Name: Dikshant Joshi

Lab 3 : Causal discovery & Panel data analysis

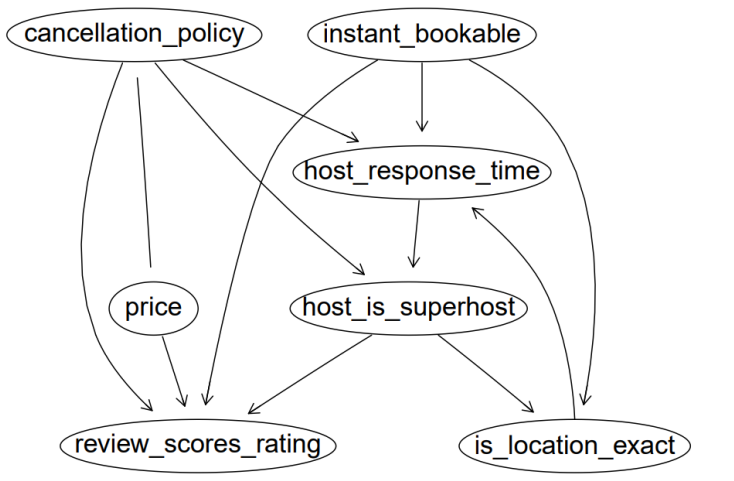
Part A

Question 1: **In the Airbnb data, develop causal discovery graphical models using the PC algorithm and another algorithm of your choice that relaxes the causal sufficiency assumption. As you develop the causal graphs, make a use case for why you are developing the graphs.**

I was interested in understanding the factors that influence review score rating of an Airbnb listing. For the purpose I took following variables mentioned with why I found them important:

Variables Used	Importance
Cancellation policy	➢ This could influence rating, as guests might be more likely to leave positive reviews if they feel they have more flexibility to cancel their booking.
Review score rating	➢ This could influence Price of an Airbnb as listings with higher ratings might be perceived as more desirable and command higher prices.
Price	➢ Price could have directly influence on review score rating as guests might be more likely to leave positive reviews if they feel they received good value and had a comfortable stay.
Instant Bookable(Yes/No), **Is location exact(True/False),** **Host response time,** **Host is superhost(True/False)**	➢ These variables could directly/indirectly influence review score rating as guests might be more likely leave positive reviews if host prior response is good which in turn can influence if host is superhost, Airbnb can be booked instantly and location is exact

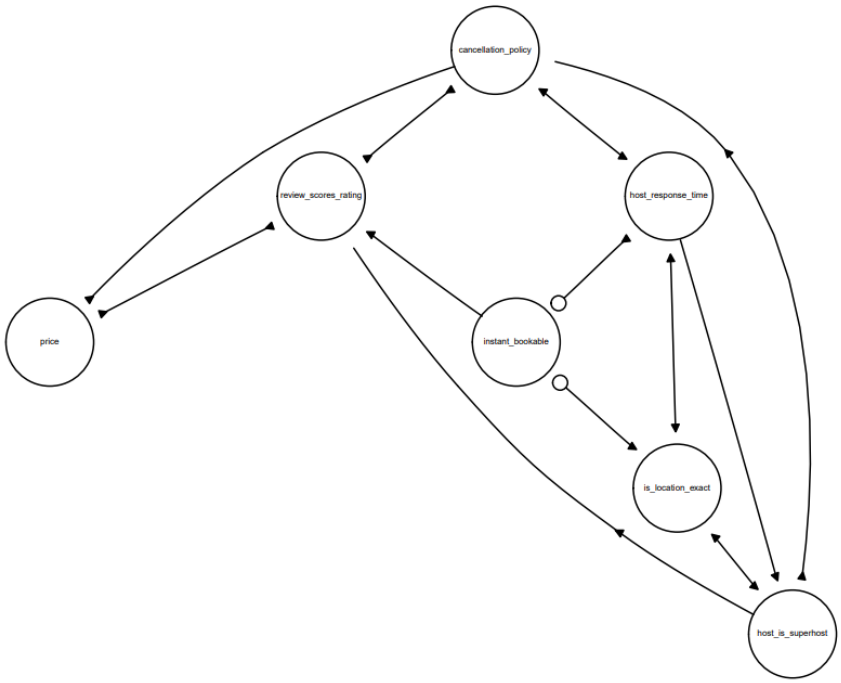
Question 2: **Report the CPDAG the PC algorithm creates. Does it make sense? Why or why not?**



Name: Dikshant Joshi

The CPDAG makes sense because cancellation policy, price directly influence review score rating, host is superhost makes direct influence on review rating. It can be possible that if for an Airbnb price got increased the review rating might come down. We can also see the relationship between host_is_superhost, host_response_time and instant_bookable. If an Airbnb is instant bookable host response time will be quick which could lead to host being a superhost and increase the review score of an Airbnb. Also if the location of Airbnb is exact the host_response_time will be quick. Instant bookable is also influencing review rating directly.

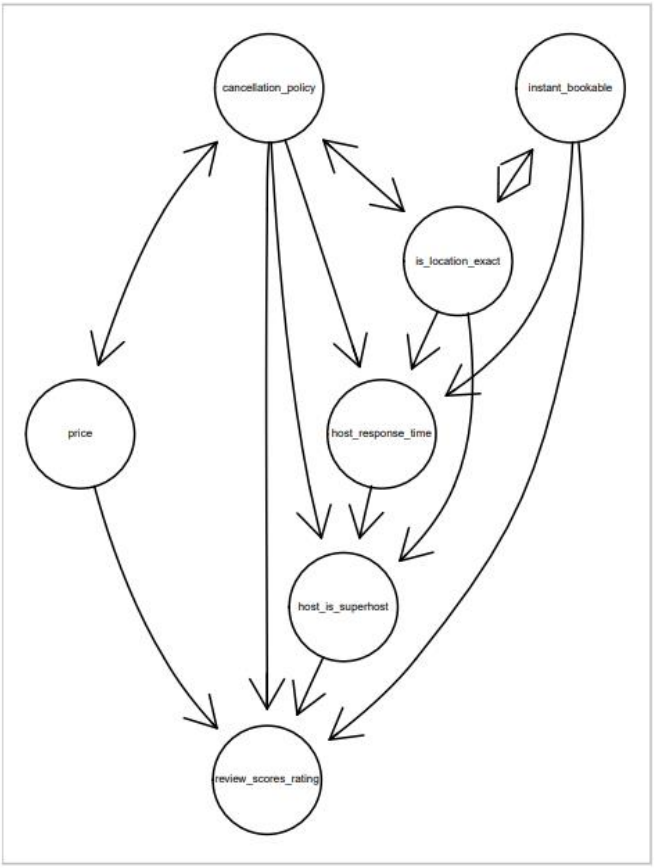
Question 3: **Report the PAG from the algorithm you chose. Compare and explain the differences.**



We can see the difference between PAG and CPDAG . Price and review rating are influencing each other which is the case with cancellation policy and rating as well. For instant bookable there are some unidentified directions. Everything else looks pretty much reasonable. Between host_is_superhost and cancellation policy there can be some other variable influencing.

Name: Dikshant Joshi

Question 4: **In a single paragraph, discuss your findings and list your key takeaways and lessons learned from the analysis. How do you think you can use causal discovery in the future?**



I also use GES to check the CPDAG. This pretty much is the same graph which we got using PAC. My finding is that review score rating is influenced by cancellation policy, price, host_is_superhost, instant_bookable directly. Key takeaway is the graph clearly explains the relationship between instant_bookable, is_location_exact, host_response_time, host_is_superhost. We can use these causal links to drive our decision as to where should Airbnb put efforts so that the ratings of there listings will be high.

Part B

Question 1: **Rebuild the model from Lab II using the minimum required variables* and add price. Provide a numerical interpretation of the effect size of the treatment in plain English.* Except for state and city. Do not add state and city to your model (yet).**

```
# Estimate the model using feols without fixed effect
cf <- feols(high_booking ~  -1 + treatment + price + accommodates + bedrooms
            + minimum_nights + review_scores_rating +
              factor(room_type), data = df1)

cf$coeftable[1,]
```

```
##              Estimate  Std. Error t value     Pr(>|t|)
## treatment 0.05077591 0.003066011 16.5609 1.655389e-61
```

Name: Dikshant Joshi

The Estimate tells us the percentage increase in the high_booking variable which is 5% because of the treatment which is safety measures taken by Airbnb after Halloween night shooting at a "party house" in Orinda, California.

Question 2: **Is the effect size consistent with what the causal forest reported in Lab II?* Why or why not do you think it is? Which one do you think is a more reliable estimate of CATE?**

*** The effect size from the causal forest was "95%CI for the ATE: 0.071 +/- 0.047"**

I think the effect size 0.05+/-0.003 from feols() is consistent with 95%CI for the ATE: 0.071 +/- 0.047 as our effect size lies within the 95% confidence interval which is (0.024,0.118). I think effect size calculated using causal forest is more reliable than the one calculated here using feols as we are not controlling for fixed effect which is why feols() estimate can be biased treating the relationship between all observation same.

Question 3: **Add fixed effects to the model at Airbnb level (using IDs). Has the effect size changed? If so, how do you explain the change in the effect size after adding fixed effects? Now, is this a better estimate than what the causal forest reported earlier? Why or why not?**

```
#Estimating the model using feols with fixed effect
cf2 <- feols(high_booking ~  -1 + treatment + price + accommodates + bedrooms
             + minimum_nights + review_scores_rating +
               factor(room_type)|id, data = df1)
```

```
## The variable 'factor(room_type)Shared room' has been removed because of collinearity (s
```

```
cf2$coeftable[1,]
```

```
##            Estimate  Std. Error  t value      Pr(>|t|)
## treatment 0.0879689 0.002424142 36.28868 4.525355e-286
```

I think the effect size calculated here provides better estimate than the causal forest as now we are controlling for fixed effect (ID) which will help us to capture unobserved differences in the outcome variable that are specific to each group and it will take time invariate variables into account avoiding bias.

Question 4: **In the model with fixed effects, fixest clusters the standard errors by group by default. What does this clustering mean in this domain for this problem (so, don't ask ChatGPT)?**

I think it means that for each group we will get one standard error of the coefficients treating within group observation as correlated and across group observation independent from each other.

Name: Dikshant Joshi

Question 5: **Between the models with and without fixed effects, we observe some dramatic changes in the statistical significance levels (and sometimes magnitudes) of other coefficient estimations such as Accommodates. How do you explain such dramatic differences? Please provide a reasoning**

	Estimate <dbl>	Std. Error <dbl>	t value <dbl>	Pr(>|t|) <chr>	<fctr>
treatment	0.050776	0.003066	16.56090	< 2.2e-16	***
price	-0.000031	0.000013	-2.35302	0.01862360	*
accommodates	-0.030435	0.001023	-29.76314	< 2.2e-16	***
bedrooms	0.035371	0.002455	14.40505	< 2.2e-16	***
minimum_nights	-0.000124	0.000100	-1.24227	0.21413877	
review_scores_rating	0.000199	0.000057	3.47886	0.00050379	***
factor(room_type)Entire home/apt	0.269071	0.006654	40.43462	< 2.2e-16	***
factor(room_type)Hotel room	0.067835	0.025758	2.63352	0.00845204	**
factor(room_type)Private room	0.248516	0.006274	39.60751	< 2.2e-16	***
factor(room_type)Shared room	0.188193	0.011503	16.36054	< 2.2e-16	***

1-10 of 10 rows

	Estimate <dbl>	Std. Error <dbl>	t value <dbl>	Pr(>|t|) <chr>	<fctr>
treatment	0.087969	0.004996	17.606842	< 2.2e-16	***
price	-0.000034	0.000039	-0.872879	0.3827649	
accommodates	-0.008304	0.004006	-2.072790	0.0382357	*
bedrooms	0.004128	0.012883	0.320430	0.7486542	
minimum_nights	0.000500	0.000266	1.880623	0.0600730	.
review_scores_rating	-0.000204	0.000076	-2.682756	0.0073224	**
factor(room_type)Entire home/apt	-0.016501	0.071008	-0.232375	0.8162548	
factor(room_type)Hotel room	-0.109146	0.074219	-1.470595	0.1414547	
factor(room_type)Private room	-0.087650	0.069510	-1.260967	0.2073713	

This can be because previously we were not controlling for ID which in turn was providing us the estimates treating the relationship between all observation same and ignoring the fact that there might be time invariate variables within each group and variables like bedrooms, accommodates remain constant over time and fixating on ID removes that bias. Hence there is dramatic difference between the estimates with & without fixed effects.

Question 6: **Add state and city as fixed effects to your model (in addition to the ID fixed effects). Has the effect size for treatment changed? Why or why not? Please provide an explanation.**

```
#Estimating the model using feols with fixed effect as id+city+state
cf3 <- feols(high_booking ~  -1 + treatment + price + accommodates + bedrooms
             + minimum_nights + review_scores_rating +
               factor(room_type)|id+factor(city)+factor(state), data = df1)

cf3$coeftable[1,]
```

```
##             Estimate  Std. Error  t value     Pr(>|t|)
## treatment 0.08797062 0.002425082 36.27531 7.31653e-286
```

Name: Dikshant Joshi

I think the after adding state and city as fixed effects to our model the effect size did not change dramatically. This could be because by fixating on ID,city & state we are fixating for certain variables that do not change within time,city and state but airbnb's within each city and state could have different variable values and won't be constant across each state & city.

Question 7: **Remove all fixed effects and develop a model explaining price instead of high_booking.* What is the effect of the treatment on price? Do you think this is a good model to make such an inference? Why or why not? Now, add the ID fixed effects. How about now? * Just flip the locations of high_booking and price in the model formulation (DV vs. IV).**

```
#Estimating the model using feols without fixed effect on price
cf4 <- feols(price ~  -1 + treatment + high_booking + accommodates + bedrooms
             + minimum_nights + review_scores_rating +
               factor(room_type), data = df1)

cf4$coeftable[1,]
```

```
##            Estimate Std. Error   t value     Pr(>|t|)
## treatment -4.706111   0.786169 -5.986132 2.157182e-09
```

	Estimate <dbl>	Std. Error <dbl>	t value <dbl>	Pr(>|t|) <chr>	<fctr>
treatment	-4.706111	0.786169	-5.98613	2.1572e-09	***
high_booking	-2.029961	0.862703	-2.35302	1.8624e-02	*
accommodates	13.697199	0.259083	52.86801	< 2.2e-16	***
bedrooms	38.344662	0.616103	62.23740	< 2.2e-16	***
minimum_nights	-0.140983	0.025633	-5.50011	3.8061e-08	***
review_scores_rating	-0.087563	0.014664	-5.97126	2.3632e-09	***
factor(room_type)Entire home/apt	82.463090	1.697147	48.58926	< 2.2e-16	***
factor(room_type)Hotel room	125.314792	6.582600	19.03728	< 2.2e-16	***
factor(room_type)Private room	25.159835	1.618718	15.54307	< 2.2e-16	***
factor(room_type)Shared room	-12.815972	2.949677	-4.34487	1.3951e-05	***

I think this is not a good model because we are not fixating on ID and there are certain variables that are time invariate like bedrooms which are very unlikely to change across time within each ID. And hence not fixating over ID will treat relation between variables across all observation same. For eg: accommodates, bedrooms seems to be of high significance on price which seems inappropriate as price might change on effect of treatment or across time but bedrooms, accommodates are very unlikely to change.

```
#Estimating the model using feols with fixed effect as id on price
cf5 <- feols(price ~  -1 + treatment + high_booking + accommodates + bedrooms
             + minimum_nights + review_scores_rating +
               factor(room_type)|id, data = df1)
```

```
## The variable 'factor(room_type)Shared room' has been removed because of collinearity (se
```

```
cf5$coeftable[1,]
```

```
##            Estimate Std. Error  t value   Pr(>|t|)
## treatment -0.5939765  0.3503789 -1.69524 0.09003361
```

Name: Dikshant Joshi

	Estimate	Std. Error	t value	Pr(>|t|)	
	<dbl>	<dbl>	<dbl>	<chr>	<fctr>
treatment	-0.593976	0.909884	-0.652804	5.1391e-01	
high_booking	-0.694760	0.797364	-0.871321	3.8361e-01	
accommodates	1.085229	1.466429	0.740049	4.5930e-01	
bedrooms	16.962354	8.929980	1.899484	5.7550e-02	.
minimum_nights	-0.169835	0.043046	-3.945409	8.0599e-05	***
review_scores_rating	-0.189814	0.016535	-11.479246	< 2.2e-16	***
factor(room_type)Entire home/apt	46.147108	13.908526	3.317901	9.1252e-04	***
factor(room_type)Hotel room	29.635596	12.961851	2.286371	2.2268e-02	*
factor(room_type)Private room	23.470776	12.482746	1.880257	6.0123e-02	.

I think this is a good model as now we fixated over ID which took into account time invariate variables(bedrooms,accommodates) making them insignificant. Treatment didn't have a significant effect of price. Review score rating being significant make sense because rating changes over time.

Name: Dikshant Joshi

Question 8: **During your analysis until Q7, was there a suspicious coefficient? If so, which one is it? What would be a quick way to check and resolve your suspicion about its estimation?**

The suspicious coefficient for me was bedroom being significant in the model without fixed effect as id. One way is to plot the data and see if there is any significant effect. If we don't group on id we get positive relation between bedroom and price as Fig 1 shows

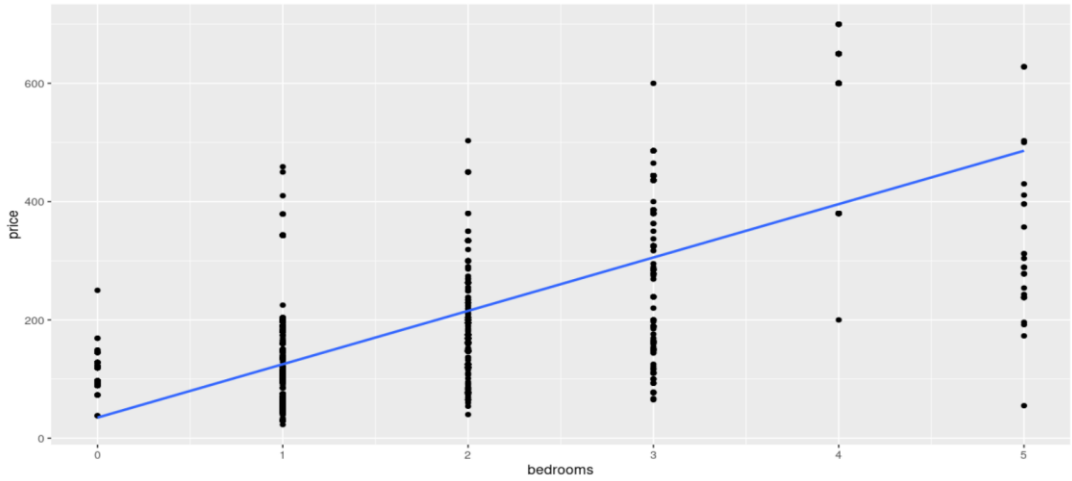


Fig 1

But as we group them with id it shows that there is not strict positive relationship between bedroom and prices. For certain ID'S there is no relation as the prices are increasing even when bedrooms are not changing(Fig2, Fig3). In Fig3 we can clearly see that for some ID'S same number of bedrooms price is changing.

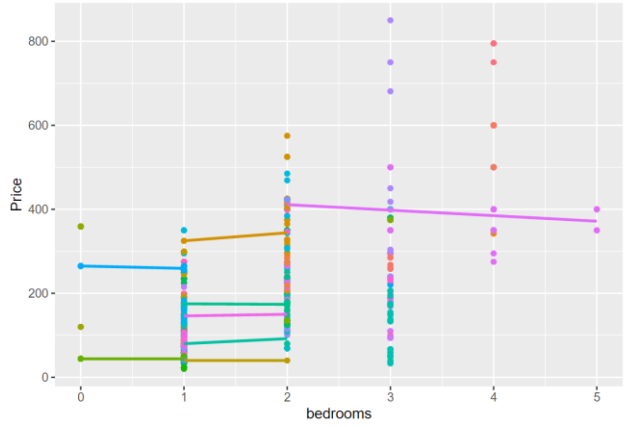


Fig 2

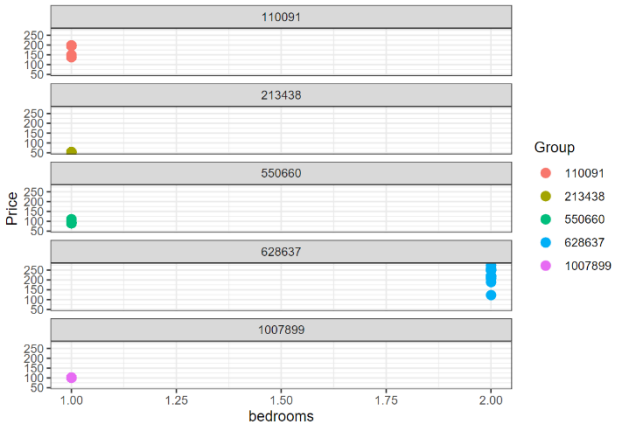


Fig 3