Wine Reviews Capstone Summary

Jacqueline Canty

2024-02-22

Welcome! Let's explore some data together.

1.Summary

After watching the documentary "Somm", about 4 individuals trying to pass the nearly impossible Master Sommelier exam, written and directed my Jason Wise on Netflix, a fellow "techie" decided to make something out of it for the advancement of future generations' technology.

This is information from a dataset created by the user @zachthoutt on Kaggle from data scraped from WineEnthusiast in June of 2017. The idea behind this data set is that the descriptions and ratings from world class sommeleirs would allow machine learning models to identify the wine through a blind tasting without ACTUALLY tasting it.
The creators overall goal was ".. to create a model that can identify the variety, winery, and location of a wine based on a description."

This analysis uses this data to try and find the best overall wines with the information provided.

2. Ask

Business Task:

Summarize the data to show which wines are considered the best at specific price points to generate a marketing strategy to increase sales.

Stakeholders:

- Major heads of wineries

- Marketing teams at previously mentioned wineries

- The general public above legal age

Dataset info:

- Originally 3 separate files, but for the purpose of this analysis, only the first original dataset will be used.
- Scraped using this code

Data Credibility:

- The original data was scraped and created by the same owner - not generated by an accredited software or company

- The original data was collected in 2017 - Not current
- There is possible sampling bias - I did not use all 3 of the files. SEE PREPARE

3. Prepare

Dataset Info:

- 1 cvs file scraped and created by a user on Kaggle found here
- We will focus on the type, price and amount of points received.

Dataset Credibility:

- Data is limited to the ratings from WineEnthusiust
- Data is not current - From 2017
- Possible sampling bias as the ratings are from regular indivduals as well as sommeleirs and **not enough of either to represent the population as a whole**

These factors in mind, this case study will take a targeted approach to be able to extract the data we need.

4. Process

Download and Clean the Data

I chose to begin with Excel. This proved unwise as the dataset has approx. 150k rows. Excel is best used with smaller datasets of around 100 or so rows.

To make things simpler, I removed the duplicates.

Next, I deleted the column I wouldn't be needing : region_2 and the column with a blank header

After that, I removed all of the blank cells.

Finally, I changed the values of the cells in the "price" column to currency AND filtered the results to only show wines with **95 points or higher.**
Note: I plan on using this data again in the future for additional analysis expanding on this project.

Oh no! It's not ready yet! As I'm looking through my data, I see that there are special characters i.e #, $, % in places where normal letters should be.

To take care of this manually, the function would look something like this :
**SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUTE(SUBSTITUT
"e"),"Ã¢", "a"),"Ã¡","a"),"Ã³","o"),"Ã‰", "E"),"Ã¨", "e"),"Ã", "i"),"Ã¶", "o"),"Ã©", "a"),"Ã´", "o"),"Ã¼", "u")**

But, that's a lot.
So, I used the Find & Replace function instead.

Taking it a step further, I combined the "region_1" and "province" columns to make the new "region_1" column using this formula :
CONCAT(E:E," , ",F:F)

For future analysis sake, I also changed the column header "*points*" to just "points".

And just like that, our 150k row dataset is now 9,226 rows. ***AND*** **my computer is running much faster now.**

5. Analyze

Transforming the data with SQL!

I chose to transfer the dataset over to BigQuery because SQL allows for analysis of large datasets such as this to be done quicker and easier than Excel.

Now, I'm no Master Sommeleir, but I do know that were are 3 major types of wine: *Red, White, and Sparkling*

So, using BigQuery, I've added a column that helps people like me know which type of wine is red, white or sparkling. **Here's how:**

- I filtered the data to show me the distinct types of wine using GROUP BY- This returned 37 types
- I researched of those 37 types, which was red, white, or sparkling
- I established the unique values within the 37 types for this code:

The CAST function saves the query results as a new table, but thats not helpful to me so lets combine the new table with the results and the old table. I accomplished this using the JOIN function

I now have a new table with the row "red_or_white".

I want to see if I have an outliers in my data as far a pricing, so I use this code:

I see that I *do* have outliers, but for the sake of this analysis, I believe that this data is beneficial so I will not remove it. I will use it for my analysis.

6. Share

Visualizing Data with R & Tableau

With this newly cleaned data, we can now show the answer to the question of what wines are best through visualization!

Tableau is one of favorite visualization tools for one reason,* it's simplicity.*

I was able to create these amazing visualizations that are clear and concise. They are user-centric with interactive capabilities. You can enjoy them HERE And HERE

But, I want to add a little bit more to my visualizations. **Let's move over to R!**

I want to specify different price ranges for stakeholders to choose from while allowing them to see the different types of wines.

First, I create a data frame customizing the price ranges.

This differentiates the names and values within the price ranges created. This perform a mutation to create the "price_range" variable.

```r
library(dplyr)
wine_reviews <- wine_reviews %>%
  dplyr::mutate(price_range = case_when(
  price <= 49 ~ "Less Than $49",
  price >= 50 & price <= 100 ~ "$50-$100",
  price >= 101 & price <= 200 ~ "$101-$200",
  price >= 201 & price <= 400 ~ "$201-$400",
  price >= 401 ~ "Over $400"))
```

```
price_range_colors <- c("Less Than $49" = "green",
                        "$50-$100" = "green4",
                        "$101-$200" = "yellow",
                        "$201-$400" = "orange",
                        "Over $400" = "red")
```

Now that we have the ranges looking great, let's plot them!

```
library(ggplot2)

splot <- ggplot(data = wine_reviews) +
   geom_point(aes(x= points, y= price, color = price_range, shape = red_or_white))+
  scale_color_manual(values = price_range_colors) +
  labs(x = "Points", y = "Price" , color = "Price Range", shape = "Red or White")  +     facet_wrap(~re
    xlim(95,100)
```

6. Act (Conclusion)

Results

- We found that **Napa,California** has the best options for high quality, inexpensive wine across the three major types.

Recommendation

- **For owners of wineries-** Consider partnering with a Napa winery listed to study the creation process or create a collaboration to increase sales and marketing strategies.
- **For the general public-** Consider shipping the wine to your house if you're not near Napa, California or look into finding brands located in the Napa, California region at your local store.

Additional Comments-

- California as a whole has the highest ratings and the **Most** affordable wine options.

- Note that this data was not indicative of **ALL** wine reviews and wineries. The sample size is very targeted.

Thank you so much for joining me on this data journey!

All comments, questions, and suggestions are highly appreciated. I'm happy to share my journey with you.