91/100
Good job

Ivan Calderoni
DSA 5103 Intelligent Data Analytics
Homework 3
Due: September 30, 2017

Problem 1: Glass Identification:

1.a) A quick glance inside the dataset "Glass" reveals a total of ten variables, Refractive Index (RI), 8 elements, and Type (1-7, each corresponding to different types of glasses). Here's a quick look:

```
> head(Glass)
       RI    Na   Mg   Al    Si    K   Ca Ba   Fe Type
1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00    1
2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00    1
3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00    1
4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00    1
5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00    1
6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26    1
```

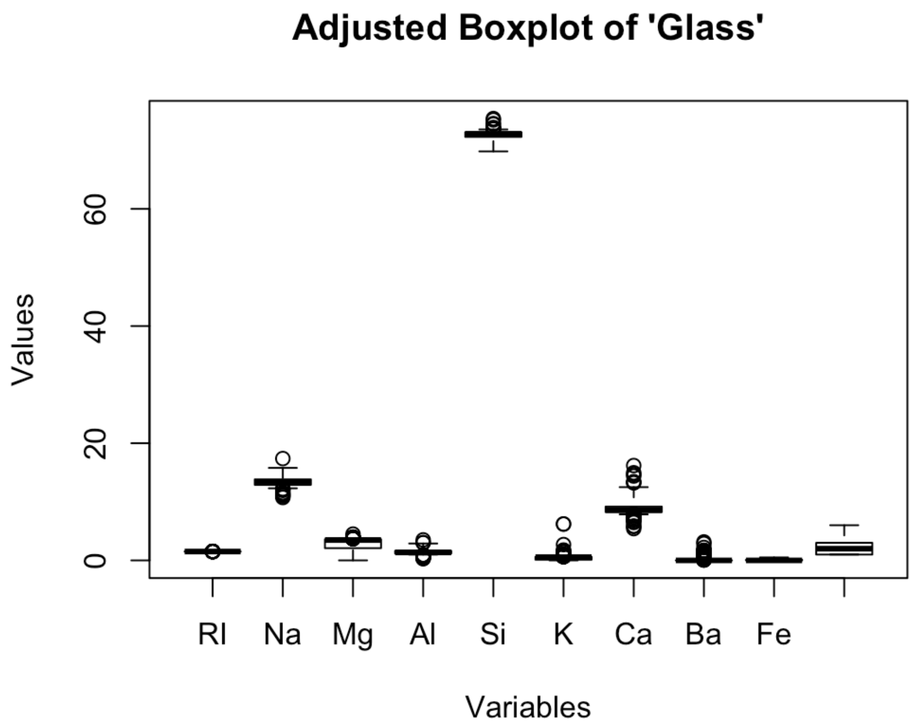
Using the describe() function from the psych library provides us with some useful statistics. Of interest, the standard deviation (sd), skew and kurtosis.

```
> describe(Glass)
      vars   n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
RI       1 214  1.52 0.00   1.52    1.52 0.00  1.51  1.53  0.02  1.60     4.72 0.00
Na       2 214 13.41 0.82  13.30   13.38 0.64 10.73 17.38  6.65  0.45     2.90 0.06
Mg       3 214  2.68 1.44   3.48    2.87 0.30  0.00  4.49  4.49 -1.14    -0.45 0.10
Al       4 214  1.44 0.50   1.36    1.41 0.31  0.29  3.50  3.21  0.89     1.94 0.03
Si       5 214 72.65 0.77  72.79   72.71 0.57 69.81 75.41  5.60 -0.72     2.82 0.05
K        6 214  0.50 0.65   0.56    0.43 0.17  0.00  6.21  6.21  6.46    52.87 0.04
Ca       7 214  8.96 1.42   8.60    8.74 0.66  5.43 16.19 10.76  2.02     6.41 0.10
Ba       8 214  0.18 0.50   0.00    0.03 0.00  0.00  3.15  3.15  3.37    12.08 0.03
Fe       9 214  0.06 0.10   0.00    0.04 0.00  0.00  0.51  0.51  1.73     2.52 0.01
Type*   10 214  2.54 1.71   2.00    2.31 1.48  1.00  6.00  5.00  1.04    -0.29 0.12
```

An adjusted boxplot was created to quickly visualize the distributions of the bulk data and to look at any possible outliers without making any parametric assumptions of the data. From the adjusted boxplot (showed on the next page), and in addition to individual histograms and the data from above (particularly skew and kurtosis), it appears that K, Ca and Ba have the most skewed distributions.

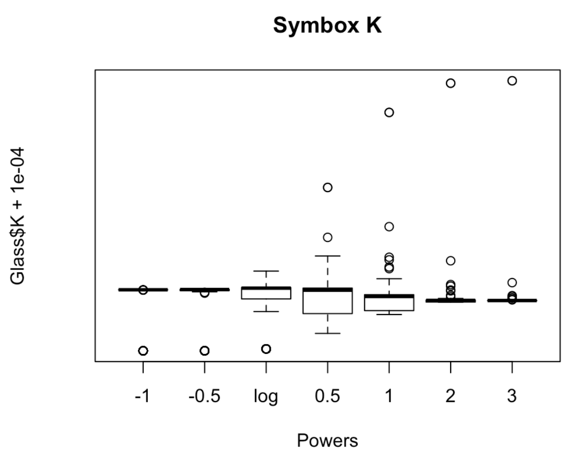
Alexander Rodríguez Castillo: ok

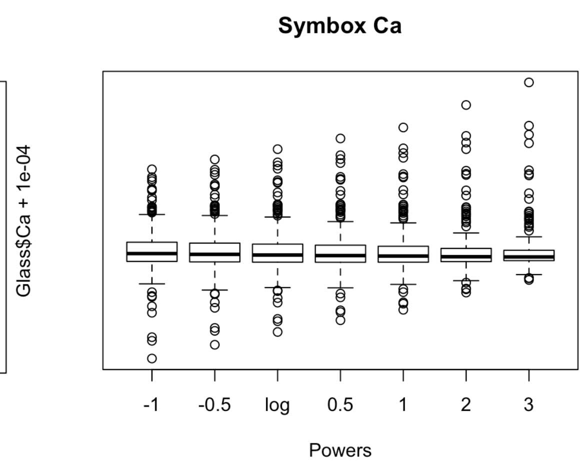
Alexander Rodríguez Castillo: ok

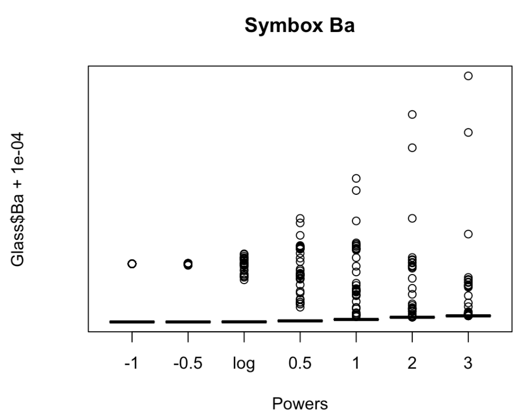


1.b) K, Ca and Ba were selected to be transformed to find out if their distributions can benefit from the transformations.

i)







Alexander Rodríguez Castillo: ok

The case could be made that a “Powers” transformation could be made for the variable K around 0.0 or log. For Ca around -1.0, and for Ba around -0.5.

ii) By using the boxcox method, we get the following optimal lambdas:

Element	Lambda
K	0.5
Ca	-1
Ba	0.0

Alexander Rodríguez Castillo: ok

1.c) It tells us that we can use PC1 through PC5 and explain about 90% of the variability.

Alexander Rodríguez Castillo: ok

```
> GlassPCA <- prcomp(Glass[,1:9], scale = TRUE)
> summary(GlassPCA)
Importance of components%s:
                         PC1    PC2    PC3    PC4    PC5     PC6    PC7     PC8     PC9
Standard deviation     1.585 1.4318 1.1853 1.0760 0.9560 0.72639 0.6074 0.25269 0.04011
Proportion of Variance 0.279 0.2278 0.1561 0.1286 0.1016 0.05863 0.0410 0.00709 0.00018
Cumulative Proportion  0.279 0.5068 0.6629 0.7915 0.8931 0.95173 0.9927 0.99982 1.00000
```

1.d) Principal Component Analysis (PCA) is an unsupervised learning technique where all the variables are treated independent from others. If a dataset has independent variables, one should use PCA. On the other hand, Linear Discriminant Analysis (LDA) is a supervised technique that takes into account class information. So if variables affect one another (for example: number of hours studying, number of hours of sleep before a test, and test scores), it is better to use LDA.

Alexander Rodríguez Castillo: -3 this is too general, you conclusion comparing PCA and LDA should consider the objective of doing these two feature reduction techniques, which is glass classification

```
> table(GlassPredict, Glass[,10])

GlassPredict  1  2  3  5  6  7
           1 52 17 11  0  1  1
           2 15 54  6  5  2  2
           3  3  0  0  0  0  0
           5  0  3  0  7  0  1
           6  0  2  0  0  6  0
           7  0  0  0  1  0 25
```

Problem 2: Missing Data:

2.a) First, deleted all rows with na value(s) – missing value(s). Listwise Deletion Coefficients:

```
> ListwiseDeletionCoefficients
      (Intercept)              year   countryIndonesia       countryKorea    countryMalaysia
    -2.650433e+02       3.580765e-01      -1.900660e+02      -2.254931e+02      -2.318437e+02
     countryNepal    countryPakistan countryPhilippines   countrySriLanka    countryThailand
    -2.270878e+02      -1.616933e+02      -2.103454e+02      -2.168838e+02      -2.014832e+02
           polity                pop             gdp.pc           intresmi             signed
    -1.902494e-01      -2.111286e-01       2.910265e-04       2.929493e-01      -1.288913e+00
           fiveop              usheg
    -1.579368e+01       9.582074e+00
```

Alexander Rodríguez Castillo: ok

2.b) First, computed the mean for every column that had missing value(s) and then use those values for imputation.

```
> meanImputationCoefficients
   (Intercept)           year countryIndonesia     countryKorea  countryMalaysia
  1.633387e+03  -7.938926e-01    -4.620179e+01    -5.937894e+01    -5.531281e+01
  countryNepal countryPakistan countryPhilippines countrySriLanka countryThailand
 -4.631776e+01  -1.440892e+01    -5.033981e+01    -4.536998e+01    -4.141807e+01
        polity            pop           gdp.pc         intresmi           signed
 -2.111236e-01  -2.628999e-02     5.922484e-04    -6.674644e-01     2.872480e+00
        fiveop          usheg
  2.254838e+00  -1.988981e+01
```

Alexander Rodríguez Castillo: ok

2.c) This was the particular model used for multiple imputation:

```
> imputationMethod <- c(year = "rf", country = "mean", tariff = "pmm", polity = "rf", pop = "sample",
gdp.pc = "rf", intresmi = "mean", signed = "rf", fiveop = "rf", usheg = "pmm")
```

rf = random forest; pmm = predictive mean matching

Alexander Rodríguez Castillo: ok

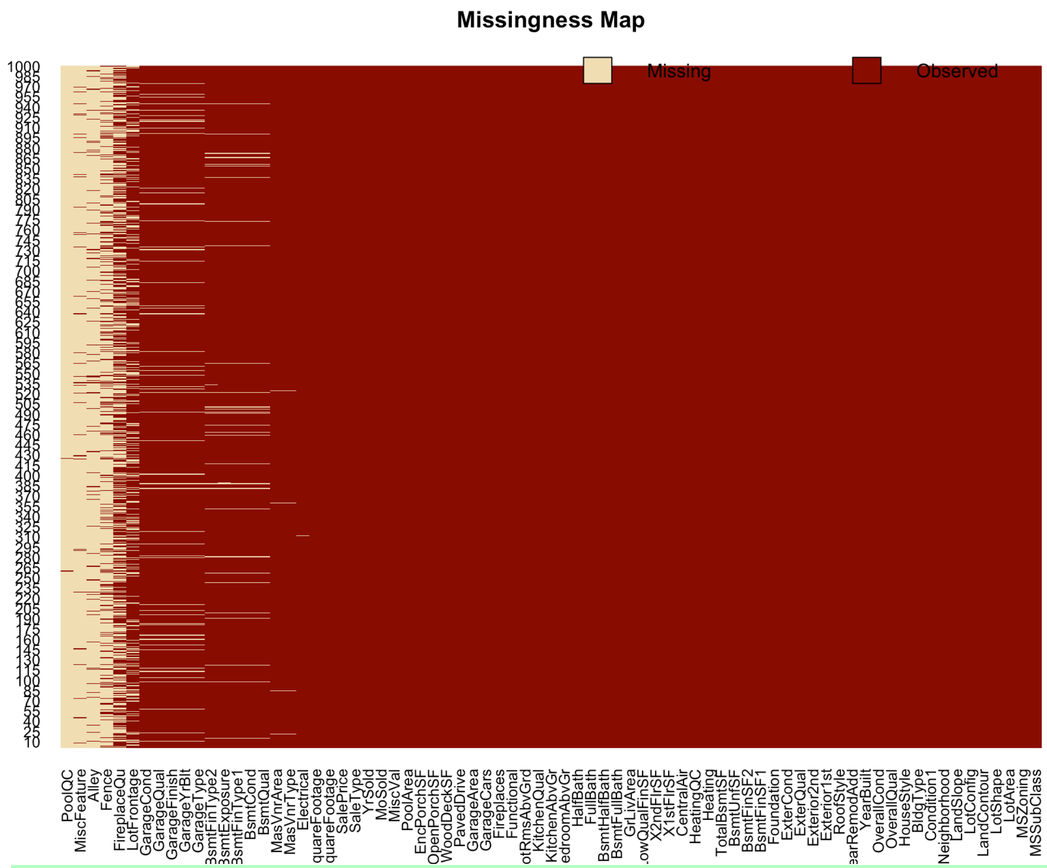
2.d) The coefficients for Listwise Deletion, Mean Imputation and Multiple Imputation were combined to easily compare results.

	ListwiseDeletionCoefficients	meanImputationCoefficients	MultipleImputationCoefficients
(Intercept)	-2.650433e+02	1.633387e+03	1.932475e+03
year	3.580765e-01	-7.938926e-01	-8.354816e-01
countryIndonesia	-1.900660e+02	-4.620179e+01	-1.081984e+02
countryKorea	-2.254931e+02	-5.937894e+01	-1.422197e+02
countryMalaysia	-2.318437e+02	-5.531281e+01	-1.380054e+02
countryNepal	-2.270878e+02	-4.631776e+01	-1.255774e+02
countryPakistan	-1.616933e+02	-1.440892e+01	-8.705671e+01
countryPhilippines	-2.103454e+02	-5.033981e+01	-1.269665e+02
countrySriLanka	-2.168838e+02	-4.536998e+01	-1.258212e+02
countryThailand	-2.014832e+02	-4.141807e+01	-1.197762e+02
polity	-1.902494e-01	-2.111236e-01	1.379656e-01
pop	-2.111286e-01	-2.628999e-02	-1.097251e-01
gdp.pc	2.910265e-04	5.922484e-04	7.816201e-04
intresmi	2.929493e-01	-6.674644e-01	-1.082014e+00
signed	-1.288913e+00	2.872480e+00	-2.164550e-02
fiveop	-1.579368e+01	2.254838e+00	-6.964024e+00
usheg	9.582074e+00	-1.988981e+01	-7.921580e+01

Alexander Rodríguez Castillo: -2 comments about these values? the comparison doesn't end with showing a table

Problem 3: House Prices Data:

3.a) To explore the data I first had to refer to the "housingVariables.pdf" to get an idea of the different variables in this particular dataframe and to understand what they mean. I wanted to find the completeness of the dataframe so I started by displaying a missingness map:



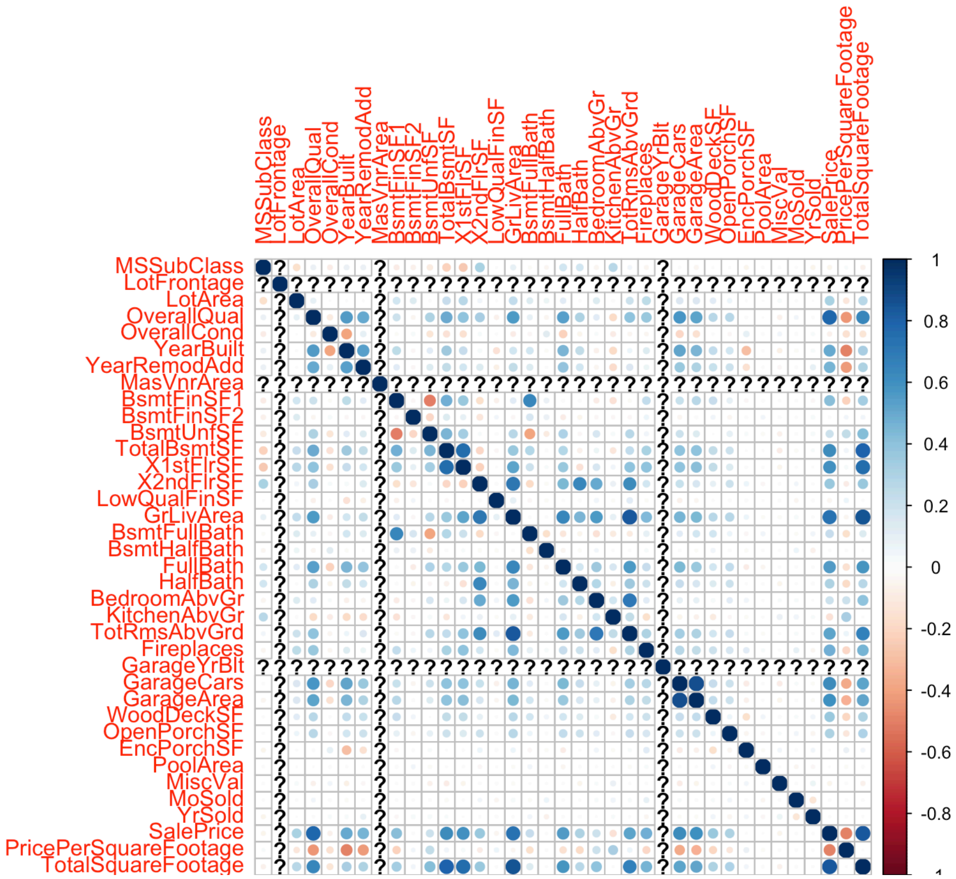
One can see that PoolQC, MiscFeature, Alley, Fence, FireplaceQu, and LotFrontage have at least 25% or more of missing values.

Alexander Rodríguez Castillo: ok

```
> which(colSums(is.na(housingData))/nrow(housingData) >= 0.25)
      Alley FireplaceQu      PoolQC       Fence MiscFeature 
          5          53          66          67          68 
```

Then, I proceeded to find out which numeric variables might have a high correlation with the variable "Sale Price." Some of the numeric variables that have a high correlation with "Sale Price" include:

OverallQual, GrLivArea, GarageCar, GarageArea, and a variable that I created which will be explained in 2b, TotalSquareFootage.



3.b) The features I created are TotalSquareFootage which is needed to then calculate PricePerSquareFootage.

The features use to calculate TotalSquareFootage were variables that count towards the gross living area (GLA). This does not include unfinished living spaces such as patios or garages; only finished spaces.

Alexander Rodríguez Castillo: ok

TotalSquareFootage = (housingData$TotalBsmtSF + housingData$X1stFlrSF + housingData$X2ndFlrSF)

PricePerSquareFootage:

Alexander Rodríguez Castillo: -2 it's ok, but you could have done more here

PricePerSquareFootage = (housingData$TotalBsmtSF + housingData$X1stFlrSF + housingData$X2ndFlrSF) / (housingData$SalePrice)

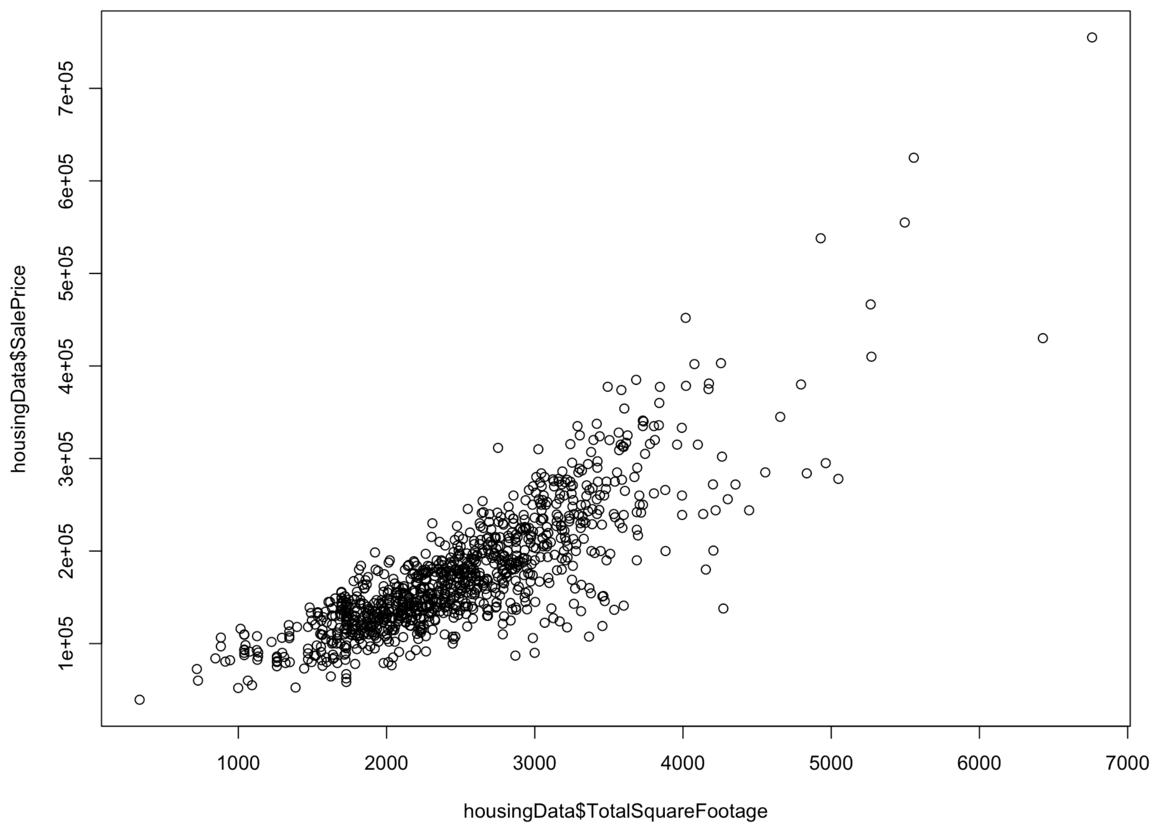
3.c) There are many variables people look at when purchasing a new home such as neighborhood, sale price, school district, time it takes to commute to work, etc.

Usually after some time, with these and other variables in mind, one narrows the house hunt to a few options. It is of importance to know how much one would pay per square footage.

Alexander Rodríguez Castillo: ok

This dataframe did not contain a variable that gave you this information. I calculated the TotalSquareFootage with an accurate sum of "Finished Space" square footage only (i.e. one could lay large amounts of patio bricks or lay a slab of concrete in the backyard and add that on to the total square footage of the house and list that number on zillow.com or use that to increase the selling price).

One would expect that with higher the square footage of the house, the higher the sale price:



Problem 4: kaggle.com – A LITTLE MORE DATA UNDERSTANDING:

4.a) The Titanic dataset was picked for this problem; which is an introductory competition to those interested in data science. The competition requires the user to analyze different features such as gender, age, economic status, etc. to predict if one would survive the tragedy we all know happened to the Titanic.

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

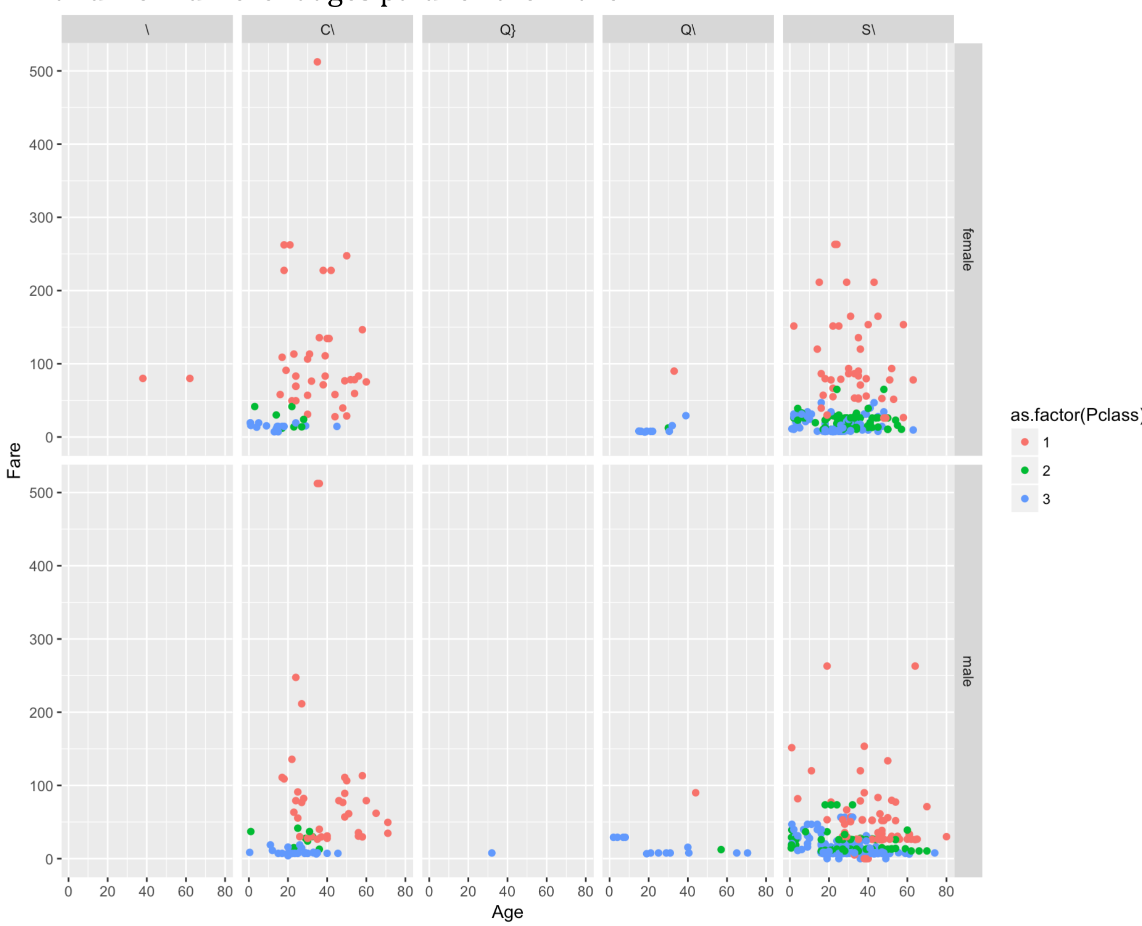
Alexander Rodríguez Castillo: ok

4.b) I am using the train dataset to find out what is in it:

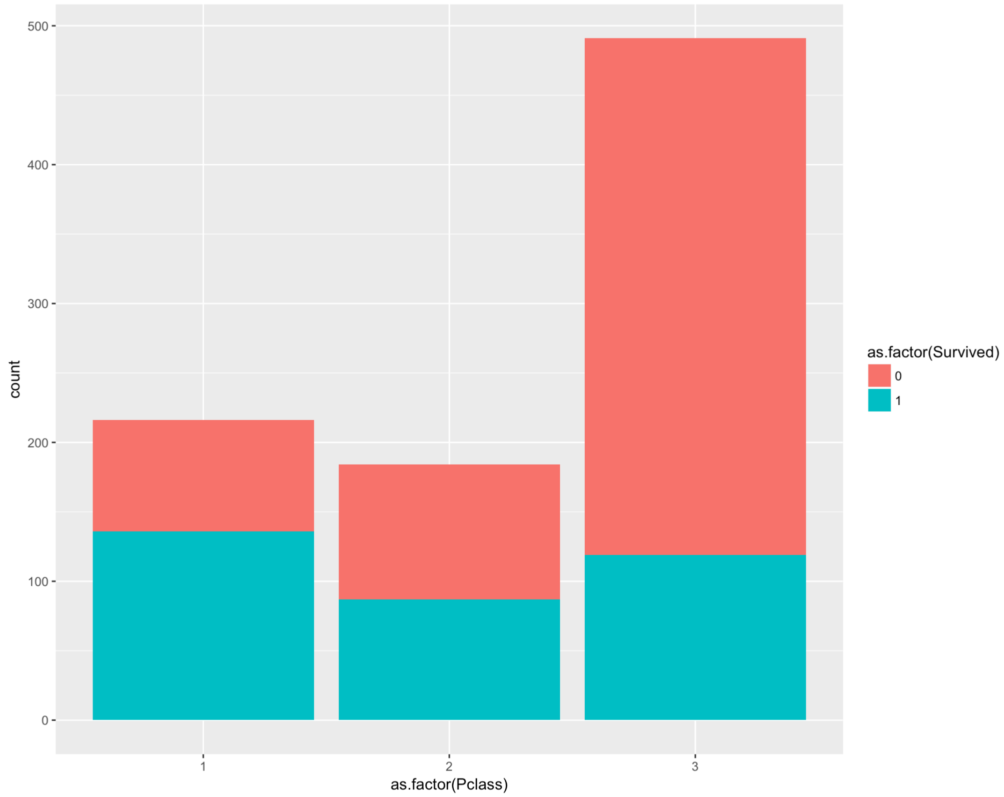
- 891 rows
- 12 columns = 12 features
- Descriptive (interesting) statistics:

```
> describe(Titanic)
            vars   n   mean     sd median trimmed    mad  min    max  range  skew kurtosis   se
PassengerId    1 891 446.00 257.35 446.00  446.00 330.62 1.00 891.00 890.00  0.00    -1.20 8.62
Survived       2 891   0.38   0.49   0.00    0.35   0.00 0.00   1.00   1.00  0.48    -1.77 0.02
Pclass         3 891   2.31   0.84   3.00    2.39   0.00 1.00   3.00   2.00 -0.63    -1.28 0.03
Name*          4 891 446.00 257.35 446.00  446.00 330.62 1.00 891.00 890.00  0.00    -1.20 8.62
Sex*           5 891   1.65   0.48   2.00    1.68   0.00 1.00   2.00   1.00 -0.62    -1.62 0.02
Age            6 714  29.70  14.53  28.00   29.27  13.34 0.42  80.00  79.58  0.39     0.16 0.54
SibSp          7 891   0.52   1.10   0.00    0.27   0.00 0.00   8.00   8.00  3.68    17.73 0.04
Parch          8 891   0.38   0.81   0.00    0.18   0.00 0.00   6.00   6.00  2.74     9.69 0.03
Ticket*        9 891 339.52 200.83 338.00  339.65 268.35 1.00 681.00 680.00  0.00    -1.28 6.73
Fare          10 891  32.20  49.69  14.45   21.38  10.24 0.00 512.33 512.33  4.77    33.12 1.66
Cabin*        11 891  18.63  38.14   1.00    8.29   0.00 1.00 148.00 147.00  2.09     3.07 1.28
Embarked*     12 891   4.34   1.18   5.00    4.55   0.00 1.00   5.00   4.00 -1.40     0.15 0.04
```

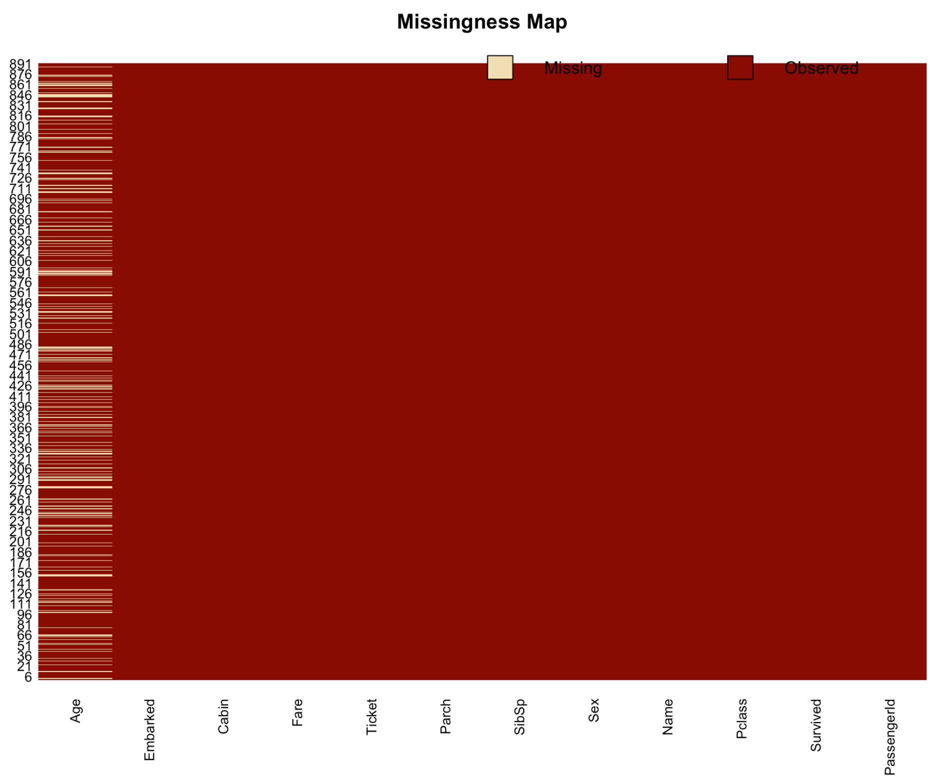
- This chart shows how much passengers from different classes, departing from different ports (Titanic made a few stops before embarking to the USA), and from different ages paid for their fare.



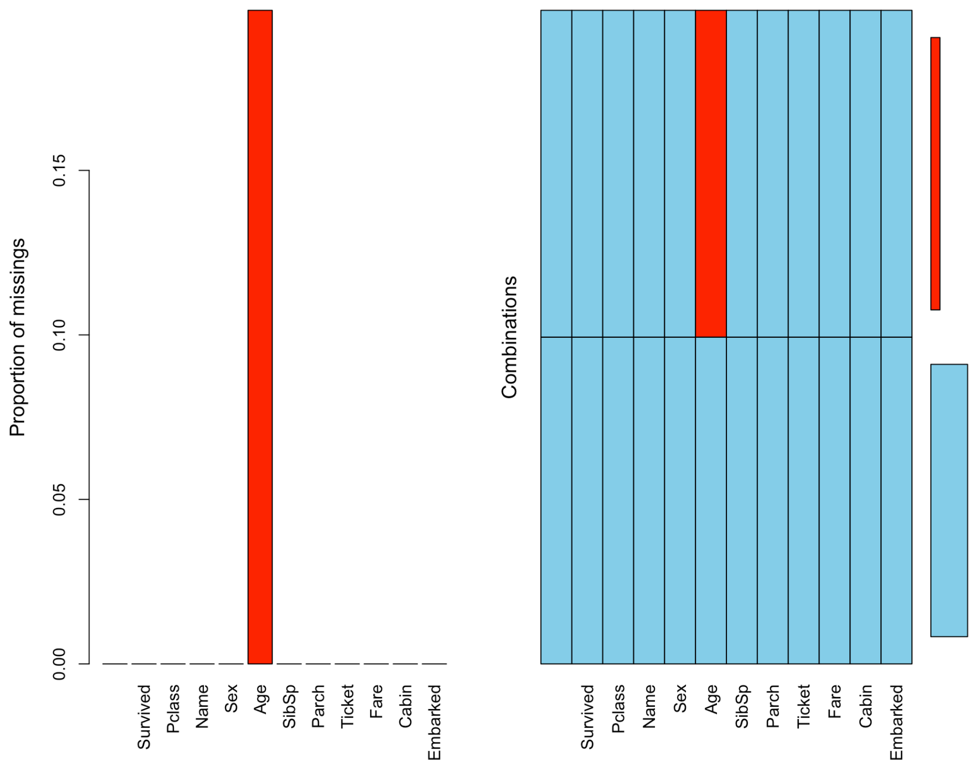
- Also of interest is visualizing who survived the wreck separated by class group (first, second and third class):



- This data is missing some values:



Alexander Rodríguez Castillo: -2 you are not explaining properly your plots



- An adjusted box plot is created to quickly visualize distribution and possible outliers: