**Esikatselu**

Kolme satunnaista riviä dataframesta:

Id Entity Code Year Democracy score

2529 Turkey TUR 2008 5.69

1276 Jamaica JAM 2012 7.39

1599 Mauritania MRT 2017 3.82

Kolumnien määrä: 4

Entity (Maa)

Code (Maakoodi)

Year (Vuosi)

Democracy score (Demokraattinen pisteytys)

Kolumnien tietotyypit:

Entity object

Code object

Year int64

Democracy score float64

Rivien määrä: 2765

Uniikkien maiden määrä poissulkien manner-alueet: 168

Maailmassa on 195 maata, joten dataframesta puuttuu tieto 27 maasta. (13.8%)

**Puuttuvan datan käsittely**

Entity, Year tai Democracy scoressa täyttöaste on 100%. Ei tyhjiä arvoja (null)

Code osiossa on 96 tyhjää arvoa (null) ja ne johtuvat siitä, että manner-alueilla ei ole maakoodia.

Seuraavat manner-alueet on sisällytetty dataframeen:

Europe, Africa, Oceania, Asia, South America, North America

Tuloksia on lisätty näille manner-alueille vuosilta [2006, 2008 ja 2010-2023]

Kun alamme laskea keskiarvoja maiden välillä ja visuaalisoimaan dataa, meidän tulee poistaa manner-alueet dataframesta, sillä ne väärentävät laskutoimituksia myöhemmässä vaiheessa.

Visuaalisoimme tämän poistetun datan erikseen myöhemmin.

***Dublikaatit***

Duplikaatteja ei löydy. Kävin sen läpi seuraavalla komennolla:

# Select relevant columns: 'Entity', 'Year', 'Code', and 'Democracy score'

selected\_columns = df[['Entity', 'Year', 'Code', 'Democracy score']]

# Check for duplicate rows based on these columns

duplicates = selected\_columns[selected\_columns.duplicated()]

**Datan jakaminen ämpäreihin (bins)**

Käytin cut työkalua ja jaoin datan seuraaviin ämpäreihin:

bins = [0, 3, 5, 7, 9, 10]

labels = ['Very Low', 'Low', 'Moderate', 'High', 'Very High']

**Ääripäät datassa**

**Entity Year Democracy score Democracy score Binned**

**Afghanistan 2022 0.320000 Very Low**

**Afghanistan 2021 0.320000 Very Low**

**Norway 2015 9.930000 Very High**

**Norway 2016 9.930000 Very High**

**Sotke vähän dataa ja korjaa se**

df.loc[df['Entity'] == 'North Korea', 'Democracy score'] = 15.0

**Entity Year Democracy score Democracy score Binned**

Norway 2013 9.930000 Very High

Norway 2014 9.930000 Very High

--

--

North Korea 2014 15.000000 NaN

North Korea 2011 15.000000 NaN

North Korea 2012 15.000000 NaN

North Korea 2015 15.000000 NaN

A graph of a bar chart

Description automatically generated with medium confidence

Nyt jokin ei täsmää. Puutteelliselle tai pilalle menneelle datalle tulee tehdä jotain.

**Datan poistaminen**

* Pohjois-Koreasta on vain 15 riviä dataframessa ja sen osuus on 0,0039%. Voisimme poistaa Pohjois-Korean kokonaan, sen vaikuttamatta suuresti dataframeen.

**Datan arvojen muuttaminen keskiarvoksi**

* Jos Pohjois-Koreaa ei voi poistaa dataframesta, voisimme muuttaa sen arvon keskiarvoksi.

**Datan arvojen eriyttäminen**

* Voimme erityttää Pohjois-Korean arvot erilliseen datasettiin.

**Datan arvojen muuttaminen tunnetuksi arvoksi**

* Tässä tapauksessa tiedämme, että Pohjois-Korean arvoja on manipuloitu.
* Käyn poistamassa aiemman muutoksen, sekä Pohjois-Korean arvo palaa alkuperäiseksi.

**Perustilastot Numphy**

Mean of Democracy Score: 5.480996332783815

Median of Democracy Score: 5.76

Standard Deviation of Democracy Score: 2.2139269532165926

Variance of Democracy Score: 4.901472554178904

Minimum Democracy Score: 0.32

Maximum Democracy Score: 9.93

Sum of Democracy Scores: 14628.779212200001

Product of Democracy Scores: inf