Statistical Programming for Data Science: An investigation on the Airbnb price per night in Amsterdam

XXX

23 May 2024

Contents

1 Dataset

1.1 Introduction

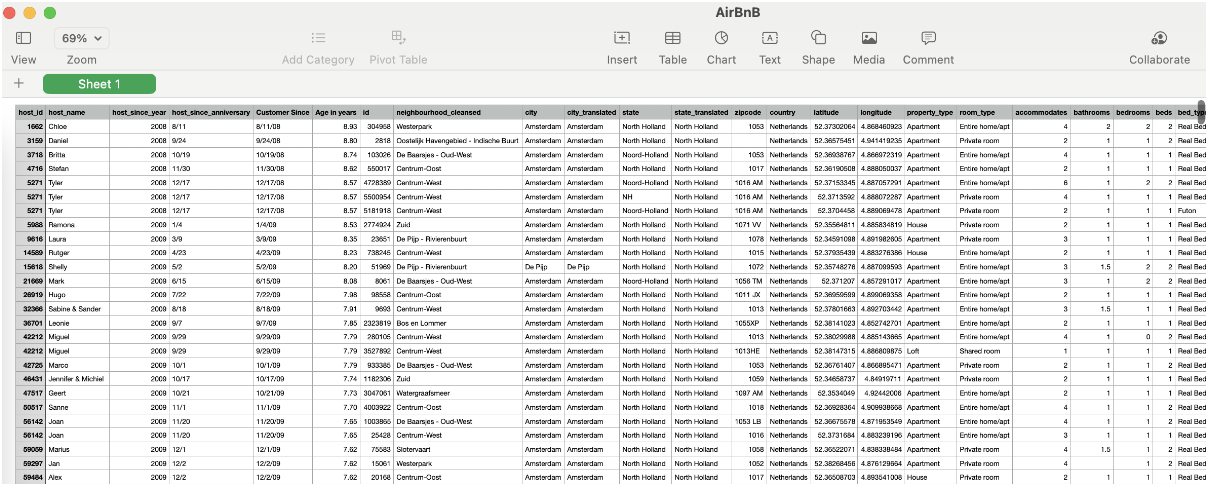
As a sharing economy product, Airbnb experiences a rapid development in recent decades. It is an alternative hotel business that shares the accommodation with others, short period or long period. There are lots of research on the price per night that Airbnb costs.In Chen Yong and Xie [2017], a wide array of utility-bearing attributes of Airbnb listings and the effects of these attributes on consumers' valuation in United States are measured. It provides a comprehensive study on the pricing of Airbnb listed properties and the results explain how the factors, i.e., listing functionality, attributes of hosts, customers reviews and market conditions affect the price. Another research in Cai et al. [2019] focuses on the market of Hong Kong. Five groups' variables were collected, i.e., listing attributes, host attributes, rental policies, listing reputation, and listing location to investigate the determinant of Airbnb price. Some use ordinary least square regression with geographically-weighted. which is introduced in Voltes-Dorta and Sánchez-Medina [2020], to study the factors that affect the price for different room types, i.e., entire room or private room.

In this project, we are going to build a suitable model that can explain the relationship between rental price per night of apartment in Netherlands, mainly in city Amsterdam, posted in Airbnb and several characteristics related to the apartment. Specially, to find the determinants of the price from the room features, e.g., number of bathrooms, bedrooms; host response rate, and the ratings received from the customers.

1.2 Description of dataset

The dataset to be analyzed is collected from https://data.world/cannata/gaairbnb and is named "AirBNB.csv". In the raw dataset, there are 7833 observations on 41 variables. The selected variables to be analyzed are price,accommodates,bathrooms,bedrooms,room_type,host_response_rate, review_scores_rating. The description and type of each variable are listed as follows. - price: continuous variable, the price per night posted on website.
- accommodates: discrete variable, the number of guests that the property can accept.
- bathrooms: continuous variable, the number of bathrooms the property has.
- bedrooms: discrete variable, the number of bedrooms the property has.
- room_type: nominal variable, the feature of the shared property, and there are three types, "Entire home/apt", "Private room" and "Shared room". - host_response_rate: continuous variable, indicating the response frequency of the host when receiving message.
- review_scores_rating: discrete variable, indicating the reputation of the shared property.

The screenshot of dataset is displayed in Figure 1.
As for the data cleaning, we propose to filter out the observations related to property type "apartment" first, then select the necessary variables. At last removing the observations with missing values and changing the format or type of the variables.



host_id	host_name	host_since_year	host_since_anniversary	Customer Since	Age in years	id	neighbourhood_cleansed	city	city_translated	state	state_translated	zipcode	country	latitude	longitude	property_type	room_type	accommodates	bathrooms	bedrooms	beds	bed_typ
1662	Chloe	2008	8/11	8/11/08	8.93	304958	Westerpark	Amsterdam	Amsterdam	North Holland	North Holland	1053	Netherlands	52.37302064	4.868460923	Apartment	Entire home/apt	4	2	2	2	Real Bed
3159	Daniel	2008	9/24	9/24/08	8.80	2818	Oostelijk Havengebied - Indische Buurt	Amsterdam	Amsterdam	North Holland	North Holland		Netherlands	52.36575451	4.941419235	Apartment	Private room	2	1	1	2	Real Bed
3718	Britta	2008	10/19	10/19/08	8.74	103026	De Baarsjes - Oud-West	Amsterdam	Amsterdam	Noord-Holland	North Holland	1053	Netherlands	52.36938767	4.866972319	Apartment	Entire home/apt	4	1	1	1	Real Bed
4716	Stefan	2008	11/30	11/30/08	8.62	550017	Centrum-Oost	Amsterdam	Amsterdam	North Holland	North Holland	1017	Netherlands	52.36190508	4.888050037	Apartment	Entire home/apt	2	1	1	1	Real Bed
5271	Tyler	2008	12/17	12/17/08	8.57	4728389	Centrum-West	Amsterdam	Amsterdam	Noord-Holland	North Holland	1016 AM	Netherlands	52.37153345	4.887057291	Apartment	Entire home/apt	6	1	2	2	Real Bed
5271	Tyler	2008	12/17	12/17/08	8.57	5500954	Centrum-West	Amsterdam	Amsterdam	NH	North Holland	1016 AM	Netherlands	52.3713592	4.888072287	Apartment	Private room	4	1	1	1	Real Bed
5271	Tyler	2008	12/17	12/17/08	8.57	5181918	Centrum-West	Amsterdam	Amsterdam	Noord-Holland	North Holland	1016 AM	Netherlands	52.3704458	4.889069478	Apartment	Private room	2	1	1	1	Futon
5988	Ramona	2009	1/4	1/4/09	8.53	2774924	Zuid	Amsterdam	Amsterdam	North Holland	North Holland	1071 VV	Netherlands	52.35564811	4.885834819	House	Private room	2	1	1	1	Real Bed
9616	Laura	2009	3/9	3/9/09	8.35	23651	De Pijp - Rivierenbuurt	Amsterdam	Amsterdam	North Holland	North Holland	1078	Netherlands	52.34591098	4.891982605	Apartment	Private room	3	1	1	1	Real Bed
14589	Rutger	2009	4/23	4/23/09	8.23	738245	Centrum-West	Amsterdam	Amsterdam	North Holland	North Holland	1015	Netherlands	52.37935439	4.883276386	House	Entire home/apt	2	1	1	1	Real Bed
15618	Shelly	2009	5/2	5/2/09	8.20	51969	De Pijp - Rivierenbuurt	De Pijp	De Pijp	North Holland	North Holland	1072	Netherlands	52.35748276	4.887099593	Apartment	Entire home/apt	3	1.5	2	2	Real Bed
21669	Mark	2009	6/15	6/15/09	8.08	8061	De Baarsjes - Oud-West	Amsterdam	Amsterdam	Noord-Holland	North Holland	1056 TM	Netherlands	52.371207	4.857291017	Apartment	Entire home/apt	3	1	2	2	Real Bed
26919	Hugo	2009	7/22	7/22/09	7.98	98558	Centrum-Oost	Amsterdam	Amsterdam	North Holland	North Holland	1011 JX	Netherlands	52.36959599	4.899069358	Apartment	Entire home/apt	2	1	1	1	Real Bed
32366	Sabine & Sander	2009	8/18	8/18/09	7.91	9693	Centrum-West	Amsterdam	Amsterdam	North Holland	North Holland	1013	Netherlands	52.37801663	4.892703442	Apartment	Entire home/apt	3	1.5	1	1	Real Bed
36701	Leonie	2009	9/7	9/7/09	7.85	2323819	Bos en Lommer	Amsterdam	Amsterdam	North Holland	North Holland	1055XP	Netherlands	52.38141023	4.852742701	Apartment	Entire home/apt	2	1	1	1	Real Bed
42212	Miguel	2009	9/29	9/29/09	7.79	280105	Centrum-West	Amsterdam	Amsterdam	North Holland	North Holland	1013	Netherlands	52.38029988	4.885143665	Apartment	Entire home/apt	4	1	0	2	Real Bed
42212	Miguel	2009	9/29	9/29/09	7.79	3527892	Centrum-West	Amsterdam	Amsterdam	North Holland	North Holland	1013HE	Netherlands	52.38147315	4.886809875	Loft	Shared room	1	1	1	1	Real Bed
42725	Marco	2009	10/1	10/1/09	7.79	933385	De Baarsjes - Oud-West	Amsterdam	Amsterdam	North Holland	North Holland	1053	Netherlands	52.36761407	4.866895471	Apartment	Private room	2	1	1	2	Real Bed
46431	Jennifer & Michiel	2009	10/17	10/17/09	7.74	1182306	Zuid	Amsterdam	Amsterdam	North Holland	North Holland	1059	Netherlands	52.34658737	4.84919711	Apartment	Private room	2	1	1	1	Real Bed
47517	Geert	2009	10/21	10/21/09	7.73	3047061	Watergraafsmeer	Amsterdam	Amsterdam	North Holland	North Holland	1097 AM	Netherlands	52.3534049	4.92442006	Apartment	Entire home/apt	2	1	1	1	Real Bed
50517	Sanne	2009	11/1	11/1/09	7.70	4003922	Centrum-Oost	Amsterdam	Amsterdam	North Holland	North Holland	1018	Netherlands	52.36928364	4.909938668	Apartment	Entire home/apt	4	1	1	2	Real Bed
56142	Joan	2009	11/20	11/20/09	7.65	1003865	De Baarsjes - Oud-West	Amsterdam	Amsterdam	North Holland	North Holland	1053 LB	Netherlands	52.36675578	4.871953549	Apartment	Entire home/apt	4	1	1	2	Real Bed
56142	Joan	2009	11/20	11/20/09	7.65	25428	Centrum-West	Amsterdam	Amsterdam	North Holland	North Holland	1016	Netherlands	52.3731684	4.883239196	Apartment	Entire home/apt	3	1	1	1	Real Bed
59059	Marius	2009	12/1	12/1/09	7.62	75583	Slotervaart	Amsterdam	Amsterdam	North Holland	North Holland	1058	Netherlands	52.36522071	4.838338484	Apartment	Private room	4	1.5	1	2	Real Bed
59297	Jan	2009	12/2	12/2/09	7.62	15061	Westerpark	Amsterdam	Amsterdam	North Holland	North Holland	1052	Netherlands	52.38268456	4.876129664	Apartment	Private room	4		1	2	Real Bed
59484	Alex	2009	12/2	12/2/09	7.62	20168	Centrum-Oost	Amsterdam	Amsterdam	North Holland	North Holland	1017	Netherlands	52.36508703	4.893541008	House	Private room	2	1	1	1	Real Bed

Figure 1: Screenshot of the dataset

1.3 Three proposed research questions

1.3.1 Q1

The first proposed question: "Are the average prices per night the same for different room type?"

1.3.2 Q2

The second proposed question: "Is the price per night related to acommodates and how is the effect?

1.3.3 Q3

The third proposed question: "What are the other varibales having impact on the price per night of the apartment?

2 Data Import and Cleaning

```r
# import dataset and filter out apartment
tb<-read.csv("AirBnb.csv") %>%
  filter(property_type=="Apartment")

# select the necessary variables
tb.selected<-tb %>%
  dplyr::select(price,accommodates,bathrooms,bedrooms,room_type,
                host_response_rate,review_scores_rating)

# change the type of some variables
tb.selected$price<-parse_number(tb.selected$price)
tb.selected$host_response_rate<-as.numeric(tb.selected$host_response_rate)

# remove the observations having missing values
tb.clean<-tb.selected %>% na.omit()
```

3 Data Analysis/Report

3.1 Q1

The objective is to analyze whether the Airbnb posted price per night of apartment are different among different room types. Since the room type is a categorical variable, an one-way ANOVA approach is suitable. Before conducting any statistical analysis, a descriptive summary for the price is tabulated in Table 1. It is found that the average price (128.63) for the Entire home or entire apartment is much higher than that for private room (68.76) and shred room (55.96). Meanwhile the variability of the price for entire home is also the highest. The boxplot displayed in Figure 2 gives a direct comparison of the distribution of price for each room type.
The ANOVA analysis yields the p-value is below 0.05. And it is concluded that the average price are significantly differnet among different room types.

Table 1: Summary statistics for price per night for different room types

room_type	average	SD
Entire home/apt	128.63	60.62
Private room	68.76	28.86
Shared room	55.96	28.34

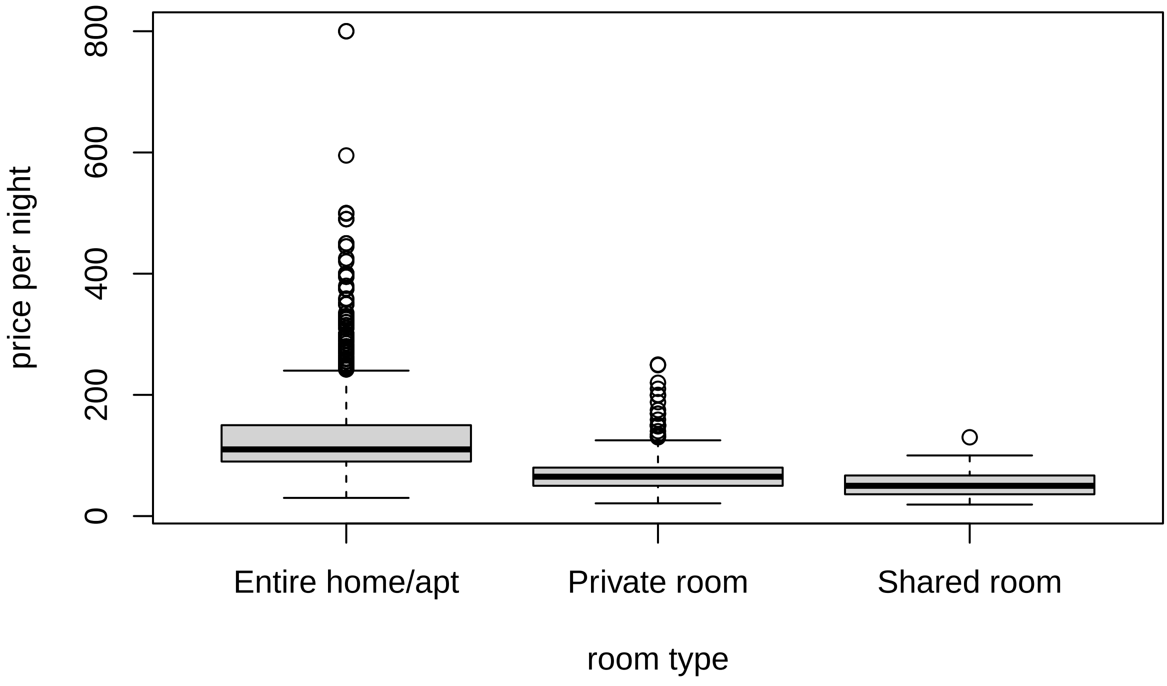


Figure 2: Boxplot for price per night

3.2 Q2

The object is to find the relationship between price per night and the number of guests that the shared property can hold. Figure 3 displays the scatter plot between the two variables. It is noted that there is a increasing trend for the price when accommodates value increases. And it seems the relationship is linear. Simple linear regression is a model that describes the relationship between one dependent and one independent variable using a straight line. Through the fitted model, the estimated coefficient on each variable indicates the association between response variable price and the predictor. The regression model is

$$price = \beta_0 + \beta_1 * accommodates + \epsilon \qquad (model\ 1)$$

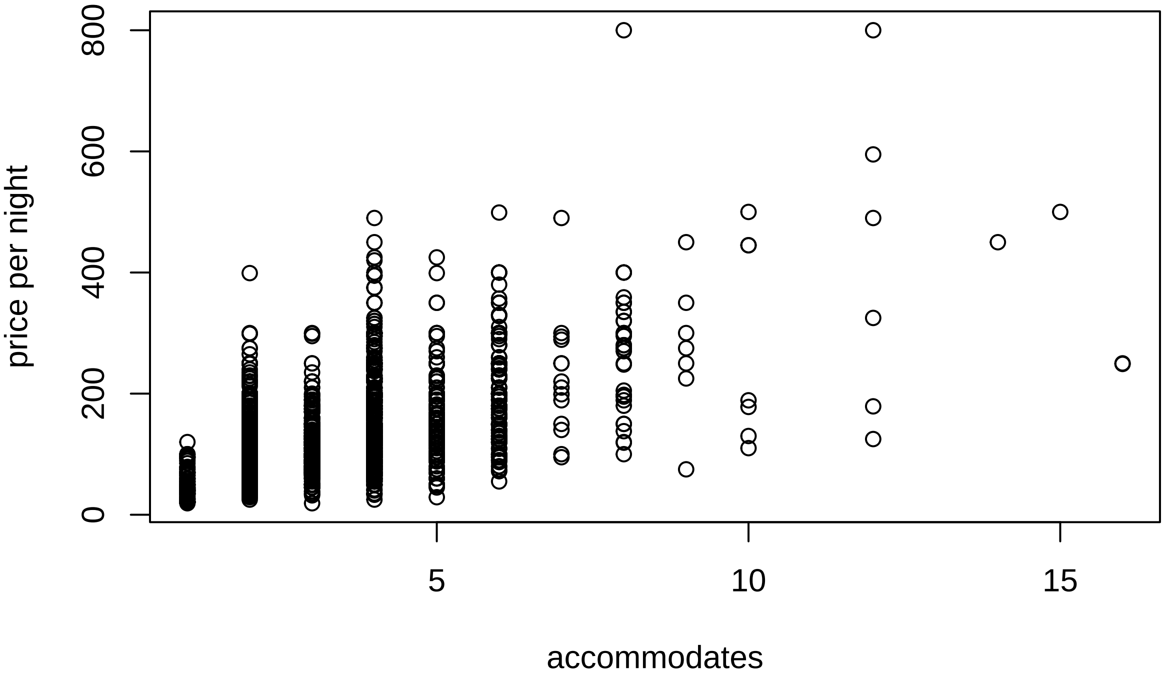


Figure 3: Scatter plot between price and accommodates

Table 2: Regression summary for model 1

term	estimate	std.error	statistic	p.value
(Intercept)	46.601	1.772	26.304	0
accommodates	24.605	0.541	45.496	0

Table 2 lists the regression summary. It is found that the accommodates has significant effect on the performance of price. The coefficient of determination is around 0.31, that about 31% variation of the price can be explained by variable accommodates.

3.3 Q3

The object is to add more independent variables to the simple model, and to find besides accommodates, what are the other determinants of the price. Table 3 presents the regression summary. We can notice that except the variable host response rate, all the other estimated coefficients are statistically significant at 5% level. In general, accommodates, bathrooms, bedrooms and review scores rating have positive effect on the prices. As for the categorical variable room type, there are three levels. In the model, entire room serves as the baseline level, therefore, the negative coefficients on private room and shared room means holding other variables constant, the entire rooms cost the highest price per night.

For the assumptions assessment for the linear regression, Figure 4 displays the residuals diagnostics. The left panel shows there is no obvious pattern of the points. But the right panel, the QQ plot tells majority of the points are align with the diagonal line but some deviations on both tails. Considering the large sample size, the normality assumption is considered moderately hold.

Table 3: Regression summary for model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-18.095	10.000	-1.809	0.070
accommodates	13.249	0.675	19.630	0.000
bathrooms	38.141	2.811	13.568	0.000
bedrooms	20.127	1.344	14.976	0.000
room_typePrivate room	-37.469	2.041	-18.360	0.000
room_typeShared room	-47.818	9.664	-4.948	0.000
host_response_rate	7.209	4.505	1.600	0.110
review_scores_rating	0.317	0.096	3.302	0.001

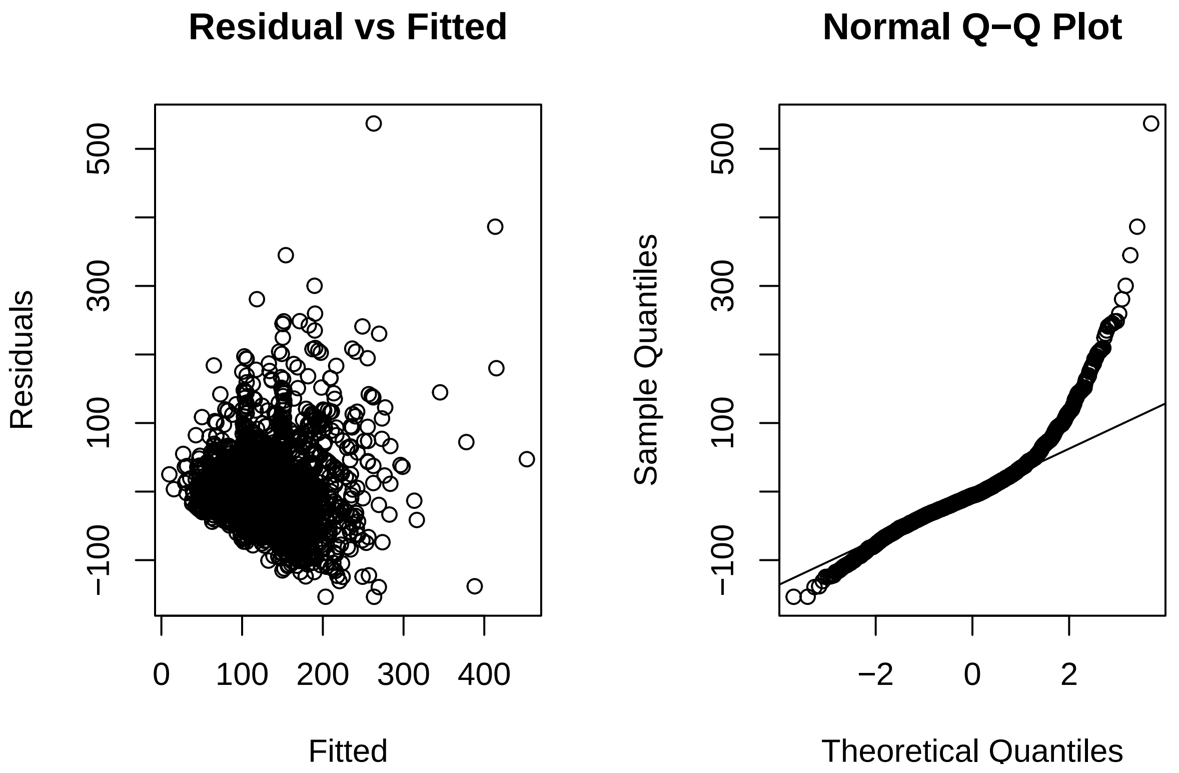


Figure 4: Residuals plots for model 2

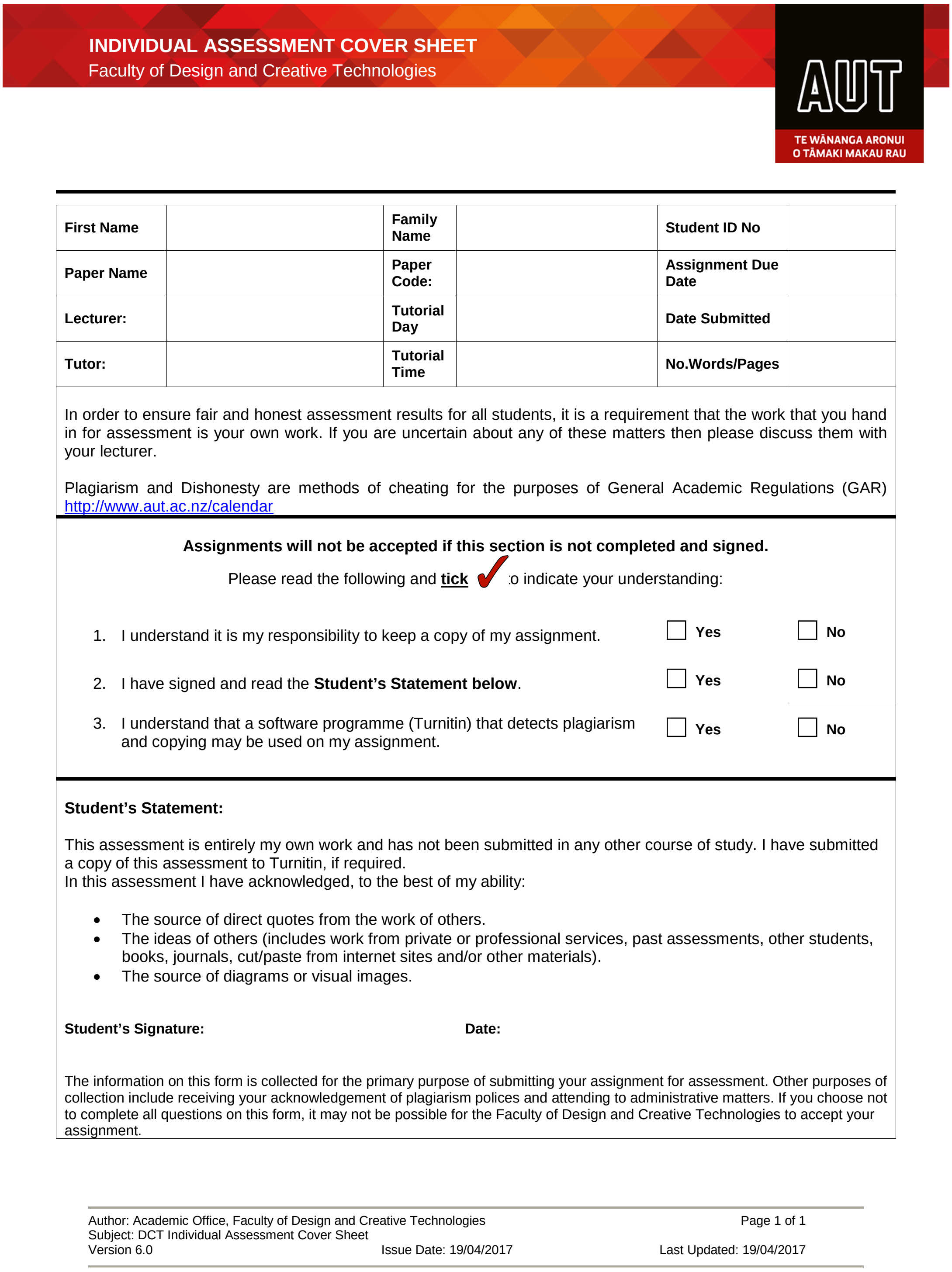
References

Yuan Cai, Yongbo Zhou, Noel Scott, et al. Price determinants of airbnb listings: evidence from hong kong. *Tourism Analysis*, 24(2):227–242, 2019.

Chen Yong Chen Yong and K Xie. Consumer valuation of airbnb listings: a hedonic pricing approach. 2017.

Augusto Voltes-Dorta and Agustín Sánchez-Medina. Drivers of airbnb prices according to property/room type, season and location: A regression approach. *Journal of Hospitality and Tourism Management*, 45:266–275, 2020.

4 Appendix: Individual Assignment Coversheet



INDIVIDUAL ASSESSMENT COVER SHEET
Faculty of Design and Creative Technologies

AUT
TE WĀNANGA ARONUI
O TĀMAKI MAKAU RAU

First Name		**Family Name**		**Student ID No**	
Paper Name		**Paper Code:**		**Assignment Due Date**	
Lecturer:		**Tutorial Day**		**Date Submitted**	
Tutor:		**Tutorial Time**		**No.Words/Pages**	

In order to ensure fair and honest assessment results for all students, it is a requirement that the work that you hand in for assessment is your own work. If you are uncertain about any of these matters then please discuss them with your lecturer.

Plagiarism and Dishonesty are methods of cheating for the purposes of General Academic Regulations (GAR) http://www.aut.ac.nz/calendar

Assignments will not be accepted if this section is not completed and signed.

Please read the following and **tick** ✓ to indicate your understanding:

1. I understand it is my responsibility to keep a copy of my assignment. □ **Yes** □ **No**
2. I have signed and read the **Student's Statement below**. □ **Yes** □ **No**
3. I understand that a software programme (Turnitin) that detects plagiarism and copying may be used on my assignment. □ **Yes** □ **No**

Student's Statement:

This assessment is entirely my own work and has not been submitted in any other course of study. I have submitted a copy of this assessment to Turnitin, if required.
In this assessment I have acknowledged, to the best of my ability:

- The source of direct quotes from the work of others.
- The ideas of others (includes work from private or professional services, past assessments, other students, books, journals, cut/paste from internet sites and/or other materials).
- The source of diagrams or visual images.

Student's Signature: **Date:**

The information on this form is collected for the primary purpose of submitting your assignment for assessment. Other purposes of collection include receiving your acknowledgement of plagiarism polices and attending to administrative matters. If you choose not to complete all questions on this form, it may not be possible for the Faculty of Design and Creative Technologies to accept your assignment.

Author: Academic Office, Faculty of Design and Creative Technologies Page 1 of 1
Subject: DCT Individual Assessment Cover Sheet
Version 6.0 Issue Date: 19/04/2017 Last Updated: 19/04/2017

5 Appendix: R Environment

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.3
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dyli
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dyli
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Asia/Shanghai
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] broom_1.0.5 readr_2.1.4 knitr_1.45  dplyr_1.1.4
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.4       cli_3.6.1         rlang_1.1.2       xfun_0.41
##  [5] highr_0.10        purrr_1.0.2       generics_0.1.3    glue_1.6.2
##  [9] backports_1.4.1   htmltools_0.5.7   hms_1.1.3         fansi_1.0.5
## [13] rmarkdown_2.25    evaluate_0.23     tibble_3.2.1      tzdb_0.4.0
## [17] fastmap_1.1.1     yaml_2.3.7        lifecycle_1.0.4   compiler_4.3.2
## [21] pkgconfig_2.0.3   tidyr_1.3.0       rstudioapi_0.15.0 digest_0.6.33
## [25] R6_2.5.1          tidyselect_1.2.0  utf8_1.2.4        pillar_1.9.0
## [29] magrittr_2.0.3    withr_2.5.2       tools_4.3.2
```