IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ryan Desfosses
11 Oct 22

Outline


Executive Summary
• (Start)
Introduction
Methodology
• Data Collection
• Visual Analytics
• Predictive Analysis
Results
• Visualization
• SQL
• Interactive Map
• Dashboard
• Predictive Analysis
Conclusion
Appendix
• (Finish)

Executive Summary

- The object of the presentation is to come to a conclusion if the imaginary company SpaceY can predict the landing of the SpaceX's first stage rocket.

- To accomplish this previous SpaceX landing results were acquired from numerous sources and used to train Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors supervised machine learning models.

- In the end all 4 machine learning models returned a scored 0.833 when being run against the test set.

Introduction

- SpaceY needs to determine if the if a well know competitor will be able to land the first stage of a rocket.
- What factors determine if the rocket will successfully land?
- Are the factors reliable enough to predict if the rocket will land?

Section 1

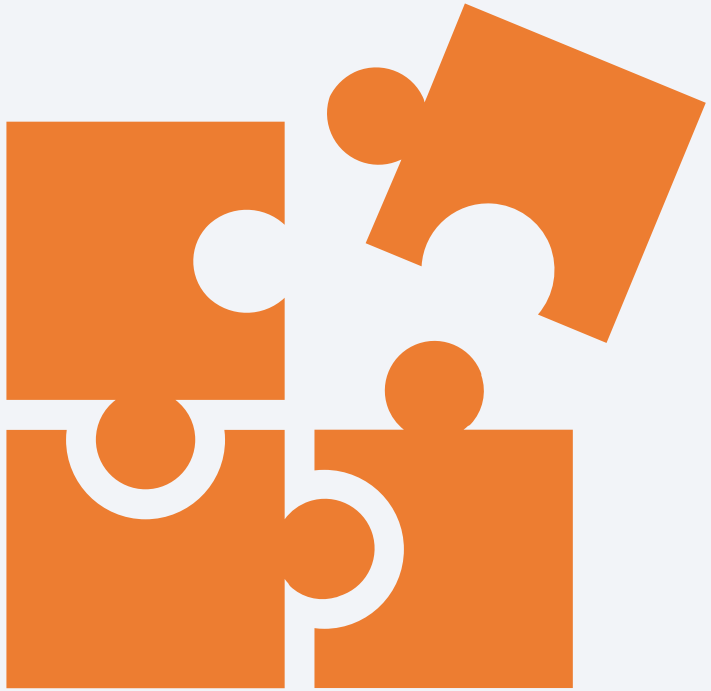
Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Acquired data via http requests
- Perform data wrangling
  - Converted outcomes into training labels where 1 was represented a successful landing and a 0 for an unsuccessful landing
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection



Data Collection Libraries

Below are the main Python Libraries used to collect the data requires for analysis

Requests

Download HTML

Pandas

Structure and Manipulate Data

Beautifulsoup4

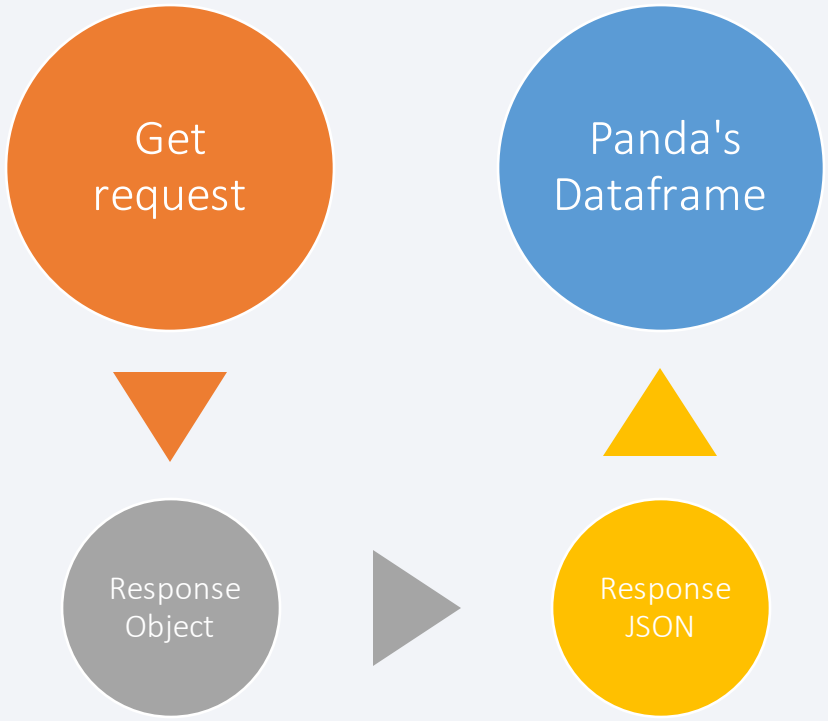
Parse HTML

IBM_DB_SA

interface to IBM Data Servers

Data Collection – SpaceX API

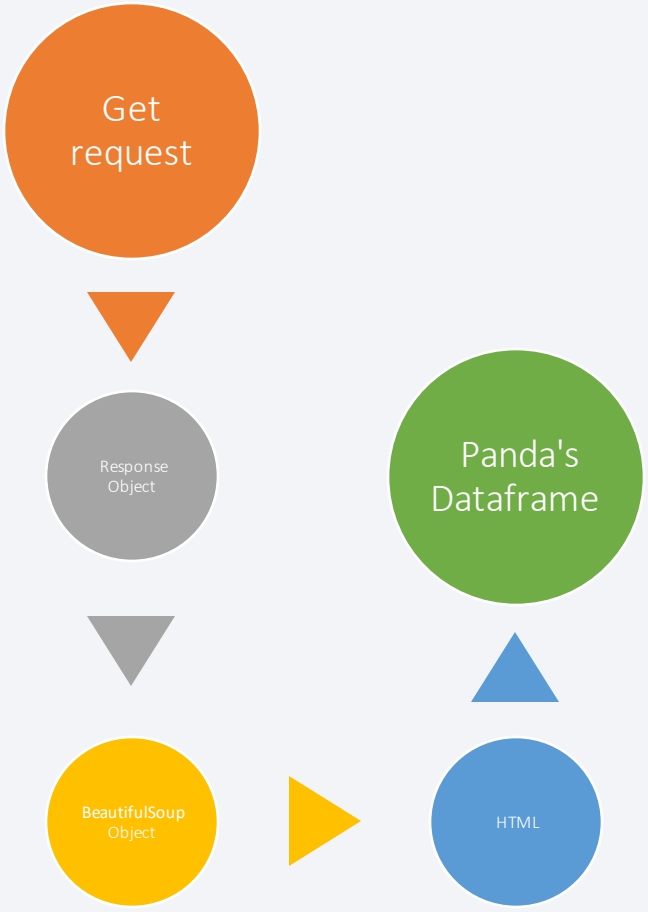
- Python uses the Request library to send a http get request to SpaceX API, which returns the desired data in a JSON format. The request library then saves the data in a request object that gets passed to Pandas library to be converted into a dataframe.



https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

- Python uses the Request library to send a http get request to Wikipedia html page saves as a response object.
- The Beautiful Soup library then converts to a beautiful Soup Object where **find** method can be used to attract the html.
- Then the text containing the desired data is parsed and saved inside a Pandas dataframe.



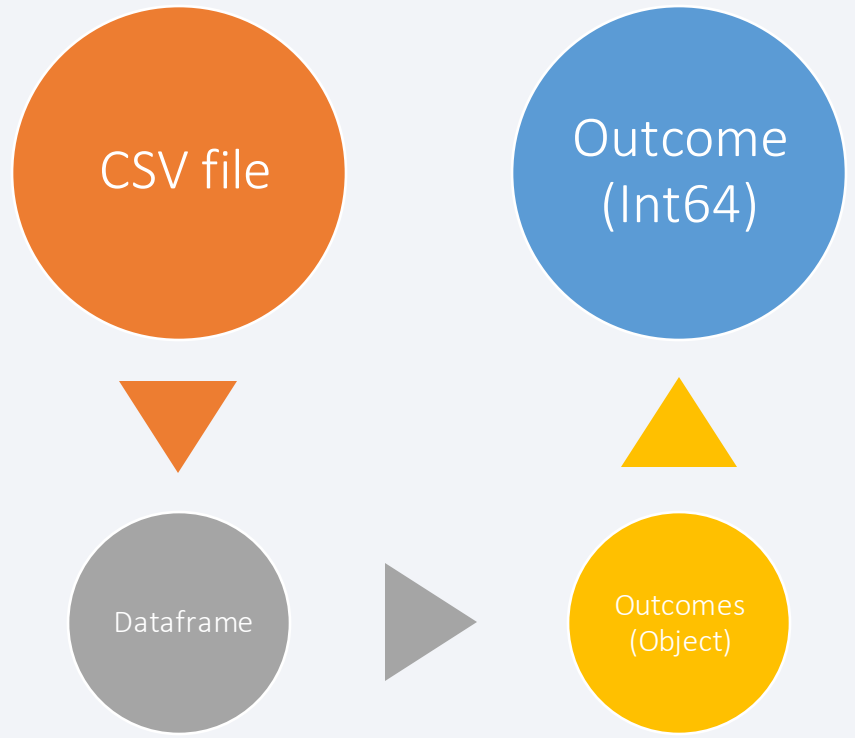
https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

- Pandas is used to read the csv file containing the landing outcomes and store the data in a dataframe, which is they used convert the outcome objects (strings) into integers.

- Once all the outcomes are converted the average is collected by using the **mean** method.

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/labs_jupyter_spacex_Data_wrangling.ipynb



EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
  - Visualize the relationship between Flight Number and Launch Site
  - Visualize the relationship between Payload and Launch Site
  - Visualize the relationship between success rate of each orbit type
  - Visualize the relationship between FlightNumber and Orbit type
  - Visualize the relationship between Payload and Orbit type

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was acheived.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass.

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

Source: folium Docs

- Map Objects
  - Lines – to show distance and mark the equator
  - Markers – to label points
  - Circles – list the number of points in the area

- Compared
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - success rate vs. orbit type
  - Flight Number vs. Orbit type
  - Payload vs. Orbit type

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Charts
  - Pie chart showed the success rate of the different launch sites
  - Scatter Chart that compared the payload mass and booster version with the launch outcomes
- Interactions
  - Pie charts
    - Offered a drop down to select: All, KSC LS-39A, CCAFS LC-40, VAFB SLC-4E, and CCAFS SLC-40
  - Scatter plot
    - Slider to select different payload size

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/spacex_dash_app.py

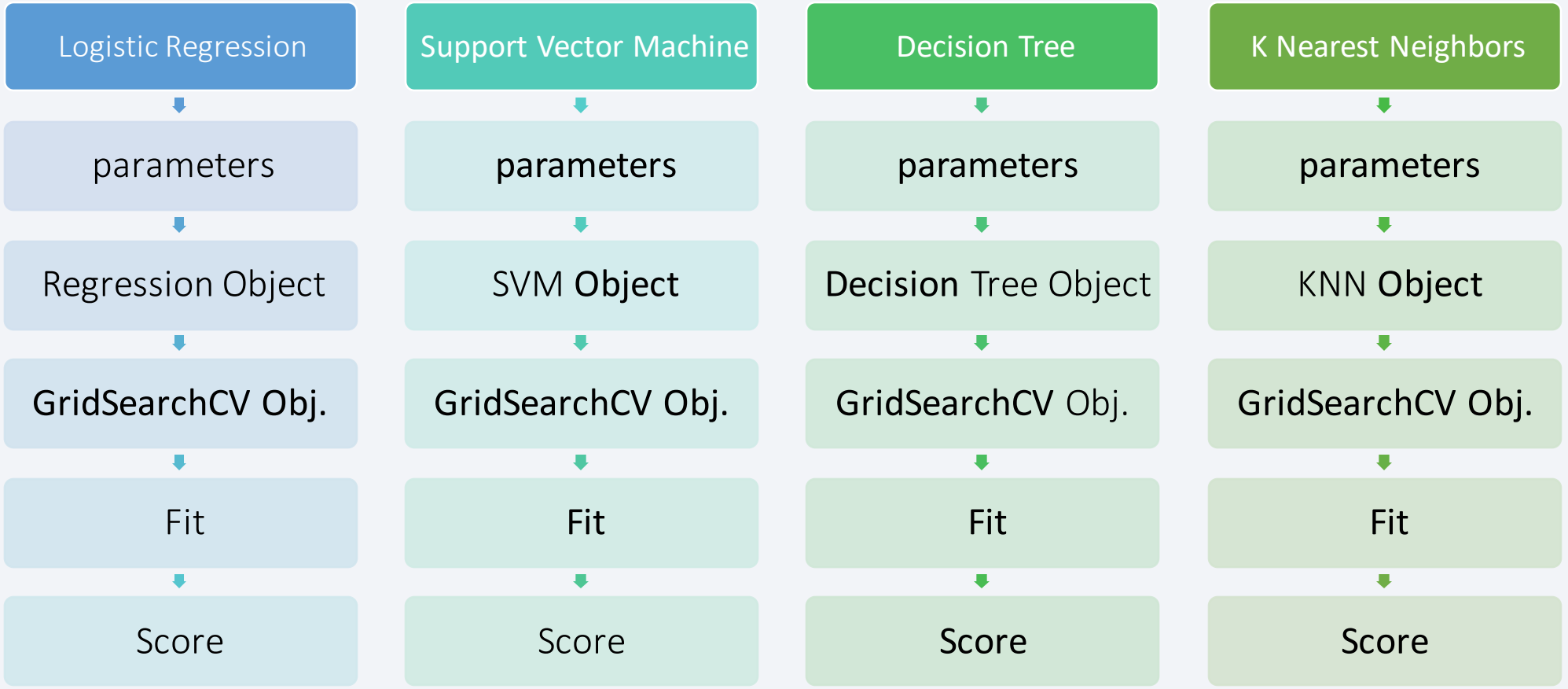
Predictive Analysis (Classification)

Once the X and Y data have been split into test and train sets. All four machine learning methods where fitted and scored.

(Chart on next Slide)

https://github.com/rdesfo/IBMAppliedDataScienceCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Predictive Analysis (Classification) Chart

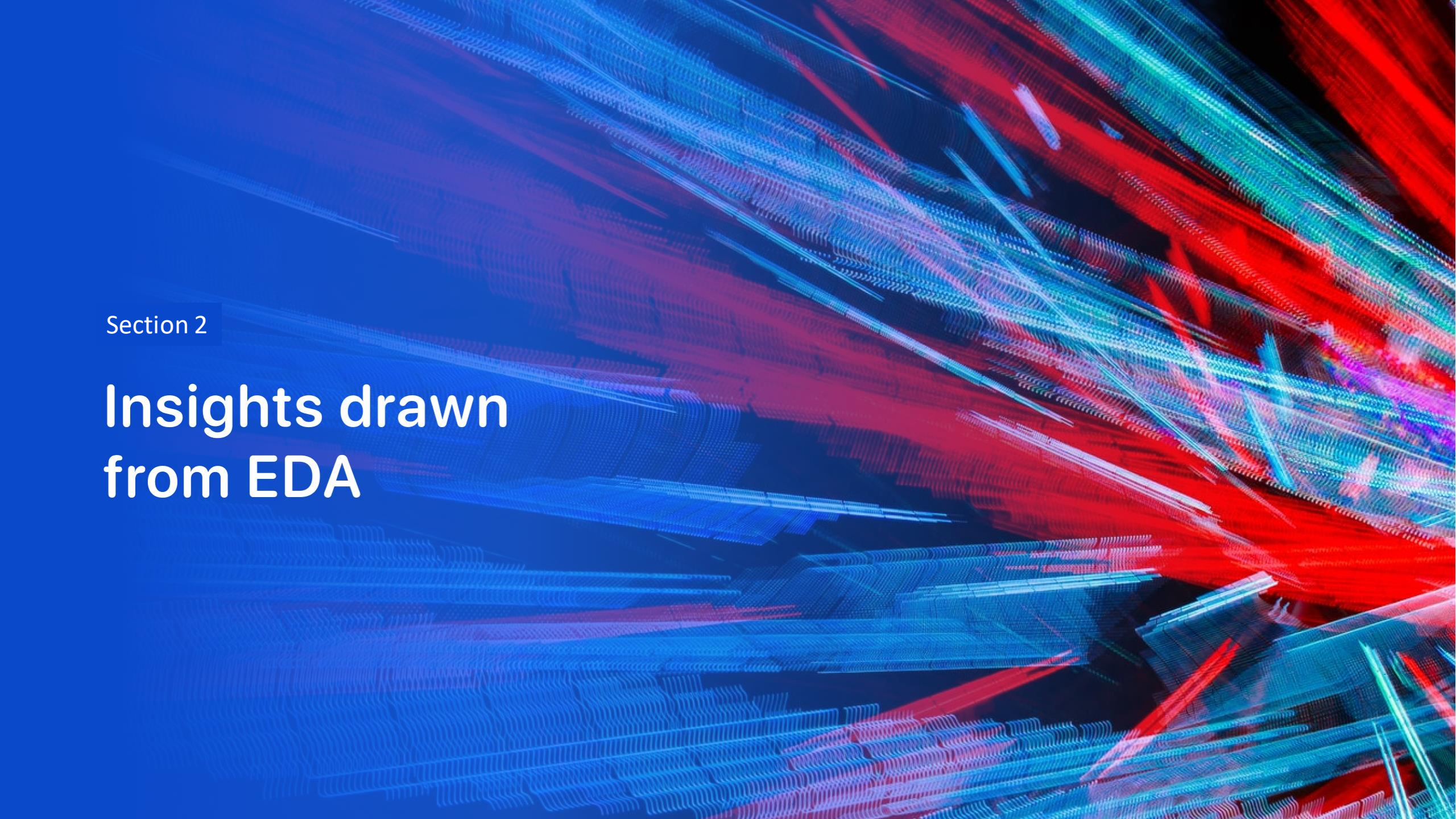


Results

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

Section 2

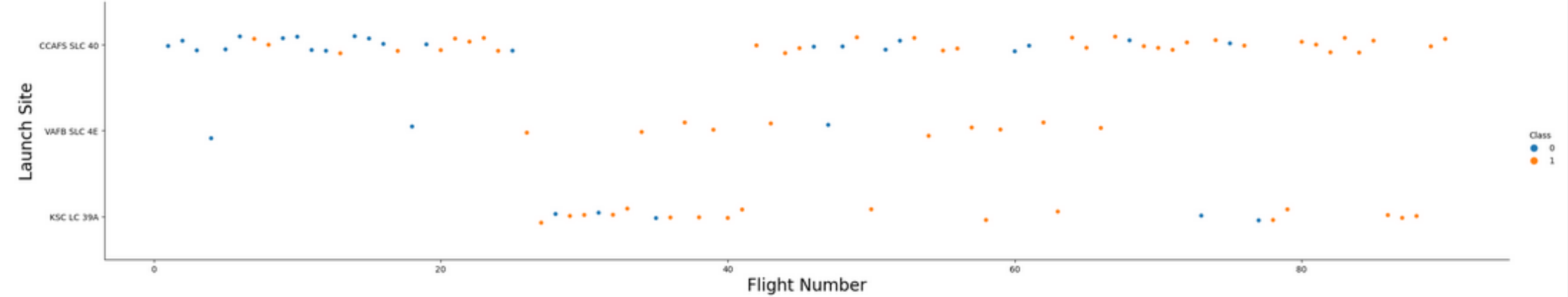
Insights drawn from EDA



Flight Number vs. Launch Site

The orange dots represent a successful landing and increase as the flight numbers increase.

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

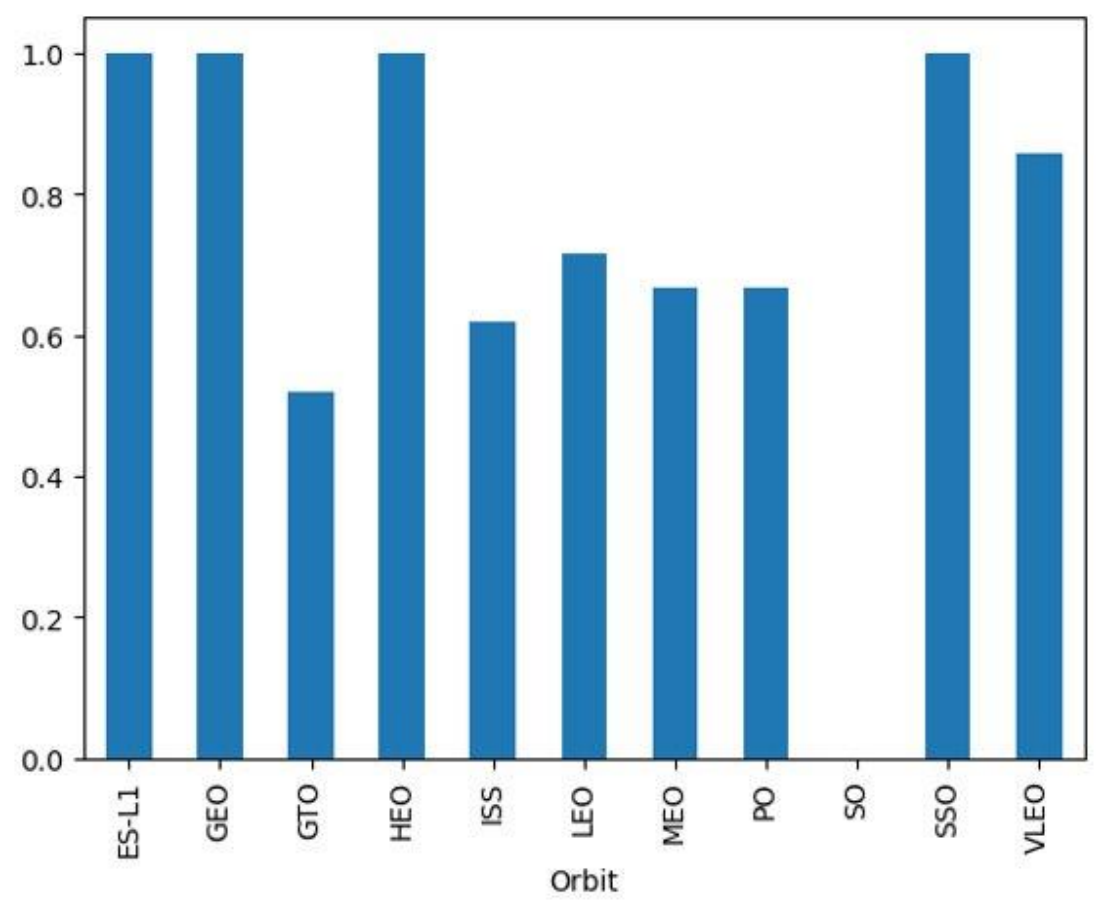


Payload vs. Launch Site

The VAFB-SLC launch site is used less often and doesn't attempt any payloads greater than 10000kg.

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the launch site, and hue to be the class value
sns.catplot(y="LaunchSite", x="PayloadMass", hue="Class", data=df, aspect = 5)
plt.xlabel("PayLoad Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

Success Rate vs. Orbit Type

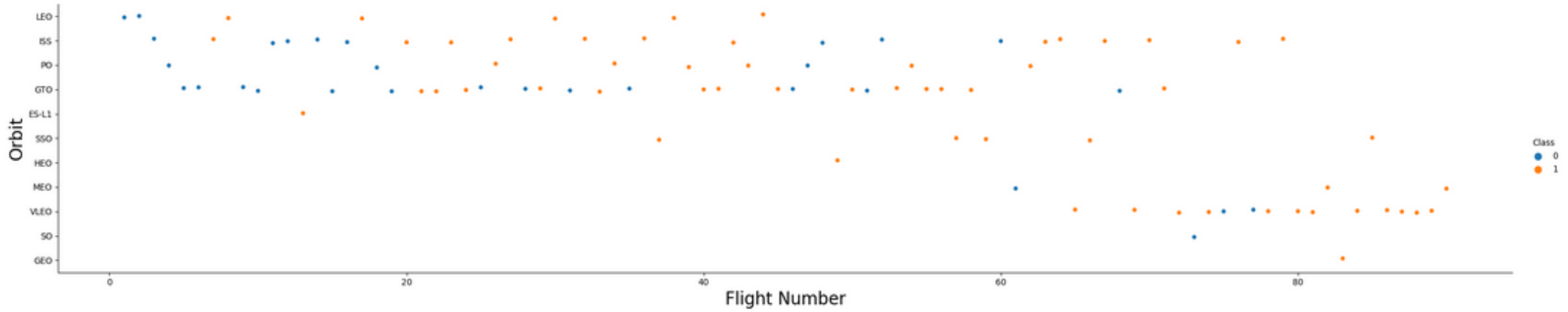


- ES-L1, GEO, and HEO orbits all had a perfect records
- GTO, ISS, LEO, MEO, and PO orbits had a success rate between 40 and 80 percent.
- SO had zero successful orbits.

Flight Number vs. Orbit Type

At the top, LEO through GTO are more common in the earlier flight numbers, which SO and VLEO are more common in the later flight numbers

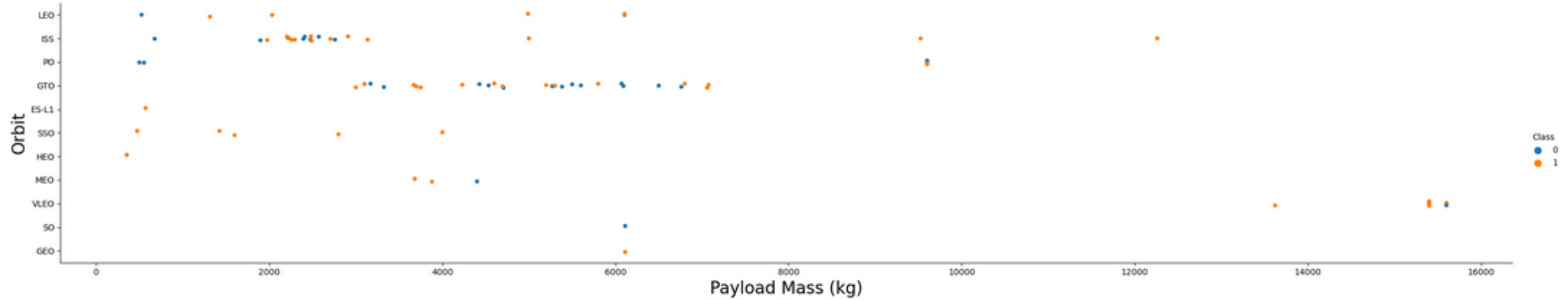
```
# Plot a scatter point chart with x axis to be FlightNumber and y axis to be the Orbit, and hue to be the class value
sns.catplot(x="FlightNumber", y="Orbit", hue="Class", data=df, aspect=5)
plt.xlabel("Flight Number", fontsize=20)
plt.ylabel("Orbit", fontsize=20)
plt.show()
```



Payload vs. Orbit Type

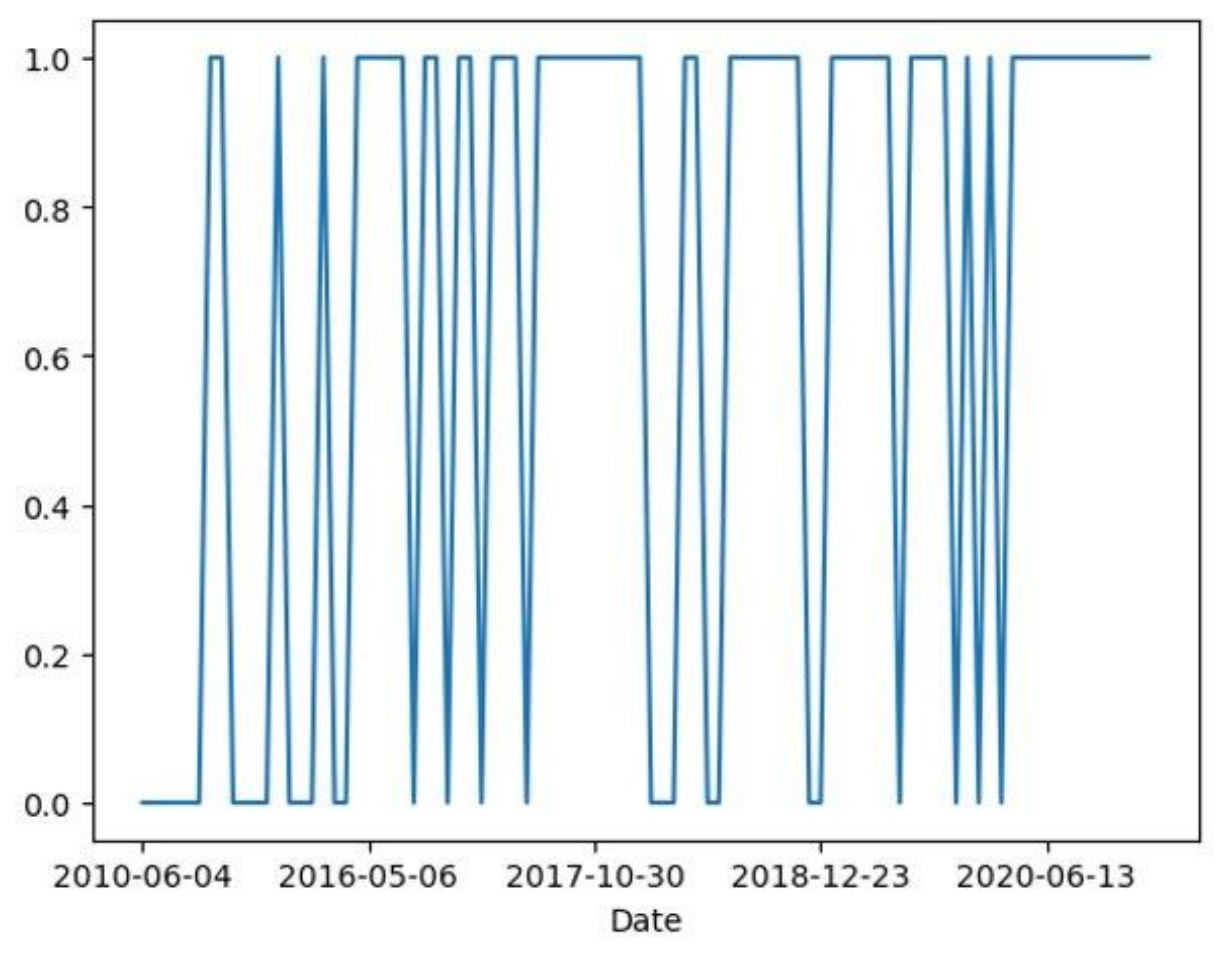
- GTO seems to be a common orbit for payloads between 2000kg and 8000kg, but experiences mix results
- LEO through MEO become less frequent after 8000kg

```
# Plot a scatter point chart with x axis to be Payload and y axis to be the Orbit, and hue to be the class value
sns.catplot(x="PayloadMass", y="Orbit", hue="Class", data=df, aspect=5)
plt.xlabel("Payload Mass (kg)",fontsize=20)
plt.ylabel("Orbit",fontsize=20)
plt.show()
```

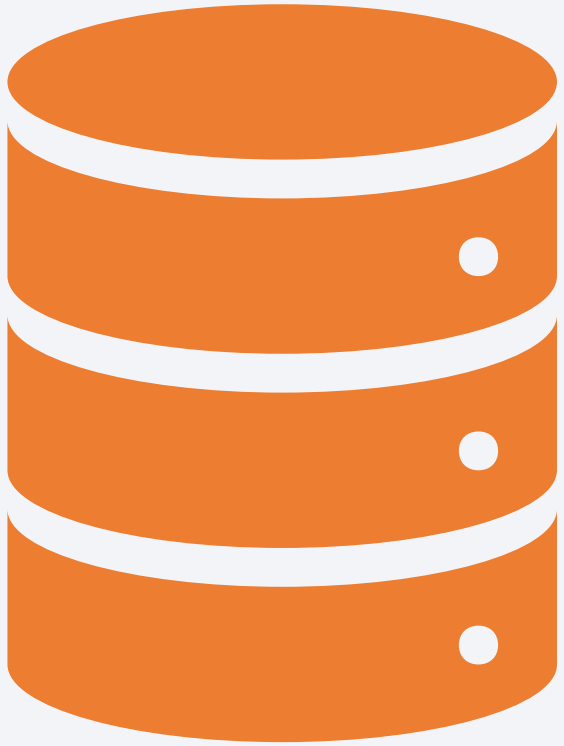


Launch Success Yearly Trend

- The Success rate since 2013 kept increasing till 2020
- The bottom of the chart illustrates the outcome change as well with the lines from multiple failure become points as the failures become less prevalent.



SQL Results



All Launch Site Names

- Selects the launch_site column and removes all the duplicates with the DISTINCT function to remove allthe duplicates

```
%%sql
select DISTINCT launch_site
from SPACEXTBL;
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.
Done.
```

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

Launch Site Names Begin with 'CCA'

The function above uses the **LIKE** and LIMIT functions to match the 'CCA' pattern and confine the output to 5 results

```
%%sql
select *
from SPACEXTBL
where launch_site LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

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

- Uses the SUM function to add all the payload masses together and renames the column 'TOTAL_PAYLOAD_MASS'

```
%%sql
select SUM(payload_mass__kg_) as TOTAL_PAYLOAD_MASS
from SPACEXTBL
where customer LIKE 'NASA (CRS)'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- Uses the **AVG** function to average the payload mass carried by the F9 v1.1 booster

```
%%sql
select AVG(payload_mass__kg_) as F9v1_1_Average_PAYLOAD_MASS
from SPACEXTBL
where booster_version LIKE 'F9 v1.1'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

f9v1_1_average_payload_mass
2928

First Successful Ground Landing Date

- Finds the earliest date by using the **MIN** function

```
%%sql
select MIN(DATE) as FIRST_Successful_groudpad_landing
from SPACEXTBL
where landing__outcome LIKE 'Success (ground pad)'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

first_successful_groudpad_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query uses the **BETWEEN** function to limit the output to values greater than 4000 and less than 6000.

```
%%sql
select booster_version
from SPACEXTBL
where (payload_mass__kg_ BETWEEN 4000 AND 6000) and
landing__outcome LIKE 'Success (drone ship)'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

- Uses the **OR** function to match both the 'Failures' and 'Success' patterns

```
%%sql
select COUNT(landing__outcome) as SUCCESSES_AND_FAILURES
from SPACEXTBL
where landing__outcome LIKE 'Failure%' OR
      landing__outcome LIKE 'Success%'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

successes_and_failures
71

Boosters Carried Maximum Payload

- Uses a subquery to find the largest payload and match the booster version that have carried that size payload.

```
%%sql
select DISTINCT booster_version
from SPACEXTBL
where payload_mass__kg_ LIKE (select MAX(payload_mass__kg_)
from SPACEXTBL)
```

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

2015 Launch Records

Uses the **AND** function to match both '2015' and 'Failure (drone ship)' patterns

```
%%sql
select landing__outcome, booster_version, launch_site
from SPACEXTBL
where YEAR(DATE) LIKE '2015' and
      landing__outcome LIKE 'Failure (drone ship)'
```

```
 * ibm_db_sa://ynf02037:***@125f9f61-9715-46f9-9399-c8177b21803b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:30426/BLUDB
Done.
```

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

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

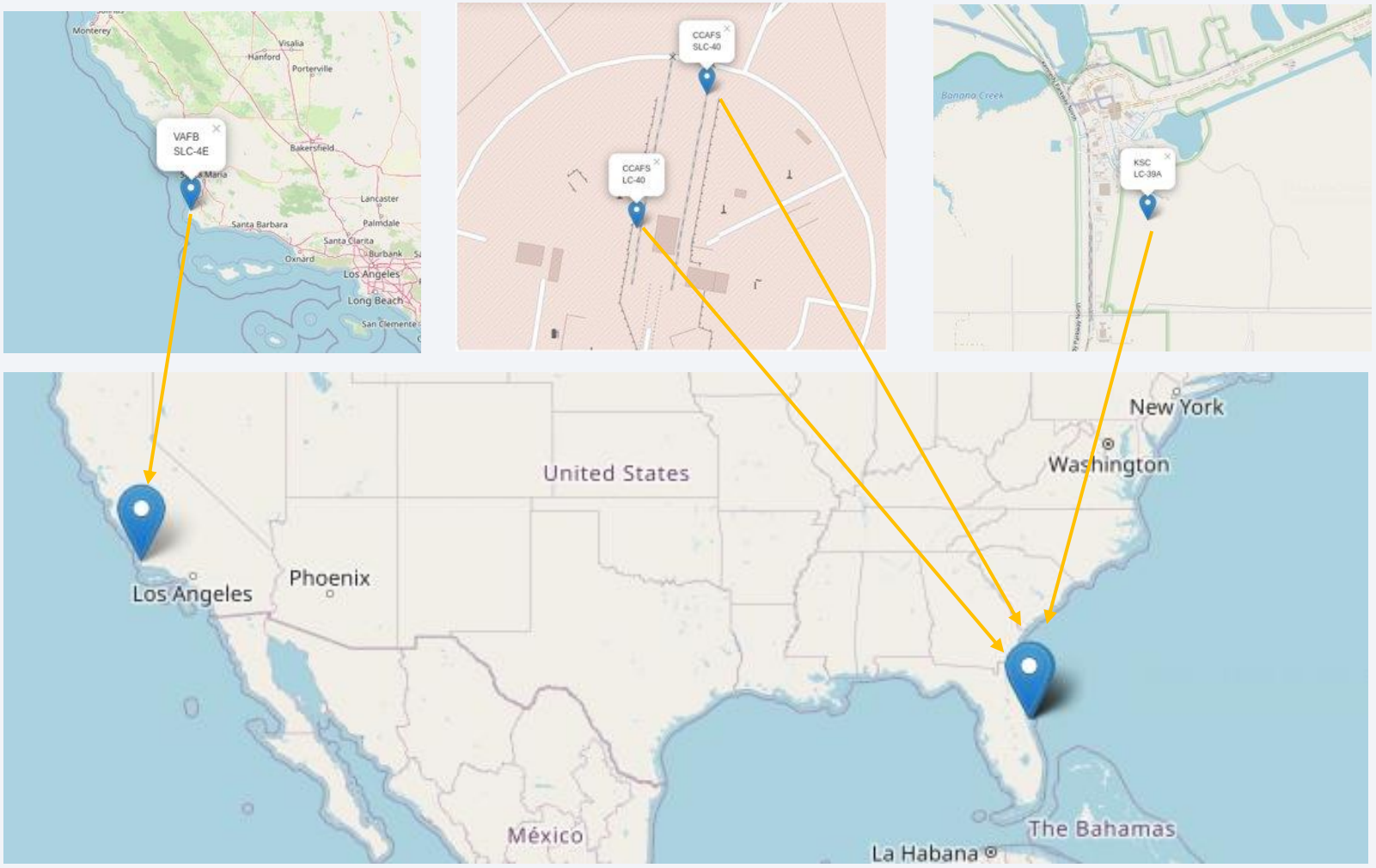
Uses the **Group BY**, **Order BY**, and **Count** functions to collect all count the number of records associated with the landing outcome and sort by descending order.

```
%%sql
select landing__outcome, COUNT(*) as LANDING_OUTCOME_COUNT
from SPACEXTBL
Group BY landing__outcome
ORDER BY LANDING_OUTCOME_COUNT DESC
```

landing__outcome	landing_outcome_count
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

Launch Sites Locations

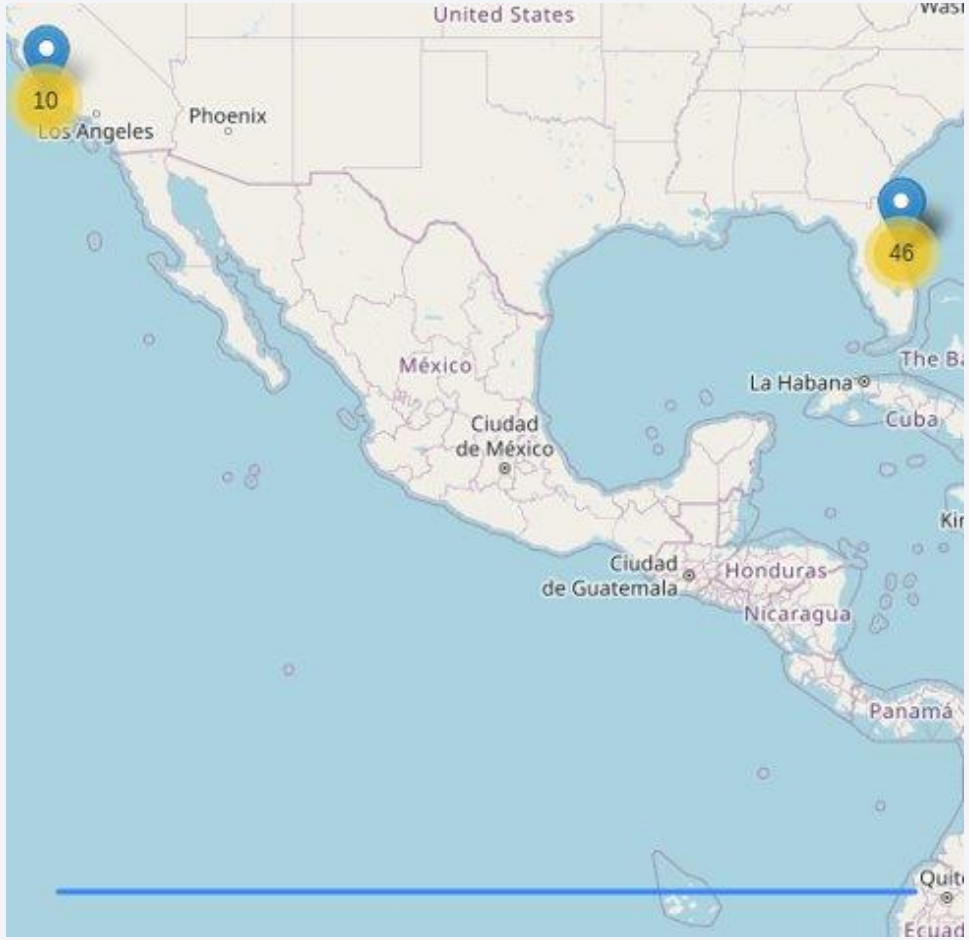


- All four launch sites are Southern region of the USA and close to the coast.

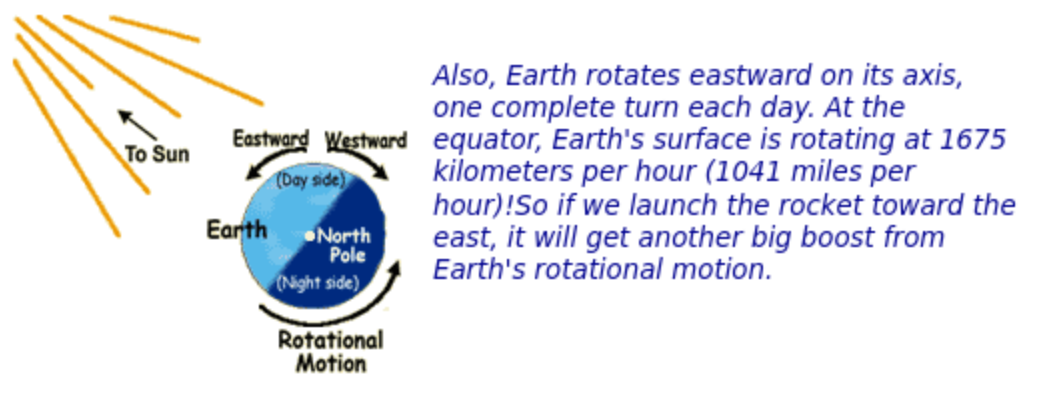
- KSC LC-39A, CCAFS SLC-40, and CCAFS LC-40 very close proximity to each other, while VAFB SLC-4E is on the other site of the country

Number of Launches based on coastline

- The map clearly shows that Florida has a larger number of launches by a large margin.



Launching to the East



Source: spaceplace.nasa.gov

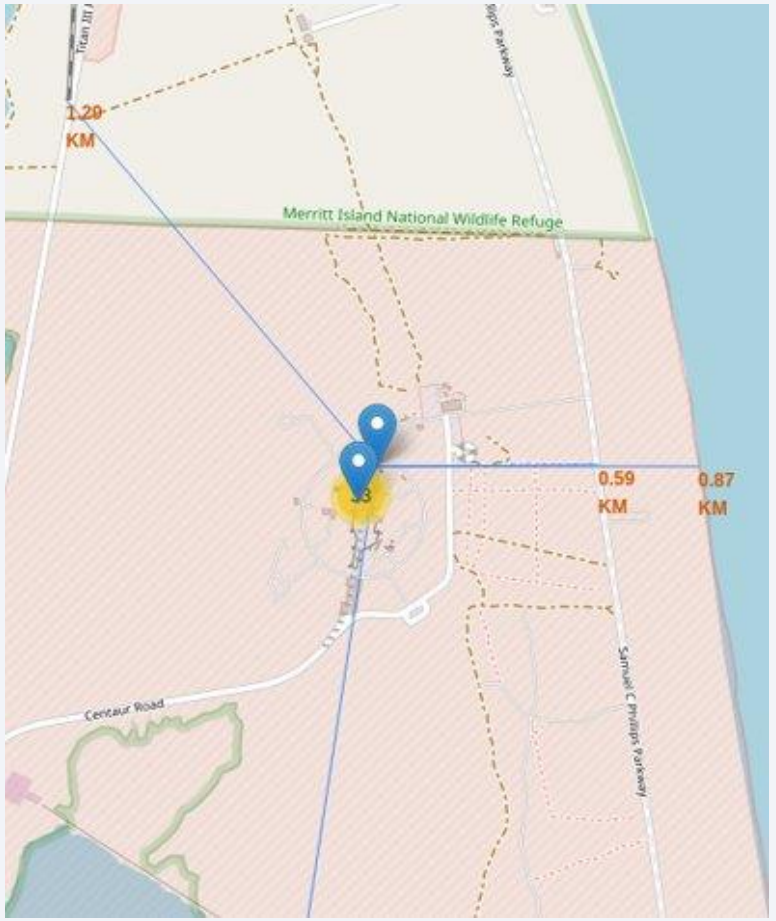
- Launch sites in Florida allow the rockets to go over the ocean rather than land
- Since the sites in Florida are farther south than VAFB SLC-4E they should benefit more from the earth's rotational motion.

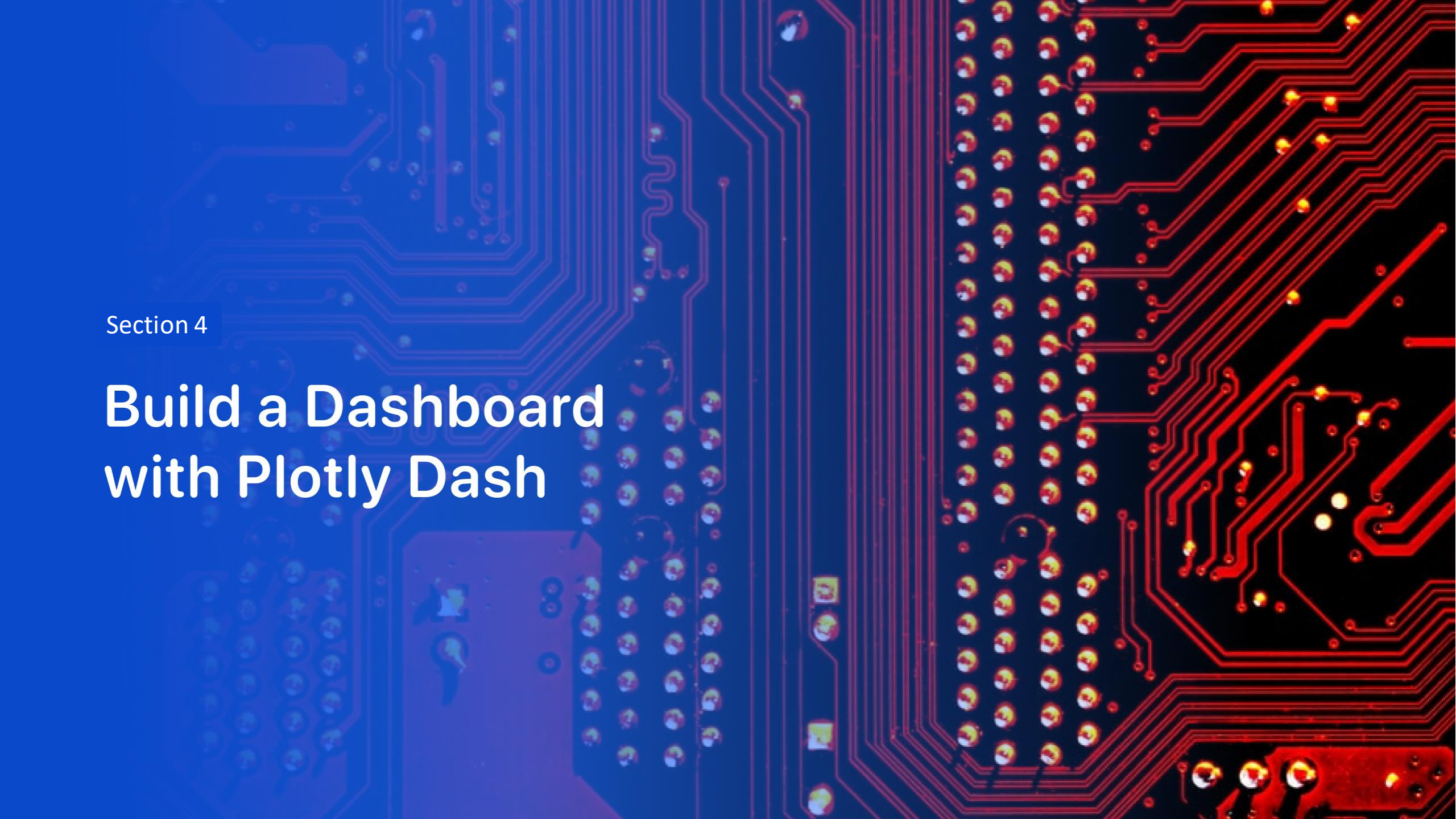
Launch Sites and Land marks

The launch sites where kept farther away from cities then highways and railways.

Left: shows CCAFS SLC-40 launch site is 18.21 KM from Cape Canaveral

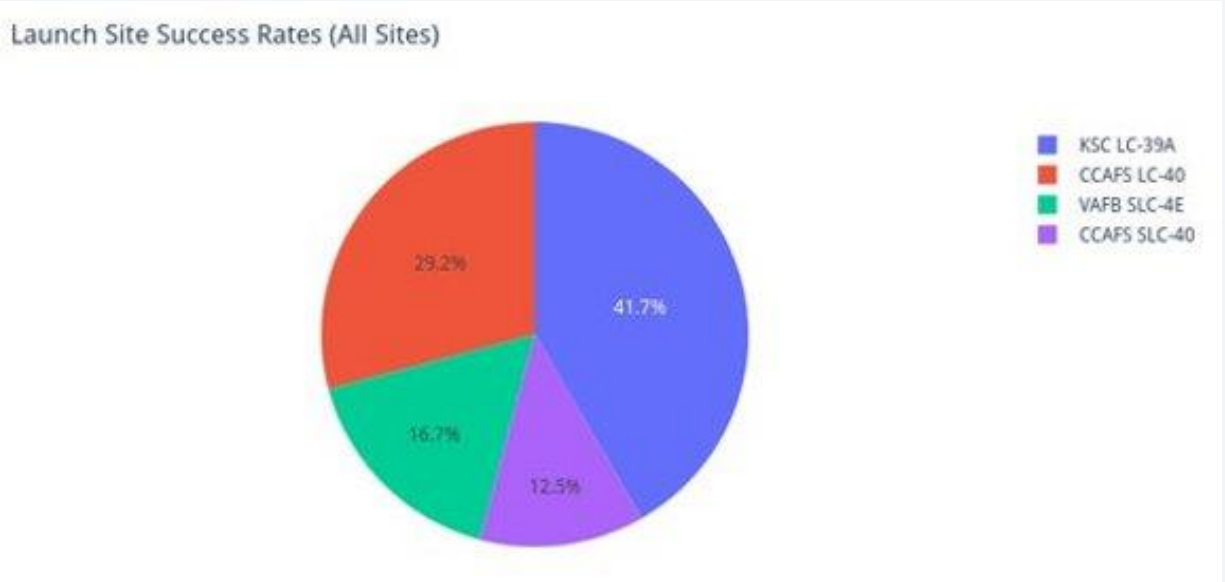
Right: shows CCAFS SLC-40 launch site only 0.59 KM from a highway and 1.29 KM from a railway.




Section 4
Build a Dashboard with Plotly Dash

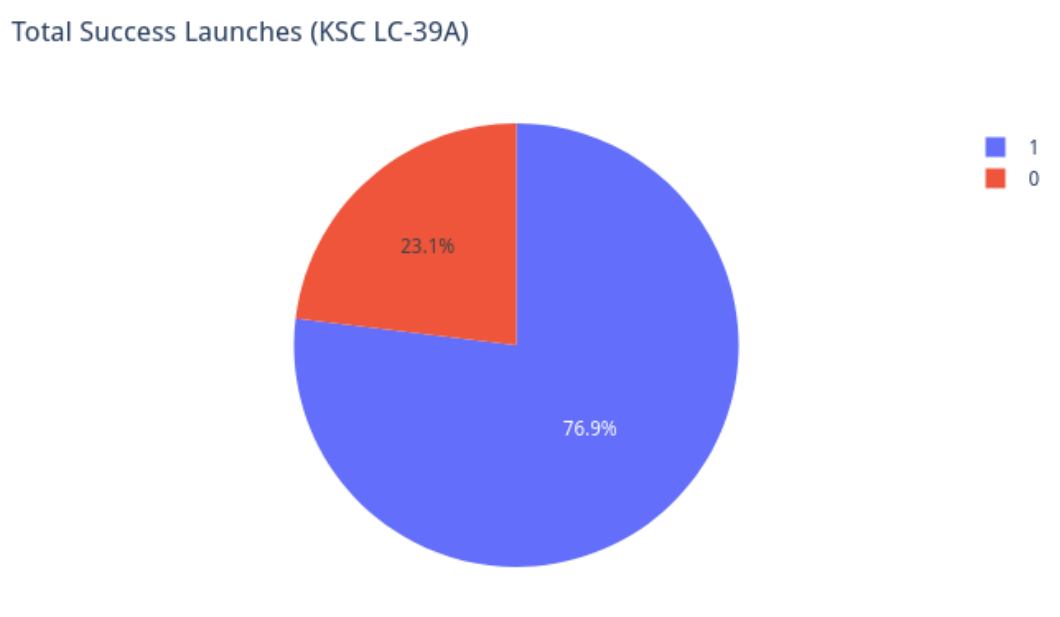
Launch site Success Rates

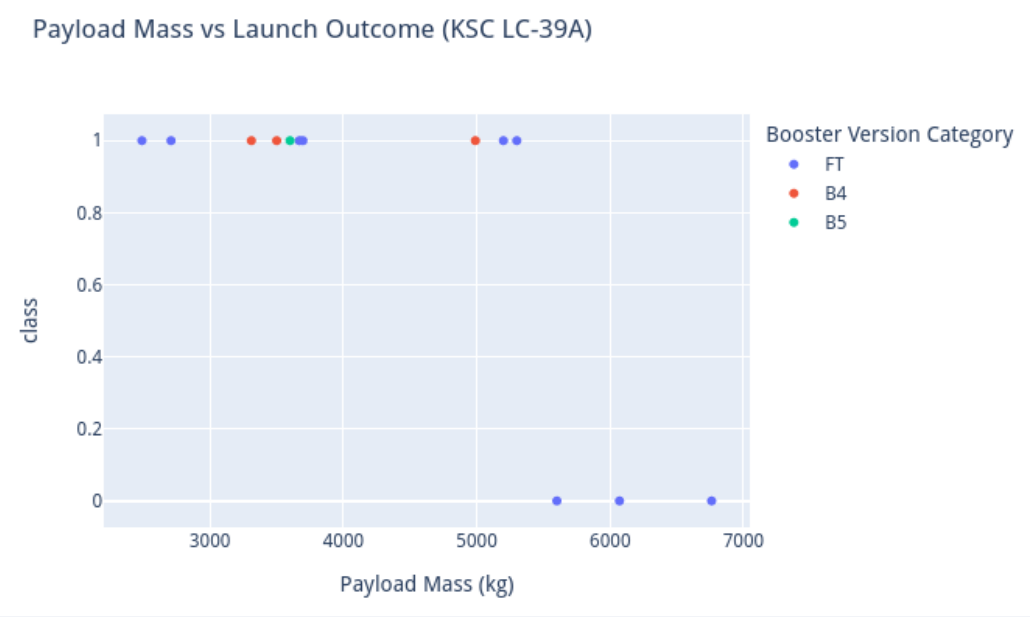
- The KSC LC-39A clearly had the highest success rate out for all the launch sites.
- CCAFS SSL-40 had the lowest success rate.



Highest Launch Success Ratio

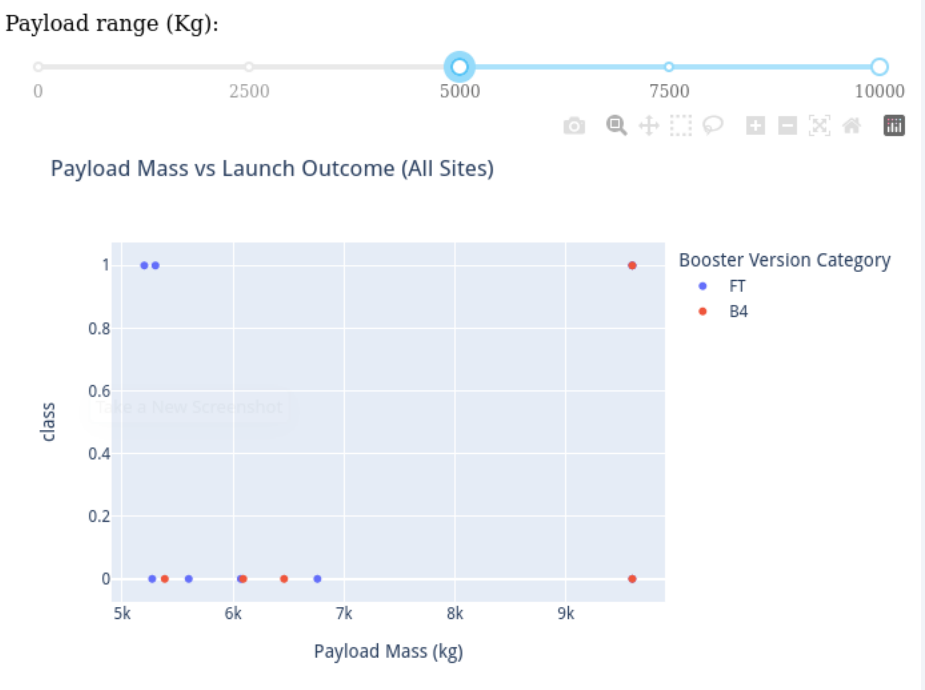
- KSC LC-39A had the highest launch success ratio out of the four launch sites.
- One factor the likely to have contributed to the launch sites success is the fact that most of its payload mass was under 5000kg.





Successful Outcome Trend

After 5000kg the successful outcomes decrease sharply.

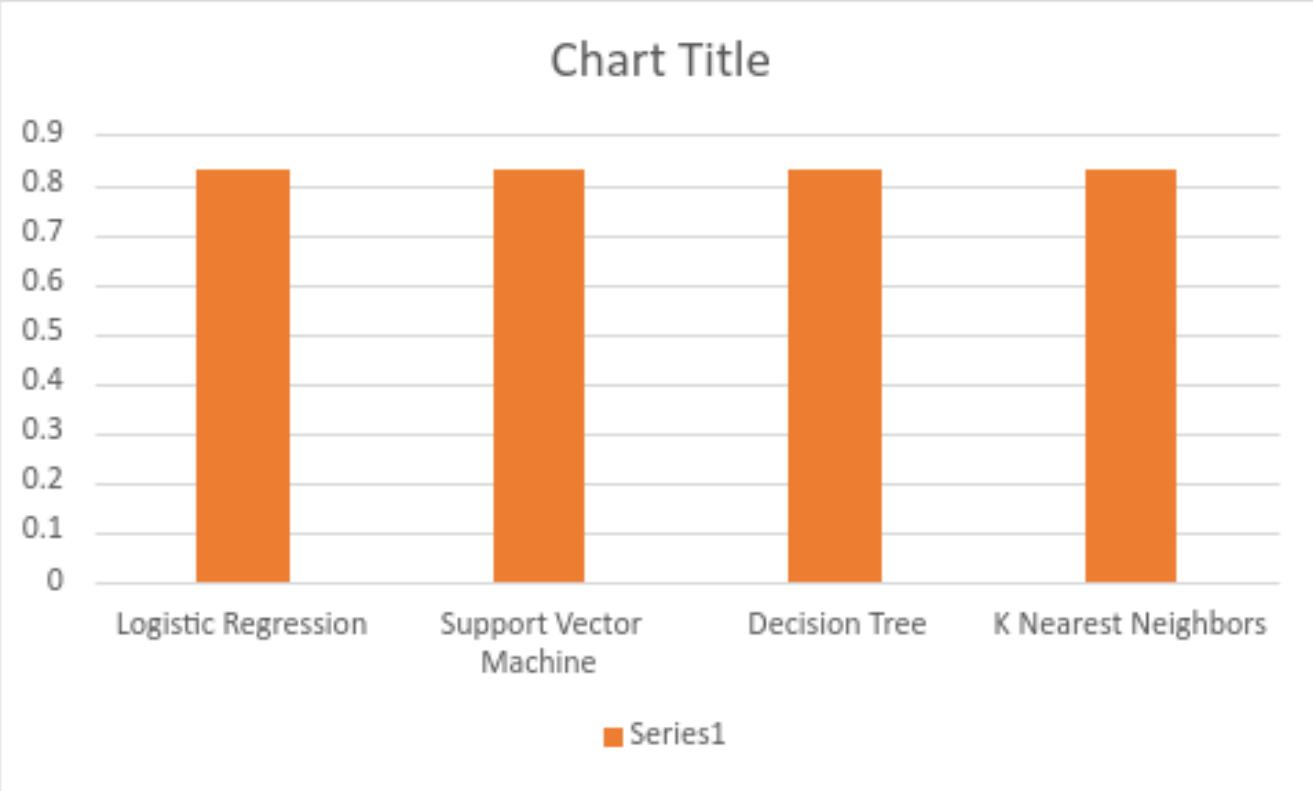


Section 5

Predictive Analysis (Classification)

Classification Accuracy

All four supervised machine learning methods had the same accuracy of 0.833



Confusion Matrix

The Confusion Matrix compares the predicted outcomes with the true values.

Logistic Regression, Support Vector Machine, Decision Tree, and K Nearest Neighbors also produced the same Confusion Matrix.

Even the model does a decent job predicting the outcomes. The upper right hand corner has a problem where the model predicted 'land' 3 times when the true label was 'did not land'. These incorrect labels are known as 'False Positives'.

Conclusions

Based on the finding covered throughout this presentation 83% success rate of our model provides a reasonable level of certainty in predicting the outcome of SpaceX launch.

To summarize the key finding found throughout the various analysis:

- launch sites location tend to close to the equator and along the coastline.
- The success rate decreases with payload size increasing
- B4 booster version was the only one booster that had a successful outcome with a payload over 9000Kg

Therefore, SpaceY should be able to use the models to make a competitive bid against SpaceX.

Appendix

Data Collected from:
- SpaceX API <https://api.spacexdata.com>
- Wikipedia <https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches>

Jupyter Notebooks used for to generate charts can be found here
- https://github.com/rdesfo/IBMAppliedDataScienceCapstone

Powerpoint Template
- Data Science Capstone PowerPoint template

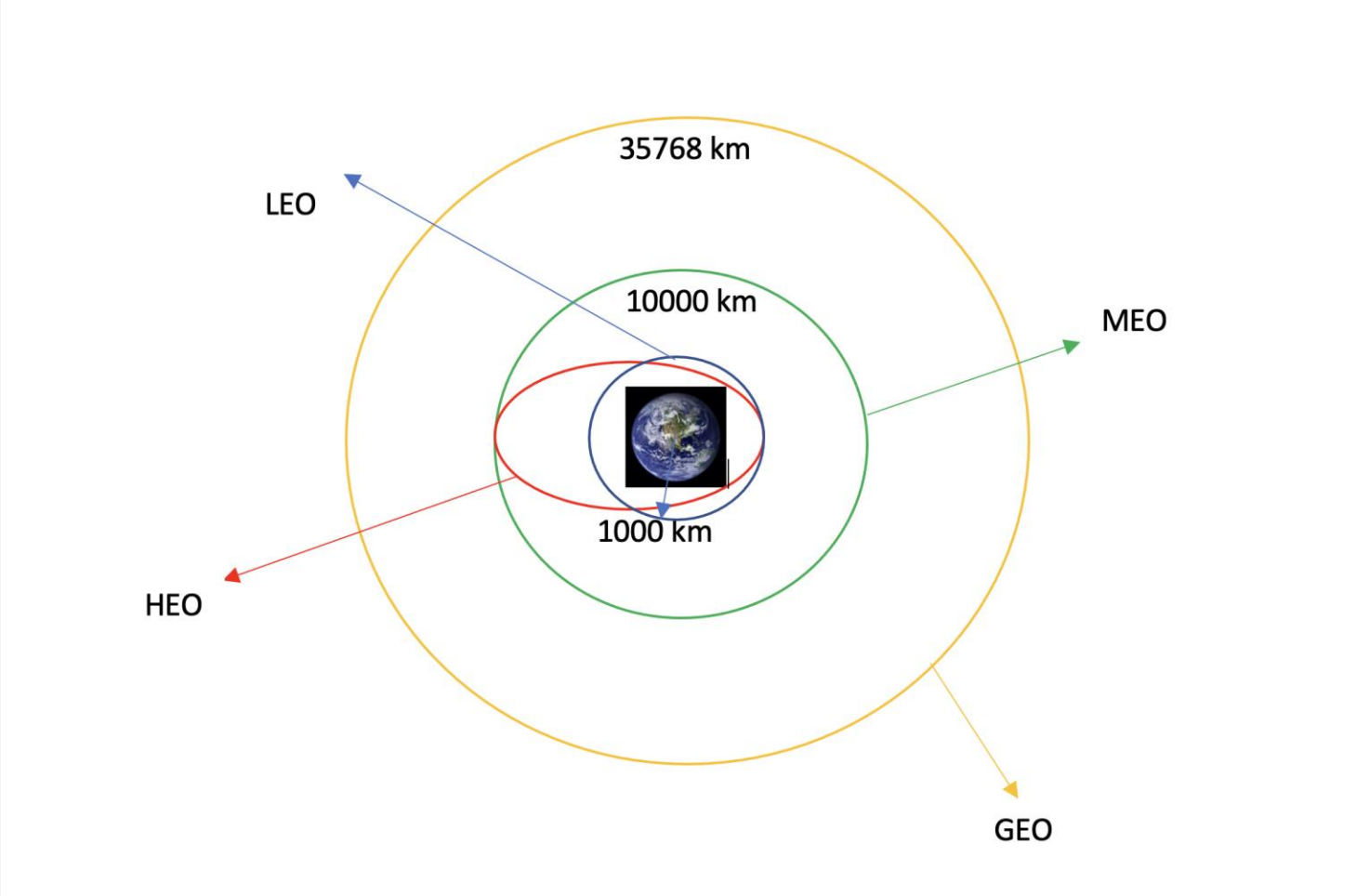
Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Appendix – Booster Version Outcomes

Booster Version Category	Landing Outcome	Landing Outcome Count
B4	Success	6
	Failure	5
B5	Success	1
	Failure	0
FT	Success	16
	Failure	8
V1.0	Success	0
	Failure	5
v1.1	Success	1
	Failure	14

- The FT version has the highest number of successful outcomes/landings.
- The v1.X categories collectively only have one successful landing
- B5 only had one attempt and it was successful

Appendix – Orbit Examples



Thank you!