Winning Space Race with Data Science

Dmytro Bogdan
19 March 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- SpaceX has made their launcher Falcon 9 reusable
- This led to a reduction in cost per kg.
- They are still not as reliable as Soyuz or Ariane-5.
- Falcon9 booster's successful recovery depends on features such as:
    - Orbit
    - Payload mass
    - Booster versions
    - The geographical location of launching sites
- Based on these features, the best Machine Learning supervised classification model developed in this report, predicted booster recovery outcomes with an accuracy close to 94%.

Introduction

- Project background and context:
  - SpaceX is a successful private manufacturer of rockets
  - Falcon 9 is meant to be least expensive: launches at $62 million—$100 million less than some competitors
  - Re-usability of the expensive Stage 1 rockets assist should make it much cheaper
  - Landing success of Stage 1 rockets correlates with payload, orbit, booster versions and launch sites.
- Problems to find answers to
  - To compete with SpaceX, we wish determine the price of each launch. Since it is based on re-usability of the Stage 1 rochet, we will use machine learning to predict whether Stage 1 rockets will land successfully based on publicly available data.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - **Data collection from open Data base and Wikipedia (Falcon9, Ariane-5)**
- Perform data wrangling
  - Handled missing values and derived new column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - **Classification Models development and validations. Selection of best predictive model.**

Data Collection – Wiki and SpaceX API

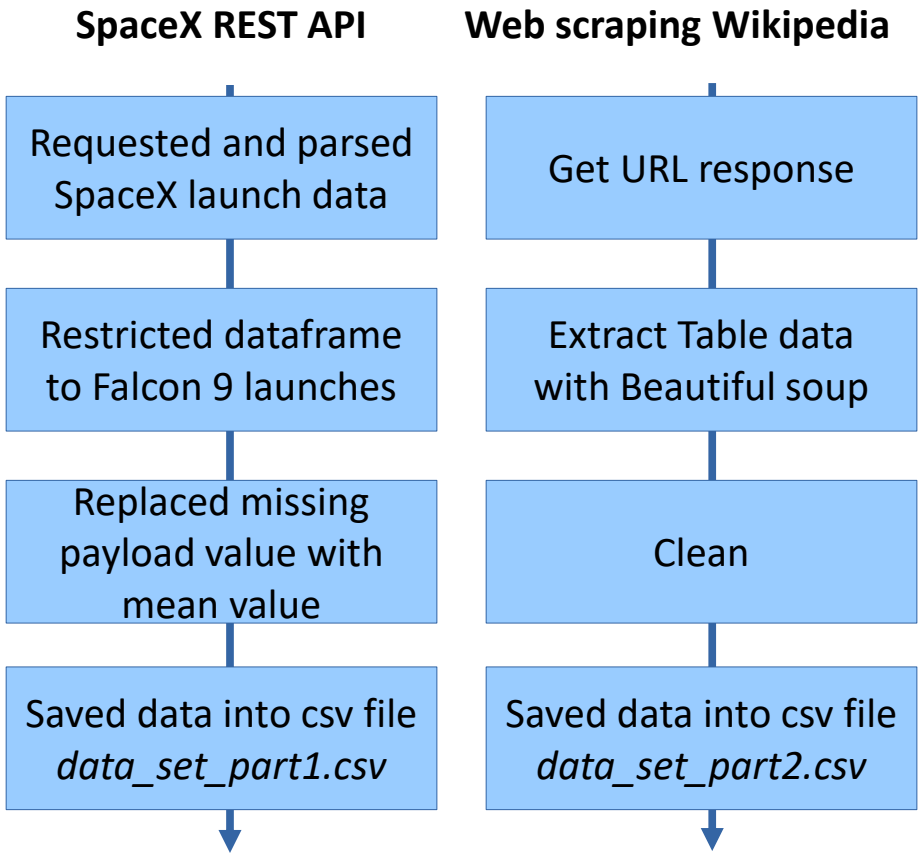
Data was collected from:

-open source SpaceX REST API

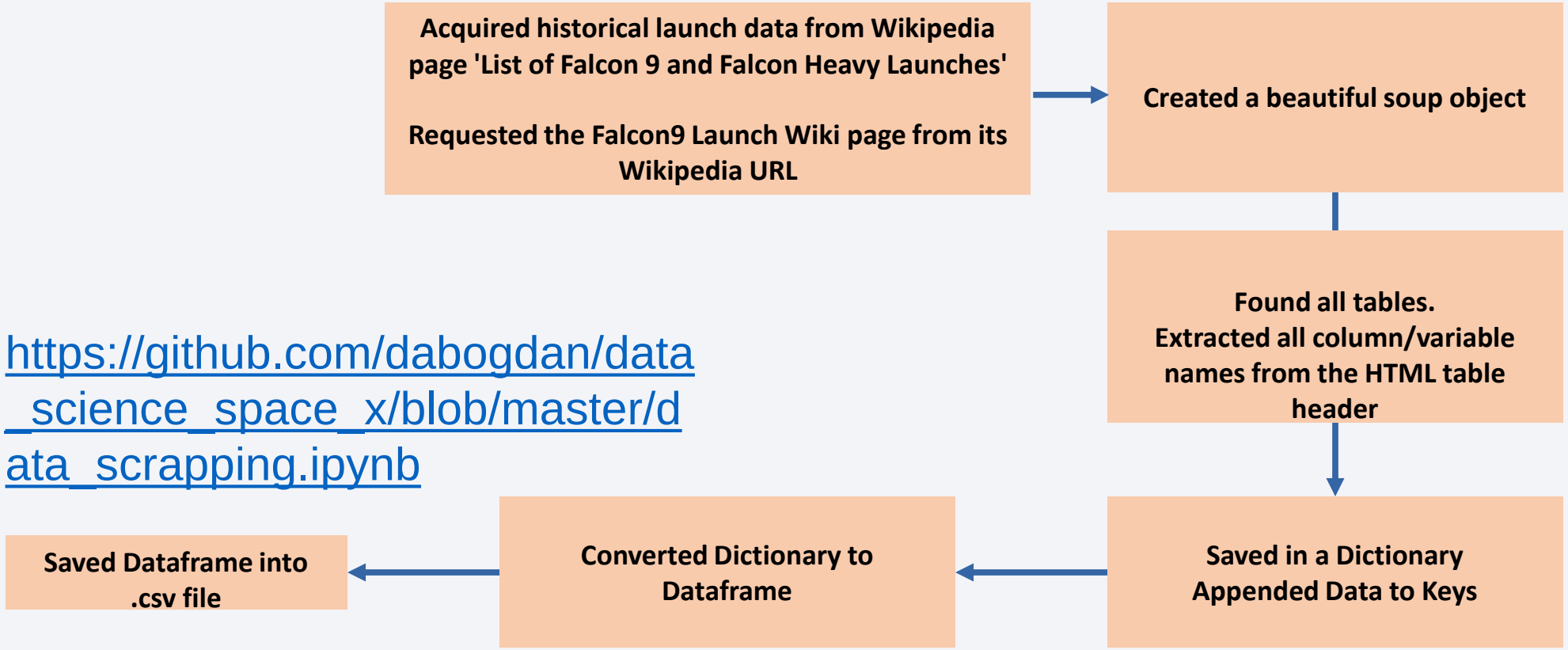
-webscraping Falcon9 launch data in Wikipedia

-webscraping Ariane5 launcher data in Wikipedia ***for reference graphs only.*** (not prepared for ML)

-



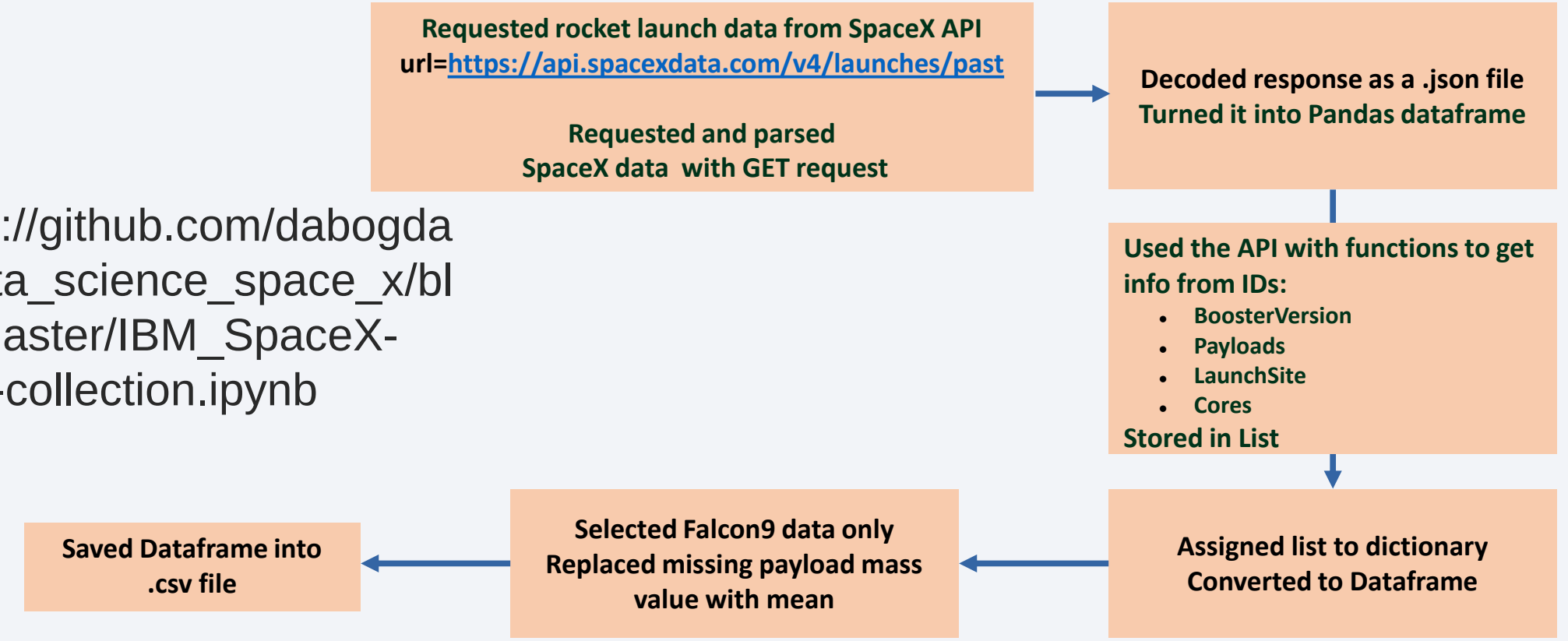
Data Collection – WIKI



- https://github.com/dabogdan/data_science_space_x/blob/master/data_scrapping.ipynb

Data Collection - Scraping

- https://github.com/dabogdan/data_science_space_x/blob/master/IBM_SpaceX-data-collection.ipynb



Data Wrangling

https://github.com/DrStef/Applied_Data_Science_Capstone/blob/main/GH_labs-jupyter-spacex-Data%20wrangling_v3.ipynb

Dataframe
From SpaceX API

Identified missing values → Replaced missing “PayloadMass” value with mean value

SpaceX dataset
Further Data wrangling

Identified:
- null values for each feature
- numerical and categorical features

Calculated:
- Number of launches on each site
- number and occurrence of each orbit
- number and occurrence of missions outcome per orbit type

Created a set of 1 stage booster landing outcomes
0 True ASDS: successful landing on a drone ship
1 None None: failure to land
2 True RTLS: successful landing to a ground pad
3 False ASDS: failed landing on a drone ship
4 True Ocean: successful landing, specific region of the ocean
5 False Ocean: failed landing, specific region of the ocean
6 None ASDS: failure to land
7 False RTLS: failed landing to a ground pad

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Class
49	50	2018-05-11	Falcon 9	3750.00	GTO	KSC LC 39A	True ASDS	1
47	48	2018-04-02	Falcon 9	2760.00	ISS	CCAFS SLC 40	None None	0
50	51	2018-06-04	Falcon 9	5383.85	GTO	CCAFS SLC 40	None None	0
44	45	2018-01-31	Falcon 9	4230.00	GTO	CCAFS SLC 40	True Ocean	1
11	12	2015-01-10	Falcon 9	2395.00	ISS	CCAFS SLC 40	False ASDS	0

Created a Training label: 'Class'
Class = 0: booster landing failure
Class = 1: booster landing success

Data Wrangling

https://github.com/DrStef/Applied_Data_Science_Capstone/blob/main/GH_labs-jupyter-spacex-Data%20wrangling_v3.ipynb

Dataframe
From SpaceX API

Identified missing values → Replaced missing "PayloadMass" value with mean value

SpaceX dataset
Further Data wrangling

Identified:
- null values for each feature
- numerical and categorical features

Calculated:
- Number of launches on each site
- number and occurrence of each orbit
- number and occurrence of missions outcome per orbit type

Created a set of 1 stage booster landing outcomes
0 True ASDS: successful landing on a drone ship
1 None None: failure to land
2 True RTLS: successful landing to a ground pad
3 False ASDS: failed landing on a drone ship
4 True Ocean: successful landing, specific region of the ocean
5 False Ocean: failed landing, specific region of the ocean
6 None ASDS: failure to land
7 False RTLS: failed landing to a ground pad

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Class
49	50	2018-05-11	Falcon 9	3750.00	GTO	KSC LC 39A	True ASDS	1
47	48	2018-04-02	Falcon 9	2760.00	ISS	CCAFS SLC 40	None None	0
50	51	2018-06-04	Falcon 9	5383.85	GTO	CCAFS SLC 40	None None	0
44	45	2018-01-31	Falcon 9	4230.00	GTO	CCAFS SLC 40	True Ocean	1
11	12	2015-01-10	Falcon 9	2395.00	ISS	CCAFS SLC 40	False ASDS	0

Created a Training label: 'Class'
Class = 0: booster landing failure
Class = 1: booster landing success

EDA with Data Visualization

- We visualized the following:
  - **relationship between Flight Number and Launch Site with a scatter point chart**
  - **relationship between Payload and Launch Site (scatter point chart)**
  - **relationship between success rate of each orbit type (bar chart)**
  - **relationship between FlightNumber and Orbit type (scatter point chart)**
  - **relationship between Payload and Orbit type (scatter point chart)**
  - **launch success yearly trend (line chart)**
- https://github.com/dabogdan/data_science_space_x/blob/master/EDA_data_vizualisation.ipynb

EDA with SQL

- Summary of the SQL queries performed:
  - names of the unique launch sites in the space mission
  - for records where launch sites begin with the string 'KSC’
  - total payload mass carried by boosters launched by NASA (CRS)
  - average payload mass carried by booster version F9 v1.1
  - date where the succesful landing outcome in drone ship was acheived.
  - names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
  - total number of successful and failure mission outcomes
  - names of the booster_versions which have carried the maximum payload mass
  - records which will display the month names, succesful landing_outcomes in ground pad, booster versions, launch_site for the months in year 2017
  - count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order
- https://github.com/dabogdan/data_science_space_x/blob/master/EDA-SQL.ipynb

Build an Interactive Map with Folium

- Summary of map objects such as markers, circles, lines added to a folium map:
  - all launch sites on a map
  - success/failed launches for each site on the map
  - distances between a launch site to its proximities
- Reasoning:
  - The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and we could discover some of the factors by analyzing the existing launch site locations.
- https://github.com/dabogdan/data_science_space_x/blob/master/IBM_SpaceX_presentation_Launch_site_locations.ipynb

Build a Dashboard with Plotly Dash

- We have added the plots/graphs and interactions to a dashboard:
  - Select site with dropdown menu
  - Select payload range with slider
  - Scatterplot shows correlation between payload and success
  - Pie chart shows proportion of successful launches
- https://github.com/dabogdan/data_science_space_x/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Summary of how we built, evaluated, improved, and found the best performing classification model:
  - Convert target column into numpy array Y
  - Standardize the data in X
  - Split into training and test sets
  - For each model:
    - a.Set search parameters
    - b.Create GridSearchCV
    - c.Fit model with training data
    - d.Identify best parameters
    - e.Check accuracy on training data
    - f.Calculate accuracy on test data
    - g.Compare predictions with actual scores
- https://github.com/dabogdan/data_science_space_x/blob/master/IBM_SpaceX_Machine%20Learning%20Prediction.ipynb

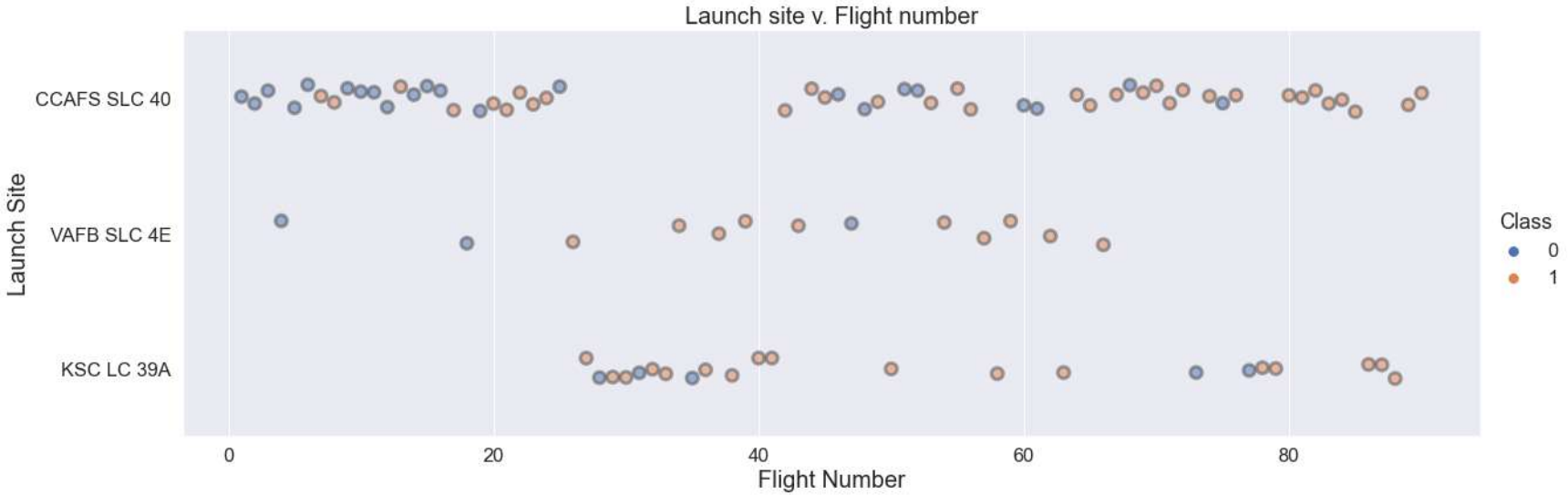
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Launch Site v. Flight Number

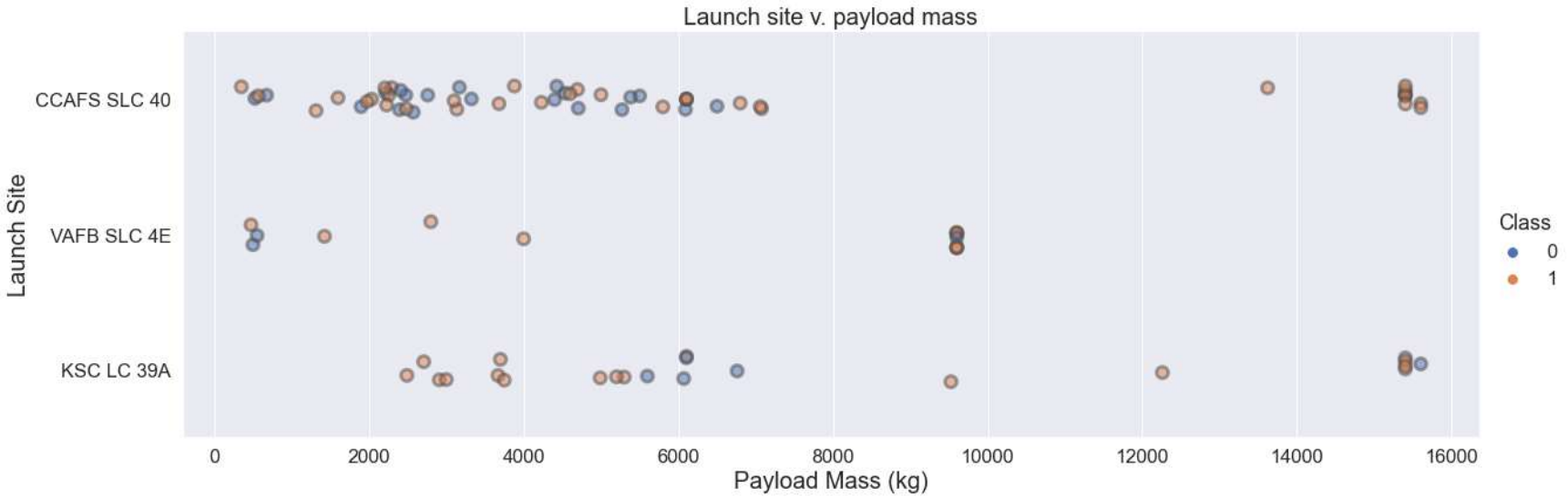


The chart displays valuable info about:

- Chronology: flight numbers
- Number of flights per launch site
- Success/Failure per launch site

- Cape Canaveral CCAFS-SLC 40 is the most used launch site.
- CCAFS-SLC 40 concentrates most of failures , particularly **in the early stage of Falcon9 project**.
- Given CCAFS-SLC 40 southern location, most “risky” GTO and GEO launches may take place there.
- Additional info needed: orbit, payload mass

Launch Site v. Payload

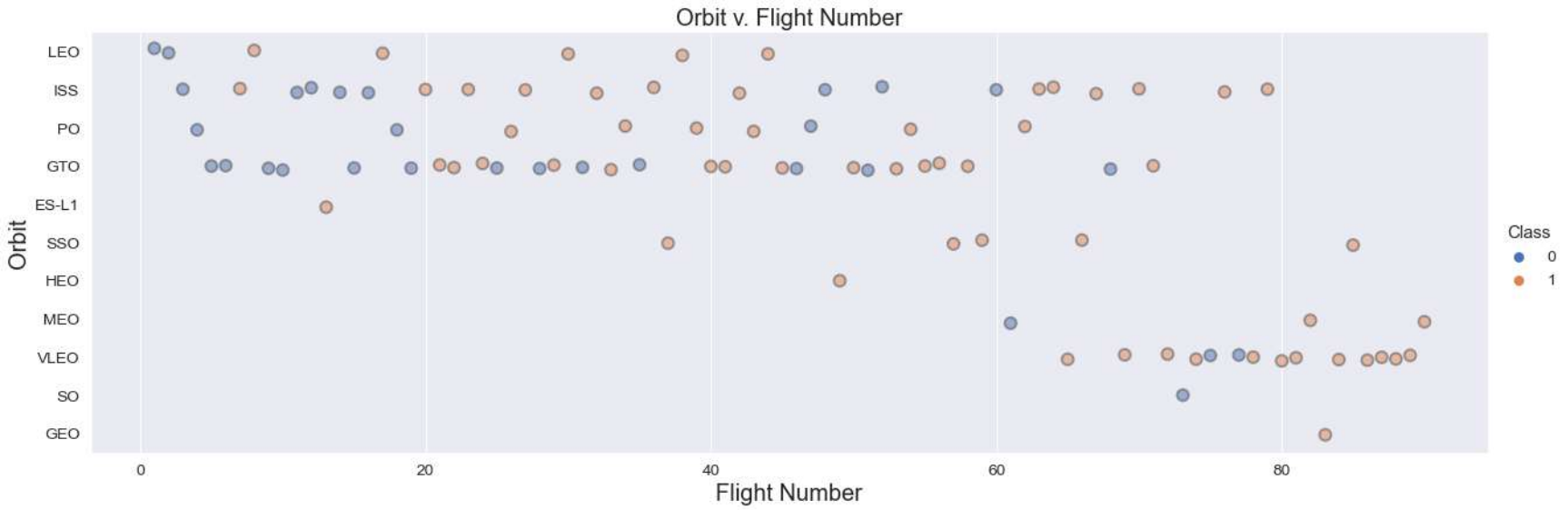


The chart brings additional info:

- Payload mass per launch site
- Success/Failure per payload mass

- Given Falcon9 specifications, heavy payloads > 10000 kg are sent to low/medium orbits LEO/MEO only.
- It looks like the percentage of failures is lower for heavy payload. Which would indicate that low orbits are less risky to the success of the mission (recovery of booster).
- Light payloads are not necessarily all sent to GTO/GEO.
- More information is needed for extracting some correlation: success rate v. payload/orbit

Orbit Type v. Flight Number



The chart brings additional info:

- Number of flights per Orbit.
- Success rate per orbit

- The number of flights for: GEO, SO, HEO, ESL-1, MEO is not significant for concluding about success rate.
- PO, SSO, ISS, VLEO are low orbits
- GTO is a transfer orbit to GEO.

It looks like GTO are higher risk missions, low orbits are lower risk.

We confirm with the following histogram.

Success Rate vs. Orbit Type

Remarks:

- GTO is a transfer orbit to GEO. Low thrust engines of the payload (satellite) complete the orbiting phase.
- We ignore results: GEO, SO, HEO, ESL-1, MEO. The number of flights is not significant.

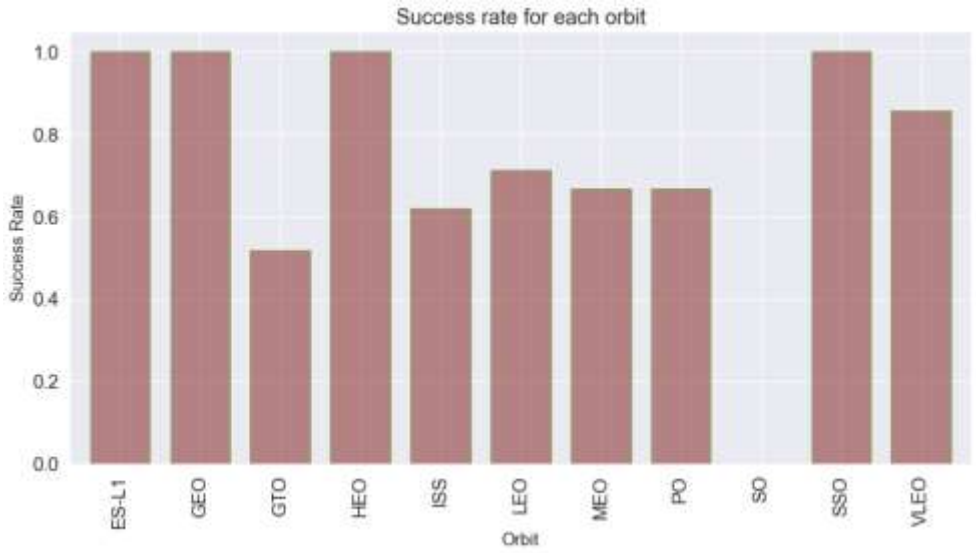
GTO sees the lowest success rate as suggested in previous slide.
SSO (polar low orbit) **the highest one.**

Success rate may strongly depend on both:

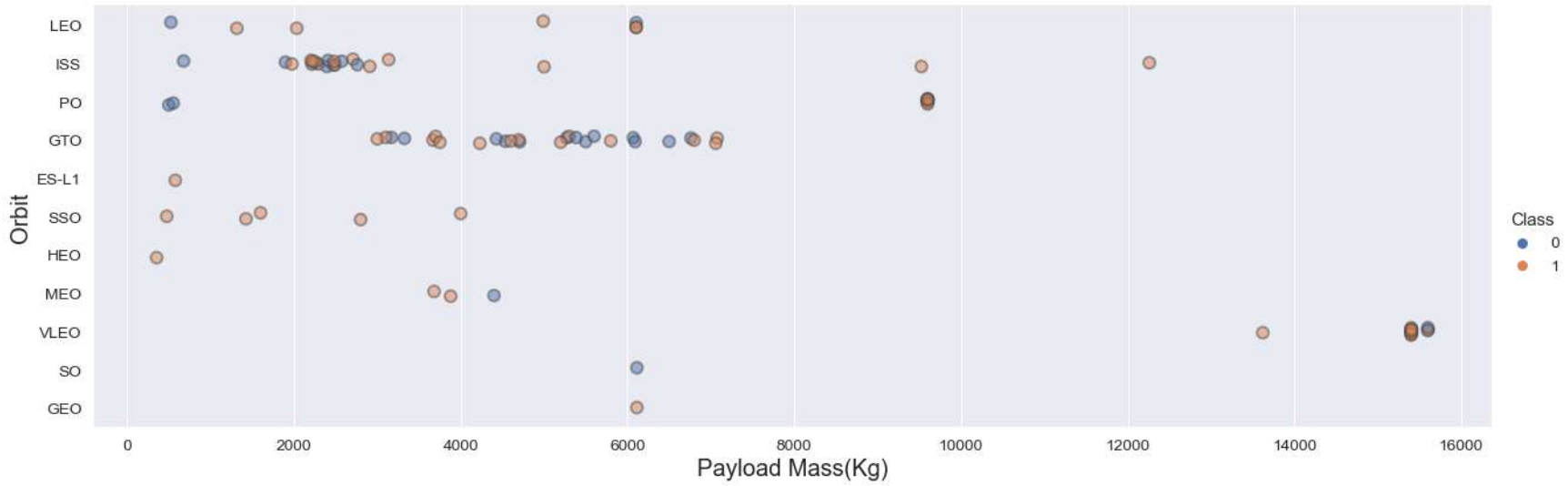
- payload mass
- orbit.

meaning the amount of energy deployed at lift-off, that may induce ***strong noise/vibrations that are known to damage satellites****.
Vibrations could damage some of the booster electronics, inertial guidance systems... and cause booster recovery/landing failure.
We also need additional info about payload mass/orbit. Fortunately it is available.

* https://adsabs.harvard.edu/full/1996ESASP.386..237F



Orbit Type v. Payload



The chart brings final info about "Orbit v. Payload". It describes the distribution "success rate v. (payload, orbit)"
Main trends:

- Maximum success rate with: low orbit except (ISS) and low payload mass
- ISS: based on "Orbit Type v. Flight Number" 5/8 failures occurred in the early stage of Falcon 9 project. When Falcon 9 reliability was low.
- Between 2000 and 7500 kg, success rate seems to be evenly distributed for GTO.
- Independently of payload mass, GTO is a risky "orbit" affecting missions success rate. Falcon 9 reliability improves over time, but there are still recent failed booster recovery after GTO launches.

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes (from SQL queries).

Here success is defined based on properly launching/orbiting payload. Success rate is very high: ~99% like Ariane-5.

Nevertheless, Falcon9 maintains a competitive advantage in terms of cost per kg compared with classic launchers like Ariane-5, **only if the reusable booster is recovered.**

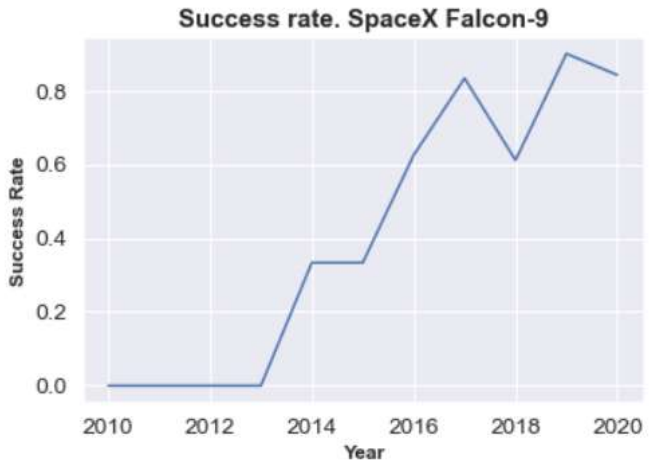
Therefore Falcon9 “success rate” in this report is defined after successful booster recovery (landing).

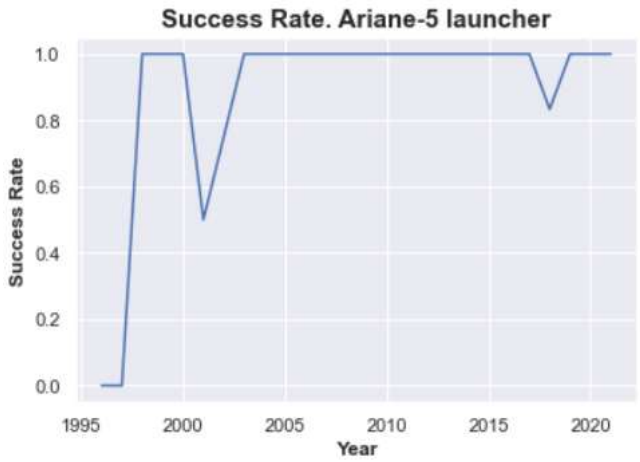
```
#  sql query
qsf= """Select (Select Count(Mission_Outcome) from spacex_v11 where Mission_Outcome like '%Success%')
        as Successful_Missions,
        (Select Count(Mission_Outcome) from spacex_v11
        where Mission_Outcome like '%Failure%') as Failed_Missions """


success_failure= pd.read_sql_query(qsf,conn)
print(success_failure)
```

```
   Successful_Missions  Failed_Missions
0                  100                1
```

Launch Success Yearly Trend





Falcon 9 reliability significantly improves over time .

Success rate, **here defined after successful booster recovery for Falcon9,** depends on:

- Payload mass
- Orbit
- + other factors we investigate next

- Independently of payload mass, orbits, **Ariane 5 has a close to 100% success rate** for 82 flights since 2003.
- Falcon9 average booster recovery success rate si 66%.
- Success rate currently sufficient for SpaceX financial viability.

All Launch Site Names

Before starting launch sites analysis, we list the names of all launch sites and some launch records (from SQL queries).

```
df_unique_launchsites=pd.read_sql_query("Select distinct Launch_Site  from spacex_v11 ",conn)
print(df_unique_launchsites)
```

```
    Launch_Site
0   CCAFS LC-40
1   VAFB SLC-4E
2    KSC LC-39A
3  CCAFS SLC-40
```

There are 4 distinct launch sites

5 records where launch sites begin with `CCA`

```
df_launchsites_CCA5=pd.read_sql_query("Select * from spacex_v11 where Launch_Site Like 'CCA%' Limit 5",conn)
df_launchsites_CCA5
```

	id	Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	1	2010-04-06	0 days 18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
1	2	2010-08-12	0 days 15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of...	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2	3	2012-05-22	0 days 07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
3	4	2012-10-08	0 days 00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
4	5	2013-03-01	0 days 15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

```
#  For validation purposes... sum in df_NASA_CRS  'PAYLOAD_MASS__KG_'   column
df_NASA_CRS=pd.read_sql_query("Select * from spacex_v11  where Customer='NASA (CRS)'",conn)
print(df_NASA_CRS.head(2))
print('----')
print('Total payload mass, customer= NASA (CRS):', df_NASA_CRS['PAYLOAD_MASS__KG_'].sum(),' kg')
```

```
   id        Date      Time (UTC) Booster_Version  Launch_Site       Payload  \
0   4  2012-10-08 0 days 00:35:00  F9 v1.0  B0006  CCAFS LC-40  SpaceX CRS-1
1   5  2013-03-01 0 days 15:10:00  F9 v1.0  B0007  CCAFS LC-40  SpaceX CRS-2

   PAYLOAD_MASS__KG_      Orbit    Customer Mission_Outcome Landing_Outcome
0                500  LEO (ISS)  NASA (CRS)         Success      No attempt
1                677  LEO (ISS)  NASA (CRS)         Success      No attempt
----
Total payload mass, customer= NASA (CRS): 45596  kg
```

```
#  Based on SQL only...
sql_nasa_crs_mass= """ Select sum(PAYLOAD_MASS__KG_) as 'Total payload mass (kg) NASA CRS'
                    from spacex_v11
                    where Customer='NASA (CRS)' """
payload_NASA_CRS=pd.read_sql_query(sql_nasa_crs_mass,conn)
print(payload_NASA_CRS)
```

```
   Total payload mass (kg) NASA CRS
0                           45596.0
```

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.

```
payload_F9v11=pd.read_sql_query("Select avg(PAYLOAD_MASS__KG_) as 'avg mass (kg)' from spacex_v11 where Booster_Version='F9 v1.1'",conn)
print(payload_F9v11)
```

```
   avg mass (kg)
0         2928.4
```

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad

```
min_Date_success_landing=pd.read_sql_query("select min(Date) from spacex_v11 where Landing_Outcome = 'Success (ground pad)'",conn)
print(min_Date_success_landing)
```

```
    min(Date)
0  2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

- Present your query result with a short explanation here

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Recent Full Thrust (FT) boosters exhibit the highest success rate on drone ship landing. Including with GTO flights.
- It bodes well if SpaceX introduces a concept like “Sea Launch” for GTO launch

```
#  sql query
q_boost_succ= """ select  Booster_Version from spacex_v11 where Landing_Outcome = 'Success (drone ship)'
                  and PAYLOAD_MASS__KG_ > 4000
                  and PAYLOAD_MASS__KG_ < 6000 """
```

```
Booster_success_landing=pd.read_sql_query(q_boost_succ,conn)

print(Booster_success_landing)
```

```
  Booster_Version
0     F9 FT B1022
1     F9 FT B1026
2  F9 FT  B1021.2
3  F9 FT  B1031.2
```

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
#  sql query
q_failed_landing= """ Select Date, Booster_Version, Launch_Site, Landing_Outcome
                    from spacex_v11
                    where Landing_Outcome = 'Failure (drone ship)'
                    and Date like '%2015%'    """
```

```
fail_drone= pd.read_sql_query(q_failed_landing,conn)
fail_drone.head(5)
```

	Date	Booster_Version	Launch_Site	Landing_Outcome
0	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
1	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
#  sql query
q_count_landing= """ Select Landing_Outcome, count(*) as count_landings
                     from spacex_v11
                     where Date between '2010-06-04' and '2017-03-20'
                     group by Landing_Outcome
                     order by count_landings desc  """
```

```
count_landing= pd.read_sql_query(q_count_landing,conn)
count_landing.head(10)
```

	Landing_Outcome	count_landings
0	No attempt	10
1	Failure (drone ship)	5
2	Success (drone ship)	5
3	Controlled (ocean)	3
4	Success (ground pad)	3
5	Uncontrolled (ocean)	2
6	Failure (parachute)	1
7	Precluded (drone ship)	1

Section 3

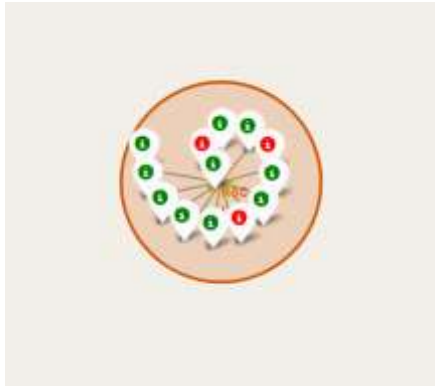
Launch Sites Proximities Analysis

All launch sites

- VAFB SLC-4E: Vandenberg Space Launch Complex 4 (CA)
- KSC-LC29A: Kennedy Space Center - Merritt Island (FL)
- CCAFS-LC40: Cape Canaveral Launch Complex 40 (FL)
- CCAF-SLC40: Cape Canaveral Space Launch Complex 40(FL)

Falcon 9 Success/Failed launches for each site

Vandenberg Space Launch Complex 4 (CA)
VAFB SLC-4E



Kennedy Space Center (FL)
KSC LC 39A

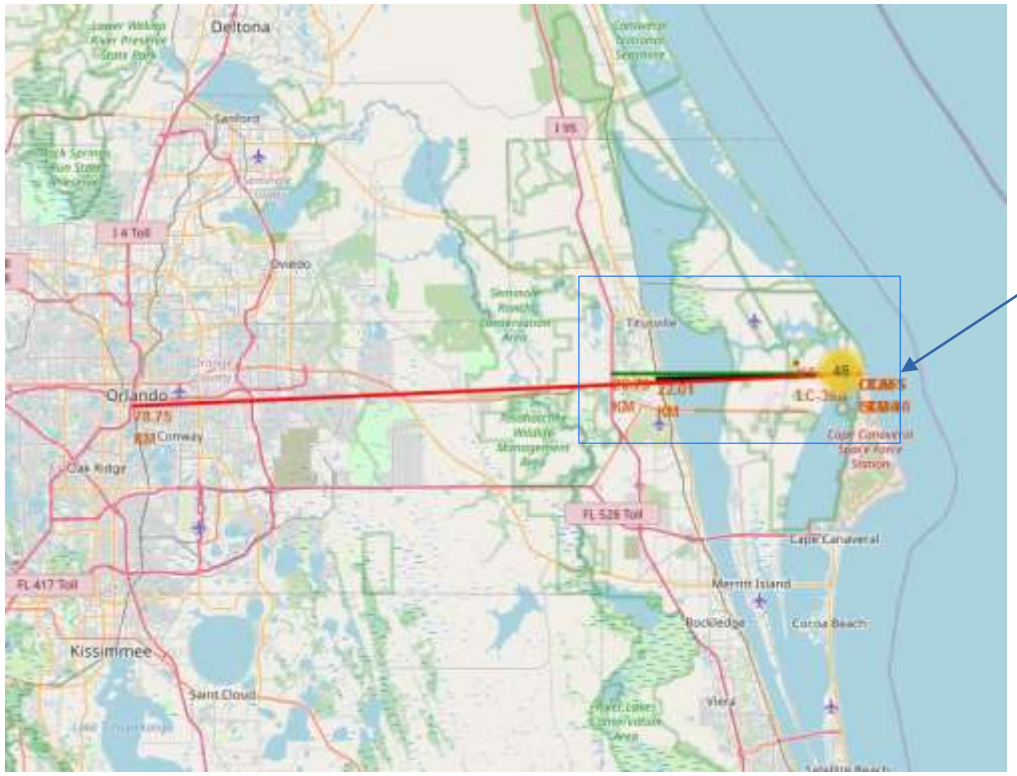
Cape Canaveral (FL)
CCAFS-LC40

Cape Canaveral (FL)
CCAFS-SLC40

Launch Site	class	
CCAFS LC-40	0	19
	1	7
CCAFS SLC-40	0	4
	1	3
KSC LC-39A	0	3
	1	10
VAFB SLC-4E	0	6
	1	4

Table: Synthesis of launches outcomes
Class 0= failure
Class 1= success

Distances between a launch site to its proximities



Distance from CCAFS_SLC40 to:

- **Closest coast: ~900 m**
- **Florida East Coast Railway: 22.0 km**
- **Highway I 95: 26.8 km**
- **Orlando: 78.75 km**

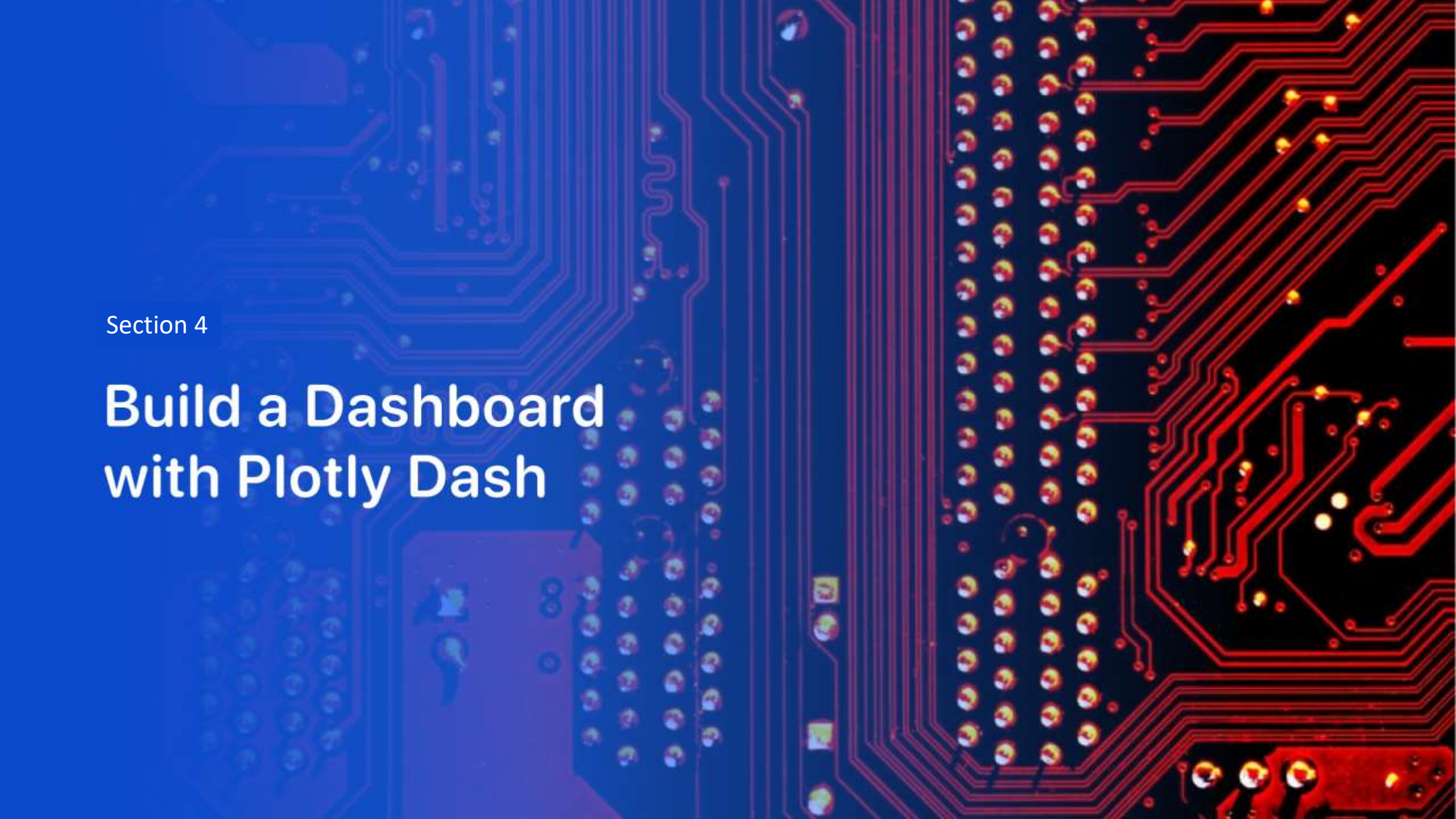
Launch sites are close to coasts. For safety issues if launcher is lost in the early stage of the flight.

Rockets are launched:

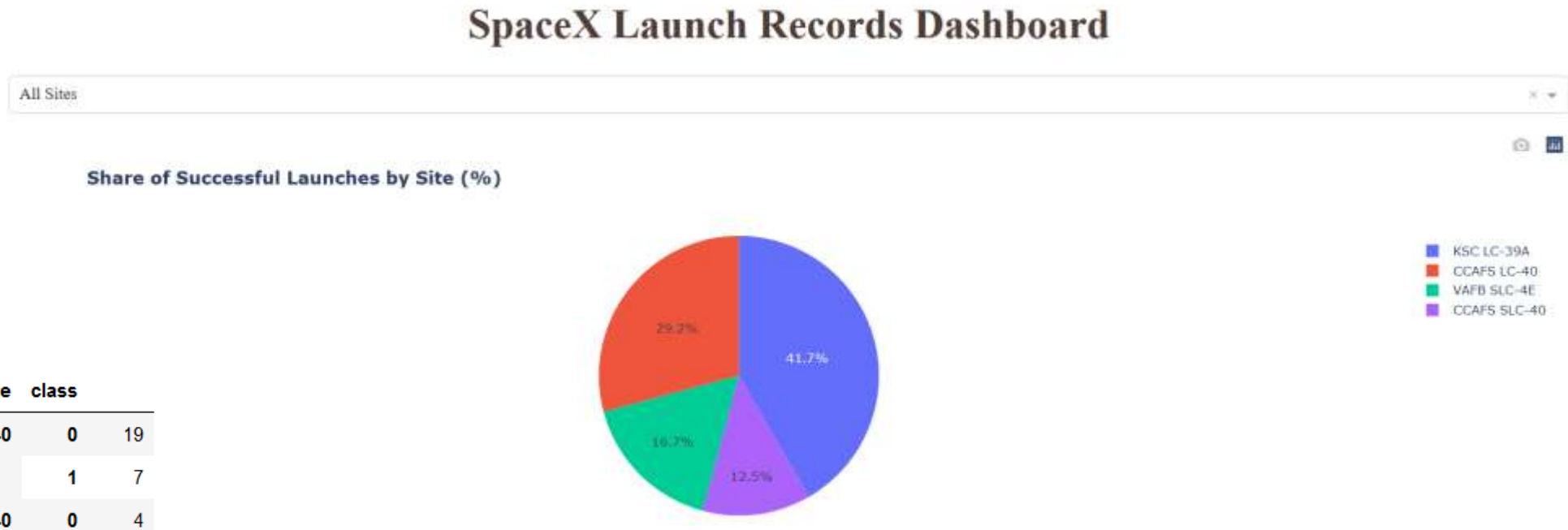
- From West to East over the ocean in Florida.
- North or South bound over the ocean in California. (Polar orbits only)

Launch sites are relatively far from populated areas for protecting population from serious incidents at lift off: explosion on the launch pad.

https://github.com/dabogdan/data_science_space_x/blob/master/IBM_SpaceX_presentation_Launch_site_locations.ipynb


Section 4
Build a Dashboard with Plotly Dash

SpaceX Falcon 9: Launch success count for all sites



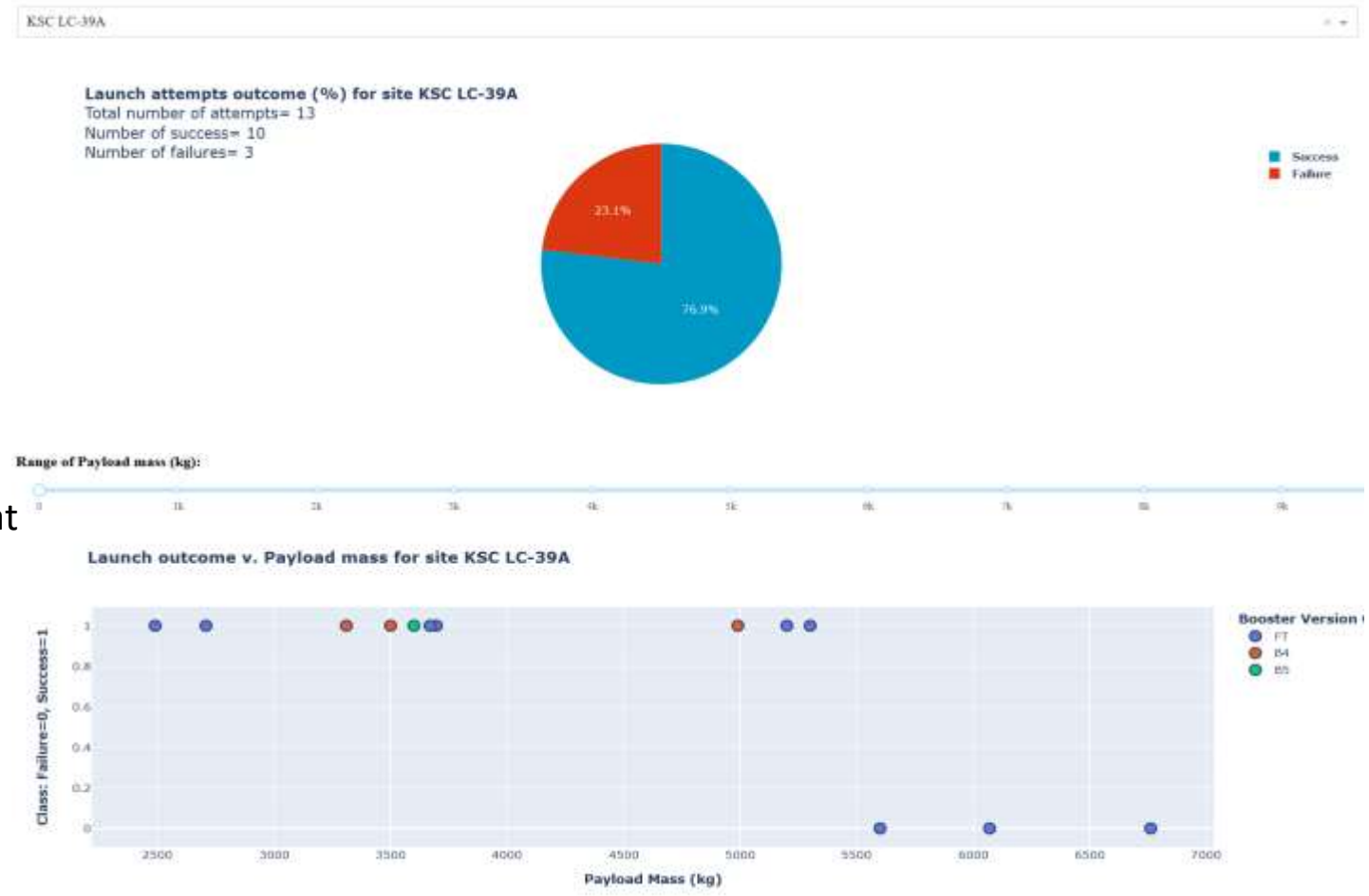
Launch Site	class	
CCAFS LC-40	0	19
	1	7
CCAFS SLC-40	0	4
	1	3
KSC LC-39A	0	3
	1	10
VAFB SLC-4E	0	6
	1	4

The dahsboard allows an interactive visualization and analysis of Falcon successful launches.
It completes scattered charts.
Here for all sites, we can identify the launch site with highest success rate.
Kennedy Space Center in Florida.

SpaceX Falcon9 Launch site with highest launch success ratio

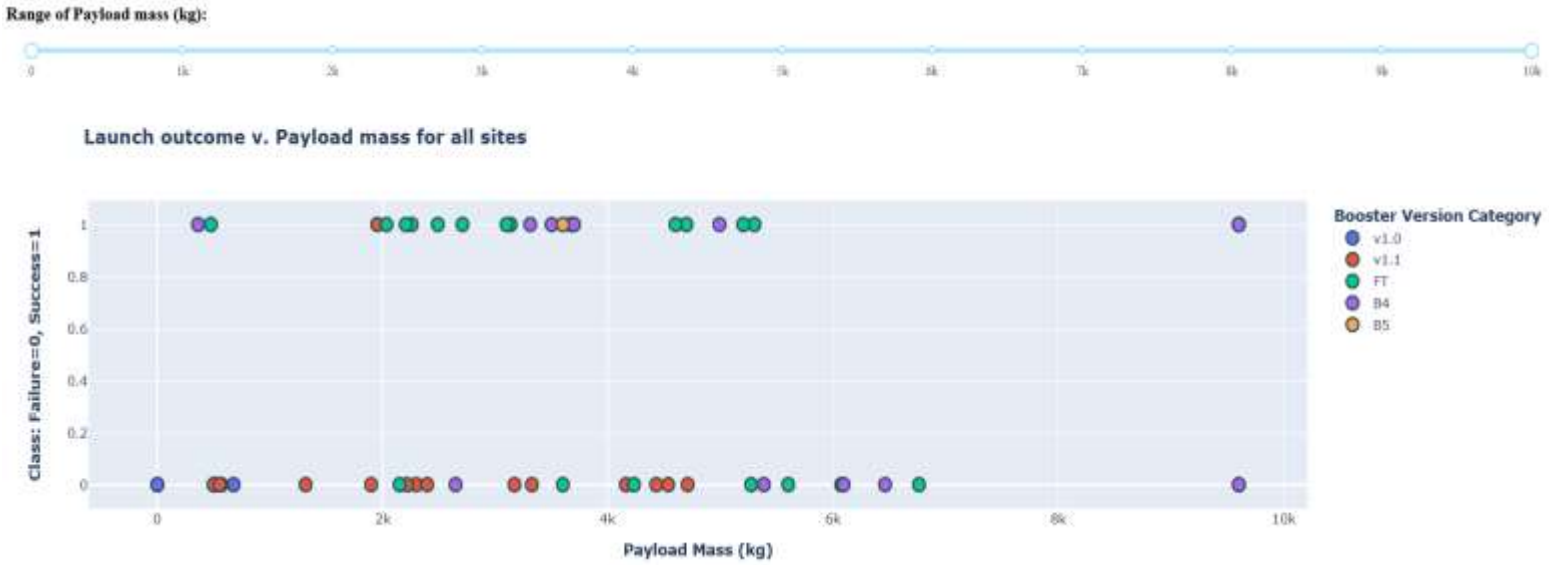
KSC LC-39A
Kennedy Space Center in Florida.
13 flights, 10 successful missions.

- Heavy payload are “high risk”
- Success does not seem to depend upon boosters versions with low mass payload <5500kg.
- B5 and FT are the most reused launchers. Data is not sufficient, but may indicates that they are as reliable as 1 time launchers.



Launch outcome v. Payload mass (all sites)

- V1.0 and v1.1 are early launchers with low reliability.
  Landing legs, were pioneered on the Falcon 9 v1.1 version, but that version never landed intact.
  They were phased out in 2015.
- FT: “Full Thrust” is the next generation and has the highest success rate for payload mass under 6 tons. Including with “drone landing” (see details in next slide).
- Many FT flights are done with reused launchers. And show good reliability.
- Heavy payload are “high risk”.



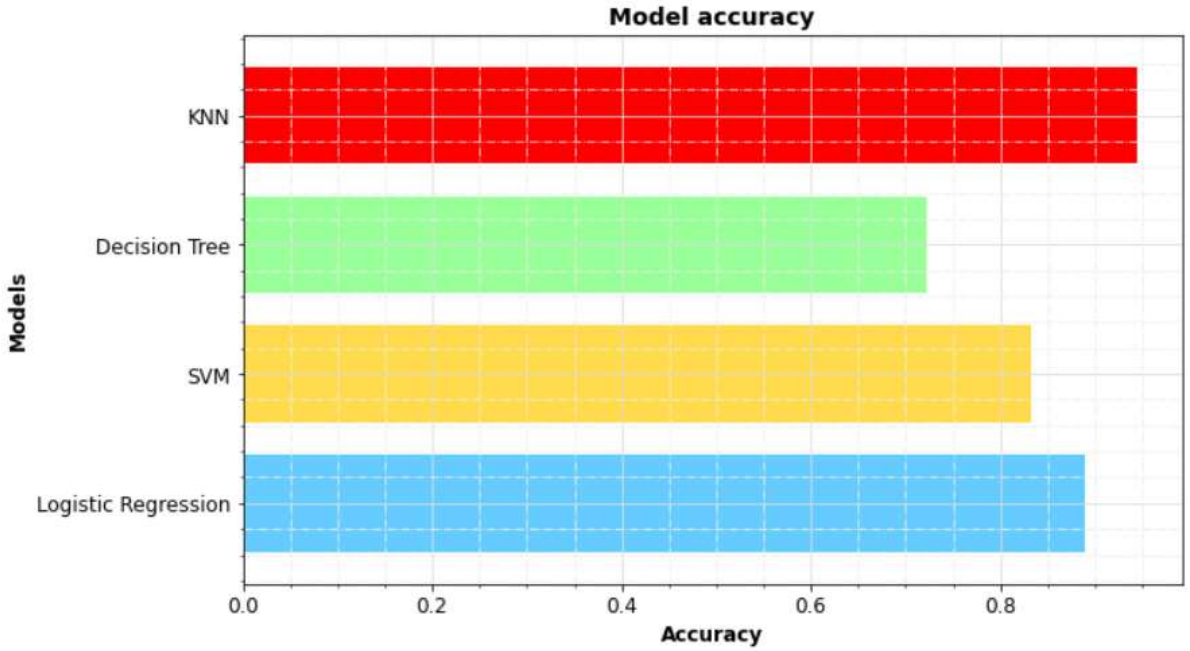
Section 5

Predictive Analysis (Classification)

Classification Accuracy

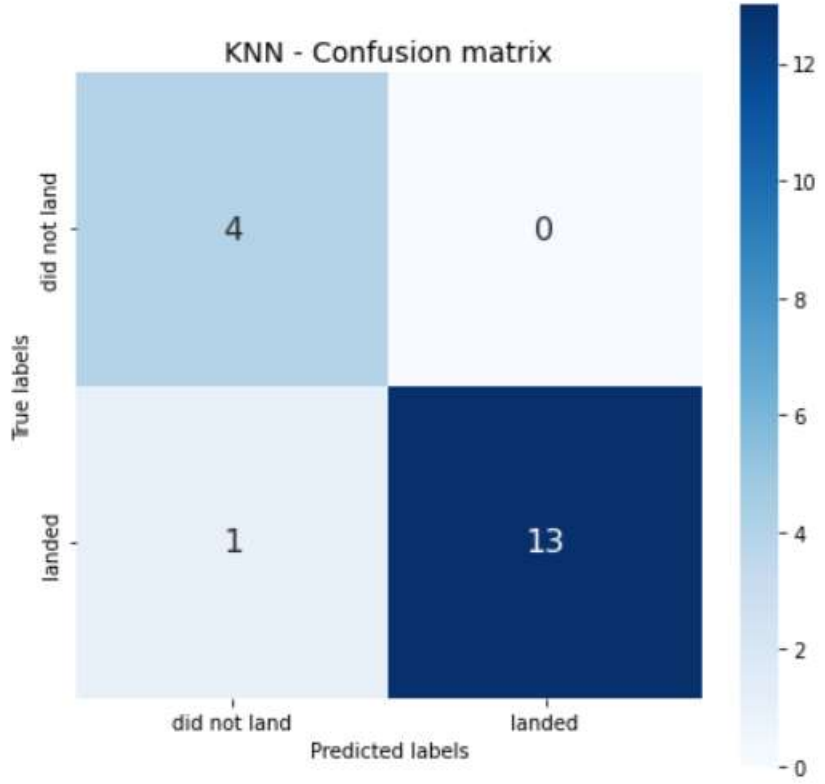
Classification Accuracy with test set.
Results with “train test split” random_state=3

- Optimization of SVM and LR hyper-parameters was refined for increasing accuracy with train set.
-
- It did not necessarily improved accuracy with test set.
- Test set is too small.
- In our case, KNN exhibits the best accuracy: ~94%



Confusion Matrix

- k-nearest neighbors algorithm (k-NN) is the best “predictor”
- The model perfectly predicts mission failure
- 1 false negative for successful booster landing (recovery)



Conclusions

- Constant increase of success booster recovery rate since 2013
- Site KSC LC-39A has greatest success rate
- Optimal payload range: 1,500-4,000Kg
- Optimal classification model: Decision Tree
- Even with recent recovery failure for GTO/GEO, SpaceX will maintain a sufficient lead in terms of cost per kg. vs Ariane5 and the new Ariane 6. Starship may well crush further the competition if similar success rate is achieved, however the success rate will gradually improve.

Thank you!