Work From Home and the Post-COVID U.S. Housing Market

STAT 222: Final Report

Andrea Padilla

May 15, 2023

1 Introduction

In the last few years the landscape of the workplace has changed drastically; COVID-19 caused shocks in policies related to work environments, and workplace culture has shifted as well. As young adults hoping to enter the workforce and looking to potentially move to different cities, we are personally interested in work-from-home prevalence, income, and the housing market in these areas. Work from home percentages have increased nationwide but on a smaller level, what areas have seen the most growth in the proportion of people working from home? Can we explain this shift using some combination of characteristics specific to the region? After some initial modeling, we shifted gears and further investigated what WFH percentages can tell us about the housing market. We ultimately believe this research is most useful for urban planners and policymakers who have the ability to create more housing and implement rent control.

In reading the literature we hoped to find ideas for potential predictors. For example, an article by Althoff, Eckert, et al.[1] discusses how the business sector shapes the economy of cities. They examine how much of the business sector is shifting to remote work and how this could affect local economy due to a loss of patronage in consumer services. They also note a positive relationship between population density and remote work. Another article, by Adam Ozimek[3], discusses how the shift towards remote work will affect population. The takeaway from this article is that the ability to work from home reduces the demand for housing in cities with a high cost of living. These two articles gave us several ideas on variables to include in our research, such as industries, housing costs, population, and income.

2 Data Description

Our data source is the American Communities Survey (ACS) [2], done by the Census Bureau, which collects demographic information annually on topics in-

cluding income, housing, and employment. About 1 in 32 households is selected to take part in the ACS, in comparison to the Census which collects data on every household. Although it only represents a sample of the population, it provides a detailed picture of American life from year to year.

The US Census Bureau has a publicly available API which enables any user to download only the data they're interested in, rather than the full survey. We decided to work with Metropolitan/Micropolitan Statistical Areas (MSAs) as our level of granularity because we are particularly interested in looking at cities. Because the ACS is such a rich source, it was not necessary to merge with other data sources but it did limit our choice of variables. We do not have any predictors that could indicate traffic congestion, apart from work from home (WFH) percentage. The options for industry are also a bit broad and don't line up exactly with our interests. Intuitively, we anticipate that areas with a large tech industry would have a high WFH percentage. However, tech doesn't cleanly fit into any of the categories provided by the ACS. We decided to include both the information sector and the finance sector as a compromise. We also included the agriculture sector as a comparison; we anticipate the agriculture sector will have a negative relationship with WFH percentage.

Because we are investigating how COVID changed the workplace, it was necessary to look at data from 2021 as well as previous years. The ACS was not conducted in 2020, but with that having been such an unpredictable year, we can still get a clear before and after of the WFH landscape and the housing market by using data from 2017-2019 and 2021.

3 EDA

Before doing any statistical analysis or modeling, we can take a look at the data in a few different visuals. Figure 1 shows the average rates of WFH in the 3 years before the pandemic in all MSAs. From this we can note which regions had high WFH rates before then and think about why that is. The 3 MSAs with highest WFH percentages pre-COVID are not big cities. They are all micropolitan areas, more specifically with populations between 15,000 and 25,000. Truckee and Clearlake are two popular California outdoor destinations both 100 miles from Sacramento. Faribault and Northfield are about 50 miles from Minneapolis. From this we can infer that before COVID-19, rural areas had the largest proportion of people working from home.

Now looking at Figure 2, the regions with leading WFH rates post-COVID are all major metropolitan areas. This tracks with the literature and our personal experiences during the first year of the pandemic. Many people living in cities were ordered to work remotely to slow the spread of COVID in areas that are densely populated. On the other hand, another narrative is that large amounts of people are leaving larger cities, moving to suburban areas, and working from home there. These two ideas aren't necessarily contradictory; both stories may be true.

We must also consider how our predictors correlate. A complete correlation

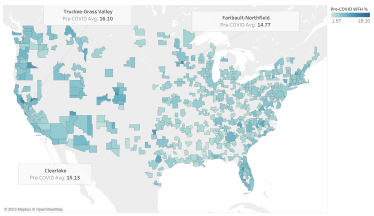

Figure 1: Percentage of population working from home pre-COVID

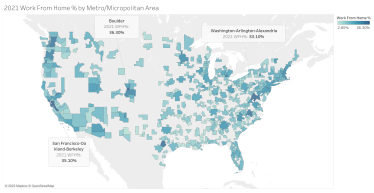

Figure 2: Percentage of population working from home in 2021

matrix would be difficult to include in a paper so I've included a smaller matrix in the Appendix, Figure 10, with our outcome variable and 4 predictors. We can see that there may be some colinearity with income and rent, which makes sense. Higher rent prices requires tenants with higher incomes. This is something to look out for once we get to modeling. We can try to remove one of the two, it's possible that having one is sufficient.

4 Initial Models

We began with a simple regression on all of our selected covariates: working population, information industry, agriculture industry, finance industry, median rent, median income, and unemployment percentage. Since the working population and the median income are large quantities, we put those on a log scale. We are not interested in the difference that a $1 salary increase makes; we can instead look at the effect of a percent change in income by taking the log. The results of this regression can be seen in Figure 11 in the Appendix. All predictors were significant and the adjusted R^2 was 0.64 but we were unsure whether that was truly indicative of goodness of fit or simply because there were so many predictors. We were also concerned about colinearity between median rent and median income. This model also uses the 2021 WFH percentage as its outcome variable which may not be what we want to explain. We are more interested in the change in proportion of people working from home, which is where we introduce a more robust outcome variable.

```
lm(formula = perc_diff ~ log(working_pop) + info_perc + med_rent +
    finance_perc + ag_perc + log(med_inc) + unemp_perc, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3199 -1.9281 -0.3442  1.6076 15.0981

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.186e+02  1.332e+01  -8.906  < 2e-16 ***
log(working_pop)  1.105e+00  1.701e-01   6.496 2.12e-10 ***
info_perc         1.299e+00  2.578e-01   5.037 6.76e-07 ***
med_rent          8.811e-05  9.421e-04   0.094 0.925529
finance_perc      3.008e-01  8.699e-02   3.458 0.000595 ***
ag_perc          -2.928e-01  6.346e-02  -4.614 5.12e-06 ***
log(med_inc)      9.929e+00  1.266e+00   7.845 2.96e-14 ***
unemp_perc        1.729e-01  8.100e-02   2.135 0.033295 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.102 on 467 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.6058,    Adjusted R-squared:  0.5999
F-statistic: 102.5 on 7 and 467 DF,  p-value: < 2.2e-16
```

Figure 3: Simple linear regression using percent difference as outcome

In order to control for the pretrend, we changed our outcome variable to be the change in WFH percentage, from the pre-COVID average. We took the average WFH percentage from 2017-2019 for each MSA and used that to center the data. Then we performed a regression with this new outcome variable.

We get similar results from this regression. The adjusted R^2 is a little lower, but still over 0.5. There is still the concern of colinearity, and median rent is the only predictor that was not significant which we found to be odd.

At this point we tried a few machine learning techniques like decision trees and bagging. We ultimately did not continue with those but from bagging we saw that median rent was the most important feature, which contradicted what we saw in our regression above. We wanted to look further into this variable and the housing market in general in order to obtain some result that may be more actionable.

5 PCA

After our simple initial models we decided to incorporate more data related to housing in an effort to make that aspect of our work more robust. This data was readily available from the ACS API and it was simple to add on to our existing data set. A few of these additional variables include median SMOC (selected monthly owner costs), average household size, and percentage of population paying more than 35% of income on rent. We then had about 12 variables related to housing which we decided to reduce through principal component analysis. By using PCA we hoped to have a more complete picture of the housing market of an MSA. The loadings of the first and second principal components are in figure 4.

The first component relates mainly to rent, housing value, and proportion of income spent on rent. The second principal component deals more with the costs associated with owning a home. A biplot and scree plot are available in the appendix that explore the results from the PCA as well. We then created an index using the projection of our housing data onto principal component one

Variable	Comp.1	Comp.2
rentedhousing_perc	0.26	0.01
renter_householdsize	0.22	0.17
mortgaged_housing	0.34	-0.54
mortgaged_SMOC	0.32	-0.54
rent_lt15perc	-0.36	-0.24
rent_15to20perc	-0.22	-0.25
rent_20to25perc	-0.01	-0.27
rent_25to30perc	0.03	-0.02
rent_30to35perc	0.07	-0.04
rent_gt35perc	0.32	0.44
medhousing_value	0.41	0.03
med_rent	0.47	0.03

Figure 4: Loadings of PC1 and PC2

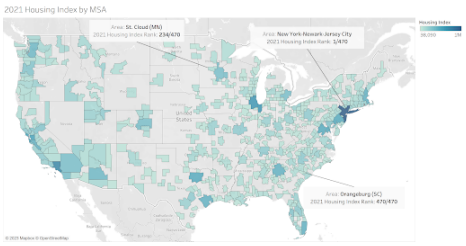

Figure 5: Map of Housing Index

to give us one metric to explain the variation in housing for all MSAs. We can visualize this index in a few ways. In figure 5 we can see how MSAs across the country perform on this metric. For example, the MSA containing New York City ranks highest, which tells us that this metric captures disproportionately high rent costs relative to income, high demand, and a competitive housing market.

This is also exemplified by the plot in figure 6. There's certainly a positive relationship between the housing index and WFH percentage, though it does not look linear. This takes us to our final models, where we hope to model the housing index based on WFH percentage and other predictors.

6 Final Models

We have two potential regression models using the housing index created by PC1. The first model has this index as its outcome variable and the other non-housing related features as predictors. Results of this regression are shown in Figure 8. This model gave us an adjusted R-squared of 0.67 which shows a high amount of correlation between the index and the covariates. Most covariates were significant, though we were confused by those that weren't, or that weren't as significant as we expected. We expected WFH percentage and percent working in finance to have higher significance as we saw in previous models.

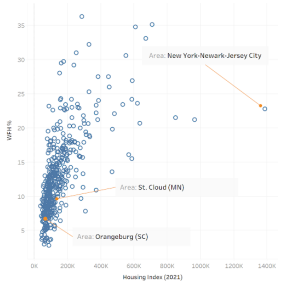

Figure 6: Housing Index and WFH Percentage

```
Call:
lm(formula = PC1_index ~ log(working_pop) + info_perc + finance_perc +
    ag_perc + log(med_inc) + unemp_perc + home_perc, data = acs_full_index)

Residuals:
    Min      1Q  Median      3Q     Max
-125937  -42303   -6769   28054  846026

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2657906.6   183098.6 -14.516  < 2e-16 ***
log(working_pop)    64968.7     2689.9  24.153  < 2e-16 ***
info_perc           23124.3     4021.0   5.751 1.20e-08 ***
finance_perc        -2781.7     1454.9  -1.912   0.0562 .
ag_perc              1384.5      929.0   1.490   0.1365
log(med_inc)       179674.8    17649.0  10.180  < 2e-16 ***
unemp_perc           5959.7     1358.7   4.386 1.28e-05 ***
home_perc             927.3      526.3   1.762   0.0784 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 72180 on 953 degrees of freedom
Multiple R-squared:  0.6732,    Adjusted R-squared:  0.6708
F-statistic: 280.4 on 7 and 953 DF,  p-value: < 2.2e-16
```

Figure 7: Regression on Housing Index

Additionally, the coefficient for finance was negative which was unexpected.

Since we were not fully satisfied with this model we decided to also model the change in the housing index using the same covariates, but only from the year 2019, since we are looking to use this model for prediction.

The results from this regression tell us that the percentage of people working from home and the median income in 2019 positively contribute to the increase in housing costs during COVID. However, the adjusted R-squared for this model is much lower than any previous model. This puts into question the ability of this model to accurately predict changes in housing costs.

7 Discussion

There is definitely room for improvement in our final models. Something we realized immediately after presenting our work is that we did not scale our data before performing our principal component analysis. Given the differences in unit and scale between some of our housing variables, (some are percentages, others are costs associated with housing,) this may have made a large difference. The reasoning there may have been that we would lose interpretability, but we lose interpretability through PCA regardless.

```
Call:
lm(formula = PC1_diff ~ log(working_pop_2019) + info_perc_2019 +
    finance_perc_2019 + ag_perc_2019 + log(med_inc_2019) + unemp_perc_2019 +
    home_perc_2019, data = merged_data)

Residuals:
    Min      1Q  Median      3Q     Max
-171761   -6763   -1321    5537  177951

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -201301.1    91731.7  -2.194  0.02865 *
log(working_pop_2019)   -2635.2     1347.6  -1.956  0.05107 .
info_perc_2019            556.8     1952.6   0.285  0.77565
finance_perc_2019        -741.6      682.0  -1.087  0.27740
ag_perc_2019             -243.3      426.7  -0.570  0.56874
log(med_inc_2019)       21750.6     8893.7   2.446  0.01480 *
unemp_perc_2019           535.8      804.8   0.666  0.50585
home_perc_2019           1652.3      599.0   2.759  0.00601 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26090 on 509 degrees of freedom
Multiple R-squared:  0.04503,   Adjusted R-squared:  0.03189
F-statistic: 3.428 on 7 and 509 DF,  p-value: 0.001358
```

Figure 8: Regression on Change in Housing Index

The results from our last model leave much to be desired. It's possible they would be improved if we redid the PCA with scaled data. There is also a gap in reasoning for interpreting our final model. The data for 2019 was used in order to predict future housing costs, but we have seen that that cannot be done particularly from 2019 to 2020 due to sudden phenomena like COVID. Given the adjusted R-squared values in these last two models, I am inclined to prefer the regression on the housing index itself. I think it holds more predictive power, and we can still use it to think about the change in housing costs.

8 Conclusion

The housing index created through PCA gives us a more robust look at the housing market than only having a few housing related variables. Given more time we would love to continue trying to model it using the rest of our data. Overall we have seen the relationship and influence that COVID work from home policies have had on the housing market. Using our index, we can recommend that policymakers implement rent control for MSAs that are scoring higher than some threshold. What that threshold is exactly is a difficult question and would require more thought and care, but it could be anything above the third quartile or even the median to be conservative. Similarly we can recommend that policymakers and urban planners create more housing in these areas to meet demand.

9 Appendix

All code for the work referenced in this paper can be found on the team's GitHub repository: https://github.berkeley.edu/paige-park/222.git

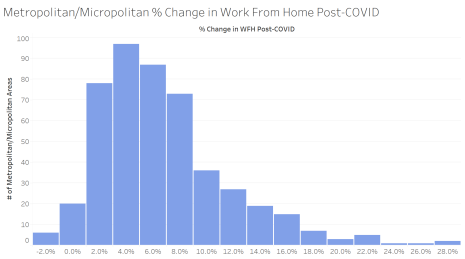

Figure 9: Percent change in WFH population

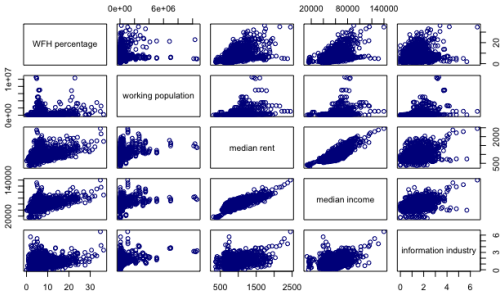

Figure 10: Correlation matrix

```
lm(formula = home_perc ~ log(working_pop) + info_perc + med_rent +
    finance_perc + ag_perc + log(med_inc) + unemp_perc + pre_avg,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2934 -1.9289 -0.3348  1.6123 15.0952

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.185e+02  1.334e+01  -8.882  < 2e-16 ***
log(working_pop)  1.112e+00  1.741e-01   6.386 4.12e-10 ***
info_perc       1.300e+00  2.581e-01   5.035 6.85e-07 ***
med_rent        2.503e-05  1.001e-03   0.025 0.980056
finance_perc    2.981e-01  8.826e-02   3.377 0.000794 ***
ag_perc        -2.920e-01  6.364e-02  -4.589 5.73e-06 ***
log(med_inc)    9.912e+00  1.270e+00   7.804 3.98e-14 ***
unemp_perc      1.726e-01  8.110e-02   2.128 0.033829 *
pre_avg         1.017e+00  8.854e-02  11.483  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.105 on 466 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.7202,    Adjusted R-squared:  0.7154
F-statistic: 149.9 on 8 and 466 DF,  p-value: < 2.2e-16
```

Figure 11: Simple regression with all predictors

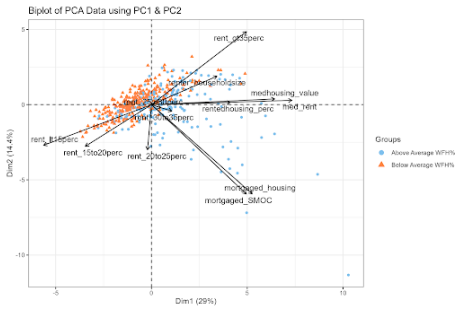

Figure 12: Biplot of PC1 and PC2 on housing variables

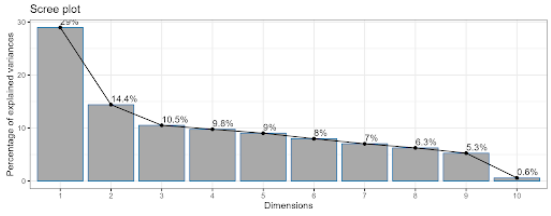

Figure 13: Scree plot from PCA

References

[1] Lukas Althoff et al. "The Geography of Remote Work". In: *Regional Science and Urban Economics* 93 (2022), p. 103770. ISSN: 0166-0462. DOI: https://doi.org/10.1016/j.regsciurbeco.2022.103770. URL: https://www.sciencedirect.com/science/article/pii/S0166046222000011.

[2] US Census Bureau. *American Community Survey (ACS)*. Mar. 2023. URL: https://www.census.gov/programs-surveys/acs.

[3] Adam Ozemik. *How remote work is shifting population growth across the U.S.* July 2022. URL: https://eig.org/how-remote-work-is-shifting-population-growth-across-the-u-s/.