STAT5101 Foundations of Data Science Assignment 1

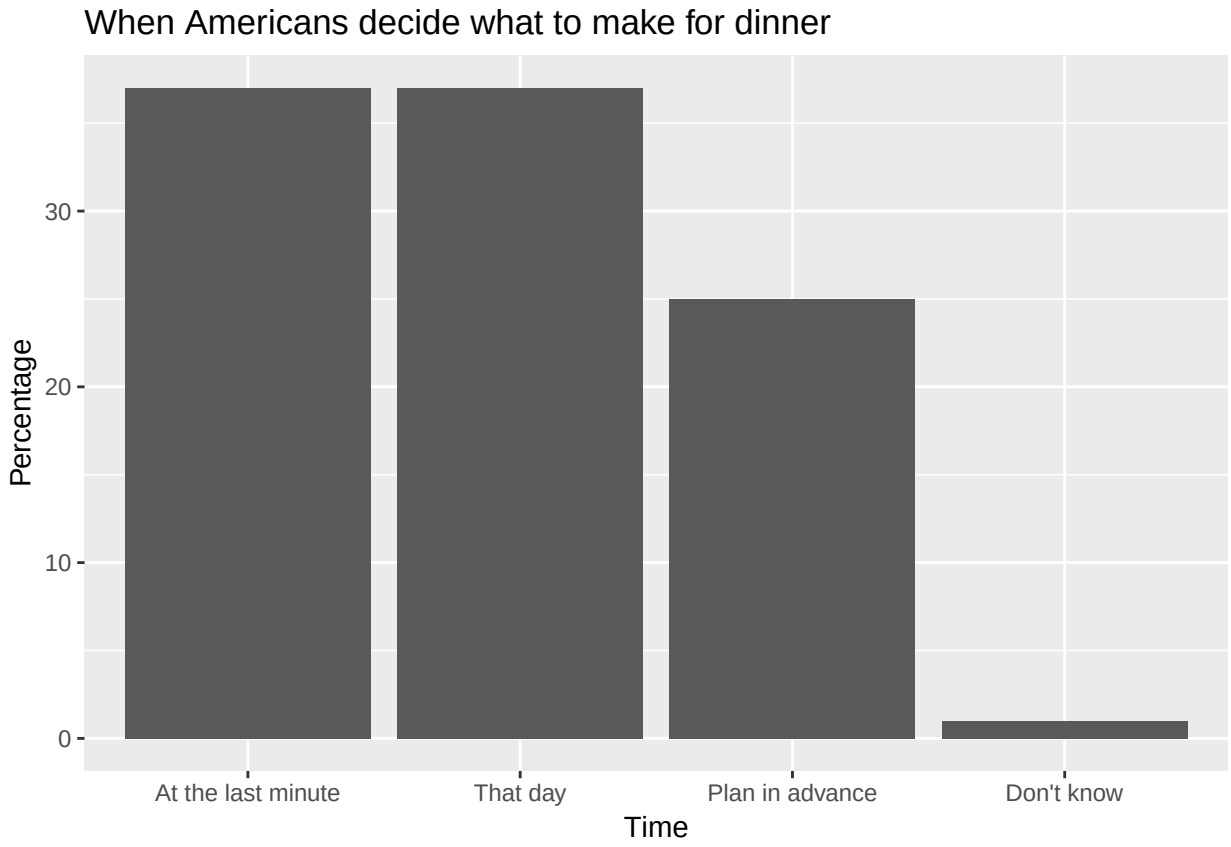
Yiu Chung WONG 1155017920

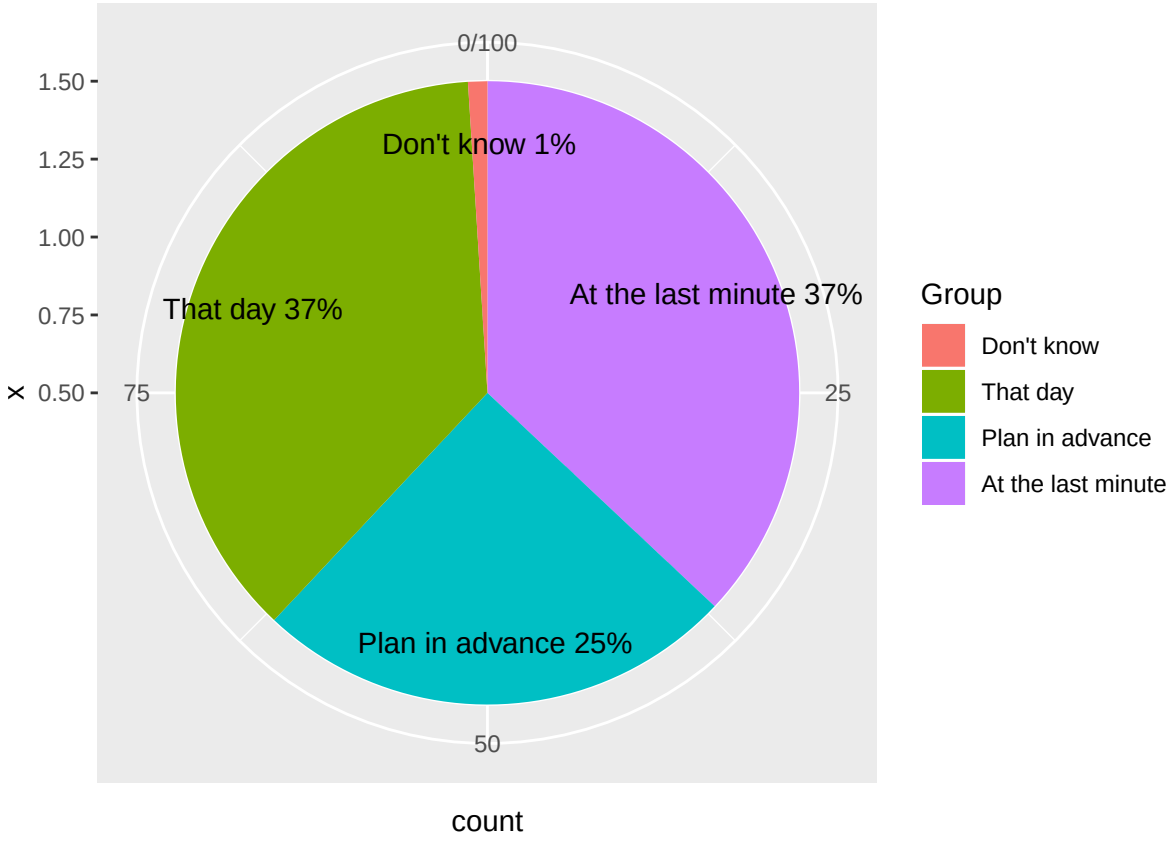
1. When do Americans decide what to make for dinner? An online survey indicated the following:

```
dinner <- data.frame(Time_int = 1:4,
                     Time = c("At the last minute",
                              "Plan in advance",
                              "That day",
                              "Don't know"),
                     Percentage = c(37, 25, 37, 1))
dinner$Time <- with(dinner, factor(dinner$Time, levels=dinner[order(-Percentage), ]$Time))
dinner
```

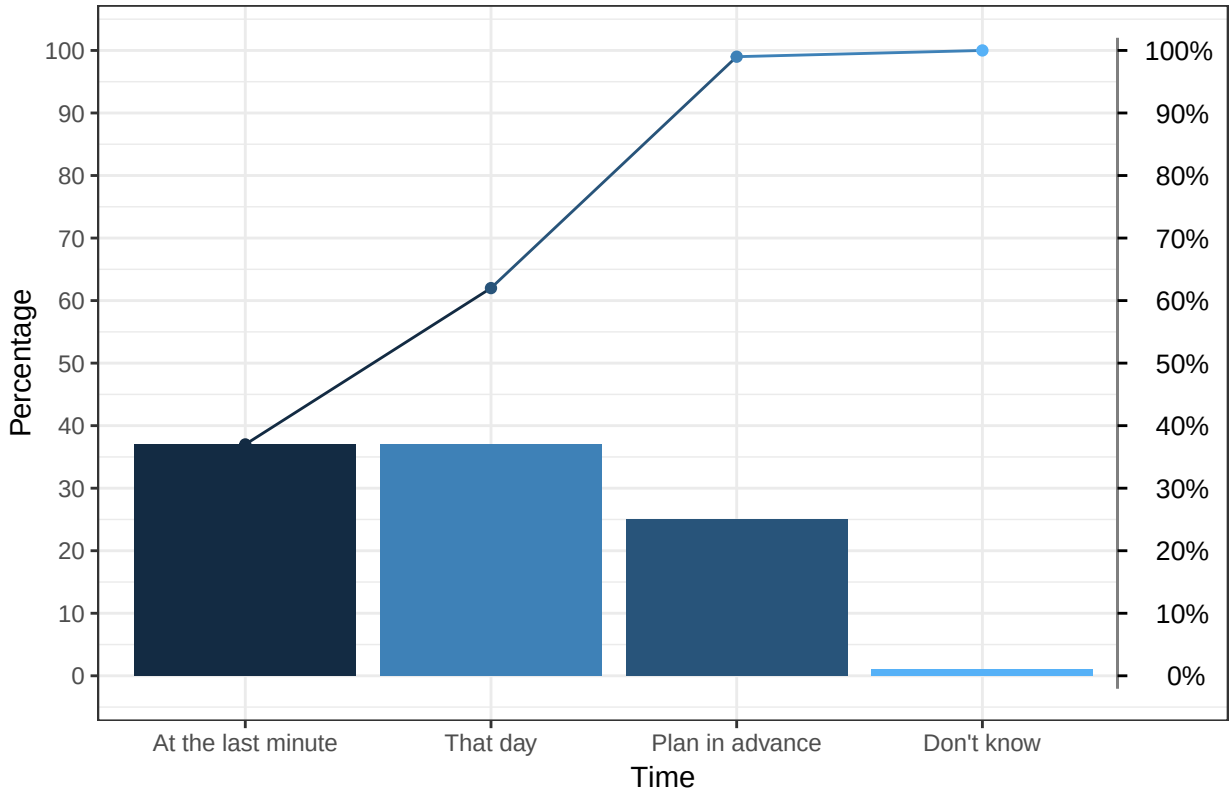
```
##   Time_int               Time Percentage
## 1        1 At the last minute         37
## 2        2    Plan in advance         25
## 3        3           That day         37
## 4        4         Don't know          1
```

(a) Construct a bar chart, a pie chart, and a Pareto diagram.





When Americans decide what to make for dinner



(b) Which graphical method do you like the best to portray these data? Provide one good feature of the selected method

Pareto diagram. Can easily see the portion of each category consumes as well as the ordering.

2. The following data represent the battery life, in shots, for three pixel digital cameras:

```
battery <- c(30, 18, 14, 17, 38, 46, 26, 35, 38, 12, 11)
```

(a) Place the data into an ordered array.

```
battery[order(battery)]
```

```
##  [1] 11 12 14 17 18 26 30 35 38 38 46
```

(b) Construct a stem-and-leaf display

```
stem(battery)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 12478
##   2 | 6
##   3 | 0588
##   4 | 6
```

(c) For the ordered array and the stem-and-leaf display, which method provides more information? Discuss

The stem-and-leaf display provides more information. Not only does the stem-and-leaf display displays ordering, it also illustrates the distrubution of the dataset.

(d) Compute the mean, minimum, maximum, median, first quartile, and third quartile using R.(Hint: apply R commands mean and summary on the vector)

```
summary(battery)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.00   15.50   26.00   25.91   36.50   46.00
```

(e) Compute the mode, variance, standard deviation, range, and coefficient of variatio

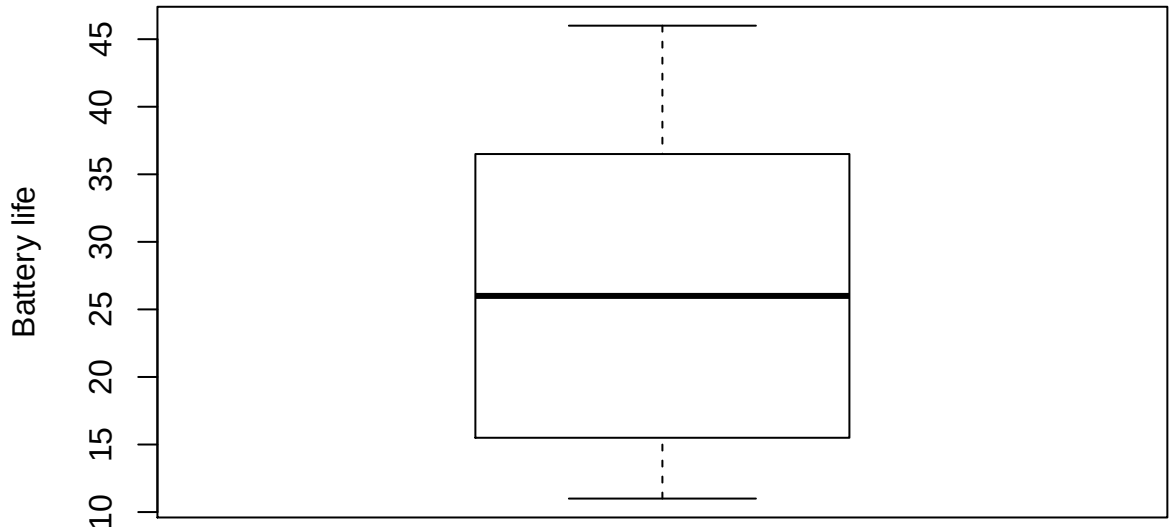
```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
e = list(getmode(battery),
         var(battery),
         sd(battery),
         range(battery),
         scales::percent(sd(battery)/mean(battery))
         )
```

```r
names(e) <- c("mode", "variance", "standard deviation", "range", "coefficient of variatio")
e
```

```
## $mode
## [1] 38
##
## $variance
## [1] 149.4909
##
## $`standard deviation`
## [1] 12.22665
##
## $range
## [1] 11 46
##
## $`coefficient of variatio`
## [1] "47.2%"
```

(f) Form the box-and-whisker plot.

```r
boxplot(battery, ylab ="Battery life")
```



3.

```r
steel <- readxl::read_excel("STEEL.xls", sheet = 1)
incre = (0.00550--0.00350)/6
breaks = seq(-0.00350, 0.00550, by=incre)
steel_cut = cut(steel$Error, breaks)
```

(a) Construct the frequency distribution and the percentage distribution. (use 6 categories between -0.00350 and 0.00550)

```r
data.frame(Frequency = as.numeric(table(steel_cut)),
           Percentage = scales::percent(as.numeric(table(steel_cut))/100),
           Cum_percent = cumsum(table(steel_cut)))
```

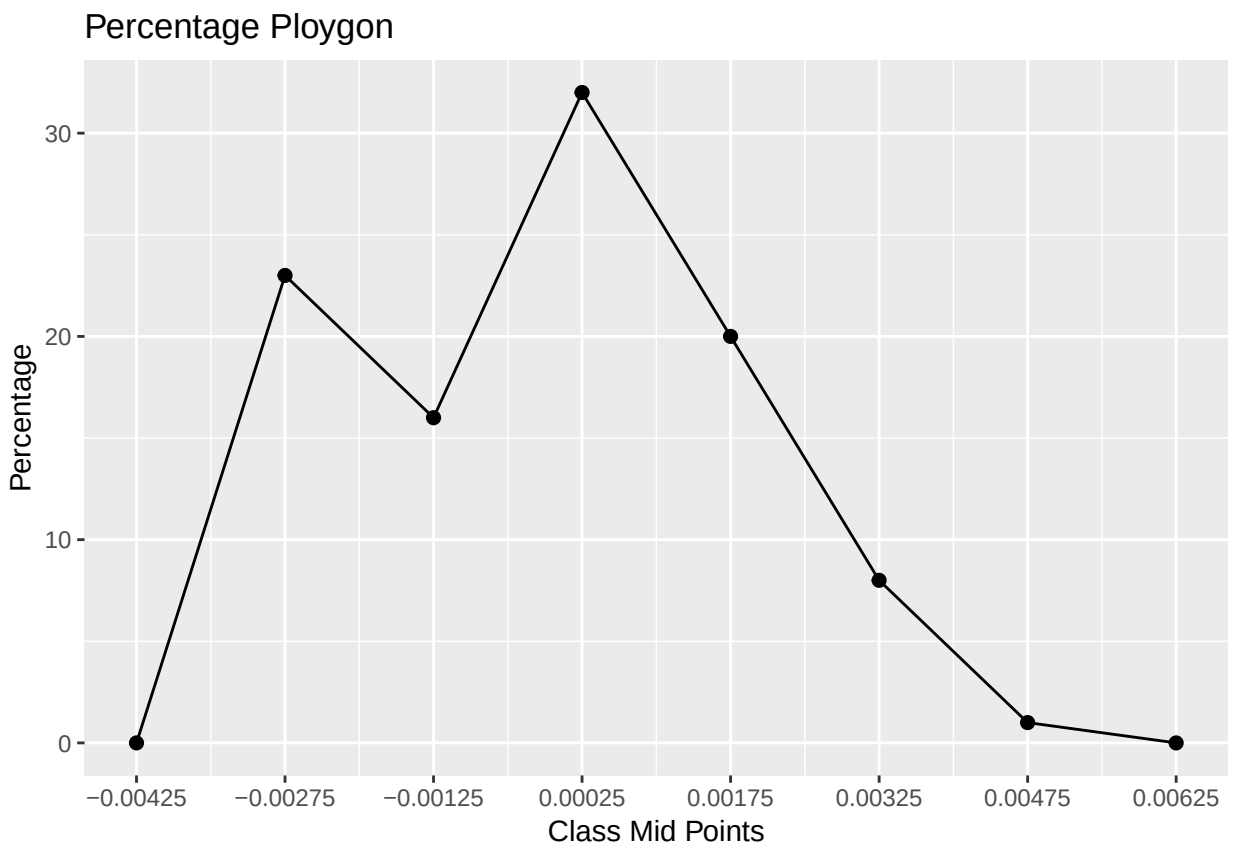
```
##                  Frequency Percentage Cum_percent
## (-0.0035,-0.002]        23        23%          23
## (-0.002,-0.0005]        16        16%          39
## (-0.0005,0.001]         32        32%          71
## (0.001,0.0025]          20        20%          91
## (0.0025,0.004]           8         8%          99
## (0.004,0.0055]           1         1%         100
```

(b) Plot the percentage histogram.

```
ggplot(data = steel, aes(x = Error)) +
        geom_histogram(binwidth = incre) +
        labs(title="Percentage Histogram.",
             x ="Classes", y = "Percentage")
```

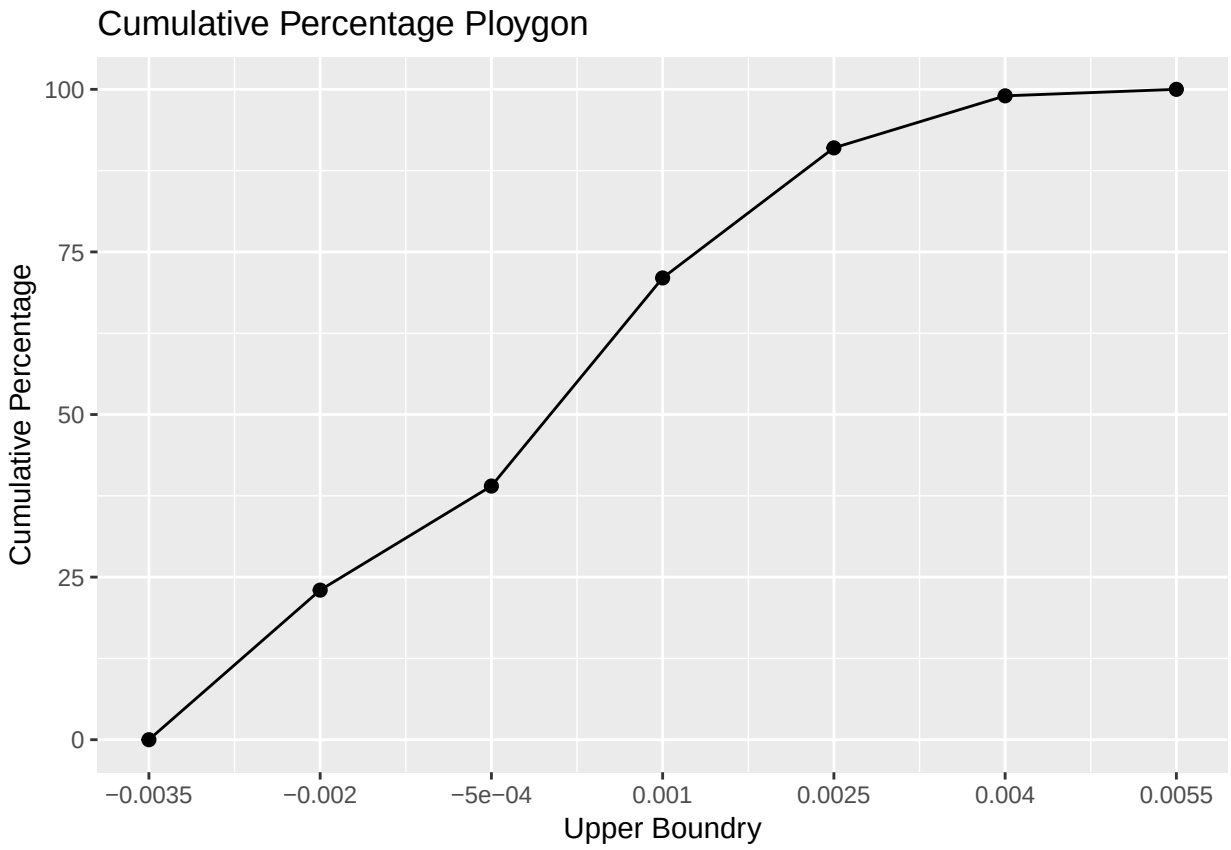
(c) Plot the percentage polygon.

```
mid <- seq(breaks[1], by = incre, length.out = 8) - incre/2

p_poly <- data.frame(mid = mid,
                     freq = c(0, as.numeric(table(steel_cut)), 0))

ggplot(p_poly, aes(mid, freq))  +
        geom_point(data=p_poly, mapping=aes(x = mid, y=freq), size=2) +
        geom_line() +
        scale_x_continuous(breaks = p_poly$mid) +
        labs(title="Percentage Ploygon", x ="Class Mid Points", y = "Percentage")
```



(d) Plot the cumulative percentage ploygon.

```
cp_ploygon = data.frame(upper = breaks,
                        cp = c(0, as.numeric(cumsum(table(steel_cut)))))

ggplot(cp_ploygon, aes(upper, cp))  +
        geom_point(data=cp_ploygon, mapping=aes(y=cp), size=2) +
        geom_line() +
        scale_x_continuous(breaks = cp_ploygon$upper, labels = cp_ploygon$upper) +
        labs(title="Cumulative Percentage Ploygon",
        x ="Upper Boundry", y = "Cumulative Percentage")
```



4.

```
battery2 <- readxl::read_excel("BATTERIES2.xls", sheet = 1)
```

(a) Compute the coefficient of correlation r.

```
cor(battery2$`Price ($)`, battery2$CCA)
```

```
## [1] 0.4837564
```

(b) What conclusions can you reach about the relationship between the cold-cranking amps and the price?

They are moderately correlated; when one goes up, the other also goes up slightly.

5. Construct a side-by-side bar chart for three age groups.

```
media = rep(c("Local TV", "National TV", "Radio", "Local newspaper", "Internet"), each = 3)
count = c(107, 119, 133, 73, 102, 127, 75, 97, 109, 52, 79, 107, 95, 83, 76)
age = rep(c("Under 36", "36-50", "50+"), 5)

news  <- data.frame(media, count, age)

ggplot(news, aes(media, count)) +
```

```
        geom_bar(aes(fill = age),
                 width = .4,
                 position = position_dodge(width=0.5),
                 stat="identity") +
        labs(title="\nNumber of people across media categories\n(grouped by age)",
        x ="Media", y = "Count")
```