

Space Launch History

Group 1 ETL Project

Introduction

This project was executed to meet the Extract Transform and Load (ETL) Module requirements for the SMU Data Visualization Course, Winter 2019 Cohort. The members of Group 1 were Bethany Lindberg, Gael Ruta Gatera, Hao Bai and Miche Maniguet.

The goal of the project was to collect data sets documenting space launch data over the period from 1957 through 2018 and to perform ETL on those datasets.

The team pushed three primary types of information to Github. These included original or input dataset, the output dataset, and the source code dataset.

Extract

As we were brainstorming database ideas for the Extract Transform and Load project for Wednesday May 1st's class we discussed the SpaceX API as possible source of data from the postman development environment. However, most of the work was already done for the data that they had so we decided not to use it. As a team we decided to stay within the Aerospace spectrum as we had interesting discussions coming from the aforementioned database and saw a world of possibilities.

The first dataset you will see in the Space_ETL file is the SpaceX csv file from Kaggle.

Description: The owner of this dataset obtained the information by scraping the SpaceX and NASA websites respectively. This dataset contains a record of every single SpaceX launch from 2006 until the time this dataset was uploaded (2017). Information includes but not limited to location, date/time and outcome (success or failure), and payload carried. The dataset was downloaded in the form of comma separated value (csv) the extracting was done using the the read_csv() function from pandas. Code can be seen in the .ipynb file.

Link: <https://www.kaggle.com/spacex/spacex-missions/downloads/spacex-missions.zip/1>

The second dataset is the NASA facilities that we got from Kaggle.

Description: The dataset is originally from data.nasa.gov and was edited to remove the contact information of each facility. The version we downloaded on Kaggle shows that it was uploaded two years ago (2017). While the original file's last update was June 27th 2018. When one gets on the original files website, there is an option to contact the owner of the dataset in order to download it. Most likely because the information in the dataset can get in the hands of the wrong people. Besides that fact, the data.nasa.gov shows that as of May 12th 2019, the dataset viewed 4898 times and downloaded 1207

times. The dataset was downloaded in the form of comma separated value (csv) and the extracting was done using the `read_csv()` function from pandas. Code can be seen in the .ipynb file.

Original Source of dataset: <https://data.nasa.gov/Management-Operations/NASA-Facilities/gvk9-iz74>

Link: <https://www.kaggle.com/nasa/nasa-facilities>

The third dataset was obtained from a Github user by the name of r-spacex

Description: The Github user obtained the data from <https://documenter.getpostman.com>. The dataset from this API is similar as the first source of SpaceX data from Kaggle but included more information and especially recent more information. The dataset was downloaded in the form of a Json file. The extracting was done using the `read_json()` function from pandas. Code can be seen in the .ipynb file.

Link: <https://github.com/r-spacex/SpaceX-API>

The fourth dataset was obtained from [spacelaunchreport.com](http://www.spacelaunchreport.com)

Description: The team came across this website as we were brainstorming for ideas. This website had lots of information and the team initially looked all over the site for some sort of excel or csv file that we could use but could not find any. We looked on the left hand side of the website and saw many tabulated launch logs. We then proceeded to scrape the website for the launch log of every single year from 1957 to 2019 and put into one excel csv file. Scraping definitely helped us as we could not let any valuable data slip through our hands.

Link: <http://www.spacelaunchreport.com/>

The fifth dataset was obtained from quite an unusual source. It came from the Union of Concerned Scientists. This dataset was exclusively of satellites. From the nature of the organization, this dataset was most likely collected by the organization in order to keep track of space debris. This dataset contain information of satellites that was launched from 1974. It includes the name of the satellite, UN registry code, Perigee, Apogee (both are orbital positioning information), and purpose. The dataset was downloaded in the form of comma separated value (csv) and the date format did not have to be changed as it was already in the desired format. The extracting was done using the `read_csv()` function from pandas. This particular dataset had some characters that UTF-8 could not recognized therefore the "cp1252" encoding type had to be used. The code can be seen in the .ipynb file.

Link: <http://www.spacelaunchreport.com/>

The sixth dataset was obtained from a online report which includes launch log data since 1957. Launch log data was obtained through internet searches from the following sources as .txt file first.

Unfortunately, the original data file is not comma-delimited. It was transformed into .csv file through importing the .txt file into Excel and setting up the column width manually. The extracting was done using the `read_csv()` function from pandas. Date format was reset to match that in other datasets. The code can be seen in the .ipynb file.

Link: <https://planet4589.org/space/log/launchlogy.txt>

Transform

Successfully extracting all the data made transforming portion of the process easier for our team. Each csv was transformed into a dataframe and either “.column()” or “.head()” methods were used in order to check whether the files were extracted properly and to look at the names of the columns and make appropriate changes.

SpaceX

The first and third dataset both had information regarding SpaceX launches and were both extracted. However, since the API dataset was a larger dataset and had more recent information, it was the one that the team chose to load. This particular dataset came in the form of a json file and was the more challenging to transform.

NASA Facilities

The Nasa facilities data was transformed to have the latitude and longitude data put in separate columns and zip code removed from the latitude column. This would make extracting either the coordinates or the cities easier for use in an API based on location. The dates were also formatted with null protection to be compliant with the MYSQL DATE format requirements. To finish, the columns were renamed to make them SQL friendly and superfluous columns were dropped.

UCS Satellites

The UCS Satellites data set was transformed to have the numerical and date columns in the appropriate format to be loaded into a MYSQL database. As on other data sets, columns were given SQL appropriate names and unneeded columns were dropped prior to being loaded.

Launch Log

The launch log data set was transformed to have the numerical and date columns in the appropriate format to be loaded into a MYSQL database. NA values in the date column were filled. As on other data sets, columns were given SQL appropriate names and unneeded columns were dropped prior to being loaded

.

Load

The transformed data was transferred to mySQL. A relational database was chosen due to the fact that a number of separate data collections were obtained, so a loading the components would allow further analysis and evaluation.

After the data had been transformed, the data was examined to create the table schemas. These can be found in SQL_create.sql. If this script is not run prior to loading the tables into MYSQL, all columns in all tables will have a text format.

Summary Queries

Some queries were written in MySQL workbench to get an idea of the data prior to data exploration in Jupyter Notebooks.

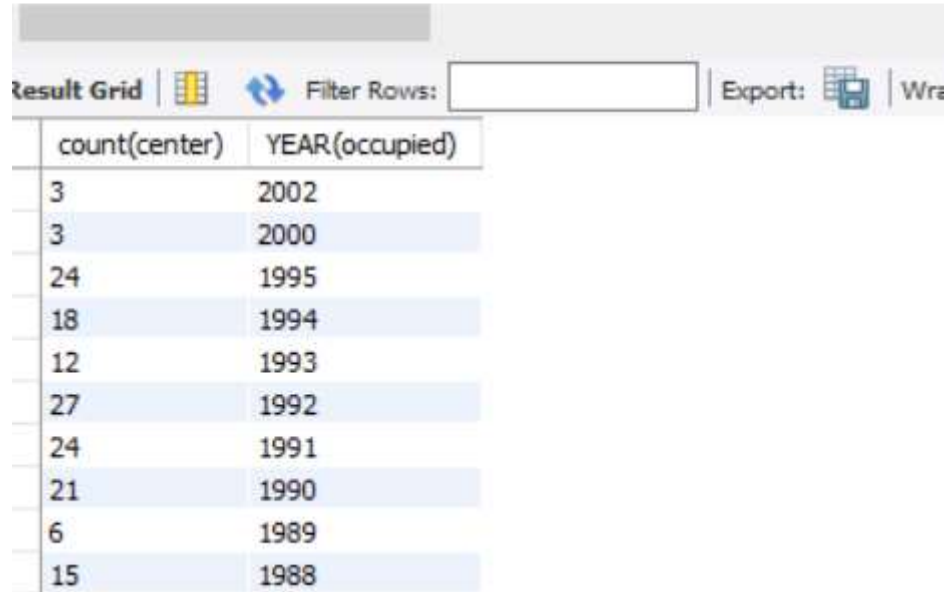
Total Global Launches By Date

```
10 • SELECT count(launch_date),YEAR(launch_date)
11 FROM launch_log
12 GROUP BY YEAR(launch_date);
13
```

result Grid		Filter Rows:	Export:	Wrap Cell C
count(launch_date)	YEAR(launch_date)			
414	2019			
1386	2018			
1323	2017			
663	2016			
708	2015			
768	2014			
621	2013			
396	2012			
387	2011			

Summary Total NASA Launches By Date

```
6 • SELECT count(center),YEAR(occupied)
7 FROM nasa_facilities
8 GROUP BY YEAR(occupied);
9
```



The screenshot shows a database interface with a 'Result Grid' tab. Above the grid is a toolbar with icons for a grid, a refresh button, a 'Filter Rows' input field, an 'Export' button with a download icon, and a 'Write' button. The result grid contains two columns: 'count(center)' and 'YEAR(occupied)'. The data is as follows:

count(center)	YEAR(occupied)
3	2002
3	2000
24	1995
18	1994
12	1993
27	1992
24	1991
21	1990
6	1989
15	1988

Summary and Evaluation

The most important visualization was the chart showing number of launches over time. Below is that chart and screenshot of the Jupyter Python Notebook and associated code used to produce the chart.

```

: launch_log["Year"] = launch_log["launch_date"].map(lambda x: int(x.strftime('%Y')))
launch_log2 = launch_log[launch_log["Year"] != 2019]

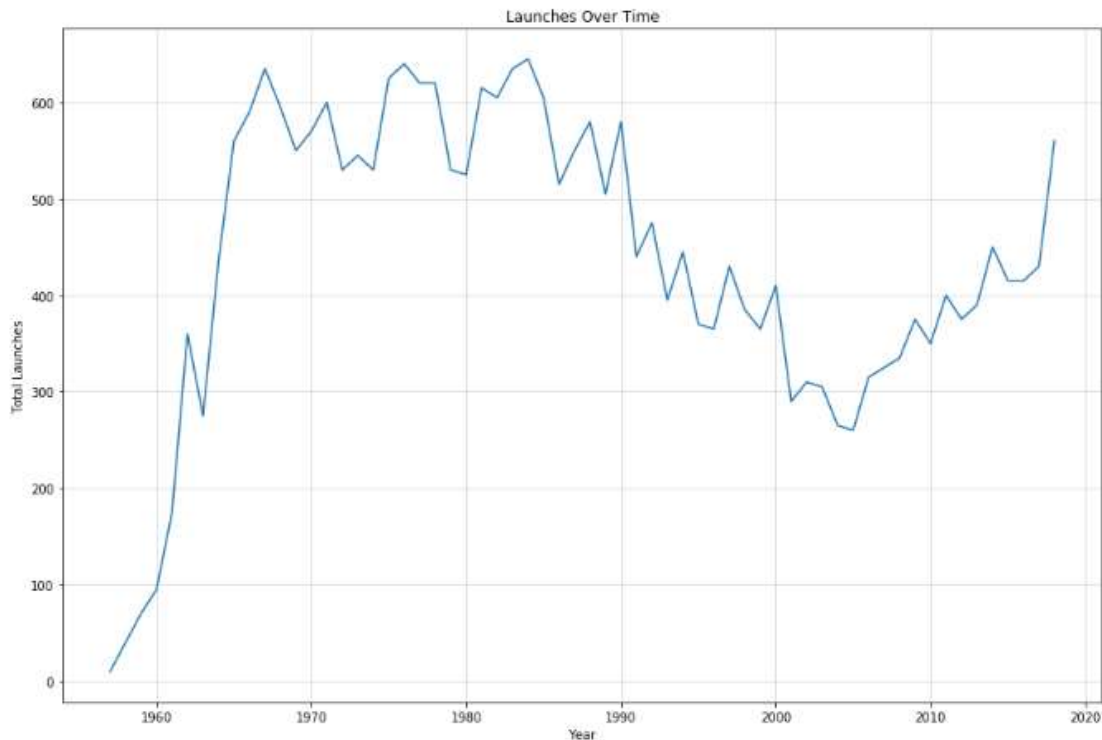
x_axis = launch_log2.groupby('Year')['Year'].max()
launches = launch_log2.groupby('Year')['Year'].count()
plt.figure(figsize = (15,10))

plt.plot(x_axis, launches)
plt.xlabel("Year")
plt.ylabel("Total Launches")
plt.grid(True,alpha=.5)
# plt.xticks(range(1957,2018,5))

plt.title("Launches Over Time")

# plt.savefig('.\\Output\\LaunchesOvertime.png')
plt.show()

```



This chart is highly significant in that it shows a dramatic drop in space launches exactly coincident with the end on the Cold War in 1991. Post 1991, a 15 year long decline in launches prevailed.

Since the Cold War is typically describes as having been over the period from 1947 to 1991, the number of launches over that period of 25 years, with an approximate 10 year rapid ramp up period, and an approximately 15 year ramp down period, is completely consistent with a common sense expectation for the number of launches over that 50 year span.

This one chart raises a number of questions however, in light of the almost 70 years of data that it represents.

While data is not without its limitations, cursory analysis suggests that the world is going through a possible resurgence of increase interest in space exploration, scientific study and possible

weaponization. The satellite dataset shows that the purposes of these launches include communication, observation, research, education and others. Users of these satellites are civil, government and military as well as commercial with most satellites likely to be a collaborative effort between the public and private sector. The satellite dataset also shows a large participation around the globe with over 100 countries launching at least 1 satellite between 1974 and 2019.

With all of our datasets showing a trend upward in competition and innovation, many more questions have been raised that beg for further exploration and data.

Some examples of questions that we are interested in exploring include:

Who were the primary entities that were launching and how many launches per year during the boom period of launches? This would demonstrate that during the cold war period (1947-1991), NATO & The Warsaw Pact were driving launch numbers. In 2018 we have not matched the peak of the cold war but the trend shows we may within the next few years. Is the resurgence that began in 2006 reflective of a modern arms race that parallels the Cold War? Who are the players driving the current resurgence? What are the reasons for their launch missions? Are the stated or publicly available purpose for launch true to their actual purposes?

All of these questions and answers would be guided by the goal of the data mining and exploration and would lead to next steps on further building the database we started on during the course of this project. With concrete answers to these and other questions we would be able to make recommendations about whether entering the space exploration industry would be worthwhile from a profit or scientific standpoint.