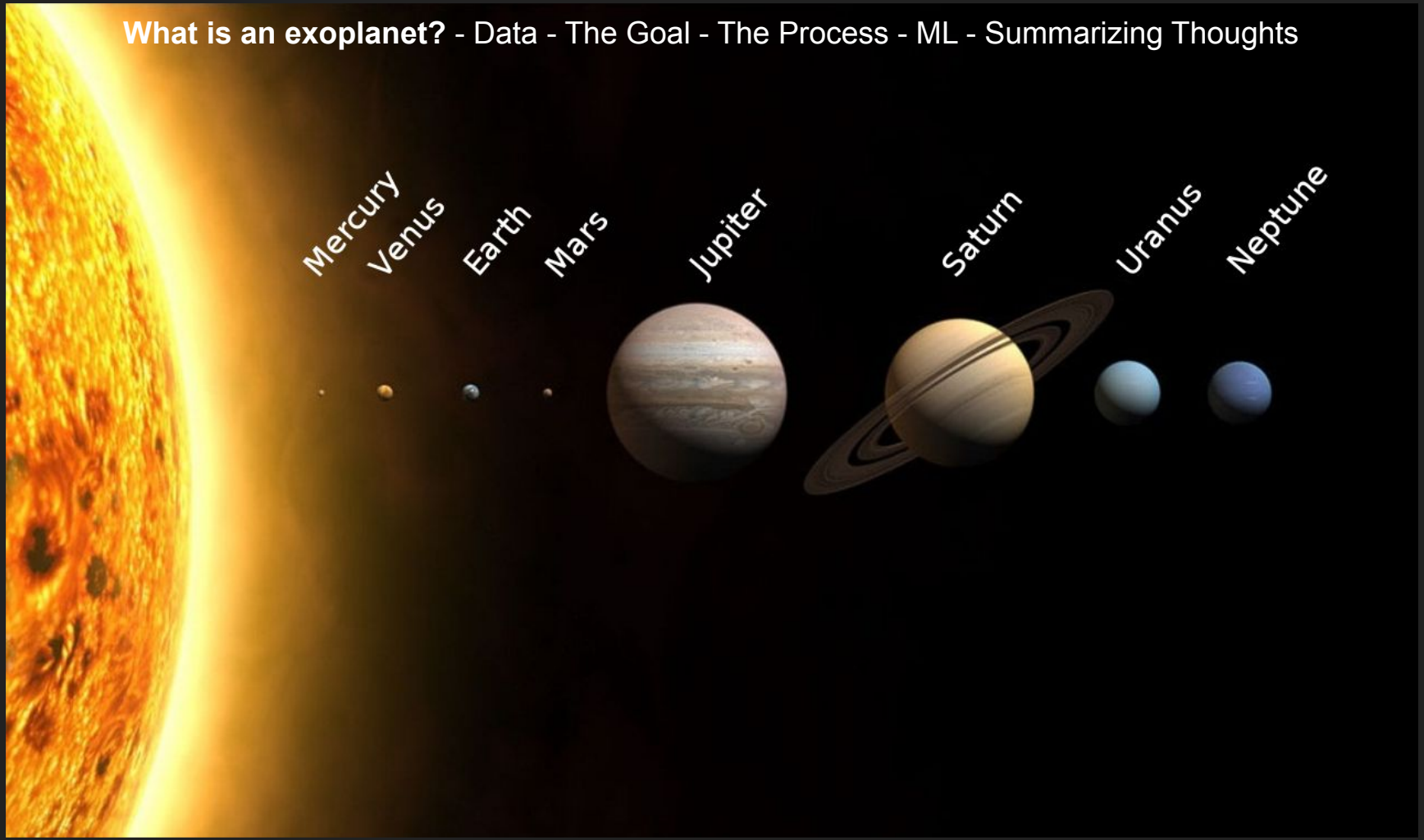
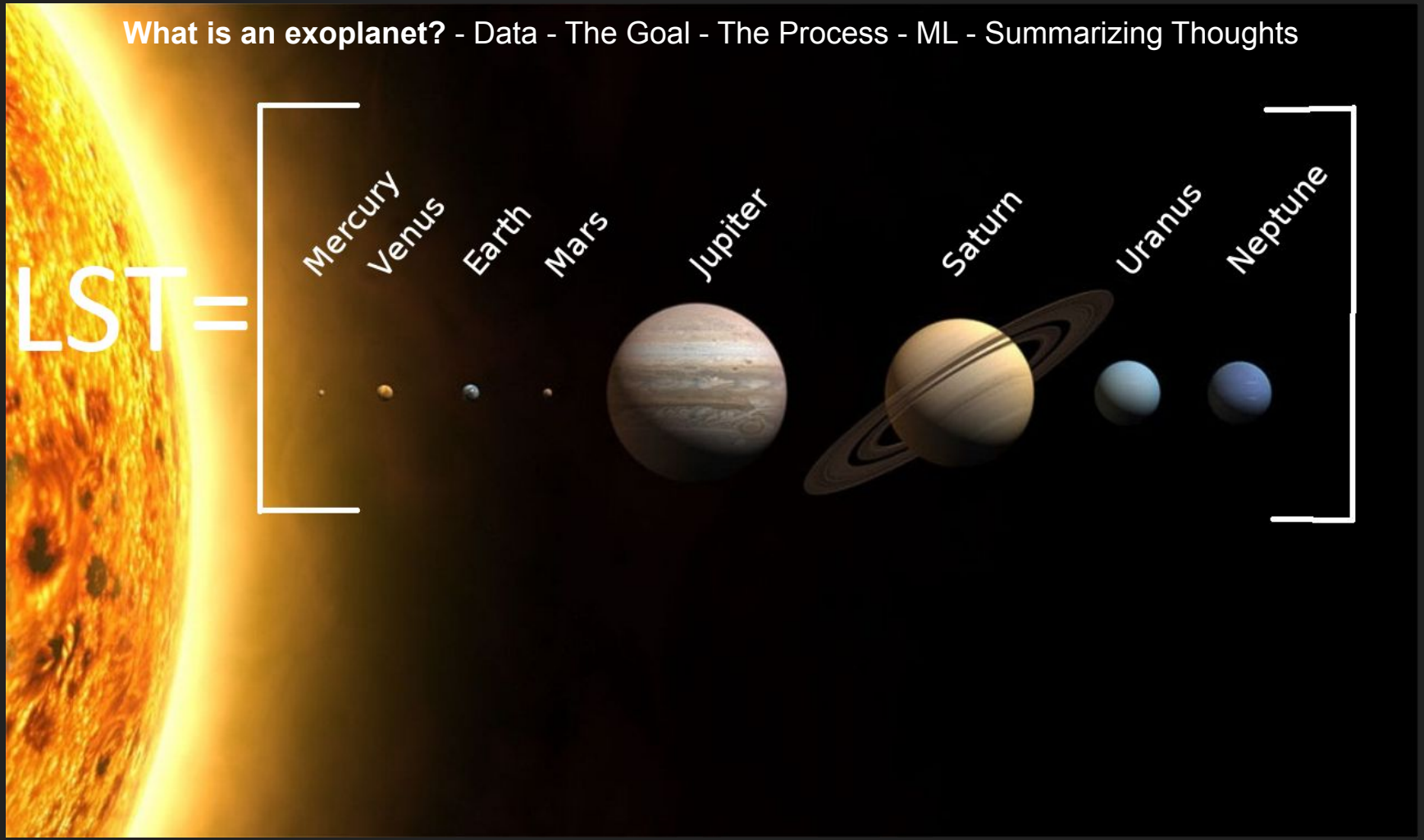
Finding Home

Finding Home

A Kevin Björk Production

What is an exoplanet? - Data - The Goal - The Process - ML - Summarizing Thoughts




LST=
Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune


LST =
Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
EXO NOT IN LST ⇨ TRUE

The Data

- Before cleaning: 4048 rows x 112 columns
- After cleaning: 4046 rows x 32 columns

The Data

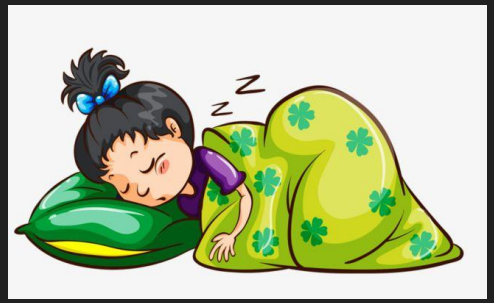
- Before cleaning: 4048 rows x 112 columns
- After cleaning: 4046 rows x 32 columns
- Originates from the Planetary Habitability Laboratory
  - Managed by the University of Puerto Rico

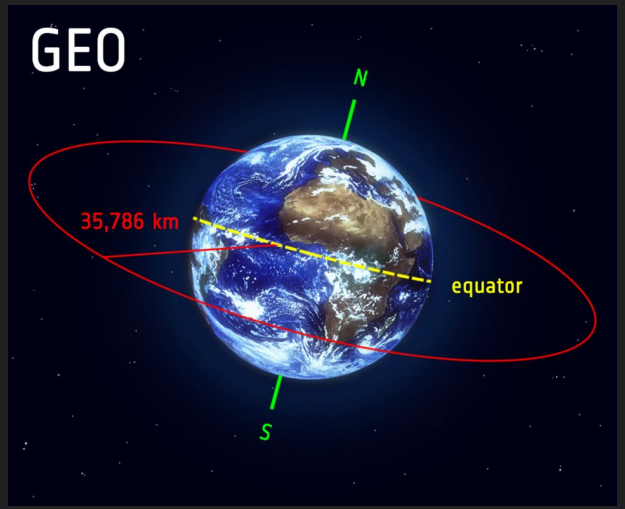
The Data

- Before cleaning: 4048 rows x 112 columns
- After cleaning: 4046 rows x 32 columns
- Originates from the Planetary Habitability Laboratory
  - Managed by the University of Puerto Rico
- Contains information about exoplanets and the stars they orbits


STAR
NUTRITION
SWEDISH
HIGH PROTEIN & CARBOHYDRATE BLEND
FOR RECOVERY & MUSCLE BUILDING
SUPREME
MASS
20
4050g

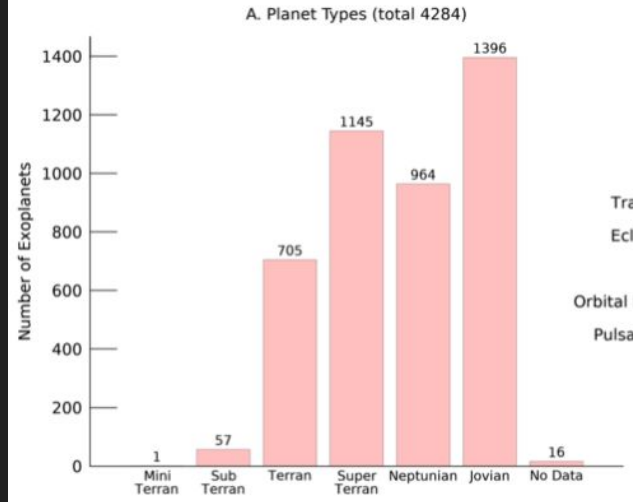

COLD
WARM
HOT

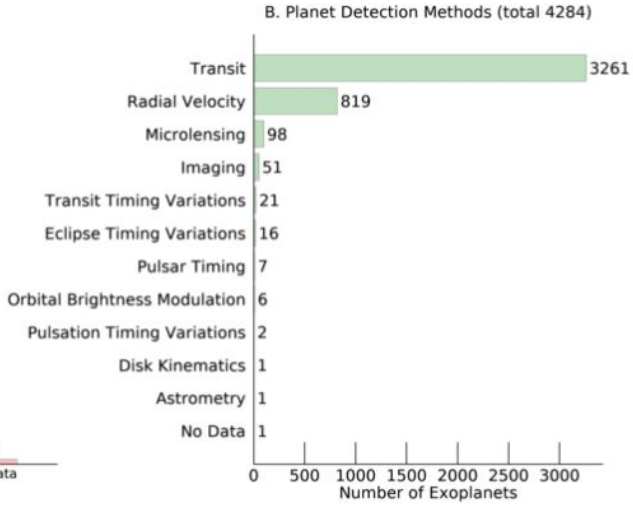


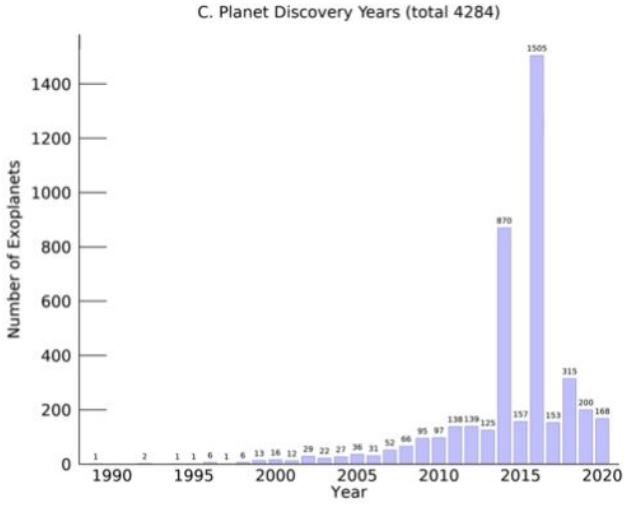

GEO
N
35,786 km
equator
S

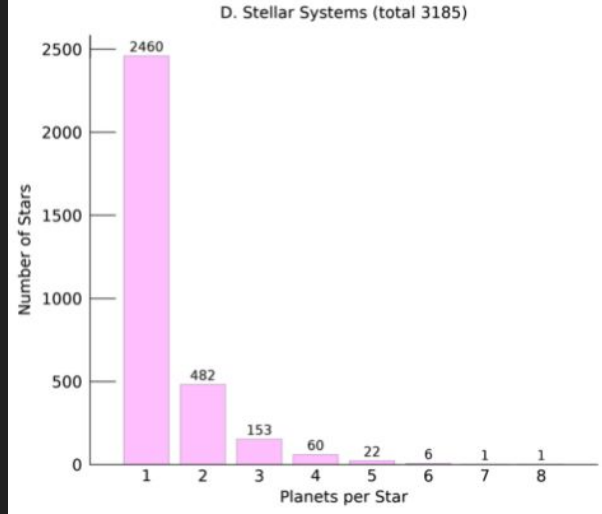
The Data

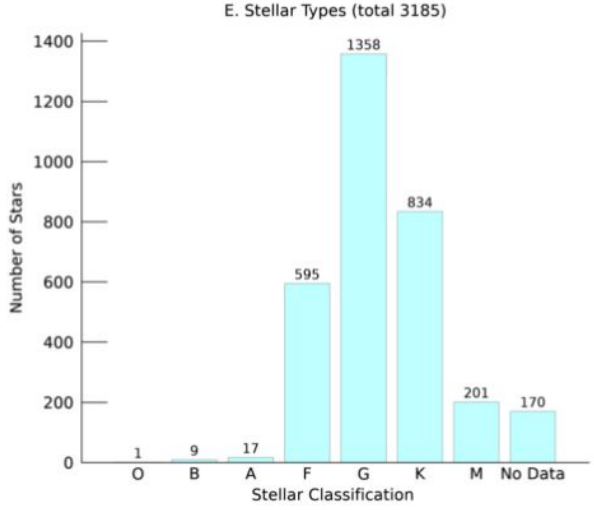
- Before cleaning: 4048 rows x 112 columns
- After cleaning: 4046 rows x 32 columns
- Originates from the Planetary Habitability Laboratory
  - Managed by the University of Puerto Rico
- Contains information about exoplanets and the stars they orbits
- Unbalanced, ca 1 % of planets are habitable

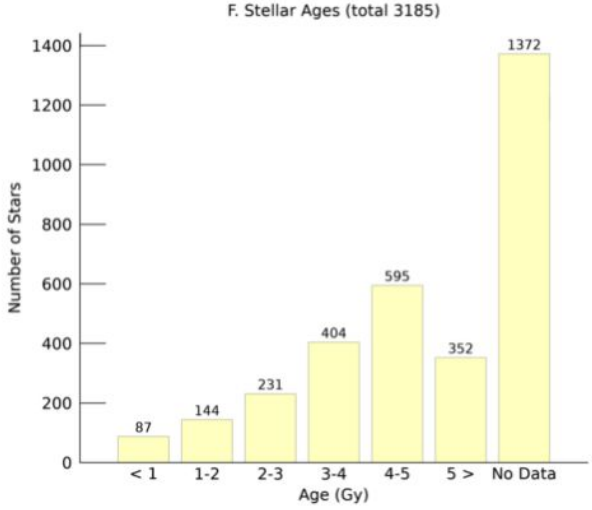

A. Planet Types (total 4284)
Number of Exoplanets
1400
1200
1000
800
600
400
200
0
1
57
705
1145
964
1396
16
Mini Terran
Sub Terran
Terran
Super Terran
Neptunian
Jovian
No Data


B. Planet Detection Methods (total 4284)
Transit 3261
Radial Velocity 819
Microlensing 98
Imaging 51
Transit Timing Variations 21
Eclipse Timing Variations 16
Pulsar Timing 7
Orbital Brightness Modulation 6
Pulsation Timing Variations 2
Disk Kinematics 1
Astrometry 1
No Data 1
0 500 1000 1500 2000 2500 3000
Number of Exoplanets


C. Planet Discovery Years (total 4284)
Number of Exoplanets
1400
1200
1000
800
600
400
200
0
1990
1995
2000
2005
2010
2015
2020
Year


D. Stellar Systems (total 3185)
Number of Stars
2500
2000
1500
1000
500
0
2460
482
153
60
22
6
1
1
1 2 3 4 5 6 7 8
Planets per Star


E. Stellar Types (total 3185)
Number of Stars
1400
1200
1000
800
600
400
200
0
1
9
17
595
1358
834
201
170
O B A F G K M No Data
Stellar Classification


F. Stellar Ages (total 3185)
Number of Stars
1400
1200
1000
800
600
400
200
0
87
144
231
404
595
352
1372
< 1 1-2 2-3 3-4 4-5 5 > No Data
Age (Gy)

GOAL:

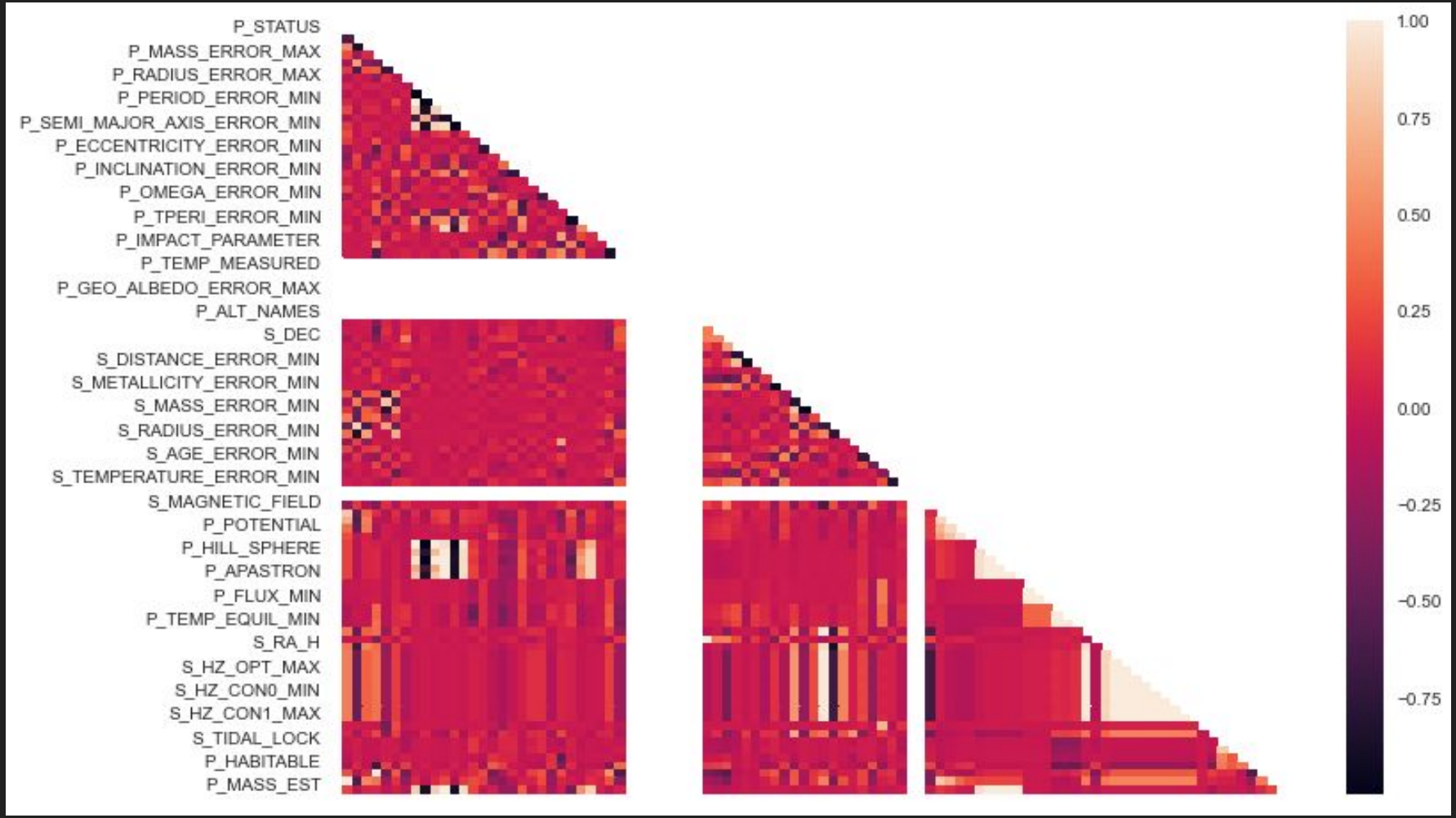
GOAL:

Is the planet habitable?

The process: DAE

The process: DAE

- Lots and lots of columns have high correlation


P_STATUS
P_MASS_ERROR_MAX
P_RADIUS_ERROR_MAX
P_PERIOD_ERROR_MIN
P_SEMI_MAJOR_AXIS_ERROR_MIN
P_ECCENTRICITY_ERROR_MIN
P_INCLINATION_ERROR_MIN
P_OMEGA_ERROR_MIN
P_TPERI_ERROR_MIN
P_IMPACT_PARAMETER
P_TEMP_MEASURED
P_GEO_ALBEDO_ERROR_MAX
P_ALT_NAMES
S_DEC
S_DISTANCE_ERROR_MIN
S_METALLICITY_ERROR_MIN
S_MASS_ERROR_MIN
S_RADIUS_ERROR_MIN
S_AGE_ERROR_MIN
S_TEMPERATURE_ERROR_MIN
S_MAGNETIC_FIELD
P_POTENTIAL
P_HILL_SPHERE
P_APASTRON
P_FLUX_MIN
P_TEMP_EQUIL_MIN
S_RA_H
S_HZ_OPT_MAX
S_HZ_CON0_MIN
S_HZ_CON1_MAX
S_TIDAL_LOCK
P_HABITABLE
P_MASS_EST
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

$$g_p = G\frac{m_p}{r_p^2}$$

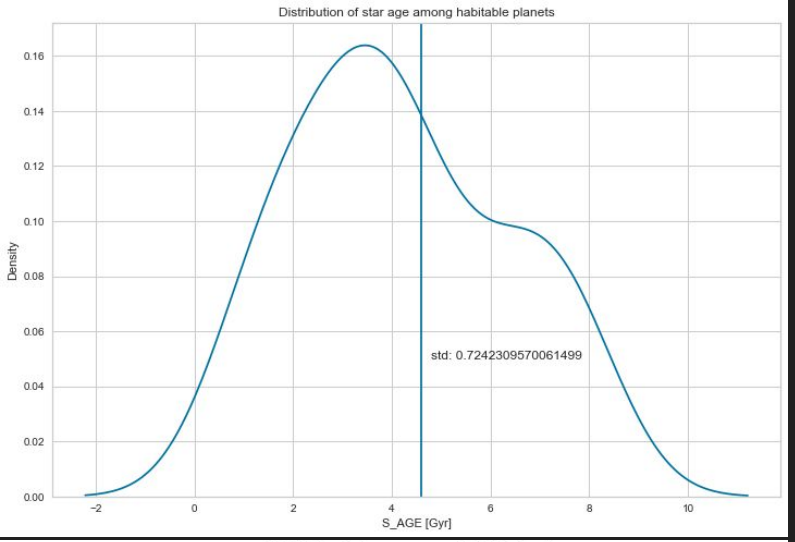
$$L_s = 4\pi r_s^2 \sigma T^4$$

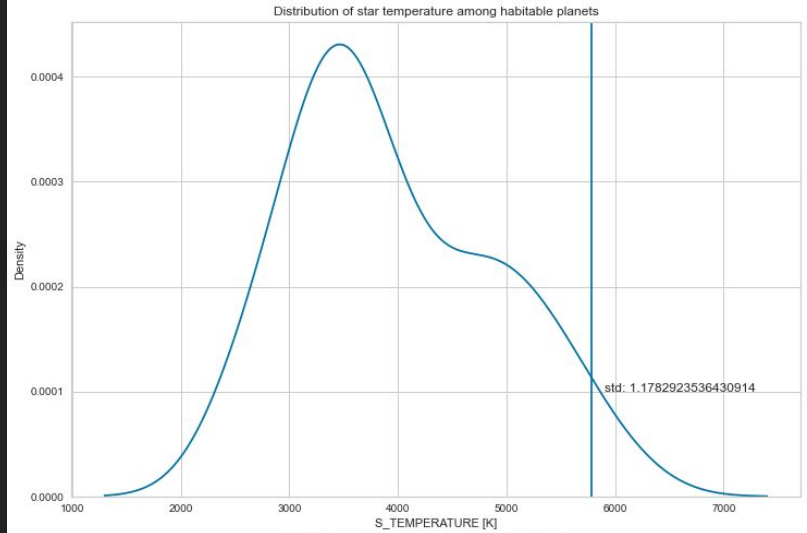
The process: DAE

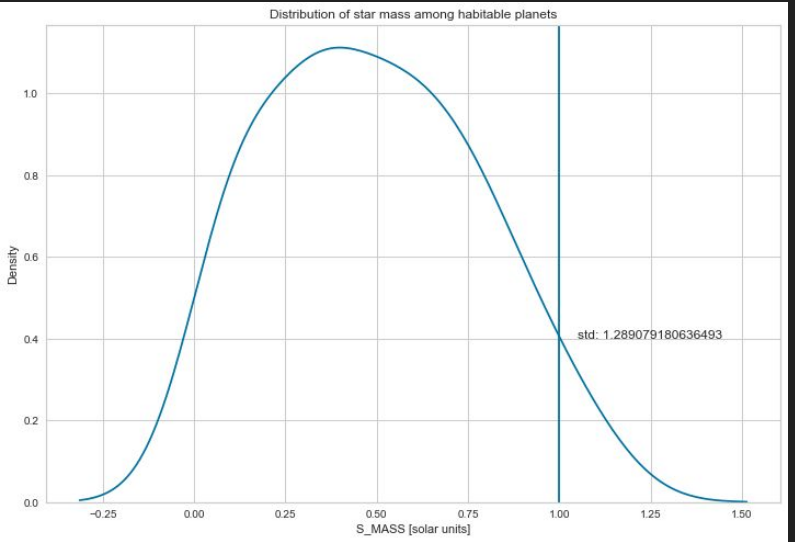
- Lots and lots of columns have high correlation
- Feature Creation: SSI

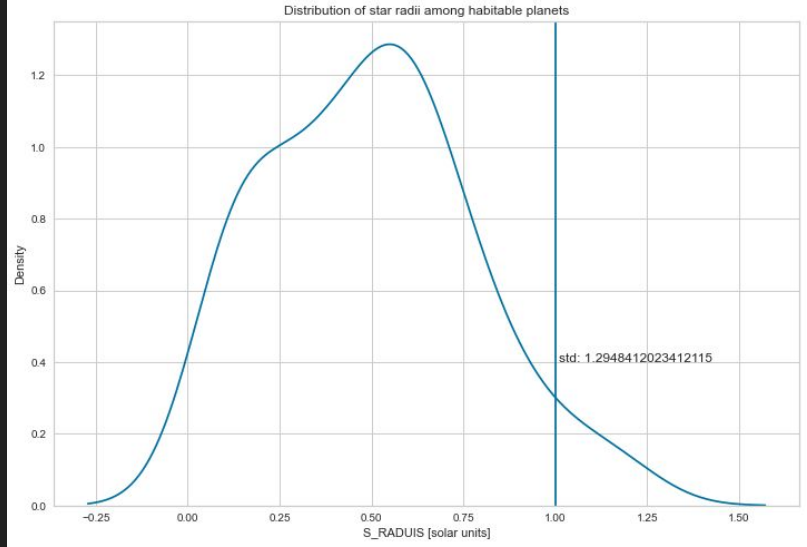
$$SSI = \sum_{i} \sqrt{\left(S_{exo,i} - S_{sun,i}\right)^2}$$

$$S_i = [Mass, Radius, ...]$$


Distribution of star age among habitable planets
Density
S_AGE [Gyr]
std: 0.7242309570061499


Distribution of star temperature among habitable planets
Density
S_TEMPERATURE [K]
std: 1.1782923536430914

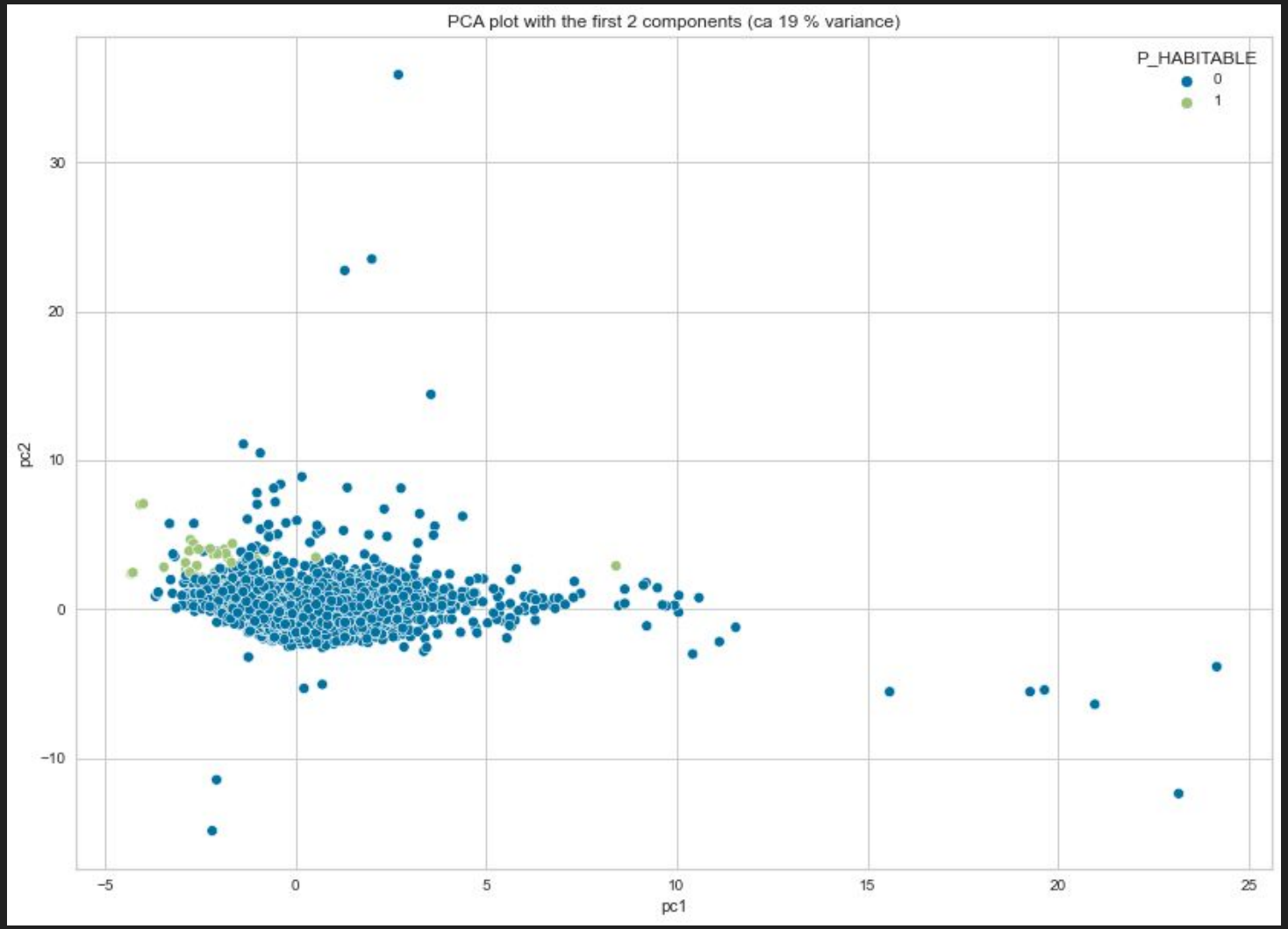

Distribution of star mass among habitable planets
Density
S_MASS [solar units]
std: 1.289079180636493

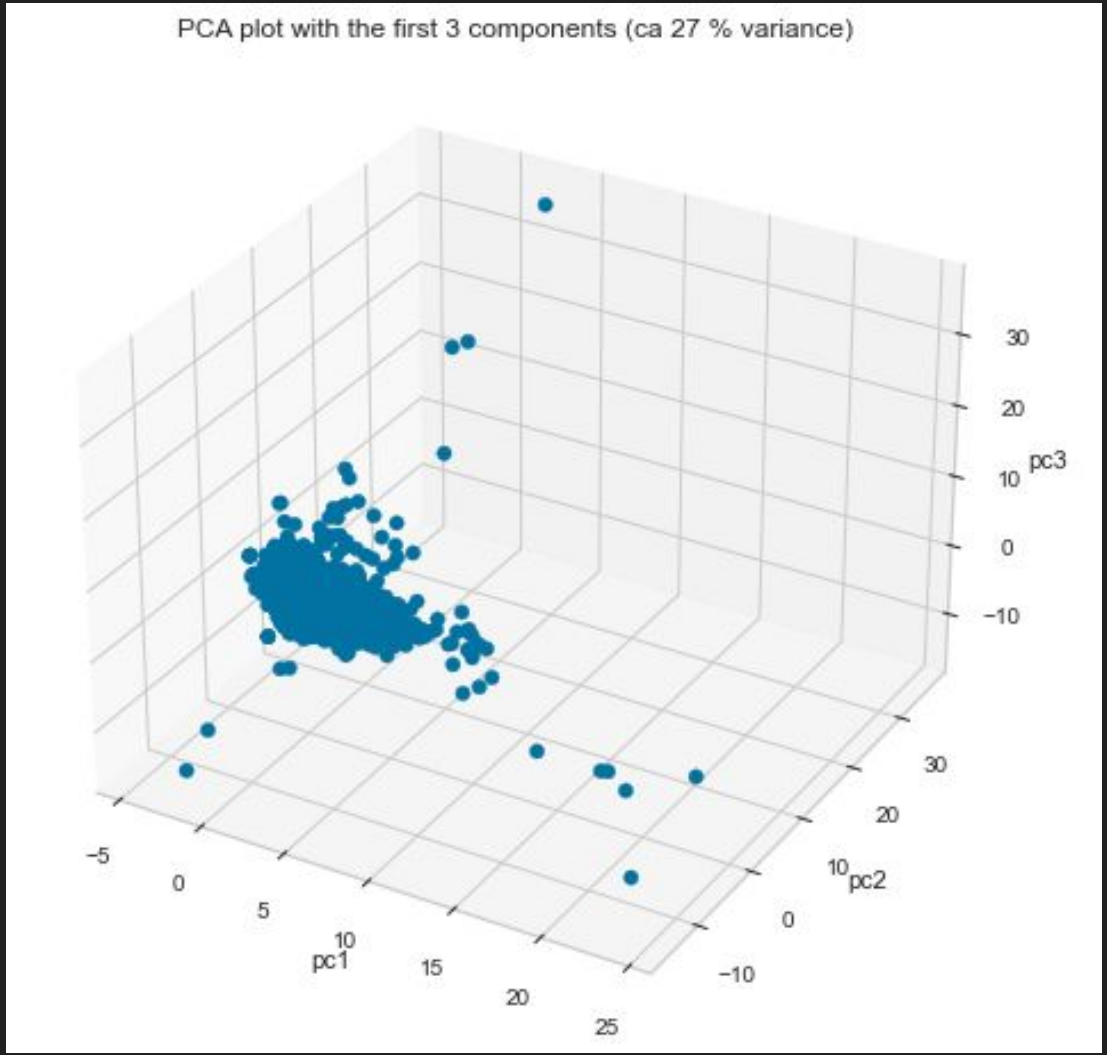

Distribution of star radii among habitable planets
Density
S_RADIUS [solar units]
std: 1.2948412023412115

The process: EDA - Habitable vs Inhabitable

The process: EDA - Habitable vs Inhabitable

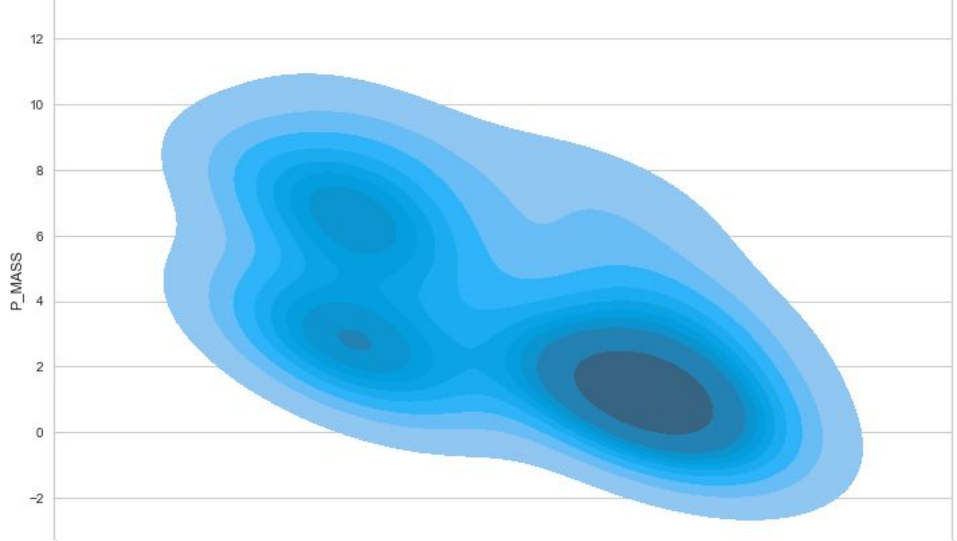
- PCA

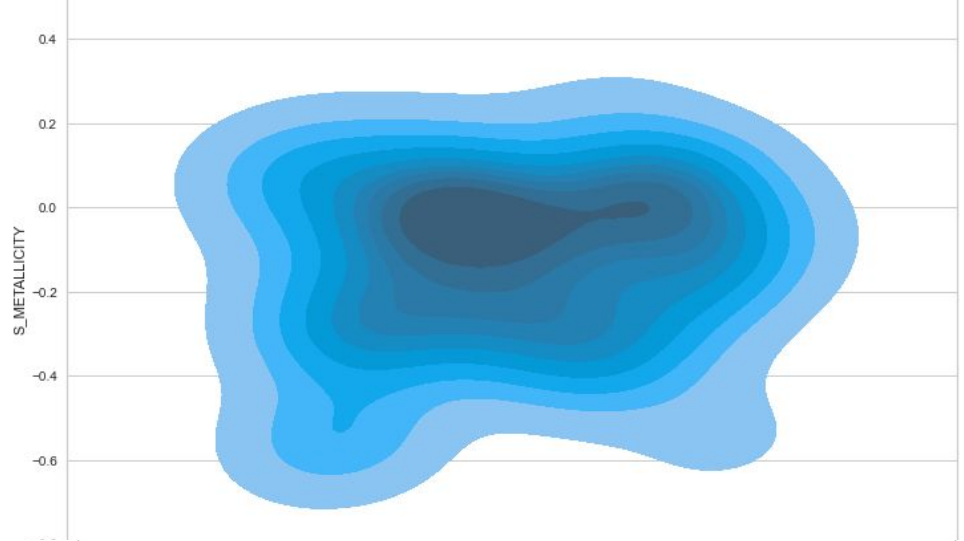

PCA plot with the first 2 components (ca 19 % variance)
P_HABITABLE
0
1
pc2
pc1
30
20
10
0
−10
−5
0
5
10
15
20
25

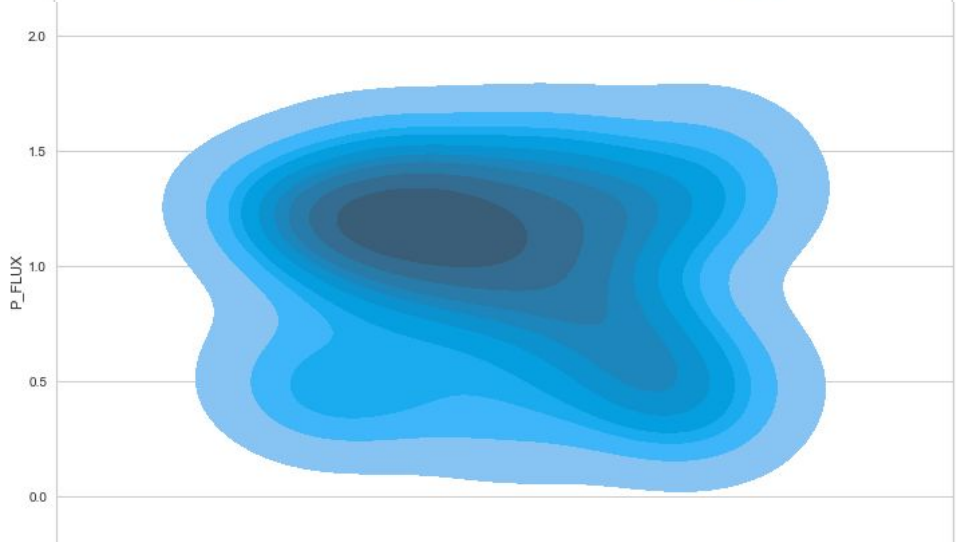

PCA plot with the first 3 components (ca 27 % variance)
30
20
10
0
−10
pc3
30
20
10
0
−10
pc2
−5
0
5
10
15
20
25
pc1

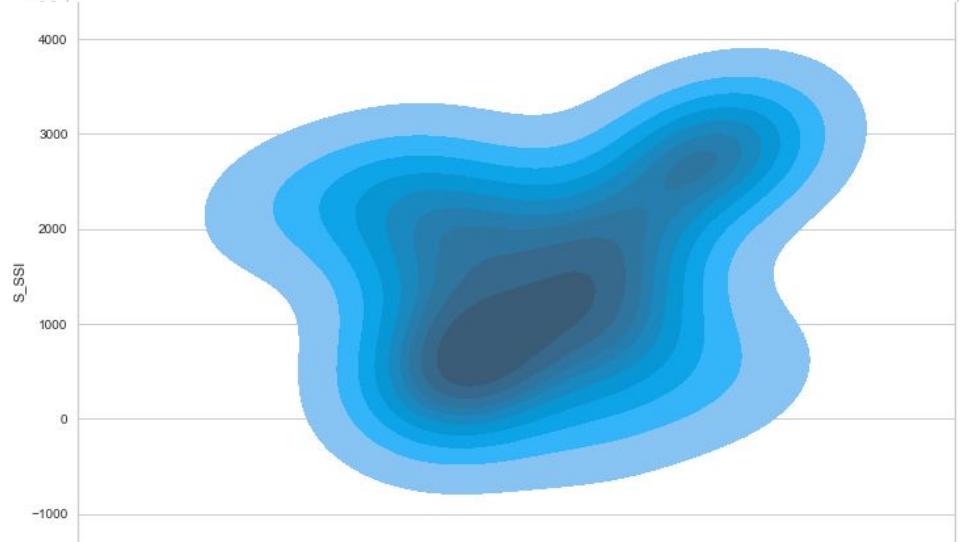
The process: EDA - Habitable vs Inhabitable

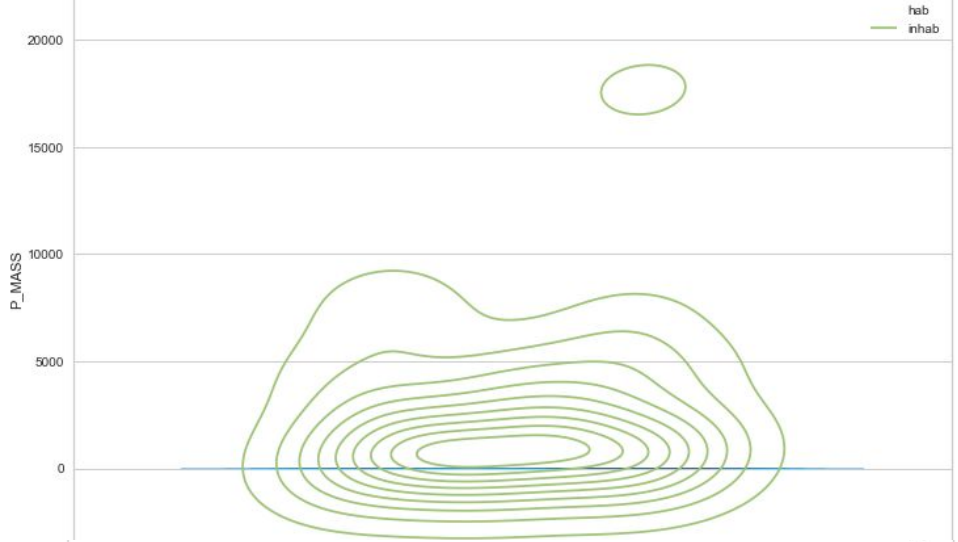
- PCA
- What specifically makes the hab/inhab planets differ?

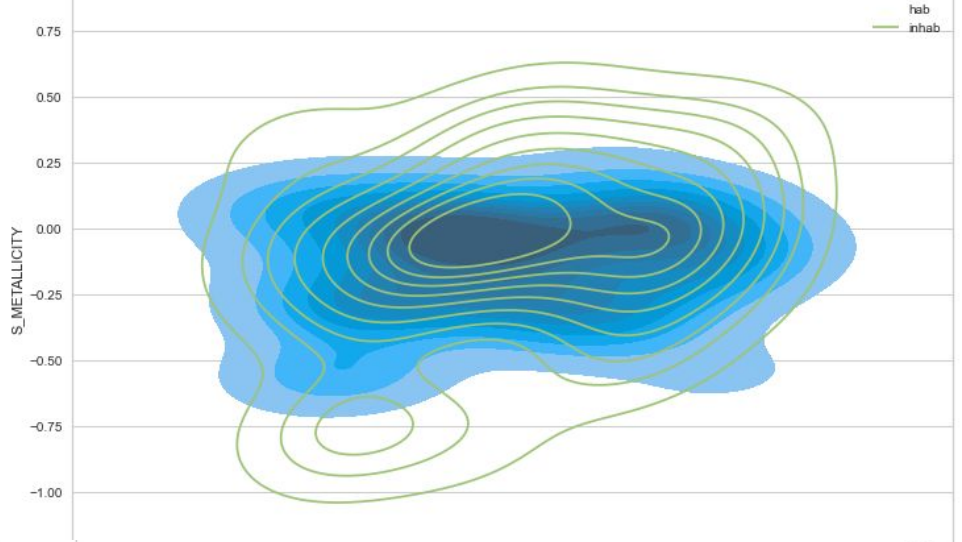

P_MASS
12
10
8
6
4
2
0
−2

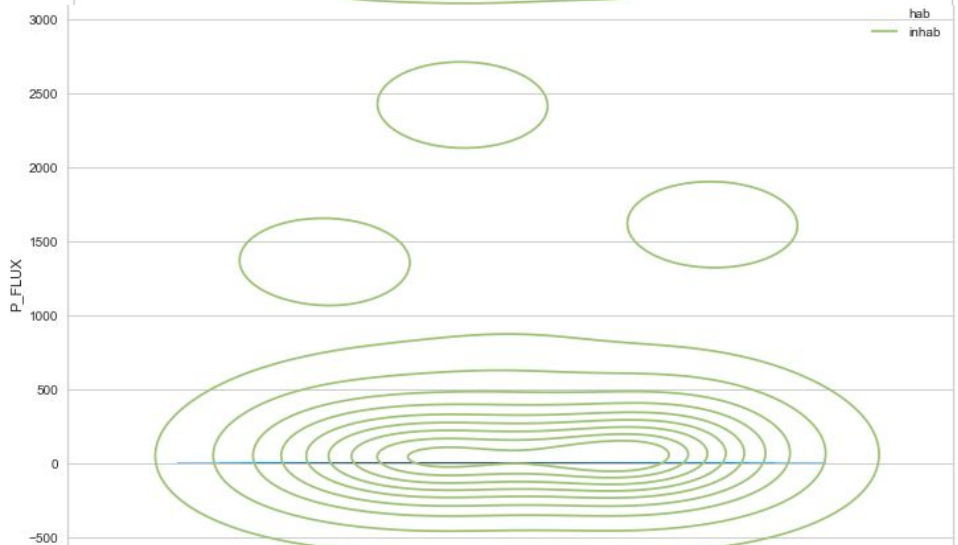

S_METALLICITY
0.4
0.2
0.0
−0.2
−0.4
−0.6

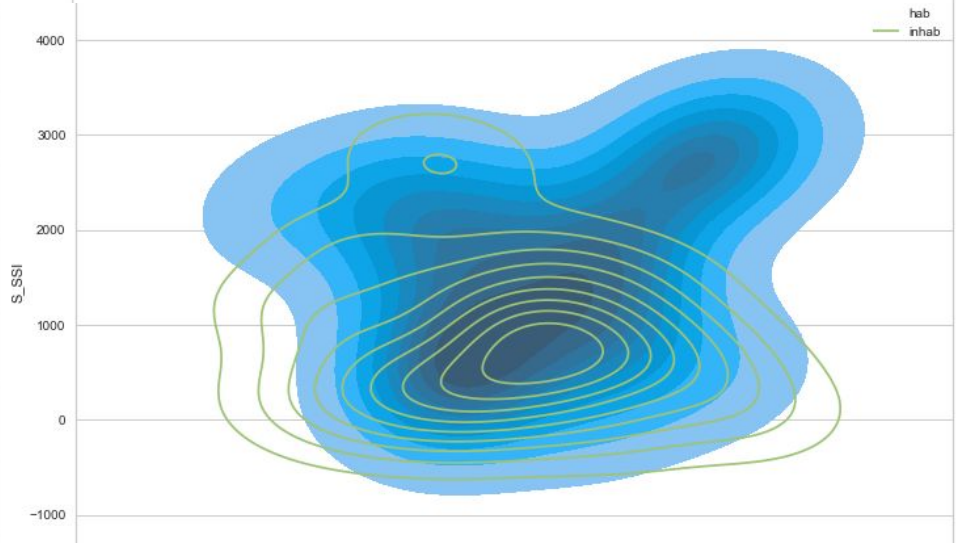

P_FLUX
2.0
1.5
1.0
0.5
0.0


S_SSI
4000
3000
2000
1000
0
−1000


hab
inhab
P_MASS
20000
15000
10000
5000
0


hab
inhab
S_METALLICITY
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


hab
inhab
P_FLUX
3000
2500
2000
1500
1000
500
0
−500

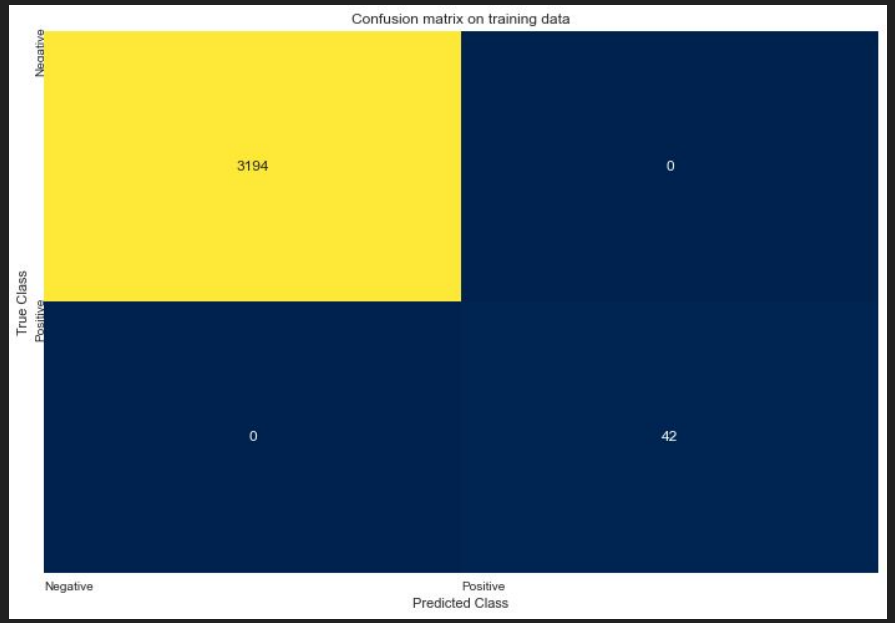

hab
inhab
S_SSI
4000
3000
2000
1000
0
−1000

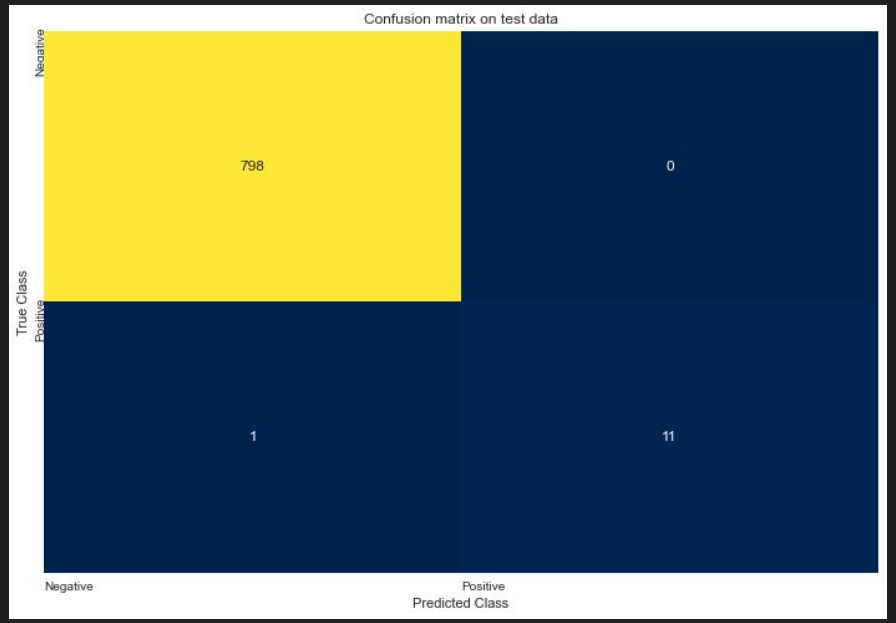
The process: EDA - Habitable vs Inhabitable

- PCA
- What specifically makes the hab/inhab planets differ?
  - Answer: hab are more earth-like

ML model: target var = P_HABITABLE

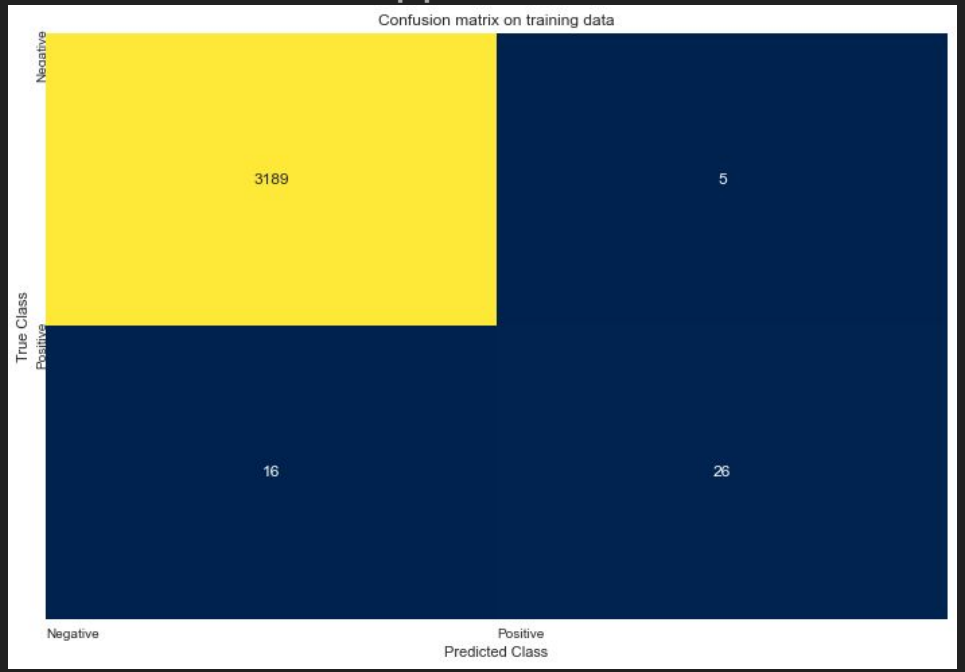
- Estimator: LogisticalRegression
- Scorer: kappa





ML model: target var = P_HABITABLE (only star data)

- Estimator: RandomForestClassifier
- Scorer: kappa

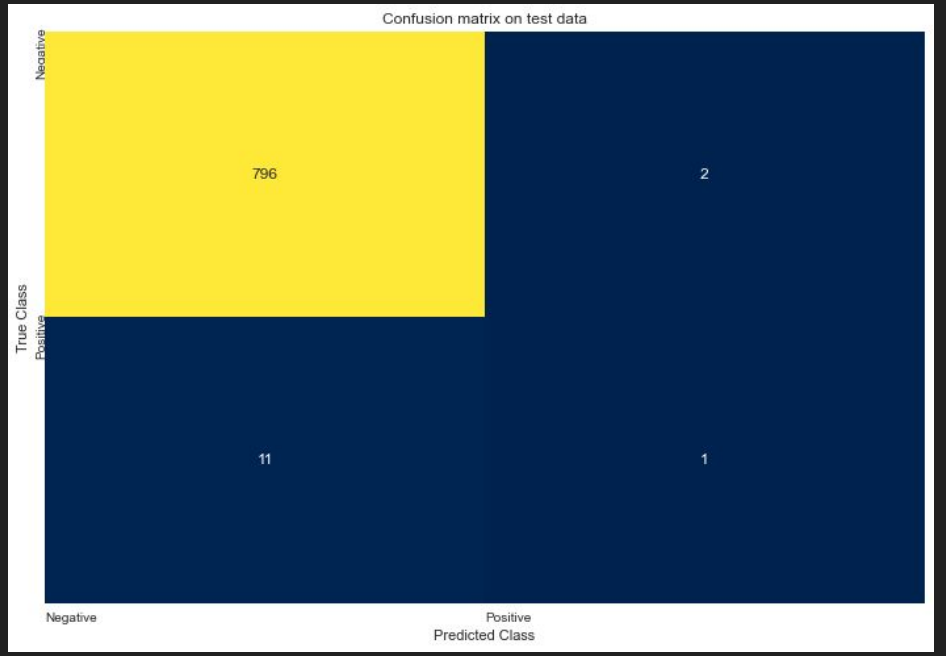


Precision: 84 %

ML model: target var = P_HABITABLE (only star data)

- Estimator: RandomForestClassifier
- Scorer: kappa

Precision: 33 %



Summarizing thoughts

Summarizing thoughts

- A model such as this might not be as inapplicable as one might think

Summarizing thoughts

- A model such as this might not be as inapplicable as one might think
- Reducing complexity is important

Summarizing thoughts

- A model such as this might not be as inapplicable as one might think
- Reducing complexity is important
- Problem with feature creation but it was useful

Summarizing thoughts

- A model such as this might not be as inapplicable as one might think
- Reducing complexity is important
- Problem with feature creation but it was useful
- Try resampling next time

The End