Analysis

Visualization

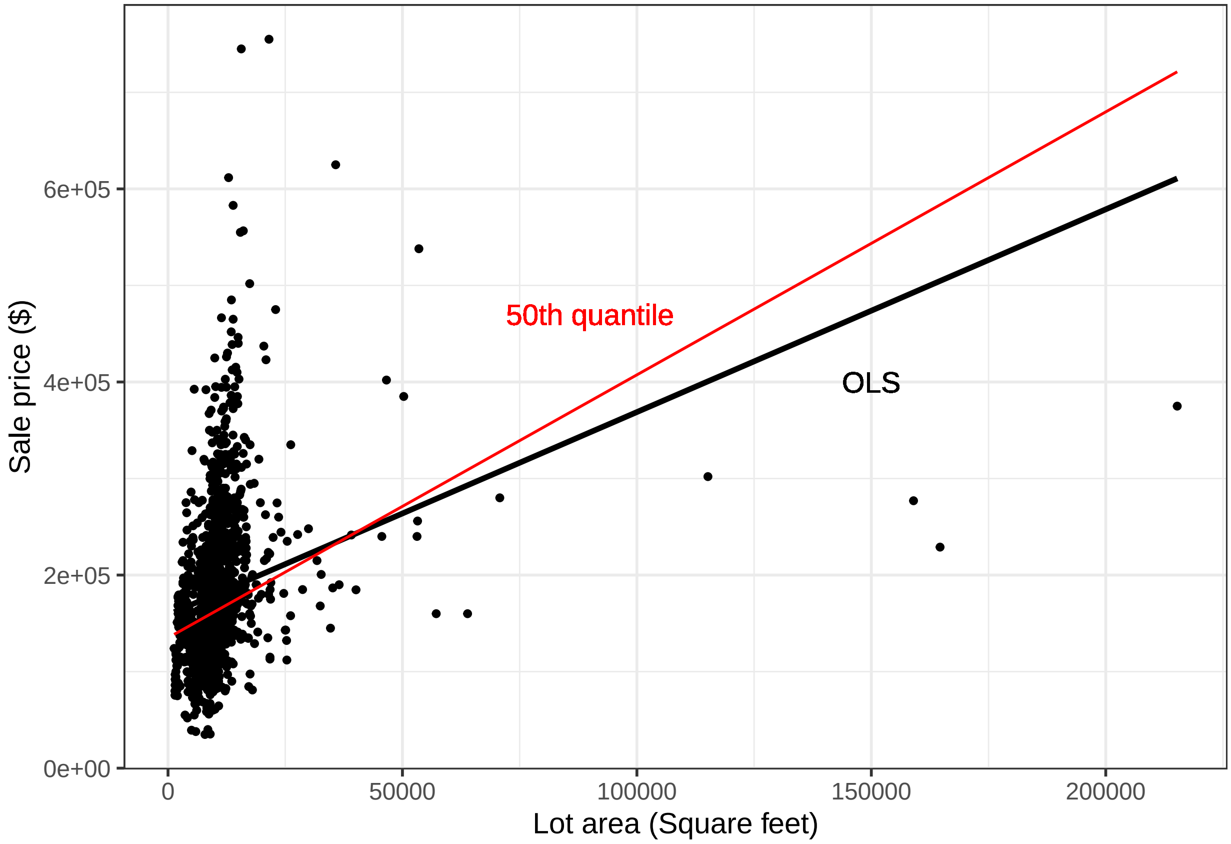
```r
df <- read.csv("TrainData.csv") |>
  na.omit() |>
  distinct()

# head(df2)
# colnames(df2)
```

Visualizing quantile regression vs OLS

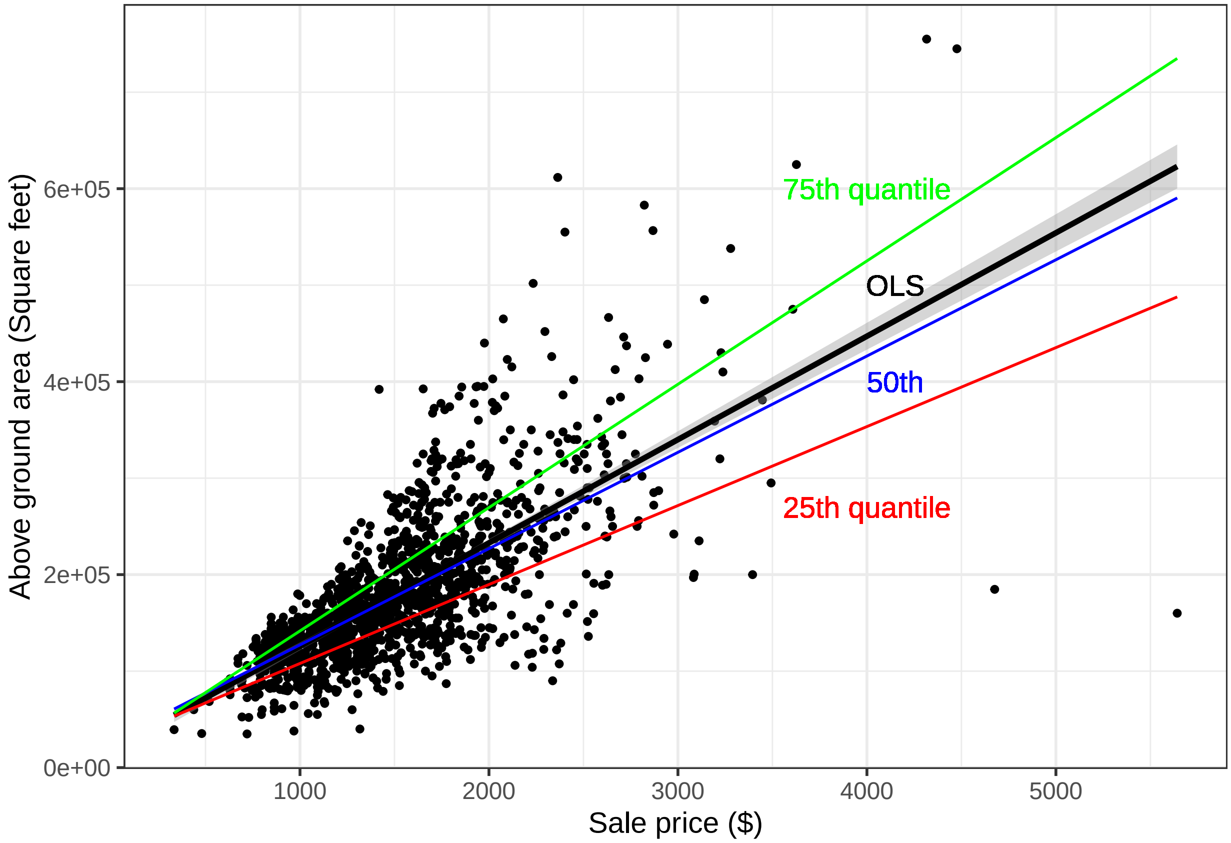
```r
df |> ggplot(aes(y = SalePrice, x = LotArea)) +
  geom_point(size = 0.9) +
  geom_smooth(method = lm, se = F, color = "black") +
  geom_text(aes(y = 400000, x = 150000, label = "OLS"), color="black") +
  geom_quantile(quantiles=0.5, color="red") +
  geom_text(aes(y = 470000, x = 90000, label = "50th quantile"), color="red") +
  ylab("Sale price ($)") +
  xlab("Lot area (Square feet)") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Smoothing formula not specified. Using: y ~ x
```



```r
df |> ggplot(aes(y = SalePrice, x = GrLivArea)) +
  geom_point(size = 0.9) +
  stat_smooth(method = lm, color = "black") +
  geom_text(aes(x = 4150, y = 500000, label = "OLS"), color="black") +
  geom_quantile(quantiles=0.25, color="red") +
  geom_text(aes(x = 4000, y = 270000, label = "25th quantile"), color="red") +
  geom_quantile(quantiles=0.5, color="blue") +
  geom_text(aes(x = 4150, y = 400000, label = "50th"), color="blue") +
  geom_quantile(quantiles=0.75, color="green") +
  geom_text(aes(x = 4000, y = 600000, label = "75th quantile"), color="green") +
  xlab("Sale price ($)") +
  ylab("Above ground area (Square feet)") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
```



Model creation

QR model

```r
qreg_model50 = rq(data=df, SalePrice ~ GrLivArea + LotArea + TotRmsAbvGrd + as.factor(LotShape) + as.fac
```

```
## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique
```

```
summary(qreg_model50)
```

```
## Warning in summary.rq(qreg_model50): 3 non-positive fis
```

```
##
## Call: rq(formula = SalePrice ~ GrLivArea + LotArea + TotRmsAbvGrd +
##     as.factor(LotShape) + as.factor(Foundation), tau = 0.5, data = df)
##
## tau: [1] 0.5
##
## Coefficients:
##                             Value        Std. Error   t value      Pr(>|t|)
## (Intercept)                  36326.81296   3853.84854      9.42611      0.00000
## GrLivArea                       96.66934      4.02708     24.00481      0.00000
## LotArea                          0.99940      0.32815      3.04561      0.00236
## TotRmsAbvGrd                 -6476.18114   1080.95132     -5.99119      0.00000
## as.factor(LotShape)IR2       -5084.13375   7841.20685     -0.64839      0.51684
## as.factor(LotShape)IR3      -21074.80675   7616.42154     -2.76702      0.00573
## as.factor(LotShape)Reg      -11065.07360   2020.92512     -5.47525      0.00000
## as.factor(Foundation)CBlock  21252.40678   1709.40460     12.43264      0.00000
## as.factor(Foundation)PConc   53311.16094   2618.05941     20.36285      0.00000
## as.factor(Foundation)Slab   -16867.20619   5378.30454     -3.13616      0.00175
## as.factor(Foundation)Stone   14561.54748  13561.64146      1.07373      0.28312
## as.factor(Foundation)Wood    -2008.81877   9022.14216     -0.22265      0.82384
```

OLS model

```
ols = lm(data=df, SalePrice ~ GrLivArea + LotArea + TotRmsAbvGrd + as.factor(LotShape) + as.factor(Foun
summary(ols)
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea + LotArea + TotRmsAbvGrd +
##     as.factor(LotShape) + as.factor(Foundation), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -422488  -26194    -805   20461  326538
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.005e+04  7.267e+03   2.759  0.00587 **
## GrLivArea                    9.893e+01  4.538e+00  21.801  < 2e-16 ***
## LotArea                      9.173e-01  1.425e-01   6.438 1.64e-10 ***
## TotRmsAbvGrd                -4.313e+03  1.396e+03  -3.089  0.00205 **
## as.factor(LotShape)IR2      -2.009e+03  8.113e+03  -0.248  0.80446
## as.factor(LotShape)IR3      -6.936e+04  1.603e+04  -4.328 1.61e-05 ***
## as.factor(LotShape)Reg      -1.342e+04  2.809e+03  -4.777 1.96e-06 ***
## as.factor(Foundation)CBlock  2.094e+04  4.497e+03   4.656 3.52e-06 ***
## as.factor(Foundation)PConc   6.679e+04  4.541e+03  14.708  < 2e-16 ***
## as.factor(Foundation)Slab   -1.426e+04  1.067e+04  -1.336  0.18170
## as.factor(Foundation)Stone  -3.396e+03  2.021e+04  -0.168  0.86658
## as.factor(Foundation)Wood   -5.553e+02  2.842e+04  -0.020  0.98441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48410 on 1448 degrees of freedom
## Multiple R-squared:  0.6315, Adjusted R-squared:  0.6287
## F-statistic: 225.6 on 11 and 1448 DF,  p-value: < 2.2e-16
```

Model evaluation

Mean absolute error

Root mean squared error

Variance of error

Min/max error