# **Exploring Relationships of Workplace Satisfaction and Venue Type and Density Near The Workplace**

Kylen Cruz

July 22, 2019

# **Introduction**

### **Background**

Amazon has recently brought to light the influence a company can bring to a location. This made national news and maintained itself in the news for an extended period of time. Amazon themselve had to look into each city and the advantages they bring. While supplying the needs of the company will be of the utmost importance, workplace satisfaction would also be a beneficial metric to look into. Higher workplace satisfaction may be associated with lower turnover rate and higher employee productivity.

### **Problem**

Choosing a location may be have the benefit of looking at venue density in order to increase workplace satisfaction, as well as the information provided on websites like Kununu. Looking at venues and venue density in a series of radii to gauge whether or not it affects workplace opinion on a variety of company review sites is the goal of this project.

### **Interest**

Picking an office location has many pieces involved with it. Looking into the local economy and being able to gauge an increase in workplace satisfaction from that is important. It may make choosing a location to start at easier from not only an analytic view, but a personable view. Two locations may be similar, but an extra coffee shop or park may be just the thing the make one be the better choice. On the inverse, if having venues nearby

The following is intended to be a scrape and analysis of employee satisfaction and venue analysis. This may be important knowledge to have if branching into new locations from a point or starting a new business and optimize location. This is not intended to be the sole resource on workplace locations if there is a significant difference. Instead it should be used to influence workplace satisfaction in addition to other tools.

# **Data**

### **Sources**

There are 2 Major data sources. The first is Kununu, an employer review platform. There are many categories that companies are rated on, including Company Culture and Work-Life Balance . You can look at the reviews at this url: <https://www.kununu.com/us/dell-usa>. There are comments and explanations as well as the rating system. The ratings, location, and company are the only parts being used. The two companies picked are Dell Technologies and Oracle Corporation. Both are large companies with multiple locations. They are also both technology companies. This is done intentionally to mitigate possible confounding factors. The data will be scraped automatically.

The second category is data from foursquare. The information to be taken is simply the count of venues based on the category specific call. There are 5 radii to used. Data from foursquare will include venue data in the following radii: 100000 Meters 10000 Meters 1000 Meters 500 Meters 250 Meters

This is to judge venue density. Foursquare API does not provide venue data beyond 100 Venues per call. The intent is to gauge venue density in the near area of the office space. To ensure that each radius is not limited by this, if venues return at the maximum it will be assumed that the venue density is consistent with the next highest radius. New York City is most likely to have issues in this area due to its density. This is not without issue, smaller areas are more prone to variance. Looking at density looks to include business density, which may be important to employee satisfaction. Venue type was maintained and turned into a ratio relative to other venues. Venue type may be more important than just the amount and density of venues. A third data source was used. Business locations were pulled from the company websites. Dell did not have a scrape friendly website. The addresses were coded manually. Oracle was able to be scraped for business locations. It is important to note, there is an issue with the Oracle location data. Many of the addresses were not able to be processed by the Bing geocoder. This is an issue as Foursquare used Longitude and Latitude to use their API. Some data was usable. It is not clear whether this is an issue of the data or the Bing geocoder. It is important to know this issue relative to the current data and if processing further data.

### **Data Cleaning**

The first step, used for simplicity and time, is dropping any Kununu scrape data that doesn't fit in the expected. The scrape data was organized using the BeautifulSoup library. Most reviews followed a basic, but a more precise filtering of the data is likely available. All the data was put into a Pandas data frame and needed to be cast as either integer or float. The location data needed to be uniform in whitespacing, some spot spelling corrections, and be only the city name. Without this, statistical analysis will be less accurate and combining the data sets will be more difficult without any benefit. Kununu doesn't have strict formatting rules on location, so there may be incorrect spellings or vague locations. From Foursquare itself, data is received from the API and turned into a ratio. The category type is know of each venue. The more difficult portion of the process of finding the venues are the business locations. Many business have their locations available on their website. The current method uses a web scrape where it looks efficient. Otherwise, inputting the addresses by hand is the alternative. A second issue is errors presented by the Geocoder. Geocoding was performed through the Bing geocoding services. The geolocator failed to find many addresses, and it is unknown where the issue is starting. Multiple failed addresses have been tested using the Bing Maps services, and each has returned a location with longitude and latitude. Taking the time to ensure these are accurate would be beneficial to the data. Review that do not have a matching foursquare location will instead be associated with the mean venue values based on company. This will be an issue for a possible skew of the data

### **Feature Selection**

Kununu uses 18 different categories. Each category will be used as well as the mean of each review, which will be called Total Rating. The focus will be on the Total Rating, but information and insights can still be gleaned from the individual review data. Foursquare data that feature at least one non zero value will be included. There is an issue with the sample size from Foursquare, and due to external factors were not fixed prior to this report’s completion. Venue location offers well over 50 features, only 10 will be used in the Machine Learning section of this paper.

# **Exploratory Data Analysis**

### **Calculation of target variable**

Workplace satisfaction involves a multitude of variables. The current focus will be the Averaged Total Rating from Kununu ,the sum of scores divided by the number of scores. This is how Kununu calculates. For scraper simplicity, making the calculations instead of scraping is a negligible time difference. For industrial organization psychology reasons, the individual rating will be kept ready to swapped in place of the Total Rating. Targeting specific areas can be useful for aiming for specific goals. Examples pointing to this will be made later.

**Relationship between satisfaction ratings and themself**

An employee satisfied with one factor of their employer will generally be satisfied with the other factors of their employer. Alternatively, if low satisfaction in one area, then the expectation is that all other areas will be low as well. While it is not surprising there is a strong positive correlation (Pearson’s r, r = 0.832, p < 0.05 ) between the Total Rating and Gender Equality as one feeds into the other (Figure 1) . All of the satisfaction rating have at least a moderate correlation and are significant (r > 0.3, p < 0.05). The weakest connection is between Challenging Work and Job Security(r = 0.412,) and the strongest is between Internal Communication and Company Culture (r=0.798.) It would be an assumption to say that improving any one of the metrics would improve all of the metrics, but the relationship is there and should be monitored.



Figure 1. Scatter plot of the relationship of Gender Equality and Total Ratings

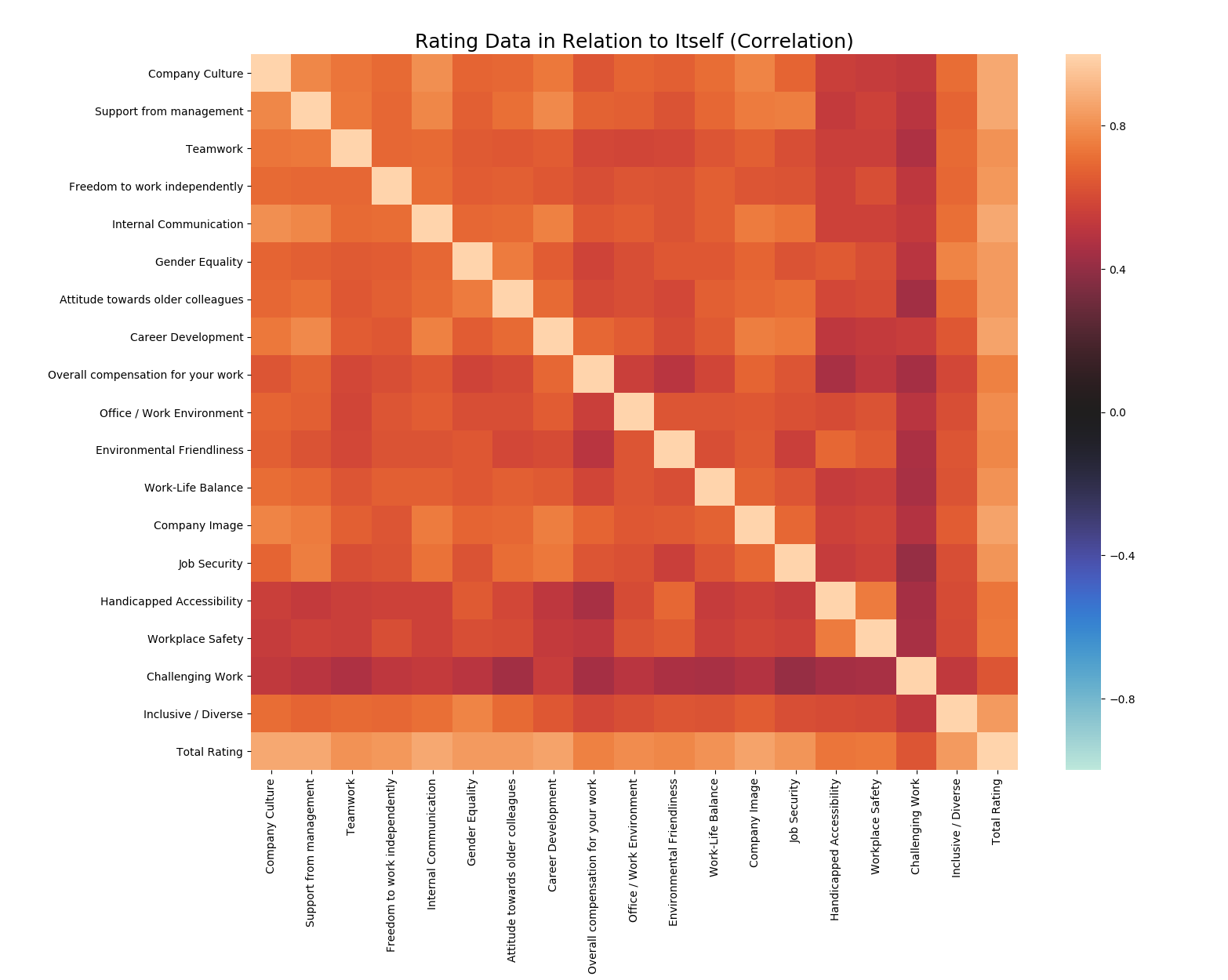


Figure 1. Heatmap showing Pearson’s r correlation between all Kununu workplace ratings

**Relationship between city and total rating**

A key hypothesis that there is a relationship between city location and satisfaction ratings. Total rating was used for the testing. The correlation between the rating will be able to provide a picture of the individual values that make up the total rating. Defining a city by its geographical location may be important, but what the city contains also is important. Looking first as a city before looking at what makes it up is critical as a step in understanding. Testing occurred for cities that had more than 30 reviews available. These cities were Round Rock, Redwood City, Hopkinton, Boston, and Austin. Testing shows there is a significant difference between the cities (one way ANOVA, p < 0.01). Looking closer, it was shown that only one city, Boston, was significantly different (t-test, p< 0.01). The greatest difference was between Round Rock (X-bar = 3.339) and Boston(X-bar = 4.246), with difference of average of 0.90 on a 5 point scale, an 18.1% difference. Within the same state in Hopkinton (X-bar = 3.420) there is a difference of 0.83 points, a 16.5% difference.

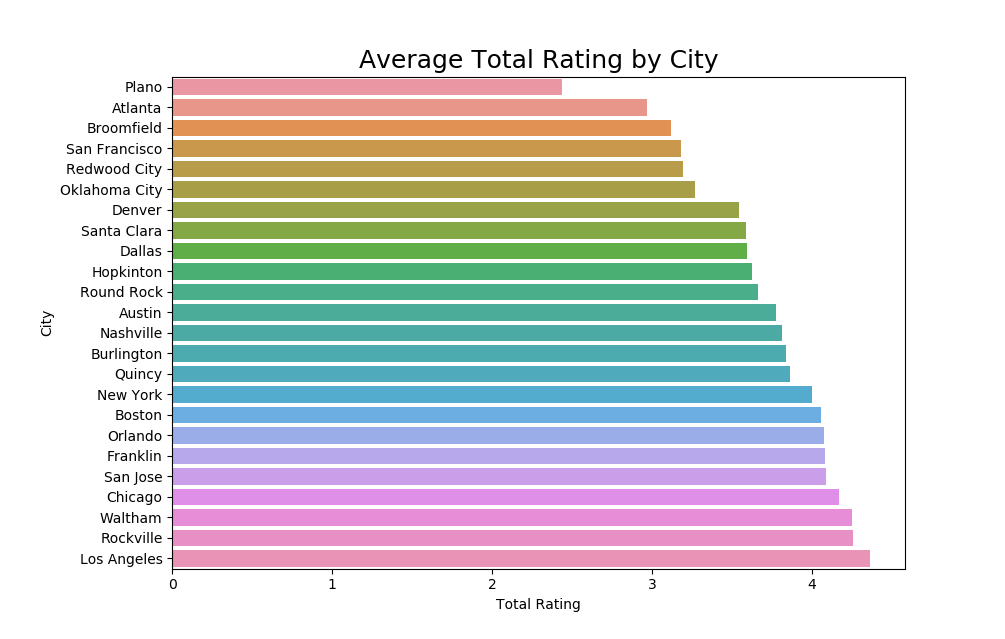
****

Figure 2. Average Total Rating with Cities

**Relationship between venue ratios and total ratings**

The overarching hypothesis of this survey is that venues and workplace satisfaction are significantly related in hopes to use the data to create a tool to help with business location decisions. Correlations were done on two versions of the data. One done using just reviews and venue data that share the same city. This data does not show a significant correlation (p > 0.05). The other uses the previously mentioned set and then takes the average foursquare data based on company and fills the missing venue data on the company. This has shown only 1 significant correlation between Total Rating and Vegetarian / Vegan Restaurants (r = -0.080, p < 0.05). A weak correlation that should not be used as either a primary or secondary factor in choosing a new business location. It should be noted, there are significant but weak correlations for individual ratings.

**Relationship between venue density and total ratings**

In order to look at venue density, foursquare calls were set up to check if the maximum venue return allowed by foursquare. It is not ideal, but it allows insight whether a location in a low venue count area has an effect on workplace satisfaction. It also allows for the follow up insight of a possible confound. The data shows no significant correlation between Venue Density and Total Ratings (p > 0.05). Once again, there are specific rating that are affected and should be noted for possible machine learning applications.

**Predictive Modeling**

There is an issue in creating a model connecting Venue Density and Total Ratings. The data only shows weak, if any, relationship between the two data sets. Regression models were created in order to possibly infer a more subtle relationship not obvious from statistical testing. One issue may stem from a relative lack of variety of venue data. Due to time restraints, the data was not expanded. The framework for the regression modelling should allow for faster implementation of regression models from improved data quality. Classification methods is not being considered. Classifying expected outcomes as being more suitable would be possible through regression. Clustering may lead to further intuitions about the review data, but venue data remains a problem,

**Regression Models**

A multiple linear regression model was used along with polynomial preprocessing up to the 9th degree. Venue data that had more than 5 significantly related were picked, as well as Vegan / Vegetarian Restaurant due to its significance with Total Ratings. Multiple Linear Regression appears to be the best fit, but an average error may span more than 50% of the data points with a MAE of 0.96. For reference, this is on a 5 point scale with a standard deviation of 1.06. This outcome was not unexpected. Regression models were not useful as a tool for predicting Total Rating from the Venue data

**Conclusion**

Analysis of Workplace Satisfaction Data and Venue has produced mixed results in terms of significance. There may or may not be a lack of significance between Venue Type Ratios/ Venue Density and Total Ratings. One should be wary of inferring too much about that particular relationship in this particular study as there were issues with data variety. There are intuitions to be garnered from the data aside from the primary goal of the study. The exclusively positive and significant nature of Rating Data within itself suggests any improvement in these metrics will generally improve the ratings of all other metrics. If this relationship is the case, that would allow companies to focus on improving a target area and have a generally improved opinion about the workplace. Additionally, the significance of Boston in relation to the rating data is extremely curious. Neither of the two companies have listed a Boston location, but there were over 30 reviews for the city from both companies. The city of Hopkinton did not share the same benefit in rating even though it shares a state. It may be suggestable to look for a confounding element affecting these rating. Regression models did not prove useful in this instance, but issues related to data may have been the key problem for the models.

**Future Directions**

The number one change to the study should be an improvement to the Venue Data. A more reliable geolocator data would be first on the list. The amount of errors returned by Bing Services outnumbered the successful calls. The specific issue isn’t clear, but it is very much a problem. Fine tuning venue data either by automating Foursquare API calls to generate radius with all of the venues included, or tuning the radius using smaller steps to include more venues in total and have a more representative ratio. Both of these introduce the logistical issue being call intensive for an API that can have a cost associated with those calls. This can be circumvented if there are other sources of Venue data.

These are the primary concerns, and after they are taken care of other steps can be done to be able to gain more insight to any relationships within the data. Expanding the scope of companies to possibly account for the companys’ innate ratings. Access to internal Workplace Satisfaction data that may have better accuracy in both rating and location. This study also looked at two companies in the tech industry, and may not have any generalization to other industries such as Fast Food. Employee benefits are also a factor that would be good to include. These are factors that may be expected to have an effect on the general workplace satisfaction, and further research may yield further insights into relationships and significance in the data.