STAT 344 Project

(November 11, 2022)

Chenyue Qian(Group Leader)	**Student Number: 74299629**
Hanxi Chen	**Student Number: 96132642**
Jun Zheng	**Student Number: 10827335**
Tianyang Wang	**Student Number: 50797802**
Xuanye He	**Student Number: 32501744**

Project Writing:

Chenyue Qian (Section 2 and 3)
Hanxi Chen(Section 4 and Part II)
Jun Zheng and Tianyang Wang (Section 1)
Xuanye He (Part II)

Logic and Spell Checking:

Chenyue Qian, Tianyang Wang, Hanxi Chen

LaTex Formatting:

Chenyue Qian, Jun Zheng, Tianyang Wang

Data Finding and Exploration:

Chenyue Qian, Xuanye He

R-Code Writing:

Chenyue Qian (Most of the Coding)
Hanxi Chen(Data Visualization)

Part I

1 Introduction

1.1 Objectives

For the housing market, housing bubble is usually a hotly-debated question for individuals and can pose significant risks to home sellers. In Greater Melbourne (referred as Melbourne in the following context), there was a housing bubble from 2016 to the beginning of 2018 (Zhou, 2019). In that period, Melbourne's traded houses' prices were experiencing tremendous changes: first rose in an abnormal way, followed by a quick decrease at the end (*What Is the Housing Bubble? Definition, Causes and Recent Example*, 2020), which is a very unusual and unstable period for the Melbourne housing market. In this case, we are interested in the characteristics of houses sold during that period. Due to the scope of this paper, we decided to examine from two perspectives: house price and the number of car spots.

Therefore, our first characteristic of interest is to learn more about the house price in Melbourne during this period (we would refer this period as "the 16-18 housing bubble" in the later context). More specifically, we would obtain quantitative data on the price of traded houses in Melbourne during the 16-18 housing bubble and use those data to estimate the average price of traded houses in Melbourne at that time.

In addition, we believe that the number of car spots per household is a reflective characteristic of the house. As more than half of the families in Melbourne have two or more cars (City of Melbourne, n.d.), many house buyers would need at least two car spots when buying a house. Hence, our second characteristic of interest is to estimate the proportion of trade houses in Melbourne with enough car spots (at least two) during the 16-18 housing bubble. Since we consider having two or more parking spaces as a satisfactory case in this context, we refer "the proportion of trade houses in Melbourne with enough car spots car spots during the 16-18 housing bubble" as "car spot satisfaction rate" in later context for simplicity purpose.

1.2 Background

Being the largest city in the Australian state of Victoria, Melbourne is the second-most populous city in Oceania. An abundance of entertainment, unique nightlife, as well as gorgeous architecture and spectacular scenery, make Melbourne the most livable city in Australia (MacFarlane, 2022; Wong, 2021). These attributes let Melbourne attract an enormous number of immigrants each year, and has the 10th largest immigrant population among global metropolitan areas (Demographics of Melbourne, 2022). In this case, it also has an active housing market because of the high demand for house-purchasing brought by those immigrants.

This high demand also makes Melbourne's housing market become fluctuate for some specific period of time, i.e. housing bubble would arise in some years. The housing bubble is a phenomenon that house price goes up successively because of an increase in demand. Later at some point, demand decreases along with supply increases, resulting in a sharp drop in prices, indicating the burst of the bubble (*What Is the Housing Bubble? Definition, Causes and Recent Example*, 2020). In the 16-18 housing bubble, Melbourne's housing market experienced this full process of a large increase in house prices and a drop in price at the very end (Zhou, 2019). Thus, analyzing the mean house prices can give us some insights into house prices in Melbourne during the 16-18 housing bubble. Also, this estimate of the mean price can become an important reference and maybe even the benchmark for the house price in the future period of the housing bubble in Melbourne.

Furthermore, as mentioned in Objective, the demand for automobiles in Melbourne is significant. Therefore, the number of car spots would be an important characteristics of houses in that period. In this case, this study allows us to draw insights into the proportion of houses in Melbourne that can satisfy the parking demand for this large group of people.

2 Data Collection and Summaries

2.1 About the Dataset

For this project, as we directly treat an existing dataset as the whole population, we would first describe the dataset here.

Our dataset was found from Kaggle (Melbourne Housing Market), which includes information regarding all the traded houses in Melbourne during the 16-18 house price bubble. These data were scraped by Tony Pino from the Domain, which is an Australian commercial real estate portal. As there are some NA values inside it, we first remove those data and then consider the rest as the population, which is the traded houses over the house price bubble. Also, as we considered houses with $\geq$ 2 car spots as a variable of interest, we created a binary variable named Car Spot Satisfaction based on such criteria by using the information from the variable Car. All the variables related to our study along with their corresponding descriptions are included in the table below.

Variable Name	Description
House Price	House Price in Australian dollars
Type	House Types: h,u,t. Where h stands for houses, cottages, villas, semis, terraces; u stands for apartments, duplex; t stands for townhouses
Rooms	Number of rooms in the house
Bedroom2	Number of bathrooms in the house
Bathroom	Number of bathrooms in the house
Car	Number of car spots
Car Spot Satisfaction	1 for houses with ≥ 2 car spots, 0 otherwise

Then, based on this dataset, we could get information about our target population and the parameters of interest.

Target Population: The target population is all the houses in Melbourne that have been traded during the 16-18 housing bubble. From the whole dataset, we get information about the total number of population, which is $N = 20423$.

Parameters of Interest: 1. The mean price of the houses traded during the 16-18 housing bubble in Melbourne in Australian Dollars.(mean house price) 2. Proportion of traded houses with two or more parking spaces during the 16-18 housing bubble in Melbourne.(car spot satisfaction rate)

2.2 Sampling Method and Procedure

In this subsection, we would decide our two sampling methods: Simple Random Sample and Stratified Random Sample, focusing on how we set up the sample size, how we choose the auxiliary variable, how we determine the strata, and how we get the sample.

2.2.1 Simple Random Sample

- **Sample Size**

In order to use R to draw the sample from data, we would need to decide the sample size n in advance. In this study, we are interested in two parameters from the population, one being binary and one being continuous. Since we did not have reference to any prior study, we cannot make a good guess for the sample variance (s^2_{guess}). Therefore, we choose to get the sample size based on the binary data (car spot satisfaction).

To make our estimation have high accuracy, we demand a relatively small width for the 95% confidence interval. Hence, we decide the width to be 10%. In this case, the half-width (δ) would be 5% and $z_{\alpha/2}$ is 1.96, and we use the conservative guess about the population proportion, i.e. setting $p_{guess} = 0.5$. Besides, the accessibility to the population size provides $N = 20423$, which allows us to implement FPC here. With this information, formula and consequent calculation can be applied to get sample size n:

$$n_0 = \frac{(z_{\alpha/2})^2 \times p_{guess} \times (1 - p_{guess})}{\delta^2} = \frac{1.96^2 \times 0.5 \times (1 - 0.5)}{0.05^2} = 384.16$$

$$n = \frac{n_0}{1 + n_0/N} = \frac{384.16}{1 + 384.16/20423} = 377.067 \rightarrow n = 378$$

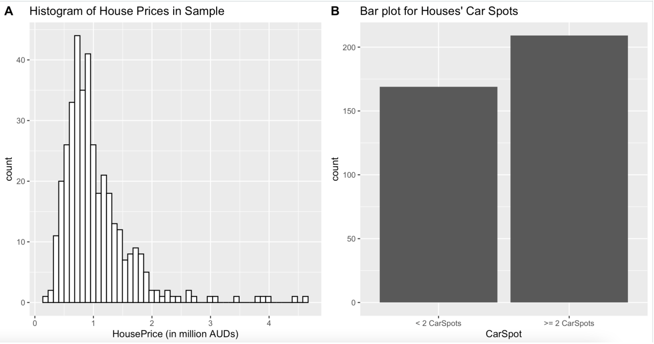
In this case, a total number of 378 houses is necessary for our sample in order to achieve a 95% confidence interval with a width less or equal to 10%.

- **Auxiliary Variable**

For the simple random sample, our group decides to use two different estimations, which are vanilla and ratio estimation. Hence, before we draw the sample, it is necessary to decide the auxiliary variable in advance. By taking a close look at three possible auxiliary variables our data provides, we choose to use Rooms as our auxiliary variable. This is because more rooms usually mean larger houses, and would consequently lead to higher prices along with more car spots. This relationship has also been shown by the linear regression model made by Zhang (2021).

- **Sampling Process**

The sample could now be drawn from the population using R (code details provided in Appendix, and we set seed to 20 for reproducibility). Here, we give an overview of our simple random sample.



2.2.2 Stratified Sample

- **Strata Selection**

In order to draw the stratified sample, the first step is to decide how we choose each stratum, which is crucial to achieve accurate estimations in a stratified sample. For this project, we would use the housing types as our stratification criteria since we believe that housing types have enormous impacts on both the number of car spots and house prices. For example, the price of a house would usually be higher than the price of an apartment due to a larger house area of a house (Wowa, n.d.). Meanwhile, a villa would usually provide more parking spots compared to an apartment. In this case, housing type would be a great choice for both house price and Car Satisfaction Rate. Hence, we would have three strata, which are "h", "t" and "u", each representing a set of different types of houses as explained in Section 2.1.

- **Sample Size**

For the total sample size n, we would use the same number as the simple random sample, which is $n = 378$. We would decide the sample size in each stratum here. In the following context, we would use subscripts h, t and u to represent our three different strata.

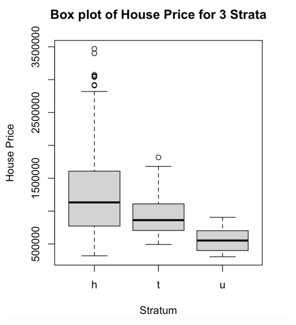
For this project, as all the data are from the same city Melbourne, we assume the variances of house prices and car spot satisfaction and sampling cost are the same for each stratum. Hence, we end up using the proportional allocation for choose sample size in each stratum. Then, achieving information about the total number of houses for each type, i.e. $N_h = 15364$, $N_t = 1577$ and $N_u = 3482$ from our dataset, we can applying the formula:

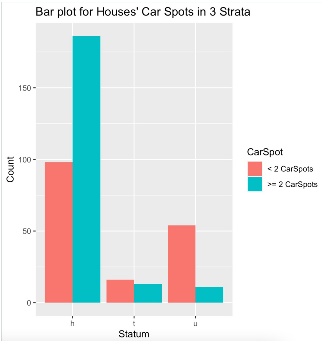
$$\frac{n_i}{n} = \frac{N_i}{N}, \text{ for stratum i}$$

So we get the sample size: $n_h = 284$, $n_t = 29$, $n_u = 65$.

- **Sampling Process**

After getting the sample size for each stratum, we can draw three simple random samples in each stratum using R (code details provided in the Appendix, and we set seed to 1000 for reproducibility). Here, we give an overview of our stratified sample.





3 Data Analysis

For the data analysis, we would estimate both mean house price and car spot satisfaction rate via two different sampling methods: simple random sampling and stratified sampling, as illustrated in the sections below.

3.1 House Price

3.1.1 Simple Random Sample

Based on previous study planning, we have $n = 378$ and $N = 20423$. In order to perform vanilla estimation, we achieve information about sample mean and sample variance of house price ($\bar{y}_s$ and s_s^2, respectively) from our sample. For ratio estimation, we assume we have additional information about the population mean of rooms in houses. In this case, we have both population mean and sample mean of rooms ($\bar{x}_p$ and $\bar{x}_s$, respectively).

Vanilla:

Estimate:

$$\bar{y}_s = 1046703$$

SE:

$$SE(vanilla) = \sqrt{(1-\frac{n}{N})\frac{s_s^2}{n}} = \sqrt{(1-\frac{378}{20423}) \times \frac{363539382759}{378}} = 30723.67$$

95% CI:

$$\bar{y}_s \pm 1.96 \times SE(vanilla) = [986484.1, 1106921]$$

Ratio:

Estimate:

$$\frac{\bar{y}_s}{\bar{x}_s} \times \bar{x}_p = 1107566$$

SE:

$$s_e^2 = \frac{1}{n-1}\sum_{i\in S} e_i^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - (\frac{\bar{y}_s}{\bar{x}_s}) \times x_i)^2 = 2.9287 \times 10^{11}$$

$$SE(ratio) = \sqrt{(1-\frac{n}{N}) \times \frac{s_e^2}{n}} = \sqrt{(1-\frac{378}{20423}) \times \frac{2.9287\times 10^{11}}{378}} = 27576.34$$

95% CI:

$$\frac{\bar{y}_s}{\bar{x}_s} \times \bar{x}_p \pm 1.96 \times SE(ratio) = [1053516, 1161616]$$

3.1.2 Stratified Sample

According to Section 3, we have $n_h = 284$, $n_t = 29$, $n_u = 65$ and $N_h = 15364$, $N_t = 1577$, $N_u = 3482$. Then, we calculated the sample mean of house price for each stratum ($\bar{y}_{S_h}, \bar{y}_{S_t}, \bar{y}_{S_u}$), along with the variance ($s^2_{S_h}, s^2_{S_t}, s^2_{S_u}$).

Estimate:

$$\bar{y}_{str} = \sum_{i\in\{h,t,u\}} \left(\frac{N_i}{N}\right) \bar{y}_{S_i} = \frac{15364}{20423} \times 1274948 + \frac{1577}{20423} \times 938936.8 + \frac{3482}{20423} \times 570319 = 1128867$$

SE:

$$SE^2[\bar{y}_{S_h}] = (1-\frac{n_h}{N_h})\frac{s^2_{S_h}}{n_h} = (1-\frac{284}{15364}) \times \frac{3.9358\times 10^{11}}{284} = 1360233925$$

$$SE^2[\bar{y}_{S_t}] = (1-\frac{n_t}{N_t})\frac{s^2_{S_t}}{n_t} = (1-\frac{29}{1577}) \times \frac{1.1307\times 10^{11}}{29} = 3827120921$$

$$SE^2[\bar{y}_{S_u}] = (1-\frac{n_u}{N_u})\frac{s^2_{S_u}}{n_u} = (1-\frac{65}{3482}) \times \frac{2.9492\times 10^{10}}{65} = 445256586$$

$$\begin{aligned} SE[\bar{y}_{str}] &= \sqrt{\sum_{i\in\{h,t,u\}} (\frac{N_i}{N})^2 SE^2[\bar{y}_{S_i}]} \\ &= \sqrt{(\frac{15364}{20423})^2 \times 1360233925 + (\frac{1577}{20423})^2 \times 3827120921 + (\frac{3482}{20423})^2 \times 445256586} \\ &= 28382.58 \end{aligned}$$

95% CI:

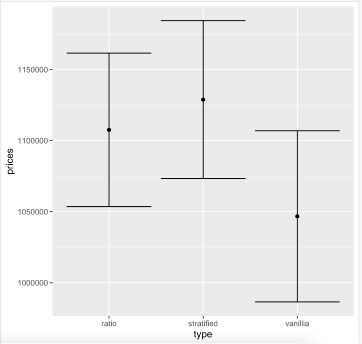
$$\bar{y}_{str} \pm 1.96 \times SE[\bar{y}_{str}] = [1073238, 1184497]$$

3.1.3 Interpretation

For the vanilla estimate, our estimate of the true mean house price is 1046703, with the standard error being 30723.67. For ratio estimation, after assuming we have the knowledge about the population mean of rooms, we obtain the value of the estimate to be 1107566 and a standard error of 27576.34. Although the sample correlation between House Price(y_i) and the number of rooms(x_i) is not a very strong positive (only about 0.45), the standard error of the ratio estimate has decreased quite a lot compared to vanilla estimate. For stratified sample, we can see that its estimate of true mean house price is 1128867 with a standard error of 28382.58, which gives a similar standard error as the ratio estimates.

By looking at the plot below, we can see that all three confidence intervals have some overlaps with each other, showing that all three might cover the true mean house price. But it is clear that the stratified estimate and the ratio estimate give more similar confidence intervals, showing a signal that these two intervals might be more trustworthy. The reason could be that stratified sample helps us avoid unrepresentative sample and the ratio estimates incorporate auxiliary variables to adjust the vanilla estimate to make it represent the whole population better. With the above information, we would choose the ratio estimate as our estimation for the true mean house price since its standard error is the smallest and its confidence interval has a bigger overlap with the other two confidence intervals. Thus, we would estimate the true mean house price to be 1107566 with margin of error equals to $1.96 \times SE(ratio)$, which is 54049.63.

Meanwhile, some other characteristics of these three estimates are worth noting. For simple random sample, it is usually more difficult to achieve compared with stratified sample in reality. For ratio estimate, although it provides a better estimate in our study, sometimes we did not have additional population information about the auxiliary variable. Besides, despite the higher accuracy of the estimate by stratified sample, it has a more complicated computation process compared with the other two, especially the vanilla estimate.



3.2 Car Spot Satisfaction

3.2.1 Simple Random Sample

Similar to the house price, we have $n = 378$ and $N = 20423$. Meanwhile, we use our sample car spot satisfaction rate ($\hat{p}$), achieved from the column Car Spot Satisfaction (z_i) to calculate estimate of true car spot satisfaction rate.

Vanilla:

Estimate:

$$\hat{p} = \frac{1}{n}\sum z_i = 0.5529$$

SE:

$$SE(vanilla) = \sqrt{(1-\frac{n}{N})\frac{\hat{p}\times(1-\hat{p})}{n}} = \sqrt{(1-\frac{378}{20423})\times\frac{0.5529\times(1-0.5529)}{378}} = 0.0253$$

95% CI:

$$\hat{p} \pm 1.96 \times SE(vanilla) = [0.5033, 0.6026]$$

Ratio:

Estimate:

$$\frac{\hat{p}}{\bar{x}_s} \times \bar{x}_p = \frac{0.5529}{2.896825} \times 3.06527 = 0.5851$$

SE:

$$s_e^2 = \frac{1}{n-1}\sum_{i\in S} e_i^2 = \frac{1}{n-1}\sum_{i\in S}(z_i - (\frac{\hat{p}}{\bar{x}_s}) \times x_i)^2 = 0.2074$$

$$SE(ratio) = \sqrt{(1-\frac{n}{N})\times\frac{s_e^2}{n}} = \sqrt{(1-\frac{378}{20423})\times\frac{0.2074}{378}} = 0.0232$$

95% CI:

$$\frac{\hat{p}}{\bar{x}_s} \times \bar{x}_p \pm 1.96 \times SE(ratio) = [0.5396, 0.6305]$$

3.2.2 Stratified Sample

Similarly, for stratified samples, we have $n_h = 284$, $n_t = 29$, $n_u = 65$ and $N_h = 15364$, $N_t = 1577$, $N_u = 3482$. And we calculated sample car spot satisfaction rate for each stratum from our sample, which is $\hat{p}_{str,h}$, $\hat{p}_{str,t}$, $\hat{p}_{str,u}$.

Estimate:

$$\hat{p}_{str} = \sum_{i\in\{h,t,u\}}(\frac{N_i}{N})\hat{p}_{str,i} = \frac{15364}{20423}\times 0.6549 + \frac{1577}{20423}\times 0.4483 + \frac{3482}{20423}\times 0.1692 = 0.5562$$

SE:

$$SE^2[\hat{p}_{str,h}] = (1-\frac{n_h}{N_h})\frac{\hat{p}_{str,h}\times(1-\hat{p}_{str,h})}{n_h} = (1-\frac{284}{15364})\times\frac{0.6549\times(1-0.6549)}{284} = 7.8105\times 10^{-4}$$

$$SE^2[\hat{p}_{str,t}] = (1-\frac{n_t}{N_t})\frac{\hat{p}_{str,t}\times(1-\hat{p}_{str,t})}{n_t} = (1-\frac{29}{1577})\times\frac{0.4483\times(1-0.4483)}{29} = 8.3716\times 10^{-3}$$

$$SE^2[\hat{p}_{str,u}] = (1-\frac{n_u}{N_u})\frac{\hat{p}_{str,u}\times(1-\hat{p}_{str,u})}{n_u} = (1-\frac{65}{3482})\times\frac{0.1692*(1-0.1692)}{65} = 2.1226\times 10^{-3}$$

$$SE[\hat{p}_{str}] = \sqrt{\sum_{i\in\{h,t,u\}}(\frac{N_i}{N})^2 SE^2[\hat{p}_{str,i}]}$$

$$= \sqrt{(\frac{15364}{20423})^2\times 7.8105\times 10^{-4} + (\frac{1577}{20423})^2\times 8.3716\times 10^{-3} + (\frac{3482}{20423})^2\times 2.1226\times 10^{-3}}$$

$$= 0.0235$$

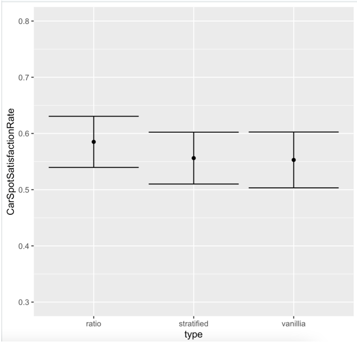
95% CI:

$$\hat{p}_{str} \pm 1.96 \times SE[\hat{p}_{str}] = [0.5100, 0.6023]$$

3.2.3 Interpretation

For the vanilla estimate, its estimate of the true car spot satisfaction rate is 0.5529 with standard error of 0.0253. For ratio estimate, it has an estimate equals to 0.5851 and a standard error equals to 0.0232. Meanwhile, the stratified sample provides an estimate and standard error of 0.5562 and 0.0235 respectively. Although the standard error for ratio estimate and estimate from the stratified sample are better, it does not improve a lot compared with vanilla estimate. The reason could be that the strata we choose or auxiliary variable we use does not influence the car spots that much compared with its influence to the house price. Hence, this shows that the effectiveness of ratio estimate largely depends on whether there is an appropriate auxiliary variable, and the estimation by stratified sample requires us to carefully choose strata to reduce more within-strata variance.

By looking at the plot of confidence intervals below, all three of them have a large overlap, showing that all of them could be quite trustworthy. For this project, we would choose the estimation from the stratified sample since it has a relatively smaller standard error along with greater overlaps with the other two confidence intervals. Thus, we would estimate the true car spot satisfaction rate to be 0.5562 with margin of error equals to 1.96 × SE for stratified sample, which is 0.046118.



4 Conclusion and Discussion

4.1 Conclusion

Through drawing samples and doing estimation for our parameters of interests, we now learn more about the characteristics of houses sold in Melbourne during 16-18 housing bubble. By examining the characteristics of those houses in two different ways, we conclude that: the true mean house price would be around 1107566 AUDs with a margin of error equals to 54049.63 AUDs 19 times out of 20, and approximate 55.62% of houses would have enough car spots (at least 2 car spots) to meet most individuals' demands with a margin of error equals to 4.6118% 19 times out of 20.

4.2 Discussion

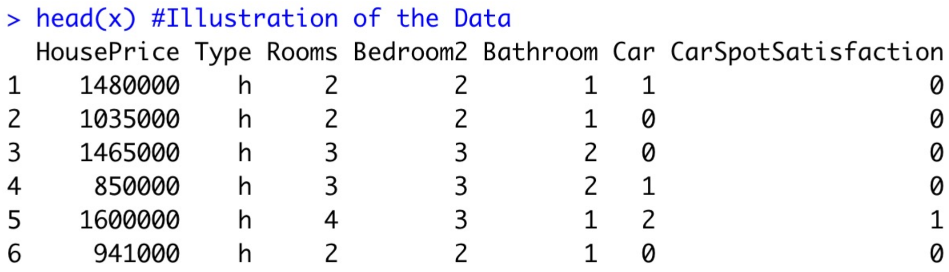
Although our project provides us with information about the characteristics of traded house during 16-18 housing bubble, there are some limitations. First of all, we believe that the conclusion would only be suitable for analyzing the traded houses during different housing bubble periods in Melbourne. This is because housing markets' characteristics would be different in disparate cities. Larger cities would have a higher average house price; developing countries and developed countries would have housing markets behave differently. Also, we assume that the variances in different strata for stratified sample are the same. If this is not the real situation, our stratified estimate may not be an optimal choice. What's more, for the analysis of our binary variable (car spot satisfaction rate), the decrease in standard error by using stratified sample and ratio estimation is not remarkable. This indicates that there might be a better choice for choosing stratum or auxiliary variables. Therefore, we believe the estimation might be improved by using other variables (not appeared in our dataset) in performing these two estimations. Meanwhile, as we remove some NA values, our dataset may not be a perfect representation of the whole population. If there is more completed dataset, the results might be improved.

Part II: Key Message of "The Emperor's New Tests"

In this paper, by starting with an example of statistical allegory, Perlman and Wu indicated the importance of using statistical intuition when assessing the effectiveness of a statistical method. The story centered around a statistical test, which is called the likelihood ratio test (LRT). With the background that the LRT has been criticized as allegedly inferior for the past two decades, a young statistician developed a New Test that was first thought as superior to the LRT by creating less biased results. However, these results are actually unwarranted since the New Test ignored the significance of variability and uncertainty. Then, it ended up being rejected and categorized as scientifically unacceptable. Hence, it is clear that the decision is made mainly based on common sense and intuition in statistics. Even though the test leads to a better result in some sense, saying something is warranted when it actually has a high level of variability is clearly against statistical intuition. In this case, Perlman and Wu demonstrated that statistical intuition constitutes a vital part of statistical science and should be incorporated into the analysis to help individuals gain a more comprehensive understanding of the subject or issue at hand.

5 Appendix

5.1 Data



```
> head(x) #Illustration of the Data
  HousePrice Type Rooms Bedroom2 Bathroom Car CarSpotSatisfaction
1    1480000    h     2        2        1   1                   0
2    1035000    h     2        2        1   0                   0
3    1465000    h     3        3        2   0                   0
4     850000    h     3        3        2   1                   0
5    1600000    h     4        3        1   2                   1
6     941000    h     2        2        1   0                   0
```

Pino, T. (2018, October 15). *Melbourne Housing Market.* Kaggle. https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market?select=Melbourne_housing_FULL.csv

5.2 References

City of Melbourne. (n.d.). *Number of cars per household | City of Melbourne | Community profile.* Retrieved November 8, 2022, from https://profile.id.com.au/melbourne/car-ownership

Demographics of Melbourne. (2022). In *Wikipedia.* https://en.wikipedia.org/wiki/Demographics_of_Melbourne

MacFarlane, R. (2022, July 26). *10 Awesome Reasons to Live in Melbourne.* Insider Guides. https://insiderguides.com.au/reasons-live-melbourne/

What Is the Housing Bubble? Definition, Causes and Recent Example. (2020, December 25). Investopedia. https://www.investopedia.com/terms/h/housing_bubble.asp

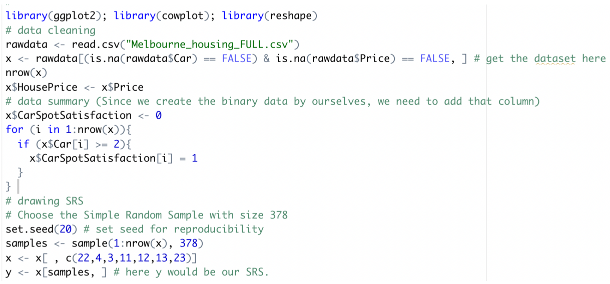
Wong, H. (2021, February 19). *What Makes Melbourne Attractive?* Melbourne Tours. https://melbournecitytour.com.au/blog/what-makes-melbourne-attractive/

Wowa. (n.d.). *Vancouver Housing Market.* Retrieved November 7, 2022, from https://wowa.ca/vancouver-housing-market

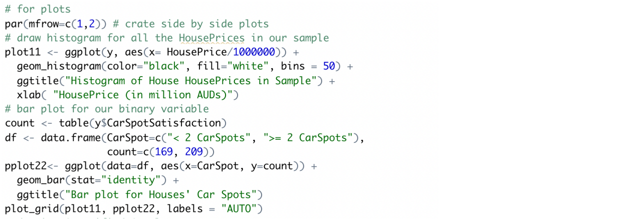
Zhang, Q. (2021, October 29). *Housing Price Prediction Based on Multiple Linear Regression.* Hindawi. https://www.hindawi.com/journals/sp/2021/7678931/

Zhou, N. (2019, January 2). *Australian house prices falling at fastest rate in a decade. The Guardian. What Is the Housing Bubble? Definition, Causes and Recent Example.* (2020, December 25). Investopedia. https://www.investopedia.com/terms/h/housing_bubble.asp

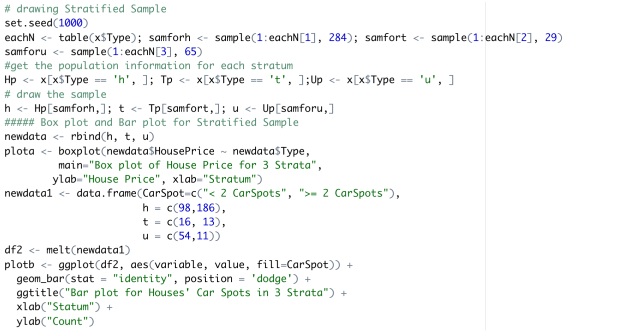
5.3 R code



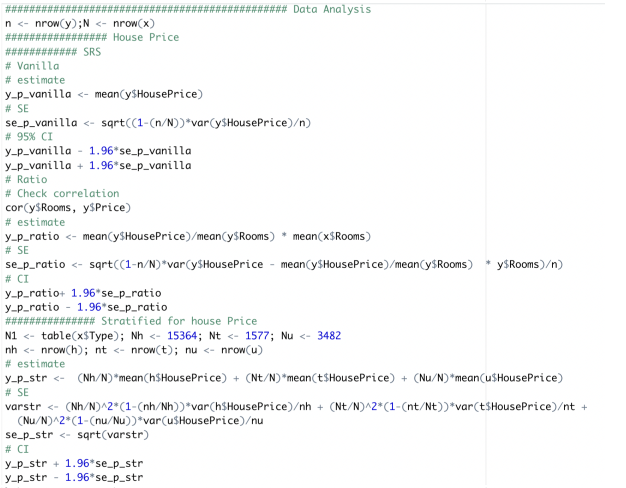
```r
library(ggplot2); library(cowplot); library(reshape)
# data cleaning
rawdata <- read.csv("Melbourne_housing_FULL.csv")
x <- rawdata[(is.na(rawdata$Car) == FALSE) & is.na(rawdata$Price) == FALSE, ] # get the dataset here
nrow(x)
x$HousePrice <- x$Price
# data summary (Since we create the binary data by ourselves, we need to add that column)
x$CarSpotSatisfaction <- 0
for (i in 1:nrow(x)){
  if (x$Car[i] >= 2){
    x$CarSpotSatisfaction[i] = 1
  }
}
# drawing SRS
# Choose the Simple Random Sample with size 378
set.seed(20) # set seed for reproducibility
samples <- sample(1:nrow(x), 378)
x <- x[ , c(22,4,3,11,12,13,23)]
y <- x[samples, ] # here y would be our SRS.
```



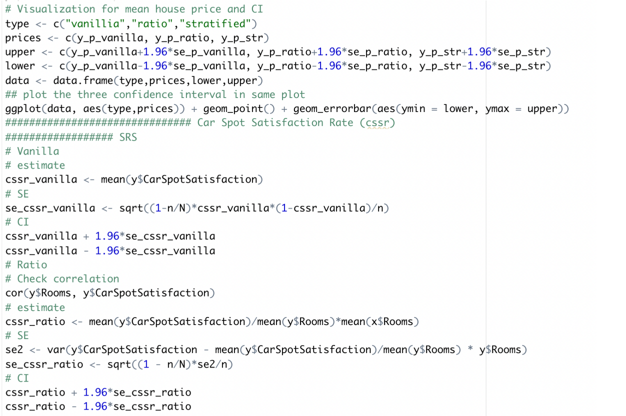
```
# for plots
par(mfrow=c(1,2)) # crate side by side plots
# draw histogram for all the HousePrices in our sample
plot11 <- ggplot(y, aes(x= HousePrice/1000000)) +
  geom_histogram(color="black", fill="white", bins = 50) +
  ggtitle("Histogram of House HousePrices in Sample") +
  xlab( "HousePrice (in million AUDs)")
# bar plot for our binary variable
count <- table(y$CarSpotSatisfaction)
df <- data.frame(CarSpot=c("< 2 CarSpots", ">= 2 CarSpots"),
                 count=c(169, 209))
pplot22<- ggplot(data=df, aes(x=CarSpot, y=count)) +
  geom_bar(stat="identity") +
  ggtitle("Bar plot for Houses' Car Spots")
plot_grid(plot11, pplot22, labels = "AUTO")
```



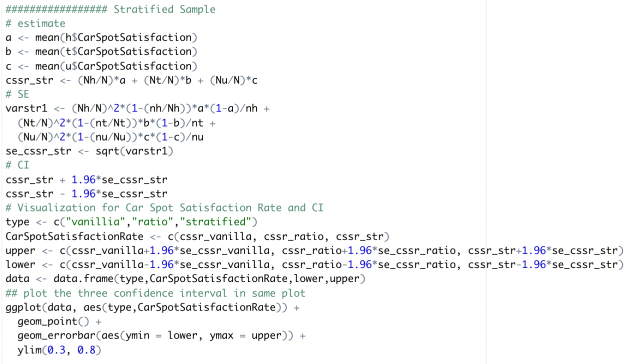
```
# drawing Stratified Sample
set.seed(1000)
eachN <- table(x$Type); samforh <- sample(1:eachN[1], 284); samfort <- sample(1:eachN[2], 29)
samforu <- sample(1:eachN[3], 65)
#get the population information for each stratum
Hp <- x[x$Type == 'h', ]; Tp <- x[x$Type == 't', ];Up <- x[x$Type == 'u', ]
# draw the sample
h <- Hp[samforh,]; t <- Tp[samfort,]; u <- Up[samforu,]
##### Box plot and Bar plot for Stratified Sample
newdata <- rbind(h, t, u)
plota <- boxplot(newdata$HousePrice ~ newdata$Type,
         main="Box plot of House Price for 3 Strata",
        ylab="House Price", xlab="Stratum")
newdata1 <- data.frame(CarSpot=c("< 2 CarSpots", ">= 2 CarSpots"),
                       h = c(98,186),
                       t = c(16, 13),
                       u = c(54,11))
df2 <- melt(newdata1)
plotb <- ggplot(df2, aes(variable, value, fill=CarSpot)) +
  geom_bar(stat = "identity", position = 'dodge') +
  ggtitle("Bar plot for Houses' Car Spots in 3 Strata") +
  xlab("Statum") +
  ylab("Count")
```



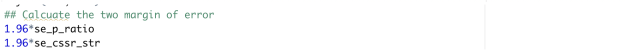
```
############################################### Data Analysis
n <- nrow(y);N <- nrow(x)
################# House Price
############ SRS
# Vanilla
# estimate
y_p_vanilla <- mean(y$HousePrice)
# SE
se_p_vanilla <- sqrt((1-(n/N))*var(y$HousePrice)/n)
# 95% CI
y_p_vanilla - 1.96*se_p_vanilla
y_p_vanilla + 1.96*se_p_vanilla
# Ratio
# Check correlation
cor(y$Rooms, y$Price)
# estimate
y_p_ratio <- mean(y$HousePrice)/mean(y$Rooms) * mean(x$Rooms)
# SE
se_p_ratio <- sqrt((1-n/N)*var(y$HousePrice - mean(y$HousePrice)/mean(y$Rooms)  * y$Rooms)/n)
# CI
y_p_ratio+ 1.96*se_p_ratio
y_p_ratio - 1.96*se_p_ratio
############### Stratified for house Price
N1 <- table(x$Type); Nh <- 15364; Nt <- 1577; Nu <- 3482
nh <- nrow(h); nt <- nrow(t); nu <- nrow(u)
# estimate
y_p_str <-  (Nh/N)*mean(h$HousePrice) + (Nt/N)*mean(t$HousePrice) + (Nu/N)*mean(u$HousePrice)
# SE
varstr <- (Nh/N)^2*(1-(nh/Nh))*var(h$HousePrice)/nh + (Nt/N)^2*(1-(nt/Nt))*var(t$HousePrice)/nt +
  (Nu/N)^2*(1-(nu/Nu))*var(u$HousePrice)/nu
se_p_str <- sqrt(varstr)
# CI
y_p_str + 1.96*se_p_str
y_p_str - 1.96*se_p_str
```



```
# Visualization for mean house price and CI
type <- c("vanillia","ratio","stratified")
prices <- c(y_p_vanilla, y_p_ratio, y_p_str)
upper <- c(y_p_vanilla+1.96*se_p_vanilla, y_p_ratio+1.96*se_p_ratio, y_p_str+1.96*se_p_str)
lower <- c(y_p_vanilla-1.96*se_p_vanilla, y_p_ratio-1.96*se_p_ratio, y_p_str-1.96*se_p_str)
data <- data.frame(type,prices,lower,upper)
## plot the three confidence interval in same plot
ggplot(data, aes(type,prices)) + geom_point() + geom_errorbar(aes(ymin = lower, ymax = upper))
################################ Car Spot Satisfaction Rate (cssr)
################## SRS
# Vanilla
# estimate
cssr_vanilla <- mean(y$CarSpotSatisfaction)
# SE
se_cssr_vanilla <- sqrt((1-n/N)*cssr_vanilla*(1-cssr_vanilla)/n)
# CI
cssr_vanilla + 1.96*se_cssr_vanilla
cssr_vanilla - 1.96*se_cssr_vanilla
# Ratio
# Check correlation
cor(y$Rooms, y$CarSpotSatisfaction)
# estimate
cssr_ratio <- mean(y$CarSpotSatisfaction)/mean(y$Rooms)*mean(x$Rooms)
# SE
se2 <- var(y$CarSpotSatisfaction - mean(y$CarSpotSatisfaction)/mean(y$Rooms) * y$Rooms)
se_cssr_ratio <- sqrt((1 - n/N)*se2/n)
# CI
cssr_ratio + 1.96*se_cssr_ratio
cssr_ratio - 1.96*se_cssr_ratio
```



```
################ Stratified Sample
# estimate
a <- mean(h$CarSpotSatisfaction)
b <- mean(t$CarSpotSatisfaction)
c <- mean(u$CarSpotSatisfaction)
cssr_str <- (Nh/N)*a + (Nt/N)*b + (Nu/N)*c
# SE
varstr1 <- (Nh/N)^2*(1-(nh/Nh))*a*(1-a)/nh +
  (Nt/N)^2*(1-(nt/Nt))*b*(1-b)/nt +
  (Nu/N)^2*(1-(nu/Nu))*c*(1-c)/nu
se_cssr_str <- sqrt(varstr1)
# CI
cssr_str + 1.96*se_cssr_str
cssr_str - 1.96*se_cssr_str
# Visualization for Car Spot Satisfaction Rate and CI
type <- c("vanillia","ratio","stratified")
CarSpotSatisfactionRate <- c(cssr_vanilla, cssr_ratio, cssr_str)
upper <- c(cssr_vanilla+1.96*se_cssr_vanilla, cssr_ratio+1.96*se_cssr_ratio, cssr_str+1.96*se_cssr_str)
lower <- c(cssr_vanilla-1.96*se_cssr_vanilla, cssr_ratio-1.96*se_cssr_ratio, cssr_str-1.96*se_cssr_str)
data <- data.frame(type,CarSpotSatisfactionRate,lower,upper)
## plot the three confidence interval in same plot
ggplot(data, aes(type,CarSpotSatisfactionRate)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  ylim(0.3, 0.8)
```



```
## Calcuate the two margin of error
1.96*se_p_ratio
1.96*se_cssr_str
```