Velo.com Case

Ali Ladha

2024-10-11


```
#loading packages
pacman::p_load(tidyverse, gt)
#loading data
velo <- read_csv('/Users/aliladha/Documents/Files/School Work/College/Graduate/IS 6489-001 Stats & Pred Analytics/velo.csv')
```

```
## Rows: 3483 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (4): checkout_system, device, country, gender
## dbl (3): customer_id, purchases, spent
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

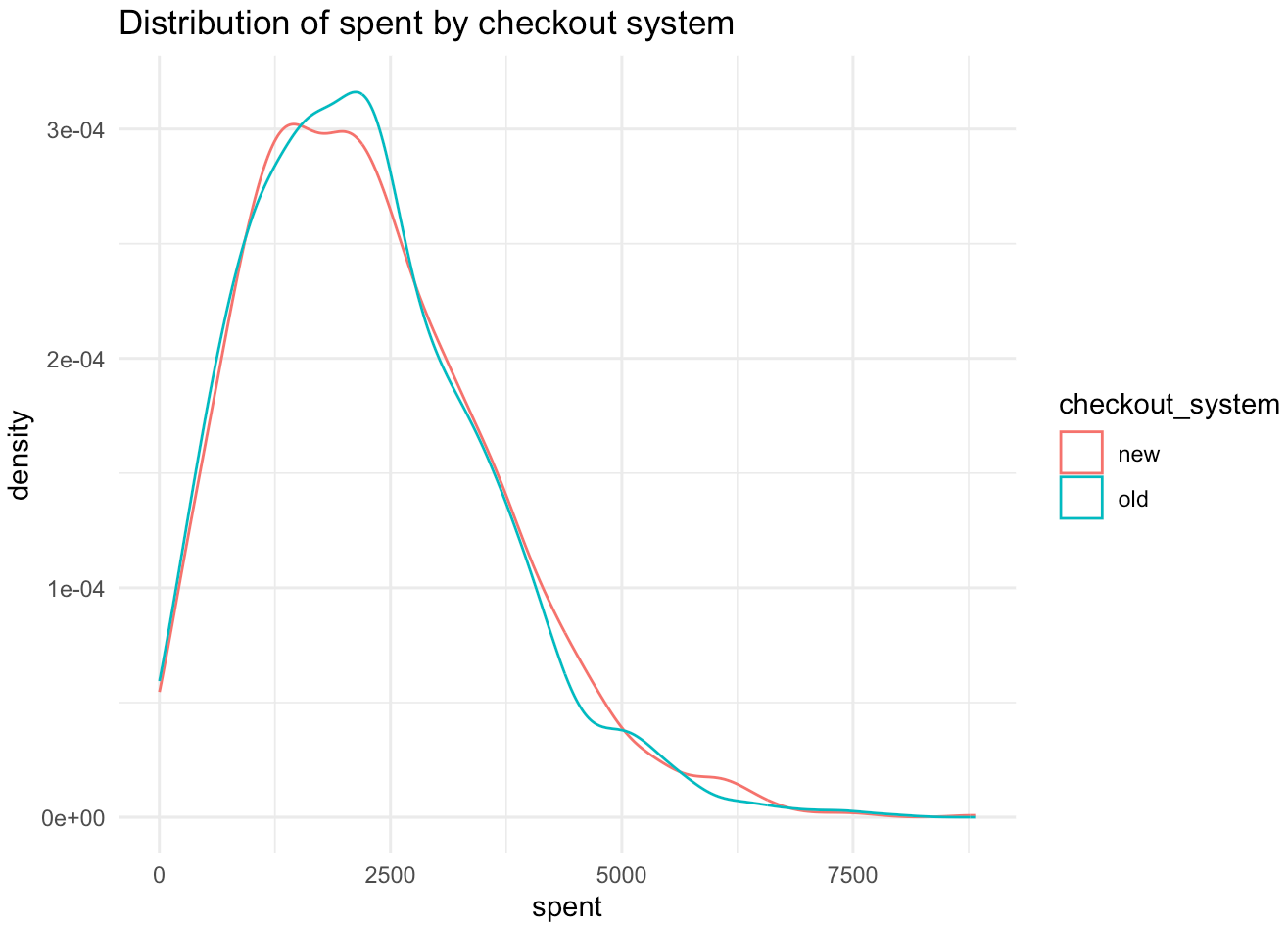
```
velo <- velo %>% filter(checkout_system == 'new' | checkout_system == 'old') %>% mutate(checkout_system = factor(checkout_system))
```

Q1

```
# Plot the distribution of spent by checkout_system. Below you will use a t-test to compare these distributions statistically. However, a t-test assumes normally distributed data. In other words, it assumes that the mean is a good measure of central tendency for comparing the distributions. Do these assumptions seem valid in this case? Why or why not?
#
# Note:
#
# You could compare the two distributions using histograms but a density plot works better. (A boxplot is also an option.)
#
# Make sure to include a plot title.


ggplot(data = velo, aes(x = spent, col = checkout_system)) +
  geom_density() +
```

```
  theme_minimal() +
  labs(title = 'Distribution of spent by checkout system')
```



Based on the density plot, median is a better measure of central tendency for comparing the distributions as the curve is skewed to the right. Regardless, we can still complete the analysis using a T Test as the curve still resembles a normal distribution curve.

Q2

```
# Create a summary table of spent by checkout_system with the following statistics:
#
# n
# mean
# median
# standard deviation
# total (a sum of all spending)
# the lower bound of a 95% confidence interval for the mean
# the upper bound of a 95% confidence interval for the mean
# Your table should have 2 rows and 8 columns.
```

```
#adding SE to to the summary table
velo %>% group_by(checkout_system) %>%
  summarize(n = n(),
                  mean = mean(spent),
```

```
                  median = median(spent),
                  sd = sd(spent),
                  total = sum(spent),
                  se = sd/sqrt(n),
                  lowerCI = (mean - 2*se) %>% round(2),
                  upperCI = (mean + 2*se) %>% round(2))
```

```
## # A tibble: 2 × 9
##   checkout_system     n  mean median    sd    total    se lowerCI upperCI
##   <fct>           <int> <dbl>  <dbl> <dbl>    <dbl> <dbl>   <dbl>   <dbl>
## 1 new              1828 2280.  2100. 1316. 4167638.  30.8   2218.   2341.
## 2 old              1655 2217.  2091. 1277. 3669381.  31.4   2154.   2280.
```

Q3

```
# Is average spending significantly higher in the treatment group? (The treatment group co
nsists in the customers using the new checkout system.) Answer this question using a 2 sam
ple, 2-tailed t-test with alpha set at .05. (Note that these are the default settings for
the t.test() function when vectors are supplied for the x and y arguments.

t.test(x = filter(velo, checkout_system == 'old')$spent,
       y = filter(velo, checkout_system == 'new')$spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(velo, checkout_system == "old")$spent and filter(velo, checkout_system ==
"new")$spent
## t = -1.4272, df = 3464.4, p-value = 0.1536
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -148.93475   23.45215
## sample estimates:
## mean of x mean of y
##  2217.148  2279.890
```

With a T statistic of approximetely 1.43 and a pvalue of 0.15 (which is above 0.05), it is likely that the differences in the mean of spending for the checkout systems occurred due to chance. Furthermore, the confidence interval is within the range of 0 which proves this point too. However, we should look into the analysis further by filtering specifically for the mobile device type as that is where the bug is occurring.

Q4

```
# First, create another summary table of spent by checkout_system and device. Include thes
e same statistics:
#
# n
# mean
```

```
# median
# standard deviation
# the lower bound of a 95% confidence interval for the mean
# the upper bound of a 95% confidence interval for the mean
# The table should have 4 rows and 8 columns.
#
# Based on this information (as well as Sarah's observation, noted in the case descriptio
n, that the glitch in the checkout system seemed more prevalent for mobile users), an addi
tional statistical comparison of new and old among just mobile users seems warranted.
#
# Second, make the comparison using a 2 sample, 2-tailed t-test with alpha set at .05. Rep
ort your results. (Note that a t-test can only compare two groups. Therefore, you will nee
d to subset the data before making the comparison.)
# ```

#adding SE too to the summary table
velo %>% group_by(checkout_system, device) %>%
  summarize(n = n(),
                   mean = mean(spent),
                   median = median(spent),
                   sd = sd(spent),
                   se = sd/sqrt(n),
                   total = sum(spent),
                   lowerCI = (mean - 2*se) %>% round(2),
                   upperCI = (mean + 2*se) %>% round(2))
```

```
## `summarise()` has grouped output by 'checkout_system'. You can override using
## the `.groups` argument.
```

```
## # A tibble: 4 × 10
## # Groups:   checkout_system [2]
##   checkout_system device       n  mean median    sd    se  total lowerCI upperCI
##   <fct>           <chr>    <int> <dbl>  <dbl> <dbl> <dbl>  <dbl>   <dbl>   <dbl>
## 1 new             computer   829 2228.  2058. 1303.  45.2 1.85e6   2137.   2318.
## 2 new             mobile     999 2323.  2145. 1326.  42.0 2.32e6   2239.   2407.
## 3 old             computer   857 2256.  2147. 1274.  43.5 1.93e6   2169.   2344.
## 4 old             mobile     798 2175.  2027. 1279.  45.3 1.74e6   2084.   2265.
```

```
t.test(x = filter(velo, checkout_system == 'old' & device == 'mobile')$spent,
       y = filter(velo, checkout_system == 'new' & device == 'mobile')$spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(velo, checkout_system == "old" & device == "mobile")$spent and filter(vel
o, checkout_system == "new" & device == "mobile")$spent
## t = -2.399, df = 1733.1, p-value = 0.01655
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -269.13848  -27.01302
## sample estimates:
## mean of x mean of y
##  2174.920  2322.996
```

```
# t test for computer
t.test(x = filter(velo, checkout_system == 'old' & device == 'computer')$spent,
       y = filter(velo, checkout_system == 'new' & device == 'computer')$spent)
```

```
##
##  Welch Two Sample t-test
##
## data:  filter(velo, checkout_system == "old" & device == "computer")$spent and filter(v
elo, checkout_system == "new" & device == "computer")$spent
## t = 0.45431, df = 1678.9, p-value = 0.6497
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -94.62766 151.67893
## sample estimates:
## mean of x mean of y
##  2256.469  2227.944
```

By comparing only the mobile as the device with the old and new checkout, the t statistic is approximetely 2.4. The pvalue is 0.01655 which is less than 0.05. Therefore the likelihood of having the observed differences in mean is significantly less than 5% by chance. A null hypothesis for this situation is that there is no difference between the checkout systems for mobile devices. Likewise, an alternative hypothesis is that there is a difference between the two checkout systems for mobile devices. Due to the low p value and the fact that the confidence interval does not include 0, we can reject the null hypothesis for mobile devices.

When comparing only the computer as teh device with the old and new checkout, the t statistic is around 0.45. The pvalue is 0.65 which is greater than 0.05. Therefore there is no difference between the checkout systems for computer devices. Due to the high p value and that the confidence interval range includes 0, we fail to reject the null hypothesis for computer devices.

Q5

```
#What course of action should Sarah recommend to the management at velo.com? Please incorp
orate your analytic results from above in fashioning an answer.
```

The course of action that Sarah should recommend to management is to implement the new checkout system. This is even more easier to accomplish given that the new system has already been developed and the majority of the sunk costs have already been accounted for.

Based on the analysis, the pvalue for mobile devices regarding the old checkout vs the new checkout is 0.01655. Which means that the differences in the average spending is 1.66% due to chance. There is no significant difference for the checkout system if a computer is used, since the pvalue is higher than 0.05.

Regardless, the mean spending for mobile devices in the new checkout system is approximately: $2,323. Whereas the median is: $2145. the mean spending for mobile devices in the old checkout system is approximately: $2175 Whereas the median is: $2027 Both means and medians are higher in the new checkout system for mobile devices. However the sample size for both the new and old checkout systems are slightly different. So if evaluating the mean, we can obtain a weighted mean calculation if further analysis is required.

Likewise other than immediate profit factors, adding the new checkout system will decrease customer frustration since the previous checkout system had a bug which erased checkout carts due to customization. Customer frustration can decrease referrals from customers, whereas satisfied customers can increase referrals and establish customer loyalty leading to continued business.