University of Hertfordshire UH

Clustering and Data Fitting

Name: Rimsha Liaqat

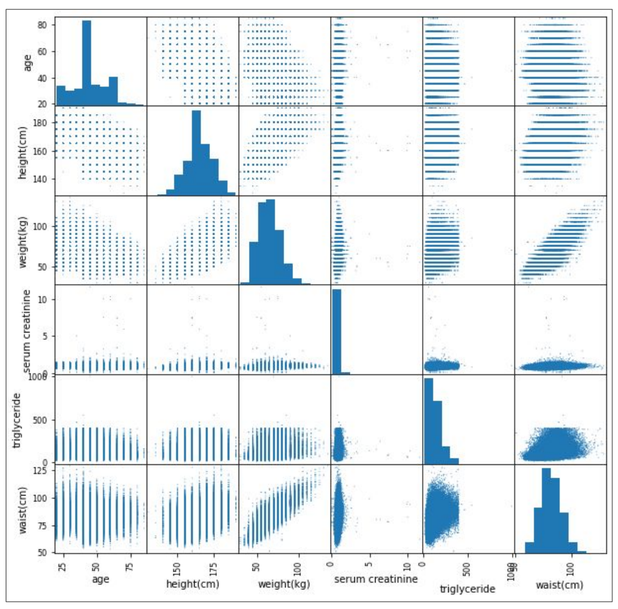
Student ID: 22021651

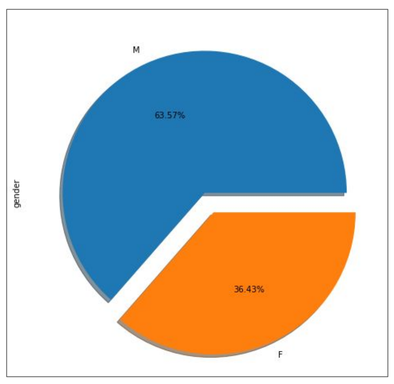
Github Repository Link: https://github.com/rl22aas/Clustering-Assignment.git

Smoking, distribution of smoking, and properties of a smoker

Summary

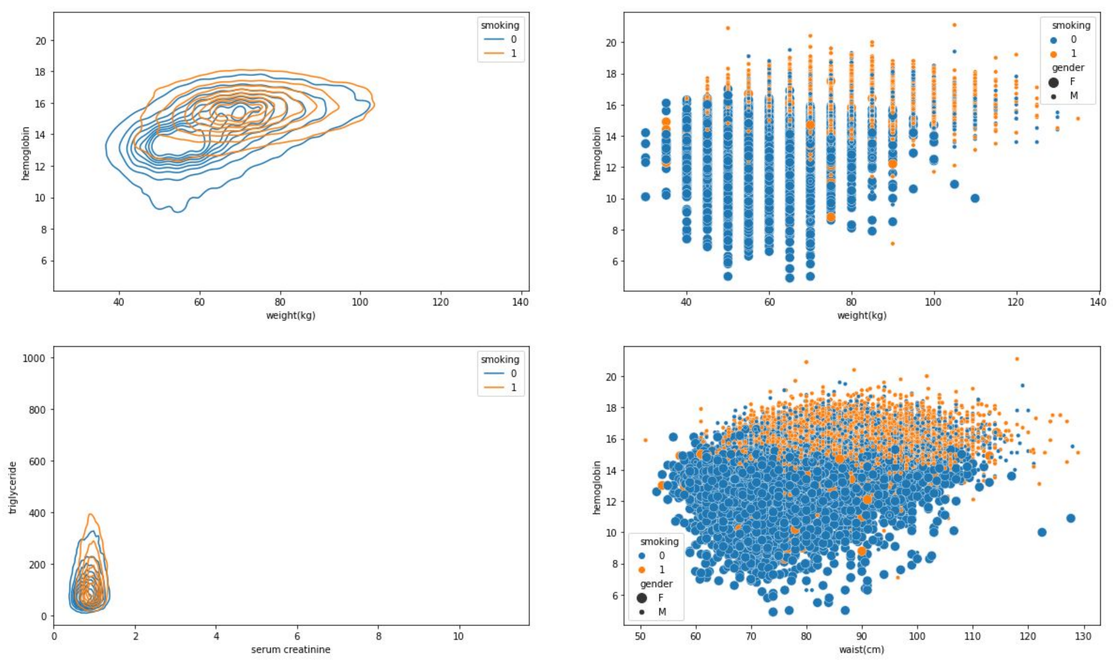
- The current era is the most busiest era in this world. Everyone is busy in earning, working whole day, and not caring about physical and mental health. This leads the current generation towards anxiety.
- To get rid of this anxiety, youth prefer to smoke some cigarettes. This is much harmful than anything because it nearly affects each and every organ of the human body[1].
- I study the characteristics of a smoker using an open-source dataset from Kaggle and try to represent how to get to know that someone is a smoker or not.
- The following scatter matrix plot display the relationship between different properties of human body.
- Further, smoking is not only the concern about men nowadays. It is somelike equal for both men and women.





Introduction

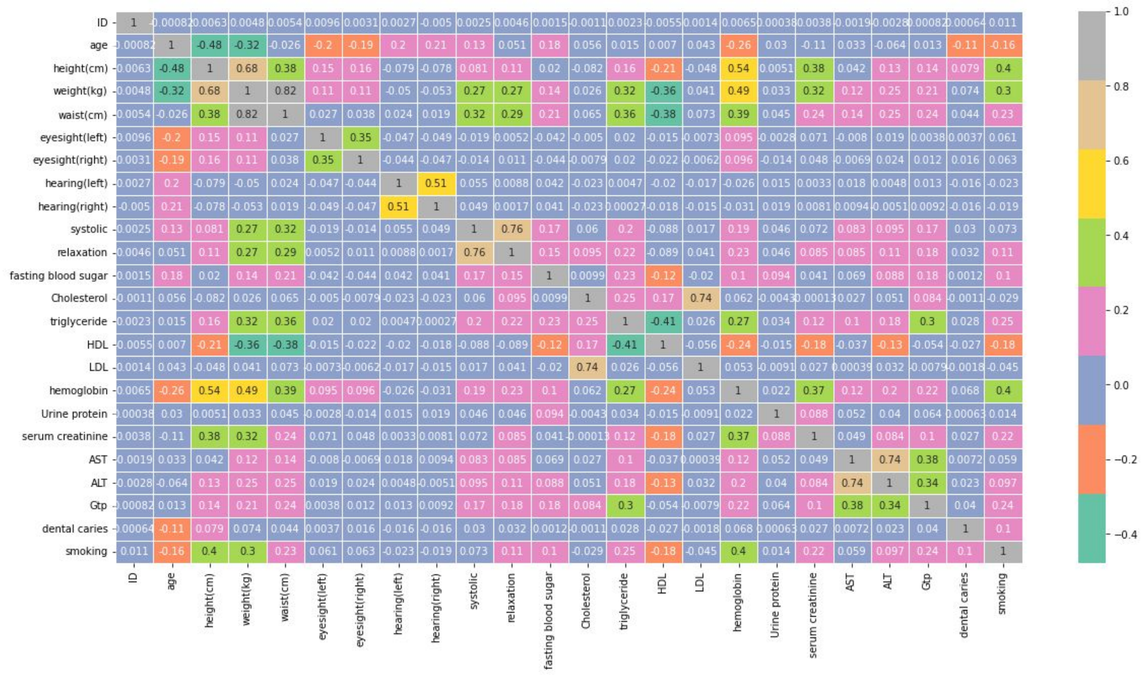
- This research has been carried out to highlight the characteristics of a smoker.
- This can help us to give some helpful advices to the smoking person that will may be lead him to leave this bad habit of smoking.
- To get some extra information about the relation between different entities of dataset, I created some scatter plots and KDE plots (kernel density estimate) to get some idea about how the data is depending on each other.



- In the preceding plot, I have drawn scatter plots against weight and hemoglobin, and waist and hemoglobin.
- The Orange color represents the smokers and blue color represents non-smokers.
- For the distribution of gender, I have created bigger dots for male and smaller dots for females.
- We can draw a horizontal line from 14 units of hemoglobin to distinguish between smokers and non-smokers.

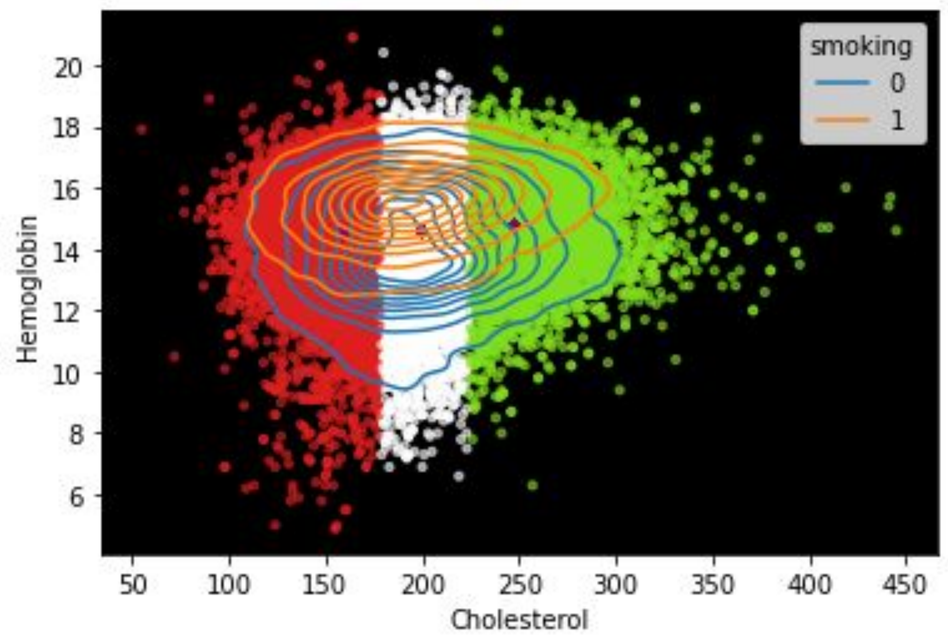
Smoking Data Correlation

- To get the insights of data and the correlation between each characteristic of data, I have created a correlation plot using the columns of smoking dataset.
- In the following image, we can observe that there most of the parameters of the dataset are dependent to each other.



Change in Human Body

- In the following plot, scatter plot, we have fit a KDE plot (kernel density estimate) to represent which area of the plot smokes.
- The yellow area represents the smokers and the blue area of the plot represents the non-smoking persons.
- The persons who smoke has larger hemoglobin than the persons who do not smoke.



References

- [1] https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm

University of Hertfordshire