IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Tetiana K******h>
<31.07.2022>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
  - Insights drawn with EDA
  - Launch site proximities analysis
  - Build a Dashboard with Plotly Dash
  - Predictive analysis (classification)
- Conclusion
- Appendix

Executive Summary

- For solving the problem, we collected data from open sources, performed data wrangling, performed exploratory data analysis using visualization and SQL, performed interactive visual analytics and then made predictive analysis using classification models.
- The predicting landing outcome model was built. The best results were obtained with the decision tree model, its accuracy is 0.8889.

Introduction

- A rocket first stage successful landing and later reusing can reduce the cost of launch. To estimate economics of launching we should be able to estimate if first stage lands successfully.
- On the base of open data, we want built model predicting if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API, web-scraping
- Perform data wrangling
  - Filtering necessary information, renumerating, replacing missing values, calculating labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Standardize data, splitting into train and test set, find hyperparameters for Logistic regression, SVM, decision tree and KNN and then find the best model

Data Collection

- Data were collected from open sources by SpaceX API and web-scraping wikipedia.
- First possible way:
- SpaceX API ==> launches + extra info about launches details ==> filtering and extracting significant ==> building resulting pandas DataFrame
- Second possible way:
- Web-scraping Wikipedia ==> building pandas DataFrame

Data Collection – SpaceX API

- Collected data have information about: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Full process is here Data Collection - SpaceX API

Main data:

https://api.spacexdata.com/v4/launches/past

Extra details:

https://api.spacexdata.com/v4/rockets

https://api.spacexdata.com/v4/launchpads

https://api.spacexdata.com/v4/payloads

https://api.spacexdata.com/v4/cores

Data Collection - Scraping

- Loading wikipedia page and parsing its tables
- Full process is here Data Collection - Scraping

Load web-page ==>

parse tables ==>

fill dataframe with parsed information

Data Wrangling

- Calculate statistics of launches
- Creating labels (0/1 i.e. failure/success) for outcomes

- Full process is here Data Wrangling

EDA with Data Visualization

- We investigate

dependence of outcomes on (Payload, Flight Number), (LaunchSite, Flight Number), (Payload, Launch Site), orbit, (orbit, Flight Number), (Payload, orbit);

success yearly trend,

- Full info is available here EDA with data visualization.

EDA with SQL

Performed SQL queries extract info about:

- launch sites
- payload mass for some boosters and customers
- success rate (total, for different launch pads, for different payload)
- the most heavy loaded boosters

- Full info is available here EDA with SQL

Build an Interactive Map with Folium

- To folium map were added markers for launch site for each launching with success status, and lines for proximity (the best proximity of each type) with distance.
- We use Circle to indicate the position, Marker to sign, Line to show the closest distance and grouped markers to MarkerCluster cause we have many objects with the same coordinates.

- Full info is available here interactive map with Folium map

Build a Dashboard with Plotly Dash

- On dashboard we can see pie chart for success rate for launch sites as well as success status scatter plot on payload mass for given launch site and payload interval. We can see results for different boosters on scatter plot.
- Plotly Dash python application and its input dataset

Predictive Analysis (Classification)

After data standardizing, they were splitting into train and test set.
Then, using GridSearchCV, we found hyperparameters for Logistic regression, SVM, decision tree and KNN.

Then, we find the best model.

Depending on runtime, all the models have the same score, or sometimes the best model was the decision tree.

So, we can think that the best model is decision tree.

- Full info is available here predictive analysis

Results

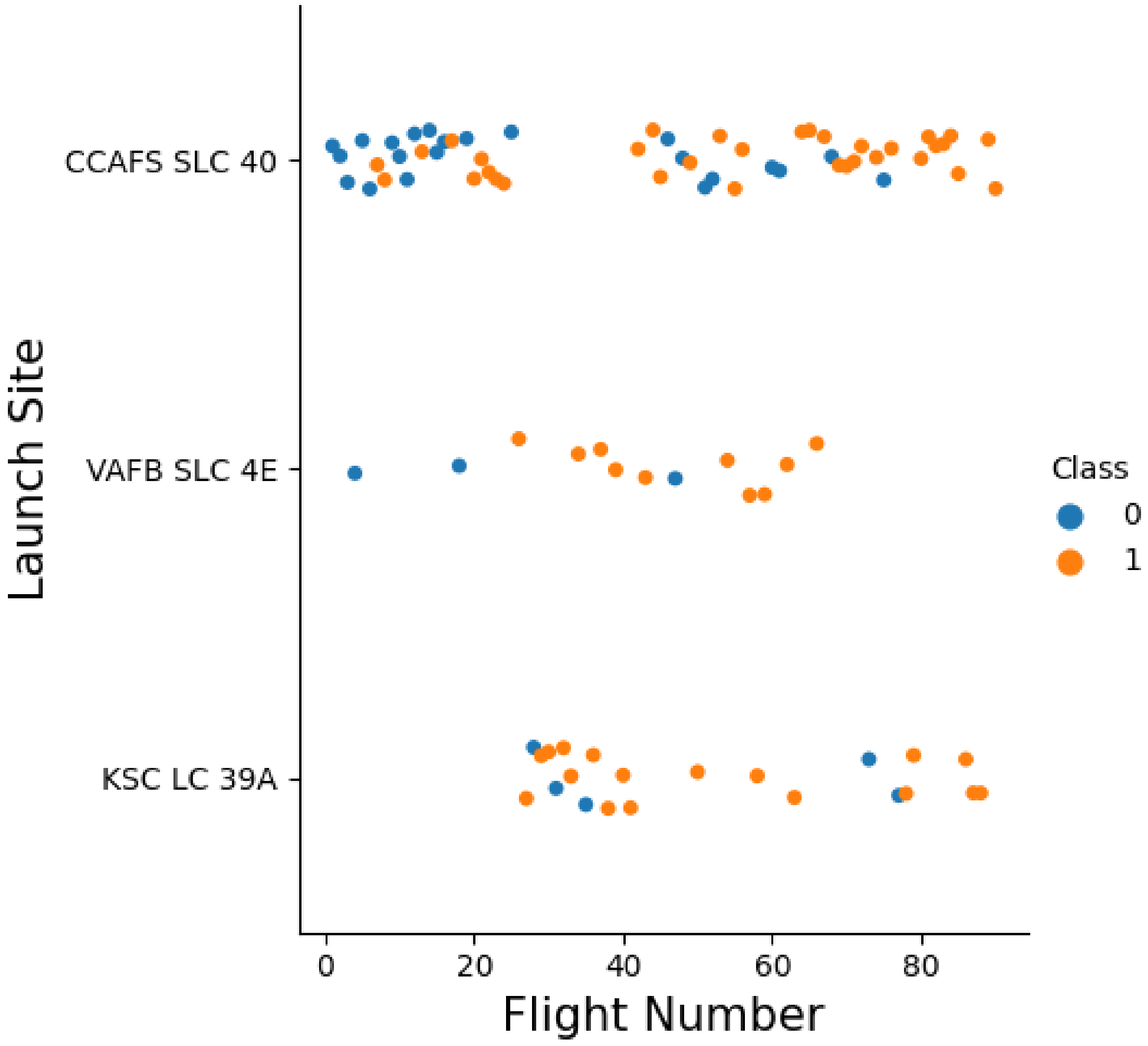
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

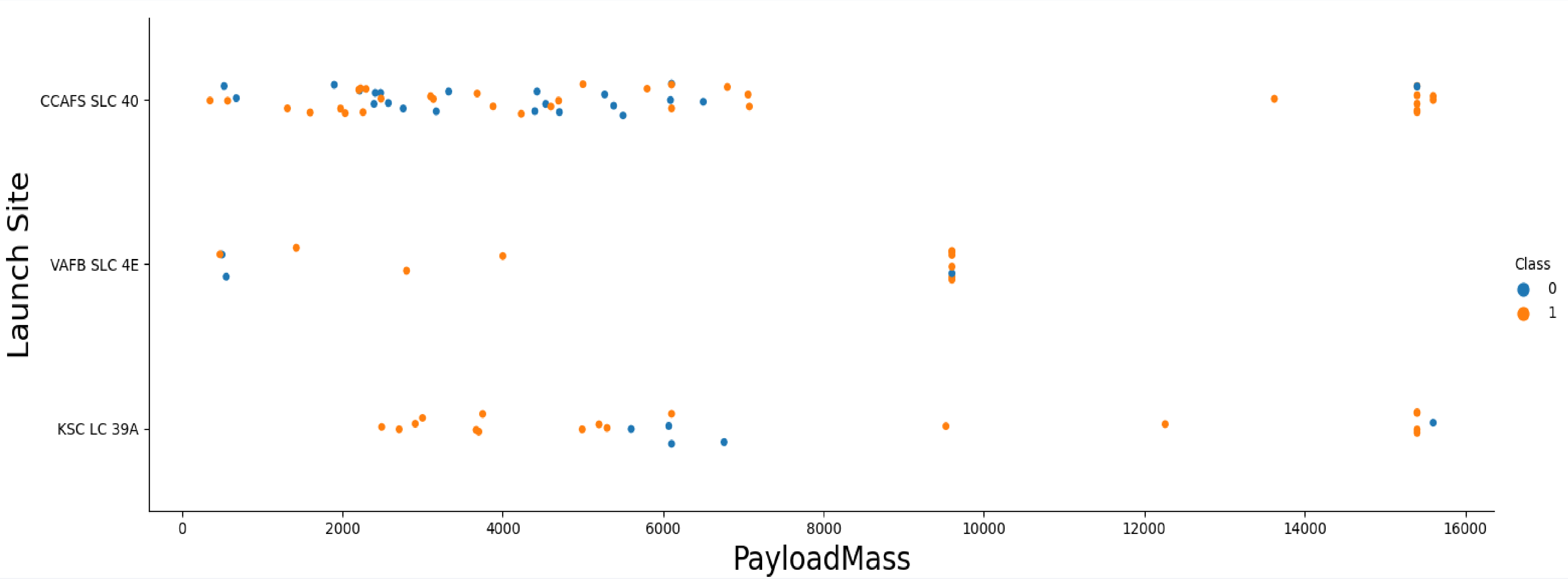
Flight Number vs. Launch Site

- For each launch site the success rate is grow with time



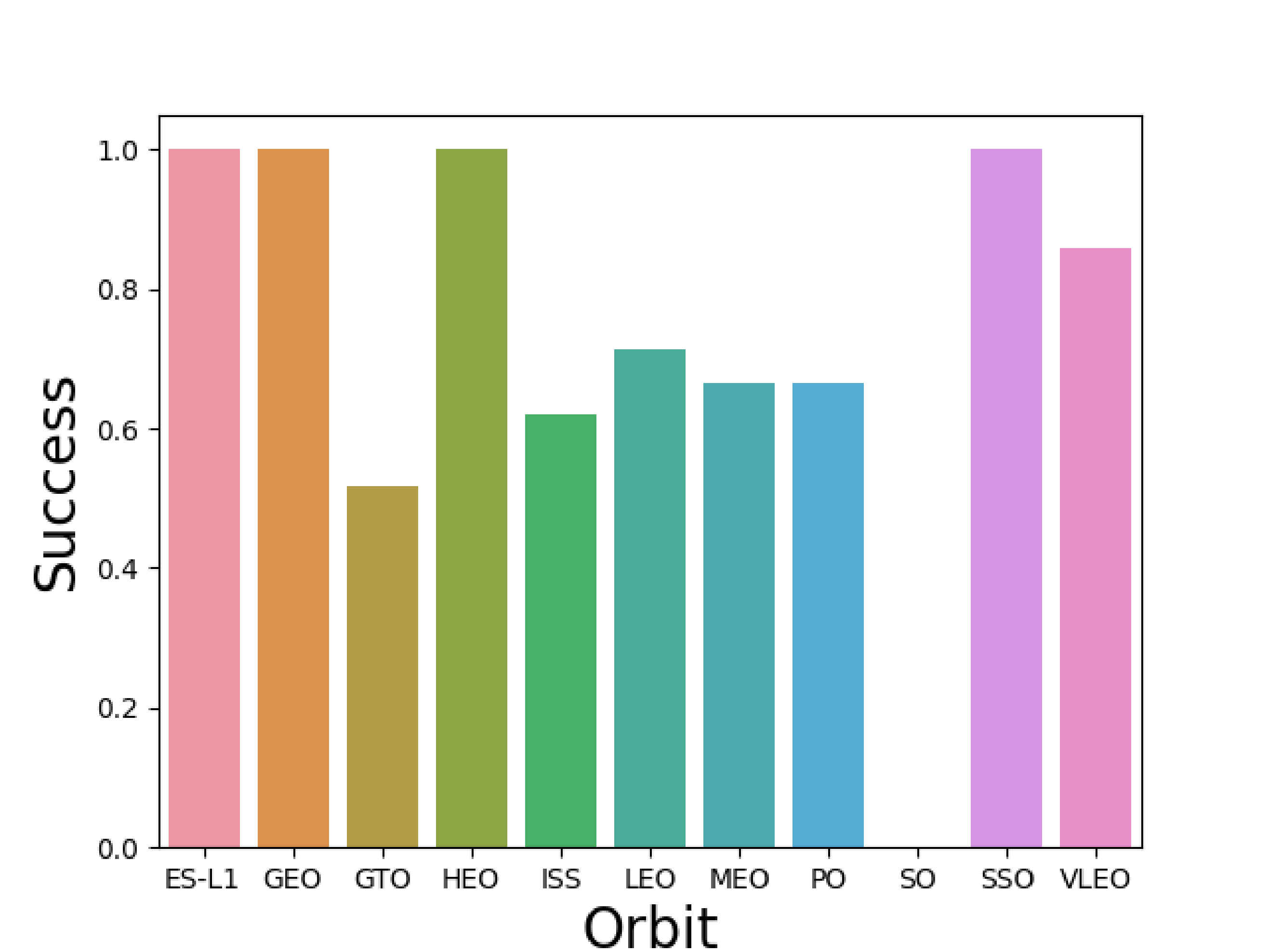
Payload vs. Launch Site

- With increasing payload the success rate is increasing too



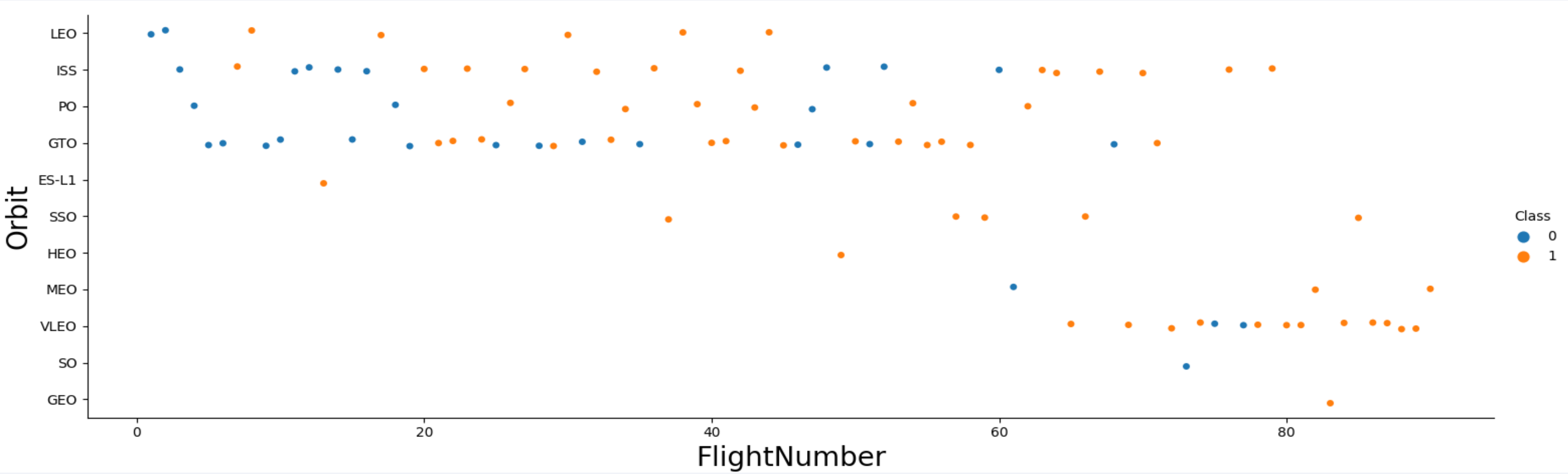
Success Rate vs. Orbit Type

- Success rate depends on orbit



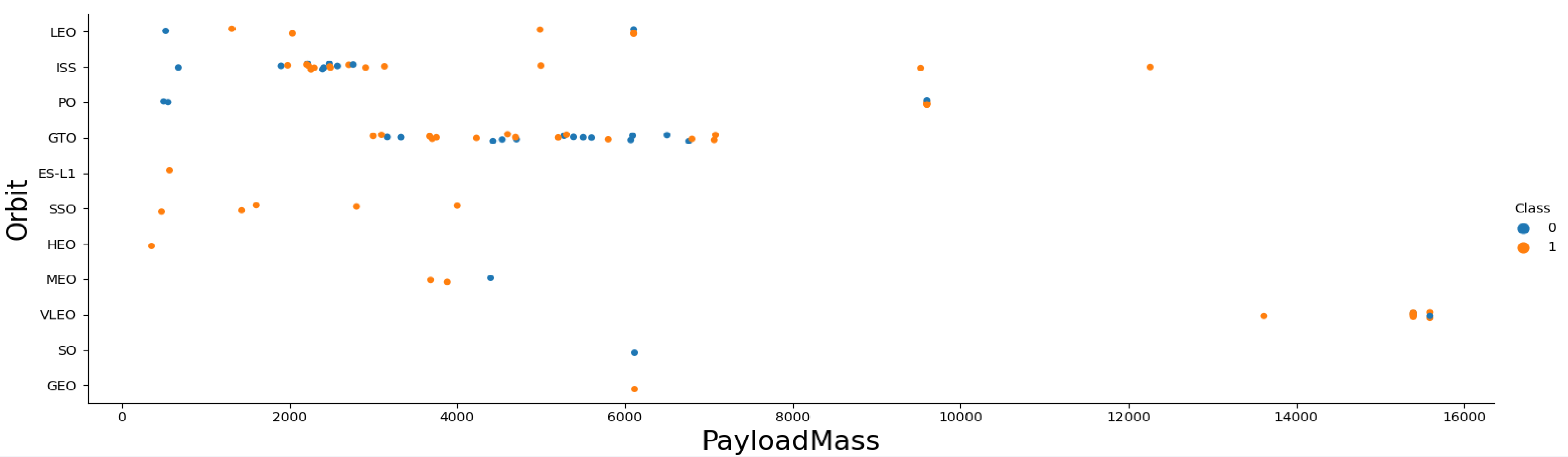
Flight Number vs. Orbit Type

- Set of popular orbits depends on time.



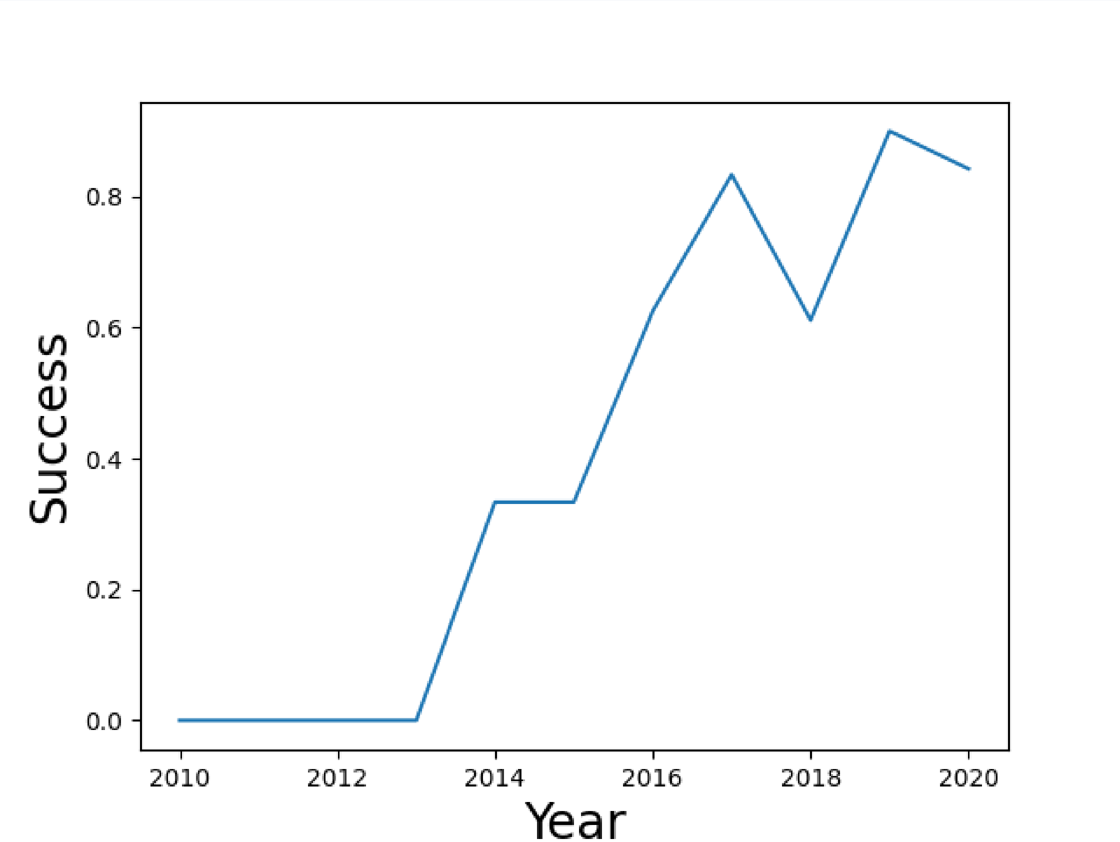
Payload vs. Orbit Type

- There is dependency between mass and orbit



Launch Success Yearly Trend

- Success rate is growing with experience (year)



All Launch Site Names

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

We have 4 launch sites.

Their names are abbreviatures.

Launch Site Names Begin with 'KSC'

DATE	TIME_UTC	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING_OUTCOME
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- 5 records with Launch site begins with KSC

Total Payload Mass

- The total payload carried by boosters from NASA is equal 45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1 is equal 2928.4

First Successful Ground Landing Date

- The first successful landing outcome on drone ship was 2016-04-08.

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

BOOSTER_VERSION
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Successful 100
- Failure 1
- Total 101

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

BOOSTER_VERSION
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2017 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

month	LANDING_OUTCOME	LAUNCH_SITE	BOOSTER_VERSION
02	Success (ground pad)	KSC LC-39A	F9 FT B1031.1
05	Success (ground pad)	KSC LC-39A	F9 FT B1032.1
06	Success (ground pad)	KSC LC-39A	F9 FT B1035.1
08	Success (ground pad)	KSC LC-39A	F9 B4 B1039.1
09	Success (ground pad)	KSC LC-39A	F9 B4 B1040.1
12	Success (ground pad)	CCAFS SLC-40	F9 FT B1035.2

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

- successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

LANDING_OUTCOME	The number of
Success (drone ship)	5
Success (ground pad)	3

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

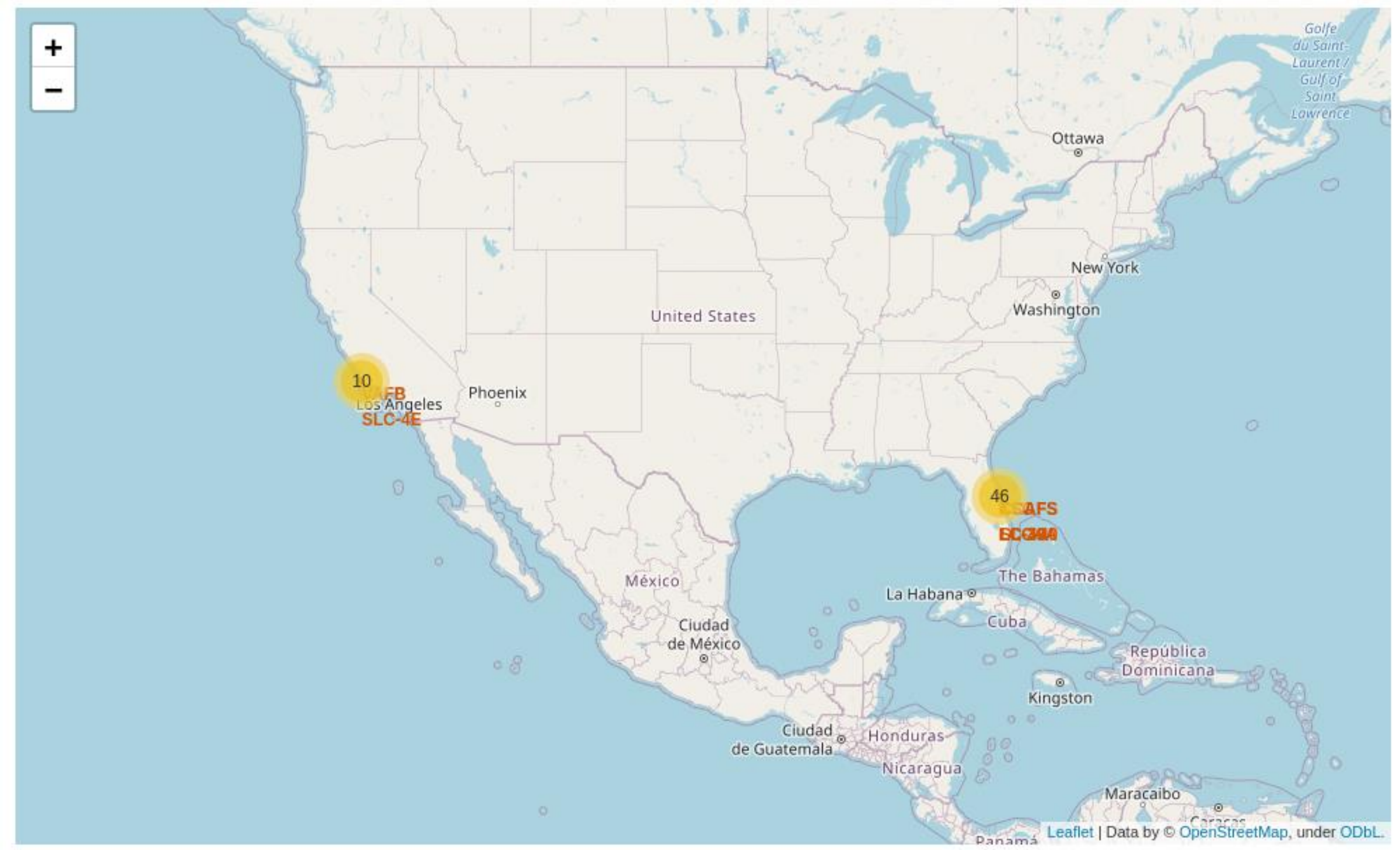
- successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order

DAT	TIME_UTC	BOOSTER_VERSION	LAUNCH_SITE	PAYLOAD	PAYLOAD_MASS	ORBIT	CUSTOMER	MISSION_OUTCOME	LANDING_OUTCOME
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

Folium Map with launch sites

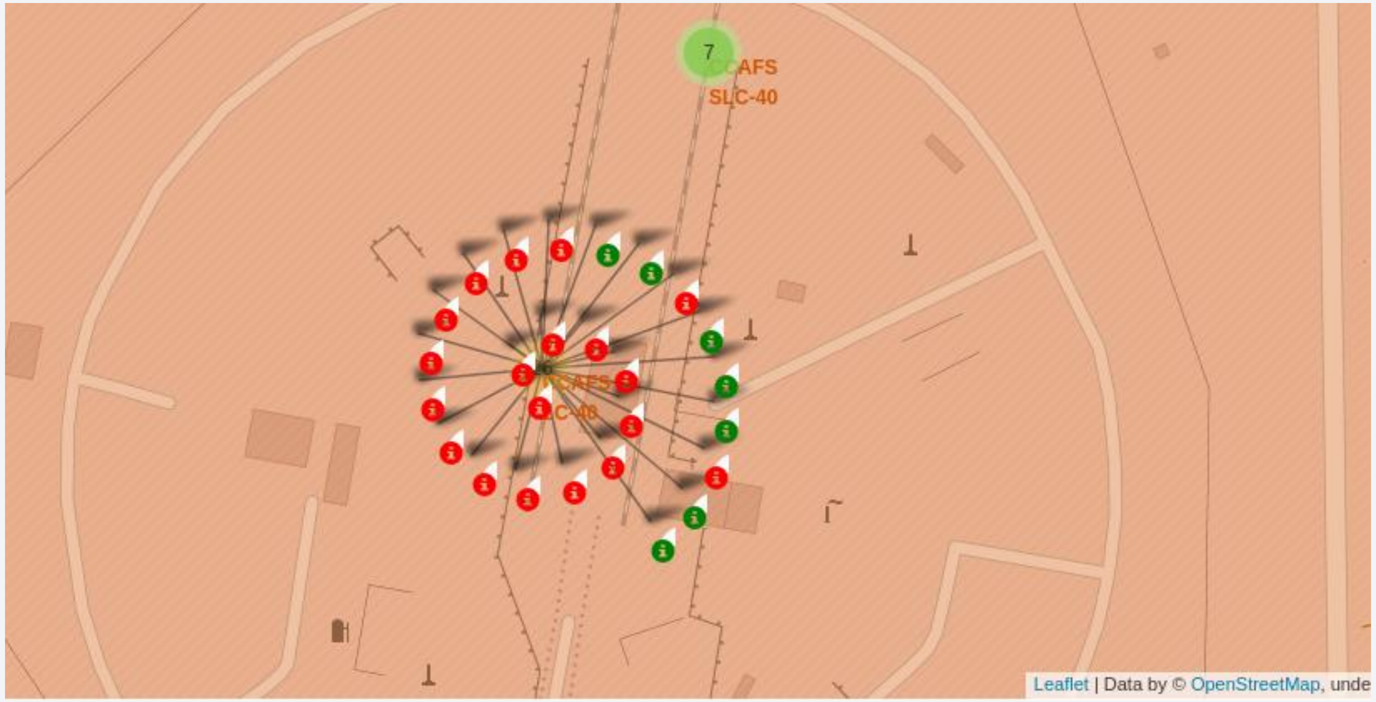


The numbers show the counts of successful landings

Folium Map with launch sites

- more close view: red labels are for failure, green ones are for success

- The success rate is growing
- with time



The closest proximities to sites

- The minimal distance from sites to road (magenta) and coastline (blue)

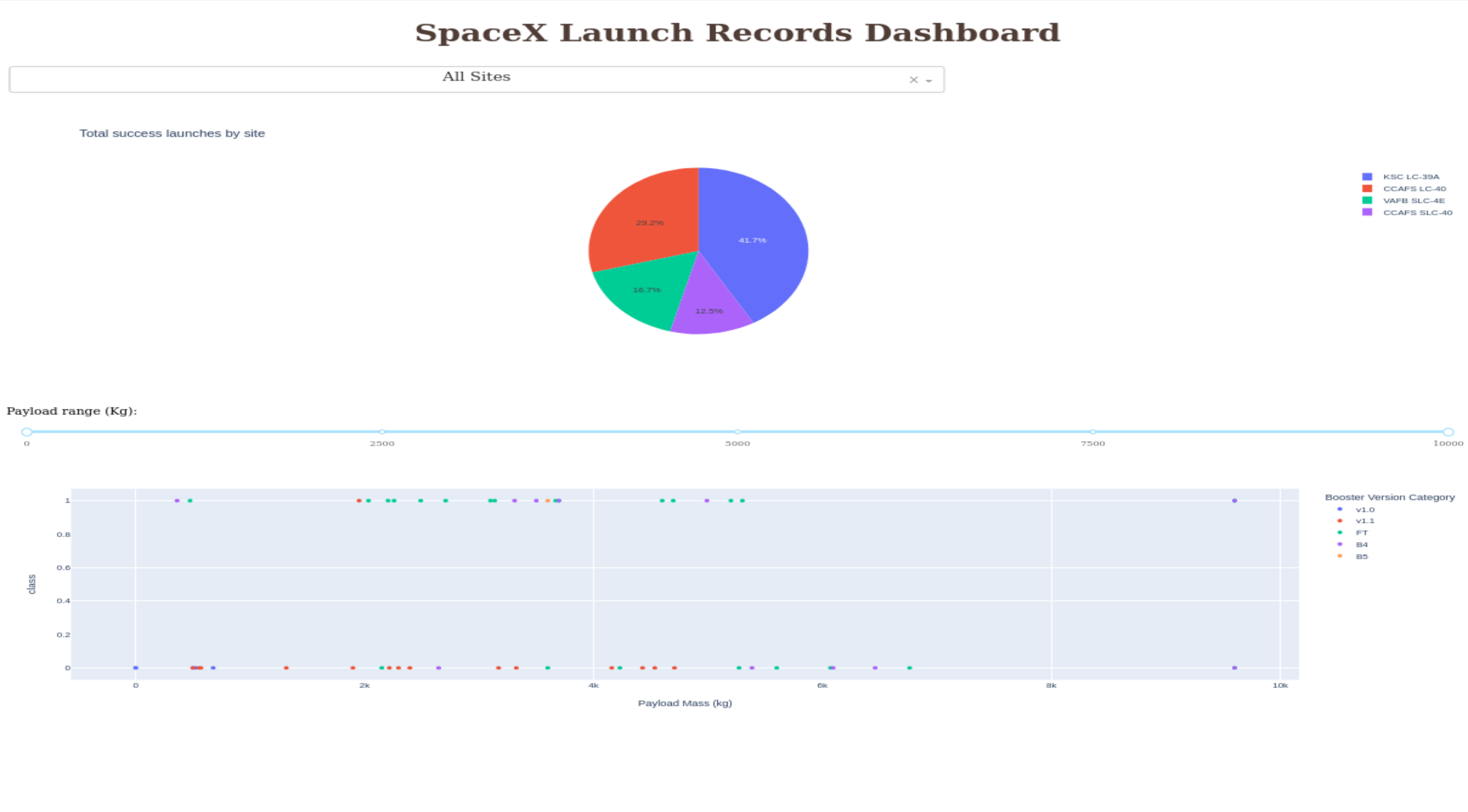
All sites are close to railway or highway, but quite far away from cities

Section 4

Build a Dashboard with Plotly Dash

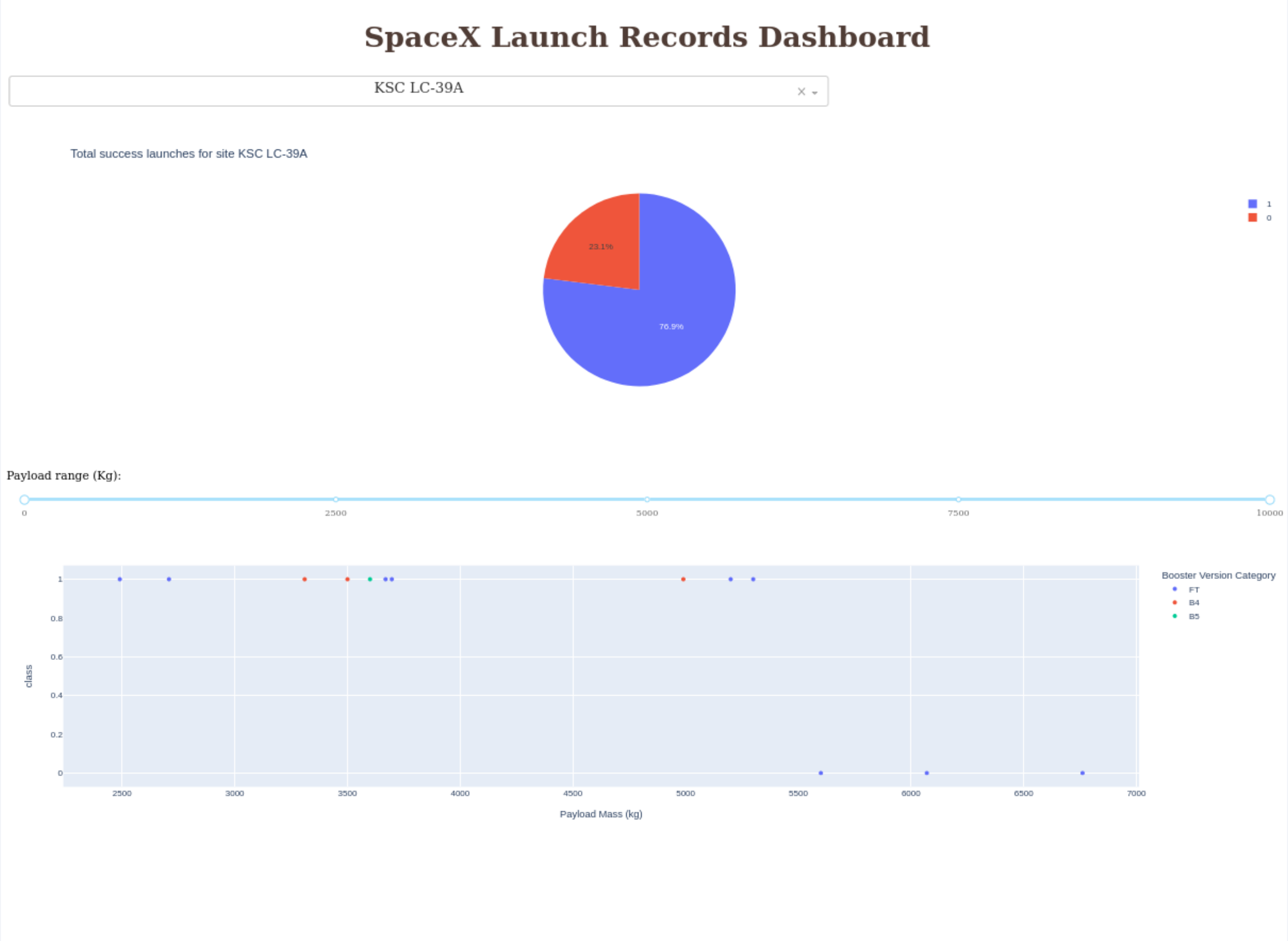
Success status to all sites

- Atlantic coasts are the best



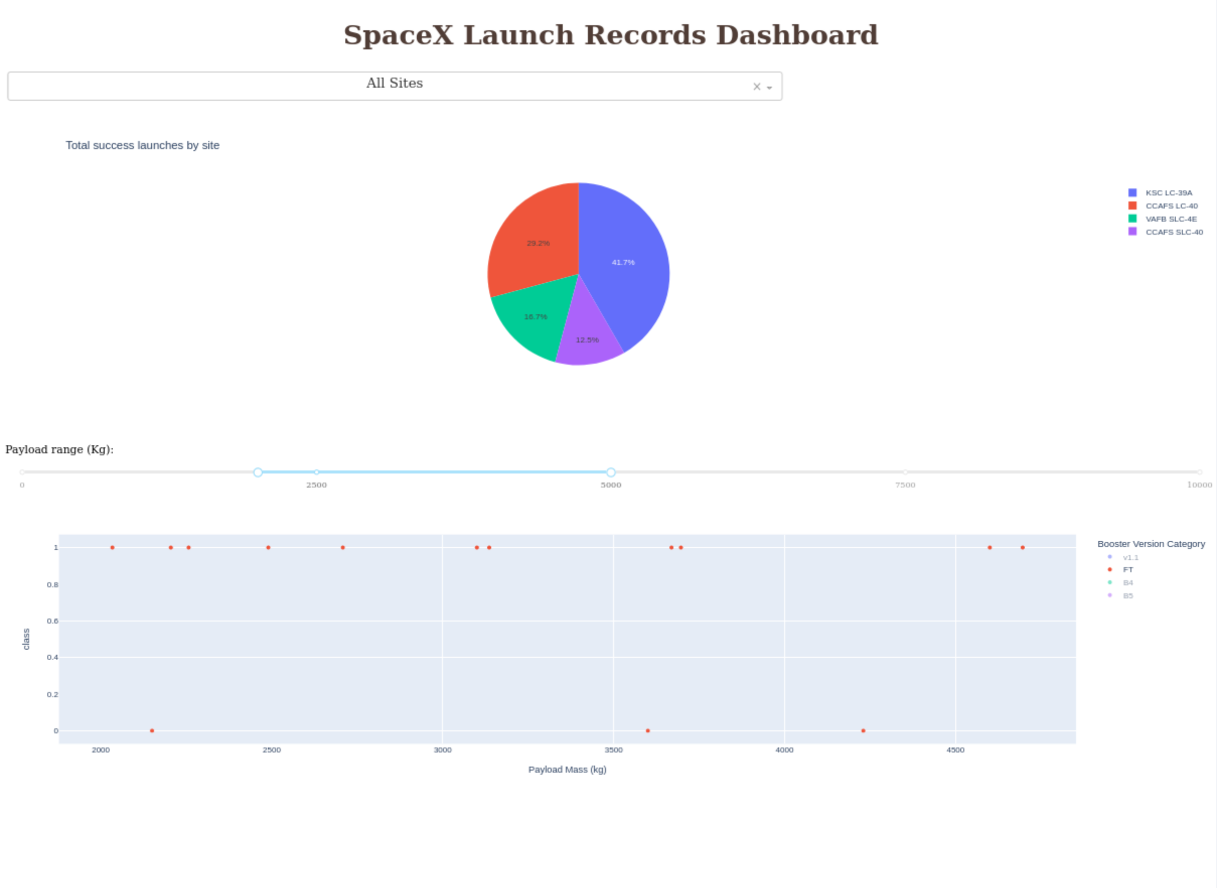
Success status to the best site (KSC)

- It's not a good idea to launch heavy rockets from that site



Boosters and Payloads

- FT booster is quite successful in range 2000-5000 kg

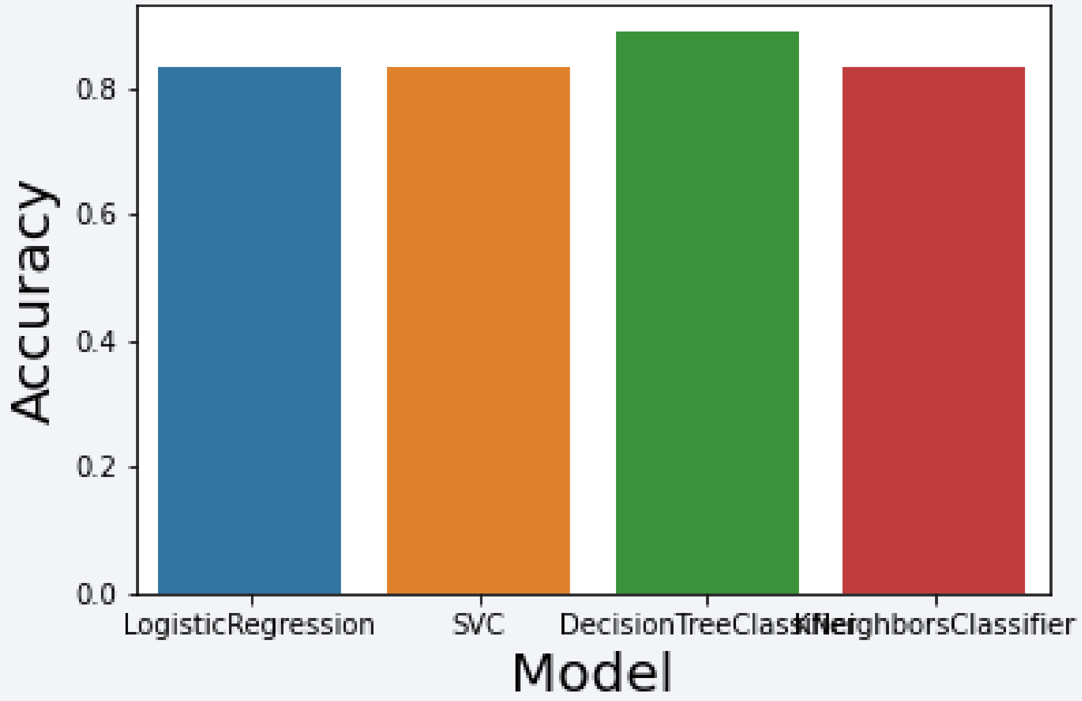


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Builded models accuracy
- The best result has decision tree
- DecisionTree uses random algorithm for splitting => result depends on runtime



Confusion Matrix

- On test data sometimes we have false positives

Conclusions

- We collected and wrangled the launch data
- We tried to establish dependencies between the data columns: dependencies exists
- We tried models of different types and found the best one
- Due to randomness in learning tree classifier we run this process not once to obtain the best result
- Out model has an accuracy 0.8889

Appendix

- Full code and some of used collected data can be found here

Thank you!