Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

1) Set up data as defined in problem:

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/RestaruantTips.csv")
> tiper=100*mydata$Tip/mydata$Bill
> newdata=subset(mydata,tiper<40)
> newtiper=100*newdata$Tip/newdata$Bill
```

The code above shows we started with 157 rows of data, and when we delete the two largest tippers our new data set has 155 rows of data. Note that we have to create a new variable for

a) Using the box plot rule, how many Tip values are considered outliers (use the original data set).

➔ Using the boxplot rule, **9** tip values are considered outliers.

```
> length(boxplot.stats(mydata$Tip)$out)
[1] 9
```

b) Using the box plot rule, how many tiper (tip percentage) values are considered outliers (use the original data set).

➔ Using the boxplot rule, **8** tiper values are considered outliers.

```
> length(boxplot.stats(tiper)$out)
[1] 8
```

c) Using the original data set, what is the correlation between dinner bill and tip?

➔ The correlation between dinner bill and tip is **0.9150592** indicating an **extremely strong** linear correlation.

```
> cor(mydata$Bill, mydata$Tip)
[1] 0.9150592
```

d) Using the data set with the two largest tip percentages removed, what is the correlation between dinner bill and tip? Is this number the same as from part (c)? Explain.

➔ Using the data set with the two largest tip percentages removed, the correlation between dinner bill and tip is **0.9462058** indicating an **extremely strong** linear correlation.

```
> cor(newdata$Bill,newdata$Tip);
[1] 0.9462058
```

The correlation between these two variables is now higher. This behavior is expected since we removed the outliers. **Outliers lie away from the best-fit line so they weaken the correlation**. On removing the outliers, the subset of points now lie closer to the best fit line and hence have a higher correlation.

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

2) Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction and form.

a) T-shirts at a store: price each, number sold.

➔ Explanatory variable: number sold
Response variable: price each
Reason: The price of each t-shirt depends on the number of t-shirts sold.
Scatterplot: A moderate to substantial inverse (downward sloping) correlation.
Reason: As the number of t-shirts sold go up, the price of each t-shirt usually go down.

b) Real estate: house price, house size (square footage).

➔ Explanatory variable: house size
Response variable: house price
Reason: The price of a house depends on the size of the house.
Scatterplot: A medium to strong direct (upward sloping) correlation.
Reason: Keeping all other variables constant, the price of a house typically goes up with the size of the house.

c) Economics: Interest rates, number of mortgage applications.

➔ Explanatory variable: number of mortgage applications
Response variable: interest rates
Reason: The interest rates are dependent on the number of mortgage applications.
Scatterplot: A moderate to substantial direct (upward sloping) correlation.
Reason: Keeping all other variables constant, the interest rates typically go up as the number of mortgage applications go up.

d) Employees: Salary, years of experience.

➔ Explanatory variable: years of experience
Response variable: salary
Reason: The salary is dependent on the number of years of experience.
Scatterplot: A moderate to substantial direct (upward sloping) correlation.
Reason: Keeping all other variables constant, the salary typically goes up as the number of years of experience goes up.

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

3) This question moves us in the direction of understanding that just because two variables are uncorrelated does not mean they are independent.

a) Explain in words what a correlation of 0 implies.

➔ A correlation of 0 implies the **absence of a linear relationship** between the two variables. There still may be another relationship present.

b) Load the blas data set into R and find the correlation of X and Y

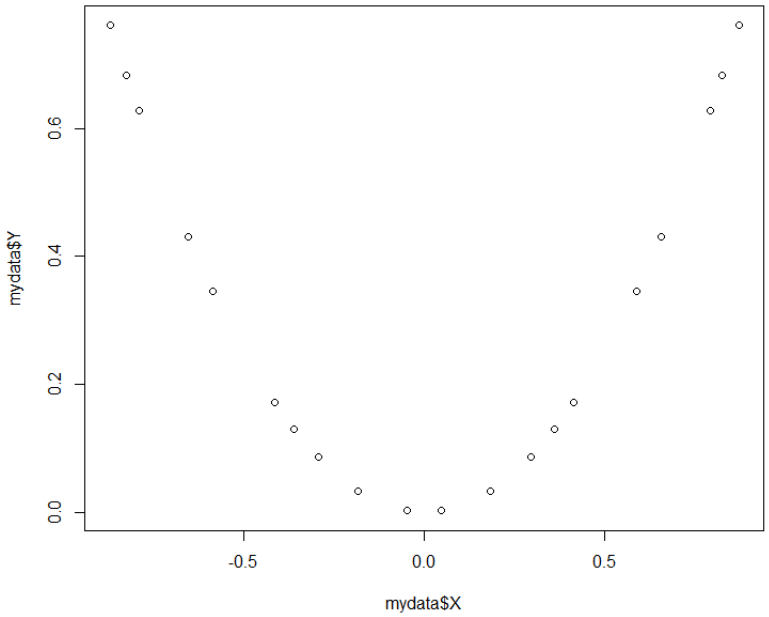
```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/blas.csv")
```

➔
```
> # Load the blas data set into R
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/blas.csv")
> # Find the correlation of  X and Y
> cor(mydata$X, mydata$Y)
[1] 1.041004e-20
```

The correlation can be rounded to **0**.

c) Plot the data-does it agree with your definition?

➔
```
> plot(mydata$X, mydata$Y)
```



We can see that there is a **parabolic relation** between X and Y. This agrees with our definition that the **relationship is not linear but another relation might exist**.

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

4) It has been noted that there is a positive correlation between the U.S. economy and the height of women's hemlines (distance from the floor of the bottom of a skirt or dress) with shorter skirts corresponding to economic growth and lower hemlines to periods of economic recession. Comment on the conclusion that economic factors cause hemlines to rise and fall.(for historical references see for example http://www.edelmanfinancial.com/education-center/articles/t/the-relationship-between-hemlines-and-the-stock-market)

   ➔ **Correlation measures association, not causation.** Correlation does not imply causation so we cannot conclude that economic factors cause hemlines to rise and fall just because there is an association between the two.

Karan A. Bhandarkar

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

5) We have state by state data (plus Washington, DC) on percentage of residents over the age of 25 who have at least a bachelor's degree and median salary. Load this data into R with the command
`mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/bach.csv")`

   a) What is the correlation between these two variables?

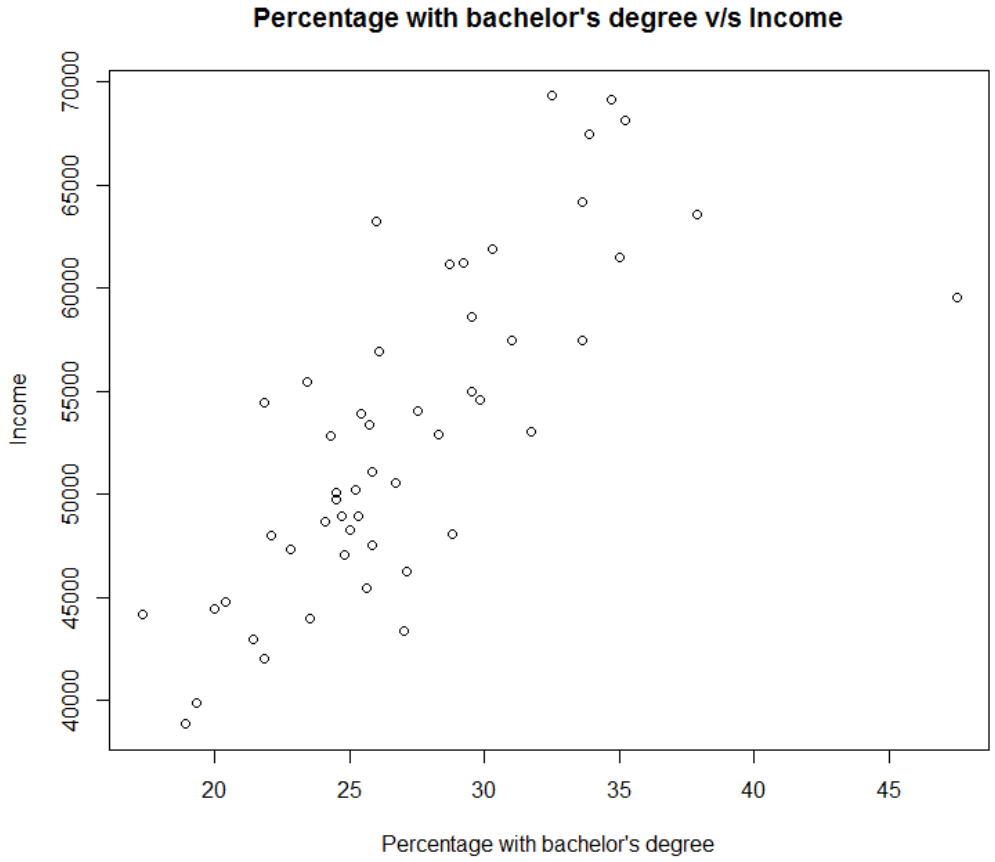
➔ The correlation between these two variables is **0.754167**

```
> cor(mydata$bach, mydata$income)
[1] 0.754167
```

   b) Produce a scatter plot of the data with percentage with bachelor's degree on the X axis. Notice the outlier? Who does that point belong to? Can you think of any reasons why this location might have a high percentage of residents with a bachelor's degree but a lower than expected median income?

➔

```
> # Produce a scatter plot of the data with percentage with bachelor's degree
on the X axis.
> plot(mydata$bach, mydata$income, main = 'Percentage with bachelor\'s degree
v/s Income', + xlab = 'Percentage with bachelor\'s degree', ylab = 'Income')
```



➔ The outlier belongs to **District of Columbia.** One major reason could be that most of the people work in the federal government where incomes are not competitive.

```
> # Pick out the outlier
> outlier = subset(mydata, mydata$bach > 45)
```

c) Remove the outlier point found in (b) and recalculate the correlation. How do the two correlation values compare? What does this illustrate about correlation?

➔ The **new correlation is higher** than the previous correlation.

```
> # Remove the outlier
> newdata = subset(mydata, mydata$bach < 45)
> # Calculate the new correlation
> cor(newdata$bach, newdata$income)
[1] 0.8205775
```

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

6) Fill in the blanks (show your work).

```
> describe(cbind(x1,x2,x3),skew=FALSE)
   vars  n    mean      sd    min   max   range     se
x1    1 74 6165.26 2949.50 3291.0 15906 12615.0 342.87
x2    2 74   39.65    4.40   31.0    51    20.0   0.51
x3    3 74    2.99    0.85    1.5     5     3.5   0.10


> cor(cbind(x1,x2,x3))
          x1        x2        x3
x1 1.0000000 0.3096174 0.1145056
x2 0.3096174 1.0000000 0.4244646
x3 0.1145056  blank 1  1.0000000


> cov(cbind(x1,x2,x3))
             x1          x2          x3
x1    Blank 2   4017.557201 285.7209367
x2    4017.5572   19.354313   1.5797853
x3     285.7209    1.579785   0.7157071
```

Blank 1 = **0.4244646**

➔ Correlation of X with Y is the same as correlation of Y with X
Blank 1 is COR(x3, x2) which is COR(x2,x3)

Blank 2 = **8699550.25**

➔ COV(X,X) = VAR(X)
Blank 2 corresponds to COV(x1,x1) which is VAR(x1)
From the describe function, SD(x1) = 2949.50
Therefore, COV(x1,x1) = VAR(x1) = $[SD(x1)]^2 = (2949.50)^2 = 8699550.25$

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

7) Set up data as defined in problem:

library(quantmod)
source("http://people.fas.harvard.edu/~mparzen/stat104/getstockdata1.R")

a) What company does each symbol represent? Go to finance.yahoo.com to find out. While you're on the yahoo finance page, also write down the Beta yahoo has for each stock.

➔ MRK: Merck & Co., Inc. (β = 0.94)
NKE: Nike Inc (β = 0.41)
SPY: SPDR S&P 500 ETF Trust (β = 1)
YUM: YUM! Brands, Inc. (β = 0.69)

b) Find the Beta for each stock. That is run a regression of each stock return as the Y variable and index returns as the X variable. Beta is the slope from this regression. How do your calculated betas compare to the reported Betas from yahoo finance?

➔ The calculated betas are almost the same as betas from yahoo finance. This probably because of a time period difference of maybe a month.

	Calculated Beta	Yahoo Finance Beta
MRK	0.938397	0.94
NKE	0.426591	0.41
SPY	1	1
YUM	0.660682	0.69

```
> fit = lm(mrkret~spyret)
> coef(fit)
 (Intercept)       spyret
-0.001119201  0.938396637
> fit = lm(nkeret~spyret)
> coef(fit)
(Intercept)      spyret
0.007400234 0.426590705
> fit = lm(spyret~spyret)
Warning messages:
1: In model.matrix.default(mt, mf, contrasts) :
  the response appeared on the right-hand side and was dropped
2: In model.matrix.default(mt, mf, contrasts) :
  problem with term 1 in model.matrix: no columns are assigned
> coef(fit)
(Intercept)
0.008074977
> fit = lm(yumret~spyret)
> coef(fit)
(Intercept)      spyret
0.008366875 0.660681751
```

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

c) What is the standard deviation for the returns of the stocks (just do sd(mkret)for example) ? Rank them from lowest to highest standard deviation.

➔ sd(spyret) < sd(mrkret) < sd(yumret) < sd(nkeret)

```
> sd(mrkret)
[1] 0.0447971
> sd(nkeret)
[1] 0.05410423
> sd(spyret)
[1] 0.02886557
> sd(yumret)
[1] 0.05141241
```

d) Rank the stocks based on their Beta values (smallest to largest). Is the order the same as if you ranked them on their standard deviations from smallest to largest?[there are many risk measures wall street uses so no reason why one ranking is the same as another.]

➔ beta(NKE) < beta(YUM) < beta(MRK) < beta(SPY)
No, the order is not the same as if ranked on their standard deviation.

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

8) We have data on frozen pizza sales (in pounds) and average price ($/unit) from Dallas Texas for 39 recent weeks. Load the class survey data into R using the command
`mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/pizzasales1.csv")`

a) Using price as the explanatory variable and sales as the response variable, run a regression and write down the linear equation relating sales to price from the output.

➔ $y = 141865.53 - 24369.49x$
Where y – frozen pizza sales
x – average price

```
> fit = lm(mydata$sales~mydata$price)
> coef(fit)
 (Intercept) mydata$price
   141865.53    -24369.49
```

b) What does the slope mean in this context?

➔ The slope is the drop in frozen pizza sales for every dollar increase in price.

c) What does the y-intercept mean in this context? Is it meaningful?

➔ The y-intercept is the frozen pizza sales at the price of $0. It is not meaningful since it's outside the scope of the samples.

d) What do you predict the sales to be if the average price charged was $3.50 for a pizza?

➔ Using the equation from a)
$y = 141865.53 - 24369.49 * 3.50 = 56572.31$

e) If the sales for a price of $3.50 turned out to be 60,000 pounds, what would the residual be?

➔ Residual = Observed Value – Predicted Value
= 60000-56572.31 = 3427.685

f) Show that the slope coefficient for the regression model can also be calculated using the equation $b_1 = r\frac{s_y}{s_x}$

➔ Calculate the value of b_1

```
> cor(mydata$price, mydata$sales) * sd(mydata$sales)/sd(mydata$price)
[1] -24369.49
```

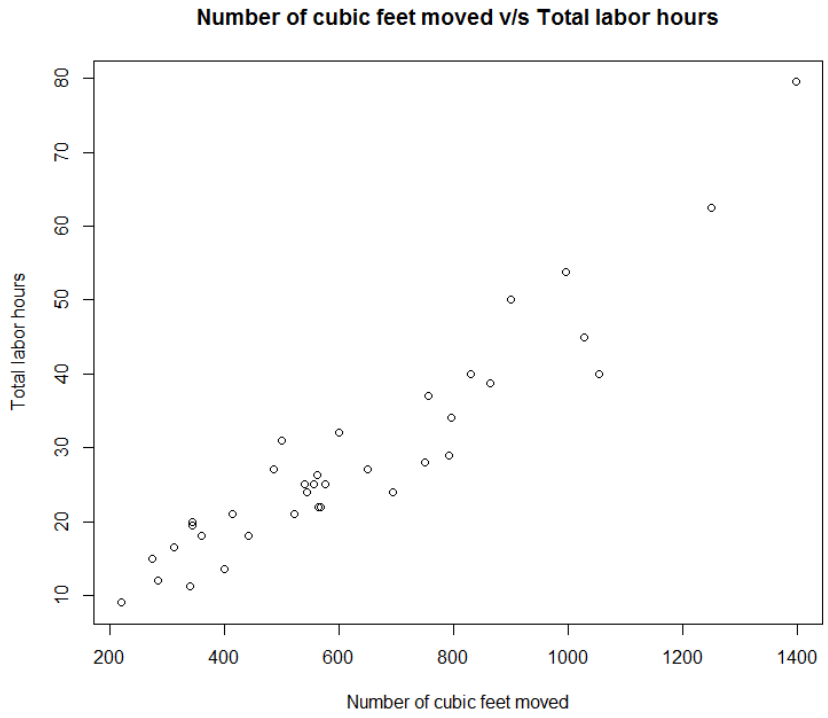
Slope coefficient calculated using the equation $b_1 = r\frac{s_y}{s_x}$ matches our earlier calculation.

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

9) The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours(Y). In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as the independent variable (X) and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the hours worked. The data may be loaded into R as follows
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/moving.csv")
Use R to answer the questions below.

a) Create a scatter diagram of the data.

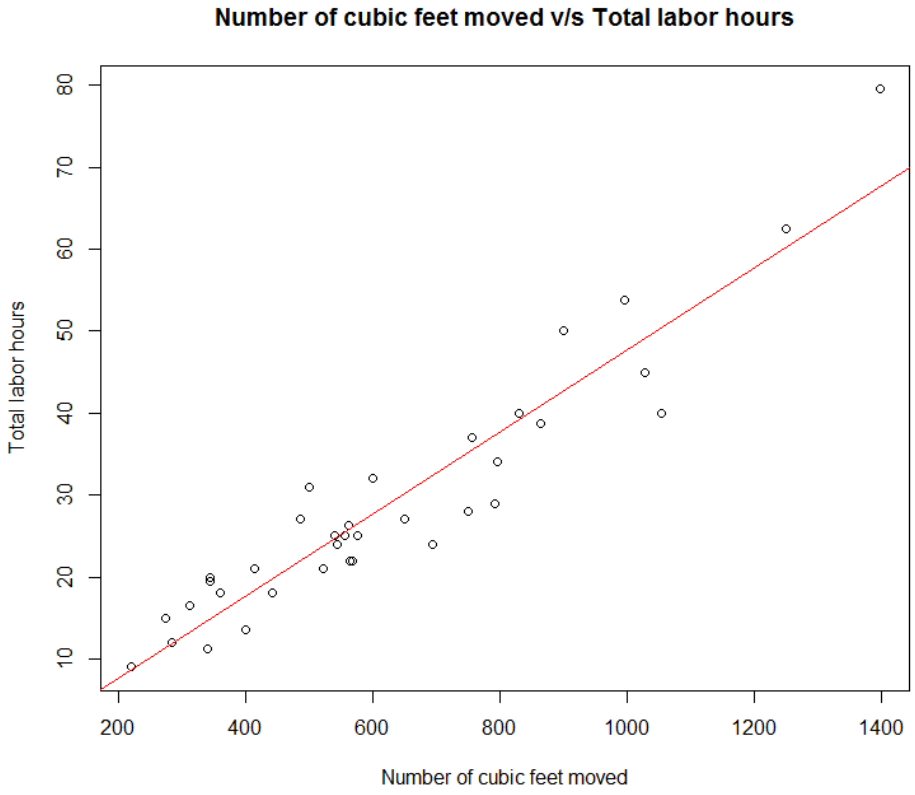
```
> plot(mydata$feet, mydata$hours, main = "Number of cubic feet moved v/s Total l
abor hours", xlab = 'Number of cubic feet moved', ylab = 'Total labor hours')
```



b) Fit a least squares regression line to this data and interpret the slope.

```
> fit = lm(mydata$hours~mydata$feet)
> abline(fit, col = 'red')
```

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics



c) Predict the labor hours for a 500 cubic feet move using the estimated regression equation developed in part (b).

➔ A 500 cubic feet move is estimated to take 22.67 labor hours.

```
> coef(fit)
(Intercept) mydata$feet
-2.36966013  0.05008027
```

$y = -2.36966013 + 0.05008027x = -2.36966013 + 0.05008027 * 500 = \mathbf{22.67047}$

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

10) A fair six-sided die is rolled.

a) What are the possible outcomes of this event?

➔ Possible outcomes = {1,2,3,4,5,6}

b) Calculate the probability of rolling a prime number.

➔ Success = {2,3,5}
Probability = $\frac{\{2,3,5\}}{\{1,2,3,4,5,6\}} = \frac{3}{6} = 0.50$

c) Calculate the probability of rolling an even number.

➔ Success = {2,4,6}
Probability = $\frac{\{2,4,6\}}{\{1,2,3,4,5,6\}} = \frac{3}{6} = 0.50$

d) What is the probability of rolling a number greater than seven?

➔ 7 is not a possible outcome so the probability is 0

11) The probability that a driver is speeding on a stretch of road is 0.27. What is the probability that a driver is not speeding?

➔ Probability of a complement event = 1- probability of event
probability that a driver is not speeding = 1- probability that a driver is speeding
probability that a driver is not speeding = 1-0.27 = **0.73**

Karan A. Bhandarkar

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

12) A department store manager has monitored the numbers of complaints received per week about poor service. The probabilities for numbers of complaints in a week, established by this review, are shown in the table. Let *A*be the event "There will be at least one complaint in a week," and B the event "There will be less than 10 complaints in a week."

NUMBER OF COMPLAINTS	0	1-3	4-6	7-9	10-12	More than 12
PROBABILITY	.15	.29	.16	?	.14	.06

a) Find the value of ?

➔ Total probability must be equal to 1
? = 1 – (0.15 + 0.29 + 0.16 + 0.14 + 0.06)
? = 0.2

b) Find the probability of *A.*

➔ P(A) = P(at least one complaint) = 1 – P(no complaints) = 1-0.15 = **0.85**

c) Find the probability of *B.*

➔ P(B) = P(less than 10 complaints) = 1-P(10 or more complaints) = 1-(0.14+0.06) = **0.8**

d) Find the probability of the complement of *A.*

➔ P(A') = 1-P(A) = 1-0.85 = **0.15**

e) Find the probability of *A* or *B.*

➔ P(A or B) = P(A) + P(B) - P(A and B) = 0.85 + 0.8 – (0.29+0.6+0.2) = **0.56**

f) Find the probability of *A* and *B.*

➔ P(A and B) = P(at least one AND less than 10 complaints) = 0.29+0.16+0.20 = **0.65**

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

13) Answer the following questions using the following joint probability table

	No wind	Some wind	Strong wind	Storm
No rain	0.1	0.2	0.05	0.01
Light rain	0.05	0.1	0.15	0.04
Heavy rain	0.05	0.1	0.1	0.05

a) Find the marginal probability P(light rain).

➔ 0.05 + 0.1 + 0.15 + 0.04 = **0.34**

b) Find the marginal probability P(strong wind).

➔ 0.05 + 0.15 + 0.1 = **0.3**

c) Find the conditional probability P(heavy rain | strong wind)

➔ P(heavy rain AND strong wind)/ P(strong wind) = 0.1/(0.1+0.15+0.05) = **0.333**

d) Find the conditional probability P(some wind | light rain)

➔ P(some wind AND light rain)/ P(light rain) = 0.1/(0.05+0.1+0.15+0.04) = **0.294**

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

14) Read the pdf document on the website entitled Birthday Problems. Then answer the following question (question 3 on page 199 of the document):
A small class contains 6 students. What is the chance that at least two have the same *birth month*?

➔ There's a **77% chance** that at least two have the same birth month.

Probability that two students in the class of 6 have different birth months =
(12 * 11 * 10 * 9 * 8 * 7)/(12* 12 * 12 * 12 * 12 * 12) = 0.2228009

Probability that at least two students in the class of 6 have the same birth months =
1 - Probability that two students in the class of 6 have different birth months

Probability that at least two students in the class of 6 have the same birth months = 1-0.2228009
= 0.7771991

Karan A. Bhandarkar

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

15) Set up data as defined in the problem.

```
> 1:6
[1] 1 2 3 4 5 6
> sample(1:6)
[1] 6 4 2 1 3 5
> sample(1:6,1)
[1] 2
> sample(1:6,10,replace=TRUE)
 [1] 3 5 1 1 1 6 1 4 6 1
>
> Roll1Die=function(n)sample(1:6, n, rep=T)
> Roll1Die(5)
[1] 6 2 4 5 6
>
> die1=Roll1Die(100)
> die2=Roll1Die(100)
> diesum=die1+die2
> prop.table(table(diesum))
diesum
   2    3    4    5    6    7    8    9   10   11   12
0.02 0.07 0.12 0.10 0.16 0.16 0.09 0.10 0.12 0.05 0.01
```

Question a: There is something unusual about the probability table above-what is it?
We can find the probability of rolling a seven by the following R command

```
> sum(diesum==7)/100
```

a) `[1] 0.2`

➔ The probability table generate in the question seems to have the outcome 12 missing. The sum does add up to 1 thought so 12 just had no matches in the generated sample. The sample I generated did have 12.

Question b:
Using
http://www.mathcelebrity.com/2dice.php?gl=1&pl=7&opdice=1&rolist=+&dby=&ndby=&montect, what is the probability that the sum of two dice equals 7? Is our simulated example close?

➔ 1/6 = 0.167
Yes, the given simulated answer and my simulated answer are close.

Question c: Increase the number of dice rolls to 10000 each time. What is the new simulated probability that the sum equals 7?

➔ The new probability that the sum equals 7 is **0.1660**

```
> die1=Roll1Die(10000)
> die2=Roll1Die(10000)
> diesum=die1+die2
> prop.table(table(diesum))
diesum
     2      3      4      5      6      7      8      9     10     11     12
0.0271 0.0594 0.0838 0.1096 0.1398 0.1660 0.1361 0.1082 0.0876 0.0546 0.0278
```

Homework 2
STAT 104 - Introduction to Quantitative Methods for Economics

Question d: Using 10000 rolls for each time, what is the simulated probability that the value of dice 1 equals the value of dice 2? This can be done in R using the command sum(die1==die2). What is the true probability of the dice equaling each other? You can find this from the weblink above.

➔ Simulated probability that the value of dice 1 equals the value of dice 2 = 1628/10000 = **0.1628**

```
> sum(die1==die2)
[1] 1628
```

True probability from weblink = 1/6 = 0.167

Karan A. Bhandarkar