Lab Assessment - 4

Name	:	Hemant H Kumar
Reg No	:	16BCE2004
Instructor's Name	:	Prof. Meenakshi S P
Course	:	CSE3020 (Data Visualisation)
Slot	:	L39+L40
Date of Submission	:	28 March 2019

Data Visualisation Lab Experiment – 4

(Network Data techniques)

Question 1

Problem 10

1) Take a text document. Show the word cloud for the document.

CODE and OUTPUT

```
# Install
#install.packages("tm")  # for text mining
#install.packages("SnowballC") # for text stemming
#install.packages("wordcloud") # word-cloud generator
#install.packages("RColorBrewer") # color palettes
# Load
library("tm")
library("SnowballC")
library("wordcloud")
library("RColorBrewer")

# Read the Gettysburg Address by Abraham Lincoln
filePath <- "https://raw.githubusercontent.com/timburks/gott/
master/test/gettysburg-address.txt"
text <- readLines(filePath)

# Load the data as a corpus
docs <- Corpus(VectorSource(text))

inspect(docs)
```

```
> inspect(docs)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 37

 [1] THE GETTYSBURG ADDRESS:
 [2]
 [3]
 [4] Four score and seven years ago our fathers brought forth on this
 [5] continent a new nation, conceived in liberty and dedicated to the
 [6] proposition that all men are created equal. Now we are engaged in
 [7] a great civil war, testing whether that nation or any nation so
 [8] conceived and so dedicated can long endure. We are met on a great
 [9] battlefield of that war. We have come to dedicate a portion of
[10] that field as a final resting-place for those who here gave their
[11] lives that that nation might live. It is altogether fitting and
[12] proper that we should do this. But in a larger sense, we cannot
[13] dedicate, we cannot consecrate, we cannot hallow this ground.
[14] The brave men, living and dead who struggled here have consecrated
[15] it far above our poor power to add or detract. The world will
```

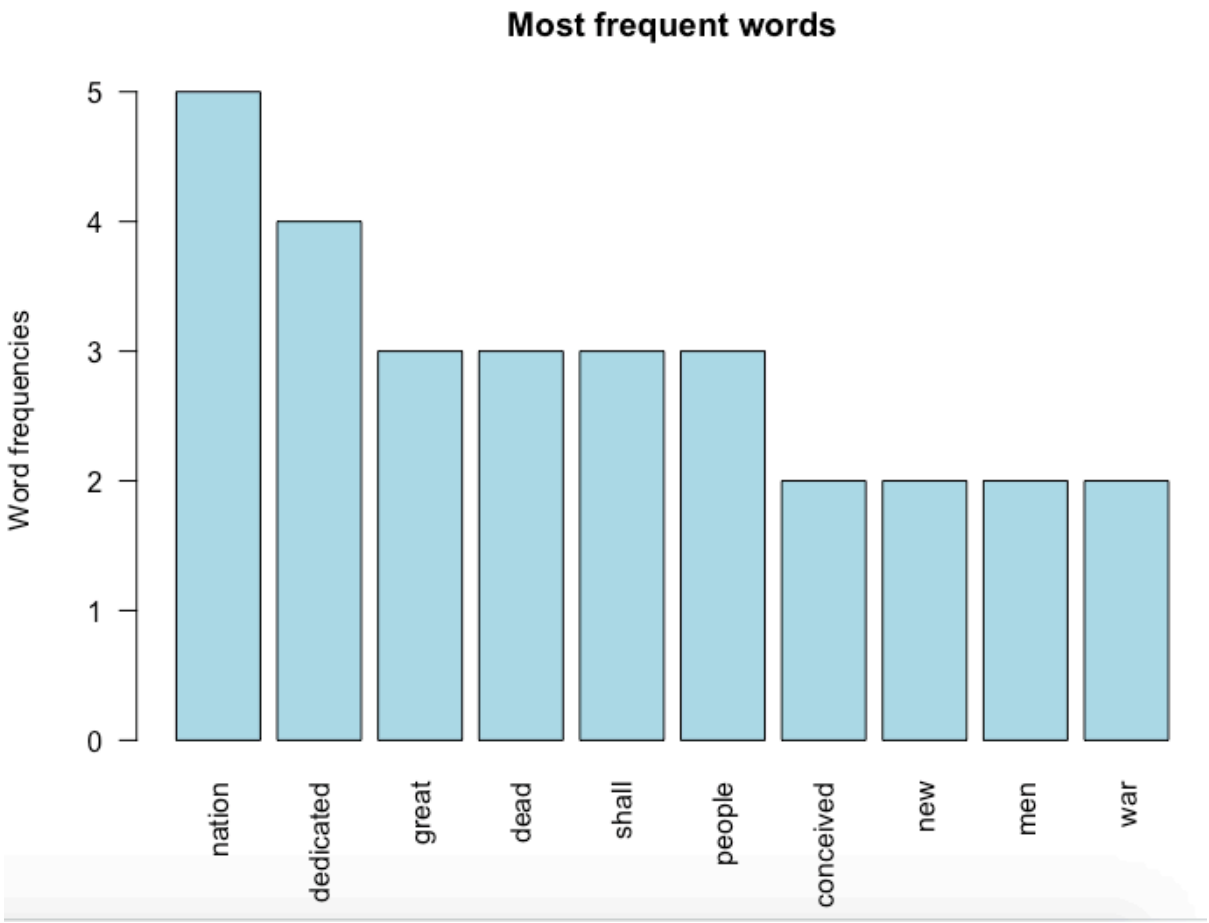
```
toSpace <- content_transformer(function (x , pattern )
gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")

# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
# docs <- tm_map(docs, stemDocument)

dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

```
> head(d, 10)
                word freq
nation       nation    5
dedicated dedicated    4
great         great    3
dead           dead    3
shall         shall    3
people       people    3
conceived conceived    2
new               new    2
men               men    2
war               war    2
```

```
set.seed(1234)
```

```
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

```
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col ="lightblue", main ="Most frequent words",
        ylab = "Word frequencies")
```



Question 2

2) Take a timeseries data. Show the three components of the timeseries data.

CODE and OUTPUT

```
#install.packages("forecast")
#install.packages("colortools")
library(ggplot2)
library(forecast)
library(dplyr)
library(colortools)

setwd("/Users/hemanthkumar/Desktop/Kannan 6th Sem/
CSE3020_DataViz/Lab/Experiment4")

monthly_milk <- read.csv("monthly_milk.csv")  # Milk
production per cow per month
daily_milk <- read.csv("daily_milk.csv")  # Milk production
per cow per milking

head(monthly_milk)
```

```
> head(monthly_milk)
       month milk_prod_per_cow_kg
1 1962-01-01                265.05
2 1962-02-01                252.45
3 1962-03-01                288.00
4 1962-04-01                295.20
5 1962-05-01                327.15
6 1962-06-01                313.65
```

```
head(daily_milk)
```

```
> head(daily_milk)
             date_time milk_prod_per_cow_kg
1 1975-01-01 05:00:00              11.21745
2 1975-01-01 17:00:00              10.67182
3 1975-01-02 05:00:00              10.90791
4 1975-01-02 17:00:00              11.03970
5 1975-01-03 05:00:00              12.53303
6 1975-01-03 17:00:00              10.69446
```

```
class(monthly_milk)

class(monthly_milk$month)

# Coerce to `Date` class
monthly_milk$month_date <- as.Date(monthly_milk$month, format
= "%Y-%m-%d")

# Check it worked
class(monthly_milk$month_date)

format(monthly_milk$month_date, format = "%Y-%B-%u")
class(format(monthly_milk$month_date, format = "%Y-%B-%u"))
```

```
> format(monthly_milk$month_date, format = "%Y-%B-%u")
  [1] "1962-January-1"    "1962-February-4"   "1962-March-4"     "1962-April-7"     "1962-May-2"
  [6] "1962-June-5"       "1962-July-7"       "1962-August-3"    "1962-September-6" "1962-October-1"
 [11] "1962-November-4"   "1962-December-6"   "1963-January-2"   "1963-February-5"  "1963-March-5"
 [16] "1963-April-1"      "1963-May-3"        "1963-June-6"      "1963-July-1"      "1963-August-4"
 [21] "1963-September-7"  "1963-October-2"    "1963-November-5"  "1963-December-7"  "1964-January-3"
 [26] "1964-February-6"   "1964-March-7"      "1964-April-3"     "1964-May-5"       "1964-June-1"
 [31] "1964-July-3"       "1964-August-6"     "1964-September-2" "1964-October-4"   "1964-November-7"
 [36] "1964-December-2"   "1965-January-5"    "1965-February-1"  "1965-March-1"     "1965-April-4"
 [41] "1965-May-6"        "1965-June-2"       "1965-July-4"      "1965-August-7"    "1965-September-3"
 [46] "1965-October-5"    "1965-November-1"   "1965-December-3"  "1966-January-6"   "1966-February-2"
 [51] "1966-March-2"      "1966-April-5"      "1966-May-7"       "1966-June-3"      "1966-July-5"
 [56] "1966-August-1"     "1966-September-4"  "1966-October-6"   "1966-November-2"  "1966-December-4"
 [61] "1967-January-7"    "1967-February-3"   "1967-March-3"     "1967-April-6"     "1967-May-1"
 [66] "1967-June-4"       "1967-July-6"       "1967-August-2"    "1967-September-5" "1967-October-7"
 [71] "1967-November-3"   "1967-December-5"   "1968-January-1"   "1968-February-4"  "1968-March-5"
 [76] "1968-April-1"      "1968-May-3"        "1968-June-6"      "1968-July-1"      "1968-August-4"
 [81] "1968-September-7"  "1968-October-2"    "1968-November-5"  "1968-December-7"  "1969-January-3"
 [86] "1969-February-6"   "1969-March-6"      "1969-April-2"     "1969-May-4"       "1969-June-7"
 [91] "1969-July-2"       "1969-August-5"     "1969-September-1" "1969-October-3"   "1969-November-6"
```

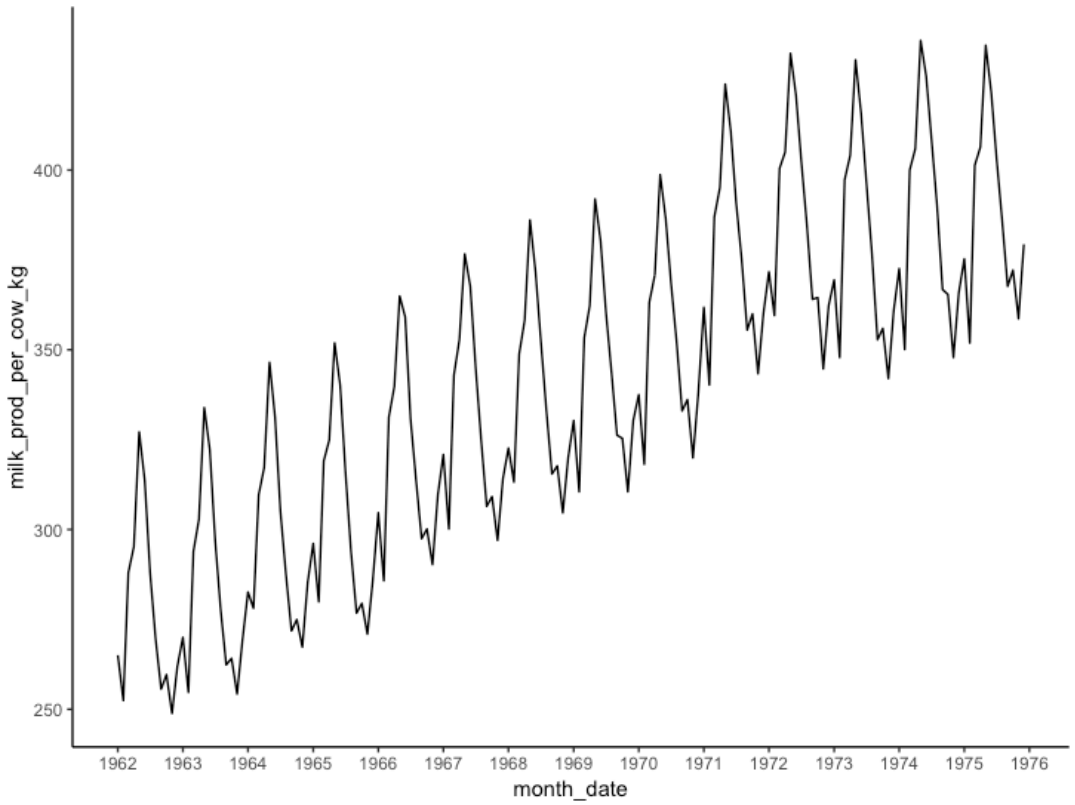
```
class(daily_milk$date_time)

daily_milk$date_time_posix <- as.POSIXct(daily_milk$date_time,
format = "%Y-%m-%d %H:%M:%S")

class(daily_milk$date_time_posix)

monthly_milk$bad_date <- format(monthly_milk$month_date,
format = "%d/%b/%Y-%u")
head(monthly_milk$bad_date)  # Awful...
class(monthly_milk$bad_date)  # Not in `Date` class

monthly_milk$good_date <- as.Date(monthly_milk$bad_date,
format = "%d/%b/%Y-%u")

head(monthly_milk$good_date)
```

```
> head(monthly_milk$bad_date)  # Awful...
[1] "01/Jan/1962-1" "01/Feb/1962-4" "01/Mar/1962-4" "01/Apr/1962-7" "01/May/1962-2" "01/Jun/1962-5"
```

```
class(monthly_milk$good_date)
```

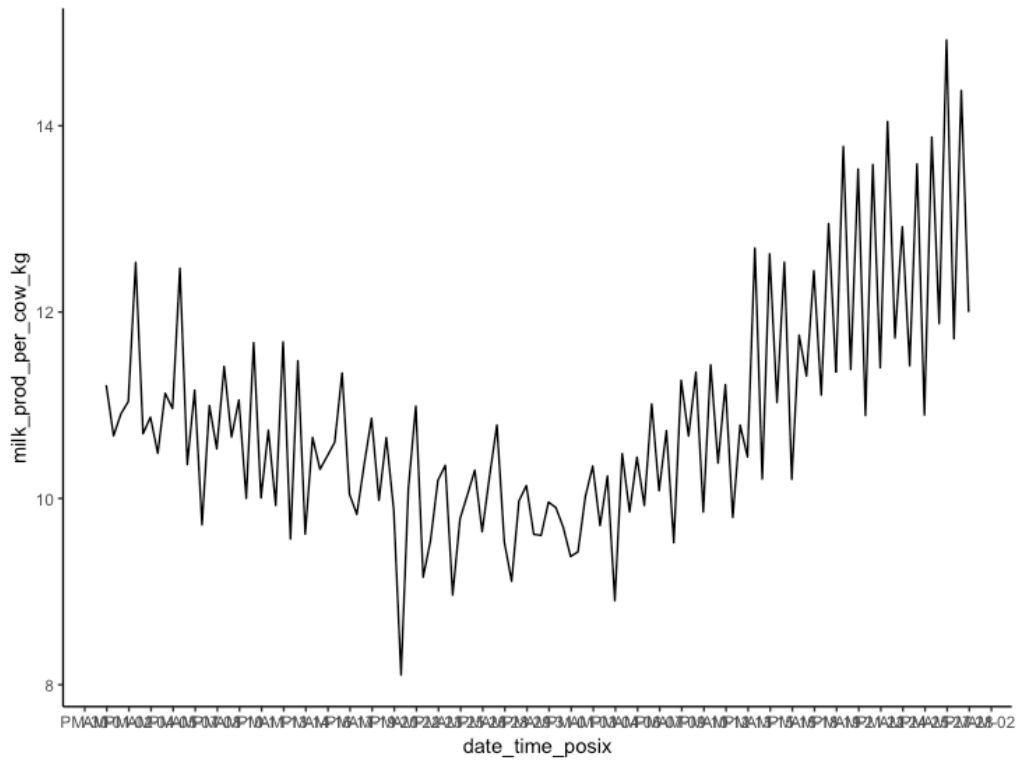
Visualising Time Series Data

```
#VISUALISING TIME SERIES DATA

(time_plot <- ggplot(monthly_milk, aes(x = month_date, y =
milk_prod_per_cow_kg)) +
    geom_line() +
    scale_x_date(date_labels = "%Y", date_breaks = "1 year") +
    theme_classic())
```




```
(time_plot_2 <- ggplot(daily_milk, aes(x = date_time_posix, y
= milk_prod_per_cow_kg)) +
    geom_line() +
    scale_x_datetime(date_labels = "%p-%d", date_breaks = "36
hour") +
    theme_classic())
```

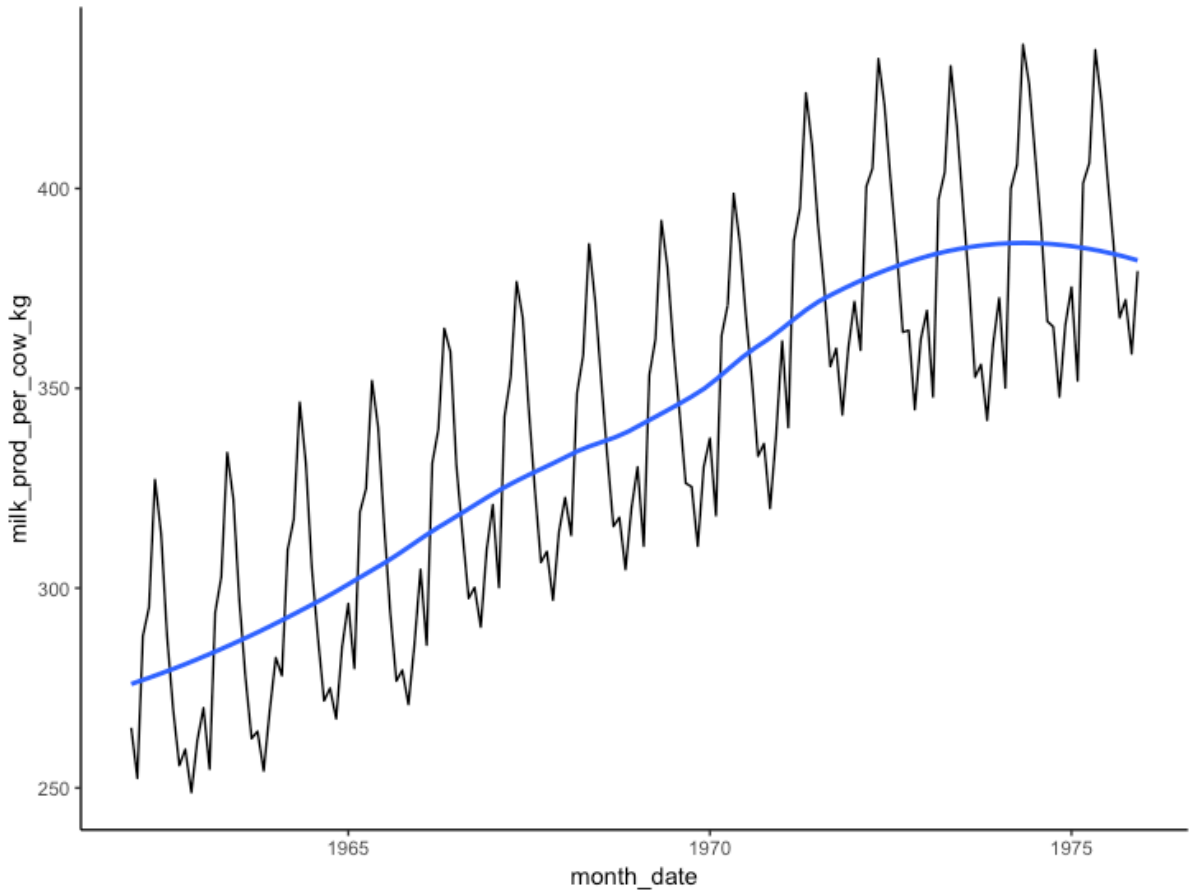
,



STATISTICAL ANALYSIS OF TIME SERIES DATA

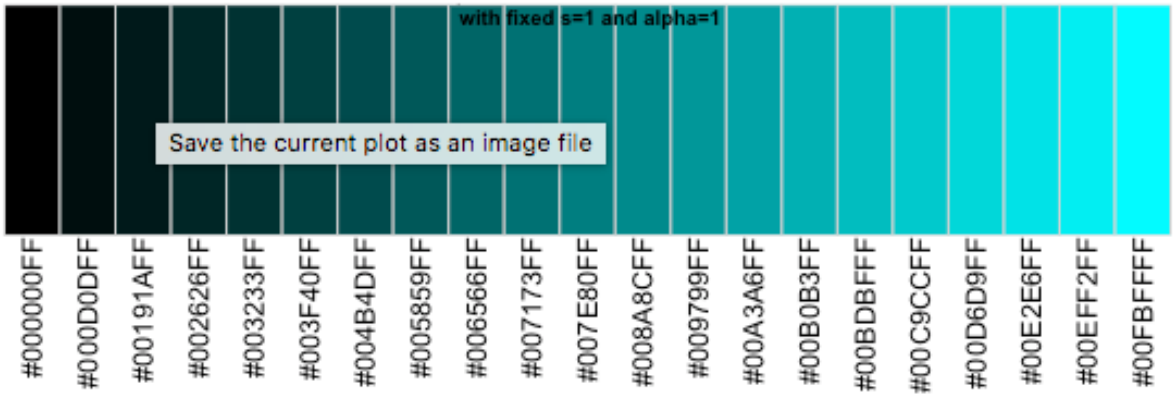
1. Decomposition

```
#STATISTICAL ANALYSIS OF TIME SERIES DATA
#1. Decomposition

(decomp_2 <- ggplot(monthly_milk, aes(x = month_date, y =
milk_prod_per_cow_kg)) +
    geom_line() +
    geom_smooth(method = "loess", se = FALSE, span = 0.6) +
    theme_classic())
```

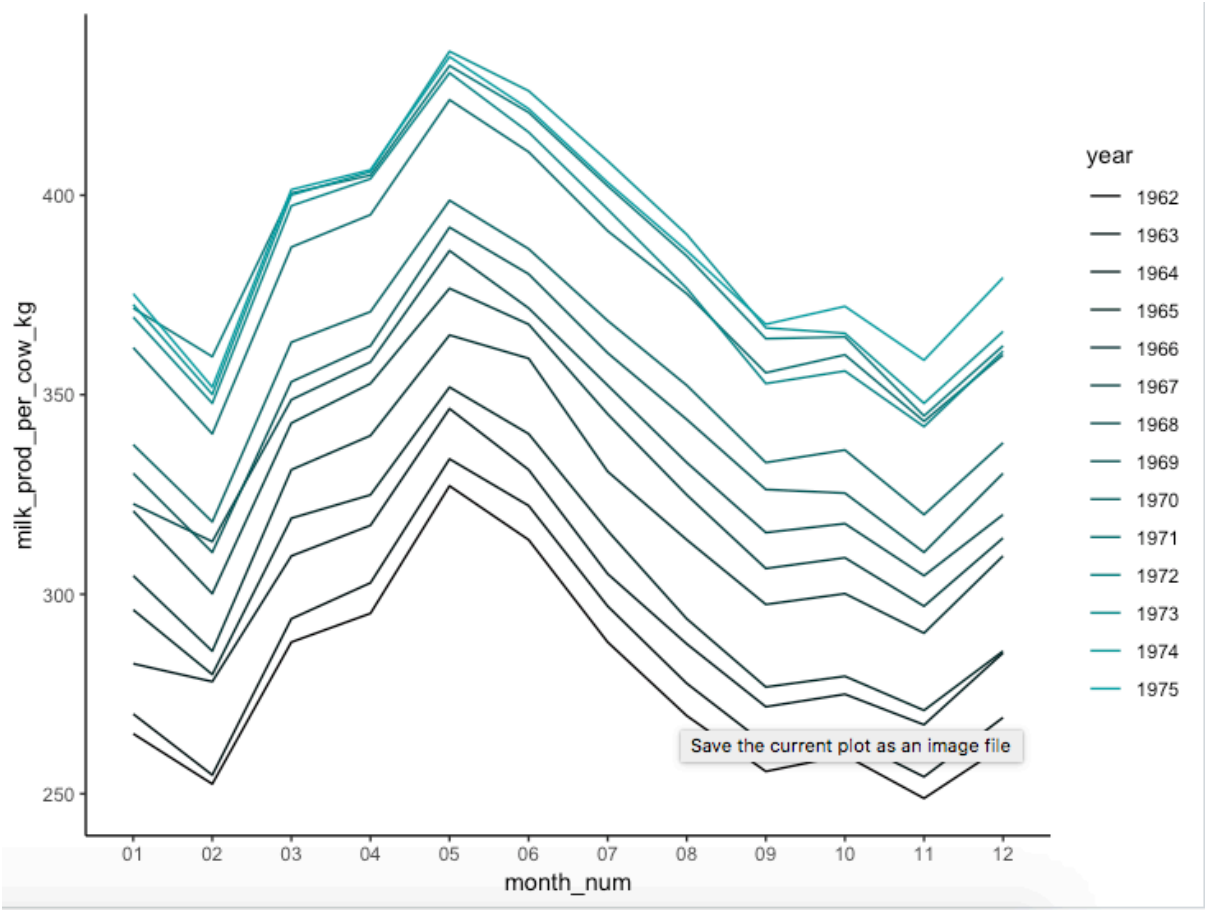



```
# Extract month and year and store in separate columns
monthly_milk$year <- format(monthly_milk$month_date, format =
"%Y")
monthly_milk$month_num <- format(monthly_milk$month_date,
format = "%m")

# Create a colour palette using the `colortools` package
```

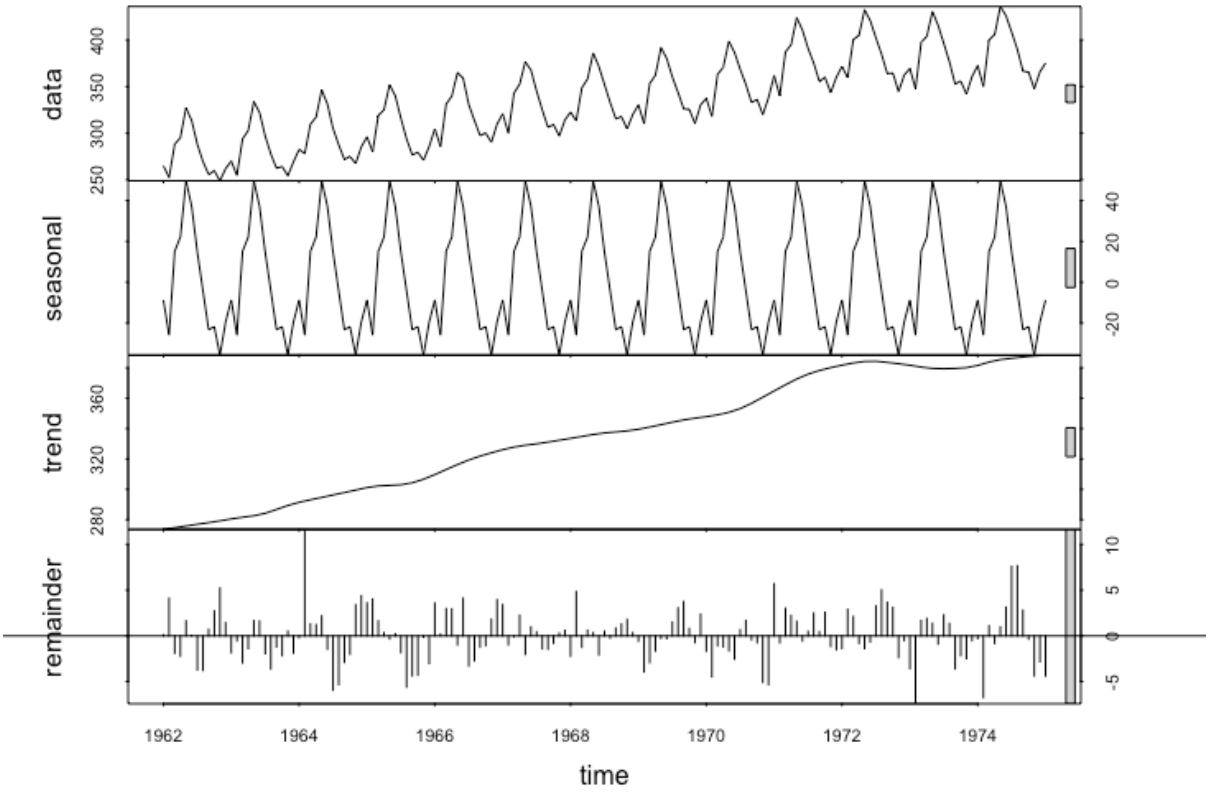
```
year_pal <- sequential(color = "darkturquoise", percentage =
5, what = "value")

# Make the plot
(seasonal <- ggplot(monthly_milk, aes(x = month_num, y =
milk_prod_per_cow_kg, group = year)) +
    geom_line(aes(colour = year)) +
    theme_classic() +
    scale_color_manual(values = year_pal))
```



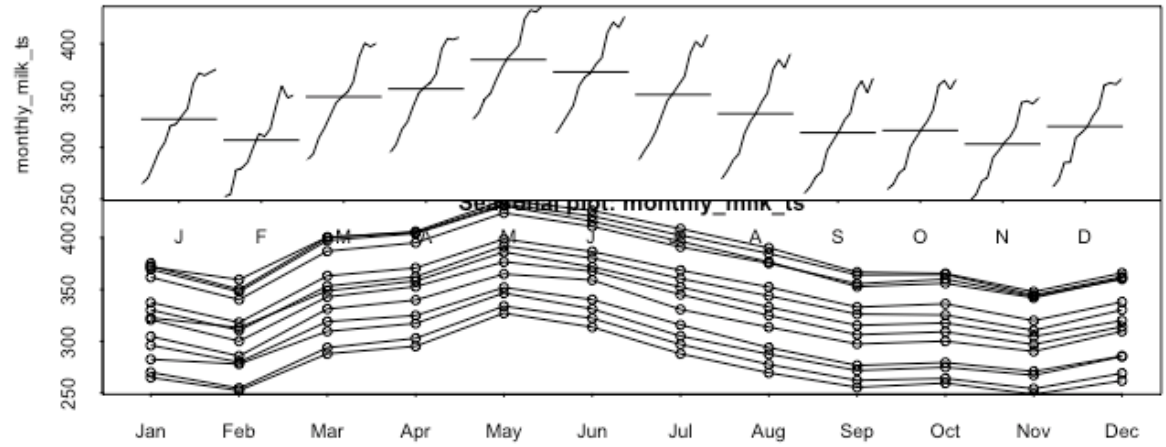

```
# Transform to `ts` class
monthly_milk_ts <- ts(monthly_milk$milk_prod, start = 1962,
end = 1975, freq = 12)  # Specify start and end year,
measurement frequency (monthly = 12)

# Decompose using `stl()`
monthly_milk_stl <- stl(monthly_milk_ts, s.window = "period")

# Generate plots
plot(monthly_milk_stl)  # top=original data, second=estimated
seasonal, third=estimated smooth trend, bottom=estimated
irregular element i.e. unaccounted for variation
```




```
monthplot(monthly_milk_ts, choice = "seasonal")  # variation in milk production for each month
```



```
seasonplot(monthly_milk_ts)
```

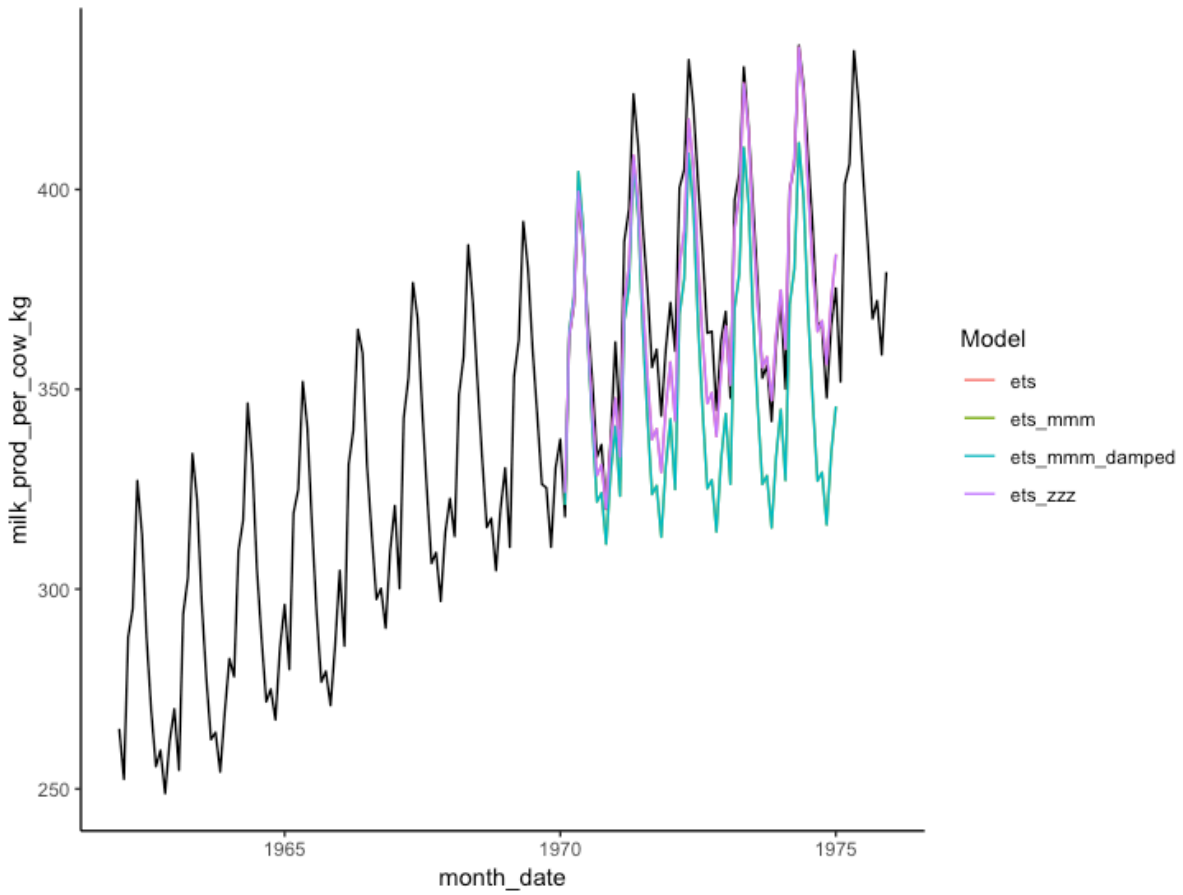


2. Forecasting

```
#2. Forecasting

monthly_milk_model <- window(x = monthly_milk_ts, start =
c(1962), end = c(1970))
monthly_milk_test <- window(x = monthly_milk_ts, start =
c(1970))

# Creating model objects of each type of ets model
milk_ets_auto <- ets(monthly_milk_model)
milk_ets_mmm <- ets(monthly_milk_model, model = "MMM")
milk_ets_zzz<- ets(monthly_milk_model, model = "ZZZ")
milk_ets_mmm_damped <- ets(monthly_milk_model, model = "MMM",
damped = TRUE)

# Creating forecast objects from the model objects
milk_ets_fc <- forecast(milk_ets_auto, h = 60)  # `h = 60`
means that the forecast will be 60 time periods long, in our
case a time period is one month
milk_ets_mmm_fc <- forecast(milk_ets_mmm, h = 60)
milk_ets_zzz_fc <- forecast(milk_ets_zzz, h = 60)
milk_ets_mmm_damped_fc <- forecast(milk_ets_mmm_damped, h =
60)

# Convert forecasts to data frames
```

```
milk_ets_fc_df <- cbind("Month" =
rownames(as.data.frame(milk_ets_fc)),
as.data.frame(milk_ets_fc))  # Creating a data frame
names(milk_ets_fc_df) <- gsub(" ", "_", names(milk_ets_fc_df))
# Removing whitespace from column names
milk_ets_fc_df$Date <- as.Date(paste("01-",
milk_ets_fc_df$Month, sep = ""), format = "%d-%b %Y")  #
prepending day of month to date
milk_ets_fc_df$Model <- rep("ets")  # Adding column of model
type

milk_ets_mmm_fc_df <- cbind("Month" =
rownames(as.data.frame(milk_ets_mmm_fc)),
as.data.frame(milk_ets_mmm_fc))
names(milk_ets_mmm_fc_df) <- gsub(" ", "_",
names(milk_ets_mmm_fc_df))
milk_ets_mmm_fc_df$Date <- as.Date(paste("01-",
milk_ets_mmm_fc_df$Month, sep = ""), format = "%d-%b %Y")
milk_ets_mmm_fc_df$Model <- rep("ets_mmm")

milk_ets_zzz_fc_df <- cbind("Month" =
rownames(as.data.frame(milk_ets_zzz_fc)),
as.data.frame(milk_ets_zzz_fc))
names(milk_ets_zzz_fc_df) <- gsub(" ", "_",
names(milk_ets_zzz_fc_df))
milk_ets_zzz_fc_df$Date <- as.Date(paste("01-",
milk_ets_zzz_fc_df$Month, sep = ""), format = "%d-%b %Y")
milk_ets_zzz_fc_df$Model <- rep("ets_zzz")

milk_ets_mmm_damped_fc_df <- cbind("Month" =
rownames(as.data.frame(milk_ets_mmm_damped_fc)),
as.data.frame(milk_ets_mmm_damped_fc))
names(milk_ets_mmm_damped_fc_df) <- gsub(" ", "_",
names(milk_ets_mmm_damped_fc_df))
milk_ets_mmm_damped_fc_df$Date <- as.Date(paste("01-",
milk_ets_mmm_damped_fc_df$Month, sep = ""), format = "%d-%b
%Y")
milk_ets_mmm_damped_fc_df$Model <- rep("ets_mmm_damped")

# Combining into one data frame
forecast_all <- rbind(milk_ets_fc_df, milk_ets_mmm_fc_df,
milk_ets_zzz_fc_df, milk_ets_mmm_damped_fc_df)
```

```
# Plotting with ggplot
(forecast_plot <- ggplot() +
    geom_line(data = monthly_milk, aes(x = month_date, y =
milk_prod_per_cow_kg)) +  # Plotting original data
    geom_line(data = forecast_all, aes(x = Date, y =
Point_Forecast, colour = Model)) +  # Plotting model forecasts
    theme_classic())
```




```
accuracy(milk_ets_fc, monthly_milk_test)
```

```
> accuracy(milk_ets_fc, monthly_milk_test)
                    ME      RMSE      MAE         MPE     MAPE      MASE        ACF1 Theil's U
Training set 0.01864308  2.726367 2.094068 0.001162836 0.675237 0.2190178 0.006994282        NA
Test set     6.49724181 10.870286 8.643747 1.687599938 2.292374 0.9040464 0.816443289 0.4875223
```

```
accuracy(milk_ets_mmm_fc, monthly_milk_test)
```

```
> accuracy(milk_ets_mmm_fc, monthly_milk_test)
                    ME      RMSE      MAE       MPE      MAPE      MASE       ACF1 Theil's U
Training set  0.3733287  3.108712  2.3070 0.1120838 0.7331154 0.2412883 0.01037896        NA
Test set     23.9447361 26.571021 24.4991 6.3609498 6.5099209 2.5623519 0.87773301  1.201562
```

```
accuracy(milk_ets_zzz_fc, monthly_milk_test)
```

```
> accuracy(milk_ets_zzz_fc, monthly_milk_test)
                    ME      RMSE      MAE         MPE     MAPE      MASE        ACF1 Theil's U
Training set 0.01864308  2.726367 2.094068 0.001162836 0.675237 0.2190178 0.006994282        NA
Test set     6.49724181 10.870286 8.643747 1.687599938 2.292374 0.9040464 0.816443289 0.4875223
```

```
accuracy(milk_ets_mmm_damped_fc, monthly_milk_test)
```

```
> accuracy(milk_ets_mmm_damped_fc, monthly_milk_test)
                    ME      RMSE      MAE       MPE      MAPE      MASE       ACF1 Theil's U
Training set  0.3733287  3.108712  2.3070 0.1120838 0.7331154 0.2412883 0.01037896        NA
Test set     23.9447361 26.571021 24.4991 6.3609498 6.5099209 2.5623519 0.87773301  1.201562
```

3. Extracting Values from Forecast

```
#3. Extracting Values from Forecast

milk_ets_fc_df %>%
  filter(Month == "Jan 1975") %>%
  select(Month, Point_Forecast)
```

```
> milk_ets_fc_df %>%
+   filter(Month == "Jan 1975") %>%
+   select(Month, Point_Forecast)
     Month Point_Forecast
1 Jan 1975       383.7897
```

```
milk_ets_zzz_fc_df %>%
  filter(Month == "Jan 1975") %>%
  select(Month, Point_Forecast)
```

```
> milk_ets_zzz_fc_df %>%
+   filter(Month == "Jan 1975") %>%
+   select(Month, Point_Forecast)
     Month Point_Forecast
1 Jan 1975       383.7897
```