Ganar la carrera espacial con **ciencia de datos**

Lilén Frisón
25/01/2023

Ir al repositorio

Esquema de contenido

- Resumen ejecutivo
- Introducción
- Metodología
- Resultados
- Conclusión
- Anexos

Resumen ejecutivo

- Para este proyecto asumí el papel de un científico de datos que trabaja para una nueva compañía de cohetes: **Space Y**.
- Mi trabajo fue predecir si mi competencia, **Space X**, reutilizará la **primera etapa**. Primera etapa es el factor que posibilita que los lanzamientos de cohetes sean relativamente económicos para esta empresa. En lugar de usar ciencia espacial para predecir esto, entrené un modelo de aprendizaje automático con información pública de la competencia que recopilé y procesé.
- **Objetivo**: Predecir el aterrizaje de la primera etapa de Space X Falcon 9.
- **Desarrollo**: Como punto de partida recopilé información de mi competencia de varias fuentes, valiendome de una API y de web scraping. Luego procesé estos datos para aumentar su calidad y poder hacer un correcto análisis con el fin de obtener información relevante. Fueron necesarias las visualizaciones y distintas técnicas para comprender cómo se relacionan las variables. Finalmente creé, evalué y refiné modelos predictivos para responder adecuadamente al objetivo del proyecto.
- Esta presentación tiene el fin de mostrar el desarrollo de todo lo anteriormente descrito.

Introducción

En su sitio web, **Space X** anuncia lanzamientos de **cohetes Falcon 9** con un costo de 62 millones de dólares, siendo que otros proveedores cuestan más de 165 millones de dólares cada uno. Gran parte del ahorro se debe a que Space X puede reutilizar la primera etapa. **Por lo tanto, si puedo determinar si la primera etapa aterrizará con éxito, y por ende será reutilizada, puedo determinar el costo de un lanzamiento.** Esta información es útil si una empresa alternativa quiere hacer una oferta contra Space X para el lanzamiento de un cohete, como es el caso de **Space Y**.

Recordatorio: El éxito de un lanzamiento en este contexto significa que este lanzamiento tuvo como resultado un aterrizaje éxitoso.

De modo tal que:

Lanzamiento exitoso = Primera etapa aterrizada con éxito

Sección 1:
Metodología

Metodología

Resumen ejecutivo:

- **Recolección de datos:** Se obtuvieron datos a través de dos vías: la API de Space X y una página de Wikipedia a la cual se le aplicó web scraping.
- **Disputa de datos:** Se hizó un pequeño análisis y se determinaron etiquetas de entrenamiento con los resultados de los aterrizajes.
- **EDA con SQL y visualización de datos:** Se sometieron los datos a diversas consultas SQL y visualizaciones con el fin de aumentar la comprensión de los mismos.
- **Anlálisis visual interactivo con Folium y Plotly Dash:** Se analizaron los sitios de lanzamiento y sus proximidades a través de visualizaciones con mapas, y se creó un tablero para interactuar con los datos a través de una interfáz con parámetros y gráficos.
- **Análisis predictivo con modelos de clasificación:** Se ajustaron cuatro modelos de clasificación con GridSearchCV y se los comparó para encontrar el de mayor efectividad.

Recolección de datos

El proceso de recopilación de datos involucró:

- Solicitudes a la API de Space X, dando como resultado un dataset.
- Web scraping a una página de Wikipedia de Space X, también dando como resultado un dataset final.

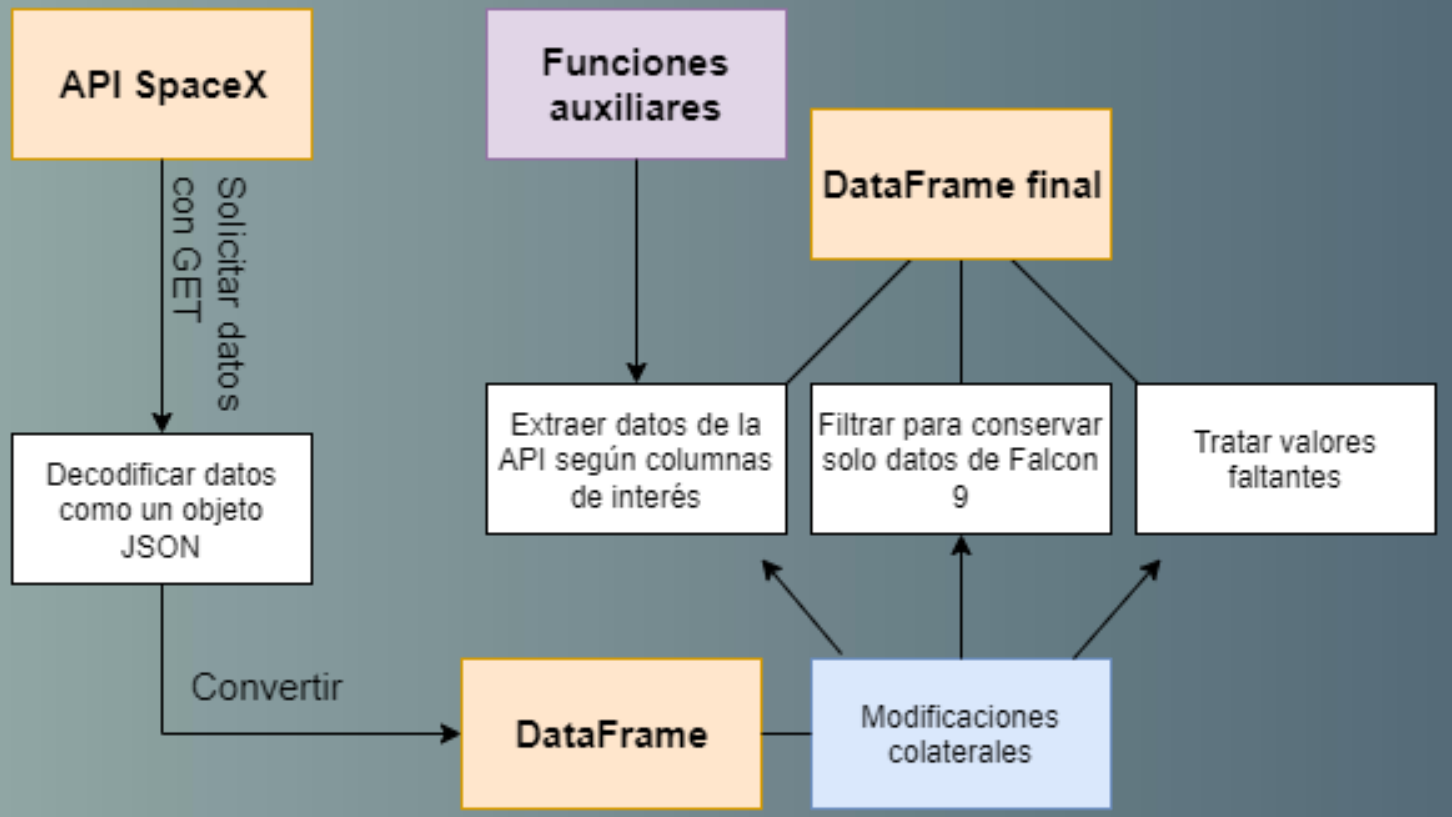
A continuación, se mostrarán diapositivas con el flujo de la recopilación de datos para ambas técnicas.

Recolección de datos – API de SpaceX

- Se utilizaron IDs para extraer información de la API a través de funciones auxiliares.

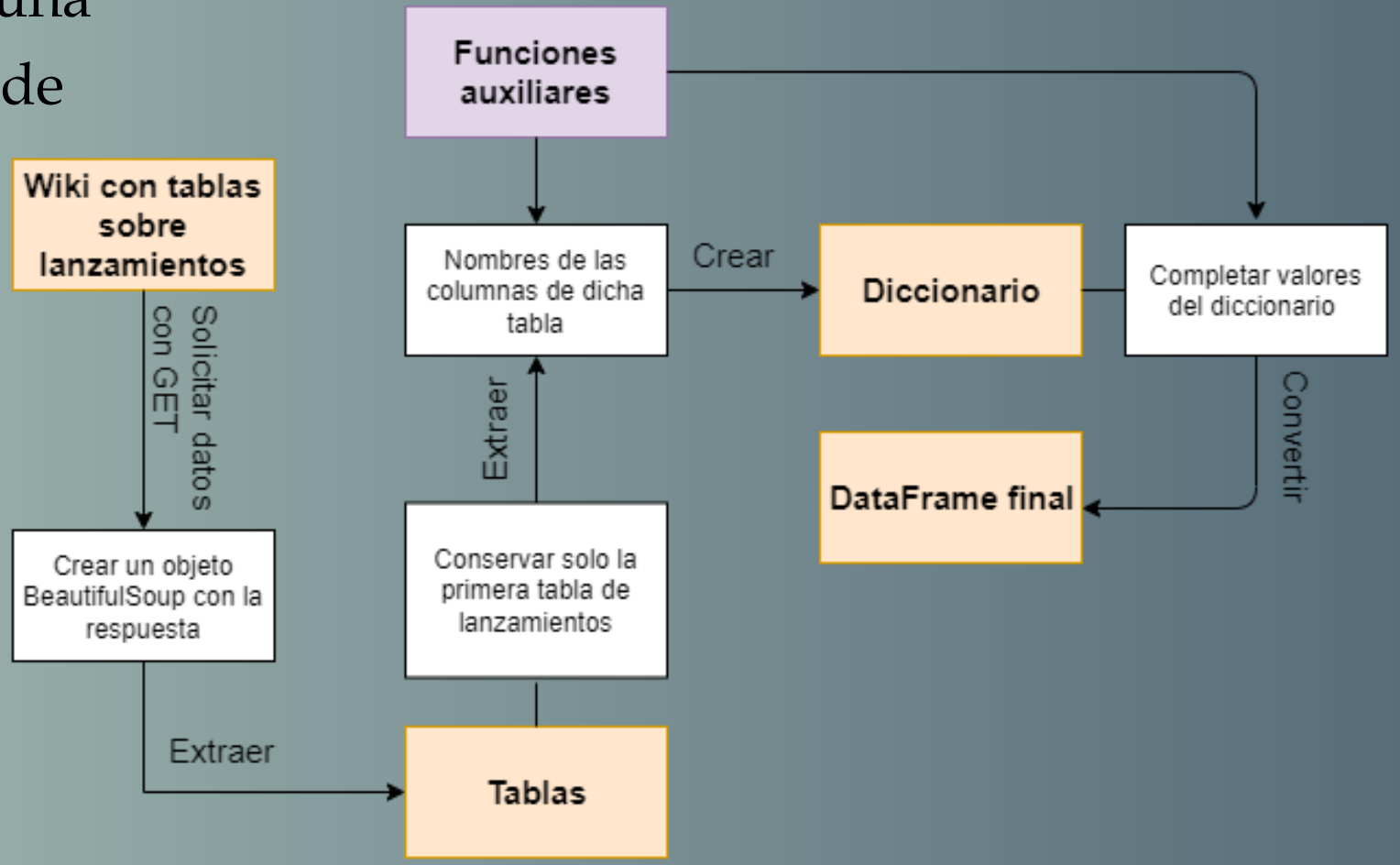
Luego se aplicaron modificaciones colaterales al dataset final.

- Ir al cuaderno



Recolección de datos – web scraping

- Se extrajo de Wikipedia una tabla HTML con registros de lanzamientos de cohetes Falcon 9.
- Se parseó la tabla y convirtió en un DataFrame de Pandas.
- [Ir al cuaderno]



Disputa de datos

- Tuvo lugar un análisis de datos donde se identificaron los valores faltantes a través de porcentajes, y también qué variables categóricas y numéricas conforman el dataset. Además, se calculó lo siguiente: el numéro de lanzamientos por zona, las orbitas a las que apuntan los lanzamientos junto con su número de ocurrencia, y también los **tipos de aterrizajes** con su número de ocurrencia.
- Los resultados de la analítica sirvieron para crear una función capaz de devolver una lista los resultados de los aterrizajes, **donde 0 representa un aterrizaje fallido y 1 representa que el aterrizaje fue exitoso.** Agregué esta lista al dataset como una nueva columna llamada **Class,** generando asi las **etiquetas de entrenamiento** necesarias para el modelo predictivo de este proyecto.
- Ir al cuaderno

EDA con SQL

- El primer paso fue Importar las librerías necesarias y crear una conexión con la **instancia db2** que contiene el dataset sobre el cual se ejecutarán **sentencias SQL**.
- Una vez este cuaderno jupyter tuvo acceso a la base de datos, se la sometió a diversas consultas SQL **con el fin de obtener cierta información y comprender más a detalle los datos**.
- Se obtuvo información sobre los aterrizajes fallidos y exitosos, los lugares e despegue, se filtraron datos por fecha, orden y cadenas de texto específicas, entre otras cosas.
- **Ir al cuaderno**

EDA con visualización de datos

- Tuvo lugar una **análisis y visualización de datos para explorar las relaciones entre las variables del dataset en relación a los resultados de lanzamiento**, que se representan con un 0 si han fracasado y con un 1 si han tenido éxito.
- Del ejercicio anterior se obtuvieron nociones sobre las variables que mayor influencia tienen sobre la tasa de éxito de un lanzamiento. Se creó entonces un DataFrame de **características** con estas variables. A las columnas multicategóricas de este dataset se les aplicó la codificación One-Hot. El dataset final quedó compuesto solo por columnas numéricas de tipo float69.
- **Ir al cuaderno**

Análisis visual interactivo con Folium

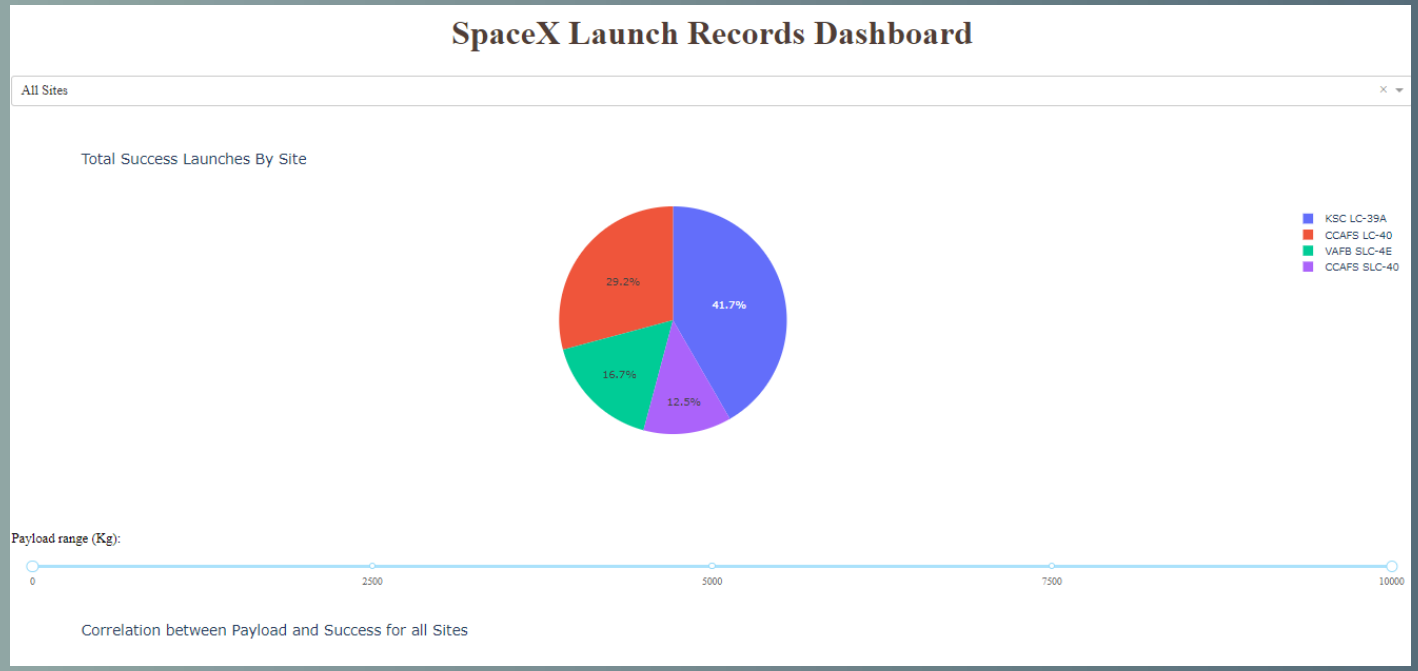
- Se utilizó un **dataset aumentado con latitud y longitud** agregadas para cada sitio de lanzamiento.
- Estas **coordenadas** sirvieron para marcar todos los sitios de lanzamiento en un mapa. También se visualizaron los lanzamientos éxitosos y fallidos para cada sitio, y se calcularon las distancias entre sus proximidades (ciudades, carreteras, ferrocarriles, costas).
- El objetivo de este análisis visual fue **descubrir patrones geográficos sobre los sitios de lanzamiento**.
- Ir al cuaderno

Tablero con Plotly Dash

- Creé una **aplicación**

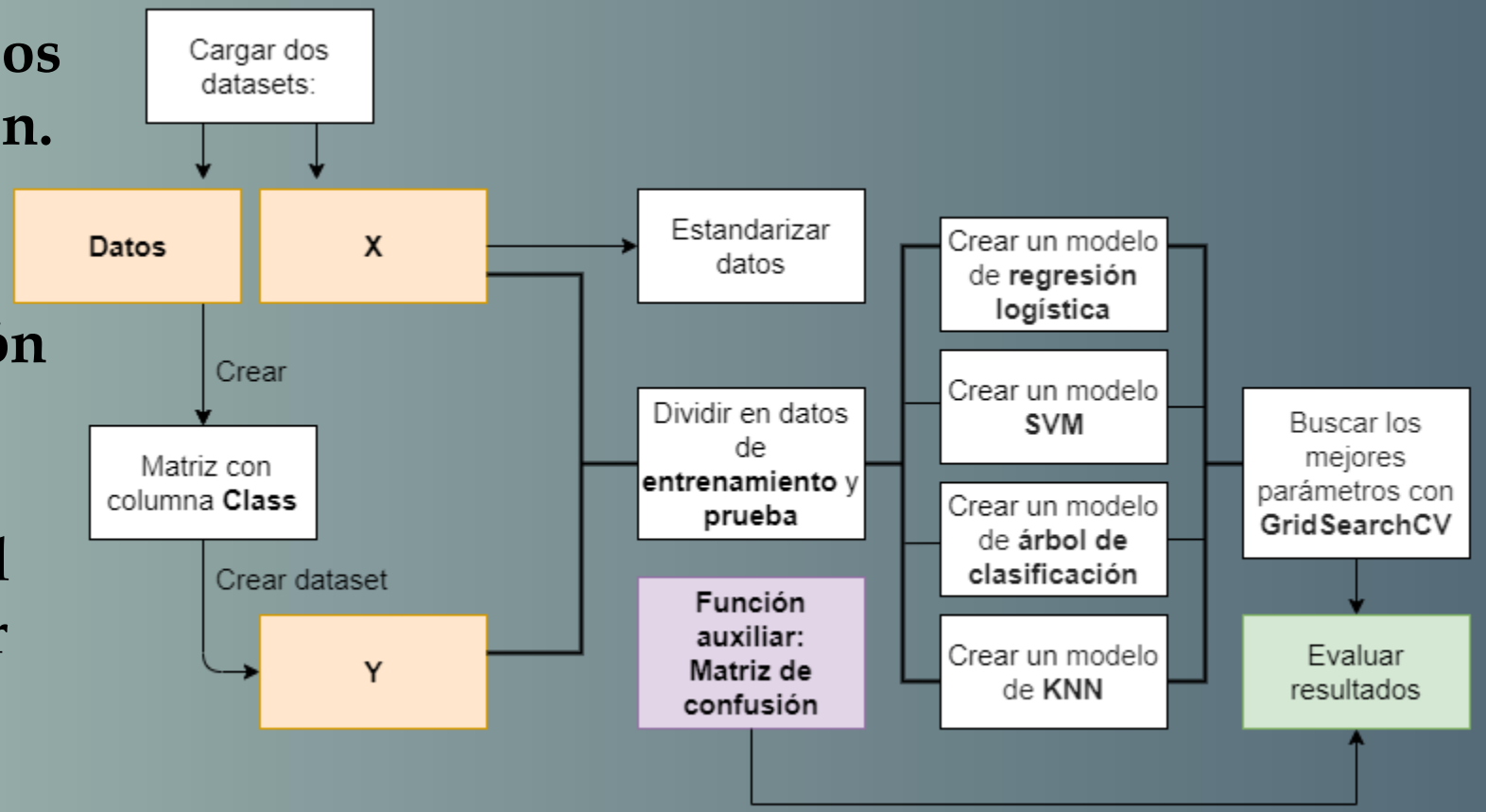
Plotly Dash para que los usuarios puedan realizar análisis visuales interactivos con los datos de los lanzamientos de Space X **en tiempo real**.

- Ir al cuaderno
- Ir al código de la app

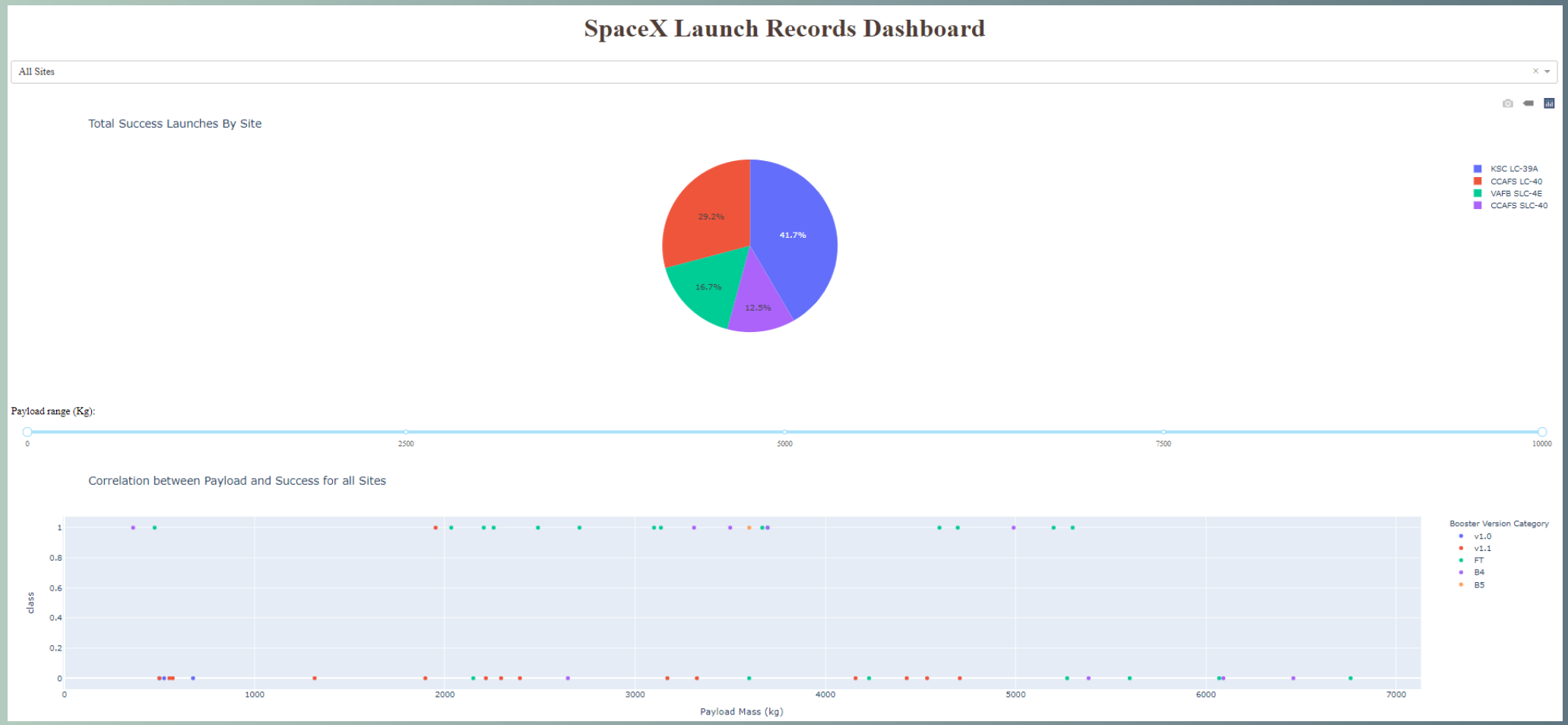


Análisis predictivo (clasificación)

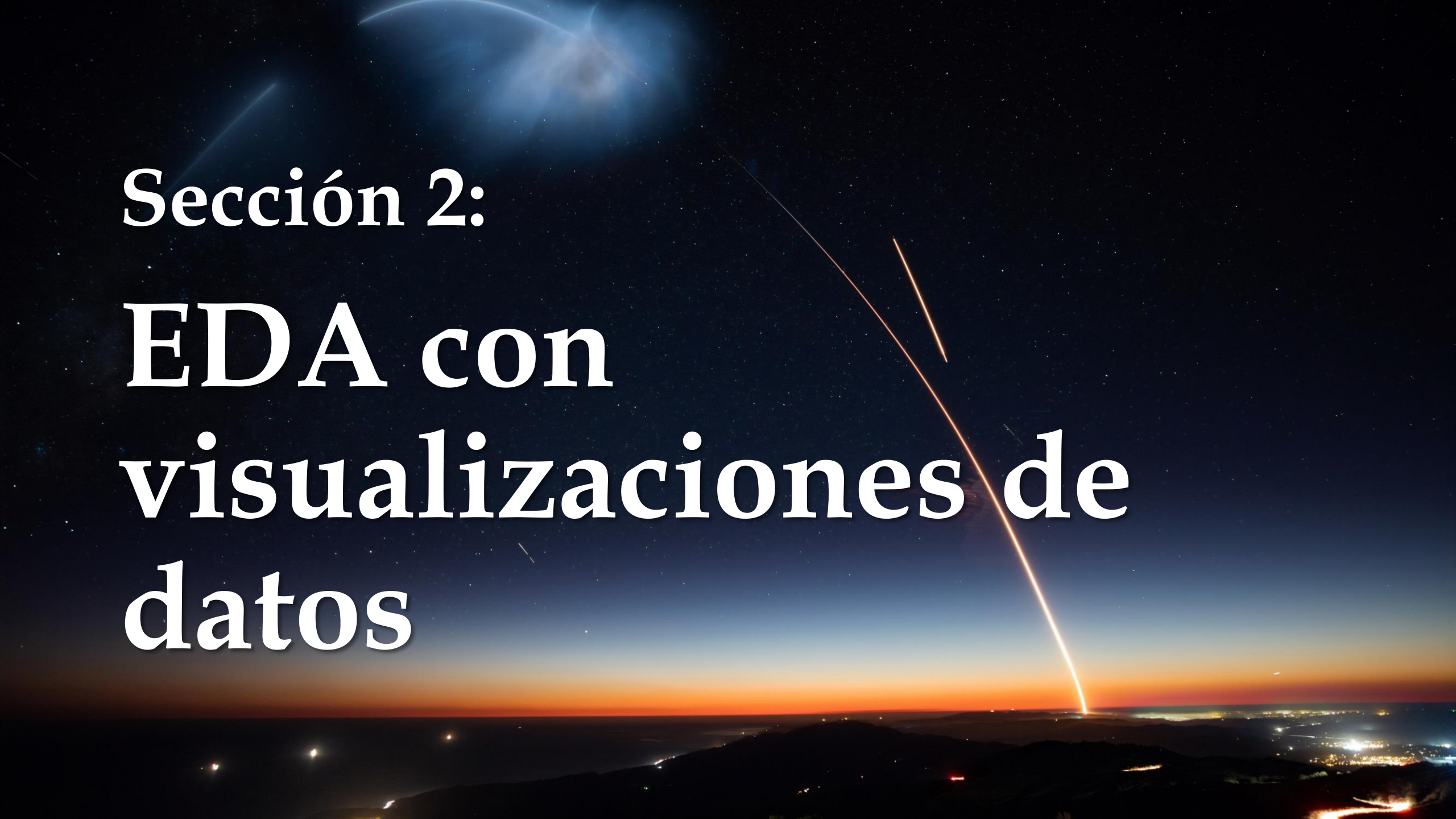
- Se buscaron los mejores hiperparámetros para **4 tipos de modelos de clasificación.** Todo esto usando **GridSearchCV**. Luego se evaluó y calculó la **precisión** de estos modelos sobre los datos de prueba.
- Finalmente se evaluó cuál es el **mejor modelo** a partir de la comparación de sus resultados.
- [Ir al cuaderno]



Resultados



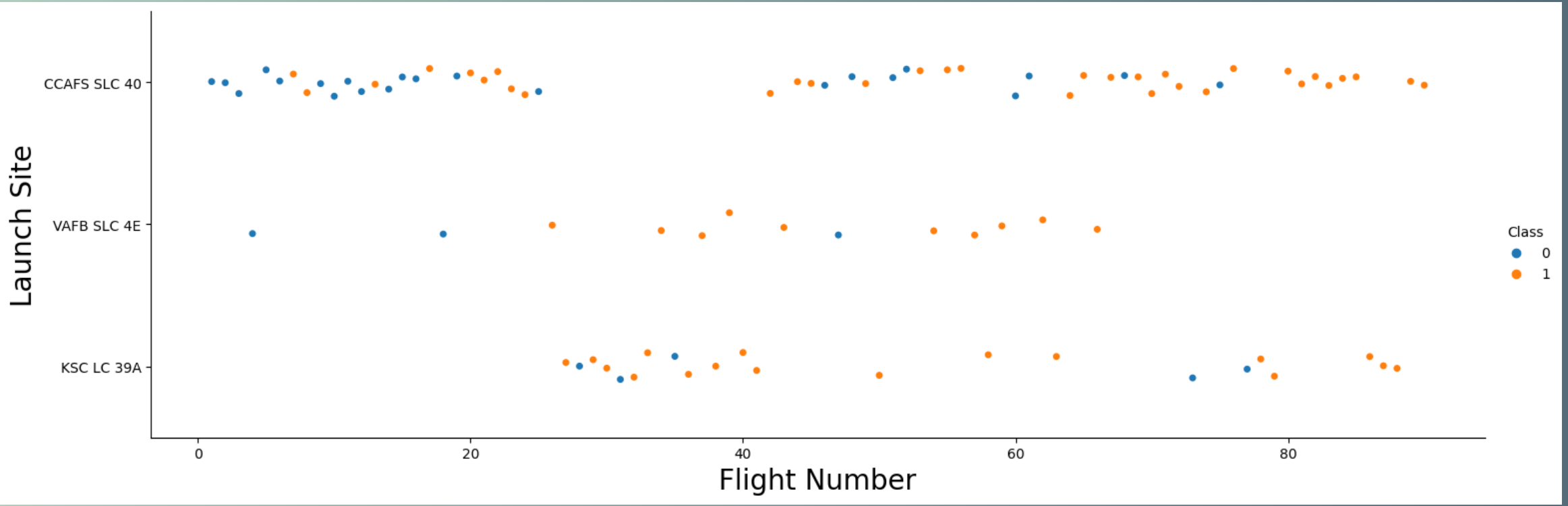
La app de Plotly.
A continuación se proundizará en los resultados de las otras secciones.



Sección 2:

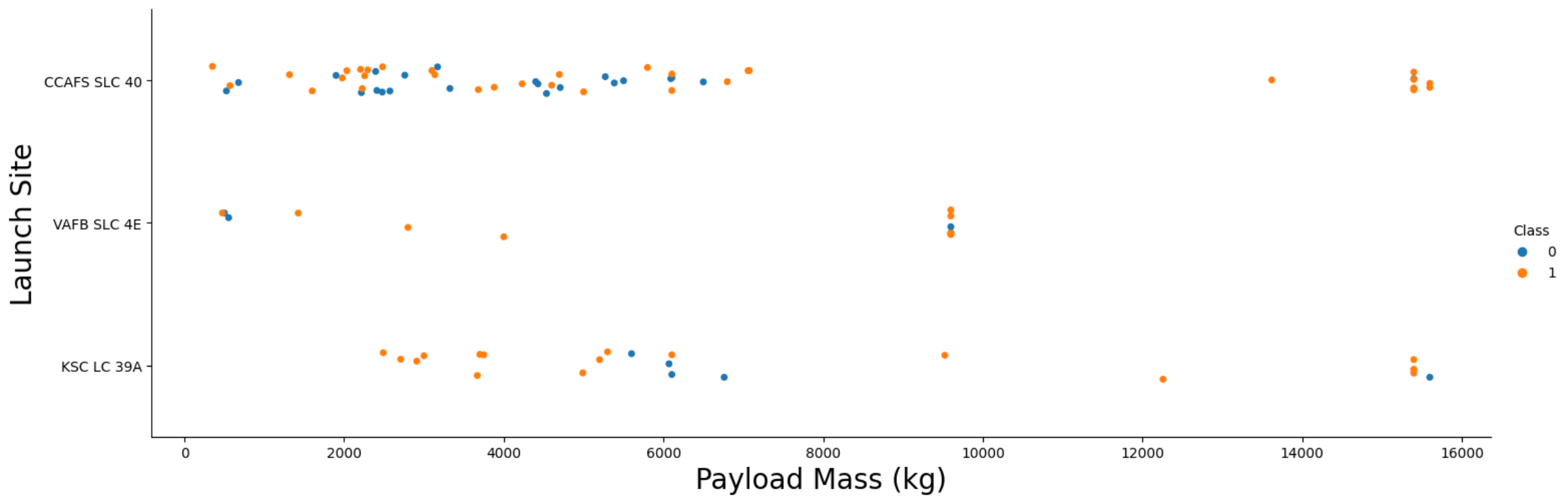
EDA con visualizaciones de datos

Número de vuelos vs. sitio de lanzamiento (Flight Number vs. Launch Site)



Naranja: Lanzamientos éxitosos
Azul: Lanzamientos fallidos
Interpretación: El gráfico sugiere que hay un aumento en la tasa de éxito a lo largo del tiempo, y que el sitio de lanzamiento CCAFS SLC 40 contiene la mayor cantidad de vuelos. También vale la pena considerar el vuelo nro. 20, ya que parece ser que a partir de este es cuando la tasa de éxito comienza a aumentar para todos los sitios de lanzamiento.

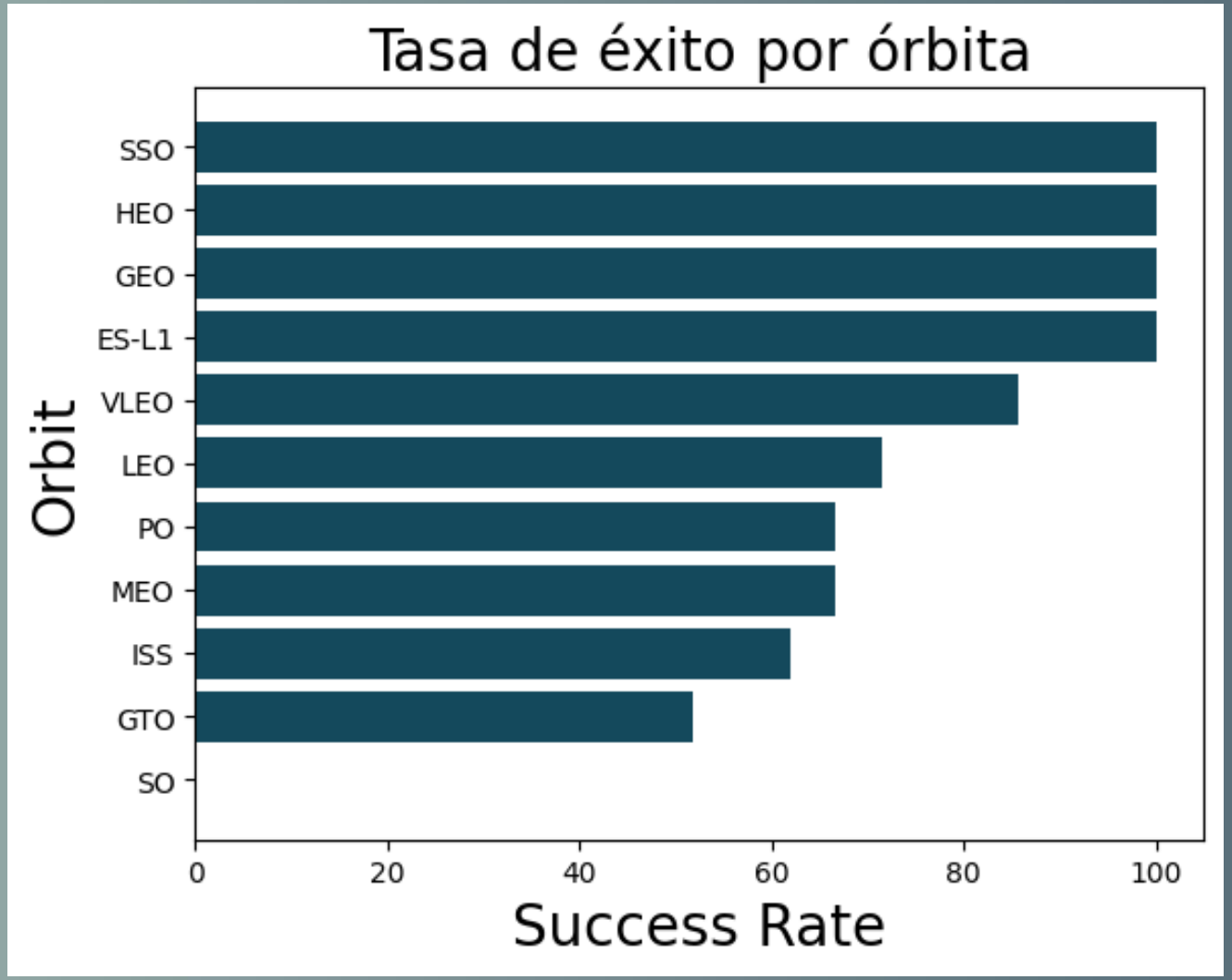
Carga útil vs. sitio de lanzamiento (Payload vs. Launch Site)



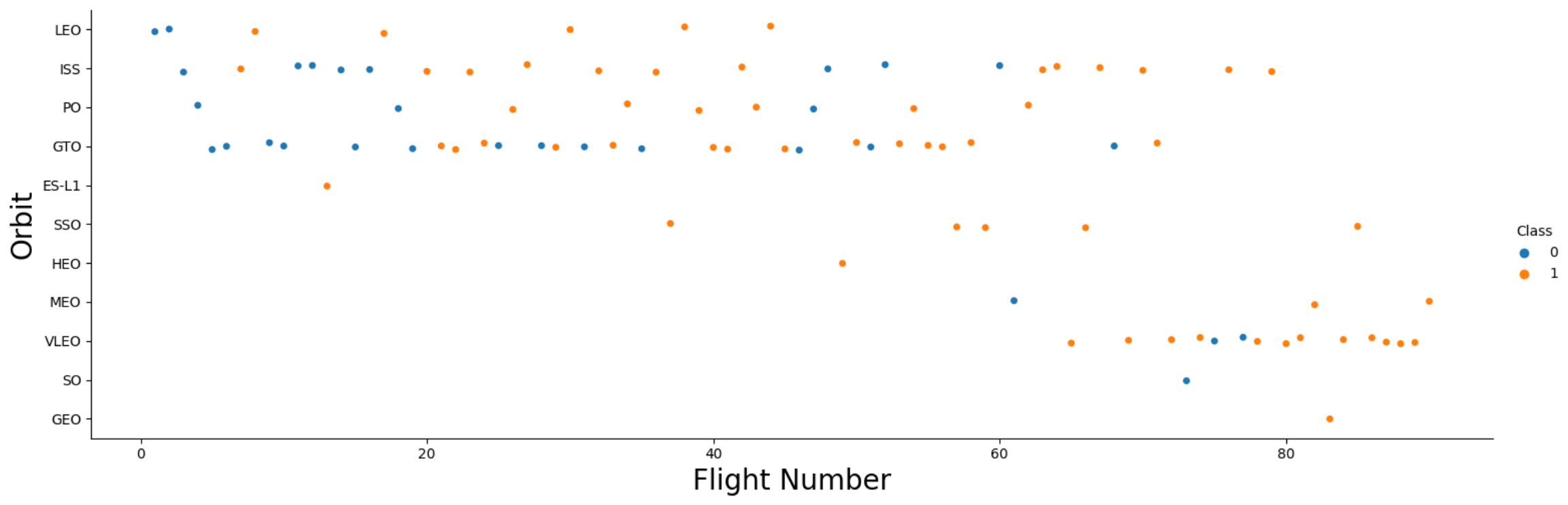
Interpretación: El gráfico sugiere que la masa de carga útil suele estar entre 0-6000 kg., y que cada sitio de lanzamiento suele manejar distintos rangos de kg. de carga. Además, para el sitio VAFB SLC 4E no se lanzan cohetes con una masa de carga pesada (superior a 10000).

Tasa de éxito vs. tipo de órbita (Success Rate vs. Orbit Type)

Interpretación: El gráfico sugiere que SSO, HEO, GEO Y ES-L1 tienen una tasa de éxito del 100%, mientras que el resto de órbitas se encuentran en un rango de éxito entre el 40 y el 80% aprox. Sin embargo, hay una excepción: la órbita SO, que tiene una tasa de éxito del 0%, es decir, nula.

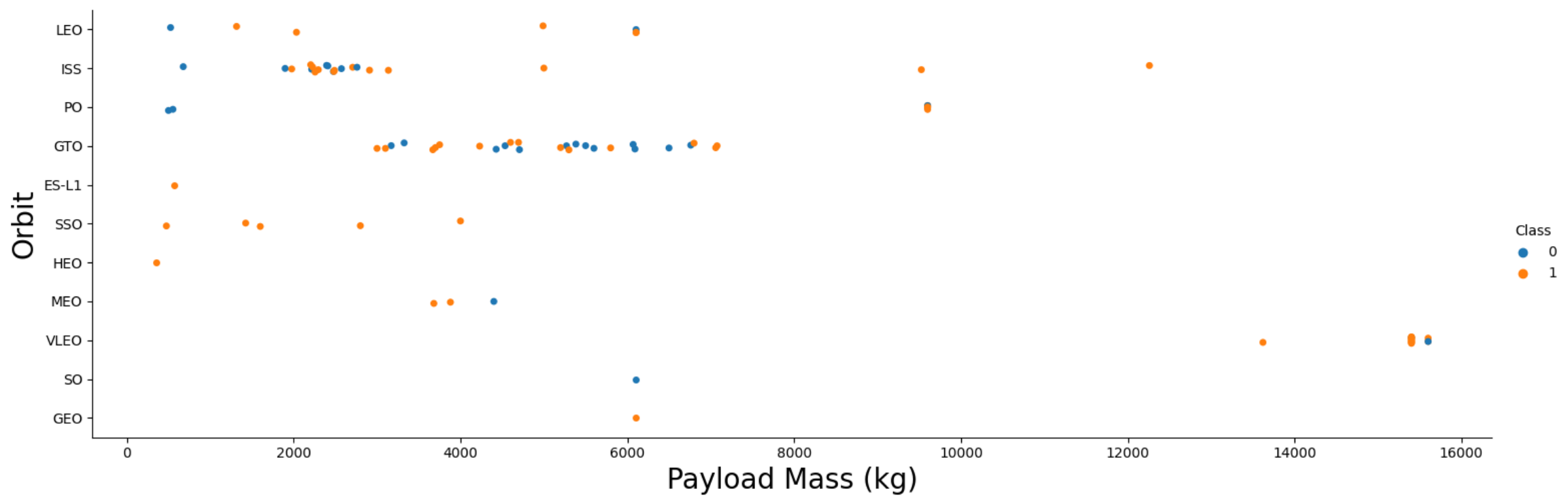


Número de vuelos vs. tipo de órbita (Flight Number vs. Orbit Type)



Interpretación: El gráfico sugiere que en la órbita LEO es donde más se dan lanzamientos éxitosos en relación al número de vuelos.

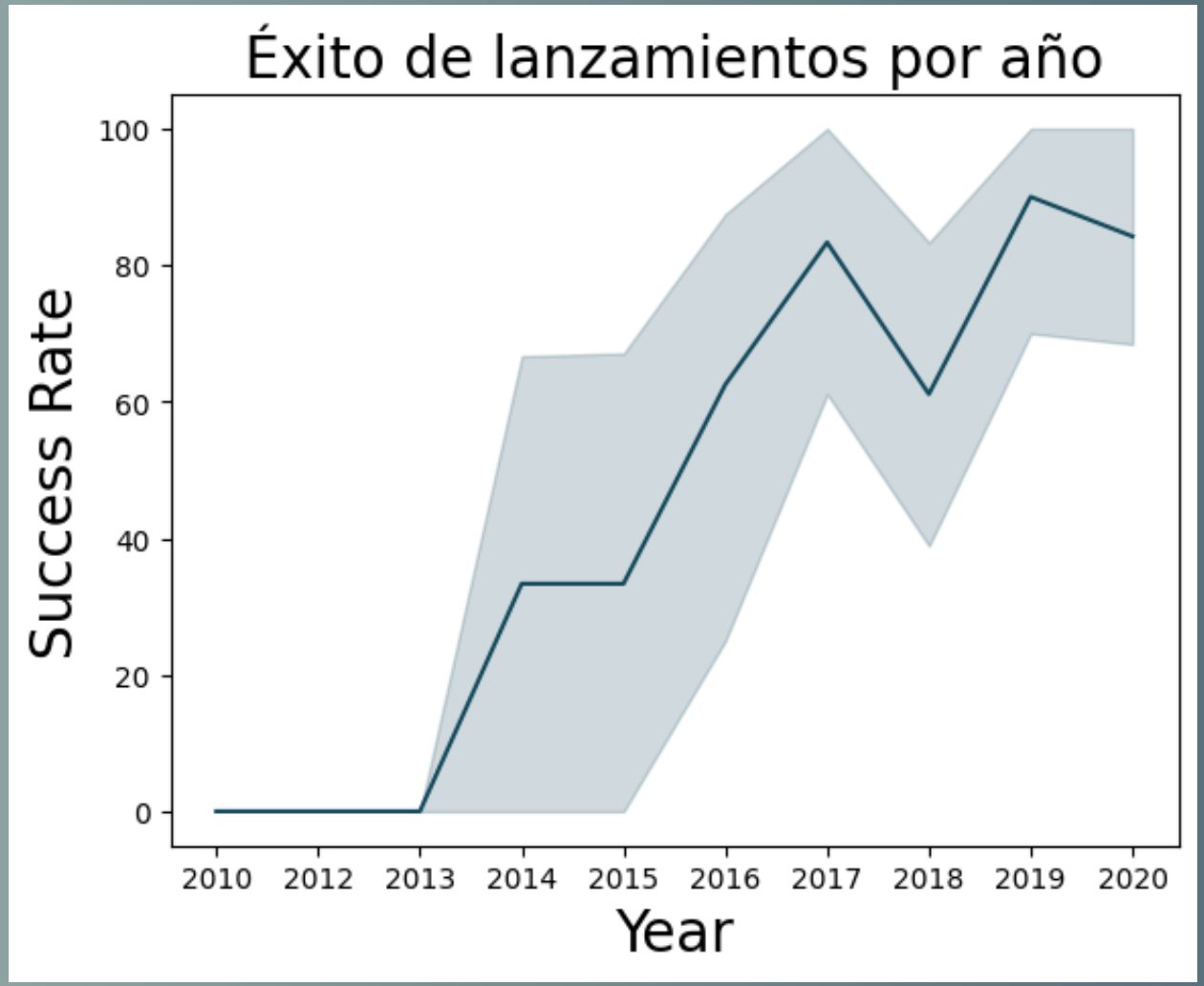
Carga útil vs. tipo de órbita (Payload vs. Orbit Type)



Interpretación: El gráfico sugiere que, con cargas pesadas, los lanzamientos éxitosos influencian más a PO, LEO e ISS. Para GTO no se distingue bien, ya que tanto la tasa de aterrizaje positiva como la negativa estan presentes. Sin embargo, La órbita más exitosa en relación a la carga es VLEO, que tiene valores positivos hasta pasando los 14000 kg.

Tasa de lanzamientos éxitosos por año (Launch Success Yearly Trend)

Interpretación: El gráfico sugiere que la tasa de éxito comenzó a aumentar a partir del año 2013, y que de ahi en adelante el éxito de los lanzamientos cubre más o menos un 80% en un periodo de 10 años.




EDA con SQL

Todos los nombres de los sitios de lanzamiento

Esta consulta SQL devuelve una tabla con los nombres de los sitios de lanzamiento únicos.

```
%%sql
SELECT DISTINCT launch_site
FROM SPACEXDATABASE;
```

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-
Done.
```

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

Nombres de los sitios de lanzamiento comenzados con 'CCA'

```
%%sql
SELECT *
FROM SPACEXDATABASE
WHERE launch_site like 'CCA%'
LIMIT 5;
```

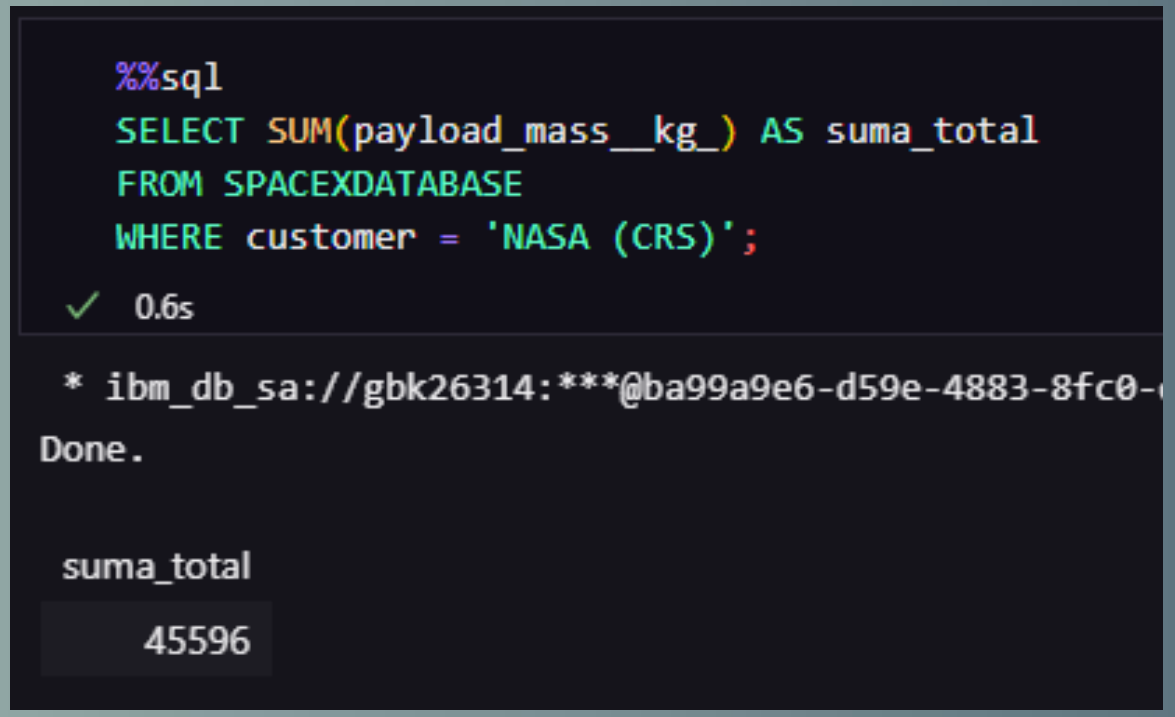
✓ 1.2s Python

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/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 de carga útil lanzada por la NASA

Esta consulta SQL devuelve la suma total de la masa de carga útil transportada por propulsores que han sido lanzados por la NASA (CRS).



Promedio de la carga útil transportada por F9 v1.1

Esta consulta SQL devuelve el promedio de la masa de carga útil transportada por el propulsor F9 v1.1

```
%%sql
SELECT AVG(payload_mass__kg_) AS Promedio
FROM SPACEXDATABASE
WHERE booster_version = 'F9 v1.1';
```

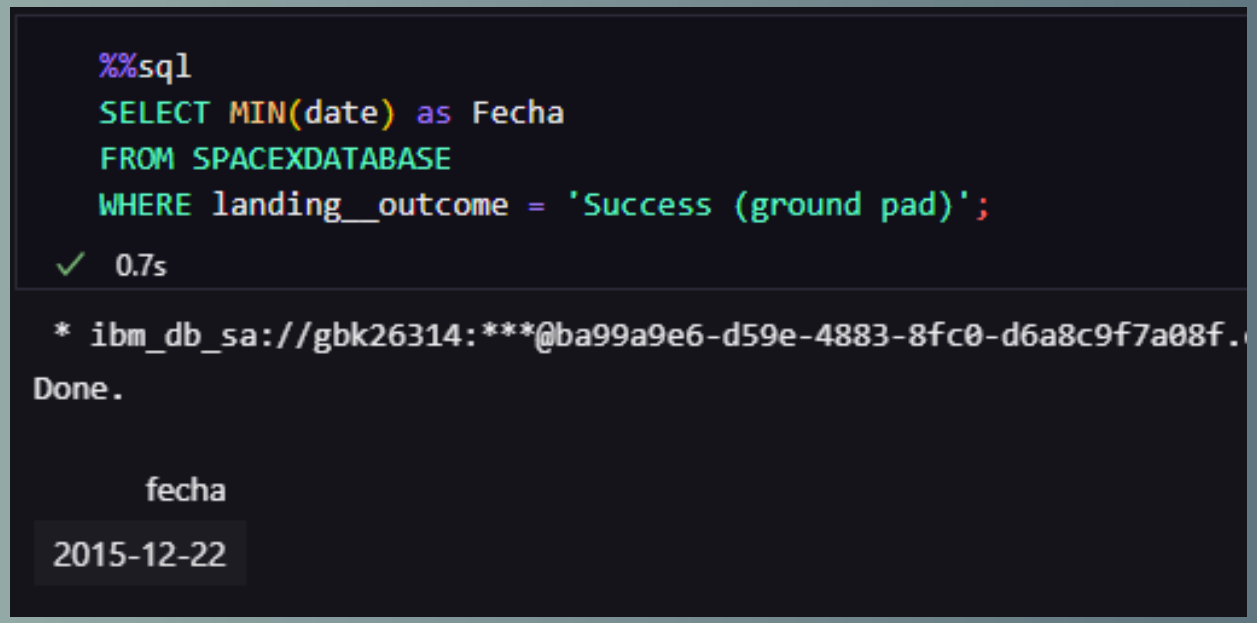
✓ 0.6s

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-
Done.
```

promedio
2928

Primera fecha de aterrizaje terrestre exitoso

Esta consulta SQL devuelve la fecha en que se logró el primer aterrizaje exitoso en la tierra.



Aterrizajes exitosos de naves no tripuladas con carga útil entre 4000 y 6000

Esta consulta SQL devuelve una tabla con los propulsores que tienen éxito en naves no tripuladas, y que tienen una masa de carga útil superior a 4000 kg. pero inferior a 6000 kg.

```
%%sql
SELECT booster_version
FROM SPACEXDATABASE
WHERE (mission_outcome like 'Success')
AND (landing__outcome = 'Success (drone ship)')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000);
```

✓ 0.6s

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-
Done.
```

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

Número total de resultados exitosos y fallidos de la misión

Esta consulta SQL devuelve los resultados exitosos y fallidos de las misiones con su número de ocurrencia.

```
%%sql
SELECT mission_outcome, COUNT(*) AS Total
FROM SPACEXDATABASE
GROUP BY mission_outcome;
```

✓ 0.5s

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-
Done.
```

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

Propulsores con la máxima carga útil

Esta consulta SQL devuelve una tabla con los propulsores que han transportado la masa máxima de carga útil.

```
%%sql
SELECT booster_version
FROM SPACEXDATABASE
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEXDATABASE);
```

✓ 0.6s

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.
Done.
```

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

Records de lanzamientos 2015

```
%%sql
SELECT landing__outcome, booster_version, launch_site
FROM SPACEXDATABASE
WHERE landing__outcome = 'Failure (drone ship)' AND EXTRACT(YEAR FROM date) = 2015;
```

✓ 0.5s

```
* ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.
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

Esta consulta SQL devuelve los registros que corresponden al año 2015 con aterrizajes fallidos de naves no tripuladas, sus propulsores y sitios de lanzamiento.

Resultados de los aterrizajes en un rango entre 2010-06-04 y 2017-03-20

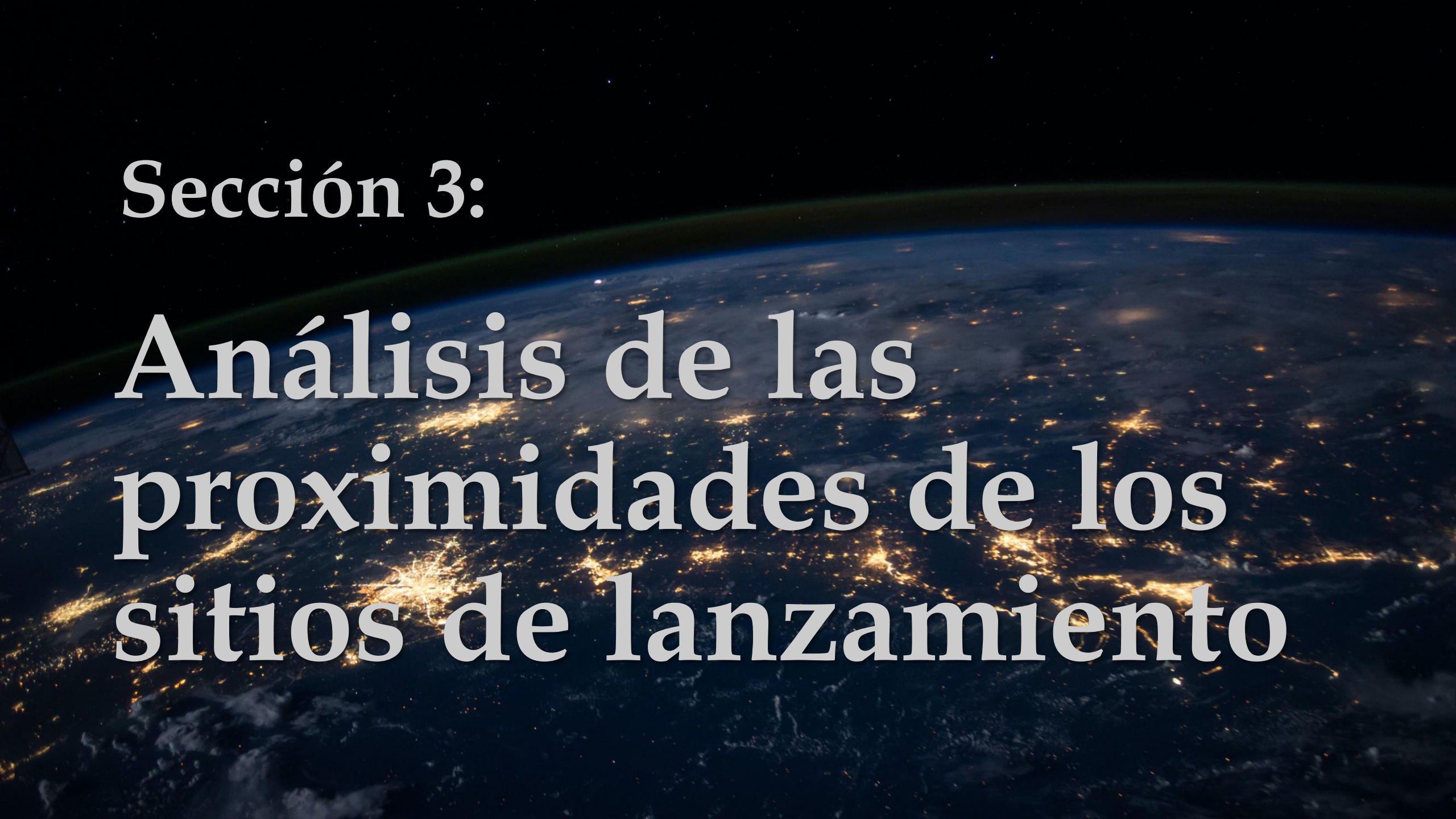
```
%%sql
SELECT landing__outcome, COUNT(*) as count
FROM SPACEXDATABASE
WHERE date BETWEEN '2010-06-04' AND '2017-03-20' AND landing__outcome IN ('Failure (drone ship)', 'Success (ground pad)')
GROUP BY landing__outcome
ORDER BY count DESC;
```

✓ 1.3s Python

```
 * ibm_db_sa://gbk26314:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

landing__outcome	COUNT
Failure (drone ship)	5
Success (ground pad)	3

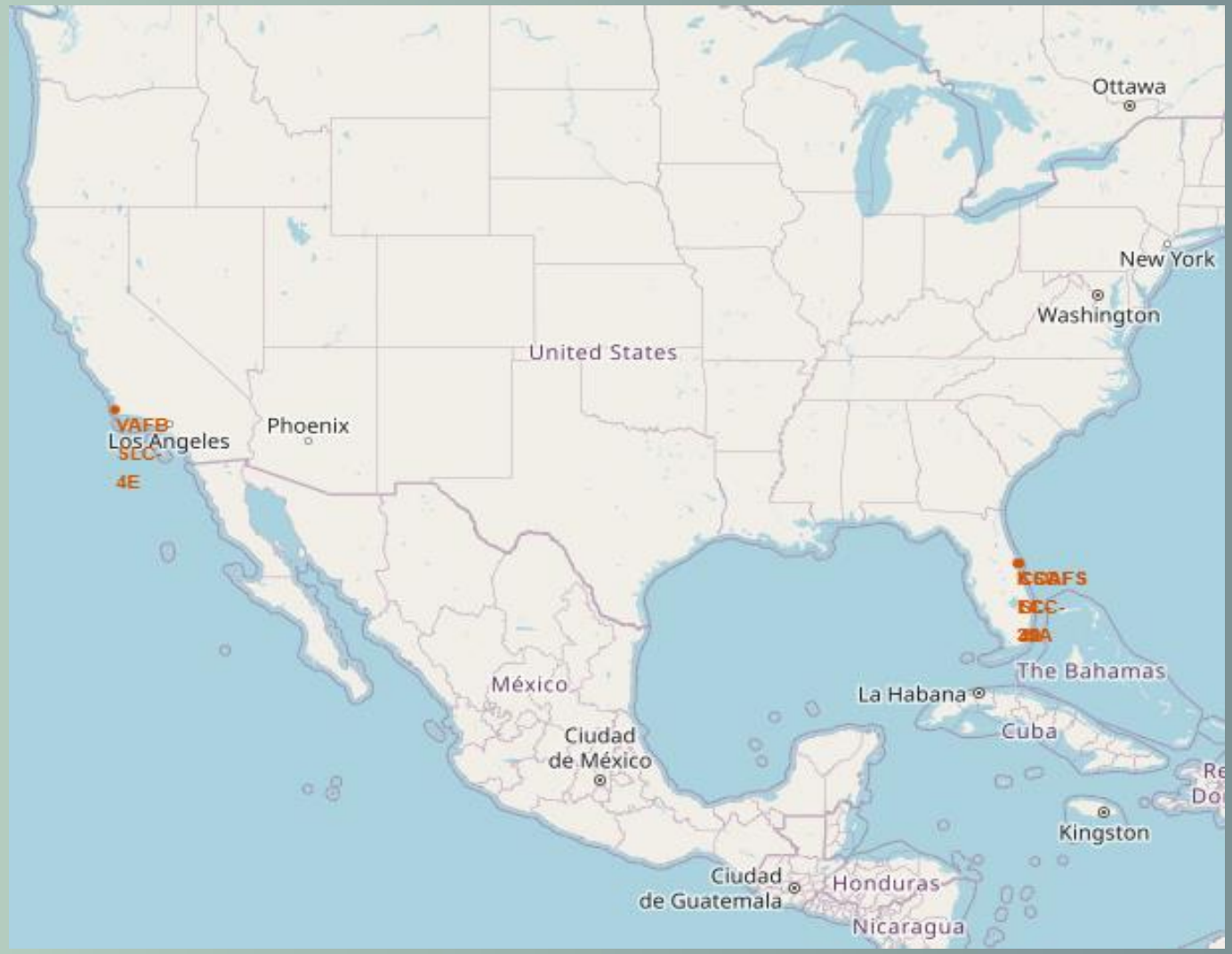
Esta consulta SQL devuelve el recuento de los resultados de los aterrizajes (Failure (drone ship) y Success (ground pad)) entre la fecha 2010-06-04 y 2017-03-20, en orden descendente.



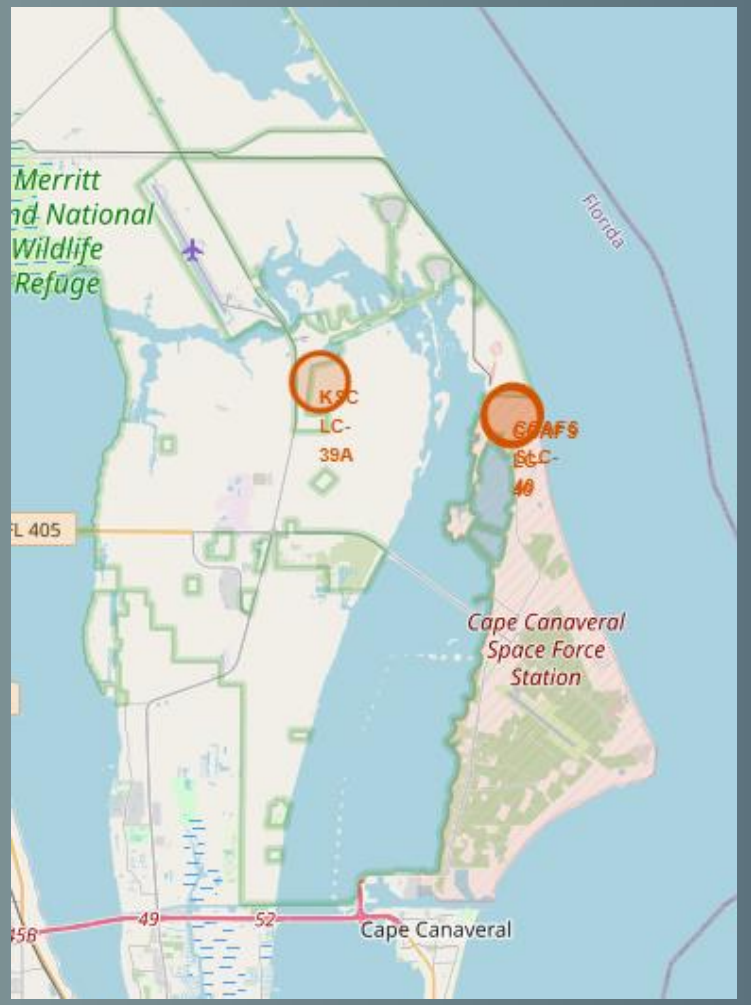
Sección 3:

Análisis de las proximidades de los sitios de lanzamiento

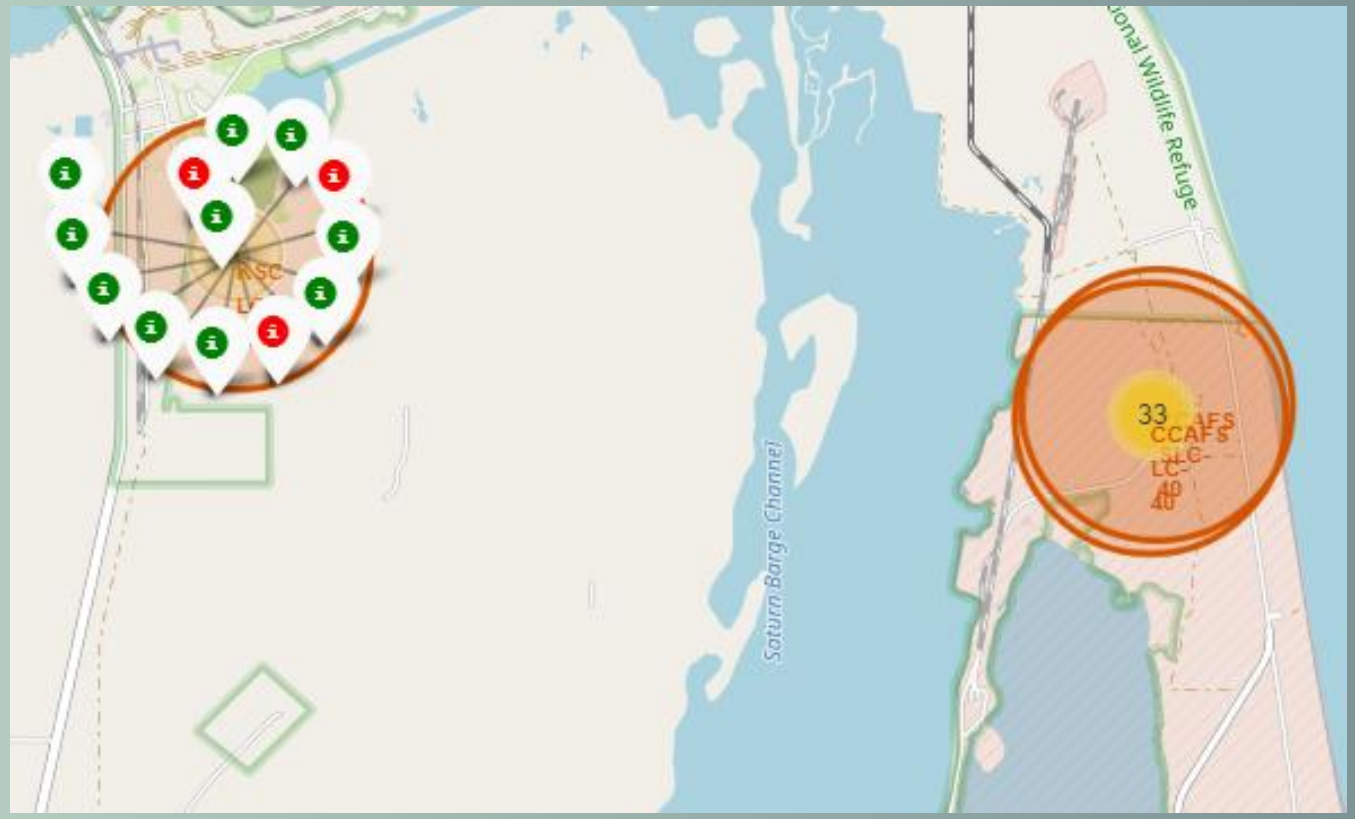
Mapa con los sitios de lanzamiento



Zoom:



Marcadores con los lanzamientos exitosos/fallidos para cada sitio en el mapa



En este ejemplo, del lado izquierdo de la imagen, se pueden ver marcadores verdes y rojos para representar los aterrizajes exitosos y fallidos del sitio KSC LC-39A.

Proximidades de los sitios de lanzamiento

En este ejemplo, se pueden ver las proximidades del sitio CCAFS SLC-40, representadas con 4 líneas azules que apuntan hacia la ciudad, carretera, ferrocarril y costa más cercana.


Sección 4:
Crear un tablero con Plotly Dash

Gráfico con el recuento de lanzamientos exitosos para todos los sitios

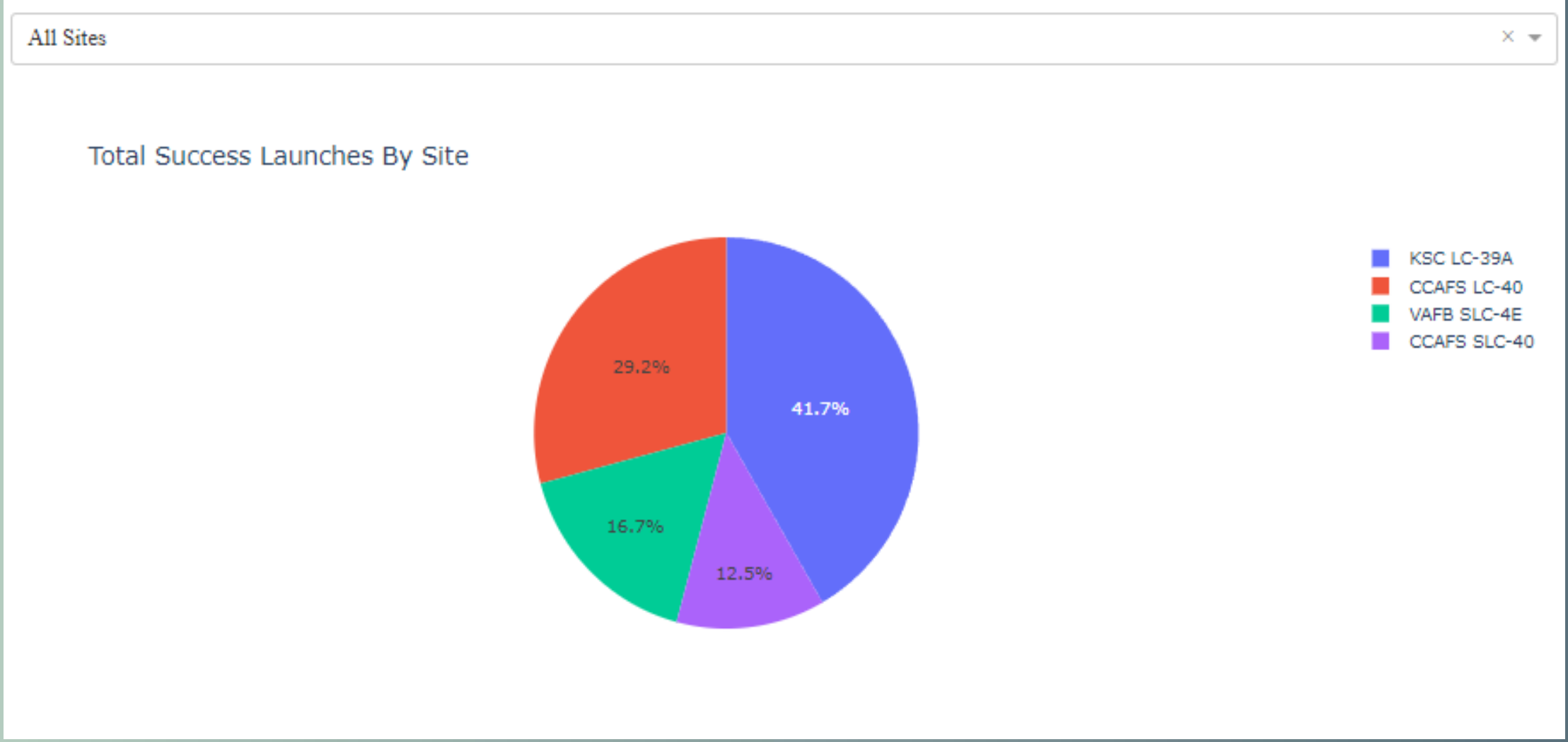
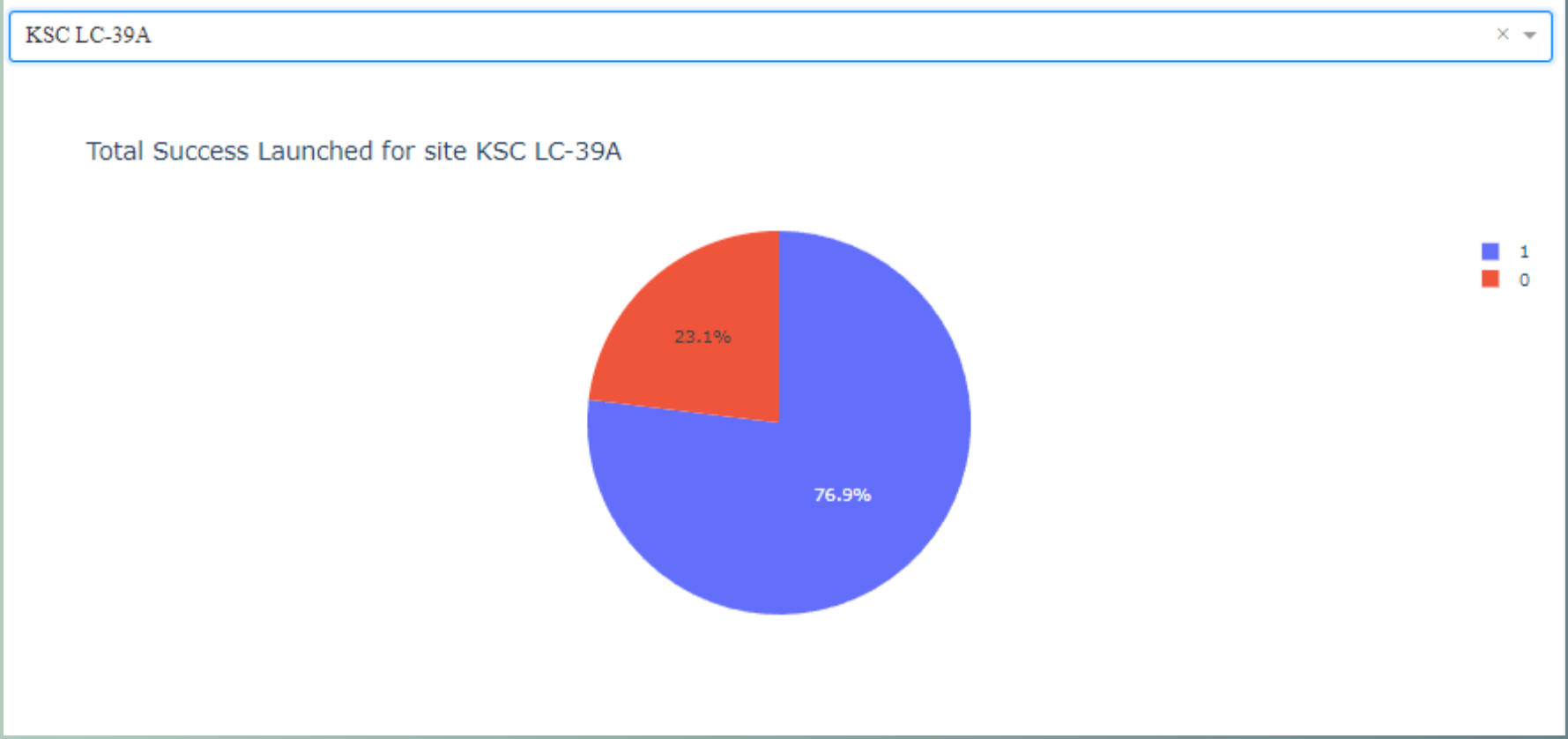


Gráfico del sitio de lanzamiento con la mayor tasa de éxito



Como se puede ver, el sitio KSC LC-39A tiene la tasa de éxito más alta.

Gráfico de dispersión de la carga útil frente al resultado del aterrizaje para todos los sitios

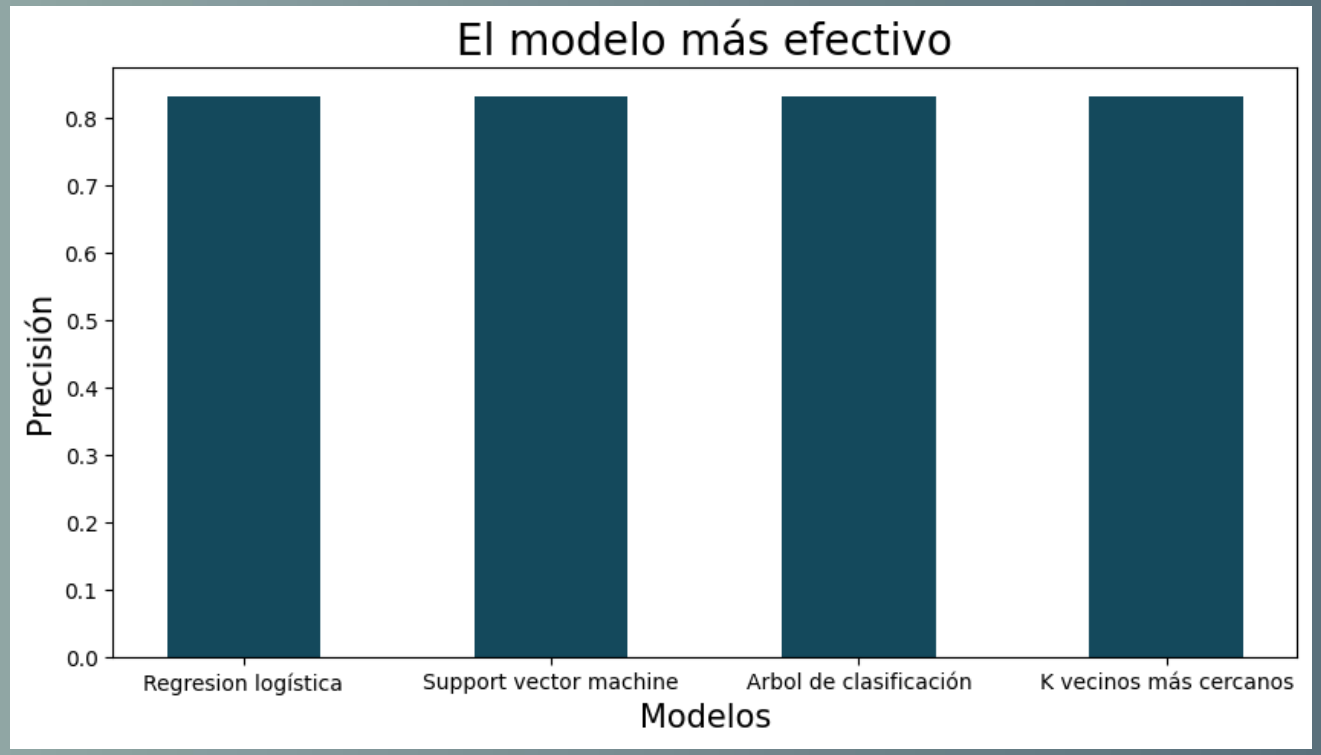
El gráfico presentado se obtuvo a partir de alterar los valores de la barra de carga útil.

Sección 5:

Análisis predictivo con algoritmos de clasificación

Encontrar el modelo que mejor funciona

Todos los modelos tienen prácticamente la misma precisión en el conjunto de prueba: **83, 33%.** Esto significa que los modelos tienen la misma capacidad para predecir correctamente la clase a la que pertenece una muestra en el conjunto de datos de prueba. Esto puede deberse a varias razones, como que los modelos son muy similares en términos de su arquitectura y parámetros, o que el conjunto de datos de prueba es muy fácil de clasificar. Sin embargo, no necesariamente significa que los modelos sean igualmente buenos en general, ya que pueden tener diferentes habilidades en conjuntos de datos diferentes o en situaciones diferentes.

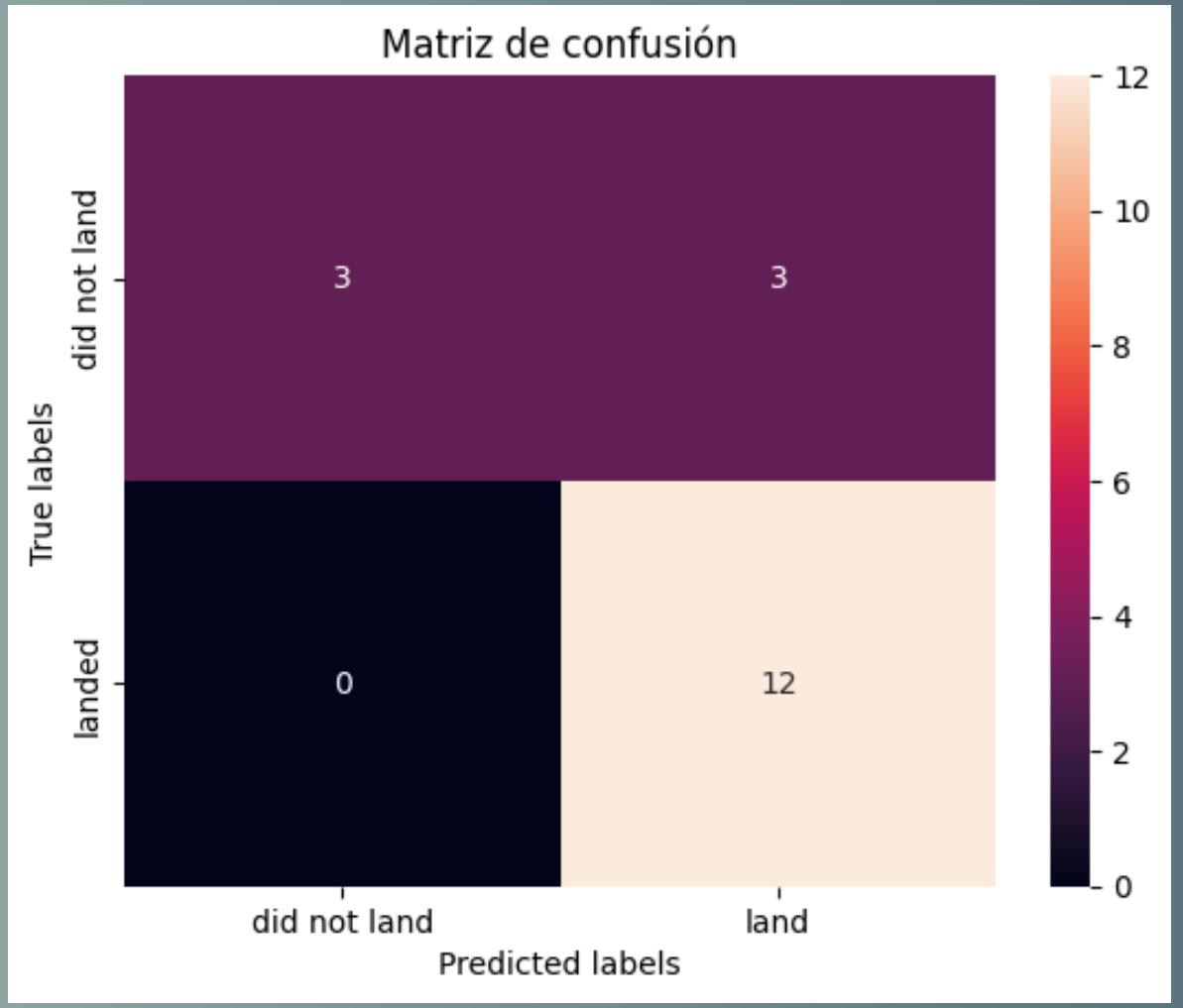


Evaluar los modelos: matriz de confusión

Dado que todos los modelos obtuvieron los mismos resultados para con el conjunto de prueba, la matriz de confusión es la misma en todos ellos.

Interpretación:

- Los modelos predijeron 12 aterrizajes exitosos de manera acertada.
- Los modelos predijeron 3 aterrizajes fallidos de manera acertada
- Los modelos predijeron 3 aterrizajes exitosos de manera incorrecta. Es decir arrojaron 3 falsos positivos.



Conclusiones

- Se completó con éxito el objetivo del proyecto, creandose modelos predictivos para que Space Y pueda competir contra Space X a través de los precios de lanzamiento de cohetes.
- Sin embargo estos modelos deben ser re-entrenados con más datos para aumentar su precisión.

Anexos

- **Github** del proyecto: https://github.com/LilenFr/IBM-SpaceY-es
- Para una mejor visualización, ver proyecto a través de **nbviewer**:

  https://nbviewer.org/github/LilenFr/IBM-SpaceY-es/tree/master/
- **Curso** que dictó el proyecto: https://www.coursera.org/learn/applied-data-science-capstone/home/week/1
- **Tutores**: **Yan Luo**, **Joseph Santarcangelo**


¡Gracias!