The relation of Whisk price with age and alcohol content

- **Graph Objective**

The graph objective is finding the relationship of whisky price with age and alcohol content. For this data set, I think that we should process the price and year, the price and alcohol content of whiskey separately.

In fact, I tried to use the Newton–Raphson method and the three-point method and the bulging rule to analyze the relationship between the whiskey price and alcohol content, and I found that there is almost no relationship between the price and alcohol content. In the results obtained by the three-point method, the model explains 2.7% the variability of the response data. So I chose to focus on the relationship between whiskey price and years.

- **Data management**

This data set mainly contains 3 features, price, years and alcohol content.

Some whiskey prices are measured by liters or by set. I choose to check the whiskey information on the Internet to calculate the actual price, if there is no information about that whiskey, I'll delete this record.

Both the year and the alcohol content contain many missing values. Because I analyze alcohol content and price separately, these missing values have no effect. I divided the data set into price and year, price and alcohol content two tables. Then, I delete the null records.

- **EDA analysis**

First, I process the kernel density estimator on original data and found that all features are not normally distributed. And I recorded the skewness of each features. The years and alcohol content are mild positive skewed. The price is highly positive skewed.

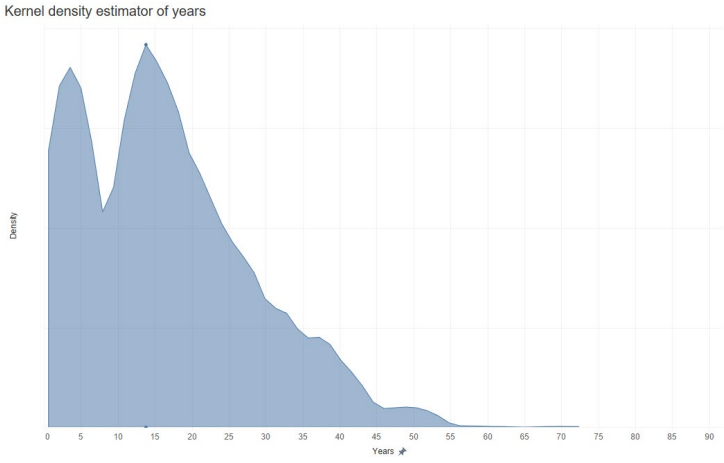

Figure1.Kernel density estimator of years.

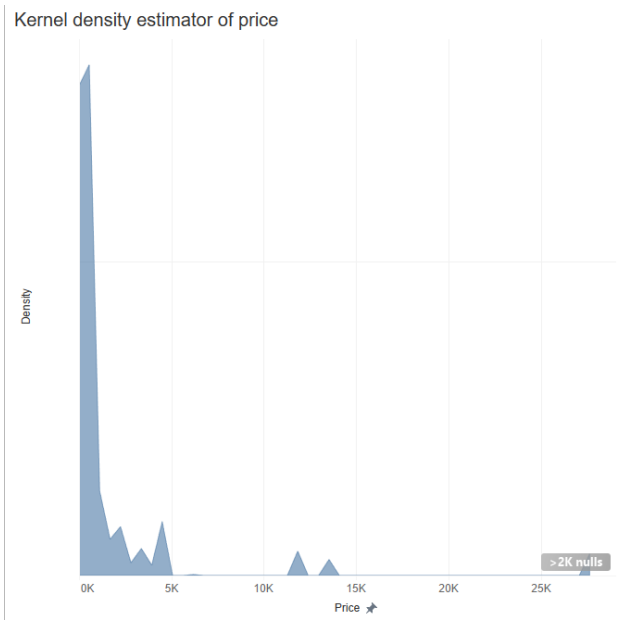

Figure2. Kernel density plot of price

And after the processing of the relationship, I test the ln price by Quantile Normality plot.

- **Transformation analysis**

  First, I exclude the alcohol content determinants. Because the regression and R-squared of the model shows the relationship between price and alcohol content is indescribable. And the optimization transformation of the alcohol content is terrible.

  Then, I observed with Kernel density estimator that neither price nor years are normally distributed. Because I mainly analyze the global relationship, I use Epanechnikov kernel.

  Third, I use the Newton–Raphson method and the three-point method and the bulging rule to analyze the relationship between the whiskey price and years. The three-point method model performance better. And the bulging rule indicates that I could use the ln price or power of x to do the normality. The R-squared indicates the ln price has better effect.

  Fourth, I need to verify that ln price conforms to a normal distribution. I chose to use Quantile Normality plot to verify.

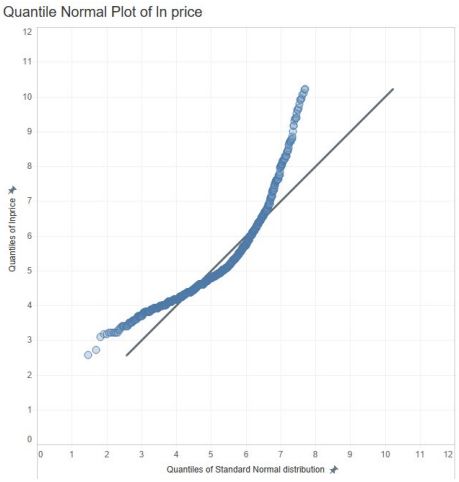

Figure3. Quantile Normality plot of ln price

In fact, ln price does not follow the normal distribution, especially near the extreme value.

I also tried some different power of years. However, the R-squared is not satisfied than the original years. Finally, I used regression to express the relationship between price and years.

The final model is ln price = 0.0459887*Years + 4.3291.

- **Visual Implantations and retinal variables**

In the final dashboard, use different colors to represent the price and years charts for easy distinction. Using dark border to describe the area edge of kernel density plot. Using the text box to mark the detailed information of the model

In the QN diagram, normality is shown in gray and distinguished from ln price.

- **Insight and conclusion**

There is a positive correlation between whiskey price and year, the higher the year, the more expensive the price.

The model is ln price = 0.0459887*Years + 4.3291. It means this model explains 22% the variability of the response data. Many price variations are difficult to explain using only the years.

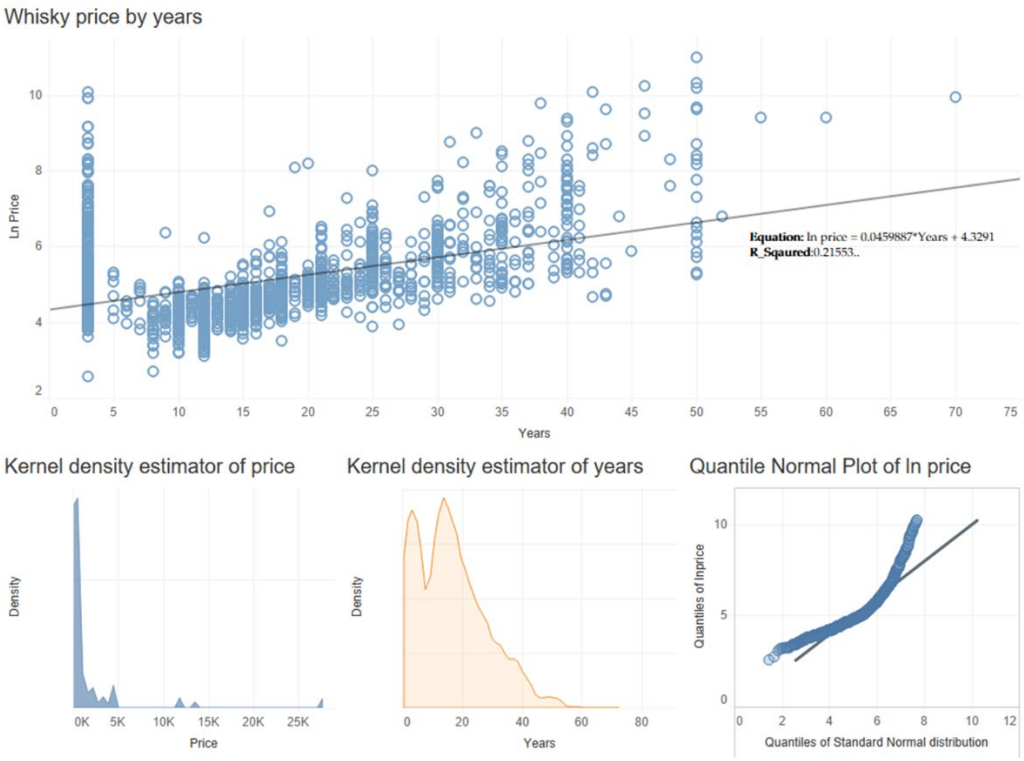

Figure 4. Dashboard