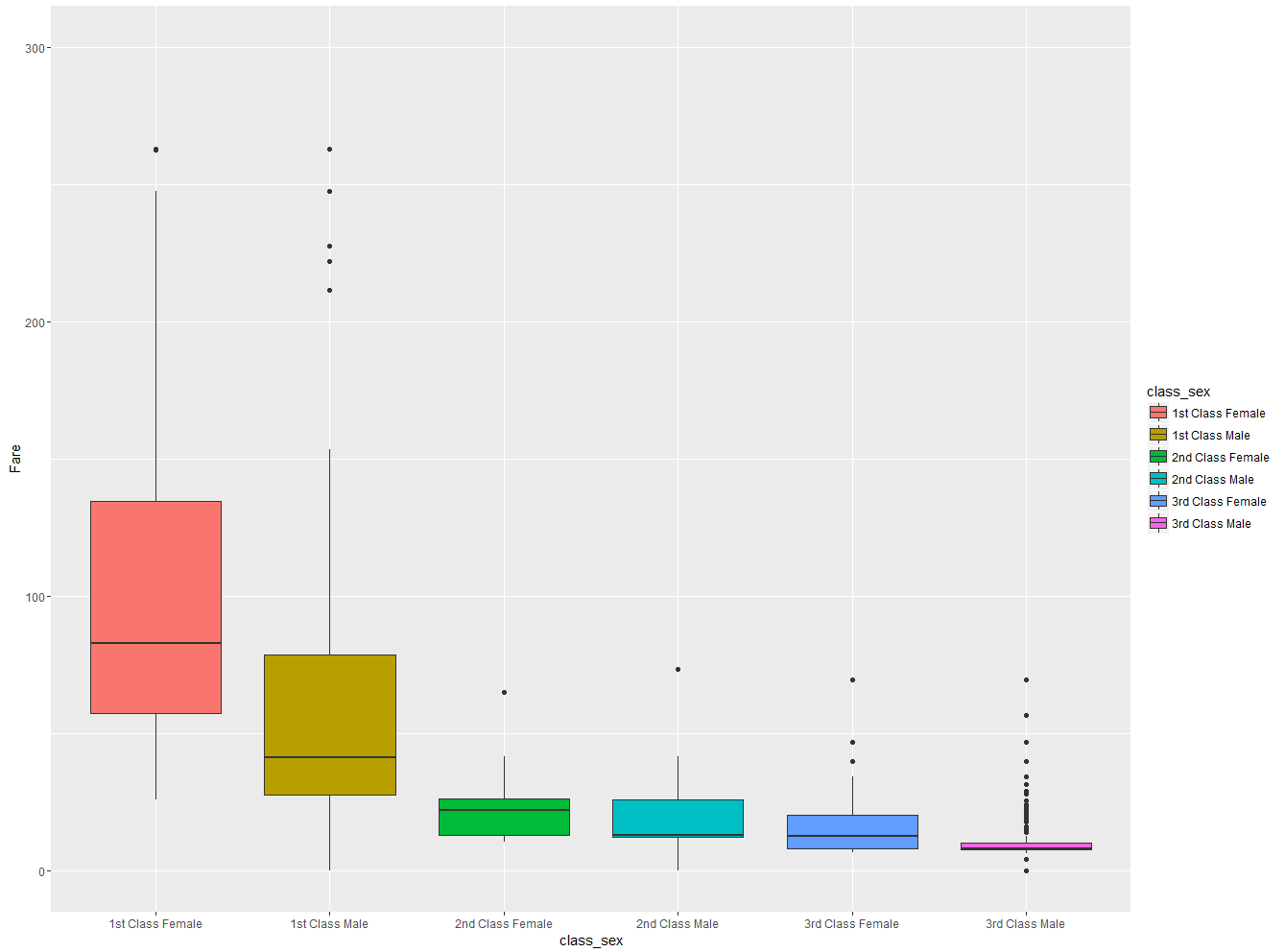
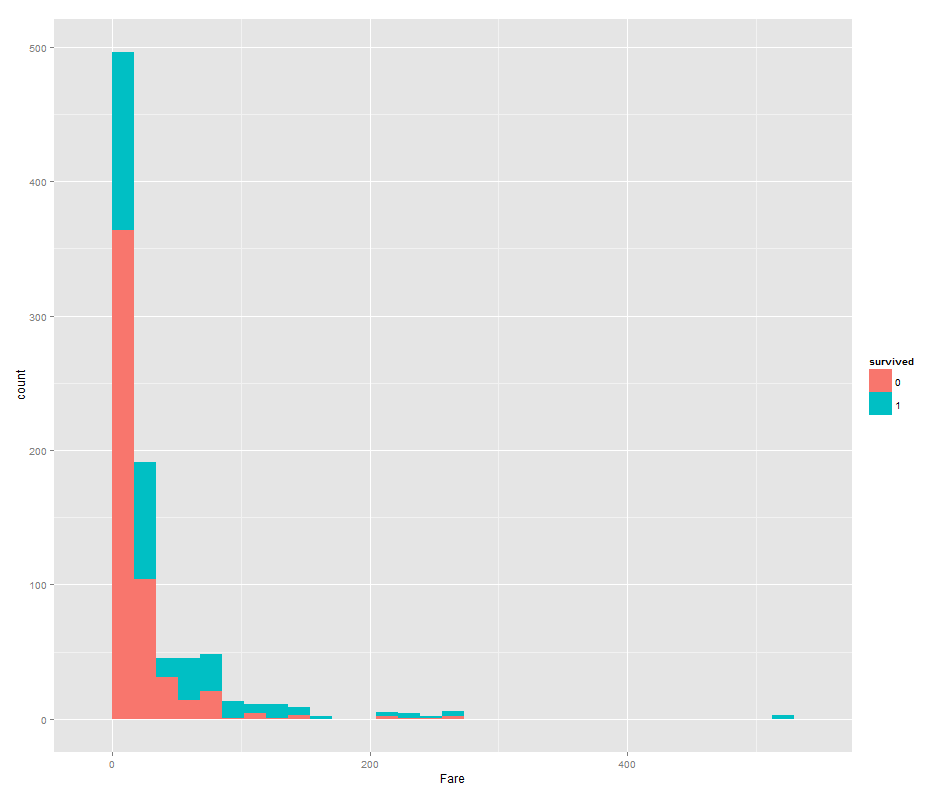
**Assignment 4**

**Katherine Rodgers and John Merranko**

For our box and whisker plot, we decided to first formulate a new variable called “class\_sex,” which categorizes observations by passenger class and sex. We then box-plotted fare by class/sex. As you can see, there is an obvious trend such that passengers in better classes paid higher fares, and there is also a less obvious trend such that females within any given class appeared to pay higher fares than men within the same class.

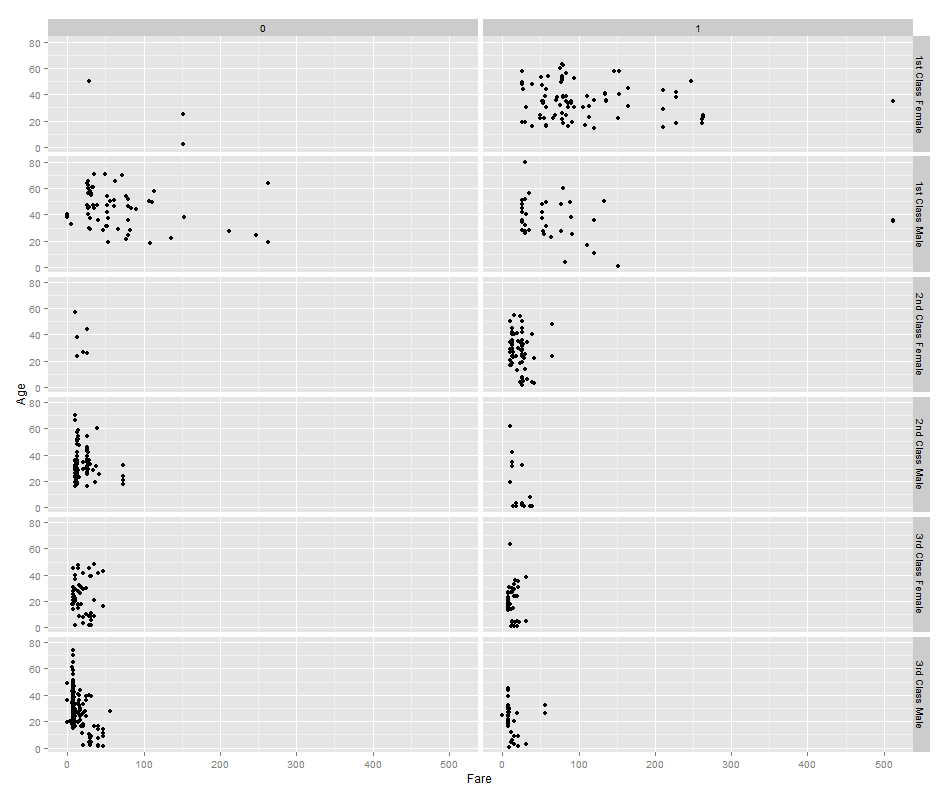


For our histogram, we sought to analyze survival proportions across the full distribution of passenger fares. As you can see, the proportion of survivors in the low end of the fare distribution seemed to be quite low compared to the proportion of survivors in the high end of the fare distribution.



We sought to expand our analysis from the previous graph to other possible correlates of survival probability using a facet grid with class variables for survival and class/sex and the continuous variables age and fare. The results indicated several relationships between the plotted variables:

1. As demonstrated earlier, lower fares were associated with lower chance of survival.
2. As demonstrated earlier, lower passenger class was associated with lower chance of survival.
3. Women were more likely to survive than men, particularly in the better classes. This indicates a possible interactive relationship between class and gender on survival (logistic regression actually confirms this as shown on the next page).
4. Older age was associated with lower chance of survival, particularly among men and lower class passengers. This indicates a possible interactive relationship between age and gender on survival (logistic regression confirms this), as well as a possible interactive relationship between age and class on survival (logistic regression does not confirm this).



Logistic Regression Output:

summary(glm(data=train,Survived~Sex+factor(Pclass)+factor(Pclass)\*Sex,family="binomial"))

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.4122 0.5868 5.815 6.06e-09 \*\*\*

Sexmale -3.9494 0.6161 -6.411 1.45e-10 \*\*\*

factor(Pclass)2 -0.9555 0.7248 -1.318 0.18737

factor(Pclass)3 -3.4122 0.6100 -5.594 2.22e-08 \*\*\*

Sexmale:factor(Pclass)2 -0.1850 0.7939 -0.233 0.81575

Sexmale:factor(Pclass)3 2.0958 0.6572 3.189 0.00143 \*\*

summary(glm(data=train,Survived~Age+Sex+Age\*Sex,family="binomial"))

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.59380 0.31032 1.913 0.05569 .

Age 0.01970 0.01057 1.863 0.06240 .

Sexmale -1.31775 0.40842 -3.226 0.00125 \*\*

Age:Sexmale -0.04112 0.01355 -3.034 0.00241 \*\*

summary(glm(data=train,Survived~Age+factor(Pclass)+Age\*factor(Pclass),family="binomial"))

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.24252 0.49128 4.565 5.00e-06 \*\*\*

Age -0.04044 0.01143 -3.537 0.000405 \*\*\*

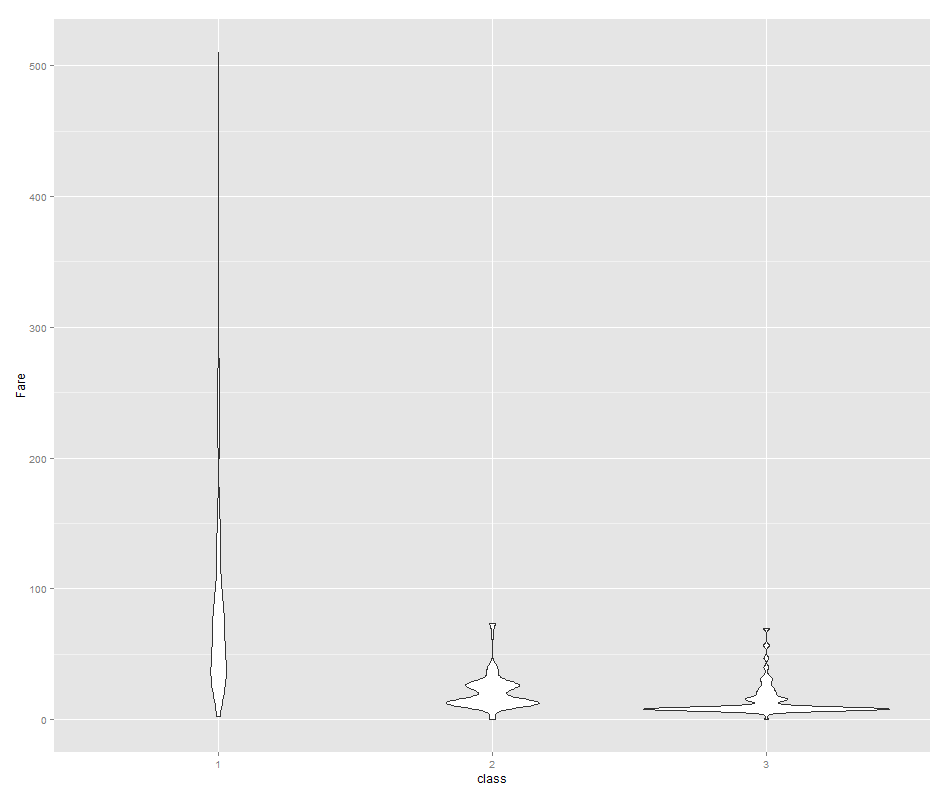
factor(Pclass)2 -1.05325 0.63260 -1.665 0.095921 .

factor(Pclass)3 -2.40716 0.56489 -4.261 2.03e-05 \*\*\*

Age:factor(Pclass)2 -0.00236 0.01687 -0.140 0.888748

Age:factor(Pclass)3 -0.00172 0.01605 -0.107 0.914654

For our violin plot, we sought to analyze the distribution of fare by passenger class. Results indicate that not only were first class fares typically higher than the other two classes, the variance of first class fares were also much greater as illustrated by the tails of the kernel densities (the actual variances are also shown below). Thus, it seems that first class offered a much greater range of passenger fares.



|  |  |
| --- | --- |
| **Passenger Class** | **Variance of Fare** |
| First | 6143.5 |
| Second | 180.0 |
| Third | 138.7 |

Lastly, for our heatmap, we sought to analyze the correlations between the variables in the titanic dataset. We used Spearman correlations given that many of the variables are non-normal/ordinal, and while we recognize that the variables “sex” and “survived” were dichotomous and thus more appropriately analyzed using tetrachoric correlation, we thought it would still be interesting to include these via Spearman correlation alongside all the others in the heatmap. The strongest correlations are described below:

1. Higher fare was positively associated with survival, better passenger class, and higher number of family members.
2. Higher number of parents and children was associated with higher number siblings and spouses (may suggest that passengers tended to have lots of family aboard or no family at all).
3. Older age was mildly associated with better passenger class.
4. Males were less likely to survive.
5. Better passenger class was positively associated with survival.

