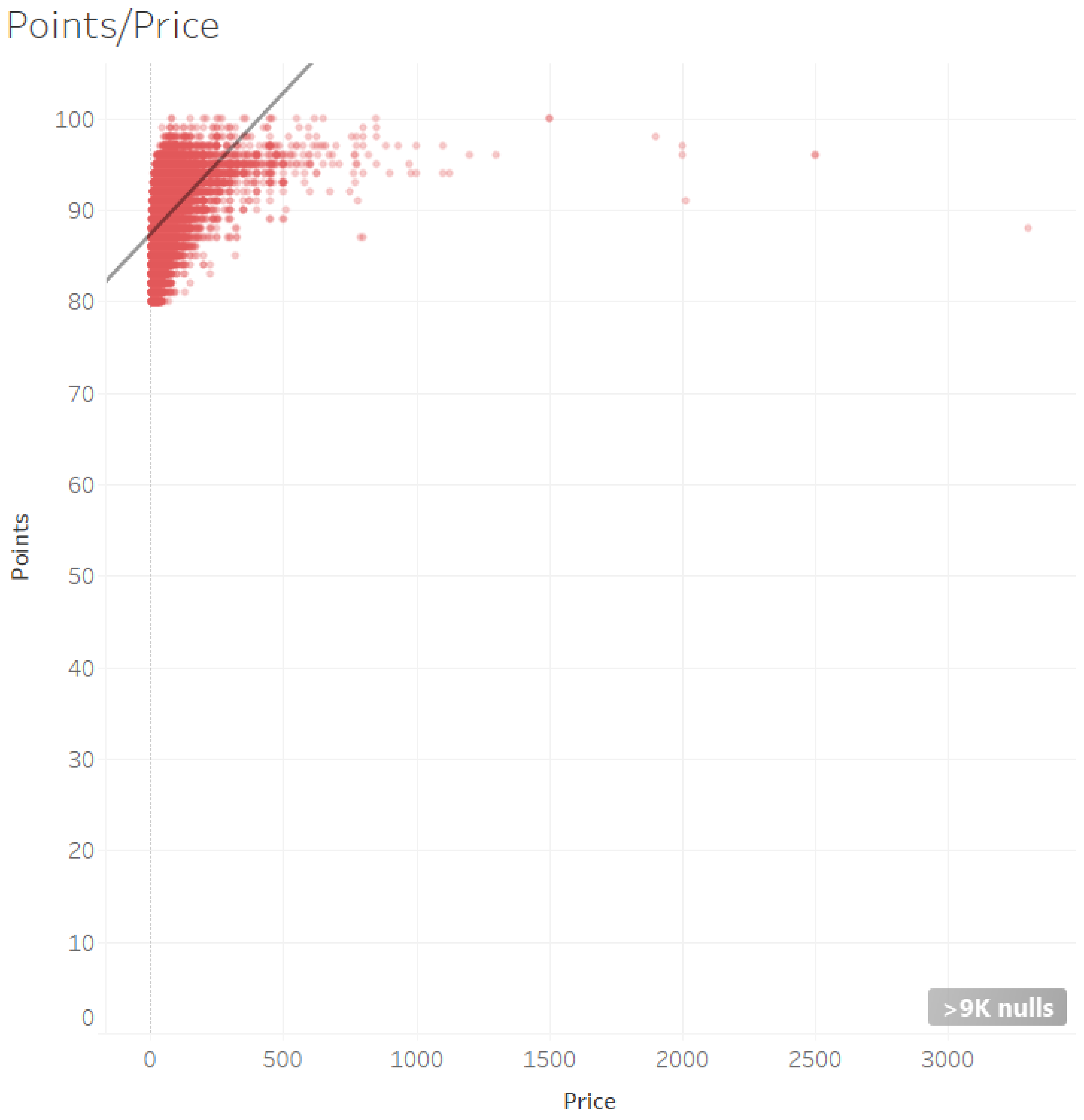
Exploratory Data Analysis

I chose a dataset on wine reviews that I found on Kaggle. recently, I've delevoped a taste for certain wines so I am curious to learn more about which wines have the most value. I am also interested to see if the human palate can distinguish the difference between a "high quality" wine and a cheap wine because taste is subjective. I am hypothesizing that while there may be a slight increase in score per increase in price, there will be a lot of variability.

Initial Question:

Do higher wine prices increase the wine's rating?

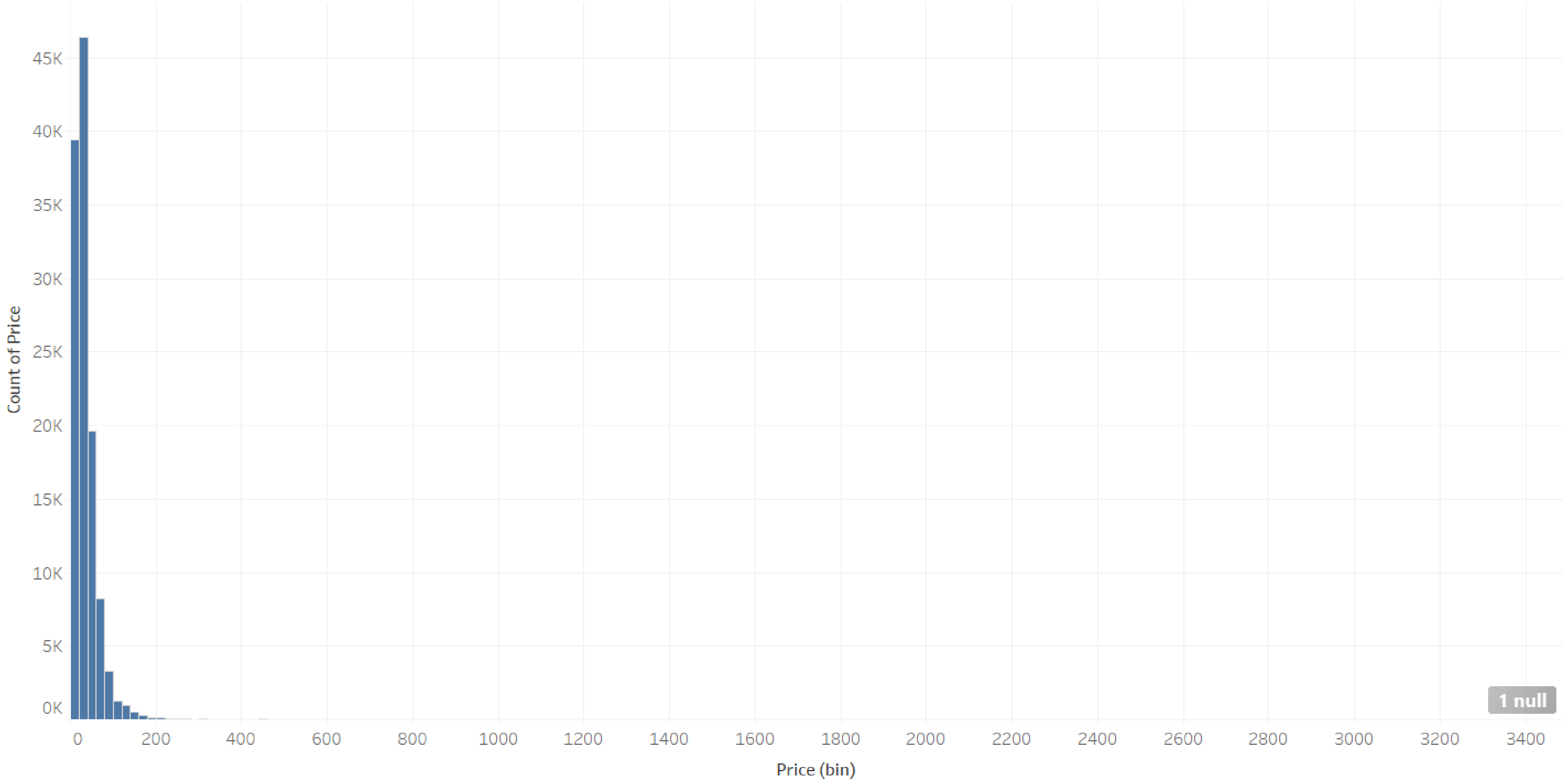
I created a scatterplot to visualize the points per price to see if there is a positive correlation between the two. While it is safe to say that there is a positive trend between the two variables, there also appears to be diminishing returns for the increase in quality to price anywhere beyond about $500



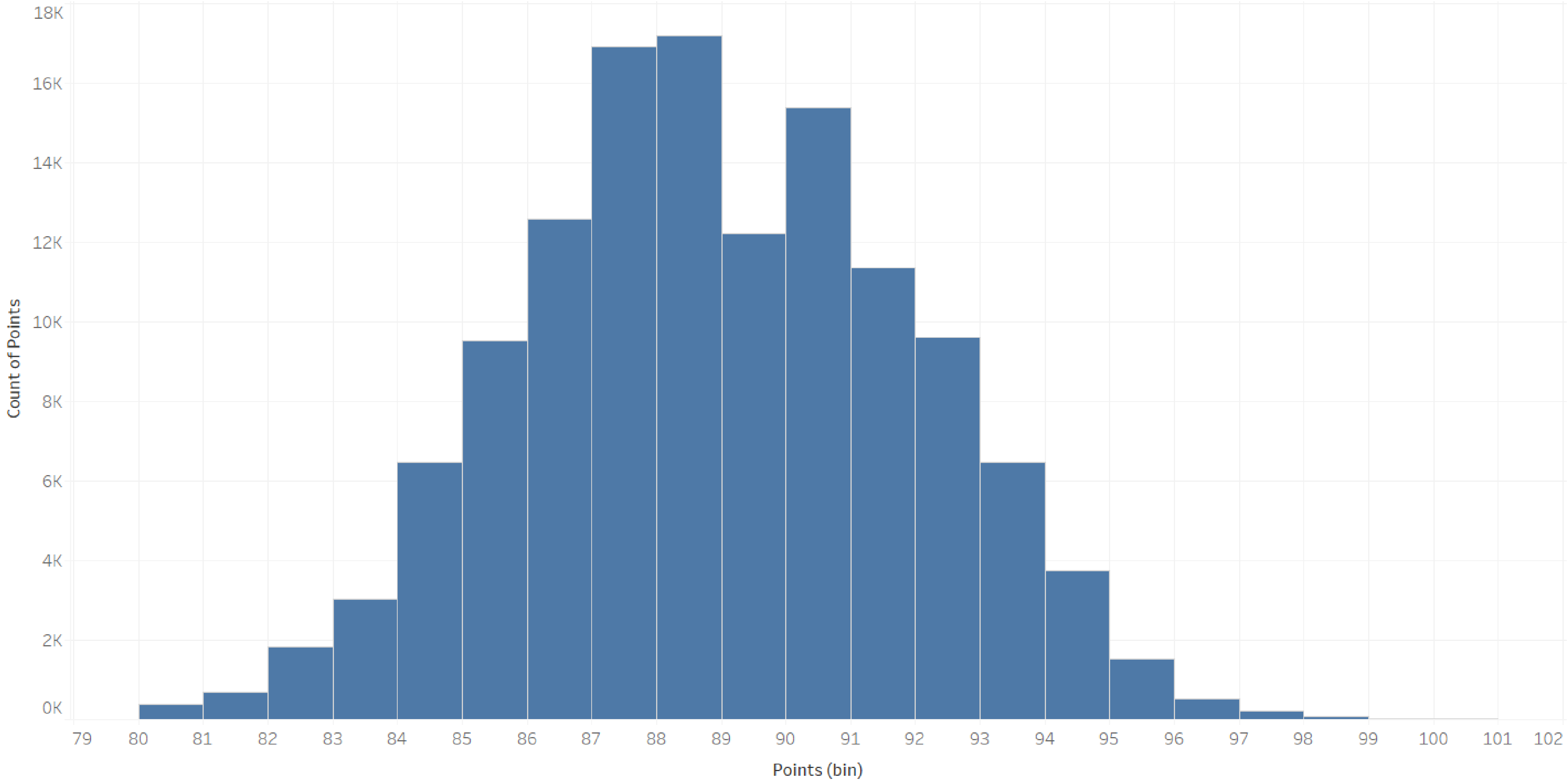
While the positive trend between price and points was apparent, there were a few outliers and low priced wines that scored 100. I didn't think price alone was a strong enough indicator of the wine's point value, so I decided to figure out what these outlier wines had in common that gave them higher points compared to the higher priced wines.

Before diving deeper into the exploratory data analysis I needed to check the distributions of the points and the prices. I created a histogram of the counts of each price (bins of 20) and points (bins of 1). The price had a right skewed distribution so I opted to using median for future price analysis. The points had a normal looking distribution so I decided that using mean would be appropriate for future analyis.

price hist

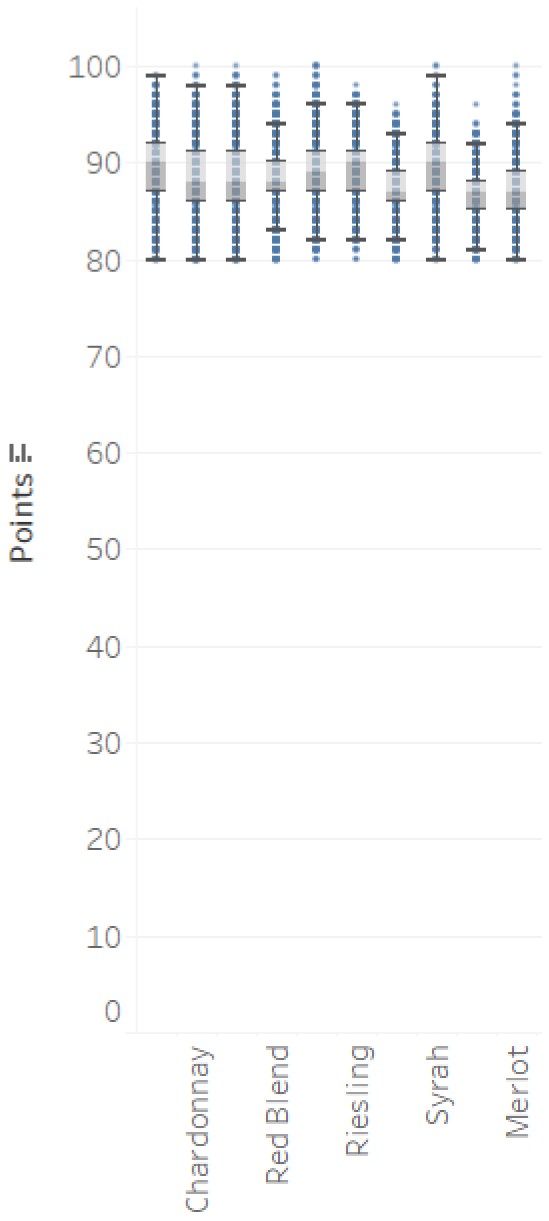


point hist



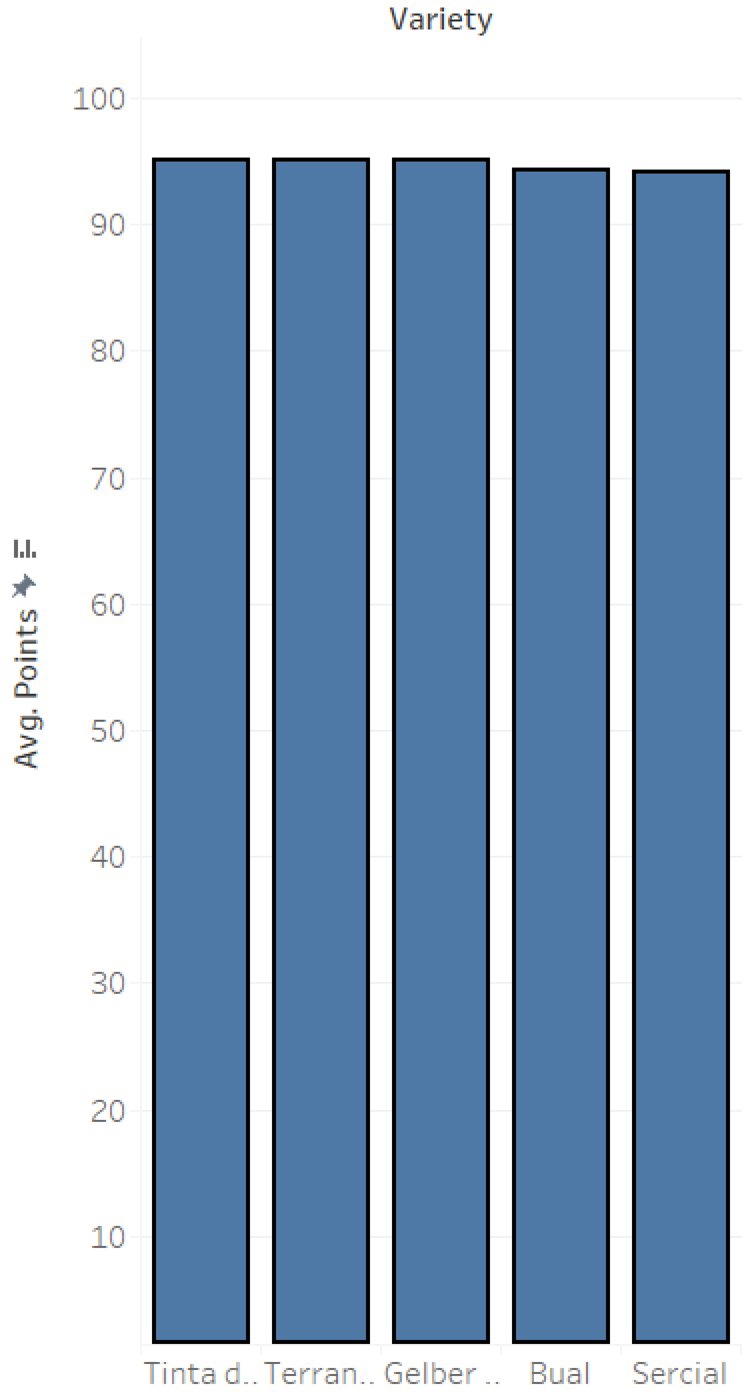
I started by looking at the points awarded per variety of wine. I used a boxplot and sorted by descending points to discover the "best" varieties of wine. The top 5 wines that appeared were Chardonnays, Red Blends, Rieslings, Syrahs, and Merlots.

points/variety box

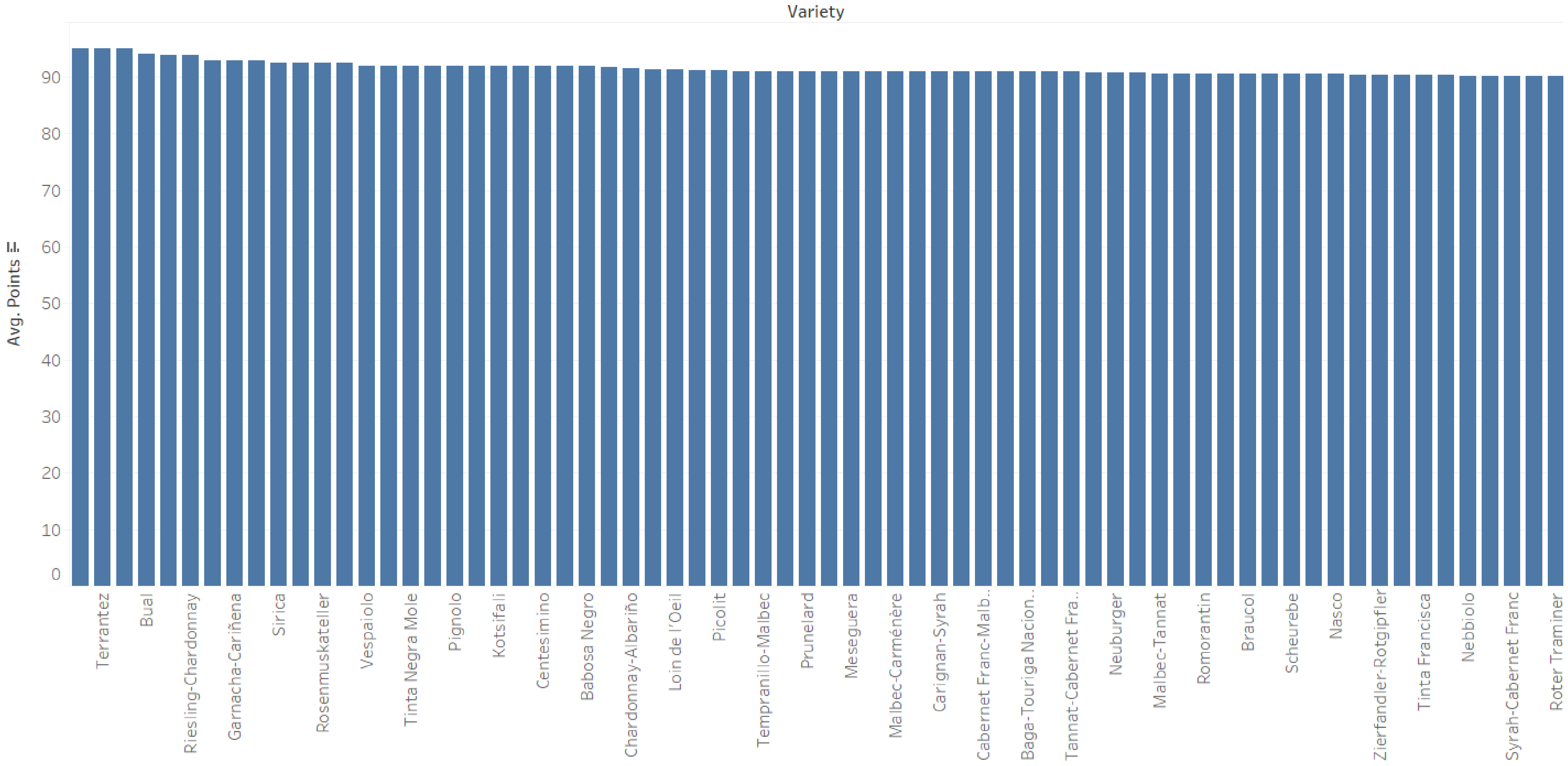


I attributed this discovery to the high frequency of these common types of wines, which gives for more results which could increase the chances of better scoring wines appearing. Following this line of thought I figured that looking and sorting by the mean alone would be more appropriate. The top 5 scoring wines by mean were Tinta del Pais, Terrantez, Gelber Traminer, Bual, and Sercial.

avg Points/Variety

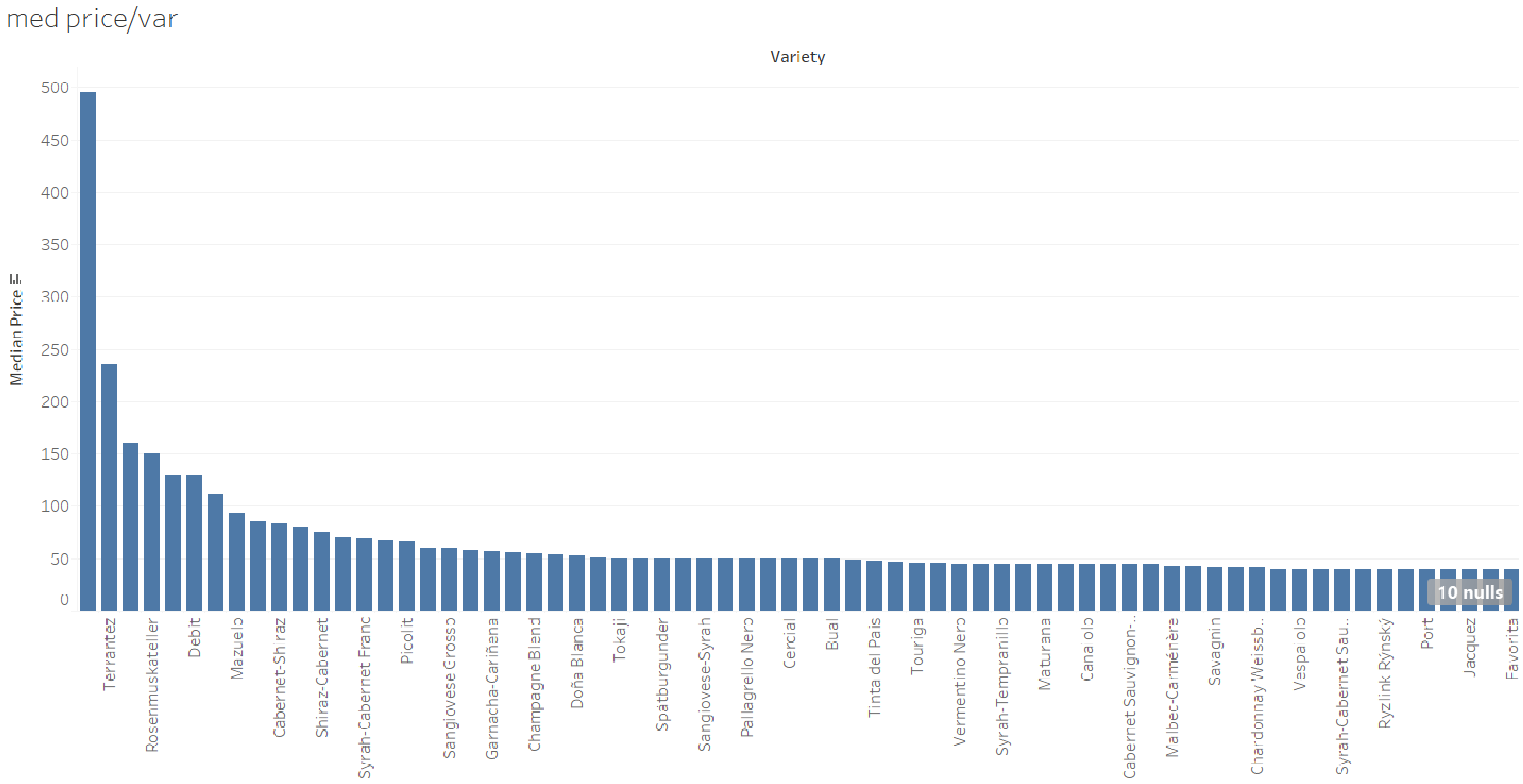


avg pt/var full



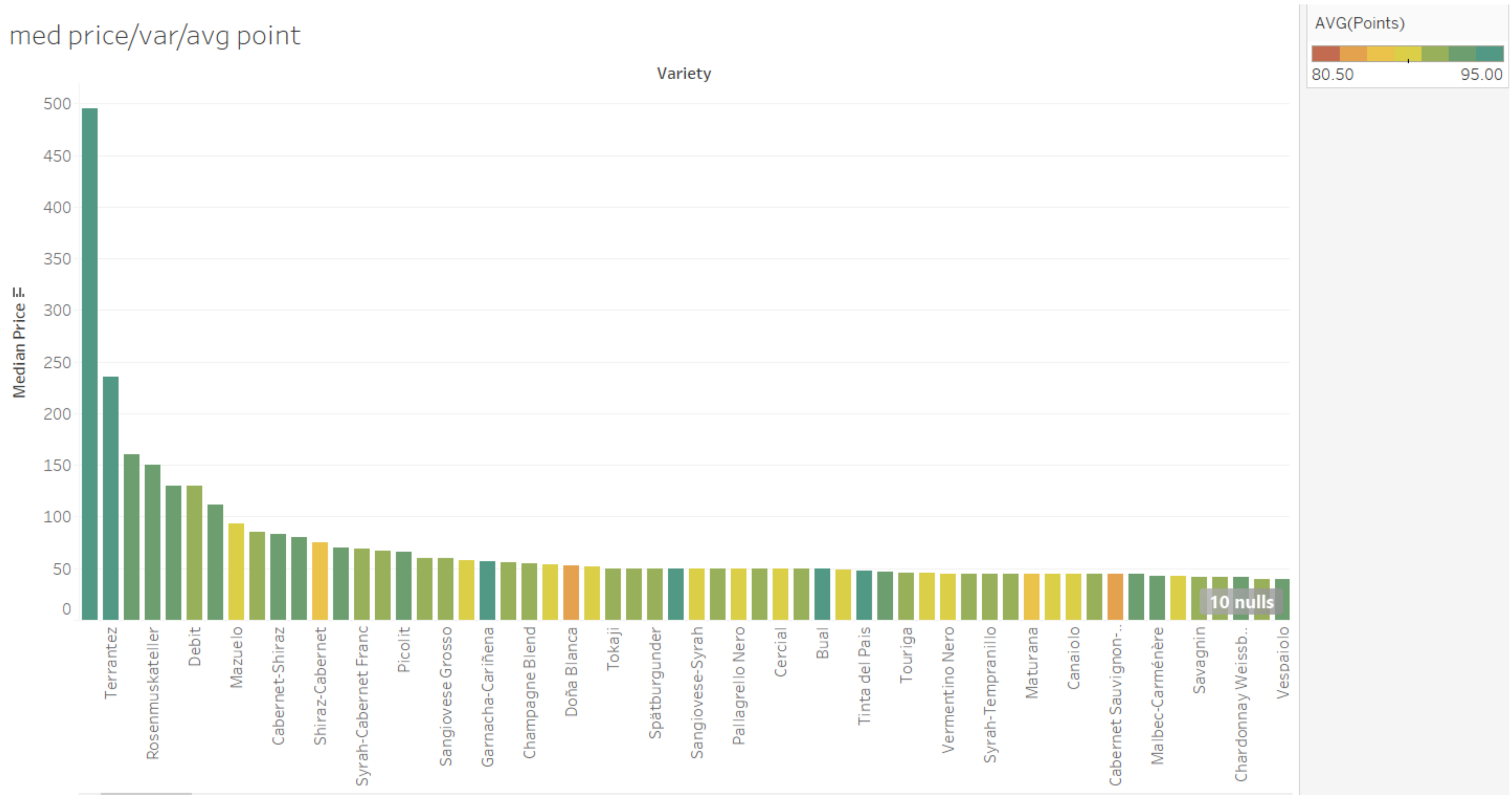
if we compare this to the median price per variety, the top results are Ramisco, Terrantez, Francisa, Rosenmuskateller, and Malbec-Cabernet.

we also see this distribution:



If we compare the point/variety graph to the price/variety graph we can see a sharp difference in prices compared to the relatively small changes in points.

I decided the best way to display three variables (median price, mean points, and each variety of wine) would be to use the median price/variety graph and encode the respective points to each bar as a color.



In this final graph we can see that there is no apparent correlation between the price and points of the wine varieties, as there are some dark green bars on the right tail of the graph where the price is cheaper.