CS 591: Data Science | Final Project

Data Overview:

planets.csv - http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets

habitable.csv - http://phl.upr.edu/projects/habitable-exoplanets-catalog/data

Planets Dataframe:

Size: 25 columns/attributes
1830 rows/confirmed planets

This is a dataframe of NASA's current list of confirmed exoplanets – or planets found outside of our own Solar System. There are 1830 rows in this dataframe, meaning that there are 1830 confirmed exoplanets as of the moment I downloaded the table. I have chosen to use 25 of the many attributes/measurements of the planets.

Missing Values:
There are many missing values for the columns:

- pl_mass_jupiter
- pl_rad_jupiter
- pl_density
- pl_orbincl
- pl_rad_earth
- pl_eq_temp
- pl_earthMass
- pl_rad_solar
- pl_transit_duration
- r_dist_sRad
- r_pRad_sRad

This is not surprising because of our limited technology of detecting exoplanets. As such, many times we are merely able to detect the physical presence of the planet but are unable to detect specific details or calculations.

I will deal with this missing data by excluding planets as needed when comparing this dataset to the Habitable dataset.

Irregularity:
Start KOI-351 had 7 confirmed exoplanets around it.

```
average of  pl_in_system          =  2.15081967213
average of  moon_in_sy            =  0.0
average of  pl_orbper             =  1068.58277608
average of  pl_orbsmax            =  7.16752578251
average of  orb_per_days          =  0.0810318415301
average of  pl_mass_jupiter       =  0.763666442623
average of  pl_rad_jupiter        =  0.280771584699
average of  pl_density            =  0.45768147541
average of  pl_orbincl            =  18.3777688525
average of  pl_rad_earth          =  3.14513879781
average of  pl_eq_temp            =  160.430601093
average of  pl_earthMass          =  242.139624836
average of  pl_rad_solar          =  0.028849726776
average of  pl_transit_duration   =  0.0905092786885
average of  r_dist_sRad           =  4.07721289344
average of  r_pRad_sRad           =  0.0110754480874
average of  star_dist             =  185.674333333
average of  st_eff_temp_K         =  4679.51114754
average of  st_mass               =  0.716513661202
average of  st_rad                =  1.31699453552
average of  discovery             =  2011.37540984

(min, max) of  pl_in_system       = ( 1 ,  7 )
(min, max) of  pl_orbper          = ( 0.09070629 ,  730000.0 )
(min, max) of  pl_orbsmax         = ( 0.0044 ,  2500.0 )
(min, max) of  orb_per_days       = ( 0.002 ,  0.9332 )
(min, max) of  pl_mass_jupiter    = ( 7e-05 ,  28.5 )
(min, max) of  pl_rad_jupiter     = ( 0.027 ,  3.0 )
(min, max) of  pl_density         = ( 0.03 ,  77.7 )
(min, max) of  pl_orbincl         = ( 17.6 ,  90.76 )
(min, max) of  pl_rad_earth       = ( 0.303 ,  33.6 )
(min, max) of  pl_eq_temp         = ( 50.0 ,  3320.0 )
(min, max) of  pl_earthMass       = ( 0.008 ,  9057.77 )
(min, max) of  pl_rad_solar       = ( 0.003 ,  0.308 )
(min, max) of  pl_transit_duration = ( 0.0145 ,  0.7472 )
(min, max) of  r_dist_sRad        = ( 2.253 ,  346.0 )
(min, max) of  r_pRad_sRad        = ( 0.0036 ,  0.584 )
(min, max) of  star_dist          = ( 1.35 ,  8500.0 )
(min, max) of  st_eff_temp_K      = ( 575.0 ,  57000.0 )
(min, max) of  st_mass            = ( 0.02 ,  3.09 )
(min, max) of  st_rad             = ( 0.04 ,  51.1 )
(min, max) of  discovery          = ( 1989 ,  2015 )
```

Habitable Dataframe:

Size: 9 columns/attributes
19 rows/confirmed habitable planets

This is a dataframe of the confirmed habitable exoplanets that have been found. This means that the 19 planets in this list have conditions prime enough to harbor and sustain some sort of life (or of life as we know it). This table came with 9 columns including one named "ESI." ESI stands for Earth Similarity Index which is a percentage of how similar the planet is to Earth based on stellar flux and size.

Missing Data:

Some planets on this list were not in the Planets dataset. This is because they were just added and not confirmed to be official exoplanets yet. I have chosen to delete these from the dataset I will officially use.

Irregularities:

No irregularities since these planets had to already have consistent measurements with respect to Earth and themselves to be considered habitable.

```
Habitable Average ========================
Mass (Earth)               =  8.43157894737
Radius (Earth)             =  1.76842105263
Equilibrium temperature (K)=  242.210526316
Period                     =  112.973684211
Distance (light years)     =  926.842105263

(min,max) of  Mass_Me      = ( 1.3 ,  26.8 )
(min,max) of  Radius_Re    = ( 1.1 ,  2.5 )
(min,max) of  Flux_Fe      = ( 0.29 ,  1.43 )
(min,max) of  T_K          = ( 188 ,  277 )
(min,max) of  Period_days  = ( 25.6 ,  289.9 )
(min,max) of  Distance_ly  = ( 13 ,  2540 )
(min,max) of  ESI          = ( 0.6 ,  0.88 )
```

Columns of Planet DataFrame:

planet_name	name of exoplanet
hostStar	name of star exoplanet orbits
pl_letter	planet letter of its star system
pl_dismethod	method planet was discovered by
pl_in_system	number of planets in its star system
moon_in_sy	number of moons in its star system
pl_orbper	time the planet takes to make a complete orbit
pl_orbsmax	longest diameter of an elliptic orbit
orb_per_days	fraction of the orbit completed per day
pl_mass_jupiter	mass of planet in terms of Jupiter's mass
pl_rad_jupiter	radius of planet in terms of Jupiter's radius
pl_density	density of planet
pl_orbincl	orbital inclination of planet
pl_rad_earth	radius of planet in terms of Earth's radius
pl_eq_temp	equilibrium temperature of planet (Kelvin)
pl_earthMass	mass of planet in terms of Earth's mass
pl_rad_solar	radius of
pl_transit_duration	transit duration in days
r_dist_sRad	ratio of the distance between star and planet to star radius
r_pRad_sRad	ratio of the planet's radius to the star's radius
star_dist	distance to star system in parsecs
st_eff_temp_K	star's effective temperature (Kelvin)
st_mass	star's mass in terms of the sun's mass
st_rad	star's radius in terms of the sun's radius
discovery	year exoplanet was discovered

Columns of Habitable DataFrame:

Name	name of planet
Type	type/classification of planet
Mass_Me	mass of plannet in terms of Earth's mass
Radius_Re	radius of planet in terms of Earth's radius
Flux_Fe	flux of planet in solar units
T_K	equilibrium temperature of planet (Kelvin)
Period_days	period of planet's orbit (days)
Distance_ly	distance to planet's system (light years)
ESI	Earth Similarity Index

Hypotheses:

1. A very small percentage of confirmed exoplanets fall even remotely close to measurements of exoplanets confirmed to be habitable.
2. There is at least one outlier star in the planets dataframe that should not have the attributes to host an exoplanet but in fact does.
3. The distance to a star (and thus time within the universe) makes a difference on how many exoplanets are able to form around it.

When finding the range of habitability, I will cross reference the planets from the Habitable dataframe and the Planets dataframe to find more information on those planets. I will also compare their measurements and averages to the other confirmed exoplanets. I will use linear and logistic regression to find a prediction on the various attributes and their habitability. I intend to find the intensity of how different measurements are related to each other.