The Problem Statement

On April 14, 1912, the RMS Titanic struck an iceberg in the North Atlantic Ocean and sank. Of the 2,224 people on board, only 706 survived.

The goal of this exercise is to predict survivors on the Titanic based on nine input variables, described below. We are provided two datasets: (1) train.csv, containing 891 records and (2) test.csv, containing 418 records. The two datasets are provided with the intent that models are formulated using the train dataset and model performance is evaluated on the test dataset.

Variable Notes

pclass: A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way… Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way… Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

source: https://www.kaggle.com/c/titanic/data

About the data

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket	class 1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in year	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Test.csv first 6 row.

I will provide test.csv and train.csv full file

```
> head(test)
  PassengerId Pclass                                          Name    Sex  Age SibSp Parch  Ticket    Fare Cabin Embarked
1         892      3                              Kelly, Mr. James   male 34.5     0     0  330911  7.8292              Q
2         893      3              Wilkes, Mrs. James (Ellen Needs) female 47.0     1     0  363272  7.0000              S
3         894      2                     Myles, Mr. Thomas Francis   male 62.0     0     0  240276  9.6875              Q
4         895      3                              Wirz, Mr. Albert   male 27.0     0     0  315154  8.6625              S
5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0     1     1 3101298 12.2875              S
6         897      3                    Svensson, Mr. Johan Cervin   male 14.0     0     0    7538  9.2250              S
>
```

Train.csv

```
> head(train)
  PassengerId Survived Pclass                                                Name    Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1     0        A/5 21171  7.2500              S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                              Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250              S
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                            Allen, Mr. William Henry   male  35     0     0           373450  8.0500              S
6           6        0      3                                    Moran, Mr. James   male  NA     0     0           330877  8.4583              Q
>
```

Exploratory Data Analysis

Missing Values

The first step is to find any and all missing data in the train and test sets.

Missing values can be treated using following methods

1.Deletion

2. Mean/ Mode/ Median Imputation (which I follow in this session)

Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

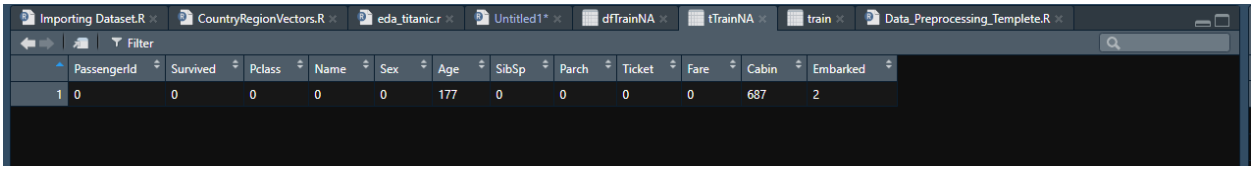
3. Prediction Model: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and va`riable with missing values is treated as target variable. Next, we create a model to predict target variable based on other attributes of the training data set and populate missing values of test data set.We can use regression, ANOVA, Logistic regression and various modeling technique to perform this. But this method has some drawbacks.

Train Dataset

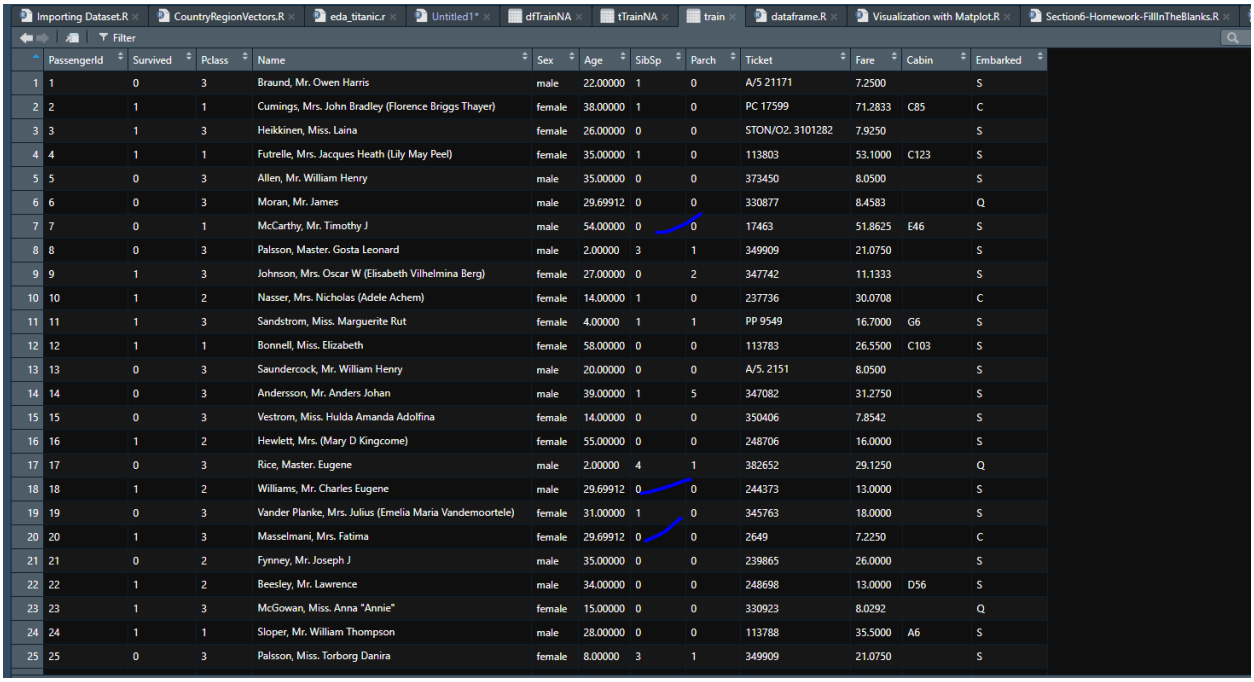
The test.csv dataset also had 3 columns with missing values: Age, Fare, and Cabin. Like the train dataset, the Cabin variable is sparse, with over 78% subjects missing values. The Age variable is populated for over 79% of the train subjects, and likely has good predictive power so it will likely be beneficial to impute values for subjects missing Age. The Fare variable is missing for a single subject. While Fare may not be an obvious predictor for survival, the fact that the dataset is over 99% complete for this variable indicates that it is a good candidate for imputation.

I can't taking care of missing data which value is string which i can that is bellow(Rscript).

Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence B	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May F	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhe	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William Henry	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Johan	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
19	0	3	Vander Planke, Mrs. Julius (Emelia Mari	female	31	1	0	345763	18		S
20	1	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0	2	Fynney, Mr. Joseph J	male	35	0	0	239865	26		S
22	1	2	Beesley, Mr. Lawrence	male	34	0	0	248698	13	D56	S
23	1	3	McGowan, Miss. Anna "Annie"	female	15	0	0	330923	8.0292		Q
24	1	1	Sloper, Mr. William Thompson	male	28	0	0	113788	35.5	A6	S
25	0	3	Palsson, Miss. Torborg Danira	female	8	3	1	349909	21.075		S
26	1	3	Asplund, Mrs. Carl Oscar (Selma August	female	38	1	5	347077	31.3875		S
27	0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	0	0	0	0	177	0	0	0	0	687	2



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00000	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00000	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00000	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00000	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00000	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	29.69912	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00000	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00000	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00000	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00000	1	0	237736	30.0708		C
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00000	1	1	PP 9549	16.7000	G6	S
12	12	1	1	Bonnell, Miss. Elizabeth	female	58.00000	0	0	113783	26.5500	C103	S
13	13	0	3	Saundercock, Mr. William Henry	male	20.00000	0	0	A/5. 2151	8.0500		S
14	14	0	3	Andersson, Mr. Anders Johan	male	39.00000	1	5	347082	31.2750		S
15	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00000	0	0	350406	7.8542		S
16	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.00000	0	0	248706	16.0000		S
17	17	0	3	Rice, Master. Eugene	male	2.00000	4	1	382652	29.1250		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	29.69912	0	0	244373	13.0000		S
19	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00000	1	0	345763	18.0000		S
20	20	1	3	Masselmani, Mrs. Fatima	female	29.69912	0	0	2649	7.2250		C
21	21	0	2	Fynney, Mr. Joseph J	male	35.00000	0	0	239865	26.0000		S
22	22	1	2	Beesley, Mr. Lawrence	male	34.00000	0	0	248698	13.0000	D56	S
23	23	1	3	McGowan, Miss. Anna "Annie"	female	15.00000	0	0	330923	8.0292		Q
24	24	1	1	Sloper, Mr. William Thompson	male	28.00000	0	0	113788	35.5000	A6	S
25	25	0	3	Palsson, Miss. Torborg Danira	female	8.00000	3	1	349909	21.0750		S

Variable Features

PassengerId

PassengerId is a primary key for each row of data in the train and test sets. This variable will not be included in any of the predictive models.

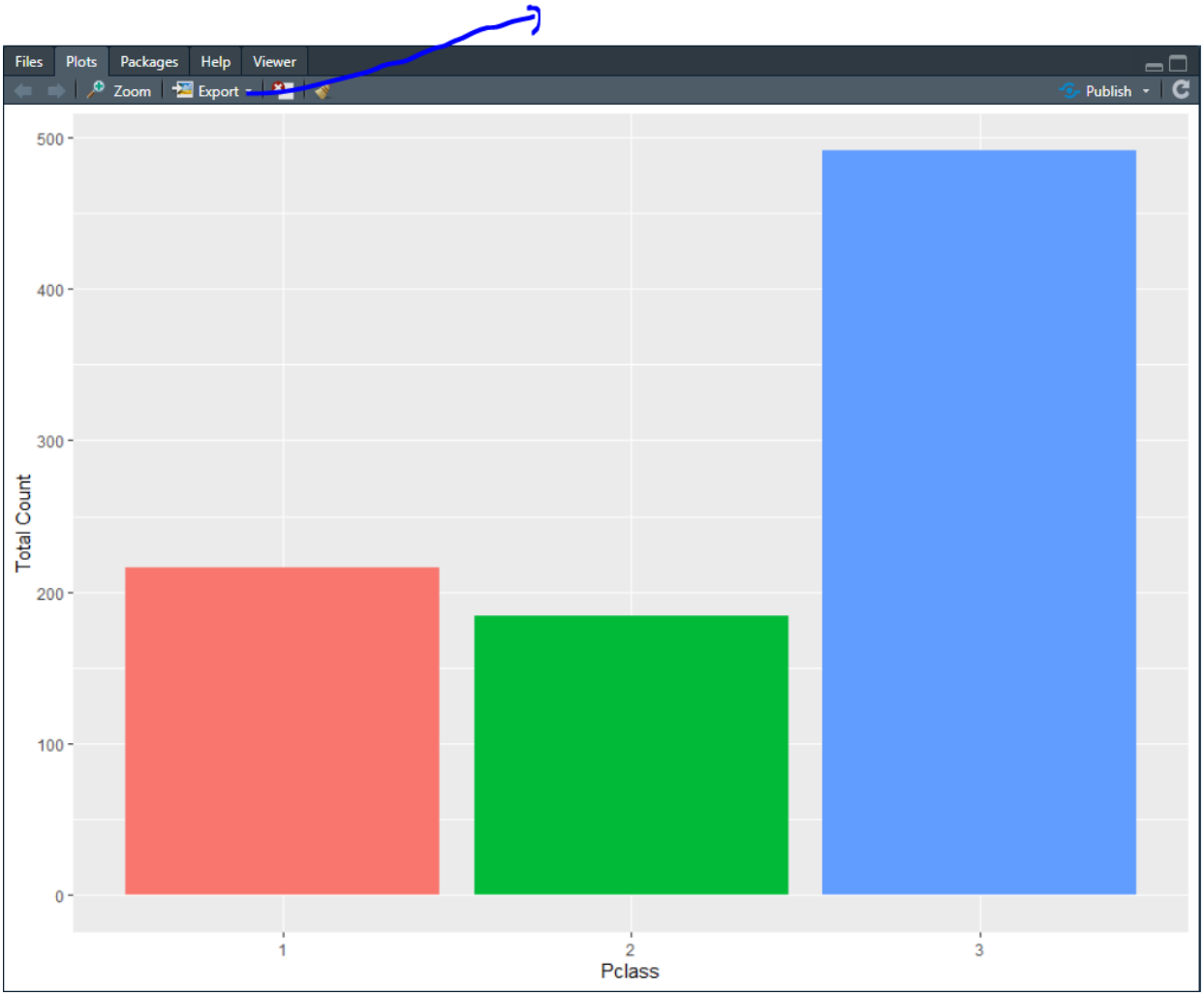
Survived

Survived is the class we're trying to predict.

Pclass

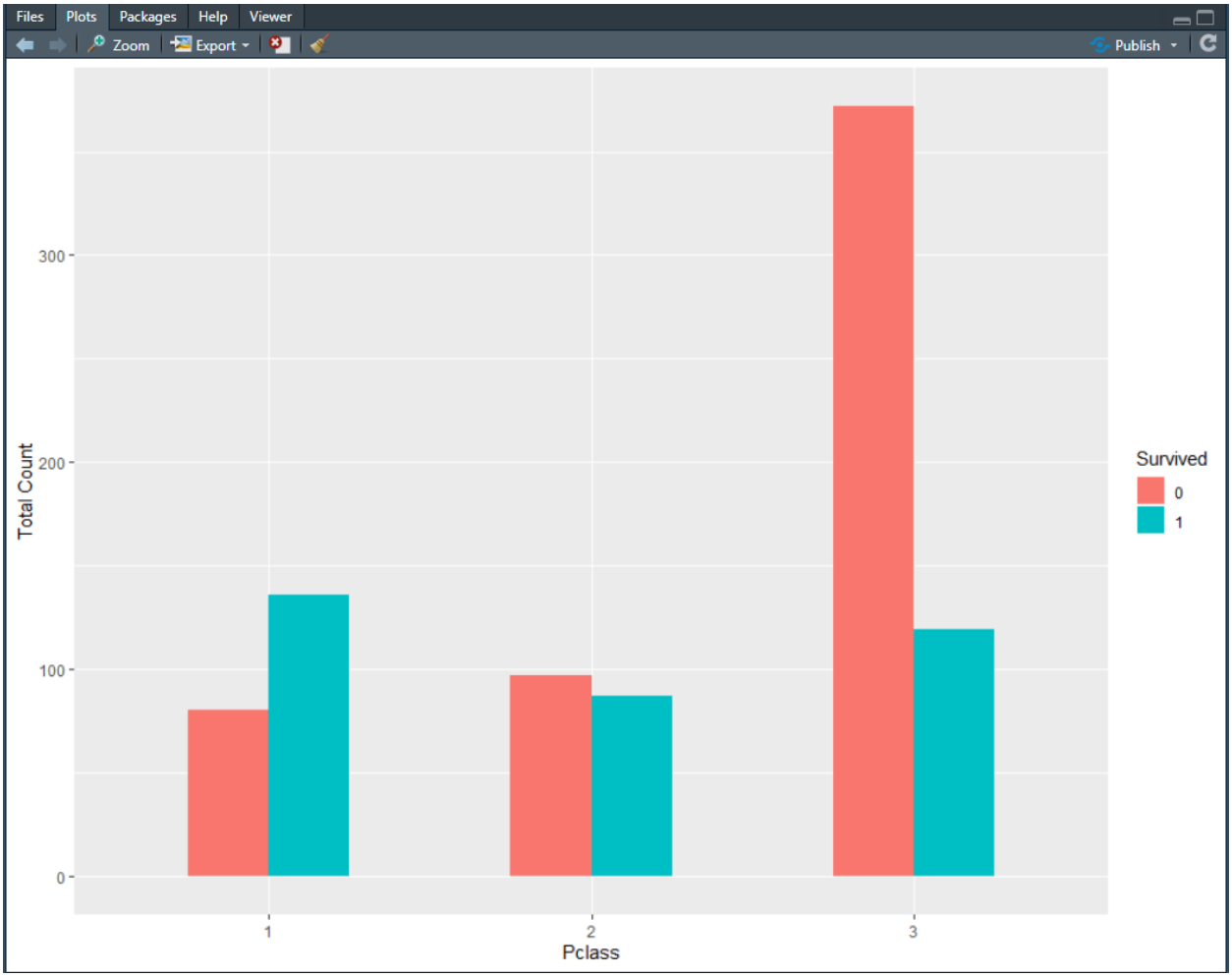
Passenger class is either 1st, 2nd, or 3rd.

N.B:In the upper marked there is an Export option to save your plot.But I don't ,because I am lazy.i do it my practice session.



The Pclass variable shows that most passengers in the train set held a 3rd class ticket. 491 of the 89 1 passengers were 3rd class, more than 1st and 2nd class combined.

If the Survived variable is plotted as a function of passenger class, it appears that Pclass will be a predictor for survivability. A higher percentage of first class passengers survived than died, contrary to the overall trend, whereas a far higher percentage of 3rd class passengers died than survived.(for 2nd plot)



Name

It may seem that the passenger name is a lot like the PassengerId in that each name acts as a sort of primary key into the data and using name as a model feature would not generalize well. However, the name field could possibly provide value. The first twenty names in the train dataset:

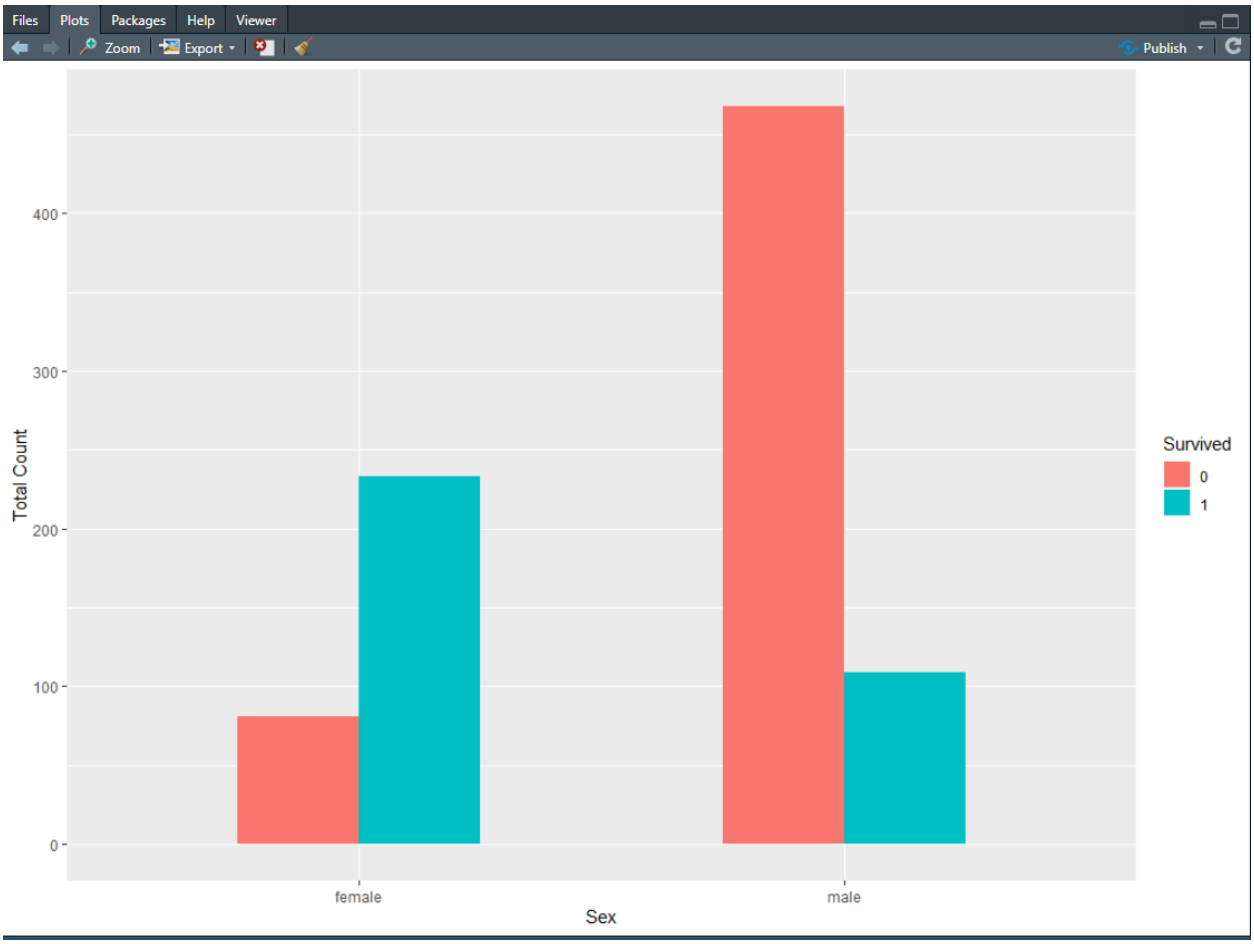
```
> head(as.character(train$Name),n=20);
 [1] "Braund, Mr. Owen Harris"                             "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
 [3] "Heikkinen, Miss. Laina"                              "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
 [5] "Allen, Mr. William Henry"                            "Moran, Mr. James"
 [7] "McCarthy, Mr. Timothy J"                             "Palsson, Master. Gosta Leonard"
 [9] "Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)"   "Nasser, Mrs. Nicholas (Adele Achem)"
[11] "Sandstrom, Miss. Marguerite Rut"                     "Bonnell, Miss. Elizabeth"
[13] "Saundercock, Mr. William Henry"                      "Andersson, Mr. Anders Johan"
[15] "Vestrom, Miss. Hulda Amanda Adolfina"                "Hewlett, Mrs. (Mary D Kingcome) "
[17] "Rice, Master. Eugene"                                "Williams, Mr. Charles Eugene"
[19] "Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)" "Masselmani, Mrs. Fatima"
>
```

Each name begins with a surname before a comma and a title. If the passenger is a married woman, her maiden name appears in parentheses. Some of the titles may help infer age (Master. and Miss.) and surnames could help determine extended family travelling together, even if they've purchased s eparate tickets and are not in the same cabin. Additionally, the presence of diacritical marks in a na me could indicate that the passenger is a non-English speaker who might have had difficulty underst anding instructions or the gravity of the situation.

Sex

Sex is an unordered factor, male or female.

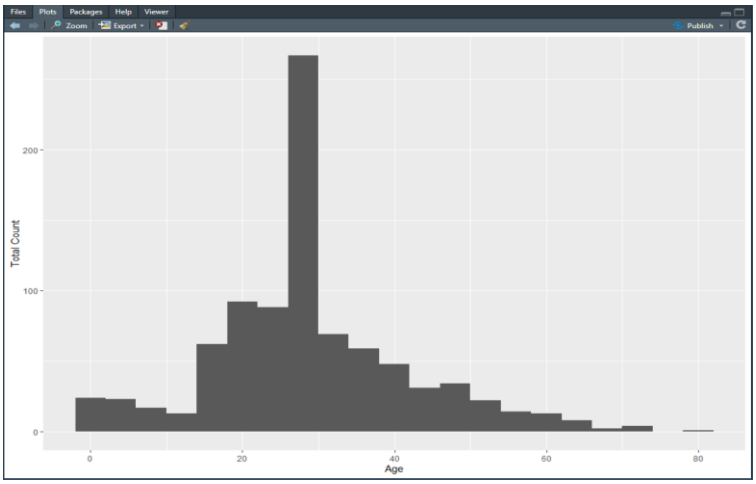
The Sex variable shows that most passengers in the training set were male (nearly 2/3 male). 577 of the 891 passengers were male, or 65%.

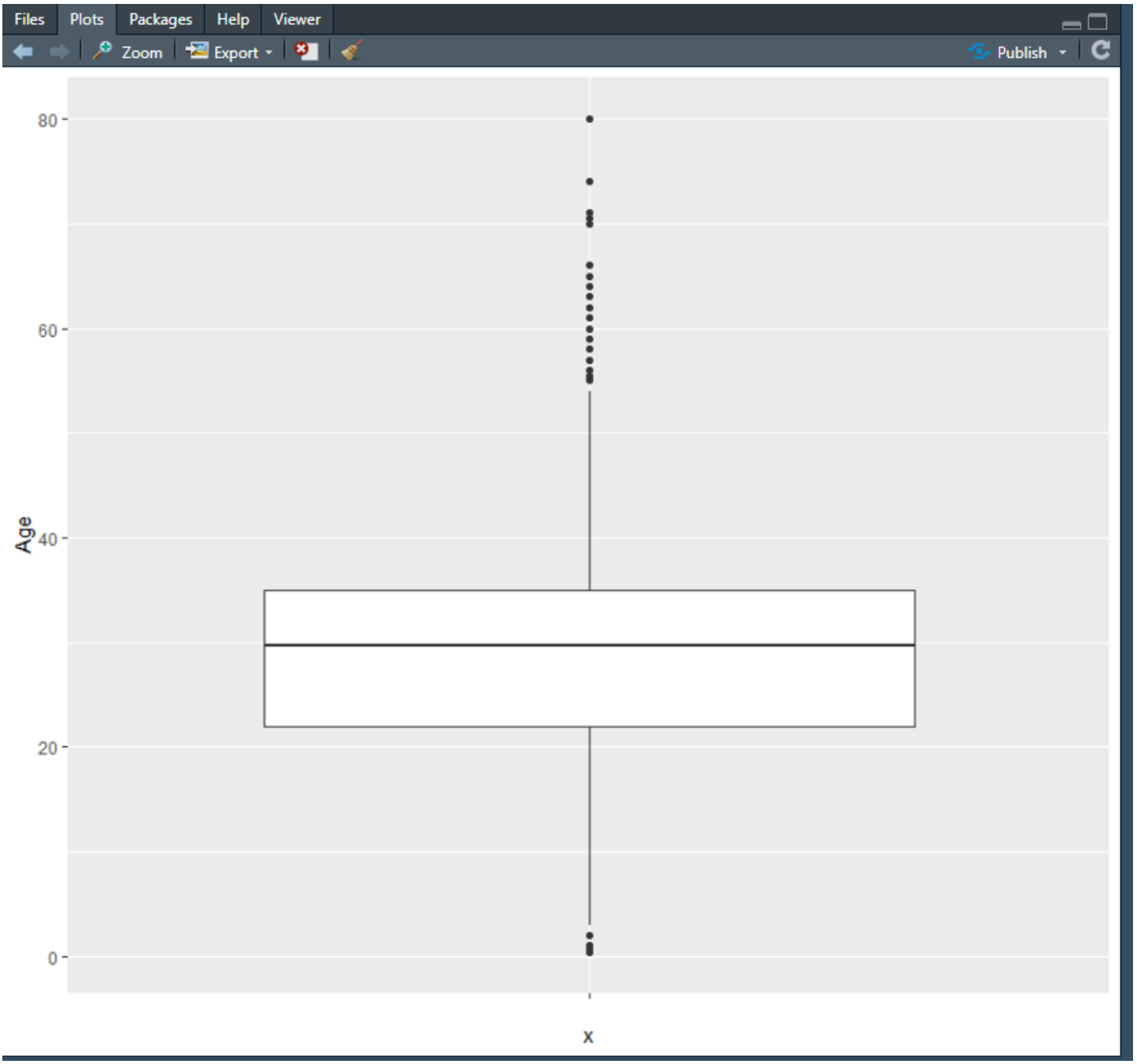


plotted as a function of Sex, it appears that Sex will be a strong predictor for survivability. Of the 314 female passengers in the training set, 233, or 74% survived. On the other hand, of the 577 male pas sengers, only 109 survived, or 19%.

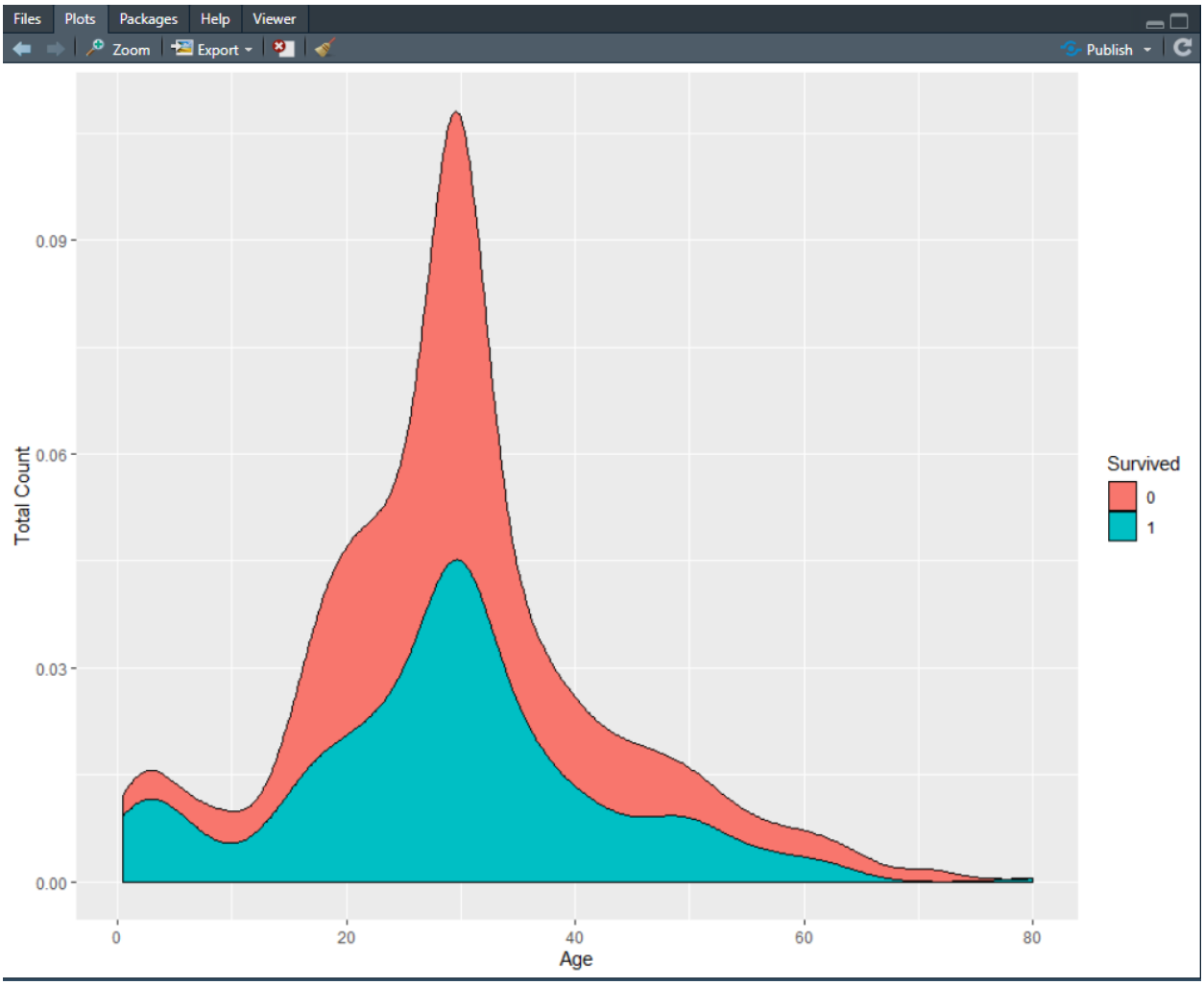
Age

there are age values for 80% of the training subjects, missing for 177 passengers. The distribution of ages is slightly skewed right, with a median of 28 years and a mean of 29.7 years.




Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
80
60
40
20
0
Age
x

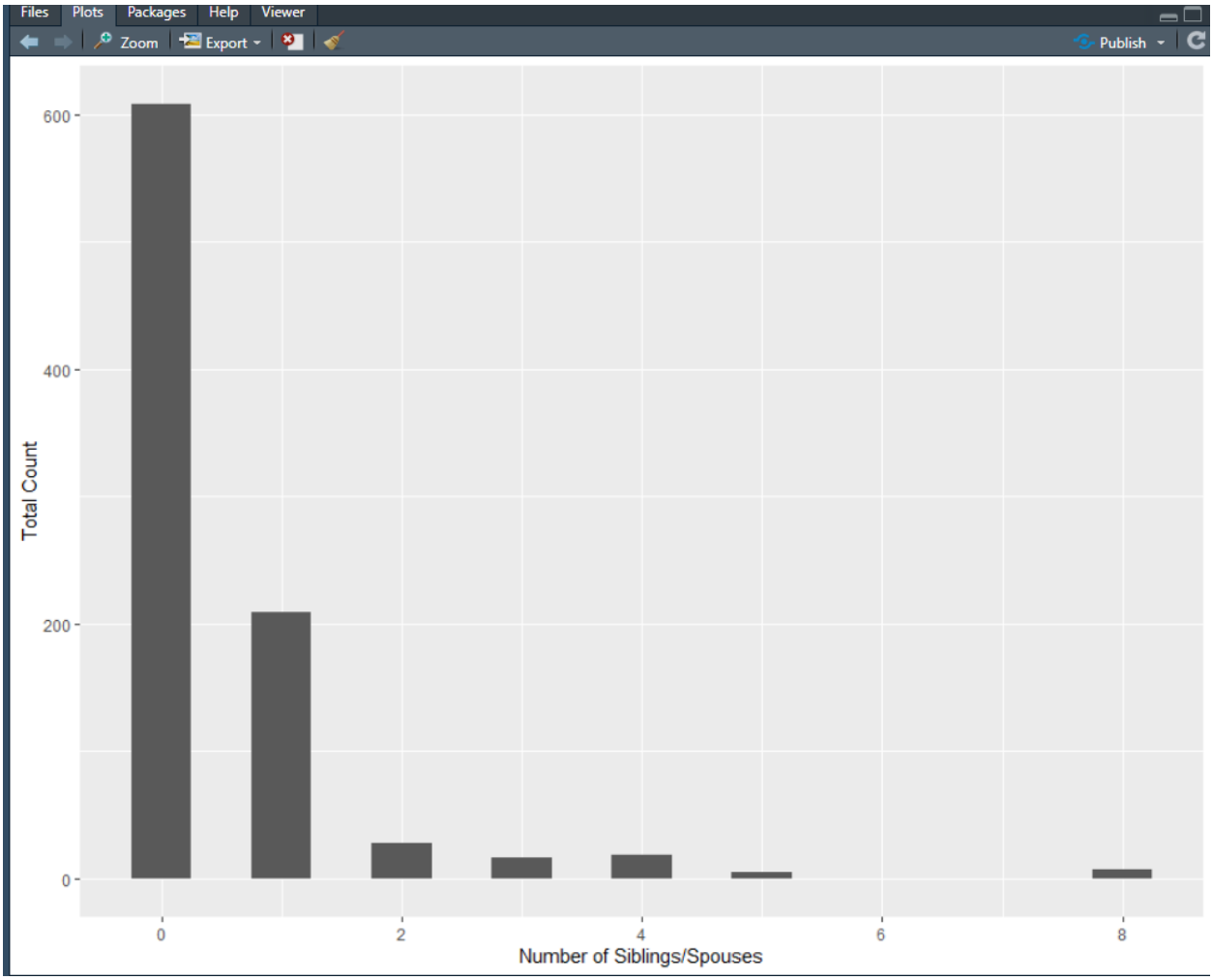
Age as a predictor of survivability



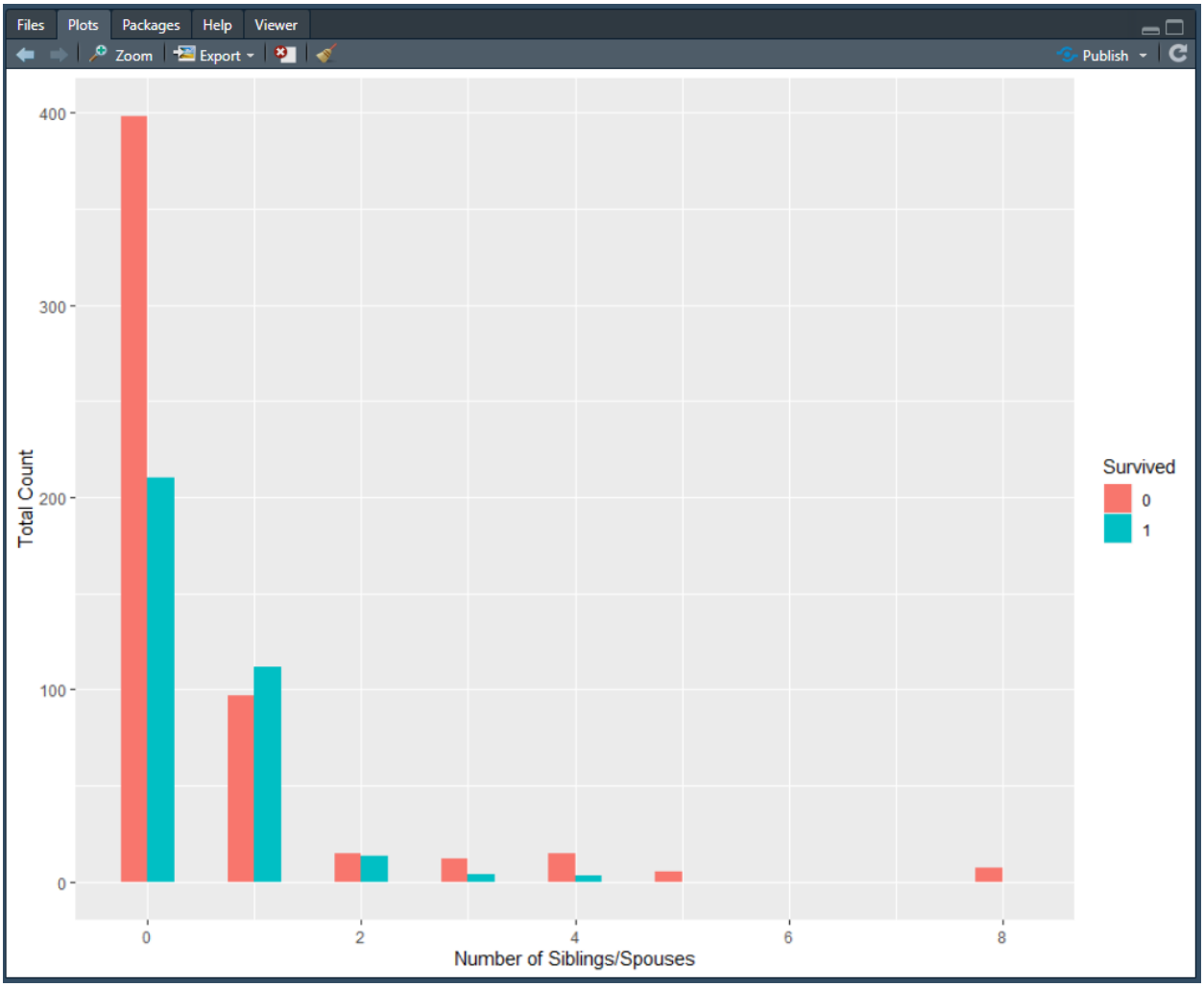
If we plot the Survived value as a function of the Age density, we see that there is a higher likelihood of younger passengers surviving over older passengers. Up until the mid-to-late teens, a training set passenger is more likely to survive than die, so age is likely to be a useful predictor for survivability.

SibSp

This variable is unique in that it is combination of number of siblings or “1” if the passenger had a spouse on board the ship. In some cases, it is not clear what the SibSp variable is encoding when a “1” is found - is the passenger travelling with a sibling or a spouse? A SibSp of 2 or more is indicative of siblings. It will likely be beneficial to disambiguate this variable into separate ‘Siblings’ and ‘Spouses’ variables.



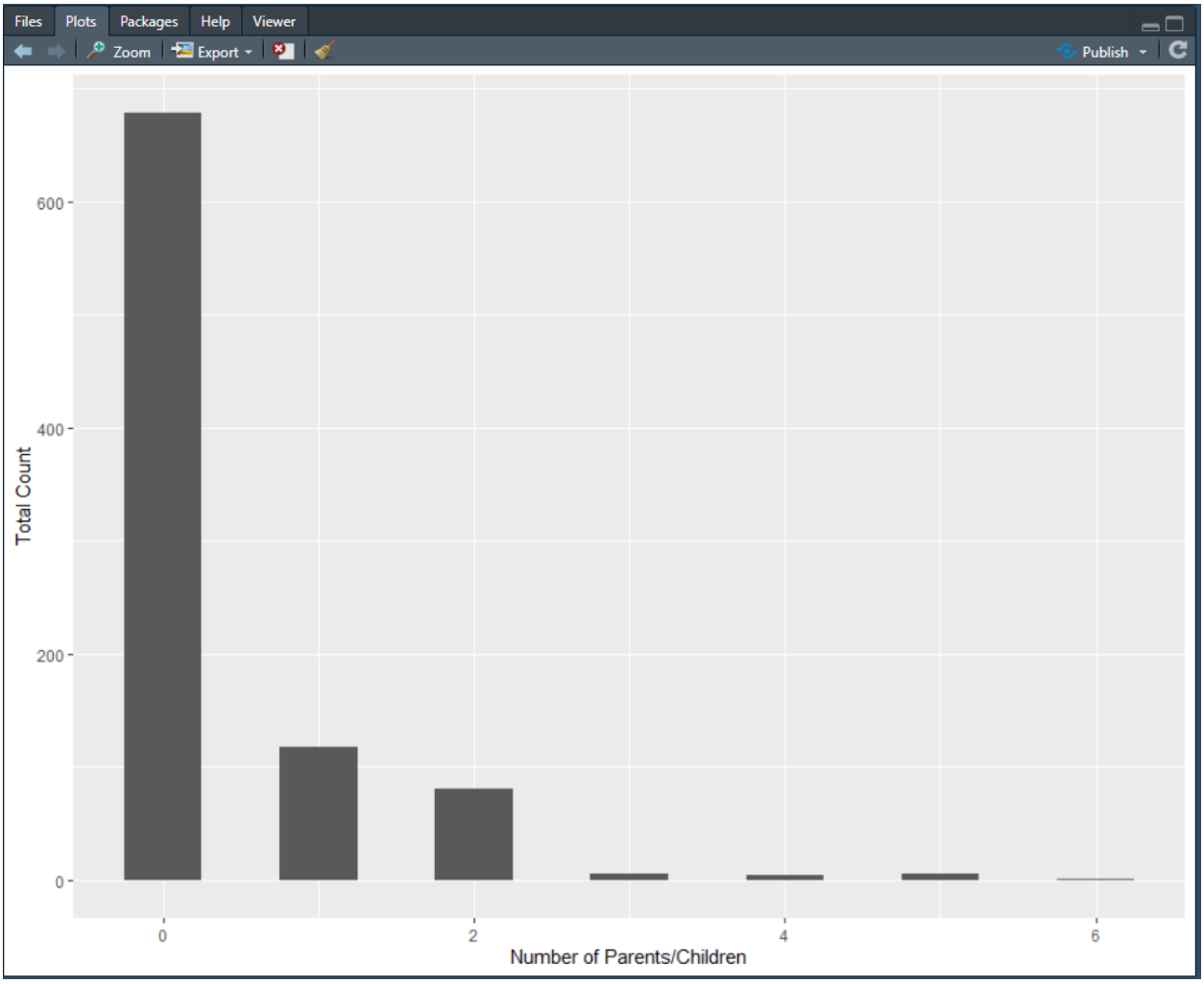
Most passengers in the training set (68%) had neither a spouse or sibling on board. 23% of the traini ng set passengers had a single sibling or spouse on board. The remaining 9% of the passengers ha d two or more (presumably) siblings on board.



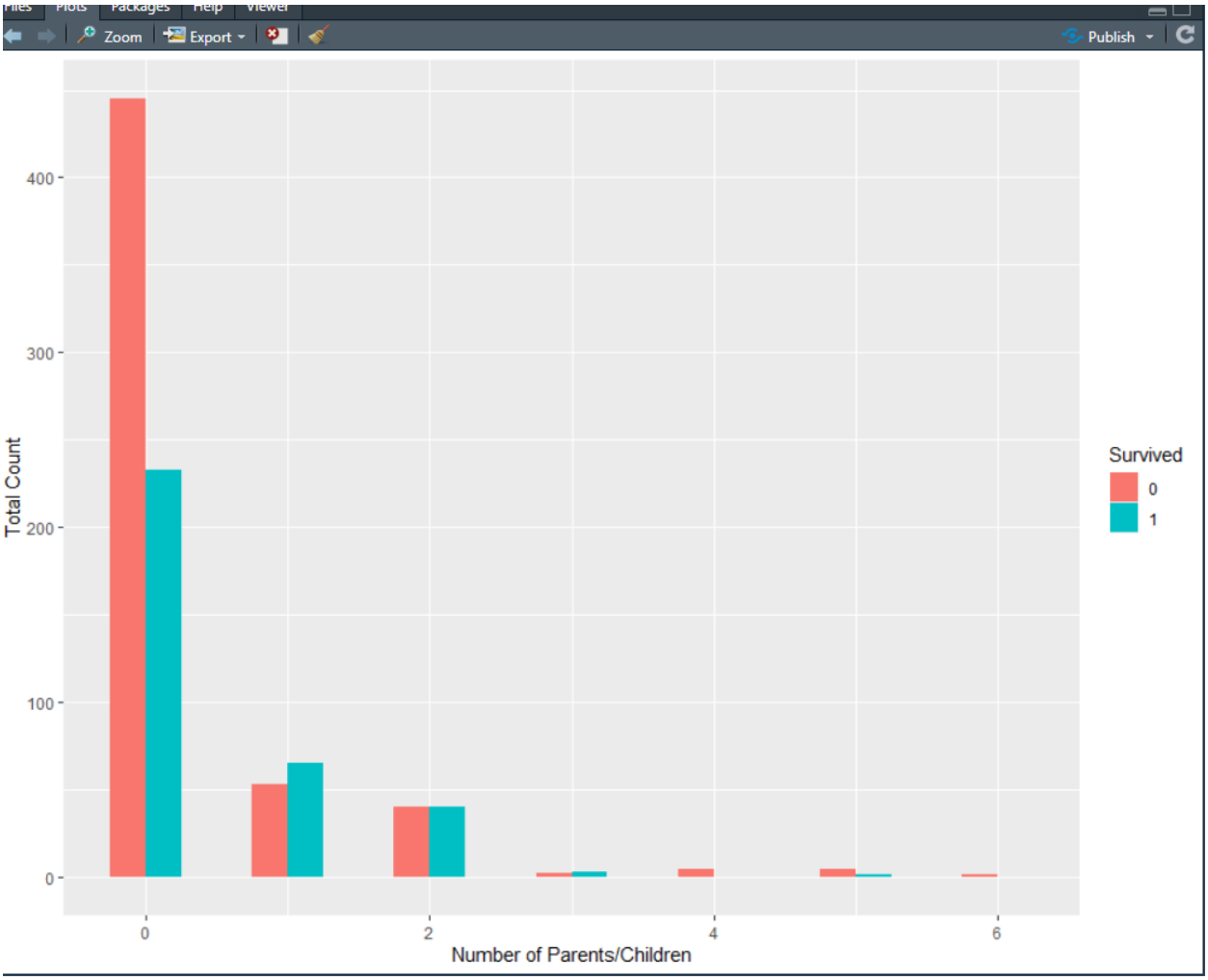
Survivability does appear to trend with the number of siblings/spouses on board. A passenger havin g no siblings or spouses is most likely to have died, whereas a passenger with one or two siblings/sp ouses has around a 50% likelihood of surviving. The remaining cases of three to eight siblings are lik ely too few from which to draw inferences individually, so it might make sense to pool the SibSp valu es as follows: 0, 1, 2, 3> to avoid overfitting to specific training cases. A close examination of the sev en instances of the SibSp variable in which SibSp equals 8 reveals that all the subjects were from th e same family and were in the same cabin. Predicting that all families of size 8 will perish is unlikley t o generalize well.

Parch

Similar to SibSp, this variable convolves two separate pieces of data: the number of parents and the number of children this passenger has on board. In some cases, it is not clear what the Parch variable is encoding when a “1” is found - is the passenger travelling with a parent or a child? This can be inferred if the Age variable is present for the passenger, but if the Age is missing, it will be ambiguous and may need further analysis. Perhaps the passenger’s title (Mr., Miss., Master) could help. Like SibSp, it will likely be beneficial to disambiguate this variable into separate ‘Parents’ and ‘Children’ variables.



Most passengers in the training set (76%) had neither a parent or child on board. 13% of the training set passengers had a single parent or child on board. About 9% of the passengers in the training set (80) had two or more parents or children on board. The remaining 2% of the passengers had either 3 , 4, 5, or 6 (presumably) children on board.



Survivability does appear to trend with the number of parents/children on board. A passenger having no parents or children is most likely to have died, whereas a passenger with one or two parents/child ren has around a 50% likelihood of surviving. The remaining cases of three to six children are likely t oo few from which to draw inferences individually, so it might make sense to group the Parch values as follows: 0, 1, 2, 3>= to avoid overfitting to specific training cases.

Ticket

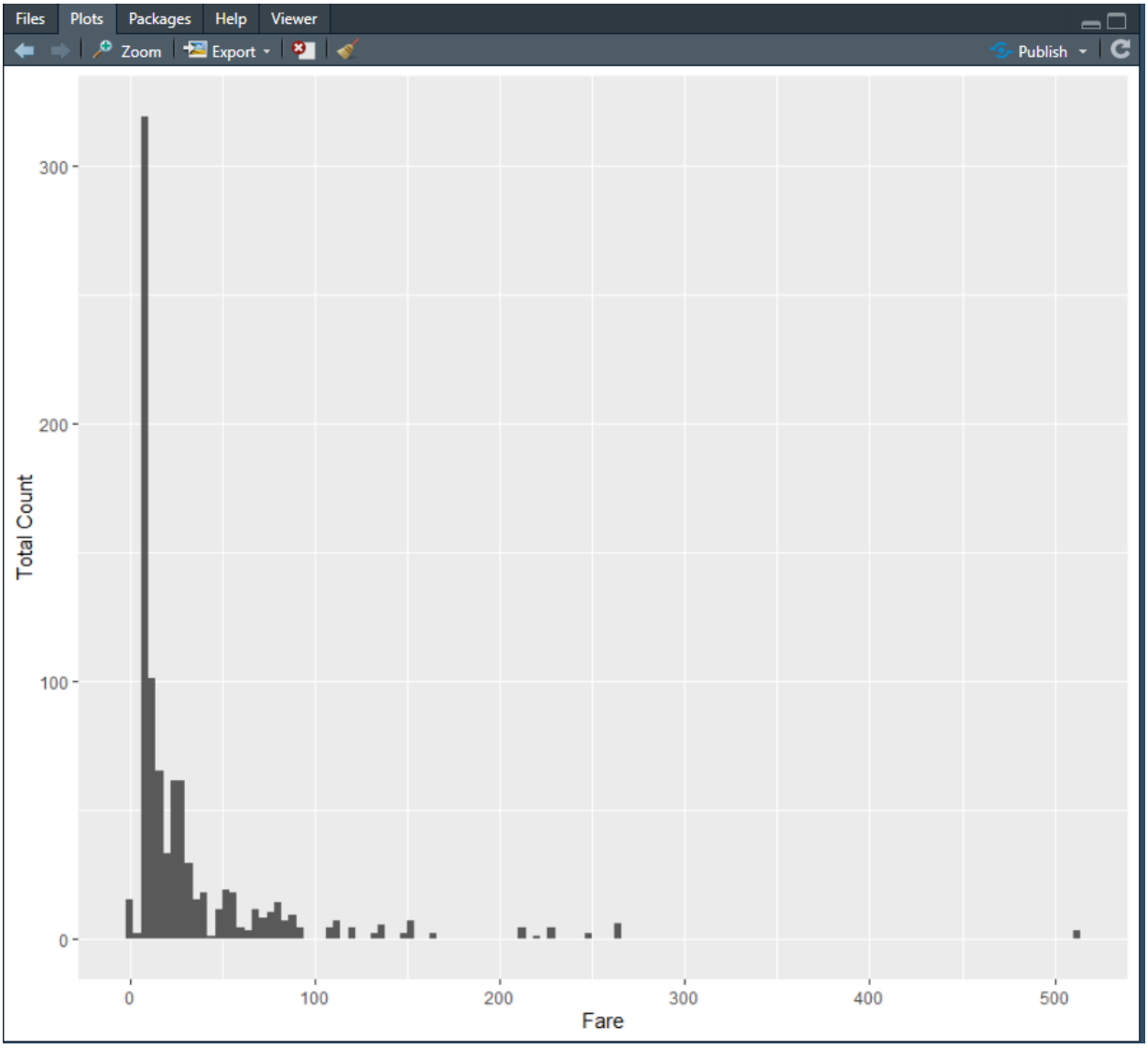
The entries in the Ticket column do not seem to be of a uniform format. Some ticket entries are just numbers - ranging from 693-392096. Other ticket entries have character prefixes like “C.A.” or “SOTON/O2”, followed by a (presumably) ticket number.

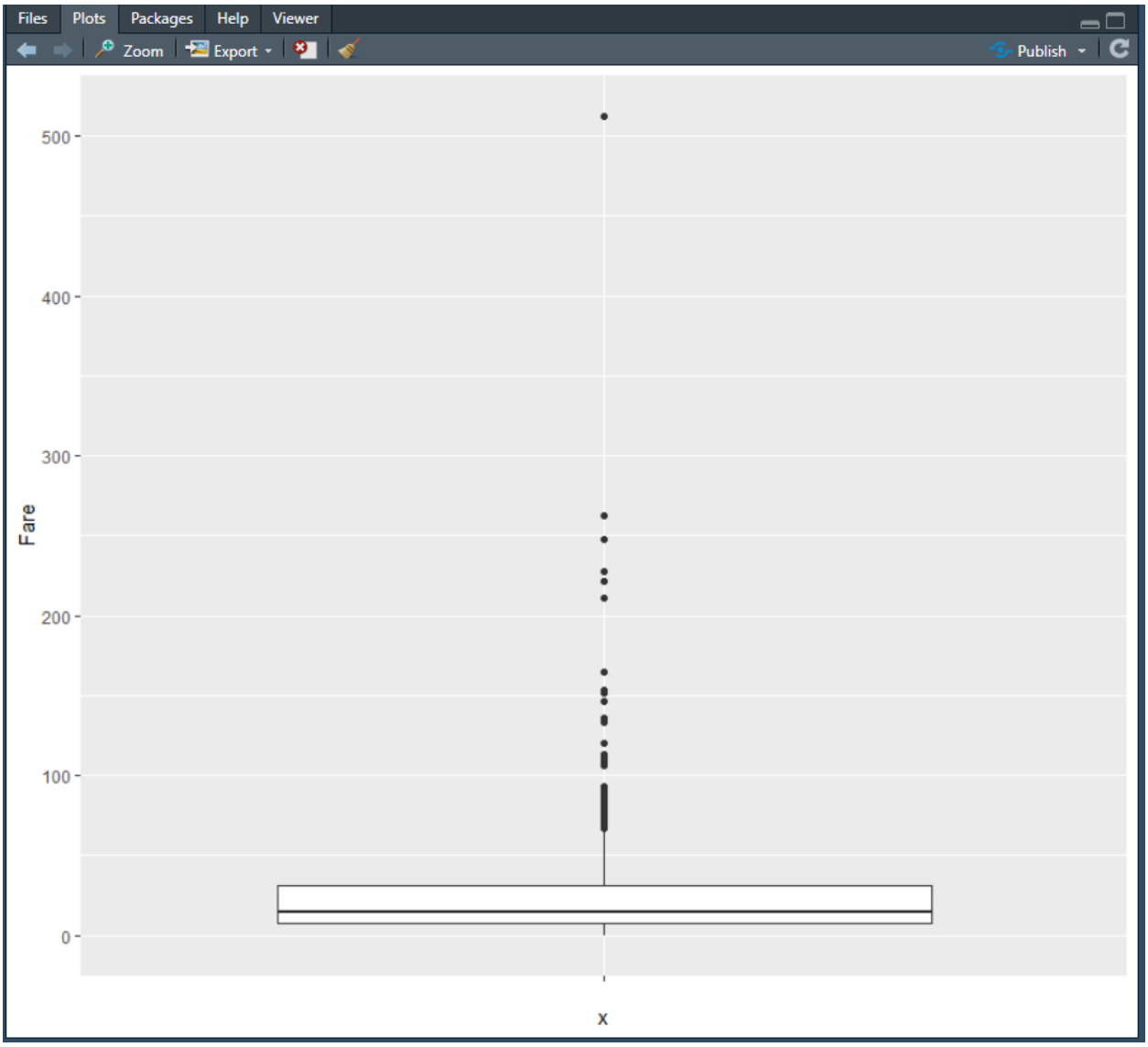
```
> head(as.character(train$Ticket),n=20);
 [1] "A/5 21171"        "PC 17599"         "STON/O2. 3101282" "113803"          "373450"          "330877"
 [7] "17463"            "349909"           "347742"           "237736"          "PP 9549"         "113783"
[13] "A/5. 2151"        "347082"           "350406"           "248706"          "382652"          "244373"
[19] "345763"           "2649"
```

An inspection of the set of tickets shows that presumed families tend to share a single ticket number. The first impression is that the ticket number seems an unlikely predictor for survivability and could l ead to overfitting the training set. However, the ticket number might help populate the missing Cabin information - and Cabin might be a good predictor for survivability.

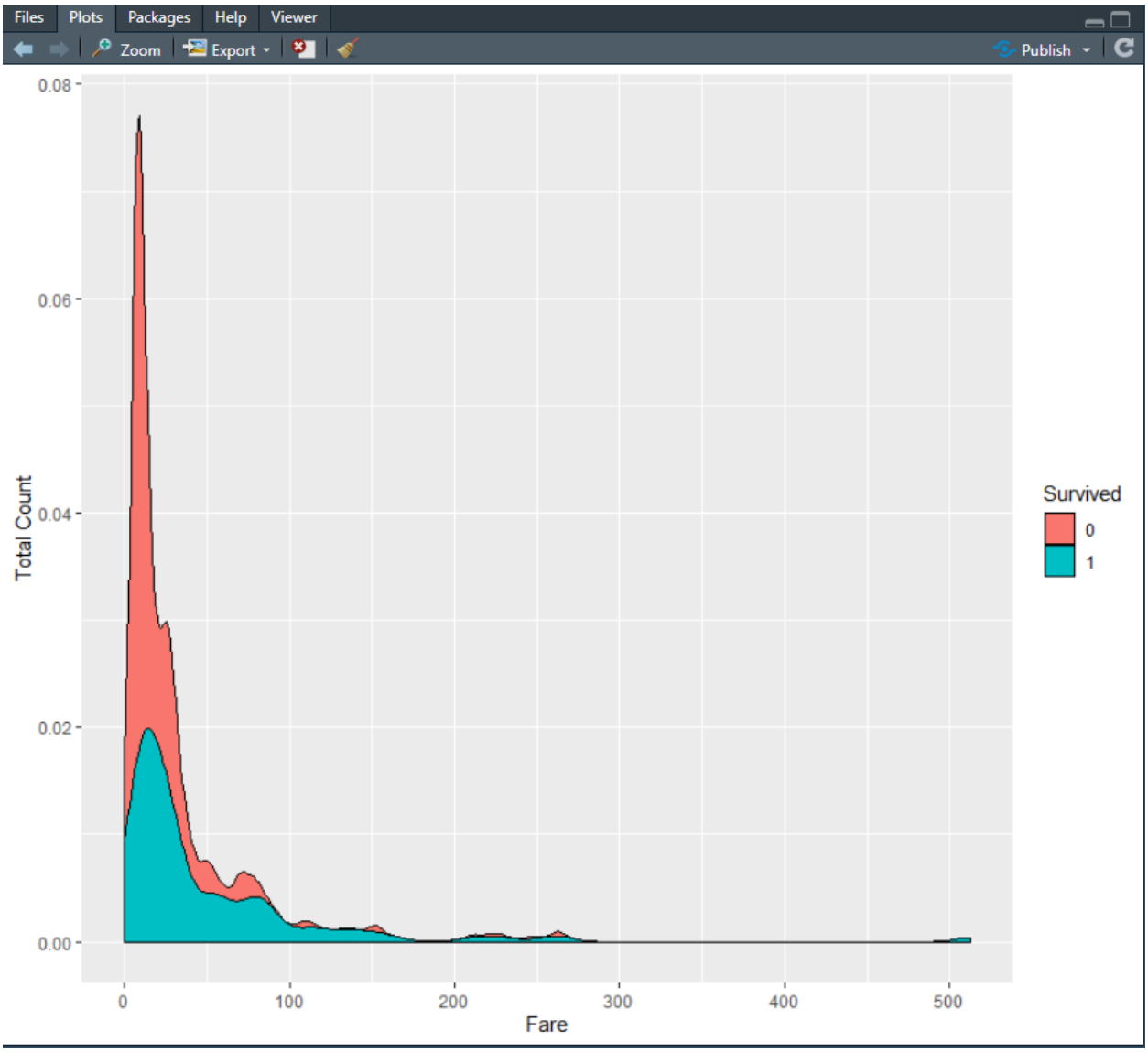
Fare

The fare (price paid per ticket) ranges from 0 to 512.3292. The units are unclear, but are likely in English pounds. The distribution is skewed to the right, with a median of 14.4542 and a mean of 32.20421. A log transform of the data may be necessary to normalize the distribution of fares. However, first the fare for passenger must be determined. It appears to be the case that individual ticket numbers are not assigned per passenger, but rather a single ticket number is given to the purchaser of an allotment of tickets. That is, families travelling together seem to be under the same ticket with the same fare. So, it may be necessary to get to create an “Amount Paid per Passenger” feature that takes into account the number of people for which a fare was purchased on a single ticket.



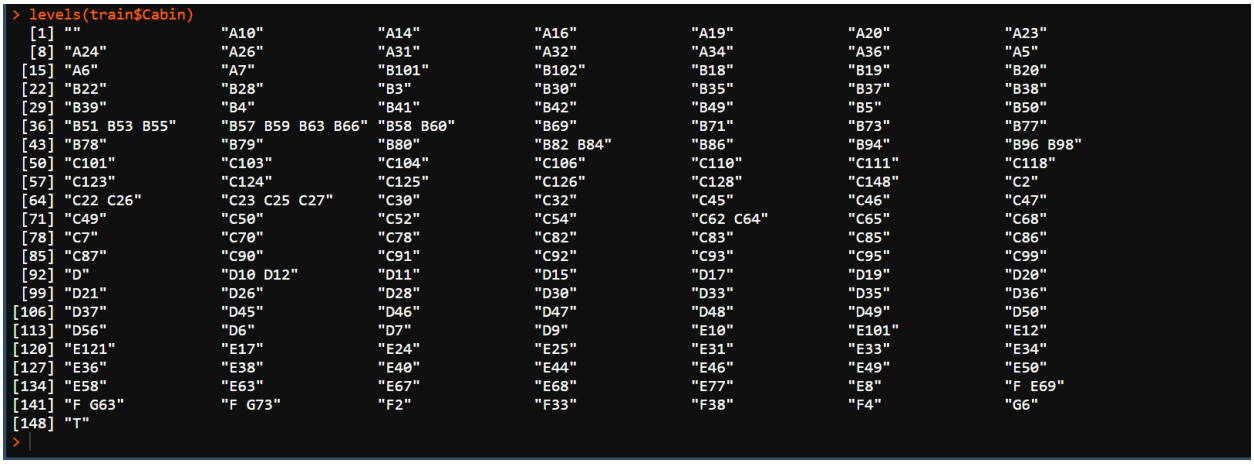


Fare could conceivably be an important factor in determining survivability. Perhaps the higher paying passengers received the first opportunity to board lifeboats. Or perhaps, those higher paying passen gers were more initially unwilling to leave their more comfortable accomodations for the plebian cond itions aboard a lifeboat. Fare as a predictor of survivability:



Cabin

The Cabin feature could be another strong predictor for survivability. Perhaps cabins located nearest the lifeboats afforded the best survivability. But, the Cabin variable has many empty values. The empty values could mean that the information was not captured or it could mean that not all passengers received cabins and stayed in other accomodations. Being assigned a cabin could be a proxy for one's social status and wealth. If so, the Pclass variable might be co-linear.



The cabin name mostly adheres to the rule of a single letter A-F,G,T, followed by a number up to 3 d igits. There are cases where a passenger has multiple cabins, each separated by whitespace. The b eginning letter of each cabin could denote a deck or particular region of the ship - which could help w ith predicting survivability. Alternatively, the number of the cabin could be more informative than the beginning letter. Perhaps cabins “A19” and “B19” are located right next to one another, for instance.

Some Resources(link)

1. https://www.udemy.com/course/r-programming/?utm_source=adwords&utm_medium=udemyads&utm_campaign=DataScience_v.PROF_la.EN_cc.ROW_ti.5336&utm_content=deal4584&utm_term=_._ag_85469003754_._ad_395279056262_._kw_._dec_._dm_._pl_._ti_dsa-774930027489_._li_9069458_._pd_._&matchtype=b&gclid=Cj0KCQiAwP3yBRCkARIsAABGiPp4MNAsWWATgBm1oXYACbH_KduSTnCjz6m28QEACJDXC1ZgcjUQ9WEaAmOEEALw_wcB Basically I complete this course. If anyone w anna this course ,I will provide.In this course actually “Advanced Visua lization” part more details ggplot and other part also describe metaplot . You can say that it’s a course which is best for neophytes of Learning R programing.

2. https://www.udemy.com/course/machinelearning/ From this course I complet e the “Data preprocessing part”.

3.Issing value related https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17

4.kaggle.com

N.B:I also help from internet , but I do it fully manually.