**Fall**

16

**UMBC**

Dialysis Facilities Mortality Rate Prediction

08

**Fall**

**Project Name: Facilities’ Mortality Rate Prediction**

**Data Set Location (URL):**<https://data.medicare.gov/data/dialysis-facility-compare>

**n:** **6404**

**p: ~10**

**Response Variable:** *Standardized Mortality Ratio (SMR)*

**Predictor Variables:**

|  |  |
| --- | --- |
| Patient.survival.category.text | Patient survival category (Better, Worse, or As Expected) |
| Rate.of.hospital.readmission.category.text | Patient Readmission Category |
| Rate.of.hospital.readmission.data.availability.code | Lists whether the facility had sufficient readmission data available or the reason for why the data is not available. |
| Lists.the.facility.s.standardized.transfusion.ratio..facility. | Lists the facility’s Standardized Transfusion Ratio (FACILITY). |
| Offers.in.center.peritoneal.dialysis | Indicates whether the facility offers in-center peritoneal dialysis. |
| Offers.home.hemodialysis.training | Indicates whether the facility offers home hemodialysis |
| Standardized.readmission.ratio | Lists the facility’s Standardized Readmission Ratio (FACILITY). |
| Standardized.hospitalization.ratio | Lists the facility’s Standardized Hospitalization Ratio (FACILITY). |
| Hypercalcemia.data.availability.code | Lists whether the facility had sufficient hypercalcemia data available or the reason for why the data is not available. |
| Patient.hospitalization.category.text | Patient hospitalization category. |

**Descriptive Analysis:**

*Summary Statistics, Histograms/Barcharts*

**1. Patient.survival.category.text**

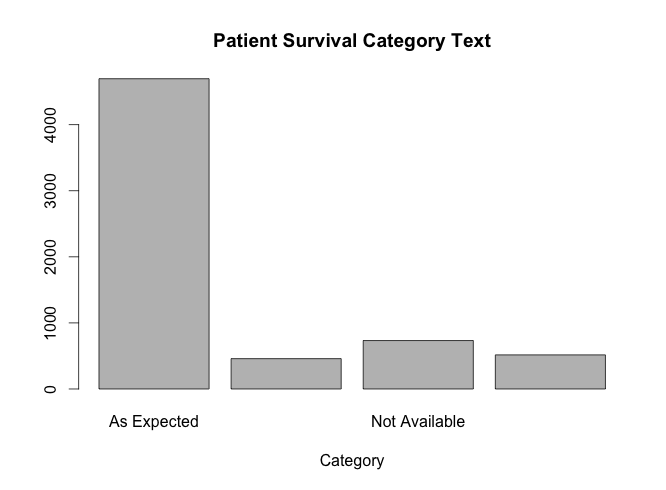
As Expected :4696

Better than Expected: 459

Not Available : 733

Worse than Expected : 516

Standard Deviation: 0.9801666



**2. Rate.of.hospital.readmission.category.text**

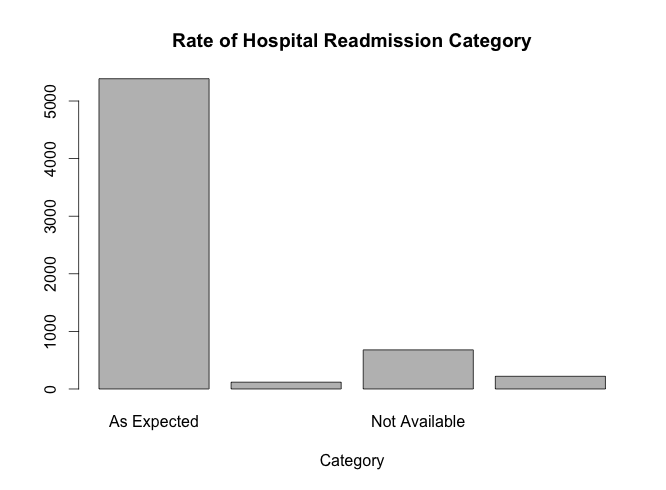
As Expected :5387

Better than Expected: 118

Not Available : 679

Worse than Expected : 220

Standard Deviation: 0.8003548



**3. Rate.of.hospital.readmission.data.availability.code**

Min. : 1.00

1st Qu.: 1.00

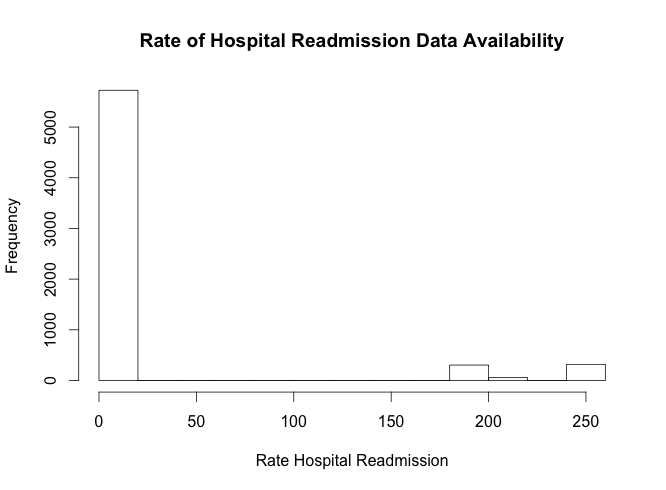
Median : 1.00

Mean : 24.93

3rd Qu.: 1.00

Max. :258.00

Standard Deviation: 70.14821



**4.Patient.hospitalization.category.text**

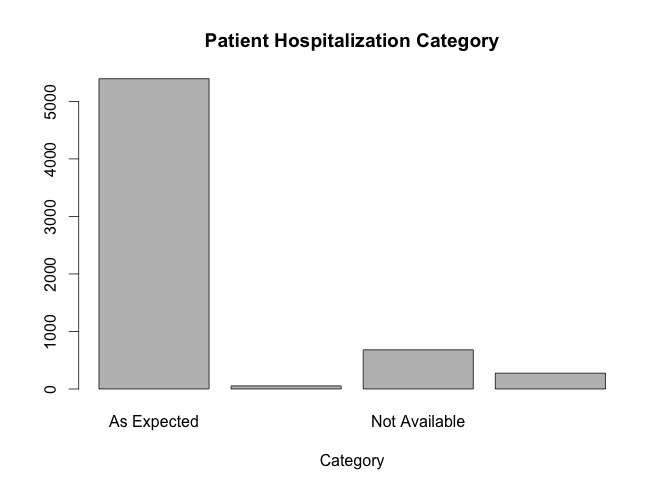
As Expected :5394

Better than Expected: 54

Not Available : 681

Worse than Expected : 275

Standard Deviation: 0.8354168

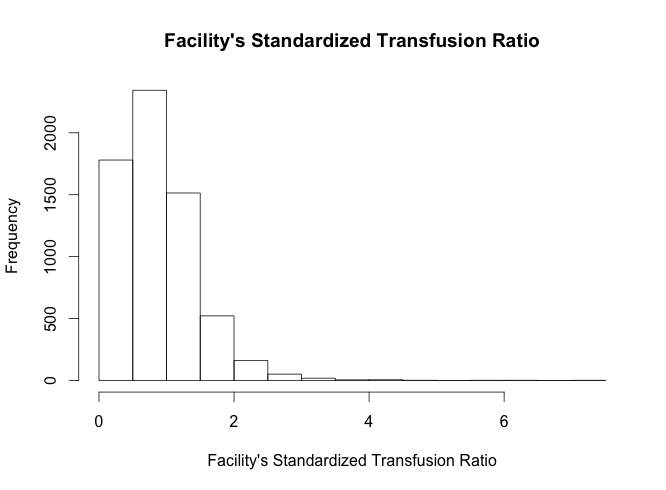


**5.Lists.the.facility.s.standardized.transfusion.ratio..facility.**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.4500 0.8100 0.8382 1.1700 7.1500

Standard Deviation: 0.6181983

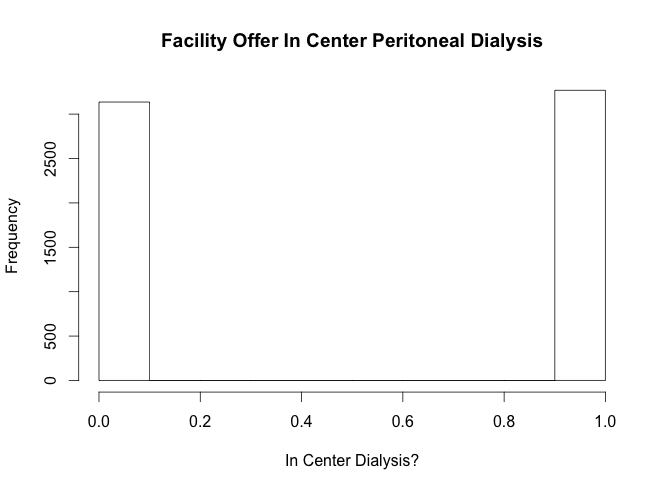


**6.Offers.in.center.peritoneal.dialysis**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 1.0000 0.5103 1.0000 1.0000

Standard Deviation: 0.4999328

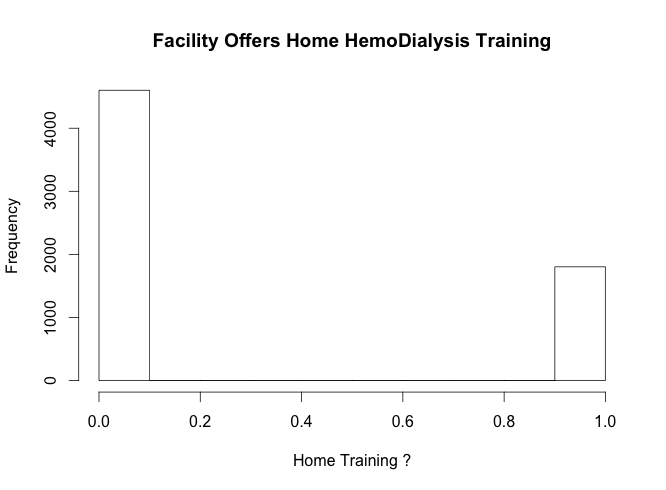


**7. Offers.home.hemodialysis.training**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 0.0000 0.2815 1.0000 1.0000

Standard Deviation: 0.4497867

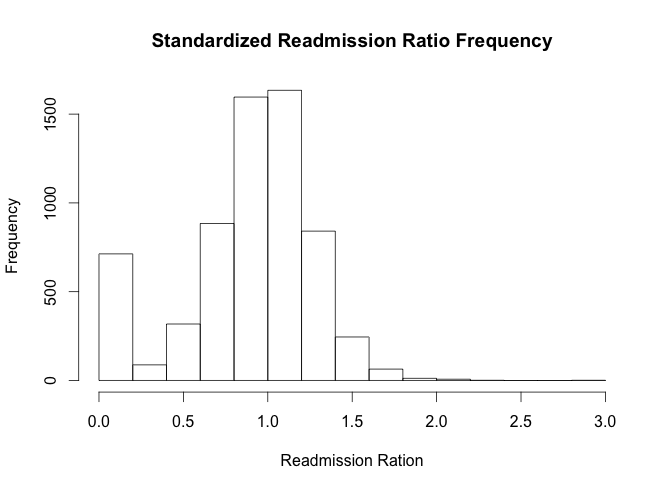


**8- Standardized.readmission.ratio**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.7300 0.9600 0.8862 1.1400 2.8600

Standard Deviation: 0.4018263

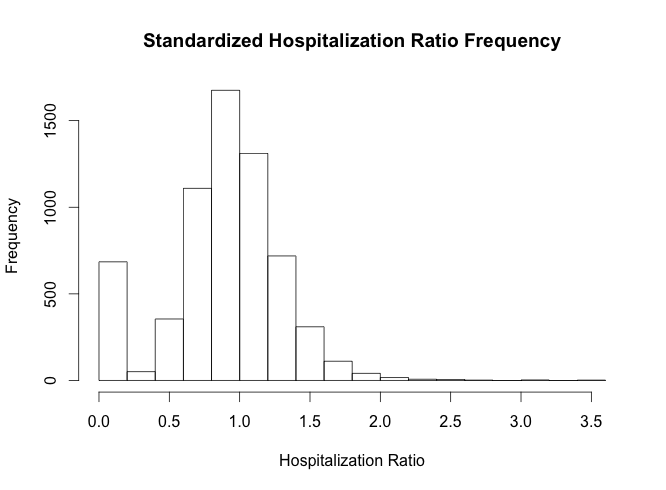


**9- Standardized.hospitalization.ratio**

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.7200 0.9200 0.8906 1.1400 3.4800

Standard Deviation: 0.4219122

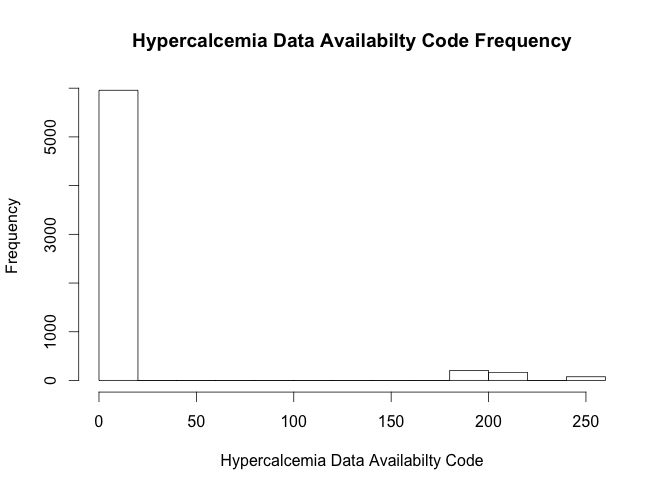


**10 - Hypercalcemia.data.availability.code**

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 1.00 1.00 15.62 1.00 258.00

Standard Deviation: 53.636



**Analysis Plan:**

**I - Variable Selection:**

Throughout the deliverable, we went over the variables selection process. We evaluated each variable to the response variable (Mortality Rate) and ranked the P-values obtained for each variable. We picked the variables with the lowest P-values as they showed how connected or related they were to our response variables. We picked 10 lowest P-values variables for future analysis. To choose the predictors, we used the “lm” function as follow:

> lm.fit = lm(Standardized.mortalization.ratio ~ Patient.survival.category.text , data = MYDASET)

> lm.fit

> attach(MYDATASET)

> summary(lm.fit)

By running the above function on each predictor against our response variable, we were able to get the p-value of the predictor and rank them accordingly.

Once the predictors obtained, we went ahead to give the descriptive statistics of our variables. Running summary(MYDATASET$Predictor) was useful to get the minimum, maximum, mean, median values of each predictors. The 1st and 3rd quartile values were also retrieved by that function. To see the disparity of values among each other, we calculated the Standard deviation of each predictor by running SD(MYDATASET$Predictor) .

Once the descriptive statistics were retrieved, we ran the histograms and barplots of our predictors. Looking at the histogram is helpful for our analysis as it shows us how our data is distributed for each predictor. Looking at those graphs can give us a initial judgment about of future predictive analysis.

**II - Initial Variable Assessment:**

After selecting the variables for our analysis, We made an initial assessment of our predictors to our response variable that we discuss hereafter:

* **Patient survival category:** This predictor is strongly related to our response variable as if a facility report a patient survival category of worse that national average, it is fair to say it is a good indicator of the mortality rate
* **Patient Readmission:** this predictor shows us that the mortality rate is somewhat related to the level at which a facility readmit old patients who are reinfected.
* f**acility’s Standardized Transfusion Ratio:** The analysis shows us that the mortality rate of also related to that facility standardized transfusion ratio. The poorer the transfusion ratio, the greater the mortality rate in that facility.
* **Offers.in.center.peritoneal.dialysis:** This predictor shows us that facilities that offer in center peritoneal dialysis are somewhat related to the mortality rate of those facilities.
* **Offers.home.hemodialysis.training:** This predictors tells us that the mortality rates is connected to the fact that a facility offer home hemodialysis or not.
* **Standardized Hospitalization Ratio:** This predictor shows us that the hospitalization ratio of a given facility is strongly related to the mortality rate observed in that facility.
* **Hypercalcemia.data.availability:** This predictor tells us if the mortality rate in a given facility is related to the availability of Hypercalcemia data. In other words, we are interested in finding out if the lack of Hypercalcemia data is related to the high mortality rate.

**III - Future Work**

Going forward, we are planning on using the variables selected above to build a model that will be helpful in predicting the mortality rate of a dialysis facility given some additional information about that facility. We will proceed by cleaning up our dataset to just include the variables we selected, and think about how we will treat missing data in our dataset. We will decide whether we have to delete the rows with missing information or replace it with default value. After that process, we will start building a simple linear regression, then many multivariate linear regressions to find out which variables give us the best prediction. We will test our models on some testing dataset that we will obtain from our original dataset. As the course progresses, we will also build some logistic models as well as doing some bootstrapping and measure the performance of each model on a gain chart and select the model with the optimal prediction.