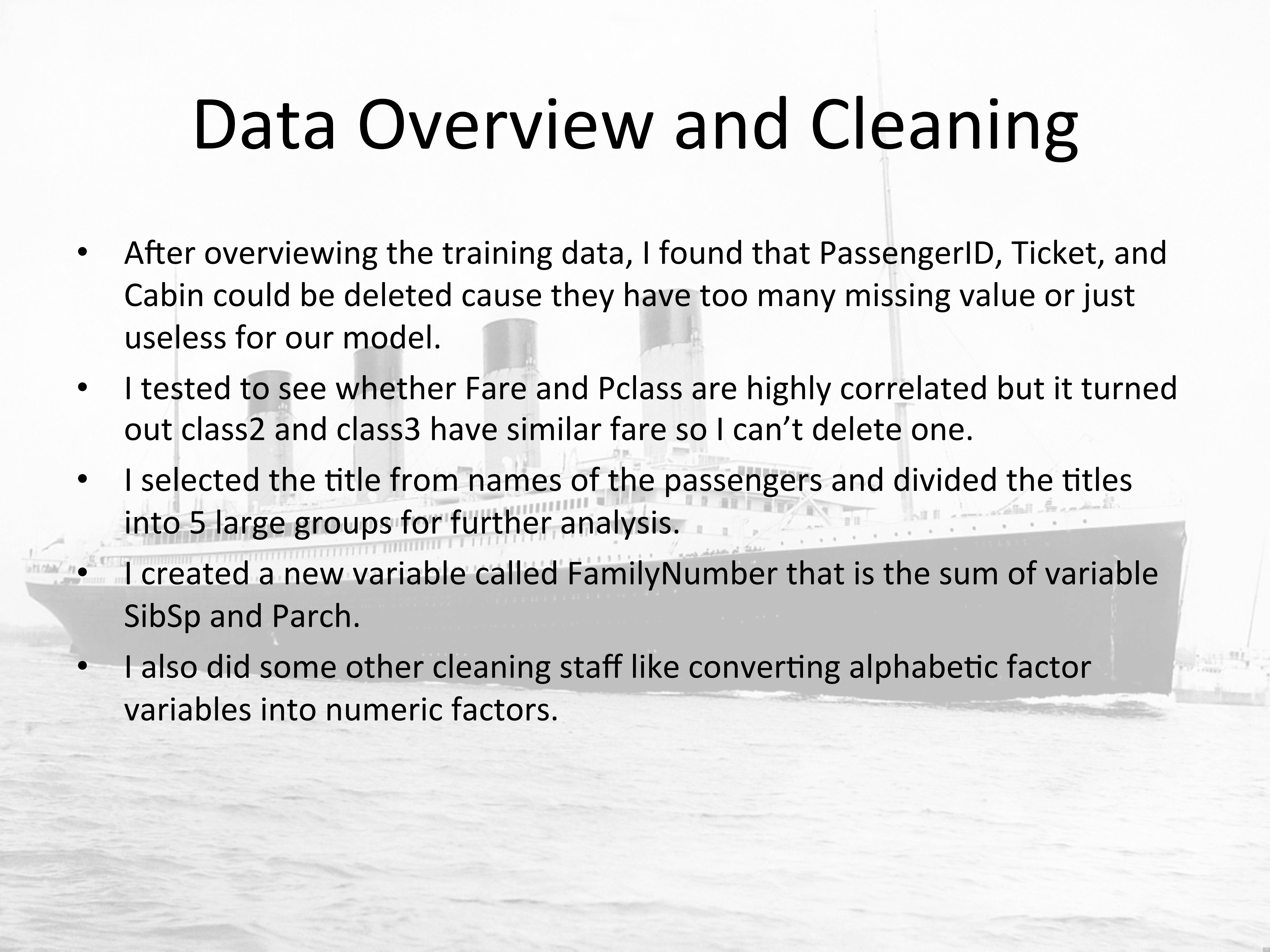
Titanic: Machine Learning from Disaster

Admission Project for Statistics AI Study Group

By: Zonghao Li

UNI: zl2613

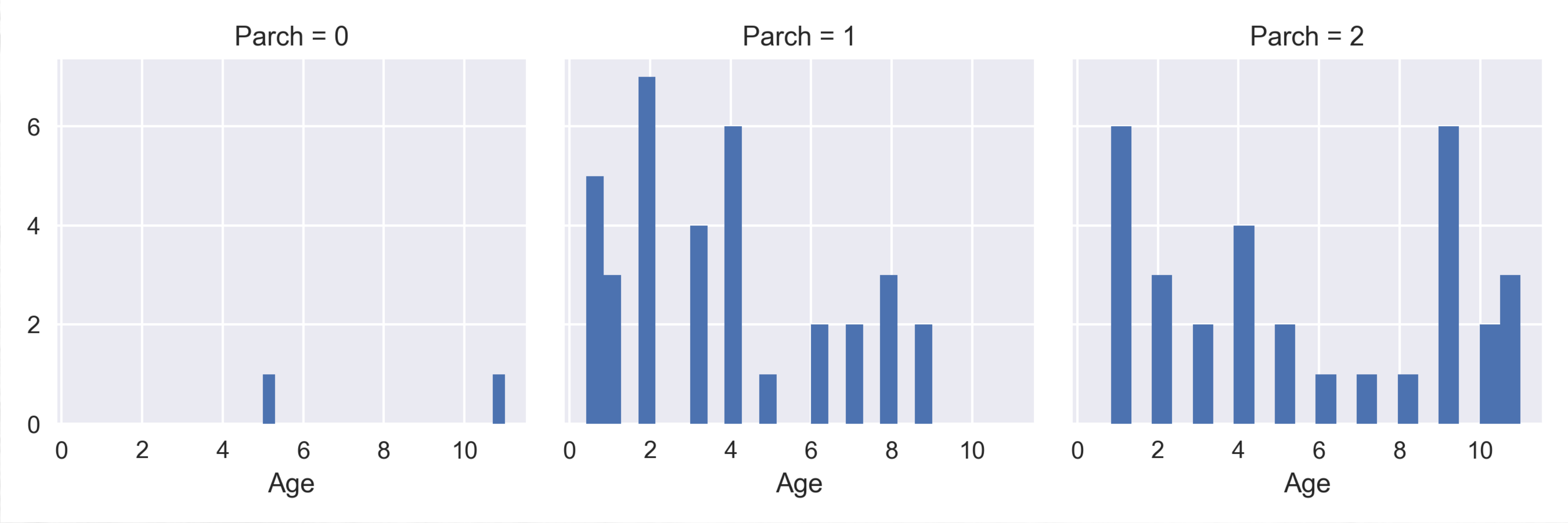


Data Overview and Cleaning

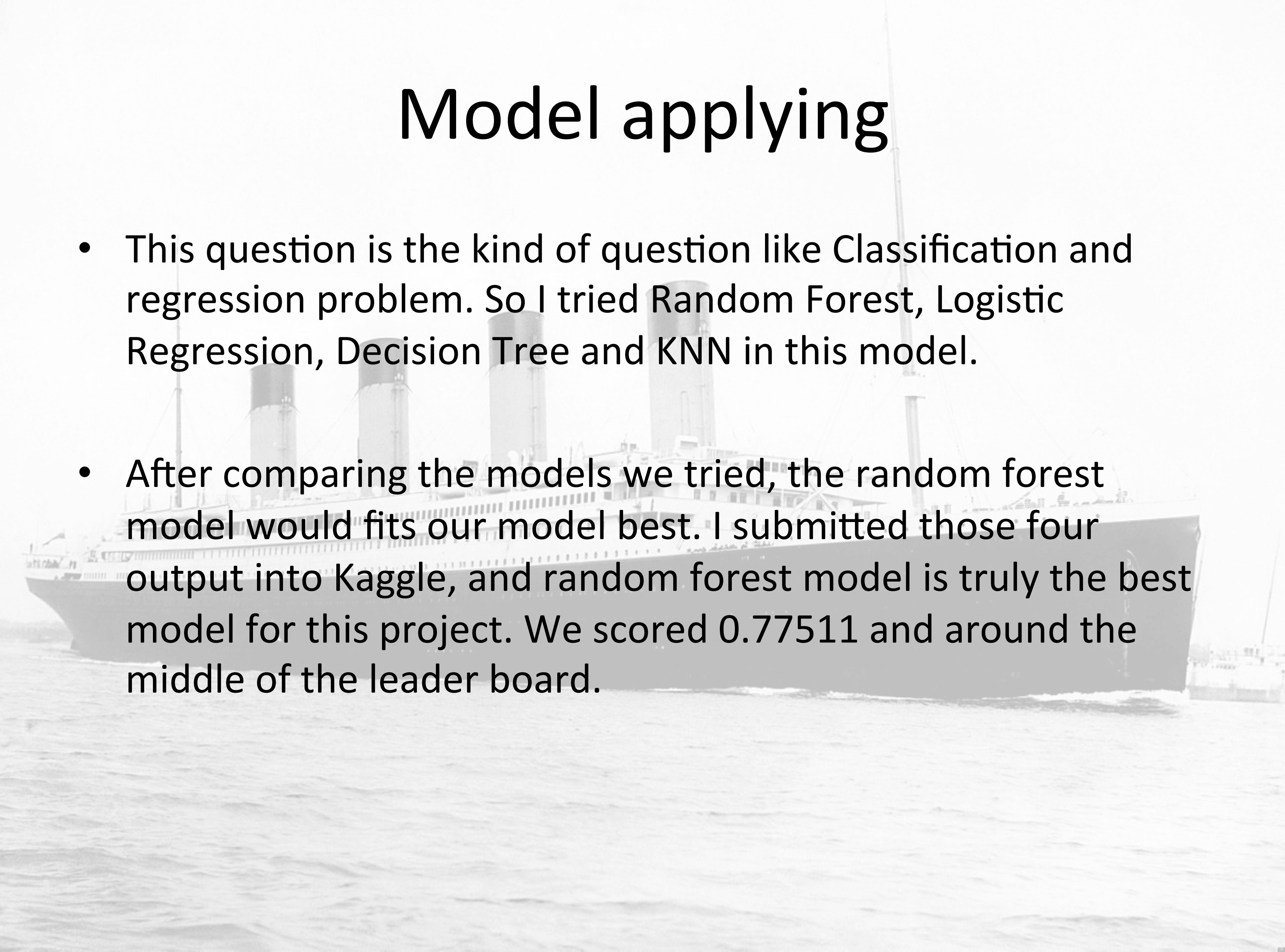
- After overviewing the training data, I found that PassengerID, Ticket, and Cabin could be deleted cause they have too many missing value or just useless for our model.
- I tested to see whether Fare and Pclass are highly correlated but it turned out class2 and class3 have similar fare so I can't delete one.
- I selected the title from names of the passengers and divided the titles into 5 large groups for further analysis.
- I created a new variable called FamilyNumber that is the sum of variable SibSp and Parch.
- I also did some other cleaning staff like converting alphabetic factor variables into numeric factors.

Missing Value

- For missing data in Embarked and Fare, I just fill them with the mean of each variable cause they just have seldom missing values.
- When dealing with Age, I didn’t just fill in the median or mean of the age column but use Parch and Title to get more accurate information.
- For missing age passenger with title “Master” which means boys and young men in English, I assume those to be children and filled with mean age of children.
- I analyzed relationship between Age and Parch and found that almost all passengers with age under 12 have at least 1 Parch which fits our common sense that little child would unlikely to travel alone.
- So I use the complement of this condition to assume passengers with missing name and 0 Parch to be adults and filled missing values with mean age of adults.



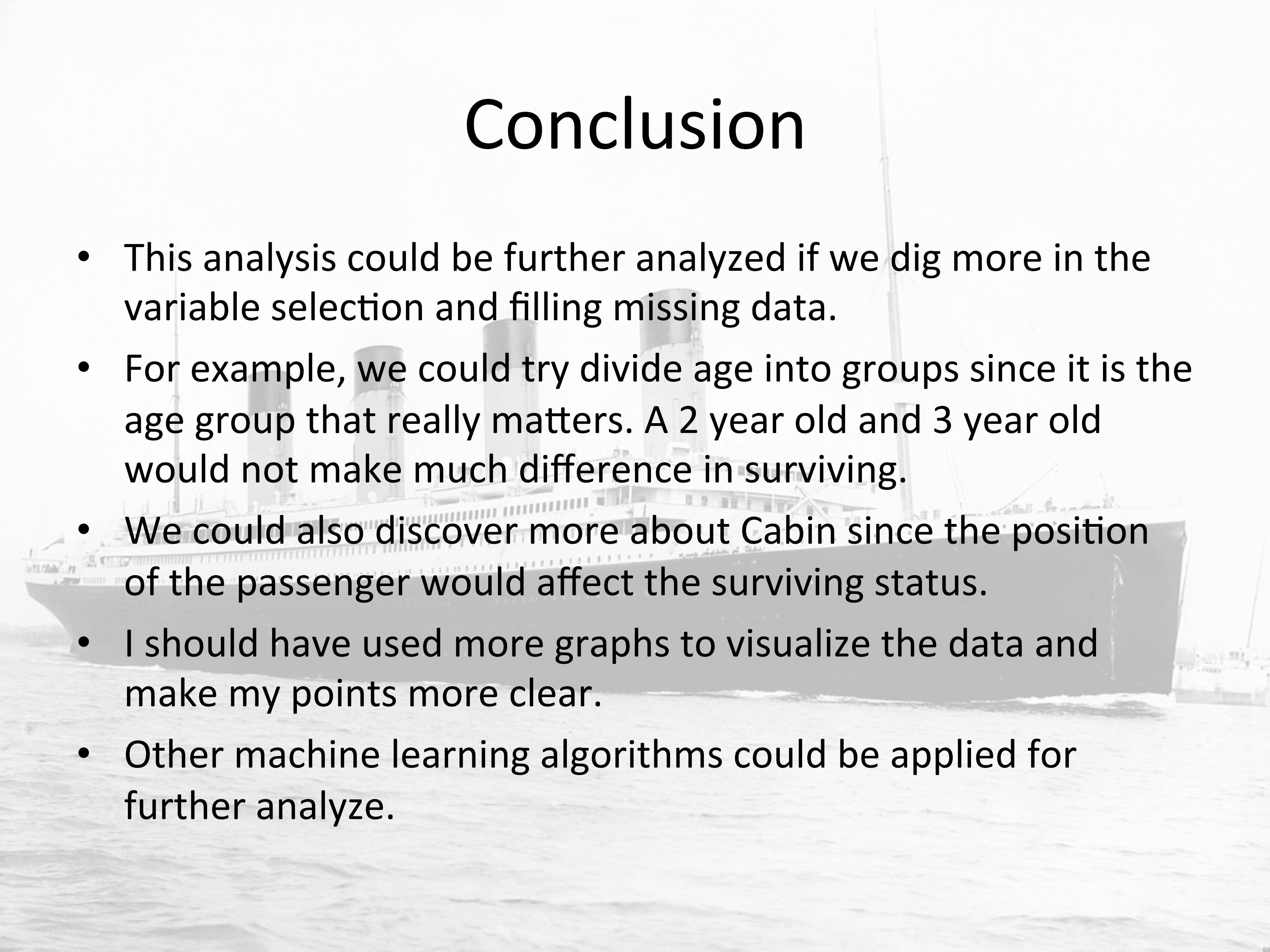
Model applying



- This question is the kind of question like Classification and regression problem. So I tried Random Forest, Logistic Regression, Decision Tree and KNN in this model.

- After comparing the models we tried, the random forest model would fits our model best. I submitted those four output into Kaggle, and random forest model is truly the best model for this project. We scored 0.77511 and around the middle of the leader board.

Conclusion



- This analysis could be further analyzed if we dig more in the variable selection and filling missing data.
- For example, we could try divide age into groups since it is the age group that really matters. A 2 year old and 3 year old would not make much difference in surviving.
- We could also discover more about Cabin since the position of the passenger would affect the surviving status.
- I should have used more graphs to visualize the data and make my points more clear.
- Other machine learning algorithms could be applied for further analyze.