Capstone 2 Milestone Report
Sam Fisher

Disclaimer: This project uses public data from Pew Research Center's "Information Engaged and Information Wary" survey.

Pew Research Center bears no responsibility for interpretations presented or conclusions reached based on analysis of the data.

Problem setup

Our client, edugames.org, is a fictional but plausible company that is interested in creating online mathematical puzzle-solving games for seniors. They have traditionally created sudoku puzzles and other physical media but are interested in creating a new offering for seniors in the online space. Their hope is that this new product will offer an enjoyable avenue for intellectual engagement and growth during one's professional retirement.

The current thinking is that the game could be a mobile app as well as a web application, with a mobile first design and social media integration. They are curious if a mobile-first online game can find traction with a senior demographic. They are also curious if this may vary amongst segments of the senior population. In order to study this, they want to gain a general sense of the level of eagerness//skepticism that senior citizens hold towards online media.

Can they expect irrational exuberance, or should they anticipate a cautiously optimistic audience? We can expect growth mindset to correlate with the desire to engage with an educational product. Can we find a demographic subset that demonstrates growth mindset as well as a high degree of willingness to engage with technology?

The Plan

Summarizing Knowns:

We will begin by conveying the implications to the problem already explored in Pew's 2017 report on technology usage amongst seniors. These insights will be included in Part 1 of the exploratory analysis.

Data Cleaning and Validation:

We will describe the dataset's structure in this secion of the report. We will then make decisions regarding null values and outliers in responses and the relevance of nuanced information. Next, we create a dictionary that maps the numerical values in the DataFrame to each response's qualitative meaning. Finally, we will create a second dictionary to map each column (survey question) name to an appropriate plot title. Phew! 100+ categorical variables is no joke.

Exploratory Data Analysis Part 1:

We will then conduct EDA phase 1 by plotting the distributions of survey responses for senior respondents and the general population, looking for strong trends to gain directional insight into obstacles and opportunities in the senior market.

Feature Engineering & Regression Analysis (Plan A):

We will then conduct a Feature Engineering and Correlation Analysis Phase. The goal will be to prepare the data for a regression analysis between demographic covariates and the idealized "goodness of customer fit" score. The score will be based on a priori assumptions of what survey responses indicate good fit to the product. This will be a weighted sum of subgroupings of the survey response variables.

The subgrouping will be single variables that encapsulate a single qualitative dimension of the customer-fit picture, such as like "growth mindset", "openness to information", etc. These will be initially scaled to be of equal range. I will try to avoid over-weighting strongly correlated contributors to the final score by looking at the correlation and pairplot matrices between these variables.

With the target variable engineering complete, I will then look at the cross-correlation between the demographic covariates and the correlation between demographic variables and the target. I will try to find a near-orthogonal basis for the feature space that is interpretable as well as predictive. I will favor demographic features that are more easily used and understood in the marketing process.

The feature engineering and correlation analysis is sure to lead to insights that will be added to the exploratory data analysis section as "EDA Part 2".

Regression Modeling: With the features and target variable designed, I will experiment with regression models to predict for the target- favoring interpretability. I will begin with OLS linear regression, but may explore Lasso or Ridge models. I will particularly bare Lasso in mind while confirming the feature selection. Boosting predictive strength of the model beyond a reasonable baseline is mostly an unnecessary endeavor in this regard, since our primary interest is directional insight in how to market the product and our target vector is a contrived construct of our own devising.

Clustering or Matrix Factorization Analysis (Plan B)

If Plan A is unsuccessful, then we will use the results of the cluster analysis from Pew's initial study of the entire population to guide an unsupervised analysis of the senior population - focusing on finding distinct behavioral sub-groups. I can try:

- Finding the intrinsic dimensionality of the survey response subgroups (from supervised section) with PCA

- Looking for behavioral sub-groups of seniors with Non-Negative Matrix Factorization of the survey response matrix.
- Looking for sub-groups of seniors with cluster analysis, beginning with K-Means and Agglomerative Clustering and projecting the results of this into 2d visualizations with t-SNE or PCA. Apply various clustering methods, then move to most promising results.
  - I will use known methods to choose the best k or neighborhood size for each clustering model.
  - I will look for consensus across a few clustering methods to validate the results.
- I will then plot responses per cluster and begin looking for qualitative explanations and demographic correlates to the clustering, focusing on those explanations most relevant to receptivity to an online game.

Summary and Recommendations:

- I will frame the findings in terms of their relevance to the game, using verbal, numerical, and graphical arguments from modeling and EDA to explore the implications for marketing to certain sub-groups of seniors.
- I will summarize the findings in a concrete set of Go/No Go recommendations for marketing the product to various groups.

Exploratory Data Analysis

An a priori ideal product-customer fit.

Because this survey doesn't tell us anything about who will buy our product directly, we'll need to make our own assumptions about this. The ultimate goal is to find demographic covariates to our pre-supposed ideal customer that we can use as an initial template in our targeted marketing efforts. These ideas can subsequently be modified as assumptions are proven or disproven empirically. The advantage here is that public data is free- the only expense being the time to analyze it. This survey also gives us access to broader qualitative insight than we might expect from click data.

The assumptions we'll make are that:

- Our ideal customer exemplifies openness to experience and growth mindset
  - We can suppose this because these customers are likely to seek out educational experiences
  - These customers will also be willing to try a variety of product that they are new to.

- Our ideal customer has time and attention to spend playing the game and learning. They do not see themselves as highly busy and stressed.
  - The game is quite involved and requires a time commitment.
  - Our hope is that retirement brings with it plenty of time and attention as well as a desire to fill that time and attention with rewarding experiences.

- Our ideal customer is already comfortable using the internet and technology and likely engages with them frequently.
  - Using the tools needed to access the game is a pre-requisite to engaging with it.
  - It's harder to build a new habit than to build on an existing one.

In this case, looking at the data is unlikely to modify our assumptions, but it will allow us to *make some concrete predictions using them.*

Technology Adoption

Let's begin by distilling relevant facts from [Pew's Report on Senior Tech Adoption](). See the linked report for greater detail on the insights below.

According to Pew's Report,

Seniors make up a substantial and growing portion of the population.

- Seniors (65+) make up 15% of the population- about 46 million people.
- The share of seniors is projected to increase to 22% by 2025.

Tech adoption for seniors is growing in step with the general population, especially for younger, wealthier, and college educated seniors.

- Growth in internet, broadband, smartphone, tablet, and social media adoption amongst seniors closely mirrors the high rate of growth in the general population.
- The total rate of adoption of these technologies varies is a little less than 2/3 that of the general population.
- Younger, wealthier, and college educated seniors show higher rates of tech adoption in terms of internet, broadband, smartphone, tablet, and social media adoption.
- 40% of seniors own smart phones
- 25% of seniors play online games.

Seniors who do adopt internet and device usage have high levels of engagement.

- 75% of internet-using seniors go online daily.
- 76% of smart-phone owning seniors use the internet multiple times per day.

Information Constraints & Confidence

About 3 of 5 seniors report needing occasional or frequent help getting information online, compared with about 2 of 5 people in the general population.
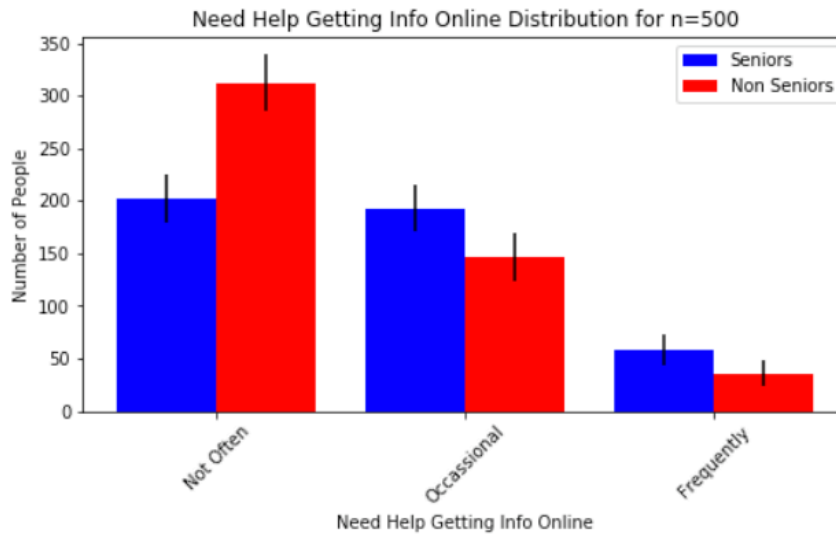

Conversely, seniors were fairly evenly divided on whether training on identifying reliable information would be helpful, perhaps demonstrating that some seniors are comfortable deciding which sources to believe.
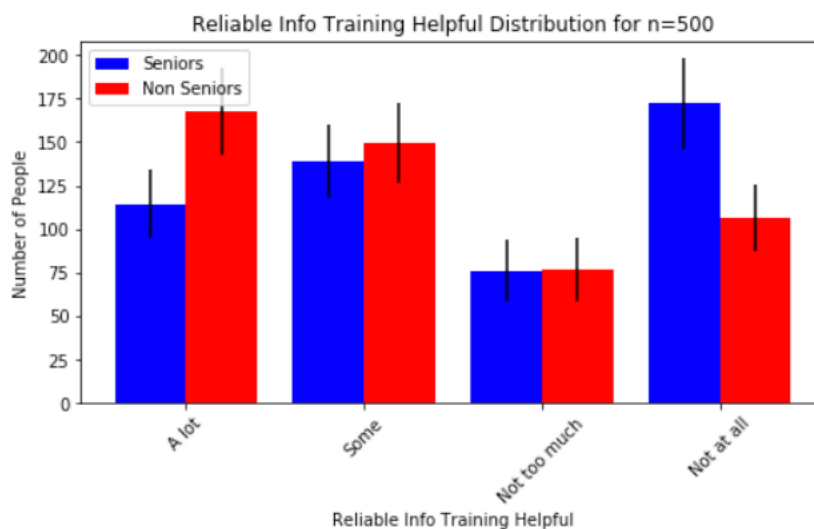
From this we can infer that there are significant differences in the distributions of individuals who exemplify "information" and "technology" comfort. Information comfort is the degree to which one is able to confidently decide upon trustworthy sources. Technological comfort is the degree to which one is confident in using devices. I would argue that technological comfort is more important to the customer fit for an educational

game product, but that a high degree of information discomfort may also preclude interested customers from participating.

Lastly, seniors are slightly more prone than non-seniors to view the current high level of information availability as stressful rather than useful. However, 3/5 seniors view it as useful- painting a promising picture overall for increasing senior engagement with digital products such as online educational games.
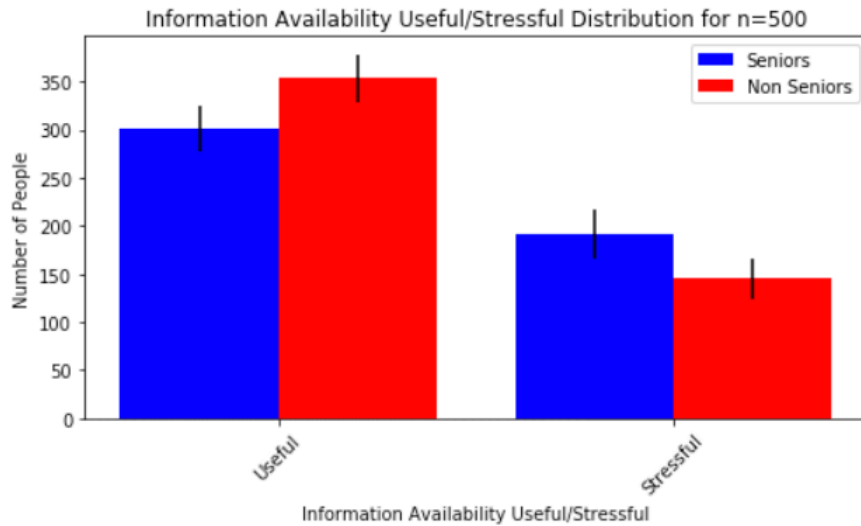
Time and Attention

We'll begin to make some comparisons between seniors and non seniors by exploring the survey dataset.

One of the best arguments for creating an educational product for seniors is the fact that they have free time, because they are retired. Here's a quick sanity check:
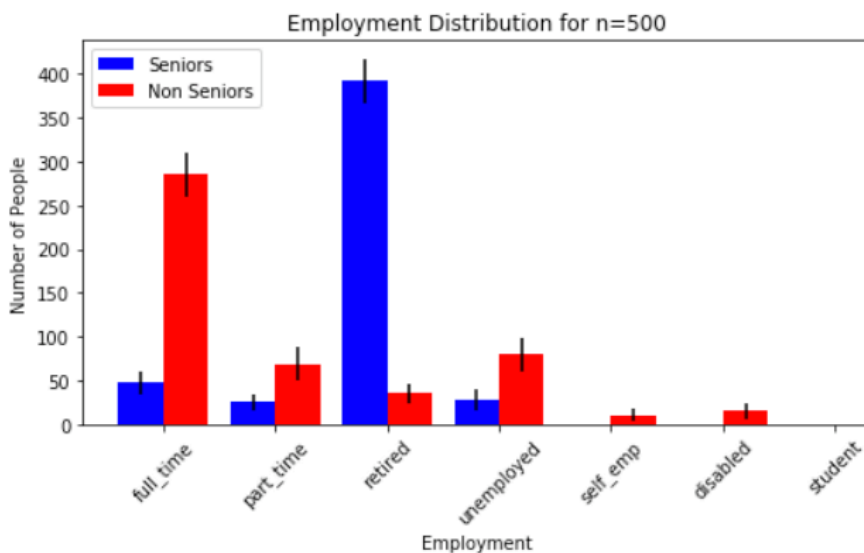
But does retirement really mean that seniors have time and attention for a new activity in their life? Let's try to answer this indirectly. When asked whether they the idea that their "fast pace of life was harmful to their health", seniors were likely to disagree.
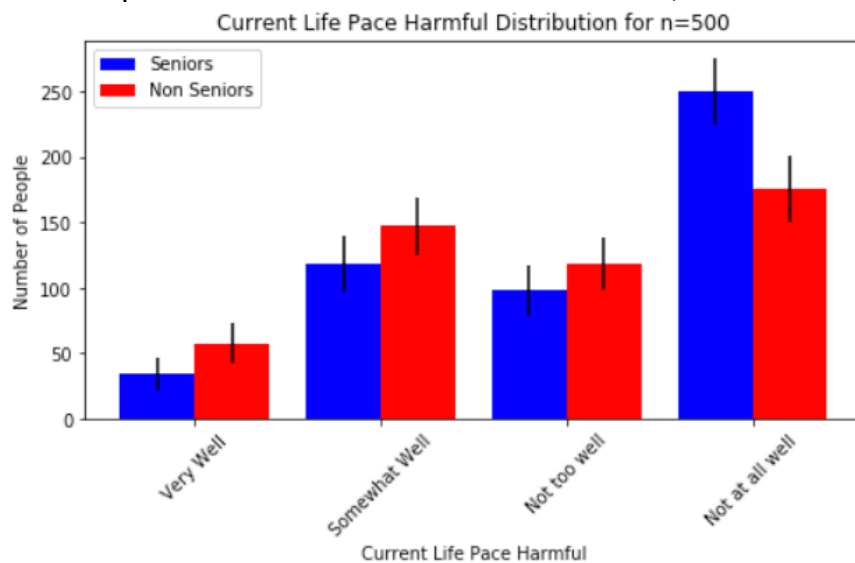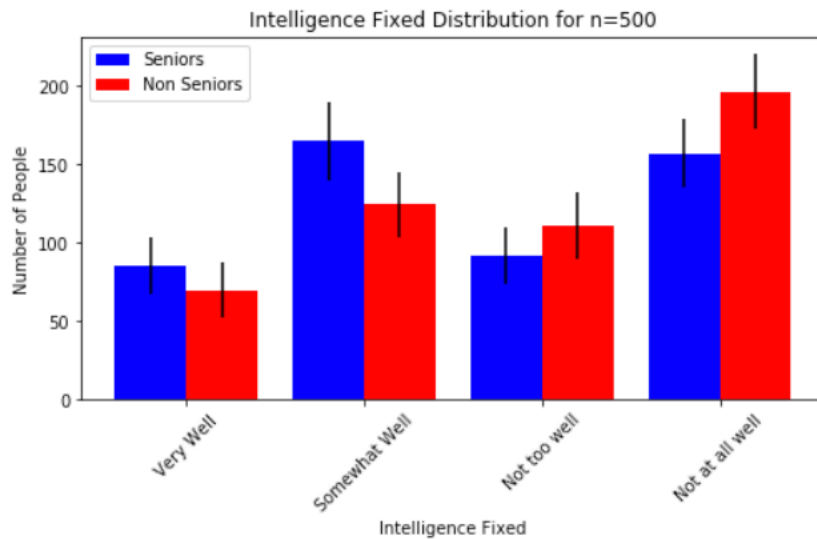
Seniors were also less likely to report that the idea that they were "usually doing two things at once" describes their views, which provides a second measure of "busy-ness".


Multi-tasker Distribution for n=500
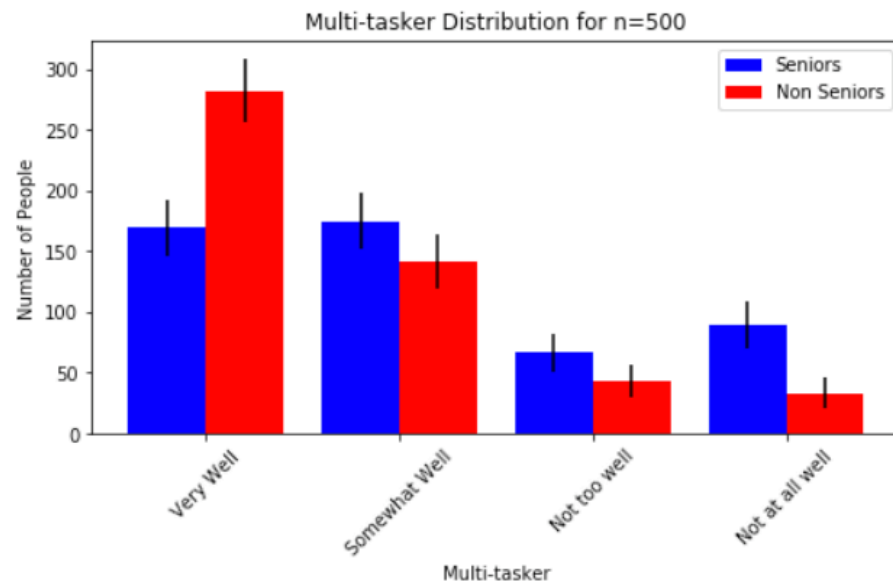
Growth Mindset

Overall, senior respondents demonstrated a slightly lower propensity towards growth mindset than non-seniors. For example, slightly more senior respondents reported that the idea that "people can learn new things but cannot change their basic level of intelligence" described their views well. To some extent, we might expect this to be the case simply because growth mindset is a newer attitude towards learning whilst fixed talent/intelligence is an older one.

Senior respondents were also slightly more likely to report that the idea "Truly smart people do not need to try hard" described their views well.
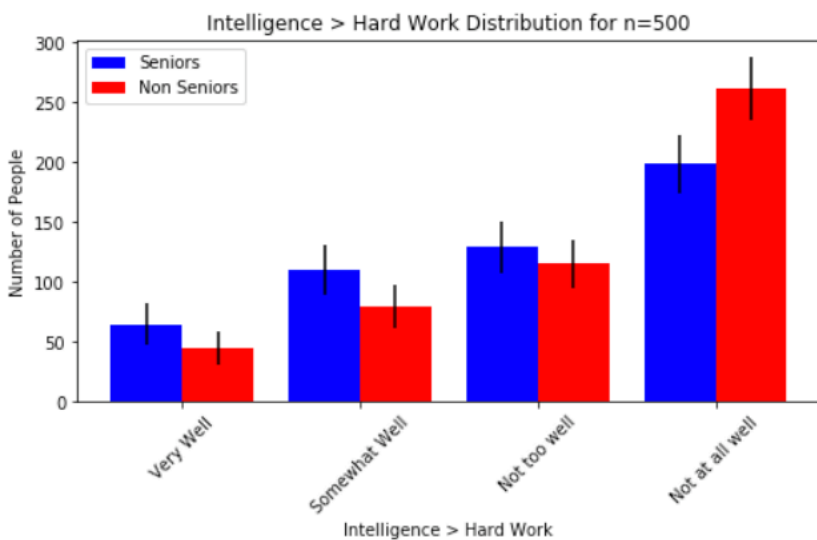
In terms of holistic openness to new experiences, the scores of seniors tended towards only slightly less preference for and openness to novelty. This bodes well for the game's product traction and provides a compelling counter-argument to those who might say seniors are globally unwilling to experiment. For example, senior respondents reported that trying new things was stressful only slightly more often than non-seniors.

Novelty Stressful Distribution for n=500