Veri Temizleme ve Eksik Değer İmputation

1. Giriş ve Amaç

Bu çalışmada, Ames Housing veri seti kullanılarak eksik değerlerin imputation işlemi gerçekleştirilmiştir. Amacımız, eksik verileri farklı yöntemlerle doldurmak ve ardından verilerin görselleştirilmesini sağlamak. Kullanılan yöntemler:

- Ortalama ve mod ile doldurma (Mean/Mode Imputation)
- KNN (K-Nearest Neighbors) ile doldurma
- Karar ağaçları (Tree-based imputation) ile doldurma

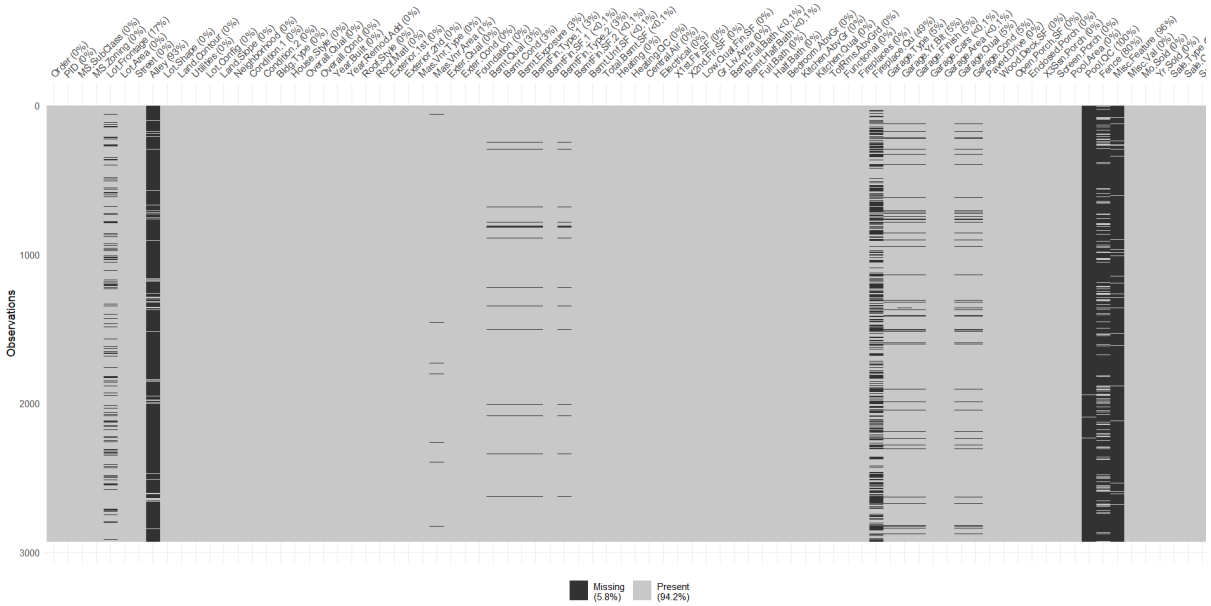
2. Veri Yükleme ve Görselleştirme

Öncelikle veri seti, `read.csv()` fonksiyonu ile yüklenmiştir:

```
data <- read.csv("D:/ames.csv")
```

Yüklenen verinin eksik değerlerini görselleştirmek için `visdat` paketinin `vis_miss()` fonksiyonu kullanılmıştır. Bu, verinin hangi sütunlarında eksik değer bulunduğunu ve eksik veri oranını göstermektedir.

```
vis_miss(data)
```



3. Eksik Değerlerin İmputation'ı

Eksik verilerin doldurulması için üç farklı yöntem uygulanmıştır:

a. Ortalama ve Mod İle Doldurma (Mean/Mode Imputation)

Bu adımda, sayısal veriler için ortalama (`mean`) ve kategorik veriler için mod (`mode`) kullanılarak eksik veriler doldurulmuştur. Bunun için `recipes` paketinin `step_impute_mean()` ve `step_impute_mode()` fonksiyonları kullanılmıştır. Bu adımlar, veri setindeki eksik değerlerin yerini almıştır.

```
rec_mean <- recipe(~ ., data = data) %>%
  step_impute_mean(all_numeric(), -all_outcomes()) %>%
  step_impute_mode(all_nominal(), -all_outcomes()) %>%
  prep()

data_mean <- bake(rec_mean, new_data = NULL)
```

b. KNN (K-Nearest Neighbors) İle Doldurma

KNN algoritması kullanılarak eksik değerler en yakın komşuların verilerine göre doldurulmuştur. Burada `VIM` paketinin `kNN()` fonksiyonu kullanılmıştır.

```
data_knn <- kNN(data, k = 5)
```

c. Karar Ağaçları (Tree-based imputation)

Verideki eksik değerlerin tahmin edilmesi için `mice` paketindeki `cart` (Classification And Regression Trees) metodu kullanılmıştır. Bu yöntemle eksik veriler karar ağaçları kullanılarak doldurulmuştur.

```
library(mice)
imp_tree <- mice(data, method = "cart", m = 1)
data_tree <- complete(imp_tree)
```

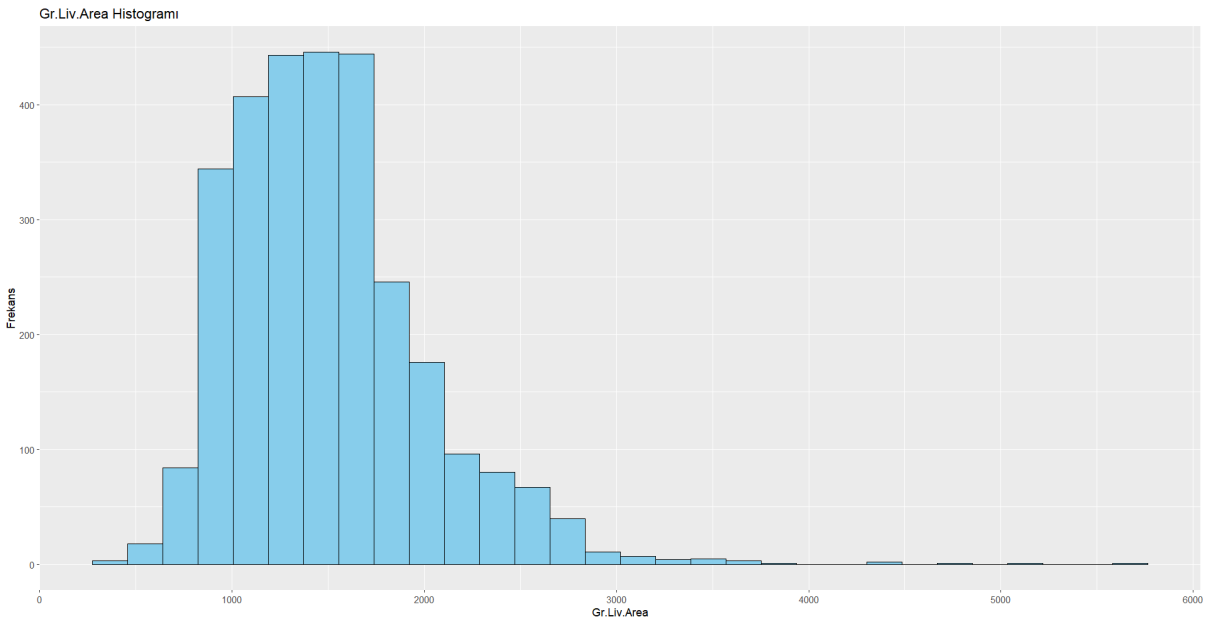
4. Veri Görselleştirme

İmputation sonrası verilerin nasıl dağıldığını görmek için histogramlar ve boxplot'lar kullanılmıştır.

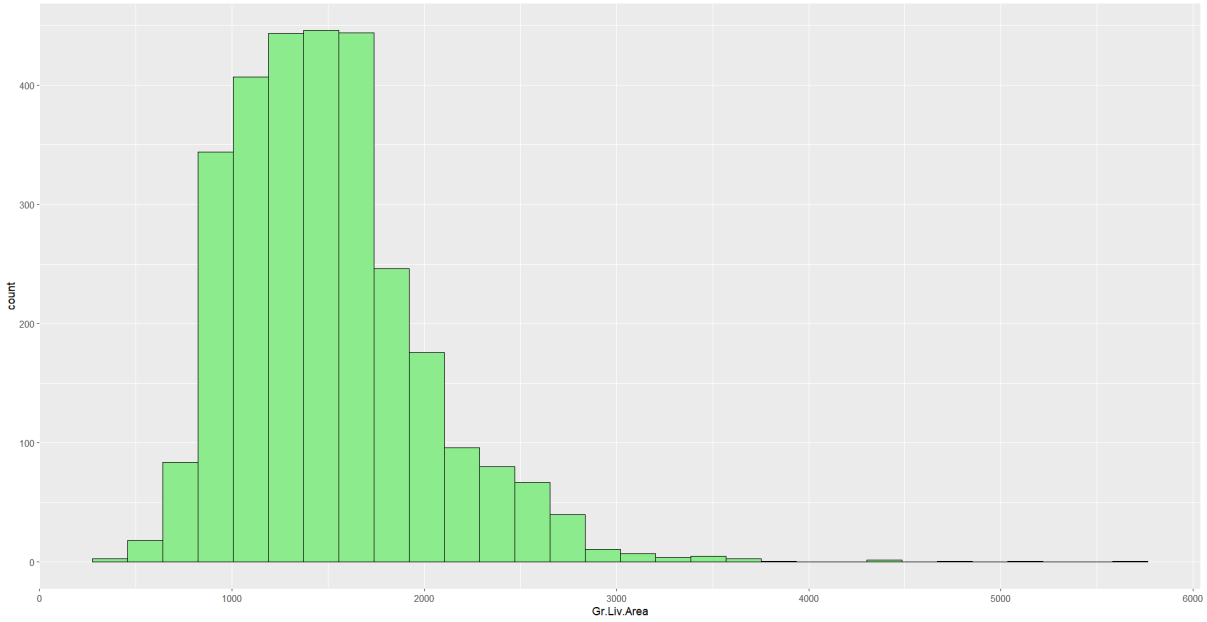
a. Histogram

`data_mean` ve `data_knn` veri setlerinden "Gr_Liv_Area" özelliği için histogramlar çizilmiştir. Histogramlar, verinin dağılımını görselleştirir ve imputation sonrası verilerin nasıl değiştiğini gözler önüne serer.

```
ggplot(data_mean, aes(x = Gr.Liv.Area)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Gr.Liv.Area Histogramı", x = "Gr.Liv.Area", y =
"Frekans")
```




```
ggplot(data_knn, aes(x = Gr.Liv.Area)) +
  geom_histogram(bins = 30, fill = "lightgreen", color = "black")
```



b. Boxplot

`data_knn` veri setindeki "Sale_Price" değişkeni için bir boxplot oluşturulmuştur. Boxplot, verinin merkezini, dağılımını ve olası uç değerleri görselleştirir.

```
ggplot(data_knn, aes(y = Sale_Price)) +
  geom_boxplot()
```

5. Sonuçlar ve Karşılaştırmalar

- **Ortalama/Mod İle Doldurma**: Sayısal verilerde ortalama, kategorik verilerde ise mod kullanarak yapılan imputation, verinin dağılımında belirgin bir değişiklik yaratmamaktadır. Ancak bu yöntem, özellikle basit ve hızlı olmasıyla dikkat çeker.

- **KNN İle Doldurma**: KNN imputation, veri noktalarına daha yakın olan komşuların bilgilerini kullanarak eksik değerleri tahmin eder. Bu yöntem, daha hassas ve veriye dayalı bir imputation sağlar. Ancak hesaplama maliyeti daha yüksektir.

- **Karar Ağaçları İle Doldurma**: Bu yöntem, daha karmaşık ve doğrusal olmayan ilişkileri modelleyerek eksik değerleri tahmin eder. KNN'ye göre daha sofistike bir yöntemdir, ancak yine de daha fazla hesaplama gücü gerektirebilir.

Görselleştirmeler (histogramlar ve boxplot'lar), imputation yöntemlerinin veri dağılımı üzerindeki etkisini net bir şekilde göstermektedir. Özellikle KNN ve karar ağacı gibi daha sofistike yöntemler, daha sağlıklı ve tutarlı sonuçlar üretmiştir.

6. Sonuçlar ve Değerlendirme

Bu çalışmada, üç farklı imputation yöntemi kullanılarak eksik veriler doldurulmuş ve sonuçlar karşılaştırılmıştır. KNN ve karar ağacı yöntemleri, daha gelişmiş ve doğru sonuçlar vermiştir. Bu yöntemler, özellikle daha karmaşık veri setlerinde daha iyi performans gösterme potansiyeline sahiptir.

Raporun Özeti:

- Eksik veriler, çeşitli imputation yöntemleri ile dolduruldu.

- Görselleştirmeler, her bir yöntem sonrası verinin dağılımını etkileyip etkilemediğini gözler önüne serdi.

- KNN ve karar ağacı, eksik verilerin doldurulmasında en iyi sonuçları verdi.

- Bu tür imputation yöntemleri, veri analizi ve modelleme aşamalarında önemli bir rol oynar ve doğruluk için kritik olabilir.