Immerse Interview Prep

Fredrick Kakembo

2022-05-7

Loading the data

```
library(ggplot2)
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
```

```
data <- read.csv("applicant_dataset_v2.csv")
```

Check the size of the data

```
dim(data)
```

```
## [1] 1100    6
```

```
colnames(data)
```

```
## [1] "id"         "agecat"     "siteid"     "BSdich"     "LAMP"
## [6] "anymalaria"
```

```
#Deal with the categorical data to factors
data$agecat <- as.factor(data$agecat)
data$siteid <- as.factor(data$siteid)
data$BSdich <- as.factor(data$BSdich)
data$LAMP <- as.factor(data$LAMP)
data$anymalaria <- as.factor(data$anymalaria)
```

```
attach(data)
```

```
head(data)
```

```
##     id               agecat siteid BSdich LAMP anymalaria
## 1 1001          >= 18 years  Jinja      0 <NA>          0
## 2 1002 5 years - < 11 years  Jinja      0    0          0
## 3 1003 5 years - < 11 years  Jinja      0    0          0
## 4 1004 5 years - < 11 years  Jinja      0    0          0
## 5 1005          >= 18 years  Jinja      0 <NA>          0
## 6 1006 5 years - < 11 years  Jinja      0 <NA>          0
```

```
# id column

#data$id
length(data$id)
```

```
## [1] 1100
```

```
length(unique(data$id))
```

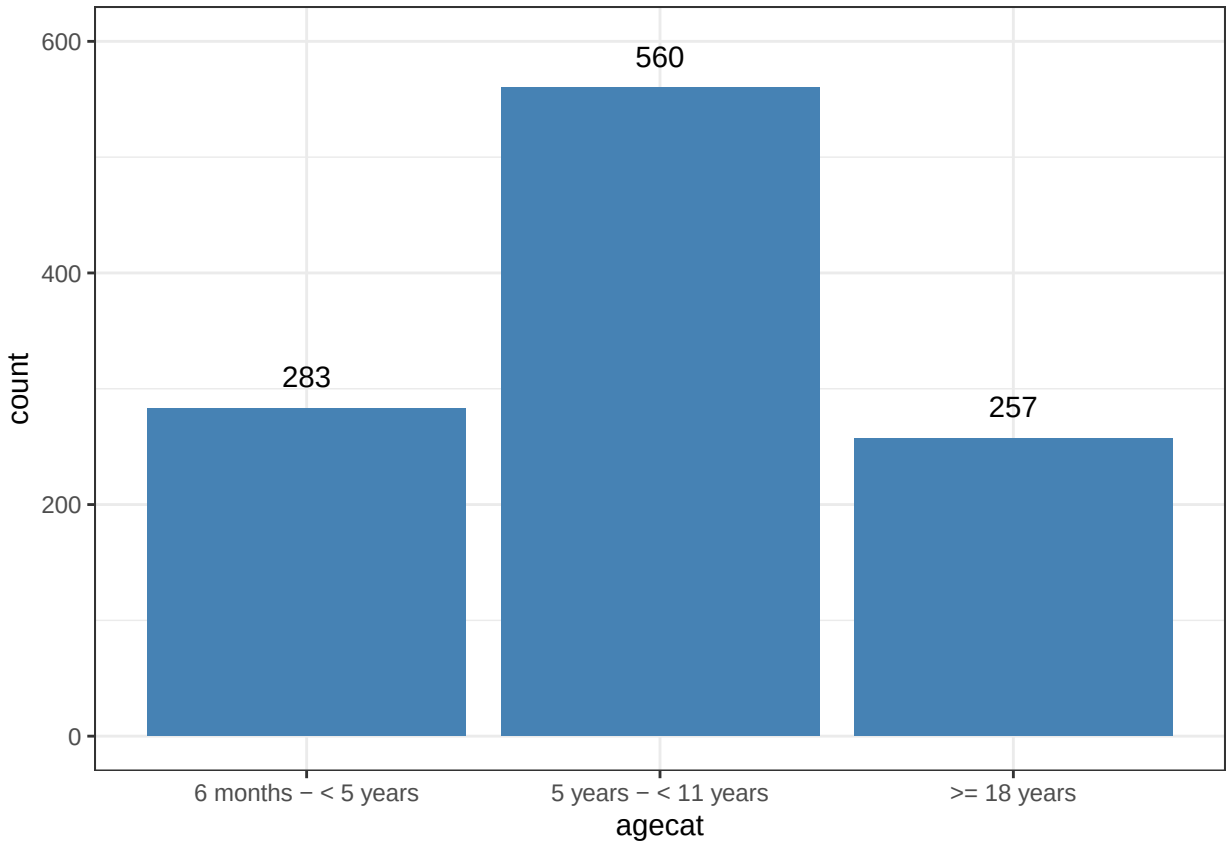
```
## [1] 1100
```

1. Age Category

```
table(agecat)
```

```
## agecat
##          >= 18 years 5 years - < 11 years 6 months - < 5 years
##                  257                  560                  283
```

```
p <- ggplot(data = data, aes(agecat)) +
  geom_bar(fill="steelblue") + ylim(c(0,600))

#Reordering the categories to start with 6months
p + scale_x_discrete(limits = c("6 months - < 5 years","5 years - < 11 years" ,">= 18 years")) +
  geom_text(stat='count', aes(label=..count..), vjust=-1)+
  theme_bw()
```



2. Site ID

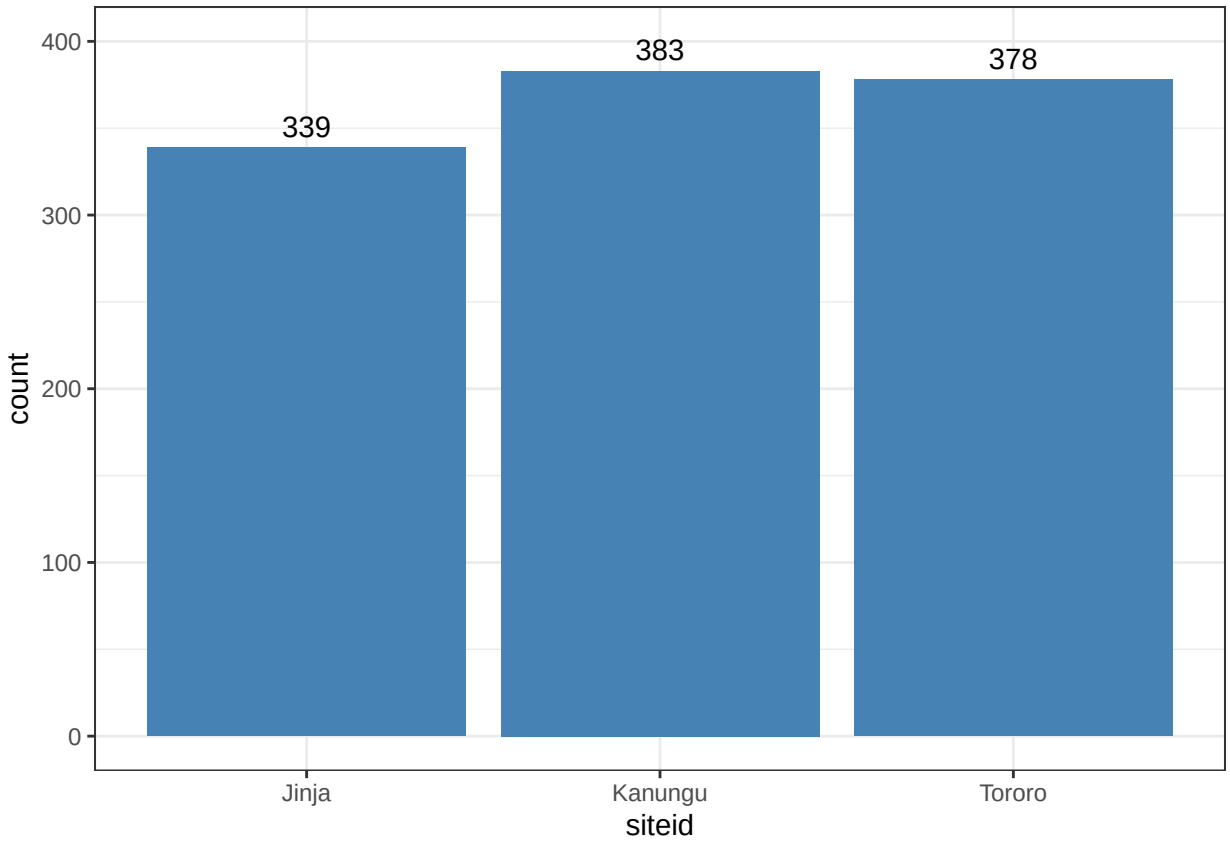
```
#Site
table(siteid)
```

```
## siteid
##   Jinja Kanungu  Tororo
##     339     383     378
```

```
sum(is.na(siteid))  #None missing
```

```
## [1] 0
```

```
ggplot(data = data, aes(siteid)) + geom_bar(fill="steelblue") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5)+
  theme_bw()+ ylim(c(0,400))
```



3. blood smear result

- 1 is positive, 0 is negative

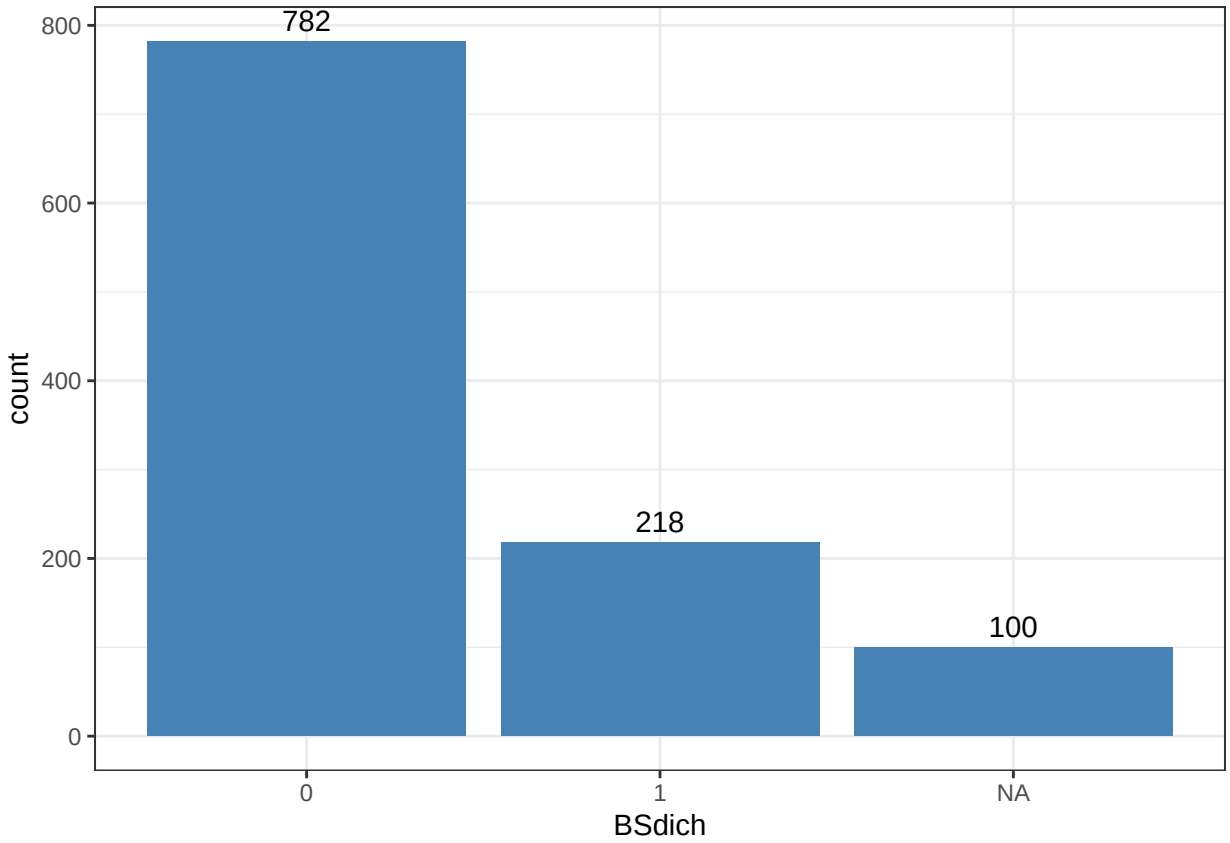
```
table(BSdich) # 1 is positive and 0 is negative
```

```
## BSdich
##   0   1
## 782 218
```

```
sum(is.na(BSdich)) #Number of missing
```

```
## [1] 100
```

```
ggplot(data = data, aes(BSdich)) + geom_bar(fill="steelblue") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) +
  theme_bw() #+ ylim(c(0,400))
```



4. LAMP

loop-mediated isothermal amplification result for submicroscopic P falciparum parasitemia. LAMP was only performed if blood smear was negative. 1 is positive, 0 is negative

```
library(dplyr)
#LAMP
data_0 <- subset(data, BSdich == 0) #Subset data for only when BS is neg
dim(data_0)
```

```
## [1] 782   6
```

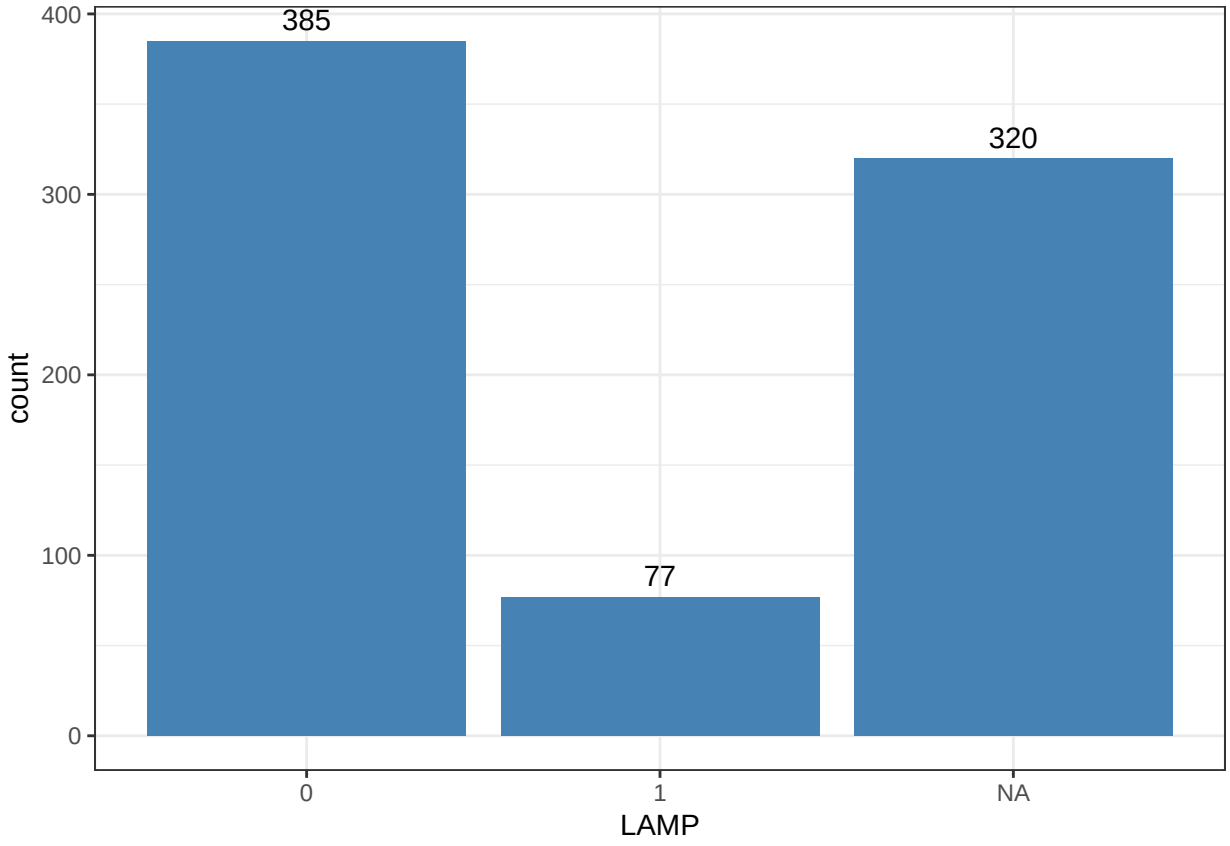
```
#View(data_0)

table(data_0$LAMP)
```

```
##
##   0   1
## 385  77
```

```
sum(is.na(data_0$LAMP)) #
```

```
## [1] 320
```

```
ggplot(data = data_0, aes(LAMP)) + geom_bar(fill="steelblue") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5)+
  theme_bw() #+ ylim(c(0,400))
```




```
### Just to be sure, LAMP was only performed when BSdich was negative
data_1 <- subset(data, BSdich == 1)
table(data_1$LAMP)
```

```
##
## 0 1
## 0 0
```

5. anymalaria

diagnosis of malaria was made on that date (clinical symptoms + positive blood smear). "1" is a diagnosis (case) of malaria, "0" is no malaria

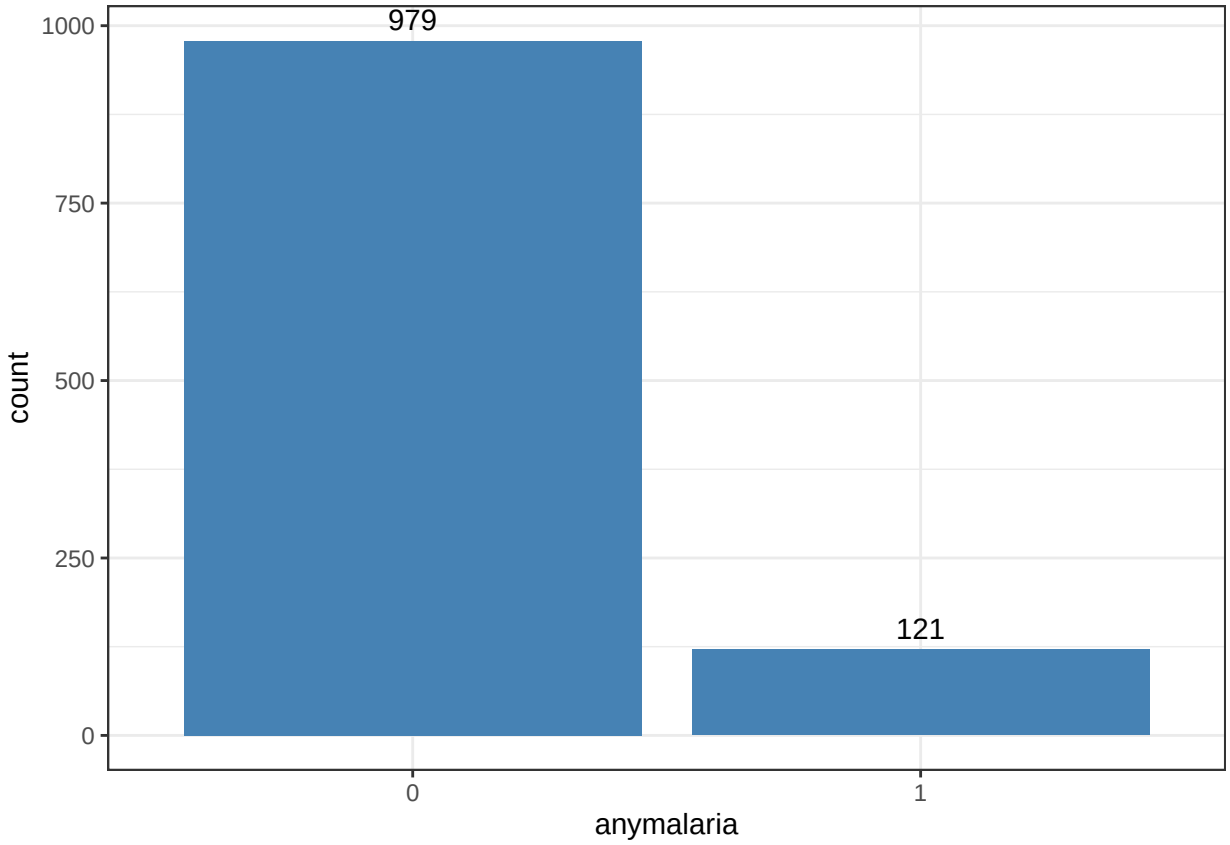
```
table(anymalaria)
```

```
## anymalaria
##   0   1
## 979 121
```

```
sum(is.na(anymalaria))
```

```
## [1] 0
```

```
ggplot(data = data, aes(anymalaria)) + geom_bar(fill="steelblue") +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5)+
  theme_bw()
```



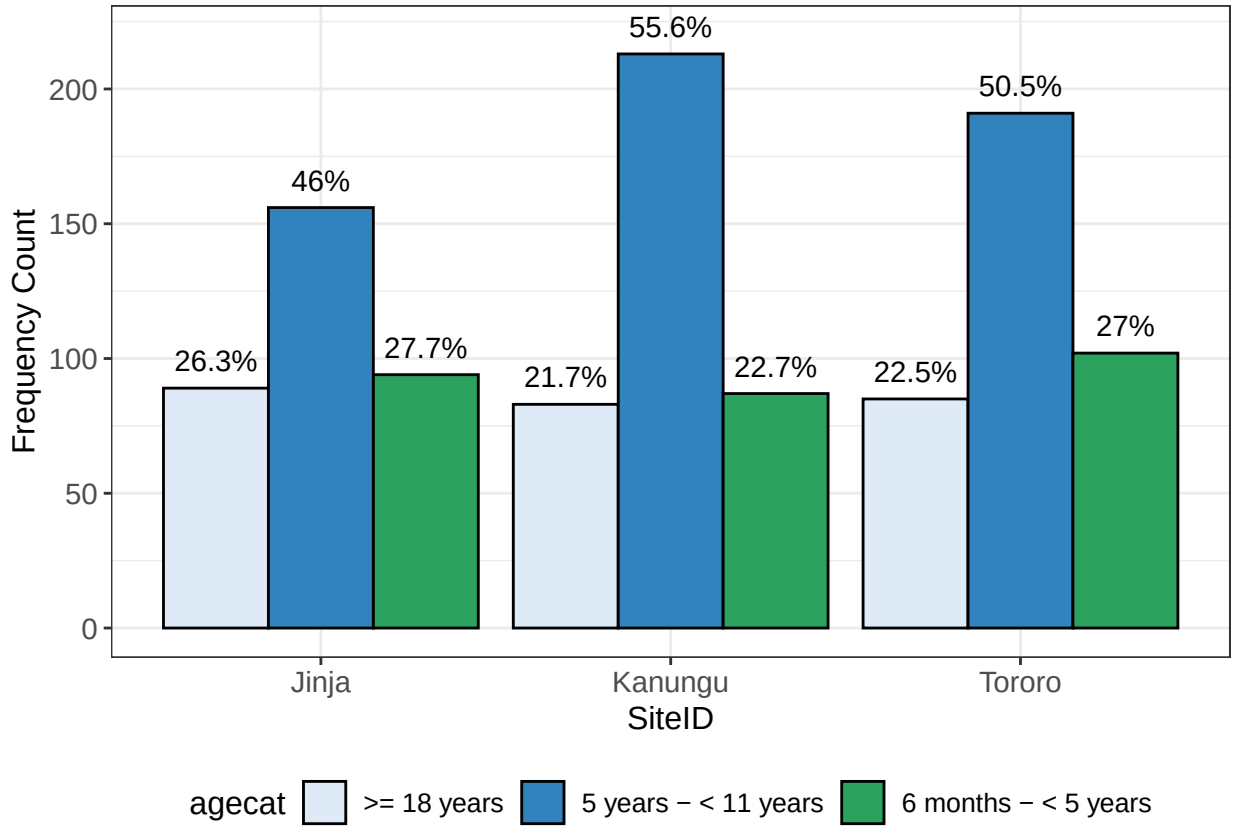
1a. Site and age

Do we have an even balance of the age categories sampled from each region?

```
#SiteID and Agecat
test2 <-   data %>%
  group_by(siteid, agecat) %>%
  summarize(t2.len = length(agecat)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'siteid'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = siteid, y = t2.len, fill = agecat)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") + ylim(0,220)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "SiteID", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd", "#2ca25f")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



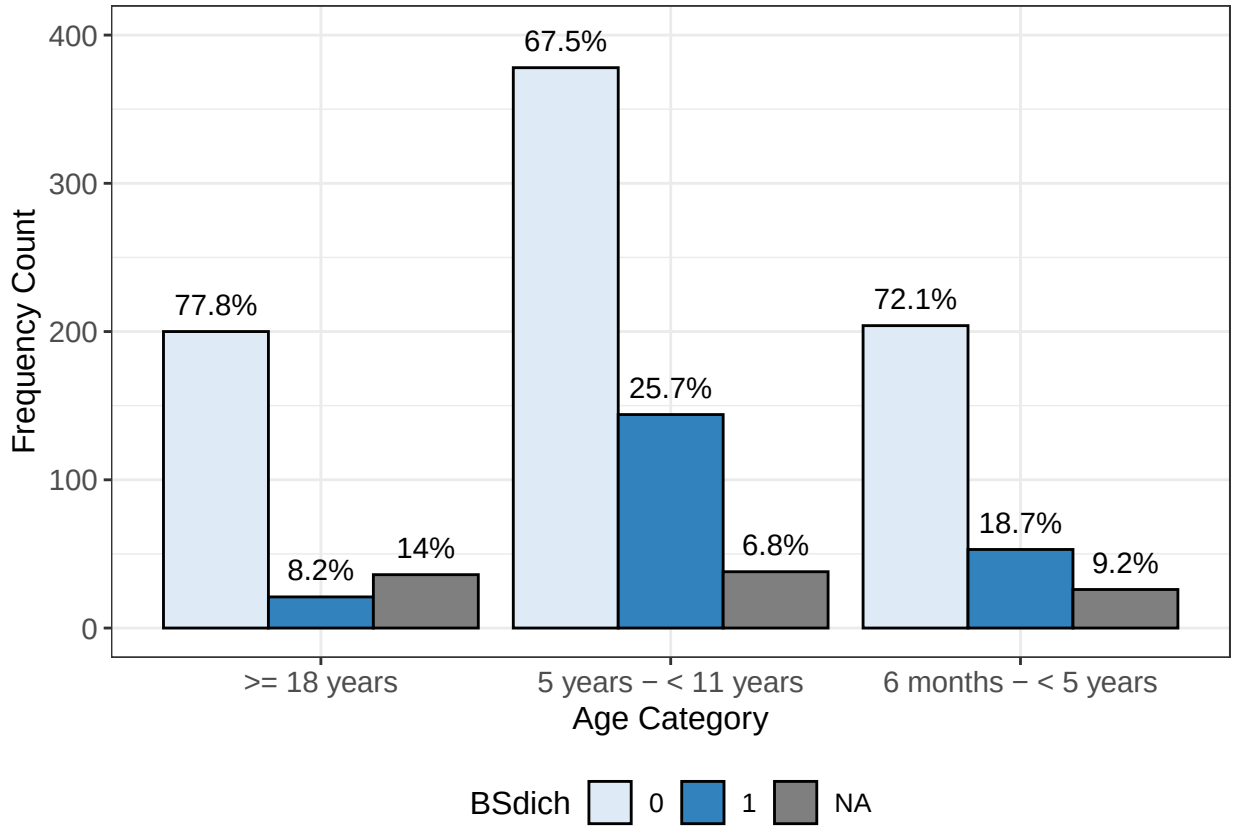
2a Age category and BSdich

Are people of a given age category likely to be positive for a BS

```
# Agecat and BSdich
test2 <-   data %>%
  group_by(agecat, BSdich) %>%
  summarize(t2.len = length(BSdich)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'agecat'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = agecat, y = t2.len, fill = BSdich)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") + ylim(0,400)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Age Category", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



Confirm if there exists a difference;

H0: There is no link / relationship between age category and results of blood smear.

H1: There is a link between age category and results of blood smear.

```
table(agecat, BSdich)
```

```
##                         BSdich
## agecat                     0   1
##   >= 18 years            200  21
##   5 years - < 11 years   378 144
##   6 months - < 5 years   204  53
```

```
c = chisq.test(agecat,BSdich)
c
```

```
##
##  Pearson's Chi-squared test
##
## data:  agecat and BSdich
## X-squared = 30.066, df = 2, p-value = 2.959e-07
```

```
c$expected
```

```
##                         BSdich
## agecat                       0       1
##   >= 18 years          172.822  48.178
##   5 years - < 11 years 408.204 113.796
##   6 months - < 5 years 200.974  56.026
```

```
c$p.value
```

```
## [1] 2.959352e-07
```

Since p-value is significant, we have enough evidence to reject the null hypothesis and conclude that there is a link (relationship) between age category and results of blood smear.

~~ P-value is significant. We are likely to have most positive cases among 4-11 years, followed by more cases in 6months-5 years and least in individuals above 18 years ~~

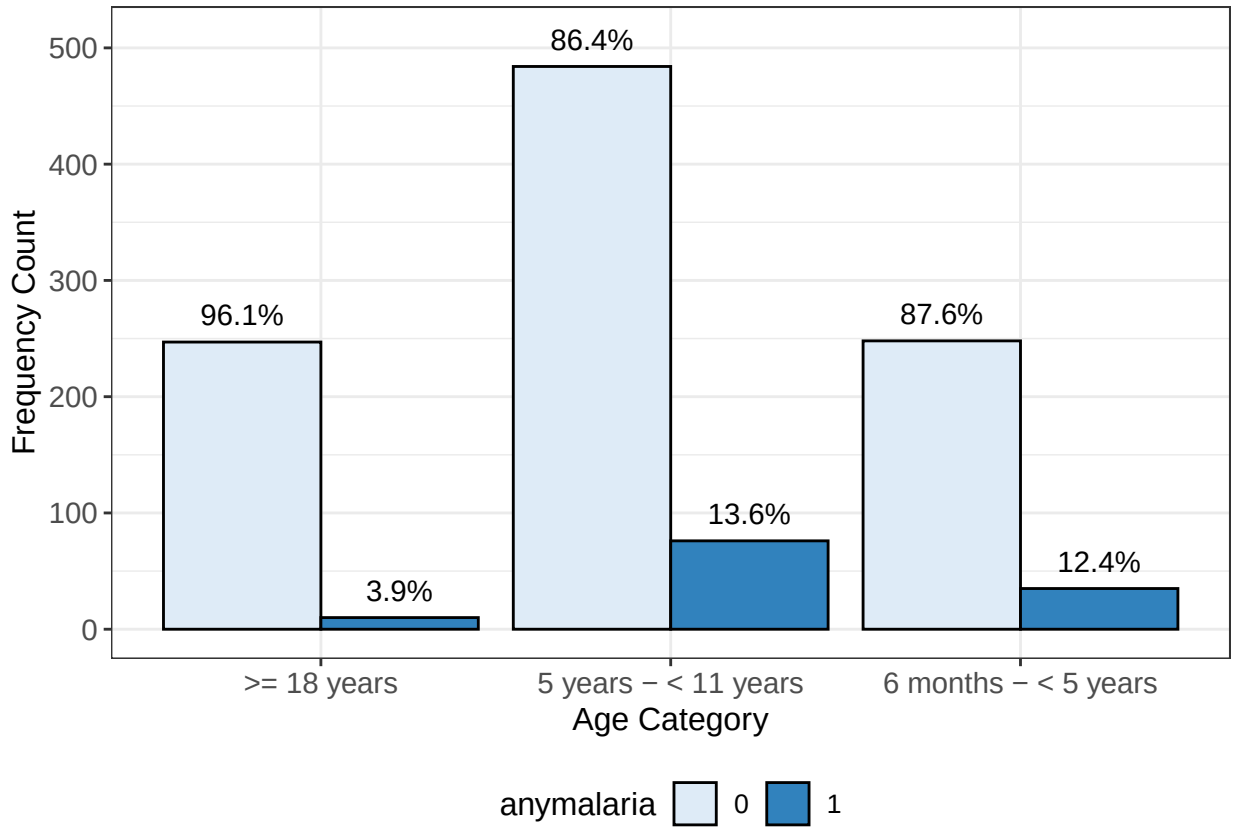
Keep in mind that "statistically significant" does not always imply "meaningful" when using the chi-square test.

2b agecat and anymalaria

```
# Agecat and anymalaria
test2 <-   data %>%
  group_by(agecat, anymalaria) %>%
  summarize(t2.len = length(anymalaria)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'agecat'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = agecat, y = t2.len, fill = anymalaria)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") + ylim(0,510)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Age Category", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



Test significance if there is any relationship between Agecat and anymalaria

```
table(agecat, anymalaria)
```

```
##                       anymalaria
## agecat                   0   1
##   >= 18 years          247  10
##   5 years - < 11 years 484  76
##   6 months - < 5 years 248  35
```

```
c = chisq.test(agecat,anymalaria)
#c
c$expected
```

```
##                       anymalaria
## agecat                      0     1
##   >= 18 years          228.73 28.27
##   5 years - < 11 years 498.40 61.60
##   6 months - < 5 years 251.87 31.13
```

```
c$p.value
```

```
## [1] 0.000151526
```

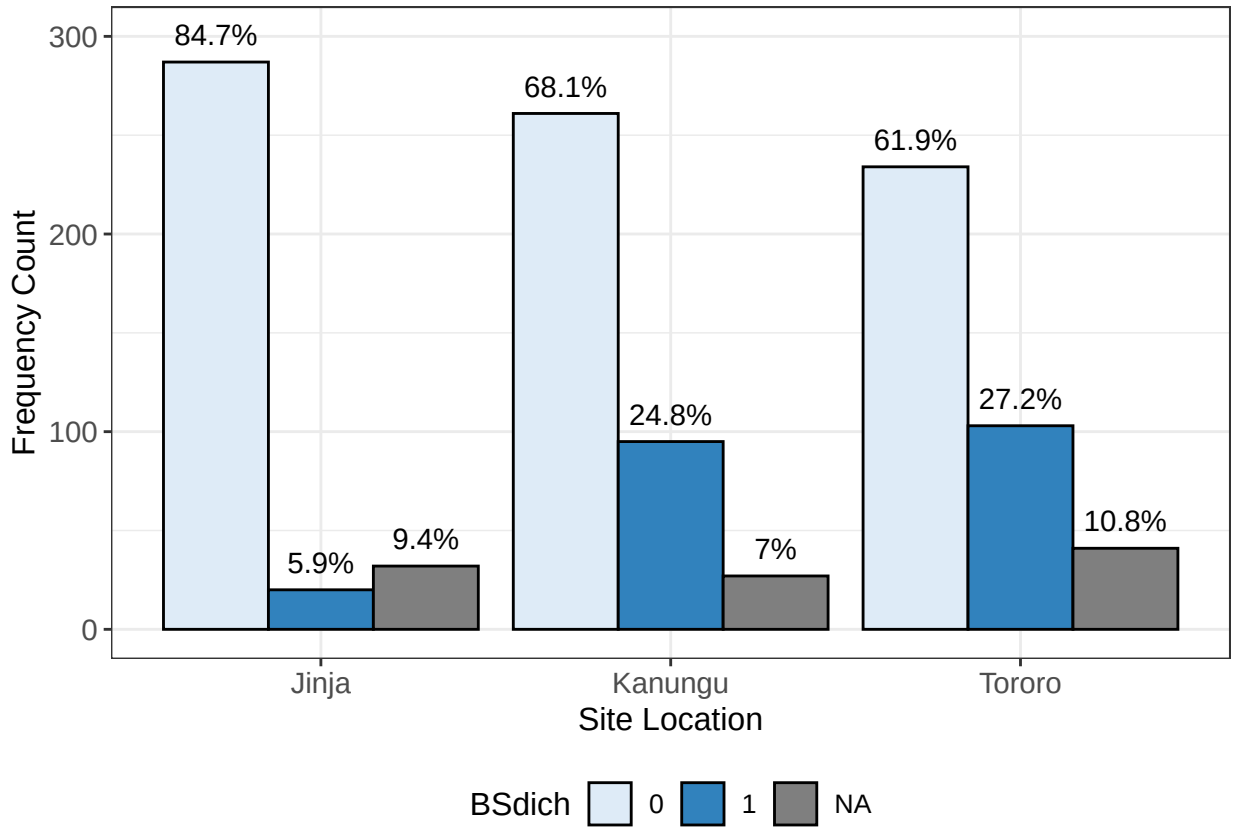
P-value is still significant, hence we can conclude there is a relationship between age category and one having malaria.

3 a siteid and BS

```
# siteid and BSdich
test2 <-   data %>%
  group_by(siteid, BSdich) %>%
  summarize(t2.len = length(BSdich)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'siteid'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = siteid, y = t2.len, fill = BSdich)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") +  ylim(0,300)
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Site Location", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



Check if this difference is significant

```
table(siteid, BSdich)
```

```
##          BSdich
## siteid      0   1
##   Jinja   287  20
##   Kanungu 261  95
##   Tororo  234 103
```

```
c = chisq.test(siteid,BSdich)
c
```

```
##
##  Pearson's Chi-squared test
##
## data:  siteid and BSdich
## X-squared = 62.242, df = 2, p-value = 3.05e-14
```

```
c$expected
```

```
##          BSdich
## siteid          0      1
##   Jinja   240.074 66.926
##   Kanungu 278.392 77.608
##   Tororo  263.534 73.466
```

```
c$p.value
```

```
## [1] 3.05015e-14
```

P-value is significant meaning there is a relationship between site location of a person and the result of the blood smear. Its more likely that people from Kanungu and Tororo will have positive blood smear than people from Jinja.

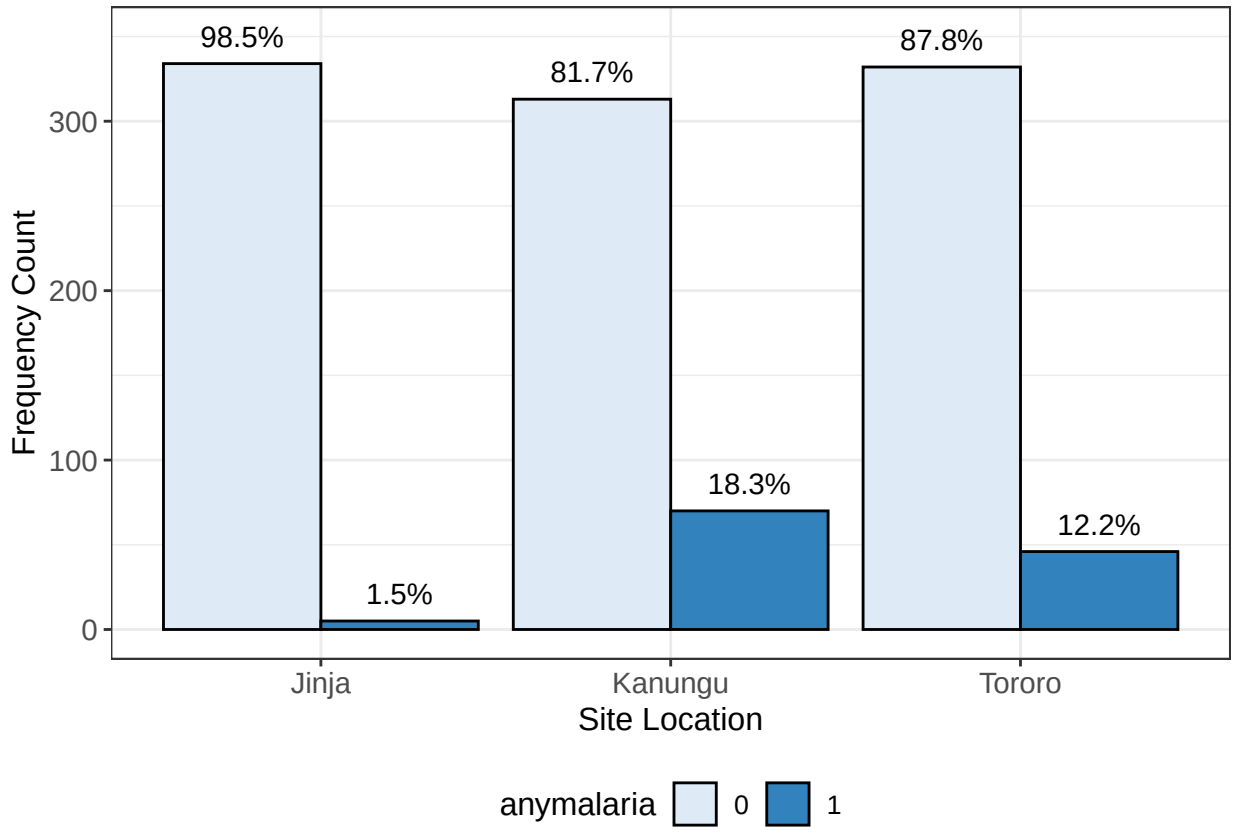
What about malaria diagnosis?

3 b. siteid and anymalaria

```
# siteid and anymalaria
test2 <-   data %>%
  group_by(siteid, anymalaria) %>%
  summarize(t2.len = length(anymalaria)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'siteid'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = siteid, y = t2.len, fill = anymalaria)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") + ylim(0,350)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Site Location", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



Confirm if the observed difference is significant

```
table(siteid, anymalaria)
```

```
##          anymalaria
## siteid      0   1
##   Jinja   334   5
##   Kanungu 313  70
##   Tororo  332  46
```

```
c = chisq.test(siteid,anymalaria)
#c
c$expected
```

```
##          anymalaria
```

```
## siteid         0     1
##   Jinja   301.71 37.29
##   Kanungu 340.87 42.13
##   Tororo  336.42 41.58
```

```
c$p.value
```

```
## [1] 3.673991e-12
```

p-value is significant, hence we conclude that there is a relation between site site location and someone having malaria.

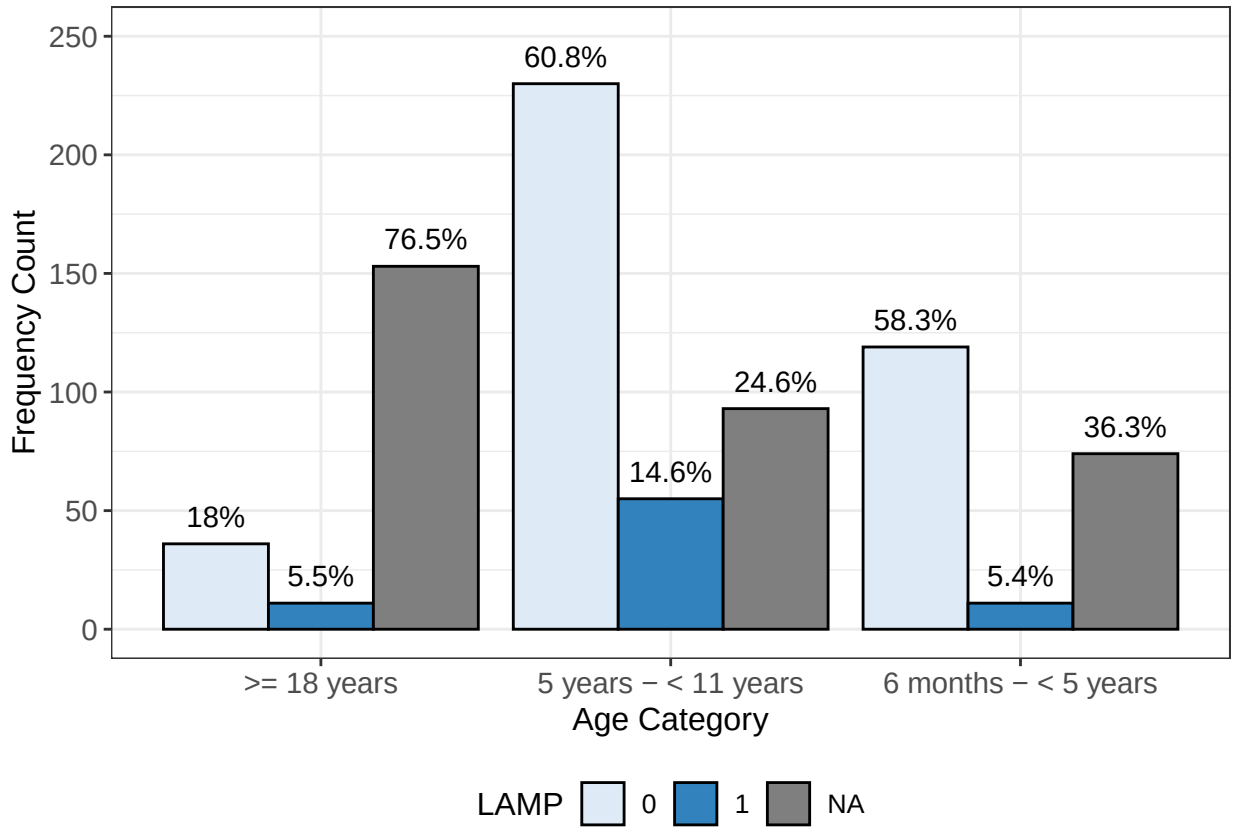
ageCat vs LAMP

Qn: Are individuals of a given age category likely be have submicroscopic results

```
detach(data)
# Agecat and LAMP
test2 <-   data_0 %>%
  group_by(agecat, LAMP) %>%
  summarize(t2.len = length(LAMP)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'agecat'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = agecat, y = t2.len, fill = LAMP)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") +  ylim(0,250)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Age Category", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```



Alot of data is missing from the group >=18

Test the significance if any

```
table(data_0$agecat, data_0$LAMP)
```

```
##
##                           0   1
##   >= 18 years            36  11
##   5 years - < 11 years  230  55
##   6 months - < 5 years  119  11
```

```
c = chisq.test(data_0$agecat, data_0$LAMP)
#c
c$expected
```

```
##                        data_0$LAMP
## data_0$agecat                    0         1
##   >= 18 years            39.16667  7.833333
##   5 years - < 11 years  237.50000 47.500000
##   6 months - < 5 years  108.33333 21.666667
```

```
c$p.value
```

```
## [1] 0.009760803
```

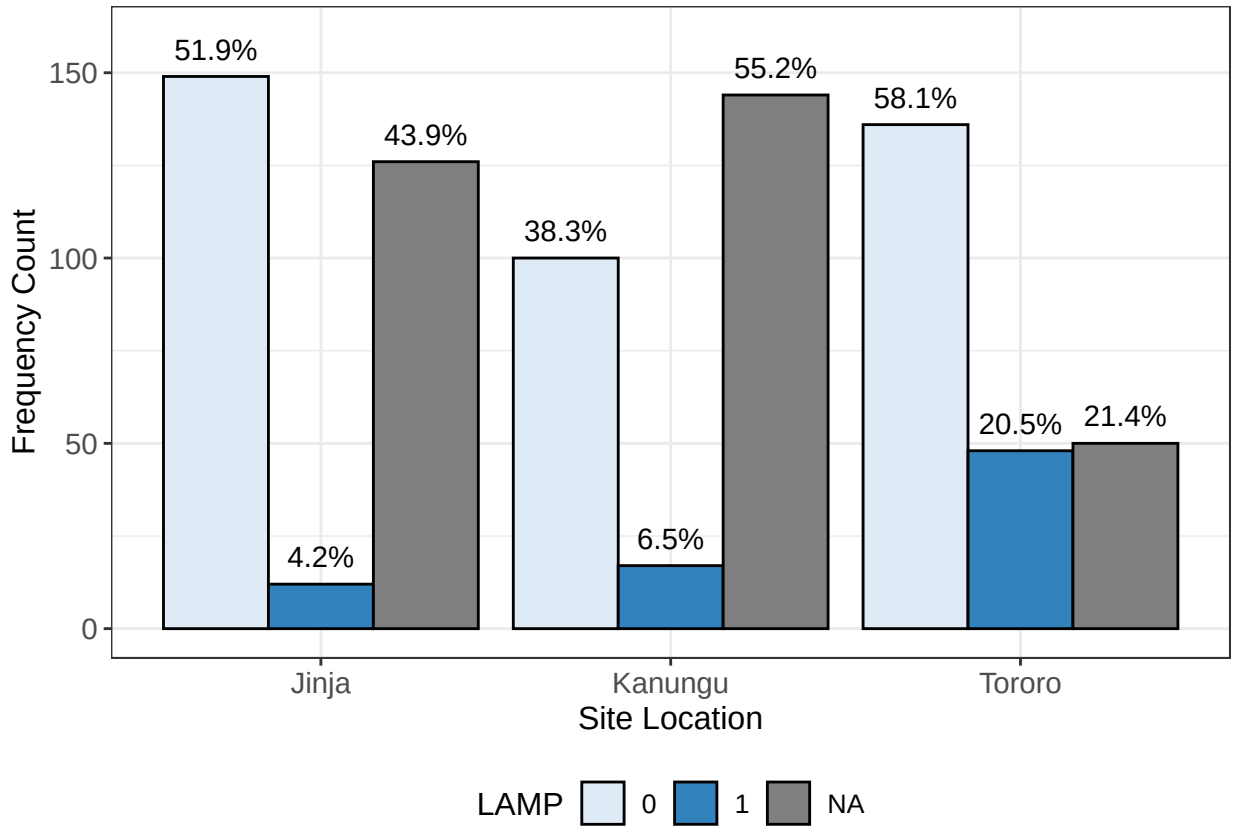
Its significance but level of significance is not that too strong like the previous

siteID vs LAMP

```
# siteid and LAMP
test2 <-   data_0 %>%
  group_by(siteid, LAMP) %>%
  summarize(t2.len = length(LAMP)) %>%
  mutate(t2.prop = round(t2.len / sum(t2.len) * 100, 1))
```

```
## `summarise()` has grouped output by 'siteid'. You can override using the
## `.groups` argument.
```

```
p22 <- ggplot(test2, aes(x = siteid, y = t2.len, fill = LAMP)) +
  geom_bar( stat = "identity", position = position_dodge(width = 0.9) , color="black") +  ylim(0,160)  +
  geom_text(aes(label = paste(t2.prop, "%", sep = "")),
            position = position_dodge(width = 0.9), vjust = -0.8) + theme_bw() +
  theme(legend.position = "bottom") + labs(x = "Site Location", y = "Frequency Count") +
  scale_fill_manual(values = c("#deebf7", "#3182bd")) +
  theme( axis.text=element_text(size=11),text=element_text(size=12))

p22
```




```
table(data_0$siteid, data_0$LAMP)
```

```
##
```

```
##             0   1
##   Jinja   149  12
##   Kanungu 100  17
##   Tororo  136  48
```

```
c = chisq.test(data_0$siteid,data_0$LAMP)
c
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_0$siteid and data_0$LAMP
## X-squared = 21.981, df = 2, p-value = 1.686e-05
```

```
c$expected
```

```
##              data_0$LAMP
## data_0$siteid        0        1
##       Jinja   134.1667 26.83333
##       Kanungu  97.5000 19.50000
##       Tororo  153.3333 30.66667
```

```
c$p.value
```

```
## [1] 1.686206e-05
```

P-value is still significant

```
dim(na.omit(data_0))
```

```
## [1] 462   6
```

```
s <- na.omit(data_0[c("siteid", "LAMP")])
dim(s)
```

```
## [1] 462   2
```

```
table(s$siteid, s$LAMP)
```

```
##
##             0   1
##   Jinja   149  12
##   Kanungu 100  17
##   Tororo  136  48
```

```
c = chisq.test(s$siteid,s$LAMP)
c
```

```
##
##  Pearson's Chi-squared test
##
## data:  s$siteid and s$LAMP
## X-squared = 21.981, df = 2, p-value = 1.686e-05
```

```
c$expected
```

```
##          s$LAMP
## s$siteid         0        1
##   Jinja   134.1667 26.83333
##   Kanungu  97.5000 19.50000
##   Tororo  153.3333 30.66667
```

```
c$p.value
```

```
## [1] 1.686206e-05
```

Note a limitatiob of chi square, When Sample size is small, the test is less trustworthy. However, with very large sample sizes, even relatively trivial relationships may be declared statistically significant.