**Abstract:**

**Introduction:**

**Methods:** The Data Analytics group from Good Health Corporation collected a total of 3682 records from 3612. Multiple linear regression was utilized to develop a model to predict length of stay using a variety of predictors.

Prior to any modeling, the data was examined through descriptive statistics and plots. This provides an idea of what they data is like and to help to identify any potential errors in the data or any necessary transformations. For categorical variables, the proportion of subjects in each category was examined and several variables were condensed into a smaller number of categories that were more evenly distributed. For example, race was collapsed into white vs non-white as many of the racial groups had low proportions of individuals. With this recategorizing, it is now much more even at 43% to 57%.

Looking at the descriptive statistics showed that there were in fact, some problematic values in the data. For example, there was an individual with a body temperature of 52.3° C, which is not biologically impossible. These values were removed from the dataset before proceeding with analysis.

As a first step in determining a suitable linear regression model for length of stay, we used automatic variable selection in SAS. Using forward, backward, and stepwise selection with an entry or removal significance level of 0.15, we obtained the same model.

**Results and Figures:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **n** | **mean** | **min** | **max** | **sd** |
| **visitid** | 3612 | 2.053175e+07 | 1.873718e+07 | 2.378690e+07 | 1.238424e+06 |
| **loshours** | 3612 | 1.311442e+02 | 1.000000e+00 | 2.111000e+03 | 1.421069e+02 |
| **losdays2** | 3612 | 5.464343e+00 | 4.166667e-02 | 8.795833e+01 | 5.921120e+00 |
| **mews\*** | 3449 | 4.107567e+00 | 1.000000e+00 | 1.400000e+01 | 1.700714e+00 |
| **ageyear** | 3612 | 6.568688e+01 | 1.800000e+01 | 1.050000e+02 | 1.869042e+01 |
| **bmi** | 2927 | 2.834929e+01 | 3.100000e+00 | 1.226500e+02 | 7.993252e+00 |
| **bpsystolic** | 3607 | 1.305527e+02 | 8.878261e+01 | 1.939630e+02 | 1.672157e+01 |
| **o2sat** | 3609 | 9.786086e+01 | 8.000000e+01 | 2.365263e+02 | 4.908312e+00 |
| **temperature** | 3610 | 3.673000e+01 | 1.185000e+01 | 5.227500e+01 | 8.994875e-01 |
| **heartrate** | 3607 | 8.007110e+01 | 3.758333e+01 | 2.425833e+02 | 1.300406e+01 |
| **respirationrate** | 3609 | 1.819603e+01 | 1.200000e+01 | 6.771795e+01 | 2.633549e+00 |
| **bpdiastolic** | 3611 | 7.251746e+01 | 2.956349e+01 | 1.544000e+02 | 9.801809e+00 |

Table 1: Descriptive Statistics of some key variables from the dataset including the number of observations, the mean, minimum, and maximum.

**Conclusion:**