Project_Step1

Elif İlayda Güntürk, Selin Ergül, Seda Metin

2023-12-30

Data Set and Details

- CalCOFI is one of the world's longest-running and most comprehensive oceanographic datasets [1]. CalCOFI collects hydrographic and biological data on regular cruises. The database we have selected contains measurements made with bottles containing seawater samples collected at CalCOFI stations from 1949 to the present day [1]. Oceanographic data includes temperature, salinity, dissolved oxygen, chlorophyll-a, nutrients and much more.
- The data has 74 variables, including the salinity scale of ocean water (Salnty), the depth at which the measurement was made (Depth), the number of bottles and casts used by CalCOFI, such as potential density, oxygen saturation, chlorophyll-a and pheopigment measurements (Cst_Cnt, Btl_Cnt), the components in the water and their values under pressure, the temperature of the water (T_degC), pH values (pH1-2).

```
bottle <- read.csv("bottle.csv", sep = ",")

summary(bottle)
```

```
##     Cst_Cnt         Btl_Cnt          Sta_ID            Depth_ID
##  Min.   :    1   Min.   :     1   Length:864863      Length:864863
##  1st Qu.: 8269   1st Qu.:216217   Class :character   Class :character
##  Median :16848   Median :432432   Mode  :character   Mode  :character
##  Mean   :17139   Mean   :432432
##  3rd Qu.:26557   3rd Qu.:648648
##  Max.   :34404   Max.   :864863
##
##      Depthm           T_degC          Salnty          O2ml_L
##  Min.   :   0.0   Min.   : 1.44   Min.   :28.43   Min.   :-0.01
##  1st Qu.:  46.0   1st Qu.: 7.68   1st Qu.:33.49   1st Qu.: 1.36
##  Median : 125.0   Median :10.06   Median :33.86   Median : 3.44
##  Mean   : 226.8   Mean   :10.80   Mean   :33.84   Mean   : 3.39
##  3rd Qu.: 300.0   3rd Qu.:13.88   3rd Qu.:34.20   3rd Qu.: 5.50
##  Max.   :5351.0   Max.   :31.14   Max.   :37.03   Max.   :11.13
##                   NA's   :10963   NA's   :47354   NA's   :168662
##      STheta           O2Sat         Oxy_µmol.Kg         BtlNum
##  Min.   : 20.93   Min.   : -0.1    Min.   : -0.43   Min.   : 0.0
##  1st Qu.: 24.96   1st Qu.: 21.1    1st Qu.: 60.92   1st Qu.: 5.0
##  Median : 26.00   Median : 54.4    Median :151.06   Median :10.0
##  Mean   : 25.82   Mean   : 57.1    Mean   :148.81   Mean   :10.5
##  3rd Qu.: 26.65   3rd Qu.: 97.6    3rd Qu.:240.38   3rd Qu.:16.0
##  Max.   :250.78   Max.   :214.1    Max.   :485.70   Max.   :25.0
##  NA's   :52689    NA's   :203589   NA's   :203595   NA's   :746196
##      RecInd        T_prec          T_qual           S_prec
##  Min.   :3.0   Min.   :1.000   Min.   :6.0      Min.   :2.00
##  1st Qu.:3.0   1st Qu.:2.000   1st Qu.:6.0      1st Qu.:2.00
##  Median :3.0   Median :2.000   Median :6.0      Median :3.00
##  Mean   :4.7   Mean   :2.017   Mean   :7.5      Mean   :2.72
##  3rd Qu.:7.0   3rd Qu.:2.000   3rd Qu.:9.0      3rd Qu.:3.00
##  Max.   :7.0   Max.   :3.000   Max.   :9.0      Max.   :3.00
##                NA's   :10963   NA's   :841736   NA's   :47354
##      S_qual           P_qual           O_qual           SThtaq
##  Min.   :6.0      Min.   :6        Min.   :6.0      Min.   :6.0
##  1st Qu.:6.0      1st Qu.:9        1st Qu.:9.0      1st Qu.:9.0
##  Median :9.0      Median :9        Median :9.0      Median :9.0
##  Mean   :7.9      Mean   :9        Mean   :8.8      Mean   :8.5
##  3rd Qu.:9.0      3rd Qu.:9        3rd Qu.:9.0      3rd Qu.:9.0
##  Max.   :9.0      Max.   :9        Max.   :9.0      Max.   :9.0
##  NA's   :789949   NA's   :191108   NA's   :680187   NA's   :799040
##      O2Satq           ChlorA           Chlqua           Phaeop
##  Min.   :2.0      Min.   : 0.0     Min.   :8        Min.   :-3.9
##  1st Qu.:9.0      1st Qu.: 0.0     1st Qu.:9        1st Qu.: 0.0
##  Median :9.0      Median : 0.2     Median :9        Median : 0.1
##  Mean   :8.8      Mean   : 0.5     Mean   :9        Mean   : 0.2
##  3rd Qu.:9.0      3rd Qu.: 0.4     3rd Qu.:9        3rd Qu.: 0.2
##  Max.   :9.0      Max.   :66.1     Max.   :9        Max.   :65.3
##  NA's   :647066   NA's   :639591   NA's   :225697   NA's   :639592
##      Phaqua           PO4uM            PO4q           SiO3uM
##  Min.   :8        Min.   :0.0      Min.   :4        Min.   :  0.0
##  1st Qu.:9        1st Qu.:0.5      1st Qu.:9        1st Qu.:  3.1
```

```
##  Median :9        Median :1.6      Median :9        Median : 18.0
##  Mean   :9        Mean   :1.6      Mean   :9        Mean   : 26.6
##  3rd Qu.:9        3rd Qu.:2.5      3rd Qu.:9        3rd Qu.: 41.5
##  Max.   :9        Max.   :5.2      Max.   :9        Max.   :196.0
##  NA's   :225693   NA's   :451546   NA's   :413077   NA's   :510772
##      SiO3qu           NO2uM             NO2q             NO3uM
##  Min.   :4        Min.   :0.0      Min.   :4        Min.   :-0.4
##  1st Qu.:9        1st Qu.:0.0      1st Qu.:9        1st Qu.: 0.6
##  Median :9        Median :0.0      Median :9        Median :18.1
##  Mean   :9        Mean   :0.0      Mean   :9        Mean   :17.3
##  3rd Qu.:9        3rd Qu.:0.0      3rd Qu.:9        3rd Qu.:30.0
##  Max.   :9        Max.   :8.2      Max.   :9        Max.   :95.0
##  NA's   :353997   NA's   :527287   NA's   :335389   NA's   :527460
##       NO3q            NH3uM             NH3q           C14As1
##  Min.   :4        Min.   : 0.0     Min.   :4.00    Min.   : -0.2
##  1st Qu.:9        1st Qu.: 0.0     1st Qu.:9.00    1st Qu.:  0.9
##  Median :9        Median : 0.0     Median :9.00    Median :  2.6
##  Mean   :9        Mean   : 0.1     Mean   :8.95    Mean   :  9.8
##  3rd Qu.:9        3rd Qu.: 0.1     3rd Qu.:9.00    3rd Qu.:  8.0
##  Max.   :9        Max.   :15.6     Max.   :9.00    Max.   :584.5
##  NA's   :334930   NA's   :799901   NA's   :56564   NA's   :850431
##      C14A1p           C14A1q          C14As2           C14A2p
##  Min.   :1.0      Min.   :8       Min.   : -0.2    Min.   :1.0
##  1st Qu.:1.0      1st Qu.:9       1st Qu.:  0.9    1st Qu.:1.0
##  Median :1.0      Median :9       Median :  2.6    Median :1.0
##  Mean   :1.3      Mean   :9       Mean   :  9.8    Mean   :1.3
##  3rd Qu.:2.0      3rd Qu.:9       3rd Qu.:  8.1    3rd Qu.:2.0
##  Max.   :2.0      Max.   :9       Max.   :948.3    Max.   :2.0
##  NA's   :852103   NA's   :16258   NA's   :850449   NA's   :852121
##      C14A2q          DarkAs           DarkAp           DarkAq
##  Min.   :8       Min.   :0.0      Min.   :1        Min.   :8
##  1st Qu.:9       1st Qu.:0.1      1st Qu.:2        1st Qu.:9
##  Median :9       Median :0.1      Median :2        Median :9
##  Mean   :9       Mean   :0.2      Mean   :2        Mean   :9
##  3rd Qu.:9       3rd Qu.:0.2      3rd Qu.:2        3rd Qu.:9
##  Max.   :9       Max.   :6.9      Max.   :2        Max.   :9
##  NA's   :16240   NA's   :842214   NA's   :844406   NA's   :24423
##      MeanAs           MeanAp           MeanAq         IncTim
##  Min.   : -0.2    Min.   :1.0      Min.   :8       Length:864863
##  1st Qu.:  1.0    1st Qu.:1.0      1st Qu.:9       Class :character
##  Median :  2.5    Median :1.0      Median :9       Mode  :character
##  Mean   :  8.4    Mean   :1.3      Mean   :9
##  3rd Qu.:  7.0    3rd Qu.:2.0      3rd Qu.:9
##  Max.   :948.3    Max.   :2.0      Max.   :9
##  NA's   :842213   NA's   :844406   NA's   :24424
##      LightP          R_Depth           R_TEMP         R_POTEMP
##  Min.   : 0.0     Min.   :   0.0   Min.   : 1.44   Min.   : 0.00
##  1st Qu.: 0.3     1st Qu.:  46.0   1st Qu.: 7.68   1st Qu.: 7.74
##  Median : 1.8     Median : 125.0   Median :10.06   Median :10.10
##  Mean   :18.4     Mean   : 226.8   Mean   :10.80   Mean   :10.84
##  3rd Qu.:24.0     3rd Qu.: 300.0   3rd Qu.:13.88   3rd Qu.:13.92
```

```
##  Max.   :99.9     Max.   :5351.0   Max.   :31.14   Max.   :31.14
##  NA's   :846212                    NA's   :10963   NA's   :46047
##    R_SALINITY       R_SIGMA           R_SVA          R_DYNHT
##  Min.   : 4.57   Min.   : 20.93   Min.   :  0.4   Min.   :0.00
##  1st Qu.:33.49   1st Qu.: 24.96   1st Qu.:143.7   1st Qu.:0.13
##  Median :33.86   Median : 25.99   Median :203.2   Median :0.34
##  Mean   :33.84   Mean   : 25.81   Mean   :220.9   Mean   :0.43
##  3rd Qu.:34.20   3rd Qu.: 26.64   3rd Qu.:299.8   3rd Qu.:0.64
##  Max.   :37.03   Max.   :250.78   Max.   :683.4   Max.   :3.88
##  NA's   :47354   NA's   :52856    NA's   :52771   NA's   :46657
##       R_O2           R_O2Sat           R_SIO3           R_PO4
##  Min.   :-0.01    Min.   : -0.10   Min.   :  0.0    Min.   :0.0
##  1st Qu.: 1.36    1st Qu.: 21.20   1st Qu.:  3.1    1st Qu.:0.5
##  Median : 3.44    Median : 54.50   Median : 18.0    Median :1.6
##  Mean   : 3.39    Mean   : 57.19   Mean   : 26.6    Mean   :1.6
##  3rd Qu.: 5.50    3rd Qu.: 97.60   3rd Qu.: 41.5    3rd Qu.:2.5
##  Max.   :11.13    Max.   :214.10   Max.   :196.0    Max.   :5.2
##  NA's   :168662   NA's   :198415   NA's   :510764   NA's   :451538
##      R_NO3            R_NO2            R_NH4            R_CHLA
##  Min.   :-0.4     Min.   :0.0      Min.   : 0.0     Min.   : 0.0
##  1st Qu.: 0.6     1st Qu.:0.0      1st Qu.: 0.0     1st Qu.: 0.0
##  Median :18.1     Median :0.0      Median : 0.0     Median : 0.2
##  Mean   :17.3     Mean   :0.0      Mean   : 0.1     Mean   : 0.5
##  3rd Qu.:30.0     3rd Qu.:0.0      3rd Qu.: 0.1     3rd Qu.: 0.4
##  Max.   :95.0     Max.   :8.2      Max.   :15.6     Max.   :66.1
##  NA's   :527452   NA's   :527279   NA's   :799881   NA's   :639587
##     R_PHAEO           R_PRES           R_SAMP           DIC1
##  Min.   :-3.9     Min.   :   0.0   Min.   :  0.0    Min.   :1949
##  1st Qu.: 0.0     1st Qu.:  46.0   1st Qu.:200.0    1st Qu.:2028
##  Median : 0.1     Median : 126.0   Median :206.0    Median :2171
##  Mean   : 0.2     Mean   : 228.4   Mean   :162.1    Mean   :2153
##  3rd Qu.: 0.2     3rd Qu.: 302.0   3rd Qu.:214.0    3rd Qu.:2254
##  Max.   :65.3     Max.   :5458.0   Max.   :424.0    Max.   :2368
##  NA's   :639588                    NA's   :742857   NA's   :862864
##       DIC2             TA1              TA2              pH2
##  Min.   :1969     Min.   :2182     Min.   :2198     Min.   :7.9
##  1st Qu.:2009     1st Qu.:2230     1st Qu.:2229     1st Qu.:7.9
##  Median :2266     Median :2244     Median :2248     Median :7.9
##  Mean   :2168     Mean   :2256     Mean   :2279     Mean   :7.9
##  3rd Qu.:2316     3rd Qu.:2278     3rd Qu.:2316     3rd Qu.:8.0
##  Max.   :2364     Max.   :2435     Max.   :2437     Max.   :8.0
##  NA's   :864639   NA's   :862779   NA's   :864629   NA's   :864853
##       pH1         DIC.Quality.Comment
##  Min.   :7.6      Length:864863
##  1st Qu.:7.9      Class :character
##  Median :7.9      Mode  :character
##  Mean   :7.9
##  3rd Qu.:8.0
##  Max.   :8.0
##  NA's   :864779
```

Data Summary

As we mentioned, 74 variables and 800k objects. It was difficult to conduct meaningful data analysis. Therefore, in order to facilitate our analysis, we provide meaningful and interconnected variables.

We collected it into a data frame. Variables in this data frame are T_degC, Salnty, O2ml_L, STheta, Depthm.

- **T_degC**: In this dataset, the maximum water temperature in degrees Celsius is 31.14 and the lowest temperature is 1.44. There are 10963 NAs in this column.
- **Salinity (Practical Salinity Scale 1978)**: Salinity was measured in this column. The lowest is 28.43 and the highest is 37.03. The number of invalid data is 47354.
- **O2ml_L**: Milliliters oxygen per liter of seawater measurements are given in this column. The lowest value is -0.1 and the maximum value is 11.13. The NA number is 168662.
- **STheta**: This column contains the measurement of Potential Density (Sigma Theta), Kg/M³. The lowest value is 20.93 and the highest value is 250.78. The total number of NAs is 52689.
- **Depthm:** Depth in meters is measured in this column. The lowest value is 0 and the highest value is 5351.0. There is no NA value.

The reason for selecting these 5 variable in the data set is that they are compatible with each other. We anticipated that these data would generate meaningful graphs in regression and analysis.

```
column_list <- c("T_degC", "Salnty", "O2ml_L", "STheta", "Depthm")

bottle_new <- select(bottle, column_list)
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## ℹ Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(column_list)
##
##   # Now:
##   data %>% select(all_of(column_list))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

- head(): Displays the first few lines of a data frame or vector. By default, the 'head()' function displays the first six rows of the dataset, but how many rows can be displayed optionally.

```
###We had seen the first 6 rows
head(bottle_new)
```

```
##   T_degC Salnty O2ml_L STheta Depthm
## 1  10.50 33.440     NA 25.649      0
## 2  10.46 33.440     NA 25.656      8
## 3  10.46 33.437     NA 25.654     10
## 4  10.45 33.420     NA 25.643     19
## 5  10.45 33.421     NA 25.643     20
## 6  10.45 33.431     NA 25.651     30
```

- tail(): Displays the last few lines of a data frame or vector. By default, the 'tail()' function displays the last six rows of the dataset, but specifying how many rows to display is optional.

```
###We had seen the last 6 rows
tail(bottle_new)
```

```
##        T_degC  Salnty O2ml_L   STheta Depthm
## 864858  5.818 34.2382  0.366 26.98477    521
## 864859 18.744 33.4083  5.805 23.87055      0
## 864860 18.744 33.4083  5.805 23.87072      2
## 864861 18.692 33.4150  5.796 23.88911      5
## 864862 18.161 33.4062  5.816 24.01426     10
## 864863 17.533 33.3880  5.774 24.15297     15
```

Data Cleaning

- Data cleaning is the process of identifying, correcting or removing erroneous, missing or inconsistent data in a data set. This process is important to achieve accurate results in data analysis because NA (not available) values can negatively affect the accuracy and reliability of the analysis. The first step of data cleaning helps us identify missing values using the is.na() function. In addition, the sum(is.na()) function is used to find the total number of NA values.The na.omit() function removes NA values from the data dataset. The output of na.omit(data) gives a new dataset containing non-NA values.
- The duplicate function is used to ignore repeated values in the data set. This eliminates the amount of deviation in the mean value because repeated data can affect the mean value.

```
anyDuplicated(bottle_new)
```

```
## [1] 822
```

```
bottle_new <- distinct(bottle_new, .keep_all = TRUE)
```

- When we came to NA omitting, which is another step of our data cleaning, we learned that our NA number is 2945102 with the function sum(is.na(bottle)) to learn the number of NAs in our data set. At first, we thought of replacing the NAs with means, but we preferred to omit the NAs because we had too much data.

```
bottle_new <- na.omit(bottle_new)
```

One Variable Analysis

- One variable analysis is used to visually understand the distribution, central tendency and variability of a single variable in a data set. In this dataset, we chose to use T_degC, Salnty, O2ml_L, Depthm variables. Since these variables have numeric values, we visualized the data using histograms. We created a frequency table within 7 variables to determine the distribution of variables in the data set, how often a particular value occurs, or the relationships between different values of the variable.

1.Degrees of Temperature Graph

```
ggplot(bottle_new,aes(x=T_degC))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Degrees of Temperature")+
  xlab("Degree Celsius")+
  ylab("Number of Measurment")
```

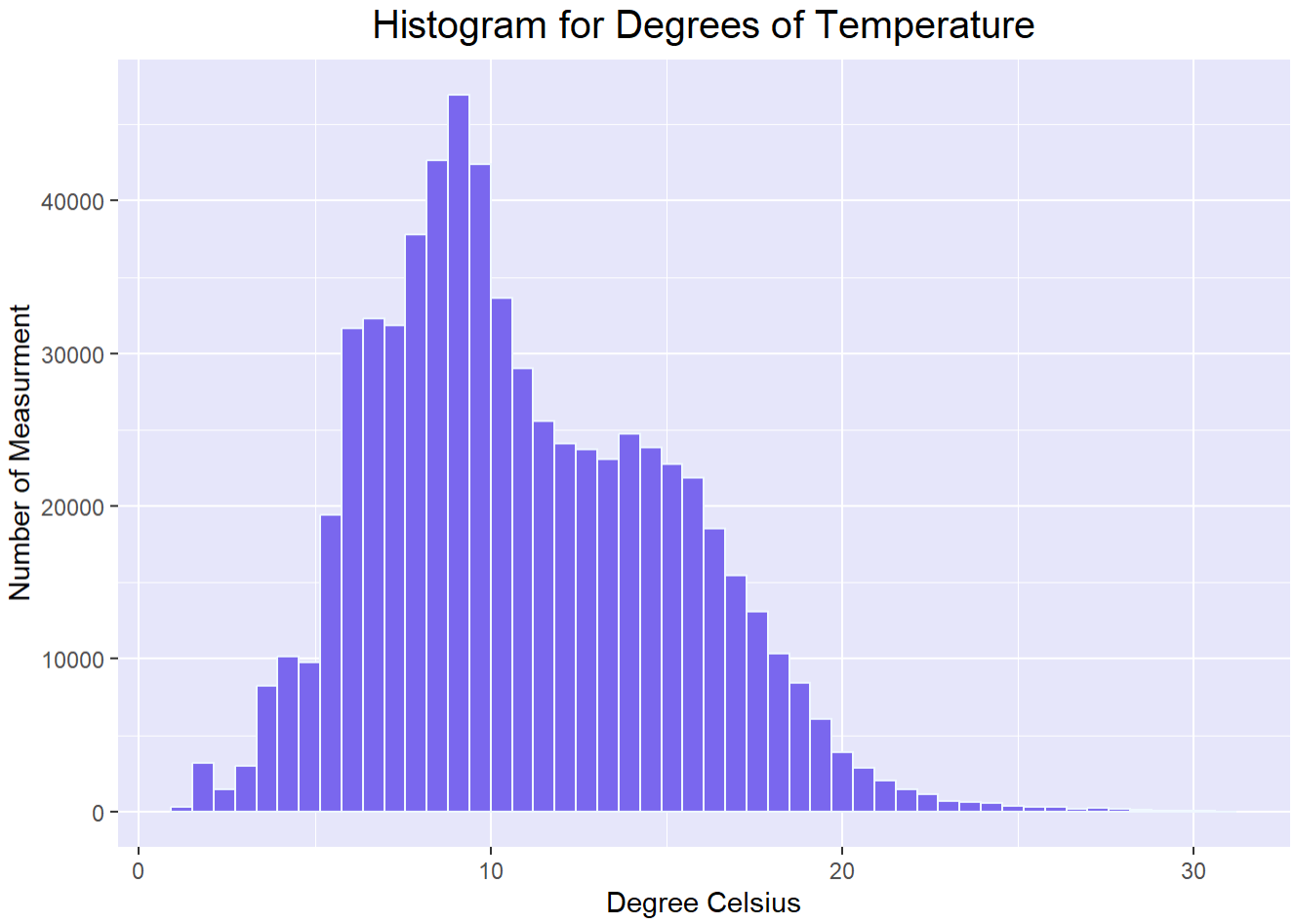


Fig.1: Histogram for Degrees of Temperature

- The number of measurements is higher when the temperature is about 10 degrees Celsius. As the temperature increases, the number of measurements decreases. This inference was made using the histogram plot.

Note: `bins` specifies the bin width to use in the histogram. Bins divide data values into specific intervals and show the frequency of values in each interval. The parameter `bins` determines the width of these bins. This can drastically change the appearance and interpretation of the histogram. In the first histogram we created, the bar width was “bins=30”. After that, we changed the bins value to 50, which reduced the width of the bars. It made the values in the dataset more visible.

```
ggplot(bottle_new, aes(y= bottle_new$T_degC)) +
  geom_boxplot()
```

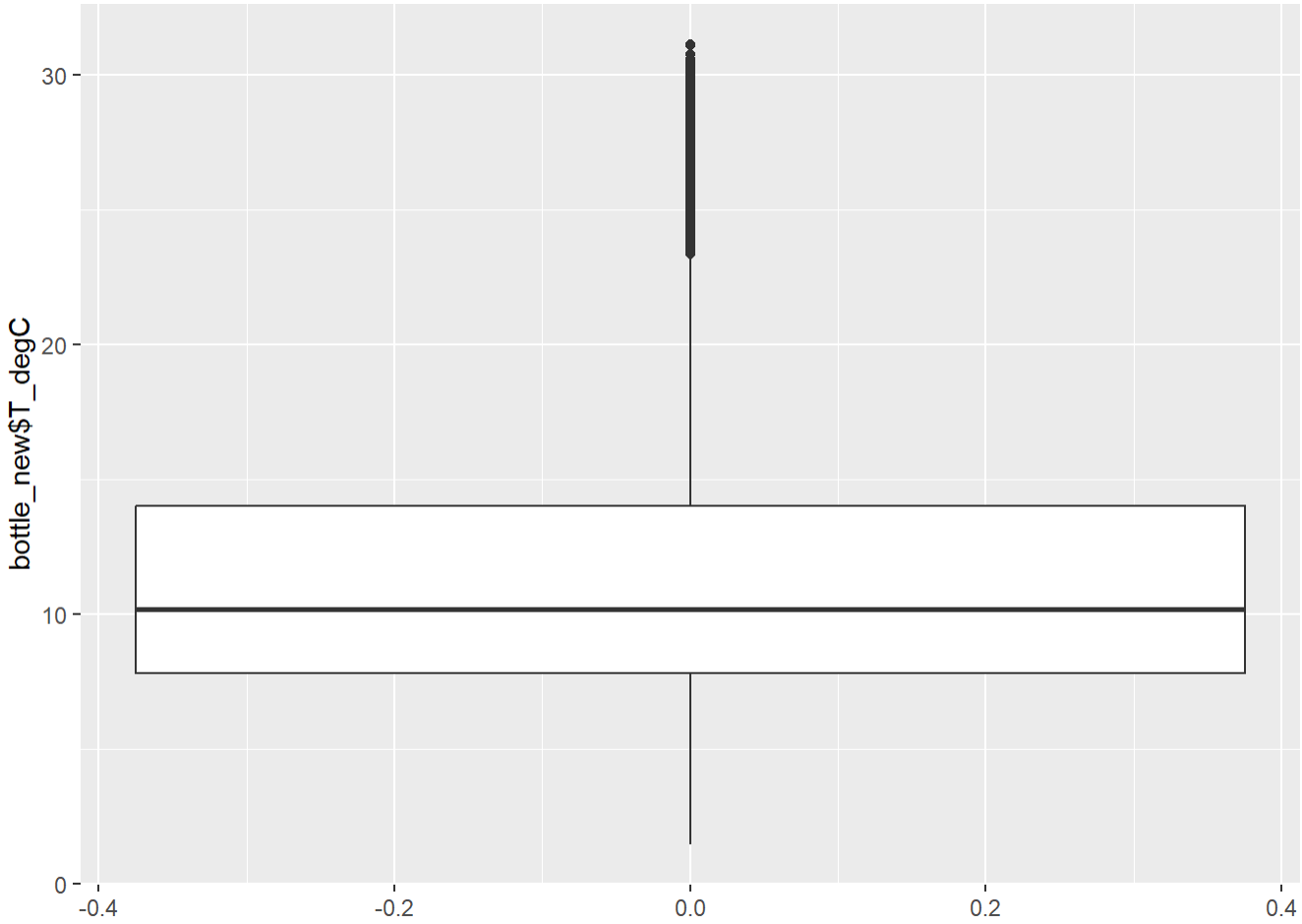
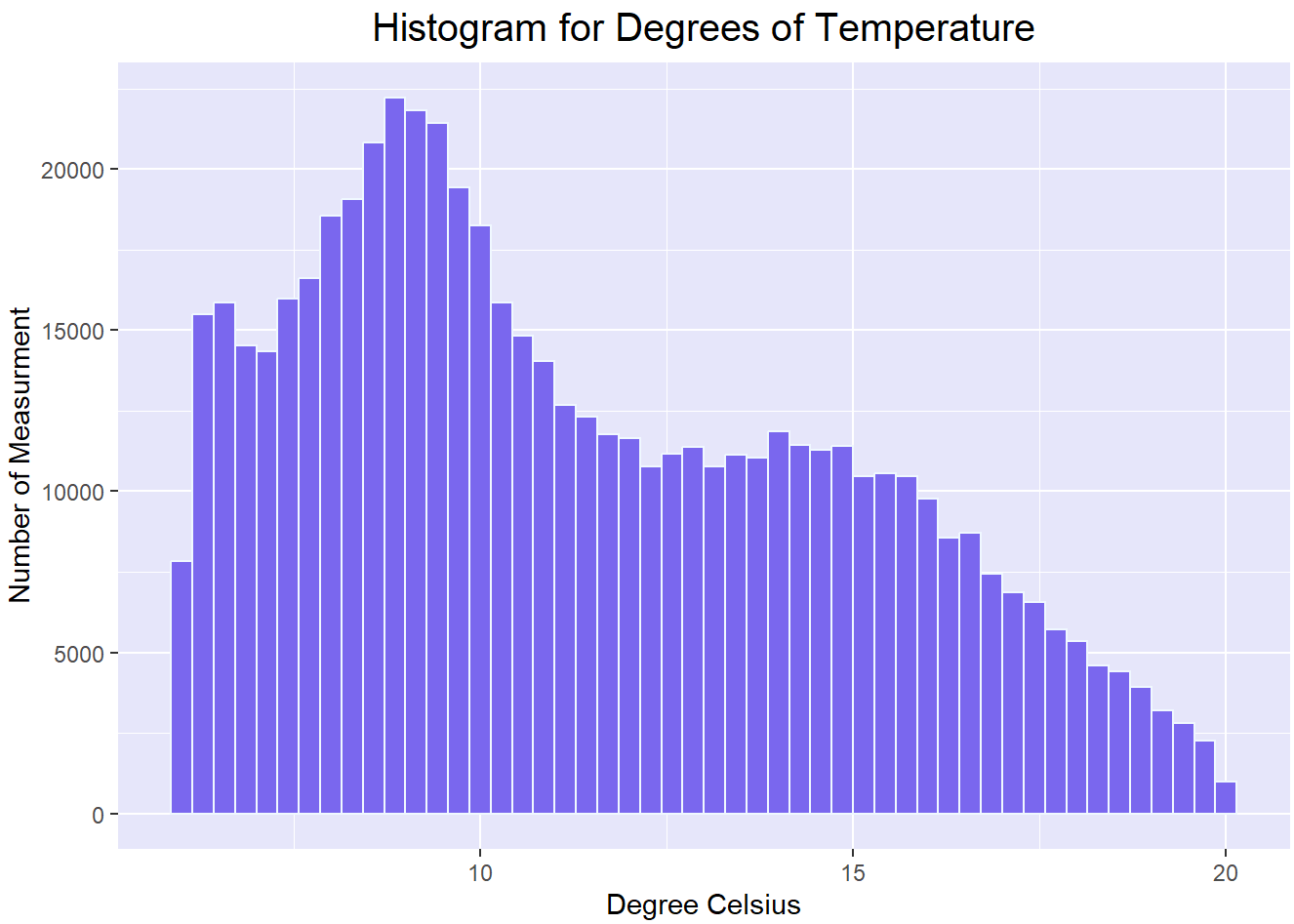


Fig.2: Boxplot for Degrees of Temperature

```
filtered_data <- filter(bottle_new, bottle_new$T_degC >= 6 & bottle_new$T_degC <= 20)

ggplot(filtered_data,aes(x=T_degC))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.50, size = 15)) +
    ggtitle("Histogram for Degrees of Temperature")+
  xlab("Degree Celsius")+
  ylab("Number of Measurment")
```



2.Scale of Salinity Graph

```
ggplot(bottle_new,aes(x=Salnty))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Scale of Salinity") +
 xlab("Salinity Scale")+
 ylab("Number of Measurment")
```

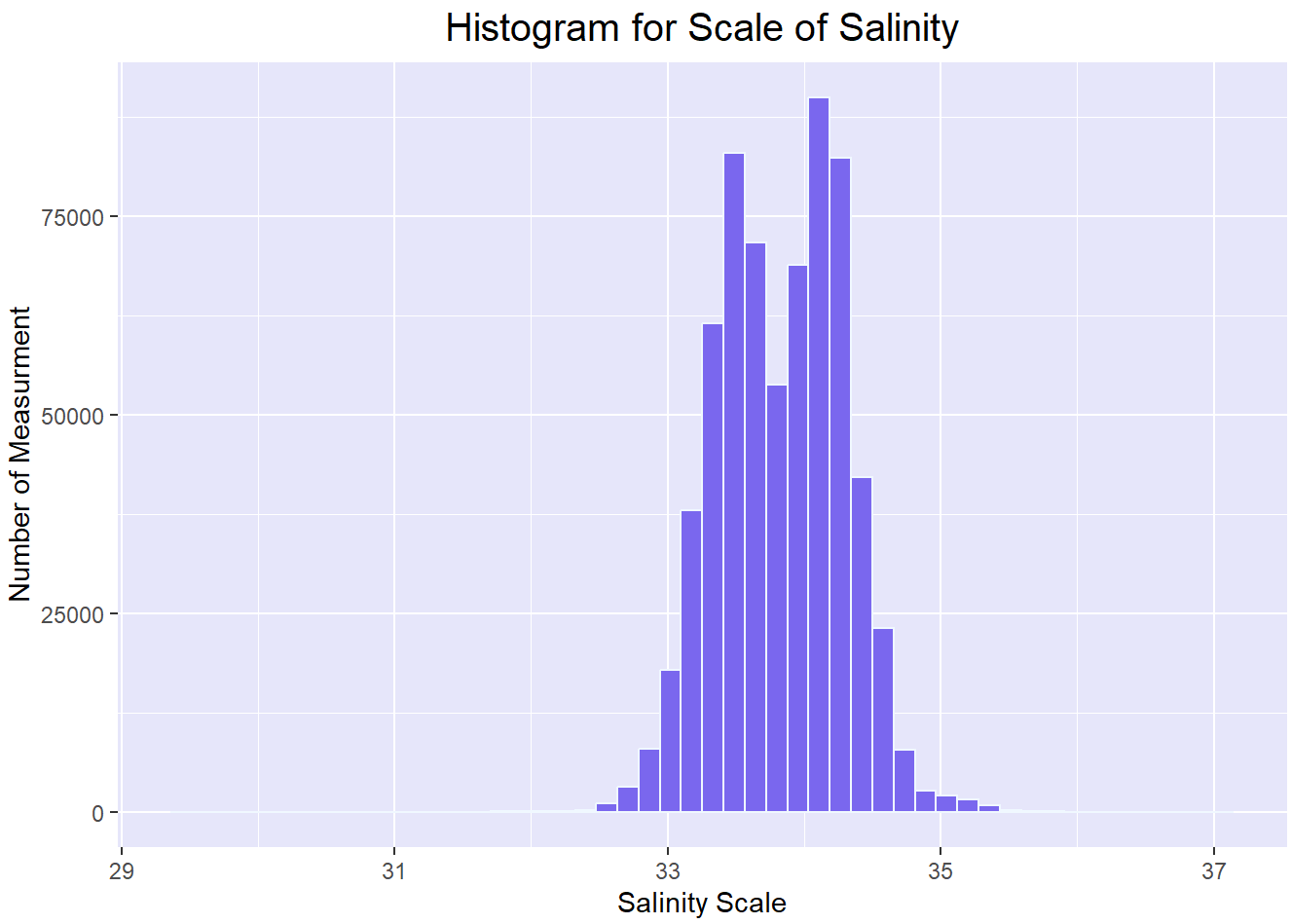


Fig.3: Histogram for Salinity Scale

```
ggplot(bottle_new, aes(y= bottle_new$Salnty)) +
  geom_boxplot()
```

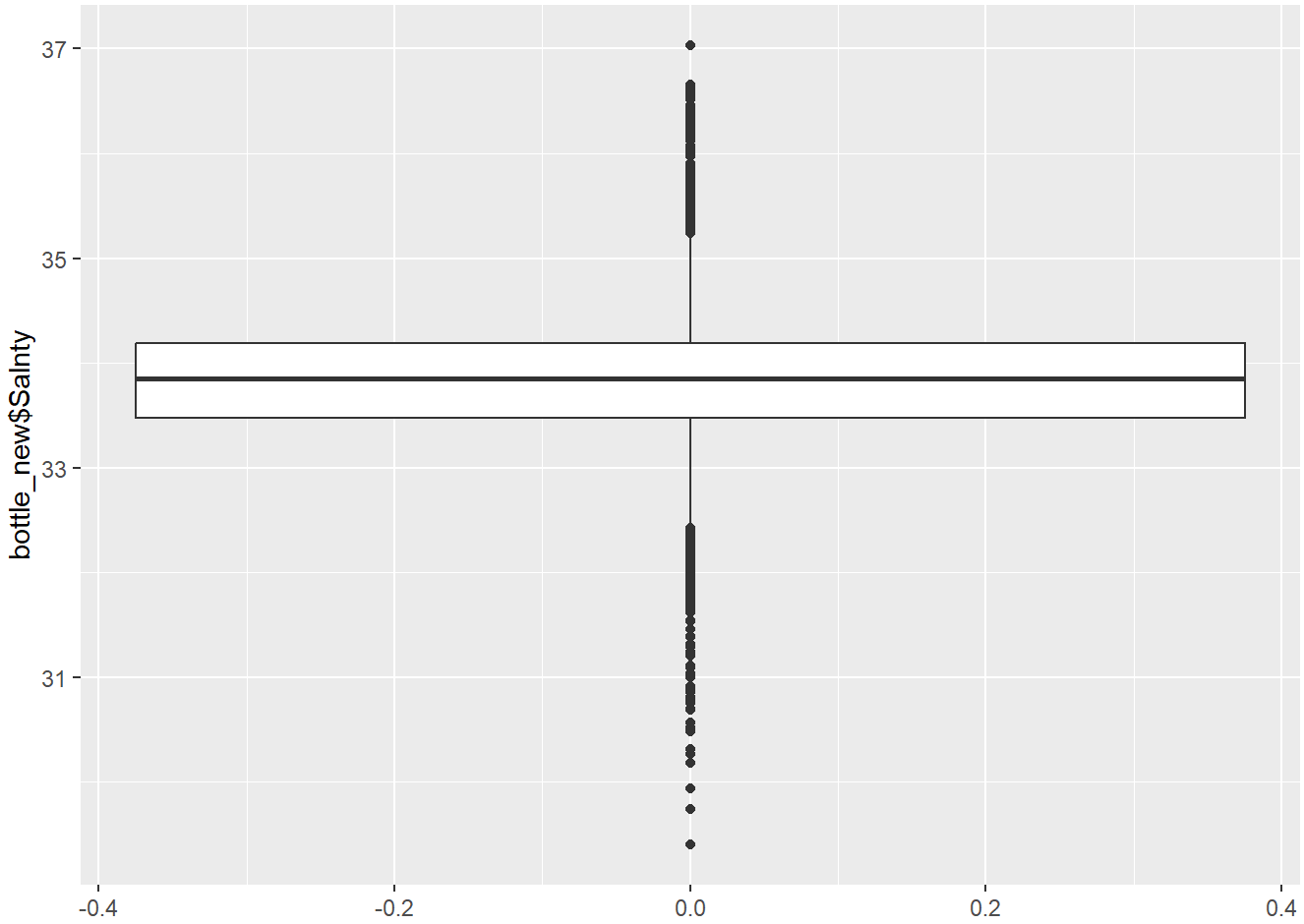


Fig.4: Boxplot for Salinity Scale

- The data on the salinity scale is highly concentrated between 32.5 and 35, and the values outside this range are considered outliers. Therefore, they will be removed during the detailed analysis, as they may lack accuracy and skew the results.

```
filtered_data <- filter(bottle_new, bottle_new$Salnty >= 32.5 & bottle_new$Salnty <= 35)
ggplot(filtered_data,aes(x=Salnty))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.50, size = 15)) +
  ggtitle("Histogram for Scale of Salinity") +
   xlab("Salinity Scale")+
  ylab("Number of Measurment")
```

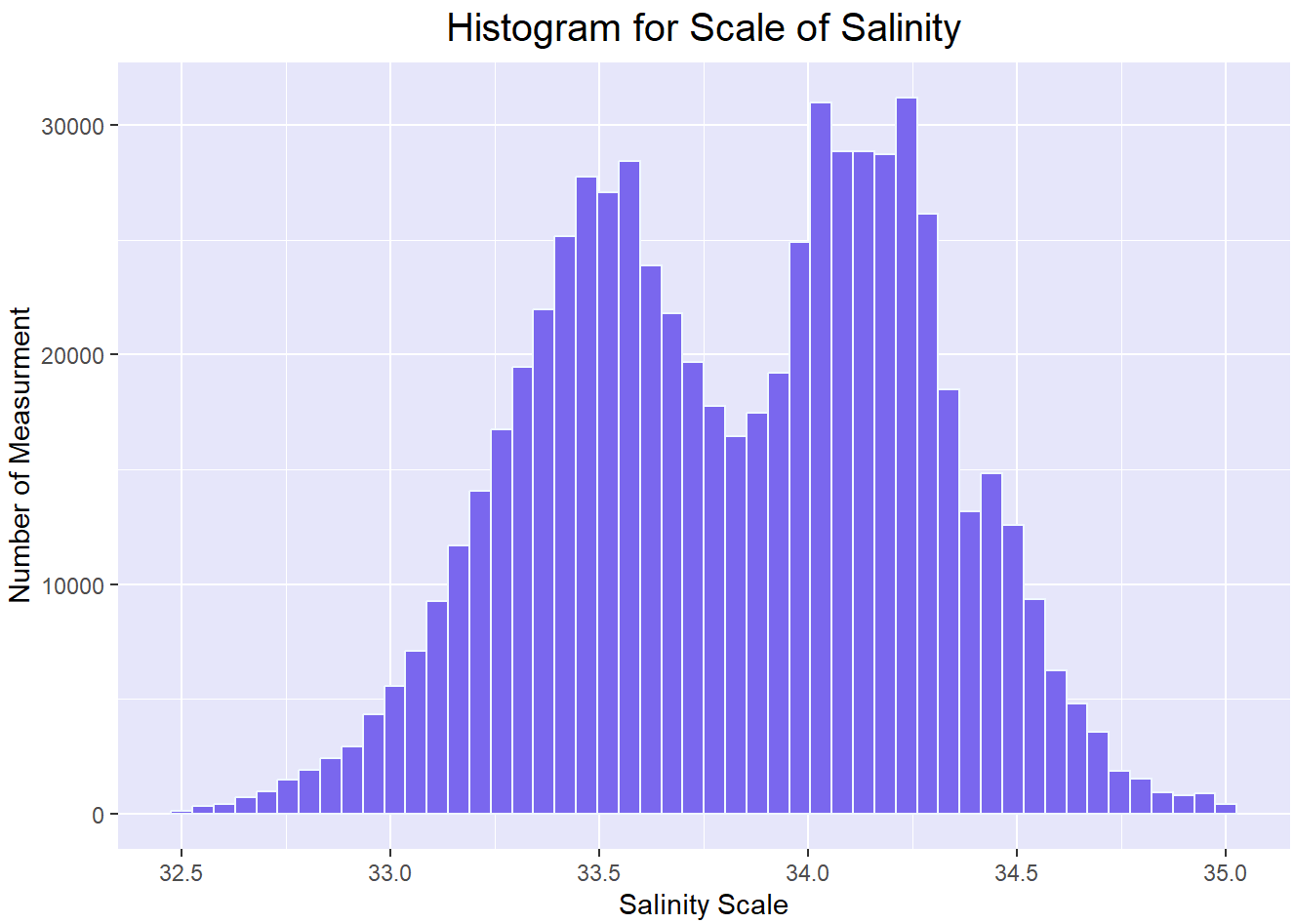


Fig.5: New Histogram for Salinity Scale

- When we apply outiler to the salinity scale, we can make a detailed examination in a more limited area. According to this graph, the number of measurements between 34 and 34.5 is higher than the other salinity scales. In the previous graph without outlier, we said that the measurements between 32.5 and 35 were more than the others. Since this graph is more detailed, it is more accurate.

3.Milliliters Oxygen per liter of Seawater Graph

```
ggplot(bottle_new,aes(x=O2ml_L))+
  geom_histogram(fill="#7a67ee",color="#7a67ee") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram Graph for Milliliters Oxygen per liter of Seawater")+
   xlab("Milliliters Oxygen per liter of Seawater")+
  ylab("Number of Measurment")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

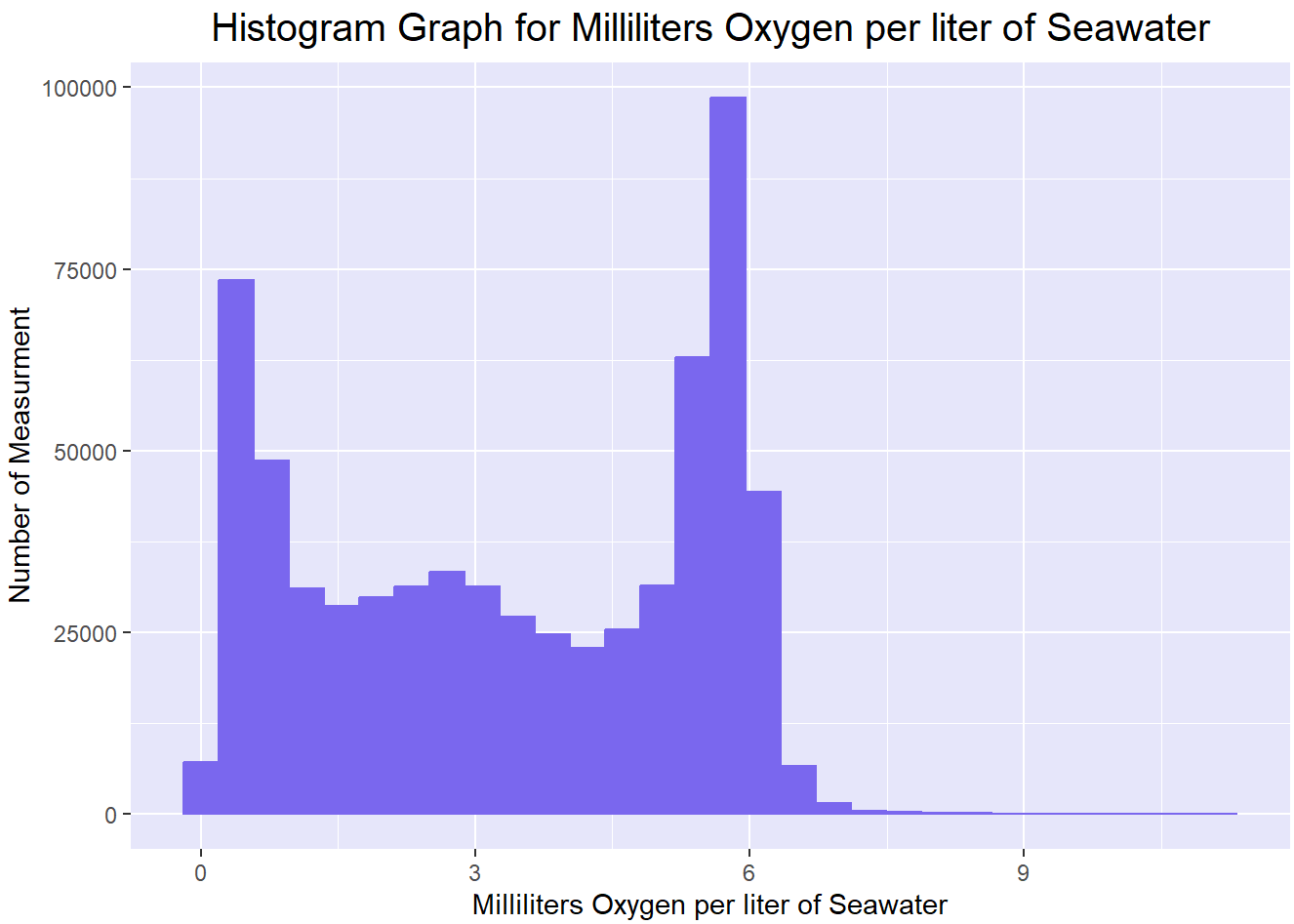


Fig.6: Histogram for Milliliters Oxygen per liter of Seawater

```
ggplot(bottle_new, aes(y= bottle_new$O2ml_L)) +
  geom_boxplot()
```

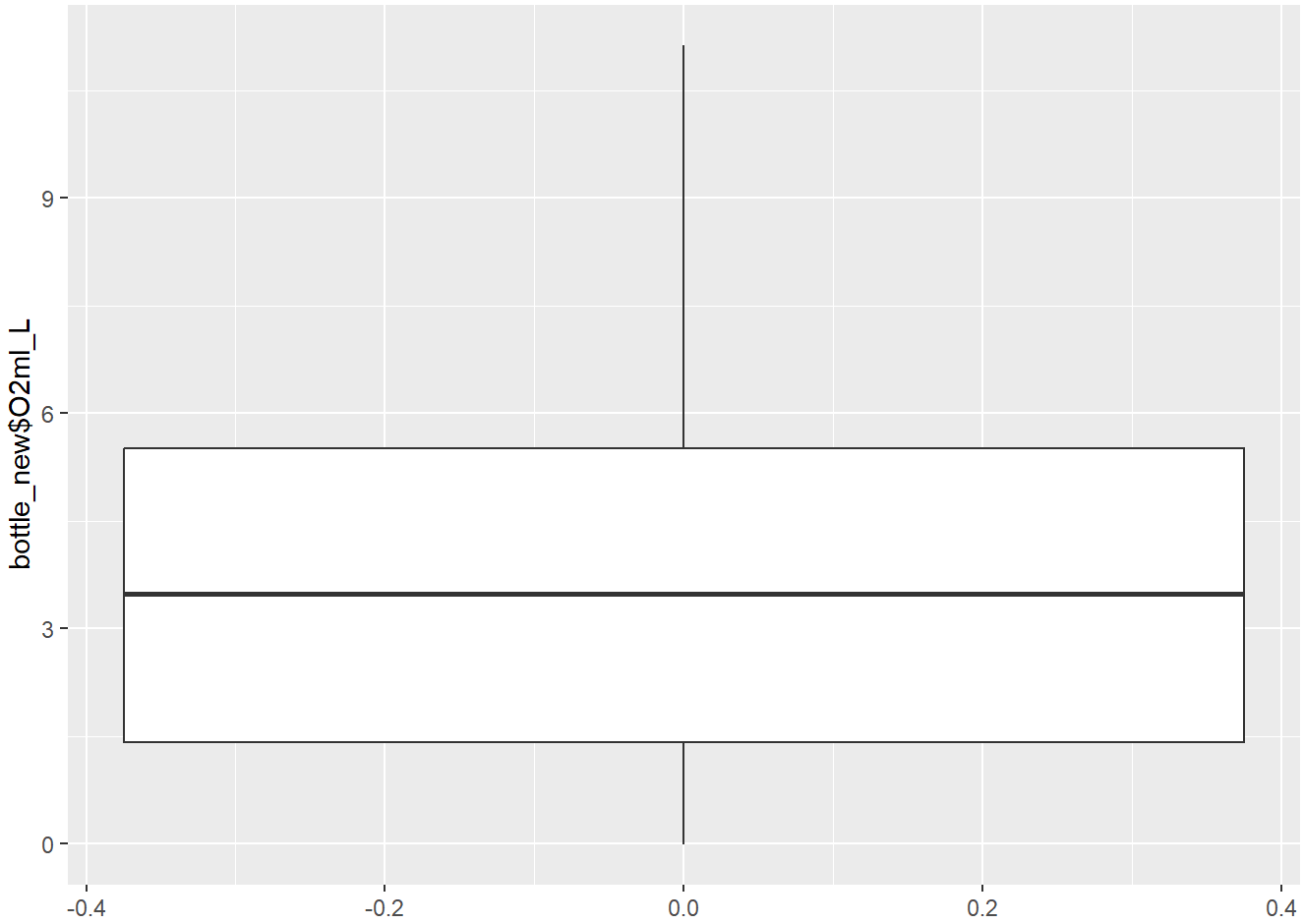


Fig.7: Boxplot for Milliliters Oxygen per liter of Seawater

- As evident in the Graph, our values show a decrease after 6 ml/O2, and we need to eliminate these outliers.

```
filtered_data <- filter(bottle_new, bottle_new$O2ml_L < 6.5)

ggplot(filtered_data,aes(x=O2ml_L))+
  geom_histogram(fill="#7a67ee",color="#7a67ee") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram Graph for Milliliters Oxygen per liter of Seawater")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

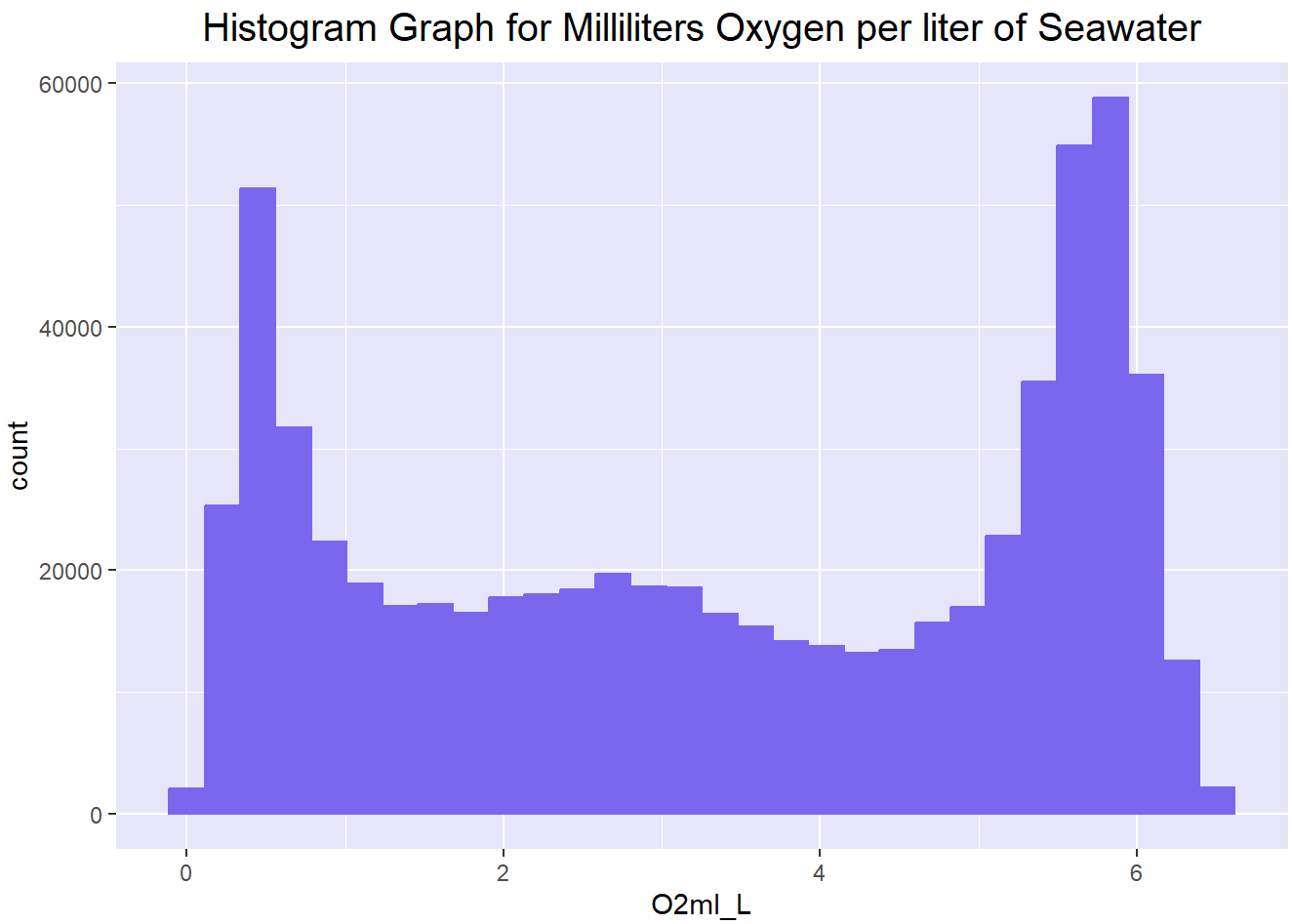


Fig.8: New Histogram for Milliliters Oxygen per liter of Seawater

- More measurements were made when the amount of oxygen per liter of seawater was 6 milliliters. After 6 milliliters the measured values decreased.

4.Bottle Depth in Meters Graph

```
ggplot(bottle_new,aes(x=Depthm))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Bottle Depth in Meters")+
  xlab("Bottle depth in meters")+
  ylab("Number of measurment")
```

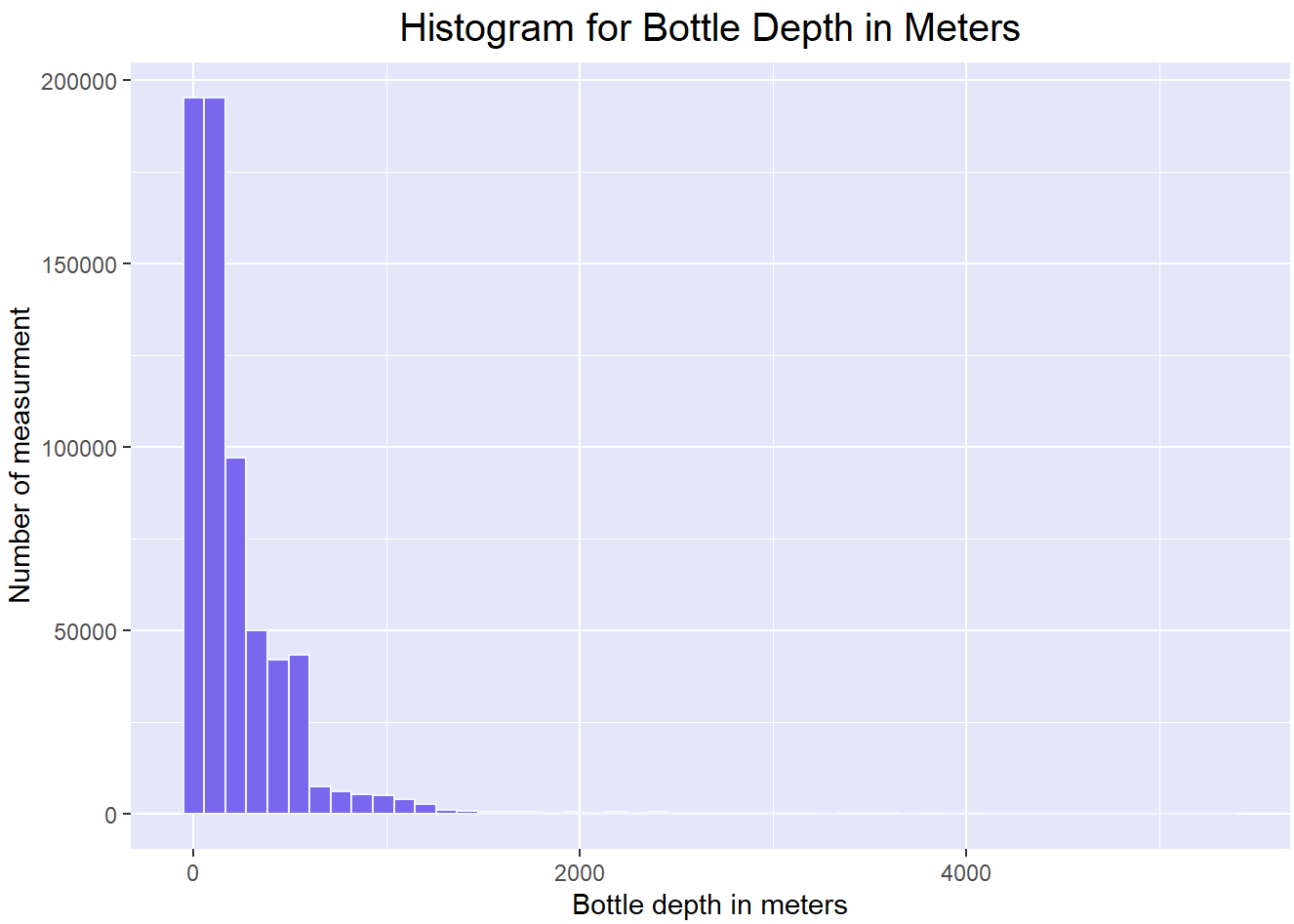


Fig.9: Histogram for Bottle Depth in Meters

- Measurements were made when the depth was 0 and 2000. As the depth approaches 0, the number of measurements is higher.

```
ggplot(bottle_new, aes(y= bottle_new$Depthm)) +
  geom_boxplot()
```

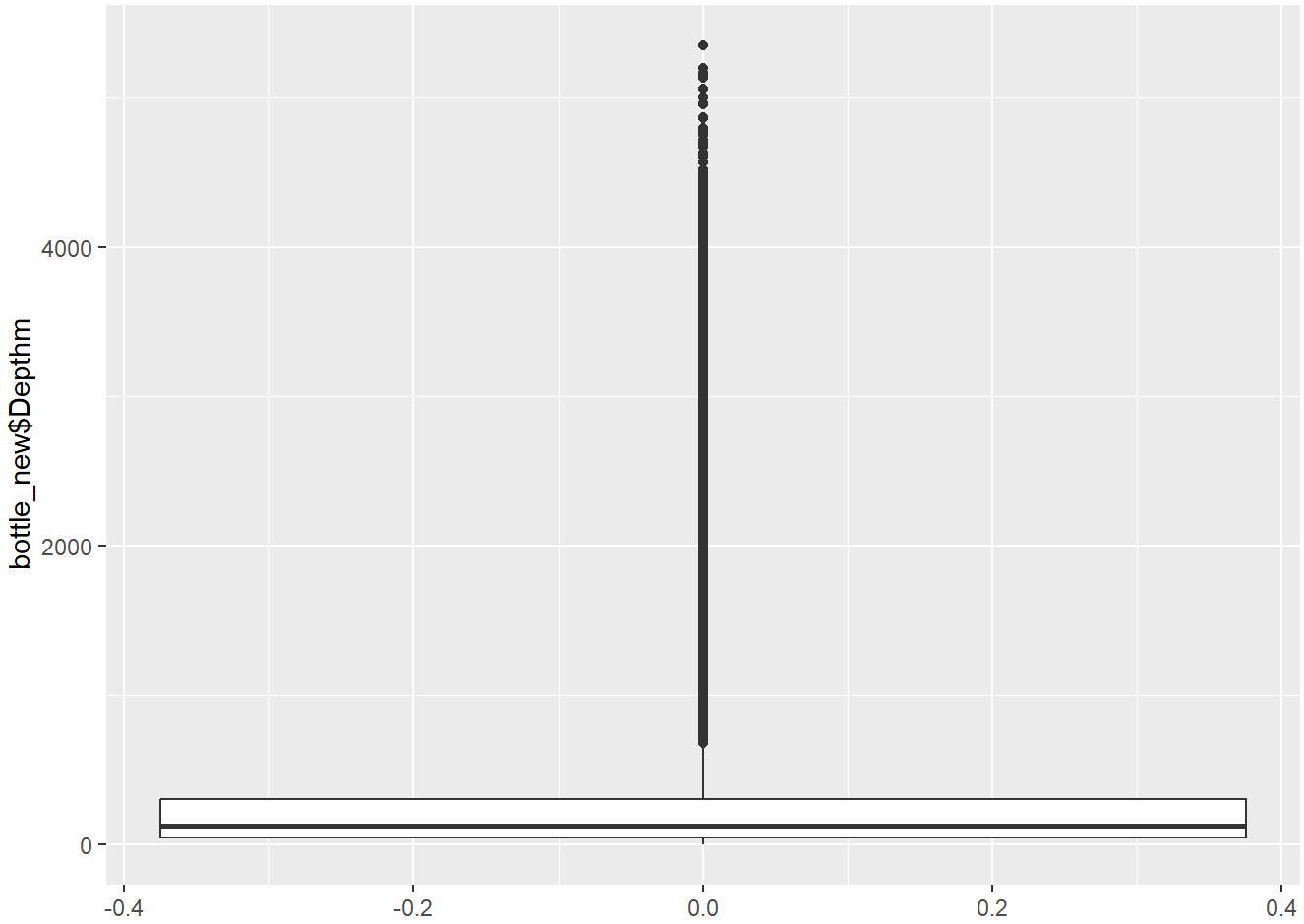


Fig.10: Boxplot for Bottle Depth in Meters

```
filtered_data <- filter(bottle_new, bottle_new$Depthm < 600)

ggplot(filtered_data,aes(x=Depthm))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Bottle Depth in Meters")+
  xlab("Bottle depth in meters")+
  ylab("Number of measurment")
```

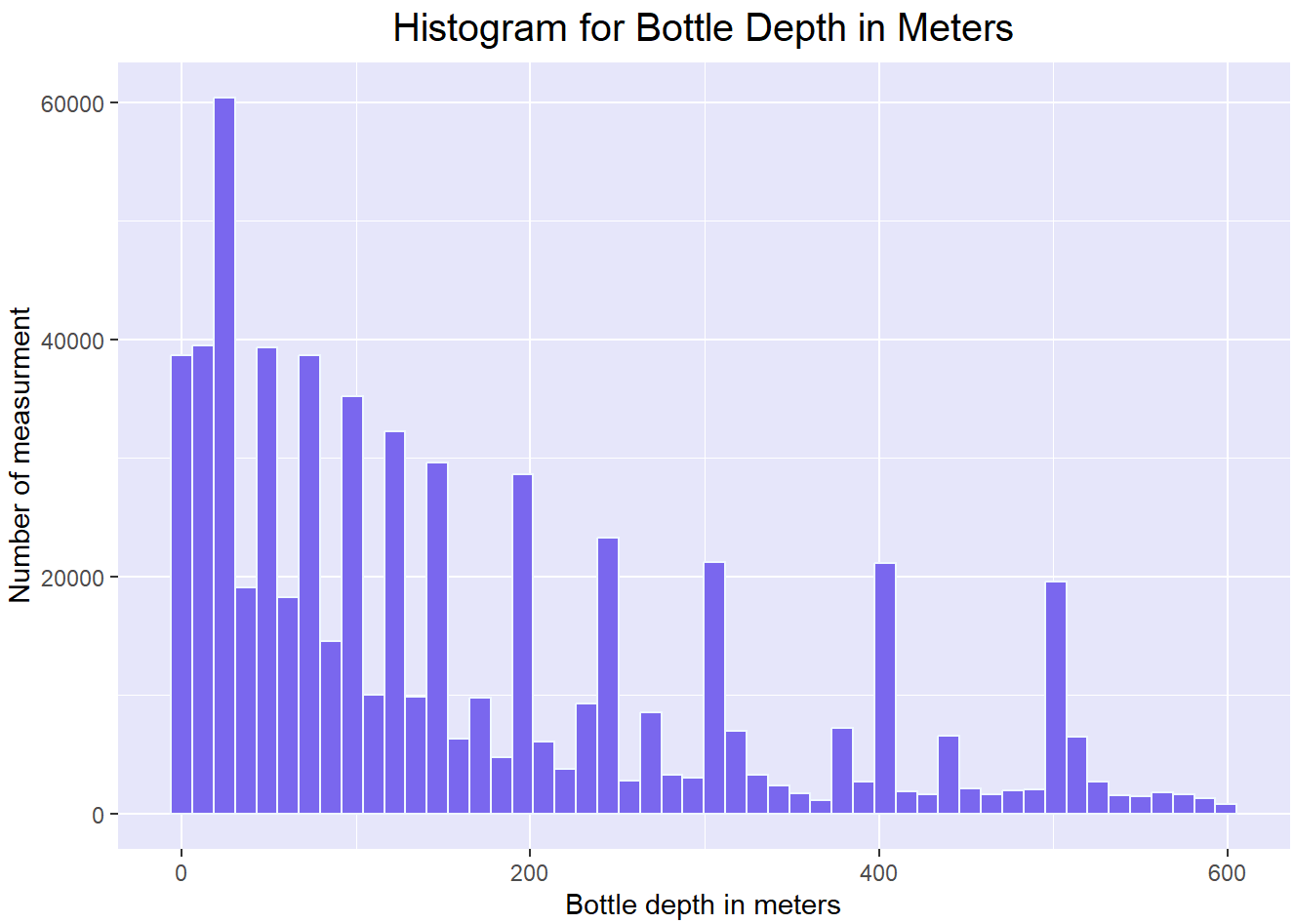


Fig.11: New Histogram for Bottle Depth in Meters

5.Potential Density

```
ggplot(bottle_new,aes(x=STheta))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Potential Density")+
  xlab("Potential Density (Kg/M³)")+
  ylab("Number of measurment")
```

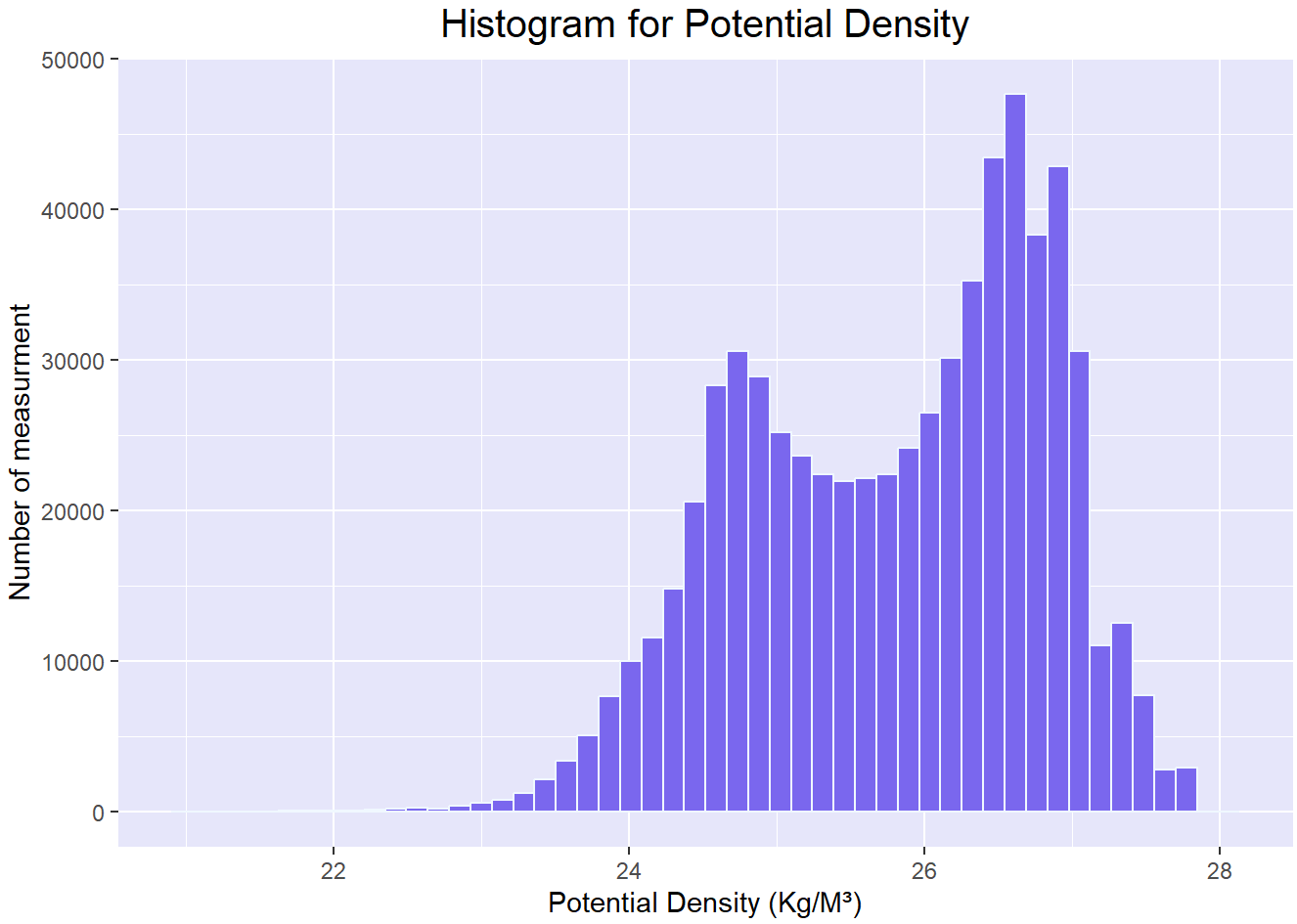


Fig.12: Histogram for Potential Density

```
filtered_data <- filter(bottle_new, bottle_new$STheta >= 24 & bottle_new$STheta <= 27)

ggplot(filtered_data,aes(x=STheta))+
  geom_histogram(bins = 50, fill="#7a67ee",color="#f0f8ff") +
  theme(panel.background = element_rect(fill = "#e6e6fa"), plot.title = element_text(hjust =
0.5, size = 15)) +
  ggtitle("Histogram for Potential Density")+
  xlab("Potential Density (Kg/M³)")+
  ylab("Number of measurment")
```

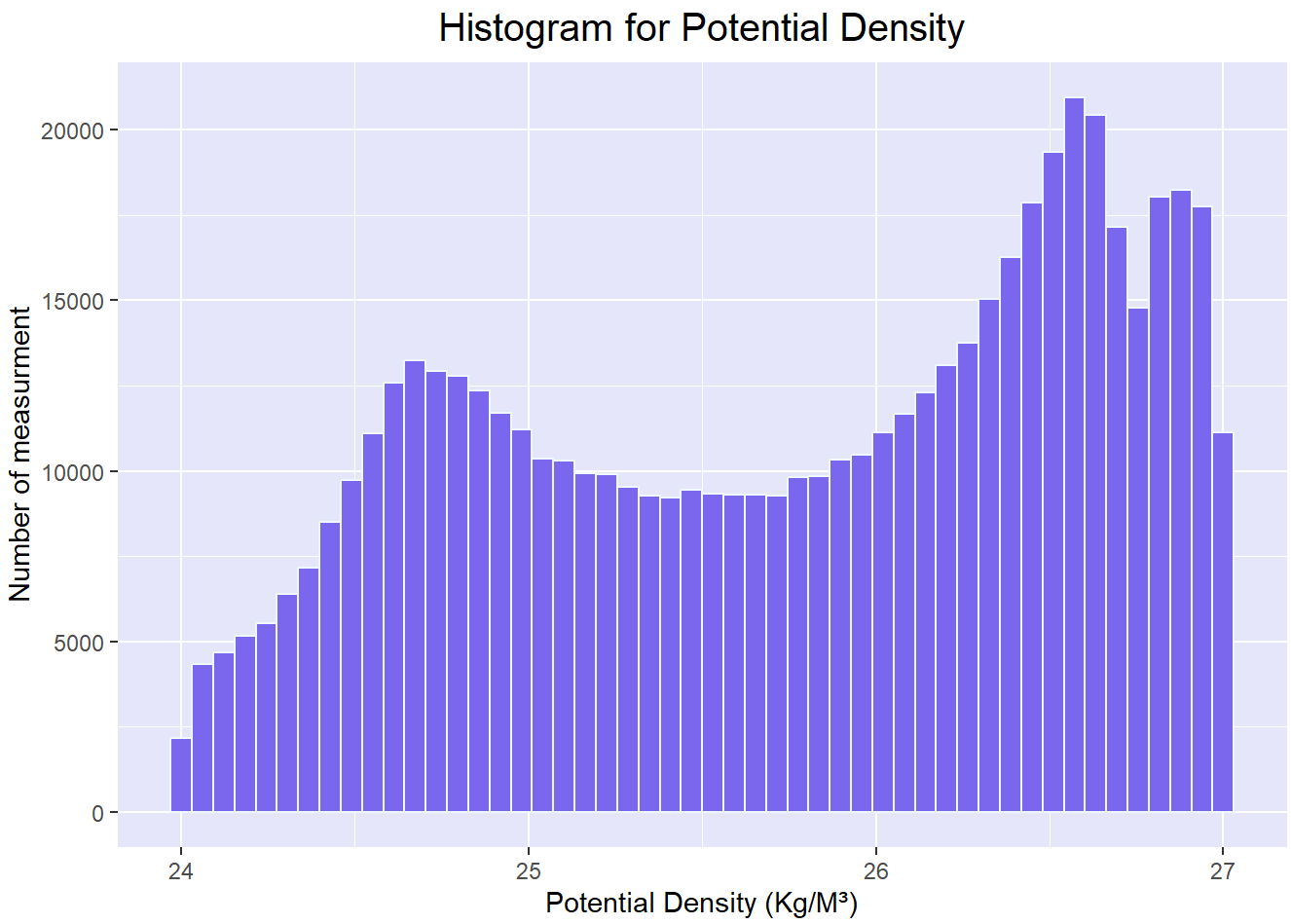


Fig.13: New Histogram for Potential Density

- Due to the residual values, filtering was done between 24 and 27 values and a graph was created again.

Two Variable Analysis

Two variable analysis is used to visualize the relationship between two variables. The x-axis represents one variable and the y-axis represents the other variable. The points, lines, or columns between these values show the relationship between the two variables. In this two variable analysis, we created T_degC - Salnty, T_degC - Depthm, Depthm - Salnty, O2ml_L - T_degC graphs.

1.Temperature - Scale of Salinity

```
ggplot(data = filtered_data, aes(x=T_degC, y=Salnty)) +
  geom_point(color='#81E0F7')+
  theme(panel.background = element_rect(fill = "#f2fbfa"))+
  ggtitle("Scatterplot Graph for Temperature - Scale of Salinity")+
  ylim(31.5,36)+
  xlab("Temperature")+
  ylab("Scale of Salinity")
```

```
## Warning: Removed 44 rows containing missing values (`geom_point()`).
```

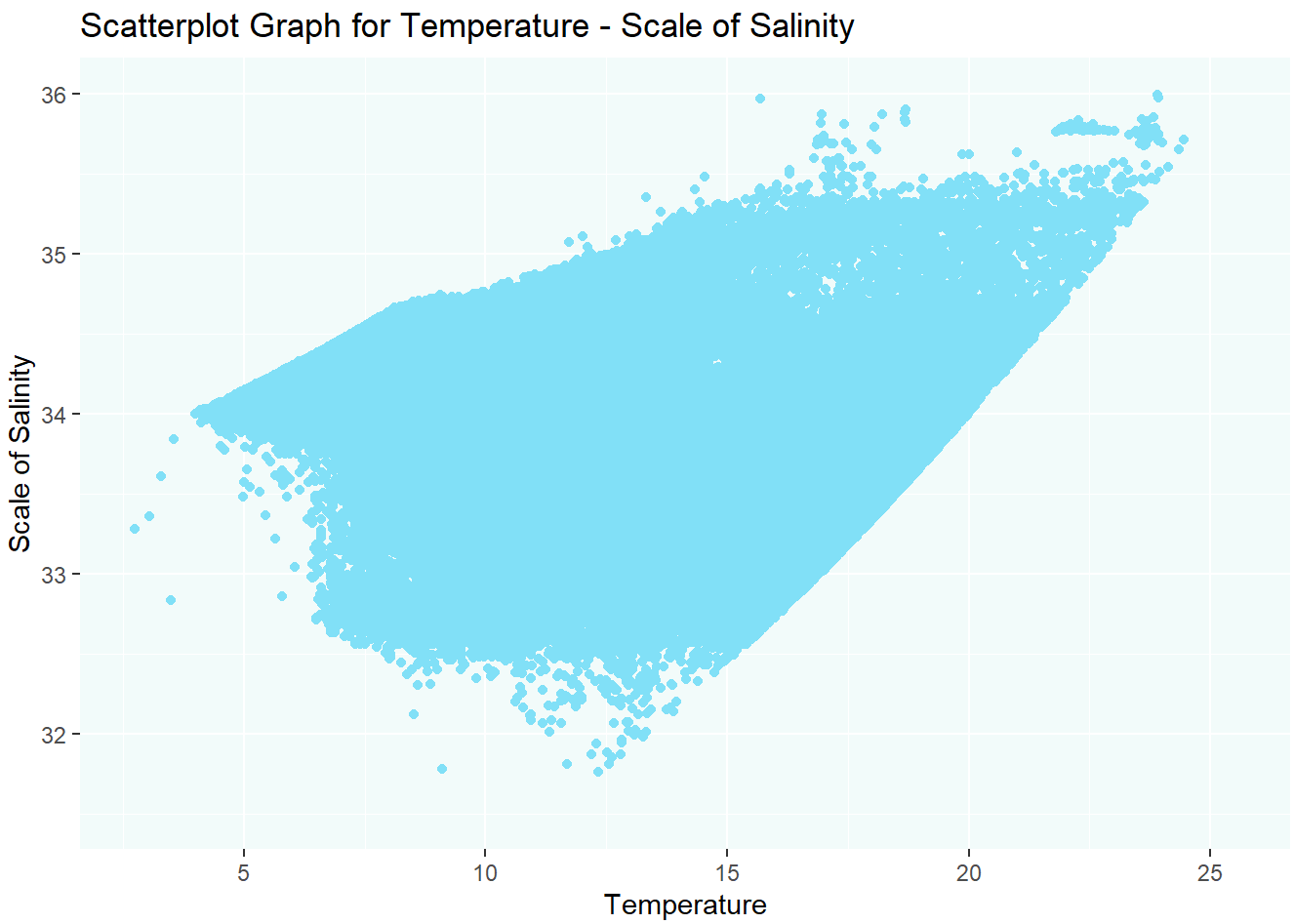


Fig.14: Scatterplot Graph for Temperature - Scale of Salinity

- If the salinity scale increases, the temperature increases. If the temperature decreases, the salinity scale decreases. As a result, there is a linear relationship between temperature and salinity scale. These variables affect each other positively.

2.Depth in Meters - Temperature

```
ggplot(data = filtered_data, aes(x=Depthm, y=T_degC)) +
  geom_point(color = "#CB6D51")+
  theme(panel.background = element_rect(fill = "#fbeee6"))+
  ggtitle("Scatterplot Graph for Depth in Meters - Temperature")+

  xlab("Depth in Meters")+
  ylab("Temperature")
```

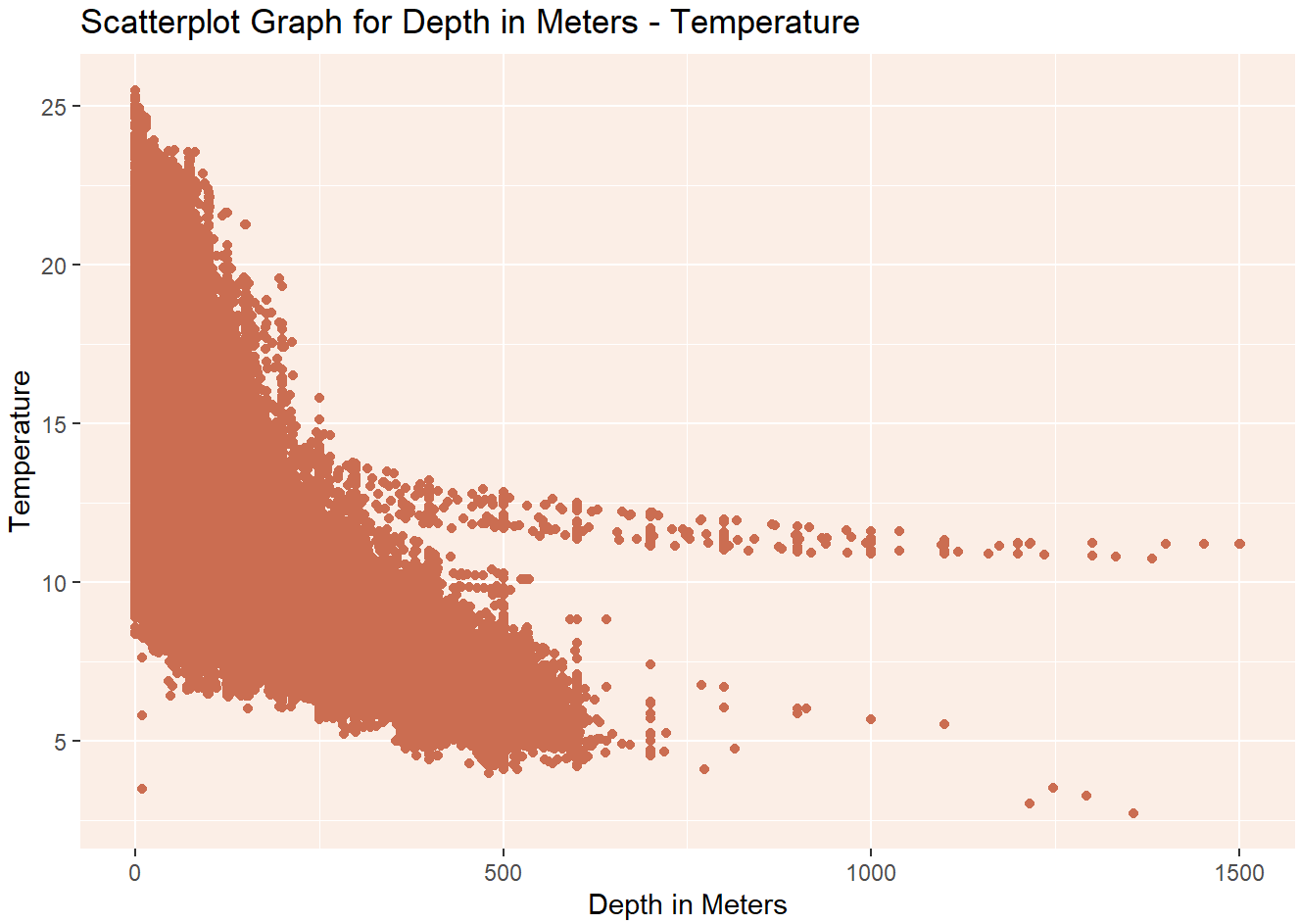


Fig.15: Scatterplot Graph for Depth in Meters - Temperature

- If the depth increases, the temperature decreases. If the depth is more than 4000, the temperature approaches 0 degrees Celsius. If the temperature increases, the depth decreases. There is an inverse relationship between these two variables. In other words, they affect each other negatively.

3.Scale of Salinity - Depth in Meters

```
ggplot(data = filtered_data, aes(x=Depthm, y=Salnty)) +
  geom_point(color='#151B54')+
  theme(panel.background = element_rect(fill = "#ebf5fb"))+
  ggtitle("Scatterplot Graph for Depth in Meters - Scale of Salinity ")+
  ylim(31.5,35.5)+
  xlab("Depth in Meters")+
  ylab("Scale of Salinity")
```

```
## Warning: Removed 220 rows containing missing values (`geom_point()`).
```

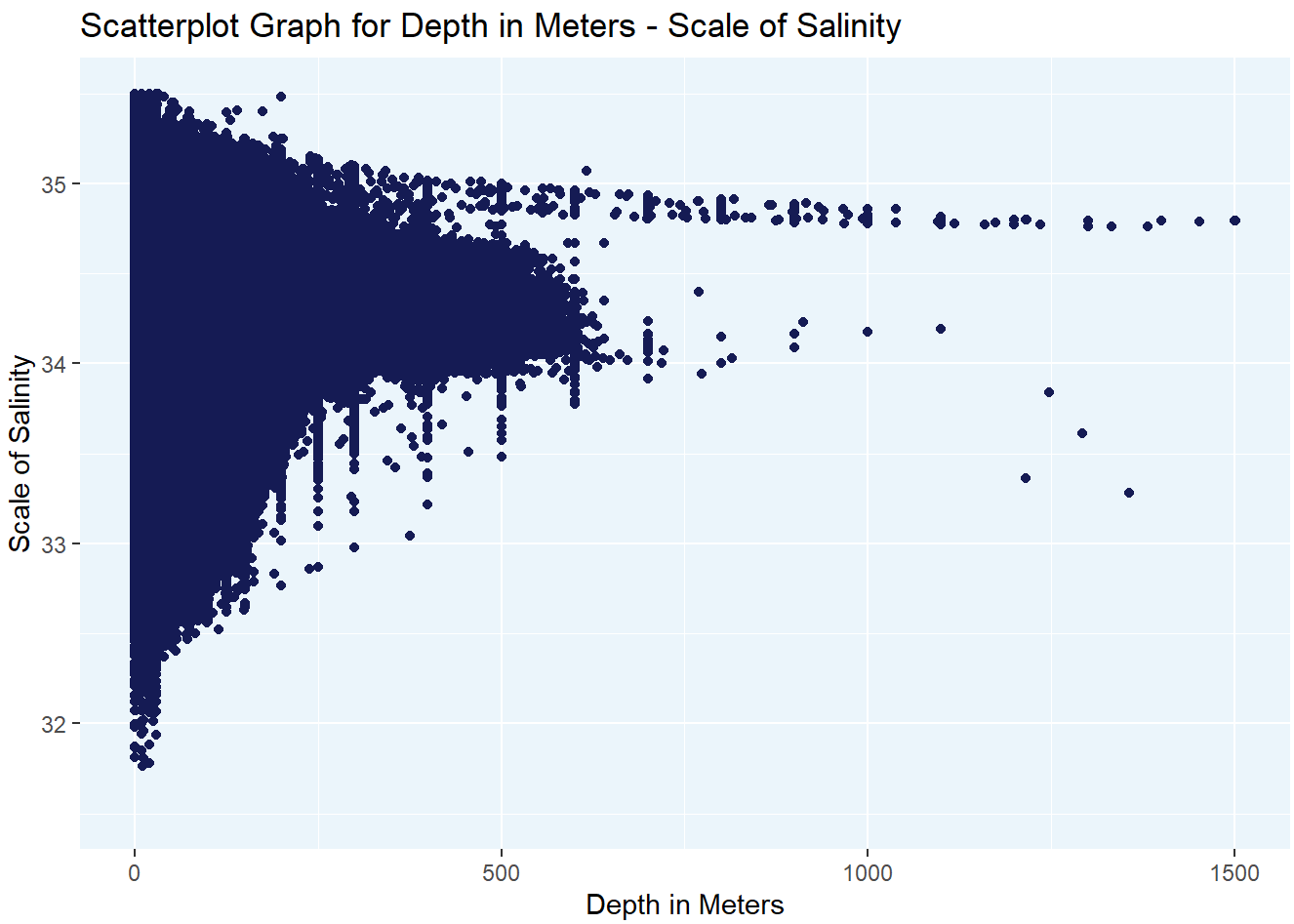


Fig16: Scatterplot Graph for Scale of Salinity - Depth in Meters

- The salinity scale can take any value when the depth is 0, but as seen in the graph, the measurement is at 1000 meters and above when the salinity is 34-35.

4.Milliliters Oxygen per liter of Seawater - Temperature

```
ggplot(data = filtered_data, aes(x=T_degC, y=O2ml_L)) +
  geom_point(color='#347C17')+
  theme(panel.background = element_rect(fill = "#eafaf1"))+
  ggtitle("Scatterplot Graph for Temperature - Milliliters Oxygen per liter of Seawater")+
  xlab("Temperature")+
  ylab("Milliliters Oxygen per liter of Seawater")
```

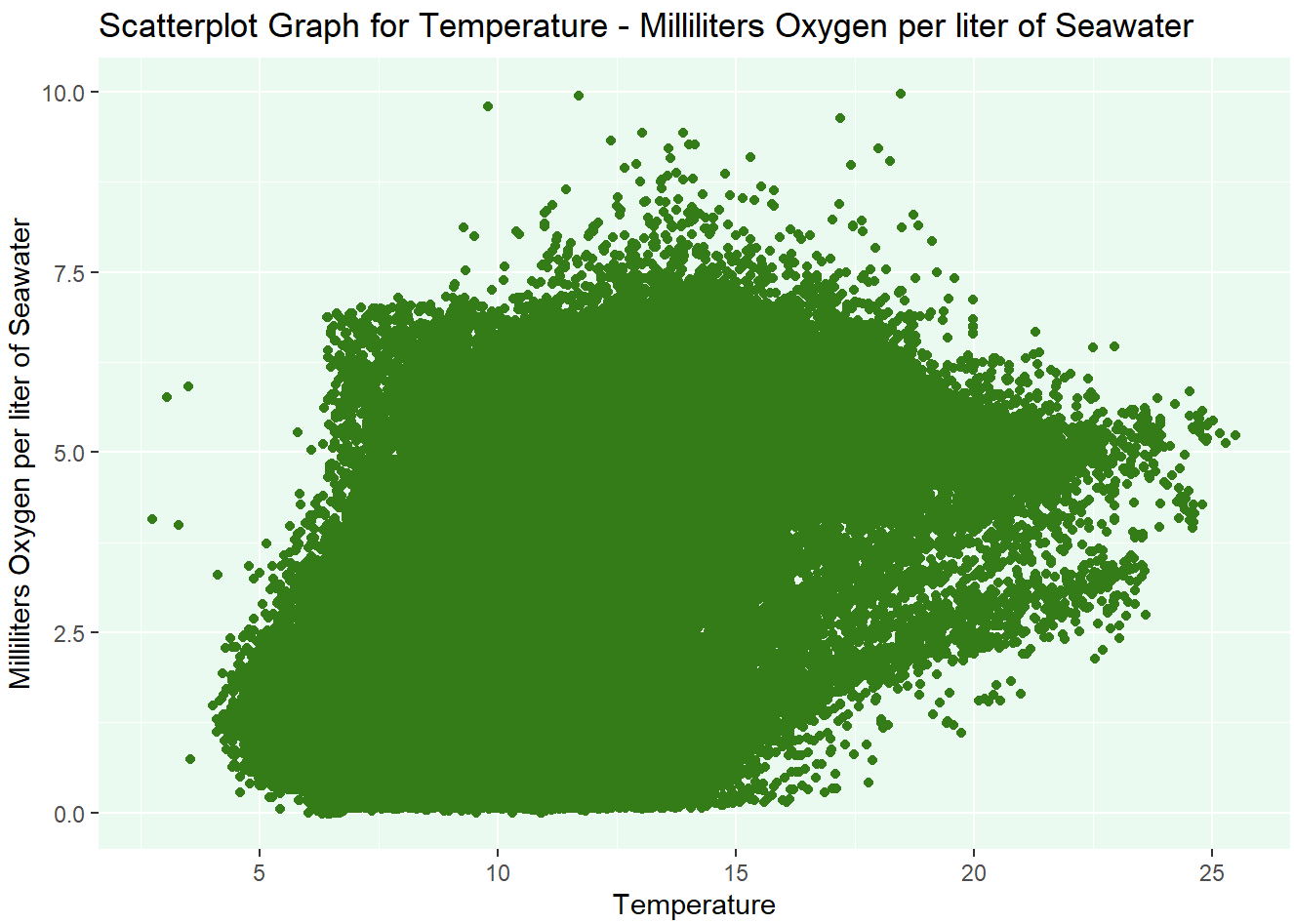


Fig17: Scatterplot Graph for Temperature - Milliliters Oxygen per liter of Seawater

```
ggplot(data = filtered_data, aes(x=T_degC, y=O2ml_L)) +
  geom_point(color='#7D0552')+
  theme(panel.background = element_rect(fill = "#fbf2f9"))+
  ylim(3, 9)+
  xlim(10, 30)+
  ggtitle("Scatterplot Graph for Temperature - Milliliters Oxygen per liter of Seawater")+
  xlab("Temperature")+
  ylab("Milliliters Oxygen per liter of Seawater")
```

```
## Warning: Removed 292883 rows containing missing values (`geom_point()`).
```

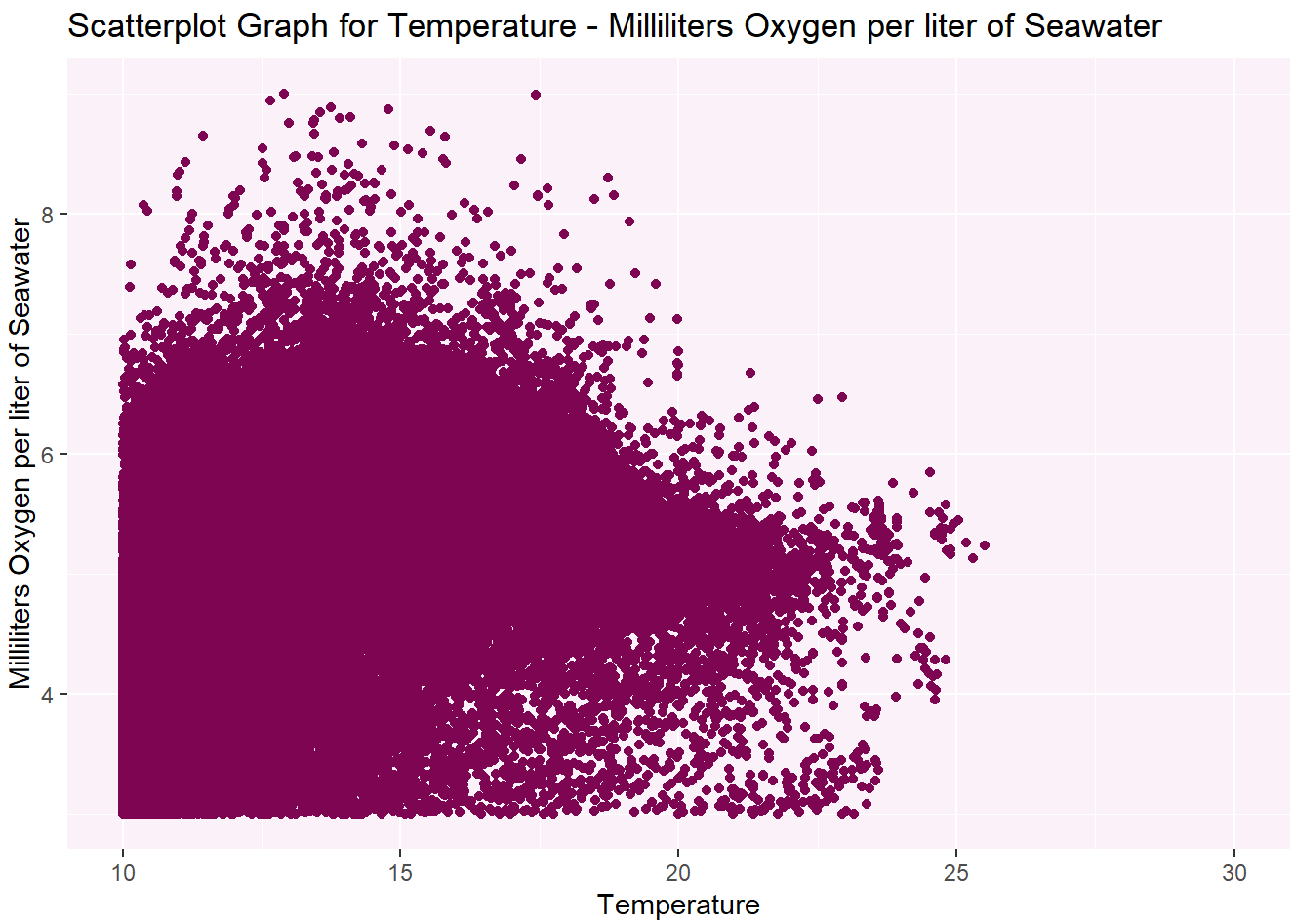


Fig18: Limited Scatterplot Graph for Temperature - Milliliters Oxygen per liter of Seawater

- At temperatures of 0-10 degrees, the amount of oxygen can take values of 3, 6, and 9. In this case, it is difficult to establish a relationship between them. Therefore, we made restrictions on X and Y coordinates. We evaluated the x-axis in the range of 0 and 9. We evaluated the y-axis in the range of 10 and 30 degrees and the point graph was created. According to this graph, when the temperature increases, the amount of oxygen increases or decreases. As a result, these two variables do not affect each other. As a result of the restriction, no inverse or linear relationship was found, but visualization was made in a narrower data set.

Regression and Correlation Analysis

- Regression analysis, as a statistical technique, is a method that examines how a dependent variable is related to one or more independent variables. This analysis is used to model and predict the relationship between variables. Its main purpose is to understand the effect of independent variables on the dependent variable.
- Correlation of Definition The cor() function in R calculates the Pearson correlation coefficient to measure the relationship between two or more variables. The Pearson correlation coefficient is a numerical measure of the linear relationship between two variables. The range of values is between -1 and 1. -1 indicates a complete negative relationship, 0 indicates no relationship, and 1 indicates a complete positive relationship.
- The reason why we chose temperature as the dependent variable is that we want to measure whether it

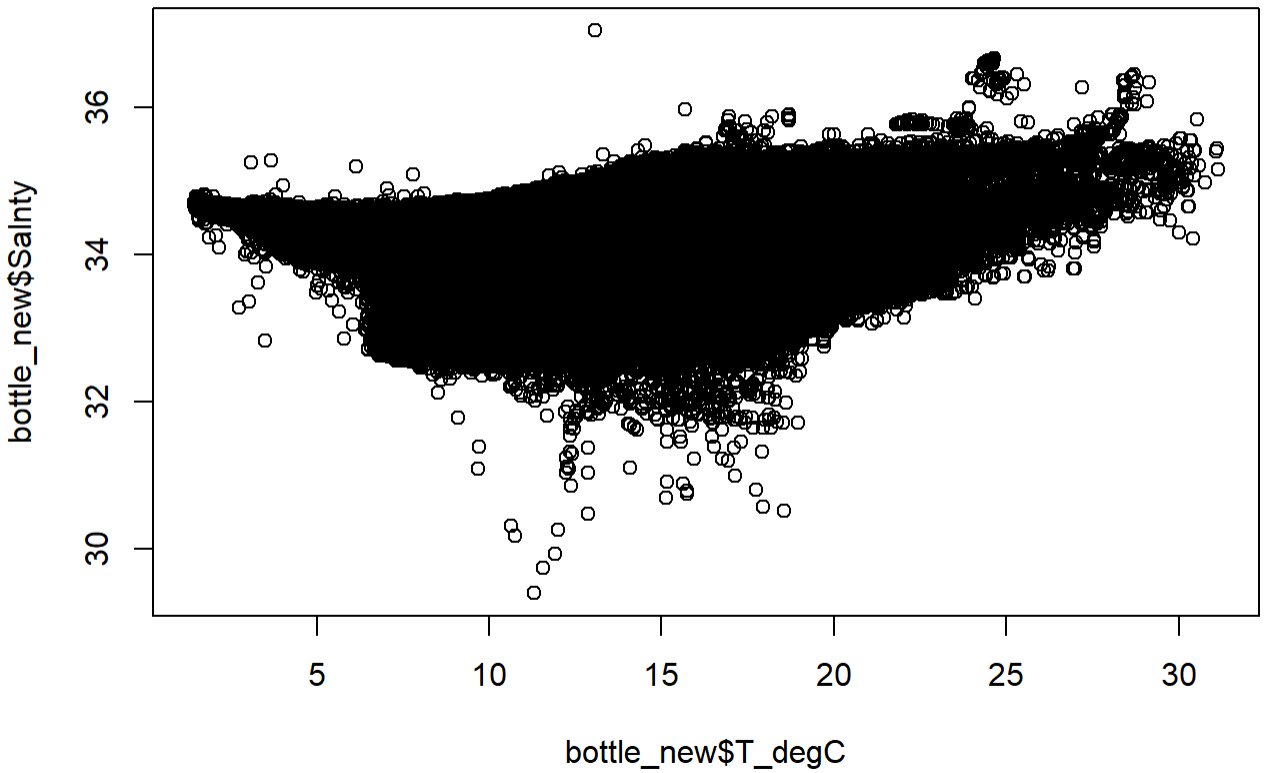
is in a dependency relationship with the other independent variables, and if so, to what extent, and it is the most appropriate variable for this.

1. Temperature - Scale of Salinity

```
cor.test(bottle_new$T_degC,bottle_new$Salnty)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bottle_new$T_degC and bottle_new$Salnty
## t = -475.39, df = 660240, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.506773 -0.503179
## sample estimates:
##        cor
## -0.5049782
```

```
plot(bottle_new$T_degC,bottle_new$Salnty)
```




```
cor(bottle_new[1:5])
```

```
##            T_degC     Salnty     O2ml_L     STheta     Depthm
## T_degC  1.0000000 -0.5049782  0.7949001 -0.9640874 -0.6669908
## Salnty -0.5049782  1.0000000 -0.8243646  0.7011695  0.5649549
## O2ml_L  0.7949001 -0.8243646  1.0000000 -0.8895904 -0.5916209
## STheta -0.9640874  0.7011695 -0.8895904  1.0000000  0.6661799
## Depthm -0.6669908  0.5649549 -0.5916209  0.6661799  1.0000000
```

Fig.19: Relationship Between Temperature - Scale of Salinity

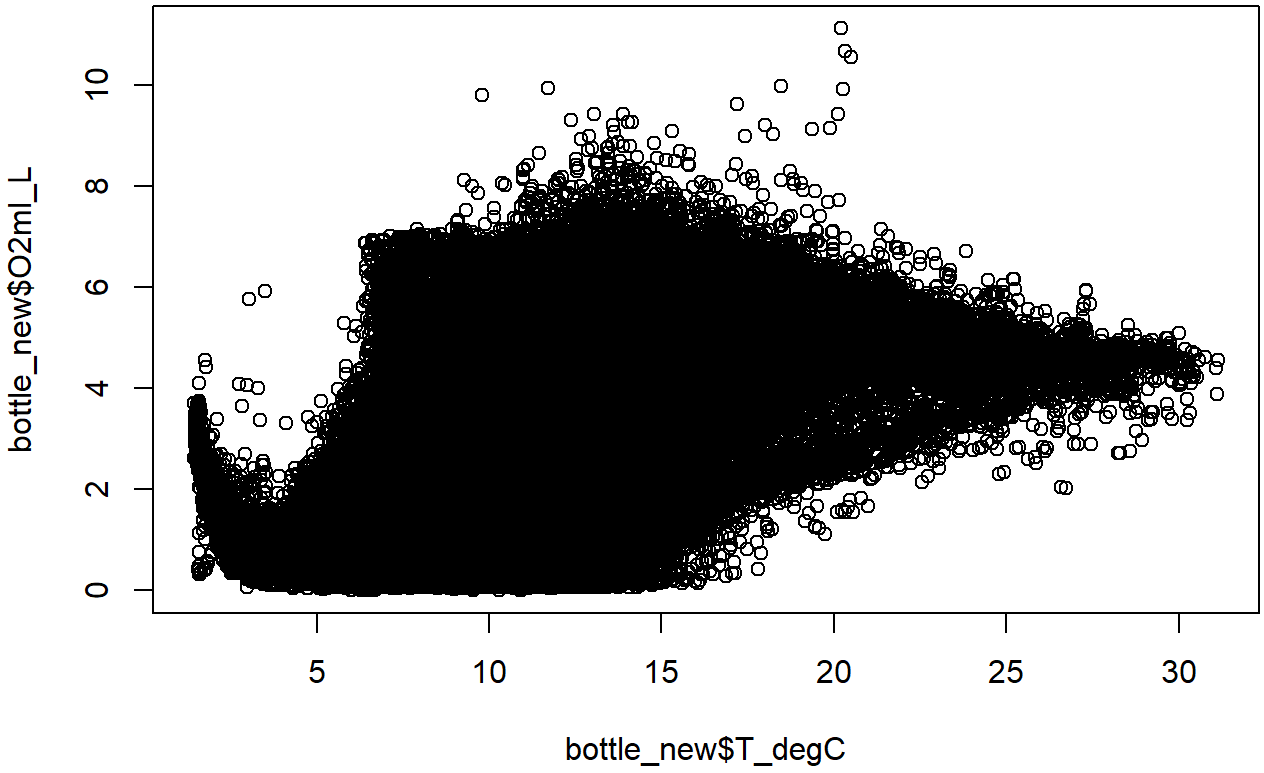
- The correlation coefficient of -0.5049812 shows that there is an inverse relationship between temperature and salinity variables. That is, as the temperature increases, salinity decreases. This relationship is quite strong because the correlation coefficient is close to -0.5. P-value indicates whether the correlation coefficient is statistically significant. If the p-value is less than 0.05, the correlation coefficient is considered to be statistically significant. In this case, a p-value < 2.2e-16 indicates that the correlation coefficient is statistically very significant.
- The 95% confidence interval shows the possible range of the true value of the correlation coefficient. In this case, the 95% confidence interval between -0.5067761 and -0.5031821 indicates that the correlation coefficient is close to -0.5.
- As a result, it can be said that there is an inverse and quite strong relationship between temperature and salinity variables. This relationship means that as the temperature increases, salinity decreases.

2. Temperature - Milliliters Oxygen per liter of Seawater

```
cor.test(bottle_new$T_degC,bottle_new$O2ml_L)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bottle_new$T_degC and bottle_new$O2ml_L
## t = 1064.5, df = 660240, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7940104 0.7957864
## sample estimates:
##       cor
## 0.7949001
```

```
plot(bottle_new$T_degC,bottle_new$O2ml_L)
```




```
cor(bottle_new[1:5])
```

```
##            T_degC     Salnty     O2ml_L     STheta     Depthm
## T_degC  1.0000000 -0.5049782  0.7949001 -0.9640874 -0.6669908
## Salnty -0.5049782  1.0000000 -0.8243646  0.7011695  0.5649549
## O2ml_L  0.7949001 -0.8243646  1.0000000 -0.8895904 -0.5916209
## STheta -0.9640874  0.7011695 -0.8895904  1.0000000  0.6661799
## Depthm -0.6669908  0.5649549 -0.5916209  0.6661799  1.0000000
```

Fig.20: Relationship Between Temperature - Milliliters Oxygen per liter of Seawater

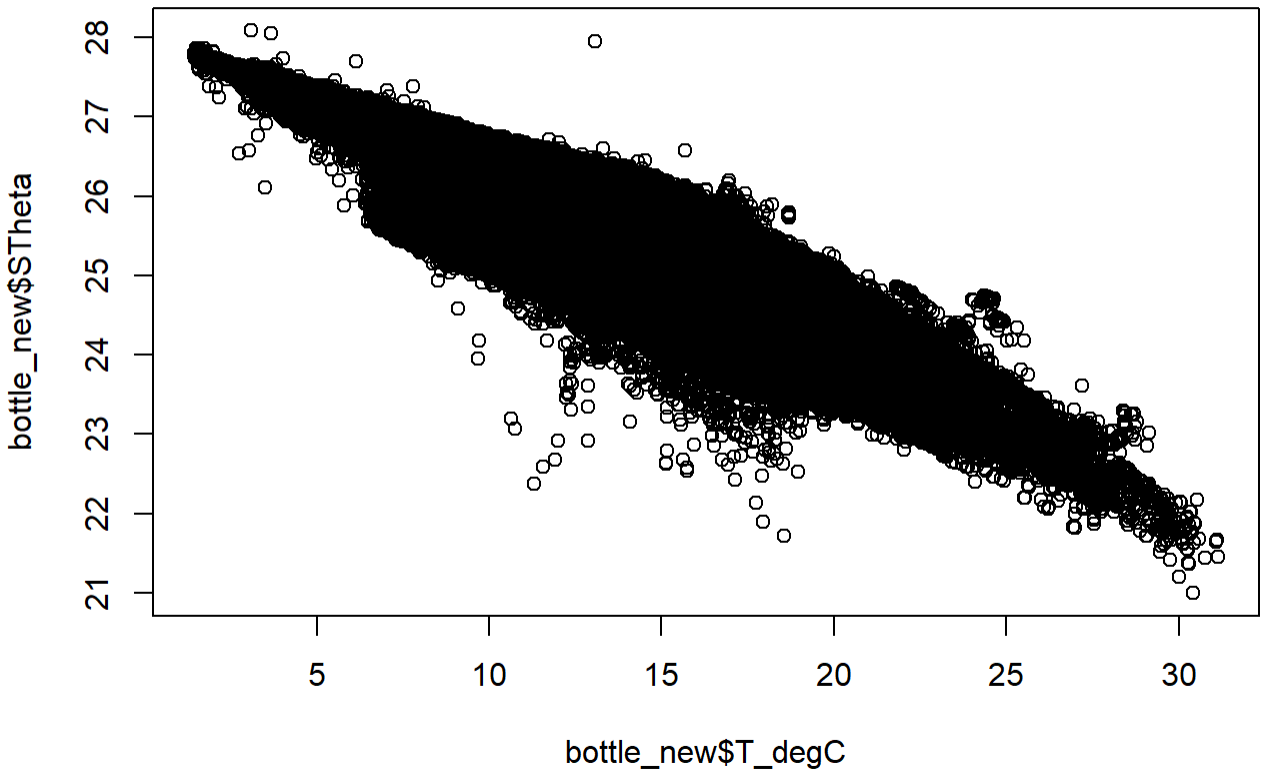
- Correlation coefficient is 0.7949021. This indicates that there is a strong positive correlation between the temperature and oxygen concentration variables. This means that as the temperature increases, the oxygen concentration also increases.

3. Temperature - Potential Density

```
cor.test(bottle_new$T_degC,bottle_new$STheta)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bottle_new$T_degC and bottle_new$STheta
## t = -2949.6, df = 660240, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9642571 -0.9639169
## sample estimates:
##        cor
## -0.9640874
```

```
plot(bottle_new$T_degC,bottle_new$STheta)
```




```
cor(bottle_new[1:5])
```

```
##            T_degC     Salnty     O2ml_L     STheta     Depthm
## T_degC  1.0000000 -0.5049782  0.7949001 -0.9640874 -0.6669908
## Salnty -0.5049782  1.0000000 -0.8243646  0.7011695  0.5649549
## O2ml_L  0.7949001 -0.8243646  1.0000000 -0.8895904 -0.5916209
## STheta -0.9640874  0.7011695 -0.8895904  1.0000000  0.6661799
## Depthm -0.6669908  0.5649549 -0.5916209  0.6661799  1.0000000
```

Fig.21: Relationship Between Temperature - Potential Density

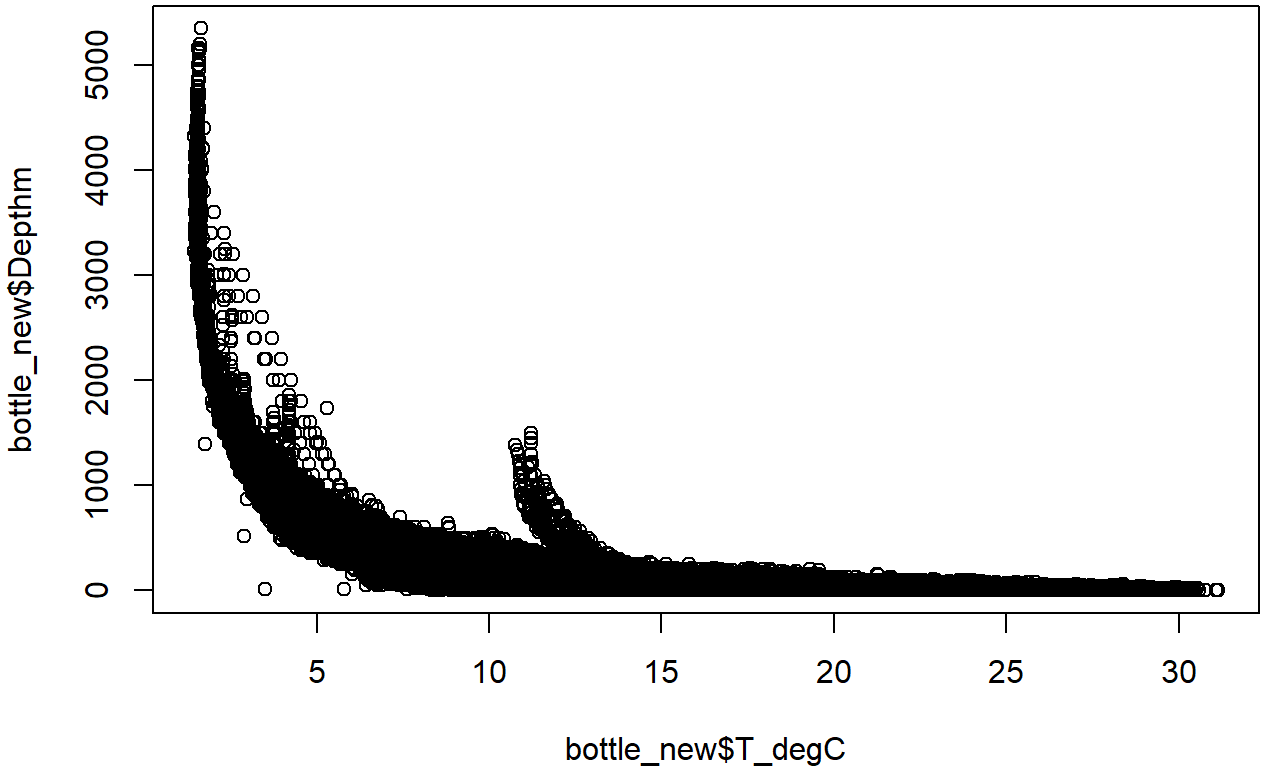
- The value of -0.9640877 indicates a very strong negative relationship between the two variables. The coefficient can take a value between -1 and +1, the closer to -1 the stronger the negative relationship, and the closer to +1 the stronger the positive relationship. A value of -0.9640877 indicates a near perfect negative correlation.

4. Temperature - Depth

```
cor.test(bottle_new$T_degC,bottle_new$Depthm)
```

```
##
##  Pearson's product-moment correlation
##
## data:  bottle_new$T_degC and bottle_new$Depthm
## t = -727.4, df = 660240, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6683277 -0.6656496
## sample estimates:
##        cor
## -0.6669908
```

```
plot(bottle_new$T_degC,bottle_new$Depthm)
```




```
cor(bottle_new[1:5])
```

```
##            T_degC     Salnty     O2ml_L     STheta     Depthm
## T_degC  1.0000000 -0.5049782  0.7949001 -0.9640874 -0.6669908
## Salnty -0.5049782  1.0000000 -0.8243646  0.7011695  0.5649549
## O2ml_L  0.7949001 -0.8243646  1.0000000 -0.8895904 -0.5916209
## STheta -0.9640874  0.7011695 -0.8895904  1.0000000  0.6661799
## Depthm -0.6669908  0.5649549 -0.5916209  0.6661799  1.0000000
```

Fig.22: Relationship Between Temperature - Depth

- Correlation coefficient is -0.6669926. This shows that there is a negative correlation between the two variables. That is, when one variable increases, the other variable decreases.

Linear Regression

- Linear regression establishes a linear relationship model that explains the effect of independent variables on the dependent variable.
- Dependent Variable (Y): The main variable to be predicted in the model. It depends on other variables that affect the value of this variable.
- Simple Linear Regression Model: $Y=\beta0+\beta1X+\epsilon$ 23

- Independent Variables (X1, X2, …): Variables that are expected to have an effect on the dependent variable. The values of these variables are used to predict the value of the dependent variable.
- β0: Intercept term, the point where the line crosses the Y-axis. β1,β2,β3,…,βn: Regression coefficients, representing the effects of the independent variable. X 1,X 2,…,X n: Values of the independent variables.
- ϵ: The error term is a simplifying error term of the model that does not reflect real-world complexities.

```
library(dplyr)
model1 <- lm(T_degC ~ Depthm, data=bottle_new)
summary(model1)
```

```
##
## Call:
## lm(formula = T_degC ~ Depthm, data = bottle_new)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.336 -2.334 -1.094  1.703 37.428
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.292e+01  4.746e-03  2721.8   <2e-16 ***
## Depthm      -9.102e-03  1.251e-05  -727.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.147 on 660240 degrees of freedom
## Multiple R-squared:  0.4449, Adjusted R-squared:  0.4449
## F-statistic: 5.291e+05 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
coeff1 <- model1$coefficients
intercept1 <- coeff1[1]
slope1 = coeff1[2]
confint(model1, level=.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) 12.907751263 12.926354124
## Depthm      -0.009126557 -0.009077507
```

```
ggplot(bottle_new, aes(x=Depthm, y=T_degC)) +
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = intercept1, slope = slope1, color= "#371f63")
```

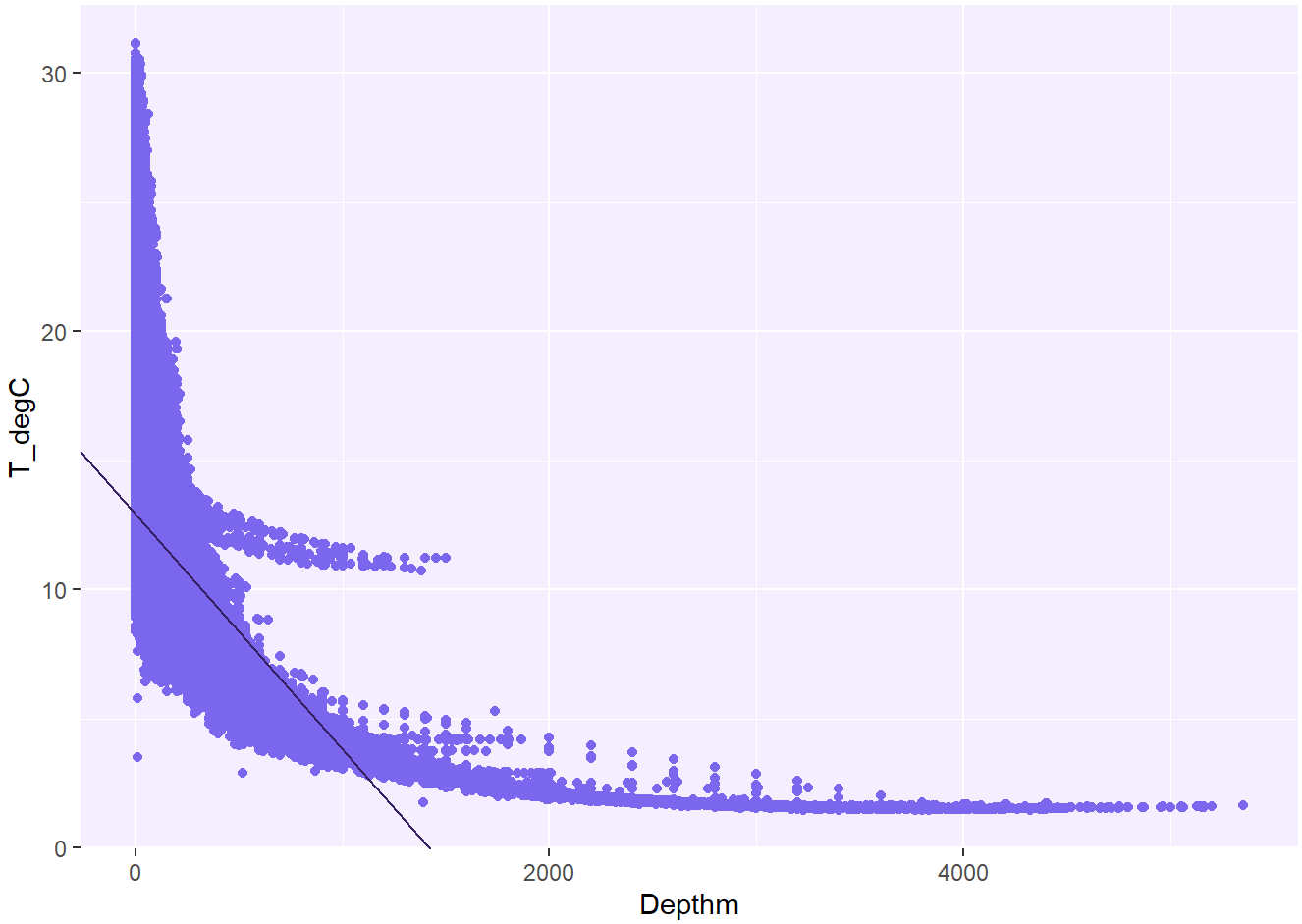


Fig.23: Temperature - Depth Graph with Slope of Model1

- Since the p-values for both coefficients are much smaller than 0.001, we can say that there is a statistically significant relationship between depth and temperature.
- Multiple R-squared: Indicates that the model can explain 44.49% of the variance in temperature variability.
- Adjusted R-squared: Provides a more realistic estimate by correcting the R-squared value in models with many variables.
- Gives 95% confidence intervals for where the true values of the coefficients are likely to be.
- It can be concluded that temperature decreases in a statistically significant way as depth increases.

```
ggplot(bottle_new, aes(x=Depthm, y=resid(model1)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0, slope = 0, color = '#371f63')
```

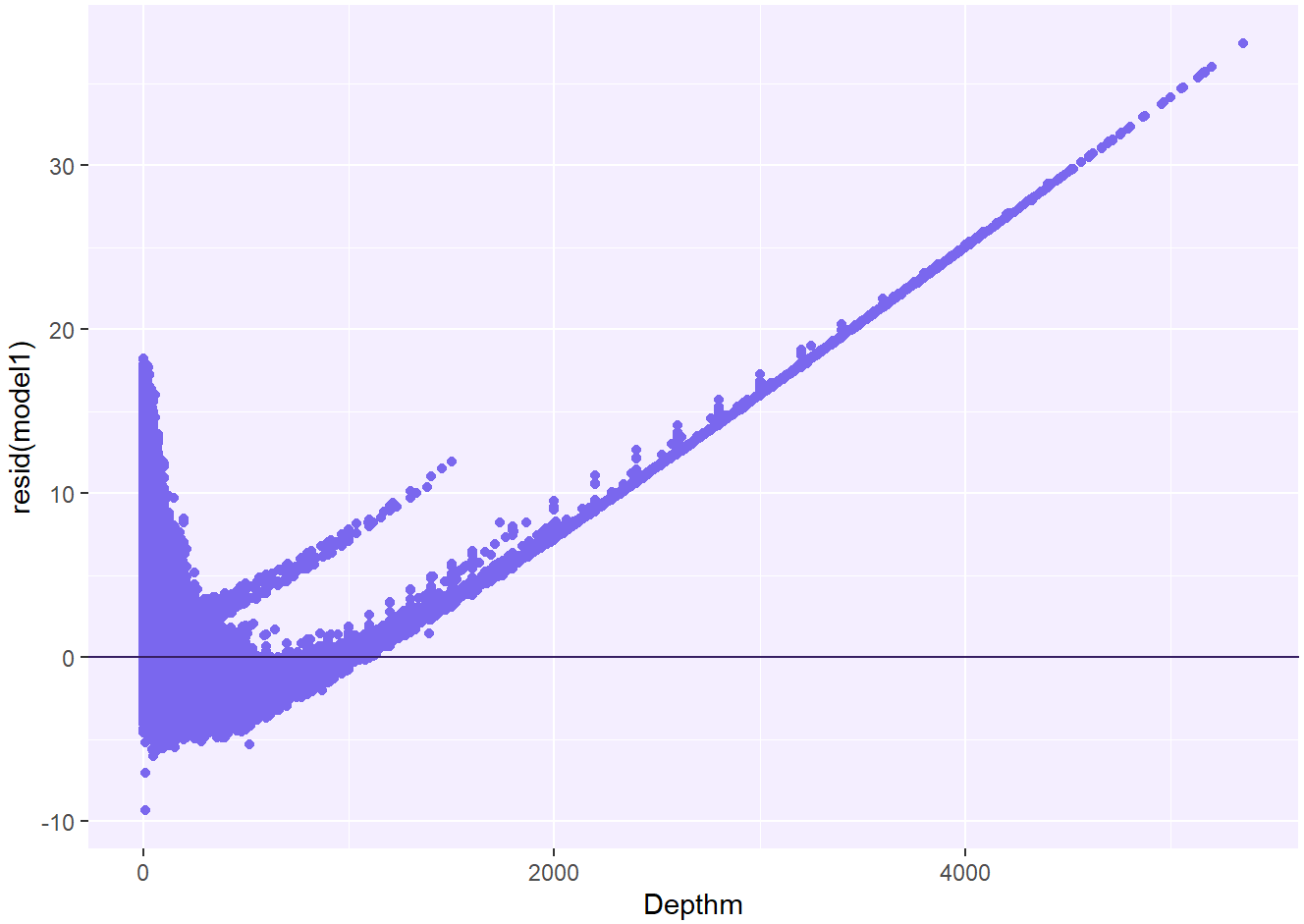


Fig.24: Depth - Residual(Model1) Graph

- In regression analysis, the term residual refers to the difference between predicted values and observed values. That is, when a regression model attempts to measure the effect of an independent variable on a dependent variable, residuals are used to assess how well that model fits real-world data.
- The distribution of the residuals is very close to a normal distribution. This is a positive sign about the assumptions of the model. A few small deviations can be seen in the graph. These deviations indicate that the model is not perfect. In conclusion, this graph shows that the model1 regression model has a good fit and its predictions are reliable.

```
model2 <- lm(T_degC ~ Salnty, data=bottle_new)
summary(model2)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.147  -2.287  -1.038   1.485  29.868
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167.659774   0.329738   508.5   <2e-16 ***
## Salnty       -4.632779   0.009745  -475.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 660240 degrees of freedom
## Multiple R-squared:  0.255,  Adjusted R-squared:  0.255
## F-statistic: 2.26e+05 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
coeff2 <- model2$coefficients
intercept2 <- coeff2[1]
slope2 = coeff2[2]
confint(model2, )
```

```
##                 2.5 %     97.5 %
## (Intercept) 167.013498 168.306049
## Salnty       -4.651879  -4.613678
```

```
ggplot(bottle_new, aes(x=Salnty, y=T_degC)) +
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = intercept2, slope = slope2, color="#371f63")
```

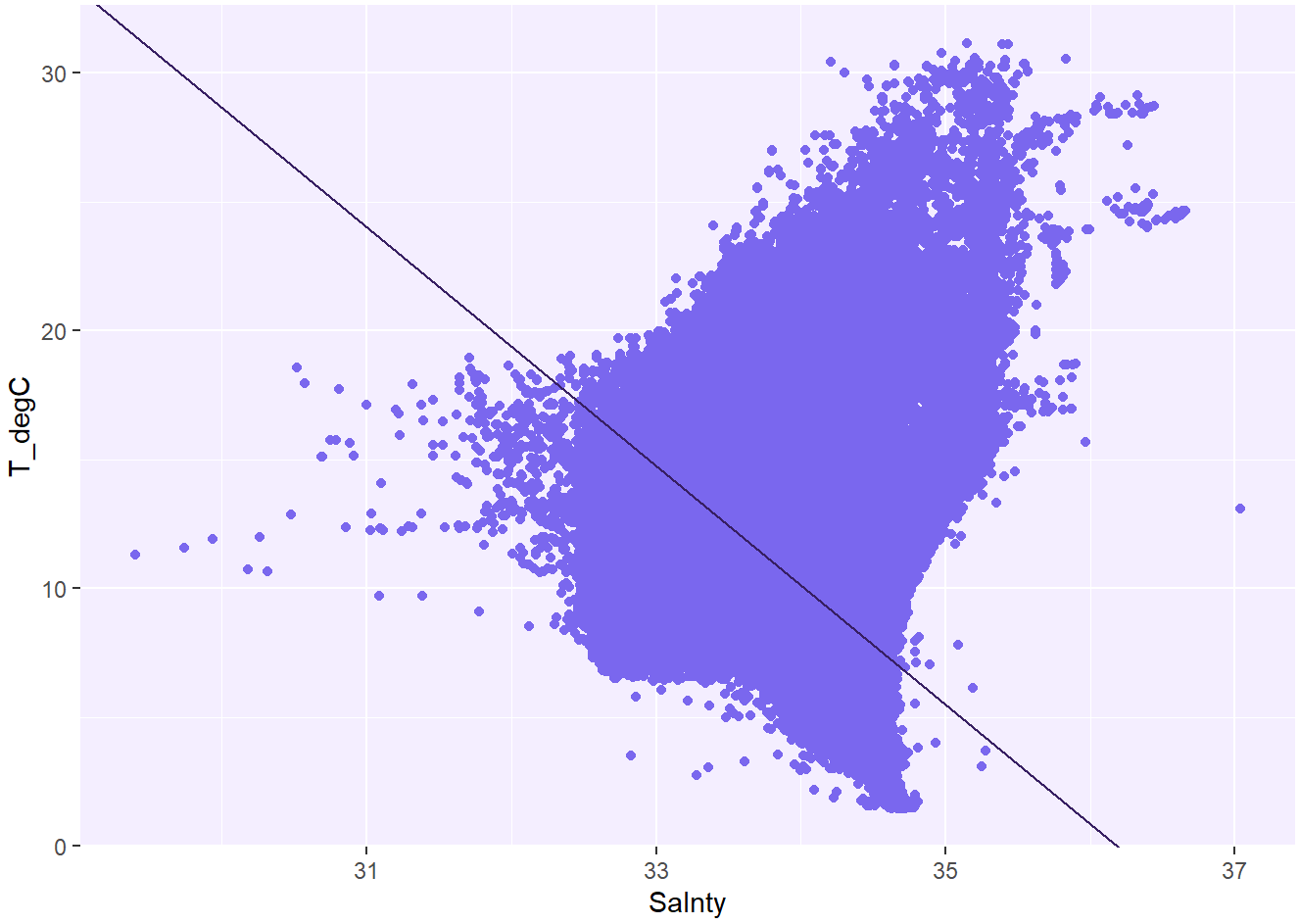


Fig.25: Temperature - Salinity Graph with Slope Model2

- There is a strong negative relationship between salinity and temperature. As salinity increases, temperature decreases. The model can explain about 25.57% of the change in temperature.

```
ggplot(bottle_new, aes(x=Salnty, y=resid(model2)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0, slope = 0, color = '#371f63')
```

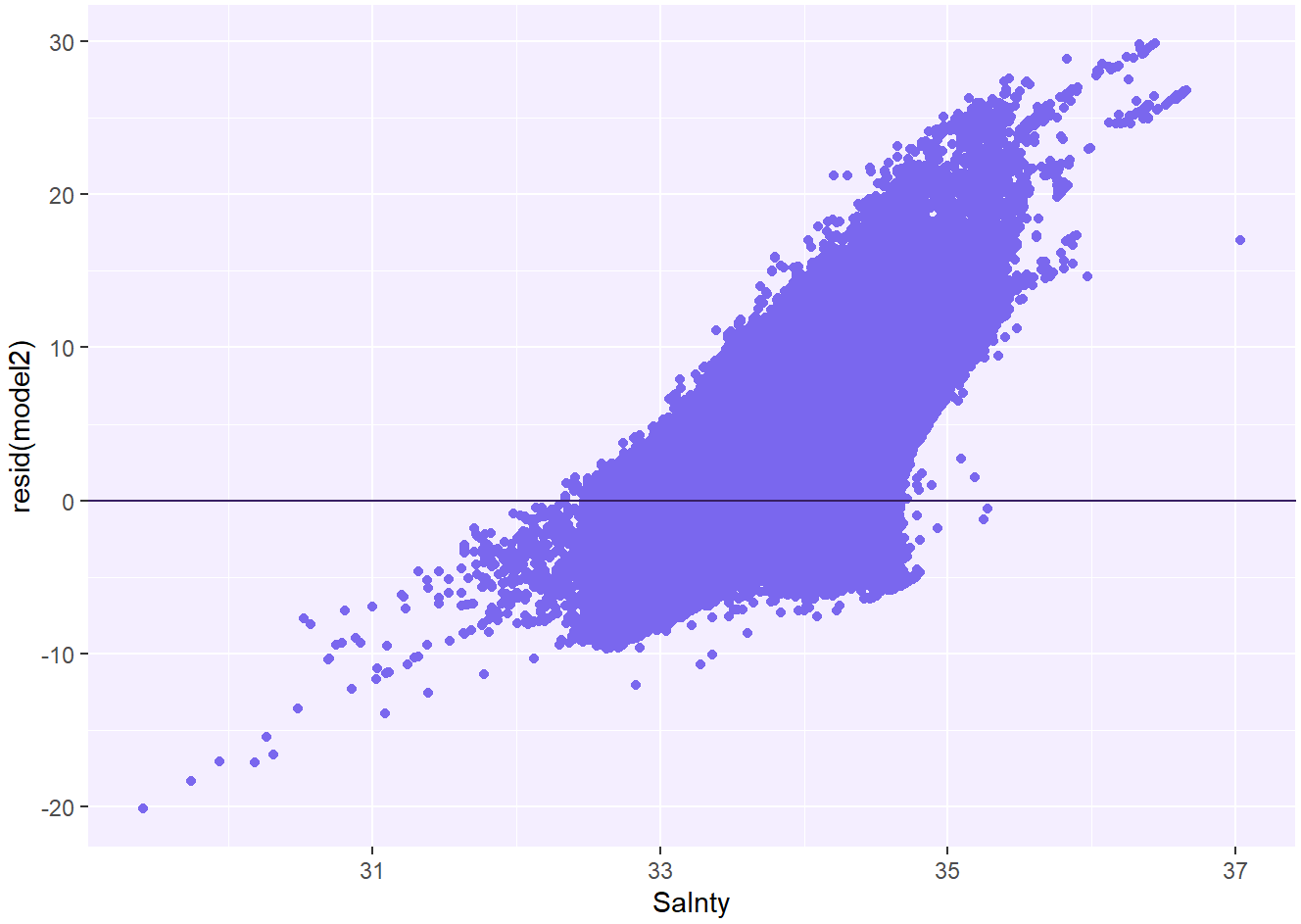


Fig.26: Salinity - Residual(Model2) Graph

- The graph shows that the residuals are approximately normally distributed. This indicates that the model represents the data well and its predictions are reliable. The residuals are concentrated in the center of the range. This indicates that the model's predictions are generally accurate. However, the graph shows that a few residuals are outside the range. This could mean that the model makes errors in some cases. Model2 shows that the regression model represents the data well and its predictions are reliable.

```
model3 <- lm(T_degC ~ O2ml_L, data=bottle_new)
summary(model3)
```

```
##
## Call:
## lm(formula = T_degC ~ O2ml_L, data = bottle_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7002  -1.5116  -0.4103   0.9381  19.4479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.373746   0.006091   882.2   <2e-16 ***
## O2ml_L      1.623300   0.001525  1064.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.562 on 660240 degrees of freedom
## Multiple R-squared:  0.6319, Adjusted R-squared:  0.6319
## F-statistic: 1.133e+06 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
coeff3 <- model3$coefficients
intercept3 <- coeff3[1]
slope3 = coeff3[2]
confint(model3, )
```

```
##                2.5 %   97.5 %
## (Intercept) 5.361807 5.385685
## O2ml_L      1.620311 1.626289
```

```
ggplot(bottle_new, aes(x=O2ml_L, y=T_degC)) +
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = intercept3, slope = slope3, color="#371f63")
```

Fig.27: Temperature - O2 Graph with Slope of Model3

Residual standard error: 2.562 indicates the typical error of the model.

- Multiple R-squared: 0.6319 indicates that oxygen content explains 63.19% of the temperature variability.
- Adjusted R-squared: 0.6319 is the R-squared value adjusted for model complexity.
- F-statistic: 1.133e+06 indicates that the model is statistically significant overall.

There is a positive linear relationship between oxygen content and temperature. As the amount of oxygen increases, the temperature is expected to increase.

```
ggplot(bottle_new, aes(x=O2ml_L, y=resid(model3)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0, slope = 0, color = '#371F63')
```

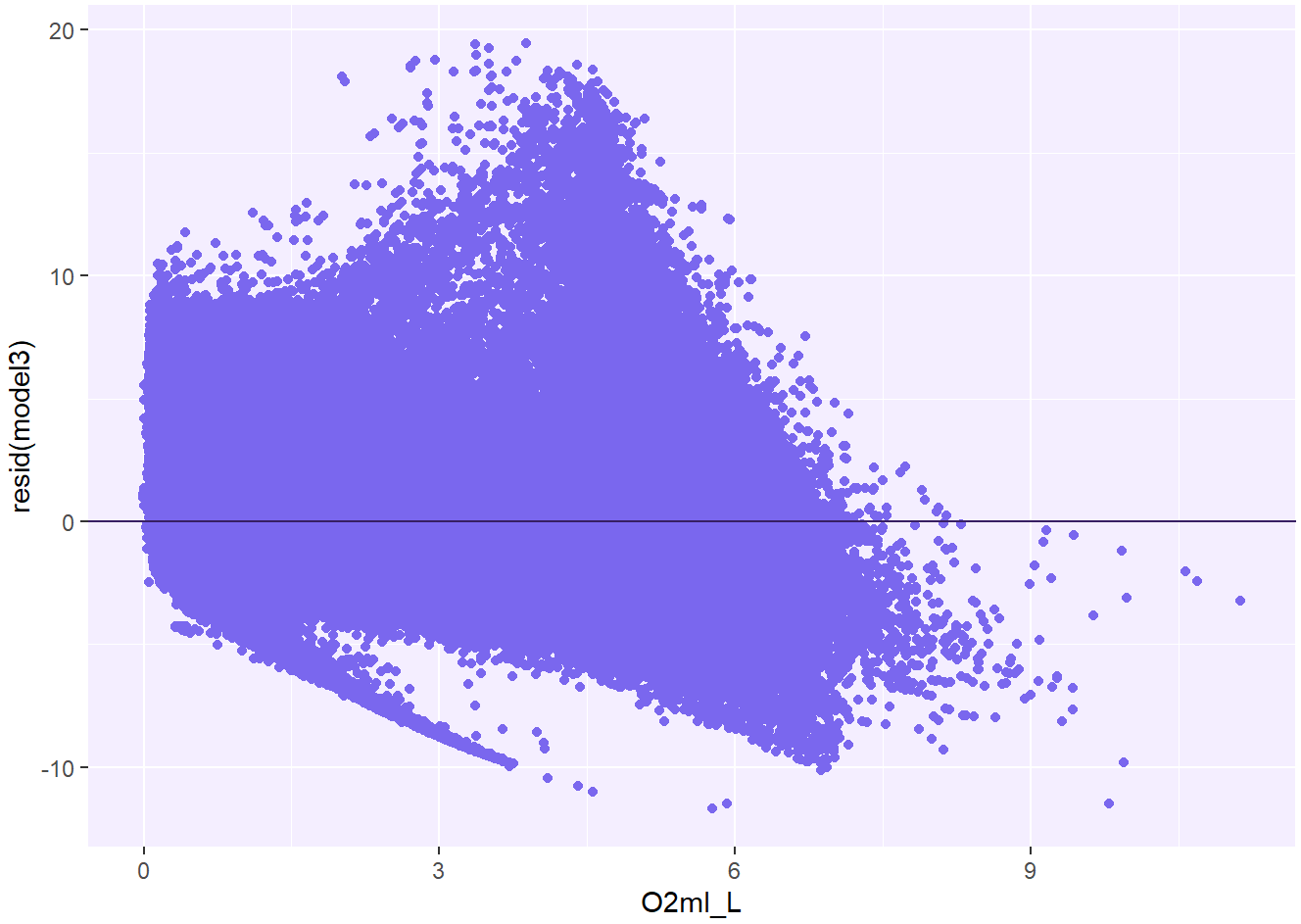


Fig.28: O2 - Residual(Model3) Graph

- It shows that the residuals are approximately normally distributed. This means that the model provides a good fit. The graph also shows some deviations, which may mean that the model has difficulty in predicting some data. Overall, the graph shows that the model provides a good fit and its predictions are accurate.

```
model4 <- lm(T_degC ~ STheta, data=bottle_new)
summary(model4)
```

```
##
## Call:
## lm(formula = T_degC ~ STheta, data = bottle_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4493  -0.5794  -0.1044   0.3843  10.8596
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.981929   0.035306    3257   <2e-16 ***
## STheta       -4.033826   0.001368   -2950   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.122 on 660240 degrees of freedom
## Multiple R-squared:  0.9295, Adjusted R-squared:  0.9295
## F-statistic: 8.7e+06 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
coeff4 <- model4$coefficients
intercept4 <- coeff4[1]
slope4 = coeff4[2]
confint(model4, )
```

```
##                  2.5 %     97.5 %
## (Intercept) 114.912729 115.051128
## STheta       -4.036506  -4.031145
```

```
ggplot(bottle_new, aes(x=STheta, y=T_degC)) +
    geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = intercept4, slope = slope4, color="#371f63")
```

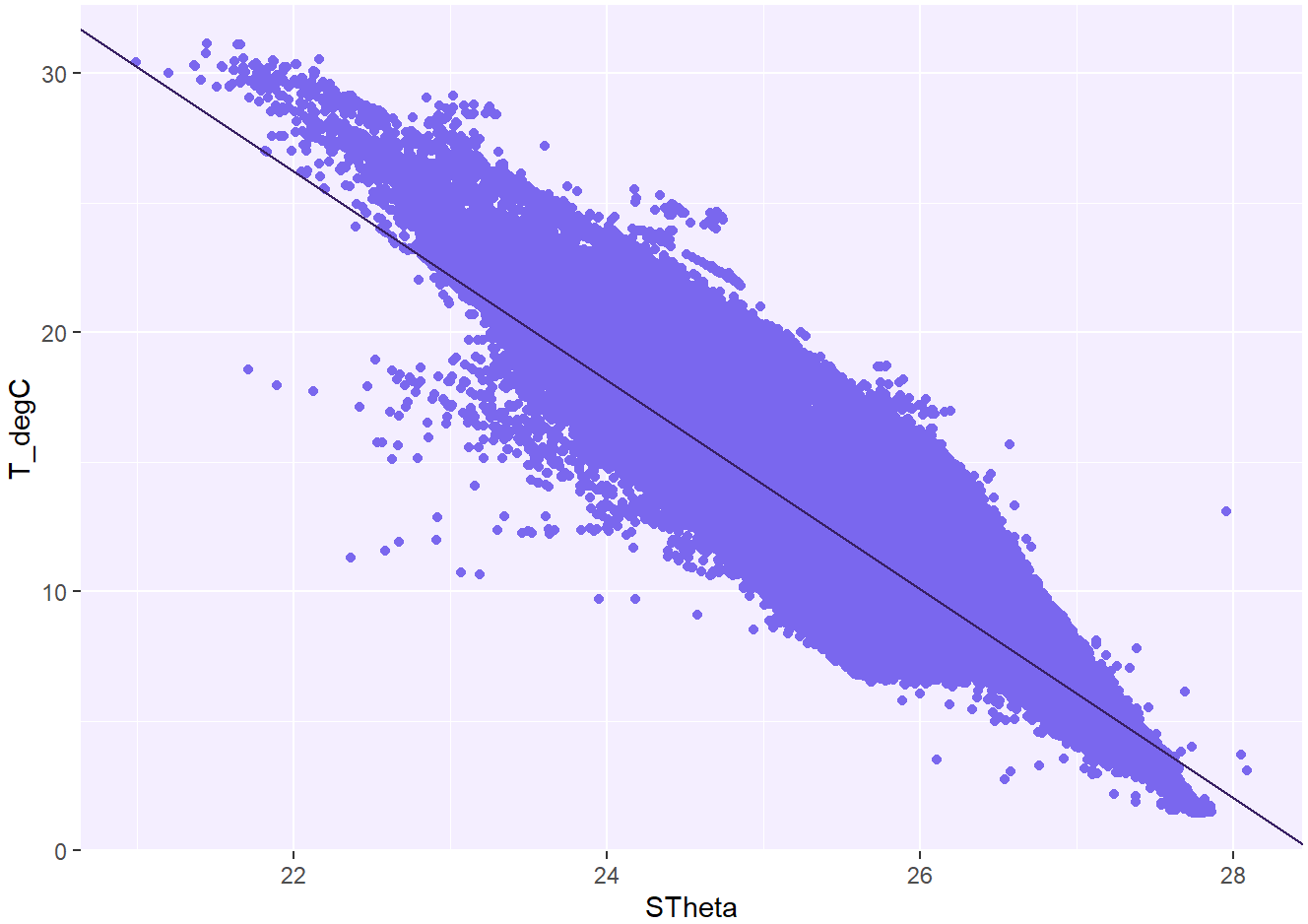


Fig.29: Temperature - Potential Density Graph with Slope of Model4

Adjusted R-squared: 0.9295: The R-squared value adjusted for the number of variables in the model is also 0.9295, which indicates a high fit.

- There is a strong, negative and linear relationship between STheta and T_degC.
- As the STheta value increases, the T_degC value is expected to decrease.
- The model can predict T_degC values with relatively high accuracy.

```
ggplot(bottle_new, aes(x=STheta, y=resid(model4)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
    geom_abline(intercept = 0, slope = 0, color = '#371F63')
```

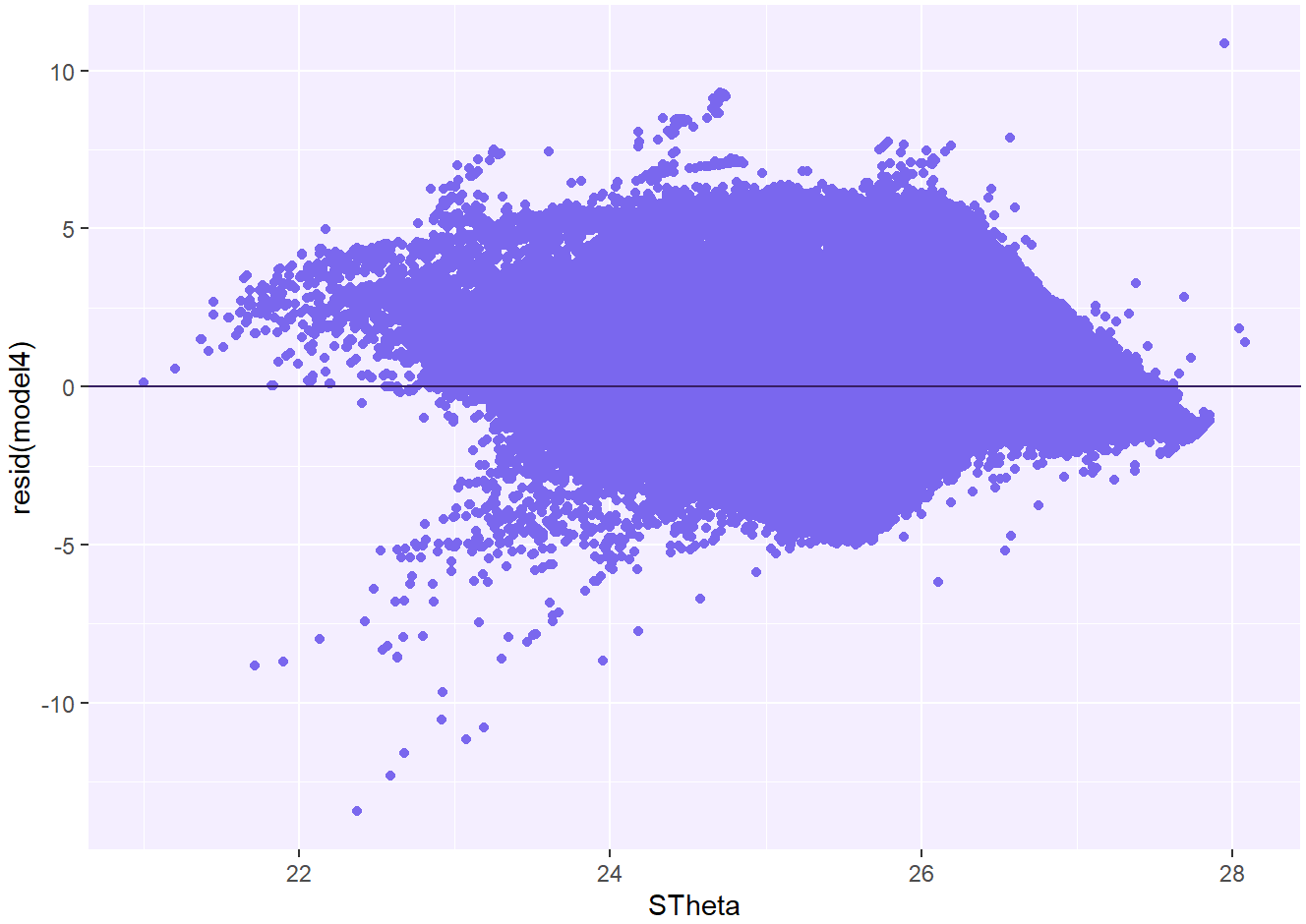


Fig.30: Potential Density - Residual(Model4) Graph

- The residuals show an approximately normal distribution. This indicates that the model provides a good.

Multiple Linear Model

- Multiple linear regression is a statistical model in which a dependent variable is explained by more than one independent variable. This model uses the independent variables to predict the mean of the dependent variable.

```
catmodel1 <- lm( T_degC ~ Salnty + O2ml_L, data = bottle_new)
summary(catmodel1)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty + O2ml_L, data = bottle_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3820  -1.2446  -0.1135   0.9876  18.1531
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.429e+02  3.750e-01  -381.1   <2e-16 ***
## Salnty       4.304e+00  1.088e-02   395.5   <2e-16 ***
## O2ml_L       2.413e+00  2.422e-03   996.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.304 on 660239 degrees of freedom
## Multiple R-squared:  0.7024, Adjusted R-squared:  0.7024
## F-statistic: 7.791e+05 on 2 and 660239 DF,  p-value: < 2.2e-16
```

- Multiple R-squared: 0.7024

The model explains 70.24% of the variation in T_degC.

- Adjusted R-squared: 0.7024

The R-squared value adjusted for the complexity of the model indicates that the model is not over-fitting.

- F-statistic: 7.791e+05, p-value < 2.2e-16

Indicates that the model is generally significant.

- It shows that both salinity and oxygen content have a positive and significant relationship with temperature.

```
catmodel2 <- lm( T_degC ~ Salnty + O2ml_L + STheta + Depthm, data = bottle_new)
summary(catmodel2)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty + O2ml_L + STheta + Depthm, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9102 -0.1513  0.0198  0.1350  4.6759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289e+00  6.658e-02   139.5   <2e-16 ***
## Salnty       3.542e+00  1.627e-03  2177.4   <2e-16 ***
## O2ml_L       1.451e-01  5.556e-04   261.2   <2e-16 ***
## STheta      -4.588e+00  9.796e-04 -4683.9   <2e-16 ***
## Depthm      -1.536e-03  1.822e-06  -842.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 660237 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9938
## F-statistic: 2.653e+07 on 4 and 660237 DF,  p-value: < 2.2e-16
```

- The p-values of all variables are much smaller than 0.001, meaning that all variables in the model are statistically significant. This indicates that there are strong relationships between temperature and these variables.
- R-squared: 0.9938. This is a very high value, indicating that the model can explain 99.38% of the temperature change.
- Adjusted R-squared: 0.9938. This is also very high, indicating that the model is not affected by over-fitting of the data.
- The model explains the temperature quite well and it is seen that Salinity, Oxygen Content, Potential Density and Depth have significant effects on temperature.
- The model can be used to predict new temperature values.

Conclusion

- The catmodel2 is able to explain the temperature better because it includes more explanatory variables.
- The high R-square value and low residual standard error of the catmodel2 indicate that the model fits the observed data very well.
- The addition of STheta and Depthm variables to the model made a significant contribution to the prediction of temperature.
- The catmodel2 should be preferred over the catmodel1 in predicting temperature due to its higher explanatory power and better fit.

Transformation

It is used to correct the distribution of variables in the data set or to provide model assumptions. These

transformations may include mathematical operations such as logarithmic and square root. Especially in models such as regression analysis, transformations are often used to approximate the normal distribution of variables or to ensure homoscedasticity of errors. In addition, transformations to eliminate outliers or to make the distribution symmetric can increase the reliability of statistical analyses.

```
ggplot(bottle_new, aes(x=Salnty,y=T_degC))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))
```

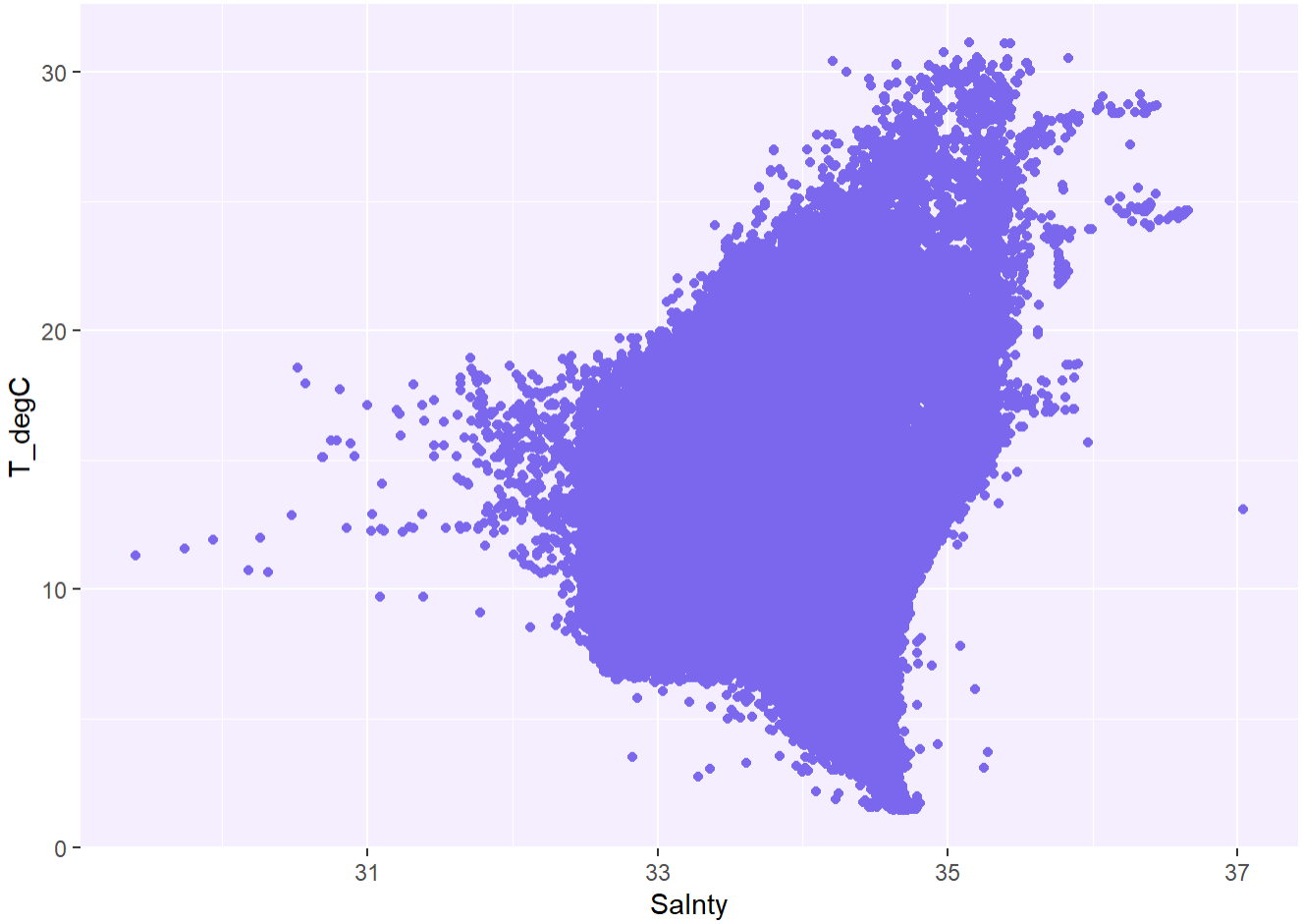


Fig.31: Temperature - Salinity Graph

```
ggplot(bottle_new, aes(x=Salnty,y=sqrt(T_degC)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))
```

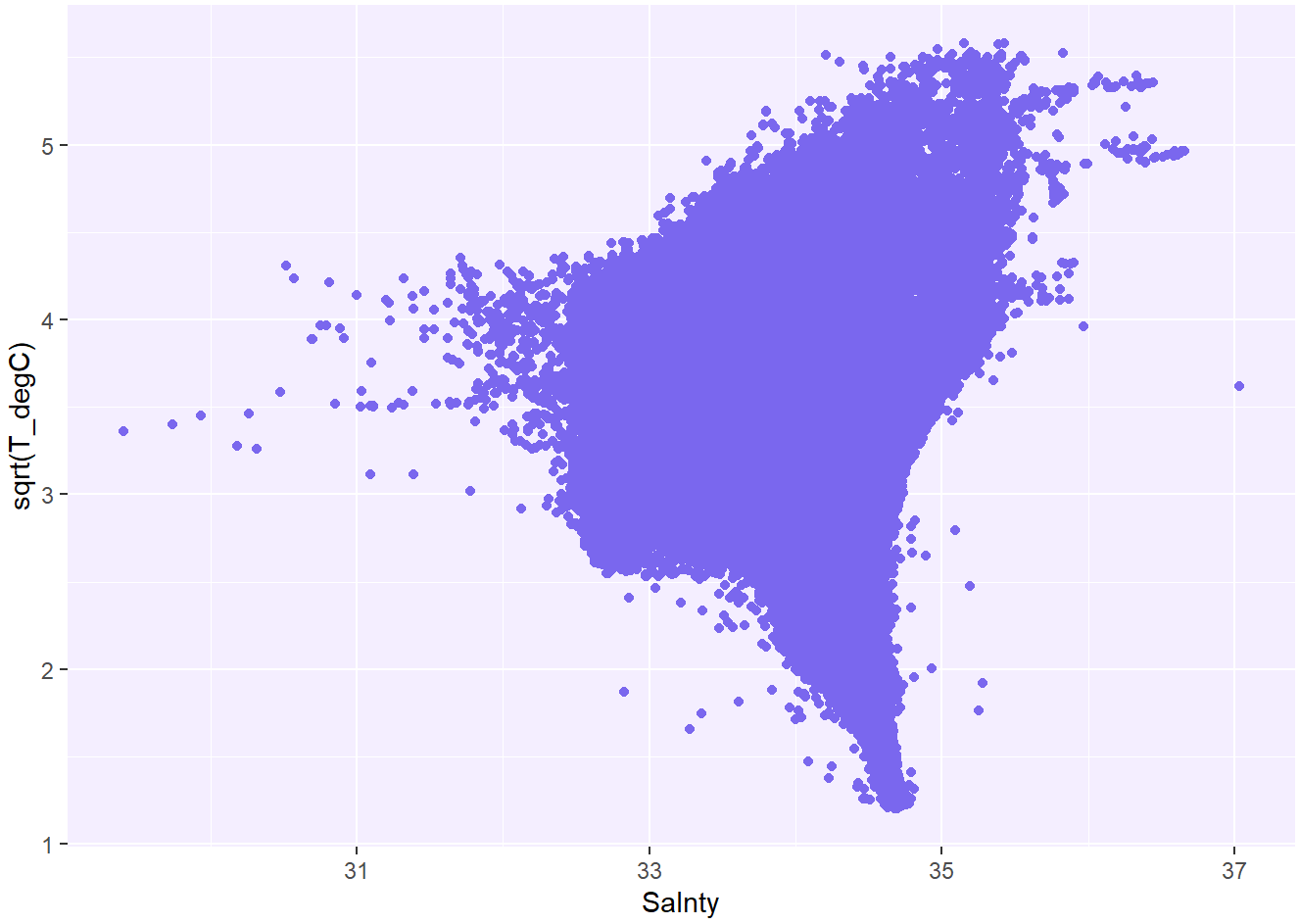


Fig.32: Squared Temperature - Salinity Graph

```
ggplot(bottle_new, aes(x=Salnty,y=log(T_degC)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))
```

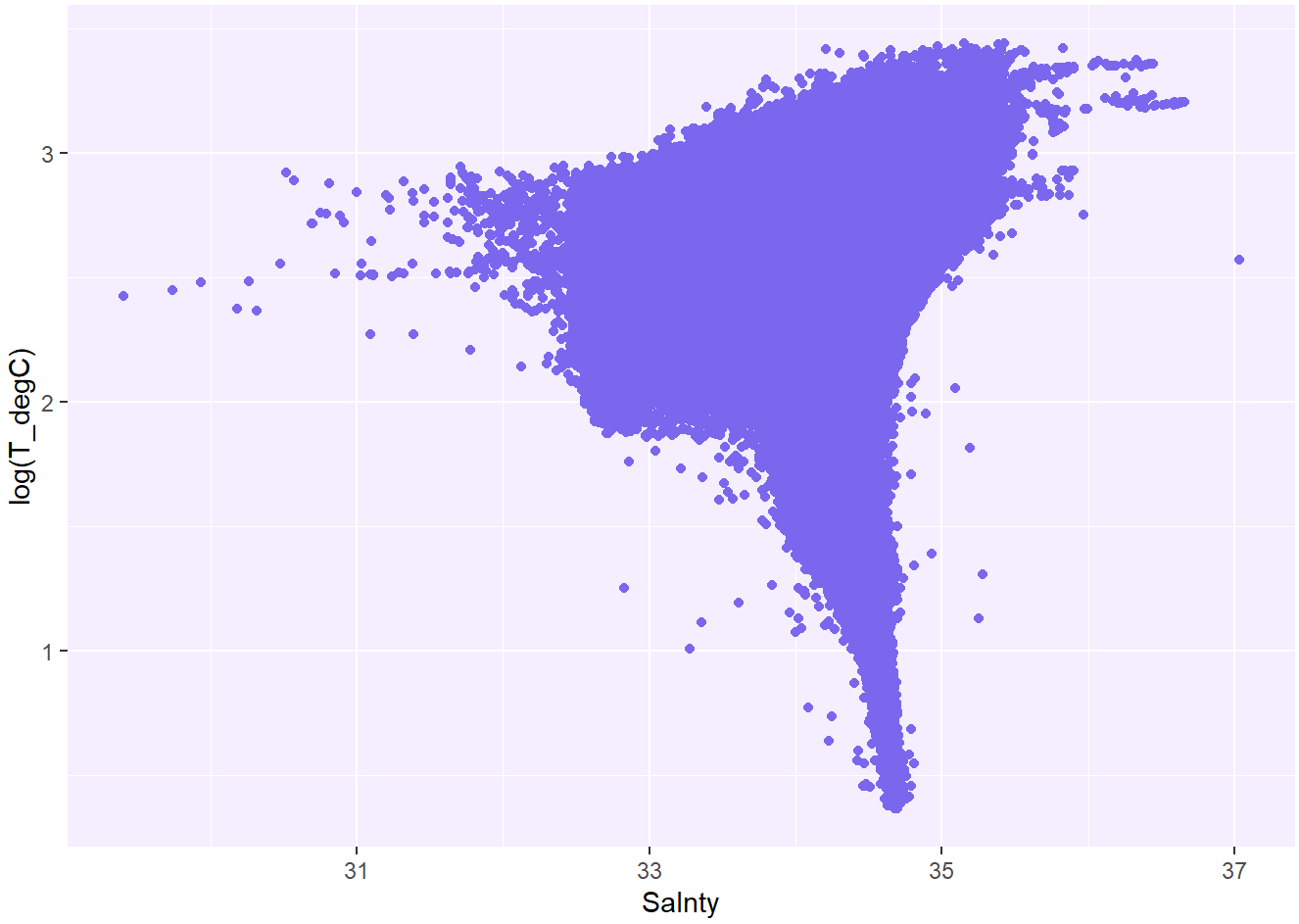


Fig.33: Log. Temperature - Salinity Graph

```
model_1<-lm(T_degC~Salnty, data=bottle_new)
summary(model_1)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.147  -2.287  -1.038   1.485  29.868
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167.659774   0.329738   508.5   <2e-16 ***
## Salnty       -4.632779   0.009745  -475.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 660240 degrees of freedom
## Multiple R-squared:  0.255,  Adjusted R-squared:  0.255
## F-statistic: 2.26e+05 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
ggplot(bottle_new,aes(x=Salnty,y=resid(model_1)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0,slope=0,color="#371F63")
```

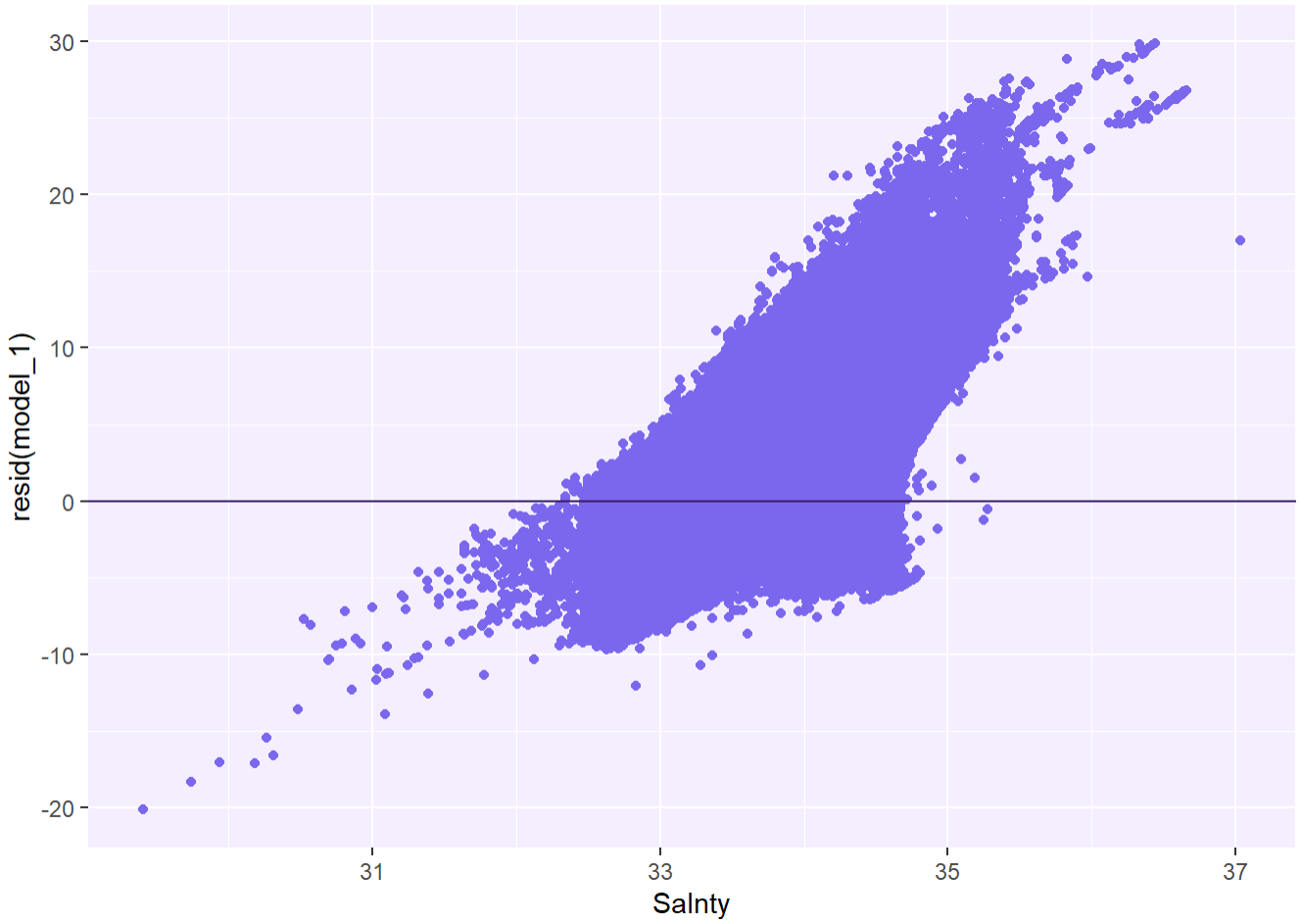


Fig.34: Temperature - Salinity Graph with Residual

```
model_2<- lm(log(T_degC)~Salnty,data=bottle_new)
summary(model_2)
```

```
##
## Call:
## lm(formula = log(T_degC) ~ Salnty, data = bottle_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22199 -0.17666 -0.03839  0.15244  2.42454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.1598351  0.0311895   646.4   <2e-16 ***
## Salnty      -0.5276187  0.0009218  -572.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3448 on 660240 degrees of freedom
## Multiple R-squared:  0.3316, Adjusted R-squared:  0.3316
## F-statistic: 3.276e+05 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
ggplot(bottle_new,aes(x=Salnty,y=resid(model_2)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0,slope=0,color="#371F63")
```

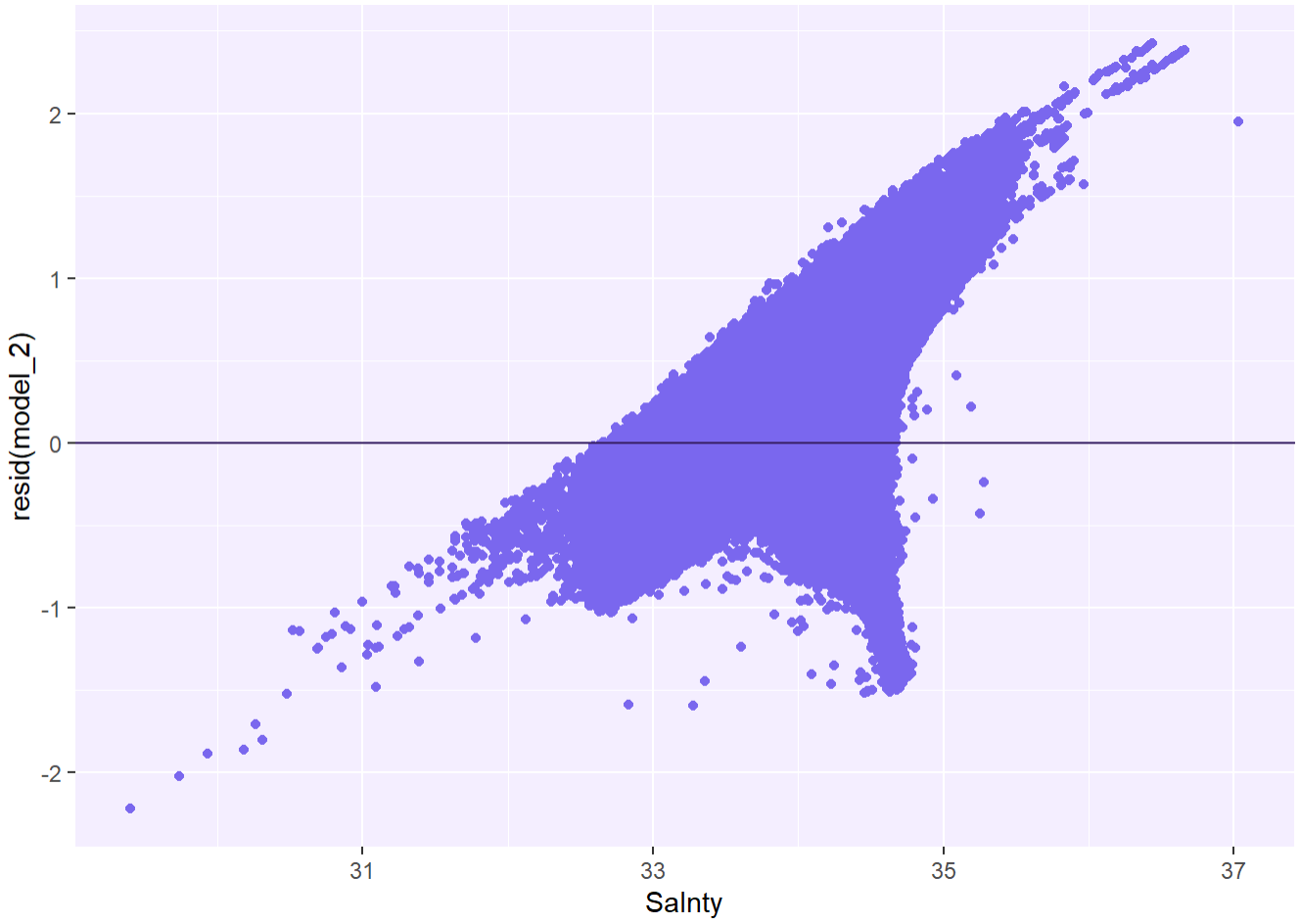


Fig.35: Log. Temperature - Salinity Graph with Residual

```
model_3<- lm(sqrt(T_degC)~Salnty,data=bottle_new)
summary(model_3)
```

```
##
## Call:
## lm(formula = sqrt(T_degC) ~ Salnty, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2959 -0.3195 -0.1109  0.2375  4.1287
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.330598   0.048893   599.9   <2e-16 ***
## Salnty      -0.771145   0.001445  -533.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5405 on 660240 degrees of freedom
## Multiple R-squared:  0.3014, Adjusted R-squared:  0.3014
## F-statistic: 2.848e+05 on 1 and 660240 DF,  p-value: < 2.2e-16
```

```
ggplot(bottle_new,aes(x=Salnty,y=resid(model_3)))+
  geom_point(color="#7a67ee")+
  theme(panel.background = element_rect(fill = "#f4eeff"))+
  geom_abline(intercept = 0,slope=0,color="#371f63")
```

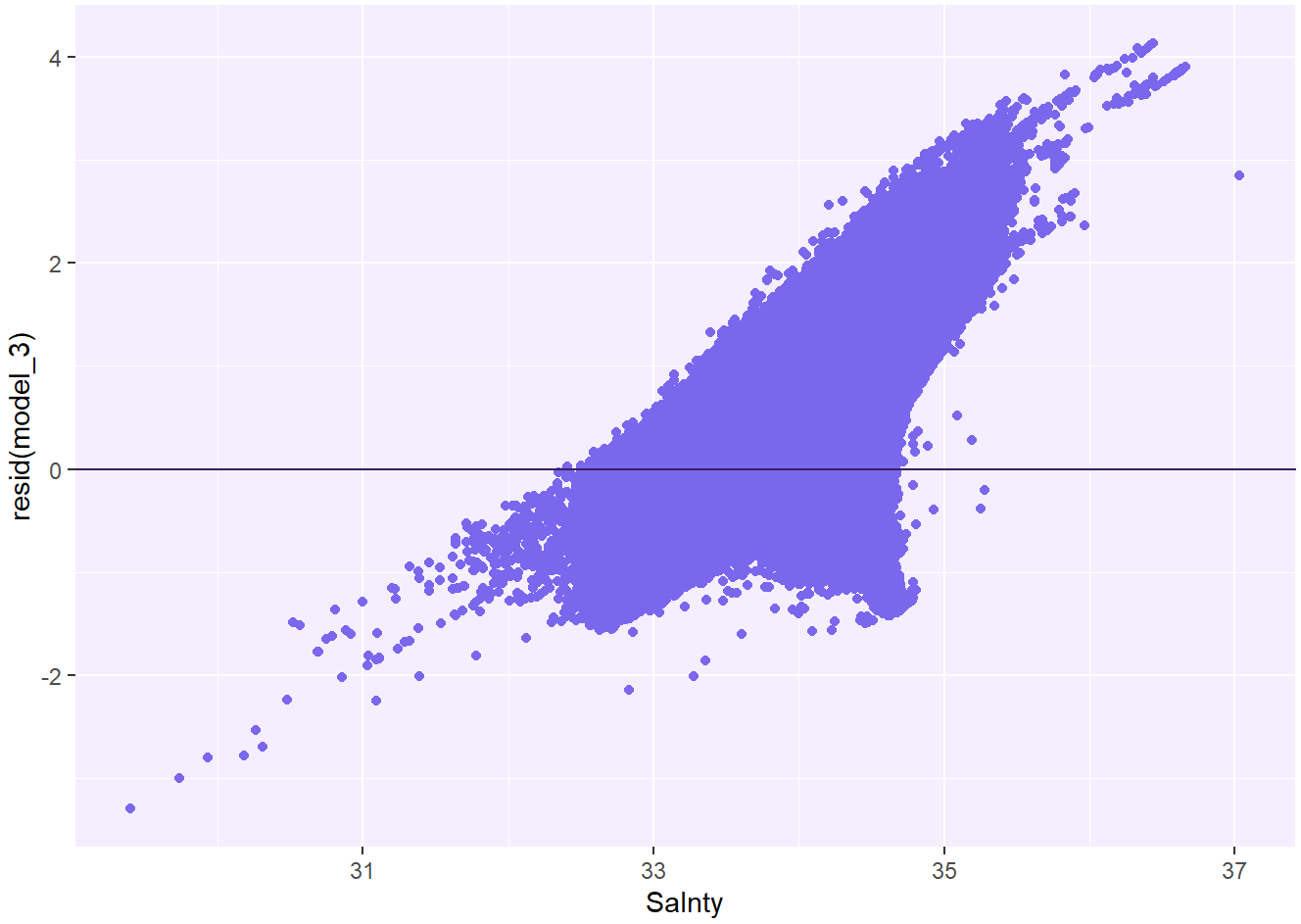


Fig.36: Squared Temperature - Salinity with Residual

- In conclusion, The fact that the logarithmic model provides the best fit indicates that the relationship between temperature and salinity is non-linear. This means that changes in salinity have a greater effect on temperature. Among the three models, the log(T_degC) ~ Salnty model provides the best fit as it has the highest R-square value and the lowest residual standard error.

Step-Wise Regression

1. Backward Selection

- This method constructs a model starting with all variables and then tries to remove the insignificant variables from the model one by one. By looking at the AIC (Akaike Information Criterion) values, it determines which variable should be removed.

```
max_model<-lm(data=bottle_new, T_degC~Salnty+O2ml_L+STheta+Depthm)
summary(max_model)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty + O2ml_L + STheta + Depthm, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9102 -0.1513  0.0198  0.1350  4.6759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289e+00  6.658e-02   139.5   <2e-16 ***
## Salnty       3.542e+00  1.627e-03  2177.4   <2e-16 ***
## O2ml_L       1.451e-01  5.556e-04   261.2   <2e-16 ***
## STheta      -4.588e+00  9.796e-04 -4683.9   <2e-16 ***
## Depthm      -1.536e-03  1.822e-06  -842.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 660237 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9938
## F-statistic: 2.653e+07 on 4 and 660237 DF,  p-value: < 2.2e-16
```

```
bw_step <- step(max_model, direction = "backward")
```

```
## Start:  AIC=-1455520
## T_degC ~ Salnty + O2ml_L + STheta + Depthm
##
##          Df Sum of Sq     RSS      AIC
## <none>                  72825 -1455520
## - O2ml_L  1      7525   80351 -1390596
## - Depthm  1     78316  151142  -973445
## - Salnty  1    522963  595788   -67813
## - STheta  1   2419903 2492728   877157
```

```
summary(bw_step)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty + O2ml_L + STheta + Depthm, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9102 -0.1513  0.0198  0.1350  4.6759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289e+00  6.658e-02   139.5   <2e-16 ***
## Salnty       3.542e+00  1.627e-03  2177.4   <2e-16 ***
## O2ml_L       1.451e-01  5.556e-04   261.2   <2e-16 ***
## STheta      -4.588e+00  9.796e-04 -4683.9   <2e-16 ***
## Depthm      -1.536e-03  1.822e-06  -842.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 660237 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9938
## F-statistic: 2.653e+07 on 4 and 660237 DF,  p-value: < 2.2e-16
```

- AIC: AIC is a model selection criterion that balances model fit and model complexity. Lower AIC values indicate better models.
- As a result of backward selection, no variables were excluded from the model. This indicates that all variables are statistically significant and important for predicting water temperature.
- The R-squared values of the model (0.9938) are very high, indicating that the model can explain a large proportion of the variance in water temperature.
- The F-statistic and p-value of the model also support that the model is statistically significant.

2. Forward Selection

- This modeling technique is used to find the explanatory variables that best fit a model. The process starts with an empty model in which no variables are included. At each step, the variable that provides the best improvement when added to the model is selected.
- The improvement is usually measured by a model selection criterion such as AIC (Akaike Information Criterion). Lower AIC values indicate better model fit.

```
min_model <- lm(data = bottle_new, T_degC ~ 1)


fw_step <- step(min_model, direction = "forward" ,
                scope= list(lower = min_model, upper = max_model))
```

```
## Start:  AIC=1902304
## T_degC ~ 1
##
##          Df Sum of Sq      RSS     AIC
## + STheta  1  10945452   830632  151582
## + O2ml_L  1   7440910  4335174 1242520
## + Depthm  1   5238906  6537178 1513711
## + Salnty  1   3002936  8773148 1707947
## <none>                11776084 1902304
##
## Step:  AIC=151582.4
## T_degC ~ STheta
##
##          Df Sum of Sq    RSS     AIC
## + Salnty  1    677445 153187 -964572
## + O2ml_L  1    222205 608427  -53955
## + Depthm  1     12954 817678  141207
## <none>                830632  151582
##
## Step:  AIC=-964572.4
## T_degC ~ STheta + Salnty
##
##          Df Sum of Sq    RSS      AIC
## + Depthm  1     72836  80351 -1390596
## + O2ml_L  1      2045 151142  -973445
## <none>                153187  -964572
##
## Step:  AIC=-1390596
## T_degC ~ STheta + Salnty + Depthm
##
##          Df Sum of Sq   RSS      AIC
## + O2ml_L  1    7525.4 72825 -1455520
## <none>                80351 -1390596
##
## Step:  AIC=-1455520
## T_degC ~ STheta + Salnty + Depthm + O2ml_L
```

```
summary(fw_step)
```

```
##
## Call:
## lm(formula = T_degC ~ STheta + Salnty + Depthm + O2ml_L, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9102 -0.1513  0.0198  0.1350  4.6759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289e+00  6.658e-02   139.5   <2e-16 ***
## STheta      -4.588e+00  9.796e-04 -4683.9   <2e-16 ***
## Salnty       3.542e+00  1.627e-03  2177.4   <2e-16 ***
## Depthm      -1.536e-03  1.822e-06  -842.6   <2e-16 ***
## O2ml_L       1.451e-01  5.556e-04   261.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 660237 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9938
## F-statistic: 2.653e+07 on 4 and 660237 DF,  p-value: < 2.2e-16
```

- The initial model is empty (contains only the intercept term). The AIC value is 1905589.
- The variable STheta is added because it reduces the AIC value the most (to 151251.8). Salnty variable is added, further reducing the AIC value (to -964885.2).
- The variable Depthm is added, reducing the AIC value again (to -1390056).
- Finally, the variable O2ml_L is added, bringing the AIC value to its lowest level (-1455444).

3. Both Direction

- Both Direction refers to an approach to stepwise regression in R that allows flexibility in adding and removing variables from the model. In this approach, we try to find the best model by both adding and removing variables from the model.

```
both_step <- step(max_model, direction = "both")
```

```
## Start:  AIC=-1455520
## T_degC ~ Salnty + O2ml_L + STheta + Depthm
##
##          Df Sum of Sq     RSS      AIC
## <none>                  72825 -1455520
## - O2ml_L  1      7525   80351 -1390596
## - Depthm  1     78316  151142  -973445
## - Salnty  1    522963  595788   -67813
## - STheta  1   2419903 2492728   877157
```

```
summary(both_step)
```

```
##
## Call:
## lm(formula = T_degC ~ Salnty + O2ml_L + STheta + Depthm, data = bottle_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9102 -0.1513  0.0198  0.1350  4.6759
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.289e+00  6.658e-02   139.5   <2e-16 ***
## Salnty       3.542e+00  1.627e-03  2177.4   <2e-16 ***
## O2ml_L       1.451e-01  5.556e-04   261.2   <2e-16 ***
## STheta      -4.588e+00  9.796e-04 -4683.9   <2e-16 ***
## Depthm      -1.536e-03  1.822e-06  -842.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 660237 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9938
## F-statistic: 2.653e+07 on 4 and 660237 DF,  p-value: < 2.2e-16
```

- both_step <- step(max_model, direction = “both”): This line starts the step-wise regression and finds the best model with direction “both”.
- Start: AIC = -1455444: The AIC (Akaike Information Criterion) value of the initial model is -1455444. AIC is a criterion used to assess model fit. Lower AIC values indicate better models.
- T_degC ~ Salnty + O2ml_L + STheta + Depthm: The initial model tries to explain the variable T_degC with the variables Salnty, O2ml_L, STheta and Depthm.
- In the model comparison table, the model with the lowest AIC value (-1455444) is the initial model. Therefore, step-wise regression did not make any changes in the model. As a result of the step-wise regression process, the initial model was found to be the best model. This model can explain T_degC variable with Salnty, O2ml_L, STheta and Depthm variables quite well (R-squared = 0.9938).

References

[1] “Bottle Database,” calcofi.org. https://calcofi.org/data/oceanographic-data/bottle-database/ (https://calcofi.org/data/oceanographic-data/bottle-database/)