

## Prerequisites for a high-rating restaurant

Group Member: Tianqi Jiang, Yifan Xie, Bonan Yuan

### 1. Research agenda

On **October 10<sup>th</sup>**, the three members in our group have gathered together and discussed the proposal for our project. We unified our opinions that we want our project to be realistically applicable which means that our project should be useful in real scenarios and able to create values or benefits. We did brainstorming and each of us came up with few proposed projects by focusing on the urban issues. By considerate comparisons, we chose to make further exploration with restaurants ratings because people always consider the restaurant rating as an import initiative for them to dine there. For long time, rating is deemed as subjective comment by the customer such as tasting and serving, but we would like to explore more beyond those subjective things and be able to predict the ratings with other prerequisites such as location and category of a restaurant.

On **Oct 12<sup>th</sup>**, two days after two days after we proposed our initial plan for the project, we started our first step of the project: data gathering. We would like to find out data available and data unavailable and then make changes to modify our project based on the data gathering condition. The data gathering process will then be further discussed in the Data Gathering section of this paper.

On **Nov 7<sup>th</sup>**, a week before the final project due date, we started our data clean process. The raw data is a JSON file with huge amount of information in every aspect of

a restaurant in dictionary format. Therefore, we need to clean the data to get the information we need for our analysis and then transform the data in appropriate data frame that we can use.

On **Nov 11<sup>th</sup>**, we finished writing the python script for data cleaning and started using R for further analysis.

On **Nov 13<sup>th</sup>**, we used R to select the correct regressors that will affect our ratings and build the regression model to evaluate the prediction.

On **Nov 15<sup>th</sup>**, we finalized the model and write final report for our regression model.

## **2. Data Gathering**

After we decided to focus our project on the restaurant ratings, the ideal data source that we thought about is Yelp. We should be able to write script to acquire the data through Yelp public API [1]. However, when we tried to acquire the data using API portal, the API would not return any restaurant data if we tried to set the limit of shops for certain zip code above 20. However, if we only use the restaurant's data for the top 20 restaurants in certain area, we could not generate an appropriate regression model for rating prediction because the top data itself is biased. [Fig 10]

Therefore, we need more complete and unbiased data to conduct our further analysis. After a complete research of the Yelp datasets available online, we found a competition held by Yelp called 'Yelp Dataset Challenge', In this dataset, Yelp provides a complete dataset that contains Restaurant Name, categories, hours, stars(ratings), price

range, noiselevel, delivery and full address of that restaurant for many major cities in United States [2].

From the ‘Yelp Data Challenge’ dataset, we selected Pittsburgh as our targeted city. Aside from ‘Yelp Data Challenge’, we need some more demographic data such as population and average income by zip code in Pittsburgh to assist our research. We use the dataset from zipatlas for the demographic data. After a complicated process of cleaning and merging, we filtered out the data we need and combined yelp datasets and demographic dataset together. Finally, we outputted the data to csv and started using R for model analysis.

### **3. Hypothesis**

Our Hypothesis is that by knowing the prerequisites such as location of the restaurant, category of the cuisines and other variables, we will be able to predict the ratings of the restaurant. We picked up the following variables that will hypothetically affect our regression model:

1. Categories of cuisines
2. Open time & close time of the restaurant
3. Price range( measured as 1 to 4 for \$ to \$\$\$\$ , \$: 0-\$10, \$\$: \$10-\$30, \$\$\$: \$30-\$50, \$\$\$\$: >\$50)
4. Noise level(quite, average, loud, very loud)
5. Delivery status (yes/no)
6. Status of average income where the restaurant located
7. Population where the restaurant located

8. Competition (calculated number of restaurants of same category in certain area)

We will dump all the variables into regression model and test if they are validated regressors in our regression model.

## **4. Methodology**

### **4.1 More Data Cleaning:**

The raw dataset provided by yelp is not quite clear with the categorization. For example, a restaurant could have a category called [Bars", "American (Traditional)", "Nightlife", "Lounges", "Restaurants"]. However, it is not the category we want for our regression model. Therefore, we wrote a python script to re-categorize the category for the restaurant. We name 14 categories for our test target restaurants. They are American, Burgers, Chinese, Deli, Italian, Japanese, Mexican, Middle Eastern, Other, Pizza, Sandwiches, Seafood, Steakhouses, Thai. And we got the pie plot of all the categories [Fig 1]. Since category 'other' is not representative, we filtered out the datasets with 'other' as their categories.

Another altering needs to be done with datasets is discarding the restaurants ratings with few reviews. A rating with only few reviews is highly biased, which should be considered as outliers in analysis. According to the histogram of frequency of reviews ,we cleaned out the restaurants with reviews lower than 5. After a series of data cleaning, we finally generated our clean data with 703 sets of data in 13 categories.

## **4.2 Regressors selection**

### **4.2.1 Round 1 Selection**

Now we have 8 regressors, which are hypothetically related to our regression model. However, not every hypothesized regressor will mathematically fit into our regression model. In R, we use Generalized Additive Model (GAM) and run the ANOVA analysis with all 9 regressor. From the parametric terms [Fig 3], we select the regressors with P-Values  $< 0.05$  or around  $0.05$ , which are competition, categories, delivery, noise level, price range and population. And we rejected other 3 regressors with P-values  $> 0.05$ , which are open&close time and average income.

### **4.2.2 Round 2 Selection**

Now we have 6 regressors left. Once again, we throw them into the GAM and run our analysis. This time, as we expected, every regressor corresponds to a P-Value  $< 0.05$ [Fig 4]. However, many regressors in our regression are not quantifiable variables but levels such price range, categories and noise level. Therefore, we need to look into regression intensity enforced by each level of each regressor. From the summary of regression model [Fig 5], we notice that although the price range does have effect on our model, there is no significant difference between different levels of price range. Therefore, we can remove price range as one of our regressors because different levels of price range do not make difference to the final rating.

## **4.3 Final model and validation**

After the process of regressor selection, we finalize our model with 6 regressors, which are competition, category, noise level, delivery and population[Fig 6]. For validation, since there are only 703 sets of data left for Pittsburgh after data cleaning, if

we divide the data into test set and validation set, we would not enough data sets to evaluate our model. Therefore, we have to use histogram of residuals and normal QQ plot to visually validate our regression model. From the graphs [Fig 9], we can observe that histogram of residual roughly follows the normal distribution and QQ -plot is approximately linear. Therefore, the two graphs demonstrate that our regression model is valid.

## **5. Results**

Our final model contains 5 regressors[Fig 6]. Here, we can interpret each variable and understand how this variable would affect the final rating of a restaurant by using the summary [Fig 7] .

### **5.1 Competition**

The coefficient for the variable competition is  $-0.00814$ , which mean more restaurants of same category of cuisine in certain area will lower the ratings. For example, if someone opens a Thai restaurant where this area already has so many Thais, then the new restaurant will probably get lower rating. This regressor is easy to interpret since more competition makes customers pickier on their tastings and cavil when rating.

### **5.2 Category**

It is quite obvious that people in a city have their unique preference on certain category of cuisine. In Pittsburgh, it is definitely Mid-Eastern Food. From the summary[Fig 7], Mid-Eastern Food has coefficient of  $0.3989$ , which means if someone opens a Mid-Eastern Food restaurant, the restaurant will have  $0.3989$  points as a plus in rating comparing to restaurants in other categories.

### **5.3 Noise Level**

From the summary [Fig 7], the noise does lower the ratings for restaurant and the louder the noise is, the more points are going to be deducted from the rating. It looks like people from Pittsburgh do prefer quiet dining environments.

### **5.4 Delivery**

From the summary [Fig 7], if the restaurant provides delivery, delivery service will lower their rating by 0.179. It is easy to understand because no matter how tasty the food originally was, after the process of delivery, it never tastes as well as it was when it came right out of kitchen. And raters who order deliveries are either busy or tired going out, which will be another factor to deteriorate the rating of the restaurant providing delivery service.

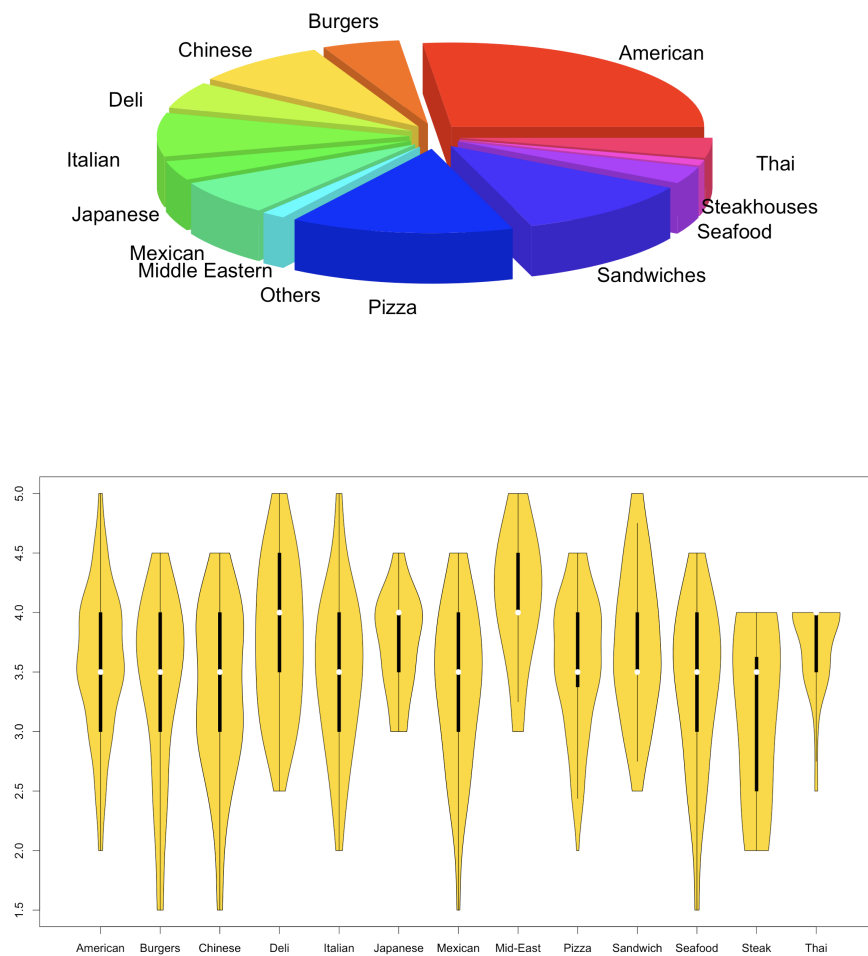
### **5.5 Population**

From the summary [Fig 7], the coefficient of population is  $-5.474\text{e-}06$ . It looks like a small coefficient. However, after multiplying the population, it will also become an important factor. Therefore, if someone wants to open a high rating restaurant, suburbs maybe a good choice because of the less population there.

# Appendix

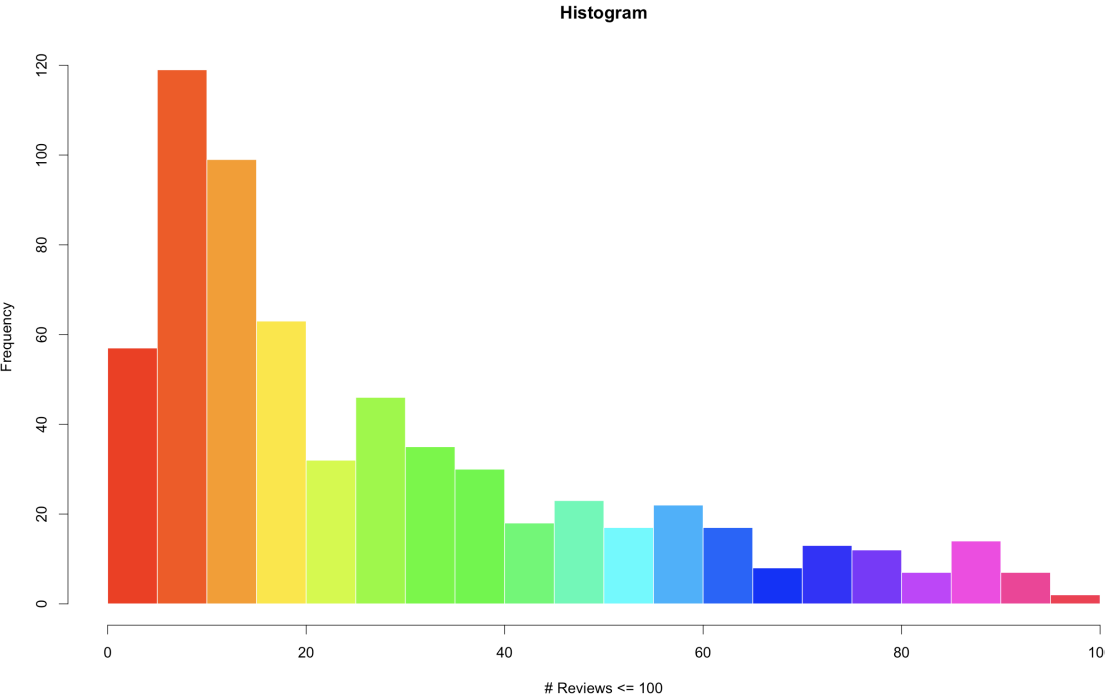
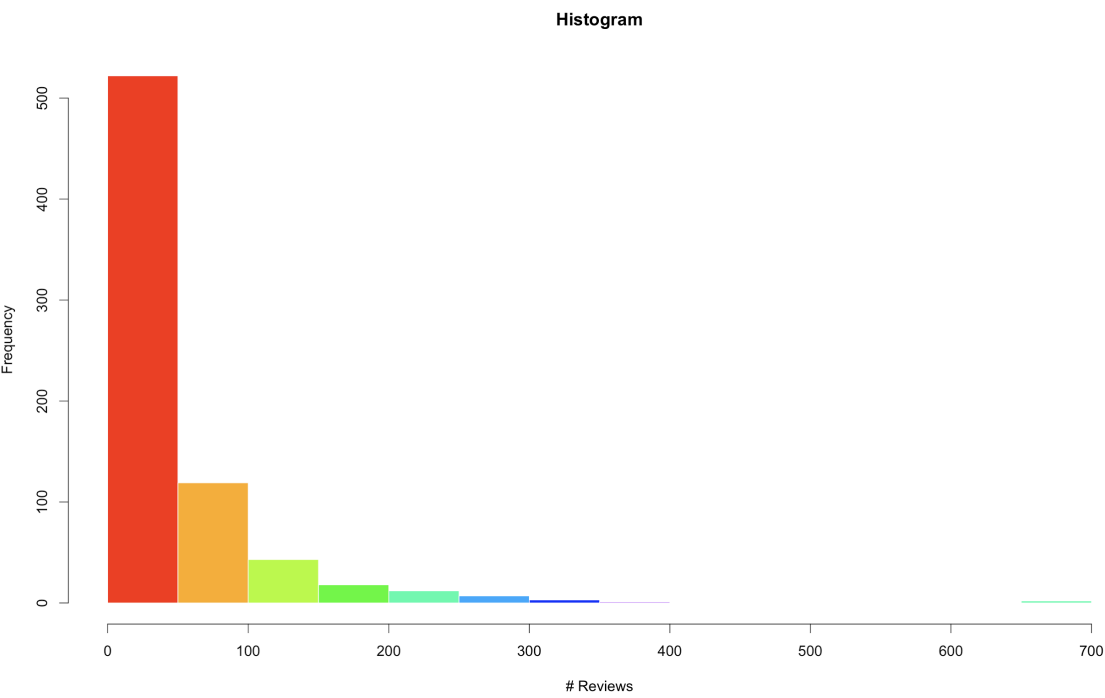
## Graphs:

[Fig 1] Pie plot of categories and violin plot of stars vs. categories





[Fig 2] Histogram of Frequency of reviews



[Fig 3] Generalized Additive Model (GAM)

ANOVA analysis

Formula:

```
stars ~ Competitorcount + factor(categories) + factor(opentime) +  
      factor(closetime) + factor(priceRange) + factor(noiselevel) +  
      factor(delivery) + avg_Income + population
```

Parametric Terms:

|                    | df | F     | p-value  |
|--------------------|----|-------|----------|
| Competitorcount    | 1  | 4.180 | 0.04130  |
| factor(categories) | 12 | 3.733 | 1.75e-05 |
| factor(opentime)   | 27 | 1.179 | 0.24403  |
| factor(closetime)  | 29 | 0.872 | 0.66199  |
| factor(priceRange) | 4  | 2.162 | 0.07178  |
| factor(noiselevel) | 3  | 2.785 | 0.04006  |
| factor(delivery)   | 1  | 6.727 | 0.00971  |
| avg_Income         | 1  | 0.474 | 0.49137  |
| population         | 1  | 3.772 | 0.05254  |

[Fig 4] revised model

Formula:

```
stars ~ Competitorcount + factor(categories) + factor(priceRange) +  
      factor(noiselevel) + factor(delivery) + population
```

Parametric Terms:

|                    | df | F     | p-value  |
|--------------------|----|-------|----------|
| Competitorcount    | 1  | 7.642 | 0.00585  |
| factor(categories) | 12 | 4.388 | 8.82e-07 |
| factor(priceRange) | 4  | 3.778 | 0.00476  |
| factor(noiselevel) | 3  | 3.643 | 0.01255  |
| factor(delivery)   | 1  | 7.393 | 0.00671  |
| population         | 1  | 4.632 | 0.03172  |

[Fig 5]Model Summary

Formula:

stars ~ Competitorcount + factor(categories) + factor(priceRange) +  
factor(noiselevel) + factor(delivery) + population

Parametric coefficients:

|                      | Estimate   | Std. Error | t value | Pr(> t ) |     |
|----------------------|------------|------------|---------|----------|-----|
| (Intercept)          | 4.058e+00  | 3.275e-01  | 12.392  | < 2e-16  | *** |
| Competitorcount      | -9.093e-03 | 3.290e-03  | -2.764  | 0.005853 | **  |
| factor(categories)1  | -2.644e-01 | 1.182e-01  | -2.237  | 0.025611 | *   |
| factor(categories)10 | -3.658e-01 | 1.385e-01  | -2.640  | 0.008463 | **  |
| factor(categories)11 | -8.499e-01 | 2.440e-01  | -3.484  | 0.000525 | *** |
| factor(categories)12 | 8.944e-02  | 1.323e-01  | 0.676   | 0.499302 |     |
| factor(categories)2  | -2.394e-01 | 1.023e-01  | -2.339  | 0.019624 | *   |
| factor(categories)3  | 9.693e-02  | 1.209e-01  | 0.802   | 0.422765 |     |
| factor(categories)4  | -1.534e-01 | 9.484e-02  | -1.618  | 0.106216 |     |
| factor(categories)5  | 8.367e-02  | 1.308e-01  | 0.639   | 0.522707 |     |
| factor(categories)6  | -2.668e-01 | 1.042e-01  | -2.561  | 0.010637 | *   |
| factor(categories)7  | 4.369e-01  | 1.954e-01  | 2.236   | 0.025666 | *   |
| factor(categories)8  | 3.758e-02  | 8.817e-02  | 0.426   | 0.670051 |     |
| factor(categories)9  | 9.923e-02  | 8.818e-02  | 1.125   | 0.260818 |     |
| factor(priceRange)1  | -7.239e-02 | 3.115e-01  | -0.232  | 0.816275 |     |
| factor(priceRange)2  | -1.633e-01 | 3.120e-01  | -0.524  | 0.600772 |     |
| factor(priceRange)3  | -1.867e-02 | 3.234e-01  | -0.058  | 0.953968 |     |
| factor(priceRange)4  | 6.260e-01  | 3.819e-01  | 1.639   | 0.101625 |     |
| factor(noiselevel)1  | -1.234e-01 | 5.848e-02  | -2.110  | 0.035190 | *   |
| factor(noiselevel)2  | -1.595e-01 | 8.765e-02  | -1.820  | 0.069134 | .   |
| factor(noiselevel)3  | -4.504e-01 | 1.514e-01  | -2.974  | 0.003038 | **  |
| factor(delivery)1    | -1.691e-01 | 6.217e-02  | -2.719  | 0.006708 | **  |
| population           | -5.475e-06 | 2.544e-06  | -2.152  | 0.031724 | *   |

---

[Fig 6] Final Revised Model

Formula:

stars ~ Competitorcount + factor(categories) + factor(noiselevel) +  
factor(delivery) + population

Parametric Terms:

|                    | df | F     | p-value  |
|--------------------|----|-------|----------|
| Competitorcount    | 1  | 7.284 | 0.00712  |
| factor(categories) | 12 | 4.021 | 4.66e-06 |
| factor(noiselevel) | 3  | 4.629 | 0.00325  |
| factor(delivery)   | 1  | 8.203 | 0.00431  |
| population         | 1  | 4.578 | 0.03272  |

Final Revised model Summery [Fig 7]

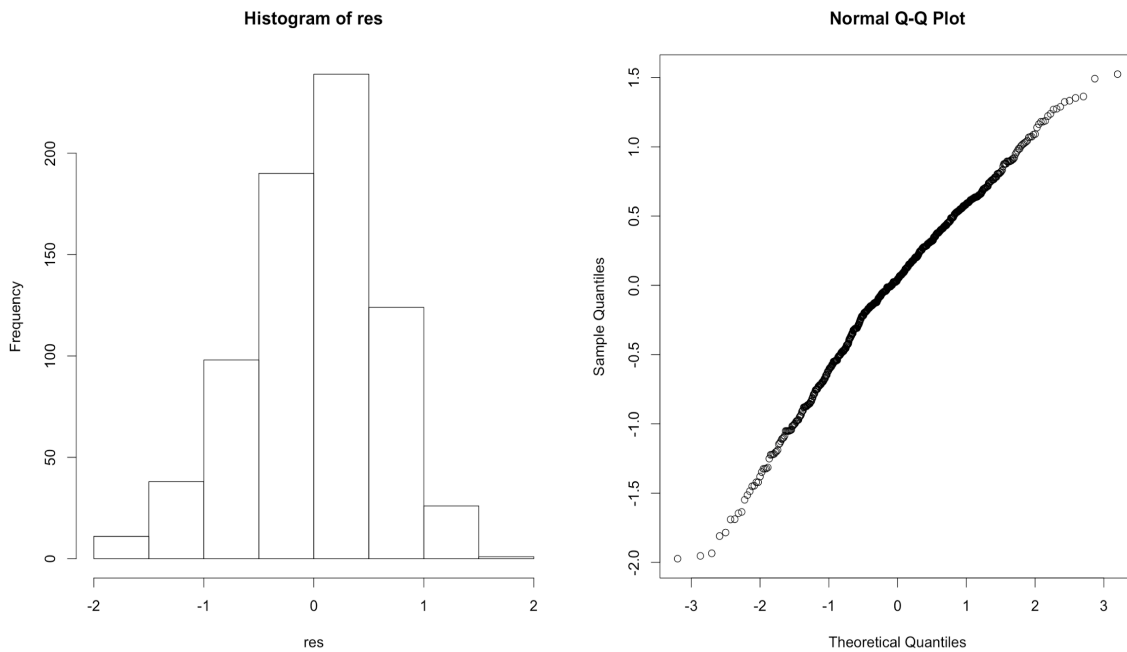
Formula:

stars ~ Competitorcount + factor(categories) + factor(noiselevel) +  
factor(delivery) + population

Parametric coefficients:

|                      | Estimate   | Std. Error | t value | Pr(> t ) |     |
|----------------------|------------|------------|---------|----------|-----|
| (Intercept)          | 3.970e+00  | 1.015e-01  | 39.103  | < 2e-16  | *** |
| Competitorcount      | -8.914e-03 | 3.303e-03  | -2.699  | 0.00712  | **  |
| factor(categories)1  | -2.748e-01 | 1.182e-01  | -2.325  | 0.02035  | *   |
| factor(categories)10 | -3.266e-01 | 1.382e-01  | -2.363  | 0.01841  | *   |
| factor(categories)11 | -5.167e-01 | 2.273e-01  | -2.274  | 0.02329  | *   |
| factor(categories)12 | 3.622e-02  | 1.320e-01  | 0.274   | 0.78387  |     |
| factor(categories)2  | -2.535e-01 | 1.019e-01  | -2.488  | 0.01307  | *   |
| factor(categories)3  | 1.224e-01  | 1.165e-01  | 1.050   | 0.29405  |     |
| factor(categories)4  | -1.410e-01 | 9.462e-02  | -1.490  | 0.13658  |     |
| factor(categories)5  | 9.223e-02  | 1.317e-01  | 0.700   | 0.48402  |     |
| factor(categories)6  | -2.712e-01 | 1.031e-01  | -2.629  | 0.00875  | **  |
| factor(categories)7  | 3.989e-01  | 1.963e-01  | 2.032   | 0.04254  | *   |
| factor(categories)8  | 4.985e-02  | 8.545e-02  | 0.583   | 0.55985  |     |
| factor(categories)9  | 1.198e-01  | 8.270e-02  | 1.449   | 0.14781  |     |
| factor(noiselevel)1  | -1.538e-01 | 5.821e-02  | -2.642  | 0.00844  | **  |
| factor(noiselevel)2  | -1.974e-01 | 8.739e-02  | -2.259  | 0.02418  | *   |
| factor(noiselevel)3  | -4.772e-01 | 1.514e-01  | -3.152  | 0.00169  | **  |
| factor(delivery)1    | -1.790e-01 | 6.250e-02  | -2.864  | 0.00431  | **  |
| population           | -5.474e-06 | 2.558e-06  | -2.140  | 0.03272  | *   |

[Fig 8] Histogram of residuals and Normal QQ Plot




[Fig 9] Min, Median, Mean of the cleaned dataset

```
> summary(yelp$stars)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.500   3.000   3.500   3.572   4.000   5.000
>
```

[Fig 10] Yelp API

### Yelp API

- ✓ Search over 50 million local businesses from 32 countries
- ✓ Enhance your app with Yelp ratings, reviews, photos and much more
- ✓ Simple and fast API with powerful category and geo search filters



#### Yelp Developers

- Manage API access
- API console
- Documentation
- Display requirements
- Support group
- Code samples

#### API v2.0

|                 |                                  |
|-----------------|----------------------------------|
| Consumer Key    | 0q-vxnYX0xFTg33vtHb4Lg           |
| Consumer Secret | vHFM9wCl6tqrhOrfBxdoX8zeqA       |
| Token           | ZFalrkQTzLLqr6nCAIQ_UR7_srTN5r6w |
| Token Secret    | bdmZe_Xrf4nNXStQGy6EdSwq85l      |

[Generate new API v2.0 token/secret](#)

[Fig 11] Yelp Dataset Challenge

## Yelp Dataset Challenge

### Yelp Dataset Challenge rides again! Round 6 is here.

We've had 5 rounds, over \$35,000 in cash prizes awarded, *hundreds of academic papers written*, and we are excited to see round 6.

Our dataset for this iteration of the challenge is the same as the last iteration - we're sure there are plenty of interesting insights still waiting there for you. If you want the latest check-ins and reviews, don't worry, we'll have them for you in 2016 (along with some new attributes if you're good). This set includes information about local businesses in 10 cities across 4 countries. This treasure trove of local business data is waiting to be mined and we can't wait to see you push the frontiers of data science research with our data.



### Reference:

1. Yelp API: [https://www.yelp.com/developers/manage\\_api\\_keys](https://www.yelp.com/developers/manage_api_keys)
2. Yelp Dataset Challenge: [http://www.yelp.com/dataset\\_challenge](http://www.yelp.com/dataset_challenge)
3. Demographic average income data:  
<http://zipatlas.com/us/pa/pittsburgh/zip-codecomparison/median-household-income.htm>