ETL Project

By: Todd Schanzlin, John Swierczynski, & Erin Lee

Extract

- Source: https://www.kaggle.com/datasets
- For this project we used 4 csv files and loaded them into DataFrames:
  - Astronauts.csv
    - Biographical and mission data for each NASA astronaut from 1959
  - Space_walks.csv
    - US and Russian extra-vehicular activity (“EVA”) from 1965
  - Space_missions.csv
    - Equipment, mission, and launch location / date information from 1957
  - Global_space_launches.csv
    - Equipment, mission, and launch location / date information from 1957

In [3]:

```
astr_df = pd.read_csv('CSVs/astronauts.csv')
spacewalks_df = pd.read_csv('CSVs/space_walks.csv')
spacemissions_df = pd.read_csv('CSVs/space_missions.csv')
global_launches_df = pd.read_csv('CSVs/global_space_launches.csv')
```

Transform

- We first cleaned the data by dropping unwanted columns.

Drop Columns

In [8]:
```
1 spacemissions_df.drop(labels=['Unnamed: 0', 'Unnamed: 0.1'], axis=1, inplace=True)
2 global_launches_df.drop(labels=['DateTime','Year','Month','Day','Time'], axis=1, inplace=True)
```

Transform

- We finished cleaning our data by renaming columns

Renaming Columns

```
global_launches_df.columns = ['company_name', 'location', 'detail', 'status_rocket', 'rocket',
       'status_mission', 'country_of_launch', 'company_country_origin',
       'private_or_state', 'date']

spacemissions_df.columns = ['company_name', 'location', 'date', 'detail', 'status_rocket',
       'rocket', 'status_mission']

spacewalks_df.columns = ['eva#', 'country', 'crew', 'vehicle', 'date', 'duration', 'purpose']

astr_df.columns = ['name', 'year', 'group', 'status', 'birth_date', 'birth_place',
       'gender', 'alma_mater', 'undergraduate_major', 'graduate_major',
       'military_rank', 'military_branch', 'space_flights',
       'space_flight_hours', 'space_walks', 'space_walks_hours', 'missions',
       'death_date', 'death_mission']
```

Transform

- We then converted the dates so that they would be in the same format
- To verify that all entries were converted, we checked for null values
- The spacewalks df had 44 unconverted “null” entries

```
## Counting number of unconverted dates
```

```
global_launches_df.date.isna().sum()
```

```
0
```

```
spacemissions_df.date.isna().sum()
```

```
0
```

```
spacewalks_df.date.isna().sum()
```

```
44
```

Date Conversions

```
date_df = pd.DataFrame({"Global Launches Dates": global_launches_df.Date,
                        "Spacemissions Dates": spacemissions_df.Datum,
                       "Spacewalks Dates": spacewalks_df.Date})
date_df.head()
```

	Global Launches Dates	Spacemissions Dates	Spacewalks Dates
0	07/08/2020	Fri Aug 07, 2020 05:12 UTC	06/03/1965
1	06/08/2020	Thu Aug 06, 2020 04:01 UTC	March 16-17, 1966
2	04/08/2020	Tue Aug 04, 2020 23:57 UTC	06/05/1966
3	30/07/2020	Thu Jul 30, 2020 21:25 UTC	07/19/1966
4	30/07/2020	Thu Jul 30, 2020 11:50 UTC	07/20/1966

```
## Splicing string to get rid of time and timezone
spacemissions_df.Datum = spacemissions_df.Datum.apply(lambda x: x[0:16])
```

```
# Built-in Pandas datetime function

global_launches_df.Date = pd.to_datetime(global_launches_df.Date, format='%d/%m/%Y', errors = "coerce")

spacemissions_df.Datum = pd.to_datetime(spacemissions_df.Datum, format= '%a %b %d, %Y', errors='coerce')

spacewalks_df.Date = pd.to_datetime(spacewalks_df.Date, format= "%m/%d/%Y", errors="coerce")
```

```
# Re-running date_df to verify conversions were successful
date_df = pd.DataFrame({"Global Launches Dates": global_launches_df.Date,
                        "Spacemissions Dates": spacemissions_df.Datum
                       ,"Spacewalks Dates": spacewalks_df.Date})
date_df.head()
```

	Global Launches Dates	Spacemissions Dates	Spacewalks Dates
0	2020-08-07	2020-08-07	1965-06-03
1	2020-08-06	2020-08-06	NaT
2	2020-08-04	2020-08-04	1966-06-05
3	2020-07-30	2020-07-30	1966-07-19
4	2020-07-30	2020-07-30	1966-07-20

Load

- We used pandas to sql to load our data into a Postgres database
- We also imported the dataframes into a SQLite database

Creating connection to space_db and converting dataframes to sql tables

Postgres:

```
password = os.environ.get('postgres_password')
```

```
engine = create_engine(f"postgresql://postgres:{password}@localhost:5432/space_db")
```

```
global_launches_df.to_sql("global_launches", engine)

spacemissions_df.to_sql("space_missions", engine)

spacewalks_df.to_sql("spacewalks", engine)

astr_df.to_sql("astronauts", engine)
```

SQLite:

```
sqlite_engine = create_engine('sqlite:///space_db.sqlite')
```

```
global_launches_df.to_sql("global_launches", sqlite_engine)

spacemissions_df.to_sql("space_missions", sqlite_engine)

spacewalks_df.to_sql("spacewalks", sqlite_engine)

astr_df.to_sql("astronauts", sqlite_engine)
```

Astronauts

index bigint	name text	year double precision	group double precision	status text	birth_date text	birth_place text	gender text	alma_mater text	undergraduate_major text	graduate_major text	military_rank text	military_branch text
0	Joseph...	2004	19	Active	5/17/1967	Inglewood, CA	Male	University of Cali...	Geology	Geology	[null]	[null]
1	Loren ...	[null]	[null]	Retired	3/7/1936	Lewiston, MT	Male	Montana State U...	Engineering Physics	Solar Physics	[null]	[null]
2	James ...	1984	10	Retired	3/3/1946	Warsaw, NY	Male	US Military Acad...	Engineering	Aerospace Engineeri...	Colonel	US Army (Retired)
3	Thoma...	1987	12	Retired	5/20/1951	St. Louis, MO	Male	University of Mis...	Applied Mathematics	Applied Mathematics	Colonel	US Air Force (Retired)
4	Buzz Al...	1963	3	Retired	1/20/1930	Montclair, NJ	Male	US Military Acad...	Mechanical Engineering	Astronautics	Colonel	US Air Force (Retired)
5	Andrew...	1987	12	Retired	8/4/1955	Philadelphia, PA	Male	Villanova Univer...	Mechanical Engineering	Business Administrat...	Lieutenant Colonel	US Marine Corps (Re...
6	Joseph...	1967	6	Retired	6/27/1937	Crawsfordsville,...	Male	DePauw Universi...	Mathematics & Physics	Physics	[null]	[null]
7	Scott D...	1995	15	Retired	8/15/1959	Lincoln, IL	Male	University of Illin...	Aeronautical & Astronautica...	Aeronautical Enginee...	Captain	US Navy (Retired)
8	William...	1963	3	Retired	10/17/1933	Hong Kong	Male	US Naval Acade...	Nuclear Engineering	Nuclear Engineering	Major General	US Air Force Reserve...
9	Clayton...	1998	17	Retired	2/23/1959	Omaha, NE	Male	Hastings College...	Physics	Aerospace Engineeri...	[null]	[null]

space_flights bigint	space_flight_hours bigint	space_walks bigint	space_walks_hours double precision	missions text	death_date text	death_mission text
2	3307	2	13	STS-119 (Dis...	[null]	[null]
1	190	0	0	STS 51-F (Ch...	[null]	[null]
2	334	0	0	STS-28 (Colu...	[null]	[null]
4	814	4	29	STS-41 (Disc...	[null]	[null]
2	289	2	8	Gemini 12, A...	[null]	[null]
3	906	0	0	STS-46 (Atla...	[null]	[null]
2	313	2	12	ST-5 (Columb...	[null]	[null]
4	1236	0	0	STS-90 (Colu...	[null]	[null]
1	147	0	0	Apollo 8	[null]	[null]
2	4005	6	38	STS-117/120...	[null]	[null]
2	594	0	0	STS-89 (Ende...	2/1/2003	STS-107 (Columbia)

Global Launches

index bigint	company_name text	location text	detail text	status_rocket text	rocket text	status_mission text	country_of_launch text	company_country_origin text	private_or_state text	date timestamp without time zone
0	SpaceX	LC-39A, Ken...	Falcon 9 ...	StatusActive	50.0	Success	USA	USA	P	2020-08-07 00:00:00
1	CASIC	Site 9401 (S...	Long Mar...	StatusActive	29.75	Success	China	China	S	2020-08-06 00:00:00
2	SpaceX	Pad A, Boca...	Starship ...	StatusActive	[null]	Success	USA	USA	P	2020-08-04 00:00:00
3	Roscosmos	Site 200/39,...	Proton-M...	StatusActive	65.0	Success	Kazakhstan	Russia	S	2020-07-30 00:00:00
4	ULA	SLC-41, Cap...	Atlas V 5...	StatusActive	145.0	Success	USA	USA	P	2020-07-30 00:00:00
5	CASIC	LC-9, Taiyua...	Long Mar...	StatusActive	64.68	Success	China	China	S	2020-07-25 00:00:00
6	Roscosmos	Site 31/6, B...	Soyuz 2....	StatusActive	48.5	Success	Kazakhstan	Russia	S	2020-07-23 00:00:00
7	CASIC	LC-101, Wen...	Long Mar...	StatusActive	[null]	Success	China	China	S	2020-07-23 00:00:00
8	SpaceX	SLC-40, Cap...	Falcon 9 ...	StatusActive	50.0	Success	USA	USA	P	2020-07-20 00:00:00
9	JAXA	LA-Y1, Tane...	H-IIA 202...	StatusActive	90.0	Success	Japan	Japan	S	2020-07-19 00:00:00

Space Missions

index bigint	company_name text	location text	date timestamp without time zone	detail text	status_rocket text	rocket text	status_mission text
0	SpaceX	LC-39A, Ken...	2020-08-07 00:00:00	Falcon 9 ...	StatusActive	50.0	Success
1	CASC	Site 9401 (S...	2020-08-06 00:00:00	Long Mar...	StatusActive	29.75	Success
2	SpaceX	Pad A, Boca...	2020-08-04 00:00:00	Starship ...	StatusActive	[null]	Success
3	Roscosmos	Site 200/39,...	2020-07-30 00:00:00	Proton-M...	StatusActive	65.0	Success
4	ULA	SLC-41, Cap...	2020-07-30 00:00:00	Atlas V 5...	StatusActive	145.0	Success
5	CASC	LC-9, Taiyua...	2020-07-25 00:00:00	Long Mar...	StatusActive	64.68	Success
6	Roscosmos	Site 31/6, B...	2020-07-23 00:00:00	Soyuz 2....	StatusActive	48.5	Success
7	CASC	LC-101, Wen...	2020-07-23 00:00:00	Long Mar...	StatusActive	[null]	Success
8	SpaceX	SLC-40, Cap...	2020-07-20 00:00:00	Falcon 9 ...	StatusActive	50.0	Success
9	JAXA	LA-Y1, Tane...	2020-07-19 00:00:00	H-IIA 202...	StatusActive	90.0	Success

Spacewalks

index bigint	eva# double precision	country text	crew text	vehicle text	date timestamp without time zone	duration text	purpose text
0	1	USA	Ed White	Gemini IV	1965-06-03 00:00:00	0:36	First U.S. EV...
1	2	USA	David ...	Gemini VIII	[null]	0:00	HHMU EVA ...
2	3	USA	Eugene...	Gemini IX-A	1966-06-05 00:00:00	2:07	Inadequate r...
3	4	USA	Mike C...	Gemini X	1966-07-19 00:00:00	0:50	Standup EV...
4	5	USA	Mike C...	Gemini X	1966-07-20 00:00:00	0:39	Retrieved M...
5	6	USA	Richar...	Gemini XI	1966-09-13 00:00:00	0:44	Attached tet...
6	7	USA	Richar...	Gemini XI	1966-09-14 00:00:00	2:10	Standup EV...
7	8	USA	Buzz Al...	Gemini XII	1966-11-12 00:00:00	2:29	Standup EV...
8	9	USA	Buzz Al...	Gemini XII	1966-11-13 00:00:00	2:06	Attached tet...
9	10	USA	Buzz Al...	Gemini XII	1966-11-14 00:00:00	0:55	Standup EV...

Queries

```sql
SELECT spacewalks.crew, spacewalks.country,
    space_missions.company_name, spacewalks.vehicle,
    spacewalks.date
FROM spacewalks
JOIN space_missions
ON spacewalks.date = space_missions.date
ORDER BY date ASC;

SELECT * FROM spacewalks
WHERE date BETWEEN '1969-01-01' AND '1969-12-30';

SELECT * FROM space_missions
WHERE date BETWEEN '1969-01-01' AND '1969-12-30'
order by date asc;


SELECT * FROM global_launches
WHERE country_of_launch = 'USA'
AND company_name = 'NASA'
ORDER BY date ASC;

SELECT * FROM astronauts
WHERE missions LIKE '%Apollo%';

SELECT * FROM spacewalks
WHERE purpose LIKE '%First%';
```

NASA Astronauts (1959-Present)

- We used the Pandas groupby function to explore the top 20 most common graduate majors among all NASA astronauts

graduate_major	counts
Aeronautical Engineering	27
Aerospace Engineering	21
Medicine	16
Physics	15
Mechanical Engineering	13
Electrical Engineering	8
Aeronautics & Astronautics	7
Aviation Systems	6
Astronomy	6
Engineering Management	5
Astronautics	5
Mechanical Engineering	5
Aeronautics	4
Aeronautical Systems	4
Public Administration	3
Business Administration	3
Ocean Engineering	3
Astronautical Engineering	3
Nuclear Engineering	3
Chemical Engineering	3

Company Responsible for Launch

- Similarly, we found the top 20 companies who have sent the greatest amount of spacecraft into space

company_name	counts
RVSN USSR	1777
Arianespace	279
General Dynamics	251
CASC	251
NASA	203
VKS RF	201
US Air Force	161
ULA	140
Boeing	136
Martin Marietta	114
SpaceX	100
MHI	84
Northrop	83
Lockheed	79
ISRO	76
Roscosmos	55
ILS	46
Sea Launch	36
ISAS	30
Kosmotras	22

Spacewalk Firsts

We were also able to search the data using SQLAlchemy to observe US Spacewalk firsts

In [43]:

```
walk_firsts = session.query(SpaceWalks.date, SpaceWalks.crew, SpaceWalks.purpose).\
    filter(SpaceWalks.purpose.like('%First%'))

for walk in walk_firsts:
    print(walk[0])
    print(walk[1])
    print(walk[2])
    print("--------------------------------------------------------------------------------")
```

```
1965-06-03 00:00:00
Ed White
First U.S. EVA. Used HHMU and took  photos.  Gas flow cooling of 25ft umbilical overwhelmed by vehicle ingress work and helm
et fogged.  Lost overglove.  Jettisoned thermal gloves and helmet sun visor
--------------------------------------------------------------------------------
1969-03-06 00:00:00
Russ Schweickart
Lunar module based.  Took photos.  Evaluated foot restraint and handrails.  Retrieved thermal experiment samples.  First use
of PLSS followed by recharge demo after EVA
--------------------------------------------------------------------------------
1969-07-20 00:00:00
Neil Armstrong|Buzz Aldrin
First to walk on the moon.  Some trouble getting out small hatch.  46.3 lb of geologic material collected.  EASEP seismograp
h and laser reflector exp deployed.  Solar wind exp deployed & retrieved.  400 ft (120m) circuit on foot.  Dust issue post E
VA
--------------------------------------------------------------------------------
1971-07-31 00:00:00
David Scott|James Irwin
Collected 169 lb of geologic material.  ALSEP exp deployed.  First use of the lunar rover.  Covered 6.2 mile (10.3 km) circu
it
--------------------------------------------------------------------------------
1971-08-05 00:00:00
Al Worden
First transearth EVA.  Retrieved 2 camera film cassettes
--------------------------------------------------------------------------------
1982-11-14 00:00:00
Bill Lenoir|Joe Allen
Suit fan and O2 regulator failures prevented first Shuttle EVA.  No spare suit onboard
```