PCA_feature-based-visualization

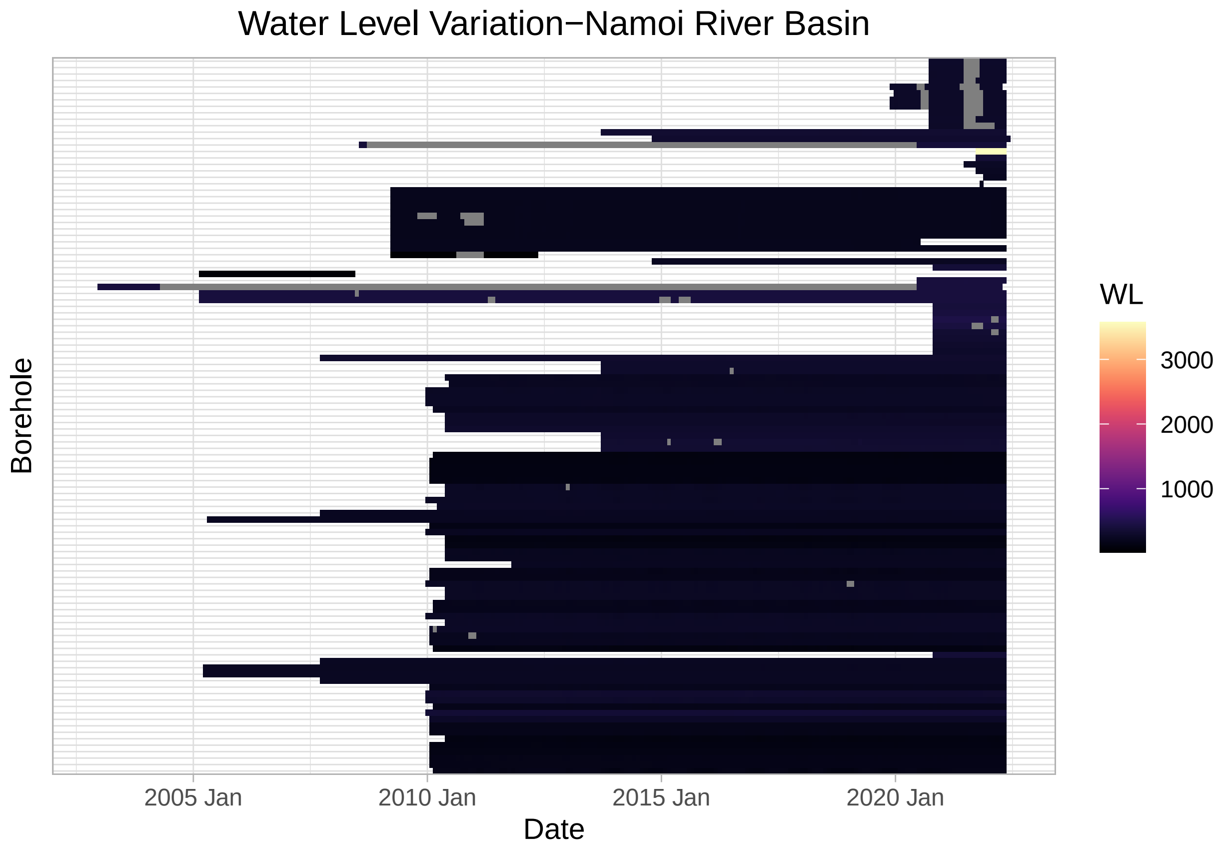
```r
#Load data for water level variation of all boreholes in Namoi River Basin
dt <- read_csv("TSvizDATA.csv")
```

```
## New names:
## Rows: 12726 Columns: 11
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (3): Date, Borehole, ...11 dbl (2): WL, No of Covariates lgl (6): ...5, ...6,
## ...7, ...8, ...9, ...10
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
```

```r
dt$Date <- tsibble::yearmonth(dt$Date, format="%m/%d/%Y")

dt<-select(dt, c(WL,Date,Borehole))
```

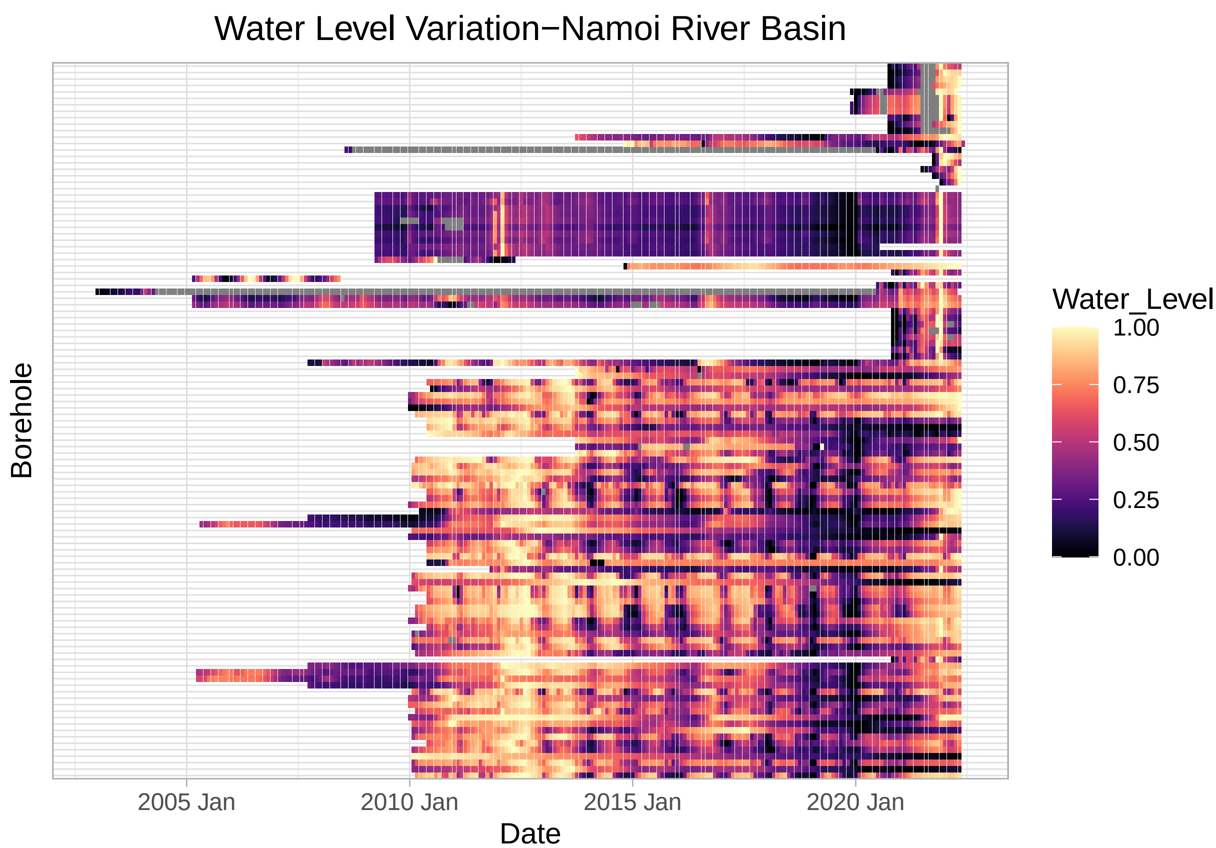
Lets see how the water level of each of the boreholes varies by Heat map Visualization.

```r
int_plt<-ggplot(dt, aes(x=Date, y=Borehole)) +
    geom_tile(aes(fill = WL)) +
    scale_fill_viridis(option="magma") +
    labs(title = "Water Level Variation-Namoi River Basin",y = "Borehole") +
    theme_light()+ theme(legend.position = 'right',axis.ticks.y =     element_blank(),axis.text.y = eler

int_plt
```



Water levels in Namoi River Basin Region varies in a huge range. Hence, it is impossible to identify the pattern variation much more clearly. Hence, we apply the minimax transformation for each series separately to get an idea about water level variations.

```
borehole_WL <- dt %>% select(WL, Borehole)
# Transform the data into 0-1 range to see the pattern much more clearly.
borehole_WL2 <- borehole_WL %>%
  group_by(Borehole) %>%
  summarise(
    MaxWLByBorehole = max(WL, na.rm = T),
    MinWLByBorehole = min(WL, na.rm = T)
  ) %>%
  arrange(Borehole)

New_DF<-merge(dt,borehole_WL2,by='Borehole')
New_DF$Water_Level <- (New_DF$WL-New_DF$MinWLByBorehole)/(New_DF$MaxWLByBorehole-New_DF$MinWLByBorehole)

HMT2<-ggplot(New_DF, aes(x=Date, y=Borehole)) +
    geom_tile(aes(fill = Water_Level)) +
    scale_fill_viridis(option="magma") +
    labs(title = "Water Level Variation-Namoi River Basin",y = "Borehole") +
    theme_light()+ theme(legend.position = 'right',axis.ticks.y =     element_blank(),axis.text.y = eler


HMT2
```



From this heat map visualization, it is impossible to get an idea about time series pattern variations. Hence, next we are going to consider feature based visualization.

Feature-based visualization

```
# Extracting time series features from water level variation time series
library(feasts)
library(tsibble)
# Consider the data of boreholes without missing values
data_pca <- read_csv("PCAdata.csv")
```

```
## New names:
## Rows: 13436 Columns: 5
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Date, Borehole dbl (2): WL, ...5 lgl (1): ...4
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...4`
## * `` -> `...5`
```

```
data_pca<-select(data_pca, c(WL,Date,Borehole))
Month<-tsibble::yearmonth(data_pca$Date)
BH<-data_pca$Borehole
Water<-data_pca$WL
data_all<-tibble(Month,BH,Water)
data_all<-data_all%>% as_tsibble(key = BH,index=Month)
```

```r
pca_features <-data_all %>% features(Water, feature_set(pkgs = "feasts"))
```

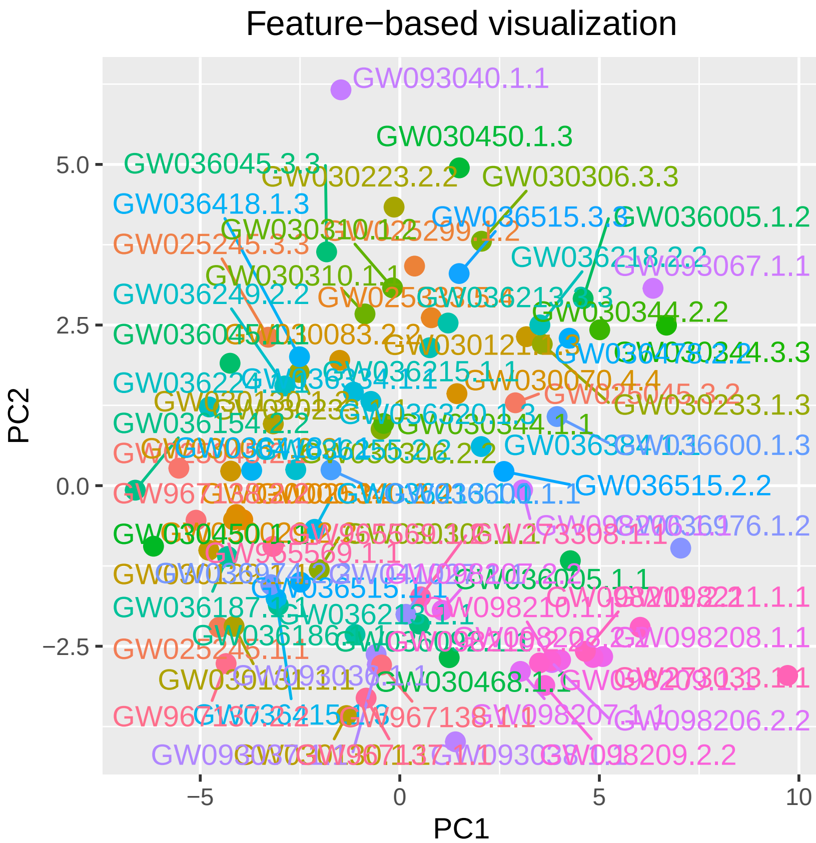
```
## Warning: `n_flat_spots()` was deprecated in feasts 0.1.5.
## Please use `longest_flat_spot()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.

## Warning: 1 error encountered for feature 3
## [1] missing values in object

## Warning: 11 errors (1 unique) encountered for feature 16
## [11] length of `x` is too short for `.size`.

## Warning: 1 error encountered for feature 20
## [1] invalid time series parameters specified
```

```r
pca_features
```

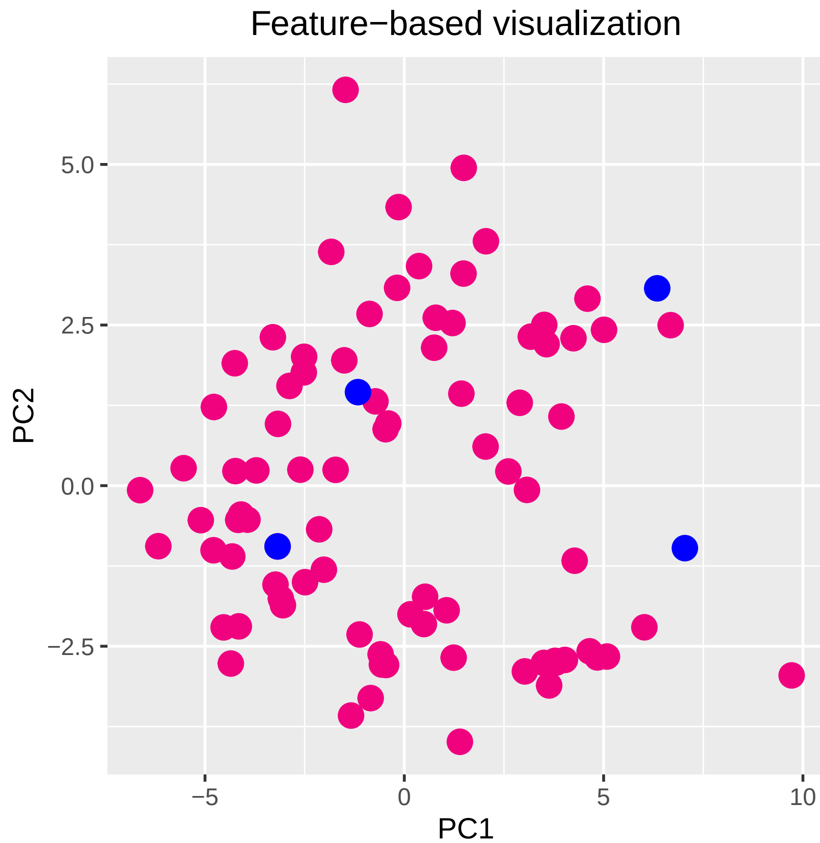
```
## # A tibble: 100 x 49
##    BH          trend_strength seasonal_streng~ seasonal_peak_y~ seasonal_trough~
##    <chr>                <dbl>            <dbl>            <dbl>            <dbl>
##  1 GW021266.3~          0.614            0.615                4               11
##  2 GW025045.2~          0.988            0.589                8                2
##  3 GW025045.3~          0.724            0.610                6                0
##  4 GW025245.1~          0.999            0.120                9                2
##  5 GW025245.3~          0.908            0.631                6                1
##  6 GW025299.1~          0.793            0.511                3                8
##  7 GW025333.5~          0.837            0.703                6                0
##  8 GW025340.2~          0.965            0.526                7                1
##  9 GW030000.1~          0.970            0.435                8                2
## 10 GW030029.2~          0.969            0.253                0                5
## # ... with 90 more rows, and 44 more variables: spikiness <dbl>,
## #   linearity <dbl>, curvature <dbl>, stl_e_acf1 <dbl>, stl_e_acf10 <dbl>,
## #   acf1 <dbl>, acf10 <dbl>, diff1_acf1 <dbl>, diff1_acf10 <dbl>,
## #   diff2_acf1 <dbl>, diff2_acf10 <dbl>, season_acf1 <dbl>, pacf5 <dbl>,
## #   diff1_pacf5 <dbl>, diff2_pacf5 <dbl>, season_pacf <dbl>,
## #   zero_run_mean <dbl>, nonzero_squared_cv <dbl>, zero_start_prop <dbl>,
## #   zero_end_prop <dbl>, lambda_guerrero <dbl>, kpss_stat <dbl>, ...
```

```r
# Removing the boreholes which contains "NA" as time series features
pca_features <- pca_features %>% filter(`BH`!= "GW021266.3.4")%>%     filter(`BH`!= "GW093000.1.1")%>%
  filter(`BH`!= "GW030136.1.1")%>%
  filter(`BH`!= "GW036215.1.3")%>%
  filter(`BH`!= "GW093002.1.1")%>%
  filter(`BH`!= "GW093008.1.1")%>%
  filter(`BH`!= "GW093013.1.1")%>%
  filter(`BH`!= "GW093021.1.1")%>%
  filter(`BH`!= "GW093028.1.1")%>%
  filter(`BH`!= "GW093029.1.1")%>%
  filter(`BH`!= "GW093033.1.1")%>%
  filter(`BH`!= "GW093039.1.1")%>%
  filter(`BH`!= "GW093042.1.1")

# Apply PCA
library(broom)
library(dplyr)
```

```r
pcsW <- pca_features %>%
  dplyr::select(-c(1,2)) %>%
  dplyr::select(-c("zero_run_mean","zero_start_prop","zero_end_prop","bp_pvalue","lb_pvalue")) %>%
  prcomp(scale = TRUE) %>%
  augment(pca_features)

# Labelling for giving different colours for selected boreholes

SorN<-c(rep("N",47),"Y",rep("N",12),"Y",rep("N",5),"Y",rep("N",15),"Y",rep("N",4))

AA<-pcsW

AA$YN<-SorN
library(ggplot2)
library(ggrepel)

Feature_variatn_plot1<-pcsW %>%
  ggplot(aes(x = .fittedPC1, y = .fittedPC2, col = BH, label=BH))     + geom_point(size=3) +
  theme(aspect.ratio = 1)+
  geom_text_repel(aes(label=BH), max.overlaps = Inf)+
  labs(x="PC1", y = "PC2",title = "Feature-based visualization")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")

Feature_variatn_plot2<-AA %>%
  ggplot(aes(x = .fittedPC1, y = .fittedPC2, col = YN))     + geom_point(size=4) +
  theme(aspect.ratio = 1)+
  labs(x="PC1", y = "PC2",title = "Feature-based visualization")+
  theme(plot.title = element_text(hjust = 0.5),legend.position="none")+  scale_color_manual(breaks = c(
                                                                                              values=c("bl
Feature_variatn_plot1
```




```
Feature_variatn_plot2
```



Feature−based visualization
PC2
5.0
2.5
0.0
−2.5
−5
0
5
10
PC1