Housing Prices

Mine Çetinkaya-Rundel

Introduction

In this analysis, we build a model predicting sale prices of houses based on data on houses
that were sold in the Duke Forest neighborhood of Durham, NC around November 2020. Let's
start by loading the packages we'll use for the analysis.

```
1  library(openintro)  # for data
2  library(tidyverse)  # for data wrangling and visualization
3  library(knitr)      # for tables
4  library(broom)      # for model summary
```

We present the results of exploratory data analysis in Section and the regression model in
Section .

Exploratory data analysis

The data contains 98 houses. As part of the exploratory analysis let's visualize and summarize
the relationship between areas and prices of these houses.

Data visualization

Figure 1 shows two histograms displaying the distributions of `price` and `area` individually.

```
1  ggplot(duke_forest, aes(x = price)) +
2    geom_histogram(binwidth = 50000) +
3    labs(title = "Histogram of prices")
4
5  ggplot(duke_forest, aes(x = area)) +
6    geom_histogram(binwidth = 250) +
7    labs(title = "Histogram of areas")
```

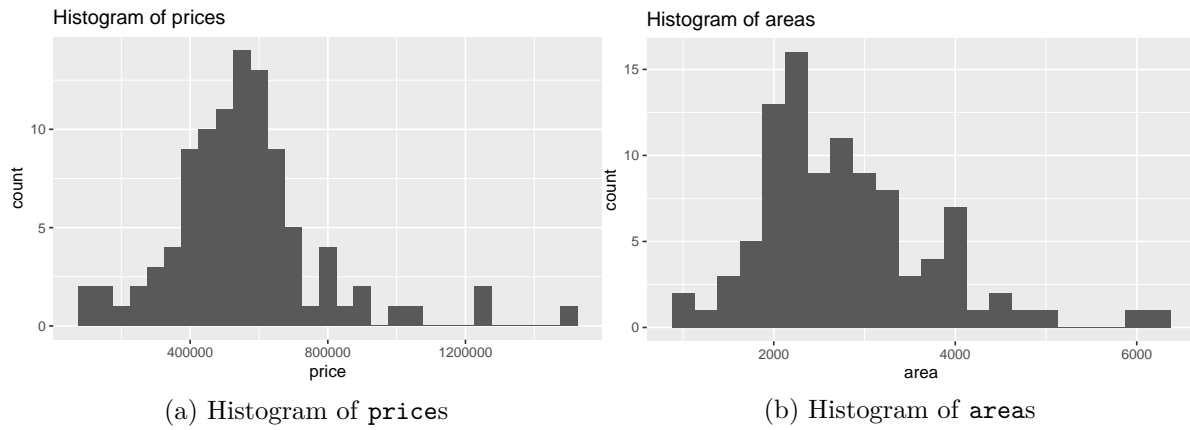(a) Histogram of `prices`   (b) Histogram of `areas`

Figure 1: Histograms of individual variables

Figure 2 displays the relationship between these two variables in a scatterplot.

```
1  ggplot(duke_forest, aes(x = area, y = price)) +
2    geom_point() +
3    labs(title = "Price and area of houses in Duke Forest")
```
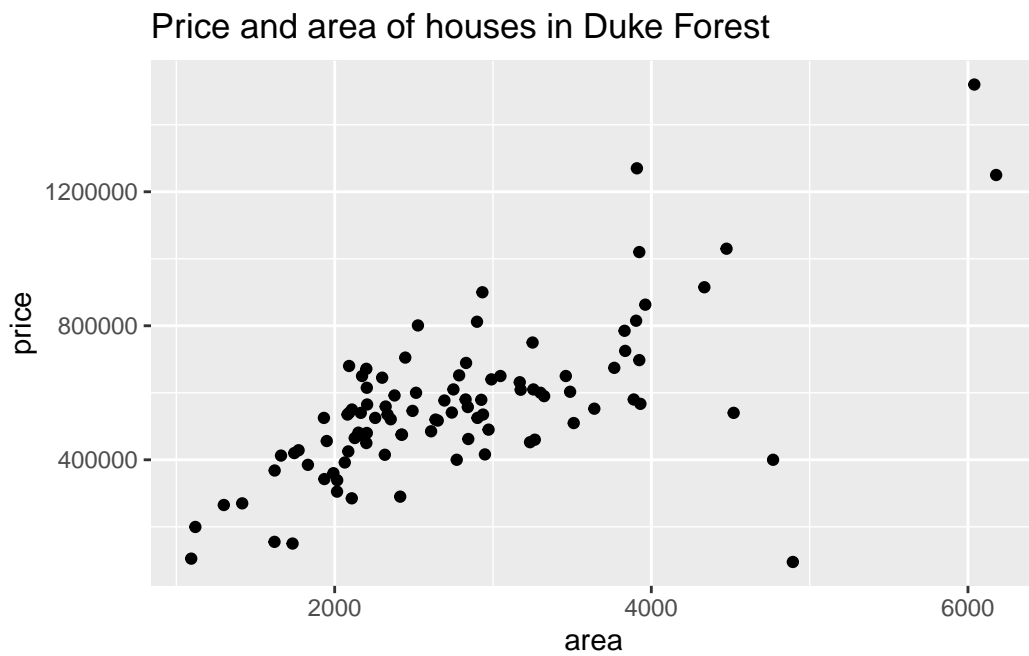

Figure 2: Scatterplot of price vs. area of houses in Duke Forest

Summary statistics

Table 1 displays basic summary statistics for these two variables.

```
1  duke_forest %>%
2    summarise(
3      `Median price` = median(price),
4      `IQR price` = IQR(price),
5      `Median area` = median(area),
6      `IQR area` = IQR(area),
7      `Correlation, r` = cor(price, area)
8    ) %>%
9    kable(digits = c(0, 0, 0, 0, 2))
```

Table 1: Summary statistics for price and area of houses in Duke Forest

Median price	IQR price	Median area	IQR area	Correlation, r
540000	193125	2623	1121	0.67

Modeling

We can fit a simple linear regression model of the form shown in **?@eq-slr**.

[ADD EQUATION HERE]

Table 2 shows the regression output for this model.

```
1  price_fit <- lm(price ~ area, data = duke_forest)
2
3  price_fit %>%
4    tidy() %>%
5    kable(digits = c(0, 0, 2, 2, 2))
```

Table 2: Linear regression model for predicting price from area

term	estimate	std.error	statistic	p.value
(Intercept)	116652	53302.46	2.19	0.03
area	159	18.17	8.78	0.00