Bachelor's Thesis Computing Science

Radboud University Nijmegen

Predicting wine prices using weather data

Author:
Elianne Heuer
s4320514

First supervisor/assessor:
prof. dr. ir. A.P. de Vries

Second assessor:
dr. Y. Shapovalova

August 20, 2023

Abstract

Wine is a very popular drink with a large associated market. Previous research has observed the influence of weather on the price and quality of wine. Especially, previous research has shown that winter rain and a warm and dry growing season is beneficial for wine quality, while rain in the harvest period has a negative influence on quality. In this research the correlation between weather and wine quality was analysed and a predictive model based on support vector regression for the wine prices and ratings of French red wines was created. To start, a wine dataset containing data from the website Vivino has been coupled with weather data from 2016, 2017 and 2018 from the North West (NW) of France and the South East (SE) of France, based on their respective locations. The wine data contained, for each wine, the name, region of the grapes, year and rating, while the weather data contained ground station measurements of precipitation, temperature, wind speed and humidity. The correlation between wine price and rating and the weather data was calculated. In the results the assumed positive influence of winter rain was indeed observed for wines from the NW of France, but not for wines from the SE of France. A positive correlation between temperature in the growing and harvest season has been observed for wines made in the SE of France, but not for wines made in the NW of France. Additionally, a negative correlation between rain in the harvest season and wine quality and price was not observed. The predictive model based on support vector regression only performed well on the data from the NW of France and predicting the natural logarithm of the wine price. The predictive models for wine price performed mediocrally and the predictive models for wine ratings performed even worse. In future research these models may be improved by using more accurate weather data of more years and wine ratings from a different source.

Contents

Chapter 1

Introduction

Wine. An incredibly popular and (sometimes) very expensive drink with a huge culture and economy surrounding it. And where some people just buy a bottle of wine now and then to combine with a nice dinner, other people invest their life savings into buying what they think are good bottles of wine, to keep these wines for a few years in a temperature and humidity controlled storage, so they (hopefully) have significantly increased in price and can be sold at a profit margin large enough to make the effort of keeping this wines in heavily controlled conditions worth it. But how do they choose which wines to keep? Often you hear people (jokingly) say "Oh 2016, a good wine year", but is there such a thing? Or is there no such thing as a "good year" and does the wine price or quality get determined by other factors?

Thinking about this relation leads to the following research questions: Is there a correlation to be found between weather data and wine prices or wine ratings? And, is it possible to predict red wine prices and ratings based on weather data?

Since investing in wine can seem quite arbitrary, you buy a wine, keep it for some time and hope for the best - maybe you look at a review of a wine expert of this wine, but who knows if they are right? - many people in this investment business would be interested to know if there are more accurate predictions of wine quality before they invest in a wine. In this day and age, machine learning and AI has often been used to predict all sorts of things based on gathered data. So if these techniques could be used to find a link between a "good year", i.e. there was a year with good weather for wine, and wine price or quality, then investing in wine could become a less daunting decision and maybe not as high risk of an investment.

1.1 Related work

In previous research many factors that influence wine price and wine ratings have been laid out. Where ratings significantly reflect specific weather conditions that also determine the quality of a wine [5, 20, 11], price is a bit more complicated. Wine price is influenced by the age of the wine, expert ratings (minor influence), the number of bottles of that wine on the market, harvest yield, the region the wine is from, the grapes that were used and the reputation of the winery that produced the wine [11, 20]. However, Ginsburgh, Monzak and Monzak (2013), as well as Ashenfelter (2008) have determined that the variation in average prices for Bordeaux and Medoc can be, respectively, for 66% and 80% be explained by fluctuations in weather variables, making them the most important contributing factor in quality and price determination [3, 14, 4]. The idea that weather has an impact on wine quality has been studied several times: some of these studies focus on specific iconic wines and examine their vintage variations [7]; a few examine multiple wines in a single region over time [23]; some look at the impact of different weather factors on a particular grape variety [15]; and others have looked at the matter in different countries [17].

Several papers define the weather variables that are an important factor in determining wine quality. These weather variables are generally defined to be the following: the amount of rain in the winter months from October to March (dormant period), the amount of rain in August and September (harvest season) and the average temperature during the whole growing season from April to September [3, 7]. There are some small additions to the influential weather variables in some prior studies: Corsi & Ashenfelter (2019) split this last variable into two variables: the average temperature in spring from March to July and the average temperature in summer from August to September [10]; Ramirez (2008) uses the average temperature from April to May, from June to July, August to September and the average precipitation of January and February, April and May, June and July and August and September [23]; Oczkowski (2016) also defined the average difference of minimum and maximum temperature per day during the growing season as an influential variable, as well as wind direction and strength [22].

An optimal climate for growing grapes that will produce high quality wines, for Bordeaux at least, has the following characteristic: the growing season gets preceded by a wet winter, this is then followed by a warm spring and summer, additionally the summer was also a dry one [10, 3, 14, 16]. Furthermore, in Greece it has been observed that wine quality ratings become higher when the maximum temperature is higher and the conditions are dry during the growing season [19], but on the other hand, other wine regions have an optimal growing season temperature and beyond this temperature,

the quality of the wine decreases again [17]. Another negative factor to wine quality, which was observed in Germany, could be the number of days the soil freezes in the growing season and in the harvesting period [21]. It was also observed in Germany that the quality of the wine increases with a rising trend of average and minimum temperatures during the growing seasons.

So, different research projects have analyzed the relation between wine price and quality and weather, often using regression methods, for the wines from the Napa region in California [23], the wines from Australia [25, 22], the wines from Germany [6, 21], the wines from the Bordeaux and Médoc regions in France [5, 16, 9, 4, 14], the wines from Switzerland [20], the wines from Greece [19] and all over the world [17]. Others have also tried to make a predictive model. Some of these predictive models were based on multiple linear regression and predicted both the wine price and rating [7, 3] whereas the model created by Corsi and Ashenfelter used an ordered probit model and only predicted the wine ratings. These models obtained a good prediction of the wine prices over a longer time and the model created by Corsi and Ashenfelter obtained an average accuracy of 0.5 in predicting wine ratings. In more recent times, researchers have started to use machine learning to predict wine prices and quality. Yeo, Fletcher & Shawe-Taylor (2015) have used Gaussian process regression and multi-task-learning to predict wine prices, their best model achieving an accuracy of 0.7 [27]. Roucher, Aristodemou & Tietze (2022) also created a predictive model for long term wine prices based on weather parameters using Local Least Squares kernel regression (LLS) [24].

1.1.1 Hypothesis

In this study, I analyze the correlations between different weather variables and short term prices and wine ratings. I use these variables to create a predictive model using SVR (Support Vector Regression) for wine prices and a predictive model for wine ratings. Because of the heavy influence of weather on wine prices, I expect to see a positive correlation between amount of rain in the winter & the temperature in the growing season and ratings & prices. Furthermore, I expect to see a negative correlation with rating and price and the amount of precipitation in the harvest season. Since hard wind can damage vines, I expect a reduction in wine quality if a vineyard has suffered a storm. Additionally, humidity can also lead to mildew in vines, loss in harvest yield and thus also loss of quality. For the predictive model I expect that the price and rating can be predicted for different wine types.

In the following chapter (chapter 2) I discuss any preliminary knowledge you need to understand this paper. I describe my research process in chapter 3, including the data description, data processing, correlation results and

the results of the predictive model. I finish with a conclusion and discussion in chapter 4.

Chapter 2

Preliminaries

2.1 Haversine distance formula

The Haversine distance formula is used to calculate the direct distance between two positions on a sphere. It uses the latitudes and longitudes of these two positions to calculate the distance. The haversine distance formula is given in Equation 2.1:

$$d = 2r\mathtt{sin}^{-1}(\sqrt{\mathtt{sin}^2(\frac{\Phi_2 - \Phi_1}{2}) + \mathtt{cos}(\Phi_1)\mathtt{cos}(\Phi_2)\mathtt{sin}^2(\frac{\lambda_2 - \lambda_1}{2})}) \quad (2.1)$$

where r is the earth's radius, Φ_1 is the latitude of position 1, Φ_2 is the latitude of position 2, λ_1 is the longitude of position 1, λ_2 is the longitude of position 2. The result d is the distance in km between point 1 and 2 [13]. Because the formula is calibrated on a perfect sphere, but the earth is an oval shape, the formula is not completely accurate.

2.2 Correlation

Correlation refers to an association between two variables. For example, people who are taller tend to have more body weight, this is a positive correlation: as one variable gets higher, the other does too. The correlation strength has a scale from 1 to -1, where 1 refers to a perfect positive correlation, 0 means no correlation and -1 refers to a perfect negative correlation, see Figure 2.2. There are different types of correlation: Pearson's correlation, Kendall's correlation and Spearman's correlation. Pearson's correlation is a measure of linear association, whereas Kendall's correlation and Spearman's correlation are both meant as a measure of correlation for ordinal/ranked data. In this research we have used the Pearson correlation coefficient, which is calculated using the formula given in Equation 2.2:

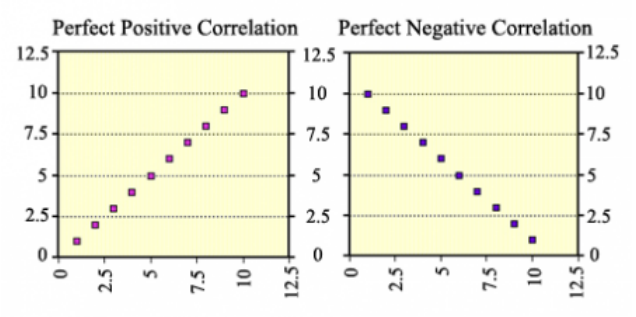

Figure 2.1: Graphs displaying two perfect correlations, left a perfect positive correlation ($r = 1$), right a perfect negative correlation ($r = -1$) [8]

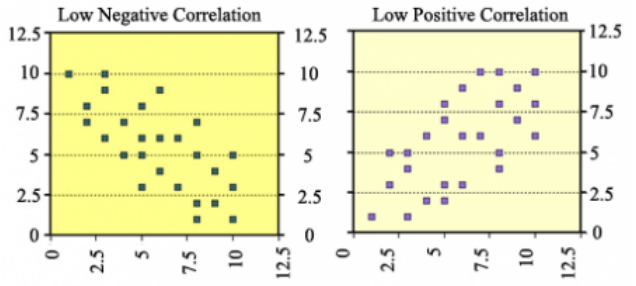

Figure 2.2: Graphs displaying two weak correlations, left a weak positive correlation, right a weak negative correlation [8]

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}} \tag{2.2}$$

Where x and y are two vectors of the same length, m_x and m_y are the means of x and y, respectively. If r is either greater than 0.5 or smaller than -0.5 the correlation is considered to be strong, an r value between 0.3 and 0.5 or -0.3 and -0.5 is moderate and an r value between 0 and 0.3 or 0 and -0.3 is a weak correlation [26].

The significance of a correlation between two variables can be tested by using the t test, where a significant correlation indicates that the correlation is not found by random chance. The result of this t test is a p-value. If this p-value is smaller than 0.05 it can be stated that a relationship between two variables is statistically significant, which indicates that they are not independent assuming both variables follow a normal distribution.

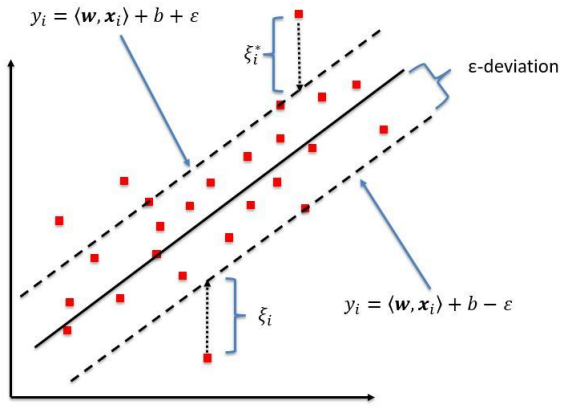

Figure 2.3: Graph displaying an epsilon insensitive tube (between the dotted lines) [18]

2.3 Support vector regression

Regression is both a descriptive and predictive data analysis and prediction method. If there are two correlated variables, one independent variable x and a dependent variable y, y can be expressed in x. Regression is a method to create a formula where variable y is expressed using variable x, this is called the regression equation. Using this regression equation values of y can be predicted for new values of x. Regression can be used to predict a continuous variable. For prediction tasks regression can be combined with support vector machine. The Support Vector Machine (SVM) is an algorithm used for classification and regression, and it can analyze linear and nonlinear relations between variables. If it is combined with regression it can be used to predict continuous variables, whereas more basic methods of support vector machine can only be used to predict binary or categorical data.

Support Vector Regression (SVR) tries to find a line that best describes a relationship between two variables, which is the same as linear regression, but for SVR this line is fit by also using several value thresholds, called an "epsilon-insensitive tube" (see Figure 2.3). This tube contains the maximum error the model is allowed to have; if datapoints lie outside this tube, then they are labeled as support vectors. The tolerance for these points can be tuned by a hyperparameter C. Additionally, the line to which the data is fit, does not have to be linear. SVR uses a kernel, that can be linear, polynomial or RBF (nonlinear). The RBF kernel has an additional tune-able parameter: Gamma. Gamma determines the influence of a single training sample. To evaluate the performance of a predictive model two variables are important: R-squared and the root mean squared error (RMSE). R-squared represents

how well the regression model fits the data. R-squared is a value between 0 and 1; if the value is 1, it means that all the variability in the target variable can be explained by the model; if it is 0, then none of the variability is explained by the model [12]. The mean squared error represents how close the regression line is to the data points. It is calculated using the following equation:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \tag{2.3}$$

Where n is the number of data points, y_i is a observed value and $\hat{y}_i$ is the accompanying predicted value. From the mean squared error we can calculate its square root, the root mean squared error (RMSE), which provides an estimate of the error through the average distance between actual values of the datasets and their accompanying predicted values [28].

Chapter 3

Research

3.1 Data

3.1.1 Data Gathering

Two types of data had to be identified for this study: weather data and wine data.

Weather data of a wine country had to be found. The most suitable weather dataset was a dataset with weather data of France, gathered by Météo-France [2]. Météo France is the national French meteorological institute, who share part of their data via their open source initiative, MeteoNet, a website created for Data Scientists to experiment with data science methods on weather data.

Wine data was gathered via two different methods. The first part of the final dataset was downloaded from Kaggle, the data was gathered using a webscraper on the popular wine website Vivino. On Vivino people can order wines online from different wine vendors and they can give specific wines a rating and a review. To expand this first dataset, I also gathered data from Vivino with my own webscraper. This webscraper was written in Python with the libraries `requests`, `json`, `urllib.request` and `pandas`. The scraper sends a search request to Vivino with specific constraints: the country of origin of the wine should be France, the results should be ordered by price in ascending order, a variable minimum price is given and the maximum price of the wine can be 500 euros. After sending these requests it receives a json file with the results from the website, then these results are written into a pandas dataframe.

3.1.2 Data Description

Weather data

The weather dataset contains weather data from 2016, 2017 and 2018 and has been split into one folder for the NW of France, anod another one for

the SE of France. Both folders have the same structure and contain data split into 5 main categories: data gathered from ground observations, data gathered by radar, data gathered by satellite and parameters for weather models. In this research we only use the data gathered from the ground observations. Ground observations are measured by observation stations or "ground stations", these have several sensors for measuring different weather parameters: wind direction, wind speed, precipitation, humidity, dew point, temperature and pressure. Every 6 minutes each weather parameter is measured. Every CSV file in this category contains ground observations from one year, so there is one file with data from the year 2016, one from the year 2017 and one from 2018. The dataset contains the columns described in Table 3.2.

Wine data

The wine data was collected from the website Vivino, as described in the previous subsection. Since the website Vivino does not have a filter for collecting wines from a specific year, the data of red wines were gathered from all years. Scraping resulted in a dataset with the columns described in Table 3.3. The dataset from Kaggle has the same columns minus the "wine brand name" column. The data on Kaggle contained separate datasets for red, white, rose and sparkling wine, but in this research I habe only collected a red wine dataset.

Table 3.1: Data description of wine data from the webscraper.

Column	**Description**
wine brand name	The name of the wine brand
wine name	The name of the wine bottle
year	Harvest year of the grapes of the wine
region	The grape growing area
winery	The name of the winery that bottled the wine
price	The cost of a bottle of the wine in euros
rating	The average rating of the wine (on a scale from 1 to 5)
no. ratings	The total number of ratings of the wine
country	The country the wine is from

Table 3.2: Data description of weather data from ground stations [1].

Column	Description	Unit
number_sta	Ground station ID	–
lat	Latitude NW France: $46.28 < \text{lat} < 50.96$ SE France: $41.37 < \text{lat} < 46.23$	decimal degrees (10^{-1} °)
lon	Longitude NW France: $-5.06 < \text{lon} < 2.0$ SE France: $2.0 < \text{lon} < 9.54$	decimal degrees (10^{-1} °)
height_sta	Station height	meters (m)
date	The moment of measurement	format: 'YYYY-MM-DD HH:mm:ss'
dd	Wind direction	degrees (°)
ff	Wind speed	m.s^{-1}
precip	Precipitation between current date and previous date	kg.m^2
hu	Humidity	percentage (%)
td	Dew point	Kelvin (K)
t	Temperature	Kelvin (K)
psl	Pressure reduced to sea level	Pascal (Pa)

3.1.3 Data Cleaning

One of the first steps that had to be taken, was cleaning up the data. The wine datasets were loaded into Jupyter notebook and concatenated. Next, the wines not made in France were filtered out and the dataset was split into separate datasets for wines made in 2016, 2017 or 2018. Then the library GeoPy was used to find the location, latitude and longitude of each of the wines. Some locations could not be found by GeoPy, so these had to be changed manually. The next step was filtering out locations with latitudes and longitudes falling outside of the scope of the weather data, see Table 3.2. The results were put into new CSV files.

Then the weather data was cleaned. Quite a few columns contained null values, so these were filled by first sorting the dataframes by ground station ID, then backward filling the null values and to be sure all values were covered the remaining null values were forward filled.

3.1.4 Data Processing

Processing weather data

We exclude the following weather parameters: wind direction (already reflected in humidity, rainfall and temperature), dew point (also expressed using temperature and humidity), and pressure (already captured by temperature and rainfall). So we consider wind speed, humidity precipitation and temperature, since these parameters have also been marked as affecting grape growth and quality in previous research. Should we confirm the findings from prior research, we can always consider other variables for inclusion. From the wind speed we include the maximum wind speed in the growing season, since a storm can damage grapevines and reduce the yield. High humidity for a longer period can cause mildew or other molds on grapes, which reduces the quality of grapes and the yield, so we look at the average humidity for each month. For temperature and precipitation we will mainly look at the average temperature and total amount of precipitation in the growing season, the harvest season and, for precipitation, we consider the amount of precipitation in the winter months, since a warm and dry harvest season, warm growing season and wet winter are all related to higher quality wines. For the temperature we further calculate the average nightly temperature and the average daylight temperature. The results are put into separate dataframes, one for each year.

Table 3.3: Data description of processed precipitation data.

Column	**Description**
wine name	The name of the wine bottle
year	Harvest year of the grapes of the wine
region	The grape growing area
winery	The name of the winery that bottled the wine
price	The cost of a bottle of the wine in euros
rating	The average rating of the wine (on a scale from 1 to 5)
no. ratings	The total number of ratings of the wine
country	The country the wine is from

Coupling ground stations and wine locations

As described in the previous section, locations were added to the wine data. These locations were needed to couple the wine data to their closest weather station. To get a list of unique weather station to couple to the wine locations, the `number_sta`, `lat` and `lon` columns from each year in the weather dataset and from each weather parameter dataset were selected, duplicate rows were dropped and duplicate ground stations were removed. Then the Haversine Distance formula was used to calculate the distance between each wine and the unique weather stations per year, this resulted in a list of weather stations and distances (in km), from this list the three closest weather stations were selected with their distances and added to the wine datasets. Finally, the wine datasets were coupled to each of the datasets calculated from each of the weather parameters (as described in the previous subsubsection). Since some ground stations had missing data for some of the weather parameters, each unique list of ground stations was different for each of the weather parameters.

Coupling weather parameters to wine locations

The final step in the data preparation was coupling the data from the specific weather parameters to the wines. To couple the data inner joins were used on each of the ground station IDs (the closest ground station, second closest and third closest) in the wine dataframes for each with the ground station IDs in the dataframes of the weather parameters and the abbreviated dataframes of the weather parameters. This was done for each of the parameters to avoid creating one dataframe with a high number of columns. To reduce the number of columns of these dataframes, only data of the closest weather station was kept, if this weather station was closer by than 20km, otherwise the average value of the weather parameter of all three weather stations was calculated and kept.

3.2 Correlation

The analysis of correlation between different variables is based on scatter-plots and correlation tables between the rating, price & ln(price) and the weather parameters. The following subsections analyze the relation between each of the weather parameters and the wine rating, price and ln(price). A visualization of the data through the form of scatter plots can be found in appendix A.

3.2.1 Standard wine parameters

Table 3.4 presents the correlation coefficients and p-values between the standard wine parameters (rating, price, ln(price) and year). As described in the related work section in the introduction, this data for both NW and SE France shows that the wine rating and wine price have a strong positive correlation (0.60) which is very statistically significant ($p \leq 0.001$). The year has a moderate negative correlation with all the other parameters: rating, price and ln(price). This can be explained by the fact that as wines get older, they "mature", they get a different flavor and generally become more expensive, so a younger wine from 2018 has a higher chance to have a lower price right now than a wine from 2017 or 2016.

Table 3.4: Linear correlation coefficients between the standard wine parameters, on the top the correlation coefficients (r) and accompanying p-values for wines made in the NW region of France, at the bottom the correlation coefficients and p-values for wines made in the SE region of France.

	NW							
	rating		**price**		**ln(price)**		**year**	
	r	p-value	r	p-value	r	p-value	r	p-value
rating	1	0	0.60	$1.2 \cdot 10^{-12}$	0.63	$3.1 \cdot 10^{-14}$	-0.29	$1.8 \cdot 10^{-3}$
price			1	0	0.91	$3.2 \cdot 10^{-44}$	-0.36	$6.9 \cdot 10^{-5}$
ln(price)					1	0	-0.42	$2.9 \cdot 10^{-6}$
year							1	0
	SE							
	rating		**price**		**ln(price)**		**year**	
	r	p-value	r	p-value	r	p-value	r	p-value
rating	1	0	0.61	$3.4 \cdot 10^{-47}$	0.67	$4.4 \cdot 10^{-60}$	-0.12	$9.8 \cdot 10^{-3}$
price			1	0	0.92	$7.1 \cdot 10^{-184}$	-0.31	$9.7 \cdot 10^{-12}$
ln(price)					1	0	-0.31	$2.1 \cdot 10^{-11}$
year					-		1	0

3.2.2 Precipitation

Table 3.5 shows correlation coefficients and p-values between the wine rating, price, ln(price) and precipitation variables. In the Table we can see a, statistically significant, moderate positive correlation between winter rain (precipW) and wine rating for NW France. The correlation between winter rain and price is weaker, but still statistically significant. For the NW of France the Table does not show a significant negative correlation between harvest rain and wine rating/price, which was observed in related research. The correlation coefficients for SE France and precipitation are quite different. Winter precipitation looks to have a weak negative correlation, that is statistically significant, with wine price and rating. Rain in the growing period also has a weak negative correlation with wine price and rating. But, also the data from SE France does not show a significant negative correlation of rain in the harvest with a lower wine quality. The correlation coefficients of the precipitation in the harvest are all not statistically significant, so the negative impact of rain in the harvest has not been refuted for these wines.

Table 3.5: Linear correlation coefficients between the standard wine parameters and the precipitation parameters, on the top the correlation coefficients (r) and accompanying p-values for wines made in the NW region of France, at the bottom the correlation coefficients and p-values for wines made in the SE region of France. precipW refers to the amount of precipitation from January to March, precipG refers to the amount of precipitation from April to September, precipH refers to the amount of precipitation in September and October.

	NW					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
precipW	0.35	$1.2 \cdot 10^{-4}$	0.21	$2.2 \cdot 10^{-1}$	0.25	$6.3 \cdot 10^{-3}$
precipG	-0.082	$3.8 \cdot 10^{-1}$	-0.050	$5.9 \cdot 10^{-1}$	-0.044	$6.4 \cdot 10^{-1}$
precipH	0.11	$2.6 \cdot 10^{-1}$	0.037	$7.8 \cdot 10^{-1}$	0.024	$8.0 \cdot 10^{-1}$
	SE					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
precipW	-0.20	$1.8 \cdot 10^{-5}$	-0.26	$2.1 \cdot 10^{-8}$	-0.28	$1.6 \cdot 10^{-9}$
precipG	-0.26	$1.2 \cdot 10^{-8}$	-0.24	$2.5 \cdot 10^{-7}$	-0.27	$8.7 \cdot 10^{-9}$
precipH	-0.076	$1.1 \cdot 10^{-1}$	-0.081	$8.6 \cdot 10^{-2}$	-0.078	$9.9 \cdot 10^{-2}$

3.2.3 Wind speed

Table 3.6 presents the correlations between the maximum wind speed in the growing and harvest season and wine rating and price. For the NW of France the maximum wind speed has a weak negative correlation with wine rating and price, on the other hand, for the SE the maximum wind speed has a very weak positive correlation with wine rating and price.

Table 3.6: Linear correlation coefficients between the standard wine parameters and the wind parameter, on the top the correlation coefficients (r) and accompanying p-values for wines made in the NW region of France, at the bottom the correlation coefficients and p-values for wines made in the SE region of France. ffmax refers to the highest measured wind speed in the growing season (from April to September).

	NW					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
ffmax	-0.19	$3.9 \cdot 10^{-2}$	-0.22	$1.8 \cdot 10^{-2}$	-0.24	$9.7 \cdot 10^{-3}$
	SE					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
ffmax	0.18	$1.4 \cdot 10^{-4}$	0.084	$7.4 \cdot 10^{-2}$	0.12	$9.3 \cdot 10^{-3}$

3.2.4 Humidity

Table 3.7 summarizes the correlation coefficients and p-values between the monthly humidity and wine price and ratings. Interestingly, for NW France, higher humidity looks to have a weak positive correlation with wine price and rating, while for SE France a higher humidity has a weak negative correlation with wine price and rating.

Table 3.7: Linear correlation coefficients between the standard wine parameters and the humidity parameters, on the top the correlation coefficients (r) and accompanying p-values for wines made in the NW region of France, at the bottom the correlation coefficients and p-values for wines made in the SE region of France. hu1-hu10 refers to month 4-10 in the year, which are the growing and harvest seasons.

	NW					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
hu4	0.26	$4.2 \cdot 10^{-3}$	0.26	$4.7 \cdot 10^{-3}$	0.22	$1.5 \cdot 10^{-2}$
hu5	0.26	$4.1 \cdot 10^{-3}$	0.28	$2.1 \cdot 10^{-3}$	0.23	$1.5 \cdot 10^{-2}$
hu6	0.27	$2.8 \cdot 10^{-3}$	0.24	$1.0 \cdot 10^{-2}$	0.19	$4.2 \cdot 10^{-2}$
hu7	0.28	$2.3 \cdot 10^{-3}$	0.30	$1.2 \cdot 10^{-3}$	0.24	$8.9 \cdot 10^{-3}$
hu8	0.27	$3.9 \cdot 10^{-3}$	0.22	$1.6 \cdot 10^{-2}$	0.16	$9.4 \cdot 10^{-2}$
hu9	0.31	$8.0 \cdot 10^{-4}$	0.38	$2.4 \cdot 10^{-5}$	0.33	$3.6 \cdot 10^{-4}$
hu10	0.19	$3.7 \cdot 10^{-2}$	0.32	$4.9 \cdot 10^{-4}$	0.28	$2.2 \cdot 10^{-3}$
	SE					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
hu4	-0.27	$7.2 \cdot 10^{-9}$	-0.20	$1.4 \cdot 10^{-5}$	-0.25	$7.7 \cdot 10^{-8}$
hu5	-0.26	$3.6 \cdot 10^{-8}$	-0.25	$1.2 \cdot 10^{-7}$	-0.29	$5.2 \cdot 10^{-10}$
hu6	-0.24	$3.4 \cdot 10^{-7}$	-0.20	$1.2 \cdot 10^{-5}$	-0.24	$2.6 \cdot 10^{-7}$
hu7	-0.24	$2.4 \cdot 10^{-7}$	-0.18	$1.4 \cdot 10^{-4}$	-0.22	$2.5 \cdot 10^{-6}$
hu8	-0.22	$1.5 \cdot 10^{-6}$	-0.19	$3.9 \cdot 10^{-5}$	-0.23	$8.2 \cdot 10^{-7}$
hu9	-0.22	$2.2 \cdot 10^{-6}$	-0.16	$6.0 \cdot 10^{-4}$	-0.19	$4.7 \cdot 10^{-5}$
hu10	-0.22	$2.1 \cdot 10^{-6}$	-0.14	$2.4 \cdot 10^{-3}$	-0.19	$6.5 \cdot 10^{-5}$

3.2.5 Temperature

Table 3.8 shows the correlation coefficients and p-values between the average temperature in the growing season, harvest season and the first (April, May) and second half (June, July) of the growing season and wine price and ratings, including the daily and nightly temperatures. With the log of the price the temperature in NW France in the harvest season has a weak positive correlation and the nightly temperature in the harvest season has a moderate, statistically significant, positive correlation with the natural logrithm of price. On the other hand the temperature in April and May (for the full day, but also separately daily and nightly), has a weak negative correlation with wine rating and price. The data for SE France looks quite different. The average temperature, also nightly and daily, in the growing period, harvest period and first and second half of the growing period, has a weak, statistically significant, positive relation with wine rating and a very weak correlation with price and ln(price).

Table 3.8: Linear correlation coefficients and p-values of the standard wine parameters and the temperature parameters for wines made in the NW and SE region of France. tG refers to the average temperature in the growing season (April-September), tH refers to the average temperature in the harvest season (August-October), tQ1 refers to the average temperature in April and May, tQ2 refers to the average temperature in June and July. The additions "D" and "N" to the parameter names refer to the daytime (between dawn and dusk) and nighttime (between dusk and dawn), respectively.

	NW					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
tG	-0.12	$2.0 \cdot 10^{-1}$	0.0084	$9.3 \cdot 10^{-1}$	0.066	$4.8 \cdot 10^{-1}$
tH	-0.035	$7.1 \cdot 10^{-1}$	0.11	$2.2 \cdot 10^{-1}$	0.18	$5.1 \cdot 10^{-2}$
tQ1	-0.21	$2.7 \cdot 10^{-2}$	-0.082	$3.8 \cdot 10^{-1}$	-0.059	$5.3 \cdot 10^{-1}$
tQ2	-0.16	$8.1 \cdot 10^{-2}$	-0.036	$7.0 \cdot 10^{-1}$	0.014	$8.9 \cdot 10^{-1}$
tGD	-0.12	$2.0 \cdot 10^{-1}$	0.0093	$9.2 \cdot 10^{-1}$	0.063	$5.0 \cdot 10^{-1}$
tHD	-0.039	$6.8 \cdot 10^{-1}$	0.089	$3.4 \cdot 10^{-1}$	0.15	$1.1 \cdot 10^{-1}$
tQ1D	-0.18	$5.3 \cdot 10^{-2}$	-0.041	$6.6 \cdot 10^{-1}$	-0.013	$8.9 \cdot 10^{-1}$
tQ2D	-0.17	$7.1 \cdot 10^{-1}$	-0.046	$6.3 \cdot 10^{-1}$	0.0016	$9.9 \cdot 10^{-1}$
tGN	-0.098	$3.0 \cdot 10^{-1}$	0.0037	$9.7 \cdot 10^{-1}$	0.065	$4.9 \cdot 10^{-1}$
tHN	0.19	$3.9 \cdot 10^{-2}$	0.22	$1.7 \cdot 10^{-2}$	0.32	$5.3 \cdot 10^{-4}$
tQ1N	-0.24	$8.8 \cdot 10^{-3}$	-0.19	$3.7 \cdot 10^{-2}$	-0.19	$4.0 \cdot 10^{-2}$
tQ2N	-0.13	$1.6 \cdot 10^{-1}$	-0.0012	$9.9 \cdot 10^{-1}$	0.052	$5.8 \cdot 10^{-1}$

	SE					
	rating		**price**		**ln(price)**	
	r	p-value	r	p-value	r	p-value
tG	0.23	$1.2 \cdot 10^{-6}$	0.12	$8.6 \cdot 10^{-3}$	0.15	$1.6 \cdot 10^{-3}$
tH	0.19	$4.2 \cdot 10^{-5}$	0.11	$1.7 \cdot 10^{-2}$	0.14	$2.3 \cdot 10^{-3}$
tQ1	0.21	$1.1 \cdot 10^{-5}$	0.059	$2.1 \cdot 10^{-1}$	0.082	$8.2 \cdot 10^{-2}$
tQ2	0.26	$3.8 \cdot 10^{-8}$	0.14	$2.4 \cdot 10^{-3}$	0.17	$3.4 \cdot 10^{-4}$
tGD	0.24	$3.8 \cdot 10^{-7}$	0.14	$2.7 \cdot 10^{-3}$	0.15	$1.3 \cdot 10^{-3}$
tHD	0.22	$2.8 \cdot 10^{-6}$	0.15	$2.0 \cdot 10^{-3}$	0.16	$5.8 \cdot 10^{-4}$
tQ1D	0.22	$2.0 \cdot 10^{-6}$	0.076	$1.1 \cdot 10^{-1}$	0.089	$6.0 \cdot 10^{-2}$
tQ2D	0.25	$5.2 \cdot 10^{-8}$	0.15	$1.8 \cdot 10^{-3}$	0.16	$5.4 \cdot 10^{-4}$
tGN	0.16	$4.7 \cdot 10^{-4}$	0.071	$1.3 \cdot 10^{-1}$	0.12	$1.4 \cdot 10^{-2}$
tHN	0.13	$6.8 \cdot 10^{-3}$	0.10	$3.1 \cdot 10^{-2}$	0.14	$4.0 \cdot 10^{-3}$
tQ1N	0.12	$1.4 \cdot 10^{-2}$	0.0014	$9.8 \cdot 10^{-1}$	0.042	$3.7 \cdot 10^{-1}$
tQ2N	0.22	$2.7 \cdot 10^{-6}$	0.11	$2.2 \cdot 10^{-2}$	0.15	$1.1 \cdot 10^{-3}$

3.3 Prediction

The prediction research question is investigated based on three different datasets, one dataset with the wines from the NW region of France, another dataset with the wines from the SE region of France and a final dataset with both wines from SE and NW France, these datasets were combined with the weather parameters that were mentioned and described in the previous sections. From these datasets we predict the wine price, the (natural) logarithm of the price and the wine rating, using separate models for each of these parameters. The datasets are split into a training and test dataset in a 80 to 20 percent ratio.

The models were based on support vector regression and used the RBF kernel. The parameters for each of the models, the data the model was trained on (NW or SE) and the value predicted are shown in Table 3.9. The parameters were chosen by using the grid search algorithm, an algorithm that finds parameters that optimize the accuracy of a model.

In Table 3.10 we can see the evaluation scores, expressed in r-squared, mean squared error (MSE) and root mean squared error (RMSE), of the different models. The fourth model fits best with an R^2 of 0.52, so the line equation calculated using the RBF kernel is reasonably representative of the data, and the RMSE is low (0.37) compared to the values of ln(price) (which all lie between 1.5 and 4.5).

The models that predict the price (model 1, 2 and 3) have a mediocre r-squared score, their RMSE is quite high. Model 5 and 6 that predict ln(price) and model 7, 8 and 9 that predict the wine rating do not perform well. The RMSE of model 5 and 6 is not that high, but the r squared score is negative, which can happen if a dataset is split into training and test dataset and the RMSE is calculated over the test set. The RMSE of model 7, 8 and 9 are also not that high, but especially the r squared score of model 8 and 9 are very low, so these models do not perform well at all. This could indicate several things. On one hand this could mean that the wine ratings on the website Vivino are not representative of the wine quality, that the ideal weather for each of the different red wine types is so different that a single model is not capable of predicting the rating or price for these wines or that the weather data does not describe the actual weather conditions of the grape growing process. Additionally, previous research focused on specific vineyards & wineries over a longer period in time and the weather conditions these vineyards & wineries recorded in those years, so it could be that weather data over 3 years provides insufficient training data for predicting wine quality and prices.

Table 3.9: Support Vector Regression model parameters.

	Region	**Predict**	**C**	ϵ	γ
Model 1	NW	Price	10	0.1	0.1
Model 2	SE	Price	10	5	0.01
Model 3	NW & SE	Price	10	0.5	0.01
Model 4	NW	ln(Price)	10	0.5	0.001
Model 5	SE	ln(Price)	1	1.5	0.001
Model 6	NW & SE	ln(Price)	1	1.5	0.001
Model 7	NW	Rating	1	0.5	0.001
Model 8	SE	Rating	1	1	0.001
Model 9	NW & SE	Rating	1	1	0.001

Table 3.10: Support Vector Regression results, on the left the R-squared score, in the middle the mean squared error (MSE) and on the right the root mean squared error (RMSE).

	R^2	MSE	RMSE
Model 1	0.31	162	12.7
Model 2	0.44	173	13.2
Model 3	0.17	290	17.0
Model 4	0.52	0.14	0.37
Model 5	-0.027	0.42	0.65
Model 6	-0.019	0.45	0.67
Model 7	-0.021	0.034	0.18
Model 8	-0.42	0.084	0.29
Model 9	-0.98	0.12	0.35

Chapter 4

Conclusions

4.1 Correlation

In the introduction we formulated the following hypotheses. We expected to find a positive correlation between winter rain and wine price and rating, but this positive correlation was only observed in a moderate form for wines made in the NW region of France. For the wines made in the SE region this led to a weak negative correlation. Another hypothesis stated that harvest rain would have a negative correlation with wine price and rating, but this was not observed in this analysis. Higher temperature in the growing and harvest season only seemed beneficial for wines made in the SE region of France; in the NW this correlation was weakly negative. Interestingly, for the NW region of France humidity had a weak positive correlation with wine price and rating while this was a weak negative correlation in the South East region of France. Lastly, the maximum wind speed only had a weak negative correlation with wine rating and price in the North West region of France, this correlation was weak positive in the South East region of France.

4.2 Prediction

It was only possible to create a reasonably well performing model for the natural logarithm of the price of the wines from North West France. The predictive models for wine prices all had a mediocre performance, but with more data and parameter fine-tuning might be able to more accurately predict wine prices. The predictive models for wine ratings all had very low r-squared values, so they did not fit the data. So, the wine ratings of the website Vivino are difficult to predict based on weather data.

4.3 Future research

It still seems possible to predict wine prices and quality for different wine types based on weather data using one model. In future research, if there is more accurate weather data available over more years, a well performing model based on support vector regression could be created. Furthermore, it could be interesting to extended the weather and wine data could with crop yields in the years the wines are produced to create a better performing model.

Bibliography

[1] https://meteofrance.github.io/meteonet/english/data/summary/.

[2] https://meteonet.umr-cnrm.fr/dataset/data/.

[3] Orley Ashenfelter. Predicting the quality and prices of bordeaux wine. *The Economic Journal*, 118(529):F174–F184, 2008.

[4] Orley Ashenfelter, David Ashmore, and Robert Lalonde. Bordeaux wine vintage quality and the weather. *Chance*, 8(4):7–14, 1995.

[5] Orley Ashenfelter and Gregory V Jones. The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics*, 8(3):285–293, 2013.

[6] Orley Ashenfelter and Karl Storchmann. Using hedonic models of solar radiation and weather to assess the economic effect of climate change: The case of mosel valley vineyards. *The Review of Economics and Statistics*, 92(2):333–349, 2010.

[7] Raymond P Byron and Orley Ashenfelter. Predicting the quality of an unborn grange. *Economic Record*, 71(1):40–53, 1995.

[8] John M Chambers. *Graphical methods for data analysis*. CRC Press, 2018.

[9] Jean-Michel Chevet, Sébastien Lecocq, and Michael Visser. Climate, grapevine phenology, wine production, and prices: Pauillac (1800–2009). *American Economic Review*, 101(3):142–146, 2011.

[10] Alessandro Corsi and Orley Ashenfelter. Predicting italian wine quality from weather data and expert ratings. *Journal of Wine Economics*, 14(3):234–251, 2019.

[11] Gustavo Ferro and Ignacio Benito Amaro. What factors explain the price of top quality wines? *International Journal of Wine Business Research*, 30(1):117–134, 2018.

[12] Jim Frost. How to interpret r-squared in regression analysis, Jul 2023.

[13] Kenneth Gade. A non-singular horizontal position representation. *The journal of navigation*, 63(3):395–417, 2010.

[14] Victor Ginsburgh, Muriel Monzak, and Andras Monzak. Red wines of medoc: What is wine tasting worth? *Journal of Wine Economics*, 8(2):159–188, 2013.

[15] John W Haeger and Karl Storchmann. Prices of american pinot noir wines: climate, craftsmanship, critics. *Agricultural economics*, 35(1):67–78, 2006.

[16] Gregory V Jones and Robert E Davis. Climate influences on grapevine phenology, grape composition, and wine production and quality for bordeaux, france. *American journal of enology and viticulture*, 51(3):249–261, 2000.

[17] Gregory V Jones, Michael A White, Owen R Cooper, and Karl Storchmann. Climate change and global wine quality. *Climatic change*, 73(3):319–343, 2005.

[18] Tania Kleynhans, Matthew Montanaro, Aaron Gerace, and Christopher Kanan. Predicting top-of-atmosphere thermal radiance using merra-2 atmospheric data with deep learning. *Remote Sensing*, 9(11):1133, 2017.

[19] Georgios C Koufos, Theodoros Mavromatis, Stefanos Koundouras, Nikolaos M Fyllas, Serafeim Theocharis, and Gregory V Jones. Greek wine quality assessment and relationships with climate: Trends, future projections and uncertainties. *Water*, 14(4):573, 2022.

[20] Philippe Masset, Alexandre Mondoux, and Jean-Philippe Weisskopf. Fine wine pricing in a small and highly competitive market. *International Journal of Wine Business Research*, 35(1):164–186, 2023.

[21] Britta Niklas. Impact of annual weather fluctuations on wine production in germany. *Journal of Wine Economics*, 12(4):436–445, 2017.

[22] Edward Oczkowski. The effect of weather on wine quality and prices: An australian spatial analysis. *Journal of Wine Economics*, 11(1):48–65, 2016.

[23] Carlos D Ramirez. Wine quality, wine prices, and the weather: Is napa "different"? *Journal of Wine Economics*, 3(2):114–131, 2008.

[24] Aymeric Roucher, Leonidas Aristodemou, and Frank Tietze. Predicting wine prices based on the weather: Bordeaux vineyards in a changing climate. *Frontiers in Environmental Science*, 10:1020867, 2022.

[25] Günter Schamel and Kym Anderson. Wine quality and varietal, regional and winery reputations: hedonic prices for australia and new zealand. *Economic Record*, 79(246):357–369, 2003.

[26] Shaun Turney. Pearson correlation coefficient (r): Guide amp; examples, Jun 2023.

[27] Michelle Yeo, Tristan Fletcher, and John Shawe-Taylor. Machine learning in fine wine price prediction. *Journal of Wine Economics*, 10(2):151–172, 2015.

[28] Zach. How to interpret root mean square error (rmse), May 2021.

Appendix A

Data visualization: Scatter Plots

A.1 Precipitation Scatter Plots

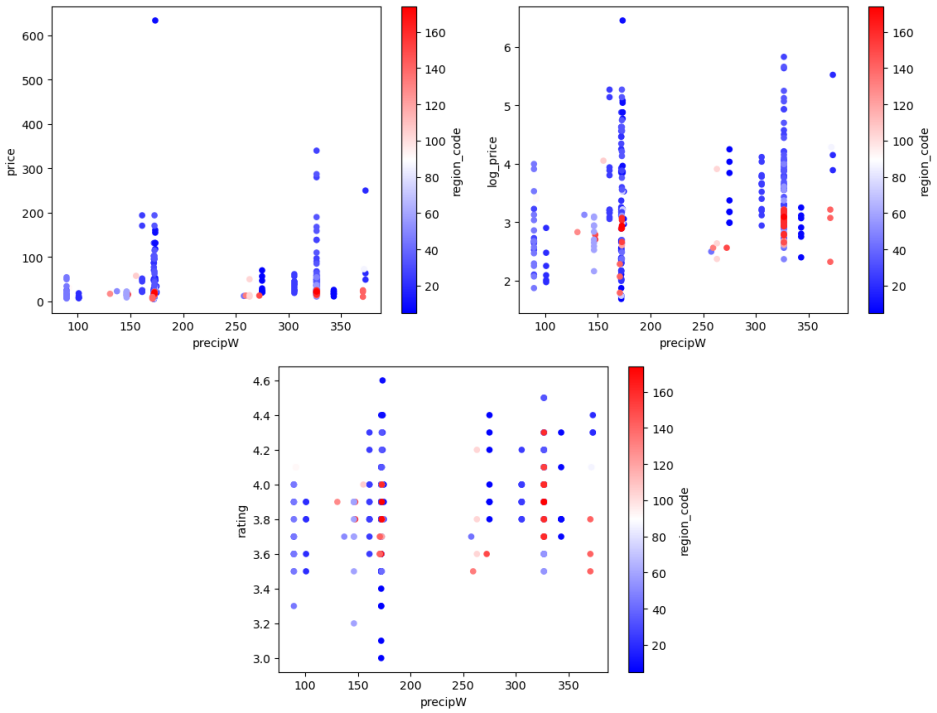

Figure A.1: Scatterplots of total winter precipitation. On the top left winter rain vs. price, on the top right winter rain vs. ln(price) and on the bottom winter rain vs. wine rating

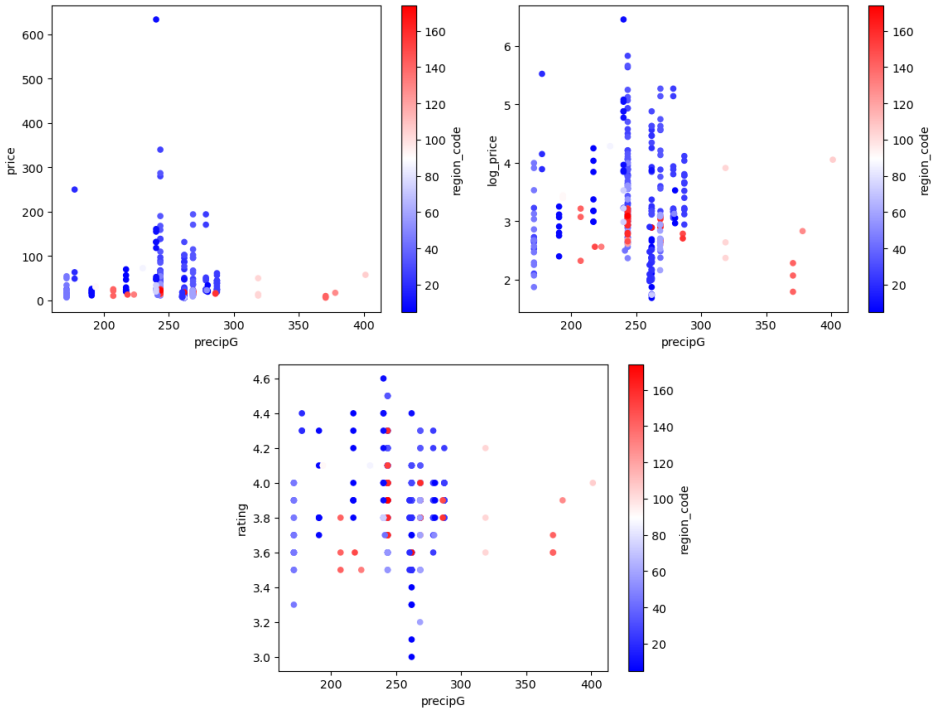

Figure A.2: Scatterplots of total growing season precipitation. On the top left growing season rain vs. price, on the top right growing season rain vs. ln(price) and on the bottom growing season rain vs. wine rating

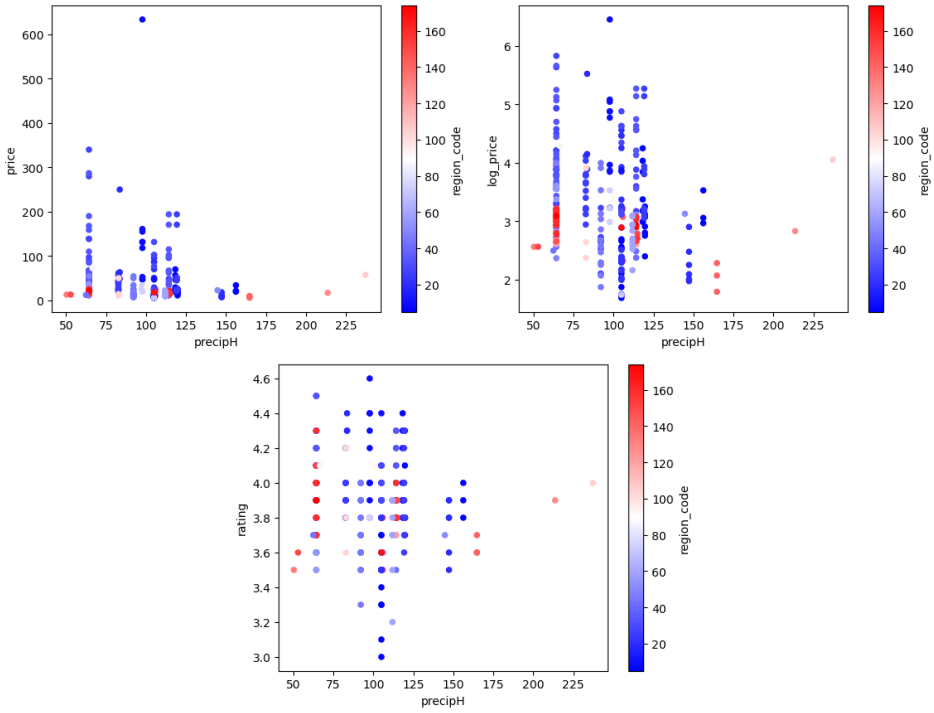

Figure A.3: Scatterplots of total harvest season precipitation. On the top left harvest rain vs. price, on the top right harvest rain vs. ln(price) and on the bottom harvest rain vs. wine rating

A.2 Wind speed Scatter Plots

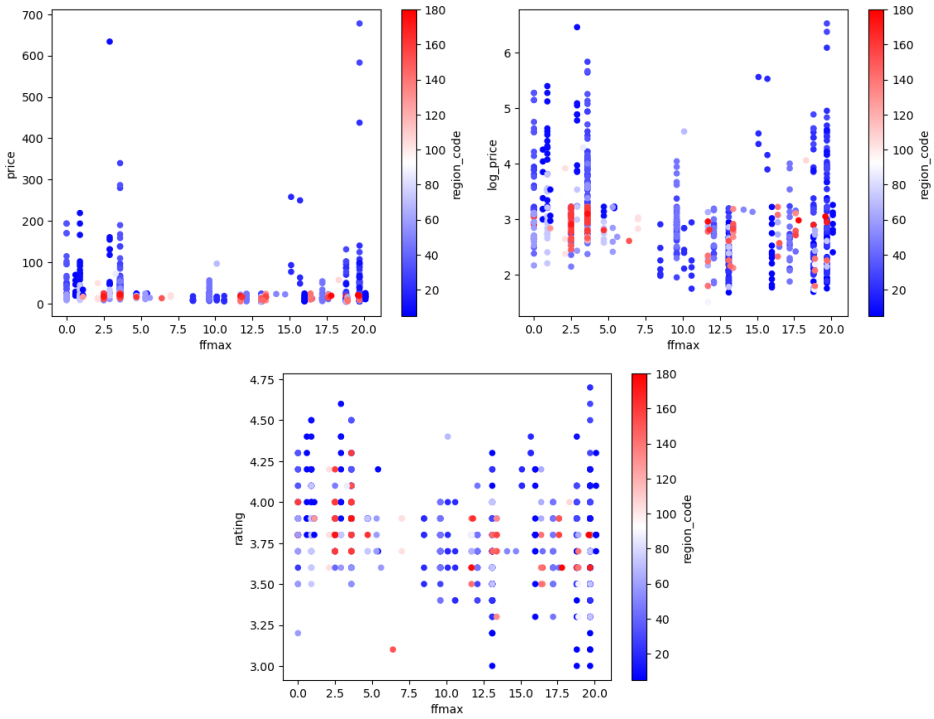

Figure A.4: Scatterplots of maximum wind speed. On the top left maximum wind speed vs. price, on the top right maximum wind speed vs. ln(price) and on the maximum wind speed vs. wine rating

A.3 Temperature Scatter Plots

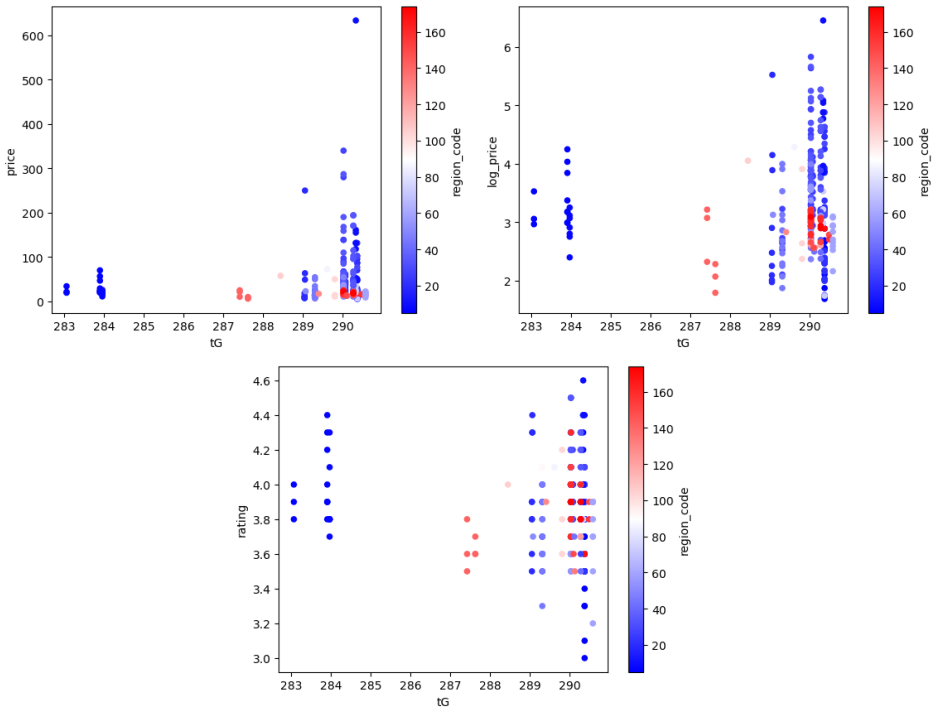

Figure A.5: Scatterplots of average growing season temperature. On the top left growing season temperature vs. price, on the top right growing season temperature vs. ln(price) and on the bottom growing season temperature vs. wine rating

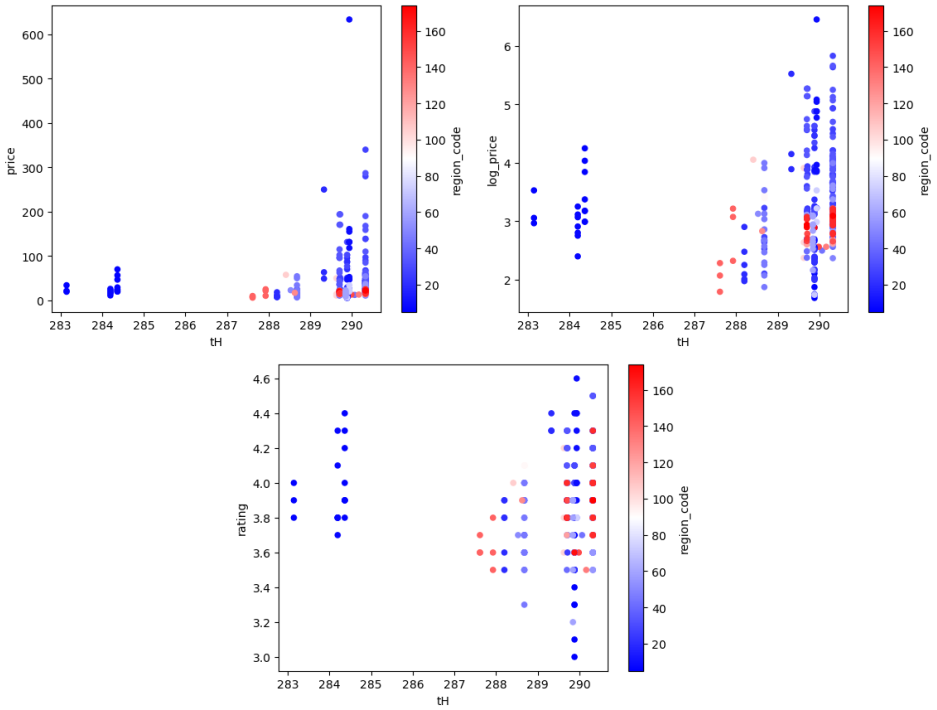

Figure A.6: Scatterplots of average harvest season temperature. On the top left harvest season temperature vs. price, on the top right harvest season temperature vs. ln(price) and on the bottom harvest season temperature vs. wine rating

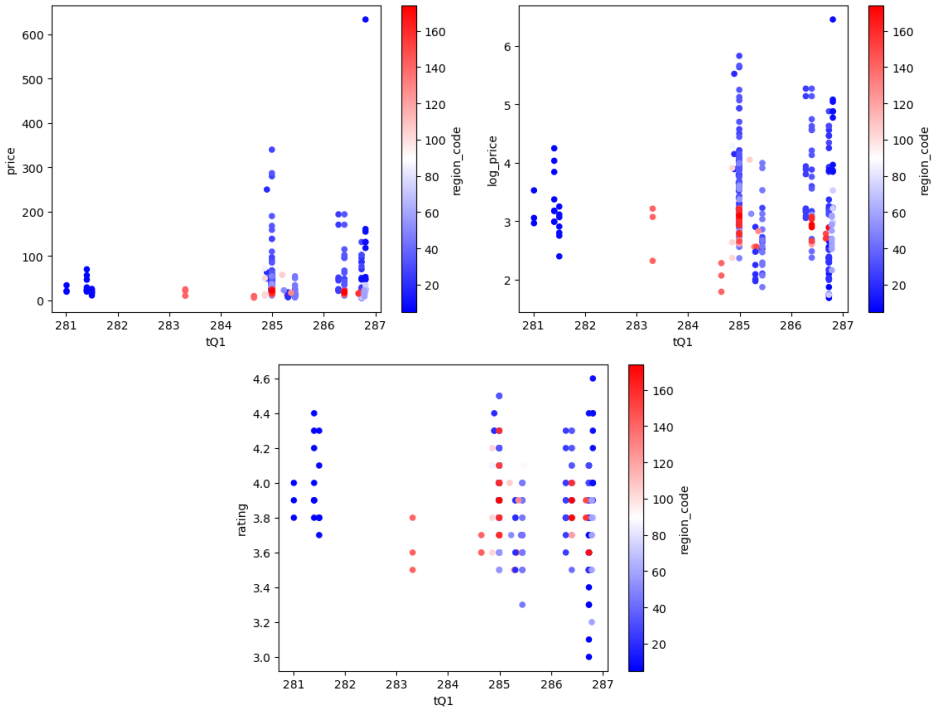

Figure A.7: Scatterplots of average temperature of quarter 1 (April and May). On the top left temperature of quarter 1 vs. price, on the top right temperature of quarter 1 vs. ln(price) and on the bottom temperature of quarter 1 vs. wine rating

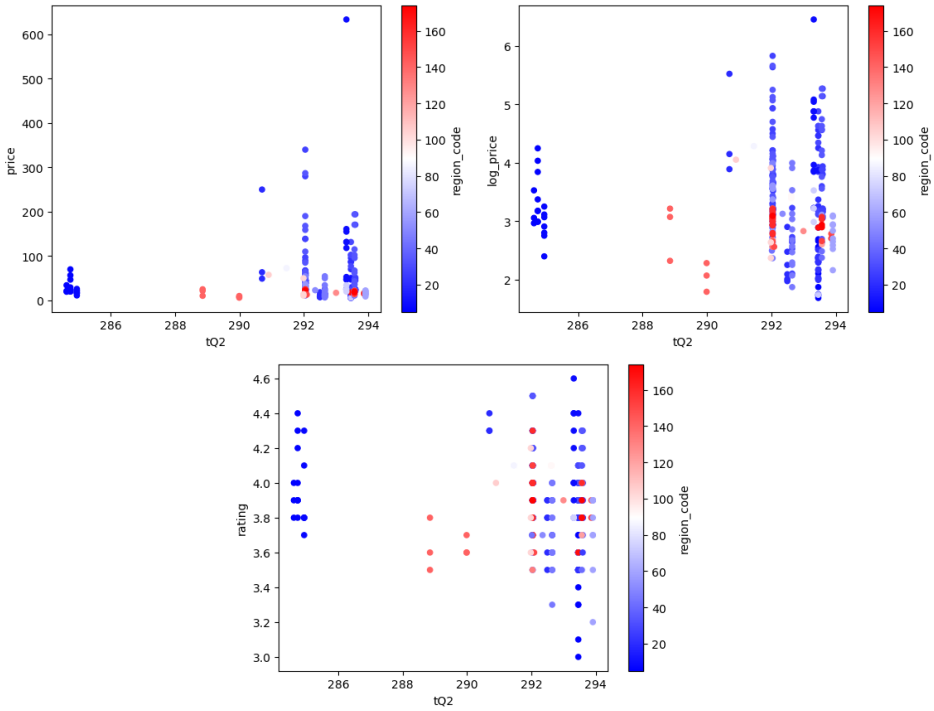

Figure A.8: Scatterplots of average temperature of quarter 2 (June and July). On the top left temperature of quarter 2 vs. price, on the top right temperature of quarter 2 vs. ln(price) and on the bottom temperature of quarter 2 vs. wine rating