Capstone for John Hopkins Data Science- Yelp Final Project

ckchan

November 18, 2015

Introduction

Generate an analysis of which franchises should be opened in cities in Arizona. Use a population census database of Arizona, www.arizona-demographics.com/cities_by_population and the yelp database. Franchises that are under represented with respect to population will imply a business opportunity for that franchise in that city in Arizona.

I will use a Mean Reversion model to generate a list of these franchises and cities in Arizona that are under represented in cities in Arizona

Methods

1. Read in Arizona's population and Yelp's business dataset into their respective dataframes. There will defintely be a need to prepare and clean up certain data
2. Arizona's cities population representation will form the mean. If any franchise representation in that city is lower than that mean by a certain threhold, it implies that business is under represented in that city and thus signifies a business opportunity for that franchise

```
azpop<-data.frame(read.csv("arizonapop.csv")) ##Read Arizona's population from external csv file
colnames(azpop)<-c("city","population")
azpop2<-subset(azpop,population>=200000)#Work on cities with a sizable population in Arizona
azpop2<-tbl_df(azpop2)
```

```
azbiz<-stream_in(file("yelp_academic_dataset_business.json")) ##Read in the Yelp business json file
```

```
## opening file input connection.
## closing file input connection.
```

```
azbiz1<-merge(azbiz,azpop2)## 'city' is the common key to be joined
tabbizperc<-round(prop.table(table(azbiz1$city))*100,2)
tabbizperc<-tbl_df(data.frame(tabbizperc))
colnames(tabbizperc)<-c('city','bizperc')
tabbizperc
```

```
## Source: local data frame [7 x 2]
##
##         city bizperc
##       (fctr)   (dbl)
## 1   Chandler    9.67
## 2    Gilbert    6.54
## 3   Glendale    7.13
## 4       Mesa   12.16
## 5    Phoenix   43.57
## 6 Scottsdale   20.92
## 7     Tucson    0.01
```

Distribution of businesses in the Yelp database broken down by cities followed by percentages. Businesses in Tucson is only 0.01%. It might be possible that Yelp's Tucson data isn't complete. So we ignore Tucson and focus on major cities with population greater than 200,000

```
azpop2<-tbl_df(subset(azpop,population>=200000 & city!="Tucson")) ##repeat, this time w
ithout Tucson
azbiz1<-merge(azbiz,azpop2)## 'city' is the common key to be joined
tabbizperc<-tbl_df(data.frame(round(prop.table(table(azbiz1$city))*100,2)))
colnames(tabbizperc)<-c('city','bizperc')
merget1<-tbl_df(merge(azpop2,tabbizperc))
merget1[4]<-merget1[2]/sum(merget1$population)*100  #create new column for population%
colnames(merget1)<-c('city','population','bizperc','popuperc')
merget1
```

```
## Source: local data frame [6 x 4]
##
##         city population bizperc  popuperc
##       (fctr)      (int)   (dbl)     (dbl)
## 1   Chandler     254276    9.67  8.580712
## 2    Gilbert     239277    6.54  8.074560
## 3   Glendale     237517    7.13  8.015168
## 4       Mesa     464704   12.16 15.681743
## 5    Phoenix    1537058   43.57 51.869037
## 6 Scottsdale     230512   20.92  7.778780
```

Notice that Scottsdale has too huge (20.92% vs 7.78%) a percentage of businesses compared to the percentage of population among the 6 shortlisted cities. We repeat the process and remove Scottsdale from the shortlisted cities.

```
azpop2<-tbl_df(subset(azpop,population>=200000 & city!="Tucson" & city!="Scottsdale"))
#repeat without Tucson and Scottsdale
azpop2perc<-mutate(azpop2,perc=population/sum(azpop2$population))
azbiz1<-merge(azbiz,azpop2)## 'city' is the common key to be joined
tabbizperc<-tbl_df(data.frame(round(prop.table(table(azbiz1$city))*100,2)))
colnames(tabbizperc)<-c('city','bizperc')
merget1<-tbl_df(merge(azpop2,tabbizperc))
merget1[4]<-merget1[2]/sum(merget1$population)*100 #create new column for population%
colnames(merget1)<-c('city','population','bizperc','popuperc')
```

The above dataframe shows the total number of businesses percentages vs population percentages for the shortlisted 5 cities of Arizona that we will be focusing on. We will next iteratively test the various frachise brands in Chandler, Gilbert, Glendale, Mesa and Phoenix for any divergences in franchise outlets percentage relative to the population percentage

Results

```
merget1
```

```
## Source: local data frame [5 x 4]
##
##       city population bizperc  popuperc
##     (fctr)      (int)   (dbl)     (dbl)
## 1 Chandler     254276   12.23  9.304487
## 2  Gilbert     239277    8.27  8.755642
## 3 Glendale     237517    9.02  8.691240
## 4     Mesa     464704   15.38 17.004485
## 5  Phoenix    1537058   55.10 56.244145
```

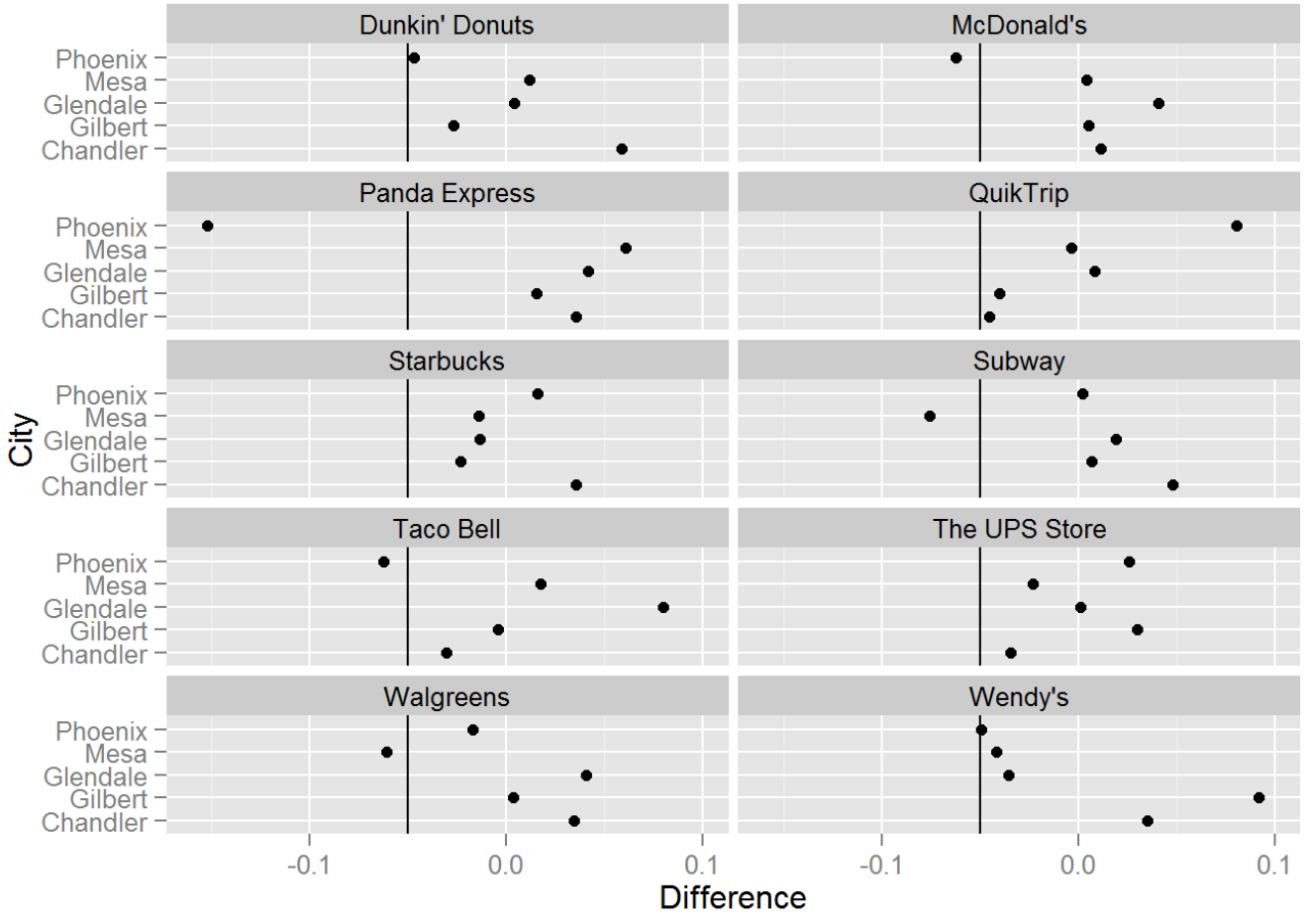
Popuperc column of the dataframe gives the mean. A positive difference from this figure implies there's too many businesses while a negative value would imply possible business opportunities. Chandler seems to have too many businesses with respect to the population percentage while Mesa would probably have some business opportunities. The next step is to further breakdown individual franchises per city to see which franchise to open and where to open it.

```
azbizdf<-tbl_df(data.frame(azbiz1[,c(1,8)]))#obtain the top 10 franchises in the state
azbiztabledf<-tbl_df(data.frame(table(azbizdf$name)))
azbiztabledf<-tbl_df(data.frame(prop.table(table(azbizdf$name))))
azbiztop10<-top_n(azbiztabledf,10,Freq)
azbizdftop10<-data.frame(azbiztop10) #convert to dataframe, use later in table generato
r function
azbizdftop10namesonly<-tbl_df(data.frame(azbizdftop10[,1]))
azbizdftop10namesonly
```

```
## Source: local data frame [10 x 1]
##
##    azbizdftop10...1.
##                (fctr)
## 1     Dunkin' Donuts
## 2         McDonald's
## 3      Panda Express
## 4           QuikTrip
## 5          Starbucks
## 6             Subway
## 7          Taco Bell
## 8      The UPS Store
## 9          Walgreens
## 10           Wendy's
```

These are the 10 franchises that we will iteratively test, at a city level, if their percentage representation lags behind the population percentage

```
generateTable<-function(a)
{
  resulttab<-NULL
  for (i in 1:a)
  {
    test1<-tapply(azbizdf$name==azbizdftop10[i,1],azbizdf$city,sum)
    test1<-prop.table(test1)
    test1<-cbind(azbizdftop10namesonly[i,1],test1)
    test1<-cbind(row.names(test1),test1) #add in city name
    resulttab<-rbind(resulttab,test1)
  }
  return (resulttab)
}
results1<-generateTable(10)
colnames(results1)<-c('city','name','bizperc')
results2<-merge(azpop2perc,results1)
results2$difference<-results2$bizperc-results2$perc
results3<-results2[,c(4,1,6)]
ggplot(data=results3, aes(x=city,y=difference)) + geom_point(data = results3, size = 2)
+ylim(-0.16, 0.1) + geom_hline(yintercept=-0.05)+ coord_flip() +  facet_wrap( ~ name,nc
ol=2 ) +  labs(x="City",y="Difference")
```



Discussion/Conclusions

I've regard franchises as business opportunities if they lag behind the population percentage mean by 5%. Thus from the plot, it can be seen that the following Franchise - City are underpresented

1. McDonald's - Phoenix
2. Panda Express - Phoenix
3. Subway - Mesa
4. Taco Bell - Phoenix
5. Walgreens - Mesa
6. Wendy's - Phoenix

In conclusion, this study provides a statistical way of analyzing which franchises is too far from the mean. Assuming that this relationship will mean revert to the population mean, it will imply a better risk/reward compared to opening other type of franchises in these cities.

However further analysis has to be done on socio-economic variables which will inevitably affect business decisions.

Thank you for reading through this report.