Titanic data set in Google sheet

Class #3, Feb 28 2023

Barbora Hladká hladka@ufal.mff.cuni.cz

Titanic in Kaggle

- Use machine learning to automatically predict which passengers survived the Titanic shipwreck, i.e. predict a target value of Survived
- Data in supervised machine learning
  - training set
  - test set

Titanic data set

- https://www.kaggle.com/c/titanic
- training data `train.csv`
- development test data `test.csv`
- `csv` = Comma Separated Values format

```
PassengerId,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 Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",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
```

Data analysis of `train.csv`

- We uploaded `train.csv` to Google drive, open this link
- It is useful to know the story of Titanic when analyzing the data (Wikipedia)

Table

is a way how to organize data using rows and columns

	1	…		*m*
1				
…				
n				

m columns = *m* attributes, *n*(+1) rows = *n* examples

Attributes

= are properties of objects that we can observe or measure. Their values can be of several types

- numerical
  - either discrete or continuous
- categorical
  - any list of discrete values, non-numerical
- binary (0/1, Yes/No)
  - can be viewed as a kind of categorical

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q

$m = 12$

1. PassengerId a passenger's unique identifier
2. Survived 0/1 binary
3. Pclass categorical
4. Name categorical
5. Sex binary
6. Age numerical
7. SibSp number of siblings/spouses aboard numerical discrete
8. Parch number of parents/children aboard numerical discrete
9. Ticket ticket number categorical
10. Fare passenger fare (British pound) numerical
11. Cabin cabin number categorical
12. Embarked port of embarkation categorical

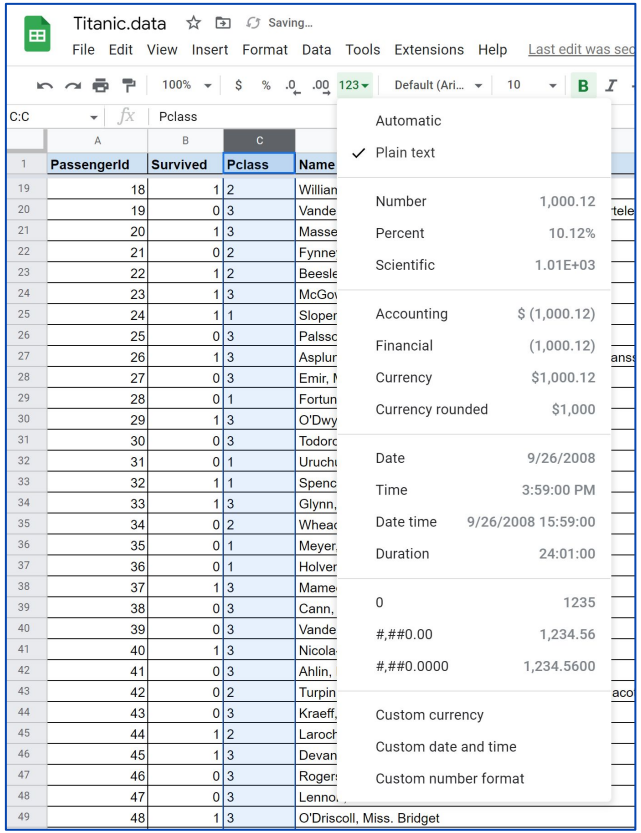
Passengers

= rows in the table, 891 passengers in `train.csv`, i.e. *n* = 891

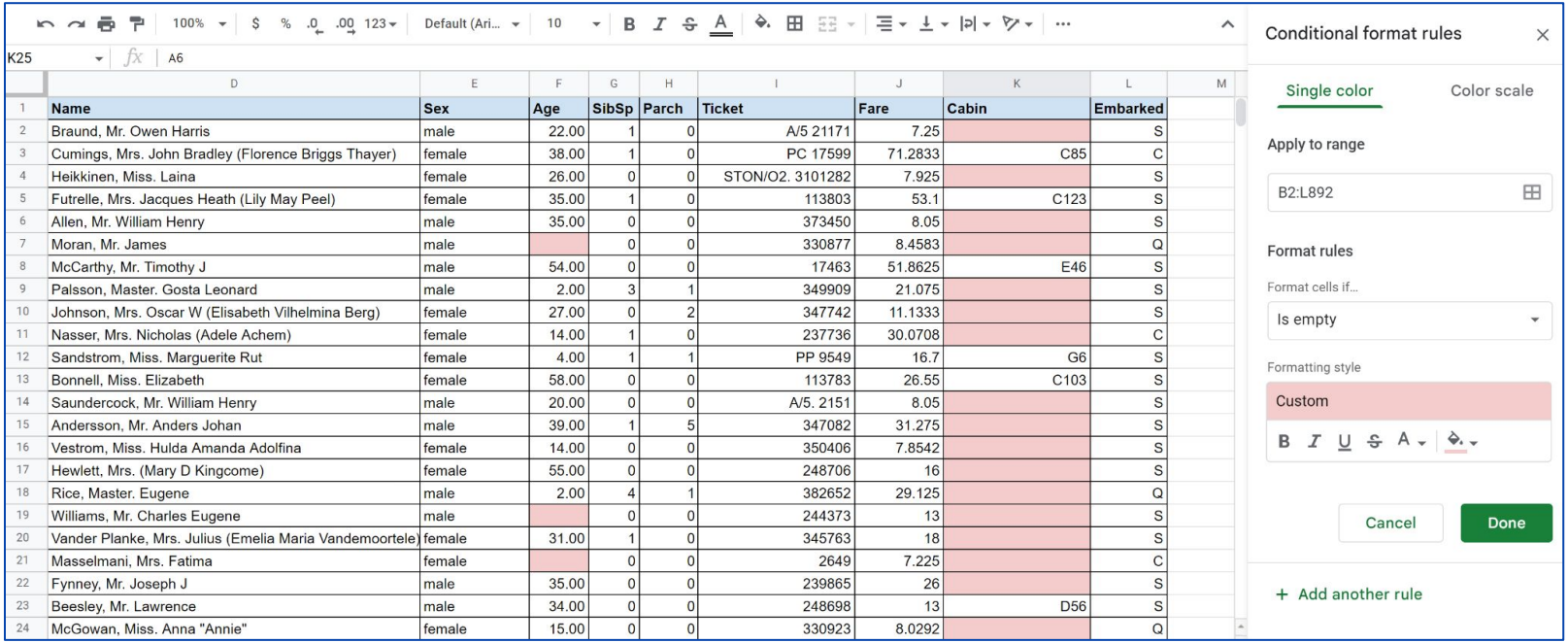
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q

Formatting of attribute values

- Check the attribute values and edit their formats if needed



Highlight missing values `Format > Conditional formatting`



	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.25		S
3	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
4	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.925		S
5	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1	C123	S
6	Allen, Mr. William Henry	male	35.00	0	0	373450	8.05		S
7	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
9	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.075		S
10	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
11	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C
12	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7	G6	S
13	Bonnell, Miss. Elizabeth	female	58.00	0	0	113783	26.55	C103	S
14	Saundercock, Mr. William Henry	male	20.00	0	0	A/5. 2151	8.05		S
15	Andersson, Mr. Anders Johan	male	39.00	1	5	347082	31.275		S
16	Vestrom, Miss. Hulda Amanda Adolfina	female	14.00	0	0	350406	7.8542		S
17	Hewlett, Mrs. (Mary D Kingcome)	female	55.00	0	0	248706	16		S
18	Rice, Master. Eugene	male	2.00	4	1	382652	29.125		Q
19	Williams, Mr. Charles Eugene	male		0	0	244373	13		S
20	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.00	1	0	345763	18		S
21	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
22	Fynney, Mr. Joseph J	male	35.00	0	0	239865	26		S
23	Beesley, Mr. Lawrence	male	34.00	0	0	248698	13	D56	S
24	McGowan, Miss. Anna "Annie"	female	15.00	0	0	330923	8.0292		Q

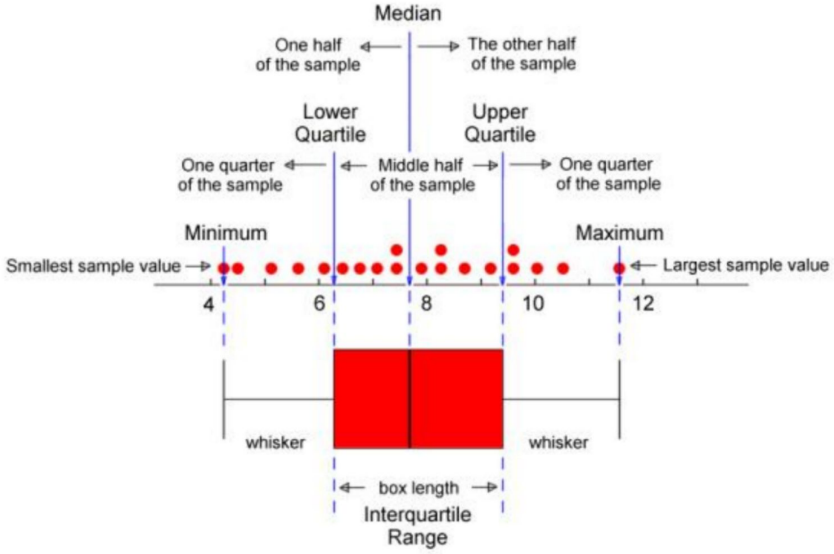
missing.values.age

Replace missing values of Age

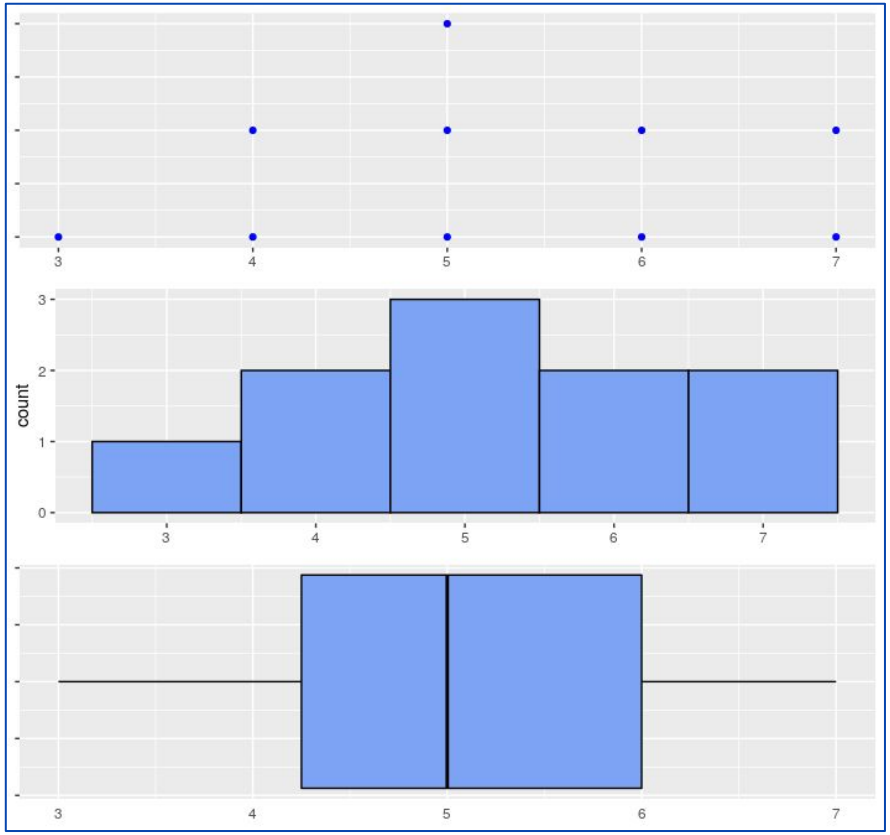
- Think about missing values carefully
- Focus on Age `range, criterion`
  - # of missing values `=COUNTIF(B5:B895,"")`
  - replace its missing values with e.g. the median value

Five number summary and box plot

- min
- 1st quartile
- median
- 3rd quartile
- max



Box plot and histogram



missing.values.age

Replace missing values of Age

- Think about missing values carefully
- Focus on Age `range, criterion`
  - # of missing values `=COUNTIF(B5:B895,"")`
  - replace its missing values with e.g. the median value
- Age - see the cell `missing.values.age!B1` `=MEDIAN(B3:B894)`

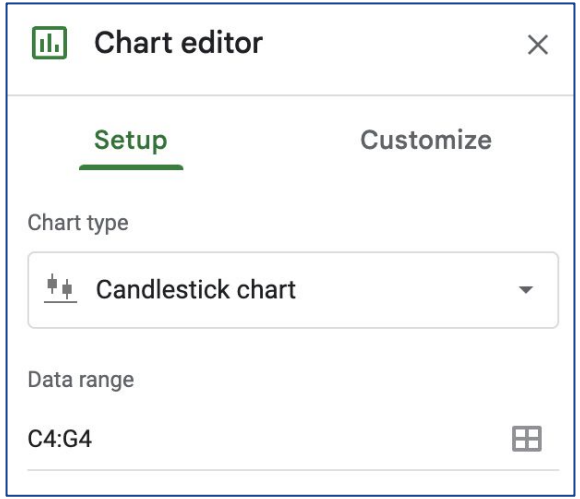
Replace missing values of Age

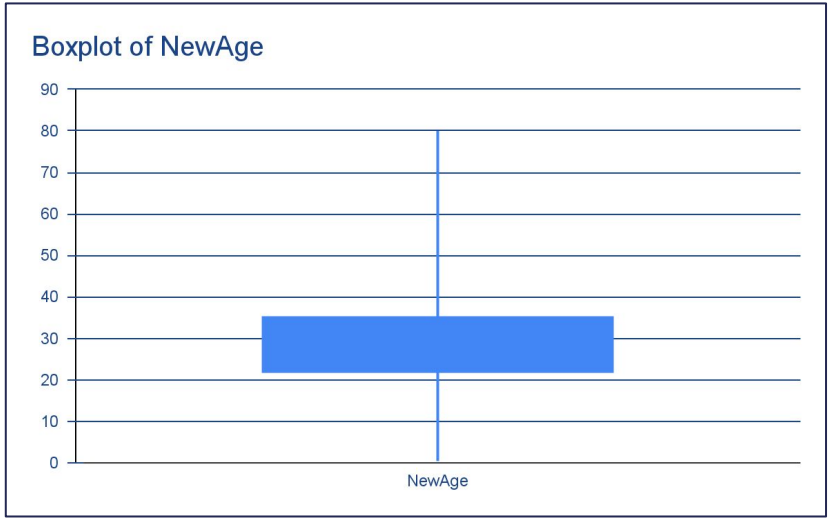
- Create a new attribute NewAge
  - replace Age missing values with the median value
  - keep others same
  - `=IF(B4="",$B$1,B4)`

if `criterion` then `Action1`
else `Action2`

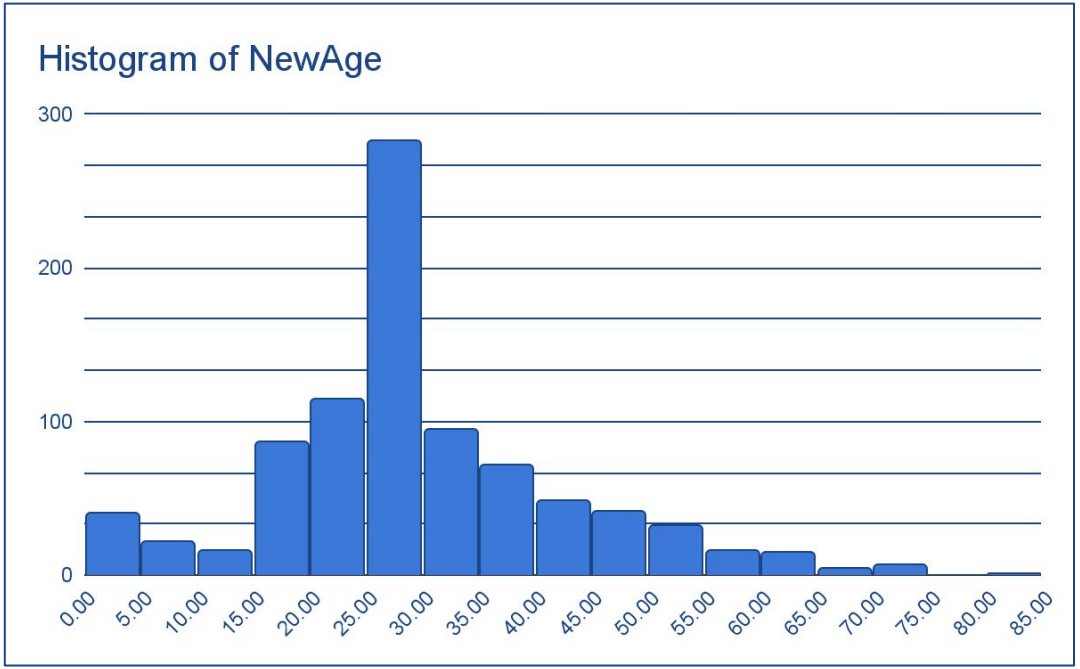
Boxplot of NewAge

- Insert > Chart





Histogram of NewAge



train.csv

Extract titles

- Name
- Copy Name column from `train.csv` to `extract.titles`
- Do `Data > Split text to columns` twice

Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkinson)
Ostby, Mr. Engelhart Cornelius
Woolner, Mr. Hugh
Rugg, Miss. Emily
Novel, Mr. Mansouer
West, Miss. Constance Mirium
Goodwin, Master. William Frederick
Sirayanian, Mr. Orsen
Icard, Miss. Amelie
Harris, Mr. Henry Birkhardt
Skoog, Master. Harald
Stewart, Mr. Albert A

Pivot table a.k.a. contingency table

shows the relationship between categorical attributes.

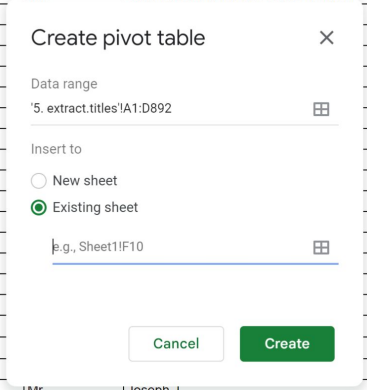
		Embarked			
		C	Q	S	Sum
Sex	female	75	36	203	314
	male	95	41	441	577
				Total sum	891

Pivot table :: Titles

- Count the number of occurrences of each title

Insert > Pivot table > Insert to Existing sheet

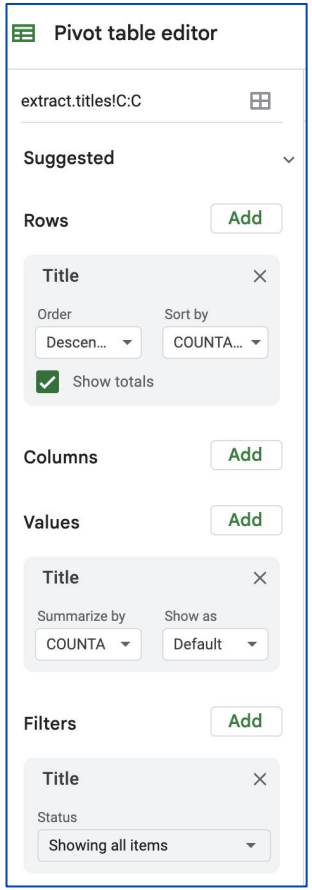
	A	B	C	D
1	Name		Title	
2	Braund, Mr. Owen Harris	Braund	Mr	Owen Harris
3	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	Cumings	Mrs	John Bradley (Florence Briggs Thayer)
4	Heikkinen, Miss. Laina	Heikkinen		
5	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Futrelle		
6	Allen, Mr. William Henry	Allen		
7	Moran, Mr. James	Moran		
8	McCarthy, Mr. Timothy J	McCarthy		
9	Palsson, Master. Gosta Leonard	Palsson		
10	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	Johnson		
11	Nasser, Mrs. Nicholas (Adele Achem)	Nasser		
12	Sandstrom, Miss. Marguerite Rut	Sandstrom		
13	Bonnell, Miss. Elizabeth	Bonnell		
14	Saundercock, Mr. William Henry	Saundercock		
15	Andersson, Mr. Anders Johan	Andersson		
16	Vestrom, Miss. Hulda Amanda Adolfina	Vestrom		
17	Hewlett, Mrs. (Mary D Kingcome)	Hewlett		
18	Rice, Master. Eugene	Rice		
19	Williams, Mr. Charles Eugene	Williams		
20	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	Vander Planke		
21	Masselmani, Mrs. Fatima	Masselmani		
22	Fynney, Mr. Joseph J	Fynney	Mr	Joseph J
23	Beesley, Mr. Lawrence	Beesley	Mr	Lawrence



`extract.title`

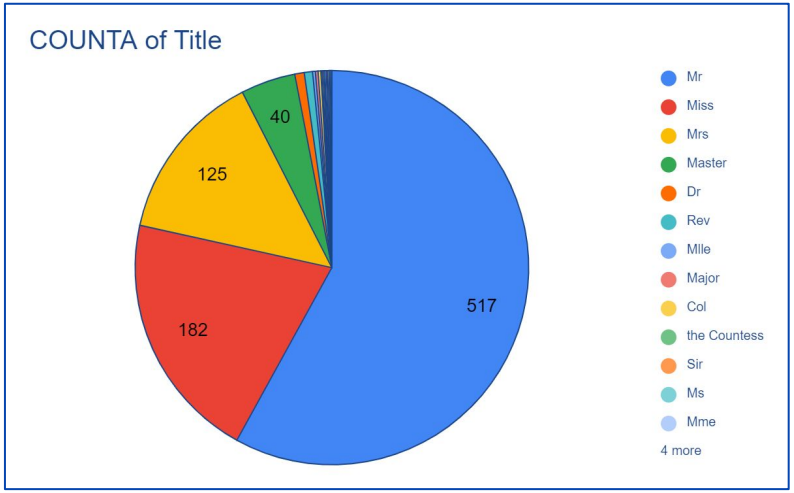
Pivot table :: Titles

- Count the number of occurrences of each title
  `Pivot table editor`
- `COUNTA` counts all categorical values in a dataset



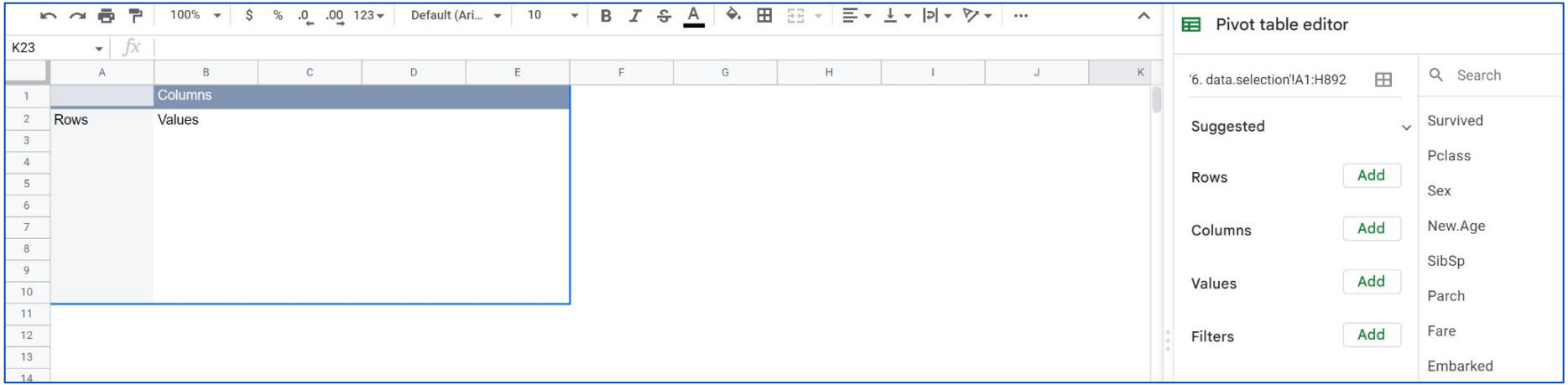
Pivot table :: Titles

- Highlight the pivot table and `Insert > Chart`



Pivot tables :: Sex and Embarked

- data.selection
- Insert > Pivot table > new sheet
- Rename new sheet > sex.embarked



COUNTA of Pcla	Embarked			
Sex	C	Q	S	Grand Total
female	75	36	203	314
male	95	41	441	577
Grand Total	**170**	**77**	**644**	**891**

Edit

Rows Add

Sex

Order: Ascen... Sort by: Sex

Show totals

Columns Add

Embarked

Order: Ascen... Sort by: Embar...

Show totals

Values Add

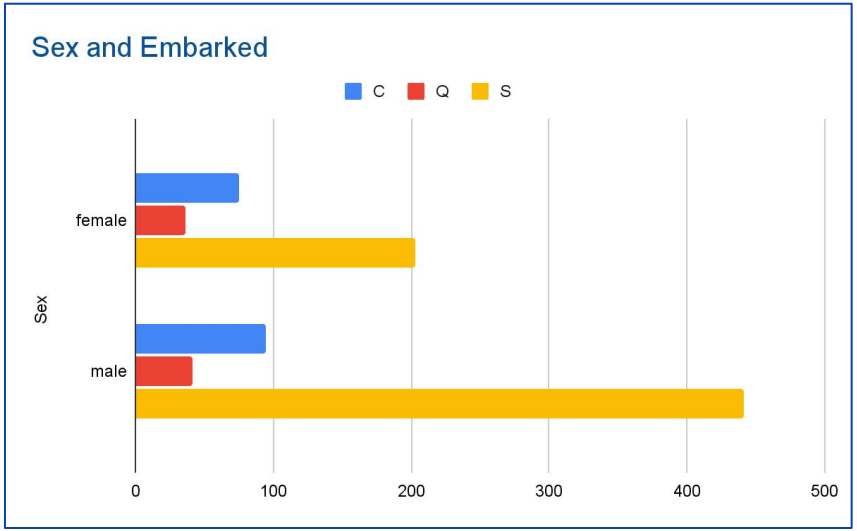
Pclass

Summarize by: COUN... Show as: Default

sex.embarked

Pivot tables :: Sex and Embarked

- Highlight the pivot table and `Insert > Chart`



Pivot table :: Sex and Class and Survived

- `data.selection`
- `Insert > Pivot table >` new sheet
- Rename new sheet `> sex.class.survived`
- Add rows, columns, values
- Add survival rates
  - see the cells `B9:J9`
  - use `round(B5/B6,2)`

Pivot table :: Hometown

- 1st class passengers and their hometowns - data from Wikipedia https://en.wikipedia.org/wiki/Passengers_of_the_Titanic
- `data.wikipedia`
- `Insert > Pivot table >` existing sheet
- `Insert > Chart > Geo chart with markers`

remove.duplicates

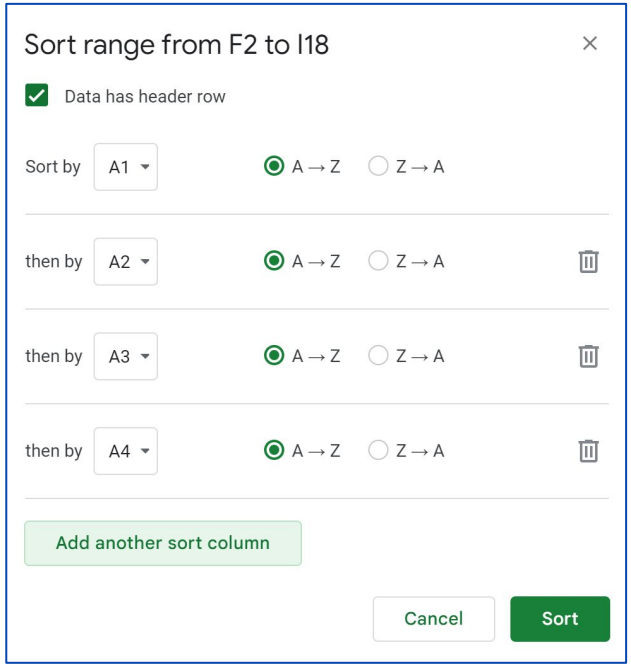
Remove duplicates

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	original data					sorted data					w/o duplicates			
2	A1	A2	A3	A4		A1	A2	A3	A4		A1	A2	A3	A4
3	0	3	F	22.00		0	1	F	54.00		0	1	F	54.00
4	1	1	F	38.00		0	3	F	14.00		0	3	F	14.00
5	0	3	F	39.00		0	3	F	22.00		0	3	F	22.00
6	0	3	M	2.00		0	3	F	22.00		0	3	F	35.00
7	0	3	F	35.00		0	3	F	35.00		0	3	F	39.00
8	0	3	M	28.00		0	3	F	39.00		0	3	M	2.00
9	0	1	F	54.00		0	3	F	39.00		0	3	M	20.00
10	0	3	M	2.00		0	3	M	2.00		0	3	M	28.00
11	1	3	4	27.00		0	3	M	2.00		1	1	F	38.00
12	1	2	F	14.00		0	3	M	20.00		1	1	M	58.00
13	0	3	F	22.00		0	3	M	28.00		1	2	F	14.00
14	1	1	M	58.00		1	1	F	38.00		1	2	F	55.00
15	0	3	M	20.00		1	1	M	58.00		1	3	4	27.00
16	0	3	F	39.00		1	2	F	14.00					
17	0	3	F	14.00		1	2	F	55.00					
18	1	2	F	55.00		1	3	4	27.00					
19														

Remove duplicates

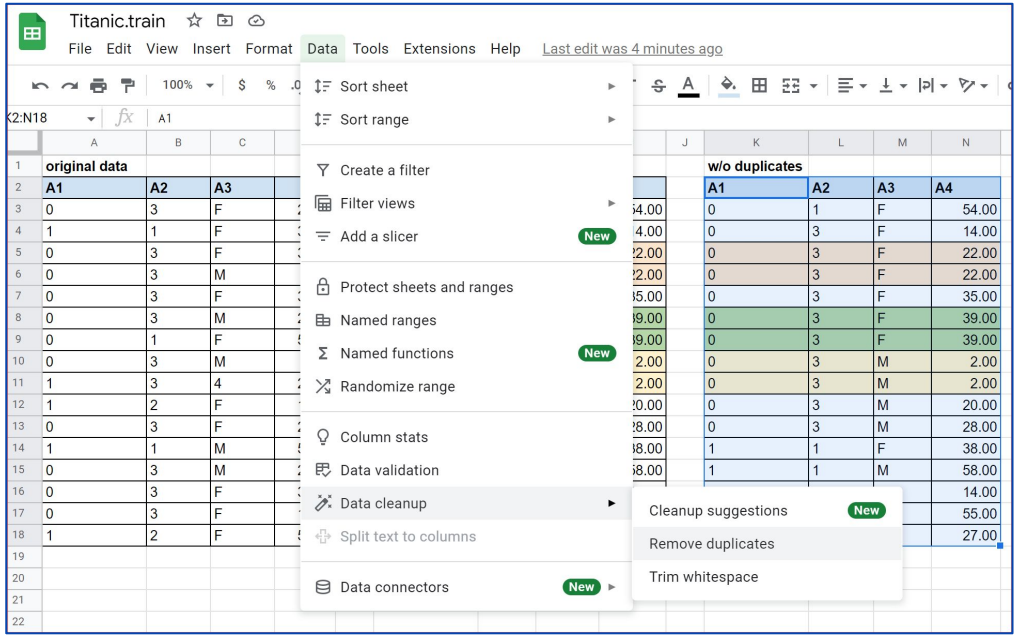
- Copy the original data and sort them

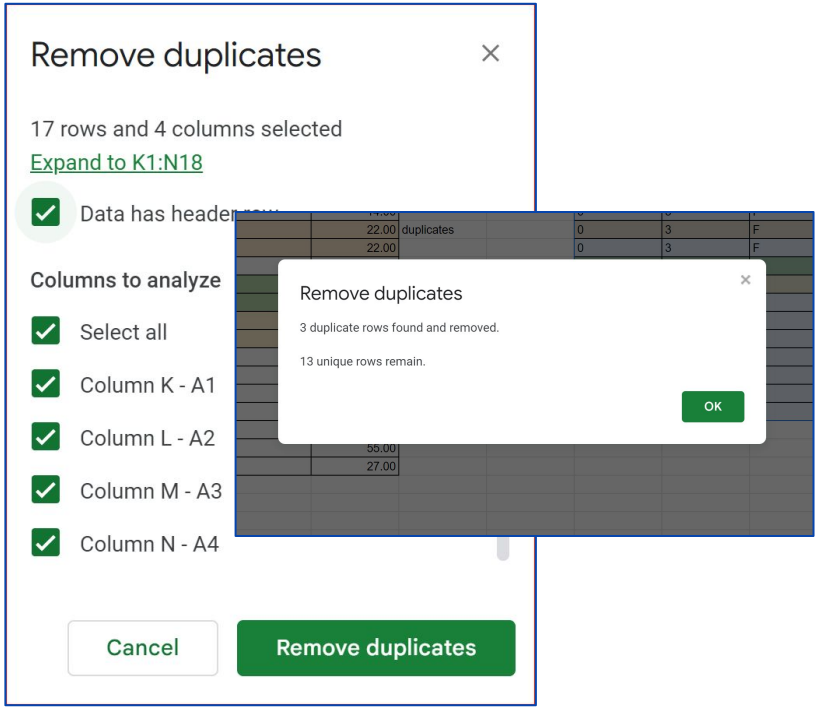
	A	B	C	D	E	F	G	H	I
1	original data					sorted data			
2	A1	A2	A3	A4		A1	A2	A3	A4
3	0	3	F	22.00		0	1	F	54.00
4	1	1	F	38.00		0	3	F	14.00
5	0	3	F	39.00		0	3	F	22.00
6	0	3	M	2.00		0	3	F	22.00
7	0	3	F	35.00		0	3	F	35.00
8	0	3	M	28.00		0	3	F	39.00
9	0	1	F	54.00		0	3	F	39.00
10	0	3	M	2.00		0	3	M	2.00
11	1	3	4	27.00		0	3	M	2.00
12	1	2	F	14.00		0	3	M	20.00
13	0	3	F	22.00		0	3	M	28.00
14	1	1	M	58.00		1	1	F	38.00
15	0	3	M	20.00		1	1	M	58.00
16	0	3	F	39.00		1	2	F	14.00
17	0	3	F	14.00		1	2	F	55.00
18	1	2	F	55.00		1	3	4	27.00



Remove duplicates

- There are three duplicates. Be sure they represent the same real world objects (e.g., passengers).
- Make a copy of the sorted data and remove the duplicates


Titanic.train
File Edit View Insert Format Data Tools Extensions Help
Last edit was 4 minutes ago
Sort sheet
Sort range
Create a filter
Filter views
Add a slicer New
Protect sheets and ranges
Named ranges
Named functions New
Randomize range
Column stats
Data validation
Data cleanup
Split text to columns
Data connectors New
Cleanup suggestions New
Remove duplicates
Trim whitespace
original data
w/o duplicates


Remove duplicates
17 rows and 4 columns selected
Expand to K1:N18
Data has header row
Columns to analyze
Select all
Column K - A1
Column L - A2
Column M - A3
Column N - A4
Cancel
Remove duplicates
Remove duplicates
3 duplicate rows found and removed.
13 unique rows remain.
OK