# Eat, Rate and Love — An Exploration of Yelp

Gurui Huang, Boyan Wu, Maoyi Song, Hanyi Chen
MSBA324 Web and Social Analytics | Instructor: Nabanita Talukdar

## INTRODUCTION

Yelp is a company that published crowd-sourced reviews about local business and it also provides online reservation service (Yelp about us, 2018). User may use Yelp application to search local business, such as restaurants or schools and use 5 starts rating system to submit their reviews.

**Data Clean**- Process missing data and use SQL to build new table to analyze.

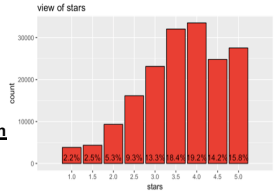
**Overall Analysis** – Descriptive Analysis about "Stars" and "Reviews".

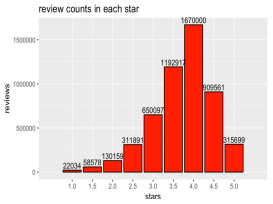
**Regression Analysis** – Using ANOVA and T-test to find out if independent variable "check in numbers" has difference when meeting different dependent variables.

## OVERALL ANALYSIS

**1. High star rating business exploration & suggestion.**

why stars important:




| Stars | Review Rate |
|---|---|
| 1.0 | 0.4% |
| 1.5 | 1.1% |
| 2.0 | 2.5% |
| 2.5 | 5.9% |
| 3.0 | 12.4% |
| 3.5 | 22.7% |
| 4.0 | 31.7% |
| 4.5 | 17.3% |
| 5.0 | 6% |

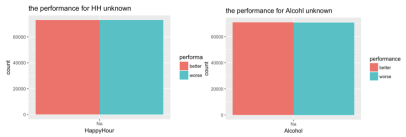
Insight: business with higher star attracts more reviewers, and may has more clients in general;

**2. Exploration on business attributes to find out possible improvement direction.**

- Attribute selection from data exploration: attributes selected as number of validate data greater than 5000 to be meaningful;

| Variable Name | Available Data | Remark |
|---|---|---|
| "dog allowed" | 6,005 | Pets friendly |
| "Happy Hour" | 6,182 | |
| "Credit Card Acceptance" | 23,581 | Payment method |
| "wheelchair accessible" | 20,947 | Assistance facility |
| "alcohol" | 10,412 | |

- Metric determination: assign "better" (star greater than 3) and "worse" (star 3 or less) performance to business.
- NA's affection: difference in performance (stars) will affect the interpretation's power.




Insight:
No effect groups: "Happy Hour" and "Alcohol".
Effect exist group: "pets friendly", "assistant facility", "accept credit card".

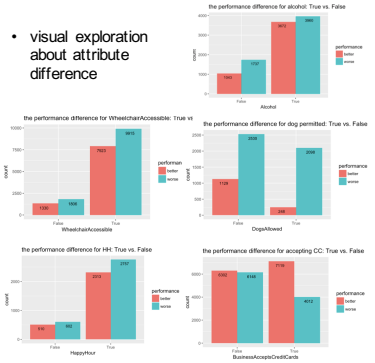


- T-test to verify if missing data

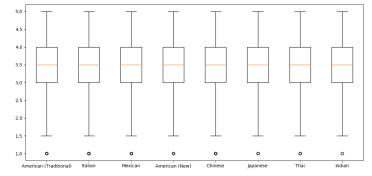
Null Hypothesis: No star's difference in attribute unknown business.

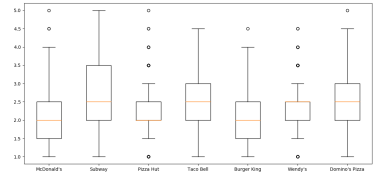
| Attributes: | P-value | Conclusion |
|---|---|---|
| HH | 0.97 | Fail to reject null hypothesis; no difference |
| Alcohol | 0.25 | |
| Wheelchair Accessibility | 1.18e-7 | Reject null hypothesis; difference exists |
| Dog Allowed | 1.59e-12 | |
| Credit Card Accepts | 2.2e-16 | |

- visual exploration about attribute difference




## SPECIFIC CASE ANALYSIS




- T-test (star & check-in database)

| Attributes | Conclusion |
|---|---|
| MC VS Subway<br><br>T-statistic: -13.856<br>P-value: 7.38 | H0: McDonald's star is lower than Subway. Accept H0. The result is significant, so McDonald's stars have significant difference with Subway. |
| MC VS Burger King<br><br>T-statistic: -1.57<br>P-value: 0.12 | H0: MCs star is lower than Burger King. Reject H0. The result is not significant; different exist |
| Weekdays VS Weekend<br><br>p-value < 2.2e-16 | H0: the check-in numbers between weekdays and weekends are the same. Reject H0; different exist |

## REGRESSION ANALYSIS

★ Model 1: checkins~hour+weekday (original data, numeric)
Model 2: checkins~hour+weekday (cleaned data, string)

```
Call:
lm(formula = checkins ~ hour + weekday, data = df4)

Residuals:
    Min      1Q  Median      3Q     Max
 -6.04   -3.22   -2.23   -0.60 1471.96

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.54404    0.03142 144.644  < 2e-16 ***
hour1:00     0.86640    0.03876  22.354  < 2e-16 ***
hour10:00    -1.83436   0.08566 -21.414  < 2e-16 ***
hour11:00    -2.08651   0.07401 -28.194  < 2e-16 ***
hour12:00    -1.94857   0.06248 -31.190  < 2e-16 ***
hour13:00    -1.92195   0.05401 -35.583  < 2e-16 ***
hour14:00    -1.61303   0.04806 -33.566  < 2e-16 ***
hour15:00    -1.45832   0.04311 -33.832  < 2e-16 ***
hour16:00    -1.21893   0.03990 -30.534  < 2e-16 ***
hour17:00    -1.23475   0.03840 -32.156  < 2e-16 ***
hour18:00    -0.70039   0.03750 -18.673  < 2e-16 ***
hour19:00    0.01715    0.03743   0.458  0.64697
hour2:00     1.34523    0.04074  33.023  < 2e-16 ***
hour20:00    -0.38530   0.03766 -10.231  < 2e-16 ***
hour21:00    -0.81377   0.03750 -21.701  < 2e-16 ***
hour22:00    -0.86436   0.03723 -23.220  < 2e-16 ***
hour23:00    -0.60084   0.03721 -16.141  < 2e-16 ***
hour3:00     1.05528    0.04350  24.260  < 2e-16 ***
hour4:00     0.55441    0.04758  11.653  < 2e-16 ***
hour5:00     0.12418    0.05388   2.490  0.01277 *
hour6:00     -0.09071   0.06193  -1.465  0.14305
hour7:00     -0.62109   0.06855  -6.143 8.11e-10 ***
hour8:00     -0.97171   0.07696 -12.626  < 2e-16 ***
hour9:00     -1.49585   0.08559 -17.474  < 2e-16 ***
weekday1     -0.08201   0.02679  -3.061  0.00219 **
weekday6     0.93862    0.02501  37.528  < 2e-16 ***
weekday7     1.14668    0.02590  44.271  < 2e-16 ***
weekday4     -0.32355   0.02615 -12.451  < 2e-16 ***
weekday2     -0.38869   0.02645 -14.697  < 2e-16 ***
weekday3     -0.37428   0.02624 -14.262  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = checkins ~ hour + weekday, data = df3)

Residuals:
    Min      1Q  Median      3Q     Max
 -5.02   -3.00   -2.66   -0.66 1471.98

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.75660    0.01319  360.51  <2e-16 ***
hournoon      -2.02374    0.02066  -97.94  <2e-16 ***
hournight     -1.09022    0.01980  -69.10  <2e-16 ***
weekdayweekend  1.26693    0.01517   83.49  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.9 on 3911214 degrees of freedom
Multiple R-squared:  0.004636,  Adjusted R-squared:  0.004635
F-statistic:  5941 on 3 and 3911214 DF,  p-value: < 2.2e-16
```

ANOVA

```
Analysis of Variance Table

Model 1: checkins ~ hour + weekday
Model 2: checkins ~ hour + weekday
  Res.Df     RSS  Df Sum of Sq     F    Pr(>F)
1 3911188 754540324
2 3911214 755639651 -26 -1099326 219.17 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In Model1 both main effects and the interaction are significant. However, most of the coefficients are smaller than 1. Then, we simplified the model (Model2) to weekdays, weekend morning, noon and night.
- Results show the both model are significant. ANOVA compared. P<0.05 Model 1 is the best fit.

## CONCLUSION

| Attributes | Conclusions | Suggestion |
|---|---|---|
| "Dog Allowed" | Large loss in better group when businesses are considered to be pets friendly. | Tips for business owner: it may not be a good idea to allow dogs in order to achieve higher star. |
| "Happy Hour" | More business tends to have HH; the difference for two levels increases when business have "Happy Hour" | Tips for business owners if HH applicable: HH might not result in increasing the chance to gain higher star level; Test design needed by Yelp to design promotions/make suggestions relating to having HH for business. |
| "Credit Card Acceptance" | Large loss in better business group measured for business dose accept credit card. | Tips for business owners: it may be a good idea to accept credit card in order to achieve higher star. |
| "Wheelchair accessible" | More business tends to have wheelchair friendly environment; the difference for two levels increases when business do have wheelchair accessible environment. | Not too much business meaning. |
| "Alcohol" | Difference of performance decreasing significantly as having the alcohol. | It is a good idea to have alcohol served in business to increase the chance |
| "Check-in" | Check-in numbers are negative correlation with weekdays, but positive correlation with weekends. | May be good idea to build brand effect, word of mouth; increasing the number of promotion and customer benefit during weekend. |

Reference
Yelp dataset Retrieved from: https://www.kaggle.com/yelp-dataset/yelp-dataset.
Yelp about us (2018) Retrieved from https://www.yelp.com/about