Statistical Methods and Data Analysis Coursework

MAP501

Preamble

```
library("tidyverse")
library("here")
library("janitor")
library("lindia")
library("rio")
library("magrittr")
library("pROC")
library("car")
library("nnet")
library("caret")
library("lme4")
library("AmesHousing")

Ames <- make_ames()
theme_set(theme_classic())
```

1. Data Preparation

1a.

Import the soccer.csv dataset as “footballer_data”. (2 points)

```
footballer_data <- read_csv(here("data", "soccer.csv"))
```

1b.

Ensure all character variables are treated as factors and where variable names have a space, rename the variables without these. (3 points)

```
footballer_data <- footballer_data %>%
  clean_names() %>%
  mutate_at(vars(full_name, position, current_club, nationality), list(factor))
```

1c.

Remove the columns birthday and birthday_GMT. (2 points)

```
footballer_data <- footballer_data %>%
  select(-birthday, -birthday_gmt)
```

1d.

Remove the cases with age<=15 and age>40. (2 points)

```
footballer_data <- footballer_data %>%
  filter(age > 15 & age <= 40)
```

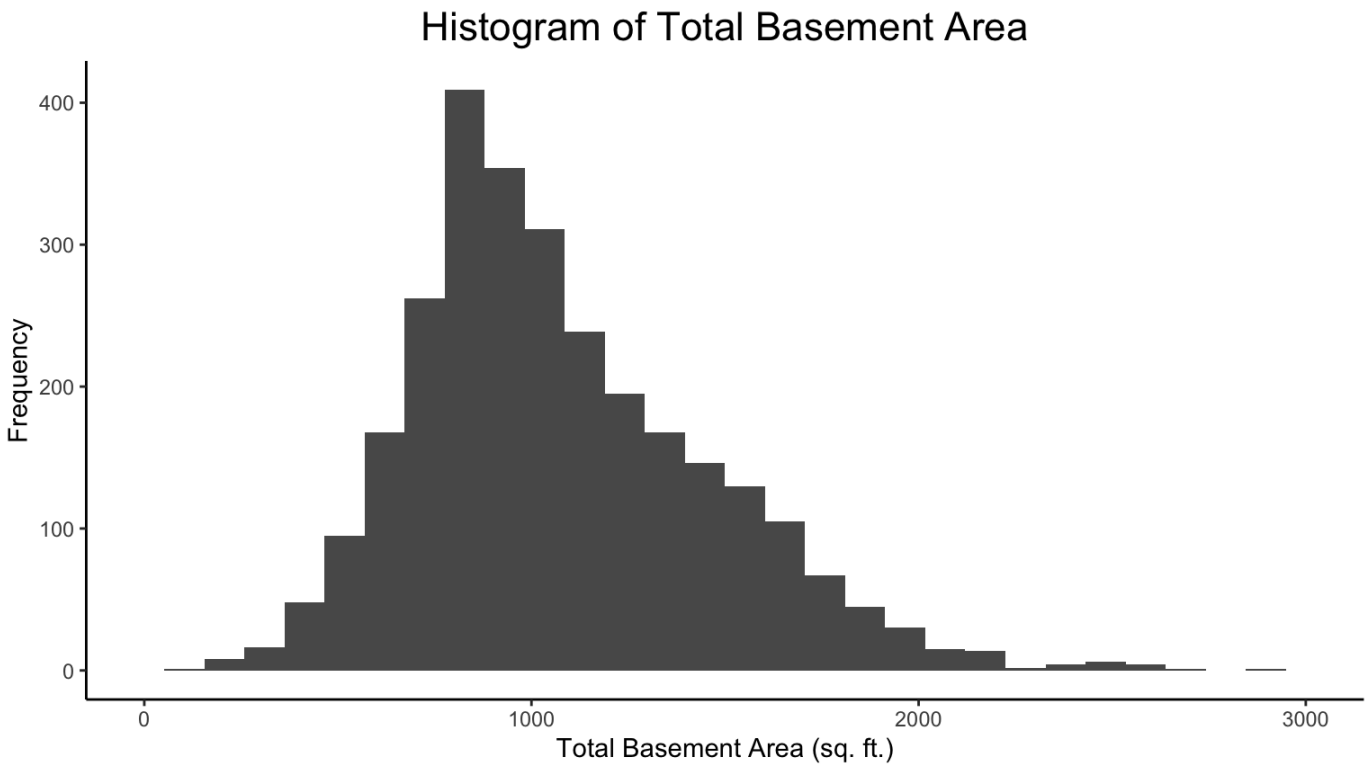
2. Linear Regression

In this problem, you are going to investigate the response variable Total_Bsmt_SF in “Ames” dataset through linear regression.

2a.

By adjusting x axis range and number of bars, create a useful histogram of Total_Bsmt_SF on the full dataset. Ensure that plot titles and axis labels are clear. (4 points)

```
Ames %>%
  ggplot(mapping = aes(x = Total_Bsmt_SF)) +
  geom_histogram(bins = 30) +
  theme(plot.title = element_text(hjust = 0.5,
                                  size  = rel(1.5))) +
  labs(title = "Histogram of Total Basement Area",
       x = "Total Basement Area (sq. ft.)",
       y = "Frequency") +
  scale_x_continuous(limits = c(0, 3000))
```



2b.

Using “Ames” dataset to create a new dataset called “Ames2” in which you remove all cases corresponding to:

2bi.

MS_Zoning categories of A_agr (agricultural), C_all (commercial) and I_all (industrial),

2bii.

BsmtFin_Type_1 category of “No_Basement”.

2biii.

Bldg_Type category of “OneFam”

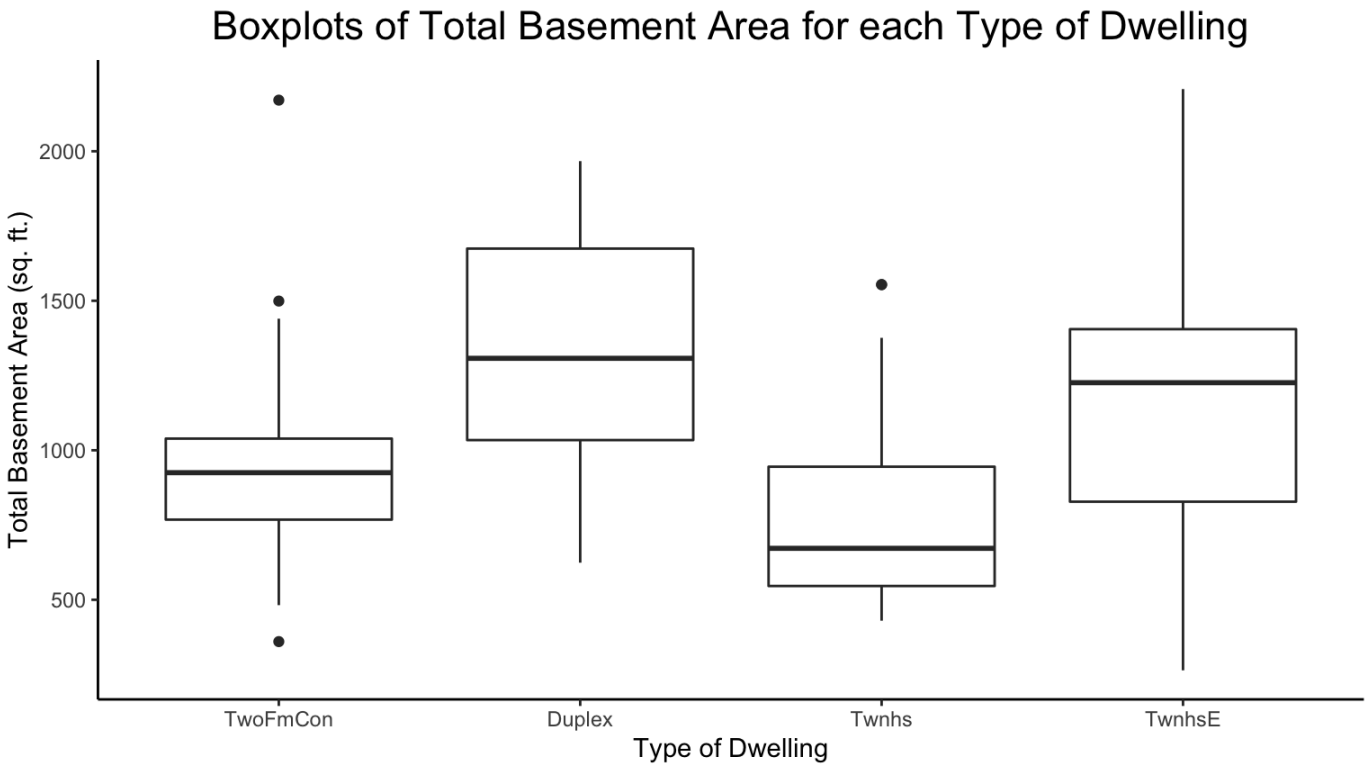
and drop the unused levels from the dataset “Ames2”. (4 points)

```
Ames2 <- Ames %>%
  filter(MS_Zoning != "A_agr" & MS_Zoning != "C_all" & MS_Zoning != "I_all" &
         BsmtFin_Type_1 != "No_Basement" & Bldg_Type != "OneFam") %>%
  droplevels()
```

2c.

Choose an appropriate plot to investigate the relationship between Bldg_Type and Total_Bsmt_SF in Ames2. (2 points)

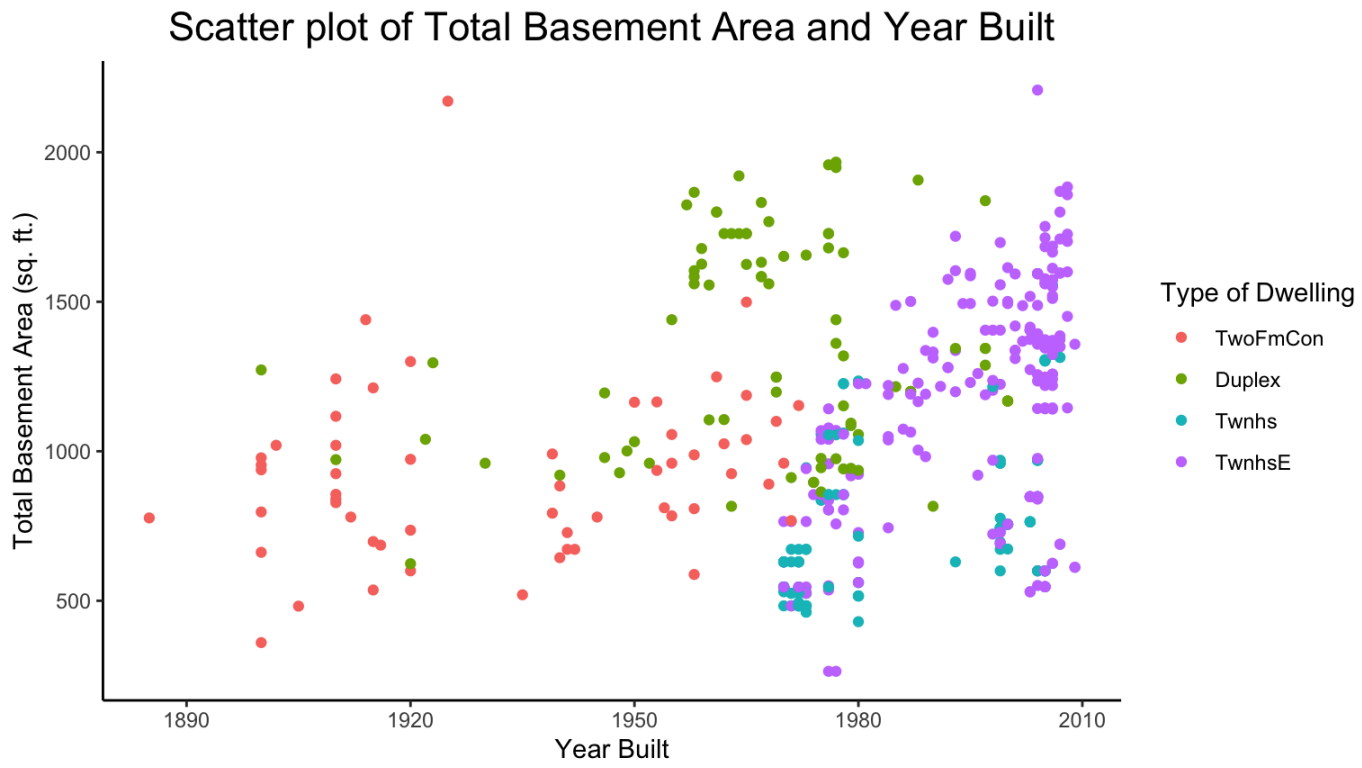
```
Ames2 %>%
  ggplot(mapping = aes(x = Bldg_Type, y = Total_Bsmt_SF)) +
  geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5,
                                  size  = rel(1.5))) +
  labs(title = "Boxplots of Total Basement Area for each Type of Dwelling",
       x = "Type of Dwelling",
       y = "Total Basement Area (sq. ft.)")
```



2d.

Choose an appropriate plot to investigate the relationship between Year_Built and Total_Bsmt_SF in Ames2. Color points according to the factor Bldg_Type. Ensure your plot has a clear title, axis labels and legend.

```
Ames2 %>%
  ggplot(mapping = aes(x = Year_Built,
                       y = Total_Bsmt_SF,
                       colour = Bldg_Type)) +
  geom_point() +
  theme(plot.title = element_text(hjust = 0.5, size  = rel(1.5))) +
  labs(title = "Scatter plot of Total Basement Area and Year Built",
       x = "Year Built",
       y = "Total Basement Area (sq. ft.)",
       colour = "Type of Dwelling")
```



What do you notice about how Basement size has changed over time?

There is a positive linear relationship. Total basement area increases over time.

Were there any slowdowns in construction over this period? When? Can you think why? (4 points)

The clearest slowdown in construction was during The Great Depression (1929 - 1939). On the plot there is a clear halt/slowdown in basement construction during this period. Other periods of a slowdown in basement construction may have been during the two world wars, World War I (1914–1918) and World War II (1939–1945). However, this is less evident in the plot.

2e.

Why do we make these plots? Comment on your findings from these plots (1 sentence is fine). (2 points)

We make these plots to visualize the relationship between the the variables before creating our model. In these plots we may be able to visualise and test for linearity and they may indicate the model may not be worth creating.

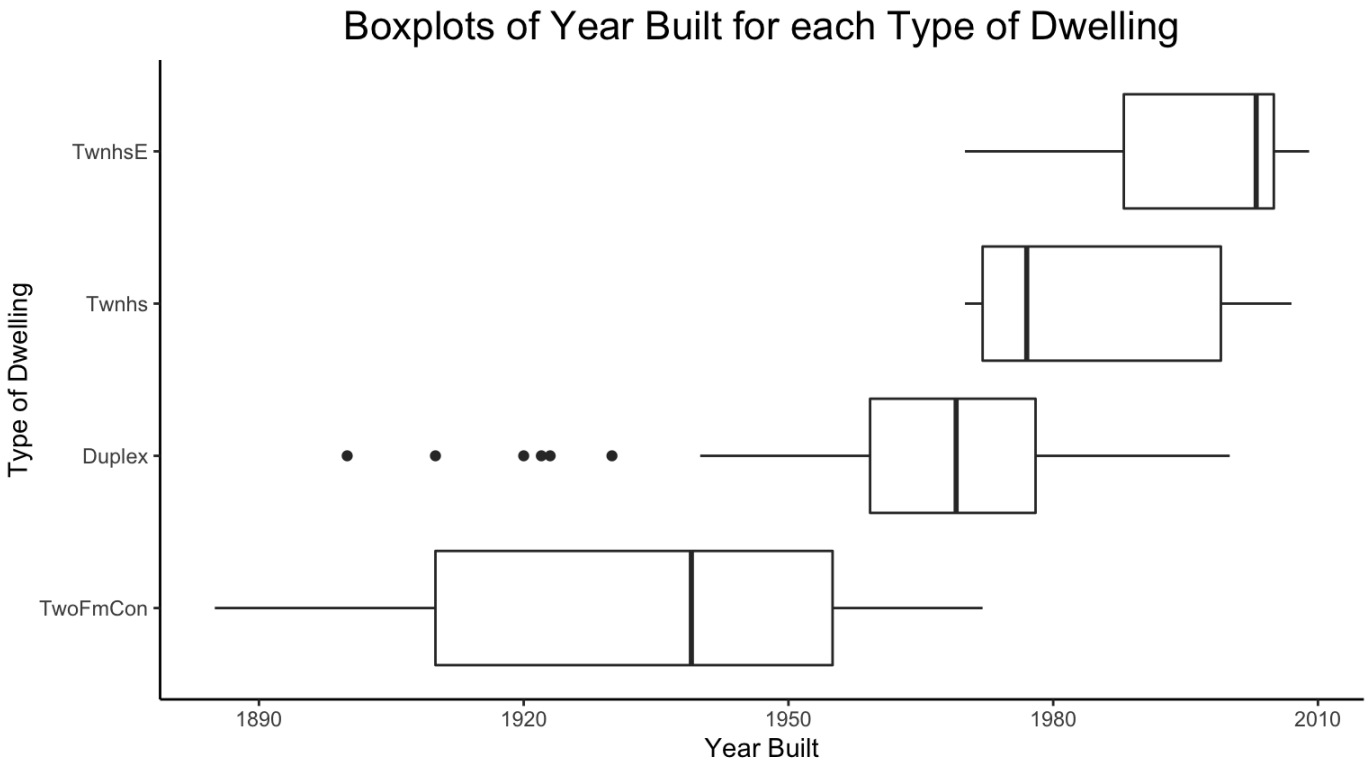
In the boxplot some types of dwelling have a higher average (median) total basement area than others. For example, duplex and extended townhouse have a higher average (median) total basement area than two family conversion and townhouse.

In the scatter plot there is a positive linear relationship in the total basement area compared to the year it was built. In addition, certain types of dwelling were built at different time periods. For example, townhouse and extended townhouse only started being built around the 1970s.

2f.

Now choose an appropriate plot to investigate the relationship between Bldg_Type and Year_Built in Ames2.

```
Ames2 %>%
  ggplot(mapping = aes(x = Year_Built, y = Bldg_Type)) +
  geom_boxplot() +
  theme(plot.title = element_text(hjust = 0.5,
                                  size  = rel(1.5))) +
  labs(title = "Boxplots of Year Built for each Type of Dwelling",
       x = "Year Built",
       y = "Type of Dwelling")
```



Why should we consider this? What do you notice? (3 points)

We should consider this plot because in the scatter plot we observed that there may be a relationship between the type of dwelling and the year it was built.

I notice that different types of dwelling tended to be built in different time periods and that they follow on from each other.

2g.

Use the lm command to build a linear model, linmod1, of Total_Bsmt_SF as a function of the predictors Bldg_Type and Year_Built for the “Ames2” dataset. (2 points)

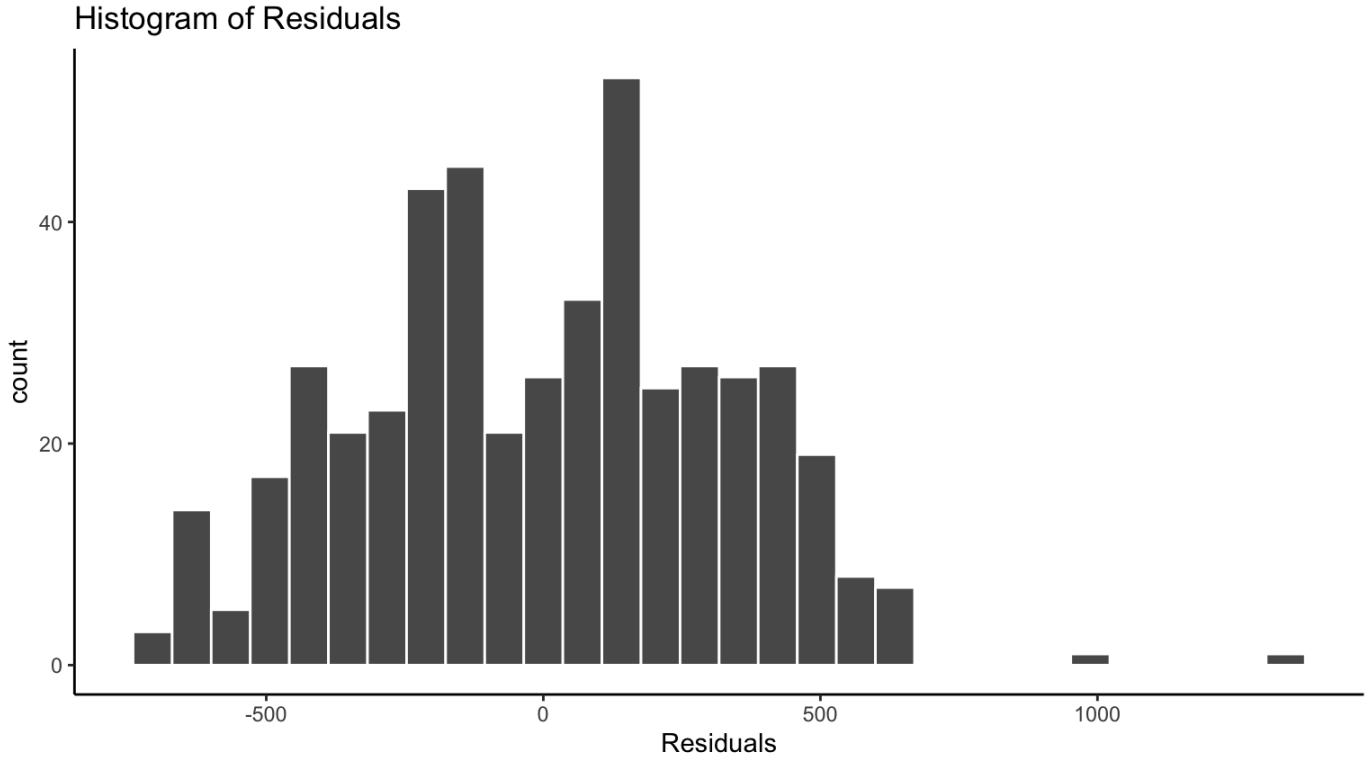
```
linmod1 <- lm(Total_Bsmt_SF ~ Bldg_Type + Year_Built, data = Ames2)
summary(linmod1)
```

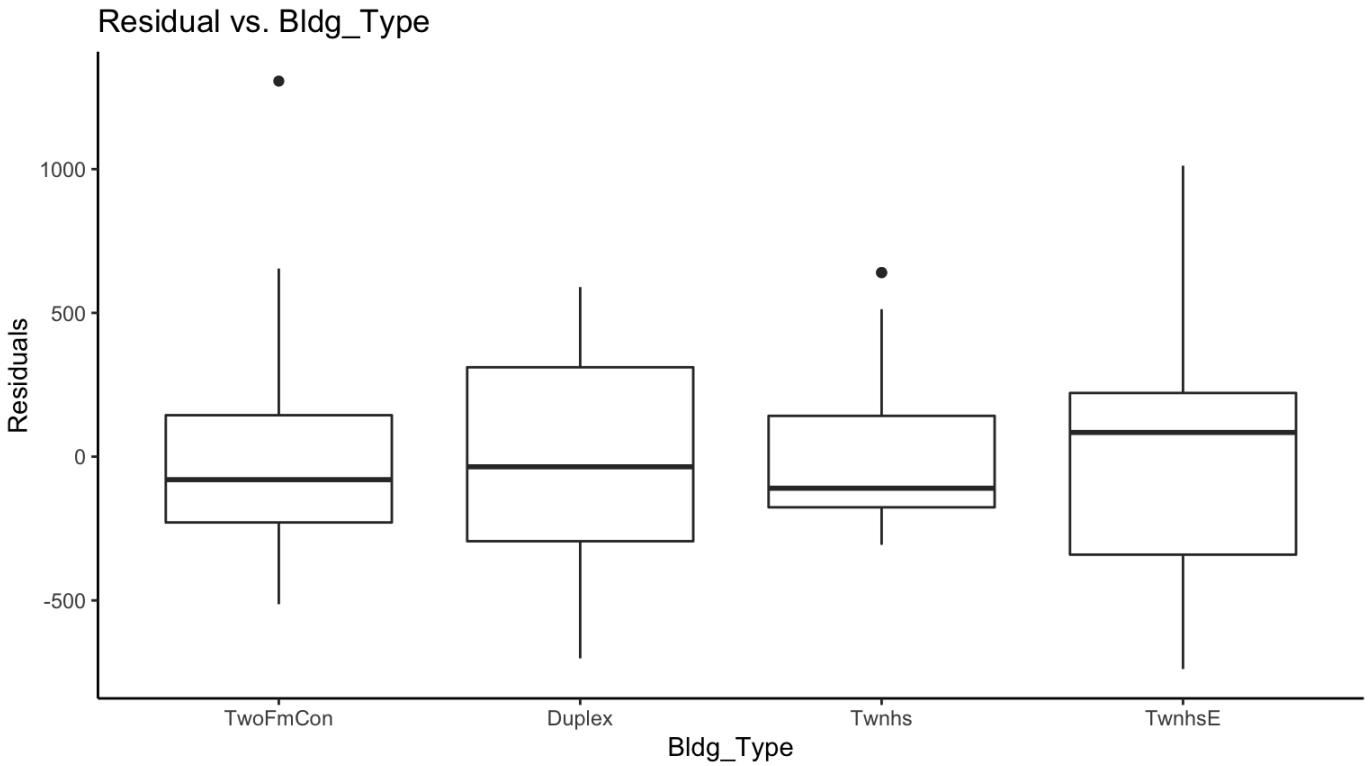
```
Call:
lm(formula = Total_Bsmt_SF ~ Bldg_Type + Year_Built, data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-738.53 -223.35    7.68  238.36 1306.23

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.293e+04  1.833e+03  -7.054 6.30e-12 ***
Bldg_TypeDuplex  1.870e+02  6.504e+01   2.875  0.00422 **
Bldg_TypeTwnhs  -5.314e+02  7.252e+01  -7.327 1.04e-12 ***
Bldg_TypeTwnhsE -2.349e+02  7.678e+01  -3.059  0.00235 **
Year_Built       7.166e+00  9.478e-01   7.560 2.15e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 327.1 on 467 degrees of freedom
Multiple R-squared:  0.3339,    Adjusted R-squared:  0.3282
F-statistic: 58.54 on 4 and 467 DF,  p-value: < 2.2e-16
```

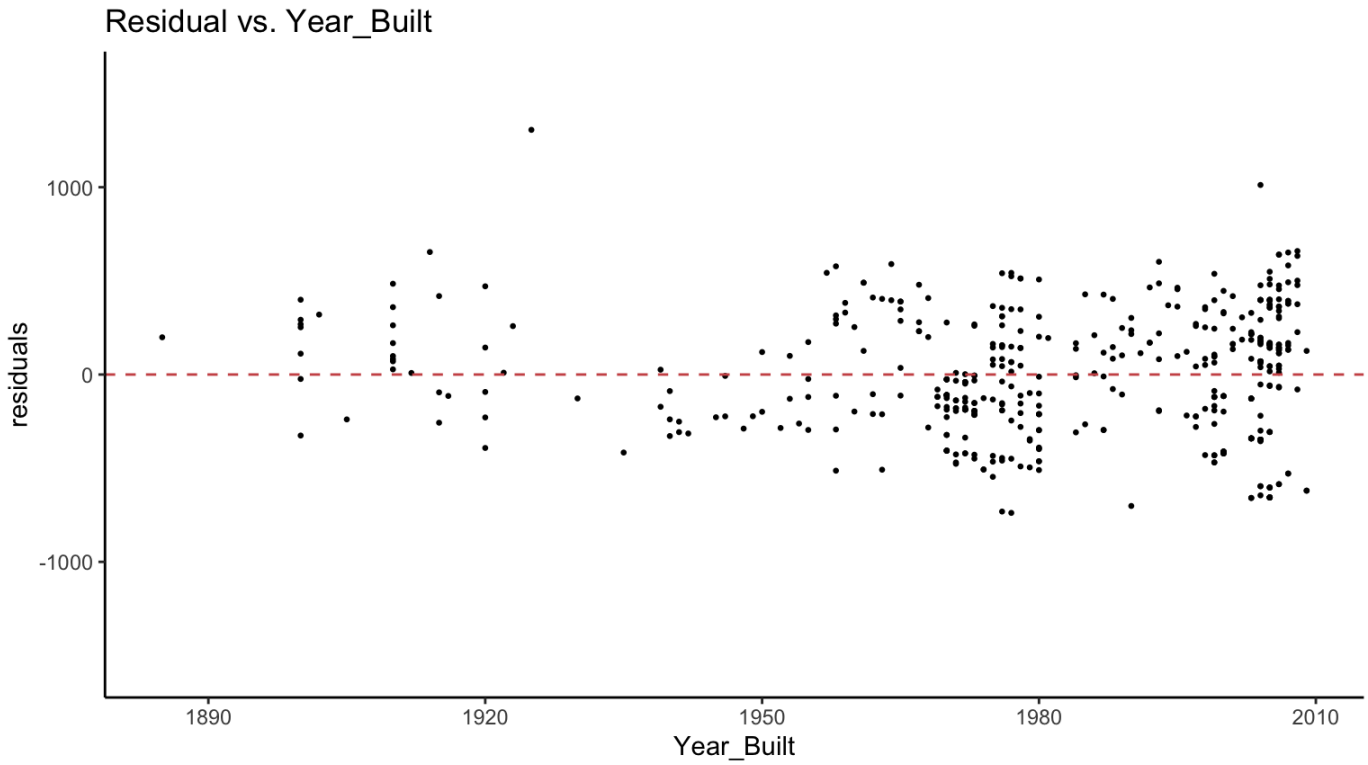
2h.

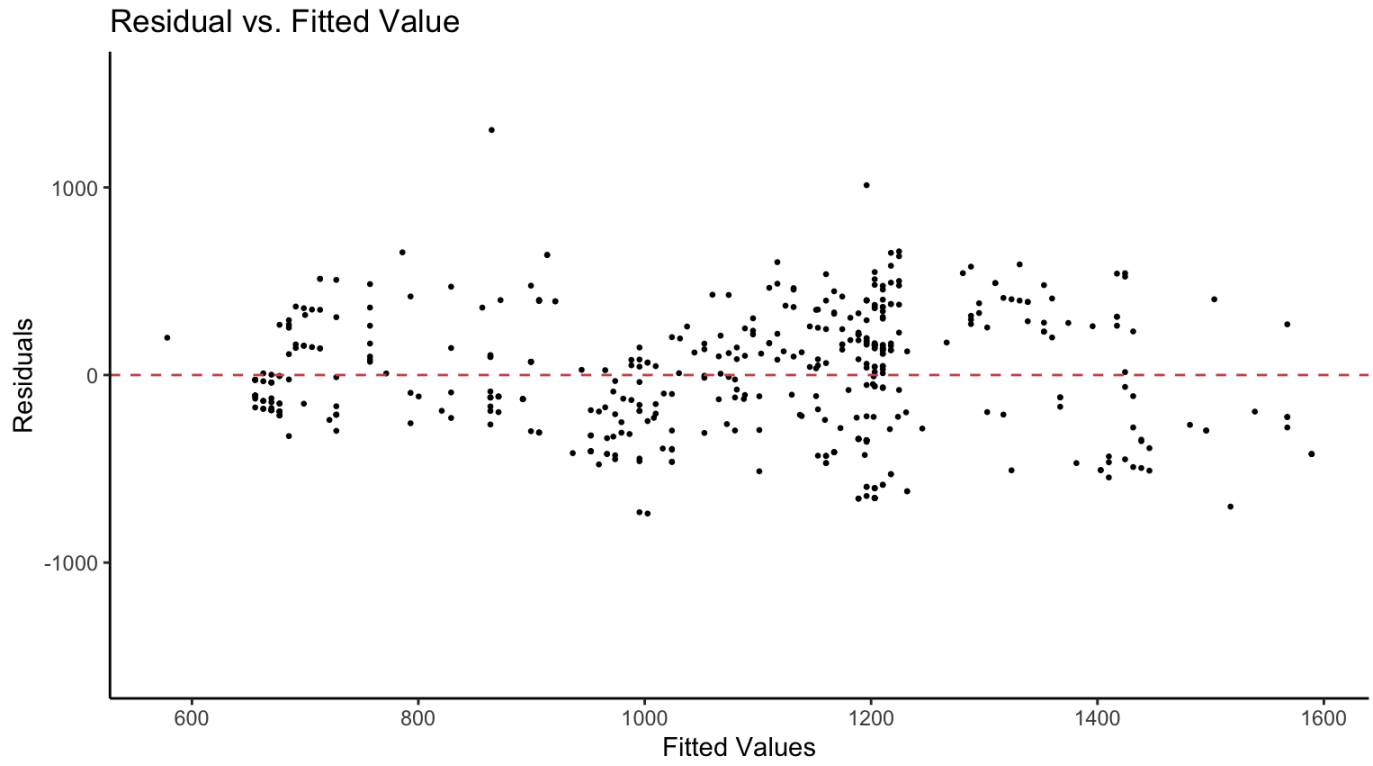
State and evaluate the assumptions of the model. (6 points)

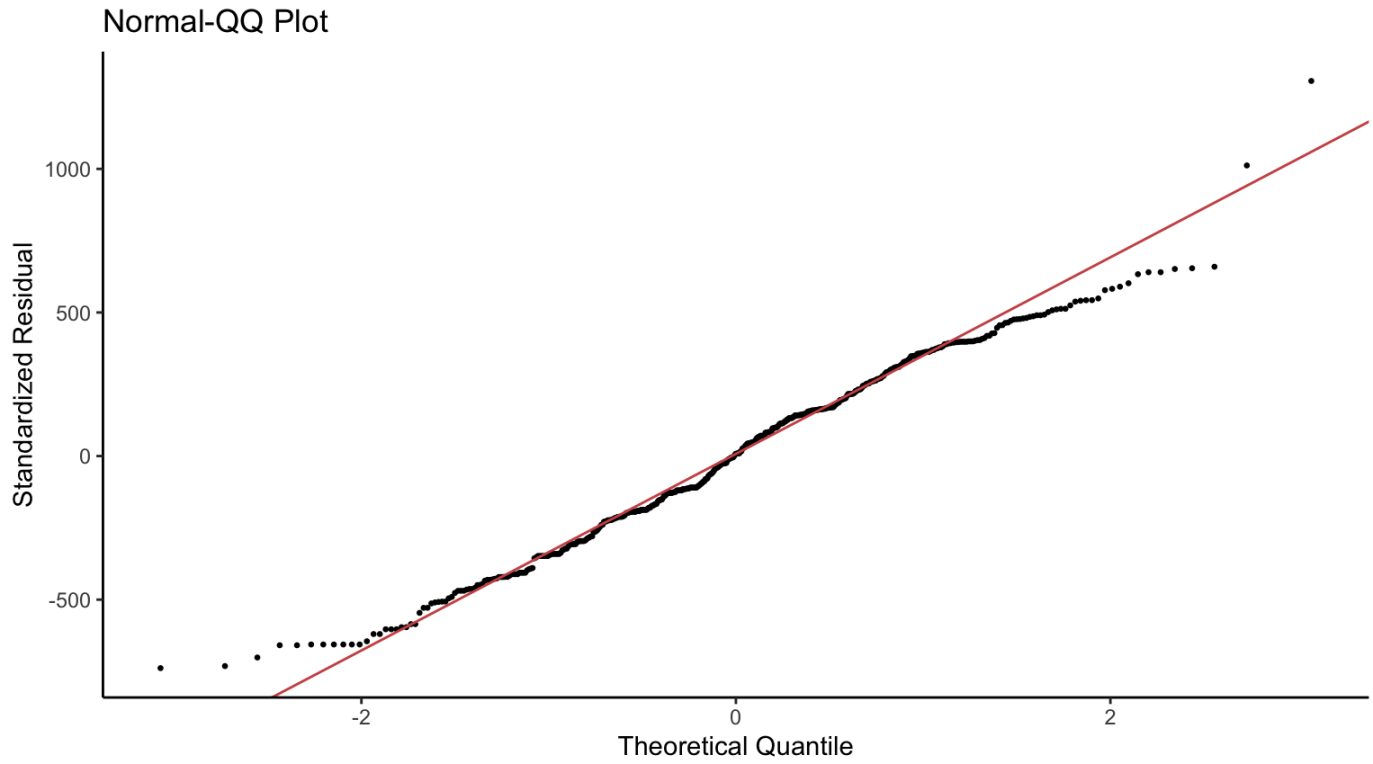
```
linmod1 %>%
  gg_diagnose(max.per.page = 1)
```

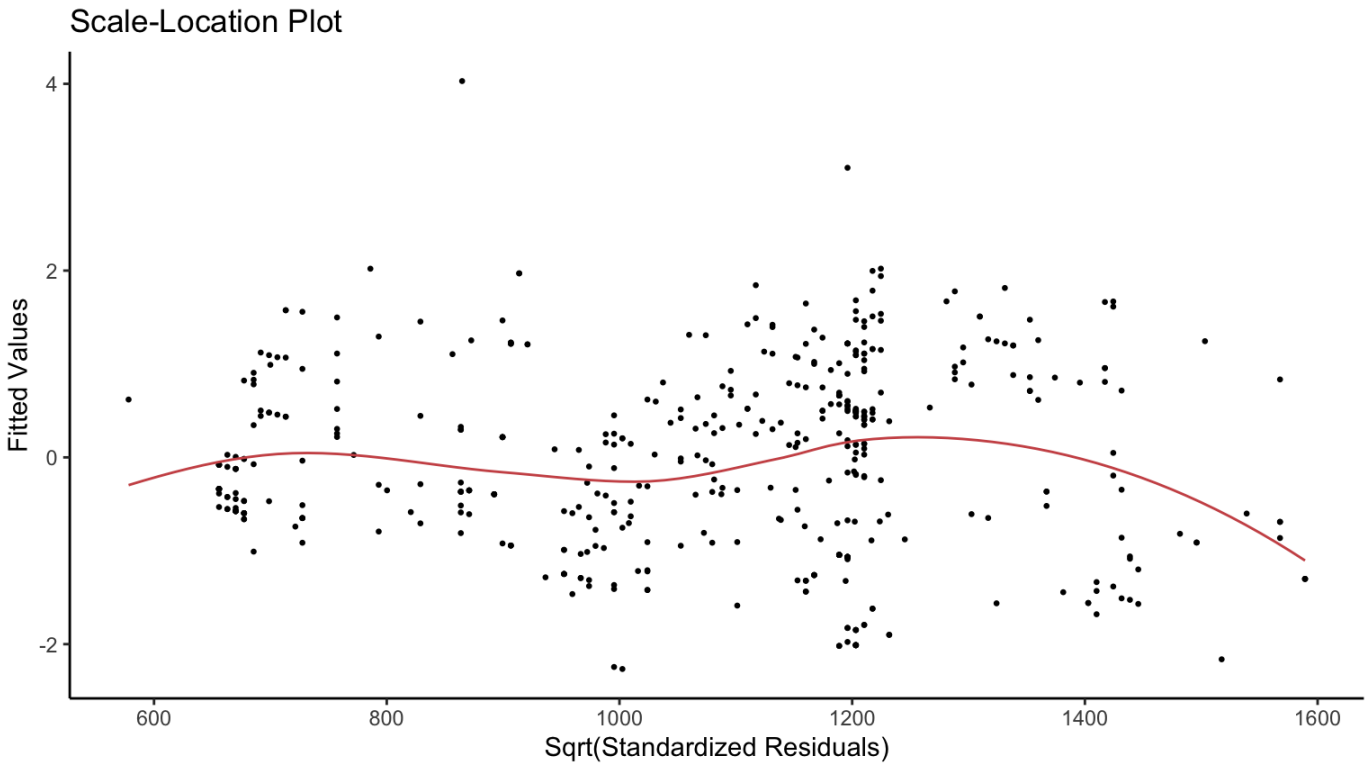


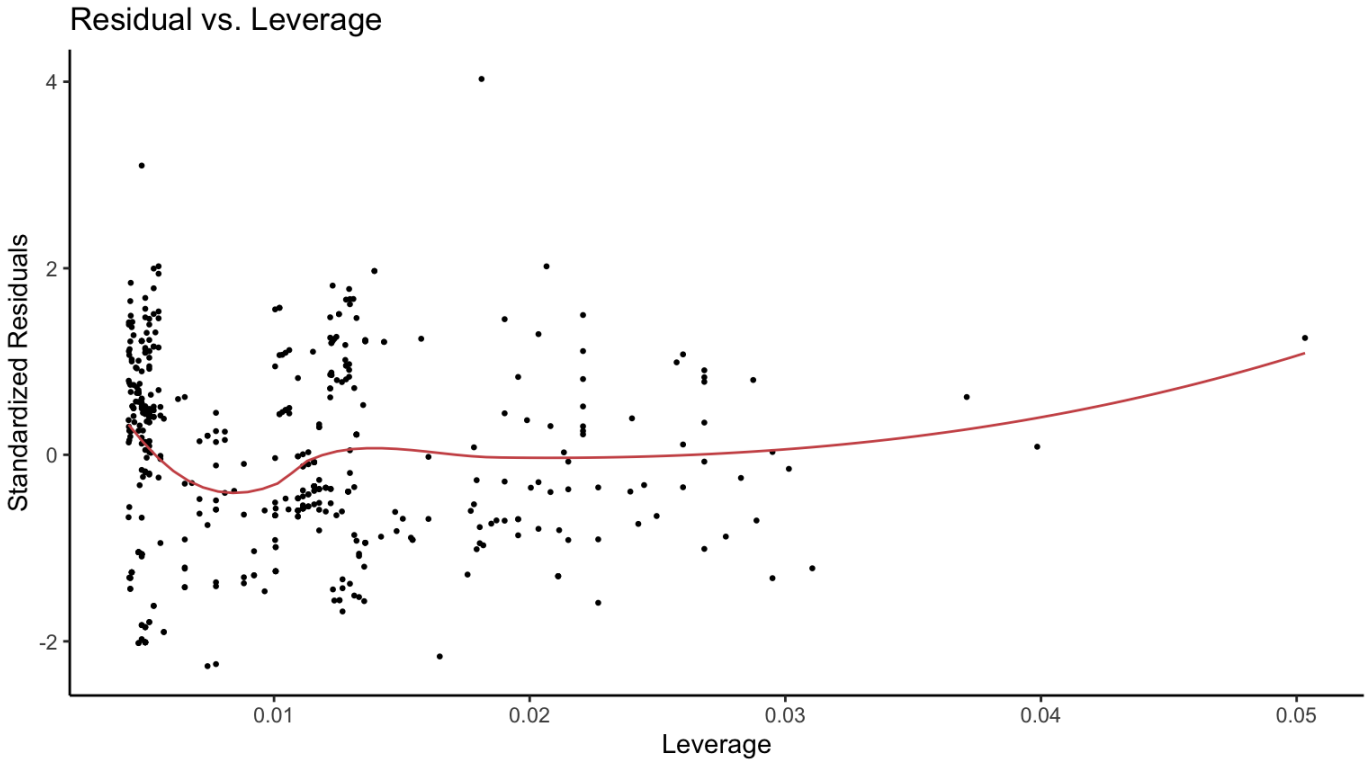

Residual vs. Bldg_Type
Residuals
1000
500
0
-500
TwoFmCon
Duplex
Twnhs
TwnhsE
Bldg_Type

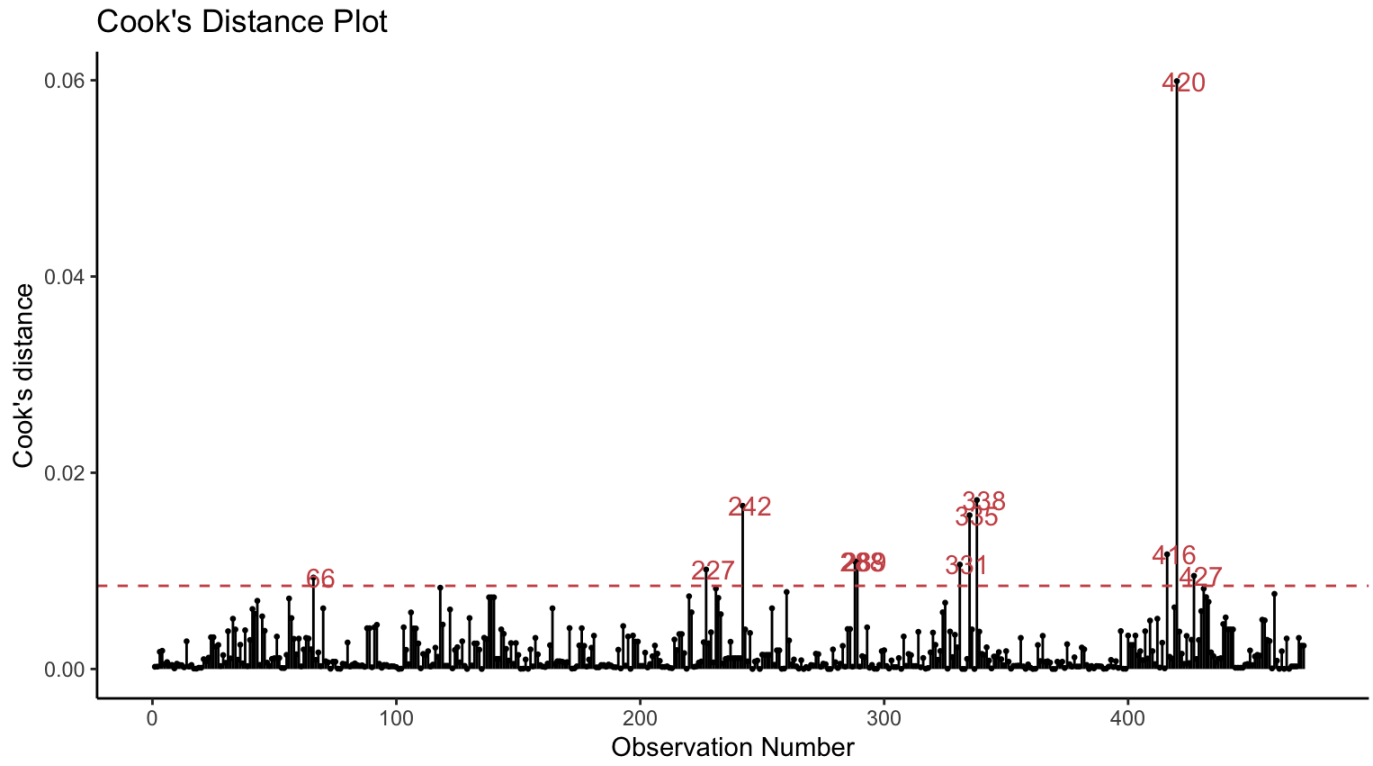

Residual vs. Year_Built
residuals
1000
0
-1000
1890
1920
1950
1980
2010
Year_Built


Residual vs. Fitted Value
Residuals
1000
0
-1000
600
800
1000
1200
1400
1600
Fitted Values


Normal-QQ Plot
Standardized Residual
1000
500
0
-500
-2
0
2
Theoretical Quantile


Scale-Location Plot
Fitted Values
4
2
0
-2
600
800
1000
1200
1400
1600
Sqrt(Standardized Residuals)


Residual vs. Leverage
Standardized Residuals
4
2
0
-2
0.01
0.02
0.03
0.04
0.05
Leverage



Assumptions:

1. Linearity

This assumption is evaluated before the linear model is created (when plotting the variables). The scatter plot before showed that there is a positive linear relationship between total basement area and the year it was built. So we can assume linearity.

2. Normality of residuals

The histogram of residuals looks Gaussian (there is a bell curve shape). There is a lack of bars to the right of the graph, but on the whole it is good.

The qqplot of residuals on the whole looks good. The points in the middle stay very close to the red line. The points on the edges deviate from the red line a bit, but not majorly.

So we can assume normality of residuals.

3. Homoscedasticity of residuals

The boxplot of residuals vs type of dwelling looks ok. The interquartile ranges are fairly similar, but some are larger than others (duplex and extended townhouse are both larger than two family conversion and townhouse)

The scatter plot of residuals vs year built looks ok as well. There are equal points above and below the red line (0). However, there are fewer points on the left than on the right. This may indicate slight heteroscedasticity.

2i.

Use the lm command to build a second linear model, linmod2, for Total_Bsmt_SF as a function of Bldg_Type, Year_Built and Lot_Area. (2 points)

```
linmod2 <- lm(Total_Bsmt_SF ~ Bldg_Type + Year_Built + Lot_Area,
              data = Ames2)
summary(linmod2)
```

```
Call:
lm(formula = Total_Bsmt_SF ~ Bldg_Type + Year_Built + Lot_Area,
    data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-810.32 -212.07   -5.72  233.88 1232.65

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.176e+04  1.828e+03  -6.435 3.08e-10 ***
Bldg_TypeDuplex  2.378e+02  6.529e+01   3.642 0.000301 ***
Bldg_TypeTwnhs  -4.120e+02  7.745e+01  -5.319 1.62e-07 ***
Bldg_TypeTwnhsE -1.265e+02  8.035e+01  -1.575 0.115942
Year_Built       6.509e+00  9.476e-01   6.868 2.09e-11 ***
Lot_Area         7.793e-03  1.960e-03   3.977 8.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 322.1 on 466 degrees of freedom
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.3489
F-statistic: 51.48 on 5 and 466 DF,  p-value: < 2.2e-16
```

2j.

Use Anova and Adjusted R-squared to compare these two models, and decide which is a better model. (6 points)

```
summary(linmod1)
summary(linmod2)
anova(linmod1,linmod2)
```

```
Call:
lm(formula = Total_Bsmt_SF ~ Bldg_Type + Year_Built, data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-738.53 -223.35    7.68  238.36 1306.23

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.293e+04  1.833e+03  -7.054 6.30e-12 ***
Bldg_TypeDuplex  1.870e+02  6.504e+01   2.875  0.00422 **
Bldg_TypeTwnhs  -5.314e+02  7.252e+01  -7.327 1.04e-12 ***
Bldg_TypeTwnhsE -2.349e+02  7.678e+01  -3.059  0.00235 **
Year_Built       7.166e+00  9.478e-01   7.560 2.15e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 327.1 on 467 degrees of freedom
Multiple R-squared:  0.3339,    Adjusted R-squared:  0.3282
F-statistic: 58.54 on 4 and 467 DF,  p-value: < 2.2e-16


Call:
lm(formula = Total_Bsmt_SF ~ Bldg_Type + Year_Built + Lot_Area,
    data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-810.32 -212.07   -5.72  233.88 1232.65

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.176e+04  1.828e+03  -6.435 3.08e-10 ***
Bldg_TypeDuplex  2.378e+02  6.529e+01   3.642 0.000301 ***
Bldg_TypeTwnhs  -4.120e+02  7.745e+01  -5.319 1.62e-07 ***
Bldg_TypeTwnhsE -1.265e+02  8.035e+01  -1.575 0.115942
Year_Built       6.509e+00  9.476e-01   6.868 2.09e-11 ***
Lot_Area         7.793e-03  1.960e-03   3.977 8.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 322.1 on 466 degrees of freedom
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.3489
F-statistic: 51.48 on 5 and 466 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Model 1: Total_Bsmt_SF ~ Bldg_Type + Year_Built
Model 2: Total_Bsmt_SF ~ Bldg_Type + Year_Built + Lot_Area
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    467 49980160
2    466 48339705  1   1640455 15.814 8.099e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

linmod1 Adjusted R-squared = 0.3282

linmod2 Adjusted R-squared = 0.3489

Comparison Anova p-value = 0.00008099 (< 0.05)

Linear model 2 has a slightly higher Adjusted R-squared, it also is significantly better according to the Anova (as the p value is less than 0.05). Therefore Linear model 2 is the better model.

2k.

Construct a confidence interval and a prediction interval for the basement area of a Twnhs built in 1980, with a lot Area of 7300. Explain what these two intervals mean. (6 points)

```
predict(linmod2, newdata = data.frame(Bldg_Type = "Twnhs",
                                      Year_Built = 1980,
                                      Lot_Area = 7300),
        interval = "confidence")

predict(linmod2, newdata = data.frame(Bldg_Type = "Twnhs",
                                      Year_Built = 1980,
                                      Lot_Area = 7300),
        interval = "prediction")
```

```
       fit      lwr      upr
1 768.7589 702.1423 835.3755
       fit      lwr      upr
1 768.7589 132.3605 1405.157
```

A confidence interval is the mean of your estimate plus and minus the variation in that estimate. You can be 95% confident that the coefficients fall between these ranges. Therefore the confidence intervals mean that we can be 95% confident that the basement area of a townhouse built in 1980, with a lot area of 7300 (768.76 sq. ft.) falls between 702.14 and 835.38 sq. ft.

A prediction interval defines a range of values within which a response is likely to fall given a specified value of a predictor. The prediction band will take into account both the confidence intervals and the variance in the residuals. The prediction interval is always larger than the confidence interval. Therefore the prediction intervals mean that we can be 95% confident that the next new observation of a basement area of a townhouse built in 1980, with a lot area of 7300 (768.76 sq. ft.) falls between 132.36 and 1405.16 sq. ft.

2l.

Now build a linear mixed model, linmod3, for Total_Bsmt_SF as a function of Year_Built,MS_Zoning and Bldg_Type. Use Neighborhood as random effect.

```
linmod3 <- lmer(Total_Bsmt_SF ~ Year_Built + MS_Zoning + Bldg_Type +
                (1|Neighborhood), data = Ames2)
summary(linmod3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Total_Bsmt_SF ~ Year_Built + MS_Zoning + Bldg_Type + (1 | Neighborhood)
   Data: Ames2

REML criterion at convergence: 6566.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1502 -0.5804 -0.0394  0.6330  4.4359

Random effects:
 Groups       Name        Variance Std.Dev.
 Neighborhood (Intercept) 35128    187.4
 Residual                 68517    261.8
Number of obs: 472, groups:  Neighborhood, 27

Fixed effects:
                                  Estimate Std. Error t value
(Intercept)                      -4890.652   2262.148  -2.162
Year_Built                           2.876      1.151   2.499
MS_ZoningResidential_High_Density  148.504    211.630   0.702
MS_ZoningResidential_Low_Density   288.369    198.824   1.450
MS_ZoningResidential_Medium_Density 109.234   197.814   0.552
Bldg_TypeDuplex                    264.530     59.046   4.480
Bldg_TypeTwnhs                     -63.140     87.020  -0.726
Bldg_TypeTwnhsE                    105.171     83.438   1.260

Correlation of Fixed Effects:
            (Intr) Yr_Blt MS_ZR_H MS_ZR_L MS_ZR_M Bld_TD Bld_TT
Year_Built  -0.996
MS_ZnnR_H_D -0.158  0.081
MS_ZnnR_L_D -0.169  0.085  0.912
MS_ZnnR_M_D -0.142  0.061  0.901   0.963
Bldg_TypDpl  0.378 -0.395  0.014  -0.017  -0.001
Bldg_TypTwn  0.546 -0.570  0.004   0.059  -0.006   0.617
Bldg_TypTwE  0.602 -0.626 -0.005   0.052  -0.010   0.662  0.911
```

What is the critical number to pull out from this, and what does it tell us? (4 points)

critical number = standard deviation (SD) of neighborhood = 187.4

This tells us that the overall random effect of neighborhood on total basement area has an SD of 187.4 sq. ft.

This means that neighborhood affects the total basement area by ± 93.7 sq. ft.

2m.

Construct 95% confidence intervals around each parameter estimate for linmod3.

```
confint(linmod3)
```

```
                                        2.5 %     97.5 %
.sig01                                114.916972  253.19244
.sigma                                244.221572  278.77404
(Intercept)                         -9699.022207 -595.99553
Year_Built                              0.691417    5.33084
MS_ZoningResidential_High_Density    -254.190137  549.38121
MS_ZoningResidential_Low_Density      -91.829020  665.77648
MS_ZoningResidential_Medium_Density  -266.136920  487.72182
Bldg_TypeDuplex                       145.073466  377.00462
Bldg_TypeTwnhs                       -249.148897  102.22240
Bldg_TypeTwnhsE                       -68.266514  263.18536
```

What does this tell us about the significant of the random effect? (3 points)

.sig01 = confidence interval for random effect

This tells us that the random effect is significant because the confidence interval does not cross 0.

2n.

Write out the full mathematical expression for the model in linmod2 and for the model in linmod3. Round to the nearest integer in all coefficients with modulus > 10 and to three decimal places for coefficients with modulus < 10. (4 points)

```
summary(linmod2)
```

```

Call:
lm(formula = Total_Bsmt_SF ~ Bldg_Type + Year_Built + Lot_Area,
    data = Ames2)

Residuals:
    Min      1Q  Median      3Q     Max
-810.32 -212.07   -5.72  233.88 1232.65

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.176e+04  1.828e+03  -6.435 3.08e-10 ***
Bldg_TypeDuplex  2.378e+02  6.529e+01   3.642 0.000301 ***
Bldg_TypeTwnhs  -4.120e+02  7.745e+01  -5.319 1.62e-07 ***
Bldg_TypeTwnhsE -1.265e+02  8.035e+01  -1.575 0.115942
Year_Built       6.509e+00  9.476e-01   6.868 2.09e-11 ***
Lot_Area         7.793e-03  1.960e-03   3.977 8.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 322.1 on 466 degrees of freedom
Multiple R-squared:  0.3558,    Adjusted R-squared:  0.3489
F-statistic: 51.48 on 5 and 466 DF,  p-value: < 2.2e-16
```

Linmod2 Equation:

$$\begin{aligned}\text{TotalBasementArea} \sim N(&-11760 \\ &+ 240 \times \text{isDwellingTypeDuplex} \\ &+ -412 \times \text{isDwellingTypeTownhouse} \\ &+ -127 \times \text{isDwellingTypeExtendedTownhouse} \\ &+ 6.509 \times \text{YearBuilt} \\ &+ 0.00779 \times \text{LotArea}, 322.1)\end{aligned}$$

```
summary(linmod3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Total_Bsmt_SF ~ Year_Built + MS_Zoning + Bldg_Type + (1 | Neighborhood)
   Data: Ames2

REML criterion at convergence: 6566.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1502 -0.5804 -0.0394  0.6330  4.4359

Random effects:
 Groups       Name        Variance Std.Dev.
 Neighborhood (Intercept) 35128    187.4
 Residual                 68517    261.8
Number of obs: 472, groups:  Neighborhood, 27

Fixed effects:
                                  Estimate Std. Error t value
(Intercept)                      -4890.652   2262.148  -2.162
Year_Built                           2.876      1.151   2.499
MS_ZoningResidential_High_Density  148.504    211.630   0.702
MS_ZoningResidential_Low_Density   288.369    198.824   1.450
MS_ZoningResidential_Medium_Density 109.234   197.814   0.552
Bldg_TypeDuplex                    264.530     59.046   4.480
Bldg_TypeTwnhs                     -63.140     87.020  -0.726
Bldg_TypeTwnhsE                    105.171     83.438   1.260

Correlation of Fixed Effects:
            (Intr) Yr_Blt MS_ZR_H MS_ZR_L MS_ZR_M Bld_TD Bld_TT
Year_Built  -0.996
MS_ZnnR_H_D -0.158  0.081
MS_ZnnR_L_D -0.169  0.085  0.912
MS_ZnnR_M_D -0.142  0.061  0.901   0.963
Bldg_TypDpl  0.378 -0.395  0.014  -0.017  -0.001
Bldg_TypTwn  0.546 -0.570  0.004   0.059  -0.006   0.617
Bldg_TypTwE  0.602 -0.626 -0.005   0.052  -0.010   0.662  0.911
```

Linmod3 Equation:

$$\begin{aligned}\text{TotalBasementArea} \sim N(&-4891 \\ &+ 2.876 \times \text{YearBuilt} \\ &+ 149 \times \text{isResidentialZoneHighDensity} \\ &+ 288 \times \text{isResidentialZoneLowDensity} \\ &+ 109 \times \text{isResidentialZoneMediumDensity} \\ &+ 265 \times \text{isDwellingTypeDuplex} \\ &+ -63 \times \text{isDwellingTypeTownhouse} \\ &+ 105 \times \text{isDwellingTypeExtendedTownhouse} \\ &+ U, 261.8)\end{aligned}$$

3. Logistic Regression

3a.

Do the following:

3ai.

Create a new dataset called “Ames3” that contains all data in “Ames” dataset plus a new variable “excellent_heating” that indicates if the heating quality and condition “Heating_QC” is excellent or not. (2 points)

```
Ames3 <- Ames %>%
  mutate(excellent_heating = case_when(
    Heating_QC == "Excellent" ~ "True",
    Heating_QC != "Excellent" ~ "False")) %>%
  mutate(excellent_heating = as.logical(excellent_heating))
```

3aii.

In “Ames3” dataset, remove all cases “3” and “4” corresponding to the Fireplaces variable. Remove all cases where Lot_Frontage is greater than 130 or smaller than 20. Drop the unused levels from . (2 points)

```
Ames3 <- Ames3 %>%
   filter(Fireplaces != 3 & Fireplaces != 4 &
          Lot_Frontage <= 130 & Lot_Frontage >= 20) %>%
  droplevels()
```

3aiii.

Save “Fireplaces” as factor in “Ames3” dataset (1 point)

```
Ames3 <- Ames3 %>%
  mutate(Fireplaces = as.factor(Fireplaces))
```

3aiv.

Construct a logistic regression model glmod for excellent_heating as a function of Lot_Frontage and Fireplaces for the dataset “Ames3”. (2 points)

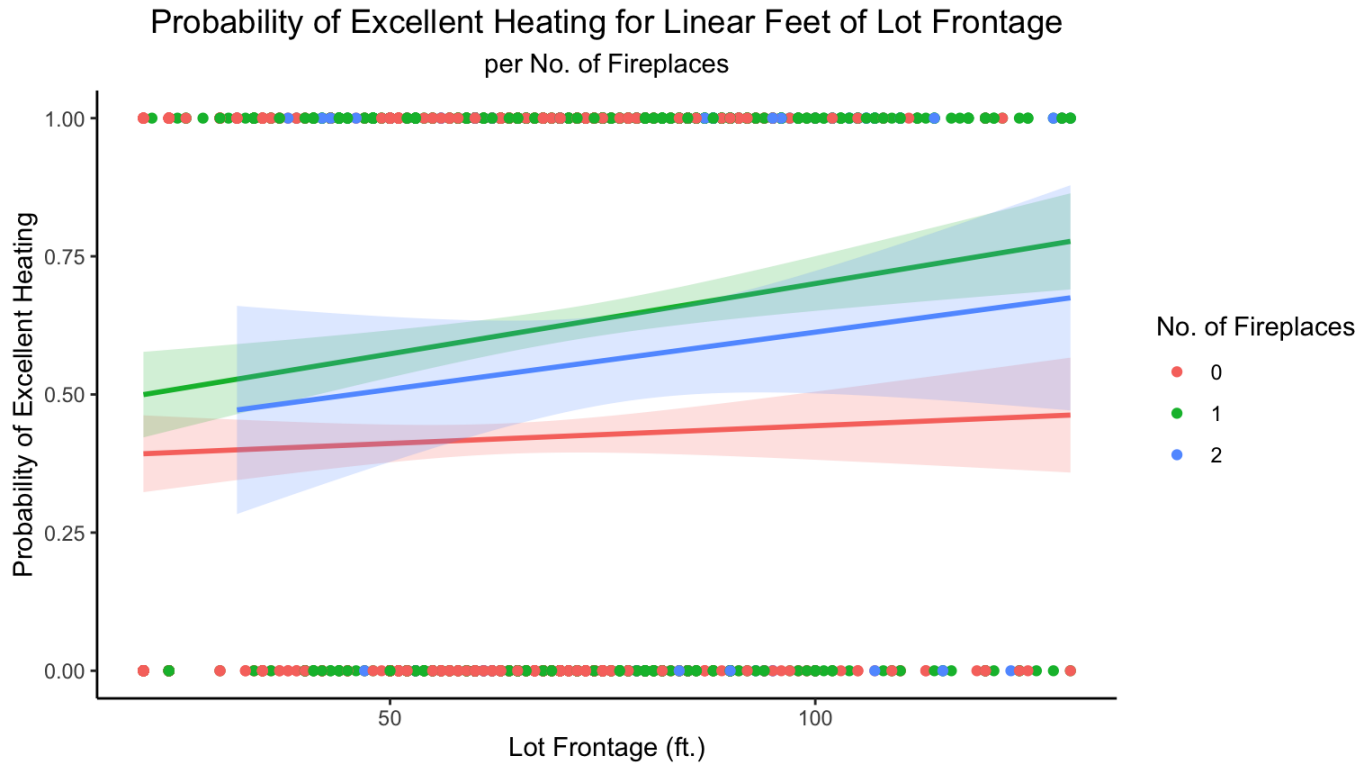
```
glmod <- glm(excellent_heating ~ Lot_Frontage + Fireplaces,
             family = "binomial", data = Ames3)
summary(glmod)
```

```

Call:
glm(formula = excellent_heating ~ Lot_Frontage + Fireplaces,
    family = "binomial", data = Ames3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5932  -1.0761   0.8695   1.0442   1.4503

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.769387   0.147977  -5.199    2e-07 ***
Lot_Frontage  0.007018   0.002137   3.285  0.00102 **
Fireplaces1   0.796183   0.088528   8.994  < 2e-16 ***
Fireplaces2   0.494887   0.176308   2.807  0.00500 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3324.2  on 2399  degrees of freedom
Residual deviance: 3213.3  on 2396  degrees of freedom
AIC: 3221.3

Number of Fisher Scoring iterations: 4
```

3b.

Construct confidence bands for the variable excellent_heating as a function of Lot_Frontage for each number of Fireplaces (hint: create a new data frame for each number of Fireplaces). Colour these with different transparent colours for each number of Fireplaces and plot them together on the same axes. Put the actual data on the plot, coloured to match the bands, and jittered in position to make it possible to see all points. Ensure you have an informative main plot title, axes labels and a legend. (7 points)

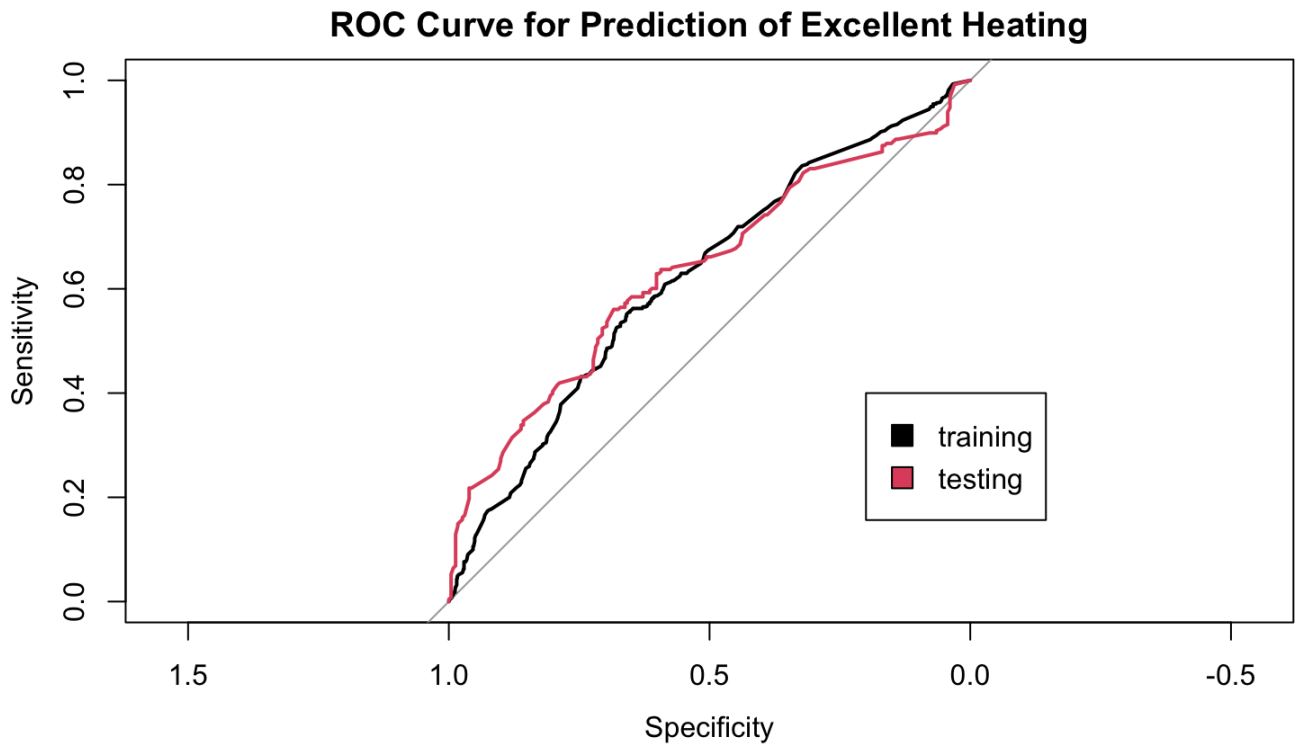
```
# Ames3_ci1 <- Ames3 %>%
#   filter(Fireplaces == 0) %>%
#   droplevels()
#
# Ames3_ci2 <- Ames3 %>%
#   filter(Fireplaces == 1) %>%
#   droplevels()
#
# Ames3_ci3 <- Ames3 %>%
#   filter(Fireplaces == 2) %>%
#   droplevels()


# ilink <- family(glmod)$linkinv
#
# lf <- with(Ames3, data.frame(Lot_Frontage = seq(
#                                  min(Ames3$Lot_Frontage),
#                                  max(Ames3$Lot_Frontage), length = 100)))
#
# lf <- cbind(lf, predict(glmod, lf,
#                                type = "link", se.fit = TRUE)[1:2])
#
# lf <-transform(lf, Fitted = ilink(fit),
#                            Upper = ilink(fit + (1.96 * se.fit)),
#                            Lower = ilink(fit - (1.96 * se.fit)))
#
# ggplot(Ames3, aes(x = Lot_Frontage,
#                       y = as.numeric(as.factor(excellent_heating)) - 1),
#                       colour = fireplaces) +
#   geom_ribbon(data = lf, aes(ymin = Lower, ymax = Upper, x = Lot_Frontage),
#               fill = "steelblue2", alpha = 0.2, inherit.aes = FALSE) +
#   geom_line(data = lf, aes(y = Fitted, x = Lot_Frontage)) +
#   geom_point() +
#   labs(y = "Probability of Excellent Heating", x = "Lot_Frontage")


ggplot(Ames3, aes(colour = Fireplaces,
                  x = Lot_Frontage,
                  y = as.numeric(as.factor(excellent_heating)) - 1)) +
   geom_smooth(method = "glm", alpha = 0.2,
               aes(fill = Fireplaces), show.legend = FALSE) +
   geom_point() +
   theme(plot.title = element_text(hjust = 0.5,
                                   size  = rel(1.25)),
         plot.subtitle = element_text(hjust = 0.5,
                                   size  = rel(1))) +
   labs(
    title = "Probability of Excellent Heating for Linear Feet of Lot Frontage",
    subtitle = "per No. of Fireplaces",
    y = "Probability of Excellent Heating",
    x = "Lot Frontage (ft.)",
    colour = "No. of Fireplaces")
```



3c.

Split the data using set.seed(120) and rebuild the model on 80% of the data. Cross validate on the remaining 20%. Plot the ROCs for both data and comment on your findings. (6 points)

```
set.seed(120)

training_sample <- c(Ames3$excellent_heating) %>%
  createDataPartition(p = 0.8, list = FALSE)

train_data <- Ames3[training_sample, ]
test_data <- Ames3[-training_sample, ]


train_model <- glm(excellent_heating ~ Lot_Frontage + Fireplaces,
                   family = "binomial", data = train_data)


predtrain <- predict(train_model, type="response")

predtest <- predict(train_model, newdata = test_data, type = "response")


roctrain<-roc(response = train_data$excellent_heating,
              predictor = predtrain,
              plot = TRUE,
              main = "ROC Curve for Prediction of Excellent Heating",
              auc = TRUE)

roc(response = test_data$excellent_heating, predictor = predtest,
    plot = TRUE, auc = TRUE, add = TRUE, col = 2)

legend(0.2, 0.4, legend = c("training", "testing"), fill = 1:2)
```




```
Call:
roc.default(response = test_data$excellent_heating, predictor = predtest,     auc = T
RUE, plot = TRUE, add = TRUE, col = 2)

Data: predtest in 231 controls (test_data$excellent_heating FALSE) < 248 cases (test_
data$excellent_heating TRUE).
Area under the curve: 0.6337
```

The ROC curves shows that the black (training) ROC and the red (testing) ROC are very similar and they overlap. This is a sign that the training model is a good fit with the testing data. There is a slight bit of overfitting near the top of the curves (at around 0.8 - 1.0 sensitivity), however this is very minimal and the training model is therefore a very good fit.

4. Multinomial Regression

4a.

For the dataset “Ames”, create a model multregmod to predict BsmtFin_Type_1 from Total_Bsmt_SF and Year_Remod_Add. (3 points)

```
multregmod <- multinom(BsmtFin_Type_1 ~ Total_Bsmt_SF + Year_Remod_Add,
                       data = Ames)
summary(multregmod)
```

```
# weights:  28 (18 variable)
initial  value 5701.516737
iter  10 value 4611.614897
iter  20 value 4159.256253
iter  30 value 4153.561922
iter  40 value 4150.324236
iter  50 value 4146.549270
iter  60 value 4144.509436
iter  70 value 4144.474970
final  value 4144.474825
converged
Call:
multinom(formula = BsmtFin_Type_1 ~ Total_Bsmt_SF + Year_Remod_Add,
    data = Ames)

Coefficients:
            (Intercept) Total_Bsmt_SF Year_Remod_Add
BLQ           34.465254  6.282504e-05   -0.017706708
GLQ         -105.324418  1.030040e-03    0.052676145
LwQ           39.566891  1.243787e-05   -0.020550529
No_Basement    4.876103 -1.729079e-01    0.004007987
Rec           56.710979  1.596801e-06   -0.028929851
Unf          -29.377212 -6.987213e-04    0.015514051

Std. Errors:
             (Intercept) Total_Bsmt_SF Year_Remod_Add
BLQ         6.503585e-08  0.0002288432   1.248219e-04
GLQ         5.006809e-08  0.0001690380   9.871691e-05
LwQ         7.902828e-08  0.0002807542   1.516674e-04
No_Basement 5.427383e-06  0.0001760992   1.072864e-02
Rec         6.504934e-08  0.0002315851   1.245335e-04
Unf         4.870842e-08  0.0001739219   9.386388e-05

Residual Deviance: 8288.95
AIC: 8324.95
```

4b.

Write out the formulas for this model in terms of P(No_Basement), P(Unf) P(Rec),P(BLQ), P(GLQ), P(LwQ), You may round coefficients to 3 dp. (4 points)

$$\text{P(NoBasement)} = 4.876 + -0.173 \times \text{TotalBasementArea} + 0.004 \times \text{YearRemodelled/YearConstructed}$$

$$\text{P(Unf)} = -29.377 + -0.0007 \times \text{TotalBasementArea} + 0.016 \times \text{YearRemodelled/YearConstructed}$$

$$\text{P(Rec)} = 56.711 + 0.000002 \times \text{TotalBasementArea} + -0.029 \times \text{YearRemodelled/YearConstructed}$$

$$\text{P(BLQ)} = 34.465 + 0.00006 \times \text{TotalBasementArea} + -0.018 \times \text{YearRemodelled/YearConstructed}$$

$$\text{P(GLQ)} = -105.324 + 0.001 \times \text{TotalBasementArea} + 0.053 \times \text{YearRemodelled/YearConstructed}$$

$$\text{P(LwQ)} = 39.567 + 0.00001 \times \text{TotalBasementArea} + -0.021 \times \text{YearRemodelled/YearConstructed}$$

4c.

Evaluate the performance of this model using a confusion matrix and by calculating the sum of sensitivities for the model. Comment on your findings. (4 points)

```
multitable <- table(Ames$BsmtFin_Type_1, predict(multregmod, type = "class"))

names(dimnames(multitable)) <- list("Actual", "Predicted")

multitable


SSens <- multitable[1,1] / sum(Ames$BsmtFin_Type_1 == "ALQ") +
         multitable[2,2] / sum(Ames$BsmtFin_Type_1 == "BLQ") +
         multitable[3,3] / sum(Ames$BsmtFin_Type_1 == "GLQ") +
         multitable[4,4] / sum(Ames$BsmtFin_Type_1 == "LwQ") +
         multitable[5,5] / sum(Ames$BsmtFin_Type_1 == "No_Basement") +
         multitable[6,6] / sum(Ames$BsmtFin_Type_1 == "Rec") +
         multitable[7,7] / sum(Ames$BsmtFin_Type_1 == "Unf")

SSens

CCR <- (multitable[1,1] +
        multitable[2,2] +
        multitable[3,3] +
        multitable[4,4] +
        multitable[5,5] +
        multitable[6,6] +
        multitable[7,7]) / length(Ames$BsmtFin_Type_1)

CCR
```

```
            Predicted
Actual        ALQ BLQ GLQ LwQ No_Basement Rec Unf
  ALQ           1   0 117   0           0  18 293
  BLQ           0   0  50   0           0  30 189
  GLQ           1   0 579   0           0   2 277
  LwQ           1   0  38   0           0  30  85
  No_Basement   0   0   0   0          80   0   0
  Rec           3   0  31   0           0  46 208
  Unf           6   0 291   0           0  76 478
[1] 2.397785
[1] 0.4040956
```

The sum of sensitivities = 2.398

The correct classification rate = 0.404

This value seems high but there are seven categories so in actual fact its not that high. Maximising this number will produce the best model.

The correct classification rate gives us the overall number of correct classifications.

5. Poisson/quassipoisson Regression

5a.

For the “footballer_data” dataset, create a model appearances_mod to predict the total number of overall appearances a player had based on position and age. (2 points)

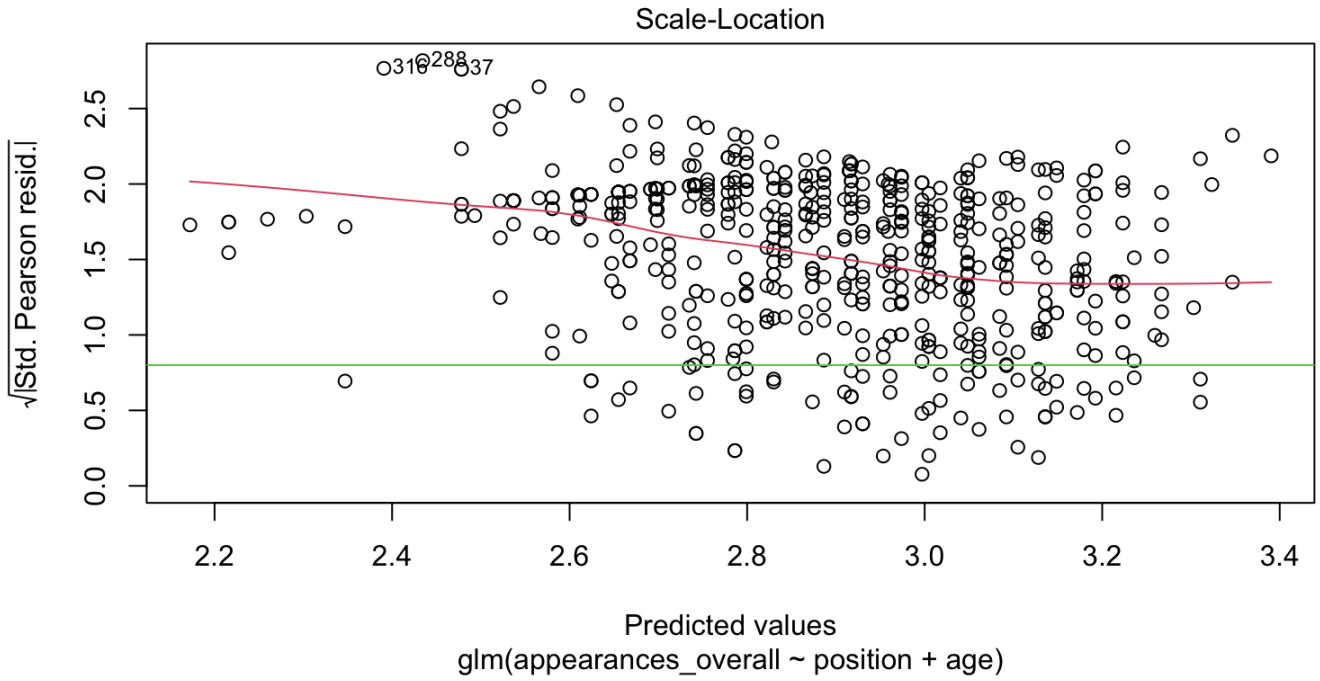
```
appearances_mod <- glm(appearances_overall ~ position + age,
                       data = footballer_data, family = "poisson")
summary(appearances_mod)
```

```
Call:
glm(formula = appearances_overall ~ position + age, family = "poisson",
    data = footballer_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.5377  -3.5215   0.0351   2.1892   6.1853

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.575316   0.074884  21.037  < 2e-16 ***
positionForward     0.110606   0.027448   4.030 5.59e-05 ***
positionGoalkeeper -0.364605   0.040780  -8.941  < 2e-16 ***
positionMidfielder  0.118259   0.023309   5.074 3.90e-07 ***
age                 0.043704   0.002392  18.275  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6539.7  on 564  degrees of freedom
Residual deviance: 6114.4  on 560  degrees of freedom
AIC: 8417.1

Number of Fisher Scoring iterations: 5
```

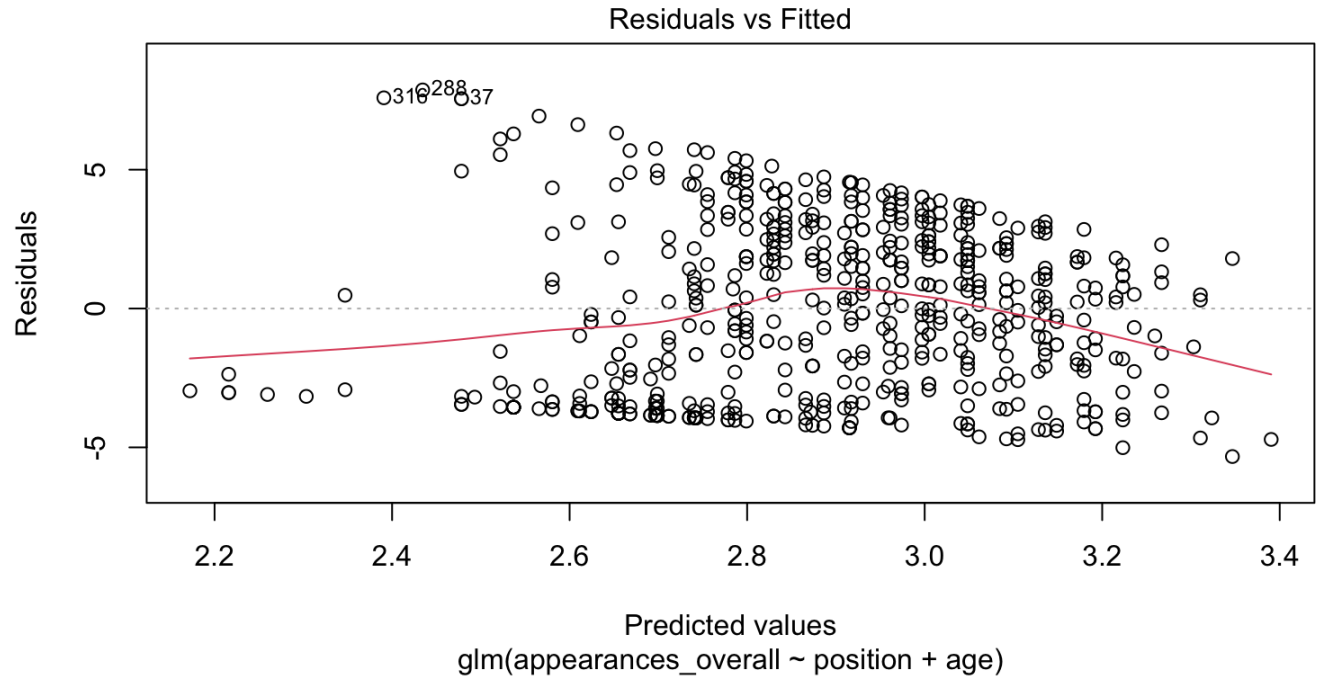
5b.

Check the assumption of the model using a diagnostic plot and comment on your findings. (3 points)

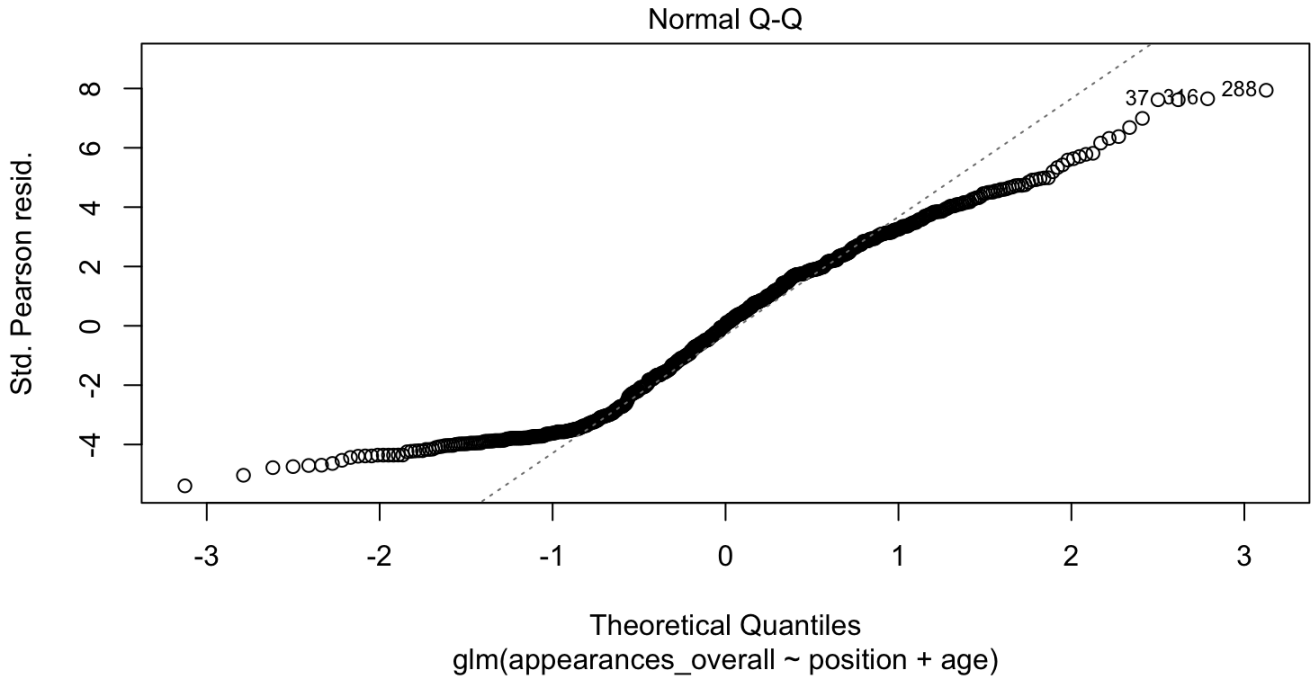
```
plot(appearances_mod, which = 3)
abline(h = 0.8, col = 3)
```




```
plot(appearances_mod, which = 1)
```




```
plot(appearances_mod, which = 2)
```



1. Dispersion

The Scale-Location plot shows that the red line is not flat and it is not close to the green line. The whole red line is above the green line, this means that there is clear overdispersion in the data. This means that all of the important predictors have likely not been accounted for in this model.

2. Linearity

The Residuals vs Fitted plot shows that the red line is not flat and has a peak. It does not follow the black line very well and therefore we cannot assume linearity.

3. Distribution

The Normal Q-Q plot shows us that the middle of the plot is good. the points are along the dotted line. However, the points deviate a lot from the edges so overall the plot is not great.

5c.

What do the coefficients of the model tell us about which position has the most appearances?

The position midfielder has the most appearances because it has the largest positive coefficient (0.118)

How many times more appearances do forwards get on average than goalkeepers? (3 points)

log forward = 0.110606

log goalkeeper = -0.364605

0.110606 - -0.364605 = 0.475211

e^0.475211 = 1.608

Forwards get 1.608 more appearances on average compared to goalkeepers.