STA2453 EDA

Yufei Liu

2025-02-09

```r
dataset_summary <- function(df, name) {
  cat("\nDataset:", name, "\n")
  print(str(df))
  print(summary(df))
  print(colSums(is.na(df)))  # Missing values count
}

dataset_summary(data013_flux, "TIC 0131799991")
```

```
##
## Dataset: TIC 0131799991
## 'data.frame':    13372 obs. of  2 variables:
##  $ time       : num  1517 1517 1517 1517 1517 ...
##  $ pdcsap_flux: num  NA NA NA NA NA NA NA NA NA NA ...
## NULL
##       time       pdcsap_flux
##  Min.   :1517   Min.   :2449
##  1st Qu.:1522   1st Qu.:2484
##  Median :1527   Median :2492
##  Mean   :1529   Mean   :2493
##  3rd Qu.:1537   3rd Qu.:2501
##  Max.   :1542   Max.   :3058
##                 NA's   :338
##        time pdcsap_flux
##           0         338
```

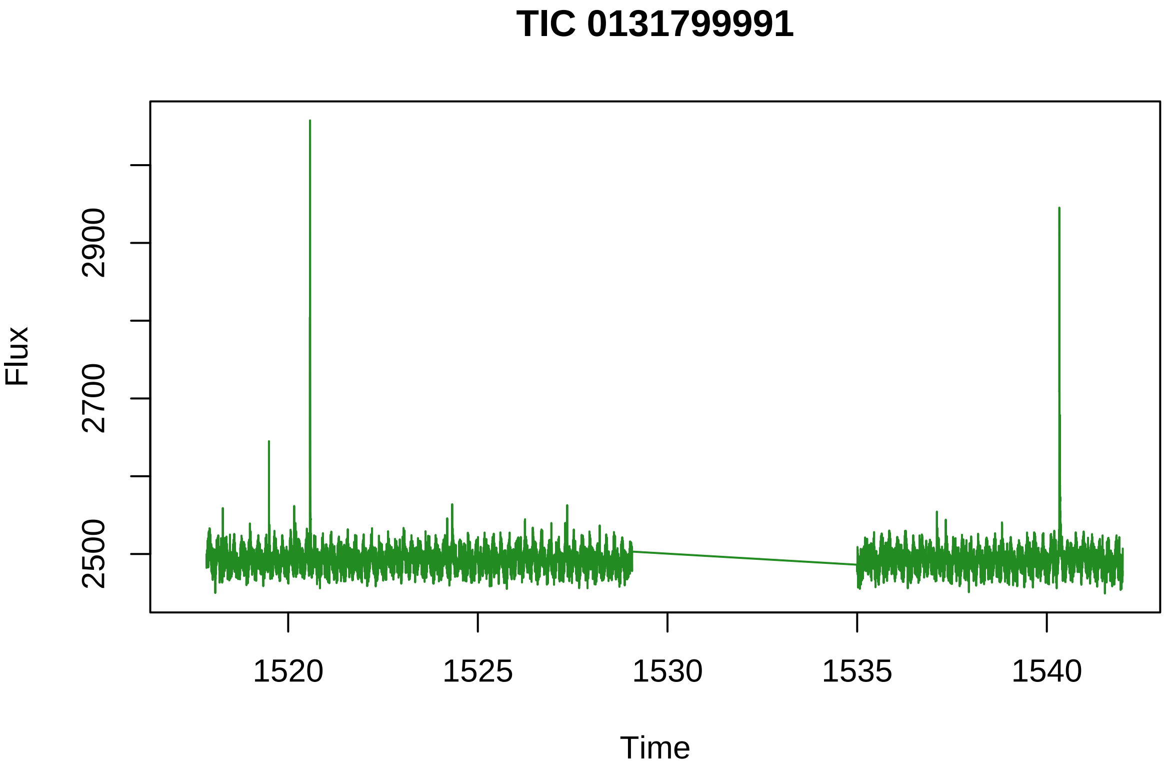
```r
dataset_summary(data129_flux, "TIC 129646813")
```

```
##
## Dataset: TIC 129646813
## 'data.frame':    18279 obs. of  2 variables:
##  $ time       : num  1325 1325 1325 1325 1325 ...
##  $ pdcsap_flux: num  808 819 816 817 817 ...
## NULL
##       time       pdcsap_flux
##  Min.   :1325   Min.   :792.0
##  1st Qu.:1332   1st Qu.:813.0
##  Median :1338   Median :817.3
##  Mean   :1339   Mean   :817.5
##  3rd Qu.:1346   3rd Qu.:821.7
##  Max.   :1353   Max.   :975.8
##                 NA's   :91
##        time pdcsap_flux
##           0          91
```

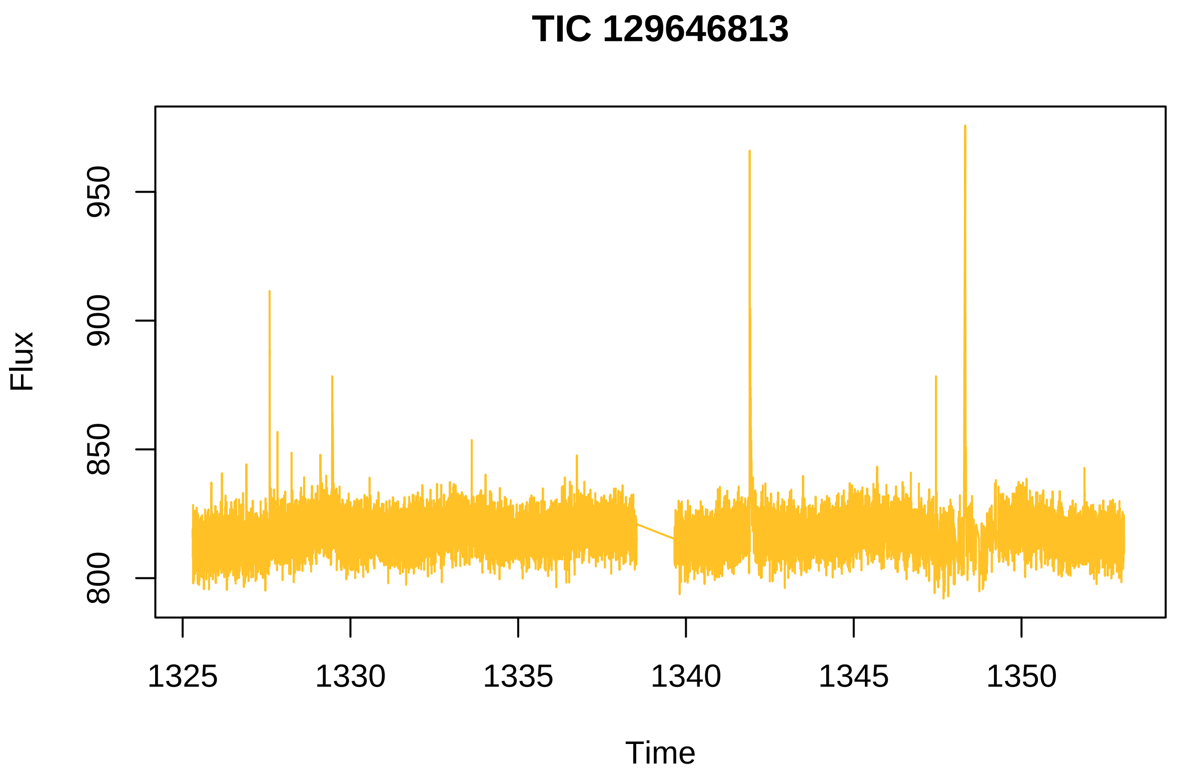
```
dataset_summary(data031_flux, "TIC 031381302")
```

```
##
## Dataset: TIC 031381302
## 'data.frame':    17719 obs. of  2 variables:
##  $ time       : num  1438 1438 1438 1438 1438 ...
##  $ pdcsap_flux: num  NA NA NA NA NA NA NA NA NA NA ...
## NULL
##       time       pdcsap_flux
##  Min.   :1438   Min.   :1531
##  1st Qu.:1444   1st Qu.:1558
##  Median :1452   Median :1564
##  Mean   :1451   Mean   :1564
##  3rd Qu.:1458   3rd Qu.:1571
##  Max.   :1464   Max.   :1679
##                 NA's   :686
##        time pdcsap_flux
##           0         686
```

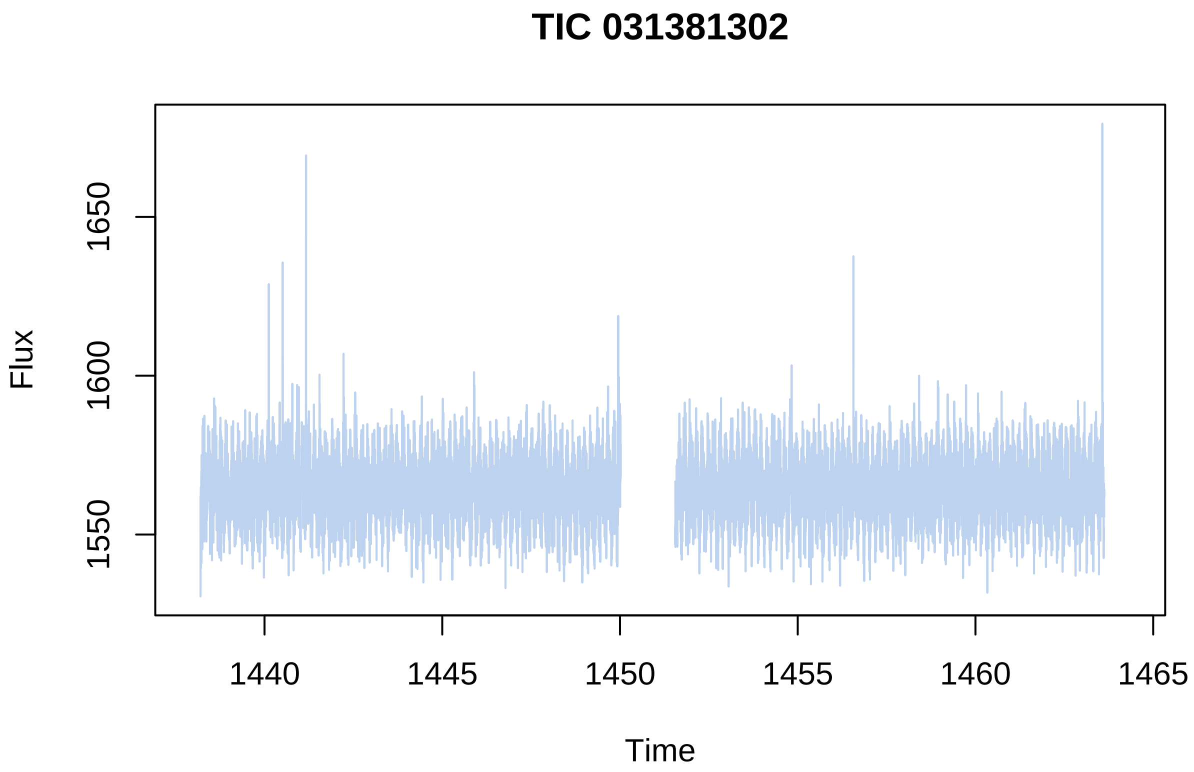
```
plot(data013_flux$time, data013_flux$pdcsap_flux, type = "l", col = "forestgreen",
     xlab = "Time", ylab = "Flux", main = "TIC 0131799991")
```



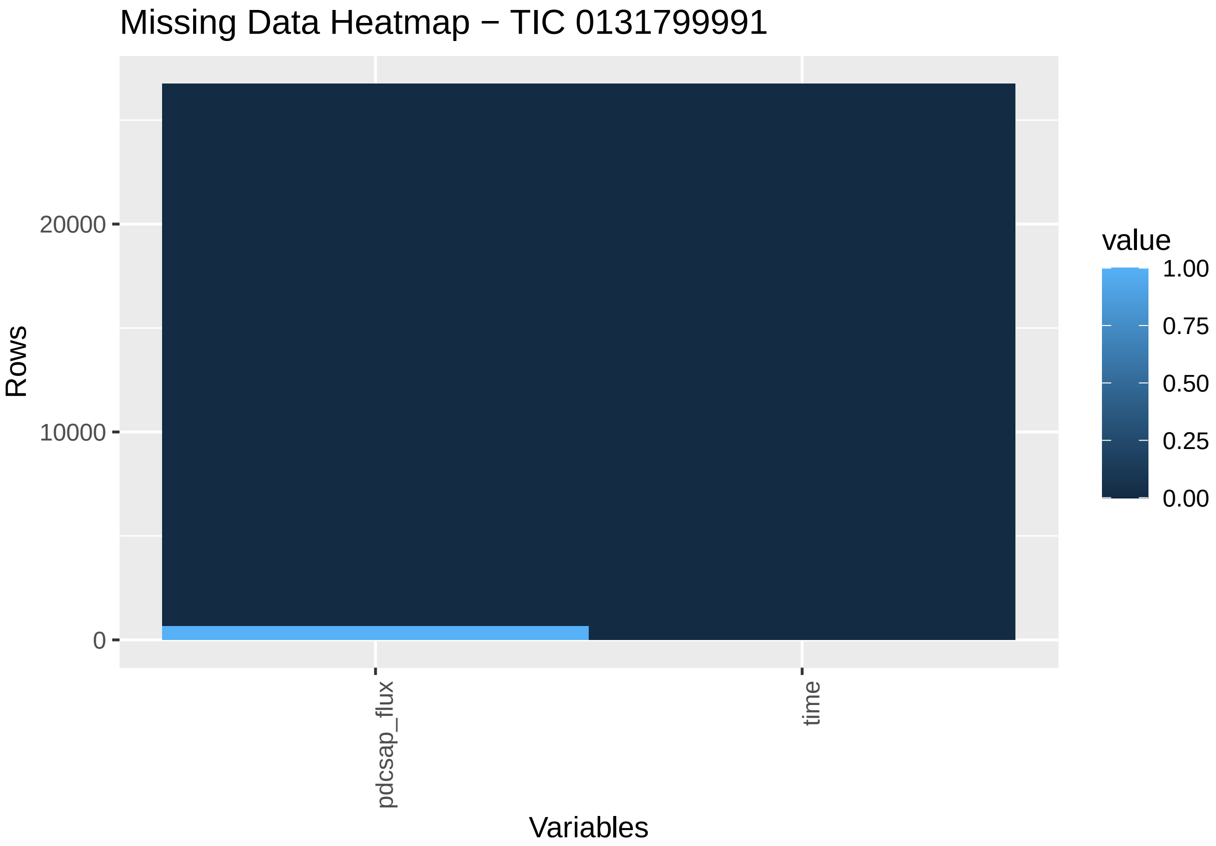
```
plot(data129_flux$time, data129_flux$pdcsap_flux, type = "l", col = "goldenrod1",
     xlab = "Time", ylab = "Flux", main = "TIC 129646813")
```



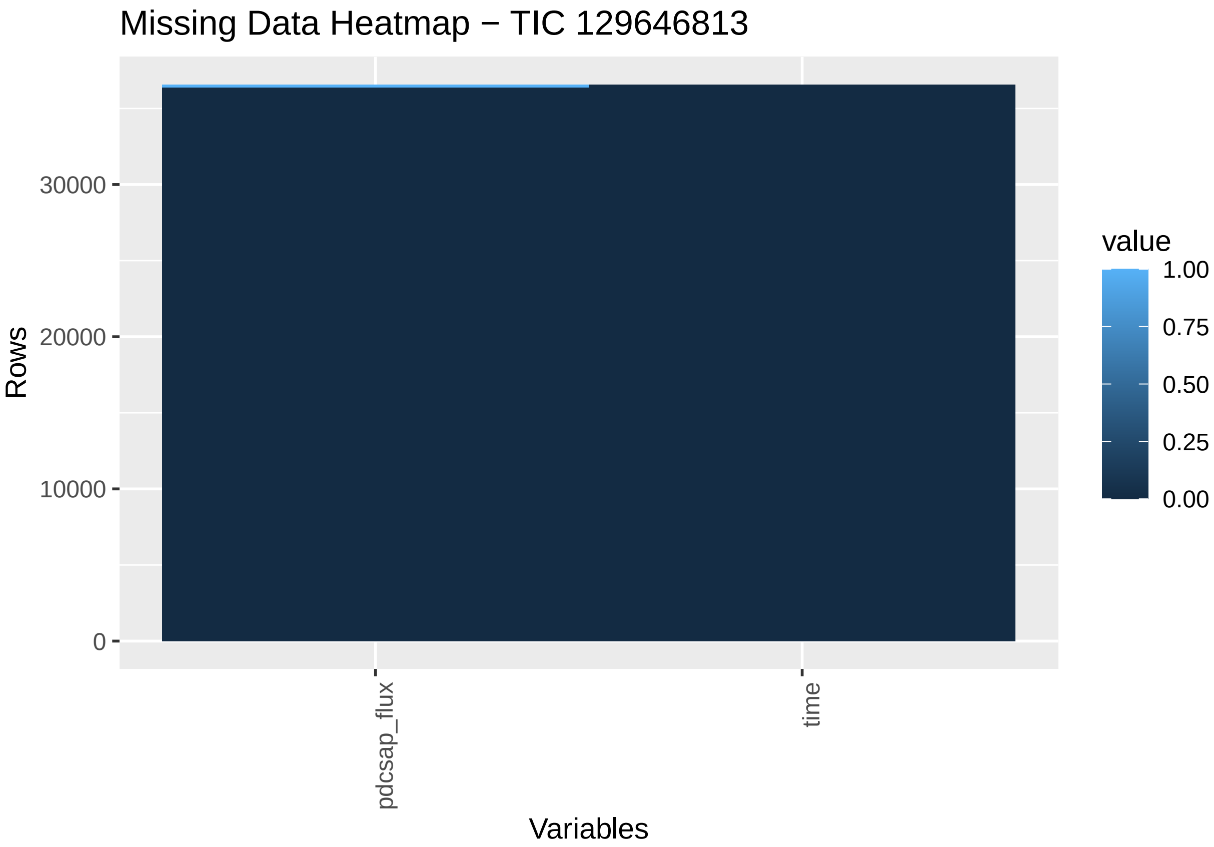
```r
plot(data031_flux$time, data031_flux$pdcsap_flux, type = "l", col = "lightsteelblue2",
     xlab = "Time", ylab = "Flux", main = "TIC 031381302")
```



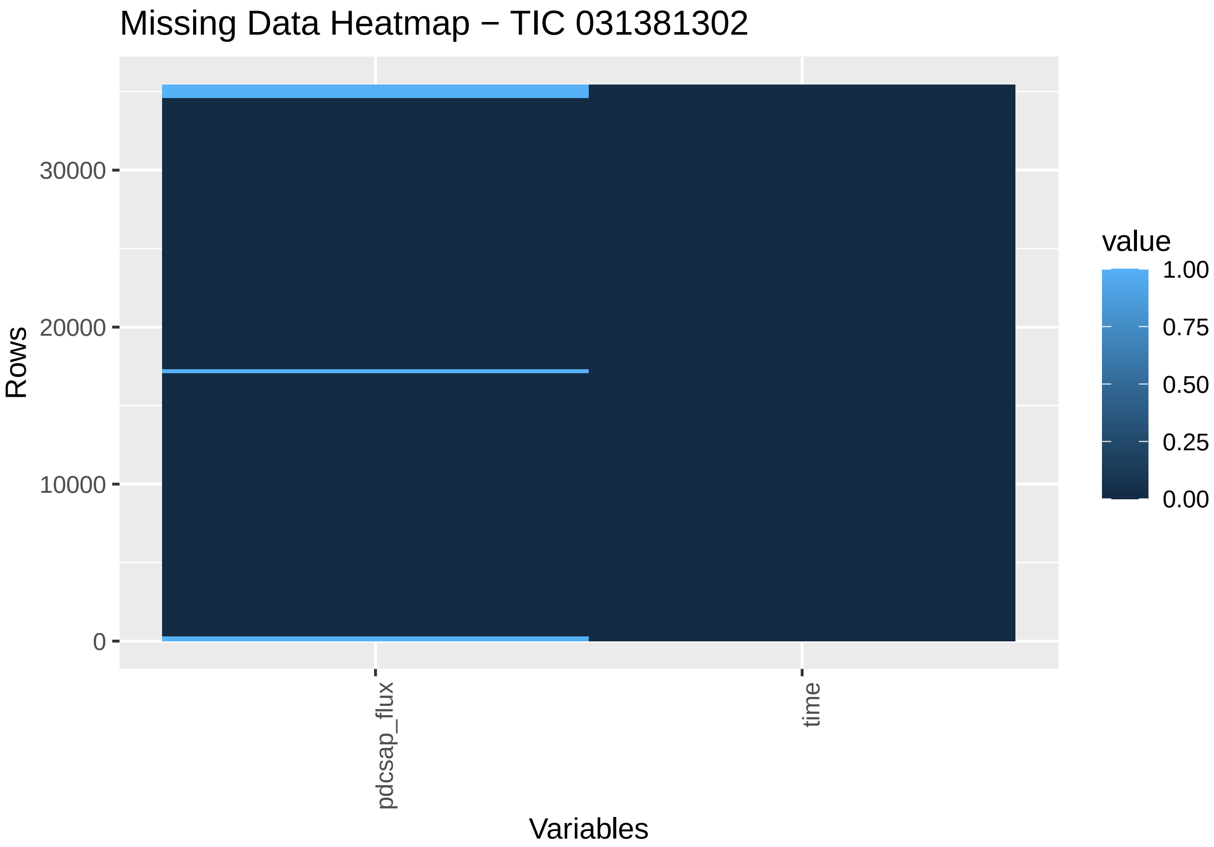
```r
missing_data_plots <- function(df, name) {
  # Heatmap of missing values
  missing_df <- df %>% mutate_all(~ifelse(is.na(.), 1, 0)) %>% pivot_longer(everything())
  ggplot(missing_df, aes(x=name, y=as.numeric(row.names(missing_df)), fill=value)) +
    geom_tile() +
    labs(title=paste("Missing Data Heatmap -", name), x="Variables", y="Rows") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
}

missing_data_plots(data013_flux, "TIC 0131799991")
```



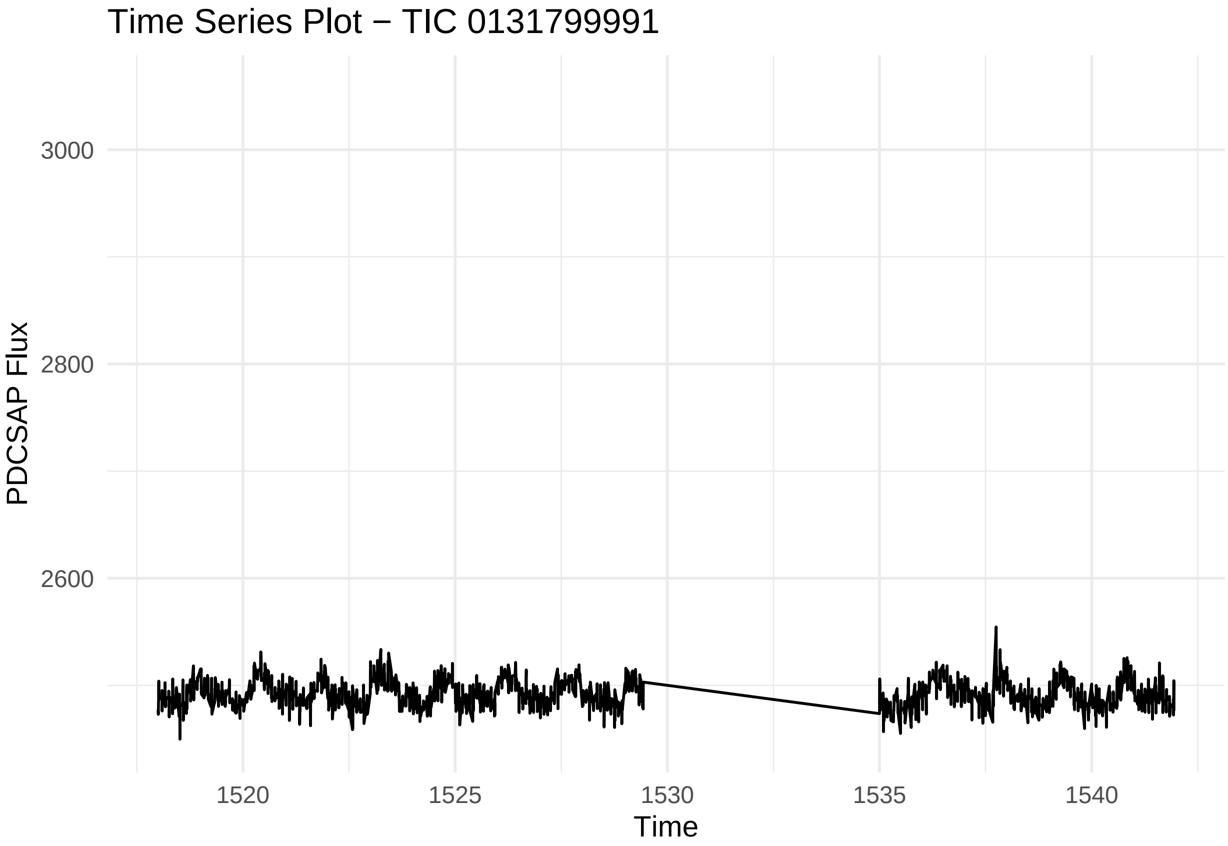
```
missing_data_plots(data129_flux, "TIC 129646813")
```

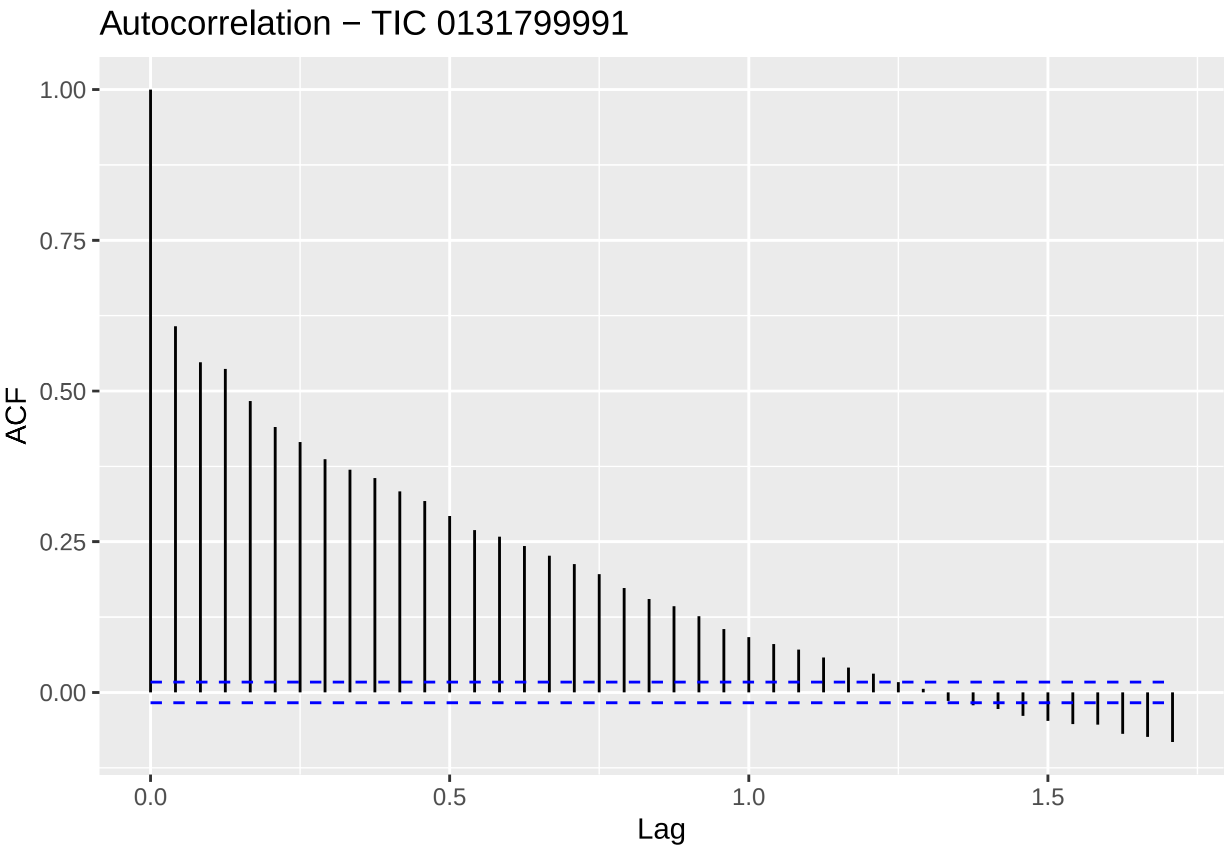


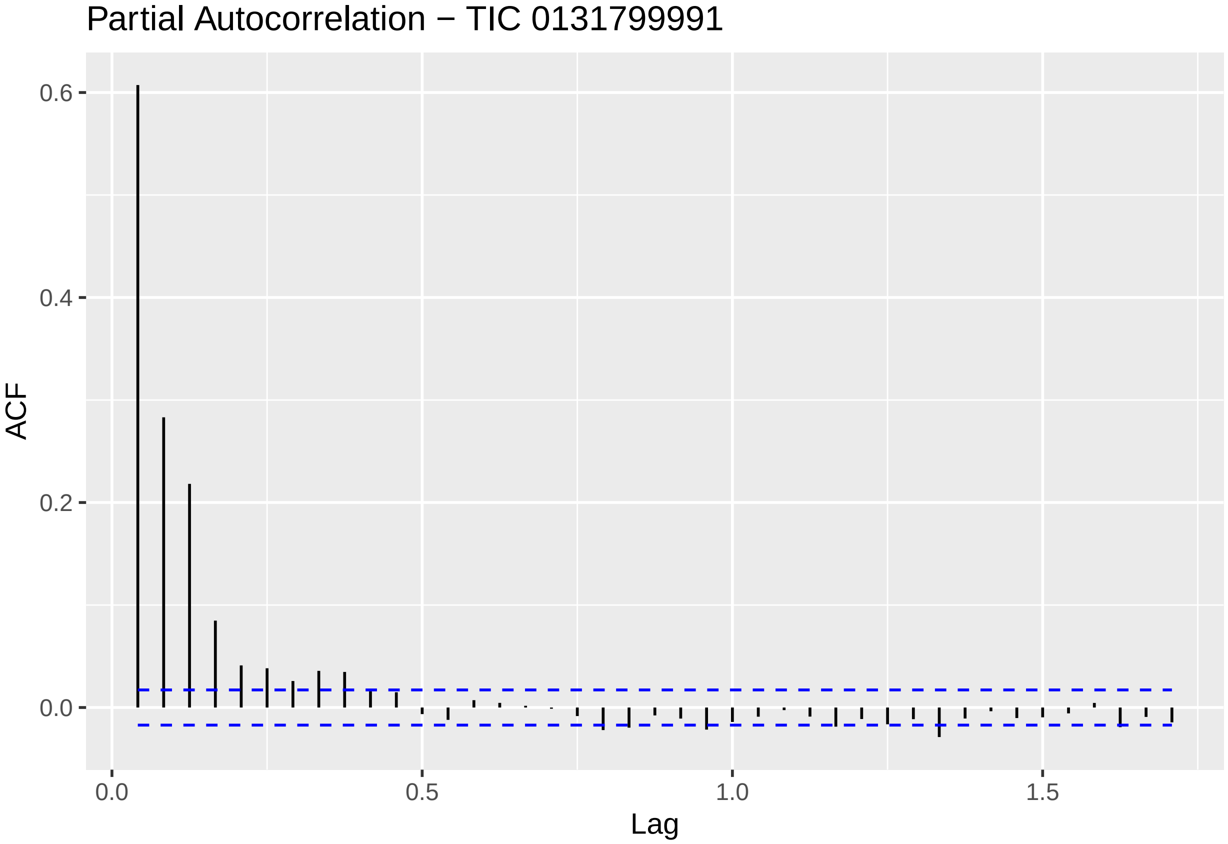
```
missing_data_plots(data031_flux, "TIC 031381302")
```

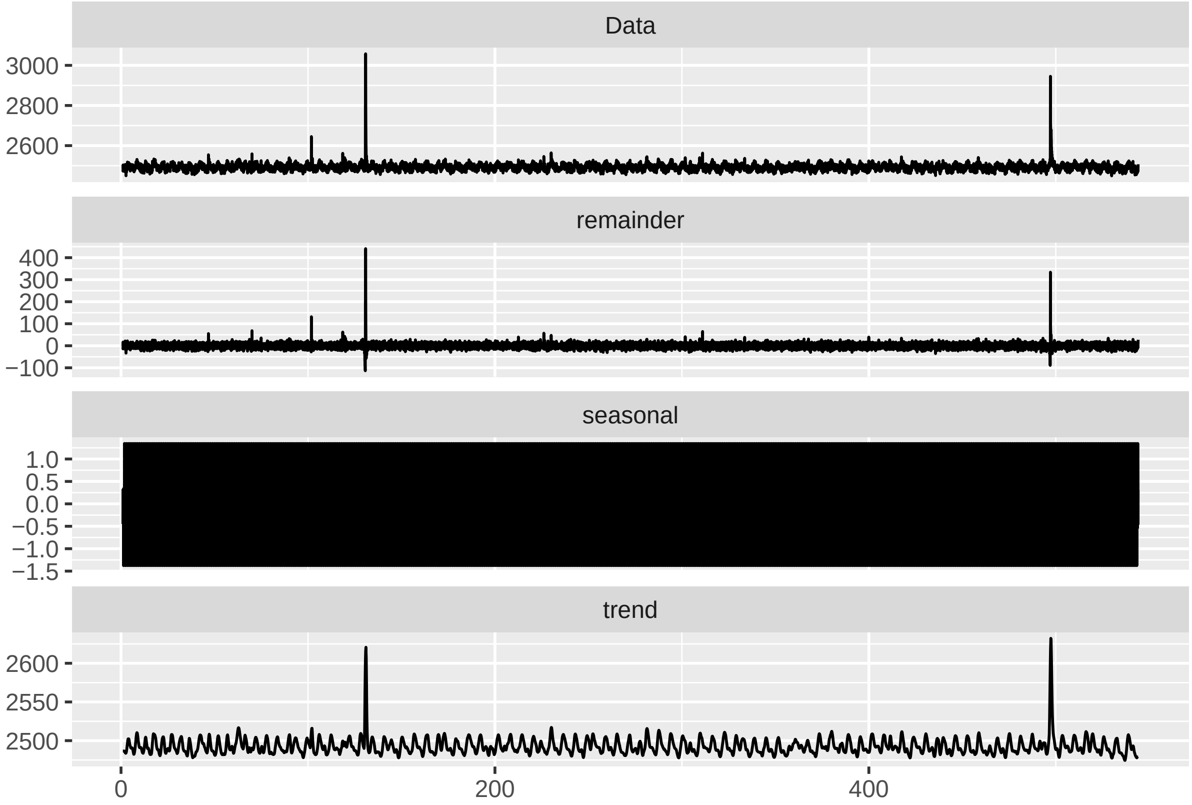


```r
time_series_analysis <- function(df, name) {
  # Ensure the time column is treated as a date
  df$time <- ymd(df$time)
  df <- df %>% arrange(time)

  # Remove missing values in time series data
  df <- df %>% drop_na(pdcsap_flux)

  # Plot time series
  ts_plot <- ggplot(df, aes(x = time, y = pdcsap_flux)) +
    geom_line() +
    labs(title=paste("Time Series Plot -", name), x="Time", y="PDCSAP Flux") +
    theme_minimal()
  print(ts_plot)

  # ACF and PACF
  ts_data <- ts(df$pdcsap_flux, frequency = 24)
  acf_plot <- autoplot(acf(ts_data, plot=FALSE)) + ggtitle(paste("Autocorrelation -", name))
  pacf_plot <- autoplot(pacf(ts_data, plot=FALSE)) + ggtitle(paste("Partial Autocorrelation -", name))
  print(acf_plot)
  print(pacf_plot)

  # Time Series Decomposition
  # decomposed <- decompose(ts_data, type="multiplicative")
  # decomposed_plot <- autoplot(decomposed)
  # print(decomposed_plot)
  decomposed_stl <- stl(ts_data, s.window="periodic")
  decomposed_stl_plot <- autoplot(decomposed_stl)
  print(decomposed_stl_plot)
}

time_series_analysis(data013_flux, "TIC 0131799991")
```

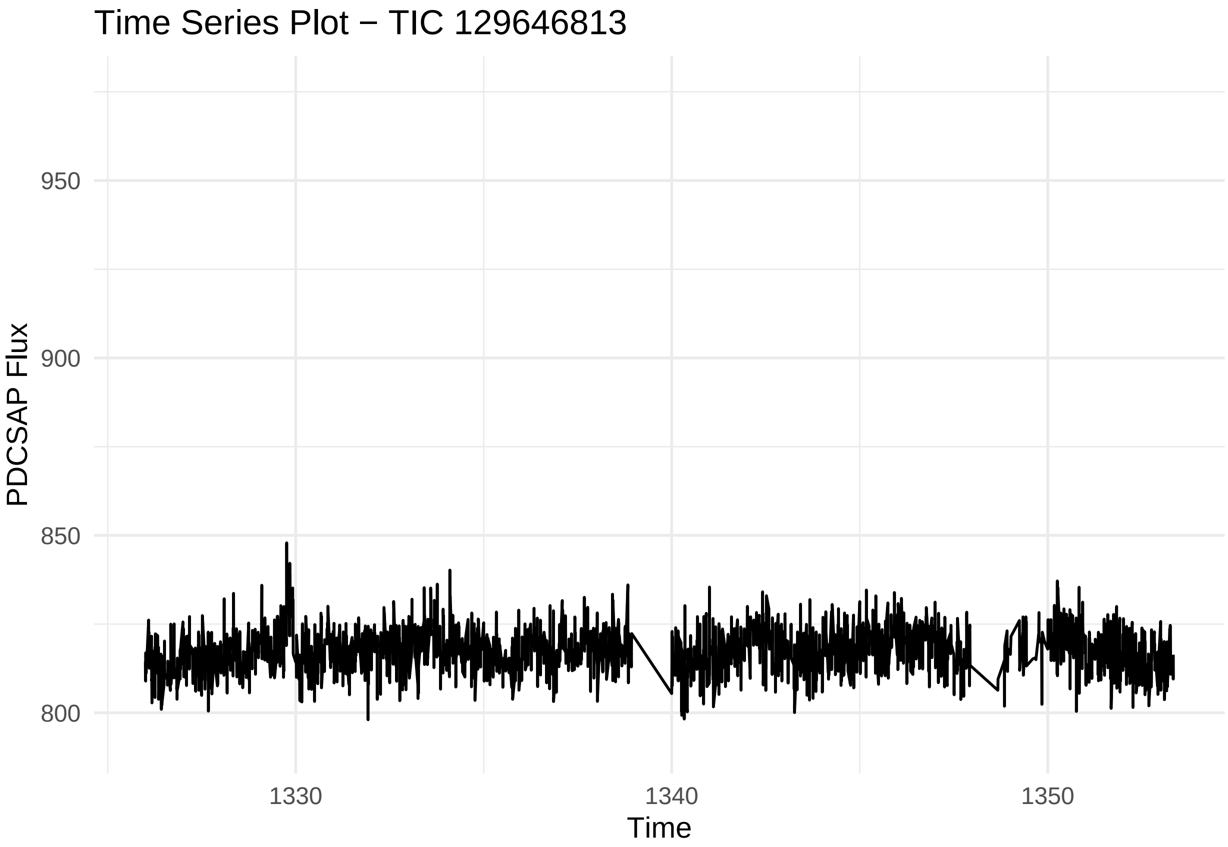


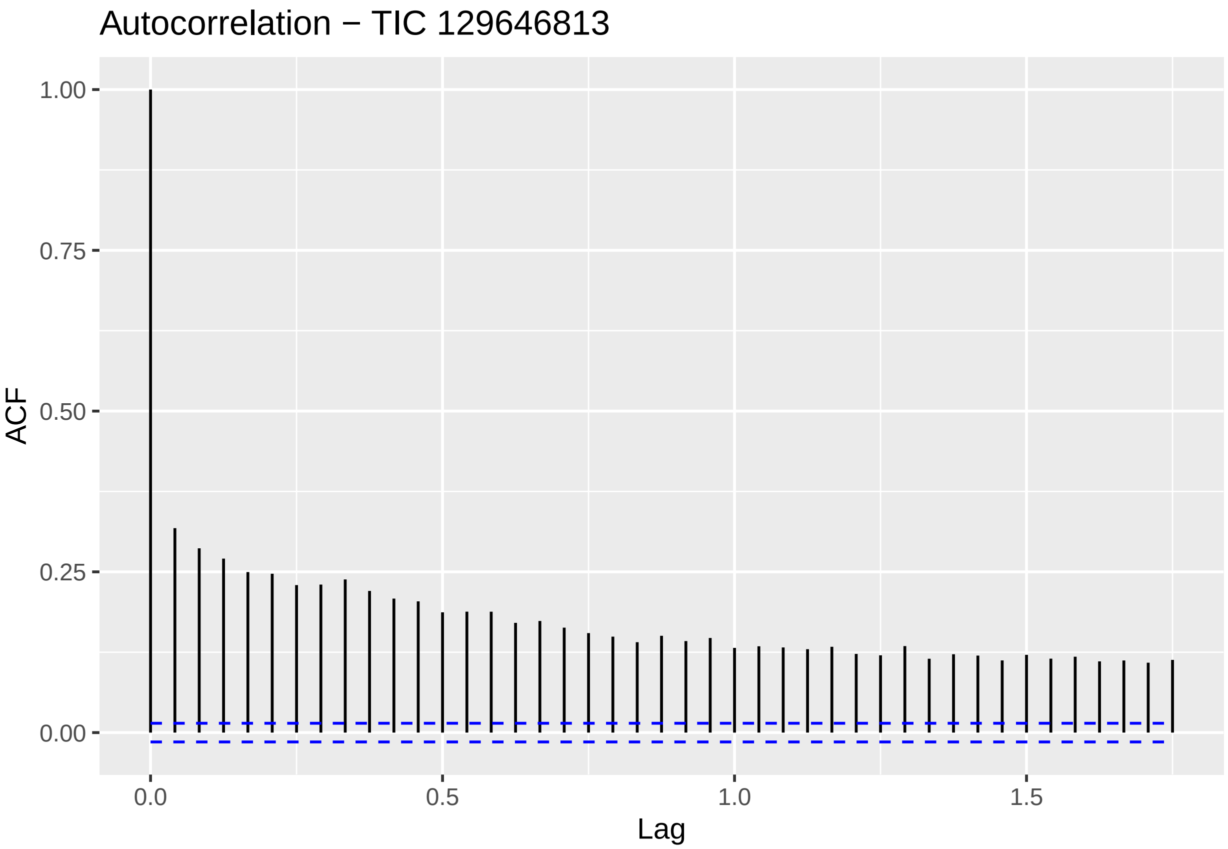

Autocorrelation − TIC 0131799991

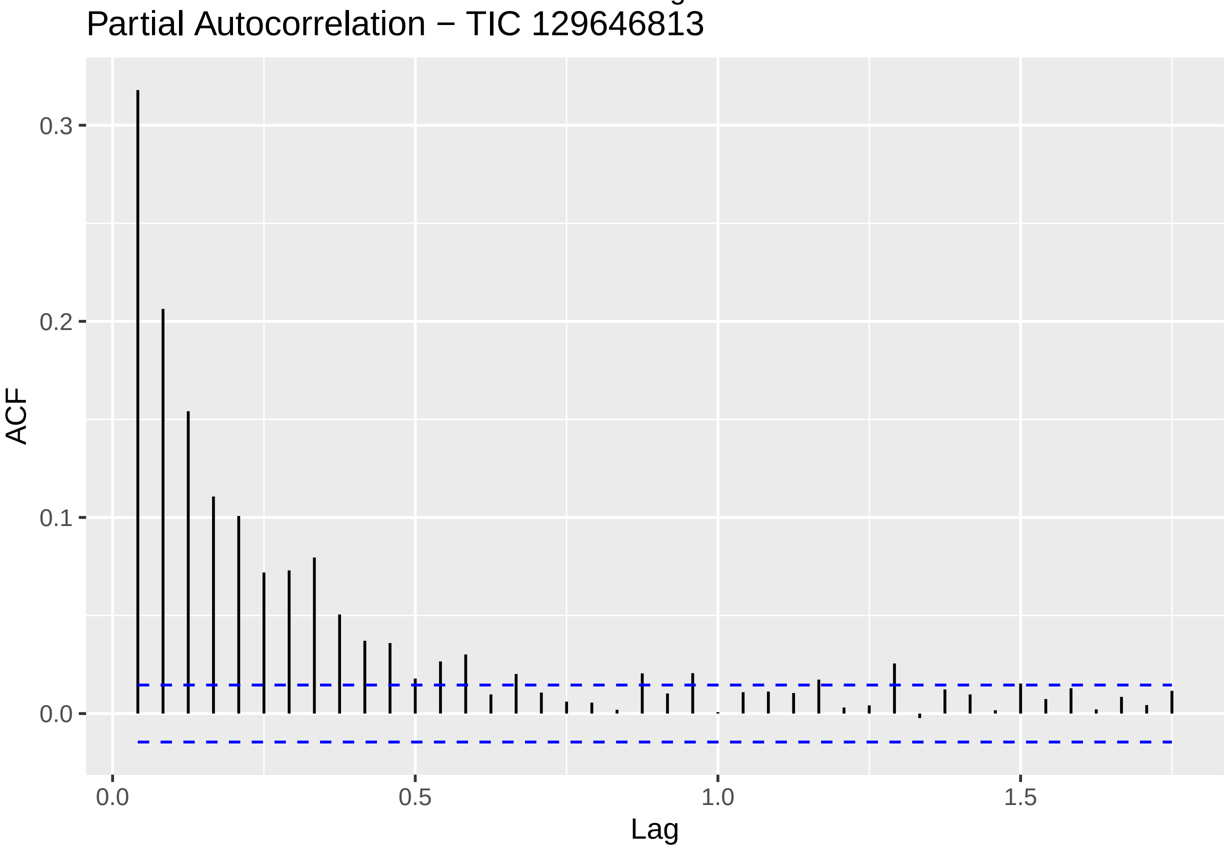

Partial Autocorrelation − TIC 0131799991

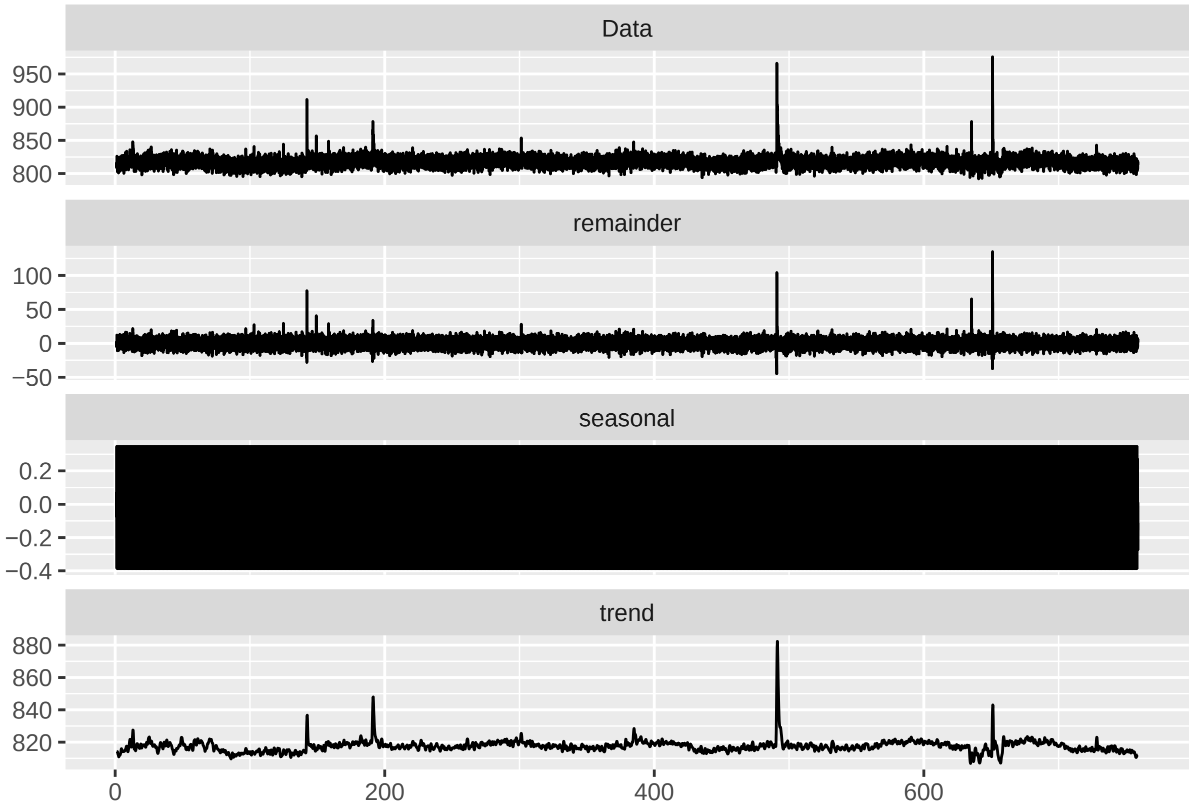


```
time_series_analysis(data129_flux, "TIC 129646813")
```

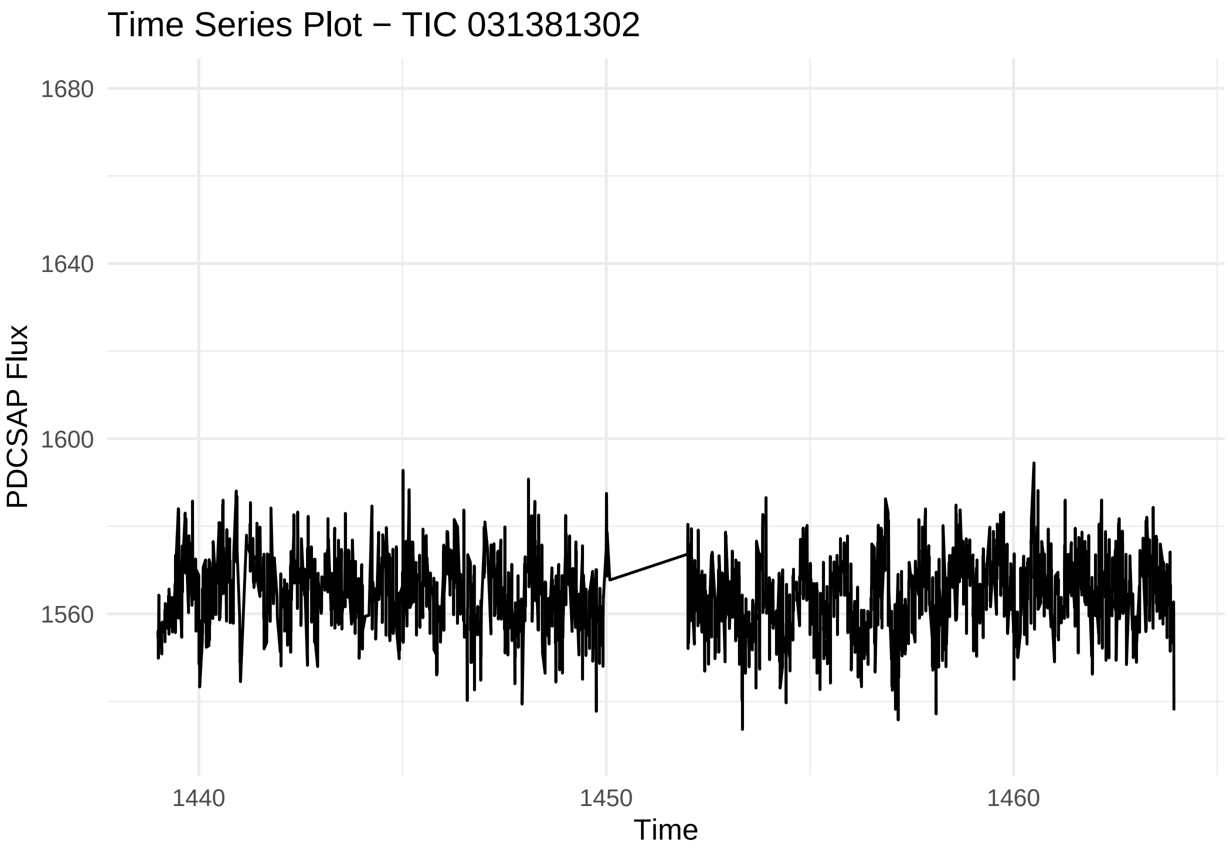


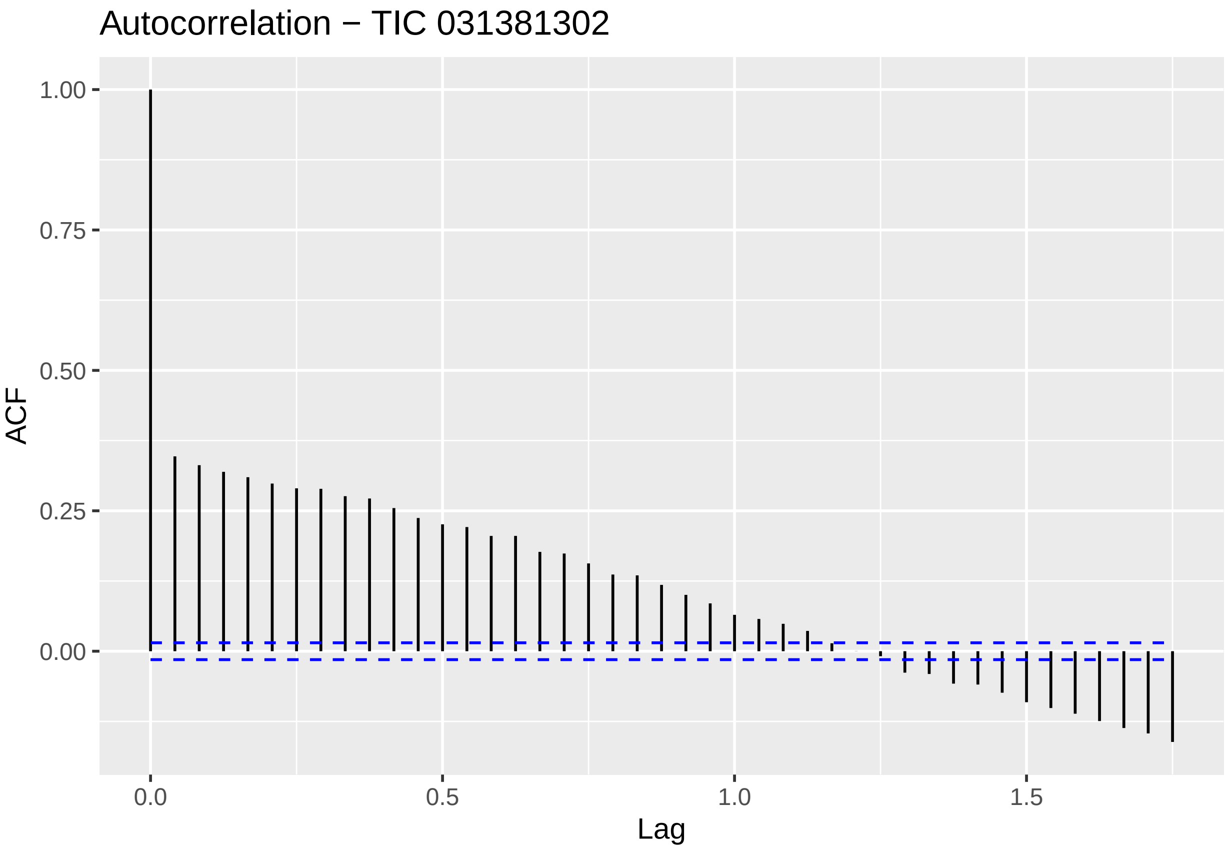

Autocorrelation − TIC 129646813
ACF
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
Lag

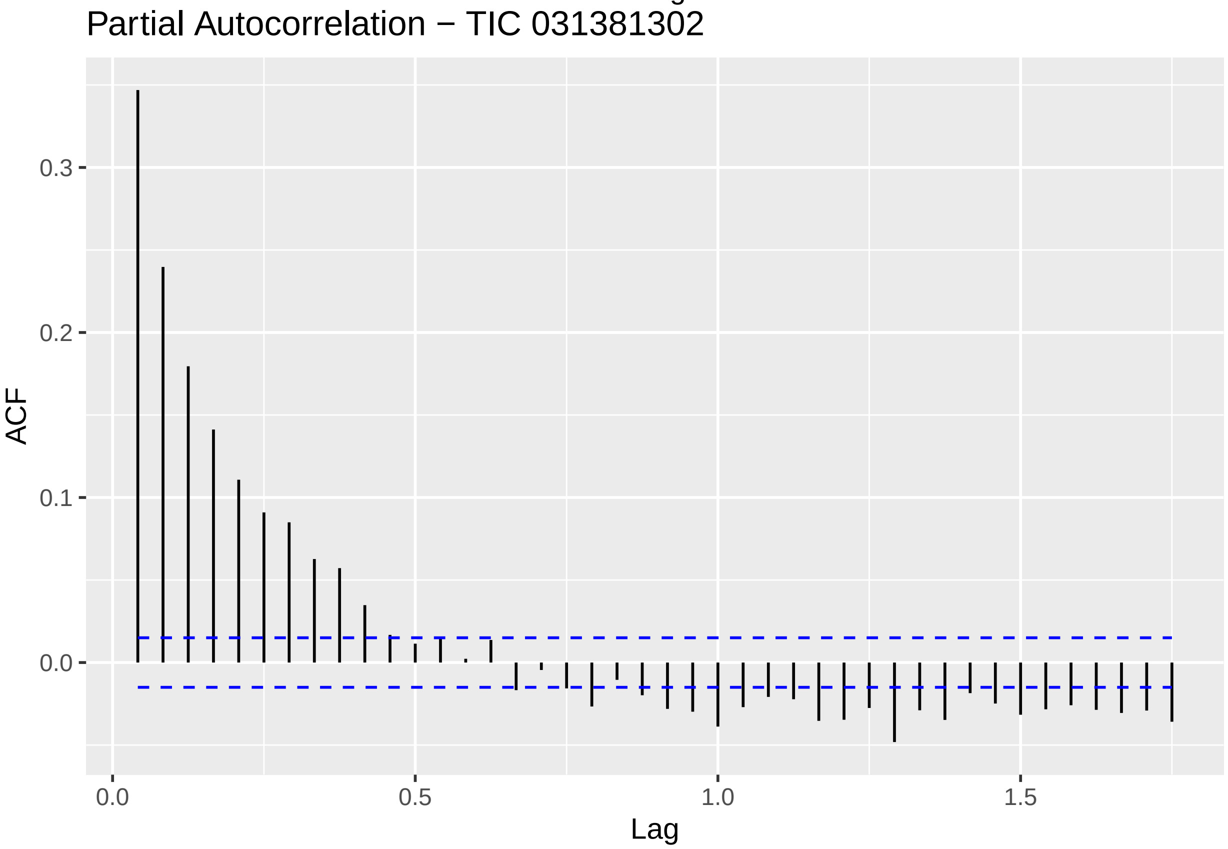

Partial Autocorrelation − TIC 129646813
ACF
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
Lag

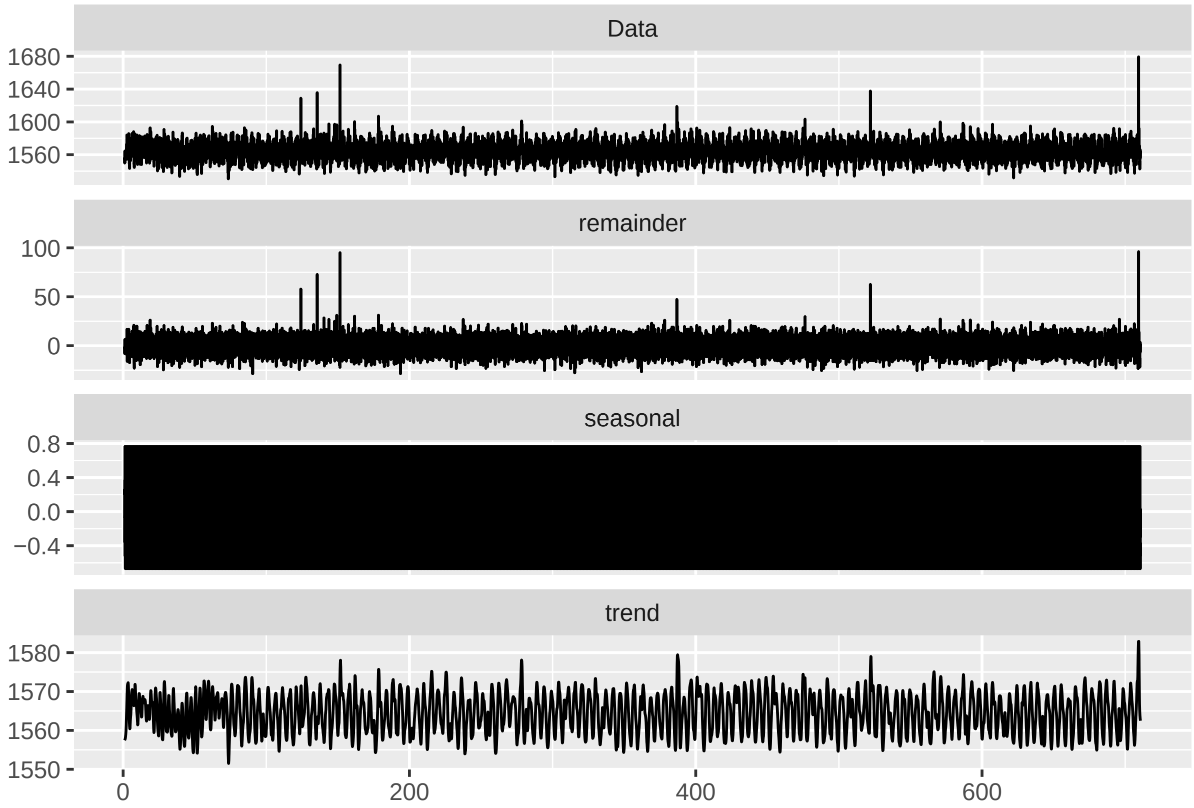


```
time_series_analysis(data031_flux, "TIC 031381302")
```

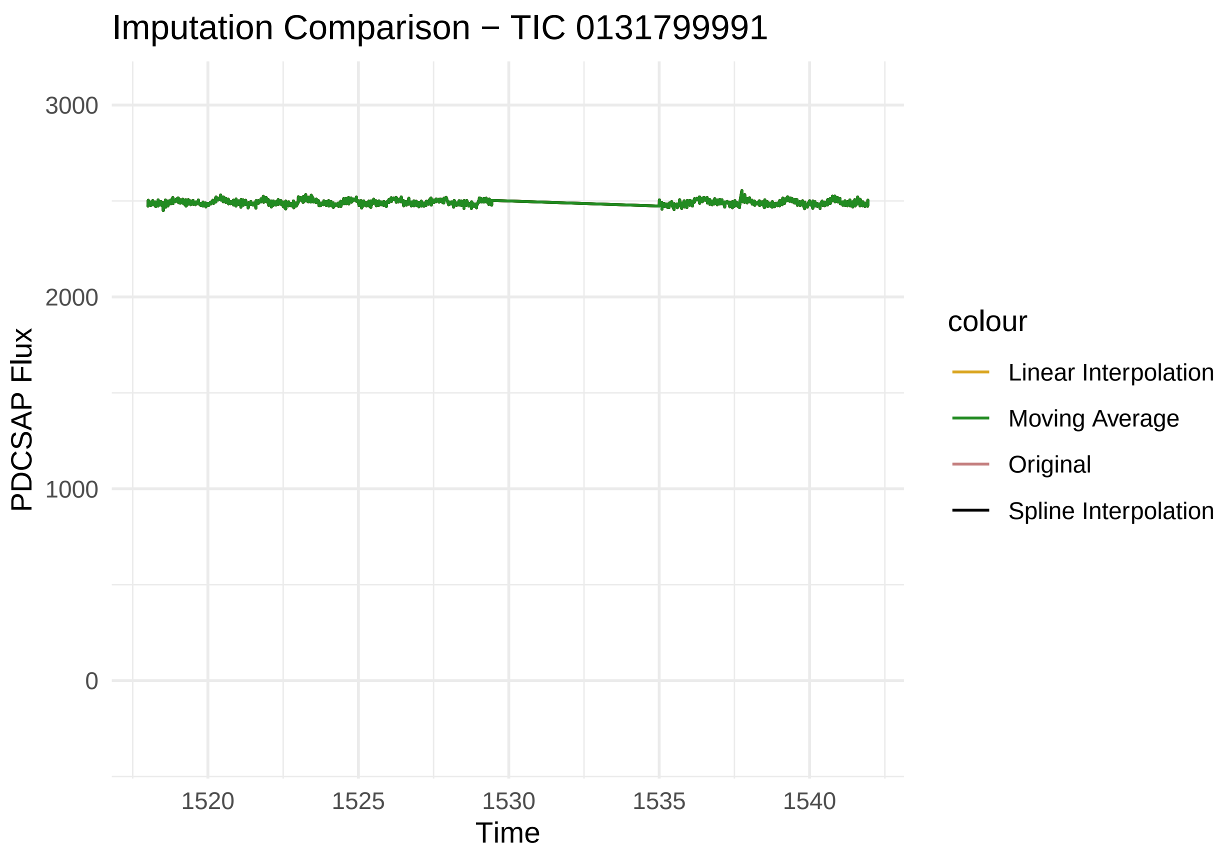



Autocorrelation − TIC 031381302
ACF
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
Lag

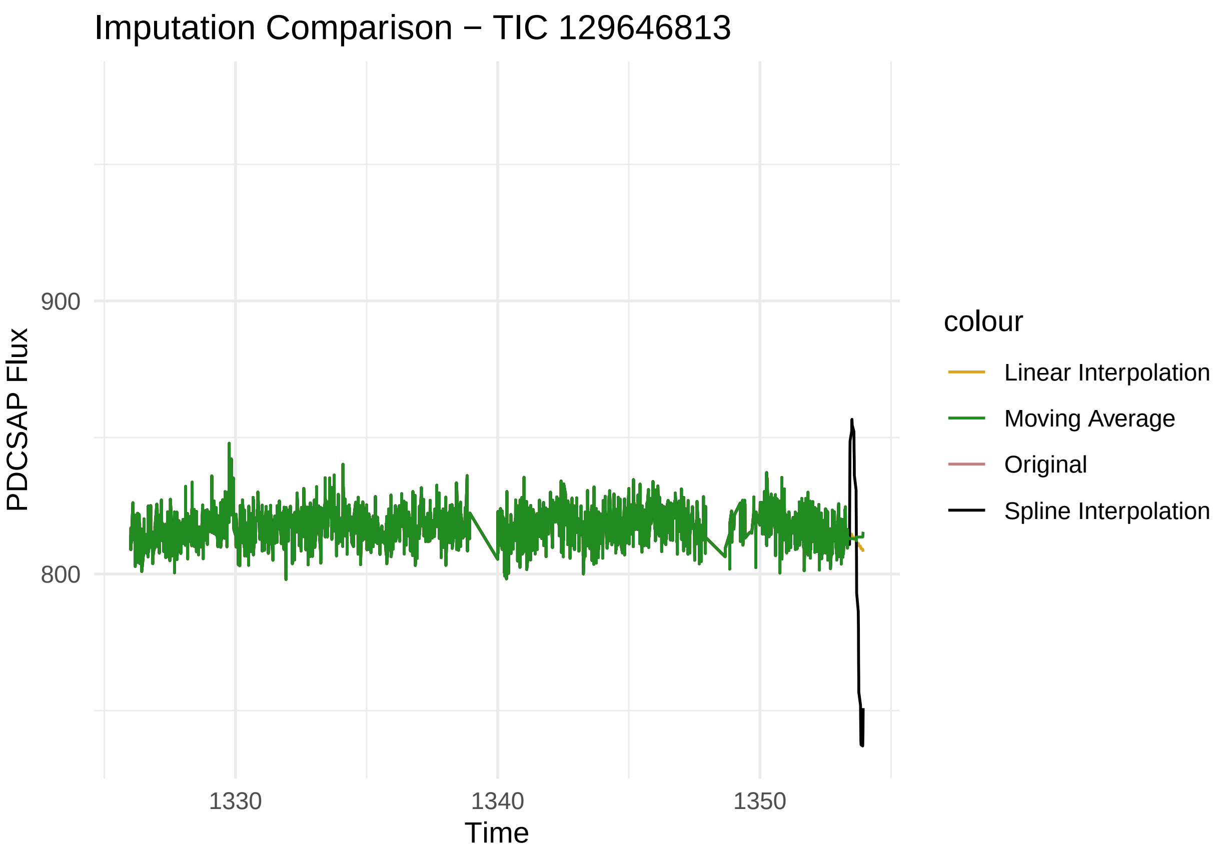

Partial Autocorrelation − TIC 031381302
ACF
0.3
0.2
0.1
0.0
0.0
0.5
1.0
1.5
Lag



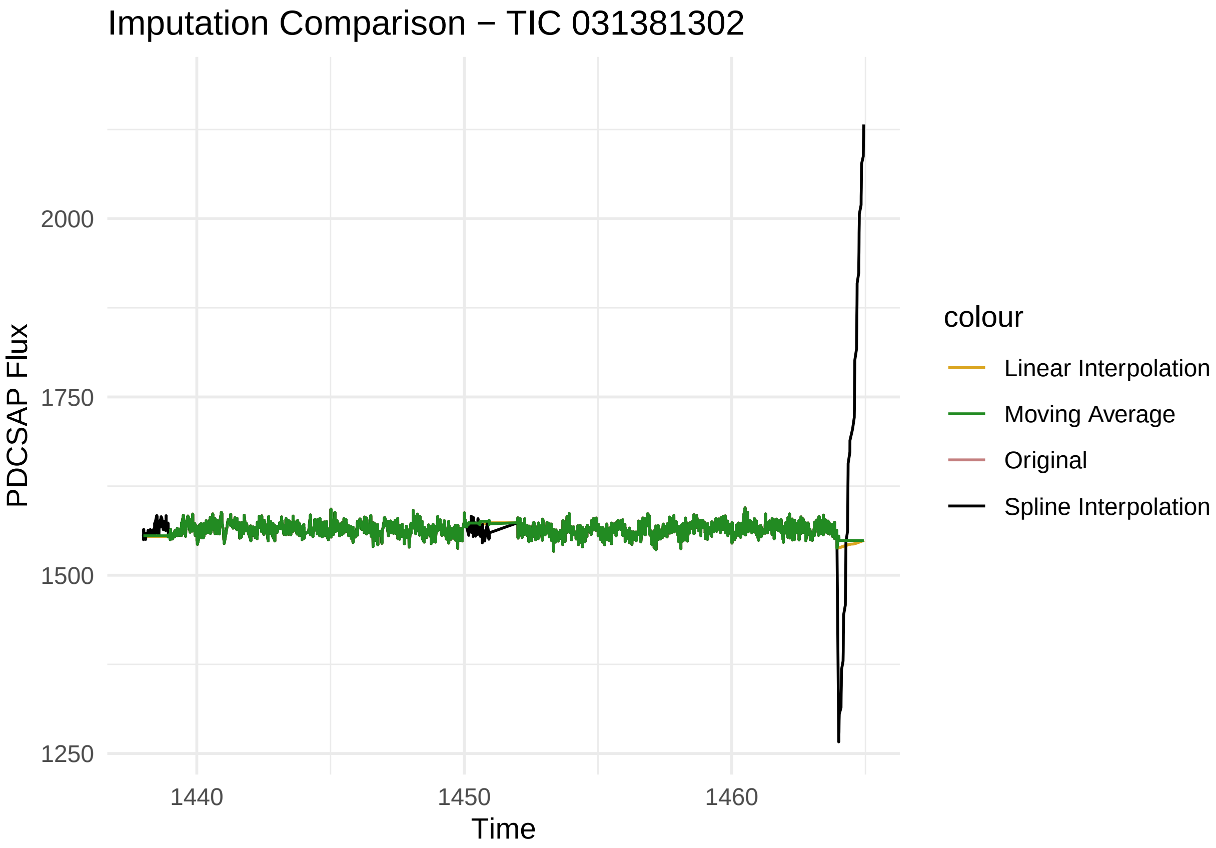
```r
impute_missing_values <- function(df, name) {
  # Ensure the time column is treated as a date
  df$time <- ymd(df$time)
  df <- df %>% arrange(time)

  # Convert to time series
  ts_data <- ts(df$pdcsap_flux, frequency=12)  # Adjust frequency if needed

  # Imputation methods
  df$pdcsap_flux_linear <- na_interpolation(ts_data, option="linear")  # Linear Interpolation
  df$pdcsap_flux_spline <- na_interpolation(ts_data, option="spline")  # Spline Interpolation
  df$pdcsap_flux_ma <- na_ma(ts_data, k=5, weighting="simple")  # Moving Average

  # Plot after imputation
  ggplot(df, aes(x = time)) +
    geom_line(aes(y = pdcsap_flux, color="Original"), alpha=0.5) +
    geom_line(aes(y = pdcsap_flux_linear, color="Linear Interpolation")) +
    geom_line(aes(y = pdcsap_flux_spline, color="Spline Interpolation")) +
    geom_line(aes(y = pdcsap_flux_ma, color="Moving Average")) +
    labs(title=paste("Imputation Comparison -", name), x="Time", y="PDCSAP Flux") +
    theme_minimal() +
    scale_color_manual(values=c("goldenrod", "forestgreen", "darkred", "black"))
}

impute_missing_values(data013_flux, "TIC 0131799991")
```



```
impute_missing_values(data129_flux, "TIC 129646813")
```



```
impute_missing_values(data031_flux, "TIC 031381302")
```


Imputation Comparison − TIC 031381302
PDCSAP Flux
2000
1750
1500
1250
1440
1450
1460
Time
colour
Linear Interpolation
Moving Average
Original
Spline Interpolation