Question 1: Regress Score on Calories, Type, and Fat. Write down the fitted model. What is the interpretation of the coefficient of Type in this regression? Is that coefficient statistically significant? Explain.

Answer:
Fitted model: Score hat = -148.8173 + 0.7430 (Calories) + 15.6344 (Type) + -3.8914 (Fat)
Type interpretation: After accounting for all other predictors, for every unit of increase in Type, the Score increases by 15.6344.
This coefficient isn't statistically significant because the p-value of 0.0651 is greater than 0.05.

R code:
imod1 = lm(Score ~ Calories + Type + Fat, data = pizza)
summary(imod1)

R output:

```
Call:
lm(formula = Score ~ Calories + Type + Fat, data = pizza)

Residuals:
   Min     1Q Median     3Q    Max
-40.63  -7.75   3.95  15.29  26.24

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -148.8173    77.9854  -1.908   0.0679 .
Calories       0.7430     0.3066   2.424   0.0229 *
Type          15.6344     8.1033   1.929   0.0651 .
Fat           -3.8914     2.1381  -1.820   0.0807 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 25 degrees of freedom
Multiple R-squared:  0.2873,    Adjusted R-squared:  0.2018
F-statistic:  3.36 on 3 and 25 DF,  p-value: 0.03464
```

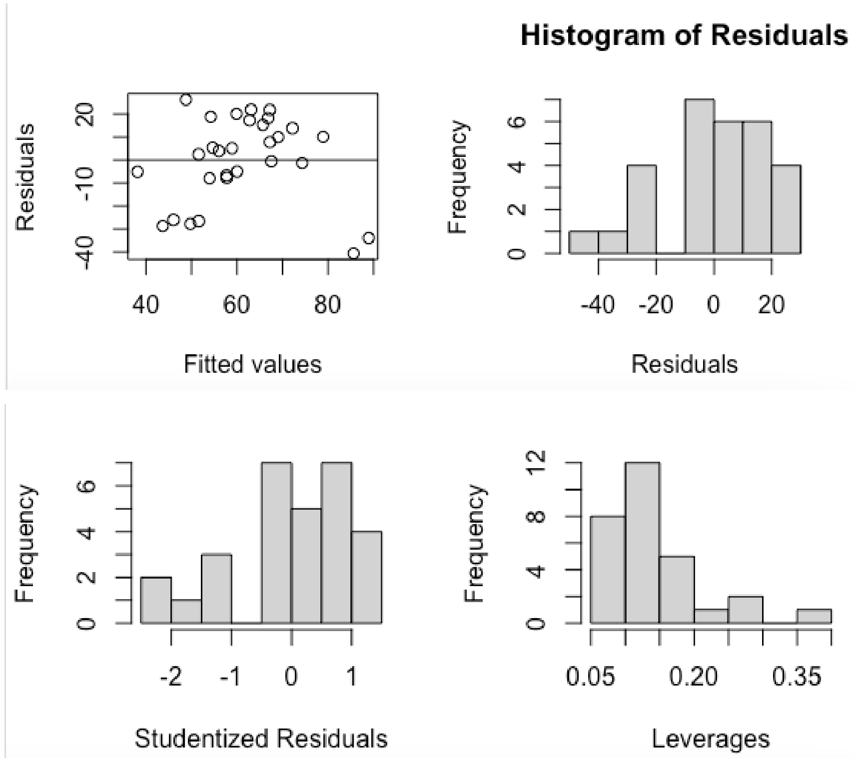
Question 2: Based on the model obtained previously, plot its residuals against predicted values. Which pizza has the most unusual residual? Also, find its standardized residual and leverage.

Answer: Michelina Pizza has the most unusual residual. The standardized residual is -2.282 and leverage is 0.19.

R code:

```
plot(imod1$fitted.values, imod1$residuals, xlab = 'Predicted', ylab = 'Residuals')
abline(0 ,0)
idx = which.max(abs(imod1$residuals))
pizza[idx, ]
rstandard(imod1)[idx]
hatvalues(imod1)[idx]
```

R output:



Question 3: Please use Cook's distances to identify influential cases for the regression model fitted previously. Show their Cook's distances.

Answer:

From the boxplot and the histogram, there are three influential points. They are Healthy Choice pepperoni, Reggio, and Michelina pizzas. Their Cook's distances are 0.454, 0.447 and 0.306.

R code:

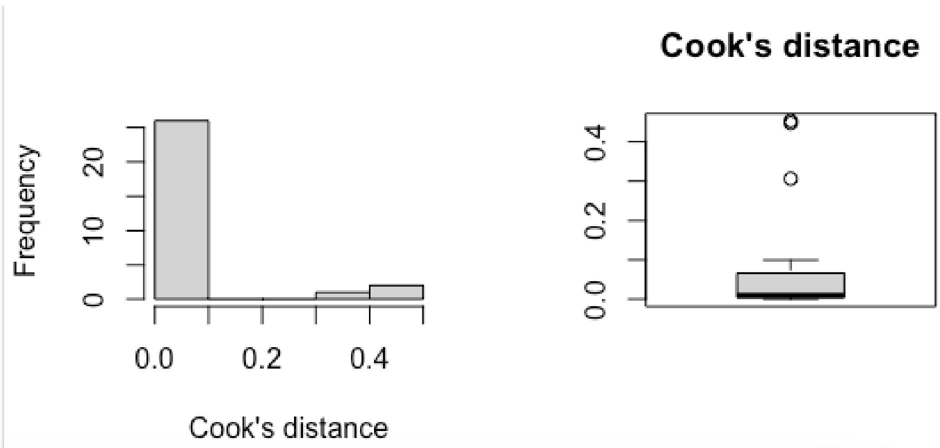
```
par(mfrow = c(1,2))
boxplot(cooks.distance(imod1))
hist(cooks.distance(imod1), main = '', xlab = 'Cooks Distance')
idx.cook1 = order(cooks.distance(imod1), decreasing = TRUE)[1]
 idx.cook2 = order(cooks.distance(imod1), decreasing = TRUE)[2]
```

idx.cook3 = order(cooks.distance(imod1), decreasing = TRUE)[3]
pizza[c(idx.cook1, idx.cook2, idx.cook3),]

R output:

```
##                           Brand Score Cost Calories Fat Type
## 29 Healthy_Choice_pepperoni    15 1.62      280   4    0
## 12                   Reggio    55 1.02      367  13    1
## 16                Michelina    45 1.28      394  19    1
```

```
##        29        12        16
## 0.4536052 0.4471159 0.3059690
```



Question 4: Let's remove all the influential cases from the dataset and refit the multiple regression model in Problem 1. Compare the new model to the old one based on their summaries. Check the assumptions for the new model.

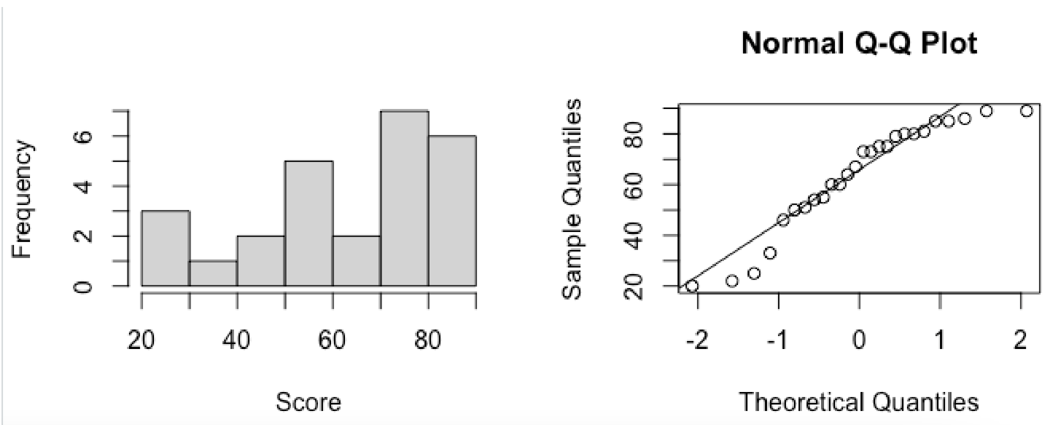
Answer:
The new model seems to be much better than the old one. It has a higher R^2 , a more significant overall F test, and more significant t-tests. From the residual plot, there is a fan shape that is symmetric around the horizontal line at 0. Therefore, the Equal Variance assumption is violated, but the Linearity assumption is satisfied. Based on the histogram and Q-Q plot of residuals, the Normality assumption is also satisfied. Because it is a random sample, the Independence assumption is met as well.

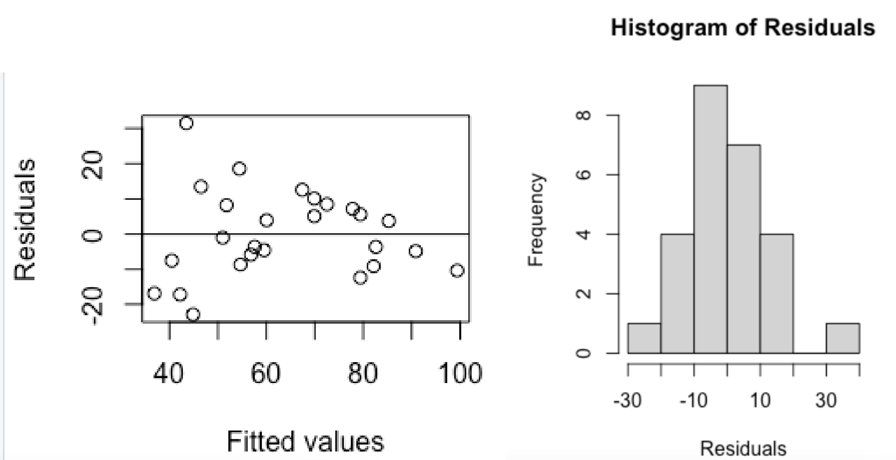
R code:
pizza.new = pizza[-c(idx.cook1, idx.cook2, idx.cook3),]
imod2 = lm(Score ~ Calories + Type + Fat, data = pizza.new)
summary(imod2)
plot(imod2$fitted.values, imod2$residuals, xlab = 'Predicted', ylab = 'Residuals')
abline(0, 0)

```
par(mfrow = c(1, 2))
hist(imod2$residuals, main = "", xlab = 'Residuals')
qqnorm(imod2$residual)
qqline(imod2$residuals)
```

<u>R output:</u>

```
> influen #3 influential points.
       12        16        29
0.4471159 0.3059690 0.4536052
```





```
Call:
lm(formula = Score ~ Calories + Type + Fat, data = new.pizza)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.878  -8.372  -2.298   7.960  31.503 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -351.9436    65.4809  -5.375 2.14e-05 ***
Calories       1.5951     0.2559   6.234 2.84e-06 ***
Type          18.1209     6.3100   2.872  0.00886 ** 
Fat           -9.8278     1.7557  -5.598 1.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.04 on 22 degrees of freedom
Multiple R-squared:  0.6639,	Adjusted R-squared:  0.6181 
F-statistic: 14.48 on 3 and 22 DF,  p-value: 1.99e-05
```

Question 5: Check collinearity for the model fitted in Problem 4. Does there exist any serious collinearity? If does, could you find the reason?

Answer:
Yes, there exists a serious collinearity in the model because the VIF measures for Calories and F at are both greater than 10. It's because those two variables are highly correlated (r = 0.9585)

R code:
library(car)
vif(m2)
cor(new.pizza$Calories, new.pizza$Fat) #highly correlated! value is 0.9585401

R output:

```
> vif(m2)
Calories     Type      Fat 
12.52500  1.48502 12.37517 
```

```
> cor(new.pizza$Calories, new.pizza$Fat) #highly correlated! value is 0.9585401
[1] 0.9585401
```

Question 6: We now use the full dataset pizza. Do we need to consider the interaction between Calories and Type? Explain. Add an interaction term to the model if you think it is necessary, and fit the new model. Interpret the resulting coefficient of the interaction term.

Answer:
New model: yhat = -361.3576 + 1.4056(Calories)-6.5009(Fat) + 288.0736(Type) + 15.6173(Cost) - 0.7806(Calories*Type)

Yes, we need to consider and keep the interaction because it greatly improved R^2 and adjusted R^2, the F-statistic increased, residual standard error decreased, and the interaction term is statistically significant. Therefore, the fitted model is statistically useful.
Interpretation of coefficient of interaction term: The effect of Type on Score depends on Calories, as Type is a dummy variable; when Type is 0, the coefficient for Calories*Type is 0. When Type is 1, then the coefficient for Calories*Type is -0.7806.

R code:
summary(lmList(Score ~ Calories | Type, data = pizza))
imod3 = lm(Score ~ Calories + Type + Fat + Type*Calories , data = pizza) summary(imod3)

R output:
#Without interaction term.

```
> summary(lm(Score ~ Calories+Fat+Type+Cost,data=pizza))

Call:
lm(formula = Score ~ Calories + Fat + Type + Cost, data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.795  -6.791   4.066  16.876  25.684 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -149.5069    79.5528  -1.879   0.0724 .
Calories       0.7635     0.3241   2.356   0.0270 *
Fat           -4.0808     2.3206  -1.759   0.0914 .
Type          15.2647     8.4057   1.816   0.0819 .
Cost          -3.3028    13.8879  -0.238   0.8140  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.17 on 24 degrees of freedom
Multiple R-squared:  0.289,	Adjusted R-squared:  0.1705 
F-statistic: 2.439 on 4 and 24 DF,  p-value: 0.07458
```

#With interaction term

```
> summary(lm(Score ~ Calories+Fat+Type+Cost+Calories*Type, data=pizza))

Call:
lm(formula = Score ~ Calories + Fat + Type + Cost + Calories *
    Type, data = pizza)

Residuals:
    Min      1Q  Median      3Q     Max
-32.518 -14.485   2.827  11.791  22.799

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -361.3576    93.6488  -3.859 0.000799 ***
Calories         1.4056     0.3378   4.161 0.000377 ***
Fat             -6.5009     2.0985  -3.098 0.005074 **
Type           288.0736    84.2085   3.421 0.002337 **
Cost            15.6173    13.1054   1.192 0.245543
Calories:Type   -0.7806     0.2401  -3.251 0.003519 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.06 on 23 degrees of freedom
Multiple R-squared:  0.5129,	Adjusted R-squared:  0.407
F-statistic: 4.843 on 5 and 23 DF,  p-value: 0.003595
```

<u>Appendix</u>

```
m1 <-lm(Score ~ Calories+Type+Fat, data=pizza)
summary(m1)

par(mfrow = c(1,2)) #residual plot
plot(m1$fitted.values, m1$residuals, xlab = 'Fitted values', ylab = 'Residuals')
abline(0, 0)
hist(m1$residuals, main = "Histogram of Residuals", xlab = 'Residuals')
std.res = rstandard(m1)
lev = hatvalues(m1)
par(mfrow = c(1,2))
hist(std.res, xlab = 'Studentized Residuals', main = "")
hist(lev, xlab = 'Leverages', main = "")
pizza[which.max(cooksd),]
cooksd = cooks.distance(m1)
cooksd
par(mfrow = c(1,2))
hist(cooksd, xlab = "Cook's distance", main = '')
boxplot(cooksd, main = "Cook's distance") #There are 3 outliers.
pizza[which.max(cooksd),] #row 29
cooksD <- cooks.distance(m1)
```

```
influen <- cooksD[(cooksD > (3 * mean(cooksD)))]
influen #3 influential points.
names(influen) # 12,16,29
new.pizza <- pizza[-c(12,16,29), ] #remove outlier
new.pizza #removed outliers.

m2 <-lm(Score ~ Calories+Type+Fat, data=new.pizza)
summary(m2)
par(mfrow = c(1, 2))
hist(new.pizza$Score, xlab = "Score", main = "")
qqnorm(new.pizza$Score)
qqline(new.pizza$Score)
plot(m2$fitted.values, m2$residuals, xlab = 'Fitted values', ylab = 'Residuals')
abline(0, 0)
hist(m2$residuals, main = "Histogram of Residuals", xlab = 'Residuals')

library(car)
vif(m2)
cor(new.pizza$Calories, new.pizza$Fat)

summary(lm(Score ~ Calories+Fat+Type+Cost,data=pizza))
summary(lm(Score ~ Calories+Fat+Type+Cost+Calories*Type, data=pizza))
```