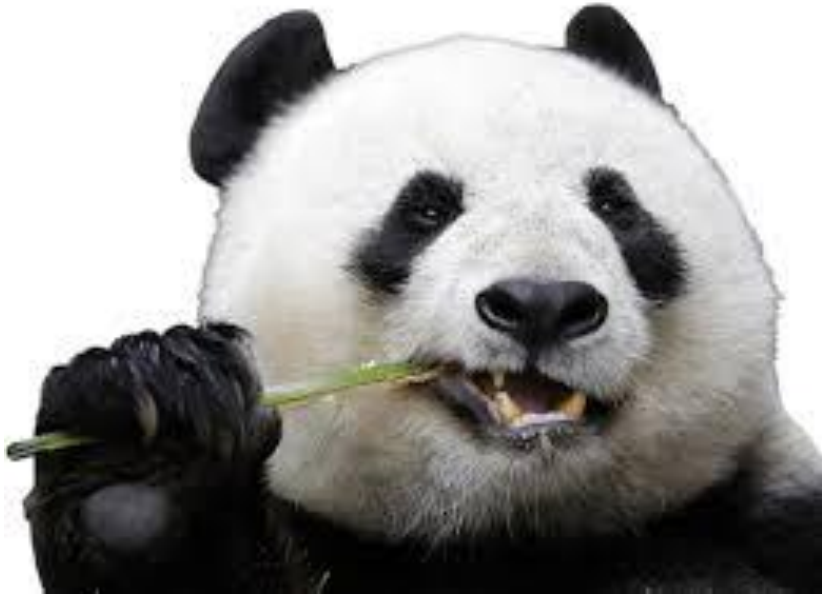
Project #2

Blanca Monreal

Ironhack - Data Analytics

Jun - Dec 2020



Mi semana:

1. **Recolectar información para dibujar una narración o una hipótesis sugerente que dinamizara el proyecto:**

- Buscar memes, efemérides y curiosidades sobre los tiburones (y sobre la película Tiburón ;)
- Informarme acerca del contexto de la tabla y el proceso de investigación de ataques de tiburones (aquí y aquí)
- Identificar otras investigaciones paralelas sobre ataques de tiburones que pudieran servir de reto

2. **Familiarizarme más con el proceso formal de EDA**

- Entender las fases y su orden para obligarme a ser metódica
- Relacionar las fases con los principales métodos de pandas
- Explicármelo como si tuviera 10 años

3. **Dominar ~~el mundo~~ Pandas**

- Pelearme con la teoría de date - time
- Acercarme a la lógica de la sintaxis (álgebra relacional)
- Ensayo - error

Preparación de los datos:

1. **Loading data**: Generate a DataFrame
    - Check the format of the original data (file or a python data structure?)
2. **Preview the data**: Getting familiar with the data
    - Check column names, index, data types, amount of null or nan, shape, sample, nunique - groupby
3. **Cleaning data**: Removing unnecessary or erroneous data
    - Assign column names
    - Define index
    - Manage missing values (nulls and nans)
    - Manage duplicate records
    - Clean values by column: Special characters, noise in strings, detect and manage the incorrect values, date/times format...
4. **Explore the data:** Generate visual insights
    - Visual exploration: Data visualisation
    - Statistical exploration: Math and statistics
5. **Draft investigation questions:** Design the investigation
6. **Reshape the df to the questions of your interest:**
    - Extreme values or outliers, drop - append columns and rows
7. **Transforming data formats**
    - Replace values (strings, parts of strings, apply a function) replace, map, l-rstrip, apply
    - Concatenating pandas series
    - Merge, pivot
    - Adding knowledge to your dataset using map function or apply
    - Discretizing continuous data, and finally about dummy variables and one-hot encoding.
8. **Publish your conclusions:** Reply your own questions
    - Plotting your clean data set (Histograms, Box plot, Scatter plot..)

Shark Attack EDA WorkFlow (part 1/3):

1. **Loading data**: He añadido “encoding” y “parse_dates” como parámetros
2. **Preview the data**: type, info, shape, head, sample, columns, count, index, nunique
3. **Clean the data**:
    - Re emmarcardo DF a re_shark (6 columns) y reshaped_sharks (drop row na) por la gran cantidad de nan y de columnas misteriosas
    - Nombres de columnas e índice dados por defecto correctos :)
    - Limpiar columnas: ¿Qué problemas hay y cómo queremos resolverlos?

        1. **Countries column**: groupby+count+duplicated.any, nunique, sort_values
            - Hay países agrupados repetidos → borrar las filas con países -no lo he conseguido como quería-
            - Hay mucho país con valores poco relevantes → nos quedamos con los top 3
        2. **Sex column:** groupby.count,
            - Hay varios sexos que necesitan ser unificados → quedarnos con 2 sexos, M y F
        3. **Type column**: describe(), unique(), groupby() & count()
            - Hay dos outliers → Desprendernos de los outliers
            - Hay dos tipos que podrían significar lo mismo (“boat” y “boating”) → Quedarnos con sólo un tipo “boat” / “boating”: cambiar el string de Type a “Boat” a “Boating” con un replace

Shark Attack EDA WorkFlow (part 2/3):

4. **Activity column:** value_counts, groupby() & count(), sort_values
    - En el top 15 hay actividades repetidas
    - Hay actividades

5. **Fatality column:** unique, count,
    - Hay valores con ruido (“ N”) → Cambiar éstas celdas para que se agrupen con “N”
    - Hay separados los valores nan de los unknowns; cuando nos hacen el mismo servicio → Fusionar unknowns y nan
    - Quedarnos con sólo Y, N y unknown

6. **Date column**:
    - Hay ruido en las celdas (texto random) → Limpiar el ruido con replace
    - El formato de los datos es tipo “object” → Convertir el formato de los datos a “date” -no lo he conseguido como quería-
    - Los formatos D - M - Y no están alineados: no todos tienen los mismos “ingredientes” ni formato → Alinear el formato a Mmm - YYYY
    - Hay valores antiguos que no nos interesan → Cargarnos los valores anteriores a 1950

Shark Attack EDA WorkFlow (part 3/3):

4. **Draft final investigation question:** Cuál es el perfil y bajo qué circunstancias hay probabilidades de morir?

7. **Publish your conclusions:**

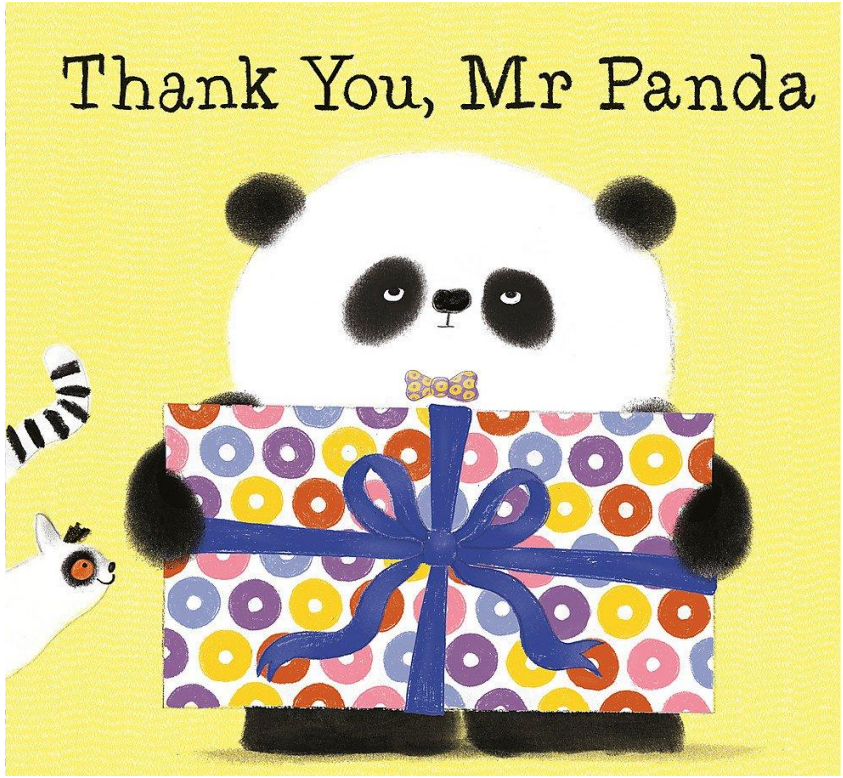
Cuál es el perfil y bajo qué circunstancias hay probabilidades de morir?

- **Type** : Unprovoked
- **Country :** AUSTRALIA
- **Activity :** Swimming
- **Sex :** M
- **Fatal (Y/N) :** Y

Bonus: Cuál es el perfil y bajo qué circunstancias hay más probabilidades de morir en España?

- **Type :** Unprovoked
- **Country :** SPAIN
- **Activity :** Bathing
- **Sex :** M
- **Fatal (Y/N)** : Y

Gracias!



Un aplauso para el señor de Kaggle
con un notebook precioso <3