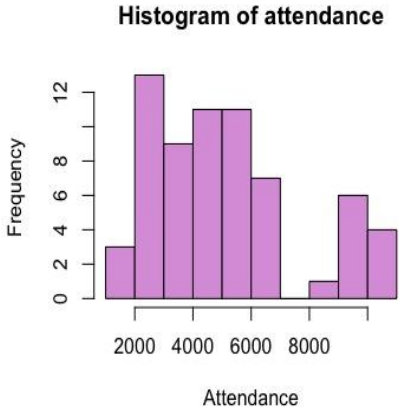
ISOM-670 Business Statistics

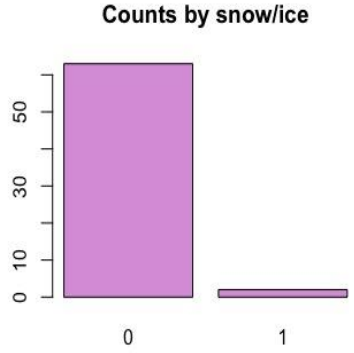
Group Assignment 2

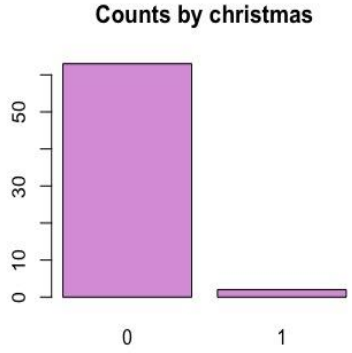
Team 2: Yifei, Jie, Eleanor & Susie

A First Look at Our Data - Descriptive Statistics & Graphs

```
> summary(Ndata)
   attendance        adult           school          snow/ice          temp          sunshine        christmas         advertising
 Min.   : 1407   Min.   :5.500   Min.   :0.000   Min.   :0.00000   Min.   :38.50   Min.   : 7.50   Min.   :0.00000   Min.   : 105
 1st Qu.: 3009   1st Qu.:5.500   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:48.38   1st Qu.:43.75   1st Qu.:0.00000   1st Qu.:2195
 Median : 4538   Median :5.500   Median :5.000   Median :0.00000   Median :61.00   Median :57.50   Median :0.00000   Median :4237
 Mean   : 5105   Mean   :5.631   Mean   :3.585   Mean   :0.03077   Mean   :61.31   Mean   :55.57   Mean   :0.03077   Mean   :4564
 3rd Qu.: 6169   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:0.00000   3rd Qu.:76.88   3rd Qu.:70.88   3rd Qu.:0.00000   3rd Qu.:6786
 Max.   :10830   Max.   :6.000   Max.   :5.000   Max.   :1.00000   Max.   :87.62   Max.   :91.12   Max.   :1.00000   Max.   :9732
```



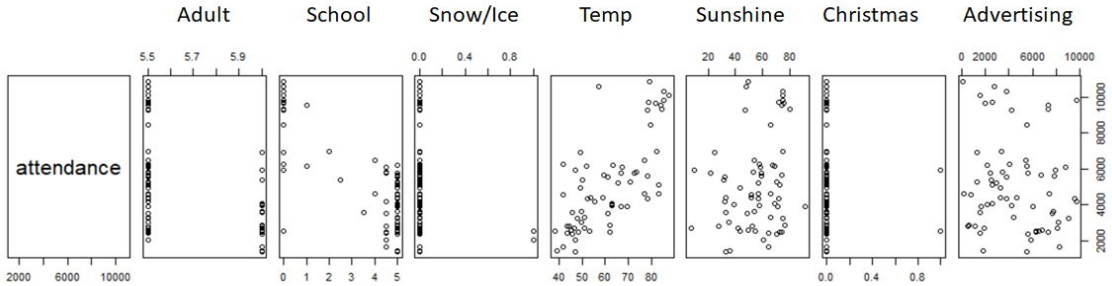




Observed from the descriptive statistics and graphs, there are no missing values in the data. The distributions of *attendance*, *adult*, *temp* and *advertising* are slightly skewed to the right, while the ones of *school* and *sunshine* bias toward the left.

Notably, both of the categorical variables *snow/ice* and *christmas* have a very unbalanced distribution, with the vast majority of value equaling “0”(only 2 of the entries are “1” in each case).

Relationship between *attendance* and Independent Variables



	attendance	adult	school	snow/ice	temp	sunshine	christmas	advertising
attendance	1.0000000	-0.4045292	-0.7734841	-0.2005705	0.7105618	0.2384953	-0.0628370	-0.1194839
adult	-0.4045292	1.0000000	0.1460357	-0.1060348	-0.3939207	-0.3591709	0.0966788	-0.1292790
school	-0.7734841	0.1460357	1.0000000	-0.1158652	-0.3372894	-0.0831032	-0.3111999	0.0534080
snow/ice	-0.2005705	-0.1060348	-0.1158652	1.0000000	-0.2266299	-0.0357111	0.4841270	0.1083775
temp	0.7105618	-0.3939207	-0.3372894	-0.2266299	1.0000000	0.4740354	-0.2258641	-0.0632674
sunshine	0.2384953	-0.3591709	-0.0831032	-0.0357111	0.4740354	1.0000000	-0.2862793	0.0275418
christmas	-0.0628370	0.0966788	-0.3111999	0.4841270	-0.2258641	-0.2862793	1.0000000	0.1749587
advertising	-0.1194839	-0.1292790	0.0534080	0.1083775	-0.0632674	0.0275418	0.1749587	1.0000000

Based on the scatterplots and the correlation matrix between the dependent variable *attendance* and all independent variables, we could clearly spot that *attendance* is highly positively correlated with *temp*, while there is an apparent negative relationship between *attendance* and *school* on the other hand.

Moreover, certain correlations are found among some of the independent variables, such as *christmas* and *snow/ice*, *sunshine*, and *temp*, which make sense in practice. Relatively significant levels of correlation are highlighted in the matrix for later reference.

Independent Variables Selection

- Start with a full regression model of *attendance* against all independent variables

```
> full = lm(attendance ~ .,data=Ndata)
> summary(full)

Call:
lm(formula = attendance ~ ., data = Ndata)

Residuals:
     Min       1Q   Median       3Q      Max
-1666.64  -611.73   -61.69   497.98  2183.47

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.820e+04  3.490e+03   5.216 2.65e-06 ***
adult       -2.333e+03  5.608e+02  -4.159 0.000109 ***
school      -8.231e+02  6.135e+01 -13.417  < 2e-16 ***
`snow/ice`  -2.264e+03  7.456e+02  -3.036 0.003608 **
temp         6.841e+01  9.780e+00   6.995 3.25e-09 ***
sunshine    -1.610e+01  7.227e+00  -2.228 0.029874 *
christmas   -1.636e+03  8.142e+02  -2.010 0.049183 *
advertising -4.216e-02  4.093e-02  -1.030 0.307342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 859.4 on 57 degrees of freedom
Multiple R-squared:  0.898,     Adjusted R-squared:  0.8854
F-statistic: 71.65 on 7 and 57 DF,  p-value: < 2.2e-16
```

- Parameter elimination process

(1) **remove *date***
(2) **remove *advertising***: very little statistical significance indicated by t-stats in the full linear regression model; high p-value when regressing *attendance* against *advertising* alone
(3) **remove *christmas***: only two data points representing Christmas; high correlation between *christmas* and *snow/ice*
(4) **remove *sunshine***: high correlation between *sunshine* and *temp*; lower explanatory power than *temp*
(5) Consequently, we include *adult*, *school*, *snow/ice*, and *temp* as independent variables in our forecasting model

Our Forecasting Model

```
> summary(lm(attendance~adult+school+temp+`snow/ice`))

Call:
lm(formula = attendance ~ adult + school + temp + `snow/ice`)

Residuals:
     Min       1Q   Median       3Q      Max
-1543.75  -754.01    36.21   568.32  2525.64

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16111.266   3519.166   4.578 2.41e-05 ***
adult       -2134.751    569.698  -3.747 0.000404 ***
school       -792.161     59.441 -13.327  < 2e-16 ***
temp           64.407      9.396   6.855 4.43e-09 ***
`snow/ice`  -3075.682    699.561  -4.397 4.57e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 904.8 on 60 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.873
F-statistic:   111 on 4 and 60 DF,  p-value: < 2.2e-16
```

We end up with this model because (a) it reduces correlation among independent variables; (b) it provides stronger statistical performance compared to any other model we have tried (see Appendix).

This model, regressing *attendance* against the four predictors *school*, *temp*, *adult*, and *snow/ice*, has the least residual standard error and the highest R-squared with all p-values far below 0.05.

Note:
- For a variety of regression models with other combinations of independent variables we have explored and examined before attaining this model, please see the Appendix.
- We also tried the “parameter addition” approach (starting with a simple model) besides the “parameter elimination” approach. Both approaches converged to the similar result above.

Relationship Between *price* and *attendance*

```
> summary(lm(attendance~adult+school+temp+`snow/ice`))

Call:
lm(formula = attendance ~ adult + school + temp + `snow/ice`)

Residuals:
     Min       1Q   Median       3Q      Max
-1543.75  -754.01    36.21   568.32  2525.64

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16111.266   3519.166   4.578 2.41e-05 ***
adult       -2134.751    569.698  -3.747 0.000404 ***
school       -792.161     59.441 -13.327  < 2e-16 ***
temp           64.407      9.396   6.855 4.43e-09 ***
`snow/ice`  -3075.682    699.561  -4.397 4.57e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 904.8 on 60 degrees of freedom
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.873
F-statistic:   111 on 4 and 60 DF,  p-value: < 2.2e-16
```

- According to our forecasting model, we conclude that price and attendance are **negatively** correlated.
- Holding other variables constant, for every $1 increase in the price of adult ticket, we expect the attendance to decrease by 2135.
- This generally conforms to the economic principle of demand, in which demand drops as price goes up.

Appendix: Attempted Models During Exploration (a)

```
> model1 <- lm(attendance~adult)
> summary(model1)

Call:
lm(formula = attendance ~ adult)

Residuals:
   Min     1Q Median     3Q    Max
 -3689  -1520   -476    694   5118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    31221       7444   4.194 8.73e-05 ***
adult          -4638       1321  -3.511 0.000831 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2340 on 63 degrees of freedom
Multiple R-squared:  0.1636,    Adjusted R-squared:  0.1504
F-statistic: 12.33 on 1 and 63 DF,  p-value: 0.0008311
```

```
> model2 <- lm(attendance~adult+school)
> summary(model2)

Call:
lm(formula = attendance ~ adult + school)

Residuals:
    Min      1Q  Median      3Q     Max
-6248.2  -990.9   112.1  1047.1  2065.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27551.05    4618.26   5.966 1.27e-07 ***
adult       -3415.78     825.89  -4.136 0.000108 ***
school       -896.07      88.43 -10.134 8.91e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1447 on 62 degrees of freedom
Multiple R-squared:  0.6851,    Adjusted R-squared:  0.675
F-statistic: 67.46 on 2 and 62 DF,  p-value: 2.763e-16
```

```
> model3 <- lm(attendance~adult+`snow/ice`)
> summary(model3)

Call:
lm(formula = attendance ~ adult + `snow/ice`)

Residuals:
    Min      1Q  Median      3Q     Max
-3423.7 -1502.7  -299.7   636.3  4968.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    33017       7276   4.538 2.67e-05 ***
adult          -4937       1290  -3.827 0.000305 ***
`snow/ice`     -3592       1642  -2.188 0.032449 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2273 on 62 degrees of freedom
Multiple R-squared:  0.2236,    Adjusted R-squared:  0.1985
F-statistic: 8.928 on 2 and 62 DF,  p-value: 0.0003915
```

```
> model4 <- lm(attendance~school+temp)
> summary(model4)

Call:
lm(formula = attendance ~ school + temp)

Residuals:
    Min      1Q  Median      3Q     Max
-3234.9  -657.9   -22.9   416.7  3186.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2366.713    731.336   3.236  0.00195 **
school      -739.400     68.990 -10.717 9.58e-16 ***
temp          87.902      9.737   9.028 6.65e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1075 on 62 degrees of freedom
Multiple R-squared:  0.8264,    Adjusted R-squared:  0.8208
F-statistic: 147.6 on 2 and 62 DF,  p-value: < 2.2e-16
```

Appendix: Attempted Models During Exploration (b)

```
> model5 <- lm(attendance~school+adult)
> summary(model5)

Call:
lm(formula = attendance ~ school + adult)

Residuals:
    Min      1Q  Median      3Q     Max
-6248.2  -990.9   112.1  1047.1  2065.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27551.05    4618.26   5.966 1.27e-07 ***
school       -896.07      88.43 -10.134 8.91e-15 ***
adult       -3415.78     825.89  -4.136 0.000108 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1447 on 62 degrees of freedom
Multiple R-squared:  0.6851,    Adjusted R-squared:  0.675
F-statistic: 67.46 on 2 and 62 DF,  p-value: 2.763e-16
```

```
> model6 <- lm(attendance~school+temp+adult)
> summary(model6)

Call:
lm(formula = attendance ~ school + temp + adult)

Residuals:
    Min      1Q  Median      3Q     Max
-3645.4  -635.8     3.4   476.5  2949.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11847.43    3857.83   3.071  0.00318 **
school       -736.88      66.25 -11.122 2.61e-16 ***
temp           78.59      10.06   7.810 9.35e-11 ***
adult       -1583.99     633.78  -2.499  0.01515 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1032 on 61 degrees of freedom
Multiple R-squared:  0.8426,    Adjusted R-squared:  0.8348
F-statistic: 108.8 on 3 and 61 DF,  p-value: < 2.2e-16
```

```
> model7 <- lm(attendance~temp+adult)
> summary(model7)

Call:
lm(formula = attendance ~ temp + adult)

Residuals:
    Min      1Q  Median      3Q     Max
-3183.4  -949.7  -542.9  1052.1  5721.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7698.78    6627.49   1.162    0.250
temp          113.03      16.53   6.839 4.08e-09 ***
adult       -1691.28    1093.79  -1.546    0.127
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1781 on 62 degrees of freedom
Multiple R-squared:  0.5233,    Adjusted R-squared:  0.5079
F-statistic: 34.03 on 2 and 62 DF,  p-value: 1.062e-10
```

```
> model8 <- lm(attendance~school+adult)
> summary(model8)

Call:
lm(formula = attendance ~ school + adult)

Residuals:
    Min      1Q  Median      3Q     Max
-6248.2  -990.9   112.1  1047.1  2065.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27551.05    4618.26   5.966 1.27e-07 ***
school       -896.07      88.43 -10.134 8.91e-15 ***
adult       -3415.78     825.89  -4.136 0.000108 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1447 on 62 degrees of freedom
Multiple R-squared:  0.6851,    Adjusted R-squared:  0.675
F-statistic: 67.46 on 2 and 62 DF,  p-value: 2.763e-16
```

Thank you!