Example Manuscript Template for a Data Analysis Project

Doreen Kibuule Kalembe

The structure below is one possible setup for a manuscript, or a general data analysis project (including the course project). Adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

Warning: package 'here' was built under R version 4.4.2

**Authors**

* First Autor (ORCID: 0000-0000-1234-5678)
* Second Author
* Third Autor
* Last Author

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Another Affiliation, Athens, GA, USA.
3. Yet another Affiliation.

These authors contributed equally to this work.

Corresponding author: some@email.com

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

MADA PROJECT PART 1:

TITLE: Sex-based analysis/Differences in TB Immune Activation.

This project is aimed at looking at the differences in the Tuberculosis immune activation between males and females.( Sex-based analysis/Differences in TB Immune Activation Research Question)

we shall use data from a tuberculosis public repository. • Data source: Tidyverse https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/AYOFEU with a Sample size of 60 individuals. The data includes 60 different individual that accepted to take part in the study.

The key variables that we shall look at are;

Key variables: - CD4 Immune activation count - Disease severity indicators(Extent of lung disease andNumber of lungs involved) -Fat in kgs - Sex/Gender - BMI - Age of the participant

#load packages  
library("readxl")  
library("readr")  
library("dplyr")

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library("ggplot2")  
library("tidyverse")

Warning: package 'tidyverse' was built under R version 4.4.2

Warning: package 'stringr' was built under R version 4.4.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ lubridate 1.9.3 ✔ tibble 3.2.1  
✔ purrr 1.0.2 ✔ tidyr 1.3.1

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("here")  
  
# follow through with the code folder to see the #variables in the data and how i selected the #variables make this data set.   
#location of the data  
data\_location <- here::here("data","processed-data","Tb\_immune\_data.rds")  
  
#loading the data  
data<- read\_rds(data\_location)  
  
#summary of the data  
summary(data)

ID Sex Participant age BMI   
 Min. :50004 Length:61 Min. :17.00 Min. :15.55   
 1st Qu.:50020 Class :character 1st Qu.:23.00 1st Qu.:18.09   
 Median :50035 Mode :character Median :28.00 Median :19.25   
 Mean :50035 Mean :30.02 Mean :19.83   
 3rd Qu.:50050 3rd Qu.:35.00 3rd Qu.:21.47   
 Max. :50067 Max. :70.00 Max. :28.15   
 NA's :1 NA's :1   
 Dose Fat in kg LBM in kg Fat %   
 Min. :29.61 Min. :-96.640 Min. :-41.98 Min. :-177.647   
 1st Qu.:29.99 1st Qu.: 3.554 1st Qu.: 39.13 1st Qu.: 6.629   
 Median :30.00 Median : 8.704 Median : 43.43 Median : 16.685   
 Mean :29.99 Mean : 9.615 Mean : 43.02 Mean : 17.950   
 3rd Qu.:30.00 3rd Qu.: 13.364 3rd Qu.: 49.82 3rd Qu.: 25.987   
 Max. :30.15 Max. : 80.979 Max. :151.04 Max. : 207.639   
 NA's :1 NA's :1 NA's :1 NA's :1   
 LBM in % CD4+ CD4 Immune activation count  
 Min. :-107.64 Min. : 8.636 Min. : 84.0   
 1st Qu.: 74.01 1st Qu.:185.006 1st Qu.: 914.8   
 Median : 83.32 Median :317.941 Median : 1384.5   
 Mean : 82.05 Mean :337.364 Mean : 1948.1   
 3rd Qu.: 93.37 3rd Qu.:484.889 3rd Qu.: 2057.5   
 Max. : 277.65 Max. :785.896 Max. :17872.0   
 NA's :1 NA's :1 NA's :1

Proposed Analysis Method

• Compare immune activation levels between males and females

• Correlate immune markers with disease severity measures/Correlation between activation levels and disease severity

• Use linear regression to analyze relationships between variables

• Control for potential confounding factors

These are some of our Expected Outcomes.

• Check and Identify any sex-based differences in immune response.

• Quantify the relationship between immune activation and disease severity.

• Propose any insights for sex-specific treatment approaches

# 2. Introduction

## 2.1 General Background Information

Tuberculosis (TB) remains a significant global public health challenge. According to the World Health Organization (WHO), reports show that 10 million new cases and 1.4 million deaths in 2023 (1) were as a result of Tuberculosis. The spread of M. tuberculosis, as the pathogen of TB, has long been hypothesized to occur more often in the household than in the community(2). While most of the M.Tuberculosis transmission in the community is attributed to the unknown contact networks that occurs between individuals, this accounts upto 90% of the transmission. The recent outbreak of Tuberculosis in the united states clearly explains the contact network phenomena. Overall, TB accounts for one quarter of the annual deaths that happen worldwide. An estimated 10.0 million people fell ill with TB in 2018. Geographically, Southeast Asia ranked highest (44%), followed by Africa (24%) and the Western Pacific (18%), with lesser percentages in the Eastern Mediterranean (8%), the Americas (3%), and Europe (3%). Among women of reproductive years, TB is the largest cause of death (3). Global Strategies made by WHO, aim at ending the TB epidemic, with targets to reduce the number of deaths caused by TB by 95% and to decrease new cases by 90% from 2015 to 2035(4).

Mycobacterium tuberculosis infection is acquired mostly through sharing air space with an individual who has active tuberculosis and inhaling droplet aerosols produced by that person. Since this bacterium is transmitted via the respiratory route to susceptible individuals in the community, the household of an infectious person is an environment for intense transmission of M. tuberculosis where most new cases of infection and diseases might be detected and treated. According some reports,they suggest that there are differences in the responses to tuberculosis between men and women. These reports highlight greater barrier to early detection and treatment of tuberculosis to be greater for women than men.

The World Health Organisations, among others, attributed to the differences in the tuberculosis trends between men and women to access to care and treatment, pregnancy effects and gender rules that diminish the social capital of women hence lowering incidence among them(5). Not forgetting that biological mechanisms might influence these differences, previous studies show that men have a higher incidence and prevalence of disease and this might lead to bad treatment outcomes as compared to women.

Since the balance of fat and lean tissue differ in men and women especially at the time of presentation, the objective of this study will be to find out whether the more immune activation there is, the more infected / sicker is the person.

## 2.2 Description of data and data source

The data we used for this project was got from the TRAC dataverse and is fully accessible to everyone. This is considered a safe place where the the Tuberculosis data from the TRAc project is kept and can be used by anyone who wants contribute to TB research. More details about the data can be found [here]!(https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/AYOFEU). The data is publicaly available and can be accessed by anyone.

## 2.3 Questions/Hypotheses to be addressed

Since the balance of fat and lean tissue differ in men and women especially at the time of tuberculosis presentation, the objective of this study will be to find out whether there are differences in the Tuberculosis immune activation between males and females. We shall hypothesize on the more immune activation there is, the more infected / sicker the person.

# 3. Methods

Study Design

This study utilized a cohort design analyzing data from 2021. We included all patients that were 15 years and older. The candidate must have had TB as their first episode, not on any hormonal therapy and does not take any hormonal birth control. The participant was not supposed to have diabetes melitus, cancer or HIV.Similarly,the exclusion criteria applied. All participants that were below the age of 15, reported TB as their second episode, took any form of hormanl therapy or biirth control , had diabetes melitus, cancer or HIV were excluded from the study.

The variables of interest were extracted from the TRAC dataset and exported to an MS Excel sheet. The outcome variable was CD4+ immune activatin counts. The primary independent variable was sex (cis-gender) and this was categorized as male and female. The covariates were age at diagnosis,BMI, Fat in Kg, lean body mass in kg(LBM in kg), CD4+ and CD4 immune activation counts.

## 3.1 Schematic of workflow

Chromosomal sex and hormones play a role in the modulation of immune responses to Mtb and thereby contribute to the general outcome of disease. The first cells that encounter Mtb are known as Macrophages. These initiate a local inflammatory reaction which may differ between the sexes(6). This causes the long term containment of TB infection to differ between men and women since their adaptive immune responses are different. (**tbfig-1?**)

|  |
| --- |
| Figure 1: Impact of sex differences on the immune response to Mtb and their influence disease outcome. |

## 3.2 Data import and cleaning

We imported the data from MS Excel to R. We explored the data and found some missing values for some variables. We decided to drop one observation since the person lacked the primary exposure and did not have most of the details we needed for our final analysis. We categorized the participants age into four categories. That is 17-<28, 28-<40, 40-<60 and >=60.

## 3.3 Statistical analysis

The final cleaned data set contains some negative fat values. Negative values occur when the deuterium dose has not had sufficient time to fully equilibrate with body water, or the dose was not completely consumed. This is seen as low deuterium enrichment, resulting in an overestimation of the size of the body water pool and hence high FFM (FFMI = LBM / height2)and low % body fat.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

We explored the data and using figures and summary tables. We used box plots, histograms and and correlation plots to see whether we can find something interesting in the data. Scatter plots were used to see the trends in the data and also see whether there are some interesting relationships between the continuous variables in the data.

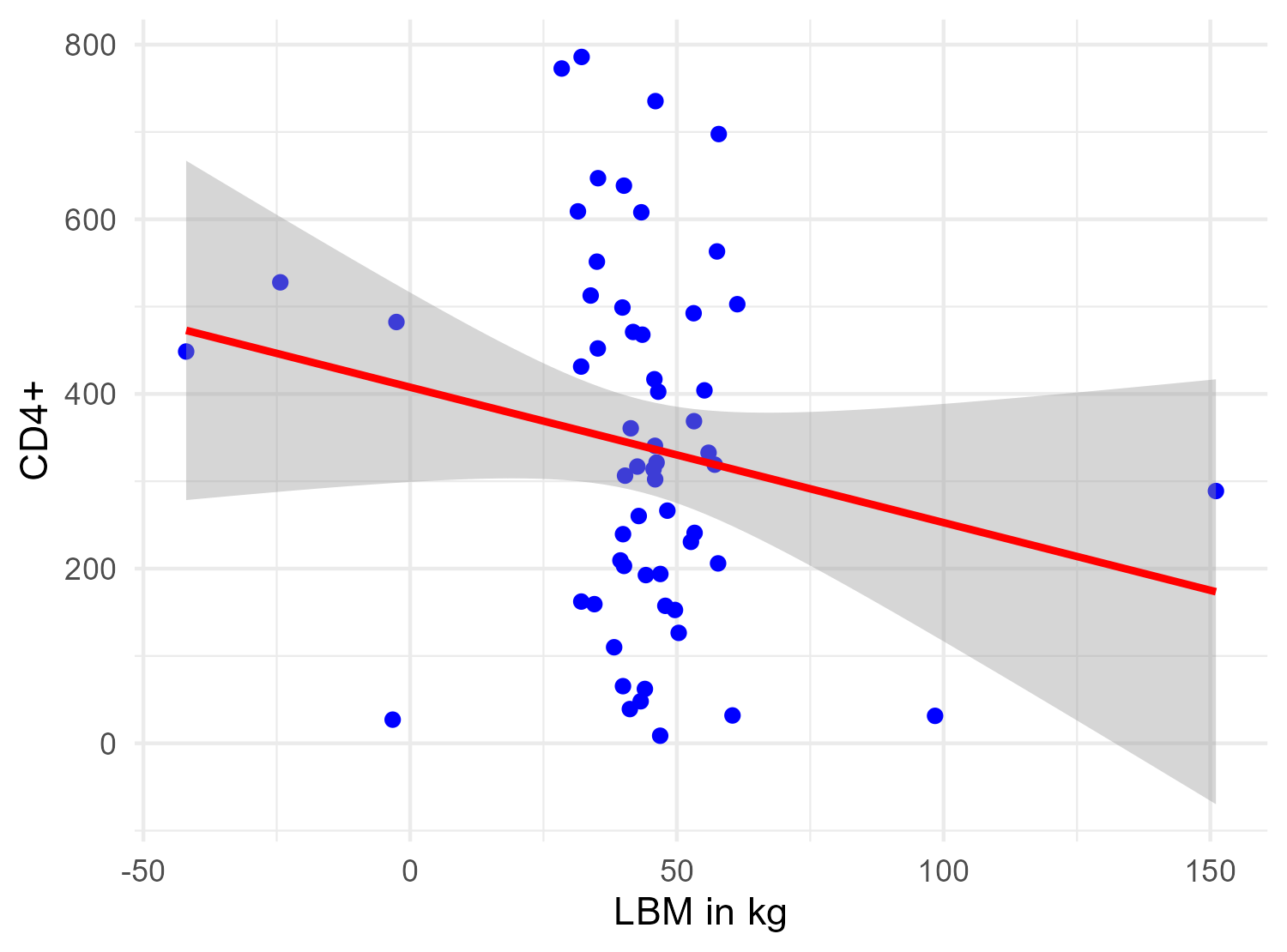
[Table 1](#tbl-TABLE1) shows a summary of the data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data descriptive summary table.   |  | F | M | Overall | | --- | --- | --- | --- | |  | (N=30) | (N=30) | (N=60) | | CD4+ |  |  |  | | Mean (SD) | 394 (217) | 289 (181) | 341 (205) | | Median [Min, Max] | 449 [27.0, 786] | 278 [8.64, 735] | 319 [8.64, 786] | | Missing | 1 (3.3%) | 0 (0%) | 1 (1.7%) | | CD4 Immune activation count |  |  |  | | Mean (SD) | 1750 (1340) | 2160 (3170) | 1960 (2430) | | Median [Min, Max] | 1520 [94.0, 7350] | 1210 [84.0, 17900] | 1380 [84.0, 17900] | | Missing | 1 (3.3%) | 0 (0%) | 1 (1.7%) | | Participant age |  |  |  | | Mean (SD) | 29.9 (12.8) | 30.2 (6.83) | 30.0 (10.1) | | Median [Min, Max] | 26.0 [17.0, 70.0] | 28.5 [19.0, 48.0] | 28.0 [17.0, 70.0] | | Fat in kg |  |  |  | | Mean (SD) | 18.9 (22.4) | 0.348 (21.9) | 9.62 (23.8) | | Median [Min, Max] | 10.9 [-0.192, 81.0] | 4.37 [-96.6, 21.3] | 8.70 [-96.6, 81.0] | | LBM in kg |  |  |  | | Mean (SD) | 31.1 (21.0) | 54.9 (21.1) | 43.0 (24.1) | | Median [Min, Max] | 39.6 [-42.0, 45.9] | 50.0 [38.2, 151] | 43.4 [-42.0, 151] | |

## 4.2 Basic statistical analysis

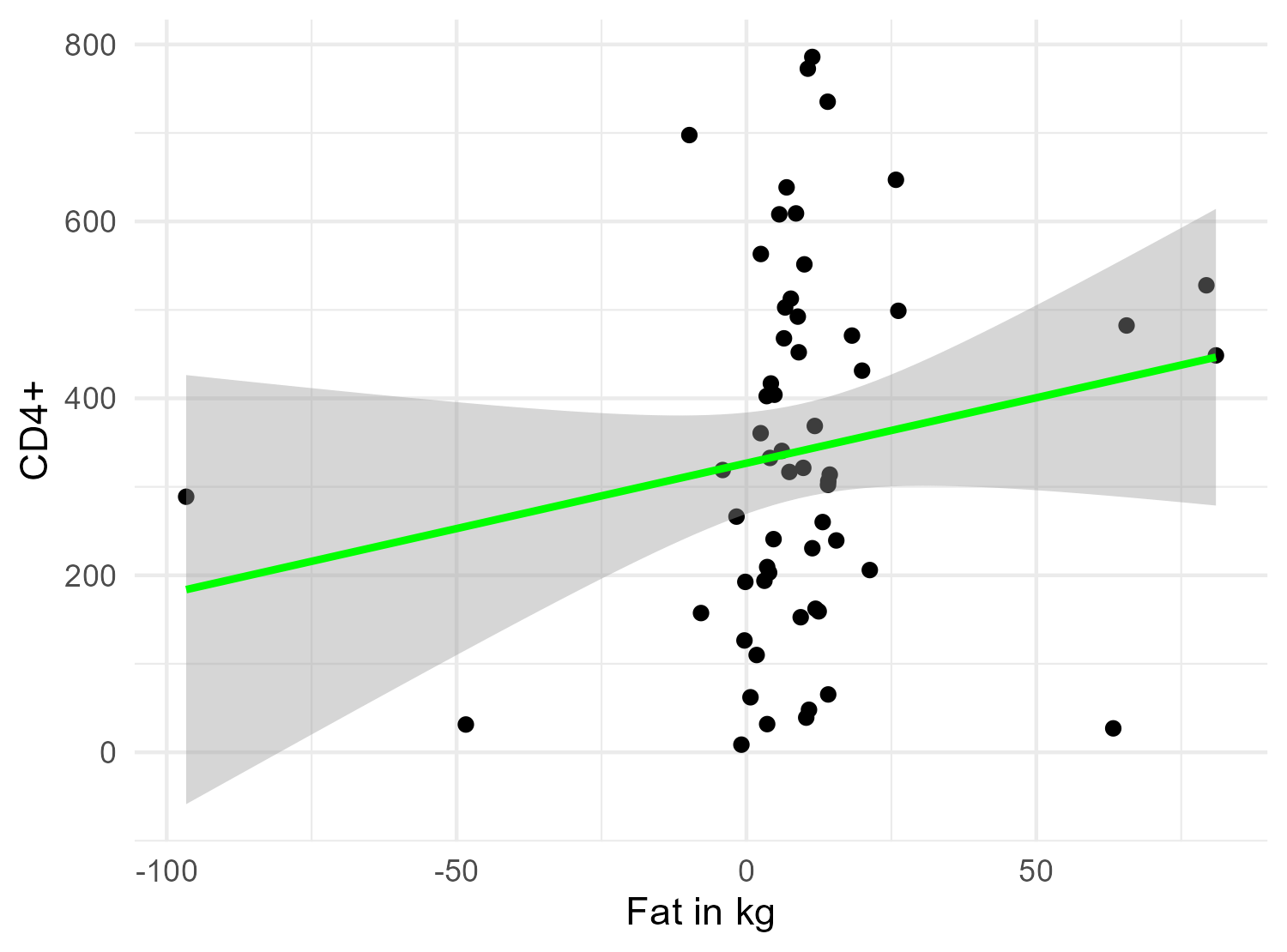
We used simple linear model to inspect whether there was an association between our main outcome CD4 immune activation count and the the primary predictor sex. To improve the model, we added in other predictors that could influence the immune activation count by performing a multi-variant linear regression anlaysis. This was done make sure we minimize confounding.

(**fig1-result?**) shows a scatter plot of lean body mass in kg and CD4



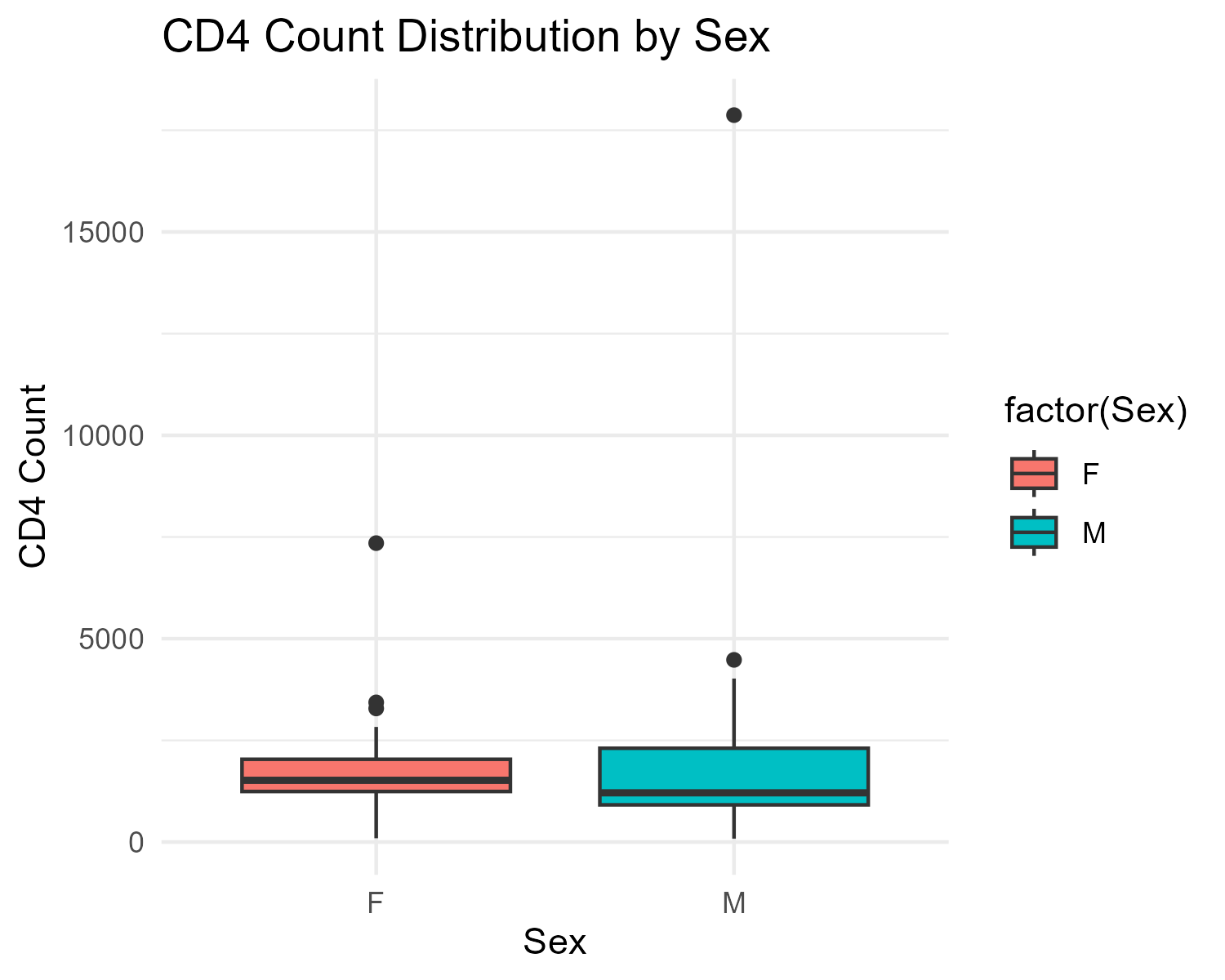
Scatter plot of CD4+ cells versus LBM in kg.

(**fig2-result?**) shows a scatter plot of CD4 vs Fat in kg.



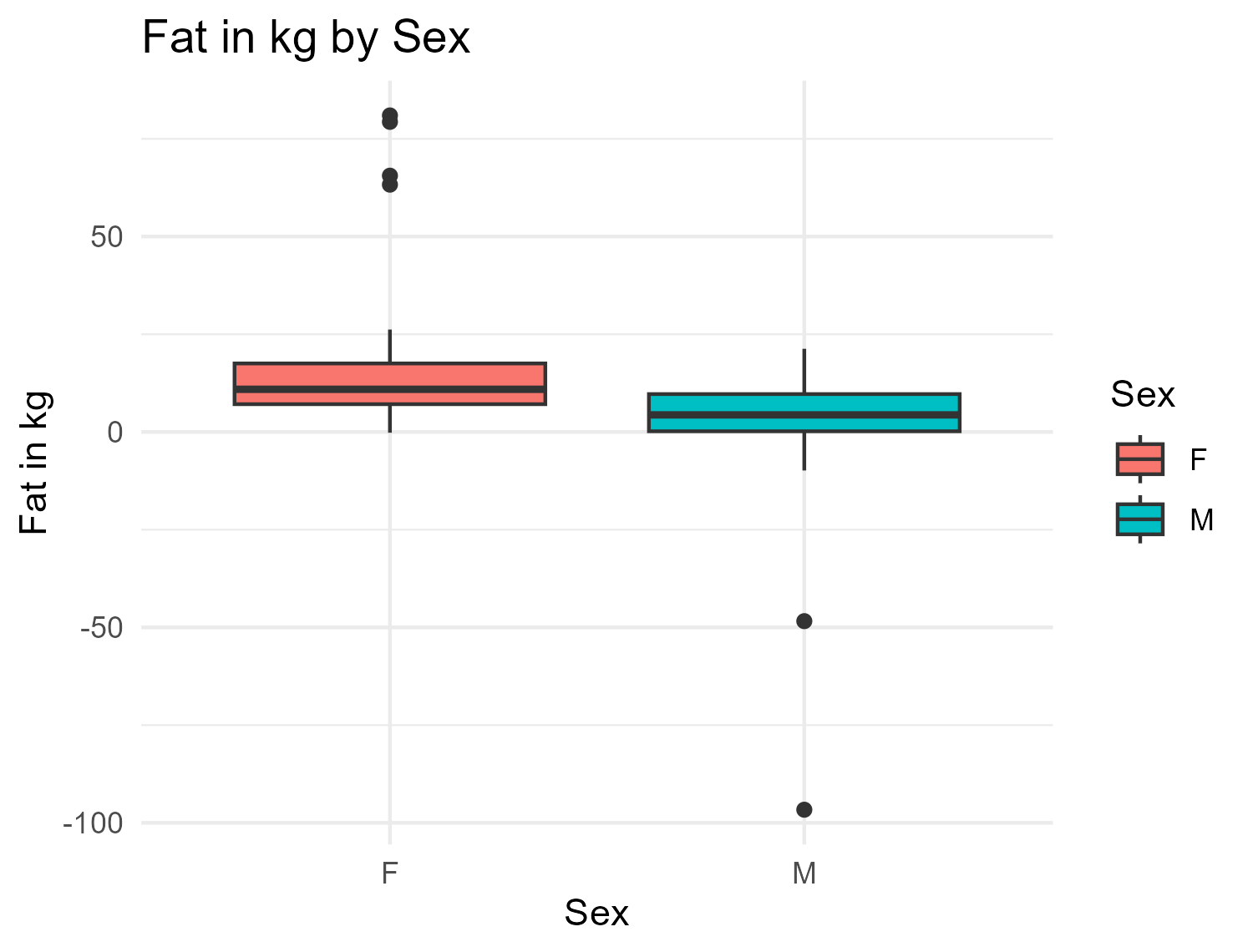
Scatter plot of CD4+ cells versus Fat in kg.

(**plot1-result?**) a boxplot of the cd4 immune activation count by sex.



Box plot to show the distribution of cd4 count immune activation counts among males and females.

(**plot2-result?**) a boxplot to show the distributions of fat in kilograms among males females.



Box plot showing Fat in kg among males and females.

## 4.3 Full analysis

In order to take care of confounding, we performed a multi-regression analysis and the results of the analysis are summarized in the table below.

This [Table 2](#tbl-model_table2) shows a summary of a multi-variant linear regression model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model summary table 2.   |  | Estimate | Std. Error | t value | Pr(>|t|) | | --- | --- | --- | --- | --- | | (Intercept) | 3759.632173 | 2567.133417 | 1.4645254 | 0.1490704 | | relevel(factor(Sex), ref = “F”)M | 928.102409 | 804.136468 | 1.1541603 | 0.2537114 | | Participant age | -6.818317 | 31.975088 | -0.2132384 | 0.8319748 | | BMI | -174.848153 | 177.081620 | -0.9873874 | 0.3280250 | | LBM in kg | 3.933569 | 60.703367 | 0.0647998 | 0.9485818 | | Fat in kg | 13.122848 | 58.955129 | 0.2225904 | 0.8247263 | | CD4+ | 3.215944 | 1.623752 | 1.9805632 | 0.0529418 | |

We went throught the steps of model building. First, we tested the null model which comprised of our main outcome and sex as the predictor. We did a stepwise selection using a foward direction approach with a our full model with all the interaction effects. We found that adding the interaction terms did not improve the model performance. We then considered the null model in comparison with the full model but this time round without interaction effects but all the predictors. Using the same procedure of stepwise selection using the foward direction chose the a simple linear model with just cd4+ as the only additional predictor. We performed a simple linear regression on the data between the CD4 immune activation count as the main outcome and sex plus cd4+ as predictors. We considered females to be our reference level.

This [Table 3](#tbl-model_table3) shows a summary of a final linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 3: Linear model fit table 3.   |  | Estimate | Std. Error | t value | Pr(>|t|) | | --- | --- | --- | --- | --- | | (Intercept) | 535.366751 | 765.700145 | 0.6991859 | 0.4873299 | | relevel(factor(Sex), ref = “F”)M | 735.337679 | 643.915291 | 1.1419789 | 0.2583223 | | CD4+ | 3.074691 | 1.583202 | 1.9420715 | 0.0571646 | |

# 5. Discussion

Our findings highlight a notable difference in CD4 immune activation count rates between males and females, with males exhibiting higher risk of TB infection and disease than females. The results show that men have a 735.34 increased immune activation count as compared to women. This result aligns with previous studies that have reported sex-based disparities in health outcomes and mortality risks as a result of TB(6),(7). Some studies still attribute the increased risk among males to be as a result of biological differences or differences in the accessibility of resources, especially healthcare. This is generally because women have better prognosis than men and they have better health seeking behavior as compared to their male counterparts.

Research shows that most women have higher fat repositories than men(8) and some studies show that the higher the fat percentage an individual has, the lower the risk of progression of the tuberculosis infection to disease hence lowering mortality rates among females compared to men. In addition, women tend to have more lean body mass compared to men hence having a lower chance of suffering from adverse effects of TB compared to men.

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

# 6. References

1. Santos M de LSG, Vendramini SHF, Gazetta CE, et al. Poverty: Socioeconomic characterization at tuberculosis. *Revista Latino-Americana de Enfermagem* [electronic article]. 2007;15(spe):762–767. (<http://dx.doi.org/10.1590/S0104-11692007000700008>)

2. Martinez L, Shen Y, Mupere E, et al. Transmission of mycobacterium tuberculosis in households and the community: A systematic review and meta-analysis. *American Journal of Epidemiology* [electronic article]. 2017;185(12):1327–1339. (<http://dx.doi.org/10.1093/aje/kwx025>)

3. Dye C, Garnett GP, Sleeman K, et al. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. *The Lancet* [electronic article]. 1998;352(9144):1886–1891. (<http://dx.doi.org/10.1016/S0140-6736(98)03199-7>)

4. Murgia Y, Sepulcri C, Crupi L, et al. Tuberculosis infection control: Experiences and considerations from a web based tool implementation. In: *pHealth 2024*. IOS Press; 2024(<http://dx.doi.org/10.3233/SHTI240053>)

5. Whalen C. Sex differences in the presentation of tuberculosis in uganda. 2023;(<https://dataverse.unc.edu/citation?persistentId=doi:10.15139/S3/AYOFEU>)

6. Hertz D, Schneider B. Sex differences in tuberculosis. *Seminars in Immunopathology* [electronic article]. 2018;41(2):225–237. (<http://dx.doi.org/10.1007/s00281-018-0725-6>)

7. Marçôa R, Ribeiro AI, Zão I, et al. Tuberculosis and gender – factors influencing the risk of tuberculosis among men and women by age group. *Pulmonology* [electronic article]. 2018;24(3):199–202. (<http://dx.doi.org/10.1016/j.pulmoe.2018.03.004>)

8. Karastergiou K, Smith SR, Greenberg AS, et al. Sex differences in human adipose tissues – the biology of pear shape. *Biology of Sex Differences* [electronic article]. 2012;3(1). (<http://dx.doi.org/10.1186/2042-6410-3-13>)