Colin Ashburn

CS373 HW1

1. a) {TH, TTH, TTTH, TTTTH, …, $T^{i-1}H$}; Probability of heads after i tosses: $½^i$
b) Set of outcomes is {TH, TTTH, TTTTTH, TTTTTTTH, …, $T^{i-1}H$ (where i % 2 == 0)}
This set of outcomes is half of the original set of outcomes, so **P(E) = ½**

2. E = Sum of dice is odd
F = At least one die is 1
G = Sum of dice is 5

P(E) = {1, 3, 5, 7, 9, 11} / {1, 2, 3, 4, …, 11, 12} = 6/12 = 1/2
P(F) = {11, 12, 13, 14, 15, 16, 21, 31, 41, 51, 61} / {11, 12, 13, …, 64, 65, 66} = 11/36
P(G) = {14, 23, 32, 41} / {11, 12, 13, …, 64, 65, 66} = 4/36 = 1/9

a) P(E ∩ F) = {12, 14, 16, 21, 41, 61} / 36 = 6/36 = **⅙**
b) P(E U F) = P(E) + P(F) - P(E ∩ F) = (18 + 11 - 6) / 36 = **23/36**
c) P(F U G) = P(F) + P(G) - P(F ∩ G) = (11 + 4 - 2) / 36 = **13/36**
d) P(E U !F) = P(E) + P(!F) - P(E ∩ !F) = (18 + 25 - 12) / 36 = **31/36**
e) P(E U F U G) = P(E) + P(F) + P(G) - P(E ∩ F) - P(E ∩ G) - P(F ∩ G) + P(E ∩ F ∩ G) =
(18 + 11 + 4 - 6 - 4 - 2 + 2) / 36 = **23/36**

3.

4. P(F) = 0.52
P(CS) = 0.05
P(F ∩ CS) = 0.0055

a) P(F | CS) = P(F ∩ CS) / P(CS) = **0.11**
b) P(CS | F) = P(CS ∩ F) / P(F) = **0.0106**
c) P(CS | F) = P(F | CS) * P(CS) / P(F) = 0.15 * 0.05 / 0.57 = **0.0132**

5.

a) E(H) = p*n; E(T) = (1 - p)*n
E(Xn) = E(H) - E(T) = pn - (1-p)n = pn - n + pn = **2pn - n**
b) Var(Xn) = Var(H) - Var(T)
c) $E(X_3)$ = 2p(3) - 3 = **6p - 3**
$Var(X_3)$ =

Part 3

```
> yelp <- read.csv(file="yelp.csv", header = TRUE, quote = "\"", comment.char = "")
> names(yelp)
 [1] "business_id"         "name"              "fullAddress"       "city"              "state"
 [6] "latitude"            "longitude"         "stars"             "reviewCount"       "checkins"
[11] "open"                "neighborhoods"     "categories"        "alcohol"           "noiseLevel"
[16] "attire"              "priceRange"        "delivery"          "ambience"          "parking"
[21] "dietaryRestrictions" "waiterService"     "smoking"           "outdoorSeating"    "caters"
[26] "recommendedFor"      "goodForGroups"     "goodForKids"
```

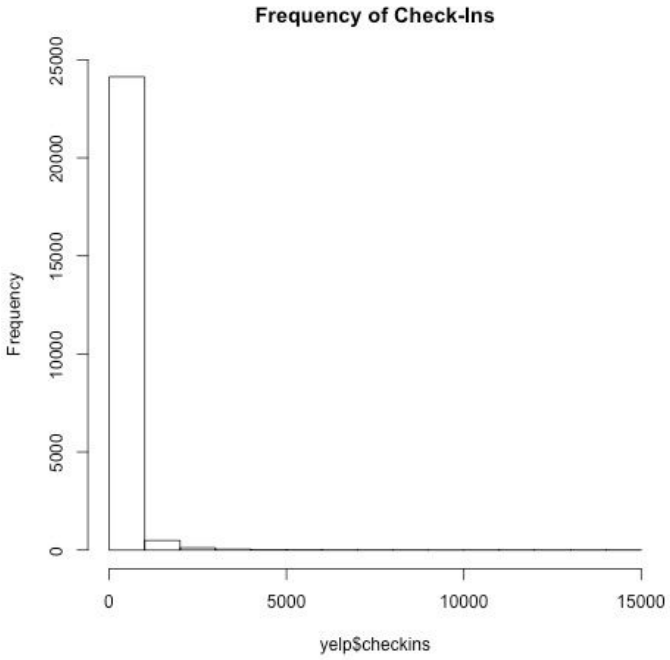
```
> summary(yelp)
                business_id            name                                                                   fullAddress              city            state
 __etvGuL2dh_a1LOT0gNYQ:     1   Starbucks :  407   Bellagio Las Vegas\n3600 S Las Vegas Blvd\nThe Strip\nLas Vegas, NV 89109          :    21   Las Vegas : 5256   AZ      :9301
 __kNfrrGoUXoF-BYciMU_Q:     1   McDonald's:  275   Las Vegas, NV                                                                      :    17   Phoenix   : 3072   NV      :6296
 __Y2jjdCFHvq3rzSbpDBlw:     1   Subway    :  256   5000 S Arizona Mills Cir\nTempe, AZ 85282                                          :    14   Charlotte : 1993   QC      :2389
 _-1EgXrkOlKajCsmasuEgg:     1   Walgreens :  158   3131 Las Vegas Blvd. South\nThe Strip\nLas Vegas, NV 89109                         :    13   Pittsburgh: 1467   NC      :2370
 _-6I6VXjr-NiwIBa_1uI4A:     1   Taco Bell :  148   Monte Carlo Hotel and Casino\n3770 Las Vegas Blvd S\nThe Strip\nLas Vegas, NV 89109:    13   Scottsdale: 1296   PA      :1613
 _-9pMxBWtG_x8l4rHWBasg:     1   Wendy's   :  113   2000 E Rio Salado Pkwy\nTempe, AZ 85281                                            :    12   Montral   : 1267   WI      :1089
 (Other)               :24807   (Other)   :23456   (Other)                                                                            :24723   (Other)   :10462   (Other):1755
    latitude        longitude            stars         reviewCount         checkins         open               neighborhoods                                     categories
 Min.   :32.88   Min.   :-115.370   Min.   :1.000   Min.   :   3.00   Min.   :    3   Mode :logical   []             :15727   ['Mexican', 'Restaurants']                : 1331
 1st Qu.:33.54   1st Qu.:-114.977   1st Qu.:3.000   1st Qu.:   8.00   1st Qu.:   16   FALSE:3580      ['The Strip']:   816   ['Food', 'Coffee & Tea']                  :  844
 Median :36.03   Median :-111.924   Median :3.500   Median :  18.00   Median :   48   TRUE :21233     ['Southeast']:   639   ['Pizza', 'Restaurants']                  :  831
 Mean   :37.53   Mean   : -97.298   Mean   :3.544   Mean   :  49.03   Mean   :  166                   ['Downtown'] :   533   ['Chinese', 'Restaurants']                :  776
 3rd Qu.:40.41   3rd Qu.: -80.807   3rd Qu.:4.000   3rd Qu.:  48.00   3rd Qu.:  155                   ['Westside'] :   526   ['Burgers', 'Fast Food', 'Restaurants']:  549
 Max.   :55.99   Max.   :   8.549   Max.   :5.000   Max.   :4578.00   Max.   :14203                   ['Eastside'] :   447   ['Restaurants', 'Italian']                :  509
                                                                                                      (Other)      :  6125   (Other)                                   :19973
         alcohol         noiseLevel        attire         priceRange        delivery              ambience                 parking                     dietaryRestrictions
              :    3             : 7947          : 7005   Min.   :1.000   Mode :logical   ['casual']:7878   ['lot']          :10348                                  :24696
 beer_and_wine: 2497   average   :10957   casual:17129   1st Qu.:1.000   FALSE:14471                :7875   []               : 6675   ['vegan']                      :   45
 full_bar     : 7565   loud      : 1622   dressy:   640   Median :2.000   TRUE :3093      []         :6348   ['street']       : 3046   ['vegetarian']                 :   23
 none         :14748   quiet     : 3562   formal:    39   Mean   :1.631   NA's :7249      ['divey'] : 716                     : 2456   []                             :   20
                       very_loud:   725                   3rd Qu.:2.000                   ['trendy']: 567   ['garage']       :  907   ['dairy-free', 'vegetarian']:    7
                                                          Max.   :4.000                   ['classy']: 320   ['street', 'lot']:  364   ['vegan', 'vegetarian']        :    5
                                                          NA's   :903                     (Other)   :1109   (Other)          : 1017   (Other)                        :   17
 waiterService        smoking      outdoorSeating     caters                  recommendedFor  goodForGroups    goodForKids
 Mode :logical          :21862   Mode :logical    Mode :logical                         :7859   Mode :logical    Mode :logical
 FALSE:6208      no     :  904   FALSE:10989      FALSE:6503      []                    :4932   FALSE:2054       FALSE:506
 TRUE :10351     outdoor: 1415   TRUE :8698       TRUE :5932      ['lunch']             :4324   TRUE :17078      TRUE :1283
 NA's :8254      yes    :  632   NA's :5126       NA's :12378     ['dinner']            :2553   NA's :5681       NA's :23024
                                                                  ['lunch', 'dinner']:1966
                                                                  ['breakfast']         :1004
                                                                  (Other)               :2175
```

```
> summary(yelp$noiseLevel)
           average      loud     quiet very_loud 
     7947    10957      1622      3562       725 
> summary(yelp$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   3.000   3.500   3.544   4.000   5.000 
```

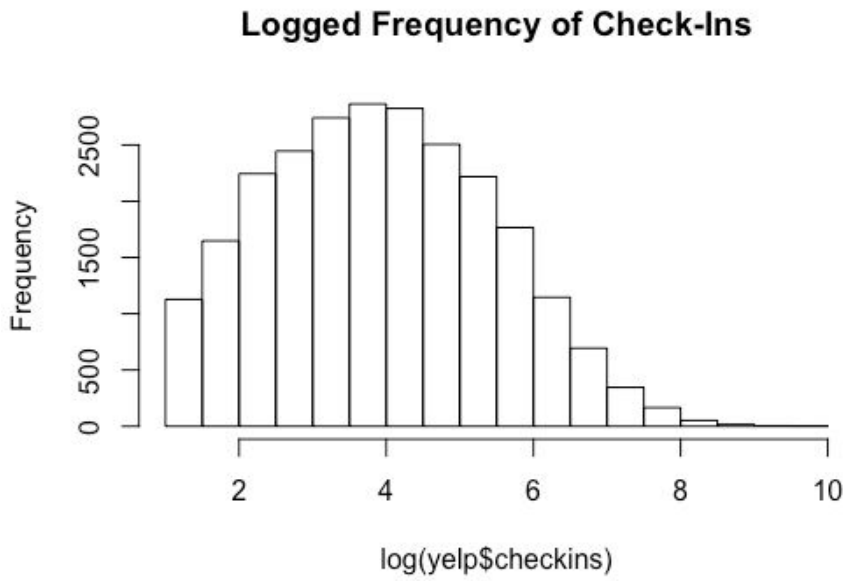
Part 4

a)

```
> jpeg('foo.jpg')
> hist(yelp$checkins, main = "Frequency of Check-Ins")
> dev.off()
RStudioGD 
        2 
```



b)



c) The second plot seems to be similar to a uniform distribution, which is an expected distribution of data. Because the second plot is the log() of the data, it suggests that the data points change exponentially (decreasing). This means that there are many many restaurants with very few check-ins and very few restaurants who have many check-ins.

Part 5

a)

```
> #Part 5
> yelp <- cbind(yelp, isAmerican=grepl("American", yelp$categories), goodForDinner=grepl("dinner", yelp$recom
mendedFor))
> summary(yelp$isAmerican)
   Mode   FALSE    TRUE
logical   21456    3357
> summary(yelp$goodForDinner)
   Mode   FALSE    TRUE
logical   19670    5143
```

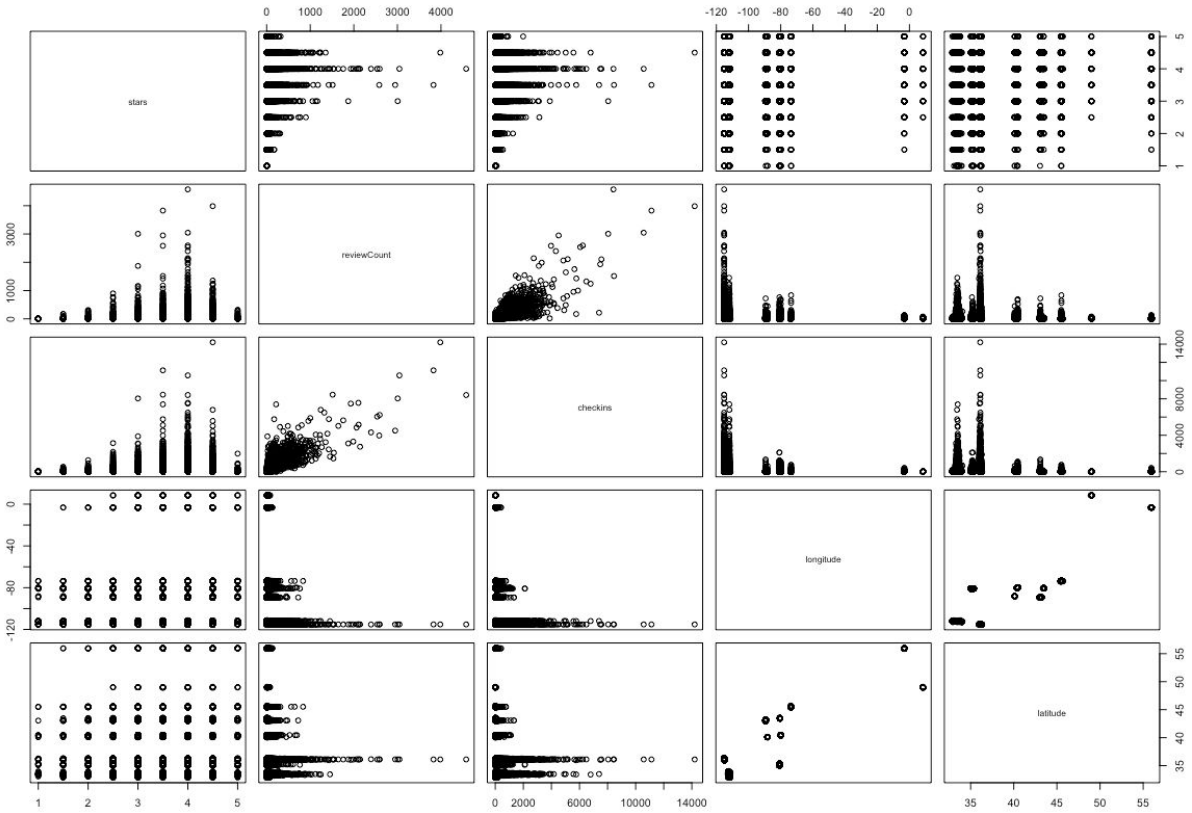
b)

```
> print("Review Count Original/Modified", quotes = FALSE)
[1] "Review Count Original/Modified"
> quantile(yelp$reviewCount)
  0%  25%  50%  75% 100%
   3    8   18   48 4578
> quantile(yelp$reviewCount[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
  0%  25%  50%  75% 100%
   3    4    5    7    8
```

c)

```
> summary(yelp$reviewCount[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   4.000   5.000   5.247   7.000   8.000
> summary(yelp$stars[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   3.500   3.418   4.000   5.000
> summary(yelp$attire[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
      casual dressy formal
  3248   3581    107     24
> summary(yelp$priceRange[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.000   1.000   1.000   1.546   2.000   4.000     825
> summary(yelp$delivery[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
   Mode   FALSE    TRUE    NA's
logical    2899     693    3368
> summary(yelp$goodForKids[yelp$reviewCount <= quantile(yelp$reviewCount)[2]])
   Mode   FALSE    TRUE    NA's
logical      15      31    6914
```

Part 6



ReviewCount and checkIns seem to be highly correlated and have almost a linear relationship. It is relatively expected--the more people who check in at (go to) a restaurant, the more people would review that restaurant.

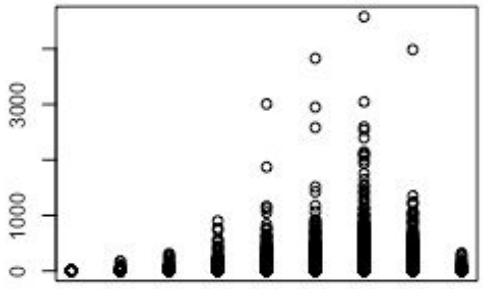
b)

```
> cor(yelp[,c("checkins", "reviewCount", "longitude", "latitude")])
              checkins reviewCount  longitude    latitude
checkins     1.0000000  0.82749365 -0.1789531 -0.15260462
reviewCount  0.8274936  1.00000000 -0.1294142 -0.09850936
longitude   -0.1789531 -0.12941420  1.0000000  0.88110176
latitude    -0.1526046 -0.09850936  0.8811018  1.00000000
```

The highest positive correlation is between longitude and latitude. This doesn't necessarily make sense, because longitude and latitude are independent, and locations aren't based on increases in both. The highest negative correlation is longitude and checkins. This is also not necessarily expected--aside from the location of where there are many restaurants, this correlation seems to just be because of sampling bias.

Part 7

a) Stars vs. ReviewCount



b) Stars and ReviewCount are both discrete. A Scatterplot is relatively appropriate for viewing this relationship.

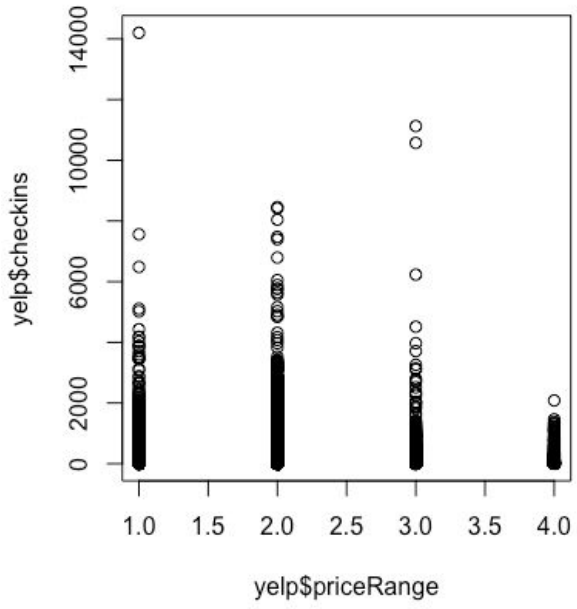
c) Stars is associated with higher ReviewCount.

d) I predict that the number of Stars increases with the number of ReviewCount, because if people rate a restaurant highly, they likely would star it as well.

e) Directional relational

---

a)



b) They are both discrete, so a scatter plot is appropriate for this relationship.

c) PriceRange is associated with Checkins

d) I predict that Checkins is inversely related to PriceRange, because more people visit restaurants with a lower price range.

e) Directional Relational