Data Pre-Processing

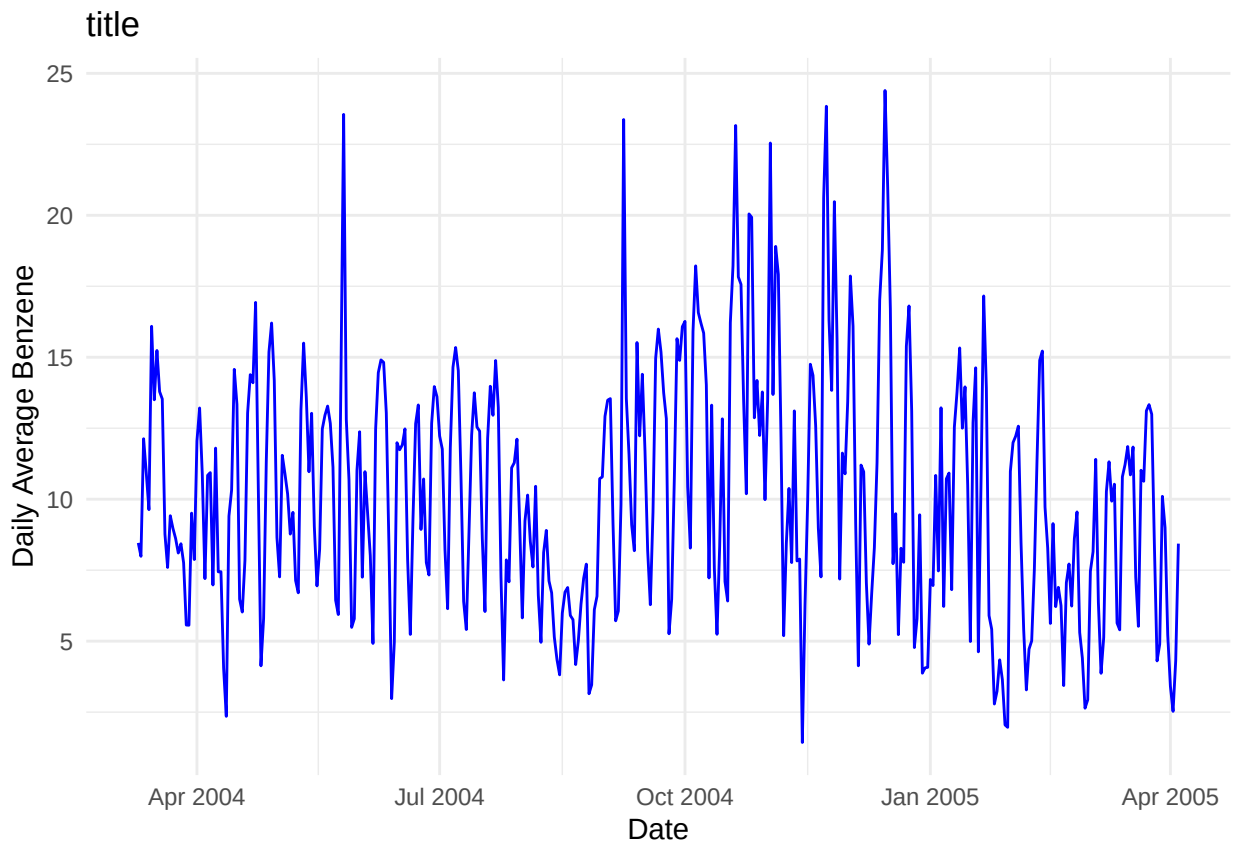
Introduction

This notebook outlines the data pre-processing steps for the air quality dataset, focusing on preparing the data for modeling. The steps include checking for stationarity, performing optional feature engineering, and splitting the data into training and testing sets.

Check Stationarity

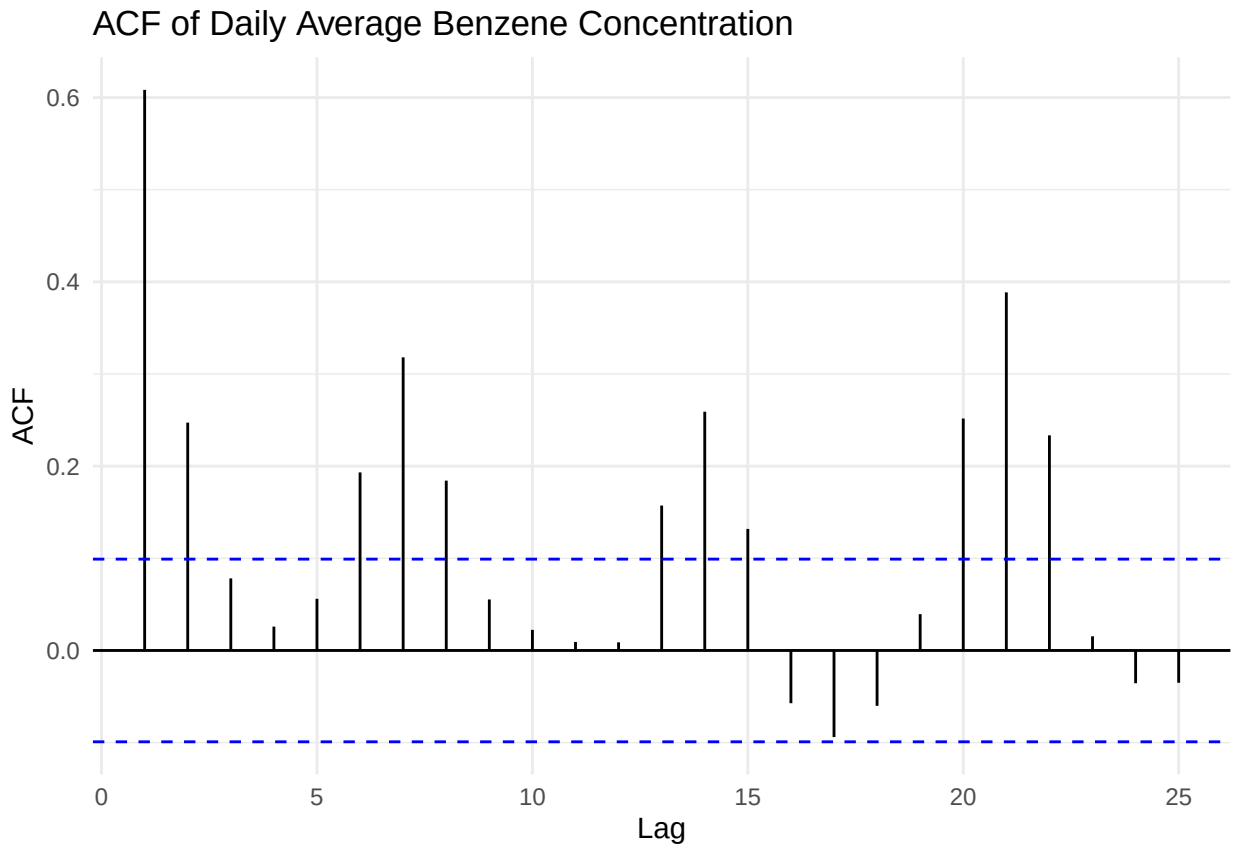
Check the Stationarity of the Benzene Concentration Time Series

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



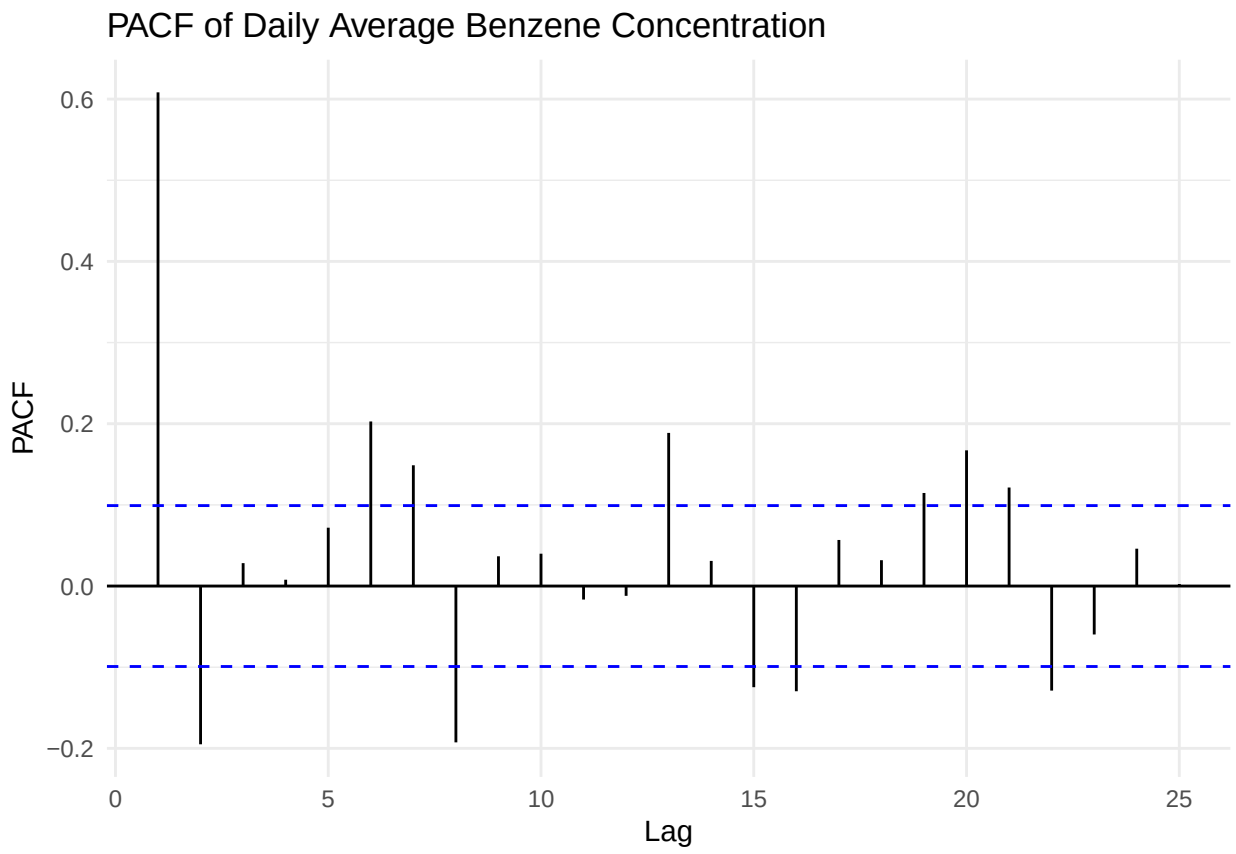

```
plot_acf(data_daily, "daily_avg_benzene", "Daily Average Benzene Concentration")
```

```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `main`
```

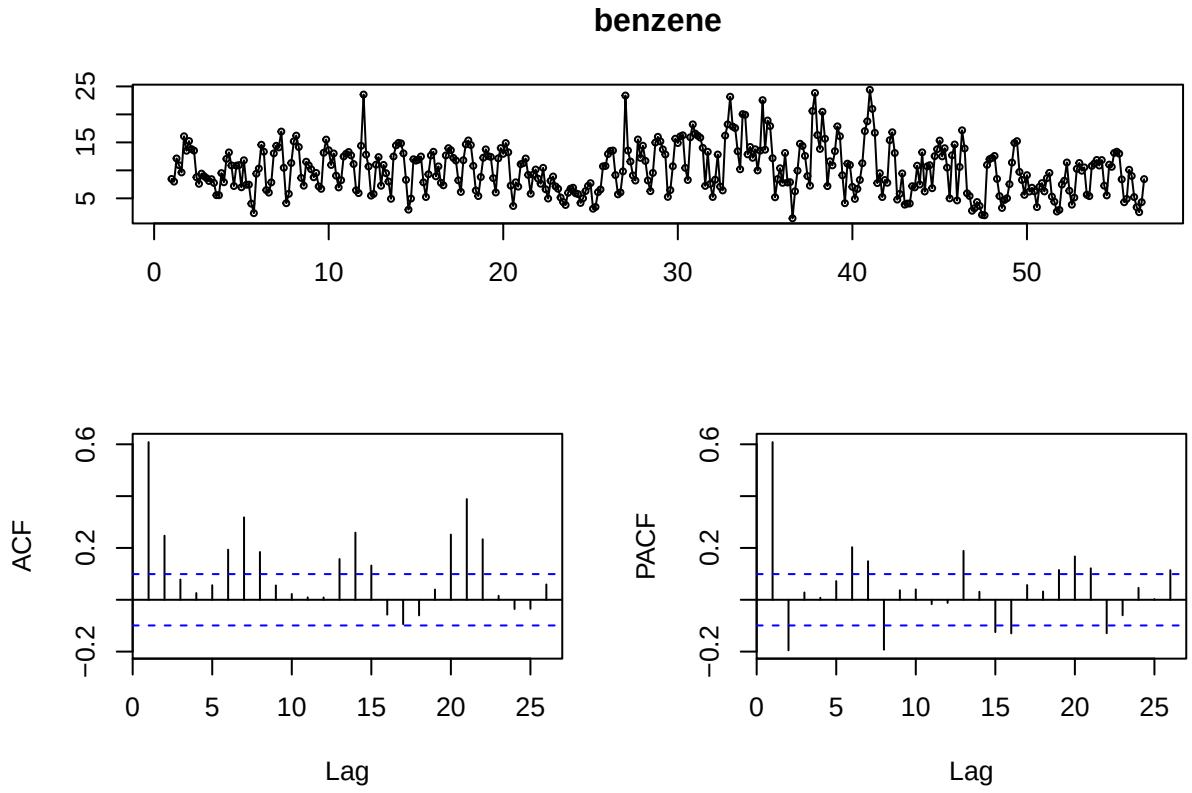



```
plot_pacf(data_daily, "daily_avg_benzene", "Daily Average Benzene Concentration")
```

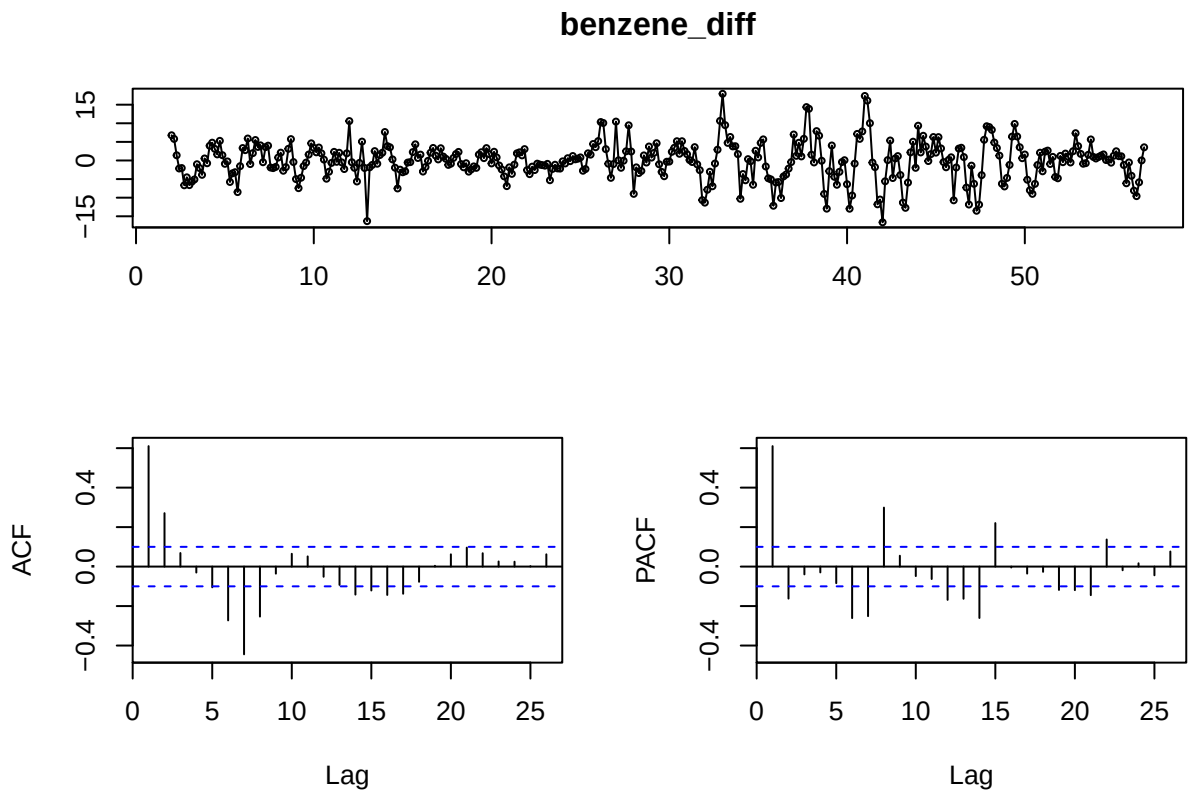
```
## Warning in ggplot2::geom_segment(lineend = "butt", ...): Ignoring unknown
## parameters: `main`
```



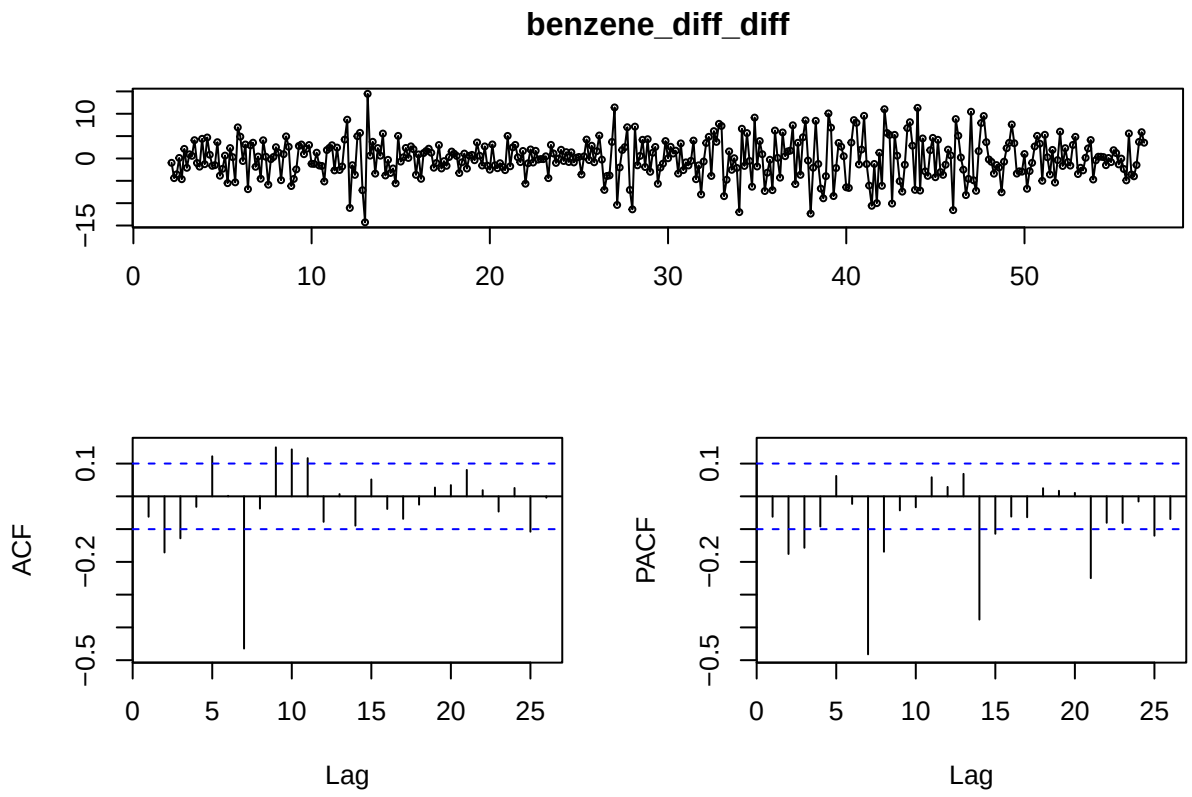

```
benzene <- ts(data_daily["daily_avg_benzene"], frequency = 7)
benzene_diff <- diff(benzene, lag = 7)
benzene_diff_diff <- diff(benzene_diff, lag = 1)
tsdisplay(benzene)
```




```
tsdisplay(benzene_diff)
```




```
tsdisplay(benzene_diff_diff)
```




```
hypothesis_tests_diff <- conduct_hypothesis_testing(as.data.frame(benzene_diff), "daily_avg_benzene")
hypothesis_tests_diff_diff <- conduct_hypothesis_testing(as.data.frame(benzene_diff_diff), "daily_avg_b

print(hypothesis_tests_diff$adf_test) # Stationary
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  df[[column]]
## Dickey-Fuller = -7.3043, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
print(hypothesis_tests_diff$kpss_test) # Stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  df[[column]]
## KPSS Level = 0.02191, Truncation lag parameter = 5, p-value = 0.1
```

```
print(hypothesis_tests_diff_diff$adf_test) # Stationary
```

```
##
##  Augmented Dickey-Fuller Test
```

```
##
## data:  df[[column]]
## Dickey-Fuller = -13.619, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
print(hypothesis_tests_diff_diff$kpss_test) # Stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  df[[column]]
## KPSS Level = 0.012833, Truncation lag parameter = 5, p-value = 0.1
```

Check the Stationarity of the Benzene Concentration Time Series

```
## Warning in adf.test(df[[column]], alternative = "stationary"): p-value smaller
## than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  df[[column]]
## Dickey-Fuller = -5.1861, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  df[[column]]
## KPSS Level = 0.40337, Truncation lag parameter = 5, p-value = 0.0757
```

Feature Engineering

Create Interaction Feature Between Temperature and Humidity

```
#data <- feature_engineering(data)
```

Split Data into Training and Testing Sets

Split the Data Into Training and Testing Sets Based on a Specified Datetime

```
split_datetime <- "2005-03-10 00:00:00"  # Specify Datetime for the Split
split <- split_data(data_daily, split_datetime)
train_data <- split$train
test_data <- split$test
head(train_data)
```

```
## # 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
```

```
head(test_data)
```

```
## # A tibble: 6 x 3
##   date       daily_avg_benzene datetime
##   <date>                 <dbl> <dttm>
## 1 2005-03-11             10.5  2005-03-11 00:00:00
## 2 2005-03-12              5.64 2005-03-12 00:00:00
## 3 2005-03-13              5.40 2005-03-13 00:00:00
## 4 2005-03-14             10.8  2005-03-14 00:00:00
## 5 2005-03-15             11.2  2005-03-15 00:00:00
## 6 2005-03-16             11.9  2005-03-16 00:00:00
```

Save Clean Dataset

Save the Clean Dataset to for Modeling

```
ts_data <- ts(data_daily["daily_avg_benzene"], frequency = 7)
ts_train_data <- ts(train_data["daily_avg_benzene"], frequency = 7)
ts_test_data <- ts(test_data["daily_avg_benzene"],start = c(53,3) , frequency = 7)
ts_train_data_s <- diff(diff(ts(train_data["daily_avg_benzene"], frequency = 7), lag = 7), lag = 1)
ts_test_data_s <- diff(diff(ts(test_data["daily_avg_benzene"],start(end(ts_train_data_s)), frequency = 
```

```
save(ts_train_data, file = "~/Downloads/ts_train_data")
save(ts_test_data, file = "~/Downloads/ts_test_data")
```