Averis AI/Data Science Assesment Solution

Contents

Note

This solution was developed using the following version of R:

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          3
## minor          6.1
## year           2019
## month          07
## day            05
## svn rev        76782
## language       R
## version.string R version 3.6.1 (2019-07-05)
## nickname       Action of the Toes
```

Libraries

Uncomment this cell to install packages:

```
# install.packages("tidyverse")
# install.packages("moments")
# install.packages("dbscan")
# install.packages("forecast")
# install.packages("lmtest")
# install.packages("tidytext")
```

Question 1

Question

1. A customer informed their consultant that they have developed several formulations of petrol that gives different characteristics of burning pattern. The formulations are obtaining by adding varying levels of additives that, for example, prevent engine knocking, gum prevention, stability in storage, and etc. However, a third party certification organisation would like to verify if the formulations are significantly different, and request for both physical and statistical proof. Since the formulations are confidential information, they are not named in the dataset.

   Please assist the consultant in the area of statistical analysis by doing this;

   a. A descriptive analysis of the additives (columns named as "a" to "i"), which must include summaries of findings (parametric/non-parametric). Correlation and ANOVA, if applicable, is a must.

   b. A graphical analysis of the additives, including a distribution study.

   c. A clustering test of your choice (unsupervised learning), to determine the distinctive number of formulations present in the dataset.

(refer attachment : ingredients.csv)

Solution

Read the ingredients.csv file:

```
df <- read_csv("../data/ingredient.csv", col_types="ddddddddd")

head(df)
```

Solution to Q1a

Some desriptive statistics:

```
df %>%
  gather(additive, measure) %>%
  group_by(additive) %>%
  summarize(mean = mean(measure),
            median = median(measure),
            `standard deviation` = sd(measure),
            min = min(measure),
            max=max(measure),
            skewness = skewness(measure),
```

```
            kurtosis = kurtosis(measure),
            count = n())
```

One-way ANOVA:

```
df %>%
  gather(additive, measure) %>%
  aov(measure ~ additive, .) %>%
  summary
```

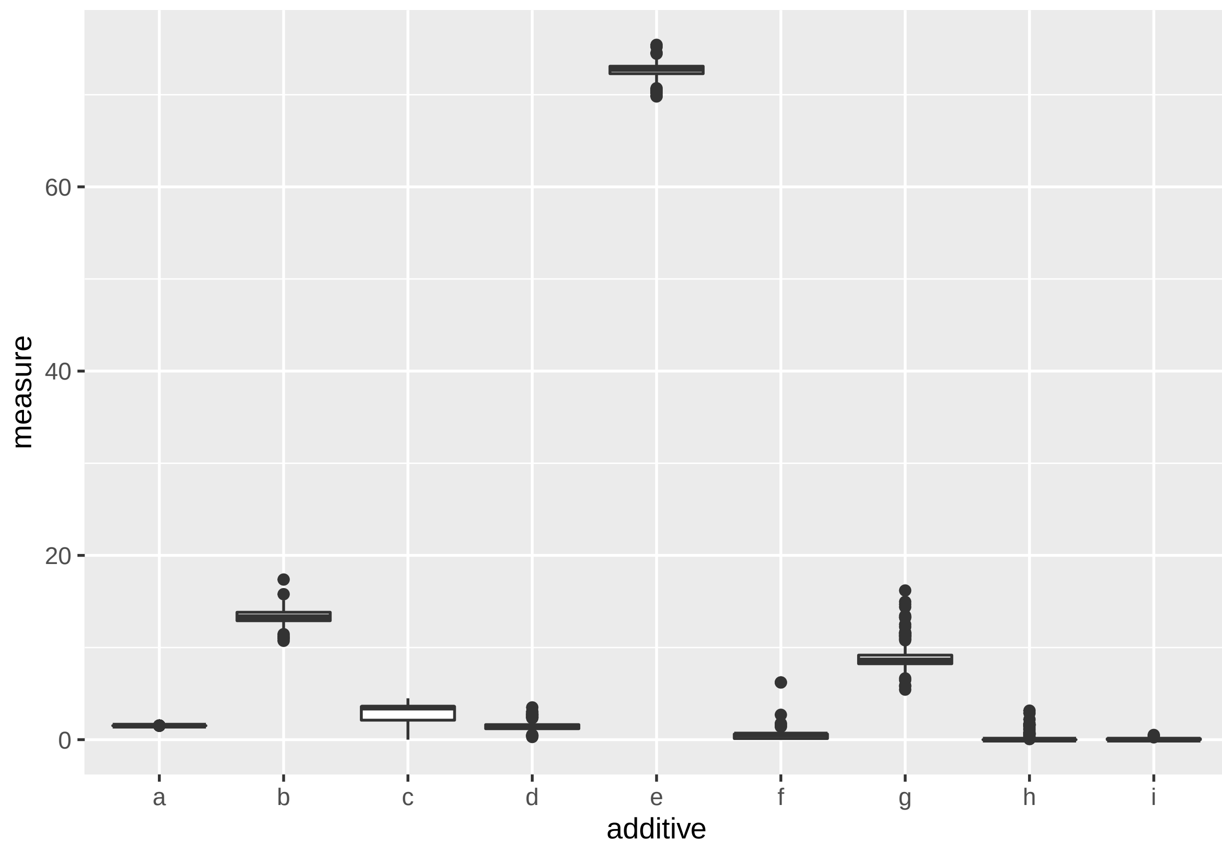
```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## additive       8 943261  117908  168332 <2e-16 ***
## Residuals   1917   1343       1                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The descriptive statistics show that the measures of each additive are very skewed and have very heavy tails relative to the normal distribution. Additive is stands out the most as the mean value of its measures is very different compared to the other additives. This is consistent with the results from the One-way ANOVA test.

Solution to Q1b

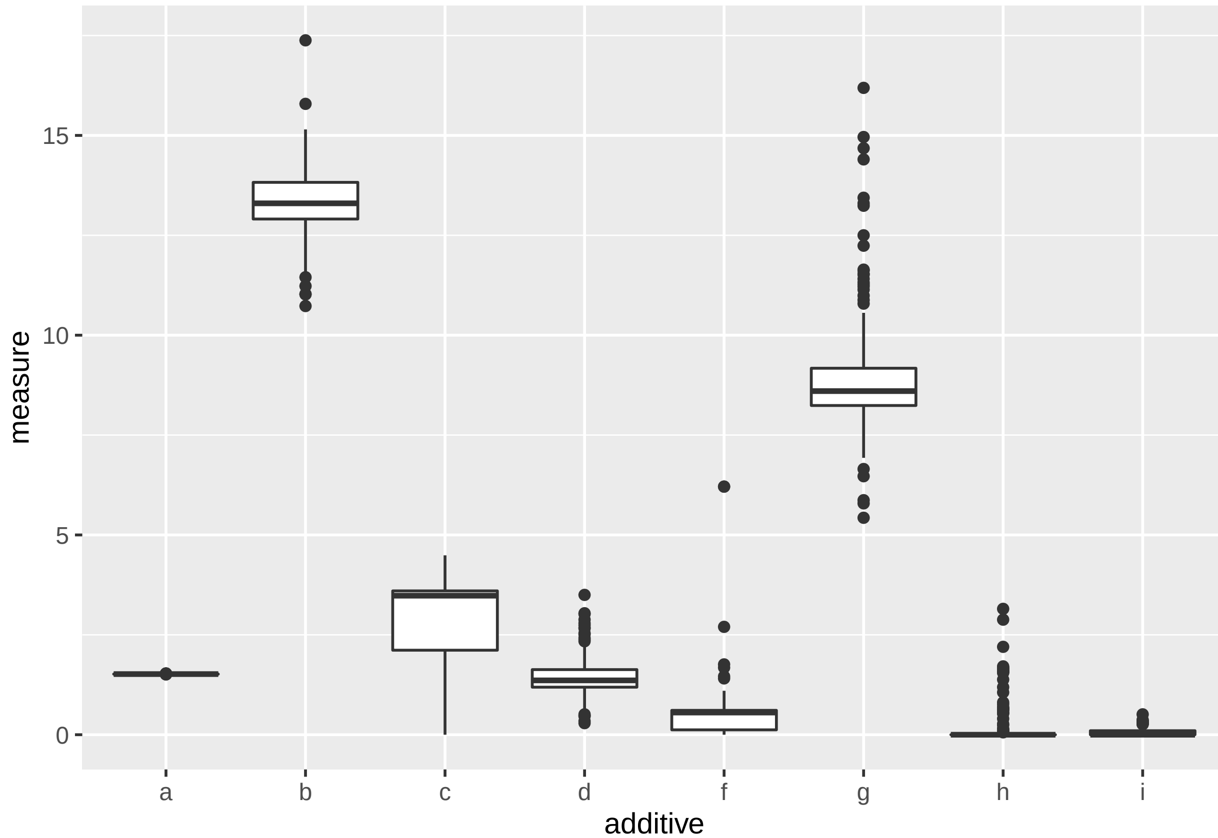
Boxplot of each additive:

```
df %>%
  gather(additive, measure) %>%
  ggplot(aes(x=additive, y=measure)) +
  geom_boxplot()
```



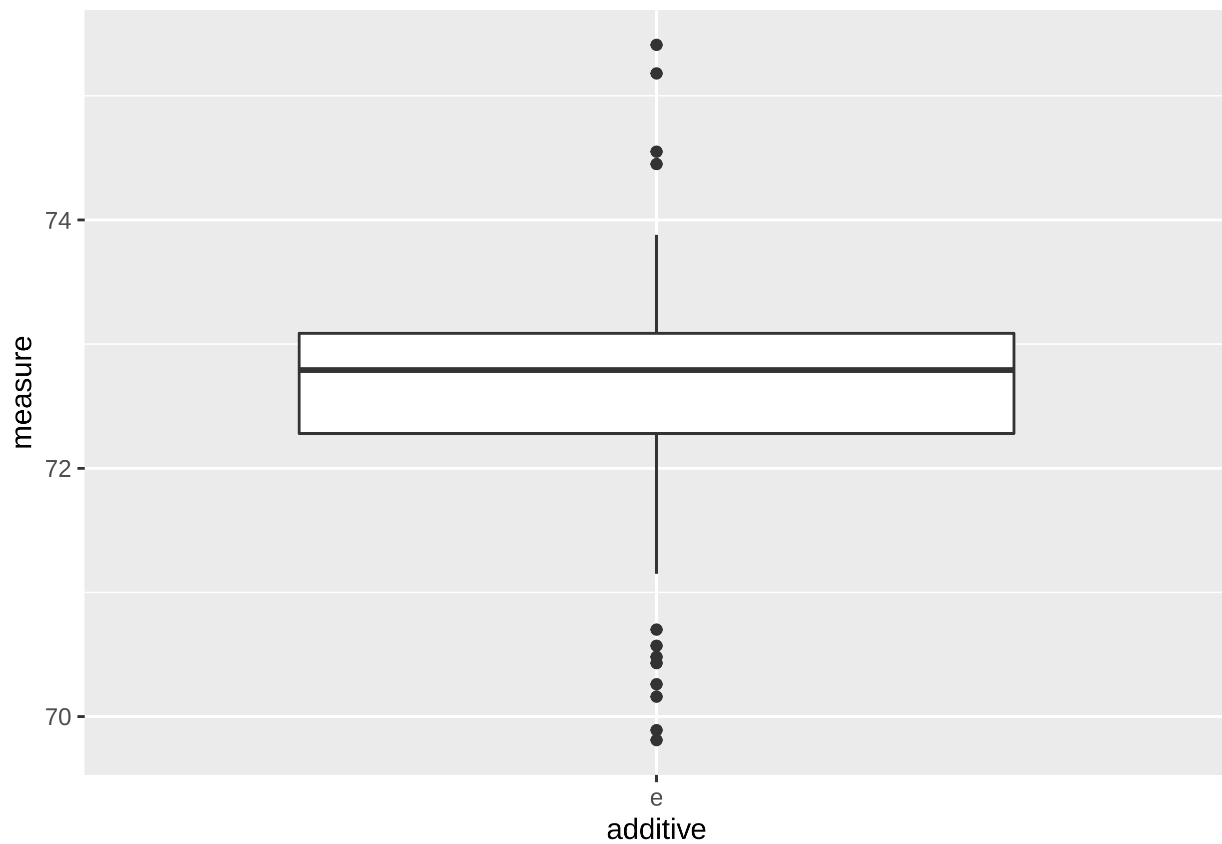
The same boxplot but exculding additive “e”:

```
df %>%
  select(-e) %>%
  gather(additive, measure) %>%
  ggplot(aes(x=additive, y=measure)) +
  geom_boxplot()
```



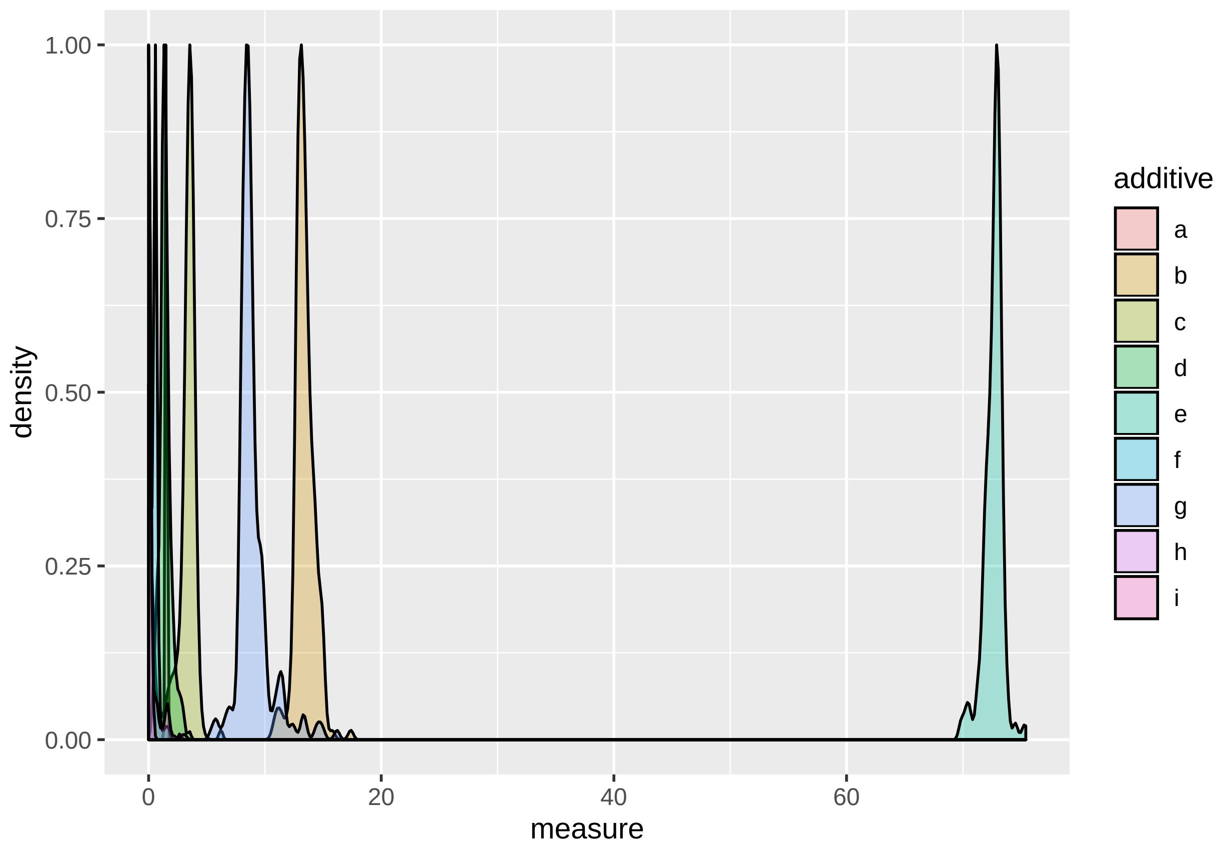
Boxplot of just additive "e":

```
df %>%
  select(e) %>%
  gather(additive, measure) %>%
  ggplot(aes(x=additive, y=measure)) +
  geom_boxplot()
```



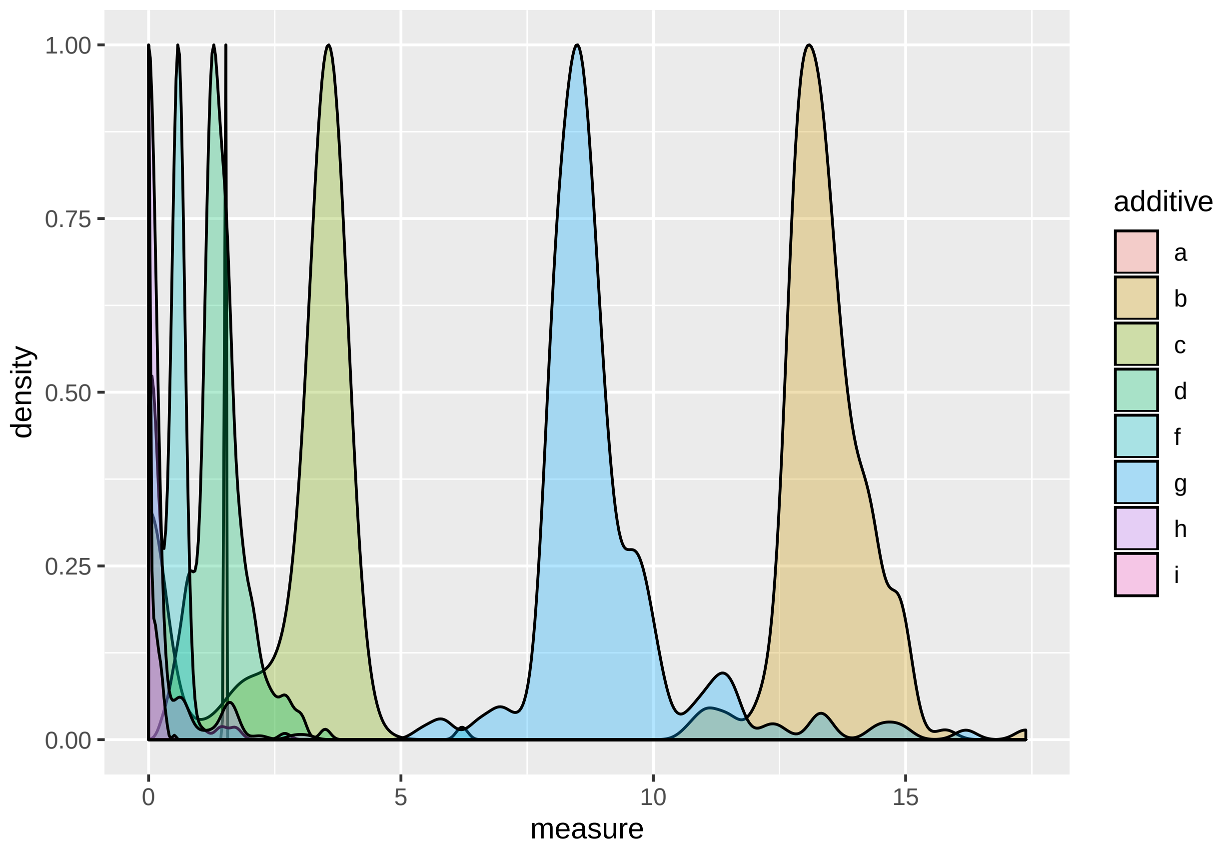
A density plot of all the additives:

```
df %>%
  gather(additive, measure) %>%
  ggplot(aes(x=measure, y=..scaled.., group=additive, fill=additive)) +
  geom_density(alpha=0.3) +
  ylab("density")
```



The same density plot but excluding additive “e”:

```
df %>%
  select(-e) %>%
  gather(additive, measure) %>%
  ggplot(aes(x=measure, y=..scaled.., group=additive, fill=additive)) +
  geom_density(alpha=0.3) +
  ylab("density")
```



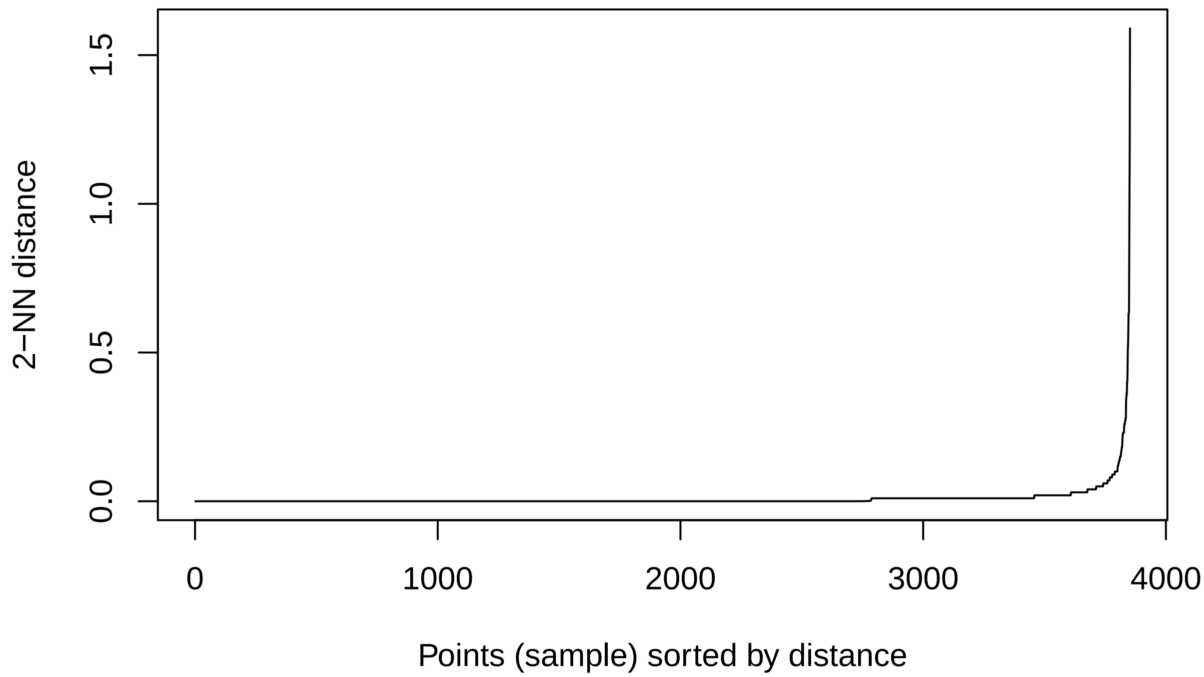
The graphical analysis gives the same findings as with the descriptive analysis in the previous answer.

Solution to Q1c

We will use dbscan to determine the number of clusters.

Find out a suitable value for the `eps` parameter using the k-NN plot for k = dim + 1:

```
df %>%
  gather(additive, measure) %>%
  select(measure) %>%
  kNNdistplot(k=2)
```



We will set the value for `eps` to 0.2.

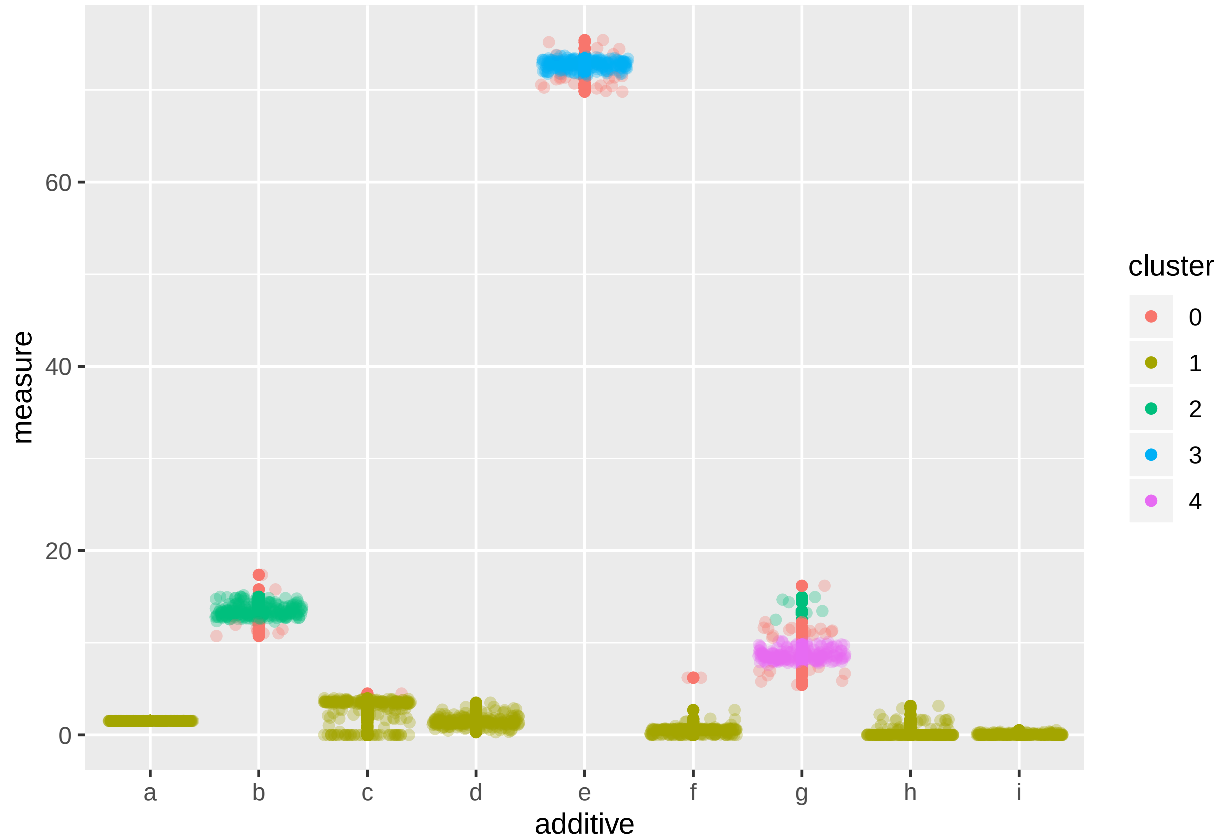
```
cl <- df %>%
  gather(additive, measure) %>%
  select(measure) %>%
  dbscan(eps=0.20, minPts = 10)

cl
```

```
## DBSCAN clustering for 1926 objects.
## Parameters: eps = 0.2, minPts = 10
## The clustering contains 4 cluster(s) and 59 noise points.
##
##    0    1    2    3    4
##   59 1281  210  193  183
##
## Available fields: cluster, eps, minPts
```

Visualize the clusters:

```
df %>%
  gather(additive, measure) %>%
  mutate(cluster = cl$cluster) %>%
  mutate(cluster = factor(cluster)) %>%
  ggplot(aes(x = additive, y = measure, color = cluster, group = cluster)) +
  geom_point() +
  geom_jitter(alpha = 0.3)
```



This clustering technique found 4 distinct clusters (cluster 0 is noise). This is consistent with the graphical analysis in the previous answer.

Question 2

Question

2. A team of plantation planners are concerned about the yield of oil palm trees, which seems to fluctuate. They have collected a set of data and needed help in analysing on how external factors influence fresh fruit bunch (FFB) yield. Some experts are of opinion that the flowering of oil palm tree determines the FFB yield, and are linked to the external factors. Perform the analysis, which requires some study on the background of oil palm tree physiology.

   (refer attachment palm_ffb.csv)

Solution

Solution to Q2

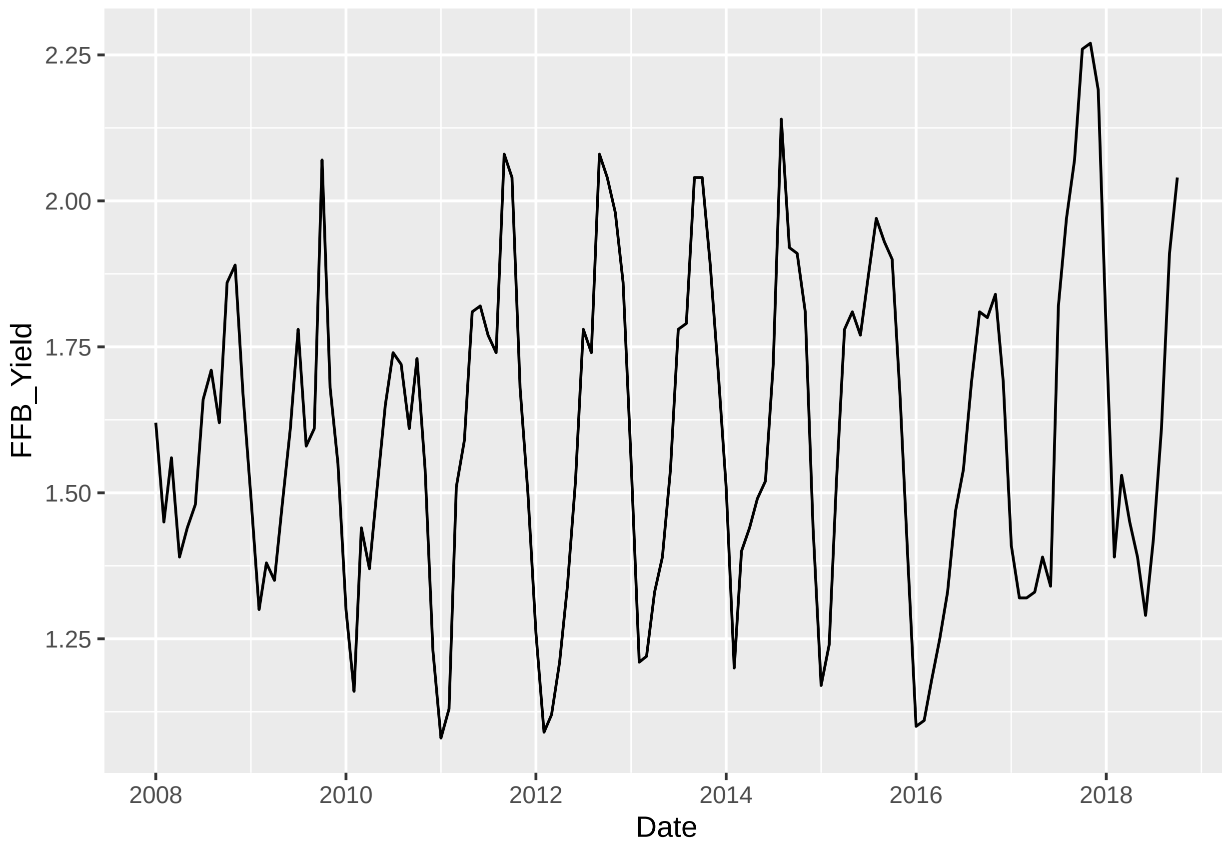
Read the data:

```r
df <- read_csv("../data/palm_ffb.csv", col_types="cdddddddd") %>%
  mutate(Date = dmy(Date))

head(df)
```

Plot the `FFB_Yield` trend:

```r
df %>%
  ggplot(aes(x = Date, y = FFB_Yield)) +
  geom_line()
```



Clearly, there is a seasonal component.

Fit an ARIMA with linear regression model:

```r
y <- ts(df$FFB_Yield, start = c(2008, 1), end = c(2018, 10), frequency = 12)


# trial and error results in the following set of variables giving the best model in terms of AIC
x <- df %>%
  select(-Date, -FFB_Yield, -HA_Harvested, -SoilMoisture, -Precipitation, -Min_Temp, -Max_Temp) %>%
  as.matrix()

model <- auto.arima(y, xreg = x)
summary(model)
```

```
## Series: y
## Regression with ARIMA(1,0,0)(2,1,1)[12] errors
```

```
##
## Coefficients:
##          ar1    sar1     sar2     sma1  Average_Temp  Working_days
##       0.7166  0.2472  -0.1903  -0.8459       -0.0727        0.0167
## s.e.  0.0645  0.1459   0.1213   0.2010        0.0288        0.0091
##
## sigma^2 estimated as 0.01554:  log likelihood=73.84
## AIC=-133.67   AICc=-132.65   BIC=-114.28
##
## Training set error measures:
##                       ME      RMSE        MAE        MPE     MAPE
## Training set 0.002452675 0.1157051 0.08461976 -0.3755017 5.307308
##                   MASE       ACF1
## Training set 0.4911526 0.07402313
```

Test the model's coefficients for statistical significance:

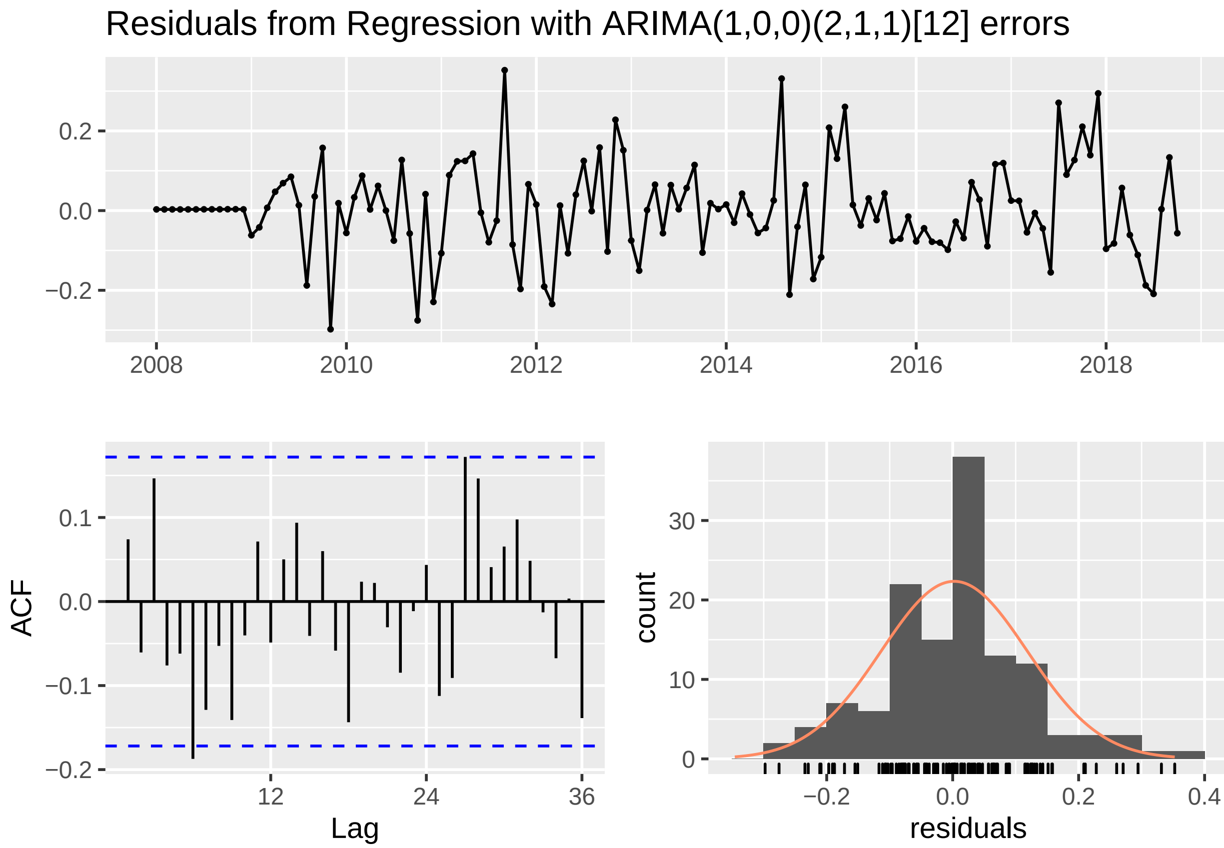
```r
coeftest(model)
```

```
##
## z test of coefficients:
##
##                Estimate Std. Error z value  Pr(>|z|)
## ar1           0.7165822  0.0645009 11.1096 < 2.2e-16 ***
## sar1          0.2471600  0.1459137  1.6939   0.09029 .
## sar2         -0.1902972  0.1213430 -1.5683   0.11682
## sma1         -0.8459350  0.2009671 -4.2093 2.561e-05 ***
## Average_Temp -0.0727380  0.0288473 -2.5215   0.01169 *
## Working_days  0.0166702  0.0091392  1.8240   0.06815 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All coefficients are significant at the 5% level.

Check the residuals for autocorrelation:

```r
checkresiduals(model)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,0,0)(2,1,1)[12] errors
## Q* = 25.047, df = 18, p-value = 0.1236
##
## Model df: 6.   Total lags used: 24
```

The ACF plot shows only 1 significant spike at lag 6 but the Ljung-Box test does not lead us to conclude (at the 5% significance level) that the residuals are not independently distributed. Also, the other plots show that the residuals look like white noise. Therefore, we can conclude that the residuals are independent.

Conclusion:

Given the provided data, the analysis shows that most of the variation in the FFB yield can be explained by the timeseries itself. FFB yield has a 12 month cycle and it's value at month n is a function of the value at month $n-1$. Accounting for this effect, `Working_days` has a positive effect on FFB yield (increasing FFB yield by 0.02 per additional day) while `Average_Temp` has a negative effect on FFB yield (decreasing FFB yield by 0.07 per additional 1 Celcius) holding all other factors constant.

Question 3

Question

3. Feed the following paragraph into your favourite data analytics tool, and answer the following;

    a. What is the probability of the word "data" occurring in each line ?

    b. What is the distribution of distinct word counts across all the lines ?

    c. What is the probability of the word "analytics" occurring after the word "data" ?

===

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data -- with the difference that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category. Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals -- all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources. At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis. Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

Solution

Read the data:

```
df <- read_lines("../data/q3_paragraph.txt") %>%
  enframe(name="line", value="text")
```

```
df
```

Calculate total lines:

```
n_lines = nrow(df)
```

Solution to Q3a

Calculate probabiliy of the word "data" appearing in a line:

```
prob_data <- df %>%
  mutate(has_data = str_detect(text, "data")) %>%
  mutate(has_data = as.numeric(has_data)) %>%
  pull(has_data) %>%
  mean
```

The probability of the word "data" occuring in each line is:

```
# assuming independence
prob_data^n_lines
```

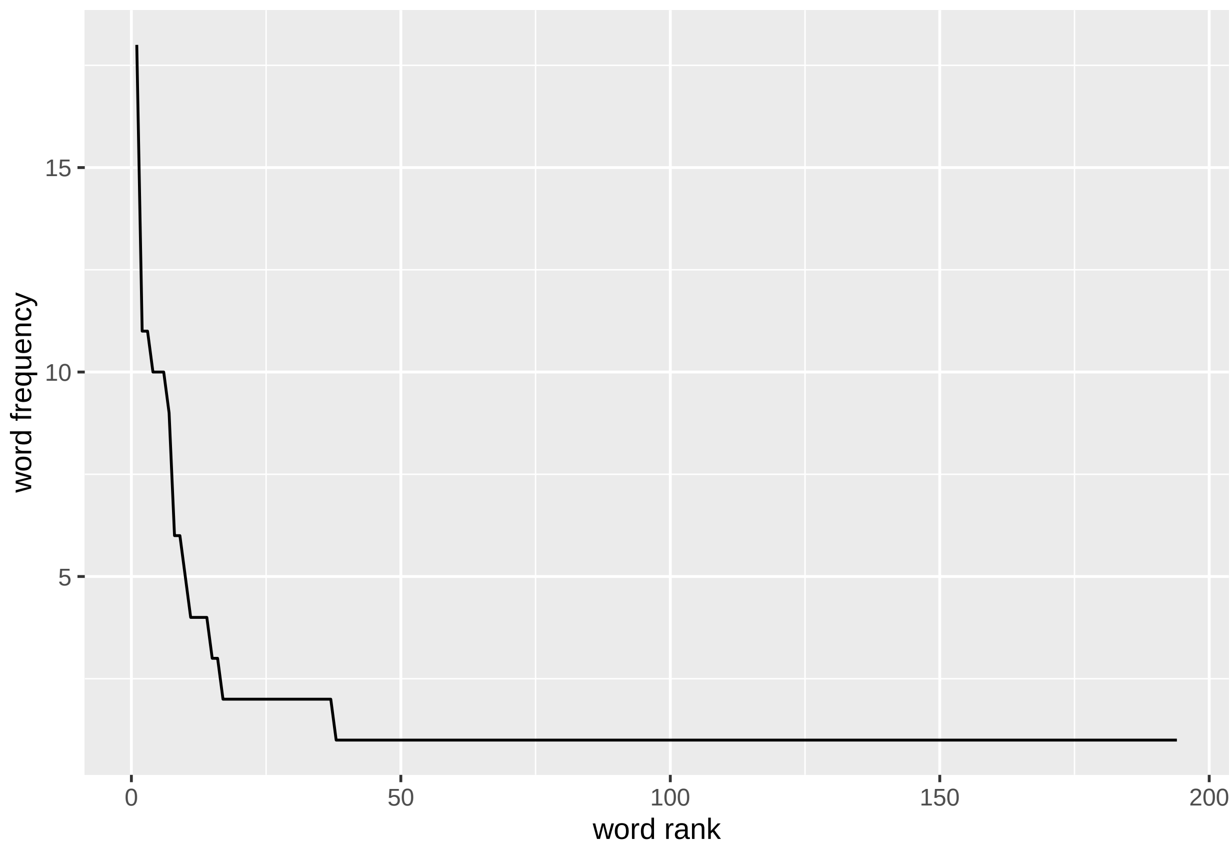
```
## [1] 1.61692e-06
```

Solution to Q3b

```

word_count <- df %>%
  unnest_tokens("word", "text") %>%
  count(word, sort=TRUE) %>%
  rowid_to_column("rank")

word_count %>%
  ggplot(aes(x = rank, y = n)) +
  geom_line() +
  xlab("word rank") +
  ylab("word frequency")
```



The distribution of distinct word counts across all the lines follows a approximately a power law.

Solution to Q3c

```
# put the paragraph into a single line
para <- df %>%
  unnest_tokens("word", "text") %>%
  group_by("word") %>%
  summarize(text = str_c(word, collapse = " ")) %>%
  select("text")

# check that last word is not data
last_word <- para %>%
  pull(text) %>%
  str_extract("\\b\\w+$")

stopifnot(last_word != "data")

bigrams <- para %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)

freq_data_and_analytics <- bigrams %>%
  filter(bigram == "data analytics") %>%
  count %>%
  pull(n)
```

```
freq_data <- bigrams %>%
  filter(str_detect(bigram, "^data")) %>%
  count %>%
  pull(n)
```

The probability of the word “analytics” occuring after the word “data” is:

```
freq_data_and_analytics/freq_data
```

```
## [1] 0.3333333
```