Introducción:

En este análisis se enfatizó en saber cuantas especies están en peligro. Para lograr resolver este principal misterio, a medida que analizamos la información nos fueron surgiendo las siguientes preguntas:

1) ¿Qué tipos de especies hay?

2) ¿Cuál es el riesgo de extinción para cada grupo de especie (categorías)?

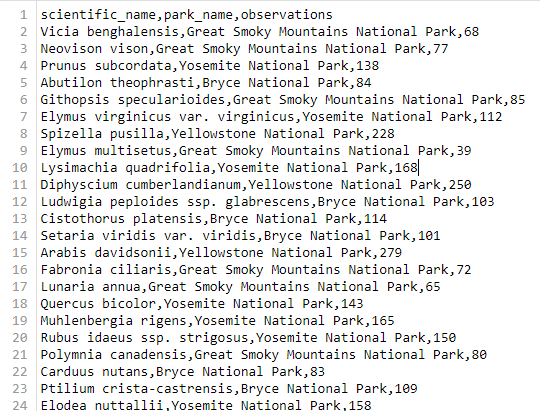
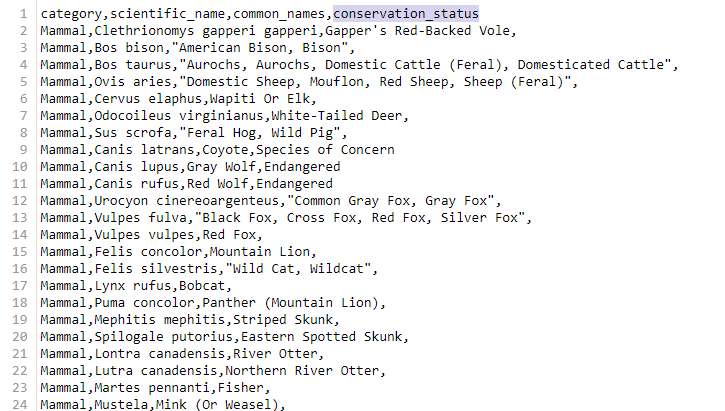
3) ¿Alguna categoría de especies se observava más?

4) ¿Está relacionada la cantidad de avistamientos de seres vivos con su propia calificación de riesgo de extinción?

5) ¿Fue la predicción de la IA lo suficientemente buena? analizando la proporción de peligro en las especies

Cuerpo:

Lo inicial fue reconocer qué datos teníamos frente a nosotros. Por un lado estaba el csv 'observations' y por otro lado estaba el csv 'species\_info' que contaban con la siguiente información:

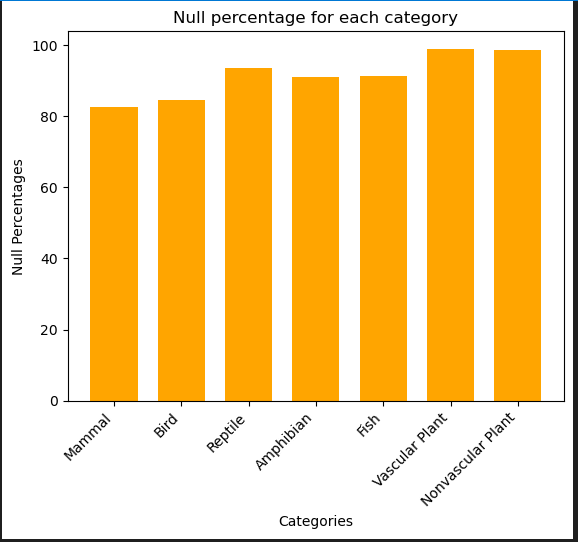


Nótese que tienen una columna en común que comunica a los archivos.

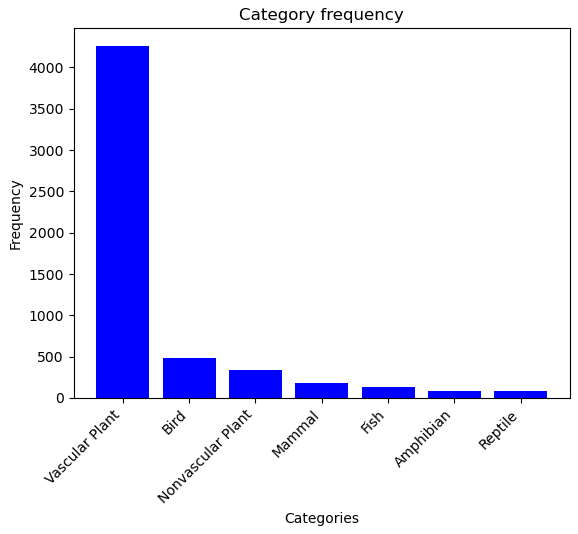
Primero iniciamos limpiando los datos, donde el csv 'observations' contenía valores duplicados, había una gran cantidad de especies por lo cual deberíamos enfocar nuestro estudio en base a categorías de especies. No habían nulos ni strings que no representen agregado de información, los tipos de datos eran los correctos, tampoco habían valores que se consideren atípicos y cabe aclarar que cada parque nacional contribuyó con la misma cantidad de filas, haciendo un total de 23296 registros ‘sucios’. Para solucionar los problemas simplemente eliminamos los datos duplicados, recortamos el final de ‘National Park’ en el nombre de cada parque porque era repetitivo y al final reseteamos los índices.

Seguido a esto, procedimos a limpiar los datos de nuestra segunda fuente de información de unas 5824 filas nada más, donde encontramos cerca de un 97% de columnas con el dato acerca del estado de la especie (en recuperación, especies preocupantes, amenazadas o en peligro de extinción), el cual era justo el de nuestro interés. Lo positivo es que no había información duplicada, o eso creíamos al inicio ya que al momento de analizar si se repetía el nombre de alguna especie (sin necesidad de que se repita toda la fila) nos dimos cuenta que al inicio estábamos equivocados. Al final nos decidimos en eliminar las filas que por lo menos repitan el nombre científico de la especie junto a su estado de riesgo y la categoría.

Si bien seguimos en la limpieza de la tabla de especies cambio de párrafo porque ahora vamos a hablar específicamente acerca de estos tantos valores nulos. Al ver los datos de manera aleatoria no veíamos ningún patrón, así que nos adentramos con nuestro primer análisis en busca de una relación entre el porcentaje de datos nulos por categoría:



Con este gráfico vimos que si bien los datos faltantes no estaban distribuídos de manera equitativa (por estar siempre en la misma columna) nos dimos cuenta que dentro de la misma columna parecían estar distribuídos equitativamente. El problema es que no estábamos seguros de esto último, motivo por el cual nos vimos forzados a comparar este gráfico con la cantidad de filas de información que teníamos de cada categoría de especies:

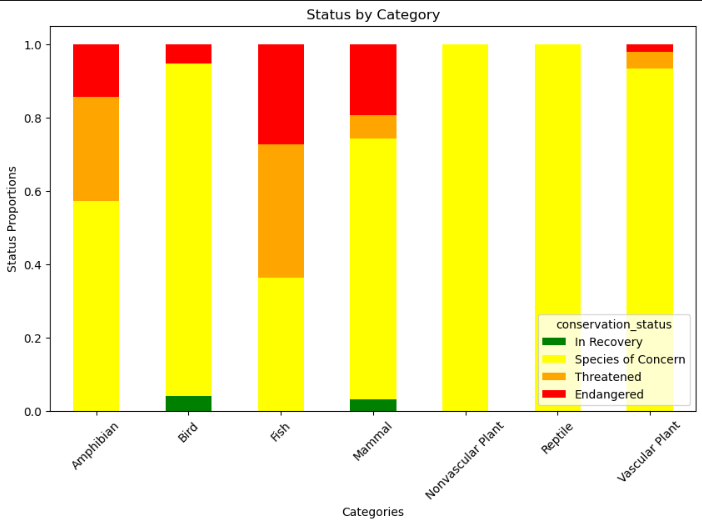


No es necesario ser un prodigio de las estadísticas para darnos cuenta que las plantas vasculares tienen demasiados más campos, pero tiene un porcentaje similar de datos nulos comparado al resto de tipos de seres vivos.

Parecía no haber un patrón para saber como rellenar ese 97% de datos faltantes, así que analizamos el tipo de dato nulo que teníamos. Debido a que solo una columna tenía nulos y distribuídos dentro de la misma se encontraban de manera aleatoria concluímos que estábamos ante un caso de MAR = ‘Missing At Random’. Esto era una buena noticia, ya que la columna de los datos incompletos era categórica (no tiene números, tiene textos clasificativos) podíamos usar un modelo predictivo en base a Inteligencia Artificial como última opción.

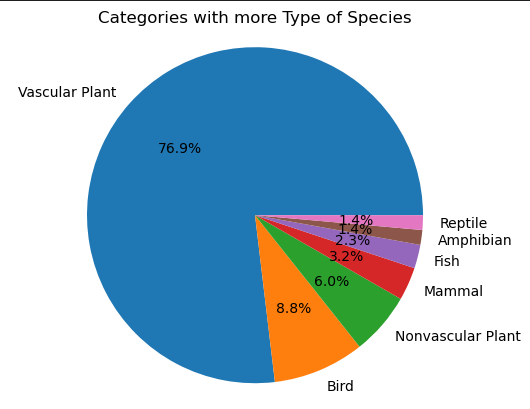
Previo a precipitarnos a una predicción con una muestra completa de tan solo el 3% era preferible buscar relación entre otras variables para ver si podíamos rellenar estos campos vacíos que tanto anhelábamos y para eso empezamos a respondernos algunas preguntas del inicio, mencionadas en la introducción, para ver si de a poco encontrábamos información de valor.

El siguiente análisis que realizamos con la información limpia (sin nulos) fue el de ver como se encontraba cada categoría de especies en relación al peligro:



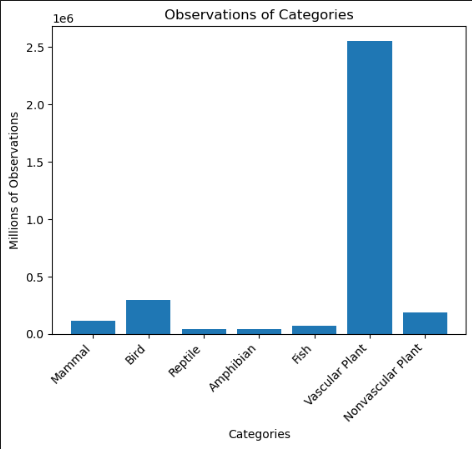
Este gráfico lo compararemos con varias variables diferentes para buscar esa tan deseada relación entre el peligro de cada categoría con algún otro dato. Dándonos a entender que los peces, mamíferos y anfíbeos son los más afectados, pero esto es con el 3% de la información de los parques, queremos saber qué es lo que en verdad está pasando.

Como primer candidato pensamos a la variedad de especies de cada categoría. Tal vez las categorías con más tipos diferentes de especies tenían mejores chances de sobrevivir, motivo por el cual calculamos qué categorías tenían una mayor variedad de especies. Recalco que todos los gráficos se hicieron con la información limpia sin contar los campos nulos, exceptuando los gráficos donde el título indica que es ‘Predictivo’:

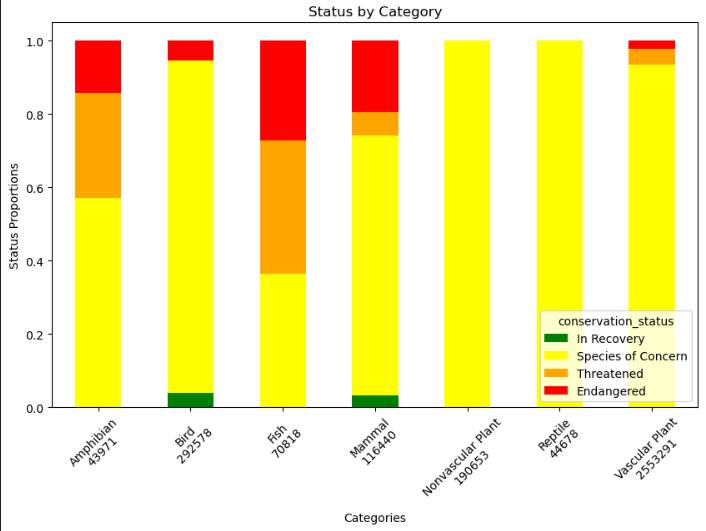


Nuestra conclusión al respecto fue rápida, los dos tipos de plantas y los reptiles eran los menos amenazados, pero las plantas vasculares eran las de mayor variedad de especies, los otros tipos de plantas se encontraban en el medio y los reptiles eran los menos diversificados cuando se habla de tener diferentes especies.

Sin cruzarse de brazos debíamos intentar con otras alternativas, y luego vino en cuestión la cantidad de observaciones. Tal vez las categorías que menos se ven en los parques eran las más amenazadas, motivo por el cual al inicio creamos un gráfico que revisa la cantidad de avistamientos que se hicieron de cada tipo de especies:

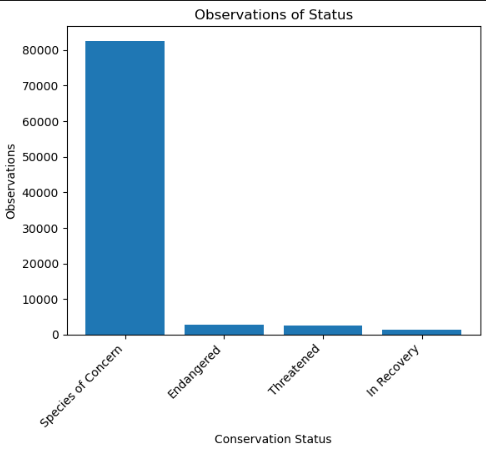


Como las plantas vasculares tienen un bajo riesgo y en este gráfico tenían una buena cantidad de avistamientos dentro de los parques por diferentes personas parecía que estábamos encontrando una correlación, así que nos mandamos en juntar un par de gráficos para comparar mejor la posible relación entre el riesgo de cada categoría y las observaciones de las mismas:



El gráfico es similar al que mostramos hace unas páginas atrás, solo que ahora debajo de cada categoría indicamos las observaciones respectivamente. Si nos enfocamos en los reptiles y las plantas no vasculares podemos notar que están en riesgos bajos y similares, pero una categoría es unas 4 veces más fácil de encontrar en la naturaleza, así que estábamos descartando otra hipótesis.

Al inicio dijimos que haríamos un análisis en base a las categorías porque teníamos muchas especies diferentes y sería confuso relacionar tanta información, pero como nos estábamos quedando sin alternativas razonables nos apuntamos a revisar las observaciones en los parques para cada especie en particular:

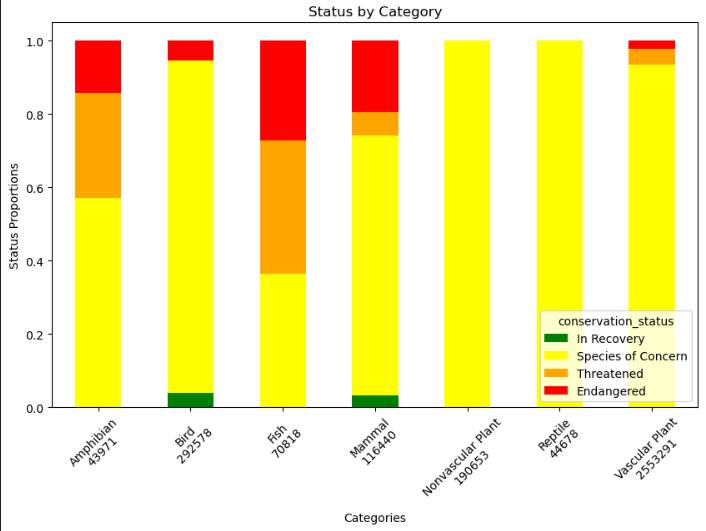


Este dato era llamativo, ya que indicaba que casi todas las especies que se ven se encuentran en un riesgo moderado de extinción clasificado como ‘especies preocupantes’. Como la mayoría de datos eran enormemente de este tipo, decidimos de una vez por todas ayudarnos con manipulación múltiple en base a IA, para que esta misma relacione todas las variables y tal vez llegase a la misma conclusión que nosotros.

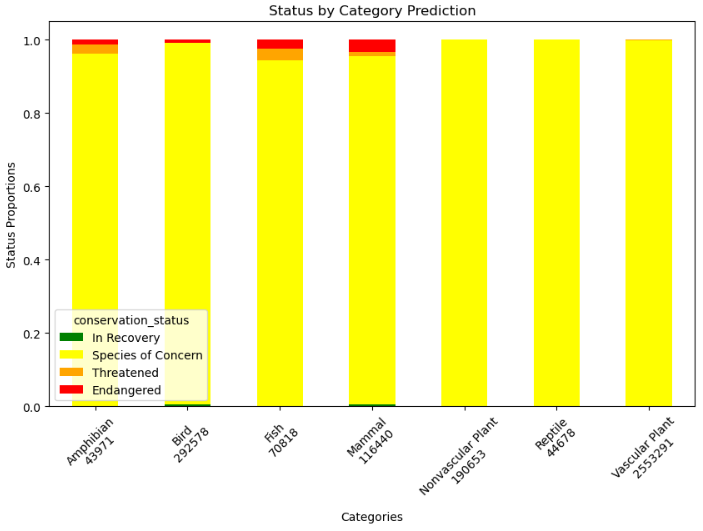
Esta última tarea nos hizo traducir todas nuestra bases de datos de especies a números, donde cáda número representa una categoría y con ella el modelo de aprendizaje automático podría sugerir elementos para agregar dentro de las casillas faltantes. Decirlo es más sencillo que hacerlo, en un momento dudamos si deberíamos pivotar la tabla haciendo que en lugar de tener una columna de nombre científico haya una columna por cada posible valor y poner un 1 o un 0 en cada fila dependiendo si coincide o no. Si bien pivotear las tablas es una técnica popular para complementarla con esta herramienta que queremos usar, al final nos dimos cuenta que generaríamos muchas columnas y optamos por tener un diccionario que relaciona cada diferente string con un número diferente, luego mandamos los registros sin datos nulos para entrenar la IA de la herramienta IterativeImputer.

Para nuestra suerte la IA vió el mismo patrón que nosotros, completando los datos faltantes de la columna ‘conservation\_status’ con el valor de ‘Species of Concern’. Lo bueno de usar este tipo de modelos es que no sesga los datos, lo malo es que se requiere tener nulos tipo MAR en variables categóricas (justo nuestro caso) donde la cantidad de registros con nulos sea menor al 10%, cuando nosotros teníamos la mayoría de nuestra información anulada.

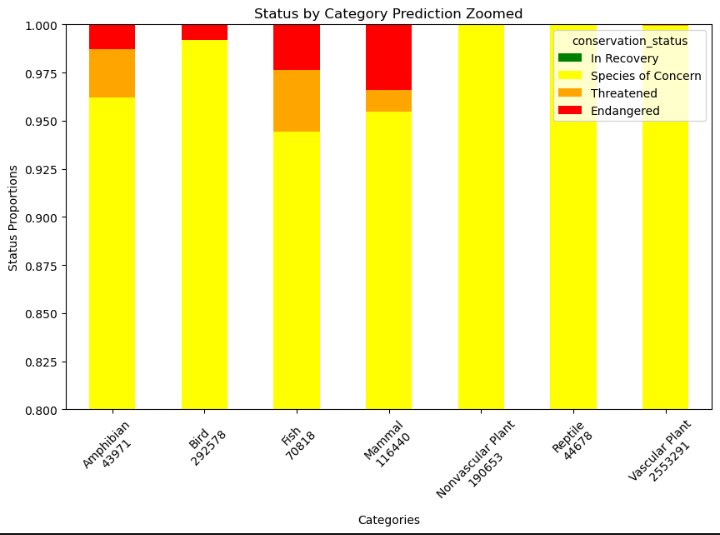
Si bien la precisión de completar la información de esta forma no es perfecta, es la más adecuada para esta situación, sugiriendo un nuevo gráfico para relacionar el riesgo con cada tipo de especie.



Este es el gráfico hecho con los datos no nulos.



Este es el gráfico estimativo



Este es el mismo gráfico que el anterior, pero enfocando en las proporciones más alarmantes.

No está demás volver a aclarar que estos últimos tres gráficos fueron los únicos que utilizaron la información completada en base a los patrones encontrados, lo cual resta precisión pero nos da una idea de qué podemos esperar.

Conclusión:

No hay una categoría de especies que esté en niveles alarmantes. Las plantas y los reptiles demuestran estar en una posición más segura, pero tenemos pocos datos para afirmar con seguridad lo dicho, por lo cual se recomienda una nueva base de datos para poder afirmar con una mayor seguridad la situación. Hemos intentado de muchas formas predecir la situación actual de las categorías de especies, si bien cada especie en particular podría saber en qué categoría se encuentra, no hemos hallado un fuerte vínculo entre el riesgo de extinción y otros datos como la diversidad de especies, la cantidad de seres vivos de cada categoría o la tendencia de riesgo de las observaciones de las especies a nivel general.

Apéndice:

Se resalta como se usó el modelo de manipulación múltiple, siendo ‘species’ un Data Frame de la librería ‘pandas’ de Python con los datos limpios pero conservando las filas nulas dentro de la columna ‘conservation\_status’.

#to use the IA we need to convert our DataFrames data types into numbers. We need de df\_test with no nulls and de original with all data

species\_reduced = species.drop('common\_names', axis=1)#we don't need this column

#I can pivote the table, but for each different name i would have a column, so i preffer to replace each name with a number

def make\_dictionary(column):

i=0

dictionary = {}

for name in species\_reduced[column].unique():

dictionary.update({name:i})

i+=1

return dictionary

category\_dicc = make\_dictionary('category')

name\_dicc = make\_dictionary('scientific\_name')

status\_dicc = make\_dictionary('conservation\_status')

mother\_dicc = {'category':category\_dicc, 'scientific\_name':name\_dicc, 'conservation\_status':status\_dicc}

species\_num\_original = species\_reduced#select interesting columns

for column in species\_num\_original.columns:

species\_num\_original[column] = species\_num\_original[column].apply(lambda name: mother\_dicc[column][name])

#all columns are int type, but we have to put again null values:

species\_num\_original['conservation\_status'] = species\_num\_original['conservation\_status'].apply(lambda num: None if num==0 else num)

species\_num\_test = species\_num\_original[~species\_num\_original.conservation\_status.isna()]

#let's generate the new data frame with complete entry

imp =IterativeImputer(max\_iter=1000, random\_state=0)

imp.fit(species\_num\_test)

species\_num\_complete =pd.DataFrame(np.round(imp.transform(species\_num\_original),0), columns=['category', 'scientific\_name', 'conservation\_status'])

#lets get the categories names again

def getKey(column, number):

for key, value in mother\_dicc[column].items():

if value == number:

return key

return 'error'

species\_completed = species\_num\_complete.copy()

for column in species\_num\_complete.columns:

species\_completed[column] = species\_num\_complete[column].apply(lambda value: getKey(column, value))

#resulted Data Frame called speices\_completed