5240 Workshop 10

Digvijay Jondhale 0862899

1. Loading and Exploring data ....

```
library(dplyr)
```

```
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
```

```
awards <- read.csv(file='ws10-exercise-awards.csv')
head(awards,5)
```

```
   id num_awards prog math
1  45          0    3   41
2 108          0    1   41
3  15          0    3   44
4  67          0    3   42
5 153          0    3   40
```

```
str(awards)
```

```
'data.frame':   200 obs. of  4 variables:
 $ id        : int  45 108 15 67 153 51 164 133 2 53 ...
 $ num_awards: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prog      : int  3 1 3 3 3 1 3 3 3 3 ...
 $ math      : int  41 41 44 42 40 42 46 40 33 46 ...
```

```
summary(awards)
```

```
       id          num_awards        prog            math      
 Min.   :  1.00   Min.   :0.00   Min.   :1.000   Min.   :33.00  
 1st Qu.: 50.75   1st Qu.:0.00   1st Qu.:2.000   1st Qu.:45.00  
 Median :100.50   Median :0.00   Median :2.000   Median :52.00  
 Mean   :100.50   Mean   :0.63   Mean   :2.025   Mean   :52.65  
 3rd Qu.:150.25   3rd Qu.:1.00   3rd Qu.:2.250   3rd Qu.:59.00  
 Max.   :200.00   Max.   :6.00   Max.   :3.000   Max.   :75.00  
```

From the summary above, the data contained no outliers or extreme values for any of the columns.

2. Data Preprocessing

Conversion of Prog to a factor

```
awards <- awards %>%
  mutate(prog = factor(prog, levels = c(1,2,3),
                       labels = c("General","Academic","Vocational")))

str(awards)
```

```
'data.frame':   200 obs. of  4 variables:
 $ id        : int  45 108 15 67 153 51 164 133 2 53 ...
 $ num_awards: int  0 0 0 0 0 0 0 0 0 0 ...
 $ prog      : Factor w/ 3 levels "General","Academic",..: 3 1 3 3 3 1 3 3 3 3 ...
 $ math      : int  41 41 44 42 40 42 46 40 33 46 ...
```

Removing Null Values

```
colSums(is.na(awards))
```

```
        id num_awards        prog        math
         0          0           0           0
```

The data contained no Null values, but the prog variable was in integer which was converted to a suitable data type factor.

Data is ready for analysis !

3. Poisson regression model

```
poisson_dist <- glm(num_awards ~ prog + math,
                    family = poisson(link = "log"), data = awards)


summary(poisson_dist)
```

```
Call:
glm(formula = num_awards ~ prog + math, family = poisson(link = "log"),
    data = awards)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.24712    0.65845  -7.969 1.60e-15 ***
progAcademic    1.08386    0.35825   3.025  0.00248 **
progVocational  0.36981    0.44107   0.838  0.40179
math            0.07015    0.01060   6.619 3.63e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 287.67  on 199  degrees of freedom
Residual deviance: 189.45  on 196  degrees of freedom
AIC: 373.5

Number of Fisher Scoring iterations: 6
```

4. Summary

All the aboe programs (proAcademic,proVocational and math variable) are compared to General type program. From the test it was observed that There is no difference in number of awards won between proVocational($p = 0.40$) and proGeneral as $p > 0.05$, and they both have very low number of awards. Students who took proAcademic had a higher number of awards than General students. Also math score was a significant factor influencing the number of awards won.