The Findings of SpaceX Project

Jackie

2022 Jun 13

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- SpaceX has four most common launch sites, KSC LC-39A, CCAFS SLC-40, VAFB SLC-4E, CCAFS LC-40.
- Launch sites locate in Florida and Los Angeles.
- The Landing success rate is about 60% - 70%, where KSC LC-39A has the highest success rate.
- Most Payload Mass ranges from 2k to 6k.
- In order to predict the landing outcome, KNN is the optimum algorithms.

INTRODUCTION

- Compares to normal cost of 165M dollars, SpaceX costs only 62M dollars due to reusable first stage.
- SpaceX has numbers of **launch sites** like CCAFS, SLC-40 etc.
- The **payload mass** ranges from 500 kg to 15,600 kg
- The **orbits** are GTO, ISS, VLEO etc. The most common orbit is GTO, about 27 times.
- We wants to **predict the landing outcomes** using a group of factors including launch site, payload mass, orbits etc.

METHODOLOGY

1. **Data Collection**: Collect the data via api and webscraping.
2. **Data Preparation:** Use pandas module to normalize, standardize, and clean the collected data. Load the data from IBM DB2 databases, further analyze data by executing SQL queries.
3. **Data Visualization:** Visualize the data pattern by using matplotlib, seaborn modules. Plus, create the map to visually show the launch sites in United States by using Folium module. Last, deliver the interactive, visual board about the data patterns.
4. **Modeling:** Use logistic regression, SVM, decision tree to model the data, predict the landing outcome.

Data Collection and Wrangling Methodology

Data Collection

- **Api collection** – I use the requests module to gain the response from the *spacexdata* site. Then use the within pandas function json_normalize to process the data into data frame.
- **Webscraping** – I scrape the data from wikipage of Falcon 9 launches. Then use the bs4 module to extract the table, columns, rows of the launch related data.

Data Wrangling

- **Handle the missing data –** identify the missing data, and use the mean value to replace the missing data.
- **Create the "Class" col –** define the bad outcomes, and create the new column Class to describe the landing outcome.

EDA and Visualization Methodology

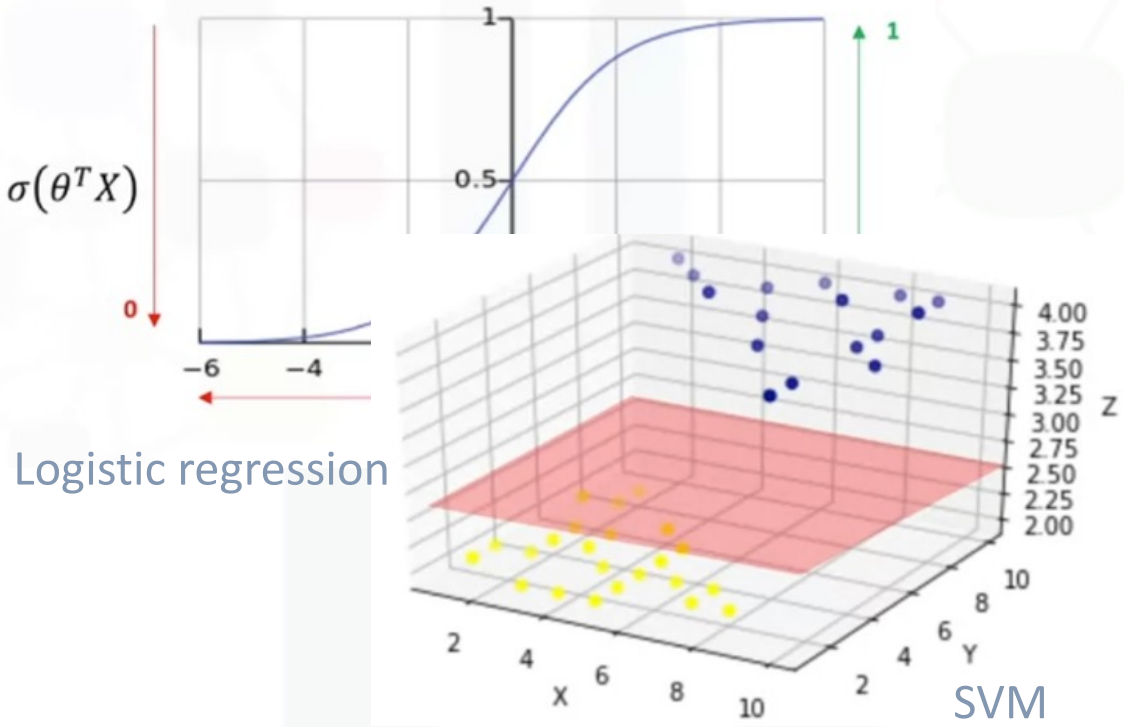
Data Exploration

- **SQL queries:** connect to the IBM DB2 database, use the SQL queries languages to generally discover the patterns of SpaceX launches, such as average payload mass, total # of success launches etc.

Data Visualization

- **General graphs:** use the matplotlib and seaborn modules to describe the relationships between factors. The graph types include scatter plot and bar chart.
- **Interactive map**: use the folium module to show the various launch sites locations, density, and the distance between launch site and railways etc.
- **Interactive visualization boards**: use the plotly dash to show the relationship between success rate and launch sites, payload mass etc.

Predictive Analysis Methodology



Logistic regression

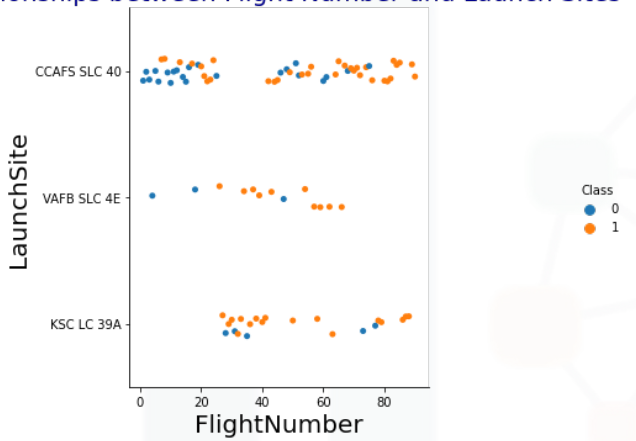
SVM

Machine Learning

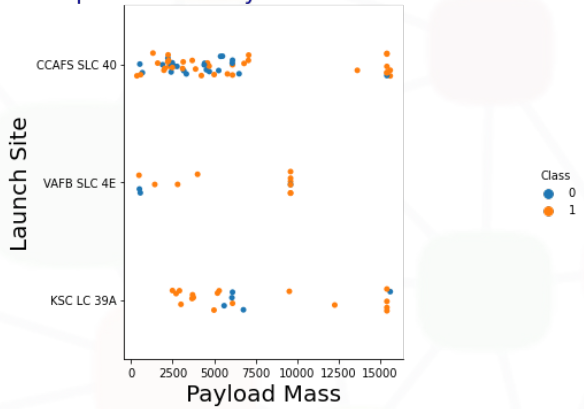
- **Machine leaning algorithms:** split the data set to train and test data sets, adopt the grid search methos, separately use the **logistic regression, SVM, decision tree** algorithms to predict the landing outcome. Review the accuracy of test data to decide the best approach.

EDA with Visualization Results

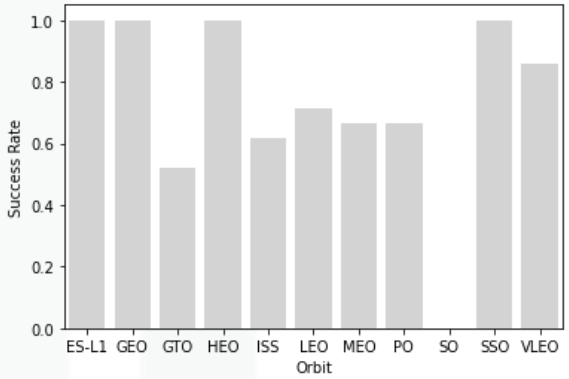
Relationships between Flight Number and Launch Sites



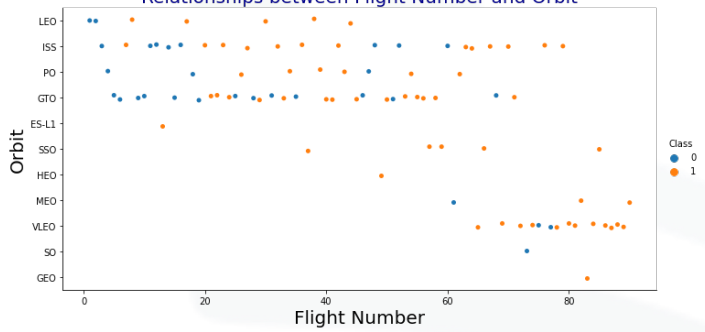
Relationships between Payload Mass and Launch Sites



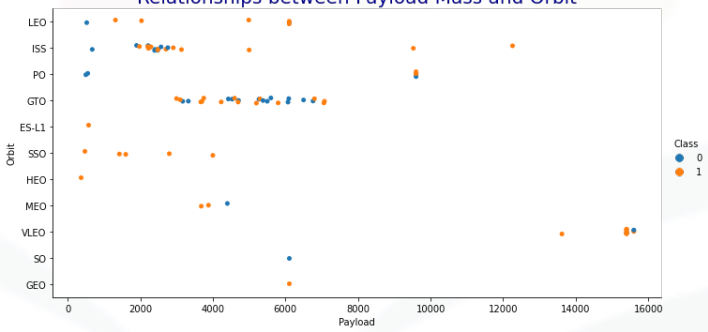
Different Orbits Success Rate



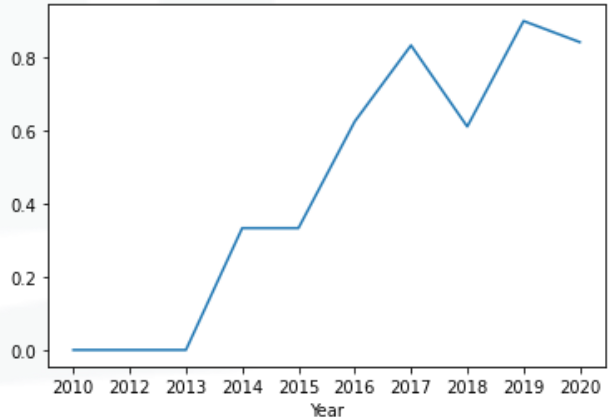
Relationships between Flight Number and Orbit



Relationships between Payload Mass and Orbit



Success Rate over Years



EDA with SQL Results

For some reason, my local JupyterNotebook could not import the IBM db2 module. Therefore, I went to the IBM DB2 on cloud and run the SQL queries directly on web.

```
-- 1. Display the names of the unique launch sites  in the space mis
select unique launch_site from spacex;

-- 2. Display 5 records where launch sites begin with the string 'CC
select launch_site, date
from spacex
where launch_site like 'CCA%'
limit 5;

-- 3. Display the total payload mass carried by boosters launched by
select sum(payload_mass__kg_)
from spacex
where customer = 'NASA (CRS)';

-- 4. Display average payload mass carried by booster version F9 v1.
select avg(payload_mass__kg_)
from spacex
where booster_version = 'F9 v1.1';

-- 5. List the date when the first successful landing outcome in gro
select min(date)
from spacex
where landing__outcome = 'Success (ground pad)';
```

LAUNCH_SITE
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

q1

LAUNCH_SITE	DATE
CCAFS LC-40	2010-06-04
CCAFS LC-40	2010-12-08
CCAFS LC-40	2012-05-22
CCAFS LC-40	2012-10-08
CCAFS LC-40	2013-03-01

q2

1
45596

q3

1
2928

q4

1
2015-12-22

q5

BOOSTER_VERSION
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

q6

MISSION_OUTCOME	TIMES
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

q7

```
-- 6.List the names of the boosters which have success in drone ship
select booster_version
from spacex
where landing__outcome ='Success (drone ship)'
and payload_mass__kg_ between 4000 and 6000;

-- 7. List the total number of successful and failure mission outcom
select mission_outcome, count(*) as times
from spacex
group by mission_outcome;

-- 8. List the names of the booster_versions which have carried the
select booster_version
from spacex
where payload_mass__kg_ =
    (select max(payload_mass__kg_) from spacex);

-- 9. List the failed landing_outcomes in drone ship, their booster
select booster_version, launch_site
from spacex
where landing__outcome = 'Failure (drone ship)' and year(date) = 201

-- 10.Rank the count of landing outcomes (such as Failure (drone shi
select landing__outcome, count(*)
from spacex
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome;
```

BOOSTER_VERSION
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5

q8

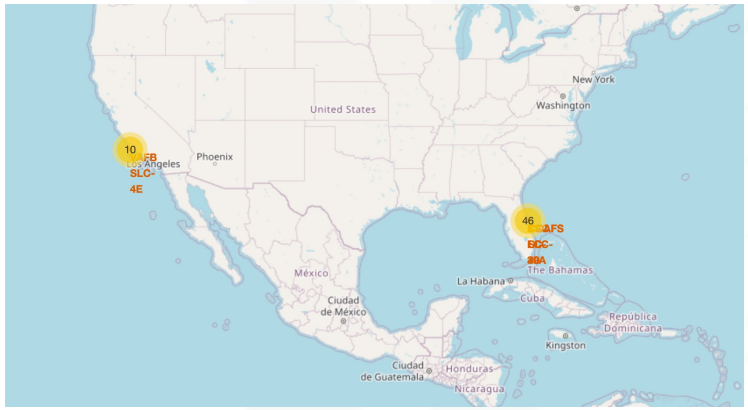
BOOSTER_VERSION	LAUNCH_SITE
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

q9

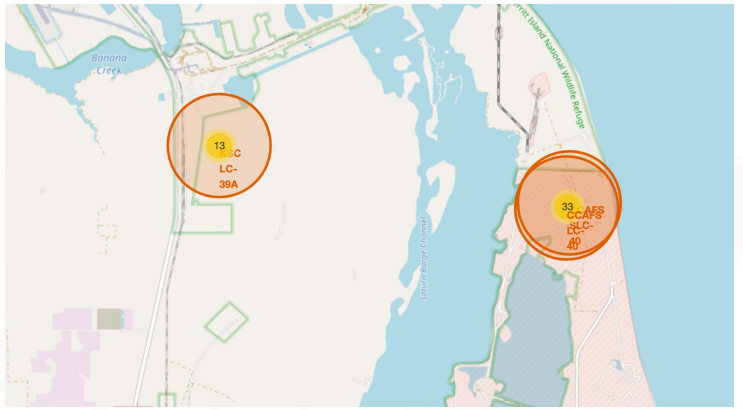
LANDING__OUTCOME ↑↓	2
Controlled (ocean)	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	10
Precluded (drone ship)	1

q10

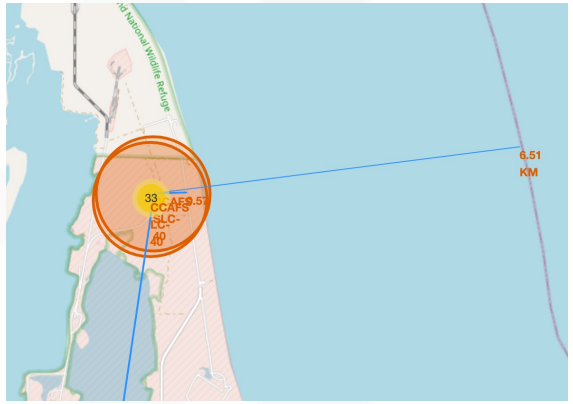
Interactive Map Results



Launch sites spread over the United States.

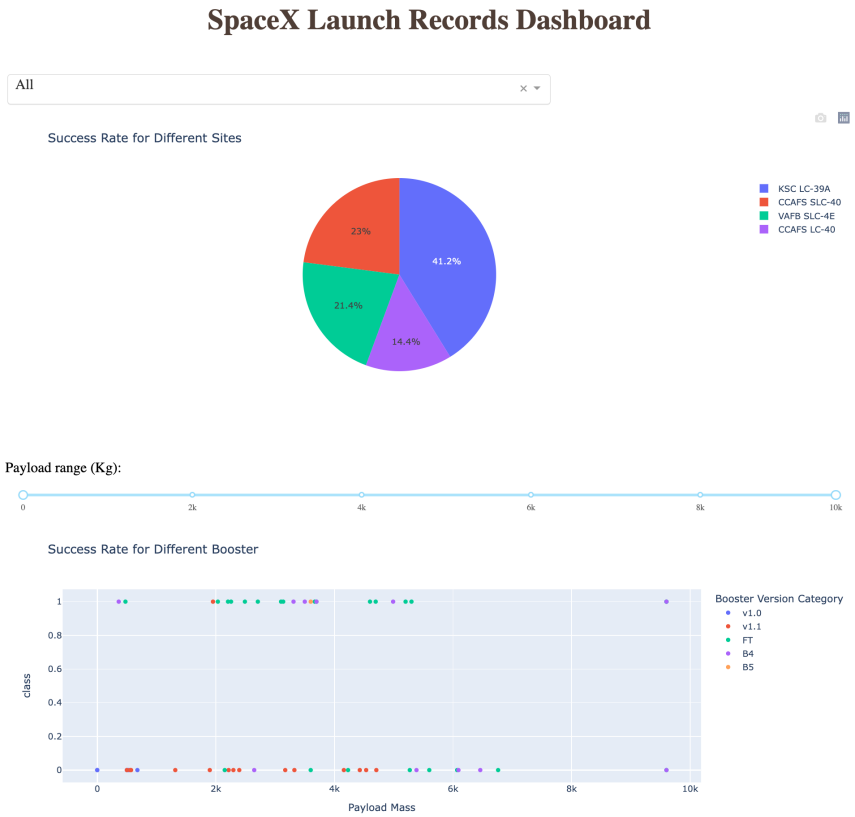


46 launch sites in Florida.

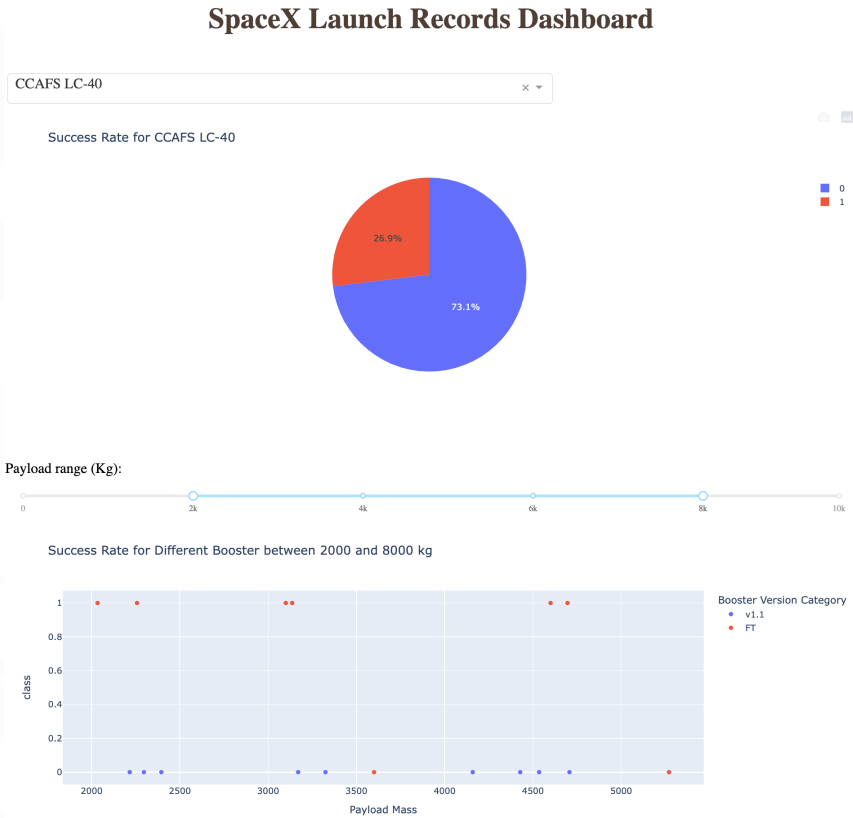


Launch sites is closet to highway, in this example, this specific site is close to Samuel C Philips Pkwy

Plotly Dash Results

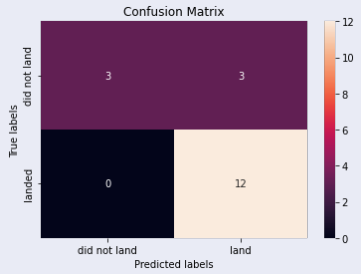
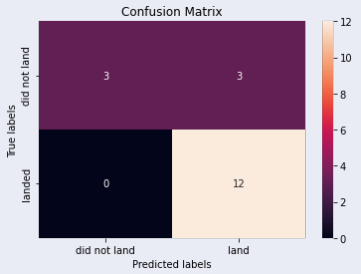
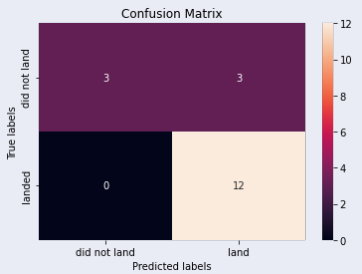


With all launch sites and payload mass range from 0 to 10k



With launch site CCAFS LC-40 and payload mass range from 2k to 8k

Predictive Analysis Results

	Logistic Regression	SVM	Decision Tree	KNN
Best Parameters	'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'	'C': 1.0, 'gamma': 0.031622776601683 79, 'kernel': 'sigmoid'	'criterion': 'entropy', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'	'algorithm': 'auto', 'n_neighbors': 1, 'p': 1
Best score	0.84643	0.84821	0.87679	0.9
Score	0.83334	0.83334	0.83334	1.0
Confusion matrix				

DISCUSSION

- If SpaceX wants to decide the next launch site, where should the location be?
- Why is that most payload mass fall into the 2k to 6k Category?

Conclusions

- Most launch sites locate in **Florida and Los Angeles.**
- The landing success rate is **increasing** over the years.
- In order to predict the landing outcome, **KNN** is the optimum algorithms.
- Common payload mass range from **2k to 6k**. Average payload mass is **3k**.