How much is my stuff worth?

An investigative look at the Mercari Marketplace and building a model to suggest product prices

Agenda

AN INTRODUCTORY LOOK AT THE TASK AT HAND

What is the current issue? Why is it a difficulty?

DELVING INTO THE DATA

Taking a further look at the data to look for trends or issues

THE SOLUTION

Building a model to complete this task

WHAT IS MERCARI?

- Mercari's main product is a community-powered shopping app, currently operating in Japan and the United States
- First Japanese company to reach unicorn status
- Individuals can buy and sell items with ease on their smartphones
- Users of the app have the freedom to choose the price while listing the item

THE PROBLEM

What price should I list my product for?

Small differences can make a huge difference in product pricing.

Product A: "Vince Long-Sleeve Turtleneck Pullover Sweater, Black, Women's, size L, great condition."

Product B: "St. John's Bay Long-Sleeve Turtleneck Pullover Sweater, size L, great condition"

One of these products is worth $335, and the other is $9.99.

Products have variables such as seasonality, are heavily influenced by brand names and while most products depreciate over time, some increase in price.

THE TASK

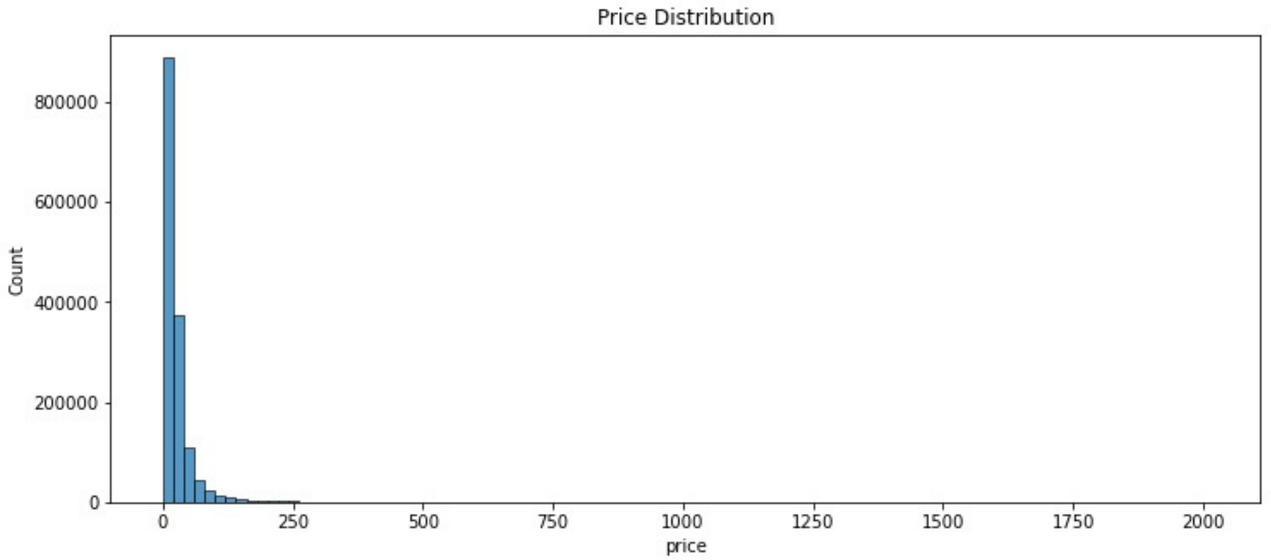
Create an algorithm that will predict the price of what a product should be listed as.

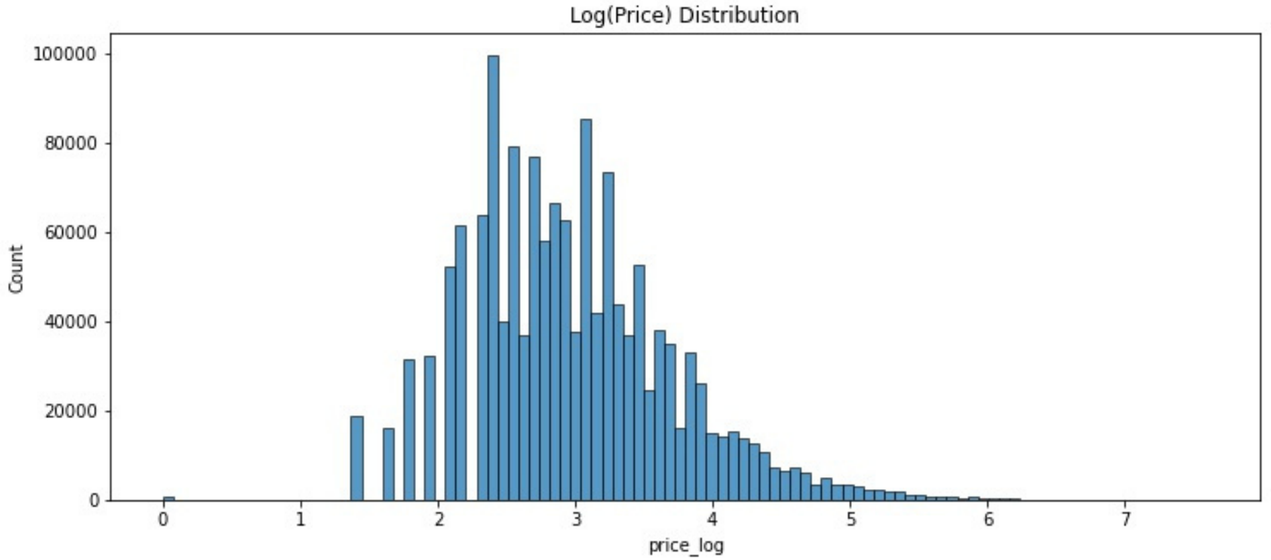
The data will include user-inputted descriptions, including details like product category name, brand name and item condition.

A snapshot of the data can be seen below:

	name	item_condition_id	category_name	brand_name	price	shipping	item_description
0	MLB Cincinnati Reds T Shirt Size XL	3	Men/Tops/T-shirts	NaN	10.0	1	No description yet
1	Razer BlackWidow Chroma Keyboard	3	Electronics/Computers & Tablets/Components & Parts	Razer	52.0	0	This keyboard is in great condition and works like it came out of the box. All of the ports are tested and work perfectly. The lights are customizable via the Razer Synapse app on your PC.
2	AVA-VIV Blouse	1	Women/Tops & Blouses/Blouse	Target	10.0	1	Adorable top with a hint of lace and a key hole in the back! The pale pink is a 1X, and I also have a 3X available in white!
3	Leather Horse Statues	1	Home/Home Décor/Home Décor Accents	NaN	35.0	1	New with tags. Leather horses. Retail for [rm] each. Stand about a foot high. They are being sold as a pair. Any questions please ask. Free shipping. Just got out of storage
4	24K GOLD plated rose	1	Women/Jewelry/Necklaces	NaN	44.0	0	Complete with certificate of authenticity

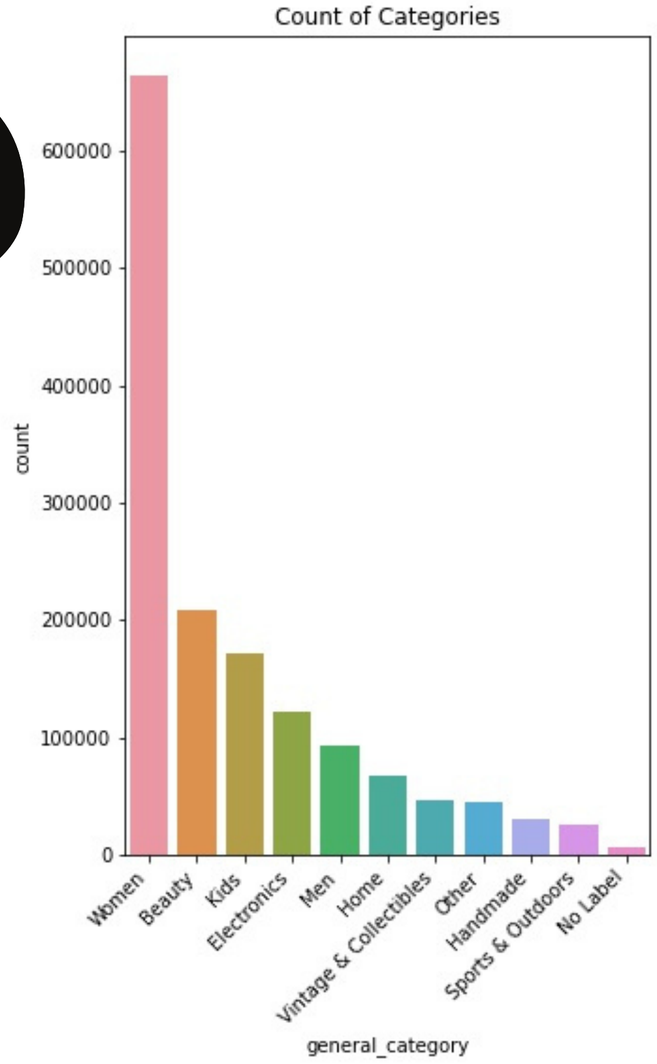
Price Issues?

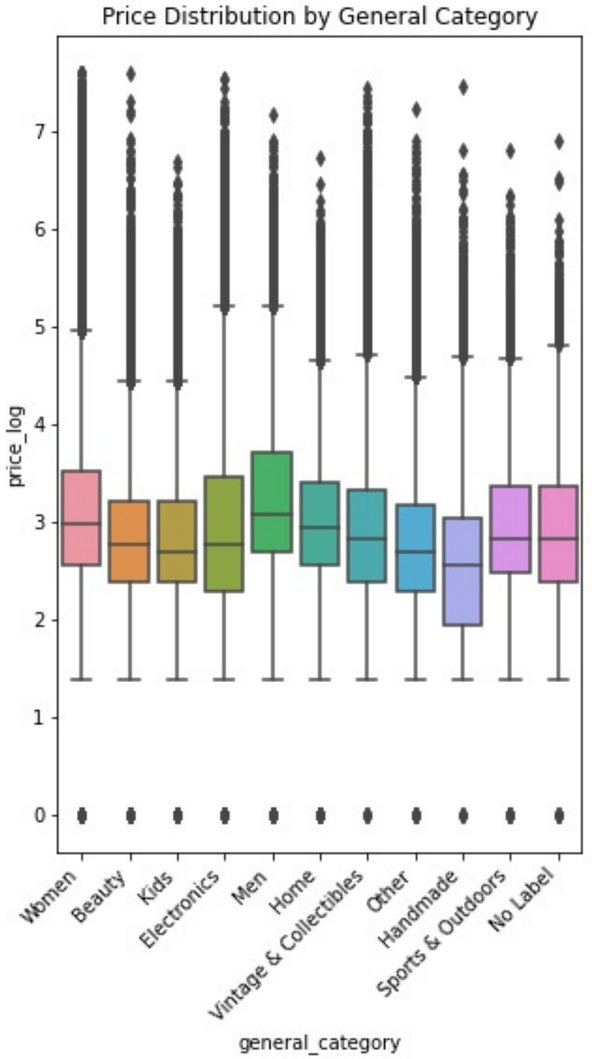




- A skewed distribution can result in less reliable predictions
- Majority of products are in the $0 to $250 range, however some go up to $2000
- Taking the logarithm of the price results in a normal distribution, greatly improving predictability

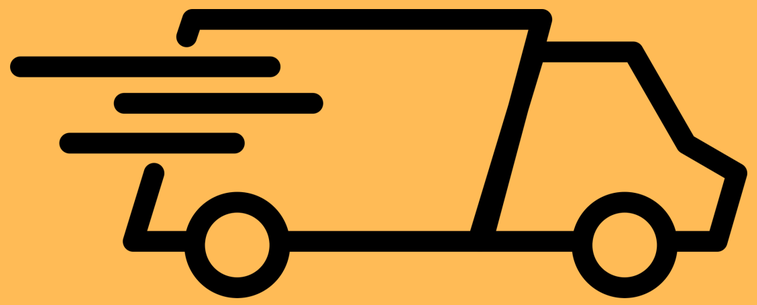
General Categories





- The 'Women' category has by far the most products
- A fairly similar price distribution between categories. No one category is much more expensive or much cheaper than the others

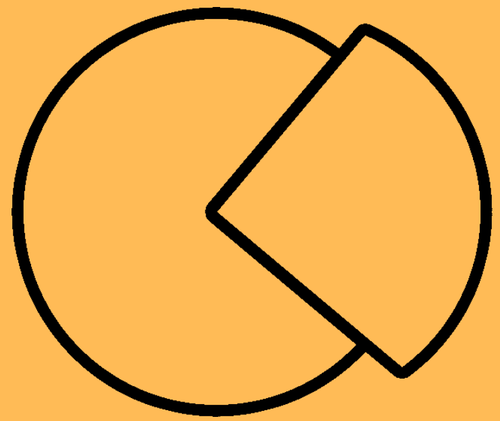
Other Interesting Statistics

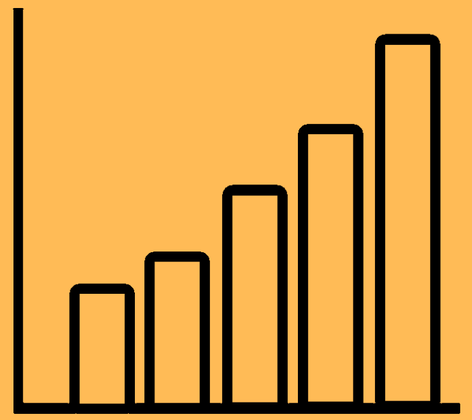


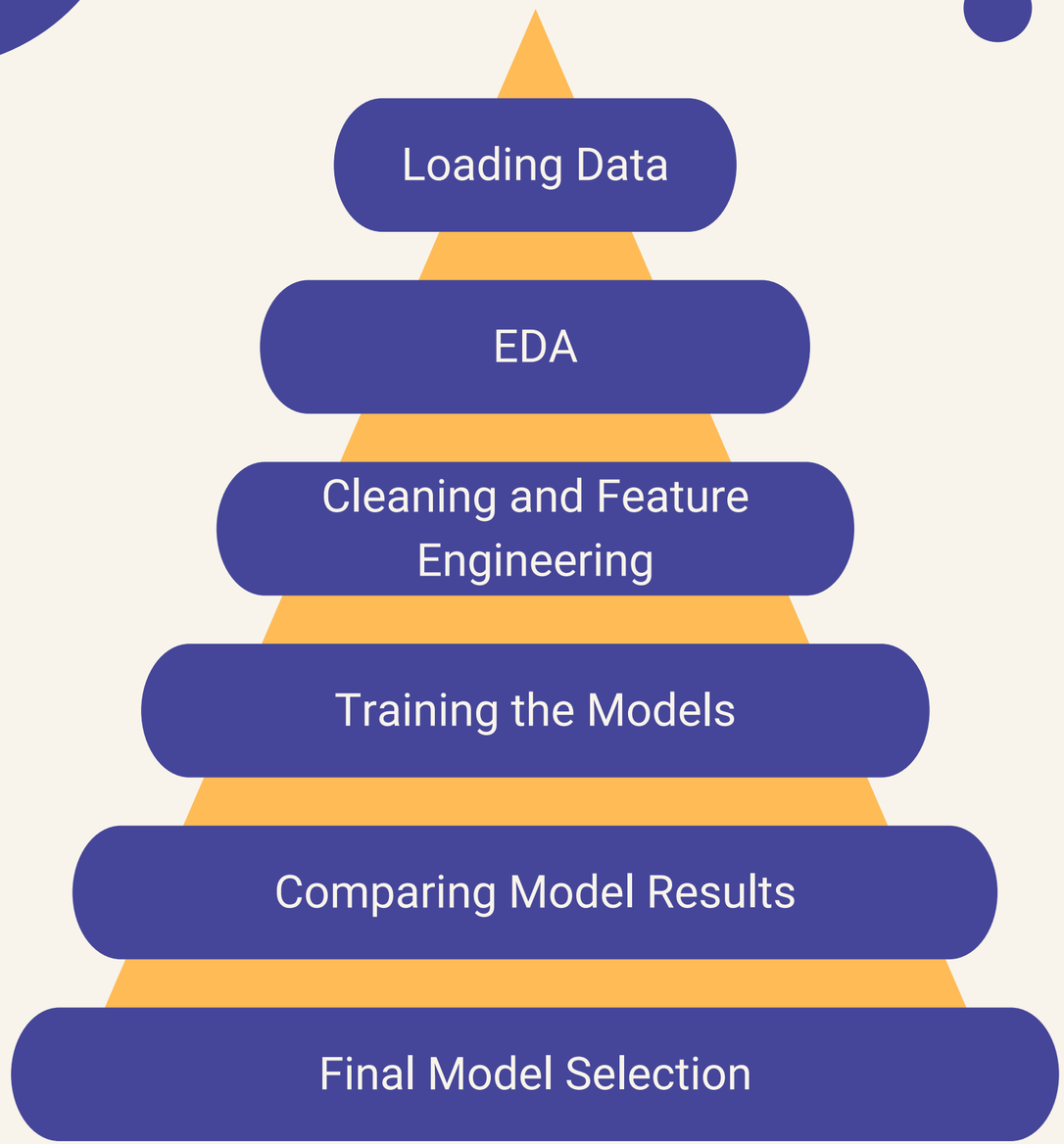
44.7% of the sellers paid for Shipping Costs

43.2% of items were in brand new condition

The two most expensive items were two Chanel bags for just over $2000







Developing the Model

- As the data was user inputted, there were many missing values in optional sections. To counteract this, a brand finder function was produced
- Conducted Natural Language Processing on the names and descriptions of the products
- As this would result in a huge amount of features, some methods were too computationally expensive
- Comparing model results and hyperparameter tuning to further improve the model
- Final model selection was completed and predictions were made

MODEL COMPARISON

The models would be evaluated by the Root Mean Squared Logarithmic Error (RMSLE).

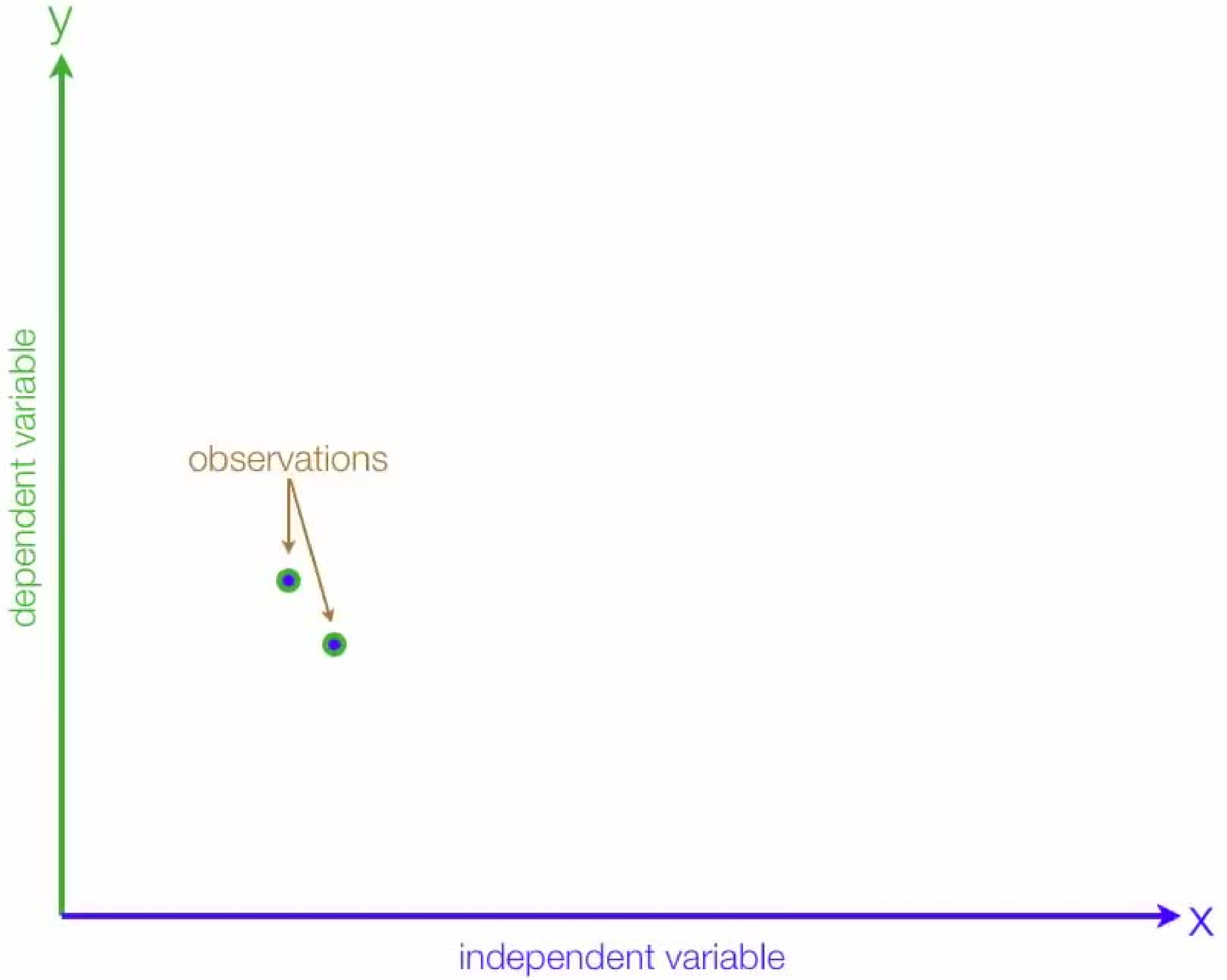
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

It is the Root Mean Squared Error of the **log-transformed predicted** and **log-transformed actual values**.

Model	RMSLE	Time Taken
Ridge	**0.4589**	19 Mins 42 Secs
SGD	0.50948	**16 Secs**
SVR	0.50596	22 Secs

MODEL SELECTION

Although the final RMSLE was fairly similar between the 3, the Ridge Regression was picked, despite the longer computational expense.


y
dependent variable
observations
x
independent variable

LINEAR VS RIDGE REGRESSION

What's the difference?

- Linear regression establishes a relationship between a dependent variable (Y) and one or more independent variables (X) using a best fit straight line (Y = MX + C)
- Due to a large number of features in the data, Ridge includes data shrinkage through regularisation
- This also solves the issue that many Linear Regression models have, of multicollinearity

FINAL RESULTS AND THOUGHTS

- As this was a previous Kaggle competition, comparisons could be made to other scores
- The winning RMSLE for this competition was 0.37758
- Our current RMSLE of 0.4589 would be in the top 30% on the leaderboard
- Could potentially improve this by utilising deep learning as this was a popular solution for this problem

THANK YOU!

ANY QUESTIONS?