Exploratory data analysis

Contents

```
library(tidyverse)
library(ggplot2)


# While reading from csv file, replacing all "nd" with NA.
# Once, done all numeric columns like Temperature_Kelvin, etc. automatically
# got converted to dbl.

df <- read_csv("~/V2-global-bleaching-and-environmental-data.csv", na = "nd")
head(df)
```

```
## # A tibble: 6 x 62
##   Site_ID Sample_ID Data_Source Latitude_Degrees Longitude_Degrees Ocean_Name
##     <dbl>     <dbl> <chr>                 <dbl>             <dbl> <chr>
## 1    2501  10324336 Donner                23.2              -82.5 Atlantic
## 2    3467  10324754 Donner               -17.6             -150.  Pacific
## 3    1794  10323866 Donner                18.4              -64.6 Atlantic
## 4    8647  10328028 Donner                17.8              -64.6 Atlantic
## 5    8648  10328029 Donner                17.8              -64.6 Atlantic
## 6    2180  10324021 Donner                 9.82             -75.9 Atlantic
## # ... with 56 more variables: Reef_ID <chr>, Realm_Name <chr>,
## #   Ecoregion_Name <chr>, Country_Name <chr>, State_Island_Province_Name <chr>,
## #   City_Town_Name <chr>, Site_Name <chr>, Distance_to_Shore <dbl>,
## #   Exposure <chr>, Turbidity <dbl>, Cyclone_Frequency <dbl>, Date_Day <dbl>,
## #   Date_Month <dbl>, Date_Year <dbl>, Depth_m <dbl>, Substrate_Name <chr>,
## #   Percent_Cover <dbl>, Bleaching_Level <chr>, Percent_Bleaching <dbl>,
## #   ClimSST <dbl>, Temperature_Kelvin <dbl>, Temperature_Mean <dbl>, ...
```

```
colnames(df)
```

```
##  [1] "Site_ID"
##  [2] "Sample_ID"
##  [3] "Data_Source"
##  [4] "Latitude_Degrees"
##  [5] "Longitude_Degrees"
##  [6] "Ocean_Name"
##  [7] "Reef_ID"
##  [8] "Realm_Name"
##  [9] "Ecoregion_Name"
## [10] "Country_Name"
## [11] "State_Island_Province_Name"
## [12] "City_Town_Name"
```

```
## [13] "Site_Name"
## [14] "Distance_to_Shore"
## [15] "Exposure"
## [16] "Turbidity"
## [17] "Cyclone_Frequency"
## [18] "Date_Day"
## [19] "Date_Month"
## [20] "Date_Year"
## [21] "Depth_m"
## [22] "Substrate_Name"
## [23] "Percent_Cover"
## [24] "Bleaching_Level"
## [25] "Percent_Bleaching"
## [26] "ClimSST"
## [27] "Temperature_Kelvin"
## [28] "Temperature_Mean"
## [29] "Temperature_Minimum"
## [30] "Temperature_Maximum"
## [31] "Temperature_Kelvin_Standard_Deviation"
## [32] "Windspeed"
## [33] "SSTA"
## [34] "SSTA_Standard_Deviation"
## [35] "SSTA_Mean"
## [36] "SSTA_Minimum"
## [37] "SSTA_Maximum"
## [38] "SSTA_Frequency"
## [39] "SSTA_Frequency_Standard_Deviation"
## [40] "SSTA_FrequencyMax"
## [41] "SSTA_FrequencyMean"
## [42] "SSTA_DHW"
## [43] "SSTA_DHW_Standard_Deviation"
## [44] "SSTA_DHWMax"
## [45] "SSTA_DHWMean"
## [46] "TSA"
## [47] "TSA_Standard_Deviation"
## [48] "TSA_Minimum"
## [49] "TSA_Maximum"
## [50] "TSA_Mean"
## [51] "TSA_Frequency"
## [52] "TSA_Frequency_Standard_Deviation"
## [53] "TSA_FrequencyMax"
## [54] "TSA_FrequencyMean"
## [55] "TSA_DHW"
## [56] "TSA_DHW_Standard_Deviation"
## [57] "TSA_DHWMax"
## [58] "TSA_DHWMean"
## [59] "Date"
## [60] "Site_Comments"
## [61] "Sample_Comments"
## [62] "Bleaching_Comments"
```

```r
# For our processing from case to case basis, we have to omit all rows with NA
df_month_temp <- df  %>%
  select(Date_Month, Temperature_Kelvin) %>%
```

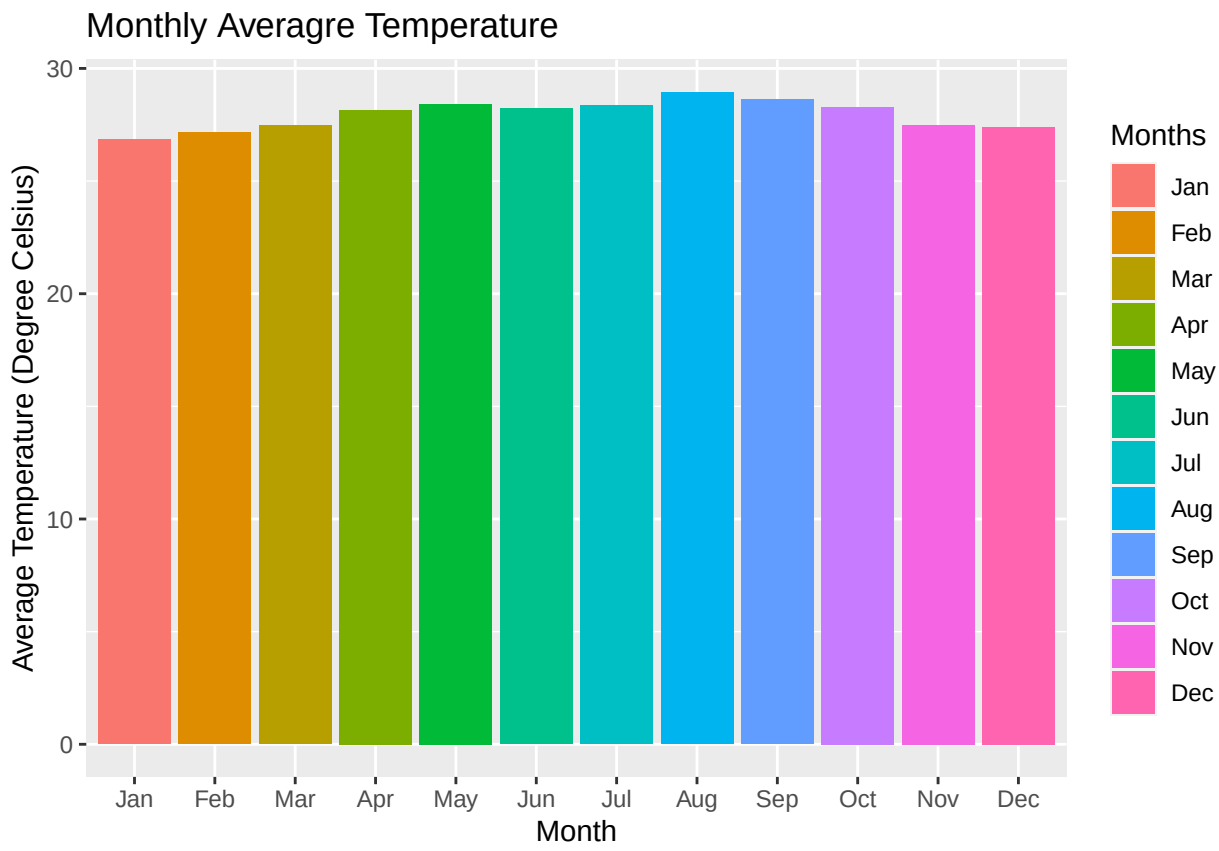
```
  na.omit()

head(df_month_temp)
```

```
## # A tibble: 6 x 2
##   Date_Month Temperature_Kelvin
##        <dbl>              <dbl>
## 1          9               302.
## 2          3               303.
## 3          1               299.
## 4          4               300.
## 5          4               300.
## 6          8               303.
```

```
# In df_month_and_avg_temp dataset, adding another column Months(i.e. Jan, Feb, ...)
df_month_and_avg_temp <- df_month_temp %>%
  group_by(Date_Month) %>%
  summarise(Average_Temperature_Centrigrate = mean(Temperature_Kelvin, na.rm=T) - 273.15) %>%
  mutate(Months = factor(month.abb[Date_Month], levels = month.abb))

print(df_month_and_avg_temp)
```

```
## # A tibble: 12 x 3
##    Date_Month Average_Temperature_Centrigrate Months
##         <dbl>                           <dbl> <fct>
##  1          1                            26.8 Jan
##  2          2                            27.1 Feb
##  3          3                            27.5 Mar
##  4          4                            28.2 Apr
##  5          5                            28.4 May
##  6          6                            28.2 Jun
##  7          7                            28.4 Jul
##  8          8                            28.9 Aug
##  9          9                            28.6 Sep
## 10         10                            28.3 Oct
## 11         11                            27.5 Nov
## 12         12                            27.4 Dec
```

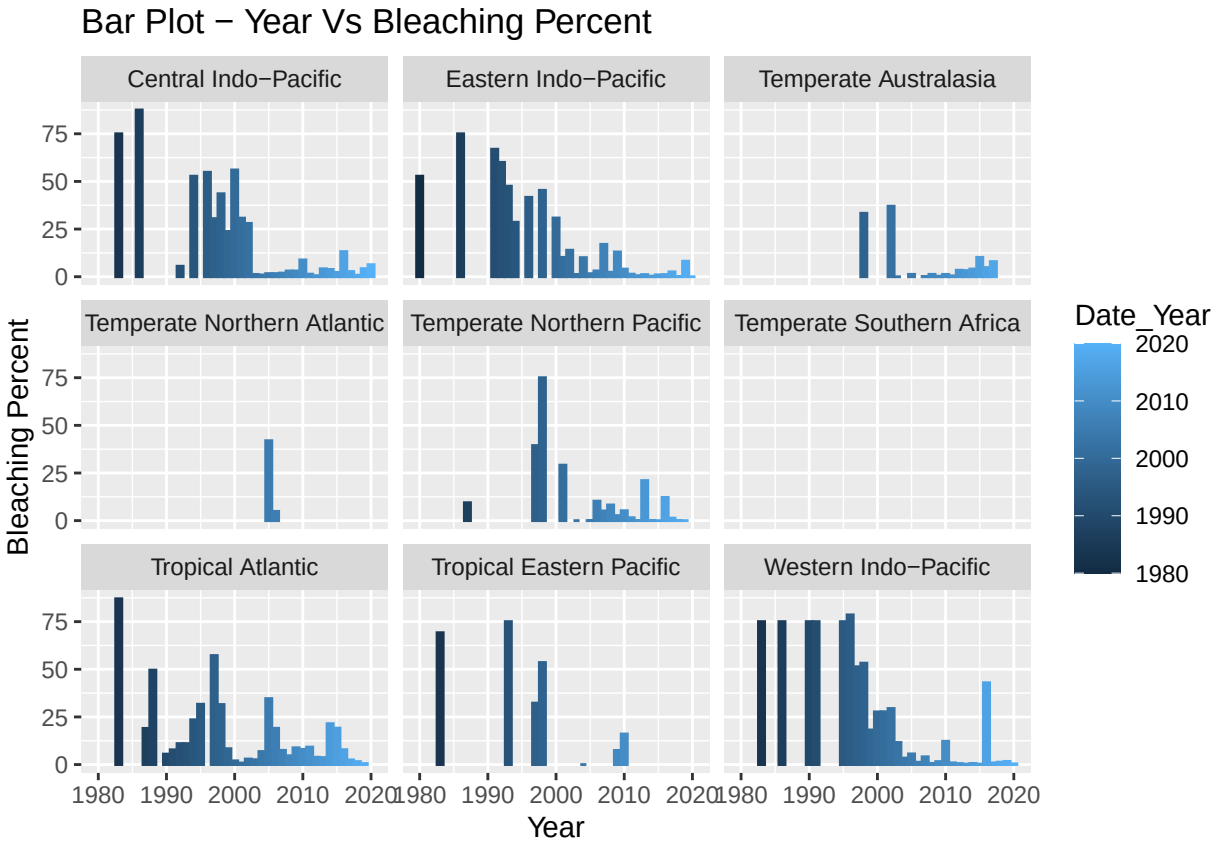
```
# Bar Plot (geom_bar) of Months Vs. Average Temperature
ggplot(df_month_and_avg_temp, aes(x = Months, y = Average_Temperature_Centrigrate, fill = Months)) +
  geom_bar(na.rm=T, stat = "identity", width = 0.90) +
  labs(x = 'Month', y = 'Average Temperature (Degree Celsius)', title = 'Monthly Averagre Temperature')
```



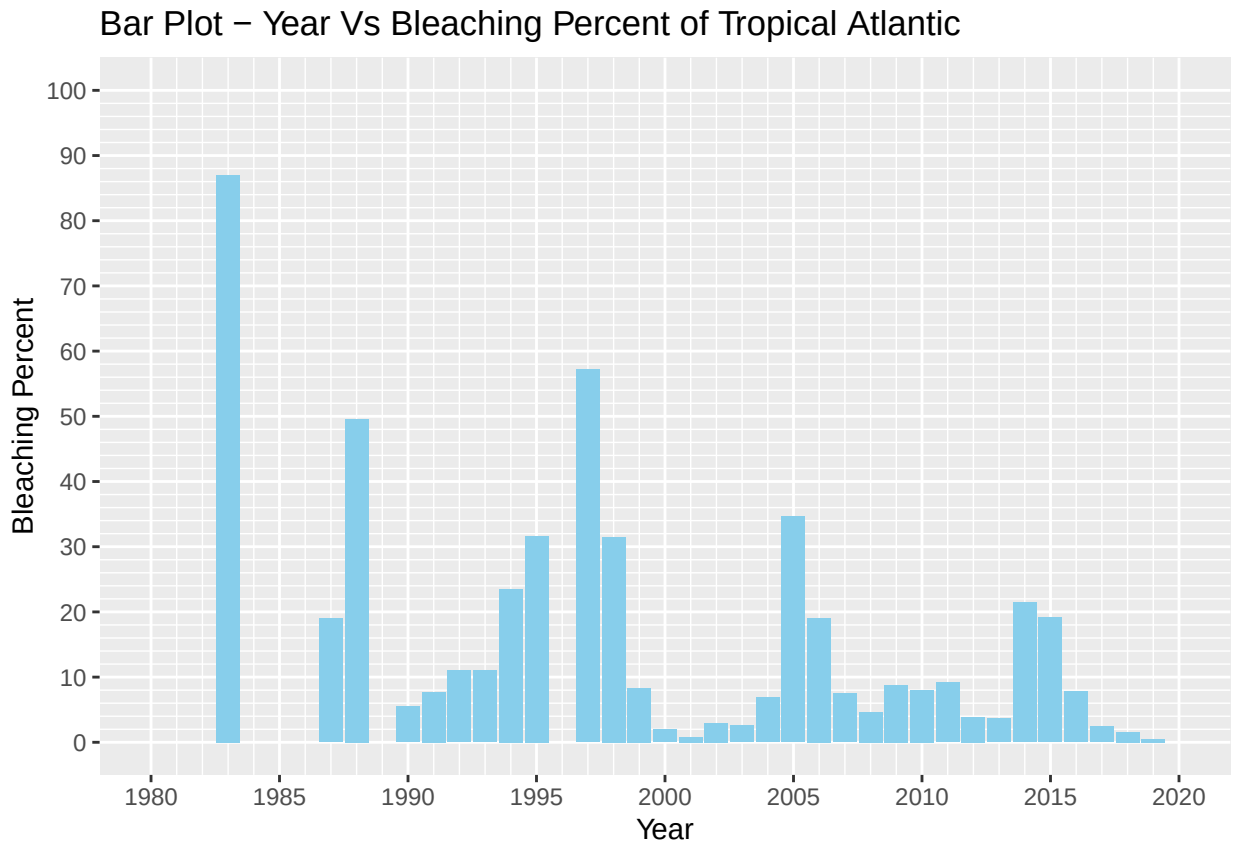

```
# To know all Realm_Name
unique(df$Realm_Name)
```

```
## [1] "Tropical Atlantic"           "Eastern Indo-Pacific"
## [3] "Western Indo-Pacific"        "Central Indo-Pacific"
## [5] "Temperate Northern Pacific"  "Tropical Eastern Pacific"
## [7] "Temperate Australasia"       "Temperate Northern Atlantic"
## [9] "Temperate Southern Africa"
```

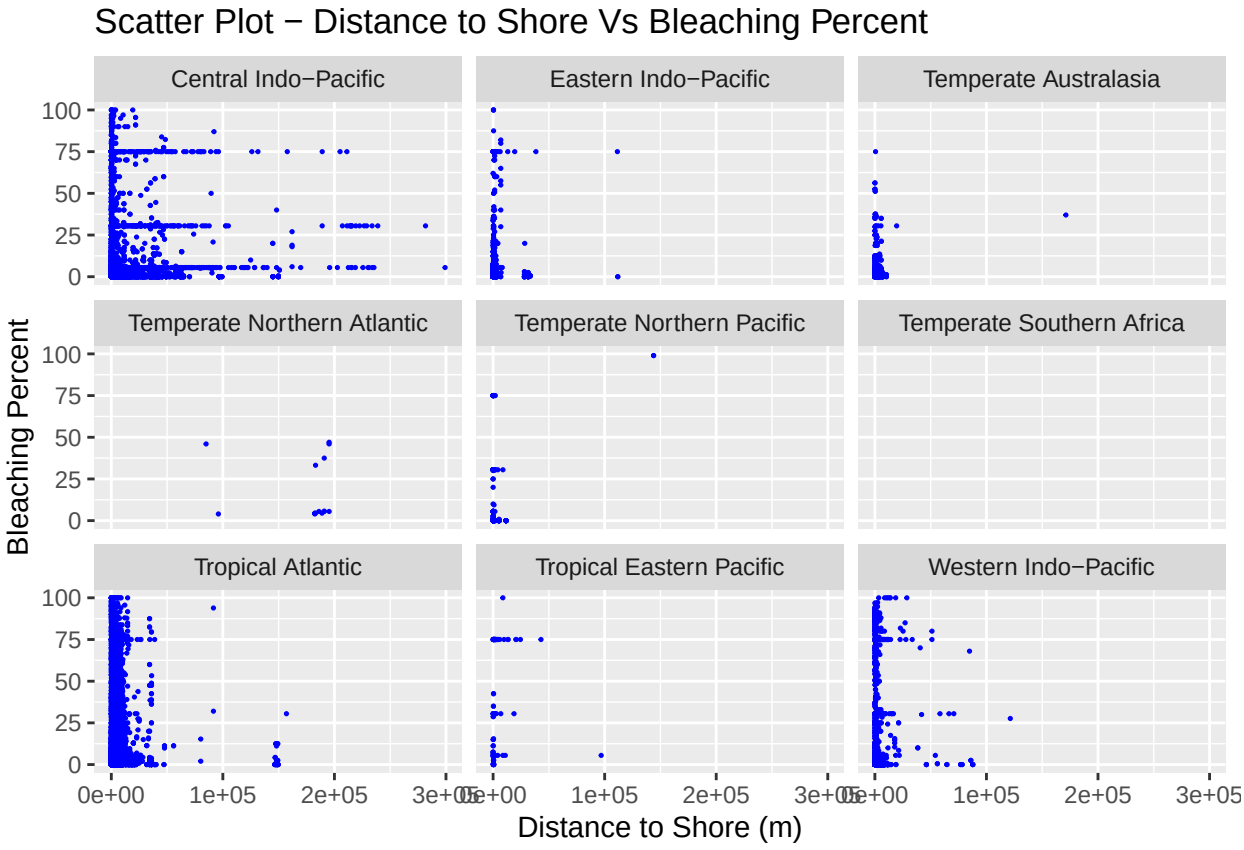
```
# Selecting Realm_Name, Date_Year and Percent_Bleaching.
# Then taking the average (ignoring NAs) of Percent_Bleaching for a combination of Realm_Name and Date_
df_year_percent_bleach_realm <- df %>%
  select(Realm_Name, Date_Year, Percent_Bleaching) %>%
  group_by(Realm_Name, Date_Year) %>%
  summarize(Average_Percent_Bleaching = mean(Percent_Bleaching, na.rm=T))

# Bar Plot (geom_bar) of Percent Bleaching Vs Year for each Realm_Name
ggplot(data=df_year_percent_bleach_realm, aes(x=Date_Year, y=Average_Percent_Bleaching, fill = Date_Yea
  geom_bar(na.rm=T, stat="identity") +
  facet_wrap(~Realm_Name) +
  labs(x = 'Year', y = 'Bleaching Percent', title = 'Bar Plot - Year Vs Bleaching Percent')
```



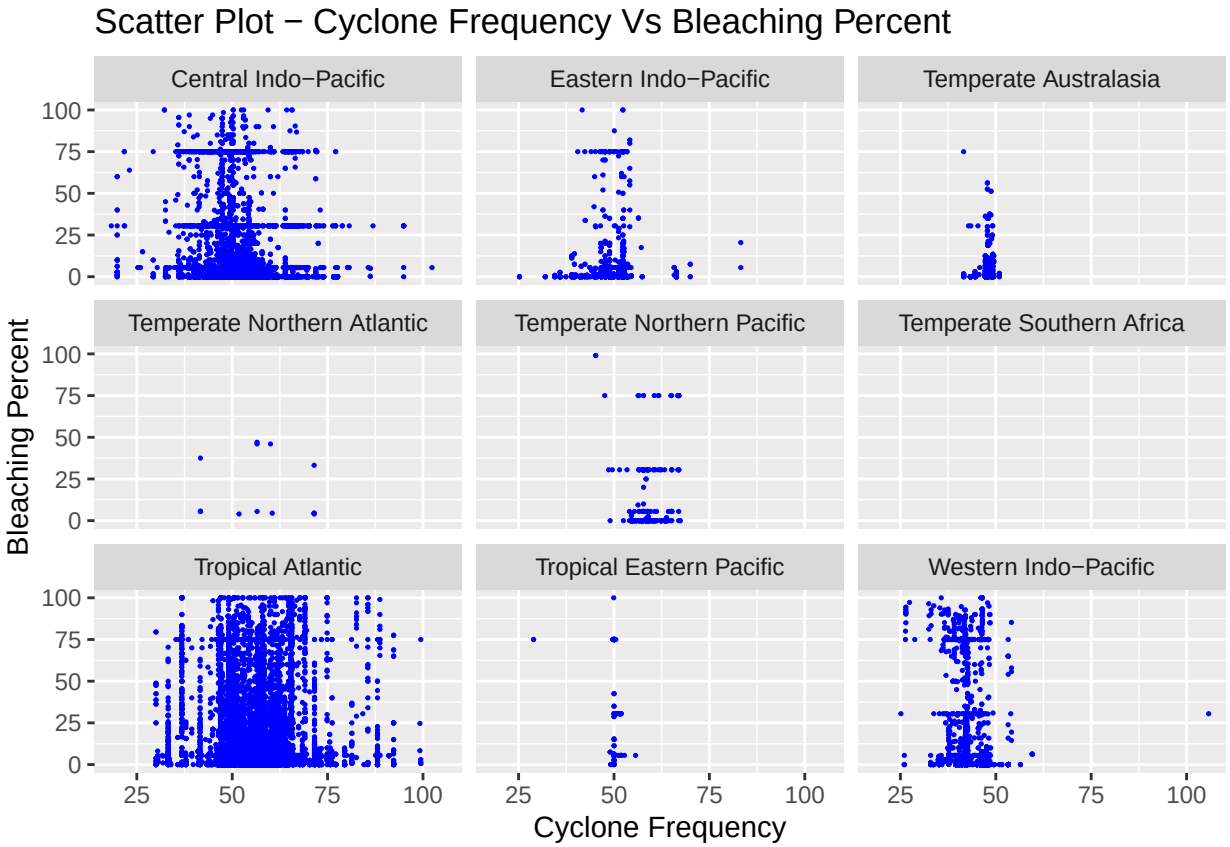

```r
df_Tropical_Atlantic <- df_year_percent_bleach_realm %>%
  filter(Realm_Name == 'Tropical Atlantic')

# Bar Plot (geom_bar) of Percent Bleaching Vs Year for Tropical Atlantic
ggplot(data=df_Tropical_Atlantic, aes(x=Date_Year, y=Average_Percent_Bleaching)) +
  geom_bar(na.rm=T, stat="identity", fill = 'skyblue') +
  scale_x_continuous(limits = c(1980,2020), breaks = seq(1980,2020,5), minor_breaks = seq(1980, 2020,1)
  scale_y_continuous(limits = c(0,100), breaks = seq(0,100,10), minor_breaks = seq(0,100,2)) +
  labs(x = 'Year', y = 'Bleaching Percent', title = 'Bar Plot - Year Vs Bleaching Percent of Tropical A
```



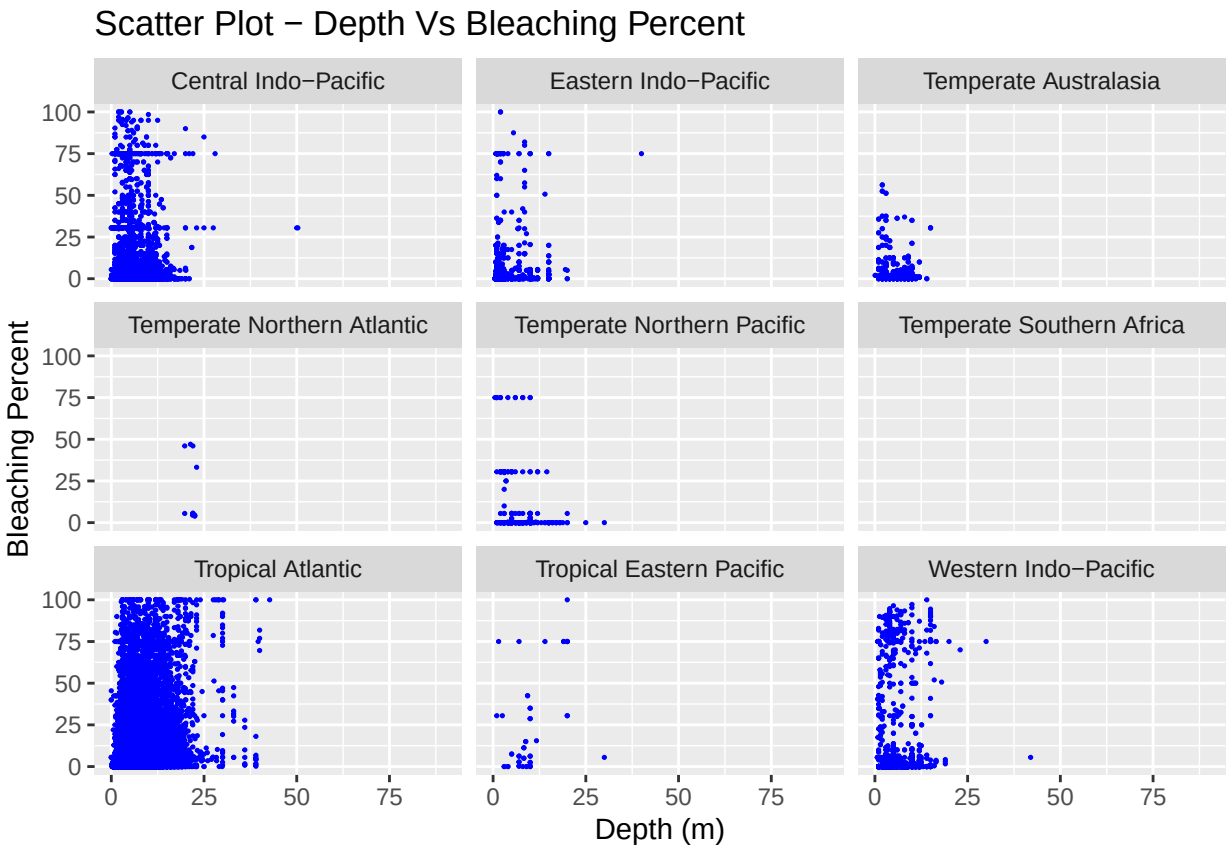

```
# Scatter Plot of Distance_to_Shore Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Distance_to_Shore, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Distance to Shore (m)', y = 'Bleaching Percent', title = 'Scatter Plot - Distance to Shore
```



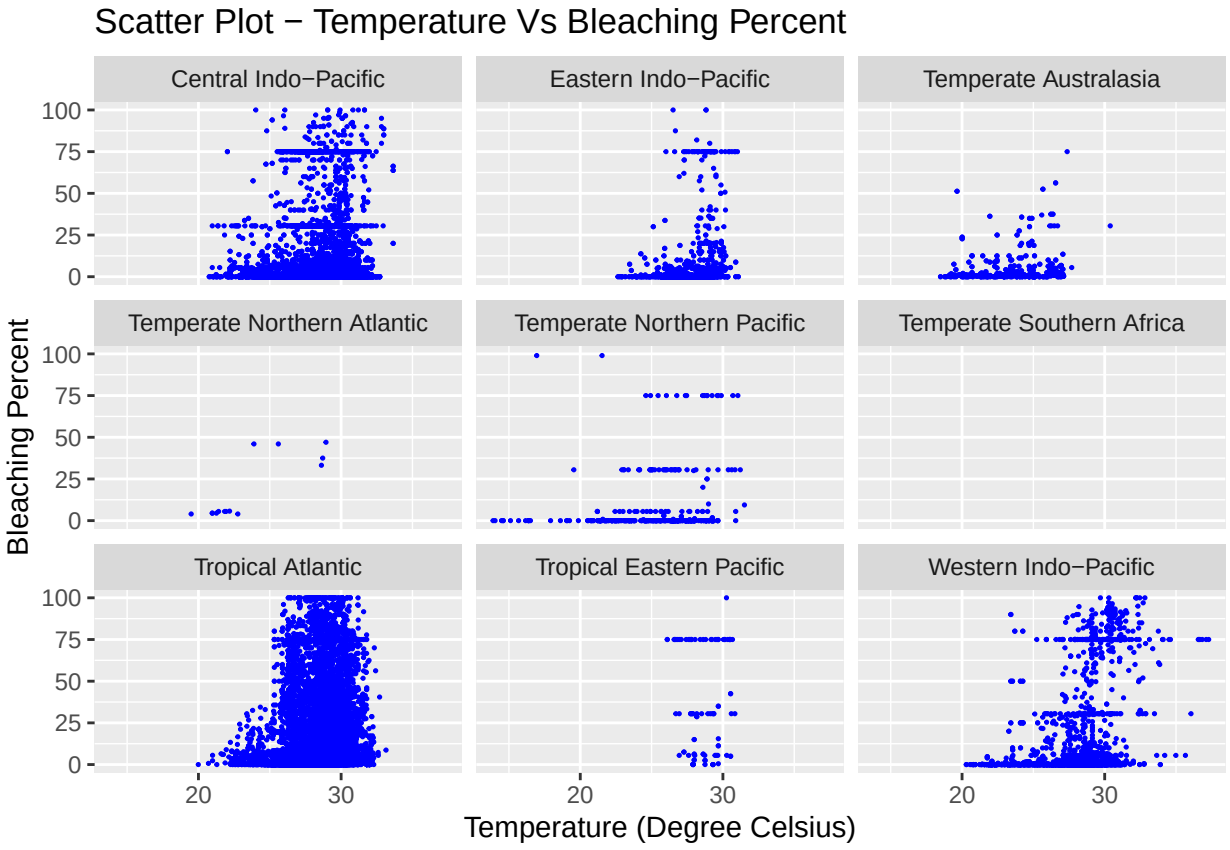

```
# Scatter Plot of Cyclone_Frequency Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Cyclone_Frequency, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Cyclone Frequency', y = 'Bleaching Percent', title = 'Scatter Plot - Cyclone Frequency Vs B
```



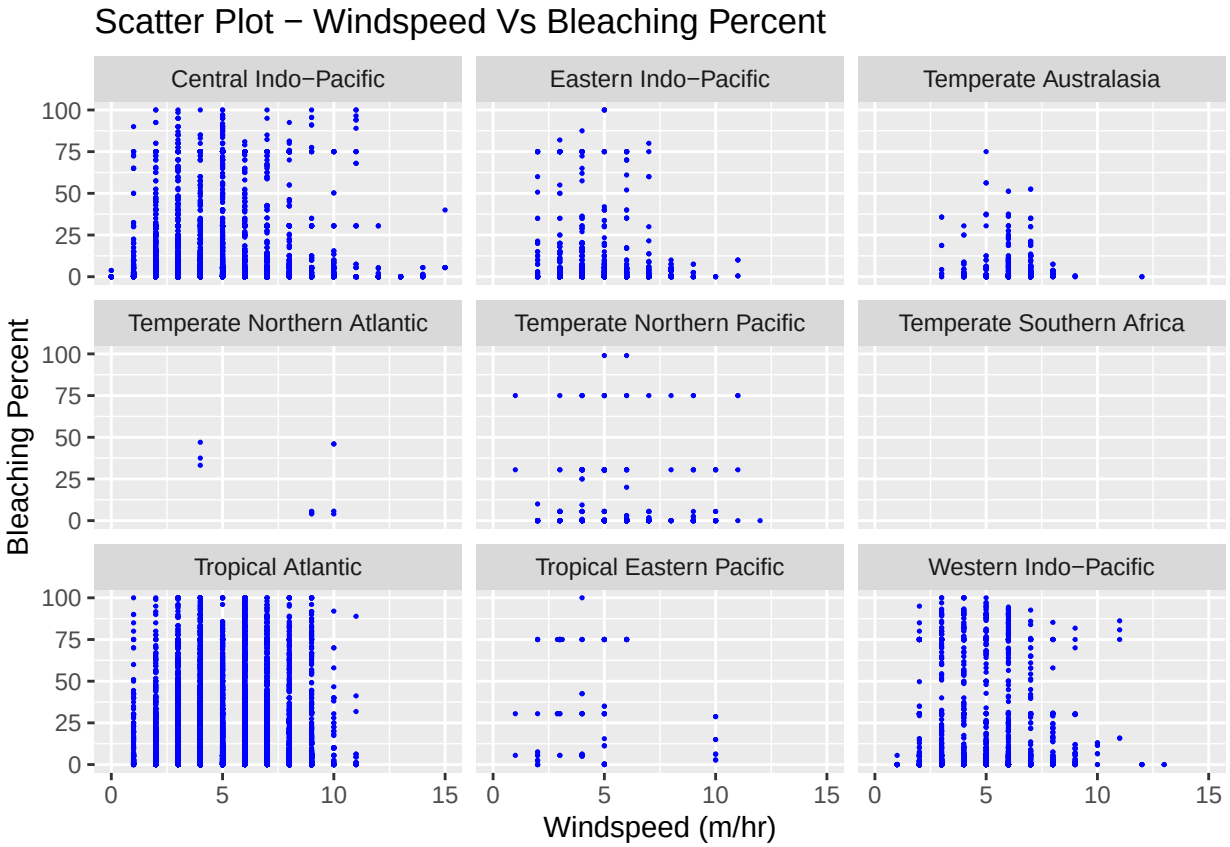

```
# Scatter Plot of Depth_m Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Depth_m, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Depth (m)', y = 'Bleaching Percent', title = 'Scatter Plot - Depth Vs Bleaching Percent')
```



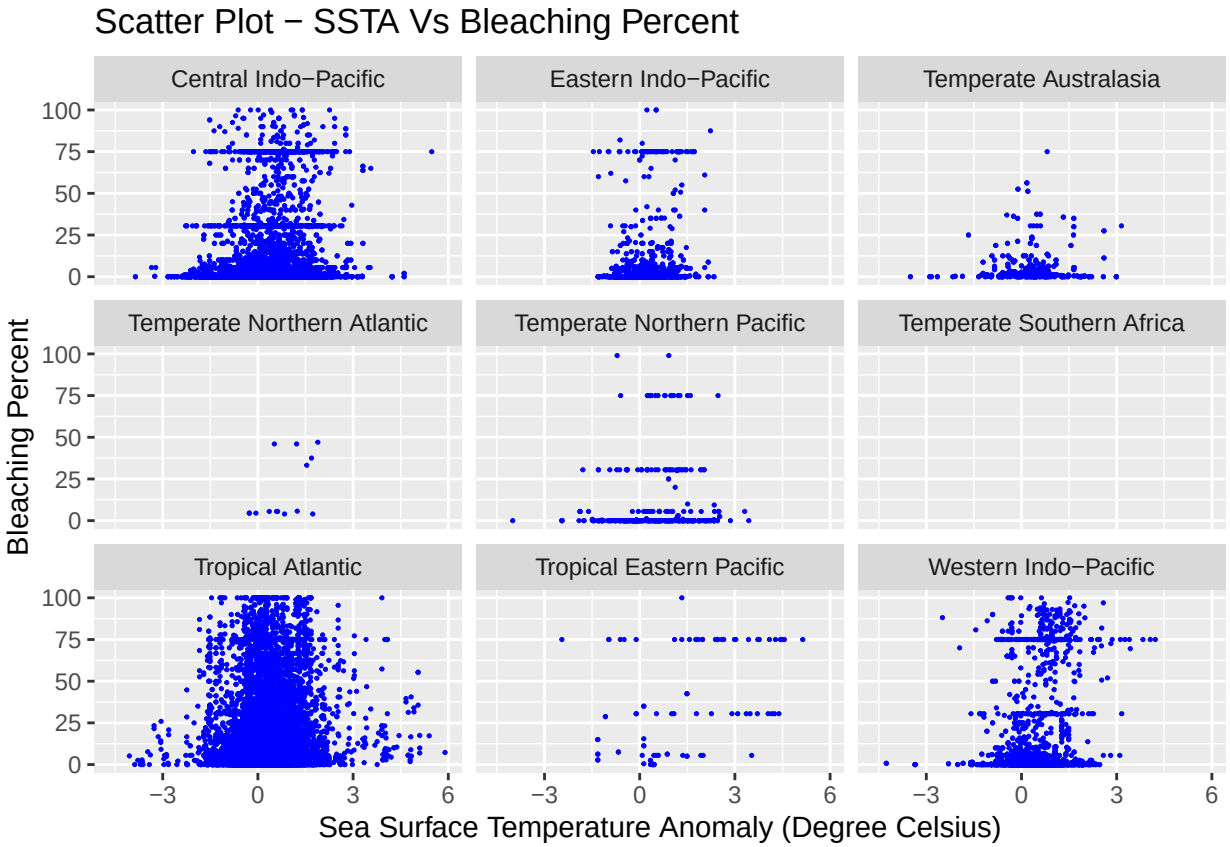

```
# Scatter Plot of Temperature Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Temperature_Kelvin - 273.15, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Temperature (Degree Celsius)', y = 'Bleaching Percent', title = 'Scatter Plot - Temperature
```



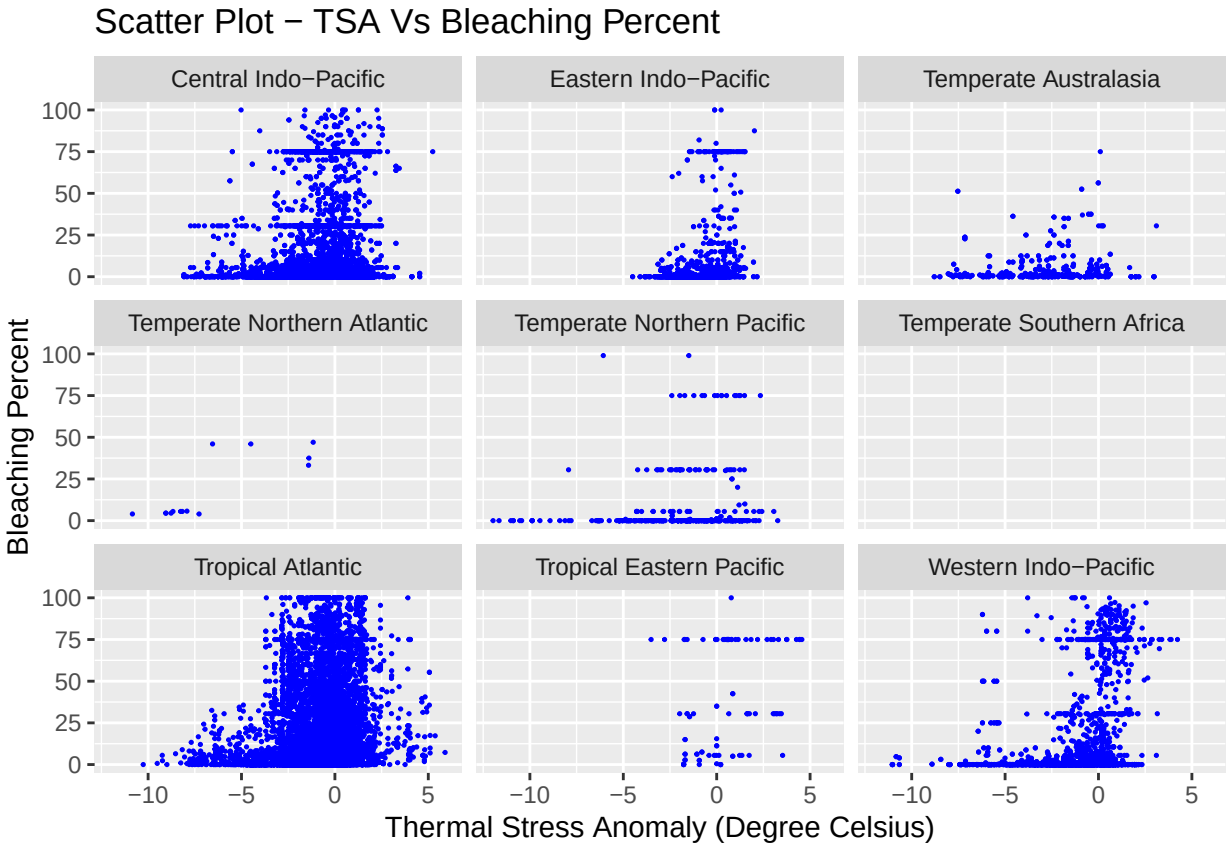

```
# Scatter Plot of Windspeed Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Windspeed, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Windspeed (m/hr)', y = 'Bleaching Percent', title = 'Scatter Plot - Windspeed Vs Bleaching
```



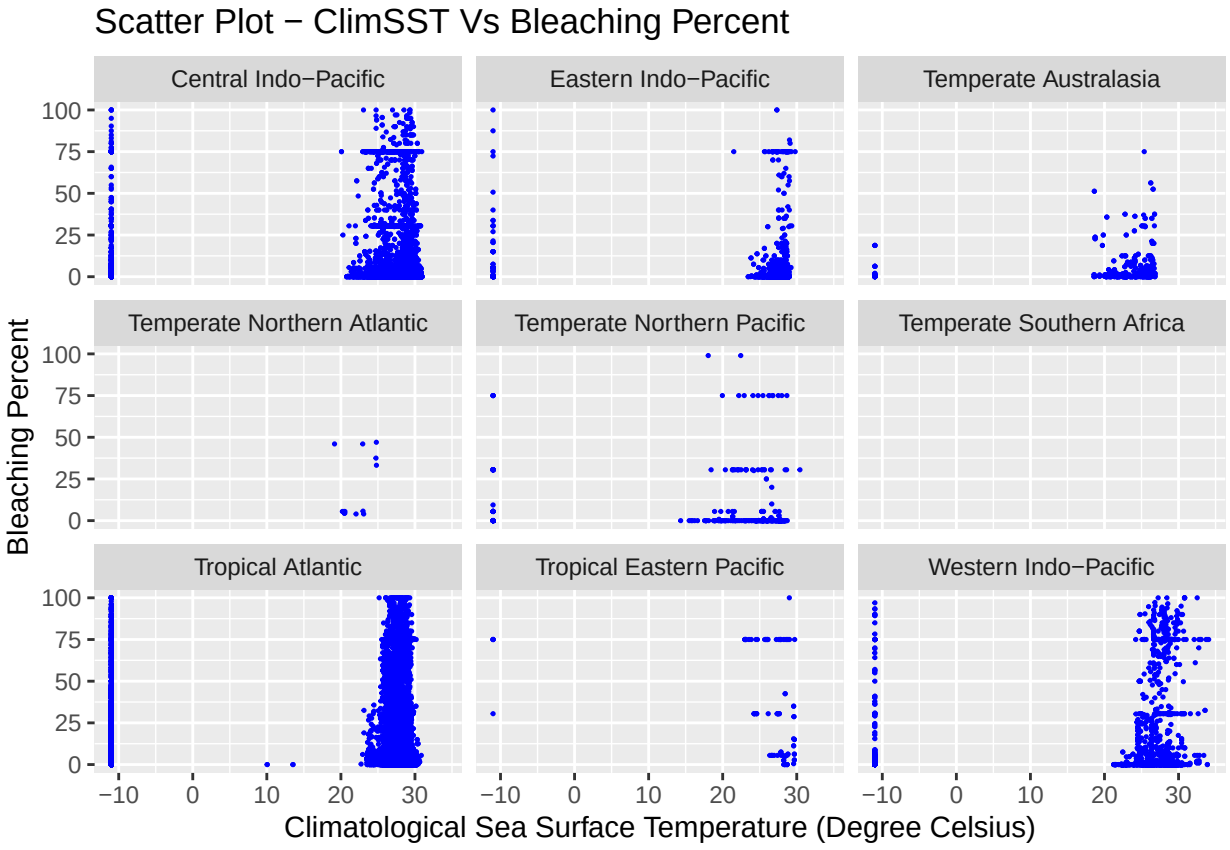

```
# Scatter Plot of SSTA Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=SSTA, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Sea Surface Temperature Anomaly (Degree Celsius)', y = 'Bleaching Percent', title = 'Scatte
```



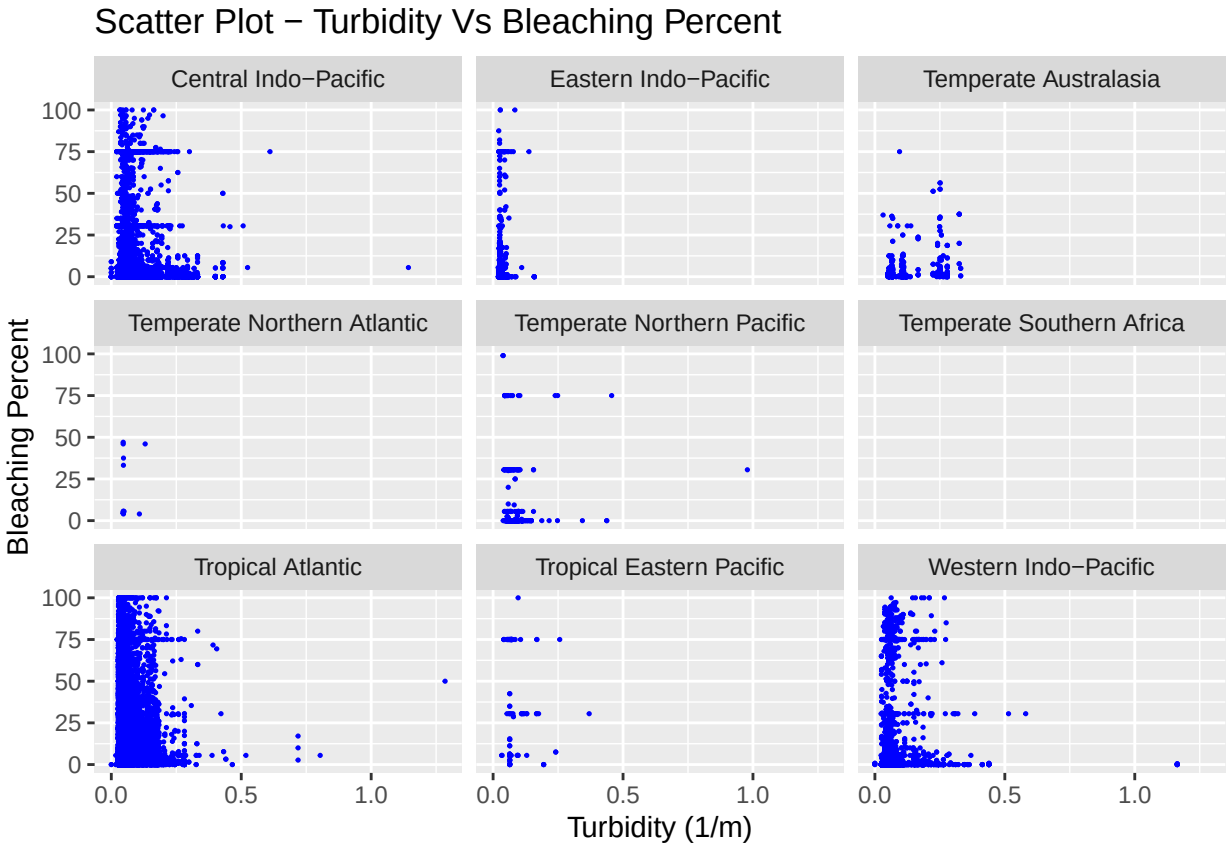

```
# Scatter Plot of TSA Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=TSA, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Thermal Stress Anomaly (Degree Celsius)', y = 'Bleaching Percent', title = 'Scatter Plot - '
```



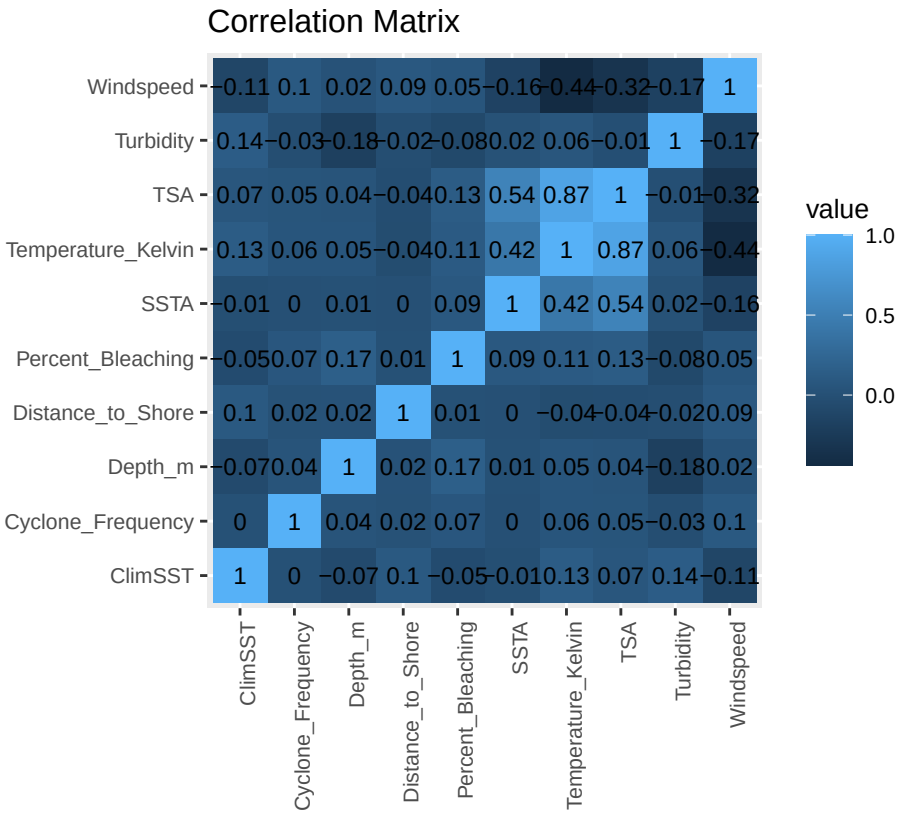

```
# Scatter Plot of ClimSST Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=ClimSST - 273.15, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Climatological Sea Surface Temperature (Degree Celsius)', y = 'Bleaching Percent', title =
```




```
# Scatter Plot of Turbidity Vs Percent_Bleaching for all Realm_Name
ggplot(df, aes(x=Turbidity, y=Percent_Bleaching)) +
  geom_point(na.rm=T, color = 'blue', size = 0.25) +
  facet_wrap(~Realm_Name) +
  labs(x = 'Turbidity (1/m)', y = 'Bleaching Percent', title = 'Scatter Plot - Turbidity Vs Bleaching P
```




```
# Visualizing Correlation Matrix for important columns
df_cor <- df %>% select(Percent_Bleaching, Distance_to_Shore,
                        Cyclone_Frequency, Depth_m, Temperature_Kelvin,
                        Windspeed, SSTA, TSA, ClimSST, Turbidity)

df_cor %>% na.omit() %>% select_if(is.numeric) %>% cor %>%
  as.data.frame %>% rownames_to_column %>% pivot_longer(-1) %>%
  ggplot(aes(rowname, name, fill=value)) + geom_tile() +
  geom_text(aes(label=round(value,2)), size=3.3) +
  theme(text = element_text(size=10), axis.text.x = element_text(angle=90, hjust=1)) +
  labs(x = '', y = '', title = 'Correlation Matrix') +
  coord_fixed()
```



Percent Bleaching, is positively correlated with Depth_m, TSA, Temperature_Kelvin, SSTA, Cyclone_Frequency, Windspeed and Distance_to_Shore in this order (i.e. Depth_m is maximum and Distance_to_Shore is minimum).

Percent Bleaching is negatively correlated with Turbidity and ClimSST (Turbidity is negatively maximum and ClimSST is negatively minimum).