3 types of outlier:

1. **Global**
   Values that are completely different from the overall data group
2. **Contextual**
   Normal data points under certain conditions but become anomalies under other conditions ( common in time-data series)
3. **Collective**
   Group of abnormal points that follow similar patterns and are isolated from the rest of the populations

Ways to deal with outliers:

1. **Delete**
   If you sure when the outliers are mistake, typos or errors
2. **Reassign**
   If the dataset is small and/or it will be used for modelling or machine learning.
   Common ways to reassign:
   - Create a floor and ceiling at a quantile
   - Impute the mean or average
3. **Leave**
   If the dataset only will be used for analysis, EDA and nothing else or the dataset is resistant to outliers

**Common threshold for outliers**

$$Upper\ Limit = Third\ Quartile + 1.5\ \times Interquartile\ Range$$

$$Lower\ Limit = First\ Quartile - 1.5\ \times Interquartile\ Range$$

**Categorial data**

- Data that uses words or quality rather than number
- Many data models and algorithms don't work well with categorical data
- Common ways to change categorial data to numerical:
  - **Dummy variables**
    Values of 0 or 1
  - **Label encoding**
    Each category is assigned with a unique number
    The data will be simpler to clean, join, group, takes less storage and algorithm/model will typically runs smoother
    Suitable for large datasets
  - **One hot encoding**
    Each category is represented by 0 or 1
    Suitable for small datasets

Common label encoding python functions:
df.astype()
.cat.codes
pd.get_dummies()
LabelEncoder() (scikit-learn.preprocessing)

**Input Validation**
The practice of thoroughly analyzing and double-checking to make sure data is complete, error-free, and high-quality.