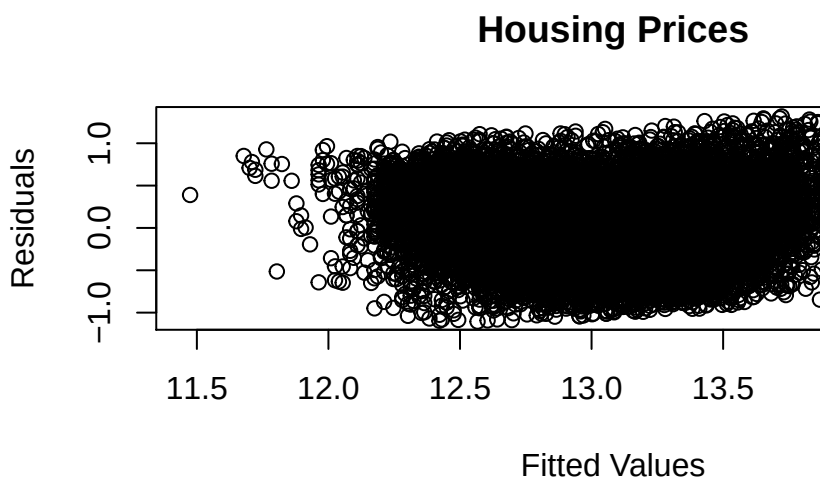
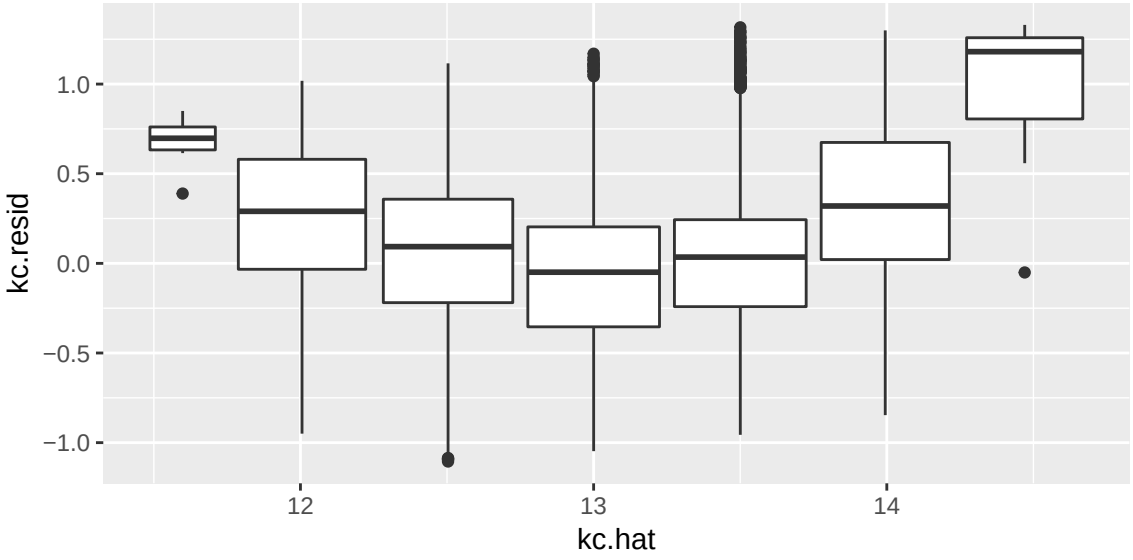
Math 158 Final Project: Part 2

Nick George and Spencer Louie

We're interested in testing whether or not theire is a positive relation between housing price and the square



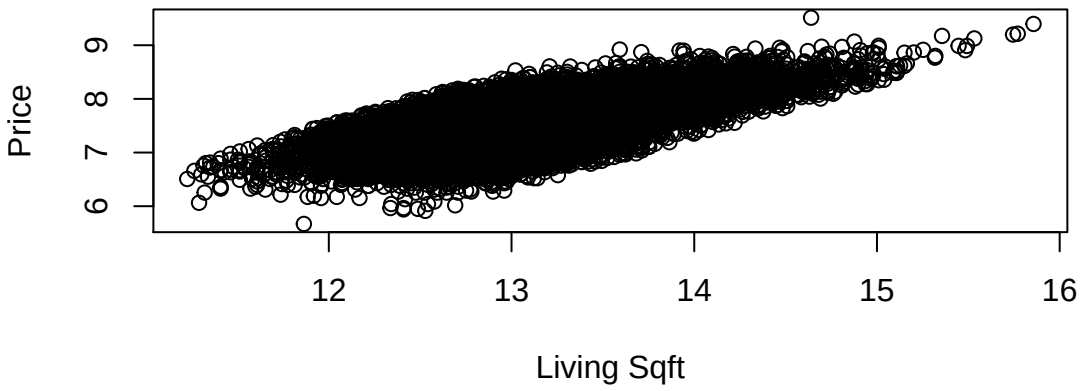
footage of living area. So $H_0 : \beta_1 \leq 0$ and $H_a : \beta_1 > 0$.



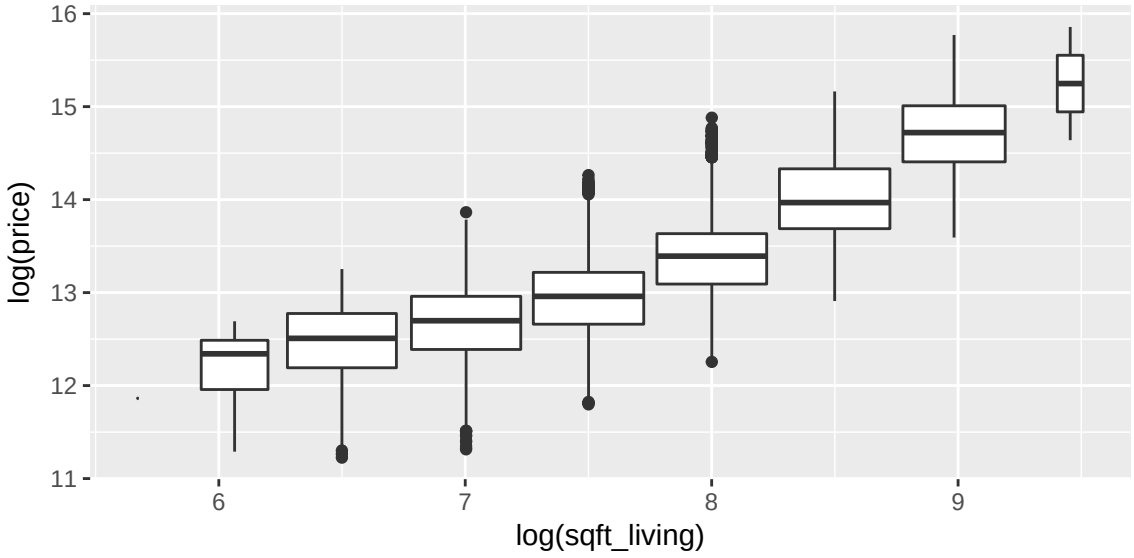

```
##              term estimate   std.error statistic p.value
## 1     (Intercept) 6.729916 0.047061982  143.0011       0
## 2 log(sqft_living) 0.836771 0.006223257  134.4587       0
```

By taking log-log transformations of our data we were able to get a residual plot that seems to satisfy some of our necessary assumptions. The residuals appear to be fairly equally distributed positively and negatively and the variance seems to be constant across the fitted values.

The p-value is roughly 0 and the t-stastic is quite large at $134/2 = 67$ so we are able to reject the null hypothesis and say that there is some positive relation between housing price and living square footage. This implies that a doubling in square footage would be associated with a $2^{\cdot}836 = 1.786$ multiplicate change in the median of



price.




```
##        fit      lwr     upr
## 1 12.66269 11.90087 13.4245
```

This is a prediction interval for 1200 square feet of living space. So 95% of the log price values are between 11.9 and 13.42.

```
##        fit      lwr      upr
## 1 12.66269 12.65505 12.67033
```

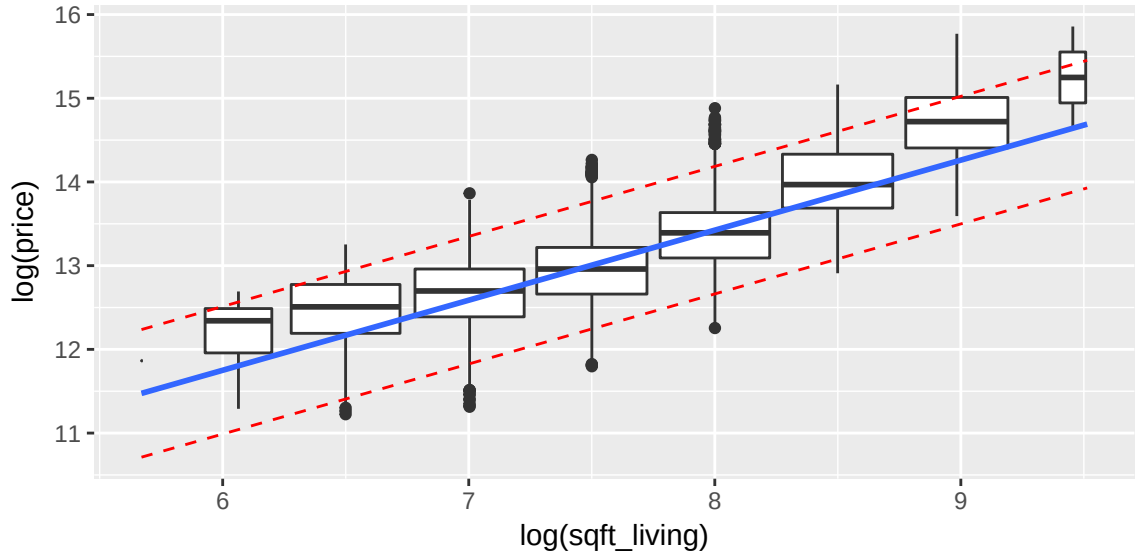
This is a confidence or mean interval for 1200 square feet of living space. We are 95% confident that the mean log price value for 1200 square feet of living space is between 12.65 and 12.67.

```
##
## Call:
## lm(formula = log(price) ~ log(sqft_living), data = kc_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10511 -0.29300  0.01262  0.25701  1.33011
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.729916   0.047062   143.0   <2e-16 ***
## log(sqft_living) 0.836771   0.006223   134.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3886 on 21611 degrees of freedom
## Multiple R-squared:  0.4555, Adjusted R-squared:  0.4555
## F-statistic: 1.808e+04 on 1 and 21611 DF,  p-value: < 2.2e-16
```

The R squared of the model is .4555 which means that about 45% of the variance in log price is explained by log living square footage.

```
## Warning in predict.lm(kc.lm, interval = "predict"): predictions on current data refer to _future_ res
```

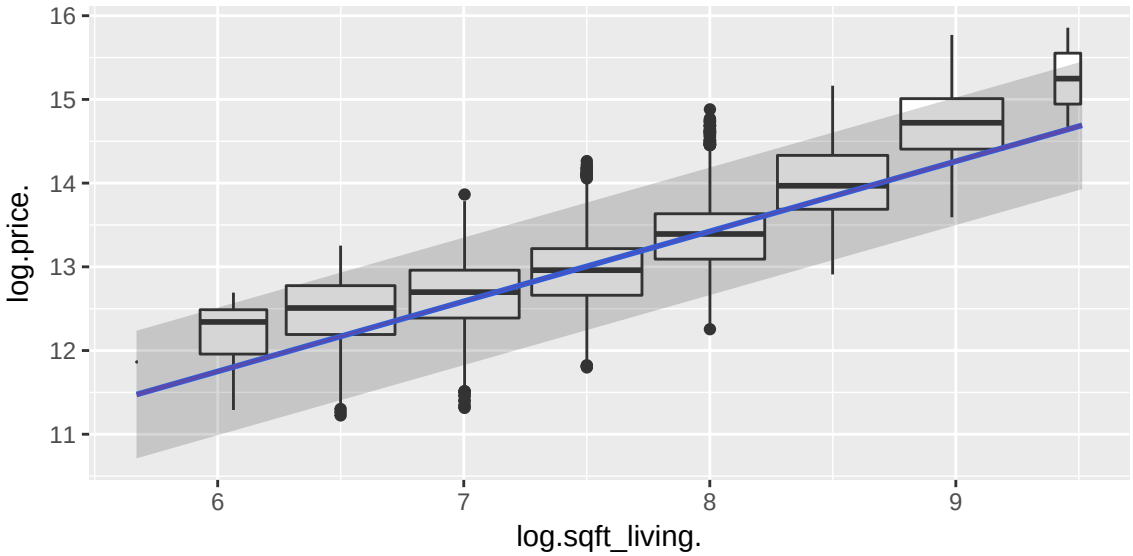



```
library(broom)
crit_val <- qt(.975, glance(kc.lm)$df.resid)
kc_gl <- broom::glance(kc.lm)
kc_sig <- dplyr::pull(kc_gl, sigma)
kc_pred <- broom::augment(kc.lm) %>% mutate(.se.pred = sqrt(kc_sig^2 + .se.fit^2)) %>% mutate(lower_PI =
kc_pred %>% head()
```

```
##   log.price. log.sqft_living.  .fitted     .se.fit      .resid
## 1   12.30998         7.073270 12.64862 0.003975310 -0.33864068
## 2   13.19561         7.851661 13.29996 0.003241177 -0.10434431
## 3   12.10071         6.646391 12.29142 0.006215684 -0.19071054
## 4   13.31133         7.580700 13.07323 0.002650366  0.23810398
## 5   13.14217         7.426549 12.94424 0.002753574  0.19792932
## 6   14.01845         8.597851 13.92435 0.007034601  0.09410321
##           .hat    .sigma      .cooksd .std.resid  .se.pred lower_PI
## 1 1.046233e-04 0.3886504 3.972411e-05 -0.8713750 0.3886686 11.88680
## 2 6.954918e-05 0.3886566 2.506955e-06 -0.2684894 0.3886618 12.53815
## 3 2.557787e-04 0.3886551 3.081006e-05 -0.4907649 0.3886980 11.52955
## 4 4.650486e-05 0.3886539 8.728279e-06  0.6126607 0.3886573 12.31143
## 5 5.019727e-05 0.3886549 6.510293e-06  0.5092890 0.3886580 12.18244
```

```
## 6 3.276163e-04 0.3886567 9.609827e-06  0.2421692 0.3887119 13.16244
##   upper_PI lower_CI upper_CI
## 1 13.41044 12.64083 12.65641
## 2 14.06176 13.29361 13.30631
## 3 13.05330 12.27924 12.30361
## 4 13.83502 13.06803 13.07842
## 5 13.70604 12.93884 12.94963
## 6 14.68625 13.91056 13.93814
```

```
ggplot(kc_pred, aes(x = log.sqft_living., y= log.price.)) + geom_boxplot(aes(group=cut_width(log.sqft_li
```




```
num_int <- 3
crit_Bonf <- qt((1-.975)/num_int, glance(kc.lm)$df.resid)
crit_WH <- sqrt(2*qf(.95, num_int, glance(kc.lm)$df.resid))
```

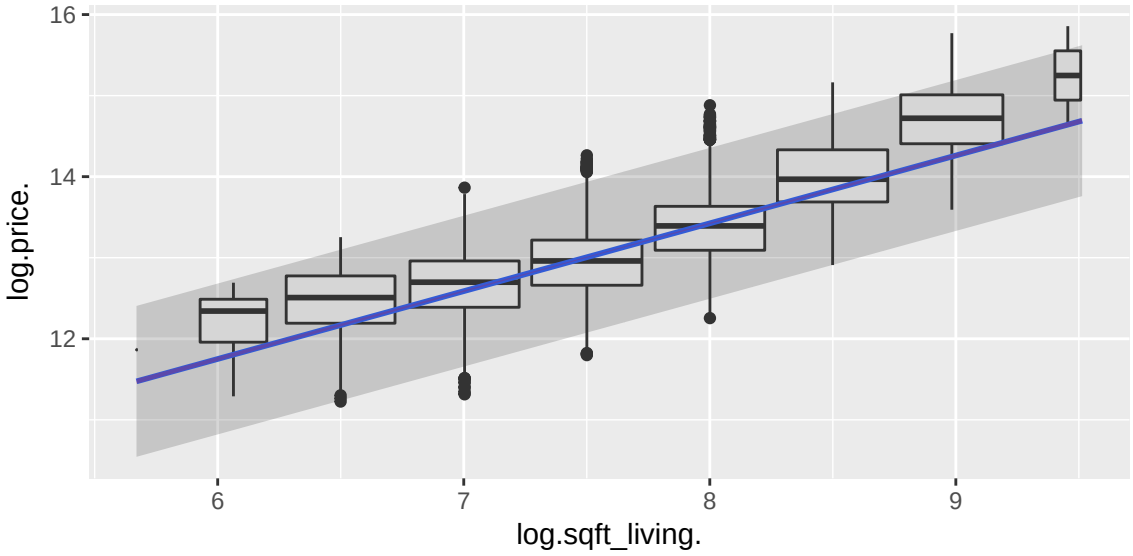
Bonf - Model

```
kc_pred_Bonf <- broom::augment(kc.lm) %>% mutate(.se.pred = sqrt(kc_sig^2 + .se.fit^2)) %>% mutate(lowe
kc_pred_Bonf %>% head()
```

```
##   log.price. log.sqft_living.  .fitted     .se.fit      .resid
## 1   12.30998         7.073270 12.64862 0.003975310 -0.33864068
## 2   13.19561         7.851661 13.29996 0.003241177 -0.10434431
## 3   12.10071         6.646391 12.29142 0.006215684 -0.19071054
## 4   13.31133         7.580700 13.07323 0.002650366  0.23810398
## 5   13.14217         7.426549 12.94424 0.002753574  0.19792932
## 6   14.01845         8.597851 13.92435 0.007034601  0.09410321
##           .hat    .sigma      .cooksd .std.resid  .se.pred lower_PI
## 1 1.046233e-04 0.3886504 3.972411e-05 -0.8713750 0.3886686 13.57916
## 2 6.954918e-05 0.3886566 2.506955e-06 -0.2684894 0.3886618 14.23048
## 3 2.557787e-04 0.3886551 3.081006e-05 -0.4907649 0.3886980 13.22203
## 4 4.650486e-05 0.3886539 8.728279e-06  0.6126607 0.3886573 14.00374
## 5 5.019727e-05 0.3886549 6.510293e-06  0.5092890 0.3886580 13.87475
## 6 3.276163e-04 0.3886567 9.609827e-06  0.2421692 0.3887119 14.85499
##   upper_PI lower_CI upper_CI
## 1 11.71809 12.65814 12.63911
## 2 12.36944 13.30772 13.29220
```

```
## 3 11.36082 12.30630 12.27654
## 4 12.14272 13.07957 13.06688
## 5 12.01372 12.95083 12.93764
## 6 12.99371 13.94119 13.90751
```

```
ggplot(kc_pred_Bonf, aes(x = log.sqft_living., y= log.price.)) + geom_boxplot(aes(group=cut_width(log.s
```



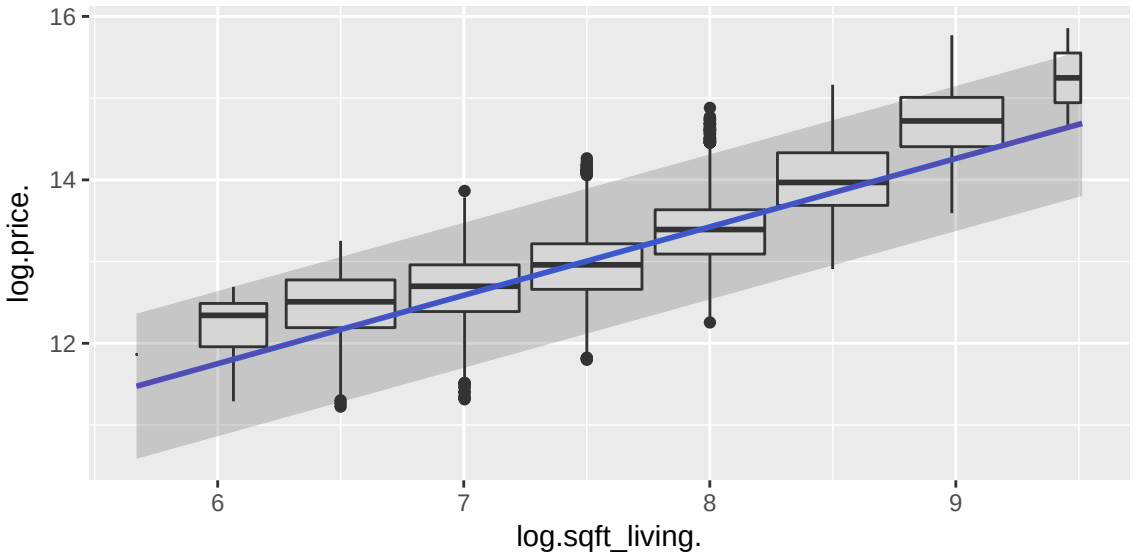
WH -

Model

```
kc_pred_WH <- broom::augment(kc.lm) %>% mutate(.se.pred = sqrt(kc_sig^2 + .se.fit^2)) %>% mutate(lower_
kc_pred_WH %>% head()
```

```
##   log.price. log.sqft_living.  .fitted     .se.fit      .resid
## 1   12.30998         7.073270 12.64862 0.003975310 -0.33864068
## 2   13.19561         7.851661 13.29996 0.003241177 -0.10434431
## 3   12.10071         6.646391 12.29142 0.006215684 -0.19071054
## 4   13.31133         7.580700 13.07323 0.002650366  0.23810398
## 5   13.14217         7.426549 12.94424 0.002753574  0.19792932
## 6   14.01845         8.597851 13.92435 0.007034601  0.09410321
##           .hat    .sigma      .cooksd .std.resid  .se.pred lower_PI
## 1 1.046233e-04 0.3886504 3.972411e-05 -0.8713750 0.3886686 11.76142
## 2 6.954918e-05 0.3886566 2.506955e-06 -0.2684894 0.3886618 12.41277
## 3 2.557787e-04 0.3886551 3.081006e-05 -0.4907649 0.3886980 11.40415
## 4 4.650486e-05 0.3886539 8.728279e-06  0.6126607 0.3886573 12.18604
## 5 5.019727e-05 0.3886549 6.510293e-06  0.5092890 0.3886580 12.05705
## 6 3.276163e-04 0.3886567 9.609827e-06  0.2421692 0.3887119 13.03704
##   upper_PI lower_CI upper_CI
## 1 13.53583 12.63955 12.65770
## 2 14.18715 13.29256 13.30736
## 3 13.17870 12.27723 12.30561
## 4 13.96041 13.06718 13.07928
## 5 13.83142 12.93795 12.95052
## 6 14.81165 13.90829 13.94041
```

```
ggplot(kc_pred_WH, aes(x = log.sqft_living., y= log.price.)) + geom_boxplot(aes(group=cut_width(log.sqf
```


16
14
12
log.price.
6
7
8
9
log.sqft_living.