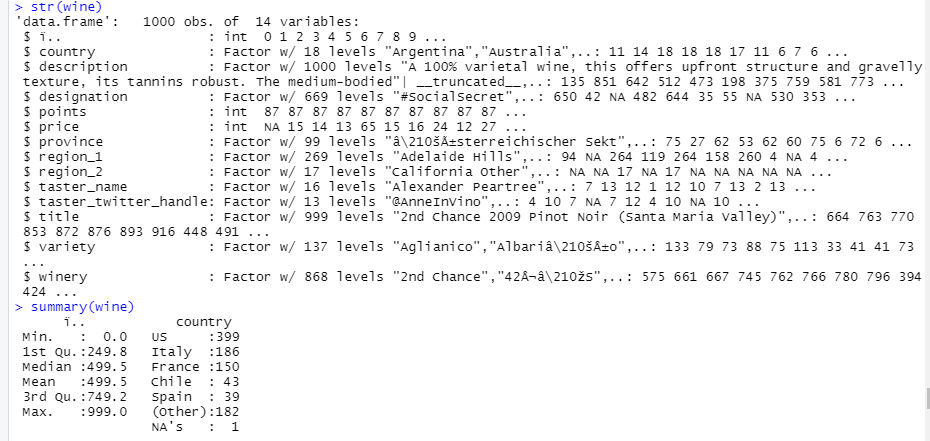
Final Project — Milestone 1

Milestone 1

**Output Summary Report**

**Name**: Rupali S. Metkari  
**Date**: 12th March 2021



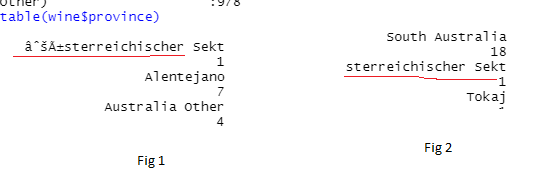
First, we load Wine\_tasting.csv data using ‘read.csv ()’ function. This data contains total 1000 obs. and 14 variables. In this data describe wine taste. In this data only price and points are numeric data, other variables are text data as shown above.

**Dataset variables description:**

1. Title: the title of the wine review, which often contains the vintage if we are interested in extracting that feature Variety: the type of grapes used to make the wine (i.e., Pinot Noir)
2. Description: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
3. Country: the country that the wine is from.
4. Province: The province or state that the wine is from.
5. Region 1: the wine growing area in a province or state (i.e., Napa)
6. Region 2: sometimes there are more specific regions specified within a wine growing area (i.e., Rutherford inside the Napa Valley), but this value can sometimes be blank.
7. Winery: the winery that made the wine Designation: the vineyard within the winery where the grapes that made the wine are from Price: the cost for a bottle of the wine
8. Taster Name: name of the person who tasted and reviewed the wine Taster.
9. Twitter Handle: Twitter handle for the person who tasted and reviewed the wine Acknowledgements.
10. Variety: This is explaining the type of wine.

**Cleaning data:**

In this dataset there are lot of missing values and in some entries data has special characters which is not required. It may be result of invalid entries from form. In the first step, I replaced blanks by ‘NA’ then filled data using ‘fill ()’ function in selected columns using the next or previous entry. For removing special character, I used ‘gsub’ function for replace special character from the text data. This way I made sure only proper data is there and every entry is formatted in similar way. Other way of cleaning data would be removing such values, but given the data contained 1000rows and many of them had mistakes, removing those rows would make the usable data very less. Hence, I choose to modify the data. If the data would have thousands of rows, removing some of the unusable data would have been a better choice.



As shown in above image (Fig 1) we can see in province column special character in text using ‘gsub’ function remove this special character output in (Fig 2). I used this function for other variables like region\_1, region\_2, taster\_name, taster\_twitter\_handle, title, variety. For price column I filled missing values by mean.

Descriptive statistical analysis and find subsets of data:

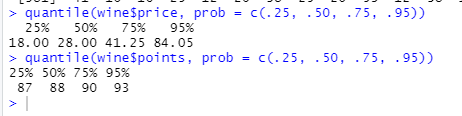


Fig 3

Quantile function used for how much of data lies between a certain value. In the quantiles, the 25th percentile is called as lower quartile, 50th percentile is called as Median and the 75th Percentile is called as the upper quantile. In Fig 3 shows quantile find for points and price column.

Find rank for wine price:

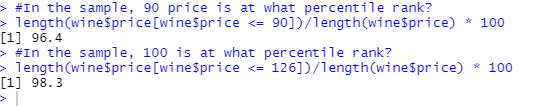


Fig 4

Fig 4 shows in sample 90 and 165 price is what percentile rank.

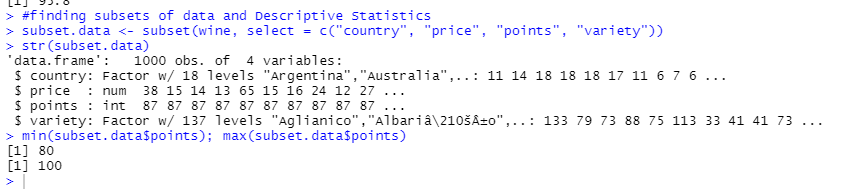
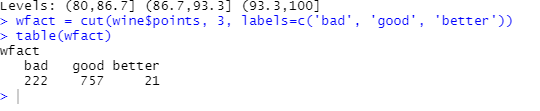
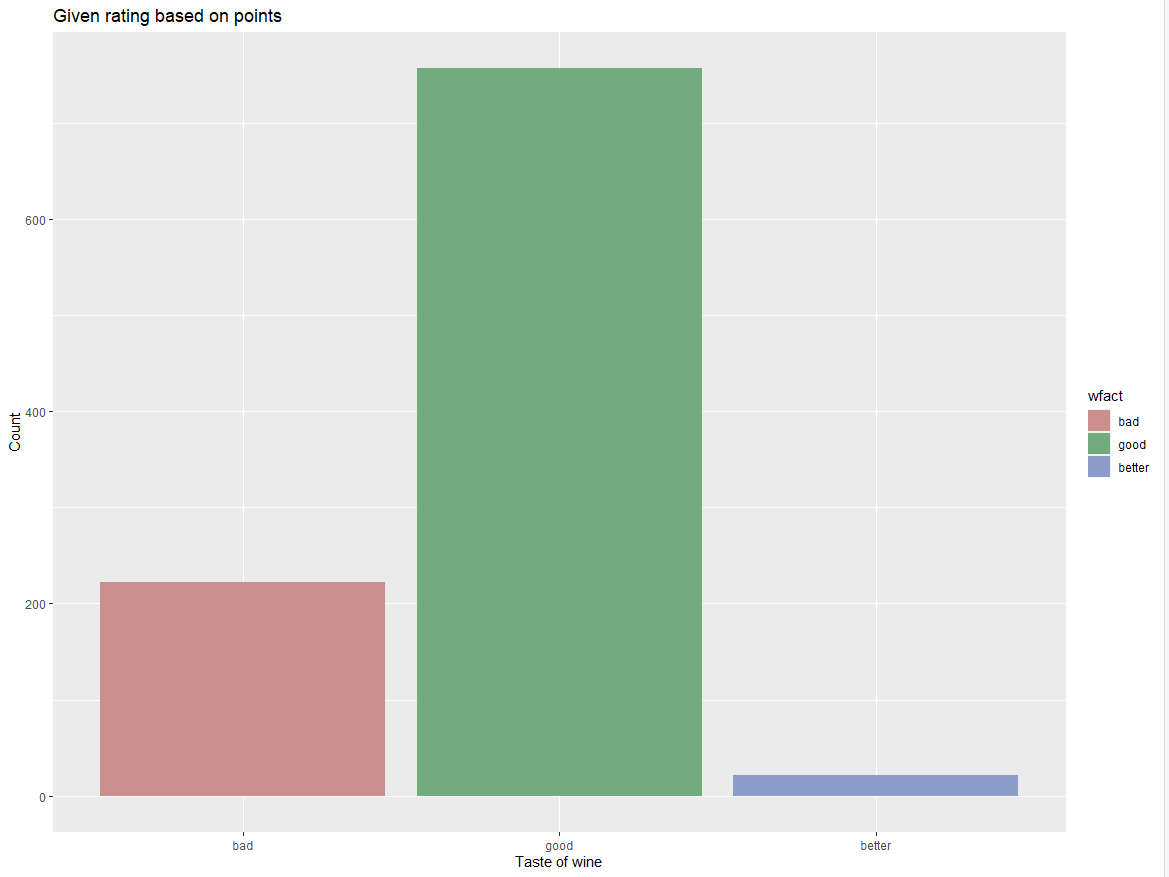


Fig 5

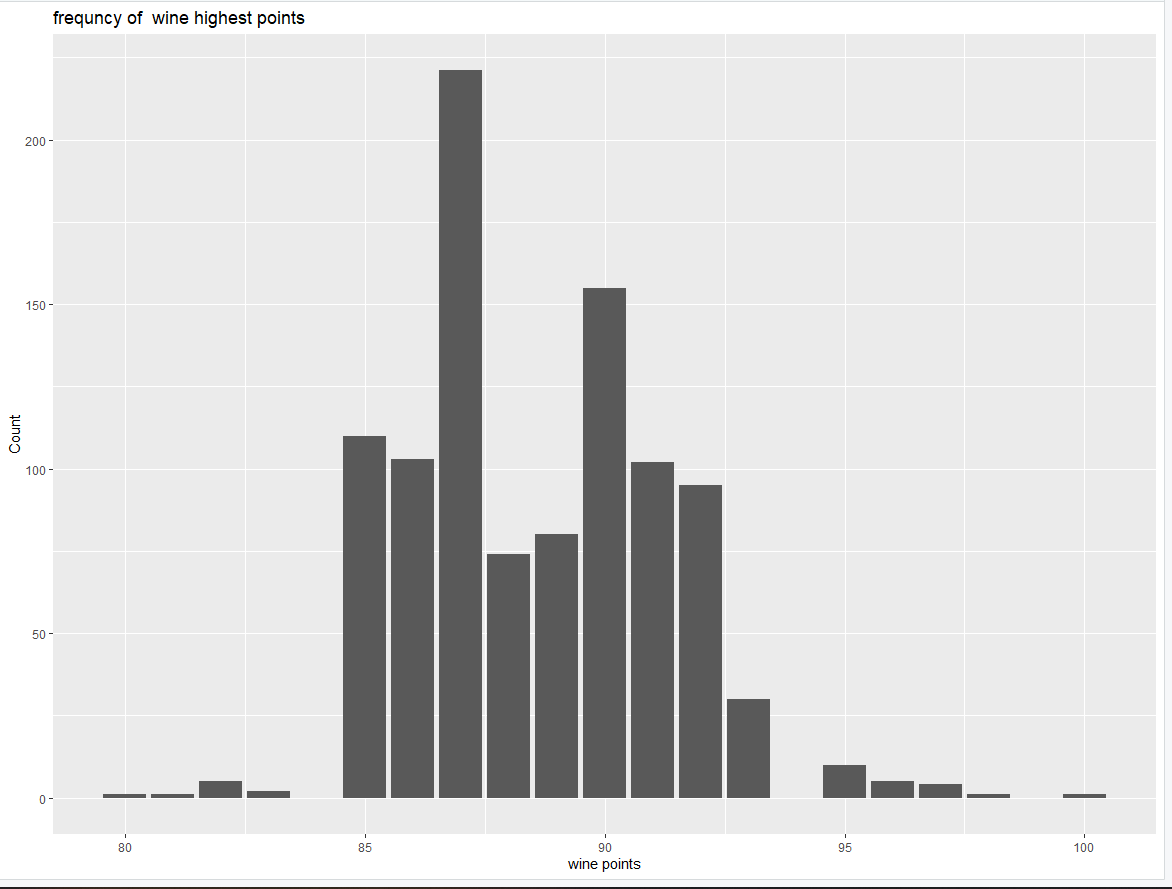
Now, finding subset of wine data (subset.data) including variables country, price, points, variety. Using ‘min’ and ‘max’ function find lowest and highest points for wine (Fig 5).



From the above diagram, I created three range for bad, good and better based on the levels given in the data. And the below figure shows the count of these ranges. We can conclude from the below diagram that most of the wines fall under good category.

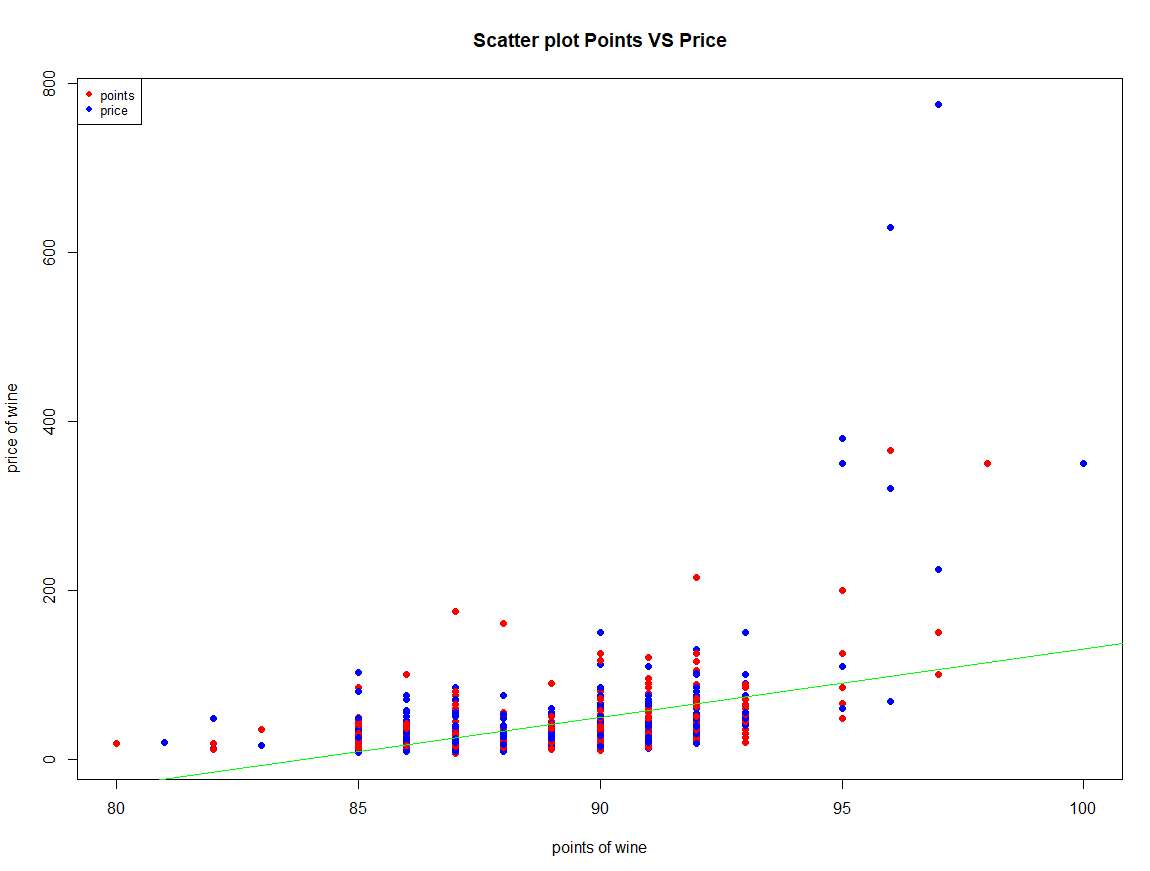


*Fig 6.*



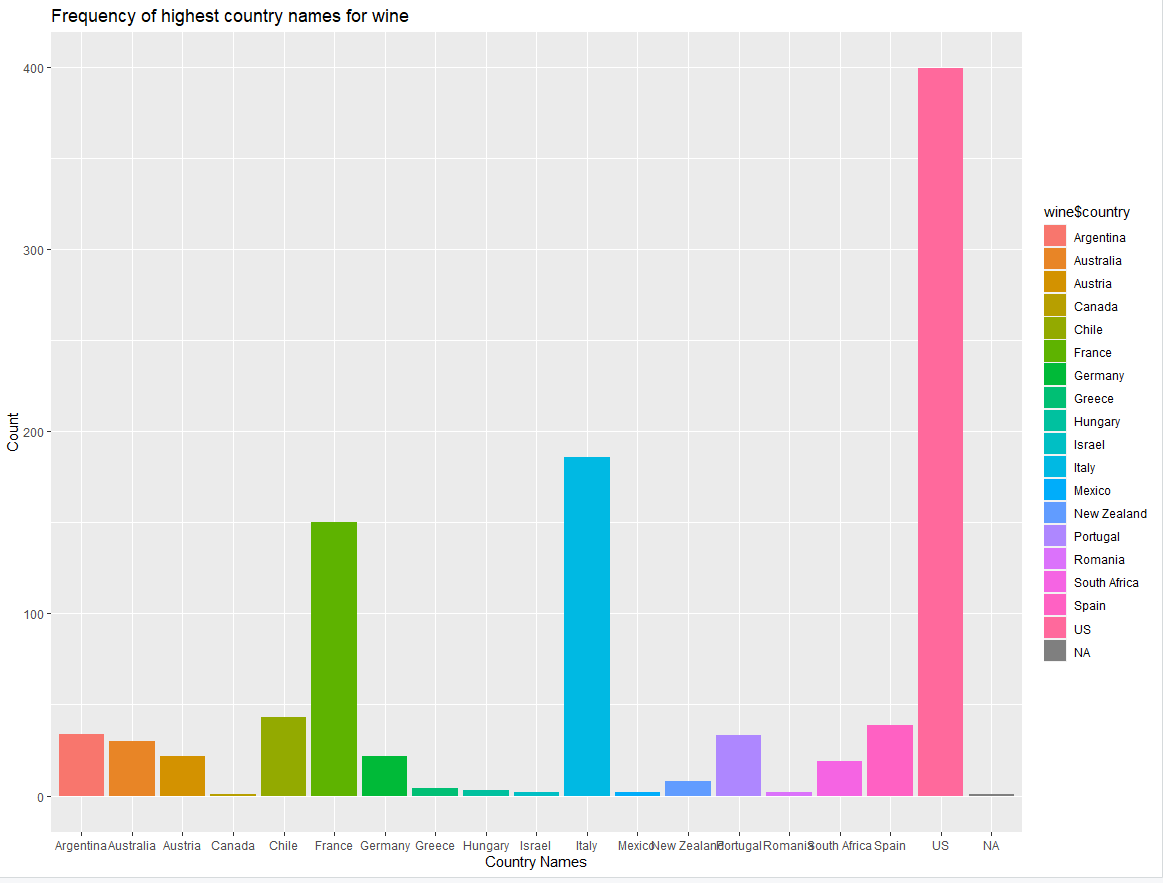
*Fig 7.*

In the above dig. (fig 7) we can see the frequency of points for wine. We can say most common range of points is 85 – 93. With 87 point being highest followed by 90.



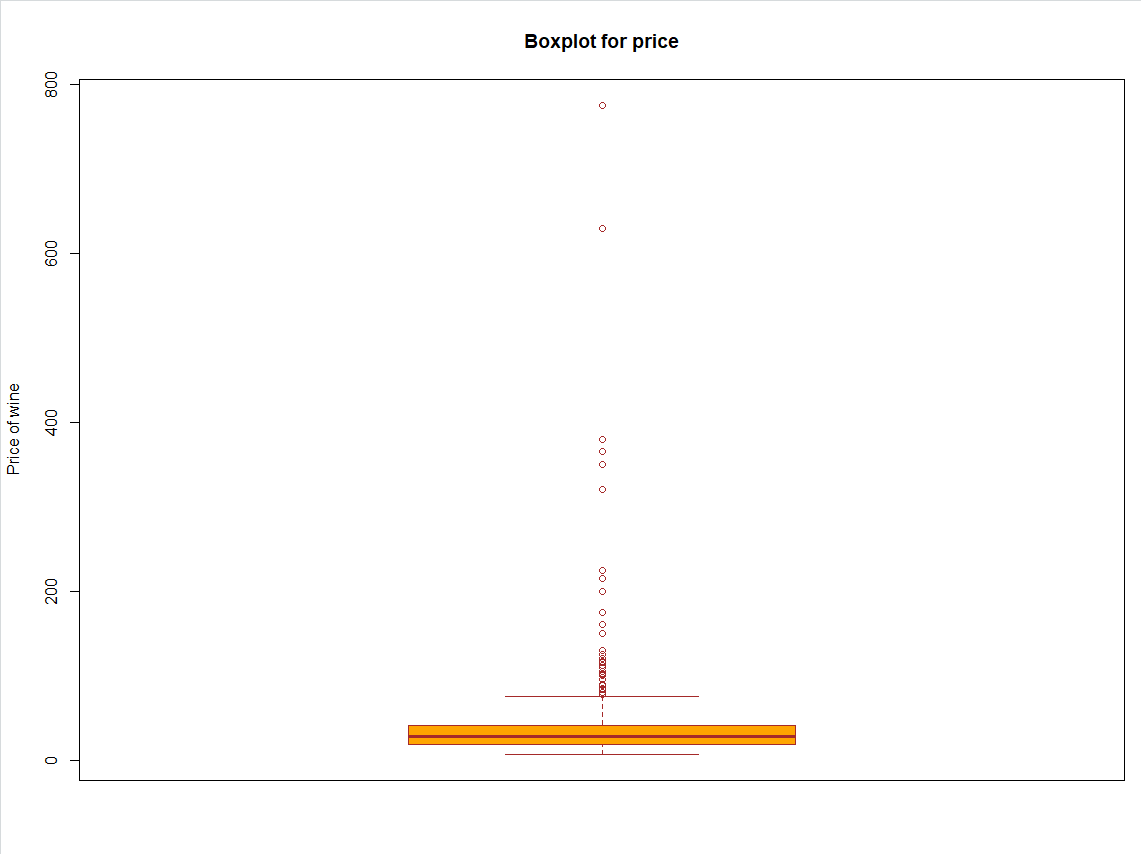
*Fig 8.*

In Fig 8. We have plotted a scatter plot of correlation between wine points and price of wine. We can see that wine who were rated from 85-93 had a similar price range, and wine with points higher than 95 priced more than 95. We plotted a regression line and can see the line closely fits the common price and point range.



*Fig 9.*

In the above fig. we can see the frequency of wine from which country it is from. Wine from US forms the majority of the data, followed by Italy and France.



*Fig 10.*

In the above fig, we create a boxplot for price to find out the range of the price of wine and any outlines. From above graph we can say that most wines are priced lesser than 80 dollars, but there are few wines which are expensive and go to 800 dollars.

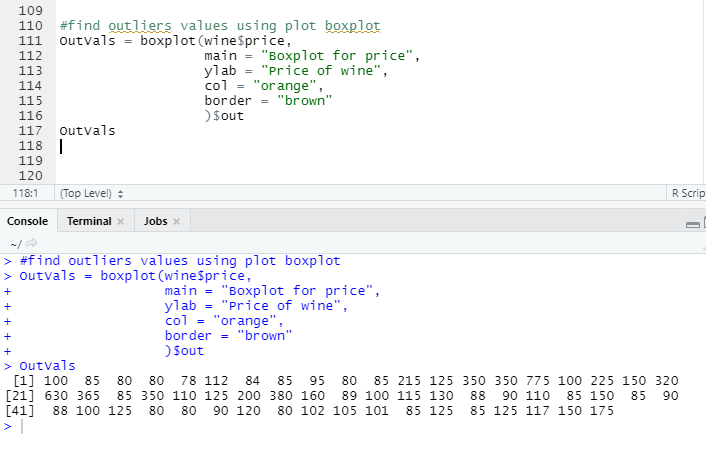


Fig 11.

From the above figure, we can see the outliners for price.

References

1. Allan G. Bluman Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2018 by McGraw-Hill Education. Elementary Statistics A Step-by-Step Approach, 10th edition
2. Robert I. Kabacoff (2011). *R in action Data analysis and graphics with R*. Shelter Island, New York: Publisher.
3. Fill missing values, Article(2018), tidyr: <https://tidyr.tidyverse.org/reference/fill.html>
4. Quantiles in R(2014) , dummies: <https://www.dummies.com/programming/r/how-to-check-quantiles-in-r/>
5. Subsetting data, Quick-R : [*https://www.statmethods.net/management/subset.html*](https://www.statmethods.net/management/subset.html)
6. Remove special character(2011), stackoverflow: [*https://stackoverflow.com/questions/10294284/remove-all-special-characters-from-a-string-in-r*](https://stackoverflow.com/questions/10294284/remove-all-special-characters-from-a-string-in-r)
7. Convert a character to numeric in R (2020), Statistics Globe: [*https://statisticsglobe.com/convert-character-to-numeric-in-r/*](https://statisticsglobe.com/convert-character-to-numeric-in-r/)