**DAY 1 WORKSHEET - 1.1 – 1.3**

**1.1   Getting to know your data**

**1.1a   Explore our dataset**

**Questions:**

Try out the following commands to get to know the data:

1. nrow(raw\_data) - How many entries does the data frame have? 1599
2. names(raw\_data) - What are the names of the first 3 columns? month, date, sex
3. dim(raw\_data) - What are the dimensions of our data? 1599 rows, 9 columns
4. summary(raw\_data) - What is the mean age of the organisms infected? 8.004
5. head(raw\_data) - In which region did the 1st case occur? Shinyanga region
6. tail(raw\_data) - In which region did the last case occur? Tanga region
7. str(raw\_data) - Which variables are numerical (num)? y, x, density
8. unique(raw\_data$species) - What types of species do we have in the data? human, dog, jackal, cat, lion
9. table(raw\_data$species) - Which species has the most cases? dog (n=727)
10. ?names() What is the first argument for the function names()? x – an R object

**1.2   Data subsetting and summarising**

**1.2a   Subsetting data**

**Questions:**

What do each of these lines of code subset the data for? 

1. select(raw\_data, x, y, age) columns named x, y, and age
2. filter(raw\_data, region %in% c("Mara", "Pwani", "Dar-es-salaam")) rows from Mara, Pwani or Dar-es-Salaam
3. filter(raw\_data, age >= 30) rows greater than or equal to 30 years of age

**1.2b   Summarising data**

**Questions:**

What do each of these lines of code summarise?

1. summarise(raw\_data, n\_males = length(which(sex=="M"))) number of males (n=809)
2. summarise(raw\_data, total\_age = sum(age)) all ages added together (n=12,798)

How are these two tables different?

df\_1 <- group\_by(raw\_data, sex)

summarise(df\_1,

n\_records = length(sex),

mean\_age = mean(age))

df\_2 <- group\_by(raw\_data, region, sex)

summarise(df\_2,

n\_records = length(sex),

mean\_age = mean(age))

df\_1 is grouped by only the sex column, which breaks down n\_records and mean\_age by sex.

df\_2 is grouped by both region and sex columns, which breaks down n\_records an mean\_age by region and sex.

**1.2c   Mutating your data and using the pipe**

**Questions:**

What do each of these lines of code add as a new column?

1. raw\_data <- mutate(raw\_data, "new\_col\_1" = substr(date, 1, 4)) substring of date to give year
2. raw\_data <- mutate(raw\_data, "new\_col\_2" = ifelse(density >= 500,"High", "Low")) text variable for density

Fill in the blanks for the following lines of code in your R script:

1. Subset for only records with wildlife (i.e. jackal, lion)

raw\_data %>%

filter(species %in% c(“jackal”, “lion”)

1. Subset for humans, and summarise the mean age per region

raw\_data %>%

filter(species == “human”) %>%

group\_by(region) %>%

summarise(mean\_age =mean(age))

**1.3   Building exploratory plots**

**1.3a   Practice plotting with ggplot2**

**Questions:**

Fill in the blanks for the following lines of code in your R script:

1. Create a barplot of species and colour by species.

ggplot() +

geom\_bar(data=raw\_data, aes(x=species, fill = species)) +

theme\_classic()

1. Add a title, x and y axis labels to the barplot of species

ggplot() +

geom\_bar(data=raw\_data, aes(x=species, fill =species)) +

xlab(“Species”) +

ylab(“Cases”) +

ggtitle(“Disease cases by Species”)+

theme\_classic()