UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

CC-3074 Data Mining

EJERCICIO No. 1
Introducción a Procesamiento de Datos

Este ejercicio guiado tiene como objetivo que el estudiante se familiarice con la herramienta y plataforma de su elección para procesamiento de datos. Durante el ejercicio abordará diferentes actividades comunes en los proyectos de minería de datos como lo son: la carga de un dataset, la manipulación del mismo y la visualización de resultados.

Preparación del ambiente

Previo a iniciar el ejercicio, el estudiante deberá realizar una investigación de las diferentes herramientas y plataformas que existen para trabajar proyectos de data science y de data mining. La elección de la misma será importante para el resto del curso ya que será conveniente que los siguientes proyectos los trabaje sobre la misma. Algunas herramientas, lenguajes o plataformas que puede considerar para este propósito (pero sin limitarse a ellas) son las siguientes:

- Python
- R
- Julia
- Spark

Definición del objetivo y carga del Dataset

Para este ejercicio particular, el objetivo está deteminado de manera previa. El objetivo principal será determinar cómo el comportamiento de los ingresos de las grandes compañías en los Estados Unidos ha cambiado históricamente. Para ello contamos con un dataset que se ha extraído del archivo público de la revista Fortune dentro de un CSV para facilitar su exploración. Este CSV será provisto por su instructor y deberá cargarlo dentro de su ambiente de trabajo.

Exploración de los Datos

Dependiendo de la herramienta que haya seleccionado los comandos que utilizará para esta actividad del ejercicio pueden ser diferentes, sin embargo en cualquier herramienta debería poder lograr lo siguiente:

- Cargar el dataset descargado dentro de una variable que lo represente y sobre la cual después pueda ejecutar acciones y transformaciones (por ejemplo en Pandas/Python esto puede lograrse con el comando `pandas.read_csv('fortune500.csv')`).
- Visualizar unas filas del dataset para hacerse una idea del contenido del mismo, deberá poder comprender la estructura que tiene (columnas) y unos cuantos datos de cada una.

```
[6]: df.head()
```

[6]:

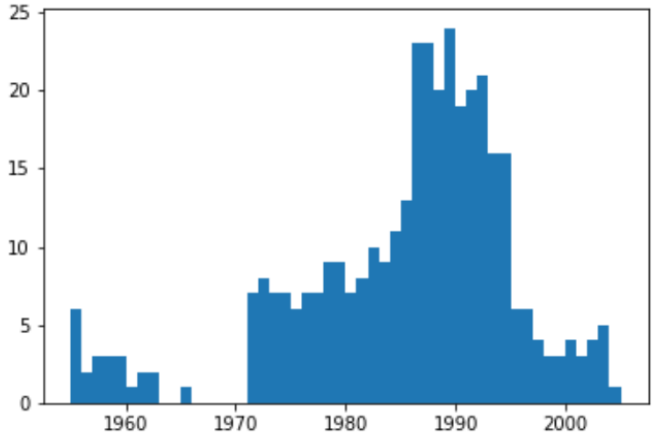
	Year	Rank	Company	Revenue (in millions)	Profit (in millions)
0	1955	1	General Motors	9823.5	806
1	1955	2	Exxon Mobil	5661.4	584.8
2	1955	3	U.S. Steel	3250.4	195.4
3	1955	4	General Electric	2959.1	212.6
4	1955	5	Esmark	2510.8	19.1

- Para un análisis posterior, los nombres de las columnas con mayúsculas, espacios y caracteres especiales no son ideales. Cambiemos los nombres de las columnas dentro de nuestro dataframe a los siguientes:
    - year, rank, company, revenue, profit
- ¿Cuántos registros hay en el dataframe?, en teoría el dataframe cuenta con los últimos 50 años de las top 500 empresas de los Estados Unidos, por lo que el número de registros debería rondar por los 25,000. Verifiquemos que esto sea así.
- Revisemos los tipos de dato que la herramienta que estamos utilizando está asignando a cada una de las columnas que cargamos y verifiquemos que sean lo que esperamos: las primeras dos deberían ser enteros y las últimas dos de tipo float. En caso no sea así identifiquemos en nuestra herramienta cómo lo podemos arreglar.

Preparación de la data

Luego de que hicimos una pequeña exploración sobre la data, en el último punto de la sección anterior puede que nos hayamos topado con un problema particular. La columna **profit** cuenta con algunos registros que no se están interpretando como números flotantes. ¿Qué debemos hacer con ellos?

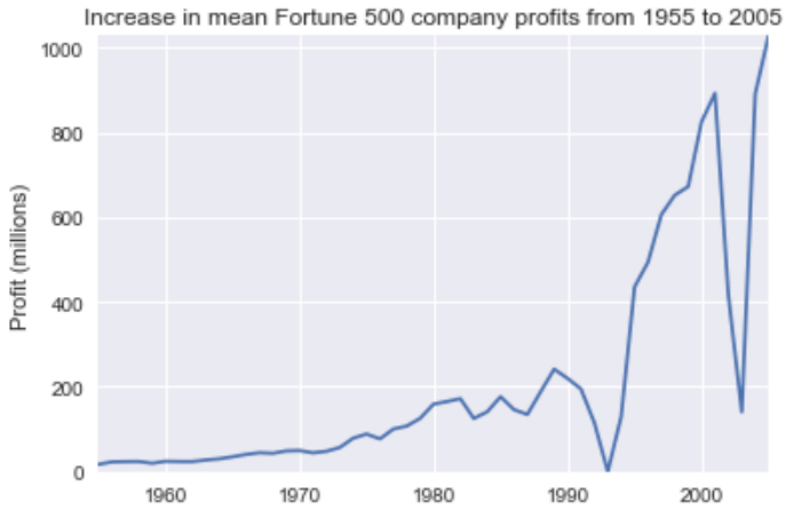
- Lo primero que vamos a hacer es identificarlos. Investigue con la herramienta que está utilizando cómo puede desplegar los casos que no se están logrando convertir a enteros para poder tener una idea de contra qué nos estamos enfrentando.
- Luego de ejecutar el punto anterior deberá lograr identificar que hay varios registros que tienen el valor "N.A.". ¿Es este el único tipo de registros que nos están dando problema, o habrá otro que debamos considerar?
- Luego de determinar que sólo son estos los casos, hagamos un conteo de cuántos son para poder decidir qué haremos con ellos.
- Son 369 casos. Para un dataset de 25,500 registros estos representan menos del 1.5% del total, por lo que podemos considerar removerlos del estudio. Antes de proceder con su exclusión, utilicemos nuestra herramienta para visualizar su distribución en una gráfica para descartar que todos los casos pertenezcan a un solo año o a un grupo pequeño.



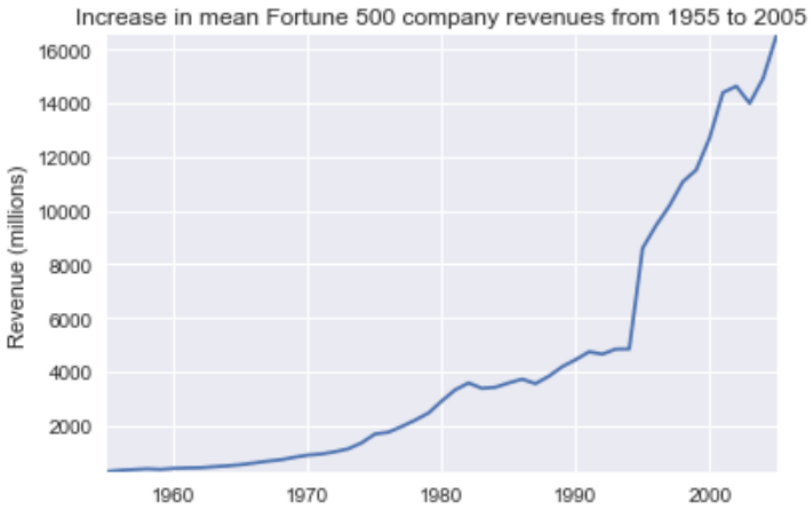
- Luego de graficar podemos determinar que la mayoría de los años con incidencias, estas se quedan por debajo de 20 (un 4% de esos años) y para los 4 años en los que sí sobrepasa los 20 no llegan a 25 (5%) por lo que consideramos que sí es válido excluirlos del dataset para nuestro propósito.
- Apliquemos el filtro y procedamos a confirmar que se ejecutó haciendo un nuevo conteo sobre el dataframe.

Visualización de la data

Lo siguiente que vamos a hacer es atacar el objetivo que nos planteamos y lo vamos a hacer graficando la ganancia (profit) promedio que han tenido las empresas a lo largo de los años.



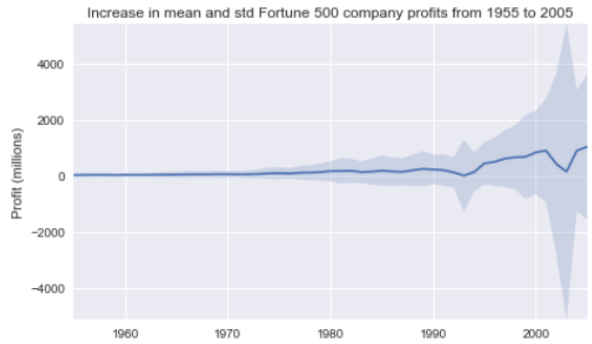
Parece que podemos identificar un crecimiento exponencial, sin embargo hay unas caídas importantes que podemos visualizar. Para entender un poco más sobre estas caídas veamos cómo se comportaron los ingresos (revenue) en el mismo tiempo.

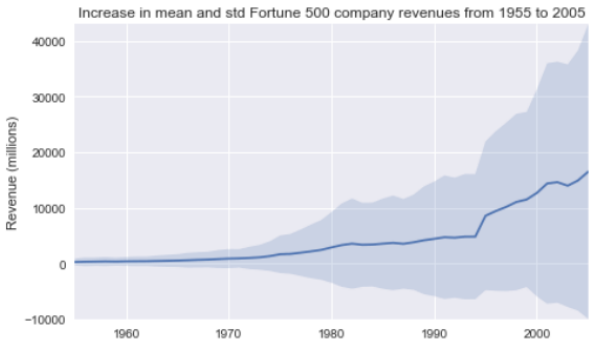


Esta nueva visualización nos cuenta algo más de la historia que estamos construyendo. Los ingresos no fueron golpeados tan fuertemente en esas dos recesiones.

Sería interesante poder visualizar junto con la media en una superposición con la desviación estándar sumada y restada a la misma para tener una mejor idea del

comportamiento de las empresas. Esta visualización requerirá skills más avanzados para lograrla, por lo que no se sienta desanimado si no logra generarla. El instructor compartirá luego el código que él utilizó para construir las propias.





En las gráficas podemos ver una desviación estándar enorme, y podemos ver cómo mientras unas empresas están generando billones otras los están perdiendo y que esa desigualdad ha crecido de manera muy acelerada a partir de las décadas de los 80s y 90s.

¿Quizá podríamos comparar sólo el top 10% y sus ingresos y ganancias sean menos volátiles?

Preguntas como esta podríamos seguir haciéndonos para tratar de comprender mejor el comportamiento de las empresas y sus ingresos y consecuentemente mejorar nuestra aproximación al objetivo que nos planteamos. Para motivo de este ejercicio vamos a dejar el análisis hasta acá pero se motiva a hacer una exploración más extensa si así lo desea el estudiante.