Statistical Techniques for Astronomy

Priya Hasan & Najam Hasan
MANUU, Hyderabad, India
priya.hasan@gmail.com
hasan.najam@gmail.com

Changing perspective

If your experiment needs statistics, you ought to have done a better experiment.

- Ernest Rutherford

God does not play dice with the Universe.

- Albert Einstein

Lies, damned lies and statistics.

- Benjamin Disraeli

We have come to the realisation that in the present Era of rich datasets it is essential to have sophisticated Statistical techniques and methodologies to analyse and interpret data.

Statistics has changed meaning over centuries:

- Originally it referred to collection & compilation of data.
- In 19^{th} century it accrued the goal of mathematical interpretation of data.
- Contemprary statistics is an amaglam of science, technology and art.

- C.R. Rao (Statistics & Truth)

Astronomy too has changed over the years:

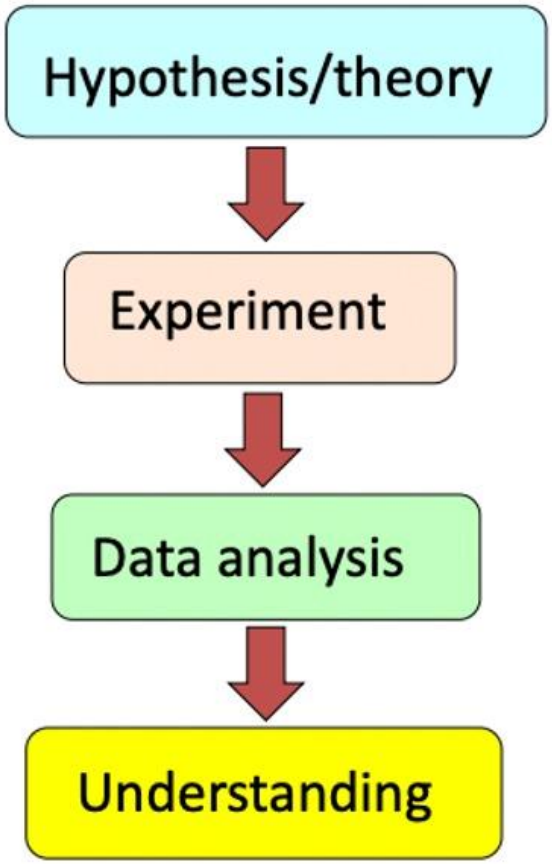
From data poor to data rich science.

We live in era of big astronomical survey:

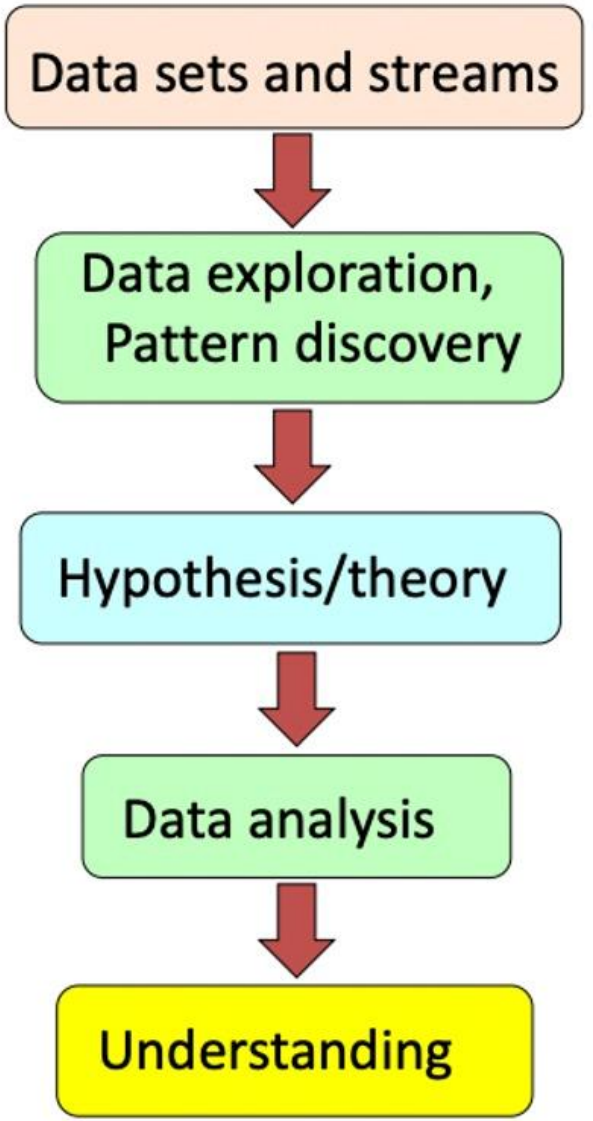
- **2dF Galaxy Redshift Survey (2dfGRS) – redshift survey conducted by the Anglo-Australian Observatory between 1997 - 2002**
- **Sloan Digital Sky Survey (SDSS) – an optical and spectroscopic survey, 2000–2006**
- **DEEP2 Redshift Survey (DEEP2) – Keck Telescopes to measure redshift of 50,000 galaxies**
- **SAGES Legacy Unifying Globulars and GalaxieS (SAGES Legacy Unifying Globulars and GalaxieS Survey (SLUGGS) survey[8] – a near-infrared spectro-photometric survey of 25 nearby early-type galaxies (2014)**
- **Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) – an extra-galactic and stellar spectroscopic survey**
- **Optical Gravitational Lensing Experiment (OGLE) – large-scale variability sky survey (in I and V bands), 1992-present**
- **2MASS**
- **GAIA (designed for astrometry: positions, distances and motions of stars with unprecedented precision – 1 billion stars)**

It has become essential for astronomers to handle the big datasets and interpret the complex astrophysical problems by using sophisticated statistical techniques and technology.

Hypothesis-driven science



Data-driven science



The two approaches are complementary

Virtual Observatory

- A virtual observatory (VO) is a collection of interoperating data archives and software tools which utilize the internet to form a scientific research environment in which astronomical research programs can be conducted.

- The main goal is to allow transparent and distributed access to data available worldwide. This allows scientists to discover, access, analyze, and combine nature and lab data from heterogeneous data collections in a user-friendly manner.

- The IVOA (International Virtual Observatory Alliance) is a standards body created by the VO projects to develop and agree the vital interoperability standards upon which the VO implementations are constructed.

- Wikipedia

Statistics in Astronomy

- Probability theory
- Time Series analysis
- ANOVA
- Correlation & Regression
- Principal Component analysis
- Clustering
- Hypothesis tests
- Point estimation for use in the analysis of modern astronomical data.
- Least squares
- Maximum likelihood, and Bayesian approaches to statistical inference
- Resampling methods - bootstrap
- Goodness of fit

Probability:

Question 1:

We flip an unbaised coin 3 times

All 3 times we get **heads**

What would you expect in the next flip?

Question 2:

These rocks have balanced for about 2,500 million years.

Is it safe to stand under them?

Is it safe to stand under them?

Statistics is **divided into two main branches:**

- **Descriptive statistics**

  analysis of data that helps to describe, show and summarize data in a meaningful way.

- **Inferential statistics**

  is used to make predictions by taking any group of data in which you are interested.

Statistics – Bayesian vs. Non-Bayesian

Similarities:

- Both forms of statistics use the likelihood function.
- Both forms of statistics are used for statistical inference.
- Both forms of statistics can be used for data analysis and visualization.

Bayesian refers to a statistical approach named after Reverend Thomas Bayes,

He formulated a theorem that describes how to update probabilities based on new data.
In a Bayesian framework:

1. Prior beliefs are represented as a probability distribution (the prior).
2. New data is observed.
3. Bayes' theorem is applied to update the prior beliefs, yielding a posterior distribution.

This approach is used in various fields, including machine learning, data science, and statistics,

Aim:

- Update beliefs based on new data
- Make predictions
- Estimate model parameters
- Perform model selection

Probability in Astronomy:

Before 1987, four naked-eye supernovae had been recorded in ten centuries.

What, before 1987, was the probability of supernova happening in the twentieth century?

Answers:

i. Supernovae are physical events, in principle, can be accurately determined. They are not Random events. Probability is meaningless!

ii. From frequentist point of view: 4/10

(assumption: supernovae were equally likely to be reported in the past 10 centuries)

iii. Apriori argument: In principle we might know the stellar mass function, life-time of a star as a function of mass and star formation rate in the galaxy and the detection efficiency – further take into account metallicity, obscuration due to dust cloud etc.

Now if we sight supernova in 1987

Approach i: One supernova does not affect another (Can't predict)

Approach ii: Revise the probability to 5/10

Approach iii: Will improve models with fresh data.

Revisiting the Supernova problem:

Data: 4 Supernovae in 10 centuries

Our prior on ϱ is uniform between 0 to 1

Suitable Model: Binomial distribution
(occurs or doesn't occur)

Prosterior probability is

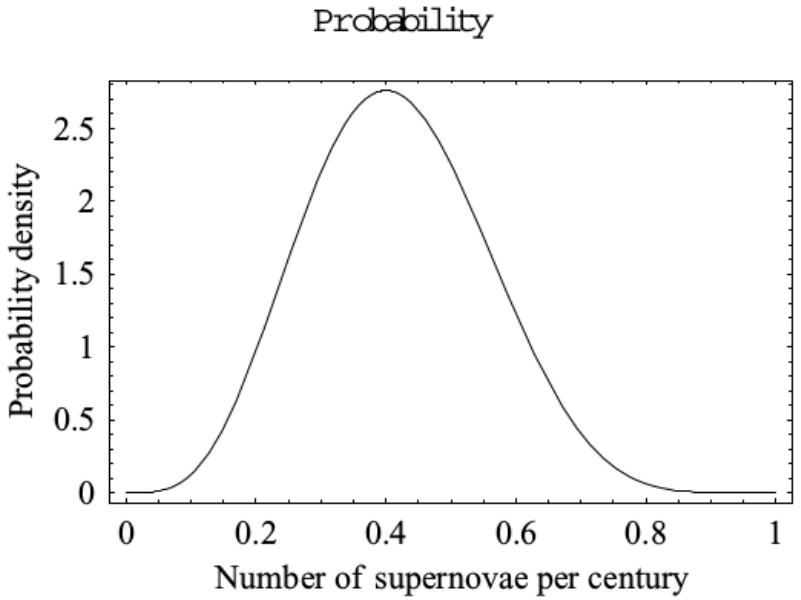
$$\text{prob}(\rho \mid \text{data}) \propto \binom{10}{4} \rho^4 (1-\rho)^6 \times \text{prior on } \rho.$$

Normalise

$$\int_0^1 \text{prob}(\rho \mid \text{data})\, \mathrm{d}\rho = 1$$

$$\int_0^1 \binom{10}{4} \rho^4 (1-\rho)^6 \, \mathrm{d}\rho$$

$$= \frac{\Gamma(10)\Gamma(4)}{\Gamma(14)} = B[5,7]$$



In general $\text{prob}(\rho \mid \text{data}) = \dfrac{\rho^n (1-\rho)^{m-n}}{B[n+1, m-n+1]}$ (for n = 4, m = 5)

Problem: Before 1987, four naked-eye supernovae had been recorded in ten centuries. In 1987 one more supernova was observed. Compare by non-Bayesian and Bayesian methods calculate the probability of supernova happening in the twentieth century?

Results Comparison

Method	Before 1987	After 1987
Frequentist	32.97%	36.5%
Bayesian	37.91%	40.7%

Key Observations

- The Bayesian approach gives higher probabilities as it accounts for uncertainty in λ
- Both methods show increased probability after the 1987 observation
- The frequentist method relies solely on observed frequencies

Principal component analysis (PCA)

When many variables are present – PCA the best correlation searcher

In a sample of *'N'* objects with *'n'* parameters measured for them

PCA is used to find
- what is correlated with what
- which variables produce primary correlation, which secondary....

The task for PCA:

Given a sample of *N* objects with *n* measured variables x_n for each, find a new set of ξ_n variables that are orthogonal.

Each one a linear combination of the original variables:

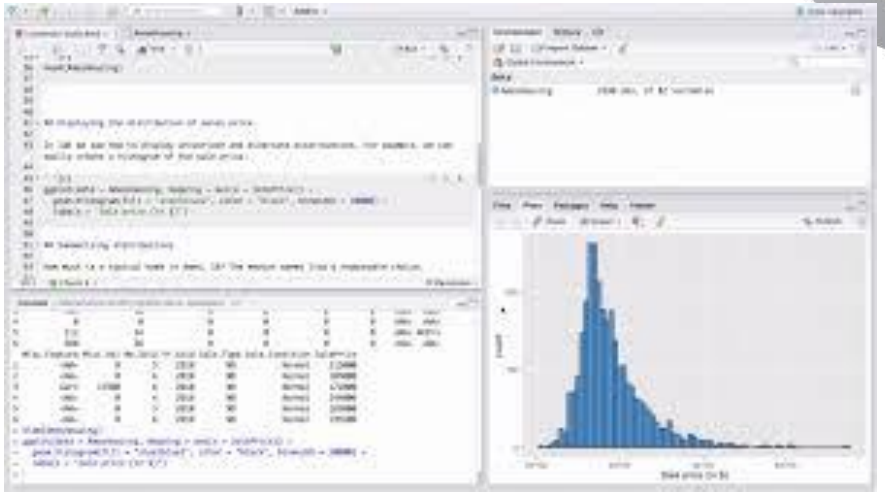
$$\xi_i = \sum_{j=1}^{n} a_{ij} x_j$$

with values of a_{ij} *such* that the smallest number of new variables accounts for as much of the variance as possible.

The ξ_i are the principal components.

Statistical Data Analysis – Tools:

- SPSS
- R
- VO tools
- Python



VO Tools

VOTable display and analysis

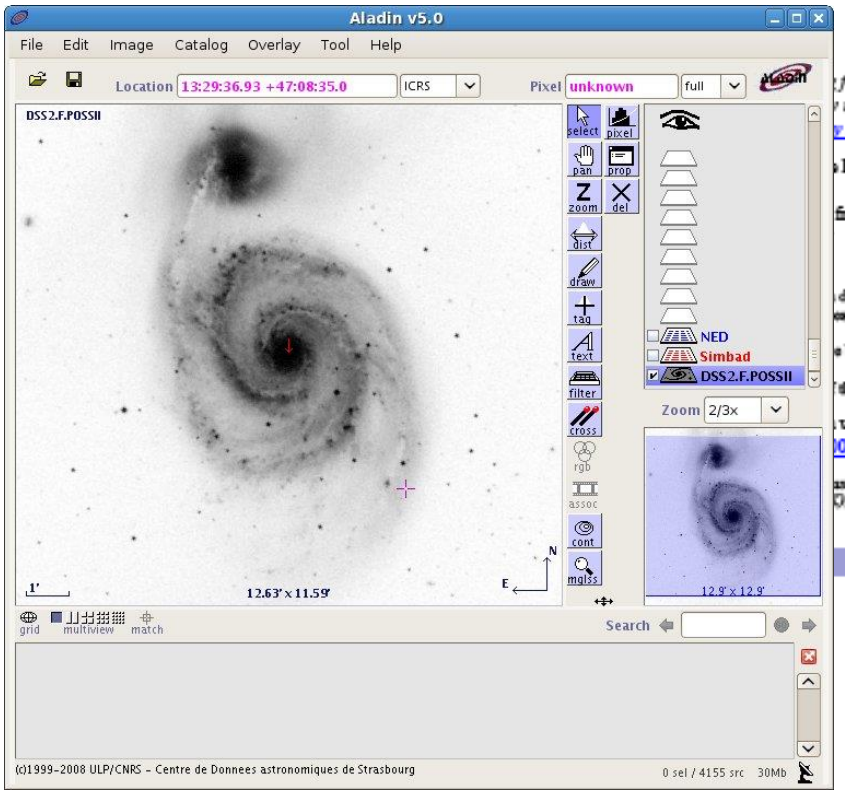
VOPlot, Topcat

Image display and analysis
Aladin, ESASky

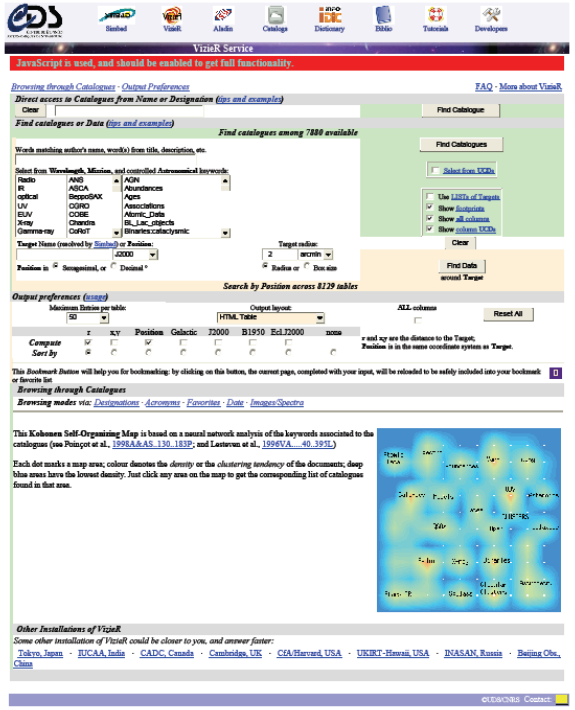
Other standard display tools for downloaded data
Spectrum display and analysis
VOSpec, SpecView

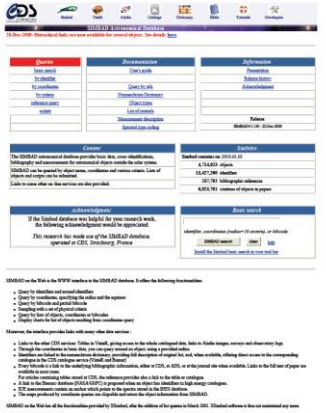
Aladin

is an interactive software sky atlas allowing the user to visualize digitized astronomical images, superimpose entries from astronomical catalogues or databases.



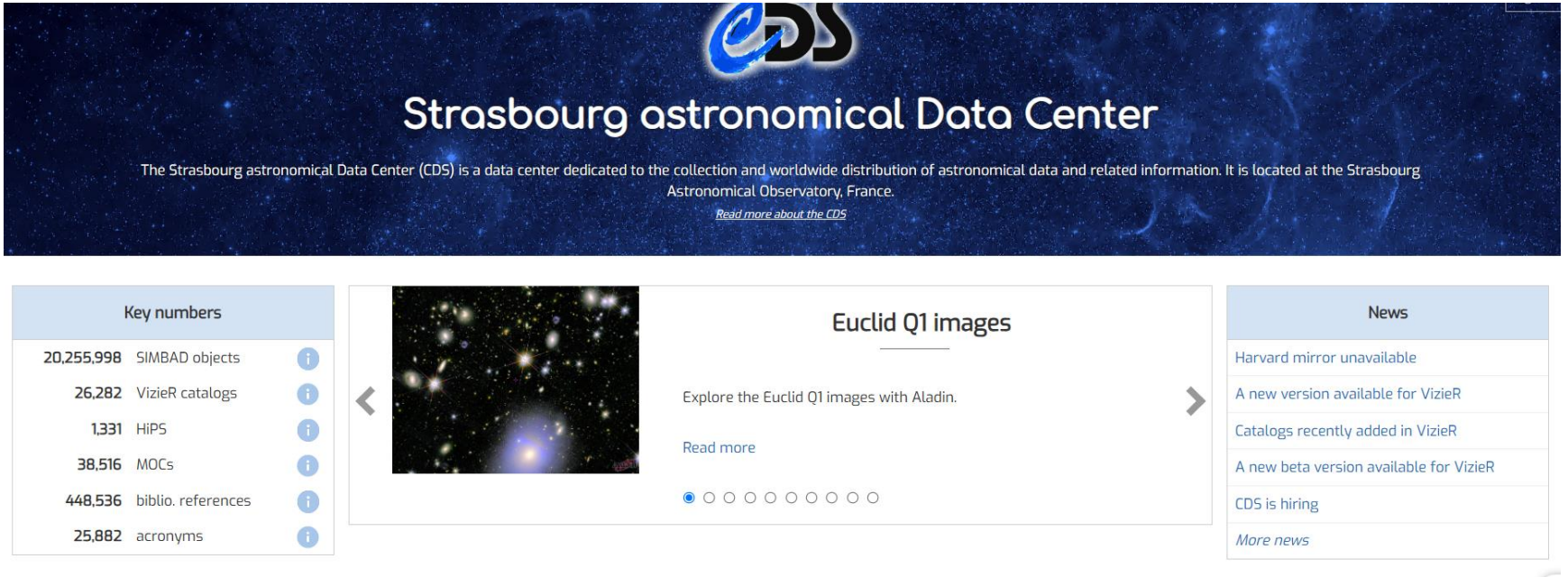
Access to data ..... CDS





Catalogues

*The **Strasbourg astronomical Data Center (CDS)** collects and distributes astronomical data catalogues, related to observations of stars and galaxies, other galactic and extragalactic objects, solar system bodies and atomic data.*



Simbad contains on *2025.05.28*

20,255,998 objects

68,571,124 identifiers

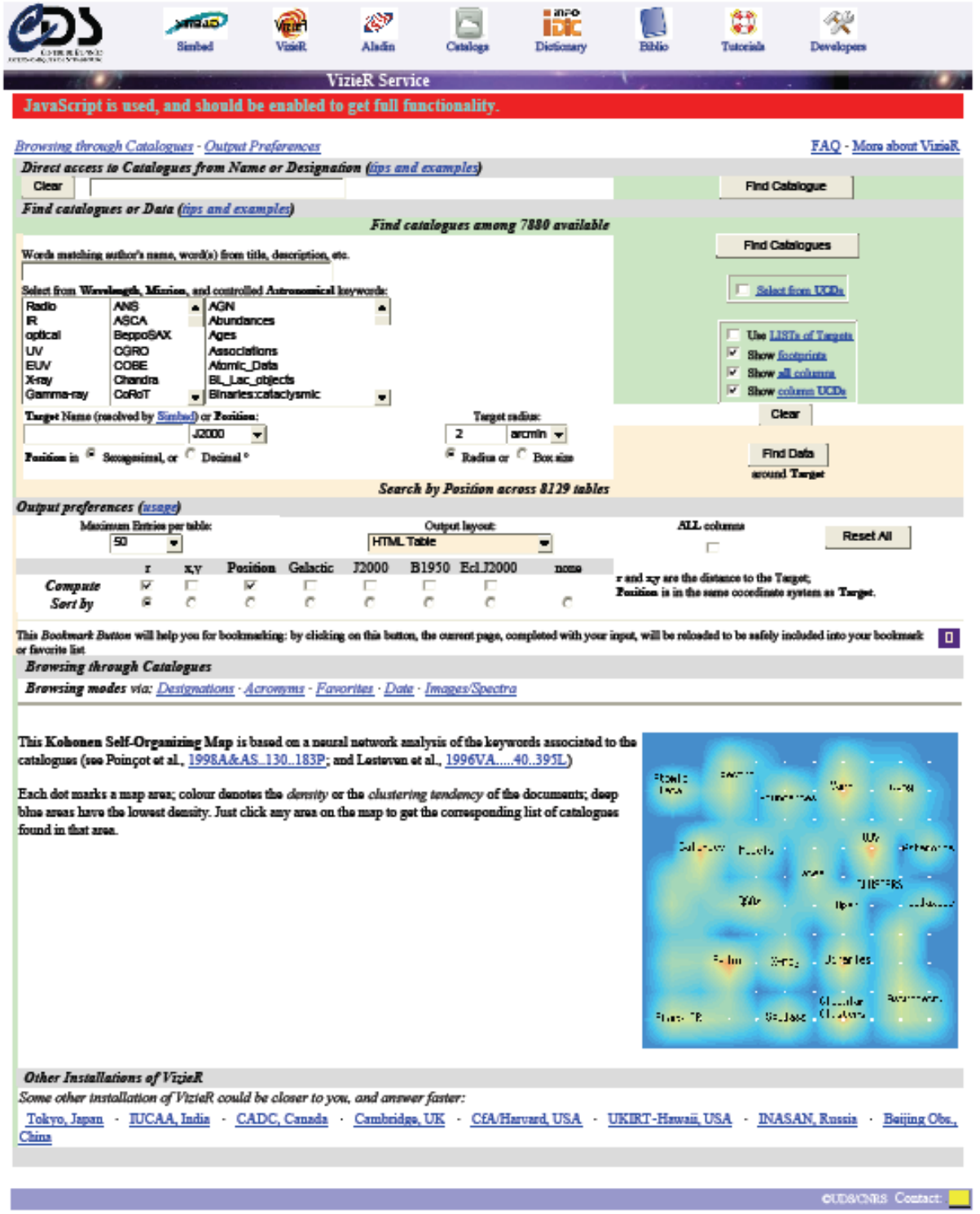
448,536 bibliographic references

46,118,463 citations of objects in papers

15,697 acronyms described for Simbad

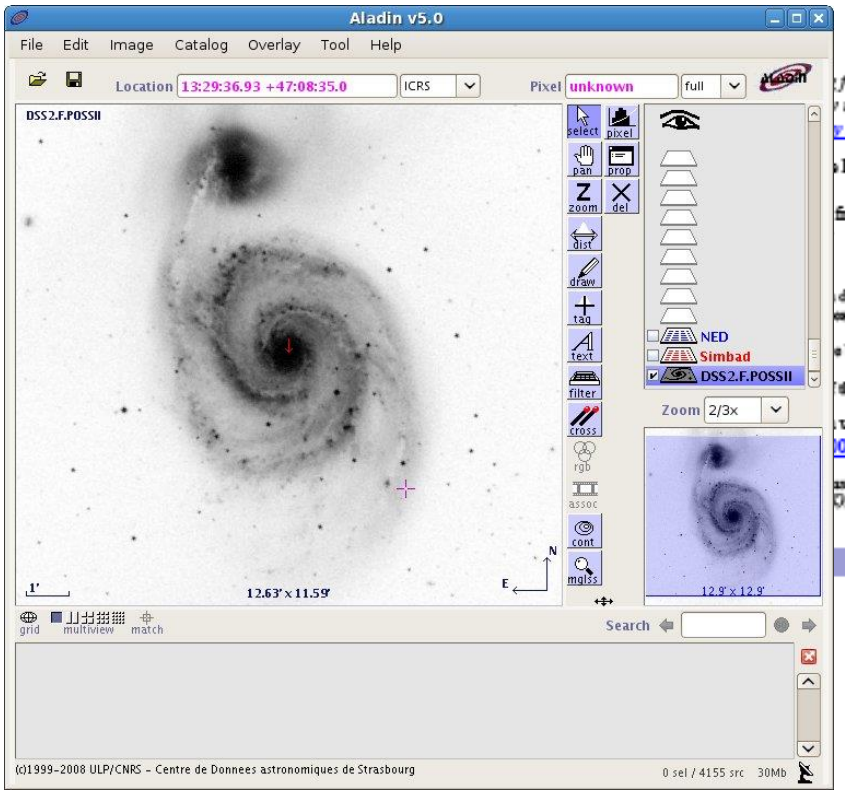
Vizier Services

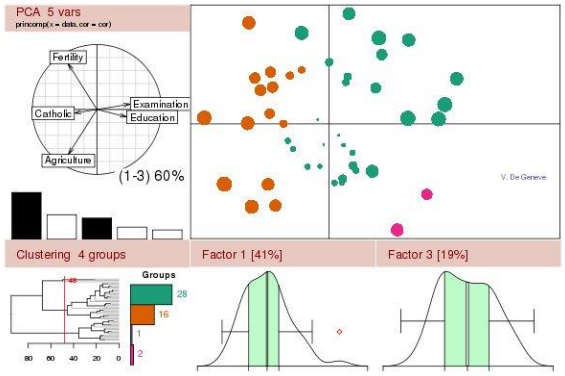
Query service for catalogues, bibliography, cross-matching across catalogues, NED objects, Simbad objects.



Aladin

is an interactive software sky atlas allowing the user to visualize digitized astronomical images, superimpose entries from astronomical catalogues or databases.





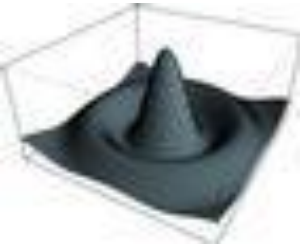
R

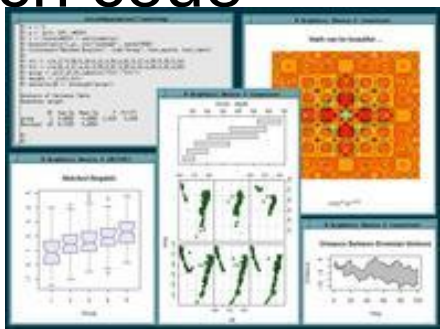
R is a free software environment for statistical computing and graphics.

It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues.

R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.





Plotting Tools

- **Topcat:** An interactive graphical viewer and editor for tabular data.
- **Python:** Matplotlib, Seaborn, Plotly…….
- **VOPlot** A tool for visualizing astronomical data.
- **VOMegaPlot** or handling large number of points (in the range of millions), similar to VOPlot and both these tools have certain common functionality.
- **STILTS**:processing of tabular data; the package has been designed for, but is not restricted to, astronomical tables such as object catalogues.
- **VOConvert** A tool for converting files from one format to another. It supports following file format conversions: (1) ASCII to VOTable (2) FITS to VOTable and (3) VOTable to ASCII.

Using Python

```
In [1]: matplotlib inline
```

```
In [3]: import pandas as pd
        from ydata_profiling import ProfileReport
        import matplotlib.pyplot as plt
        import seaborn as sns
```

```
In [4]: plt.style.use('ggplot')
```

```
In [5]: data=pd.read_csv('dias2018.csv')
```

```
In [6]: data
```

Out[6]:

	LII	BII	LD	RGC	LOG_AGE	PM_RA	PM_DEC	NUM_CLUSTER_STARS	RAD_VEL	RAD_VEL_ERROR	METAL
0	306.5592	-21.4418	4.012222	8262.035772	9.300	-6.30	-5.41	67	0.00	0.00	0.000
1	313.5830	-30.7574	5.752015	6566.127096	9.700	0.96	-4.82	37	0.00	0.00	0.000
2	292.4817	-21.6535	1.128074	8592.322145	6.900	-30.00	27.80	18	16.10	0.50	0.000
3	307.2214	-16.3409	3.197567	7567.324146	9.350	-2.81	-3.68	122	0.00	0.00	-0.050
4	300.2173	-15.6085	0.596600	8478.627331	6.602	-39.50	-1.00	10	13.00	5.00	0.000
5	291.1373	-35.7047	1.622333	8923.234812	9.300	4.39	3.83	22	0.00	0.00	0.000
6	317.7172	-19.3386	6.498056	9881.110039	9.315	-1.93	-7.56	238	0.00	0.00	0.000
7	305.9925	-8.6218	4.459963	10344.985938	9.100	-2.49	-0.93	77	0.00	0.00	0.000
8	283.7728	-38.2608	7.975018	8007.584768	9.480	2.86	5.71	120	4.96	3.60	0.000
9	308.6803	-8.2906	11.629630	7001.344056	8.300	-3.48	-3.06	670	0.00	0.00	0.000

In [7]:
```
data.describe()
```

Out[7]:

	LII	BII	LD	RGC	LOG_AGE	PM_RA	PM_DEC	NUM_CLUSTER_
count	2166.000000	2166.000000	2166.000000	2166.000000	2166.000000	2166.000000	2166.000000	2166.000000
mean	190.859480	-0.797745	4.799574	8778.562255	7.737534	-1.872461	-1.226519	288.974146
std	98.873882	8.289514	5.558965	1938.267074	2.281574	4.260588	4.251714	802.063735
min	0.009400	-79.261200	0.000000	1369.671399	0.000000	-69.590000	-45.000000	0.000000
25%	107.977725	-2.596550	1.604889	7711.184526	7.506250	-3.417500	-3.380000	23.000000
50%	204.348550	-0.403800	3.198148	8540.524642	8.450000	-1.570000	-1.340000	76.000000
75%	276.256400	1.630850	5.882848	9704.337318	8.900000	0.000000	0.677500	203.000000
max	359.994200	84.591400	75.705981	23360.246914	10.100000	36.390000	27.800000	14341.000000

In [8]:
```
pp.ProfileReport(data)
```

Overview

Dataset info

Number of variables	11
Number of observations	2166
Total Missing (%)	0.0%
Total size in memory	186.2 KiB
Average record size in memory	88.0 B

Variables types

Numeric	11
Categorical	0
Boolean	0
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Warnings

Rejected 0
Unsupported 0

Warnings

- LD has 132 / 6.1% zeros Zeros
- LOG_AGE has 154 / 7.1% zeros Zeros
- METALLICITY has 1868 / 86.2% zeros Zeros
- NUM_CLUSTER_STARS has 65 / 3.0% zeros Zeros
- PM_DEC has 64 / 3.0% zeros Zeros
- PM_RA has 65 / 3.0% zeros Zeros
- RAD_VEL has 1464 / 67.6% zeros Zeros
- RAD_VEL_ERROR has 1493 / 68.9% zeros Zeros

Variables

BII
Numeric

Distinct count	2151	**Mean**	-0.79774
Unique (%)	99.3%	**Minimum**	-79.261
Missing (%)	0.0%	**Maximum**	84.591
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

LOG_AGE
Numeric

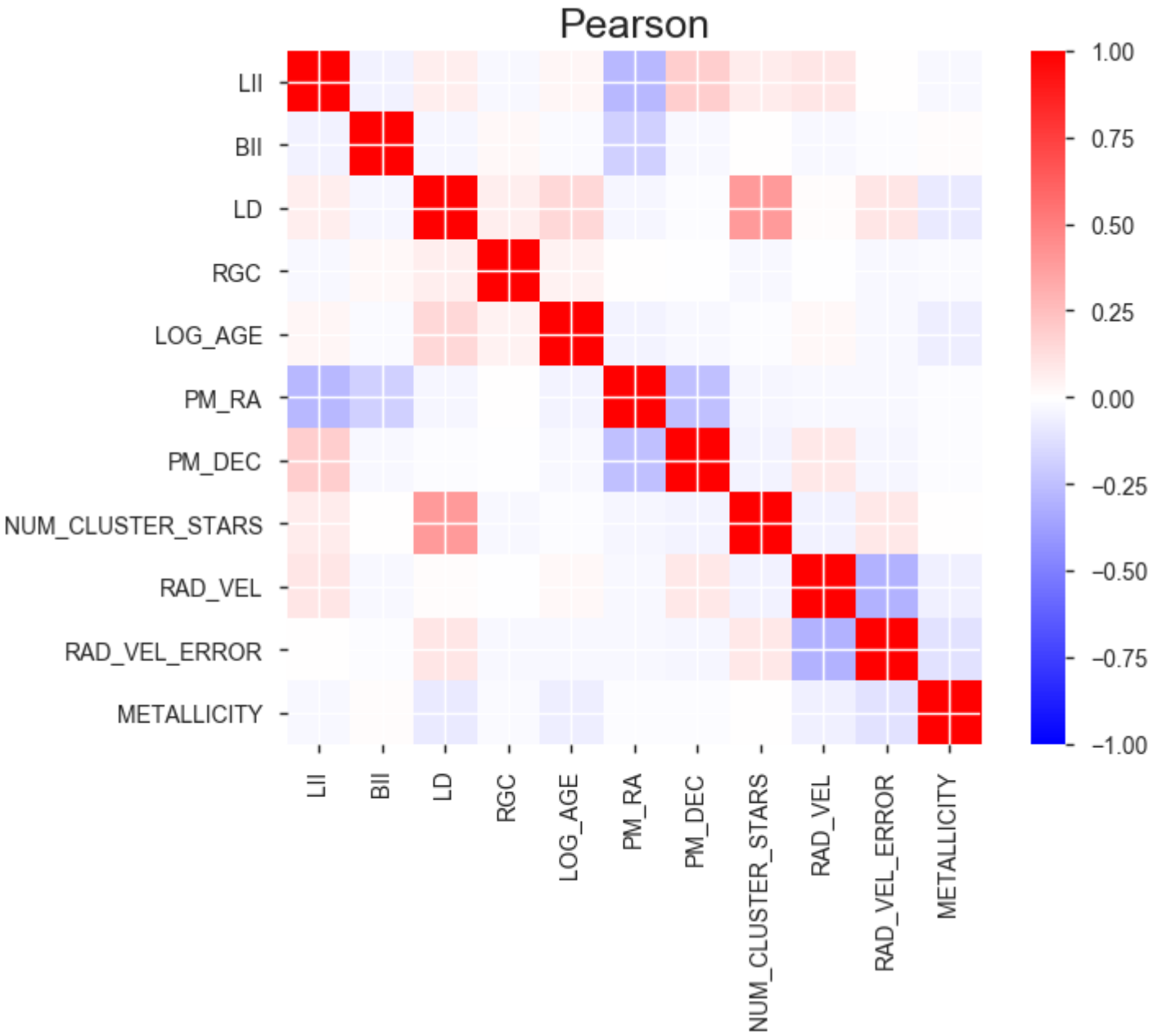
Distinct count	587	**Mean**	7.7375
Unique (%)	27.1%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	10.1
Missing (n)	0	**Zeros (%)**	7.1%
Infinite (%)	0.0%		
Infinite (n)	0		

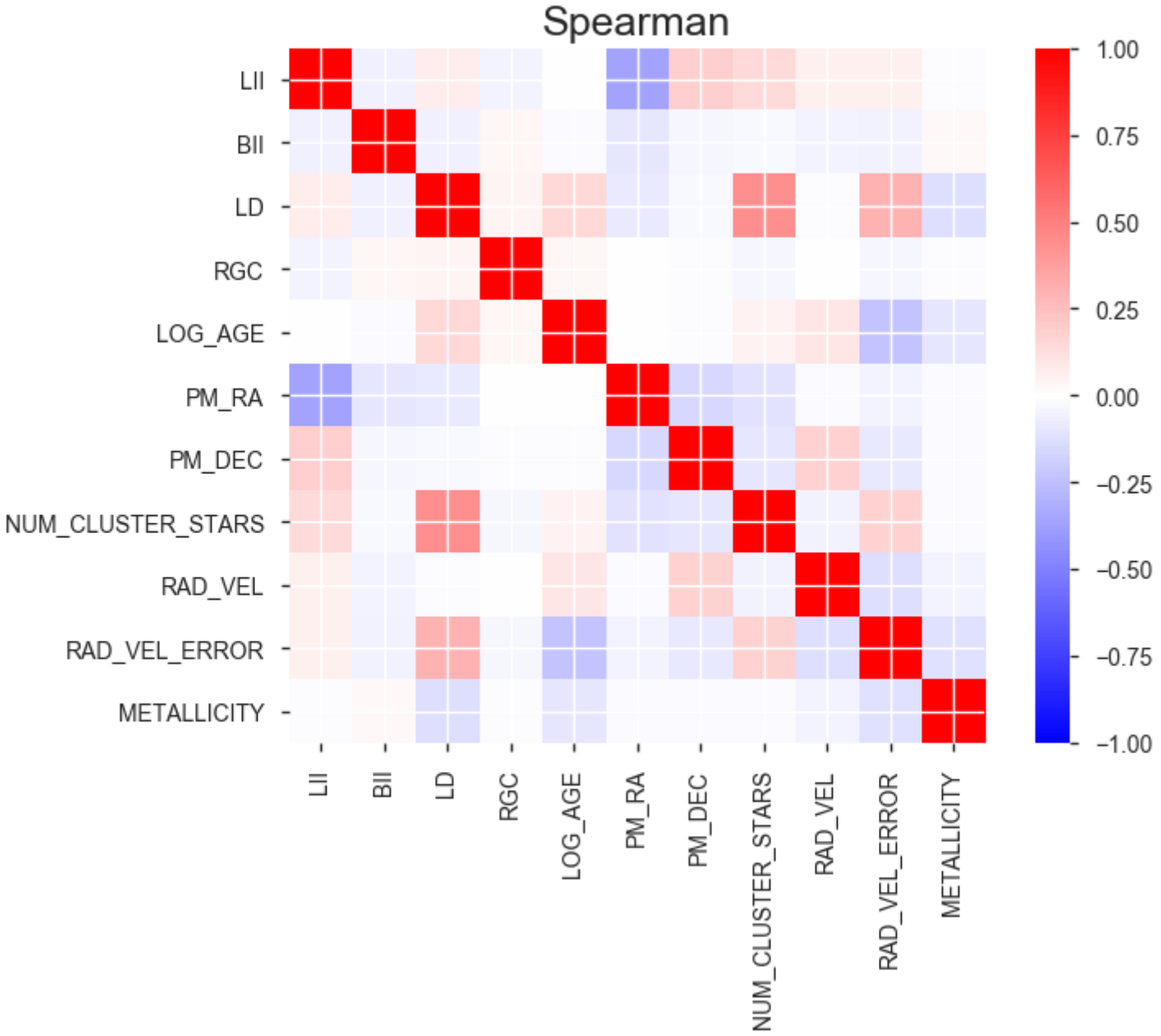
Toggle details

METALLICITY
Numeric

Distinct count	128	**Mean**	-0.01663
Unique (%)	5.9%	**Minimum**	-1.544
Missing (%)	0.0%	**Maximum**	0.46
Missing (n)	0	**Zeros (%)**	86.2%

Correlations




Spearman
LII
BII
LD
RGC
LOG_AGE
PM_RA
PM_DEC
NUM_CLUSTER_STARS
RAD_VEL
RAD_VEL_ERROR
METALLICITY
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Session with python ……..

Thank You