Data Project #1
QAMO 4700-001

Executive Summary:

Our team was tasked with the job of aiding a web-based learning platform called Creative Share by conducting thorough market research on two demographics, Texas and California. The goal of Creative Share in these two areas was to introduce a new business model that includes live classes and workshops. In order to gauge the likeliness of someone purchasing a live class or workshop we had to quantify a given person's level of creativity. It was predetermined that a person who is considered a creative individual is more likely to purchase a live class or workshop from Creative Share. We were then interested in discovering relationships between how a person's likelihood of being creative is impacted by their gender, age, race, and employment status. The goal was to see if there were meaningful relationships between these independent variables listed and someone's level of creativity and to define those differences while also comparing the two different markets of Texas and California. Additionally, we wanted to test if any other explanatory variables in our data could help us better identify a less biased model. After several regression tests and a lot of trial and error, we discovered interesting conclusions from the data and our research. We were able to determine that was is no significant relationship between a person's creativity level and their age or employment status. Rather, we found that a person's creativity level is better determined by looking at a person's gender, race, number of children, as well as whether or not they are in the labor force. These four independent variables above were the explanatory variables that we decided on for our final regression model.

Data:

To conduct our research, we used data sourced from the Integrated Public Use Microdata Series or IPUMS. The data we collected was from February 2018 and February 2020 and included over twenty variables for each sample. Before we began any sort of regression testing, we wanted to limit our data to only the demographic of people that we are interested in. From the information about Creative Share provided, we know that people who tend to take live classes on the web-based platform tend to be highly educated. Now, generally when people say they are highly educated they mean a master's degree or more but for our market research, we determined that our highly educated people in the sample were people who have a 4-year Bachelor's degree or more. The thought behind this is people who complete a Bachelor's degree are already much more educated than the average person and we do not want to exclude them from our sample. Additionally, we wanted to put bounds on the ages of our data samples. Our thought process was that a teenager is very unlikely to be purchasing online art classes because they are already in school and likely have access to art classes for free. Teenagers would also not have the high-level education that most consumers of the courses tend to have. Additionally, we do not anticipate elderly people purchasing online art classes through Creative Share due to technological barriers and the lack of interest in learning something new from a high-level course. The result of our thought process was limiting the age range from 20 years old to 54 years old. We chose the age 54 because we thought that 54 is an age when people are at the back end of their career or approaching retirement, and are less likely to be learning new things through online courses. This age range clearly defines prime-age adults who would be interested in learning new topics. The age range is demonstrated in the histogram above (Figure 1) with the red lines being our cutoffs. Through this visualization, it can be seen why we needed to limit the data to these ranges to weed out those to who the course would not apply. The final change to the data we made was eliminating all of the data points for

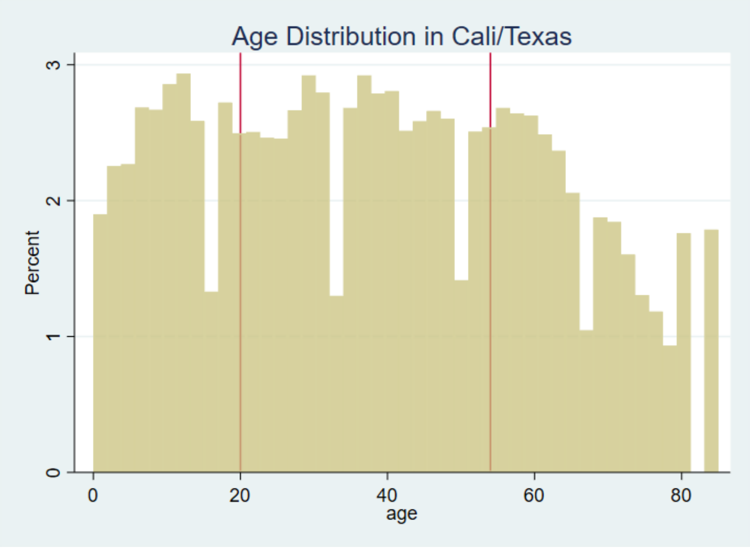


Figure 1: Histogram of Age Distribution in California and Texas

people who showed no sign of being creative person. We did this because what we are focused on is addressing the variables that determine if a person is creative and the people who show no signs of creativity are not useful in our market research. To do this, we dropped all of the people who (1) do not write stories, poems, or plays, (2) do not paint, draw, sculpt, or make prints, (3) do not take photographs, or make movies or videos, and (4) do not weave, crochet, quilt, or sew. To further understand the data we were working with, we wanted to analyze the race demographics of the sample to see how we should handle the testing. The best way for us to visualize this was to have Stata output a labeled pie chart (Figure 2). Immediately we noticed how the sample was dominated by white people, nearly 70%, and there was not a lot of diversity in our sample. This could be an issue with the sampling process or a coincidence. This allowed us to create a race variable that accounted for if an individual was white or not in order for us to acknowledge this racial distribution in our further regression testing.

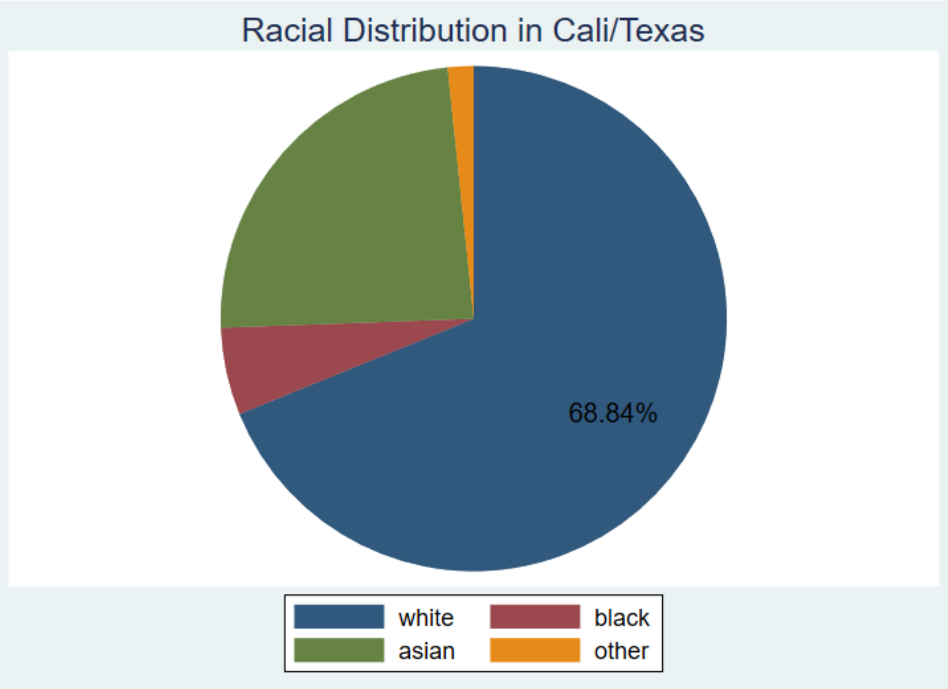


Figure 2: Pie Chart of Racial Distribution in California and Texas

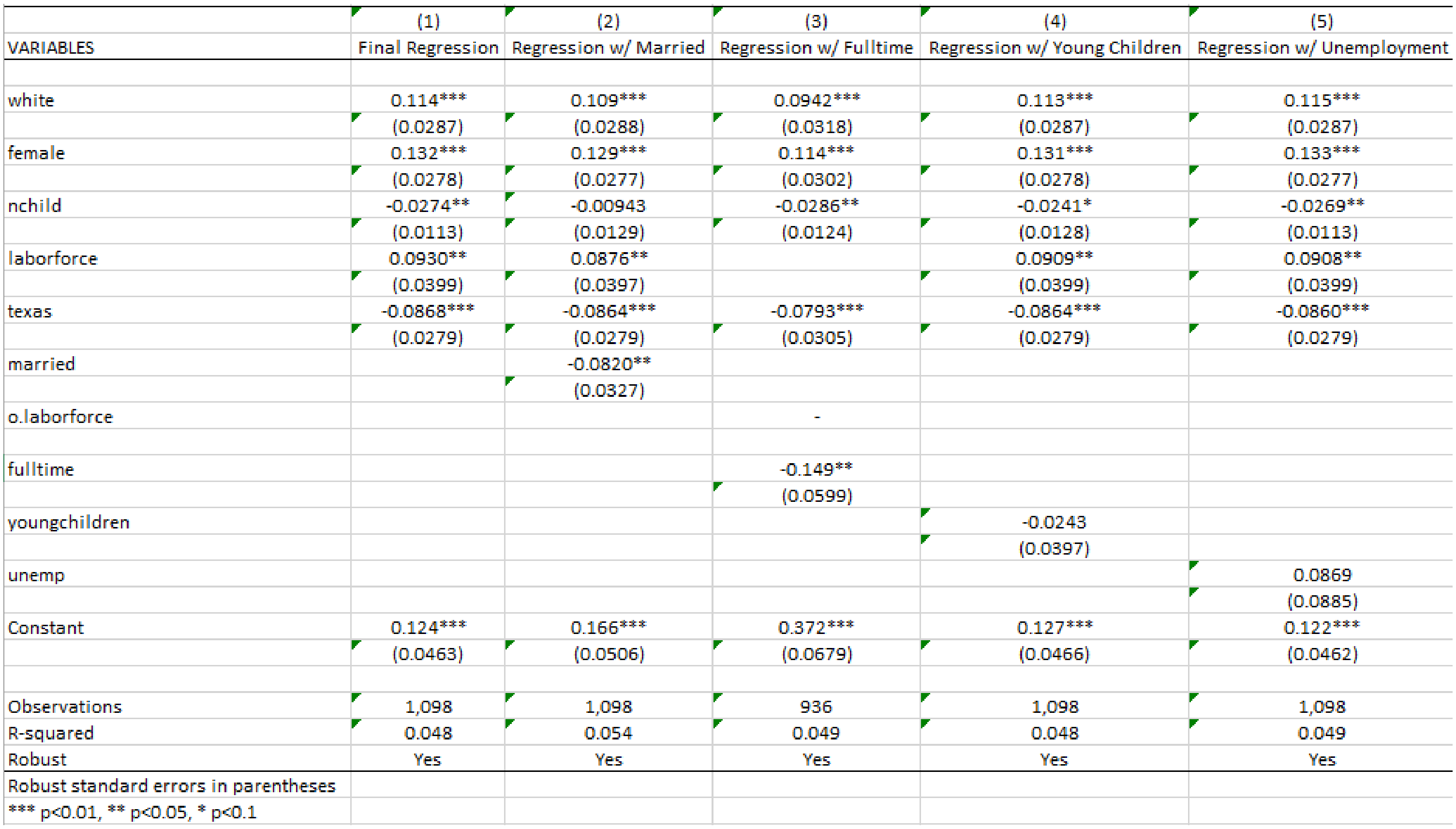
The data manipulation we did before starting our market research helped us better understand the true relationships at hand and determine the variables that have no relationship with a person's level of creativity. With these determinations, we were able to build out our final regression model.

Model:

Our final regression model consisted of five separate variables that all were factors in how we determined if an individual is considered creative or not. The final regression is laid out below (also shown in Figure 3):

Creative **=** 0.124 + 0.114(White) + 0.132(Female) - 0.027(#ofChildren) + 0.093(Labor Force) - 0.087(Texas)

While this does not convey much other than lengthy formula at a first glance, this model allows for the ability to get a real estimate of whether or not a person is considered creative with just a few variables. With our initial criteria of selecting only individuals that are between our selected age range, educational history, and feedback on their creativity status, we can explain what this regression conveys. Beginning with the dependent variable creative, this variable is considered a dummy variable. This means that creativity either equals 0 or 1 depending on the feedback from the individual. Creativity would be set to 1 if an individual indulges in any one of the creative activities that were listed earlier (painting, weaving, writing, etc.). This regression model now uses other independent variables to describe how changing one of these variables would change the likelihood that creative is set equal to 1. Next, we will define the constant term in our equation, this means that a person is predicted to have a 12.4% chance to be considered creative with all of our other variables equaling 0. This would mean that this specific person would be a non-white male with no children who lives in California and is out of the labor force. With this regression, we can predict that a specific individual's likelihood of being creative which can estimate to be 12.4%. The rest of the variables are more self-explanatory. The independent variable white is a dummy variable that predicts, if an individual is white, they are 11.4 percentage points more likely to be considered creative while holding all other things constant. Percentage points are a way to indicate the amount of change on a number rather than percent which defines the amount of change from a previous number. Percentage points give us a streamlined process to explain how a change in a variable impacts the likelihood of a person being creative, so we will use these to describe our variables. The female, labor force and Texas variables all act similarly to how the dummy variable white works. These three variables are all dummy variables as well and have corresponding coefficients that when they equal one, affect the likelihood that a person is creative. For females, this variable is equal to one when an individual is a female; and zero if they are a male. For the labor force, it equals 1 when an individual is defined to be in the labor force and zero otherwise. The Texas dummy variable is a bit different as when this variable is equal to zero, this is stating that the individual lives in California. This is because the overall goal of this model is to define differences between people in these two states, therefore this variable encapsulates that. The last variable is the # of children and this is not a dummy variable, but one that behaves as normal variables do. As the number of children an individual has in their home increases, so does the variable. Therefore, the more children a person has the less likely they are to be a creative person according to our model. Through various tests and a lot of trial and error, we determined that this model we created functioned in a way that could be of use to Creative Share in determining potential clients for their online courses and workshops.



	(1)	(2)	(3)	(4)	(5)
VARIABLES	Final Regression	Regression w/ Married	Regression w/ Fulltime	Regression w/ Young Children	Regression w/ Unemployment
white	0.114***	0.109***	0.0942***	0.113***	0.115***
	(0.0287)	(0.0288)	(0.0318)	(0.0287)	(0.0287)
female	0.132***	0.129***	0.114***	0.131***	0.133***
	(0.0278)	(0.0277)	(0.0302)	(0.0278)	(0.0277)
nchild	-0.0274**	-0.00943	-0.0286**	-0.0241*	-0.0269**
	(0.0113)	(0.0129)	(0.0124)	(0.0128)	(0.0113)
laborforce	0.0930**	0.0876**		0.0909**	0.0908**
	(0.0399)	(0.0397)		(0.0399)	(0.0399)
texas	-0.0868***	-0.0864***	-0.0793***	-0.0864***	-0.0860***
	(0.0279)	(0.0279)	(0.0305)	(0.0279)	(0.0279)
married		-0.0820**			
		(0.0327)			
o.laborforce			-		
fulltime			-0.149**		
			(0.0599)		
youngchildren				-0.0243	
				(0.0397)	
unemp					0.0869
					(0.0885)
Constant	0.124***	0.166***	0.372***	0.127***	0.122***
	(0.0463)	(0.0506)	(0.0679)	(0.0466)	(0.0462)
Observations	1,098	1,098	936	1,098	1,098
R-squared	0.048	0.054	0.049	0.048	0.049
Robust	Yes	Yes	Yes	Yes	Yes
Robust standard errors in parentheses					
*** p<0.01, ** p<0.05, * p<0.1					

Figure 3: Spreadsheet of regressions throughout the research process

Analysis:

Throughout our research process, we went through countless different variables and regressions to finalize our last model. We wanted to include as many variables as we could to make our regression as refined as possible, but some of these variables would make our regression not statistically significant. If a regression is not statistically significant then we would not be able to make predictions about an individual with said regression. Some variables that we tried that ended up hurting our regression were dummy variables for being married, working full-time, having young children, and being unemployed (all of these variables are shown in their respective regressions in Figure 3). Each of these variables ended up hurting some other piece of our final regression, or just did not contribute enough to the overall model so they were omitted. With the variable married, when added to our regression, this makes our variable of the number of children not statistically significant. We concluded that we would rather account for children in a person’s life rather than if they were married because it was easier to quantify the impact of kids than marriage. For example, when an individual has a child or brings another child under their roof, they tend to spend the majority of their time focusing on the well-being and upbringing of their children. When a person gets married, it is not defined by where their time goes. There could be a situation where being married means that this individual spends their free time with their new spouse and building a relationship together, or there could be a marriage where the new spouse lightens the load for the surveyed individual and they have more time and energy to pursue their creativity. With this uncertainty, we decided that having children would better match our fit for the regression and it helped keep the regression more statistically significant as a whole to not include the married variable.

The next variables that we tried to introduce into our model were the unemployed and full-time work dummy variables. The full-time variable accounts for only those who are in the labor force and is set to 1 for full-time and 0 for part-time. When adding this into our regression, it ended up omitting our laborforce variable altogether while also reducing the overall significance of our regression. We came to the conclusion that knowing if an individual is in the labor force would be much more important a distinction than if they were working part-time or full-time. Similarly, the unemployed variable also ended up reducing the significance of our regression model, specifically on the white and female dummy variables. This could be because unemployment would be accounted for in our laborforce variable, or in a multitude of different ways. In the same fashion as with the full-time variable, we chose the route of using a variable that better fits our regression and described a more important distinction over the unemployed variable. The young children variable follows the same path of hurting the overall significance and just being too specific of a variable for it to be placed in the regression over another variable. Unfortunately, these 4 separate variables could not make our final regression. They had logical reasons for why we believed they would improve our overall analysis, but due to reducing the overall significance of our regression and specific negative implications for each separate variable, we could not use these independent variables in our final regression.

After eliminating the variables that were hurting our overall regression, we ended up with our final model. This model is built with only five independent variables that all have an impact on the likelihood that a person is considered creative. All five of our independent variables had a high statistical significance with only two of the variables having a p-value greater than 1%. A p-value is how you quantify the statistical significance of a particular variable and for our case, and most regression cases, a p-value of 5% is desired, which all of our variables easily achieved. Our constant term was also able to have a p-value of below 1% which meant that we could confidently say that holding all things equal, an individual is predicted to start with a 12.4% likelihood of being considered creative. The other variables that positively impacted the chances of being considered creative were if the individual was a female, if they were white, and if they were in the labor force. Being a female had the largest magnitude of impact followed closely behind by the race of the individual. Each of these variables impacted the overall likelihood by close to or over 10 percentage points which is a massive increase. For our other two variables Texas and the number of children, harmed creativity. The number of children variable is predicted to decrease the likelihood that a person is creative by about 2.5 percentage points for every child still living in the same home as the individual who was surveyed. The Texas variable is a more interesting one. At a first glance, it seems as though if a person lives in Texas they are predicted to be 8.5 percentage points less likely to be considered a creative person. While this is true, there is some confusion about if the individual lives in California. For a California resident, the Texas variable is set to zero which makes sense, but this does not mean that being a resident in California has zero impact on the likelihood of creativeness. All this means is that living in Texas has an 8.5 percentage point decrease in the likelihood of creativeness. The impact of living in California is accounted for in the constant term. This regression was able to create a modeling system between 5 variables that lets us see the likelihood of a given person being creative given a few different parameters.

Conclusion:

The Creative Share company could find great use in our model and market research to better launch their new business model that includes live classes and workshops. Firstly, we have helped narrow down their target audience when it comes to advertising or marketing campaigns. Additionally, we are confident that the variables we chose for our final model are variables that have a direct relationship with how creative a person is. To aid Creative Share in having a successful launch of their new business model, they should begin with a trial launch focussed on one key demographic: white women who live in California, are in the labor force and have no children. We found that this type of person tends to be highly creative. If this trial launch works well and new customers are acquired, then launching fully in California could be a great next step. The reason California would be better than Texas is that on average, people in California tend to be more creative than people in Texas (- 0.087(Texas)). Things that Creative Share should not focus on when choosing who to market/target their new live courses and workshops are people's unemployment status, marital status, and if they work full-time. We found these variables to not have any relationship with a person's creativity level and therefore have no relationship with whether or not they will purchase the live classes. Creative Share is predicted to achieve the best outcomes by following our recommendations for their target audience.

Contributions:

The two of us spent a lot of time in the library together doing a ton of regression testing and trial and error to get our model where we wanted it. We then spent some time editing the do file to get it to a solid format and went back and forth with each other there. After that, we communicated closely to finish writing the memo. We would each write different sections and have the other partner come through and edit/add pieces to our analysis throughout the memo. Finally, we finished with one final edit and run through to make sure that the memo was up to the standard we like. I would say we split the work up well.