Data Analytics: Final Report

Joshua Cooper | Student ID: 16098824 | Group C

Abstract

Data, when collected over time, often exhibit seasonality. This is especially prevalent in data related to economic activity. For instance, an ice cream vendor can expect sales to be much higher in the summer than in the winter. It is important for the vendor to be commercially aware of this fact in order to simultaneously exploit seasonality and be careful not to attribute trends caused by seasonal effects to business decisions.

Businesses on Yelp are likely to be interested in change over time in their user ratings. A deterioration in average star rating might prompt changes in business strategy, and a rise in ratings might make management think a particular strategy, deployed at a lucky time, is successful. However, people's social behaviour may be influenced by seasonal factors, and this may feed into a change in overall ratings. Therefore, businesses need to be aware of seasonality in their ratings, so that they can make informed business decisions grounded in sound reasoning.

This report aims to address the question: Is there seasonality in star ratings? Our null hypothesis is that there is no seasonality in the data, and we conduct statistical experiments to determine whether the null hypothesis holds. The alternative hypothesis is that there is seasonality present in the data. There is a variety of statistical tests that can be deployed to address this problem, including ANOVA or Kruskal-Wallis tests on the set of monthly rating distributions, depending on whether our data are normally distributed. We also test correlation of star ratings over time with weather.

Results indicate that seasonality is prevalent in the data, and that we consequently can reject the null hypothesis. Our results also point to weather conditions at the time of reviewing being able to explain a small amount of variance in ratings. This highlights the need for businesses to be cognisant of any temporally driven patterns in star ratings, in order to effectively determine appropriate business strategy.

Computer Science Department
University College London
30 March 2017

First, we present a discussion of the dataset. After this, we conduct some exploratory analysis, which includes fitting distributions to empirical data. Finally, we present our analysis of the posed question regarding seasonality,

Contents

Throughout this report, numerical results are generally reported to three significant figures. For integer numerical results, some figures are presented to more than three significant figures. However, in these rare cases, the overall emergent themes arising from the reported results should remain clear.

1 Overall description of the dataset

The Yelp dataset contains, in the broadest sense, data related to customers reviewing businesses. The dataset consists of five files, each containing a different type of data. The five files correspond to:

- **Business** - businesses in a select group of cities which have reviews on the Yelp website
- **Review** - reviews pertaining to the list of businesses
- **Checkin** - recorded accounts of users visiting businesses
- **Tip** - tips written by users for other users
- **User** - a list of users of the website

The `business` file contains one entry for each business in the data set. A rich list of attributes is also provided, including features such as opening times, ambience, and music, for approximately 86,000 businesses. The `review` file contains over 2.5 million individual reviews. Each `review` entry contains the review itself (text and ratings), the business ID for which the review was submitted, and the unique ID of the user who submitted the review. The `checkin` file contains approximately 60 thousand checkins. The `tip` file contains well over half a million tips given by users, and each tip record has an associated business ID, a user ID, the actual text tip, and various other fields. Lastly, the `user` file contains approximately 700 thousand users with, for each user, counts of various types of interactions with other users (e.g. compliments). Each entry in the file also lists other users that the primary user is *friends* with, average stars awarded in reviews, and number of submitted reviews.

Each field in the data is either meta-data (for example, the date of the review), key-data (which we outline below), or text data (reviews/tips). The files are linked via shared fields; namely, `user_id`, `business_id`, `review_id`. These form the keys used to merge and traverse across tables, and the three keys are present in the following files.

- `user_id`: review, tip, user
- `business_id`: business, review, checkin, tip
- `review_id`: review

The data needed a modest amount of cleaning, since many city names existed in the data set in more than one form. For example, *Las Vegas* data needed merging with *North Las Vegas* data.

2 Exploring the dataset

First, we present some exploratory analysis of the dataset. This consists of two parts. Firstly, we show how star ratings vary across a small subset of cities, and discuss the main statistical properties of the distributions of star ratings. Second, we look at the distribution of number of reviews per user and city, and attempt to fit established distributions to these samples. Plots are generated using `Matplotlib` in `Python` (*see appendix 1*).

2.1 Analysing star ratings across three cities

For the three chosen cities (Las Vegas, Edinburgh, and Montreal), the histogram plot is below:

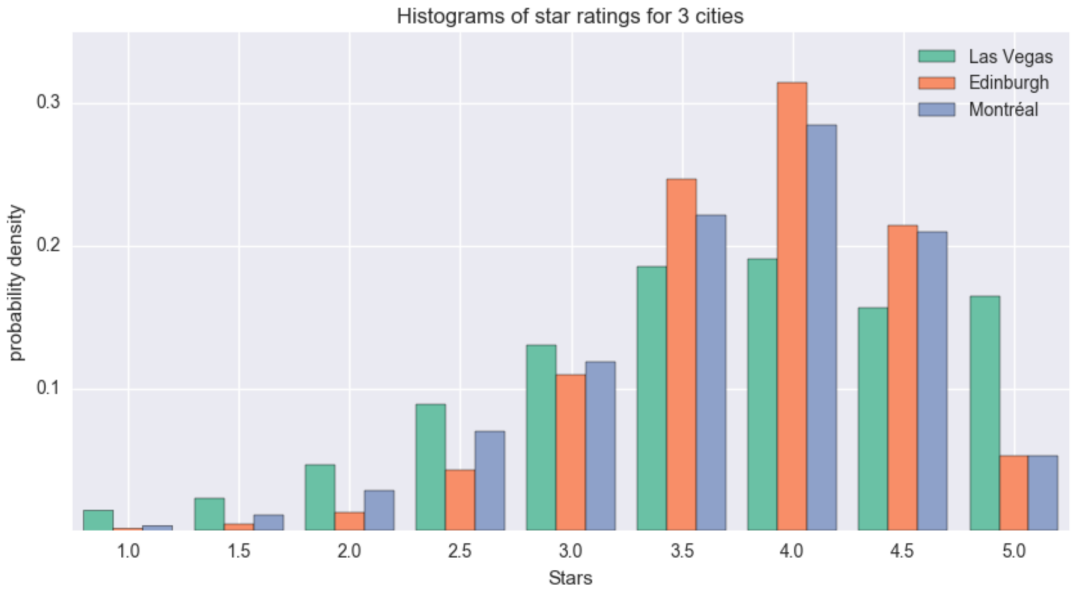

Figure 1: Histograms of three cities

Below we show moments for each distribution. Moments are calculated using `Pandas` (*see appendix 1*).

Table 1: Moments for three cities

City	Mean	Variance	Skewness	Kurtosis (excess)
Las Vegas	3.70	0.94	-0.54	-0.27
Edinburgh	3.82	0.44	-0.64	0.74
Montreal	3.73	0.58	-0.71	0.41

According to our sample, Edinburgh restaurants have the highest mean rating, and variance is lower in Edinburgh than it is in Montreal and Las Vegas. Skewness is similar in all cities, with Montreal having the highest negative skewness (distribution is skewed to the left side of the mean). Two of three distributions have small excess kurtosis, while the Las Vegas distribution is slightly platykurtic. This is intuitively appealing given the flatter shape of the Las Vegas distribution in the histogram. Given that the histograms all look fairly similar, it is not surprising that the moments do not differ significantly between cities.

2.2 Analysing hotel star ratings across three cities

We analyse hotels in our three cities next:

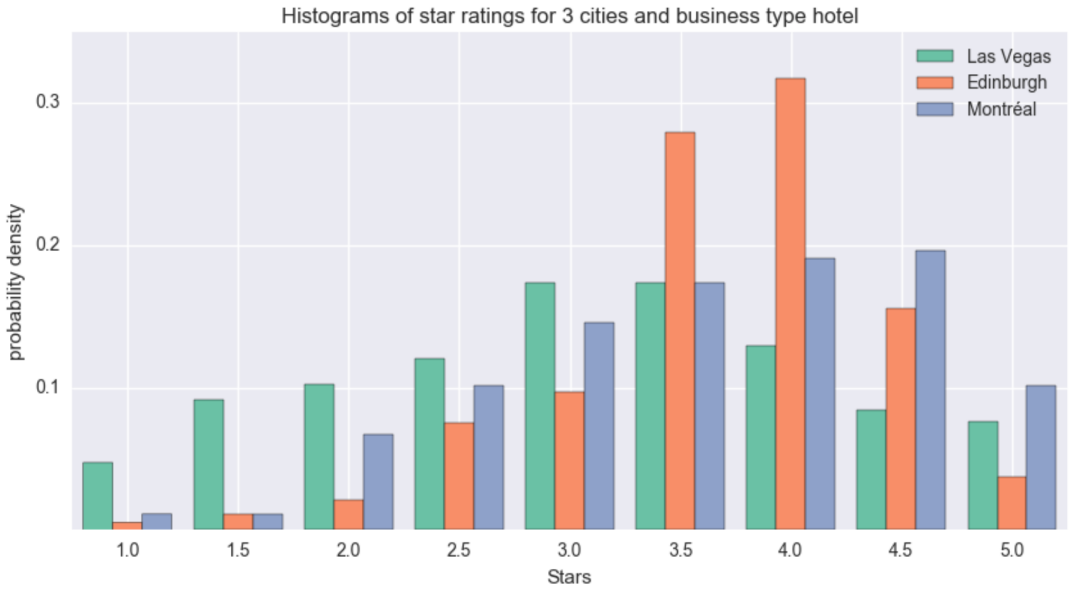

Figure 2: Histograms of hotels in three cities

Again, we now show moments for these distributions

Table 2: Moments for hotels in three cities

City	Mean	Variance	Skewness	Kurtosis (excess)
Las Vegas	3.10	1.19	-0.09	-0.79
Edinburgh	3.68	0.52	-0.80	1.05
Montreal	3.62	0.86	-0.45	-0.44

Visual inspection of the histogram and the table of moments suggest that these three distributions differ significantly from one another. We can clearly see from the means of the distributions and by visually inspecting the histogram that Las Vegas hotels are rated more poorly on average than hotels in the other two cities. The distribution also looks more evenly spread in Las Vegas, which is consistent with higher variance. Given that the distribution looks quite symmetric, it is unsurprising to see skewness close to zero. This is not to say that the Montreal and Edinburgh distributions look the same. Despite similar means, variance is slightly different, Edinburgh is more negatively skewed, and, interestingly, Edinburgh has strongly positive excess kurtosis. This makes sense given the relatively small variance and presence of counts in the left tail of the distribution. On the contrary, the Montreal data show negative excess kurtosis.

2.3 Fitting distributions to number of reviews per user or business

For this task, we focus on the city of Las Vegas. First, we plot the data in log-log scale.

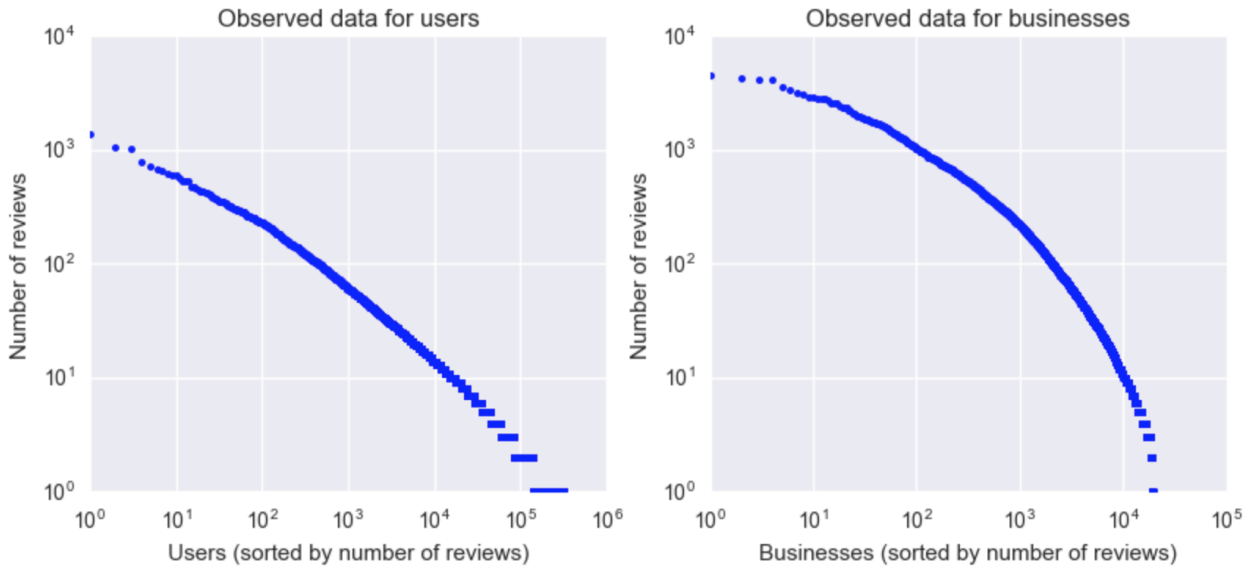

Figure 3: Empirical data

The step-like shape at the right side of the user plot corresponds to the fact that many users have written only a handful of reviews, with each "step" corresponding to a specific amount of reviews. However, it is notable that although many users have one, two, or three reviews submitted, for businesses, the sharper drop off in the bottom right of the plot implies there are relatively fewer businesses with very low review counts. We now show histograms (in log-log scale) and the moments for each distribution next.

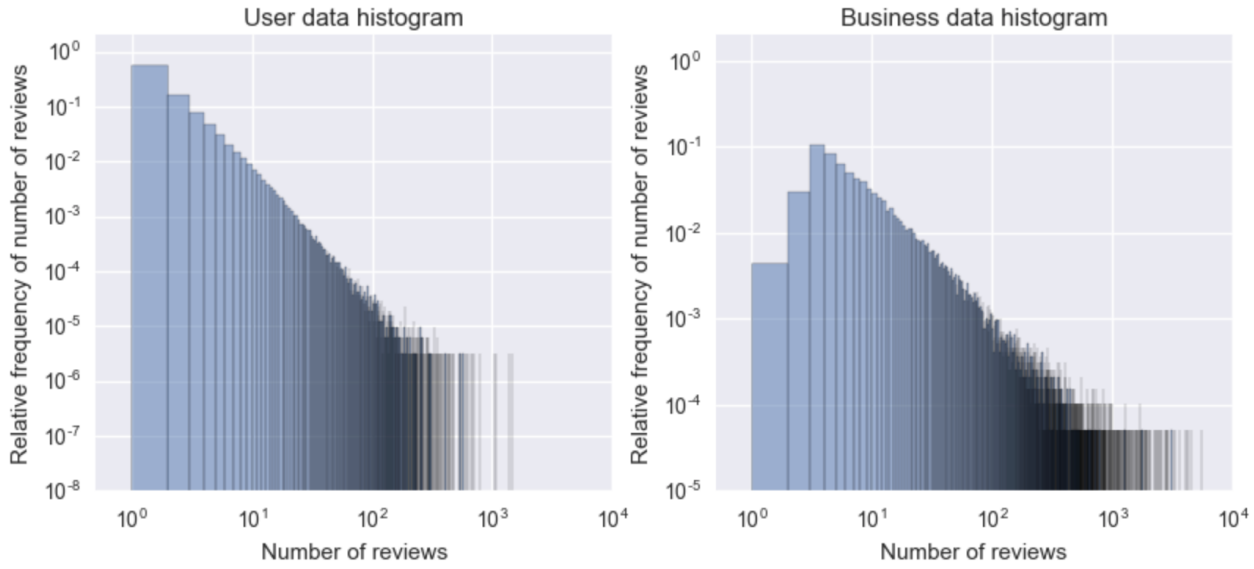

Figure 4: Histograms of empirical distributions

We can see from the above histograms that there is quite a good exponential fit to each distribution up

to a certain point on the x-axis, but that the tails do not conform to this linear (in log-log space) trend.

Table 3: Moments for review count distributions

	User	Business
Mean	3.26	53.43
Variance	112.36	30,579.53
Skewness	39.46	11.44
Kurtosis (excess)	3,314.9	203.96

With such high kurtosis, we clearly cannot expect our data to be generated by a Gaussian distribution (*see appendix 2*). Therefore, we turn to more exotic distributions to seek a good fit.

To fit distributions to these data, the `SciPy Stats` library within `Python` was used (*see appendix 1*). The library uses Maximum Likelihood Estimation to find optimal parameters for a distribution, given data. The library also has an implementation for the Kolmogarov-Smirnov test, which was used for this report. A script was written in `Python` that fitted a range of distributions to our data and highlighted distributions with better goodness of fit, according to the Kolmogarov-Smirnov test. Certain distributions look appealing in terms of their fit to the data. For instance, we now show the Weibull-min distribution, fitted to both distributions. More detail around the Weibull-min distribution can be found in *appendix 2*. However, no individual distribution was found that was a sound fit for the data, with Kolmogorov–Smirnov test statistics that were equivalent to extremely low p-values.

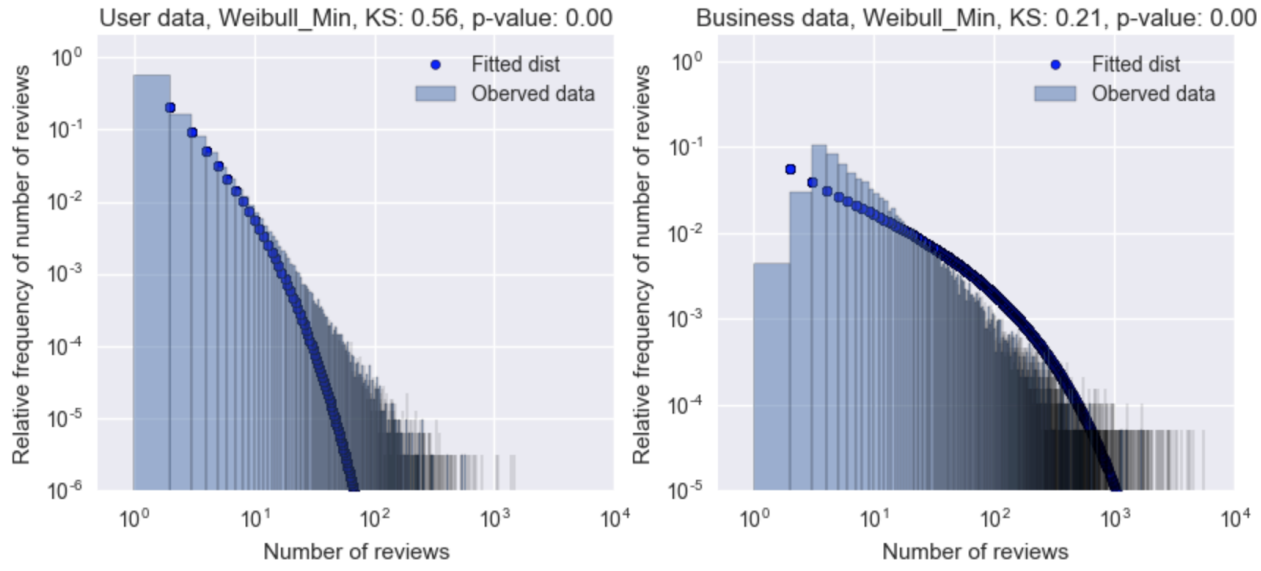

Figure 5: Weibull-min pdf

The same distributions are shown in their complementary cumulative form (again in log-log space).

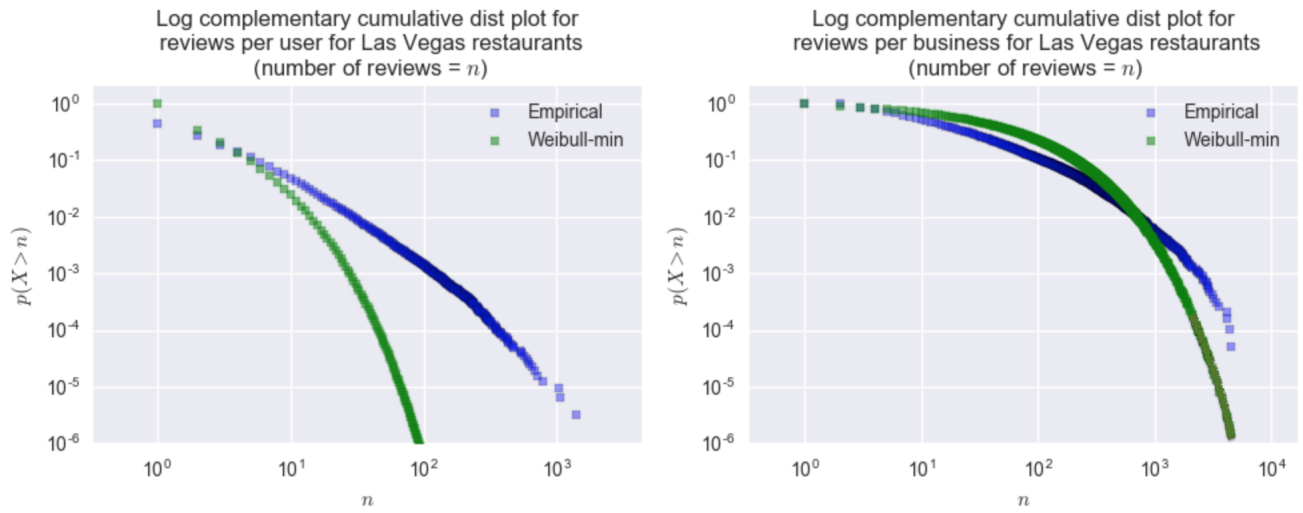

Figure 6: Weibull-min ccdf

Now, we investigate the potential powerlaw behaviour in the tail (*see appendix 2*). Having explored different values for x-min while trying to map the Pareto distribution to our data, the following Pareto distribution parameters were found. The `Powerlaw` library was used (*see appendix 1*) to calculate parameters.

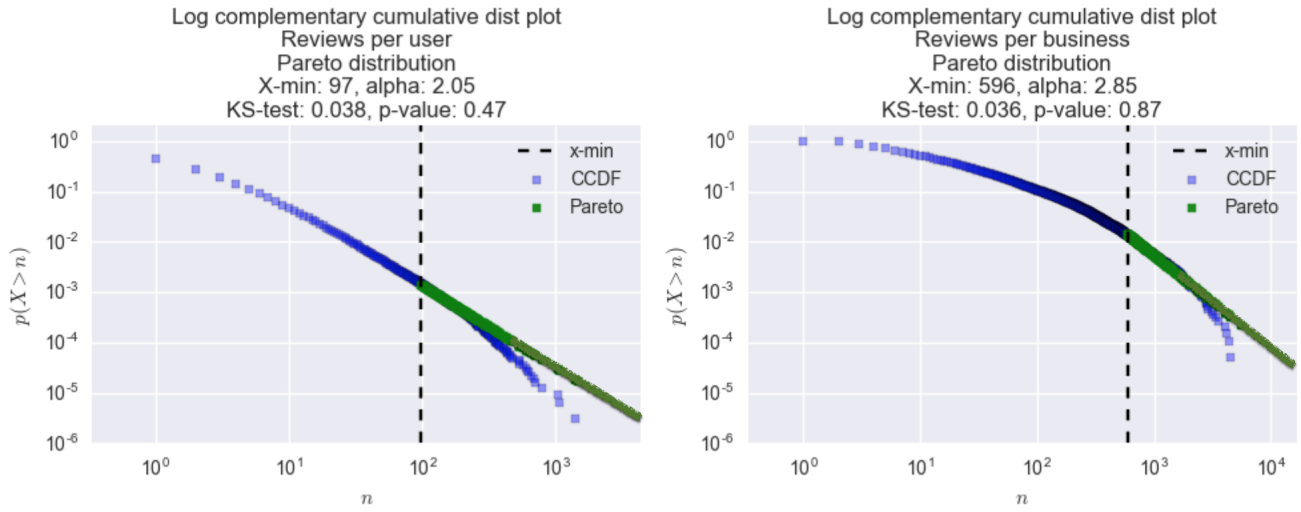

Figure 7: Pareto fitting on the tails

The specific parameter values used are shown in the plot titles. Also note that the alpha parameter is for the discrete case. The value of alpha for the user distribution suggests very heavy tailed data, since alpha is only marginally above 2. However, we have finite variance. The business data is not as heavy tailed, with a higher alpha. Both sets of parameters generally fit the tails of the data well, with the p-value for the tail of the business data being particularly robust. We are now left with the bodies of the two distributions.

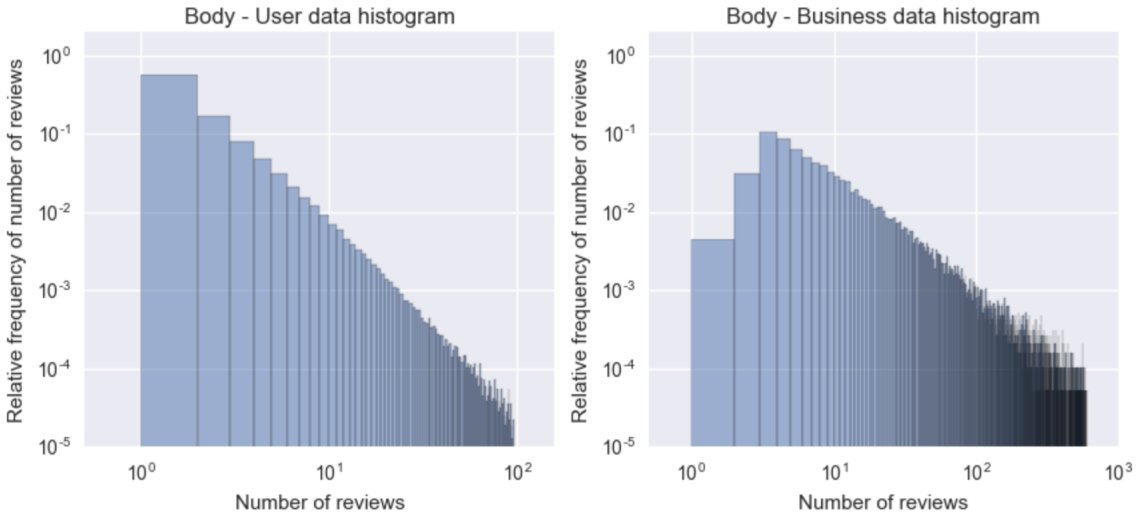

Figure 8: Body histograms

There are several options available for fitting the bodies of the distributions. No individual distribution was found to be a good fit for either the user or business data. An alternative method would be to divide the data into more mutually exclusive buckets and find appropriate distributions for each bucket. However, this time-consuming approach is not taken here. Instead, we simply use the empirical counts of the user data to create a well-defined distribution on the discrete variable of *number of reviews per user* on the body of the distribution. This is sound from a maximum likelihood perspective and, in concatenation with the tail Pareto distribution, forms a valid probability mass function (since this is a discrete variable). Below is a plot of this approach applied to the review-per-user distribution in the form of complementary cumulative distribution.

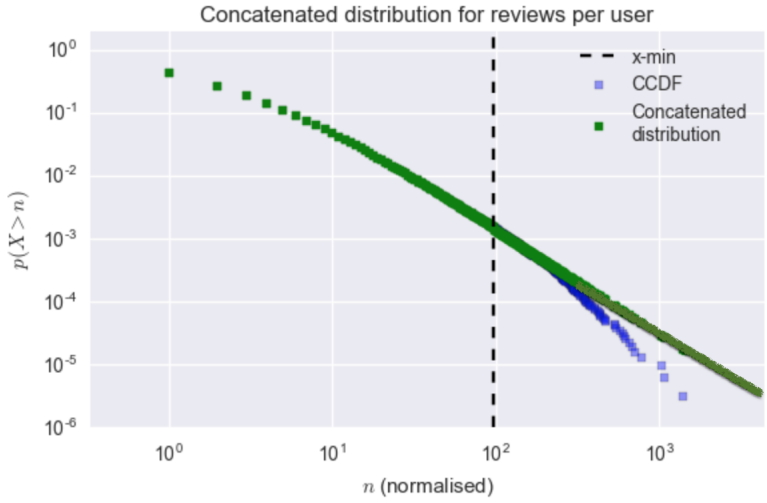

Figure 9: Concatenated distribution for reviews per user

However, this technique only works on the user data, since all values below x-min were observed in the empirical data, giving non-zero probabilities for each value. For the business data, there were many values below x-min for which there were no observations, which would lead this approach to assign zero probability to these values.

2.4 Conclusions

Clearly, our two distributions are very heavy tailed, and anyone wishing to perform statistical analysis using these distributions needs to take be cognisant of this, and take steps to allow for this in any analysis conducted.

3 Investigating seasonality in star ratings

3.1 Can we detect seasonality by visual inspection?

We wish to determine whether seasonality exists in star ratings. Below is an example of quarterly seasonality, created using made up data, where the seasonality itself is clearly visible.

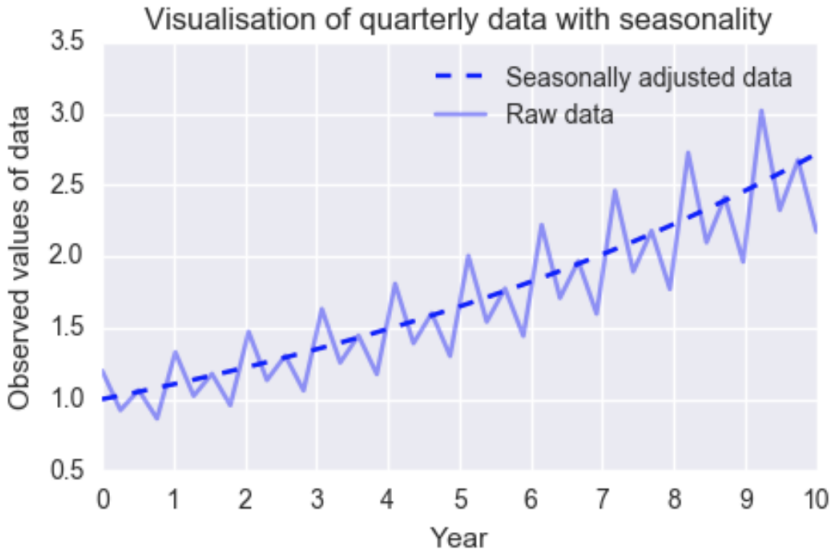

Figure 10: A demonstration of seasonality

Now we show a similar plot using star ratings from the Yelp dataset. The data is obtained by grouping reviews across the entire dataset by month, then taking the average rating per month. Only the years 2011-2015 are shown due to previous years having significantly fewer data available.

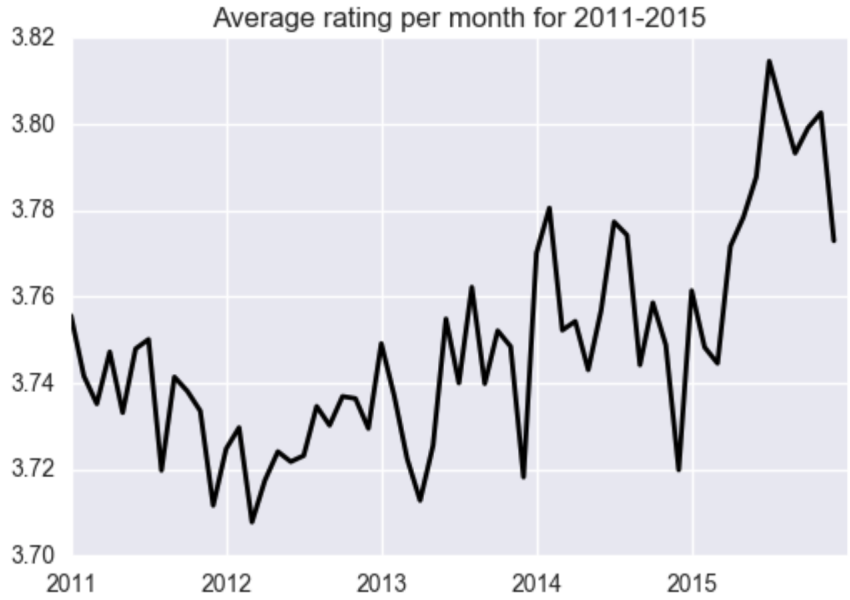

Figure 11: Star ratings over time

Clearly, if there is any seasonality, it is much less obvious than in our fictional case. Next, we group all reviews by month alone, and plot the mean of each monthly distribution, to see whether this gives us any indication of seasonality.

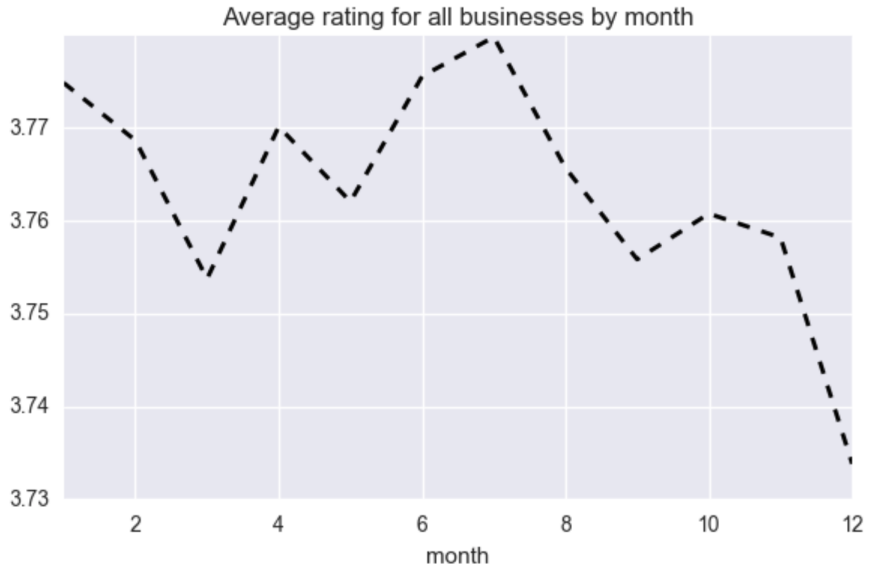

Figure 12: All star ratings grouped by month

We can observe that there is some volatility, and in particular, December ratings seem worse than ratings submitted during the rest of the year. Now we plot the same thing for some of most commonly reviewed

cities.

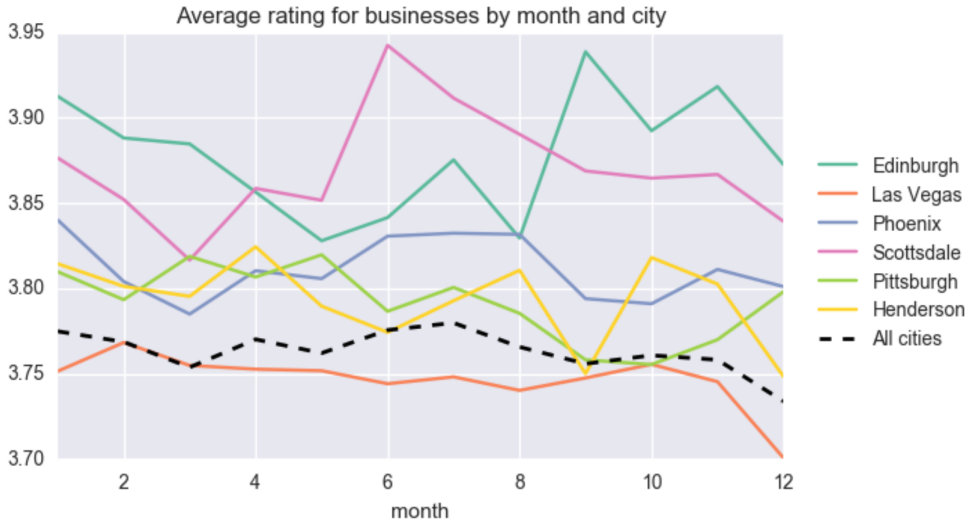

Figure 13: All star ratings grouped by month and city

Cities clearly do not behave uniformly with regard to seasonality, if there is indeed seasonality present.

3.2 Can we detect seasonality using statistical tests?

We now formalise the notion of seasonality.

Definition: Seasonality exists if samples of ratings taken over different time intervals are significantly different, as indicated by statistical tests.

There are several tests we can choose from the identify seasonality under this definition. One of the most established is the one-way ANOVA test. This test attempts to determine whether samples from independent groups have significantly different means. However, the test is based on the assumption that the underlying group distributions are of a Gaussian form. Therefore, we first need to determine whether this is a valid assumption to make for star ratings. Since we are testing twelve monthly distributions, we below show histograms for all twelve months.

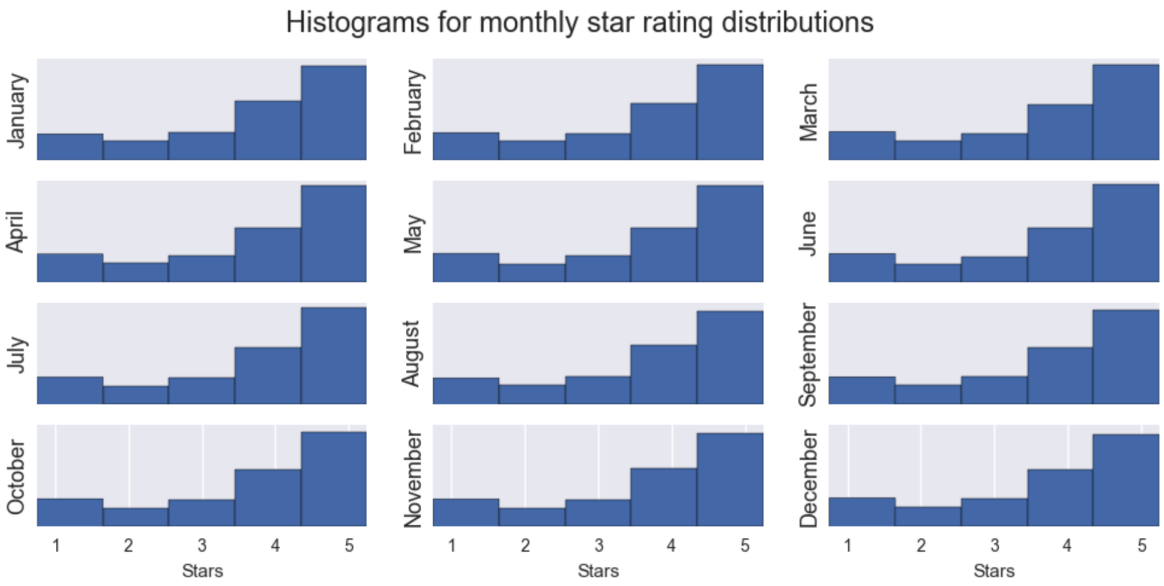

Figure 14: Histograms for monthly star rating distributions

All twelve distributions look very unlikely to be of a Gaussian form, given that the histogram is not monotonically increasing as the ratings increase towards the mean. However, we can formalise this by using the `SciPy Stats` library in `Python` to find the MLE Gaussian distribution parameters for each month, then check using the Kolmogarov-Smirnov test that the Gaussian distribution is indeed a poor fit. The results of this are presented below, and the p-values pertain to the null hypothesis that the data are Gaussian distributed.

Table 4: Testing monthly samples for Gaussian goodness of fit

Month	Mean	Variance	KS statistic	p-value
January	3.77	1.37	0.242	0.000
February	3.77	1.39	0.241	0.000
March	3.75	1.40	0.239	0.000
April	3.77	1.40	0.241	0.000
May	3.76	1.41	0.241	0.000
June	3.78	1.40	0.243	0.000
July	3.78	1.38	0.241	0.000
August	3.77	1.37	0.242	0.000
September	3.76	1.39	0.240	0.000
October	3.76	1.39	0.240	0.000
November	3.76	1.38	0.241	0.000
December	3.73	1.40	0.239	0.000

In all cases, we reject the null hypothesis of our data being generated by a Gaussian distribution. This shows us that it is inappropriate to make an assumption of these data being generated by a Gaussian distribution, which means that the one-way ANOVA test is not appropriate. The one-way Kruskal-Wallis test provides a non-parametric alternative to the one-way ANOVA, and does not rely on normality assumptions, so is therefore appropriate for this experiment. The test allows us to determine, given a desired statistical significance, whether a group of samples plauibly are generated from the same distribution. The null hypothesis of the Kruskal Wallis test states that the samples are generated by identical distributions. Therefore, if the null hypothesis can be rejected at our chosen 5% significance level, then at least two of

the monthly distributions differ statistically significantly. Using the `SciPy Stats` library in `Python`, we observe the following Kruskal-Wallis test statistics.

$$\text{KW} \approx 296 \qquad \text{and} \qquad p\text{-value} \approx 0.000$$

This means that we can reject the null hypothesis that the monthly samples are identically distributed. However, we can also test this hypothesis for each individual city. Below, results for the 30 most commonly reviewed cities are shown. Cities in which the p-value is less than 5% are highlighted in red. For these cities, under our previous definition, we have seasonality. For cities highlighted in blue, we fail to reject the null hypothesis that the monthly samples are all generated from the same distribution.

Table 5: Kruskal-Wallis p-values for most commonly reviewed cities

	P-value	# data
Las Vegas	0.0	1,032,055
Phoenix	0.0	366,308
Scottsdale	0.0	196,556
Charlotte	0.0	148,038
Tempe	0.001	105,776
Pittsburgh	0.009	104,396
Henderson	0.038	98,577
Montreal	0.019	77,031
Mesa	0.0	75,927
Chandler	0.209	73,842
Madison	0.415	58,090
Gilbert	0.26	54,775
Glendale	0.0	45,075
Edinburgh	0.0	30,792
Peoria	0.024	22,390
Surprise	0.021	14,177
Champaign	0.001	13,282
Goodyear	0.104	12,990
Avondale	0.209	9,525
Queen Creek	0.842	8,219
Cave Creek	0.046	6,399
Matthews	0.239	6,300
Urbana	0.152	5,589
Middleton	0.017	4,117
Fort Mill	0.076	4,115
Waterloo	0.018	3,723
Concord	0.086	3,423
Karlsruhe	0.093	3,332
Fountain Hills	0.044	3,209
Pineville	0.631	3,123

This demonstrates the need for business owners to look at an appropriate subset of the overall data when trying to determine whether they specifically are affected by seasonality. The results above verify that the data are heterogeneous. We can also run a similar analysis on types of business. The results for this are shown below.

Table 6: Kruskal-Wallis p-values for most commonly reviewed business types

	P-value	# data
Restaurants	0.0	1,630,712
Nightlife	0.0	352,386
Food	0.0	399,601
Bars	0.0	431,734
Breakfast & Brunch	0.0	179,951
Mexican	0.0	166,805
Arts & Entertainment	0.045	162,739
Shopping	0.0	151,744
Hotels & Travel	0.0	146,141
Event Planning & Services	0.009	145,371
Pizza	0.0	142,206
Beauty & Spas	0.0	140,769
Italian	0.0	133,552
Sandwiches	0.016	114,497
Burgers	0.001	113,332
Hotels	0.0	146,141
Japanese	0.012	94,302
Sushi Bars	0.001	90,027
Steakhouses	0.0	89,011
Coffee & Tea	0.0	85,074
Seafood	0.011	84,377
Chinese	0.03	82,643
Automotive	0.0	75,070
Casinos	0.001	72,750
Active Life	0.0	69,295
Home Services	0.0	65,074
Health & Medical	0.013	63,501
Fast Food	0.257	59,334

Here, only for fast food restaurants do we fail to reject the null hypothesis that monthly samples are generated from the same distribution. To explore seasonality further, we can analyse each pair of months. Since the Kruskal-Wallis test is an extension of the Mann-Whitney U test, which determines whether two samples are significantly differently distributed, we can use the Mann-Whitney U test to explore inter-month relationships. P-values for each pair of months for all reviews in the dataset are shown below. In each case, our null hypothesis assumes, for each pair of samples, that the samples both arise from the same distribution. Again, `Python`'s `SciPy Stats` library was used for calculations.

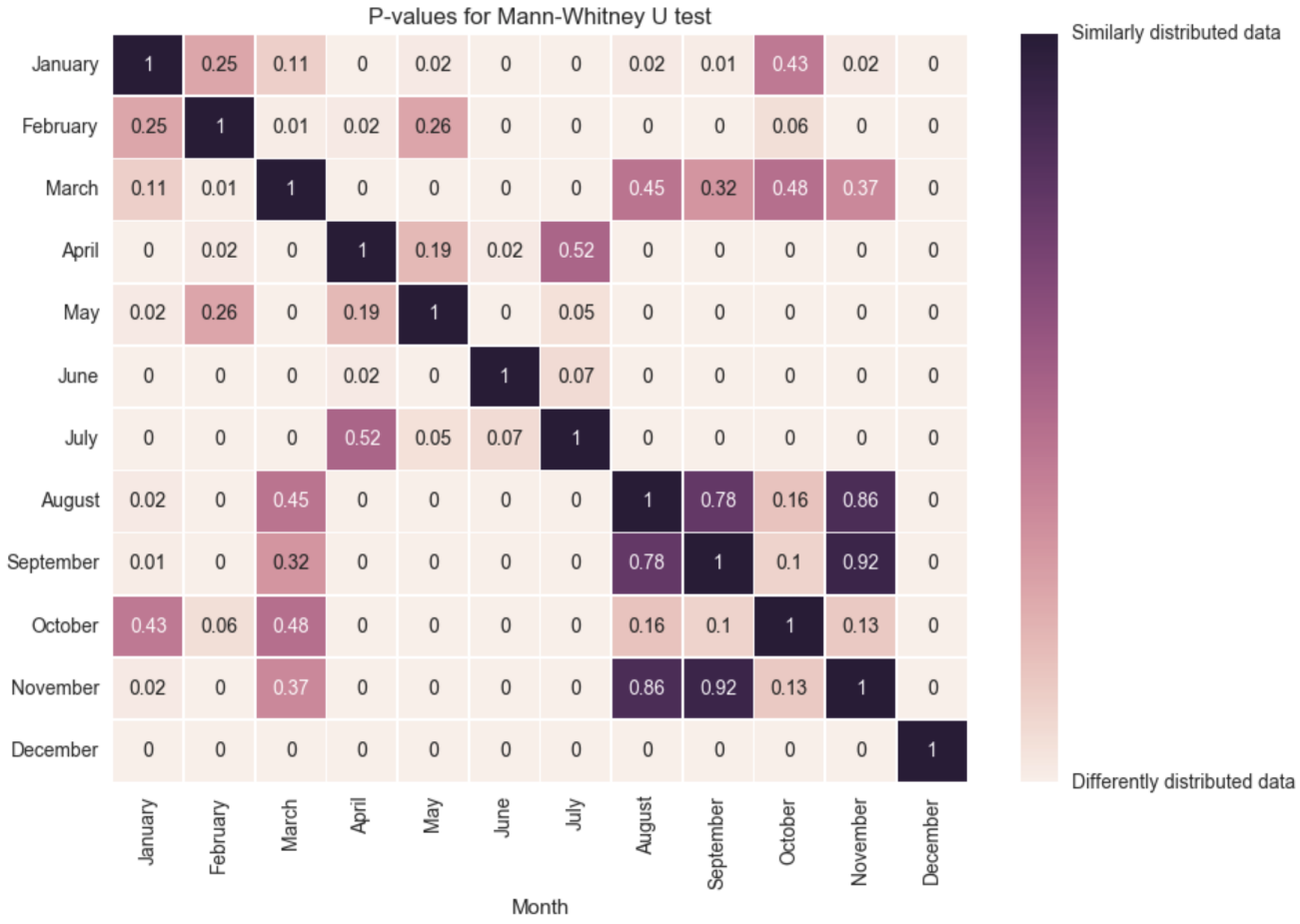

Figure 15: Mann-Whitney U test p-values on pairs of months

We can observe from the abundance of zeros that most pairs of months seem to have samples that are generated by significantly different distributions. In fact, of all 66 pairs of months, 48 pairs, representing 73% of the total, seem to be generated from two different distributions. This strongly indicates that seasonality is present in the dataset as a whole. It is also worth noting the zeroes along the bottom row of the matrix, which indicate that December ratings are not similarly distributed to any other month, which is consistent with our earlier observation that December ratings seem markedly worse than ratings submitted during other months. We can also perform this analysis for subsets of the total dataset and once again filter by business type and location, but this analysis is not presented here due to how many pages of this report it would consume.

3.3 Can we find other variables to explain seasonality?

Our aim here is to find variables which are seasonal by nature, and to investigate whether these seasonal variables have any explanatory power over ratings. An obvious place to start is looking at the weather. A reasonable hypothesis is that a reviewer's mood may be affected by weather, and this may impact star rating awarded. It stands to reason that in winter, given generally harsher weather conditions, there may be a negative bias to ratings, and in the summer, the converse may hold. To conduct this experiment, weather data were downloaded using Weather Underground's API service. Since each daily weather reading per

location required its own API call, downloading all weather data for each location in the dataset was not scalable, so daily temperature data for only the ten most commonly reviewed cities was downloaded for the years 2011-2015. Approximately 73% of reviews in the dataset fall inside the time period 2011-2015, and of those reviews, 85% are reviews of businesses in one of the top ten most commonly reviewed cities.

For this experiment, average daily temperature was correlated (using Pearson correlation coefficient) with average star rating for each city in the selection over the five years. The results are shown below. The cities for which the correlation is statistically significant at a 5% significance level are shown in red. Using Spearman rank correlation instead of the Pearson approach yielded similar results.

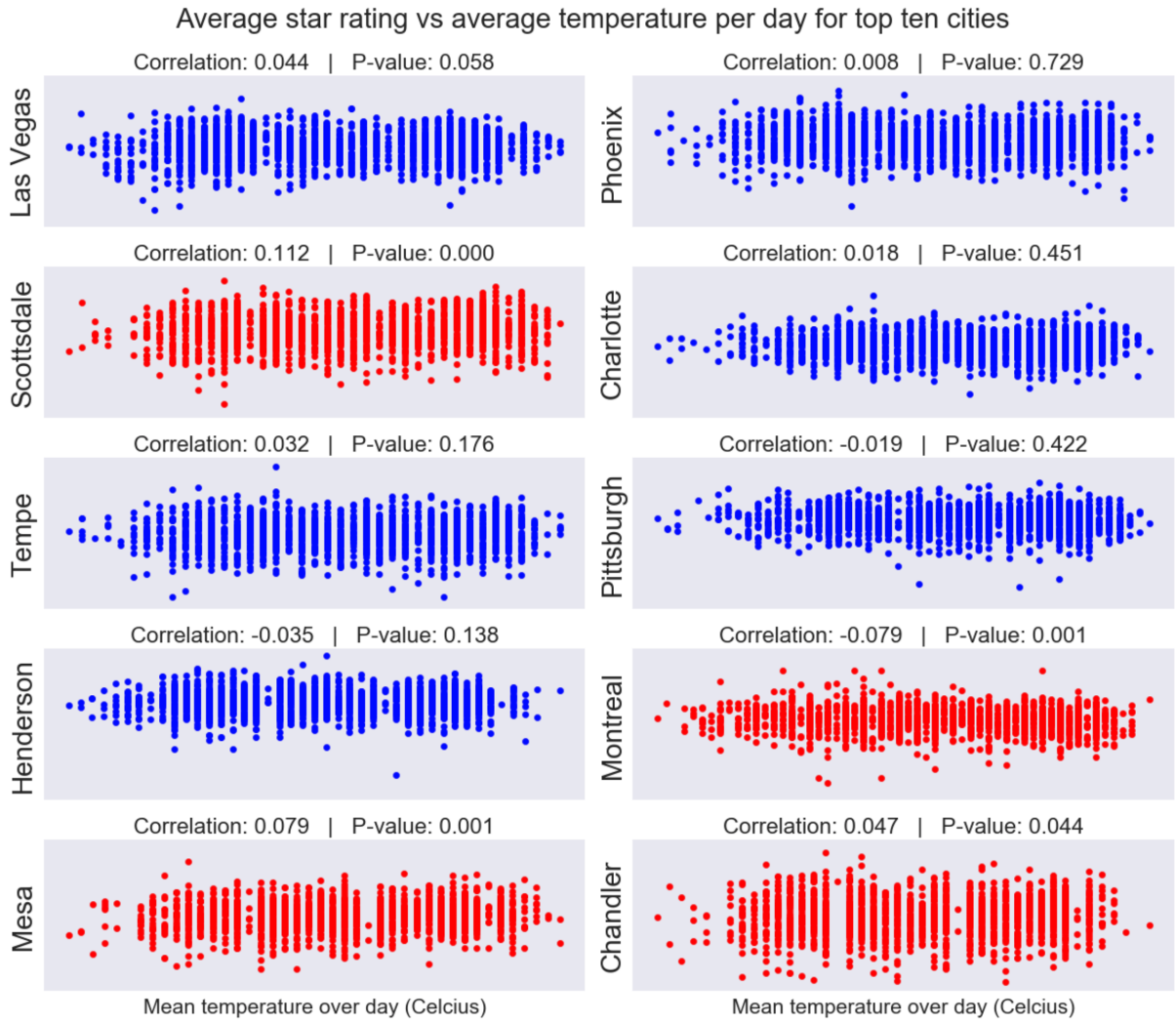

Figure 16: Scatter plots: temperature and star ratings per city

We can also run this analysis over all ten cities, and correlate all observed pairs of average temperature and average star rating.

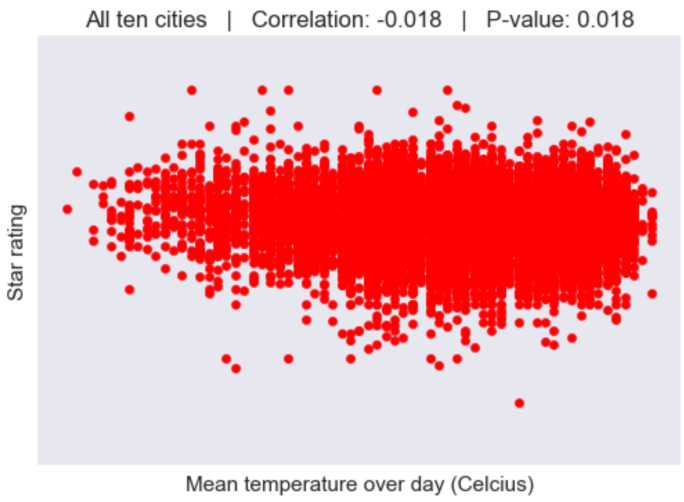

Figure 17: Scatter plot showing average temperature vs star rating aggregated over all cities

Interestingly, although only one city in the plot on the previous page displayed negative correlation, the correlation for the aggregated data does indeed exhibit negative correlation behaviour, and is statistically significant at a 5% level. On the other hand, it is worth noting that coefficients of determination are all small, with Scottsdale having the largest coefficient of determination, at 0.0125. However, the above plot is not very useful, since it masks the fact that distributions in cities display abundant heterogeneity. For instance, a Scottsdale business owner should expect more favourable reviews with higher temperatures, and a glance at the aggregated relationship would be misleading.

3.4 Conclusions

According to our definition, there seems to be a significant amount of seasonality in the data. However, while it is tempting to look at the entire dataset for signs of seasonality, the key for business owners is to consider the heterogeneity of the dataset and look at seasonality only in relevant subsets of the data. This might mean filtering by both business type and location. Once a stakeholder arrives at the conclusion that there is seasonality, a seasonal adjustment may have to be made. This would essentially allow the business owner to filter away a certain amount of "noise". Running a business has the potential to be a very stressful occupation. As a business owner, knowing how much of one's change in Yelp ratings is down to seasonal effects, and how much is therefore down to genuine change in customer experience, has the potential to take one source of stress off the table.

However, there are limitations in this approach. It may be the case that the majority of reviews are not written on the same day that the reviewer experienced the services of the business. There may also be other factors related to weather, such as humidity and precipitation, that correlate with ratings. Additionally, temperature data was rounded to the nearest integer, and thus the variable's true continuous nature has been compromised in the data collection period. More precise data may be more powerful with respect to discovering relationships between variables. Furthermore, the relationship between rating and temperature may be non-linear. This is analogous to the way that higher temperatures only indicate more voluminous ice cream sales up until the point that it is too hot for consumers to venture outside. Non-linearity has the potential to render correlation ineffective.

There are also additional data that can be used to further detect seasonality. For example, tourism data may be able to explain some of the variation in star ratings.

4 Appendices

.1 Software packages used

.1.1 `Python`

All computations and visualisations were generated using the `Python` programming language. Several Python libraries were used to support these computations and visualisations.

.1.2 `NumPy`

The `NumPy` library supports efficient matrix- and vector- like computation, as seen in languages like `R` and `MATLAB`.

.1.3 `Pandas`

This library allows flexible manipulation of tabular data, and this analysis was carried out in its entirety with data stored as `Pandas.Dataframe` objects.

.1.4 `SciPy.Stats`

`SciPy.Stats` functionality allows the user to fit data to a wide range of statistical distributions using Maxmimum Likelihood Estimation. The library also provides many statistical test implementations, including all tests performed in this report.

.1.5 `Matplotlib` & `Seaborn`

These libraries provide plotting functionality, and were used to generate all plots in this report.

.1.6 `Powerlaw`

This library provides a range of plotting and parameter estimation tools for Powerlaw distributions.

.2 Probability distributions used in this report

.2.1 Weibull-min (also known as Frechet right)

This distribution is characterised by the following cumulative distribution function:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp(-(\frac{x}{\lambda})^{\alpha}), & x \geq 0 \end{cases}$$

.2.2 Gaussian distribution (univariate)

This distribution is defined using two parameters, μ and λ. The probability density function is shown below.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

.2.3 Pareto distribution

This distribution is characterised by the following cumulative distribution function. Parameters are x_m and α.

$$F(x) = \begin{cases} 1 - (\frac{x_m}{x})^{\alpha}, & x \geq x_m \\ 0, & x < x_m \end{cases}$$

Once x_m is determined, α is readily solvable via

$$\alpha = \frac{N}{\sum_{i=1}^{N} \ln \frac{x_i}{x_m}}$$