Business Help From Yelp※

Insights into Yelp's Open Dataset
Jan 2020

David Yu

Data Source: Yelp Open Dataset
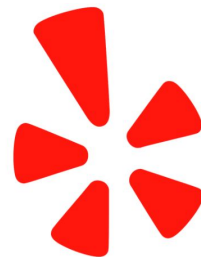
6,685,900 reviews

192,609 businesses

200,000 pictures
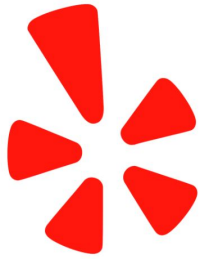
10 metropolitan areas

1,223,094 tips by 1,637,138 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 192,609 businesses
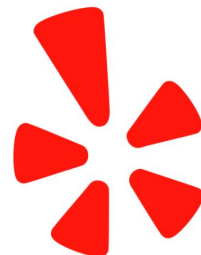
Source: yelp.com/dataset

Executive Summary

1. Elite users have **different** user engagements

2. Funny Feedback = Cool Feedback

3. There is a 10am Slump in ratings and review word dip

Summary Yelp Open Dataset Used:

File Name	Number of Entries	Attributes
business.json	192609	names, stars, reviews_count, city, state, attributes, categories
checkin.json	161950	business_id, dates
photo.json	200000	caption, label
⭐ review.json	5376719	review_id, user_id, business_id, stars, useful, funny, cool, text, date
tip.json	1223094	text, date, compliment_count
⭐ user.json	1637138	review_count, useful, funny, cool, fans, avg_stars, compliment_hot ...

Users Are Important

Dataset **users.json** contains **1.6 Million** Entries

Users write *reviews* and give *compliments*
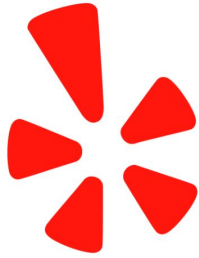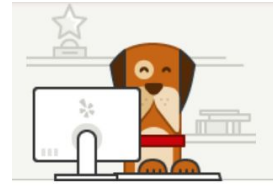
But they don't do both all the time!
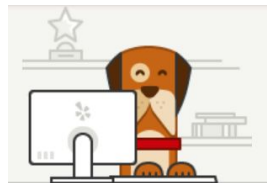
Users Are Important

Most people do combinations of feedbacks...

Elite users are yearly qualified professional promoters

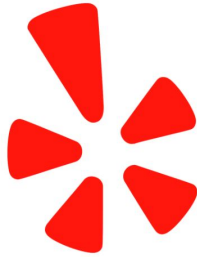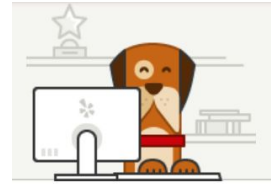They generate **send** reviews and

    receive compliments q

★ Write a Review

★★★★☆ 4/23/2019

| 💡 Useful | 😀 Funny | 😎 Cool |

Choose Your Compliment Type:

- 🏆 Thank You
- ✏️ Good Writer
- 📄 Just a Note
- ⭐ Write More
- 📷 Great Photos
- 😀 You're Funny
- ❤️ Cute Pic
- 🔥 Hot Stuff
- 👤 Like Your Profile
- 😎 You're Cool
- 📋 Great Lists

Send Cancel

Users Activities: Sent

From **user.json**, user sent attributes are *sum* of..

Written reviews Count ———————

Hawaiian, Poke, Su

★ Write a Review

Useful

Funny

Cool

💡 Useful   😀 Funny 1   😎 Cool

Users Activites: Received
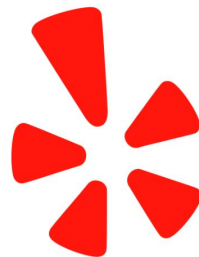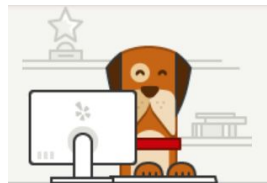
User **received** attributes are *sum* of compliments (stared):

Pearson Correlation:

Sent:
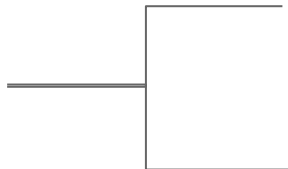
Received:
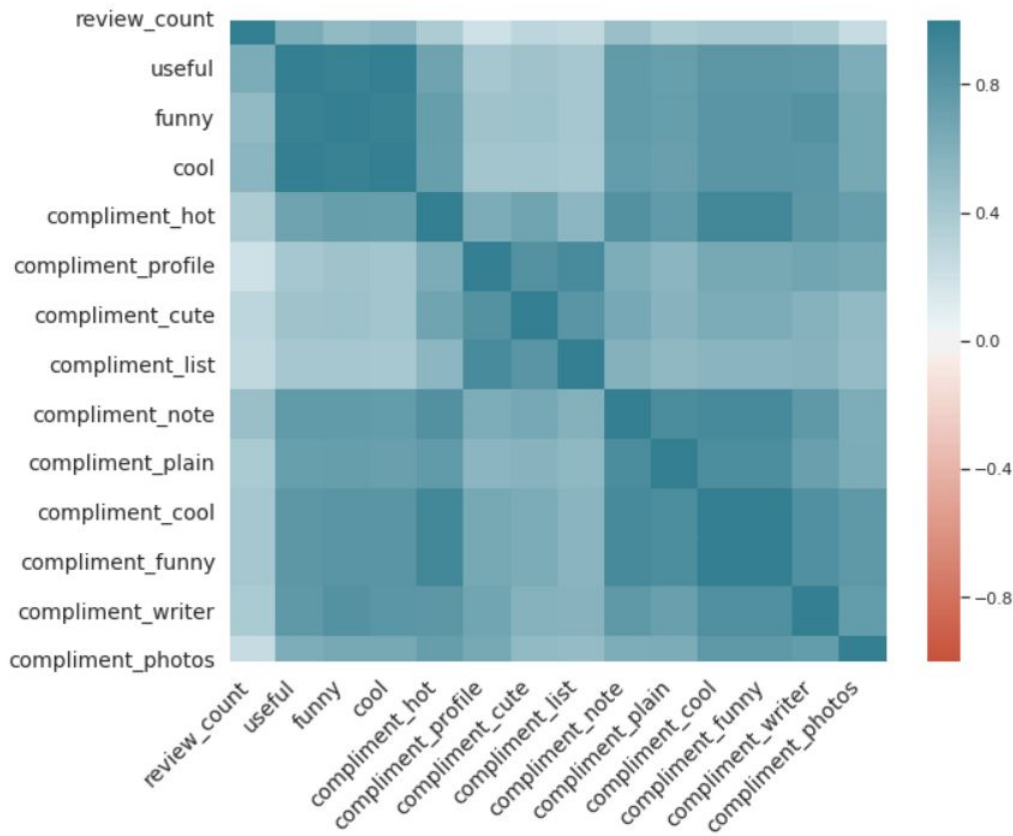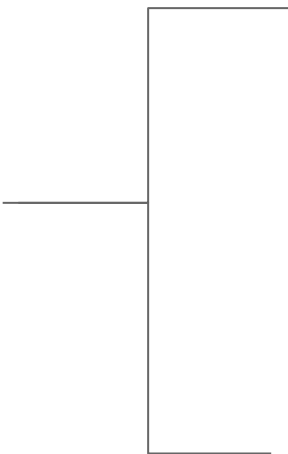
Take away:
Sent and Received
are positively
correlated

How Elite Are You?

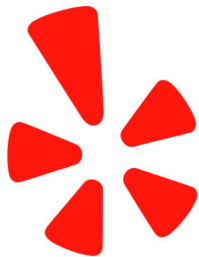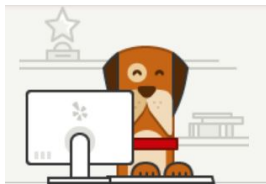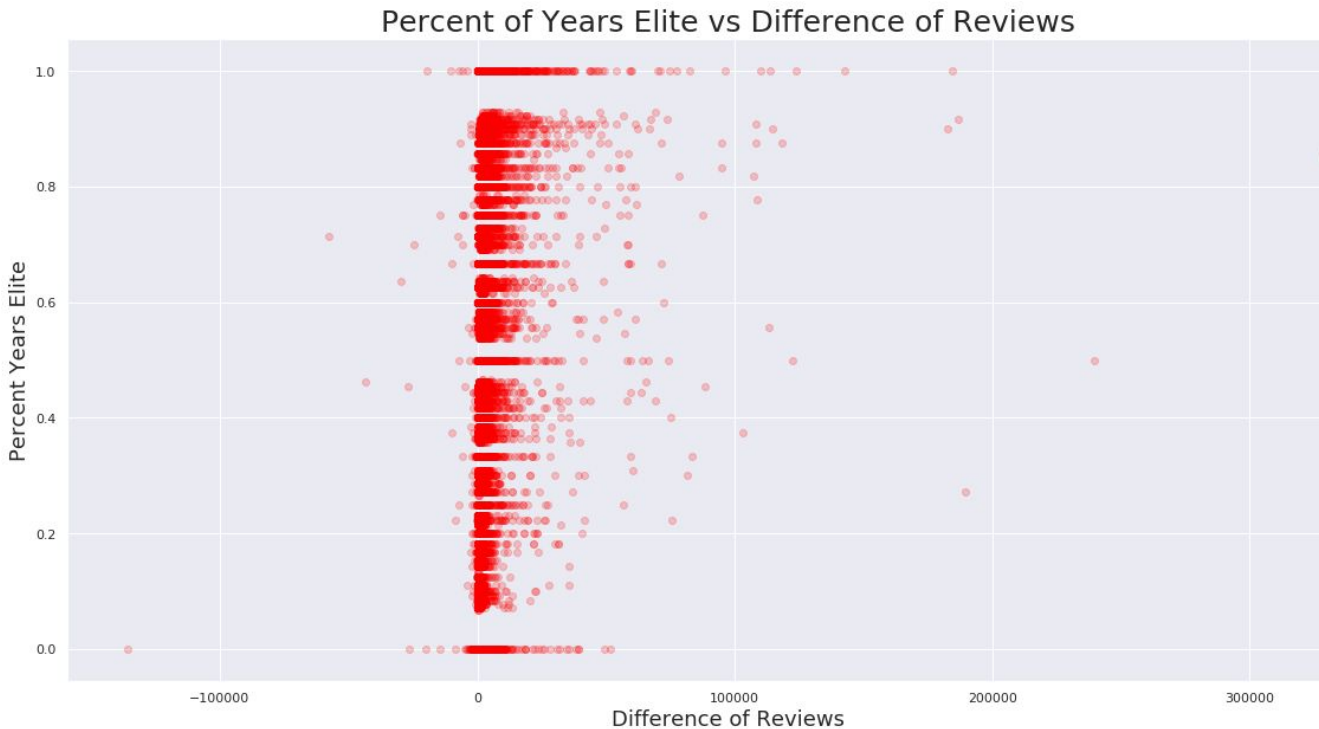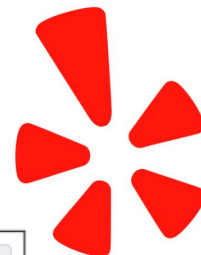Assume user **engagement** is measured by **difference**

of *reviews* and

compliments.

Is there a **difference**

amongst

percent elites?

Percent of Years Elite vs Difference of Reviews

How Elite Are You?
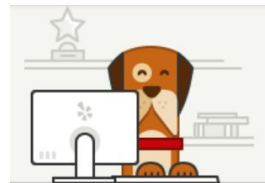
Density Plot with Multiple Elite Status

Dist. Assumption:
Normal b/c large
Sample size
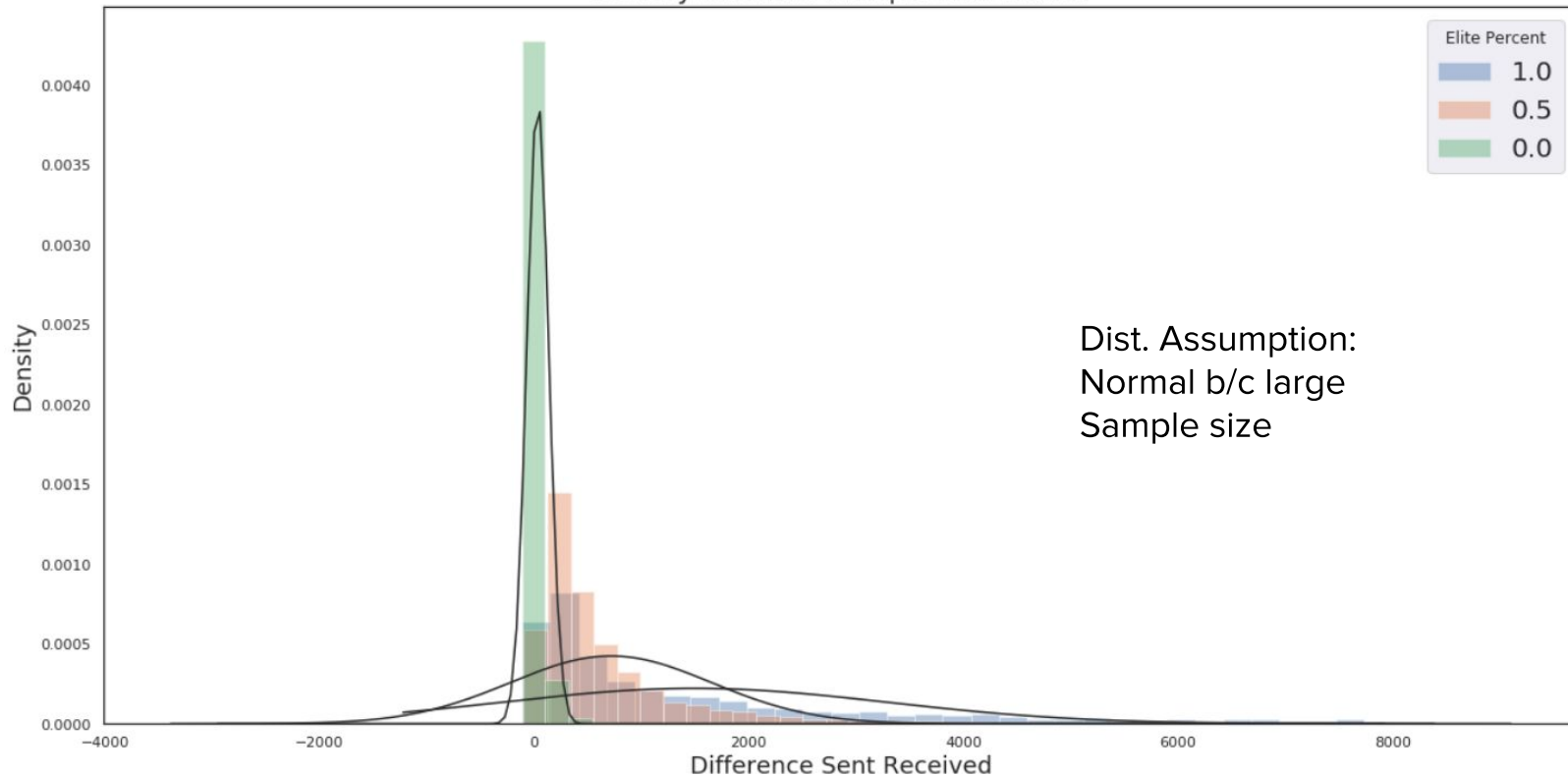
Hypothesis: Elite vs Half Elite
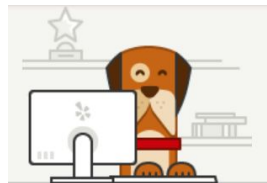
Null

$$H0: \mu_{\text{Elite}} = \mu_{\text{Half Elite}}$$

Alternative

$$HA: \mu_{\text{Elite}} \neq \mu_{\text{Half Elite}}$$

Results:

```
z statistic: -1.9168
p-value: 0.0553

DO NOT REJECT H0. Means are
statistically similar.
```

Warning: Conclusion subject to Type I error because p-value is close to alpha set to 0.05

Hypothesis: Elite vs Not Elite

Null

$$H0: \mu_{\text{Elite}} = \mu_{\text{Not Elite}}$$

Alternative

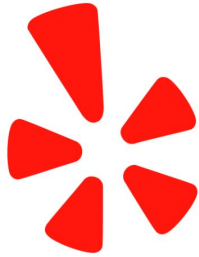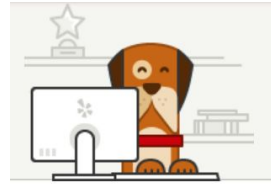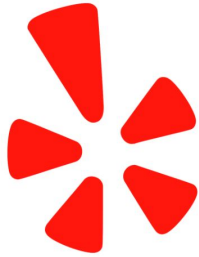$$HA: \mu_{\text{Elite}} \neq \mu_{\text{Not Elite}}$$

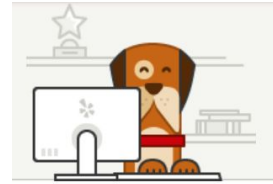**Results:**

```
z statistic: 11.2716
p-value: 0.0000

REJECT H0. Means are
statistically different.
```

Hypothesis: Half Elite vs Not Elite

Null

H0: $\mu_{\text{Half Elite}} = \mu_{\text{Not Elite}}$

Alternative

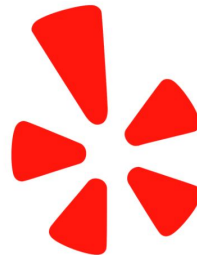HA: $\mu_{\text{Half Elite}} \neq \mu_{\text{Not Elite}}$

Results:

```
z statistic: 17.9632
p-value: 0.0000

REJECT H0. Means are
statistically different.
```

Proportion z Hyp. Testing

Null

    H0: $p_{Funny} = p_{Cool}$

Alternative
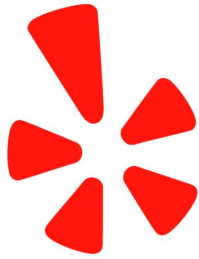
    HA: $p_{Funny} \neq p_{Cool}$

Results:

```
z statistic: -1.1642
p-value: 0.2443

DO NOT REJECT H0. Proportions
are statistically similar.
```

Conclusion:
Average Proportions of Funny and
Cool sent are similar.

Pallet Changer:

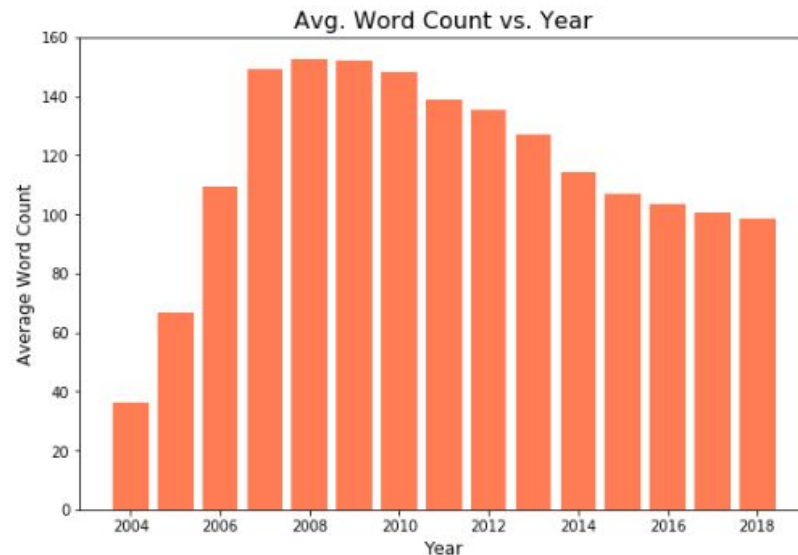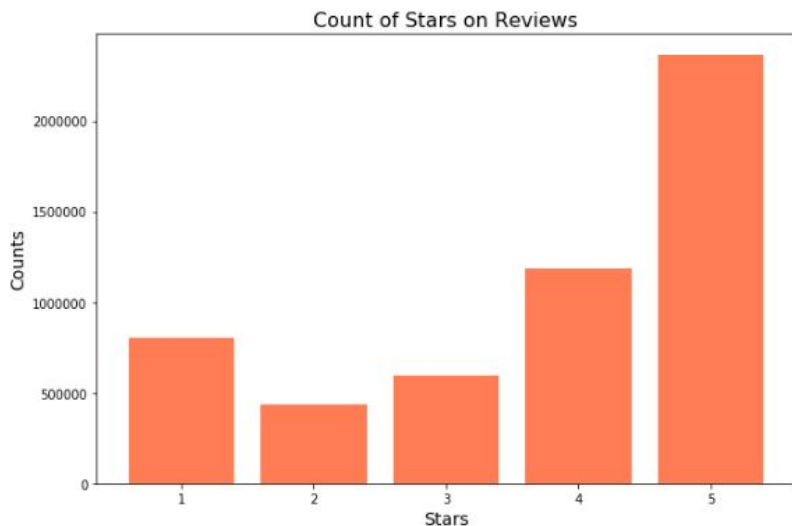# Reviews Star Struck?
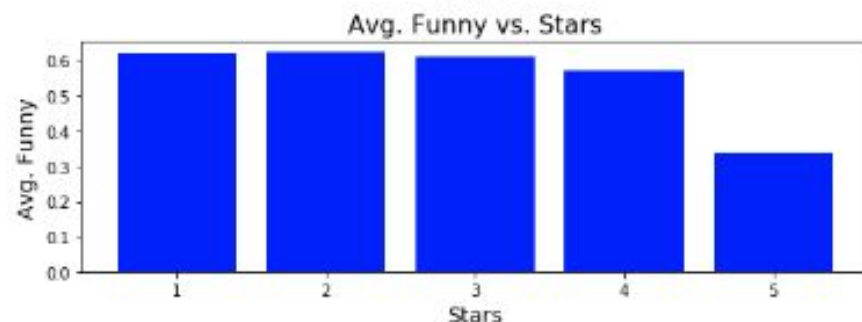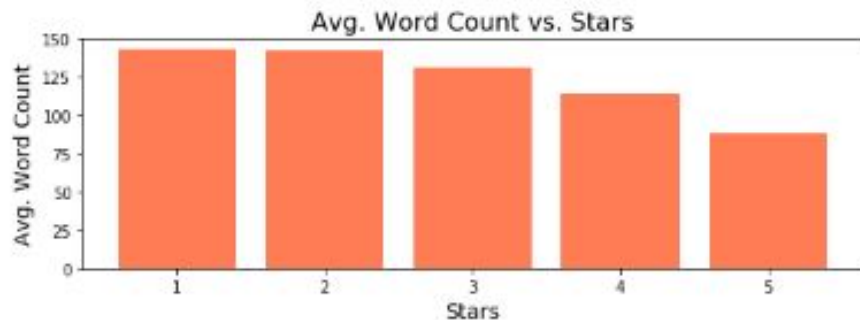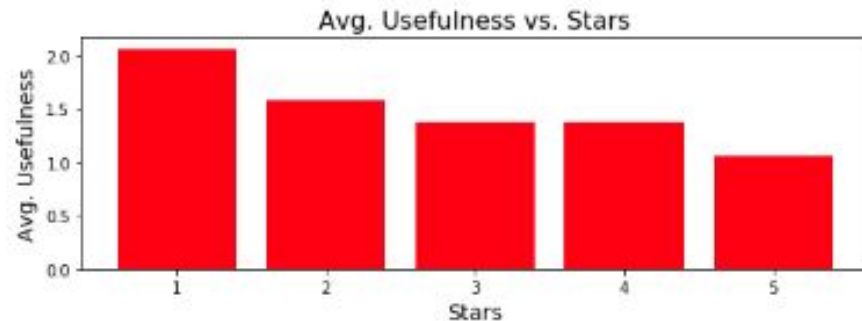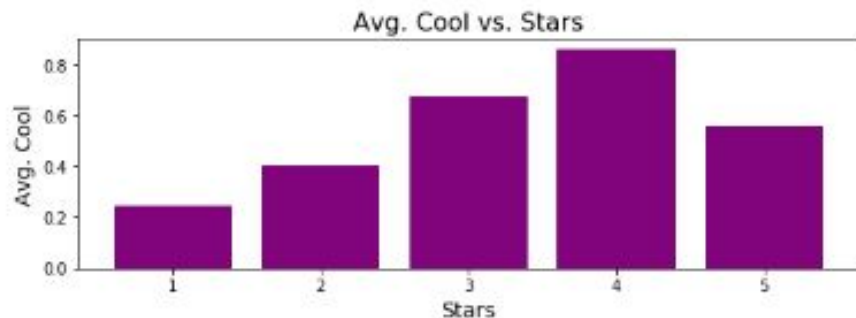
The **review.json** file has 5.4 Million entries

Here we will focus on stars and time stamp given on reviews

Reviews Star Struck?
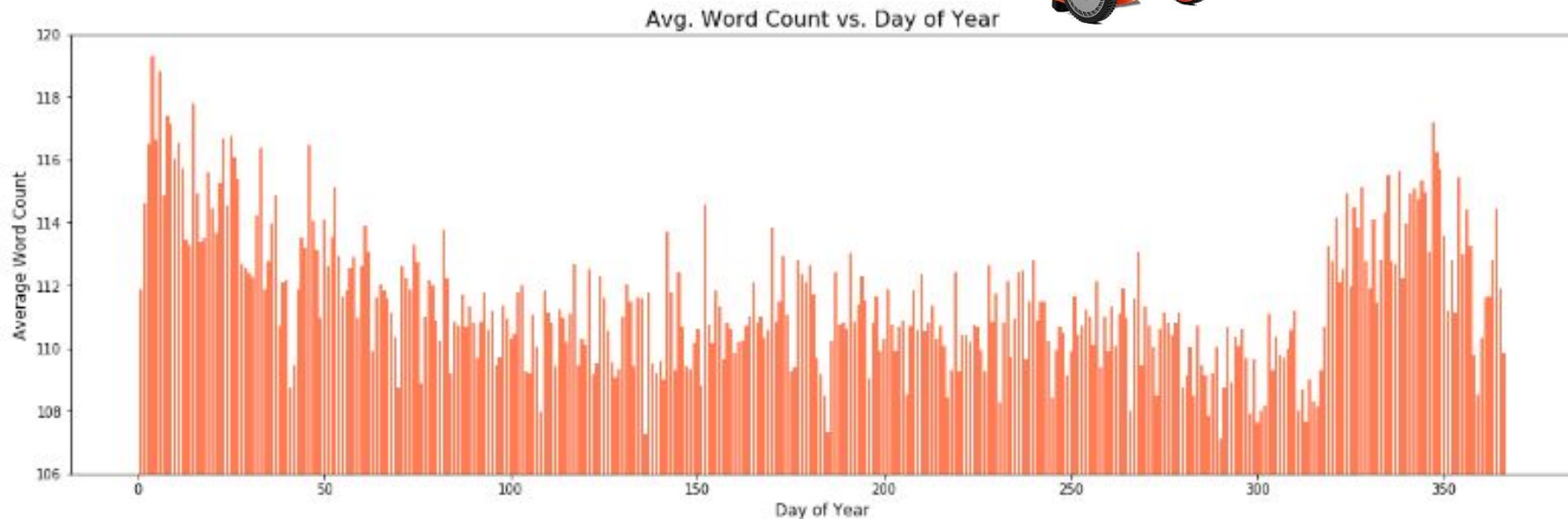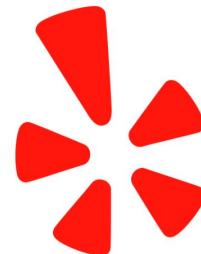
See any *patterns*?

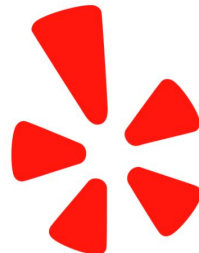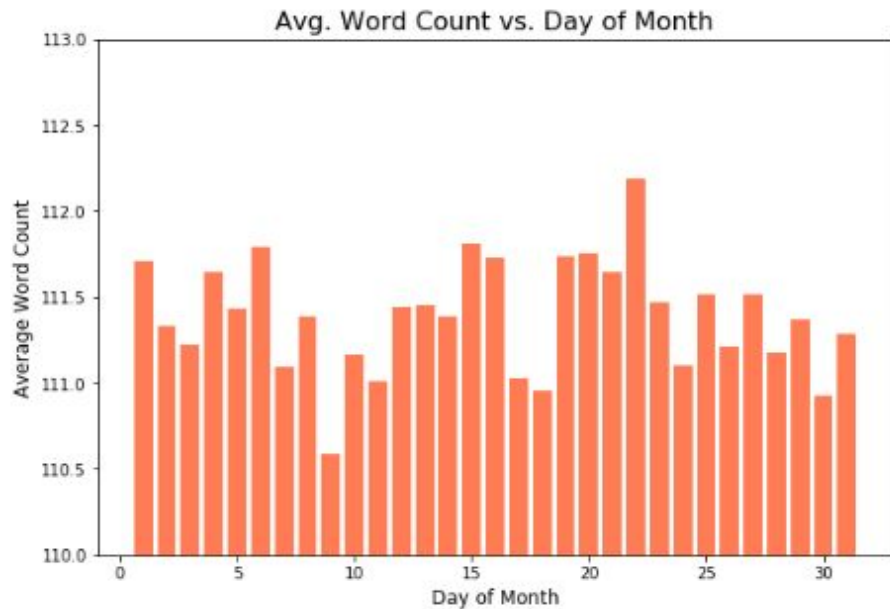# Word Count and Time

Time to trim some counts!

Avg. Word Count vs. Day of Year

Marginal Avg. Word Count

Notice the y-axis!

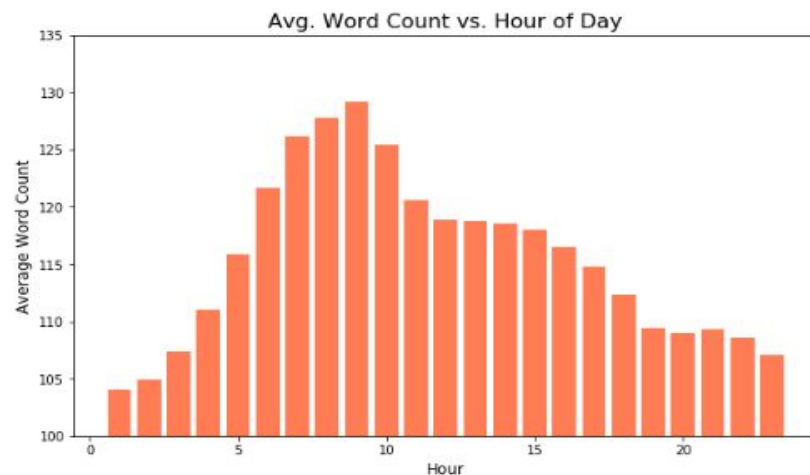Avg. Word Count vs. Weeks of The Year

Avg. Word Count vs. Day of Month
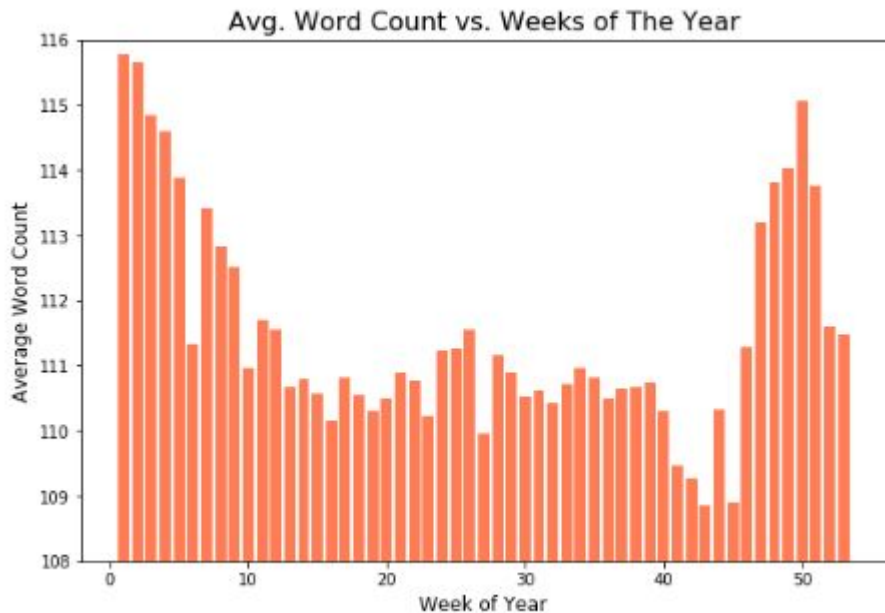
Word Count and Time

Hour of the day is important! See the 10am slump?
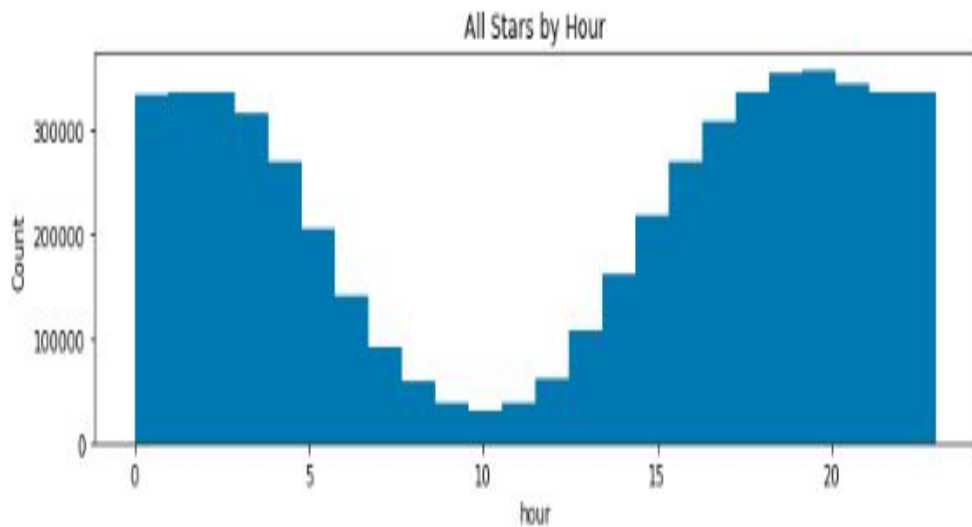
Chi Square Hyp. Testing
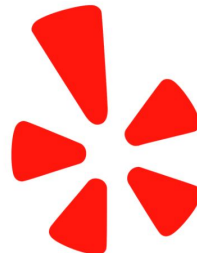
Null

H0: **NO** statistical relationship

exist between *star count*

and *hour of the day*

Alternative

HA: Statistical relationship **exists**

Results:

```
Chi Sqaure Statistics: 10051.1773
Critical Value: 1232.0735
p-value: 0.0000


Reject H0. There is A relationship
between the two categorical variables
```

Conclusion:
Statistical relationship exists
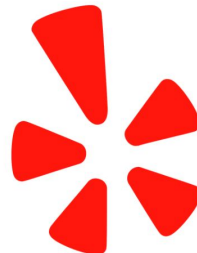between star count and hour of the
day

Chi Square Hyp. Testing

Null

H0: **NO** statistical relationship

exist between *STAR* count

and *MINUTE* in the hour

Alternative

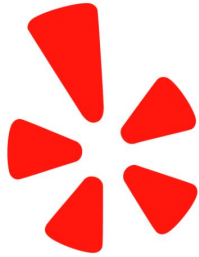HA: Statistical relationship **exists**

Results:

```
Chi Sqaure Statistics: 238.3721
Critical Value: 1232.0735
p-value: 1.0000


Cannot Reject H0. There is NO
relationship between the two catgorical variables
```

Conclusion:
No statistical relationship exist between star count and minute in the hour

Conclusion:

1. Elite users have **different** user engagements

2. Proportion of Funny Feedback = Proportion of Cool Feedback

3. There is a 10am Slump in ratings and review word dip