ProblemSet1

Math 286

Aaron Graybill

2021-02-18

Data Description:

In following section I read in the data and find its dimension:

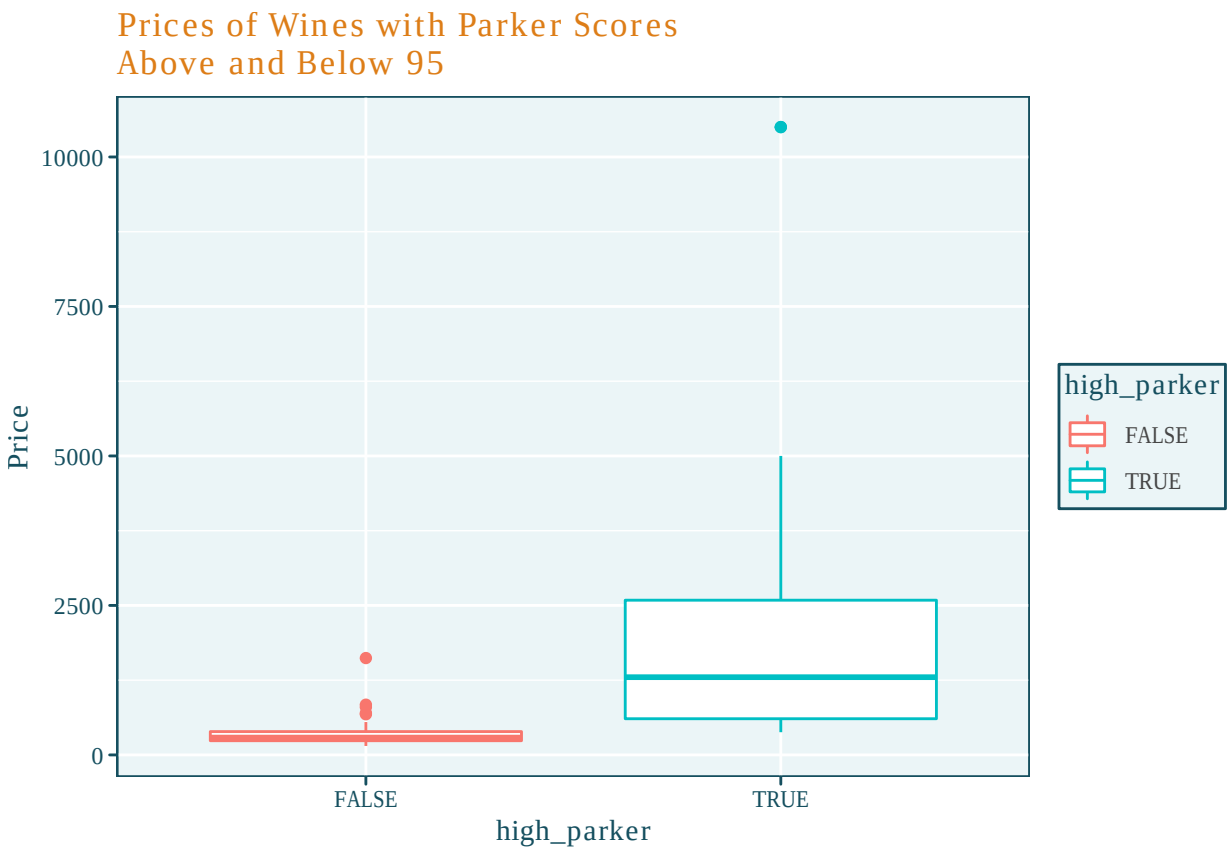
```
df <-
  read.csv(here("Data","Bordeaux.csv"))
df_dims <-
  dim(df)
```

Using the above computation, there are 72 observations, and 9 variables.

Statistical Analysis

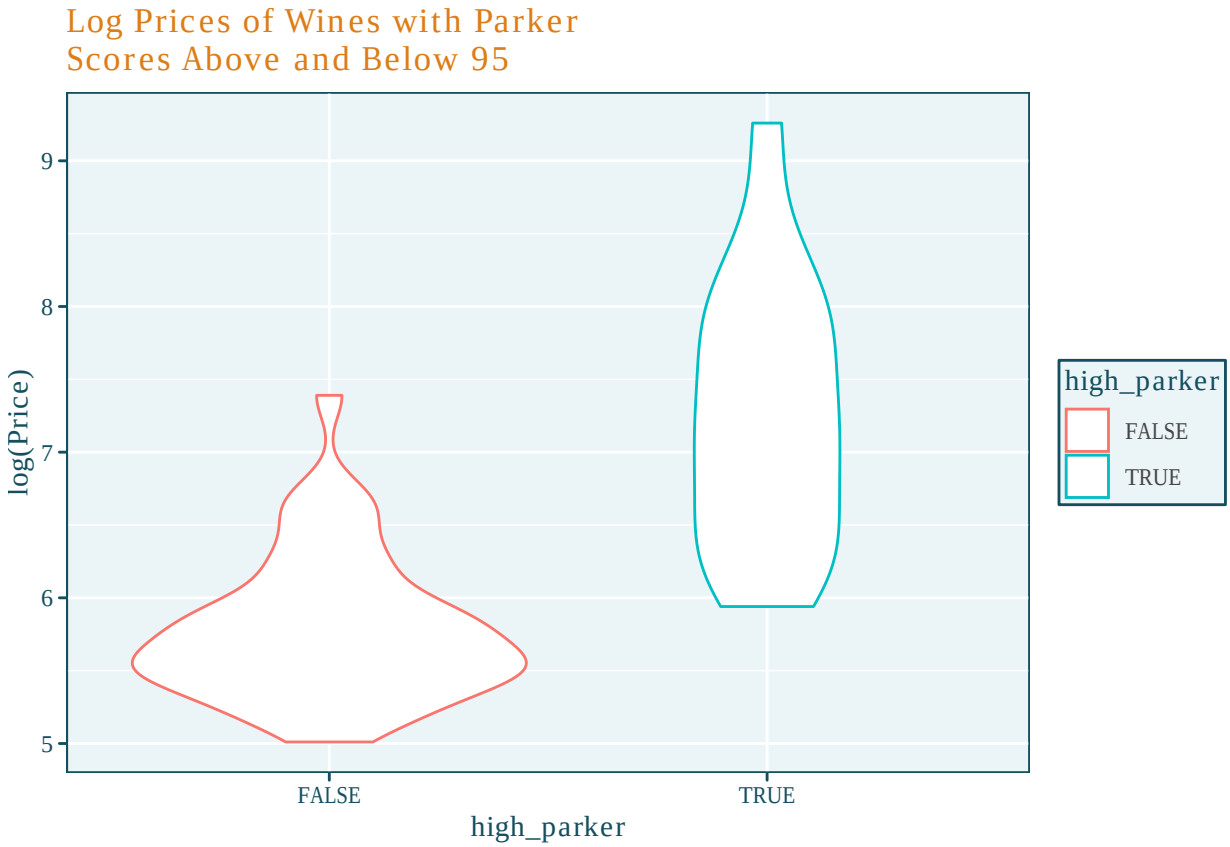
First I will visualize the distribution of prices of wines with Parker scores below 95 and above 95 respectively.

```
df <-
  df %>%
  mutate(high_parker=case_when(
    ParkerPoints>=95~TRUE,
    TRUE~FALSE
  ))
ggplot(df,aes(x=high_parker,y=Price,col=high_parker))+
  geom_boxplot()+
  ggtitle("Prices of Wines with Parker Scores\nAbove and Below 95 ")+
  theme_custom()
```



The large outliers make it hard to see that relationship in detail, for completeness, let's look at the distribution of log prices:

```
ggplot(df,aes(x=high_parker,y=log(Price),col=high_parker))+
  geom_violin()+
  ggtitle("Log Prices of Wines with Parker\nScores Above and Below 95 ")+
  theme_custom()
```



I mostly included that because the density plot of log price looks like a wine bottle for high-Parker-score wines!

That digression aside, the graphs seem to indicate that wines with high Parker scores are generally more expensive. Let's see if we can back that with some summary statistics:

```
df %>%
  group_by(high_parker) %>%
  summarise(Min.=min(Price),
            `1st Qu.`=quantile(Price)[2],
            Median=quantile(Price)[3],
            Mean=mean(Price),
            `3rd Qu.`=quantile(Price)[4],
            Max.=max(Price),
            `sd`=sd(Price)
            ) %>%
  knitr::kable()
```

high_parker	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
FALSE	150	237.5	290	376.500	390.0	1620	263.4369
TRUE	380	605.0	1300	2187.656	2587.5	10500	2483.2909

So the measures of central tendency seem to indicate the wines with high Parker scores also have higher prices. That being said, the standard deviation of the wines with high Parker scores are quite high, so we should run a t-test to ensure that the difference in means is significance.

```r
high_parker <-
  df %>%
  filter(high_parker==T) %>%
  pull(Price)
low_parker <-
  df %>%
  filter(high_parker==F) %>%
  pull(Price)
#un-paired two sided t-test
test <- t.test(high_parker,low_parker)
test
```

```
##
##  Welch Two Sample t-test
##
## data:  high_parker and low_parker
## t = 4.1073, df = 31.559, p-value = 0.000264
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   912.458 2709.855
## sample estimates:
## mean of x mean of y
##  2187.656   376.500
```

The above results show that the p-value from the test is $p = 0.0003$. This is less than the stated threshold of $p^* = 0.005$. We can interpret these results in the following way. The t-test concluded that there is a significant difference between the mean prices of wines with a Parker score ≥ 95 than those wines with scores < 95.