SÜLEYMAN DEMİREL ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
VERİ MADENCİLİĞİ FİNAL SINAV SORULARI

ADI SOYADI:…….NO:……

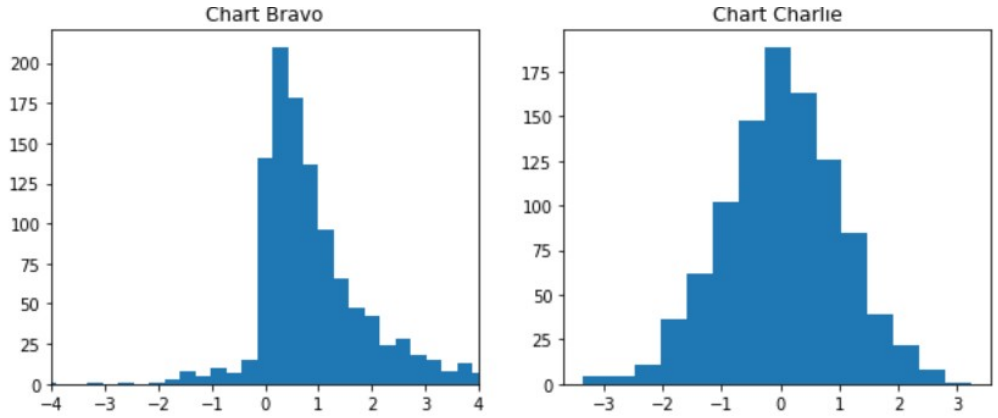
Sınav Yönergesi:
1. Sınav Süresi 30 dakikadır
2. Soruların puanları üzerinde belirtildiği gibidir.

1. Airbnb'de veri bilimci olduğunuzu ve belirli bir bölgedeki mülkler için en uygun gecelik kiraları önerecek bir model oluşturmak istediğinizi varsayalım.

a) Ne tür bir model kullanırsınız? Neden?
b) Output ne olurdu?
c) Hangi input'ları kullanırsınız? (en az iki tanesini belirtin)

a) Regresyon modelleri kullanılır. Sayısal değerleri tahmin etmede regresyon modelleri kullanılır.
b) Output gecelik kiradır.
c) Yatak odası sayısı, metrekare cinsinden büyüklük ve toplu taşımaya, şehir merkezine veya gece hayatı seçeneklerine olan mesafe olabilir.

2. Aşağıdaki grafiklerden hangisinin normal dağılıma sahip verilerden çizilmesi en olasıdır? Grafiklerin her biri için, grafiğin neden normal dağılıma sahip verileri görüntüleyip görüntülemediğini açıklayın?



a) Grafik Bravo büyük ölçüde sağa çarpıktır ve bu nedenle normal dağılıma sahip verilerden çizilmemiştir.
b) Chart Charlie simetriktir ve çan eğrisi şeklinde eşit olarak dağılmıştır ve gerçekten de normal dağılımın rastgele bir örneğinden alınmıştır.

3. Aşağıdaki tablo, bebeklerin rahimde üç tür 'müzik' dinlediği, daha sonra emekleme/yürüme ilerlemelerinin gözlemlendiği ve erken, zamanında veya geç olarak kategorize edildiği bir çalışmayla ilgilidir. Araştırmanın amacı; rahimde müzik dinleme ile bebeklerde emekleme/yürüme ilerleme süresi arasında istatistiksel olarak anlamlı bir ilişki olup olmadığını belirlemektir.

	Emekleme/Yürümede İlerleme			**Örnek Boyutu**
Rahimdeki Müzik	**Erken**	**Zamanında**	**Geç**	**N**
Mozart (Piano Sonata)	%50	%30	%20	63
Philip Glass ((minimalist music))	%40	%35	%25	60
Beyaz Gürültü ve Sessizlik	%20	%25	%55	44

Yukarıdaki tabloda satır yüzdeleri ve örnek boyutları verilmektedir. Örneğin, 'Mozart' dinleyen bebeklerinin %50'si emekleme/yürüme'de 'Erken' ilerleme kaydetmiştir ve toplam 63 'Mozart' dinleyen bebek bulunmaktadır.

a) Sağlanan satır yüzdelerine dayanarak, aşağıdaki frekans tablosunu en yakın tam sayıya yuvarlayarak doldurun.

	Emekleme/Yürümede İlerleme		
Rahimdeki Müzik	**Erken**	**Zamanında**	**Geç**
Mozart (Piano Sonata)	**31**	**19**	**13**
Philip Glass ((minimalist music))	**24**	**21**	**15**
Beyaz Gürültü ve Sessizlik	**9**	**11**	**24**

b) Bu araştırma için gerekli hipotezleri yazınız ? Bu hipotezleri test etmek için hangi istatistiksel test kullanılmalıdır yazınız?

H_O : Rahimde dinelenen müzik ile Emekleme/Yürüme arasında bir ilişki yoktur.
H_1 : Rahimde dinelenen müzik ile Emekleme/Yürüme arasında bir ilişki vardır.

χ2 testi kullanılır.

4. Bir alandaki aykırı değerler(outliers) IQR kullanılarak belirlenebilir. Q1 – 1,5 IQR'nin altına düşen veya Q3 + 1,5 IQR'nin üzerinde kalan veri noktaları aykırı değerlerdir. Bu bilgi dahilinde kendisine bir dataframe parametre olarak gönderildiğinde o dataframe'deki sayısal alanlara ait aykırı değerleri bir dictionary olarak geri döndüren fonskiyonu yazınız? Örnek Çıktı : {'Yas' : [8, 12, 5], 'Maas': [25000, 28000, 1000]}

```
import numpy as np


def get_outliers(incoming_df):
    result = {}
    num_df = incoming_df.select_dtypes(include=np.number)

    for col in num_df.columns:
        q3, q1 = np.percentile(incoming_df[col], [75 ,25])
        iqr = q3 - q1
        outliers = df[col][(df[col] < q1 - 1.5 * iqr) | (df[col] > q3 + 1.5 * iqr)].tolist()
        result[col] = outliers
    return result
```