Capstone 1

Exploration of Coronavirus data, Data from Duolingo, and Brazilian Housing models

The spread of Coronavirus

In 2019 a novel coronavirus was discovered in China

The novel virus has been named SARS-CoV-2

The disease it causes has been named Covid-19

The virus is fairly contagious and relatively lethal

It has disrupted markets on a global scale and has been declared a pandemic by WHO

Predicting the spread of Coronavirus may help save lives

Coronavirus and temperature

Coronaviruses are typically less stable at higher temperatures

I wanted to understand if the ambient air temperature in a region was a predictor determining the growth rate of the novel coronavirus

I downloaded data on March 17th from the European Center for Disease Control

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

I analyzed Covid-19 case growth rates among countries that had 50 or more recorded cases - Amounted to 45 countries

Coronavirus and temperature

Cases of COVID-19 show an exponential growth rate in the population

- I fitted exponential models to the portions of the curves that exhibited exponential growth

I then examined the relationship between the ambient temperature in a country and the growth rate of COVID-19

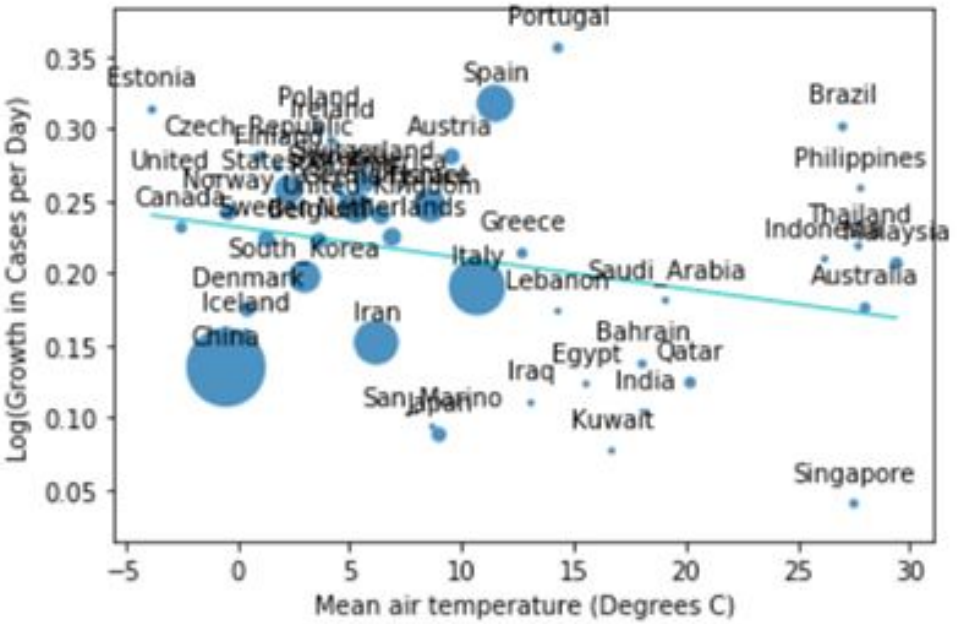
Coronavirus and temperature: Results

R2 = 0.08

P = 0.06

No strong relationship between growth rate and temperature

Notice that many countries that fall below the curve put in place strong social distancing orders



Duolingo and second language acquisition

Duolingo is a website that promotes second language acquisition and learning

I have been interested in examining if it is possible to predict how intensely a user will interact with the website Duolingo

The following slides show some plots, and statistical summaries I have generated as part of my exploratory data analysis

Duolingo data wrangling

The data I downloaded from duolingo was available as .csv file from the following website

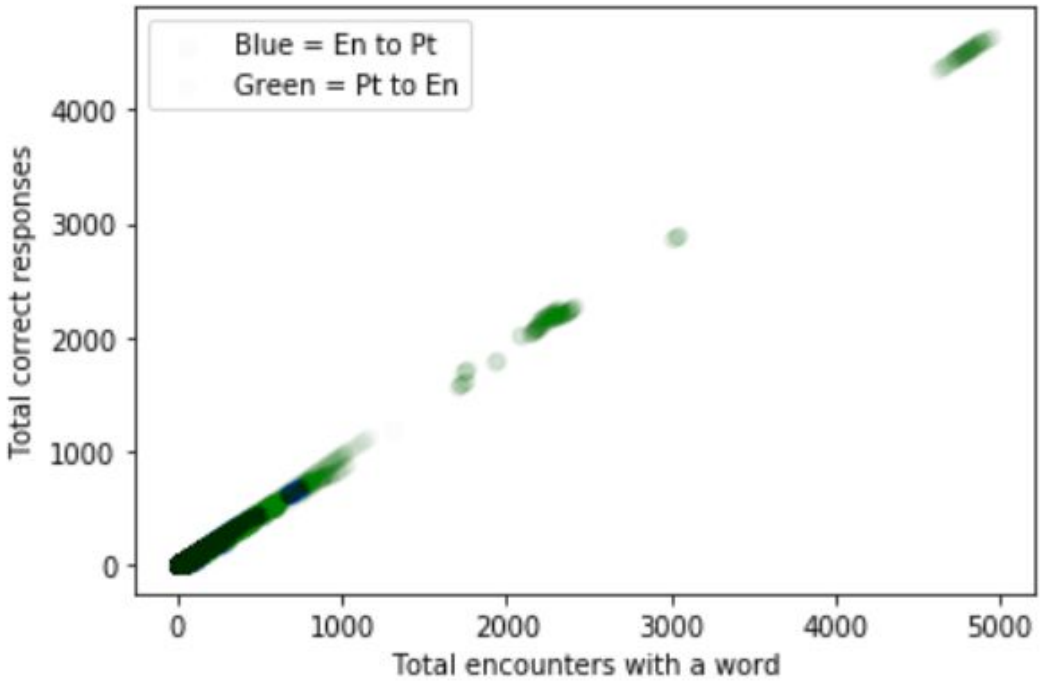
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/N8XJME

The file itself was well taken care of, and required very little reshaping for the exploratory analysis phase

I began the exploratory phase of analysis by examining how many times a user correctly identified a word and how many times a user had encountered a word

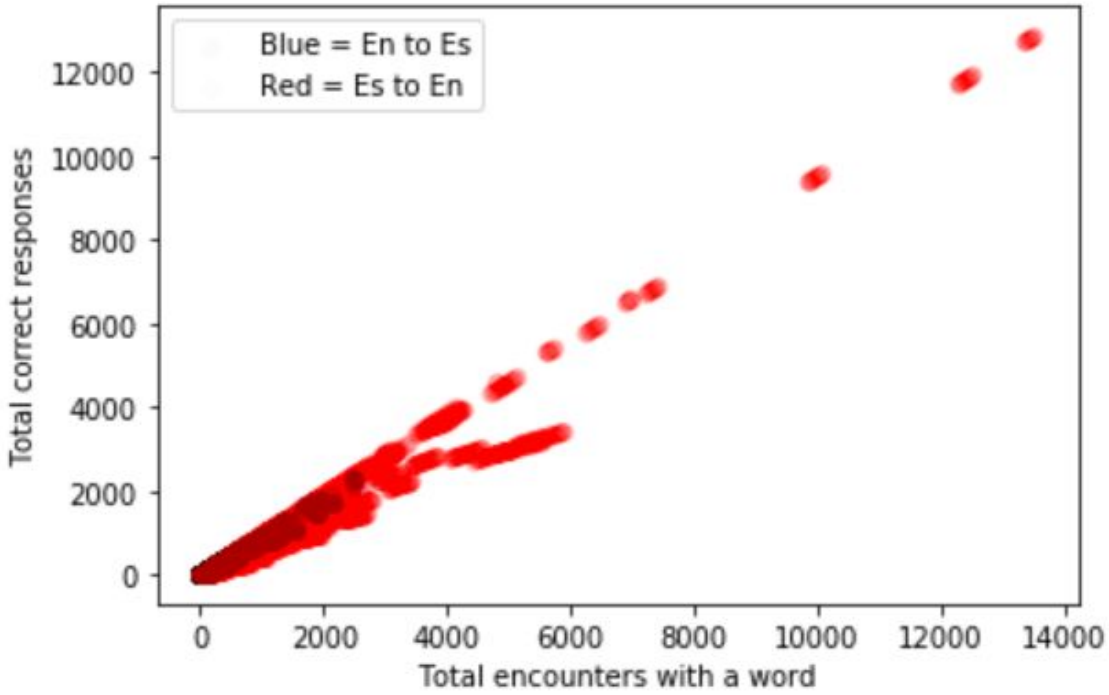
Duolingo: En to Pt and Pt to En

English learners from Portuguese appear to show more persistence (as measured by total encounters)



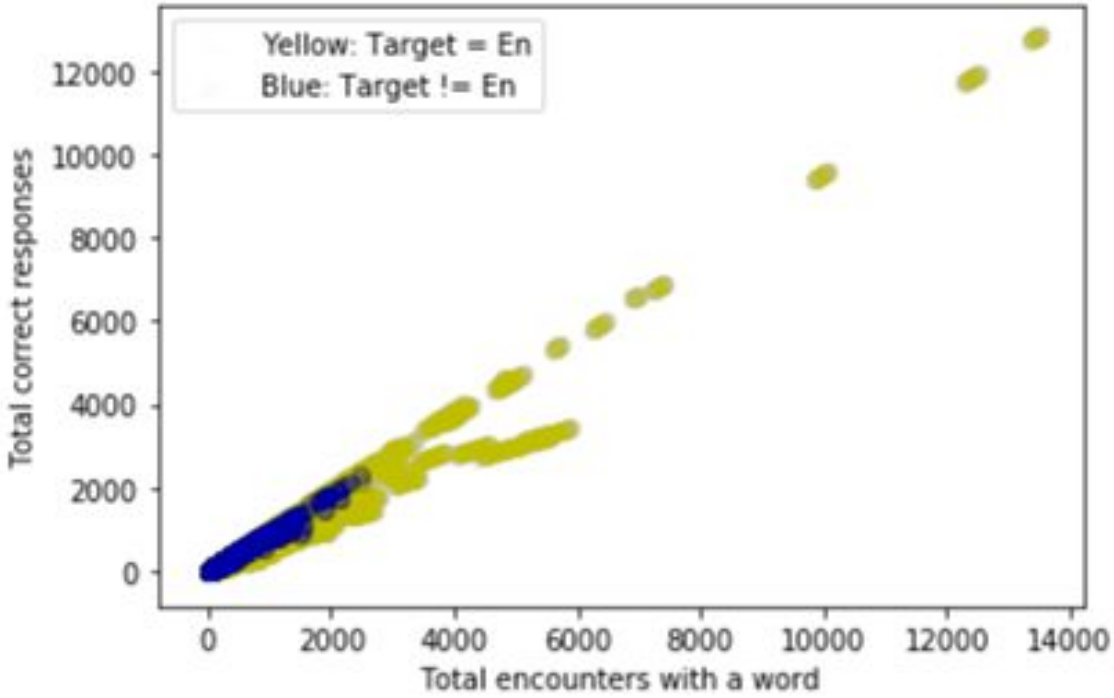
Duolingo: En to Es and Es to En

English learners from Spanish appear to show more persistence (as measured by total encounters)



Duolingo: En to Target and Target to En

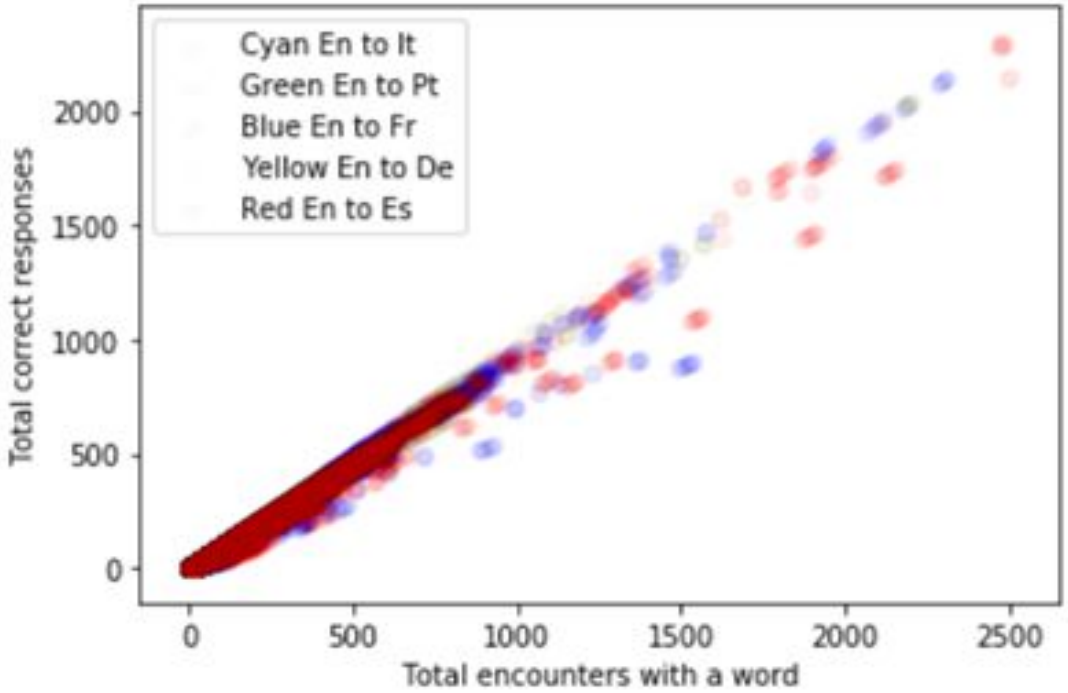
English learners from any target language appear to show more persistence (as measured by total encounters)



Duolingo: Native English Users

Do native English speakers show differences with different target languages?

The resulting plot does not obviously show any differences. Let's use an ANOVA and Tukey HSD



Duolingo: Native English Users ANOVA, Tukey HSD

ANOVA - English learners show differences with different target languages

The only target languages that did not show differences with Portuguese and Spanish

Multiple Comparison of Means - Tukey HSD, FWER=0.01

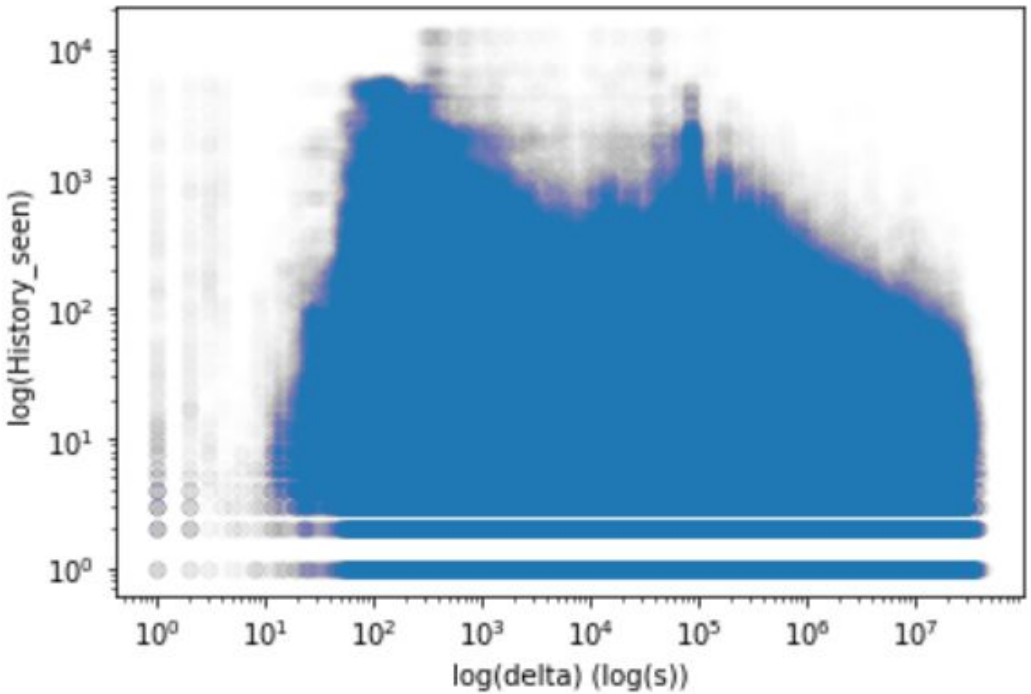
group1	group2	meandiff	p-adj	lower	upper	reject
de	es	1.1043	0.001	0.9833	1.2252	True
de	fr	7.4926	0.001	7.3576	7.6275	True
de	it	-1.7342	0.001	-1.9046	-1.5638	True
de	pt	0.8976	0.001	0.6566	1.1386	True
es	fr	6.3883	0.001	6.2773	6.4993	True
es	it	-2.8385	0.001	-2.9906	-2.6864	True
es	pt	-0.2067	0.0269	-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

Duolingo: Time between practice sessions

Time between sessions may also be important for predicting user intensity

The resulting plot shows parabolic distribution

Time between sessions likely important future modeling



Duolingo: Gauging user intensity

I created metrics for user intensity by:

1) Measuring the total number of times a user completed a training session

2) Measuring the total duration of a user on duolingo

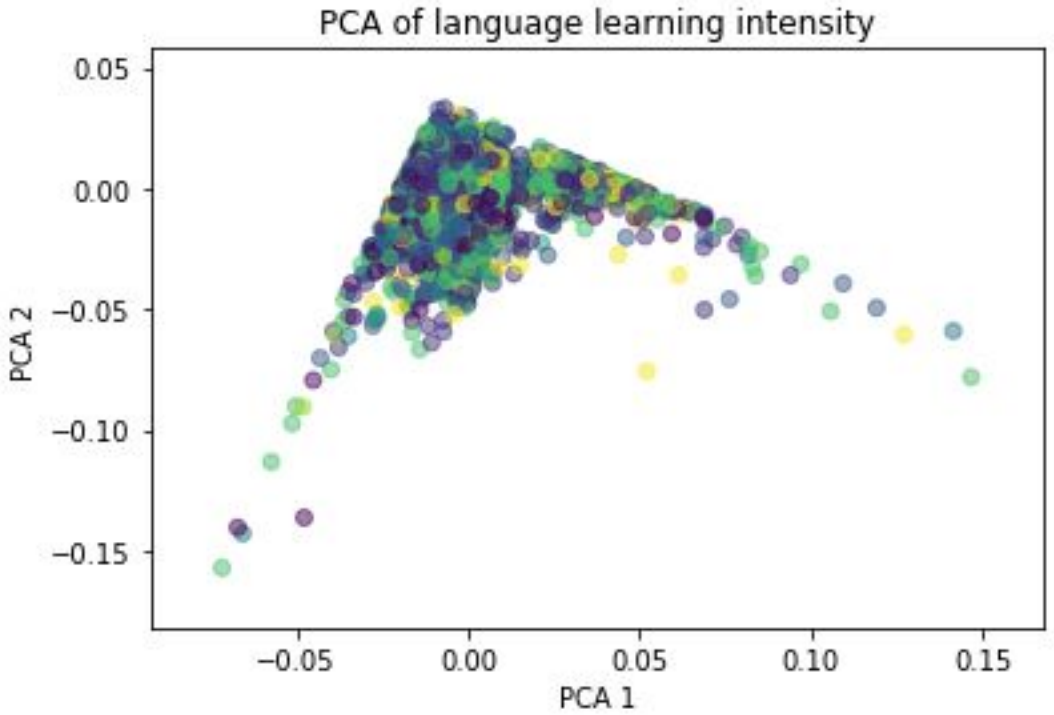
3) Measuring the slope of accumulated training sessions over time

4) Measuring the r^2 of the relationship between training sessions and time

I then examined if these metrics clustered with language acquisition pairs

Duolingo: Visual clustering examination

PCA analysis did not show strong clustering by language learning acquisition pairs



Duolingo: Support vector machines

SVM failed to accurately classify the users by language acquisition pair

Similar to predicting where someone in the U.S. is from based on their height

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

Brazilian Housing Data

The availability of housing affects every person

Understanding the factors that affect housing can help people spot deals in the market

Brazil is a fascinating country and I speak Portuguese so was interested in a Brazilian housing dataset posted on Kaggle

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

Brazilian Housing Data: Wrangling + Model

I was interested in predicted the rent from variables in the dataset

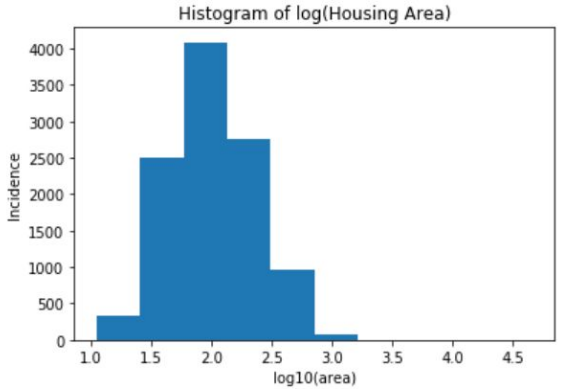
1) Number of rooms,
2) Housing area,
3) Number of bathrooms
4) The city

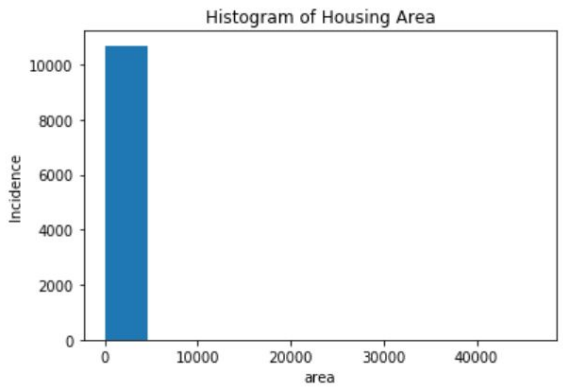
Brazilian Housing Data: Wrangling + Model

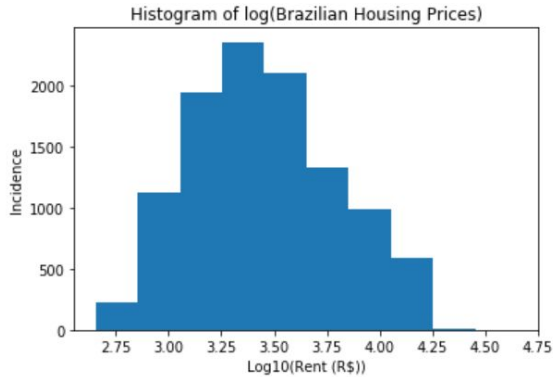
Log transformations

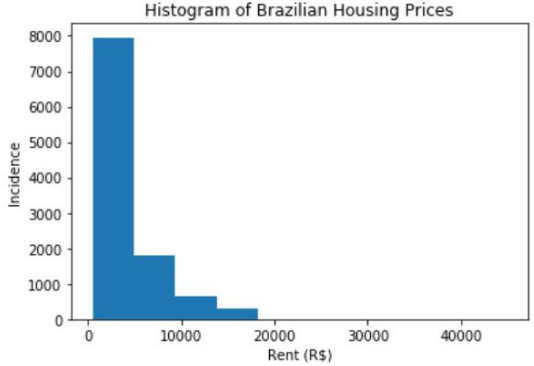
Rent, Number of Rooms, Housing area were all roughly log normal

Taking the log of the data produced roughly normal distributions









Brazilian Housing Data: Predictions

Model accounted for 58 % of the variance in the data for training data

Predicted well on test set