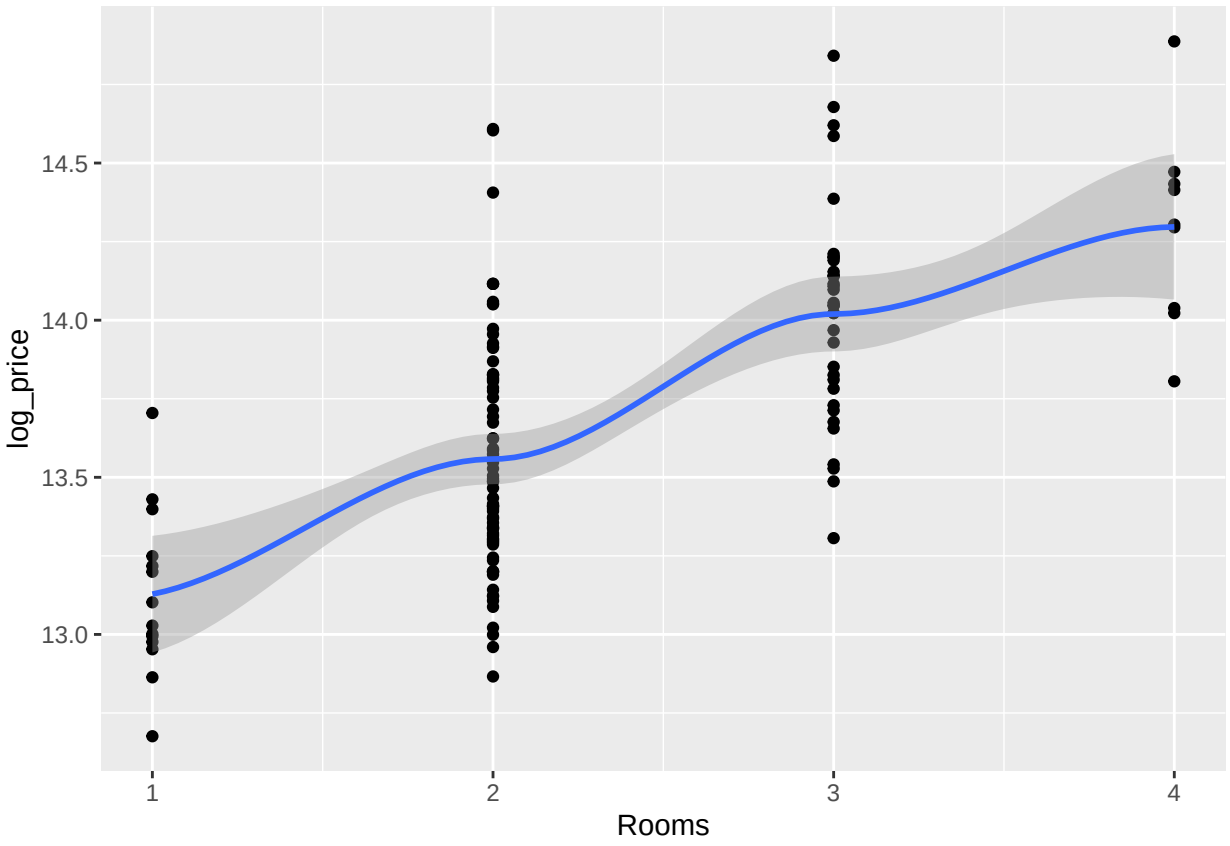
GAM Regression - Melbourne Housing

Sivaram Ainkaran

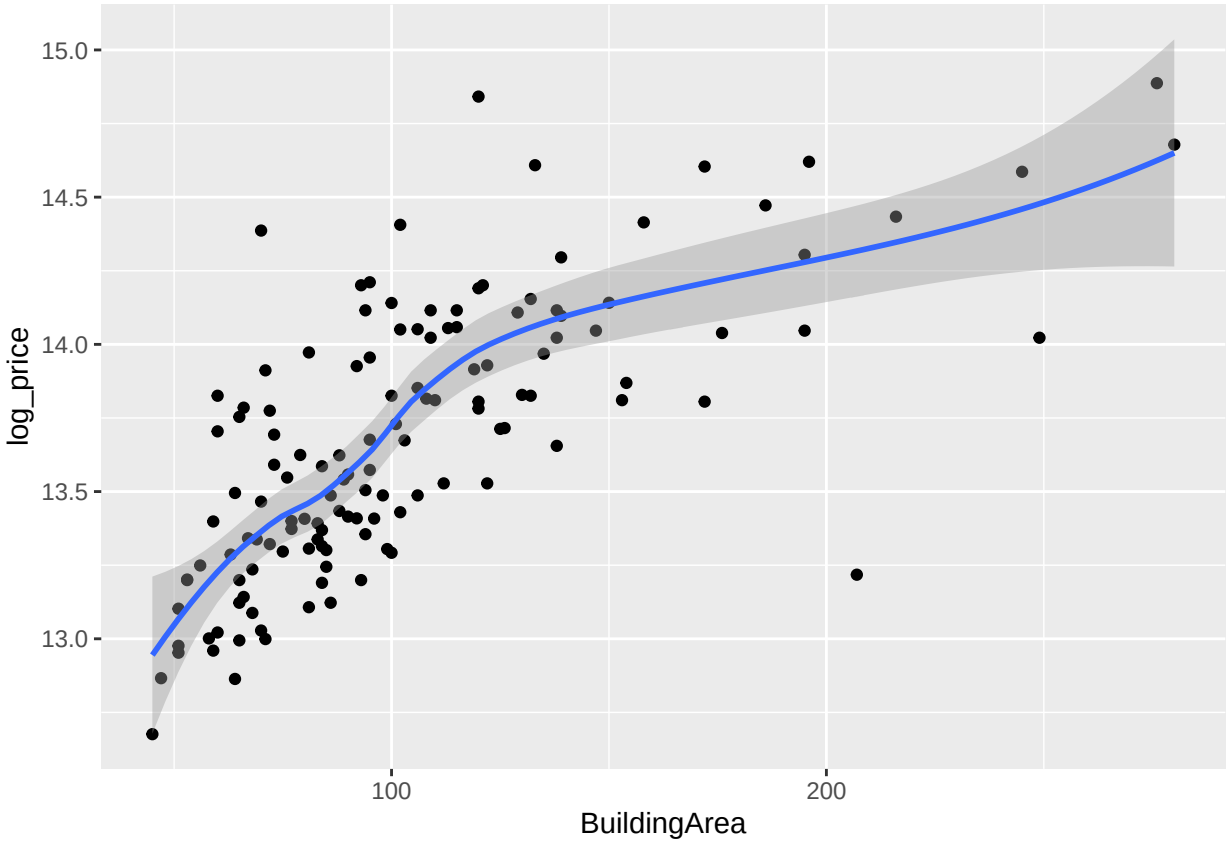
19/10/2021

Question 1

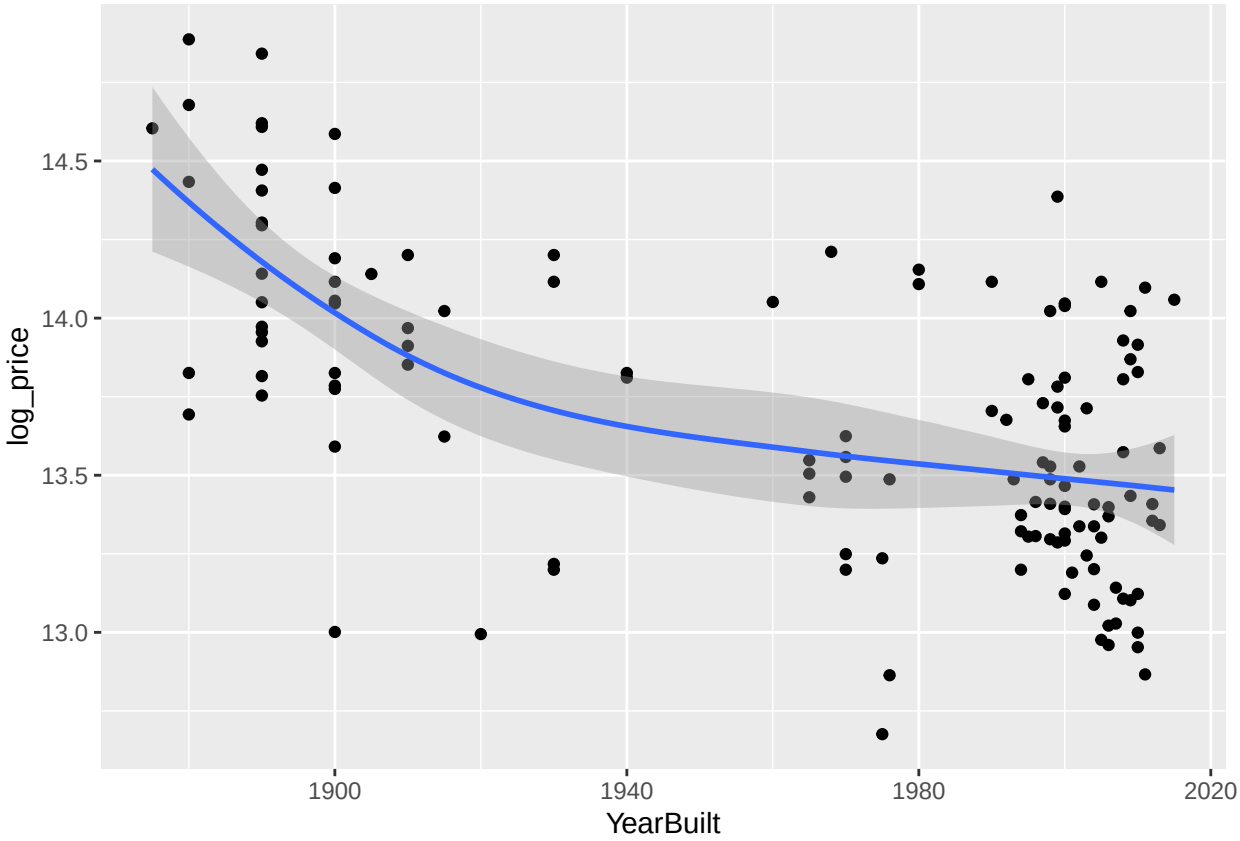
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

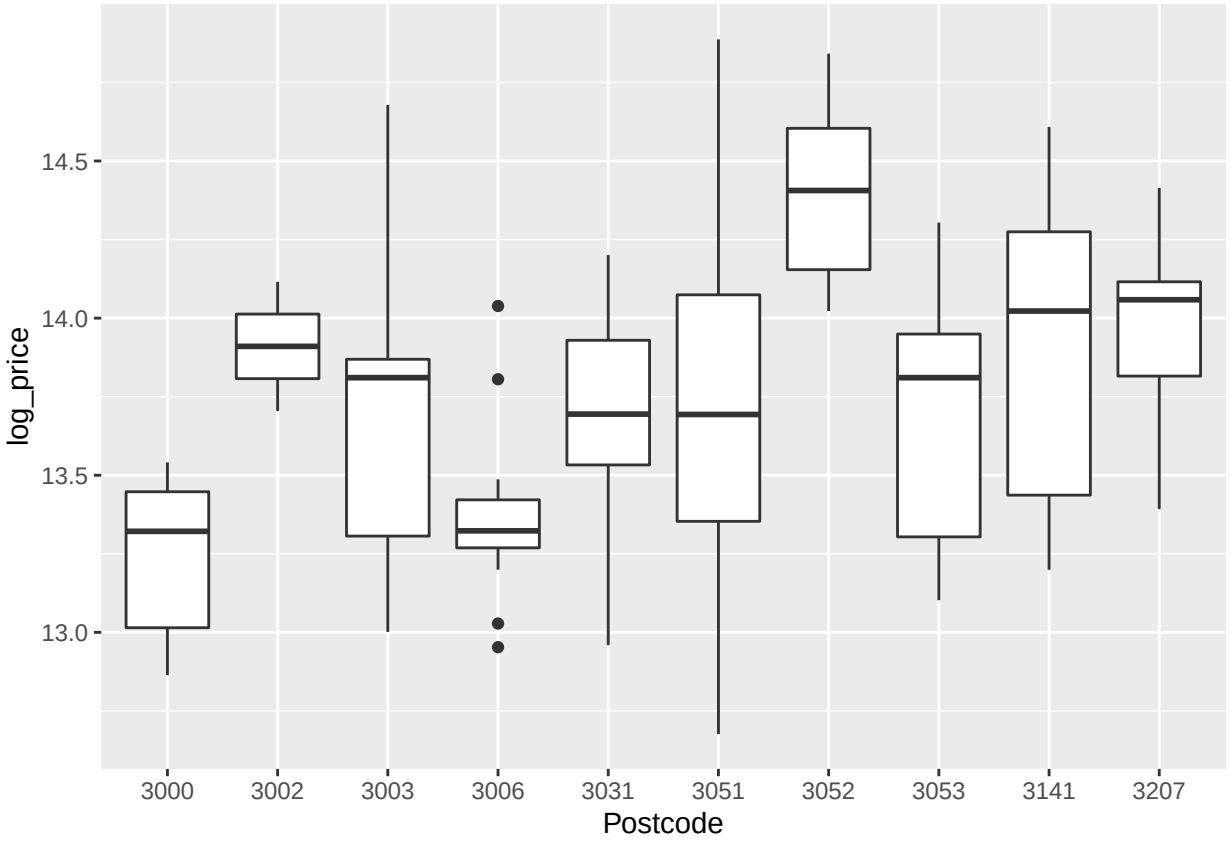



```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


log_price
14.5
14.0
13.5
13.0
1900
1940
1980
2020
YearBuilt



From these plots, we can see that "Rooms" and the response variable have a rather linear relationship with means of log_price travelling through the same line. We can see that "BuildingArea" does not have a linear relationship with the logarithm of "Price" as it folows more of a logarithmic curve. "YearBuilt" also does not have a strictly linear curve, following a more negative logarithmic curve, with alot of points scattered far from the curve. "Postcode" does not seem to have any linear effect on "log_price" with different means and variances for each postcode.

Question 2

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log_price ~ Rooms + poly(BuildingArea, 2, raw = TRUE) + YearBuilt +
##     Postcode
##
## Parametric coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.849e+01  8.604e-01  21.492  < 2e-16 ***
## Rooms                              1.575e-01  3.593e-02   4.384 2.54e-05 ***
## poly(BuildingArea, 2, raw = TRUE)1  1.120e-02  1.946e-03   5.753 7.03e-08 ***
## poly(BuildingArea, 2, raw = TRUE)2 -2.582e-05  6.122e-06  -4.217 4.87e-05 ***
## YearBuilt                          -3.131e-03  4.275e-04  -7.325 3.21e-11 ***
## Postcode3002                        6.755e-01  1.544e-01   4.374 2.65e-05 ***
## Postcode3003                        2.093e-02  8.056e-02   0.260  0.79546
## Postcode3006                       -1.211e-02  7.924e-02  -0.153  0.87875
```

```
## Postcode3031                          6.124e-02  6.886e-02   0.889  0.37563
## Postcode3051                          2.007e-01  6.638e-02   3.024  0.00307 **
## Postcode3052                          4.571e-01  9.352e-02   4.888 3.24e-06 ***
## Postcode3053                          1.736e-01  8.256e-02   2.103  0.03762 *
## Postcode3141                          2.768e-01  9.581e-02   2.890  0.00459 **
## Postcode3207                          2.941e-01  8.911e-02   3.300  0.00128 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.808   Deviance explained = 82.7%
## GCV = 0.046779  Scale est. = 0.041818  n = 132
```

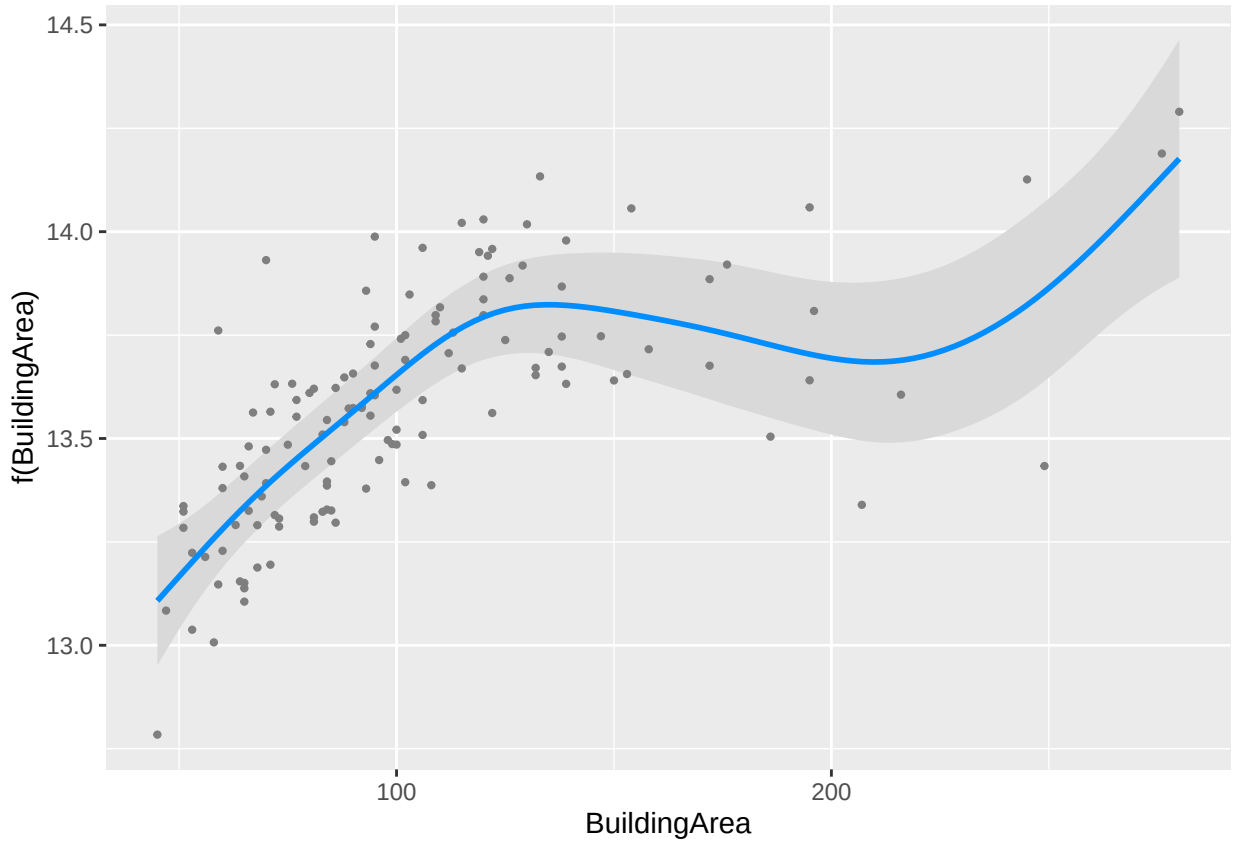
We can see that in this case, "log_price" has a negative quadratic relationship with "BuildingArea" since the coefficient for the "x^2" term is -2.582*10^-5 and the coefficient for the "x" term is 0.01120.

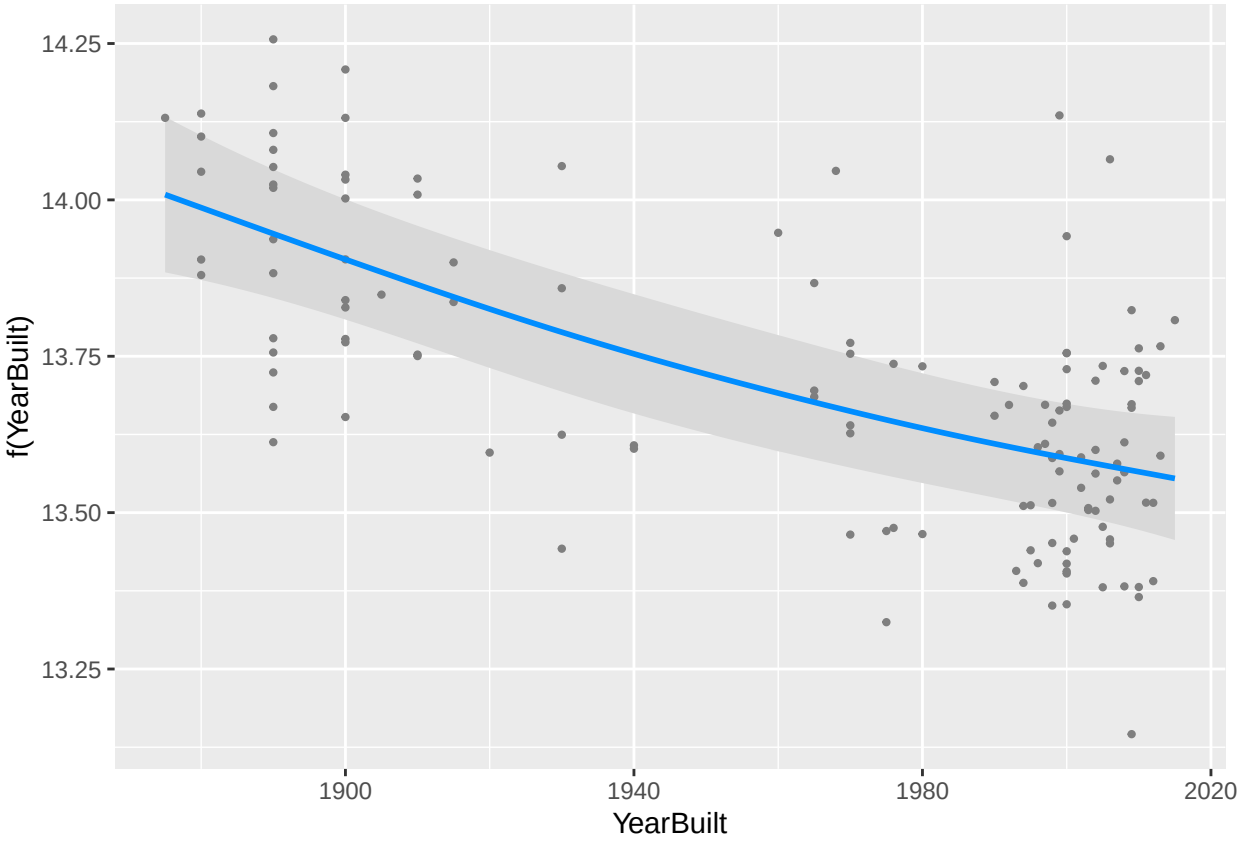
Question 3

```
## poly(BuildingArea, 2, raw = TRUE)1
##                           216.8185
```

By taking the derivative of this quadratic with the terms above and solving it by setting this equal to 0, we get a "BuildingArea" of 216.82 square metres yielding the highest price.

Question 4





We can see that "BuildingArea" follows a cubic curvature with an upward trend up to around 125 square meters where the price of the building stabilizes and begins to drop, followed by an increase in price yet again, at around 210 square meters. "YearBuilt" on the other hand follows an almost linear curve downwards, with prices decreasing as the year it was built increases. This could be due to the increase in material availability, better tools and better infrastructure as a whole we have to build houses today.

Question 5

```
## Analysis of Deviance Table
##
## Model 1: log_price ~ Rooms + s(BuildingArea) + s(YearBuilt) + Postcode
## Model 2: log_price ~ Rooms + s(BuildingArea) + YearBuilt + Postcode
##   Resid. Df Resid. Dev       Df  Deviance      F Pr(>F)
## 1    113.12     3.8276
## 2    113.88     3.8725 -0.76692 -0.044904 1.7524 0.1854
```

Since the p-value obtained from this F-test (0.1854) is much greater than 0.05, we do not reject the null hypothesis and we can say that "YearBuilt" is not significantly non-linear.

Question 6

```
##   house_pricediff
## 1          209266
```

Here we use "fit_add2" with the linear term for "YearBuilt" since it is not significantly non-linear. We can see that the predicted difference in price for these 2 houses is $209,266.