Prediction of price based on the quality of the wine using different Machine Learning models

Emma Bats[2620868], Janey Kok[2623750], Chiara Kraag[2617517], Asif Anwar[2660561], and Joshua Reimer[2579500]

Vrije Universiteit, De Boelelaan 1105, 1081 HV, Amsterdam
e.bats@student.vu.nl, j.m.a.kok@student.vu.nl, c.a.kraag@student.vu.nl, a.anwar@student.vu.nl, j.k.reimer@student.vu.nl
Group 33

Abstract. This research paper aims to compare the price prediction of wine by different machine learning techniques; Multi-layer Perceptron (MLP), Gradient Boosted Regression (GBR) trees, Linear Regression (LR) and Support Vector Regression (SVR). A large data set of 141.617 wines with ratings between 80 and 100 out of 100 and prices in USD was obtained from kaggle.com [1], preprocessed, and split into training, validation and test sets. The validation results clearly indicate that the MLP neural network performs best, closely followed by the GBR. However, since the GBR trees have a tendency to overfit, the MLP technique is considered superior. Testing shows that it does not perform particularly well, achieving an R of only 51 percent, which may be explained by the filtering of the data set by high ratings and/or a lack of correlation between the features and the price of wines in general.

Keywords: Multi-layer Perceptron · Gradient Boosting Decision Tree · Linear Regression · Support Vector Regression.

1 Introduction

When you buy a bottle of wine, you need to pay for it. But is the price fair given the quality? In this research, the aim is to predict the price of wine based on the quality and origin data, using different Machine Learning (ML) algorithms: linear regression (LR), support vector regression (SVR), multi-layer perceptron (MLP) regression and gradient boosted regression (GBR). The data that is used to train the Machine Learning algorithm is from Kaggle [1] and contains a list of 141,617 different wines and the following features: country, province, region_1, region_2, description, quality, title/name of the wine and its given price. The final aim is to find the model that best predicts the price of wine based on the available features. Therefore the main research question discussed in this paper will be:

Which machine learning algorithm is most efficient for predicting wine pricing ranges based on the quality of the wine based on the feature data set?

2 Data inspection and preparation

2.1 The Data set

The data set contains 15 different features in total and 141.617 observations. These features are: unnamed, country, description, designation, points, price, province, region_1, region_2, taster_twitter_handle, title, variety, winery and vintage (year of wine). The following features are removed since they don't have a correlation with either wine quality or wine price: unnamed, taster_twitter_handle and title. After further inspecting the data frame, it now contains 12 features and 11,2% in missing values across all features. The data contains three numeric features, of which 'points' is indicated as an integer, while 'price' and 'vintage' are represented as floats. The other seven features are categorical and classified as object types. The categorical features are: description, designation, province, region_1, region_2, variety and winery and have a high cardinality. The price is highly skewed and the features 'designation' and 'region_2' have the most missing values. The data contains 45 different countries, the most frequent of which are: US, France and Italy. The feature 'points' is approximately normally distributed with a standard deviation of 3.0793 and a mean of 88.566 (see Appendix Fig. 1).

2.2 Correlations

The data set is tested for correlations between features using both the Pearson and Spearman correlation coefficients.

The Pearson correlation is a coefficient of the mean of the slopes of 2 linear lines. These lines, l1 is the linear regression of y on x and l2 is the linear regression of x on y, both have a slope. The coefficient is the geometric mean of these 2 slopes, with a range from -1 (negative relationship) to +1 (positive relationship). In Fig. 2 (see Appendix), the positive relationship is red, while the negative relationship is blue. The relationship between the same 2 features is always equal to +1.

The Spearman correlation is a coefficient used to measure the extent to which 2 variables are related with each other. This correlation is especially useful when 2 variables don't have the same scale, using Spearman, these variables can be moved to the same scale. The coefficient has the same ranking and interpretation of this ranking as the Pearson correlation (see Fig. 3 in Appendix) [2].

Both indicate a correlation between the features 'points' and 'price' and 'vintage' and 'price'.

2.3 Feature Extraction

In total, only 61 rows can be observed not having a value for 'country'. Considering the size of the data set, these values will be omitted. The 'designation' feature has 43.245 missing values in total, equal to 28.4%. The feature contains mixed information for marketing purposes, which are both too unique and too unreliable to use, therefore the feature will be dropped. The feature 'points' does not have any missing values. Initial analysis shows that 'points' is an important feature in the data set. It has a linear relationship with the 'price' feature and will be vital for the machine learning models (see Fig. 4). However, the points are only 80 or higher as the data set only contains wines rated 80 points out of 100, or higher. The feature 'price' has 6,8% missing values, which will be omitted since 'price' is the target feature. The missing values are linked to French, Italian and Portuguese wines.

2.4 Creating Feature Classification

For the machine learning algorithm classification, a categorical feature for 'price', 'price category', was introduced, dividing it equally into five categories: bronze, silver, gold, platinum and diamond. The categories have been given bounds based on an equal division of wines, sorted by 'price' (see Fig. 5). Within the 'price' feature there are some outliers (see Fig. 6): a total of 69 wines that have a price above 500 US dollars. These rows have been removed to achieve a better rendering of the distribution.

2.5 Handling missing data

The 'province' feature has no missing data and was confirmed on inspection to contain valid data. Hence, no action is required for this feature. The 'region_2' feature is empty for nearly 50% of all entries. There are two options to solve this. The feature can be removed, or the null values can be replaced with the values of 'region_1', of which the latter option was chosen. For the feature 'taster_name' there is no good method for replacement, so the missing were all replaced with "others". If these replacements have any impact on the price, this feature must be deleted. After all these replacements and cleaning of the data, the data frame does not have any missing values and outliers are bounded to the acceptable range of 0-500 USD.

2.6 Data Analysis and Features Encoding

After cleaning the data, a further analysis was performed to better understand the data. The average price was compared to the wine description length (see Fig. 7), and a strong correlation could be observed: the longer the description, the higher the price.

Feature encoding is the process of changing a categorical feature into a numeric feature. This is useful when a feature is highly dimensional and has the advantage of being compatible with numeric methods [3]. The feature encoding is performed for the categorical 'country' feature, increasing the number of features by 43, the number of countries. The next feature is 'province', for which we create a dictionary relating the province to its number of occurrences in the feature. The same method is then used for the other categorical features 'region_1', 'region_2' and 'winery'.

2.7 Split into Train, Test and Validation data

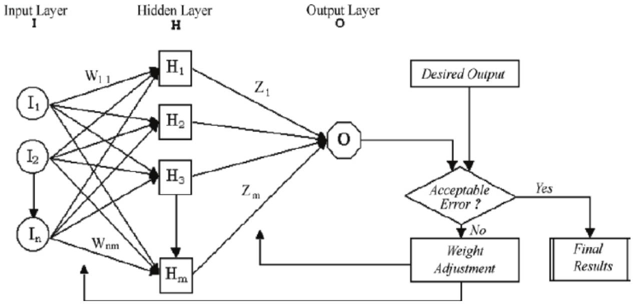
The original data does not contain a train and test set. Therefore that data has been split into arrays in random train and test subsets. This is a Quick utility that wraps input validation and application to input data into a single call for splitting. For the scaling and normalization of the data several utility functions transform the classes into a new feature vector that is more suitable for the downstream estimators. The data set will be transformed into standardization and this provides benefits for the learning algorithms. The outliers that will exist in the data frame will be more appropriate. The standardization process is required for many machine learning estimators since we want to predict the price [4].

3 Experimental Setup

This research makes use of 4 different models: Multi-layer perceptron, Gradient Boosted Regression tree, Linear Regression and Support Vector Machine regression. Each model will be explained in detail below. Every model used the same train set and eventually the same test set.

3.1 Multi-layer Perceptron

Also known as MLP, the multi-layer perceptron is a model that contains multiple layers of neurons that communicate using weighted connections between neurons. This network of layers consists of 3 parts, the input layer, the hidden layer and the output layer. There are no connections between neurons of the same layer, but the neurons are fully connected to neurons in adjacent layers. The weights of the weighted connections indicate the extent of a correlation between neurons. The picture below shows the different layers of neurons and how they are connected to each other. [5, 6].



3.2 Gradient Boosted Regression Tree

The second model is the Gradient Boosted Regression tree, which is a machine learning algorithm used for supervised problems like classification and regression. The algorithm is based on decision tree models and uses a method called boosting wherein different models are combined into one model. In this technique, the focus is on iteratively training predictors to learn from the mistakes of the previous predictors. The algorithm produces an ensemble of a weak prediction model into another, stronger prediction model. This is done to predict the target label of the decision tree wherein the most trees have a depth larger than 1 [7].

3.3 Linear Regression

Linear regression is about the relationship between the dependent variable and one or more independent variables (features) in a linear way. This linear regression is defined as a set of regression parameters and a random variable, each of which are assigned a constant weight per feature. There are multiple options to determine the random variable, of which the least-squares method is the most common. A good model then has parameters that minimize the cumulative sum of the least-squares errors [8].

3.4 Support Vector Regression

The fourth and final model is Support vector regression (SVR), a supervised machine learning model that uses the regression algorithm. An SVR training algorithm builds a model that predicts new values to one category. A support vector machine then takes the data points and outputs and the hyperplane will separate the tags. This hyperplane will be the decision boundary. The SVR finds a hyperplane to fit within the data and define the acceptable error in the model. The SVR minimizes the coefficients (as l2-norm of the coefficient vector) and the error term is handled in the constraints, where the absolute error is set to the maximum error(epsilon). The epsilon is then tuned to gain the desired accuracy for the model with the use of the formula [9, 10]:

Minimise:

$$MIN\frac{1}{2}\left\|w\right\|^2$$

Constraints:

$$|y_i - w_i x_i| \leq \varepsilon$$

3.5 Model evaluation

In order to evaluate the performance of the models, the performance scores below were used. Each matrix/score is described below.

The coefficient of determination R^2 is well established in classical regression analysis. Its definition as the proportion of variance 'explained' by the regression model makes it useful as a measure of success of predicting the dependent variable from the independent variables. The value remains within 0 to 1 where close to 1 means less residual error [11, 12].

Finding the residual errors is a good way to evaluate the models performance in machine learning projects. The residual refers to the gap between target value and predicted value. There are different measures of the residual errors. In this test the mean squared error (MSE), mean absolute error (MAE) and square root mean squared error (RMSE) were used. The formulas of each score are shown below [11, 12].

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

$$\mathrm{MSE} = \frac{1}{m} \sum_{i} \left(\hat{\boldsymbol{y}}^{(\mathrm{test})} - \boldsymbol{y}^{(\mathrm{test})} \right)_i^2$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}$$

4 Results

4.1 Performance of the models

Table 1 shows the performance of each model based on the selected performance matrices by using the validation data set. From the table, it can observed that the Linear Regression (LR) model performs poorly against all evaluation measures, the Gradient Boosted Regression (GBR) model has performed relatively well, the Support Vector Regression (SVR) model shows poor performance for all measures similar to the LR model, while we can notice that values are better than Linear models and similar with the Multi-Layer Perceptron (MLP) performs slightly better than the GBR model for all performance measures.

Model	R^2**score**	**MSE**	**RMSE**	**MAE**
LR	0.288218	730.265497	27.023425	16.200241
GBR	0.533152	478.970795	21.885401	12.361965
SVR	0.285782	732.763829	27.069611	13.416374
MLP	0.549919	461.768181	21.488792	12.142759

Table 1. Performance results of MLP, GBR, LR, SVR models using validation data set with different performance metrics.

4.2 Performance of the test data

Since the MLP model performs best, it is the best choice for a final model. While the GBR model performs similarly, it performs slightly worse for all measures, and decision trees have a higher likelihood of overfitting.

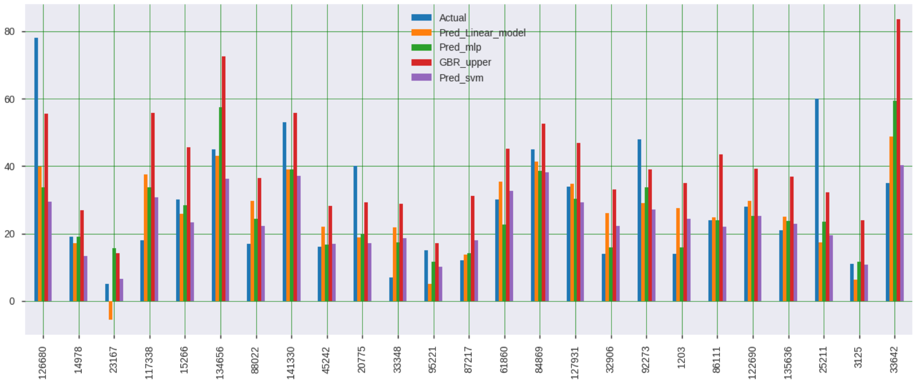
Table 4.2 shows the performance of the MLP model with the test dataset. The MLP model performs slightly worse when looking at the determination coefficient, (51% vs 55%), but performs better for the error measures MSE, RMSE and MAE.

Model	R^2**score**	**MSE**	**RMSE**	**MAE**
MLP	0.510660	185.325107	13.613416	9.807378

Table 2. Performance results of MLP model with the test data

5 Discussion

The overall performance of all models is poor, both in terms of explanatory power and for measures of residual error. The figure below shows the results of all the models. The x-axis represents the different wines (id-name), the y-axis represent the price of the wine.



None of the models achieve a determination coefficient above 51% during testing, and the residual errors are quite high as can be seen from table 4.1.

The MLP model performed moderately well in the test data set compared to the validation set for most of the measures of residual error. Although the R^2-score is only 51% for the MLP model, the MSE has improved by almost 250 points, indicating a significant improvement despite its low explanatory power.

For future studies, a complete data set could be used, i.e. including wines with ratings below 80 points. The MLP neural network performed best for this data, hence it can be assumed to be the most suitable type of model for this data set. Instead of finding exact wine prices, the model could be redesigned to predict price categories, which might improve performance and be of more practical use. If combined with a quality predictor based on chemical composition of a wine, the model might serve to suggest a pricing range for vendors and buyers of wines.

6 Conclusion

This paper attempted to answer the following research question:

Which machine learning algorithm is most efficient for predicting wine pricing ranges based on the quality of the wine based on the feature data set?

When training the four models, various measures of error can be used to intuit which model best predicts the price of wine: the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE) among others. For all these measures, the multi-layer perceptron (MLP) neural network at 100 layers and 500 iterations performs best, followed closely by the gradient boosted regression tree (GBR). The linear regression (LR) model and support vector regression (SVR) model achieve nearly identical scores for each metric. The results of the test data set shows that the multi-layer perceptron (MLP) neural network works the best. The performance of the measurement gives the best performance. Although, the result is close to GRB, which has a tendency for overfitting.

Overall, the data set is big, which explains why the results of the training set and the results of the test set are almost the same. This statement only holds for the models which were selected for testing with the test set.

Coming back to the research question, none of the models performed remarkably well. It can be concluded that MLP is most suited compared to GBR, LR and SVR, and that features such as rating, geographical origin and vintage year do not have significantly predict the price of the wine.

However, this research has some limitations. The data set only contains only high ranking wines (80-100 out of 100), meaning it does not contain any common, mediocre or bad wines. Hence, the conclusion can be said to hold only for good to excellent wines.

Another limitation is the number of models used in this research. Due to time restrictions, only four different machine learning models were tested and the search for optimal hyperparameters limited. Given more time to select a better model and fine-tune its hyperparameters, better results might be achieved.

References

1. Wine Reviews, https://www.kaggle.com/zynicide/wine-reviews. Last accessed 3 March 2020
2. Artusi, R., Verderio, P., Marubini, E.: Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval. The International journal of biological markers **17**(2), 148–151 (2002)
3. Chowhan, S. S., Shinde, G. N.: Evaluation of statistical feature encoding techniques on iris images. In 2009 WRI World Congress on Computer Science and Information Engineering **7**, 71–75 (2009)
4. Raschka, S.: About Feature Scaling and Normalization (and the effect of standardization for Machine Learning algorithms). Polar Political Legal Anthropology Re **30**(1), 67–89 (2014)
5. Ruck, D. W., Rogers, S. K., Kabrisky, M.: Feature selection using a multilayer perceptron. Journal of Neural Network Computing **2**(2), 40-48 (1990)
6. Khalafi, H.: A Literature Survey of Neutronics and Thermal-Hydraulics Codes for Investigating Reactor Core Parameters. Artificial Neural Networks as the VVER-1000 Core Predictor **6**(5), 103–122 (2011)
7. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma W., Liu.: Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems **2**(5),31469–4154 (2017)
8. Yan, X., Su, X.: Linear regression analysis: theory and computing. World Scientific.(2009)
9. Chun-Hsin W., Jan-Ming H. Lee D.T.: Travel-time prediction with support vector regression. Transactions on Intelligent Transportation Systems **5**(4), 276–281 (2004)
10. Ni, K. S., Nguyen, T. Q.: Image superresolution using support vector regression. Transactions on Image Processing **16**(6), 1596–1610 (2007)
11. Nagelkerke, N.J.: A note on a general definition of the coefficient of determination. Biometrika, **78**(3), 691–692 (1991)
12. Chai, T. and Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development **7**(3), 1247–1250 (2014)

7 Appendix

Link to code:
`https://colab.research.google.com/drive/1mHKtlWpgIvtdlgSVymJxRXbLwH1jwViA`

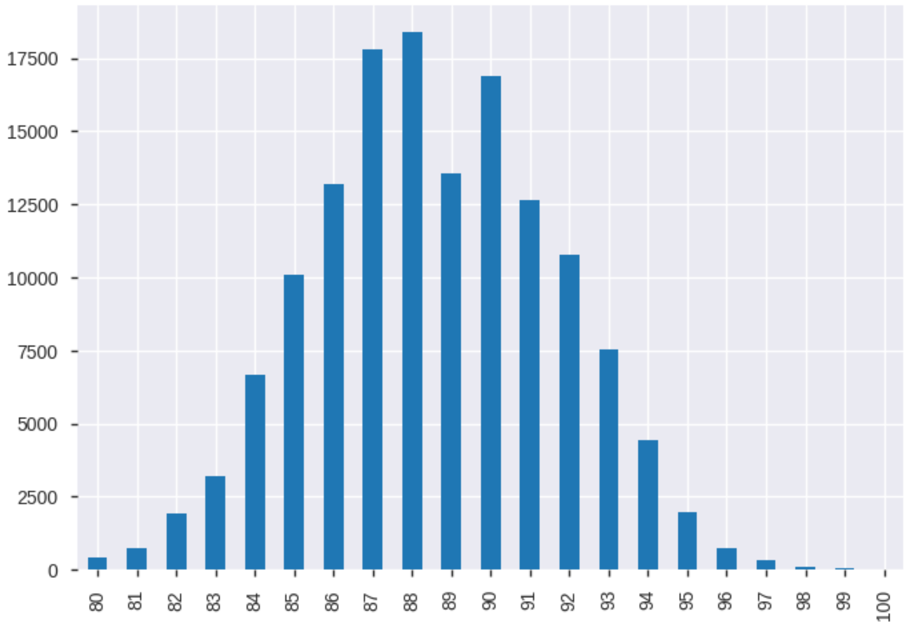


Fig. 1. Distribution of points

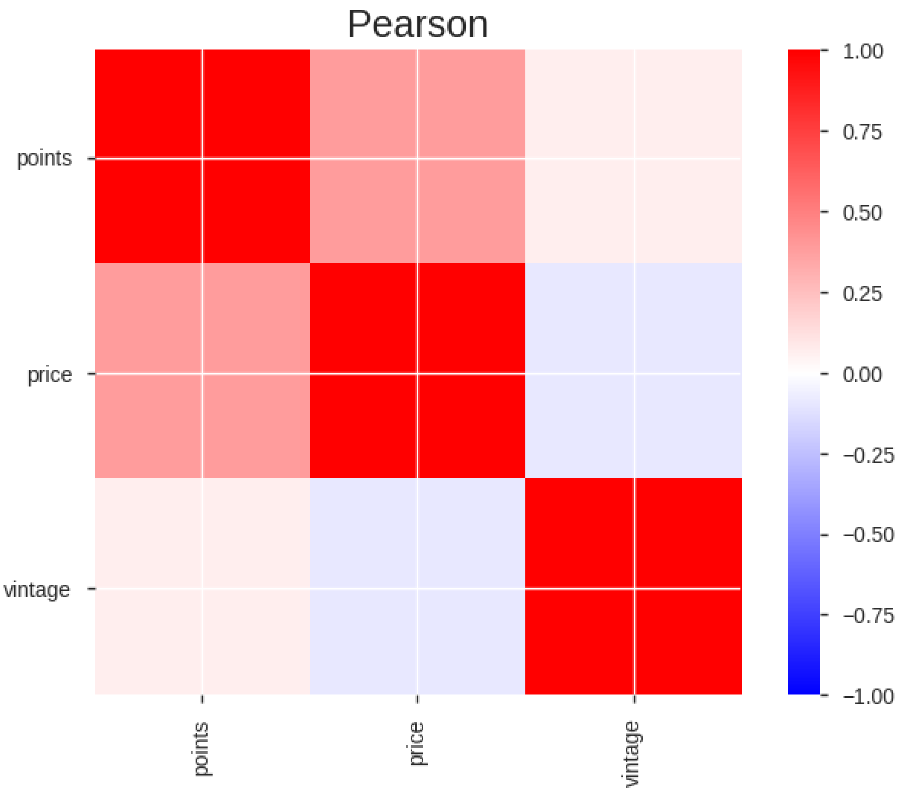


Fig. 2. Pearson correlation

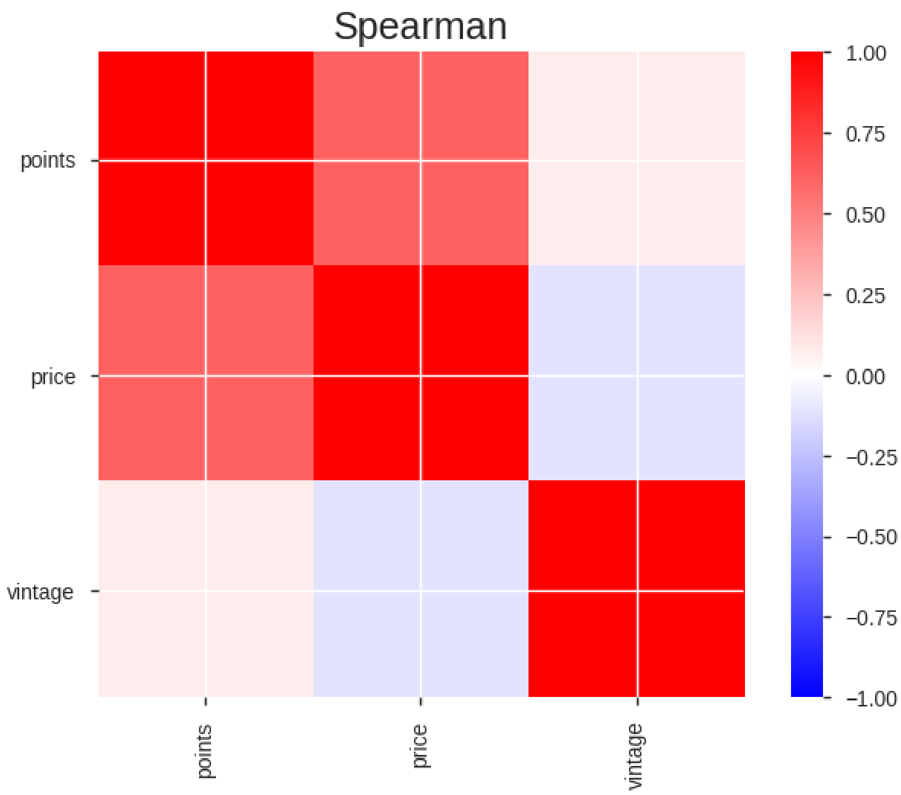


Fig. 3. Spearman correlation

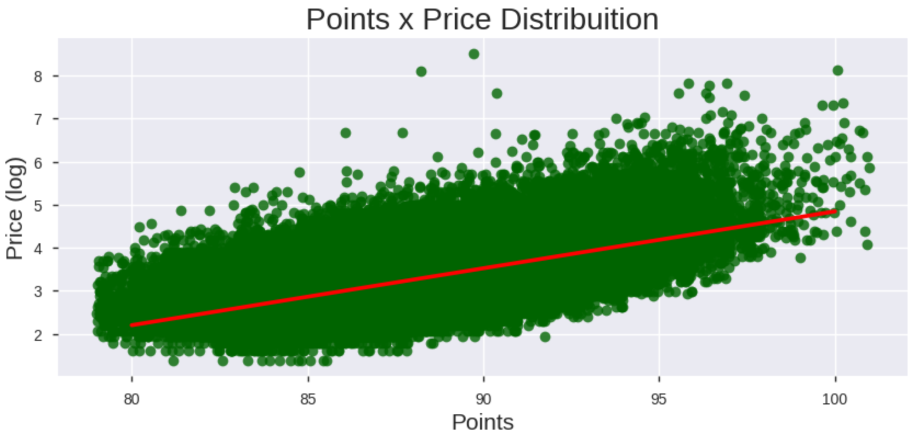


Fig. 4. Distribution Points and Price

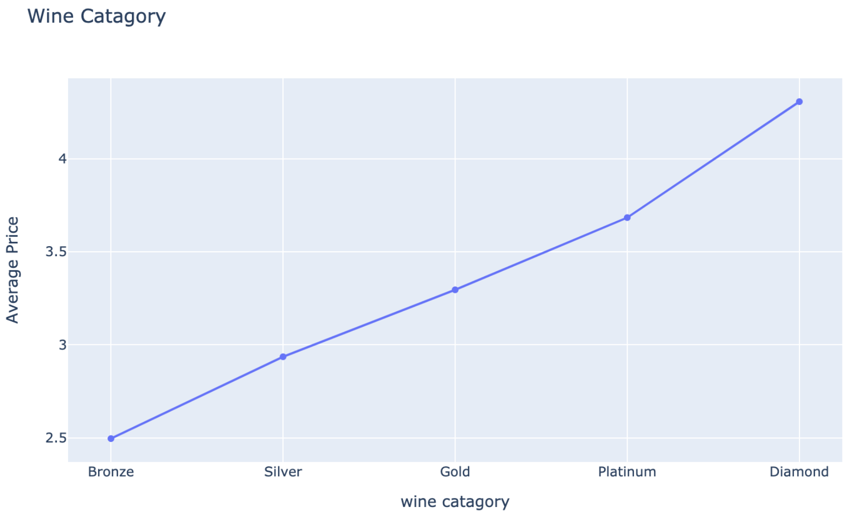


Fig. 5. Distribution of Classification

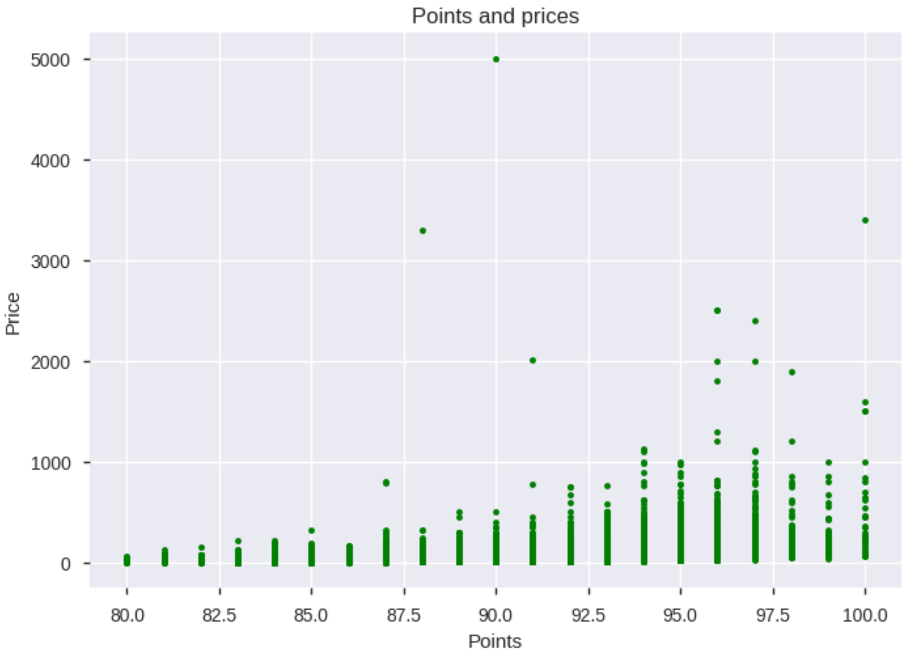


Fig. 6. Outliers of the price

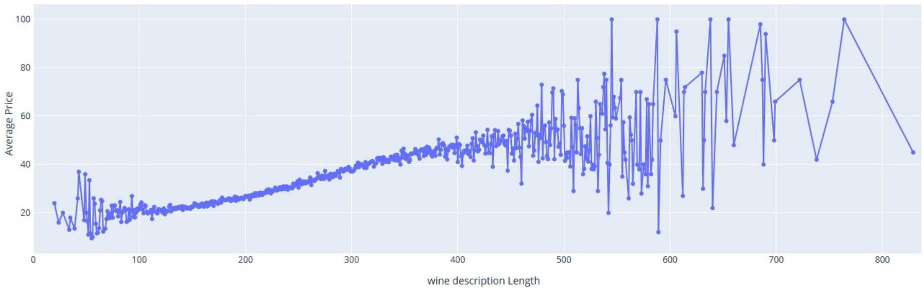


Fig. 7. Average Price (y-axis) by Wine description length (x-axis)