# 2. Vizualizace dat

## Zadání:

V jednom ze cvičení jste probírali práci s moduly pro vizualizaci dat. Mezi nejznámější moduly patří matplotlib (a jeho nadstavby jako seaborn), pillow, opencv, aj. Vyberte si nějakou zajímavou datovou sadu na webovém portále Kaggle a proveďte datovou analýzu datové sady. Využijte k tomu různé typy grafů a interpretujte je (minimálně alespoň 5 zajímavých grafů). Příklad interpretace: z datové sady pro počasí vyplynulo z liniového grafu, že v létě je vyšší rozptyl mezi minimální a maximální hodnotou teploty. Z jiného grafu vyplývá, že v létě je vyšší průměrná vlhkost vzduchu. Důvodem vyššího rozptylu může být absorpce záření vzduchem, který má v létě vyšší tepelnou kapacitu.

## Řešení:

Jako zdroj dat jsem využil Kaggle a jejich obsáhlou databázi datasetů. Dataset který jsem se nakonec rozhodl využít „[Spotify and Youtube](https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube)“ dataset. Dataset obsahuje 20000 unikátnách songů. Sloupců je dohromady 28 s informacemi které se týkají, počtu zhlédnutí a streamů, generované sloupce vypovídající o jejich vlastnostech a dalších. Zajímalo mě hned několik témat mezi nimi bylo i využití generovaných sloupců ke kategorizaci.

Na vizualizaci dat jsem využil matplotlib a pandas. Vytvořil jsem 5 grafů (2 další jsem bral jako nedokonalé a rozhodl jsem se je vynechat). Program začíná načtením dat, přičemž následují jednotlivé funkce, které definují jednotlivé grafy. Za každou funkcí se zavolají. To mi sloužilo hlavně k rychlému debuggingu, abych pouze ukazoval grafy, na kterých zrovna pracuji.

Obsah obrázku text, diagram, snímek obrazovky, design

Popis byl vytvořen automaticky

Z grafu lze vyčíst, že většina songů v datasetu je zařazená v albech poté v menším počtu následují singly a ve velicé malé míře jsou to jen kompilace.

A screen shot of a computer program

Description automatically generated

Obsah obrázku snímek obrazovky, text, diagram, kruh

Popis byl vytvořen automaticky

Z grafu lze vyčíst pár kategorií do kterých jsem songy zařadil. Kvůli nejednotnosti, jestli song je určité kategorie, jsem za pomocí této funkce zjišťoval přes prahy, jestli je daný song určitého typu, jelikož sloupce obsahují hodnoty „jistoty“ která je v rozmezí od 0-1.

Mnoho songů je nezařazených, ale určitě pokud by se pohrálo správně s prahy, tak by jich mohlo být o dost méně a však by to bylo za cenu jistoty, že song je určitého typu.

Obsah obrázku text, snímek obrazovky, Písmo

Popis byl vytvořen automatickyA screen shot of a computer code

Description automatically generated

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

Graf, který ukazuje průměrný zhlednutí a počet streamů s viditelným podílem Spotify vs YouTube. Lze vidět, že méně známé songy se vice poslouchají na Spotify mezitím co ty známější jsou více a více spouštěny na YouTube.

A screen shot of a computer code

Description automatically generated

Obsah obrázku text, snímek obrazovky, řada/pruh, Vykreslený graf

Popis byl vytvořen automatickyObsah obrázku text, snímek obrazovky, řada/pruh, diagram

Popis byl vytvořen automaticky

Graf, na kterém jsem chtěl ukázat podíl zhlédnutí a lajků které song na YouTube dostal. Jde vidět, jak absurdně malý podíl to je (podíl je lajků na: 0.6993400313173325 %). I přes malou čitelnost grafu jsem se ho rozhodl ponechat, aby byla vidět rozdíl mezi tím kolik lidí video zhlédlo a kolik jich jej lajknulo

A computer code on a black background

Description automatically generated

Obsah obrázku text, snímek obrazovky, číslo, Vykreslený graf

Popis byl vytvořen automaticky

Tímto grafem jsem chtěl zjistit, jestli je nějaká korelace mezi délkou songů a popularity. Lze vidět, že na délce songu zas tolik nezáleží, ale lze vidět mírné výkyvy u méně známějších písniček

A computer code on a black background

Description automatically generated

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

U tohohle grafu mě zajímalo, jestli hlučnost autoři použili ke generování měřítka energetičnosti. Lze vidět nějaká korelace mezi těmito dvěma vlastnostmi, přesto lze vidět spoustu výjimek jako například obrovský nárůst a pokles v půlce grafu

A screen shot of a computer code

Description automatically generated