1. Motivation

1.1 *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

The dataset was created to study the difference in the average ratings of red wines from multiple regions. The goal was to create a feasible dataset that researchers can use to study whether certain regions, that are popular for the production of red wines, produce higher rated wines than other regions. We aim to fill this gap by scraping an online listing of wines. The outcomes of this study can be used by consumers, in deciding from which region to buy a bottle of wine from. . We chose to scrape the content of Vivino.com as this website offers the opportunity to obtain this data as the website is the largest listing of wines and includes the opportunity to leave a review and is openly accessible for the public.

1.2 *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

This dataset is created by Teun Geurts, Ilse van den Bosch, Julie Habets, Marit Verbruggen and Eveline Huckriede as part of the course Skills: Online Data Collection and Management. This course is taught by Hannes Datta at Tilburg School of Economics and Management as part of the Marketing Analytics Master's program.

1.3 *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

Besides time and effort, there was no funding needed for the creation of this dataset.

2. Composition

2.1 *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of in- stances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

Every instance in our dataset represents a listing of a bottle of red wine on the Vivino.com website that is scraped at that moment. For our final dataset, this means all the wines listed on Vivino.com on October 15th. Furthermore there are no interactions between the instances, but instances could belong to the same manufacturer.

2.2 *How many instances are there in total (of each type, if appropriate)?*

Vivino.com displays a lot of different bottles of wine. To be precise, 49,422 different wines are available.

2.3 *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

Since the total number of wines is too big for the scope of our research, we narrowed down our sample by only taking the red wines displayed on Vivino.com into account. We decided to focus on the red wines as it is the main category on Vivino.com (the majority of the listed wines are red wines). In total, 30,126 red wines are available on Vivino.com. Then, conforming to the goal of our research, we filter on the six regions that are available in the filter bar displayed on Vivino.com. These regions are Bordeaux, Bourgogne, Napa Valley, Piemonte, Rhone Valley and Toscana. By using this filter, the total of red wines to be scraped has already dropped to 13,682, this is because small regions are not included in the filter function of www.vivino.com. Therefore we only scrape the 6 most common wine regions. Even though it is possible to scrape all these 13,682 red wines, it would still take too much time. As the scraper overwrites the output every time it scrapes, the marginal time of scraping a wine increases a lot with every wine added (for more information see question 3.2). Because of the fact that the scraper scrapes all wines when an extra 25 wines are visible, this would mean 519 scraping rounds are needed if we want to scrape all 13,682 wines. Given that each wine is scraped within 0,2 seconds, the time it would take to scrape all these wines is approximately 8 days (see calculations in appendix 1). Because we run our scraper on a laptop, it won't be able to run for days. Therefore, we decided to create an even narrower sample, which allows us to scrape as many red wines as possible within our desired time frame of a workday. Ultimately, we narrowed our sample down to the 500 cheapest wines from each region. The disadvantage of this sample is that we

[1*] https://arxiv.org/abs/1803.09010

won't be able to draw conclusions on all wines listed on Vivino.com. However, the scraper is built in a way that it is relatively easy to convert the scraper to be able to scrape other types of wines too.

2.4 *What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.*

Every instance (wine) in our dataset includes a name, region, price, number of ratings and the average rating. In total there are 6 regions included in our dataset. The rating variable is the average of all ratings people left on the specific wine on Vivino.com. This is a number between 1 and 5.

2.5 *Is there a label or target associated with each instance? If so, please provide a description.*

Every instance is labeled by the variable region, so we can filter every instance by its region.

2.6 *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

Since we scrape the catalogue page, some prices are not available. Instead of a price, the site gives "Sold out". We transform these variables into our dataset as ("Sold out, no price available").

2.7 *Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

In our dataset the relationship between wines from different regions has been made explicit through the region assigned to each instance. We did not include relationships between different wines from the same manufacturers as it was not within the scope of our research. This could be an entry for future research.

2.8 *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

There are no recommended data splits in our dataset. We tested the scraper step-by-step and performed the web scraper after we knew exactly that our web scraping tool runs as desired. However, no data splits were necessary for this process.

2.9 *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they ex- isted at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The web scraper we built does, by itself, rely on the web structure of Vivino.com. A small likelihood exists that Vivino.com will remove open-access to their pages, and they might update their HTML-layout which might lead to errors in the web scraper. The dataset itself is not self-contained as data on the website could change (average rating could in-/decrease and/or new listings will be added). There are no archival versions of the complete dataset, as Vivino.com is continually being updated.

2.10 *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

Our dataset does not contain data that might be considered confidential, as all the data scraped is publicly available on the Vivino.com website. There is

also no authentication or login needed to access the data we scrape.

2.11 *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

As we only scrape the total review rating, no individual reviewer could be identified by our data. If we would have conducted a sentiment analysis based on reviews, for example, this would have been the case. Furthermore, besides the full anonymity of identities within our data, we do not see problems regarding offensivity of our data or the conclusions of our research.

2.12 *Does the dataset relate to people?* ***If not, you may skip the remaining questions in this section.***

Our dataset does not relate to people.

2.13 *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

Not applicable on our dataset.

2.14 *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

Not applicable on our dataset.

2.15 *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

Not applicable on our dataset.

3. Collection Process

3.1 *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

We scraped the data directly from the Vivino.com listing page. All the data is directly observable, but it comes in big numbers which makes our scraper still very valuable. The fact that the data is directly observable minimized issues with validation/verification .

3.2 *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software pro- gram, software API)? How were these mechanisms or procedures validated?*

As Vivino.com does not have an API, we built a web scraping tool ourselves. We used the programming language Python to extract all the wines from the website. Within Python we used the selenium module to be able to navigate through the website. A substitute for selenium could have been BeautifulSoup, but since the BeautifulSoup scraping tool is too robust for Vivino.com, the website denies access when BeautifulSoup is used. On the contrary, Selenium lets us scrape the page as if we are users: this withholds us from authentication problems and blocks from the website.
Our technical extraction plan actually consists of three parts: navigating, collecting data and transforming data. The starting point of our scraper is the homepage of Vivino: www.vivino.com. Thus, the first important step is to navigate the scraper towards the correct page to extract data from. We manage this by defining a prepare_window function. Ultimately, the output of this function is a maximized ChromeDriver window that contains the Wines page of Vivino.com, filtered on: Red Wines, Full Price Range, All Average Ratings and sorted by price from low to high. To get to this page we search for the correct buttons via xpath navigation. By using xpath navigation instead of text navigation to navigate through the website, we make our scraper useful for a bigger public. Now the scraper is not only limited to the English version of Vivino, but

also for other languages! Then, we use the built-in click() function of selenium to click on the right buttons which ultimately leads to the final page as described.
A second important step in our scraper is the scrolling process. Since the first window of the catalogue page only shows 25 wines, it would be impossible to scrape all the other wines without scrolling. Thus, a scroller is necessary. We built the scroller by using a for-loop which scrolls in range +1000 a given amount of time. Within this loop, we also collect our data.
The data collection process happens within the scroller loop. That would mean that every time we scroll, the site will also be scraped. This would make the scraper very inefficient and time consuming, thus we created an if-statement which checks whether the amount of wines visible in the current(scrolled) window is higher than the amount of wines visible in the previous window. If this is the case, the scraper starts collecting data from all the wines visible in the current window. Ultimately, four variables are defined: the name of the wine, the price, the number of ratings and the average rating. We define classes and attributes for each variable and then create a for-loop that iterates in the range of how many wines are visible in the current window. This means that the scraper extracts these variables from every listed wine visible in the current window. Then, the scraper writes all these variables per wine directly into a csv file. This csv file is named according to the region the scraper is collecting data from.
However, this unfortunately still leads to some duplicate data. That is where the data transformation process comes into place. To be able to filter out all duplicate data, we create a variable "scraping_round" into our scrolling loop. This variable goes up by +1 when a new scraping round starts. Then, after the data collection is done, we can filter the output csv files by filtering on the maximum (a.k.a. last) scraping round, containing the data in the last (and biggest) window. This leads to 6 new csv files per region that contain filtered data without any duplicates. Since the non-filtered files are still in the output folder, we let the computer delete these files automatically, because these will not be of any use any more. Then, the last step in our data transformation process is to merge all output files together into one big output file available for analysis. This is done by a merge function which merges all the filtered files together. Ultimately, we are left with one big output file that contains all the collected data from the different regions. We also keep the 6 loose filtered files per Region in the output folder, since this could be of any use in the future: for example researchers interested in data per region. The final dataset is stored in the same folder as the scraper file.

3.3 *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

Due to time constraints, we could only scrape 500 red wines per region. This will allow us to scrape within a working day. As 500 wines per region translates to approximately ¼ of the total wines on the webpage, we believe that it was sufficiently large to show any large differences within the regions. Besides, these 500 red wines are the cheapest red wines per region. This gives an extra dimension to our research. (E.g. which region offers the best cheapest wines).

3.4 *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

All team members were involved in the data collection process and none of them was financially compensated for their efforts, as the project was part of the course Online Data Collection and Management which is compulsory in their education. This did allow for many learning opportunities.

3.5 *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time- frame in which the data associated with the instances was created.*

We did not perform a time series analysis as the content on Vivino.com is quite static, so we decided to build a one-time scraper for the website. An application for future research could be to see whether the results differ over time.

3.6 *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

No ethical review process was conducted.

3.7 *Does the dataset relate to people?* ***If not, you may skip the remaining questions in this section.***

Our dataset does not relate to people.

3.8 *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

Not applicable.

3.9 *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

Not applicable.

3.10 *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

Not applicable.

3.11 *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

Not applicable.

3.12 *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

Not applicable.

4. Preprocessing, cleaning, labeling

4.1 *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remain- der of the questions in this section.*

The output which the scraper delivers by itself is 6 output files containing the name, price, number of reviews and average ratings of the wines of the different Regions. These output files contain a lot of duplicate rows which we could not prevent from being written into the csv file. Since duplicate rows are not of value for our research this was a reason for us to filter the data in these first output files. The reason behind the duplicate rows within our output files is the fact that a new scraping round overwrites the previous scraping round. Consequently, we managed to filter the files by implementing a variable named "scraping_round" into our scraper. The "scraping_round" variable is an integer value that starts at 1 and goes up by +1 for every time it scrapes the website again. The variable "scraping_round" is also written into our csv file per wine that is scraped. We finally filter the csv file on maximum (a.k.a. last) scraping round and thus delete all rows obtained in previous scraping rounds. The last scraping round represents the scraping round where the 500 wines are scraped, and thus contains all the wines we want in the filtered output files per region.

After we obtain the filtered output files per region, we transform these 6 output files into one big output file containing all the data together. Conform the goal of our research, a merged dataset is more convenient to work with. We can now compare variables within one dataset instead of variables within six separate files. The retrieval of one big dataset is done by merging the six files together. This is an easy job since all the files contain the same header and layout. We also keep the filtered files per region in the output folder, since they could be of use in future research projects.

Within our dataset, we also transform some variables before we write them to a csv file. Since we scrape the

catalogue page, some prices are not available. Instead of a price, the site gives "Sold out". We transform these variables into our dataset as ("Sold out, no price available). Also, some prices are given as "available online from €(price)". To make the output data from our scraper more useful for further analysis, we delete the euro sign and then transform this variable to one float value that contains the price (conforms to the other wine prices in our output data). Besides some prices that were not available on the catalogue page, we did not face any issues regarding missing values. Lastly, the number of ratings on Vivino.com is given as "x ratings". In our scraper we transform this into one integer value that represents the number of ratings, without the string "ratings" attached. When writing to the csv files, a header is created to label columns into meaningful and clear variable names. Finally, the average ratings appeared as a character variable in the scraped data file. To allow for the analysis, we transformed this to a numeric variable.

4.2 *Was the "raw" data saved in addition to the prepro- cessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

The main part of the transformation of the variables happens before the variables are written into a csv file. Thus, this is not saved in a separate file and is not accessible for other people. However, the transformation of the average ratings happens afterwards in the software program R. Other temporary output files are also not saved into the output folder: we delete these files already in our Python code since we do not see any valuable use of the data containing a lot of duplicate rows. Deleting the unnecessary data keeps the output folder more neat and clear.

4.3 *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

Not applicable.

5. Uses

5.1 *Has the dataset been used for any tasks already? If so, please provide a description.*

Next to our own analysis, our dataset has not been used for any other tasks yet. In our own analysis we use our dataset to be able to calculate the mean average rating of red wines per region.

5.2 *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

There is no repository that links to any or all papers or systems that use the dataset as of now.

5.3 *What (other) tasks could the dataset be used for?*

Besides research purposes mentioned in the other questions of this documentation, our dataset can also be used by enterprises or regular consumers. Enterprises such as entrepreneurs, restaurants and regular consumers can make use of this scraper to explore which regions overall have the best red wines. Entrepreneurs can see which wine regions are most likely to score the highest among consumers and gain inspiration on where to set up a wine business. Restaurants can easily offer the best rated wines in their restaurant, without going through the hassle of computing the mean of each region themselves. Regular consumers can do this as well, but they can also use the dataset of this scraper for completely different purposes as well. Regular customers can base their choice for a 'wine holiday' on the outcome of this data.

5.4 *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam congue rhoncus erat. Vestibulum sed nulla sed est malesuada vehicula. Etiam malesuada luctus aliquet. Aenean quis erat tincidunt, consequat orci vel, luctus risus. Nunc vel diam sit amet lorem interdum mattis.

5.5 *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset can be seen as good support in making business decisions. However, this dataset should not be used solely when making business decisions about wines, but it should be supported by other documentation specifically targeted to the business decision.

```python
total_time_seconds = 0
for numbers in range(25, 13000, 25):
    total_time_seconds += numbers*0.2
    total_time_hours = total_time_seconds/3600
    total_time_days = total_time_hours/24
print("Total time in hours:", total_time_hours)
print("Total time in days:", total_time_days)
```

```
Total time in hours: 187.41666666666666
Total time in days: 7.809027777777778
```