Analyzing the Financial Outcomes of Renting vs. Buying a Home in San Francisco Research Design

University of California, Berkeley | School of Information

DATASCI 201 Research Design and Applications for Data and Analysis - Spring 2022

April 2, 2022

Team #4: Ivy Chan, Isabelle Engelberg, Mwanga Malcom Lupampa

Intended Audience

Individuals who are considering buying or renting a home in San Francisco

Overview

Determining whether to buy or rent a home involves a complex decision-making process that can be further convoluted by an array of conflicting research (Olick, 2022; Kolomatsky, 2021; Beracha & Johnson, 2021). While multiple companies have attempted to create user-friendly calculators to compare the cost of renting or buying over time (NerdWallet, 2022; Realtor.com, 2022; The New York Times, 2014), these sites have been shown to miss key financial factors (Cox & Followill, 2018).

Our research aims to build on the existing literature and tools referenced above by creating and analyzing a more complete dataset on the total costs of renting or owning a home using survey data. For the purpose of this study, we will focus solely on the financial costs of renting vs. buying a home in order to limit the dependent variables and produce a single, tangible outcome. We will therefore omit the impact of intangible factors such as stability, financial predictability, freedom to renovate, pride of ownership, the flexibility of moving, etc. from our analysis. Additionally, we will not consider price appreciation or depreciation since the house price fluctuates based on the unpredictable housing market condition.

To conduct this analysis, we will identify two randomized, equivalent groups. One made up of renters and one made up of homeowners, and compare the total costs accumulated after a period of four years. We will use the most recent data available (2018-2021). We have chosen to focus on this time period as it represents our best estimation of a housing market life cycle phase (Kaiser, 1997). While real estate life cycles can vary widely, they are estimated to last around 18 years. As the real estate life cycle is made up of four phases (recovery, expansion, hyper supply, and recession), we can infer that each phase lasts around 4.5 years (Kaiser, 1997). Additionally, limiting our study to only one geographical location (San Francisco) will allow us to perform an in-depth analysis within our budget and time constraints. San Francisco is an ideal location for this analysis as it is a population-dense, diverse, metropolitan location with a wide variety of housing options (McCann, 2021; San Francisco Planning Department, 2018).

This research will act as an essential resource for individuals who are in the process of deciding to buy or rent a home in San Francisco, while also providing a framework for those looking to analyze the rent vs. buy decision in other locations.

Research Question

Does renting a home cost more than buying a home in San Francisco over four years - 2018-2021?

- By controlling a host of variables (see below variables section for details), our research focuses on comparing the total costs of renting and buying and ignores the intangible factors.

Potential Sub-Questions:

- Do one-time costs or recurring costs contribute more to the difference in the cost of renting vs. buying?
- How have the costs of renting and buying changed over the four-year time period?

Study Design

We plan on conducting an observational study relying on cross-section data collected through various sources for the 4 year time period. We consider this approach because:

- It's inexpensive and quick to develop a working model (Wang & Cheng, 2020).
- We are only interested in residents living in the San Francisco area that are likely to move within a period of 2 to 4 years.
- There is no need for follow-up (Wang & Cheng, 2020). Market data changes have been influenced by the recent covid-19 pandemic where home prices in the area increased significantly but rent prices went drastically down with rent falling roughly 24% according to an article by guardian.com (Canon, G. 2021, Jan). This also further supports our assumption of not requiring a follow-up. This trend is likely to change if we go back to the 'work from the office' work model.

We will design and create a survey (University of Michigan, 2018) to get accurate costs of renting or buying a home. The survey in the form of paper mail-in surveys to collected addresses from partner companies like apartments.com and zillow.com.

For the paper mails, we will send surveys in the form of questionnaires with a prepaid mailing envelope which users can mail back after filling in the forms. The survey requirements will be as below:

- An explanation of what the survey is and the intended use of the data will be given in the introduction. An explanation will be provided that participation is voluntary and by submitting the survey users give us permission to use the collected data.
- Participants will be informed that mailing the form back to us is consent for participation and compensation will be in the form of a gift card mailed back to users.

- Participants will also be advised that in accordance with the California Privacy Consumer's Act (CPRA)any collected data will only be stored in an encrypted form for up to a year and participation in the survey is the acknowledgment of this requirement. (Yates, W. 2021, Feb)
- The data from the returned surveys will be converted into a digital format that we can further automate in our data analysis pipeline.
- No personal information will be collected on the actual form besides the mailing address for the gift card. Only general data on the costs of renting or owning a home will be collected on the forms. Responses will be predetermined and stratified so as to avoid the risk of re-identification.

Because we will not collect any personal data besides names and home addresses, the responses to questions on costs will be in the form of stratified multiple choice answers, and digitally converted data from the surveys will be anonymous, we do not see any possible harm to survey participants as de-anonymization techniques would be difficult to reproduce the data. We also don't see the need of using an Institutional Research Board (IRB) board.

For the data collection, we will partner with apartment.com and zillow.com to obtain two randomized sample groups - rent and buy. Both groups have comparable demographics, types of dwellings, number of bedrooms, etc. The only difference between these two groups is rent or buy.

The data holding the returned mail-in surveys will have participant names and home addresses. The name will be redacted immediately after mailing back the gift cards and the un-redacted file encrypted and stored for the time we will have communicated to participants and then all and any personal data will be deleted.

Data

Recruitment

We will partner with zillow.com to identify randomized samples of homeowners in San Francisco and combine them with the county record to find the owners' names and contact information. To comply with Belmont Report's principles (Office for Human Research Protections, 2021), we will contact the subjects to obtain agreements to participate in the study.

We will partner with apartments.com to obtain randomized renters' samples in San Francisco. We will then contact the subjects for agreement to participate in the study.

Sample

We will use the population size, margin of error, and confidence level to estimate the sample size. Using an online sample size calculator from qualtrics.com, we computed the sample size of 384 based on the population of San Francisco 874,000, the standard margin of error of 5%, and the confidence level of 95% (Determining sample size, n.d.). In addition, according to TownCharts.com, 62% of households rent homes in San Francisco (San Francisco, California Housing Data, n.d.). Based on these two pieces of information, we plan to collect a sample size of 238 renters and 146 homeowners in our study.

To establish causality between rent/buy and the outcome variable (cost of living), we will conduct an independent variable balance check to confirm these two groups are equivalent in various characteristics (Please see the Variables section for more information on the independent variables). However, if the independent variable balance check fails, we encounter the selection risk that we will explain in the Potential Risks section.

Variables

The observed outcome of the study

The study will observe and compare the total costs associated with renting or buying a house over four years.

In our survey for renters, we will collect the following data to calculate the cost of renting: monthly rent, security deposit, renter's insurance, utilities, and any additional fees (pet fees, etc.). These are the key variables for the cost of renting calculation.

In our survey for homeowners, we will collect the following data to calculate the cost of homeownership: monthly mortgage payment, mortgage insurance, utilities, property tax, HOA (homeowner association) fee, marginal income tax rate, interest rate, down payment, purchase price, closing costs, and maintenance costs. These are the key variables for the cost of homeownership calculation.

The variable in research

The variable in research is a binary variable with only two values: Renting or Buying. By controlling a host of variables, we strive to make it the only variable that is different between the renting and buying groups.

The independent variables for control

We compare renting and buying costs by controlling a host of independent variables, sometimes called covariates. These variables improve the precision of the study result and allow us to confirm that we have two equivalent groups of samples for comparison. According to middleprofessors.com, "In observational designs, covariates might be added to a model to 1) increase predictive ability, 2) because the researcher is interested in specific conditional effects, or 3) to eliminate confounding.". (Adding Covariates to a linear model, n.d.).

We organized our independent variances into two groups: 1) Demographics and 2) House Types

In the demographics, we consider these independent variables: age, marital status, size of household, household income, based on some articles and previous studies on home-buying criteria or mortgage application criteria ("Buying a house before vs. after marriage," 2021; "First time home buyer save California homeownership rate," 2019; "Rent vs. Buy - what's right for you?" n.d; "Rent or Buy?", n.d; Bostock, Carter, and Tse, n.d). Including independent variables can increase the precision of the model by decreasing the standard errors. In addition, it

ensures that our two randomized samples are equivalent. Finally, they open up opportunities for further causality experiments if we discover a statistically significant relationship between an independent variable and the outcome.

For house types, we consider these independent variables: type of dwelling (such as House, Condo, Apartment, Townhouse), number of bedrooms, number of bathrooms, inclusion of garage, inclusion of basement/attic, and square footage (which is a determinant for bedrooms, baths, garage, and basement). We select these variables based on the common house search criteria on zillow.com, apartment.com, and craigslist.org. Since we want to obtain two equivalent randomized groups, we must have a good mixture of different housing types.

Statistical Methods

We plan to use the common statistics language R and the linear regression package to generate the model. The linear model generates a best-fit model and estimates the coefficients of the variable in the study (rent/buy) and control variables. The coefficient of the variable rent/buy is the result of the difference between renting and buying costs. Please see an example model output from Princeton University in Appendix A. ("Using stargazer to report regression output", n.d.)

In addition, we will calculate the p-value and set statistical significance at less than 0.05. We will also compute the standard errors and confidence interval to determine the accuracy of the effect of the study - the difference in the cost of renting and buying.

Potential Risks

Privacy and security

Because we will collect all this data and store it in one place for analysis, there is still a risk of de-anonymizing if this dataset were to be leaked, per an article on techcrunch.com which states that methods used in anonymizing data can not protect against re-identification of complex data sets with personal information, per a study done by researchers at two European universities. (Lomas, 2019). Even though our collection of Personal Information is limited to names and addresses, this data together with other data sets in one place poses a high risk to our participants. So we will make sure the data is always encrypted and access to the encrypted dataset is only granted to team members working on the project. Also, most data will only be stored for the length and validity of the study. Personal data like home addresses and names will be stored for a shorter period of time so we need to assess better what a reasonable time frame is.

Biases

Omitted Variable Bias

One of the challenges of an observational study is the omitted variable bias. We strive to compare renting and buying costs by obtaining two equivalent groups and controlling a host of variables. However, there are always some omitted variables that may affect the outcome. For example, some may argue that a house near public transportation can cause a higher renting

but not buying cost. To mitigate the risks, we reference over ten articles on rent vs. buy and compile a list of criteria and controlling variables.

Selection bias (non-equivalent groups)

Because rentals are mostly apartments and homes are stand-alone units it may be hard to compare things like the square footage of the home as parking spaces are not always accounted for in apartment homes or condos. Also, there are variations we may not be able to capture that determine home costs. To mitigate this risk we will assume square footage to include all livable and unit-attached common spaces as documented by our data sources.

Attrition bias

Because we are going to collect some financial data even if stratified to help with anonymity, there is still a large potential of users opting out of the survey due to reluctance of providing the data.

As suggested by Pritha Bhandari, an experiment researcher, we will use various techniques to reduce attrition: secure commitment and agreement, provide compensation for participation, send routine reminders, and set up an online and easy-to-use questionnaire to reduce the effort to participate. (Bhandari, 2021)

Cost of surveys

Compensation of participants

Due to concerns about minimum wage and the Belmont principle of beneficence (Office for Human Research Protections, 2021), we would need to work with both our legal representative and our accounting representative to determine what is fair compensation for participants who participate in the San Francisco area while also keeping the compensation within our budget.

Data collection and survey data analog to digital conversion costs

Converting the survey data into a digital format that will allow for automated analysis of the data will be a challenge. We would need to run a dummy trial for the conversion of a complete form and determine the cost of the process, then expand it to the number of actual surveys we will do. Also, costs of new data may not be immediately available as we may have some bad data points in the samples, but we will make our best effort to estimate this cost.

Deliverables

Phase 1 - Participant Recruitment and Survey Design (2 months)

In Phase 1, we will complete the participant recruitment for our survey while finalizing the survey design. Deliverables for this phase include:

- An up-to-date list of 238 renters' and 146 homeowners' addresses and emails in the San Francisco area.
- A validated online survey.

- A written explanation of the survey and its use.
- A written consent from the subjects.

Phase 2 - Data Collection (3 weeks)

In Phase 2, we will complete the data collection by way of an online survey. Deliverables for this phase include:

- Completed housing surveys. Participants will be invited by email to complete the survey.
- A comprehensive list of participants who declined to participate.

Phase 3 - Data Analysis and Report (2 months)

In Phase 3, we will complete the data analysis and compile a report of our findings. Deliverables for this phase include:

- A written report with the sections:
    - Introduction
    - Method
    - Results
    - Discussion
    - References
- A PowerPoint presentation summarizing the report.

Phase 4 - Post-Analysis (2 weeks)

Following Phase 3, we will complete a post-mortem analysis of our study. Deliverables for this phase include:

- A summary of lessons learned.
- Opportunities for future work from this study.

Statement of Contribution

Ivy

All team members participated in discussions of the project, editing and giving feedback on each section and all project deliverables. Ivy researched and wrote the Data, Sample, Variables, and Statistical Methods.

Our team cooperated and contributed to selecting the topic, defining the scope, identifying the variables, and providing critics to the sections in the project within a tight schedule. I appreciated that all of us were very responsive and open to discussion. Our original scope was to compare the costs and benefits between rent and buying. However, the scope was enormous and contained too many intangible factors. Therefore, we narrowed the scope to only an observational study on the financial cost comparison. But an observational study is prone to bias and hard to establish causality. So I will structure the project as an experiment if we have more time.

Isabelle

All team members participated in discussions of the project, editing and giving feedback on each section and all project deliverables. Isabelle researched and wrote the Overview, Research Question, and Deliverables.

Our team did a good job continuously refining our research proposal to 1) limit the scope to the most reasonable and clear questions and outcomes and 2) justify each aspect of the proposal. If we were to do this project again, I would like to include more qualitative research through the study to get a more well-rounded understanding of the difference in renting vs. buying and all of the factors that go into it.

Malcolm

All team members participated in discussions of the project, editing and giving feedback on each section and all project deliverables. Malcolm researched and wrote the Study Design and Potential risks.

I feel we have done a good job of determining the most necessary variables. Given more time and resources, I would put similar homes in similar buckets and also make sure to collect all necessary data consistent with each home as a data point. I would then randomly sample from each group an equal number of data points along with their corresponding data. This would account for consistency in the data where certain properties are maintained along with each data point due to sampling being done once.

References

Beracha, E., & Johnson, K. H. (2021). Lessons from Over 30 Years of Buy versus Rent Decisions: Is the American Dream Always Wise? *Real Estate Economics, 40*(2), 217-247.

Bhandari, P (2021). Attrition Bias | Examples, Explanation, Prevention. Retrieved from https://www.scribbr.com/research-bias/attrition-bias/

Canon, G. (2021, Jan). San Francisco rents are plummeting – but its housing crisis could get worse. https://www.theguardian.com/us-news/2021/jan/09/san-francisco-rents-housing-crisis-covid-pandemic

Cox, A., & Followill, R. (2018). To Rent or Buy? A 30-Year Perspective. *Journal of Financial Planning, 31*(5), 48-55. Retrieved from Financial Planning Association.

How to Determine the Correct Sample Size. (2022, March 30). Qualtrics. https://www.qualtrics.com/experience-management/research/determine-sample-size/

Kaiser, R. (1997). The Long Cycle in Real Estate. *Journal of Real Estate Research, 14*(3), 233-257.

Kolomatsky, M. (2021, June 10). *Renting Is Cheaper Than Buying, Almost Everywhere*. Retrieved from The New York Times: https://www.nytimes.com/2021/06/10/realestate/renting-cheaper-than-buying.html

Lomas, N (2019, July) Researchers spotlight the lie of ‘anonymous’ data. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/

McCann, A. (2021, February 17). *Most & Least Ethnically Diverse Cities in the U.S.* Retrieved from WalletHub: https://wallethub.com/edu/cities-with-the-most-and-least-ethno-racial-and-linguistic-diversity/10264

Mendoza, C. (2019, October 4). *Zillow: Millennials are moving every two years for their careers*. Retrieved from MPA Magazine: https://www.mpamag.com/us/news/general/zillow-millennials-are-moving-every-two-years-for-their-careers/179718

NerdWallet. (2022). *Rent vs Buy Calculator: Should I Rent or Buy?* Retrieved from NerdWallet: https://www.nerdwallet.com/mortgages/rent-vs-buy-calculator

Office for Human Research Protections (OHRP). (2021, June 16). *The Belmont Report*. HHS.gov. Retrieved March 28, 2022, from Office for Human Research Protections (OHRP): https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Olick, D. (2022, January 6). *Should you rent or buy a home? Prices are surging in both cases, which makes it complicated*. Retrieved from CNBC: https://www.cnbc.com/2022/01/06/should-you-rent-or-buy-a-home-the-answer-is-getting-more-complicated.html

Realtor.com. (2022). *Rent Vs. Buy Calculator - Buying or Renting a House*. Retrieved from Realtor.com: https://www.realtor.com/mortgage/tools/rent-or-buy-calculator/

San Francisco, California Housing Data. Retrieved from https://www.towncharts.com/California/Housing/San-Francisco-city-CA-Housing-data.html

San Francisco Planning Department. (2018). *Housing Needs and Trends Report and Housing Affordability Strategy.* Memo, San Francisco.

The New York Times. (2014, May 21). *Is It Better to Rent or Buy?* Retrieved from The New York Times: https://www.nytimes.com/interactive/2014/upshot/buy-rent-calculator.html

University of Michigan. (2018). *Practical Tools for Designing and Weighting Survey Samples | SpringerLink*. Retrieved from Springer Link: https://link.springer.com/book/10.1007/978-3-319-93632-1?noAccess=true&error=cookies_not_supported&code=c9d5c8b9-7970-4701-8fd6-33ff3acd876f#toc

Using stargazer to report regression output and descriptive statistics in R (n.d). Retrieved from https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

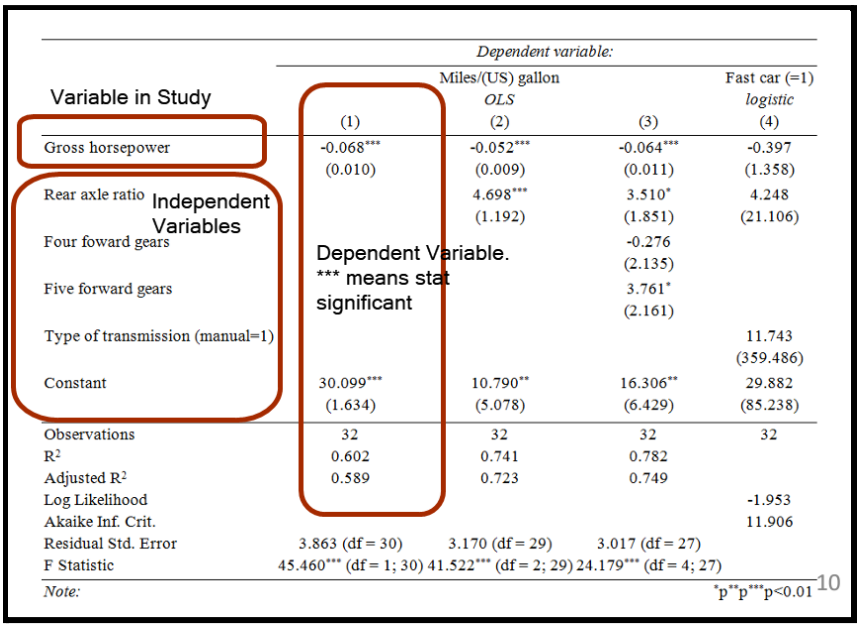
Walker, J. C. A. (2018). Chapter 14 Adding covariates to a linear model | Elements of Statistical Modeling for Experimental Biology. Retrieved from: https://www.middleprofessor.com/files/applied-biostatistics_bookdown/_book/adding-covariates-to-a-linear-model.html

Wang and Cheng (2020, July). An Overview of Study Design and Statistical Considerations https://journal.chestnet.org/action/showPdf?pii=S0012-3692%2820%2930462-1

Yates, W. (2021, Feb). The CPRA's Storage Limitation Requirement is Coming—Practical Tips for Shoring Up Your Record Retention Practices to Comply. https://www.jdsupra.com/legalnews/the-cpra-s-storage-limitation-9898179/

Appendix

Appendix A: Example model output.



Variable in Study	*Dependent variable:*			
	Miles/(US) gallon			Fast car (=1)
		OLS		*logistic*
	(1)	(2)	(3)	(4)
Gross horsepower	-0.068***	-0.052***	-0.064***	-0.397
	(0.010)	(0.009)	(0.011)	(1.358)
Rear axle ratio		4.698***	3.510*	4.248
		(1.192)	(1.851)	(21.106)
Four foward gears			-0.276	
			(2.135)	
Five forward gears			3.761*	
			(2.161)	
Type of transmission (manual=1)				11.743
				(359.486)
Constant	30.099***	10.790**	16.306**	29.882
	(1.634)	(5.078)	(6.429)	(85.238)
Observations	32	32	32	32
R^2	0.602	0.741	0.782	
Adjusted R^2	0.589	0.723	0.749	
Log Likelihood				-1.953
Akaike Inf. Crit.				11.906
Residual Std. Error	3.863 (df = 30)	3.170 (df = 29)	3.017 (df = 27)	
F Statistic	45.460*** (df = 1; 30)	41.522*** (df = 2; 29)	24.179*** (df = 4; 27)	
Note:				*p**p***p<0.01

Independent Variables

Dependent Variable. *** means stat significant

10