Cosmos (Or How I Learned to Love the Night Sky)

Space. The final frontier. These are the stars that top scientific minds have concluded that exoplanets are orbiting around them. Using this database, students, teachers, amateur star gazers, stoners who watch too much Science Channel, and parents who want to get their children excited about the great unknown can explore data and prepare themselves to take that one small step for a man, one giant leap for mankind.

Over the last decade or two, scientists have gone from theorizing that exoplanets - that is, planets that orbit suns not our own - exist, to theorizing that not only do they exist, but it's likely that every star in the sky has planets orbiting them. The methods they use to locate and identify these strange new worlds vary, from observing dips in the star's light as the body crosses in front of its parent star to measuring the gravitational impact that an orbiting body has on it's parent star, but the end result is the same: a long list of new, unexplored worlds to dazzle and inspire the imaginations of generations to come.

There is just one issue, and that is the availability of the data specifically for an audience that is looking for an introduction to extrasolar solar systems and space exploration. For our project, Group Cosmic Degenerates decided to step in and fulfill this need.

Before we begin, we did want to acknowledge one major limitation that we encountered at the outset of our project and it's one that may surprise you. From the priests of Ancient Egypt to the Oracles at Delphi to the Druids of Celtic Britain, people have been cataloguing and observing the locations and movements of the celestial bodies. There is literally thousands of years of data from civilizations all across the globe to parse through associated with the stars in the sky. There is, undoubtedly, a plethora of information we could have included given more time to thoroughly explore the data and study some of the more technical aspects that scientists have developed to measure stellar properties, however we feel that this is a great starter database for cadets of StarFleet.

Without further ado, our ETL Project: "Cosmos: Or How I Learned to Love The Sky".

Extract

There are numerous sources out there for information regarding both exoplanets and the stars in the sky. We decided to use NASA's Exoplanet API (https://exoplanetarchive.ipac.caltech.edu/docs/data.html), which describes what stars astronomers have focused on in their search for extrasolar worlds and gives a lot of detailed information about the star including position in the night sky, luminosity, and star type. From Kaggle, we were able to find Kepler data regarding exoplanets and their properties. The Kepler data is as surefire source as the NASA data, as the Kepler mission's primary goal is to discover and ascertain information about exoplanets. The data was available in many types including csv, json, and others. We went with csv files.

21	# COLUMN st_disterr1: Distance Upper Unc. [pc]												
22	# COLUMN st_disterr2: Distance Lower Unc. [pc]												
23	# COLUMN st_vmag: V-band [mag]												
24	# COLUMN st_vmagerr: V-band Unc. [mag]												
25	# COLUMN st_bmv: B-V [mag]												
26	# COLUMN st_bmverr: B-V Unc. [mag]												
27	# COLUMN st_spttype: Spectral Type												
28	# COLUMN st_lbol: Stellar Bolometric Luminosity [Solar]												
29	# COLUMN st_lbolerr: Stellar Bolometric Luminosity Unc. [Solar]												
30	#												
31	star_name	hip_name	hd_name	gj_name	tm_name	st_exocatflag	st_coronagflag	st_starshadeflag	st_wfirstflag	st_lbtiflag	st_rvflag	st_ppnum	rastr
32	HIP 171	HIP 171	HD 224930	GJ 914 A	2MASS J00021014+	1	0	1	0	0	0	0	00h02m0
33	HD 142	HIP 522	HD 142	GJ 4.2 A	2MASS J00061921-	1	1	0	0	0	1	2	00h06m1
34	HIP 544	HIP 544	HD 166	GJ 5	2MASS J00063674+	1	0	1	0	0	1	0	00h06m3
35	HIP 677	HIP 677	HD 358		2MASS J00082326+	1	1	0	0	0	0	0	00h08m2
36	HIP 746	HIP 746	HD 432	GJ 8	2MASS J00091071+	1	1	0	0	0	0	0	00h09m1
37	HIP 910	HIP 910	HD 693	GJ 10	2MASS J00111587-	1	1	0	0	1	0	0	00h11m1
38	HIP 1599	HIP 1599	HD 1581	GJ 17	2MASS J00200446-	1	1	1	0	0	0	0	00h20m0
39	HIP 2021	HIP 2021	HD 2151	GJ 19	2MASS J00254416-	1	1	1	0	0	0	0	00h25m3
40	HIP 2072	HIP 2072	HD 2262	GJ 20	2MASS J00261219-	1	1	0	0	0	0	0	00h26m1
41	HIP 2081	HIP 2081	HD 2261	GJ 9013	2MASS J00261699-	1	1	0	0	0	0	0	00h26m1

Transform

Once we had our datasets, we began the process of getting the data ready for queries. This involved removing rows with null or incomplete values, filtering out columns that we felt didn't add anything meaningful to the conversation, and determining keys that would be used to relate our cleaned exoplanet table to our cleaned star table. We also changed the datatypes of the values in columns so our target users will be able perform aggregate queries when necessary.

In [1]:
```
import pandas as pd
```

In [2]:
```
place = pd.read_csv('Resources/kepler.csv/kepler.csv')
```

In [3]:
```
place.head()
```

Out[3]:

	# name	planet_status	mass	mass_error_min	mass_error_max	mass_sini	mass_sini_error_min	mass_sini_error_max	radius	radius_error_min	...	sta
0	OGLE-2016-BLG-1469L b	Confirmed	13.60	3.00	3.00	NaN	NaN	NaN	NaN	NaN	...	
1	11 Com b	Confirmed	19.40	1.50	1.50	19.40	1.50	1.50	NaN	NaN	...	
2	11 Oph b	Confirmed	21.00	3.00	3.00	NaN	NaN	NaN	NaN	NaN	...	
3	11 UMi b	Confirmed	10.50	2.47	2.47	10.50	2.47	2.47	NaN	NaN	...	
4	14 And b	Confirmed	5.33	0.57	0.57	5.33	0.57	0.57	NaN	NaN	...	

5 rows × 98 columns

In [8]:
```
exoplanet = place[['# name','planet_status','mass','orbital_period','semi_major_axis','eccentricity','discovered','detection_
```

In [20]:
```
exoplanet_df = exoplanet.dropna(axis='rows').reset_index()
```

Load

We had our sources from which we collected our raw data, we transformed that data into usable things, now it was time to create our database. We opted to use SQL for a few reasons. First and foremost, it is - as far as we know, at least - unparalleled when it comes to ease of use and performing relational data queries. Using our database “Cosmos”, students, teachers, and anyone with an inkling to explore the vast poisonous radioactive wasteland that is outer space and look for diamonds in that rough. For more information, the included ReadMe file has a more detailed breakdown of the database. Below are a few example queries that students can run while exploring the database. One provides the name of the exoplanet along with when it was discovered and it’s sun’s stellar type. The other is the average mass of all discovered exoplanets after 1996.

```sql
SELECT exoplanets.planet_name, exoplanets.discovered, exoplanets.star_name, host_stars.class
FROM exoplanets
JOIN host_stars ON exoplanets.star_name=host_stars.star_name;

SELECT AVG(mass) as "Average Mass"
FROM exoplanets;

SELECT sum(mass) as "Total Mass for Planets Found after 1996"
FROM exoplanets
WHERE discovered >= 1996;

SELECT AVG(exoplanets.orbital_period) as "Average Orbital Period"
FROM exoplanets
WHERE exoplanets.detection_type = 'Radial Velocity';

```

	planet_name character varying	discovered numeric	star_name character varying	class character varying
1	14 Her b	2002	14 Her	K0V
2	51 Peg b	1995	51 Peg	G5V
3	55 Cnc b	1996	55 Cnc	G8V
4	55 Cnc c	2002	55 Cnc	G8V
5	55 Cnc d	2002	55 Cnc	G8V
6	55 Cnc e	2004	55 Cnc	G8V
7	55 Cnc f	2007	55 Cnc	G8V

	Total Mass for Planets Found after 1996 numeric
1	4052.797856644

Finitione

There is little to no doubt that mankind’s trajectory into the future lies above us. Using the Cosmos database we created, anyone from young students and teachers to old people near death will be able to query aspects of both exoplanets and the suns that rise in their sky. The only question left is “Where to start?”, to which our answer is “Second star on the right, and straight on ‘til morning.”