VISUALIZACIÓN DE DATOS Y STORYTELLING

Análisis y Visualización de Datos Categóricos con Python

Tecnológico de Monterrey

Cuando trabajamos con **estadísticas**, es importante reconocer los **diferentes tipos de datos**: numéricos (discretos y continuos), categóricos y ordinales. Los datos no son más que observaciones del mundo en que vivimos, por tanto, los mismos pueden venir en diferentes formas, no solo numérica.

Por ejemplo, si le preguntáramos a nuestros amigos ¿cuántas mascotas tienen? nos podrían responder: 0, 1, 2, 4, 3, 8; esta información por sí misma puede ser útil, pero para nuestro análisis de mascotas, nos podría servir también otro tipo de información, como por ejemplo el género de cada uno de nuestros amigos; de esta forma obtendríamos la siguiente información: hombre, mujer, mujer, mujer, hombre, mujer. Como vemos, podemos incluir a los datos dentro de tres categorías fundamentales: datos cuantitativos o numéricos, **datos cualitativos o categóricos** y **datos ordinales**.

Datos cuantitativos

Los datos cuantitativos **son representados por números**; estos números van a ser significativos si representan la medida o la cantidad observada de cierta característica. Dentro de esta categoría podemos encontrar, por ejemplo: cantidades de dólares, cuentas, tamaños, número de empleados, y kilómetros por hora.

- Con los **datos cuantitativos**, se puede hacer todo tipo de tareas de procesamiento de datos numéricos, tales como sumarlos, calcular promedios, o medir su variabilidad.
- Asimismo, vamos a poder dividir a los datos cuantitativos en **discretos y continuos**, dependiendo de los valores potencialmente observables.

Revisemos a continuación cada uno de ellos:

- Los ***datos discretos*** solo van a poder asumir un valor de una lista de números específicos. Representan ítems que pueden ser contados; todos sus posibles valores pueden ser listados. Suele ser relativamente fácil trabajar con este tipo de dato.
- Los ***datos continuos*** representan mediciones; sus posibles valores no pueden ser contados y sólo pueden ser descritos usando intervalos en la recta de los números reales. Por ejemplo, la cantidad de kilómetros recorridos no puede ser medida con exactitud, puede ser que hayamos recorrido 1.7 km o 1.6987 km; en cualquier medida que tomemos del mundo real, siempre puede haber pequeñas o grandes variaciones.

Generalmente, los **datos continuos** se suelen redondear a un número fijo de decimales para facilitar su manipulación.

Datos cualitativos

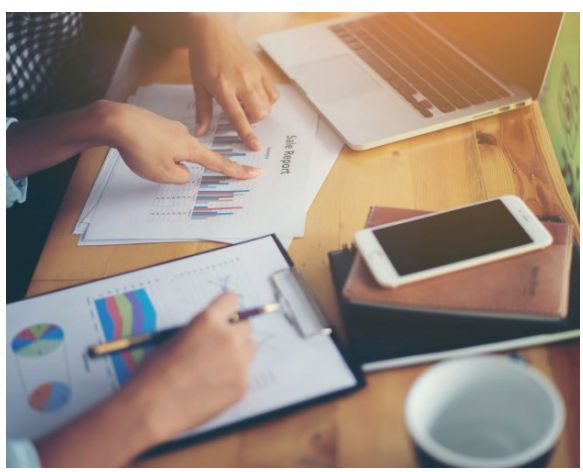
Si los datos nos dicen en cual de determinadas categorías no numéricas nuestros ítems van a caer, entonces estamos hablando de **datos cualitativos o categóricos**; ya que los mismos van a representar determinada cualidad que los ítems poseen. Dentro de esta categoría vamos a encontrar datos como: el género de una persona, el estado civil, la ciudad natal, o los tipos de películas que le gustan. Los datos categóricos pueden tomar valores numéricos (por ejemplo, “1” para indicar “masculino” y “2” para indicar “femenino”), pero esos números no tienen un sentido matemático.

Datos ordinales

Una **categoría intermedia** entre los dos tipos de datos anteriores, son los datos ordinales. En este tipo de datos, va a existir un orden significativo, vamos a poder clasificar un primero, segundo, tercero, etc. es decir, que podemos establecer un ranking para estos datos, el cual posiblemente luego tenga un rol importante en la etapa de análisis. Los datos se dividen en categorías, pero los números colocados en cada categoría tienen un significado. Por ejemplo, la **calificación de un restaurante** en una escala de 0 (bajo) a 5 (más alta) estrellas representa datos ordinales. **Los datos ordinales** son a menudo tratados como datos categóricos, en el sentido que se suelen agrupar y ordenar. Sin embargo, a diferencia de los datos categóricos, los números sí tienen un significado matemático.

Análisis de datos categóricos con Python

Para **ejemplificar el análisis**, vamos a utilizar nuestras habituales librerías científicas NumPy, Pandas, Matplotlib y Seaborn. También vamos a utilizar la librería pydataset, la cual nos facilita cargar los diferentes conjuntos de datos (dataset) para analizar.



La idea es realizar un **análisis estadístico** sobre los datos de los sobrevivientes a la tragedia del Titanic y poder visualizarlo.

La tragedia del Titanic

El **hundimiento del Titanic** es uno de los naufragios más infames de la historia. El 15 de abril de 1912, durante su viaje inaugural, el Titanic se hundió después de chocar con un iceberg, matando a miles de personas. Esta **tragedia sensacional** conmocionó a la comunidad internacional y condujo a mejores normas de seguridad aplicables a los buques. Una de las razones por las que el naufragio dio lugar a semejante cantidad de muertes fue que no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque hubo algún elemento de suerte involucrada en sobrevivir al hundimiento, algunos grupos de personas tenían más probabilidades de sobrevivir

que otros, como las mujeres, los niños y la clase alta.

El siguiente **dataset** proporciona información sobre el destino de los pasajeros en el viaje fatal del **trasatlántico Titanic**, que se resume de acuerdo con el **nivel económico (clase), el sexo, la edad y la supervivencia.**

```
# importando modulos necesarios
%matplotlib inline

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from pydataset import data

# parametros esteticos de seaborn
sns.set_palette("deep", desat=.6)
sns.set_context(rc={"figure.figsize": (8, 4)})

# importando dataset
titanic = data('titanic')
# ver primeros 10 registros
titanic.head(10)
```

El **fragmento de código** anterior daría como **resultado** los siguientes datos:

	class	age	sex	survived
1	1st class	adults	man	yes
2	1st class	adults	man	yes
3	1st class	adults	man	yes
4	1st class	adults	man	yes
5	1st class	adults	man	yes
6	1st class	adults	man	yes
7	1st class	adults	man	yes
8	1st class	adults	man	yes
9	1st class	adults	man	yes
10	1st class	adults	man	yes

El problema con datos como estos, y en general con la mayoría de las tablas de datos, es que nos presentan mucha información y no nos permiten ver que, es lo que realmente sucede o sucedió. Por tanto, se debería procesar de alguna manera para hacernos una imagen de lo que los datos realmente representan y nos quieren decir; y qué mejor manera para hacernos una imagen de algo que utilizar las **visualizaciones**. Una buena visualización de los datos puede revelar cosas que es probable que no podamos ver en una tabla de números y nos ayudará a pensar con claridad acerca de los patrones y relaciones que pueden estar escondidos en los datos. También nos va a ayudar a encontrar las características y patrones más importantes o los casos que son realmente excepcionales y no deberíamos de encontrar.

Tablas de frecuencia

Para hacernos una **imagen de los datos**, lo primero que tenemos que hacer es **agruparlos**. Al armar diferentes grupos nos vamos acercando a la comprensión de los datos. La idea es ir agrupar las cosas que parecen ir juntas, para poder ver cómo se distribuyen a través de las diferentes categorías. Para los datos categóricos, agrupar es fácil; simplemente debemos contar el número de ítems que corresponden a cada categoría y agruparlos. Una forma en la que podemos agrupar el conjunto de datos del Titanic es contando las diferentes clases de pasajeros. Podemos organizar estos conteos en una tabla de frecuencia, que registra los totales y los nombres de las categorías utilizando la función **value_counts** que nos proporciona la librería Pandas del siguiente modo:

```
# tabla de frecuencia de clases de pasajeros
pd.value_counts(titanic['class'])
```

El **código anterior** arrojaría el siguiente resultado:

```
3rd class    706
1st class    325
2nd class    285
dtype: int64
```

Contar la cantidad de apariciones de cada categoría puede ser útil, pero a veces puede resultar más útil saber la fracción o proporción de los datos de cada categoría, así que podríamos entonces dividir los recuentos por el total de casos para obtener los porcentajes que representa cada categoría.

Una tabla de frecuencia relativa muestra los porcentajes, en lugar de los recuentos de los valores en cada categoría. Ambos tipos de tablas muestran cómo los casos se distribuyen a través de las categorías. De esta manera, ellas describen la distribución de una variable categórica, ya que enumeran las posibles categorías y nos dicen con qué frecuencia se produce cada una de ellas.

```
# tabla de frecuencia relativa de pasajeros
100 * titanic['class'].value_counts() / len(titanic['class'])
```

El código anterior arrojaría el siguiente resultado:

```
3rd class    53.647416
1st class    24.696049
2nd class    21.656535
dtype: float64
```

Gráficos de pastel y barras

Ahora que ya conocemos a las **tablas de frecuencia** ya estamos en condiciones de crear visualizaciones que realmente nos den una imagen de los datos, sus propiedades y sus relaciones. En este punto, debemos ser sumamente cuidadosos, ya que una mala visualización puede llegar a distorsionar nuestra comprensión, en lugar de ayudarnos. Las **mejores visualizaciones de datos** siguen un principio fundamental llamado el **principio del área**.

Este principio nos dice que el área ocupada por cada parte del gráfico se debe corresponder con la magnitud del valor que representa. Violaciones del principio de área son una forma común de mentir con estadísticas. Dos gráficos útiles que podemos utilizar para representar nuestros datos y que cumplen con este principio son el **gráfico de barras y el gráfico de pastel**.

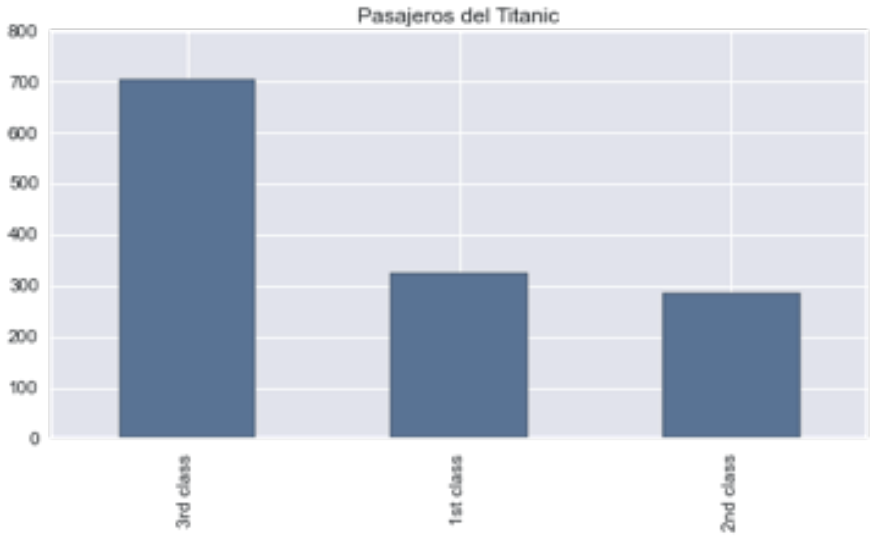
Gráfico de barras

- El **gráfico de barras** nos ayuda a darnos una impresión visual más precisa de la **distribución de nuestros datos**.
- La **altura de cada barra** muestra el recuento de su categoría. Las barras tienen el mismo ancho, por lo que sus alturas determinan sus áreas, y estas áreas son proporcionales a los recuentos en cada categoría.
- De esta forma, podemos ver fácilmente que había más del doble de pasajeros de tercera clase, que de primera o segunda clase.
- Los gráficos de barras hacen que este **tipo de comparaciones** sean fáciles y naturales.

Veamos cómo podemos crearlos de forma sencilla utilizando el método plot dentro de un DataFrame de la librería de Pandas.

```
# Gráfico de barras de pasajeros del Titanic
plot = titanic['class'].value_counts().plot(kind='bar',
                                            title='Pasajeros del Titanic')
```

El código anterior nos arrojaría el siguiente resultado:



Si quisiéramos enfocarnos en la ***proporción relativa*** de los pasajeros de cada una de las clases, simplemente podemos sustituir a los recuentos con porcentajes y utilizar un gráfico de barras de frecuencias relativas.

```
# gráfico de barras de frecuencias relativas.
plot = (100 * titanic['class'].value_counts() / len(titanic['class'])).
plot(
kind='bar', title='Pasajeros del Titanic %')
```

El **código anterior** nos arrojaría el siguiente resultado:

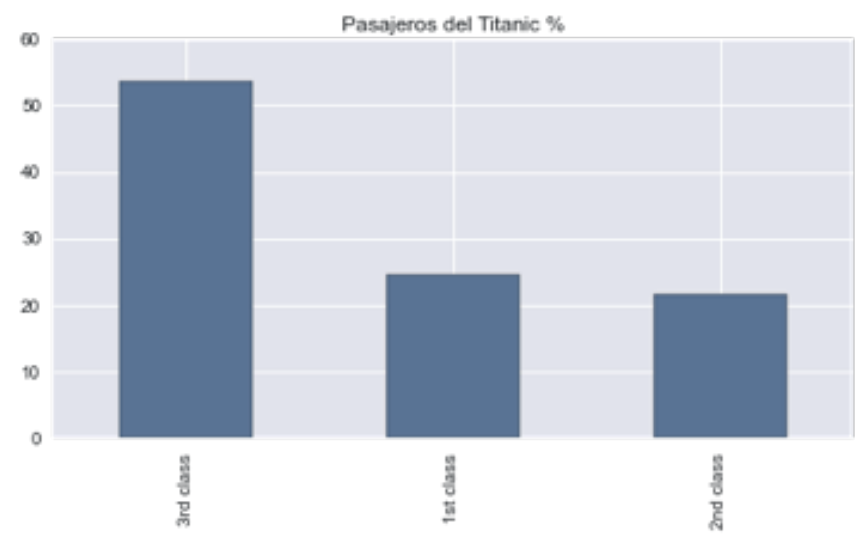


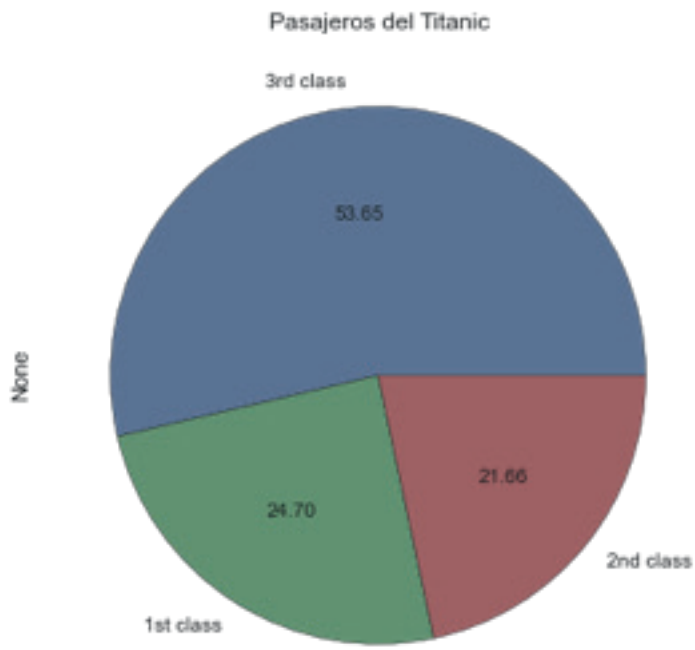
Gráfico de tartas

El gráfico de tarta muestra el total de casos como un círculo y luego corta este círculo en piezas cuyos tamaños son proporcionales a la fracción que cada categoría representa sobre el total de casos. Los gráfico de tarta dan una impresión rápida de cómo todo un grupo se divide en grupos más pequeños.

Lo podríamos graficar del siguiente modo, también utilizando el método *plot* de la librería de Pandas:

```
# Gráfico de tarta de pasajeros del Titanic
plot = titanic['class'].value_counts().plot(kind='pie', autopct='%.2f',
                                            figsize=(6, 6),
                                            title='Pasajeros del Titanic')
```

El código anterior nos arrojaría el siguiente resultado:



Como se puede apreciar, con el **gráfico de tarta** no es tan fácil determinar que los pasajeros de tercera clase son más que el doble que los de primera clase; tampoco es fácil determinar si hay más pasajeros de primera o de segunda clase. Para este **tipo de comparaciones**, son mucho más útiles los gráficos de barras.

Tecnológico de Monterrey