LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

Chapter 0: Additional Material

Bernd Bischl, Christoph Molnar

Department of Statistics – LMU Munich

Winter term 2017/18

BOSTON HOUSING DATA SET

A widely used dataset to benchmark algorithms is the Boston housing dataset. The data was originally published 1978 by David Harrison and Daniel Rubinfeld in **Hedonic Housing Prices and the Demand for Clean Air**.
This paper investigates the methodological problems associated with the use of housing market data to **measure the willingness to pay for clean air**.

BOSTON HOUSING DATA SET

Example Data: Boston Housing

Variable	Description
medv	median value of owner-occupied homes in USD 1000's
crim	per capita crime rate by town
zn	prop. of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the prop. of blacks by town
lstat	percentage of lower status of the population

506 obs., 13 features, 'medv' numerical target.

BOSTON HOUSING DATA SET

Importing the Data

We use Open ML (R-Package) to download the dataset in a machine-readable format and convert it into a 'data.frame':

```
## # A tibble: 506 x 14
##       CRIM    ZN INDUS CHAS    NOX    RM   AGE   DIS RAD     TAX PTRATIO     B LSTAT  MEDV
##  *   <dbl> <dbl> <dbl> <fct> <dbl> <dbl> <dbl> <dbl> <fct> <dbl>   <dbl> <dbl> <dbl> <dbl>
##  1 0.00632  18    2.31 0     0.538  6.58  65.2  4.09 1       296    15.3  397.  4.98  24
##  2 0.0273    0    7.07 0     0.469  6.42  78.9  4.97 2       242    17.8  397.  9.14  21.6
##  3 0.0273    0    7.07 0     0.469  7.18  61.1  4.97 2       242    17.8  393.  4.03  34.7
##  4 0.0324    0    2.18 0     0.458  7.00  45.8  6.06 3       222    18.7  395.  2.94  33.4
##  5 0.0690    0    2.18 0     0.458  7.15  54.2  6.06 3       222    18.7  397.  5.33  36.2
##  6 0.0298    0    2.18 0     0.458  6.43  58.7  6.06 3       222    18.7  394.  5.21  28.7
##  7 0.0883   12.5  7.87 0     0.524  6.01  66.6  5.56 5       311    15.2  396. 12.4   22.9
##  8 0.145    12.5  7.87 0     0.524  6.17  96.1  5.95 5       311    15.2  397. 19.2   27.1
##  9 0.211    12.5  7.87 0     0.524  5.63 100    6.08 5       311    15.2  387. 29.9   16.5
## 10 0.170    12.5  7.87 0     0.524  6.00  85.9  6.59 5       311    15.2  387. 17.1   18.9
## # ... with 496 more rows
```

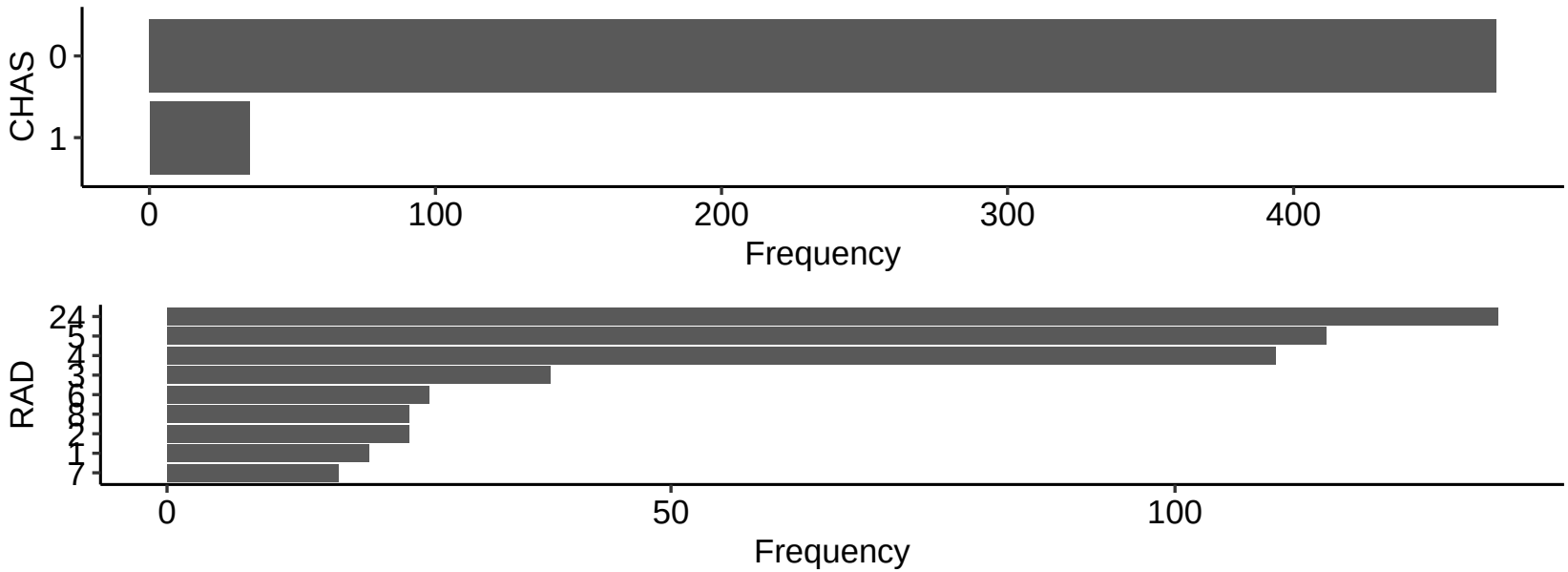
BOSTON HOUSING DATA SET

Exploratory Data Analysis

Factor variables

variable	missing	n_unique	top_counts
CHAS	0	2	0: 471, 1: 35, NA: 0
RAD	0	9	24: 132, 5: 115, 4: 110, 3: 38

Barplots of discrete features



BOSTON HOUSING DATA SET

Exploratory Data Analysis

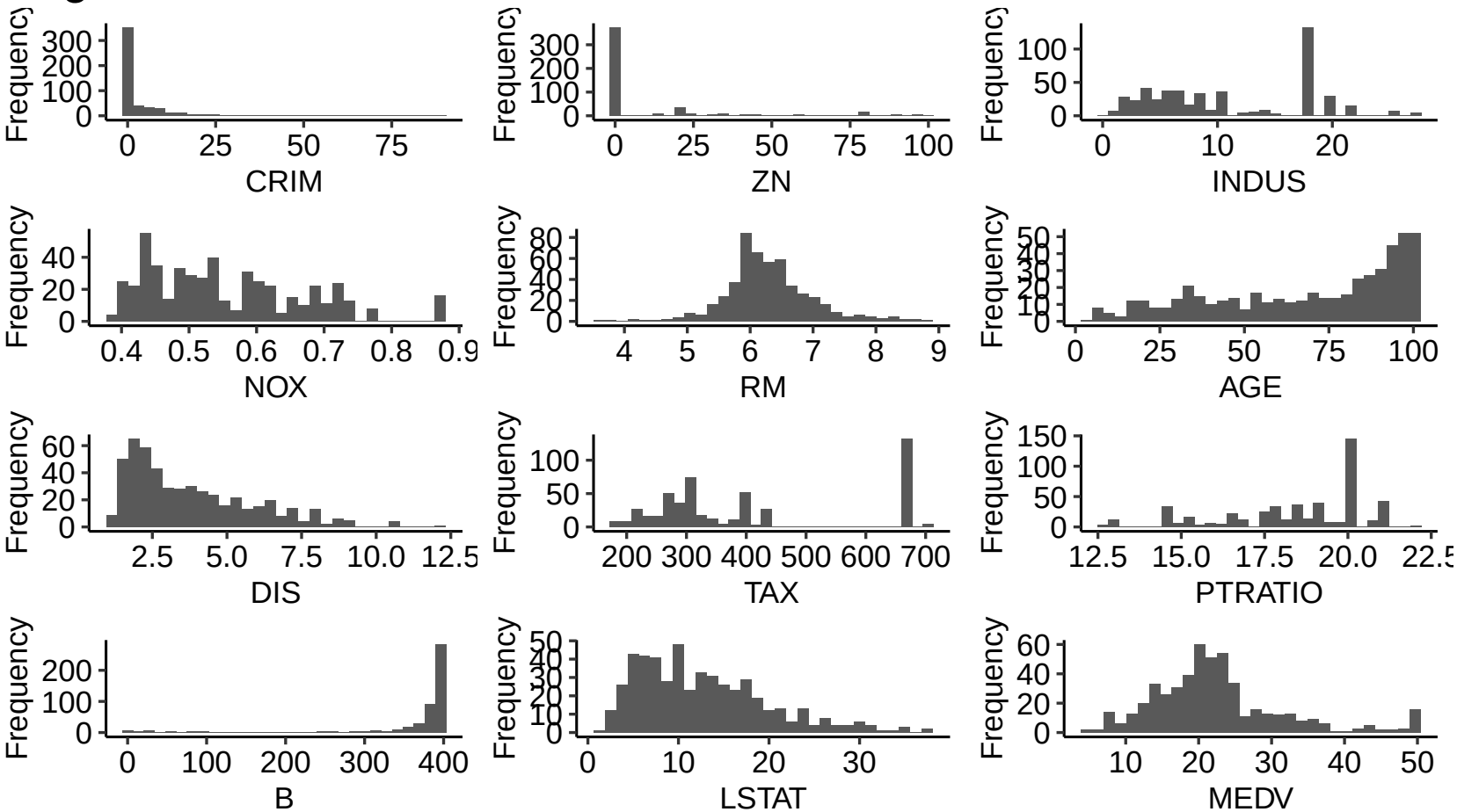
Numeric variables

variable	missing	mean	sd
AGE	0	68.57	28.15
B	0	356.67	91.29
CRIM	0	3.61	8.6
DIS	0	3.8	2.11
INDUS	0	11.14	6.86
LSTAT	0	12.65	7.14
MEDV	0	22.53	9.2
NOX	0	0.55	0.12
PTRATIO	0	18.46	2.16
RM	0	6.28	0.7
TAX	0	408.24	168.54
ZN	0	11.36	23.32

BOSTON HOUSING DATA SET

Exploratory Data Analysis

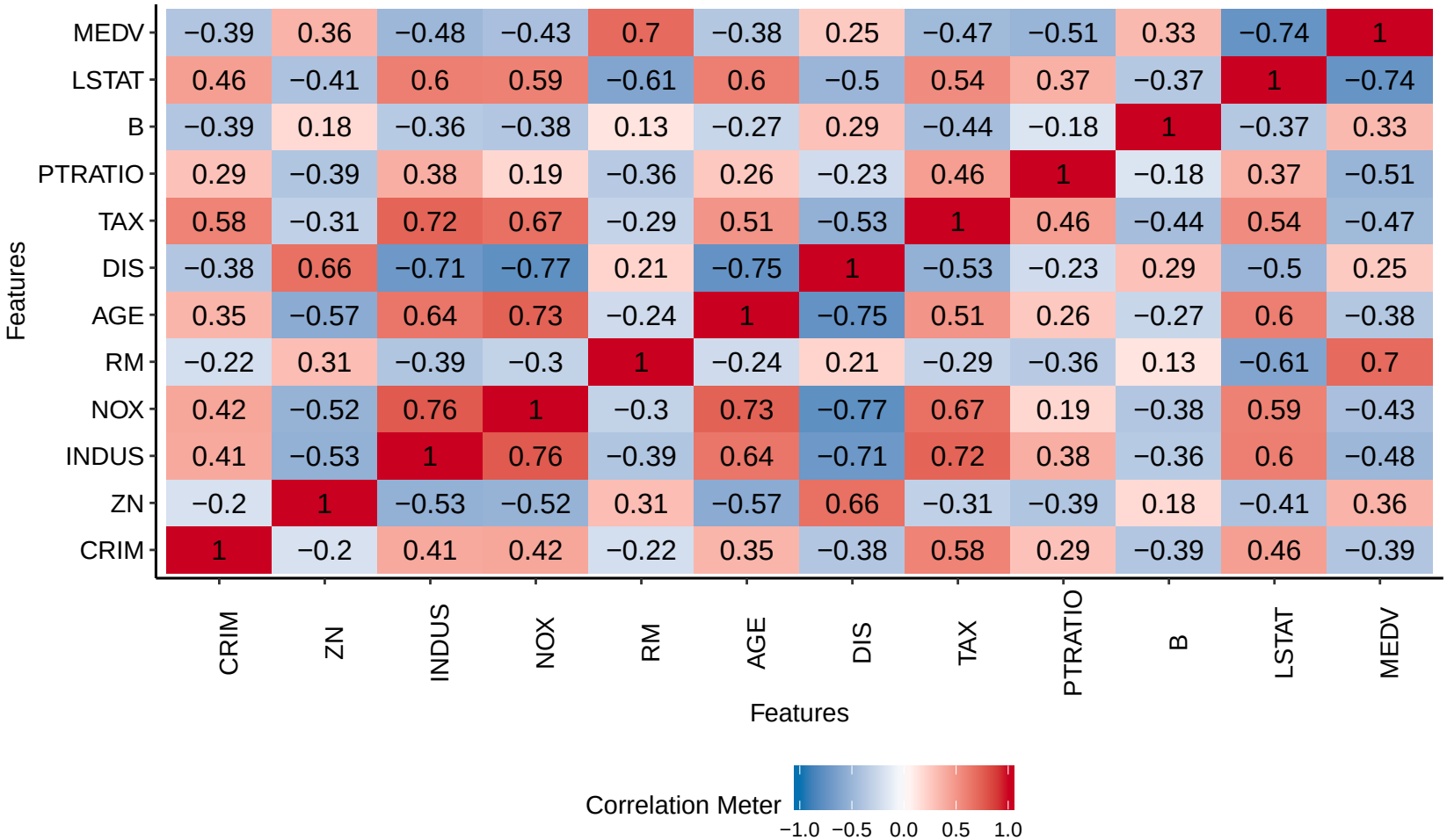
Histograms of numerical features



BOSTON HOUSING DATA SET

Exploratory Data Analysis

It is always useful to check the correlation among variables:



IRIS DATA SET

The iris dataset was introduced by the statistician Ronald Fisher and is one of the most frequent used datasets. Originally it was designed for linear discriminant analysis.
The set is a typical test case for many statistical classification techniques and has its own **wikipedia page**.

Setosa

Versicolor

Virginica

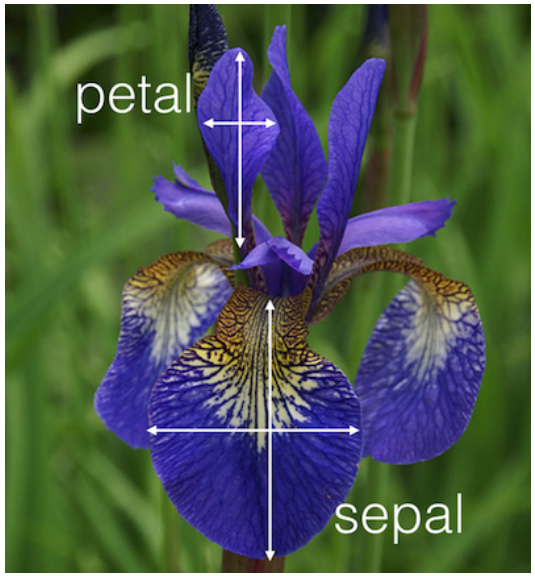
Source:
`https://en.wikipedia.org/wiki/Iris_flower_data_set`

IRIS DATA SET

Importing the Data

We use OpenML (R-Package) to download the dataset in a machine-readable format and convert it into a 'data.frame':

- 150 iris flowers (50 setosa, 50 versicolor, 50 virginica), species should be predicted.
- Sepal length / width and petal length / width in [cm].



Source: `https://holgerbrandl.github.io/kotlin4ds_kotlin_night_frankfurt//krangl_example_report.html`

IRIS DATA SET

```
## # A tibble: 150 x 5
##    sepallength sepalwidth petallength petalwidth class
##  *       <dbl>      <dbl>       <dbl>      <dbl> <fct>
##  1         5.1        3.5         1.4        0.2 Iris-setosa
##  2         4.9        3           1.4        0.2 Iris-setosa
##  3         4.7        3.2         1.3        0.2 Iris-setosa
##  4         4.6        3.1         1.5        0.2 Iris-setosa
##  5         5          3.6         1.4        0.2 Iris-setosa
##  6         5.4        3.9         1.7        0.4 Iris-setosa
##  7         4.6        3.4         1.4        0.3 Iris-setosa
##  8         5          3.4         1.5        0.2 Iris-setosa
##  9         4.4        2.9         1.4        0.2 Iris-setosa
## 10         4.9        3.1         1.5        0.1 Iris-setosa
## # ... with 140 more rows
```

IRIS DATA SET

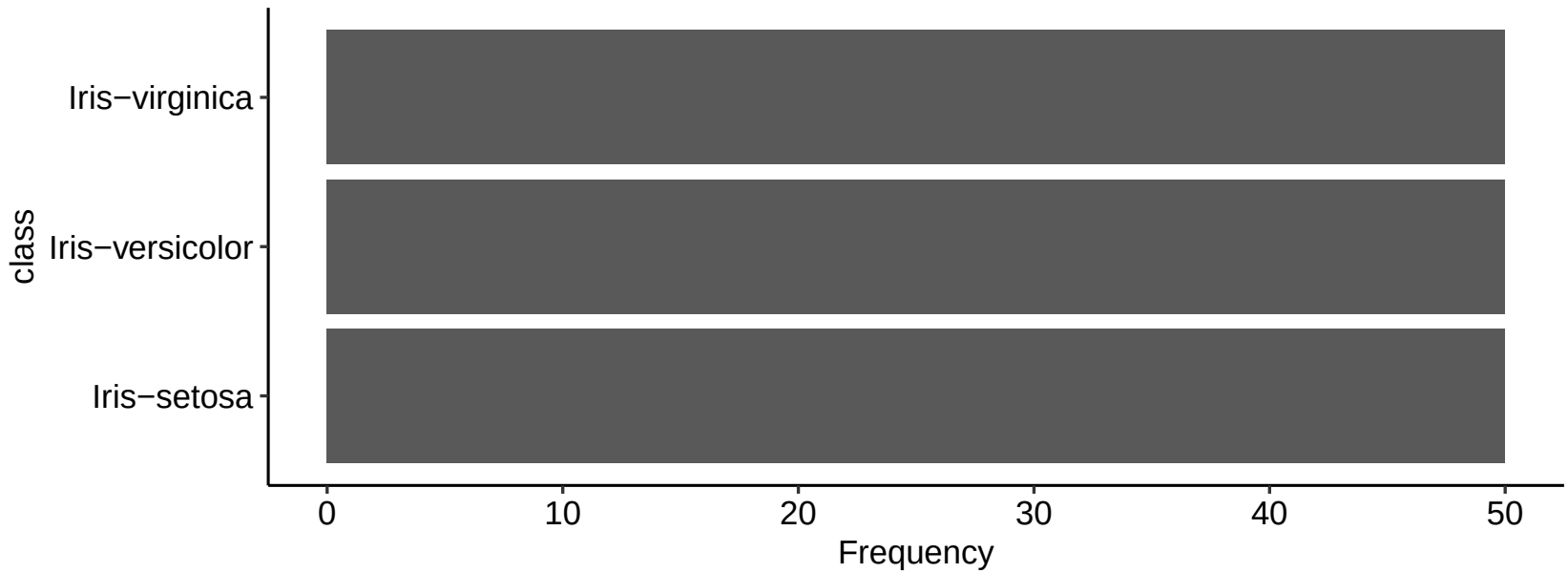
Exploratory Data Analysis

Factor variables

variable	missing	n_unique	top_counts
class	0	3	Iri: 50, Iri: 50, Iri: 50, NA: 0

Exploratory Data Analysis

Barplots of discrete features



IRIS DATA SET

Exploratory Data Analysis

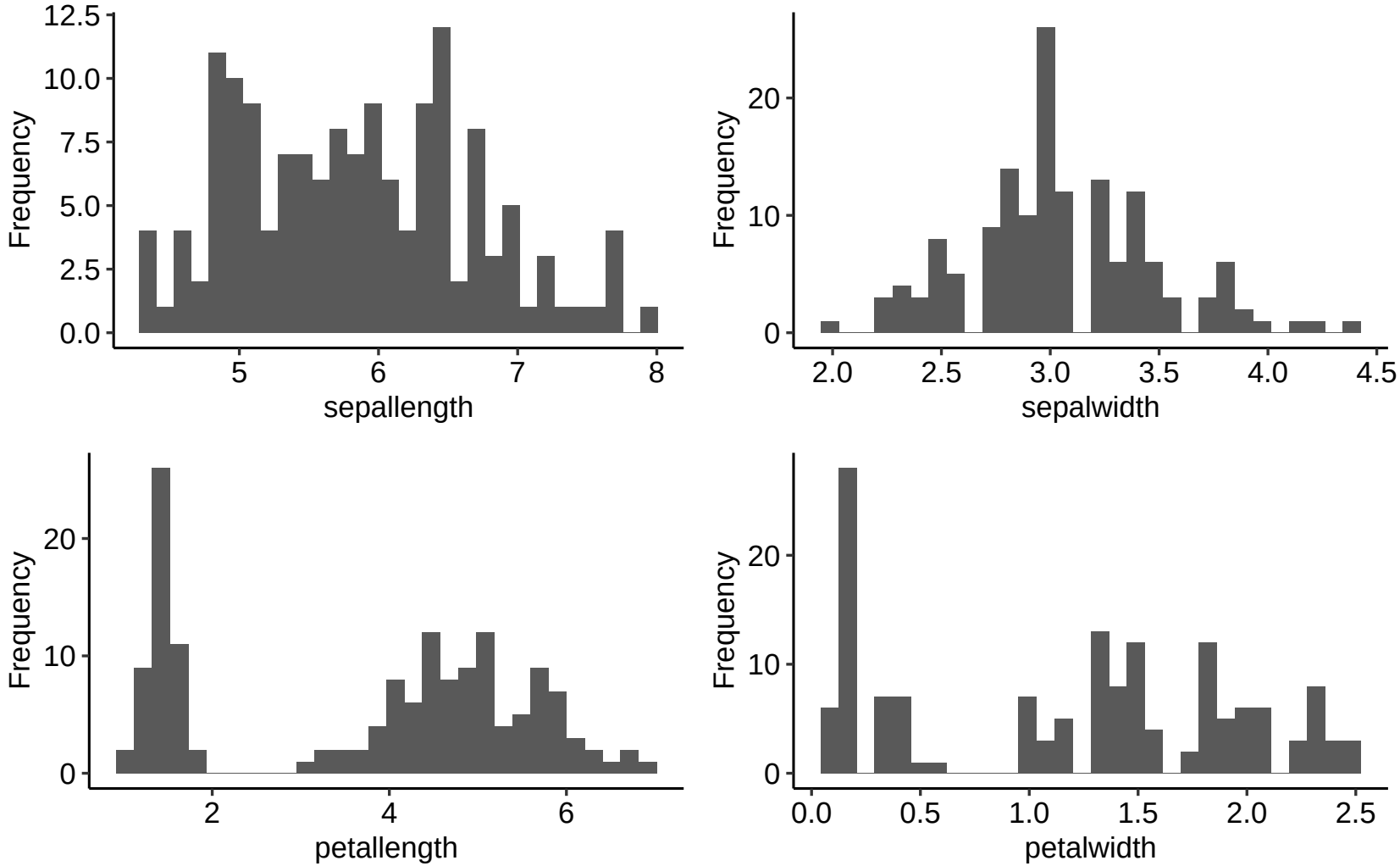
Numeric variables

variable	missing	mean	sd
petallength	0	3.76	1.76
petalwidth	0	1.2	0.76
sepallength	0	5.84	0.83
sepalwidth	0	3.05	0.43

IRIS DATA SET

Exploratory Data Analysis

Histograms of numerical features



SPAM DATA SET

A data set collected at Hewlett-Packard Labs, that classifies 4601 **e-mails as spam or non-spam** (variable 'class'). The spam dataset is one of the datasets used in **The Elements of Statistical Learning** by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Besides the option to import it from **OpenML** it also comes as an example dataset in the packages **ElemStatLearn** and **kernlab**.

SPAM DATA SET

- 'class' : 0 = no spam, 1 = spam
- 'word_freq_*': 48 features corresponding to the relative frequency of a specific word in an e-mail
- 'char_freq_*': 6 features that measures the percentage of a sequence of specific characters occurs relative to the total number of characters
- 'capital_run_length_[average, longest, total]': 3 features indicating the average, longest, and sum of uninterrupted sequence of capital letters

SPAM DATA SET

Importing the Data

We use Open ML (R-Package) to download the dataset in a machine-readable format and convert it into a 'data.frame':

- 58 features (e. g. specific word frequencies)
- 4,601 observations
- too much to display as a table

SPAM DATA SET

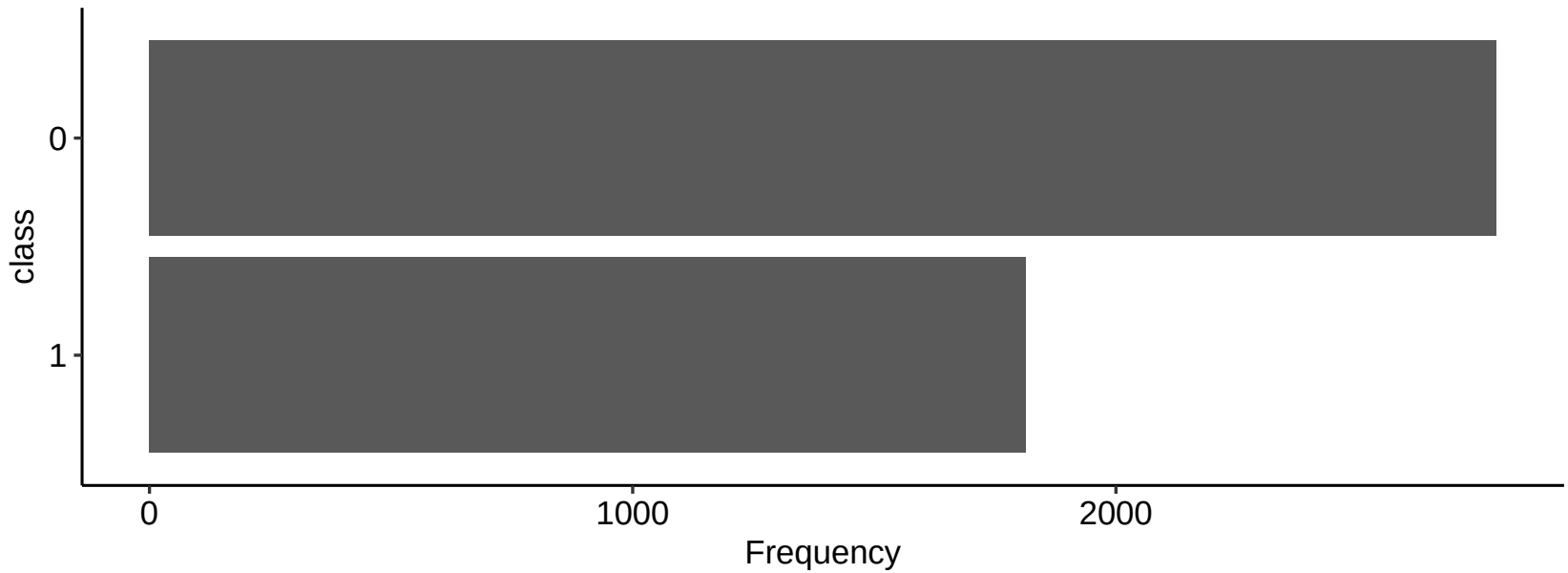
Exploratory Data Analysis

Factor variables

variable	missing	n_unique	top_counts
class	0	2	0: 2788, 1: 1813, NA: 0

Exploratory Data Analysis

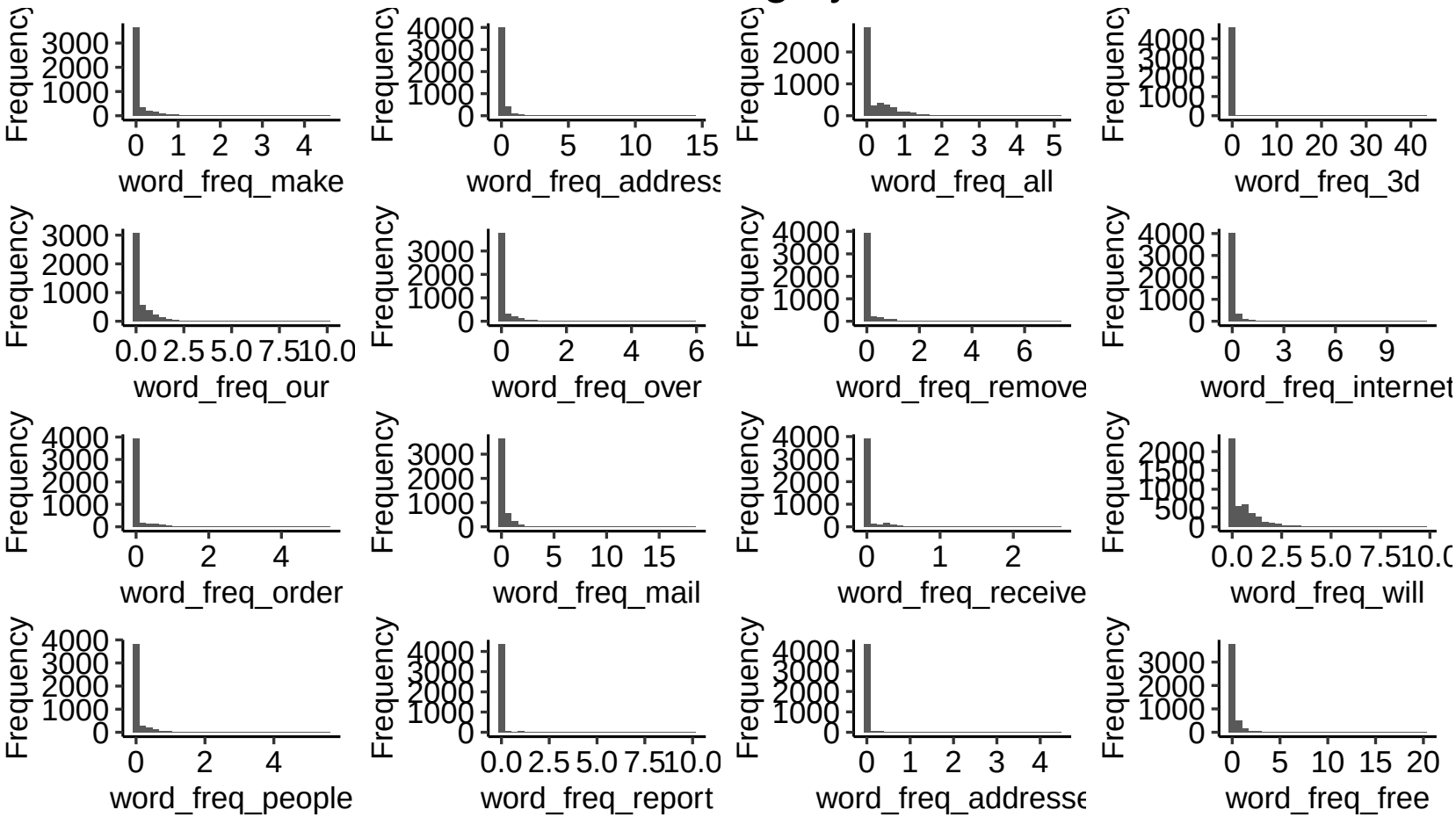
Barplots of discrete features



SPAM DATA SET

Exploratory Data Analysis

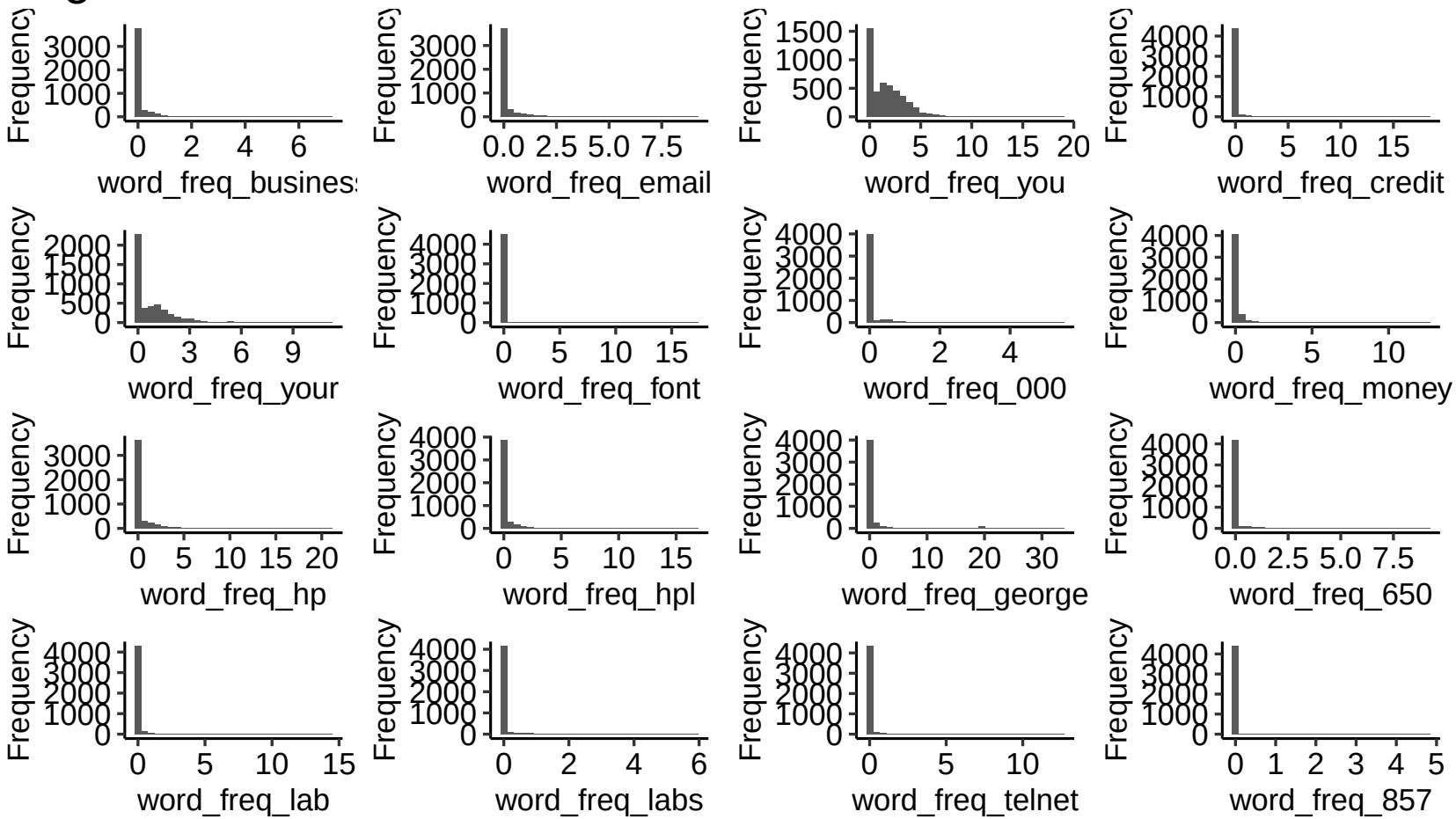
The distribution of most variables is highly skewed:



SPAM DATA SET

Exploratory Data Analysis

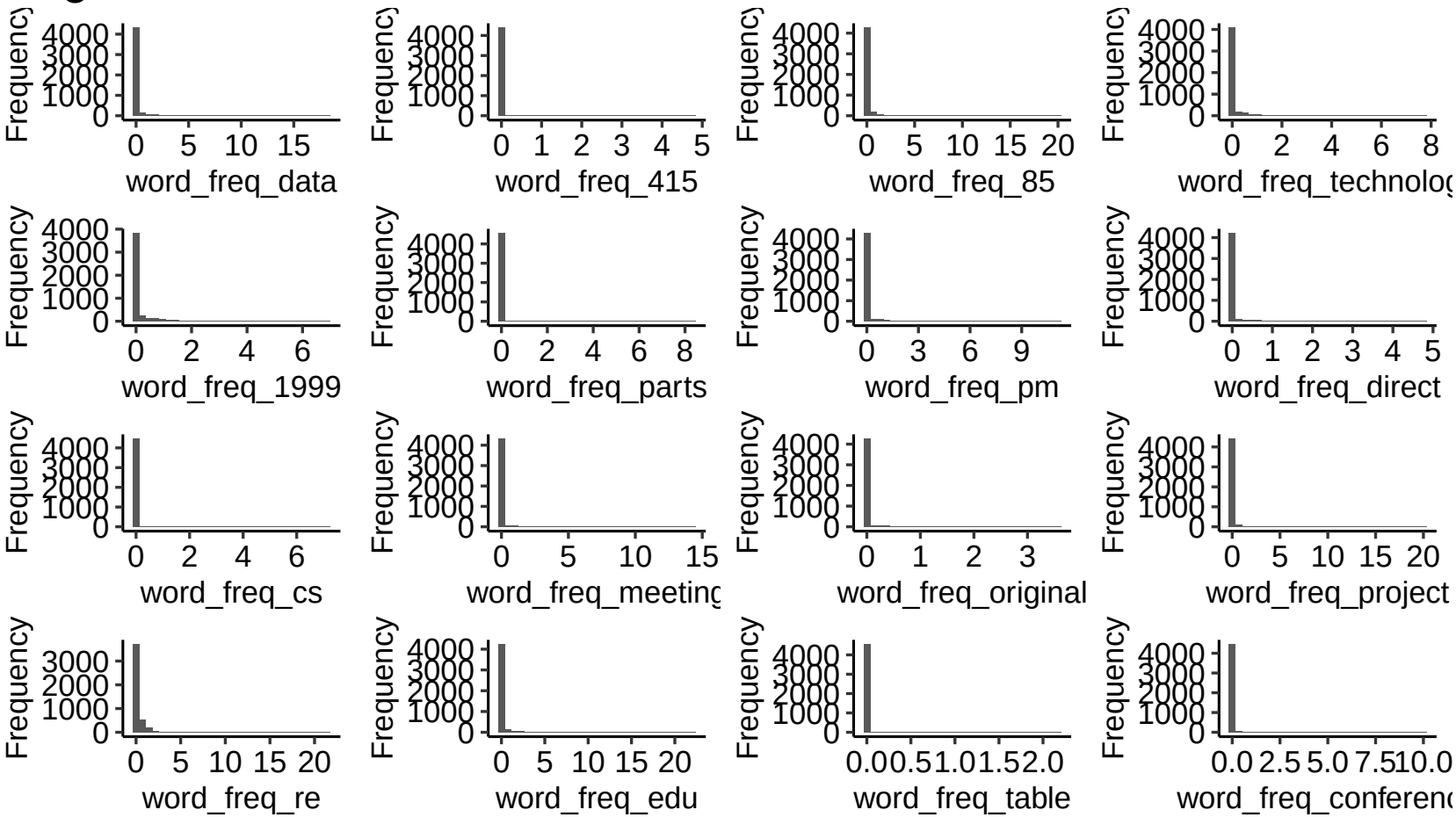
Histograms of numerical features



SPAM DATA SET

Exploratory Data Analysis

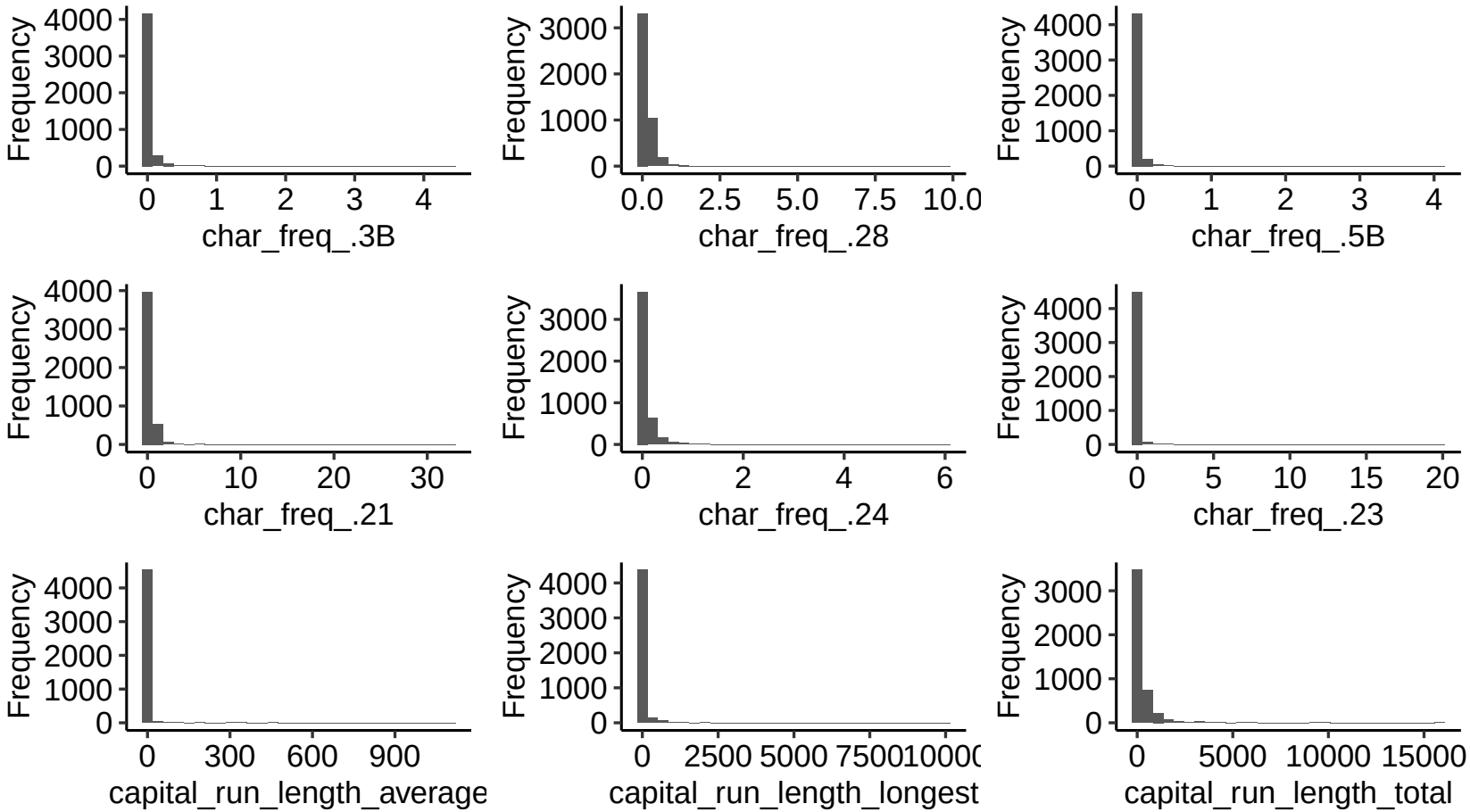
Histograms of numerical features



SPAM DATA SET

Exploratory Data Analysis

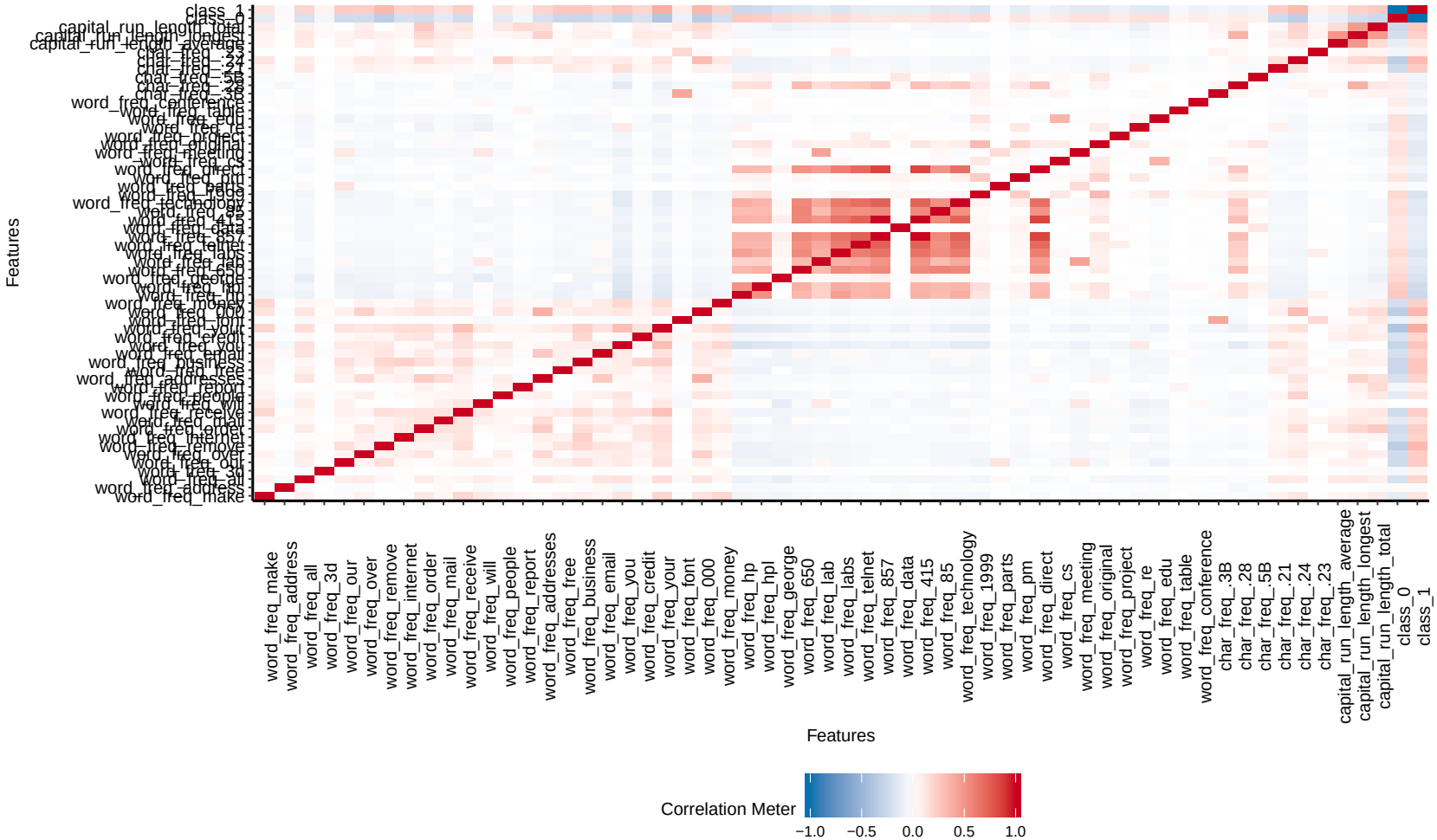
Histograms of numerical features



SPAM DATA SET

Exploratory Data Analysis

Let's take a look at the correlation among the variables:



TITANIC DATA SET

The original Titanic dataset, describing the **survival status of individual passengers**(1309) on the Titanic. The titanic data does not contain information from the crew, but it does contain actual ages of half of the passengers. The principal source for data about Titanic passengers is the Encyclopedia Titanica.
One of the original sources is Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd. It includes a passenger list created by many researchers (edited by Michael A. Findlay).

TITANIC DATA SET

Variable	Description
survived	0 = No, 1 = Yes
pclass	1 = 1st; 2 = 2nd; 3 = 3rd
name	First and last Name
sex	Sex
Age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
Ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation C = Cherbourg; Q = Queenstown; S = Southampton
body	Body Identification Number
boat	Boat number
home.dest	Home destination

TITANIC DATA SET

Importing the Data

We use Open ML (R-Package) to download the dataset in a machine-readable format and convert it into a 'data.frame':

```
## # A tibble: 1,309 x 14
##    pclass survived name  sex      age sibsp parch ticket  fare cabin embarked boat   body
##  *  <dbl> <fct>    <chr> <fct>  <dbl> <dbl> <dbl> <chr>  <dbl> <chr> <fct>    <chr> <dbl>
##  1      1 1        Alle~ fema~ 29         0     0 24160  211.  B5    S        2        NA
##  2      1 1        Alli~ male   0.917     1     2 113781 152.  C22 ~ S        11       NA
##  3      1 0        Alli~ fema~  2         1     2 113781 152.  C22 ~ S        <NA>     NA
##  4      1 0        Alli~ male  30         1     2 113781 152.  C22 ~ S        <NA>    135
##  5      1 0        Alli~ fema~ 25         1     2 113781 152.  C22 ~ S        <NA>     NA
##  6      1 1        Ande~ male  48         0     0 19952   26.6 E12   S        3        NA
##  7      1 1        Andr~ fema~ 63         1     0 13502   78.0 D7    S        10       NA
##  8      1 0        Andr~ male  39         0     0 112050   0   A36   S        <NA>     NA
##  9      1 1        Appl~ fema~ 53         2     0 11769   51.5 C101  S        D        NA
## 10      1 0        Arta~ male  71         0     0 PC 17~  49.5 <NA>  C        <NA>     22
## # ... with 1,299 more rows, and 1 more variable: home.dest <chr>
```

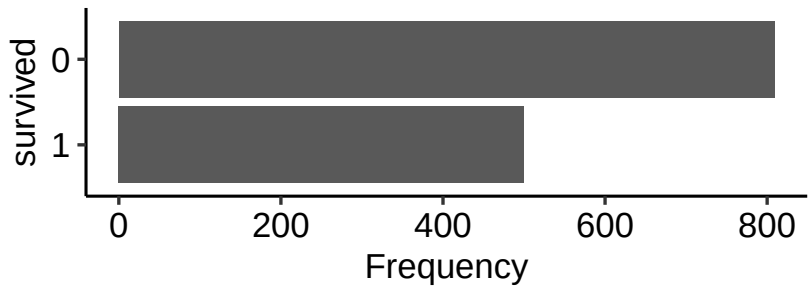
TITANIC DATA SET

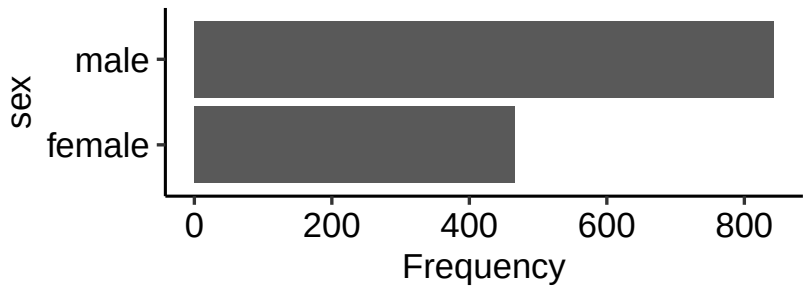
Exploratory Data Analysis

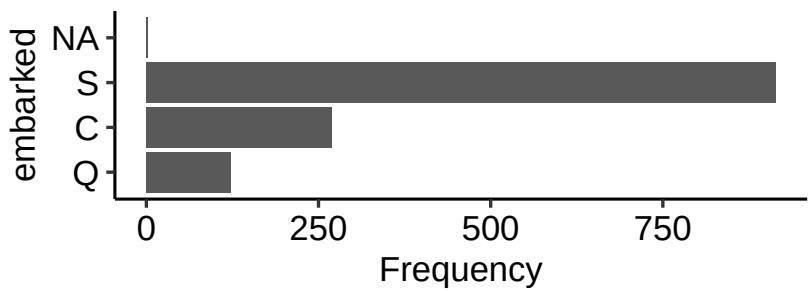
Factor variables

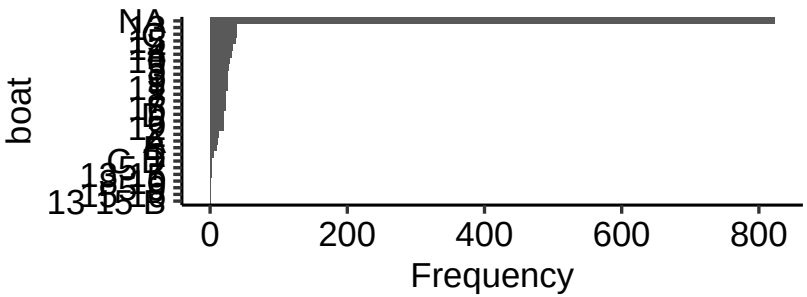
variable	missing	n_unique	top_counts
embarked	2	3	S: 914, C: 270, Q: 123, NA: 2
sex	0	2	mal: 843, fem: 466, NA: 0
survived	0	2	0: 809, 1: 500, NA: 0

Barplots of discrete features









TITANIC DATA SET

Exploratory Data Analysis

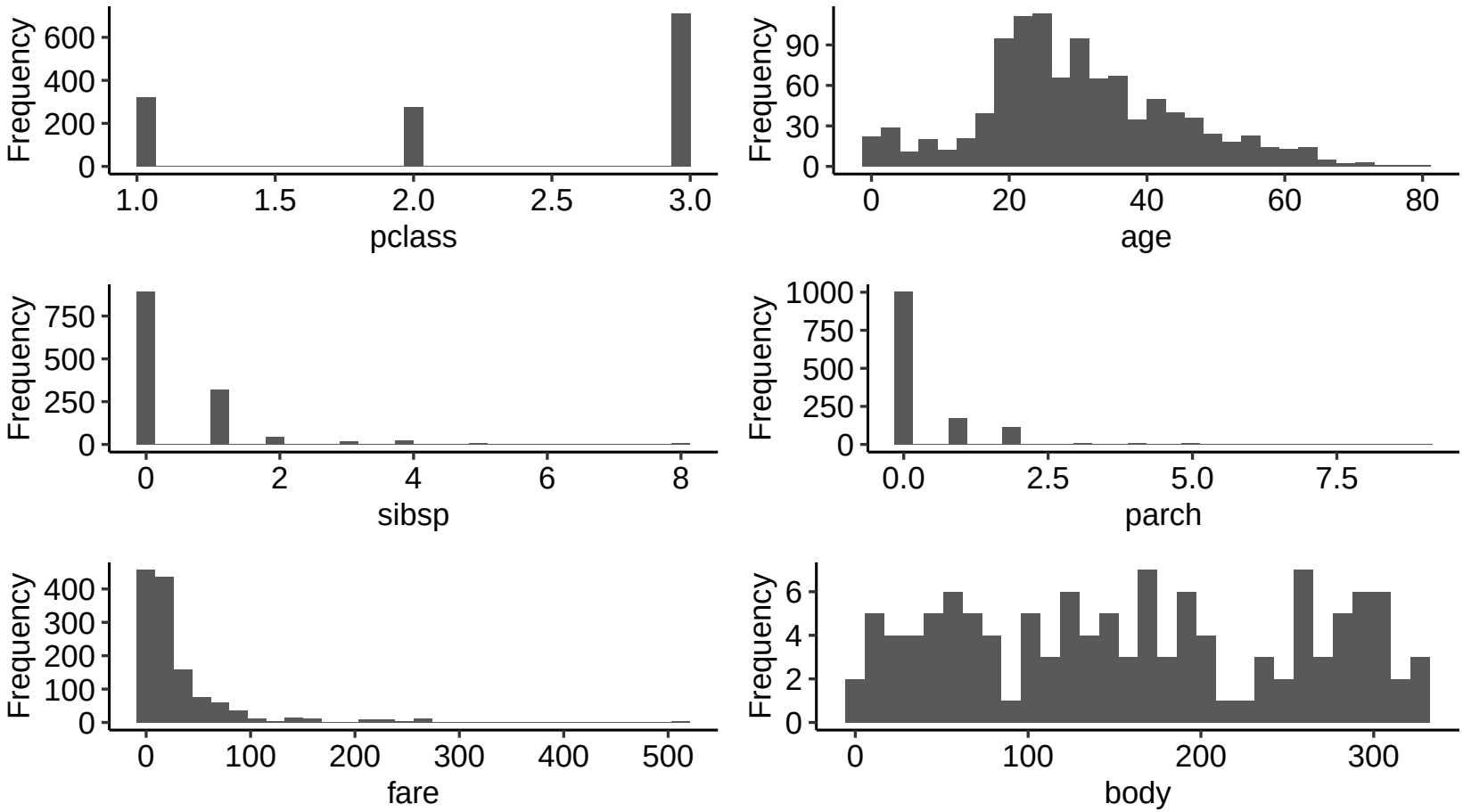
Numeric variables

variable	missing	mean	sd
age	263	29.88	14.41
body	1188	160.81	97.7
fare	1	33.3	51.76
parch	0	0.39	0.87
pclass	0	2.29	0.84
sibsp	0	0.5	1.04

TITANIC DATA SET

Exploratory Data Analysis

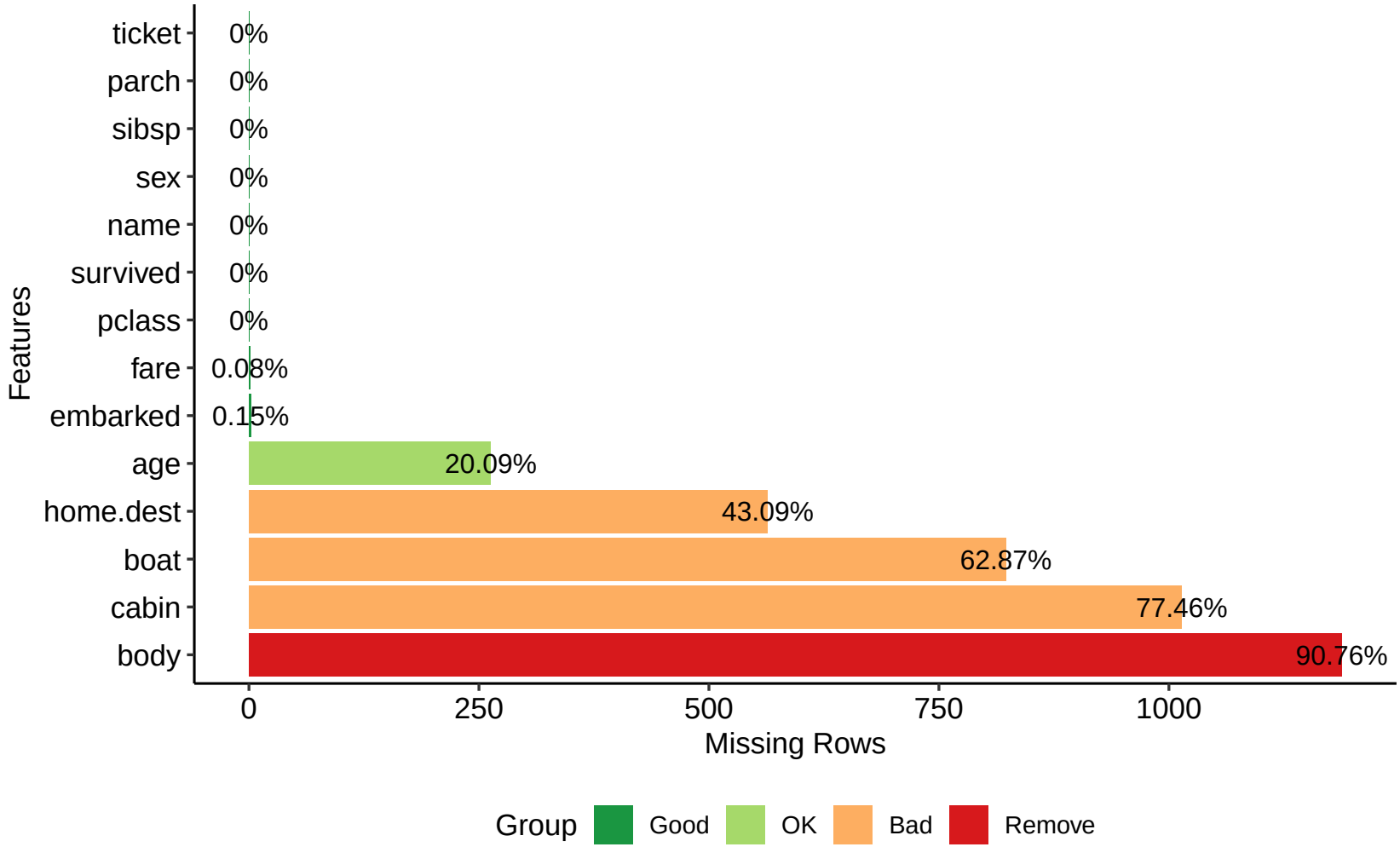
Histograms of numerical features



TITANIC DATA SET

Exploratory Data Analysis

Seems we have quite some missing observations. Let's take a closer look:



TITANIC DATA SET

Exploratory Data Analysis

It is always useful to check the correlation among variables: