Modeling Lung Capacity in Pediatric Population: A Regression-Based Approach to Assessing Respiratory Development

Sana Siddiqui, MD

Introduction

Pediatric lung capacity is a vital clinical marker used to assess respiratory health and growth in children, especially those at risk for chronic pulmonary diseases such as asthma. While spirometry is the gold standard for measuring lung function, it is often impractical for very young children who cannot reliably perform the test. As a result, there is a clinical need for alternative methods to estimate lung capacity using routinely collected health data.

In this case study, we developed a linear regression model to predict pediatric lung capacity based on common clinical indicators: age, height, gender, smoking status, and birth method. Our goal was to create a practical tool to help clinicians proactively identify children at risk for respiratory conditions when spirometry is not feasible.

Data Cleaning

The data was cleaned in a separate R script: Scripts/Data_Cleaning.R

Cleaning steps included:

- Standardizing column names
- Assessing for missing data
- Removing duplicate rows

Data Overview/Library

The data includes the following variables:

- lung capacity: measured in liters
- age: years
- height: inches
- gender
- smoking status: yes or no depending on if the child smokes
- caesarean: yes or no depending on birth method of the child

Data Overview

lung_cap	age	height	smoke	gender	caesarean
6.475	6	62.1	no	male	no
10.125	18	74.7	yes	female	no
9.550	16	69.7	no	female	yes
11.125	14	71.0	no	male	no
4.800	5	56.9	no	male	no
6.225	11	58.7	no	female	no

Exploratory Data Analysis

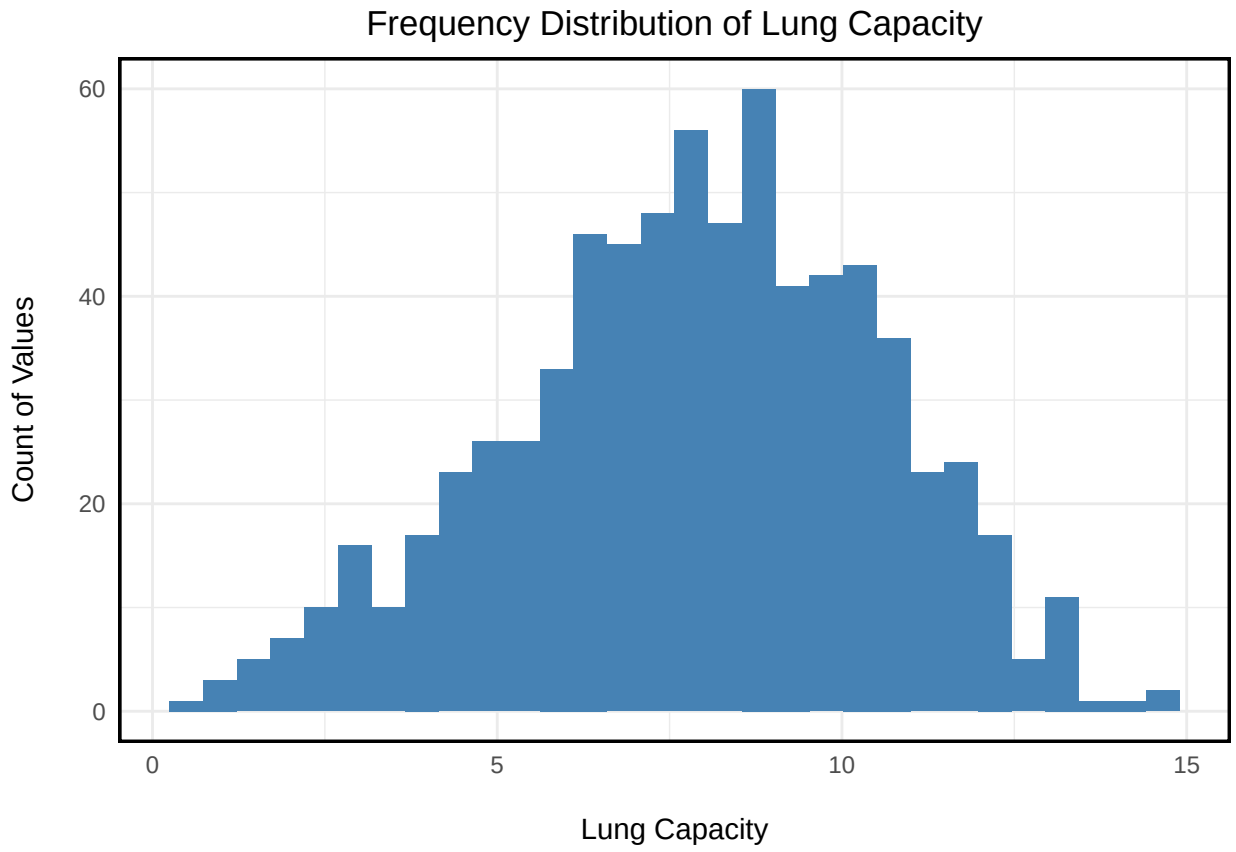
Summary Statistics

```
## Rows: 725
## Columns: 6
## $ lung_cap  <dbl> 6.475, 10.125, 9.550, 11.125, 4.800, 6.225, 4.950, 7.325, 8.~
## $ age       <dbl> 6, 18, 16, 14, 5, 11, 8, 11, 15, 11, 19, 17, 12, 10, 10, 13,~
## $ height    <dbl> 62.1, 74.7, 69.7, 71.0, 56.9, 58.7, 63.3, 70.4, 70.5, 59.2, ~
## $ smoke     <chr> "no", "yes", "no", "no", "no", "no", "no", "no", "no", "no",~
## $ gender    <chr> "male", "female", "female", "male", "male", "female", "male"~
## $ caesarean <chr> "no", "no", "yes", "no", "no", "no", "yes", "no", "no", "no"~
```

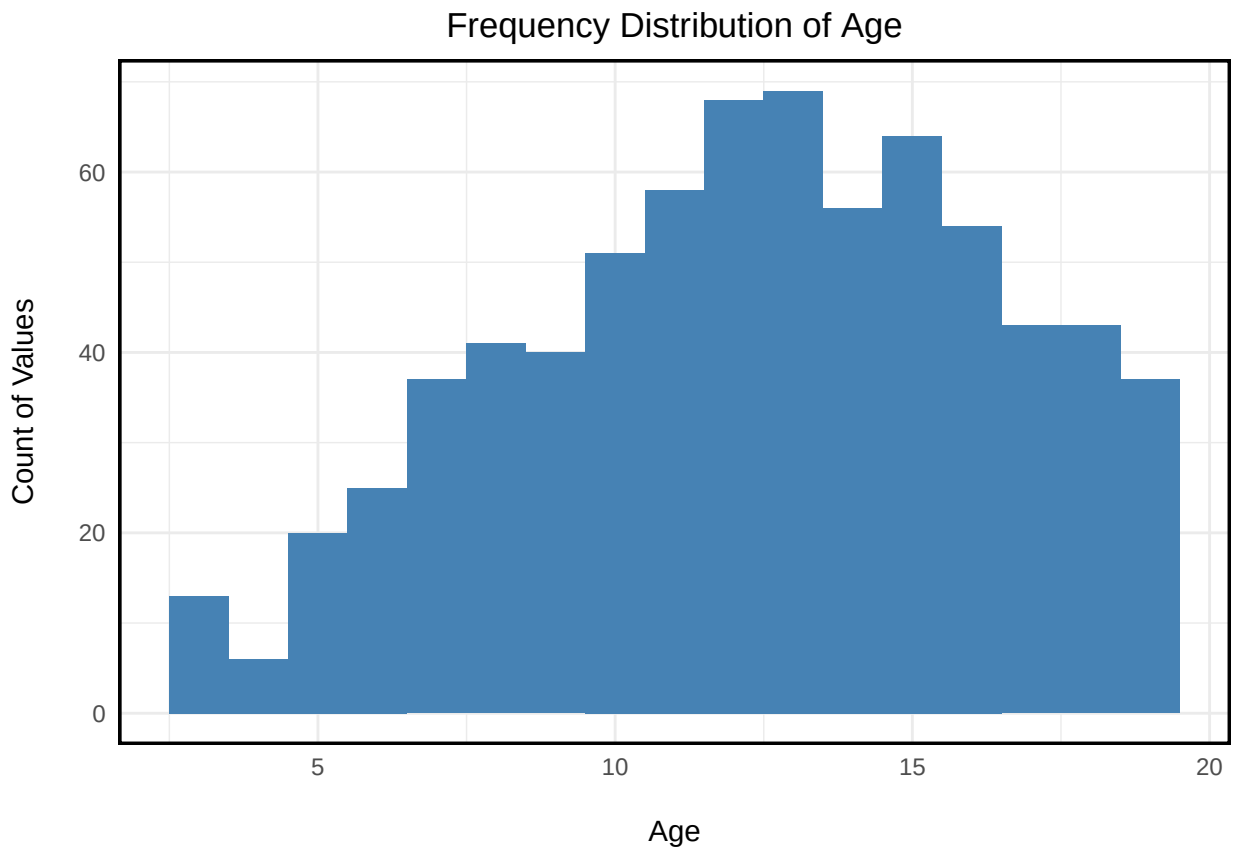
```
##     lung_cap           age             height         smoke          
##  Min.   : 0.507   Min.   : 3.00   Min.   :45.30   Length:725        
##  1st Qu.: 6.150   1st Qu.: 9.00   1st Qu.:59.90   Class :character  
##  Median : 8.000   Median :13.00   Median :65.40   Mode  :character  
##  Mean   : 7.863   Mean   :12.33   Mean   :64.84                     
##  3rd Qu.: 9.800   3rd Qu.:15.00   3rd Qu.:70.30                     
##  Max.   :14.675   Max.   :19.00   Max.   :81.80                     
##     gender           caesarean        
##  Length:725         Length:725        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##
```

Distributions of Continuous Variables

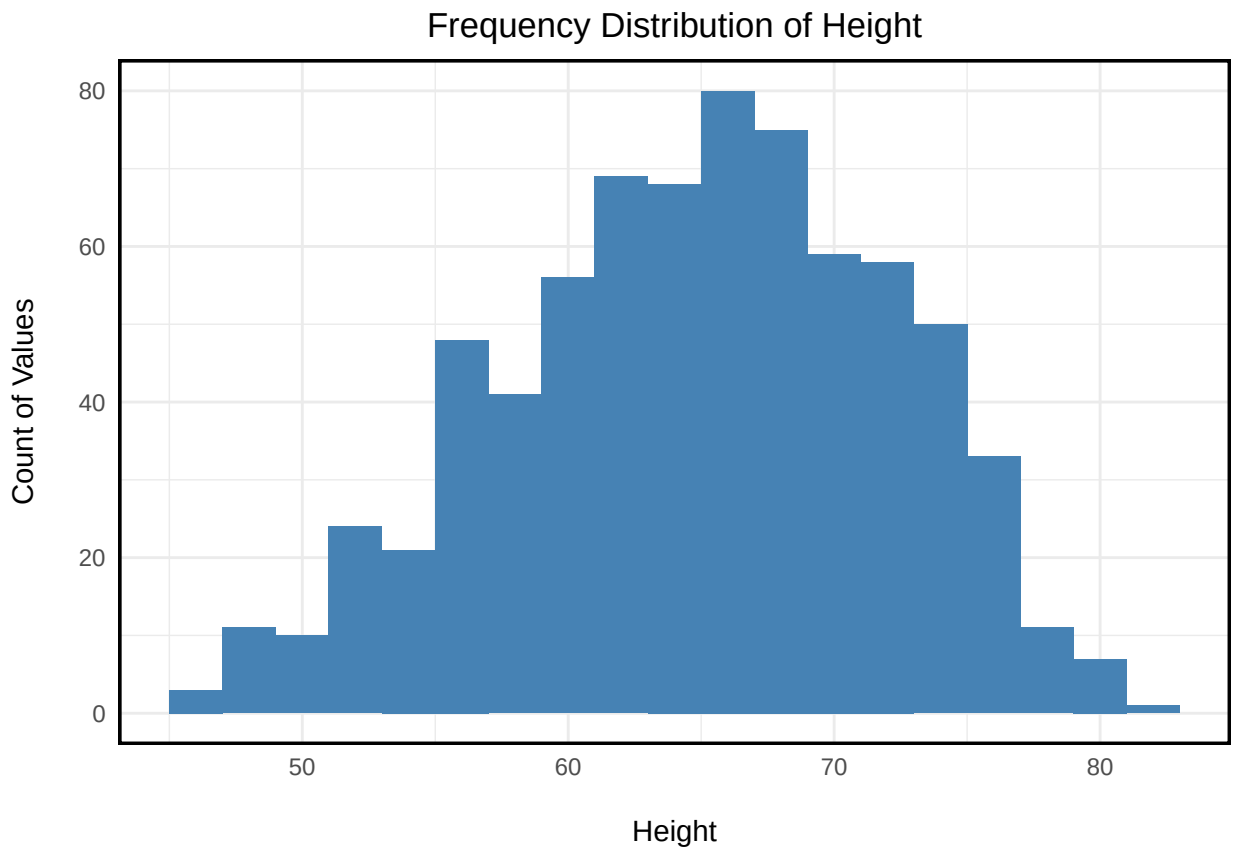
The histogram below displays the distribution of lung capacity values in the dataset. The distribution appears approximately normal, with most values clustered around the mean, indicating a central tendency typical of a Gaussian distribution



Similarly, the distribution of age values appears to be approximately normal with most values clustered around the mean.



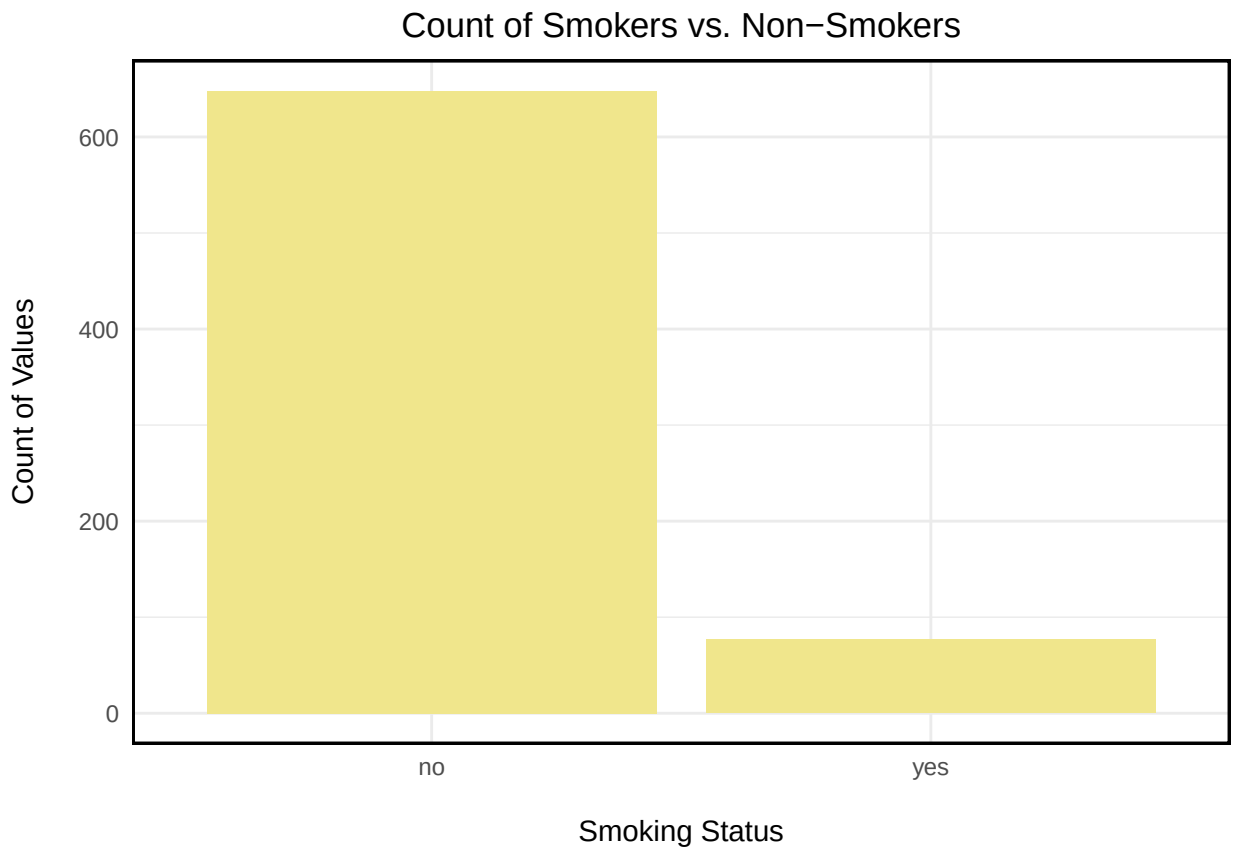
Finally, the histogram above reflects the distribution of height values and also appears to be approximately normal.



Distributions of Categorical Variables

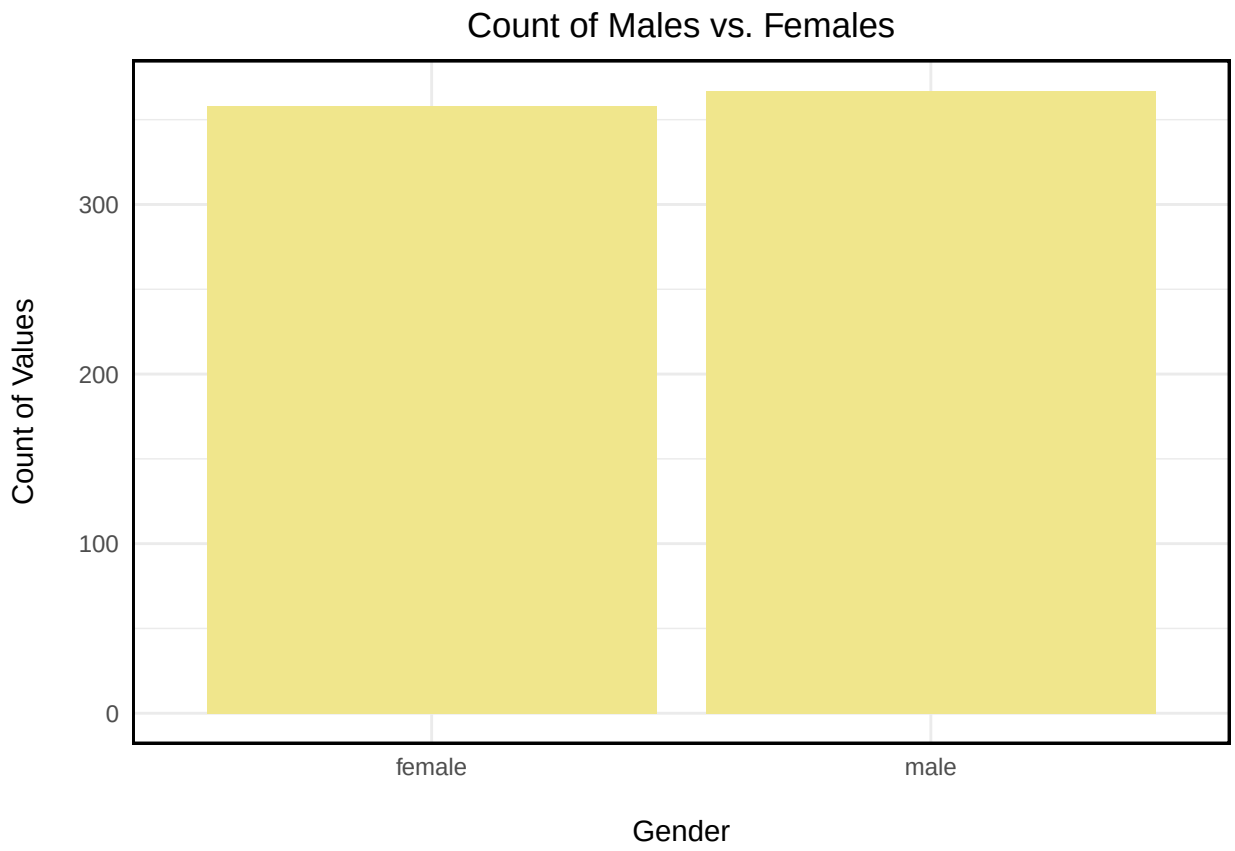
Smoking Status

The bar graph below shows that majority patients (over 95%) in this dataset were nonsmokers.



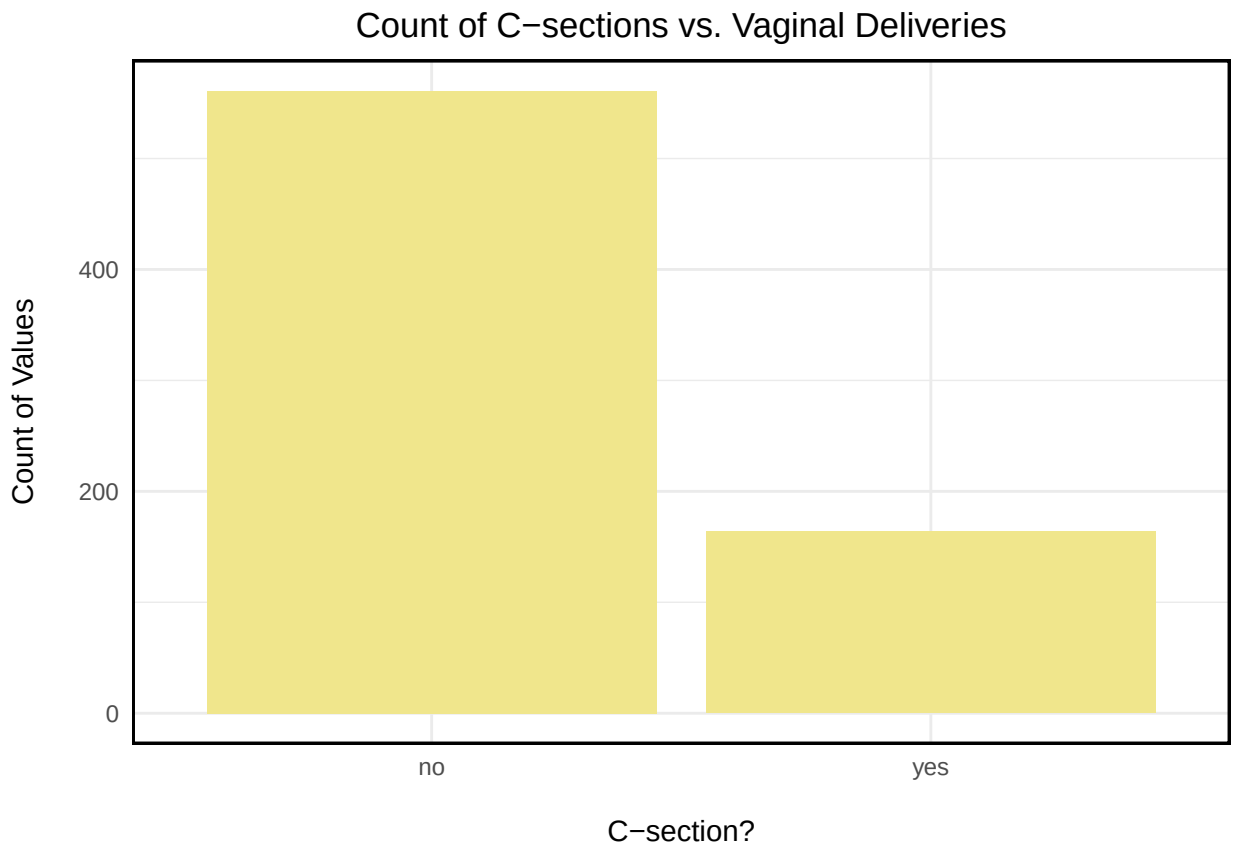
Gender

The bar graph below shows that we have an almost equal split between male and female patients in the dataset out of a total of 725 patients.



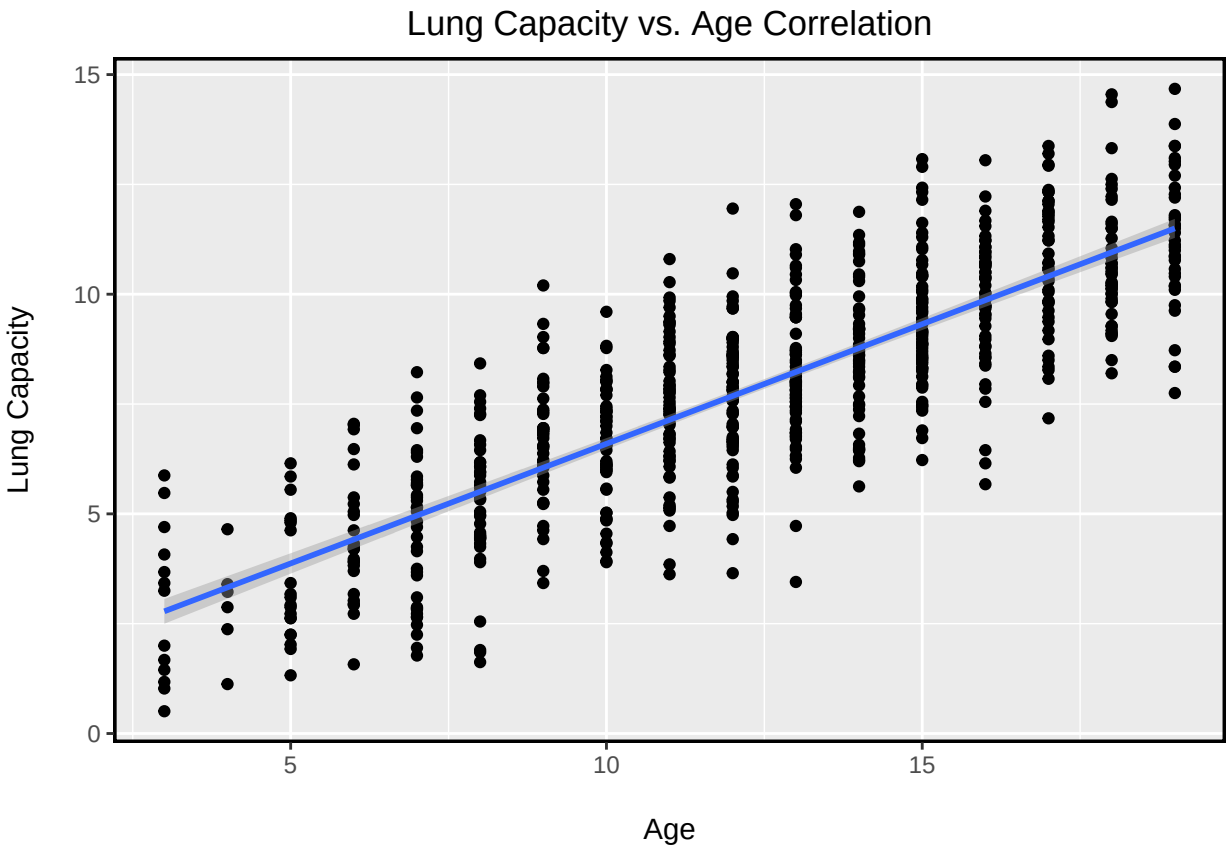
Caesarean Status

The bar graph below shows that majority of the patients in this dataset were delivered vaginally ($> 80\%$ of patients) while a small subset were delivered via C-section.



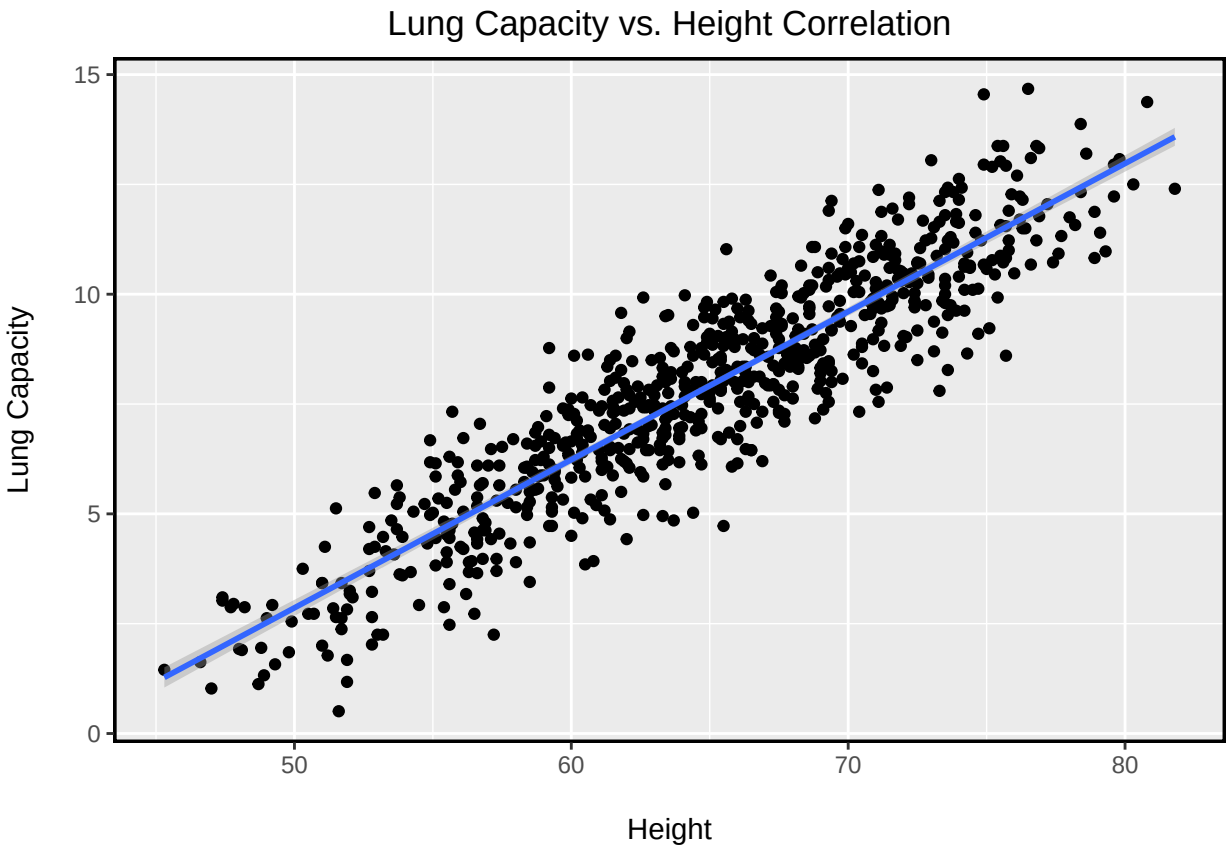
Relationship between Lung Capacity and other variables

Lung Capacity and Age



There is a positive correlation between lung capacity and age. Lung capacity tends to increase with age.

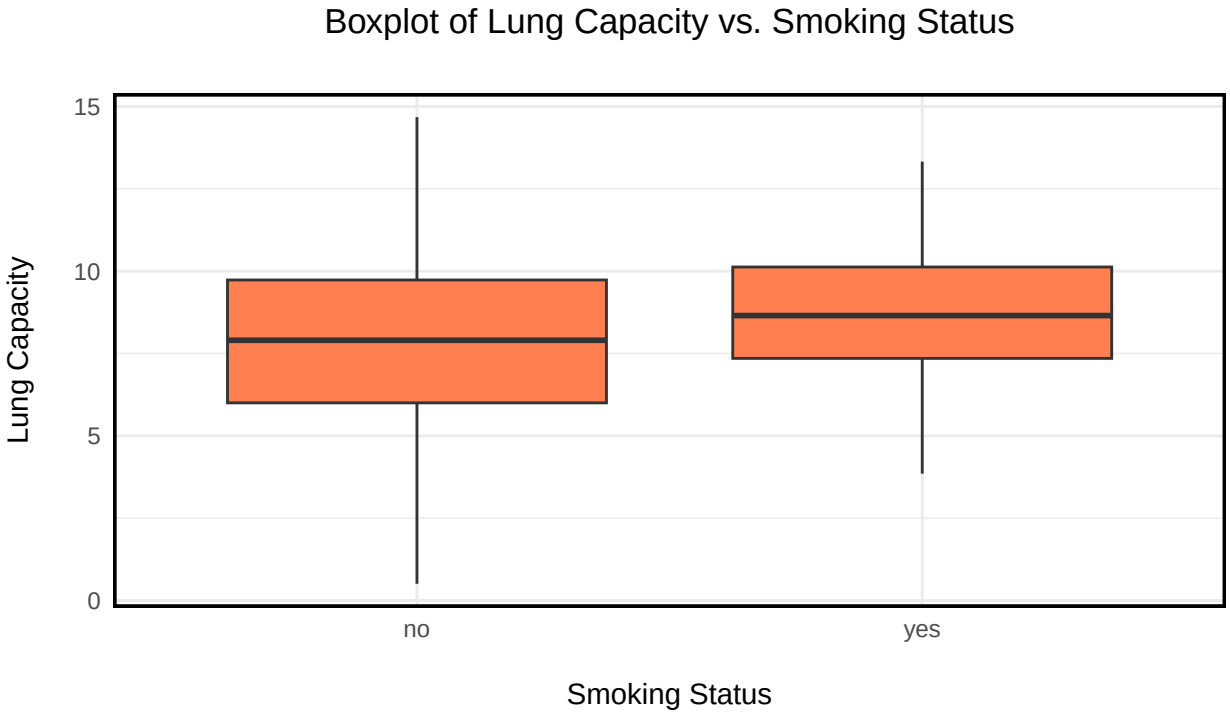
Lung Capacity and Height



There is a positive correlation between lung capacity and age. Lung capacity tends to increase with age.

Lung Capacity By Smoking Status

The boxplot below illustrates the distribution of lung capacity values among smokers and non-smokers. Given that the dataset contains a larger proportion of non-smokers, the wider spread in lung capacity values for this group is expected due to the greater sample size



The following table is used to understand the smoking vs. non smoking groups better.

- The group of non smokers has a similar number of male and females
- The smoking group also has a similar number of males and females

smoke	gender	min_age	max_age	avg_age	n
no	female	3	19	12	314
no	male	3	19	12	334
yes	female	10	19	15	44
yes	male	10	19	15	33

The next table explores the smoking group even further to understand the differences in lung capacities between the two groups. The boxplot above showing the distribution of values between smokers and non-smokers shows a slightly greater median in the smoking group. This is counter-intuitive; we would expect a lower lung capacity in the smoking group.

smoke	min_lung_cap	max_lung_cap	avg_lung_cap	n
no	0.507	14.675	8	648
yes	3.850	13.325	9	77

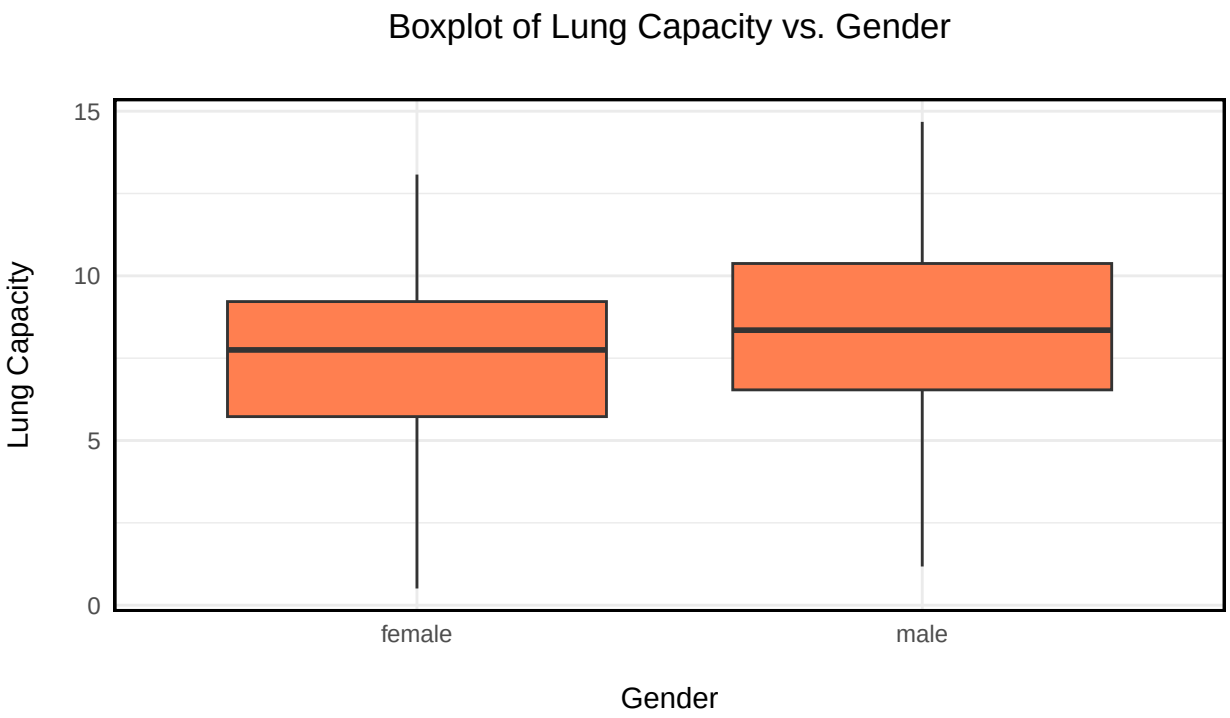
We observe that the number of smokers in the dataset is significantly smaller than the number of non-smokers. Additionally, the average age of smokers is higher than that of non-smokers. Since lung capacity

generally increases with age, we would expect older children—regardless of smoking status—to exhibit higher lung capacity.

It's important to note that teenagers are more likely to engage in smoking than younger children. However, the negative effects of smoking on lung capacity typically develop over time and may not be immediately apparent in the early years of smoking. Therefore, the boxplot results align with expectations: as children grow older, they are both physically developing and increasingly exposed to risk behaviors such as smoking. Yet, the short-term impact of smoking may be masked by ongoing growth, making the harmful effects less visible at this stage.

Lung Capacity By Gender

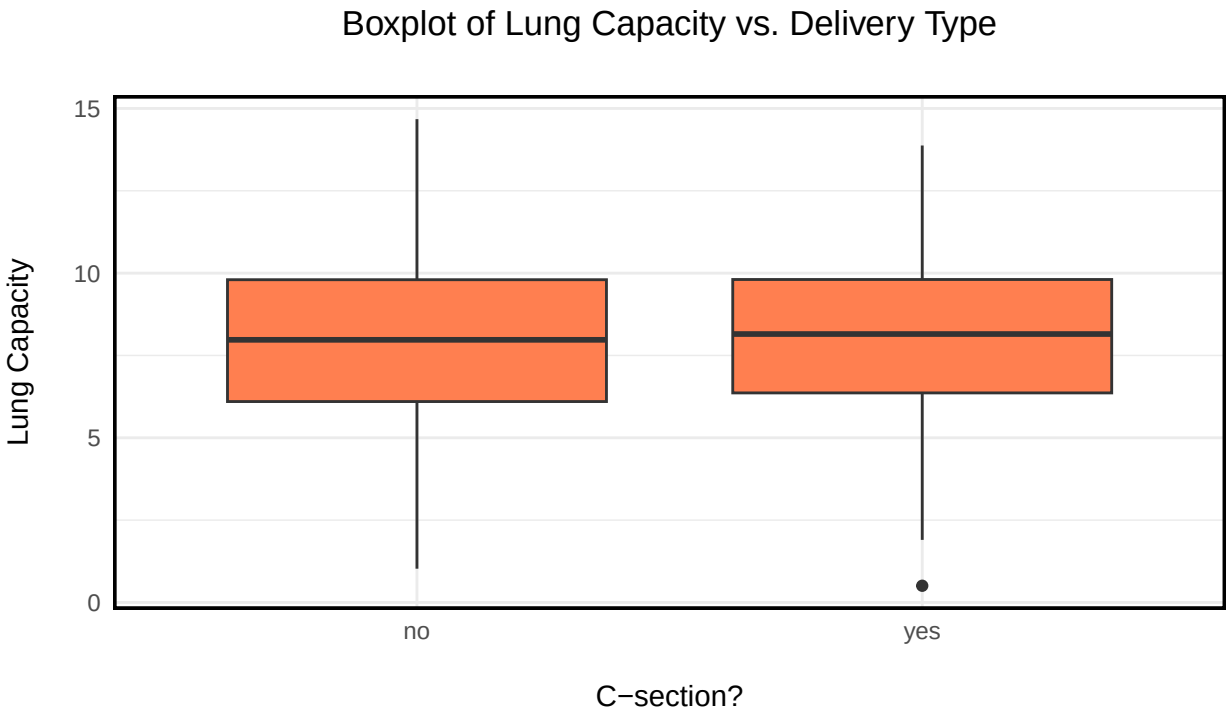
The boxplot below illustrates the distribution of lung capacity among males and females. While the overall variability in lung capacity appears comparable between the two groups, the female group shows a slight left skew. This suggests that lung capacity values are more tightly clustered between the first and third quartiles, with a few lower values pulling the median closer to the upper quartile



gender	min_age	max_age	avg_age	n
female	3	19	12	358
male	3	19	12	367

Lung Capacity By Delivery Type

The boxplot below displays distribution of lung capacity in children born born via C-section and via vaginal delivery. While majority children in this dataset were born vaginally, there is comparable variability in lung capacity amongst both groups of children. The medians are almost identical.



To predict lung capacity in the pediatric population, a linear regression model was selected as the most appropriate modeling approach based on the characteristics of the dataset.

The model satisfies key assumptions of linear regression:

- Linearity: Lung capacity shows a linear relationship with both age and height.
- No multicollinearity: All predictor variables have Variance Inflation Factors (VIF) well below the threshold of 5, indicating minimal correlation between them.
- Homoskedasticity: Residuals exhibit constant variance across the range of fitted values.

Linear Regression

Multicollinearity / VIF

```
##       age    height     smoke    gender caesarean
##  3.620270  3.655950  1.050188  1.105652  1.004598
```

All VIFs are < 5, meaning no significant multicollinearity amongst predictors.

Linear Regression

```
##
## Call:
```

```
## lm(formula = lung_cap ~ age + height + smoke + gender + caesarean,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3388 -0.7200  0.0444  0.7093  3.0172
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.32249    0.47097 -24.041  < 2e-16 ***
## age            0.16053    0.01801   8.915  < 2e-16 ***
## height         0.26411    0.01006  26.248  < 2e-16 ***
## smokeyes      -0.60956    0.12598  -4.839 1.60e-06 ***
## gendermale     0.38701    0.07966   4.858 1.45e-06 ***
## caesareanyes  -0.21422    0.09074  -2.361   0.0185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 719 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8532
## F-statistic: 842.8 on 5 and 719 DF,  p-value: < 2.2e-16
```

Results Interpretation

- All predictors have p-values < 0.05 and are significant at the 5% level of significance, we can safely assume all predictors have
- Adjusted R-squared value is posiitve and very close to 1.
- Overall p-value of the model is highly significant at the 5% significance level

Standardizing Model Fit

```
##
## Call:
## lm(formula = lung_cap ~ age + height + smoke + gender + caesarean,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3388 -0.7200  0.0444  0.7093  3.0172
##
## Coefficients:
##               Estimate Standardized Std. Error t value Pr(>|t|)
## (Intercept)  -11.32249           NA    0.47097 -24.041  < 2e-16 ***
## age            0.16053      0.24150    0.01801   8.915  < 2e-16 ***
## height         0.26411      0.71457    0.01006  26.248  < 2e-16 ***
## smokeyes      -0.60956     -0.07060    0.12598  -4.839 1.60e-06 ***
## gendermale     0.38701      0.07274    0.07966   4.858 1.45e-06 ***
## caesareanyes  -0.21422     -0.03369    0.09074  -2.361   0.0185 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 719 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8532
## F-statistic: 842.8 on 5 and 719 DF,  p-value: < 2.2e-16
```

What is standardization and why did we standardize the model's coefficients?

Standardization refers to the process of transforming variables to a common scale — typically by converting them to z-scores (i.e., mean = 0, standard deviation = 1).

In the context of linear regression, this allows us to normalize the scale of the predictors (independent variables), so that the resulting coefficients are directly comparable.

This is useful because even though multiple predictors may be statistically significant, we cannot directly compare the magnitude of their unstandardized coefficients — since they may be measured in different units (e.g., age in years vs. height in cm).

By using standardized coefficients (e.g., via lm_beta()), we can determine which predictors have a greater relative effect on the outcome variable, because they are all now on the same unitless scale.

Results of Linear Regression

- All predictors are significant at the 5% significance level
- Height and Age are the strongest predictors of lung capacity, with standardized coefficients of approximately 0.74 and 0.21, respectively.
- This means that for every 1 standard deviation increase in height or age, the lung capacity increases by 0.74 SDs and 0.21 SDs respectively.
- The adjusted R-squared is 0.85 with a significant p-value < 0.001 indicating that 85% of the variation in lung capacity can be explained by the predictors used in this model i.e. age, height, smoking status, gender, and caeserean status.

Predictive Modeling

```
##
## Call:
## lm(formula = lung_cap ~ age + height + smoke + gender + caesarean,
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5157 -0.6889  0.0383  0.6848  2.8708
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.54080    0.50774 -22.730  < 2e-16 ***
## age            0.15012    0.01920   7.818 2.56e-14 ***
## height         0.26902    0.01081  24.893  < 2e-16 ***
## smokeyes      -0.53205    0.13421  -3.964 8.28e-05 ***
## gendermale     0.36828    0.08604   4.280 2.19e-05 ***
## caesareanyes  -0.19731    0.09745  -2.025   0.0434 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9868 on 576 degrees of freedom
## Multiple R-squared:  0.8629, Adjusted R-squared:  0.8617
## F-statistic: 725.1 on 5 and 576 DF,  p-value: < 2.2e-16
```

```
## [1] 1.146868
```

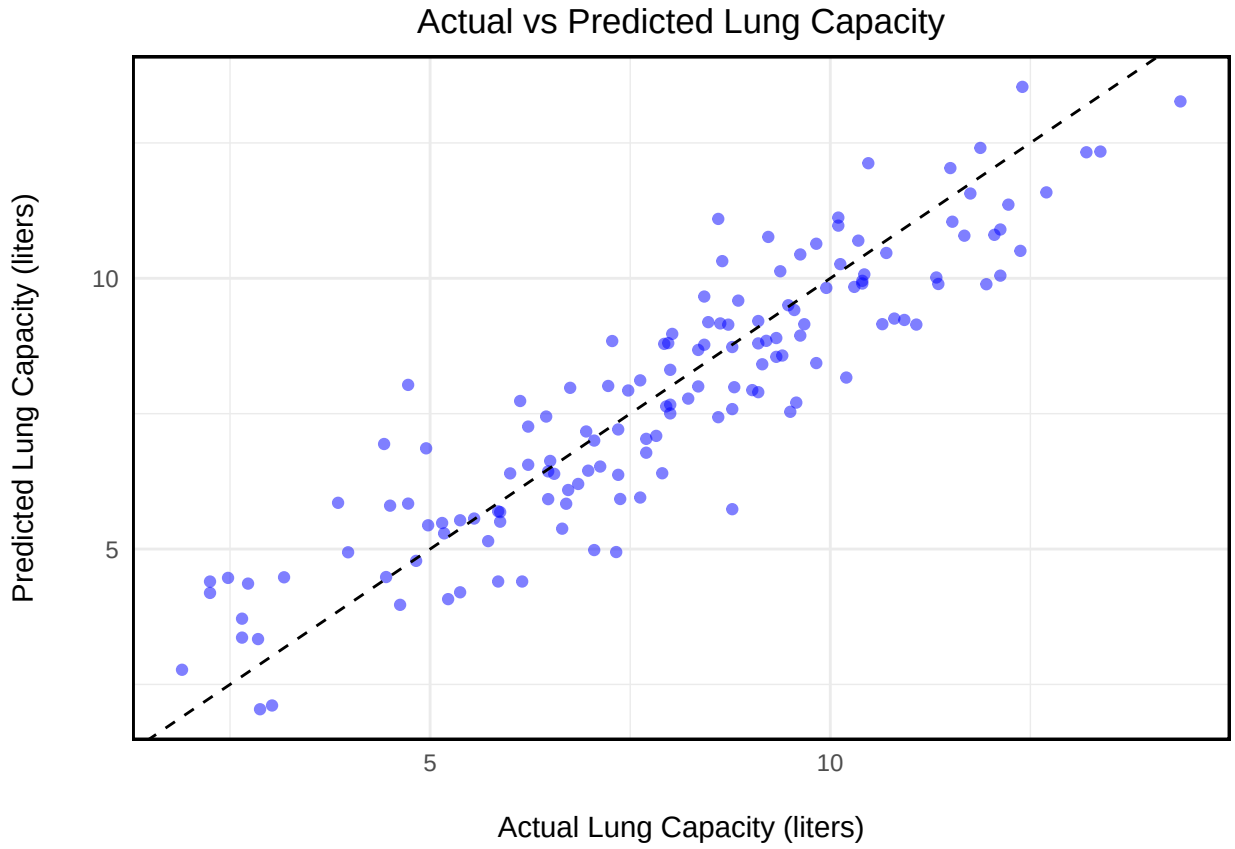
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.507   6.150   8.000   7.863   9.800  14.675
```

```
## [1] 2.662008
```

Prediction Analysis

- All predictors in the model are statistically significant with p-values < 0.05
- The adjusted R-squared is 0.86 and is statistically significant; all predictors captured in this model explain 86% of the variability seen in lung capacity
- The root mean squared error is 1.14, meaning on average, the model's lung capacity prediction is off by 1.14 liters. In order to evaluate this, we compare RMSE to the standard deviation (SD) of lung capacity of our entire dataset. As seen above in the summary statistics, the SD of lung capacity is 2.66.

Overall, this indicates that the model provides considerably more accurate predictions than a naive model based on the mean, and it explains a meaningful amount of the variance in lung capacity.



The scatterplot above indicates a well-performing predictive model. There is a strong positive relationship between actual and predicted lung capacity values, with most points closely aligning along the best-fit trend line and minimal deviation.

Conclusion

Pediatric lung capacity is a powerful tool that helps clinicians assess overall pulmonary function as the child grows, especially in children with chronic pulmonary diseases. It can be used as a preventative tool in populations of children that are at highest risk for chronic conditions, most commonly asthma. While the gold standard technique of accurately capturing lung capcity is pulmonary function testing via spirometry, very young children are unable to complete this test purely based on their age. Therefore, if we are able to use other characteristics about the child's health to predict lung capacity and assess it against the average capacity of children within the respective age group, we can proactively identify children who are at risk of developing pulmonary conditions.

In this case study, we aimed to utilize the most commonly tracked elements from clinical pediatric visits including age, height, gender, smoking status of the child, and birth method to predict lung capacity.

We used linear regression to predict lung capacity based on the predictors desacribed above. Our model and predictors were all statistically significant with an adjusted R-squared of 0.85; 85% of the variability in the lung capacity within our dataset can be explained by the predictors utilized in the model to make predictions. In order to assess the accuracy of the model, we calculated the root mean squared error (RMSE) to compare our model to a brute force approach of using the mean and standard deviation of lung capacity within the original datsaet to calculate lung capacity of any given child. The RMSE was 1.14 liters and standard deviation of lung capacity in the original datset was 2.66 liters. We can strongly conclude that our model, despite being off by 1.14 liters on average from actual lung capacity predicts lung capacity better than utilizing a naive approach of using the standard deviation alone.