p1

February 27, 2023

1 Project #1:

What Determines Business' Yelp Ratings in Toronto?
by Thomas Keough

1.1 Introduction

Customers can easily access reviews for virtually anything on the internet. Yelp is a popular place to look for said reviews when visiting restaurants, hair salons, dentists, and other businesses of all kinds. Both businesses and customers should be interested in identifying what causes a business to earn good reviews, which this analysis seeks to answer. Business owners can optimize their perception by understanding which factors contribute most to positive reviews. Customers can make better decisions regarding which businesses to visit by understanding which metrics are most indicative of good service.

This analysis using data acquired from Kaggle.com. The data was initially published by Yelp as part of a dataset challenge. It was last updated in August 2017 and includes data from 11 cities in four countries, including the USA and Canada. Included in the dataset are files pertaining to business attributes, hours of operation, reviews, check-ins, tips, and users. This analysis merges the files for business attributes, hours of operation, and check-ins. This analysis is concerned with businesses in Toronto, Ontario. The dependent variable is business ratings, measured in stars on Yelp. The covariates chosen include business' total number of Yelp reviews, total weekly hours of operation, and daily average number of Yelp check-ins.

The results of this analysis demonstrate that the number of reviews, hours of operation, and number of check-ins are all positively correlated with business ratings. For future analyses, the direction of causality and the relevance of these covariates in determining Toronto business ratings on Yelp should be identified.

1.2 Data Cleaning

(i) Setting Up The Project

Y: Business Rating ('stars')
This analysis seeks to identify the determinants of business ratings using quantitative data gathered from Yelp. The outcome variable is the business rating, which is referred to as "stars" in the dataset. This variable takes a value between 1 and 5. It is discrete data; potential values must be multiples of 0.5. This analysis includes three covariates of interest.

X1: Number of Reviews ('review_count')
The first covariate is the total number of reviews for each business. It is discrete. The review

system is an extremely prominent feature of Yelp and is used frequently, so the number of reviews should account for a significant proportion of the variation in business ratings. There are two valid hypotheses for the correlation between the number of reviews with business ratings. For instance, a business that has exceptionally good service can encourage Yelp users to leave positive reviews. This would suggest that the number of reviews is positively associated with business ratings. However, the opposite effect is also feasible; exceptionally poor service may encourage Yelp users to leave negative reviews, meaning the number of reviews would be negatively correlated with business ratings. The dominant hypothesis, should one exist, will manifest itself through an empirical analysis of the relationship between review count and ratings.

X2: Number of Operating Hours per Week ('wk_op_hours')
The second covariate is the number of operating hours in a week for a business. Businesses differ largely in hours of operation on a day-to-day basis, so using the weekly total of operating hours will account for those differences across the days of the week. This variable will provide insight on the effect of customer access to businesses, which may be a component in evaluating business quality from the perspective of customers. There are several hypotheses for the direction of the relationship between operating hours and business ratings. It could be that businesses that are open longer throughout the week earn more reviews because they can serve more customers. Therefore, the direction of the relationship would be dependent on the ambiguous effect of the number of reviews on business ratings. Alternatively, businesses that are open for fewer hours per week may instill a sense of exclusivity in its customers. For example, fine dining venues, nightclubs, and other businesses that have limited weekly operating hours may observe more positive reviews because attendees feel exclusive.

X3: Average Number of Check-ins per Day ('daily_checkin_avg')
The third covariate is the average number of Yelp check-ins per day. It is important to note that this calculation of the daily check-in average ignores days wherein businesses had zero check-ins. In other words, it is the average number of check-ins for days that had at least one check-in. "Checking in" is a Yelp feature that allows users to inform their Yelp following of the businesses they visit. When a user "checks in" to a business, their attendance is published on their profile. Users can also earn badges and special offers at businesses they check in at. These two components provide increased incentives for Yelp users to visit high-quality businesses; users can show off their attendance at a popular business and potentially earn discounts in the process. Comparatively, poor-quality businesses would have fewer check-ins since users would not want to publish their attendance at them. Therefore, this variable accounts for customer perception of businesses. Therefore, check-ins should be positively correlated with a business's ratings in theory. An OLS regression that utilizes this covariate and the number of reviews would provide insight on how the number of reviews affects a business's rating while holding constant the customers' perception of the business. This would provide insight on the causality of business ratings.

(ii) Data Cleaning

[ ]:
```
import pandas as pd
%matplotlib inline
```

[ ]:
```
# load in data
df = pd.read_csv('/Users/thomas/Documents/schoolwork/eco225/yelp_data/
  ↪yelp_business.csv')
```

```
df2 = pd.read_csv('/Users/thomas/Documents/schoolwork/eco225/yelp_data/
 ↪yelp_checkin.csv')
df3 = pd.read_csv('/Users/thomas/Documents/schoolwork/eco225/yelp_data/
 ↪yelp_business_hours.csv')

# make copies
bsn_df = df.copy()
checkin_df = df2.copy()
hours_df = df3.copy()
```

[ ]:
```
# build Toronto businesses dataframe:

# make a df for Toronto businesses
tor_bsn = bsn_df.loc[bsn_df['city'] == 'Toronto']

# select columns
tor_bsn = tor_bsn[['business_id', 'name', 'stars', 'review_count',␣
 ↪'categories']]

# generate a restaurant dummy
tor_bsn['restaurant'] = tor_bsn['categories'].str.contains('Restaurant').
 ↪astype(int)

# Incorporate check-in data:

# find average daily checkins for each business
all_checks = checkin_df.groupby('business_id').mean().
 ↪rename(columns={'checkins': 'daily_checkin_avg'})

# inner join on tor_bsn to get checkin data for each business in Toronto
tor_bsn = tor_bsn.merge(all_checks, on='business_id')

# Incorporate weekly operating hours:

# build a function to clean hours_df

def count_hours(operating_hours: str) -> float:
    """Returns a float given a string of the form "HH:MM-HH:MM" that contains a␣
 ↪business's operating hours."""

    # if business is closed on the given day:
    if operating_hours == 'None':
        return 0.0

    # '2000-01-01' is needed in the string that is converted to datetime to␣
 ↪avoid being out of pandas' accepted date range
```

```
    hours = operating_hours.split('-')
    opn = pd.to_datetime('2000-01-01 ' + hours[0])
    close = pd.to_datetime('2000-01-01 ' + hours[1])

    # if the business is open past midnight:
    if close < opn:
        end_date = '2000-01-02'
    else:
        end_date = '2000-01-01'

    # calculate a timedelta that accounts for hours of operation past midnight
    window = pd.to_timedelta(pd.to_datetime(f'{end_date} ' + hours[1]) - pd.
 ↪to_datetime('2000-01-01 ' + hours[0]))

    # return a float reprsenting the number of hours open in the given day
    return window.total_seconds() / (60 * 60)

# merge hours_df with tor_bsn to avoid running count_hours on unneeded␣
 ↪businesses
tor_bsn = tor_bsn.merge(hours_df, on='business_id')

# generate total weekly operating hours for each Toronto restaurant
tor_bsn['wk_op_hours'] = 0

# sum operating hours for each day - monday thru sunday are columns 8-14
for day in tor_bsn.columns[-8:-1]:
    tor_bsn['wk_op_hours'] += tor_bsn[day].apply(count_hours)

# remove daily operating hours columns
tor_bsn = tor_bsn.drop(columns=['monday', 'tuesday', 'wednesday', 'thursday',␣
 ↪'friday', 'saturday', 'sunday'])

# order columns
tor_bsn = tor_bsn[['business_id', 'name', 'stars', 'review_count',␣
 ↪'daily_checkin_avg', 'wk_op_hours', 'restaurant', 'categories']]

# Toronto dataframe is now clean
tor_bsn
```

[ ]:

	business_id	name	stars	\
0	l09JfMeQ6ynYs5MCJtrcmQ	"Alize Catering"	3.0	
1	lHYiCS-y8AFjUitv6MGpxg	"Starbucks"	4.0	
2	VSGcuYDV3q-AAZ9ZPq4fBQ	"Sportster's"	2.5	
3	1K4qrnfyzKzGgJPBEcJaNQ	"Chula Taberna Mexicana"	3.5	
4	AtdXq_gu9NTE5rx4ct_dGg	"DAVIDsTEA"	4.0	
…	…	…	…	
14845	sEAKw3MZkER1u_1fzIeD3g	"Gol Take-Out"	4.0	

```
14846  1HplwLVbBid-Bgw1sEPGFg       "Dumpling Melody Bistro"    2.0
14847  dWoAayHRyIrkk1dcvBxv3Q           "Art Ink Collective"    3.5
14848  SvW3WsatQWvR8c1iwAD_QA             "Urban House Cafe"    4.0
14849  nGjEV4bnODPk8bcb0C6Aig  "Sweet Serendipity Bake Shop"    4.5

       review_count  daily_checkin_avg  wk_op_hours  restaurant  \
0                12           1.000000         91.0           1
1                21           4.577586        115.5           0
2                 7           2.125000         70.0           0
3                39           1.333333        103.5           1
4                 6           1.272727         70.5           0
…               …                 …           …          …
14845            15           1.000000          0.0           1
14846            12           1.125000          0.0           1
14847             3           1.000000         43.0           0
14848            32           2.000000         91.0           1
14849            22           1.000000         45.0           0

                                              categories
0                            Italian;French;Restaurants
1                                    Food;Coffee & Tea
2                           Bars;Sports Bars;Nightlife
3          Tiki Bars;Nightlife;Mexican;Restaurants;Bars
4                          Coffee & Tea;Food;Tea Rooms
…                                                  …
14845  Food;Restaurants;International Grocery;Ethnic …
14846                                Restaurants;Chinese
14847  Shopping;Beauty & Spas;Piercing;Art Galleries;…
14848  Nightlife;Restaurants;Sandwiches;Bars;Canadian…
14849                                    Bakeries;Food

[14850 rows x 8 columns]
```

1.3 Summary Statistics

[ ]:
```
# generate summary statistics for y, x1, x2, x3
tor_bsn[['stars', 'review_count', 'wk_op_hours', 'daily_checkin_avg']].
 ↪describe()
```

[ ]:
```
              stars  review_count   wk_op_hours  daily_checkin_avg
count  14850.000000  14850.000000  14850.000000       14850.000000
mean       3.494007     28.295556     47.315771           1.522870
std        0.841782     56.253223     35.878516           1.090018
min        1.000000      3.000000      0.000000           1.000000
25%        3.000000      5.000000      0.000000           1.000000
50%        3.500000     10.000000     54.000000           1.187500
75%        4.000000     28.000000     74.500000           1.555556
```

```
max        5.000000    1494.000000     167.883333          31.042553
```

Stars

The summary statistics for *stars* indicate that the data is discrete. Moreover, it demonstrates that the minimum and maximum ratings are 1 and 5 respectively. The average review is roughly 3.5 stars. Assuming that an average business should have a rating near the midpoint between the minimum and maxmimum, the mean may indicate that there is a systematic bias towards over-rating businesses by 0.5 stars at the aggregate level. The interquartile range is between 3 and 4 stars, which further suggests this bias; it demonstrates that half of all businesses in Toronto are above the rating midpoint. In other words, only 25% of business can be considered poor-quality (i.e. a rating below 3 stars).

Number of Reviews

There appears to be a great variation in the number of reviews across businesses but it is primarily due to an enormous right skew. The standard deviation (56) is double the mean review count (28). The maximum review count of 1494 is indicative of right skew since it is 53 times larger than the average. Moreover, 75% of all Toronto businesses have 28 reviews or less. This may yield low power in a regression analysis since the vast majority of businesses have very similar review counts, despite the extreme variation suggested by the standard deviation.

Weekly Operating Hours

On average, businesses in Toronto are open for roughly 47 hours per week. The standard deviation is roughly 36 hours per week which indicates that there is a large variation in weekly operating hours across businesses. There are several signs of right skew in this variable, which may be the cause of the large standard deviation. The 50th percentile (54) is larger than the mean, which suggests that outliers on the right side of the distribution are positively biasing the mean. Moreover, this variable's lower bound of zero suggests that most of the variation would occur above the mean. The maximum value observed for this variable is 167.88, which is 0.12 hours less than the total number of hours in a week. This observation may need to be dropped from analysis if it biases the trends of interest, since the vast majority of businesses cannot operate for nearly every hour of the week.

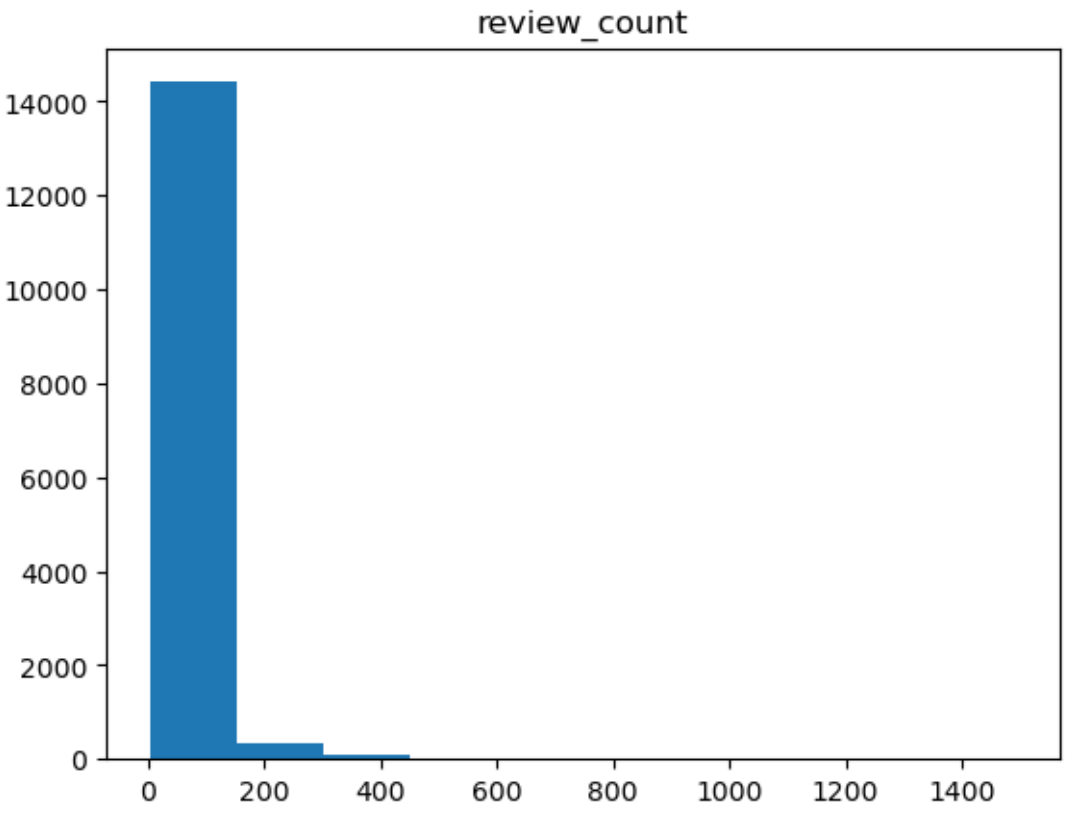
Daily Check-in Average

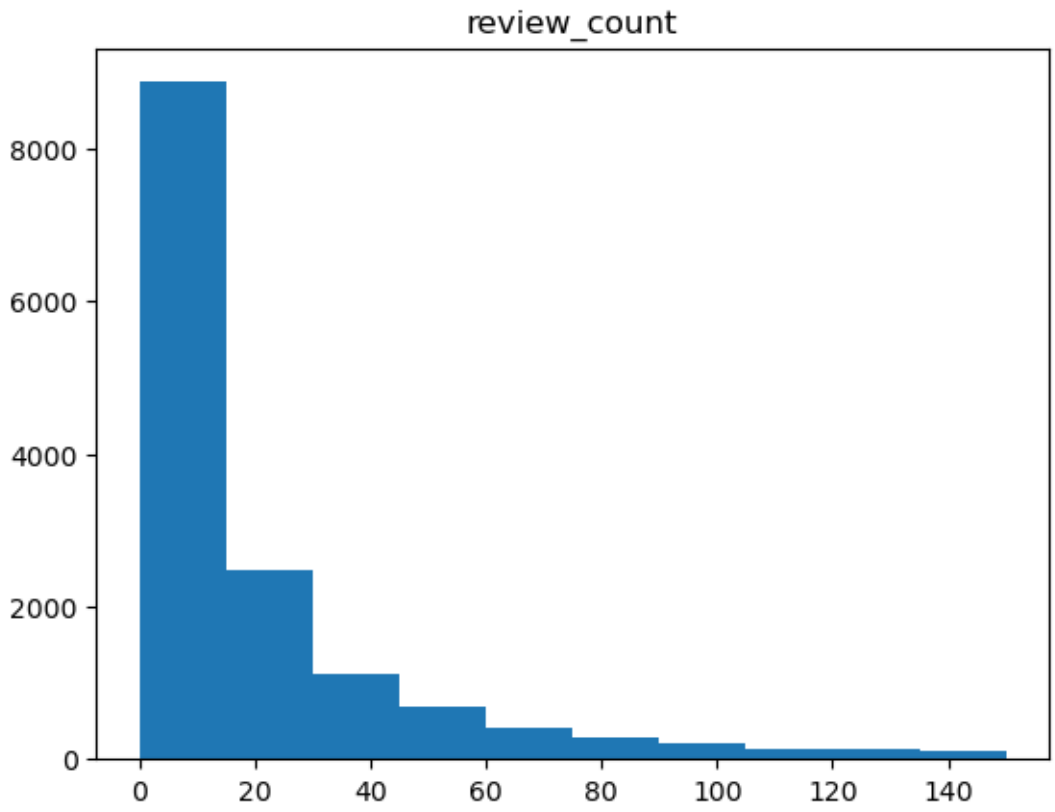
There is very little variation across businesses in their daily check-in averages. Despite the mean and standard deviations being approximately 1.52 and 1.10 respectively, 75% of the observations lie between 1 and 1.55. There is a significant right skew in this data since the maximum value observed is roughly 31 check-ins on average. This variable may prove irrelevant in a regression analysis due to the uniformity of the data: there is not enough variation to properly calculate how an outcome variable changes given a change in the daily check-in average.

1.4 Plots & Figures

[ ]:
```
# histograms of review_count
tor_bsn.hist('review_count', grid=False)
tor_bsn.hist('review_count', grid=False, range=(0,150))
```

[ ]:
```
array([[<AxesSubplot: title={'center': 'review_count'}>]], dtype=object)
```


review_count
14000
12000
10000
8000
6000
4000
2000
0
0
200
400
600
800
1000
1200
1400



As demonstrated by the summary statistics table, review_count has significant right skew. Ignoring businesses with more than 150 reviews yields a clearer picture: the distribution of reviews is unimodal with a peak between 0 and ~15. Given the lack of variation, the number of reviews may not be a significant determinant of business ratings. In other words, the variation in business ratings cannot be meaningfully explained by the number of reviews alone because the majority of businesses are very similar in this dimension. This variable must be used in a multiple regression alongside additional covariates to generate statistically significant coefficients and identify the relevant determinants of business ratings.

[ ]:
```
# relationship between business rating and review count
tor_bsn.plot(kind='scatter', x='review_count', y='stars', c='b', title='Number␣
 ↪of Reviews & Business Rating, With Outliers')
tor_bsn.plot(kind='scatter', x='review_count', y='stars', c='b', xlim=(0,700),␣
 ↪title='Number of Reviews & Business Rating, Without Outliers')
```

```
/Users/thomas/opt/anaconda3/lib/python3.9/site-
packages/pandas/plotting/_matplotlib/core.py:1114: UserWarning: No data for
colormapping provided via 'c'. Parameters 'cmap' will be ignored
  scatter = ax.scatter(
```

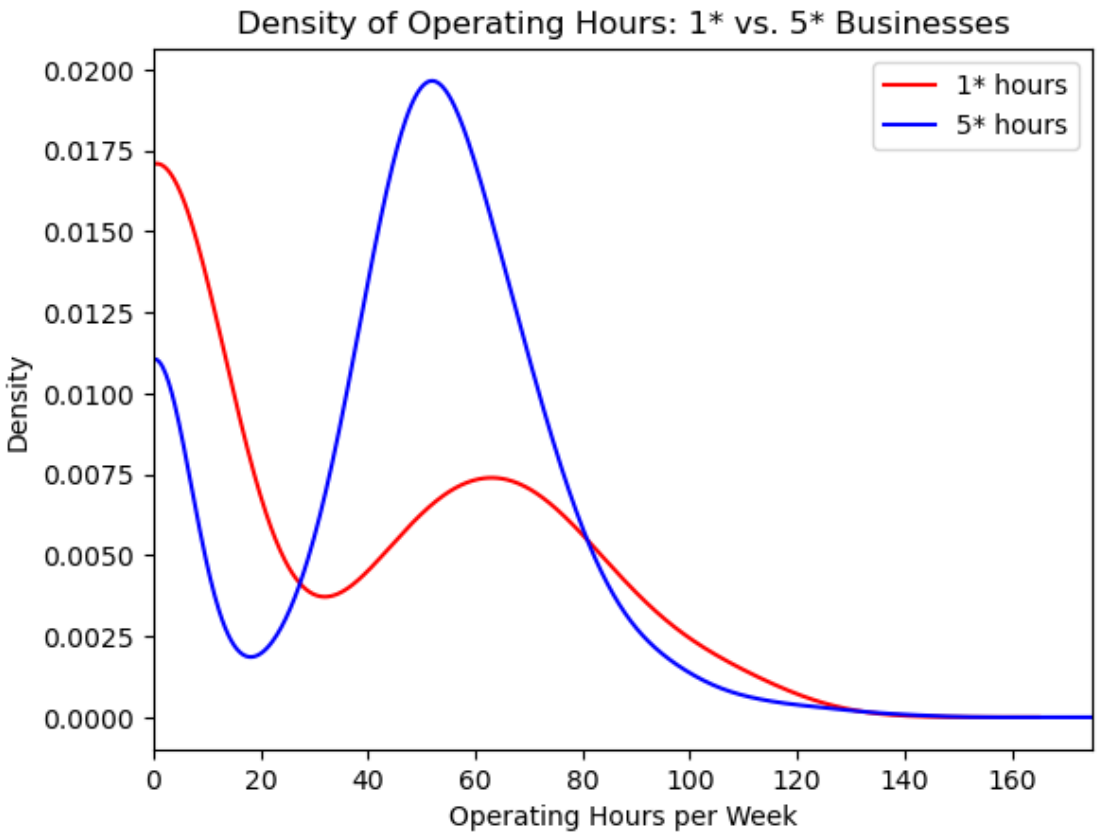
[ ]: <AxesSubplot: title={'center': 'Number of Reviews & Business Rating, Without Outliers'}, xlabel='review_count', ylabel='stars'>

This scatter plot demonstrates that a business's number of reviews is positively correlated with the business's rating. This suggests that the effect of good service dominates the effect of bad service in regard to generating reviews for businesses. If the poorest quality businesses had the greatest number of reviews, then the effect of bad service would dominate. There appears to be non-linearity in this relatationship: the marginal benefit of an additional review decreases significantly between the 200 and 300 review marks. However, this non-linearity may simply be a result of the data for *stars* being discrete. The true effect of the number of reviews on a business's rating is not clear through this plot. Controlling for additional variables, such as the number of checkins on average per day (i.e. customer perception), may present a relationship that differs from the one seen here.

[ ]:
```
# density of operating hours for businesses with 1*, 5* reviews
tor_1s = tor_bsn.groupby('stars').get_group(1).rename(columns={'wk_op_hours':␣
 ↪'1* hours'})
tor_5s = tor_bsn.groupby('stars').get_group(5).rename(columns={'wk_op_hours':␣
 ↪'5* hours'})

tor_1s['1* hours'].plot(kind='kde', c='r', legend=True, title='Density of␣
 ↪Operating Hours: 1* vs. 5* Businesses', xlim=(0,175)).set_xlabel("Operating␣
 ↪Hours per Week")
tor_5s['5* hours'].plot(kind='kde', c='b', legend=True)
```

[ ]: <AxesSubplot: title={'center': 'Density of Operating Hours: 1* vs. 5* Businesses'}, xlabel='Operating Hours per Week', ylabel='Density'>



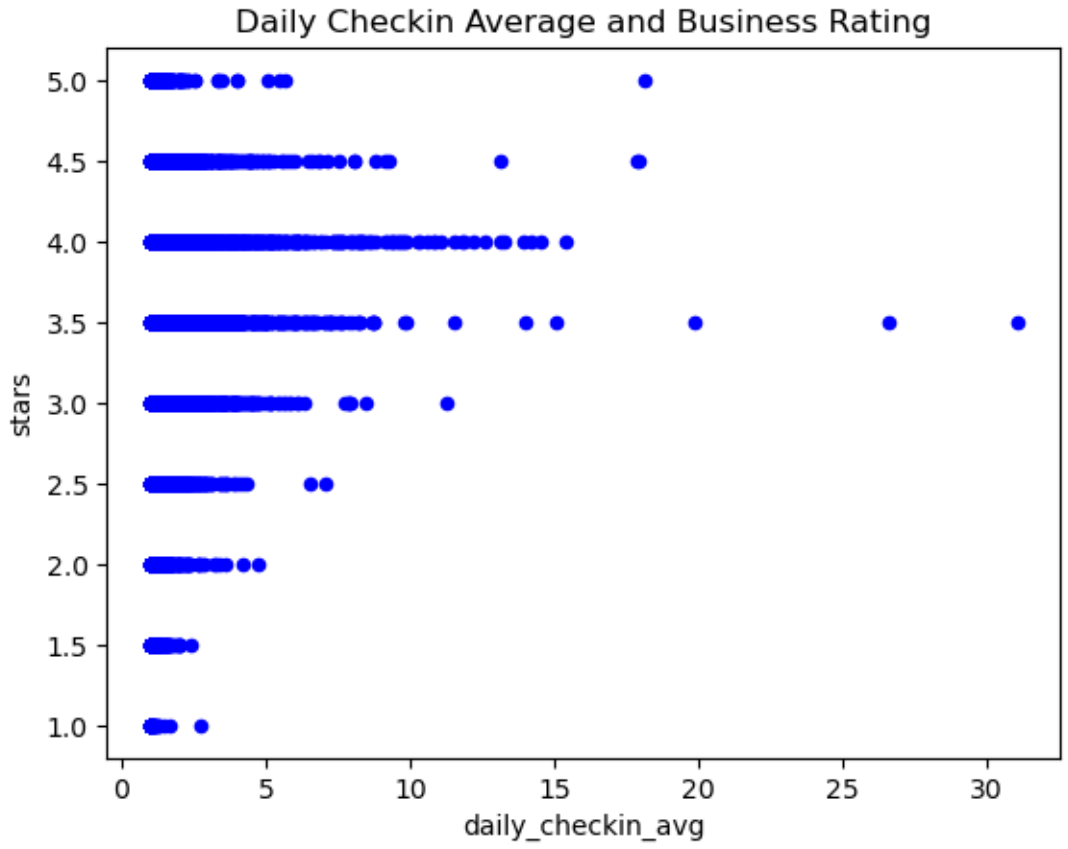
This density plot demonstrates how weekly operating hours differ across businesses at either end of the rating spectrum. Five-star businesses are more likely to have between 45-65 hours of operation in any given week, while one-star businesses are more likely to have less than 30 hours of operation. This evidences the hypothesis that customer access plays a role in determining business ratings. However, both distributions are bimodal have a peak at 0 hours of operation per week. This is likely because business hour data is missing from many businesses on Yelp. Since there is more missing data for one-star businesses, it is not certain that five-star businesses truly have more hours of operation in a week on average. Rather, it could be that five-star businesses are more likely to publish their hours of operation to Yelp which is indicative of high quality in a dimension that is unrelated to customer access. In order to identify if higher-quality businesses tend to be open for more hours in a week in general, an analysis including businesses of all ratings should be peformed.

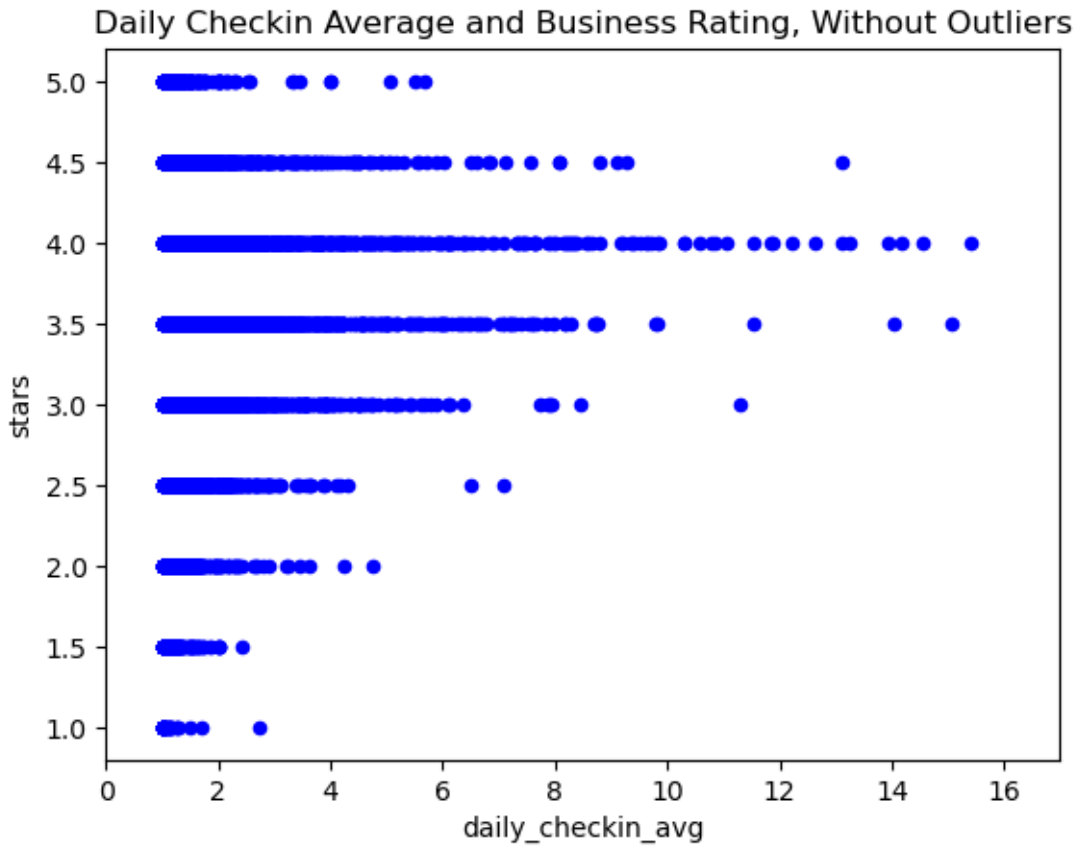
[ ]:
```
# relationship between businsess rating and checkins
tor_bsn.plot(kind='scatter', x='daily_checkin_avg', y='stars', c='b',␣
 ↪title='Daily Checkin Average and Business Rating')
```

```
tor_bsn.plot(kind='scatter', x='daily_checkin_avg', y='stars', c='b',␣
↪title='Daily Checkin Average and Business Rating, Without Outliers').
↪set_xlim(0, 17)
```

```
/Users/thomas/opt/anaconda3/lib/python3.9/site-
packages/pandas/plotting/_matplotlib/core.py:1114: UserWarning: No data for
colormapping provided via 'c'. Parameters 'cmap' will be ignored
  scatter = ax.scatter(
```

[ ]: (0.0, 17.0)





There is a clear positive association between the average number of check-ins per day and the business rating. Non-linearity also appears to be present but this may be a result of business ratings being measured discretely. There are several massive outliers beyond roughly 17 average check-ins per day, which should be excluded from a formal regression analysis as they exacerbate the non-linear trend. Overall, this relationship is very similar to that of the total number of reviews and business ratings. As a result, controlling for both variables in a regression may be redundant and reduce power. The average daily check-in variable is likely to be a worse predictor of business ratings than the count of reviews since it has relatively less variation.

Using the check-in data as a covariate may present issues with reverse causation as well. Assuming customers use the check-in feature to boast about the businesses they visit, a higher Yelp rating would increase the number of check-ins a business receives. This complicates the identification of a causal relationship between these variables because either the business rating or the average number of check-ins must be used as a dependent variable. Though the issue of reverse causation may not be easily solved with the current dataset, a multiple regression including the other covariates (review_count, wk_op_hours) would provide greater insight into the causality of business ratings.

1.5 Conclusion

The analysis demonstrates that the number of reviews, weekly hours of operation, and average number of check-ins per day all appear to be positively correlated with business ratings in Toronto

on Yelp. In other words, high-quality businesses tend to have more reviews, are open more often, and receive more check-ins on average relative to poor-quality businesses. However, this conclusion can only be drawn from each covariate's relationship with business ratings alone. In other words, the direction and magnitude of causality among these variables, if any exists, remains unknown. This analysis is unable to determine how relevant the chosen covariates are in determining business ratings.

A multiple regression model with these variables would contribute to a better understanding of the causality of business ratings. It allows for control variables which would yield more precision in determining the effects of any individual covariate. Moreover, it would generate an R^2 value to indicate how relevant the covariates are. This would provide insight on whether the total number of reviews and average number of check-ins per day are relevant predictors, considering they have limited variation.