Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of death in the United States. People believe COPD is most associated with smoking cigarettes. However, there are other beliefs that there are other factors and variables that contribute to a person developing COPD.

The purpose of the COPDGene study is to analyze the relationship between the air's volume forcefully exhaled in 1 second, also known as the Forced Expiratory Volume (FEV1) and FEV1 measured 5 years later (FEV1_phase2), and other contributing variables. I am interested in identifying which genetic variables do and do not contribute to COPD.

To find out what causes COPD, the COPDGene research group collected data from participants in their research. The research group took many measurements on men and women regarding their bodies. The measurements included a person's height, weight, how old they were, etc. However, their most important measurement was their exhaled air volume FEV1 and FEV1_phase2. I found that the smaller a person's air volume is, the more likely they would be susceptible to diseases like COPD. Generally, men had higher capacities than women. This meant that women were more likely to get COPD as opposed to men. I also found that patients who were not diagnosed with emphysema, which is another pulmonary disease, were more likely to have a higher capacity in FEV1_phase2. I also tested FEV1 and FEV1_phase2 for linear regression. My linear regression equation for y came out to be y = 0.178336 + 0.840423x(FEV1). This can be interpreted as "for every 0.178336 volume of air in FEV1, the volumes of air of FEV1_phase2 increase by 0.840423". The R^2 of this regression model is 0.7897, which means that the variation in FEV1 can explain 78.97% of the variability in FEV1_phase2. I also noted that there is a strong negative linear relationship between FEV1 and FEV1_phase2 within the next 5 years. This indicates that FEV1 is a high indicator of poor health that other factors can influence. However, it is important to note that any factors relevant to patient physiology may not fully indicate COPD.

The FEV1 distribution is symmetrical, bell-shaped, and normal, indicating that it follows the empirical rule. My visit_age histogram ranges from 40 to 80 years old, with the mode located at 55 years old. The total_lung_capacity histogram data for this is not normally distributed due to the outliers. I also tested FEV1_phase2 with other variables, where I analyzed this data with boxplots using smoking_status and compared it with divided 3 different groups. Unsurprisingly, the group with the best breathing capacity is the group with people who never smoked.

Based on the data, I have concluded that COPD is closely related to women, those who are in the age range 40 to 80, those who have quit smoking, and others who have other pulmonary diseases. I selected these factors based on their low FEV1 capacity because a lower capacity has a stronger indicator of developing COPD or other pulmonary diseases.