# 7CS519 Information Visualization

## Visualization of Smokers Vital Signs

Ntoulmperis Michail[1]

[1]107 Patision anenue & Pellinis Str, 11251, Athens, GR
UoD: 100615926, mich.ntoulmperis@mc-class.gr
MSc Big Data Analytics

**Abstract.** In the last decade, it has become common knowledge that how we process, analyze, and visualize Big Data is going to affect the way our society functions. With visualization, which is an essential procedure in data analysis, we can gain insight into the data and gather new information that can help us overcome challenges and gain a great advantage in the decision-making process of any problem. This paper focuses on the visualization techniques that we can use on Tableau and R. Tableau is an easy-to-use software that can provide interactive visualization that can explain the data but with limited capabilities in terms of the application of statistical methods and data pre-processing, while R is a command-line and open architecture interface that gives an analyst a great variety of ways that can perform visualization. For the purpose of this study, we are going to try and visualize different graphs and plots of a data set containing different vital signs of smokers and non-smokers. We are going to try and observe the different values of each variable, depending on if one person smokes while considering their gender.

**Keywords:** Smoking, Smokers, Non-Smokers, Visualization, Tableau, R, Weight, Waist Size, Blood Pressure, Triglyceride, Cholesterol, AST, ALT

## Contents

# 1 Introduction

Conventional and everyday database management, data analysis, and visualization tools are unable to keep up with Big Data. While today's scientific community hasn't concluded on a strict definition of what Big Data is, we can describe it using the ''5Vs''[1]. The ''5Vs'' of Big Data refer to the Variety, Value, Volume, Velocity, and Veracity of data that are being generated.

The high volume refers to the huge amount of information a big dataset can have[2]. The speed in which Big Data is generated relates to high velocity, while variety refers to the complexities of Big Data, meaning large datasets can contain a plethora of variables that can sometimes make preprocessing exponentially harder. For example, datasets such as the one we are going to try and visualize can contain different types of variables, that either contain numerical data, for example, the amount of Cholesterol an observation has, or categorical data such as gender which describe if the observation was male or female[3]. Furthermore, the high value of Big Data can be obvious, as we

can give meaning to its variables through various interpretations. Gender, height, or age can greatly influence if a person smokes or not[4]. Lastly, low veracity corresponds to the number of missing values and biased obtained information that large datasets can have. Uncertainty in Big Data can lead to false results which can affect the analysis[5]. In this paper, we tried to demonstrate those aspects of Big Data by picking for our study a data set with a large number of observations and variables of either qualitative or quantitative nature. This data set contains vital signs measurements from smokers and non-smokers. As visualization tools, we are going to use Tableau software and the R - Programming language in the environment of RStudio. A discussion and comparison of the previously mentioned are also going to take place while also describing their suitability to the dataset and their disadvantages regarding how well they visualize data. The main body of the study consists of the process of the visualization and commentary on the results we got. Finally, the conclusion summarizes our findings and compares the tools that we used. In the next part of the report, we are going to present the data set and describe how the preprocessing was done.

## 2 Data Presentation and Pre-Processing

As mentioned above our data set contains vital signs and physiological measurements from smokers and nonsmokers[3]. Our data set has 55692 observations and 27 variables. During the coding process with R, we made sure no missing values were present in our data because that will lead to a great amount of uncertainty, meaning it will alter our computation of plots and graphs. Another part of preprocessing is going to be deleting any unwanted data, with this we save processing power and produce faster and more reliable results as we will make sure that all data used to train the machine learning algorithms is useful to the model, maximizing its efficiency and accuracy. We removed from the data set the first column which included the ID number of each observation as it will serve no purpose in our analysis. Another variable that was deleted from the data set was the "oral" variable which included the oral status of each person, the reason for the deletion was this column had in its entirety the "True" logical value and it wasn't going to contribute to our analysis. Lastly, the values of "gender" and "tartar" variables were given the "1" or "0" value depending on their original value. This was done to make it easier if we need to apply any filter during the visualization process. Finally, several smaller datasets were created, such as "dfMale", "dfMaleSmokers" and "dfMaleNonSmokers" in order to divide the original dataset into smaller more manageable chunks. Each time a sub-dataset got created, we erased the variable that gave its characteristics, for example "dfNonSmokers" had the variables "smoking" and "gender" deleted, as it contained only the male non-smokers. To add to, due to the use of some variables, some of them were transformed from quantitative to qualitative. A description of the variables is given in **Appendix A**. Finally, we should mention that almost none of the preprocessing done in R was needed in Tableau, because the data set was almost perfect for analysis from the start, and thanks to the interactive platform of Tableau, we managed to achieve our goals.

# 3 Visualization

## 3.1 Smokers

As we can see in the graphs produced by R (**Appendix B**) and Tableau (**Appendix C**) not only the sum of the total observations of each gender is different but the number of smokers and non-smokers for each gender is drastically different. Knowing this, the number of female smokers may not give a good visualization as we have very few observations for that sub-group of the data. Despite that, because the main goal of this study is not only to visualize the data set but to compare Tableau and R, we are still going to include this sub-group in our study.

## 3.2 Cholesterol and Triglyceride

Continuing with the visualization of the data set using Tableau we produce a bar plot of the variables that described the cholesterol and triglyceride for each gender and depending on if they are smoking or not while considering the age of each observation as seen in **Appendix D** and **Appendix E.** On the other hand, as we can see in **Appendix F** we produced a similar plot in R, but since in R we can easily create sub-groups of the data, we included the age as a variable while the smoking and gender variable are hidden inside the dataset that we have defined. From these graphs, we can conclude that the fact a person smokes, greatly affects the levels of cholesterol and triglyceride, especially in males, because as previously mentioned we don't have enough data for the female smokers to be represented accurately. It is worth mentioning that as males grow older their cholesterol and triglyceride levels seam gradually fall while the levels of old smoker males almost keep the same max values. The reason we chose a violin plot and not a bar plot for the graph we computed in R, was so we could show the range of the age variable for each cholesterol and triglyceride maximum and minimum that we had.

## 3.3 Blood Pressure

For Blood Pressure since we already observed some differences in Tableau and R when we make plots with the same variables, this time we tried to gain two different insights into the data. In **Appendix G,** which was produced by R, we can see for male smokers have a high concentration of both blood pressure variables when they approach their maximum values. This means, that males who smoke tend to have more common extreme blood pressure variables, meaning systolic and relaxation blood pressure. On the other hand, as we can see in **Appendix H,** there isn't any remarkable amount of correlation between the blood pressure variables and the age variable among the male smokers and non-smokers thanks to the similar line that the maximum values draw for each bar. It is also worth mentioning that for the Tableau graph we computed the average of each variable instead of its sum. Again, for the lack of data, we can't make any legitimate commentary on the female observations.

### 3.4 The Enzymes AST and ALT

In both bar plots of **Appendix I** and **Appendix J** we can say that there is no large correlation to the age variable, meaning the height of each bar, stays relatively the same, except in the case of 40-year-old males. We can see that smokers tend to have less of both enzymes compared to smokers. But if we take into account the average value of each variable, we can see in **Appendix K** that the bar plot of the enzymes for both smokers and non-smokers almost look identical. This means that maybe, smoking doesn't affect the values of the enzymes after all.

### 3.5 Waist Size and Weight

Last but not least, we wanted to visualize any connection between obesity to smoking by trying to make graphs of waist size and weight as seen in **Appendix L** and **Appendix M**. As we can see in the graph of male smokers, high values of weight and waist size have many concentrated values, while on male non-smoker they seem to be more spread out. If we link high waist size to obesity, we can assume that people with high waist size and high weight tend to be obese, in other words, smokers tend to be more obese compared to non-smokers. From Tableau, we can confirm this because the biggest amount of concentrated values is linked to male smokers. Lastly, although we have a little number of female smokers, in this graph we can observe the same pattern as the male ones.

## 4 Conclusion, R vs Tableau Software

After some pre-processing was done so we could easily code in R and after describing what tools we are going to use we started the visualization process. In this study, we showcased that there were more male smokers than female ones. Also, we learned that while old male smokers have high values of cholesterol and triglyceride, for most of the observations it made no difference if a person smoked or not because the graphs, we made were almost identical. For blood pressure, extreme values of the variable seemed to be more concentrated for the male smoker in comparison with non-smokers, but when the average values got taken into account the graphs again showed no significant correlation between smoking and blood pressure. For the enzymes, we observed that both female and male smokers tend to not correlate with the age variable, except for when they are 40 years. Both graphs in R and Tableau showed that maybe smoking doesn't affect in great the amount of enzymes in one's body. Lastly, we found evidence of obesity being linked to smoking.

Tableau is a data visualization software that simplifies raw data into an understandable format. It was easy to use and demanded if any, little knowledge of programming. The implemented interactive environment with drag and drop actions was used to create visualizations automatically and sped up the prosses of the analysis while offering customization options, such as different graphs, colors, and more. What we mean by this is that Tableau is a less time-consuming software with a very straightforward

procedure of visualization. The dashboards and worksheets that were created with Tableau were easy to be imported into many types of formats[6].

As an environment and programming language R especially excels in statical data analysis and visualization. Thanks to its unique ways of doing calculations with great speed and efficiency using array-based numerical computations and its dynamic and reflective scripting, R can be used to Pre-Process, analyze and Visualize Big Data spectacularly. The graphical output we got was of high quality which was mainly accessed from basic libraries and some additional installed packages[7]. However, using R as a visualization tool, demands a great deal of programming knowledge. Although the libraries in R made the task easier in some sense, mastery of its syntax was needed to utilize some of its potential, which seems limitless.

In conclusion, if a dataset contains a lot of missing values and generally needs a great amount of preprocessing the R is needed to make the data set in the proper form for analysis. On the other hand, if the data set is ready for analysis from the start, then the Tableau software is the way to go, as it is way easier to use, faster, and easy to learn for anyone, even if no prior knowledge of programming is available.

## References


[1] X. Wang and Y. He, "Learning from Uncertainty for Big Data," *Ieee Syst. Man Cybern. Mag.*, no. August 2016

[2] J. Bendler, S. Wagner, T. Brandt, and D. Neumann, "Taming uncertainty in big data: Evidence from social media in urban areas," *Bus. Inf. Syst. Eng.*, vol. 6, no. 5, pp. 279–288, 2014

[3] Kukuroo3, "Find smokers by vital signs," *https://www.kaggle.com/*. https://www.kaggle.com/datasets/kukuroo3/body-signal-of-smoking (accessed May 23, 2022).

[4] N. Experiments, "Editorial Identification of Treatment Effects," vol. 1131, no. 2007, pp. 1127–1131, 2008

[5] A. Berko and V. Alieksieiev, "A Method to Solve Uncertainty Problem for Big Data Sources," *Proc. 2018 IEEE 2nd Int. Conf. Data Stream Min. Process. DSMP 2018*, pp. 32–37, 2018

[6] N. Balaji, B. H. Karthik Pai, B. Bhat, and B. Praveen, "Data Visualization in Splunk and Tableau: A Case Study Demonstration," *J. Phys. Conf. Ser.*, vol. 1767, no. 1, 2021

[7] B. T. Y. Widemann, C. F. Bolz, and C. Grelck, "The functional programming language R and the paradigm of dynamic scientific programming (Position paper)," *Lect. Notes Comput. Sci. (including Subsea. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7829 LNCS, no. January, pp. 182–197, 2013

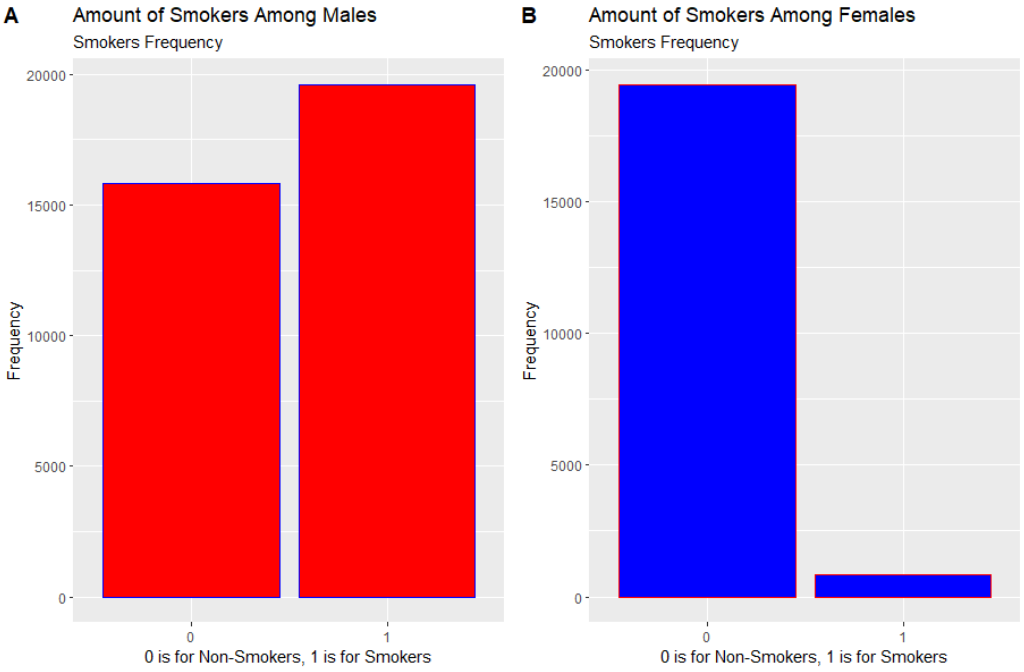
# Appendices

## Appendix A

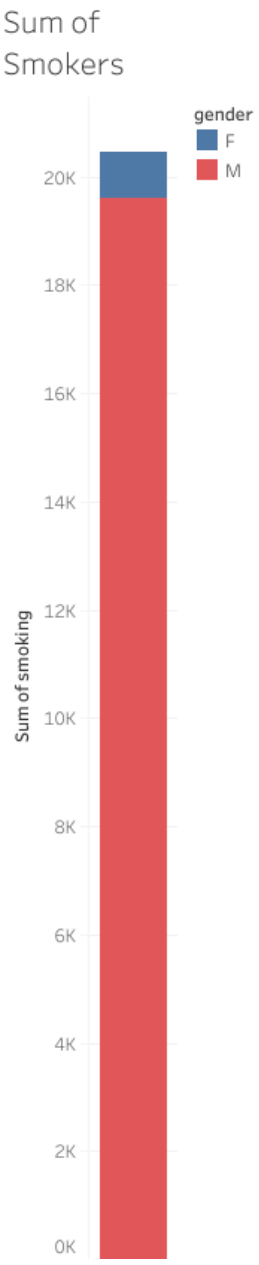
*Description of the Variables.*

- **gender :** If "1" the observation was male and "0" it was female
- **age :** The age of each observation
- **height.cm. :** The height in cm of each observation
- **weight.kg. :** The weight of each observation in kg
- **waist.cm. :** The waist size of each observation in cm
- **eyesight.left. :** A score of that describes each observations eyes sight in its left eye
- **eyesight.right. :** A score of that describes each observations eyes sight in its right eye
- **hearing.left. :** A score of that describes each observations' hearing in its right ear
- **hearing.right. :** A score of that describes each observation hearing in its left eye
- **systolic :** Blood pressure measurement of each observation.
- **relaxation :** Blood pressure measurement of each observation.
- **fasting :** Blood sugar measurement of each observation.
- **Cholesterol :** Total Cholesterol measurement of each observation.
- **triglyceride :** Triglyceride measurement of each observation.
- **HDL :** Cholesterol type measurement of each observation.
- **LDL :** Cholesterol type measurement of each observation.
- **hemoglobin :** Hemoglobin measurement of each observation.
- **Urine protein :** Urine protein measurement of each observation.
- **serum creatinine :** Serum creatinine measurement of each observation.
- **AST :** Glutamic oxaloacetic transaminase type measurement of each observation.
- **ALT :** Glutamic oxaloacetic transaminase type measurement of each observation.
- **Gtp :** γ-GTP measurement of each observation.
- **dental caries :** Number of dental carries of each observation.
- **tartar :** If "1" the observation had tartar and "0" it had not.
- **smoking :** If "1" the observation was a smoker and "0" it was a non-smoker.

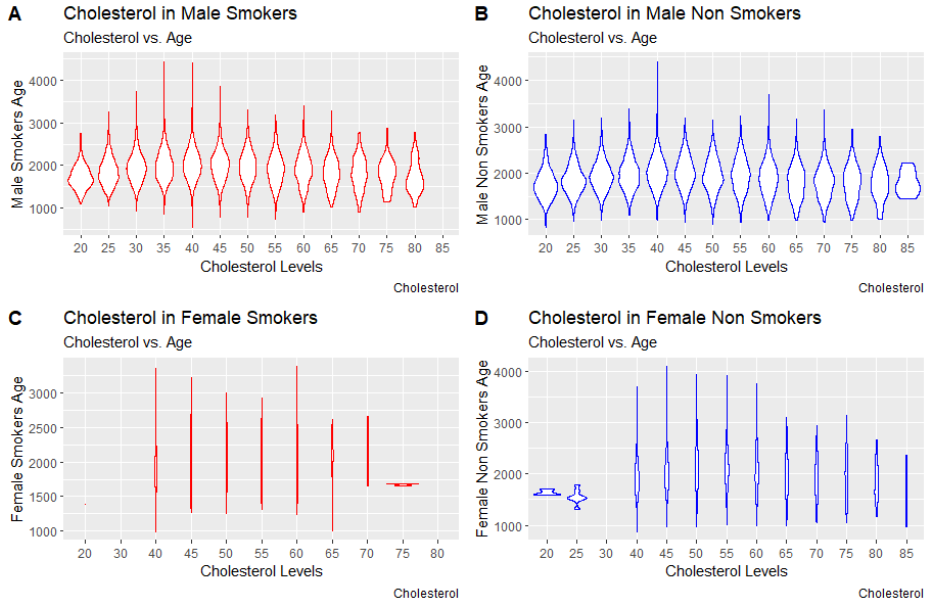
## Appendix B


A
Amount of Smokers Among Males
Smokers Frequency
Frequency
20000
15000
10000
5000
0
0
1
0 is for Non-Smokers, 1 is for Smokers
B
Amount of Smokers Among Females
Smokers Frequency
Frequency
20000
15000
10000
5000
0
0
1
0 is for Non-Smokers, 1 is for Smokers

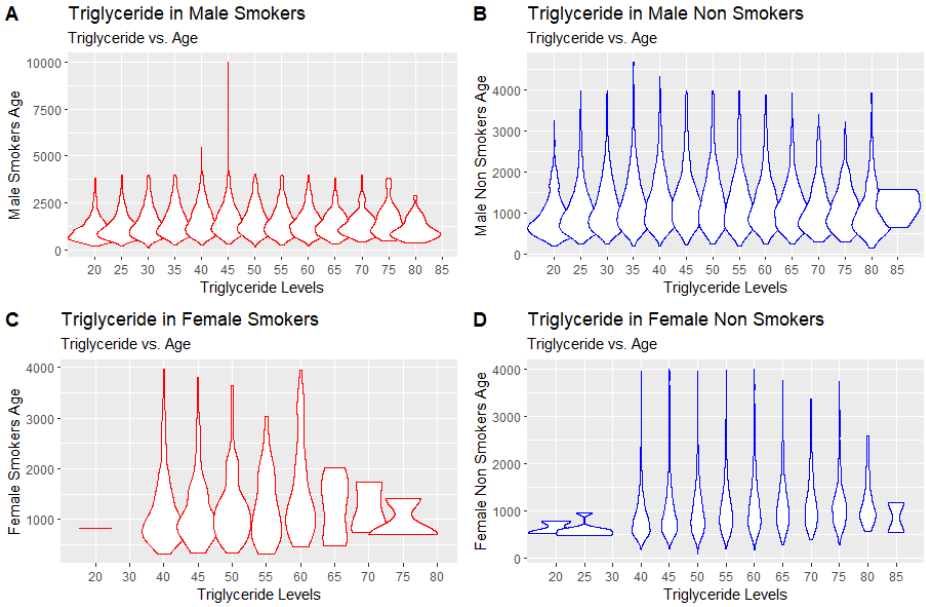
**Appendix C**


Sum of
Smokers
gender
F
M
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
Sum of smoking

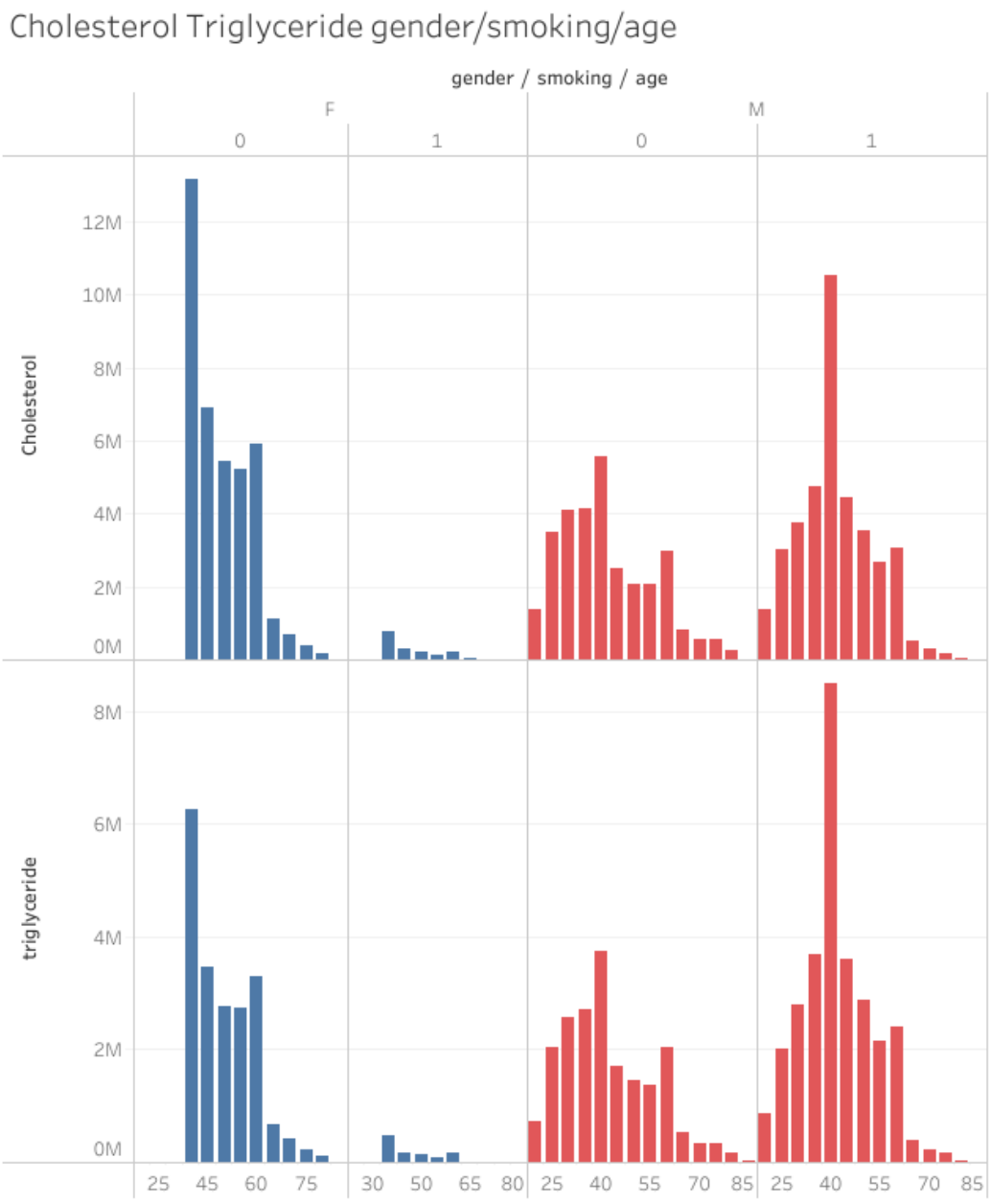
## Appendix D



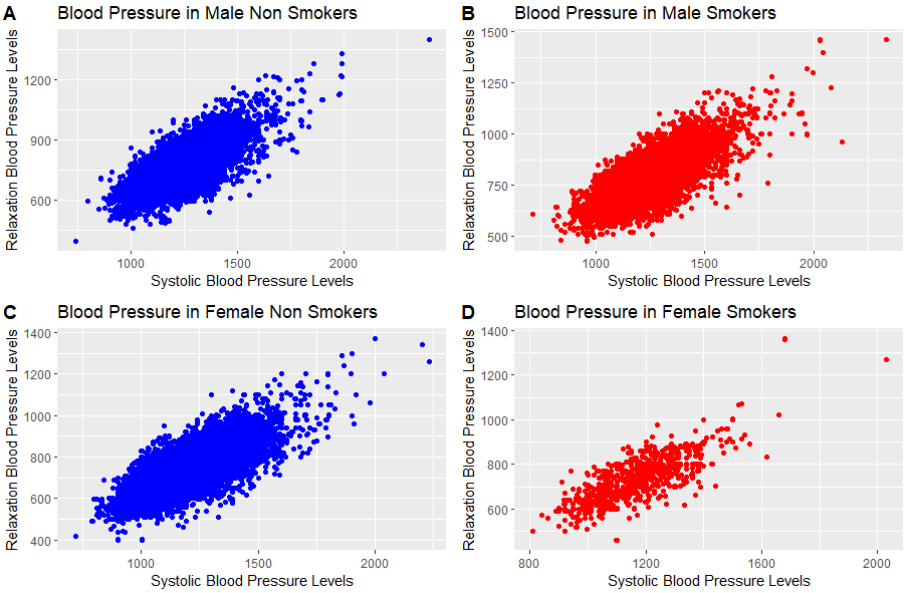

## Appendix E



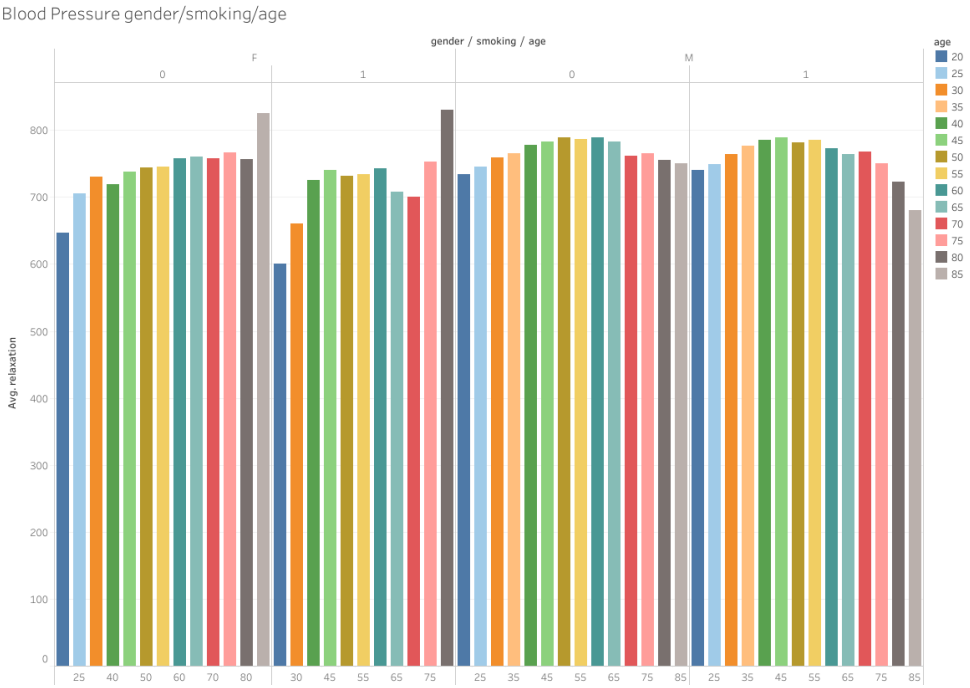
**Appendix F**


Cholesterol Triglyceride gender/smoking/age
gender / smoking / age
F
M
0
1
0
1
Cholesterol
12M
10M
8M
6M
4M
2M
0M
triglyceride
8M
6M
4M
2M
0M
25 45 60 75
30 50 65 80
25 40 55 70 85
25 40 55 70 85

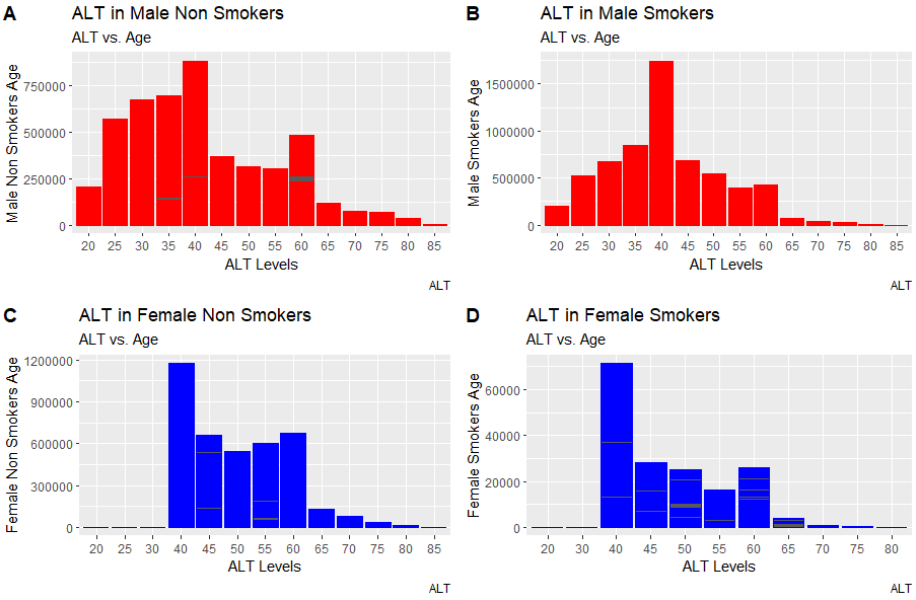
## Appendix G



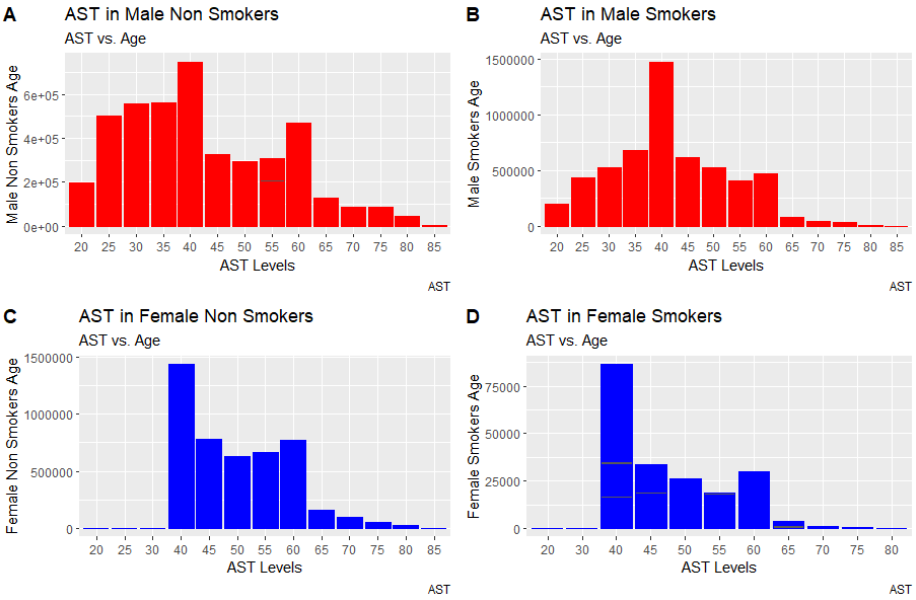

## Appendix H



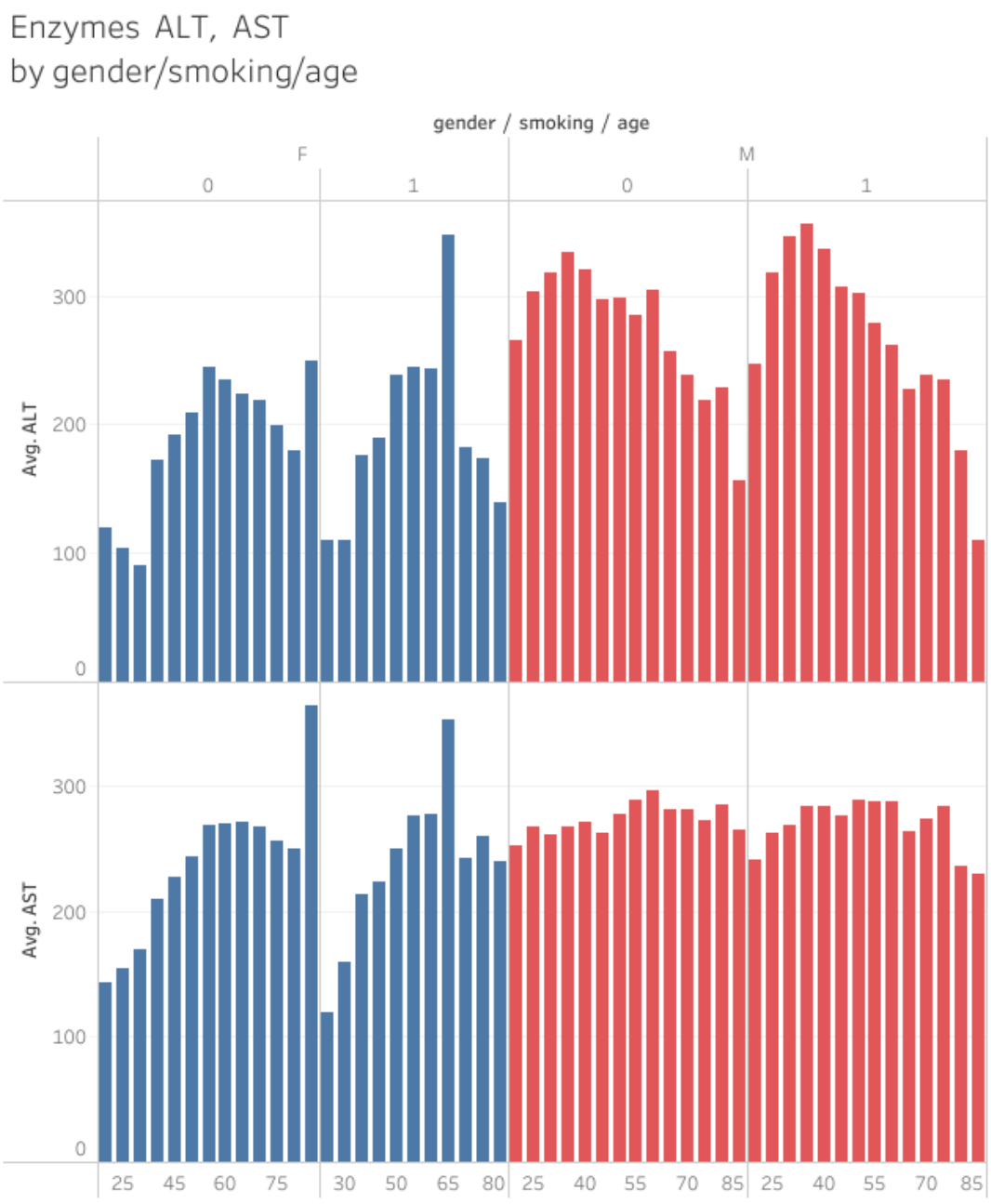
## Appendix I



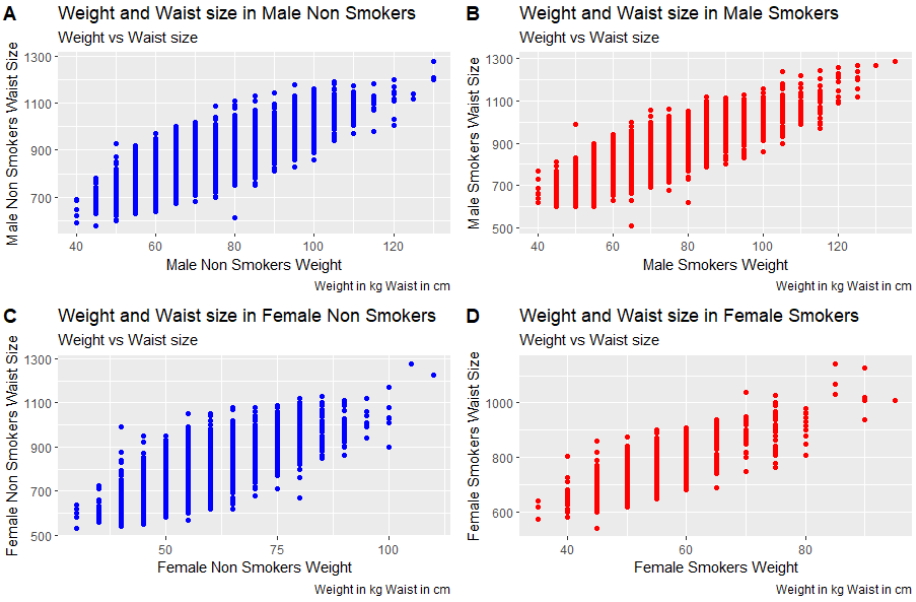

## Appendix J



**Appendix K**


Enzymes ALT, AST
by gender/smoking/age
gender / smoking / age
F
M
0
1
0
1
Avg. ALT
300
200
100
0
Avg. AST
300
200
100
0
25 45 60 75 30 50 65 80 25 40 55 70 85 25 40 55 70 85

## Appendix L




## Appendix M