SAKA003 Quantitative Methods: Multivariate Analysis
Alvaro von Borries
810207-8618

Lab Report 3 – Regression with Binary Outcome

Introduction

The task consists on building a statistical model that can accurately predict the chances of survival of the passengers of the Titanic. In particular we want to assess Kate's claim that the absence of her father Leonardo, might have affected her and her mother's chances of survival. Given the data provided, we begin by fitting a logistic model with "Survived" as the dependent variable and age, gender, number of siblings and spouses (SibSp), number of parents or children (Parch), and the ticket class as predictors.

Data Exploration

Tables 1 presents the descriptive statistics for the selected variables. It is possible to see that there are 177 missing values for the variable age. Also, there are ages below one, which could rise suspicion on coding errors, but here it will be assumed that those individuals were babies of less than 1 years of age at the moment, and the age on months (fraction of a year) was recorded.

Then Figures 1 to 3 presents histograms for age, SibSp, and Parch and they show that all three scale variables selected are skewed to the left. Further, Figure 4 to 8 present the relation between "Survived" and the predictors, where no clear relation is deductible for age, SibSp, and Parch, while there is a much clearer pattern when it comes to gender and class. The majority of women survived, while the majority of men died. Also, most people on first class survived, while the great majority on third class did not.

To include ticket class, three dummy variables were created and the models include 2nd and 3rd class variables, making 1st class the base category. Also, SibSp and Parch were converted to scale variables.

The Model

A first run has revealed that all pre-selected variables but "number of parents or children" are statistically significant. A model without this predictor shows an improvement on AIC from 533 to 514 and no loss of McFadden pseudo R^2 (0.340). A model with "Parch" transformed into a dummy variable (0 if Parch=0, and 1 if Parch>0) does not achieve any significant improvement.

Therefore, the final model includes only age, sex, number of siblings and spouses (SibSp), and class (dummies for 2nd and 3rd class) as predictors. This model correctly predicts 80.3% of the overall cases: 85.1% of the non-survival cases, and 73.1% of the survival cases.

The model is significantly better than the null model as it is shown by the Chi-square test (X^2=328, p-value=0.000), and the improvement on AIC from 831 to 514 when running the multinomial logistic model. As said, the McFadden R^2 for this model is 0.34.

Table 2 presents the statistics describing the coefficients of the predictors, where it is possible to see that the most influential predictors are gender and 3rd class dummy, since the AIC increases the most if the model omits these variables (from 514 to 699 and 616 respectively).

The final regression equation is as follows:

Log(Odds) = 1.707 – 0.045*Age + 2.628*Sex – 0.380*SibSp – 1.414*2nd_Class – 2.653*3rd_Class

The predicted probability of survival of can be calculated as follows:

- Log(Odds)_Kate = 1.707 – 0.045*4 + 2.628*1 – 0.380*0 – 1.414*0 – 2.653*1 = 1.502
    - Survival Probability = exp(1.502)/(1 + exp(1.502)) = 0.818
- Log(Odds)_Sue w.Leo = 1.707 – 0.045*20 + 2.628*1 – 0.380*1 – 1.414*0 – 2.653*1 = 0.402
    - Survival Probability = exp(0.402)/(1 + exp(0.402)) = 0.599
- Log(Odds)_Sue w/o.Leo = 1.707 – 0.045*20 + 2.628*1 – 0.380*0 – 1.414*0 – 2.653*1 = 0.782
    - Survival Probability = exp(0.782)/(1 + exp(0.782)) = 0.686

Therefore, according to this predictive model, the presence of Leonardo on the boat does not influence the probability of survival of Kate, since the number of parents or children turned out to be non-significant and was excluded. However, it turns out that the presence of Leonardo would have lowered Sue's probabilities of survival by almost 9 points.

As we have already discussed the most influential predictors were sex and ticket class (in particular holding one for 3rd class). The presence of a parent turned out to not be statistically significant, and the presence of a spouse or sibling is the less influential variable of the ones included in the final model.

Appendix

Table 1. Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Survived	891	0.00	1.00	0.38	0.487
Age	714	0.42	80.00	29.70	14.527
Gender	891	0.00	1.00	0.35	0.478
SibSp	891	0.00	8.00	0.52	1.103
Parch	891	0.00	6.00	0.38	0.836
First Class	891	0.00	1.00	0.24	0.429
Second Class	891	0.00	1.00	0.21	0.405
Third Class	891	0.00	1.00	0.55	0.498
Valid N (listwise)	714				

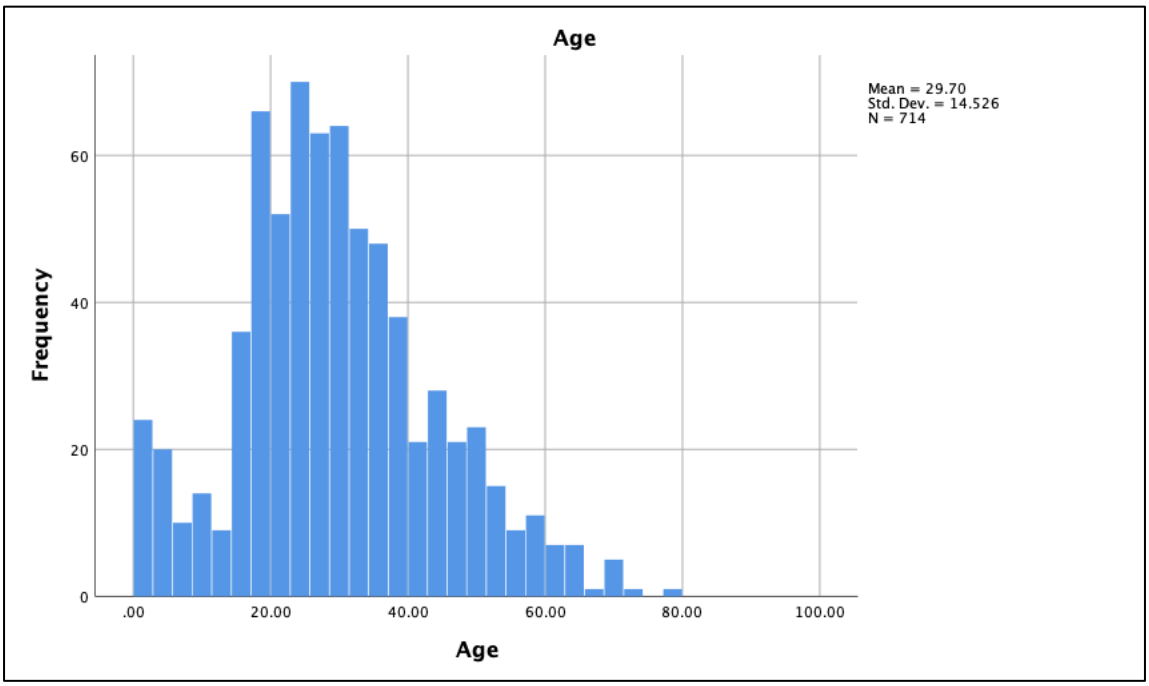

Figure 1. Histogram for Age

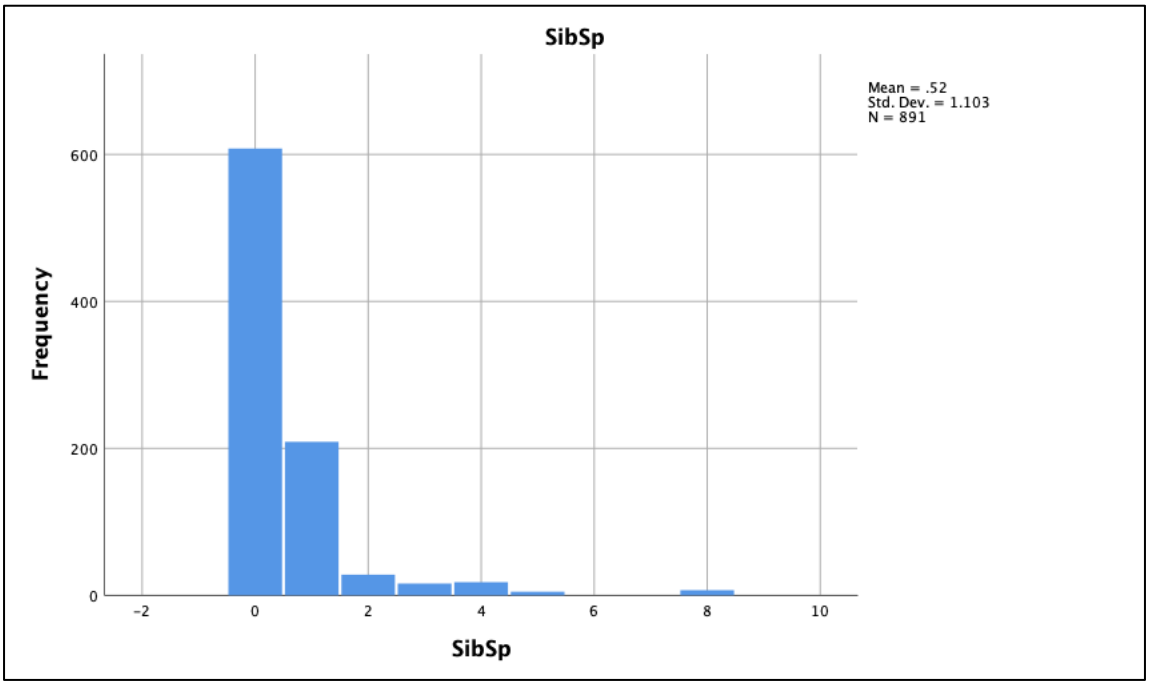

Figure 2. Histogram for Number of Siblings and Spouses (SibSp)

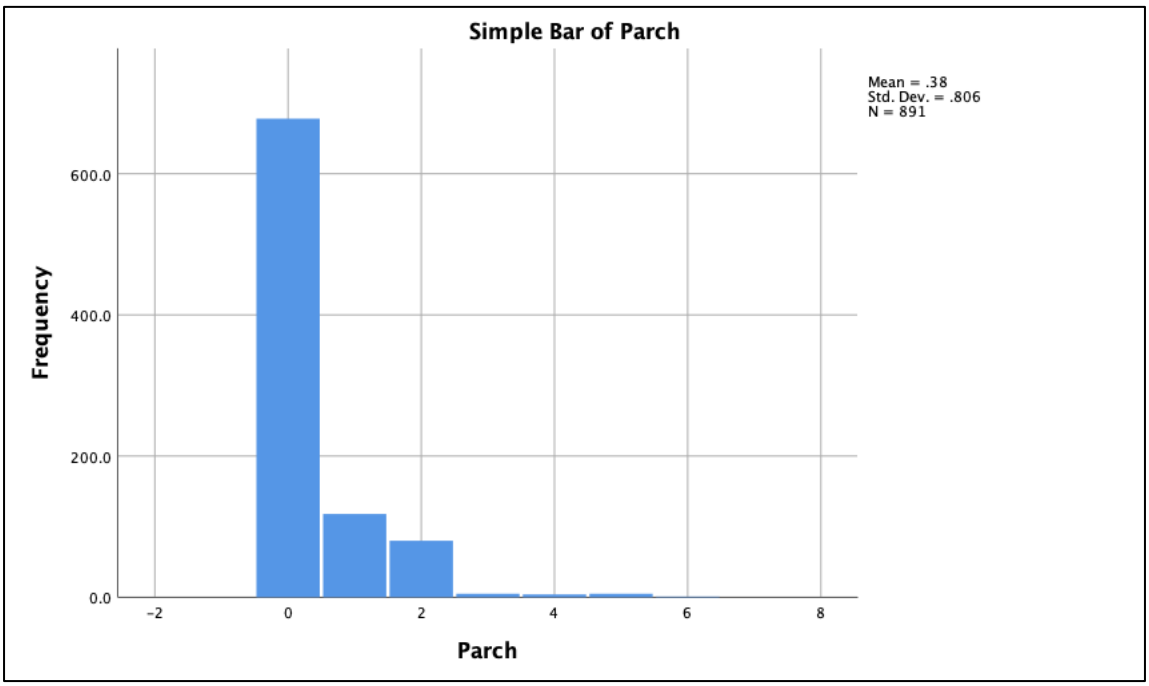

Figure 3. Histogram for Number of Parents or Children (Parch)

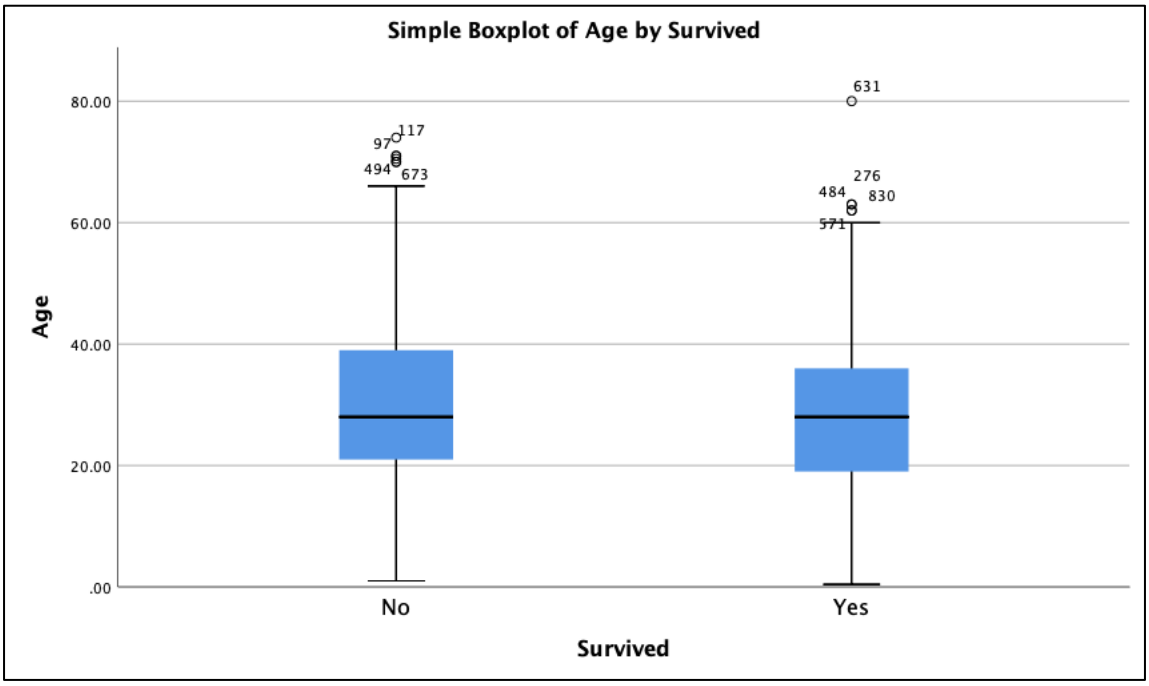

Figure 4. Boxplot Survived vs Age

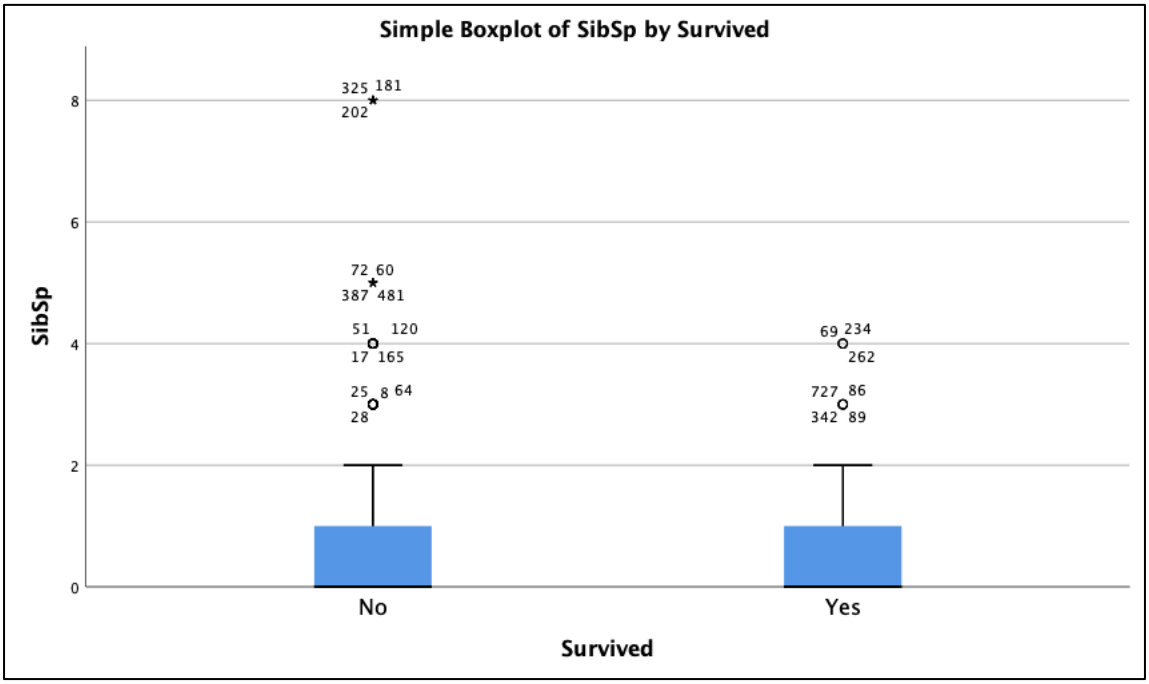

Figure 5. Boxplot Survived vs SibSp

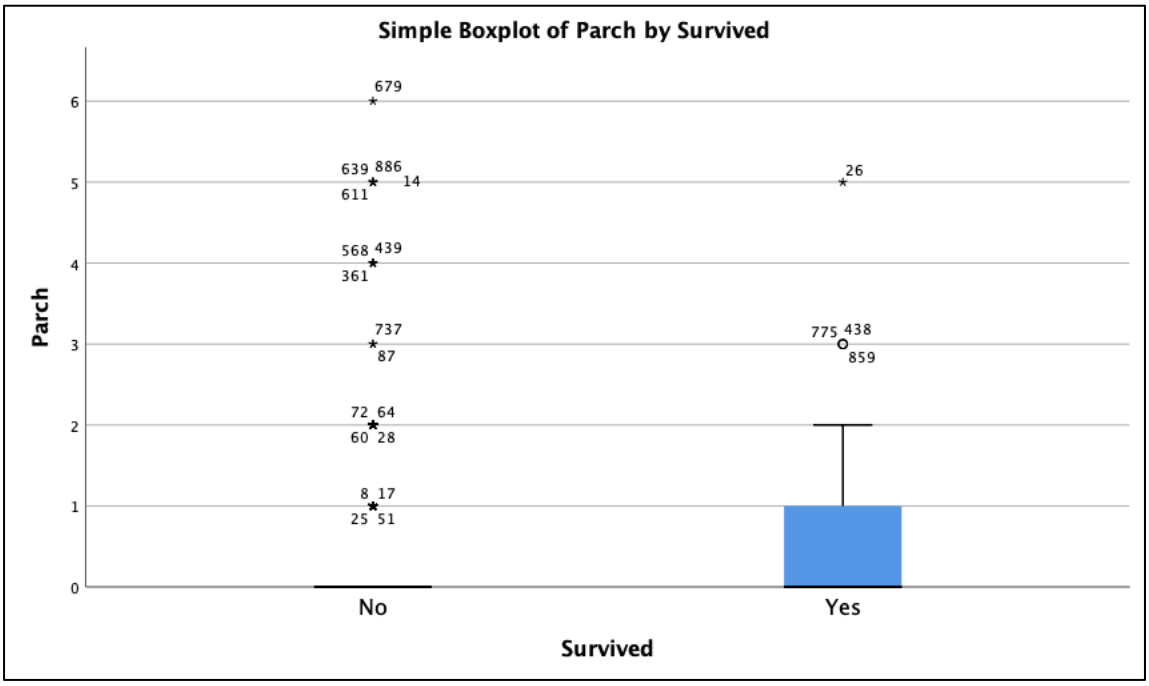

Figure 6. Boxplot Survived vs Parch

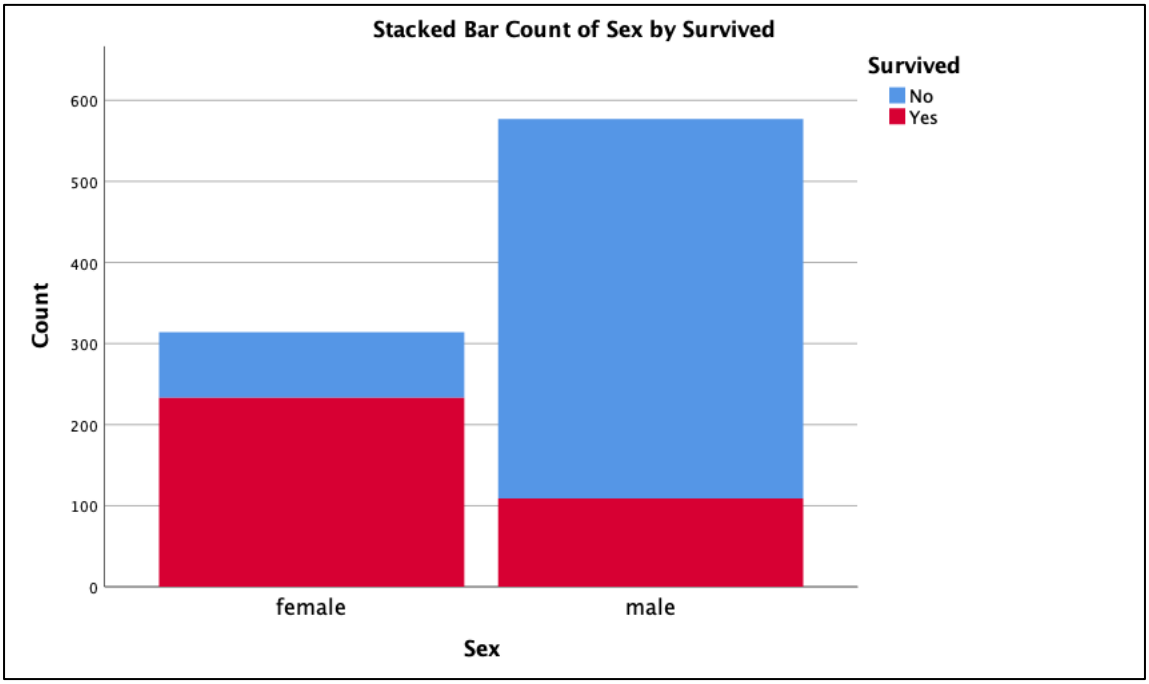

Figure 7. Bar Chart Survived vs Gender

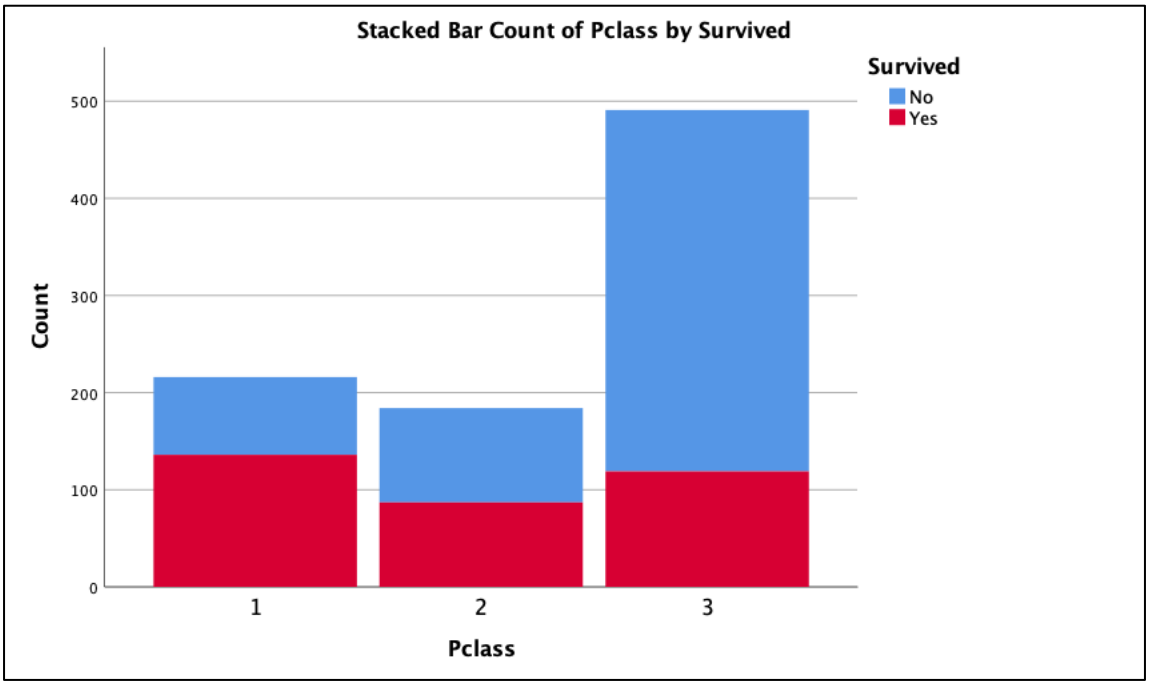

Figure 7. Bar Chart Survived vs Class

Table 2. Model Coefficients and Predictor Statistics

	b	Odds Ratio	95% CI lb of OR	95% CI ub of OR	AIC of reduced model	Chi^2	p-value
Intercept	1.71				532.05	20.01	< .001
Age	-0.05	0.96	0.94	0.97	544.88	32.84	< .001
Gender	2.63	13.84	9.09	21.09	699.20	187.15	< .001
SibSp	-0.38	0.68	0.54	0.87	522.76	10.72	0.002
Second Class	-1.41	0.24	0.14	0.43	538.35	26.30	< .001
Third Class	-2.65	0.07	0.04	0.12	616.10	104.06	< .001

Note: All Chi^2 test dfs = 1

Links

Syntax: https://github.com/avonborries/SAKA003-VT20/tree/master/Lab%203