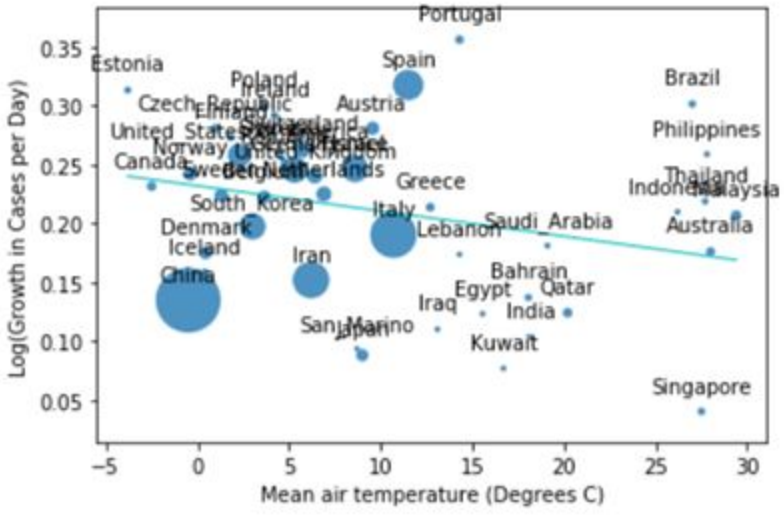
My capstone project has focussed on three different projects. The first part of the project examined whether cases of COVID-19 could be linked to differing air temperatures in different regions of the world. I also examined whether usage patterns on the language learning website duolingo could be linked to a users native language and the language they were learning. Finally, I examined the factors contributing to housing prices in Brazil.

The spread of infectious disease is a problem that affects every human on the planet. To illustrate this fact, as of April 24, 2020, there have been 196 thousand deaths from COVID-19 alone. One of the questions I wanted to address about COVID-19 was whether the COVID-19 case number reports were related to the ambient air temperature in a region. One thing that is known about corona viruses is that they become less stable at higher temperatures. I had seen some reporting that the virus was not as potent in warmer regions, but I also noticed that people were talking about total case numbers and were not looking at the growth rate of the virus. On March 17th, I downloaded a dataset of COVID-19 cases that had been recorded by each country from the European Center for Disease Control (ECDC) from the following link.

https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide.

I wanted to assess if the growth rate of COVID-19 cases is correlated with the average temperature in the region of interest. This required splitting the dataframe into a dictionary of dataframes. I then had to create new columns with total observed cases. I limited the countries I examined to those with more than 50 cases and excluded portions of the data where case number had leveled off to prevent the calculation of a grow rate that was too low. COVID-19 is an infectious disease with non-linear, and in this case exponential growth, so I fitted exponential curves to the COVID-19 growth rate data and extracted the growth rate variable from the qualifying countries. I then gathered air temperature data for each country from the month of February to analyze the growth rate of the virus against temperature. There was no strong correlation between the case-growth rate of the virus and ambient air temperature.



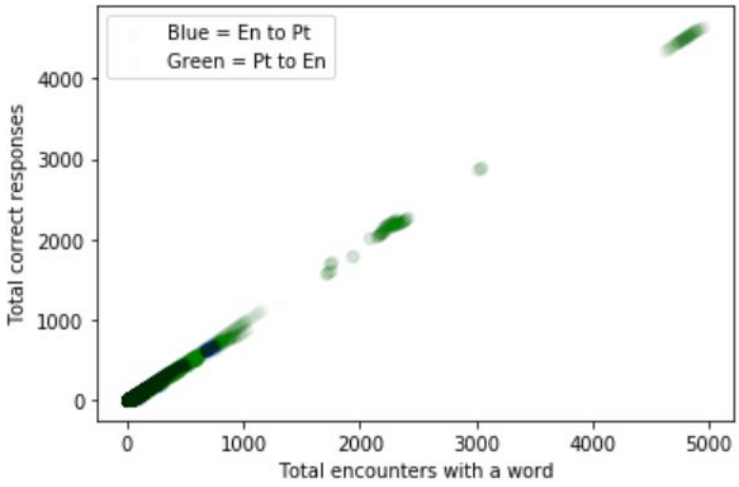
In the plot, the size of the data point is reflective of the total number of confirmed cases observed in the data with larger points being representative of more cases. China, South Korea, and Italy all plot under the regression line. These countries all put strong social distancing measures into effect by March 17th. These countries also show

greater case numbers, represented by larger circles, which could suggest that perhaps the rate of viral infection is already slowing down in these countries and the analysis failed to capture that. Many of the countries above the line show intermediate case numbers perhaps suggesting that countries fail to enact social distancing measures soon enough. The testing rate may also be contributing to the variance in the data.

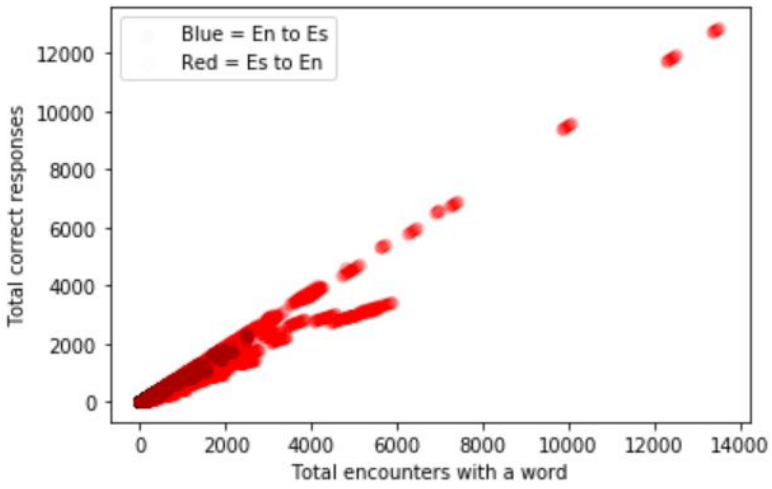
As for language acquisition, the link to the duolingo data was acquired from the following link

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N8XJME

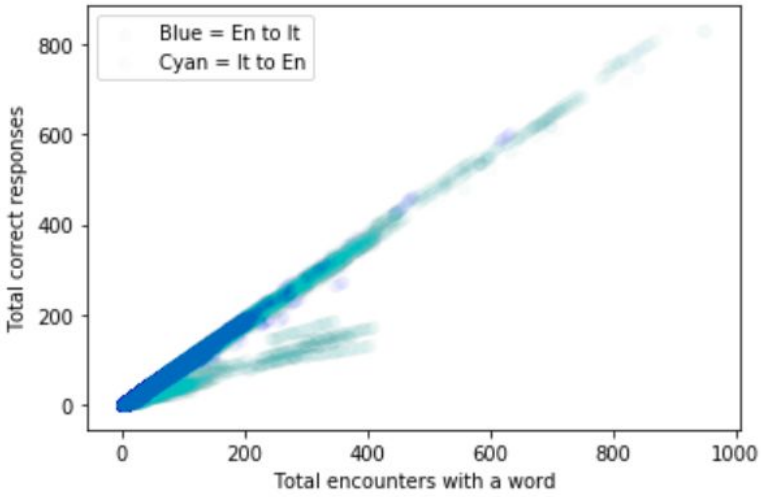
One of the questions I was interested in using the data to answer was if there were differences in platform usage between users with different target languages. This is an important question for the creators of duolingo potentially in terms of marketing. If the motivation or persistence of groups is related to the language they are trying to learn, campaigns or on site advertisements may be more effective for some users than others. I started my analysis by plotting the number of total correct responses to a word a user has accumulated over their history on the site versus the total number of times a user has seen on the site to observe if any major differences appear between groups. I started by examining Portuguse speakers learning English and English speakers learning Portuguese.



The above figure shows that Portuguese to English learners show more persistence (it may be driven by a few users), but not necessarily better rates of language acquisition. Perhaps it is more advantageous for a Portugese speaker to know English and maybe this accounts for the increase in persistence as shown by increases in the total number of times a word has been encountered. I then continued my exploratory analysis by asking if this relationship is also present for English and Spanish learners. The result of that plot is below.

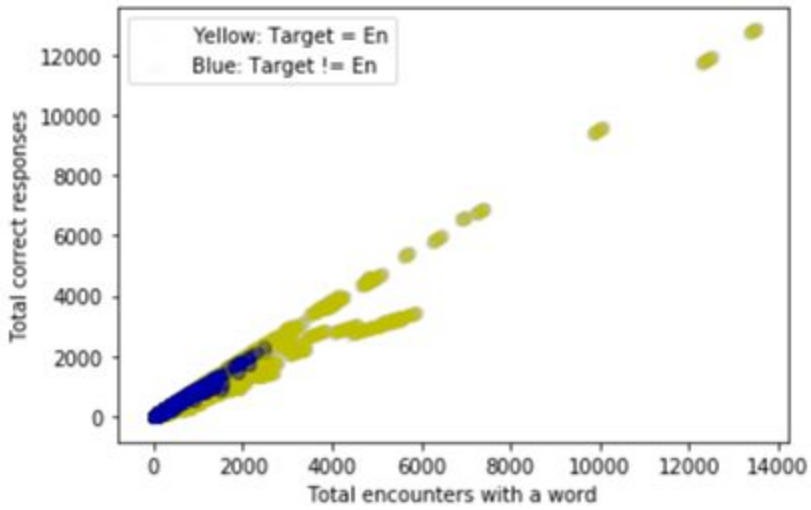


This result also held for learners of English from Italian and learners of Italian from English.

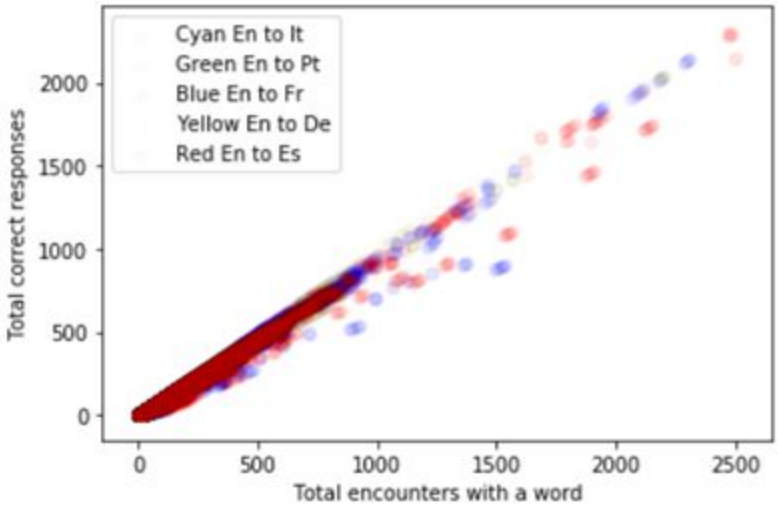


While the overall pattern of learners of English exhibiting greater intensity in their usage of the website persists for these two groups, the persistence of these two groups seems to be the lowest yet. This all seems to suggest that personal motivation to learn a language may be tied to economic reasons. There also seems to be a second population of English language learners that do not acquire the language as easily. This could suggest a lack of motivation for this group or that English may present a difficult challenge for Italian speakers.

Of the languages I have examined, users who are learning English interact with the website the most as indicated by the increase in the number of times they have seen a word compared to the number of times learners of other languages. As can be seen in the figure below, the yellow dots (a target language of English) show much higher total encounters with words than the blue dots (a target language of not English).



Additionally, among users with a native language of English, usage differs by the user's target language. The following plot makes this point difficult to observe, however.



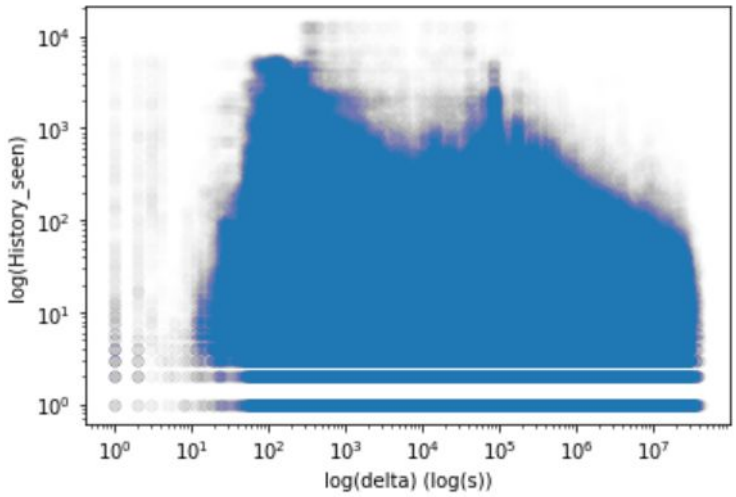
As seen in the above plot, the large number of data points obfuscates any obvious difference between total encounters with a word and the users target language. However, differences are able to be observed through ANOVA ($df = 4$, $F = 13577$, $p = 0.0$). Using Tukey's Honestly significant difference, a post-hoc test to determine which groups show two-way differences in ANOVA, all of the groups showed differences in user engagement except for English to Spanish and English to Portuguese learners (shown below).

Multiple Comparison of Means - Tukey HSD, FWER=0.01

group1	group2	meandiff	p-adj	lower	upper	reject
de	es	1.1043	0.001	0.9833	1.2252	True
de	fr	7.4926	0.001	7.3576	7.6275	True
de	it	-1.7342	0.001	-1.9046	-1.5638	True
de	pt	0.8976	0.001	0.6566	1.1386	True
es	fr	6.3883	0.001	6.2773	6.4993	True
es	it	-2.8385	0.001	-2.9906	-2.6864	True
es	pt	-0.2067	0.0269	-0.4352	0.0218	False
fr	it	-9.2268	0.001	-9.3902	-9.0633	True
fr	pt	-6.595	0.001	-6.8311	-6.3588	True
it	pt	2.6318	0.001	2.3737	2.8899	True

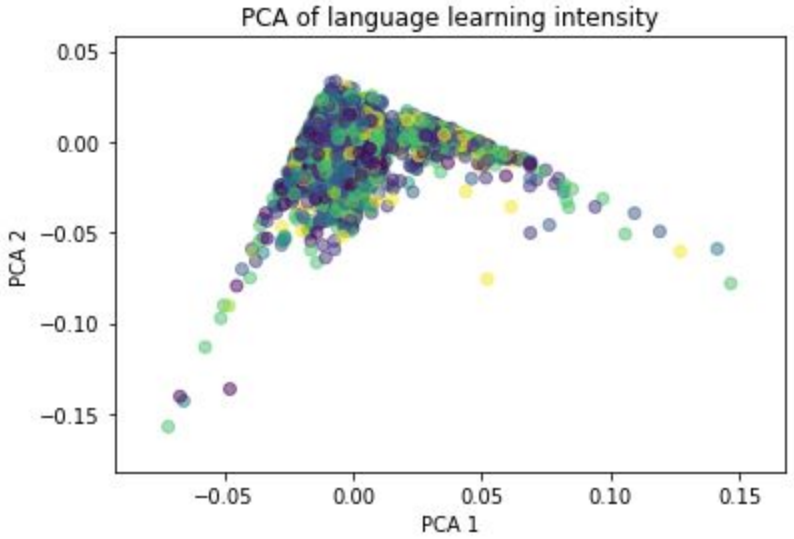
The above results show that user target language/user native language pairs are predictors of user engagement on duolingo and will likely be helpful for modeling user engagement on the site.

Another factor that may be important in modeling how strongly a user may interact with the website is the amount of time they have between practice sessions. It may be a reasonable hypothesis that smaller deltas tend to correlate with more observations of words



The above plot shows a partially decreasing relationship with the time between practice sessions and the total number of times a user has seen a word. The plot also shows that at small intervals between sessions the number of total words seen increases. So the overall relationship may be described at this preliminary stage of analysis as parabolic (or at least increasing over part of the data and decreasing over another part of the data). It appears that delta will be an important factor to consider when predicting how likely a user will be to engage with the site.

I decided that important measures to consider when measuring a user's persistence in language acquisition included the total number of training sessions, the total duration of time the user spent on the website, the slope of the relationship between practice sessions and time, and the r^2 of the relationship between practice sessions and time. I visualized the data using principal components analysis to examine the data for any obvious clustering. Visual analysis of the data by PCA revealed that the data did not show obvious clustering by language acquisition pairs.



I tried recovering or predicting the language acquisition pairs using support vector machines (SVM). The resulting confusion matrix produced by SVM when looking for 6 language acquisition pairs shows that it couldn't reasonably distinguish between different groups. The algorithm instead sought to predict the two major learning pairs present in the data.

Confusion matrix of language learning pairs							
0	72	0	0	0	56	0	0
0	205	0	0	0	113	0	0
0	99	0	0	0	69	0	1
0	46	0	0	0	28	0	0
0	13	0	0	0	15	0	0
0	161	0	0	0	161	0	2
0	26	0	0	0	13	0	0
0	46	0	0	0	44	0	0

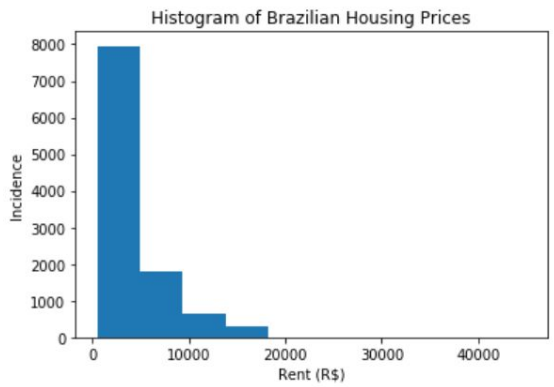
The Brazilian housing data was obtained from the website Kaggel, and can be found here:

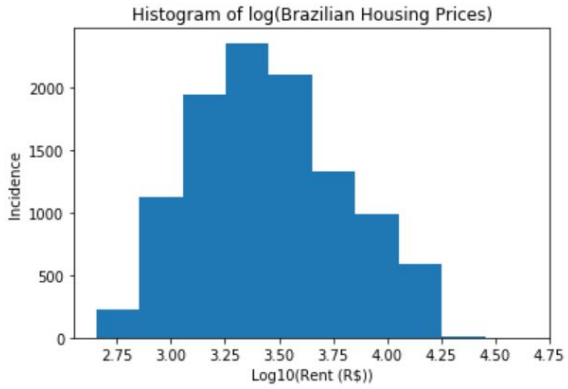
https://www.kaggle.com/rubenssjr/brasilian-houses-to-rent

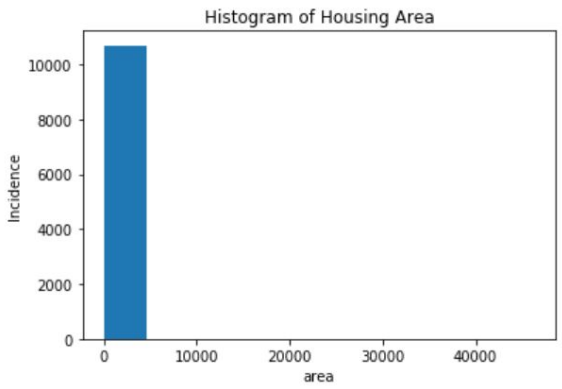
I am fascinated by Brazil and am a Lusophone. When I found the Brazilian housing dataset on Kaggle, I downloaded it. I was interested in developing a model to predict the housing prices in Brazil. I thought that a model that accounted for the city, the area of the house, the number of rooms, and nuber bathrooms should predict the rent. Since the model is seeking to predict a continuous numerical variable, I decided that multiple regression would be the best machine learning method to use.

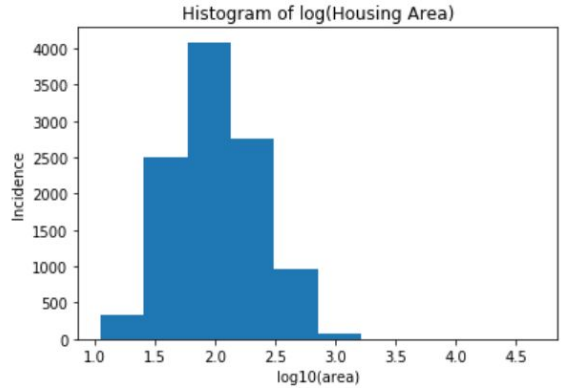
I noticed that several of these variables such as rent, the area of the house, and the number of rooms all appeared to show log-normal distributions as shown below. Taking the log of these distributions yielded distributions that were much closer to normal distributions.

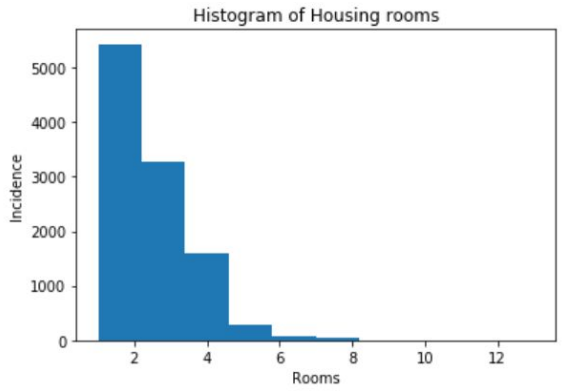
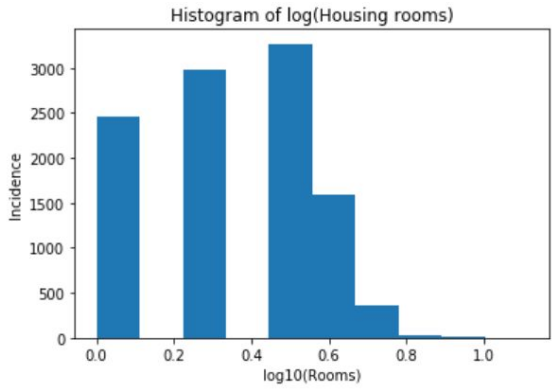
Plots showing log transformation of metrics of interest.









Plots showing log transformation of metrics of interest continued.	
	

I expected the number of bathrooms to be much less important in predicting the rent than other variables.

The model based on these variables was successful in explaining 58 % of the variance in the test data. The resulting model also performed well when making predictions on the test data, as shown in the below figure. The model may overpredict the rent at lower housing prices and slightly underpredict it at high housing prices.