DATA CLEANING AND PREPARATION

CHEATSHEET FOR LESSON 3

MISSING VALUES

- **Reasons we could have missing values in our dataset:**
  There are three main reasons that will change how we deal with the problem:
  - Missing Completely at Random (MCAR) Example: The dataset was corrupted during a transfer and some data went missing.
  - Missing at Random (MAR) Example: A dataset of car information, if we sample on old cars then we would have a lot more missing values than on brand new cars.
  - Missing not at Random (MNAR) Example: 2016 US presidential election results where the predictions were wrong because people were more hesitant to fill out fields admitting their support of Donald Trump.

- **Dealing with missing values:**

  i. Listwise Deletion :
  We delete every row that has at least one missing value.
  In python: df.dropna(inplace=True)

  ii. Dropping Variables:
  Usually applicable if a variable has more than 60% of missing values and the variable isn't essential for the next steps.
  In python: df=df.drop(['Name_of_variable'], axis=1)

  iii. Imputation: Mean, Median and Mode
  Replacing the missing values of a variable with the mean, Median or mode when the variable is numerical of course.
  In python :
  df["Name_of_variable"].fillna(df["Name_of_variable"].mean(), inplace=True)

IMBALANCED DATA

- Arises in classification tasks.
- We say we have imbalanced data if one or more classes have significantly less data points than the majority classes.
- This could be caused by the data collection process.
- **Majority Class:** The classes with a lot of examples.
- **Minority Class:** The class with a relatively small number of examples.
- Could be solved by data augmentation

DATA CLEANING AND PREPARATION

CHEATSHEET FOR LESSON 3

Example: we are trying to build a cats/dogs image classifier and we have a dataset of 1000 dog images and 10 cat images.

OUTLIERS

In this case the data isn't missing but we have some abnormal values.

- **Types of outliers:**
  a. Global Outliers:
     Global outliers are points in the dataset that are far from all the other points.
  b. Contextual Outliers:
     Values that significantly deviate from the rest of the data points in the same context.
     Example: A dataset of yearly raspatory issues deaths with a change of context because of the 2020 pandemic.
  c. Collective Outliers:
     A group of outliers that are close to each other. This is harder to detect than the previous two cases.

- **Detecting outliers:**
  a. Z-Score:
     - The z-score of an observation is a measurement of how far the data point is from the mean.
     - Formula: $z = (x - \mu) / \sigma$ where μ is the mean and σ is the standard deviation.
     - Usually we pick 2.5, 3, 3.5 or more standard deviations.
     - In python:

```
df['sales_zscore']=(df.NA_Sales-
df.NA_Sales.mean() ) / df.NA_Sales.std()
df=df[df['sales_zscore']>3]
```

CATEGORICAL DATA ENCODING

We have to adapt our variables to machine learning algorithms that only accept numeric values.

- **One Hot Encoding**

  The non-numerical variable will be converted to a numerical variable by mapping each unique label to a binary vector; the method is called one hot because for each vector only one value is non-null (hot) and the others are all zeros.

  In python: `y = pd.get_dummies(df.Genre, prefix='Genre')`