DATA 202 Homework 6

Adham Rishmawi

2022-11-28

Problem 1

Suppose We want to predict the sale price of homes

1. Which locations have the largest prices associated with it
2. What is the sale-ability of this property even with its price
3. regression because we are determining a price which fits the qualifications of a model
4. a potential good error metric can be the difference between the price predicted to sell vs the price it actually sold as made into a (Mean Squared error) percentage!
5. I would say that the expected performance of a model as a decision-maker would be great because we can use predictors such as location, size, and demand as categorical decisions and make great predictions on
6. a inappropriate error metric would be something like a MASE which measures the accuracy of forecasts (MAE divided by the Mean Absolute Error in sample one-step naive forecast)
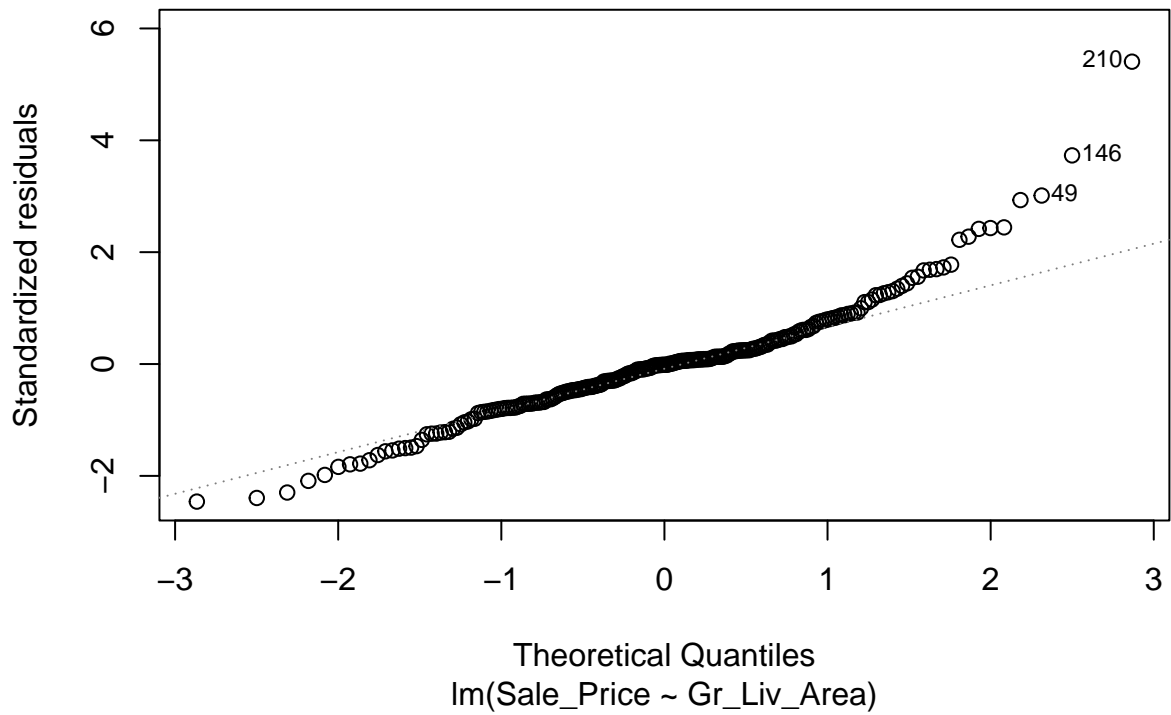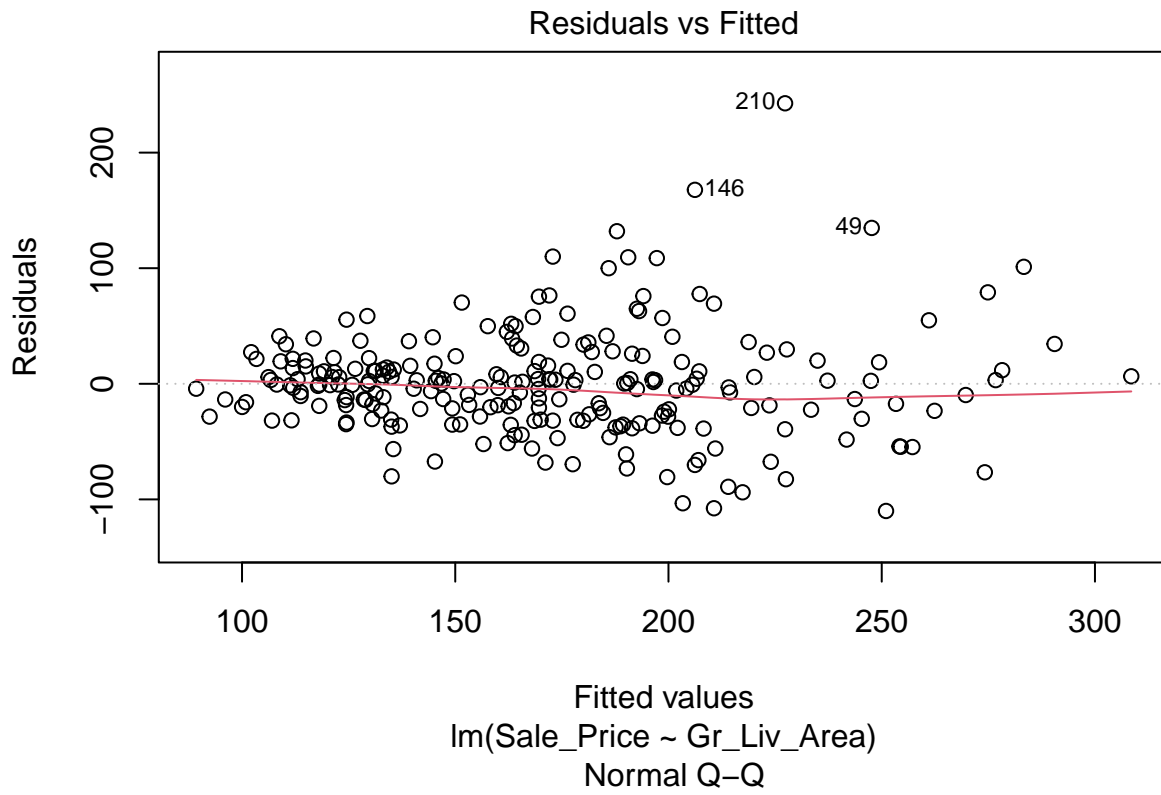
Problem 2

```
ames_home_sales <- read_builtin("ames", package = "modeldata") %>%
  mutate(Sale_Price = Sale_Price / 1000) %>%
  filter(Gr_Liv_Area < 4000, Sale_Condition == "Normal")
```

A.Hold out 10% of homes to validate the model.

```
home_split <- ames_home_sales |>
  initial_split(prop = 0.1)
home_training <- home_split|>
  training()
home_test <- home_split|>
  testing()
```

B.

```
model<- lm(Sale_Price ~ Gr_Liv_Area, data = home_training )
plot(model)
```

Residuals vs Fitted

210

146

49

Residuals

Fitted values
lm(Sale_Price ~ Gr_Liv_Area)

Normal Q–Q

210

146

49

Standardized residuals

Theoretical Quantiles
lm(Sale_Price ~ Gr_Liv_Area)

Scale–Location

Fitted values
lm(Sale_Price ~ Gr_Liv_Area)

Residuals vs Leverage

Leverage
lm(Sale_Price ~ Gr_Liv_Area)
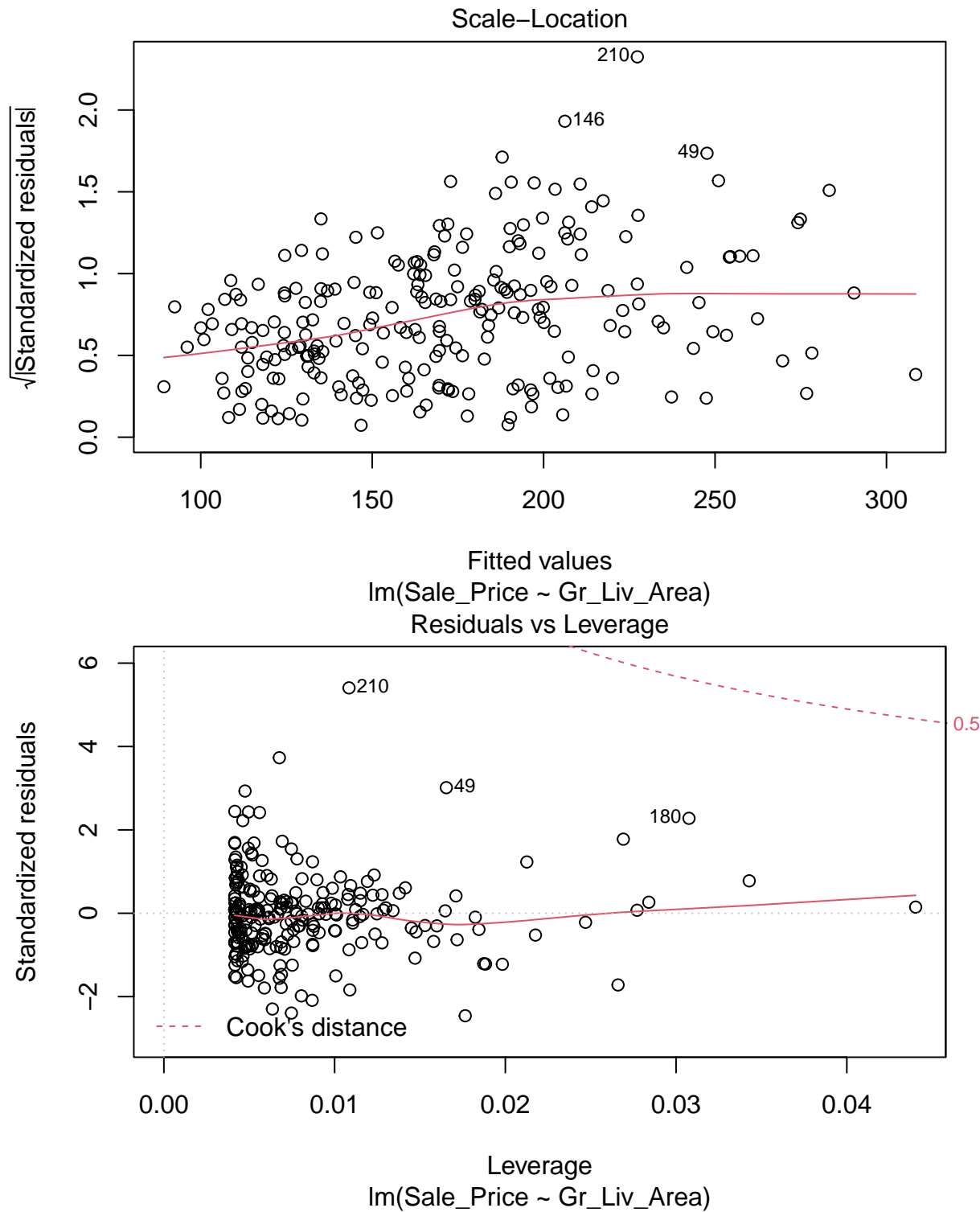
C.
MAE: the predictions are typically off by $5000-7000 MAPE: Predictions are typically off by 6-7%

D. I would expect that this decision-maker will be accurate enough to predict the prices of houses to be listed

Problem 3

```r
ames_vs_median <- ames_home_sales %>%
  mutate(sale_category = case_when(
    Sale_Price > median(Sale_Price) ~ "Above Median",
    TRUE ~ "Below Median"
  ) %>%
    as_factor() %>%
    fct_relevel("Above Median") # Make sure that "Above Median" is considered the positive class.
  )
```

for each home, we are tasked to say either "above median" or "below median" 1. Which types of houses are typically above or below the Median 2. What are the major factors to a above median property 3.This is classification because it is either above or below Median 4. a good error metric is MAPE because it utilizes percentages 5.I would say that the expect performance of this decision maker is to divide the houses by types and sizes 6.an inappropriate error metrics might be MAE because values wouldn't be useful for a categorical matters.

Problem 4

A.

```r
home_split1 <- ames_vs_median |>
  initial_split(prop = 0.1)
home_training1 <- home_split1|>
  training()
home_test1 <- home_split1|>
  testing()
```
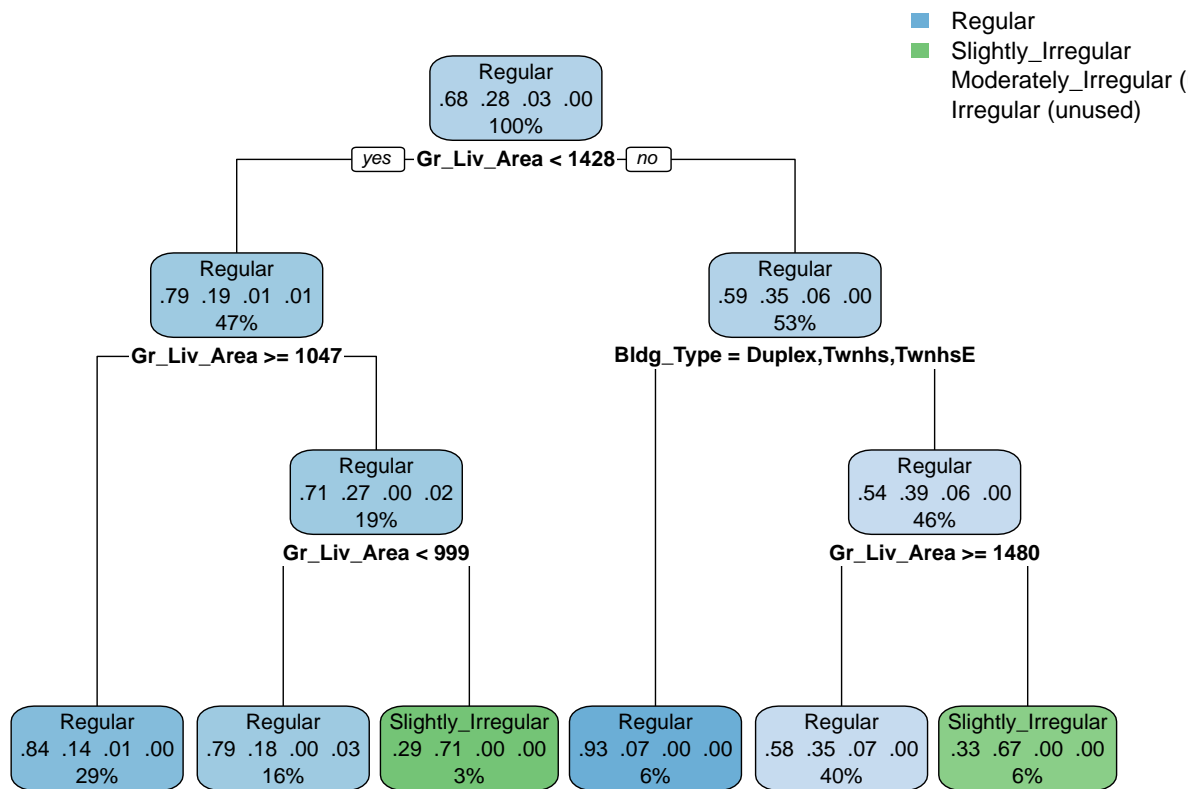
B.

```r
decision_tree_fit <- decision_tree(mode = "classification", tree_depth = 3)|>
  set_engine("rpart")|>
  fit(
    Lot_Shape~ Gr_Liv_Area + Bldg_Type,
    data = home_training1
  )
```

```r
decision_tree_fit|>
  extract_fit_engine()|>
  rpart.plot::rpart.plot(roundint = FALSE)
```

Regular ■
Slightly_Irregular ■
Moderately_Irregular (
Irregular (unused)

```
                          Regular
                      .68  .28  .03  .00
                            100%
                    yes — Gr_Liv_Area < 1428 — no

        Regular                                   Regular
    .79  .19  .01  .01                       .59  .35  .06  .00
          47%                                      53%
    Gr_Liv_Area >= 1047                 Bldg_Type = Duplex,Twnhs,TwnhsE

              Regular                                    Regular
          .71  .27  .00  .02                        .54  .39  .06  .00
                19%                                        46%
          Gr_Liv_Area < 999                       Gr_Liv_Area >= 1480
```

Regular	Regular	Slightly_Irregular	Regular	Regular	Slightly_Irregular
.84 .14 .01 .00	.79 .18 .00 .03	.29 .71 .00 .00	.93 .07 .00 .00	.58 .35 .07 .00	.33 .67 .00 .00
29%	16%	3%	6%	40%	6%

C. MAE: NA MAPE: Predictions are typically off by 6-7%

D. I would expect that this decision-maker will be accurate enough to predict the houses to be either higher or lower than the median!