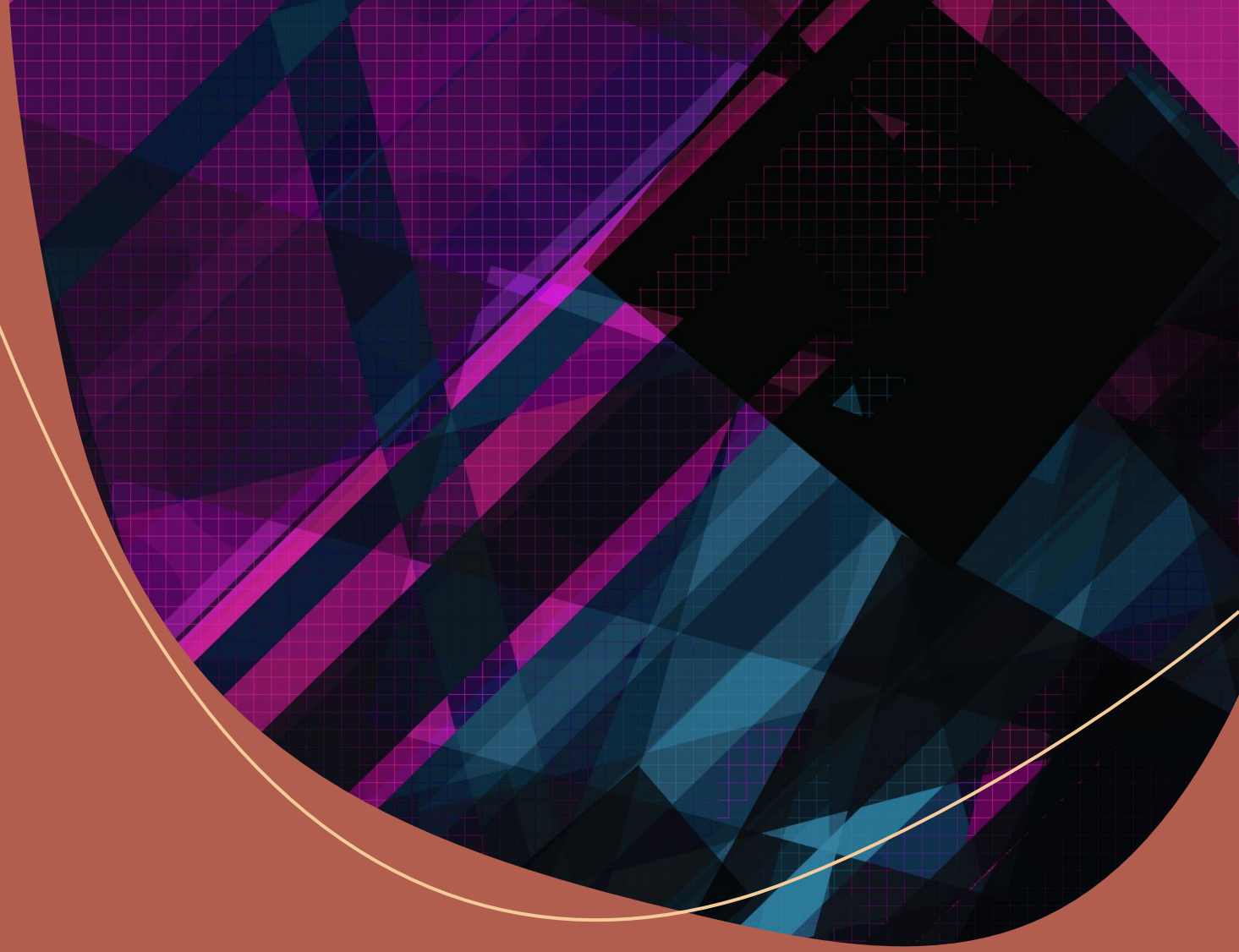
Lung Cancer prediction



DSC 530 Data exploration and Analysis

Final Term Project

Tapaswi chowdary Sriramineni

DATA SET

- The data set that we are working on throughout this project is Lung cancer prediction.
- This data is obtained from the survey that is created to find the data on the causes of the lung cancer.
- This dataset provides the data of the level of air pollution and it's effect on the lung cancer.

Variables

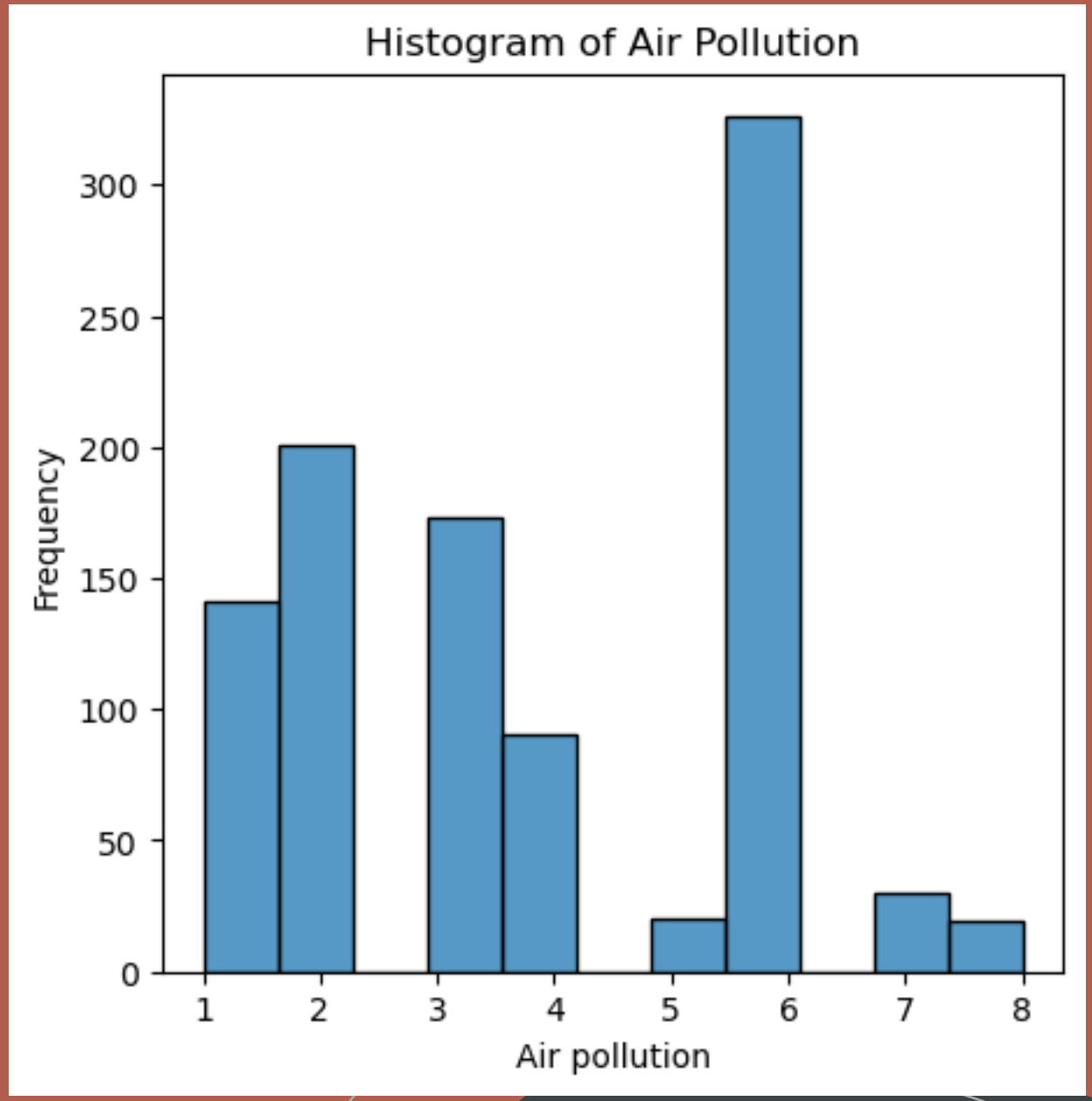
The five variables that we will be working on in the project are:

- Air pollution : The level of air pollution exposure of the patient. (Categorical)
- Alcohol use : The level of alcohol use of the patient. (Categorical)
- Dust Allergy : The level of dust allergy of the patient. (Categorical)
- Genetic Risk : The level of genetic risk of the patient. (Categorical)
- Chronic Lung Disease : The level of chronic lung disease of the patient. (Categorical)

HISTOGRAMS

Histogram of Air Pollution

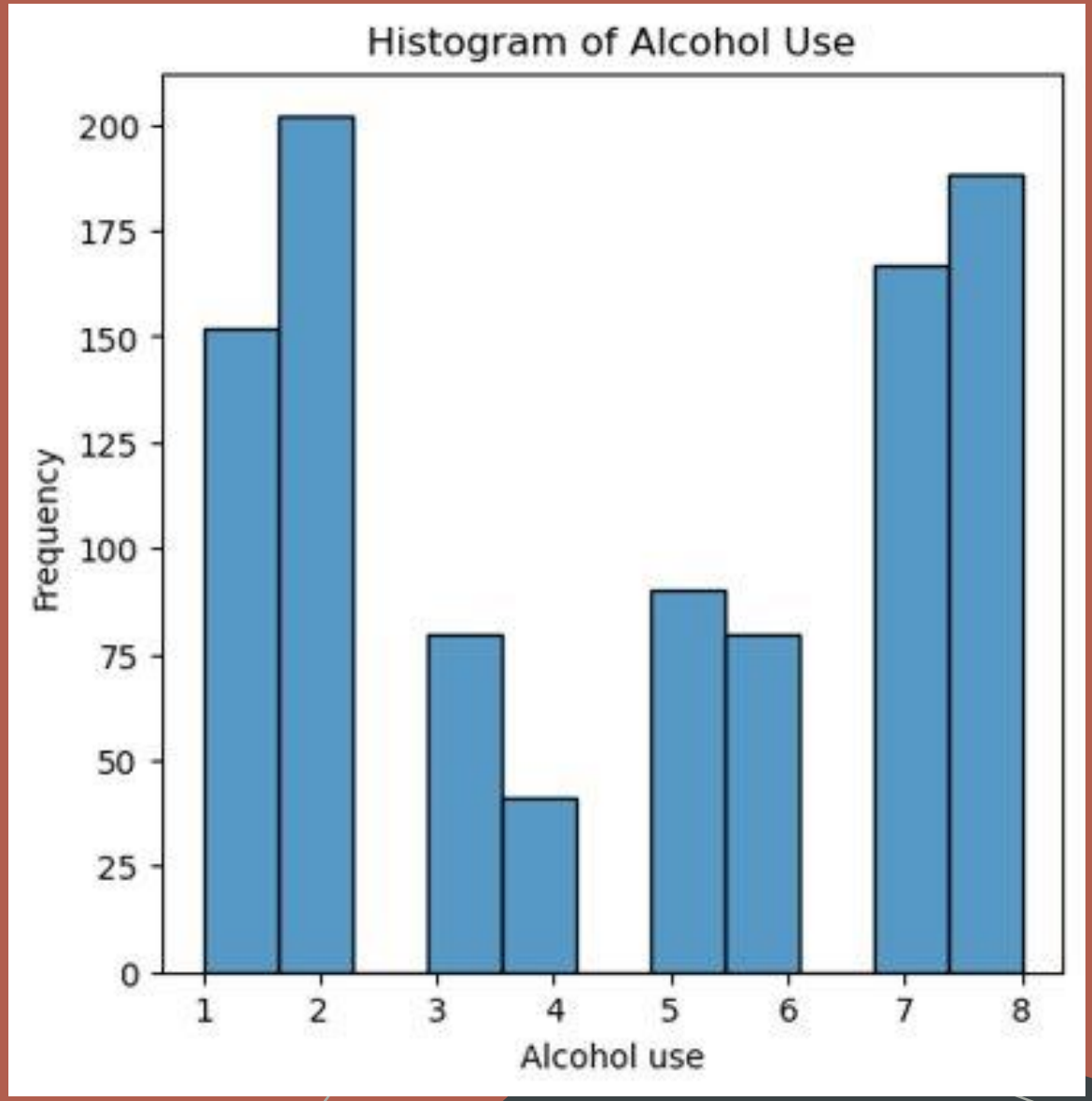
Mean: 3.84
Mode: [6]
Variance: 4.12252252
2522477
Spread(IQR): 4.0
Tail:
995 6
996 6
997 4
998 6
999 6



HISTOGRAMS

Histogram of Alcohol Use

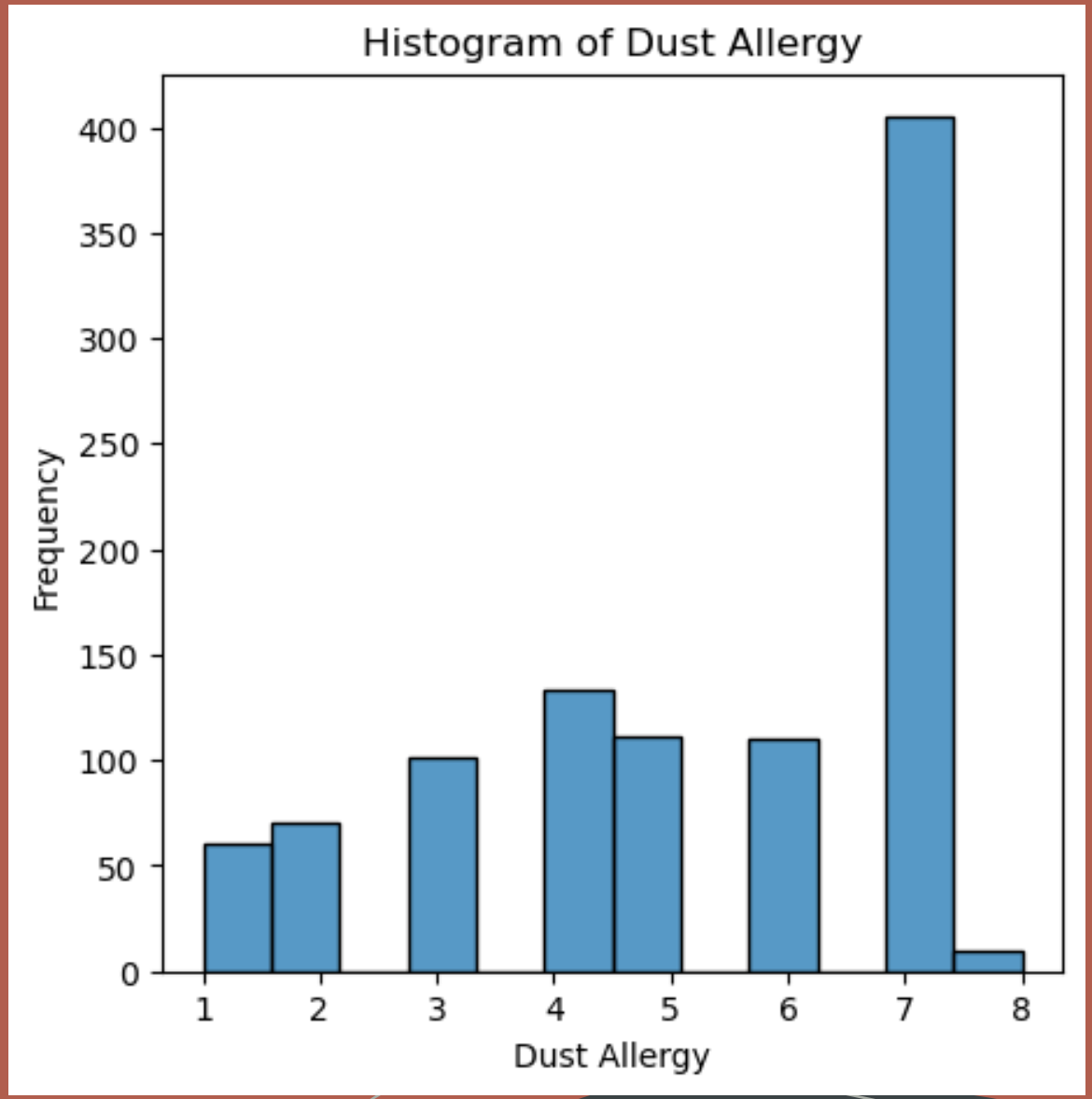
Mean: 4.563
Mode: [2]
Variance: 6.86689789789 7863
Spread(IQR): 5.0
Tail:
995 7
996 8
997 5
998 8
999 5



HISTOGRAMS

Histogram of Dust allergy

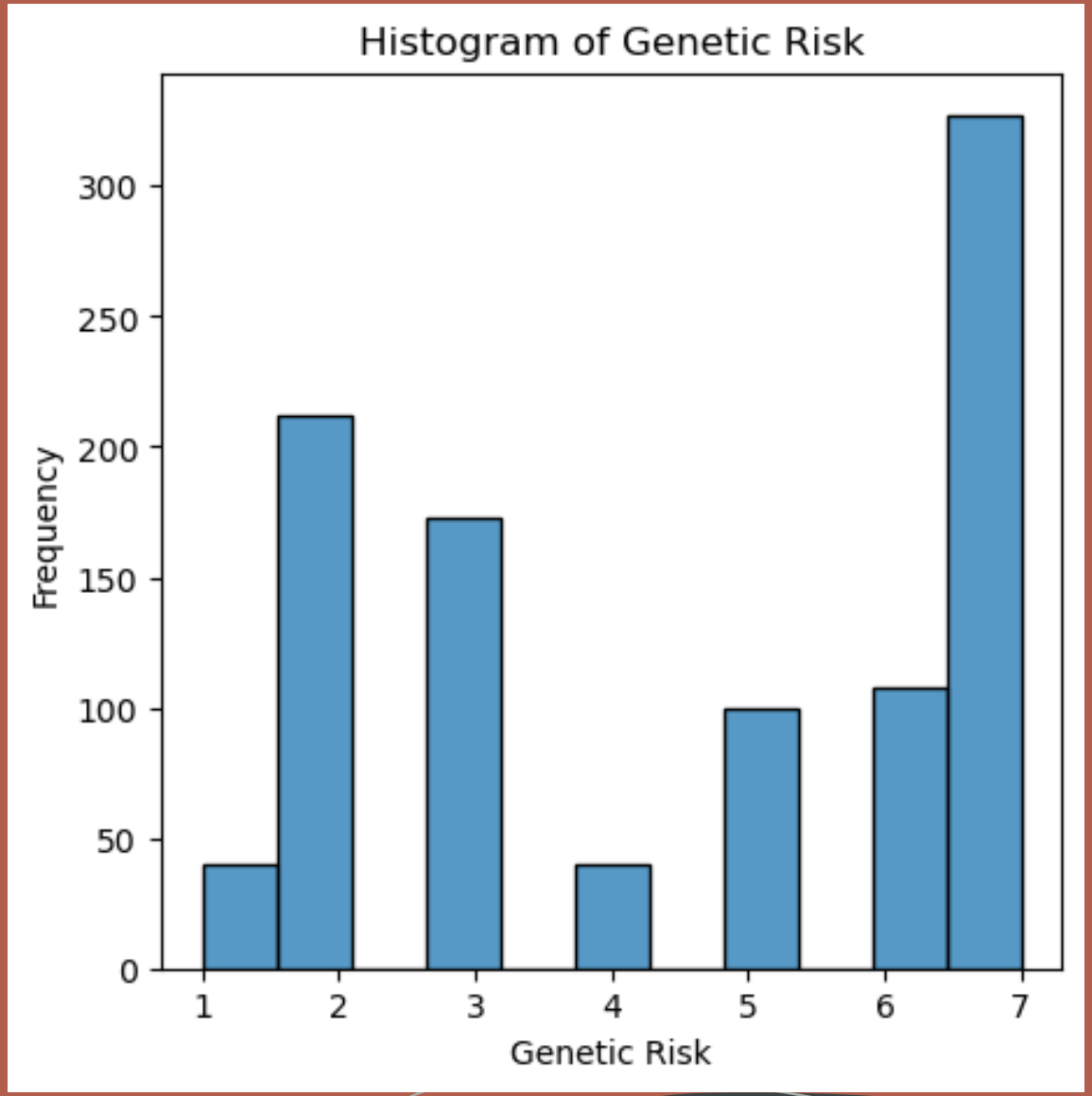
Mean: 5.165
Mode: [7]
Variance: 3.9236986986986877
Spread(IQR): 3.0
Tail:
995 7
996 7
997 6
998 7
999 6



HISTOGRAMS

Histogram of Genetic Risk

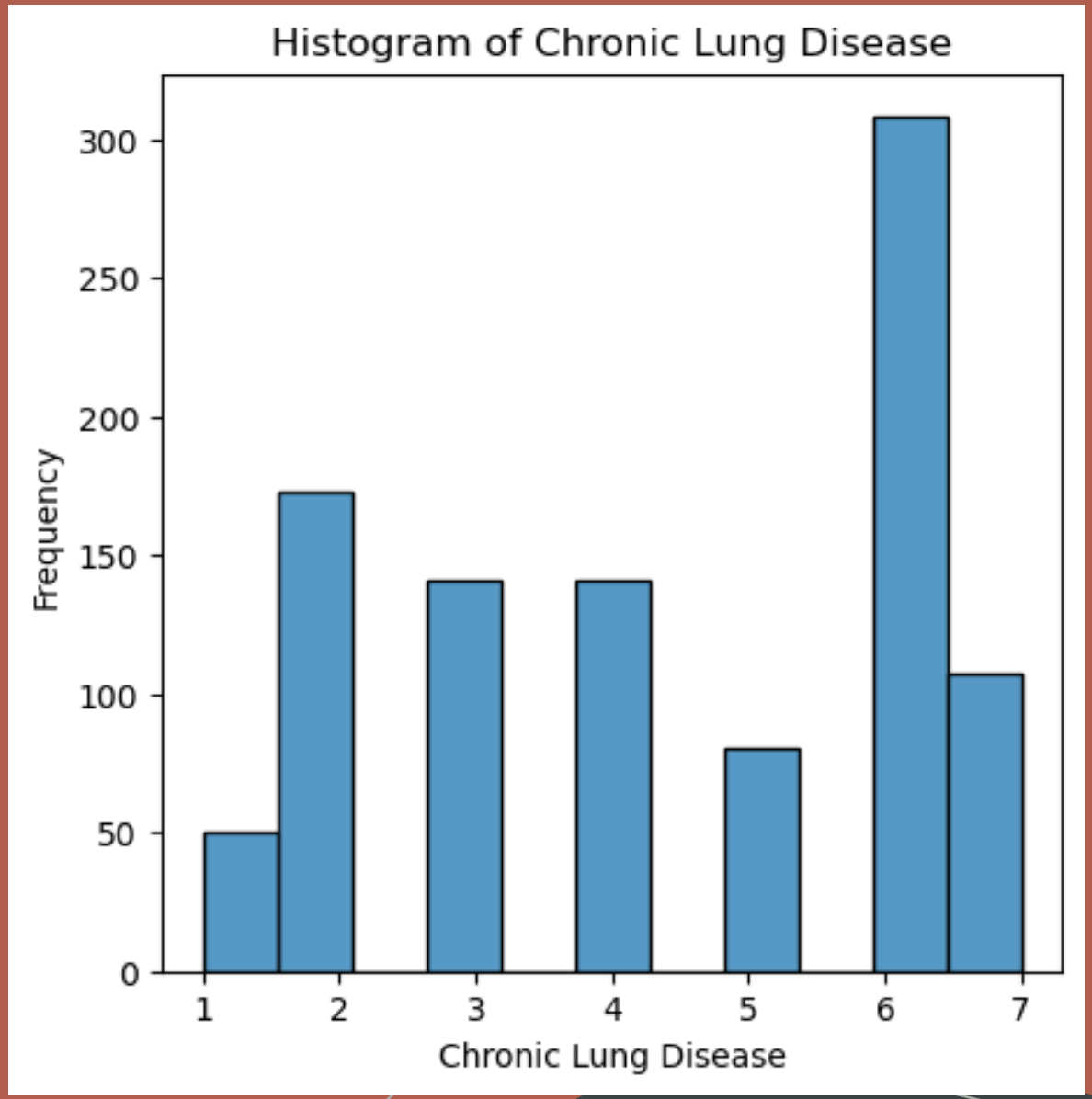
Mean: 4.58
Mode: [7]
Variance:
4.524124124124109
Spread(IQR): 5.0
Tail:
995 7
996 7
997 5
998 7
999 5



HISTOGRAMS

Histogram of Chronic Lung disease

Mean: 4.38
Mode: [6]
Variance:
3.4170170170170 18
Spread(IQR): 3.0
Tail:
995 6
996 6
997 4
998 6
999 4

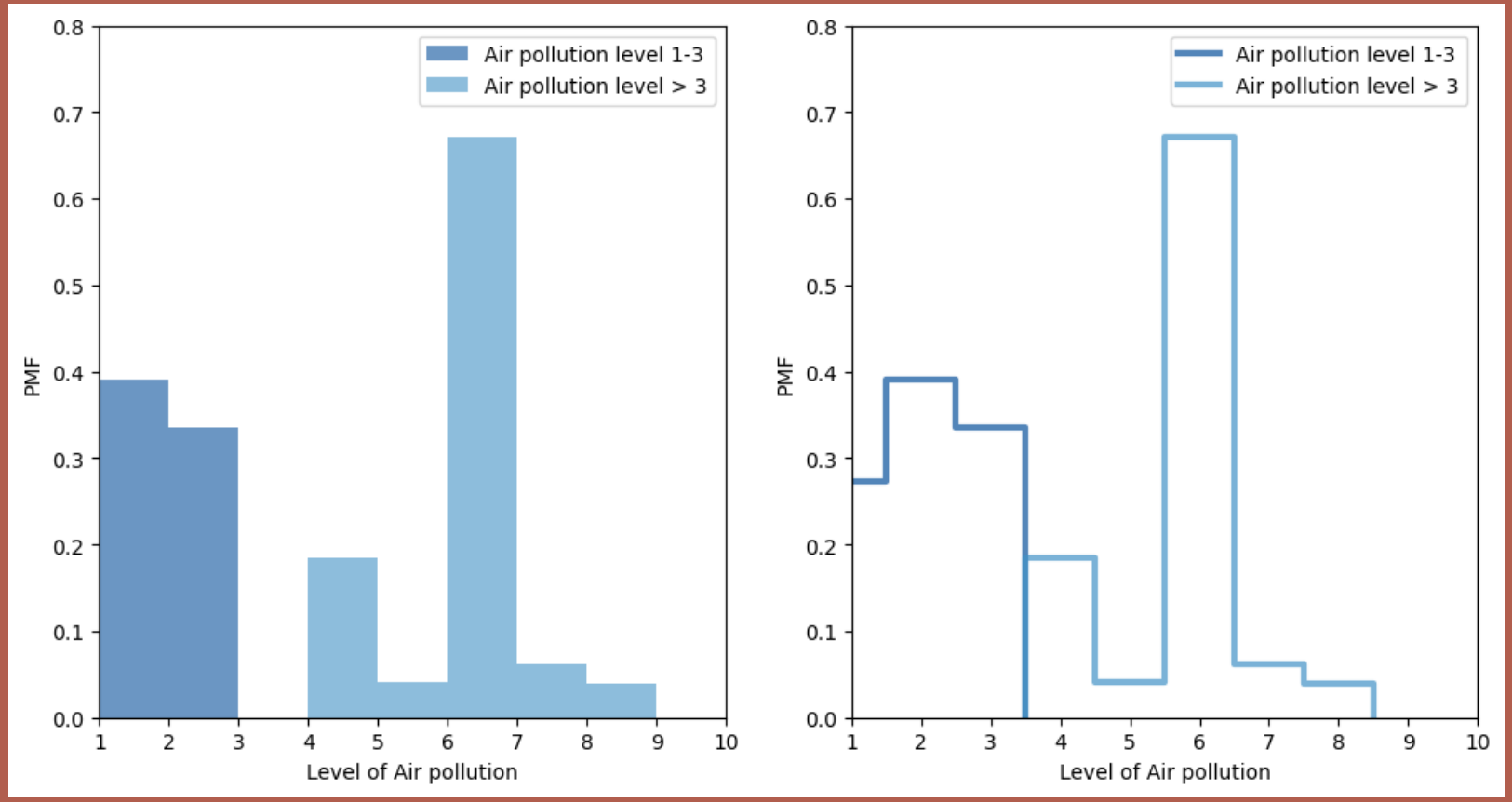


Outliers:

- These variables doesn't have any outliers. Based on the data's source, the maximum and lowest numbers in this dataset seem reasonable. Nothing out of the usual that may affect the analysis doesn't seem to exist.

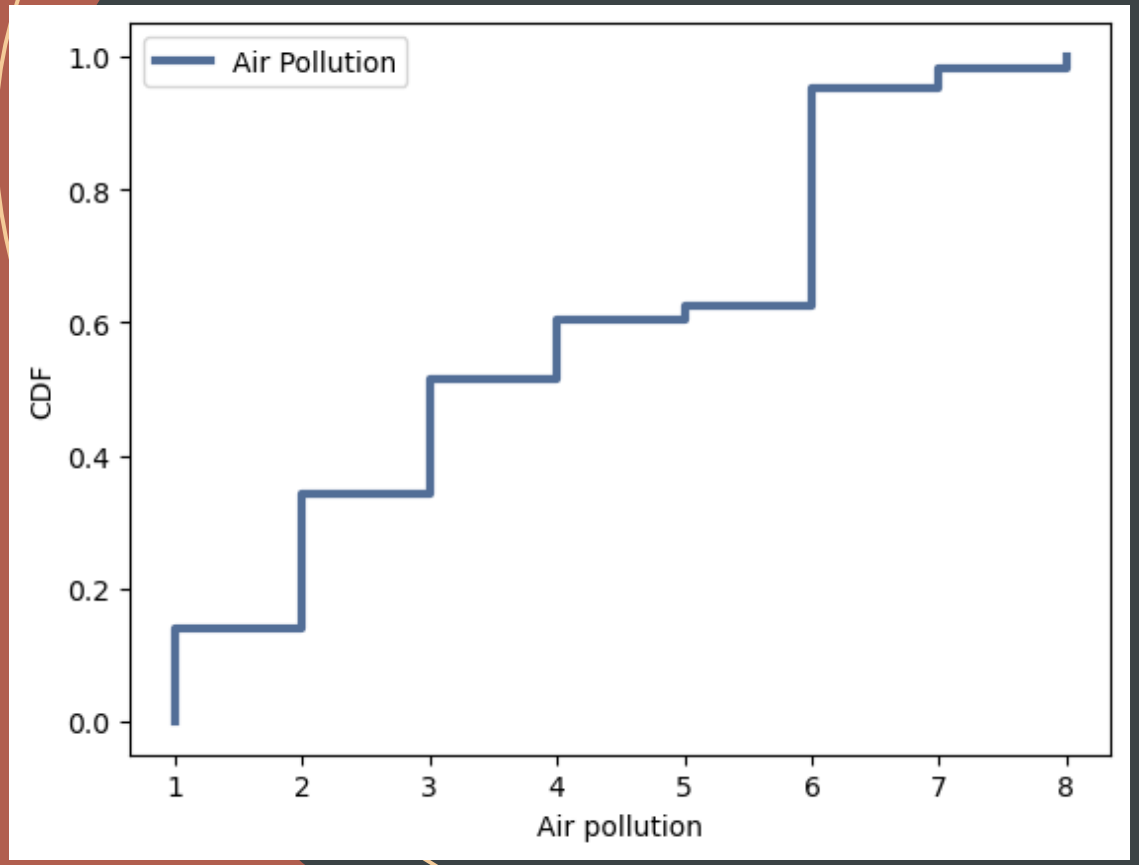
PMF

- So we are comparing 2 Scenarios of the same variable:
- Here we are considering the variable 'Air Pollution' and taking two scenarios where considering the airpollution level from 1-3 and comparing it with the rest of the range inorder to understand how it affects the lung cancer prediction



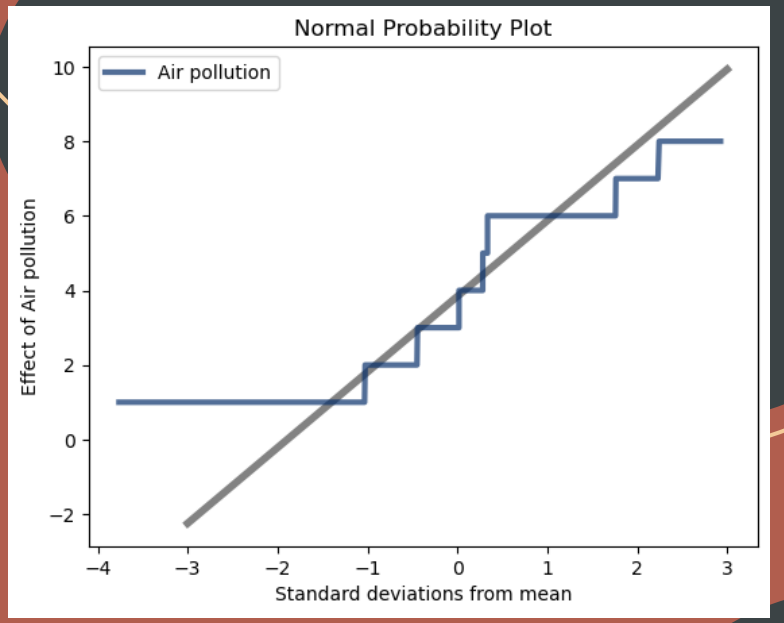
CDF

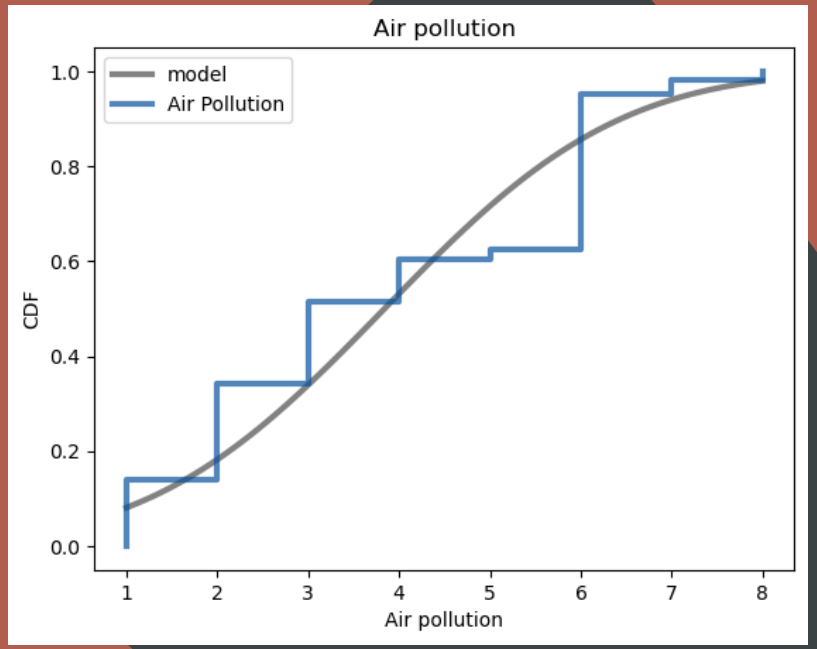
- Here we are considering the variable 'Air pollution' and performing CDF on it.
- According to the CDF the chances of having the cancer has increased along with the level of airpollution.
- The most common level of air pollution according to the CDF will be 6 which means that most of the patients in the given data have been exposed to a level 6 air pollution. We can observe that there is just 25% of the people who are effected by the air pollution under the level of 1-3 but we can observe that more than 75% has the exposure to the higher levels of air pollution.



Analytical distribution

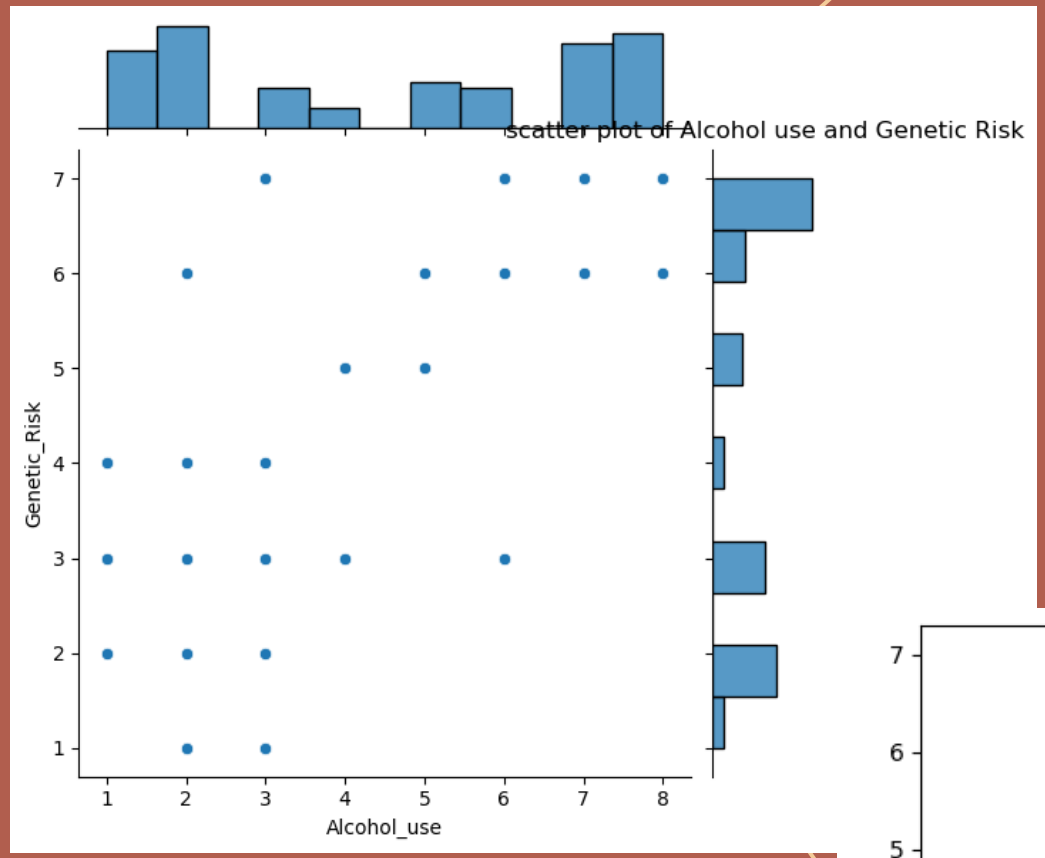
- Here we are considering Normal Distribution
- The curve is almost similar to the CDF graph which shows that is has the expected value. The highest value of the air pollution 8 is in 90th percentile rank and the lowest ranges are in 30th percentile which clearly shows that the effect of the air pollution is higher in higher ranges.





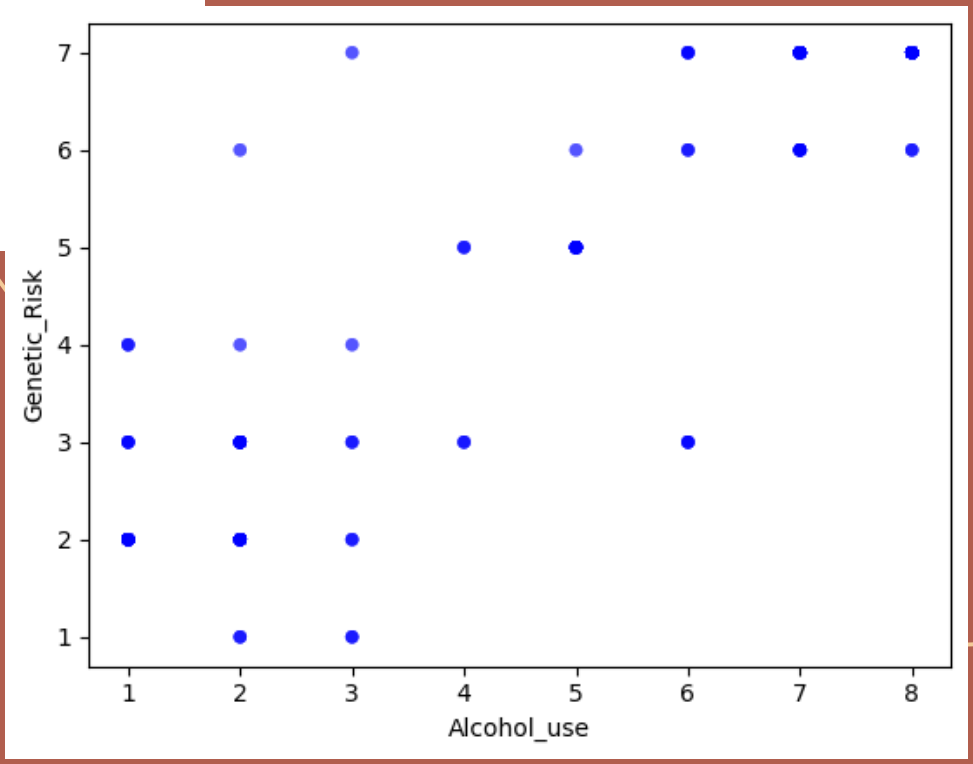
Scatter Plots

- Here we are comparing two variables 'Alcohol_use' and 'Genetic_Risk'. Using the Seaborn and Joinplot(fig 1) and thinkplot (fig 2)
- From the scatter plot we can say that these two variables have many overlapping points almost 80% of them which depicts the correlation between the two variables



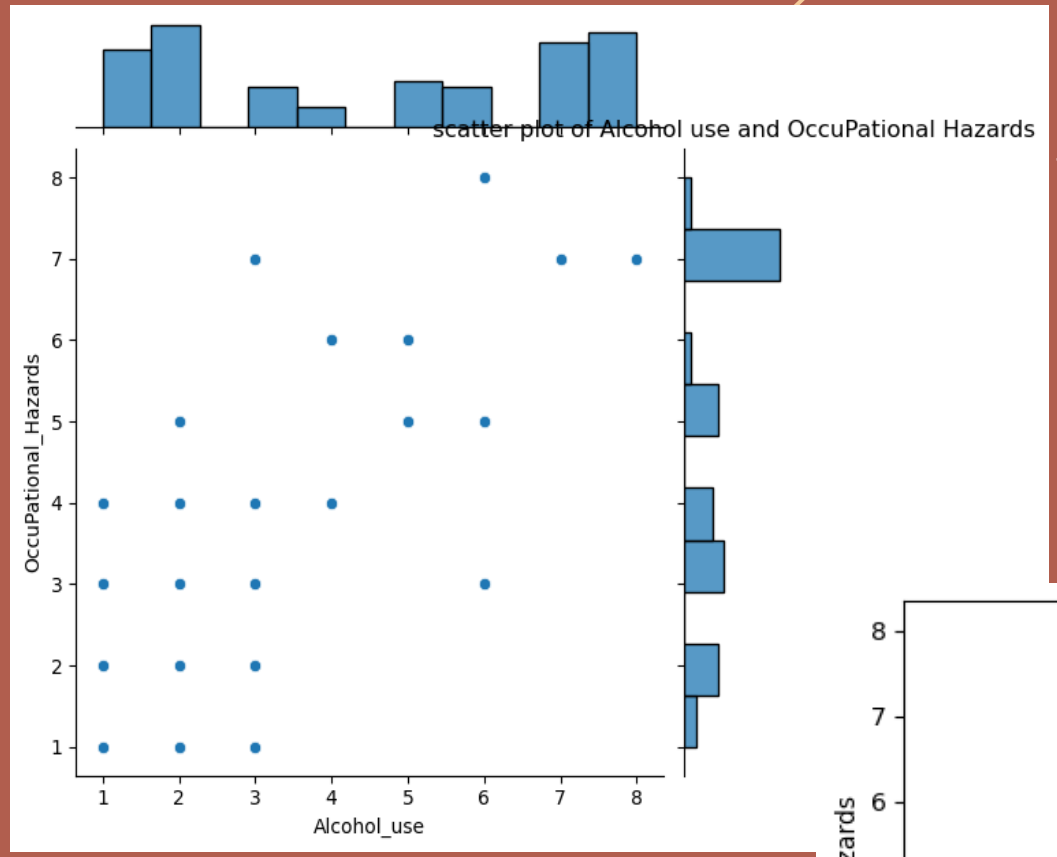
<- Fig 1

Fig 2 ->



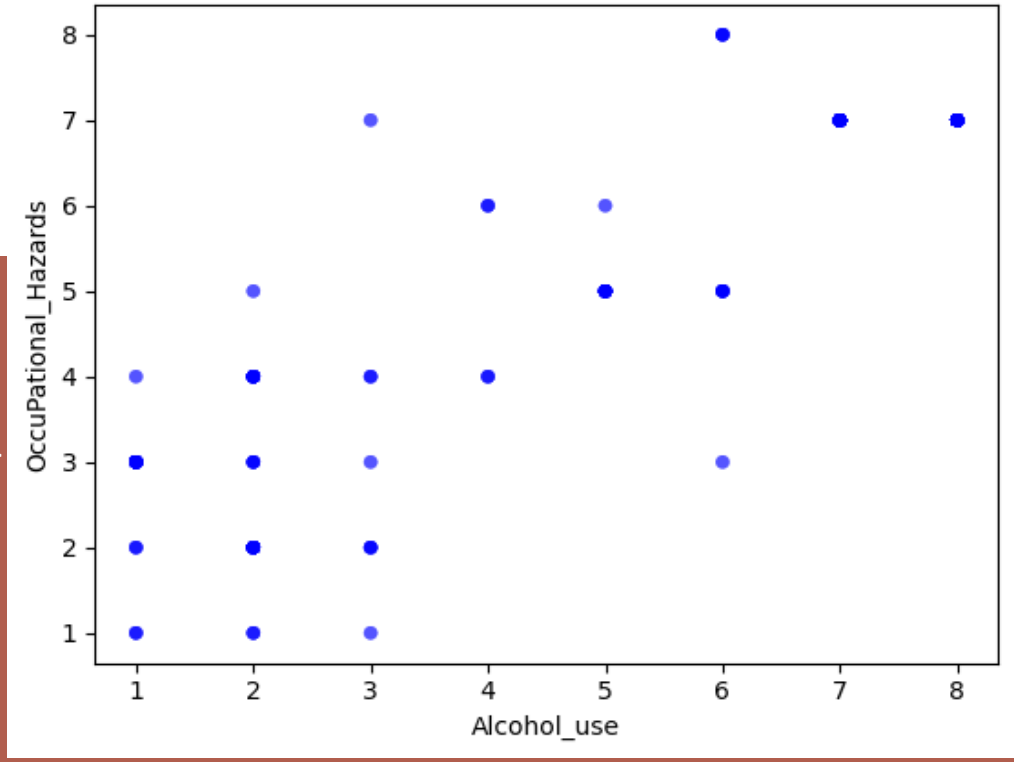
Scatter Plots

- Here we are comparing two variables 'Alcohol_use' and 'OccuPational_Hazards'. Using the Seaborn and Joinplot(fig 1) and thinkplot (fig 2)
- From the scatter plot we can say that these two variables have many overlapping points almost 90% of them which depicts the correlation between the two variables



<- Fig 1

Fig 2 ->



Hypothesis Testing

Null Hypothesis

My hypotheseis is to point out the no correlation between the Age and Airpollution variables which depicts that in the lung cancer prediction the Air pollution effect doesn't depend on the age of the person

pvalue: 0.0

The pvalue shows that in the lung cancer prediction the air pollution effect is not dependent on the age of the patient

Regression Analysis

- The formula is 'Alcohol_use ~ Genetic Risk'
- Result is :

  Intercept: -0.3867359943357907

  Slope: 1.0807283830427465

  p-value of slope estimate: 2.7193e-320

  R^2: 0.7694971870880725

- An R-squared score of 0.7695 suggests that the independent variables in the model can account for around 76.95% of the variability in the dependent variable. This number ranges from 0 to 1, with 0 indicating that the model explains no variability and 1 indicating that it fully accounts for all variability.

Data set

- Data source:

Link: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link?resource=download