# Animal Health: WBC in Golden Retrievers

EEMB146 Final Project

Amy Kuang

## Abstract

The motivation for this analysis is to understand what kind of health and or environmental factors affect the health condition of pets: in particular, I want to better understand how golden retrievers, a specific breed of dogs, are affected by health and or environmental factors. The idea for this project was inspired from the parasites and viruses on frog examples in lecture. For this project, I examined a health dataset on golden retrievers with two numerical variables (white blood cell count and age) and two categorical variables (parasite status and type of area). The first statistical test conducted was a two-way ANOVA test on exploring if the categorical variables, parasite status and type of area, have a significant effect on the white blood cell count of golden retrievers. The results found that there is no significant interaction between parasite status and type of area on white blood cell count, but, individually, the variables parasite status and type of area were found to have significant effects on white blood count. The second statistical test I conducted was a linear regression to test if age and type of area were significant predictors for white blood cell count in golden retrievers. The results found that the variables age and type of area are significant predictors for white blood cell count in golden retrievers. My broad conclusions were that simple health and environmental factors are often easily overlooked despite playing a major role in pet health.

## Introduction

The data examined in this project is randomly sampled data from the Morris Animal Foundation Golden Retriever Life Time Study. The Morris Animal Foundation Golden Retriever Life Time Study contains data for over 3000+ Golden Retrievers across the continental United States ($n = 3018$) with 15 variables. While Golden Retrievers are one of the most popular dogs in the US as family pets, service dogs, search and rescue dogs, Golden Retrievers also have an estimated life span at around 10 to 12 years — previously, the life expectancy of golden retrievers were much higher, around 16 years (Gaeng 2022). With contracting parasites and gaining chronic diseases being not uncommon for dogs, this dataset focuses on Golden Retrievers. In order to further understand risk factors for diseases such as cancer in dogs, this dataset is one of the first observational investigations of its kind in attempts to further expand knowledge of genetic factors, health, and environmental variables that contribute to disease in veterinarian medicine (Simpson 2017). In a similar study on golden retrievers using the same dataset, Kubas and her team conducted a similar study on the prediction for positive or negative result for endoparasitism using various health variables such as age, blood count, serum biochemistry, and fecal data through logistic regression (Kubas 2022). In this project, I hope to perform a simpler study to examine the effects of select key variables I believe to have a significant role on Golden Retrievers' health. In my project, after data wrangling, the dataset was reduced to a sample size of $n = 2996$ with 4 variables: age, parasite status, and type of area as factors that affect white blood cell count. I chose white blood cell count as an 'informal metric' for judging health since high white blood cell count is noted to be an indicator of infection or illness in dogs. For testing, I chose to perform two tests. The first test was a two-way ANOVA test to explore if the categorical variables, parasite status and type of area (individually), have a significant effect on the white blood cell count of golden retrievers. The second test was a linear regression to test if age and type of area were significant predictors for white blood cell count in golden retrievers.

## Exploratory Data Analysis

To visualize the data meaningfully and explore the relationships between the variables, I utilized one histogram, one scatterplot, and two boxplots. After cleaning the data, my variables are: white blood cell count of golden retrievers, parasite status (indicates if the golden retriever tested negative or positive), type of area the golden retriever is from (urban, suburban, or rural areas), and age.

**Histogram** To see the distribution of the data on white blood cell count of golden retrievers, I created a histogram. According to the histogram below, the data appears to be normal since it follows a bell-shape curve. While there is also a noticeable right skew, the data can still be considered to be normally distributed per the Central Limit Theorem. According to the Central Limit Theorem, non-normally distributed data can be approximated to follow a normal distribution when a sample size gets large. Since the sample size of the data is very large at $n = 2996$ observations, the data can be considered to follow a normal distribution. There also appears to be some outliers on the right tail, but they are not concerning.

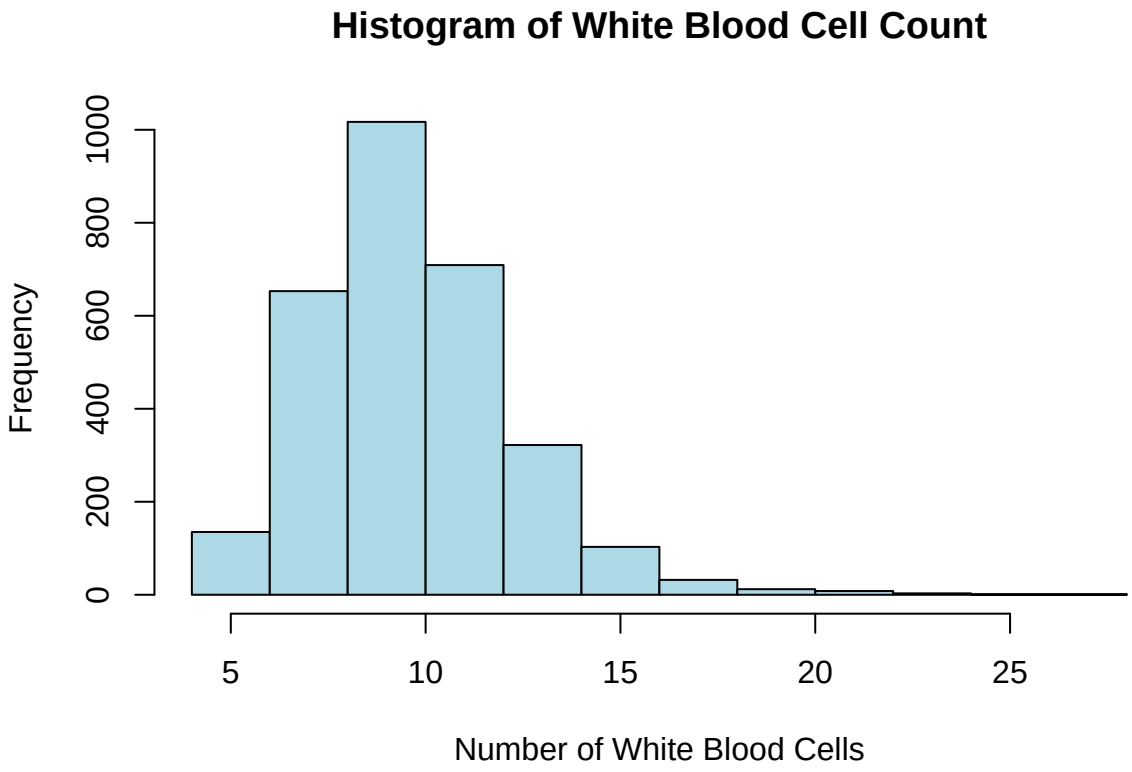



Figure 1: Figure 1. Histogram of Golden Retrievers' white blood cell count data. The data follows a bell-shaped curve that is relatively normal, but there is a skew to the right. However, since the sample size is large (n = 2996), normality can be assumed by the Central Limit Theorem.

**Scatterplot** To illustrate the age variable in the data, I created a scatterplot for white blood cell count of golden retrievers' data against age. According to the plot, there is a negative linear relationship between the number of white blood cells and age. There also appears to be some outlier points, but they are not significantly concerning to need removal.

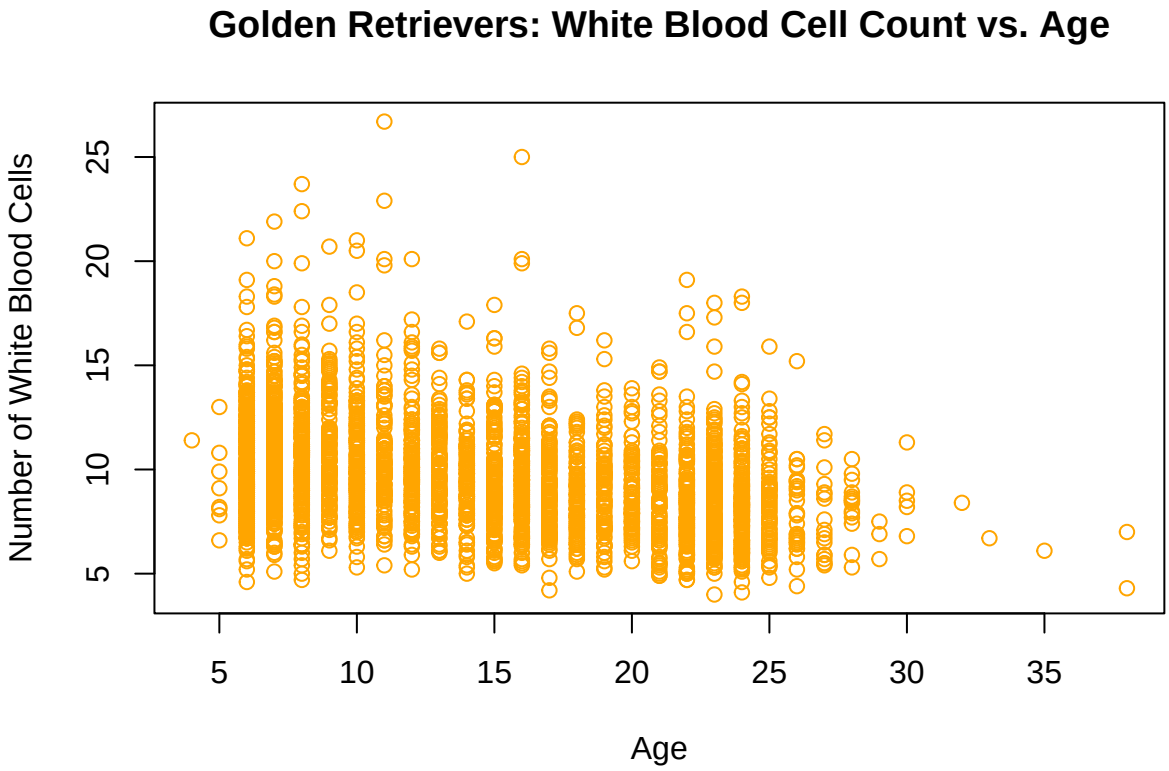



Figure 2: Figure 2. Scatterplot of golden retrievers' white blood cell count by age. There appears to be a negative linear relationship.

**Boxplots** To visualize the number of white blood cells in golden retrievers by their parasite status, I created a boxplot to compare the differences between the white blood cell count of golden retrievers with positive parasite status and the white blood cell count of golden retrievers with negative parasite status. According to the boxplot below, the mean number of white blood cells for golden retrievers that tested positive for parasites are shown to be slightly higher than the mean of number of white blood cells for golden retrievers that tested negative. Additionally, there are lots of noticeable outliers in the data that corresponds to negative parasite status compared to the positive parasite status, however, both status' outliers are not too concerning given that the sample size is significantly large ($n = 2996$).

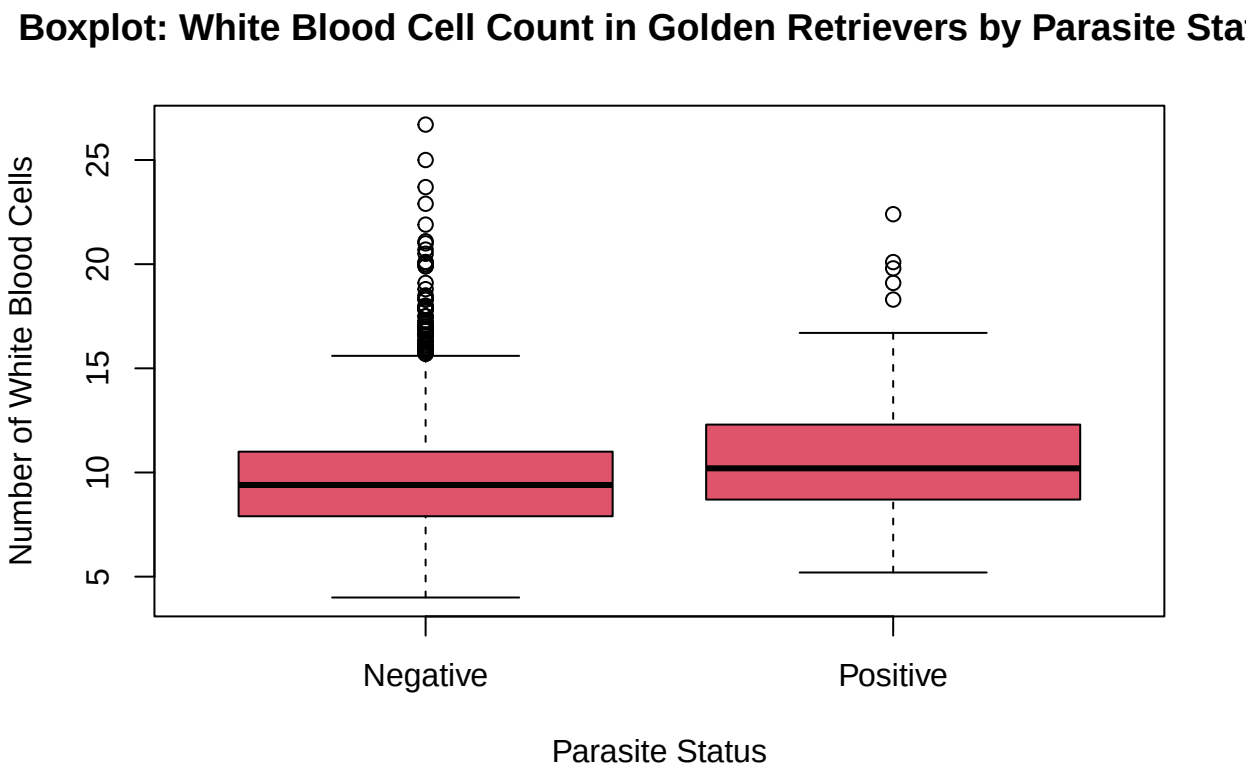



Figure 3: Figure 3. Boxplot of golden retrievers' white blood cell count data by Parasite Status. Golden retriever data with a negative parasite status (meaning no parasites) appear to have the most outliers, but this should not be concerning since the sample size is very large (n=2996).

Next, to visualize the number of white blood cells in golden retrievers by the area they are from, I created a boxplot to compare the differences between the white blood cell count of golden retrievers from urban areas, the white blood cell count of golden retrievers from suburban areas, and the white blood cell count of golden retrievers from rural areas. According to the boxplot below, the mean number of white blood cells for golden retrievers from rural areas are shown to be the highest compared to the mean of number of white blood cells for golden retrievers from urban and suburban areas. Additionally, there are lots of noticeable outliers for all three types of areas, however, these outliers are also not too concerning given that the sample size is significantly large ($n = 2996$).

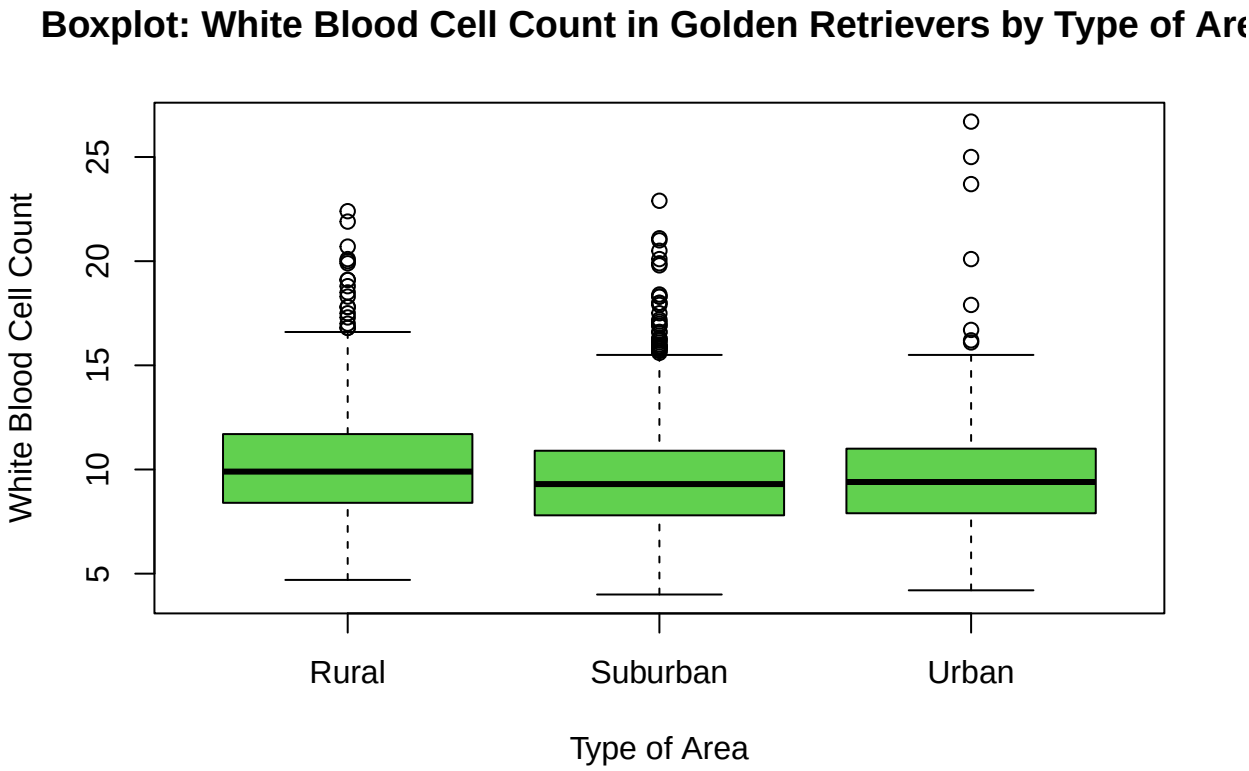



Figure 4: Figure 4. Boxplot of white blood cell count of golden retrievers by type of area. There are many outliers in each category, but this should not be concerning since the sample size is 2996.

## Statistical Methods

For my statistical tests, I will conduct a two-way ANOVA test for test 1 and a linear regression test for test 2.

My motivation for a two-way ANOVA test comes seeking to understand if both categorical variables, parasite status and a type of area, individually have a significant effect on the number of white blood cells golden retrievers. From Figure 3, the boxplot indicates that golden retrievers that tested positive for parasites have a higher mean white blood cell count compared to those that tested negative. From Figure 4, the boxplot indicates that golden retrievers in rural areas have a higher mean white blood cell count compared to other locations respectively. By testing individually, it would allow us to eliminate external factors such as wealth or level of grooming that can be afforded (since an interaction could potentially suggest that higher white blood cell count could be attributed by rural environments, affordability, or expenditure power and such).

My motivation for a linear regression test is to test if type of area and age are significant predictors for white blood cell count of golden retrievers. From Figure 4, it is rational to wonder if type of area is a significant predictor for white blood cell count of golden retrievers since golden retrievers in rural areas appear to have a higher mean of white blood cell count compared to the other areas. At the same time, from Figure 2, there is an evident negative relationship between white blood cell count and age in golden retrievers. With these two considerations, a linear regression test can help determine if these two predictors are significant.

**ANOVA** An ANOVA test is a statistical test that tells us how one or more categorical variables can affect the mean value of a quantitative variable. The result of an ANOVA test can tell us if the categorical variables have a significant effect on the quantitative variable.

In this project, I will use a two-way ANOVA test to test if parasite status and type of area have significant effects on white blood cell count in golden retrievers. Parasite status is a categorical and independent variable with two levels: positive and negative. Type of area is a categorical and independent variable with three levels: rural, suburban, and urban. White blood cell is a numerical and dependent variable.

To conduct an ANOVA test, the following three assumptions must be met: the data is from a random sample, the residuals are normally distributed, and variance of the residuals are equal (equality of variance). Residuals are measure of the difference between a predicted value and observed value, and variance refers to the spread of the data (i.e., spread of the data points if the data was plotted).

It is important to note that the white blood cell count of golden retrievers is considered normal by the Central Limit Theorem.

The random sample assumption is met since the data is a randomly sampling of data from across the continental US.

For the normality of residuals assumption, a "Normal Q-Q" plot can be used to check. If the residuals follows a linear trend line in "Normal Q-Q" plot, then the residuals are normal. In this case, the normality of residuals assumption is met as the residuals follow a linear trend line (Figure 5). Even though there are some noticeable outliers present, the residuals are still assumed to be normal since the residuals are predominately following a linear trend and since the sample size is very large ($n = 2996$) – recall the Central Limit Theorem applies to the data!

For the equality of variance for the residuals assumption, the "Residuals vs Fitted" plot can be used to check if residuals' variances are equal. If the red line in the "Residuals vs Fitted" plot is flat with no distinct pattern and the residuals are falling along the zero line, then the variance of the residuals are equal. In this case, the equality of variance is met as the red line is flat with no distinct pattern (Figure 5). The residuals are also shown to be equidistant from the red line.

Since the ANOVA assumptions are met, I hypothesize that parasite status and type of area have a significant effect on white blood cell count of golden retrievers. In other words, my null hypothesis is that type of area and parasite status (individually) have no significant effect on white blood cell count of golden retrievers; my alternate hypothesis is that type of area and parasite status (individually) do have a significant effect

on white blood cell count of golden retrievers. Using an $\alpha = 0.05$ significance level, if the p-value of my ANOVA test is greater than 0.05, then I fail to reject my null hypothesis. If the p-value is less than 0.05, then I reject my null hypothesis.

**Linear Regression** A linear regression is a linear model typically used to make predictions.

In this project, I will use a linear regression to test if type of area and age are significant in predicting white blood cell count of golden retrievers. Type of area is a categorical and independent variable with three levels: rural, suburban, and urban. Age is a numerical and independent variable. White blood cell count is a numerical and dependent variable.

To conduct a linear regression, the following three assumptions must be met: the data comes from a random sample, the response variable Y must be normally distributed with equal variance, and the residuals follow a normal distribution.

Since the data is a randomly sampling of data from across the continental US, the random sample assumption is met.

Recall, the white blood cell count data of golden retrievers is normal by Central Limit Theorem! Thus, the normality assumption for response variable Y in this case is also met.

According to the "Residuals vs. Fitted" plot, the equality of variance is also met since the red line is flat with no particular trend and the data points are equidistant from the red line (Figure 7).

According to the "Normal Q-Q" plot, the normality of residuals is also met since the residuals follow a linear trend line (Figure 7). Even though there are some noticeable outliers present, the residuals are still assumed to be normal since the residuals are predominately following a linear trend and since the sample size is very large ($n = 2996$) – recall the Central Limit Theorem applies to the data!

Since the linear regression assumptions are met, I hypothesize that type of area and age are significant predictors for white blood cell count of golden retrievers. In other words, my null hypothesis is that type of area and age (individually) are not significant predictors of white blood cell count of golden retrievers; my alternate hypothesis is that type of area and age (individually) are significant predictors of white blood cell count of golden retrievers. Using an $\alpha = 0.05$ significance level, if the p-value of my linear regression is greater than 0.05, then I fail to reject my null hypothesis. If the p-value is less than 0.05, then I reject my null hypothesis.

## Results

**ANOVA** For my two-way ANOVA test, my null hypothesis is that type of area and parasite status (individually) do not have a significant effect on white blood cell count of golden retrievers. My alternate hypothesis is that type of area and parasite status (individually) do have a significant effect on white blood cell count of golden retrievers. Using an $\alpha = 0.05$ significance level, if the p-value of my ANOVA test is greater than 0.05, then I fail to reject my null hypothesis. If the p-value is less than 0.05, then I reject my null hypothesis.

In the two-way ANOVA test, the results found that there is no significant interaction between parasite status and type of area on white blood cell count (p-value = $0.51 > 0.05$ significance level), but, individually, the variables parasite status (0.00000531) and type of area (p-value = $0.000000000207 < 0.05$ significance level) are found to have significant effects on white blood count.

Since the p-value = 0.00000531 corresponding to parasite status is less than the 0.05 significance level, I reject my null hypothesis that parasite status does not significant effect on white blood cell count in golden retrievers and conclude that parasite status does have a significant effect on white blood cell count in golden retrievers.

Since the p-value = 0.000000000207 corresponding to type of area is less than the 0.05 significance level, I reject my null hypothesis that type of area does not significant effect on white blood cell count in golden

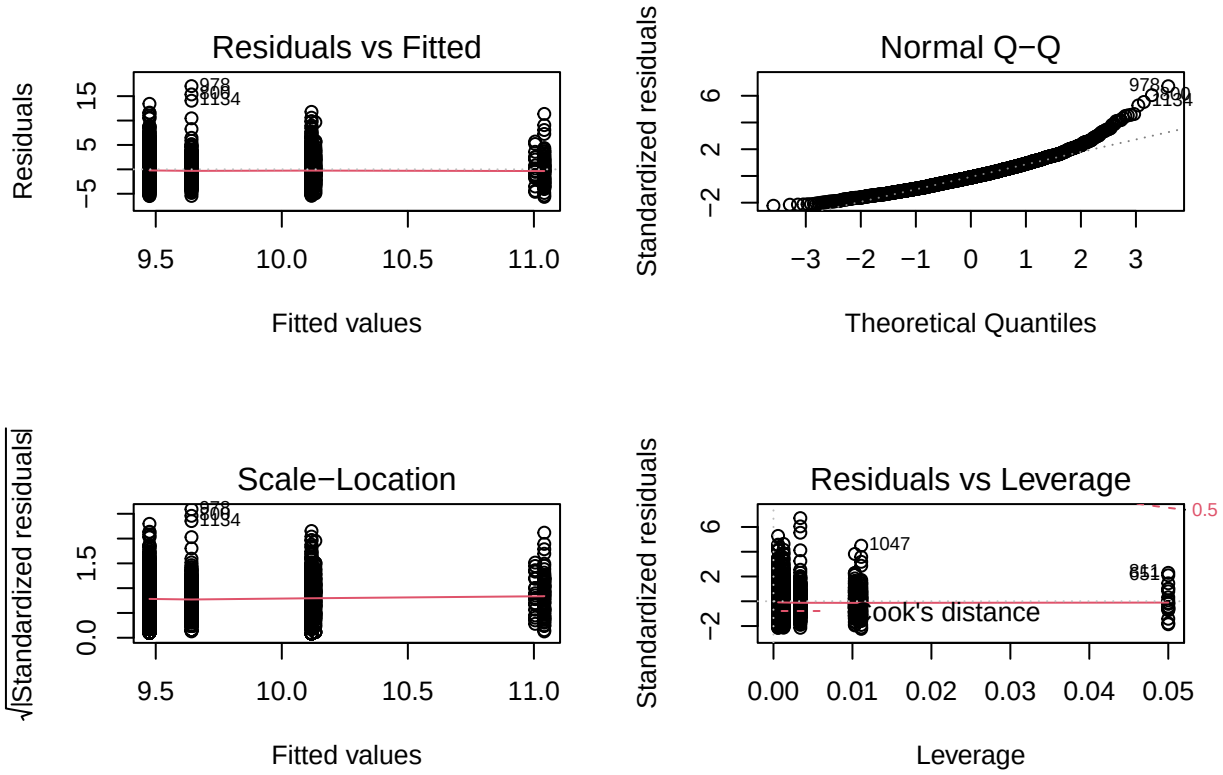



Figure 5: Figure 5. Diagnostic plots for ANOVA. The Residuals vs Fitted plot is shown to indicate equality of variance since the red line is flat with no particular pattern and the residuals look equidistant from the red line. The Normal Q-Q plot appears to show the residuals following a linear trend. While there are outliers present, normality of residuals can still be assumed since the residuals predominately follow a linear trend and because the sample size is very large, at n=2996.

retrievers and conclude that type of area does have a significant effect on white blood cell count in golden retrievers.

Additionally, since the p-value = 0.51 corresponds to the interaction between parasite status and type of area is greater than the 0.05 significance level, I would fail to reject a corresponding null hypothesis that states that the interaction between parasite status and type of area does not have a significant effect on white blood cell count.

Biologically speaking, the results find that type of area and parasite status individually does have a significant effect on white blood cell count in golden retrievers.

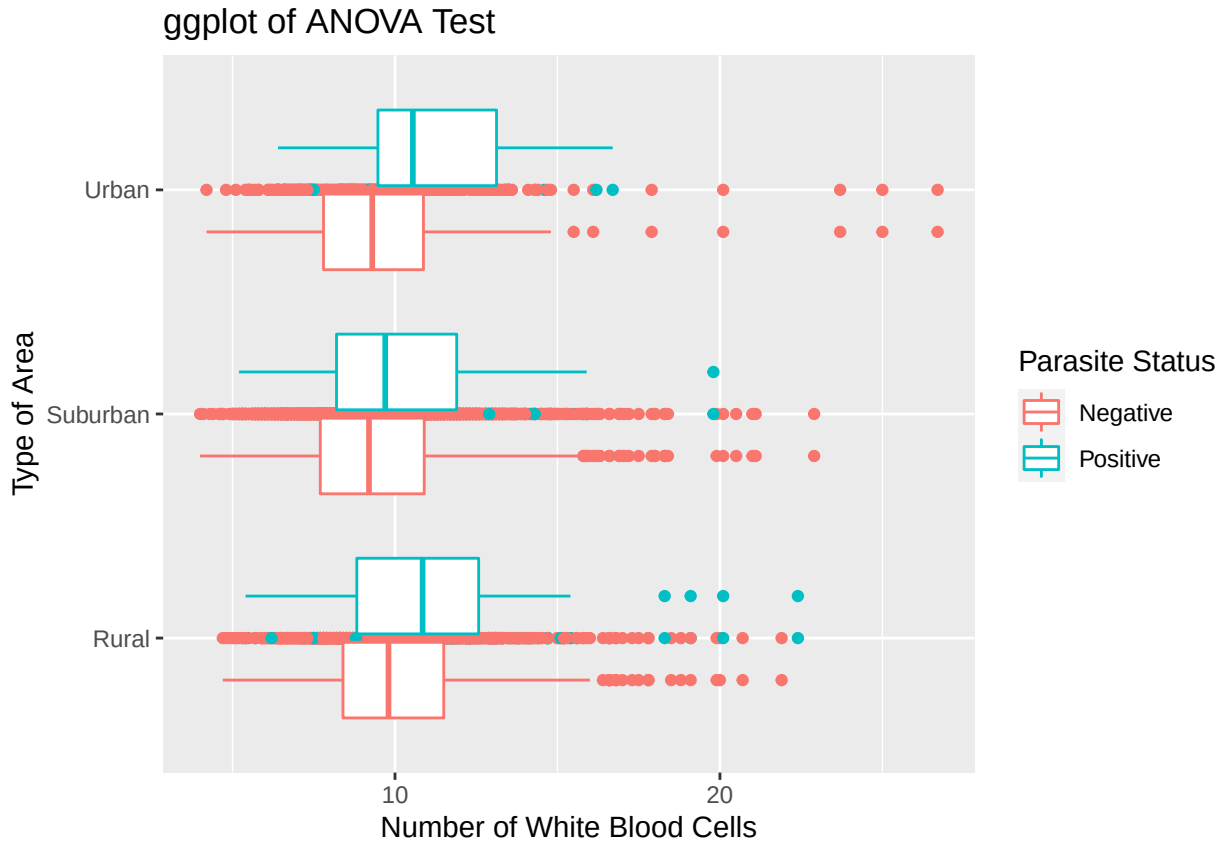



Figure 6: Figure 6. GGPLOT of white blood cell count data by type of area they live in with their parasite status. The ggplot shows that golden retrievers in rural areas have more amount of instances of testing positive for parasites compared to those in suburban and urban areas. The ggplot also indicates that golden retrievers in rural areas appear to have more white blood cell counts on average compared to those in suburban and urban areas.

**Linear Regression** For the linear regression test, my null hypothesis is that type of area and age (individually) are not significant predictors of white blood cell count of golden retrievers; my alternate hypothesis is that type of area and age (individually) are significant predictors of white blood cell count of golden retrievers. Using an $\alpha = 0.05$ significance level, if the p-value of my linear regression is greater than 0.05, then I fail to reject my null hypothesis. If the p-value is less than 0.05, then I reject my null hypothesis.

In the linear regression test, the results found that there age and all 3 levels of type of area are significant predictors for white blood cell count.

For age, since the p-value = 0.0000000000000002 is less than the 0.05 significance level, I reject my null

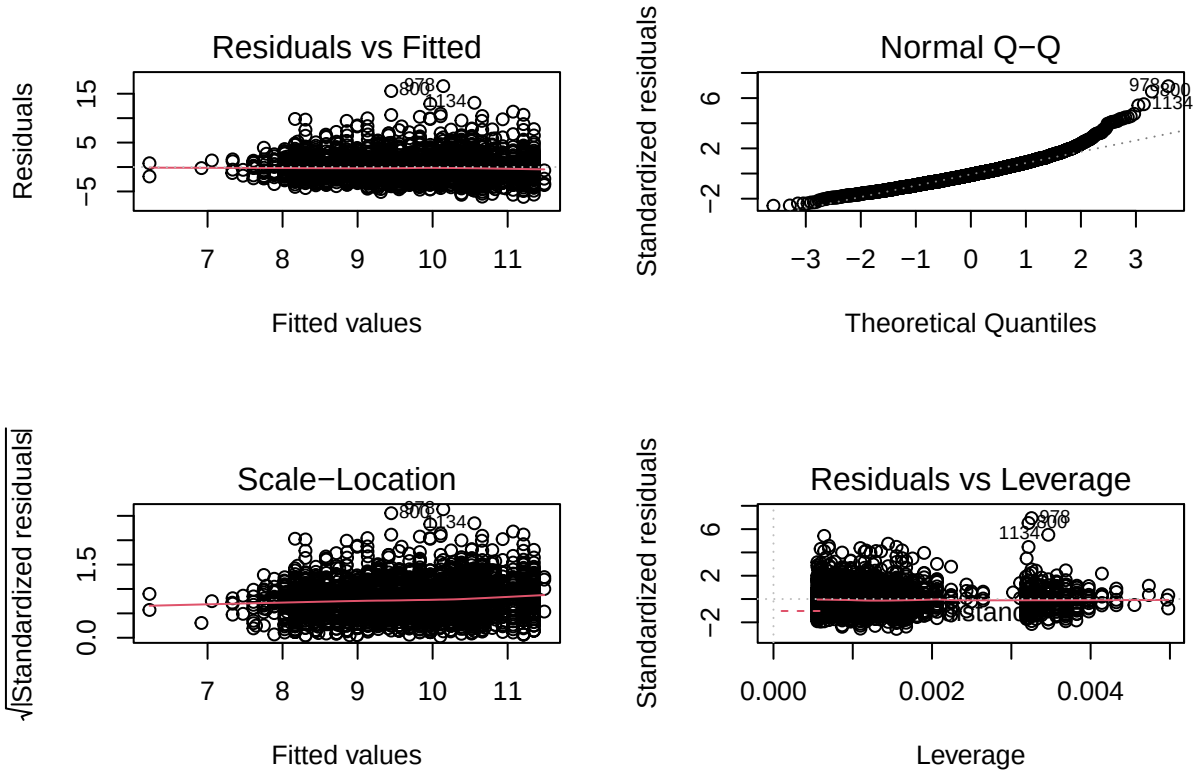



Figure 7: Figure 7. Diagnostic plots for linear regression. The Residuals vs Fitted plot is shown to indicate equality of variance since the red line is flat with no particular pattern and the residuals are equidistant from the red line. The Normal Q-Q plot appears to show the residuals following a linear trend with some outliers present. However, normality of residuals can still be assumed since the residuals predominately follow a linear trend and since the sample size is very large, n=2996.

hypothesis that age is not a significant predictor for white blood cell count and conclude that age is a significant predictor for white blood cell count in golden retrievers.

For type of area-rural, since the p-value = 0.0000000000000002 is less than the 0.05 significance level, I reject my null hypothesis that type of area (factor: rural) is not a significant predictor for white blood cell count and conclude that rural area is a significant predictor for white blood cell count in golden retrievers.

For type of area-suburban, since the p-value = 0.00000000000325 is less than the 0.05 significance level, I reject my null hypothesis that type of area (factor: suburban) is not a significant predictor for white blood cell count and conclude that rural area is a significant predictor for white blood cell count in golden retrievers.

For type of area-urban, since the p-value = 0.00106 is less than the 0.05 significance level, I reject my null hypothesis that type of area (factor: urban) is not a significant predictor for white blood cell count and conclude that urban area is a significant predictor for white blood cell count in golden retrievers.

Biologically speaking, the results found that there age and type of area (rural, suburban, and urban areas) are significant predictors for white blood cell count of golden retrievers.

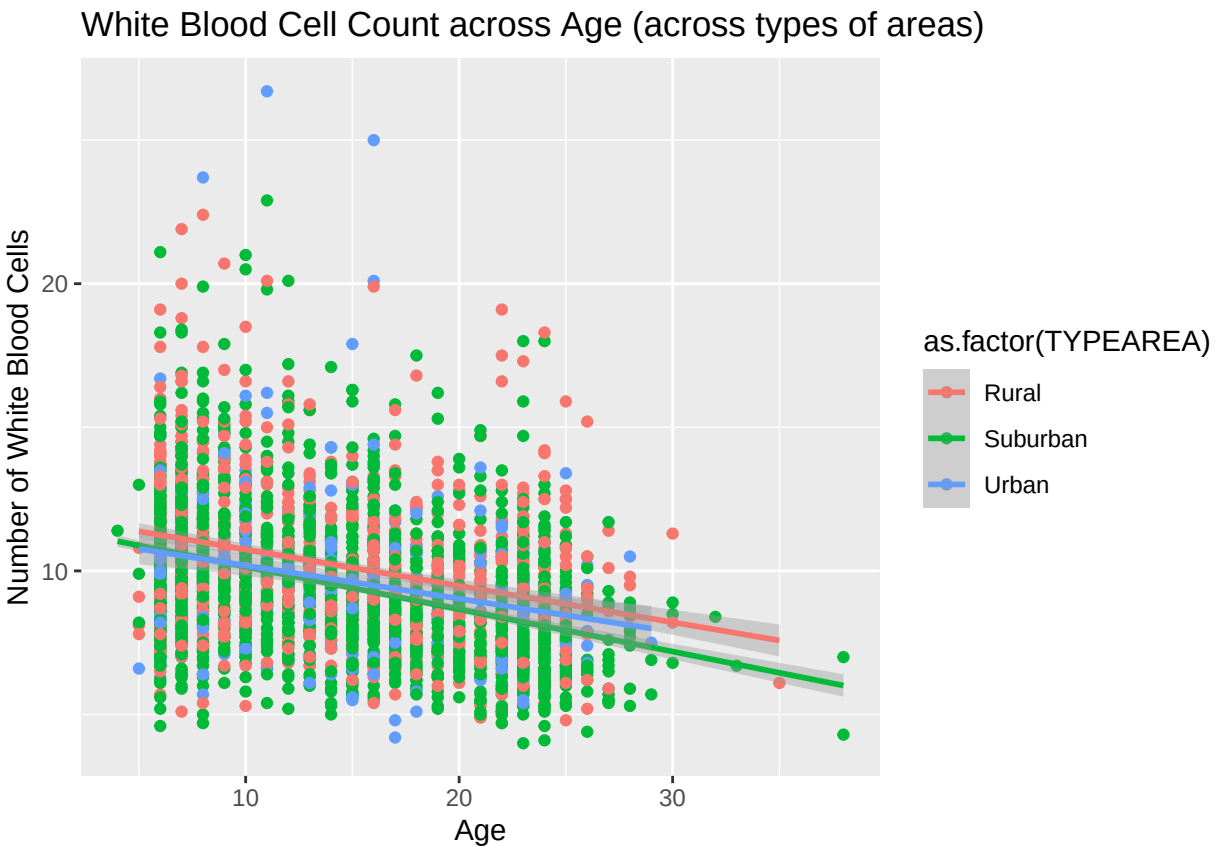



Figure 8: Figure 8. Scatterplot with linear regression lines corresponding to white blood cell count against age and type of area (shown by color). According to the scatterplot, age appears to have a negative relationship with white blood cell count. Per the colors that represent a different level for the type of area variable, golden retrievers in rural areas appear to have comparably higher predictions of white blood cell count compared to the other levels with urban following with the second highest predictions and then suburban areas with the third highest predictions of the three levels.

## Discussion:

After some data wrangling, the original dataset became a sample size of $n = 2996$ with 4 variables: age, parasite status, and type of area as factors that affect white blood cell count; I chose as white blood cell count as a dependent variable an 'informal metric' for judging health in golden retrievers. Since high white blood cell count is noted to be an indicator of infection or illness in dogs, I chose white blood cell count as a health metric where high white blood cell count was a indicator of possible disease or illness and lower or average white blood cell count was considered good or regular health.

For testing, I chose to perform two tests: 1) a two-way ANOVA test to explore if the categorical variables, parasite status and type of area, have a significant effect on the white blood cell count of golden retrievers and 2) a linear regression to test if age and type of area were significant predictors for white blood cell count in golden retrievers.

In the two-way ANOVA test results, while there is no interaction at play with parasite status and type of area, the individual variables demonstrating significant effects on white blood cell count implicitly suggests that bodily-cleanliness and or environmental cleanliness potentially play a large role in infection and illness in pets. In other words, cleaner environments may attribute to less illness and chances of infections. Broadly speaking, the results suggest that the chances of disease and illness for dogs, not from genetic factors, may be significantly reduced if their external environment is clean.

Regarding the linear regression test results, the idea of cleanliness relating to less illness is reinforced through type of area being indicated to be a significant predictor variable for white blood cell count – in particular, there is a negative relationship.

Given the nature of my sample, the amount of variables (4), and simplicity of my variables, there are definitely limitations to my analysis. For example, my analysis does not take into account for different breeds of dogs, their current physical condition (e.g., disabled or not), and their genetics causing them to be more susceptible to infection, and disease. However, with the current (whether pre-cleaned or post-cleaned) dataset, accounting for these factors may not be possible without more types of variables.

Regarding possible sources of bias or dependence, I do not believe that there is a large chance for bias. However, it is worth considering that data collection date could be a potential factor for bias in that life expectancy or rates of parasitism in dogs went up or down over the years.

To improve my study, I would want to possibly take more numeric variables as independent variables and predictors and possibly utilize a categorical variable as a dependent response variable.

If given more time and data, I would want to analyze what factors play a role in life expectancy in dogs since dogs tend to live around 2 decades at most depending on breed. By understanding factors for life expectancy better, dogs could potentially live longer! Overall, I believe the final the final takeaway from this analysis is a reaffirmation that maintaining cleanliness for dogs (and all pets) appears to be the best way to reduce chances of infection, illness, and disease caused by non-genetic or preexisting health conditions.

## References


Gaeng, J. (2022, February 23). Golden Retriever lifespan: How long do golden retrievers live? Retrieved June 8, 2022, from https://a-z-animals.com/blog/golden-retriever-lifespan-how-long-do-golden-retrievers-live/

Guy, M. K., Page, R. L., Jensen, W. A., Olson, P. N., Haworth, J. D., Searfoss, E. E., & Brown, D. E. (2015). The Golden Retriever Lifetime Study: establishing an observational cohort study with translational relevance for human health. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1673), 20140230. https://doi.org/10.1098/rstb.2014.0230

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham, Jim Hester and Jennifer Bryan (2022). readr: Read Rectangular Text Data. R package version 2.1.2. https://CRAN.R-project.org/package=readr

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Kirill Müller (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here

Kubas EA, Fischer JR, Hales EN (2022) Endoparasitism of Golden Retrievers: Prevalence, risk factors, and associated clinicopathologic changes. PLOS ONE 17(2): e0263517. https://doi.org/10.1371/journal.pone.0263517

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-153, <URL: https://CRAN.R-project.org/package=nlme>.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W. (2022) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

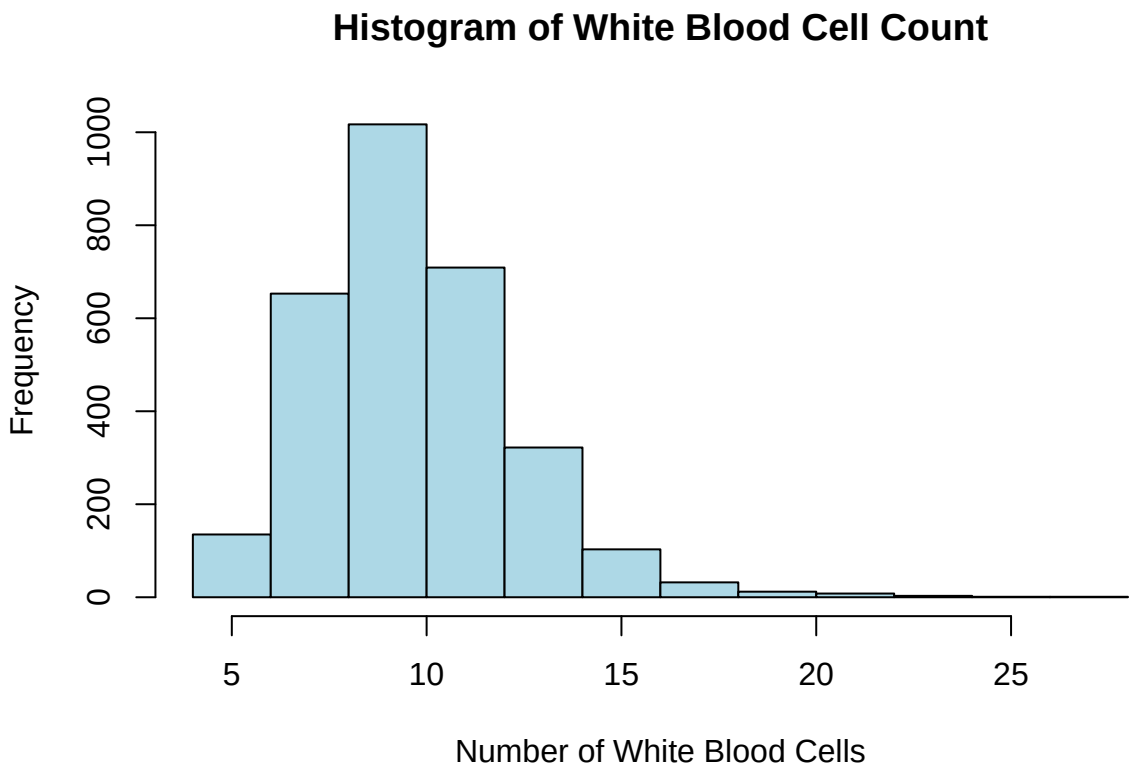
Yihui Xie (2022). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.38.

Simpson, M., Searfoss, E., Albright, S., Brown, D. E., Wolfe, B., Clark, N. K., McCann, S. E., Haworth, D., Guy, M., & Page, R. (2017). Population characteristics of golden retriever lifetime study enrollees. Canine genetics and epidemiology, 4, 14. https://doi.org/10.1186/s40575-017-0053-5

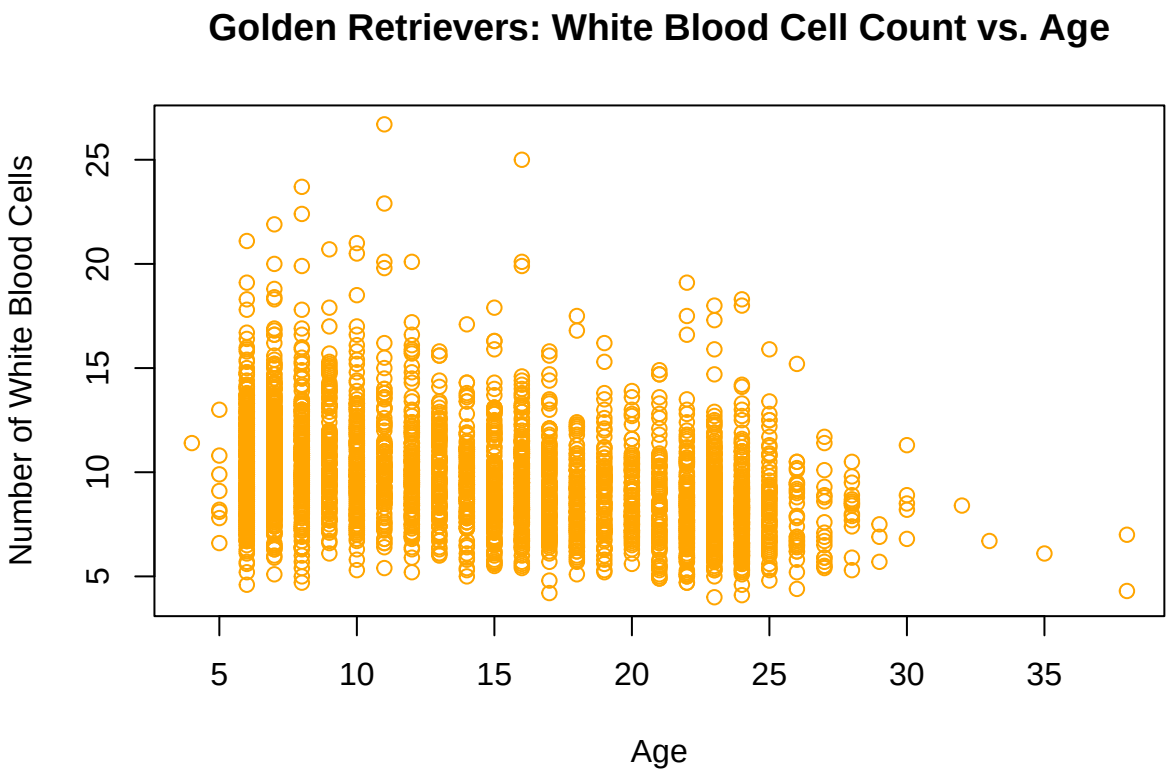
Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

## Appendix

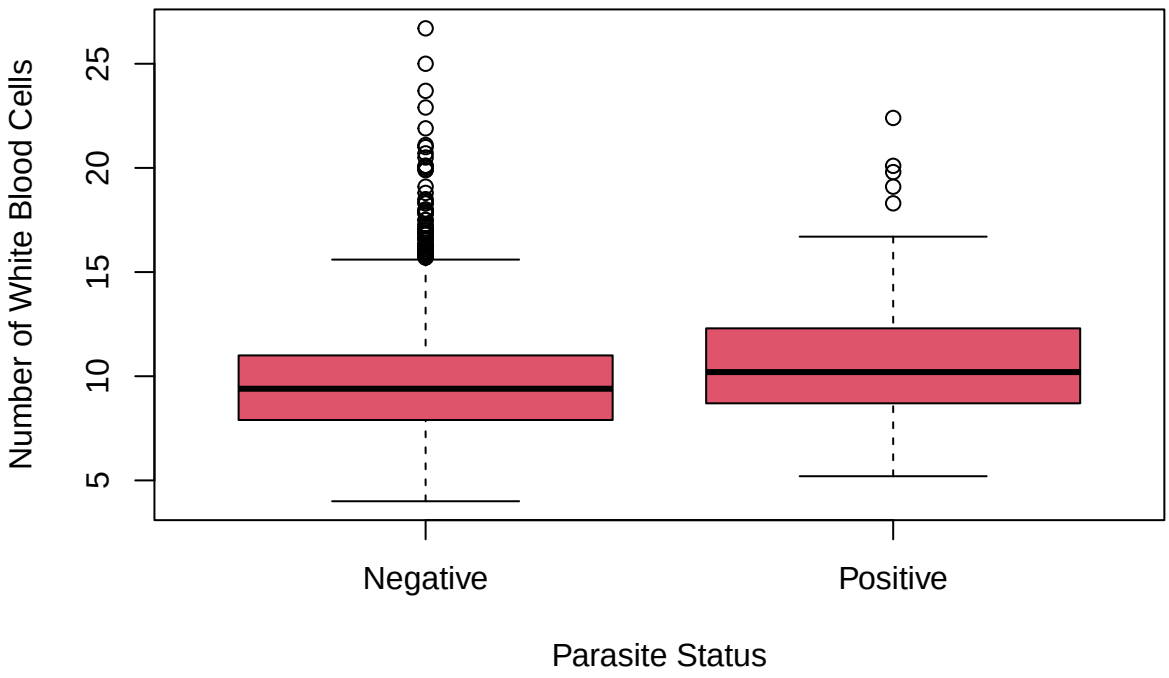
```
# Packages used
library(readr)
library(tidyverse)
library(here)
library(psych)
library(car)
library(stats)
library(ggplot2)
library(knitr)
library(nlme)

## Exploratory Data Analysis
# load in data
gr <- read_csv("golden_retriever_data.csv", show_col_types = FALSE)
# clean data
gr_clean <- gr %>% select(c(WBC, PARASITE_STATUS, TYPEAREA, AGE)) # select variables
gr_clean <- na.omit(gr_clean) # get rid of missing values

# Visualization 1: Histogram of WBC
hist(gr_clean$WBC, xlab = "Number of White Blood Cells", main="Histogram of White Blood Cell Count", co
```



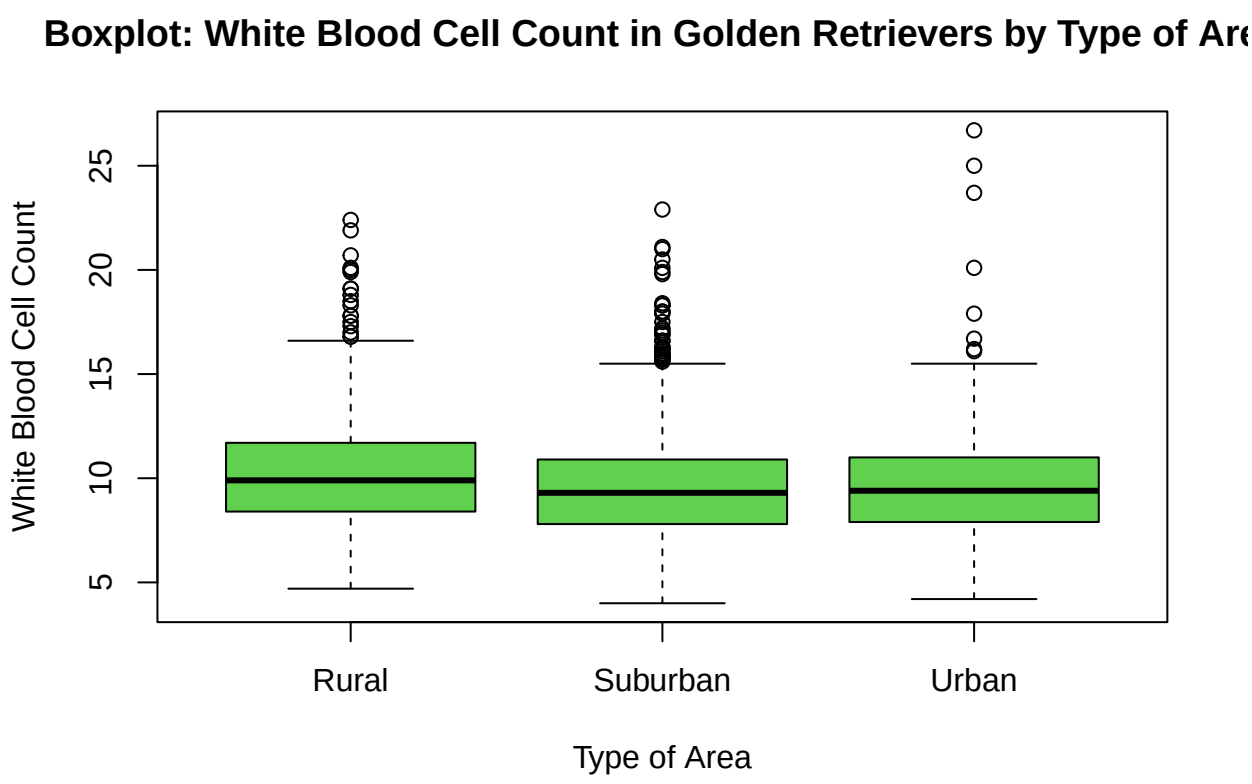

```
# Visualization 2: Scatterplot WBC~AGE
plot(gr_clean$WBC~gr_clean$AGE, xlab = "Age", ylab = "Number of White Blood Cells", main = "Golden Retri
```




```
# Visualization 3: Boxplot WBC~PARASITE_STATUS
boxplot(gr_clean$WBC~gr_clean$PARASITE_STATUS, xlab = "Parasite Status", ylab = "Number of White Blood C
```



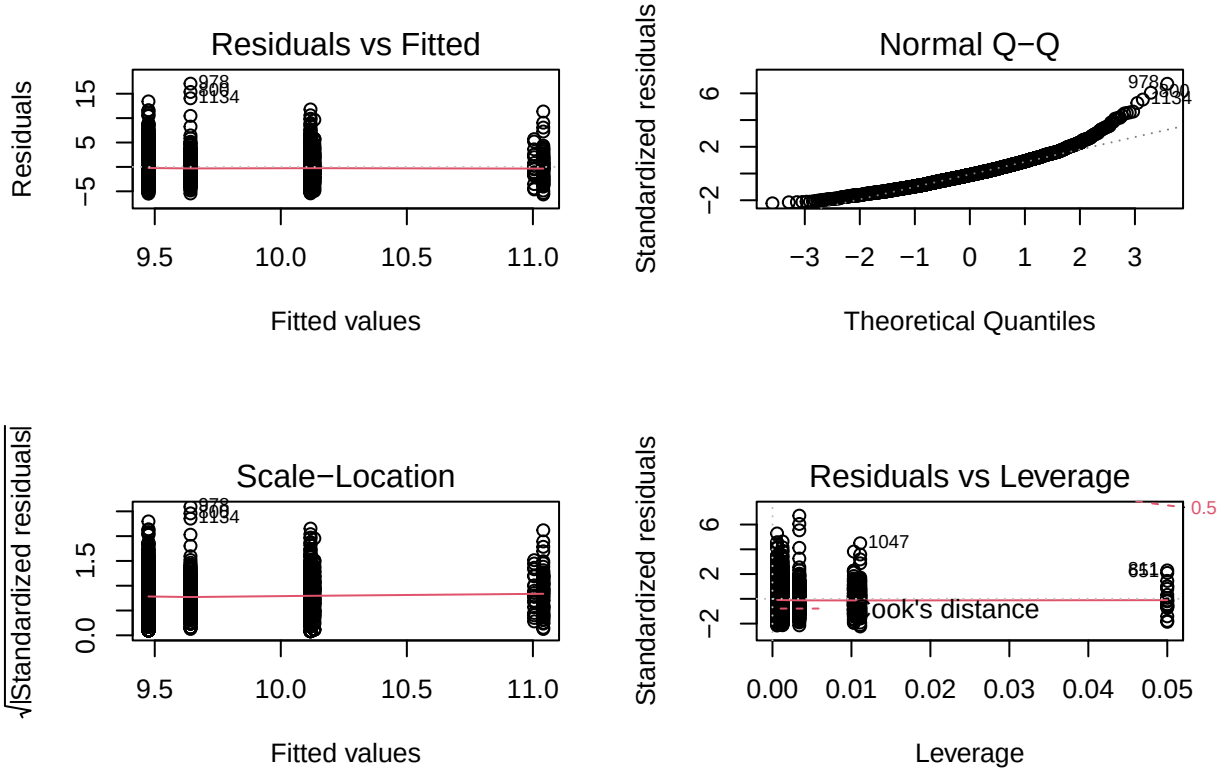

```
# Visualization 4: WBC~TYPEAREA
boxplot(gr_clean$WBC~gr_clean$TYPEAREA, xlab = "Type of Area", ylab = "White Blood Cell Count", main ="
```



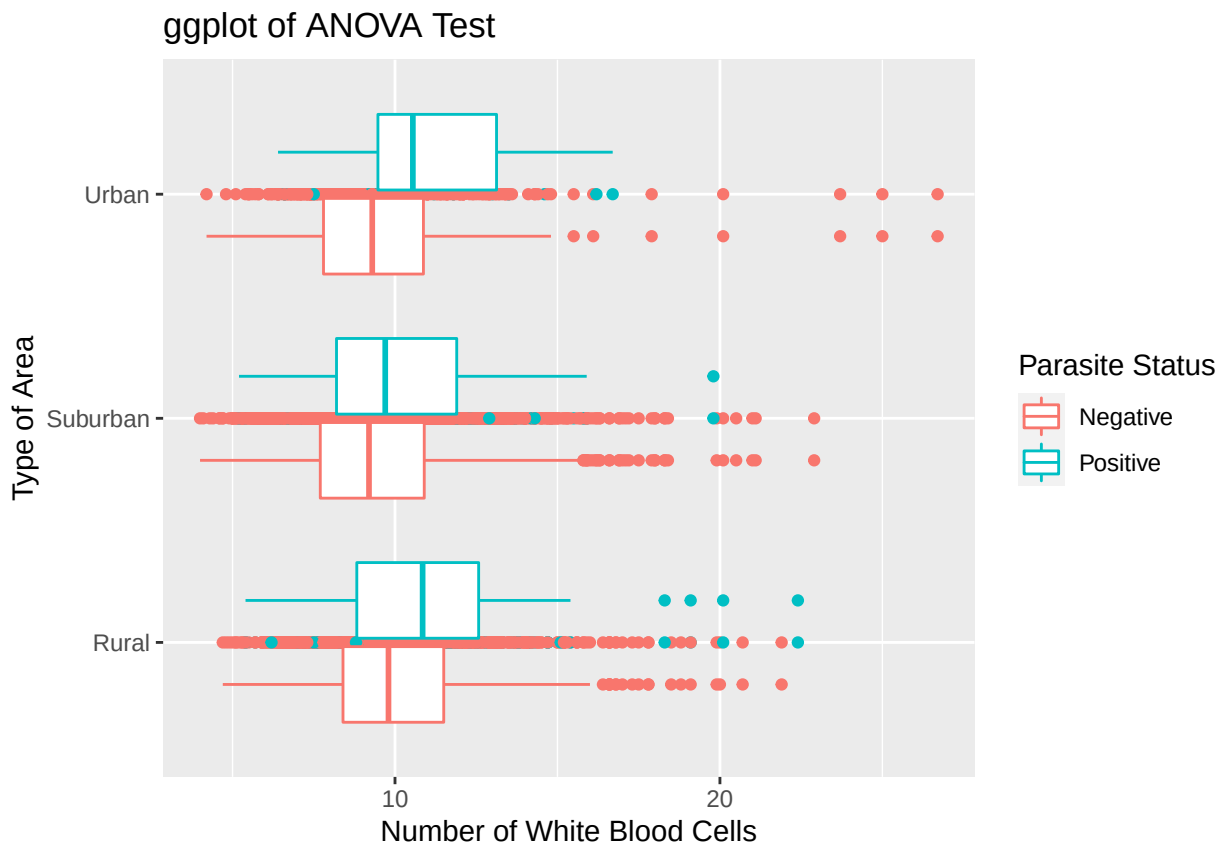

```
#### ANOVA
wbc_anova <- aov(WBC~TYPEAREA*PARASITE_STATUS, data=gr_clean) # fit ANOVA
summary(wbc_anova) # summary of ANOVA object
```

```
##                              Df Sum Sq Mean Sq F value   Pr(>F)
## TYPEAREA                      2    291  145.44  22.468 2.07e-10 ***
## PARASITE_STATUS               1    135  134.63  20.797 5.31e-06 ***
## TYPEAREA:PARASITE_STATUS      2      9    4.36   0.674     0.51
## Residuals                  2990  19356    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA diagnostics
par(mfrow=c(2,2)) # edit output view
plot(wbc_anova) # diagnostic plots
```



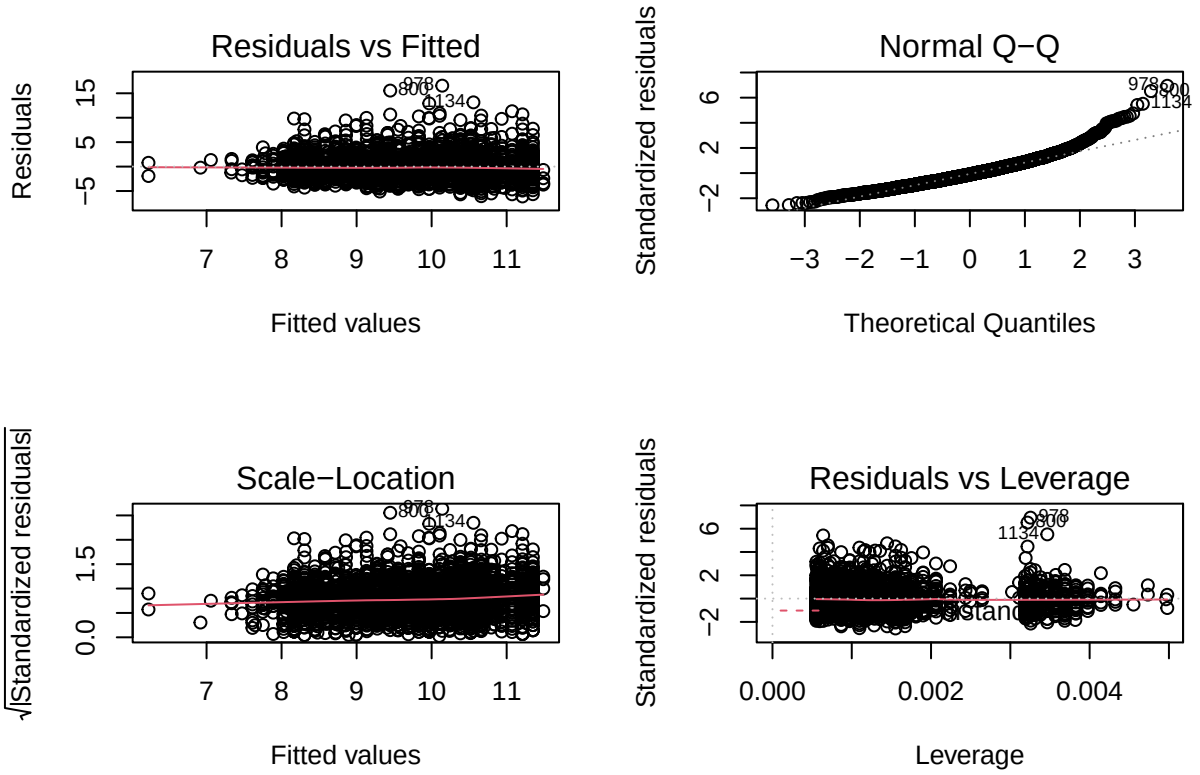

```
# ANOVA data plot
ggplot(gr_clean, aes(x=WBC, y=TYPEAREA, color= PARASITE_STATUS))+ geom_point() + geom_boxplot() +
labs(title="ggplot of ANOVA Test", x="Number of White Blood Cells", y="Type of Area", col="Parasite Sta
```



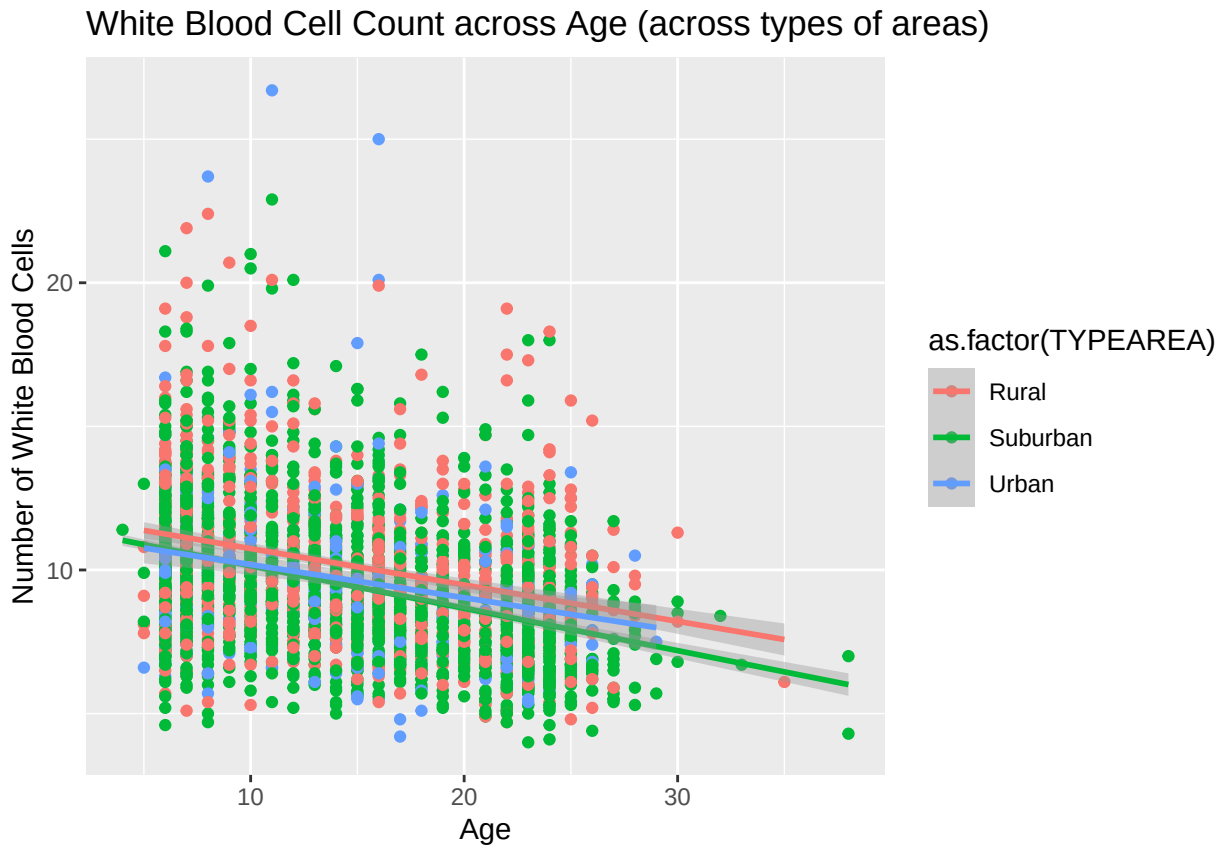

```
#### Linear Regression
# Fitting a linear regression
mod <- lm(WBC~AGE+TYPEAREA, data=gr_clean) # fitting model
summary(mod) # summary of lm object
```

```
##
## Call:
## lm(formula = WBC ~ AGE + TYPEAREA, data = gr_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1108 -1.6050 -0.2446  1.2935 16.5586
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      12.181147   0.124818  97.591  < 2e-16 ***
## AGE              -0.138616   0.006704 -20.676  < 2e-16 ***
## TYPEAREASuburban -0.688005   0.098351  -6.995 3.25e-12 ***
## TYPEAREAUrban    -0.514951   0.157096  -3.278  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.388 on 2992 degrees of freedom
## Multiple R-squared:  0.1379, Adjusted R-squared:  0.137
## F-statistic: 159.5 on 3 and 2992 DF,  p-value: < 2.2e-16
```

```
# lm diagnostics
par(mfrow=c(2,2)) # edit output view
plot(mod) # diagnostic plots
```




```
# Plot LM regression model
par(mfrow=c(1,1)) # reset output view

# Linear regression line on scatterplot
ggplot(gr_clean, aes(x=AGE, y=WBC, color=as.factor(TYPEAREA))) + geom_point() + geom_smooth(method = "lm
```




```
## References
citation("readr")
```

```
##
## To cite package 'readr' in publications use:
##
##   Hadley Wickham, Jim Hester and Jennifer Bryan (2022). readr: Read
##   Rectangular Text Data. R package version 2.1.2.
##   https://CRAN.R-project.org/package=readr
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {readr: Read Rectangular Text Data},
##     author = {Hadley Wickham and Jim Hester and Jennifer Bryan},
##     year = {2022},
##     note = {R package version 2.1.2},
##     url = {https://CRAN.R-project.org/package=readr},
##   }
```

```
citation("tidyverse")
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
```

```
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostin
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }
```

```
citation("here")
```

```
##
## To cite package 'here' in publications use:
##
##   Kirill Müller (2020). here: A Simpler Way to Find Your Files. R
##   package version 1.0.1. https://CRAN.R-project.org/package=here
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {here: A Simpler Way to Find Your Files},
##     author = {Kirill Müller},
##     year = {2020},
##     note = {R package version 1.0.1},
##     url = {https://CRAN.R-project.org/package=here},
##   }
```

```
citation("psych")
```

```
##
## To cite the psych package in publications use:
##
##   Revelle, W. (2022) psych: Procedures for Personality and
##   Psychological Research, Northwestern University, Evanston, Illinois,
##   USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {psych: Procedures for Psychological, Psychometric, and Personality Research},
##     author = {William Revelle},
##     organization = { Northwestern University},
##     address = { Evanston, Illinois},
##     year = {2022},
##     note = {R package version 2.2.5},
##     url = {https://CRAN.R-project.org/package=psych},
##   }
```

```
citation("car")
```

```
##
## To cite the car package in publications use:
##
##   John Fox and Sanford Weisberg (2019). An {R} Companion to Applied
##   Regression, Third Edition. Thousand Oaks CA: Sage. URL:
##   https://socialsciences.mcmaster.ca/jfox/Books/Companion/
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     title = {An {R} Companion to Applied Regression},
##     edition = {Third},
##     author = {John Fox and Sanford Weisberg},
##     year = {2019},
##     publisher = {Sage},
##     address = {Thousand Oaks {CA}},
##     url = {https://socialsciences.mcmaster.ca/jfox/Books/Companion/},
##   }
```

```
citation("stats")
```

```
##
## The 'stats' package is part of R.  To cite R in publications use:
##
##   R Core Team (2021). R: A language and environment for statistical
##   computing. R Foundation for Statistical Computing, Vienna, Austria.
##   URL https://www.R-project.org/.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2021},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

```
citation("ggplot2")
```

```
##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
```

```
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```

```
citation("knitr")
```

```
##
## To cite the 'knitr' package in publications use:
##
##   Yihui Xie (2022). knitr: A General-Purpose Package for Dynamic Report
##   Generation in R. R package version 1.38.
##
##   Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
##   Chapman and Hall/CRC. ISBN 978-1498716963
##
##   Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
##   Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
##   Peng, editors, Implementing Reproducible Computational Research.
##   Chapman and Hall/CRC. ISBN 978-1466561595
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

```
citation("nlme")
```

```
##
## Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). _nlme:
## Linear and Nonlinear Mixed Effects Models_. R package version 3.1-153,
## <URL: https://CRAN.R-project.org/package=nlme>.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {{nlme}: Linear and Nonlinear Mixed Effects Models},
##     author = {Jose Pinheiro and Douglas Bates and Saikat DebRoy and Deepayan Sarkar and {R Core Team}
##     year = {2021},
##     note = {R package version 3.1-153},
##     url = {https://CRAN.R-project.org/package=nlme},
##   }
```