This metric is a fundamental component of vital statistics and plays a crucial role in assessing the overall health and well-being of a society.

$$\textbf{IMR} = \frac{Number\ of\ infant\ death\ in\ a\ year}{Number\ of\ total\ live\ birth} \ \mathbf{*1000}$$

i. Number of Infant Deaths: This refers to the total number of deaths among infants under one year of age during the specified time period.

ii. Number of Live Births: This represents the total number of live births that occurred within the same time frame.

1.4.1. Significance of Infant Mortality Rate:

The significance of IMR extends beyond a mere statistical figure. It serves as a sensitive indicator of the overall health status of a population and the quality of maternal and child healthcare services. A high IMR indicates potential issues in prenatal care, maternal health, and access to essential healthcare services for infants.

Understanding IMR is crucial for policymakers, healthcare professionals, and researchers as it allows for targeted interventions and resource allocation. High IMR may point to areas where healthcare infrastructure needs improvement, education on maternal and child health should be enhanced, and socioeconomic conditions demand attention.

1.5. Factors Influencing Infant Mortality Rate

i. ***Maternal Factors:*** Maternal health plays a pivotal role in infant survival. Issues such as maternal age, nutritional status, access to prenatal care, and maternal education can impact the likelihood of a healthy infant.

ii. ***Healthcare-Related Factors****:* The availability and quality of healthcare services are critical in preventing and addressing infant mortality. Factors like access to skilled birth attendants, immunization coverage, and availability of neonatal care facilities significantly influence IMR.

iii. ***Socio-Economic Determinants:*** Socioeconomic factors, including income levels, education, sanitation, and overall living conditions, play a substantial role in infant mortality. Poverty, for instance, can limit access to healthcare and proper nutrition, leading to higher IMR.

1.5.1. Global and Regional Perspectives on Infant Mortality Rate:

The global perspective on IMR reveals significant disparities across countries and regions. Developed nations often boast lower IMRs due to advanced healthcare infrastructure, better socioeconomic conditions, and comprehensive public health programs. In contrast, developing nations may face higher IMRs due to challenges in healthcare accessibility and quality, poverty, and limited educational resources.

Infant Mortality Rate is a crucial metric that provides insights into the health and well-being of the youngest members of a population. By understanding the factors influencing IMR and implementing targeted interventions, societies can work towards ensuring the survival and thriving of infants worldwide.

1.6. Source of the Data

The foundation of any empirical research lies in the quality and reliability of its data source. In the case of our project on "*Analysing Infant Mortality in India: A Logistic Regression Analysis Approach to Uncover Influential Factors Utilizing Python, Excel, and SPSS with NFHS 5 Data sourced from Demographic Health Survey program*" we draw crucial insights and evidence from the National Family Health Survey 5 (NFHS-5). NFHS-5 serves as the primary reservoir of comprehensive and up-to-date information on various health and demographic indicators in India. This introduction aims to illuminate the significance, methodology, and key features of NFHS-5 as the cornerstone of our investigation.

1.6.1. Overview of NFHS-5:

The National Family Health Survey (NFHS) is a series of nationally representative surveys conducted in India to capture essential data related to population, health, and nutrition. As the most recent iteration, NFHS-5 promises a snapshot of India's health landscape, offering insights into vital aspects such as maternal and child health, family planning, nutrition, and more.

NFHS-5 encompasses both urban and rural areas across all states and union territories of India. This extensive coverage ensures representation from diverse geographical, cultural, and socio-economic contexts, providing a nuanced understanding of health dynamics across the nation.

The survey employs standardized questionnaires and methodologies, ensuring consistency and comparability of data. NFHS-5 is household-based, collecting information from a representative sample of households. The meticulous approach to data collection allows for a comprehensive analysis of health-related trends and patterns.

NFHS-5 explores a plethora of health and demographic indicators crucial for our study. These include maternal health indicators such as antenatal care and delivery practices, child health indicators like immunization coverage and feeding practices, family planning data, and nutritional status information.

One of the strengths of NFHS-5 lies in its ability to disaggregate data by various demographic and socio-economic factors. This enables us to delve into the intricacies of health disparities and identify vulnerable populations that may require targeted interventions.

Beyond its role as a data repository, NFHS-5 has significant implications for policymaking and program planning. The findings contribute to evidence-based decision-making, allowing policymakers to gauge the effectiveness of existing health initiatives and identify areas for improvement.

1.6.2. Accessing NFHS-5 Data

NFHS-5 data is made publicly accessible through official reports, fact sheets, and a dedicated data portal. Researchers, policymakers, and the public at large can explore and utilize this rich dataset for analyses, fostering a deeper understanding of health dynamics in India. In this project the NFHS-5 is taken from the official website of “Demographic Health Survey Program”.

1.6.3. Demographic Health Survey Program

The Demographic and Health Surveys (DHS) program is an international initiative that conducts nationally representative surveys in various countries to collect data on population, health, and nutrition.

DHS aims to provide high-quality demographic and health data that can be used for policy and program planning, monitoring, and evaluation. The surveys cover a wide range of topics, including fertility, maternal and child health, family planning, nutrition, and more.

DHS surveys are typically implemented by national statistical offices in collaboration with ICF International and are funded by the United States Agency for International Development (USAID).

1.7. Review of Literature

This section provides a comprehensive foundation for this research, The assessment of factors influencing infant mortality in India is a critical endeavor, given the profound implications for public health and societal well-being. This comprehensive literature review synthesizes existing studies to contribute to the understanding of this complex issue. The multifaceted nature of infant mortality necessitates a nuanced examination of various factors, ranging from socioeconomic determinants to healthcare accessibility, maternal health, and environmental conditions. Numerous studies have delved into the intricacies of infant mortality in the Indian context, revealing a mosaic of contributing factors. Socioeconomic disparities emerge as a recurring theme, with poverty and lack of education identified as pivotal determinants. Maternal health also plays a crucial role, as the well-being of mothers significantly influences the survival prospects of their infants. Access to healthcare services, including prenatal and postnatal care, emerges as a critical factor that demands attention in the context of reducing infant mortality rates. The present review aims to build upon the existing body of knowledge by employing a multi-dimensional approach that integrates Python, Excel, and SPSS for a more comprehensive and data-driven analysis. Leveraging these powerful analytical tools enhances the precision and depth of our study, allowing for a systematic exploration of the interplay between diverse variables influencing infant mortality. Python facilitates advanced data manipulation and visualization, Excel offers a user-friendly platform for data organization, and SPSS enables sophisticated statistical analyses. As we navigate through the existing literature, this review identifies gaps and methodological limitations in previous studies, providing a foundation for the novel contributions of our research. By synthesizing findings from a diverse array of studies and employing sophisticated analytical tools, our project seeks to uncover new insights and refine existing understandings of the multifactorial landscape contributing to infant mortality in India. The integration of Python, Excel, and SPSS not only enhances the analytical rigor of our study but also showcases the synergy between traditional epidemiological research and modern computational methodologies in addressing complex public health challenges.

1.7.1. “Predictive model of under-five mortality in developing countries: evidence from multiple indicators cluster survey Punjab” By *Muhammad Nadeem, Shahid Adil, Fatima Hunnain, Shahzada M. Naeem Nawaz*

The paper under consideration aims to contribute to the reduction of under-five mortality in developing countries, with a specific focus on Punjab. The objective is clearly stated, emphasizing the need for a comprehensive framework to address this critical issue. By focusing on a specific region, the study recognizes the importance of tailoring strategies to the unique challenges faced by different populations. The data and methodological approach employed in the research provide a robust foundation for the study's findings. Leveraging data from the Multiple Indicators Cluster Survey (MICS) Punjab conducted in 2017–18, the study utilizes binomial logistic regression to identify predictors of under-five mortality. This analytical method enhances the precision of the study, allowing for a nuanced understanding of the complex factors influencing child survival. The theoretical framework adopted in this study draws from the Mosley and Chen (1984) approach, integrating both social and biological factors that impact child survival. This comprehensive framework encompasses a range of factors, including maternal characteristics, environmental pollution, nutritional deficiencies, injuries, and health care behavior. By incorporating these diverse elements, the study provides a holistic understanding of the multifaceted nature of under-five mortality. The literature review undertaken by the authors critically examines existing research on under-five mortality in developing countries. It highlights a crucial gap in the literature—specifically, the need for fresh evidence and viable strategies tailored to the context of Punjab. This acknowledgment sets the stage for the paper's unique contribution by not only identifying gaps but also proposing a framework to address these gaps. In summary, this paper offers a valuable contribution to the field by combining a clear objective, robust methodology, a comprehensive theoretical framework, and a critical review of existing literature. The focus on Punjab adds a contextual dimension to the study, recognizing the importance of localized strategies in addressing under-five mortality in diverse developing country settings.

1.7.2. “A Statistical analysis of child mortality, An Evidence from Nigeria” By *Gbemisola Wuraola Samuel, Emmanuel O Amoo*

The research article under review presents a meticulous examination of child mortality in Nigeria, utilizing data extracted from the Nigeria Demographic and Health Survey (NDHS) 2008. The primary objective of the study is to identify predictors of child mortality among children aged 0-4 years, offering valuable insights for policy formulation aimed at reducing child mortality in Nigeria. The methodology employed in the research is robust, involving cross-tabulation and binary logistic regression techniques to scrutinize the association between child mortality and an array of socio-economic, demographic, and maternal health care variables. The study narrows its focus to currently married women, distinguishing between those who have experienced child mortality and those who have not. The literature review within the paper provides a comprehensive backdrop, encompassing global and regional trends of child mortality. Notably, the study delves into the causes and determinants of child mortality, including infectious diseases, malnutrition, maternal education, wealth index, family size, birth order, age of the mother, and place of residence. By contextualizing its findings within the broader landscape of existing research, the study adds depth and relevance to its own analysis. Furthermore, the article emphasizes the challenges and gaps in achieving Millennium Development Goal 4, which aimed to reduce child mortality by two-thirds by 2015. This critical perspective contributes to a holistic understanding of the broader implications of the study's findings. In synthesizing these key elements, the research serves as a valuable resource for policymakers, researchers, and public health practitioners seeking evidence-based strategies to address child mortality in Nigeria.

Chapter-2

METHODOLOGY

2.1. Methodology

The Merriam-Webster Dictionary defines ‘methodology’ as “a body of methods, rules, and postulates employed by a discipline” or “a particular procedure or set of procedures”.
In the field of project management, this would be a set of rules and processes that define how you manage a project.

2.2. Data Collection and Analysis

2.2.1. Data

Data is a collection of information gathered by observations, measurements, research or analysis. They may consist of facts, numbers, names, figures or even description of things. Data is organized in the form of graphs, charts or tables.

2.2.2. Types of Data

Based on the source data is classified as Primary Data and Secondary Data

i. ***Primary Data:*** It refers to the data that the investigator collects for the very first time. This type of data has not been collected either by this or any other investigator before. Primary data will provide the investigator with the most reliable first-hand information about the respondents. The investigator would have a clear idea about the terminologies used, the statistical units employed, the research methodology and the size of the sample. Primary data may either be internal or external to the organization.

ii. ***Secondary Data:*** It refers to the data that the investigator collects from another source. Past investigators or agents collect data required for their study. The investigator is the first researcher or statistician to collect this data. Moreover, the investigator does not have a clear idea about the intricacies of the data. There may be ambiguity in terms of the sample size and sample technique. There may also be unreliability with respect to the accuracy of the data.

The content of this chapter is prepared with the help of various sources whose references are given in the reference section.

Based on the nature, Data is classified as Qualitative data and Quantitative Data

i. ***Qualitative data:*** Qualitative data is defined as the data that approximates and characterizes. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods. Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon. Qualitative data is also called categorical data since this data can be grouped according to categories.

   For example, think of a student reading a paragraph from a book during one of the class sessions. A teacher who is listening to the reading gives feedback on how the child read that paragraph. If the teacher gives feedback based on fluency, intonation, throw of words, clarity in pronunciation without giving a grade to the child, this is considered as an example of qualitative data.

ii. ***Quantitative Data****:* Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data. Height in feet, age in years, and weight in pounds are examples of quantitative data.

2.2.3. Methods of Collecting Data

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories:

i. Methods of collecting primary data.
ii. Methods of collecting secondary data.

2.2.4. Methods of Collecting Primary Data

i. Direct Personal Investigation
ii. Indirect Oral Interview
iii. Mailed Questionnaire
iv. Schedules
v. Local Agencies

2.2.5. Methods of Collecting Secondary Data

The sources of secondary data

i. ***Published Sources:*** There are many national organizations, international agencies and official publications that collect various statistical data. They collect data related to business, commerce, trade, prices, economy, productions, services, industries, currency and foreign affairs. They also collect information related to various (internal and external) socio-economic phenomena and publish them. These publications contain statistical reports of various kinds. Central Government Official Publication, Publications of Research Institutions, Committee Reports and International Publications are some published sources of secondary data.

ii. ***Web Scraping*****:** It is also termed Screen Scraping, Web Data Extraction, Web Harvesting etc. It is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in a computer or to a database in table (spreadsheet) format.

iii. ***Unpublished Sources:* -** Some statistical data are not always a part of publications. Such data are stored by institutions or a private firm. Researchers often make use of these unpublished data in order to make their researches all the more original.

iv. ***Data Mining****:* Data mining refers to extracting valuable information from a large amount of data.

2.2.6. Data Mining Techniques

i. ***Classification:* -** This analysis is used to retrieve important and relevant information about data, and meta-data (it is information that provides information about the other data). This data mining method helps to classify data in different classes.

ii. ***Clustering:* -** Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

iii. ***Regression:* -** Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

iv. ***Association Rules*: -** This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

v. ***Outer detection:* -** This type of data mining technique refers to observation of data items in the data set which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

vi. ***Sequential Patterns:* -** This data mining technique helps to discover or identify similar patterns or trends in transaction data for a certain period.

vii. ***Prediction:* -** Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

2.2.7. Challenges in Implementation of Data Mining

i. Skilled: Experts are needed to formulate the data mining queries.

ii. Over-fitting: Due to the small size training database, a model may not fit future states.

iii. Data mining needs large databases which sometimes are difficult to manage

iv. Business practices may need to be modified to determine how to use the information uncovered.

v. If the data set is not diverse, data mining results may not be accurate.

2.2.8. Advantages of Data Mining

i. Data mining techniques helps companies to get knowledge-based information.

ii. Data mining helps organizations to make profitable adjustments in operation and production.

iii. Data mining is a cost-effective and efficient solution compared to other statistical data applications.

iv. Data mining helps with the decision-making process.

v. Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.

vi. It can be implemented in new systems as well as existing platforms.

vii. It is the speedy process which makes it easy for the users to analyze huge amounts of data in less time.

2.2.9. Disadvantages of Data Mining

i. There are chances that companies may sell useful information about their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.

ii. Many data mining analytics software is difficult to operate and requires advanced training to work on. Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of the correct data mining tool is a very difficult task.

iii. The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

2.3. Graphical Visualization Techniques

2.3.1. Bar Chart: A bar chart (aka bar graph, column chart) plots numeric values for levels of a categorical feature as bars. Levels are plotted on one chart axis, and values are plotted on the other axis. Each categorical value claims one bar, and the length of each bar corresponds to the bar’s value. Bars are plotted on a common baseline to allow for easy comparison of values. From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others. Since this is a fairly common task, bar charts are a fairly ubiquitous chart type.

2.3.2. Pie Chart: A pie chart (or a circle chart) is a circular statistical graphic which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented. The earliest known pie chart is generally credited to William Playfair's Statistical Breviary of 1801. Pie charts are very widely used in the business world and the mass media. However, they have been criticized, and many experts recommend avoiding them, as research has shown it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts. Pie charts can be replaced in most cases by other plots such as the bar chart, box plot, dot plot, etc.

2.3.4. Line Chart: A line chart or line graph, also known as curve chart, is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. It is similar to a scatter plot except that the measurement points are ordered (typically by their x-axis value) and joined with straight line segments. A line chart is often used to visualize a trend in data over intervals of time – a time series – thus the line is often drawn chronologically. In this case they are known as run charts.

2.3.5. Histogram: In statistics, a histogram is a graphical representation of the distribution of data. The histogram is represented by a set of rectangles, adjacent to each other, where each bar represents a kind of data. Statistics is a stream of mathematics that is applied in various fields. When numerals are repeated in statistical data, this repetition is known as Frequency and which can be written in the form of a table, called a frequency distribution. A Frequency distribution can be shown graphically by using different types of graphs and a Histogram is one among them.

2.4. Crosstabs

Cross tabulation (crosstab) is a useful analysis tool commonly used to compare the results for one or more variables with the results of another variable. It is used with data on a nominal scale, where variables are named or labelled with no specific order. They are basically data tables that present the results from a full group of survey respondents as well as subgroups. They allow us to examine relationships within the data that might not be obvious when simply looking at total survey responses.

2.5. Odd Ratio

An odds ratio (OR) is a statistic that quantifies the strength of the association between two events, A and B. The odds ratio is defined as the ratio of the odds of A in the presence of B and the odds of A in the absence of B, or equivalently (due to symmetry), the ratio of the odds of B in the presence of A and the odds of B in the absence of A. Two events are independent if and only if the OR equals 1, i.e., the odds of one event are the same in either the presence or absence of the other event. If the OR is greater than 1, then A and B are associated (correlated) in the sense that, compared to the absence of B, the presence of B raises the odds of A, and symmetrically the presence of A raises the odds of B. Conversely, if the OR is less than 1, then A and B are negatively correlated, and the presence of one event reduces the odds of the other event.

2.6. Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis. Hypothesis testing allows us to make probabilistic statements about population parameters. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed. All analysts use a random population sample to test two different hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is usually a hypothesis of equality between population parameters, H0 is the symbol for it. and the alternative hypothesis is effectively the opposite of a null hypothesis, H1 is the symbol for it. Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true. Depending on the population distribution, you can classify the statistical hypothesis into two types; Simple Hypothesis: A simple hypothesis specifies an exact value for the parameter. Composite Hypothesis: A composite hypothesis specifies a range of values.

2.6.1. Parametric Tests:

The basic principle behind the parametric tests is that we have a fixed set of parameters that are used to determine a probabilistic model that may be used in Machine Learning as well. 30 Parametric tests are those tests for which we have prior knowledge of the population distribution (i.e, normal), or if not then we can easily approximate it to a normal distribution which is possible with the help of the Central Limit Theorem.

Parameters for using the normal distribution is –

i. Mean
ii. Standard Deviation

Eventually, the classification of a test to be parametric is completely dependent on the population assumptions. There are many parametric tests available from which some of them are as follows:

i. To find the confidence interval for the population means with the help of known standard deviation.
ii. To determine the confidence interval for population means along with the unknown standard deviation.
iii. To find the confidence interval for the population variance.

2.6.2. Non-Parametric Test:

Non-parametric tests are experiments that do not require the underlying population for assumptions. It does not rely on any data referring to any particular parametric group of probability distributions. Non-parametric methods are also called distribution-free tests since they do not have any underlying population. In Non-Parametric tests, we don’t make 32 any assumptions about the parameters for the given population or the population we are studying.

2.6.3. Need of Non-Parametric Tests?

This is also the reason that nonparametric tests are also referred to as distribution-free tests. In modern days, non-parametric tests are gaining popularity and an impact of influence some reasons behind this fame is

i. The main reason is that there is no need to be mannered while using parametric tests.
ii. The second reason is that we do not require to make assumptions about the population given (or taken) on which we are doing the analysis.

2.7. Mann Whitney U Test

The Mann-Whitney U Test, also known as the Wilcoxon Rank Sum Test, is a non-parametric statistical test used to compare two samples or groups.
The Mann-Whitney U Test assesses whether two sampled groups are likely to derive from the same population, and essentially asks; do these two populations have the same shape with regards to their data? In other words, we want evidence as to whether the groups are drawn from populations with different levels of a variable of interest. It follows that the hypothesis in a Mann-Whitney U Test are :

i. The null hypothesis (H_0) is that the two populations are equal.

ii. The alternative hypothesis (H_1) is that the two populations are not equal.

Some researchers interpret this as comparing the medians between the two populations (in contrast, parametric tests compare the means between two independent groups). In certain situations, where the data are similarly shaped (see assumptions), this is valid – but it should be noted that the medians are not actually involved in calculation of the Mann-Whitney U test statistic. Two groups could have the same median and be significantly different according to the Mann-Whitney U test.

Some key assumptions for Mann-Whitney U Test are detailed below:

i. The variable being compared between the two groups must be continuous (able to take any number in a range – for example age, weight, height or heart rate). This is because the test is based on ranking the observations in each group.

ii. The data are assumed to take a non-Normal, or skewed, distribution. If your data are normally distributed, the unpaired Student's t-test should be used to compare the two groups instead.

iii. While the data in both groups are not assumed to be Normal, the data are assumed to be similar in shape across the two groups.

iv. The data should be two randomly selected independent samples, meaning the groups have no relationship to each other. If samples are paired (for example, two measurements from the same group of participants), then a paired samples t-test should be used instead.

v. Sufficient sample size is needed for a valid test, usually more than 5 observations in each group.

2.8. Chi-Square Test

i. It is a non-parametric test of hypothesis testing.

ii. As a non-parametric test, chi-square can be used:

   a) test of goodness of fit.

   b) test of independence of two variables.

iii. Helps in assessing the goodness of fit between a set of observed and those expected theoretically

iv. It makes a comparison between the expected frequencies and the observed frequencies.

v. Greater the difference, the greater is the value of chi-square.

vi. If there is no difference between the expected and observed frequencies, then the value of chi-square is equal to zero.

vii. It is also known as the “Goodness of fit test” which determines whether a particular distribution fits the observed data or not.

viii. Chi-square as a parametric test is used as a test for population variance based on sample variance.

ix. It is calculated as: $\chi^2 = \Sigma \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

2.9. Regression Analysis

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

2.9.1. Multiple Linear Regression

Multiple linear regression is used to estimate the relationship between two or more explanatory (independent) variables and one response (dependent) variable. It is also known simply as multiple regression. It is an extension of simple linear regression and allows you to estimate how a dependent variable changes as the independent variable changes

You can use multiple linear regression when you want to know:

i. How strong the relationship is between two or more independent variables and one dependent variable (e.g., how rainfall, temperature, and amount of fertilizer added affect crop growth).

ii. The value of the dependent variable at a certain value of the independent variables (e.g., the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

2.9.2. Assumptions of multiple linear regression:

i. ***Homogeneity of variance (homoscedasticity):*** the size of the error in our prediction doesn’t change significantly across the values of the independent variable, this means that the variance of residuals should be the same at each level of the independent variable.

ii. ***Independence of observations**:* the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.

i. ***Normality**:* The data follows a normal distribution.
ii. ***Linearity**:* The line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor

2.9.3. Coefficient of Determination

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

R^2 by itself can't this be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

The formula for a multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$$Y = \ is\ the\ dependent\ variable$$

$$X_1, X_2, \ldots X_p \ = \ are\ the\ independent\ variables$$

$\beta_0 = is\ the\ intercept$

$$\beta_1, \beta_2, \ldots \beta_p \ = \ are\ the\ slope\ coefficients\ for\ each\ explanatory\ variables$$

$$\varepsilon = is\ the\ error\ term$$

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

i. The regression coefficients that lead to the smallest overall model error.

ii. The t statistic of the overall model.

iii. The associated p value (how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

2.10. Logistic Regression

When we want to understand the relationship between one or more predictor variables and a continuous response variable whose distribution is normal, we often use linear regression.

However, when the response variable is categorical, one can instead use logistic regression. Logistic regression is a type of classification algorithm because it attempts to "classify" observations from a dataset into distinct categories.

Logistic regression answers the question in the form of yes / no.

Here are a few examples of when we might use logistic regression:

i. We want to use *credit score* and *bank balance* to predict whether or not a given customer will default on a loan. (Response variable = "Default" or "No default").

ii. We want to use *average rebounds per game* and *average points per game* to predict whether or not a given basketball player will get drafted into the NBA (Response variable = "Drafted" or "Not Drafted").

iii. We want to use *square footage* and the number *of bathrooms* to predict whether or not a house in a certain city will be listed at a selling price of $200k or more. (Response variable = "Yes" or "No").

2.10.1. Assumptions of Logistic Regression

i. **The response variable is binary:** It is assumed that the response variable can only take on two possible outcomes (e.g., 0 or 1, yes or no).

ii. **The observations are independent:** It is assumed that the observations in the dataset are independent of each other. That is, the observations should not come from repeated measurements of the same individual or be related to each other in any way.

iii. **No Multicollinearity:** There should be little or no multicollinearity among the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to separate their individual effects on the dependent variable.

iv. **There are no extreme outliers:** It is assumed that there are no extreme outliers or influential observations in the dataset.

v. **There is a linear relationship between the predictor variables and the logit of the response variable**: This assumption can be tested using a Box-Tidwell test.

vi. **The sample size is sufficiently large**: Logistic regression tends to perform well with a relatively large sample size. As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable.

2.10.2. The Logistic Regression Equation

Logistic regression uses a method known as maximum likelihood estimation to find an equation of the following form:

$$ln\; ln \left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

P = probability of occurring of an event E.

β_0 = Intercept term. It represents the log-odds of the event occurring when all predictor variables are zero.

$\beta_1, \beta_2, \ldots \beta_n$ = are the coefficients for each predictor variable.

$X_1, X_2, \ldots X_n$ = are predictor variables.

Thus, when we fit a logistic regression model, we can use the following equation to calculate the probability that a given observation takes on a value of 1:

$$P(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

We then use some probability threshold to classify the observation as either 1 or 0.

For example, we might say that observations with a probability greater than or equal to 0.5 will be classified as “1” and all other observations will be classified as “0.”

2.10.3. Differences between Multiple Linear Regression and Logistic Regression

i. Type of Outcome Variable:

Multiple Linear Regression: This is used when the dependent variable (outcome) is continuous and follows normal distribution. The goal is to predict a quantitative outcome based on one or more independent variables.

Logistic Regression: This is used when the dependent variable is binary or categorical, representing two classes (e.g., 0 or 1, yes or no). Logistic regression models the probability of the event occurring.

ii. Nature of Relationship:

Multiple Linear Regression: It models the linear relationship between the independent variables and the continuous dependent variable. The equation is a linear combination of the predictor variables.

Logistic Regression: It models the log-odds (or logit) of the probability of an event occurring. The relationship is not linear, and the logistic function is used to transform the linear combination of predictors into probabilities.

iii. Output Interpretation:

Multiple Linear Regression: The coefficients in multiple linear regression represent the change in the mean of the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.

Logistic Regression: The coefficients in logistic regression represent the change in the log-odds of the event happening for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.

iv. Assumptions:

Multiple Linear Regression: Assumes a linear relationship between the variables, independence of observations, homoscedasticity (constant variance of errors), and normality of errors.

Logistic Regression: Does not assume linearity, but it assumes that there is no perfect multicollinearity among the independent variables, independence of errors, and the absence of outliers.

v. Model Output:

Multiple Linear Regression: The output includes coefficients, intercept, standard errors, p-values, R-squared, and other statistics.

Logistic Regression: The output includes coefficients, intercept, standard errors, p-values, odds ratios, and sometimes measures like AIC or BIC.

2.10.4. What is Nonlinear Regression?

Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function. Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship.

The goal of the model is to make the sum of the squares as small as possible. The sum of squares is a measure that tracks how far the Y observations vary from the nonlinear (curved) function that is used to predict Y.

It is computed by first finding the difference between the fitted nonlinear function and every Y point of data in the set. Then, each of those differences is squared. Lastly, all of the squared figures are added together. The smaller the sum of these squared figures, the better the function fits the data points in the set. Nonlinear regression uses logarithmic functions, trigonometric functions, exponential functions, power functions, Lorenz curves, Gaussian functions, and other fitting methods.

i. Both linear and nonlinear regression predict Y responses from an X variable (or variables).

ii. Nonlinear regression is a curved function of an X variable (or variables) that is used to predict a Y variable.

iii. Nonlinear regression can show a prediction of population growth over time**.**

Nonlinear regression modeling is similar to linear regression modeling in that both seek to track a particular response from a set of variables graphically. Nonlinear models are more complicated than linear models to develop because the function is created through a series of approximations (iterations) that may stem from trial-and-error. Mathematicians use several established methods, such as the Gauss-Newton method and the Levenberg-Marquardt method.

Often, regression models that appear nonlinear upon first glance are actually linear. The curve estimation procedure can be used to identify the nature of the functional relationships at play in your data, so you can choose the correct regression model, whether linear or nonlinear. Linear regression models, while they typically form a straight line, can also form curves, depending on the form of the linear regression equation. Likewise, it’s possible to use algebra to transform a nonlinear equation so that it mimics a linear equation—such a nonlinear equation is referred to as “intrinsically linear.”

2.11. Trend Analysis

Trend analysis in statistics is a powerful method used to discern and understand the underlying patterns in data over time. This analytical approach is particularly valuable for identifying trends, patterns, and fluctuations within a dataset that spans different time intervals. The essence of trend analysis lies in the examination of how a variable or set of variables changes over time, aiming to reveal insights that may inform decision-making, predictions, or future planning.

To embark on trend analysis, the first step involves the careful collection of relevant data over a specified time period. This data can be sourced from diverse origins, ranging from surveys and experiments to observational studies. Once the data is acquired, a crucial step in the analytical process is data cleaning. This step ensures the accuracy of the dataset by addressing any errors, inconsistencies, or missing values that might distort the results.

Data visualization is an integral part of trend analysis, involving the graphical representation of data to discern visual patterns or irregularities. By plotting the data on graphs or charts, analysts can gain an initial understanding of the temporal evolution of the variable under investigation. Visualization serves as a preliminary step, allowing researchers to identify potential trends and outliers, prompting a more in-depth examination.

The heart of trend analysis lies in the application of various statistical techniques tailored to the nature of the data. Common methodologies include linear regression, moving averages, and exponential smoothing. These techniques quantify the observed trends, providing numerical representations of the data's trajectory over time. Linear regression, for example, helps identify the existence of a linear relationship between the independent variable (time) and the dependent variable, offering insights into the direction and strength of the trend.

Interpreting the results of trend analysis is a critical final step in deriving meaningful conclusions. Analysts must consider external factors that may influence the observed trends, ensuring a comprehensive understanding of the dataset. The insights gained from trend analysis can have far-reaching implications, informing decision-makers in fields such as finance, economics, public health, and environmental science. Ultimately, trend analysis empowers researchers and practitioners to make informed predictions and strategic decisions based on a robust understanding of how variables evolve over time.

2.12. Software and Programming languages used

2.12.1. Python

Python is a widely-used general-purpose, high-level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code.

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don’t need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc.

There are different libraries like Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly express, Psych which are used for specific purposes.

2.12.2. IBM SPSS

SPSS Statistics is a statistical software suite developed by IBM for data management, advanced analytics, multivariate analysis, business intelligence, and criminal investigation. Long produced by SPSS Inc., it was acquired by IBM in 2009. Versions of the software released since 2015 have the brand name IBM SPSS Statistics.

The software name originally stood for **Statistical Package for the Social Sciences** (**SPSS**), reflecting the original market, then later changed to Statistical Product and Service Solutions.

Its ease of use, flexibility and scalability make SPSS accessible to users of all skill levels.

2.12.3. MICROSOFT EXCEL

Microsoft Excel is a versatile and widely used data analysis tool that empowers users to organize, manipulate, and interpret data efficiently. With its intuitive spreadsheet interface, Excel facilitates tasks like sorting, filtering, and performing complex calculations. Its formula functions, pivot tables, and charting capabilities enable users to derive valuable insights from raw data, making it a go-to solution for professionals across various industries. Excel's accessibility and familiarity make it an indispensable tool for data analysis, aiding in decision-making processes and enhancing overall productivity in both personal and professional settings.

2.13. Variables Studied

These are 1644 variables in the NFHS-5 dataset , but in the current study we have incorporated only the following variables

i. ***Wealth Index:*** A wealth index is a summary measure derived from a set of variables related to household assets, amenities, and housing conditions. It aims to capture the economic well-being of individuals or households by considering a range of indicators that reflect their material wealth and living standards. this project the wealth index is categorised as (Poorest, Poor, Middle, Richer, Richest)

ii. ***Place Of Residence****:* Classified as urban and rural

iii. **Smoke**: This variable shows the status of mother who uses smoke/gutka (CATEGORICAL VARIABLES classified as Yes / No)

iv. ***Highest Education Level****:* Categorical Variable (Classified as No education=0, Primary=1, secondary=2, Higher =4)

v. ***Twin:*** This variable shows whether the child birth is twin (multiple births ie twin, triplets)
categorical variable classified as (single, second multiple, third multiple)

vi. ***Pregnancy Duration****:* This variable shows the duration of pregnancy (in months)

vii. ***Birth Weight***: The weight of a newborn infant measured in grams or pounds at the time of birth

viii. ***Total Children Ever Born***: The total number of children ever born to a woman, including both living and deceased children. (classified as 0=1, 1= 2 to 3, 2=4 or more)

ix. ***Breastfeeding****:* The practice of feeding infants with breast milk, either exclusively or in combination with other foods or liquids (categorical variable classified as (Yes/NO)

x. ***Drugs For Intestinal Parasites:*** The use of medications or drugs to treat or prevent infections caused by intestinal parasites. (Categorical Variables classified as
No=0, Yes =1, Don't know =2)

xi. ***Received Health Checkups****:* Whether an individual has undergone regular health checkups or examinations for monitoring and maintaining health.
(Categorical Variable classified as Yes / No)

xii. ***Received Nutrition****:* Whether an individual has received nutritional support or interventions, such as dietary supplements or counseling.

xiii. ***Antenatal Care****:* The healthcare and medical attention received by pregnant women before childbirth, including prenatal checkups and interventions.
(Categorical variable classified as (Yes/No))

xiv. ***Age In 5 Yr Group***: Categorization of individuals into age groups, each spanning five years, to analyse and present age-related patterns

xv. ***Preceding birth interval****:* The time duration between the birth of a previous child and the conception of the subsequent child.

xvi. ***Respondent weight****:* The weight of the survey or study participant, measured in kilograms or another relevant unit.

xvii. ***Wanted pregnancy when became pregnant****:* categorical variable classified as (Then, Later, no more)

xviii. ***Child is alive****:* Status of the child (categorical variable classified as (Yes / No)

Variable names and labels:

i. Child is alive: child_alive
ii. Wealth Index: wi
iii. Place Of Residence: residence
iv. Smoke: smoke

v. Highest Education Level: Literacy
vi. Twin: twin
vii. Pregnancy Duration: preg_dur
viii. Birth Weight: BW
ix. Total Children Ever Born: TCEB
x. Breastfeeding: breastfeed
xi. Drugs For Intestinal Parasites: drugs
xii. Received Health Checkups: health_checkup
xiii. Received Nutrition: nut_stat
xiv. Antenatal Care: ant_care
xv. Age In 5 Yr Group: Age_5_yrgp
xvi. Preceding birth interval: preceeding_bi
xvii. Respondent weight: resp_wt
xviii. Wanted pregnancy when became pregnant: wanted_preg

Chapter-3

STATISTICAL ANALYSIS

3.1. Objective

Objective of the study

The objectives of the study are:

1. To Determine predictors that have a significant effect on Infant Mortality.
2. To Assess the Association Between Maternal Health Factors and Infant Mortality.
3. To Examine association between Infant Mortality and various socio-economic disparities.
4. To Develop a predictive model for infant mortality.

3.2. Exploratory data analysis

Exploratory Analysis gives the basic idea about the data we are analyzing.

The table given below gives mean, minimum, maximum, std. deviation.

Descriptive Statistics	N	Minimum	Maximum	Mean	Std. Deviation
Birth weight in grams	101167	500	9980	2809.39	562.547
Preceeding birth Interval (months)	101167	5	221	34.07	15.400
Respondent's weight in kilograms	101167	20.0	107.8	59.001	11.0371

In our dataset the above three variables are viz; birth weight, birth internal and respondent weight continuous variables and the rest of the variables which are incorporated in the current study are categorical in nature.

Given below are the disturbance of all the categorical variables,

Type of place of residence	Frequency	Percent	Valid Percent	Cumulative Percent
Urban	18901	17.0	17.0	17.0
Rural	92484	83.0	83.0	100.0
Total	111385	100.0	100.0	

Highest educational level	Frequency	Percent	Valid Percent	Cumulative Percent
No education	21872	19.6	19.6	19.6
Primary	14140	12.7	12.7	32.3
Secondary	60734	54.5	54.5	86.9
Higher	14639	13.1	13.1	100.0
Total	111385	100.0	100.0	

Drugs consumed for intestinal parasites during pregnancy	Frequency	Percent	Valid Percent	Cumulative Percent
No	70693	63.5	63.5	63.5
Yes	39467	35.4	35.4	98.9
Don't know	1225	1.1	1.1	100.0
Total	111385	100.0	100.0	

Wealth index within state	Frequency	Percent	Valid Percent	Cumulative Percent
Poorest	27408	24.6	24.6	24.6
Poorer	26149	23.5	23.5	48.1
Middle	24082	21.6	21.6	69.7
Richer	20614	18.5	18.5	88.2
Richest	13132	11.8	11.8	100.0
Total	111385	100.0	100.0	

Did you receive any of the following benefits: Health check-ups?	Frequency	Percent	Valid Percent	Cumulative Percent
No	12772	11.5	11.5	11.5
Yes	98613	88.5	88.5	100.0
Total	111385	100.0	100.0	

Pregnancy duration	Frequency	Percent	Valid Percent	Cumulative Percent
4	232	.2	.2	.2
5	16	.0	.0	.2
6	92	.1	.1	.3
7	809	.7	.7	1.0
8	11752	10.6	10.6	11.6

9	90559	81.3	81.3	92.9
10	7925	7.1	7.1	100.0
Total	111385	100.0	100.0	

Smoke status	Frequency	Percent	Valid Percent	Cumulative Percent
No	109419	98.2	98.2	98.2
Yes	1966	1.8	1.8	100.0
Total	111385	100.0	100.0	

Child is twin	Frequency	Percent	Valid Percent	Cumulative Percent
Single birth	110411	99.1	99.1	99.1
2nd of multiple	965	.9	.9	100.0
3rd of multiple	9	.0	.0	100.0
Total	111385	100.0	100.0	

Sex of child	Frequency	Percent	Valid Percent	Cumulative Percent
Male	59803	53.7	53.7	53.7
Female	51582	46.3	46.3	100.0
Total	111385	100.0	100.0	

Child is alive	Frequency	Percent	Valid Percent	Cumulative Percent
No	2519	2.3	2.3	2.3
Yes	108866	97.7	97.7	100.0
Total	111385	100.0	100.0	

Received antenatal care for pregnancy	Frequency	Percent	Valid Percent	Cumulative Percent
No	4023	3.6	3.6	3.6
Yes	107362	96.4	96.4	100.0
Total	111385	100.0	100.0	

Wanted pregnancy when became pregnant	Frequency	Percent	Valid Percent	Cumulative Percent
Then	104084	93.4	93.4	93.4
Later	3828	3.4	3.4	96.9
No more	3473	3.1	3.1	100.0
Total	111385	100.0	100.0	

Currently breastfeeding	Frequency	Percent	Valid Percent	Cumulative Percent
No	41673	37.4	37.4	37.4
Yes	69712	62.6	62.6	100.0
Total	111385	100.0	100.0	

Age in 5-year groups	Frequency	Percent	Valid Percent	Cumulative Percent
15-19	167	.1	.1	.1
20-24	4033	3.6	3.6	3.8
25-29	14364	12.9	12.9	16.7
30-34	21314	19.1	19.1	35.8
35-39	24959	22.4	22.4	58.2
40-44	22948	20.6	20.6	78.8
45-49	23600	21.2	21.2	100.0
Total	111385	100.0	100.0	

Total children ever born	Frequency	Percent	Valid Percent	Cumulative Percent
1	37445	33.6	33.6	33.6
2 to 3	59178	53.1	53.1	86.7
4 or more	14762	13.3	13.3	100.0
Total	111385	100.0	100.0	

3.3. Graphical Visualization

A graphical representation is a visual representation of data statistics-based results using graphs, plots, and charts. This kind of representation is more effective in understanding and comparing data than seen in a tabular form. Graphical representation helps to qualify, sort, and present data in a method that is simple to understand for a larger audience. This visual representation helps in clarity, comparison, and understanding of numerical data

3.3.1. Distribution of continuous variables

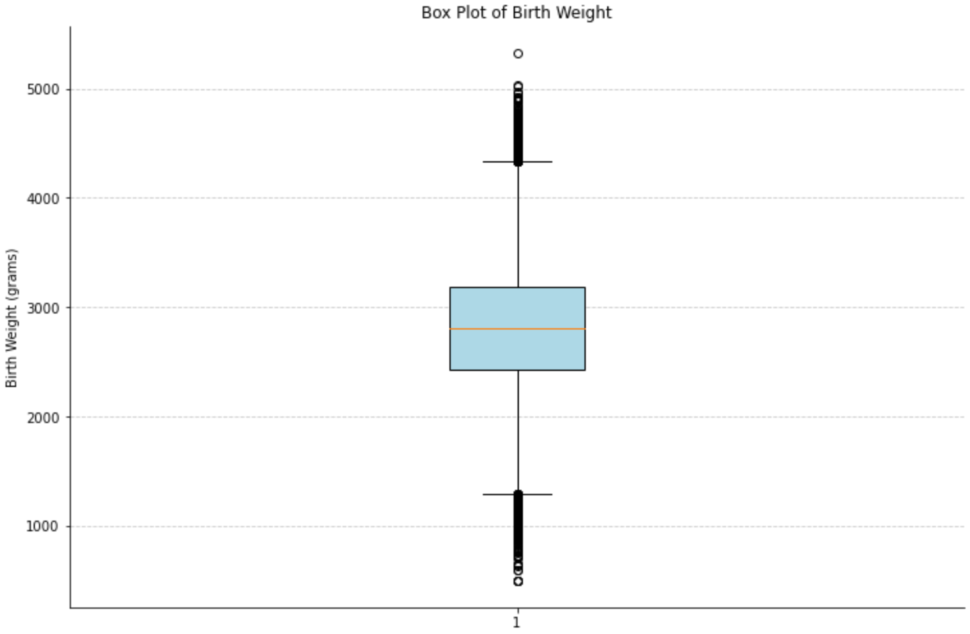


Fig 3.3.1 (a)

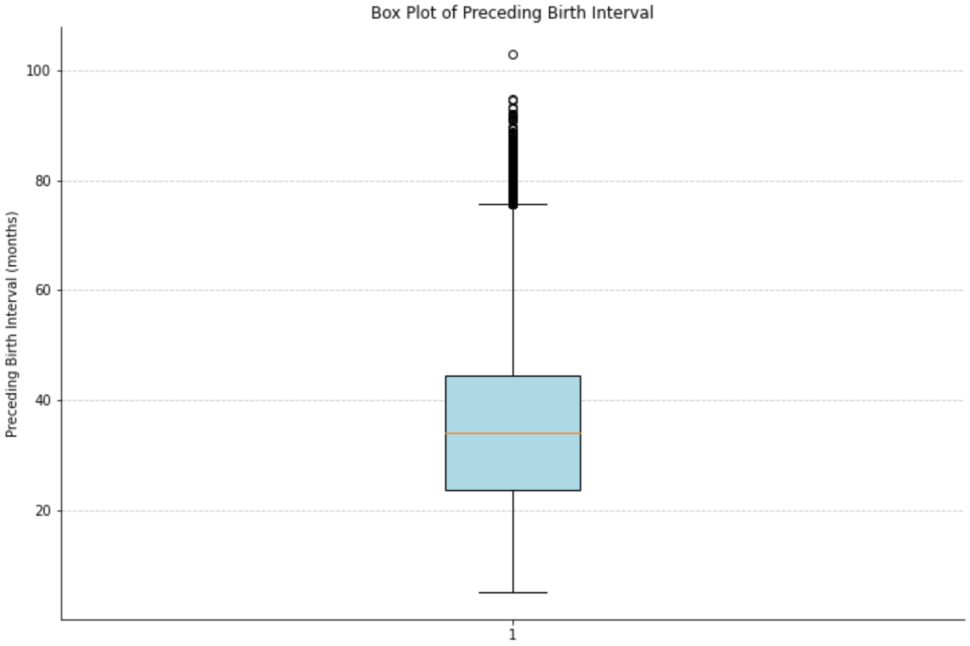


Fig 3.3.1 (b)

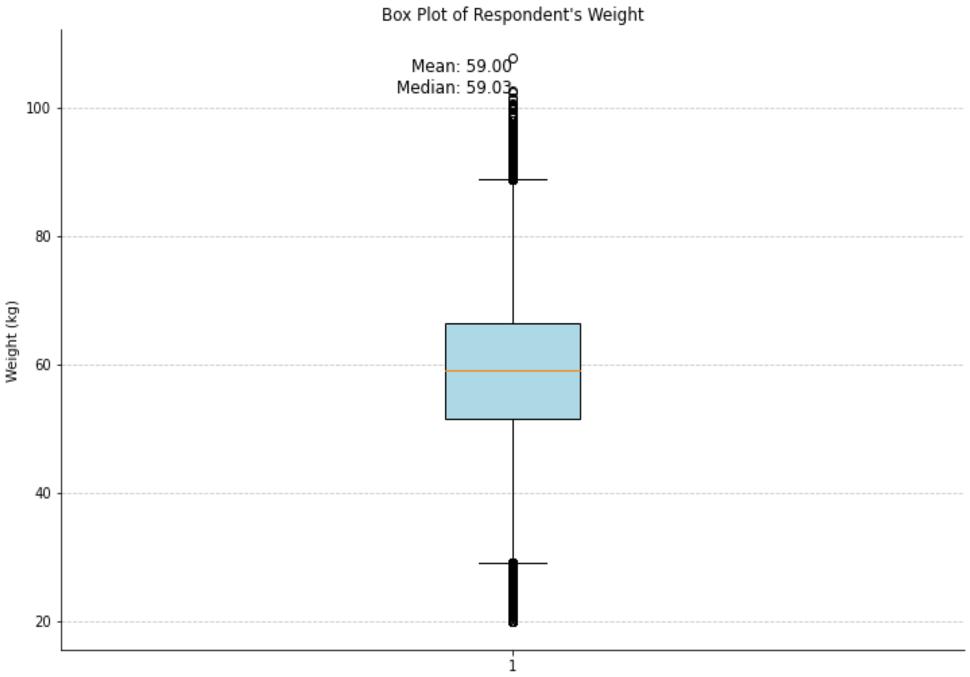


Fig 3.3.1 (c)

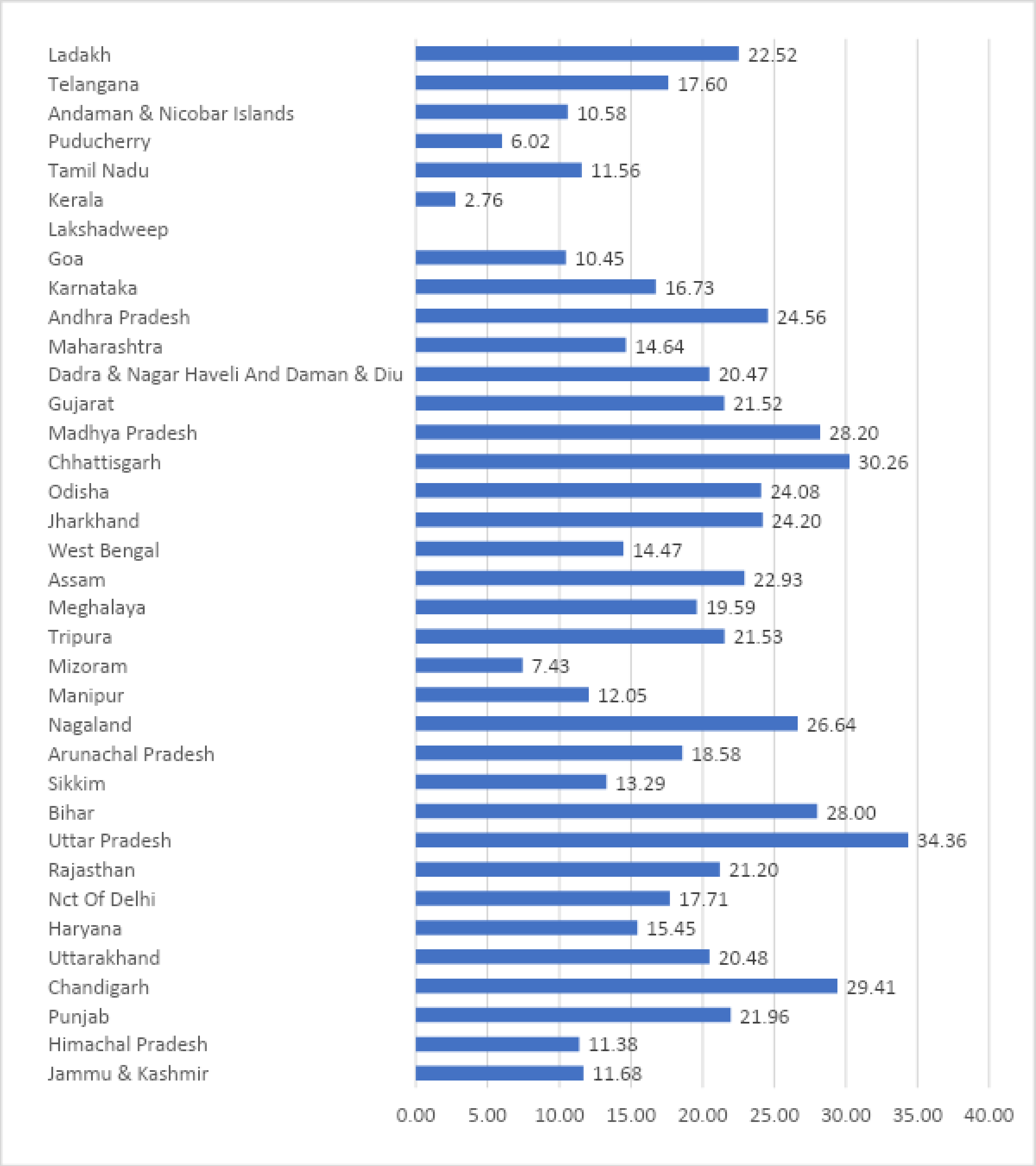
Interpretation

From **Fig 3.3.1 (a)** The median birth weight is approximately 2810 grams. The IQR is from about 2500 to 3100 grams. The lower whisker extends down to around 1500 grams, and the upper whisker goes up to approximately 4000 grams. There are outliers on both the lower and upper ends, with some birth weights below 1500 grams and some above 4000 grams

From **Fig 3.3.1 (b)** The median (Q2) is approximately 34 months. The interquartile range (IQR), which is the distance between Q1 and Q3, is roughly from 28 to 40 months. There are several outliers indicated by the individual points above the upper whisker, which extends to around 80 months. The data seems to be right-skewed, given the presence of outliers on the higher end.

From **Fig 3.3.1 (c)** The median weight is approximately 59 kg, The IQR is from around 50 to 65 kg. The whiskers extend from approximately 40 to 80 kg, indicating the range of the bulk of the data.

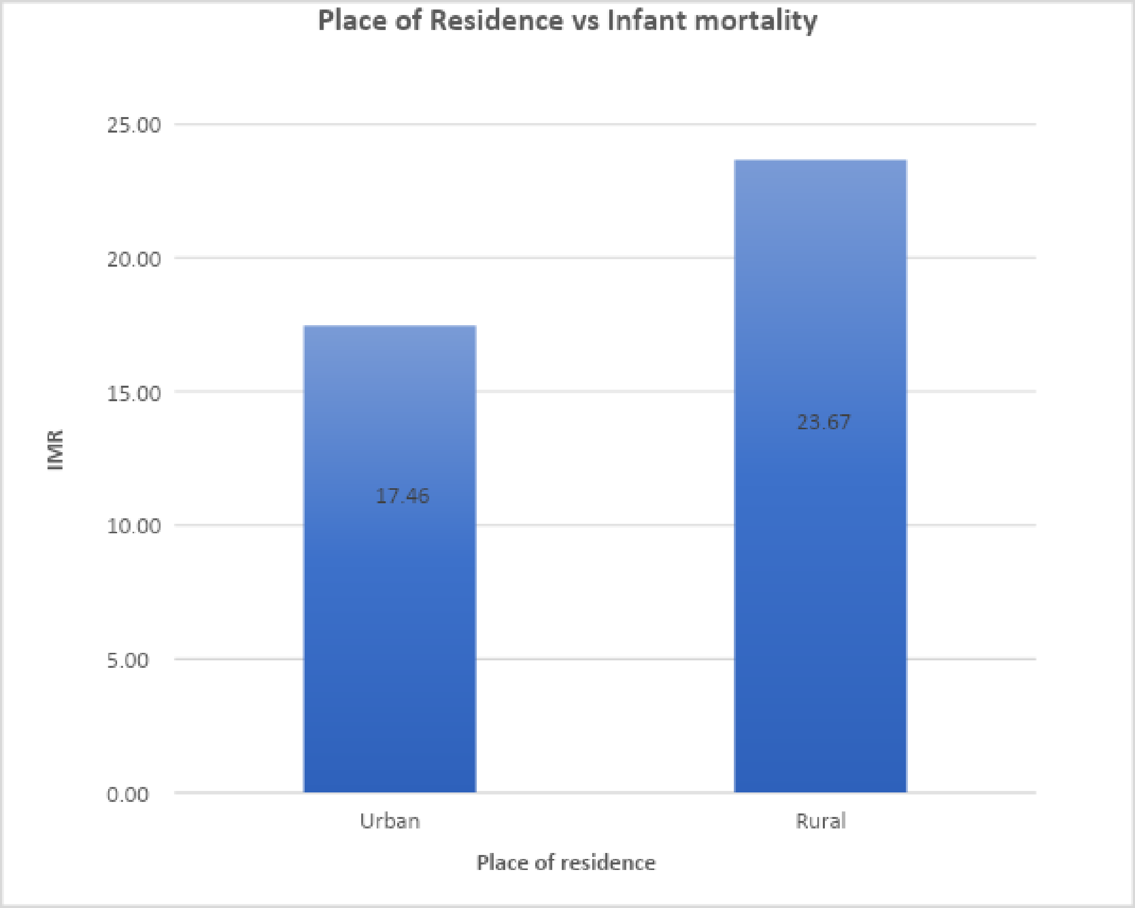
3.3.2. Infant Mortality Rate (statewise) over the years 2011-2021



Interpretation

From the graph it is evident that the state of Uttar Pradesh has the highest Infant Mortality during the years 2011-2021 followed by the states of Chhattisgarh and Madhya Pradesh. Kerala has the lowest Infant Mortality Rate followed by Mizoram, the union territories Lakshadweep also reporting less rates. (zero)

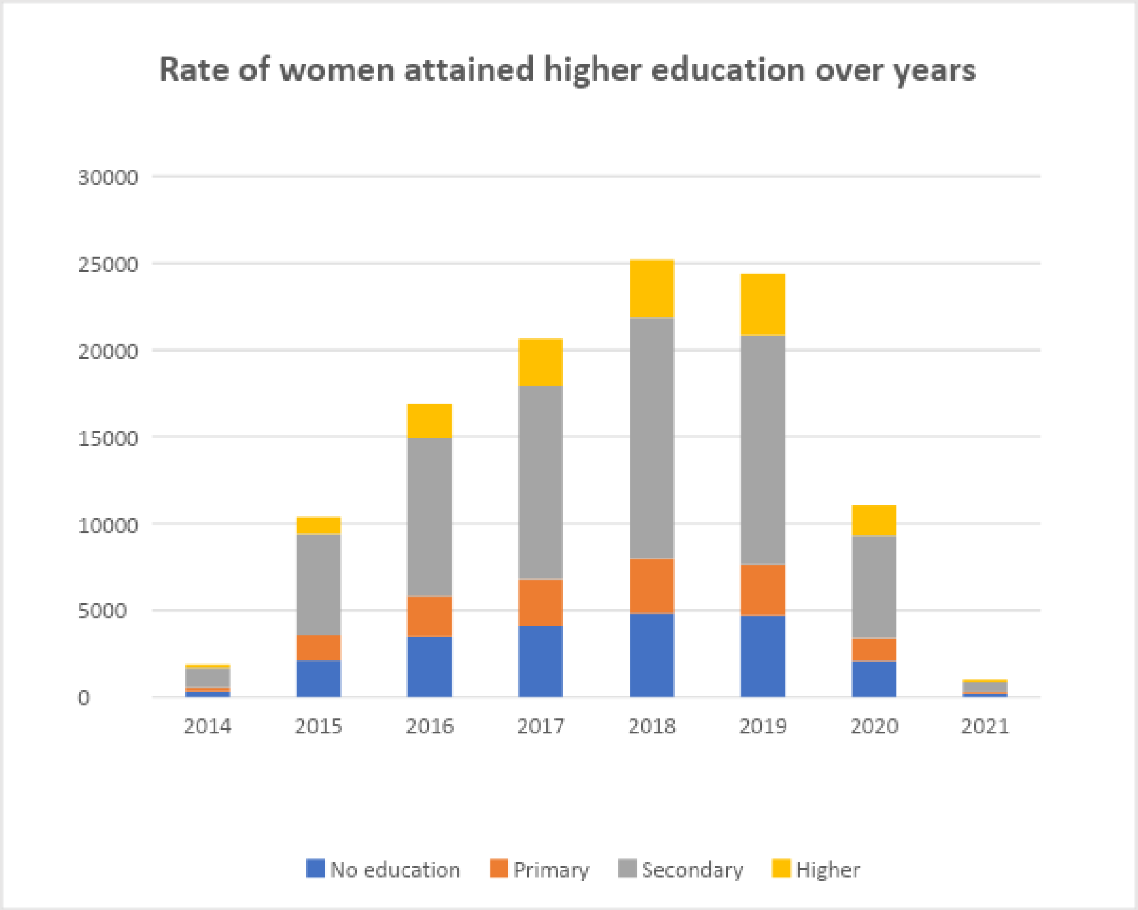
3.3.3. Place of residence and Infant Mortality Rate.



Interpretation

From the bar graph it is evident that Infant mortality in rural areas is higher than that of urban areas. The higher rate in rural areas indicates challenges in accessing quality medical services, potentially stemming from limited infrastructure or inadequate healthcare resources

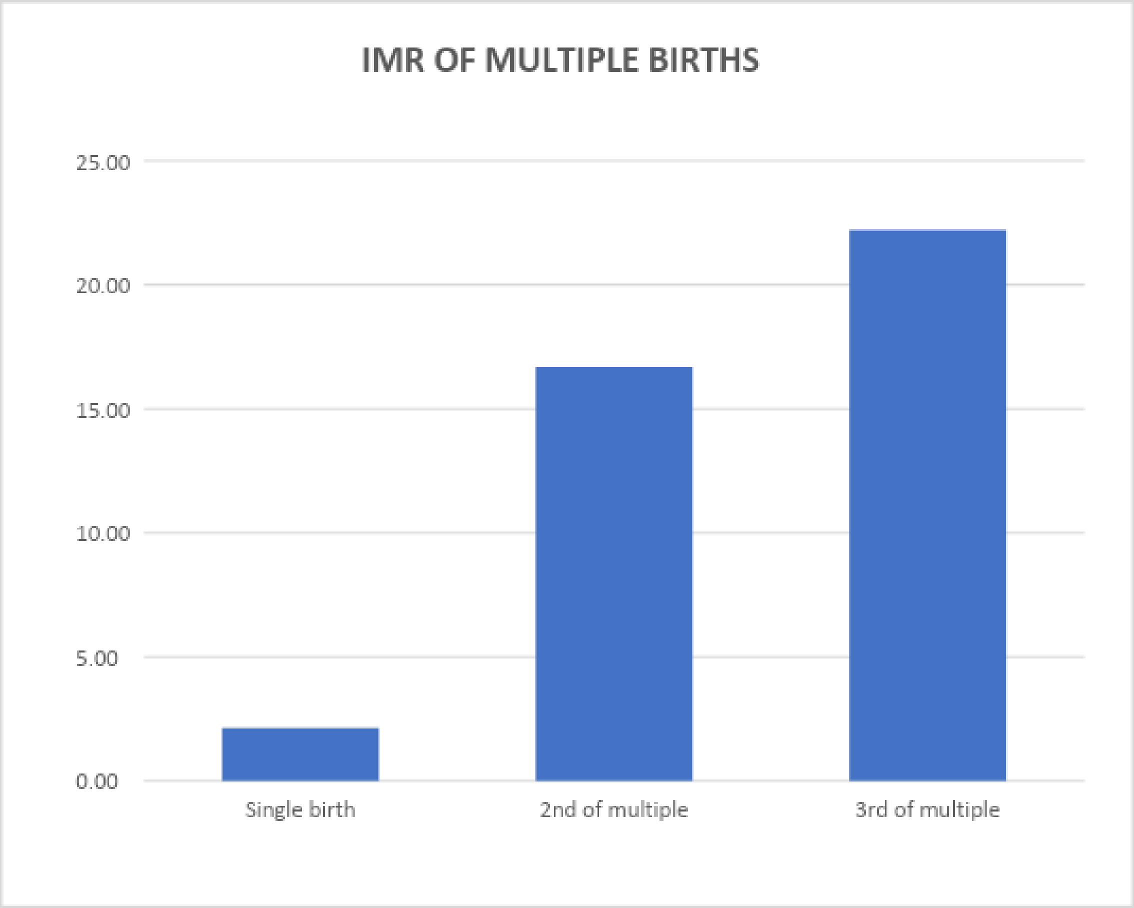
3.3.4. Rate of women attained highest education level



Interpretation

In the graph, it is evident that the percentage of women attaining higher education has been increasing year by year.

3.3.5. IMR V/s Multiple childs

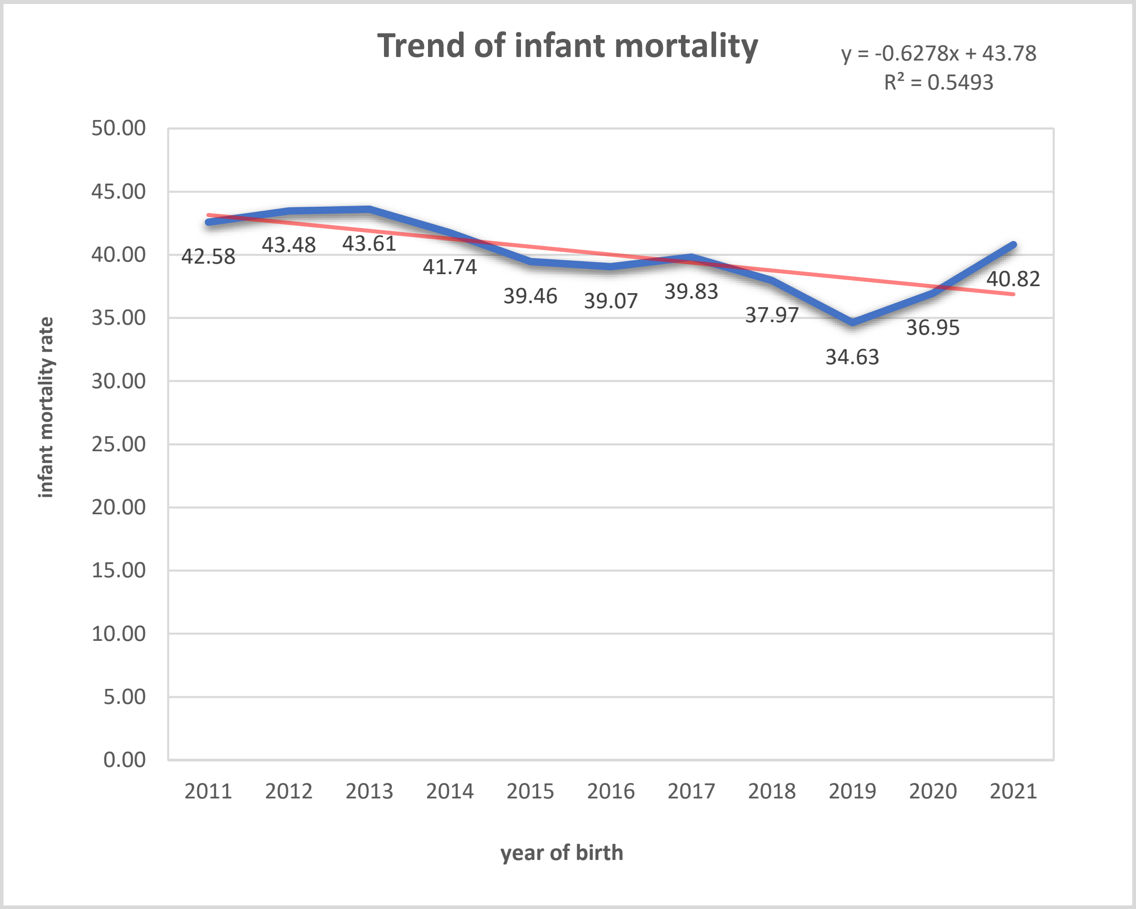


Interpretation

Single births exhibit the lowest mortality, suggesting a favourable environment for infant health. Twin births follow, indicating a slightly higher but still relatively lower risk. Conversely, multiple births, such as triplets, show the highest mortality rate.

3.4. Trend Analysis

3.4.1. Trend of infant mortality Rate in India over year 2011-2021



Interpretation

From the graph it is evident that there is a clear decrease in trend in Infant Mortality Rate from 2013 to 2019 and the rate appears to be increasing after 2019.On fitting a trend line, we get

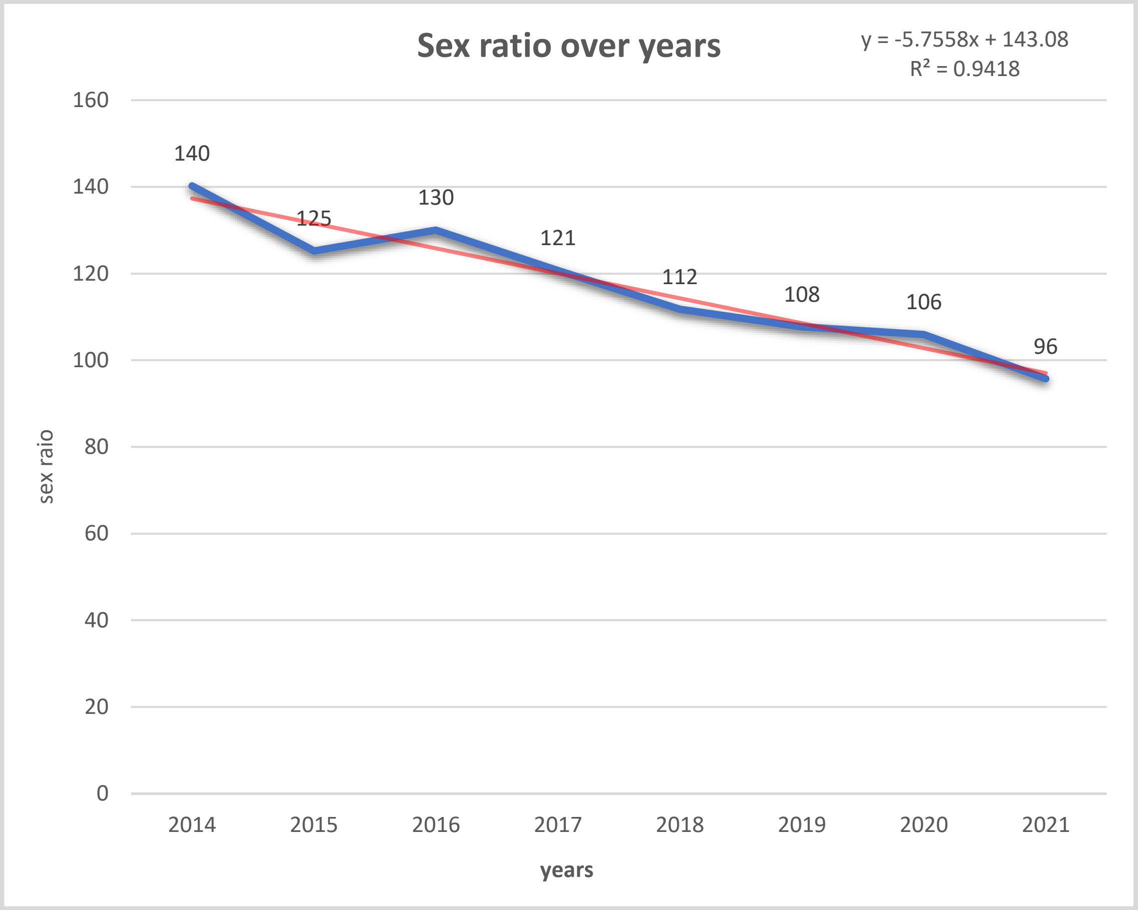
$y=-0.6278x+43.78$, R-squared = 0.5493

The negative slope (−0.6278) implies that there is a decreasing trend, which means the overall tendency of IMR is to decline with respect to time.

The intercept (43.78) is the estimated value of y when x is zero.

The R-squared value of 0.5493 indicates that approximately 54.93% of the variability in IMR can be explained by the fitted trend line.

3.4.2. Sex Ratio Over Years 2014-2021



Interpretation

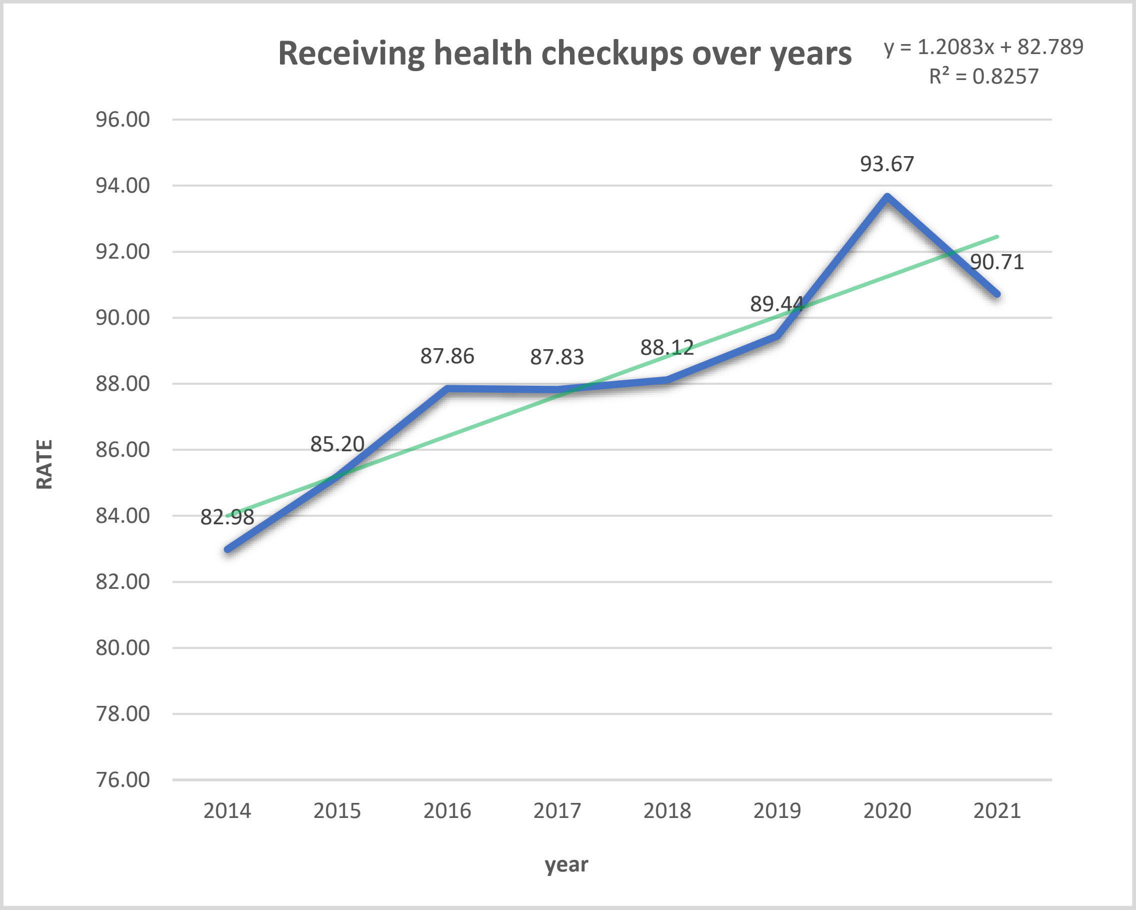
Over the period from 2014 to 2021, the sex ratio in India has witnessed a concerning trend of decline. The graph illustrates a persistent decrease in the ratio of females to males, reflecting a demographic imbalance. This decline raises crucial societal and gender-related concerns that require close attention.

y=−5.778x+143.08, R-squared = 0.9418

The strong negative slope −5.778 suggests a significant negative relationship between x and y. and a downward trend.

The high R-squared value of 0.9418 indicates that approximately 94.18% of the variability in y is explained by the linear regression model. This suggests a strong fit.

3.4.3. Trend of Receiving Health Checkups Over the years 2014-2021.



Interpretation

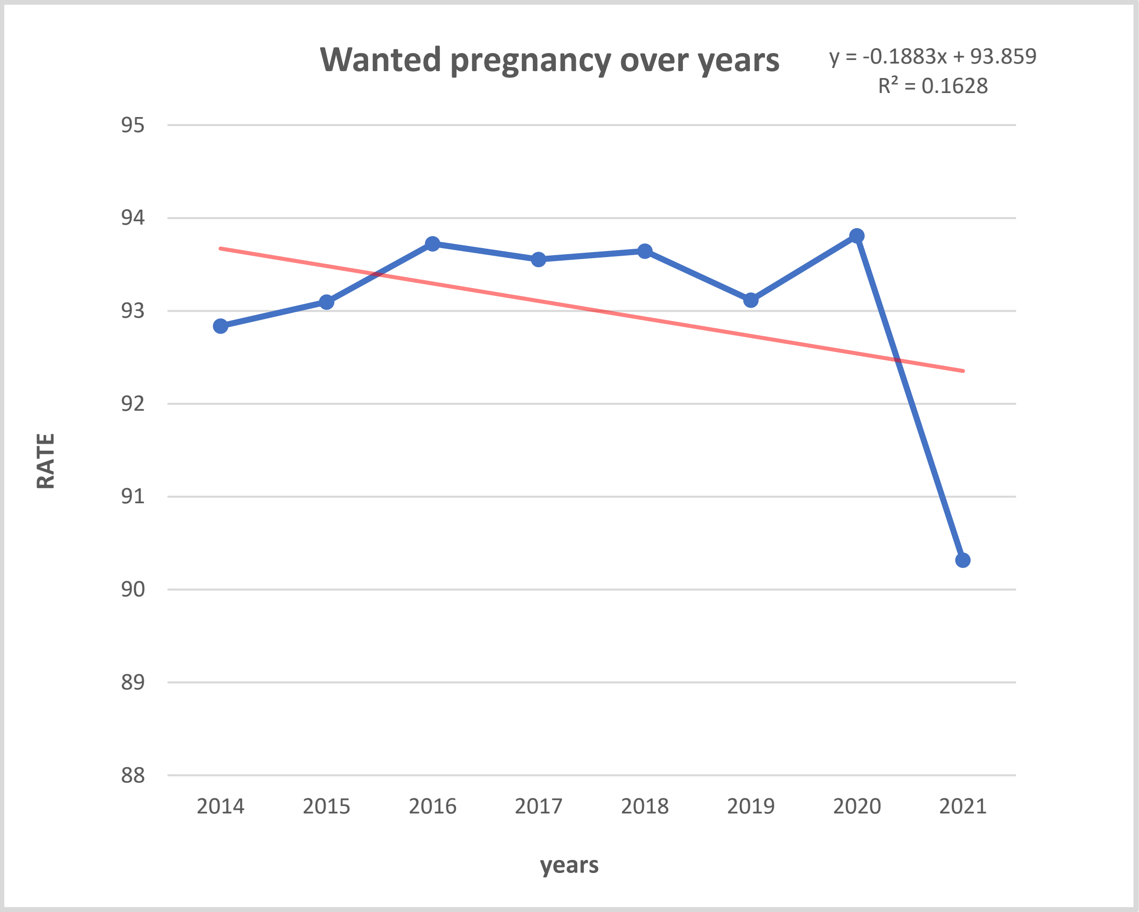
From the graph it is evident that there is a clear increasing trend in receiving health benefits for pregnant women from 2014 to 2020 and a start of decreasing trend from 2021. The trend of decreasing in that particular year can be because of the limitations in accessing healthcare at the time of COVID-19 pandemic.

y=1.2083x+82.789, R-squared = 0.8257:

The positive slope (1.2083) indicates a positive relationship between x and y. and an increasing trend.

The R-squared value of 0.8257 suggests that approximately 82.57% of the variability in y can be explained by the linear regression model.

3.4.4. Percentage of Wanted pregnancies over the years 2014 – 2021



Interpretation

From the graph we can see a drastic decrease in the percentage of wanted pregnancies in the year 2020-21 Amid the COVID-19 pandemic, women's hesitancy towards pregnancy likely stems from heightened concerns about maternal and fetal health.

$y = -0.1883x + 93.859$, R-squared = 0.1628

The small negative slope (−0.1883) suggests a weak negative relationship between x and y.

Since the data incorporated in the study have only limited data available for the year 2021.

So that the trend line for that year may be more sensitive to small changes in data points due to the lower sample size.

3.5 Comparison of Birth Weight Between Dead and Alive.

Independent Sample t-test

We intend to test the following set of Hypothesis pertaining to Birth weights of the infants using independent sample t-test

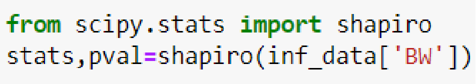
H_0: There is no significant difference between average birth weight of (survived and died) infants.

H_1: There is a significant difference between the average birth weight of the two groups.

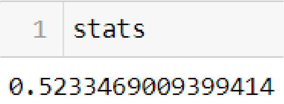
In order to apply t-test, the first and foremost assumption to check for is normality, Therefore, the normality test is applied for the variable birth weight as follows.

Normality Check

Checking the Normality using Shapiro function in python from scipy.stats package

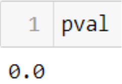


```
from scipy.stats import shapiro
stats,pval=shapiro(inf_data['BW'])
```



```
stats
```

```
0.5233469009399414
```



```
pval
```

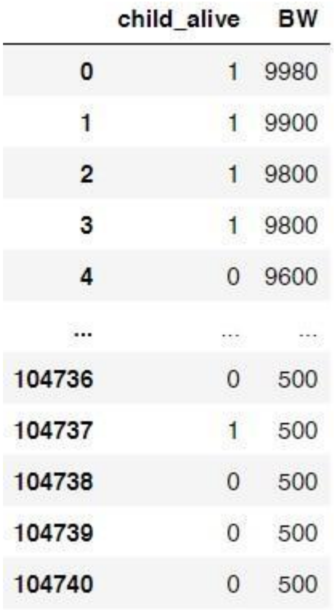
```
0.0
```

From the Shapiro test variable, we get the p value as $0 < \alpha=0.05$ (level of significance) Hence the Birth Weight data is not distributed normally Therefore we use the Non Parametric alternative of the independent sample t test which is Mann Whitney U Test. For Mann Whitney Test the set of hypothesis will be

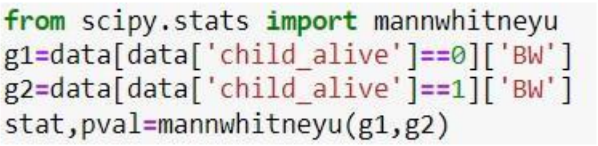
Hypothesis testing:

H_0: Median Birth Weight of survivors = Median Birth Weight of non-survivors (in grams)

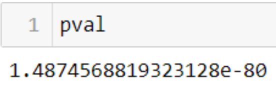
H_1: Median Birth Weight of survivors ≠ Median Birth Weight of non-survivors (in grams)



	child_alive	BW
0	1	9980
1	1	9900
2	1	9800
3	1	9800
4	0	9600
...	...	...
104736	0	500
104737	1	500
104738	0	500
104739	0	500
104740	0	500



```
from scipy.stats import mannwhitneyu
g1=data[data['child_alive']==0]['BW']
g2=data[data['child_alive']==1]['BW']
stat,pval=mannwhitneyu(g1,g2)
```



```
pval
```

```
1.4874568819323128e-80
```

Interpretation

After performing the test, we get p-value as `1.48e-8` which is less than α=0.05 (level of significance) and hence indicates a strong evidence against null-hypothesis (Median Birth Weight of non-survivors ≠ Median Birth Weight of non-survivors).

Therefore it can be a significant difference in birth weight of surviving infants and those who didn’t survive.

3.6. Comparison of Survival Probabilities

In this section, the survival probabilities of different classes/categories are compared by making use of various empirical joint and marginal distributions.

3.6.1. Type of place of Residence vs Child is Alive

Type of place of residence * Child is alive		Child is alive		Total
		No	Yes	
Type of place of residence	Urban	330	18571	18901
	Rural	2189	90295	92484
Total		2519	108866	111385

	Dead	Alive	Total
Urban	330/111385 =0.002	18571/111385 =0.166	18901/111385 =0.169
rural	2189/111385 =0.019	90295/111385 =0.810	92484/111385 =0.831
Total	2519/111385 =0.022	108866/111385 =0.976	1

$$P(Urban) = \frac{P(Death,Urban)}{P(Urban)}$$

$$= \frac{0.002}{0.169}$$

$$= 0.011$$

$$\text{P (Dead | Rural)} = \frac{P(Death,Rural)}{P(Rural)}$$

$$= \frac{0.019}{0.831}$$

$$= 0.022$$

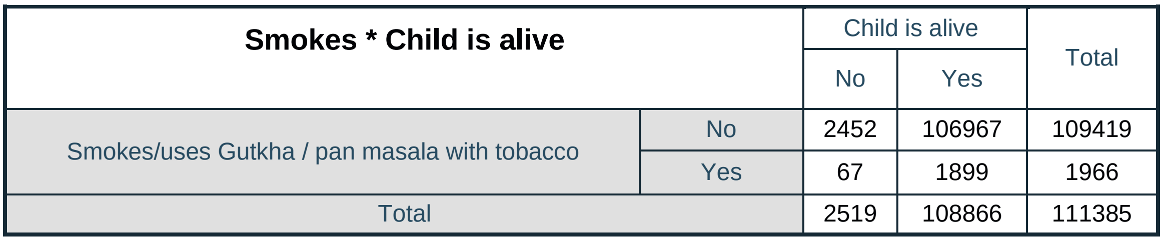
$$\frac{P(Death \ \mid Urban)}{P(Death \mid Rural)} = \frac{0.011}{0.022} = 0.5$$

$$P(Death|Urban) = 0.5 * P(Death|Rural)$$

Interpretation

An infant belonging to Urban area has 50% more chances to survive than an infant belonging to Rural area.

3.6.2. Mother's habit of smoking vs Child is Alive



Smokes * Child is alive		Child is alive		Total
		No	Yes	
Smokes/uses Gutkha / pan masala with tobacco	No	2452	106967	109419
	Yes	67	1899	1966
Total		2519	108866	111385

	Dead	**Alive**	**Total**
Non-smoker	2452/111385 =0.022	106967/111385 =0.96	109419/111385 =0.982
Smoker	67/111385 =0.0006	1899/111385 =0.017	1966/111385 =0.018
Total	2519/111385 =0.024	108866/111385 =0.976	1

$$P(Non\ Smoker) = \frac{P(Death, Non-Smoker)}{P(Non\ smoker)}$$

$$= \frac{0.022}{0.982}$$

$$= 0.0224$$

$$\text{P (Dead | Smoker)} = \frac{P(Death, Smoker)}{P(Smoker)}$$

$$= \frac{0.0006}{0.018}$$

$$= 0.03$$

$$\frac{P(Death\ \ | Non\ Smoker)}{P(Death\ | \ Smoker)} = \frac{0.0224}{0.03} = 0.74$$

$$P(Death|NonSmoker) = 0.74 * P(Death|Smoker)$$

Interpretation

From the results we can see that an infant belonging to a mother who has no smoking habit has 36% more survival chance than an infant belonging to a mother who has a smoking habit.

3.6.3. Mother wanted pregnancy vs Child is Alive

Wanted pregnancy when became pregnant * Child is alive		Child is alive		Total
		No	Yes	
Wanted pregnancy when became pregnant	Then	2307	101777	104084
	Later	109	3719	3828
	No more	103	3370	3473
Total		2519	108866	111385

Wanted Pregnancy	Dead	Alive	Total
Then	2307/111385 =0.02	101777/111385 =0.913	104084/111385 =0.934
Later	109/111385 =0.0009	3719/111385 =0.033	3838/111385 =0.0343
No more	103/111385 =0.0009	3370/111385 =0.0302	3473/111385 =0.0311
Total	2519/111385 =0.024	108866/111385 =0.976	1

$$\text{P (Dead | Wanted preg. Then)} = \frac{P(Death,Then)}{P(Wanted\ Preg.Then)}$$

$$= \frac{0.02}{0.934}$$

$$= 0.0214$$

$$\text{P (Dead | Wanted preg. Later)} = \frac{(P(Death,Later))}{P(Wanted\ Preg.Later)}$$

$$= \frac{0.0009}{0.0343}$$

$$= 0.0262$$

$$\text{P (Dead | Wanted preg. No more)} = \frac{(P(Death,No\ More))}{P(Wanted\ Preg.No\ More)}$$

$$= \frac{0.0009}{0.0311}$$

$$= 0.0298$$

$$\frac{P(Death\ \ | Wanted\ preg.Then)}{P(Death\ | Wanted\ preg.Later)} = \frac{0.0214}{0.0262} = 0.816$$

$$\frac{P(Death\ \ | Wanted\ preg.Then)}{P(Death\ | Wanted\ preg.No\ more)} = \frac{0.0214}{0.0298} = 0.71$$

$$P(Dead|Wanted\ Preg.Then) = 0.816 * P(Dead|Wanted\ Preg.Later)$$

$$P(Dead|Wanted\ Preg.Then) = 0.71 * P(Dead|Wanted\ Preg.No\ More)$$

Interpretation

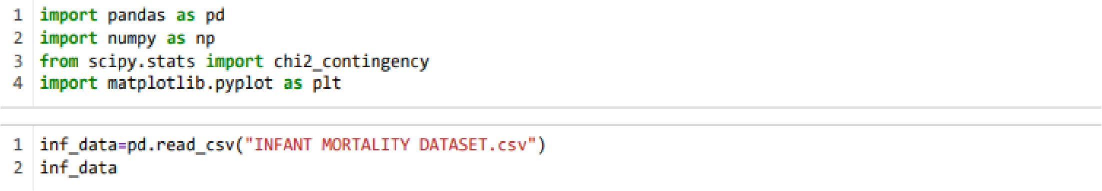
From the results it is evident that if a mother becomes pregnant by her own will, then the infant has 19% more survival chance than an infant born from a mother who had an unplanned pregnancy.

Also, the infant has 29% more chance of survival than the infant whose mother never wanted to be pregnant any more.

3.7 Exploring Associations.

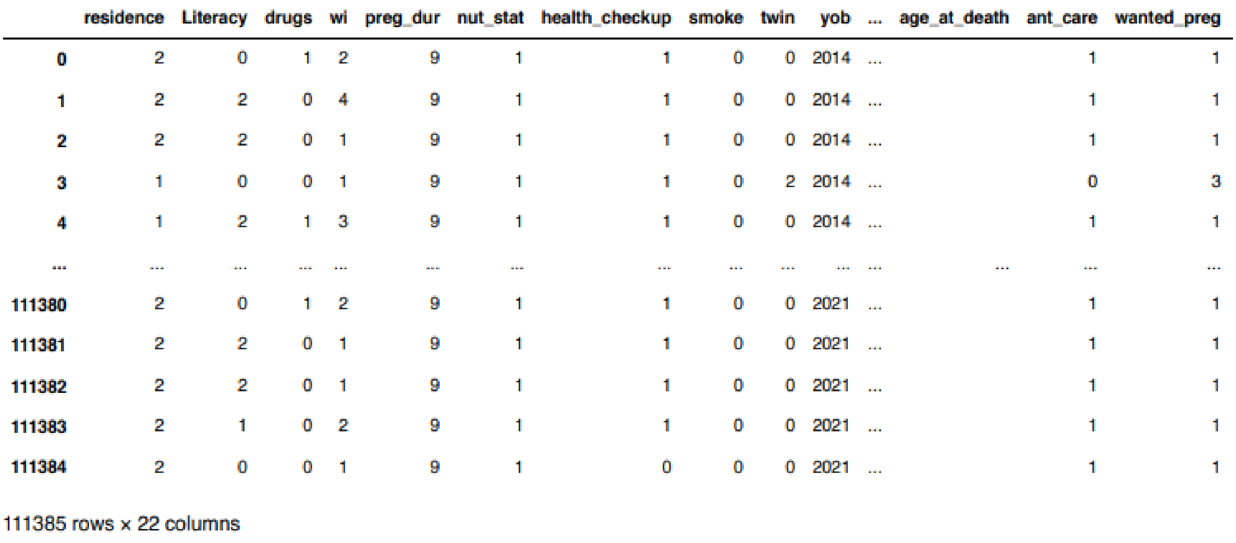
In this section Chi-Square test is used to determine whether there is any association between various categorical variables using python.

Firstly, import the necessary libraries and read the data as inf_data



```python
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
```

```python
inf_data=pd.read_csv("INFANT MORTALITY DATASET.csv")
inf_data
```



	residence	Literacy	drugs	wi	preg_dur	nut_stat	health_checkup	smoke	twin	yob	...	age_at_death	ant_care	wanted_preg
0	2	0	1	2	9	1	1	0	0	2014	...		1	1
1	2	2	0	4	9	1	1	0	0	2014	...		1	1
2	2	2	0	1	9	1	1	0	0	2014	...		1	1
3	1	0	0	1	9	1	1	0	2	2014	...		0	3
4	1	2	1	3	9	1	1	0	0	2014	...		1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
111380	2	0	1	2	9	1	1	0	0	2021	...		1	1
111381	2	2	0	1	9	1	1	0	0	2021	...		1	1
111382	2	2	0	1	9	1	1	0	0	2021	...		1	1
111383	2	1	0	2	9	1	1	0	0	2021	...		1	1
111384	2	0	0	1	9	1	0	0	0	2021	...		1	1

111385 rows × 22 columns

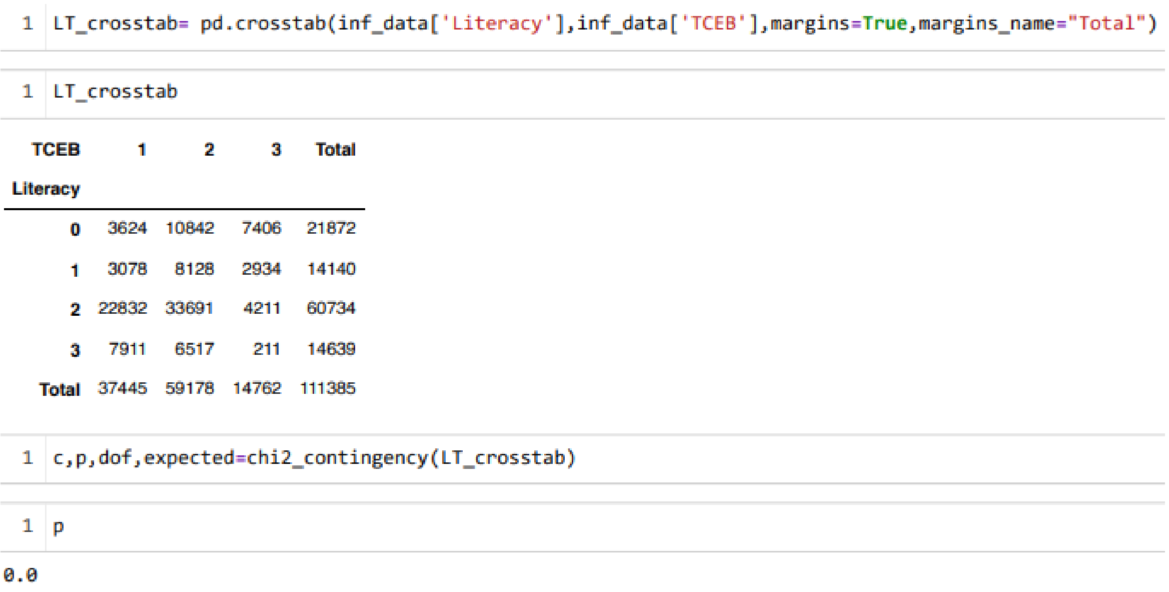
3.7.1. Literacy V/s total children ever born

H_0: Literacy and total children ever born are independent

H_1: Literacy and total children ever born are dependent

In order to check the test for independence, cross tabulate the variables literacy and total children ever born.

Test using chi2 contingency ()



```
LT_crosstab= pd.crosstab(inf_data['Literacy'],inf_data['TCEB'],margins=True,margins_name="Total")
```

```
LT_crosstab
```

TCEB	1	2	3	Total
Literacy				
0	3624	10842	7406	21872
1	3078	8128	2934	14140
2	22832	33691	4211	60734
3	7911	6517	211	14639
Total	37445	59178	14762	111385

```
c,p,dof,expected=chi2_contingency(LT_crosstab)
```

```
p
```

```
0.0
```

Interpretation

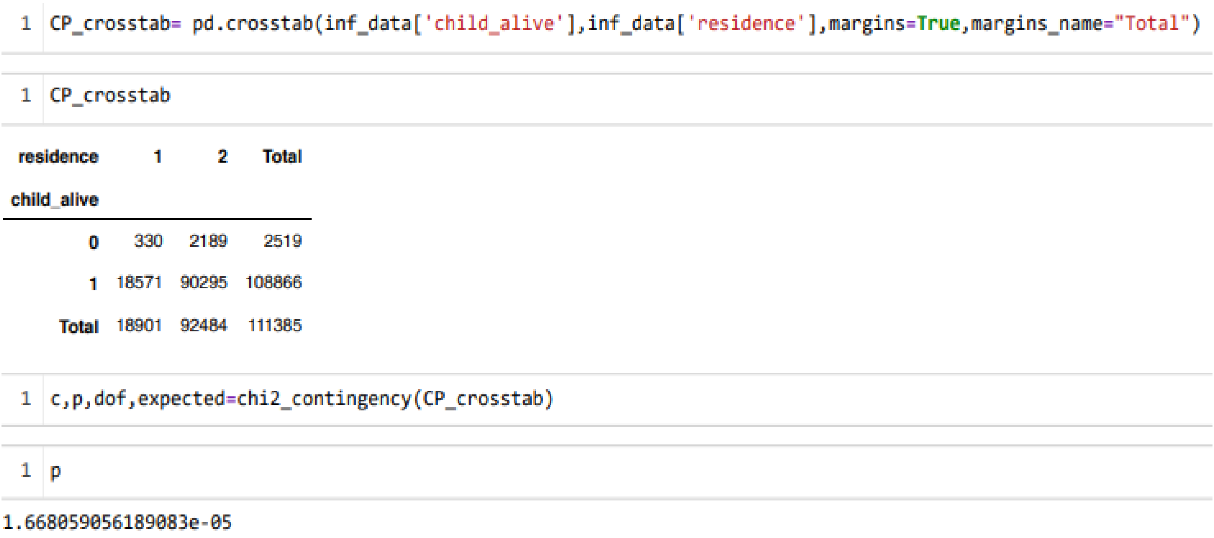
From the above table we can see that p value= 0 < alpha =0.05

i.e; We Reject the null hypothesis. So, we can say that there is an association between literacy and total children ever born

3.7.2. Checking the Independency of Child Death and Place of Residence

H_0: Child Death and Place of Residency are independent

H_1: Child Death and Place of Residency are dependent



```
CP_crosstab= pd.crosstab(inf_data['child_alive'],inf_data['residence'],margins=True,margins_name="Total")
```

```
CP_crosstab
```

residence	1	2	Total
child_alive			
0	330	2189	2519
1	18571	90295	108866
Total	18901	92484	111385

```
c,p,dof,expected=chi2_contingency(CP_crosstab)
```

```
p
```

```
1.668059056189083e-05
```

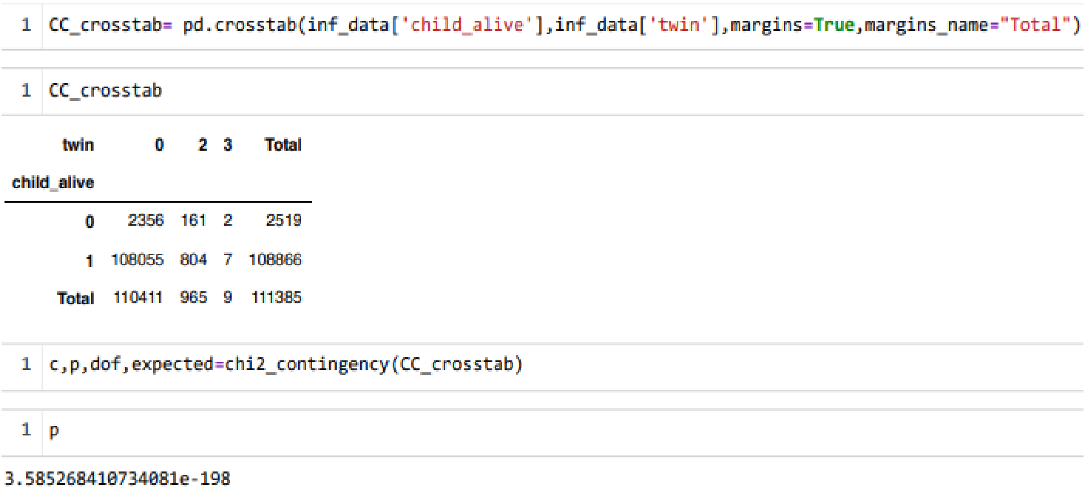
Interpretation

From the above table we can see that p value= 0 < alpha =0.05 i.e.; We Reject the null hypothesis. So, we can say that there is an association between a child alive and place of residence.

3.7.3. Child is twin v/s Child is alive

H_0**:** There is no significant association between the variable's child is twin and Child is Alive.

H_1**:** There is significant association between the variable's child being twin and Child is Alive.



```
CC_crosstab= pd.crosstab(inf_data['child_alive'],inf_data['twin'],margins=True,margins_name="Total")
```

```
CC_crosstab
```

twin	0	2	3	Total
child_alive				
0	2356	161	2	2519
1	108055	804	7	108866
Total	110411	965	9	111385

```
c,p,dof,expected=chi2_contingency(CC_crosstab)
```

```
p
```

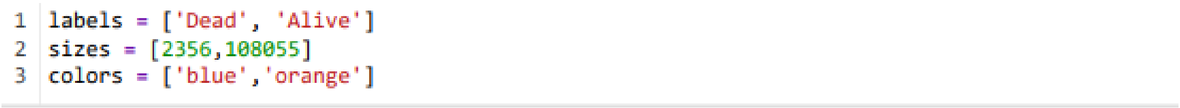
```
3.585268410734081e-198
```

Interpretation

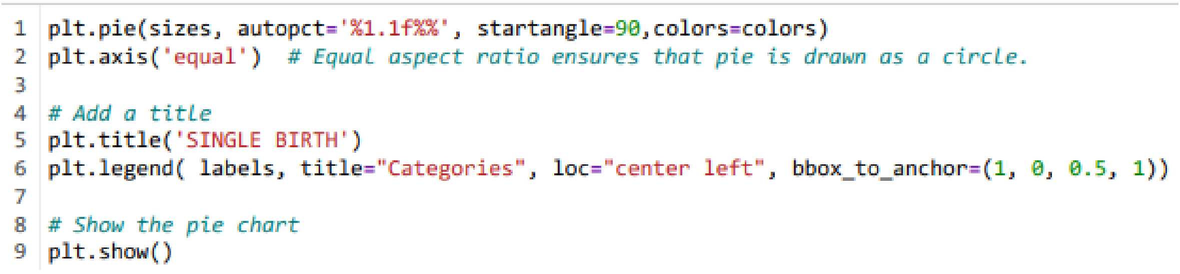
From the above table we can see that p value= 0.00 < alpha =0.05

i.e.; We Reject the null hypothesis. So, we can say that there is an association between child alive and child is twin (multiple birth)

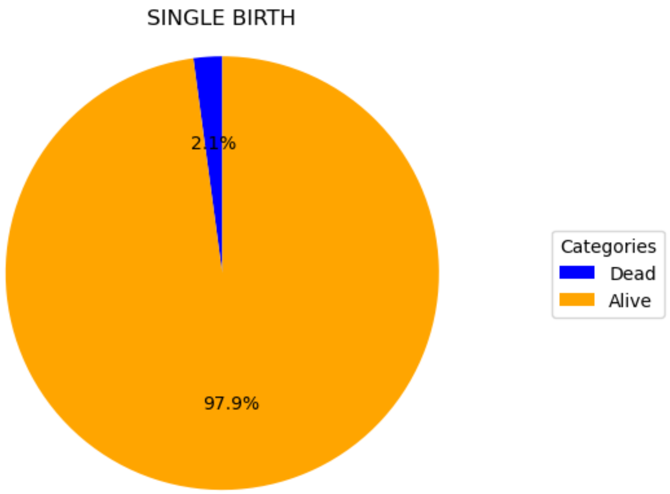
Representing the percentage of deaths in each class using pie charts



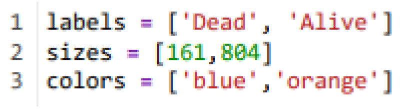
```
labels = ['Dead', 'Alive']
sizes = [2356,108055]
colors = ['blue','orange']
```



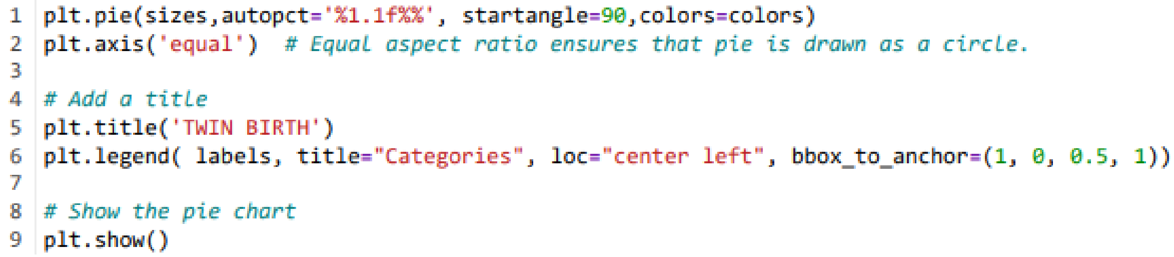
```
plt.pie(sizes, autopct='%1.1f%%', startangle=90,colors=colors)
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

# Add a title
plt.title('SINGLE BIRTH')
plt.legend( labels, title="Categories", loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))

# Show the pie chart
plt.show()
```



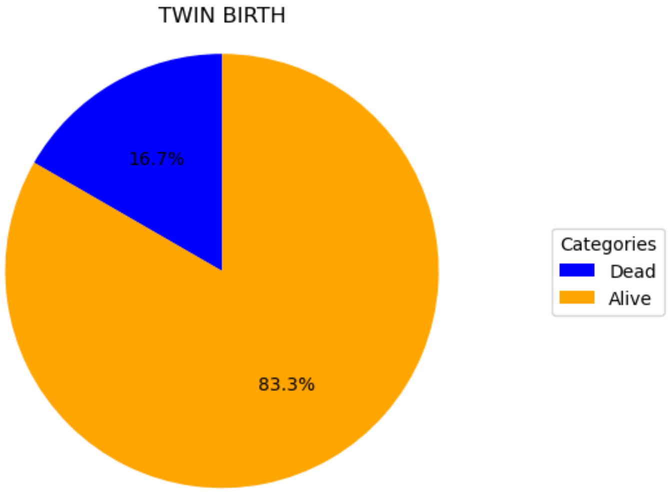
During single birth it is seen that the death rate is 2% which is very low.



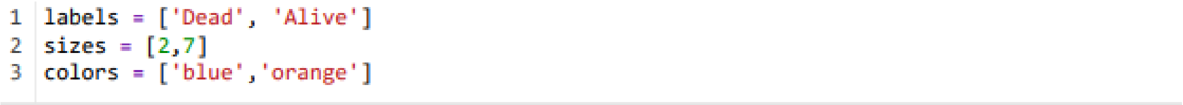
```
labels = ['Dead', 'Alive']
sizes = [161,804]
colors = ['blue','orange']
```



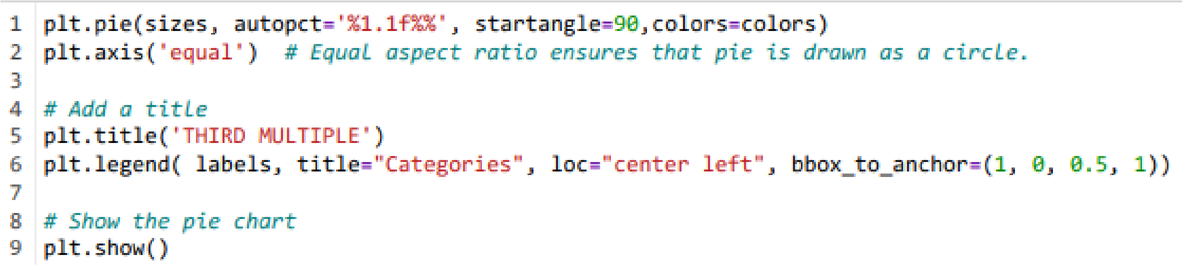
```
plt.pie(sizes,autopct='%1.1f%%', startangle=90,colors=colors)
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

# Add a title
plt.title('TWIN BIRTH')
plt.legend( labels, title="Categories", loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))

# Show the pie chart
plt.show()
```



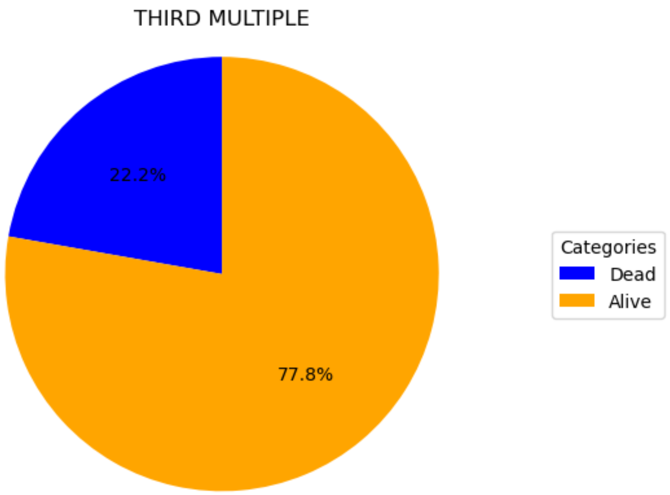
During twin birth the death rate is 17% which is higher than that of single birth.



```
labels = ['Dead', 'Alive']
sizes = [2,7]
colors = ['blue','orange']
```



```
plt.pie(sizes, autopct='%1.1f%%', startangle=90,colors=colors)
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

# Add a title
plt.title('THIRD MULTIPLE')
plt.legend( labels, title="Categories", loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))

# Show the pie chart
plt.show()
```

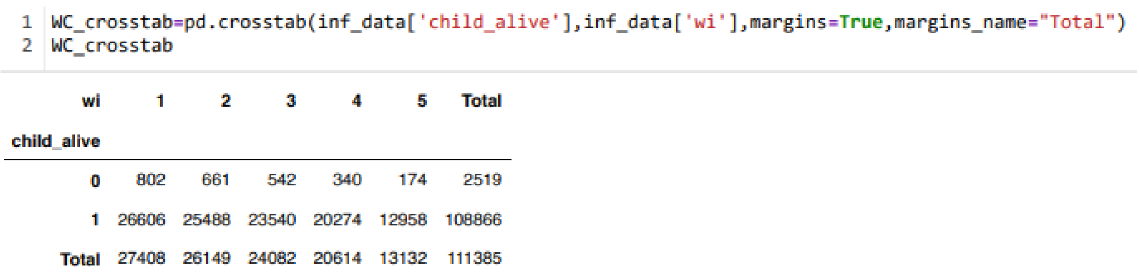


During multiple births the death rate has been increased to 22.2% which is higher than that of the other two cases.

3.7.4. Wealth index V/s child death

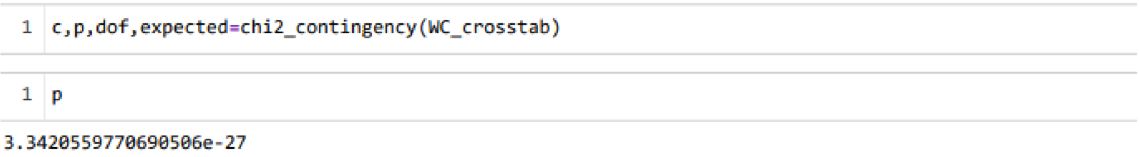
H_0: There is no significant association between wealth index groups and child births.

H_1: There is a significant association between wealth index groups and child births.



```
WC_crosstab=pd.crosstab(inf_data['child_alive'],inf_data['wi'],margins=True,margins_name="Total")
WC_crosstab
```

wi	1	2	3	4	5	Total
child_alive						
0	802	661	542	340	174	2519
1	26606	25488	23540	20274	12958	108866
Total	27408	26149	24082	20614	13132	111385



```
c,p,dof,expected=chi2_contingency(WC_crosstab)
```

```
p
```

```
3.3420559770690506e-27
```

Interpretation

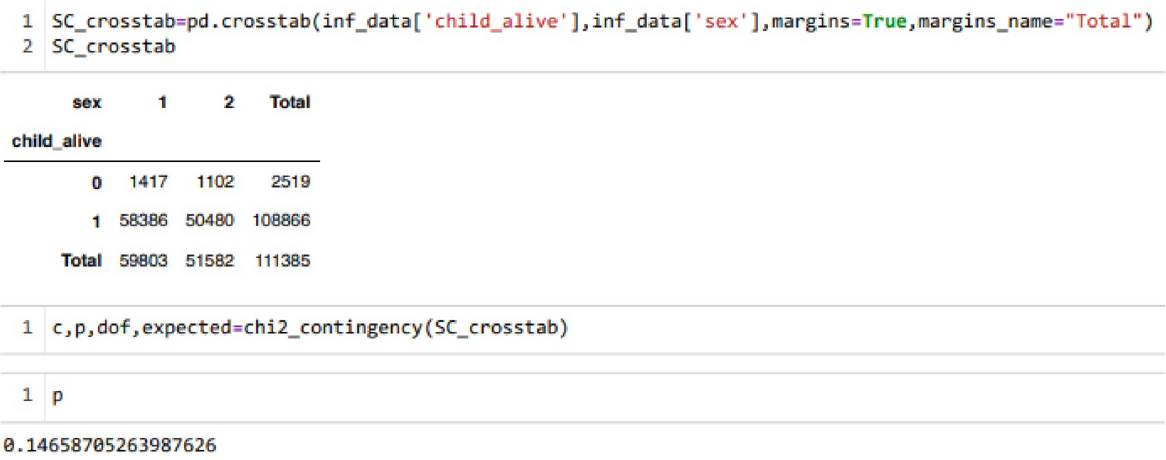
From the above table we can see that p value= 0.00 < alpha =0.05 i.e.; We Reject the null hypothesis. So, we can say that there is an association between wealth index groups and the death of the child.

3.7.5. Sex of child V/s child death

H_0: There is no significant association between child death and the sex of the child.

H_1: There is a significant association between child death and the sex of the child.

If the p-value obtained from the chi-square test is greater than the chosen significance level (commonly set at 0.05), you would fail to reject the null hypothesis. This would suggest that there is not enough evidence to conclude that there is a significant association between child death and the sex of the child.



```
SC_crosstab=pd.crosstab(inf_data['child_alive'],inf_data['sex'],margins=True,margins_name="Total")
SC_crosstab
```

sex child_alive	1	2	Total
0	1417	1102	2519
1	58386	50480	108866
Total	59803	51582	111385

```
c,p,dof,expected=chi2_contingency(SC_crosstab)
```

```
p
```

```
0.14658705263987626
```

Interpretation

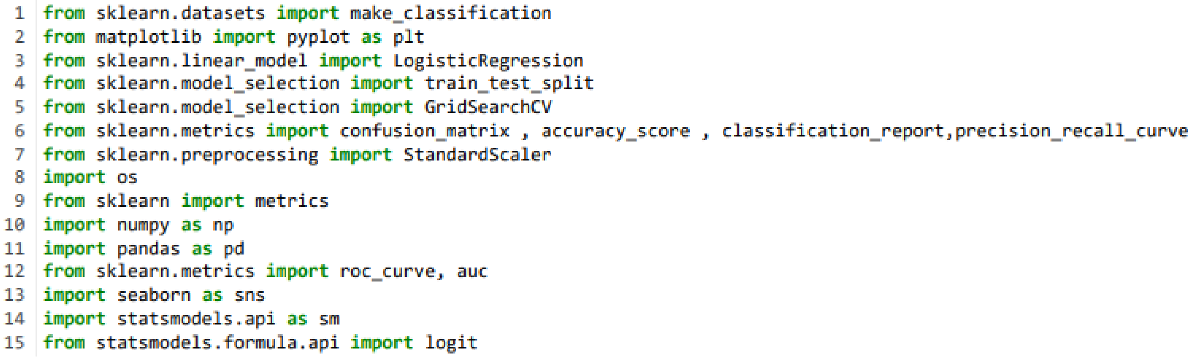
The p-value obtained from the chi-square test is greater than the chosen significance level alpha = 0.05, Thus we fail to reject the null hypothesis. This would suggest that there is not enough evidence to conclude that there is a significant association between child death and the sex of the child.

3.8 Applying Logistic Regression model using python

In the application of logistic regression, the process involves training a model to forecast binary outcomes, exemplified by predicting infant mortality. Employing Python and libraries such as NumPy, Pandas, and scikit-learn, the workflow includes importing essential libraries, loading and preparing the data, constructing the model, generating predictions, and assessing its efficacy using metrics like accuracy and the confusion matrix.

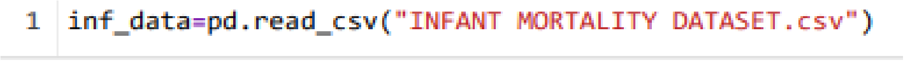
3.8.1. Performing Logistic Regression using python (step by step).

Importing necessary libraries

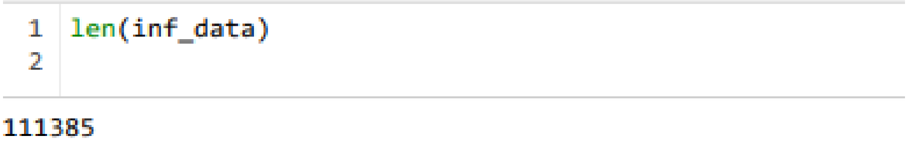


```python
from sklearn.datasets import make_classification
from matplotlib import pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import confusion_matrix , accuracy_score , classification_report,precision_recall_curve
from sklearn.preprocessing import StandardScaler
import os
from sklearn import metrics
import numpy as np
import pandas as pd
from sklearn.metrics import roc_curve, auc
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import logit
```

Reading the data as inf_data

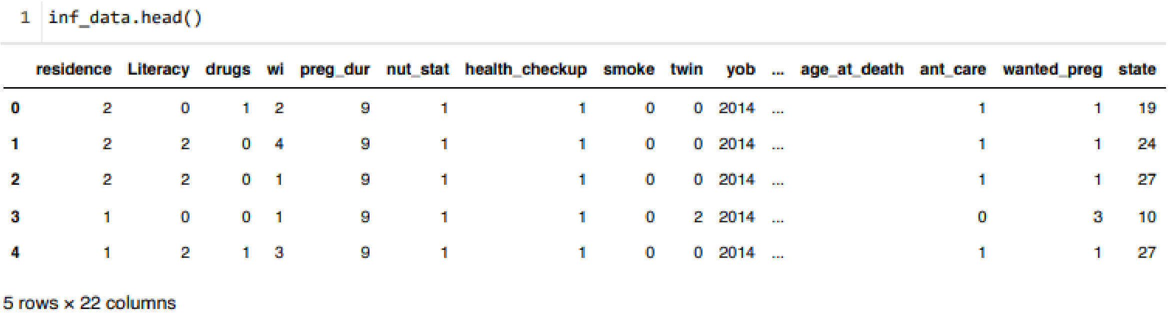


```python
inf_data=pd.read_csv("INFANT MORTALITY DATASET.csv")
```



```python
len(inf_data)

```

```
111385
```

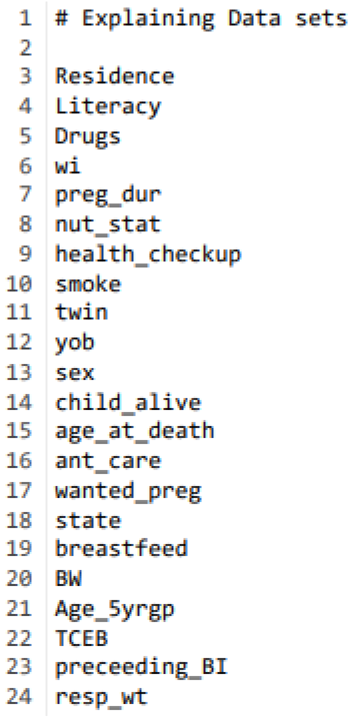


```python
inf_data.head()
```

	residence	Literacy	drugs	wi	preg_dur	nut_stat	health_checkup	smoke	twin	yob	...	age_at_death	ant_care	wanted_preg	state
0	2	0	1	2	9	1	1	0	0	2014	...		1	1	19
1	2	2	0	4	9	1	1	0	0	2014	...		1	1	24
2	2	2	0	1	9	1	1	0	0	2014	...		1	1	27
3	1	0	0	1	9	1	1	0	2	2014	...		0	3	10
4	1	2	1	3	9	1	1	0	0	2014	...		1	1	27

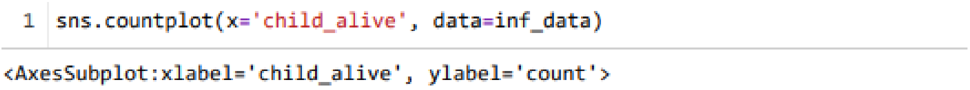
5 rows × 22 columns

Variables included in the dataset:



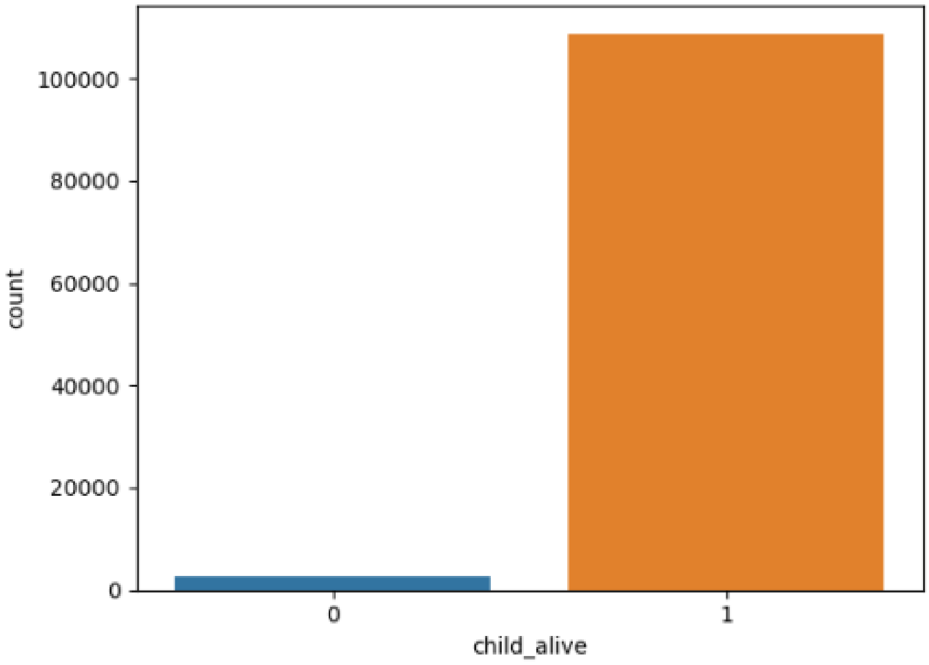
```
# Explaining Data sets

Residence
Literacy
Drugs
wi
preg_dur
nut_stat
health_checkup
smoke
twin
yob
sex
child_alive
age_at_death
ant_care
wanted_preg
state
breastfeed
BW
Age_5yrgp
TCEB
preceeding_BI
resp_wt
```

Plotting the countplot of dependent variable i.e; child is alive

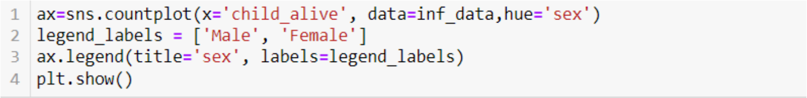


```
sns.countplot(x='child_alive', data=inf_data)
```

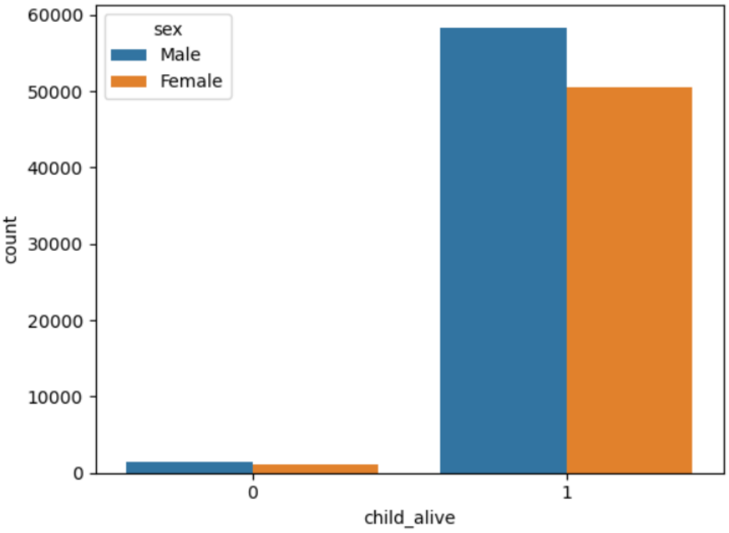
```
<AxesSubplot:xlabel='child_alive', ylabel='count'>
```



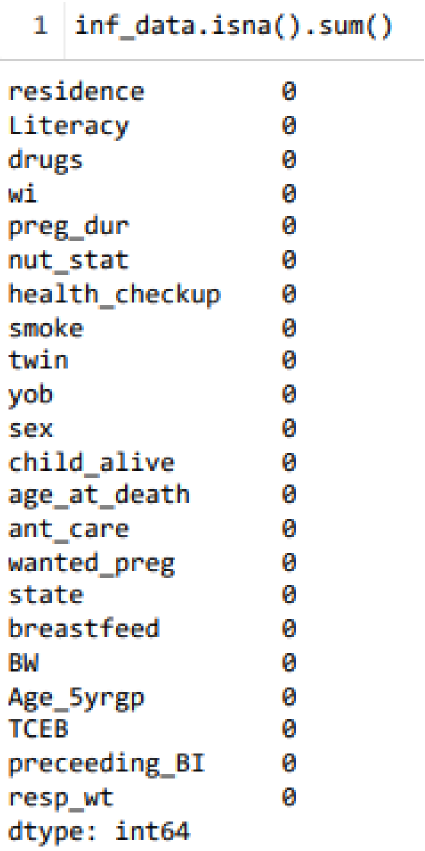
Distribution of child alive with respect to sex



```
ax=sns.countplot(x='child_alive', data=inf_data,hue='sex')
legend_labels = ['Male', 'Female']
ax.legend(title='sex', labels=legend_labels)
plt.show()
```



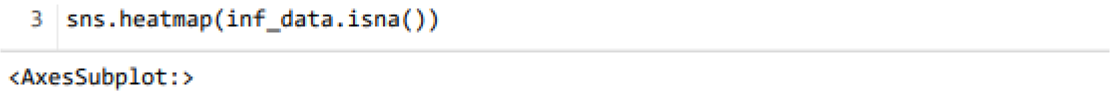
Checking for null values



```
inf_data.isna().sum()
```

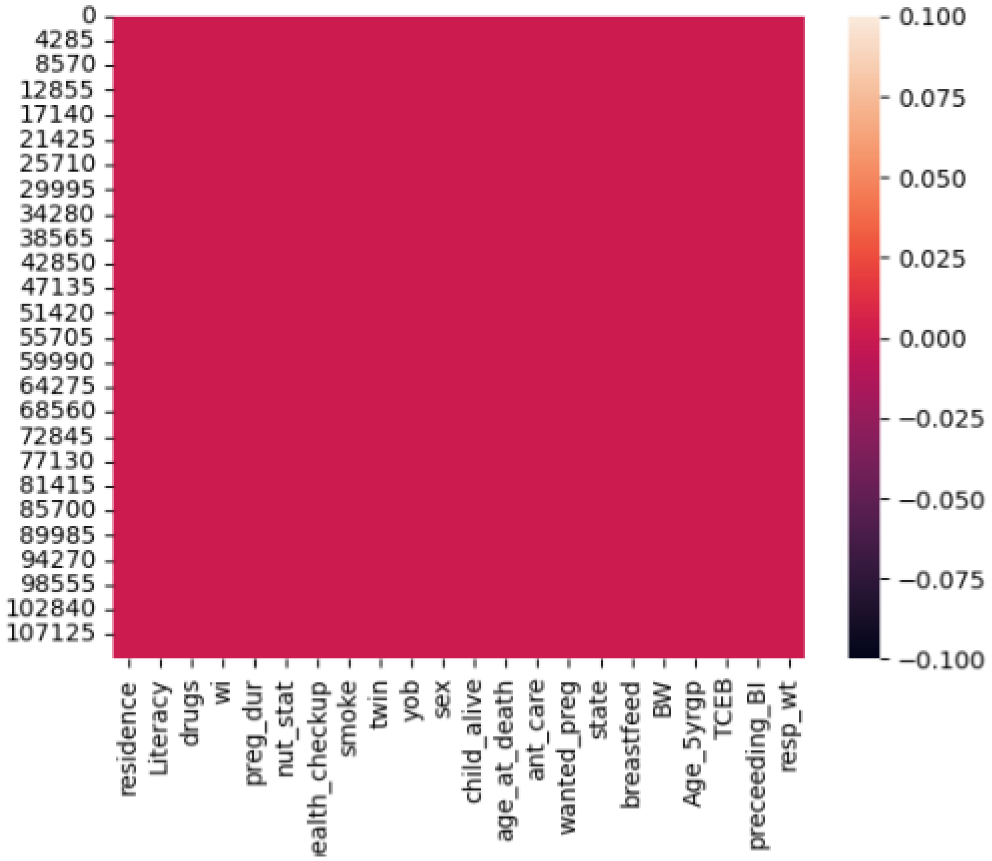
```
residence          0
Literacy           0
drugs              0
wi                 0
preg_dur           0
nut_stat           0
health_checkup     0
smoke              0
twin               0
yob                0
sex                0
child_alive        0
age_at_death       0
ant_care           0
wanted_preg        0
state              0
breastfeed         0
BW                 0
Age_5yrgp          0
TCEB               0
preceeding_BI      0
resp_wt            0
dtype: int64
```

Visualizing null values using heatmap



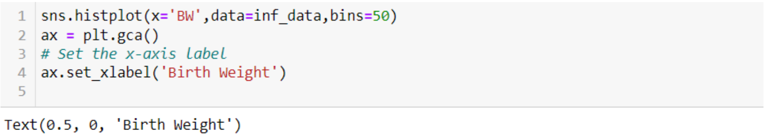
```
sns.heatmap(inf_data.isna())
```

```
<AxesSubplot:>
```



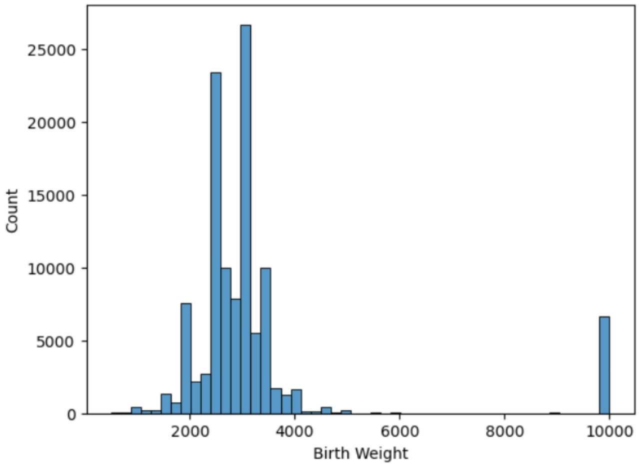
From the graph it is evident that our dataset has no null values.

3.8.2. Distribution of birth weight



```
sns.histplot(x='BW',data=inf_data,bins=50)
ax = plt.gca()
# Set the x-axis label
ax.set_xlabel('Birth Weight')

```

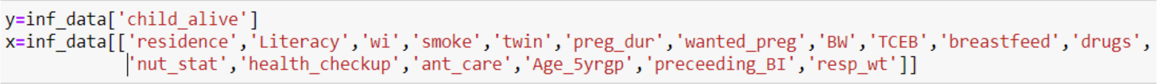
```
Text(0.5, 0, 'Birth Weight')
```



From the above distribution plot, we can see most of the infants are weighed between 2000gm and 4000gm.

There are some outliers as well like 9000gm, some rare cases which are weighed below 750gm.

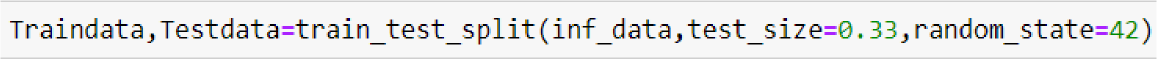
Separating dependent variables and independent variables



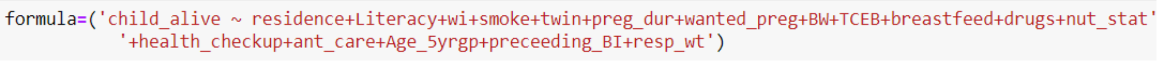
```
y=inf_data['child_alive']
x=inf_data[['residence','Literacy','wi','smoke','twin','preg_dur','wanted_preg','BW','TCEB','breastfeed','drugs',
            'nut_stat','health_checkup','ant_care','Age_5yrgp','preceeding_BI','resp_wt']]
```

3.8.3. Fitting the model using logit ()

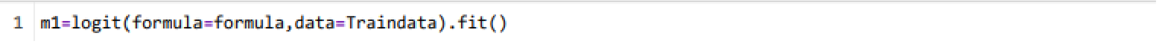
Split the dataset as testing and training datasets. a training set is used to build the model whereas a testing set is used to evaluate its performance.



```
Traindata,Testdata=train_test_split(inf_data,test_size=0.33,random_state=42)
```

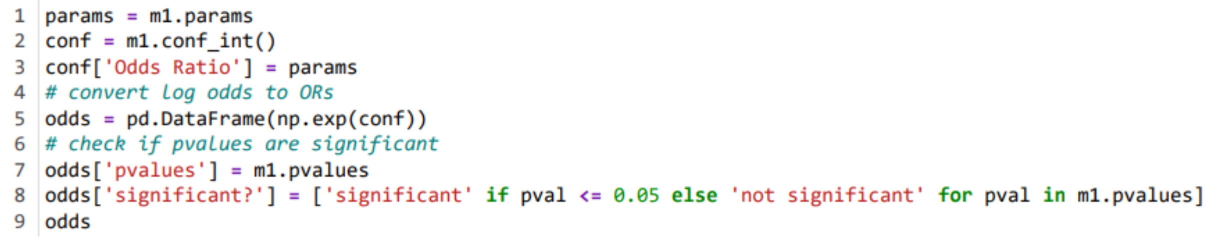


```
formula=('child_alive ~ residence+Literacy+wi+smoke+twin+preg_dur+wanted_preg+BW+TCEB+breastfeed+drugs+nut_stat'
            '+health_checkup+ant_care+Age_5yrgp+preceeding_BI+resp_wt')
```

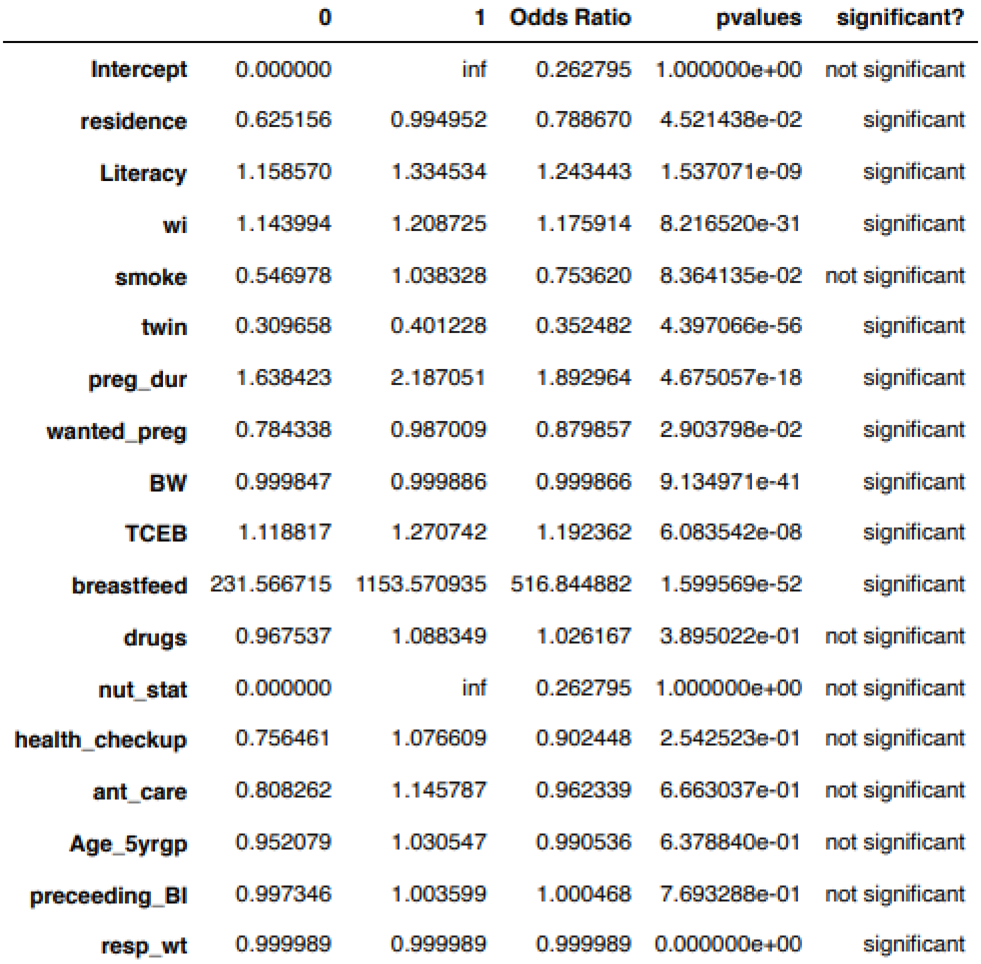


```
m1=logit(formula=formula,data=Traindata).fit()
```

To find the Odds ratio and the significance of the model use the following commands.



```
params = m1.params
conf = m1.conf_int()
conf['Odds Ratio'] = params
# convert log odds to ORs
odds = pd.DataFrame(np.exp(conf))
# check if pvalues are significant
odds['pvalues'] = m1.pvalues
odds['significant?'] = ['significant' if pval <= 0.05 else 'not significant' for pval in m1.pvalues]
odds
```



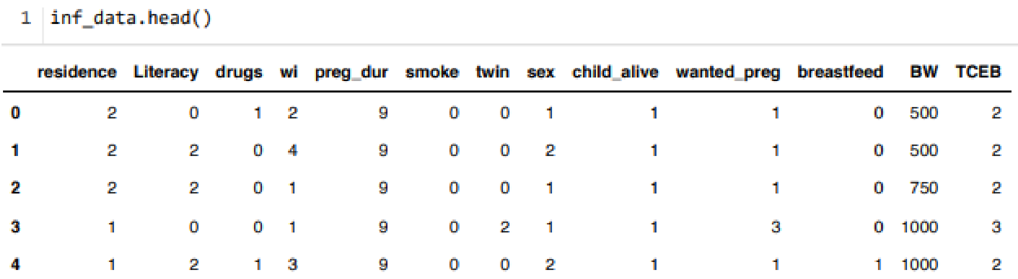
	0	1	Odds Ratio	pvalues	significant?
Intercept	0.000000	inf	0.262795	1.000000e+00	not significant
residence	0.625156	0.994952	0.788670	4.521438e-02	significant
Literacy	1.158570	1.334534	1.243443	1.537071e-09	significant
wi	1.143994	1.208725	1.175914	8.216520e-31	significant
smoke	0.546978	1.038328	0.753620	8.364135e-02	not significant
twin	0.309658	0.401228	0.352482	4.397066e-56	significant
preg_dur	1.638423	2.187051	1.892964	4.675057e-18	significant
wanted_preg	0.784338	0.987009	0.879857	2.903798e-02	significant
BW	0.999847	0.999886	0.999866	9.134971e-41	significant
TCEB	1.118817	1.270742	1.192362	6.083542e-08	significant
breastfeed	231.566715	1153.570935	516.844882	1.599569e-52	significant
drugs	0.967537	1.088349	1.026167	3.895022e-01	not significant
nut_stat	0.000000	inf	0.262795	1.000000e+00	not significant
health_checkup	0.756461	1.076609	0.902448	2.542523e-01	not significant
ant_care	0.808262	1.145787	0.962339	6.663037e-01	not significant
Age_5yrgp	0.952079	1.030547	0.990536	6.378840e-01	not significant
preceeding_BI	0.997346	1.003599	1.000468	7.693288e-01	not significant
resp_wt	0.999989	0.999989	0.999989	0.000000e+00	significant

From the above table we can say that certain variables are not significant to the model since the p-value > alpha level of significance (alpha=0.05)

Now fit a model by removing the variables which are not significant

3.8.4. Dropping not significant variables

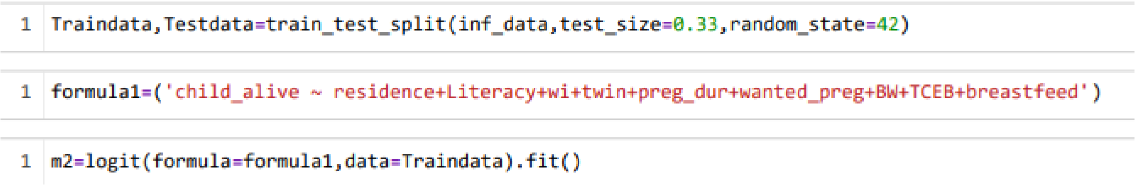
```
inf_data.drop(['health_checkup','ant_care','preceeding_BI','resp_wt','yob','state','age_at_death','nut_stat','Age_5yrgp'],
axis=1,inplace=True)
```



```
inf_data.head()
```

	residence	Literacy	drugs	wi	preg_dur	smoke	twin	sex	child_alive	wanted_preg	breastfeed	BW	TCEB
0	2	0	1	2	9	0	0	1	1	1	0	500	2
1	2	2	0	4	9	0	0	2	1	1	0	500	2
2	2	2	0	1	9	0	0	1	1	1	0	750	2
3	1	0	0	1	9	0	2	1	1	3	0	1000	3
4	1	2	1	3	9	0	0	2	1	1	1	1000	2

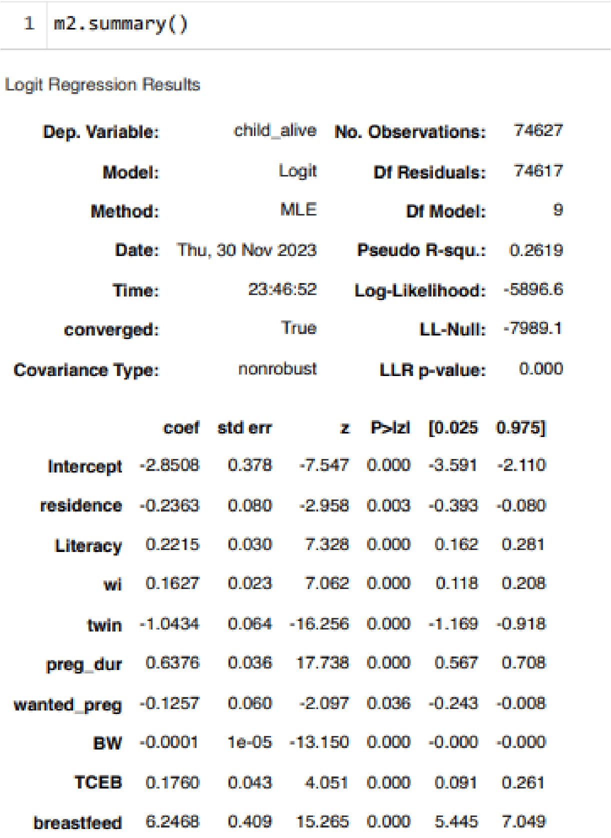
3.8.5. Again, testing the model



```
Traindata,Testdata=train_test_split(inf_data,test_size=0.33,random_state=42)
```

```
formula1=('child_alive ~ residence+Literacy+wi+twin+preg_dur+wanted_preg+BW+TCEB+breastfeed')
```

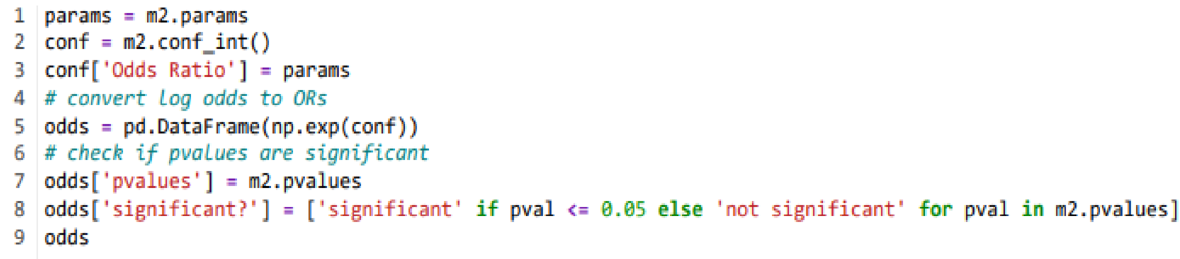
```
m2=logit(formula=formula1,data=Traindata).fit()
```



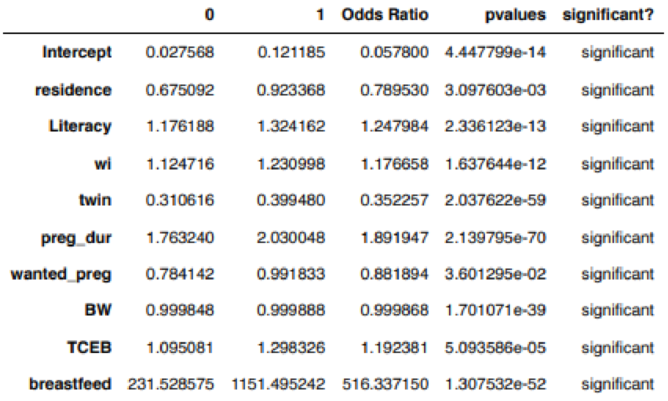
```
m2.summary()
```

```
Logit Regression Results

   Dep. Variable:        child_alive  No. Observations:    74627
           Model:              Logit      Df Residuals:    74617
          Method:                MLE          Df Model:        9
            Date:   Thu, 30 Nov 2023     Pseudo R-squ.:   0.2619
            Time:           23:46:52    Log-Likelihood:  -5896.6
       converged:               True           LL-Null:  -7989.1
 Covariance Type:          nonrobust       LLR p-value:    0.000

                 coef  std err        z  P>|z|  [0.025  0.975]
   Intercept  -2.8508    0.378   -7.547  0.000  -3.591  -2.110
   residence  -0.2363    0.080   -2.958  0.003  -0.393  -0.080
    Literacy   0.2215    0.030    7.328  0.000   0.162   0.281
          wi   0.1627    0.023    7.062  0.000   0.118   0.208
        twin  -1.0434    0.064  -16.256  0.000  -1.169  -0.918
    preg_dur   0.6376    0.036   17.738  0.000   0.567   0.708
 wanted_preg  -0.1257    0.060   -2.097  0.036  -0.243  -0.008
          BW  -0.0001    1e-05  -13.150  0.000  -0.000  -0.000
        TCEB   0.1760    0.043    4.051  0.000   0.091   0.261
  breastfeed   6.2468    0.409   15.265  0.000   5.445   7.049
```

In this case we got a Pseudo R-square value of 0.2619 thus our logistic regression model explains 26.19% variability in the response variable. This means that this model has some explanatory powers but still there exists some unexplained variability.



```
params = m2.params
conf = m2.conf_int()
conf['Odds Ratio'] = params
# convert log odds to ORs
odds = pd.DataFrame(np.exp(conf))
# check if pvalues are significant
odds['pvalues'] = m2.pvalues
odds['significant?'] = ['significant' if pval <= 0.05 else 'not significant' for pval in m2.pvalues]
odds
```



	0	1	Odds Ratio	pvalues	significant?
Intercept	0.027568	0.121185	0.057800	4.447799e-14	significant
residence	0.675092	0.923368	0.789530	3.097603e-03	significant
Literacy	1.176188	1.324162	1.247984	2.336123e-13	significant
wi	1.124716	1.230998	1.176658	1.637644e-12	significant
twin	0.310616	0.399480	0.352257	2.037622e-59	significant
preg_dur	1.763240	2.030048	1.891947	2.139795e-70	significant
wanted_preg	0.784142	0.991833	0.881894	3.601295e-02	significant
BW	0.999848	0.999888	0.999868	1.701071e-39	significant
TCEB	1.095081	1.298326	1.192381	5.093586e-05	significant
breastfeed	231.528575	1151.495242	516.337150	1.307532e-52	significant

From the above table we can see that all the variables are now significant and from the table odds ratio we can interpret the results as follows:

3.8.6. Odds ratio

Odd ratio interpretation

1. ***Residence***

An odds ratio of 0.7883 suggests that the odds of the survival of the child occurring in rural areas are 78.83% of the odds in urban areas. The odds of a child being alive are 22% lower when considering the residence variable compared to the residence variable. If the odds ratio had been 1, it would imply no difference in the odds between rural and urban areas. Since the odds ratio is less than 1 (0.7883), it indicates a negative association

2. ***Pregnancy duration*** (Reference point: 0= No education)

Odd Ratio:1.891 > 1 (positive association). For each additional unit of pregnancy duration (upto 10 months), the odds of a child being alive increase by approximately 89%

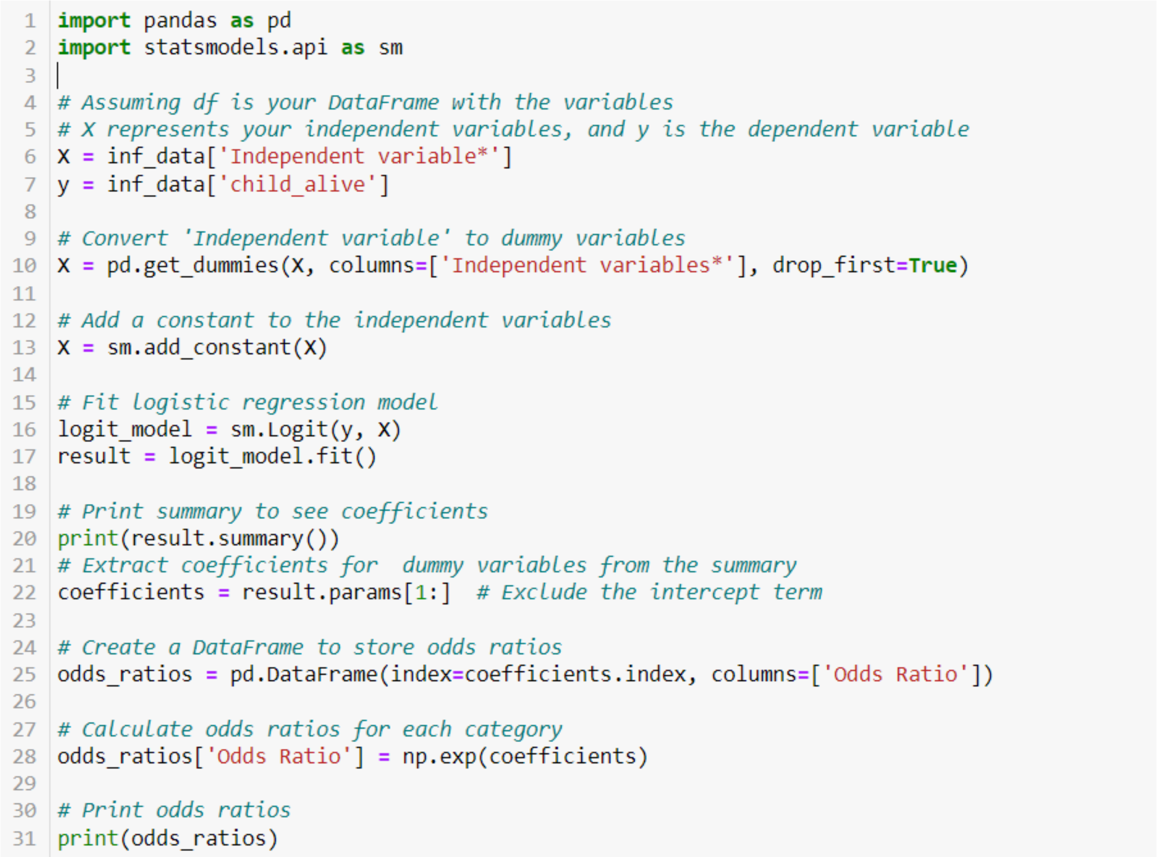
3. ***Birth weight of child***

Odd Ratio:0.999 < 1 (negative association). For a one-unit increase in birth weight, the odds of a child being alive change very little." This suggests that birth weight may not be a strong predictor of whether a child is alive or not.

4. ***Breastfeed***

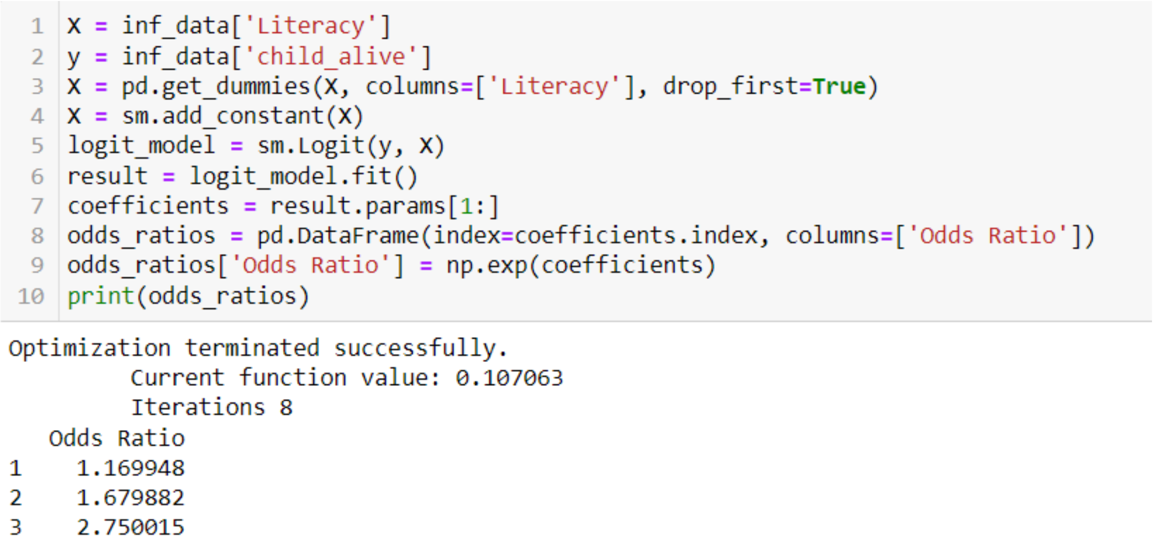
Odd Ratio: 516.337 > 1 (positive association). Higher positive value of odd ratio indicates that the odds of a child being alive are substantially higher when considering the breastfeeding variable compared to when the child was not breastfed

For variables with more than two categories we need to extract the odd ratio for each category with respect to a reference point, usually the first category. For that we can define a model to find out the following



```
import pandas as pd
import statsmodels.api as sm

# Assuming df is your DataFrame with the variables
# X represents your independent variables, and y is the dependent variable
X = inf_data['Independent variable*']
y = inf_data['child_alive']

# Convert 'Independent variable' to dummy variables
X = pd.get_dummies(X, columns=['Independent variables*'], drop_first=True)

# Add a constant to the independent variables
X = sm.add_constant(X)

# Fit logistic regression model
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Print summary to see coefficients
print(result.summary())
# Extract coefficients for  dummy variables from the summary
coefficients = result.params[1:]  # Exclude the intercept term

# Create a DataFrame to store odds ratios
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])

# Calculate odds ratios for each category
odds_ratios['Odds Ratio'] = np.exp(coefficients)

# Print odds ratios
print(odds_ratios)
```

By inserting the independent variables in the command ‘x’ we get the odd ratio within each category



```
X = inf_data['Literacy']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['Literacy'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```

```
Optimization terminated successfully.
         Current function value: 0.107063
         Iterations 8
   Odds Ratio
1    1.169948
2    1.679882
3    2.750015
```

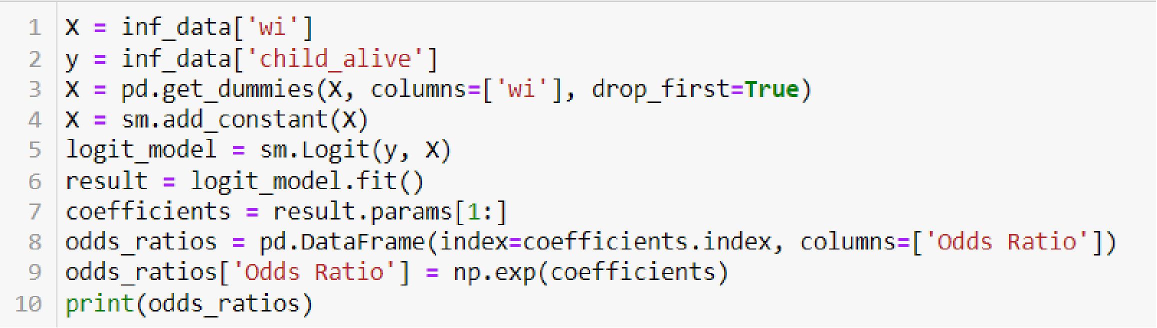
5. ***Literacy*** (Reference point: 0 = No education)

i. Odds Ratio for Category 1 (Primary Education) compared to No Education: The odds ratio is 1.169948. This suggests that, compared to individuals with no education, the odds of the event (survival of a child) for individuals with primary education are 1.169948 times higher.

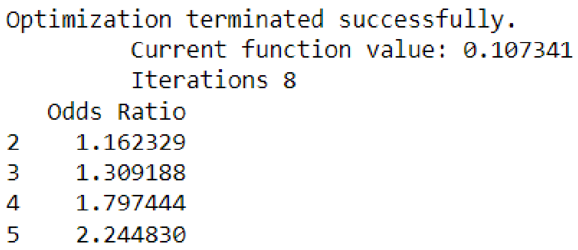
ii. Odds Ratio for Category 2 (Secondary Education) compared to No Education: The odds ratio is 1.679882. This suggests that, compared to individuals with no education, the odds of the event for individuals with secondary education are 1.679882 times higher.

iii. Odds Ratio for Category 3 (Higher Education) compared to No Education: The odds ratio is 2.750015. This suggests that, compared to individuals with no education, the odds of the event for individuals with higher education are 2.750015 times higher.

6. ***Wealth Index*** (Reference point: 1= poor)



```
X = inf_data['wi']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['wi'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```



```
Optimization terminated successfully.
         Current function value: 0.107341
         Iterations 8
   Odds Ratio
2    1.162329
3    1.309188
4    1.797444
5    2.244830
```

i. Wealth Index Category 2 (Poorer):

Odds Ratio: 1.162329

The odds of the event (child survival) for individuals in the "Poorer" category are 1.162 times the odds for individuals in the "Poor" category. In other words, there is a slight increase in the odds of child survival for those in the "Poorer" category compared to the "Poor" category.

ii. Wealth Index Category 3 (Middle):

Odds Ratio: 1.309188

The odds of the event for individuals in the "Middle" category are 1.309 times the odds for individuals in the "Poor" category. This suggests a greater increase in the odds of child survival for those in the "Middle" category compared to the "Poor" category.

iii. Wealth Index Category 4 (Richer):

Odds Ratio: 1.797444

The odds of the event for individuals in the "Richer" category are 1.797 times the odds for individuals in the "Poor" category. This indicates a substantial increase in the odds of child survival for those in the "Richer" category compared to the "Poor" category.

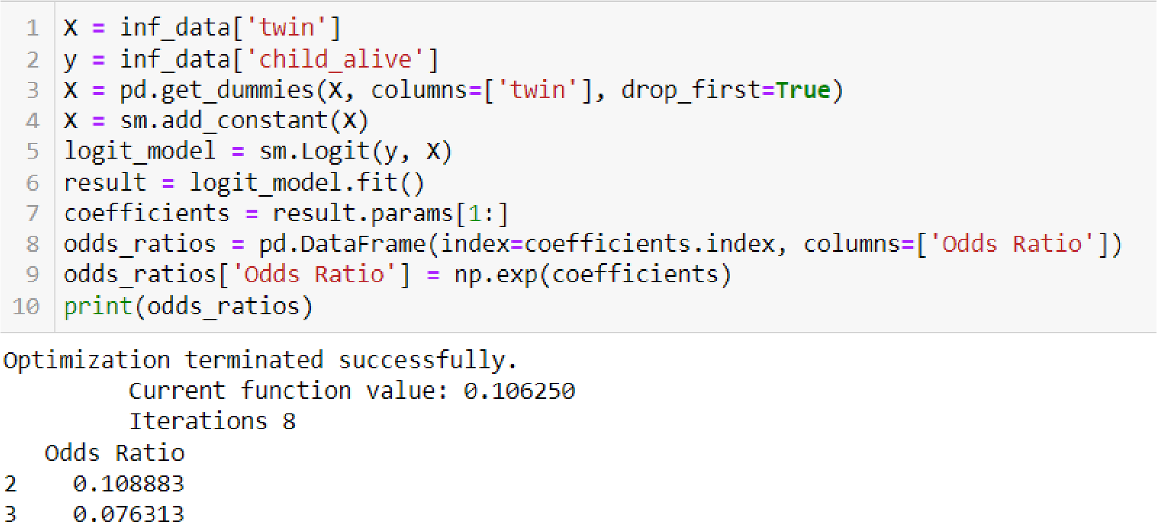
iv. Wealth Index Category 5 (Richest):

Odds Ratio: 2.244830

The odds of the event for individuals in the "Richest" category are 2.244 times the odds for individuals in the "Poor" category. This suggests a significant increase in the odds of child survival for those in the "Richest" category compared to the "Poor" category.

The wealth index category increases from "Poor" to "Poorer," "Middle," "Richer," and "Richest," there is a trend of increasing odds of child survival. This indicates a positive association between wealth index and the likelihood of child survival, with higher wealth categories being associated with higher odds of survival.

7. ***Child is Twin*** (Reference point: 0= No education)



```
X = inf_data['twin']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['twin'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```

```
Optimization terminated successfully.
         Current function value: 0.106250
         Iterations 8
   Odds Ratio
2    0.108883
3    0.076313
```

i. For twins (category 2):

Odds Ratio = 0.108883

This odds ratio of 0.108883 suggests that the odds of survival for twins are approximately 10.89% of the odds of survival for single children.

Since the odds ratio is less than 1, it implies a decrease in the odds of survival for twins compared to single children.

In practical terms, this may indicate that, all else being equal, single children have higher odds of survival compared to twins.

ii. For 3rd multiples (category 3):

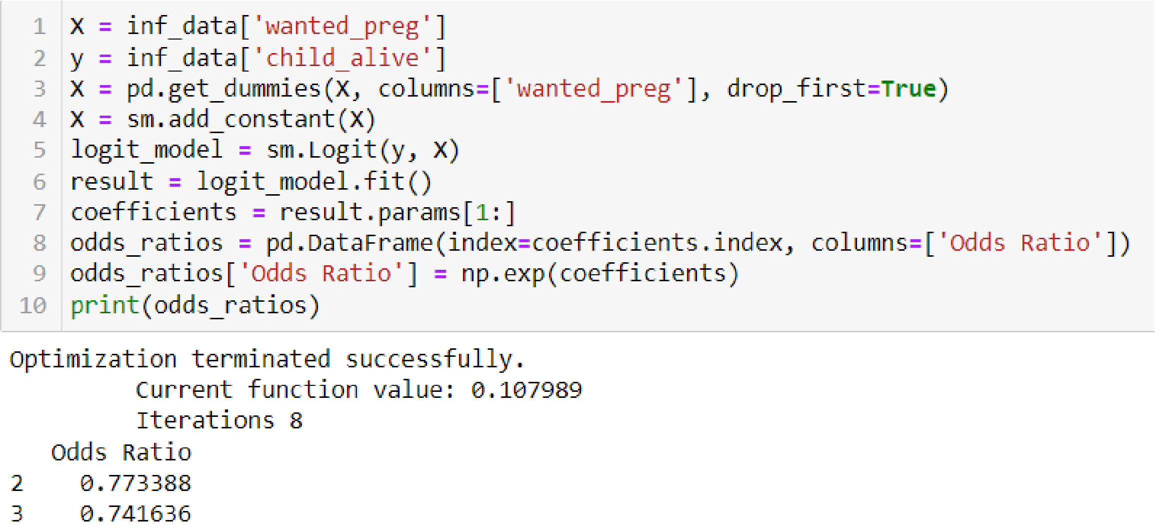
Odds Ratio = 0.076313

This odds ratio of 0.076313 suggests that the odds of survival for 3rd multiples are approximately 7.63% of the odds of survival for single children.

Similar to the previous case, since the odds ratio is less than 1, it implies a decrease in the odds of survival for 3rd multiples compared to single children.

In practical terms, this may indicate that, all else being equal, single children have higher odds of survival compared to 3rd multiples.

8. ***Wanted pregnancies When Became Pregnant*** (Reference point: 0= No education)



```
X = inf_data['wanted_preg']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['wanted_preg'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```

```
Optimization terminated successfully.
         Current function value: 0.107989
         Iterations 8
   Odds Ratio
2    0.773388
3    0.741636
```

i. Category 2 (Mother wanted pregnancy later)

The odds ratio of 0.773388 indicates that, compared to mothers who wanted the pregnancy, the odds of the child surviving are approximately 77.34% of the odds for mothers who wanted the pregnancy.

A value less than 1 suggests a decrease in the odds of child survival for mothers who wanted the pregnancy later compared to those who wanted it.

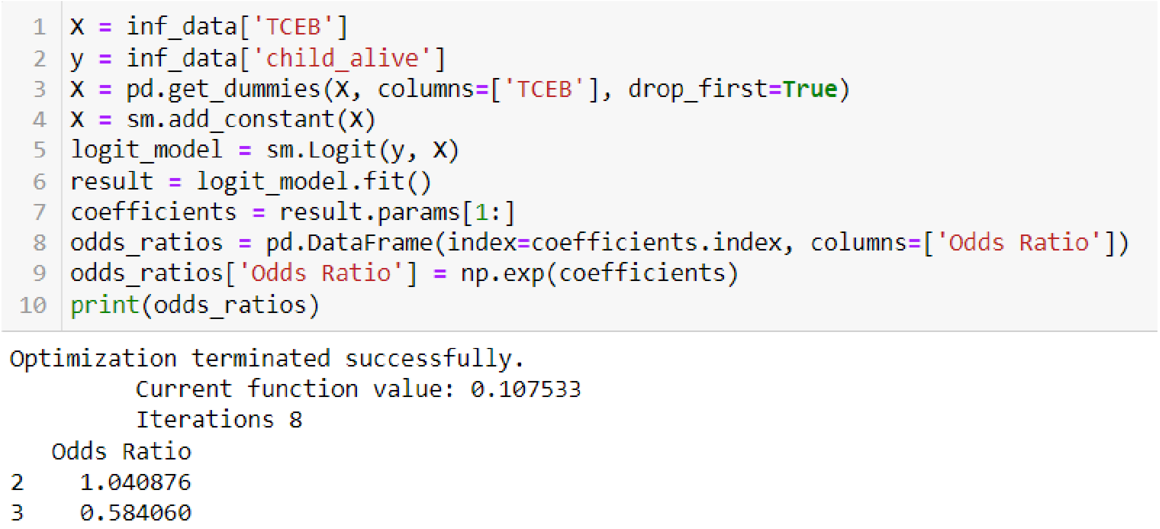
It's important to check the confidence interval associated with this odds ratio to determine if the difference is statistically significant. If the confidence interval includes 1, it may not be statistically significant.

ii. Category 3 (Mother wanted pregnancy no more): (Reference point: 0=No education)

The odds ratio of 0.741636 indicates that, compared to mothers who wanted the pregnancy, the odds of the child surviving are approximately 74.16% of the odds for mothers who wanted the pregnancy.

Similarly, a value less than 1 suggests a decrease in the odds of child survival for mothers who do not want the pregnancy anymore compared to those who wanted

9. ***Total children ever born*** (Reference point: 0= No education)



```
X = inf_data['TCEB']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['TCEB'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```

```
Optimization terminated successfully.
         Current function value: 0.107533
         Iterations 8
   Odds Ratio
2    1.040876
3    0.584060
```

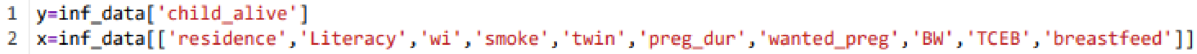
i. For the category "2 = 2 to 3 children":

   The odds ratio of 1.040876 suggests that, compared to individuals with only one child, the odds of child survival for individuals with two to three children are 1.040876 times higher.But since it is very close to 1 we can assume there is no significant difference among the odds Since the odds ratio is less than 1 (0.7883), it indicates a (negative association) It's a slight increase, but whether this increase is statistically significant would depend on the confidence interval associated with this odds ratio.

ii. For the category "3 = 4 or more children":

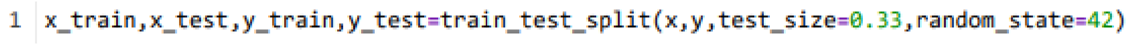
   The odds ratio of 0.584060 suggests that, compared to individuals with only one child, the odds of child survival for individuals with four or more children are 0.584060 times lower. This indicates a decrease in the odds of child survival for individuals with four or more children compared to those with only one child

3.8.7. Model Prediction Using Logistic Regression

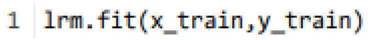


```
y=inf_data['child_alive']
x=inf_data[['residence','Literacy','wi','smoke','twin','preg_dur','wanted_preg','BW','TCEB','breastfeed']]
```

Split the data as train data and test data

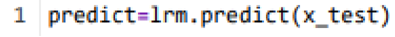


```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33,random_state=42)
```



```
lrm.fit(x_train,y_train)
```

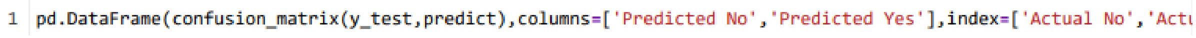
For prediction



```
predict=lrm.predict(x_test)
```

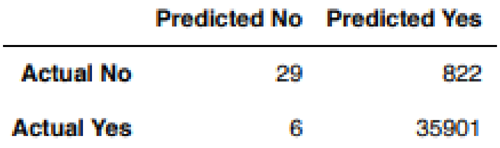
3.8.8. Testing the model

To evaluate how our model is performing.



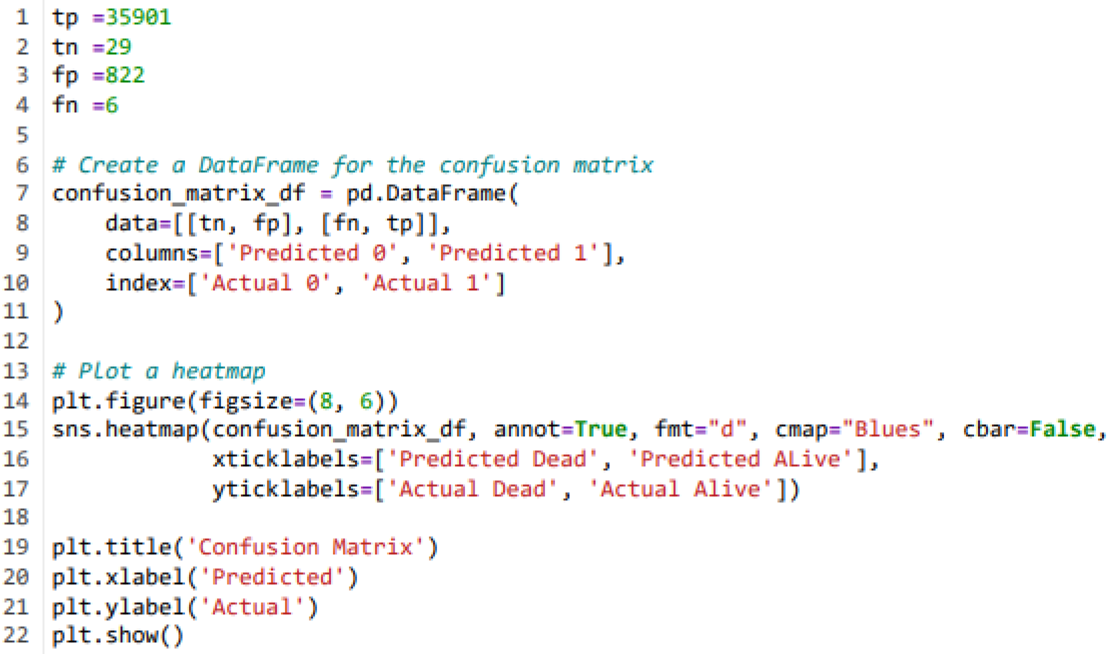
```
pd.DataFrame(confusion_matrix(y_test,predict),columns=['Predicted No','Predicted Yes'],index=['Actual No','Actu
```

Print the confusion matrix

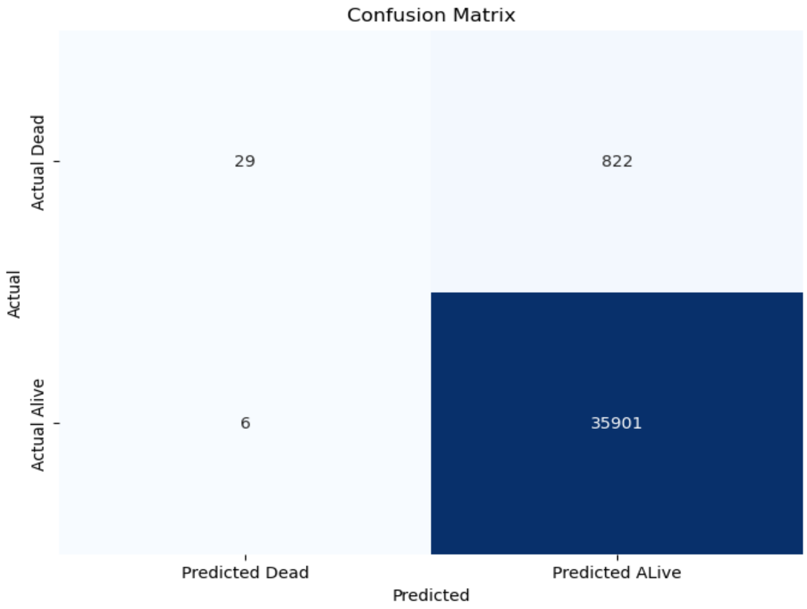


	Predicted No	Predicted Yes
Actual No	29	822
Actual Yes	6	35901

3.8.9. Visualizing the confusion matrix



```
tp =35901
tn =29
fp =822
fn =6

# Create a DataFrame for the confusion matrix
confusion_matrix_df = pd.DataFrame(
    data=[[tn, fp], [fn, tp]],
    columns=['Predicted 0', 'Predicted 1'],
    index=['Actual 0', 'Actual 1']
)

# PLot a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix_df, annot=True, fmt="d", cmap="Blues", cbar=False,
            xticklabels=['Predicted Dead', 'Predicted ALive'],
            yticklabels=['Actual Dead', 'Actual Alive'])

plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```



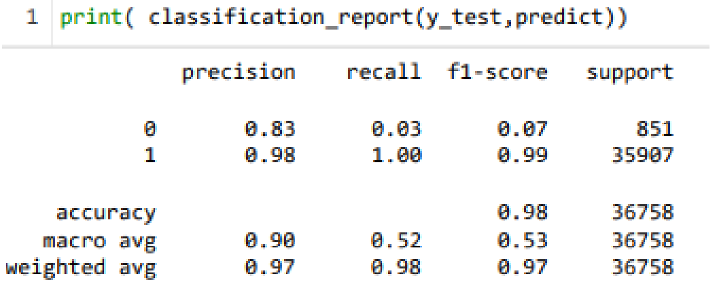
True Positive (TP): There are 35,901 instances where the model correctly predicted the positive class. This represents the number of actual positive cases that the model correctly identified.

True Negative (TN): There are 29 instances where the model correctly predicted the negative class. This represents the number of actual negative cases that the model correctly identified.

False Positive (FP): There are 822 instances where the model incorrectly predicted the positive class. This represents the number of actual negative cases that the model incorrectly identified as positive. Also known as Type I errors.

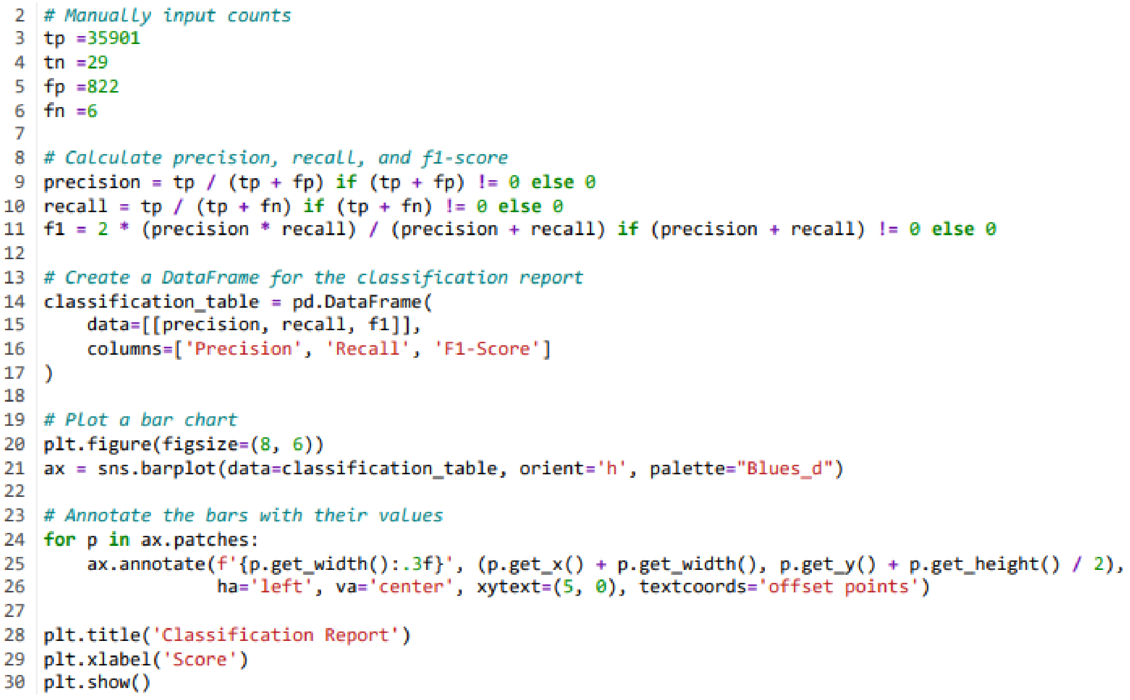
False Negative (FN): There are 6 instances where the model incorrectly predicted the negative class. This represents the number of actual positive cases that the model incorrectly identified as negative. Also known as Type II errors.

3.8.10. Print classification report in order to check the accuracy

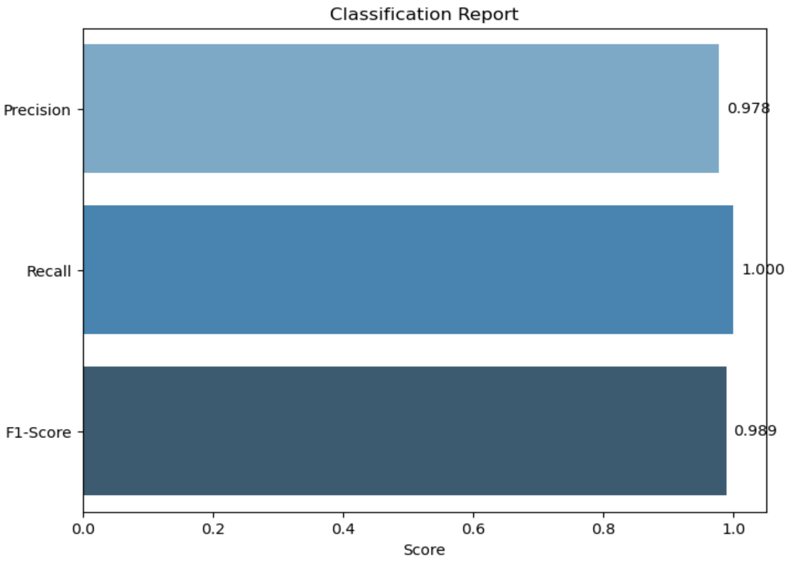


```
print( classification_report(y_test,predict))
```

```
              precision    recall  f1-score   support

           0       0.83      0.03      0.07       851
           1       0.98      1.00      0.99     35907

    accuracy                           0.98     36758
   macro avg       0.90      0.52      0.53     36758
weighted avg       0.97      0.98      0.97     36758
```



```
# Manually input counts
tp =35901
tn =29
fp =822
fn =6

# Calculate precision, recall, and f1-score
precision = tp / (tp + fp) if (tp + fp) != 0 else 0
recall = tp / (tp + fn) if (tp + fn) != 0 else 0
f1 = 2 * (precision * recall) / (precision + recall) if (precision + recall) != 0 else 0

# Create a DataFrame for the classification report
classification_table = pd.DataFrame(
    data=[[precision, recall, f1]],
    columns=['Precision', 'Recall', 'F1-Score']
)

# Plot a bar chart
plt.figure(figsize=(8, 6))
ax = sns.barplot(data=classification_table, orient='h', palette="Blues_d")

# Annotate the bars with their values
for p in ax.patches:
    ax.annotate(f'{p.get_width():.3f}', (p.get_x() + p.get_width(), p.get_y() + p.get_height() / 2),
                ha='left', va='center', xytext=(5, 0), textcoords='offset points')

plt.title('Classification Report')
plt.xlabel('Score')
plt.show()
```

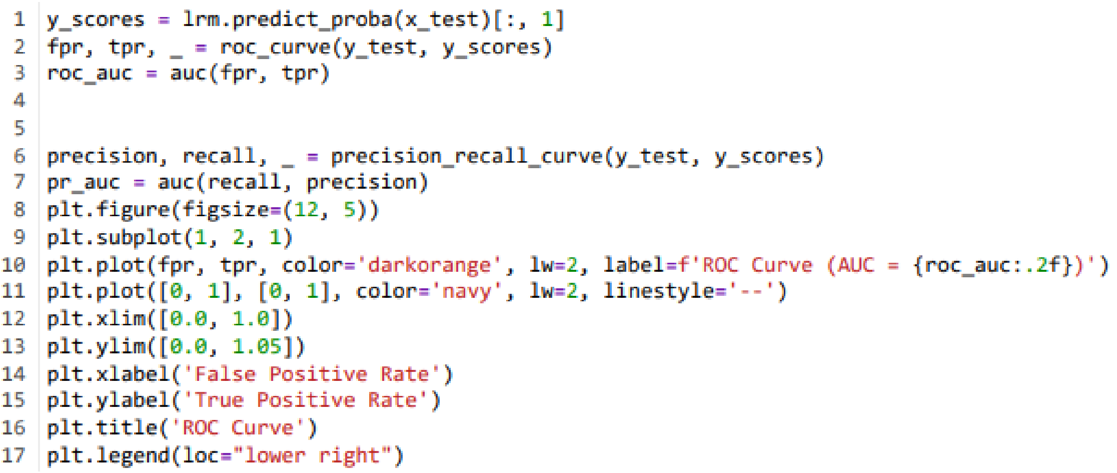


Recall = 1: This means that the model is correctly identifying all instances of the positive class. A recall of 1 indicates that there are no false negatives, and the model is capturing all positive cases.

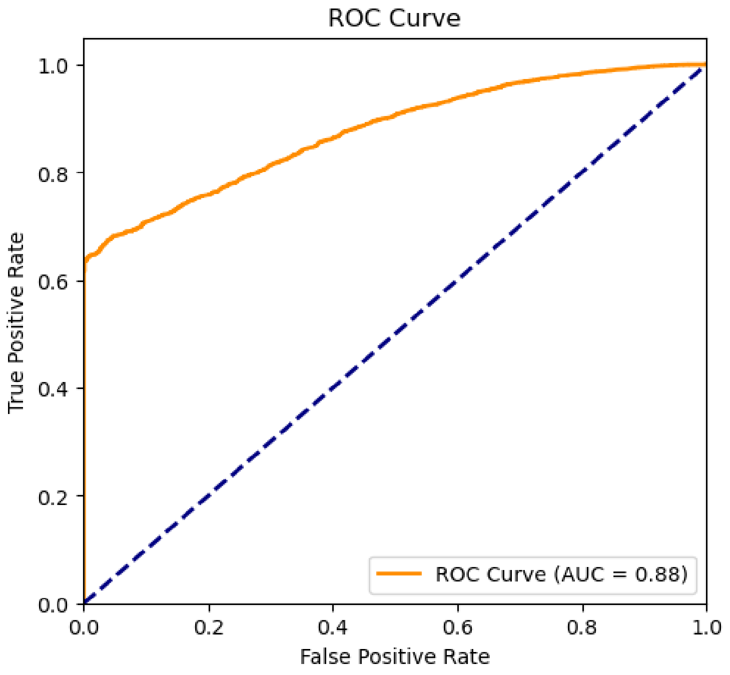
F1 Score = 0.989: The F1 score is the harmonic mean of precision and recall. A value close to 1 indicates a good balance between precision and recall. In your case, with an F1 score of 0.989, it suggests that your model is performing very well in terms of both precision and recall.

Accuracy = 0.97: The accuracy is the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances. An accuracy of 0.97 (97%) is high and indicates that your model is making accurate predictions on the overall dataset.

3.8.11. Plotting ROC curve



```
y_scores = lrm.predict_proba(x_test)[:, 1]
fpr, tpr, _ = roc_curve(y_test, y_scores)
roc_auc = auc(fpr, tpr)


precision, recall, _ = precision_recall_curve(y_test, y_scores)
pr_auc = auc(recall, precision)
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC Curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
```



An AUC (Area Under the Curve) value of 0.88 in the context of an ROC (Receiver Operating Characteristic) curve suggests that your logistic regression model has relatively good discriminatory power.

Chapter-4

CONCLUSIONS

4.1 Conclusion and Discussion

1. There is a clear decrease trend in total cases of Infant Death from 2013 to 2019 and the rate appears to increasing from 2019.

2. The State of Uttar Pradesh has the highest Infant Mortality Rate during the years 2011-2021 followed by the states of Chhattisgarh and Madhya Pradesh. Kerala have the lowest followed by Mizoram, and Lakshwadeep

3. Over the period from 2014 to 2021, the sex ratio in India has witnessed a concerning trend of decline. This decline raises crucial societal and gender-related concerns that require close attention.

4. Infant mortality in rural areas is higher than that of urban areas. The higher rate in rural areas indicates challenges in accessing quality medical services, potentially stemming from limited infrastructure or inadequate healthcare resources.

5. There is a clear increasing trend in receiving health benefits for everyone from 2014 to 2020 and a start of decreasing trend from 2021. The trend of decreasing in that particular year can be because of the limitations in accessing healthcare at the time of COVID-19 pandemic

6. There occurred a drastic decrease in the percentage of wanted pregnancies in the year 2020-21. Amid the COVID-19 pandemic, women's hesitancy towards pregnancy likely stems from heightened concerns about maternal and fetal health.

7. It is evident that the percentage of women attaining higher education has been increasing year by year.

8. Single births exhibit the lowest mortality, suggesting a favorable environment for infant health. Twin births follow, indicating a slightly higher but still relatively lower risk. Conversely, multiple births, such as triplets, show the highest mortality rate.

9. From Mann Whitney U test we conclude that there is significant difference in average birth weight between infants those who survived and those who didn't survive.

10. An infant belonging to Urban area has 50% more chance of survival than an infant belonging to Rural area.

11. An infant belonging to a mother who has no smoking habit has 36% more survival chance than an infant belonging to a mother who has a smoking habit.

12. If a mother becomes pregnant by her own will, then the infant has 19% more survival chance than an infant born from a mother who had an unplanned pregnancy.

Also, the infant has 29% more chance of survival than the infant whose mother never wanted to be pregnant any more.

After performing Chi square test for associations, we arrived in few conclusions

I. There is a significant association between literacy and total children ever born

II. There is a significant association between child alive and place of residence

III. There is a significant association between wealth index groups and death of the child.

IV. There is a significant association between child is alive and child is twin (multiple birth)

V. There is no significant association between sex of child and survival of Infant

In Logistic Regression Modelling

1. It is found that the fitted model is able to explains 26% of the variation in dependent variable, even though there are some unexplained Variation, we can go on with the model Because there are several other factors that may cause the death of the infants like anxiety level of mother during pregnancy other internal or external injuries etc. which we didn’t take under consideration.

2. Predicted the model outcome with accuracy 97% .

3. AUC (Area Under the Curve) value of 0.88 in the context of an ROC (Receiver Operating Characteristic) curve implies, Fitted logistic regression model has relatively good discriminatory power.

4.2 Recommendations

Reducing infant mortality rates in India requires a multifaceted approach that addresses various factors contributing to infant deaths. Here are some recommendations:

i. ***Improved Maternal Healthcare***: Enhance access to quality prenatal and antenatal care. Encourage early and regular prenatal check-ups to ensure optimal maternal and fetal health. Promote maternal nutrition education, particularly in rural areas where infant mortality rates are elevated, to foster healthy pregnancies.

ii. ***Skilled Birth Attendance***: Increase the availability of skilled healthcare professionals during childbirth. Encourage institutional deliveries with skilled birth attendants.

iii. ***Neonatal Care***: Strengthen neonatal care facilities, particularly in rural areas. provide training for healthcare workers in neonatal resuscitation and care.

iv. ***Immunization Programs***: Ensure widespread coverage of immunization programs to protect infants from preventable diseases. implement awareness campaigns to educate parents on the importance of vaccinations.

v. ***Promote Breastfeeding***: Encourage and support breastfeeding to enhance infants' immunity. provide lactation support and education to mothers.

vi. ***Nutrition Programs***: Implement nutrition programs to address malnutrition in infants. promote the importance of exclusive breastfeeding for the first six months.

vii. ***Water and Sanitation***: Improve access to clean water and sanitation facilities. educate communities on hygiene practices to prevent waterborne diseases.

viii. ***Education and Awareness***: Conduct public awareness campaigns on infant health, maternal care, and family planning. provide education on the importance of spacing pregnancies for maternal and child health, particularly in regions with lower literacy rates.

ix. ***Government Policies***: Implement and enforce policies that support maternal and child health. allocate sufficient resources to healthcare, especially in regions experiencing higher rates such as Uttar Pradesh, Chhattisgarh, and Madhya Pradesh.

REFERENCE

References

1. SC Guptha & VK Kapoor Fundamentals of Mathematical Statistics

   S Chand and Sons (1989)

2. Logistic Regression A self learning Text, Third Edition New York: Spring

   KleinBaum, DG (1994)

3. Regression analysis by example, 3rd edition

   Ali S. Hadi and Samprit Chatterjee

4. Introduction to Linear Regression Analysis

   (1989) Douglas Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining

5. Python for Data Analysis

   Wes McKinney (2012)

Source of data

1. Demographic Health Survey Programme: Birth Recode Data

   https://dhsprogram.com/data/dataset/India_Standard-DHS_2020.cfm?flag=0

APPENDICES

APPENDIX-A

Data

year of birth	Child is alive		Total
	No	Yes	
2014	24	1833	1857
2015	155	10216	10371
2016	292	16576	16868
2017	391	20240	20631
2018	569	24623	25192
2019	674	23701	24375
2020	377	10702	11079
2021	37	975	1012
Total	2519	108866	111385

State	Child is alive		Total
	No	Yes	
Jammu & Kashmir	19	1608	1627
Himachal Pradesh	20	1737	1757
Punjab	49	2182	2231
Chandigarh	2	66	68
Uttarakhand	53	2535	2588
Haryana	47	2996	3043
Nct Of Delhi	15	832	847
Rajasthan	135	6233	6368
Uttar Pradesh	533	14981	15514
Bihar	160	5554	5714
Sikkim	4	297	301
Arunachal Pradesh	16	845	861
Nagaland	13	475	488
Manipur	9	738	747
Mizoram	9	1202	1211
Tripura	29	1318	1347
Meghalaya	47	2352	2399
Assam	129	5498	5627
West Bengal	54	3678	3732
Jharkhand	125	5041	5166
Odisha	163	6605	6768
Chhattisgarh	167	5352	5519
Madhya Pradesh	266	9165	9431

Gujarat	119	5412	5531
Dadra & Nagar Haveli and Daman & Diu	7	335	342
Maharashtra	59	3970	4029
Andhra Pradesh	43	1708	1751
Karnataka	83	4878	4961
Goa	3	284	287
Lakshadweep	0	182	182
Kerala	4	1443	1447
Tamil Nadu	50	4277	4327
Puducherry	2	330	332
Andaman & Nicobar Islands	2	187	189
Telangana	78	4353	4431
Ladakh	5	217	222
Total	2519	108866	111385

state	Highest educational level				Total
	No education	Primary	Secondary	Higher	
Jammu & Kashmir	358	134	958	177	1627
Himachal Pradesh	55	139	1111	452	1757
Punjab	250	262	1377	342	2231
Chandigarh	9	5	43	11	68
Uttarakhand	222	219	1418	729	2588
Haryana	396	370	1661	616	3043
Nct Of Delhi	116	102	453	176	847
Rajasthan	1855	1109	2673	731	6368
Uttar Pradesh	4507	2123	6490	2394	15514
Bihar	2643	775	1975	321	5714
Sikkim	16	65	187	33	301
Arunachal Pradesh	190	97	508	66	861
Nagaland	60	81	313	34	488
Manipur	53	80	517	97	747
Mizoram	78	186	846	101	1211
Tripura	110	224	949	64	1347
Meghalaya	388	644	1256	111	2399
Assam	853	824	3650	300	5627
West Bengal	395	631	2443	263	3732
Jharkhand	1567	636	2581	382	5166
Odisha	1465	928	3851	524	6768
Chhattisgarh	1324	747	3024	424	5519
Madhya Pradesh	1844	1409	5294	884	9431

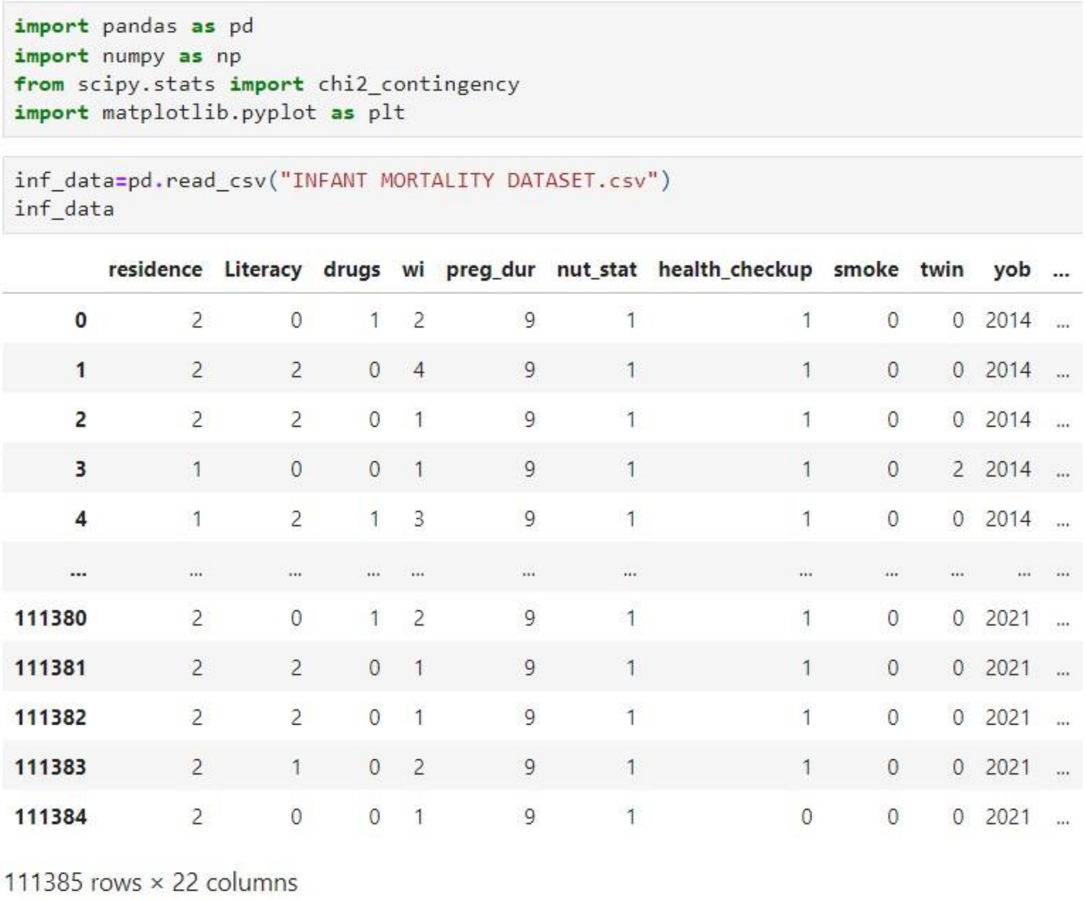
Gujarat	1114	789	3239	389	5531
Dadra & Nagar Haveli And Daman & Diu	37	33	230	42	342
Maharashtra	255	326	2875	573	4029
Andhra Pradesh	246	207	1024	274	1751
Karnataka	572	405	3273	711	4961
Goa	8	14	178	87	287
Lakshadweep	2	4	126	50	182
Kerala	2	10	841	594	1447
Tamil Nadu	76	213	2419	1619	4327
Puducherry	4	11	181	136	332
Andaman & Nicobar Islands	8	13	148	20	189
Telangana	753	309	2492	877	4431
Ladakh	41	16	130	35	222
Total	21872	14140	60734	14639	111385

state	Wealth index within state					Total
	Poorest	Poorer	Middle	Richer	Richest	
Jammu & Kashmir	396	368	353	281	229	1627
Himachal Pradesh	430	350	361	337	279	1757
Punjab	523	537	480	404	287	2231
Chandigarh	15	24	17	9	3	68
Uttarakhand	649	741	527	410	261	2588
Haryana	668	668	679	616	412	3043
Nct Of Delhi	224	216	174	148	85	847
Rajasthan	1717	1400	1272	1157	822	6368
Uttar Pradesh	3621	3605	3435	2869	1984	15514
Bihar	1326	1385	1286	1098	619	5714
Sikkim	95	91	59	39	17	301
Arunachal Pradesh	180	190	220	146	125	861
Nagaland	166	118	94	61	49	488
Manipur	221	148	156	125	97	747
Mizoram	384	333	232	172	90	1211
Tripura	334	310	278	248	177	1347
Meghalaya	710	616	504	392	177	2399
Assam	1214	1290	1244	1158	721	5627
West Bengal	982	885	871	652	342	3732
Jharkhand	1396	1238	1121	946	465	5166
Odisha	1514	1473	1351	1343	1087	6768
Chhattisgarh	1692	1174	1060	922	671	5519
Madhya Pradesh	2100	2098	2033	1880	1320	9431

Gujarat	1642	1507	1120	861	401	5531
Dadra & Nagar Haveli And Daman & Diu	51	50	70	83	88	342
Maharashtra	1168	1062	915	671	213	4029
Andhra Pradesh	396	380	384	350	241	1751
Karnataka	1156	1200	1119	929	557	4961
Goa	61	55	66	53	52	287
Lakshadweep	24	35	37	49	37	182
Kerala	322	336	304	277	208	1447
Tamil Nadu	824	1048	1049	924	482	4327
Puducherry	67	82	81	70	32	332
Andaman & Nicobar Islands	53	50	48	21	17	189
Telangana	1044	1045	1045	860	437	4431
Ladakh	43	41	37	53	48	222
Total	27408	26149	24082	20614	13132	111385

APPENDIX-B

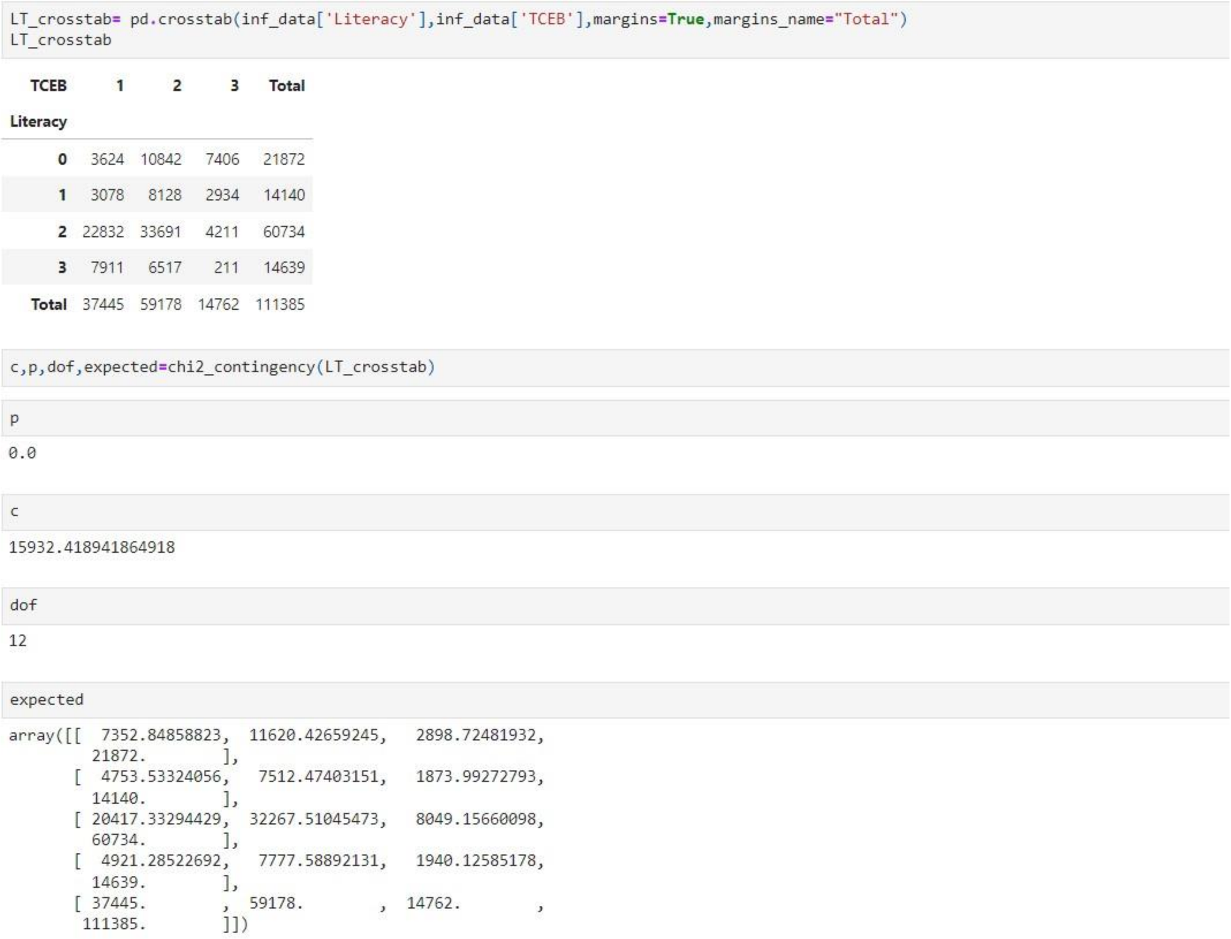
Programming commands for analysis



```
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
```

```
inf_data=pd.read_csv("INFANT MORTALITY DATASET.csv")
inf_data
```

```
        residence  Literacy  drugs  wi  preg_dur  nut_stat  health_checkup  smoke  twin   yob  ...
0               2         0      1   2         9         1               1      0     0  2014  ...
1               2         2      0   4         9         1               1      0     0  2014  ...
2               2         2      0   1         9         1               1      0     0  2014  ...
3               1         0      0   1         9         1               1      0     2  2014  ...
4               1         2      1   3         9         1               1      0     0  2014  ...
...           ...       ...    ...  ...      ...       ...             ...    ...   ...   ...  ...
111380          2         0      1   2         9         1               1      0     0  2021  ...
111381          2         2      0   1         9         1               1      0     0  2021  ...
111382          2         2      0   1         9         1               1      0     0  2021  ...
111383          2         1      0   2         9         1               1      0     0  2021  ...
111384          2         0      0   1         9         1               0      0     0  2021  ...

111385 rows × 22 columns
```



```python
LT_crosstab= pd.crosstab(inf_data['Literacy'],inf_data['TCEB'],margins=True,margins_name="Total")
LT_crosstab
```

TCEB	1	2	3	Total
Literacy				
0	3624	10842	7406	21872
1	3078	8128	2934	14140
2	22832	33691	4211	60734
3	7911	6517	211	14639
Total	37445	59178	14762	111385

```python
c,p,dof,expected=chi2_contingency(LT_crosstab)
```

```python
p
```

```
0.0
```

```python
c
```

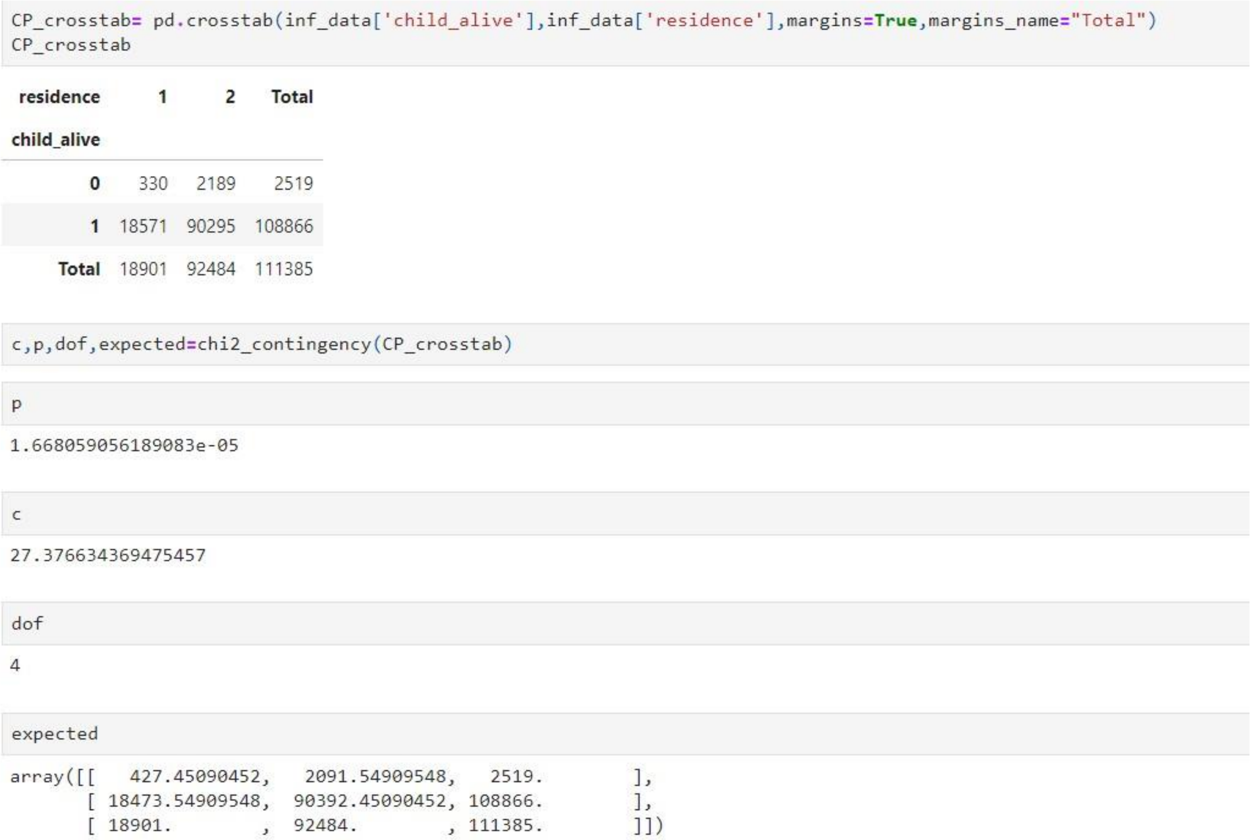
```
15932.418941864918
```

```python
dof
```

```
12
```

```python
expected
```

```
array([[  7352.84858823,  11620.42659245,   2898.72481932,
         21872.        ],
       [  4753.53324056,   7512.47403151,   1873.99272793,
         14140.        ],
       [ 20417.33294429,  32267.51045473,   8049.15660098,
         60734.        ],
       [  4921.28522692,   7777.58892131,   1940.12585178,
         14639.        ],
       [ 37445.        ,  59178.        ,  14762.        ,
        111385.        ]])
```



```python
CP_crosstab= pd.crosstab(inf_data['child_alive'],inf_data['residence'],margins=True,margins_name="Total")
CP_crosstab
```

residence	1	2	Total
child_alive			
0	330	2189	2519
1	18571	90295	108866
Total	18901	92484	111385

```python
c,p,dof,expected=chi2_contingency(CP_crosstab)
```

```python
p
```

```
1.668059056189083e-05
```

```python
c
```

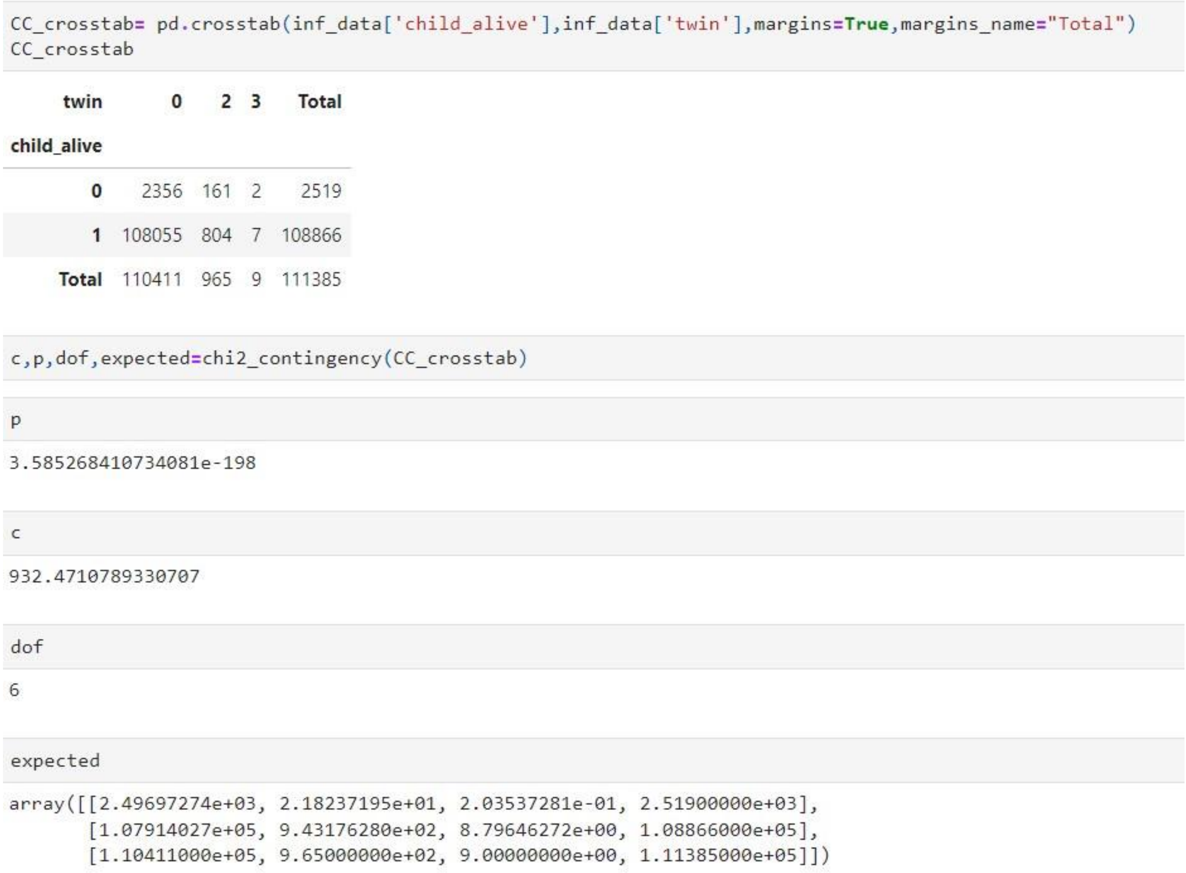
```
27.376634369475457
```

```python
dof
```

```
4
```

```python
expected
```

```
array([[   427.45090452,   2091.54909548,   2519.        ],
       [ 18473.54909548,  90392.45090452, 108866.        ],
       [ 18901.        ,  92484.        , 111385.        ]])
```



```
CC_crosstab= pd.crosstab(inf_data['child_alive'],inf_data['twin'],margins=True,margins_name="Total")
CC_crosstab
```

twin	0	2	3	Total
child_alive				
0	2356	161	2	2519
1	108055	804	7	108866
Total	110411	965	9	111385

```
c,p,dof,expected=chi2_contingency(CC_crosstab)
```

```
p
```

```
3.585268410734081e-198
```

```
c
```

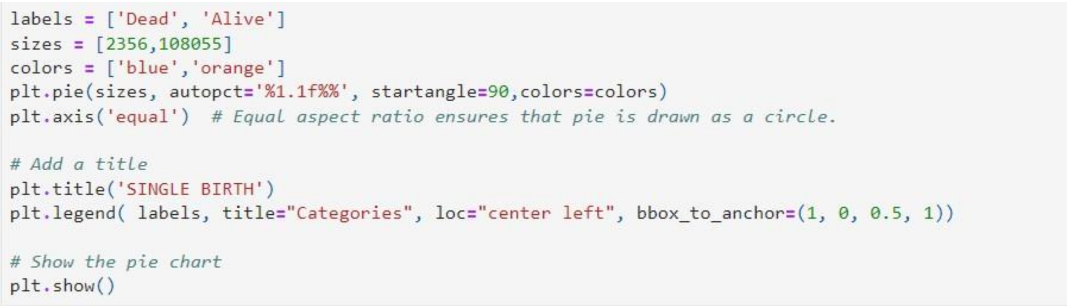
```
932.4710789330707
```

```
dof
```

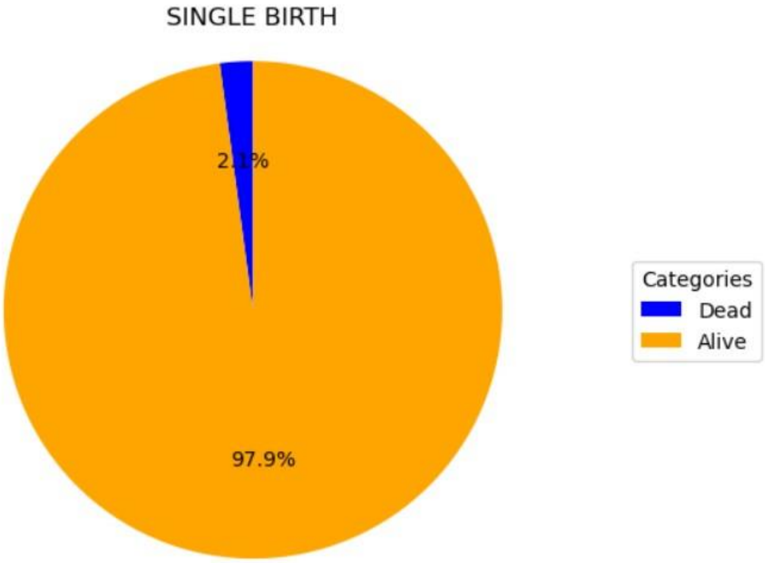
```
6
```

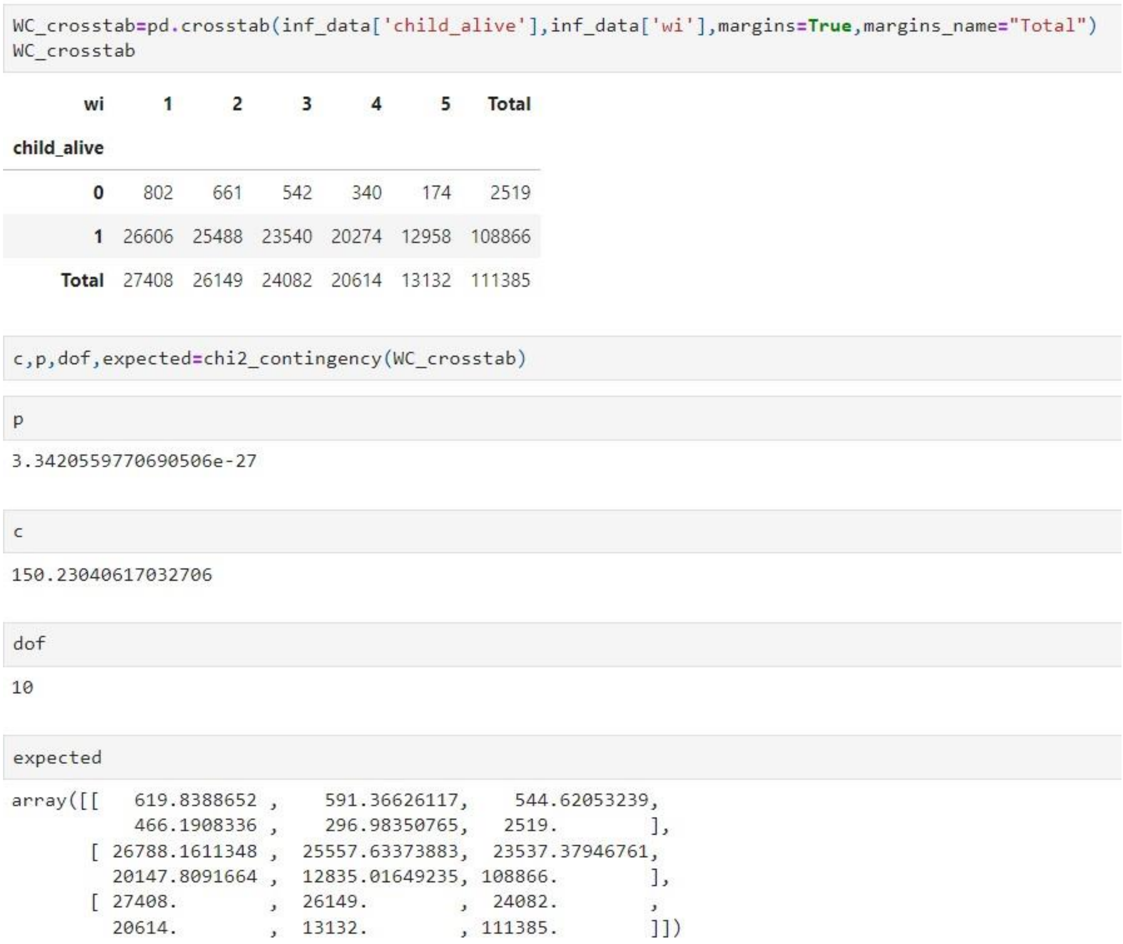
```
expected
```

```
array([[2.49697274e+03, 2.18237195e+01, 2.03537281e-01, 2.51900000e+03],
       [1.07914027e+05, 9.43176280e+02, 8.79646272e+00, 1.08866000e+05],
       [1.10411000e+05, 9.65000000e+02, 9.00000000e+00, 1.11385000e+05]])
```



```
labels = ['Dead', 'Alive']
sizes = [2356,108055]
colors = ['blue','orange']
plt.pie(sizes, autopct='%1.1f%%', startangle=90,colors=colors)
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.

# Add a title
plt.title('SINGLE BIRTH')
plt.legend( labels, title="Categories", loc="center left", bbox_to_anchor=(1, 0, 0.5, 1))

# Show the pie chart
plt.show()
```





```
WC_crosstab=pd.crosstab(inf_data['child_alive'],inf_data['wi'],margins=True,margins_name="Total")
WC_crosstab
```

wi	1	2	3	4	5	Total
child_alive						
0	802	661	542	340	174	2519
1	26606	25488	23540	20274	12958	108866
Total	27408	26149	24082	20614	13132	111385

```
c,p,dof,expected=chi2_contingency(WC_crosstab)
```

```
p
```

```
3.3420559770690506e-27
```

```
c
```

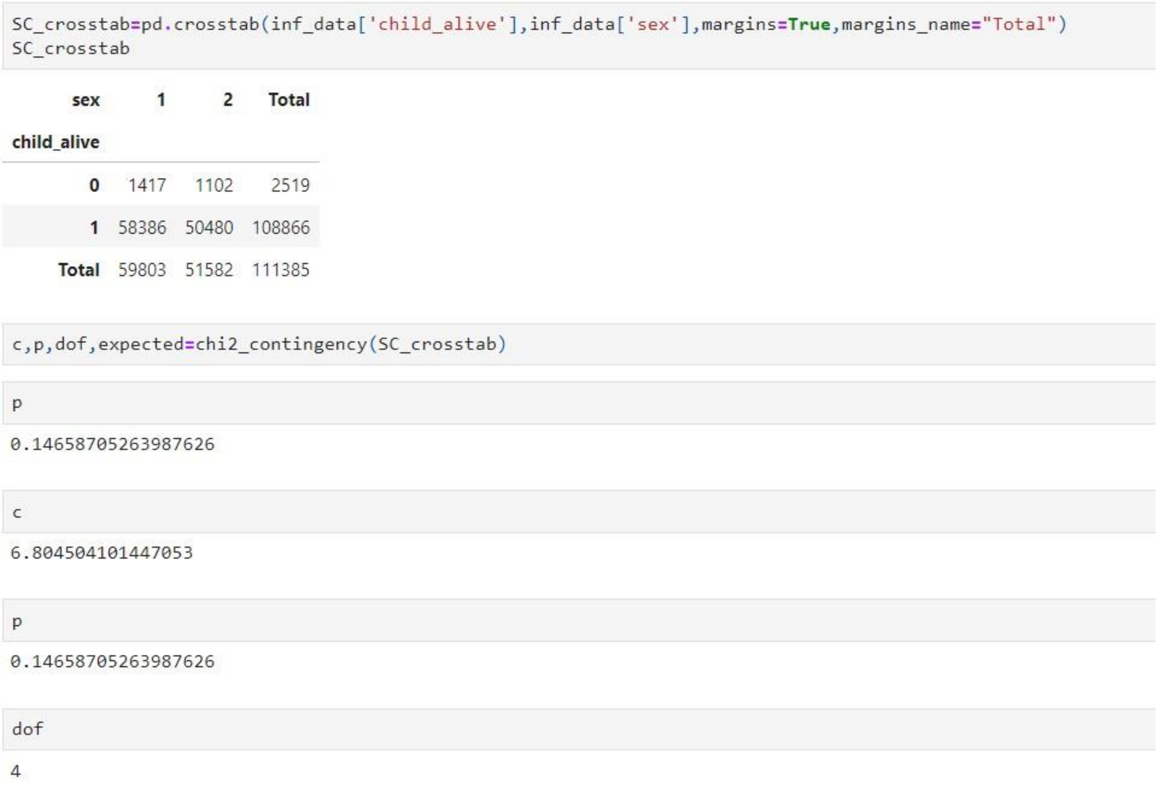
```
150.23040617032706
```

```
dof
```

```
10
```

```
expected
```

```
array([[   619.8388652 ,    591.36626117,    544.62053239,
           466.1908336 ,    296.98350765,   2519.        ],
       [ 26788.1611348 ,  25557.63373883,  23537.37946761,
         20147.8091664 ,  12835.01649235, 108866.        ],
       [ 27408.        ,  26149.        ,  24082.        ,
         20614.        ,  13132.        , 111385.        ]])
```



```
SC_crosstab=pd.crosstab(inf_data['child_alive'],inf_data['sex'],margins=True,margins_name="Total")
SC_crosstab
```

sex	1	2	Total
child_alive			
0	1417	1102	2519
1	58386	50480	108866
Total	59803	51582	111385

```
c,p,dof,expected=chi2_contingency(SC_crosstab)
```

```
p
```

```
0.14658705263987626
```

```
c
```

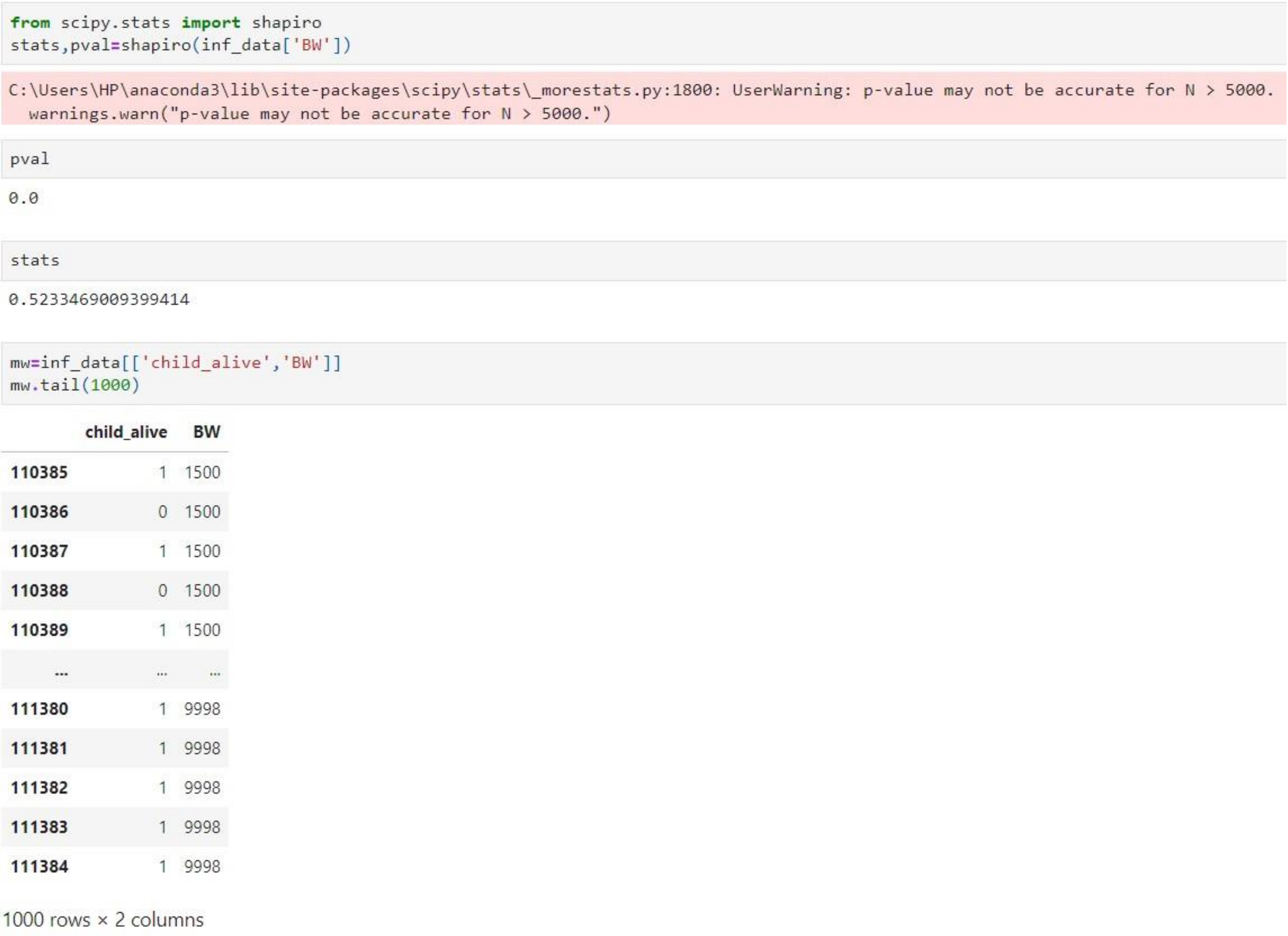
```
6.804504101447053
```

```
p
```

```
0.14658705263987626
```

```
dof
```

```
4
```



```
from scipy.stats import shapiro
stats,pval=shapiro(inf_data['BW'])
```

```
C:\Users\HP\anaconda3\lib\site-packages\scipy\stats\_morestats.py:1800: UserWarning: p-value may not be accurate for N > 5000.
  warnings.warn("p-value may not be accurate for N > 5000.")
```

```
pval
```

```
0.0
```

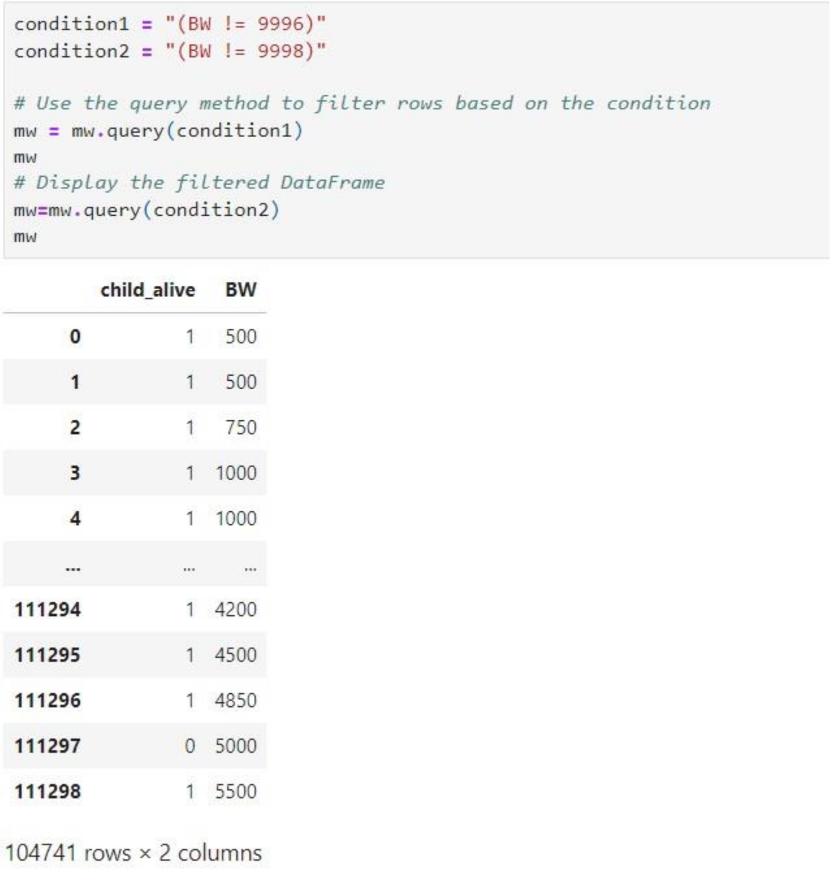
```
stats
```

```
0.5233469009399414
```

```
mw=inf_data[['child_alive','BW']]
mw.tail(1000)
```

	child_alive	BW
110385	1	1500
110386	0	1500
110387	1	1500
110388	0	1500
110389	1	1500
...	...	...
111380	1	9998
111381	1	9998
111382	1	9998
111383	1	9998
111384	1	9998

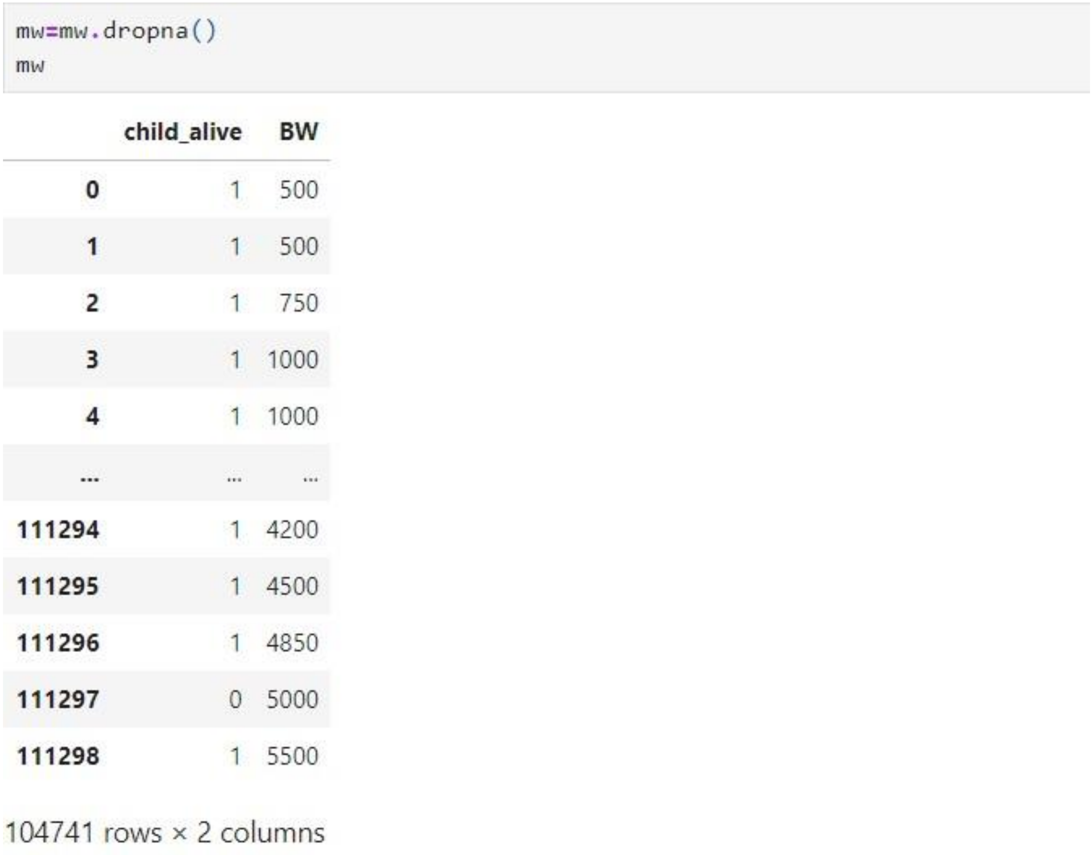
1000 rows × 2 columns



```
condition1 = "(BW != 9996)"
condition2 = "(BW != 9998)"

# Use the query method to filter rows based on the condition
mw = mw.query(condition1)
mw
# Display the filtered DataFrame
mw=mw.query(condition2)
mw
```

	child_alive	BW
0	1	500
1	1	500
2	1	750
3	1	1000
4	1	1000
...	...	...
111294	1	4200
111295	1	4500
111296	1	4850
111297	0	5000
111298	1	5500

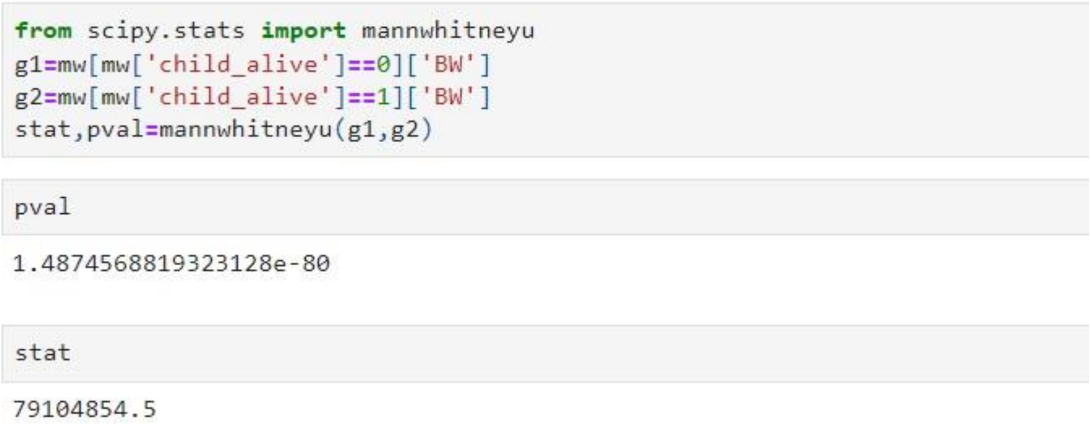
104741 rows × 2 columns



```
mw=mw.dropna()
mw
```

	child_alive	BW
0	1	500
1	1	500
2	1	750
3	1	1000
4	1	1000
...	...	...
111294	1	4200
111295	1	4500
111296	1	4850
111297	0	5000
111298	1	5500

104741 rows × 2 columns



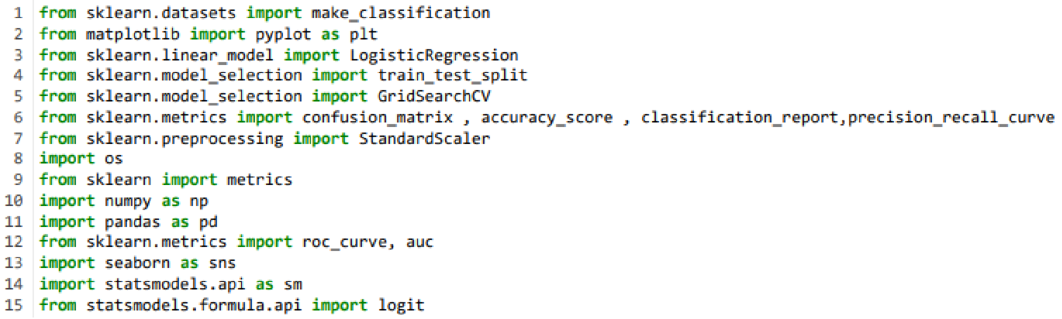
```
from scipy.stats import mannwhitneyu
g1=mw[mw['child_alive']==0]['BW']
g2=mw[mw['child_alive']==1]['BW']
stat,pval=mannwhitneyu(g1,g2)
```

```
pval
```

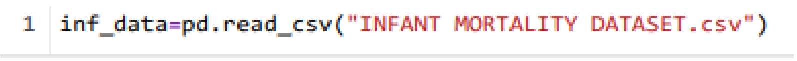
```
1.4874568819323128e-80
```

```
stat
```

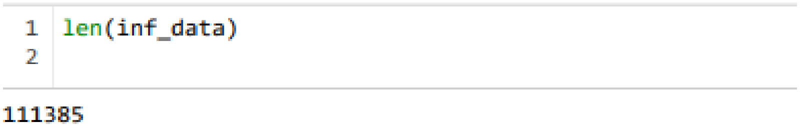
```
79104854.5
```



```
from sklearn.datasets import make_classification
from matplotlib import pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import confusion_matrix , accuracy_score , classification_report,precision_recall_curve
from sklearn.preprocessing import StandardScaler
import os
from sklearn import metrics
import numpy as np
import pandas as pd
from sklearn.metrics import roc_curve, auc
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import logit
```

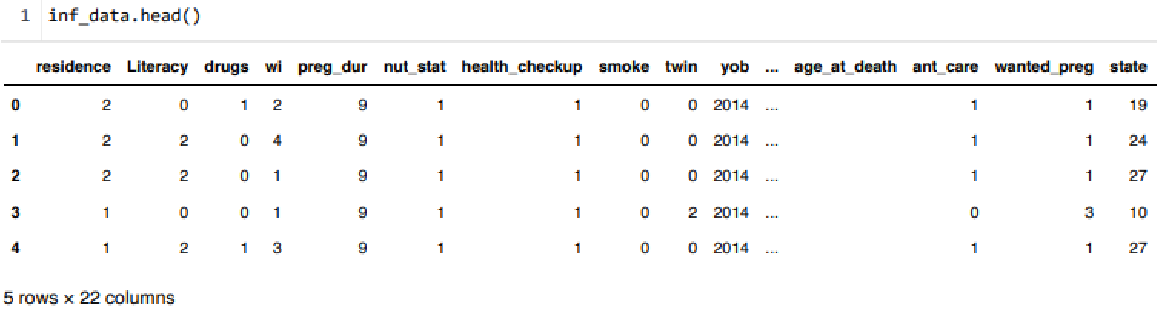


```
inf_data=pd.read_csv("INFANT MORTALITY DATASET.csv")
```



```
len(inf_data)

```

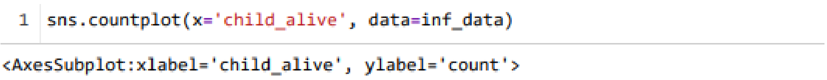
```
111385
```



```
inf_data.head()
```

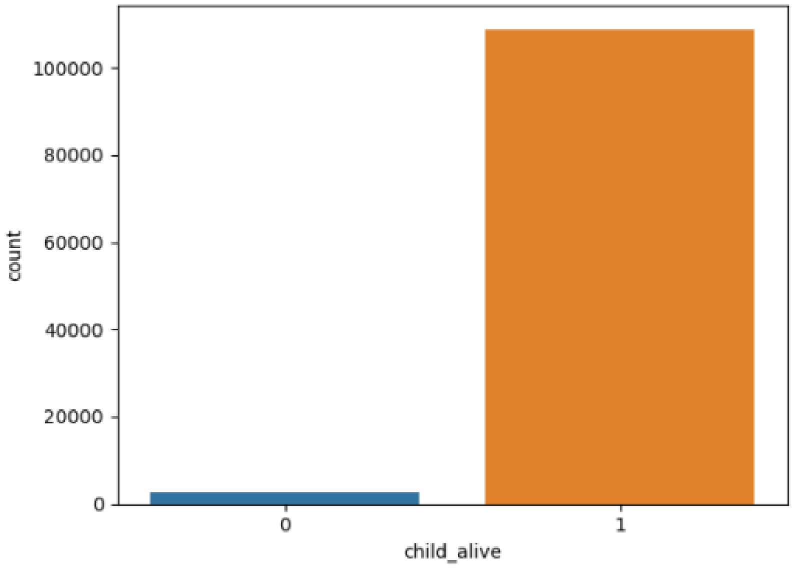
	residence	Literacy	drugs	wi	preg_dur	nut_stat	health_checkup	smoke	twin	yob	...	age_at_death	ant_care	wanted_preg	state
0	2	0	1	2	9	1	1	0	0	2014	...		1	1	19
1	2	2	0	4	9	1	1	0	0	2014	...		1	1	24
2	2	2	0	1	9	1	1	0	0	2014	...		1	1	27
3	1	0	0	1	9	1	1	0	2	2014	...		0	3	10
4	1	2	1	3	9	1	1	0	0	2014	...		1	1	27

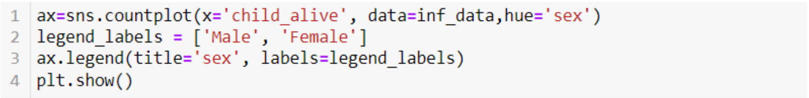
5 rows × 22 columns



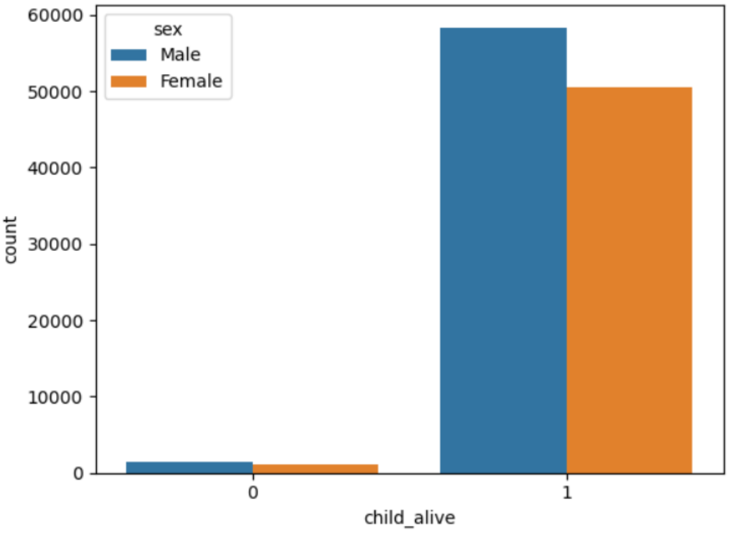
```
sns.countplot(x='child_alive', data=inf_data)
```

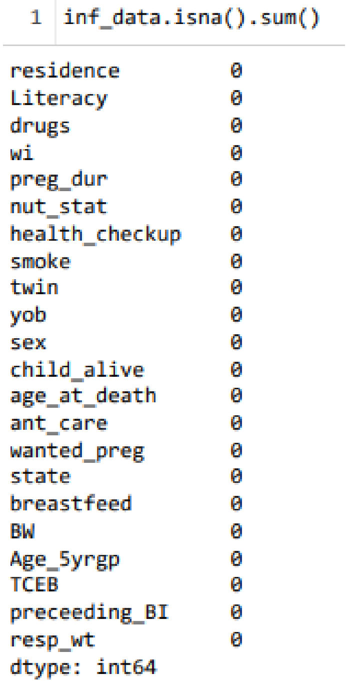
```
<AxesSubplot:xlabel='child_alive', ylabel='count'>
```





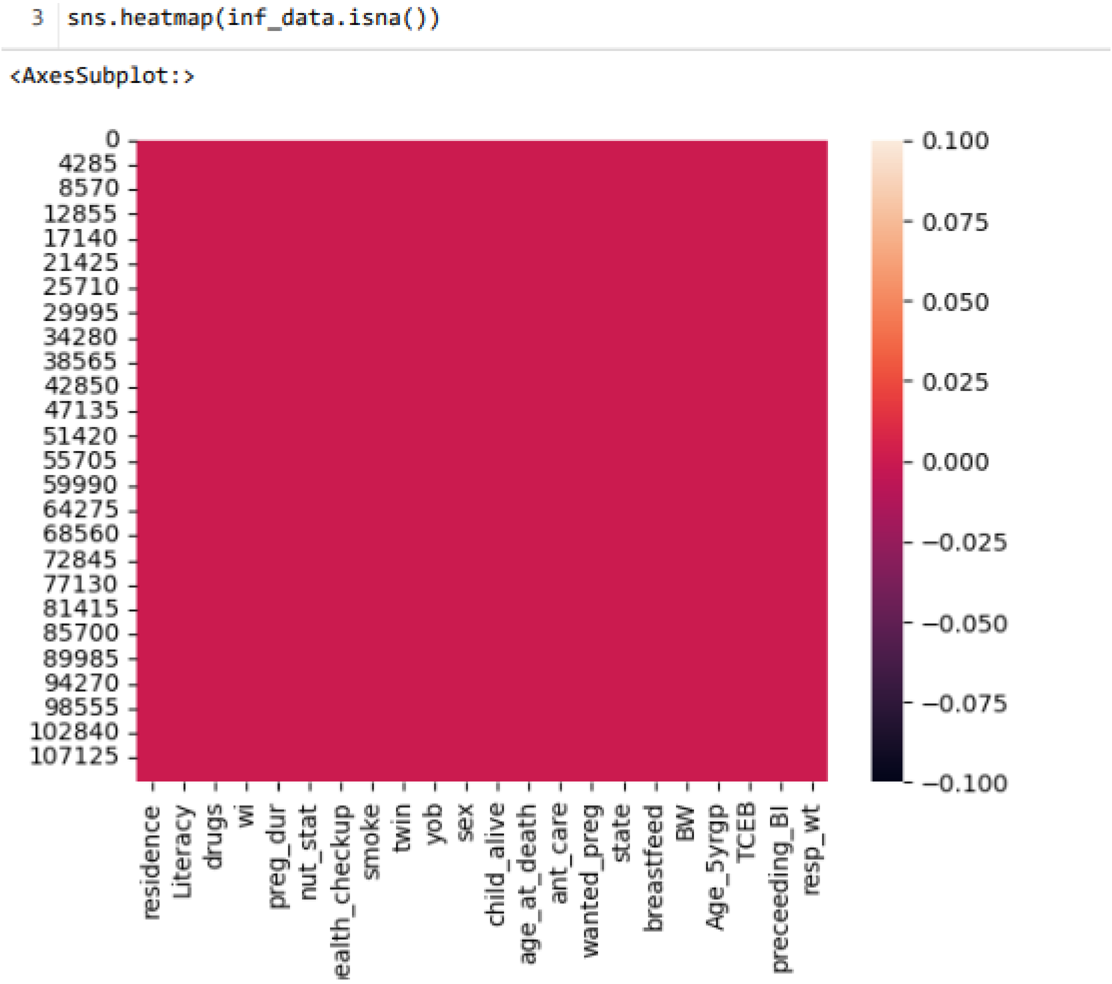
```
ax=sns.countplot(x='child_alive', data=inf_data,hue='sex')
legend_labels = ['Male', 'Female']
ax.legend(title='sex', labels=legend_labels)
plt.show()
```





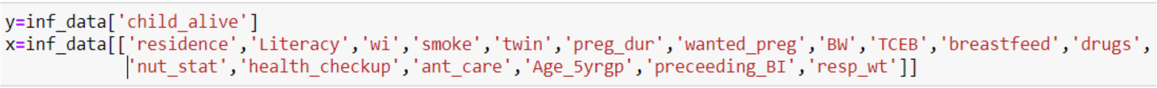
```
inf_data.isna().sum()
```

```
residence          0
Literacy           0
drugs              0
wi                 0
preg_dur           0
nut_stat           0
health_checkup     0
smoke              0
twin               0
yob                0
sex                0
child_alive        0
age_at_death       0
ant_care           0
wanted_preg        0
state              0
breastfeed         0
BW                 0
Age_5yrgp          0
TCEB               0
preceeding_BI      0
resp_wt            0
dtype: int64
```

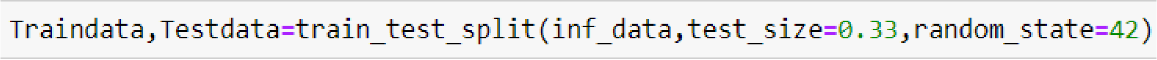


```
sns.heatmap(inf_data.isna())
```

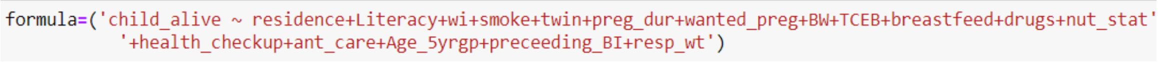
```
<AxesSubplot:>
```



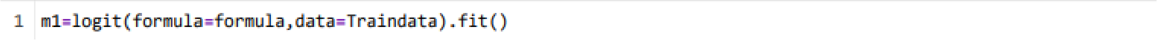
```
y=inf_data['child_alive']
x=inf_data[['residence','Literacy','wi','smoke','twin','preg_dur','wanted_preg','BW','TCEB','breastfeed','drugs',
            'nut_stat','health_checkup','ant_care','Age_5yrgp','preceeding_BI','resp_wt']]
```



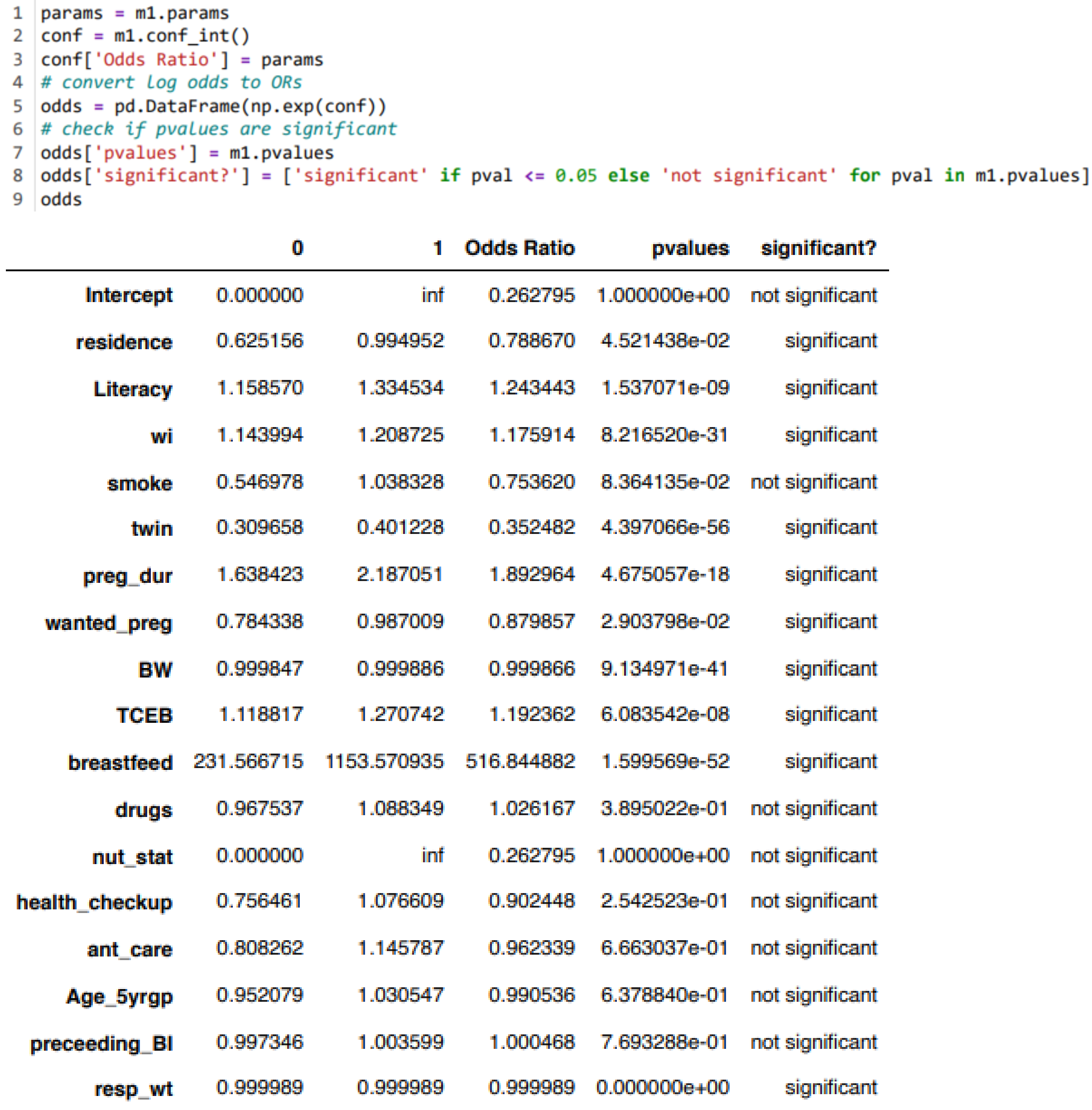
```
Traindata,Testdata=train_test_split(inf_data,test_size=0.33,random_state=42)
```



```
formula=('child_alive ~ residence+Literacy+wi+smoke+twin+preg_dur+wanted_preg+BW+TCEB+breastfeed+drugs+nut_stat'
          '+health_checkup+ant_care+Age_5yrgp+preceeding_BI+resp_wt')
```

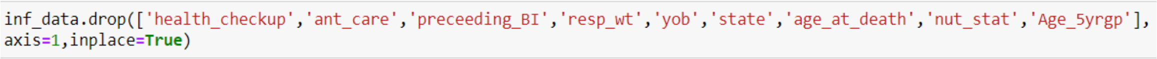


```
m1=logit(formula=formula,data=Traindata).fit()
```

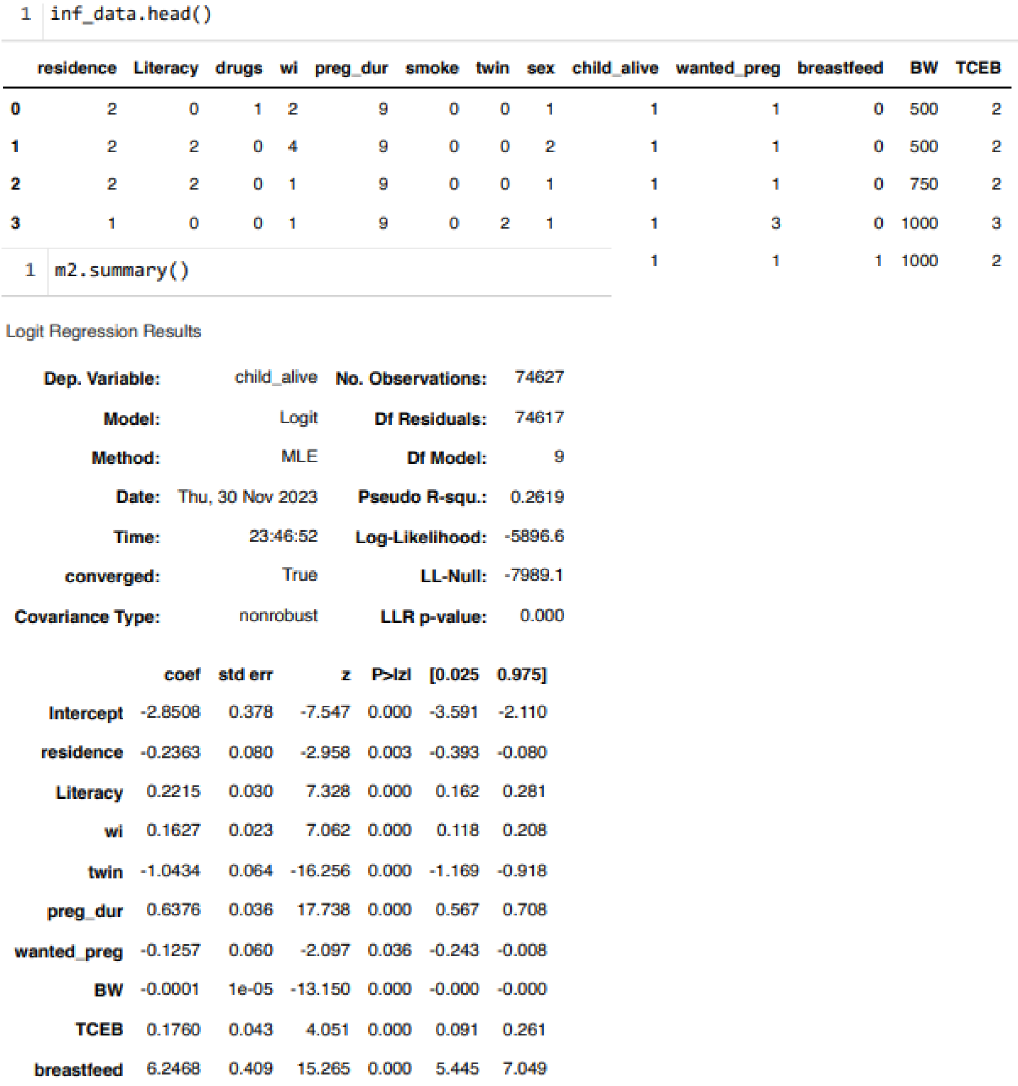


```
params = m1.params
conf = m1.conf_int()
conf['Odds Ratio'] = params
# convert log odds to ORs
odds = pd.DataFrame(np.exp(conf))
# check if pvalues are significant
odds['pvalues'] = m1.pvalues
odds['significant?'] = ['significant' if pval <= 0.05 else 'not significant' for pval in m1.pvalues]
odds
```

	0	1	Odds Ratio	pvalues	significant?
Intercept	0.000000	inf	0.262795	1.000000e+00	not significant
residence	0.625156	0.994952	0.788670	4.521438e-02	significant
Literacy	1.158570	1.334534	1.243443	1.537071e-09	significant
wi	1.143994	1.208725	1.175914	8.216520e-31	significant
smoke	0.546978	1.038328	0.753620	8.364135e-02	not significant
twin	0.309658	0.401228	0.352482	4.397066e-56	significant
preg_dur	1.638423	2.187051	1.892964	4.675057e-18	significant
wanted_preg	0.784338	0.987009	0.879857	2.903798e-02	significant
BW	0.999847	0.999886	0.999866	9.134971e-41	significant
TCEB	1.118817	1.270742	1.192362	6.083542e-08	significant
breastfeed	231.566715	1153.570935	516.844882	1.599569e-52	significant
drugs	0.967537	1.088349	1.026167	3.895022e-01	not significant
nut_stat	0.000000	inf	0.262795	1.000000e+00	not significant
health_checkup	0.756461	1.076609	0.902448	2.542523e-01	not significant
ant_care	0.808262	1.145787	0.962339	6.663037e-01	not significant
Age_5yrgp	0.952079	1.030547	0.990536	6.378840e-01	not significant
preceeding_BI	0.997346	1.003599	1.000468	7.693288e-01	not significant
resp_wt	0.999989	0.999989	0.999989	0.000000e+00	significant



```
inf_data.drop(['health_checkup','ant_care','preceeding_BI','resp_wt','yob','state','age_at_death','nut_stat','Age_5yrgp'],
axis=1,inplace=True)
```

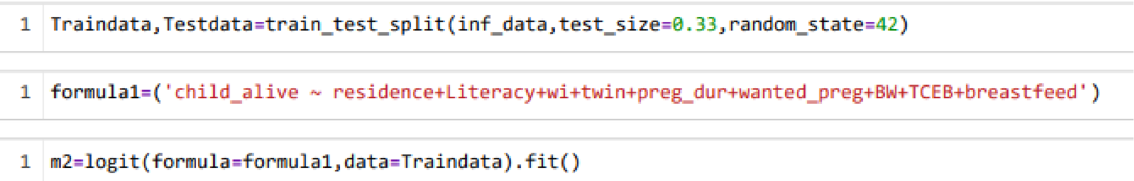


```
inf_data.head()
```

```
   residence  Literacy  drugs  wi  preg_dur  smoke  twin  sex  child_alive  wanted_preg  breastfeed    BW  TCEB
0          2         0      1   2         9      0     0    1            1            1           0   500     2
1          2         2      0   4         9      0     0    2            1            1           0   500     2
2          2         2      0   1         9      0     0    1            1            1           0   750     2
3          1         0      0   1         9      0     2    1            1            3           0  1000     3
                                                                         1            1           1  1000     2
```

```
m2.summary()
```

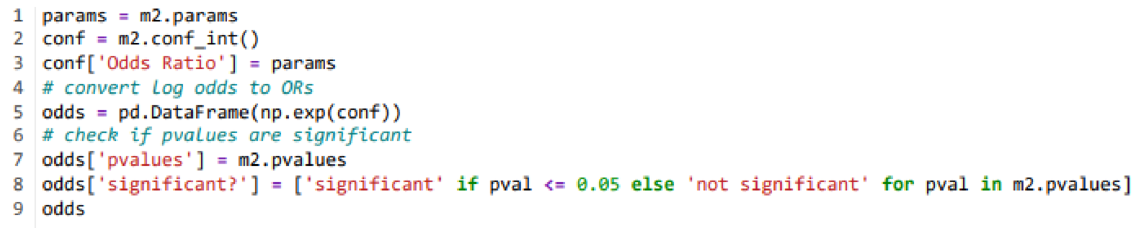
```
Logit Regression Results

   Dep. Variable:        child_alive   No. Observations:   74627
           Model:              Logit       Df Residuals:   74617
          Method:                MLE           Df Model:       9
            Date:   Thu, 30 Nov 2023     Pseudo R-squ.:   0.2619
            Time:           23:46:52     Log-Likelihood:  -5896.6
       converged:               True            LL-Null:  -7989.1
 Covariance Type:          nonrobust        LLR p-value:    0.000

                 coef   std err        z   P>|z|   [0.025   0.975]
  Intercept   -2.8508     0.378   -7.547   0.000   -3.591   -2.110
  residence   -0.2363     0.080   -2.958   0.003   -0.393   -0.080
   Literacy    0.2215     0.030    7.328   0.000    0.162    0.281
         wi    0.1627     0.023    7.062   0.000    0.118    0.208
       twin   -1.0434     0.064  -16.256   0.000   -1.169   -0.918
   preg_dur    0.6376     0.036   17.738   0.000    0.567    0.708
wanted_preg   -0.1257     0.060   -2.097   0.036   -0.243   -0.008
         BW   -0.0001     1e-05  -13.150   0.000   -0.000   -0.000
       TCEB    0.1760     0.043    4.051   0.000    0.091    0.261
 breastfeed    6.2468     0.409   15.265   0.000    5.445    7.049
```



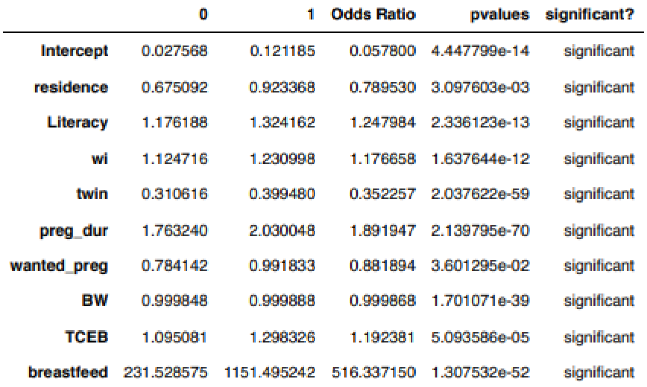
```
Traindata,Testdata=train_test_split(inf_data,test_size=0.33,random_state=42)
```

```
formula1=('child_alive ~ residence+Literacy+wi+twin+preg_dur+wanted_preg+BW+TCEB+breastfeed')
```

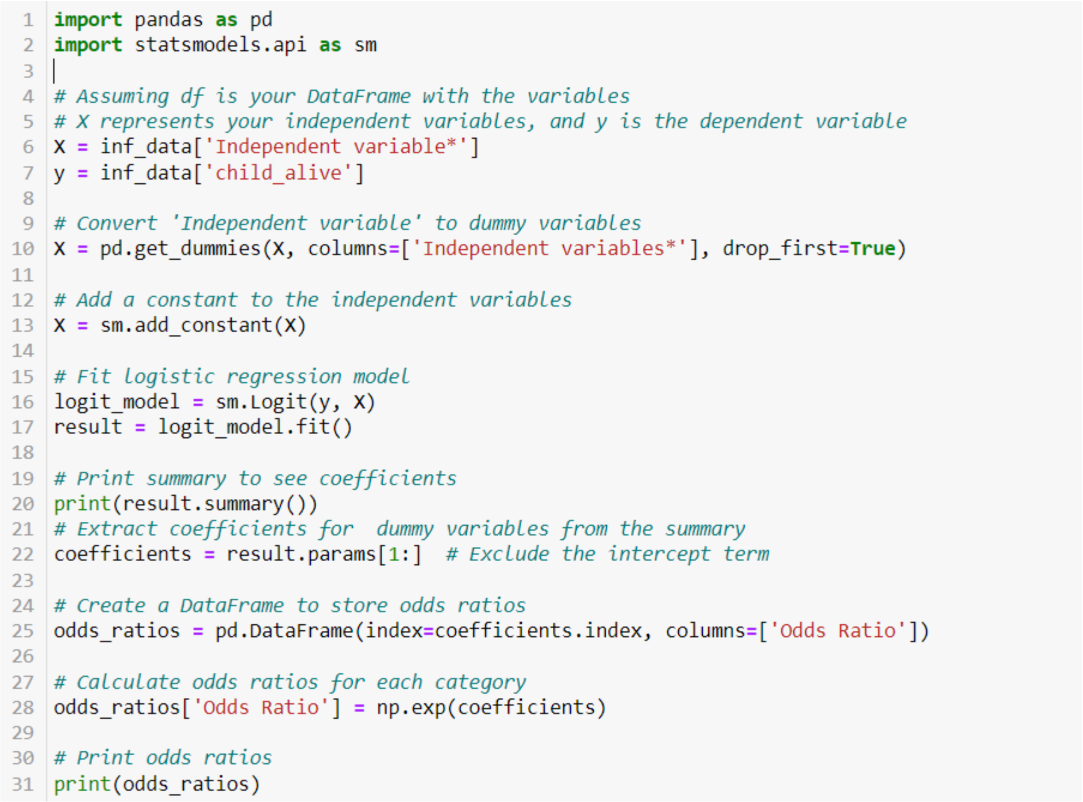
```
m2=logit(formula=formula1,data=Traindata).fit()
```



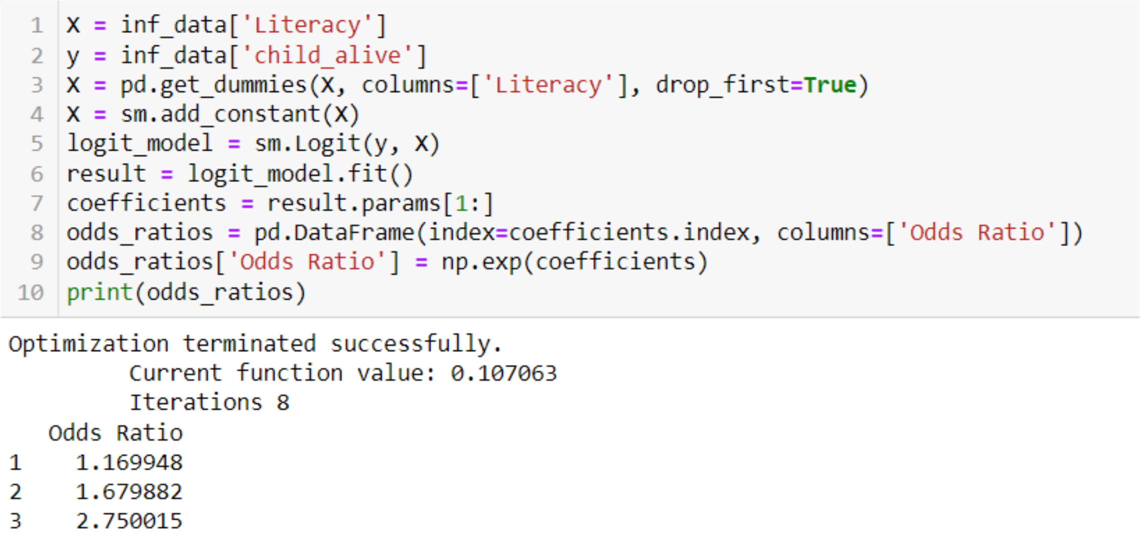
```
params = m2.params
conf = m2.conf_int()
conf['Odds Ratio'] = params
# convert Log odds to ORs
odds = pd.DataFrame(np.exp(conf))
# check if pvalues are significant
odds['pvalues'] = m2.pvalues
odds['significant?'] = ['significant' if pval <= 0.05 else 'not significant' for pval in m2.pvalues]
odds
```



	0	1	Odds Ratio	pvalues	significant?
Intercept	0.027568	0.121185	0.057800	4.447799e-14	significant
residence	0.675092	0.923368	0.789530	3.097603e-03	significant
Literacy	1.176188	1.324162	1.247984	2.336123e-13	significant
wi	1.124716	1.230998	1.176658	1.637644e-12	significant
twin	0.310616	0.399480	0.352257	2.037622e-59	significant
preg_dur	1.763240	2.030048	1.891947	2.139795e-70	significant
wanted_preg	0.784142	0.991833	0.881894	3.601295e-02	significant
BW	0.999848	0.999888	0.999868	1.701071e-39	significant
TCEB	1.095081	1.298326	1.192381	5.093586e-05	significant
breastfeed	231.528575	1151.495242	516.337150	1.307532e-52	significant

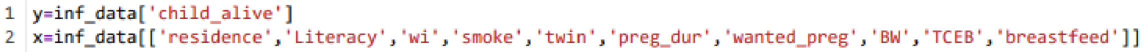


```python
import pandas as pd
import statsmodels.api as sm

# Assuming df is your DataFrame with the variables
# X represents your independent variables, and y is the dependent variable
X = inf_data['Independent variable*']
y = inf_data['child_alive']

# Convert 'Independent variable' to dummy variables
X = pd.get_dummies(X, columns=['Independent variables*'], drop_first=True)

# Add a constant to the independent variables
X = sm.add_constant(X)

# Fit logistic regression model
logit_model = sm.Logit(y, X)
result = logit_model.fit()

# Print summary to see coefficients
print(result.summary())
# Extract coefficients for  dummy variables from the summary
coefficients = result.params[1:]  # Exclude the intercept term

# Create a DataFrame to store odds ratios
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])

# Calculate odds ratios for each category
odds_ratios['Odds Ratio'] = np.exp(coefficients)

# Print odds ratios
print(odds_ratios)
```



```python
X = inf_data['Literacy']
y = inf_data['child_alive']
X = pd.get_dummies(X, columns=['Literacy'], drop_first=True)
X = sm.add_constant(X)
logit_model = sm.Logit(y, X)
result = logit_model.fit()
coefficients = result.params[1:]
odds_ratios = pd.DataFrame(index=coefficients.index, columns=['Odds Ratio'])
odds_ratios['Odds Ratio'] = np.exp(coefficients)
print(odds_ratios)
```

```
Optimization terminated successfully.
         Current function value: 0.107063
         Iterations 8
   Odds Ratio
1    1.169948
2    1.679882
3    2.750015
```



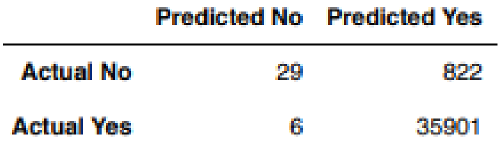
```
y=inf_data['child_alive']
x=inf_data[['residence','Literacy','wi','smoke','twin','preg_dur','wanted_preg','BW','TCEB','breastfeed']]
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.33,random_state=42)
```

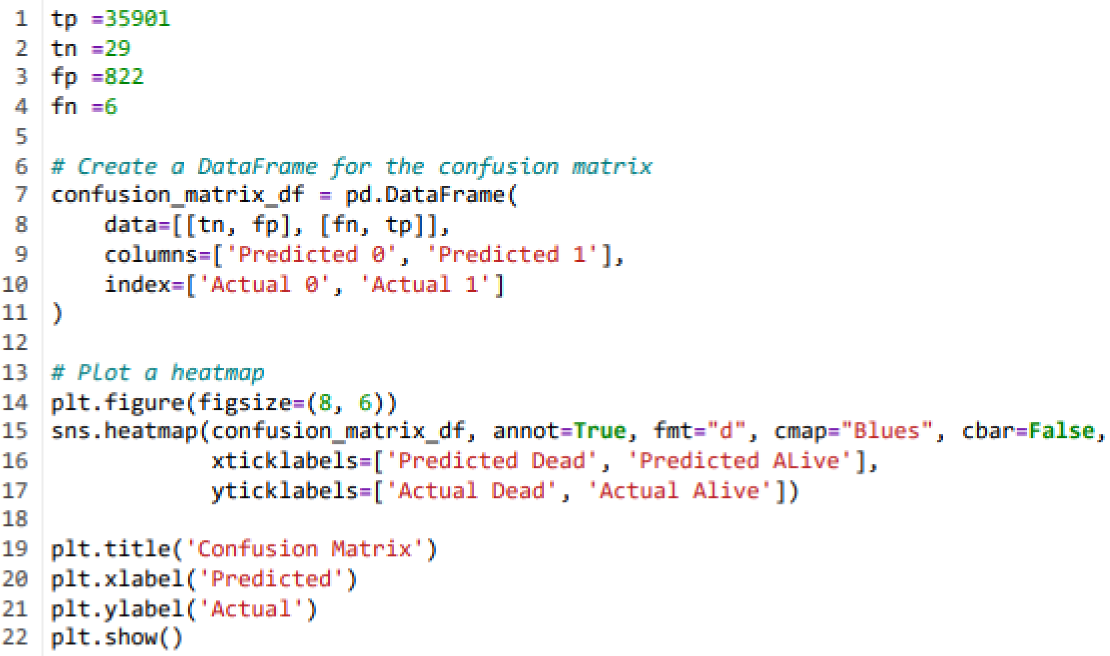
```
lrm.fit(x_train,y_train)
```

```
predict=lrm.predict(x_test)
```

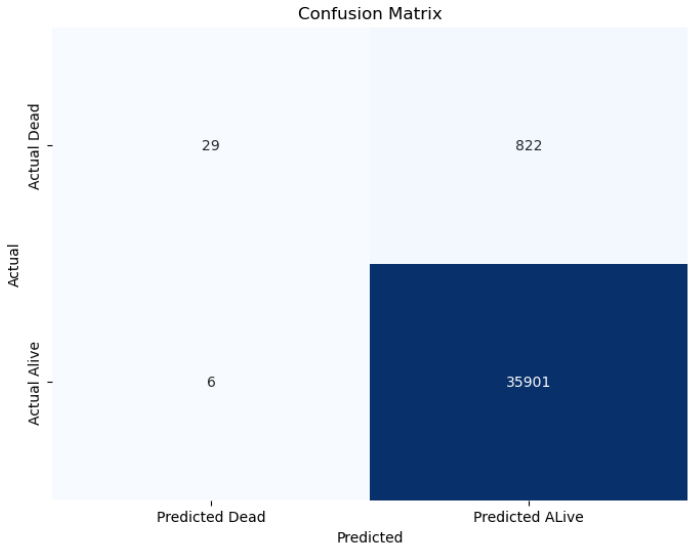
```
pd.DataFrame(confusion_matrix(y_test,predict),columns=['Predicted No','Predicted Yes'],index=['Actual No','Actu
```

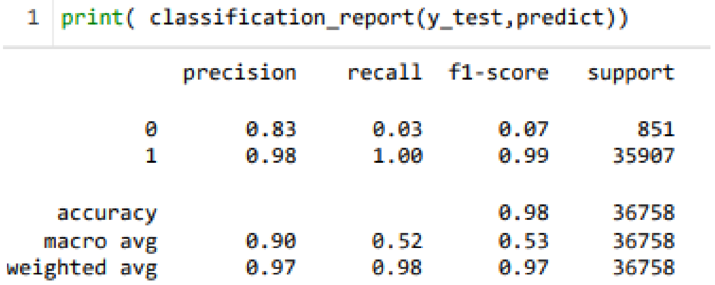


	Predicted No	Predicted Yes
Actual No	29	822
Actual Yes	6	35901



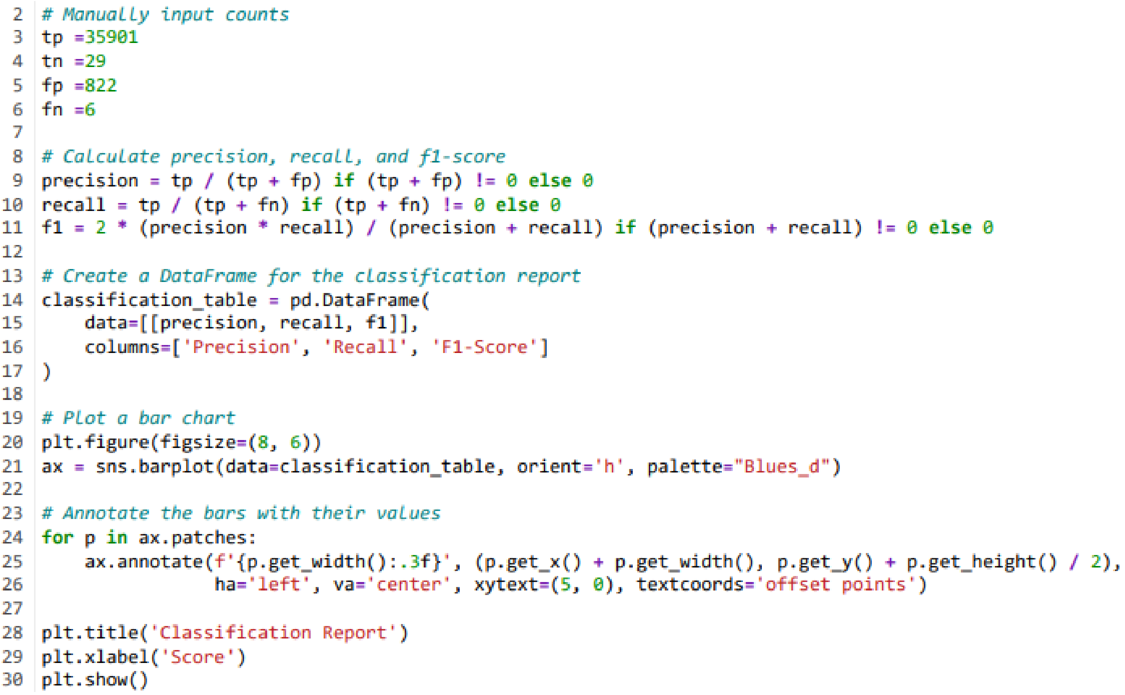
```
tp =35901
tn =29
fp =822
fn =6

# Create a DataFrame for the confusion matrix
confusion_matrix_df = pd.DataFrame(
    data=[[tn, fp], [fn, tp]],
    columns=['Predicted 0', 'Predicted 1'],
    index=['Actual 0', 'Actual 1']
)

# Plot a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(confusion_matrix_df, annot=True, fmt="d", cmap="Blues", cbar=False,
            xticklabels=['Predicted Dead', 'Predicted ALive'],
            yticklabels=['Actual Dead', 'Actual Alive'])

plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```



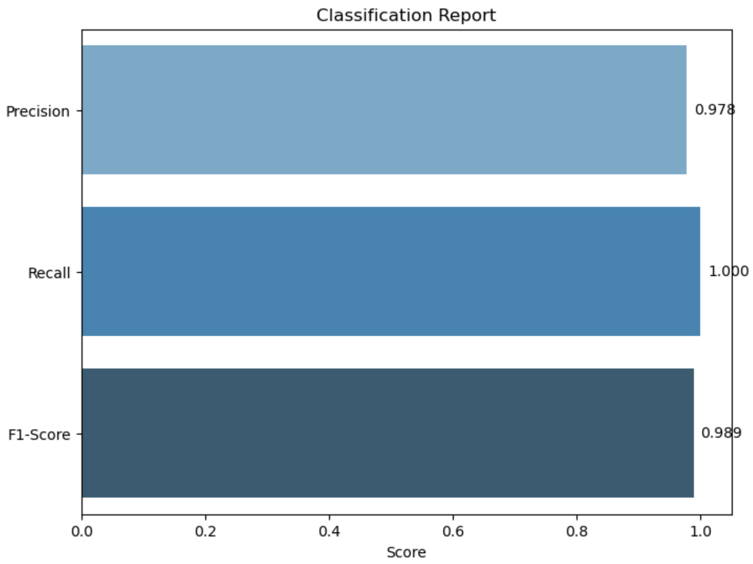


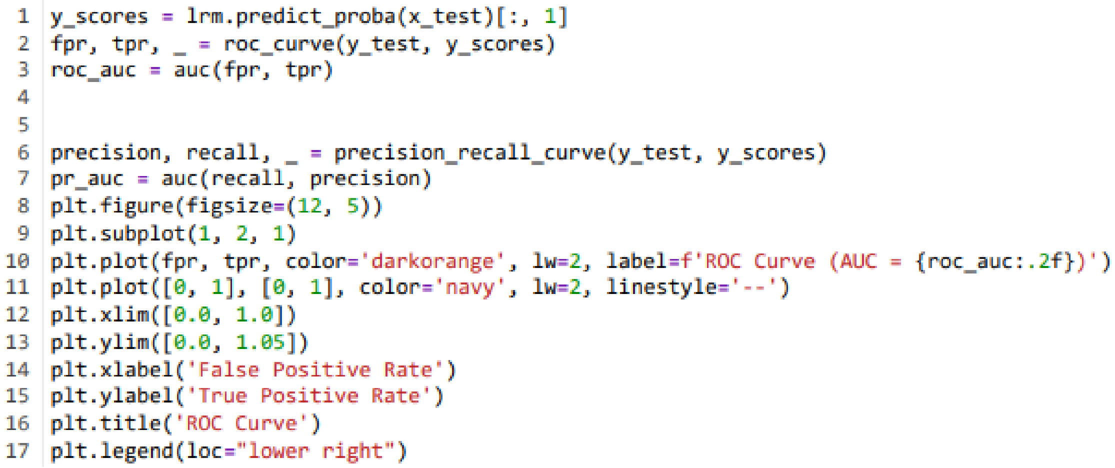
```
print( classification_report(y_test,predict))
```

```
              precision    recall  f1-score   support

           0       0.83      0.03      0.07       851
           1       0.98      1.00      0.99     35907

    accuracy                           0.98     36758
   macro avg       0.90      0.52      0.53     36758
weighted avg       0.97      0.98      0.97     36758
```

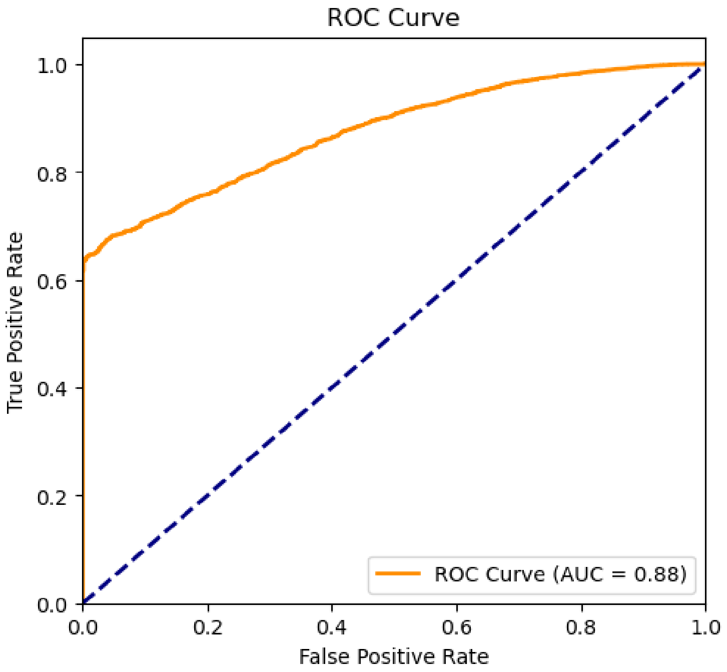


```
# Manually input counts
tp =35901
tn =29
fp =822
fn =6

# Calculate precision, recall, and f1-score
precision = tp / (tp + fp) if (tp + fp) != 0 else 0
recall = tp / (tp + fn) if (tp + fn) != 0 else 0
f1 = 2 * (precision * recall) / (precision + recall) if (precision + recall) != 0 else 0

# Create a DataFrame for the classification report
classification_table = pd.DataFrame(
    data=[[precision, recall, f1]],
    columns=['Precision', 'Recall', 'F1-Score']
)

# Plot a bar chart
plt.figure(figsize=(8, 6))
ax = sns.barplot(data=classification_table, orient='h', palette="Blues_d")

# Annotate the bars with their values
for p in ax.patches:
    ax.annotate(f'{p.get_width():.3f}', (p.get_x() + p.get_width(), p.get_y() + p.get_height() / 2),
                ha='left', va='center', xytext=(5, 0), textcoords='offset points')

plt.title('Classification Report')
plt.xlabel('Score')
plt.show()
```





```
y_scores = lrm.predict_proba(x_test)[:, 1]
fpr, tpr, _ = roc_curve(y_test, y_scores)
roc_auc = auc(fpr, tpr)


precision, recall, _ = precision_recall_curve(y_test, y_scores)
pr_auc = auc(recall, precision)
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC Curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
```


ROC Curve
True Positive Rate
False Positive Rate
ROC Curve (AUC = 0.88)