ETL Project

By: Todd Schanzlin, John Swierczynski, & Erin Lee

Extraction:

We used Kaggle to extract four different csv files. Three of our data sets are based on space launches/missions. The last csv file was a dataset of astronauts. The following files is the datasets we used from Kaggle.

- Astronauts.csv
- Space_walks.csv
- Space_missions.csv
- Global_space_launches.csv

Transformation:

After we extracted our csv files, we used pandas to read the files into DataFrames (Figure 1). From there we scanned the DataFrames to see what needed to be cleaned. To clean our data, we first dropped unwanted columns (Figure 2). In our datasets, we had different formats for dates (figure 3). We used the datetime function so that the date column in each DataFrame would be in the same format (Figure 4). We finished cleaning our data by renaming some columns.

Figure 1:

```
In [3]:
1 astr_df = pd.read_csv('CSVs/astronauts.csv')
2 spacewalks_df = pd.read_csv('CSVs/space_walks.csv')
3 spacemissions_df = pd.read_csv('CSVs/space_missions.csv')
4 global_launches_df = pd.read_csv('CSVs/global_space_launches.csv')
```

Figure 2:

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)
```

Figure 3:

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

Figure 4:

```
## 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:

Our final step was to get our cleaned data into a Database. To do this we created an engine to connect to postgres. We used pandas_to_sql formula to transfer our DataFrames into a database (figure 5).

Figure 5:

Creating connection to space_db and converting dataframes to sql tables

```
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)
```

Summary:

We used this data because we wanted to answer the following questions about space launches/missions:

- How has space launch activity changed from year to year by country, mission type or scope?
- How many space missions are moon landings, space walks, science experiments or other?
- How do you become a US Astronaut -- review of background of US Astronauts over the years and how their backgrounds have changed?

After we pulled our data into a database we were able to run a few queries to answer some of our questions (figure 6).

Figure 6:

```
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%';
```