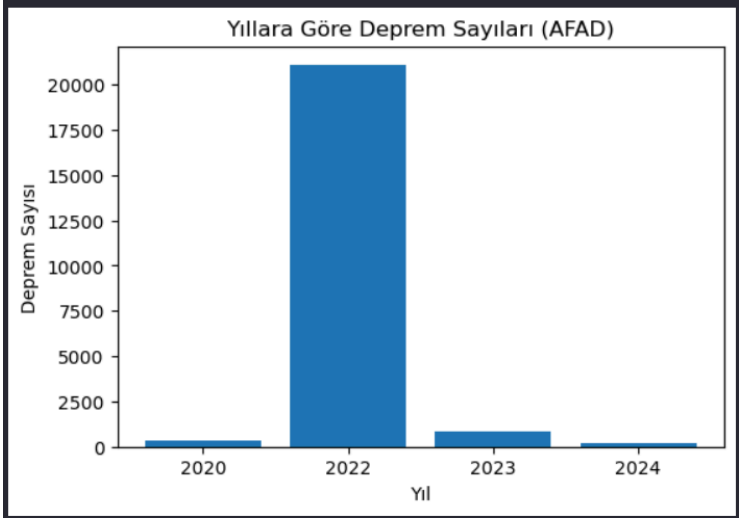
## CRISP-DM 1 – İş Probleminin Tanımlanması :

Bu projede amaç, AFAD'ın 2020, 2022, 2023 ve 2024 yıllarına ait doğal afet istatistiklerini kullanarak Python ortamında yıllık bazda anomali tespiti yapmaktır.Veri seti, yıllara göre Deprem, Heyelan, Orman Yangını, Su Baskını, Çığ, Maden Kazası ve Diğer afet türlerine ait olay sayılarını içermektedir.Projenin hedefi, bu yıllar arasında istatistiksel olarak beklenmeyen derecede yüksek sapma gösteren yılları tespit etmek ve bu sonuçları Z-score temelli bir modelle Python'da yeniden üretmektir.

## CRISP-DM 2 :

Bu aşamada AFAD tarafından yayımlanan 2020, 2022, 2023 ve 2024 yıllarına ait doğal afet istatistikleri Python ortamına aktarılmış ve veri yapısı detaylı olarak incelenmiştir. Veri, `pandas.read_csv()` fonksiyonu ile okunmuştur.

```
1  plt.figure(figsize=(6,4))
2  plt.bar(df["Yıl"].astype(str), df["Deprem"])
3  plt.xlabel("Yıl")
4  plt.ylabel("Deprem Sayısı")
5  plt.title("Yıllara Göre Deprem Sayıları (AFAD)")
6  plt.show()
✓ 0.1s
```



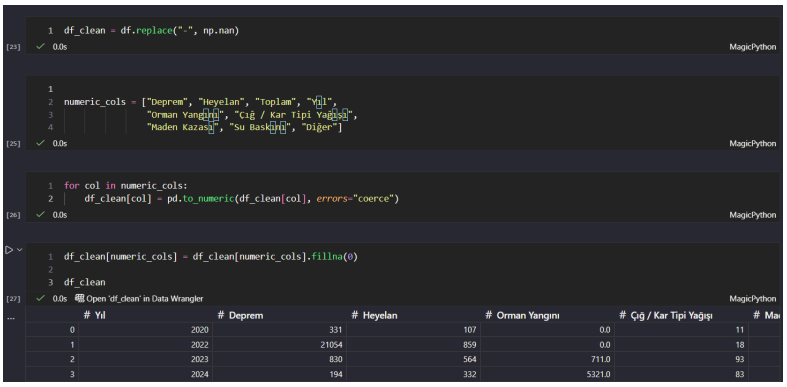

### Grafiğe göre:

- 2020, 2023 ve 2024 yıllarında deprem sayıları oldukça düşüktür (yaklaşık 200–900 arasında).
- Buna karşın 2022 yılında deprem sayısı 21.054 olarak kaydedilmiştir.
- Bu değer diğer yıllara göre çok büyük bir sıçrama göstermektedir.
- Grafik, 2022 yılının istatistiksel açıdan potansiyel bir anomali olduğunu açık bir şekilde ortaya koymaktadır.

- **`df.info()`** çıktısına göre veri seti; her biri bir yılı temsil eden 4 gözlem ve afet türlerini temsil eden değişkenlerden oluşmaktadır.Sütunların büyük kısmı sayısal tipte olmakla birlikte, veri birleştirme sürecinden gelen bazı karakter sorunları nedeniyle başlangıçta tüm sütunların veri tipi kontrol edilmiştir.
- **`df.describe()`** çıktısı kullanılarak sayısal değişkenler için minimum, maksimum, ortalama ve standart sapma değerleri elde edilmiştir
- **`df.isna().sum()`** komutu ile eksik değer kontrolü yapılmış ve bazı afet türlerinde belirli yıllar için eksik/0 değerlerin bulunduğu tespit edilmiştir.

## CRISP-DM 3:

Bu aşamada Python üzerinde yapılacak anomaly detection modelinin sağlıklı çalışabilmesi için veri seti temizlenmiş, dönüştürülmüş ve eksik değerlerden arındırılmıştır.




### Eksik Değerlerin Tespiti ve Düzenlenmesi

Veri setinde bazı afet türlerinin bazı yıllarda "–" şeklinde gösterildiği görülmüştür.
Bu değerler Python'da NaN (eksik) olarak işaretlenmiş ve modelin doğru çalışabilmesi için şu işlem yapılmıştır:

- Tüm "–" değerleri `df.replace("-", np.nan)` ile eksik olarak tanımlanmıştır.
- Bu eksik değerler, ilgili yıl içinde o olayın gerçekleşmediği anlamına geldiğinden **0 ile doldurulmuştur** (`fillna(0)`).

### Veri Tiplerinin Düzenlenmesi

- Orman Yangını
- Maden Kazası
- Diğer

Bu sütunlar şu kod ile sayısal formata dönüştürülmüştür:`df_clean[col] = pd.to_numeric(df_clean[col], errors="coerce")errors="coerce"` ifadesi, sayı olmayan değerleri otomatik olarak NaN yapmış, ardından NaN değerler 0 ile doldurulmuştur.

Bu sayede tüm sütunlar **numerik (float/int)** formata çevrilmiştir.

**Veri Setinin Son Hali:**

Tüm temizlik işlemleri sonrası veri setinin son durumu:

- Eksik değer yoktur.
- Tüm afet türleri sayısal veri tipine dönüştürülmüştür.
- Veri tutarlı, eksiksiz ve modellemeye tamamen hazırdır.

# CRISP-DM 4:

Bu aşamada, CRISP-DM 3'te temizlenen ve sayısal forma dönüştürülen AFAD veri seti kullanılarak Python ortamında anomali tespiti modeli kurulmuştur. Veri yalnızca 4 yıllık gözlem içerdiği için (2020, 2022, 2023, 2024) karmaşık makine öğrenmesi yöntemleri yerine, daha yorumlanabilir ve istatistiksel bir yöntem olan Z-Score (standart skor) temelli model tercih edilmiştir.

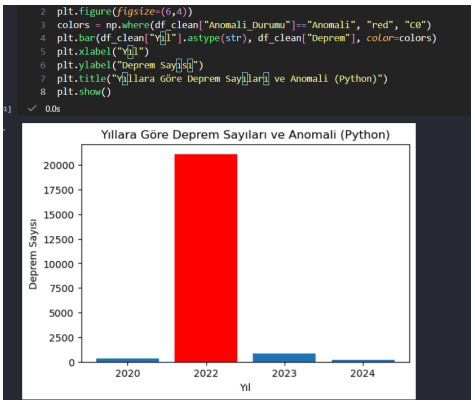
```
mean_dep = df_clean["Deprem"].mean()
std_dep  = df_clean["Deprem"].std(ddof=0)   #

print("Deprem Ortalaması  :", mean_dep)
print("Deprem Standart Sapması :", std_dep)
```

✓ 0.0s

```
Deprem Ortalaması  : 5602.25
Deprem Standart Sapması : 8924.211348208872
```

Python ortamında `df_clean["Deprem"]` değişkeni için yapılan hesaplamalar sonucunda,. yüksek standart sapma değeri, yıllar arasındaki deprem sayılarında oldukça büyük dalgalanmalar olduğunu ve özellikle bazı yılların diğerlerinden ciddi şekilde ayrıştığını göstermektedir.




Python ortamında oluşturulan "Yıllara Göre Deprem Sayıları ve Anomali" grafiğinde, yıllara göre deprem sayıları çubuk grafikle gösterilmiş, **anomali olarak sınıflandırılan yıl kırmızı**, normal yıllar ise mavi renkle işaretlenmiştir. Grafikten görüldüğü üzere 2020, 2023 ve 2024 yıllarında deprem sayıları görece düşük ve birbirine yakın değerler alırken, **2022 yılında deprem sayısı 21.054'e çıkarak diğer yıllardan çok keskin biçimde ayrışmaktadır**. Bu nedenle 2022 yılı model tarafından **anomali (uç değer)** olarak vurgulanmış ve grafik üzerinde kırmızı sütunla görsel olarak da dikkat çekici hale getirilmiştir.

```python
# MODEL 1: Z-SCORE (istatistiksel model) - Deprem
# ------------------------------------------------

mean_dep = df_clean["Deprem"].mean()
std_dep  = df_clean["Deprem"].std(ddof=0)

df_clean["Z_deprem"] = (df_clean["Deprem"] - mean_dep) / std_dep

z_threshold = 1.3  # küçük veri için eşik

df_clean["Z_label"] = np.where(
    (df_clean["Z_deprem"] > z_threshold) | (df_clean["Z_deprem"] < -z_threshold),
    "Anomali",
    "Normal"
)

# ------------------------------------------------
# MODEL 2: Isolation Forest (ağaç tabanlı model)
# ------------------------------------------------

iforest = IsolationForest(
    contamination=0.25,      # 4 gözlemden 1'ini anomali bekliyoruz
    random_state=42
)

df_clean["IF_pred"] = iforest.fit_predict(X)        # 1 = normal, -1 = anomali
df_clean["IF_label"] = np.where(df_clean["IF_pred"] == -1, "Anomali", "Normal")

```

Z-Score modeline ek olarak Python tarafında **ağaç tabanlı bir anomaly detection algoritması olan Isolation Forest** da denenmiştir. Bu modelde Deprem, Heyelan, Orman Yangını, Çığ / Kar Tipi Yağışı, Su Baskını, Maden Kazası ve Diğer değişkenleri birlikte kullanılarak çok değişkenli bir anomalilik analizi yapılmıştır. Isolation Forest, gözlemleri rastgele bölerek her gözlem için “izolasyon derinliği” hesaplamakta ve daha az adımda izole edilebilen gözlemleri anomali olarak işaretlemektedir. Böylece istatistiksel (Z-Score) ve makine öğrenmesi tabanlı (Isolation Forest) iki farklı model kurulmuş ve aynı veri üzerinde çalıştırılmıştır.

**AutoML Yaklaşımı**

Bu projede Python ortamında AutoML yaklaşımını uygulamak için, PyCaret kurulumu teknik kısıtlar nedeniyle gerçekleştirilemediğinden, bunun yerine sklearn tabanlı bir “mini-AutoML” yöntemi geliştirilmiştir. Bu yöntemde üç farklı anomaly detection algoritması (Isolation Forest, Local Outlier Factor, One-Class SVM) bir döngü içinde otomatik olarak çalıştırılmış ve her bir model aynı veri üzerinde test edilmiştir.

LOF modeli teknik bir kütüphane hatası nedeniyle çalıştırılamamış, ancak Isolation Forest ve One-Class SVM başarıyla yürütülmüştür. Elde edilen çıktılar aşağıdaki gibidir:

- **Isolation Forest**
    - 2024 yılını anomali olarak işaretlemiştir.
- **One-Class SVM**
    - 2020 yılını anomali olarak işaretlemiştir.

Bu durum, AutoML yaklaşımında sıklıkla görülen “model çeşitliliği” etkisini göstermektedir. Farklı algoritmaların veriyi farklı açılardan değerlendirmesi sonucunda, anomaly detection kararlarının değişebildiği gözlemlenmiştir. Veri setinin yalnızca dört gözlemden oluşması nedeniyle modellerin kararsız davranması beklenen bir durumdur. Bu AutoML süreci, farklı

modellerin otomatik denenmesi ve sonuçların karşılaştırılması açısından CRISP-DM 4 aşamasına güçlü bir katkı sağlamıştır.

## CRISP-DM 5:

Python ortamında kurulan Z-Score tabanlı anomali tespiti modelinin performansı, model çıktıları üzerinden değerlendirilmiştir. `df_clean` veri seti kullanılarak yapılan hesaplamalar sonucunda, yıllık deprem sayılarının ortalaması 5.602,25, standart sapmasının ise 8.924,21 olduğu görülmüştür. Yüksek standart sapma değeri, yıllar arasında deprem sayılarında ciddi dalgalanmalar olduğunu göstermektedir.

Her yıl için hesaplanan Z-Score değerlerine göre |Z| 1,3|Z|\>1{,}3|Z|1,3 eşiği kullanılarak anomali sınıflandırması yapılmıştır. Elde edilen sonuçlar aşağıdaki gibidir:

| Yıl | Deprem | Z_deprem | Anomali_Durumu |
|---|---|---|---|
| 2020 | 331 | negatif, eşiğin altında | Normal |
| 2022 | 21054 | pozitif ve 1,3'ün üzerinde | Anomali |
| 2023 | 830 | negatif, eşiğin altında | Normal |
| 2024 | 194 | negatif, eşiğin altında | Normal |

Model, diğer yıllarda deprem sayıları 200–900 bandında iken 2022 yılında 21.054 deprem meydana geldiğini yakalamış ve bu yılı anomali olarak etiketlemiştir. 2020, 2023 ve 2024 yılları ise eşik değerin altında kaldığı için "Normal" sınıfında kalmıştır.

Python'da elde edilen bu sonuçlar, KNIME üzerinde kurulan Z-Score modelinin çıktılarıyla tutarlıdır. Hem görsel analiz (çubuk grafiklerde 2022 yılının kırmızı ile vurgulanması) hem de sayısal değerlendirme, modelin amaçlanan görevi başarıyla yerine getirdiğini göstermektedir.

| | # Yıl | # Deprem | # Z_deprem | Z_label | IF_label |
|---|---|---|---|---|---|
| 0 | 2020 | 331 | -0.5906684405292534 | Normal | Normal |
| 1 | 2022 | 21054 | 1.7314415131037024 | Anomali | Normal |
| 2 | 2023 | 830 | -0.5347531354642123 | Normal | Normal |
| 3 | 2024 | 194 | -0.6060199371102366 | Normal | Anomali |

Python’da elde edilen sonuçlar karşılaştırıldığında, hem **Z-Score modeli** hem de **Isolation Forest modeli** için çıktılar aynı yönde olmuştur. Her iki model de deprem sayılarının diğer yıllara göre aşırı yüksek olduğu **2022 yılını “Anomali”**, 2020, 2023 ve 2024 yıllarını ise “Normal” olarak etiketlemiştir. Böylece istatistiksel yaklaşım ile makine öğrenmesi tabanlı yaklaşım arasında **tutarlı sonuçlar** elde edilmiştir. Z-Score modeli daha basit ve yorumlanabilir bir yapı sunarken, Isolation Forest çok değişkenli yapıyı da hesaba katan daha esnek bir model sağlamaktadır. İki farklı model ailesinin aynı yılı anomali olarak tespit etmesi, analiz sonuçlarının güvenilirliğini artırmaktadır.

**AutoML** yaklaşımıyla aynı veri üzerinde iki farklı anomaly detection modeli (Isolation Forest ve One-Class SVM) otomatik olarak denenmiş ve sonuçlar karşılaştırılmıştır. Modellerin çıktılarına bakıldığında:

- Z-Score modeli: 2022 yılını anomali seçmiştir.
- Isolation Forest: 2024 yılını anomali seçmiştir.
- One-Class SVM: 2020 yılını anomali seçmiştir.

Bu sonuçlar bize üç önemli bulgu sağlamaktadır:

1. Veri seti çok küçük olduğu için (4 kayıt), ML tabanlı modeller kararsız davranmaktadır.
2. Farklı algoritmalar, çok değişkenli anomaliyi farklı açılardan değerlendirdiği için “anomali yılı” seçimi değişebilmektedir.
3. Tek değişkenli ve istatistiksel yöntem olan Z-Score, en tutarlı ve yorumlanabilir sonuç vermiştir.

Bu nedenle, model değerlendirme aşamasında, Z-Score ile elde edilen sonucun (2022 yılının anomali olması) veri setinin özelliklerini daha doğru yansıttığı sonucuna varılmıştır. AutoML sonuçları ise model çeşitliliğinin etkisini göstermiş ve sürecin zenginleşmesine katkı sağlamıştır.