

# CS373 Homework1 – Part B

By: Siddharth Shah

3) a) `> 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"

```

b) `> summary(yelp)`

```

business_id      name
--1emggGhgog6ipd_RMb-g: 1 Starbucks : 407
--5jkZ3-nUPZxuvTcbr8Uw: 1 McDonald's : 275
-024YETnIsPQCrmSHCKLQW: 1 Subway : 256
-0bUDimSOGuv8R0Qq6J4A: 1 Walgreens : 158
-0D_CYh1D2ILkmLR0pBmNA: 1 Taco Bell : 148
-0GkcDIgVm0XzDZC8RFog: 1 Wendy's : 113
(other) :24807 (other) :23456

fullAddress      city
Bellagio Las Vegas\n3600 S Las Vegas Blvd\nThe Strip\nLas Vegas, NV 89109 : 21 Las Vegas : 5256
Las Vegas, NV : 17 Phoenix : 3072
5000 S Arizona Mills Cir\nTempe, AZ 85282 : 14 Charlotte : 1993
3131 Las Vegas Blvd. South\nThe Strip\nLas Vegas, NV 89109 : 13 Pittsburgh: 1467
Monte Carlo Hotel and Casino\n3770 Las Vegas Blvd S\nThe Strip\nLas Vegas, NV 89109: 13 Scottsdale: 1296
2000 E Rio Salado Pkwy\nTempe, AZ 85281 : 12 Montral : 1267
(other) :24723 (other) :10462

state latitude longitude stars reviewCount checkins open
AZ :9301 Min. :32.88 Min. : -115.370 Min. :1.000 Min. : 3.00 Min. : 3 Mode :logical
NV :6296 1st Qu.:33.54 1st Qu.: -114.977 1st Qu.:3.000 1st Qu.: 8.00 1st Qu.: 16 FALSE:3580
QC :2389 Median :36.03 Median : -111.924 Median :3.500 Median : 18.00 Median : 48 TRUE :21233
NC :2370 Mean :37.53 Mean : -97.298 Mean :3.544 Mean : 49.03 Mean : 166
PA :1613 3rd Qu.:40.41 3rd Qu.: -80.807 3rd Qu.:4.000 3rd Qu.: 48.00 3rd Qu.: 155
WI :1089 Max. :55.99 Max. : 8.549 Max. :5.000 Max. :4578.00 Max. :14203
(other):1755

neighborhoods categories alcohol noiseLevel
[] :15727 ['Mexican', 'Restaurants'] : 1331 : 3 : 7947
['The Strip'] : 816 ['Food', 'Coffee & Tea'] : 844 beer_and_wine: 2497 average :10957
['Southeast'] : 639 ['Pizza', 'Restaurants'] : 831 full_bar : 7565 loud : 1622
['Downtown'] : 533 ['Chinese', 'Restaurants'] : 776 none :14748 quiet : 3562
['Westside'] : 526 ['Burgers', 'Fast Food', 'Restaurants'] : 549 very_loud: 725
['Eastside'] : 447 ['Restaurants', 'Italian'] : 509
(other) : 6125 (other) :19973

attire priceRange delivery ambience parking
: 7005 Min. :1.000 Mode :logical ['casual'] :7878 ['lot'] :10348
casual:17129 1st Qu.:1.000 FALSE:14471 :7875 [] : 6675
dressy: 640 Median :2.000 TRUE :3093 [] :6348 ['street'] : 3046
formal: 39 Mean :1.631 NA's :7249 ['divey'] : 716 : 2456
3rd Qu.:2.000 ['trendy'] : 567 ['garage'] : 907
Max. :4.000 ['classy'] : 320 ['street', 'lot'] : 364
NA's :903 (other) :1109 (other) : 1017

dietaryRestrictions waiterService smoking outdoorSeating caters
:24696 Mode :logical :21862 Mode :logical Mode :logical
['vegan'] : 45 FALSE:6208 no : 904 FALSE:10989 FALSE:6503
['vegetarian'] : 23 TRUE :10351 outdoor: 1415 TRUE :8698 TRUE :5932
[] : 20 NA's :8254 yes : 632 NA's :5126 NA's :12378
['dairy-free', 'vegetarian'] : 7
['vegan', 'vegetarian'] : 5
(other) : 17

recommendedFor goodForGroups goodForKids
:7859 Mode :logical Mode :logical
[] :4932 FALSE:2054 FALSE:506
['lunch'] :4324 TRUE :17078 TRUE :1283
['dinner'] :2553 NA's :5681 NA's :23024
['lunch', 'dinner'] :1966
['breakfast'] :1004
(other) :2175

```

c) `> summary(yelp$noiseLevel)`

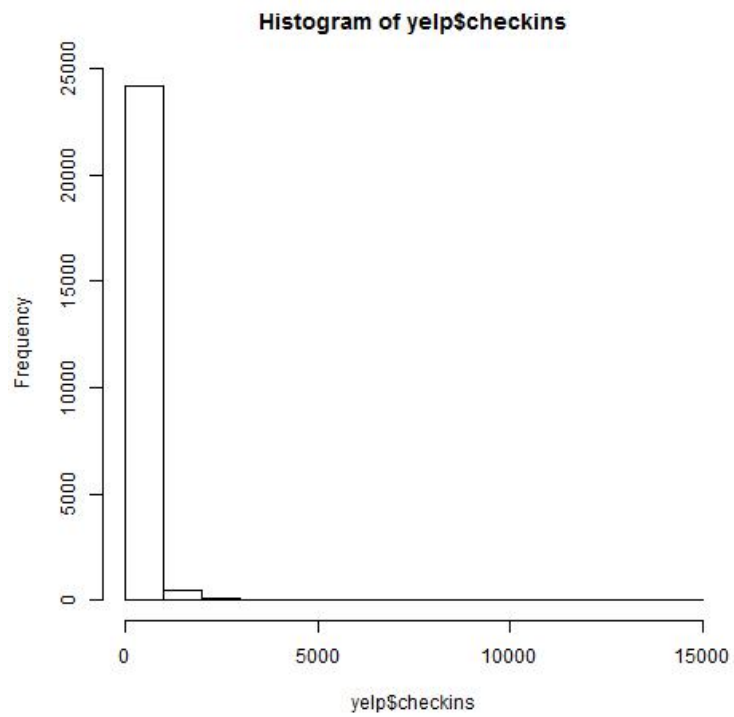
```

average loud quiet very_loud
7947 10957 1622 3562 725

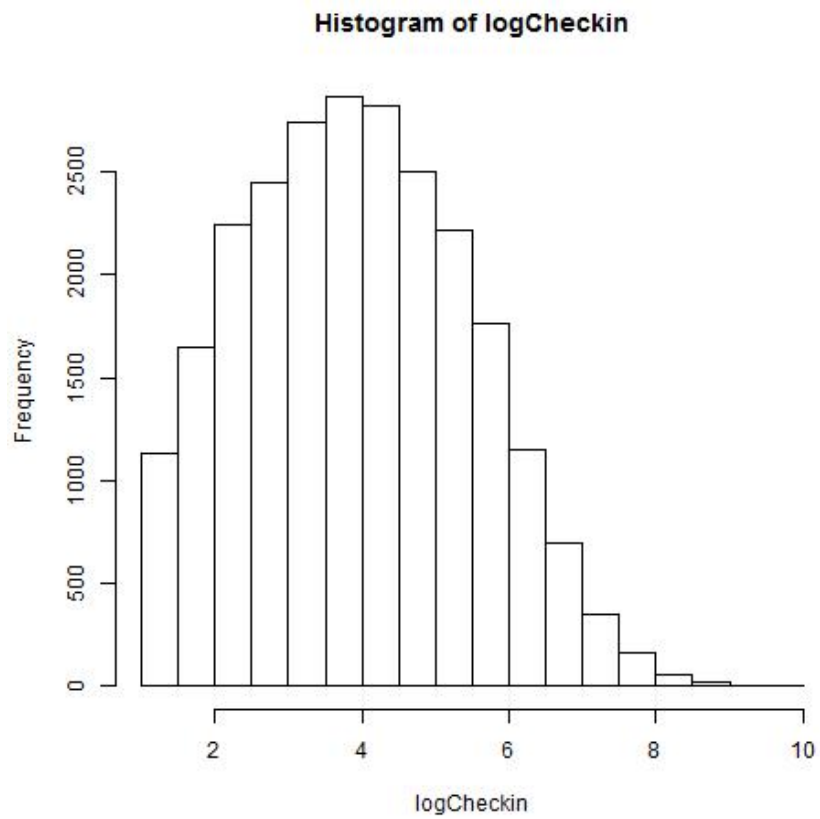
> summary(yelp$stars)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.000 3.000 3.500 3.544 4.000 5.000

```

4) a)



b)



c) Since the Yelp data has data about restaurants some very popular restaurants as well as new/upcoming or not-so-popular restaurants, the raw check-in histogram is skewed. This causes the frequency of higher check-ins diminish the expressivity of data by shadowing the less frequency data.

Using log-scale causes a normalization of the data generally, there by weighting all data equally, and also getting rid of broken data like negative check-ins if the exist. The shape is also more normal, there by creating ease of applying inference and statistical techniques.

```
5) a) > summary(yelp$isAmerican)
      Mode FALSE  TRUE
logical 21456  3357
> summary(yelp$goodForDinner)
      Mode FALSE  TRUE
logical 19670  5143

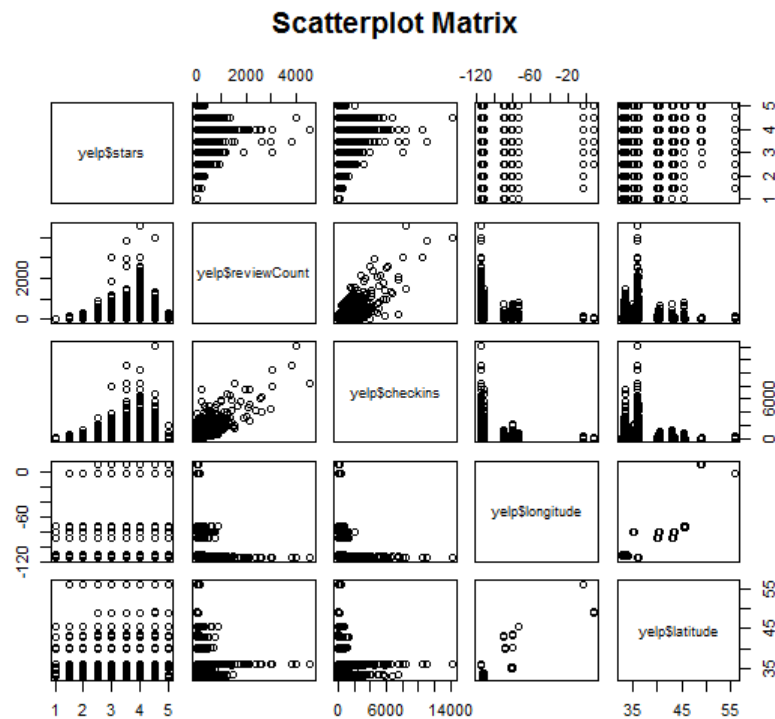
b) > quantile(yelp$reviewCount)
0%  25%  50%  75% 100%
 3    8   18   48 4578

c) > summary(yelp_subset$reviewCount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.000  4.000   5.000   5.247  7.000   8.000
> summary(yelp_subset$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000  3.000   3.500   3.418  4.000   5.000
> summary(yelp_subset$attire)
      casual dressy formal
 3248   3581    107     24
> summary(yelp_subset$priceRange)
   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_subset$delivery)
   Mode FALSE  TRUE  NA's
logical 2899   693  3368
> summary(yelp_subset$goodForKids)
   Mode FALSE  TRUE  NA's
logical 15    31   6914

> summary(yelp$reviewCount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.00  8.00  18.00  49.03  48.00 4578.00
> summary(yelp$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.000  3.000   3.500   3.544  4.000   5.000
> summary(yelp$attire)
      casual dressy formal
 7005  17129   640     39
> summary(yelp$priceRange)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
 1.000  1.000   2.000   1.631  2.000   4.000    903
> summary(yelp$delivery)
   Mode FALSE  TRUE  NA's
logical 14471  3093  7249
> summary(yelp$goodForKids)
   Mode FALSE  TRUE  NA's
logical 506   1283 23024
```

On general, since we are comparing subset to a superset, we see a decrease in means for every variable. However, for something skewed like the “reviewCount” we see a sharp decrease in the mean. We see that most priceRange data is unavailable below 1<sup>st</sup> quantile reviewCount. While the “goodForKids” data is mostly available below the reviewCount below 1<sup>st</sup> quantile. Median number of stars has less effect on the reviewCount, since it is the same for both the datasets.

6) a)



These correlations have some obviousness, like the strong co-relation between the latitude and longitude, since these 2 variables are literally at the cross-section for places available in the data. The correlation between stars and reviewCount and stars and check-ins is interesting, since this gives valuable inference to the credibility of Yelp as a platform. Why do the stars decrease after a certain number of reviewCounts? Do people see more reviews as something misleading? And why do people have less check-ins at restaurants with more stars after a certain threshold? These questions can be investigated based on the data. A striking contrast to that is the reviewCount vs. the Check-ins, that relationship seems to uniformly increasing. So that means that people who visit places more write more reviews, this means that Yelp as a platform is achieving its desired model.

```

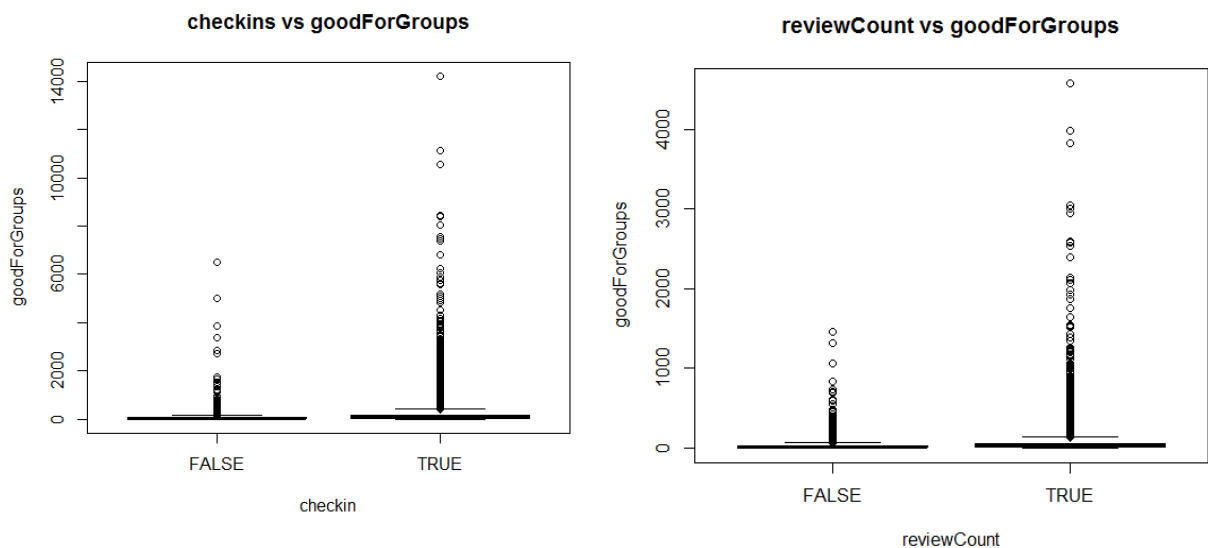
b) > cor(yelp$stars, yelp$reviewCount)
[1] 0.1070506
> cor(yelp$stars, yelp$checkins)
[1] 0.09440071
> cor(yelp$stars, yelp$longitude)
[1] 0.1174446
> cor(yelp$stars, yelp$latitude)
[1] 0.1211631
> cor(yelp$reviewCount, yelp$checkins)
[1] 0.8274936
> cor(yelp$reviewCount, yelp$longitude)
[1] -0.1294142
> cor(yelp$reviewCount, yelp$latitude)
[1] -0.09850936
> cor(yelp$checkins, yelp$longitude)
[1] -0.1789531
> cor(yelp$checkins, yelp$latitude)
[1] -0.1526046
> cor(yelp$longitude, yelp$latitude)
[1] 0.8811018

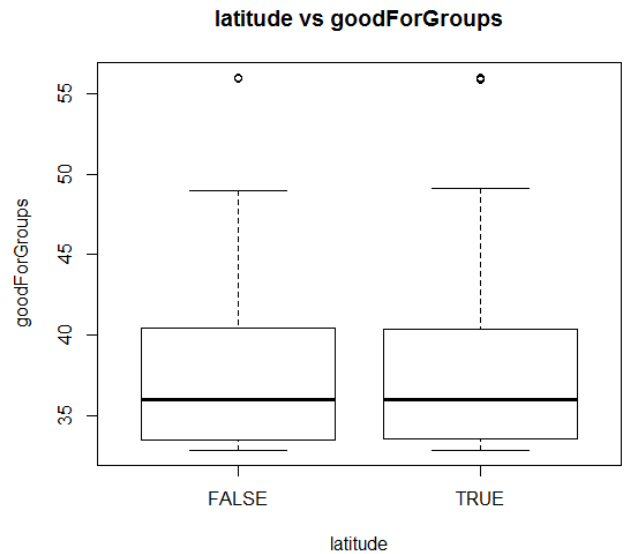
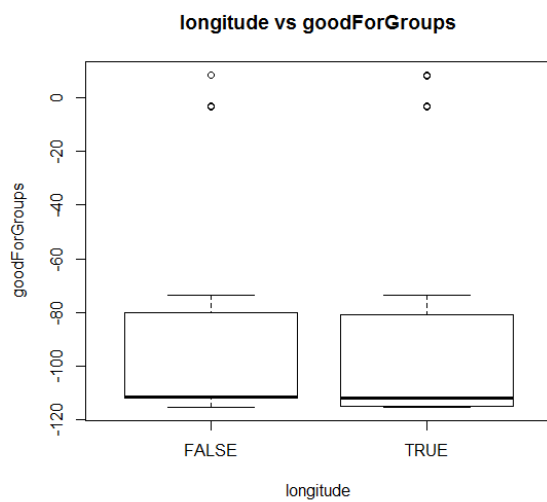
```

The largest pairwise positive correlation exists between longitude and latitude. While the largest pairwise negative correlation exists between check-ins and longitude.

Visually, these correlations comply, since we can see a linear relationship between longitudes and latitudes both ways, while check-ins and longitudes do not seem to have any correlation at all.

c)

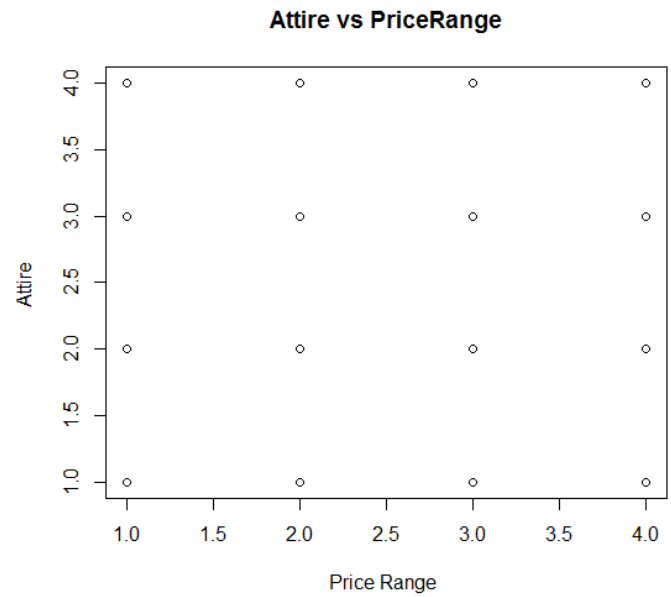
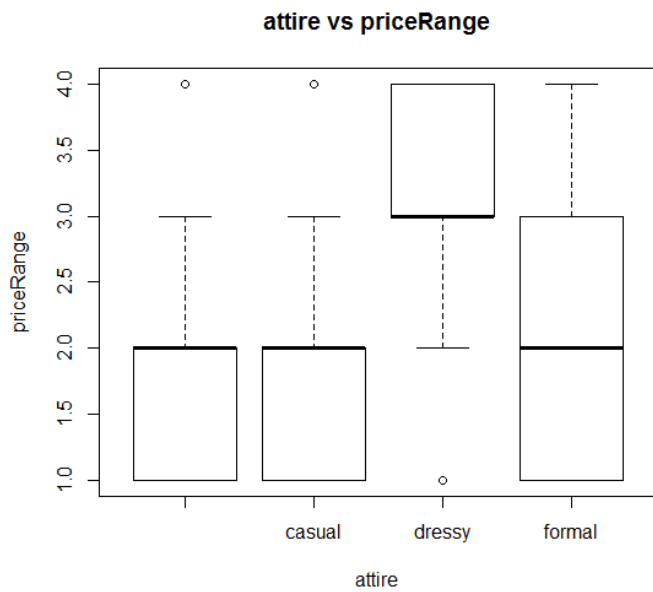




```
> tru_ch <- subset(yelp, yelp$goodForGroups == TRUE)
> quantile(tru_ch$checkins)
 0%  25%  50%  75% 100%
 3   19   59  181 14203
> quantile(tru_ch$reviewCount)
 0%  25%  50%  75% 100%
 3   10   24   61 4578
> quantile(tru_ch$longitude)
 0%      25%      50%      75%     100%
-115.36973 -115.04307 -111.92574 -80.82606  8.54856
> quantile(tru_ch$latitude)
 0%      25%      50%      75%     100%
32.87687 33.53849 36.02708 40.36092 55.99042
> false_ch <- subset(yelp, yelp$goodForGroups == FALSE)
> quantile(false_ch$checkins)
 0%  25%  50%  75% 100%
 3   11   25   66 6485
> quantile(false_ch$reviewCount)
 0%  25%  50%  75% 100%
 3    7   13   30 1453
> quantile(false_ch$longitude)
 0%      25%      50%      75%     100%
-115.328981 -112.152874 -111.840497 -80.018910  8.410954
> quantile(false_ch$latitude)
 0%      25%      50%      75%     100%
32.87918 33.51192 36.04116 40.45204 55.97743
```

The variable “checkins” exhibits most association with “goodForKids”. This is interesting because one would expect more checkins creating happy vibe in the restaurant that would make place more kid-friendly. However, this is bizarre because a lot of checkins also happen at places like bars, which aren’t kid-friendly.

7) a) Based on the data, I’d like to propose a relationship between PriceRange and Attire type.



b) These variables are discrete since these are discrete price ranges, and categorical variables corresponding to the attire type.

c) The function would relate “attire” (X) to “priceRange” (Y), and not the other way around. We can see there is no association the other way, but attire to price-range does have inferable association.

d) Based on the box plot, we can see that these variables have some inference property. We can see that more “dressy” attire relates to high priced places, and formal dressing is not just limited to high-prices places.

Is the fourth unknown type of dressing also “casual”? Since the data tightly corresponds to that of casual dressing.

Would inducing high prices, yield an environment of the restaurant where “dressy” clothing is preferred?

e) The hypothesis purely empirical, since the data is more or less empirical and so is the inference.