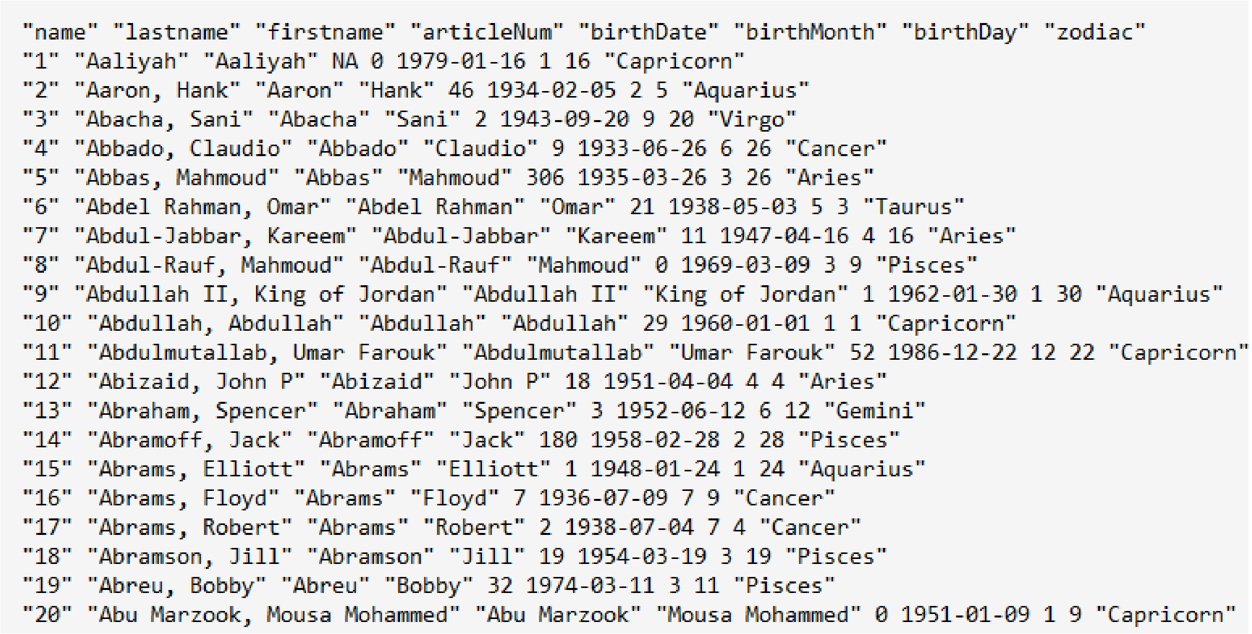
Data Dictionary

Datasets

Famous_birthdates.txt

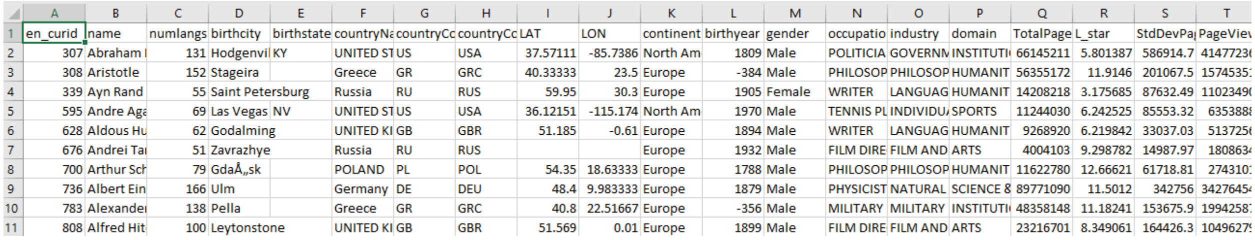
This dataset was published by github user Richard512 and was found at https://github.com/richard512/Little-Big-Data/blob/master/famous-birthdates.csv. This dataset contains over 4700 rows.



```
"name" "lastname" "firstname" "articleNum" "birthDate" "birthMonth" "birthDay" "zodiac"
"1" "Aaliyah" "Aaliyah" NA 0 1979-01-16 1 16 "Capricorn"
"2" "Aaron, Hank" "Aaron" "Hank" 46 1934-02-05 2 5 "Aquarius"
"3" "Abacha, Sani" "Abacha" "Sani" 2 1943-09-20 9 20 "Virgo"
"4" "Abbado, Claudio" "Abbado" "Claudio" 9 1933-06-26 6 26 "Cancer"
"5" "Abbas, Mahmoud" "Abbas" "Mahmoud" 306 1935-03-26 3 26 "Aries"
"6" "Abdel Rahman, Omar" "Abdel Rahman" "Omar" 21 1938-05-03 5 3 "Taurus"
"7" "Abdul-Jabbar, Kareem" "Abdul-Jabbar" "Kareem" 11 1947-04-16 4 16 "Aries"
"8" "Abdul-Rauf, Mahmoud" "Abdul-Rauf" "Mahmoud" 0 1969-03-09 3 9 "Pisces"
"9" "Abdullah II, King of Jordan" "Abdullah II" "King of Jordan" 1 1962-01-30 1 30 "Aquarius"
"10" "Abdullah, Abdullah" "Abdullah" "Abdullah" 29 1960-01-01 1 1 "Capricorn"
"11" "Abdulmutallab, Umar Farouk" "Abdulmutallab" "Umar Farouk" 52 1986-12-22 12 22 "Capricorn"
"12" "Abizaid, John P" "Abizaid" "John P" 18 1951-04-04 4 4 "Aries"
"13" "Abraham, Spencer" "Abraham" "Spencer" 3 1952-06-12 6 12 "Gemini"
"14" "Abramoff, Jack" "Abramoff" "Jack" 180 1958-02-28 2 28 "Pisces"
"15" "Abrams, Elliott" "Abrams" "Elliott" 1 1948-01-24 1 24 "Aquarius"
"16" "Abrams, Floyd" "Abrams" "Floyd" 7 1936-07-09 7 9 "Cancer"
"17" "Abrams, Robert" "Abrams" "Robert" 2 1938-07-04 7 4 "Cancer"
"18" "Abramson, Jill" "Abramson" "Jill" 19 1954-03-19 3 19 "Pisces"
"19" "Abreu, Bobby" "Abreu" "Bobby" 32 1974-03-11 3 11 "Pisces"
"20" "Abu Marzook, Mousa Mohammed" "Abu Marzook" "Mousa Mohammed" 0 1951-01-09 1 9 "Capricorn"
```

Pantheon_People.csv

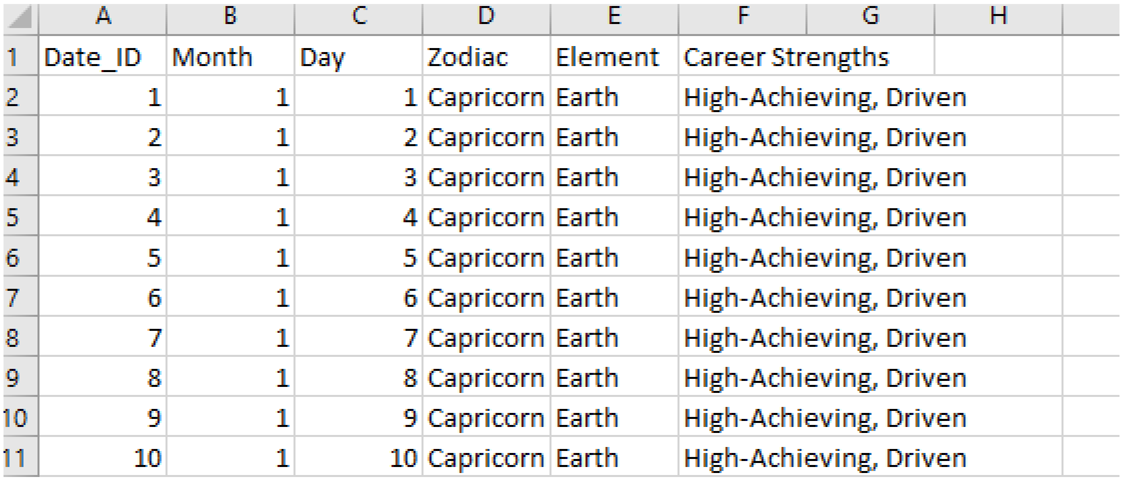
The Pantheon dataset can be found here (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28201) and is described as “A Manually Verified Dataset of Globally Famous Biographies”. Per the authors, the 11,000+ individuals were chosen based on extensive numbers of wikipedia views across several languages, and “is enriched with: (i) manually verified demographic information (place and date of birth, gender) (ii) a taxonomy of occupations classifying each biography at three levels of aggregation”.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	en_curid	name	numlangs	birthcity	birthstate	countryNa	countryCo	countryCo	LAT	LON	continent	birthyear	gender	occupatio	industry	domain	TotalPage	L_star	StdDevPa	PageViev
2	307	Abraham	131	Hodgenvil	KY	UNITED ST	US	USA	37.57111	-85.7386	North Am	1809	Male	POLITICIA	GOVERNM	INSTITUTI	66145211	5.801387	586914.7	4147723
3	308	Aristotle	152	Stageira		Greece	GR	GRC	40.33333	23.5	Europe	-384	Male	PHILOSOP	PHILOSOP	HUMANIT	56355172	11.9146	201067.5	1574535
4	339	Ayn Rand	55	Saint Petersburg		Russia	RU	RUS	59.95	30.3	Europe	1905	Female	WRITER	LANGUAG	HUMANIT	14208218	3.175685	87632.49	1102349
5	595	Andre Aga	69	Las Vegas	NV	UNITED ST	US	USA	36.12151	-115.174	North Am	1970	Male	TENNIS PL	INDIVIDU	SPORTS	11244030	6.242525	85553.32	635388
6	628	Aldous Hu	62	Godalming		UNITED KI	GB	GBR	51.185	-0.61	Europe	1894	Male	WRITER	LANGUAG	HUMANIT	9268920	6.219842	33037.03	513725
7	676	Andrei Ta	51	Zavrazhye		Russia	RU	RUS			Europe	1932	Male	FILM DIRE	FILM AND	ARTS	4004103	9.298782	14987.97	180863
8	700	Arthur Sch	79	GdaÅ„sk		POLAND	PL	POL	54.35	18.63333	Europe	1788	Male	PHILOSOP	PHILOSOP	HUMANIT	11622780	12.66621	61718.81	274310
9	736	Albert Ein	166	Ulm		Germany	DE	DEU	48.4	9.983333	Europe	1879	Male	PHYSICIST	NATURAL	SCIENCE &	89771090	11.5012	342756	3427645
10	783	Alexander	138	Pella		Greece	GR	GRC	40.8	22.51667	Europe	-356	Male	MILITARY	MILITARY	INSTITUTI	48358148	11.18241	153675.9	1994258
11	808	Alfred Hit	100	Leytonstone		UNITED KI	GB	GBR	51.569	0.01	Europe	1899	Male	FILM DIRE	FILM AND	ARTS	23216701	8.349061	164426.3	1049627

Zodiac.csv

I created this csv myself from a combination of excel formulas and readily-available internet data. Career strengths, though I ultimately did not use it, came from (https://www.rd.com/list/career-strength-according-to-zodiac-sign/)

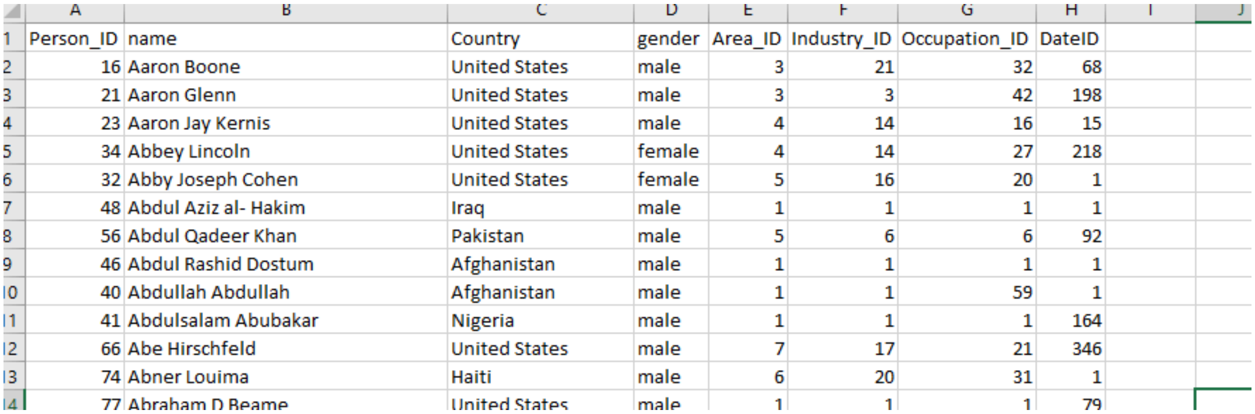


	A	B	C	D	E	F	G	H
1	Date_ID	Month	Day	Zodiac	Element	Career Strengths		
2	1	1	1	Capricorn	Earth	High-Achieving, Driven		
3	2	1	2	Capricorn	Earth	High-Achieving, Driven		
4	3	1	3	Capricorn	Earth	High-Achieving, Driven		
5	4	1	4	Capricorn	Earth	High-Achieving, Driven		
6	5	1	5	Capricorn	Earth	High-Achieving, Driven		
7	6	1	6	Capricorn	Earth	High-Achieving, Driven		
8	7	1	7	Capricorn	Earth	High-Achieving, Driven		
9	8	1	8	Capricorn	Earth	High-Achieving, Driven		
10	9	1	9	Capricorn	Earth	High-Achieving, Driven		
11	10	1	10	Capricorn	Earth	High-Achieving, Driven		

Blank_Pantheon.csv

This is a supplemental csv I created for this project. Once my tables were joined, there were only around 1200 overlapping people, and I wanted to increase the size of my dataset. This represents the people in the Famous_birthdate file who did not have a corresponding entry in the Pantheon dataset.

I provided chatGPT with the names and asked that it return their career. I then backed into the ID’s from there. This data was then compiled with the other datasets to form the People table of over 10,000 rows. Upon spot-checking the responses, they appeared to be reasonably correct, but could potentially contain errors.



	A	B	C	D	E	F	G	H
1	Person_ID	name	Country	gender	Area_ID	Industry_ID	Occupation_ID	DateID
2	16	Aaron Boone	United States	male	3	21	32	68
3	21	Aaron Glenn	United States	male	3	3	42	198
4	23	Aaron Jay Kernis	United States	male	4	14	16	15
5	34	Abbey Lincoln	United States	female	4	14	27	218
6	32	Abby Joseph Cohen	United States	female	5	16	20	1
7	48	Abdul Aziz al- Hakim	Iraq	male	1	1	1	1
8	56	Abdul Qadeer Khan	Pakistan	male	5	6	6	92
9	46	Abdul Rashid Dostum	Afghanistan	male	1	1	1	1
10	40	Abdullah Abdullah	Afghanistan	male	1	1	59	1
11	41	Abdulsalam Abubakar	Nigeria	male	1	1	1	164
12	66	Abe Hirschfeld	United States	male	7	17	21	346
13	74	Abner Louima	Haiti	male	6	20	31	1
14	77	Abraham D Beame	United States	male	1	1	1	79

Missingbirthdate.csv

This is also a supplemental csv I created to supplement my data. I provided chatGPT with the names and birthdate. I then backed into the day_ ID's from there. This data was then compiled with the other datasets to form the People table of over 10,000 rows. Upon spot-checking the responses, they appeared to be reasonably correct, but could potentially contain errors.

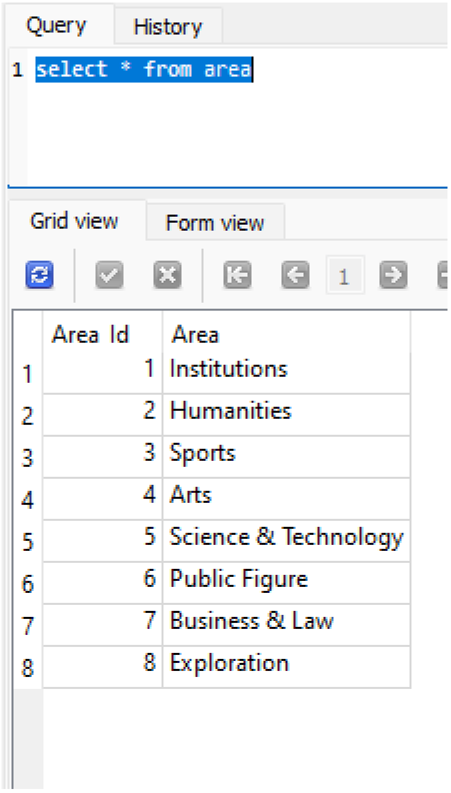
Tables

Area Table

This table represents the unique domain values from the Pantheon People dataset. I changed the name from Domain to Area to avoid conflicts with the Domain keyword in python. Industries from the Industry table all roll into a designated area. See Relationships between Area, Industry, and Occupation for more information.

Each row represents a distinct area of expertise.

Column Name	**Data Type**	**Explanation**
Area_Id	integer	Primary Key; assigned index value to join with other tables.
Area	string	The name of the area.



```
select * from area
```

	Area Id	Area
1	1	Institutions
2	2	Humanities
3	3	Sports
4	4	Arts
5	5	Science & Technology
6	6	Public Figure
7	7	Business & Law
8	8	Exploration

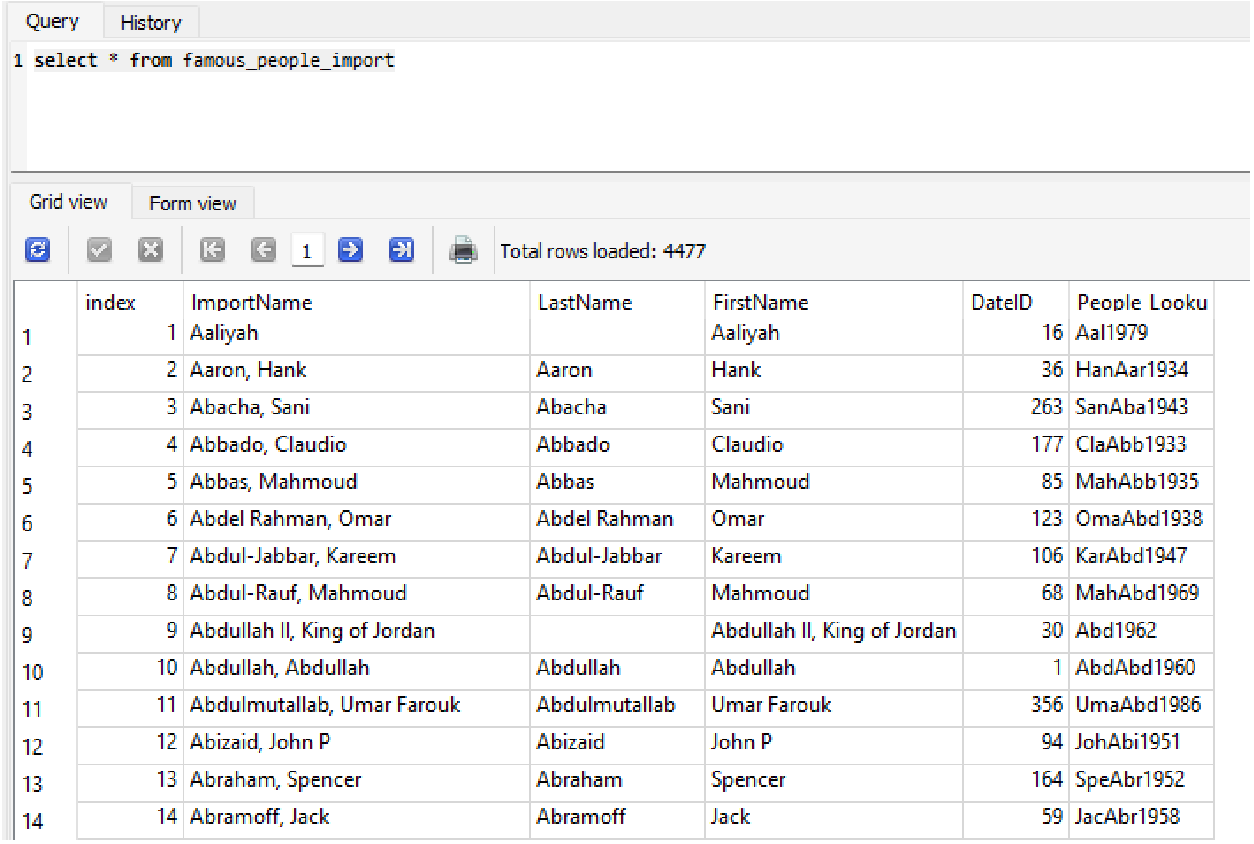
Famous_People_Import Table

This table is derived from the text file data/famous_birthdates.txt found here: https://github.com/richard512/Little-Big-Data/blob/master/famous-birthdates.csv .

Each row represents one person and their name, a value representing their birthdate, and a key I designed to match names across my different datasets.

Column Name	**Data Type**	**Explanation**
Index	int	Index of the dataframe
ImportName	string	The full name as provided in the original data file.
LastName	string	The last name as provided in the original data file.
FirstName	string	The first name as provided in the original data file
DateID	int	Foreign key; represents the date of the year on which the person's birthdate falls

People_Lookup	string	Joinable field created from the first 3 characters of the person's first name, first 3 characters of the person’s last name, and the year of their birth.



```
select * from famous_people_import
```

Total rows loaded: 4477

	index	ImportName	LastName	FirstName	DateID	People Looku
1	1	Aaliyah		Aaliyah	16	Aal1979
2	2	Aaron, Hank	Aaron	Hank	36	HanAar1934
3	3	Abacha, Sani	Abacha	Sani	263	SanAba1943
4	4	Abbado, Claudio	Abbado	Claudio	177	ClaAbb1933
5	5	Abbas, Mahmoud	Abbas	Mahmoud	85	MahAbb1935
6	6	Abdel Rahman, Omar	Abdel Rahman	Omar	123	OmaAbd1938
7	7	Abdul-Jabbar, Kareem	Abdul-Jabbar	Kareem	106	KarAbd1947
8	8	Abdul-Rauf, Mahmoud	Abdul-Rauf	Mahmoud	68	MahAbd1969
9	9	Abdullah II, King of Jordan		Abdullah II, King of Jordan	30	Abd1962
10	10	Abdullah, Abdullah	Abdullah	Abdullah	1	AbdAbd1960
11	11	Abdulmutallab, Umar Farouk	Abdulmutallab	Umar Farouk	356	UmaAbd1986
12	12	Abizaid, John P	Abizaid	John P	94	JohAbi1951
13	13	Abraham, Spencer	Abraham	Spencer	164	SpeAbr1952
14	14	Abramoff, Jack	Abramoff	Jack	59	JacAbr1958

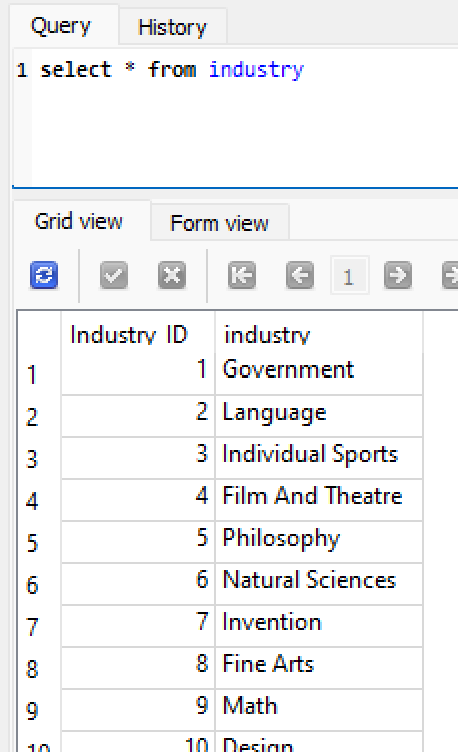
Industry Table

This table represents the unique industry values from the Pantheon People dataset. Occupations from the occupation table all roll into a designated industry, which rolls into a designated area. See Relationships between Area, Industry, and Occupation for more information.

Each row represents a distinct industry..

Column Name	**Data Type**	**Explanation**
Industry_Id	integer	Primary Key; assigned index

		value to join with other tables.
Industry	string	The name of the industry



```
select * from industry
```

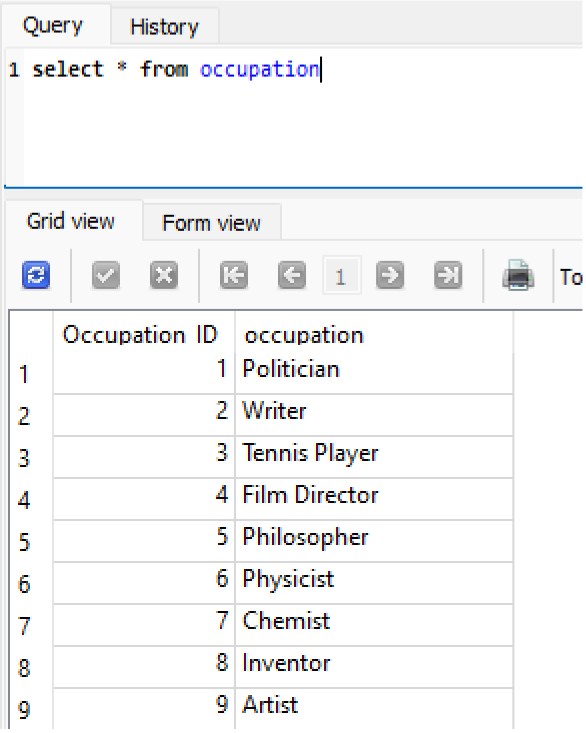
	Industry ID	industry
1	1	Government
2	2	Language
3	3	Individual Sports
4	4	Film And Theatre
5	5	Philosophy
6	6	Natural Sciences
7	7	Invention
8	8	Fine Arts
9	9	Math
10	10	Design

Occupation Table

This table represents the unique occupation values from the Pantheon People dataset. Occupations roll into a designated industry, which rolls into a designated area. See Relationships between Area, Industry, and Occupation for more information.

Each row represents a distinct occupation.

Column Name	**Data Type**	**Explanation**
Occupation_ID	integer	Primary Key; assigned index value to join with other tables.
Occupation	string	The name of the occupation



```sql
select * from occupation
```

	Occupation ID	occupation
1	1	Politician
2	2	Writer
3	3	Tennis Player
4	4	Film Director
5	5	Philosopher
6	6	Physicist
7	7	Chemist
8	8	Inventor
9	9	Artist

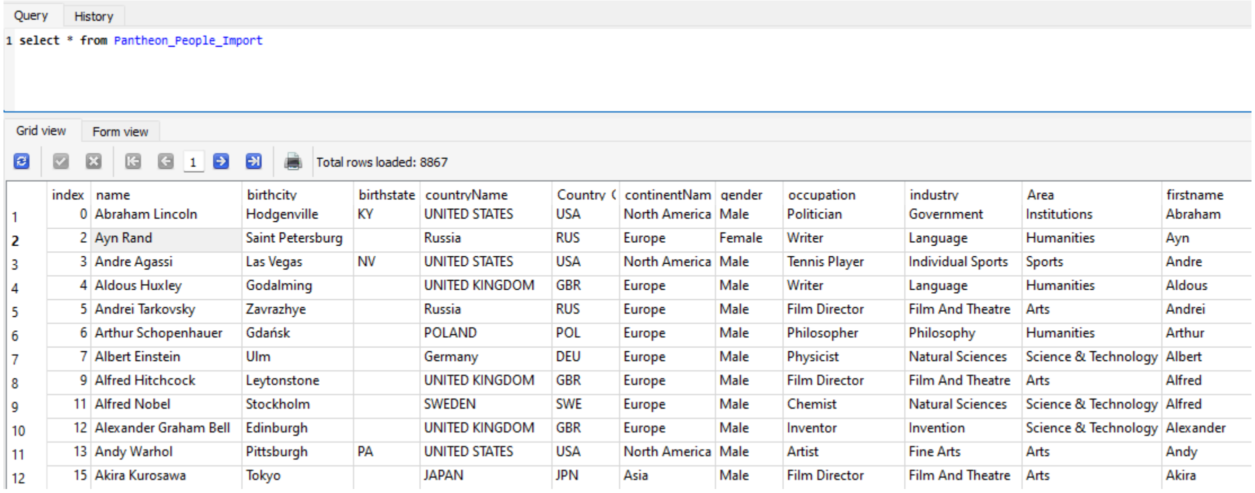
Pantheon_People_Import

This table is derived from the Harvard Pantheon 1.0 Dataset found here: (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28201)

Each row represents one person and demographic information, including the name, first, last, and middle name, country, state (if applicable), and city of birth, gender, and designated area, industry, occupation, and people look-upid.

Column Name	**Data Type**	**Explanation**
Index	integer	Index of dataframe
name	string	Person's name.
birthcity	string	City of person's birth
birthstate	string	State of person's birth; not applicable for all rows
countryName	string	Country of person's birth or nationality.
Country_Code	string	3-letter country code for the person's nationality or country of birth

continentName	string	Continent on which the country is
Gender	string	Male or female
occupation	string	Person’s occupation as designated by the Pantheon project
industry	string	Person’s industry as designated by the Pantheon project
area	string	Person’s domain as designated by the Pantheon project
firstname	string	Person’s first name; split out from name field
lastname	string	Person’s last name; split out from name field
middlename	string	Person’s middle name; split out from name field
People_Lookup	string	Joinable field created from the first 3 characters of the person's first name, first 3 characters of the person’s last name, and the year of their birth.



```
select * from Pantheon_People_Import
```

Total rows loaded: 8867

	index	name	birthcity	birthstate	countryName	Country (	continentNam	gender	occupation	industry	Area	firstname
1	0	Abraham Lincoln	Hodgenville	KY	UNITED STATES	USA	North America	Male	Politician	Government	Institutions	Abraham
2	2	Ayn Rand	Saint Petersburg		Russia	RUS	Europe	Female	Writer	Language	Humanities	Ayn
3	3	Andre Agassi	Las Vegas	NV	UNITED STATES	USA	North America	Male	Tennis Player	Individual Sports	Sports	Andre
4	4	Aldous Huxley	Godalming		UNITED KINGDOM	GBR	Europe	Male	Writer	Language	Humanities	Aldous
5	5	Andrei Tarkovsky	Zavrazhye		Russia	RUS	Europe	Male	Film Director	Film And Theatre	Arts	Andrei
6	6	Arthur Schopenhauer	Gdańsk		POLAND	POL	Europe	Male	Philosopher	Philosophy	Humanities	Arthur
7	7	Albert Einstein	Ulm		Germany	DEU	Europe	Male	Physicist	Natural Sciences	Science & Technology	Albert
8	9	Alfred Hitchcock	Leytonstone		UNITED KINGDOM	GBR	Europe	Male	Film Director	Film And Theatre	Arts	Alfred
9	11	Alfred Nobel	Stockholm		SWEDEN	SWE	Europe	Male	Chemist	Natural Sciences	Science & Technology	Alfred
10	12	Alexander Graham Bell	Edinburgh		UNITED KINGDOM	GBR	Europe	Male	Inventor	Invention	Science & Technology	Alexander
11	13	Andy Warhol	Pittsburgh	PA	UNITED STATES	USA	North America	Male	Artist	Fine Arts	Arts	Andy
12	15	Akira Kurosawa	Tokyo		JAPAN	JPN	Asia	Male	Film Director	Film And Theatre	Arts	Akira

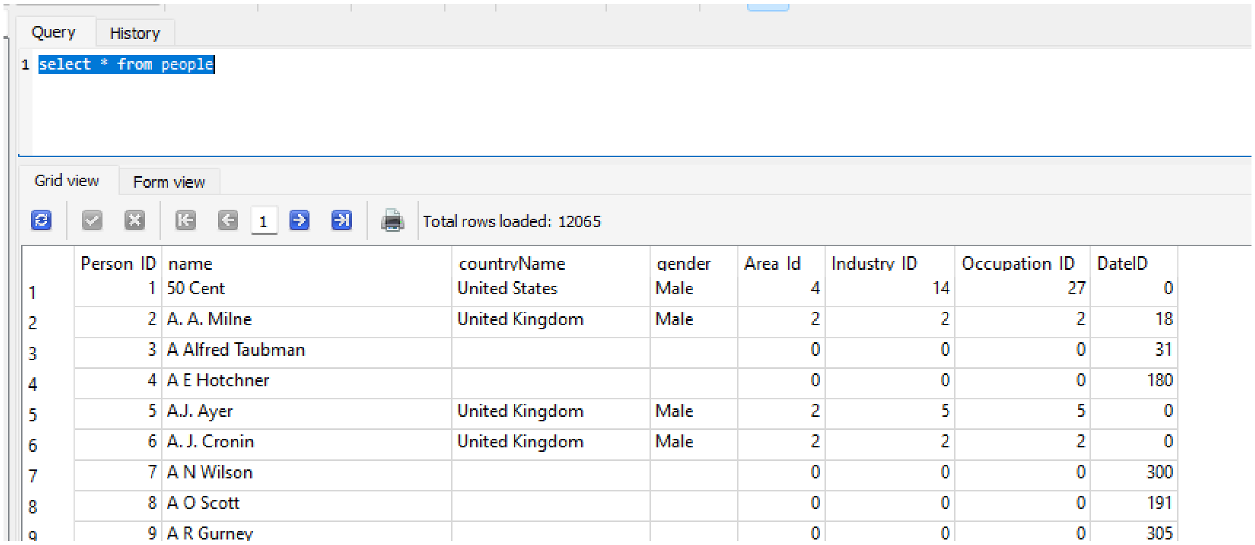
People

This is the final formatted table produced in this project. It is derived from a union of the rows in the famous person dataset and the Pantheon dataset joined on a people lookup code. It contains foreign keys for each row to link to the Area, Industry, Occupation, and Zodiac tables.

Each row represents one single person, their demographic data, including birthplace and gender, and foreign keys to identify area, industry, occupation, and zodiac sign.

Column Name	**Data Type**	**Explanation**
Person_ID	integer	Primary Key; unique ID assigned to each person.
name	string	The name of the person; for most rows this comes from the Pantheon people data.
countryName	string	Country of the person's nationality or birth; originated from Pantheon table
gender	string	Gender of the person; originated from Pantheon table
Area_Id	integer	Foreign key - joins to the Area table
Industry_ID	integer	Foreign key - joins to the Industry table
Occupation_ID	integer	Foreign key - joins to the

		Occupation table
DateID	integer	Foreign key - joins to the Zodiac table



```
1 select * from people
```

Total rows loaded: 12065

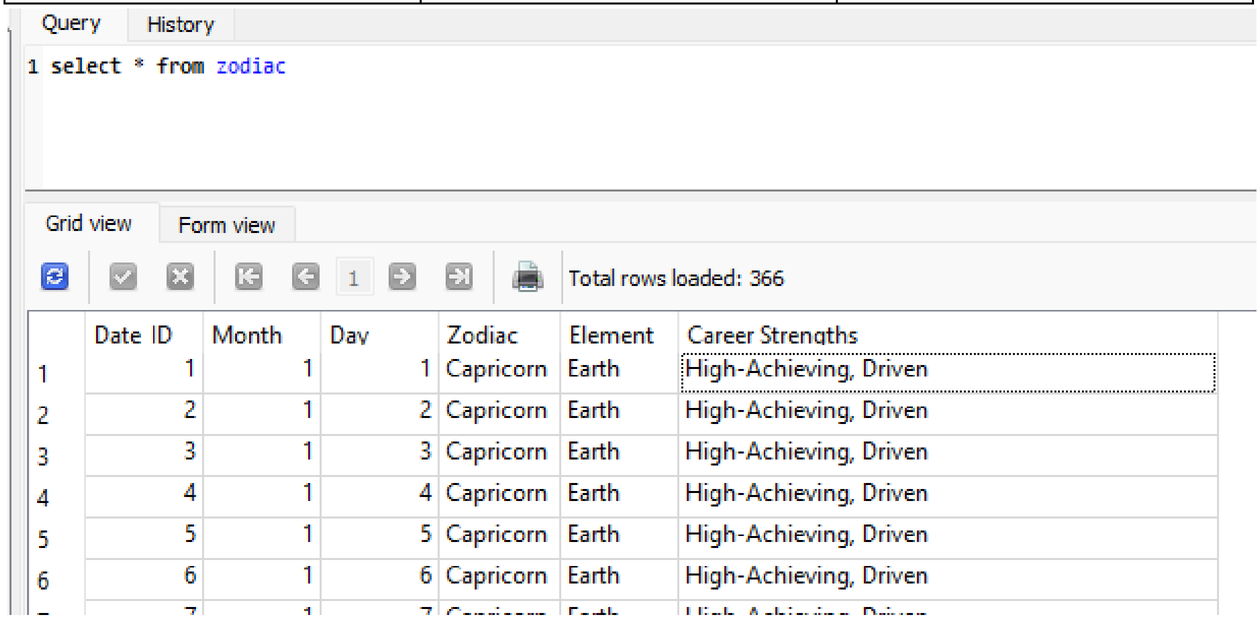
	Person ID	name	countryName	gender	Area Id	Industry ID	Occupation ID	DateID
1	1	50 Cent	United States	Male	4	14	27	0
2	2	A. A. Milne	United Kingdom	Male	2	2	2	18
3	3	A Alfred Taubman			0	0	0	31
4	4	A E Hotchner			0	0	0	180
5	5	A.J. Ayer	United Kingdom	Male	2	5	5	0
6	6	A. J. Cronin	United Kingdom	Male	2	2	2	0
7	7	A N Wilson			0	0	0	300
8	8	A O Scott			0	0	0	191
9	9	A R Gurney			0	0	0	305

Zodiac

This table is derived from the Zodiac csv file that I created for this project. Each row represents one day of the year.

Column Name	**Data Type**	**Explanation**
Date_Id	integer	Primary Key; this represents the day of the year, where January 1 = 1…December 31=366, etc.
Month	integer	Month of the year, where January = 1, February = 2, etc.
Day	integer	Day of the month, where 1 = 1st of the month, 2 = 2nd of the month, etc.
Zodiac	string	The Zodiac sign for the designated day and month.
Element	string	The element of the Zodiac sign

Career Strengths	string	Summary of career strengths, as found in (https://www.rd.com/list/career-strength-according-to-zodiac-sign/. Was ultimately not used in this project.



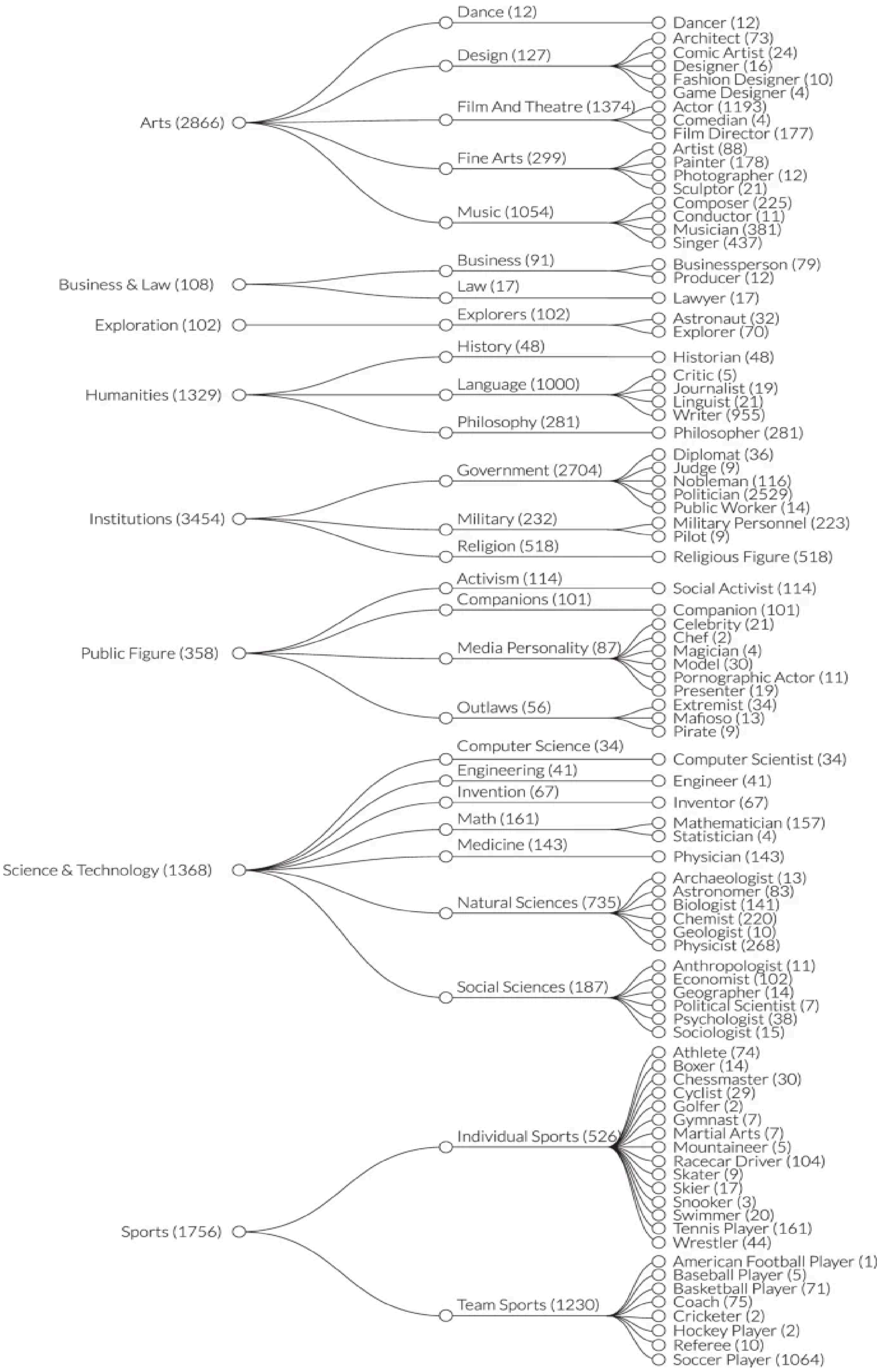
```
select * from zodiac
```

Total rows loaded: 366

	Date ID	Month	Day	Zodiac	Element	Career Strengths
1	1	1	1	Capricorn	Earth	High-Achieving, Driven
2	2	1	2	Capricorn	Earth	High-Achieving, Driven
3	3	1	3	Capricorn	Earth	High-Achieving, Driven
4	4	1	4	Capricorn	Earth	High-Achieving, Driven
5	5	1	5	Capricorn	Earth	High-Achieving, Driven
6	6	1	6	Capricorn	Earth	High-Achieving, Driven
7	7	1	7	Capricorn	Earth	High-Achieving, Driven

Relationships between Area, Industry, and Occupation

This image, taken from the abstract of the Pantheon project (https://www.nature.com/articles/sdata201575) , describes the relationships between the area, industry, and occupation of each entry. Area, on the left, represents the broadest category, with occupation, on the right, representing the most granular. It should be noted that numbers in parens represent the count of individuals in the original data, not in my project or analysis.


Arts (2866)
Dance (12)
Dancer (12)
Design (127)
Architect (73)
Comic Artist (24)
Designer (16)
Fashion Designer (10)
Game Designer (4)
Film And Theatre (1374)
Actor (1193)
Comedian (4)
Film Director (177)
Fine Arts (299)
Artist (88)
Painter (178)
Photographer (12)
Sculptor (21)
Music (1054)
Composer (225)
Conductor (11)
Musician (381)
Singer (437)
Business & Law (108)
Business (91)
Businessperson (79)
Producer (12)
Law (17)
Lawyer (17)
Exploration (102)
Explorers (102)
Astronaut (32)
Explorer (70)
Humanities (1329)
History (48)
Historian (48)
Language (1000)
Critic (5)
Journalist (19)
Linguist (21)
Writer (955)
Philosophy (281)
Philosopher (281)
Institutions (3454)
Government (2704)
Diplomat (36)
Judge (9)
Nobleman (116)
Politician (2529)
Public Worker (14)
Military (232)
Military Personnel (223)
Pilot (9)
Religion (518)
Religious Figure (518)
Public Figure (358)
Activism (114)
Social Activist (114)
Companions (101)
Companion (101)
Media Personality (87)
Celebrity (21)
Chef (2)
Magician (4)
Model (30)
Pornographic Actor (11)
Presenter (19)
Outlaws (56)
Extremist (34)
Mafioso (13)
Pirate (9)
Science & Technology (1368)
Computer Science (34)
Computer Scientist (34)
Engineering (41)
Engineer (41)
Invention (67)
Inventor (67)
Math (161)
Mathematician (157)
Statistician (4)
Medicine (143)
Physician (143)
Natural Sciences (735)
Archaeologist (13)
Astronomer (83)
Biologist (141)
Chemist (220)
Geologist (10)
Physicist (268)
Social Sciences (187)
Anthropologist (11)
Economist (102)
Geographer (14)
Political Scientist (7)
Psychologist (38)
Sociologist (15)
Sports (1756)
Individual Sports (526)
Athlete (74)
Boxer (14)
Chessmaster (30)
Cyclist (29)
Golfer (2)
Gymnast (7)
Martial Arts (7)
Mountaineer (5)
Racecar Driver (104)
Skater (9)
Skier (17)
Snooker (3)
Swimmer (20)
Tennis Player (161)
Wrestler (44)
Team Sports (1230)
American Football Player (1)
Baseball Player (5)
Basketball Player (71)
Coach (75)
Cricketer (2)
Hockey Player (2)
Referee (10)
Soccer Player (1064)

Explanation of the Zodiac

While astrology is a pseudo-science, it is one with which we are all at least somewhat familiar. Astrology combines traditions, religious beliefs, and pre-enlightenment "science" from many parts of the world and assigns meaning and causation to personality traits, acts, and events based on the movement and visibility of celestial bodies, that is, stars, planets, and other natural phenomena.

Astrological signs

Depending upon exactly which flavor of astrology one follows, the signs, their dates, and what exactly they mean may vary. Generally accepted in popular culture, however, one's sign is determined by date of birth and represented by a constellation, and has a supposed accompaniment of personality traits. Each element also has a designated element - fire, earth, water, or air. Personality traits below defined by this article.

Aries

- The Ram
- Spans mid-March to mid-April
- Fire element
- Aries are competitive, warm, bold, and lively.

Taurus

- The Bull
- Spans mid-April to mid-May
- Earth element
- Tauruses are stubborn, resolute, grounded, and resilient.

Gemini

- The Twins
- Spans mid-May to mid-June
- Air element
- Gemini are curious, witty, communicative, and youthful.

Cancer

- The Crab
- Spans mid-June to mid-July
- Water element
- Cancers are deeply emotional, sentimental, passionate, and loyal.

Leo

- The Lion
- Spans mid/end-July to mid/end-August
- Fire element
- Leos are dramatic, courageous, passionate and charismatic

Virgo

- The Maiden
- Spans mid/end-August to mid/end-September
- Earth element
- Virgos are deeply practical, analytical, detail-focused, desire to help.

Libra

- The Scale
- Spans mid/end-September to mid/end-October
- Air element
- Libras are artistic, indecisive, and desire to maintain harmony and peace.

Scorpio

- The Scorpion
- Spans mid/end-October to mid/end-November
- Water element

- Scorpios are mysterious, intense, deeply passionate, and independent.

Sagittarius

- The Archer
- Spans mid/end-November to mid December.
- Fire element
- Sagittarius are adventurous, free-spirited, playful, and constantly seeking new wisdom and experiences.

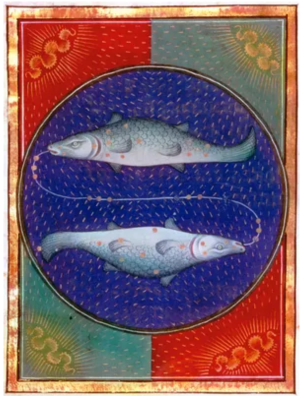
Capricorn

- The Sea Goat
- Spans mid/end-December to mid/end-January
- Earth element
- Capricorns are disciplined, dedicated, patient, and hardworking.

Aquarius

- The Water Bearer
- Spans mid/end-January to mid/end-February
- Air element
- Aquarians are truthful, intelligent, creative, and forward-thinking.

Pisces



- The FIsh
- Spans mid February to mid/end-March
- water element
- Pisces are empathetic, intuitive, understanding, and sensitive.