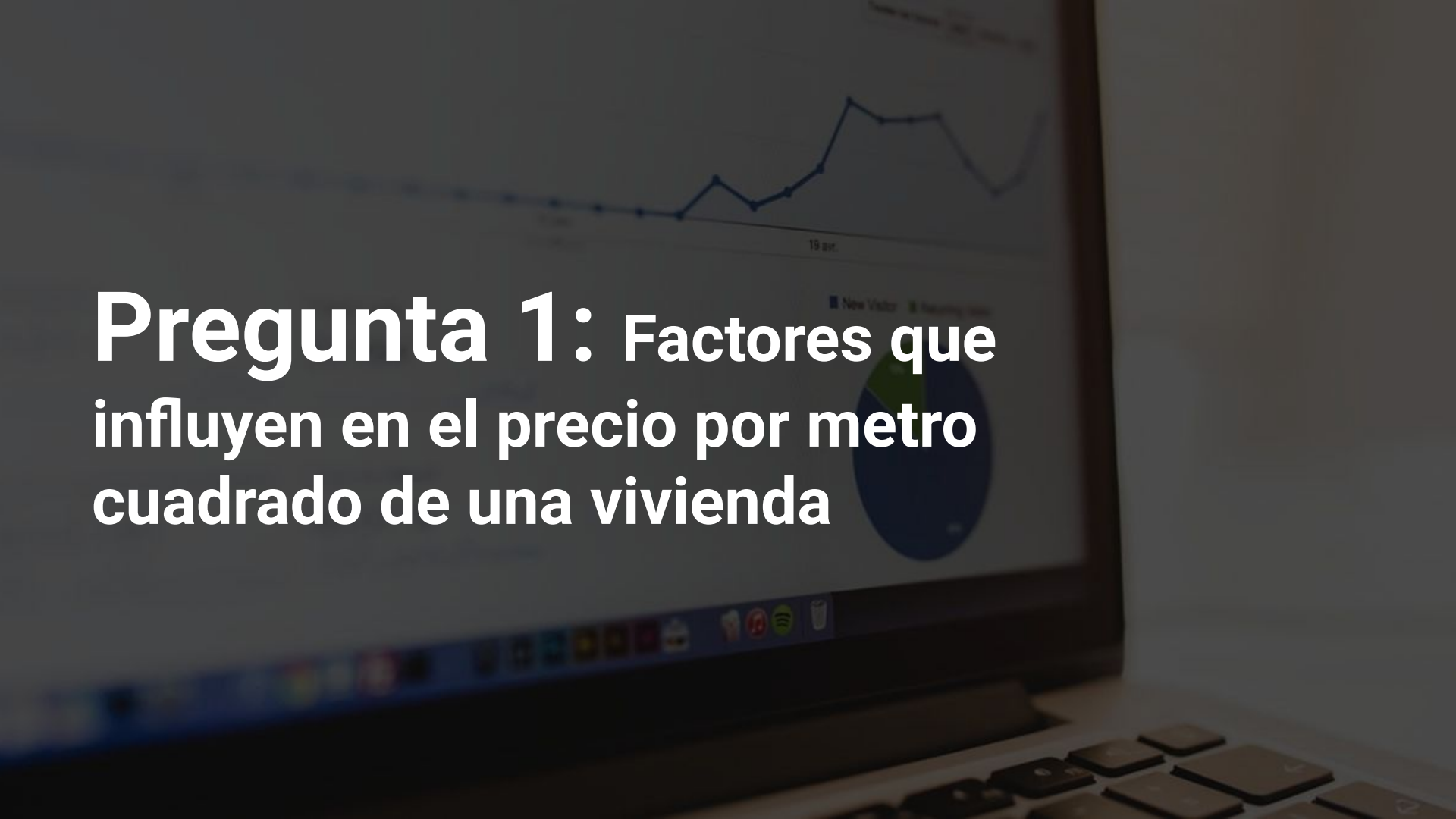


DD360

Prueba

Javier Arturo Hernández Sosa

The background image shows a laptop screen with a dark overlay. On the screen, there is a line graph with two data series: 'New Visitor' (blue line) and 'Returning Visitor' (green line). The 'New Visitor' line shows a general upward trend with some fluctuations, while the 'Returning Visitor' line is relatively flat. Below the graph is a pie chart with a large blue section and a smaller green section. The text 'Pregunta 1: Factores que influyen en el precio por metro cuadrado de una vivienda' is overlaid in white, bold font on the left side of the screen.

Pregunta 1: Factores que influyen en el precio por metro cuadrado de una vivienda

Investigación

Se realizó una investigación en internet sobre los factores que afectan el precio de un inmueble, y los siguientes factores son los más recurrentes.

- Ubicación
- Tamaño
- Orientación, Altura y Vistas
- Amenidades
- Estacionamiento



Desarrollo

- **Revisión de la información**
- **Selección de atributos** (algunos atributos contienen la misma información pero limpia, se descartaron los duplicados sucios)
- **Limpieza de algunos atributos** (monthly_fee - limpiar MXN para poder manipular los datos, location - quitar ubicaciones fuera de la CDMX, etc)
- **Agrupación de los datos en los atributos para promediar** el monto del metro cuadrado de los departamentos y hacer una comparación

NOTA: Se planteaba realizar una regresión lineal para poder ver la importancia de los datos a través de las betas del modelo, pero por los datos nulos se complicó este enfoque.

Conclusiones

Ubicación	Promedio m2
Roma norte	\$62,938
Roma sur	\$60,625
Otro	\$40,057

Disposición	Promedio m2
Frente	\$74,824
Contrafrente	\$70,866
Interno	\$63,511
Otro	\$61,411

En este caso podemos ver como las mejores zonas tienen un promedio más alto de costo por metro cuadrado y las disposiciones con mejor ubicación también tienen un alto costo por metro cuadrado.

Amenidades	Promedio m2
1	\$80,017
2	\$75,057
3	\$68,809
4	\$59,403
5	\$58,059

Amenidades	Promedio m2
6	\$60,020
7	\$59,878
8	\$68,945
Otro	\$62,683

En el tema de amenidades tenemos esta información pero no contamos con el catálogo para saber lo que significa cada valor y poder deducir si hay alguna amenidad que aumente el valor del departamento.

# Estacionamiento	Promedio m2
1	\$61,834
2	\$61,363
3	\$78,373

# Cuartos	Promedio m2
1	\$77,225
2	\$60,624
3	\$57,798
4	\$15,669

En este caso podemos ver como con más lugares de estacionamiento (3) es más caro, además podemos ver que entre más cuartos más baratos, esto puede ser debido a que los cuartos se reducen de tamaño y se vuelven más incómodos.

Tipo de departamento	Promedio m2
Loft	\$71,563
Otro	\$61,182



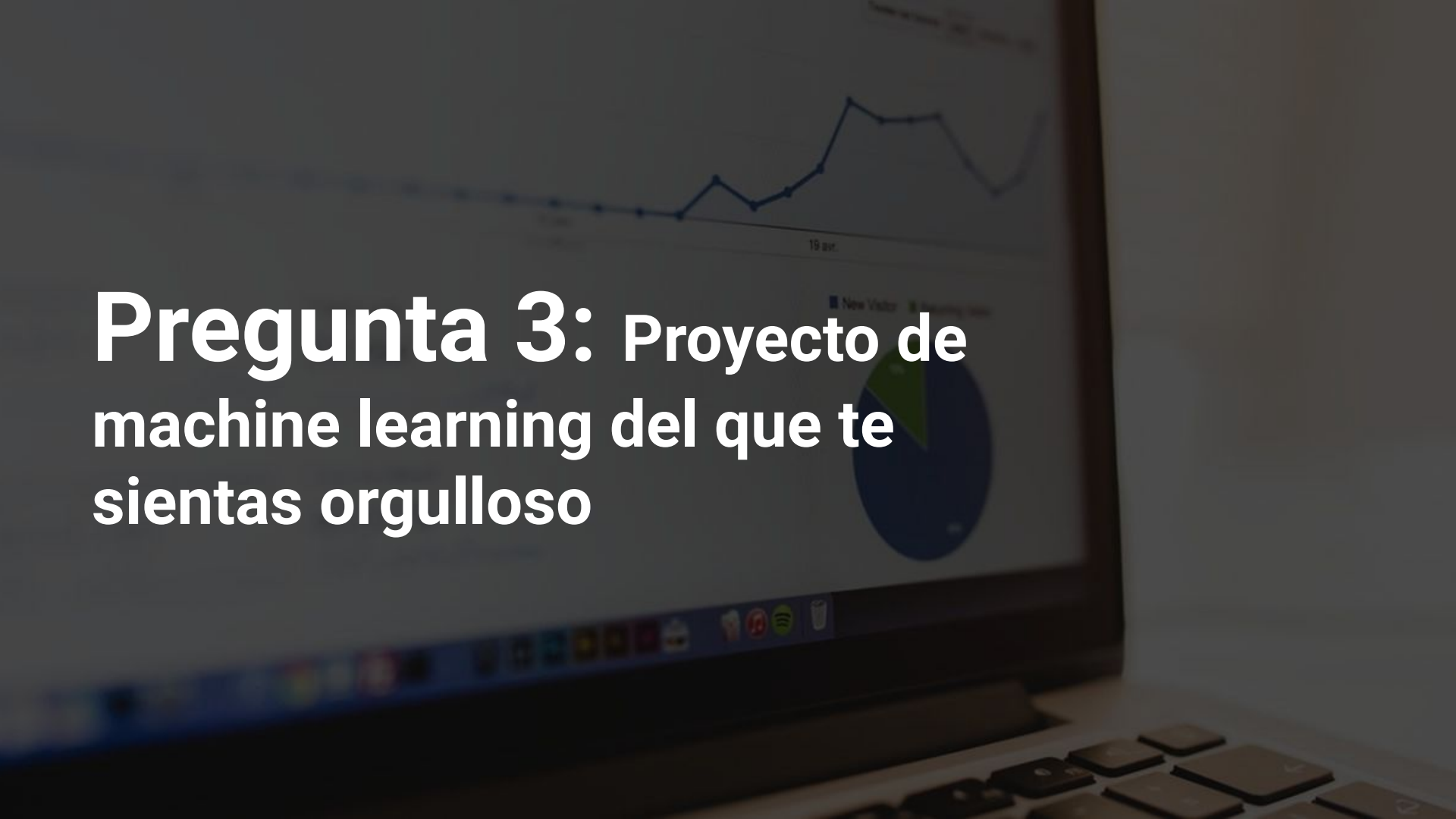
En este otro ejemplo vemos como los departamentos tipo Loft son más caros que otros departamentos.

Estas fueron algunas variables relevantes donde podemos ver una diferencia en los precios del metro cuadro, con esto podemos ver que se cumplen algunos de los datos sobre los factores que alteran el precio de los departamentos.

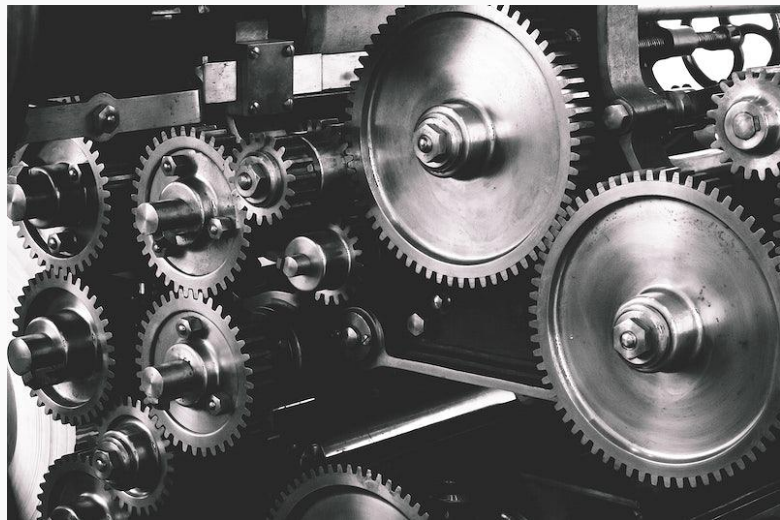
Bibliografía

- <https://sisolinmobiliarias.com/>
- <https://blog.ion.com.mx/>
- <https://es.linkedin.com/>
- <https://www.inmuebles24.com/>
- <https://www.nouhouse.mx/>

Pregunta 3: Proyecto de machine learning del que te sientas orgulloso



RETOS



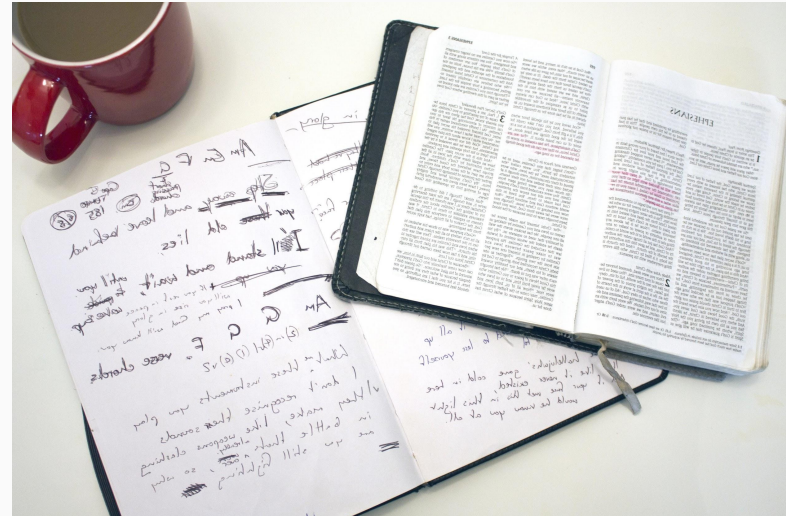
- Modelo sin documentación de código
- Proyecto sin documentación sobre reglas de negocio.
- Gran volumen de datos (1,000 GB)
- Elementos que no sabíamos que pertenecían al proyecto (dashboards, tablas, etc)

Conjunto de datos

El conjunto de datos consistía en métricas de varias partes del proceso, válvulas, nivel de gases, etc.

fecha : Variable : valor

Pero los datos venían encimados y se tenían que filtrar y ordenar por variable.



Metodología utilizada



Para todos los proyectos se utiliza SCRUM, con spring plannings y reuniones diarias (daily).

Para desarrollar el modelo, se usa el ciclo normal de ciencia de datos, reuniones de entendimiento del problema, obtención de datos, validación de datos, limpieza de datos, análisis de datos, ingeniería de datos, selección de características, selección del modelo y puesta en producción.

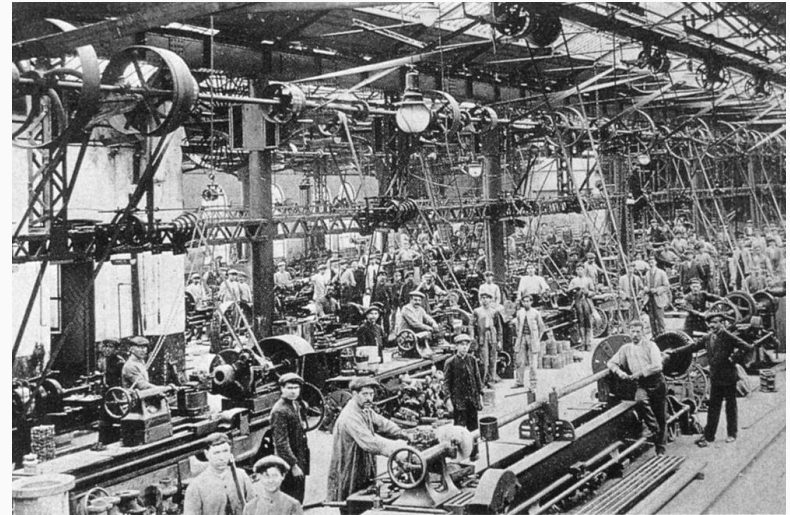
Solución de los retos

- Modelo sin documentación de código:
Entendimiento del código
- Proyecto sin documentación sobre reglas de negocio: **Buscamos algunos responsables para tener algo de luz sobre el proyecto y entender algunos procesos**
- Gran volumen de datos (1,000 GB): **Se utilizó pyspark para el procesamiento**
- Elementos que no sabíamos que pertenecían al proyecto (dashboards, tablas, etc): **Se estudiaron todos los elementos encontrados y se generó la documentación necesaria.**



Forma de incorporarlo

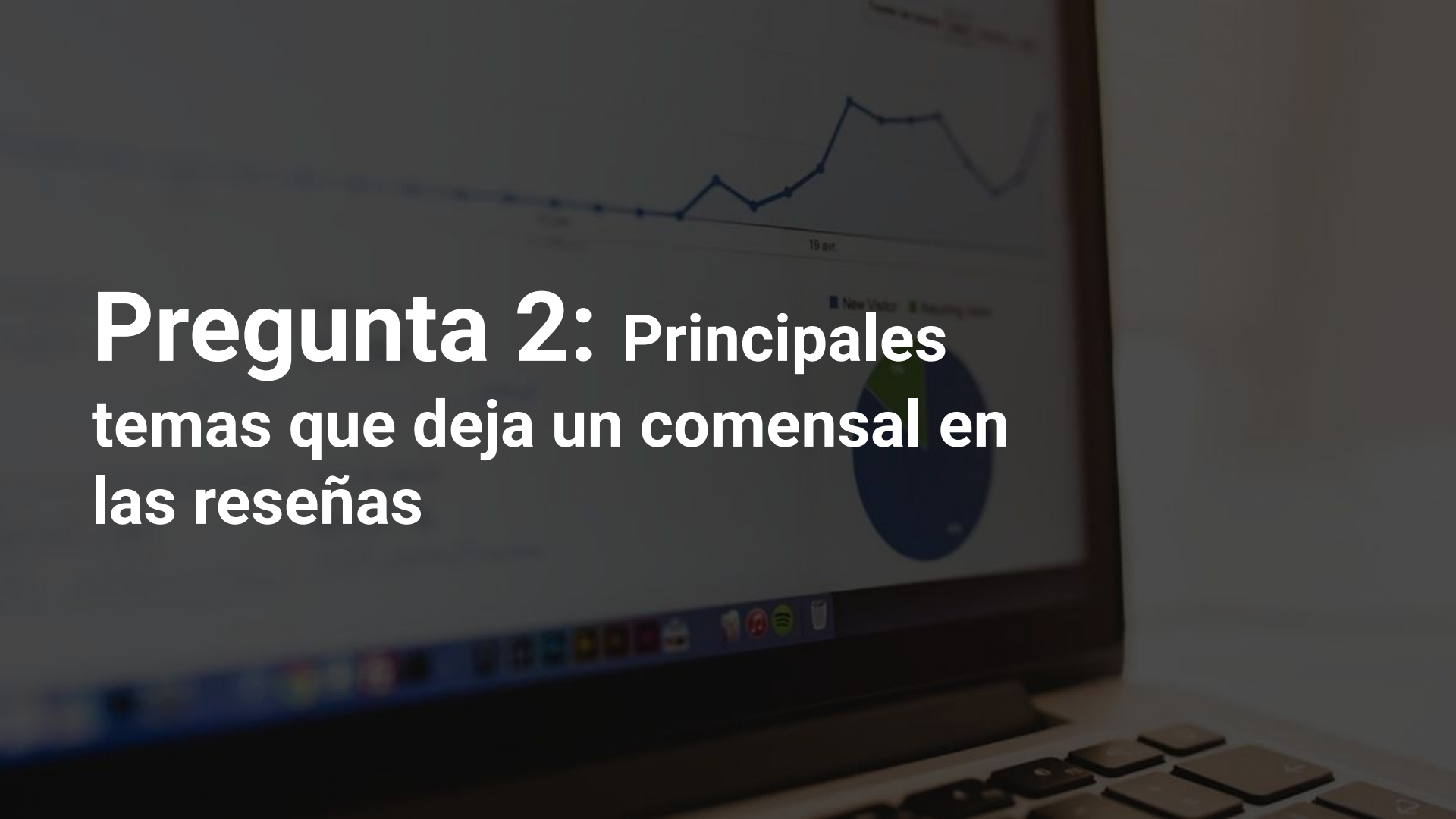
- Se tiene un dashboard donde se guardan las métricas del modelo, indicadores y bitácoras.
- En el dashboard se dan indicaciones para que el proceso termine en un rango de valores óptimos para el producto final, entonces los procesos se pueden modificar para estar en un rango que terminado el proceso se encuentre en el valor óptimo indicado.



Aprendizajes y Resultados



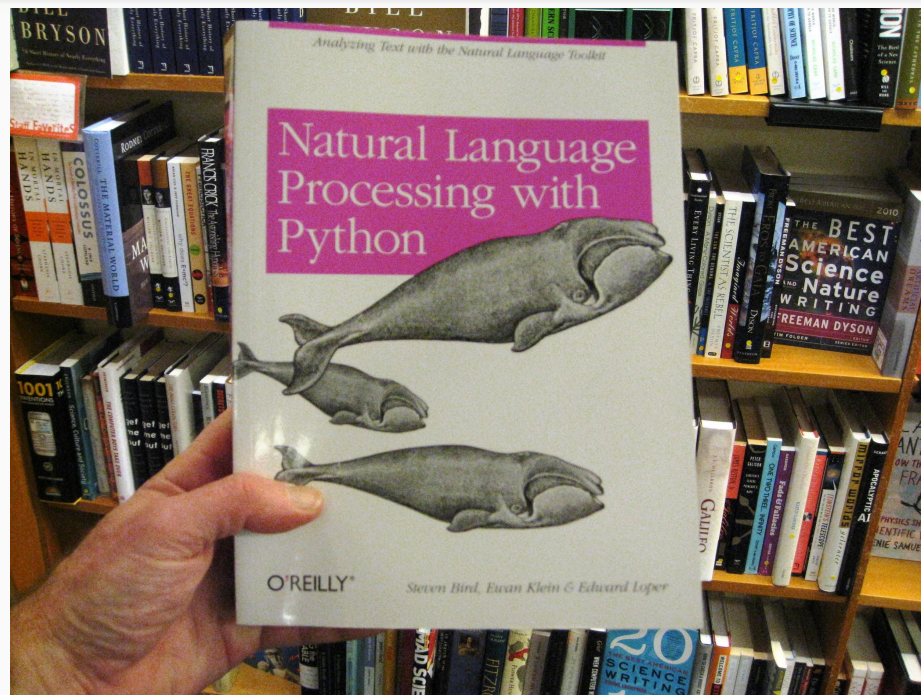
- **Utilizar pyspark** en un proceso real
- **Manipulación de un volumen grande de datos** (antes solo había procesado unos cuantos millones)
- Generar un modelo para producción de alimentos
- **Trabajar bajo mucha presión** al no tener conocimiento del problema ni del código y tener una fecha límite de entrega, al ser una recalibración el tiempo es más corto
- **Se incrementaron las ganancias** de la empresa además no se tenía mucho en cuenta la funcionalidad del dashboard y después de ver las ganancias se contrató un equipo y se capacitó para utilizar el dashboard

The background image shows a laptop screen with a dark overlay. On the screen, there is a line graph with a blue line and a pie chart with a blue and green segment. The text is overlaid on the left side of the screen.

Pregunta 2: Principales temas que deja un comensal en las reseñas

Posible solución

Este problema se podría resolver haciendo una limpieza de las reseñas, quitando las palabras que no ofrecen información (stop words) y luego tomando las palabras que más se repiten en los texto para conocer los temas de los que más se hablan en estas.



Gracias

