Tehtävä: Opiskelijaryhmän tiedot

Students:

Adham Naderi - AB8911

Onni Roivas - AB0410

Tehtävä 1: Ongelman kuvaaminen (Business Understanding)

Haetaan vastaukset kysymyksiin:

- Millainen data on kyseessä?
- Missä yhteydessä tästä datasta voi olla hyötyä?
- Millainen liiketoiminta on datan taustalla?
- Mitä kaikkea datasta voidaan oppia?
- Mitä datasta voidaan nähdä?
- Mitä datasta voidaan tunnistaa?
- Voidaanko datasta havaita jotain poikkeavaa?
- Tarvitaanko tämän lisäksi jotakin muuta lisätietoa?

Palautetaan vastaus repositoryyn annettuun palautuspäivämäärään mennessä:

- Palautus tehdään MarkDown-formaatissa (tarkenne .md), jolloin kuvat ja kaaviot linkitetään dokumenttiin.

Tehtävä 1: Vastaukset

Millainen data on kyseessä:

Kyseessä olisi veden käyttöön/kulutukseen liittyvää tietoa missä näemme, että tietoa on kerätty neljän vuoden ajalta elikkä 2012 - 2015. Todennäköisemmin tietoa on kerätty päivittäin joka kuukausi jos datan rakenteen ottaa huomioon.

Missä yhteydessä tästä datasta voi olla hyötyä?

Dataa on hyötyä monessa erillaisissa asioissa kuten:

Veden kulutuksen analysoinissa, jossa katsotaan veden kulutuksen tiettynä aikana, kuinka paljon sitä käytetään ja miten sitä käytetään.

Ennakoivassa analytiikassa, jossa voidaan nähdä ennustavia näkymiä kuinka paljon vettä käytetään jonka avulla voidaan ennakoida

Laskutuksessa, jossa yhtiö tai kiinteistön käyttäjä voivat katsoa miten paljon kulutusta on sekä mitä se kustantaa.

Näin akillisesti tämmöisiä asioita voisi tehdä mutta varmasti on paljon erillaisia keinoja myös käyttää dataa tiedon keruun.

Millainen liiketoiminta on datan taustalla?

Kyseessä on hyvin todennäköisesti jonkinlainen ravintola tai yksiö/perhe joka käyttää kiinteistöä sillä dataa ja kuvaa katsoessa huomaa seuraavat asiat:

01.01.2012 - 23.08.2013 ei ole ollut minkäänlaista toimintaa

23.08.2013-24.02.2015 veden kulutusta on ollut

24.02.2015 eteenpäin kiinteistössä ei ole taas minkäänlaista veden kulutusta.

Elikkä todennäköisesti liiketila/kiinteistö on ollut tyhjillä reilun vuoden ja sitten siinä on ollut yrittäjä/asukas puolitoista vuotta, tämän jälkeen se on ollut taas tyhillään.

Mitä datasta voidaan nähdä? Mitä datasta voidaan tunnistaa?

Datasta voidaan nähdä asiat mitkä mainitsin aiemmin elikkä veden kulutuksen analysointi ja ennakoivaa analytiikka elikkä toisin sanoen miten paljon vettä käytetään, miten käytetään ja milloin.

Voidaanko datasta havaita jotain poikkeavaa?

Datassa on kolme poikkeaavaa asiaa jotka ovat graaffissa olevat kolme piikkiä joka tarkoittaa silloin veden kulutuksen määrä on ollut jostain syystä paljon korkeampi verrattuna. Aikavälit jolloin piikit näkyvät:

12.10.2023

30.04.2014

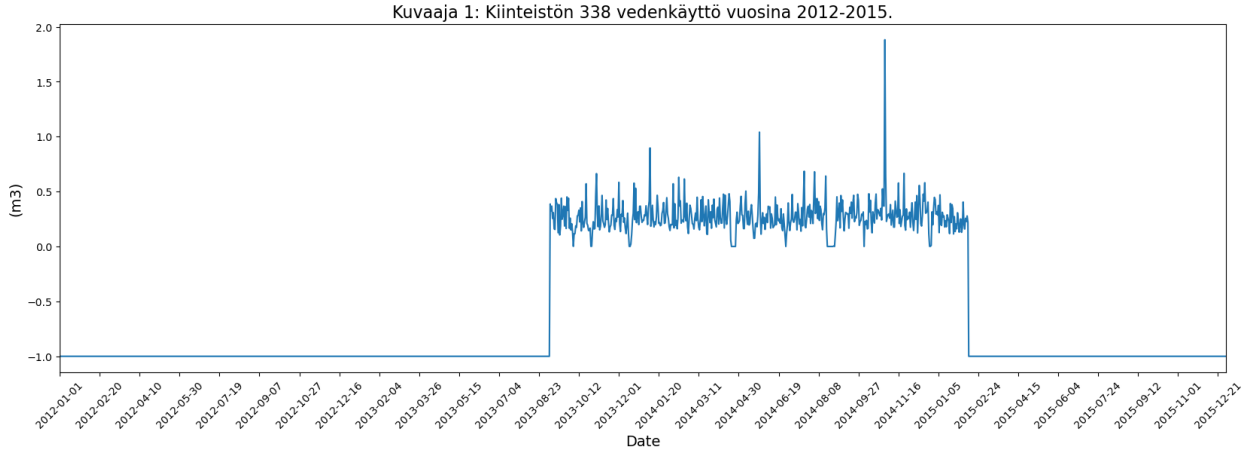
27.09.2014

Tarvitaanko tämän lisäksi jotakin muuta lisätietoa?

Varmaan ainoa lisätieto mikä olisi järkevää tietää on milloin dataa on kerätty esimerkiksi onko se päivittän kerätty vai viikottain tai kahden viikon välein. Sillä datassa huomasin sen et joskus oli 11 päivän velein ja joskus 19 päivän välein joka taas olisi hyvä tietää miksi. Mielestäni olisi hyvä asettaa yksinkertainen aika jolloin aina dataa kerätään koska se vaikuttaa analyysin tarkkuuteen ja kuvioiden ja poikkeavuuksien havaitsemiseen.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
df = pd.read_csv("Yorkshire_external_meters_2012_2015.csv",
index_col=0)

#Kuvaaja 1.1

kiinteisto = '338'  # vaihdetaan toinen kiinteistö

fig1, ax = plt.subplots(figsize=(20, 6))
ax.plot(df['date'], df[kiinteisto])
ax.set_ylabel('(m3)', size=14)
ax.set_xlabel('Date', size=14)
ax.set_xlim('2012-01-01', '2015-12-31')
ax.set_title('Kuvaaja 1: Kiinteistön ' + kiinteisto + ' vedenkäyttö
vuosina 2012-2015.', size=16)
ax.xaxis.set_major_locator(plt.MaxNLocator(30))
ax.tick_params(axis='x', rotation=45)

plt.show()
```



Tehtävä 2: Datan kuvaaminen (Data Understanding)

- Millaisia muuttujia on datassa?
- Mitä arvoa on käytetty täytearvona?
- Millaisia korrelaatioita datasta löytyy?
- Tarkistelee dataa tietyillä aikaväleillä (esim. viikko, kuukausi tai vuosi).
- Havaintojen tueksi voi liittää visualisoitua dataa
- Mieti mitä tällä datalla voidaan mahdollisesti tehdä?

Palautetaan vastaus repositoryyn annettuun palautuspäivämäärään mennessä:

- Palautus tehdään MarkDown-formaatissa (tarkenne .md), jolloin kuvat linkitetään dokumenttiin

Tehtävä 2: Vastaukset

Millaisia muuttujia on datassa?

Data sisältää useita muuttujia, joita ovat ainakin 'date', '1' - '2155'. Muuttujat ovat kokonaislukuja tai niiden yhdistelmiä. 'date' kuvaa päivämäärää, ja muut muuttujat ovat kiinteistöjen vedenkulutusta. Numeroiden '1', '2', '3', jotka ovat kiinteistöjen tunnisteita tai indeksejä.

Mitä arvoa on käytetty täytearvona?

Datassa uskoisin on käytetty täytearvona -1.0 joka merkitsee todennäköisesti minusta sitä et niinä tiettyinä päivinä dataa ei ole kerätty kiinteistöstä.

Millaisia korrelaatioita datasta löytyy?

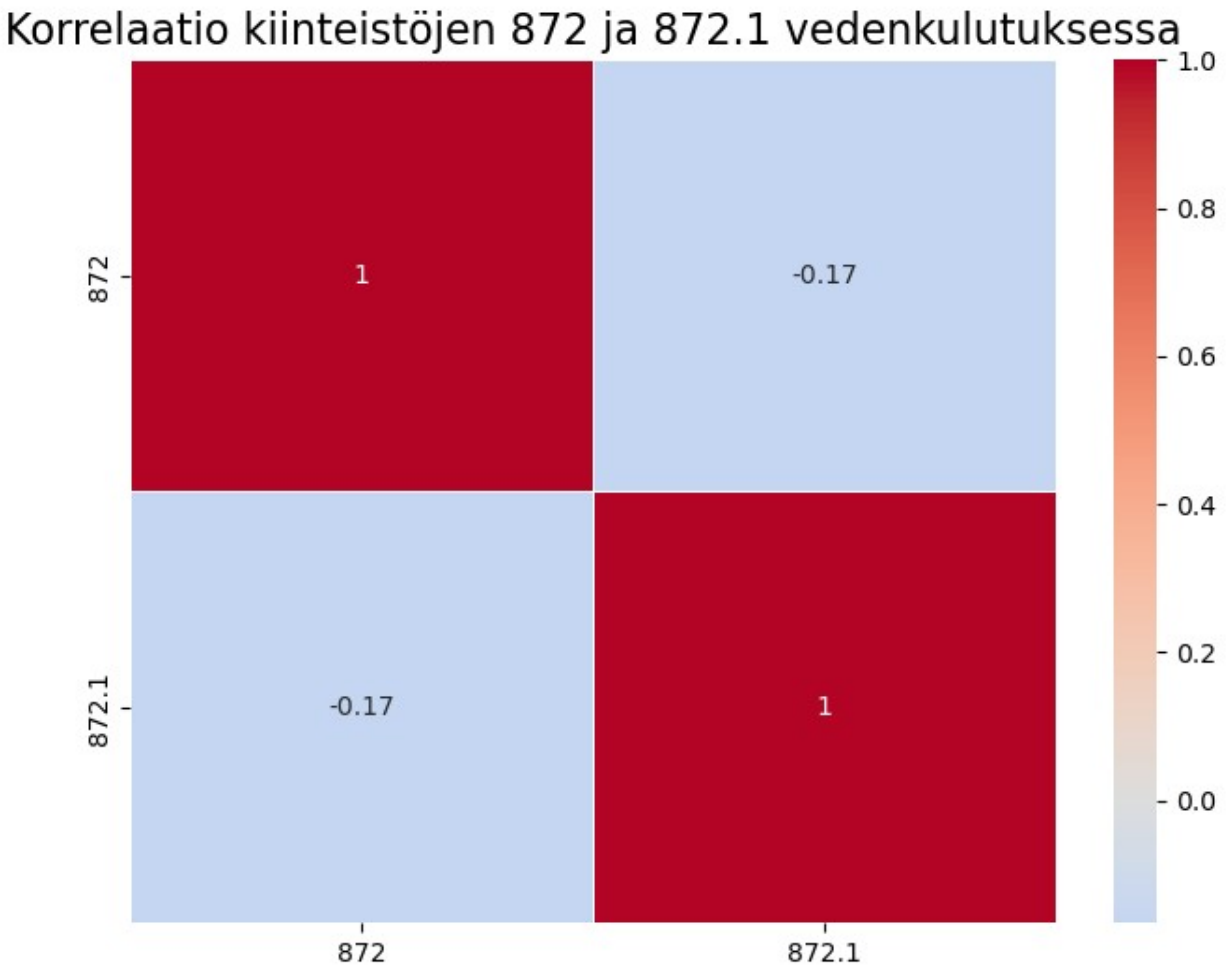
Korrelaatio ymmärrykseni mukaan kiinteistöillä ei ole minkäänlaista riippuvuutta koska arvot aika lähellä nollaa elikkä toisin sanoen ainakun toisen vedenkulutus kasvaa niin toisen kiinteistön laskee.

Mieti mitä tällä datalla voidaan mahdollisesti tehdä?

Datasta voidaan tutkia poikkeamat mitkä oli mainittu tehtävä ykkösessä, laskea mahdollisesti keskiarvot, maksimit ja minimit vedenkulutuksesta, tarkkailla päivittäis, viikottais tai kuukausittais kulutukset ja vertailla niitä kiinteistöjen välillä ja mahdollisesti katsoa onko kiinteistöillä mitään yhteyksiä.

```
#Muutin graafin piirakka muotoon sillä se oli paljon selkeämpi
ymmärtää vedenkulutuksen osuuden kiinteistöjen välillä.

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("Yorkshire_external_meters_2012_2015.csv",
index_col=0)

kiinteisto1 = '872'
kiinteisto2 = kiinteisto1 + '.1'
vuosi = '2012'


korreelaatio_df = df[[kiinteisto1, kiinteisto2]]

korrelaatio = korreelaatio_df.corr()

plt.figure(figsize=(8, 6))
sns.heatmap(korrelaatio, annot=True, cmap='coolwarm', center=0,
linewidths=0.5)
```

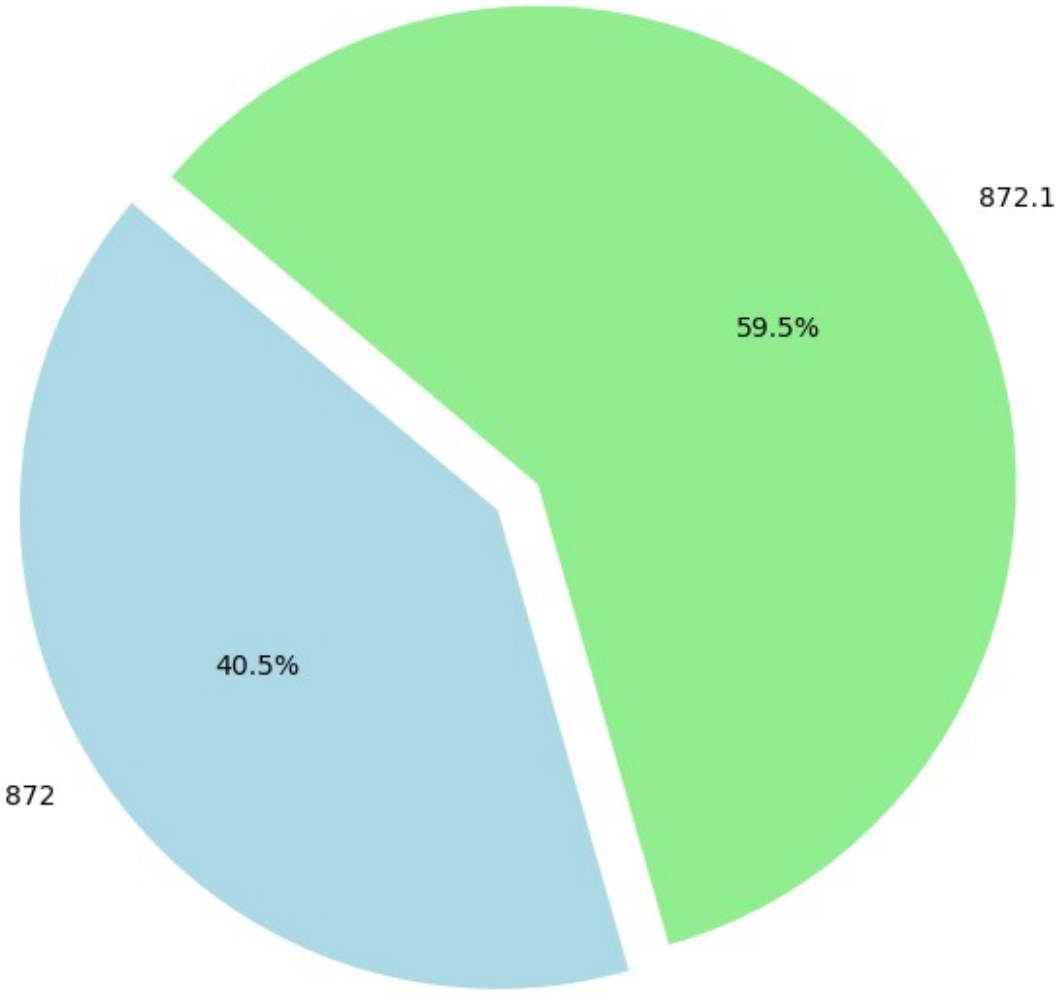
```
plt.title('Korrelaatio kiinteistöjen ' + kiinteisto1 + ' ja ' +
kiinteisto2 + ' vedenkulutuksessa', size=16)
plt.show()
```




```
#Muutin graafin piirakka muotoon sillä se oli paljon selkeämpi
ymmärtää vedenkulutuksen osuuden kiinteistöjen välillä.

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv("Yorkshire_external_meters_2012_2015.csv",
index_col=0)
df = df.replace(-1.0, np.nan)

kiinteisto1 = '872'
kiinteisto2 = kiinteisto1 + '.1'

kiinteisto_summa = df[[kiinteisto1, kiinteisto2]][df['Year'] ==
```

```
2012].sum()

labels = [kiinteisto1, kiinteisto2]
sizes_2012 = [kiinteisto_summa[kiinteisto1],
kiinteisto_summa[kiinteisto2]]
colors = ['lightblue', 'lightgreen']

fig, ax = plt.subplots(figsize=(8, 8))
ax.pie(sizes_2012, explode=(0.1, 0), labels=labels, colors=colors,
autopct='%1.1f%%', startangle=140)
ax.set_title('Vedenkulutuksen osuus kiinteistöjen ' + kiinteisto1 + '
ja ' + kiinteisto2 + ' välillä vuonna 2012', size=12)

plt.show()
```

Vedenkulutuksen osuus kiinteistöjen 872 ja 872.1 välillä vuonna 2012



Tehtävä 3: Datan valmistelu (esikäsittely, Data Preparation)

Tehtävä 3: Vastaukset

Mikä voisi selittää vedenkäytössä esiintyvät pidemmät käyttökatkot?

Kun kyseessä on minun mielestäni ravintola tai siivous-, tai pesuyhtiö niin seuraavat asiat voivat johtaa pidempiin käyttökatkoihin:

- Huoltotoimenpiteet
- Vuodot ja putkirikot
- Korjaukset tai investoinnit
- Suuret käyttömäärät
- Yllättävät ongelmat tai tekniset viat

Mikä voisi selittää käyttöpiikit kuvaajissa?

Jos otamme kuvaajan 3.1 huomioon olen laittanut siihen vuodet 2012 sekä 2013 ja kyseisten vuosien vedenkulutuksen keskiarvon sekä maksimin.

Tässä ovat arvot jotka saamme:

Vuosi 2012 Kiinteistön 892 vedenkulutuksen keskiarvo: 0.32393214285714295 m3 Kiinteistön 892 vedenkulutuksen maksimi: 0.746 m3

Vuosi 2013 Kiinteistön 892 vedenkulutuksen keskiarvo: 0.3223150684931509 m3 Kiinteistön 892 vedenkulutuksen maksimi: 0.737 m3

Tärkein elementti minkä huomaa minusta on se et kyseessä on ravintola tai jonkinlainen siivous-, ja pesuyhtiö koska perustalous ei kuluta maksimissaan 740 litraa päivässä tai keskiarvoltaan 320 litraa.

Elikkä piikit voivat kuvaajassa tarkoittaa kiireistä päivää kiinteistössä kuten esimerkiksi jos kyseessä on ravintola niin kyseessä voisi olla rallit, festarit tai juhlat. Siivous-, tai pesuyhtiö taas voisi olla kiireinen päivä.

Esiintyykö vedenkäytössä tunnistettavia malleja?

Kolmea eri kiinteistöä vertaillen seuraavat asiat esiintynyt:

Kausittaiset vaihtelut = Tietyt kiinteistöt voivat kuluttaa vettä keskimäärin enemmän tietyinä vuodenaikoina. Esimerkiksi kiinteistö 871 näyttää kuluttavan enemmän vettä viikolla 3 vuosina 2012 ja 2013.

Päivittäiset vaihtelut = Joillakin kiinteistöillä voi olla selkeitä päivittäisiä vaihteluja vedenkulutuksessa, kuten korkeampi kulutus tiettyinä kellonaikoina.

Piikit = Voi olla erityisen korkeita kulutushuippuja tai piikkejä, jotka voivat johtua erityistapahtumista

Korrelaatio = Vedenkulutus korreloi muiden tekijöiden, kuten sääolosuhteiden, käyttötarkoituksen tai asukasmäärän kanssa.

Mitä muita asioita kuvaajista on luettavissa?

Ensimmäinen kuviosarja vertailee päivittäistä vedenkulutusta 892 kiinteistön osalta vuosina 2012 ja 2013.

Toinen kuviosarja vertailee päivittäistä vedenkulutusta 892 kiinteistön osalta huhtikuussa 2012 ja heinäkuussa 2013.

Kolmas kuviosarja vertailee päivittäistä vedenkulutusta 892 kiinteistön osalta viikolla 3 vuosina 2012 ja 2013.

Elikkä kaiken kaikkiaan kuvaajat tarjoavat dataa kiinteistöjen vedenkulutuksesta jotka auttavat henkilöä ymmärtämään veden kulutuksen sekä jos kiinteistöissä esintyy jonkinlaisia poikkeamia.

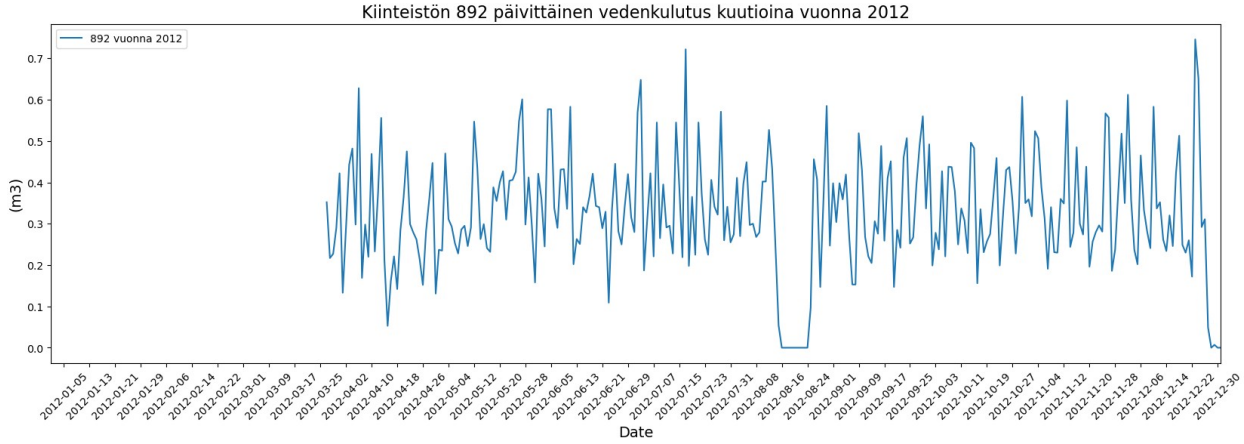
Kuvaajien **3.1-3.3** selitykset

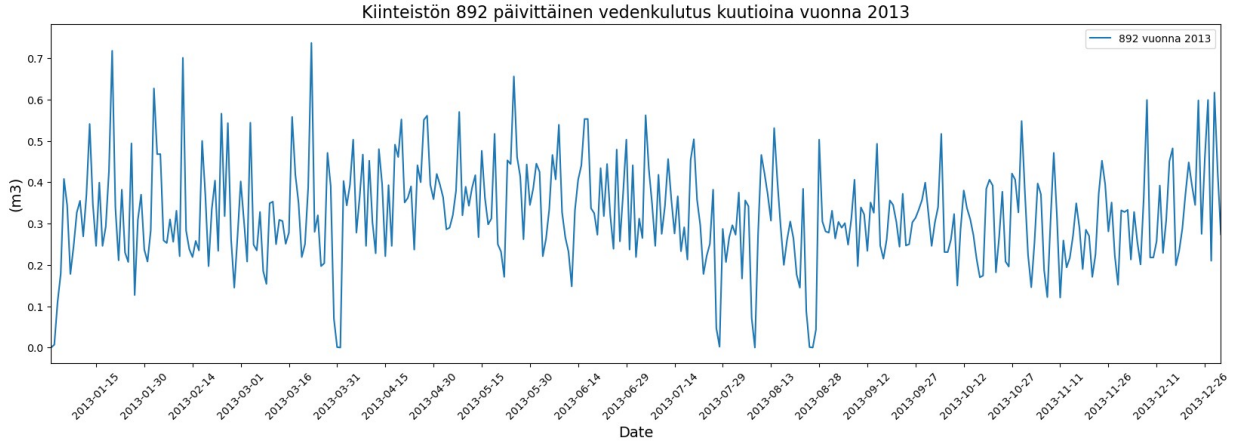
- Kuvaaja 3.1: Saman kiinteistön vedenkäyttö eri vuosina.
- Kuvaaja 3.2: Saman kiinteistön vedenkäyttö eri kuukausina.
- Kuvaaja 3.3: Saman kiinteistän vedenkäyttö eri viikkoina.

```
# Kuvaaja 3.1

import matplotlib.pyplot as plt
import pandas as pd


vuosi1 = '2012'
vuosi2 = '2013'
kiinteisto = '892'

df['date'] = pd.to_datetime(df['date'])

fig, axs = plt.subplots(2, 1, sharey=True, gridspec_kw={'hspace':
0.4}, figsize=(20, 14))

axs[0].plot(df['date'][(df['Year'] == int(vuosi1))], df[kiinteisto]
[(df['Year'] == int(vuosi1))])
axs[0].set_title('Kiinteistön ' + kiinteisto + ' päivittäinen
vedenkulutus kuutioina vuonna ' + vuosi1, size=16)
axs[0].set_ylabel('(m3)', size=14)
axs[0].set_xlabel('Date', size=14)
axs[0].set_xlim(pd.to_datetime(vuosi1 + '-01-01'),
pd.to_datetime(vuosi1 + '-12-31'))
axs[0].xaxis.set_major_locator(plt.MaxNLocator(52))
axs[0].tick_params(axis='x', rotation=45)
axs[0].legend([kiinteisto + ' vuonna ' + vuosi1])
```

```
axs[1].plot(df['date'][(df['Year'] == int(vuosi2))], df[kiinteisto]
[(df['Year'] == int(vuosi2))])
axs[1].set_title('Kiinteistön ' + kiinteisto + ' päivittäinen
vedenkulutus kuutioina vuonna ' + vuosi2, size=16)
axs[1].set_ylabel('(m3)', size=14)
axs[1].set_xlabel('Date', size=14)
axs[1].set_xlim(pd.to_datetime(vuosi2 + '-01-01'),
pd.to_datetime(vuosi2 + '-12-31'))
axs[1].xaxis.set_major_locator(plt.MaxNLocator(30))
axs[1].tick_params(axis='x', rotation=45)
axs[1].legend([kiinteisto + ' vuonna ' + vuosi2])

vesikulutus_kiinteistolta_keskiarvo = df[df['Year'] == 2012]
[kiinteisto].mean()
vesikulutus_kiinteistolta_max = df[df['Year'] == 2012]
[kiinteisto].max()

vesikulutus_kiinteistolta_keskiarvo1 = df[df['Year'] == 2013]
[kiinteisto].mean()
vesikulutus_kiinteistolta_max2 = df[df['Year'] == 2013]
[kiinteisto].max()

print("Kiinteistöjen keskiarvo sekä maksimi vuodelta 2012")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{vesikulutus_kiinteistolta_keskiarvo} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi:
{vesikulutus_kiinteistolta_max} m3")
print("\n")
print("Kiinteistöjen keskiarvo sekä maksimi vuodelta 2013")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{vesikulutus_kiinteistolta_keskiarvo1} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi:
{vesikulutus_kiinteistolta_max2} m3")

plt.show()
```

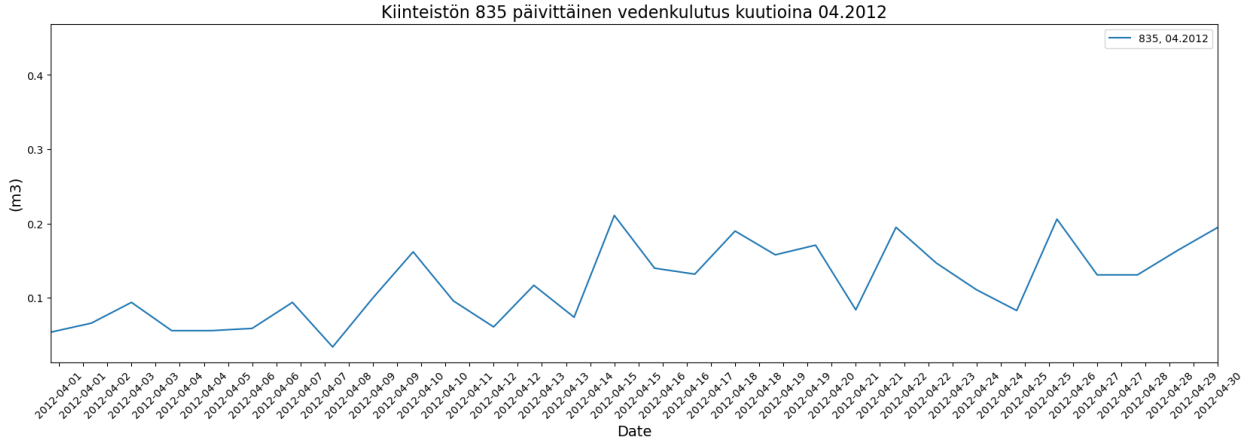
```
Kiinteistöjen keskiarvo sekä maksimi vuodelta 2012
Kiinteistön 892 vedenkulutuksen keskiarvo: 0.32393214285714295 m3
Kiinteistön 892 vedenkulutuksen maksimi: 0.746 m3


Kiinteistöjen keskiarvo sekä maksimi vuodelta 2013
Kiinteistön 892 vedenkulutuksen keskiarvo: 0.3223150684931509 m3
Kiinteistön 892 vedenkulutuksen maksimi: 0.737 m3
```

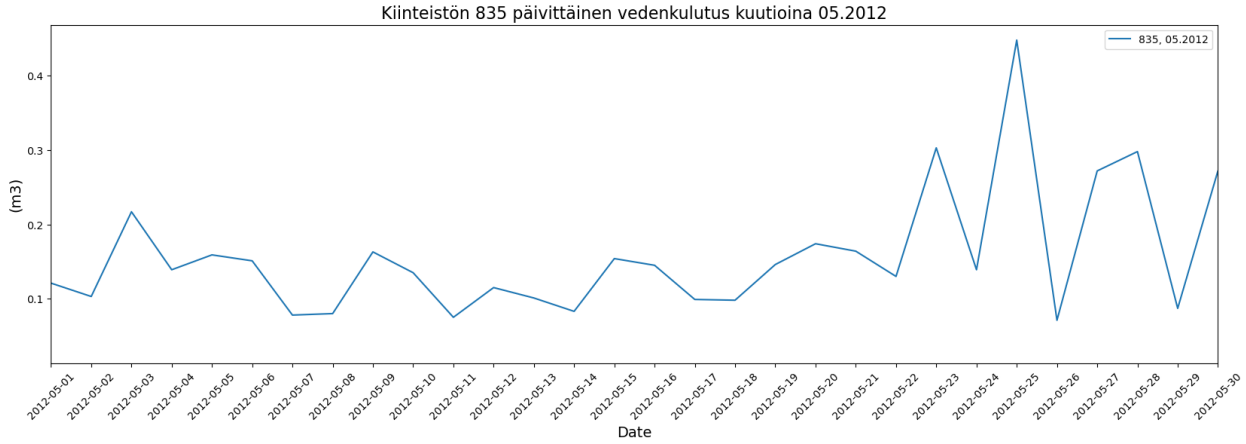





```
# Kuvaaja 3.2

vuosi1 = '2012'
kuukausi1 = '04'
vuosi2 = '2012'
kuukausi2 = '05'
kiinteisto = '835'

df['date'] = pd.to_datetime(df['date'])

fig, axs = plt.subplots(2, 1, sharey=True, gridspec_kw={'hspace':
0.4}, figsize=(20, 14))

valid_date_filter1 = (df['date'].dt.year == int(vuosi1)) &
(df['date'].dt.month == int(kuukausi1))
valid_date_filter2 = (df['date'].dt.year == int(vuosi2)) &
(df['date'].dt.month == int(kuukausi2))

axs[0].plot(df['date'][valid_date_filter1],
            df[kiinteisto][valid_date_filter1])
axs[0].set_title(
    'Kiinteistön ' + kiinteisto + ' päivittäinen vedenkulutus
```

```
kuutioina ' + kuukausi1 + '.' + vuosi1,
    size=16)
axs[0].set_ylabel('(m3)', size=14)
axs[0].set_xlabel('Date', size=14)
axs[0].set_xlim(pd.to_datetime(vuosi1 + '-' + kuukausi1 + '-01'),
pd.to_datetime(vuosi1 + '-' + kuukausi1 + '-30'))
axs[0].xaxis.set_major_locator(plt.MaxNLocator(52))
axs[0].tick_params(axis='x', rotation=45)
axs[0].legend([kiinteisto + ', ' + kuukausi1 + '.' + vuosi1])

axs[1].plot(df['date'][valid_date_filter2],
            df[kiinteisto][valid_date_filter2])
axs[1].set_title(
    'Kiinteistön ' + kiinteisto + ' päivittäinen vedenkulutus
kuutioina ' + kuukausi2 + '.' + vuosi2, size=16)
axs[1].set_ylabel('(m3)', size=14)
axs[1].set_xlabel('Date', size=14)
axs[1].set_xlim(pd.to_datetime(vuosi2 + '-' + kuukausi2 + '-01'),
pd.to_datetime(vuosi2 + '-' + kuukausi2 + '-30'))
axs[1].xaxis.set_major_locator(plt.MaxNLocator(30))
axs[1].tick_params(axis='x', rotation=45)
axs[1].legend([kiinteisto + ', ' + kuukausi2 + '.' + vuosi2])

plt.show()

kuukausi1_keskiarvo = df[valid_date_filter1][kiinteisto].mean()
kuukausi2_max = df[valid_date_filter1][kiinteisto].max()

kuukausi1_keskiarvo2 = df[valid_date_filter2][kiinteisto].mean()
kuukausi2_max2 = df[valid_date_filter2][kiinteisto].max()

print("Kiinteistö 835 keskiarvo sekä maksimi huhtikuussa 2012")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{kuukausi1_keskiarvo} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi:
{kuukausi2_max} m3")
print("\n")
print("Kiinteistö 835 keskiarvo sekä maksimi toukokuulta 2012")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{kuukausi1_keskiarvo2} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi:
{kuukausi2_max2} m3")
```

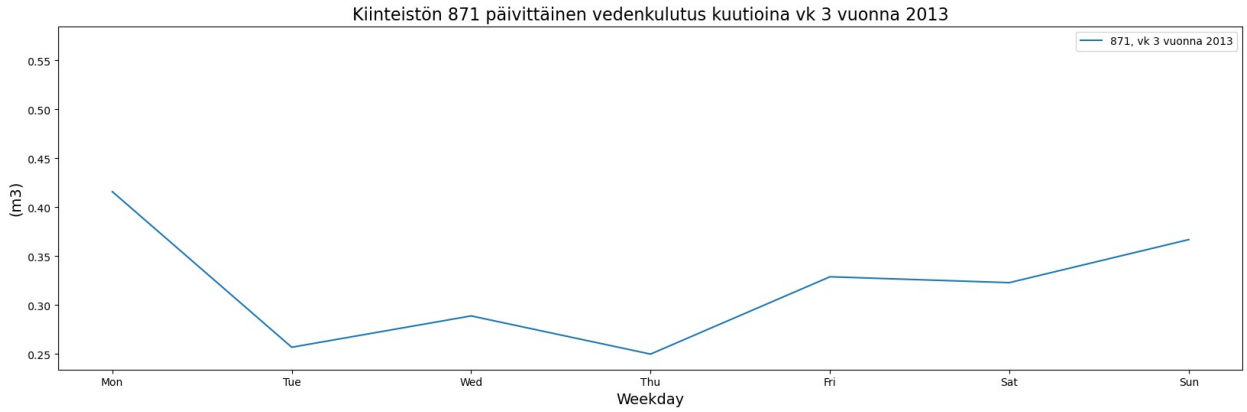


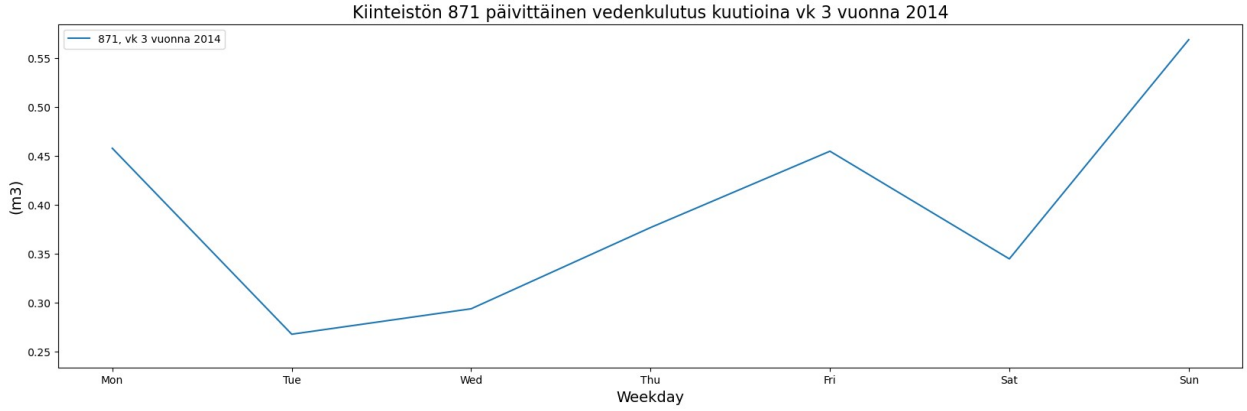



```
Kiinteistö 835 keskiarvo sekä maksimi huhtikuussa 2012
Kiinteistön 835 vedenkulutuksen keskiarvo: 0.11906666666666667 m3
Kiinteistön 835 vedenkulutuksen maksimi: 0.211 m3


Kiinteistö 835 keskiarvo sekä maksimi toukokuulta 2012
Kiinteistön 835 vedenkulutuksen keskiarvo: 0.1574838709677419 m3
Kiinteistön 835 vedenkulutuksen maksimi: 0.448 m3
```

```
# Kuvaaja 3.3

vuosi1 = '2013'
viikko1 = '3'
vuosi2 = '2014'
viikko2 = '3'
kiinteisto = '871'

fig, axs = plt.subplots(2, 1, sharey=True, gridspec_kw={'hspace':
0.4}, figsize=(20, 14))

axs[0].plot(df['Weekday'][(df['Year'] == int(vuosi1)) & (df['Week'] ==
int(viikko1))],
```

```
            df[kiinteisto][(df['Year'] == int(vuosi1)) & (df['Week']
== int(viikko1))])
axs[0].set_title(
    'Kiinteistön ' + kiinteisto + ' päivittäinen vedenkulutus
kuutioina vk ' + viikko1 + ' vuonna ' + vuosi1,     size=16)
axs[0].set_ylabel('(m3)', size=14)
axs[0].set_xlabel('Weekday', size=14)
axs[0].legend([kiinteisto + ', vk ' + viikko1 + ' vuonna ' + vuosi1])

axs[1].plot(df['Weekday'][(df['Year'] == int(vuosi2)) & (df['Week'] ==
int(viikko2))],
            df[kiinteisto][(df['Year'] == int(vuosi2)) & (df['Week']
== int(viikko2))])
axs[1].set_title(
    'Kiinteistön ' + kiinteisto + ' päivittäinen vedenkulutus
kuutioina vk ' + viikko2 + ' vuonna ' + vuosi2, size=16)
axs[1].set_ylabel('(m3)', size=14)
axs[1].set_xlabel('Weekday', size=14)
axs[1].legend([kiinteisto + ', vk ' + viikko2 + ' vuonna ' + vuosi2])

plt.show()

paiva_keskiarvo1 = df[(df['Year'] == int(vuosi1)) & (df['Week'] ==
int(viikko1))][kiinteisto].mean()
paiva_max1 = df[(df['Year'] == int(vuosi1)) & (df['Week'] ==
int(viikko1))][kiinteisto].max()

paiva_keskiarvo2 = df[(df['Year'] == int(vuosi2)) & (df['Week'] ==
int(viikko2))][kiinteisto].mean()
paiva_max2 = df[(df['Year'] == int(vuosi2)) & (df['Week'] ==
int(viikko2))][kiinteisto].max()

print("Kiinteistö 871 keskiarvo sekä maksimi päivittäinen \
nvedenkulutus kuutioina vk 3 2012")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{paiva_keskiarvo1} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi: {paiva_max1}
m3")
print("\n")
print("Kiinteistö 871 keskiarvo sekä maksimi päivittäinen \
nvedenkulutus kuutioina vk 3 2013")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen keskiarvo:
{paiva_keskiarvo2} m3")
print(f"Kiinteistön {kiinteisto} vedenkulutuksen maksimi: {paiva_max2}
m3")
```






```
Kiinteistö 871 keskiarvo sekä maksimi päivittäinen
vedenkulutus kuutioina vk 3 2012
Kiinteistön 871 vedenkulutuksen keskiarvo: 0.31871428571428567 m3
Kiinteistön 871 vedenkulutuksen maksimi: 0.416 m3


Kiinteistö 871 keskiarvo sekä maksimi päivittäinen
vedenkulutus kuutioina vk 3 2013
Kiinteistön 871 vedenkulutuksen keskiarvo: 0.39514285714285713 m3
Kiinteistön 871 vedenkulutuksen maksimi: 0.569 m3
```