### Hoang Anh Nguyen(Jolie).

Final project.

### **A. Introduction**

This project explores the relationship between the heights of parents and their children, using data from Francis Galton’s 1885 study. The focus is on identifying correlations between parental heights and child height and examining how these relationships differ between boys and girls. Height is influenced by many factors, but this project specifically investigates parental height and gender as key contributors.

The research questions are: “Is there a correlation between the heights of parents and their children?” and “How does this relationship differ between boys and girls?” To answer these questions, I analyzed data from Kaggle which provide me with the heights of fathers, mothers, and children, along with the gender of each child. Methods included calculating averages, correlations, and performing hypothesis testing like t-tests and regression analysis to better understand the relationships.

The analysis involved loading and cleaning the data, performing calculations, and visualizing results using scatter plots and bar charts. The goal was to see if taller parents tend to have taller children and whether boys are generally taller than girls. This project also discusses the limitations of the study, such as the role of other factors like nutrition or environment. The findings provide a statistical understanding of height patterns and their significance.

### **B. Background**

Height is something we often inherit from our parents, but it’s not just about genetics. Other things, like how well we eat, how healthy we are, and the environment we grow up in, also play a big role. This project focuses on figuring out how much of a child’s height is influenced by their parents’ heights, while keeping in mind that other factors might also matter.

Francis Galton’s 1885 study is a classic example of research on how traits like height are passed down through families. It collected data on the heights of fathers, mothers, and their adult children, making it possible to see patterns between parent and child heights. Galton’s work was one of the first to show how we can use numbers and statistics to understand heredity, and it’s still used today to study how traits are inherited.

Francis Galton studied this data to understand how traits like height are passed from parents to children. He focused on height because it’s easy to measure and shows clear variation. By analyzing family data, Galton discovered regression toward the mean, meaning very tall or short parents tend to have children closer to average height.

This work was part of Galton’s larger effort to use math and statistics to study heredity. He developed concepts like correlation and regression, which became key tools in understanding how traits are inherited. This dataset was a crucial step in connecting biology with statistics.

Gender is another important factor when it comes to height. Boys are generally taller than girls, especially after puberty, because of differences in hormones like testosterone, which promotes more growth in boys. This project looks at how gender impacts child height and whether boys and girls with the same parental heights end up growing to different average heights. By studying this, we can better understand how gender and parental heights together shape a child’s height.

### **C. Data and Methods**

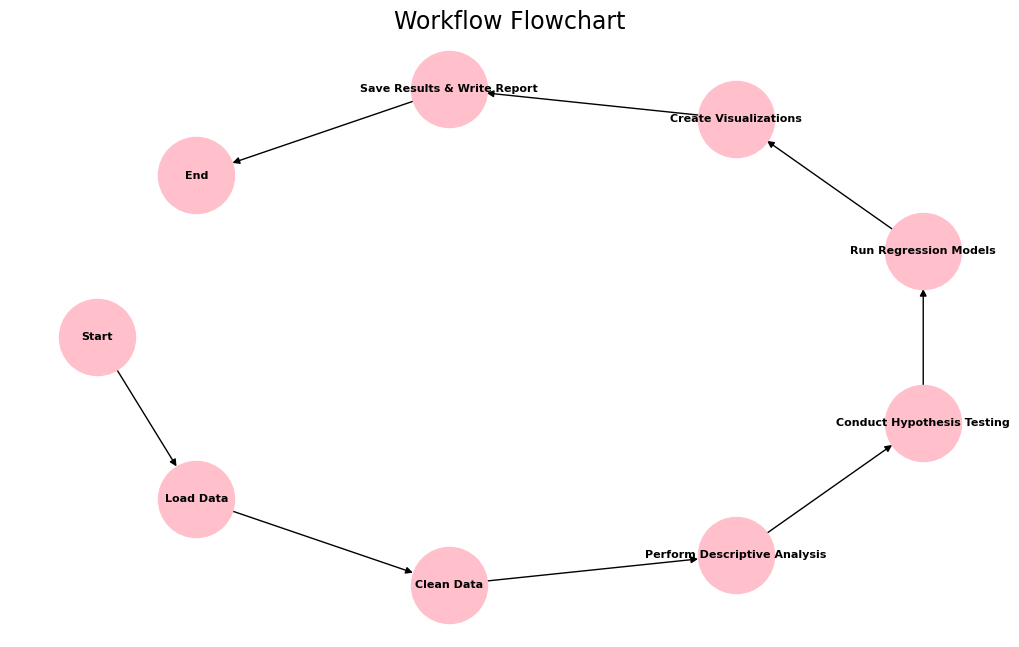
This project uses data from Francis Galton’s 1885 study, which is available on Kaggle. The dataset includes information about the heights of fathers, mothers, and their children, as well as the gender of each child. This data makes it possible to explore how much parents’ heights influence their child’s height and whether boys and girls grow differently. The dataset is neatly organized into columns such as: “Father\_height,” “Mother\_height,” “Child\_height,” and “gender,” making it easy to work with and analyze.

To start, I loaded the data into Python using Pandas to make it easier to work with. I cleaned the data by checking for any missing or incorrect values and made sure all the height measurements were in the right format. I also encoded gender as 0 for females and 1 for males so that it could be used in calculations. These steps helped prepare the dataset for analysis and made sure everything was ready for creating graphs and running statistical tests.

The analysis was done in three main steps. First, I calculated the averages (mean) and performed descriptive statistics to get an overall understanding of the data. I created a bar chart to compare the average heights of boys and girls to see if boys are generally taller. I also created scatter plots to check if there appeared to be a correlation between parental heights and child height. These visualizations helped identify basic patterns in the data.

Next, I performed hypothesis testing to look deeper into these relationships. For the first part, I used correlation coefficients and p-values to test whether I had enough evidence to say that child height is influenced by parental heights. Then, I performed a t-test and used the p-value to see if the difference in height between boys and girls was statistically significant. This helped determine whether gender plays a major role in explaining child height.

Finally, I used regression analysis to quantify these relationships. I first ran a linear regression to see how much parental heights contribute to predicting a child’s height. This showed how much the father’s and mother’s heights explain the child’s height. Then, I ran a logistic regression using gender (encoded as 0 for females and 1 for males) to see how much gender contributes to explaining child height. This combination of methods provided a clear picture of the roles of parental heights and gender in determining child height.



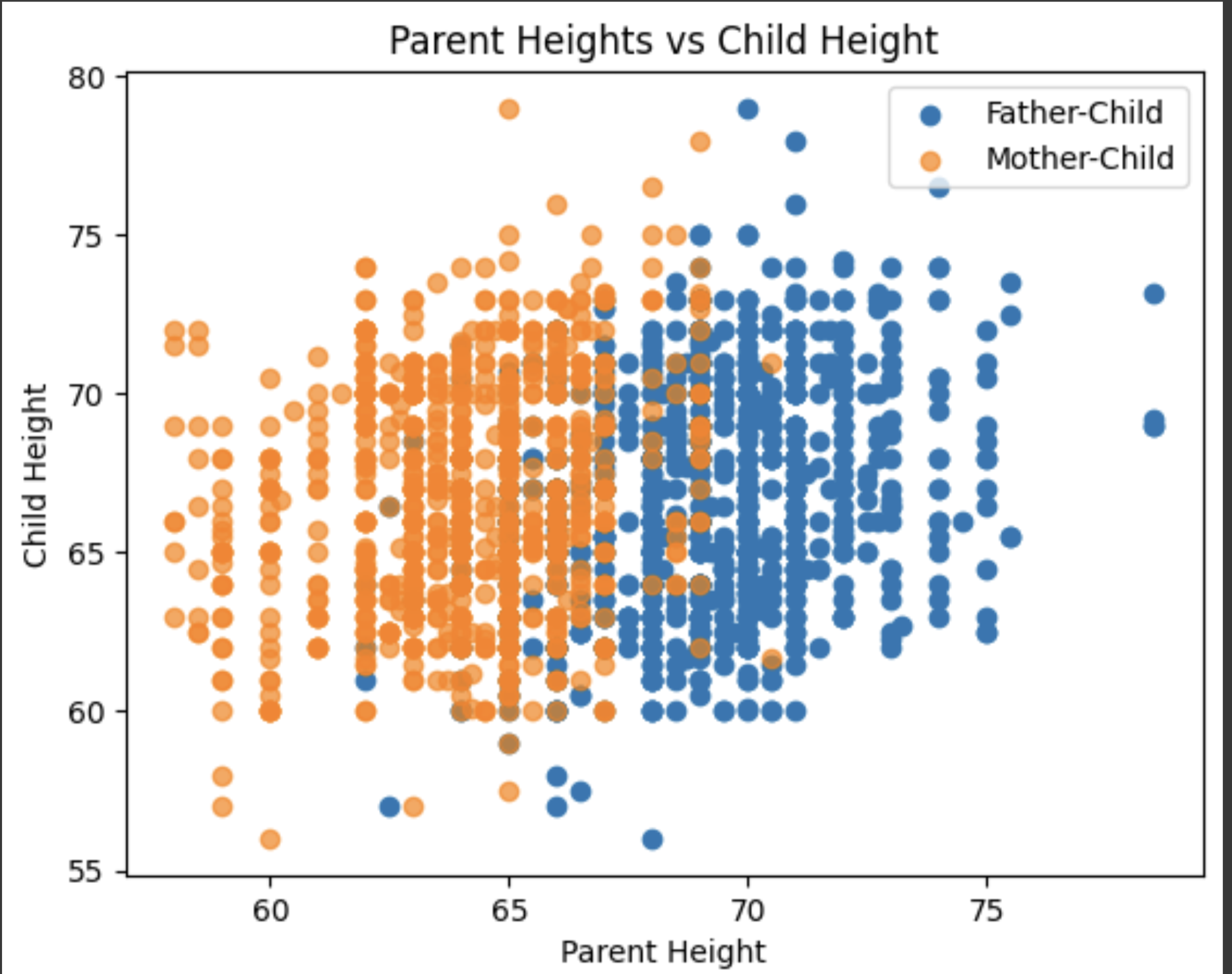
( Flowchart illustrating the workflow used in the analysis ).

### **D. Results**

#### **1. Correlation Analysis**

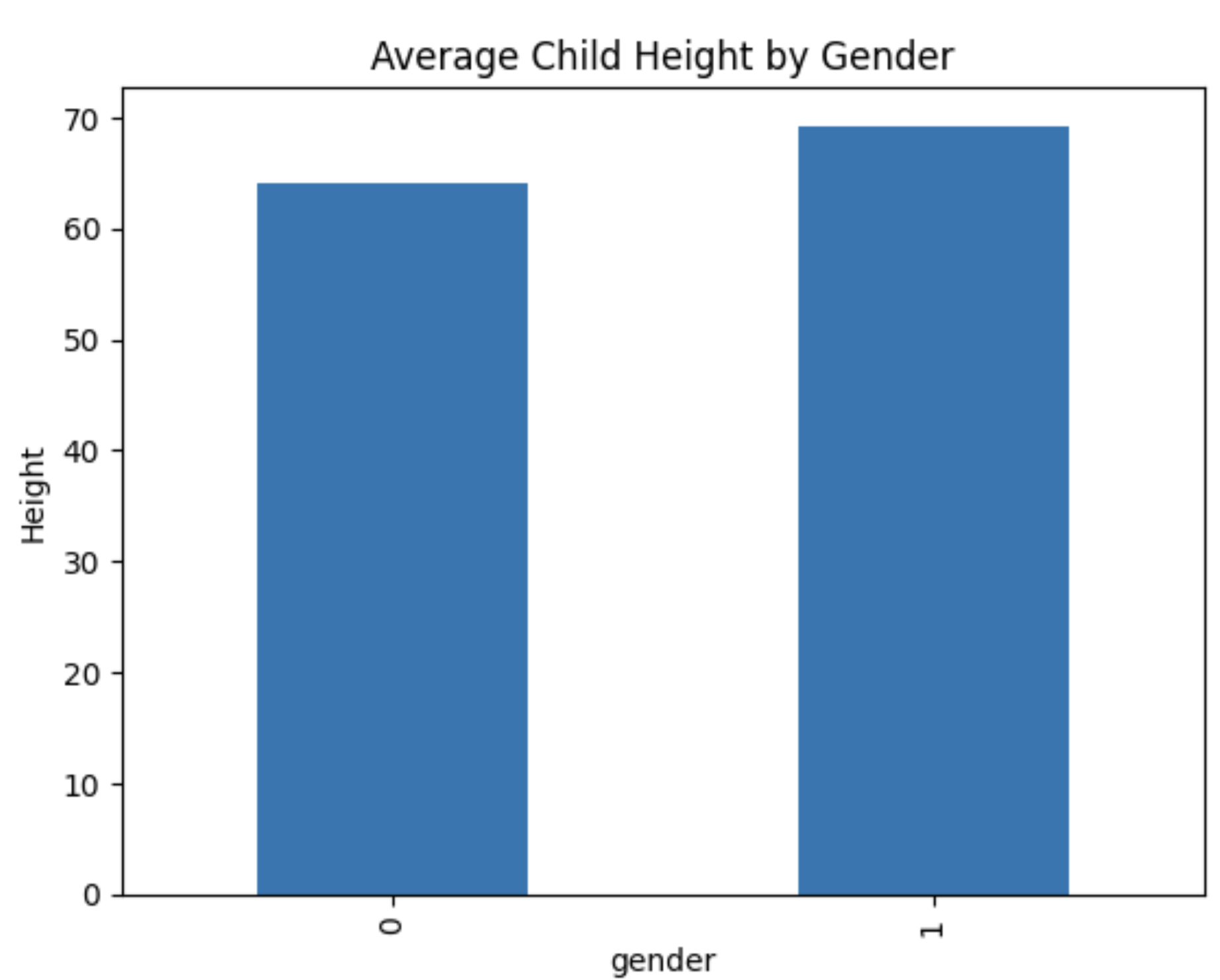
The correlation results show a weak positive relationship between both parental heights and child height:

* Father-Child Correlation: 0.266 with a p-value of 1.35e-16, indicating that the relationship is statistically significant.
* Mother-Child Correlation: 0.201 with a p-value of 5.36e-10, also statistically significant.  
  These results suggest that taller parents are likely to have taller children, but the influence is relatively weak.



This scatter plot shows the relationship between parents’ heights and their child’s height. Blue dots are for fathers, and orange dots are for mothers. There’s a slight upward trend, meaning taller parents tend to have taller children, but the spread of the points shows the relationship is weak. Other factors, like gender, likely influence child height too.

#### **2. T-Test for Gender Differences**

The t-test comparing the average heights of boys and girls showed a T-test Statistic of 31.38 with a p-value of 4.68e-148. This confirms that boys are, on average, significantly taller than girls, and the difference is statistically significant. This highlights that gender plays a key role in determining child height.

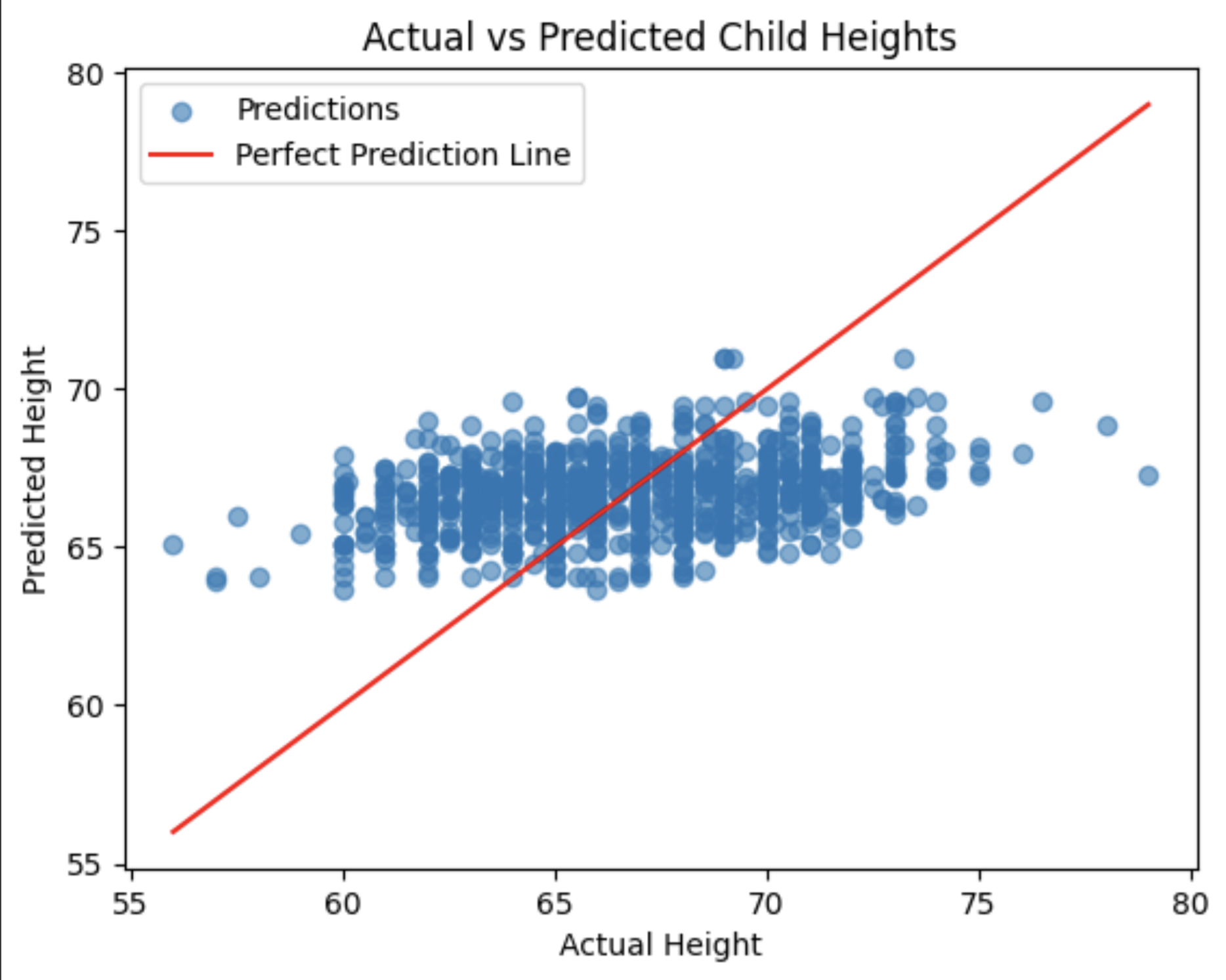
This bar chart shows the average height of children by gender. The left bar (labeled 0) represents girls, and the right bar (labeled 1) represents boys. The chart shows that boys, on average, are taller than girls. This matches the statistical tests, which confirmed that the height difference between boys and girls is significant. It highlights that gender plays an important role in determining a child’s height.

#### **3. Linear Regression Analysis**

* Parental Heights and Child Height:  
  The regression model for parental heights showed:
  1. Intercept: 22.64
  2. Father Height Coefficient: 0.368
  3. Mother Height Coefficient: 0.291
  4. R² Score: 0.105

This indicates that both parents’ heights contribute to predicting child height, but they only explain 10.5% of the variability. This means other factors, such as gender, nutrition, or environment, likely play a significant role.

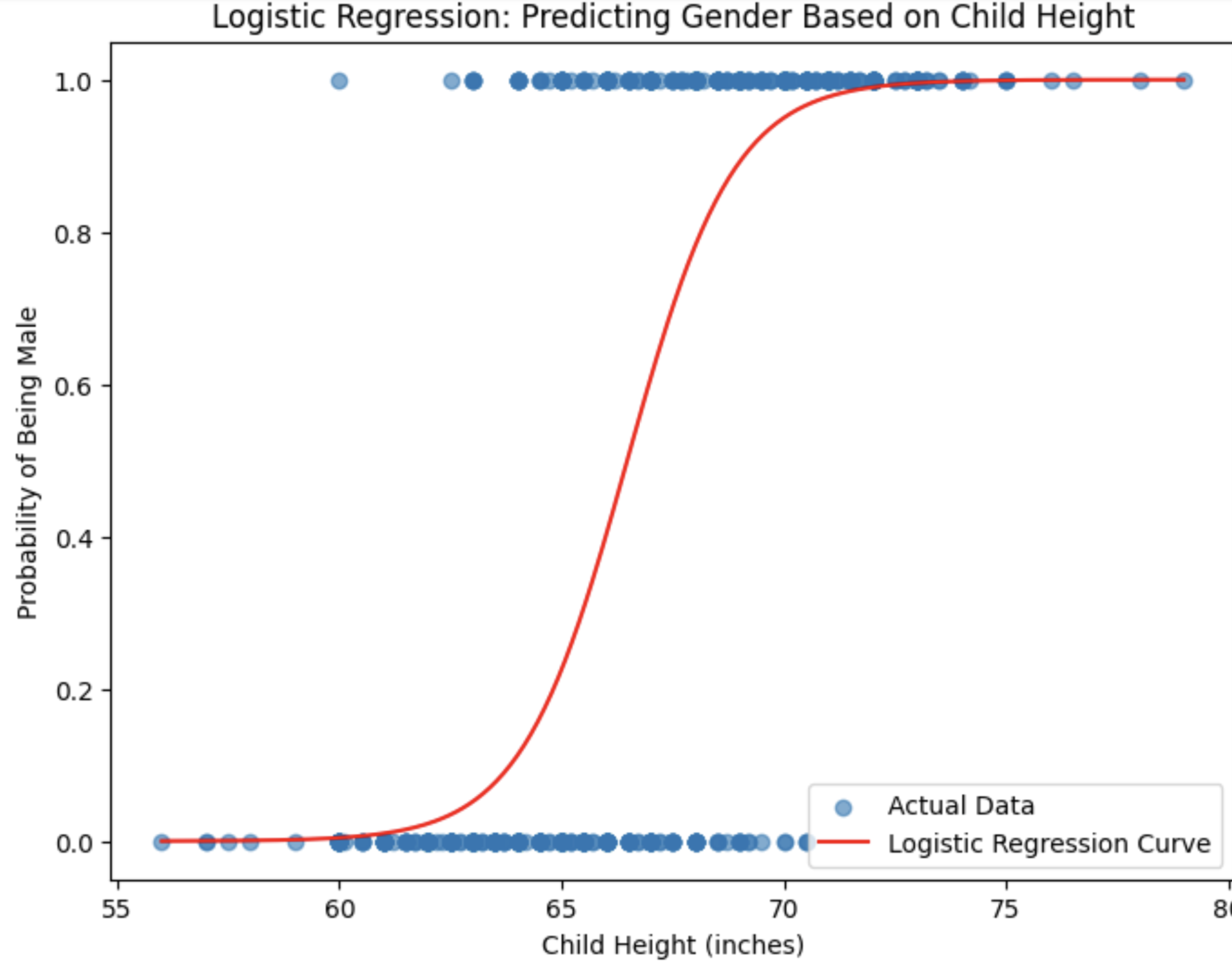
* Parental Heights and Gender Combined:  
  When gender was added to the model, the regression results showed:
  1. Intercept: 16.52
  2. Father Height Coefficient: 0.393
  3. Mother Height Coefficient: 0.318
  4. Gender Coefficient: 5.21
  5. R² Score: 0.635

The addition of gender significantly improved the model’s ability to explain child height, with the R² score increasing to 63.5%. The gender coefficient (5.21) shows that boys are, on average, about 5 inches taller than girls, even after accounting for parental heights.

This plot compares actual child heights (x-axis) with predicted heights (y-axis). The blue dots are predictions, and the red line shows perfect predictions. Most dots are close to the line, meaning the model predicts fairly well, but it’s not perfect. This reflects the R² score of 0.635, showing other factors also affect child height.

#### **4. Key Findings**

* Parental heights are positively correlated with child height, but the relationships are weak.
* Boys are significantly taller than girls, with gender having a strong influence on height.
* Combining parental heights and gender in a regression model greatly improves the ability to predict child height.



This plot shows how child height predicts the probability of being male. The blue dots are actual data points, and the red curve is the logistic regression prediction. As child height increases, the probability of being male also increases. Shorter children are more likely female (near 0 probability), while taller children are more likely male (near 1 probability). The sharp curve shows that height is a strong indicator of gender in this dataset.

These results provide a clear picture of how both hereditary and biological factors contribute to child height while acknowledging that other variables may also be important.

### **E. Conclusion**

This project showed that both parents’ heights and gender affect a child’s height. The data confirmed that taller parents tend to have taller children, but the connection is weak. The t-test showed that boys are, on average, taller than girls, and gender plays a big role in predicting height. When gender was added to the regression, the model explained a much bigger part of the variation in child height.

In short, both parental heights and gender matter, but other things like nutrition or environment might also be important. This project helped show how these factors work together to influence height. Future research could look at more details, like how health or lifestyle affects height.

**Citation:**

Fundal. (n.d.). *Galton’s Height Data - Multiple Linear Regression* [Data set]. Kaggle.  
Available at: https://www.kaggle.com/datasets/fundal/galtons-height-data-multiple-linear-regression