Data Science
Unit 3
Assess Superstep:

- **Data quality:**
- refers to the condition of a set of qualitative or quantitative variables.
- a multidimensional measurement of the acceptability of specific data sets.
- measured to determine whether data can be used as a basis for reliable intelligence extraction for supporting organizational decisions.
- **Data profiling** involves observing in your data sources all the viewpoints that the information offers.
- The main goal is to determine if individual viewpoints are accurate and complete.

- **1) Errors**
- Accept the Error: If it falls within an acceptable standard.
- if you accept the error, you will affect data science techniques and algorithms that perform classification.
- Reject the Error: Removing data is a last resort. add a quality flag and use this flag to avoid this erroneous data being used in data science techniques and algorithms that it will negatively affect.
- Correct the Error: Spelling mistakes in customer names, addresses, and locations are a common source of errors, which are methodically corrected.
- Create a Default Value: system developers assume that if the business doesn't enter the value, they should enter a default value. you discuss default values with your customer in detail and agree on an official "missing data" value.

- **2) Analysis of Data:**
- six data quality dimensions:
- 1. Completeness: completeness is specific to the business area of the data you are processing.
- For example, for personal data to be unique, you need, as a minimum, a first name, last name, and date of birth. If any of this information is not part of the data, it is an incomplete personal data entry.
- 2. Uniqueness: evaluate how unique the specific value is, in comparison to the rest of the data in that field.
- 3. Timeliness: Record the impact of the date and time on the data source. Are there periods of stability or instability?
- 4. Validity: Validity is tested against known and approved standards. It is recorded as a percentage of nonconformance against the standard.
- 5. Accuracy: Accuracy is a measure of the data against the real-world person or object that is recorded in the data source.
- 6. Consistency: This measure is recorded as the shift in the patterns in the data. Measure how data changes load after load.

- **3) Practical Actions:**
- Missing Values in Pandas:
- four basic processing concepts:
- 1. Drop the Columns Where All Elements Are Missing Values:
- (Assess-Good-Bad-01.py)
- Import sys
- import os
- import pandas as pd
- Base='C:/VKHCG'
- sInputFileName='Good-or-Bad.csv'
- sOutputFileName='Good-or-Bad-01.csv'
- Company='01-Vermeulen'
- sFileDir=Base + '/' + Company + '/02-Assess/01-EDS/02-Python'
- if not os.path.exists(sFileDir):
- os.makedirs(sFileDir)
- sFileName=Base + '/' + Company + '/00-RawData/' + sInputFileName
- print('Loading:',sFileName)
- RawData=pd.read_csv(sFileName,header=0)
- print(RawData)
- print('Rows:',RawData.shape[0])
- print('Columns:',RawData.shape[1])
- sFileName=sFileDir + '/' + sInputFileName
- RawData.to_csv(sFileName, index = False)
- TestData=RawData.dropna(axis=1, how='all')
- print(TestData)
- print('Rows:',TestData.shape[0])
- print('Columns:',TestData.shape[1])
- sFileName=sFileDir + '/' + sOutputFileName
- TestData.to_csv(sFileName, index = False)
-
- 2. Drop the Columns Where Any of the Elements Is Missing Values:
- Assess-Good-Bad-02
- TestData=RawData.dropna(axis=1, how='any')
-
- 3. Keep Only the Rows That Contain a Maximum of Two Missing Values:
- Assess-Good-Bad-03.py
- TestData=RawData.dropna(thresh=2)
-
- 4. Fill All Missing Values with the Mean, Median, Mode, Minimum, and Maximum of the Particular Numeric Column:
- Assess-Good-Bad-04.py
- TestData=RawData.fillna(RawData.mean())
- TestData=RawData.fillna(RawData.median())
- TestData=RawData.fillna(RawData.mode())
- TestData=RawData.fillna(RawData.min())
- TestData=RawData.fillna(RawData.max())