Report 1

Naomi Zubeldia

10/20/2022

Abstract

Since it was discovered, gold has been a coveted mineral. It is now more commonly in the jewelry market. There are times of year that people want to buy gold whether for a special occasion or for investment. This analysis is specifically looking for what times of year the prices of gold are higher. Since most holidays that inspires gift buying or investing are in the winter months, the prices of the months of June and December will be compared. This analysis was able to find a significant difference between the two months and make the conclusion that the prices in December are higher, which supports the hypothesis. It is advised that future analysis be made to include factors of influence such as recessions and gold weight sold.

Contents

1 Introduction

Recently, I got married and was proposed to with a diamond ring with a gold band. In the midst of ring shopping, it was fascinating how expensive precious metals and stones are. Since gold has been sought after since it was discovered, the question arose if there was a better time of year to invest in gold because of lower prices. According to a statistical study done by analysts at GoldSilver, on average gold is a slightly high price at the beginning of the year and then lowers in the spring and summer to then rise higher towards the end of the year. Jeff Clark (2021)

It is interesting how gold prices seem to fluctuate multiple times within a year especially around the colder months. Those months are around the times of year when there are more holidays, which usually incites the desire to buy gold for gifts and new year resolutions. Therefore, I chose to analyze whether gold prices are lower due to it being a colder more wintery time of year versus the spring and summer specifically by comparing the prices of June to December.

The first step of this procedure is to clean the gold monthly data so that the dates are uniform enough to group by. Then, it is filtered to pull out the prices of June and December. An exploratory data analysis will then be completed on the two groups in order to determine their characteristics.

Next, if necessary, transformations are made to fulfill the assumptions of normality for testing. Then, a parametric paired two sample t-test is completed in order to determine whether there is a difference in pricing between the two months. Nonparametric test are also completed in order to analyze the results and to considered the possible benefit of nonparametric testing in comparison to the parametric testing. Finally, conclusions are made based on the results of the tests.

Based on research, the hypothesis being tested is that the prices of gold are higher in December compared to June due to the winter being around more holidays, which usually means more interest in gold. Therefore the null and alternative hypotheses are as follows, with an alpha level of .05:

Null: $H_0 : \mu_{December} = \mu_{June}$

Alternative: $H_a : \mu_{December} > \mu_{June}$

1.1 Data

The data for the gold pricing was sourced through kaggle.com from Nicolas Ward's data post Nicolas Ward (2021), specifically the monthly set that is updated yearly. Each of the 12 months has prices with the volume of prices and the percent change from the previous year. The set starts in 1979 to 2021 with 515 values overall. June and December have 42 observations each for comparison. Also note, the measurement of the gold is in carats for the pricing.

The following tables show basic summary statistics for each group in the USD:

We see from the normal probability plot of each month in Figure 1 that they both appear to have not quite normal distributions with both looking a bit right tailed. The middle of both lines looks to have a split in the data. The histograms of the data in Figure 2 confirms that the data is bi-modal, which was not apparent from the split in the normal probability plots.

A box-plot comparison of the two months in Figure 3 shows that the medians are not much different, and there are not outliers for either to affect the results. Both are also positively skewed with most of the dispersment of data being between \$500-\$1,500. June has more dispersed data while December has a greater overall spread.

From observing the data to be skewed as well as bi-modal specifically from the histogram, a transformation needs to be preformed on both months in order to follow the normality assumption for testing.

After testing which transformation are best for the data, I found that the log transformation is what makes the data follow the most normal distribution. The following histograms show the relatively normal distributions of both months after the change.

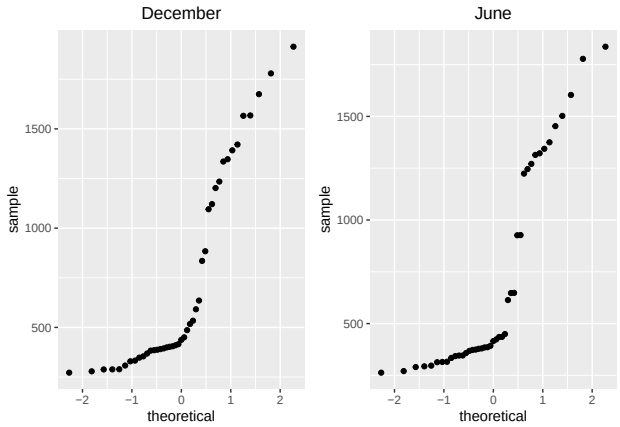

Figure 1: Normal Probability Plots of December and June Gold Prices

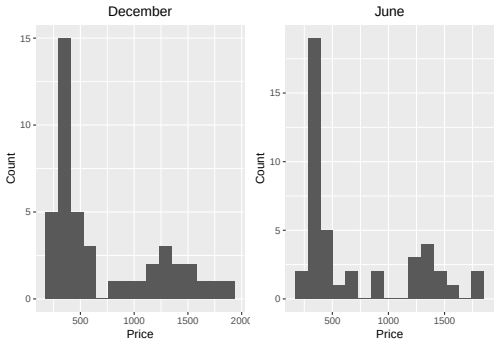

Figure 2: Histograms of December and June Gold Prices

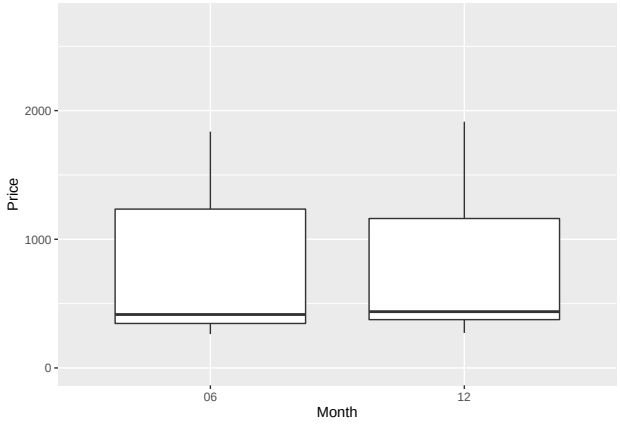

Figure 3: Boxplot comparison of the distribution of December and June Gold Pricing

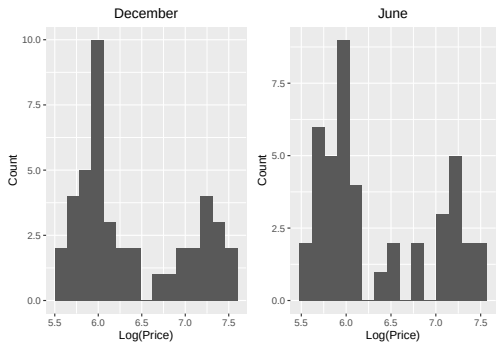

Figure 4: Histograms of December and June distributions showing Log(Price)

2 Testing

The first method used for testing the difference between the two months is the two-sample paired t-test, which assumes normality in the distributions. The paired t-test was used because the test using two sets of data that are across the same years. The paired t-test was computed using this R code. Alboukadel Kassambara (n.d.)

t.test(x,y, alternative="greater", paired= T, var.equal=T)

Since the data does not appear normal even after a transformation, nonparametric methods are used to continue testing the data under different parameters that do not require normality. The second method of testing is a nonparametric statistical method for hypothesis testing, and it is called the Wilcox-Mann Whitney Test. This method allows for skewed data to be tested as long as samples are random. Even after transforming the data, there was still skewness especially since the data was still bi-modal, so the transformed data is used in this test to make the comparison with the t-test be more accurate. This nonparametric test is computed by the following R code. *Mann-Whitney u Test in r* (n.d.)

wilcox.test(x,y, alternative="greater", paired=T)

The next method of nonparametric statistical methods is called the Permutation Test. This test allows for all possible combinations without replacement of the data to be tested and compared, which helps because the only assumption this test relies on the exchangeability of the data rather than normality. For the test, the transformed data is used since the first nonparametric test is used with it too. The Permutation test is computed by the following R code. *Paired Permutationn t-Test* (n.d.)

paired.perm.test(x-y,n.perm=10000,pval=T)

The final method of nonparametric statistical computing is the Randomization test. This method is like the Permutation test but makes combinations with replacement of the data. This test allows for more randomness in the test due to the fact that it is not limited to previous combinations of the data like the Permutation test is. With the transformed data, this test is computed by the following R code written by David C. Howell (2015).

DJ<-dataset%>%select(x,y)

diffObt <- mean(DJx) − $mean(DJ$y)

difference <- DJx − DJy

nreps <- 10000

resampMeanDiff <- numeric(nreps) for (i in 1:nreps) { signs <- sample(c(1,-1),length(difference), replace = T) resamp <- difference * signs resampMeanDiff[i] <- mean(resamp) }

diffObt <- abs(diffObt)

highprob <- length(resampMeanDiff[resampMeanDiff >= diffObt])/nreps

lowprob <- length(DJ$resampMeanDiff[DJ$resampMeanDiff <= (-1)*DJ$diffObt])/nreps

prob2tailed <- lowprob + highprob

3 Results

After computing the t-test, we find that the test statistic of 2.0755903 produces a p-value of 0.0220461, which is smaller than the alpha level of .05. Therefore, the t-test allows for the assumption that there is a statistically significant difference in the gold prices between the months of December and June.

Out of the nonparametric tests, the Mann Whitney test produced a p-value of 0.067007, which is larger than the alpha level of .05. According to this test, there is a not statistically significant difference in the gold prices in December and June over the years of 1979-2021.The Permutation test generated a p-value of 0.0397. This value is less than the level of significance, which supports that assumption that this is a significant difference between the two months' prices. Lastly, the p-value of the Randomization test is also less than the level of significance at 0.0175. This test also supports the assumption that this is a statistical difference between the prices of December and June.

4 Summary of Conclusions

Although a transformation was done to the data, the results of the t-test are not as trustworthy because the transformation did not quite achieve normality, which is seen mostly from it having two modes. The nonparametric test results prove to be a better option due to the use of ranking and permuting methods to evaluate the data, which helps with non-normal distributions. The Mann Whitney test did not find significance between the two months' prices. However, based on the results from the Permutation and Randomization tests, it appears that the ranking method in the Mann Whitney test give one combination of the data that happens to accept the null hypothesis. The other two nonparametric tests give stronger results because of the 10,000 combinations that the hypotheses are tested against. They are also better because they compare difference of means while the Mann Whitney is focused more on the ranks of the data.

After considering all the tests and p-values, the conclusion of this analysis is to reject the null hypothesis that the prices of the two months are the same. This conclusion is based on the results of the Permutation and Randomization tests having p-values less than the alpha level.We can conclude that the price of gold in December are greater than those in June since that is the time of year where people are more in the mindset to buy gold because of the holidays.

5 Future Recommendations

After a discussion with a fellow classmate, an influential variable that may have affected the pricing of gold is different years that had recessions like in 2008. A recommendation on further analysis is to pull out years that were affected by recession and compare them to others. This would be to make sure they are not affecting the results of the tests. Another recommendation off of that would be to take out the years of recession from the analysis to just eliminate that chance of interference.

There is a possibility of confounding in this analysis specifically with the weight of gold sold averagely each month. It is recommended that a future study be made with ways to limit the confounding based on gold weight, which may be averaging the weight and getting prices for those weights.

6 References

Alboukadel Kassambara. n.d. *Paired Samples t-Test in r*. STHDA. http://www.sthda.com/english/wiki/paired-samples-t-test-in-r.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Broman, Karl W. 2022. *Broman: Karl Broman's r Code*. https://github.com/kbroman/broman.

David C. Howell. 2015. *Randomization Test w Samples*. University of Vermont. https://www.uvm.edu/~statdhtx/StatPages/R/RandomizationTestsWithR/RandomMatchedSample/RandomMatchedSampleR.html.

Huntington-Klein, Nick. 2022. *Vtable: Variable Table for Variable Documentation*. https://nickch-k.github.io/vtable/.

Jeff Clark. 2021. *The Best Time of the Year to Buy Gold & Silver in 2021*. TRB Bullion. https://www.trbbullion.com/metal-investment-education/the-best-time-of-the-year-to-buy-gold-silver-in-2021/.

Mann-Whitney u Test in r. n.d. Statistical Methods. https://stat-methods.com/home/mann-whitney-u-r/.

Nicolas Ward. 2021. *Gold Historical Datasets*. kaggle. https://www.kaggle.com/datasets/nward7/gold-historical-datasets?resource=download&select=Gold_Monthly.csv.

Paired Permutationn t-Test. n.d. R Project. https://search.r-project.org/CRAN/refmans/broman/html/paired.perm.test.html.

Wickham, Hadley. 2022a. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2022b. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

7 Appendix

```
#data <- cbind(head(dec, 10), head(jun, 10))
#data
head(c,10)
```

```
## # A tibble: 10 x 3
## # Groups:   Month, Year [10]
##    Month Year  Price
##    <chr> <chr> <dbl>
##  1 06    1979   294.
##  2 12    1979   534.
##  3 06    1980   647.
##  4 12    1980   591.
##  5 06    1981   424.
##  6 12    1981   399.
##  7 06    1982   315.
##  8 12    1982   450.
##  9 06    1983   416.
## 10 12    1983   385.
```

```
knitr::opts_chunk$set(echo = TRUE)

set.seed(1234)

library(tidyverse)
library(knitr)
library(dplyr)
library(vtable)
library(gridExtra)
library(lubridate)
library(broman)
knitr::write_bib(c('knitr', 'stringr', 'dplyr', 'gridExtra','tidyverse','vtable','broman'),file = 'pack

shell("type projectreferences.bib pack.bib > projectreferences_new.bib ")
gold<- read.csv("Gold_Monthly.csv",header=T)

M<-substr(gold$Date,1,3)
Y<-substr(gold$Date,5,7)

Year<-year(as.Date(as.character(Y), format="%y"))
Month<-match(M,month.abb)

nth <- paste0(1:12, c("st", "nd", "rd", rep("th", 9)))
date<-paste(Year,Month,rep(1,length(Year)), sep="-")
gold$Date<-as.Date(date, format="%Y-%m-%d")

gold$Month<- format(gold$Date, "%m")
gold$Year<- format(gold$Date, "%Y")

separate<-gold%>%
  group_by(Month,Year)%>%
```

```
  select(Month, Year, Price)
c<-separate%>%
  filter(Month=='06' | Month=='12')

dec<- separate%>% filter(Month=='12')
jun<- separate%>% filter(Month=='06')

sumtable(dec)
sumtable(jun)


dec_norm <- ggplot() +
  stat_qq(mapping = aes(sample = dec$Price)) +
  ggtitle("December") +
  theme(plot.title = element_text(hjust = 0.5))

jun_norm <- ggplot() +
  stat_qq(mapping = aes(sample = jun$Price)) +
  ggtitle("June") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(dec_norm, jun_norm, ncol = 2)


dec_hist <- ggplot() +
  geom_histogram(mapping = aes(x = dec$Price), bins = 15) +
  ggtitle("December") +
  ylab("Count") +
  xlab("Price") +
  theme(plot.title = element_text(hjust = 0.5))

jun_hist <- ggplot() +
  geom_histogram(mapping = aes(x = jun$Price), bins = 15) +
  ggtitle("June") +
  ylab("Count") +
  xlab("Price") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(dec_hist, jun_hist, ncol = 2)


ggplot(data = c, mapping = aes(x = Month, y = Price)) +
  geom_boxplot() +
  ylab("Price") +
  ylim(0, 2700) +
  theme(plot.title = element_text(hjust = 0.5))

hist(jun$Price)
hist(1/sqrt(jun$Price))
hist(log(jun$Price))
hist(abs(jun$Price - mean(jun$Price)))
```

```
hist(dec$Price)
hist(1/sqrt(dec$Price))
hist(log(dec$Price))
hist(sqrt(dec$Price))



dec_histt <- ggplot() +
  geom_histogram(mapping = aes(x = log(dec$Price)), bins=15) +
  ggtitle("December") +
  ylab("Count") +
  xlab("Log(Price)") +
  theme(plot.title = element_text(hjust = 0.5))

jun_histt <- ggplot() +
  geom_histogram(mapping = aes(x = log(jun$Price)), bins=15) +
  ggtitle("June") +
  ylab("Count") +
  xlab("Log(Price)") +
  theme(plot.title = element_text(hjust = 0.5))
grid.arrange(dec_histt, jun_histt, ncol = 2)

jun$Price<-log(jun$Price)
dec$Price<-log(dec$Price)

montht<-t.test(dec$Price,jun$Price,paired=TRUE, alternative="greater")


separate$Price<- log(separate$Price)
jd<-separate%>%
  filter(Month=='06' | Month=='12')
monthw<-wilcox.test(1-Price~Month,data=jd, paired=T)

perm<-paired.perm.test(dec$Price-jun$Price,n.perm=10000,pval=T)

jun$'June Price'= jun$Price
dec$'December Price'= dec$Price
comb<-full_join(dec,jun, by='Year')
DJ<-comb%>%select('December Price','June Price')

diffObt <- mean(DJ$'December Price') - mean(DJ$'June Price')
difference <- DJ$'December Price' - DJ$'June Price'  #Use that order to keep most diff. positive

nreps <- 10000
#set.seed(1234)
resampMeanDiff <- numeric(nreps)
for (i in 1:nreps) {
  signs <- sample( c(1,-1),length(difference), replace = T)
  resamp <- difference * signs
  resampMeanDiff[i] <- mean(resamp)
}
diffObt <- abs(diffObt)
highprob <- length(resampMeanDiff[resampMeanDiff >= diffObt])/nreps
```

```
lowprob <- length(DJ$resampMeanDiff[DJ$resampMeanDiff <= (-1)*DJ$diffObt])/nreps
prob2tailed <- lowprob + highprob

#data <- cbind(head(dec, 10), head(jun, 10))
#data
head(c,10)
```