# ADM 2303 - Assignment 1

Renad Gharzeddine

02/05/2022

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

## 1. Child Safety Seat Survey

Canada has a Road Safety Vision of having the safest roads in the world. Yet, the leading cause of death of Canadian children remains vehicle crashes. In 2006, a national child safety seat survey was conducted by an AUTO21 research team in collaboration with Transport Canada to empirically measure Canada's progress toward achieving Road Safety Vision 20210. Child seat use was observed in parking lots and nearby intersections in 200 randomly selected sites across Canada.

**Age Groups**

```
##   X1            X2
## 1  1  Infant (0-1)
## 2  2 Toddler (1-4)
## 3  3  School (4-9)
## 4  4    Older (9+)
```

**Restraint Types**

```
##   X1             X2
## 1  R    Rear-facing
## 2  F Forward-facing
## 3  B   Booster seat
## 4  S      Seat belt
```

### 1.1 Contingency Table

> Using data table, create a 4×4 cross-tabulation (i.e., contingency or pivot table) of the children in the survey by age group (row position) and type of restraint (columnposition).

Crosstab with counts:

```
crosstab <- table(data$AgeGroup, data$RestType)# Converting data to contingency table
crosstab_margins <- addmargins(crosstab)# Adding margins (sums) to contingency table
crosstab_margins
```

```
##
##          B    F    R    S  Sum
##   1      1   52  181    0  234
##   2    117  483   49    3  652
##   3    450   98    0  325  873
##   4     16    0    0  627  643
##   Sum  584  633  230  955 2402
```

Crosstab with proportions:

```r
prop_table <- prop.table(crosstab)# Converting contingency table to proportion format
prop_table_margins <- addmargins(prop_table)# Adding margins (sums) to proportion contingency table
round(prop_table_margins, digits = 4)
```

```
##
##            B      F      R      S    Sum
##   1   0.0004 0.0216 0.0754 0.0000 0.0974
##   2   0.0487 0.2011 0.0204 0.0012 0.2714
##   3   0.1873 0.0408 0.0000 0.1353 0.3634
##   4   0.0067 0.0000 0.0000 0.2610 0.2677
##   Sum 0.2431 0.2635 0.0958 0.3976 1.0000
```

## 1.2 Data Types

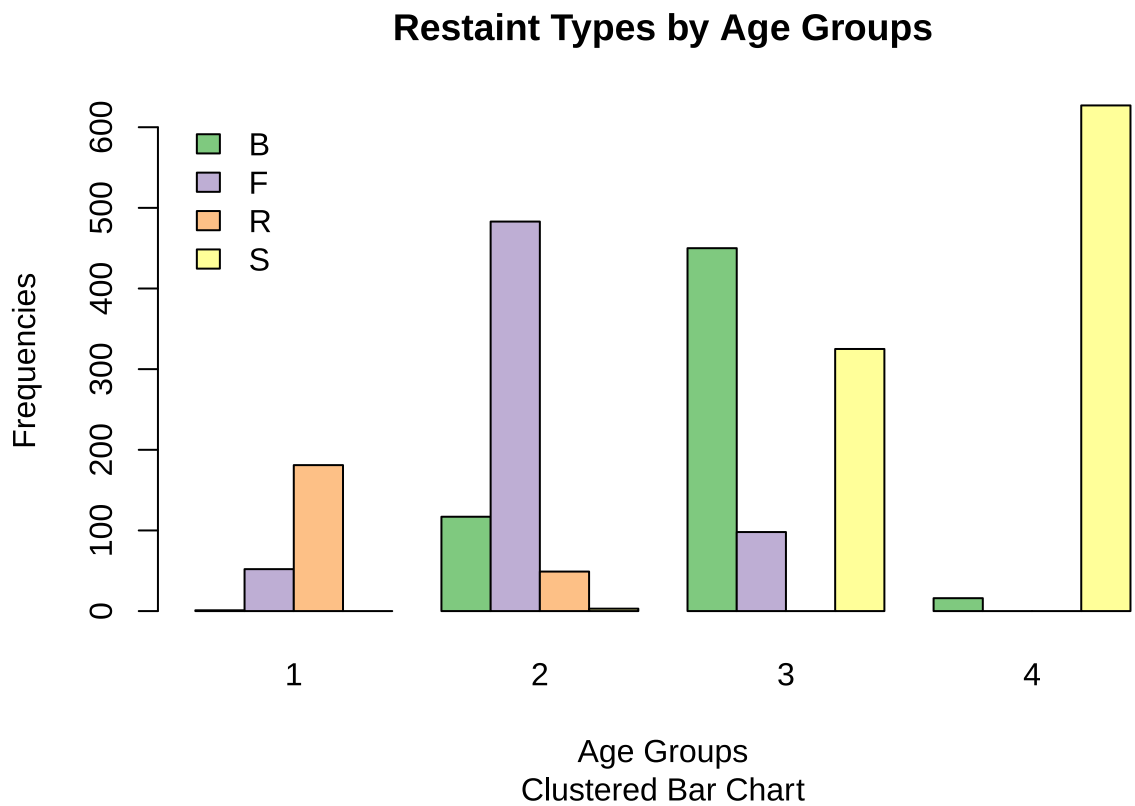
> What are the variables measured in this survey? Are they qualitative (i.e., categorical) or quantitative?

There are 2 variables measured in this survey: *AgeGroup*, *RestType*. Both variables are qualitative (or categorical) in nature; *AgeGroup* because we are looking at which age group the children in the survey belong in rather than their actual age itself, and *RestType* because we are looking at the type of safety restraint applied to each child, there are no quantifiable values. *AgeGroup* has 4 possible values: 1, 2, 3, 4; and *RestType* also has 4 possible values: B, F, R, S.

## 1.3 Side-by-Side Bar Chart

> Construct a side-by-side bar chart to compare the type of restraints at different age groups.

```r
crosstab_inverse <- table(data$RestType, data$AgeGroup) # Flipping rows and columns of the crosstab
crosstab_clustered_bars <- barplot(crosstab_inverse,
                                beside = TRUE,
                                main = "Restaint Types by Age Groups",
                                sub = "Clustered Bar Chart",
                                xlab = "Age Groups",
                                ylab = "Frequencies",
                                col = brewer.pal(n = 4, name = "Accent"),
                                legend.text = c("B","F","R","S"),
                                args.legend = list(x = "topleft",
                                                   bty = "n",
                                                   inset=c(0.01, 0)))# Clustered bar chart comparing re
```

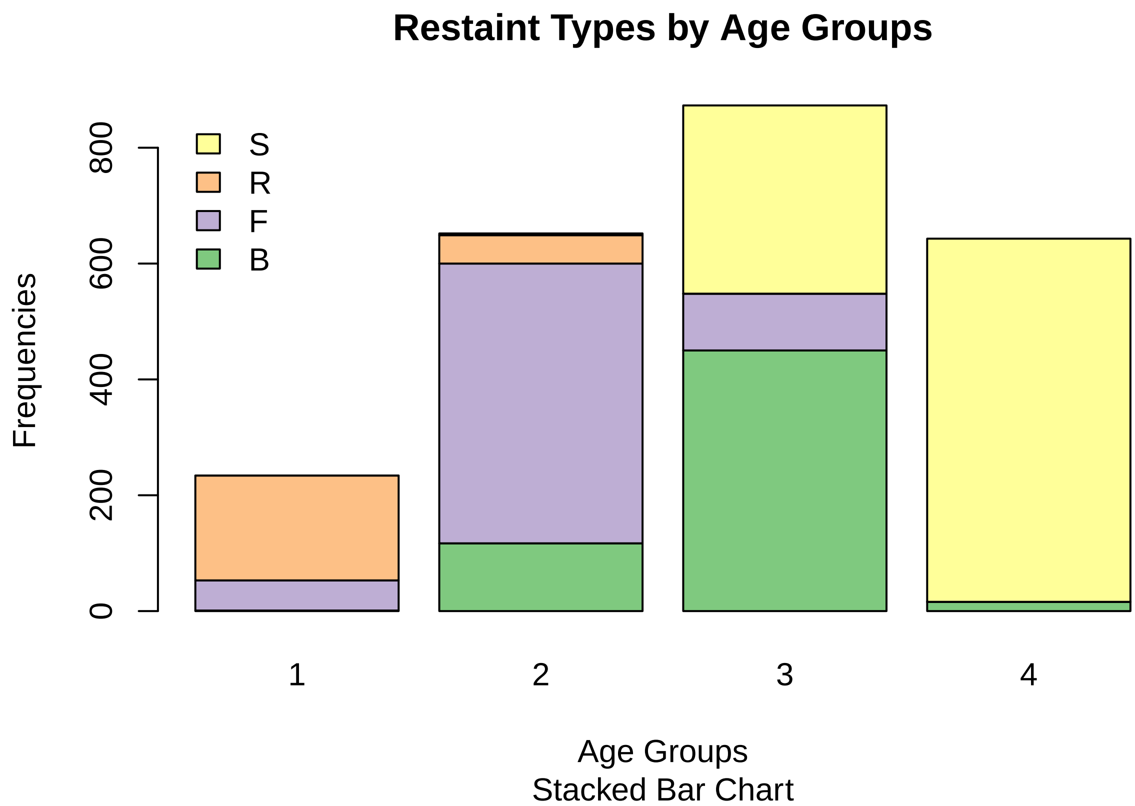



### 1.4 Pie and Stacked Bar Charts

> Construct pie charts or a stacked bar chart to compare the type of restraints at different age groups.

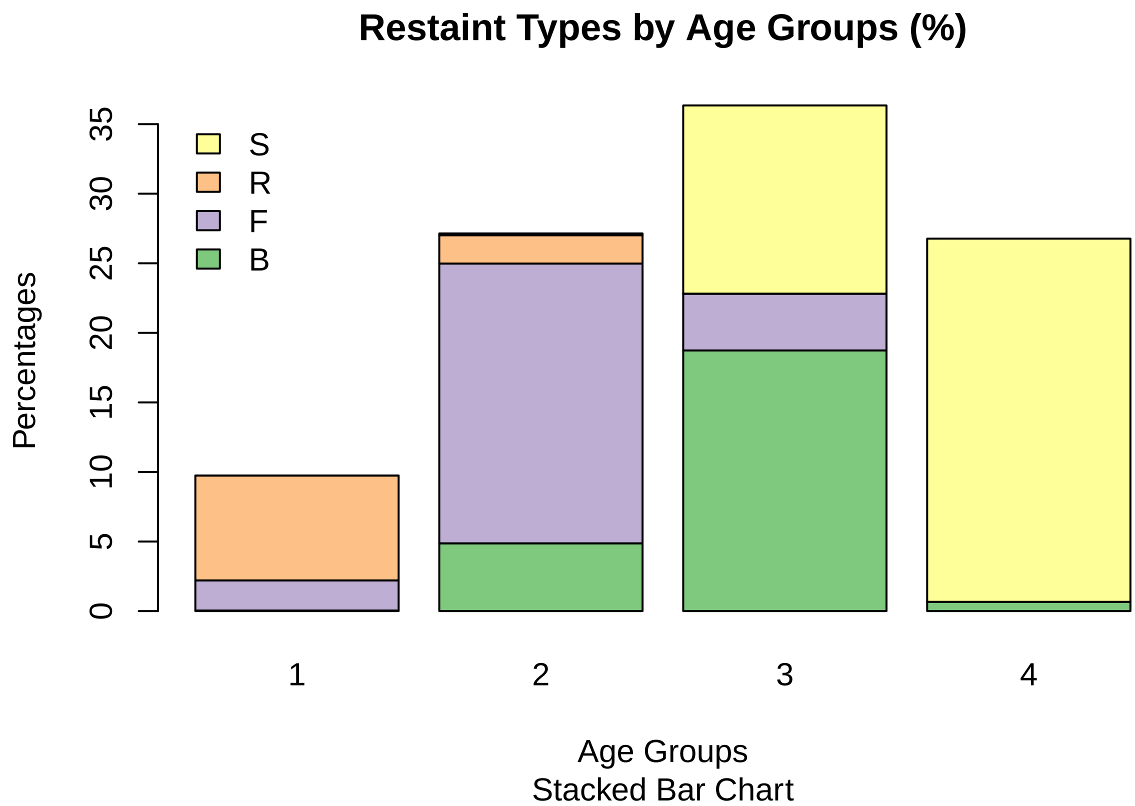
Stacked bar charts (counts):

```
crosstab_stackedbars <- barplot(crosstab_inverse,
                                main = "Restaint Types by Age Groups",
                                sub = "Stacked Bar Chart",
                                xlab = "Age Groups",
                                ylab = "Frequencies",
                                col = brewer.pal(n = 4, name = "Accent"),
                                legend.text = c("B","F","R","S"),
                                args.legend = list(x = "topleft",
                                                   bty = "n",
                                                   inset=c(0.01, 0))) # Stacked bar chart comparing res
```




Stacked bar charts (percentages):

```
prop_table_inverse <- prop.table(crosstab_inverse) * 100 # Converting proportions to percentages for be
crosstab_stackedbars <- barplot(prop_table_inverse,
                                main = "Restaint Types by Age Groups (%)",
                                sub = "Stacked Bar Chart",
                                xlab = "Age Groups",
                                ylab = "Percentages",
                                col = brewer.pal(n = 4, name = "Accent"),
                                legend.text = c("B","F","R","S"),
                                args.legend = list(x = "topleft",
                                                   bty = "n",
                                                   inset=c(0.01, 0))) # Stacked bar chart comparing res
```




### 1.5 Summary of Data

Write a short paragraph summarizing the information that can be gained by looking at these graphs.

From these graphs we can conclude that starting from age 4 to 9, children can start foregoing other restraints in exchange for seat belts while almost all children aged 9 or older don't require booster seats anymore and can simply use seat belts. As for infants, they need to be put in a rear-facing position in cars and starting from 1 to 4 years old they can be switched to front-facing positions, while few remain rear-facing and some toddlers can even be placed in booster seats.

## 2. National Cable Service

Like all companies, cable companies send stakeholders reports on their profits, dividends, and return on equity. They often supplement this information with some metrics unique to the cable business. To construct one such metric, a cable company can compare the number of households it actually serves to the number of households its current transmission lines could reach (without extending lines). The number of households that the cable company's lines could reach is called its number of cable passings, while the ratio of the number of households the cable company actually serves to its number of cable passings is called the company's cable penetration. There are various types of cable penetrations - one for cable television, one for cable internet, once for cable phone, and others. For example, National cable television penetration is a probability defined as follows:

$$\frac{\text{the number of cable passings that have National's cable television services}}{\text{the total number of cable passings}}$$

National's cable has 38 million cable passings. Let us consider National cable's two services viz. cable television service (A) and cable internet service (B). 10.9 million has only cable television service and 10.1 million has only cable internet service, while 8.2 million has both services.

## 2.1 Contingency Tables

> Create a 2 × 2 contingency table considering cable television service (A) in the row position and cable internet service in the column position (B).

Crosstab with counts:

```
cable_service_table <- data.frame(B = c(8.2, 10.1, 18.3),
                                  B_c = c(10.9, 8.8, 19.7),
                                  Sum = c(19.1, 18.9, 38),
                                  row.names = c("A", "A_c", "Sum")) # Creating "crosstab" of the data u
                                                                    # Can't use table() function becaus
cable_service_table
```

```
##        B  B_c  Sum
## A    8.2 10.9 19.1
## A_c 10.1  8.8 18.9
## Sum 18.3 19.7 38.0
```

Crosstab with proportions:

```
cable_service_table_prop <- data.frame(B = c(8.2, 10.1, 18.3),
                                       B_c = c(10.9, 8.8, 19.7),
                                       Sum = c(19.1, 18.9, 38),
                                       row.names = c("A", "A_c", "Sum")) %>%
  mutate(across(where(is.numeric)) / cable_service_table[3,3]) # Converting crosstab from frequencies t
round(cable_service_table_prop, digits = 4)
```

```
##          B    B_c    Sum
## A   0.2158 0.2868 0.5026
## A_c 0.2658 0.2316 0.4974
## Sum 0.4816 0.5184 1.0000
```

## 2.2 Probability of Union

> What is the probability that a randomly selected cable passing has either cable television service or cable internet service?

```
prob_A_u_B <- cable_service_table_prop[1,3] + cable_service_table_prop[3,1] - cable_service_table_prop[
prob_A_u_B
```

```
## [1] 0.7684211
```

## 2.3 Probability of Intersection

> What is the probability that a randomly selected cable passing does not have National's cable television service and does not have National's cable internet service?

```
prob_Ac_n_Bc <- cable_service_table_prop[2,2] # Joint probability
prob_Ac_n_Bc
```

```
## [1] 0.2315789
```

### 2.4 Mutually Exclusive Events

> Are the events cable television service and cable internet service mutually exclusive? Justify.

No, cable television service and cable internet service are NOT mutually exclusive because they can happen at the same time because it's possible for cable passings to have both (8.2 million capable passings or 21.58% of all cable passings to be exact). The occurrence of cable television service does not preclude cable internet service and vice-versa, thus they are not mutually exclusive.

We can also prove that they are not mutually exclusive since $P(A \cap B) \neq 0$, where A is cable television service and B is cable internet service, which is a requirement for 2 events to be mutually exclusive.

### 2.5 Independence of Events

> Are the events cable television service and cable internet service independent? Justify.

```
cable_service_table_prop[1,1] == cable_service_table_prop[1,3] * cable_service_table_prop[3,1]  # Since
```

```
## [1] FALSE
```

$P(A \cap B) \neq P(A) \cdot P(B)$ thus, the events A and B are not independent because they fail one of the 3 conditions for independent events.

### 2.6 Conditional Probability

> If a randomly selected cable has television service, what is the probability that it does not have cable internet service?

```
prob_Bc_given_A <- cable_service_table_prop[1,2] / cable_service_table_prop[1,3] # Join probability of
prob_Bc_given_A
```

```
## [1] 0.5706806
```

## 3. Flight Delays

Below we give two contingency tables of data from reports submitted by airlines to the U.S. Department of Transportation. The data concern the numbers of on-time and delayed flights for Delta and Frontier Airlines at five major airports.

**Delta Airlines**

```
##               OnTime Delayed Total
## Los Angeles      248      31   279
## Phoenix          110       6   116
## San Diego        106      10   116
## San Francisco    252      51   303
## Seattle          920     152  1072
## Total           1636     250  1886
```

**Frontier Airlines**

```
##               OnTime Delayed Total
## Los Angeles      231      39   270
## Phoenix         1613     138  1751
## San Diego        128      22   150
## San Francisco    107      43   150
## Seattle           67      20    87
## Total           2146     262  2408
```

We can convert the count data above into proportions to make it easier to calculate probabilities:

Delta Airlines:

```
delta_flights_prop <- data.frame(OnTime = c(248,110,106,252,920,1636),
                                 Delayed = c(31,6,10,51,152,250),
                                 Total = c(279,116,116,303,1072,1886),
                                 row.names = c("Los Angeles","Phoenix","San Diego","San Francisco","Sea
  mutate(across(where(is.numeric)) / delta_flights[6,3]) # Table for Delta Airlines
round(delta_flights_prop, digits = 4)
```

```
##               OnTime Delayed  Total
## Los Angeles   0.1315  0.0164 0.1479
## Phoenix       0.0583  0.0032 0.0615
## San Diego     0.0562  0.0053 0.0615
## San Francisco 0.1336  0.0270 0.1607
## Seattle       0.4878  0.0806 0.5684
## Total         0.8674  0.1326 1.0000
```

Frontier Airlines:

```
frontier_flights_prop <- data.frame(OnTime = c(231,1613,128,107,67,2146),
                                    Delayed = c(39,138,22,43,20,262),
                                    Total = c(270,1751,150,150,87,2408),
                                    row.names = c("Los Angeles","Phoenix","San Diego","San Francisco","S
  mutate(across(where(is.numeric)) / frontier_flights[6,3])# Table for Frontier Airlines
round(frontier_flights_prop, digits = 4)
```

```
##               OnTime Delayed  Total
## Los Angeles   0.0959  0.0162 0.1121
## Phoenix       0.6699  0.0573 0.7272
## San Diego     0.0532  0.0091 0.0623
## San Francisco 0.0444  0.0179 0.0623
## Seattle       0.0278  0.0083 0.0361
## Total         0.8912  0.1088 1.0000
```

## 3.1 Marginal Probabilities

> What percentage of all Delta Airlines flights were delayed? That is, use the data to estimate the probability that an Delta Airline flight will be delayed. Do the same for Frontier Airlines? Which airline does best overall?

```
delayed_delta_flights <- delta_flights_prop[6,2] # Probability of delayed flights for Delta Airlines
round(delayed_delta_flights, digits = 4)
```

```
## [1] 0.1326
```

```
delayed_frontier_flights <- frontier_flights_prop[6,2] # Probability of delayed flights for Frontier Ai
round(delayed_frontier_flights, digits = 4)
```

```
## [1] 0.1088
```

## 3.2 Conditional Probabilities

> For Delta Airlines, find the percentage of delayed flights at each airport. That is, use the data to estimate each of the probabilities P(delayed | Los Angeles), P(delayed | Phoenix), and so on. Then do the same for Frontier Airlines. Which airline does best at each individual airport?

```r
# Delta Airlines
p_delta_delayed_given_los_angeles <- delta_flights_prop[1,2] /
  delta_flights_prop[1,3] # P(Delta | Los Angeles)
p_delta_delayed_given_phoenix <- delta_flights_prop[2,2] /
  delta_flights_prop[2,3] # P(Delta | Phoenix)
p_delta_delayed_given_san_diego <- delta_flights_prop[3,2] /
  delta_flights_prop[3,3] # P(Delta | San Diego)
p_delta_delayed_given_san_francisco <- delta_flights_prop[4,2] /
  delta_flights_prop[4,3] # P(Delta | San Francisco)
p_delta_delayed_given_seattle <- delta_flights_prop[5,2] /
  delta_flights_prop[5,3] # P(Delta | Seattle)

# Frontier Airlines
p_frontier_delayed_given_los_angeles <- frontier_flights_prop[1,2] /
  frontier_flights_prop[1,3] # P(Frontier | Los Angeles)
p_frontier_delayed_given_phoenix <- frontier_flights_prop[2,2] /
  frontier_flights_prop[2,3] # P(Frontier | Phoenix)
p_frontier_delayed_given_san_diego <- frontier_flights_prop[3,2] /
  frontier_flights_prop[3,3] # P(Frontier | San Diego)
p_frontier_delayed_given_san_francisco <- frontier_flights_prop[4,2] /
  frontier_flights_prop[4,3] # P(Frontier | San Francisco)
p_frontier_delayed_given_seattle <- frontier_flights_prop[5,2] /
  frontier_flights_prop[5,3] # P(Frontier | Seattle)

# Dataframe to display all the conditional probabilities from 3.2 in a readable format
summary_conditional_probabilities <-
  data.frame(Delta = c(p_delta_delayed_given_los_angeles,
                       p_delta_delayed_given_phoenix,
                       p_delta_delayed_given_san_diego,
                       p_delta_delayed_given_san_francisco,
                       p_delta_delayed_given_seattle),
             Frontier = c(p_frontier_delayed_given_los_angeles,
                          p_frontier_delayed_given_phoenix,
                          p_frontier_delayed_given_san_diego,
                          p_frontier_delayed_given_san_francisco,
                          p_frontier_delayed_given_seattle),
             row.names = c("Los Angeles","Phoenix","San Diego","San Francisco","Seattle"))
round(summary_conditional_probabilities, digits = 4)
```

```
##                Delta Frontier
## Los Angeles   0.1111   0.1444
## Phoenix       0.0517   0.0788
## San Diego     0.0862   0.1467
## San Francisco 0.1683   0.2867
## Seattle       0.1418   0.2299
```

From the table above, we can conclude that Delta Airlines does better than Frontier Airlines at each airport in terms of percentage of delayed flights.

## 3.3 Simpson's Paradox

> Compare the results of part 3.1 and 3.2 i.e., the performance of both airlines? Are they aligned or contradictory? Explain.

We cannot compare the results from 3.1 to the results from 3.2 because of Simpson's Paradox.