Görev 1 (Kolay): "mtcars" Veri Setinde Basit Rastgele Bölme

Amaç: "mtcars" veri setini "%60 eğitim" ve "%40 test" setlerine bölün. Eğitim ve test setlerindeki "mpg" (miles per gallon) değişkeninin ortalamalarının benzer olduğunu gösterin.

```
>
> # Veri setini yükle
> data(mtcars)
>
> # mtcars veri setinin kullanılması
> set.seed(123)  # Tekrar üretilebilirlik için rastgelelik ayarlanıyor
> split <- initial_split(mtcars, prop = 0.6)  # mtcars veri seti %60 eğitim, %40 test olarak bölünüyor
>
> # Eğitim ve test setlerini oluşturma
> train_data <- training(split)  # Eğitim verisi
> test_data <- testing(split)  # Test verisi
>
> # Eğitim setindeki "mpg" değişkeninin ortalamasını hesaplama
> mean_train_mpg <- mean(train_data$mpg)
> cat("Eğitim Setindeki mpg Ortalaması:", mean_train_mpg, "\n")
Eğitim Setindeki mpg Ortalaması: 21.04737
>
> # Test setindeki "mpg" değişkeninin ortalamasını hesaplama
> mean_test_mpg <- mean(test_data$mpg)
> cat("Test Setindeki mpg Ortalaması:", mean_test_mpg, "\n")
Test Setindeki mpg Ortalaması: 18.69231
>
```

Açıklama: Veri setini istenilen şekilde rastgele böldüğümüzde “mpg” değişkenin ortalama değerleri farklı çıktığı gözlemlenmiştir.

Görev 2 (Normal): "PimaIndiansDiabetes" Veri Setinde Stratified Sampling

Amaç: Diyabet veri setini ("mlbench" paketinden "PimaIndiansDiabetes"), stratified sampling ile "%70 eğitim" ve "%30 test" setlerine bölün. "diabetes" değişkeninin dağılımının eğitim-test setlerinde orijinal veriyle aynı olduğunu doğrulayın.

```
>
> # Veri setini yükle
> data("PimaIndiansDiabetes")
>
> # Stratified Sampling ile verilerin dagitilmasi (%70 Eğitim, %30 Test)
> set.seed(123) # Tekrar üretilebilirlik için rastgelelik ayarlanıyor
> split <- initial_split(PimaIndiansDiabetes, prop = 0.7, strata = "diabetes")
>
> # Eğitim ve test setlerini oluştur
> train_data <- training(split)
> test_data <- testing(split)
>
> # Orijinal veri setindeki sınıf dağılımı
> table_original <- prop.table(table(PimaIndiansDiabetes$diabetes))
>
> # Eğitim setindeki sınıf dağılımı
> table_train <- prop.table(table(train_data$diabetes))
>
> # Test setindeki sınıf dağılımı
> table_test <- prop.table(table(test_data$diabetes))
>
> cat("Orijinal Veri Seti:\n", table_original,
+     "\nEğitim Seti:\n", table_train,
+     "\nTest Seti:\n", table_test, "\n")
Orijinal Veri Seti:
 0.6510417 0.3489583
Eğitim Seti:
 0.6517691 0.3482309
Test Seti:
 0.6493506 0.3506494
>
```

Açıklama: Kullanılan veri setindeki verilerin orantısız dağılmasının, eğitim ve test veri setlerinde de benzer şekilde orantısız bir dağılım yaratması olasılığına karşı, 'Stratified Sampling' yöntemi uygulanmıştır.

Görev 3 (Zor): "diamonds" Veri Setinde Cross Validation ve RMSE Hesabı

Amaç: "diamonds" veri setinde ("ggplot2" paketinde yer alır), fiyat ("price") tahmini için 5 katlı cross validation uygulayın. Her kat için RMSE hesaplayın ve ortalama RMSE'yi raporlayın.

```
> library(ggplot2)
> library(tidymodels)
>
> # Veri setini yükle
> data("diamonds")
>
> # Küçük bir örneklem alalım (hesaplamaları hızlandırmak için)
> set.seed(123)
> diamonds_sample <- diamonds[sample(nrow(diamonds), 50000), ]
>
> # 5 Katlı Cross Validation
> set.seed(123)  # Tekrarlanabilirlik için
> cv_folds <- vfold_cv(diamonds_sample, v = 5)
>
> # RMSE hesaplamak için fonksiyon
> calculate_rmse <- function(split) {
+   train_data <- analysis(split)  # Eğitim verisi
+   test_data <- assessment(split) # Test verisi
+
+   # Lineer regresyon modeli (basit model: carat ile price tahmini)
+   model <- lm(price ~ carat, data = train_data)
+
+   # Test verisi üzerinde tahmin yap
+   predictions <- predict(model, newdata = test_data)
+
+   # Gerçek fiyatlarla karşılaştırarak RMSE hesapla
+   rmse_val <- yardstick::rmse_vec(truth = test_data$price, estimate = predictions)
+
+   return(rmse_val)
+ }
>
```

```
>
> # Her kat için RMSE hesapla
> rmse_values <- map_dbl(cv_folds$splits, calculate_rmse)
>
> # Sonuçları göster
> cat("Her kat için RMSE:\n")
Her kat için RMSE:
> print(rmse_values)
[1] 1529.409 1582.605 1520.538 1515.431 1580.103
>
> # Ortalama RMSE'yi hesapla
> mean_rmse <- mean(rmse_values)
> cat("\nOrtalama RMSE:", mean_rmse, "\n")

Ortalama RMSE: 1545.617
>
```

Açıklama: Veri setimizde uyguladığımız 'Cross Validation' yöntemi ile, farklı ancak kendi verilerimizden çok da uzak olmayan kombinasyonlarla verilerimiz ayrılmıştır. Ayrılan bu verilerle oluşturulan lineer regresyon modeli üzerinde tahmin işlemleri gerçekleştirilmiş ve her bir modelin çıktıları karşılaştırılmıştır.