Capstone project

Dinh Huy Hoang

18th Nov 2015

Project Title

Check-ins a restaurant in Las Vegas city base on number of reviews, stars and its price range

Analysis the relationship between number of reviews, average stars and its price range with checkins customers of Restaurants in Las Vegas and build the Linear Regression model (statistical inference) to find out it.

Introduction

Executive Summary

Yelp was founded in 2004 to help people find great local businesses. Yelp covers a vast variety of businesses, like restaurants, bars, cafes, local events, doctors, pharmacies, hotels and so on. Users have accounts, and can also add friends on Yelp. Yelp had a monthly average of 89 million unique visitors who visited Yelp via their mobile device in Q3 2015* Yelpers have written more than 90 million reviews by the end of Q3 2015.

Over the years, Yelp has collected a huge database of users, businesses, and reviews. Yelp has opened up a lot of opportunities for analyzing and using this data for different purposes in different ways. How we can use the Yelp dataset to help Business owner to sustain the business restaurants. *The sustainability of a restaurant can be measured by numbers of customers(checkins)*, more customers it means business will be sustainable and success.

Question and the rationale for studying

With Analysis of the Yelp dataset, we will build the Linear regression model to find out the relationship between the number of reviews, average star, price range and checkins customers and answer for the question **Can number of reviews, average star and its price range decide the checkins customers of a restaurant in Las Vegas - one of top tourist destinations in the United States?**

Methods and Data

Yelp dataset Characteristics

The Yelp Dataset is Json format and after loaded we have **Review data**:1569264 reviews, **Business data**:61184 businesses, **User data**: 366715 users, **Tips data**:495107 tips and **Checkin data**:45166 checkins(Over time for each of the 61K businesses)

Almost 481K business attributes, e.g., hours, parking availability, ambience. Social network of 366715 users for a total of 2.9M social edges. And (4 Countries and 10 cities: U.K.: Edinburgh, Germany: Karlsruhe, Canada: Montreal and Waterloo, U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison). The Check-in, Review and Business data for analysis

- Checkin data: Information for all checkins for 61K businesses such as number of checkins, hourly, daily
- Business data: It includes information for the variables such as business id, location information including latitudes and longitudes, full address, and so on for each business. It also has the number of reviews that have ever been written for the business, and an average star rating across all the reviews.
- Reviews data: Each review consists of a star rating and an optional review text. The review datatype also connects the businesses they review. Each review can also get votes from other users, who can say if they find the review cool, funny, or useful

Getting, cleaning and transforming data

Checkins data

- Load checkins data and sum up total checkins each business id, we will merge with business data
- Merge checkins with Business data by business id
- Sum up checkins and review_count by city

Top five cities by the number of checkins

city	checkins	reviews	percentage
Las Vegas	2630281	669440	25.45
Phoenix	911264	222925	24.46
Scottsdale	616411	123184	19.98
Charlotte	336758	89189	26.48
Tempe	268436	65988	24.58

It shows the total number of reviews for all businesses in that city, as well as the percentage of reviews over checkins. It can be seen that for the biggest cities in the datasets, and Las Vegas is city which has top number of checkins customers

The Las Vegas is one of top tourist destinations in the United States, therefore above result is quite logic, the city has top checkins customers and top number of reviews. In this report we will focus business restaurants in that city.

Business data

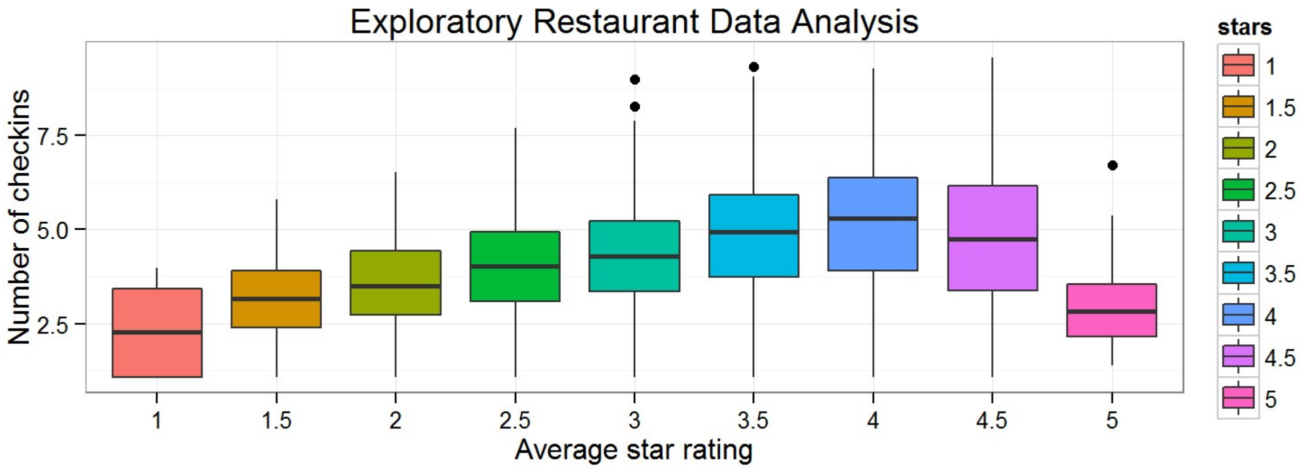
- Use the Business data after merge with checkin, Subset all city = “Las Vegas” and
- Subset one more time to get all “Restaurants” in the “categories” in “business” dataset.
- Flatten and Explore Business Attributes to get relevant attributes it may related to checkins customers such as “Good for Kids”, “Good for groups”, “Price Range”, “Takes Reservations”, “Open 24 hours”
- Transform above business attributes to numeric and replace all NA to ‘0’

Review data

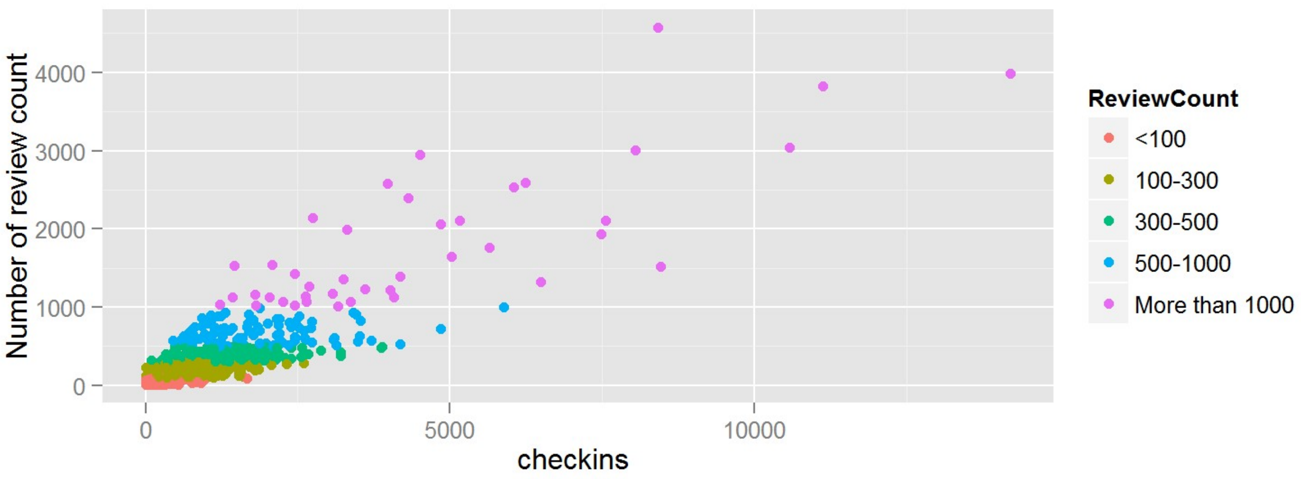
- Subset all review data for Business Restaurant in Las Vegas City
- Sum up by start_date to find the min date (the date restaurant started operated or recorded in Yelp dataset)
- Sum up by end_date to find out the max date (the latest date restaurants operate or recorded in Yelp dateset)
- Flatten and Sum up total number of votes (cool, useful, funny) for each restaurant
- Add in variable **lifespan** for each restaurant by number of dates (end_date - start_date)

Exploratory Restaurant data in Las Vegas City

After extract, transform and merge business and review data by business_id and We have a final business dataset with 4033 Restaurants and 20 variables (business_id, full_address, categories, city, state, name, review_count, stars, checkins, good4kids, good4groups, price range, reservations, open24h, start_date, end_date, funny, cool, useful and lifespan) for analysis



The boxplots of the number of checkins to each business restaurant by the average star rating of that business. It can be observed that restaurants have three to four star rating tend to have the most checkins (customers), but one and five star rating restaurants actually have fewer. This may be either due to the fact that restaurant with such a lowest or highest rating have fewer total reviews, or they are poor of services or more expensive and less likely to have a high number of customers.



The above plot is very encouraging that there is strong relationship between the number of review from customers checkins customers. The plot demonstrates that more number of reviews each restaurant, it likely to have more checkins customers.

The result of Exploratory, we find the top five cities by the number of checkins, the relationship average star of each restaurant versus checkins and the number of reviews versus checkins are very encouraging that number reviews and average star give will be a strong relationship with checkins(customers) each restaurant.

Methods

The regression models are incredibly handy statistical tools used to answers all sorts of questions according to Prof. Caffo. In the project two of three most common tasks for regression models were used:

- **Modeling** try to find a parsimonious described relationship of total number of checkins with other possible regressors such as average star rating, number of reviews, price range, so on
- **Covariation** investigate the variation (residual variation) in checkins which appears unrelated to regressors.
- **Modelling Multiple Linear Regression** model to find the relationship between reviews, average stars and its price range with checkins

In fact, there are more variable factors other than type of **stars** and **review_count** such as **good4kids, good4groups, pricerange, reservations, open24h**, and **lifespan** which may be good candidates to include in regression model. we will use the a statistical multiple linear regression model:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik}\beta_j + \epsilon_i$$

Where β_n slopes, Xni are variables and ϵ_i residuals which are normally distributed

Variables

With final business dataset mentioned above, We can easily chose variables to be used for our regression model: (**review_count, stars, checkins, good4kids, good4groups, pricerange, reservations, open24h, start_date, end_date, funny, cool, useful** and **lifespan**)

High correlation variables

We use the “cor” function in R to do further checking all variables and find out review_count has very high correlation votes (cool, useful, funny) respectively 94%, 94% and 93%. It means that when customer writes the review for each restaurant in Las Vegas and most likely they will vote for that restaurant. Therefore we can eliminate the high correlation variables (cool, useful, funny) with checkins in regression model.

The Selection Strategy

To select the best fitting model to describe the relationship between checkins and restaurant star rating. We use stepwise regression to quickly eliminate the big number of possible models and select few good models.

Applying function step() to this full model for both(forward,backward) stepwise regression we got the best regressions is (**checkins** ~ **review_count** + **stars** + **good4kids** + **good4groups** + **pricerange**) which each variable has 3 star(***). However we observe that the standard errors for both variables (**good4group** and **good4kids**) much higher than **review_count**, **stars** and **pricerange**. Therefore we can eliminate **good4group** and **good4kids** out of regression model to avoid the problem of multicollinearity.

We do further checking R-squared of lm() function to model (**checkins** ~ **review_count** + **stars** + **good4kids**), (**checkins** ~ **review_count** + **stars** + **good4groups**) and (**checkins** ~ **review_count** + **stars** + **pricerange**), found R-squared is respectively 0.7983, 0.7934 and 0.8019 while all P-value is smaller than 0.05, hence we will chose model with highest R-spared and the final model is (**checkins** ~ **review_count** + **stars** + **pricerange**)

Results

With Exploratory Restaurant Data explore, We observe that the checkins has relationship with review count and stars rating of each restaurant in Las Vegas. We use the Multiple Linear Regression to analysis and find out it is true there is a strong relationship between checkin customers and review count and star rating. We also find out the business attributes price range has some impacting to model, therefore we include it to improve the final model.

In fact, It is quite logic. The tourists are most likely to check all reviews, number of reviews, star rating, price range before they decide to checkin restaurant in Las Vegas City and we have coefficient as below.

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    62.187086 22.79869369   2.727660 6.406178e-03
## review_count    2.674459  0.02174447 122.994895 0.000000e+00
## stars          22.151974  3.52923511   6.276707 3.823810e-10
## pricerange    -93.197174  6.96322259 -13.384201 5.287613e-40
```

The Result of regression model is

$$C_i = 62.18709 + 2.67446 * Ri + 22.15197 * Si - 93.19717 * Pi$$

Where: Ci is checkins(customer) of i restaurant, Ri is review_count of each restaurant, Si is star of each restaurant, Pi is price range

The slopes of the final model showed that total number of checkins at a restaurant in Las Vegas is expected to change by almost 3 units (2.67446 to be exact) per unit change in total number of reviews on this restaurant at Yelp website, holding other regressors fixed. Similarly, when all other variables unchanged, change one unit in average star rating or price range lead to change of total checkins by 22 or 93 units respectively.

Furthermore for the model slopes, theirs confidence intervals with 95% confidence contain no zero, it mean there were relationships between total number of checkins (response) with each regressor.

```
##                    2.5 %    97.5 %
## (Intercept)    17.489040 106.88513
## review_count    2.631827   2.71709
## stars          15.232721  29.07123
## pricerange   -106.848941 -79.54541
```

Based on the intervals above, we say that “with 95% confidence, we estimate that the number of review increases results in a 2.631827 to 2.71709 increase in total checkins at a restaurant in Las Vegas, holding other variables fixed”. And other intervals also can be interpreted similarly.

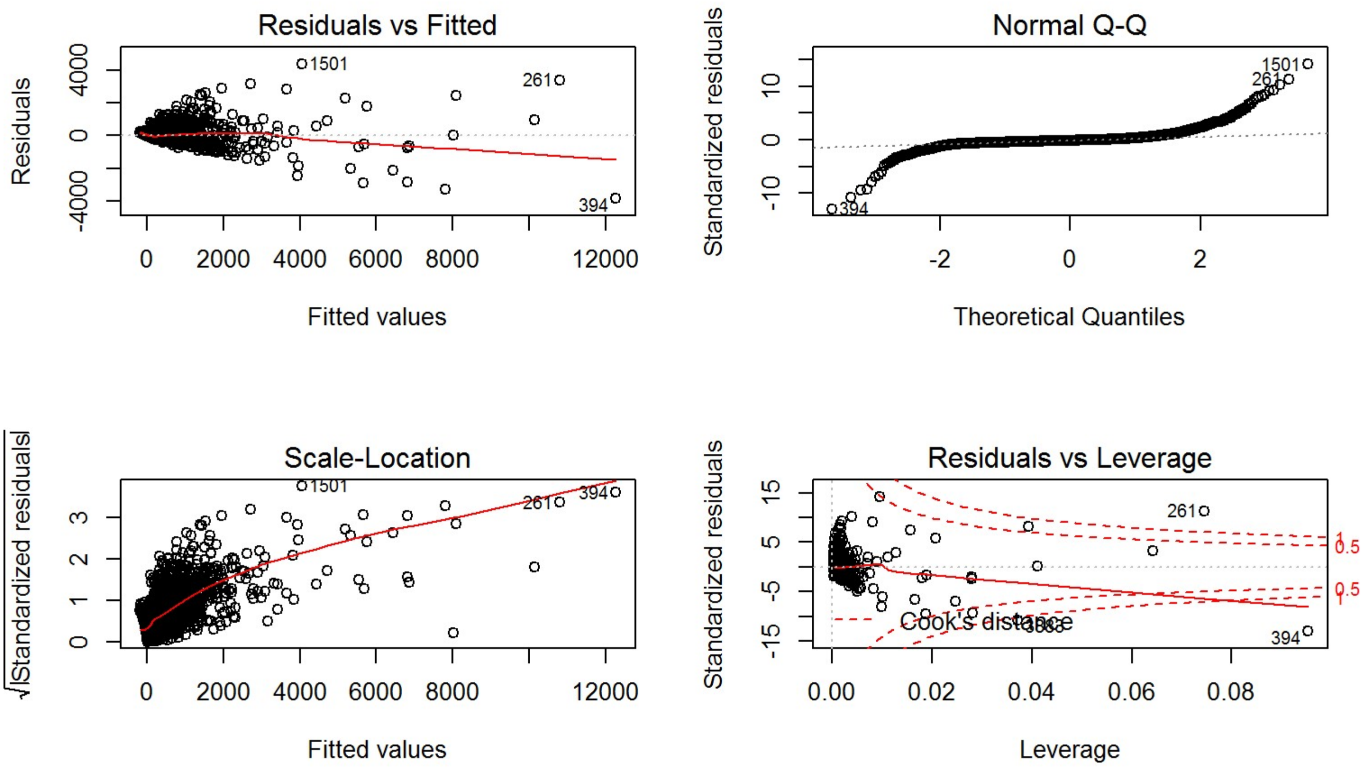
Residuals and Diagnostics

The model’s R-squared is 0.8019, Adjusted R-squared 0.8017 and and Variance Inflation Factors (VIF) showed that this model has a good fit: VIF for all variables are in acceptable range of 1.034-1.11. According to Prof. Caffo, VIF “measure how much variance inflation the variable causes to the setting where it was orthogonal to the other regressors”, For all variables, theirs individual inclusion into the model only caused variance inflation from 3% to 11%.

```
## review_count        stars   pricerange
##     1.114638     1.034122     1.096334
```

The plot of residuals versus fitted values shows that residuals are randomly located above and below 0 and the mean residuals is

-8.923187e-15 (close to 0) and the normal quantile plot of the standardized residuals shows the diagonal line which is a sign of normality of errors. It proved that our assumption that residuals should normally distributed when fitting linear regression model to the data was met.



Based on Cook’s distance plot, there 2 data points (observation 261 and 394) having the most influence with Cook’s distance value around 1. There 2 data points have total number of checkins 14203 and 8416, price ranges of 1 and 2. Using dfbeta() to evaluate how much parameter estimates changed if these 2 outliers were dropped from the data set, it found that except for Price Range slopes which can be changed by around 10%, most of change to others slope estimates are around 1% or less. So it can be concluded that the final model is quite robust regarding outliers data points.

Discussion

Inference

The results of above analysis gave **clear answer to the primary question of interest**, both qualitatively and quantitatively. The linear regression model established that total number of checkins customers in to a Las Vegas restaurant increased with increasing number of reviews on restaurant improvement of the restaurant star rating and its price range. Base on model slopes, we can estimate the increasing 3 reviews(exact 2.67446) number of reviews, it will bring in 1% more checkin customers, more number of reviews expects more customers. Similarly when rating star increases by 0.5 star rating, it will probably have 22 (exact 22.15197) different increase number of checkins customers and when price range decreases by 1 range we also have one 93 more customers checkin with 95% confidence.

Conclusion

The Checkins customers of restaurant in Las Vegas has strong relationship with number of reviews from customer, star rating of restaurants and its price range. It also means checkins customers will base on the number of reviews, stars and its price range before they decide. The sustainability of a restaurant can be measured by numbers of customers(checkins), more customers it means business will be sustainable and success.

References: Regression Models for Data Science in R - Prof. Brian Caffo (2015)
Source codes (https://github.com/hhdinhgithub/Capstone)