Data Exploration

Introduction

This notebook presents an exploratory data analysis (EDA) of the air quality dataset. The analysis involves loading the data, summarizing it, and generating visualizations to understand its structure and characteristics.

Load Data

Load the Dataset Using the load_data Function

```
file_path <- '~/Downloads/AirQualityUCI.xlsx'
data <- load_data(file_path)
```

```
## # A tibble: 6 x 15
##   Date                Time                `CO(GT)` `PT08.S1(CO)` `NMHC(GT)`
##   <dttm>              <dttm>                 <dbl>         <dbl>      <dbl>
## 1 2004-03-10 00:00:00 1899-12-31 18:00:00      2.6          1360        150
## 2 2004-03-10 00:00:00 1899-12-31 19:00:00      2            1292.       112
## 3 2004-03-10 00:00:00 1899-12-31 20:00:00      2.2          1402         88
## 4 2004-03-10 00:00:00 1899-12-31 21:00:00      2.2          1376.        80
## 5 2004-03-10 00:00:00 1899-12-31 22:00:00      1.6          1272.        51
## 6 2004-03-10 00:00:00 1899-12-31 23:00:00      1.2          1197         38
## # i 10 more variables: `C6H6(GT)` <dbl>, `PT08.S2(NMHC)` <dbl>,
## #   `NOx(GT)` <dbl>, `PT08.S3(NOx)` <dbl>, `NO2(GT)` <dbl>,
## #   `PT08.S4(NO2)` <dbl>, `PT08.S5(O3)` <dbl>, T <dbl>, RH <dbl>, AH <dbl>
## [1] "2004-03-10 18:00:00 UTC" "2004-03-10 19:00:00 UTC"
## [3] "2004-03-10 20:00:00 UTC" "2004-03-10 21:00:00 UTC"
## [5] "2004-03-10 22:00:00 UTC" "2004-03-10 23:00:00 UTC"
```

```
head(data)
```

```
## # A tibble: 6 x 16
##   Date       Time   `CO(GT)` `PT08.S1(CO)` `NMHC(GT)` `C6H6(GT)` `PT08.S2(NMHC)`
##   <date>     <chr>     <dbl>         <dbl>      <dbl>      <dbl>           <dbl>
## 1 2004-03-10 18:00~      2.6          1360        150       11.9            1046.
## 2 2004-03-10 19:00~      2            1292.       112        9.40            955.
## 3 2004-03-10 20:00~      2.2          1402         88        9.00            939.
## 4 2004-03-10 21:00~      2.2          1376.        80        9.23            948.
## 5 2004-03-10 22:00~      1.6          1272.        51        6.52            836.
## 6 2004-03-10 23:00~      1.2          1197         38        4.74            750.
## # i 9 more variables: `NOx(GT)` <dbl>, `PT08.S3(NOx)` <dbl>, `NO2(GT)` <dbl>,
## #   `PT08.S4(NO2)` <dbl>, `PT08.S5(O3)` <dbl>, T <dbl>, RH <dbl>, AH <dbl>,
## #   datetime <dttm>
```

Summarize Data

Summarize the Data to Get an Overview of Its Structure and Statistics

```
summarize_data(data)
```

```
## Rows: 9,357
## Columns: 16
## $ Date            <date> 2004-03-10, 2004-03-10, 2004-03-10, 2004-03-10, 2004-~
## $ Time            <chr> "18:00:00", "19:00:00", "20:00:00", "21:00:00", "22:00~
## $ `CO(GT)`        <dbl> 2.6, 2.0, 2.2, 2.2, 1.6, 1.2, 1.2, 1.0, 0.9, 0.6, NA, ~
## $ `PT08.S1(CO)`   <dbl> 1360.00, 1292.25, 1402.00, 1375.50, 1272.25, 1197.00, ~
## $ `NMHC(GT)`      <dbl> 150, 112, 88, 80, 51, 38, 31, 31, 24, 19, 14, 8, 16, 2~
## $ `C6H6(GT)`      <dbl> 11.881723, 9.397165, 8.997817, 9.228796, 6.518224, 4.7~
## $ `PT08.S2(NMHC)` <dbl> 1045.50, 954.75, 939.25, 948.25, 835.50, 750.25, 689.5~
## $ `NOx(GT)`       <dbl> 166, 103, 131, 172, 131, 89, 62, 62, 45, NA, 21, 16, 3~
## $ `PT08.S3(NOx)`  <dbl> 1056.25, 1173.75, 1140.00, 1092.00, 1205.00, 1336.50, ~
## $ `NO2(GT)`       <dbl> 113, 92, 114, 122, 116, 96, 77, 76, 60, NA, 34, 28, 48~
## $ `PT08.S4(NO2)`  <dbl> 1692.00, 1558.75, 1554.50, 1583.75, 1490.00, 1393.00, ~
## $ `PT08.S5(O3)`   <dbl> 1267.50, 972.25, 1074.00, 1203.25, 1110.00, 949.25, 73~
## $ T               <dbl> 13.600, 13.300, 11.900, 11.000, 11.150, 11.175, 11.325~
## $ RH              <dbl> 48.875, 47.700, 53.975, 60.000, 59.575, 59.175, 56.775~
## $ AH              <dbl> 0.7577538, 0.7254874, 0.7502391, 0.7867125, 0.7887942,~
## $ datetime        <dttm> 2004-03-10 18:00:00, 2004-03-10 19:00:00, 2004-03-10 ~
## # A tibble: 9,357 x 16
##    Date       Time  `CO(GT)` `PT08.S1(CO)` `NMHC(GT)` `C6H6(GT)` `PT08.S2(NMHC)`
##    <date>     <chr>    <dbl>         <dbl>      <dbl>      <dbl>           <dbl>
##  1 2004-03-10 18:0~      2.6         1360         150      11.9            1046.
##  2 2004-03-10 19:0~      2           1292.        112       9.40            955.
##  3 2004-03-10 20:0~      2.2         1402          88       9.00            939.
##  4 2004-03-10 21:0~      2.2         1376.         80       9.23            948.
##  5 2004-03-10 22:0~      1.6         1272.         51       6.52            836.
##  6 2004-03-10 23:0~      1.2         1197          38       4.74            750.
##  7 2004-03-11 00:0~      1.2         1185          31       3.62            690.
##  8 2004-03-11 01:0~      1           1136.         31       3.33            672
##  9 2004-03-11 02:0~      0.9         1094          24       2.34            608.
## 10 2004-03-11 03:0~      0.6         1010.         19       1.70            561.
## # i 9,347 more rows
## # i 9 more variables: `NOx(GT)` <dbl>, `PT08.S3(NOx)` <dbl>, `NO2(GT)` <dbl>,
## #   `PT08.S4(NO2)` <dbl>, `PT08.S5(O3)` <dbl>, T <dbl>, RH <dbl>, AH <dbl>,
## #   datetime <dttm>
##       Date                Time              CO(GT)        PT08.S1(CO)
##  Min.   :2004-03-10   Length:9357        Min.   : 0.100   Min.   : 647.2
##  1st Qu.:2004-06-16   Class :character   1st Qu.: 1.100   1st Qu.: 936.8
##  Median :2004-09-21   Mode  :character   Median : 1.800   Median :1063.0
##  Mean   :2004-09-21                      Mean   : 2.153   Mean   :1099.7
##  3rd Qu.:2004-12-28                      3rd Qu.: 2.900   3rd Qu.:1231.2
##  Max.   :2005-04-04                      Max.   :11.900   Max.   :2039.8
##                                          NA's   :1683     NA's   :366
##     NMHC(GT)         C6H6(GT)      PT08.S2(NMHC)       NOx(GT)
##  Min.   :   7.0   Min.   : 0.149   Min.   : 383.2   Min.   :   2.0
##  1st Qu.:  67.0   1st Qu.: 4.437   1st Qu.: 734.4   1st Qu.:  98.0
##  Median : 150.0   Median : 8.240   Median : 909.0   Median : 179.8
```

```
##  Mean   : 218.8   Mean   :10.083   Mean   : 939.0   Mean   : 246.9
##  3rd Qu.: 297.0   3rd Qu.:13.989   3rd Qu.:1116.2   3rd Qu.: 326.0
##  Max.   :1189.0   Max.   :63.742   Max.   :2214.0   Max.   :1479.0
##  NA's   :8443     NA's   :366      NA's   :366      NA's   :1639
##   PT08.S3(NOx)       NO2(GT)       PT08.S4(NO2)   PT08.S5(O3)
##  Min.   : 322.0   Min.   :  2.0   Min.   : 551   Min.   : 221.0
##  1st Qu.: 657.9   1st Qu.: 78.0   1st Qu.:1227   1st Qu.: 731.4
##  Median : 805.5   Median :109.0   Median :1463   Median : 963.2
##  Mean   : 835.4   Mean   :113.1   Mean   :1456   Mean   :1022.8
##  3rd Qu.: 969.2   3rd Qu.:142.0   3rd Qu.:1674   3rd Qu.:1273.4
##  Max.   :2682.8   Max.   :339.7   Max.   :2775   Max.   :2522.8
##  NA's   :366      NA's   :1642    NA's   :366    NA's   :366
##         T               RH              AH
##  Min.   :-1.90   Min.   : 9.175   Min.   :0.1847
##  1st Qu.:11.79   1st Qu.:35.812   1st Qu.:0.7368
##  Median :17.75   Median :49.550   Median :0.9954
##  Mean   :18.32   Mean   :49.232   Mean   :1.0255
##  3rd Qu.:24.40   3rd Qu.:62.500   3rd Qu.:1.3137
##  Max.   :44.60   Max.   :88.725   Max.   :2.2310
##  NA's   :366     NA's   :366      NA's   :366
##     datetime
##  Min.   :2004-03-10 18:00:00
##  1st Qu.:2004-06-16 05:00:00
##  Median :2004-09-21 16:00:00
##  Mean   :2004-09-21 16:00:00
##  3rd Qu.:2004-12-28 03:00:00
##  Max.   :2005-04-04 14:00:00
##
```

Missing Values

Check for Missing Values in the Dataset

```
missing_values <- count_missing_values(data)
print(missing_values)
```

```
## # A tibble: 1 x 16
##    Date  Time `CO(GT)` `PT08.S1(CO)` `NMHC(GT)` `C6H6(GT)` `PT08.S2(NMHC)`
##   <int> <int>    <int>         <int>      <int>      <int>           <int>
## 1     0     0     1683           366       8443        366             366
## # i 9 more variables: `NOx(GT)` <int>, `PT08.S3(NOx)` <int>, `NO2(GT)` <int>,
## #   `PT08.S4(NO2)` <int>, `PT08.S5(O3)` <int>, T <int>, RH <int>, AH <int>,
## #   datetime <int>
```

Impute Missing Values

Impute the Missing Values Using Linear Interpolation

```
data <- impute_missing_values(data)
```

Summarize Data After Imputation

Summarize the Data Again After Imputing the Missing Values to See the Changes

```
summarize_data(data)
```

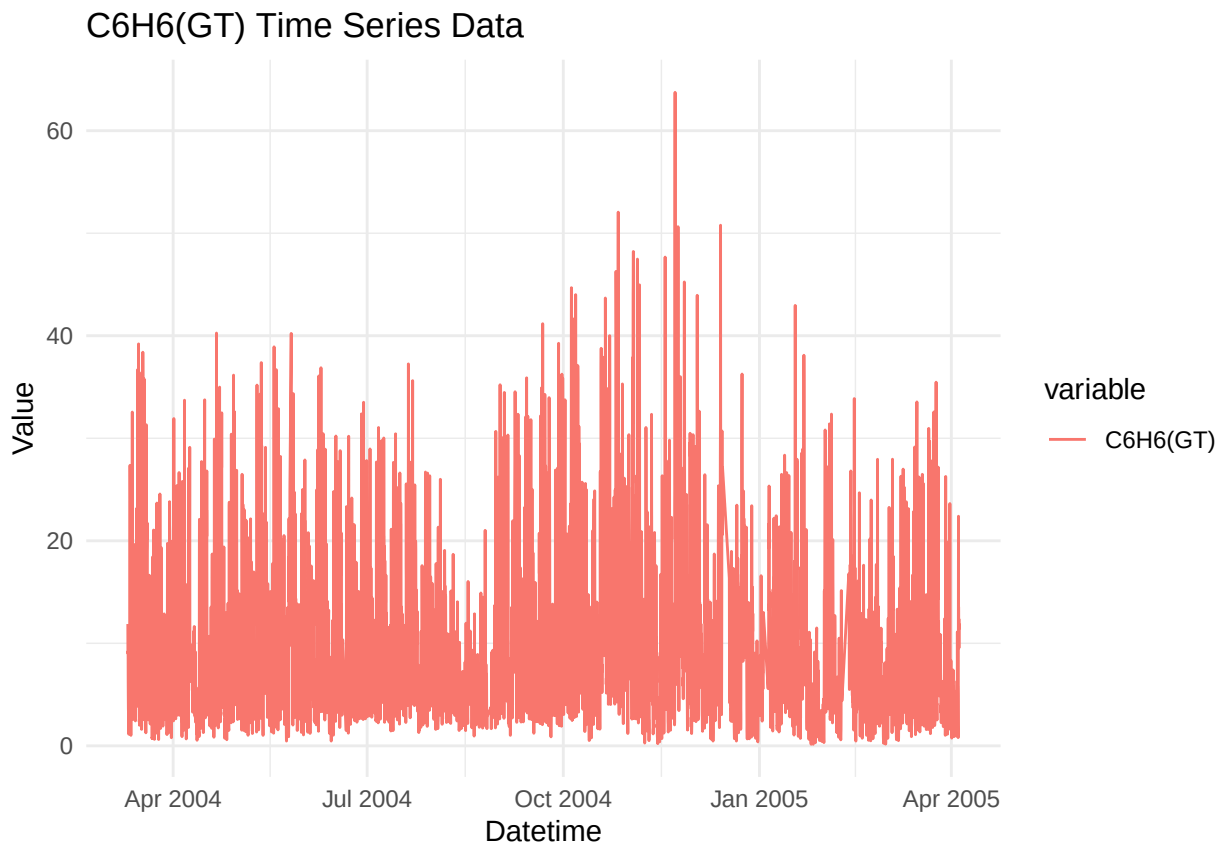
```
## Rows: 9,357
## Columns: 16
## $ Date            <date> 2004-03-10, 2004-03-10, 2004-03-10, 2004-03-10, 2004-~
## $ Time            <chr> "18:00:00", "19:00:00", "20:00:00", "21:00:00", "22:00~
## $ `CO(GT)`        <dbl> 2.60, 2.00, 2.20, 2.20, 1.60, 1.20, 1.20, 1.00, 0.90, ~
## $ `PT08.S1(CO)`   <dbl> 1360.00, 1292.25, 1402.00, 1375.50, 1272.25, 1197.00, ~
## $ `NMHC(GT)`      <dbl> 150, 112, 88, 80, 51, 38, 31, 31, 24, 19, 14, 8, 16, 2~
## $ `C6H6(GT)`      <dbl> 11.881723, 9.397165, 8.997817, 9.228796, 6.518224, 4.7~
## $ `PT08.S2(NMHC)` <dbl> 1045.50, 954.75, 939.25, 948.25, 835.50, 750.25, 689.5~
## $ `NOx(GT)`       <dbl> 166, 103, 131, 172, 131, 89, 62, 62, 45, 33, 21, 16, 3~
## $ `PT08.S3(NOx)`  <dbl> 1056.25, 1173.75, 1140.00, 1092.00, 1205.00, 1336.50, ~
## $ `NO2(GT)`       <dbl> 113, 92, 114, 122, 116, 96, 77, 76, 60, 47, 34, 28, 48~
## $ `PT08.S4(NO2)`  <dbl> 1692.00, 1558.75, 1554.50, 1583.75, 1490.00, 1393.00, ~
## $ `PT08.S5(O3)`   <dbl> 1267.50, 972.25, 1074.00, 1203.25, 1110.00, 949.25, 73~
## $ T               <dbl> 13.600, 13.300, 11.900, 11.000, 11.150, 11.175, 11.325~
## $ RH              <dbl> 48.875, 47.700, 53.975, 60.000, 59.575, 59.175, 56.775~
## $ AH              <dbl> 0.7577538, 0.7254874, 0.7502391, 0.7867125, 0.7887942,~
## $ datetime        <dttm> 2004-03-10 18:00:00, 2004-03-10 19:00:00, 2004-03-10 ~
## # A tibble: 9,357 x 16
##    Date       Time  `CO(GT)` `PT08.S1(CO)` `NMHC(GT)` `C6H6(GT)` `PT08.S2(NMHC)`
##    <date>     <chr>    <dbl>         <dbl>      <dbl>      <dbl>           <dbl>
##  1 2004-03-10 18:0~      2.6         1360         150      11.9            1046.
##  2 2004-03-10 19:0~      2           1292.        112       9.40            955.
##  3 2004-03-10 20:0~      2.2         1402          88       9.00            939.
##  4 2004-03-10 21:0~      2.2         1376.         80       9.23            948.
##  5 2004-03-10 22:0~      1.6         1272.         51       6.52            836.
##  6 2004-03-10 23:0~      1.2         1197          38       4.74            750.
##  7 2004-03-11 00:0~      1.2         1185          31       3.62            690.
##  8 2004-03-11 01:0~      1           1136.         31       3.33            672
##  9 2004-03-11 02:0~      0.9         1094          24       2.34            608.
## 10 2004-03-11 03:0~      0.6         1010.         19       1.70            561.
## # i 9,347 more rows
## # i 9 more variables: `NOx(GT)` <dbl>, `PT08.S3(NOx)` <dbl>, `NO2(GT)` <dbl>,
## #   `PT08.S4(NO2)` <dbl>, `PT08.S5(O3)` <dbl>, T <dbl>, RH <dbl>, AH <dbl>,
## #   datetime <dttm>
##       Date                Time              CO(GT)        PT08.S1(CO)
##  Min.   :2004-03-10   Length:9357        Min.   : 0.100   Min.   : 647.2
##  1st Qu.:2004-06-16   Class :character   1st Qu.: 1.100   1st Qu.: 937.5
##  Median :2004-09-21   Mode  :character   Median : 1.800   Median :1066.8
##  Mean   :2004-09-21                      Mean   : 2.131   Mean   :1102.9
##  3rd Qu.:2004-12-28                      3rd Qu.: 2.900   3rd Qu.:1238.8
##  Max.   :2005-04-04                      Max.   :11.900   Max.   :2039.8
##
##     NMHC(GT)          C6H6(GT)      PT08.S2(NMHC)       NOx(GT)
##  Min.   :   7.00   Min.   : 0.149   Min.   : 383.2   Min.   :   2.0
##  1st Qu.:  75.63   1st Qu.: 4.477   1st Qu.: 736.0   1st Qu.:  96.0
##  Median : 154.00   Median : 8.289   Median : 910.3   Median : 180.0
```

```
##  Mean   : 235.74   Mean   :10.179   Mean   : 942.0   Mean   : 241.9
##  3rd Qu.: 418.50   3rd Qu.:14.096   3rd Qu.:1119.0   3rd Qu.: 326.0
##  Max.   :1189.00   Max.   :63.741   Max.   :2214.0   Max.   :1479.0
##  NA's   :8126
##   PT08.S3(NOx)       NO2(GT)       PT08.S4(NO2)   PT08.S5(O3)
##  Min.   : 322.0   Min.   :  2.0   Min.   : 551   Min.   : 221.0
##  1st Qu.: 654.0   1st Qu.: 76.0   1st Qu.:1227   1st Qu.: 733.2
##  Median : 803.5   Median :104.8   Median :1460   Median : 970.0
##  Mean   : 832.6   Mean   :109.6   Mean   :1453   Mean   :1032.4
##  3rd Qu.: 967.5   3rd Qu.:136.4   3rd Qu.:1668   3rd Qu.:1293.0
##  Max.   :2682.8   Max.   :339.7   Max.   :2775   Max.   :2522.8
##
##        T               RH              AH
##  Min.   :-1.90   Min.   : 9.175   Min.   :0.1847
##  1st Qu.:11.72   1st Qu.:35.800   1st Qu.:0.7323
##  Median :17.57   Median :49.650   Median :0.9895
##  Mean   :18.23   Mean   :49.189   Mean   :1.0196
##  3rd Qu.:24.27   3rd Qu.:62.250   3rd Qu.:1.3067
##  Max.   :44.60   Max.   :88.725   Max.   :2.2310
##
##     datetime
##  Min.   :2004-03-10 18:00:00
##  1st Qu.:2004-06-16 05:00:00
##  Median :2004-09-21 16:00:00
##  Mean   :2004-09-21 16:00:00
##  3rd Qu.:2004-12-28 03:00:00
##  Max.   :2005-04-04 14:00:00
##
```

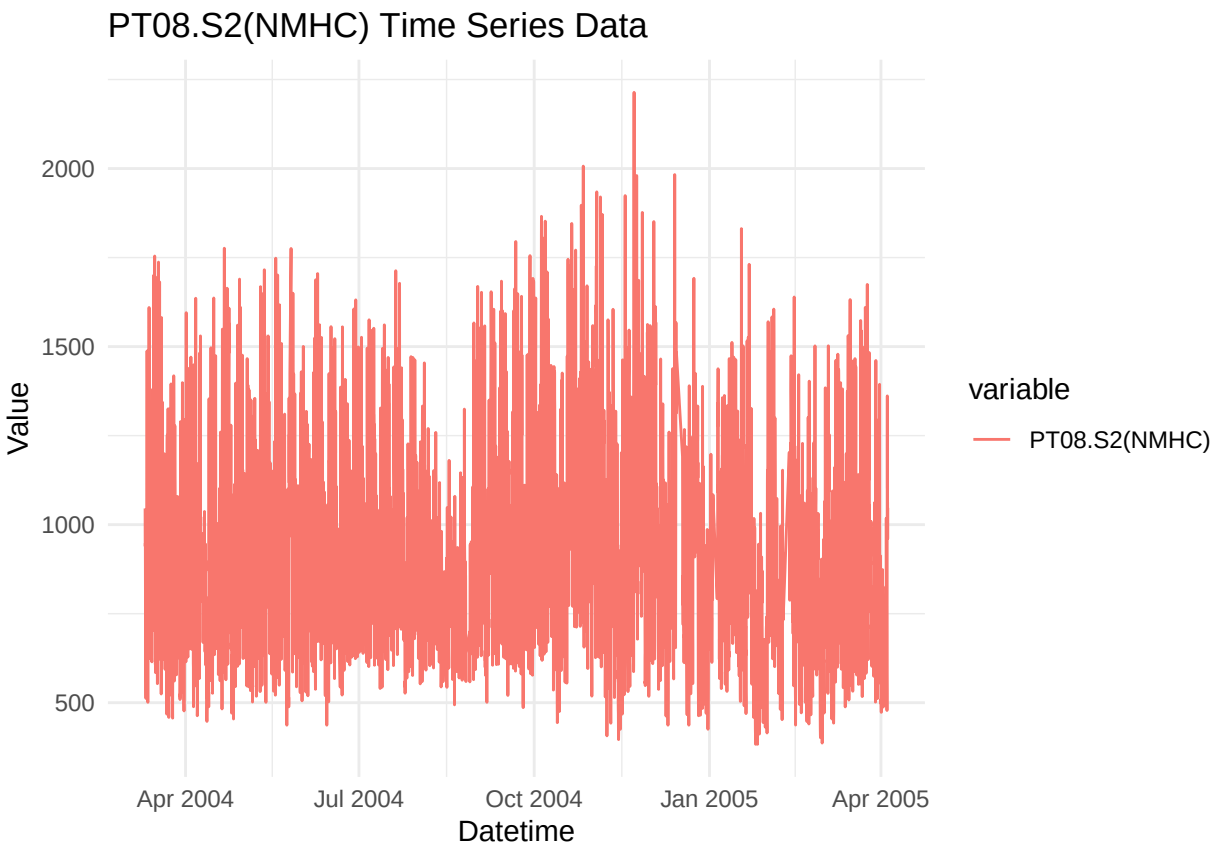
Time Series Plots

Plot the Time Series Data to Visualize the Trends and Patterns of Different Gas Concentrations and Environmental Factors

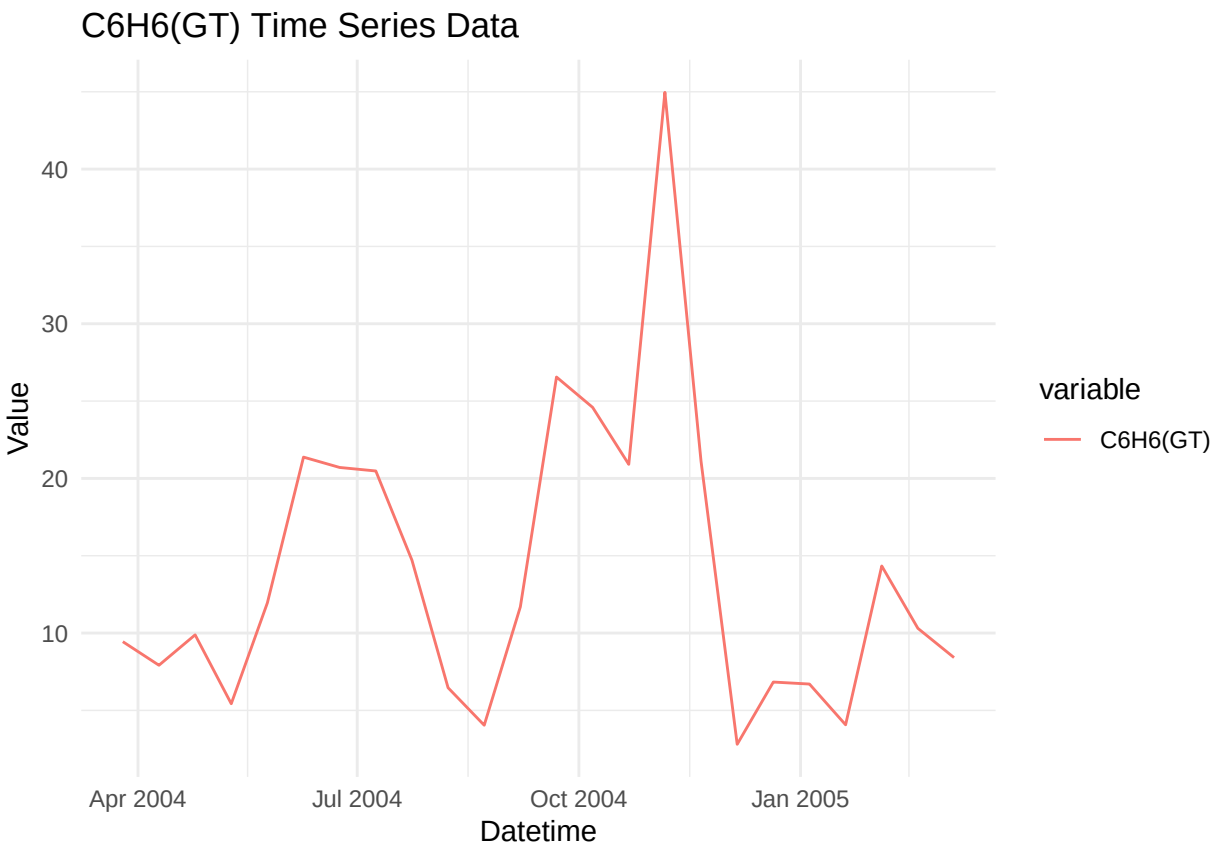
```
plot_time_series(data, 'C6H6(GT)', 'C6H6(GT) Time Series Data')
```




```
plot_time_series(data, 'PT08.S2(NMHC)', 'PT08.S2(NMHC) Time Series Data')
```



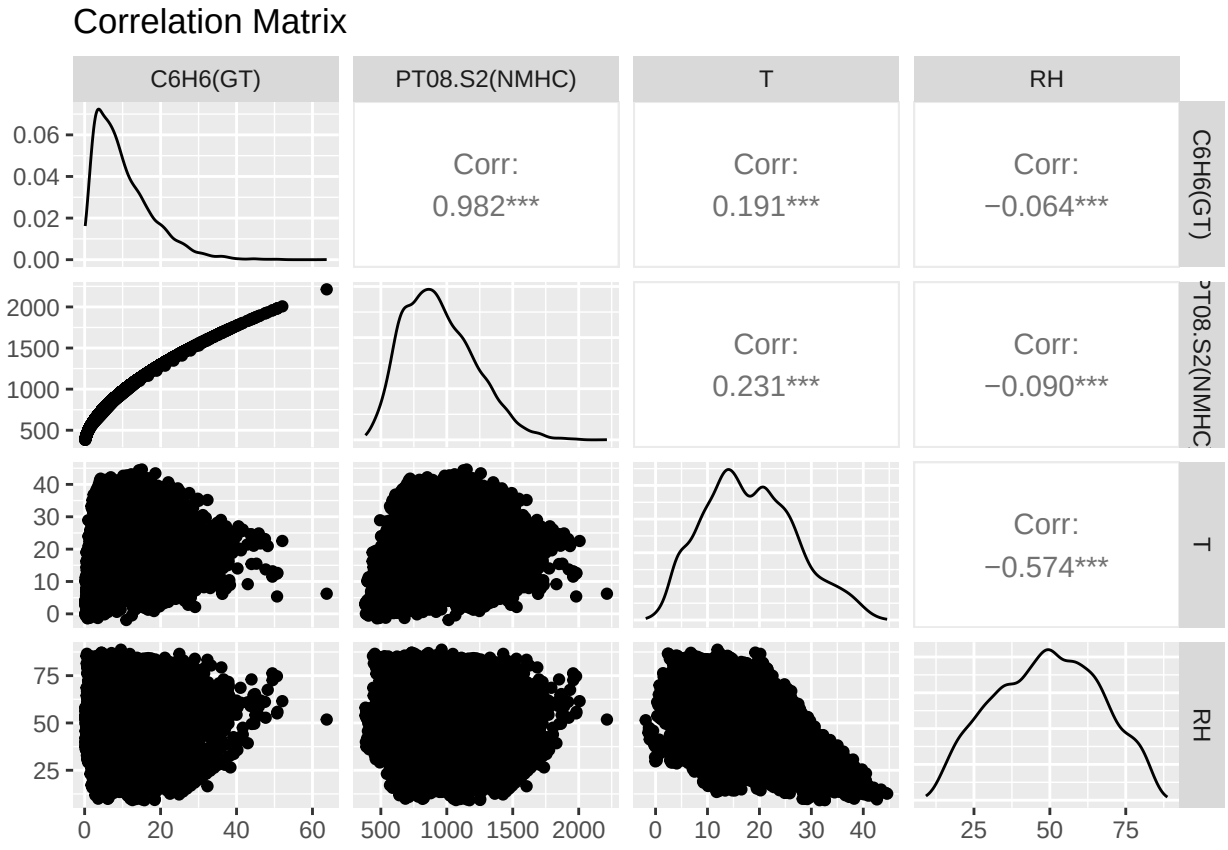

```
plot_time_series(data[(1:24*30*12),], 'C6H6(GT)', 'C6H6(GT) Time Series Data')
```



Correlation Analysis

Perform Correlation Analysis to Understand the Relationships Between Different Variables in the Dataset

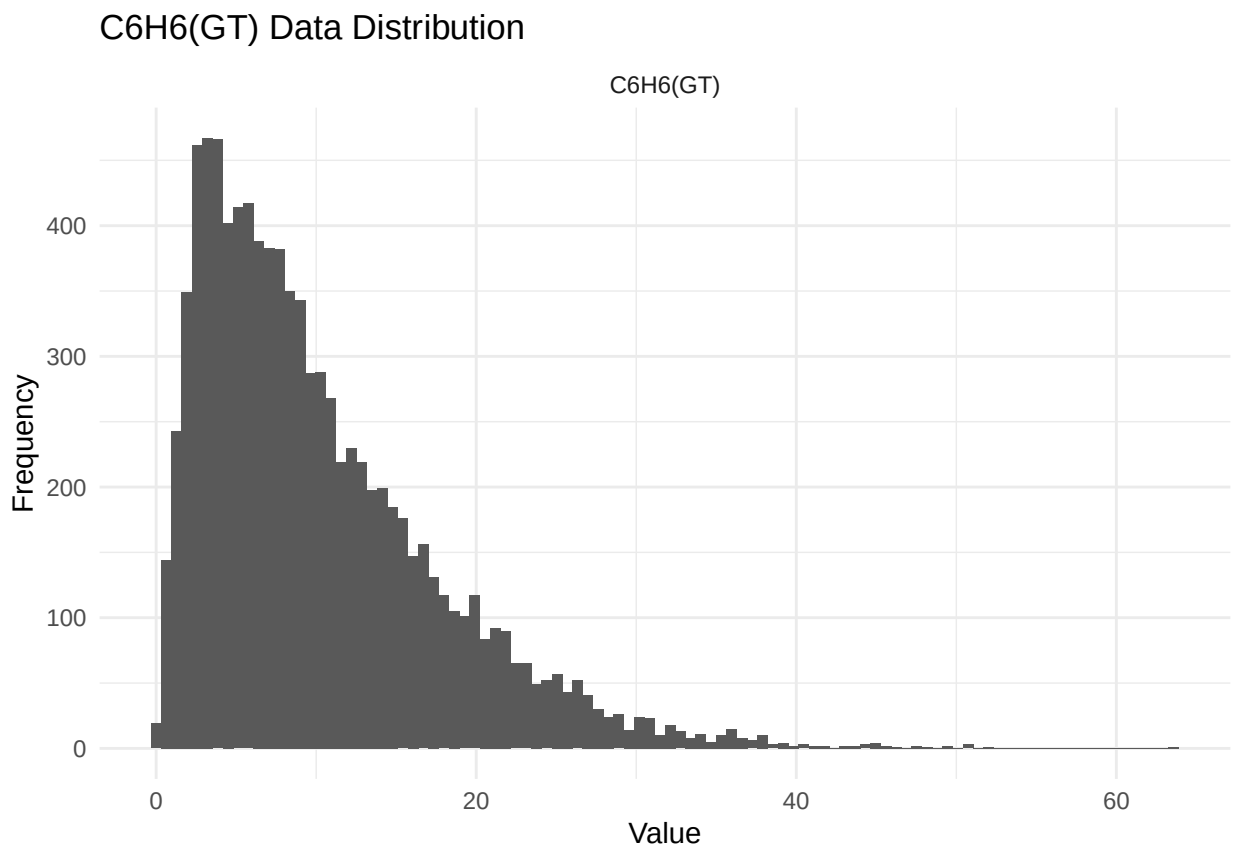
```
correlation_columns <- c('C6H6(GT)', 'PT08.S2(NMHC)', 'T', 'RH')
plot_correlations(data, correlation_columns)
```



Distribution Analysis

Visualize the Distribution of Each Variable to Understand Their Spread and Identify Any Potential Skewness or Abnormalities

```
data_distribution(data, 'C6H6(GT)', 'C6H6(GT) Data Distribution')
```



Aggregate to Daily Data (Downsampling)

Aggregate the Data to a Daily Level

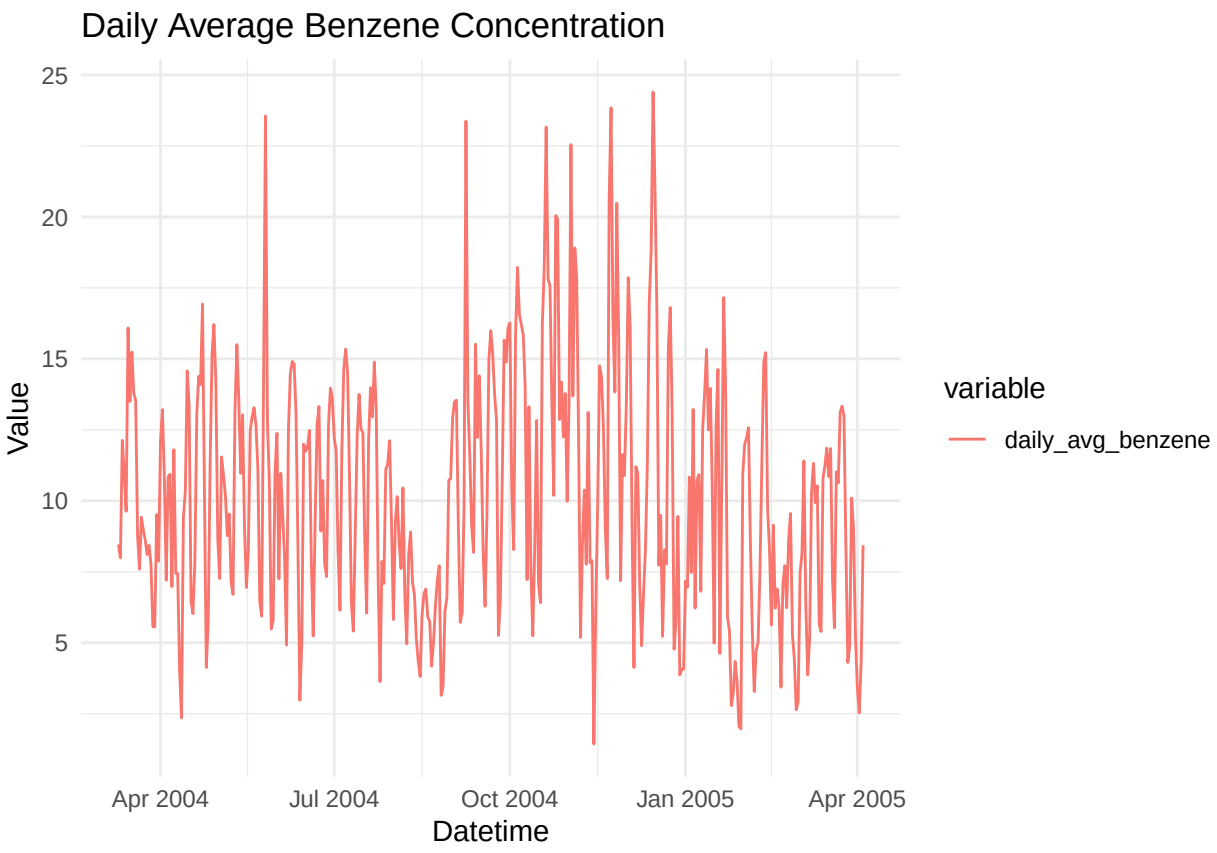
```
data_daily <- aggregate_to_daily(data)
head(data_daily)
```

```
## # A tibble: 6 x 3
##   date       daily_avg_benzene datetime
##   <date>                 <dbl> <dttm>
## 1 2004-03-10              8.46 2004-03-10 00:00:00
## 2 2004-03-11              7.99 2004-03-11 00:00:00
## 3 2004-03-12             12.1  2004-03-12 00:00:00
## 4 2004-03-13             10.9  2004-03-13 00:00:00
## 5 2004-03-14              9.63 2004-03-14 00:00:00
## 6 2004-03-15             16.1  2004-03-15 00:00:00
```

Time Series Plots

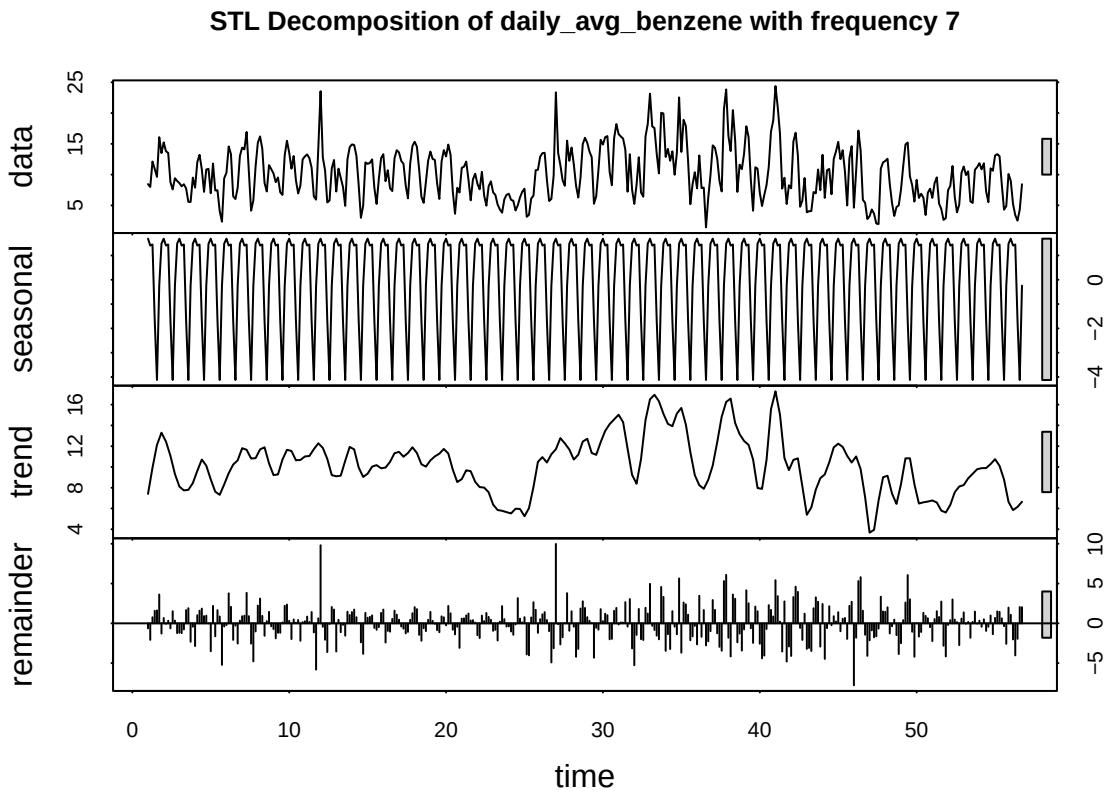
Plot the Time Series Data to Visualize the Trends and Patterns of Benzene Concentration and Other Environmental Factors

```
plot_time_series(data_daily, "daily_avg_benzene", "Daily Average Benzene Concentration")
```



Decomposition

```
plot_decomposition(data_daily, "daily_avg_benzene", freq = 7)
```

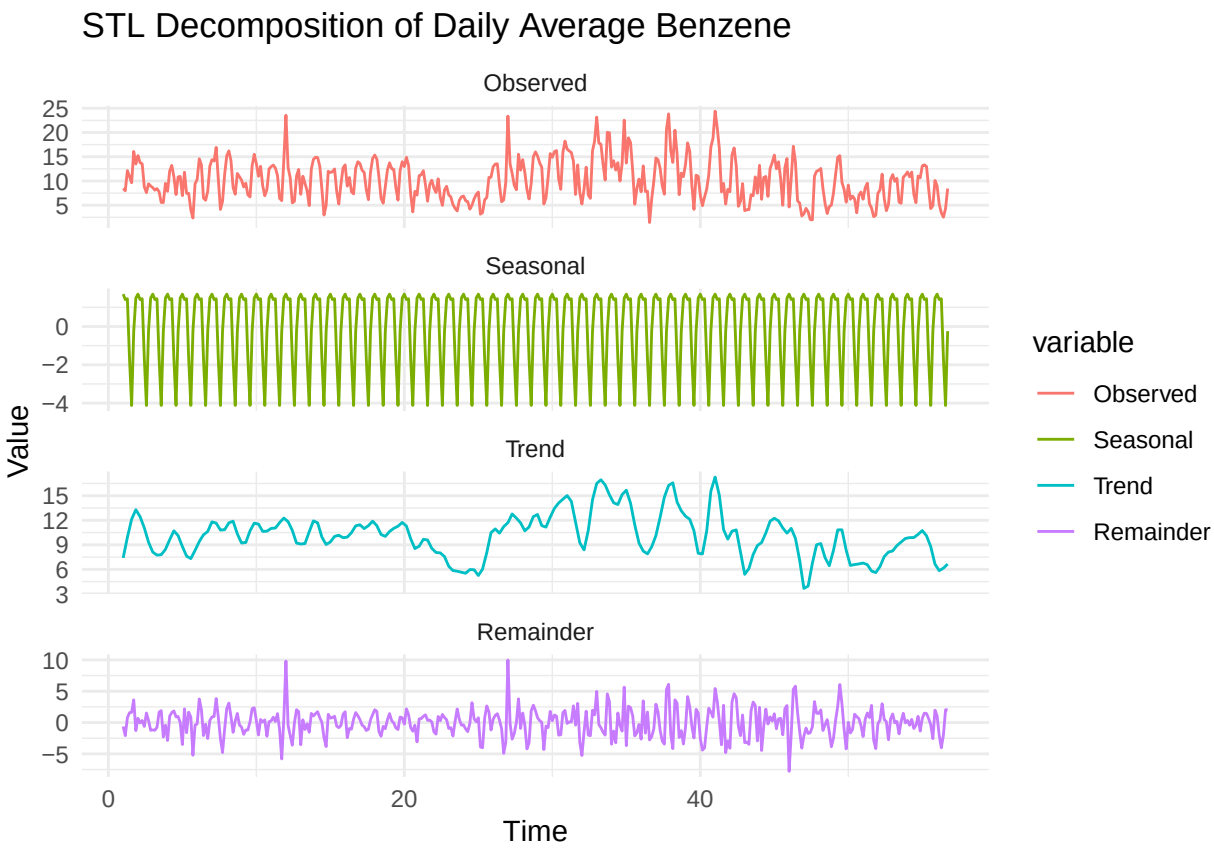


Decomposition Conclusions: * No Trend * Weekly Seasonality * Data Has a Random Noise

Decomposition - ALT Plot

```
plot_decomposition_ggplot(data_daily, "daily_avg_benzene", freq = 7)
```

```
## Don't know how to automatically pick scale for object of type <ts>. Defaulting
## to continuous.
```


STL Decomposition of Daily Average Benzene
Observed
Seasonal
Trend
Remainder
Value
Time
variable
Observed
Seasonal
Trend
Remainder