```
import pandas as pd
df_animal = pd.read_csv("data/msleep.csv")
df_animal
```

```
                          name       genus   vore          order
conservation  \
0                      Cheetah    Acinonyx  carni     Carnivora
lc
1                   Owl monkey       Aotus   omni      Primates
NaN
2               Mountain beaver  Aplodontia  herbi      Rodentia
nt
3   Greater short-tailed shrew     Blarina   omni  Soricomorpha
lc
4                          Cow         Bos  herbi  Artiodactyla
domesticated
..                         ...         ...    ...           ...
...
78                  Tree shrew      Tupaia   omni    Scandentia
NaN
79        Bottle-nosed dolphin    Tursiops  carni       Cetacea
NaN
80                       Genet     Genetta  carni     Carnivora
NaN
81                  Arctic fox      Vulpes  carni     Carnivora
NaN
82                     Red fox      Vulpes  carni     Carnivora
NaN

    sleep_total  sleep_rem  sleep_cycle  awake  brainwt   bodywt
0          12.1        NaN          NaN   11.9      NaN   50.000
1          17.0        1.8          NaN    7.0  0.01550    0.480
2          14.4        2.4          NaN    9.6      NaN    1.350
3          14.9        2.3     0.133333    9.1  0.00029    0.019
4           4.0        0.7     0.666667   20.0  0.42300  600.000
..          ...        ...          ...    ...      ...      ...
78          8.9        2.6     0.233333   15.1  0.00250    0.104
79          5.2        NaN          NaN   18.8      NaN  173.330
80          6.3        1.3          NaN   17.7  0.01750    2.000
81         12.5        NaN          NaN   11.5  0.04450    3.380
82          9.8        2.4     0.350000   14.2  0.05040    4.230

[83 rows x 11 columns]
```

```
# İstatistik Nedir?
# İstatistik, veri toplama ve analiz etme uygulama ve çalışmasıdır.
# İstatistiğin İki Ana Dalı:
# 1.Descriptive (Betimsel) İstatistik: Elde edilen veriyi özetlemek ve
açıklamak için
# kullanılır. Bu dal, örneğin bir grubun ortalama geliri ya da
```

```
notlarının dağılımı gibi,
# verinin mevcut durumunu ifade eder. Bu tür özetler ve açıklamalar
grafik, tablo ya da
# sayısal ölçülerle yapılabilir.
# Descriptive/Summary İstatistik¶
#Dört arkadaşa işe nasıl gittikleri sorulduğunda; %50'si işe arabayla,
%25'i otobüs %25'i ise bisikletle gittiğini
#belirtsin. Bunlar betimsel istatistiklerdir.
# 2.Inferential (Çıkarımsal) İstatistik: Bir örnekten yola çıkarak,
genel popülasyon hakkında
# tahminlerde bulunmayı içerir. Örneğin, bir şehirde yapılan anket
sonucunda tüm ülke hakkında
# tahmin yürütmek bu dalın konusudur. Çıkarımsal istatistik,
örneklemden elde edilen bilgiyi
# genelleyerek, popülasyon hakkında öngörüde bulunmamızı sağlar.
#Inferential İstatistik
#Bir popülasyon hakkında sonuç çıkarmak için bir örneklem kullanılır.
#Örneğin, 100 kişiye sosyal medya reklamlarını gördükten sonra kıyafet
alıp almadıkları sorulabilir
#ve burdan elde edilen sonuç tüm insanların yüzde kaçının sosyal medya
reklamı sonrası kıyafet
#aldığını anlamak için kullanılabilir.
# Veri Türleri:
# Nümerik/Nicel Veriler:
# Bu veriler sayısal değerlerden oluşur ve kendi içinde ikiye ayrılır:
# Sürekli Veriler: Herhangi bir aralık içinde ölçülebilen değerlerdir.
Örneğin hisse senedi fiyatı, günlük rüzgar hızı veya ürün kutusu
ölçüleri.
# Ayrık Veriler: Sayılabilir ve belirli aralıklarla ifade edilebilen
değerlerdir. Örneğin ürün incelemelerinin sayısı veya sınıftaki
öğrenci sayısı.
# Nümerik Verileri Görselleştirme: Nümerik veriler arasındaki ilişkiyi
görselleştirmenin yaygın bir yolu dağılım grafikleri kullanmaktır.

# Kategorik/Nitel Veriler:
# Bu veriler sayısal olmayan, kategorilere ayrılmış verilerdir:
# Nominal Veriler: Sıralama gerektirmeyen veriler. Örneğin göz rengi.
# Ordinal Veriler: Sıralanabilir kategorilere sahip verilerdir.
Örneğin, eğitim düzeyi veya memnuniyet anketlerinde kullanılan
“seviyorum”, “sevmiyorum” gibi ifadeler.

# Merkez Ölçümleri
# Veri kümelerinin merkezi eğilimlerini incelemek için üç ana ölçüm
kullanılır: Ortalama, Medyan ve Mod.
#Ortalama: Verilerin toplamının, veri sayısına bölünmesiyle elde
edilir. En çok kullanılan merkez ölçümlerindendir.
#Medyan: Verilerin sıralandığında ortada kalan değeri temsil eder.
Özellikle uç değerlerden etkilenmeyen bir merkez ölçüsüdür.
#Mod: Verilerde en sık tekrarlanan değeri ifade eder.
```

#Örneğin, bir iş yerinde aylık ortalama sipariş sayısını veya bir evin tipik maliyetini belirlemek için bu merkez ölçümleri kullanılır. Histogramlar, sayısal verileri özetlemenin ve dağılımını görselleştirmenin etkili bir yoludur.

#Yayılım Ölçüleri
Yayılım ölçümleri, verinin ne kadar geniş bir alana yayıldığını gösterir.
Range (Aralık): En büyük ve en küçük değer arasındaki farktır.
Varyans: Her bir veri noktasının ortalamaya olan uzaklığını hesaplayarak veri setindeki yayılımı gösterir. Yüksek varyans, verilerin daha geniş bir alana yayıldığını belirtir.
Standart Sapma: Varyansın kareköküdür ve verilerin ortalama etrafında nasıl kümelendiğini gösterir. Standart sapma sıfıra ne kadar yakınsa verilerin ortalama etrafında o kadar yakın kümlendiği anlaşılır.
Çeyrekler (Quartiles): Veriyi dört eşit parçaya böler. Bu parçalar, verinin dağılımını ölçmede etkilidir. Boxplot grafikleri, çeyrekleri ve aykırı değerleri görselleştirmek için kullanılır.
Interquartile Range (IQR): Çeyrekler arası mesafeyi hesaplar ve aykırı değerleri belirlemede kullanılır.
Aykırı Değerler (Outliers): Verilerden önemli ölçüde sapma gösteren noktalardır. Bir değerin aykırı olup olmadığını belirlemek için genellikle IQR kullanılır.
Q1 - 1.5IQR < data < Q3 + 1.5IQR

Mean Absolute Deviation (MAD), veri noktalarının ortalamadan ne kadar uzaklaştığını ölçen bir yayılma ölçüsüdür. Standart sapmada her uzaklığın karesi alınır; bu, daha uzak mesafelerin daha fazla "ceza" almasına yol açar ve böylece uç değerlerin etkisini büyütür.
MAD'de ise her veri noktasının ortalamaya olan mutlak uzaklığı (yani pozitif olarak kabul edilen mesafesi) alınır. Böylece her veri eşit bir "ceza" alır ve uç değerlerin etkisi daha azdır. Bu yüzden:

Standart sapma daha fazla uç değer içeren dağılımlar için daha duyarlıdır.
MAD, daha dengeli bir dağılım arayan analizler için kullanışlıdır.
İki yöntemden biri diğerinden üstün değildir, ancak standart sapma daha yaygın olarak kullanılır.

Ayrık ve Sürekli Dağılımlar
Ayrık Dağılımlar: Belirli sayıda olasılık içerir; örneğin, bir zar atıldığında her bir yüzün çıkma olasılığı eşittir.
Sürekli Dağılımlar: Sayısız olasılık vardır. Örneğin, otobüs bekleme süresi 0 ile 12 dakika arasında herhangi bir değere sahip olabilir.

Olasılık Dağılımı ve Beklenen Değer
Olasılık dağılımı, her sonucun olasılığını açıklar ve beklenen değer, olasılıklarla ağırlıklı olarak sonuçların ortalamasıdır.
Histogram gibi görselleştirmeler ile dağılımlar daha kolay anlaşılır.
Olasılık Dağılımları Neden Önemlidir?

• Riski ölçmeye ve karar alma sürecini bilgilendirmeyi sağlar.
• Hipotez testlerinde sonuçların şans eseri çıkıp çıkmadığını anlamak için

Skewness (Çarpıklık) ve Kurtosis (Basıklık)
Çarpıklık: Veri dağılımının hangi yöne kaydığını gösterir; pozitif çarpıklık sağa, negatif çarpıklık sola kaymış dağılımlardır.
Basıklık: Dağılımın uç değerlerinin sıklığını ifade eder; pozitif basıklık yüksek tepe noktasına sahipken, negatif basıklık daha geniş bir yayılma gösterir.

Merkezi Limit Teoremi (MLT)
Büyük bir örneklemde, örnek ortalamalarının dağılımı, normal dağılıma yaklaşır.
Teorem, örneklem büyüklüğü en az 30 olduğunda geçerlidir.

Hipotez Testi
Hipotez testi, popülasyonlar arasında istatistiksel bir fark olup olmadığını belirlemek için kullanılır.
Null (H0) ve alternatif (H1) hipotezler belirlenir ve sonuç p-değerine göre yorumlanır. Alfa (α) tipik olarak 0.05 olarak alınır.
ip I Hata: Null hipotezin yanlışlıkla reddedilmesi.
Tip II Hata: Null hipotezin yanlışlıkla kabul edilmesi.

Serbestlik Derecesi (Degrees of Freedom)
Bir istatistiksel hesaplamada bağımsız olarak değişebilen veri sayısıdır.
Daha düşük serbestlik derecesi t-dağılımının kuyruklarını genişletir; daha yüksek serbestlik derecesi, dağılımı normal dağılıma yaklaştırır.
Hipotez Testi
Popülasyonlar arasında fark olup olmadığını belirlemek için kullanılan bir istatistiksel test yöntemidir.
Null hipotez (H0) hiçbir fark olmadığını varsayar; alternatif hipotez (H1) fark olduğunu öne sürer. P-değeri ile sonuç değerlendirilir, alfa seviyesi ile kıyaslanır.
4. Bağımsız ve Bağımlı Değişkenler
Bağımsız Değişken: Diğer verilerden etkilenmeyen değişkendir (örn. tedavi uygulaması).
Bağımlı Değişken: Bağımsız değişkenden etkilenen veridir (örn. tedavi sonucu).
5. Deney Tasarımı (Design of Experiments)
Bir deneyde bağımsız değişkenin bağımlı değişken üzerindeki etkisini analiz etmek için veriler toplanır. Kontrollü deneylerde, katılımcılar rastgele gruplara atanarak önyargı en aza indirilir.
6. Altın Standart Deneyler
Randomize kontrollü, çift kör çalışmalardır. Rastgelelik ve körleme kullanılarak önyargı azaltılır. İlaç denemelerinde yaygın olarak görülür.
7. Gözlemsel Çalışmalar
Rastgele atamanın mümkün olmadığı durumlarda yapılır. Bu çalışmalar

ilişki kurabilir ama nedensellik belirlemez, çünkü önyargılardan etkilenebilir.
8. Uzunlamasına ve Kesitsel
Uzunlamasına: Katılımcılar zaman içinde gözlemlenir, daha güvenilir sonuçlar verir.
Kesitsel: Tek bir zaman diliminde veri toplar, hızlı ve düşük maliyetlidir.

```
df_animal.vore.value_counts() # sözel verilerde mod bilgisini verir.
```

```
vore
herbi      32
omni       20
carni      19
insecti     5
Name: count, dtype: int64
```

```
df_animal.sleep_total.mean()
```

```
10.433734939759034
```

```
df_animal.sleep_total.mode()
df_animal.sleep_total.mean()
df_animal.sleep_total.median()
```

```
istenen = df_animal.sleep_total
toplam = df_animal.sleep_total.agg('sum')
toplam
uzunluk = len(list(df_animal.sleep_total))
ortalama = toplam/uzunluk
ortalama
```

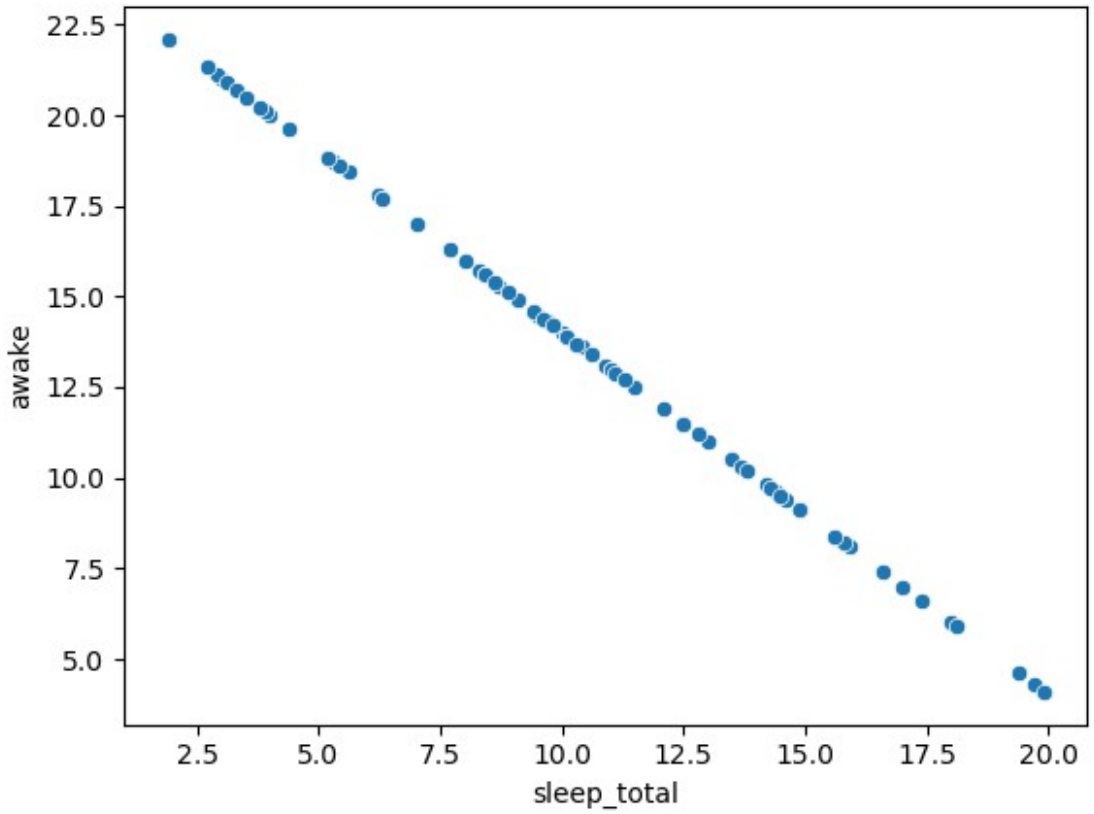
```
10.433734939759034
```

```
df_animal[df_animal['vore'] == 'carni']['awake'].agg('mode')
```

```
0    11.5
Name: awake, dtype: float64
```

```
import seaborn as sns
sns
df_animal['sleep_total'].corr(df_animal['awake'])
sns.scatterplot(x='sleep_total', y='awake', data=df_animal)
df_animal['sleep_total'].corr(df_animal['awake']) #korelasyon -0.99 olduğu için buna negatif güçlü korelasyon deriz. yani güçlü ters orantu vardır!
```

```
-0.9999985737040996
```




```
import statistics as stat
stat.mode(df_animal.vore)
```

```
'herbi'
```

```
#range sleep total
range_total = df_animal['sleep_total'].max() -
df_animal['sleep_total'].min()
range_total
```

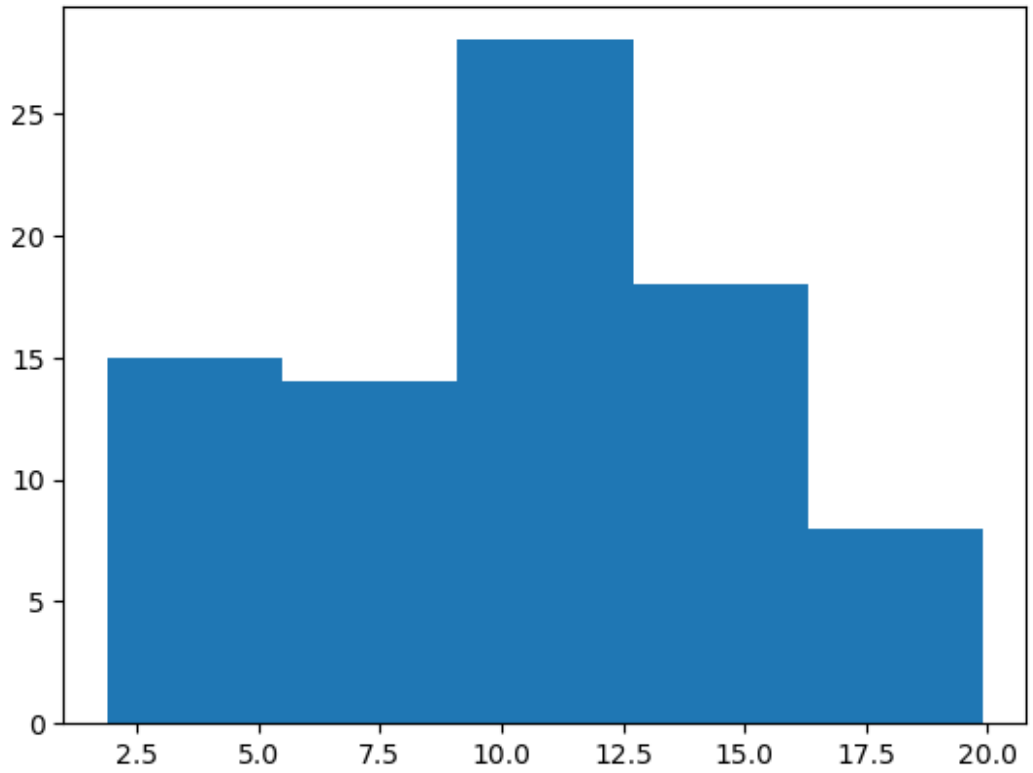
```
18.0
```

```
#sleep_total varyansı. varyans: ortalamadan sapmanın büyüklüğünü ifade
eder
import numpy as np
varyans = np.var(df_animal['sleep_total'], ddof=1)
#sleep total standart sapma: varyansın kareköküdür. ortalamadan
hepsinin uzaklığını temsil eder
np.sqrt(varyans)
np.std(df_animal['sleep_total'], ddof=1) #yukarıdaki ile aynı sonucu
verir.
#mean abs dev -> ortalama mutlak sapma
```

```
4.4503569905705795
```

```
import matplotlib.pyplot as plt
plt.hist(df_animal['sleep_total'], bins=5)
```

```
(array([15., 14., 28., 18.,  8.]),
 array([ 1.9,  5.5,  9.1, 12.7, 16.3, 19.9]),
 <BarContainer object of 5 artists>)
```




```
#quantile: veriyi belli yüzdelik bölümlere bölme işlemidir
np.quantile(df_animal['sleep_total'], 0.5) #mediandır. çünkü veriyi
ortadan ikiye böler
```

```
10.1
```

```
np.quantile(df_animal['sleep_total'], ([0.0, 0.25, 0.5, 0.75, 1.0]))
#0.0 yüzdeliği, veri setindeki minimum değeri döndürür. Yani, veri
setindeki en küçük değeri elde etmek için bu yüzdeliği
kullanabilirsiniz
# Veri setindeki en küçük değer (minimum değer).
# 0.25: İlk çeyrek veya alt çeyrek (Q1), yani veri kümesinin alt yüzde
25'lik bölümünün üst sınırını ifade eder.
# 0.5: Medyan, veri kümesinin orta değeri (Q2).
# 0.75: Üçüncü çeyrek veya üst çeyrek (Q3), yani veri kümesinin üst
yüzde 25'lik bölümünün alt sınırını ifade eder.
# 1.0: Veri setindeki en büyük değer (maksimum değer).
```

```
array([ 1.9 ,  7.85, 10.1 , 13.75, 19.9 ])
```

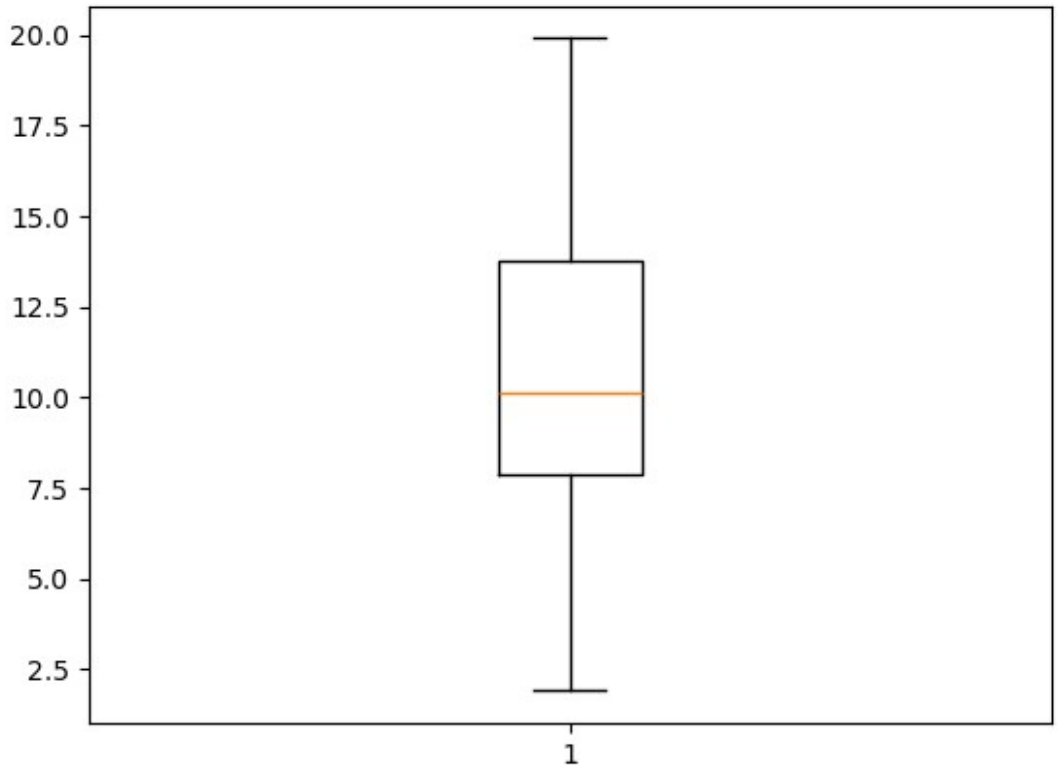
```
#bodywtnin iqr'ını hesapla, sonra aykırı değerleri ver:
# iqr (interquartile range), bir veri setinin üçüncü çeyreği (Q3, 75. yüzdelik) ile birinci çeyreği (Q1, 25. yüzdelik) arasındaki farkı ifade
# eder. Yani, veri setinin orta %50'sinin yayılımını ölçer.
# iqr değeri, veri dağılımının merkezine yakın olan bir değerler grubunun dağılımını yansıtarak, aykırı değerlerin tespitinde kullanılır çünkü
# bu değerler genellikle veri setinin uçlarında yer alır.
import numpy as np
from scipy.stats import iqr
iqr_value = iqr(df_animal['bodywt'])
iqr_value
lower = np.quantile(df_animal['bodywt'], 0.25) + 1.5 * iqr_value
upper = np.quantile(df_animal['bodywt'], 0.75) - 1.5 * iqr_value
df_animal[(df_animal.bodywt < lower) | (df_animal.bodywt > upper)]
#aykırı değerleri bulduk
```

```
                          name       genus   vore         order conservation  \
0                       Cheetah    Acinonyx  carni     Carnivora          lc
1                    Owl monkey       Aotus   omni      Primates         NaN
2                Mountain beaver  Aplodontia  herbi      Rodentia          nt
3   Greater short-tailed shrew     Blarina   omni  Soricomorpha          lc
4                           Cow         Bos  herbi  Artiodactyla  domesticated
..                          ...         ...    ...           ...         ...
78                   Tree shrew      Tupaia   omni    Scandentia         NaN
79         Bottle-nosed dolphin    Tursiops  carni       Cetacea         NaN
80                        Genet     Genetta  carni     Carnivora         NaN
81                  Arctic fox      Vulpes  carni     Carnivora         NaN
82                     Red fox      Vulpes  carni     Carnivora         NaN

    sleep_total  sleep_rem  sleep_cycle  awake  brainwt   bodywt
0          12.1        NaN          NaN   11.9      NaN   50.000
1          17.0        1.8          NaN    7.0  0.01550    0.480
2          14.4        2.4          NaN    9.6      NaN    1.350
3          14.9        2.3     0.133333    9.1  0.00029    0.019
```

```
4           4.0        0.7     0.666667   20.0  0.42300  600.000
..          ...        ...          ...    ...      ...      ...
78          8.9        2.6     0.233333   15.1  0.00250    0.104
79          5.2        NaN          NaN   18.8      NaN  173.330
80          6.3        1.3          NaN   17.7  0.01750    2.000
81         12.5        NaN          NaN   11.5  0.04450    3.380
82          9.8        2.4     0.350000   14.2  0.05040    4.230

[83 rows x 11 columns]
```

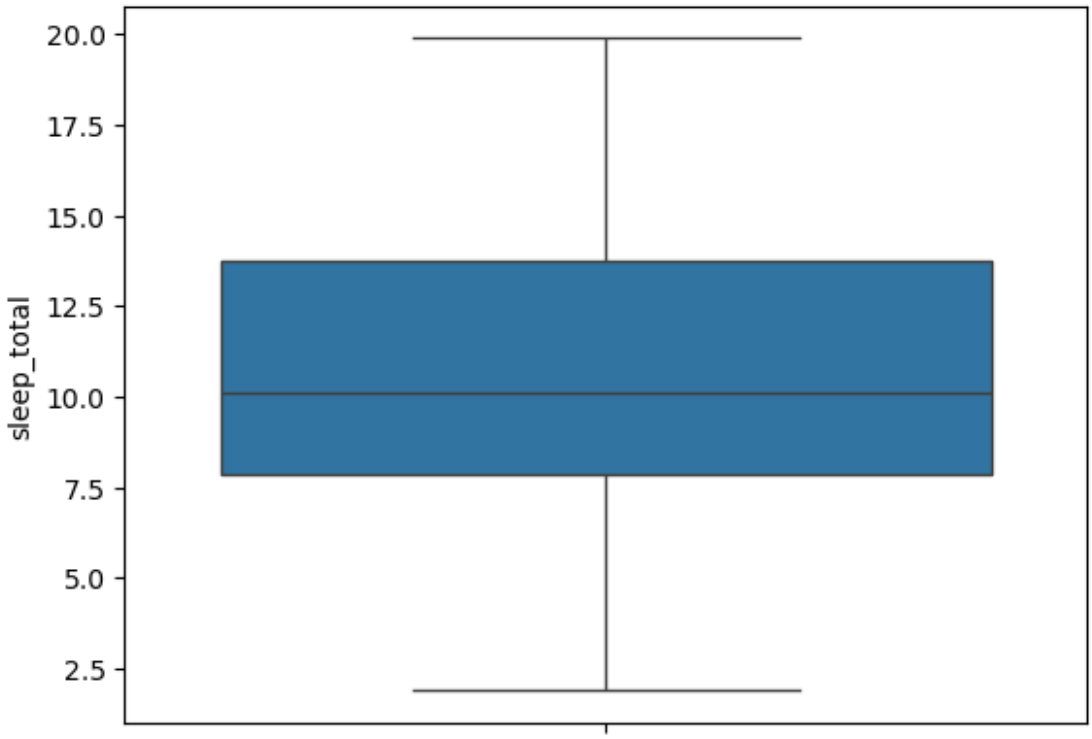
```
#boxplot çizme. Boxplot veri dağılımını ve aykırı değerleri
görselleştiren etkili bir grafik türüdür
import matplotlib.pyplot as plt
plt.boxplot(df_animal.sleep_total)
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x16a36c230>,
  <matplotlib.lines.Line2D at 0x16a36c530>],
 'caps': [<matplotlib.lines.Line2D at 0x16a36c830>,
  <matplotlib.lines.Line2D at 0x16a36ca10>],
 'boxes': [<matplotlib.lines.Line2D at 0x16a32ff20>],
 'medians': [<matplotlib.lines.Line2D at 0x16a36cce0>],
 'fliers': [<matplotlib.lines.Line2D at 0x16a36cf80>],
 'means': []}
```



```
import seaborn as sns
sns.boxplot(data=df_animal , y='sleep_total')
```

```
<Axes: ylabel='sleep_total'>
```

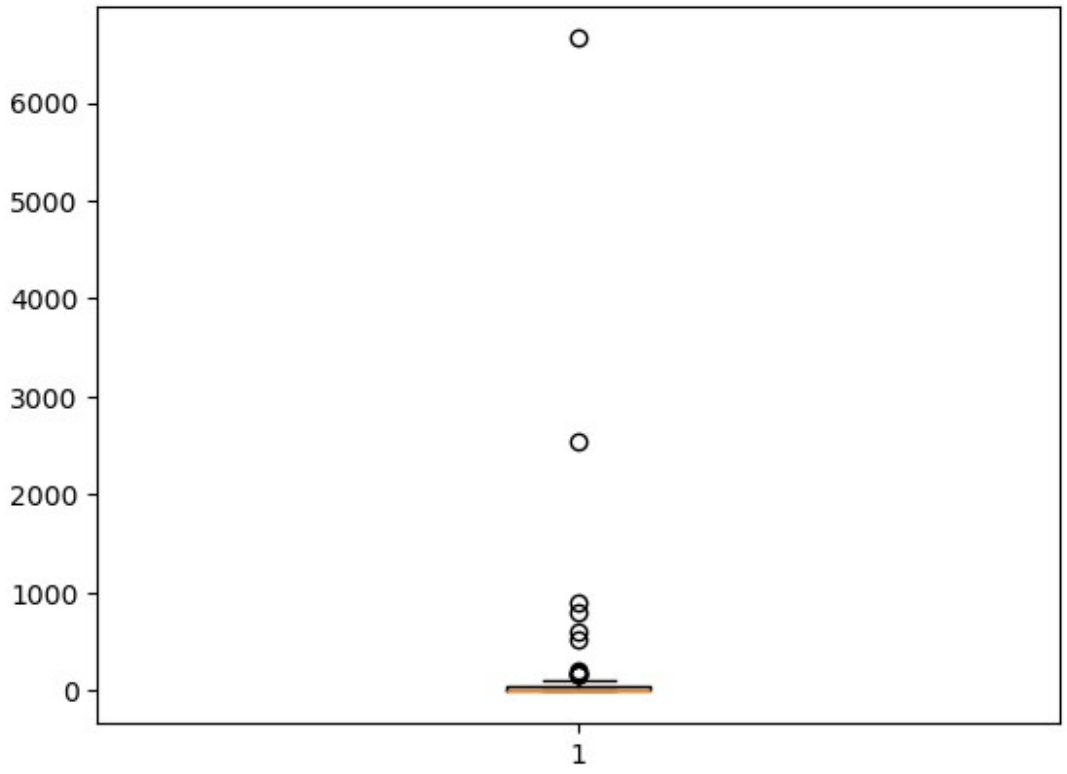



```
import plotly.express as px
px.box(df_animal, x="sleep_total")
```


```
plt.boxplot(df_animal.bodywt)
```

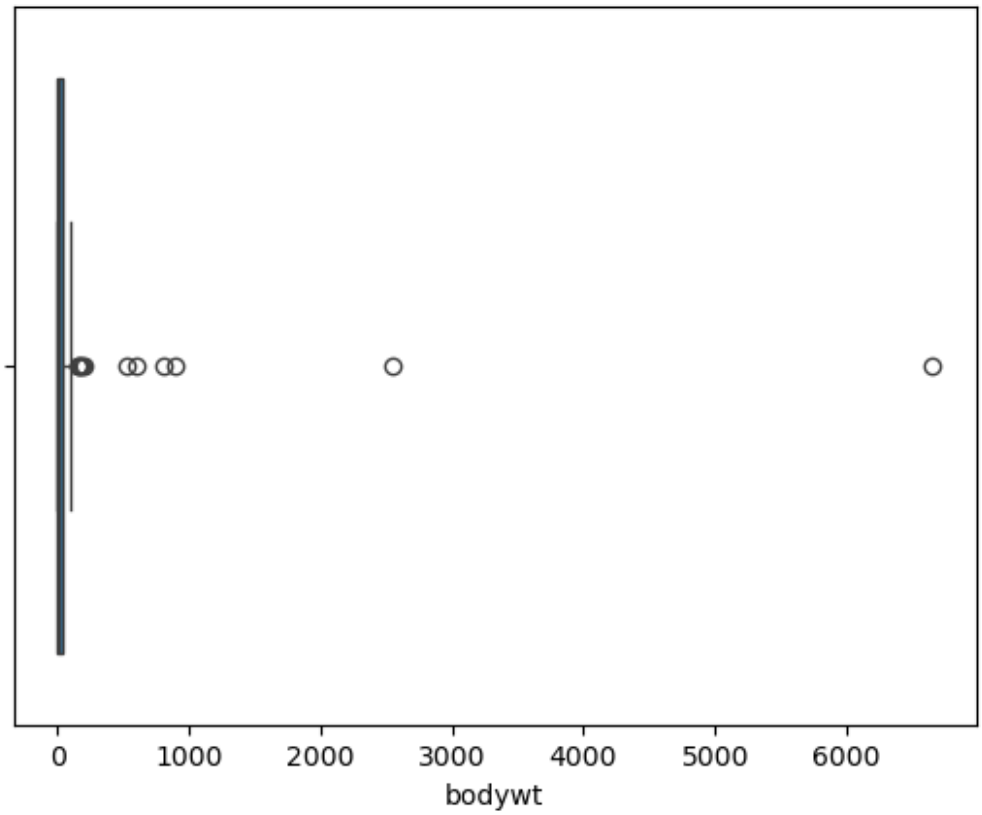
```
{'whiskers': [<matplotlib.lines.Line2D at 0x16e837f80>,
  <matplotlib.lines.Line2D at 0x16e86c290>],
 'caps': [<matplotlib.lines.Line2D at 0x16e86c590>,
```

```
 <matplotlib.lines.Line2D at 0x16e86c860>],
'boxes': [<matplotlib.lines.Line2D at 0x16e7dc770>],
'medians': [<matplotlib.lines.Line2D at 0x16e86cb60>],
'fliers': [<matplotlib.lines.Line2D at 0x16e86ce60>],
'means': []}
```

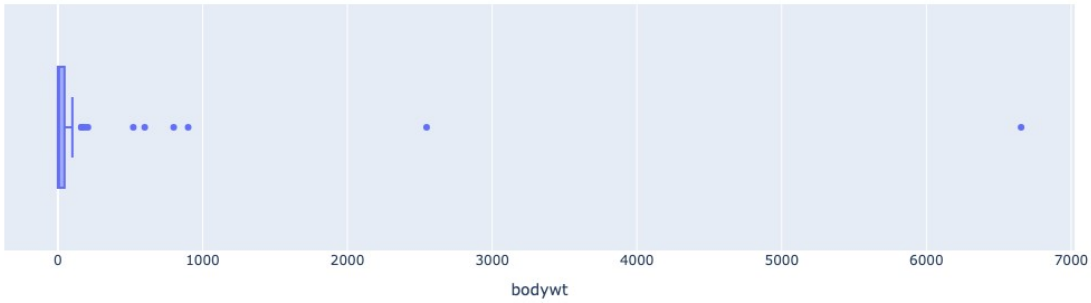



```
sns.boxplot(data=df_animal, x='bodywt')
```

```
<Axes: xlabel='bodywt'>
```




```
px.box(df_animal, x='bodywt')
```




```
import pandas as pd
df_sales = pd.read_csv("data/amir_deals.csv")
df_sales
```

	Unnamed: 0	product	client	status	amount	num_users
0	1	Product F	Current	Won	7389.52	19
1	2	Product C	New	Won	4493.01	43
2	3	Product B	New	Won	5738.09	87
3	4	Product I	Current	Won	2591.24	83

```
4             5  Product E  Current    Won  6622.97         17
..          ...        ...      ...    ...      ...        ...
173         174  Product A  Current   Lost  5835.32         23
174         175  Product D  Current    Won  6377.50         12
175         176  Product D  Current    Won  3537.61         28
176         177  Product A  Current    Won  6448.07         34
177         178  Product D      New   Lost  7320.05         72

[178 rows x 6 columns]
```

```
df_sales_users = df_sales.groupby('num_users')
['amount'].agg(sum='sum')
df_sales_users_benim = df_sales.groupby('num_users')
['amount'].agg(['sum','mean'])
df_sales_users
```

```
                sum
num_users
1          13624.50
2          40732.68
3          24858.82
4           3880.07
5          12428.48
...             ...
92          4509.96
94          4171.76
96          8180.81
98          5992.86
99         16750.45

[79 rows x 1 columns]
```

```
df_sales_users.sample(n=5, replace=True) #5 değer gelsin. replace true
olduğu için aldığını geri yerine koyar. yani aynı değer tekrar
gelebilir!
np.random.seed(42) #hep aynı değerin gelmesini sağlar ⬇
df_sales_users.sample()
```

```
               sum
num_users
44         5493.01
```

```
# week 2:
# sürekli dağılımlar: tam sayı olmadığı için yani olasılık tam sayı
ile ifade
#edilmedeiğinde bir çzigi şeklinde ifade edilir.
# TÜM OLASILIK DAĞILIMLARINDA ALTTA KALAN ALAN 1'DİR!
# bimodal dist: sürekli dağılımlar bazi değerlerinin daha yüksek
olduğu tekdüze olmayan biçimler alabilir. Bimodal Dağılım, bir veri
setinde
# iki farklı zirve veya mod bulunan bir olasılık dağılımıdır. Bimodal
```

```
terimi, genellikle iki önemli zirveye sahip olasılık dağılım grafiklerini
# tanımlamak için kullanılır. Bu dağılım, tipik olarak tek modlu (unimodal) dağılımlardan farklı olarak, iki farklı ortalamaya veya merkeze
# sahip iki grup veriyi ifade eder.

# Sürekli Düzgün Dağılım
from scipy.stats import uniform #altta kalan. başka dağılımlar için de geçerlidir bu
uniform.cdf(7,0,12)#7den küçük olma olasılığı, 0'dan 12'ye kadar!
```

```
0.5833333333333334
```

```
#7'den büyük değerler:
1 - uniform.cdf(7,0,12)
```

```
0.41666666666666663
```

```
#0 ile 10 arasında 5 tane rastgele sayı
uniform.rvs(0,10,size=5)
```

```
array([1.74954927, 9.82168343, 5.16635891, 2.60829175, 9.962537  ])
```

```
# BİNOM DAĞILIMI: bir bağımsız denemedeki başarı sayısının olasılığını tanımlar
# n= kaç defa p=başarılı olma olasılığı. dicrete yani kesikli bir dağılımdır.
# Kesikli dağılım: Belli ve sayılabilir sayıda değer alabilen dağılımdır.
# yine alan hesaplanır! expected value = ortalama değerdir -> binomda n * p 'dir!

# bir parayı atıyoruz, 0.5 olasılık var ve 1 defa atoıyoruz
from scipy.stats import binom
binom.rvs(1, 0.5, size=1)
```

```
array([1])
```

```
# 8 parayı havaya atıyoruz
binom.rvs(8, 0.5, size=1)
```

```
array([4])
```

```
# 3 tane parayı 10 kere havaya attığında kaç tane başarılı geldi?
binom.rvs(3, 0.5, size=10)
```

```
array([3, 1, 1, 2, 2, 2, 1, 3, 2, 3])
```

```
#bir tarafı daha ağır para. yüzde 25 tura
binom.rvs(3, 0.25, size=10)
```

```
array([1, 1, 1, 0, 2, 2, 0, 0, 1, 1])
```

```
# olasılık kütle fonksiyonu - kesikli dağılım. belli bir olasılıkla kesikli değer alma ihtimalini açıklar. kesikli bir rastgele değişkenin
# belirli bir değeri alma olasılığını hesaplar

#10 jetondan 7sinin yazı gelme
binom.pmf(7,10,0.5)
```

```
0.11718750000000004
```

```
#7 ve daha az gelme olasılığı
binom.cdf(7,10,0.5)
```

```
0.9453125
```

```
#7den daha fazla gelme olasılığı
1-binom.cdf(7,10,0.5)
```

```
0.0546875
```

```
# NORMAL DAĞILIM: en popüler. şekli çan eğrisi şeklindedir: tansiyon ve emeklilik yaşı. veriler normal dağılıma uydurulmaya çalışılır.
# her dağılımın cdf fonksiyonu kullanılır altta kalan alan hesaplanırken
# beklenen değeri ortalama değeridir!
# gerçek dünya verisi
# simetriktir.
# eğrinin altında kalan alan 1'e eşittik (tüm olasılık dağılımlarında)
# uçları öyle görünse de olasılık hiçbir zaman 0'a ulaşmaz
# orta çizgi ortalamayı (expected value)'yu gösterir!
# m (mü) = ort , v = standart sapma
# verilerin yüzde 68'i bir standart sapma altında ve üstündedir
# iki ss üstünde ve altında yüzde 95
# yüzde 99.7 3 ss altında ve üstündedir
# 3 ss altının da altındaysa outline (?) iqr yerine ss de kullanılabilir yani eğer veriler normal dağılıyorsa. Eğer veriler normal dağılıyorsa, veri setindeki aykırı değerleri belirlemek için üç standart sapma kuralı kullanılabilir.
# 68-95-99.7 -> 3ss ifade eder
# ort 0 ve ss 1 olan dağılıma standart-normal dağılım denir!

# ort 161 cm ve ss 7 olan , 154den kısa kadınlar
from scipy.stats import norm
norm.cdf(154,161,7)
```

```
0.15865525393145707
```

```
#154den uzun
1-norm.cdf(154,161,7)
```

```
0.8413447460685429
```

```
#ort 70 ss 10 12 veri
norm.rvs(70,10,12)
```

```
array([69.88556553, 74.13354088, 68.59101446, 78.72161859,
83.00345148,
       56.56476236, 78.50120582, 73.94339951, 69.95304005,
54.29188016,
       84.79848775, 73.96774281])
```

```
#154 157 arasındakiler
norm.cdf(157,161,7) - norm.cdf(154,161,7)
```

```
0.1251993291672192
```

```
# ppf -> ters kümülatif dağılım
# belirli bir olasılığa karşılık gelen x'i bulmaya yarar. verilen bir
olasılığa karşılık gelen değeri bulur
#kadınların yüzde 90ı 169.97den kısadır
norm.ppf(0.9, 161,7)
```

```
169.9708609588122
```

```
#kadınların yüzden 90ı şu boydan uzundur
norm.ppf((1-0.9), 161, 7)
```

```
152.0291390411878
```

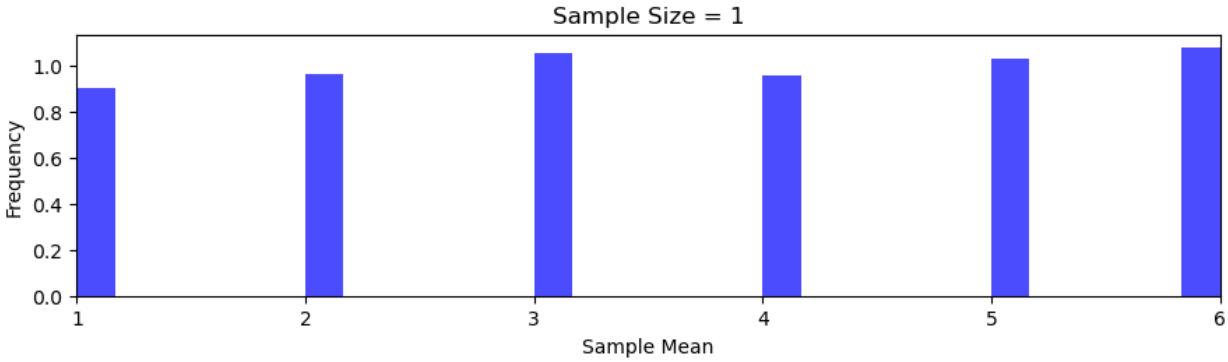
```
#10 tane random veri
norm.rvs(161,7,10)
```

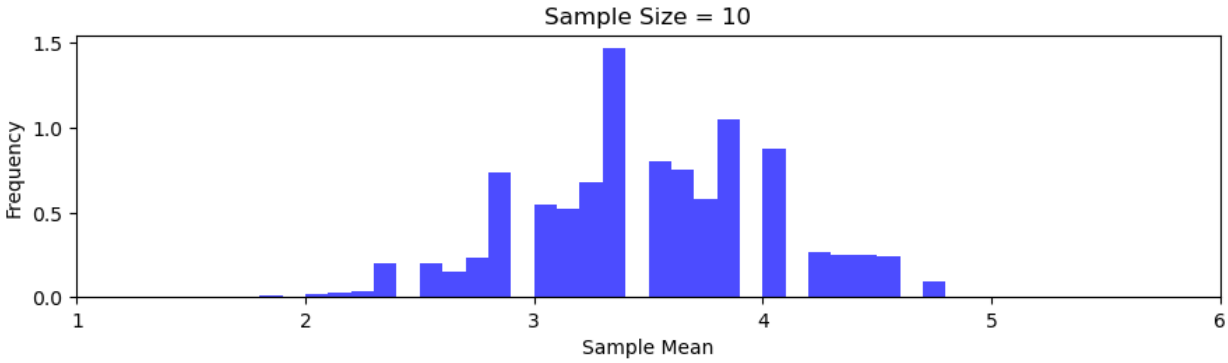
```
array([148.51724256, 159.29805878, 163.2905637 , 164.54602711,
       160.10676644, 163.39113564, 175.83220349, 171.82575729,
       154.35336668, 156.82171868])
```

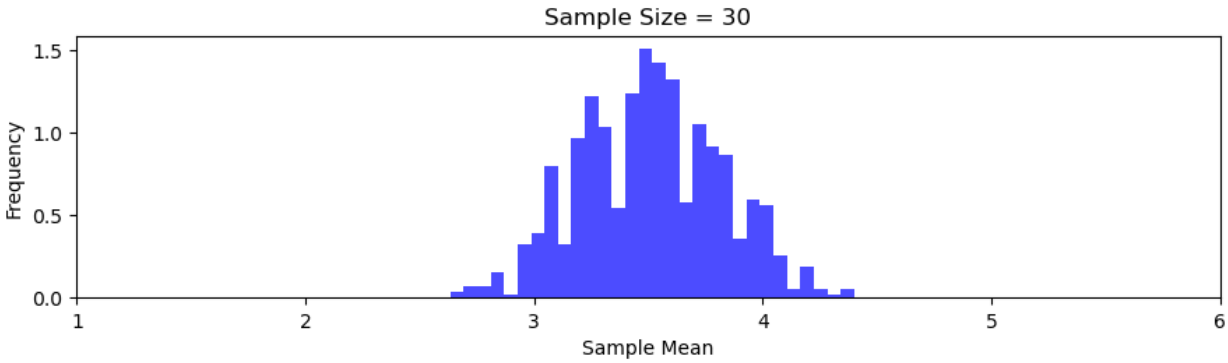
```
# çarpıklık -> veri simetrik değilse pozitif çarpık ya da negatif
çarpık
# sağdan tokat attıysam sağ çarpık -> pozitif çarpık. soldan tokat
atıysaç sol çarpık -> negatif çarpık
# basıklık -> değılımdaki aşırı değerleri açıklama. 3 çeşit
# 1. Mesokurtik (Mesokurtic):
# Mesokurtik dağılımlar, normal dağılıma benzer bir basıklığa sahiptir
(kurtosis değeri yaklaşık 3).
# Tepe ve kuyrukları, normal dağılıma kıyasla ne fazla sivriliğe ne de
düzlüğe sahiptir.
# Bu tür bir dağılım, aşırı değerler açısından ne çok riskli ne de çok
güvenli olarak değerlendirilir.
# 2. Leptokurtik (Leptokurtic):
# Leptokurtik dağılımlar, normal dağılımdan daha sivri tepeye ve daha
ağır kuyruklara sahip olan dağılımlardır (kurtosis değeri 3'ten
büyük).
# Bu tür dağılımlar, normal dağılıma göre daha fazla aşırı değer
```

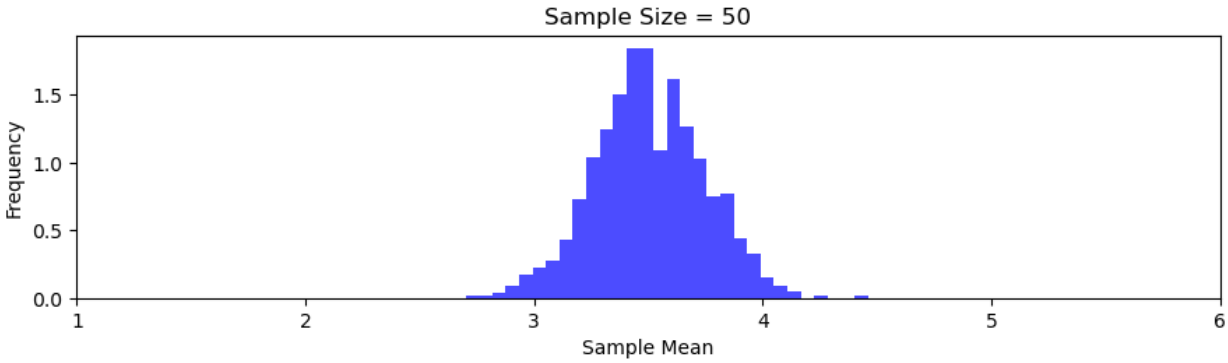
```
içerir.
# Aşırı değerlerin varlığı, potansiyel olarak riskli durumları veya
veri setindeki özel durumları işaret edebilir.
# 3. Platykurtik (Platykurtic):
# Platykurtik dağılımlar, normal dağılıma göre daha düz bir tepeye ve
daha hafif kuyruklara sahip olan dağılımlardır (kurtosis değeri 3'ten
küçük).
# Bu tür dağılımlar, normal dağılıma kıyasla daha az aşırı değer
içerir.
# Daha düz tepe, verilerin daha homojen dağıldığını ve aşırı
değerlerin daha az sıklıkta olduğunu gösterir.

# merkezi limit teoremi: örneklem büyüklüğü arttıkça normal dağılıma
yaklaşır
# bir zar atıldığında her yüzün değeri ile gelme olasılığı toplanırsa
-> 3.67 expected value
# Örneklem dağılımı (sampling distribution):
# bir popülasyondan alınan tüm olası örneklerin bir istatistik
(örneğin, ortalama, medyan veya oran) için dağılımıdır.
# Örneklem Dağılımının Özellikleri
# Ortalama: Örneklem dağılımının ortalaması, genellikle popülasyonun
gerçek ortalamasına yakındır. Bu özellik, merkezi limit teoremi ile
ilgilidir;
# örneklem büyüklüğü arttıkça örneklem ortalamasının dağılımı normal
dağılıma yaklaşır.
# Merkezi Limit Teoremi: Örneklem büyüklüğü yeterince büyük olduğunda
(genellikle n>30), örneklem ortalamalarının dağılımı,
# popülasyon dağılımı ne olursa olsun yaklaşık olarak normal dağılıma
benzer. Bu özellik, örneklem dağılımlarının birçok istatistiksel
# analizde temelini oluşturur.
#(merkezi limit teoremi örneği):
import numpy as np
import matplotlib.pyplot as plt

# Zar atma fonksiyonu, 1'den 6'ya kadar olan sayıları eşit olasılıkla
üretir
def throw_dice(n):
    return np.random.randint(1, 7, size=n)

# Örneklem ortalamalarını hesaplama
def sample_means(sample_size, total_samples):
    means = []
    for _ in range(total_samples):
        samples = throw_dice(sample_size)
        means.append(np.mean(samples))
    return means

# Zar atma simülasyonu
sample_size = [1, 10, 30, 50, 100]  # Farklı örneklem büyüklükleri
total_samples = 1000  # Her örneklem büyüklüğü için 1000 kez tekrarla
```

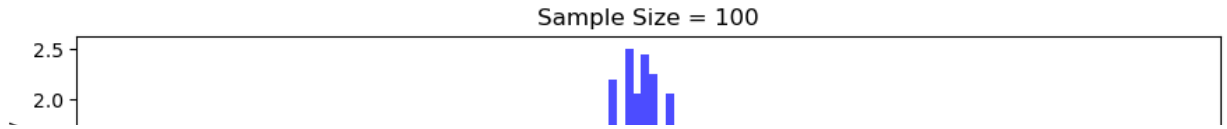
```
fig, axs = plt.subplots(len(sample_size), 1, figsize=(10, 15))
fig.tight_layout(pad=5.0)

for i, size in enumerate(sample_size):
    means = sample_means(size, total_samples)
    axs[i].hist(means, bins=30, density=True, color='blue', alpha=0.7)
    axs[i].set_title(f'Sample Size = {size}')
    axs[i].set_xlim([1, 6])
    axs[i].set_xlabel('Sample Mean')
    axs[i].set_ylabel('Frequency')

plt.show()
```


Sample Size = 1
Frequency
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
Sample Mean


Sample Size = 10
Frequency
0.0
0.5
1.0
1.5
1
2
3
4
5
6
Sample Mean


Sample Size = 30
Frequency
0.0
0.5
1.0
1.5
1
2
3
4
5
6
Sample Mean


Sample Size = 50
Frequency
0.0
0.5
1.0
1.5
1
2
3
4
5
6
Sample Mean

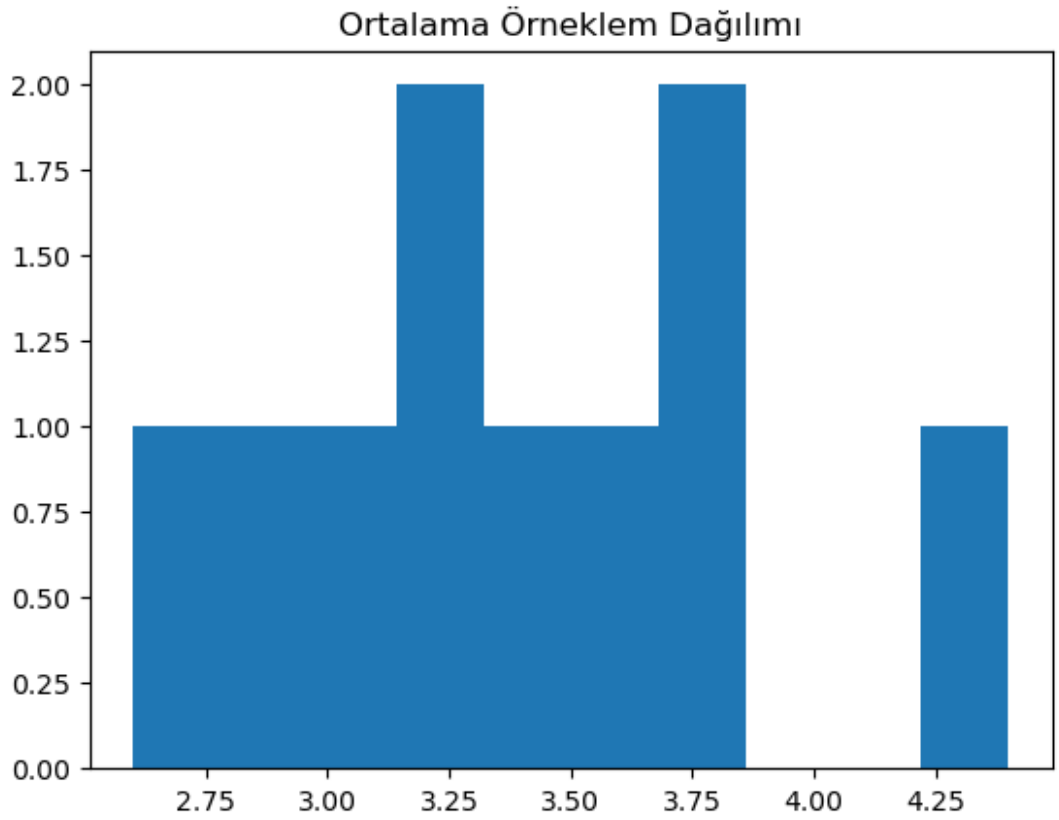

Sample Size = 100
2.0
2.5

```
import pandas as pd
import numpy as np
die= pd.Series([1,2,3,4,5,6])
sample = die.sample(n=5, replace=True)
sample
```

```
3    4
1    2
0    1
2    3
1    2
dtype: int64
```

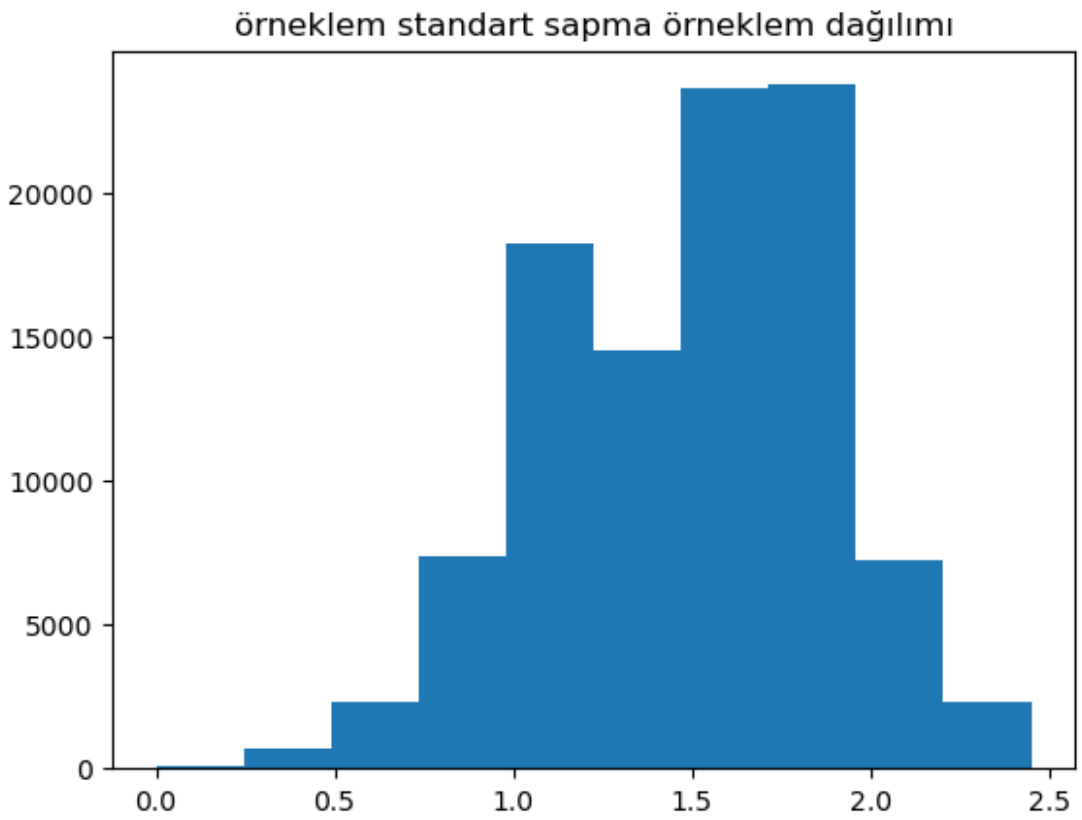
```
#yukarıdaki işlem 10 kere tekrar edilsin. ÖRNEKLEM DAĞILIMI
import matplotlib.pyplot as plt
sample_ = []
for i in range(10):
    sample = die.sample(n=5, replace=True)
    sample_.append(sample.mean()) #ortalama için yapacağız o yüzden
ortalaması
plt.title('Ortalama Örneklem Dağılımı')
plt.hist(sample_)
```

```
(array([1., 1., 1., 2., 1., 1., 2., 0., 0., 1.]),
 array([2.6 , 2.78, 2.96, 3.14, 3.32, 3.5 , 3.68, 3.86, 4.04, 4.22,
4.4 ]),
 <BarContainer object of 10 artists>)
```




```
sample_sts = []
for i in range(100000):
    sample = die.sample(5, replace=True)
    sample_sts.append(np.std(sample))
plt.title("örneklem standart sapma örneklem dağılımı")
plt.hist(sample_sts)
```

```
(array([   69.,   681.,  2284.,  7352., 18229., 14508., 23633.,
23741.,
         7219.,  2284.]),
 array([0.        , 0.24494897, 0.48989795, 0.73484692, 0.9797959 ,
        1.22474487, 1.46969385, 1.71464282, 1.95959179, 2.20454077,
        2.44948974]),
 <BarContainer object of 10 artists>)
```




```
# poisson süreci : belirli bir zaman dilimindeki ortalama olay
sayısının bilindiği ancakk olaylar arasındaki zaman veya boşluğun
rastgele olduğu bir süreçtir
# hangi aralıklarla geldiğini bilmezsiniz
# lambda ile ifade edilir
# zaman periyodu başına ortalama olay sayısı -> lambda
# dağılımın beklenen değeri -> ort !!
# discreed -> kesikli bir dağılımdır
# lamda dağılımın şeklini değiştirir -> basıklığını değiştirir yani.
# sample sayısı büyüdükçe poisson dağılımı olarak normal dağılıma
benzer -> merkezi limit teoremi
from scipy.stats import poisson
#haftada ort 8 sahiplenme olan bir yerde haftada 5 sahiplenme
poisson.pmf(5,8)
```

```
0.09160366159257921
```

```
# 5 veya daha az
poisson.cdf(5,8)
```

```
0.1912360620796254
```

```
#5ten fazla
1-poisson.cdf(5,8)
```

```
0.8087639379203746
```

```
# 10 tane rastgele değer
poisson.rvs(8, size=10)
```

```
array([ 2,  5,  8,  8, 15,  6, 13,  4,  5,  7])
```

```
# üstel dağılım -> poisson olaylarında belirli bir zaman geçme olasılığını temsil eden dağılımdır
# lambda beklenen değerdir! expected value
# poissonun aksine zaman belirttiği için continous bir distrndır
# 2 dakikada 1 bilet ise periyot 0.5tir. posiondaki varsa 1/lambdadır periyot
from scipy.stats import expon
# 2 dakikada 1 bilet .yeni bir istek için 1 dakikadan az bekleme olasılığı
expon.cdf(1,scale=2)
```

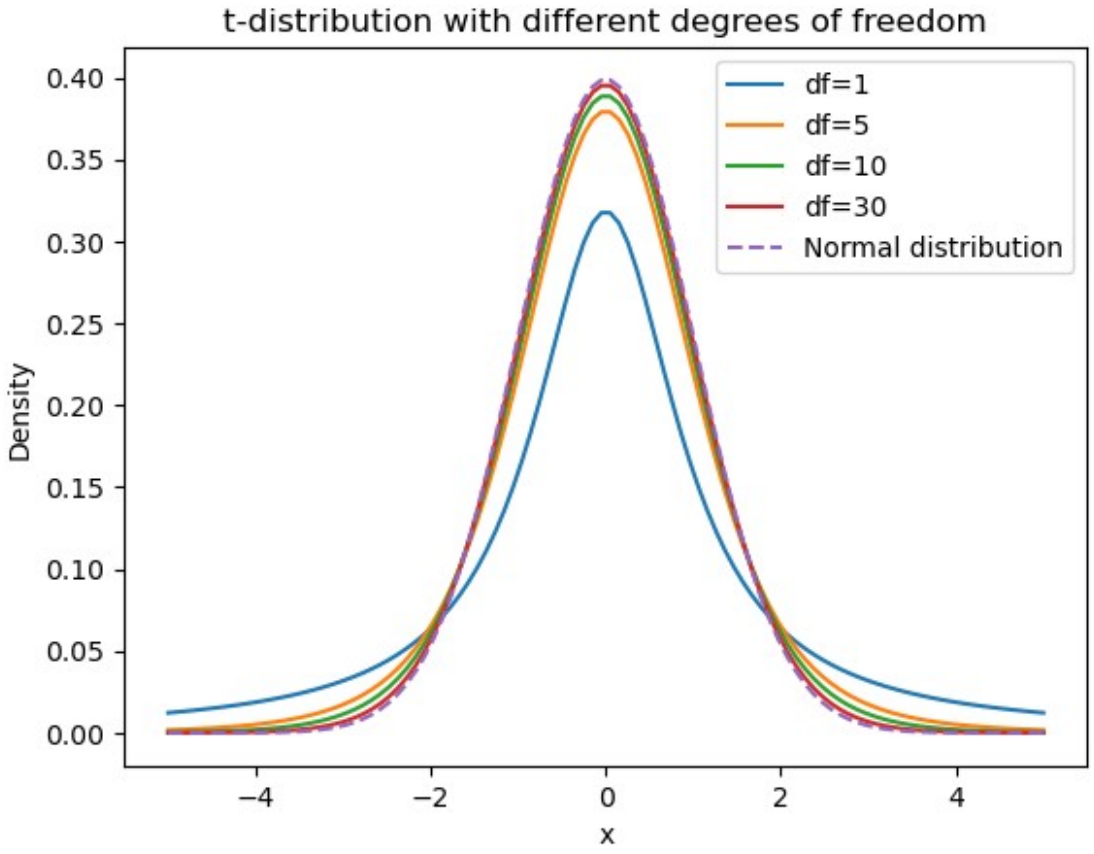
```
0.3934693402873666
```

```
#4 dakikadan fazla bekleme olasılığı
1-expon.cdf(4, scale=2)
```

```
0.1353352832366127
```

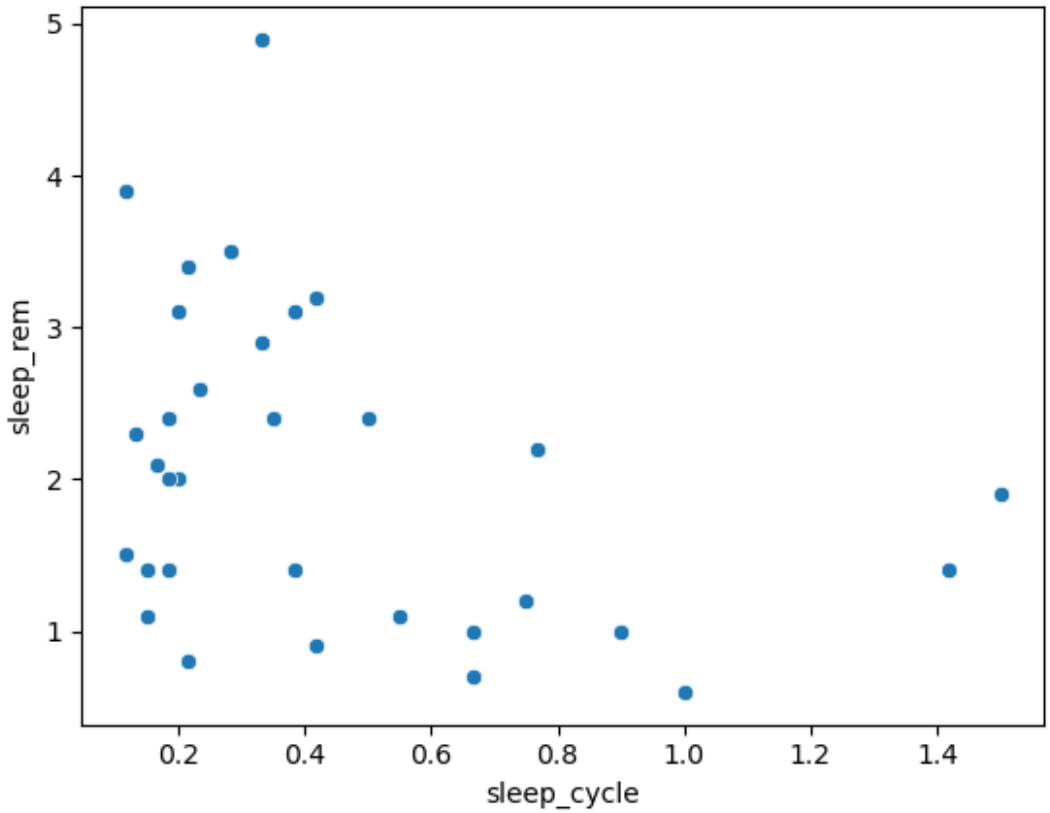
```
# t dağılımı -> küçük örneklem veya popülasyonun standart sapması bilinmiyorsa kullanılır.
# t dağılımı, ortalaması sıfır olan ve simetrik bir eğriye sahiptir. Ancak, küçük örneklem büyüklüklerinde (n < 30) normal dağılıma göre daha geniş kuyruklara sahiptir. Bu, küçük örneklemde aşırı değerlere daha fazla tolerans gösterildiği anlamına gelir.
# normale benzer.
# serbestlik derecesi -> istatistiksel bir parametreyi kullanılabilen bağımsız bir değer veya bilgi parçalarının değeri. dağılımın kuyruğunun ne kadar geniş ve düz olacağını.
# 30 veya daha yakın olduğunda normal dağılıma yaklaşır
# büyüklük çok önemli (n-1) olur. n: örneklemin büyüklüğü
# t Dağılımı: Küçük örneklem büyüklüklerinde veya popülasyon standart sapması bilinmediğinde kullanılan dağılımdır.
# Serbestlik Derecesi (n - 1): Örneklem büyüklüğüne dayalıdır ve t dağılımının şeklini belirler. Serbestlik derecesi arttıkça t dağılımı normal dağılıma yaklaşır.
# Normal Dağılıma Yakınlık: Örneklem büyüklüğü 30 veya daha fazla olduğunda, t dağılımı neredeyse normal dağılım gibi davranır.

import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

x = np.linspace(-5, 5, 100)
```

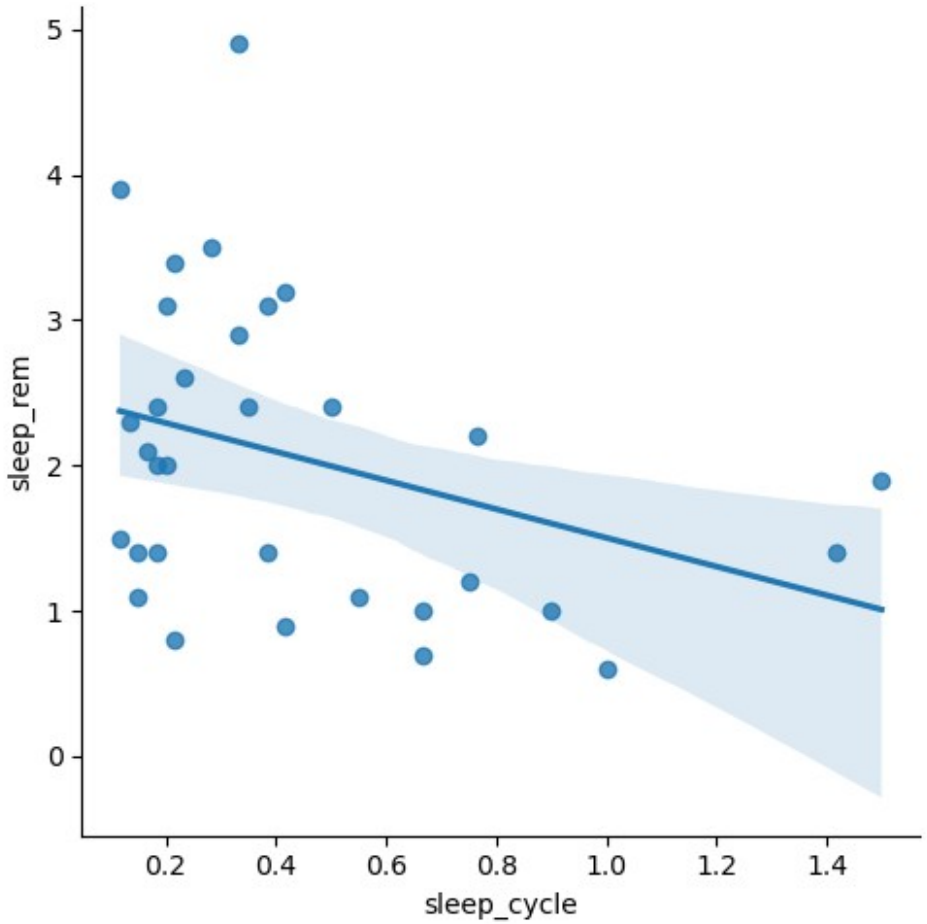
```python
# Plot for different degrees of freedom
for df in [1, 5, 10, 30]:
    plt.plot(x, stats.t.pdf(x, df), label=f'df={df}')

plt.plot(x, stats.norm.pdf(x), label='Normal distribution',
         linestyle='--')
plt.legend()
plt.title('t-distribution with different degrees of freedom')
plt.xlabel('x')
plt.ylabel('Density')
plt.show()
```




```python
# WEEK 4
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt

df_sleep = pd.read_csv("data/msleep.csv")
df_sleep['sleep_cycle'].corr(df_sleep['sleep_rem']) #negatif zayıf
ilişki
sns.scatterplot(x='sleep_cycle', y='sleep_rem', data=df_sleep)
sns.lmplot(x='sleep_cycle', y='sleep_rem', data=df_sleep)
```

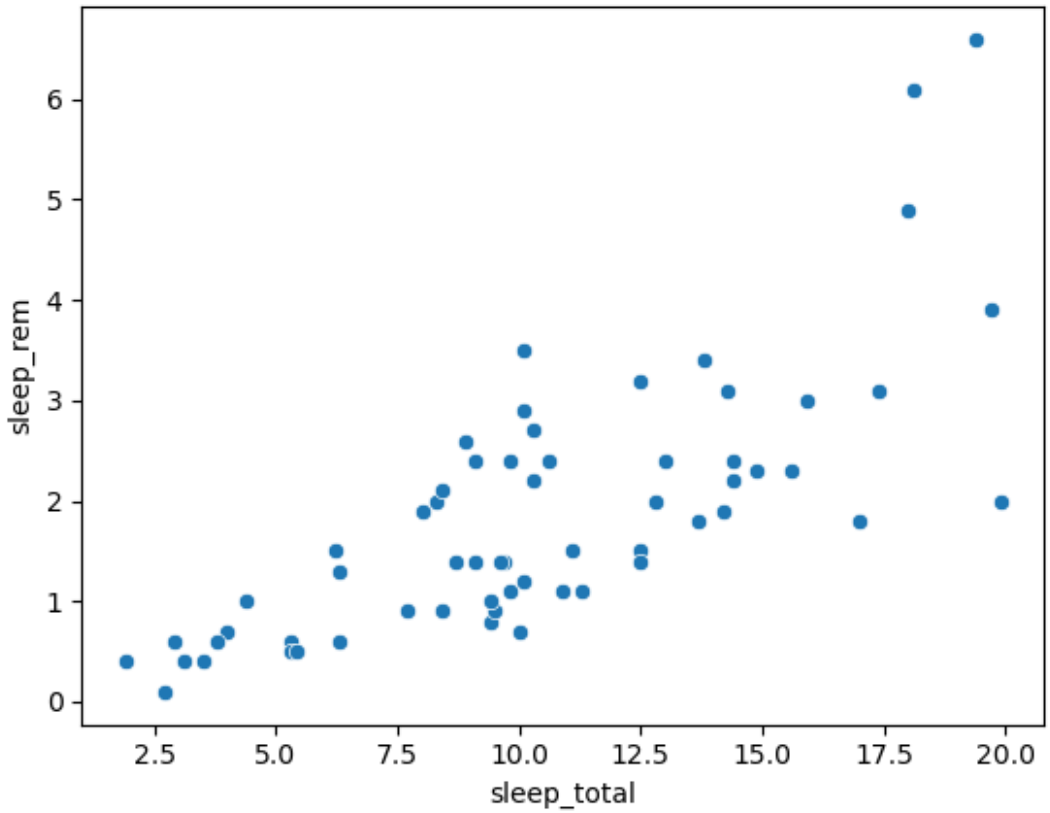
```
<seaborn.axisgrid.FacetGrid at 0x159d37ad0>
```

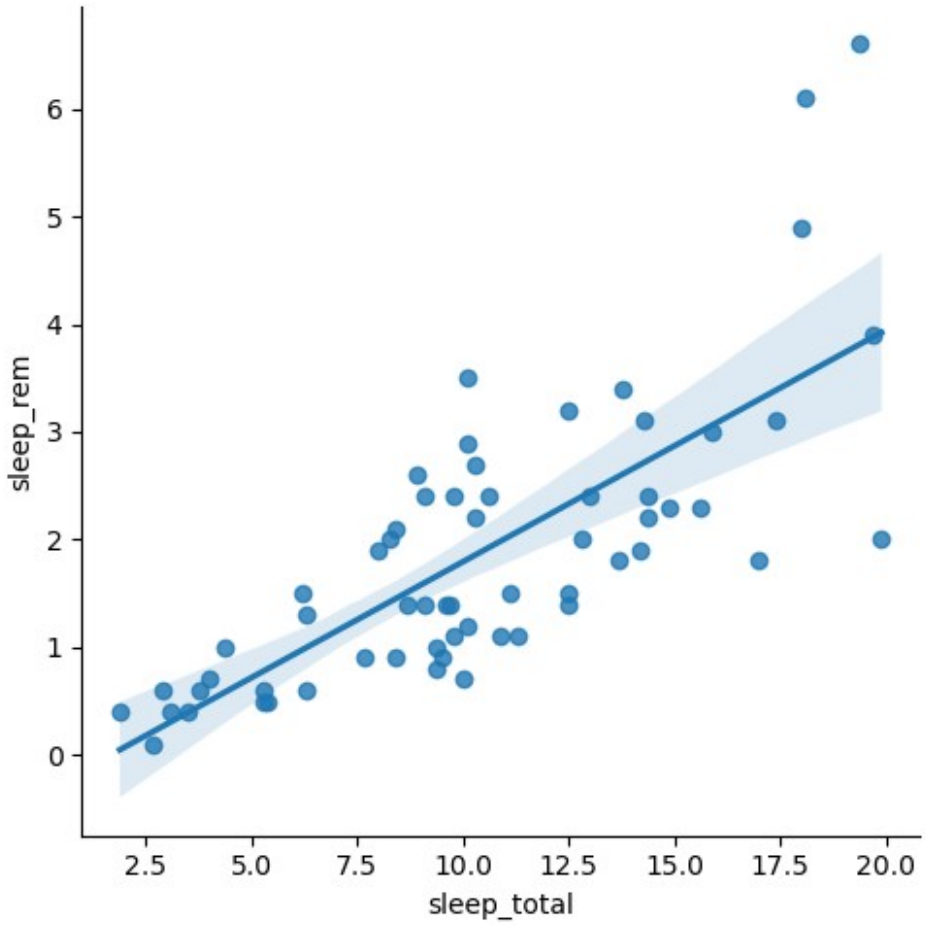





```
df_sleep['sleep_total'].corr(df_sleep['sleep_rem']) #aralarında güçlü pozitif ilişki var
sns.scatterplot(x='sleep_total', y='sleep_rem',data=df_sleep)
sns.lmplot(x='sleep_total', y='sleep_rem', data=df_sleep)
```

```
<seaborn.axisgrid.FacetGrid at 0x178097d70>
```


sleep_rem
sleep_total
0
1
2
3
4
5
6
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

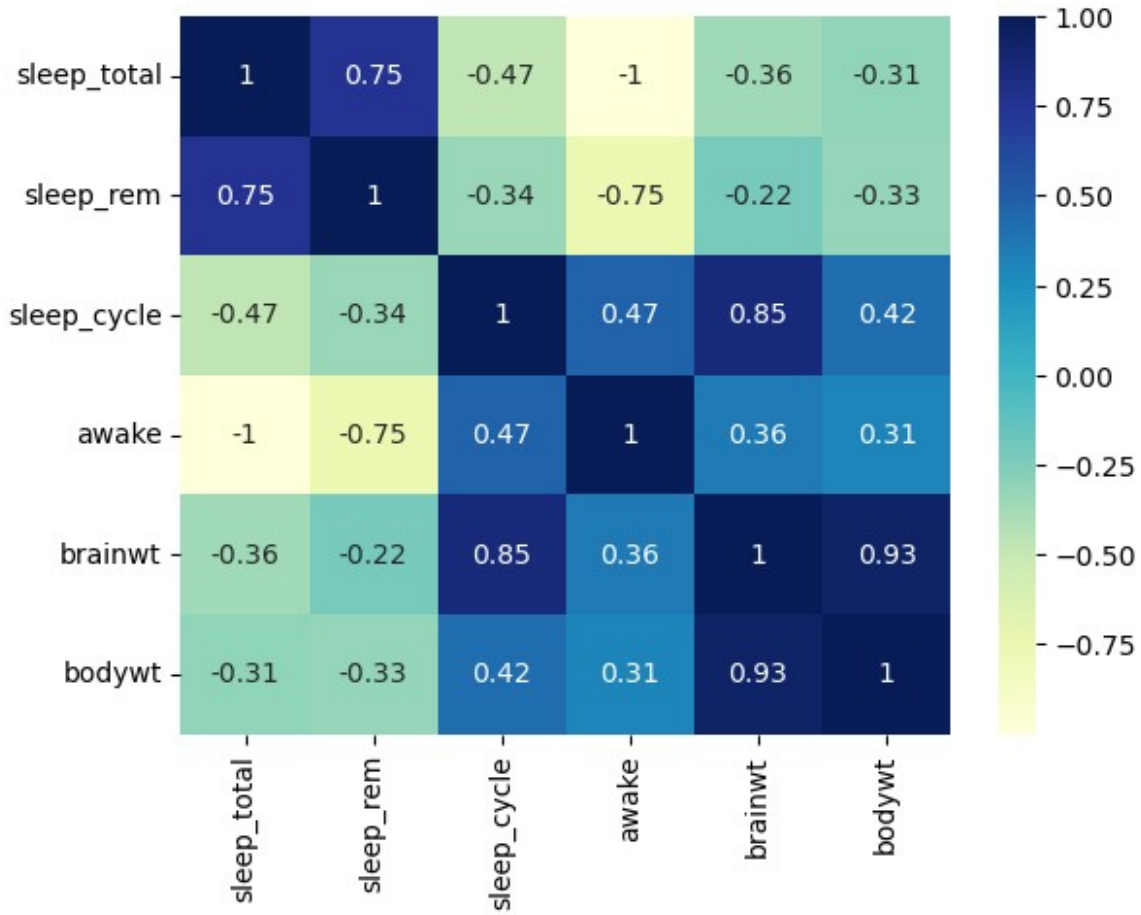



```
df_sleep.corr(numeric_only=True) #her sütünu birbiri ile corr işlemine tabii tutar. numeric_pnly sadece sayısal değelerin katılmasını sağlar bu işleme!
```

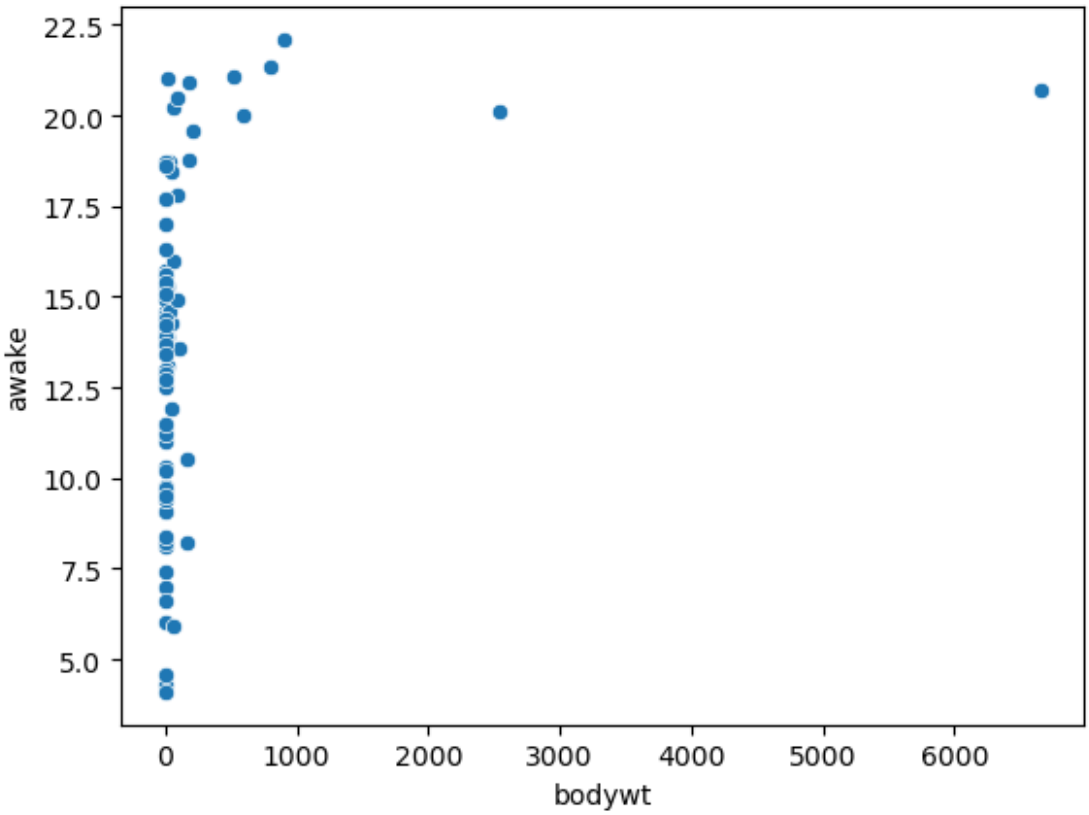
```
              sleep_total  sleep_rem  sleep_cycle     awake   brainwt
bodywt
sleep_total      1.000000   0.751755    -0.473713 -0.999999 -0.360487 -
0.312011
sleep_rem        0.751755   1.000000    -0.338123 -0.751771 -0.221335 -
0.327651
sleep_cycle     -0.473713  -0.338123     1.000000  0.473713  0.851620
0.417803
awake           -0.999999  -0.751771     0.473713  1.000000  0.360487
0.311980
brainwt         -0.360487  -0.221335     0.851620  0.360487  1.000000
0.933782
bodywt          -0.312011  -0.327651     0.417803  0.311980  0.933782
1.000000
```

```
sns.heatmap(df_sleep.corr(numeric_only=True),annot=
True,cmap="YlGnBu")
```

```
<Axes: >
```



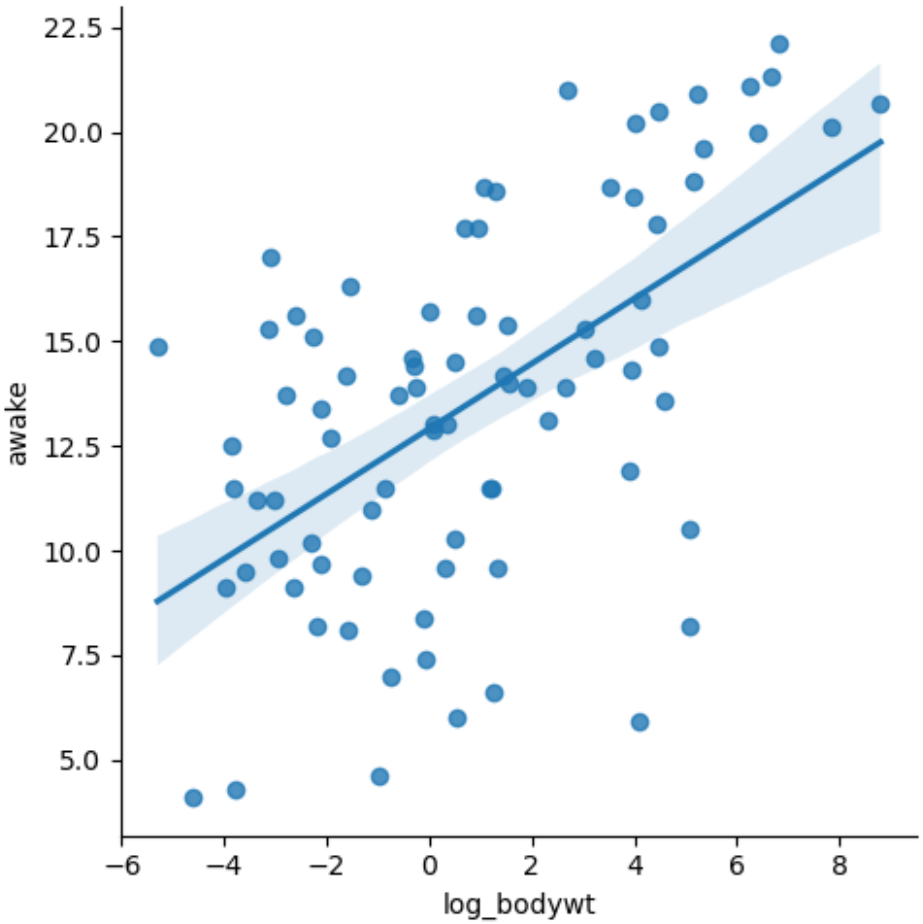

```
df_sleep['bodywt'].corr(df_sleep['awake']) #buna bakarsan aslında düşük pozitif ilişki var dersin
sns.scatterplot(x='bodywt', y='awake', data=df_sleep) #ancak grafiğe bakıldığında aralarında lineer bir ilişki yok!

<Axes: xlabel='bodywt', ylabel='awake'>
```




```
# LOGNORMAL DAĞILIM:  Log-normal dağılım, bir değişkenin logaritması
alındığında normal dağılım gösteren bir dağılımdır. Yani, bir değişken
# logaritması alındığında simetrik ve çan şeklinde bir dağılım (normal
dağılım) elde ediliyorsa, bu değişken log-normal dağılıma sahiptir.
# Lognormal dağılım ve logaritmik dönüşüm, özellikle dağılımın
asimetrik veya çarpık olduğu durumlarda ilişkiyi
# lineerleştirmek için kullanılır.
# Logaritmik dönüşüm, verileri logaritma ölçeğinde dönüştürerek,
dağılımın şeklini değiştirme amacı taşır. Özellikle, büyük değerler
ile küçük değerler
# arasındaki fark çok fazlaysa (örneğin, bodywt sütununda olduğu gibi
vücut ağırlıkları çok geniş bir aralıkta dağılıyorsa),
# logaritmik dönüşüm veriyi sıkıştırarak daha simetrik hale
getirebilir.
# Neden Logaritmik Dönüşüm Yapılır?
# Doğrusal İlişkiyi Ortaya Çıkarmak: Logaritmik dönüşüm, doğrusal
olmayan bir ilişkiyi doğrusal bir ilişki gibi göstererek analiz etmeyi

# kolaylaştırabilir.
# Aykırı Değerleri Azaltmak: Büyük değerler küçültülerek dağılımdaki
uç değerlerin etkisi azaltılır.
# Simetri Sağlamak: Veriler sağa çarpık veya asimetrik olduğunda, log
dönüşümü dağılımı daha simetrik hale getirebilir.
# Log-Normal Dağılımın Özellikleri
```

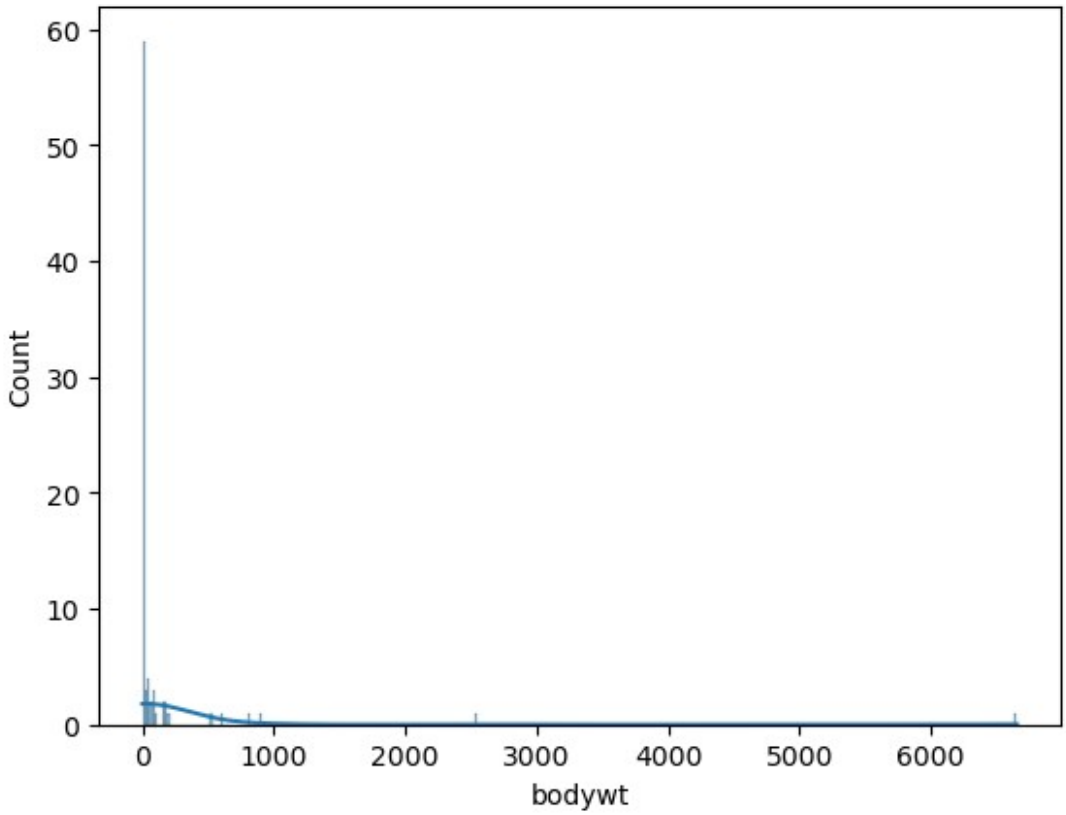
```
# Standard Sapma (s): Bu, log-normal dağılımda değişkenin
logaritmasının standart sapmasını gösterir. Yani, log(X) normal
dağılıma uyduğu
# için, log(X)'in standart sapması s olur.
# Ortalama (mean): Logaritması alınmış veri (log(X)) normal dağılıma
uyduğu için, log(X)'in ortalaması mean olur. Yani, log(X)'in
ortalaması
# aslında log-normal dağılımın ortalama değeridir.
# Ölçek Parametresi (scale = e^mean): Bu, log-normal dağılımın
genişliğini kontrol eder. Eğer mean değeri yüksekse, e^mean büyür ve
dağılım
# sağa doğru genişler. Bu durumda, log-normal dağılımda büyük
değerlerin görülme ihtimali artar.

#lineerleştirelim:
import numpy as np
import seaborn as sns
df_sleep['bodywt'].corr(df_sleep['awake'])
df_sleep['log_bodywt'] = np.log(df_sleep['bodywt'])
sns.lmplot(x='log_bodywt', y='awake', data=df_sleep)
```

```
<seaborn.axisgrid.FacetGrid at 0x159e8b2c0>
```



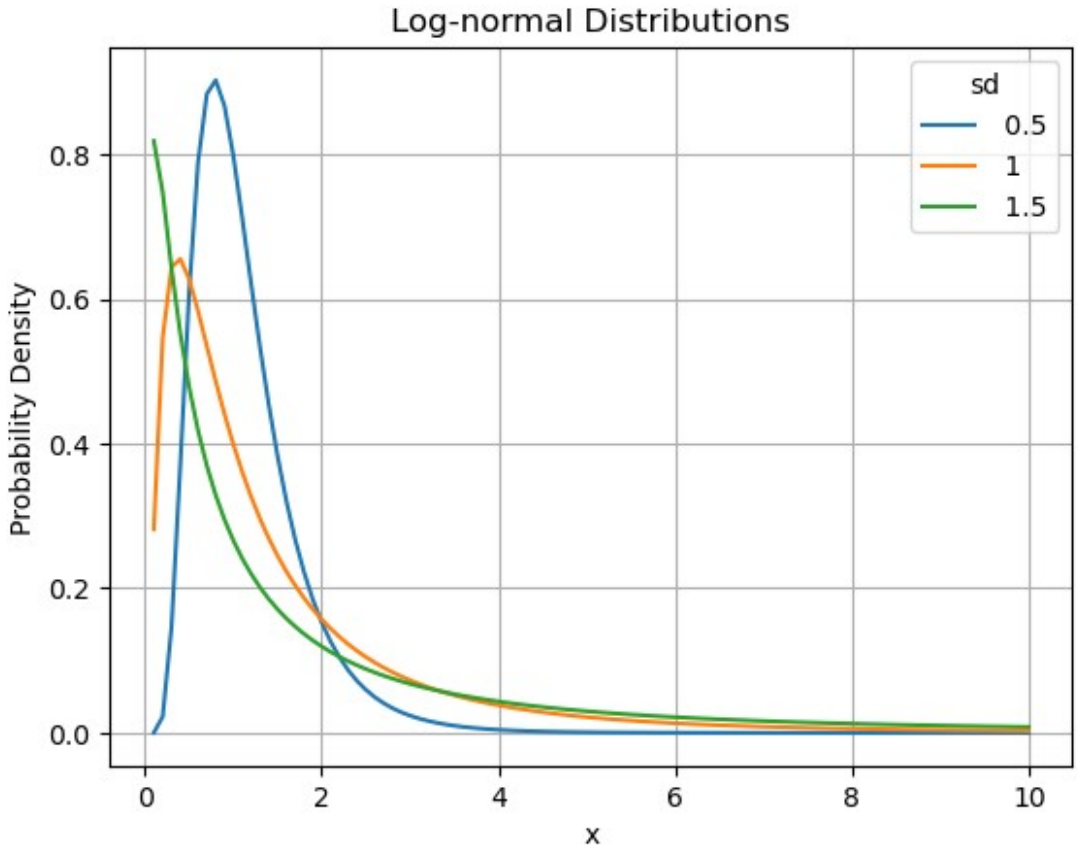

```
sns.histplot(data=df_sleep,x="bodywt", kde=True) #KDE, veri setinin
sürekli bir yoğunluk tahminini görselleştirmek için kullanılır.
#kde=True parametresi eklenmesi, histogramın üzerine bir çekirdek
yoğunluk tahmini çizilmesini sağlar. Bu yoğunluk grafiği, veri
dağılımının
#daha pürüzsüz bir şekilde görüntülenmesine olanak tanır ve özellikle
veri setindeki yoğunlukların nerede olduğunu gösterir.
```

```
<Axes: xlabel='bodywt', ylabel='Count'>
```




```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import lognorm

# X ekseni için değerler
x = np.linspace(0.1, 10, 100)


# Farklı log-normal dağılımlar için parametreler
params = [
    {'s': 0.5, 'scale': np.exp(0)},  # dar
    {'s': 1, 'scale': np.exp(0)},  # Orta
    {'s': 1.5, 'scale': np.exp(0)},  # geniş
]

for param in params:
    pdf = lognorm.pdf(x, param['s'], loc=0, scale=param['scale'])
    plt.plot(x, pdf, label=f" {param['s']}")

# Grafik ayarları
plt.title('Log-normal Distributions')
plt.xlabel('x')
plt.ylabel('Probability Density')
plt.legend(title="sd")
```

```
plt.grid(True)
plt.show()
```




```
# week 5
import pandas as pd

df_sleep = pd.read_csv("data/msleep.csv")
df_sleep
```

```
                          name       genus   vore         order
conservation  \
0                      Cheetah    Acinonyx  carni     Carnivora
lc
1                  Owl monkey       Aotus   omni      Primates
NaN
2              Mountain beaver  Aplodontia  herbi      Rodentia
nt
3   Greater short-tailed shrew     Blarina   omni  Soricomorpha
lc
4                          Cow         Bos  herbi  Artiodactyla
domesticated
..                         ...         ...    ...           ...
...
```

```
78                  Tree shrew      Tupaia   omni    Scandentia
NaN
79        Bottle-nosed dolphin    Tursiops  carni       Cetacea
NaN
80                       Genet     Genetta  carni     Carnivora
NaN
81                  Arctic fox      Vulpes  carni     Carnivora
NaN
82                     Red fox      Vulpes  carni     Carnivora
NaN

    sleep_total  sleep_rem  sleep_cycle  awake  brainwt   bodywt
0          12.1        NaN          NaN   11.9      NaN   50.000
1          17.0        1.8          NaN    7.0  0.01550    0.480
2          14.4        2.4          NaN    9.6      NaN    1.350
3          14.9        2.3     0.133333    9.1  0.00029    0.019
4           4.0        0.7     0.666667   20.0  0.42300  600.000
..          ...        ...          ...    ...      ...      ...
78          8.9        2.6     0.233333   15.1  0.00250    0.104
79          5.2        NaN          NaN   18.8      NaN  173.330
80          6.3        1.3          NaN   17.7  0.01750    2.000
81         12.5        NaN          NaN   11.5  0.04450    3.380
82          9.8        2.4     0.350000   14.2  0.05040    4.230

[83 rows x 11 columns]
```

```
# Null Hipotezi (H0)
# Tanım: Null hipotezi, bir araştırmada başlangıçta doğru kabul edilen
varsayımdır. Genellikle, "bir etki veya ilişki yoktur" şeklinde
#formüle edilir.
# Amacı: Hipotez testlerinde, null hipotez reddedilmeye çalışılır.
Eğer reddedilirse, alternatif hipoteze (H1) geçiş yapılır.

# p-Değeri (p-value)
# p-değeri, gözlemlenen verilerin null hipotezin doğru olduğu
varsayımı altında elde edilme olasılığını gösterir.
# Kullanımı: Bir hipotez testinde, p-değeri anlamlılık seviyesinden
küçükse (genellikle 0.05), null hipotezi reddederiz. Bu durumda,
# elde edilen sonucun istatistiksel olarak anlamlı olduğu düşünülür.
# Örnek: Bir ilaç testinde p-değeri 0.03 olarak hesaplandıysa, bu,
ilacın etkisiz olduğu varsayımı altında (null hipotez) bu sonucun elde

# edilme olasılığının %3 olduğu anlamına gelir.

# Anlamlılık Seviyesi (α - Alpha)
# Tanım: Anlamlılık seviyesi, bir hipotez testinde yanılma riskini
ifade eder ve genellikle %5 (0.05) olarak kabul edilir.
# Bu, test sonucunda %5 hata yapmayı kabul ettiğimiz anlamına gelir.
# Kullanımı: p-değeri, anlamlılık seviyesinden küçükse null hipotez
reddedilir.
```

```
# Örnek: Anlamlılık seviyesi 0.05 olarak belirlendiyse ve p-değeri
0.04 çıktıysa, null hipotez reddedilir. Bu durumda sonuç, %5'ten düşük
bir
# hata olasılığı ile istatistiksel olarak anlamlı kabul edilir.

# Type 1 Hatası (Yanlış Pozitif / Alpha Hatası):

# Tanım: Gerçekte doğru olan bir null hipotezi reddetme hatasıdır.
Yani, aslında bir etki yokken var olduğunu düşünmektir.
# Sonuç: Yanlış bir şekilde alternatif hipotezi kabul etmiş oluruz.
# Örnek: Yeni bir ilacın etkisiz olduğu halde etkili olduğunu düşünmek
bir Type 1 hatasıdır.

#Type 2 Hatası (Yanlış Negatif / Beta Hatası):

# Tanım: Gerçekte yanlış olan bir null hipotezi reddetmeme hatasıdır.
Yani, aslında bir etki varken yokmuş gibi düşünmektir.
# Sonuç: Yanlış bir şekilde null hipotezi kabul etmiş oluruz.
# Örnek: Yeni bir ilacın etkili olduğu halde etkisiz olduğunu düşünmek
bir Type 2 hatasıdır.

#p-Değeri  Null hipotez doğruysa gözlemlenen sonucun elde edilme
olasılığı
#Null Hipotezi  Başlangıçta doğru kabul edilen, "etki yoktur"
anlamındaki varsayım
#Anlamlılık Seviyesi (α)   Hipotez testinde kabul edilen yanılma
riski, genelde 0.05 olarak belirlenir
#Type 1 Hatası  Null hipotezi yanlışlıkla reddetmek (yanlış pozitif)
#Type 2 Hatası  Null hipotezi yanlışlıkla kabul etmek (yanlış negatif)

df_coffee = pd.read_feather("data/coffee_ratings_full.feather")
df_coffee
```

```
      total_cup_points  species                     owner
country_of_origin  \
0                90.58  Arabica                metad plc
Ethiopia
1                89.92  Arabica                metad plc
Ethiopia
2                89.75  Arabica  grounds for health admin
Guatemala
3                89.00  Arabica       yidnekachew dabessa
Ethiopia
4                88.83  Arabica                metad plc
Ethiopia
...                ...      ...                      ...
...
1333             78.75  Robusta              luis robles
Ecuador
1334             78.08  Robusta              luis robles
```

```
Ecuador
1335             77.17  Robusta               james moore     United
States
1336             75.08  Robusta             cafe politico
India
1337             73.75  Robusta             cafe politico
Vietnam

                                      farm_name lot_number
mill  \
0                                     metad plc       None       metad
plc
1                                     metad plc       None       metad
plc
2     san marcos barrancas "san cristobal cuch       None
None
3        yidnekachew dabessa coffee plantation       None
wolensu
4                                     metad plc       None       metad
plc
...                                         ...        ...
...
1333                                  robustasa   Lavado 1     our own
lab
1334                                  robustasa   Lavado 3  own
laboratory
1335                            fazenda cazengo       None    cafe
cazengo
1336                                      None       None
None
1337                                      None       None
None

             ico_number                                    company
altitude  \
0             2014/2015      metad agricultural developmet plc
1950-2200
1             2014/2015      metad agricultural developmet plc
1950-2200
2                  None                                    None  1600 -
1800 m
3                  None  yidnekachew debessa coffee plantation
1800-2200
4             2014/2015      metad agricultural developmet plc
1950-2200
...                 ...                                     ...
...
1333               None                                robustasa
None
```

```
1334               None                                  robustasa
40
1335               None                 global opportunity fund     795
meters
1336  14-1118-2014-0087                       cafe politico
None
1337                n/a                       cafe politico
None

      ...       color category_two_defects           expiration  \
0     ...       Green                   0.0      April 3rd, 2016
1     ...       Green                   1.0      April 3rd, 2016
2     ...        None                   0.0       May 31st, 2011
3     ...       Green                   2.0     March 25th, 2016
4     ...       Green                   2.0      April 3rd, 2016
...   ...         ...                   ...                  ...
1333  ...  Blue-Green                   1.0   January 18th, 2017
1334  ...  Blue-Green                   0.0   January 18th, 2017
1335  ...        None                   6.0  December 23rd, 2015
1336  ...       Green                   1.0    August 25th, 2015
1337  ...        None                   9.0    August 25th, 2015

                     certification_body  \
0     METAD Agricultural Development plc
1     METAD Agricultural Development plc
2           Specialty Coffee Association
3     METAD Agricultural Development plc
4     METAD Agricultural Development plc
...                                  ...
1333        Specialty Coffee Association
1334        Specialty Coffee Association
1335        Specialty Coffee Association
1336        Specialty Coffee Association
1337        Specialty Coffee Association

                         certification_address  \
0     309fcf77415a3661ae83e027f7e5f05dad786e44
1     309fcf77415a3661ae83e027f7e5f05dad786e44
2     36d0d00a3724338ba7937c52a378d085f2172daa
3     309fcf77415a3661ae83e027f7e5f05dad786e44
4     309fcf77415a3661ae83e027f7e5f05dad786e44
...                                        ...
1333  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1334  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1335  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1336  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1337  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                         certification_contact unit_of_measurement  \
0     19fef5a731de2db57d16da10287413f5f99bc2dd                   m
```

```
1     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
3     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
4     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
...                                       ...                  ...
1333  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1334  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1335  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1336  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1337  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
```

```
     altitude_low_meters altitude_high_meters altitude_mean_meters
0                 1950.0                2200.0                2075.0
1                 1950.0                2200.0                2075.0
2                 1600.0                1800.0                1700.0
3                 1800.0                2200.0                2000.0
4                 1950.0                2200.0                2075.0
...                  ...                   ...                   ...
1333                 NaN                   NaN                   NaN
1334                40.0                  40.0                  40.0
1335               795.0                 795.0                 795.0
1336                 NaN                   NaN                   NaN
1337                 NaN                   NaN                   NaN

[1338 rows x 43 columns]
```

```
#Popülasyon:
pts_vs_flavor_pop = df_coffee[["total_cup_points","flavor"]] # bu iki sütunu alarak yeni dataframe döndürür
pts_vs_flavor_pop
```

```
      total_cup_points  flavor
0                90.58    8.83
1                89.92    8.67
2                89.75    8.50
3                89.00    8.58
4                88.83    8.50
...                ...     ...
1333             78.75    7.58
1334             78.08    7.67
1335             77.17    7.33
1336             75.08    6.83
1337             73.75    6.67

[1338 rows x 2 columns]
```

```
pts_vs_flavor_pop_samp = pts_vs_flavor_pop.sample(10) #10 tane random veri ama hepsi biribirinden farklı çünkü replace true değil!
pts_vs_flavor_pop_samp
```

```
      total_cup_points  flavor
474              83.17    7.50
1240             78.00    6.83
215              84.17    7.83
213              84.17    7.75
295              83.75    7.58
9                88.25    8.58
1205             78.92    7.00
317              83.67    7.75
534              82.92    7.50
439              83.25    7.75
```

```
# # totalcuppointsten random 10 veri
cup_points_samp = df_coffee['total_cup_points'].sample(n=10)
cup_points_samp
```

```
256     84.00
273     83.92
1098    80.25
1022    80.92
1147    79.75
653     82.50
1257    77.25
964     81.25
725     82.33
1148    79.75
Name: total_cup_points, dtype: float64
```

```
import numpy as np
print(f"Popülasyon Ortalaması =
{np.mean(pts_vs_flavor_pop['total_cup_points'])}")
print(f"Örneklem Ortalaması (pts_vs_flavor_pop_samp) =
{np.mean(pts_vs_flavor_pop_samp['total_cup_points'])}")
print(f"Örneklem Ortalaması (cup_points_samp) =
{np.mean(cup_points_samp)}")
```

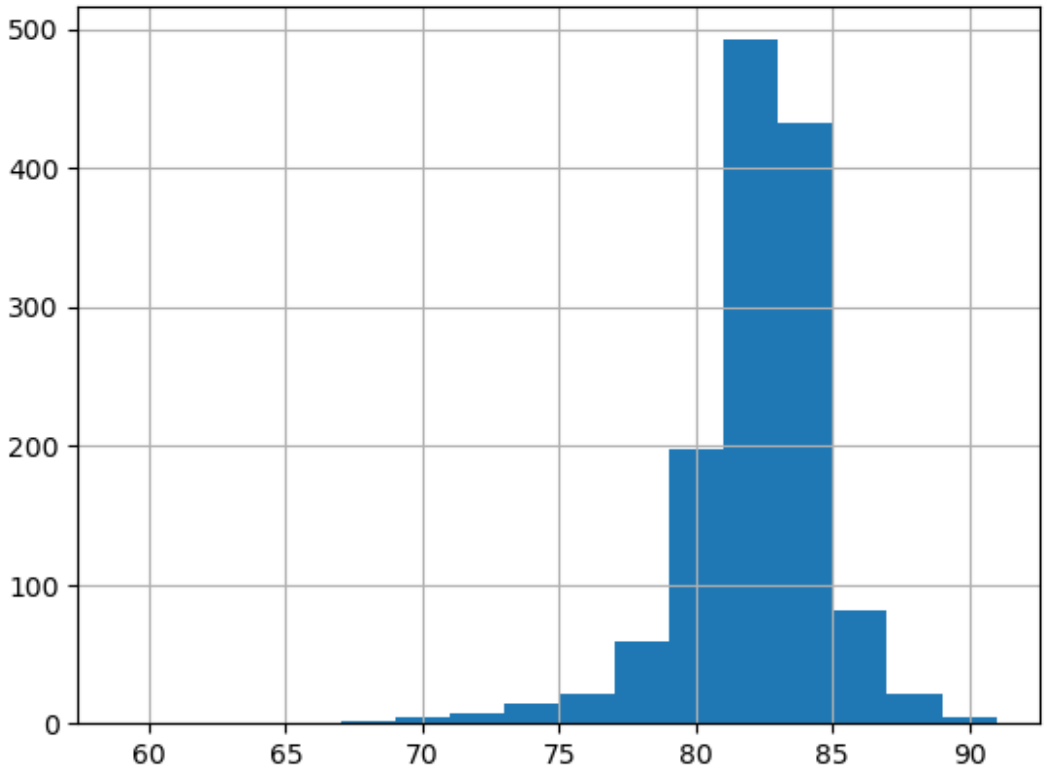
```
Popülasyon Ortalaması = 82.15120328849028
Örneklem Ortalaması (pts_vs_flavor_pop_samp) = 83.027
Örneklem Ortalaması (cup_points_samp) = 81.19200000000001
```

```
df_coffee['total_cup_points'].mean()
df_coffee.head()['total_cup_points'].mean()
```

```
89.616
```

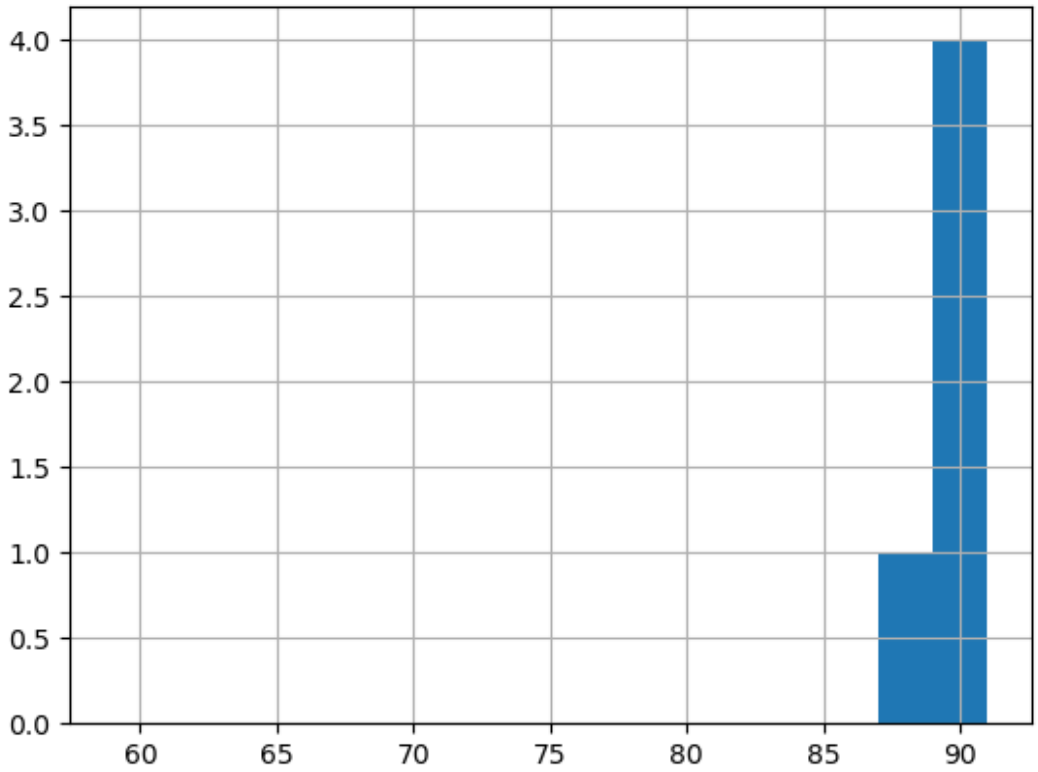
```
import matplotlib.pyplot as plt
df_coffee["total_cup_points"].hist(bins = np.arange(59,93,2))
```

```
<Axes: >
```



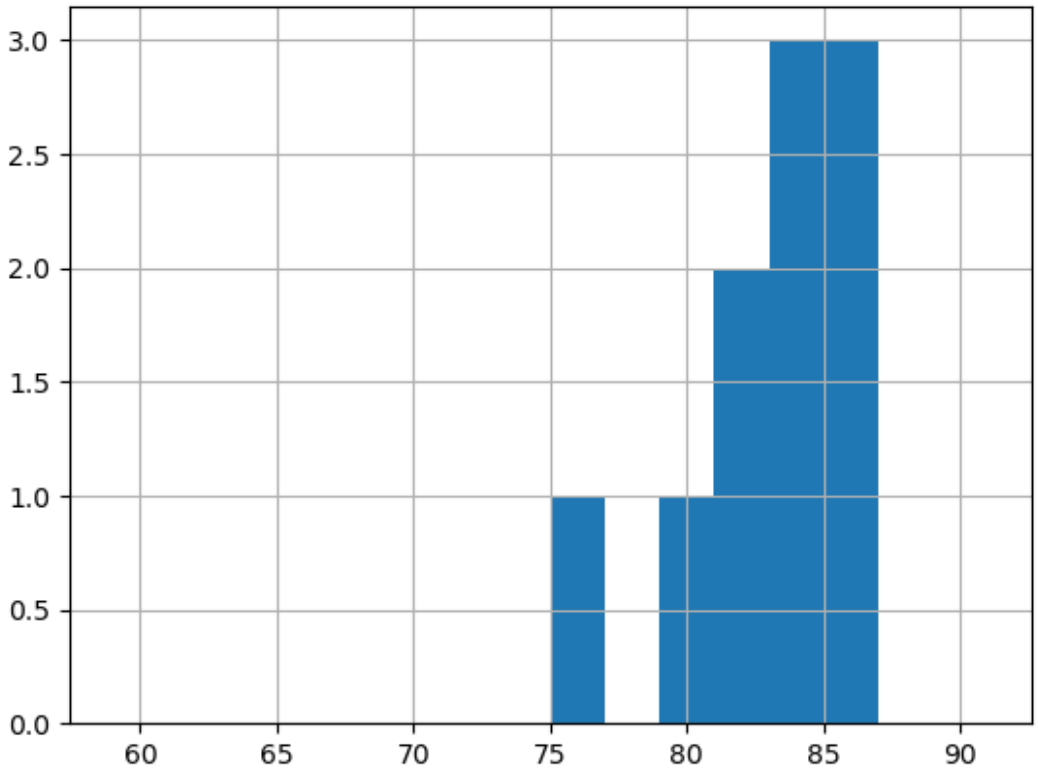

```
df_coffee.head()['total_cup_points'].hist(bins=np.arange(59,93,2))
```

```
<Axes: >
```




```
df_coffee.sample(n=10)['total_cup_points'].hist(bins =
np.arange(59,93,2))
```

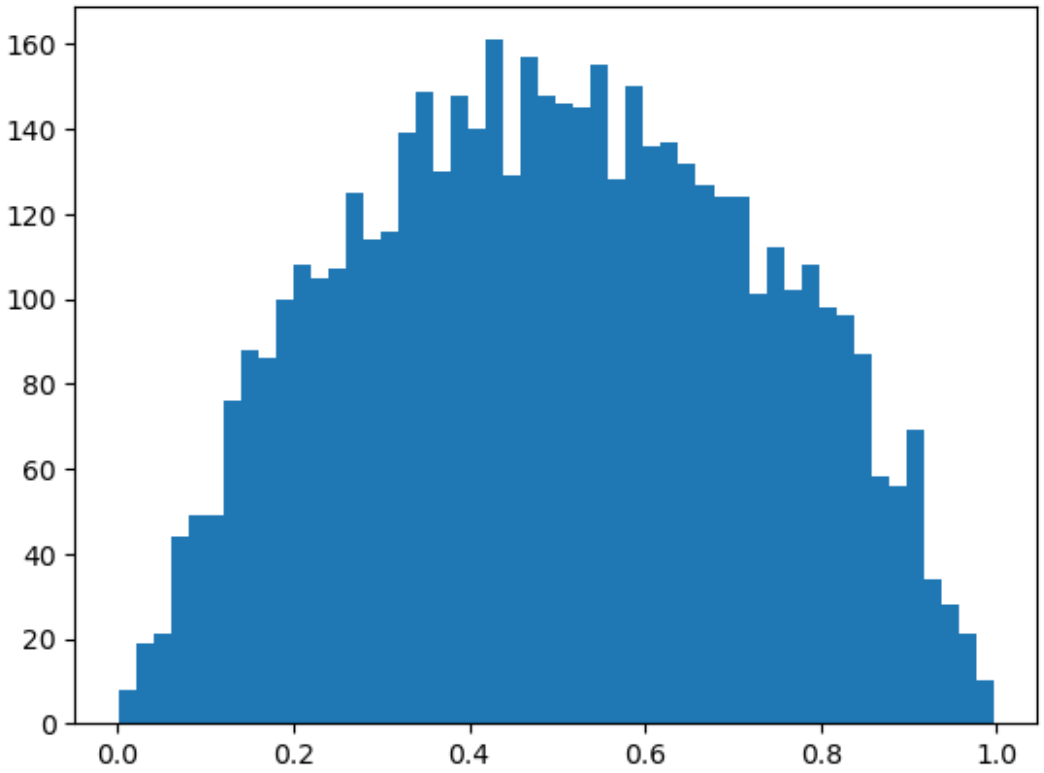
```
<Axes: >
```




```
import numpy as np
import matplotlib.pyplot as plt
randoms = np.random.beta(a=2, b=2,size=5000)
randoms
plt.hist(randoms, bins=50)
```

```
(array([  8.,  19.,  21.,  44.,  49.,  49.,  76.,  88.,  86., 100.,
108.,
        105., 107., 125., 114., 116., 139., 149., 130., 148., 140.,
161.,
        129., 157., 148., 146., 145., 155., 128., 150., 136., 137.,
132.,
        127., 124., 124., 101., 112., 102., 108.,  98.,  96.,  87.,
58.,
         56.,  69.,  34.,  28.,  21.,  10.]),
 array([0.00131336, 0.02121902, 0.04112467, 0.06103032, 0.08093597,
        0.10084162, 0.12074727, 0.14065292, 0.16055857, 0.18046422,
        0.20036987, 0.22027552, 0.24018117, 0.26008682, 0.27999247,
        0.29989813, 0.31980378, 0.33970943, 0.35961508, 0.37952073,
        0.39942638, 0.41933203, 0.43923768, 0.45914333, 0.47904898,
        0.49895463, 0.51886028, 0.53876593, 0.55867158, 0.57857724,
        0.59848289, 0.61838854, 0.63829419, 0.65819984, 0.67810549,
        0.69801114, 0.71791679, 0.73782244, 0.75772809, 0.77763374,
        0.79753939, 0.81744504, 0.83735069, 0.85725635, 0.877162  ,
        0.89706765, 0.9169733 , 0.93687895, 0.9567846 , 0.97669025,
```

```
       0.9965959 ]),
 <BarContainer object of 50 artists>)
```




```
df_coffee.sample(n=5, random_state=15242)
```

	total_cup_points	species	owner	\
488	83.08	Arabica	kona pacific farmers cooperative	
864	81.75	Arabica	sergio landa alarcon	
1274	75.58	Arabica	juan luis alvarado romero	
891	81.67	Arabica	christina dusing	
65	85.50	Arabica	essencecoffee	

	country_of_origin	farm_name	lot_number	\
488	United States (Hawaii)	None	None	
864	Mexico	finca tepictla	None	
1274	Guatemala	agropecuaria quiagral	None	
891	Mexico	ucipa santa catarina	None	
65	Panama	elida estate	None	

	mill	ico_number	\
488	None	K131353	
864	tepictla y xalapa veracruz	1104372561	
1274	beneficio ixchel	11/23/0768	
891	ecc beneficio veracruz	1506803385	
65	elida estate	290503	

```
                                company altitude  ...         color  \
488   kona pacific farmers cooperative     None  ...  Bluish-Green
864                                None     1250  ...         Green
1274              unex guatemala, s.a.     4300  ...         Green
891                                None     1200  ...         Green
65                      essence coffee     1680  ...          None

     category_two_defects            expiration
certification_body  \
488                   0.0       March 8th, 2014  Specialty Coffee
Association
864                  23.0  September 10th, 2013
AMECAFE
1274                 15.0        July 9th, 2013  Asociacion Nacional
Del Café
891                   2.0       July 27th, 2013
AMECAFE
65                    2.0        May 22nd, 2016  Blossom Valley
International

                          certification_address  \
488   36d0d00a3724338ba7937c52a378d085f2172daa
864   59e396ad6e22a1c22b248f958e1da2bd8af85272
1274  b1f20fe3a819fd6b2ee0eb8fdc3da256604f1e53
891   59e396ad6e22a1c22b248f958e1da2bd8af85272
65    fc45352eee499d8470cf94c9827922fb745bf815

                          certification_contact unit_of_measurement  \
488   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                   ft
864   0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1274  724f04ad10ed31dbb9d260f0dfd221ba48be8a95                   ft
891   0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
65    de73fc9412358b523d3a641501e542f31d2668b0                    m

     altitude_low_meters altitude_high_meters altitude_mean_meters
488                  NaN                  NaN                  NaN
864              1250.00              1250.00              1250.00
1274             1310.64              1310.64              1310.64
891              1200.00              1200.00              1200.00
65               1680.00              1680.00              1680.00

[5 rows x 43 columns]
```

```
# WEEK 6
# Tabakalı Örnekleme: Popülasyon gruplara ayrılır, her gruptan eşit
oranda örnek seçilir.
# Ağırlıklı Örnekleme: Her bir grubun popülasyondaki oranına göre
örnekler alınır.
# Küme Örnekleme: Alt gruplara ayrılmış popülasyonda belirli kümeler
```

```
seçilir.
# Sistematik Örnekleme: Belirli aralıklarla örnekler seçilir,
popülasyon sıralıdır.
import pandas as pd
df_coffee = pd.read_feather("data/coffee_ratings_full.feather")
df_coffee
```

```
      total_cup_points  species                     owner
country_of_origin  \
0                90.58  Arabica                metad plc
Ethiopia
1                89.92  Arabica                metad plc
Ethiopia
2                89.75  Arabica  grounds for health admin
Guatemala
3                89.00  Arabica       yidnekachew dabessa
Ethiopia
4                88.83  Arabica                metad plc
Ethiopia
...                ...      ...                      ...
...
1333             78.75  Robusta              luis robles
Ecuador
1334             78.08  Robusta              luis robles
Ecuador
1335             77.17  Robusta              james moore     United
States
1336             75.08  Robusta            cafe politico
India
1337             73.75  Robusta            cafe politico
Vietnam

                                     farm_name lot_number
mill  \
0                                    metad plc       None       metad
plc
1                                    metad plc       None       metad
plc
2     san marcos barrancas "san cristobal cuch       None
None
3        yidnekachew dabessa coffee plantation       None
wolensu
4                                    metad plc       None       metad
plc
...                                        ...        ...
...
1333                                 robustasa   Lavado 1     our own
lab
1334                                 robustasa   Lavado 3  own
laboratory
```

```
1335                            fazenda cazengo       None    cafe
cazengo
1336                                       None       None
None
1337                                       None       None
None

             ico_number                                  company
altitude  \
0              2014/2015      metad agricultural developmet plc
1950-2200
1              2014/2015      metad agricultural developmet plc
1950-2200
2                   None                                   None  1600 -
1800 m
3                   None  yidnekachew debessa coffee plantation
1800-2200
4              2014/2015      metad agricultural developmet plc
1950-2200
...                  ...                                   ...
...
1333                None                                robustasa
None
1334                None                                robustasa
40
1335                None                global opportunity fund     795
meters
1336  14-1118-2014-0087                           cafe politico
None
1337                n/a                           cafe politico
None

      ...       color category_two_defects           expiration  \
0     ...       Green                  0.0      April 3rd, 2016
1     ...       Green                  1.0      April 3rd, 2016
2     ...        None                  0.0       May 31st, 2011
3     ...       Green                  2.0     March 25th, 2016
4     ...       Green                  2.0      April 3rd, 2016
...   ...         ...                  ...                   ...
1333  ...  Blue-Green                  1.0   January 18th, 2017
1334  ...  Blue-Green                  0.0   January 18th, 2017
1335  ...        None                  6.0  December 23rd, 2015
1336  ...       Green                  1.0    August 25th, 2015
1337  ...        None                  9.0    August 25th, 2015

                     certification_body  \
0     METAD Agricultural Development plc
1     METAD Agricultural Development plc
2           Specialty Coffee Association
3     METAD Agricultural Development plc
```

```
4     METAD Agricultural Development plc
...                                  ...
1333        Specialty Coffee Association
1334        Specialty Coffee Association
1335        Specialty Coffee Association
1336        Specialty Coffee Association
1337        Specialty Coffee Association

                         certification_address  \
0     309fcf77415a3661ae83e027f7e5f05dad786e44
1     309fcf77415a3661ae83e027f7e5f05dad786e44
2     36d0d00a3724338ba7937c52a378d085f2172daa
3     309fcf77415a3661ae83e027f7e5f05dad786e44
4     309fcf77415a3661ae83e027f7e5f05dad786e44
...                                        ...
1333  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1334  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1335  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1336  ff7c18ad303d4b603ac3f8cff7e611ffc735e720
1337  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                         certification_contact unit_of_measurement  \
0     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
1     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
3     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
4     19fef5a731de2db57d16da10287413f5f99bc2dd                    m
...                                       ...                  ...
1333  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1334  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1335  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1336  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m
1337  352d0cf7f3e9be14dad7df644ad65efc27605ae2                    m

     altitude_low_meters altitude_high_meters altitude_mean_meters
0                 1950.0                2200.0                2075.0
1                 1950.0                2200.0                2075.0
2                 1600.0                1800.0                1700.0
3                 1800.0                2200.0                2000.0
4                 1950.0                2200.0                2075.0
...                  ...                   ...                   ...
1333                 NaN                   NaN                   NaN
1334                40.0                  40.0                  40.0
1335               795.0                 795.0                 795.0
1336                 NaN                   NaN                   NaN
1337                 NaN                   NaN                   NaN

[1338 rows x 43 columns]
```

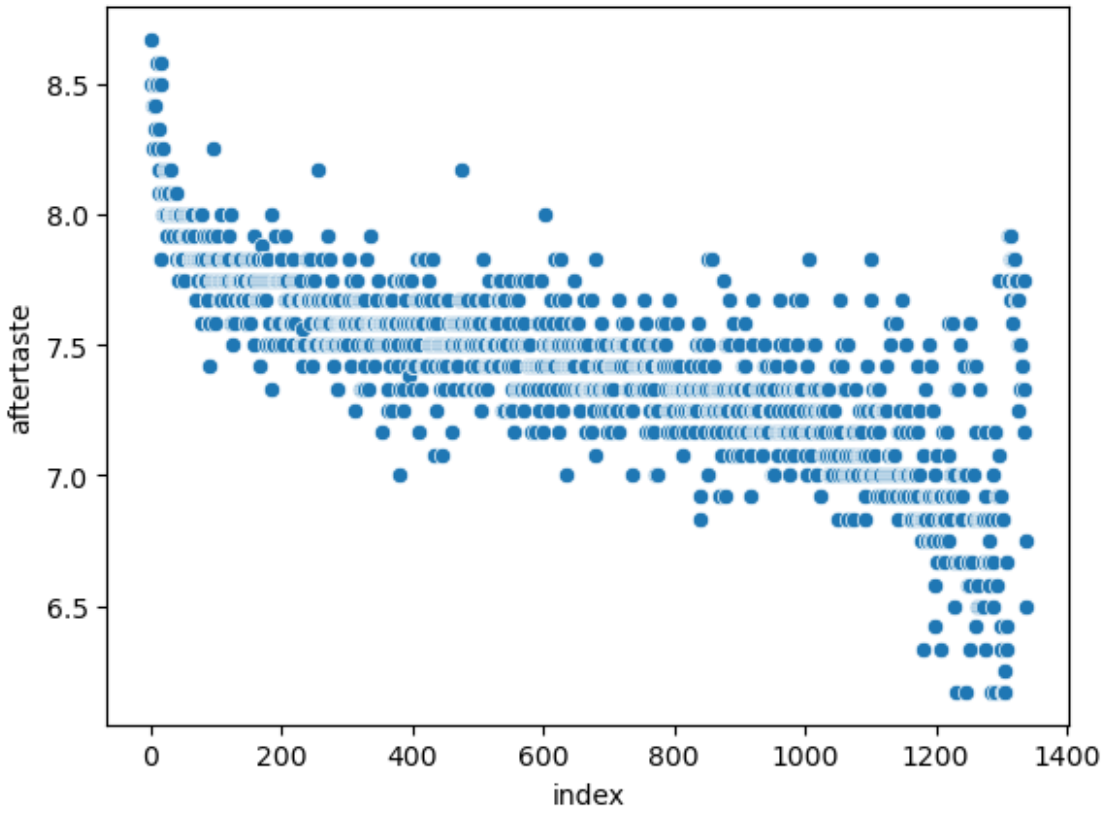
```
# SİSTEMATİK ÖRNEKLEME
sample_size = 5
pop_size = len(df_coffee) #bunun yerine
pop_size2= df_coffee.shape[0]

interval = pop_size2 // sample_size
interval

df_coffee.iloc[interval::interval]
```

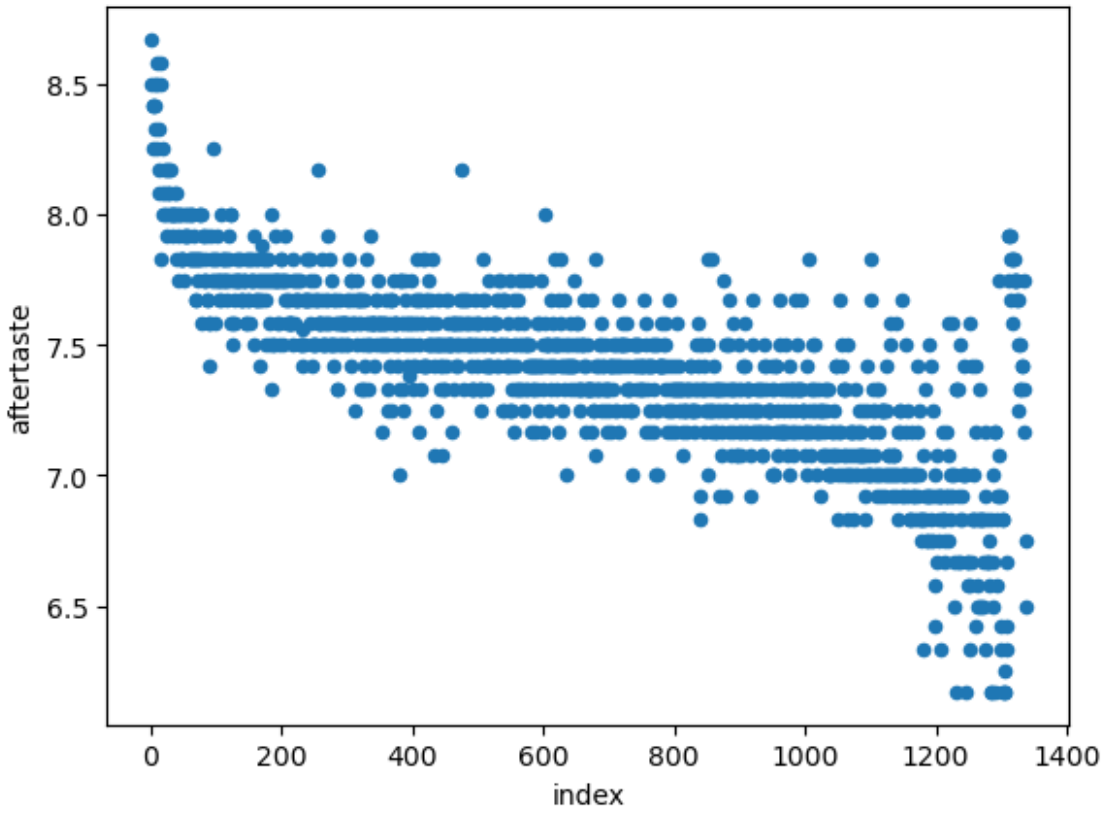
```
      total_cup_points  species                                owner  \
267              83.92  Arabica  federacion nacional de cafeteros
534              82.92  Arabica      consejo salvadoreño del café
801              82.00  Arabica            lin, che-hao krude 林哲豪
1068             80.50  Arabica          cqi taiwan icp cqi 台灣合作夥伴
1335             77.17  Robusta                         james moore

     country_of_origin
farm_name  \
267           Colombia                                          None

534        El Salvador                                santa josefita

801             Taiwan  you siang coffee farmtainan, taiwan 台灣台南優香
咖啡
1068            Taiwan                                          王秋金

1335     United States                               fazenda cazengo


     lot_number                                          mill
ico_number  \
267        None                                          None     01-
1969
534       1-198                          beneficio cuzcachapa  09-030-
273
801        None  you siang coffee farmtainan, taiwan 台灣台南優香咖啡
Taiwan
1068          1                                           non
None
1335       None                                cafe cazengo
None

                              company    altitude  ...       color  \
267   federacion nacional de cafeteros        None  ...        None
534      soc. coop. cuzcachapa de r.l.        1350  ...       Green
801              red on tree co., ltd.        600m  ...       Green
1068                            王秋金    50  ...  Blue-Green
1335          global opportunity fund  795 meters  ...        None
```

```
     category_two_defects           expiration
certification_body  \
267                   1.0     March 11th, 2016
Almacafé
534                   1.0    August 28th, 2018     Salvadoran Coffee
Council
801                   0.0      July 22nd, 2015  Specialty Coffee
Association
1068                  0.0   December 8th, 2018  Blossom Valley
International
1335                  6.0  December 23rd, 2015  Specialty Coffee
Association

                          certification_address  \
267   e493c36c2d076bf273064f7ac23ad562af257a25
534   3d4987e3b91399dbb3938b5bdf53893b6ef45be1
801   36d0d00a3724338ba7937c52a378d085f2172daa
1068  fc45352eee499d8470cf94c9827922fb745bf815
1335  ff7c18ad303d4b603ac3f8cff7e611ffc735e720

                          certification_contact unit_of_measurement  \
267   70d3c0c26f89e00fdae6fb39ff54f0d2eb1c38ab                   m
534   27b21e368fb8291cbea02c60623fe6c98f84524d                   m
801   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                   m
1068  de73fc9412358b523d3a641501e542f31d2668b0                   m
1335  352d0cf7f3e9be14dad7df644ad65efc27605ae2                   m

     altitude_low_meters altitude_high_meters altitude_mean_meters
267                  NaN                  NaN                  NaN
534               1350.0               1350.0               1350.0
801                600.0                600.0                600.0
1068                50.0                 50.0                 50.0
1335               795.0                795.0                795.0

[5 rows x 43 columns]
```

```
# yukarıda güzel kahveler aşağıda kötü kahveler var. bunun için:
df_coffee_id = df_coffee.reset_index()
sns.scatterplot(x='index', y='aftertaste', data=df_coffee_id)
#ya da
df_coffee_id.plot(x='index', y='aftertaste', kind='scatter')
```

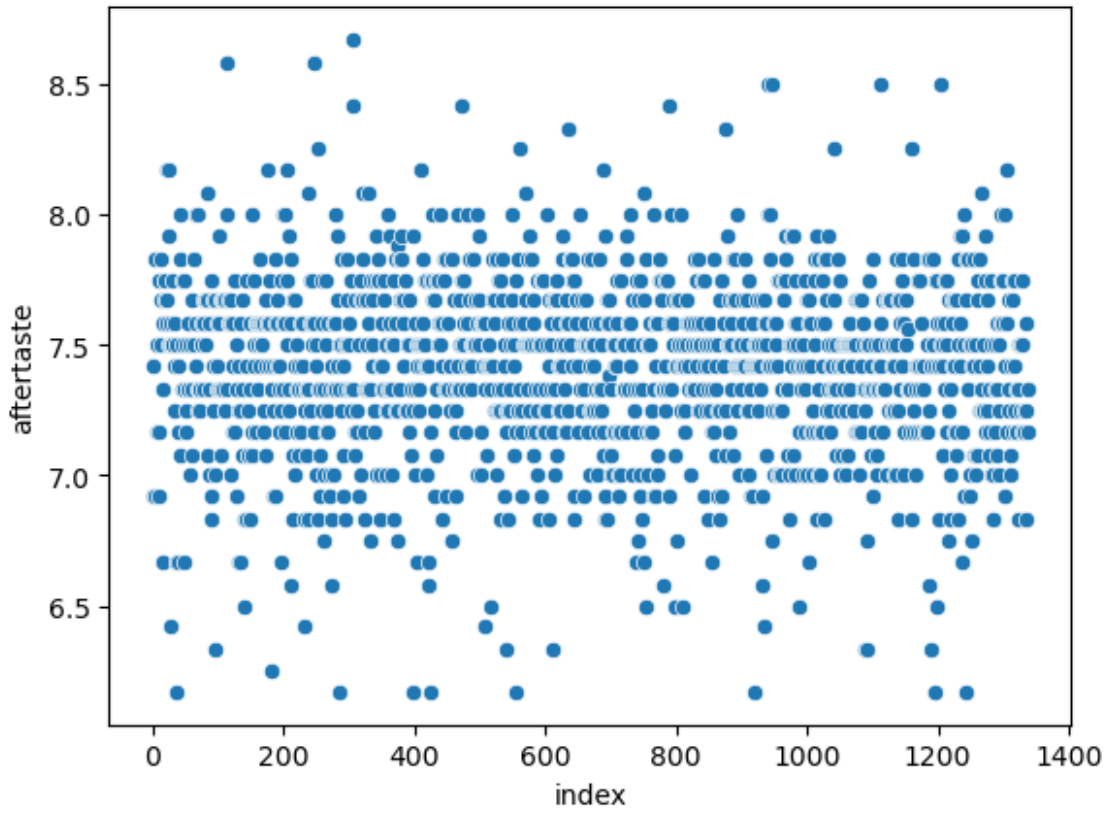
```
<Axes: xlabel='index', ylabel='aftertaste'>
```


aftertaste
8.5
8.0
7.5
7.0
6.5
0
200
400
600
800
1000
1200
1400
index


aftertaste
8.5
8.0
7.5
7.0
6.5
0
200
400
600
800
1000
1200
1400
index

```
# eşitlenmeleri için, yani güzeller üstte kötüler altta olmasın diye bu verişeri karıştırmalıyız!
shuffled = df_coffee.sample(frac=1) # tğm verilerin random şekilde karışmasını sağlar. frac=! sayesinde tüm verileri
shuffled = shuffled.reset_index(drop=True).reset_index() #ilk baştaki indexleri sil, sonra baştan idexle 2. reset ile
shuffled.plot(x='index', y='aftertaste', kind='scatter')
#ya da
sns.scatterplot(x='index', y='aftertaste', data=shuffled)
```

```
<Axes: xlabel='index', ylabel='aftertaste'>
```




```
#TABAKALI ÖRNEKLEME: alt grupları içeren bir popülasyonu örneklemeye olanak sağlayan tekniktir
top_counts = df_coffee['country_of_origin'].value_counts()
top_counts
```

```
country_of_origin
Mexico                          236
Colombia                        183
Guatemala                       181
Brazil                          132
Taiwan                           75
United States (Hawaii)           73
```

```
Honduras                            52
Costa Rica                          51
Ethiopia                            44
Tanzania, United Republic Of        40
Uganda                              36
Thailand                            32
Nicaragua                           26
Kenya                               25
El Salvador                         21
Indonesia                           20
China                               16
India                               14
Malawi                              11
United States                       10
Peru                                10
Myanmar                              8
Vietnam                              8
Haiti                                6
Philippines                          5
United States (Puerto Rico)          4
Panama                               4
Ecuador                              3
Laos                                 3
Burundi                              2
Papua New Guinea                     1
Rwanda                               1
Zambia                               1
Japan                                1
Mauritius                            1
Cote d?Ivoire                        1
Name: count, dtype: int64
```

```
# top_counts2=
df_coffee['country_of_origin'].value_counts().head(6).toList()
top_counts2 =
df_coffee['country_of_origin'].value_counts().head(6).index.tolist()
top_counts2
```

```
array(['Ethiopia', 'Guatemala', 'Brazil', 'Peru', 'United States',
       'United States (Hawaii)', 'Indonesia', 'China', 'Costa Rica',
       'Mexico', 'Uganda', 'Honduras', 'Taiwan', 'Nicaragua',
       'Tanzania, United Republic Of', 'Kenya', 'Thailand',
'Colombia',
       'Panama', 'Papua New Guinea', 'El Salvador', 'Japan',
'Ecuador',
       'United States (Puerto Rico)', 'Haiti', 'Burundi', 'Vietnam',
       'Philippines', 'Rwanda', 'Malawi', 'Laos', 'Zambia', 'Myanmar',
       'Mauritius', 'Cote d?Ivoire', None, 'India'], dtype=object)
```

```
# top countrylerin kahvelerini seç
top_counted_countries = top_counts2
top_counted_subset =
df_coffee['country_of_origin'].isin(top_counted_countries) #boolean
olarak tutar verileri
coffee_ratings_top = df_coffee[top_counted_subset] #trueları seç
coffee_ratings_top
```

```
      total_cup_points  species                          owner  \
2                89.75  Arabica       grounds for health admin
5                88.83  Arabica                    ji-ae ahn
13               87.92  Arabica       grounds for health admin
22               87.17  Arabica          roberto licona franco
25               86.92  Arabica                      nucoffee
...                ...      ...                           ...
1300             71.00  Arabica  ricardo aaron sampieri marini
1301             70.75  Arabica                kurt kappeli
1302             70.67  Arabica         volcafe ltda. - brasil
1306             68.33  Arabica       juan carlos garcia lopez
1309             59.83  Arabica      juan luis alvarado romero

           country_of_origin                                farm_name
\
2                  Guatemala  san marcos barrancas "san cristobal cuch

5                     Brazil                                       None

13    United States (Hawaii)                          arianna farms

22                    Mexico                           la herradura

25                    Brazil                       fazenda kaquend

...                      ...                                      ...

1300                  Mexico                              la morena

1301                  Mexico                                various

1302                  Brazil                                   None

1306                  Mexico                         el centenario

1309               Guatemala                        finca el limon


             lot_number
mill  \
2                  None
None
5                  None
```

```
None
13                  None                                                  None
22                  None                                                 la herradura
25                  None                                                  None
...                  ...                                                   ...
1300                None                                tlamatoca, hutusco, ver.
1301                None                                                  f.i.e.c.h.
1302  2017/2018 - Lot 2                                                   copag
1306                None  la esperanza, municipio juchique de ferrer, ve...
1309                None                                     beneficio serben

          ico_number        company       altitude  ...          color \
2               None           None  1600 - 1800 m  ...           None
5               None           None           None  ...  Bluish-Green
13              None           None        2000 ft  ...           None
22                 0           None           1320  ...          Green
25     002/1251/0073       nucoffee          1250m  ...          Green
...               ...            ...            ...  ...            ...
1300      1104351023           None           1800  ...          Green
1301  0016-2847-0001  globus coffee    1000 meters  ...          Green
1302            None  volcafe ltda.           None  ...          Green
1306      1104328663     terra mia             900  ...           None
1309      11/853/165       unicafe           4650  ...          Green

     category_two_defects            expiration  \
2                     0.0        May 31st, 2011
5                     1.0   September 3rd, 2014
13                    2.0        May 31st, 2011
22                    0.0       July 26th, 2013
```

```
25                    2.0    December 2nd, 2012
...                   ...                    ...
1300                  0.0       July 11th, 2013
1301                  1.0         May 5th, 2015
1302                 55.0    October 27th, 2018
1306                 20.0  September 17th, 2013
1309                  4.0        May 24th, 2013

                        certification_body  \
2            Specialty Coffee Association
5      Specialty Coffee Institute of Asia
13           Specialty Coffee Association
22                                AMECAFE
25                               NUCOFFEE
...                                   ...
1300                              AMECAFE
1301         Specialty Coffee Association
1302  Brazil Specialty Coffee Association
1306                              AMECAFE
1309         Asociacion Nacional Del Café

                          certification_address  \
2     36d0d00a3724338ba7937c52a378d085f2172daa
5     726e4891cf2c9a4848768bd34b668124d12c4224
13    36d0d00a3724338ba7937c52a378d085f2172daa
22    59e396ad6e22a1c22b248f958e1da2bd8af85272
25    567f200bcc17a90070cb952647bf88141ad9c80c
...                                        ...
1300  59e396ad6e22a1c22b248f958e1da2bd8af85272
1301  36d0d00a3724338ba7937c52a378d085f2172daa
1302  3297cfa4c538e3dd03f72cc4082c54f7999e1f9d
1306  59e396ad6e22a1c22b248f958e1da2bd8af85272
1309  b1f20fe3a819fd6b2ee0eb8fdc3da256604f1e53

                          certification_contact unit_of_measurement  \
2     0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
5     b70da261fcc84831e3e9620c30a8701540abc200                    m
13    0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                   ft
22    0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
25    aa2ff513ffb9c844462a1fb07c599bce7f3bb53d                    m
...                                        ...                  ...
1300  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1301  0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
1302  8900f0bf1d0b2bafe6807a73562c7677d57eb980                    m
1306  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
1309  724f04ad10ed31dbb9d260f0dfd221ba48be8a95                   ft

     altitude_low_meters altitude_high_meters altitude_mean_meters
2                 1600.00               1800.00               1700.00
5                     NaN                   NaN                   NaN
```

```
13                609.60               609.60               609.60
22               1320.00              1320.00              1320.00
25               1250.00              1250.00              1250.00
...                  ...                  ...                  ...
1300             1800.00              1800.00              1800.00
1301             1000.00              1000.00              1000.00
1302                 NaN                  NaN                  NaN
1306              900.00               900.00               900.00
1309             1417.32              1417.32              1417.32

[880 rows x 43 columns]
```

```
coffee_ratings_samp = coffee_ratings_top.sample(frac=0.1,
random_state=2021) # yukarıdaki verinin yüzde 10unu aldık
coffee_ratings_samp
```

```
      total_cup_points  species                             owner  \
1229             78.33  Arabica             pablo cervantes morelos
232              84.08  Arabica                     carcafe ltda ci
697              82.42  Arabica           jose daniel cobilt castro
865              81.75  Arabica    diego manuel woolrich ramirez
155              84.58  Arabica   exportadora de cafe condor s.a
...                ...      ...                                 ...
1282             74.83  Arabica      pablo enrique martinez gama
369              83.50  Arabica      gabriel bernardo rivas ross
852              81.83  Arabica         jacques pereira carneiro
713              82.33  Arabica        bourbon specialty coffees
806              82.00  Arabica           lin, che-hao krude 林哲豪

     country_of_origin
farm_name  \
1229           Mexico                                       llano
hermoso
232           Colombia
None
697              Mexico                                     cañada
fria
865              Mexico  arroyo triste, arroyo triste, san jose vista
h...
155           Colombia
various
...                 ...
...
1282             Mexico                                       la
orduña
369              Mexico                                    la
corralera
852              Brazil                                     sertao
farm
713              Brazil
```

```
None
806             Taiwan                                      gao chun fang
高醇坊

     lot_number                                                      mill  \
1229       None        llano hermoso, xochitonalco huautla, oaxaca
232   3-59-0503                                                     neiva
697        None                                                  huatusco
865        None  arroyo triste, arroyo triste, san jose vista h...
155        None                                  trilladora boananza
...         ...                                                      ...
1282       None                                falcafe s.a. de c.v.
369        None                      dos puentes de finca kassandra
852        None                            armazens gerais cocarive
713        None                                                  None
806        None                                   gao chun fang 高醇坊

                       ico_number  \
1229                            0
232                      3-59-0503
697                     1104558673
865   2037240, 2037150,1400213685
155                      3-68-0005
...                            ...
1282                    1104362940
369                           2484
852                 002/1352/0159
713                 002/4542/0478
806                        Taiwan

                                                    company
altitude  ...  \
1229  asociación agricola local de productores de ca...
1300  ...
232                                          carcafe ltda
442  ...
697                                                  None
1350  ...
865                                                  None
1100  ...
155                     exportadora de cafe condor s.a  1800
msnm  ...
...                                                   ...        ...  ..
.
1282                                                 None
1250  ...
369                                                  None
1400  ...
852           exportadora de cafés carmo de minas ltda
1250  ...
```

```
713                              bourbon specialty coffees
None  ...
806                                 red on tree co., ltd.  600-700
m  ...

             color category_two_defects            expiration  \
1229         Green                 47.0  September 11th, 2013
232          Green                  3.0    November 9th, 2018
697          Green                  6.0       July 11th, 2013
865          Green                  1.0   September 4th, 2013
155          Green                  6.0     October 9th, 2013
...            ...                  ...                   ...
1282         Green                 30.0      August 1st, 2013
369          Green                  0.0       July 11th, 2013
852   Bluish-Green                  0.0       March 2nd, 2014
713          Green                 10.0      April 19th, 2016
806   Bluish-Green                  0.0        June 3rd, 2014

                       certification_body  \
1229                              AMECAFE
232                              Almacafé
697                               AMECAFE
865                               AMECAFE
155                              Almacafé
...                                   ...
1282                              AMECAFE
369                               AMECAFE
852          Specialty Coffee Association
713   Brazil Specialty Coffee Association
806          Specialty Coffee Association

                         certification_address  \
1229  3e18a5ae6f5e2aabca37e025f94e1974558bf5f0
232   e493c36c2d076bf273064f7ac23ad562af257a25
697   59e396ad6e22a1c22b248f958e1da2bd8af85272
865   59e396ad6e22a1c22b248f958e1da2bd8af85272
155   e493c36c2d076bf273064f7ac23ad562af257a25
...                                        ...
1282  59e396ad6e22a1c22b248f958e1da2bd8af85272
369   59e396ad6e22a1c22b248f958e1da2bd8af85272
852   36d0d00a3724338ba7937c52a378d085f2172daa
713   3297cfa4c538e3dd03f72cc4082c54f7999e1f9d
806   36d0d00a3724338ba7937c52a378d085f2172daa

                         certification_contact unit_of_measurement  \
1229  e3212d17882b7657b3fba559b4072e552604d5d1                    m
232   70d3c0c26f89e00fdae6fb39ff54f0d2eb1c38ab                    m
697   0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
865   0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
155   70d3c0c26f89e00fdae6fb39ff54f0d2eb1c38ab                    m
```

```
...                                              ...                  ...
1282  0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
369   0eb4ee5b3f47b20b049548a2fd1e7d4a2b70d0a7                    m
852   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
713   8900f0bf1d0b2bafe6807a73562c7677d57eb980                    m
806   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m

     altitude_low_meters altitude_high_meters altitude_mean_meters
1229              1300.0               1300.0                1300.0
232                442.0                442.0                 442.0
697               1350.0               1350.0                1350.0
865               1100.0               1100.0                1100.0
155               1800.0               1800.0                1800.0
...                  ...                  ...                   ...
1282              1250.0               1250.0                1250.0
369               1400.0               1400.0                1400.0
852               1250.0               1250.0                1250.0
713                  NaN                  NaN                   NaN
806                600.0                700.0                 650.0

[88 rows x 43 columns]
```

```
coffee_ratings_samp['country_of_origin'].value_counts(normalize=True)
```

```
country_of_origin
Mexico                     0.250000
Guatemala                  0.204545
Colombia                   0.181818
Brazil                     0.181818
United States (Hawaii)     0.102273
Taiwan                     0.079545
Name: proportion, dtype: float64
```

```
#yüzde 10nunu aldık. taiwanin yuzdesi normale göre çok düşük gelmiş:
#"Her ülkeden %10 oranında örnek seçerek veri setini nasıl
oluşturursunuz ve her ülkenin veri setindeki oranını nasıl kontrol
edersiniz?"
coffee_rating_strat =
coffee_ratings_top.groupby('country_of_origin').sample(frac=0.1,
random_state=2021)
coffee_rating_strat['country_of_origin'].value_counts(normalize=True)
```

```
country_of_origin
Mexico                     0.272727
Colombia                   0.204545
Guatemala                  0.204545
Brazil                     0.147727
Taiwan                     0.090909
United States (Hawaii)     0.079545
Name: proportion, dtype: float64
```

```
#hepsinden eşit gelsin:
#"Her ülkeden 15 tane örnek seçerek veri setini eşit sayıda veriyle
nasıl oluşturursunuz?"
coffee_ratings_eq =
coffee_ratings_top.groupby('country_of_origin').sample(n=15,random_sta
te=2021)
coffee_ratings_eq['country_of_origin'].value_counts(normalize=True)
```

```
country_of_origin
Brazil                    0.166667
Colombia                  0.166667
Guatemala                 0.166667
Mexico                    0.166667
Taiwan                    0.166667
United States (Hawaii)    0.166667
Name: proportion, dtype: float64
```

```
# "Her ülkeden %10 oranında örnek seçerek veri setini nasıl
oluşturursunuz ve
# her ülkenin veri setindeki oranını nasıl kontrol edersiniz?"
coffee_istenen =
coffee_ratings_top.groupby('country_of_origin').sample(frac=0.1,
random_state=2021)
coffee_istenen['country_of_origin'].value_counts(normalize=True)
```

```
country_of_origin
Mexico                    0.272727
Colombia                  0.204545
Guatemala                 0.204545
Brazil                    0.147727
Taiwan                    0.090909
United States (Hawaii)    0.079545
Name: proportion, dtype: float64
```

```
# AĞIRLIKLI RASTGELE ÖRNEKLEME: her grubun popülasyondaki oranına göre

coffee_ratings_weight = coffee_ratings_top.copy()
condition = coffee_ratings_weight['country_of_origin'] == 'Taiwan'
coffee_ratings_weight['weight'] = np.where(condition, 2, 1)
coffee_ratings_weight[coffee_ratings_weight['weight'] == 2]
coffee_ratings_weight =
coffee_ratings_weight.groupby('weight').sample(frac=0.1)
coffee_ratings_weight['country_of_origin'].value_counts(normalize=True
)
```

```
country_of_origin
Mexico                    0.238636
Colombia                  0.227273
Guatemala                 0.193182
Brazil                    0.170455
```

```
Taiwan                      0.090909
United States (Hawaii)      0.079545
Name: proportion, dtype: float64
```

```
varieties_pop = list(df_coffee['variety'].unique())
varieties_pop
```

```
[None,
 'Other',
 'Bourbon',
 'Catimor',
 'Ethiopian Yirgacheffe',
 'Caturra',
 'SL14',
 'Sumatra',
 'SL34',
 'Hawaiian Kona',
 'Yellow Bourbon',
 'SL28',
 'Gesha',
 'Catuai',
 'Pacamara',
 'Typica',
 'Sumatra Lintong',
 'Mundo Novo',
 'Java',
 'Peaberry',
 'Pacas',
 'Mandheling',
 'Ruiru 11',
 'Arusha',
 'Ethiopian Heirlooms',
 'Moka Peaberry',
 'Sulawesi',
 'Blue Mountain',
 'Marigojipe',
 'Pache Comun']
```

```
import random
varieties_samp = random.sample(varieties_pop, k=3)
varieties_samp
```

```
['SL14', 'Java', 'Mandheling']
```

```
# "Belirli bir kahve çeşitleri listesindeki (varieties_samp) çeşitlere
ait verileri
# nasıl seçersiniz?"
v_condition = df_coffee['variety'].isin(varieties_samp)
coffee_ratings_cluster = df_coffee[v_condition]
coffee_ratings_cluster
```

```
      total_cup_points  species                                    owner
\
27                86.83  Arabica                  kabum trading company

53                85.92  Arabica                     kawacom uganda ltd

68                85.50  Arabica                         kyagalanyi ltd

71                85.42  Arabica              great lakes coffee uganda

96                85.00  Arabica                  kyagalanyi coffee ltd

116               84.83  Arabica                     kawacom uganda ltd

126               84.67  Arabica                         sanjava coffee

159               84.50  Arabica                     ecom japan limited

172               84.42  Arabica                      aulia arif syahri

231               84.13  Arabica  pt.royal pacific indah international

342               83.58  Arabica  bulamburi coffee farmers association

380               83.42  Arabica               bugisu cooperative union

420               83.25  Arabica                     kawacom uganda ltd

456               83.17  Arabica                  kabum trading company

508               83.00  Arabica                          star cafe ltd

544               82.92  Arabica                     kawacom uganda ltd

545               82.92  Arabica     nyapea coffee farmers association

619               82.67  Arabica                     kawacom uganda ltd

642               82.58  Arabica                     kawacom uganda ltd

1008              81.00  Arabica                         kyagalanyi ltd

1050              80.67  Arabica                      aulia arif syahri

1233              78.17  Arabica                         sanjava coffee


     country_of_origin                              farm_name  \
27              Uganda              chebonet (23) women coffee
53              Uganda            sipi organic coffee project
68              Uganda                              buginyanya
71              Uganda                    chesiyo farmer group
```

```
96              Uganda                          mount elgon area
116             Uganda                mt.elgon bugisu shamba 2
126          Indonesia              sran temanggung plantation
159             Uganda                     kawacom sipi project
172          Indonesia                                  darmawi
231          Indonesia                                mus, eman
342             Uganda                bulamburi coffee farmers
380             Uganda                     bulago & buginyanya
420             Uganda  kawacom uganda ltd sipi farmers group
456             Uganda                                 kaptanya
508             Uganda                           kabeywa county
544             Uganda            sipi organic coffee project
545             Uganda                  nyapea coffee farmers
619             Uganda                           bugisu shamba
642             Uganda             mt.elgon sipi falls cheema
1008            Uganda                                 kapchorwa
1050         Indonesia                                   darmawi
1233         Indonesia                                   various

                lot_number                              mill
ico_number  \
27                    None         kabum trading company
0
53                    None                        kawacom
0
68                    None         kyagalanyi coffee ltd
0
71                    None            great lakes coffee
0
96                    6133         kyagalanyi coffee ltd
None
116                   None                        kawacom
0
126                      1                   wet hulling
None
159   035/170/5071146                       kawacom  035/170/5071146

172   MANDHELING BRASTAGI                       dry mill         To be
advice
231                      1  dry mill or hulling facility
None
342                   None      bulamburi coffee farmers
0
380                   5055                           bcu
7697
420       035/170/5061178                       kawacom
035/170/5061178
456                   None                     gumutindo
0
508                   None          kucofa farmers group
```

```
0
544                   None                          kawacom
0
545                   None                           nyapea
0
619                   None                          kawacom
793
642                   None                          kawacom
0
1008                  None          kyagalanyi coffee ltd
0
1050                  None  surbakti / pt.olam indonesia
015/1691/006
1233             sran-ijen                      east java
None

                                      company    altitude  ...
color  \
27                     kabum trading company        1950  ...
Green
53                        kawacom uganda ltd   1400-1900  ...
Green
68                     kyagalanyi coffee ltd        1600  ...
Green
71                        great lakes coffee        1950  ...
Green
96                     kyagalanyi coffee ltd        1800  ...
Green
116                       kawacom uganda ltd   1400-1900  ...
Green
126               pt. shriya artha nusantara        1200  ...
Green
159                       kawacom uganda ltd        1750  ...
Green
172                       pt. olam indonesia        1400  ...    Blue-
Green
231   pt. royal pacific indah international        None  ...  Bluish-
Green
342    bulamburi coffee farmers association        1800  ...
Green
380                 bugisu cooperative union        1800  ...
Green
420                       kawacom uganda ltd        None  ...
Green
456                    kabum trading company        1800  ...  Bluish-
Green
508                           star cafe ltd        1800  ...
Green
544                       kawacom uganda ltd   1400-1900  ...
Green
```

```
545       nyapea coffee farmers association        1400  ...
Green
619                                    afca  1400- 1900  ...
Green
642                      kawacom uganda ltd   1400-1900  ...
Green
1008                   kyagalanyi coffee ltd        1700  ...
Green
1050                      pt. olam indonesia   1200-1500  ...    Blue-
Green
1233              pt. shriya artha nusantara        1300  ...
None

     category_two_defects           expiration  \
27                    1.0      June 26th, 2015
53                    0.0      June 30th, 2015
68                    1.0      June 26th, 2015
71                    5.0      June 26th, 2015
96                    1.0      July 24th, 2018
116                   1.0      June 30th, 2015
126                   7.0  November 24th, 2017
159                   1.0     March 14th, 2018
172                   3.0     March 14th, 2018
231                   0.0       May 24th, 2018
342                   5.0      June 27th, 2015
380                   1.0      July 21st, 2017
420                   1.0       May 19th, 2017
456                   0.0   October 28th, 2015
508                   0.0      June 26th, 2015
544                   0.0      June 27th, 2015
545                   4.0      June 27th, 2015
619                   0.0      June 25th, 2015
642                   0.0      June 30th, 2015
1008                  3.0      June 30th, 2015
1050                  1.0   November 9th, 2016
1233                  1.0      July 18th, 2018

                              certification_body  \
27          Uganda Coffee Development Authority
53          Uganda Coffee Development Authority
68          Uganda Coffee Development Authority
71          Uganda Coffee Development Authority
96          Uganda Coffee Development Authority
116         Uganda Coffee Development Authority
126   Specialty Coffee Association of Indonesia
159         Uganda Coffee Development Authority
172                Specialty Coffee Association
231   Specialty Coffee Association of Indonesia
342         Uganda Coffee Development Authority
380         Uganda Coffee Development Authority
```

```
420         Uganda Coffee Development Authority
456         Uganda Coffee Development Authority
508         Uganda Coffee Development Authority
544         Uganda Coffee Development Authority
545         Uganda Coffee Development Authority
619         Uganda Coffee Development Authority
642         Uganda Coffee Development Authority
1008        Uganda Coffee Development Authority
1050              Specialty Coffee Association
1233  Specialty Coffee Association of Indonesia

                         certification_address  \
27    188fe373b511e21f614564bf86aa4774270d8e04
53    188fe373b511e21f614564bf86aa4774270d8e04
68    188fe373b511e21f614564bf86aa4774270d8e04
71    188fe373b511e21f614564bf86aa4774270d8e04
96    188fe373b511e21f614564bf86aa4774270d8e04
116   188fe373b511e21f614564bf86aa4774270d8e04
126   99fa73db21b7acd9c9ceb9dd84e409d2077d55c4
159   188fe373b511e21f614564bf86aa4774270d8e04
172   36d0d00a3724338ba7937c52a378d085f2172daa
231   99fa73db21b7acd9c9ceb9dd84e409d2077d55c4
342   188fe373b511e21f614564bf86aa4774270d8e04
380   188fe373b511e21f614564bf86aa4774270d8e04
420   188fe373b511e21f614564bf86aa4774270d8e04
456   188fe373b511e21f614564bf86aa4774270d8e04
508   188fe373b511e21f614564bf86aa4774270d8e04
544   188fe373b511e21f614564bf86aa4774270d8e04
545   188fe373b511e21f614564bf86aa4774270d8e04
619   188fe373b511e21f614564bf86aa4774270d8e04
642   188fe373b511e21f614564bf86aa4774270d8e04
1008  188fe373b511e21f614564bf86aa4774270d8e04
1050  36d0d00a3724338ba7937c52a378d085f2172daa
1233  99fa73db21b7acd9c9ceb9dd84e409d2077d55c4

                         certification_contact unit_of_measurement  \
27    b7614767a5343729bbde3a2777c60ce836aed928                    m
53    b7614767a5343729bbde3a2777c60ce836aed928                    m
68    b7614767a5343729bbde3a2777c60ce836aed928                    m
71    b7614767a5343729bbde3a2777c60ce836aed928                    m
96    b7614767a5343729bbde3a2777c60ce836aed928                    m
116   b7614767a5343729bbde3a2777c60ce836aed928                    m
126   36910838db193ebdd61fa1427bac74622114c49a                    m
159   b7614767a5343729bbde3a2777c60ce836aed928                    m
172   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
231   36910838db193ebdd61fa1427bac74622114c49a                    m
342   b7614767a5343729bbde3a2777c60ce836aed928                    m
380   b7614767a5343729bbde3a2777c60ce836aed928                    m
420   b7614767a5343729bbde3a2777c60ce836aed928                    m
456   b7614767a5343729bbde3a2777c60ce836aed928                    m
```

```
508   b7614767a5343729bbde3a2777c60ce836aed928                    m
544   b7614767a5343729bbde3a2777c60ce836aed928                    m
545   b7614767a5343729bbde3a2777c60ce836aed928                    m
619   b7614767a5343729bbde3a2777c60ce836aed928                    m
642   b7614767a5343729bbde3a2777c60ce836aed928                    m
1008  b7614767a5343729bbde3a2777c60ce836aed928                    m
1050  0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                    m
1233  36910838db193ebdd61fa1427bac74622114c49a                    m

     altitude_low_meters altitude_high_meters altitude_mean_meters
27                1950.0                1950.0                1950.0
53                1400.0                1900.0                1650.0
68                1600.0                1600.0                1600.0
71                1950.0                1950.0                1950.0
96                1800.0                1800.0                1800.0
116               1400.0                1900.0                1650.0
126               1200.0                1200.0                1200.0
159               1750.0                1750.0                1750.0
172               1400.0                1400.0                1400.0
231                  NaN                   NaN                   NaN
342               1800.0                1800.0                1800.0
380               1800.0                1800.0                1800.0
420                  NaN                   NaN                   NaN
456               1800.0                1800.0                1800.0
508               1800.0                1800.0                1800.0
544               1400.0                1900.0                1650.0
545               1400.0                1400.0                1400.0
619               1400.0                1900.0                1650.0
642               1400.0                1900.0                1650.0
1008              1700.0                1700.0                1700.0
1050              1200.0                1500.0                1350.0
1233              1300.0                1300.0                1300.0

[22 rows x 43 columns]
```

```
coffee_ratings_cluster.loc[:,'variety'] =
coffee_ratings_cluster['variety'].astype('category').cat.remove_unused
_categories() #yanşışlıkla farklı category gelirse siler. cat ->
category olduğu için sr-tring olsaydı str yazılırdı

#random olarak seçtiğimiz varietylerin hepsinden bir örnek ver
coffee_ratings_cluster.groupby('variety', observed=True).sample(n=1,
random_state=2021)['variety']
coffee_ratings_cluster.groupby('variety', observed=True).sample(n=1,
random_state=2021)
```

```
      total_cup_points  species                     owner
country_of_origin  \
1233             78.17  Arabica             sanjava coffee
Indonesia
```

```
172               84.42  Arabica          aulia arif syahri
Indonesia
71                85.42  Arabica  great lakes coffee uganda
Uganda

                  farm_name           lot_number                mill  \
1233                various            sran-ijen           east java
172                 darmawi  MANDHELING BRASTAGI            dry mill
71    chesiyo farmer group                 None  great lakes coffee

        ico_number                      company altitude  ...
color  \
1233          None  pt. shriya artha nusantara     1300  ...
None
172   To be advice          pt. olam indonesia     1400  ...  Blue-
Green
71                0          great lakes coffee     1950  ...
Green

     category_two_defects        expiration  \
1233                  1.0   July 18th, 2018
172                   3.0  March 14th, 2018
71                    5.0   June 26th, 2015

                            certification_body  \
1233  Specialty Coffee Association of Indonesia
172                Specialty Coffee Association
71           Uganda Coffee Development Authority

                         certification_address  \
1233  99fa73db21b7acd9c9ceb9dd84e409d2077d55c4
172   36d0d00a3724338ba7937c52a378d085f2172daa
71    188fe373b511e21f614564bf86aa4774270d8e04

                         certification_contact unit_of_measurement  \
1233  36910838db193ebdd61fa1427bac74622114c49a                   m
172   0878a7d4b9d35ddbf0fe2ce69a2062cceb45a660                   m
71    b7614767a5343729bbde3a2777c60ce836aed928                   m

     altitude_low_meters altitude_high_meters altitude_mean_meters
1233              1300.0               1300.0                1300.0
172               1400.0               1400.0                1400.0
71                1950.0               1950.0                1950.0

[3 rows x 43 columns]
```