Recolección de datos, limpieza y validación

En las clases anteriores ya se han trabajado algunos temas particulares, desde la exploración a primera vista, importación de datos y forma de cargar dentro de un entorno.

Ademas se observó que existe un set de datos que ya esta **pre-procesado**. Esto quiere decir que ya se le aplico alguna limpieza y validación para que se pudieran trabajar con ellos. Pero en este curso abordaremos ese tema.

Así que preparemos todo nuestro set para poder trabajar con el.

Vamos allá

In [ ]:
```
# Importando librerías
import empiricaldist
import janitor
import matplotlib.pyplot as plt
import numpy as np
import palmerpenguins
import pandas as pd
import scipy.stats
import seaborn as sns
import sklearn.metrics
import statsmodels.api as sm
import statsmodels.formula.api as smf
import statsmodels.stats as ss
import session_info
```

Establecer apariencia general de las gráficas

In [ ]:
```
%matplotlib inline
sns.set_style(style='whitegrid')
sns.set_context(context='notebook')
plt.rcParams['figure.figsize'] = (11, 9.4)

penguin_color = {
    'Adelie': '#ff6602ff',
    'Gentoo': '#0f7175ff',
    'Chinstrap': '#c65dc9ff'
}
```

Cargar los datos

Datos crudos

In [ ]:
```
raw_penguins_df = pd.read_csv('dataset/penguins_raw.csv')
```

Datos Preprocesados

NOTA: Puede que no usemos estos datos

In [ ]:
```
preprocessed_penguins_df = pd.read_csv('dataset/penguins.csv')
```

Preguntas:

- ¿Cuantos años en promedio vive un pingüino?
- ¿Las hembras viven mas que los machos?
- ¿Cuál es el rasgo más importante para definir el sexo de un pingüino?
- ¿Existe una relación de tamaño en cuerpo o algún elemento en el pingüino para que sea hembra?
- De esas 3 especies ¿una de ellas es mas grande que las demás?
- ¿Existe una relación entre tamaño y edad?
- ¿En que parte o isla hay una mayor población de pingüinos?
- ¿Existe una relación entre el sexo y el lugar (isla) que predomine?
- ¿Qué especie es más común?

¿Qué es la recolección de datos?

En general es el proceso que hicieron los científicos de capturar los datos sobre la Antártida.
De manera más formal:

> "Forma de recolección de información que permite obtener conocimiento de primera mano e ideas originales sobre el problema de investigación"

Tipos de recolección de datos:

Primaria:

Son los datos que recolectamos de primera mano; a través de entrevistas, encuestas experimentos y otros. En estos casos tu eres quien decides que recolectas y que quieres medir, por lo que tenemos el control total de los datos que necesitaremos y exploraremos, para dar respuesta a las preguntas.

Secundaria:

En este tipo usaremos datos previamente recolectados por terceros (es decir una fuente primaria externa). Ejemplos son:

- datos de gobierno
- empresas

Son datos que se encuentran abiertos y podemos utilizar en un análisis. Sin embargo no tenemos control sobre los parámetros que se midieron o el objetivo inicial por el cual se recolectaron los datos.

Terciaria Son datos que se encuentran de fuentes externas, y por lo regular son vendidos por empresas.

Un ejemplo podría ser datos relacionados a comportamiento de usuarios con sus compras.

La desventaja es que no tienes control sobre los datos y puede no ayudarte con tu objetivo particular, pero son datos que siempre están disponibles.

¿Qué es la validación de datos?

> Es el proceso de asegurar la consistencia y precisión dentro de un conjunto de datos.

Es decir que no tengamos comportamientos extraños dentro de nuestros datos, por otro lado debemos recordar lo siguiente:

> Si los datos no son precisos desde el comienzo, los resultados definitivamente no serán precisos.

https://www.safe.com/what-is/data-validation/

¿Qué se debe validar para asegurar la consistencia de los datos?

Verificar el modelo de datos

Es decir si nos encontramos utilizando datos de otras personas, verifica que preguntas querías responder con este conjunto de datos. Si somos la persona que llevamos a cabo las entrevistas o los experimentos; haz demasiadas preguntas para saber si con los datos que vamos a empezar a recolectar vamos a poder responder las preguntas, y si estos datos van a tener sentido o existirá algún sesgo asociado a estas entrevistas.

Seguimiento de formato estándar de archivos

Por ejemplo cuando nosotros cargamos un archivo **CSV** el archivo está separado por cada campo por una coma.

Pero piensa en lugares donde se separan los números con comas, entonces si nosotros separamos el archivo con este tipo de datos, los datos ya no tendrían una consistencia porque habría separaciones en donde no se quieren.

Por lo que es importante verificar la extensión de los datos y el formato interno que tienen dichos datos.

Tipos de datos

Verifica que las categorías de los datos correspondan con el contenido de cada categoría; es decir si la columna dice que son datos tipo **bool**, el contenido de toda la columna debería cumplir con ello.

Rango de las variables

El rango de una variable es el conjunto de valores que puede tomar. Por ejemplo, en la calificación de un examen el rango debe ser de **5 a 10**, pero imaginemos que observamos un **-2**, este es un ejemplo de que hay un dato fuera del rango establecido y puede que por alguna razón se salga. Así que para el caso de nuestro data set, debemos hacernos algunas preguntas como:

- ¿Cómo llegaron estos datos aquí?
- ¿Será que tienen algún significado alterno que aun no conozco?
- ¿Debo preservarlos?
- ¿Debo eliminarlos?
- Entre otras preguntas.

Unicidad

Con esto nos referimos a **¿qué tan únicos son nuestros datos?**.

Imaginemos que por cada pingüino nosotros establecemos una serie de datos, y por cada pingüino tenemos un **IDENTIFICADOR**.

Otro día observamos el mismo set de datos pero tenemos un caso de un **IDENTIFICADOR DUPLICADO**, pero con diferentes datos. Esto ya no tiene sentido porque habíamos acordado que solo debía existir un **IDENTIFICADOR** de pingüino por cada fila, entonces debemos manejar este tipo de casos.

Consistencia de expresiones

Con esto nos referimos a la manera en que las personas escriben o relatan la descripción de las variables o su comportamiento. Para este caso tenemos el ejemplo de las variables temporales: ¿cómo escribes la fecha en tu país?

- México: DD/MM/AAAA
- EEUU: MM/DD/AA
- Europa (la mayoría): DD/MM/AA
- Ásia (la mayoría): AA/MM/DD
- Perú MM/DD/AAA.

Entonces trata de tomar en cuenta esto para poder entender y tener consistencia en las expresiones.

Otro ejemplo: De acuerdo en nuestro dataset hay una categoría de **genero o sexo del pingüino**, entonces debería existir solo 2 opciones, así debemos fijarnos en la consistencia de las expresiones como:

- Nada de abreviaciones.
- Uniformidad en la escritura o formato.
  - Mayúsculas en un lado y minúsculas en otro.

Esto nos ayudará a que tenga consistencia nuestros datos y sigan un patrón asociado.

Valores nulos

Algunas veces van a estar explícitos y otras veces implícitos, ¿que quiere decir esto?

Es decir que nos vamos a dar cuenta que nuestros datos nulos **realmente faltan** en el conjunto de datos, es decir vamos a ver campos vacíos que denota que los datos no están.

En otros casos el dato no aparecerá en los registros, por consiguiente es un dato faltante que no se puede trackear o no se puede hacer nada con el. En el caso de toparme con esto debo preguntarme; ¿realmente tengo todos los datos? y si me falta un dato y ese campo esta vació debo preguntarme ¿por qué está vació?, ¿será que lo puedo rellenar con otros datos?,¿será que está vació por procesos aleatorios?,¿o tiene un sentido? Estas preguntas deberemos hacerlas, pra asegurar la consistencia de los datos.