Business Analytics

Assignment-1

Name-Subham Kedia
UNI-sk4355

Answer 1

Part a. Running the summary() function shows the summary of our data set.

```
      eggs            feed         temperature
 Min.   :0.000   Min.   :18.36   Min.   :-12.61
 1st Qu.:1.418   1st Qu.:21.50   1st Qu.: 10.71
 Median :1.782   Median :22.27   Median : 21.76
 Mean   :1.773   Mean   :23.11   Mean   : 19.96
 3rd Qu.:2.174   3rd Qu.:23.30   3rd Qu.: 29.63
 Max.   :3.652   Max.   :32.60   Max.   : 48.12
```

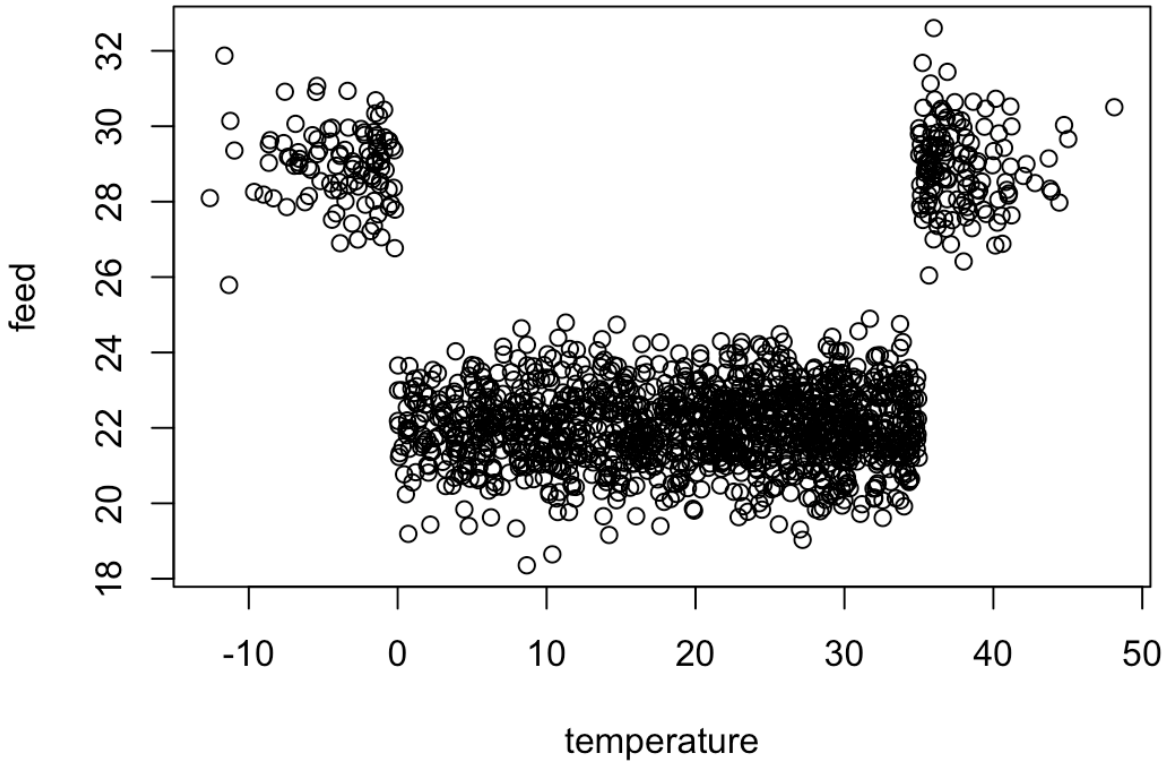
Part b. The feed coefficient is negative and significant (considering its low p- value), which means that the feed negatively affects the eggs. The p-value is low which implies that the model is significant. When we plot the graph between eggs (on the y-axis) and feed (on the y-axis), we find that on increasing feed the eggs decrease which is shown by our model as the feed coefficient is negative. The adjusted R-squared value is low which implies that the fit is not that good.

```
Call:
lm(formula = eggs ~ feed, data = eggData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54185 -0.34831 -0.02782  0.36793  1.81521

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.832768   0.113951   33.63   <2e-16 ***
feed        -0.089108   0.004897  -18.20   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5215 on 1550 degrees of freedom
Multiple R-squared:  0.176,     Adjusted R-squared:  0.1755
F-statistic: 331.1 on 1 and 1550 DF,  p-value: < 2.2e-16
```

Part c. The feed coefficient is negative and significant (considering its low p- value) but the temperature coefficient is negative and insignificant (considering its high p- value), which means that the feed negatively affects the eggs but change in temperature does not have a significant effect on the eggs. The p-value is low which implied that the model is significant. When we plot the graph between eggs (on the y-axis) and temperature (on the y-axis), we find that on increasing temperature the eggs do not show any significant change which is also shown by our model. The adjusted R-squared value is still low which implies that the fit is not that good.

```
Call:
lm(formula = eggs ~ feed + temperature, data = eggData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55172 -0.34901 -0.02884  0.36528  1.81519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.8448807  0.1160307  33.137   <2e-16 ***
feed        -0.0891043  0.0048985 -18.190   <2e-16 ***
temperature -0.0006112  0.0010969  -0.557    0.577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5216 on 1549 degrees of freedom
Multiple R-squared:  0.1762,    Adjusted R-squared:  0.1751
F-statistic: 165.6 on 2 and 1549 DF,  p-value: < 2.2e-16
```

Part d. Running the summary() function shows the summary of our data set.



Part e. One particular set of temperatures (temperatures between 0 and 35) have a positive and significant effect on the eggs because the coefficient of the categorical variable is positive and significant (low p-value as well). The adjusted R-squared value also improved as compared to the previous models and the model gives a prediction of the eggs with lower errors as compared to the previous models.

```
Call:
lm(formula = eggs ~ feed + temperature + TempCat, data = eggData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57159 -0.33950 -0.00575  0.33716  1.74597

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0452405  0.3678233   0.123  0.90213
feed         0.0386355  0.0125835   3.070  0.00218 **
temperature -0.0002184  0.0014323  -0.152  0.87883
TempCat2     0.9998608  0.1071919   9.328  < 2e-16 ***
TempCat3    -0.0472243  0.0884356  -0.534  0.59342
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5027 on 1547 degrees of freedom
Multiple R-squared:  0.2357,    Adjusted R-squared:  0.2337
F-statistic: 119.3 on 4 and 1547 DF,  p-value: < 2.2e-16
```

Part f. The last model is the best for prediction because it takes into account multiple factors and also gives a better R-squared value.

Part g. The confint() function is used with the argument level=0.95. It tells us the value between which our coefficients of our regression model will lie 95% of the times if we run the model again and again (repeatedly).

```
> confint(fit3, level=0.95)
                   2.5 %      97.5 %
(Intercept) -0.67624443 0.766725509
feed         0.01395299 0.063317940
temperature -0.00302783 0.002591056
TempCat2     0.78960407 1.210117545
TempCat3    -0.22069068 0.126241996
```

Part h. The predict() function is used with the arguments, interval="prediction" and level=0.99. It tells us the lower bound and the upper bound of the value of eggs if we predict the eggs only once and not by running the model repeatedly for a specific set of inputs to the model. It does not give us the lower bound and the upper bound of the average value of eggs which is given to us when we specify the argument interval="confidence" in our predict() function.

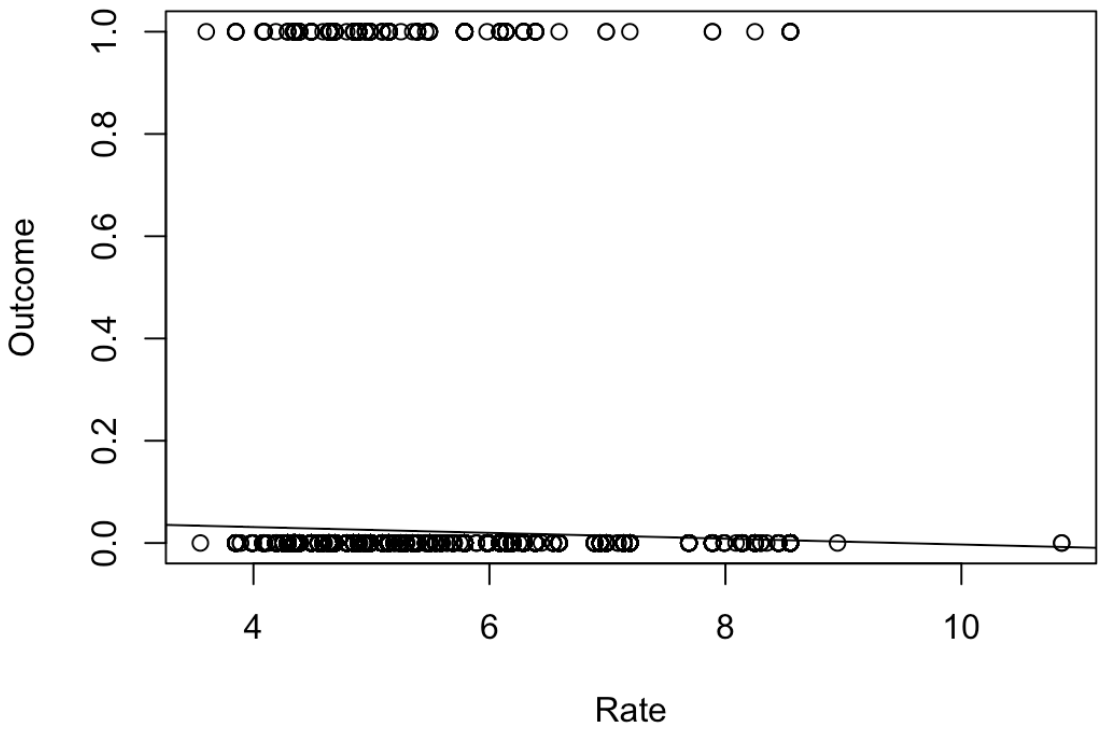
```
> predict(fit3, data.frame(feed=26, temperature=-2, TempCat="1"), interval="prediction", level=0
.99)
       fit        lwr      upr
1 1.050199 -0.2560821 2.356481
```

Part i. In the first regression we only considered the effect of feed on eggs and in the second regression we are considering the effect of temperature as well as feed, taking into account two factors which gives us a better model. In the last model we have taken into account the effect of temperature, feed and segmented temperature into three different bins which gives us a even better model for eggs. Although, there is not much difference in the first and second model because eggs is not significantly dependent on temperature but eggs is actually dependent on feed and feed is in dependent on temperature. Therefore, there is an indirect dependency of eggs on temperature.

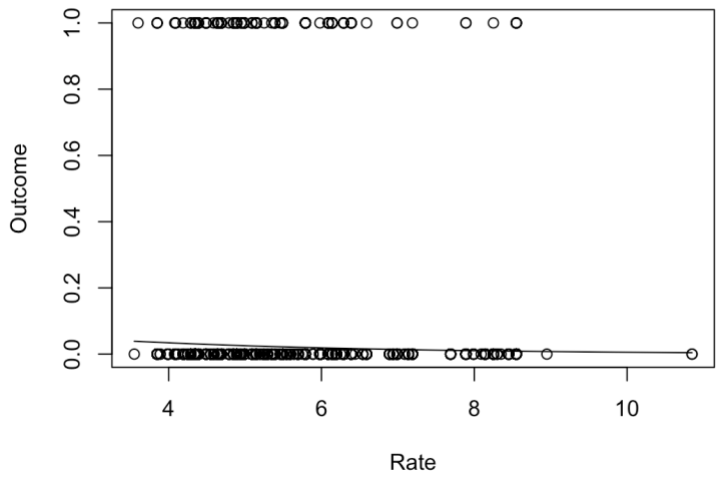
Answer 2.

Part a. The model states that higher the APR offered, lower is the probability of people accepting the APR. The APR has a negative coefficient which means that it has a negative impact on the acceptance which is evident when we a plot a graph between outcome (on the y-axis) and APR (on the x-axis). The p-value for the model is not low and the adjusted R-square is very low which shows that the model is not a very good fit for our data set.

```
Call:
lm(formula = Outcome ~ Rate, data = cardata1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.03401 -0.02657 -0.02485 -0.01947  0.99461

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.054339   0.008717   6.234 4.76e-10 ***
Rate        -0.005725   0.001575  -3.635 0.000279 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1503 on 9237 degrees of freedom
Multiple R-squared:  0.001429,  Adjusted R-squared:  0.001321
F-statistic: 13.22 on 1 and 9237 DF,  p-value: 0.000279
```



Part b. The model is not a good fit for our data set because the p-value for the model is high and the adjusted R-square is very low. The logistic regression is a better model as compared to the linear regression model because the logistic regression suggests that the terminal value of the outcome can be 0 but linear regression suggests that the value of the outcome can also be negative which is not possible according to our given data set. The linear regression and logistic regression has been compared by finding out the mean squared value of both the models individually.



```
> print (mse_fit1)
[1] 0.02259384
> print (mse_fit2)
[1] 0.02258727
```

Part c. The new model takes into account the competitor's rate as well and makes a lot more sense because the outcome will depend on the APR offered by Nomis as well as the APR offered by the competitors of Nomis. The decision or outcome does not only depend on the APR offered by Nomis because if a competitor offers a lower APR than Nomis, then the customer will not accept Nomis' APR and the outcome will therefore be effected. The fit improves if we bring the APR offered by the competitor into our linear model but not to a great extent.

```
Call:
lm(formula = Outcome ~ Rate + Competition_Rate, data = cardata1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.04035 -0.02875 -0.02384 -0.01825  0.99770

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.003190   0.017789   0.179 0.857692
Rate            -0.006271   0.001583  -3.962 7.48e-05 ***
Competition_Rate  0.012011   0.003642   3.298 0.000978 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1502 on 9236 degrees of freedom
Multiple R-squared:  0.002603,  Adjusted R-squared:  0.002387
F-statistic: 12.05 on 2 and 9236 DF,  p-value: 5.917e-06
```

Part d. The mean squared error has been computed by dividing the entire data set into train data and test data. The linear model has been trained using the train data and then the model has been implemented on the test data to predict the Outcome. The difference between the actual Outcome and the predicted Outcome is squared and is divided by the total number of observations to find the mean squared error.

```
> mse=(sum(testdata$pred1))/length(testdata$pred1)
> print(mse)
[1] 0.03342975
> #Part-(d) using the predict() function
> predct = predict(fit4, newdata=testdata)
> y=testdata$Outcome
> mse1=mean((predct-y)**2)
> print(mse1)
[1] 0.03342975
```