Module Assignment
Module 8
QMB-6304 Foundations of Business Statistics

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

Preprocessing

1. Load the file "Module 8 Assignment Data.xlsx" into R. This file contains information on 20,807 employees of the City of Chicago who received overtime pay during 2016. This will be your master data set. Variables included are:
    a. department.name: A character variable identifying the department in which the employee works. This data set includes only the five largest (by headcount) in the City of Chicago government in 2016.
    b. employee.name: The name of the employee.
    c. title: the job title of the employee
    d. january through december: the amount of overtime pay the employee received in each of the 12 months in 2016.
    e. total: the total amount of overtime pay the employee received in 2016.
    f. nummos: the number of months in 2016 in which the employee received overtime pay.
    g. over5000: a binary variable indicating whether the employee earned $5000 or more in overtime pay during 2016.
2. Using the numerical portion of your U number as a random number seed, take a random sample of 4500 cases from the master data set using the method presented in class. This will be your primary data set for analysis.

```
#Carolina Aldana Yabur
#U25124553

#Preprocessing
rm(list=ls())
setwd("C:/Users/calda/Desktop")
library(rio)
chicago=import("6304 Module 8 Assignment Data.xlsx")
colnames(chicago)=tolower(make.names(colnames(chicago)))
> names(chicago)
 [1] "department.name" "employee.name"   "title"
 [4] "january"         "february"        "march"
 [7] "april"           "may"             "june"
[10] "july"            "august"          "september"
```

```
[13] "october"          "november"         "december"
[16] "total"            "nummos"           "over5000"
set.seed(25124553)
chicago.sample=chicago[sample(1:nrow(chicago),4500),]
attach(chicago.sample)
```

Analysis

Using your primary data set:

1. Show the results of the str() command.

```
> #Analysis
> #Part 1
> str(chicago.sample)
'data.frame':  4500 obs. of  18 variables:
 $ department.name: chr  "Police" "Aviation" "Water Management"
"Police" ...
 $ employee.name  : chr  "Cardona, Maribel" "Cuadro, Yolanda"
"Urian, Toribio J" "Griffin, Laura" ...
 $ title          : chr  "Police Officer" "Laborer" "Laborer -
Apprentice" "Sergeant" ...
 $ january        : num  495 1408 2384 617 2103 ...
 $ february       : num  NA 2743 256 1234 1165 ...
 $ march          : num  903 388 2171 1234 2721 ...
 $ april          : num  NA 518 1805 1234 3313 ...
 $ may            : num  NA NA NA 1234 1598 ...
 $ june           : num  495 NA NA 617 1964 ...
 $ july           : num  990 NA 617 1234 3003 ...
 $ august         : num  2999 NA 1000 1581 2297 ...
 $ september      : num  NA 940 1803 2674 3462 ...
 $ october        : num  NA 403 617 2872 3645 ...
 $ november       : num  NA 629 3028 3027 2530 ...
 $ december       : num  NA 495 2313 1889 2031 ...
 $ total          : num  5882 7525 15996 19445 29830 ...
 $ nummos         : num  5 8 10 12 12 7 3 9 1 3 ...
 $ over5000       : num  1 1 1 1 1 0 0 1 0 0 ...
```

2. Parameterize a logistic regression model with over5000 as the dependent and department.name and nummos independent variables. Report the results of the model using the summary() command.

```
> #Part 2
>
log_sample.out=glm(over5000~department.name+nummos,data=chicago.
sample,
+                   family=binomial)
> summary(log_sample.out)
```

```
Call:
glm(formula = over5000 ~ department.name + nummos, family =
binomial,
    data = chicago.sample)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2544  -0.4936   0.1746   0.4453   2.9497

Coefficients:
                                      Estimate Std. Error z
value Pr(>|z|)
(Intercept)                           -3.82755    0.19158 -
19.978  < 2e-16 ***
department.nameFire                    0.82482    0.18496
4.459 8.22e-06 ***
department.namePolice                 -0.28941    0.17848  -
1.622 0.104890
department.nameStreets and Sanitation -1.89192    0.21494  -
8.802  < 2e-16 ***
department.nameWater Management       -0.82239    0.23507  -
3.499 0.000468 ***
nummos                                 0.69112    0.01963
35.210  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6191.6  on 4499  degrees of freedom
Residual deviance: 3059.1  on 4494  degrees of freedom
AIC: 3071.1

Number of Fisher Scoring iterations: 6
```

3. State whether you believe the Residual Deviance of your model is markedly different from the Null Deviance.

While the Null Deviance represents how well the over5000 dependent variable can be predicted by the model with only an intercept, the Residual Deviance shows how well it can be predicted by the model including the independent variables department.name and nummos. In this case, the Null Deviance is 6191.6, while the Residual Deviance is 3059.1, resulting in a difference of 3132.5. This large difference suggests that the model with department.name and nummos has the highest likelihood of being able to correctly pick probabilities of events in the over5000 variable. Therefore, we can conclude that this model is a good fit.

4. Given your model from Part 2 and ignoring p values, which variable or variable/level will have the greatest influence in increasing the modeled probability that an employee earned $5000 or more in 2016?

Ignoring the p-values, the beta coefficient of the department.nameFire variable is the highest (0.82482), which means that this variable has a greater effect on increasing the probability of employees earning $5000 or more in 2016. Therefore, employees in the Fire department were more likely to earn $5000 or more in this year.

5. Given your model from Part 2 and ignoring p values, which variable will have the greatest influence in decreasing the modeled probability that an employee earned $5000 or more in 2016?

Ignoring the p-values, the beta coefficient of the department.nameStreets and Sanitation variable is the lowest (-1.89192), which means that this variable has the least effect on increasing the probability of employees earning $5000 or more in 2016. Therefore, employees in the Streets and Sanitation department were less likely to earn $5000 or more in this year.

6. Using the *expand.grid()* command develop a prediction file with all independent variables in the Step 1 model. For independent variables in this case use the *unique()* qualifier. R will by default calculate predicted probabilities to many decimal places. For convenience in reporting round your stored predictions to only 4 decimal places. Show the predicted probabilities for ONLY the first ten cases appearing in your prediction file.

```
> #Part 6
>
chicago.predictions=expand.grid(department.name=unique(chicago.s
ample$department.name),
+
nummos=unique(chicago.sample$nummos))
> chicago.predictions$pred_prob=round(
+   predict(log_sample.out,
+           newdata=chicago.predictions,
+           type="response"),4)
```

```
> top10=head(chicago.predictions,10)
> top10
          department.name nummos pred_prob
1                  Police      5    0.3404
2                Aviation      5    0.4081
3         Water Management      5    0.2325
4  Streets and Sanitation      5    0.0942
5                    Fire      5    0.6113
6                  Police      8    0.8041
7                Aviation      8    0.8457
8         Water Management      8    0.7066
9  Streets and Sanitation      8    0.4525
10                   Fire      8    0.9260
```

7. Based on your predictions generated in Step 6, find the maximum and minimum predicted probabilities generated. State the values of the independent variables for these max and min cases.

```
> #Part 7
> max(top10$pred_prob)
[1] 0.926
> which.max(top10$pred_prob)
[1] 10
> top10[which.max(top10$pred_prob),c(1,2,3)]
   department.name nummos pred_prob
10            Fire      8     0.926
>
> min(top10$pred_prob)
[1] 0.0942
> which.min(top10$pred_prob)
[1] 4
> top10[which.min(top10$pred_prob),c(1,2,3)]
         department.name nummos pred_prob
4 Streets and Sanitation      5    0.0942
```

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is "Assignment 5" of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas.

This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.