results_paper

DPREP Team 16

17/10/2021

Results linear regression

To create the linear model, the cleaned dataset is loaded into the R script. Then, a linear model is made by using the price as the dependent variable, the regulation as the independent variable and then the different neighborhoods and room types are added as covariates, as well as the other three dummy's: host_since_dummy, superhost_dummy and identity_verified_dummy. For a more elaborate explanation on these variables, please read the paper in the source code datapreparation folder.

When running this regression, the results are as follows:

```
cleaned_data <- read.csv(file = '../../gen/data-preparation/output/cleaned_data.csv')
price_lm <- lm(price ~ regulation + old_town + etobicoke + north_york + east_york + york +
                 entire_home_apt + hotel_room + private_room +
                 host_since_dummy + superhost_dummy +
                 identity_verified_dummy, cleaned_data)

summary(price_lm)
```

```
##
## Call:
## lm(formula = price ~ regulation + old_town + etobicoke + north_york +
##     east_york + york + entire_home_apt + hotel_room + private_room +
##     host_since_dummy + superhost_dummy + identity_verified_dummy,
##     data = cleaned_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -189.5   -72.5   -35.9     4.5 13032.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                35.258     15.194   2.321 0.020318 *
## regulation                 11.000      3.432   3.205 0.001350 **
## old_town                   45.673      6.678   6.839 8.09e-12 ***
## etobicoke                  20.839      9.284   2.245 0.024789 *
## north_york                 19.185      7.655   2.506 0.012204 *
## east_york                  13.646     13.474   1.013 0.311186
## york                       -8.093     11.542  -0.701 0.483228
## entire_home_apt           114.252     13.897   8.221  < 2e-16 ***
## hotel_room                 19.185     34.009   0.564 0.572675
## private_room               20.087     13.960   1.439 0.150190
```

```
## host_since_dummy          13.273      3.503   3.789 0.000151 ***
## superhost_dummy          -19.878      3.838  -5.179 2.24e-07 ***
## identity_verified_dummy   -7.814      4.296  -1.819 0.068955 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 315.4 on 34633 degrees of freedom
## Multiple R-squared:  0.02733,    Adjusted R-squared:  0.027
## F-statistic: 81.11 on 12 and 34633 DF,  p-value: < 2.2e-16
```

Here, it is directly noticeable that regulation has a significant, positive influence on the price. All covariates, except for the identity_verified_dummy, also have a significant influence on the listing price. From this, it can be concluded that not only regulation has had a significant impact on the price, but these covariates as well. However, these effects can also go hand in hand; when the regulation was made official, the offer for several room types such as entire homes or apartments would significantly decrease, as way less people could offer such accommodation. The scarcity for this room type can also lead to an increase in price.

Note that for example, not all neighborhoods and room_type have a significant influence on the price. However, to draw the conclusion that being in a specific neighborhood and having a specific room type to offer has a significant impact on the price, having one significant dummy is already enough, as this shows that at least one neighborhood or room type is significant for the price.

Data visualization

Data retrieval plots

To obtain the data for all the plots created, several files were created for the specific plots.

General barplot

First, a general barplot will be made, which will just show the increase in price before and after the regulation. This result will be shared in the Github Readme, as it is the most concise conclusion.

The following code is used for the general barplot. The results of this data are written to a csv file to make the data easily accessible in other files.

```
# --- Load Data --- #
all_data <- read.csv(file = '../../gen/data-preparation/output/cleaned_data.csv')

# --- Data for Plot 3 --- #

regulation_price <- all_data %>%
  group_by(regulation) %>%
  summarise(price = mean(price))

write.csv(regulation_price, "../../gen/analysis/input/data_for_barplot.csv", row.names =
            FALSE)
```

Neighborhood plot

Next, also a barplot containing the price difference per neighborhood is included. The structure of this data retrieval code is a bit more complex, as the data has to be pivoted into a long format first. In the end, a data set containing the neighborhoods and the regulation, summarized by the means, has been created. This data set is once again written to a csv file.

```
## `summarise()` has grouped output by 'neighborhood'. You can override using the `.groups` argument.
```

Room type plot

Last, we also include a plot that visualizes the difference in prices before and after the regulation for the room types. The retrieval of the data has been done in a similar way.

```
# --- Load Data --- #
all_data <- read.csv(file = '../../gen/data-preparation/output/cleaned_data.csv')

# --- Data for Plot 1 --- #

roomtype_price <- all_data %>%
  group_by(room_type, regulation) %>%
  summarise(price = mean(price))
```

```
## `summarise()` has grouped output by 'room_type'. You can override using the `.groups` argument.
```

```
dir.create('../../gen/analysis/input/', recursive = T)
```

```
## Warning in dir.create("../../gen/analysis/input/", recursive = T): '..\..
## \gen\analysis\input' already exists
```

```
write.csv(roomtype_price, "../../gen/analysis/input/data_for_roomtype_plot.csv",
          row.names = FALSE)
```

Barplots

Barplot general prices

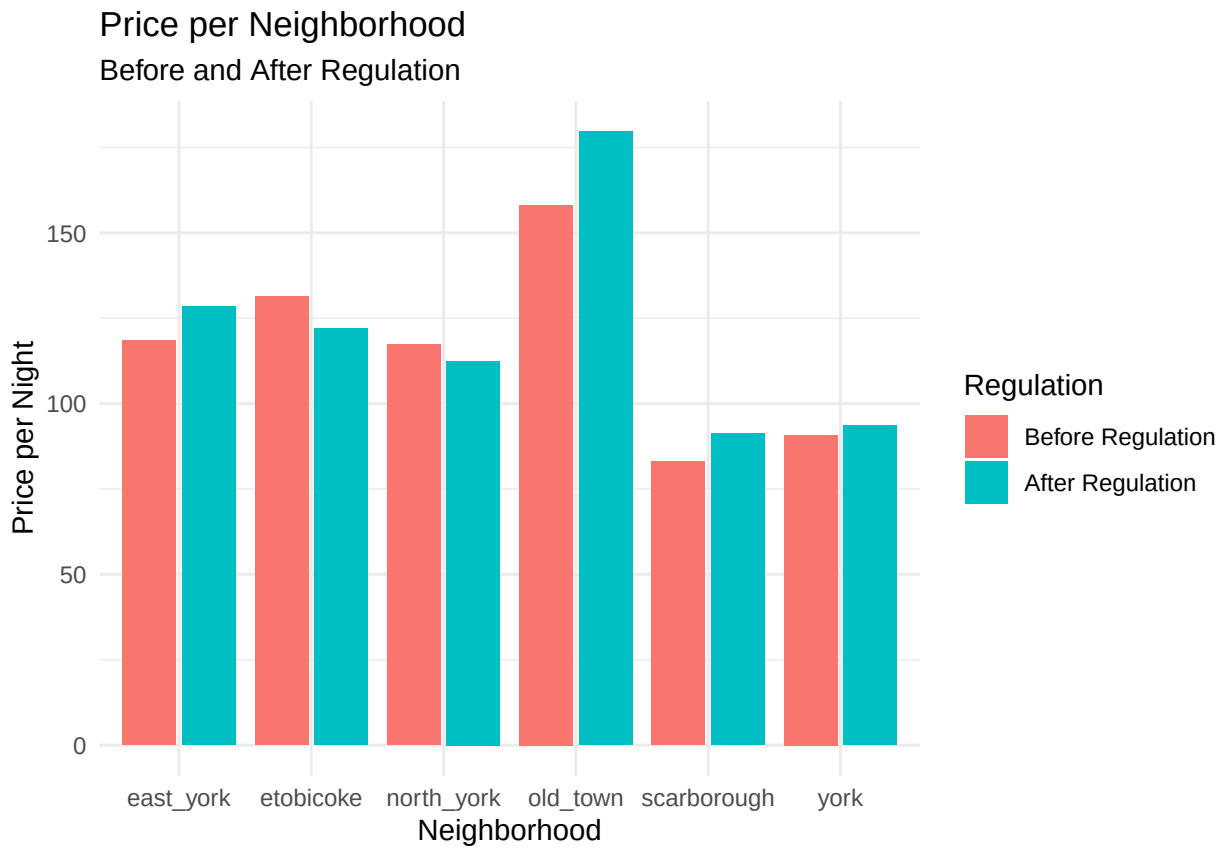
The following plot has been made based on the data retrieved. This plot visually represents the mean price increase before and after the regulation.

As you can see, the average price increased after the regulation was implemented. It is not an enormously large increase, however, in such an overheated market as the one AirBnB is in, it can definitely have a larger impact than it seems. On the next pages, the impact of the covariates will be evaluated.

Barplot neighborhoods

To show the price difference for the different neighborhoods before and after the regulation, a barplot visualization has been chosen. A barplot allows for easy scanning and it shows the differences in price between each neighborhood before and after the regulation clearly.

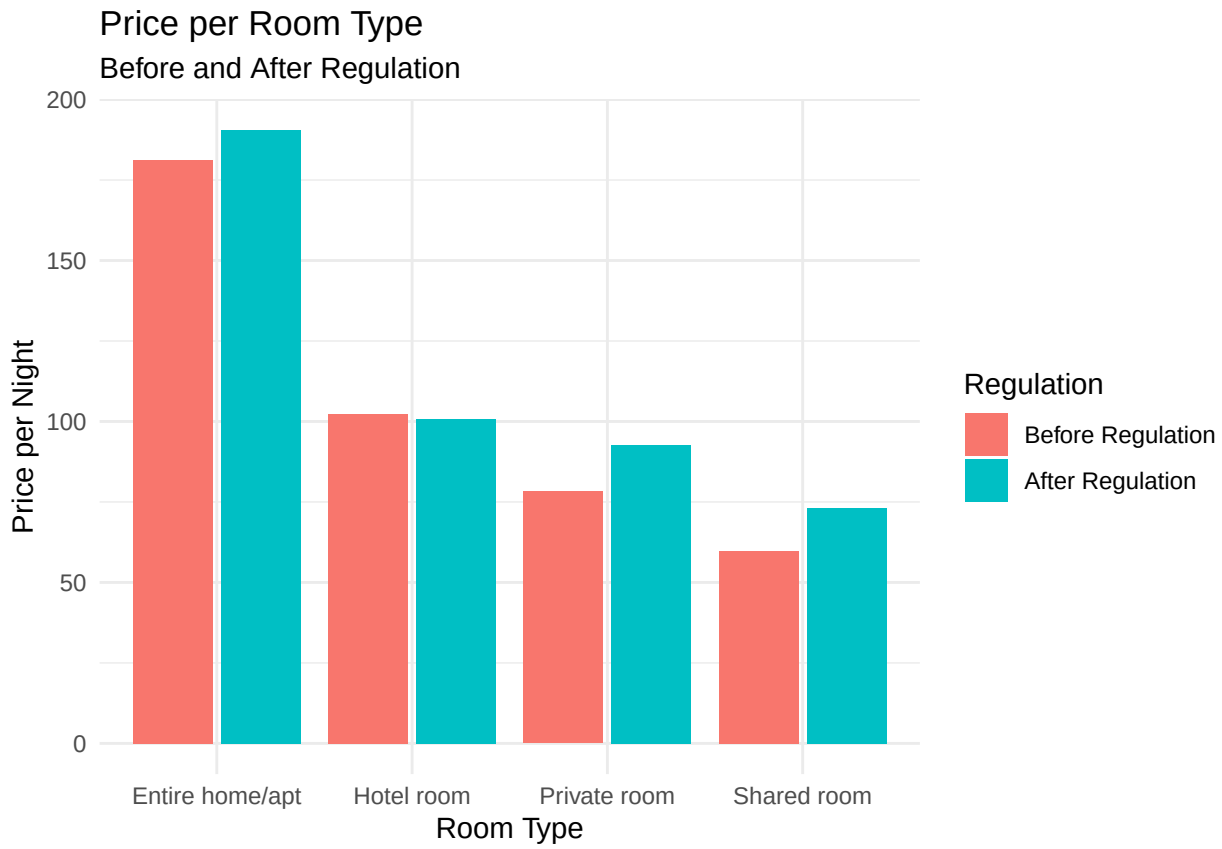
The barplot for the neighborhoods shows as follows:



You can see that the prices in the Old Town of Toronto are affected most by the regulation. Except for Etobicoke and North York prices go up slightly, but not such a large difference as for Old Town. This does make sense, as Old Town is the most important neighborhood for tourists, as the neighborhoods located within the Old Town are in the city center. Therefore, the prices can be more sensitive than for example neighborhoods in the suburbs.

Barplot room types

To visualize the price differences per room type, a barplot is chosen as well for similar reasons as for the neighborhood barplot. The barplot is plotted below.



As you can see, prices in crease after the regulation for every room type, except hotel. That does make sense, because the regulation mainly effects the entire homes or apartments, or the private/shared rooms within someones house. Hotel rooms are not included abundantly in the AirBnB listings, so that is another reason why the prices can be lower after the regulation.

Table dummies

Lastly, a table for the dummy variables has been created. Instead of a barplot, we chose for a table, since we believe that a table can more clearly represent the differences in the price for the host dummy variables. In this table, the identity_verified_dummy has not been included, as this dummy proved no significant impact on the price.

```
# --- Load Data --- #
host_data <- read.csv(file = '../../gen/data-preparation/output/cleaned_data.csv')

# --- Table 1 --- #

# Make dataset
host_price <- host_data %>%
  group_by(host_since_dummy, superhost_dummy, regulation) %>%
  summarise(price = mean(price))
```

```
## `summarise()` has grouped output by 'host_since_dummy', 'superhost_dummy'. You can override using th
```

```
# Remove irrelevant columns
host_price <- host_price[-c(1:2, 7:8),]

gridExtra::grid.table(host_price)
```

	host_since_dummy	superhost_dummy	regulation	price
1	0	1	0	133.552869750456
2	0	1	1	142.313452329125
3	1	0	0	146.823523861368
4	1	0	1	159.419264588329

As shown in the table, prices go up for both the host_since_dummy and superhost_dummy. If calculated in percentages, that is 8.6% and 6.6% for host_since_dummy and superhost_dummy respectively.

Overall conclusion

By means of this linear regression and the barplots, it can be concluded that the regulation definitely has had its effect on the listing price of AirBnB's. The regulation was not the only significant influence for the price increase, however, the effects of the regulation and the other significant impacts on the increase in price can be related to each other, as explained earlier. Therefore, it can be concluded that this regulation definitely played a role in the increase in the prices. However, the linear regression model only explains 2.7% of the total variance. Other researchers can build further on this report to find out what caused the other variance in price. The influence of COVID-19 can be a very interesting angle to look at this data.