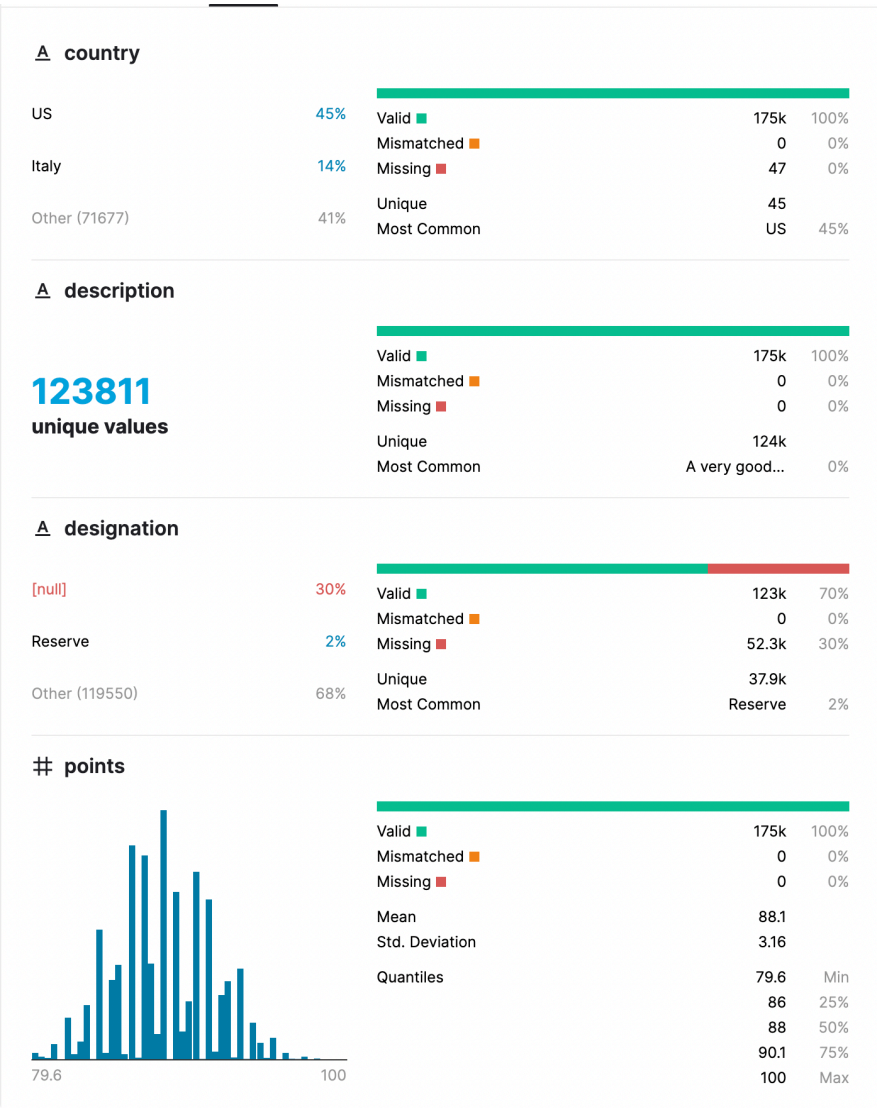
UCS663 Data Science
Lab Evaluation - I
Harsh Thakur
Roll number – 101916052
3CS10

Problem	AMMI Bootcamp Kaggle Competition
Problem Link	https://www.kaggle.com/c/nlp-getting-started
Problem Type	Linear Regression
Github Link	
Kaggle Link	https://www.kaggle.com/harshthakur178/mains-py
Libraries Used	Cuml, Cudf, Copy
Model Implemented	Linear Regression (algorithm used is SVD)
Evaluation metrics used	RMSE, MSE, R2 Score
Task Done in Code	Data cleaning and model implementation

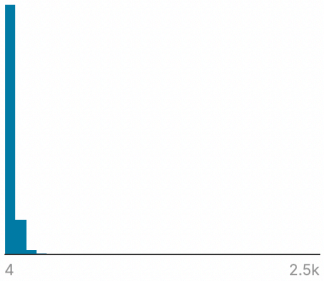
Data

There are two files provided train.csv and test.csv. The former is for training of the model and the latter is for making the predictions and then submitting for evaluation purpose.

Each sample in the train set has the following columns:



price



Valid	175k	100%
Mismatched	0	0%
Missing	0	0%
Mean	34.3	
Std. Deviation	38.4	
Quantiles	4	Min
	16	25%
	25	50%
	40	75%
	2.5k	Max

province

California	31%
Washington	7%
Other (107904)	62%

Valid	175k	100%
Mismatched	0	0%
Missing	47	0%
Unique	468	
Most Common	California	31%

region_1

[null]	16%
Napa Valley	4%
Other (139253)	80%

Valid	146k	84%
Mismatched	0	0%
Missing	28.5k	16%
Unique	1278	
Most Common	Napa Valley	4%

region_2

[null]	57%
Central Coast	9%
Other (59220)	34%

Valid	75.4k	43%
Mismatched	0	0%
Missing	99.6k	57%
Unique	18	
Most Common	Central Coast	9%

taster_name

[null]	63%
Roger Voss	8%
Other (51815)	30%

Valid	65.5k	37%
Mismatched	0	0%
Missing	109k	63%
Unique	19	
Most Common	Roger Voss	8%

taster_twitter_handle

[null]	64%
@vossroger	8%
Other (48496)	28%

Valid	62.2k	36%
Mismatched	0	0%
Missing	113k	64%
Unique	15	
Most Common	@vossroger	8%

title

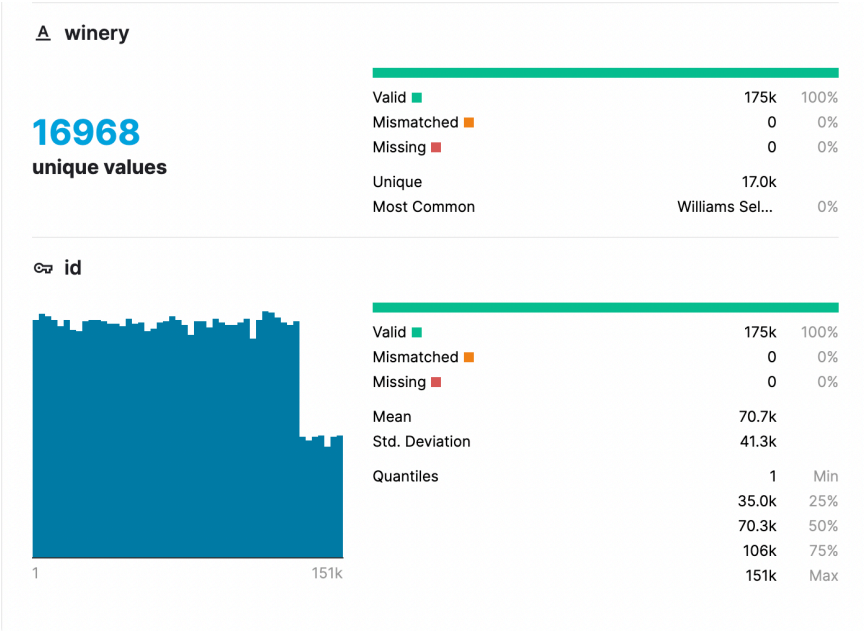
[null]	53%
Gloria Ferrer NV Sonoma Brut Sparkli...	0%
Other (82181)	47%

Valid	82.2k	47%
Mismatched	0	0%
Missing	92.8k	53%
Unique	77.4k	
Most Common	Gloria Ferrer...	0%

variety

PINOT NOIR	10%
CHARDONNAY	10%
Other (140124)	80%

Valid	175k	100%
Mismatched	0	0%
Missing	1	0%
Unique	706	
Most Common	PINOT NOIR	10%



There are a total of 175k entries and 14 columns, including the column that needs to be predicted aka the “price” column.
As the competition was completed 2 years, gaggle has stopped accepting submissions for the same so the accuracy of the following model will be evaluated using the test data made using train_test_split method (found in cuml.preprocessing)

Data

First we import the data and then we introduce the various data cleaning procedures such as encoding (using label encoder) and dropping NAN values. Then the linear regression model from viml was used as the model. The algorithm used is svd-jacobi and an r2 score of 41% was achieved.

Due to a large number of unique and null values, many rows and columns were dropped and label encoder was used to convert the string values to numerical. Then a self made standard scalar was used to scale the values between 0 and 1.
Due to the the competition being already completed, a valid submission could not be made, but if one were to made, the following would have been the result:
Previous submissions:

Team	Members	Score	Entries	Last	Code
Group 1- FCI		18.10164	11	2Y	
Group 8		18.27200	8	2Y	
Group11_YA		19.04578	11	2Y	
Group3		19.94217	10	2Y	
G2_AMMI_RW19		20.24618	6	2Y	
Group 4		20.47504	12	2Y	
Group_10		20.61727	7	2Y	
group5		20.97158	11	2Y	
Grace's Team - Group9		21.28934	7	2Y	
SAA-Group 15		21.71385	10	2Y	
over djm		22.19315	1	2Y	
Group 12-AMMI-RW		22.36314	5	2Y	
Group 6		23.17708	6	2Y	
Group Light (7)		23.72826	12	2Y	
Group 13		26.35297	7	2Y	
Group_14		32.18716	4	2Y	
Usman(worst submission)		37.22976	1	2Y	
Benchmark		41.43708			
HussamHassan		49.02042	2	2Y	

My submissions:

Submission and Description	Private Score	Public Score
Sample_Submission_NEWONE.csv 3 hours ago by Harsh Thakur 178 add submission details	34.66252	37.23361
Sample_Submission.csv 13 hours ago by Harsh Thakur 178 add submission details	36.17323	38.56175
Sample_Submission.csv 14 hours ago by Harsh Thakur 178 add submission details	36.17323	38.56175
Sample_Submission.csv 14 hours ago by Harsh Thakur 178 add submission details	41.43708	43.41744