1. Who collected the data? [25 words; 1 mark; *Consider the source of the data and its relation to the underlying data generating process.*]

The data given was collected and arranged by Murray Cox as an independent researcher, and was published in the InsideAirBnb website. 1

1. Why did they collect it? [50 words; 2 marks; *Consider the purposes for which the data was collected and how this might shape its form/content.*]

After Airbnb published a report to indicate its business in New York City on December 1, 2015, Cox noticed anomaly in the data. Therefore, he collected data from Airbnb website to illustrate that the released data from Airbnb misled the public in his report focusing on highlighting illegal renting 2

1. How was it collected? [75 words; 3 marks; *What was the method by which the data was collected and how might this shape its form/content.*]

It is basically scaped data which was extracted from Airbnb website with the help of open source technologies. As a first step, by using search and filtering tool in Airbnb website, data from specific location was listed as possible. This process was repeated many times in order to minimize any missing lists. Then, because each list contains more information in its own page, prepared program visited each one to get more information.2

1. What useful information does it contain? [100 words; 7 marks; *A general description of key fields and records identified by the student that might support subsequent analysis.*]

There are variety of information in the data. To support the initial idea why data was collected, probably the essential and useful ones are host id and room type. Host id is unique number given by Airbnb to the host and room type is an indicator in between three categories: “Entire Home/apt”, “Private Room” and “Shared Room”. In addition, there are latitude and longitude variables that point where the property located with a 150 metres accuracy1. Also, there are several indicators to tell the information about property such as description, number of bathrooms and beds, and price per night.

1. What useful information is it missing? [50 words; 7 marks; *A general description of key fields and records that a naive, but technically skilled, user might think that the data included but which it does not.*]

It is not possible to detect “purpose features”. In other words, we couldn’t get that each property is designed for touristic purpose or daily need purpose and suitability for long-term or short-term stay. Also, if there are changings in data, the reason such as complaining and outdated may not drawn

1. To what extent is the data ‘complete’? [150 words; 20 marks; *Reflecting on #4 and #5, to what extent is this data a ‘complete’ picture of the process that it claims to allow us to examine?*]

This data was simply pulled from Airbnb website without any intervention and alteration except compile process. Therefore, there is a clear distinction between looking one specific property listing in Airbnb and looking row in data which indicates this specific property. In other words, this data may not completely illustrate listings in Airbnb website. Because it is a simple extraction of data, one can easily get all information presented in data at first glance while looking each page in Airbnb website. However, as mentioned above, finding “purpose features” can be only with the help of human interpretation or advanced artificial intelligence. Also, deleted listings and its reasons, and listings in different times of year are other structural missing points that prevent it from being completed. In short, this raw data is just like the list of ingredients placed on the kitchen table for cooking.

1. What kinds of analysis would this support? [150 words; 15 marks; *Given the limitations and opportunities identified above, what kinds of analysis would this data support?*]

Although it contains missing information, good analysis from this data could be done. Firstly, as Cox did**2**, it is possible to detect illegal rents, which is the purpose of collection of data, limited to the month of August. Secondly, the relationship between review score and the number of sales can be set up, as Zervas, Proserpio and Byers did in some port of their research**3.** Thirdly, property features given in dataset maybe used to determine their effect on price per night. Fourthly, it is possible to detect possible attraction point by using latitude and longitude variables in data**4**. There can be also comparing analysis by adding another data. For example, the question of is home sharing drives up rents can be answered with the help of rent dataset**5** in August.

1. What kinds of analysis would it \*not\* support? [100 words; 15 marks; *Given the limitations and opportunities identified above, what kinds of analysis cannot be undertaken that a naive, but technically skilled, user might think were possible?*]

This data may not support a detailed analysis of factors that attract people to rent specific property from specific host. Gender, racial and visual qualities of host maybe an attractive point and even there is an image link in data, it is necessary to classify these qualities. This situation is also same for house qualities. In addition, because we do not know how long each guest stay, a detailed analysis of purpose features and its effect on other variables may not be supported.

1. Which of the uses listed in #7 are ethical? [300 words; 30 marks; *Justify the ethics of these analyses with reference to examples drawn from both #7 and #8 and to the literature.*

These analyses can ethically depend on two critical point: The protection of the rights of people that may be affected by analysis, and sustain of public trust. As in first point, we need to determine whether this analysis may harm individuals in listings because of privacy issues. In given data, it is obvious that there are some private points such as sharing property pictures, host pictures, locations etc. We need to be sure whether each host is willing to publicly share its sensitive data. Even there are several consent forms, hosts may not realize who can collect data, where it is stored and how the sensitive data may be used. Due to the priority of economy, private life may have been thrown into the background. Secondly, doing research appropriately is ethically significant. I mean that the input and output variables at the end of each analysis should be double-checked and criticized because it may mislead public. As an input, we must know that each listing in given data really exists in real-life. It is possible that due to the several factors this data maybe incorrectly pulled from Airbnb website which results in misleading information. Also, even the data is almost perfectly pulled from the website, there may be lists that are not suitable for the purpose of Airbnb. For example, hosts, who does not want to share the house, may post it on the site for advertising, spoiling and other purposes. As a result, in order to say that kinds of analysis mentioned above is ethical, privacy and public trust issues are key elements to look at. Consent of users, accuracy of methods in both getting data process and analysis process, user integrity are the ethical concepts that the researcher should talk about in the analysis.