Humberto Leon, Joao Lucena, Armando Zubillaga
DS3001 Foundations of Data Science
Professor Kyumin Lee
May 1, 2020

Project Checkpoint

Our dataset consists of about 130k wine reviews, and we are looking to see what attributes can help us predict the wine rating better. This dataset was found in Kaggle and has the following attributes:

Table 1: Wine Review Data Attributes

country	The country that the wine is from
description	The wine description given by the manufacturer
designation	The vineyard within the winery
points	The number of points WineEnthusiast rated the wine on a scale of 1-100 (Dataset consists only has score >=80)
price	The cost for a bottle of the wine
province	The province or state that the wine is from
region_1	The wine growing area in a province or state (ie Napa)
region_2	More specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley)
taster_name	Wine rater name
taster_twitter_handle	Wine rater twitter account
title	The title of the wine review, which often contains the vintage
variety	The type of grapes used to make the wine (ie Pinot Noir)
winery	The winery that made the wine

For our study, we have decided to focus on factors that have a higher chance of affecting each wine's rating. The main aspects we wanted to explore were the country, description, price, vintage(harvest year), and grape variety. We decided to stay away from very specific data such as the designation, province, regions, and winery information. For our purposes, no information on the wine rater is needed, so we decided not to use the taster name and twitter handle information.

Before any preprocessing, our numerical data had the characteristics shown on the table below. There were 129,971 data points in general. The points attribute had a mean of 88.45, with the minimum value of 80 and maximum of 100. Our price data had slightly less data points because of null values (we fixed that in the cleaning process). The range of price was $4-3330, with a mean of $35.36. We can see in this data, that the maximum value is very far from the 75 percentile value, and we need to address outliers in our cleaning process.

	points	price
count	129971.000000	120975.000000
mean	88.447138	35.363389
std	3.039730	41.022218
min	80.000000	4.000000
25%	86.000000	17.000000
50%	88.000000	25.000000
75%	91.000000	42.000000
max	100.000000	3300.000000

Figure 1: Initial Analysis of Numerical Data

For our data cleaning process, our first step was to delete duplicate entries and drop all the attributes that didn't apply. After that, we studied the data left and looked for null entries in all the remaining attributes. The description, title, and points columns did not present any null entries. However, we found out that the country attribute had 59

null entries, and variety had 1 null entry. Considering that we have about 130k samples, those entries could be erased without any major changes to our data.

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian

Figure 2: Wine Review with all attributes

	country	description	points	price	title	variety
0	Italy	Aromas include tropical fruit, broom, brimston...	87	NaN	Nicosia 2013 Vulkà Bianco (Etna)	White Blend
1	Portugal	This is ripe and fruity, a wine that is smooth...	87	15.0	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red
2	US	Tart and snappy, the flavors of lime flesh and...	87	14.0	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris
3	US	Pineapple rind, lemon pith and orange blossom ...	87	13.0	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling
4	US	Much like the regular bottling from 2012, this...	87	65.0	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir

Figure 3: Wine Review with only selected attributes

Continuing to study the null values, we came across the information that about 10% of our data did not have a price listed. Furthermore, we decided to do a boxplot of the price data and also found outliers in the data. This is a considerable amount of our data, so we replaced the null entries with the mean of the prices after doing a boxplot and getting rid of the outliers for the price category. We used the rule which says that a data point is an outlier if it is more than 1.5 * Interquartile Range(IQR) above the third quartile or below the first quartile. Figure 4 and 5 shows the price box plot of our price data before and after removing the outliers.

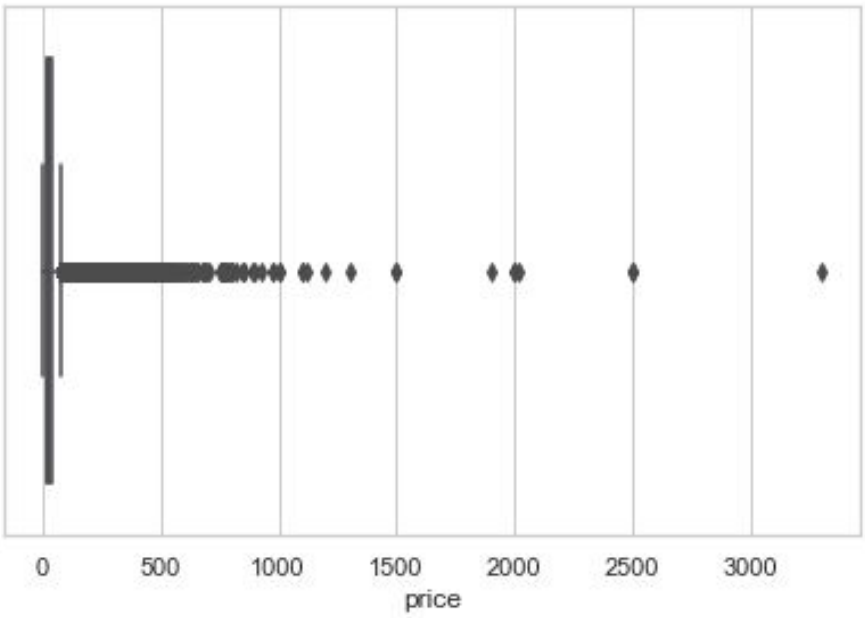

Figure 4: Initial Price Data Boxplot

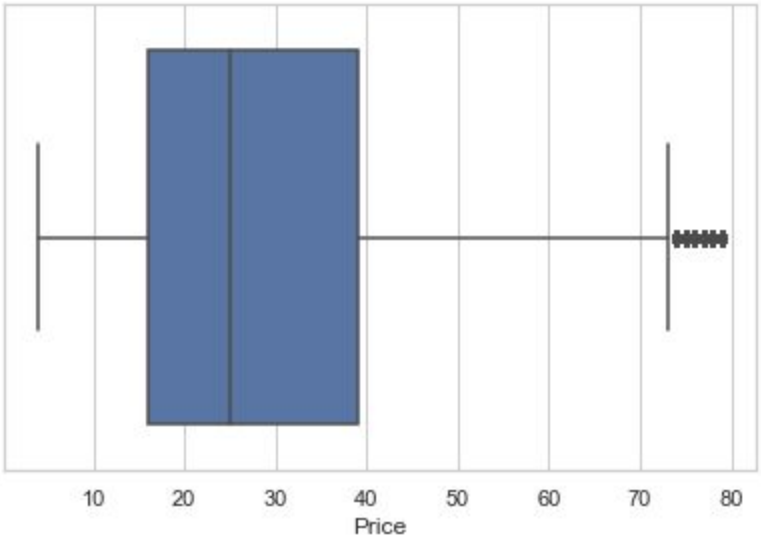

Figure 5: Price Data Boxplot without Outliers

Continuing the cleaning process, we extracted the year out of the given attribute ‘title’. About 4000 entries did not have a year in the title, those rows were removed from our data. For better interpretation of the description attribute, we found online the most commonly used words to describe wines and pasted it into a csv file. Each time a description had words in our wine description dictionary, it got a point. This way, we were able to formulate a grade system for each description.

Once we were done cleaning the data, we wanted to see how the points relate to our numerical data. First, we plotted points versus price, and we could observe some linearity (Figure 6). Other attributes were also analyzed as seen in Figures 7-10

Figure 6: Price vs Points Data

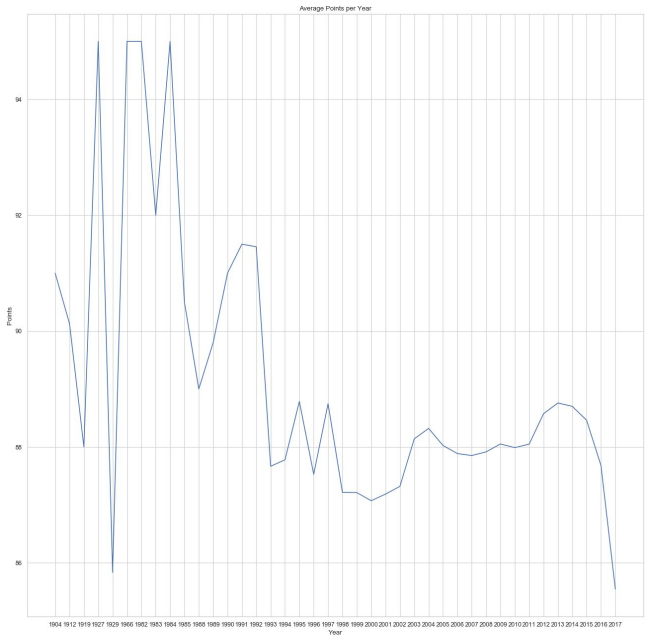

Figure 7:Average Points per Year

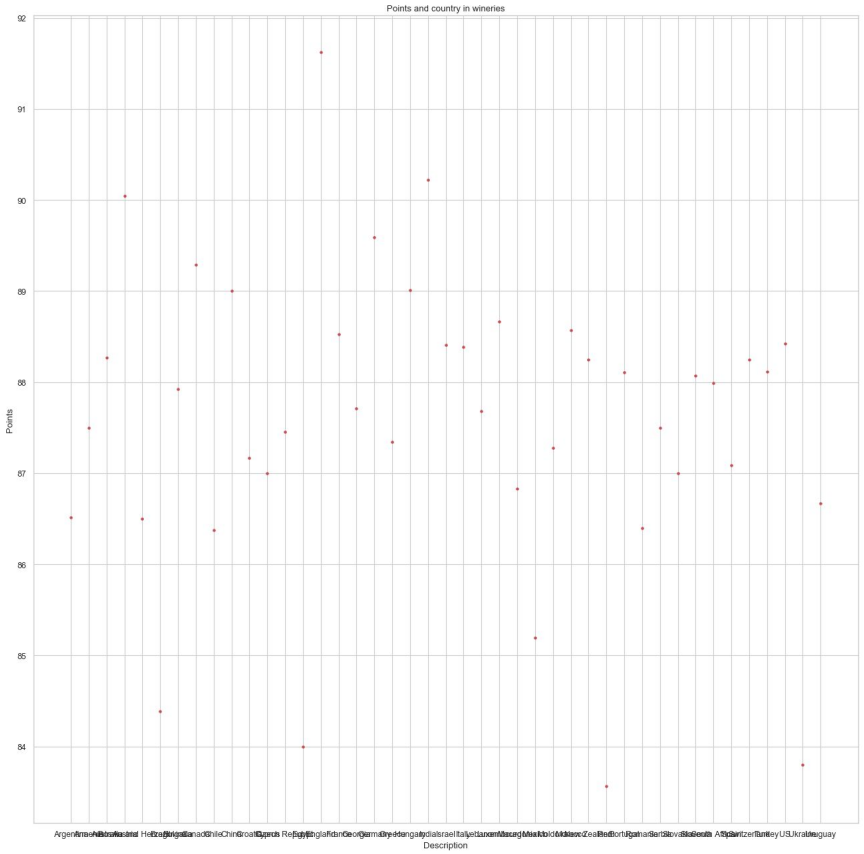

Figure 8: Average Points per Country

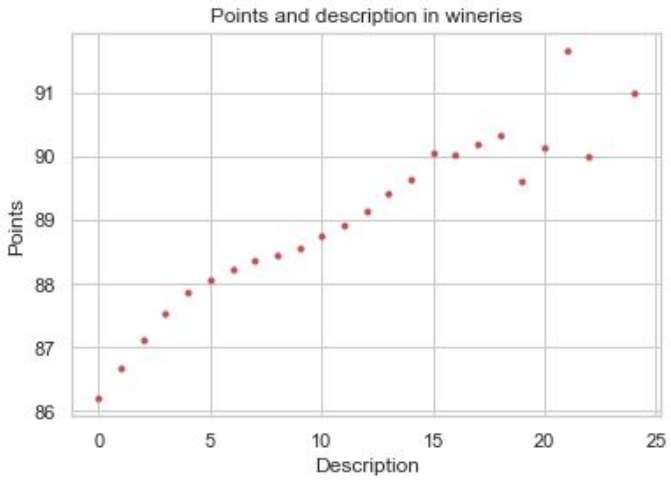

Figure 9: Points versus Description

Figure 10: Point Average versus Wine Variety

We also computed the grape variety frequency, and if that frequency was below 700 data points, it got reassigned to a category called ‘other’. This allowed us to have less types of grapes to perform the one-hot encoding later and convert categorical data into numerical data. The graph below shows Pinot Noir, Chardonnay and Red Blend as the most common grapes/grape blends in our data.

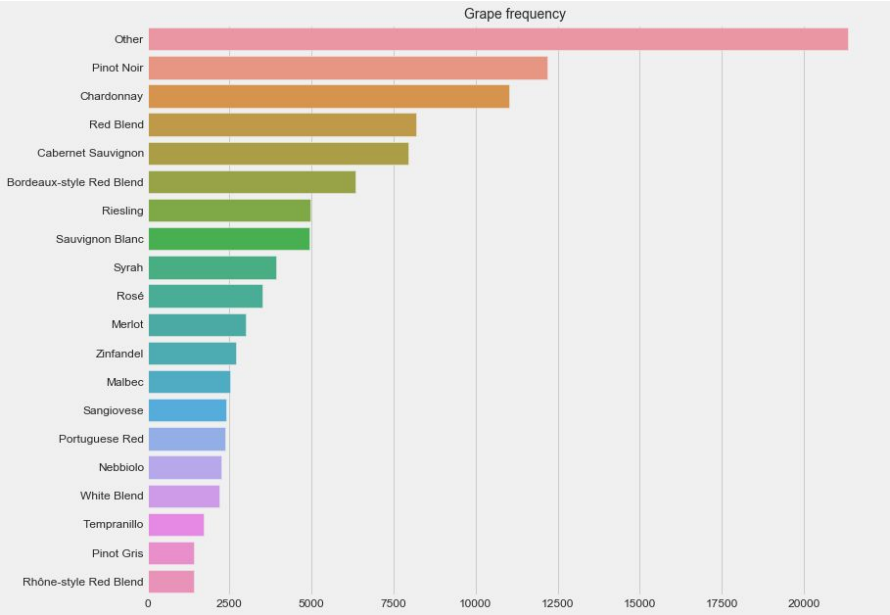

Figure 11: Variety Frequency 20 Most Frequent

To finalize our data processing, we one-hot encoded the country (Figure 12) and variety (Figure 13) data.

ıkia	country_Slovenia	country_South Africa	country_Spain	country_Switzerland	country_Turkey	country_US	country_Ukraine	country_Uruguay
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0

Figure 12: One-hot Encode for Country Attribute

variety_Sauvignon Blanc	variety_Shiraz	variety_Sparkling Blend	variety_Syrah	variety_Tempranillo	variety_Viognier	variety_White Blend	variety_Zinfandel
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 13: One-hot Encode for Variety Attribute

Next steps

After the data cleaning and preprocessing, we are going to run a correlation matrix to visualize how our attributes affect the points data in any way. Next, use bootstrap to generate the training and test dataset. This test dataset will not be touched until we have a final model ready to test. Then, we apply hierarchical cross-validation to generate another set of training data and test data using bootstrap as well. In the last pair, we are going to train and test our learning algorithms.

```
first_train = df.sample(frac = 0.7, replace = True, random_state=1)
first_test = df.sample(frac = 0.3, replace = True, random_state=1)
```

Figure 14: Bootstrapping Code in Pandas

Once we have two training and test data sets ready, we would have to normalize our numerical data. This transformation improves the numerical stability of your model generating accurate predictions. The normalization will be only applied to numerical categories.

Before applying a multi-linear regression model, we are going to test principal component analysis (PCA) to reduce some of our components to make the computation of the linear regression model much faster and efficient. Finally, train the multi-linear regression model with our bootstrap train dataset and test it on the first test dataset. We are going to determine if PCA had any impact in the model and if not research other dimension reduction models. Finally, test the model with the first test data set to see if we overfit the data or not.