Occipital Pressure Ulcer Analysis

After filtering the NIS dataset down to only those who have been diagnosed with one of the ICD-10 codes corresponding to occipital pressure ulcers (the I10\_DX[NUM] variables), further analysis is now ready.

I recoded the categorical variables to be factors in an effort to make things easier to interpret.

The data is restricted to only those patients who have been diagnosed with an occipital pressure ulcer at least once.

**EDA**

We are interested in finding the distribution of the patients who have been diagnosed with occipital pressure ulcers, so initial EDA will be performed on the dataset. First, let’s see how many people were diagnosed with occipital pressure ulcers:

## Number of observations: 2486

As we can see, there are 2486 people who were diagnosed with any stage of the ulcer at least one time.

How many are in any given stage?

occipital pressure Ulcers by Stage

|  |  |  |
| --- | --- | --- |
| STAGE | NUMBER OF OBS | PERCENT |
| unstageable | 503 | 19.80315 |
| stage 1 | 336 | 13.22835 |
| stage 2 | 621 | 24.44882 |
| stage 3 | 287 | 11.29921 |
| stage 4 | 130 | 5.11811 |
| unspecified | 663 | 26.10236 |

What are the counts by Sex? Race? (note that Native Americans are pooled in with “other” due to a small sample size)

occipital pressure Ulcers by Sex

|  |  |  |
| --- | --- | --- |
| SEX | NUMBER OF OBS | PERCENT |
| Male | 1393 | 54.842520 |
| Female | 908 | 35.748032 |
| NA | 239 | 9.409449 |

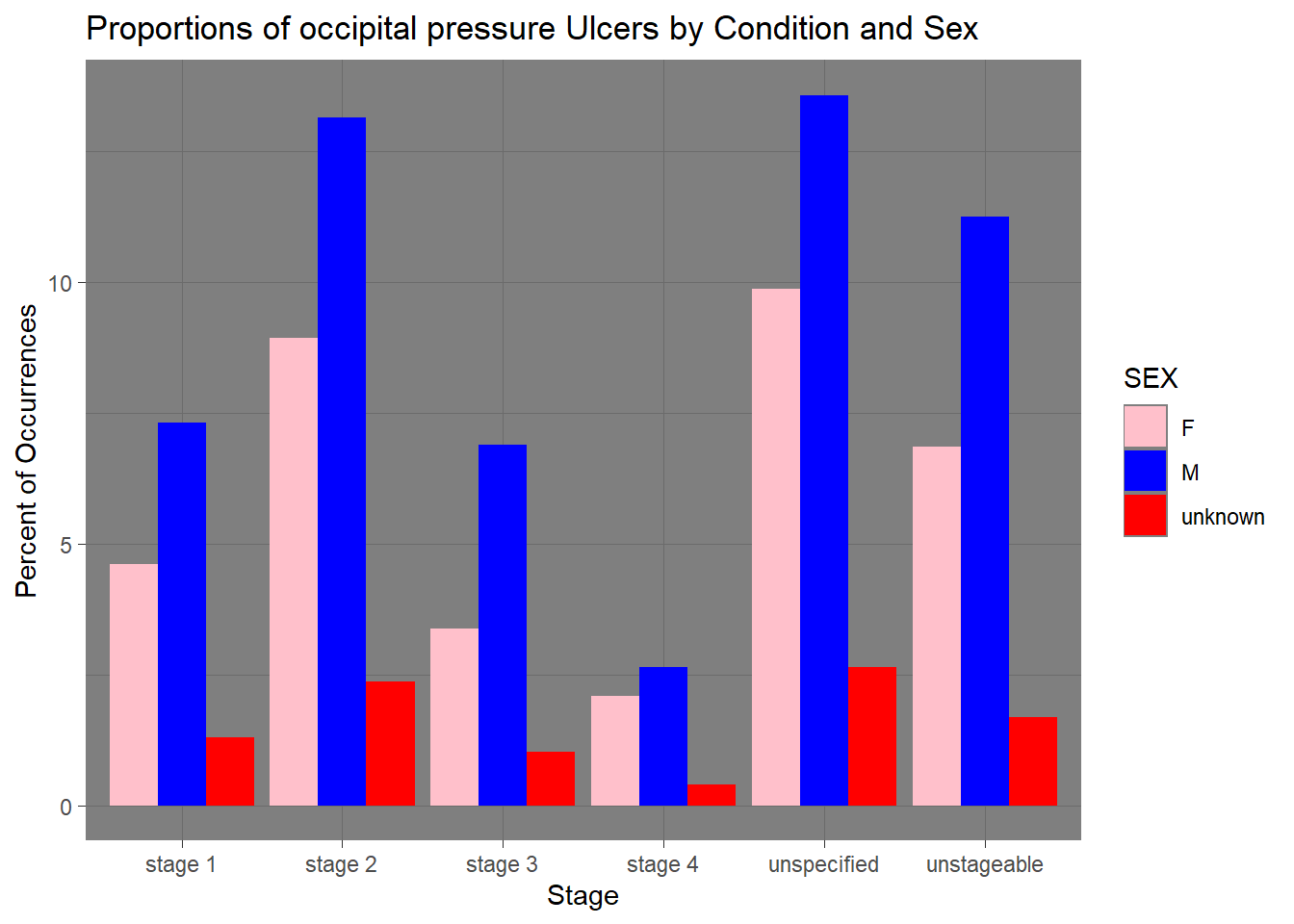
|  |  |  |
| --- | --- | --- |
| RACE | NUMBER OF OBS | PERCENT |
| White | 1482 | 58.346457 |
| unknown | 368 | 14.488189 |
| Black | 319 | 12.559055 |
| Hispanic | 229 | 9.015748 |
| Other | 81 | 3.188976 |
| Asian | 61 | 2.401575 |

How many people do not have another condition?

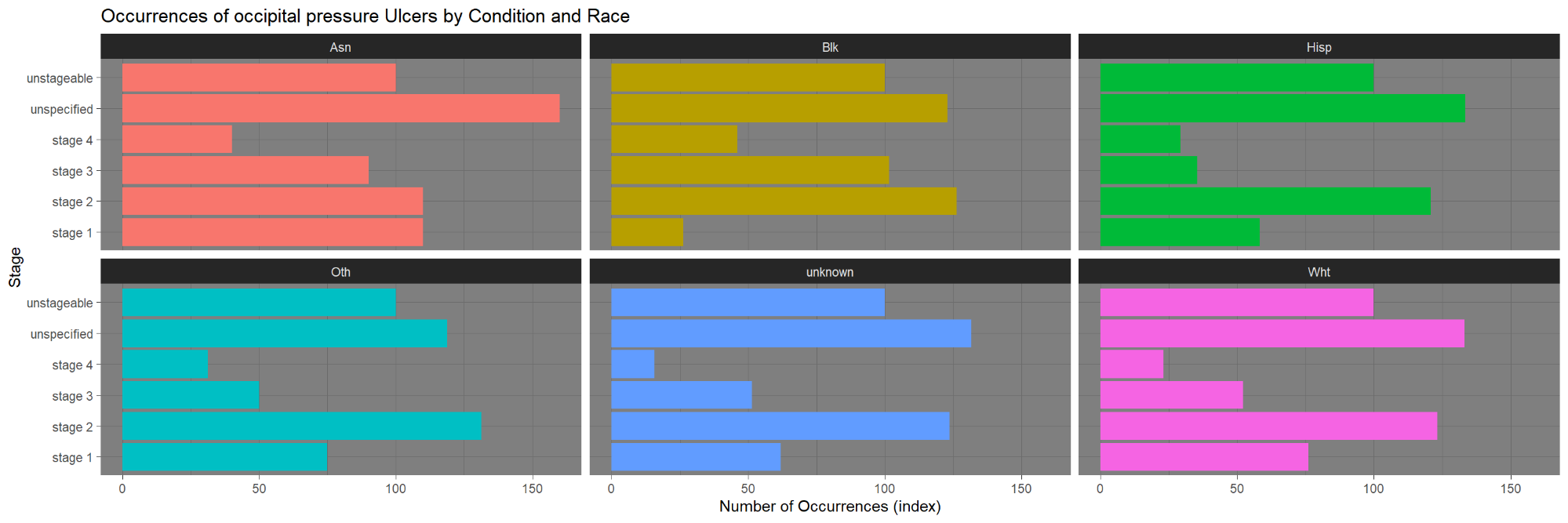
## Number of people who do not have another condition: 54

Let’s break things down by stage…

The percent of occurrences by sex:



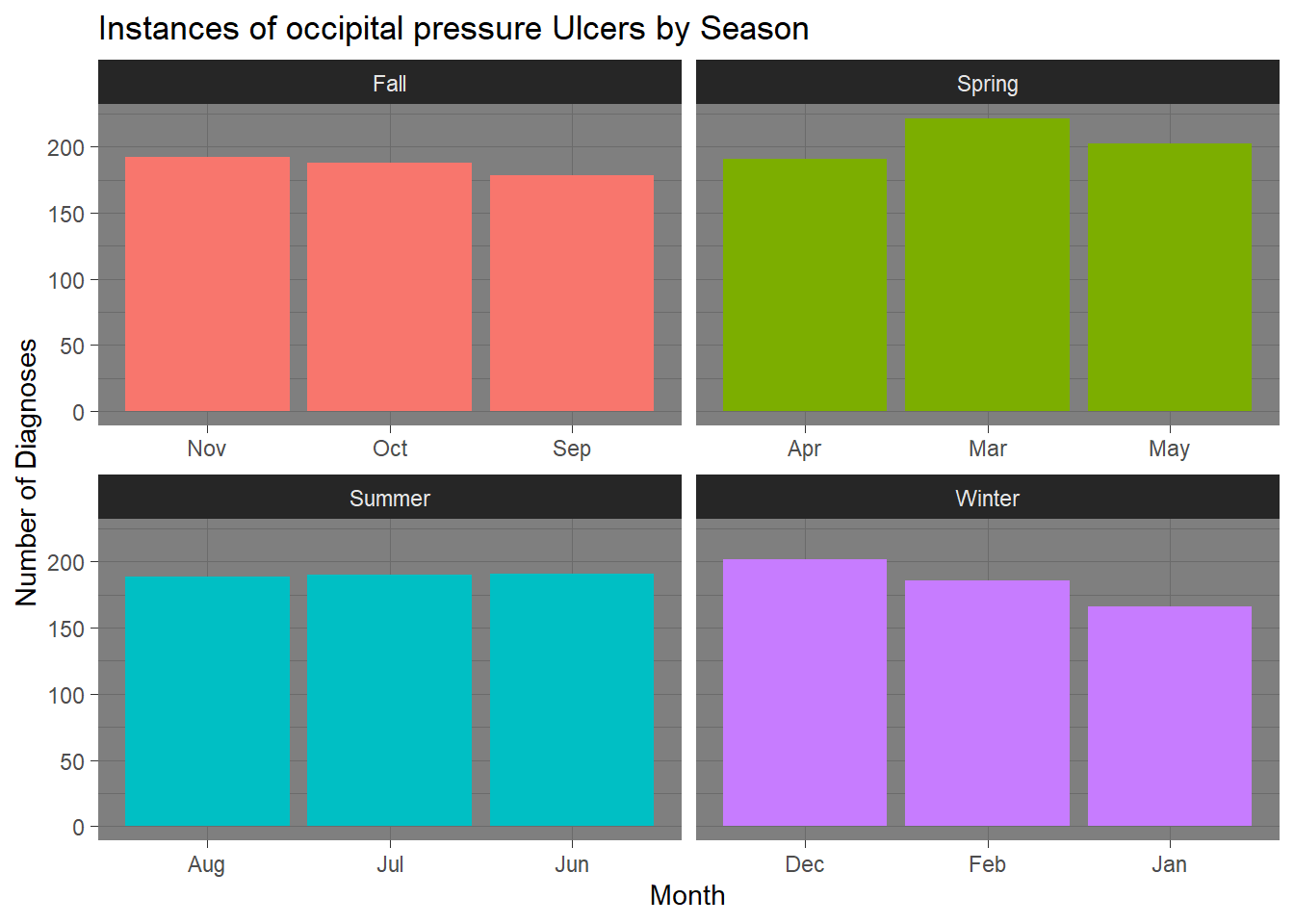
And now race:



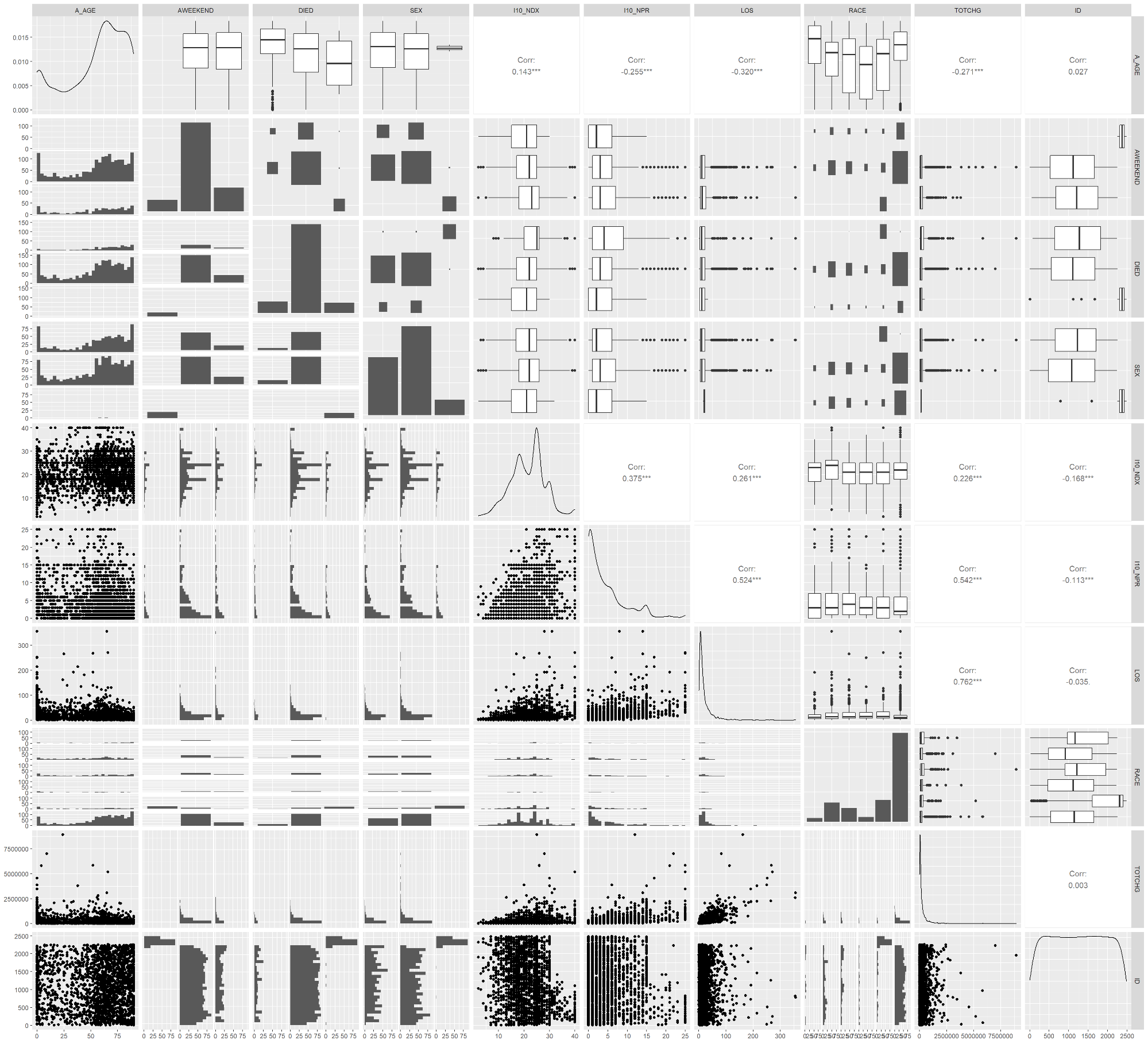
The number of occurrences were put on an index to more easily compare between groups. The index allows us to ignore differences in absolute numbers (i.e., there are more white patients in the country, so of course they will have more occurrences. As we can see, most differences between races are noticeable but not extreme. Remember this, as it is going to come up again in a few moments.

Now, let’s see how diagnoses of occipital pressure ulcers change over time. Is the diagnosis of an occipital pressure ulcer seasonal?

nis\_long %>%   
 filter(STAGE %in% ulcer\_codes,  
 AMONTH != "unknown") %>%   
 count(AMONTH, SEASON) %>%  
 ggplot() +  
 geom\_col(mapping = aes(x = AMONTH, y = n, fill = SEASON), show.legend = F) +  
 facet\_wrap(~ SEASON, scales = "free\_x") +  
 ggtitle("Instances of occipital pressure Ulcers by Season") +  
 xlab("Month") +  
 ylab("Number of Diagnoses") +  
 theme\_dark()



Diagnosis of occipital pressure ulcers does not seem to have a seasonal component. Let’s see how a combination of variables relate to one another.



**Logistic Regression**

Packages and Engine

The rsample package will be used to pull samples with stratified sampling. The stats package’s glm() function will be the engine through which the logistic regression will run. The pscl package’s pR2() function will be used to calculate McFadden’s R-squared. To check for multicollinearity, the car package will be used to calculate the variance inflation factor (VIF). To create the ROC curve and confusion matrices, the InformationValue package will be used.

Sampling

I did the cleaning in the EDA section and with Spark, so we can jump straight into sampling. Stratification on the RACE variable will be pursued.

# stratified sampling by RACE  
# 80 percent in train, 20 percent in test  
  
# setting the random number generator  
set.seed(2021)  
nis\_split <- rsample::initial\_split(nis\_long %>%   
 mutate(DIED = as.character(DIED)) %>%   
 filter(!is.na(DIED),  
 STAGE %in% ulcer\_codes,  
 DIED != "unknown") %>%   
 mutate(DIED = as.factor(DIED)),  
 prob = 0.8,  
 strata = "RACE")  
  
# Splitting into training and test sets  
nis\_train <- rsample::training(nis\_split)  
nis\_test <- rsample::testing(nis\_split)

Models

I will try several different logistic regression models on the data.

logit1 <- glm(data = nis\_train,  
 formula = DIED ~ A\_AGE + SEX + RACE + LOS,  
 family = "binomial")  
  
logit2 <- glm(data = nis\_train,  
 formula = DIED ~ SEASON + ZIPINC\_QRTL + HCUP\_ED + I10\_NPR,  
 family = "binomial")

Evaluation

Summaries

The results from ‘logit1’ are as follows:

##   
## Call:  
## glm(formula = DIED ~ A\_AGE + SEX + RACE + LOS, family = "binomial",   
## data = nis\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6416 0.3460 0.4859 0.5781 1.2135   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.205945 0.669305 6.284 3.3e-10 \*\*\*  
## A\_AGE -0.021152 0.003499 -6.044 1.5e-09 \*\*\*  
## SEXM -0.153082 0.153644 -0.996 0.31909   
## SEXunknown 10.428574 324.743727 0.032 0.97438   
## RACEBlk -0.945117 0.634673 -1.489 0.13645   
## RACEHisp -1.079835 0.643328 -1.679 0.09325 .   
## RACEOth -0.891753 0.731146 -1.220 0.22259   
## RACEunknown -0.925393 0.677735 -1.365 0.17212   
## RACEWht -0.646113 0.611197 -1.057 0.29045   
## LOS -0.007266 0.002285 -3.181 0.00147 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1304.9 on 1723 degrees of freedom  
## Residual deviance: 1257.0 on 1714 degrees of freedom  
## (1 observation deleted due to missingness)  
## AIC: 1277  
##   
## Number of Fisher Scoring iterations: 11

Only age and length of stay are significant. Interestingly, the demographic variables are not significant.

The results from ‘logit2’ are as follows:

##   
## Call:  
## glm(formula = DIED ~ SEASON + ZIPINC\_QRTL + HCUP\_ED + I10\_NPR,   
## family = "binomial", data = nis\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3653 0.4166 0.4780 0.5314 1.0192   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.54297 0.21834 11.647 < 2e-16 \*\*\*  
## SEASONSpring -0.09186 0.20733 -0.443 0.6577   
## SEASONSummer 0.19157 0.22278 0.860 0.3898   
## SEASONunknown 11.79471 377.76913 0.031 0.9751   
## SEASONWinter -0.19621 0.20576 -0.954 0.3403   
## ZIPINC\_QRTLtop 50 -0.02050 0.14909 -0.138 0.8906   
## ZIPINC\_QRTLunknown -0.38293 0.50283 -0.762 0.4463   
## HCUP\_ED1 -0.41077 0.17075 -2.406 0.0161 \*   
## HCUP\_ED2 -0.28647 0.23735 -1.207 0.2275   
## HCUP\_ED4 -1.20461 0.28689 -4.199 2.68e-05 \*\*\*  
## I10\_NPR -0.05054 0.01263 -4.002 6.28e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1305.2 on 1724 degrees of freedom  
## Residual deviance: 1270.7 on 1714 degrees of freedom  
## AIC: 1292.7  
##   
## Number of Fisher Scoring iterations: 12

The Emergency Department codes, along with some of the Emergency Department codes, are significant.

One of the most important things I learned about modeling is that proving something is only marginally helpful–disproving is powerful. I think we may be able to rule out a seasonal component to instances of occipital pressure ulcers in hospital patients. Similarly, income levels–broken down into two groups [bottom 50th percentile, top 50th percentile]–are not significant. Income levels do not improve the survival rate of occipital pressure ulcers.

McFadden’s R-square

We can calculate McFadden’s pseudo R-squared for each model to evaluate the goodness of fit. In general, a pseudo R-squared greater than 0.4 indicates a good fit.

Let’s begin with ‘logit1’:

## fitting null model for pseudo-r2  
## 0.03675786

‘logit2’?

## fitting null model for pseudo-r2  
## 0.0264248

These models perform poorly… Let’s check the VIF scores. Maybe multicollinearity is an issue?

Variance Inflation Factor

We can calculate the VIF of each predictor to see if multicollinearity is a problem. VIF scores range from 1 to infinity. A score of 1 indicates zero multicollinearity and scores greater than 5 indicate multicollinearity is an issue that needs to be addressed.

## GVIF Df GVIF^(1/(2\*Df))  
## A\_AGE 1.174299 1 1.083651  
## SEX 1.014739 2 1.003664  
## RACE 1.065084 5 1.006325  
## LOS 1.143490 1 1.069341

Excellent, the addition of new variables has little effect on the significance of the predictors.

‘logit2’?

## GVIF Df GVIF^(1/(2\*Df))  
## SEASON 1.021274 4 1.002635  
## ZIPINC\_QRTL 1.011055 2 1.002752  
## HCUP\_ED 1.057347 3 1.009337  
## I10\_NPR 1.054752 1 1.027011

multicollinearity does not seem to be an issue with the second logistic regression.

Test Data

Finally, let’s check the model’s accuracy on the test sets. First the probabilities will need to be calculated for each variable.

logit1\_test\_probs <- predict(logit1,  
 nis\_test,  
 type = "response")  
  
logit2\_test\_probs <- predict(logit2,  
 nis\_test,  
 type = "response")

We can make a confusion matrix to find the sensitivity and specificity. The threshold for determining whether someone dies is 0.5–anyone with a probability >= 0.5 is classified as dead and alive otherwise.

The confusion matrix for ‘logit1’:

## died survived  
## 1 53 521

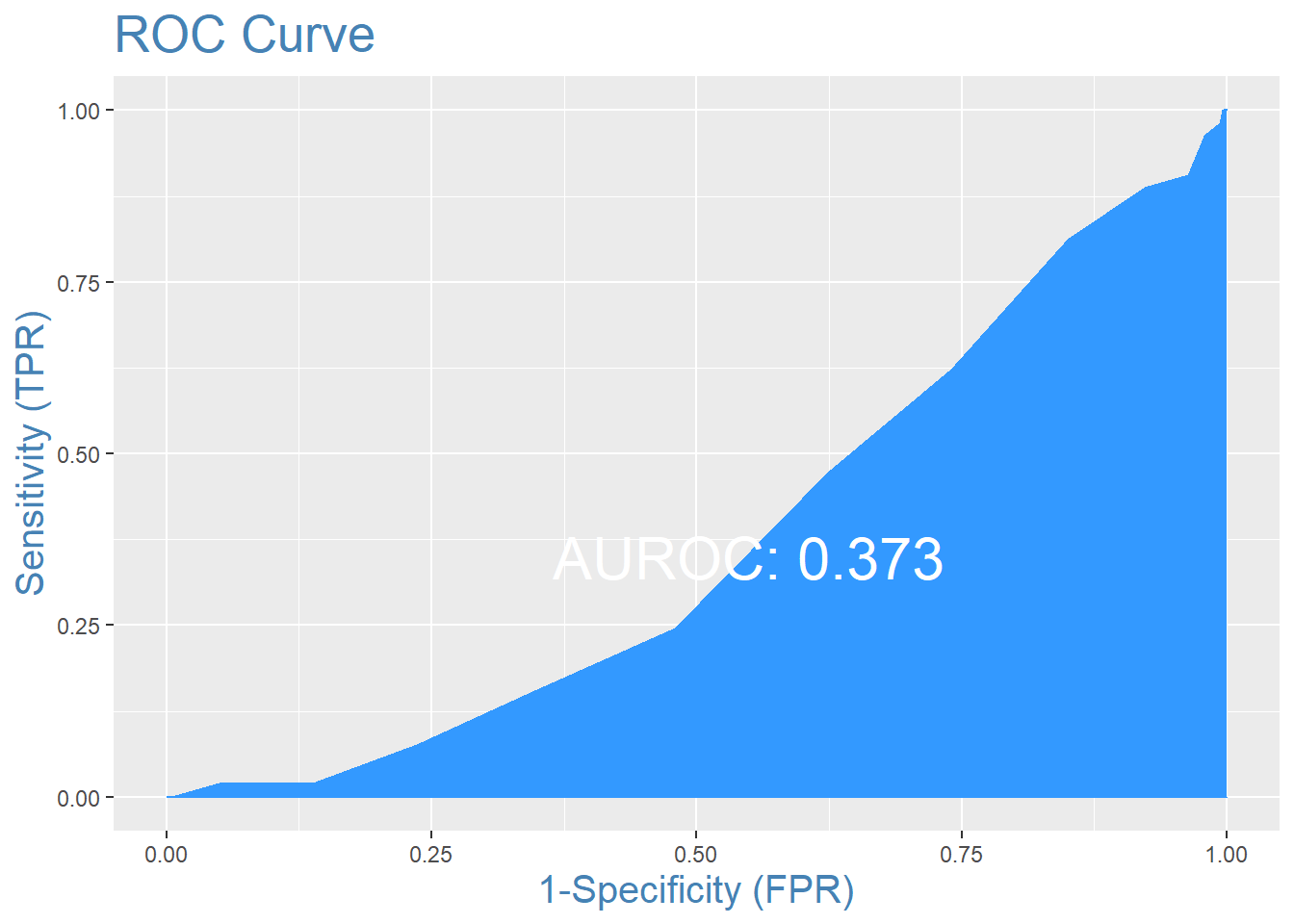
The confusion matrix for ‘logit2’:

## died survived  
## 1 53 521

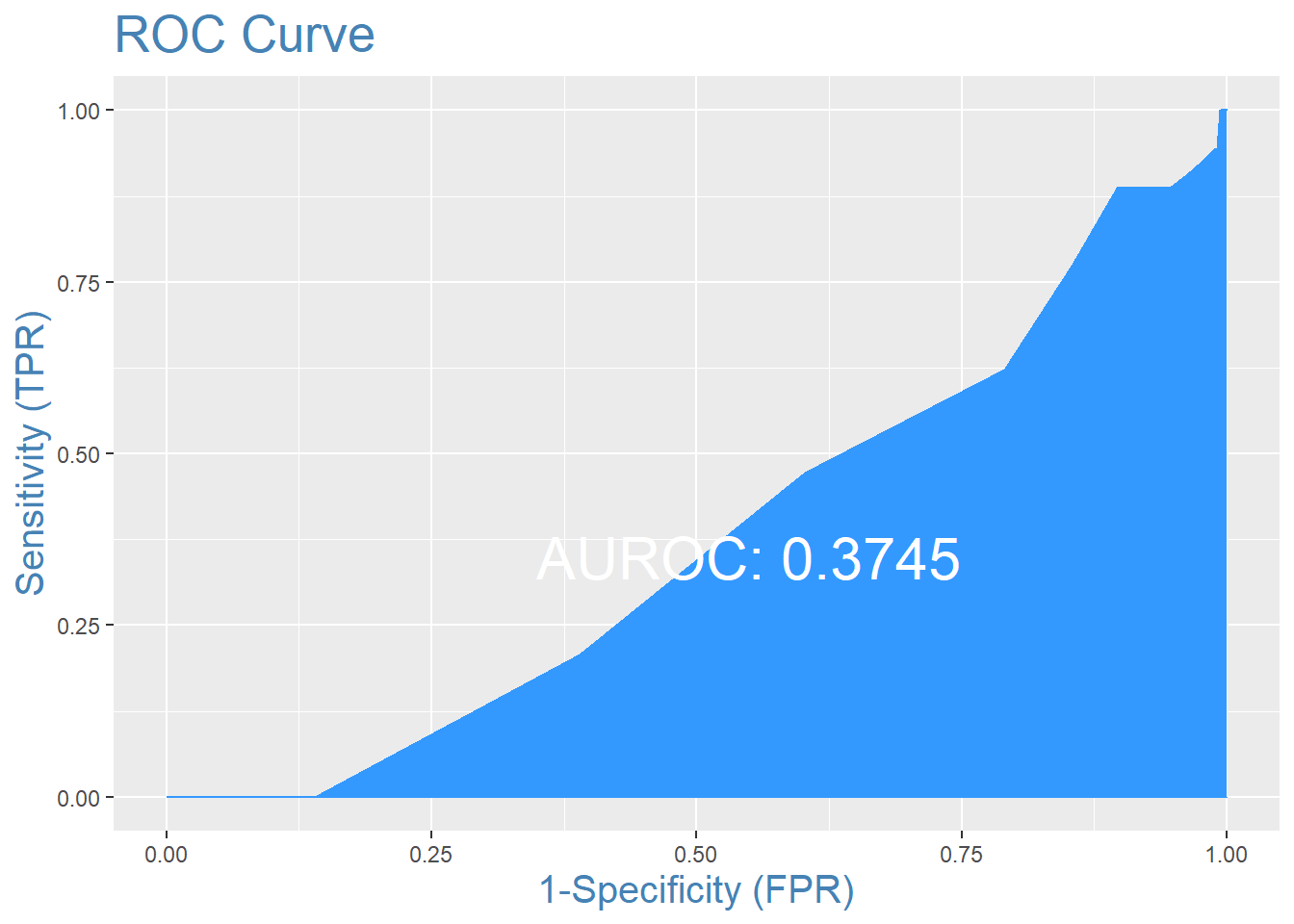
Neither of the models have any misclassifications. In case you missed it, missing values (where there was no ‘DIED’ value provided) were omitted from analysis. Since there are few significant variables, much of these predictions are likely due to overfitting.

ROC Curves

The ROC curve for ‘logit1’:



The ROC curve for ‘logit2’:



Optimization

Clearly, the two previous models were not nearly sufficient to prove anything, although they allowed us to disprove some important ideas–namely that there is a statistically significant difference between the survival rate between different racial groups, although different races are not diagnosed with different stages of occipital pressure ulcers uniformly.

The following model includes only significant variables:

logit3 <- glm(data = nis\_train,  
 formula = DIED ~ A\_AGE + I10\_NPR + TOTCHG,  
 family = "binomial")  
  
summary(logit3)

##   
## Call:  
## glm(formula = DIED ~ A\_AGE + I10\_NPR + TOTCHG, family = "binomial",   
## data = nis\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7025 0.3058 0.4637 0.5543 1.8529   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.843e+00 2.830e-01 13.580 < 2e-16 \*\*\*  
## A\_AGE -2.405e-02 3.704e-03 -6.492 8.45e-11 \*\*\*  
## I10\_NPR -4.899e-02 1.519e-02 -3.225 0.00126 \*\*   
## TOTCHG -4.921e-07 1.542e-07 -3.191 0.00142 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1246.8 on 1696 degrees of freedom  
## Residual deviance: 1176.0 on 1693 degrees of freedom  
## (28 observations deleted due to missingness)  
## AIC: 1184  
##   
## Number of Fisher Scoring iterations: 5

car::vif(logit3)

## A\_AGE I10\_NPR TOTCHG   
## 1.166809 1.467363 1.494038

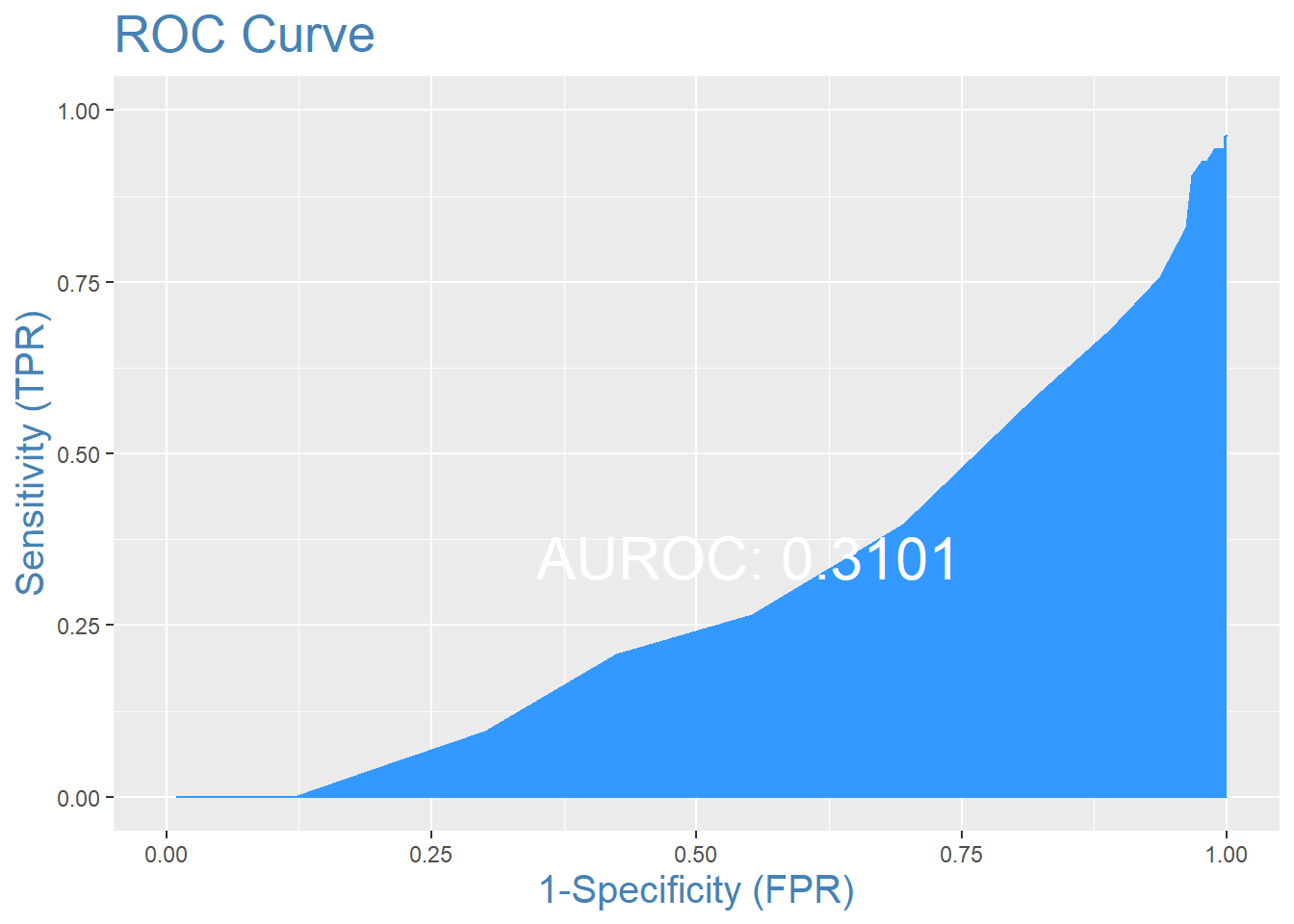
cat(pscl::pR2(logit3)["McFadden"])

## fitting null model for pseudo-r2  
## 0.05675355

logit3\_test\_probs <- predict(logit3,  
 nis\_test,  
 type = "response")  
InformationValue::confusionMatrix(nis\_test$DIED, logit3\_test\_probs, threshold = 0.5)

## died survived  
## 0 0 1  
## 1 51 515

InformationValue::plotROC(survived, logit3\_test\_probs)



Perhaps there is something else that can be done, but I am unsure of what to try. Interaction terms, transformation of the numerical variables, and creation of related variables (such as ‘SEASON’) were all tried. The pseudo R-squared and ROC curves are not very promising.

*Insignificant variables: PAY1, AMONTH, AWEEKEND, SEX, PL\_NCHS, STAGE, ZIPINC\_QRTL, TRAN\_IN, TRAN\_OUT, YEAR, DIAGNOSIS\_NUM, SEASON, RACE.*