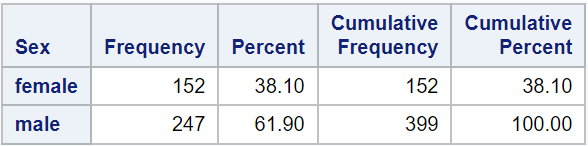
**Data Analysis on the Survival Rate for the Titanic**

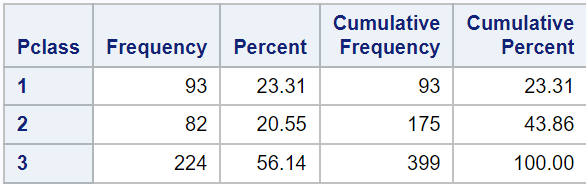
The purpose of this project is to use a sample of the data regarding passengers on the Titanic to determine if key metrics affected an individual’s probability of survival. The data analysis is done using SAS and can be accessed through the accompanying file, “Titanic.sas”

1. **Determining the demographics of the passengers:**
2. The ratio of male to female amongst the passengers.



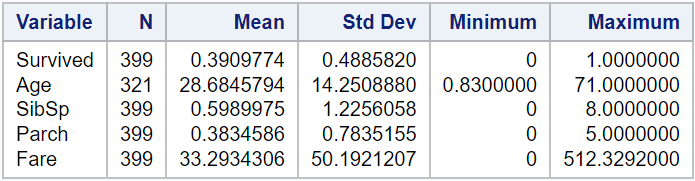
**Figure 1.a)** Number of male and female passengers and their cumulative ratio.

1. Number of passengers in each class (PClass)

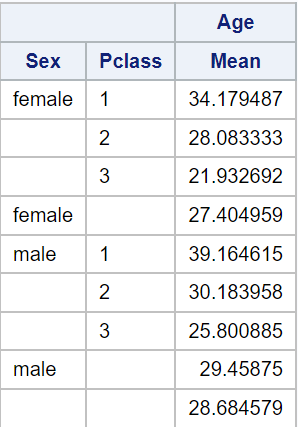


**Figure 1.b)** The number of passengers in each travel class and their cumulative ratio.

1. Average age of the passengers



**Figure 1.c)** We are only interested in the mean value of the Age variable, note that there are only 321 entries for Age instead of 399, meaning there are missing entries.

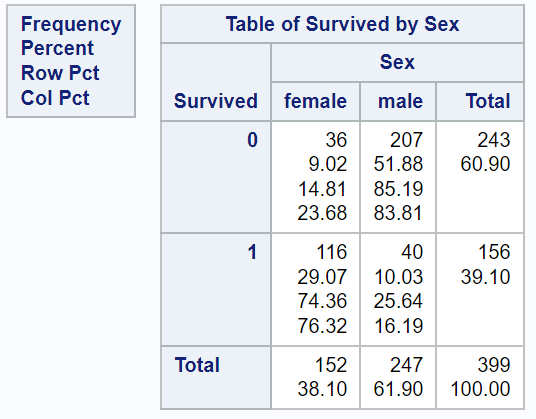
**Figure 1.d)** A further breakdown of the average passenger age by gender and class.

The higher the travel class tends to indicate older passengers, likely because they’ve had more time to earn higher incomes.

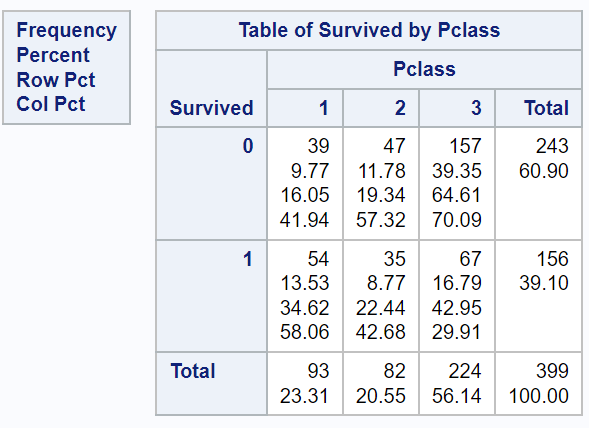
1. The minimum, maximum, average, and standard deviation of ticket fares

Refer to Figure 1.c to determine that fares had a mean price of $33.29, a Std deviation of $50.19, minimum of $0, and maximum of $512.33.

1. Analyzing survival rates

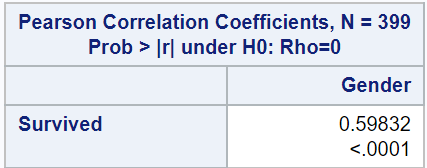
**Figure 1.1.e)** The breakdown of survival by gender. A value of 0 indicates the passenger did not survive.

It is quickly evident that male passengers had a far lower survival rate (25.64%) compared to female passengers (74.37%). When referring to figure 1.a) we see that of 152 female passengers, 116 survived as opposed to 40 of 247 male passengers.

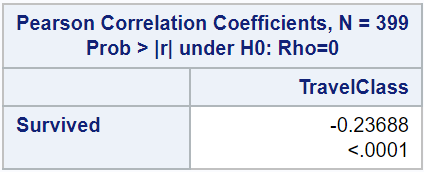
**Figure 1.2.e)** Breakdown of survival by travel class. A value of 0 indicates the passenger did not survive.

1st class travelers have a (58.06%) survival rate, 2nd class (42.68%) and 3rd class (29.91%) indicating that those traveling in higher classes tended to have higher survival rates. Referring to Figure 1.b) 1st class had 54 of 93 passengers survive, 2nd class 35 of 82 passengers survive and 3rd class 67 of 224 passengers survive.

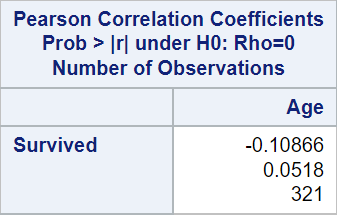
1. Examining the correlation between the survival rate and other variables.

**Figure 1.1.f)** The Pearson correlation between Survived and gender.

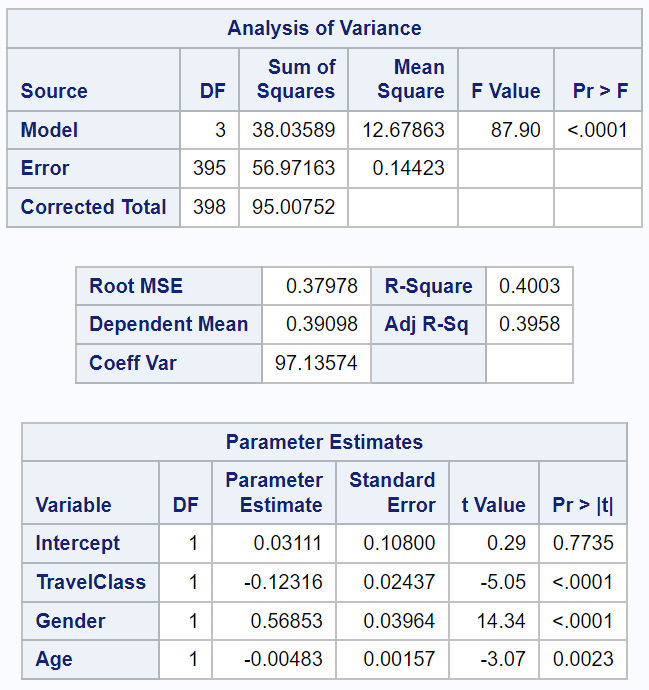
Note that we previously assigned values of 1 for males and 2 for females. There is a fairly strong positive correlation between the rate of survival and gender meaning that females were more likely to survive. This can likely be explained by the evacuation mandate of “women and children first”.

**Figure 1.2.f)** The Pearson correlation between Survived and TravelClass.

There is a slight negative correlation between one’s travel class and the rate of survival indicating that those traveling in lower classes tended to have a lower survival rate. This could maybe be explained by the cabins that lower classes would be assigned cabins deeper into the vessel.

**Figure 1.3.f)** The Pearson correlation between Survived and Age

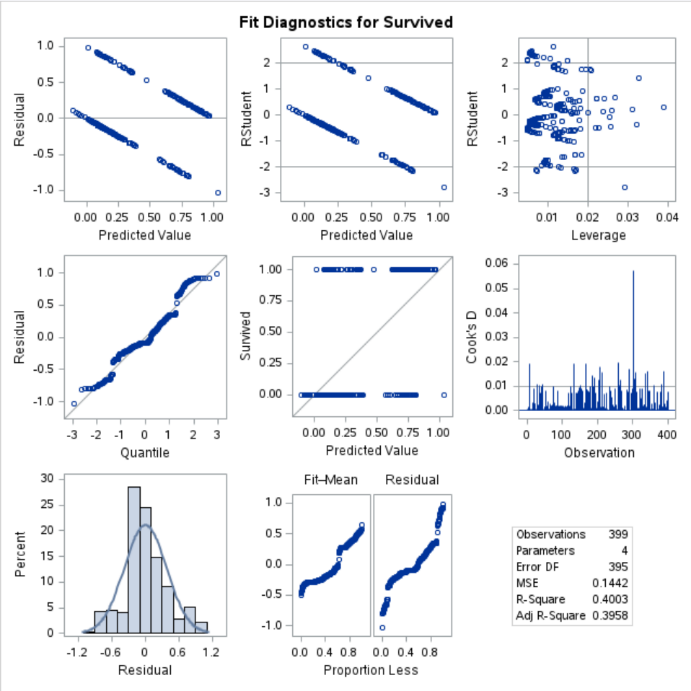
Age has a weak negative correlation against survival rate, surprisingly, despite the mantra of “women and children first”. The correlation does indicate that younger people were only slightly prioritized in evacuation compared to prioritization based on gender.

1. ****Building a linear regression model to predict probability of survival.

**Figure 2.1.a)** The result of running the linear regression procedure.

The linear regression model is then built using the coefficients for TravelClass, gender (where male = 1 and female = 2) and age with y intercept 0.031111.

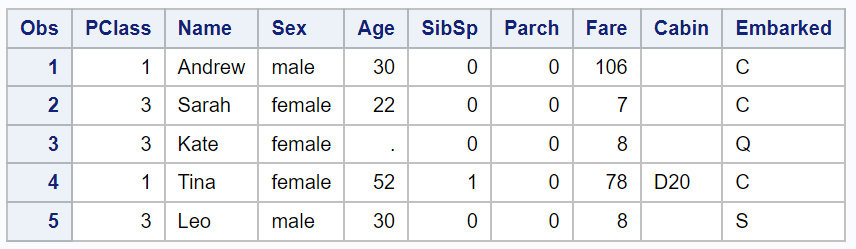
Prob(survival) = 0.56853\* Gender -0.12316\*TravelClass – 0.00483\*Age + 0.03111

****It should be stated that with an R-squared value of 0.4003, that this model is not very good as a predictor of survival rate for the passengers. Unfortunately, the other variables available in the original dataset such as SibSp (number of onboard siblings and spouses), Parch (number of parents onboard), fare (ticket fare) have very low correlation or a lot of missing data entries that would be difficult to make assumptions for. Other variables such as embarked (where the passenger embarked) are clearly meaningless.

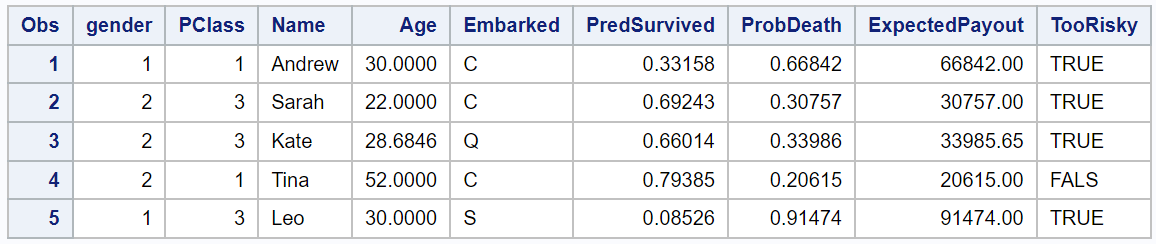
Addressing the missing entries in the age category, we use the average age in place of the missing entries which further increases the error within the model.

**Figure 2.1.b)** Visualization ofthe regression model as a predictor of survival rate.

1. **Using a small sample data set to analyze the risk of selling life insurance to the passengers:**
2. We are interested in the scenario of selling life insurance with a $100,000 payout to a small sample of passengers, right before the Titanic departs. Five people are interested in buying the insurance coverage. We want to try to use the linear regression model from our previous work to predict whether selling life insurance to the individual is worth it. Assuming a one-time premium of $25,000. (We determine this by $100,000 \* probability of death < $25,000).

**Figure 3.1.a)** A sample subset of the passenger information.

Using the linear regression model that we built earlier (Probability of survival = 0.12316\*PClass + 0.56853\*Sex - 0.00483\*Age + 0.03111). We can then find the probability of death from 1 – Probability of survival. Applying ($100,000\*Probability of death) to find the expected payout the results are shown as follows:



**Figure 3.1.b)** Dataset with appended values from the linear regression model, ProbDeath = 1 – PredSurvived and the expected payout = ProbDeath \* $100,000.

From these results it is evident that Tina carries the lowest risk with insuring.

It should be noted again that the R-square value associated with our model is 0.4003 so there remains a large amount of error within the model.

1. Comparing our model’s results to the true outcome to see if the model is an accurate predictor of risk. (Of the five passengers, Leo was the only one to perish amongst the 5 people that bought the life insurance).

Under the assumption that all 5 purchased the life insurance coverage and comparing our model in which we only sell the life insurance to Tina we clearly see that the predicted model results in a loss compared to reality.

If the model was to be trusted and we only sold the coverage to Tina and she survives, our profit is $25,000. Otherwise, if we had sold the coverage to everyone, the expected profit would be: $25,000 \* 5 – ∑ExpectedPayout ($243,673.65) = -$118,673.65

If we use the true result (Leo was the only one to die) and that all five had purchased the insurance coverage, then the profit would be $25,000 - $100,000 = $25,000.

In this case, if we had used the model to predict risk before selling coverage then the outcome would have been the same. The model, however, was not a good predictor of survival rate when compared to true outcomes.