

## Examen Final

### SISTEMAS AVANZADOS DE BASES DE DATOS

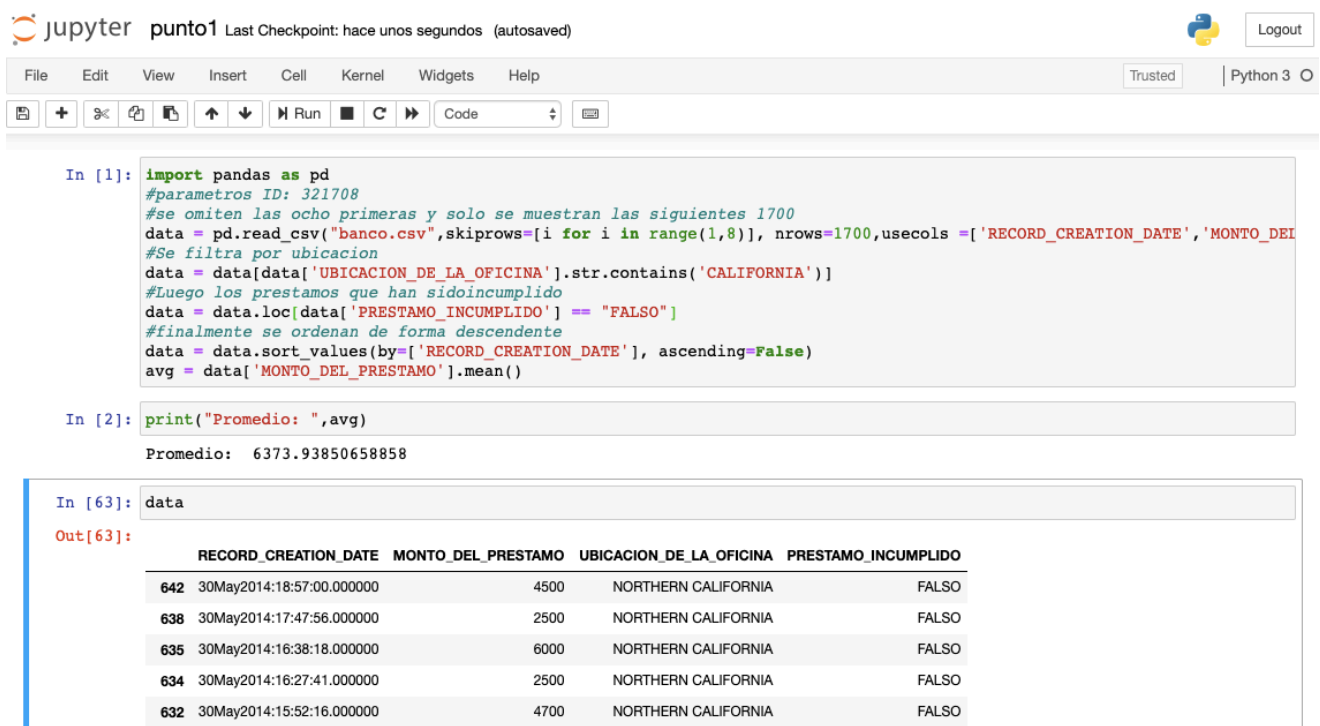
#### PARTE 1: TEORÍA DE BDA

##### 1. (1 punto)

Usted ha sido contratado por un banco regional en el oeste de los Estados Unidos para evaluar su pequeña cartera de préstamos personales. El archivo `banco.csv` (**ver archivo CSV adjunto**) es un repositorio de bases de datos que el banco le ha proporcionado. Cada fila del archivo es el registro de un préstamo en la cartera del banco.

Escriba un programa en Python (usando pandas) para tomar una fracción del dataset desde la fila (últimos dos dígitos de su ID) hasta la fila (últimos 4 dígitos de su ID, en caso de ser mayor a 8000, últimos 3 dígitos), sobre él, devuelva el número de registros que provienen de las oficinas de California, ordénelo descendientemente por la columna `RECORD_CREATION_DATE` y promedie el monto del préstamo entre los registros que hayan incumplido el préstamo. (**Adjunto capturas de ejemplo con Pandas py.**)

El desarrollo de este punto se realizó usando Jupyter, en caso de alguna duda revisar el link de github para confirmar su funcionamiento. Se utilizaron funciones propias de pandas para el ordenamiento y agrupacion del dataframe.



The screenshot shows a Jupyter Notebook interface with the following content:

```
In [1]: import pandas as pd
#parametros ID: 321708
#se omiten las ocho primeras y solo se muestran las siguientes 1700
data = pd.read_csv("banco.csv", skiprows=[i for i in range(1,8)], nrows=1700, usecols = ['RECORD_CREATION_DATE', 'MONTO_DEL_PRESTAMO', 'UBICACION_DE_LA_OFICINA', 'PRESTAMO_INCUMPLIDO'])
#Se filtra por ubicacion
data = data[data['UBICACION_DE_LA_OFICINA'].str.contains('CALIFORNIA')]
#Luego los prestamos que han sido incumplido
data = data.loc[data['PRESTAMO_INCUMPLIDO'] == "FALSO"]
#finalmente se ordenan de forma descendente
data = data.sort_values(by=['RECORD_CREATION_DATE'], ascending=False)
avg = data['MONTO_DEL_PRESTAMO'].mean()

In [2]: print("Promedio: ", avg)

Promedio: 6373.93850658858
```

Below the code, the output of the final command is shown as a table:

```
In [63]: data
```

	RECORD_CREATION_DATE	MONTO_DEL_PRESTAMO	UBICACION_DE_LA_OFICINA	PRESTAMO_INCUMPLIDO
642	30May2014:18:57:00.000000	4500	NORTHERN CALIFORNIA	FALSO
638	30May2014:17:47:56.000000	2500	NORTHERN CALIFORNIA	FALSO
635	30May2014:16:38:18.000000	6000	NORTHERN CALIFORNIA	FALSO
634	30May2014:16:27:41.000000	2500	NORTHERN CALIFORNIA	FALSO
632	30May2014:15:52:16.000000	4700	NORTHERN CALIFORNIA	FALSO

## 2. (1 punto)

Se realiza un estudio para revisar las tasas de mortalidad infantil en hospitales de todo el país. El estudio encuentra que, entre los 10 hospitales con más bajas tasas de mortalidad infantil, 8 de ellas se encuentran en condados rurales. Los hospitales rurales eran remotos, a menudo con poco personal, y tenían un volumen significativamente menor de pacientes en total, en comparación con los hospitales urbanos. Según este estudio ¿Cuál de las siguientes conclusiones sería la más lógica?

- A). Entre los 10 hospitales con las tasas de mortalidad infantil más altas la mayoría son probablemente de condados rurales
- B). Los bebés en las comunidades rurales son más saludables debido a factores externos, como la calidad del aire, la contaminación y la dieta
- C). Los bebés en las comunidades rurales probablemente reciben un mayor grado de atención, porque hay menos pacientes que atender en el hospital
- D). Si los hospitales urbanos aprendieran de los hospitales rurales y adoptaran sus prácticas, podrían proporcionar una mejor atención médica
- E). Todas las hipótesis anteriores son igualmente plausibles

## PARTE 2: PRÁCTICA

Para esta parte, utilice al menos uno de los siguientes datasets acerca de la COVID-19 alojados en notebooks de Kaggle:

<https://www.kaggle.com/koryto/countryinfo>

<https://www.kaggle.com/chrischow/demographic-factors-for-explaining-covid19>

<https://www.kaggle.com/bitsnpieces/covid19-country-data>

<https://www.kaggle.com/jieyingwu/covid19-us-countylevel-summaries>

## SIST. A. BASES DE DATOS – Examen Final

Docente: Juan Sebastián Gómez Rosas

Haremos un análisis general de la pandemia de la COVID-19 aprovechando los datos disponibles. Para esto utilizaremos Google Datastudio (con su cuenta de Google):

Herramienta Datastudio de google: <https://datastudio.google.com/navigation/reporting>

Tutorial de Datastudio:

<https://datastudio.google.com/reporting/0B5FF6JBKbNJxOWItcWo2SVVVeGc/page/DjD>

3. (1 punto) ~~Plantee un objetivo de análisis. Haga uso de la herramienta Google Datastudio~~ para montar y visualizar las columnas y registros del (los) dataset(s). Tenga en cuenta documentar las columnas y eliminar aquellas que no considere pertinentes para su análisis (pre-procesamiento).

Objetivo de análisis:

Reconocimiento de Brasil como el nuevo foco de la pandemia mundial.

4. (1 punto) Cree un reporte que permita visualizar datos básicos de sus datasets incluyendo países y cantidades por país, orígenes y aquello que usted considere pertinente, para esto válgase de las herramientas de gráficas de Datastudio y de los filtros que provee la herramienta de acuerdo con su dataset.

Enlace del reporte:

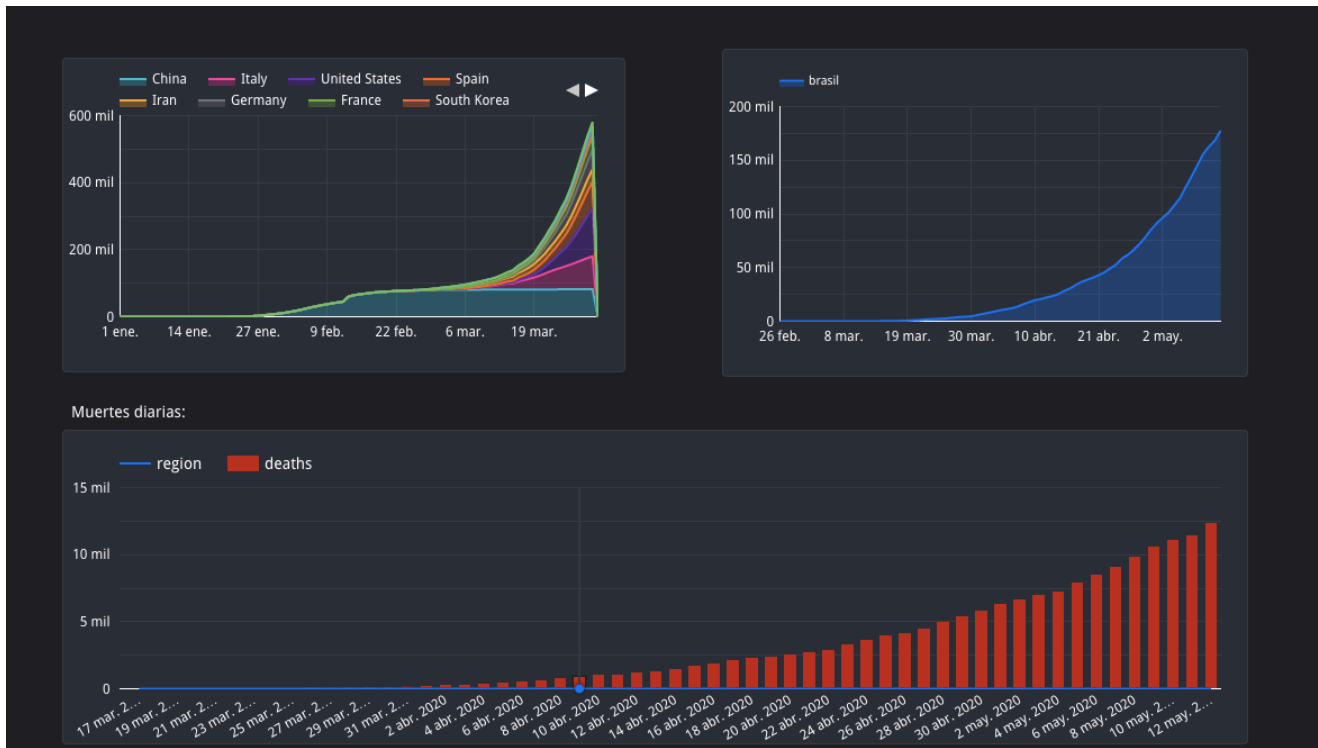
<https://datastudio.google.com/s/lukcdXOWFX8>



imagen 1-informe covid-19 en Brasil

El reporte se realiza conforme al tema de análisis, se usan diferentes graficas otorgadas por data estudio.

**5.(1.5 puntos) Anexe una página adicional al reporte en la que se pueda observar uno o varios gráficos que apunten al objetivo de análisis.**



*imagen 2-Análisis de datos*

En el análisis podemos filtrar por el país que queremos comparar a Brasil, podemos ver que Brasil se comporta como países que anteriormente fueron el foco de la pandemia del covid 19, tales como: China, Italia y Estados Unidos. Incluso su curva de crecimiento es aun mayor que la de estos países.

También vemos el crecimiento de las personas fallecidas a causa del covid-19, entre otros reportes. Se grafican las muertes por cada una de las regiones y estados de país, ubicando al sudeste como el foco de contagios en Brasil.