

BIG DATA

Jorge González Piedra

PROYECTO FINAL



Jorge González Piedra

RESPONSABLE DEL DESARROLLO DE LA APLICACIÓN

**Fecha:** 05, FEB, 2023

**DESARROLLO**

1. Un informe científico, en el que se transmitan los resultados de los análisis realizados. Aquí explicaremos paso a paso cada uno de los apartados con las conclusiones correspondientes de las tareas realizadas. Podremos incluir secciones de código si es necesario y por supuesto, los resultados de cada una de las tareas realizadas sobre los datos obtenidos a través de la ejecución del código contenido en el documento técnico.
2. Un documento técnico que tendrá el código fuente (PySpark) empleado para la resolución de cada una de las tareas. El código fuente debe ser insertado como imágenes y con un tamaño que permita leer el texto contenido en las imágenes.
3. Una presentación guardada en formato pdf. Esta presentación nos servirá para mostrar los resultados de cada una de las tareas y no contendrá código fuente sino que mostrará los resultados obtenidos siguiendo las guías de presentación que hemos visto en el módulo de proyectos *big data* y *storytelling*.
4. **Cassandra**
   1. **Recuperar registros de la aerolínea” Air China”**



Los datos se encuentras adjuntos a la entrega en la carpeta *airchina\_data\_extraction*

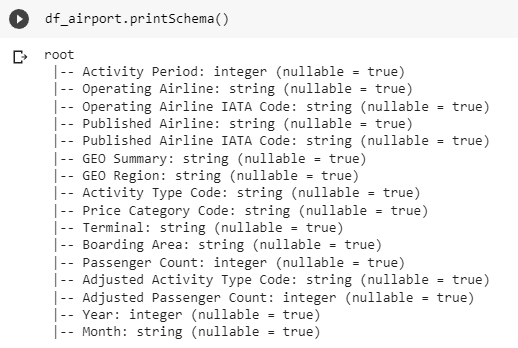
* 1. **Recuperar todos los vuelos de la compañía “Air Berlín” embarcados por la puerta “G”**



Los datos se encuentras adjuntos a la entrega en la carpeta *airberlin\_data\_extraction*

1. **PySpark**
   1. **Creación del Dataframe y análisis preliminar**
      1. **Creación dataframe**





* + 1. **Seleccionar compañías diferentes en el fichero**



* + 1. **Pasajeros de media de los vuelos de cada compañía**



* + 1. **Eliminación registros duplicados por GEO Region**

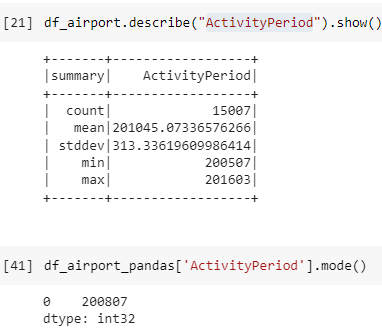


* + 1. **Volcar resultados**

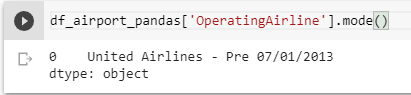


Los datos se encuentran adjuntos a la entrega en el fichero *airtraffic\_drop\_duplicates\_georegion.csv*.

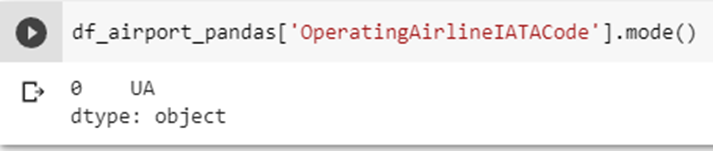
* 1. **Análisis estadístico**
     1. **Análisis descriptivo**
        1. **Activity Period**



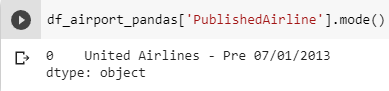
* + - 1. **Operating airline**



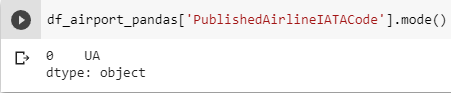
* + - 1. **Operating airline IATA code**

****

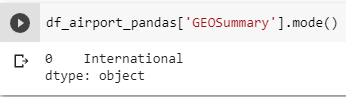
* + - 1. **Published airline**



* + - 1. **Published airline IATA code**



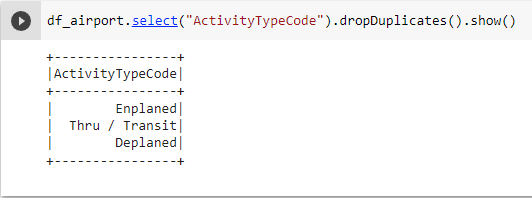
* + - 1. **GEO summary**



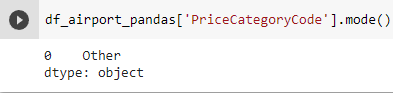
* + - 1. **GEO Region**



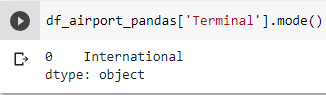
* + - 1. **Activity type code**



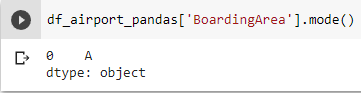
* + - 1. **Price category code**



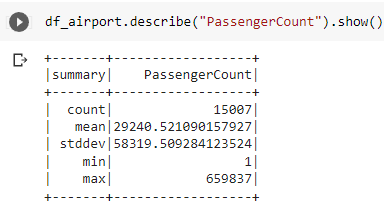
* + - 1. **Terminal**



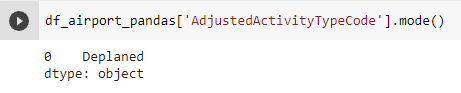
* + - 1. **Boarding área**



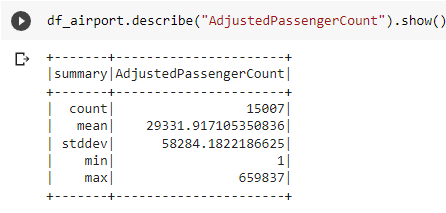
* + - 1. **Passenger count**



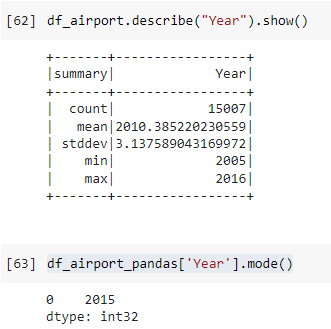
* + - 1. **Adjusted activity type code**



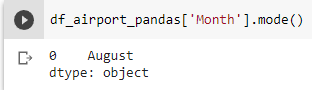
* + - 1. **Adjusted Passenger count**



* + - 1. **Year**

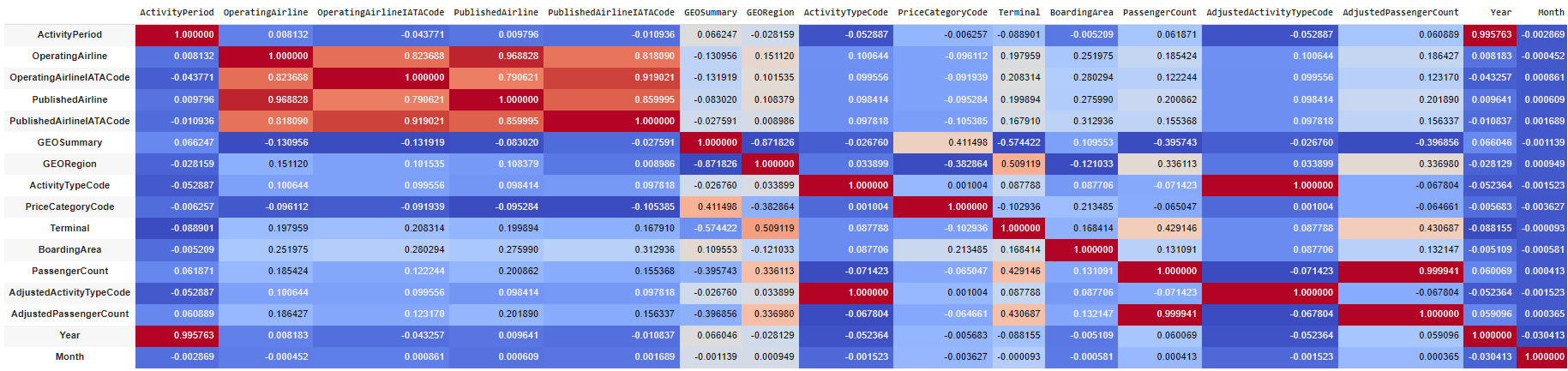


* + - 1. **Month**



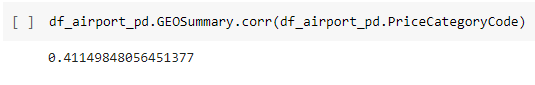
* + 1. **Análisis de correlación**



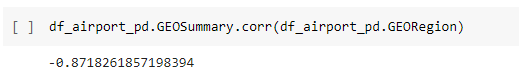


Correlaciones mas fuertes:

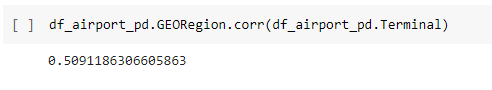
* + - 1. **GEO Summary – Price Category Code**



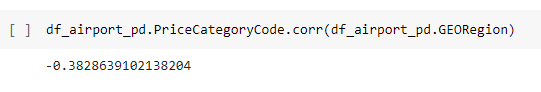
* + - 1. **GEO Summary – GEO Region**



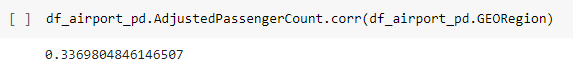
* + - 1. **GEO Region – Terminal**



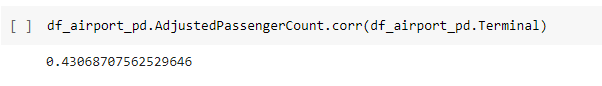
* + - 1. **GEO Region – Price Category Code**



* + - 1. **Adjusted Passenger Count – GEO Region**

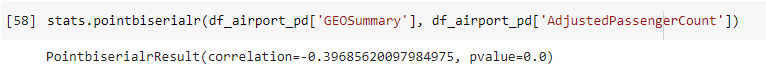


* + - 1. **Adjusted Passenger Count – Terminal**

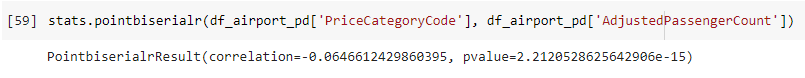


Utilizamos el **método** **Point-Biserial** para calcular de manera más precisa la correlación entre las variables dicotómicas y el número de pasajeros.

* + - 1. **GEO Summary– Adjusted passenger count**

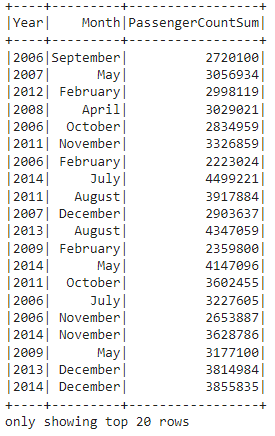


* + - 1. **Price category code – Adjusted passenger count**



* + 1. **Regresión lineal**
* Creación del nuevo DataFrame para estudiar el número de pasajeros a lo largo de los años

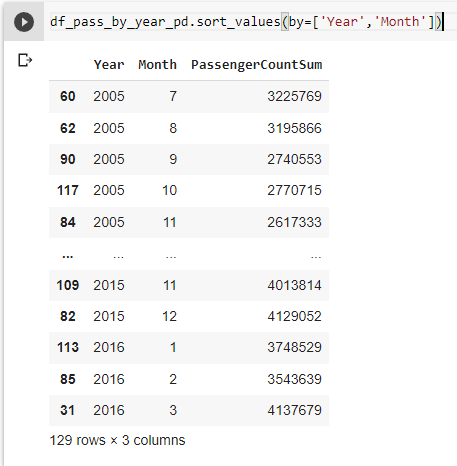




* Conversión de los valores de la columna *Month* a tipo numérico

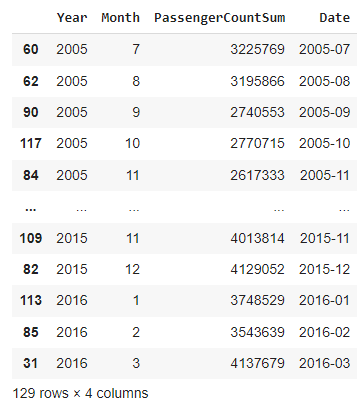


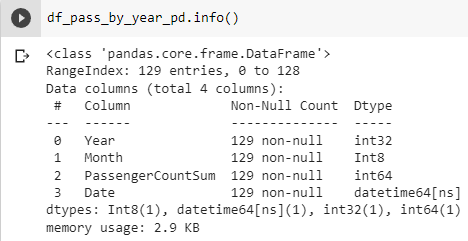




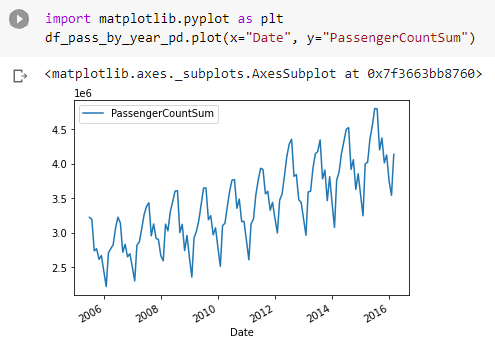
* Creación de una nueva columna para unir Year y Month en un campo de tipo fecha:







* Utilización de la librería matplotlib para la creación del gráfico

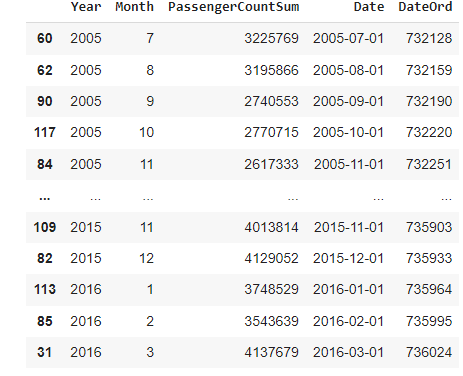


* Importación de la librería para crear posteriormente el modelo de regresión lineal

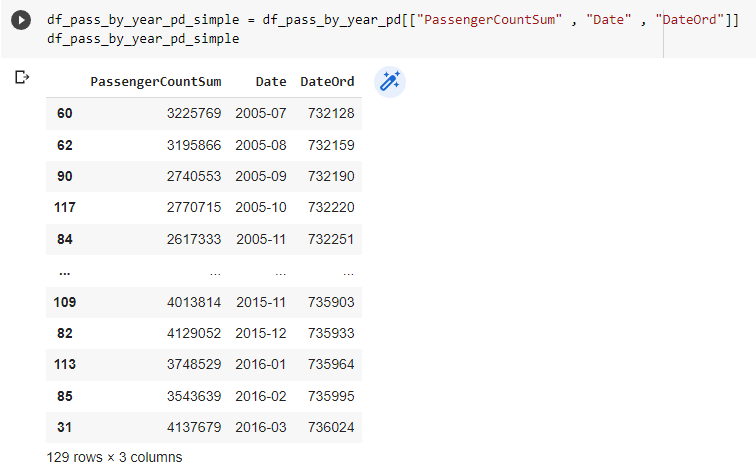


* Conversión del campo Date a un ordinal:

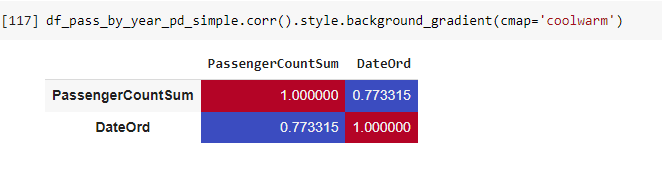


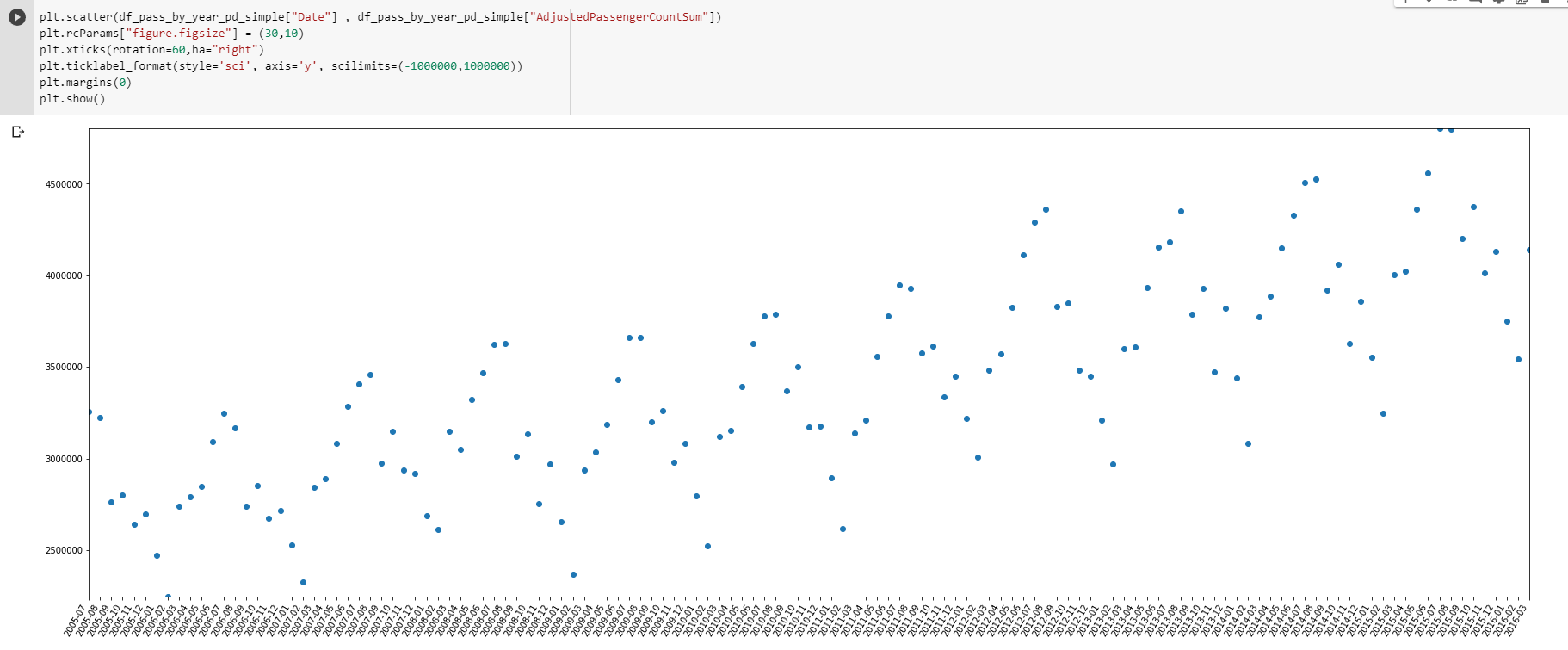


* Simplificación del DataFrame



* Correlación entre las variables del nuevo DataFrame:



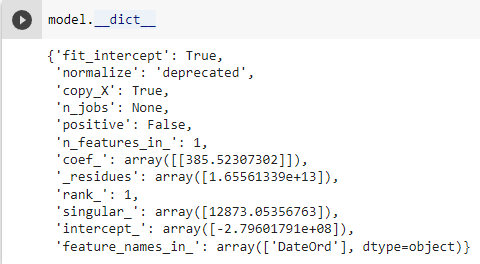


* Creación la regresión:



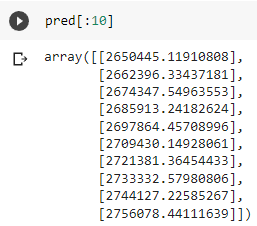






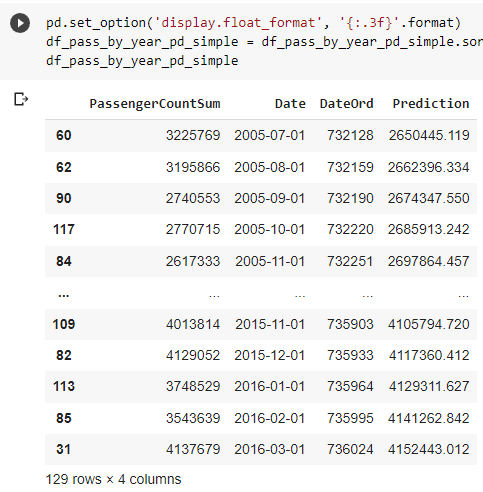
* Creación de la predicción:



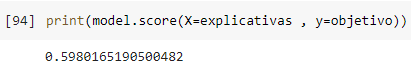


* Ordenación de los valores en función del campo DateOrd:



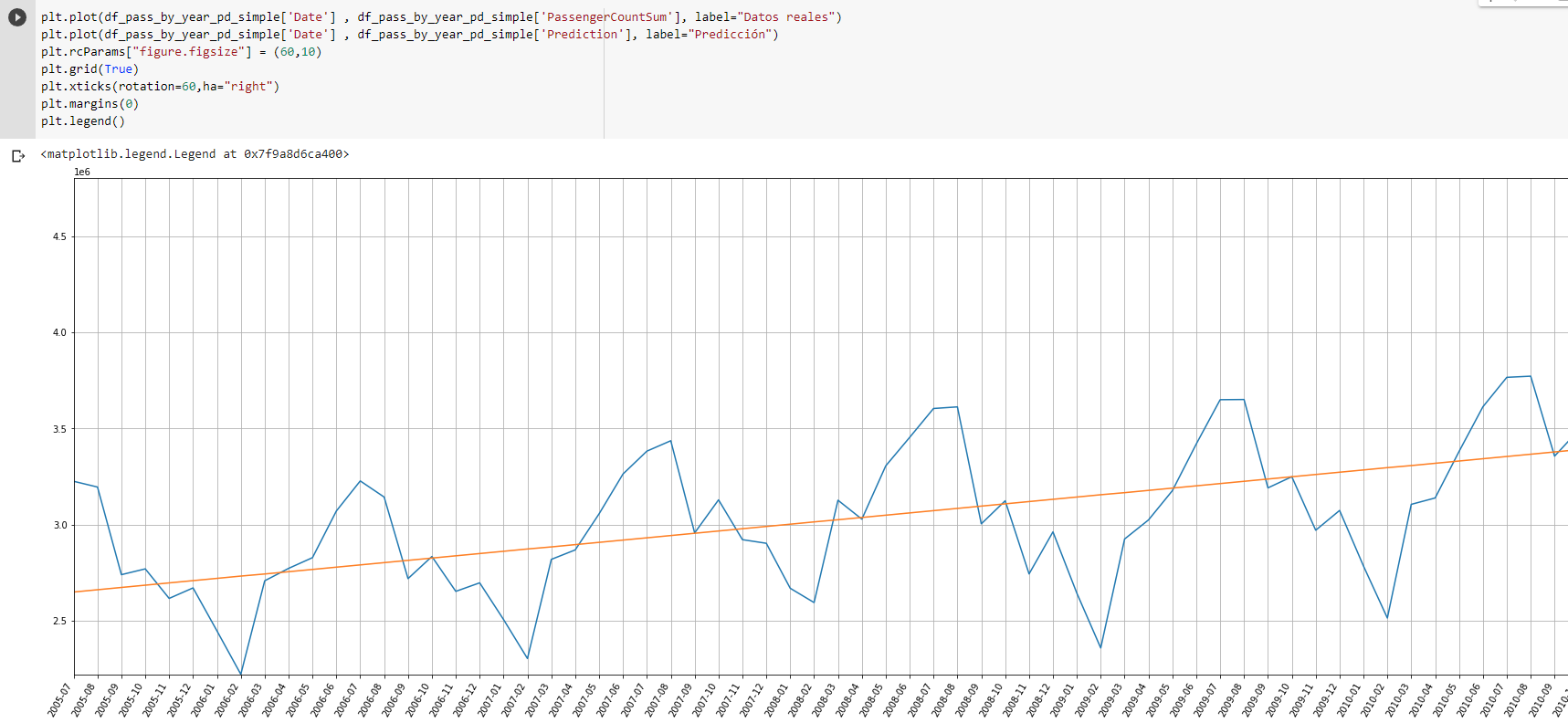


* Precisión del modelo:





* Gráfico de regresión lineal:



* Predicción de datos futuros:



