QUESTION 1)

1) **Variables used** :

- **bedrooms**: Integer variable describing number of bedrooms available, might be valuable to know if higher number of bedrooms leads to higher price
- **bathrooms**: Integer variable describing number of bathrooms available, might be valuable to know if higher number of bathrooms leads to higher price
- **accommodates**: Integer variable describing number of guests that can be accommodated in the listing. It could affect the price in a way that it may or may not increase depending on number of guests.
- **room_type**: String variable used to determine how certain types of rooms have different prices. Done by calculation of average log price for a particular type of room. Turned into numerical percentage score called **room_type_score** by calculation
- **neighbourhood**: String variable used to determine how listings in certain neighbourhoods affects price. Done by calculation of average log price for a particular neighbourhood and using that in the model to determine its effect on price.Turned into numerical score called **neighbourhood_score** by calculation
- **guests_included:** Integer variable describing number of more people that can stay in the listing without having to pay more. It is valuable in the sense that listings with greater allowances for number of guests included might have a higher price.
- **host_total_listing_count:** Integer variable of number of place listings the host has currently. This might be valuable in the aspect that a higher number of listings from a single host might be correlated to a lower price.

2) **Handling missing values**:

In order to handle missing values, we first calculated the columns with 50% or more values missing, and these columns can be dropped from the data frame since filtering them out would be difficult through row extraction or imputation. We decided to keep columns with less than 20 percent values missing as such columns can have methods of imputation applied to them or even if we filter out rows with missing values we would have a good chunk of observations remaining for our analysis. In our data, we filtered out rows with missing values in the data frame as we had very few observations missing in the data frame we put in our model.

Question 2)

#### Either split your data into training and validation sets, or just use cross validation below:

See notebook

#### Develop the models. Report all the variables and how do you clean/encode those. While the exact details are visible in the code, explain the broad choices in text:

The variables we inputted into our model were accommodates, bathrooms, bedrooms, guests_included, host_total_listings_count, room_type_scores, and neighborhood_scores. The first four of these variables were selected because we figured they'd be best reflective of a listing's size (the square footage field had too many nulls to use) which we predicted would be strongly correlated with price. We also used guests_included since AirBnB customers often split the cost of AirBnB accommodations between all members of their party staying at the listing, so a higher number of guests would suggest the ability to afford a higher listing price. We used host_total_listings_count because we predicted that hosts with more listings would offer a more professional experience than those with fewer listings, and thus be able to charge higher prices. room_type_score and neighborhood_scores merely converted the categorical data fields room_type and neighbourhood to numerical variables. These variables were calculated using pricing averages then converted to percentages for QA/readability purposes. The values in these fields could have been just as easily derived from census report averages. Further data cleaning included filtering null records, converting the price field from an object type to a float, then getting the log of those cleaned price values.

###Report the final number of observations, the estimated coefficient values, adjusted R 2 , and RMSE on validataion data (or k-fold CV) for three models:

See notebooks

### Interpret the coefficients of the reported models. Again, only interpret the most interesting/important ones, not all of those! Do the coefficient values differ between the models? Can you explain why?:

a)**Part A model**- In this model, we have taken accommodates and neighbour_scores as our independent variables on which we are trying to predict price.
Intercept : **-3.4534**, which means that if we hold both "accommodates" and 'neighbour_scores' at zero, then we would get this value for the log of price.
accommodates coefficient: **0.1496,** indicating that if we keep other variables constant, every unit increase in the number of people that can be accommodated there is an increase of 0.1496 in the log value of price.
neighbour_scores coefficient: 0.0867, indicating that if we keep other variables constant, every unit increase in the neighbourhood_score there is an increase of 0.0867 in the log value of price.

we have an **adjusted r^2 value** of **0.462**, which means after adjusting statistics we have **46.2%** of data that can be explained by this model

we have **6029** observations in total, that's the number we get after cleaning the data.

take **alpha = 0.05**, the p-value (P>|t|) tells us the effects of every variable included is considered as significant.

b) **Part B model** - In this model, we have taken all variables in our final dataframe apart from price and id as our independent variables on which we are trying to predict price.

Intercept : **-4.2002**, which means that if we hold all the values of all variables at zero, then we would get this value for the log of price.

bathrooms coefficient: **0.0495,** indicating that if we keep other variables constant, every unit increase in the number of bathrooms there is an increase of 0.0495 in the log value of price.

neighbour_scores coefficient: **0.0635,** indicating that if we keep other variables constant, every unit increase in the neighbourhood_score there is an increase of 0.0635 in the log value of price.

bedrooms coefficient: **0.1677,** indicating that if we keep other variables constant, every unit increase in the number of bedrooms there is an increase of 0.1677 in the log value of price.

we have an **adjusted r^2 value** of **0.592,** which means after adjusting statistics
we have **59.2%** of data that can be explained by this model

we have **6029** observations in total, that's the number we get after cleaning the data.

take **alpha = 0.05**, the p-value (P>|t|) tells us the effects of every variable included is considered as significant.

c) **Part C model** - In the model, we have variables that we consider might have some effects on "price",

Intercept value shows us when all variables remain zero, the resulting price would be **-3.8361**

the variables coefficient values tell us: when all other variables keep same value:
rising "**accommodates**" by one, resulting log value of price will rise by **0.1000**
rising **bathrooms** number by one, resulting log value of price will rise by **-0.0234**
rising **bedrooms** number by one, resulting log value of price will rise by **0.1516**
rising **neighbour_scores** by one, resulting log value of price will rise by **0.0910**

we have an **adjusted r^2 value** of 0.478, which means after adjusting statistics
we have **47.8%** of data that can be explained by this model

we have **6029** observations in total, that's the number we get after cleaning the data.

take **alpha = 0.05**, the p-value (P>|t|) tells us the effects of every variable included except “bathrooms” is considered as significant.

The coefficients of the variables do not remain the same throughout all the models because we are changing the number of variables used that changes the complexity of model and variables themselves have different amount of effect on the price variable, we use different methods to clean the data and quantify the different categorical data which affects the RMSE as well.

Question 3):

#### Does our model does a good job at predicting price?

Our model does a good job at predicting price. Our model had an RMSE of ~0.43. To improve accuracy, we used the log of listing prices rather than the raw price values for our model. So, a RMSE of ~0.43 for the log of listing prices means that on average, our model's prediction is only about 1.50 dollars off from the actual raw price values. We consider this score impressive given that our price actuals range from 10 to 5,400 dollars and would feel confident using our model to predict price for additional AirBnB listing datasets.

#### How will our model be useful to (a) AirBnB hosts and (b) AirBnB customers?

AirBnB hosts can use our model for determining how to price their own listing(s). A host's goal is to maximize the revenue their listing(s) bring in- offering competitive pricing relative to comparable listings in their area is conducive to this goal. If a host prices their listing at an arbitrary rate, they risk making it too expensive, causing them to consequently lose out on business to competitors. Conversely, if an owner were to underprice their listing, they may get a ton of business, but their total revenue would still be less than it could've been had they opted for a slightly higher yet still competitive price. AirBnB customers can use our model when planning travel accommodations to determine if an owner if offering them a good deal or not. They can simply plug in the fields of the place they're looking at (room type, neighborhood, number of bedrooms, etc.) and see if the predicted price is around what the owner is offering. If the predicted price is significantly lower than what the owner is offering, they'd know they should probably pass on the listing in favor of waiting for other options to go up or searching for listings that fit different parameters.

#### Did we include any other price-related variables, such as "weekly price" or "security deposit" in our model?

We did not include any pricing variables beyond daily rate because variables such as weekly price would be misleading. One reason weekly price is a poor variable is because stay lengths beyond 7 days are rare for AirBnB users- it would be unwise to derive insights such limited data. Beyond data availability constraints, it simply doesn't make sense to include price-related variables in a model like ours since they are really responding variables themselves. for nearly all cases, weekly price will be tied to daily price, usually slightly cheaper per day than the daily price. As daily price increases, weekly price should increase at roughly the same rate. The same goes for security deposits too—albeit with some rate discrepancies—but the concept is no different; as price increases so will security deposit price and vice versa.

#### Did we think our model can be used by AirBnB itself or the government?

We could see AirBnB benefiting from our model if they wanted to create a "pricing assist" service for AirBnB owners to use to better price their listings, but we reckon that the government could better leverage our model's insights. They could simply use it for reporting purposes, perhaps for rent/property value information when putting together census reports, or they could go a step further and use it to influence taxation policies. For example, say the government noticed that the hotel industry in a particular city was suffering greatly and wanted to investigate why. They could plug hotel listing data (hotel room type, hotel location, number of bedrooms, etc.) into our model with hotel fields replacing the proper AirBnB fields, and if the predicted price values were significantly less than what the hotels were currently offering, they'd know AirBnB was the culprit for the hotel industry’s troubles. The government could take responsive action and implement an AirBnB tax for the city, forcing AirBnB owners to hike their prices in order to cover the tax. With increased AirBnB prices, hotels would no longer be as relatively expensive as before and likely experience a rekindling in business.

#### Do we see any ethical issues with this work?

The largest ethical issues we see with this work pertain to the economics of the travel accommodation/rental property industry. Too many individuals possessing the knowledge our model provides regarding listing valuations could potentially trigger a "race to the bottom" in certain markets. Consider the example: If you were an AirBnB owner and discovered (after plugging your listing's information into our model) that your listing was overpriced relative to what should be expected for your area, how would you respond? You'd likely reduce your pricing, perhaps even barely below the expected pricing our model returned in order to make yourself more competitive. However what would happen if every other AirBnB owner in your city followed suit and did the exact same thing? The other owners would likely cut their prices too- causing significant drops in pricing averages across the board. If you then repeated your original process, and plugged your listing into our model again, your new expected price result would probably be even lower than your already reduced price. Should you cut your price an additional time? Maybe, but this kind of behavior is bound to spiral out of control until all owners have priced their listings so cheaply that they can't even cover their base ownership expenses, at which point the market will have failed.

Question 4):

###load the testing data arbnb-seattle-listings-test.csv. This has exactly the same structure and variables as the original dataset:

See notebook

###compute RMSE on the testing dataset. This is the ultimate goodness measure of your model. Present it prominently in your report:

Rmse we got for this model is 0.43666717142791706