Airbnb Data Exploration:

How Does COVID-19 Affect the Price of San Francisco Airbnbs?

—

Rae Godfredsen & Ally Martin

Introduction to the Dataset: Airbnb

- collects data monthly for 100+ cities worldwide
- this data: subset of SF Airbnb's collected in June 2020
    - four tables: calendar, listings, ratings, evictions

```
calendar.sample(5)
```

listing_id	date	available	price
905902	2020-08-12	t	135
23581750	2021-02-01	t	280
9258941	2021-01-20	f	275
10886231	2021-01-26	t	95
11886255	2021-02-04	f	82

Important Tables for Analysis: `calendar` & `calendar_feb`

- `calendar` table: contains price of each listing over the next 365 days
- `calendar_feb` table: contains the same information as `calendar` table except the data was collected in February 2020 (before COVID)
    - *additional table we loaded
- important columns in these tables: listing_id, price (USD), and date

Filtering `calendar` & `calendar_feb` to contain dates from July-August ...

- 'date' column is a string
- defined function 'yr' and 'mon' to convert the dates into integers
- applied function to both tables

```
#function to convert dates from string to int
def yr(date):
    year = date[0:4]
    return int(year)

def mon(date):
    month = date[5:7]
    return int(month)
```

```
#2020 June Calendar Table Filtered
calendar=calendar.with_columns('year',calendar.apply(yr,'date'),'month',calendar.apply(mon,'date'))

#2020 June Calendar Table containing data for months July and August
calendar=calendar.where('year',2020).where('month',are.between_or_equal_to(7,8))
calendar=calendar.drop('available','year')
calendar.sample(5)
```

- added a 'year' column to each table to specify 2020, then dropped this column
- added a 'month' column to each table to specify timeline July-August

`calendar`: June 2020

listing_id	date	price	month
24475148	2020-08-27	35	8
2465543	2020-08-24	99	8
33563347	2020-07-17	219	7
8221207	2020-08-15	63	8
31981354	2020-08-25	213	8

`calendar_feb`: Feb 2020

listing_id	date	price	month
21796486	2020-07-13	55	7
32750134	2020-07-23	50	7
1203849	2020-07-30	250	7
32075650	2020-07-22	150	7
1578597	2020-08-24	229	8

Hypothesis Testing & Prediction Questions

Hypothesis Testing Question: How has COVID-19 affected the average price of SF Airbnbs in July and August?

- Null Hypothesis: the price of SF Airbnbs has not changed during COVID-19
- Process: compare average prices of listings in July-August from the June (pre-COVID) and February (during COVID) predicted prices

Prediction Problem: Can an Airbnb's 'price pre-COVID' be used to predict its 'price during COVID'?

- Process: linear regression model to determine linear association

Exploration I:
Table w/ Join & Quantitative Graph

```
# Use this cell to join two datasets
calendar = calendar_feb.join(['listing_id', 'date'], calendar, [
    'listing_id', 'date'])
```

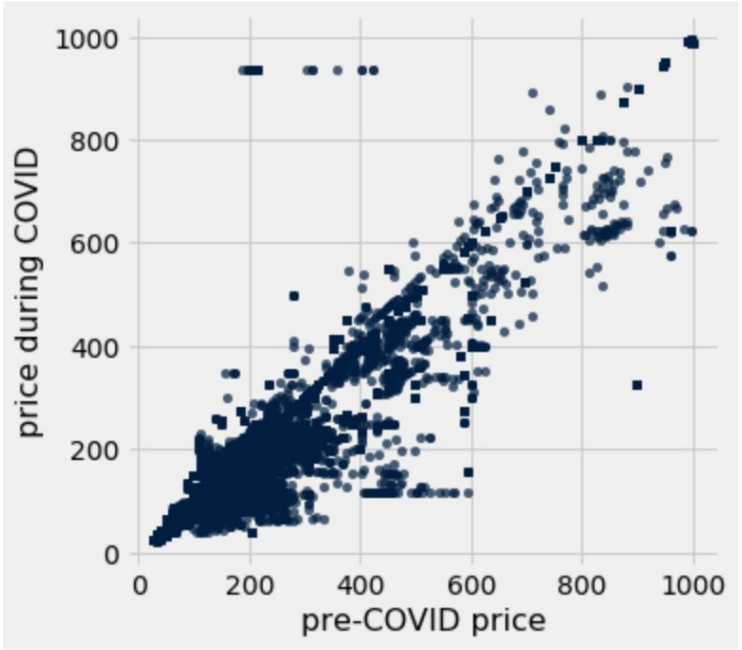
```
calendar = calendar.relabel('price','pre-COVID price').relabel(
    'price_2','price during COVID')
```

```
calendar = calendar.drop('month_2')
calendar
```

listing_id	date	pre-COVID price	month	price during COVID
5858	2020-07-01	325	7	325
5858	2020-07-02	325	7	325
5858	2020-07-03	325	7	325
5858	2020-07-04	325	7	325
5858	2020-07-05	325	7	325
5858	2020-07-06	325	7	325
5858	2020-07-07	325	7	325
5858	2020-07-08	325	7	325
5858	2020-07-09	235	7	235
5858	2020-07-10	235	7	235

... (39069 rows omitted)

```
# Use this cell to generate your quantitative plot
calendar.scatter('pre-COVID price', 'price during COVID')
```



Exploration II: Aggregated Table & Qualitative Graph

Aggregated Data Table:

```
# Use this cell to generate your aggregated data table
cal_avg=calendar.group('date',np.average).drop('listing_id average','month average')
cal_avg
```

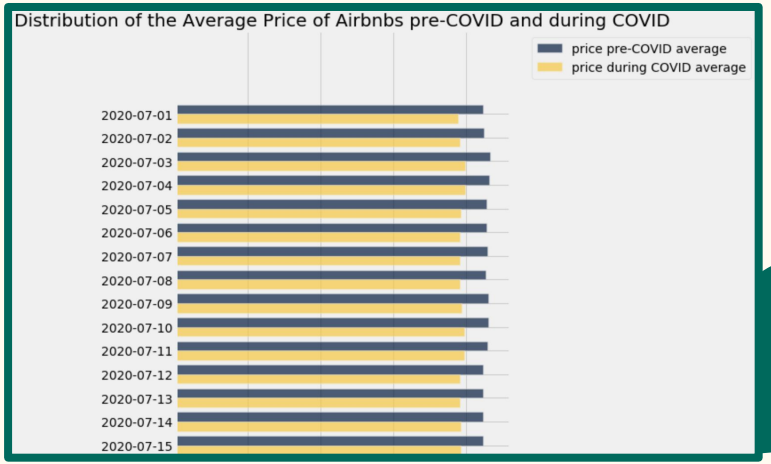
date	price pre-COVID average	price during COVID average
2020-07-01	211.666	194.911
2020-07-02	212.318	195.949
2020-07-03	216.91	199.427
2020-07-04	216.281	199.743
2020-07-05	214.557	196.652
2020-07-06	214.581	196.241
2020-07-07	215.177	196.277
2020-07-08	214.006	196.055
2020-07-09	215.641	197.451
2020-07-10	215.602	199.092

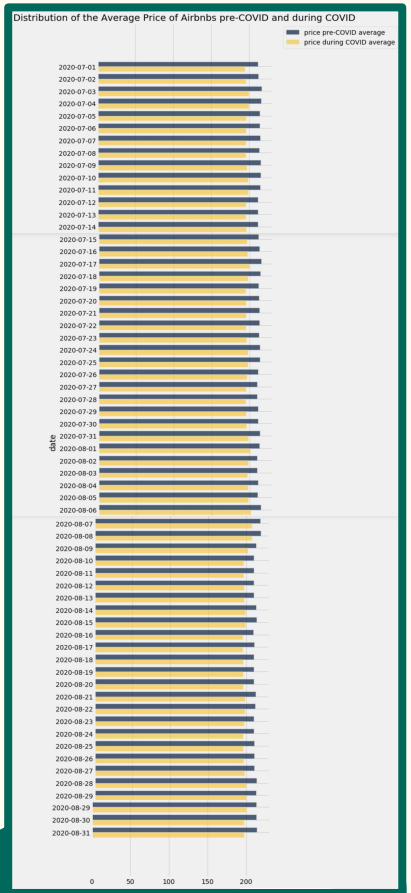
... (52 rows omitted)

Qualitative Plot:

```
# Use this cell to generate your qualitative plot
cal_avg.barh('date')
plots.title('Distribution of the Average Price of Airbnbs pre-COVID and during COVID')
```

```
Text(0.5, 1.0, 'Distribution of the Average Price of Airbnbs pre-COVID and during COVID')
```





Hypothesis Testing: how has COVID-19 affected the average price of SF Airbnbs?

- Null Hypothesis: the price of SF Airbnbs has not changed during COVID-19
- Alternative Hypothesis: the price of SF Airbnbs has decreased in response to COVID-19, and this difference is not due to random chance

Process: A/B Hypothesis Test

- A Group: avg. price during COVID (COVID = True)
- B Group: avg. price pre-COVID (COVID = False)
- Test Statistic: difference in averages of price pre-COVID and price during COVID
- Significance Cutoff: 5%

```
#calendar table w/ False value for COVID - Feb 2020
pre_covid=calendar.with_column('COVID', False).select('price pre-COVID','COVID').relabel('price pre-COVID','price')

#calendar table w/ True value for COVID - June 2020
during_covid=calendar.with_column('COVID', True).select('price during COVID','COVID').relabel('price during COVID',
                                                                                               'price')

#stacking tables
pre_covid_copy=pre_covid.copy()
price_and_covid_change=pre_covid_copy.append(during_covid)
price_and_covid_change
```

Step 1: Stacking Tables

stacked two tables to create '`price_and_covid_change`'

one table had a `COVID` column w/ False values and the `price` column included data collected before the pandemic

the other contained a `COVID` column w/ True values and the `price` column reflected data collected during the pandemic.

`price_and_covid_change`

price	COVID
325	False
325	False
325	False
325	False
325	False
325	False
325	False
325	False
235	False
235	False

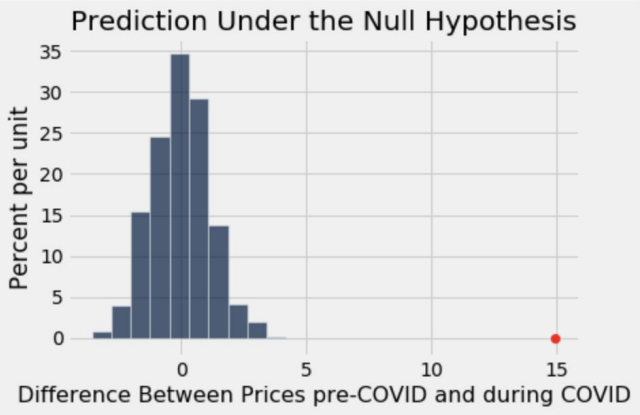
(78148 rows omitted)

A/B Test Results

Histogram Displaying 1000 Simulated Values of our Test Statistic Under the Null. Red Dot = Observed Value

```
Table().with_column('Difference Between Prices pre-COVID and during COVID',differences).hist()
plots.scatter(observed_difference, 0, color='red', s=40)
plots.title('Prediction Under the Null Hypothesis')
print('Observed Difference:', observed_difference)
```

```
Observed Difference: 14.93403106527802
```



p-value = 0

```
p_value=np.count_nonzero(differences>=observed_difference)/repetitions
p_value
```

```
0.0
```

Conclusion

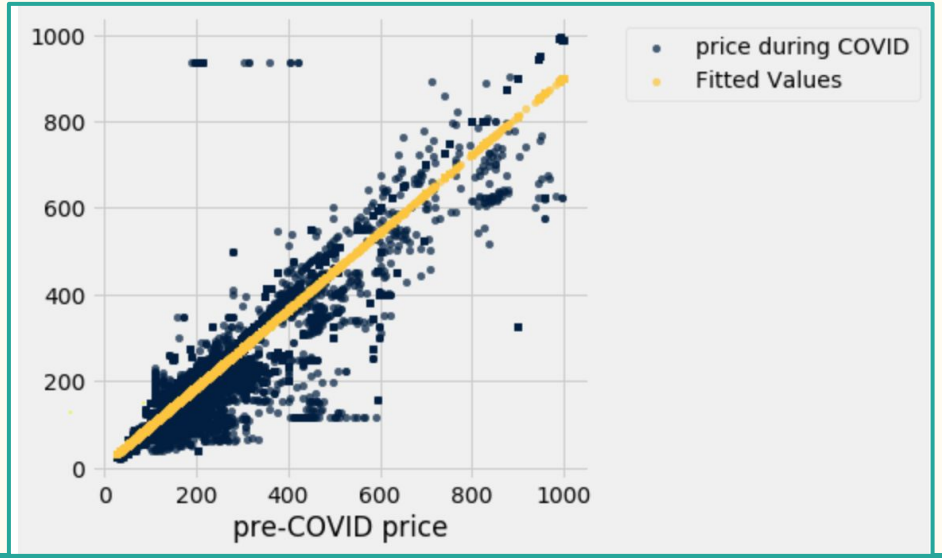
- the observed value (14.93 dollars) and the predicted behavior under the null (centered at 0) are inconsistent
- p-value = 0 (less than our p-cutoff at 5%), so <u>the data is inconsistent with the null hypothesis</u>
  - favors the alternative: avg. price of Airbnbs decreased in response to COVID-19 → difference not due to random chance

Prediction: can a SF Airbnb's pre-COVID price be used to predict its price during COVID?

- Process: Linear Regression Method
  1. Found correlation to be 0.941 = strong positive linear correlation

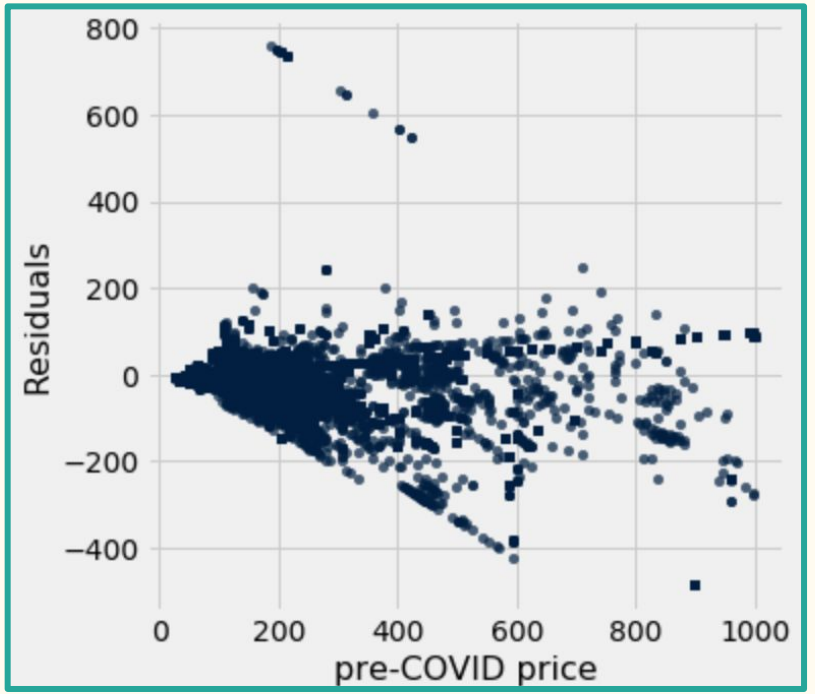
  2. Fit a linear regression line to plot

     Can see that it fits the data well



Prediction Problem: Results

3. Evaluated prediction residuals, found that our model was realistic because no pattern was observed in residual plot.



Conclusion: The linear regression method was a good fit for our data and therefore, pre-COVID price is a good predictor of the price during COVID. They have a strong linear positive association.

Conclusion

From Hypothesis Test:

- With A/B Hypothesis test we found that the average price of Airbnbs in SF decreased during COVID-19.

From Prediction Model:

- Pre-COVID price of an Airbnb in SF is a good predictor of the price during COVID of an Airbnb in SF. These variables have a strong positive linear association.

Limitations:

- Recorded prices were predicted by property owners instead of what they may have actually been listed for. Prices could change when listed from what owners predicted and listed prices would probably be a better representation of the market for Airbnbs.