Data Science Capstone project

SAGNIKA DUTTA

15.08.2021

Outline

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

Executive Summary

- We are trying to establish the cost from planned success and failure of first stage rocket landings.
- The data has been collected from the SpaceX website and Wikipedia pages using API and Web Scrapping methods
- Further, the data has been preprocessed by Data Wrangling
- The data is then analyzed and processed by :
  - EDA using SQL
  - Array of charts
  - Folium
  - Dashplots
  - Predictive Analysis
- All of these finally leads us to observe that basis the dataset, the launch sites have good geographical positioning with close proximity to features like railways, highways and coastlines.
- Further, we find that the success of first stage landings is dependent on several factors such as the launch site('KSC LC -39A'), lower payload mass, the orbit for the concerned launch, the booster version, etc.
- Also, we have analyzed several predictive models to facilitate Machine learning in the exercise for success and failure and found that while several models such as Logistic Regression, SVM and KNN provide a good accuracy in this case, the best accuracy is provided by the Decision Tree Classifier model.

Introduction

- Project background and context
  - We are analyzing rocket launch data from a company Space X for another new company Space Y, to understand the cost associated with rocket launches.
  - We will also be studying the associated parameters affecting the success of re-landing of first stages of rocket - payload, launch sites, Orbit etc.
- Problems you want to find answers
  - Impact of Payload Mass, Launch Site, Flight Number on the success of first stage landing.
  - The best Machine Learning Model to predict the success rate of any future launch's first stage landing
  - The geographical accessibility of existing launch sites

Methodology

- Data collection methodology using following techniques:
  - Data Collection via API from Space X website (cached)
  - Web Scrapping from Wikipedia page
- Perform data wrangling
  - Missing values were replaced with mean values
- Perform exploratory data analysis (EDA) using visualization and SQL
  - To find the correlation between different parameters and which are relevant vs irrelevant to our objective
- Perform interactive visual analytics using Folium and Plotly Dash
  - To visualize datasets and interpret data usability
- Perform predictive analysis using classification models
  - To ensure best possible model is used
  - With optimum parameters
  - And maximized accuracy score

Methodology

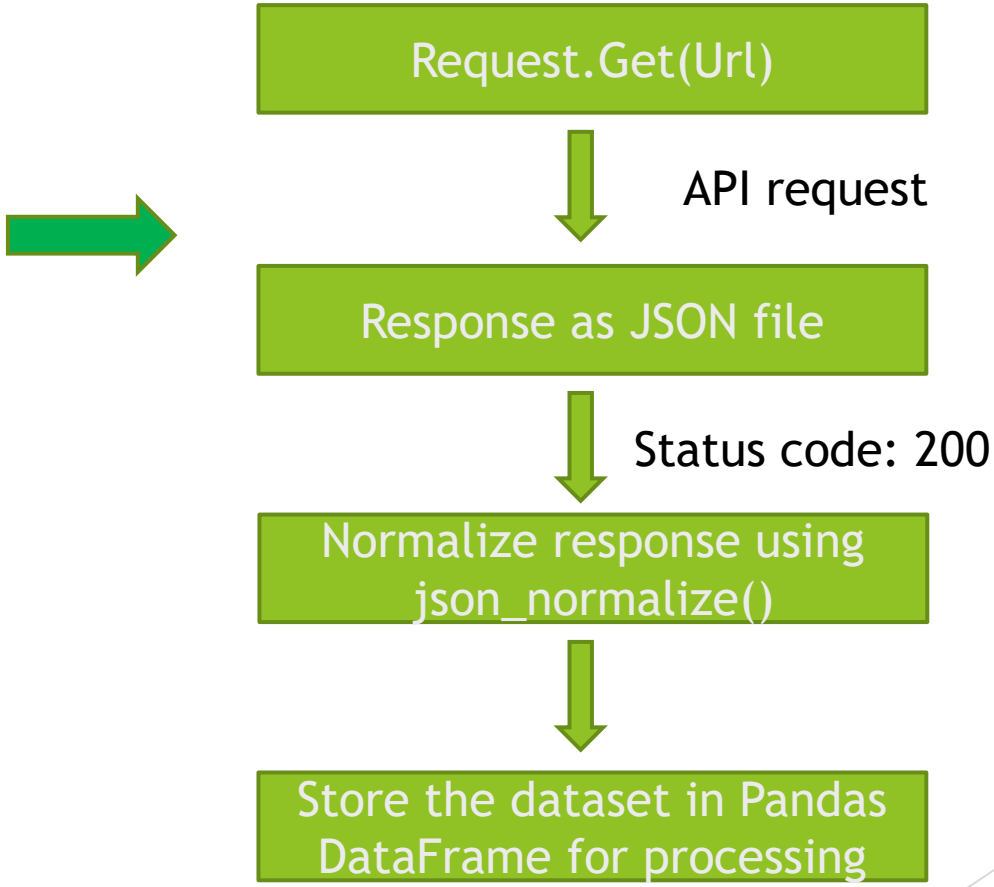
Data collection – SpaceX API

Added a flowchart of SpaceX API calls here

1. Present your data collection with SpaceX REST calls using key phrases and flowcharts

2. GitHub URL of the completed SpaceX API calls notebook as an external reference and peer-review purpose

Git Hub Link



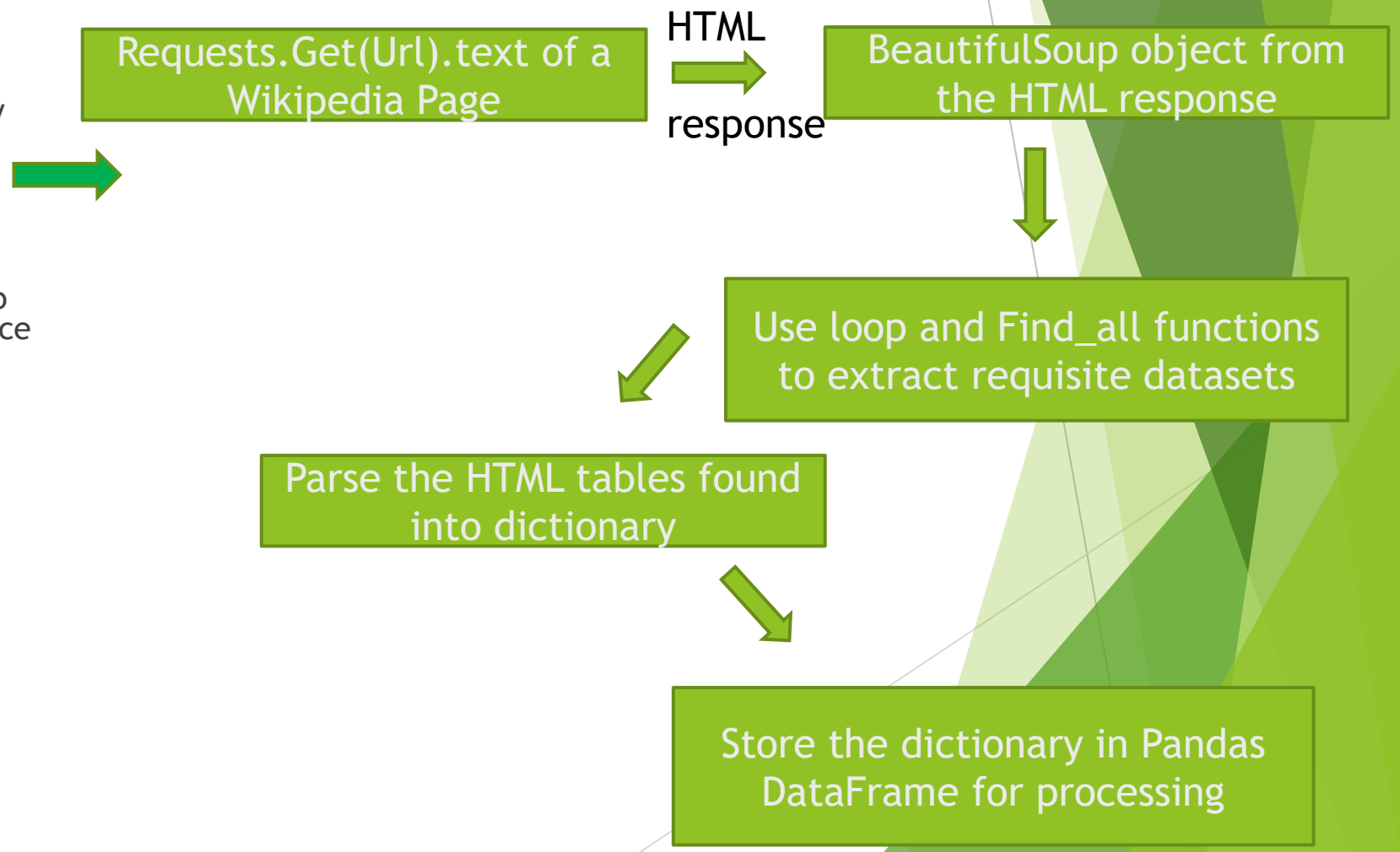
Add a flowchart of web scraping here

Data collection – Web scraping

Present your web scraping process use key phrases and flowcharts

Add the GitHub URL of the completed web scraping notebook, as an external reference and peer-review purpose

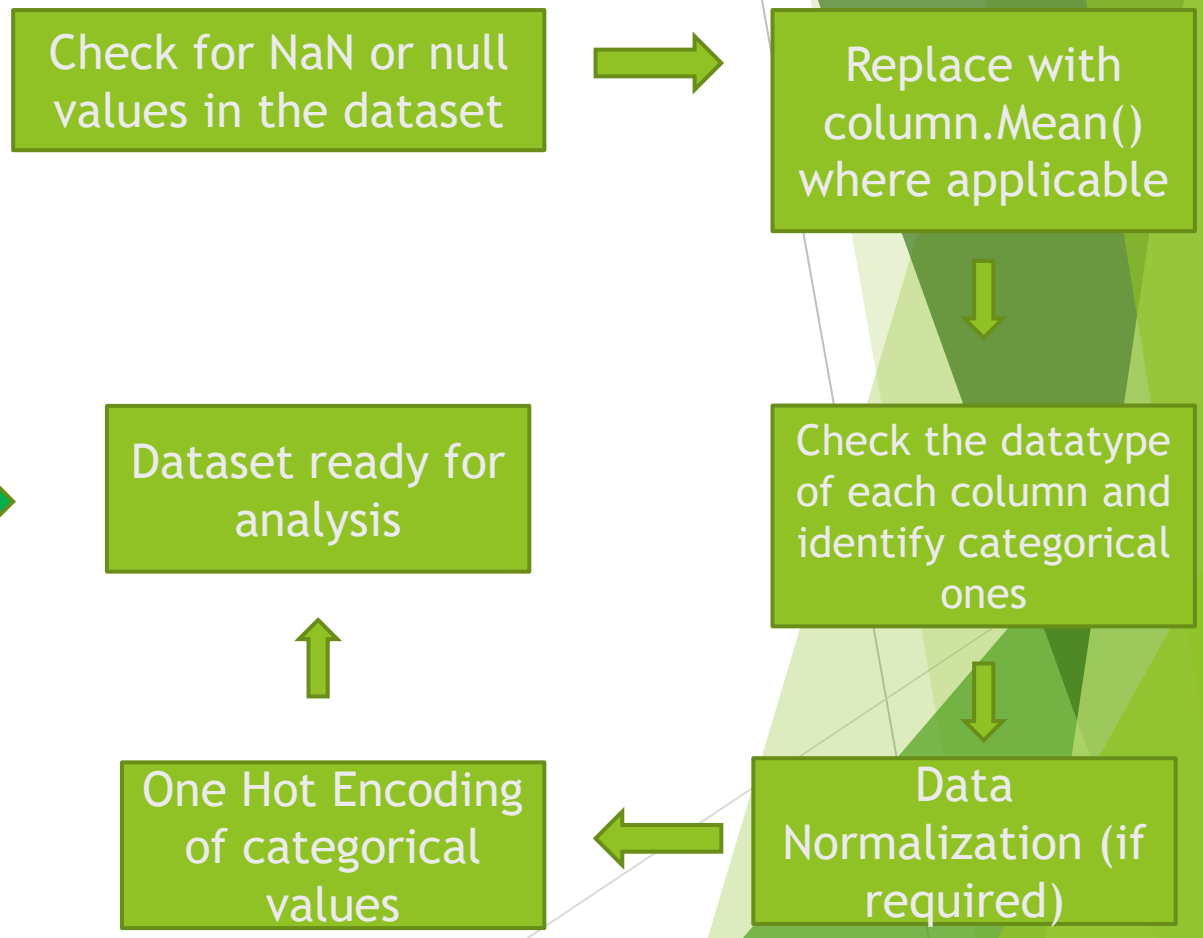
Git Hub Link



Data wrangling

- Describe how data were processed
  - *All the data collected was checked for missing values and then rectified by replacing with mean of values in the respective column*
  - *Data Formatting of various columns to improve data processing*
  - *Data normalization*
  - *One Hot encoding of categorical independent variables for ML process.*
- You need to present your data wrangling process using key phrases and flowcharts
- Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

Git Hub URL



EDA with data visualization

- Summarize what charts were plotted and why used those charts
  - Scatter plot - To see the impact of two parameters on a dependent variable via color coding
  - Bar chart – To visualize effect of one categorical parameter on a continuous one
  - Line graph – To observe trend of dependent variable based on an independent one.

- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

Git Hub URL

EDA with SQL

- Summarize performed SQL queries using bullet points
  - ‘Select’ queries to display variety of combination of columns
  - Use of ‘where’ and ‘like’ commands to use filters
  - Use of subquery to allow complex filters
  - Use of mathematical operations like AVG(), MIN(), COUNT(), etc.
  - Use of ‘Order BY’, ‘GROUP BY’, ‘RANK()’ functions to apply pivot like functions of table.
- Add the GitHub URL of your completed EDA with SQL notebook, as an external reference and peer-review purpose

Git Hub URL

Build an interactive map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Circle Markers created for each launch site
  - Marker clusters created for multiple markers generated on each site
  - Color coding of markers
  - Lines
  - Labels
- Explain why you added those objects
  - Circle Markers – to highlight area of site
  - Marker clusters – to increase visibility and comprehension of markers at a point
  - Color coding of markers to indicate success and failure in first stage landing for each site in our dataset
  - Lines – to show distance from closest railway or coastline to launchsite
  - Labels – to increase readability of maps

- Add the GitHub URL of your completed interactive map with Folium map, as an external reference and peer-review purpose

Git Hub URL

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Dropdown box
  - Pie chart
  - Range Slider
  - Scatter Plot

- Explain why you added those plots and interactions
  - Dropdown box – to select the site or sites for analysis
  - Pie chart – to provide view of success rate at each site
  - Range Slider – to choose the payload mass range for analysis
  - Scatter Plot – to depict the relation between payload mass and launch site on the success rate of landing

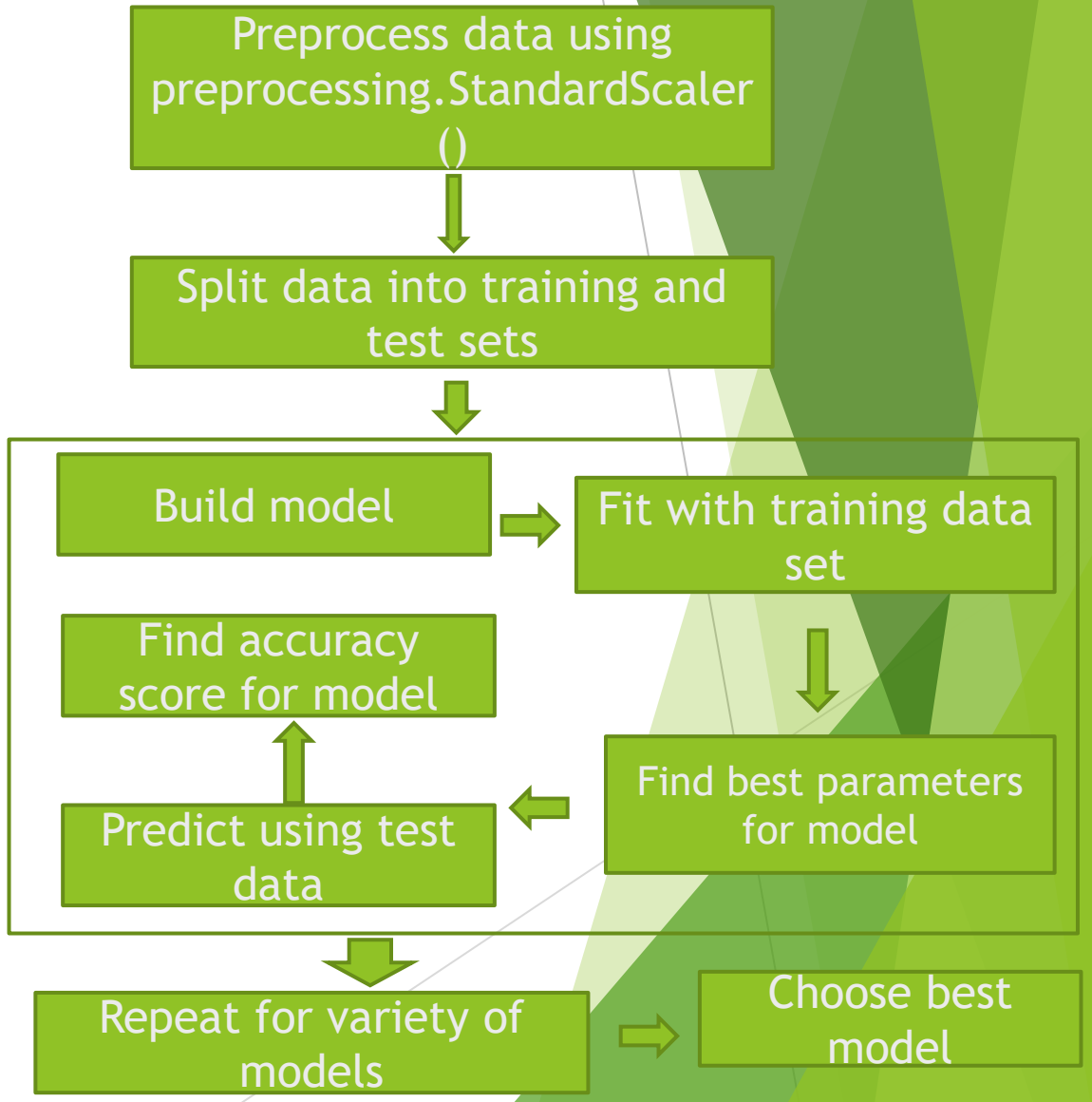
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Git Hub URL

Predictive analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  - Preprocess the data using StandardScaler()
  - Split dataset into training and test data
  - Create variety of models, fit them with training data, predict using test data.
  - Evaluate score of each model and pick model with highest accuracy
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Git Hub URL



Results

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

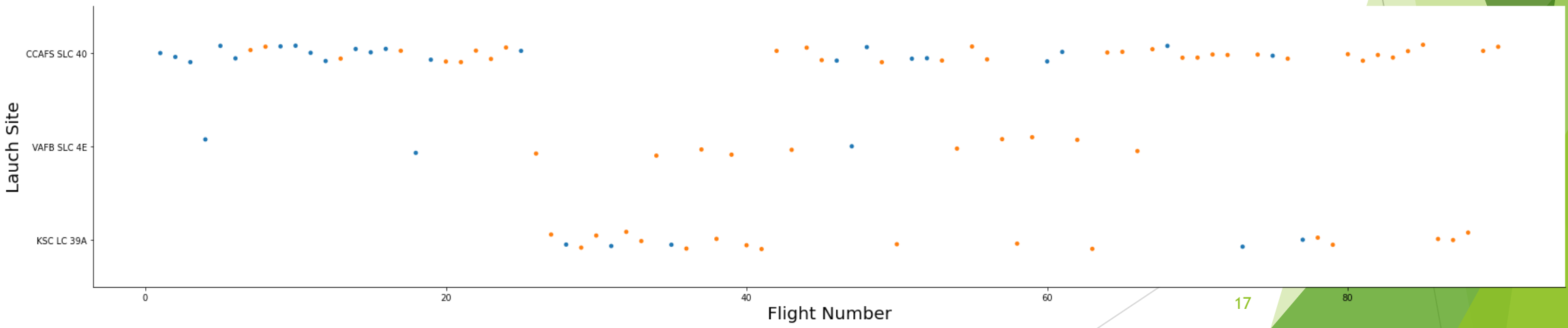
EDA with Visualization

Flight Number vs. Launch Site

Show a scatter plot of Flight Number vs. Launch Site

Show the screenshot of the scatter plot with explanations

- Below graph indicates that as flight numbers have increased chances of success has increased.
- Further, CCAFS SLC 40 is predominantly used as the launch site, so it has a lower success rate, but the other two sites have a higher success ratio due to relatively lower flight no.s

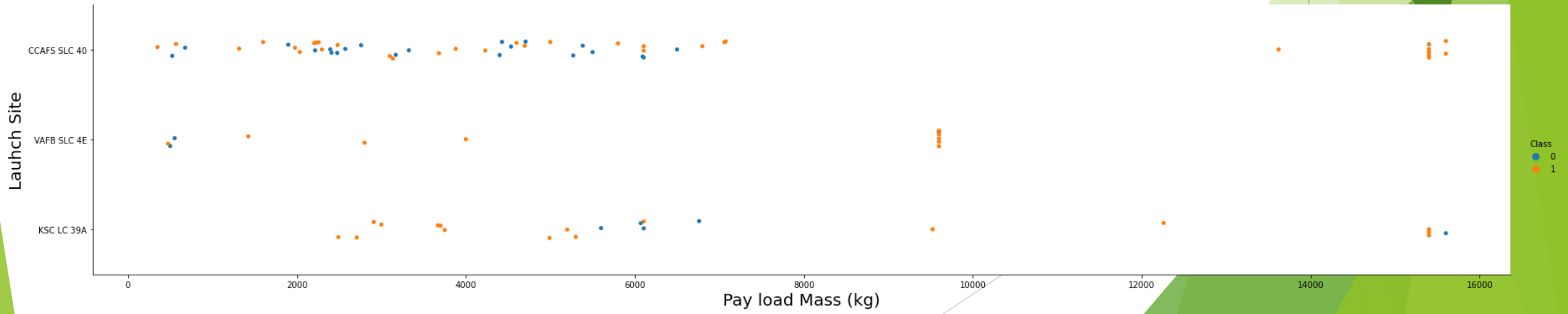


Payload vs. Launch Site

Show a scatter plot of Payload vs. Launch Site

Show the screenshot of the scatter plot with explanations

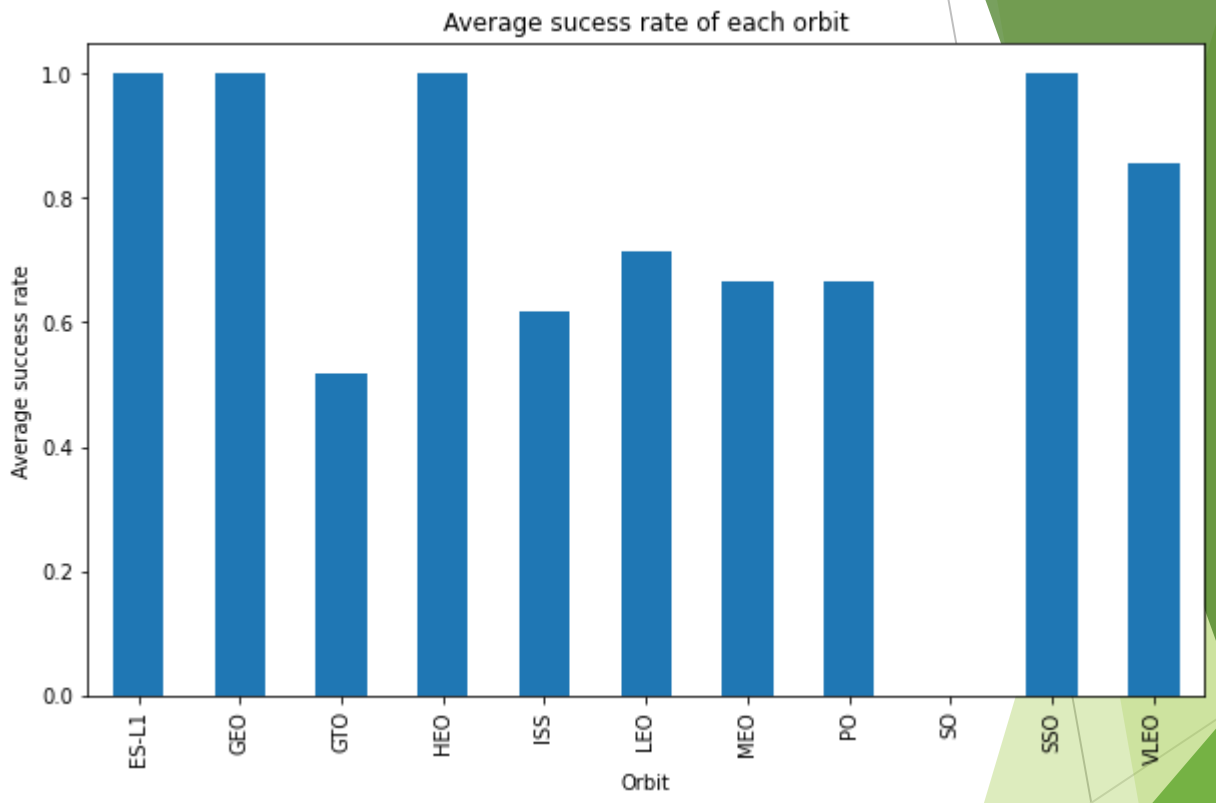
- Success rate seems to increase with higher payload mass; meanwhile CCAFS SLC 40 & KSCLC 39A outperform VAFB SLC4E at higher payload mass points
- At lower payload mass (<7000 kg), KSC LC 39 A has the best performance. Hence on an overall KSC LC 39 A seems to perform better at given payloadmass in terms of success outcome



Success rate vs. Orbit type

Show a barchart for the success rate of each orbit type

Show the screenshot of the scatter plot with explanations



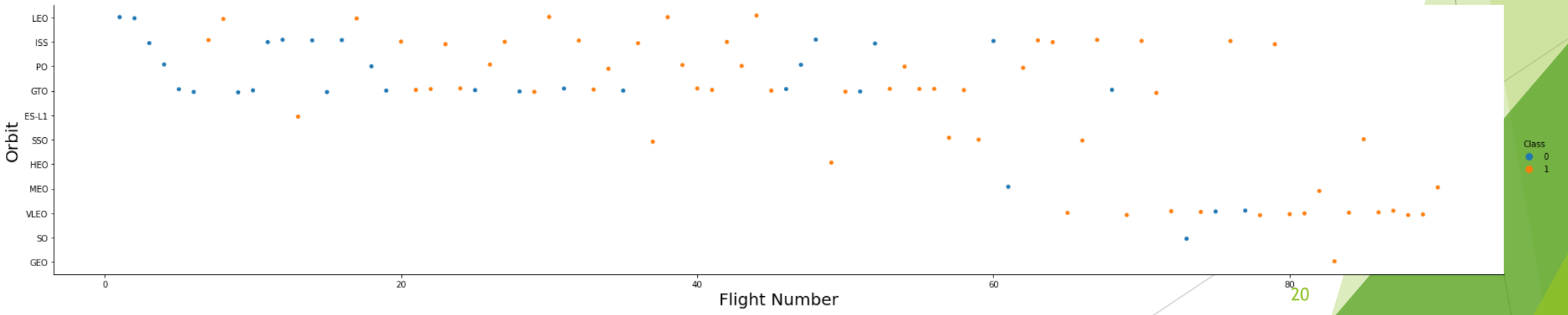
- We find that the Orbits ES-L1, GEO, HEO and SSO have 100% average success rates
- While GTO orbit has lowest average success rate of ~50%

Flight Number vs. Orbit type

Show a scatter point of Flight number vs. Orbit type

Show the screenshot of the scatter plot with explanations

- We see that the orbits VLEO, SSO, MEO and GEO have high success rates at higher flight numbers (>60)
- The orbits LEO, ISS, PO and GTO function better at mid range flight number (20-70)

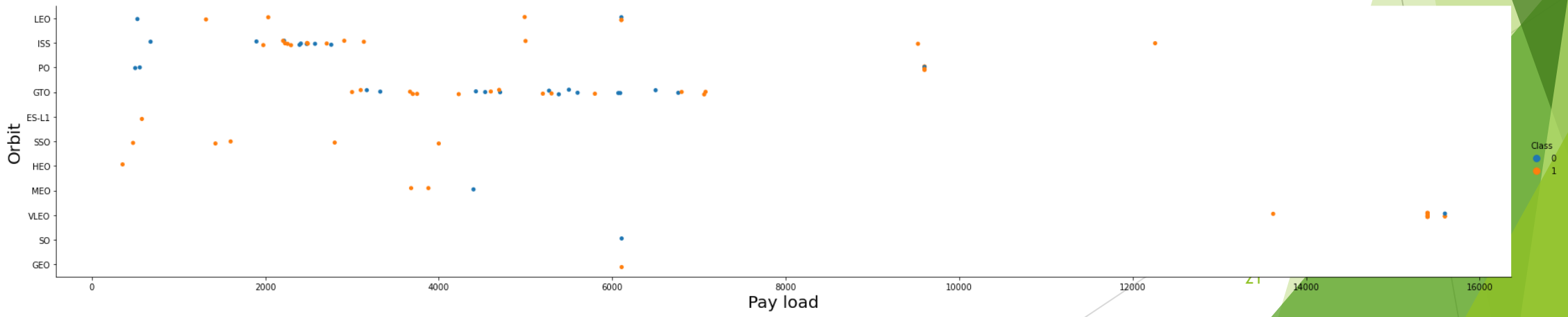


Payload vs. Orbit type

Show a scatter point of payload vs. orbit type

Show the screenshot of the scatter plot with explanations

We observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

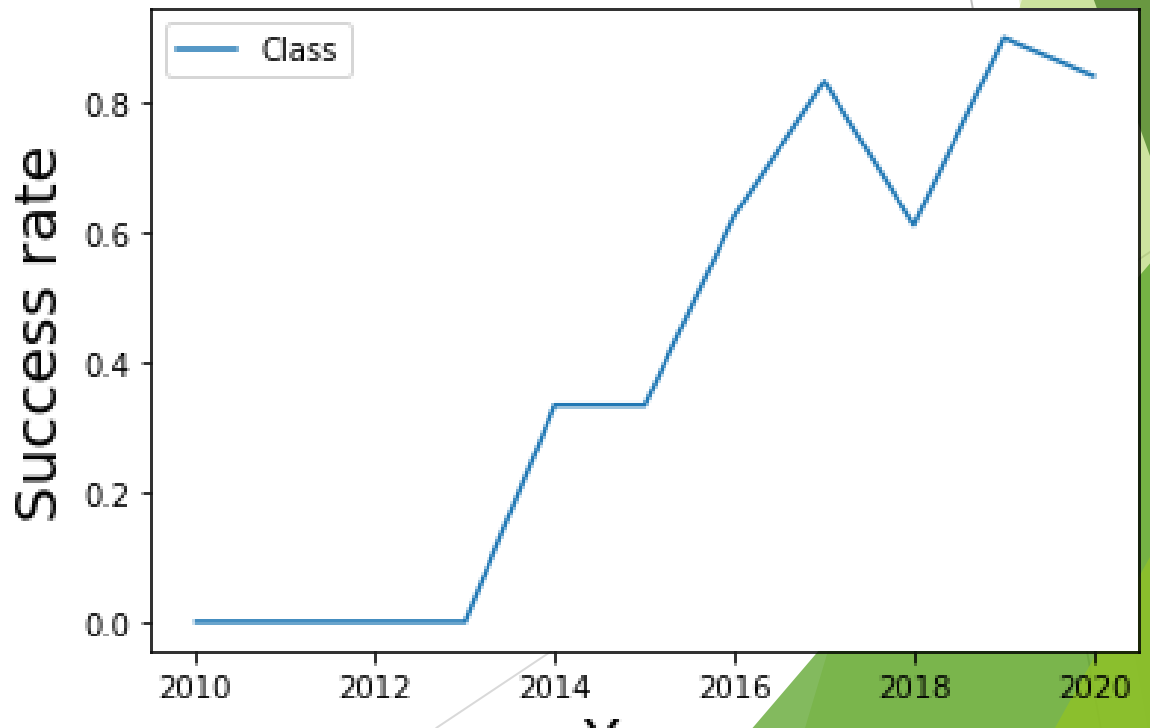


Launch success yearly trend

Show a line chart of yearly average success rate

Show the screenshot of the scatter plot with explanations

- We observe that the success rate starts rising since 2013 and kept increasing till 2020
- Success rate peaks in 2019 with above 80%



EDA with SQL

All launch site names

- Find the names of the unique launch sites

```
%sql select unique(LAUNCH_SITE) from SPACEXTBL
```

- Present your query result with a short explanation here
  - We have created a slice of the table with only unique launch site names.

Done.

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Launch site names begin with `CCA`

- Find all launch sites begin with `CCA`

```
%sql select * from SPACEXTBL where LAUNCH_SITE LIKE 'CCA%'
```

- Present your query result with a short explanation here
  - Filters the table data for launch sites starting with 'CCA', i.e., CCAFS LC- 40 and CCAFS SLC - 40

done.

Out[17]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total payload mass

- Calculate the total payload carried by boosters from NASA

```
: %sql select SUM(payload_mass__kg_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
```

- Present your query result with a short explanation here
  - Shows the sum of payload mass from NASA over the complete time period captured in the table. The total payload = 45596 Kg

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

1
45596

Average payload mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
-
```
: %sql select AVG(payload_mass__kg_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```
- Present your query result with a short explanation here
  - We have calculated the average payload mass carried by F9 v 1.1 over the complete time period captured in the table. The average payload = 2928 kg

1
2928

First successful ground landing date

- Find the date when the first successful landing outcome in ground pad

```
%sql select MIN(DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)'
```

- Present your query result with a short explanation here
  - As per the dataset available, the first successful landing in ground pad has occurred on 22nd Dec 2015.

1
2015-12-22

Successful drone ship landing with payload between 4000 and 6000

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_ Between 4000 and 6000
```

```
%sql select payload from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_ Between 4000 and 6000
```

- Present your query result with a short explanation here
  - Since we do not have a specific column for booster names we have selected a list of booster version and payloads relevant to the given condition

booster_version	payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

Total number of successful and failure mission outcomes

```
%sql select MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS count from SPACEXTBL GROUP BY MISSION_OUTCOME
```

- Present your query result with a short explanation here
  - As per the dataset available, 99 out of 101 instances have been completely successful mission outcomes, while 1 is unclear on payload status but is considered a success.

mission_outcome	COUNT
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters carried maximum payload

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

```
%sql select BOOSTER_VERSION from SPACEXTBL where payload_mass__kg_ = (SELECT  MAX(payload_mass__kg_) from SPACEXTBL)
```

- Present your query result with a short explanation here
  - We have obtained a complete list of booster versions which have carried payload mass of maximum weight.
  - This has been achieved using a sub query to select 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

2015 launch records

- 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

```
%sql select MONTHNAME(DATE) as Month_name, LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL where YEAR(DATE) = 2015 and LAND
ING__OUTCOME = 'Failure (drone ship)'
```

- Present your query result with a short explanation here
  - As directed, we have 4 columns denoting month name, landing outcome of failure in drone ship, booster versions and the corresponding launch site. These are only for relevant failures in 2015.

month_name	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank success count between 2010-06-04 and 2017-03-20

- Rank the count of successful landing_outcomes between the date 2010-06-

```
%sql SELECT landing__outcome, COUNT(landing__outcome) as COunt, RANK() OVER (ORDER BY COUNT(landing__outcome) DESC) as ORDER from SPACEXTBL where landing__outcome LIKE 'Success%' And DATE Between '2010-06-04' and '2017-03-20' GROUP BY landing__outcome
```

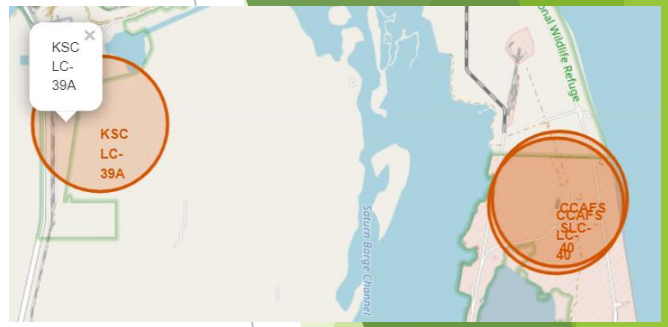
- Present your query result with a short explanation here
  - For the given time range, we find that the success in drone ship has the highest successful outcomes and the ranking in descending order of count has been shown.

landing__outcome	COUNT	ORDER
Success (drone ship)	5	1
Success (ground pad)	3	2

Interactive map with Folium

Launch Site Identification with Folium

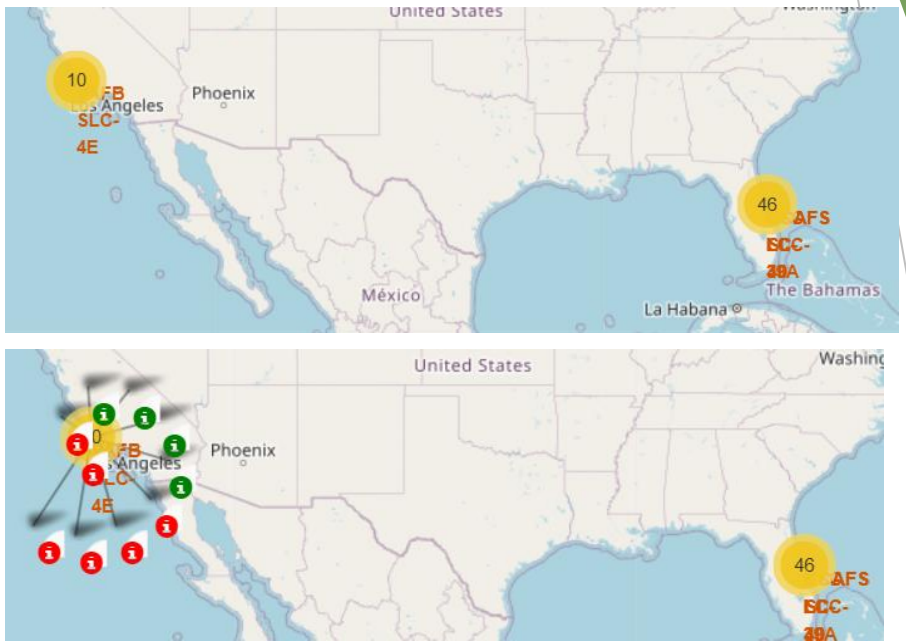
- Replace <Folium map screenshot 1> title with an appropriate title
- Show the screenshot of all launch sites' location markers on a global map
- Explain the important elements and findings on the screenshot



- We have marked the 4 unique launch sites with circle markers of radius 1000 units
- We have also provided labels with corresponding site names and pop ups of same.

Marker colored cluster map

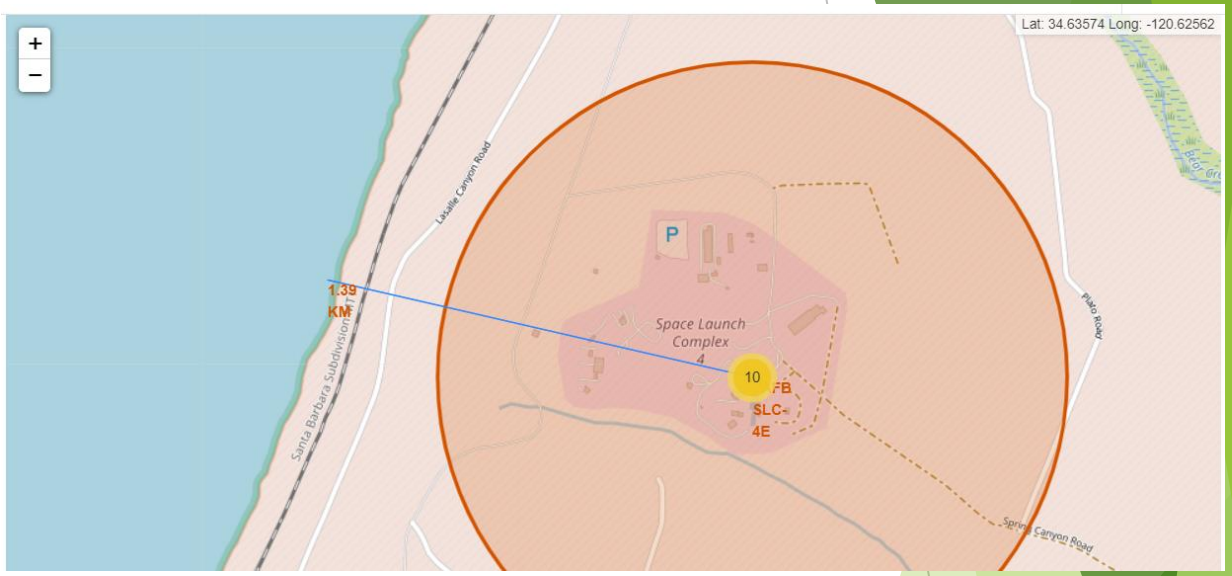
- Replace <Folium map screenshot 2> title with an appropriate title
- Show the screenshot of color-labeled launch records on the map
- Explain the important elements and findings on the screenshot



- Marker clusters have been created for visual ease
- Each launch record has a pointer on map with color code basis success or failure of corresponding data point.

Geographical proximity depiction

- Replace <Folium map screenshot 3> title with an appropriate title
- Show the screenshot of a selected launch site to its proximities such as railway, highway, coastline, with distance calculated and displayed
- Explain the important elements and findings on the screenshot

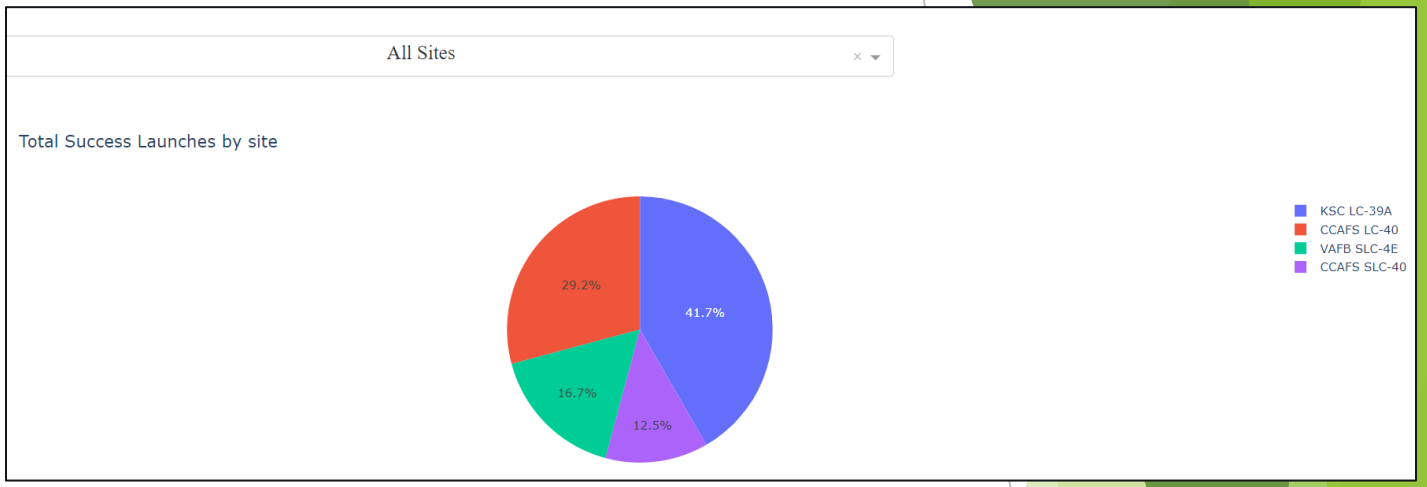


- We have calculated the distance from launch site to nearest coastline by choosing the point using mouse pointer.
- The blue line shows the straight distance from the two points on map and label depicts the distance shown by line.

Build a Dashboard with Plotly Dash

Total Success Launches by Site

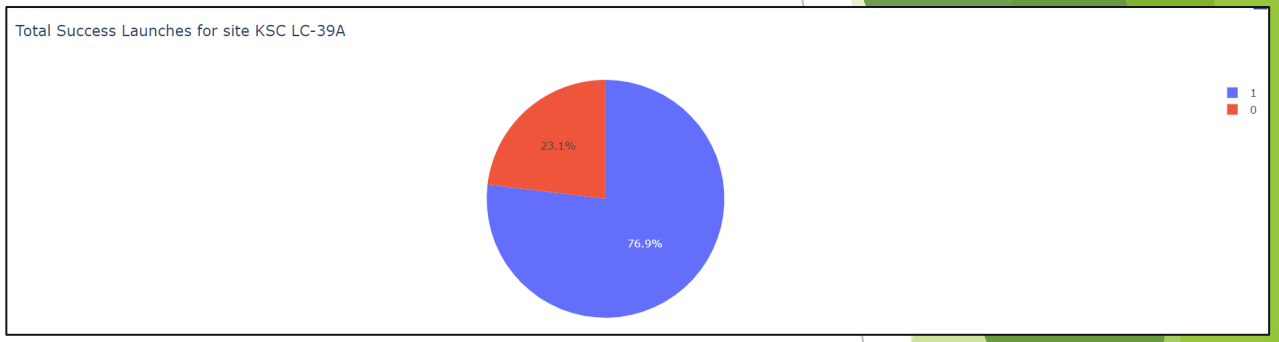
- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot



- As per the pie chart, we find that the launch site 'KSC LC -39A' has the highest success rate of 41.7% among all sites
- This is followed by CCAFS LC -40 site with 29.2%.
- However, no information on failure rate of these sites can be visualized here.

Success Launches for Individual sites

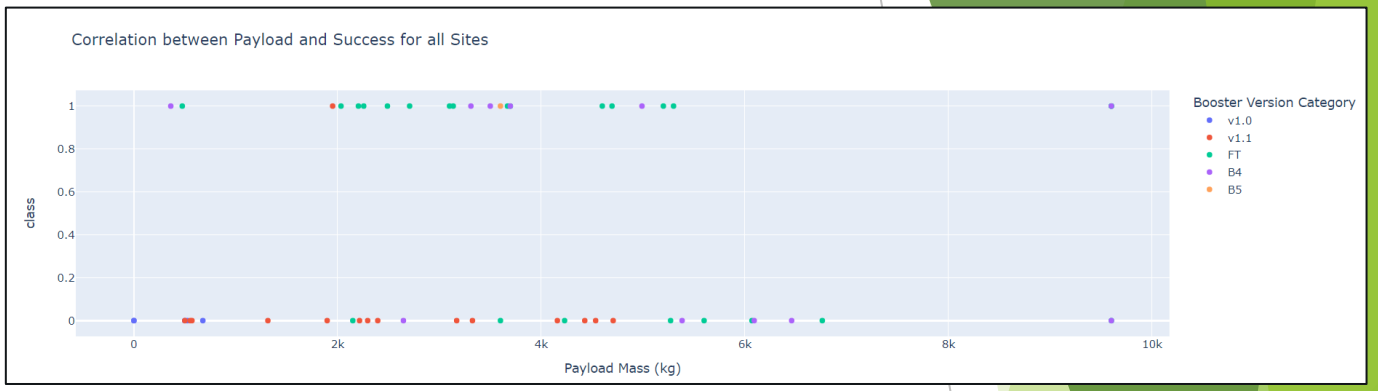
- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot



- As per the pie chart, we find that the launch site ‘KSC LC -39A’ a success rate of 76.9% vs a failure percentage of 23.1%
- This provides a concentrated view of the success rates based on individual sites only.

Correlation between payload & success

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot



- As per the scatter chart, we find that the there is a lower chance of success at lower payload mass
- Most payloads above 5500 Kg have resulted in failures across sites.

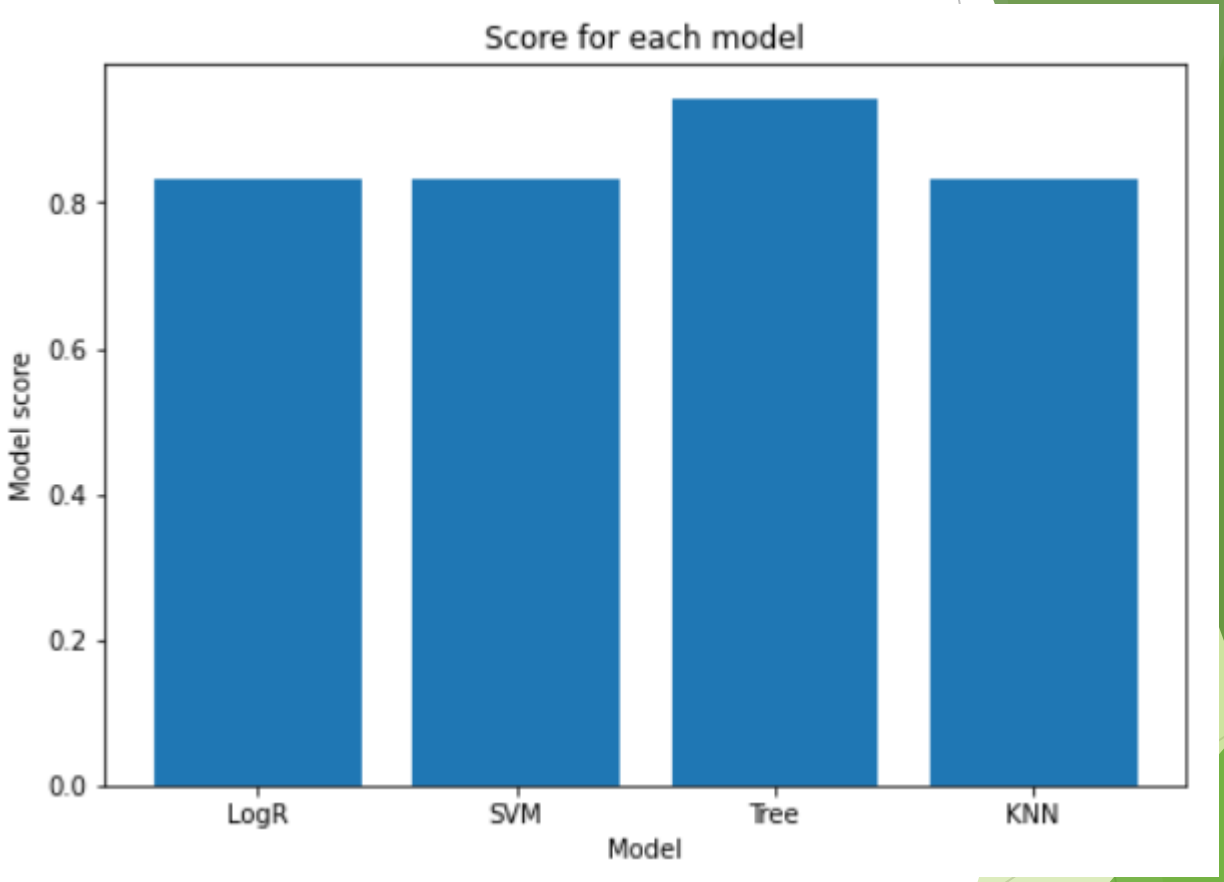
Predictive analysis (Classification)

Classification Accuracy

Visualize all the built model accuracy for all built models, in a barchart

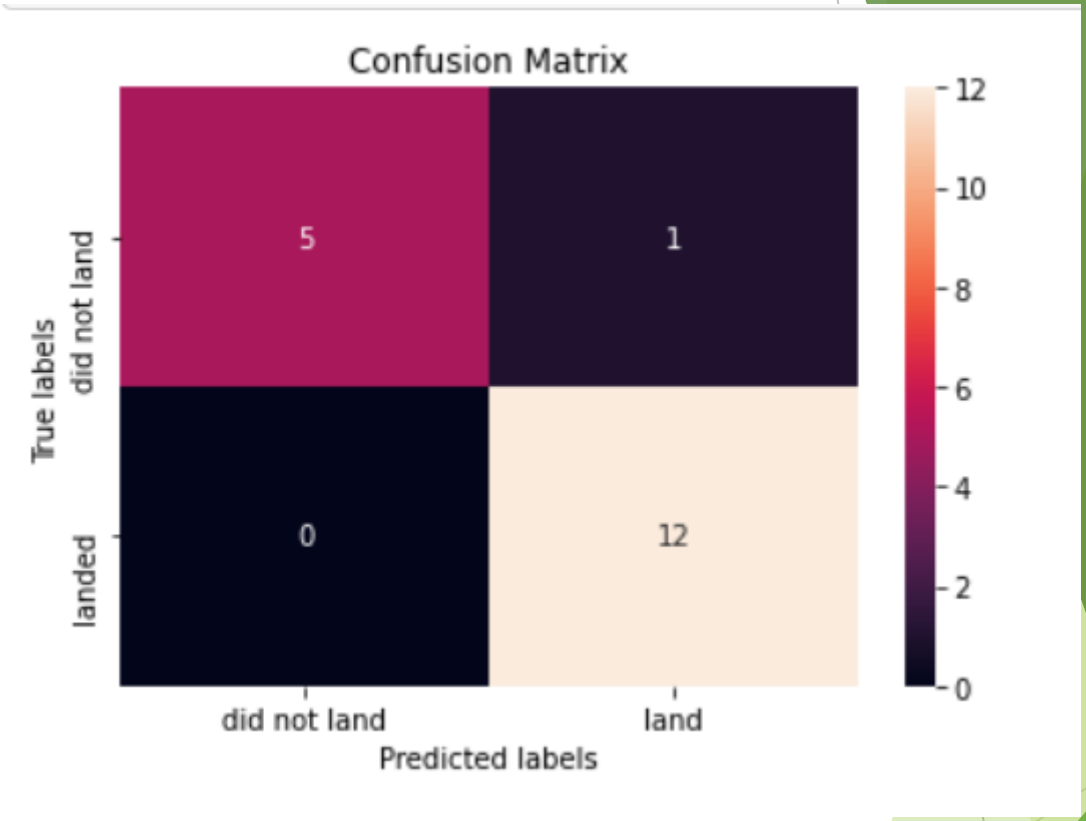
Find which model has the highest classification accuracy

- Based on the bar chart created for the accuracy score of the 4 models, we find that the Decision Tree model has the highest accuracy of 0.94



Confusion Matrix

Show the confusion matrix of the best performing model with explanation



- The confusion matrix for the Tree model depicts that it successfully predicts all 12 True positive answers for successful landings.
- However, among 6 failed landings, the model only predicts 5 instances correctly while 1 is a False positive.
- This leads to a method score of 0.94

CONCLUSION

- Visualization Analysis:
  - As flight numbers have increased chances of success has increased. Also, success rate seems to increase with higher payload mass
  - Few orbits have 100% accuracy rate such as ES-L1, GEO, HEO and SSO, while the lowest is GTO at 50%.
  - We observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.
  - We observe that the success rate starts rising since 2013 and kept increasing till 2020 with peak in 2019 at above 80%.
- Based on SQL, we find that majority of the mission outcomes are successful, i.e., 99 out of 101, which is commendable.
- Folium graph plotting:
  - The launch sites appear to be in proximity of important geographical features such as coastline (helps in bringing back landed first stage), railways (helps in goods and people transport), etc.
- Dashboard:
  - We find that the most successful launch site is 'KSC LC -39A' along with the impact of payload at various sites, i.e., lower chance of success at lower payload mass.
- Prediction Analysis:
  - We find that the Decision Tree to be the best model for evaluation of the given data set with an accuracy score of ~94%.