**ETL Project**

U.S. Space Missions: 1957 - 2020

**Julian Mejia, Lareena LIamado, Scott Provencher, Taylor Clark**

Data Visualization Bootcamp

**The Technique:**

* The sources of data that you will extract from:

https://www.kaggle.com/davidroberts13/one-small-step-for-data

https://www.kaggle.com/rosetabares/spacemissionsflightstatus

* The type of transformation needed for this data:

1. Imported two csv.files into Pandas.
2. Merged two data frames into one data frame, space\_df.
3. Converted the ‘Datum’ column data type to datetime.
4. Filtered space\_df by ‘Location’ for USA using str.slice.
5. Filtered usa\_df by ‘Datum’ to create a separate data frame for each decade.
6. Added ‘Year’ column and filled it with the year of the launch for future data manipulation.
7. SQL statement JOIN used to join the two tables together on company name.
8. SQL statement GROUP BY used to create table with columns.
9. SQL statement ORDER BY used sort table.

* The type of final production database to load the data into a relational database using Postgres.
* The final tables or collections that will be used in the production database.

Select sm.launchdate, s.location, sm.company, sm.failurereason

FROM space\_corrected as s

JOIN spacemissions sm ON (s.companyname = sm.company)

GROUP BY sm.company, s.location, sm.failurereason, sm.launchdate

Order By sm.launchdate

**Conclusion:**

Extract: Our data was recovered from two datasets on Kaggle.com Data pulled from Space\_Corrected.csv and the data was assigned to the variable called Space\_Corrected. Data pulled from SpaceMissions.csv and the data was assigned to the variable called SpaceMissions.

*#Create dataframes from csv files*

Space\_Corrected = pd.read\_csv(file1)

SpaceMissions = pd.read\_csv(file2)

Transform: We merged two data frames into one data frame using concat function; concat([Space\_Corrected, SpaceMissions] space\_df. We converted the ‘Datum’ column data type to datetime. We filtered space\_df by ‘Location’ for USA using str.slice. We filtered usa\_df by ‘Datum’ to create a separate data frame for each decade. Added ‘Year’ column and filled it with the year of the launch for future data manipulation.

Load: We chose to not assign primary keys to either table because it showed duplicate rows when we tried to do any Select statement.

We used JOIN and Group BY to join the tables Space\_corrected and SpaceMission together in Postgres:

Select sm.launchdate, s.location, sm.company, sm.failurereason

FROM space\_corrected as s

JOIN spacemissions sm ON (s.companyname = sm.company)

GROUP BY sm.company, s.location, sm.failurereason, sm.launchdate

Order By sm.launchdate