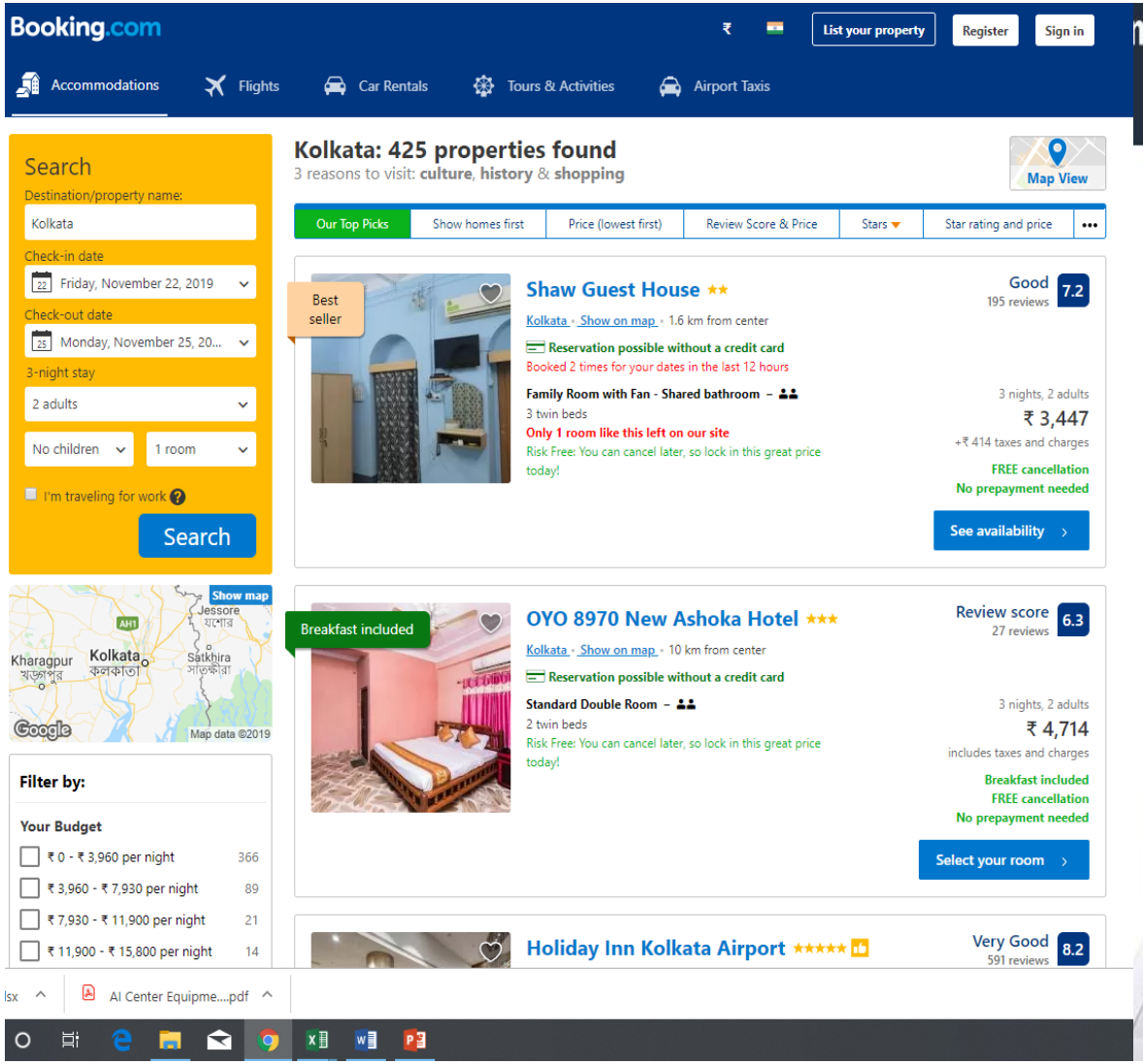
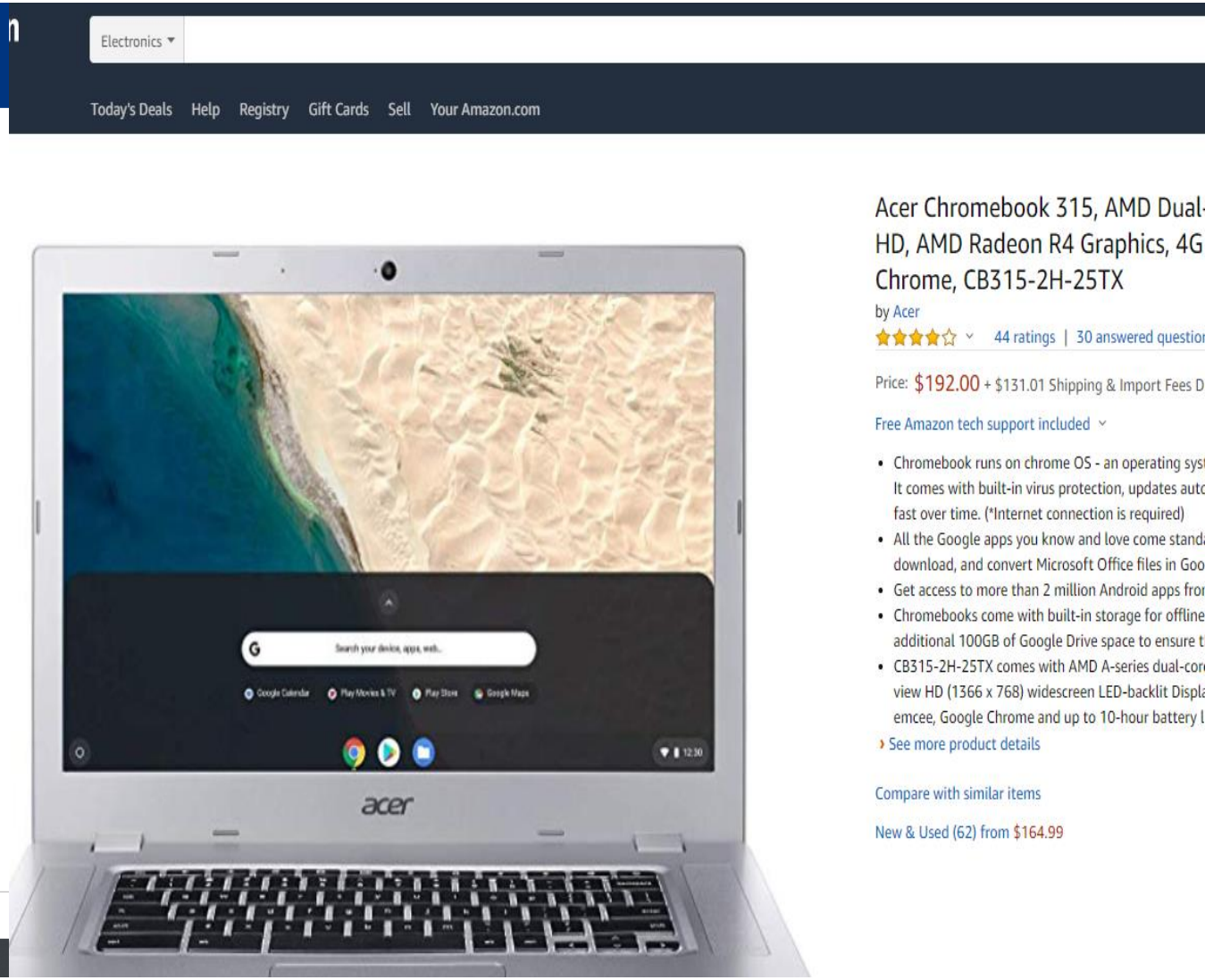
Linear Regression

Machine Learning Unit 4

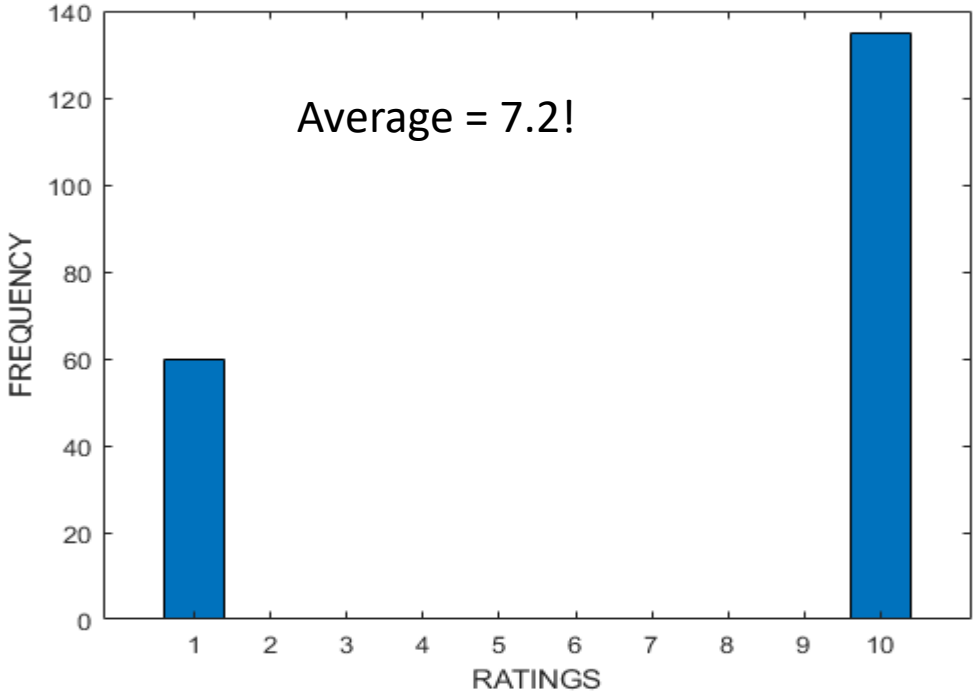
Average Ratings

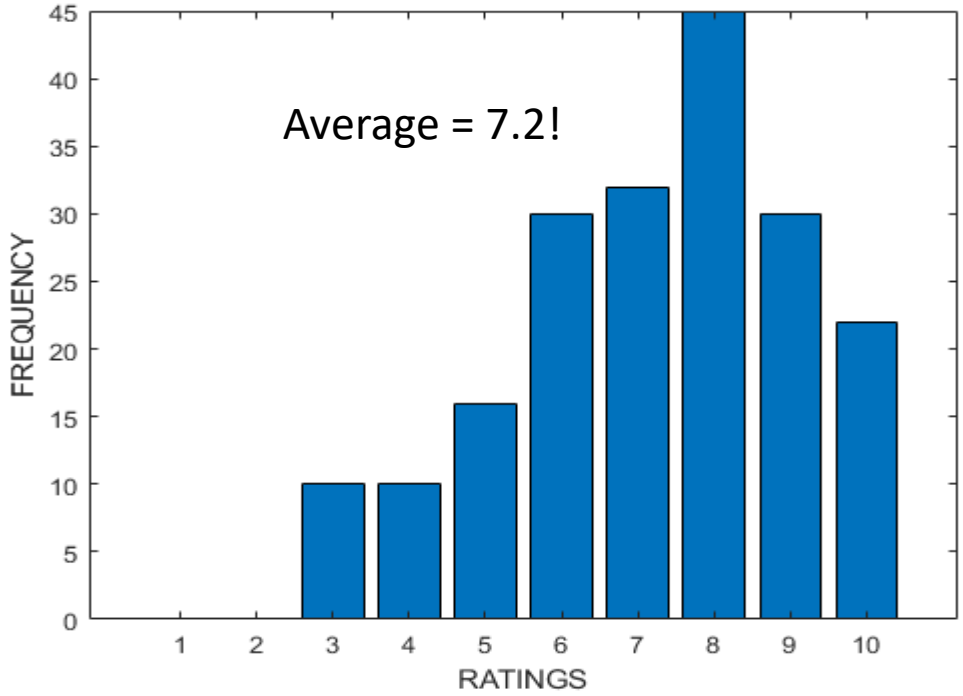




Average Ratings

- 195 reviews, on a scale of 1 to 10
- Average rating: 7.2!
- There may be large or small variance among individual reviews





How do users rate a product?

Title:

Presenter:

Date: Time:

Your job classification: ☐ Classified ☐ Professional/Technical ☐ Administrator ☐ Faculty

Please circle the appropriate response for each statement:

	Excellent	Good	Fair	Poor
1. The relevance of this topic to me was	4	3	2	1
2. The usefulness of materials was	4	3	2	1
3. The effectiveness of the presenter was	4	3	2	1
4. I expect the future usefulness of this topic to be	4	3	2	1
5. My overall evaluation of this session is	4	3	2	1

Your Account > Packaging Feedback

Rate Amazon's Packaging

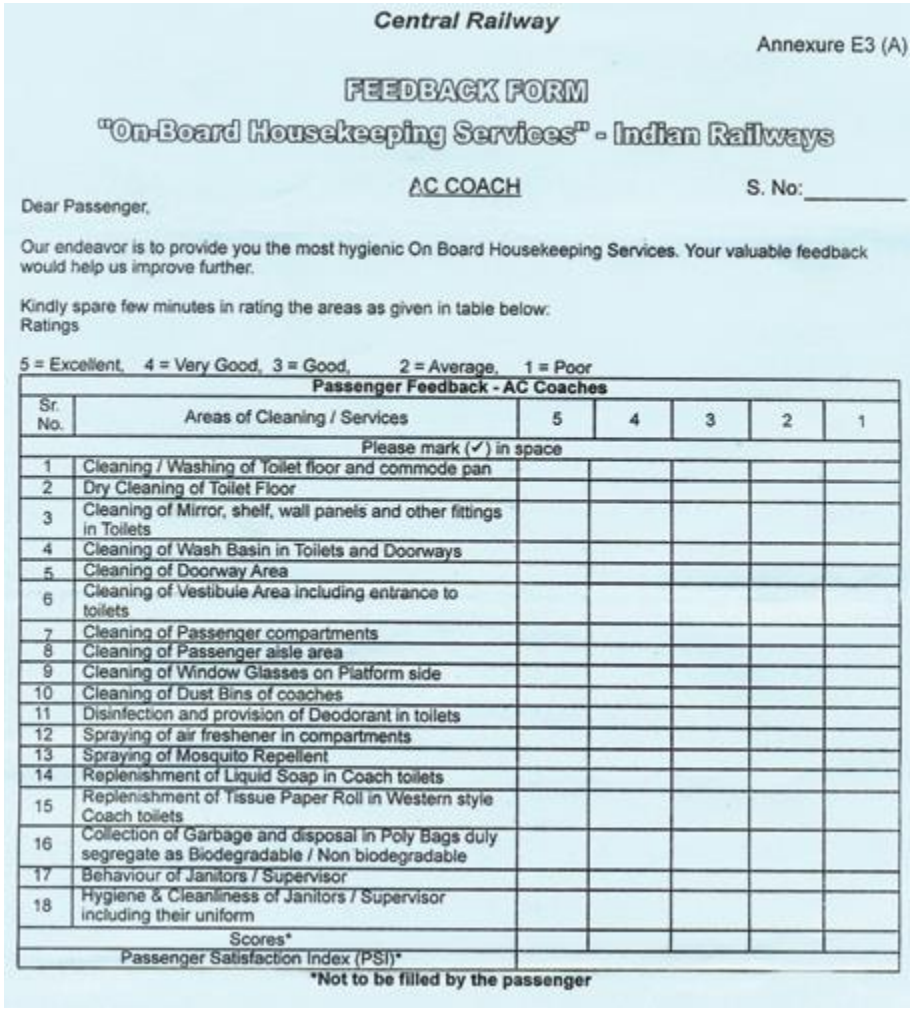
Did the packaging protect your items adequately? ☆☆☆☆☆ Protection — 1 star = Poor; 5 stars = Excellent

Was the box size and packaging appropriate for the items?

- ○ Too Small
- ○ About Right
- ○ Too Big
- ○ Way Too Big

Rate Item's Packaging

[redacted] ☆☆☆☆☆ Ease of Opening — 1 star = Very Difficult; 5 stars = Very Easy



Central Railway

Annexure E3 (A)

FEEDBACK FORM

"On-Board Housekeeping Services" - Indian Railways

AC COACH S. No:_____

Dear Passenger,

Our endeavor is to provide you the most hygienic On Board Housekeeping Services. Your valuable feedback would help us improve further.

Kindly spare few minutes in rating the areas as given in table below:
Ratings

5 = Excellent, 4 = Very Good, 3 = Good, 2 = Average, 1 = Poor

Passenger Feedback - AC Coaches

Sr. No.	Areas of Cleaning / Services	5	4	3	2	1
	Please mark (✓) in space					
1	Cleaning / Washing of Toilet floor and commode pan					
2	Dry Cleaning of Toilet Floor					
3	Cleaning of Mirror, shelf, wall panels and other fittings in Toilets					
4	Cleaning of Wash Basin in Toilets and Doorways					
5	Cleaning of Doorway Area					
6	Cleaning of Vestibule Area including entrance to toilets					
7	Cleaning of Passenger compartments					
8	Cleaning of Passenger aisle area					
9	Cleaning of Window Glasses on Platform side					
10	Cleaning of Dust Bins of coaches					
11	Disinfection and provision of Deodorant in toilets					
12	Spraying of air freshener in compartments					
13	Spraying of Mosquito Repellent					
14	Replenishment of Liquid Soap in Coach toilets					
15	Replenishment of Tissue Paper Roll in Western style Coach toilets					
16	Collection of Garbage and disposal in Poly Bags duly segregate as Biodegradable / Non biodegradable					
17	Behaviour of Janitors / Supervisor					
18	Hygiene & Cleanliness of Janitors / Supervisor including their uniform					
	Scores*					
	Passenger Satisfaction Index (PSI)*					

*Not to be filled by the passenger

How do users rate a product?

User 1:

Title:

Presenter:

Date: Time:

Your job classification: ☐ Classified ☐ Professional/Technical ☐ Administrator ☐ Faculty

Please circle the appropriate response for each statement:

	Excellent	Good	Fair	Poor	
1. The relevance of this topic to me was	4	3	2	1	3
2. The usefulness of materials was	4	3	2	1	3
3. The effectiveness of the presenter was	4	3	2	1	3
4. I expect the future usefulness of this topic to be	4	3	2	1	3
5. My overall evaluation of this session is	4	3	2	1	4

User 2:

Title:

Presenter:

Date: Time:

Your job classification: ☐ Classified ☐ Professional/Technical ☐ Administrator ☐ Faculty

Please circle the appropriate response for each statement:

	Excellent	Good	Fair	Poor	
1. The relevance of this topic to me was	4	3	2	1	3
2. The usefulness of materials was	4	3	2	1	5
3. The effectiveness of the presenter was	4	3	2	1	1
4. I expect the future usefulness of this topic to be	4	3	2	1	4
5. My overall evaluation of this session is	4	3	2	1	4

How do users rate a product?

- Each product has N features (f_1, f_2, …. , f_N)
- The rating "y_i" given by any user "i" may be a weighted average of her scores (x_{i1}, x_{i2}, … , x_{iN}) on the individual features
- The weights (w_{i1}, w_{i2}, … , w_{iN}) may vary from one user to another according to their respective priorities
- Simplest model for user rating: $y_i = \sum_j w_{ij}x_{ij} + b_i$ (b_i: bias)

How do users rate a product?

- Each product has N features (f_1, f_2, …. , f_N)
- The rating "y_i" given by any user "i" may be a weighted average of her scores (x_{i1}, x_{i2}, … , x_{iN}) on the individual features
- The weights (w_{i1}, w_{i2}, … , w_{iN}) may vary from one user to another according to their respective priorities
- Simplest model for user rating: $y_i = \sum_j w_{ij}x_{ij} + b_i$ (b_i: bias)
- Need to estimate the weights "w": M users x N features
- Too many parameters!!

How do users rate a product?

- Each product has N features (f_1, f_2, …. , f_N)
- The rating "y_i" given by any user "i" may be a weighted average of her scores (x_{i1}, x_{i2}, … , x_{iN}) on the individual features
- The weights (w_{i1}, w_{i2}, … , w_{iN}) may vary from one user to another according to their respective priorities
- Simplest model for user rating: $y_i = \sum_j w_{ij}x_{ij} + b_i$ (b_i: bias)
- Need to estimate the weights "w": M users x N features
- Too many parameters!!
- New approximate model: $y_i = \sum_j w_j x_{ij} + b$, i.e. all users have equal weights!

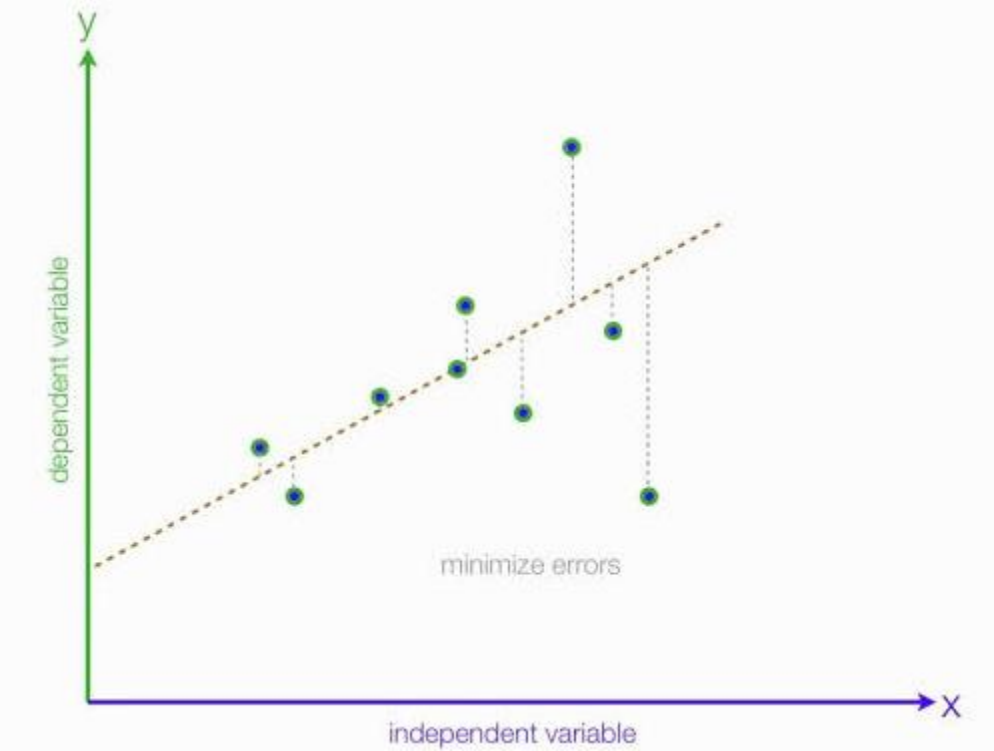
Linear Regression

- We know the feature scores "s_{ij}" and the final score "x_i"
- We want to find out the relative importance of the different features (on average)
- The answer: linear regression!
- General Recipe:

1) Define a model with parameters (w, b)
2) Define a measure on how well the model can fit the final scores
3) Choose the model parameters to improve this measure!

Linear Regression

- The model in this case: $h_i = \sum_j w_j x_{ij} + b$ (h_i: predicted rating)
- Measurement of fit: squared error loss function
- $L(y_i, h_i) = (y_i - h_i)^2 = \sum_j (y_i - w_j x_{ij} - b)^2$



Linear Regression

- The model in this case: $h_i = \sum_j w_j x_{ij} + b$ (h_i: predicted rating)
- Measurement of fit: squared error loss function
- $L(y_i, h_i) = (y_i - h_i)^2 = \sum_j (y_i - w_j x_{ij} - b)^2$

- Choose w, b to minimize total loss $\sum_i L(y_i, h_i)$ over all M users!
- Differentiate the total loss w.r.t. each variable, equate to 0, and solve an equation!

Linear Regression in one dimension

First, let us consider each product has only one feature

$$\frac{dL}{dw} = 0 \implies 2\sum_i (y_i - wx_i - b)x_i = 0$$

$$\frac{dL}{db} = 0 \implies 2\sum_i (y_i - wx_i - b) = 0$$

Solving these equations, we get

$$b = \bar{y} - w\bar{x}$$

$$w = \left(\sum_i (\tilde{x}_i)^2\right)^{-1} \left(\sum_i \tilde{x}_i \tilde{y}_i\right)$$

where $\bar{x} = \frac{1}{N}\sum_i x_i$, $\bar{y} = \frac{1}{N}\sum_i y_i$, $\tilde{x}_i = x_i - \bar{x}$

In [3]:

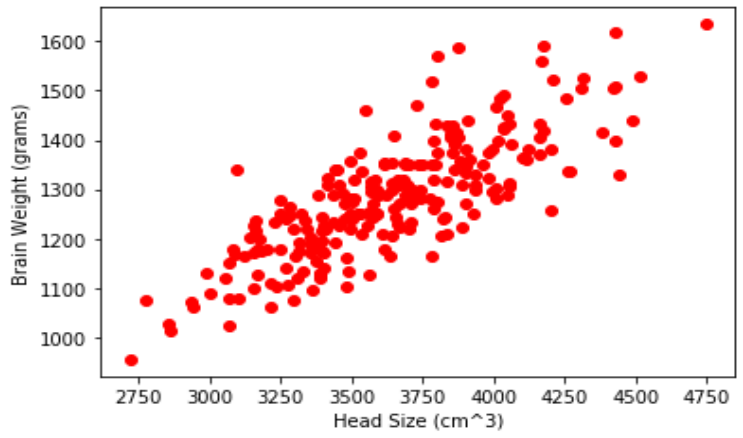
```
#initializing our inputs and outputs


#mean of our inputs and outputs
x_mean = np.mean(X)
y_mean = np.mean(Y)

#total number of values
n = len(X)

#using the formula to calculate the b1 and b0
numerator = 0
denominator = 0
for i in range(n):
    numerator += (X[i] - x_mean) * (Y[i] - y_mean)
    denominator += (X[i] - x_mean) ** 2

b1 = numerator / denominator
b0 = y_mean - (b1 * x_mean)

#printing the coefficient
print(b1, b0)
```

Python Implementation

In [2]:
```
#import libraries
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

#reading data
dataset = pd.read_csv('dataset.csv')
print(dataset.shape)
dataset.head()

X = dataset['Head Size(cm^3)'].values
Y = dataset['Brain Weight(grams)'].values

#plot the data point
plt.scatter(X, Y, color='#ff0000', label='Data Point')

# x-axis label
plt.xlabel('Head Size (cm^3)')

#y-axis label
plt.ylabel('Brain Weight (grams)')
```

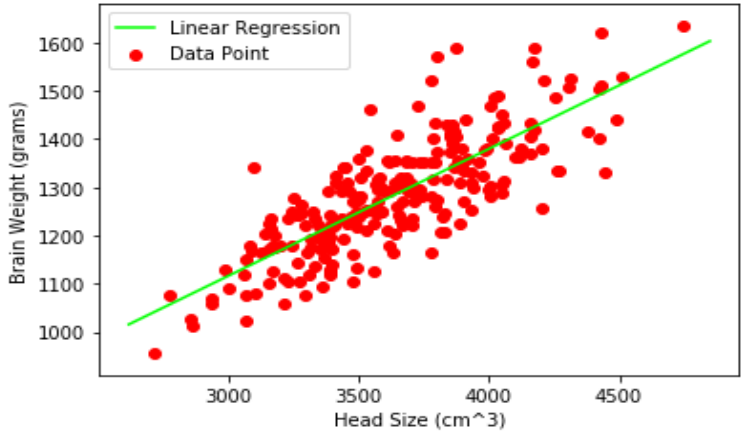
```
(237, 4)
```

Out[2]:
```
Text(0, 0.5, 'Brain Weight (grams)')
```




```
#mean of our inputs and outputs
x_mean = np.mean(X)
y_mean = np.mean(Y)

#total number of values
n = len(X)

#using the formula to calculate the b1 and b0
numerator = 0
denominator = 0
for i in range(n):
    numerator += (X[i] - x_mean) * (Y[i] - y_mean)
    denominator += (X[i] - x_mean) ** 2

b1 = numerator / denominator
b0 = y_mean - (b1 * x_mean)

#printing the coefficient
print(b1, b0)
```

In [3]:
```
#plotting values
x_max = np.max(X) + 100
x_min = np.min(X) - 100

#calculating line values of x and y
x = np.linspace(x_min, x_max, 1000)
y = b0 + b1 * x

plt.plot(x, y, color='#00ff00', label='Linear Regression') #plotting line
plt.scatter(X, Y, color='#ff0000', label='Data Point') #plot the data point
plt.xlabel('Head Size (cm^3)') # x-axis label
plt.ylabel('Brain Weight (grams)') #y-axis label

plt.legend()
plt.show()
```



Linear Regression: Vector Form

Now consider each product has D features
The feature weights $w = \{w_1, \ldots, w_D\}$ arranged as a D-dimensional vector
Each user's feature ratings $x_i = \{x_{i1}, \ldots, x_{iD}\}$ arranged as a D-dimensional vector
Predicted rating by user i: $h_i = w^T x_i + b$
Loss function $L(w, b) = \sum_i (y_i - w^T x_i - b)^2$

Linear Regression: Vector Form

$$\frac{dL}{dw} = 0 \implies 2\sum_i (y_i - w^T x_i - b)x_i = 0$$
$$\frac{dL}{db} = 0 \implies 2\sum_i (y_i - w^T x_i - b) = 0$$

Solving these equations, we get

$$b = \bar{y} - w^T \bar{x}$$
$$w = (\sum_i (\tilde{x}_i \tilde{x}_i{}^T))^{-1} (\sum_i \tilde{x}_i \tilde{y}_i)$$

where $\bar{x} = \frac{1}{N}\sum_i x_i$, $\bar{y} = \frac{1}{N}\sum_i y_i$, $\tilde{x}_i = x_i - \bar{x}$

Here, all the additions are vector additions

Average Rating Prediction

- Given a new product, we need to predict it's "average rating"
- Average rating = $mean_i(y_i)$
- According to LR model:
- predicted average rating = $mean_i(h_i)$
- $= mean_i(\sum_j w_j x_{ij} + b) = \sum_j w_j \, mean_i(x_{ij}) + b$
- We have the weights "w_j" of its features and bias "b", by linear regression for similar products
- We can find the average user ratings of each feature $mean_i(x_{ij})$, based on other products having same feature

Average Rating Prediction

- New Product: a new camera model
- Features: resolution, battery life, memory, flash, weight, size
- Weights of features: calculate by linear regression from user ratings on other cameras
- New camera resolution: 5 MP
- Average rating on resolution: 4.0
- Weight of resolution: 0.54

Model	Resolution	Mean feature rating
Camera1	5 MP	4.1
Camera2	5 MP	3.9
Camera3	10 MP	4.4
Camera4	12 MP	4.1
Camera5	6 MP	4.0
Camera6	15 MP	4.3

Average Rating Prediction

- New Product: a new camera model
- Features: resolution, battery life, memory, flash, weight, size
- Weights of features: calculate by linear regression from user ratings on other cameras
- New camera battery life: 2 years
- Average rating on battery life : 3.8
- Weight of battery life : 0.36

Model	Battery Life	Mean feature rating
Camera1	3 years	4.5
Camera2	2 years	3.6
Camera3	2 years	3.8
Camera4	1 year	3.1
Camera5	2 years	3.9
Camera6	3 years	4.3

Average Rating Prediction

- New Product: a new camera model
- Features: resolution, battery life, memory, flash, weight, size
- Weights of features: calculate by linear regression from user ratings on other cameras
- New camera memory: 5 GB
- Average rating on memory: 4.5
- Weight of memory: 0.10

- Predicted average rating

= 0.54*4.0 + 0.36*3.8 + 0.1*4.5 = 4.0!

Model	Memory	Mean feature rating
Camera1	1 GB	3.8
Camera2	1 GB	3.9
Camera3	2 GB	4.1
Camera4	3 GB	4.0
Camera5	5 GB	4.4
Camera6	5 GB	4.5