Homework 2

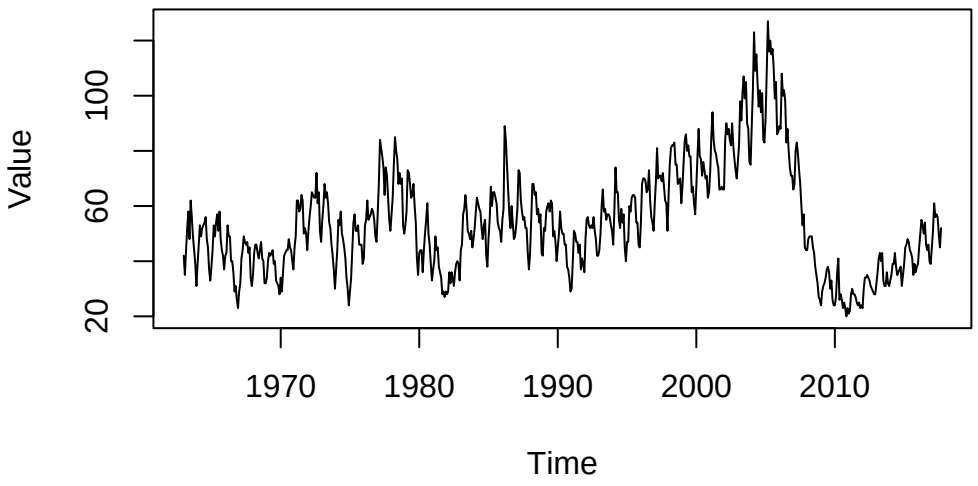
Zahlen Zbinden

2024-01-22

```
home <- home |>
          mutate(Value = as.numeric(Value)) |>
          na.omit() |>
          mutate(Period = my(Period)) |>
          mutate(Period = as.Date(Period, "%b-%Y"))
```

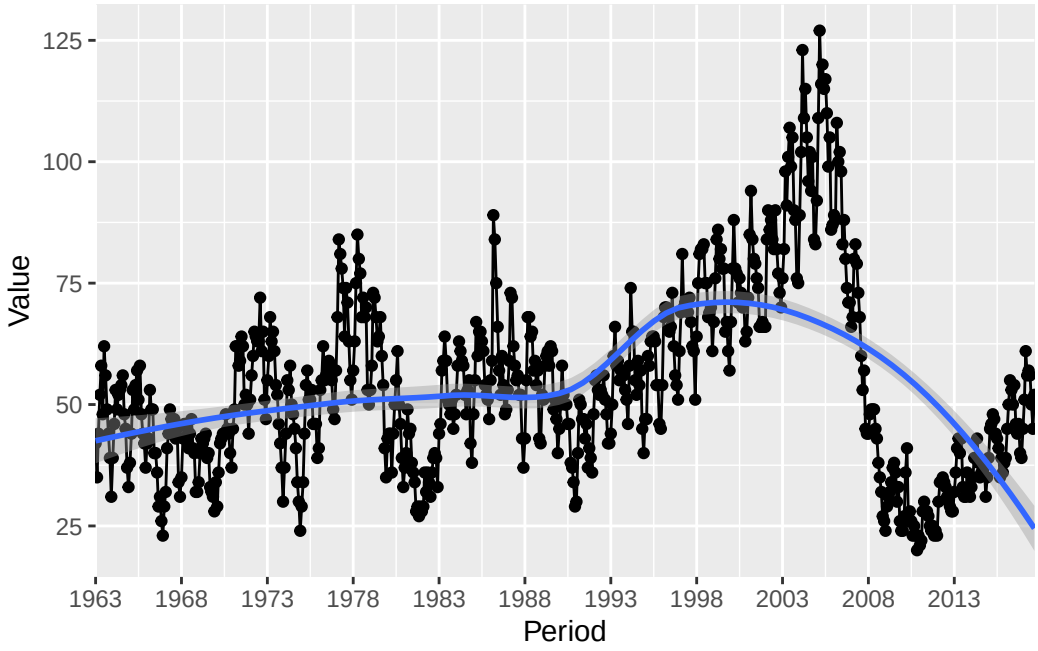
```
ts_data <- ts(
              home[,2],
              start = c(1963, 1),
              frequency = 12
            )
```

```
plot(ts_data)
```




```
home |>
  ggplot(
    aes(x = Period, y = Value)
  ) +
    geom_point(
    ) +
    geom_line(
      size = .5
    ) +
    geom_smooth() +
    scale_x_date(
      limits = as.Date(c("1963-01-01", "2017-09-01")),
      breaks = "5 years",
      date_labels = "%Y",
      expand = c(0,0)
    )
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



1. We can see an upward trend from 1963 to approximates 2000, with a sharp increase from 1990 to 2000. However after the year 2000 we see an even steeper decline, to a whole new low for the entire time period. We can see a seasonality that ramps up and down through each year of observations, most likely having to do with housing purchases going up in the summer when it is nice out and looking at houses is appealing, and going down in the winter when it is cold out and looking at houses isn't as fun. We can see from the smoothed line, that the data doesn't follow a strictly quadratic form and a linear model will not do us any good here.

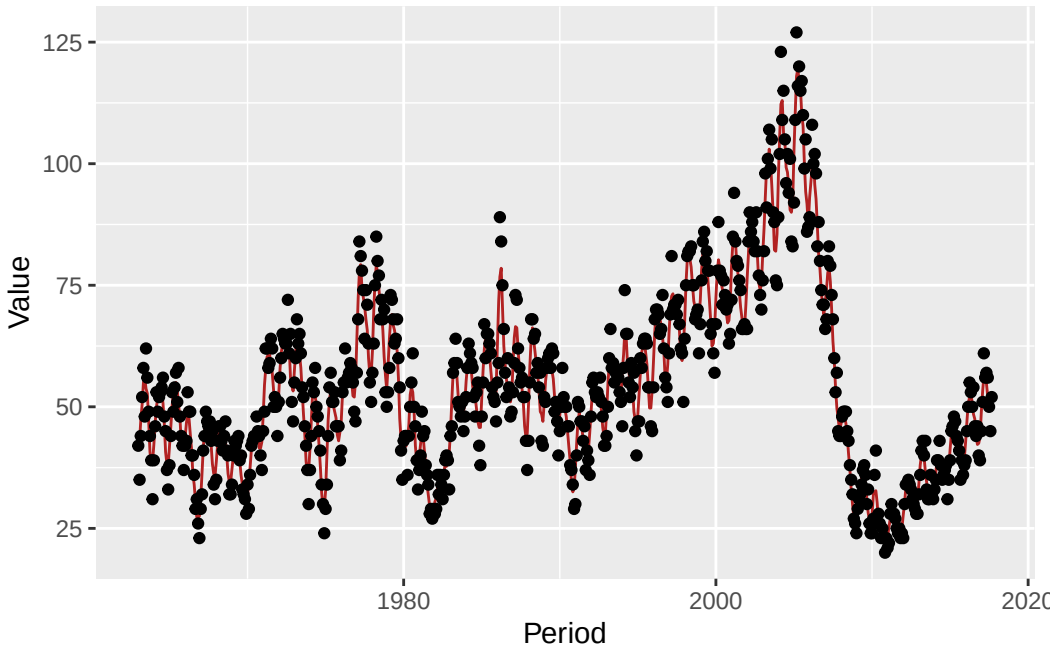
Removing Seasonality with Moving average

```
home_ma <- stats::filter(home, filter = rep(1/4, 4))
```

```
home |>
  ggplot(
    aes(x = Period, y = Value)
  ) +
    geom_line(
      aes(y = rollmean(Value, 4, na.pad = TRUE)),
      color = "firebrick"
      ) +
```

```
    geom_point()
```

```
Warning: Removed 3 rows containing missing values (`geom_line()`).
```



Removing Seasonality with Kernel Smoothing

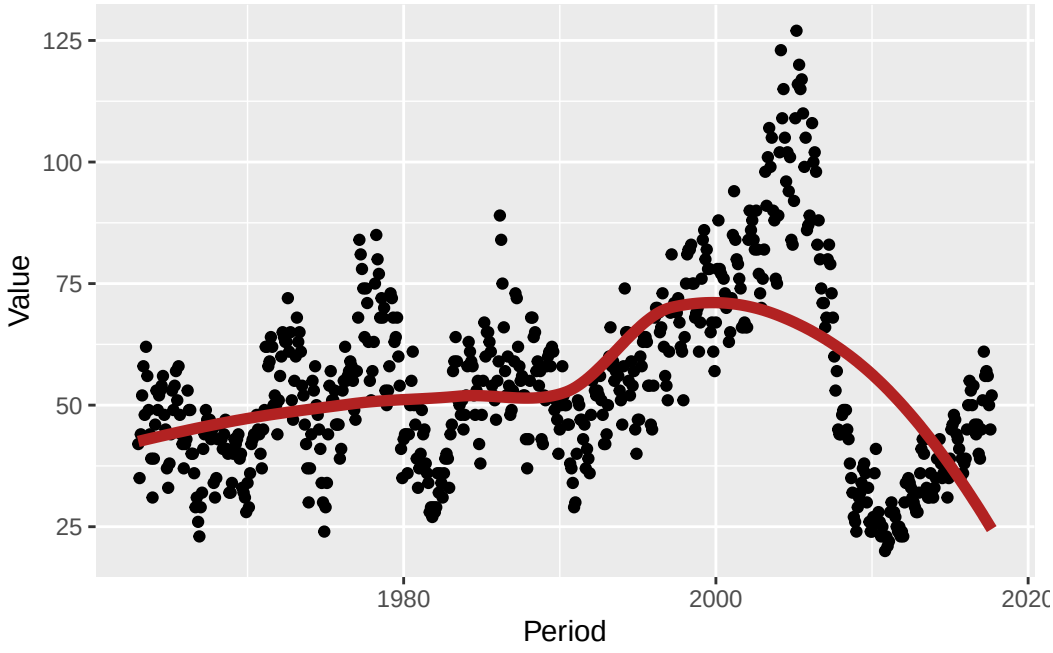
```
ts_time <- time(ts_data)
```

```
ts_loess <- loess(ts_data ~ ts_time)
ts_loess_pred <- predict(ts_loess)
```

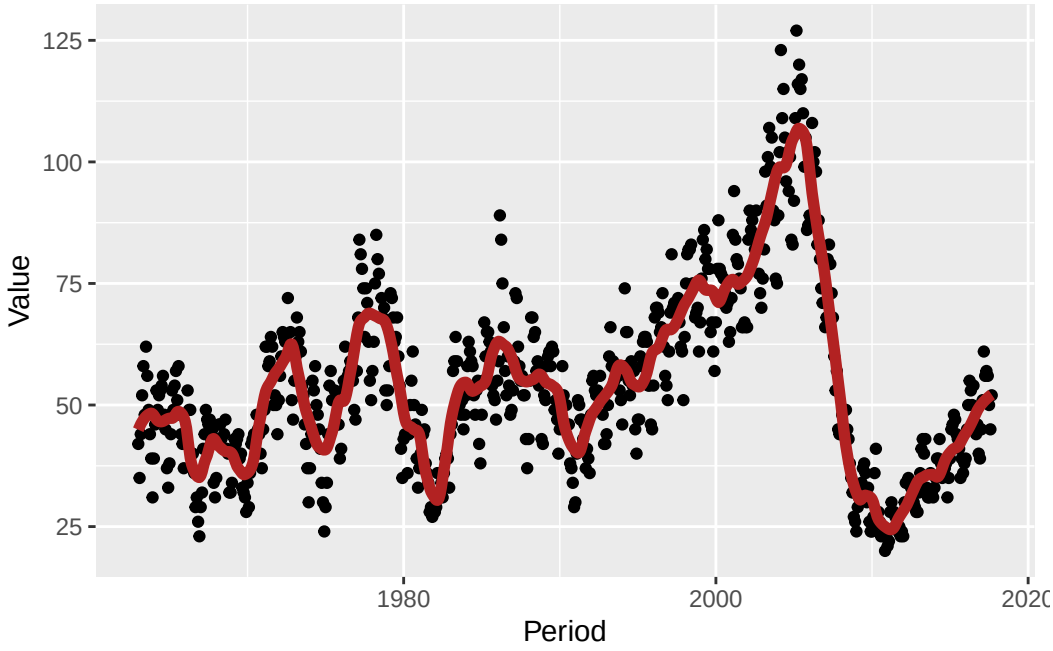
```
home$Predicted <- predict(ts_loess)
```

```
home |>
  ggplot(
    aes(x = Period, y = Value)
  ) +
    geom_point(
```

```
  ) +
  geom_line(
    aes(x = Period, y = predict(ts_loess)),
    color = "firebrick",
    size = 2
  )
```

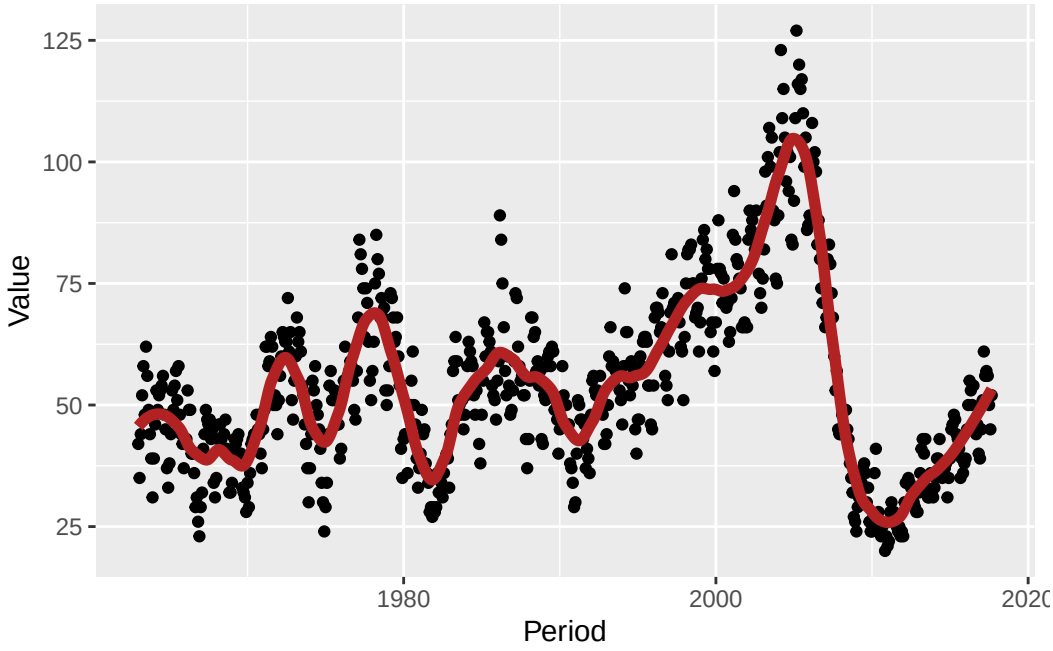



```
loess_span <- loess(ts_data ~ ts_time, span = 0.05)
ls_predict <- predict(loess_span)
home |>
  ggplot(
    aes(x = Period, y = Value)
  ) +
    geom_point() +
    geom_line(
      aes(x = Period, y = ls_predict),
      size = 2,
      color = "firebrick"
    )
```



We can see that the trend and seasonality are not as "lost" when using a very small span during the kernel smoothing operation, lets try with a slightly higher span, and see if we get a "less smooth" line than we do with the default span, but also a more descriptive "line"

```
loess_span <- loess(ts_data ~ ts_time, span = 0.1)
ls_predict <- predict(loess_span)
home |>
  ggplot(
    aes(x = Period, y = Value),
  ) +
    geom_point() +
    geom_line(
      aes(x = Period, y = ls_predict),
      size = 2,
      color = "firebrick"
    )
```

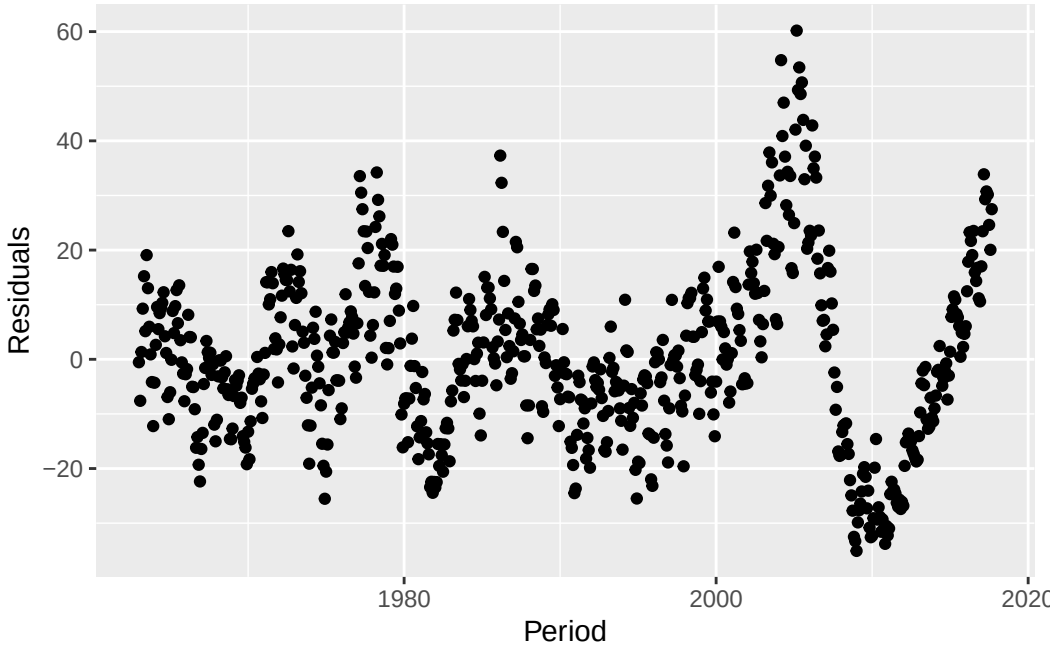


We can see in this plot that the small changes throughout the years has been lost (from the small span). With the larger value for the span we can get a much more accurate picture of the housing market as it rises and falls, without viewing the seasonal trend between summer and winter.

Lets now remove the predicted values from the original data and look at just the residuals.

```
home <- home |>
          mutate(Residuals = Value - Predicted)
```

```
home |>
  ggplot(
    aes(x = Period, y = Residuals)
  ) +
    geom_point()
```



We can see that the residuals are split pretty evenly between over and underestimating. We can see that the variance increases over time, as the residuals get larger in magnitude through time. When looking at models, like the latter kernel smoothing operation to obtain predicted values we can see two things. With a very small value the moving predictions still contain seasonality over the months of the year, with the larger non-default value we see a decrease in the seasonality of over the months and start to notice one over the course of the years. With the default value we can see that there is a slight upward trend in the first half of the data, then a strong increase, followed by a sharp decline. Overall I do not think that this data is stationary, as the statistics of the first half of the data, the third quarter, and final quarter are all completely different, with drastically different means. Further techniques will be needed to continue the disection of this data set.