ANALYZING HEIGHT PATTERNS IN FAMILIES

-A Data Analysis of Family Heights

Submitted by:

Kohana Bhalla (20231229)
Kartik Semwal (20231228)
Keshav Sharma (20231270)
Raaheel Aamir (20231240)
Varenya Arora (20231250)
Sameer (20231246)

Supervisor/Instructor:

Prof. Killa Anil Kumar.

Course and College:

B.Sc.(H) Statistics
Ramanujan College, University of Delhi

Date of Submission:

15 April 2025

INTRODUCTION

This project utilizes a comprehensive dataset of family height measurements to explore patterns and relationships between the heights of parents and their children. The dataset includes recorded heights of fathers and mothers, as well as detailed information on one or more children in each family, including their sex, age, and height.

The primary goal of this analysis is to investigate potential hereditary trends in height, understand the influence of parental height on children's height, and assess whether variables such as gender and age play significant roles in height development. This allows rich, multi-dimensional analysis that can accommodate sibling comparisons and age-based growth trends within families. By leveraging this data, we aim to examine patterns in height inheritance, evaluate how strongly child height correlates with that of the parents, and assess whether such patterns vary by gender or age group.

This analysis not only serves as a practical application of data science and statistical techniques but also contributes to understanding the broader questions around heredity and human development.

DATASET DESCRIPTION

The dataset used for this analysis provides detailed family-level data focused on **parental and child heights**. It appears to be structured to explore hereditary patterns in human height, making it suitable for studies in growth patterns and family genetics analysis.

Each record in the dataset includes:

- **Father's Height (cm).**
- **Mother's Height (cm).**
- **Children's Details,** including for each child:
    - **Sex** (0 = Female, 1 = Male)
    - **Age** (in years)
    - **Height** (in cm)

The cleaned dataset accommodates multiple children per family, with each child's sex, age and height listed sequentially in the same row. Some rows contain data for one child, while others include up to four children. The structure allows us to investigate patterns not only within individual parent-child pairs but also across siblings of different ages and sexes.

AIM OF THE PROJECT

This project aims to study and analyze the **Family** data and successfully find valuable insights on the following aspects:

1. To assess the Normality of the Height distributions of Fathers and Mothers.
2. To analyse the Relationship/Correlation between Child's Height and Parent's Height.
3. To investigate if parents' height has a significant impact on children's height.
4. To study whether the sex of the child influences their heights.
5. To study the time gap between two successive births in a family.
6. To check whether parity (No. of Children) affects height.

ANALYSIS

To assess the Normality of the height distributions of Fathers and Mothers.

In order to assess the normality of the height distributions of fathers and mothers, both Shapiro-Wilk Test and Histogram graphs are used in SPSS. The objective of this analysis is to determine whether the heights of fathers and mothers in the given dataset follow a normal distribution.

Test of Normality

To evaluate this assumption, we have used the Shapiro-Wilk Test which is specifically designed to test for normality and is a more powerful test for detecting deviations from normality, especially in smaller samples.

Let us assume the following hypothesis:

H_0: Height of parents are normally distributed.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Father height	.091	286	.000	.985	286	.004
Mother height	.068	286	.003	.981	286	.001

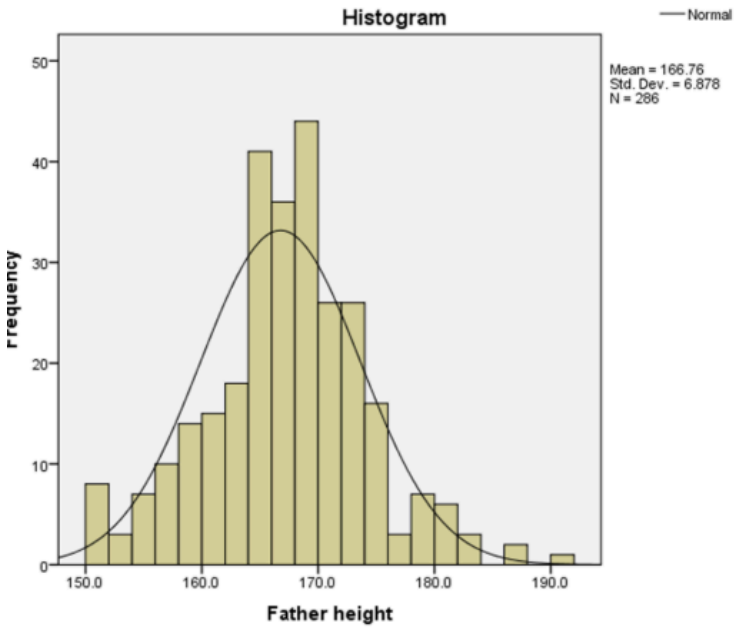
Observations:

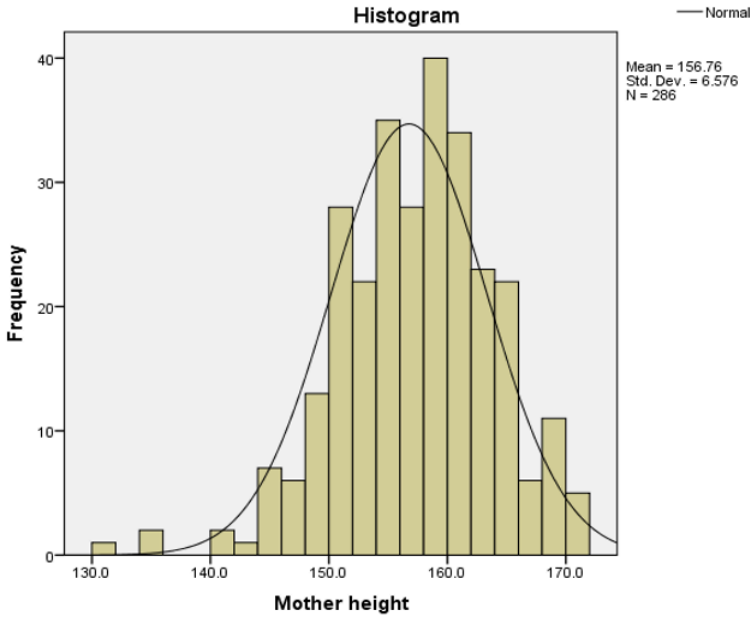
Following observations can be made from the Shapiro-Wilk Test:

- Father's Height: (p value = 0.004)
- Mother's Height: (p value = 0.001)

Since both p-values are **less than 0.05**, we reject the null hypothesis of normality. This suggests that **adult heights are not perfectly normally distributed**, although they are close.

HISTOGRAM ANALYSIS





Observations:

The plotted histograms show a roughly bell-shaped distribution for both **Father's Height** and **Mother's Height**, indicating a tendency toward normality. Overall the Histograms suggest approximate normality, the statistical tests confirm that the deviations are significant enough to reject the assumption of normality.

Results:

The Shapiro-Wilk Test showed p-values below 0.05 for both parents, indicating significant deviations from a perfect normal distribution. However, the histograms tell a slightly different story. They show bell-shaped curves, suggesting the heights tend to resemble normal distributions, even if not perfectly. Thus, the statistical results reject the assumption of strict normality but visually the data suggests a tendency toward normality since the heights are not perfectly normally distributed but they are close.

To Analyse the Relationship between Child's Height and Parent's Height:

This question aims to analyze the relationship between a child's height and the heights of their parents (father and mother) using Pearson Correlation in SPSS.
Let us assume the following hypothesis:

H_0: Heights of childrens are uncorrelated with their parents' height.

Correlation Analysis of First Child's Height

		Height Child1	Father height	Mother height
Pearson Correlation	Height Child1	1.000	.278	.108
	Father height	.278	1.000	.272
	Mother height	.108	.272	1.000
Sig. (1-tailed)	Height Child1	.	.000	.034
	Father height	.000	.	.000
	Mother height	.034	.000	.
N	Height Child1	286	286	286
	Father height	286	286	286
	Mother height	286	286	286

Observations:

- **First child and Father's height**:
  - Correlation coefficient : 0.278
  - Significance value (p- value): 0.000
  - A **weak** correlation, statistically significant. As father's height increases, the child's height tends to increase.

- **First child and Mother's height:**
  - Correlation coefficient: 0.108
  - Significance (p-value) : 0.034
  - A **negligible positive** correlation, but still statistically significant. The effect is smaller compared to father's height.

Correlation Analysis of Second Child's Height

		Height Child2	Father height	Mother height
Pearson Correlation	Height Child2	1.000	.231	.109
	Father height	.231	1.000	.219
	Mother height	.109	.219	1.000
Sig. (1-tailed)	Height Child2	.	.000	.042
	Father height	.000	.	.000
	Mother height	.042	.000	
N	Height Child2	249	249	249

	Father height	249	249	249
	Mother height	249	249	249

Observations:

- **Second child and Father's height**:
    - Correlation coefficient : 0.231
    - Significance value (p- value): 0.000
    - A **weak** but statistically significant correlation. Taller fathers tend to have taller children.

- **First child and Mother's height:**
    - Correlation coefficient: 0.109
    - Significance (p-value) : 0.042
    - A **negligible** but statistically significant correlation. Mother's height has a smaller influence on child's height compared to father's.

Correlation Analysis of Third Child's Height

		Height Child3	Father height	Mother height
Pearson Correlation	Height Child3	1.000	.059	.162
	Father height	.059	1.000	.255
	Mother height	.162	.255	1.000
Sig. (1-tailed)	Height Child3	.	.260	.039
	Father height	.260	.	.003
	Mother height	.039	.003	.
N	Height Child3	120	120	120
	Father height	120	120	120
	Mother height	120	120	120

Observations:

- **Third child and Father's height**:
    - Correlation coefficient : 0.059
    - Significance value (p- value): 0.260

- A **negligible** and not statistically significant correlation. Father's height does not show a meaningful relationship with the child's height in this sample.

- **Third child and Mother's height:**
  - Correlation coefficient: 0.162
  - Significance (p-value) : 0.039
  - A **weak positive** but statistically significant correlation. This indicates a small tendency of child height to increase as mother' height increases.

Correlation Analysis of Fourth Child's Height

		Height Child4	Father height	Mother height
Pearson Correlation	Height Child4	1.000	.013	.141
	Father height	.013	1.000	.187
	Mother height	.141	.187	1.000
Sig. (1-tailed)	Height Child4	.	.469	.202
	Father height	.469	.	.134
	Mother height	.202	.134	.
N	Height Child4	37	37	37
	Father height	37	37	37
	Mother height	37	37	37

Observations:

- **Fourth child and Father's height**:
    - Correlation coefficient : 0.013
    - Significance value (p- value): 0.469
    - **Negligible** and not statistically significant. Father's height does not meaningfully relate to the child's height in this sample.

- **Fourth child and Mother's height:**
    - Correlation coefficient: 0.141
    - Significance (p-value) : 0.202
    - A **very weak positive** correlation, not statistically significant. Slight tendency for taller mothers to have taller children, but not strong or reliable.

Results:

There is a consistent but weak to moderate correlation between child height and parental height, with father's height having a slightly greater influence than mother's across most of the children. Further, it can be observed that the height of the first and second child is influenced more by the father's height than the mother's height which is contrary to the observation that the third born and fourth born's height are more correlated to the mother's height than the father's. However, these relationships are not strong enough to be considered predictive on their own, indicating that child height is determined by multiple, complex factors beyond just parental height.

To investigate if a parent's height has a significant impact on children's height.

This question aims to analyse if there is a significant impact of parent's height (father and mother) on children's height using Regression Analysis in SPSS.

Case 1:

To analyse if a parent's height has a significant impact on the **First Child's** height.

Let us assume the following hypotheses:
H_1: There is a significant impact of father's height on First Child's height.
H_2: There is a significant impact of mother's height on First Child's height.

The hypothesis tests if a parent's height carries a significant impact on First child's height. The dependent variable - First Child's height was regressed on predicting variable - Father's height and Mother's height to test **H_1** and **H_2**.

Hypotheses	Regression weights	Beta Coefficients	R^2	F	t-value	p-value	Hypotheses supported?
H_1	Father's Height→ First child's height	0.269	0.079	12.086	4.535	0.000^b	Yes
H_2	Mother's Height→ First child's height	0.035			0.595	0.553	No

Observation:

Where father's height significantly predicted the First child's height, **F(2,283)=12.086, p<0.001**, which indicates that the fathers height can play a significant role in shaping the first child's height **(b = 0.269, p < .001)**. These results clearly direct the positive effect of the fathers height. Moreover, the **R^2 = 0.079** depicts that the model explains **7.9%** of the variance in the first child's height.

Where mothers height shows **(b = 0.169, p > .001)**. These results clearly direct the null effect of the mothers height in variance with the first child's height. Moreover, the mother's height does not signify the first child's height.

Case 2:

To analyse if a parent's height has a significant impact on the **Second Child's** height.

Let us assume the following hypotheses:
H_1: There is a significant impact of father's height on Second Child's height.
H_2: There is a significant impact of mother's height on Second Child's height.

The hypothesis tests if a parent's height carries a significant impact on the Second child's height. The dependent variable- Second child's height was regressed on predicting variable - parents height to test the hypothesis **H_1** and **H_2.**

Hypotheses	Regression weights	Beta Coefficients	R^2	F	t-value	p-value	Hypotheses supported?
H_1	Father'sHeight→Second child's height	0.217	0.057	7.412	3.420	0.001	Yes
H_2	Mother's Height→ Second child's height	0.062			0.976	0.330	No

Observation:

Where father's height significantly predicted the Second child's height, **$F(2,246)=7.412$**, **$p \leq 0.001$**, which indicates that the fathers height can play a significant role in shaping the second child's height **($b = 0.217$, $p \leq .001$)**. These results clearly direct the positive effect of the fathers height. Moreover, the **$R^2 = 0.057$** depicts that the model explains **5.7%** of the variance in the second child's height.

Where mothers height shows **(b = 0.062, p > .001)**. These results clearly direct the null effect of the mothers height in variance with the second child's height. Moreover, the mother's height does not signify the second child's height.

Case 3:

To analyse if a parent's height has a significant impact on the **Third Child's** height.

Let us assume the following hypotheses:
H_1: There is a significant impact of father's height on Third Child's height.
H_2: There is a significant impact of mother's height on Third Child's height.

The hypothesis tests if a parent's height carries a significant impact on the Third child's height. The dependent variable- Third child's height was regressed on predicting variable - parents height to test the hypothesis **H_1** and **H_2**.

Hypotheses	Regression weights	Beta Coefficients	R^2	F	t-value	p-value	Hypotheses supported?
H_1	Father's Height→ Third child's height	0.019	0.027	1.598	0.206	0.837	No
H_2	Mother's Height→ Third child's height	0.157			1.665	0.099	No

Observation:

Where father's height significantly predicted the third child's height, **F(2,117) =1.598, p >0.001**, which indicates that the fathers height can play a significant role in shaping the third child's height **(b =0.019, p >0.001)**. These results clearly direct the null effect of the father's height in variance with the third child's height. Moreover, the father's height does not signify the third child's height.

Similarly, mothers height shows **(b =0.157, p > .001)**. These results clearly direct the null effect of the mothers height in variance with the third child's height. Moreover, the mother's height does not signify the third child's height.

Case 4:

To analyse if a parent's height has a significant impact on the **Fourth Child's** height.

Let us assume the following hypotheses:
H_1: There is a significant impact of father's height on Fourth Child's height.
H_2: There is a significant impact of mother's height on Fourth Child's height.

The hypothesis tests if a parent's height carries a significant impact on the Fourth child's height. The dependent variable- Fourth child's height was regressed on predicting variable - parents height to test the hypothesis **H_1** and **H_2**.

Hypotheses	Regression weights	Beta Coefficients	R^2	F	t-value	p-value	Hypotheses supported?
H_1	Father'sHeight→Fourth child's height	-0.014	0.020	0.349	-0.79	0.937	No

H_2	Mother's Height→ Fourth child's height	0.144			0.832	0.411	No

Observation:

Where father's height significantly predicted the Fourth child's height, **F(2,34)= 0.349**, **p > 0.001**, which indicates that the fathers height can play a significant role in shaping the Fourth child's height **(b = -0.014, p >0.001)**. Negative beta indicates a negative relationship between the father's height and the fourth child's height. These results clearly direct the null effect of the father's height in variance with the fourth child's height. Moreover, the father's height does not signify the fourth child's height.

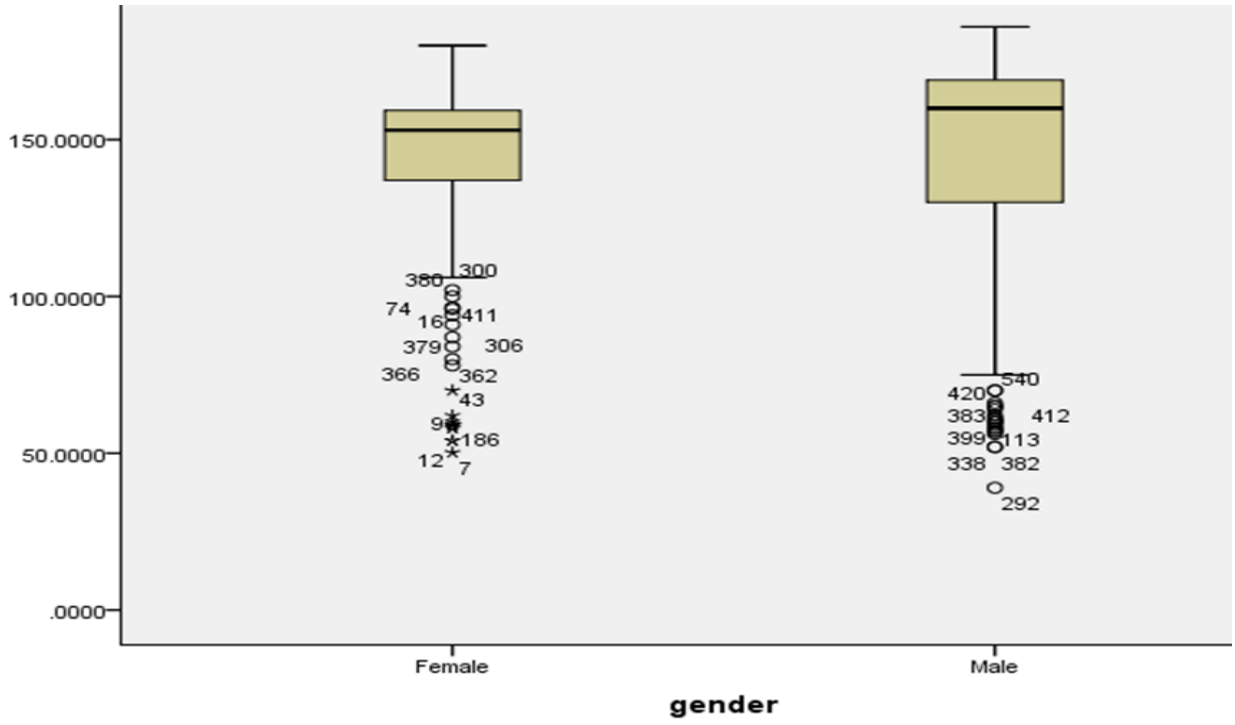
Similarly, mothers height shows **(b =0.144, p > .001)**. These results clearly direct the null effect of the mothers height in variance with the fourth child's height. Moreover, the mother's height does not signify the fourth child's height.

Results:

- Father's Height shows a significant positive effect on the heights of the first two children. Although the overall significance is weak. It does not significantly impact the third or fourth child's height.
- Mother's Height does not exhibit a significant impact on any of the children's heights.
- This could point to a lesser genetic contribution from mothers regarding height traits or a complex interplay of genes where the father's genetic influence is dominant for this trait. Another possibility is the environmental factors, which might overshadow maternal genetic contributions, play a larger role in shaping children's heights.

To study whether the sex of the child influences their heights.

This presents a comparative analysis of the height distribution among Male and Female children using **Boxplot visualization** in SPSS. Boxplots provide a concise summary of key statistical measures, including median, outliers, range and skewness. The objective is to interpret sex-based differences in height of children.



Observations:

- **<u>Central Tendency (Median Height):</u>**

The median height for male children is higher than that for females, indicating that, on an average, male children are taller than female children. This aligns with general biological growth patterns observed in children.

- **Spread (Interquartile Range-IQR):**

Both genders exhibit similar IQR's, suggesting that the variability in height is nearly the same for both the genders. However, the male boxplot is marginally taller, hinting at slightly greater variation among male children.

- **Outliers:**

Outliers are present in both gender groups. In females, a greater number and more extreme outliers, especially on the lower end, are observed. In males, fewer outliers are present, also on the lower end. This might indicate the presence of younger or unusually short female children, possible due to data entry entry errors or natural biological differences.

- **Whiskers (Range of Typical Values):**

In female children, height shows a range approximately from 70 cm to 180 cm. In male children, whiskers range from about 100 cm to 190 cm. This indicates that male children generally have a **higher range** of height values.

- **Skewness:**

Female child boxplot is negatively skewed since the median is closer to the top of the box while the male child boxplot is fairly symmetrical, possibly slightly positively skewed. Negative skewness in female boxplot means more low height values while symmetry in male heights suggests a more balanced distribution.

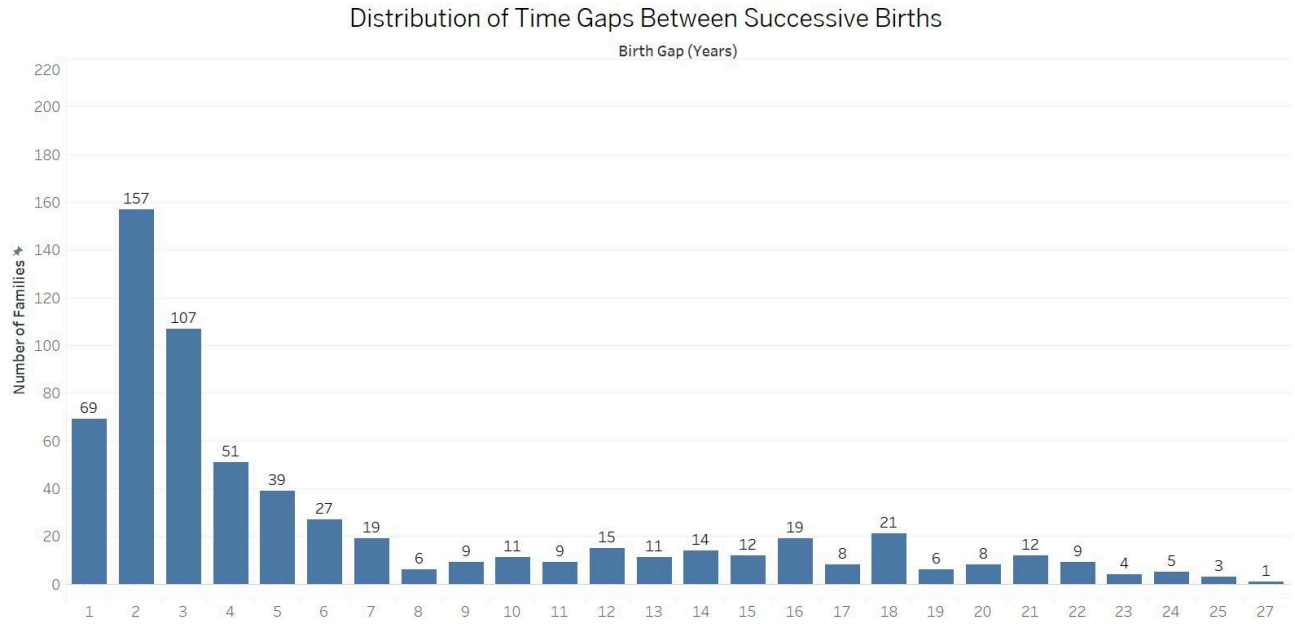
Results:

Overall, **Male children tend to be taller than Female children** on average, as indicated by a higher median and upper whisker range. Despite this, both genders display a similar spread in height, suggesting comparable growth variability. This might reflect **biological differences** in growth or could be affected by **age distribution.** Outliers are more common and extreme in the female group, particularly on the lower end. Negative skewness in female boxplot means more low height values while symmetry in male heights suggests a more balanced distribution.

To study the Time gap between Two Successive Births in a Family.

This analysis aims to determine the interval between the births of successive children within the same family. Rather than focusing solely on the number of children, the emphasis is on identifying patterns and trends in birth spacing to understand common family planning behaviours.

The raw dataset was first cleaned in Microsoft Excel to ensure accuracy and consistency. Subsequently, data visualization was performed using Tableau to identify key trends and the distribution of time gaps between births.



Observations:

- The most common time gaps between births fall within the **0** to **3-year** range.
- A **2-year** gap stands out as the most frequent, accounting for **157** cases.
- Gaps of **1 year** and **3 years** are appearing to have high frequency too, indicating a strong preference for less gaps between successive births.
- A marked decrease in frequency is observed for gaps exceeding **5 years**. While there are occasional fluctuations beyond this point,they are relatively minor.
- A small number of families exhibit birth intervals ranging from **8** to **27 years**.
- The data is right-skewed highlighting a general trend towards having children in relatively quick succession.

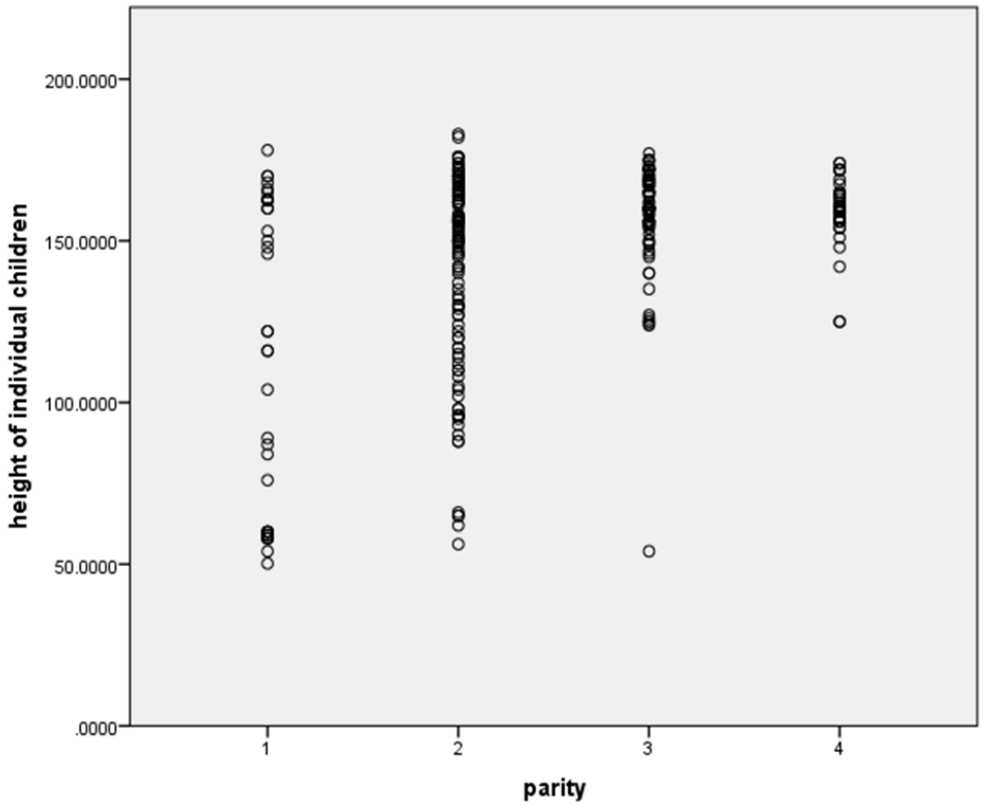
Results:

The analysis reveals a clear trend toward closely spaced births, with a significant concentration in the 1–3 year range—particularly a peak at 2 years. Extended gaps between births are uncommon and usually associated with specific life events. This pattern suggests that, in most families, decisions around childbirth are geared toward maintaining smaller age differences between siblings.

To check whether Parity (Number of Children) affects height.

This analysis aims to explore whether the number of children in a family (parity) has any association with the height of individual children.

A scatterplot analysis in SPSS was used to examine the relationship between parity (ranging from 1 to 4) and the height of individual children. The dataset includes children grouped by family size, with parity ranging from 1 to 4. Each data point represents the height of a single child from a family of a given parity.



- **X-axis** represents **parity**, a categorical variable indicating the no. of children a mother has,ranging from 1 to 4.
- **Y-axis** represents a **child's height**, a continuous variable (in cm.) ranging from approx. 0 to 200 cm.

Observations:

- **Distribution:**
    - Height values show a wide range across all parity levels.
    - Clusters of data appear between 100 cm and 180 cm for each parity.

- **Outliers:**
    - Significant low values, especially at parity levels 1 and 3 , may indicate potential outliers due to age differences or data entry issues.

- **Data Density:**
    - Parity level 2 has the most data points, suggesting it is the most common group.
    - Parity 4 has the fewest observations, possibly reducing reliability for that category.

- **Trend Analysis:**
    - There is no obvious trend or directional relationship between parity and height.
    - The data do not show consistent increases or decreases in height with higher parity levels.

Results:

From the scatterplot alone, we can draw the following preliminary conclusion:

There is no clear or consistent relationship between parity (no. of children) and the height of individual children. The heights are spread out similarly across all parity groups, and there is no obvious increasing or decreasing trend as parity increases from 1 to 4.

CONCLUSION

---

1. **To assess the Normality of the height distributions of Fathers and Mothers.**

- The Father's Height (FHT) and Mother's Height (MHT) were analyzed using **Histograms** and the **Shapiro-Wilk normality test.**
- The distribution appears roughly normal, but the p-values were below 0.05, indicating that the heights deviate slightly from a perfect normal distribution.

2. **To analyse the Relationship/Correlation between Child's Height and Parent's Height.**

- Child height shows a **weak to moderate correlation** with parental height, with father's height having slightly more influence overall.
- The height of the First and Second child is more Correlated to the father's height, whereas the third and fourth child's height is more linked to the mother's. However, these **correlations are not strong enough for prediction**, suggesting that multiple complex factors beyond parental height contribute to child height.

3. **To investigate if a parent's height has a significant impact on children's height.**

- A **Linear Regression model** was used to predict the Child's Height based on Parent's Height. Father's height has a **weak but significant positive effect** on the heights of the first two children, with no significant impact on the third or fourth.
- Mother's height shows **no significant** influence on any child. This suggests dominant paternal genetic traits or the role of environmental factors overshadowing maternal contributions.

4. **To study whether the sex of the child influences their heights.**

- Male children are generally **taller** than female children, with a higher median and upper whisker range. Both genders show **similar variability** in height, indicating comparable growth patterns.
- Female heights exhibit more extreme outliers, particularly on the lower end, and a negative skew suggests more low values. In contrast, male heights display symmetry, indicating a more balanced distribution. **Biological differences** or **age distribution** may contribute to these patterns.

5. To study the Time gap between Two Successive Births in a Family.

- The analysis reveals a clear trend toward closely spaced births, with a significant concentration in the **1–3 year range**—particularly a peak at **2 years**. Extended gaps between births are uncommon and usually associated with specific life events. This pattern suggests that, in most families, decisions around childbirth are geared toward **maintaining smaller age differences** between siblings.

6. To check whether Parity (Number of Children) affects height.

- The scatterplot suggests **no consistent relationship** between birth order (parity) and child height. Heights are similarly distributed across all parity groups, with no noticeable upward or downward trend as parity increases from 1 to 4.
