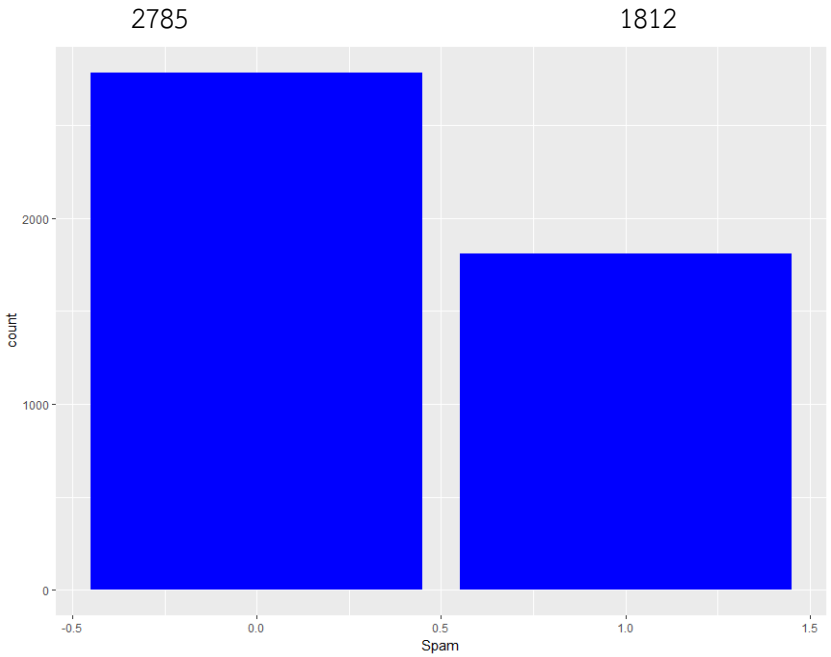
Spambase

ชุดข้อมูลนี้เกี่ยวกับอีเมล์ปกติและอีเมล์โฆษณา

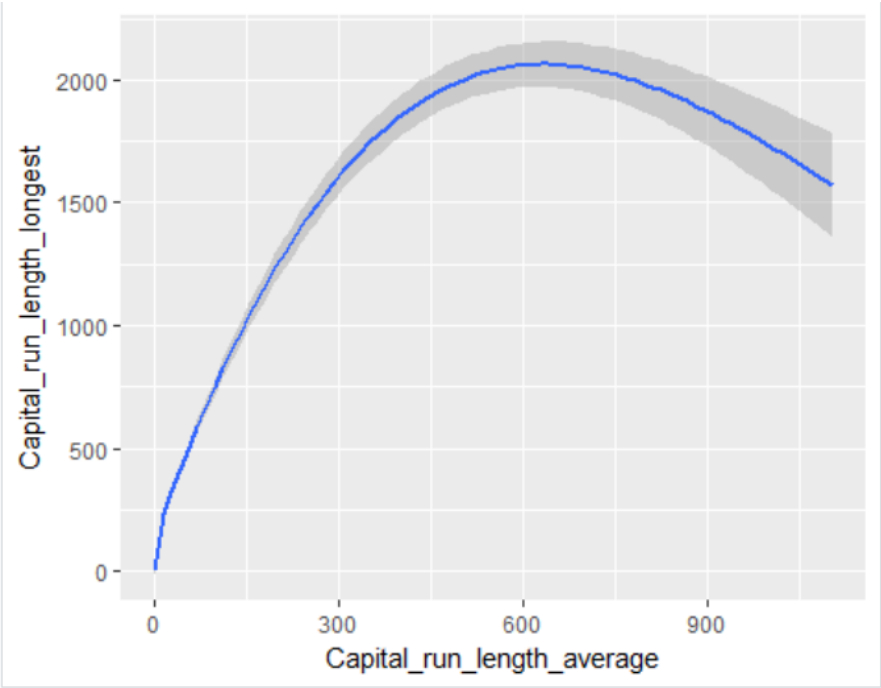
Bar graph : Email VS Spam



สรุปได้ว่า การเปรียบเทียบโดยใช้การกราฟแท่งสรุปได้ว่าจำนวนอีเมล์ปกติมีมากว่าอีเมล์ที่เป็นแสปม

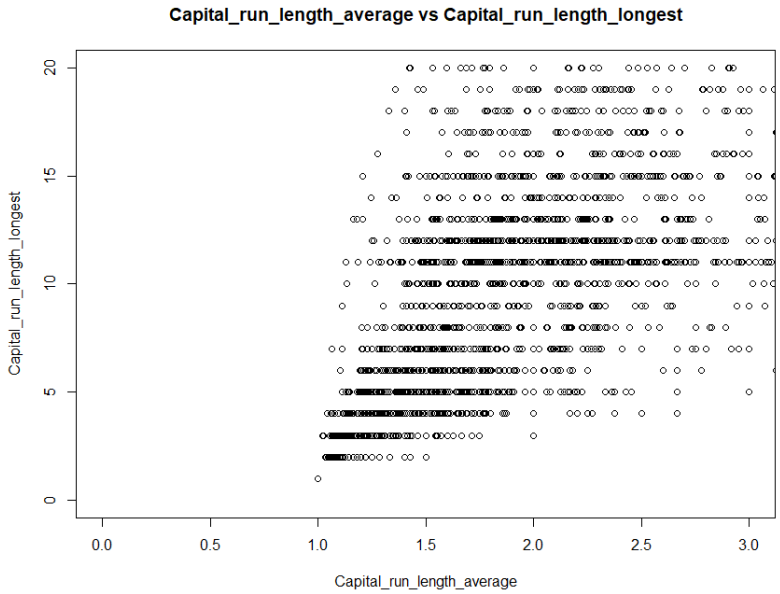
(แสปม เช่นอีเมล์โฆษณา)

Trend graph



สรุปได้ว่า ยิ่งค่าเฉลี่ยความยาวของตัวอักษรตัวมีค่ามาก ความยาวของลำดับที่ยาวที่สุดของตัวอักษรยิ่งมีค่ามาก

Scatter plot



สรุปได้ว่า มีแนวไปทางบวกระหว่างตัวแปรการค่าเฉลี่ยความยาวตัวอักษร จากทางซ้ายไปขวา ขณะที่ความยาวของลำดับที่ยาวที่สุดของตัวอักษรมีแนวโน้มเพิ่มขึ้น

คำนวณ

เปอร์เซ็นข้อมูลแสปมอยู่ที่ 39.42%
เปอร์เซ็นข้อมูลอีเมล์อยู่ที่ 60.58%
แสดงให้เห็นว่าข้อมูลชุดนี้มีอีเมล์ปกติมากกว่าอีเมล์ที่เป็นแสปมจากคอลัมน์ Spam

ค่าเฉลี่ยความถี่ที่เกิดคำในอีเมล์แสปมทั้งหมดมี 8.761 มัธยฐานอยู่ที่ 7.820
ค่ามากสุดของของความถี่ตัวอักษรทั้งหมดอยู่ที่ 61
มีส่วนเบี่ยงเบนมาตรฐานที่ 6.345
จากคอลัมนที่รวมค่าความถี่ของคำที่สร้างขึ้น wordFreq.sumAll

ประมาณค่าเฉลี่ยเป็นช่วง Interval Average

```
> mean(DATASET$Capital_run_length_total)+ c(-E,E)
[1] 265.8179 300.8874
```

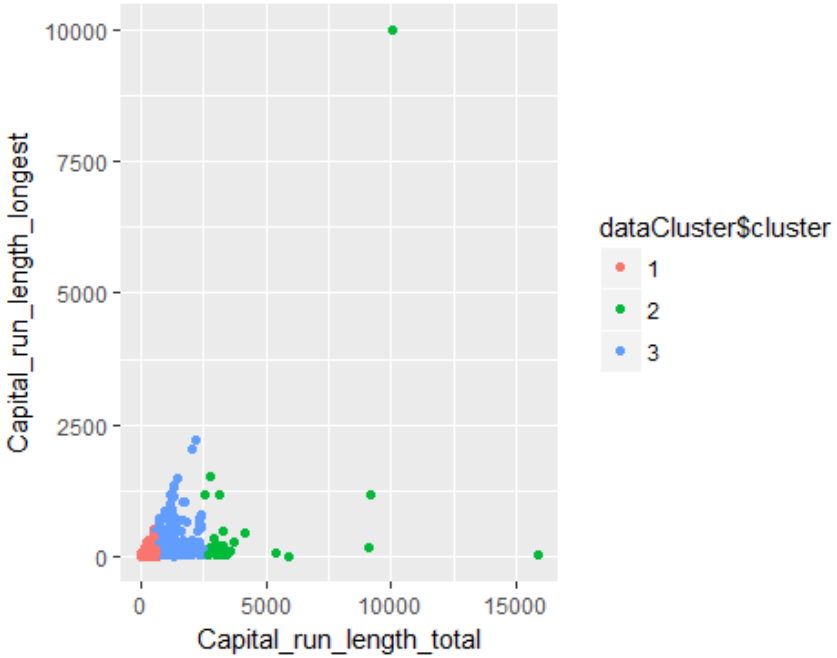
- จำนวนความยาวตัวอักษรทั้งหมดจะอยู่ที่ช่วงประมาณ 266 -301

ทดสอบสมมติฐานค่าเฉลี่ยในข้อ 3 โดยค่าทดสอบสมมุติตามความเหมาะสม

เมื่อจำนวนตัวอักษรของอีเมล์ทั้งหมดมีค่าเฉลี่ยของจำนวนตัวอักษรมากกว่า 300 ตัว จากจำนวน 4597 อีเมล์ เราพบว่าจำนวนตัวอักษรเฉลี่ยเท่ากับ 280 ตัว ่ากับ 280 ตัว ถ้าสมมติว่าส่วนเบี่ยงเบนมาตรฐานประชากรเท่ากับประชากรเท่ากับ 606 ตัว ที่ระดับนัยสำคัญ 0.05 เราสามารถปฏิเสธค่าเฉลี่ยของจำนวนตัวอักษรมากกว่า 300ตัวได้หรือไม่

ทดสอบแล้วได้ เนื่องจาก Z = -2.235519 < $Z_{1\text{-alpha}}$ =-1.644854 ดังนั้น ปริเสธ H0 นั้นคือ ค่าเฉลี่ยของจำนวนตัวอักษรมากกว่า 300ตัว

โมเดล : K Means Clustering

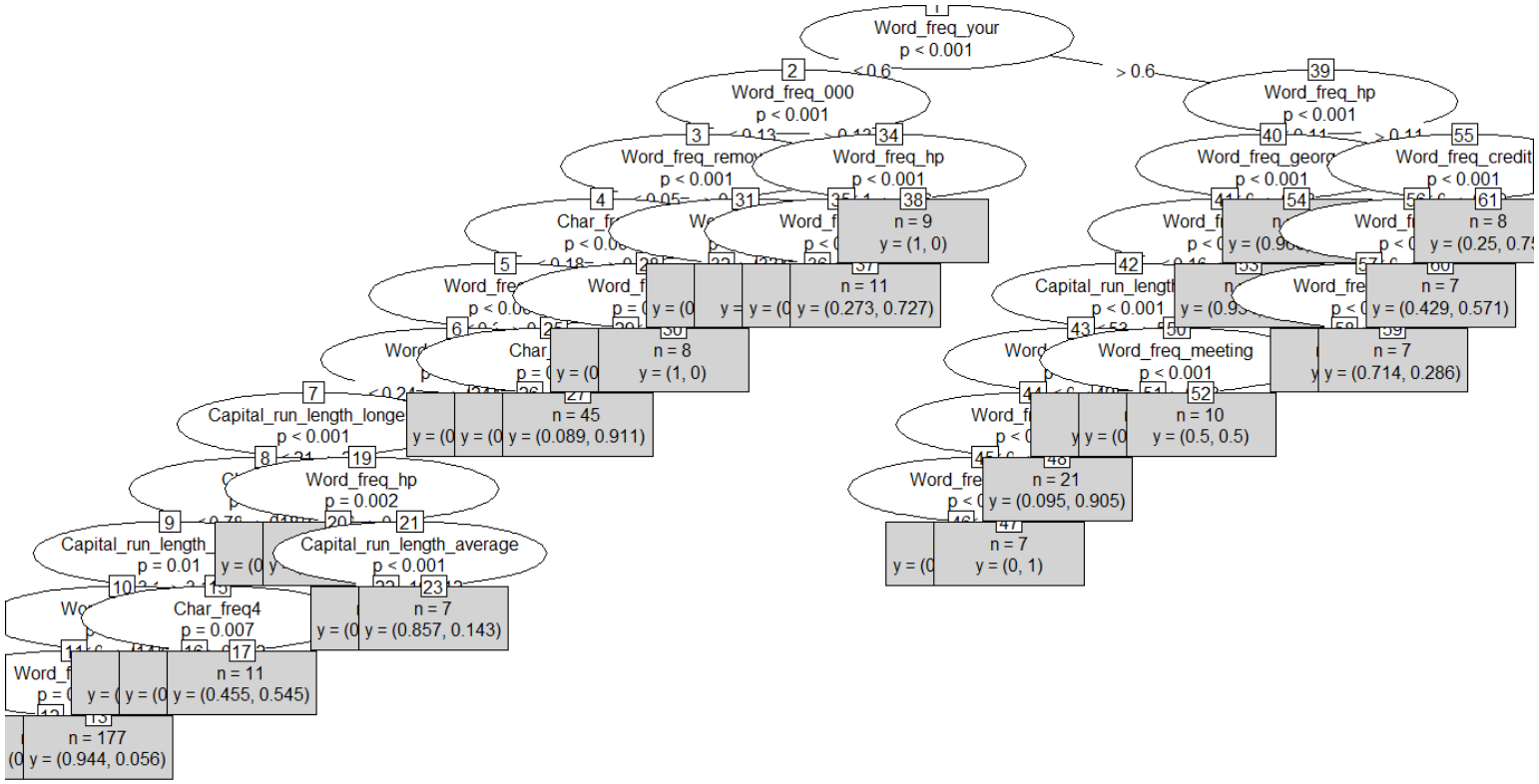


ข้อมูลนี้สามารถแบ่งได้ 3 กลุ่ม โดยใช้ข้อมูลจำนวนตัวอักษรทั้งหมดและความยาวของตัวอักษร

โมเดล : Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction email spam
     email  2651  313
     spam    134 1499

               Accuracy : 0.9028
                 95% CI : (0.8938, 0.
    No Information Rate : 0.6058
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7928
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9519
            Specificity : 0.8273
         Pos Pred Value : 0.8944
         Neg Pred Value : 0.9179
             Prevalence : 0.6058
         Detection Rate : 0.5767
   Detection Prevalence : 0.6448
      Balanced Accuracy : 0.8896

       'Positive' Class : email
```

โมเดล : Decision Tree



โมเดล : Naïve Bayes

```
Confusion Matrix and Statistics

          Reference
Prediction email spam
     email  2651  313
     spam    134 1499

               Accuracy : 0.9028
                 95% CI : (0.8938, 0.9112)
    No Information Rate : 0.6058
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7928
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9519
            Specificity : 0.8273
         Pos Pred Value : 0.8944
         Neg Pred Value : 0.9179
             Prevalence : 0.6058
         Detection Rate : 0.5767
   Detection Prevalence : 0.6448
      Balanced Accuracy : 0.8896

       'Positive' Class : email
```

โมเดล : Naïve Bayes

```
Confusion Matrix and Statistics


preds   email spam
  email   452   28
  spam    366  534

               Accuracy : 0.7145
                 95% CI : (0.6899, 0.7382)
    No Information Rate : 0.5928
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4595
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5526
            Specificity : 0.9502
         Pos Pred Value : 0.9417
         Neg Pred Value : 0.5933
             Prevalence : 0.5928
         Detection Rate : 0.3275
   Detection Prevalence : 0.3478
      Balanced Accuracy : 0.7514

       'Positive' Class : email

>
```

สรุป

- ข้อมูลนี้เป็นการวิเคราะห์ “สแปม" : โฆษณา / ผลิตภัณฑ์เว็บ แคม, ทำเงินได้อย่างรวดเร็วรูปแบบจดหมายลูกโซ่สื่อ
- วัดความยาวของลำดับตัวอักษรทติดกัน ความถี่ของแต่ละคำในอีเมล์ ความถี่ของลักษณะ และจำนวนตัวอักษรทั้งหมดมาช่วยในการวิเคราะห์ทางสถิตว่า อีเมล์นี้เป็นแสปมหรือไม่