IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

- Project background and context
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
  - We want to find what characteristic affect success rate of landing. Is it geographical feature or maybe payload mass or something else. We want to explore connection between different features and if possible develop predictive model.

Section 1

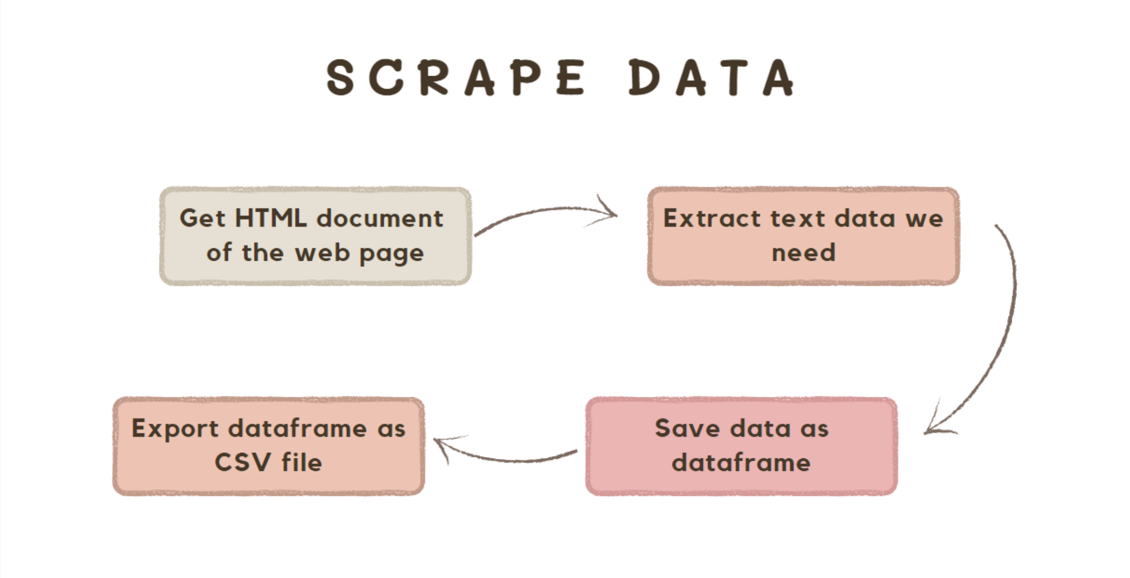
Methodology

Methodology

Executive Summary

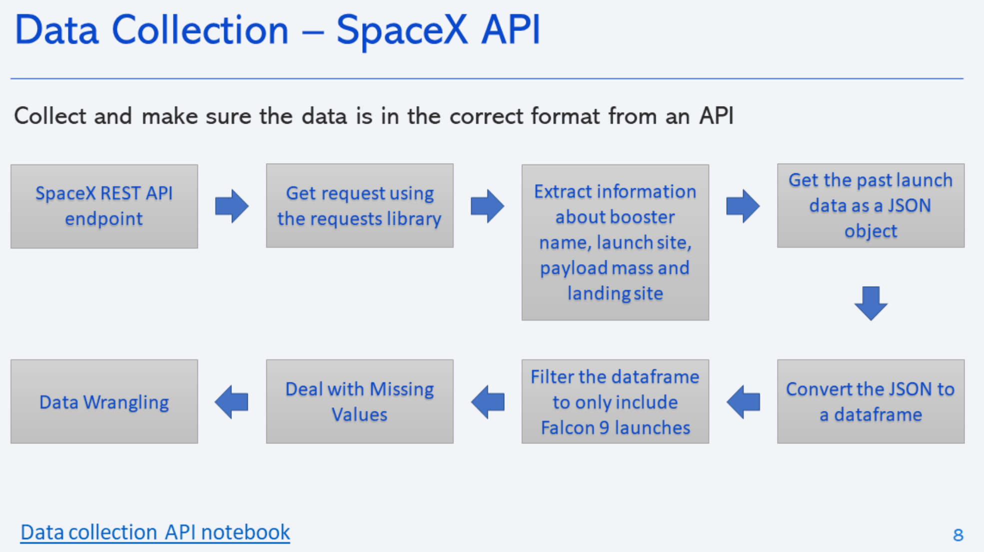
- Data collection methodology:
  - We used SpaceX API and wikipedia page to collect open data for our project
- Perform data wrangling
  - We explored data for correctness and replaced missing values with average value of corresponding feature.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - In this block we developed several models KNN-mean, TreeClassifier and LogisticRegression

Data Collection



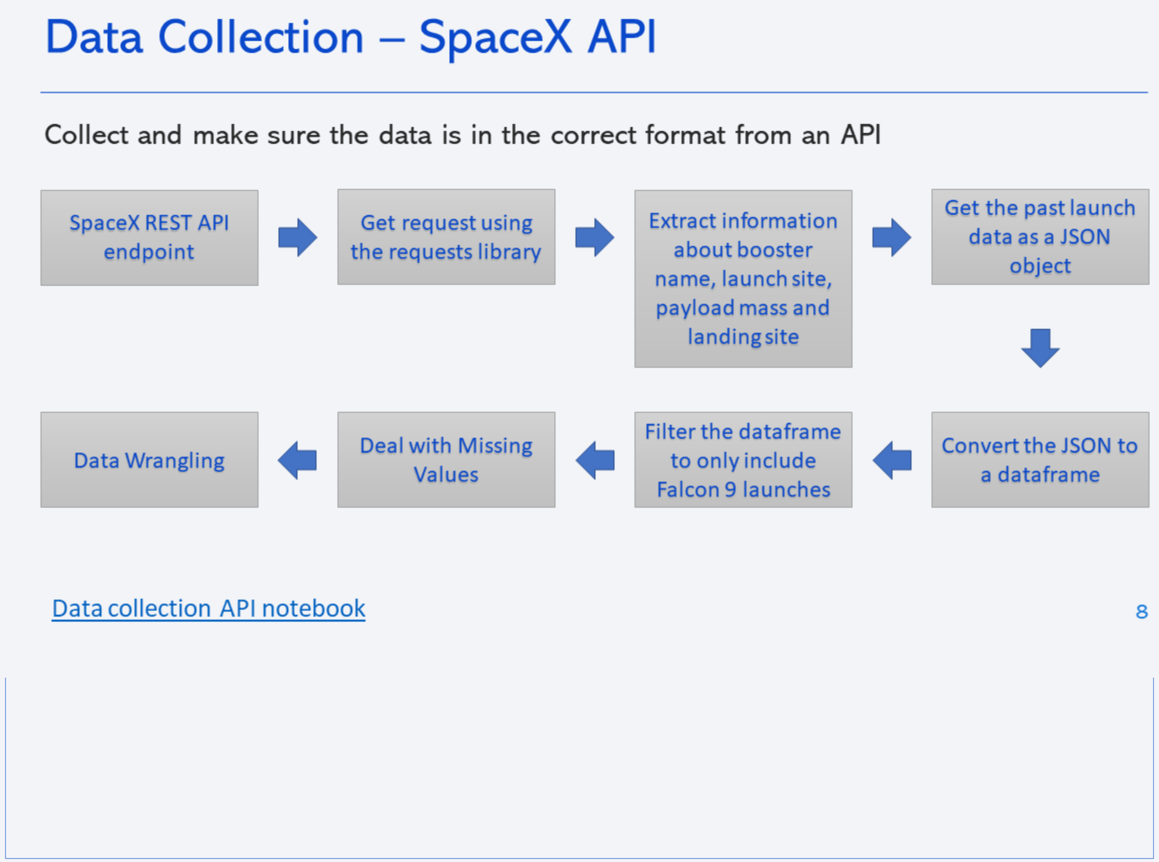
Web scrapping from Wikipedia

SpaceX API



Data Collection – SpaceX API

- We used following URL:
- https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json
- `Pyhton, JSON, Pandas.`
- For more references look git repo:
    - https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_1/Data-collection/jupyter-labs-spacex-data-collection-api.ipynb



Result of API call

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response.json())
```

[13]

Using the dataframe `data` print the first 5 rows

```
# Get the head of the dataframe
data.head()
```

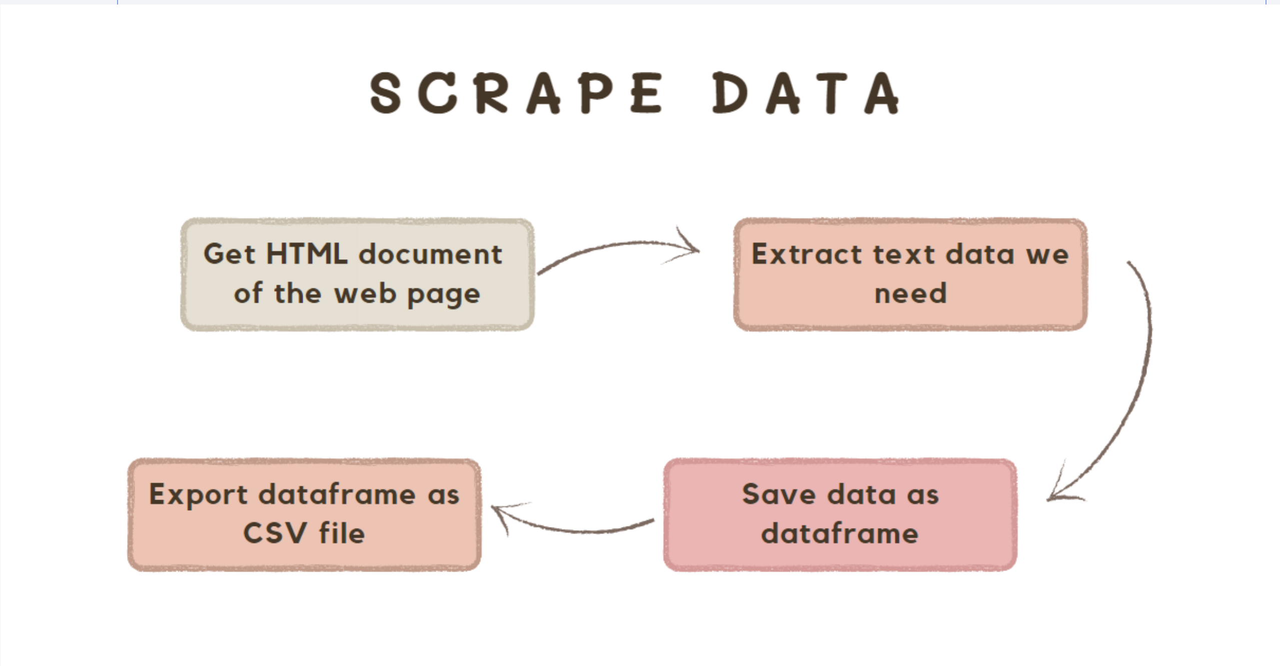
[14]

5 rows 5 rows × 42 cols Static Output

	static_fire_date_utc	static_fire_date_unix	tbd	net	window	rocket	success
0	2006-03-17T00:00:00.000Z	1.142554e+09	False	False	0.0	5e9d0d95eda69955f709d1eb	False
1	None	NaN	False	False	0.0	5e9d0d95eda69955f709d1eb	False
2	None	NaN	False	False	0.0	5e9d0d95eda69955f709d1eb	False
3	2008-09-20T00:00:00.000Z	1.221869e+09	False	False	0.0	5e9d0d95eda69955f709d1eb	True
4	None	NaN	False	False	0.0	5e9d0d95eda69955f709d1eb	True

Data Collection - Scraping

- Web Scrapping was performend on following web page:
- https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Python, BeautifulSoap, HTML-parsing, Pandas.
- For more references look git repo:
    - https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_1/Data-collection/jupyter-labs-webscraping.ipynb



Data Wrangling

Data processing stages:

- Determine what type of characteristics available, features and their data type.
- Perform simple data analysis
  - What types of orbit are used
  - How many different launch sites we have
  - Calculate number of launches on different sites
- Explore what type of outcomes are present and change them binary represantation(Success or Failure).
- Export data frame to .csv file.

GitHub: https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_1/Data-collection/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Main charts that were used:
  - **Scatter plot** – to determine relashionships between different features.
  - **Line plot** – to determine main trend of success rate through years
  - **Bar plot** – to visualize the relationship between success rate of each orbit type

GitHub:

https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_2/edadataviz.ipynb

EDA with SQL

- `%sql select distinct "Launch_Site" from SPACEXTABLE`
- `%sql select * from SPACEXTABLE where "Launch_Site" like "CCA%"  limit 5`
- `%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = 'NASA (CRS)'`
- `%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'`
- `%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'`
- `%sql select "Booster_Version", PAYLOAD_MASS__KG_  from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000`
- `%sql select "Mission_Outcome", count(*) from SPACEXTABLE group by "Mission_Outcome"`
- `%sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_  = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE)`
- `%sql select (substr("Date",6,2)) as "Month", "Landing_outcome", "Booster_Version", "Launch_Site" from SPACEXTABLE where substr("Date",0,5) = '2015' and "Landing_outcome" = 'Failure (drone ship)'`
- `%sql select "Landing_outcome", count(*) from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_outcome" order by count(*) desc`

GitHub: https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_2/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Marker, Circle, MarkerCLuster – were used to point launch sites and make it interactive.
- Lines – were used to point nearest roads, railways, cities and show distance to them

Git: https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart – to show success rate of landing
- Scatter plot – to show relationship between PayloadMass and SuccessRate
- DropDownOptions – to choose specific launch site to see success rate of given site
- RangeSlider – to choose range of PayloadMass.

Git: https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_3/spacex_dash_app.py

Predictive Analysis (Classification)

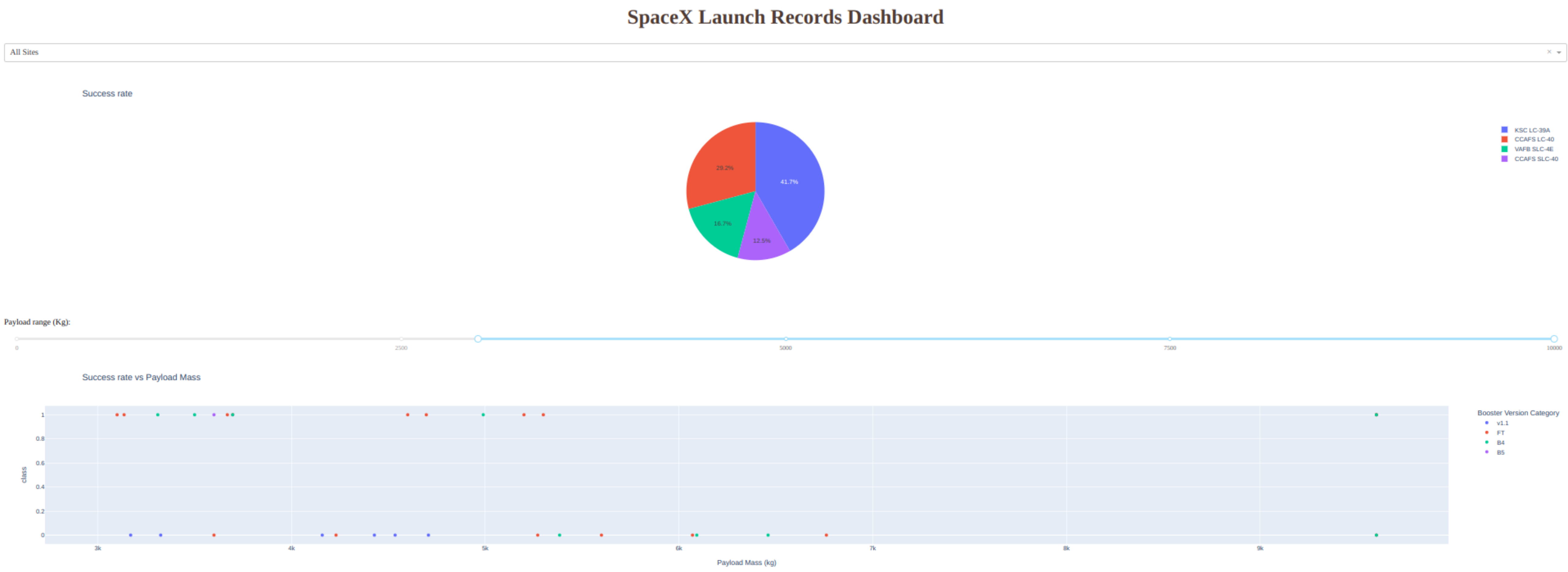
Steps:

- Data loading
- Features preprocesing, standartalization.
- Train-test split
- Models selection: LogisticRegression, SVM, TreeClassifier, KNN.
- Models training.
- Models evaluation

Git: https://github.com/Oleg-algebra/DataScienceCapstoneProject/blob/main/Module_4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

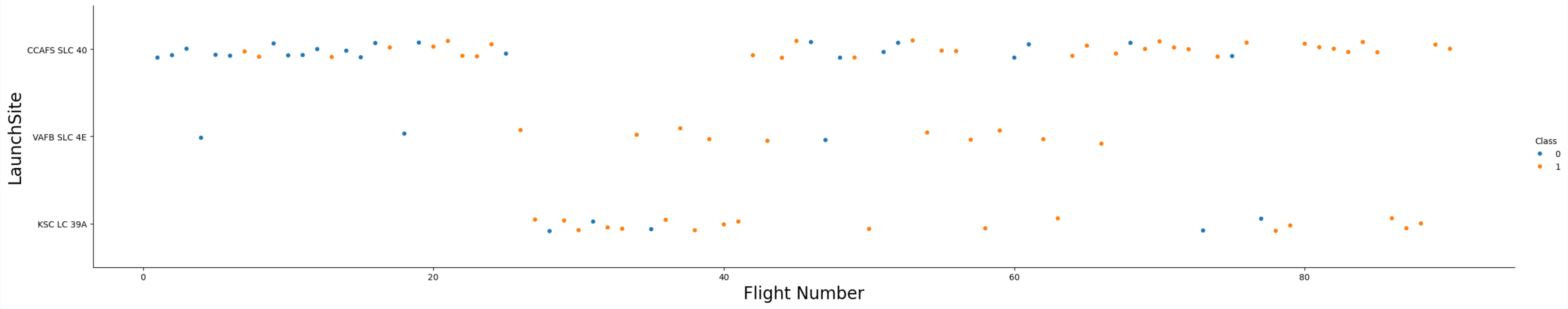
- Exploratory data analysis results
  - VAFB SLC 4E – doesn’t have rockets with payload mass greater than 10000kg.
  - Increasing trend of success rate landing from 2010 to 2020.
- Predictive analysis results
  - All models have accuracy on testing set greater than 83%
  - The best results was show by TreeClassifier and KNN models with 87% and 86% of accuracy respectively.


SpaceX Launch Records Dashboard
All Sites
Success rate
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
0
2500
5000
7500
10000
Success rate vs Payload Mass
class
1
0.8
0.6
0.4
0.2
0
3k
4k
5k
6k
7k
8k
9k
Payload Mass (kg)
Booster Version Category
v1.1
FT
B4
B5

Section 2

Insights drawn from EDA

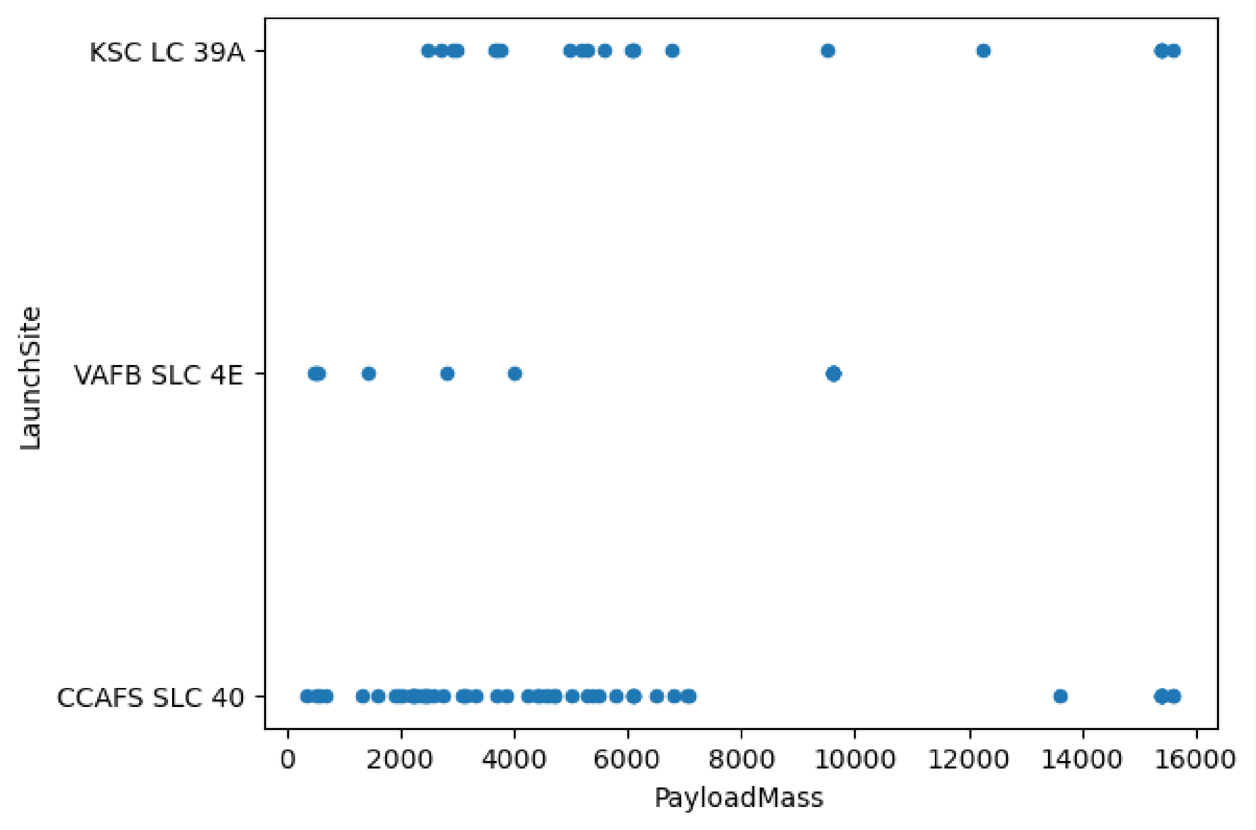
Flight Number vs. Launch Site



- Biggest numbers of landings was on launch site CCAFS SLC 40.
- The least amount of landings – VAFB SLC 4E.
- KSC LC 39A wasn’t used for the first landings. CCAFS SLC 40 is used unifomly through all period of time.
- For last 40 flight number CCAFS SLC 40 has hight success rate of landing

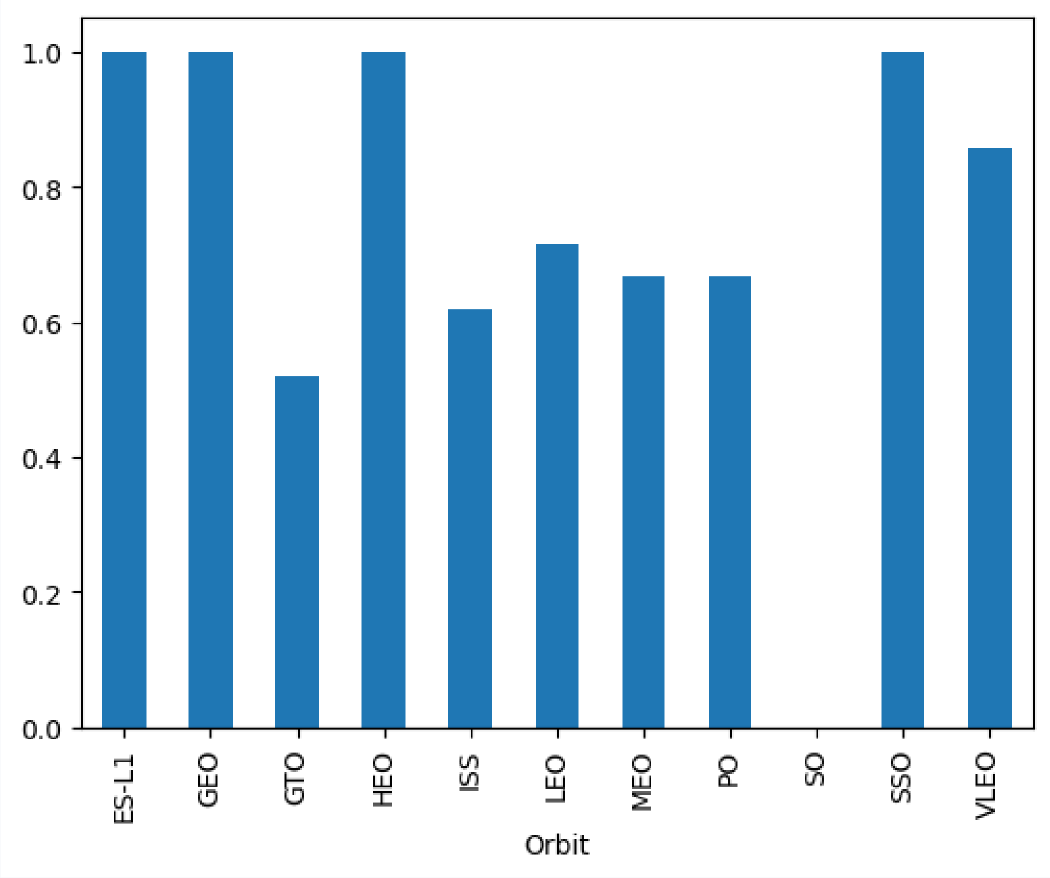
Payload vs. Launch Site

- VAFB SLC 4E doesn't have payload mass greater than 10000
- CCAFS SLC 40 commonly used for payloads smaller than 8000.
- KSC LC 39A has different types of payload record.



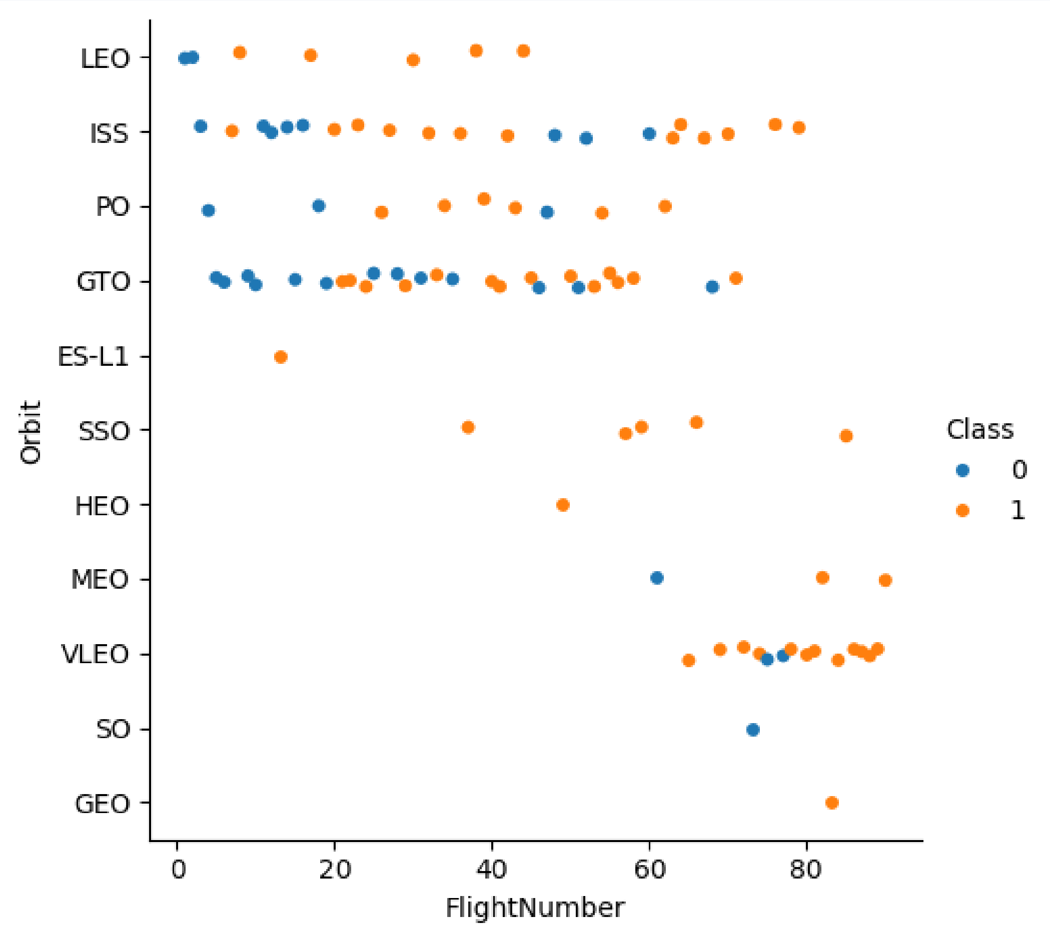
Success Rate vs. Orbit Type

- Orbits ES-L1, GEO, HEO, SSO have the highest success rate.
- SO – has the lowest success rate.
- Rest of orbits have average success rate .

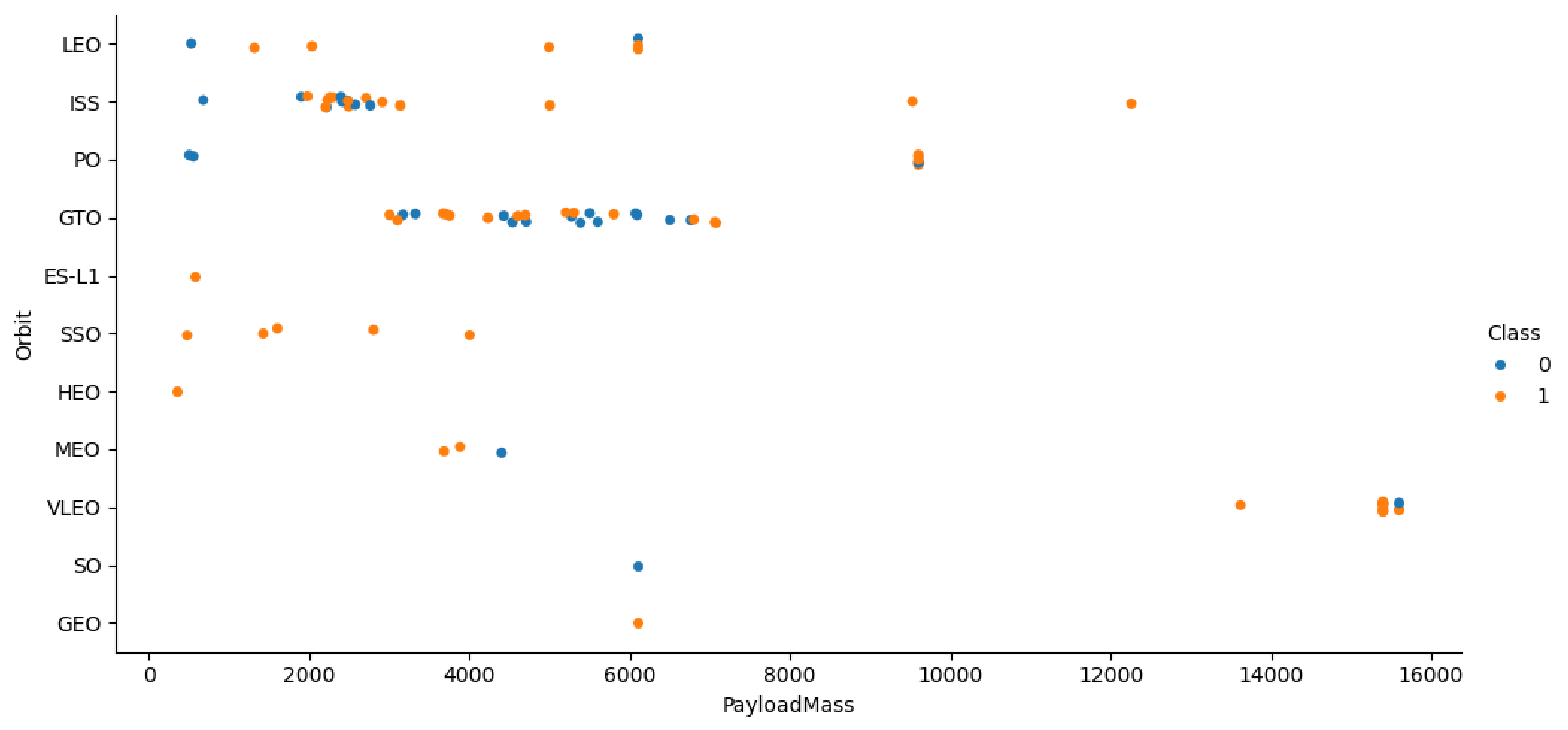


Flight Number vs. Orbit Type

- Most commonly used orbits are LEO, ISS and GTO.
- VLEO has increasing trend of usage for last 40 flight numbers.
- ES-L1, HEO, GEO, SO have only one record of flight. It explains some results from previous slide.

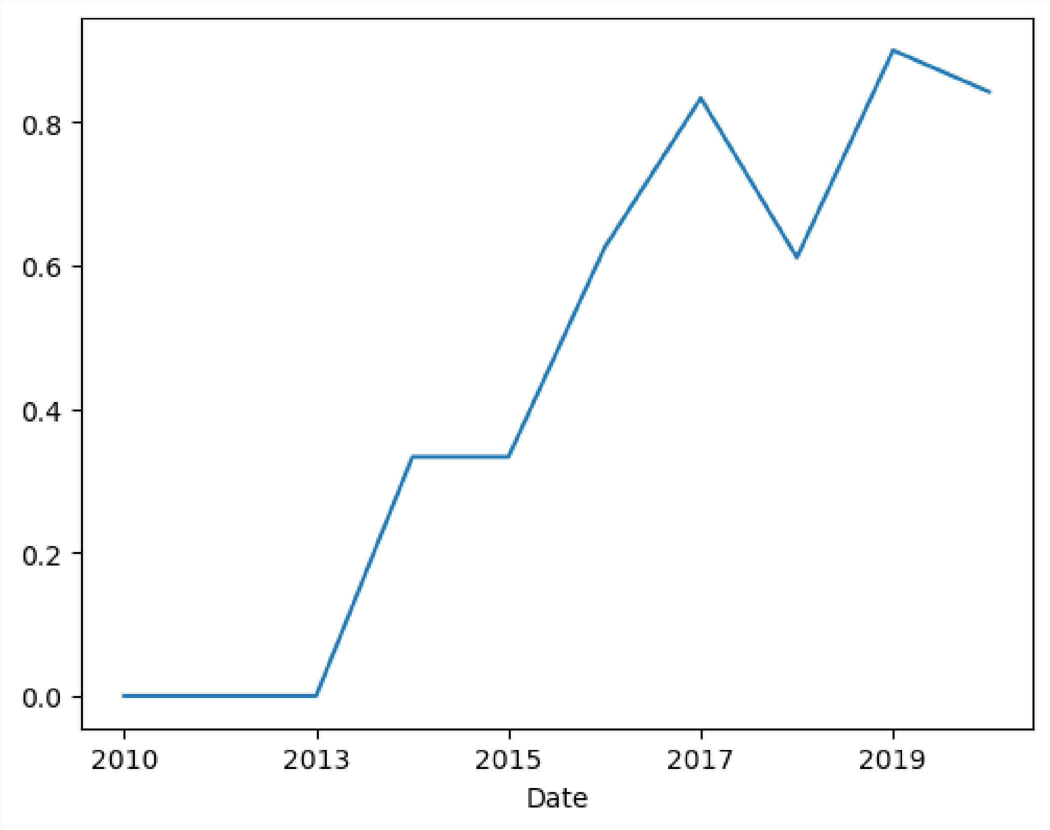


Payload vs. Orbit Type



Launch Success Yearly Trend

- From 2010 to 2013 we can see 0 success rate.
- From 2013 to 2020 we can see increasing trend of average success rate of landing.



All Launch Site Names

- Using sql magic we can that our data base have four distinct launch sites where rockets are landing.

```
In [10]: %sql select distinct "Launch_Site" from SPACEXTABLE

 * sqlite:///my_data1.db
Done.
```

Out[10]:

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

Launch Site Names Begin with 'CCA'

- We have that first landing attempt was in 2010 and ended with failure.
- From the first five records we see that there was 2 flights per year on average.

```
In [14]: %sql select * from SPACEXTABLE where "Launch_Site" like "CCA%"  limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Out
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (para
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (para
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No at
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No at
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No at

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [25]:
```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Customer" = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[25]:

sum(PAYLOAD_MASS__KG_)
45596

- From 2010 to 2020 total PayLoad is more than 45 tons.

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

In [26]:
```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where "Booster_Version" = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[26]:

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [46]:

```
%sql select min("Date") from SPACEXTABLE where "Landing_Outcome" = 'Success (ground pad)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[46]:

min("Date")
2015-12-22

- First successful landing occurred only after 5 years after first flight.

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [49]:
```
%sql select "Booster_Version", PAYLOAD_MASS__KG_  from SPACEXTABLE where "Landing_Outcome" = 'Success (drone ship
```

```
 * sqlite:///my_data1.db
Done.
```

Out[49]:

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- There are only 4 records and all of them belongs to booster F9 type.

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

In [52]:
```
%sql select "Mission_Outcome", count(*) from SPACEXTABLE group by "Mission_Outcome"
```

```
 * sqlite:///my_data1.db
Done.
```

Out[52]:

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- All boosters with max PayloadMass are boosters F9 B5 B..

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [54]:
```
%sql select "Booster_Version" from SPACEXTABLE where PAYLOAD_MASS__KG_  = (select max(PAYLOAD_MASS__KG_) from SPA
```

```
 * sqlite:///my_data1.db
Done.
```

Out[54]:

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

2015 Launch Records

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

In [60]:
```
%sql select (substr("Date",6,2)) as "Month", "Landing_outcome", "Booster_Version", "Launch_Site" from SPACEXTABL
```

```
 * sqlite:///my_data1.db
Done.
```

Out[60]:

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

There are two cases and both of them are have booster version F9 v1.1 and occurred on CCAFS LC-40 launch site.

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

- The most frequent result is no attempt for landing.
- The most frequent success landing is performed on drone ship. Also such type of landing has a biggest count of failures.

Task 10

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.

In [65]:

```
%sql select "Landing_outcome", count(*) from SPACEXTABLE where "Date" between '2010-06-04' and '2017-03-20' group
```

```
 * sqlite:///my_data1.db
Done.
```

Out[65]:

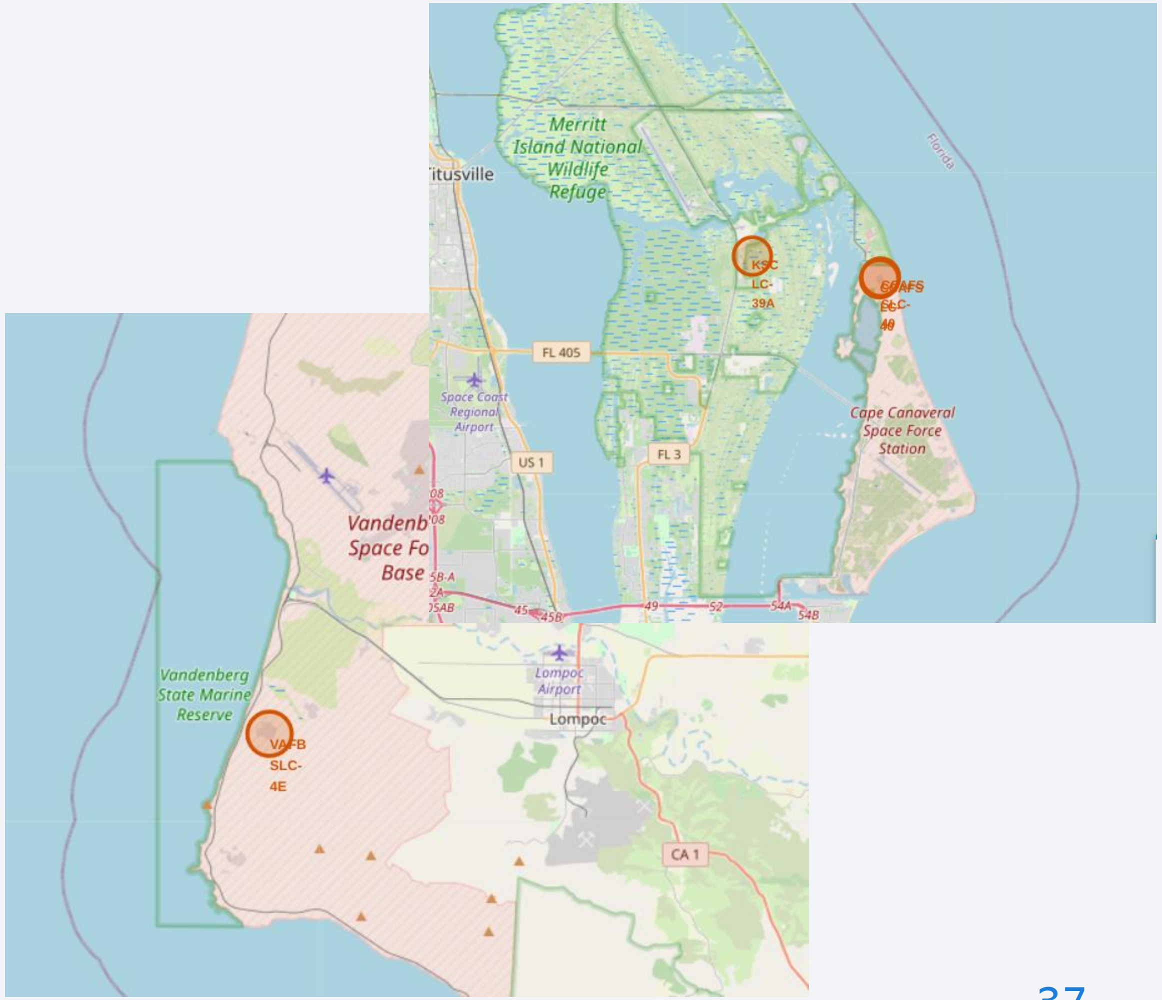
Landing_Outcome	count(*)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

Map of launching sites

- According to generated map we have two main regions.
- First region located in the east part of America and has three launching sites.
- Second region located on the west part of continent and has only one launching site.

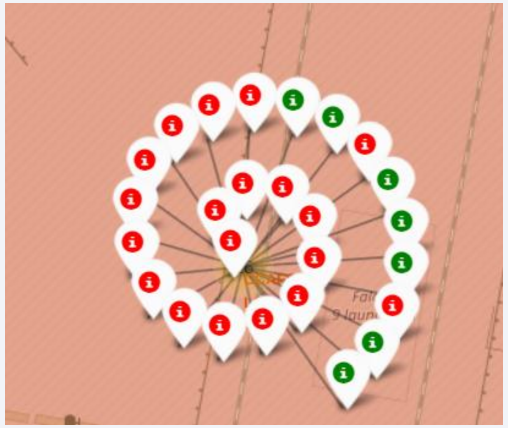


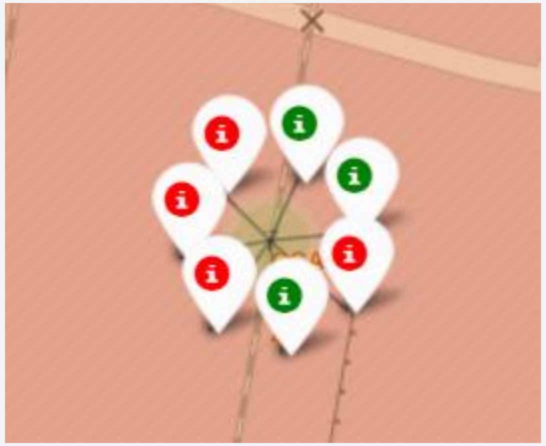
Map of landing outcomes

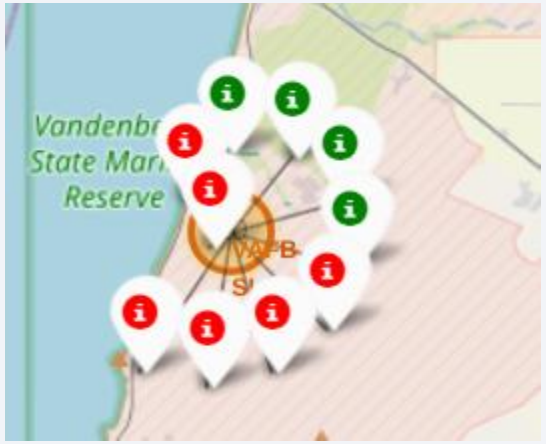
From screenshots we can see that east region has the biggest amount of landings attempts, see upper most pictures and second picture in the second row.

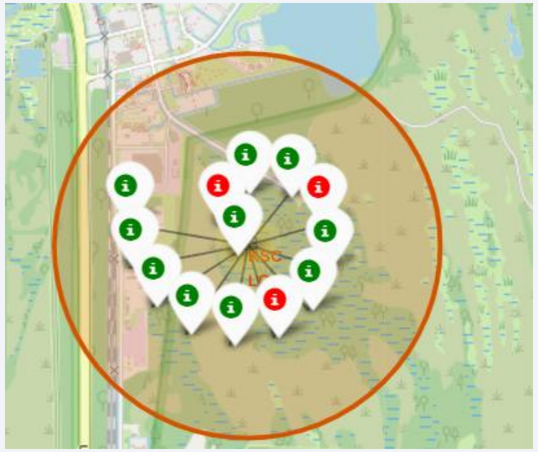
Also picture in first row and first column shows that CCAFS SLC-40 has the biggest count of landings and most of them are failures(red color).

On the contrary KSC LC-39A has the biggest amount of successful landings.

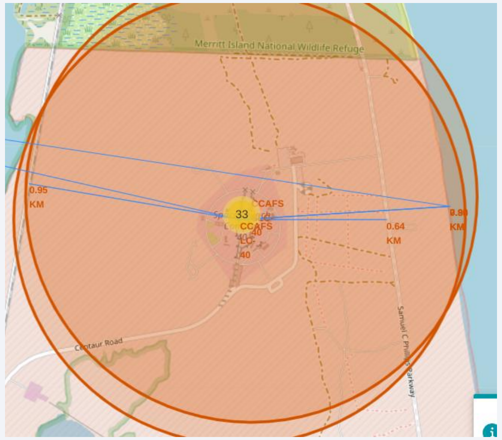


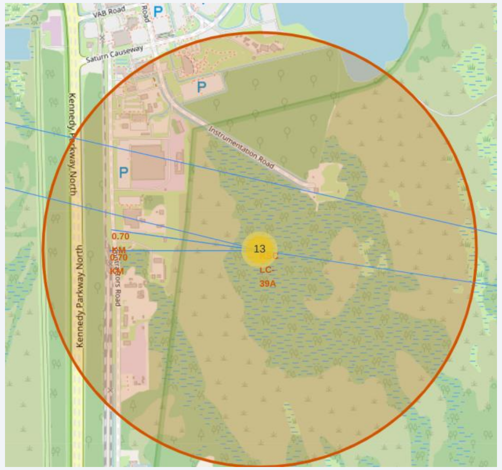






Launch sites location observation





All launch sites have railways, roads and coastlines near them, but located far from cities.

Section 4

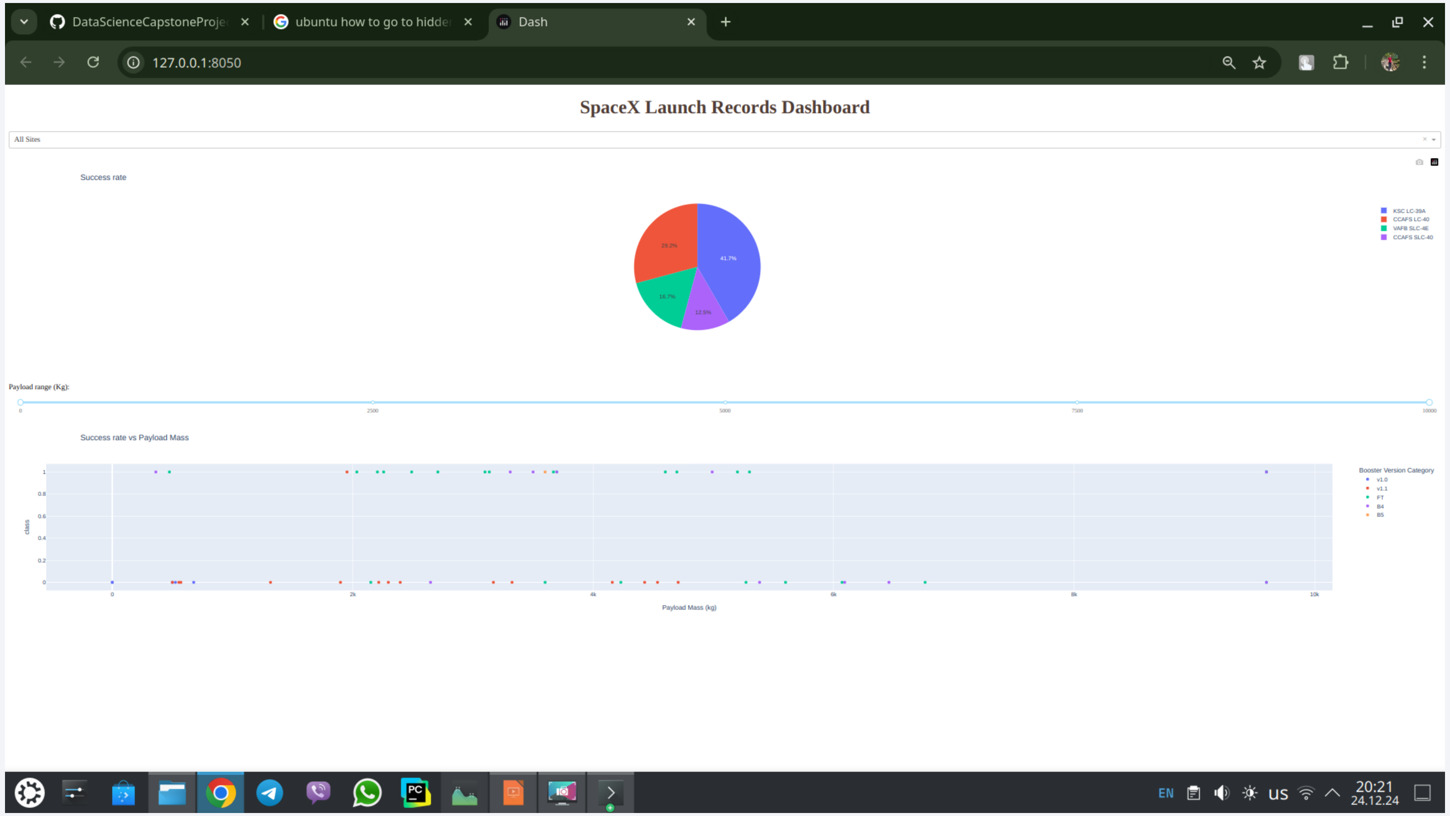
Build a Dashboard with Plotly Dash

All sites success rate

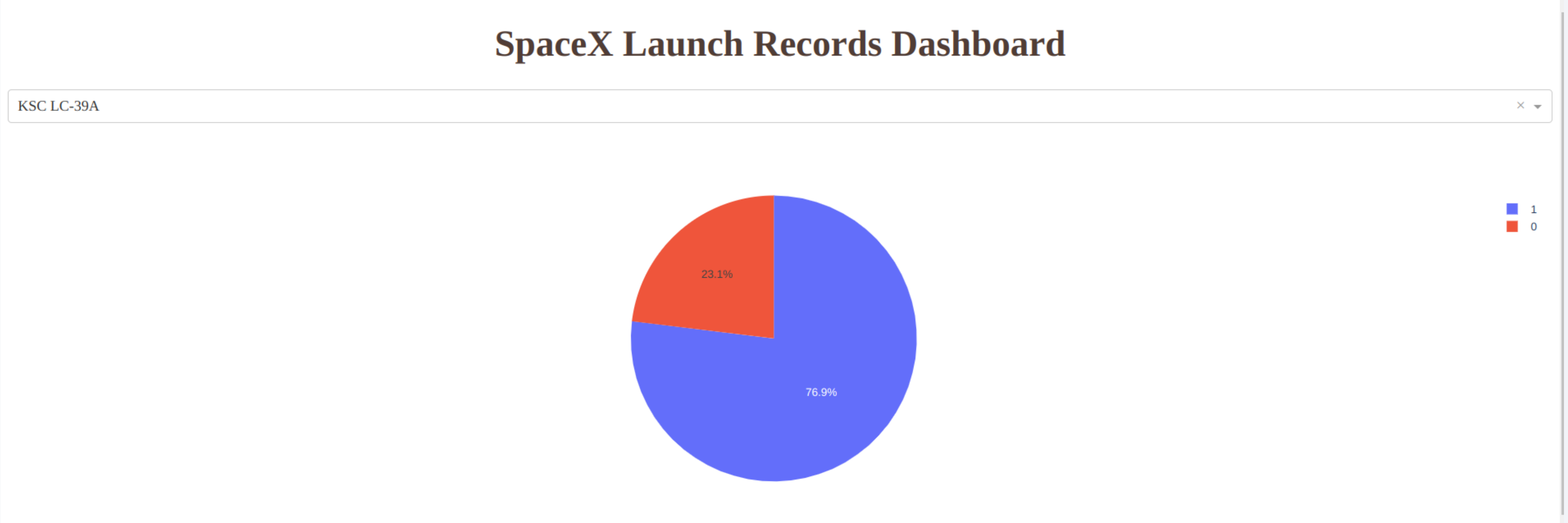
There we have 4 main elements:

- DropDownMenu to choose launching site
- PieChart
- RangeSlider to choose range of Payload Mass(kg)
- Scatter plot “Payload Mass vs Success rate”.

From piechart we can see that KSC LC-39A has the highest success rate.

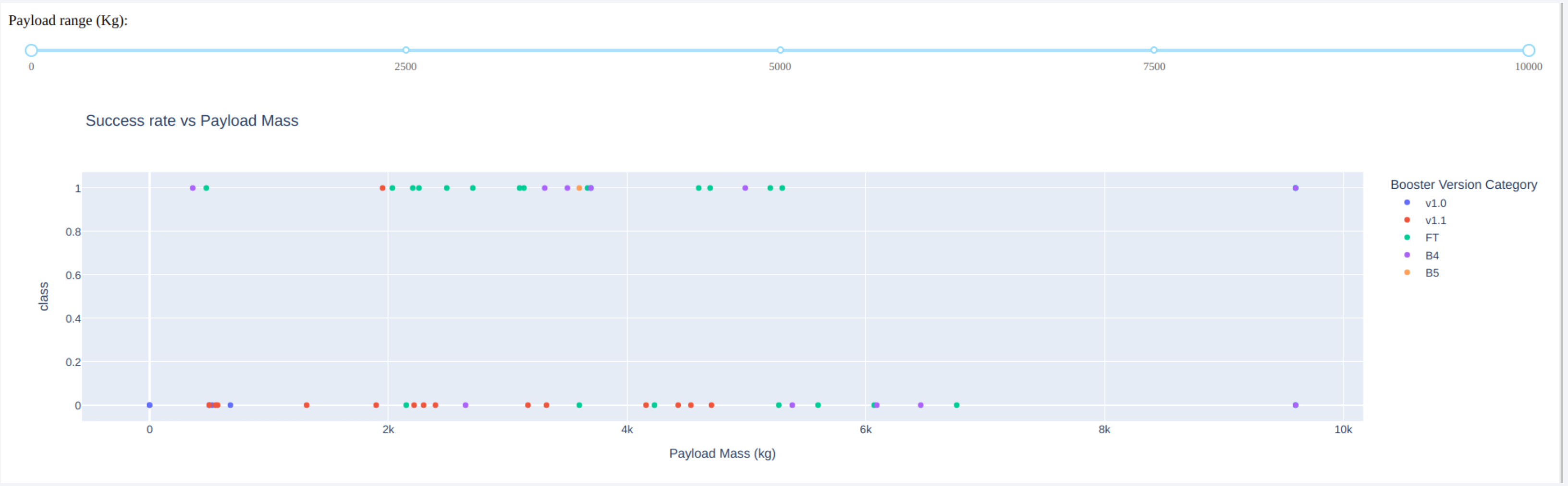


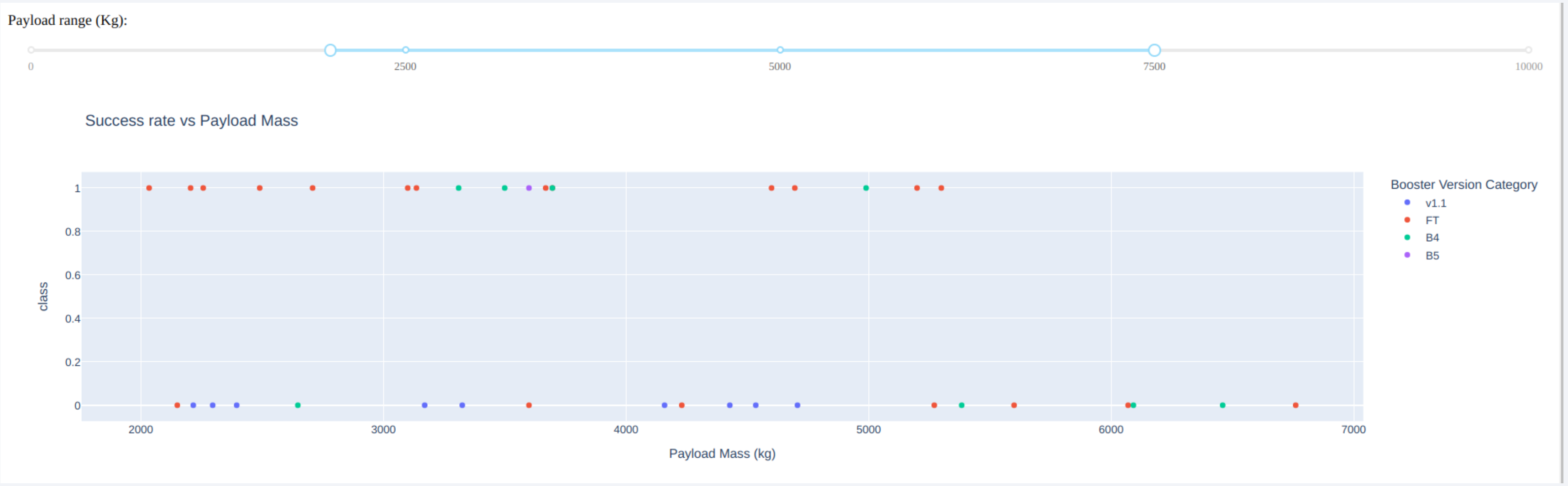
KSC LC-39A success rate

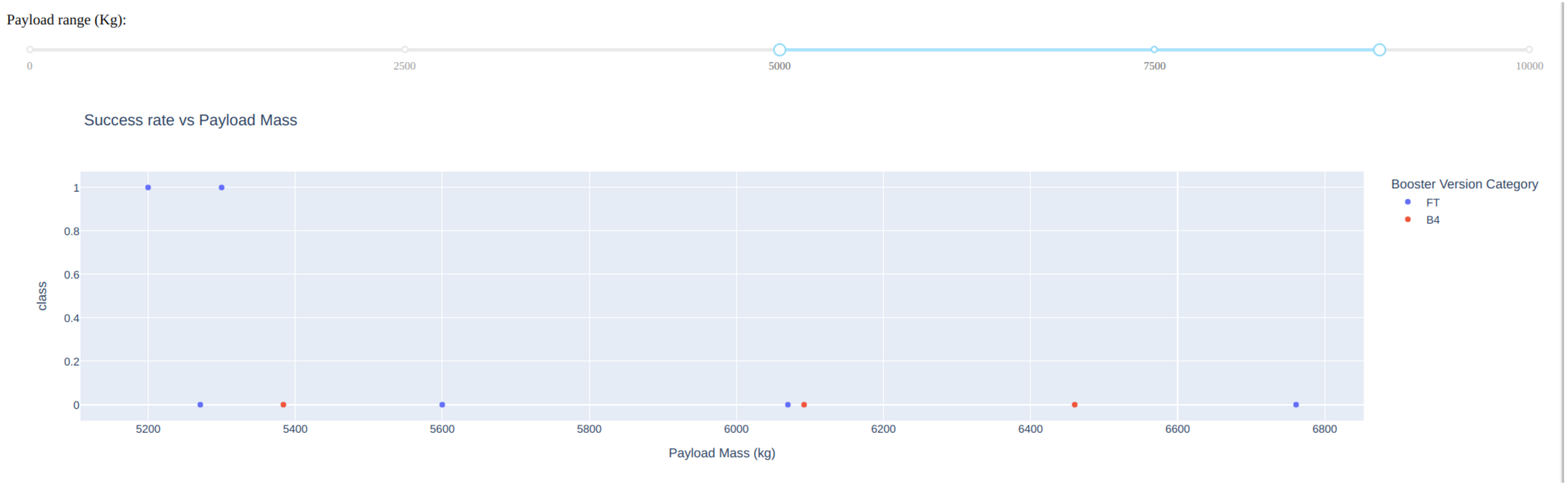


- From pie chart we can see that success rate on that site is greater than 75%

PayloadMass vs Success rate







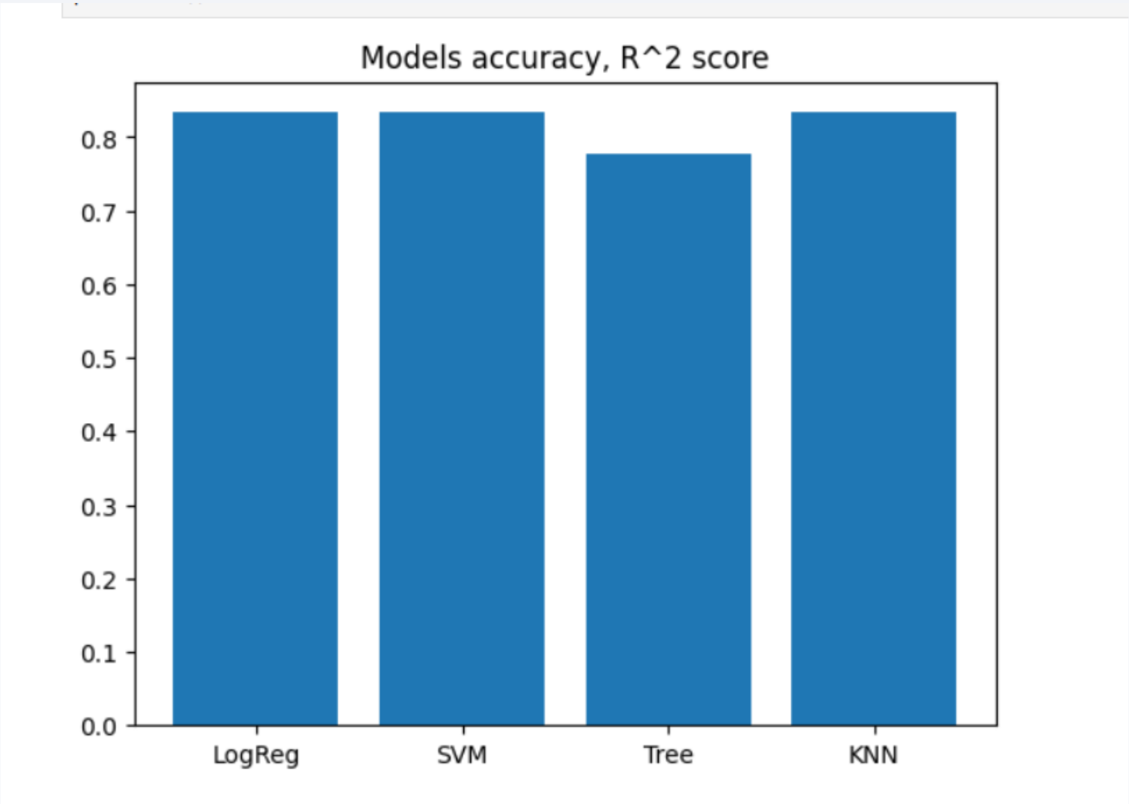
- From screenshots we can see that payload range from 5000 to 7000 have the lowest success rate
- The biggest success rate can be if payload will be in range between 2000 and 5500
- Booster version category **FT** has the highest succes rate across differrent payload masses
-

Section 5

Predictive Analysis (Classification)

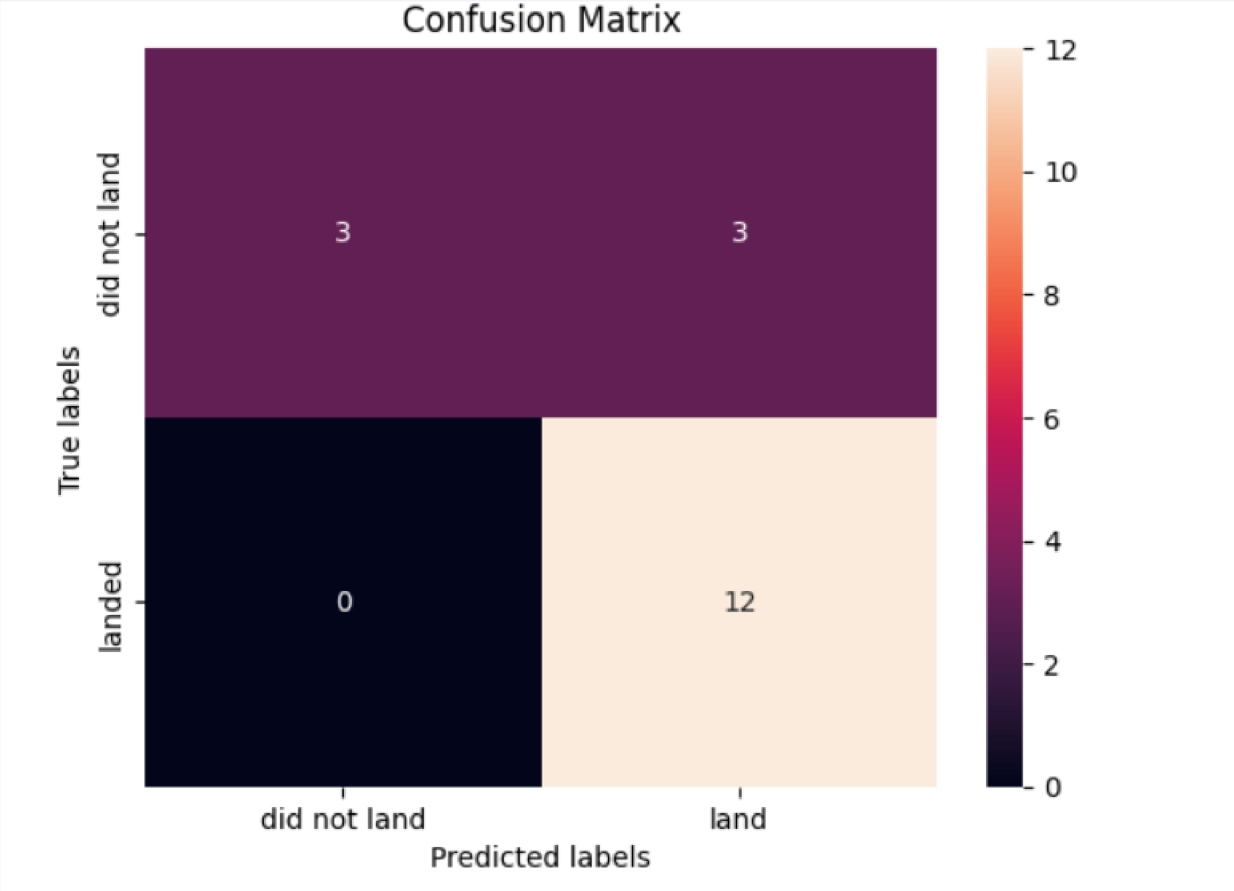
Classification Accuracy

- From a bar chart we can see that KNN and SVM have the highest R^2 score.
- On the contrary TreeClassifier has the lowest result.



Confusion Matrix

- On the screenshot on the right confusion matrix of KNN model.
- We can see that model differentiate between classes.
- There some problems with False positive error and need some improvement.
- Such model perfectly handle successful landings.



Conclusions

- VAFB SLC 4E – doesn’t have rockets with payload mass greater than 10000kg.
- Increasing trend of success rate landing from 2010 to 2020.
- All boosters with max PayloadMass are boosters F9 B5 B…
- CCAFS SLC-40 has the biggest count of landings and most of them are failures.
- The biggest success rate can be if payload will be in range between 2000 and 5500.
- KNN and SVM have the highest R^2 score.

Thank you!