44940076_Sekhar

Arjun Sekhar

08/09/2021

Question 1

Firstly, we set our working directory and read our dataset 'zooplankton.csv' and we then we note our observations accordingly.

```
setwd("~/OneDrive/Assessments/University/Year 4/Semester 2/Stat8123/Tutorials/Assignment 2")
zooplankton <- read.csv("zooplankton.csv", header = TRUE)

library(pander)
pander(head(zooplankton), caption = "Zooplankton Data")
```

Table 1: Zooplankton Data (continued below)

Tow_Number	Ship_Code	Date	Month	Year	Season	Latitude
1	AA	12-Jan-91	January	1991	1990-91	-63.28
1	AA	12-Jan-91	January	1991	1990-91	-63.28
1	AA	12-Jan-91	January	1991	1990-91	-63.29
1	AA	12-Jan-91	January	1991	1990-91	-63.29
1	AA	12-Jan-91	January	1991	1990-91	-63.29
1	AA	12-Jan-91	January	1991	1990-91	-63.29

Table 2: Table continues below

Longitude	Segment_No	Segment_Length	Calanoida_indet	Calanus_simillimus
88.38	1	5	0	0
88.2	2	5	0	0
88.02	3	5	0	0
87.83	4	5	0	0
87.64	5	5	0	0
87.45	6	5	0	0

Table 3: Table continues below

Foraminifera_indet	Fritillaria_sp	Oithona_similis	Total_abundance
0	0	0	0
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0

Foraminifera_indet	Fritillaria_sp	Oithona_similis	Total_abundance
0	0	0	1

Fluorescence	Salinity	Water_Temperature
15.18	33.91	3.38
8.97	33.91	3.33
7.81	33.9	3.4
9.34	33.9	3.29
7.49	33.8	3.04
7.06	33.69	2.93

We want to find the most abundant zooplankton species, which means that we see the following observations.

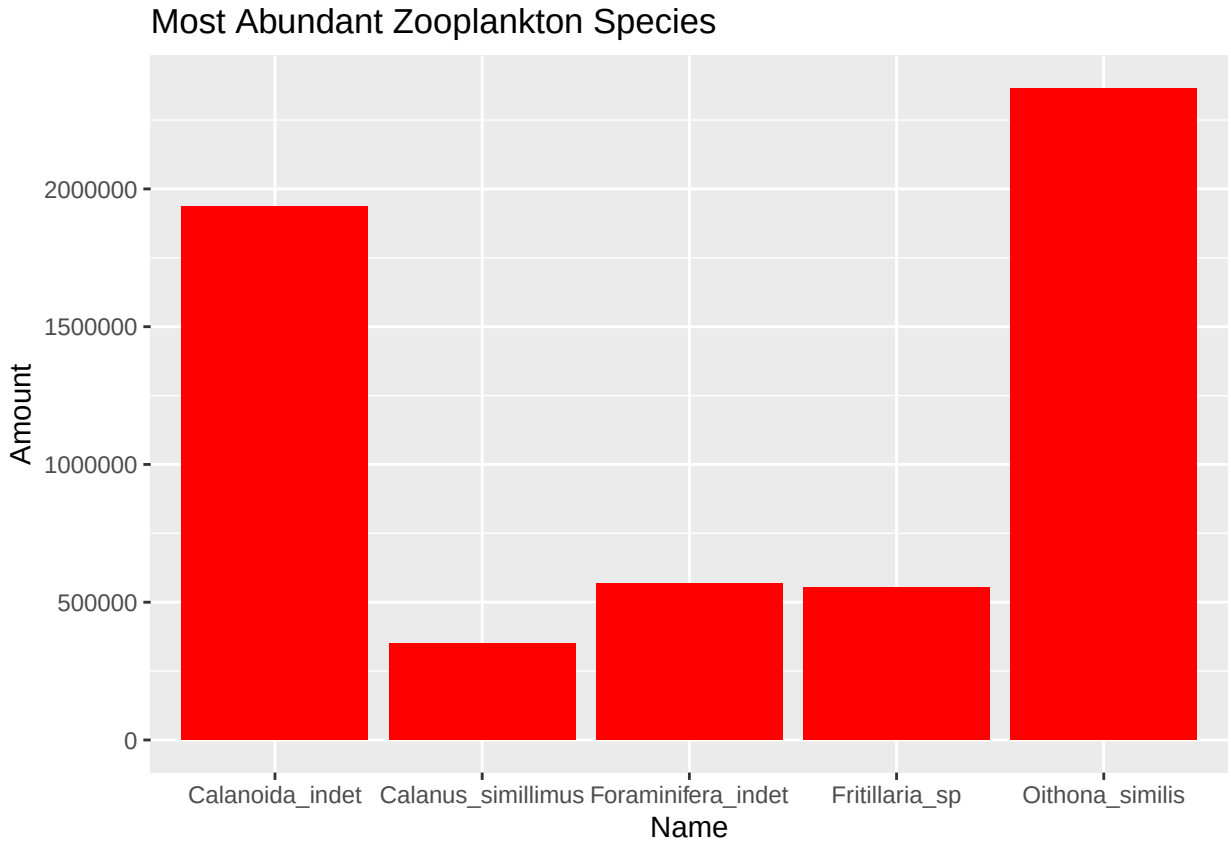
```
# Step 1: Inputting the packages
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)

# Step 2: Finding the sum of each Zooplankton species.
Amount <- c(sum(zooplankton$Calanoida_indet),
            sum(zooplankton$Calanus_simillimus),
            sum(zooplankton$Foraminifera_indet),
            sum(zooplankton$Fritillaria_sp),
            sum(zooplankton$Oithona_similis))

# Step 3: Naming the Zooplankton species names.
Name <- c("Calanoida_indet",
          "Calanus_simillimus",
          "Foraminifera_indet",
          "Fritillaria_sp",
          "Oithona_similis")

# Step 4: Creating a data frame.
pop_zooplankton <- data.frame(Name, Amount)

# Step 5: Using ggplot, we plot pop_zooplankton data frame.
ggplot(pop_zooplankton, aes(x = Name, y = Amount)) +
  geom_bar(stat = "identity", fill = "red") +
  ggtitle("Most Abundant Zooplankton Species")
```



Interpreting the graph, it is a bar graph that tells us the amount of each species of Zooplankton that were observed in the Southern Ocean. Given red is a vibrant colour, we use this as our primary colour of this graph.
From the above, we can see that the most abundant Zooplankton species is the Oithona_Similis species with over 2 250 000 species (approximately) existing in the Southern Ocean. The graphics that we have used are from the packages *ggplot2* and *tidyverse* and the rank of the species we analysed were:

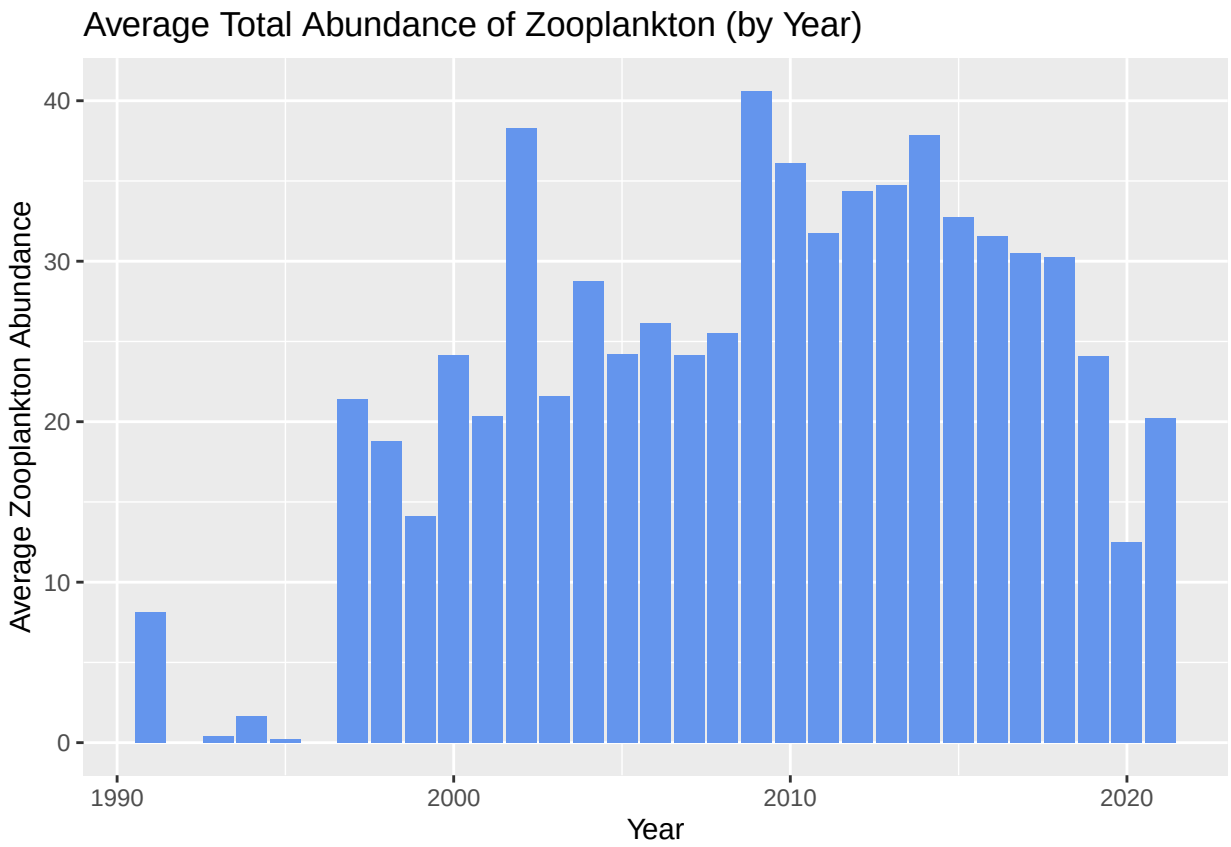
1. Oithona Similis
2. Calanoida Indet
3. Foraminifera Indet
4. Fritillaria Sp
5. Calanus Simillimus

Question 2

For the average total abundance we firstly create another bar graph using the *tidyverse* and *ggplot2* function. We do this in two stages here in RMarkdown, because we firstly want to take into account about the fact that the years had varying number of segments and lengths taken.

```
# Step 1: Load the 'Tidyverse' package.
library(tidyverse)
Total_Abundance <- zooplankton$Total_abundance
Segment_Length <- zooplankton$Segment_Length

# Step 2: Create a variable to find and summarise the mean of Total Abundance.
Avg <- zooplankton %>%
  group_by(Year) %>%
```

```
  summarise(mean_abund=mean(Total_abundance/Segment_Length))

# Step 3: We plot this using a Bar chart (geom_bar()).
ggplot(Avg, aes(x=Year, y=mean_abund)) +
  geom_bar(stat = "identity", fill = "cornflowerblue") +
  ggtitle("Average Total Abundance of Zooplankton (by Year)") +
  xlab("Year") +
  ylab("Average Zooplankton Abundance")
```

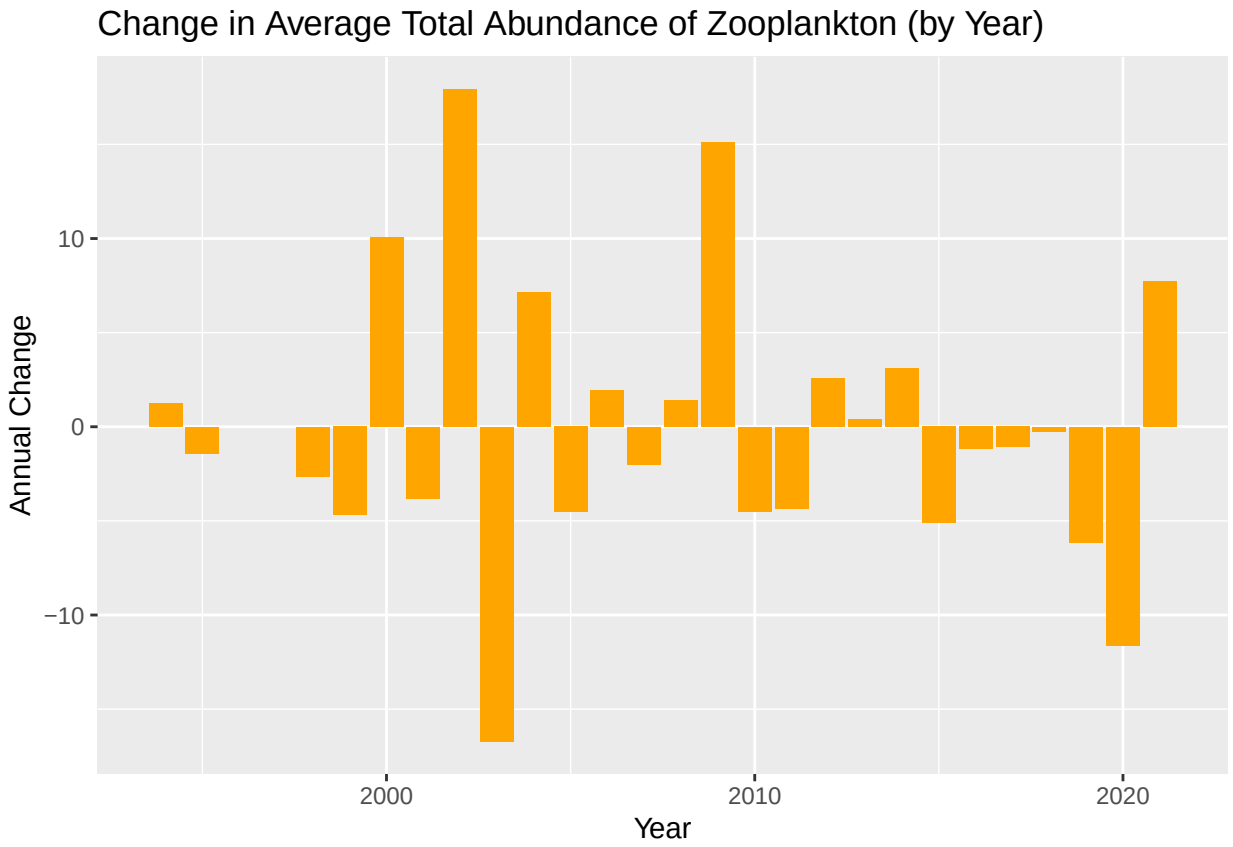


Firstly, we can see here that during the early years, there were not many observations of Zooplankton that were recorded in this data set, which resulted in the averages in those early years to be a paltry amount. However as the years wore on, particularly when it reached the 2010 decade, the average zooplankton abundance increased.

We now want to understand the change in the average total abundance by year, so we will implement another bar chart, except we want to also introduce the *merge()* and the *mutate()* function.

```
# Step 1: We want to find the change, so we take the 'Change from Previous Year'.
Avg$Prev_Year <- Avg$Year - 1

# Step 2: We create a 'Change' variable by merging and mutating the average
Change <- Avg %>%
  merge(Avg, by.x = 'Prev_Year', by.y = 'Year') %>%
  mutate(annual_change = mean_abund.x - mean_abund.y)

# Step 3: We plot this in a bar graph and notice it as a percentage change.
ggplot(Change, aes(x = Year, y = annual_change)) +
  geom_bar(stat = "identity", fill = "orange") +
  xlab("Year") +
```

```
  ylab("Annual Change") +
  ggtitle("Change in Average Total Abundance of Zooplankton (by Year)")
```

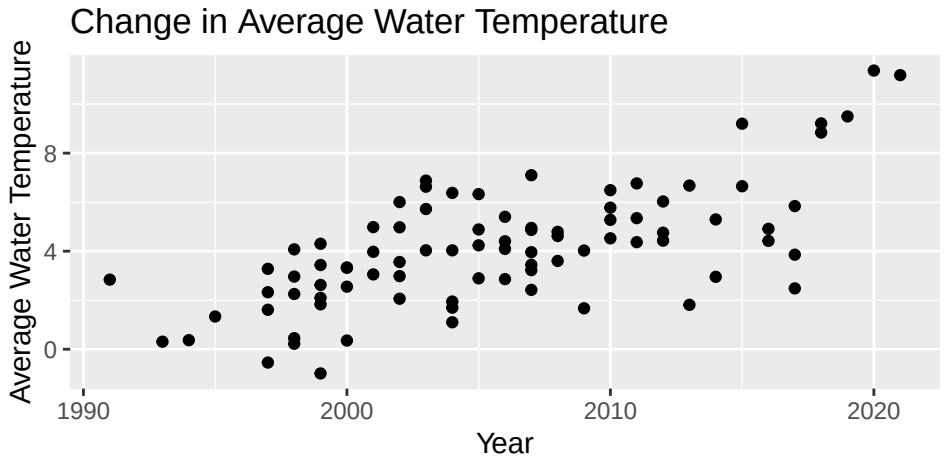


Based on this plot, we can see that given there were fewer observations of Zooplankton in the early 1990s, as the study approached the start of the new millennium, the rate of change was particularly rapid during the years 2002 (greatest increase) and 2003 (greatest decrease). In fact the 2000s decade had a particularly fluctuating cycle of abundance of Zooplankton.

Question 3

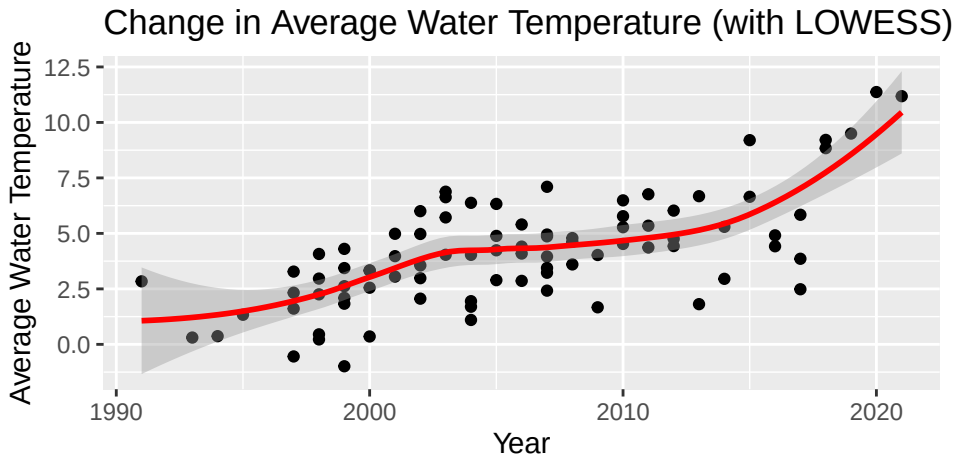
Our goal here is to find the average change in water temperature over time. We employ the *tidyverse* package once again and the intention is to present plots using 'ggplot()'.

```
# Step 1: We firstly find the mean of the Water Temperature variable.
Avg_H20 <- zooplankton %>%
  group_by(Month, Year) %>% summarise(mean_WT=mean(Water_Temperature))

# Step 2: Plot the Change in Water Temperature
library(ggplot2)
ggplot(Avg_H20, aes(x=Year,y=mean_WT)) +
  geom_point()  +
  ggtitle("Change in Average Water Temperature") +
  xlab("Year") +
  ylab("Average Water Temperature")
```



In this plot, we can see that the that the change in the water temperature appears to have an increasing trend. However, there is a problem here, as indicated in the question - we have no LOWESS Curve to prove that the overall trend is a positive trend, and this is highly important in proving any trend in a scatterplot.

```
# Step 3: Include the LOWESS Curve in the geom_smooth() bracket.
library(ggplot2)
ggplot(Avg_H20, aes(x=Year,y=mean_WT)) +
  geom_point() +
  geom_smooth(method = "loess", colour = "red") +
  ggtitle("Change in Average Water Temperature (with LOWESS)") +
  xlab("Year") +
  ylab("Average Water Temperature")
```

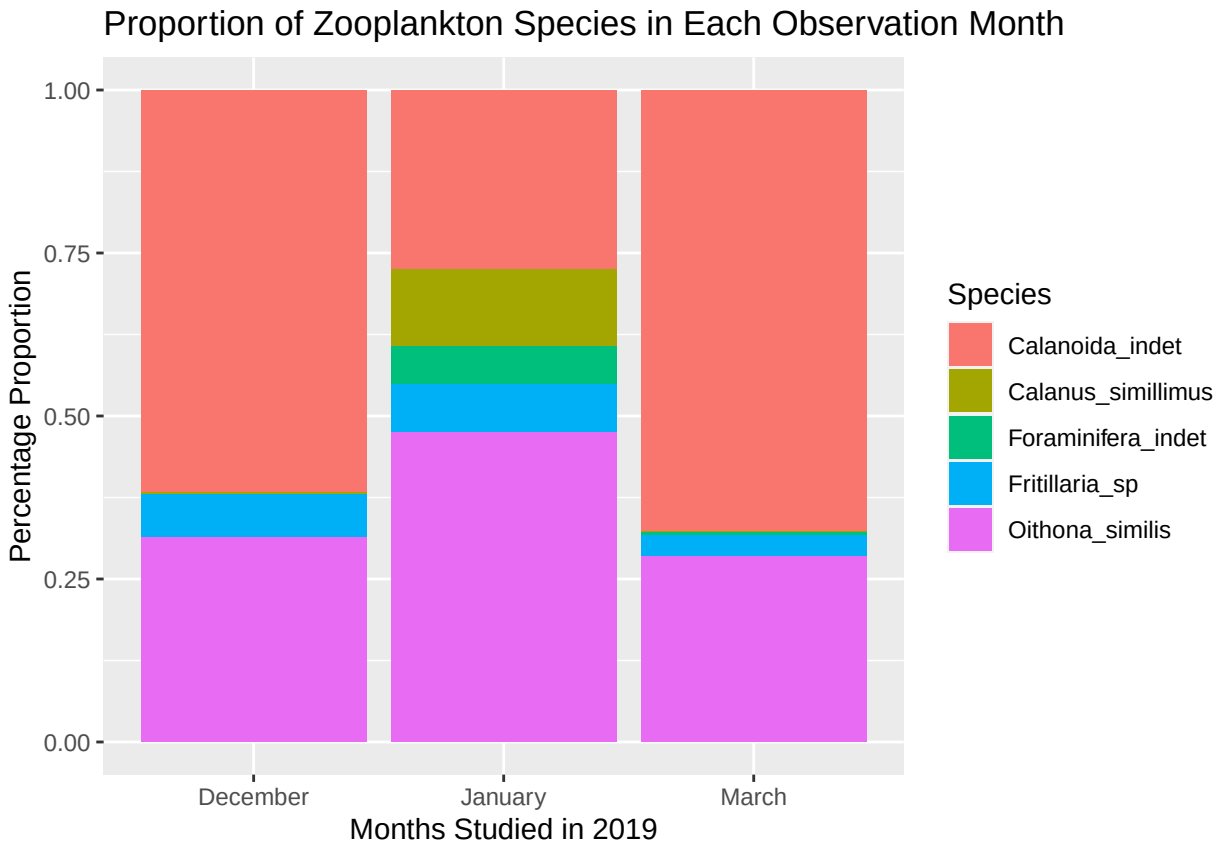


Upon including a smooth line of best fit (also known as a 'LOWESS Curve') with confidence intervals, we can see that there is actually a more increasing trend, but during certain stages over the years (particularly from 2005 to 2010), the curve flattens before incrementing on a consistent basis after around 2011.
Using a LOWESS Curve, we are able to gauge the changes taking place in relation to the average Water Temperature, and more so, we are able to look into the future and suggest that there is an increasing trend following on. More so, it is easier to prove initial judgements without the LOWESS Curve as misleading, particularly since overall it looks to be following an increasing pattern - only upon implementing the curve do we notice when the increase really takes place (since there were stages that the curve flattened).
Additionally, the grey intervals tell us the maximum and minimum intervals, which is particularly helpful for us given that we are able to use this to predict the Water Temperature changes deeper into the future.

Question 4

This question involves us displaying the proportion of each of the plankton species identified in the observation months of 2019. We employ the *geom_bar()* and the key word here is 'proportion', so when we find our results we will observe this key word to identify the most popular species in 2019.

```
# Step 1: We create our 'Name' data frame to list the Zooplankton species.
Name <- c("Calanoida_indet", "Calanus_simillimus", "Foraminifera_indet",
          "Fritillaria_sp", "Oithona_similis")

# Step 2: We create our 'Months2019' data frame to list the relevant months.
Months2019 <- c("January", "January", "January", "January", "January",
                "March", "March", "March", "March", "March",
               "December", "December", "December", "December", "December")

# Step 3: We also create a data frame to list our species.
Species <- c("Calanoida_indet", "Calanus_simillimus", "Foraminifera_indet",
             "Fritillaria_sp", "Oithona_similis", "Calanoida_indet",
             "Calanus_simillimus", "Foraminifera_indet", "Fritillaria_sp",
             "Oithona_similis", "Calanoida_indet", "Calanus_simillimus",
             "Foraminifera_indet", "Fritillaria_sp", "Oithona_similis")

# Step 4: Select the relevant variables and group them by Month.
zooplankton.2019 <- zooplankton %>% group_by(Month) %>%
  filter(Year == 2019) %>%
  summarise_at(vars(Calanoida_indet, Calanus_simillimus, Foraminifera_indet,
                    Fritillaria_sp, Oithona_similis), sum) %>%
  count() %>%
  mutate(percentage = 100*(n/sum(n)))

# Step 5: Compile the summation frequencies.
Frequency <- c(11109,4766,2382,2998,19158,3241,0,27,153,1363,850,3,0,92,431)
Percentage = 100*(Frequency/sum(Frequency))

# Step 6: Create our data frame using Months, Species, Frequency and Percentage.
zooplankton_df <- data.frame(Months2019, Species, Frequency, Percentage)

# Step 7: We use the 'fill' position to display the proportions in our bar graph.
ggplot(zooplankton_df, aes(fill = Species, x = factor(Months2019), y = Percentage)) +
  geom_bar(position = "fill", stat = "identity") +
  ggtitle("Proportion of Zooplankton Species in Each Observation Month") +
  xlab("Months Studied in 2019") +
  ylab("Percentage Proportion")
```



From our above graph, we can firstly take note of the fact that the months that recorded observations in this study were December, January and March. From that, we can note the following:

- In December, the most predominant Zooplankton is the Calanoida Indet species.
- In January, the most predominant Zooplankton was the Oithona Similis species.
- In March, the most predominant Zooplankton was the Calanoida Indet species.

By power of deduction, we can see that the same species was not the most abundant in all observation months of 2019. While it was predominantly observed, Oithona Similis clearly spoiled the show for us.
Describing the graph, the bar plot function in the ggplot package was once again of use in our study. In particular, the 'fill' position was useful because given the expectation of this task is to find the proportion, this tool helps us in deducing which Zooplankton species was predominant in each month observed.

Question 5

As the first step, we introduce our packages, which are in this case *tidyverse* and *sf*. Subsequently, we input our image of the World Map, using the st_read() function.

```
# Step 1: Load packages 'Tidyverse' and 'Sf'.
library(tidyverse)
library(sf)
```

```
## Linking to GEOS 3.8.1, GDAL 3.2.1, PROJ 7.2.1
```

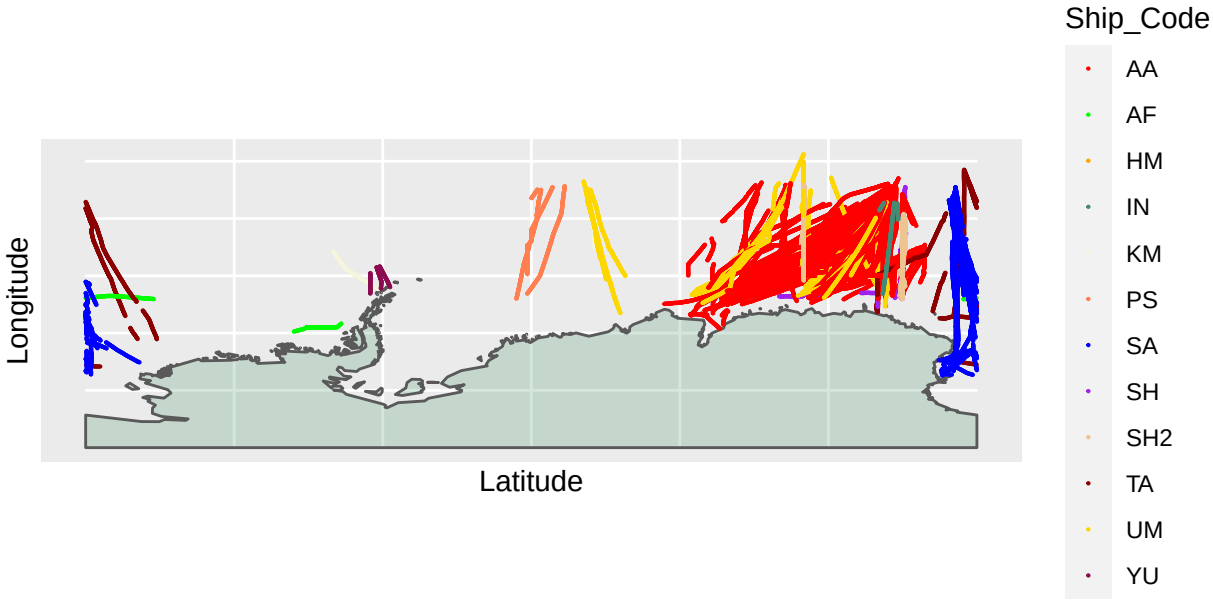
```
#Step 2: Load the world map into the system.
world_map <- st_read("TM_WORLD_BORDERS_SIMPL-0.3.shp")
```

```
## Reading layer `TM_WORLD_BORDERS_SIMPL-0.3' from data source
##   `/Users/arjun/OneDrive/Assessments/University/Year 4/Semester 2/Stat8123/Tutorials/Assignment 2/TM_
```

```
##   using driver `ESRI Shapefile'
## Simple feature collection with 246 features and 11 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -180 ymin: -90 xmax: 180 ymax: 83.57027
## Geodetic CRS:  WGS 84
```

Now here, we want to see an image of Antarctica and add the Zooplankton data, since our objective here is to find which ships made the most and the least observations.

```
# Step 3: Extract the Latitude and Longitude, assigning them as variables.
Longitude <- world_map$LON
Latitude <- world_map$LAT

# Step 4: Filter out Antarctica.
Antarctica <- filter(world_map, NAME == "Antarctica")

# Step 5: Let's plot with some amazing colours in our palatte.
ggplot(Antarctica) + geom_sf(fill = "seagreen4", alpha = 0.2) +
  geom_point(data = zooplankton, aes(x = Longitude, y = Latitude,
                                     color = Ship_Code), size = 0.1) +
  scale_fill_brewer(palette = "Set3") +
  scale_x_continuous(name = "Latitude") +
  scale_y_continuous(name = "Longitude") +
  coord_sf() +
  scale_color_manual(values=c("red","green","orange",
                              "aquamarine4","beige","coral",
                              "blue","purple","burlywood2",
                              "darkred","gold","deeppink4"),)
```




```
library(pander)
pander(table(zooplankton$Ship_Code),
       caption = "Ships that made the Most and Least Observations")
```

Table 5: Ships that made the Most and Least Observations

AA	AF	HM	IN	KM	PS	SA	SH	SH2	TA	UM	YU
30647	264	163	251	278	721	5060	2871	2996	2196	6128	142

Interpreting the graph, it we can see that there were 12 ships that navigated across several parts of the Southern Ocean. Our objective here is to determine which ship travelled the most extensively. While the colours were somewhat feint, we can see that the majority of the ships travelled more glaringly around second set of grids from the right, as this study was targeted towards the regions between Australia and Antarctica.

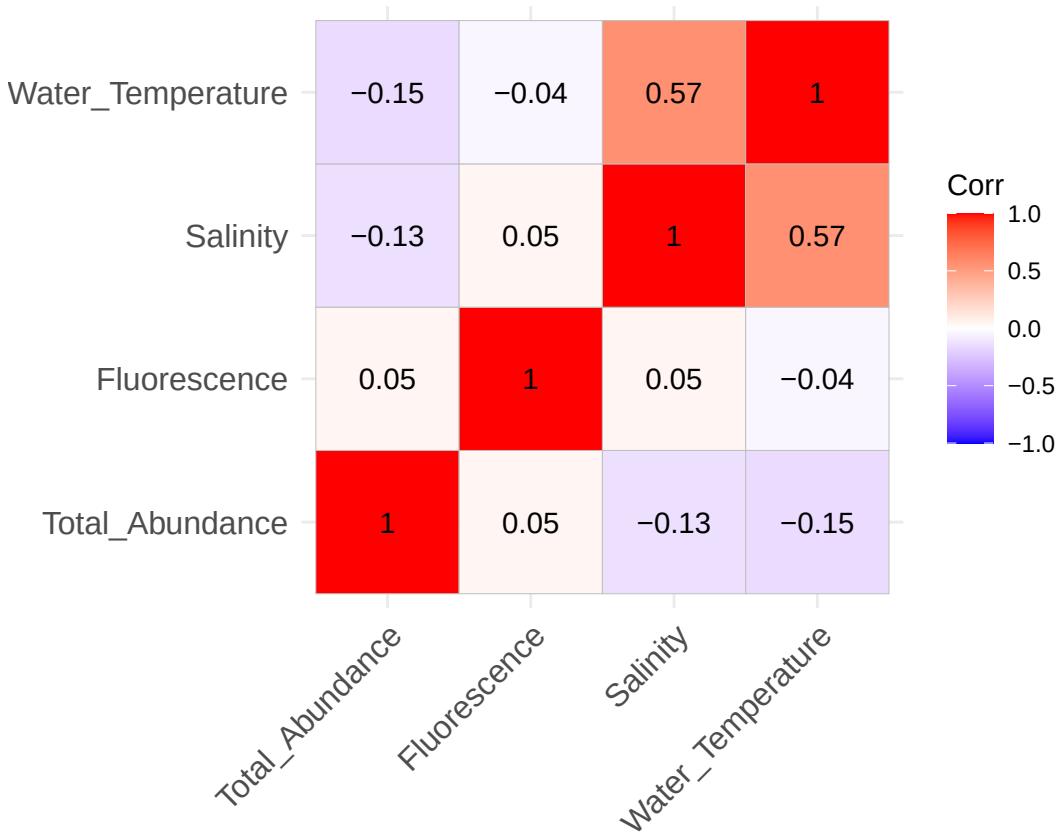
This was a messy graph to stare at, but we can see that the ship with the Code AA (red) was the one that travelled the most extensively, followed by the ship SA (blue). The ship YU was the least extensive travelling ship, which could only be found travelling along the sparingly on the left side of the plot of Antarctica.

Question 6

Research Question: Using the variables Fluorescence, Salinity and Water Temperature, determine the correlation of these variables against each other and determine the most correlated variable pair.

Approach: We intend to evaluate the correlation between the three variables Fluorescence, Salinity and Water Temperature, and understand the relationships that prevail between the variables.

```
# Step 1: Define the variables again, for the sake of creating a data frame.
Total_Abundance <- zooplankton$Total_abundance
Fluorescence <-zooplankton$Fluorescence
Salinity <- zooplankton$Salinity
Water_Temperature <- zooplankton$Water_Temperature

# Step 2: Create our data frame.
zooplankton2 <- data.frame(Total_Abundance,
                           Fluorescence,
                           Salinity,
                           Water_Temperature)


# Step 3: Make the variables numeric using dplyr.
library(dplyr)
zooplankton_df1 <- zooplankton2 %>% dplyr::select(where(is.numeric))

# Step 4: Plot the chart using 'ggcorrplot'.
library(ggcorrplot)
zooplankton_df2 <- cor(zooplankton_df1, use = "complete.obs")
ggcorrplot(zooplankton_df2, lab = TRUE)
```



Interpreting this model, we can overall see that most variable pairs have a weak positive or negative relationship.

- **Total Abundance versus Fluorescence:** This has a *weak positive correlation*, which means that for every unit increase in Fluorescence, the Total Abundance increases at a slightly higher rate.
- **Total Abundance versus Salinity:** This has a *weak negative correlation*, which means that for every unit increase in Salinity, the Total Abundance decreases at a slightly higher rate.
- **Salinity versus Water Temperature:** This has a *weak negative correlation*, which means that for every unit increase in Water Temperature, the Total Abundance decreases at a slightly higher rate.
- **Salinity versus Fluorescence:** This has a *weak positive correlation*, which means that for every unit increase in Fluorescence, the Salinity increases at a slightly higher rate.
- **Fluorescence versus Water Temperature:** This has a *neutral correlation*, which means that the Water Temperature and Fluorescence do not encounter substantial changes on each other.
- **Salinity versus Water Temperature:** This has a *moderate positive correlation*, which means that for every unit increase in Fluorescence, the Total Abundance increases at a moderately higher rate.

Overall we can see that the most correlated variable pair is the **Salinity-Water Temperature** pairing, because of the fact that it carries a correlation value of 0.57. A flaw in this system we can note in our findings is that we do not know the nature of the scatterplot in these relationships (i.e. whether any fanning occurs). This would be an important finding if we embark on creating a Statistical model of some sort, since this will reveal how (mis)leading certain findings may be. We close an eye on this because the objective of this question is to find the pair of variables with the strongest relationship. Note that we also do not include the relationships of the variables to themselves (the diagonal line of correlation 1) because correlations do not work in that manner.