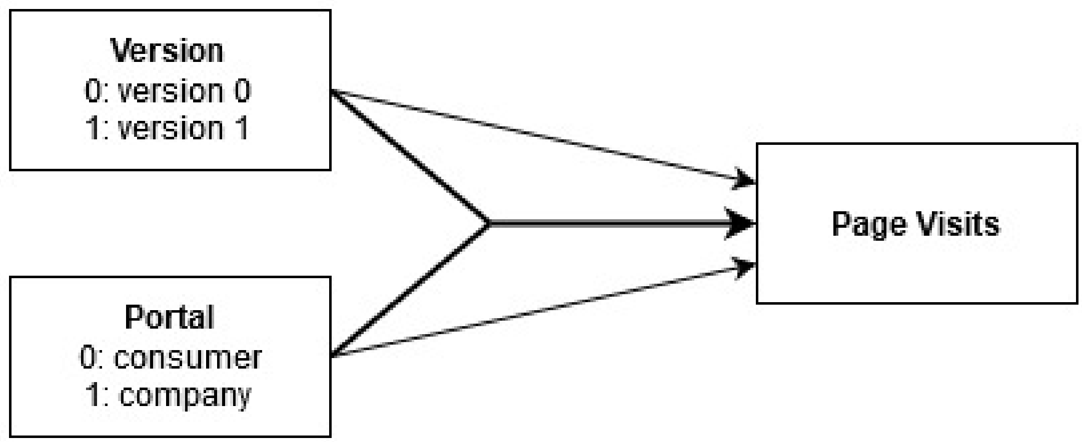
Question 2 - Website visits (between groups - Two factors)

2.1 Conceptual model



We are seeking the effect of Version (0,1), Portal (consumer, customer) and their interaction on the page visits by the user.

2.2 Visual inspection

2.2.1 Reading Data

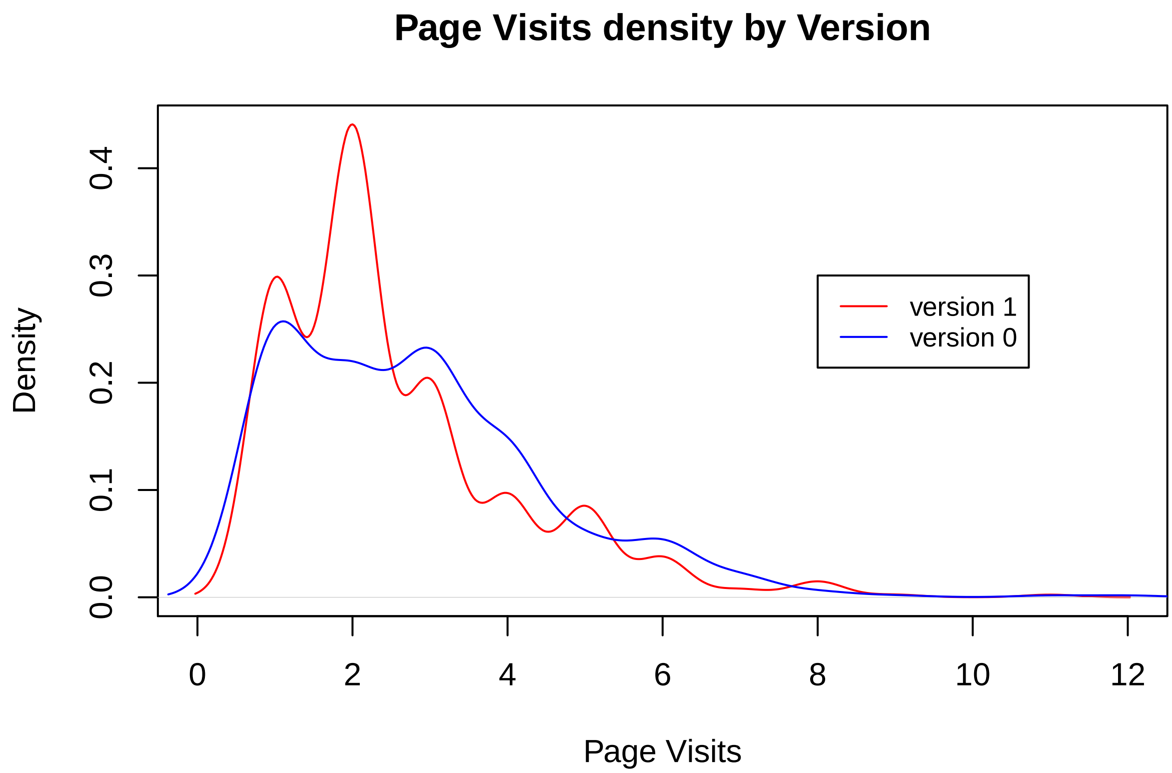
We had to use file named 2 (by the %3 rule) but we used `webvisita.csv` as the files were named as **0, 1 and 'a'** for some reason :/

```
data = read.csv('webvisita.csv')
data$version = as.factor(data$version)
data$portal = as.factor(data$portal)
summary(data)
```

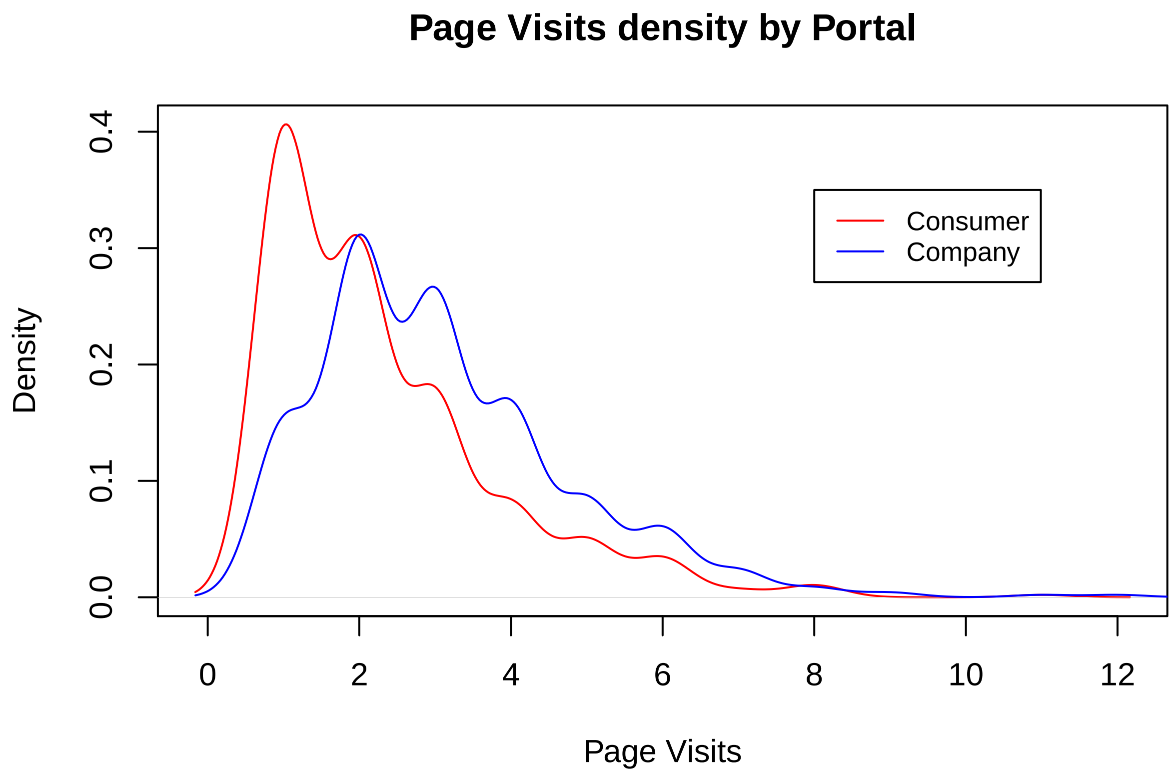
```
##       user        version portal      pages
##  Min.   :  1.0   0:525   0:497   Min.   : 1.00
##  1st Qu.:250.5   1:474   1:502   1st Qu.: 1.00
##  Median :500.0                   Median : 2.00
##  Mean   :500.0                   Mean   : 2.71
##  3rd Qu.:749.5                   3rd Qu.: 3.00
##  Max.   :999.0                   Max.   :14.00
```

2.2.2 Examine the variation in Page Visits for different factors

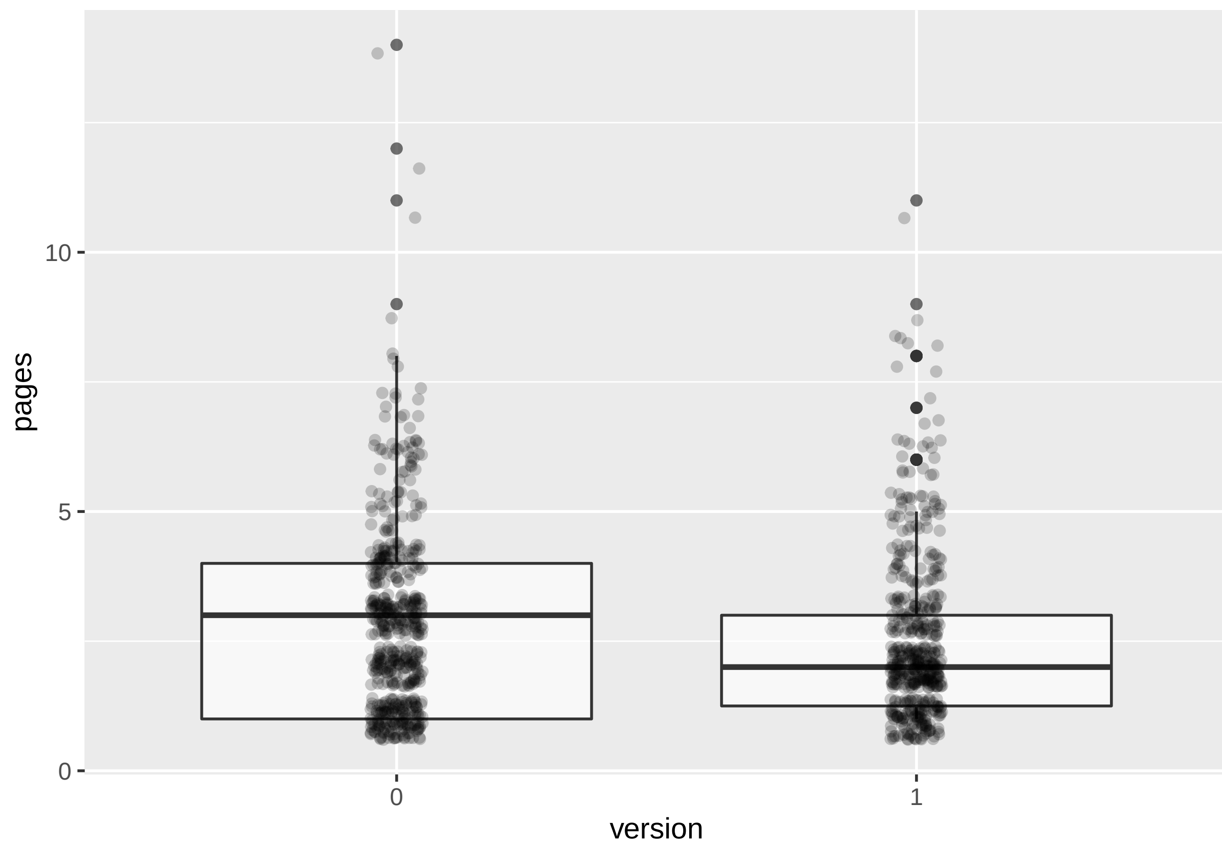
```
plot(  density(subset(data, version == 1)$pages), col='red',
       main="Page Visits density by Version",
       xlab = "Page Visits")
lines( density(subset(data, version == 0)$pages), col='blue')
legend(8, 0.3, legend=c("version 1", "version 0"), col=c("red", "blue"), lty=1, cex=0.8)
```



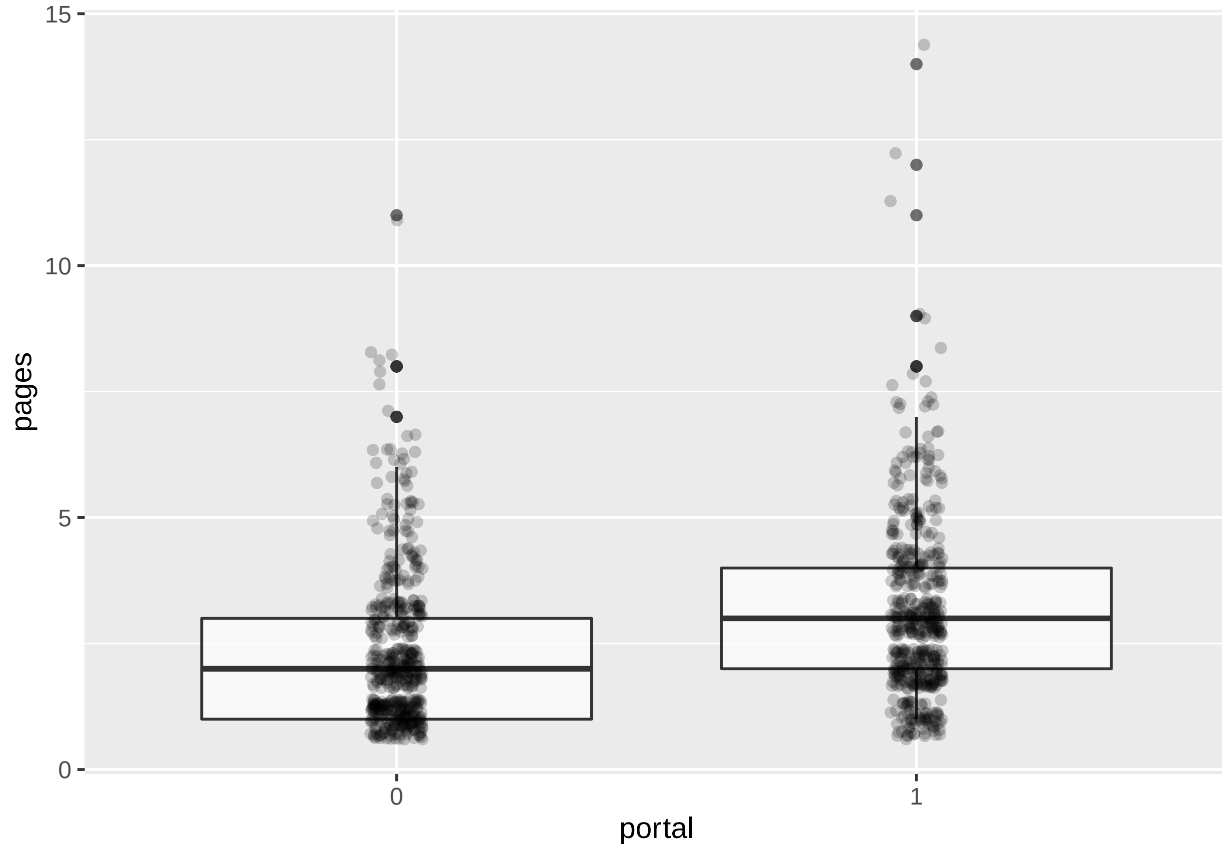

```r
plot(  density(subset(data, portal == 0)$pages), col='red',
       main="Page Visits density by Portal",
       xlab = "Page Visits")
lines( density(subset(data, portal == 1)$pages), col='blue')
legend(8, 0.35, legend=c("Consumer", "Company"), col=c("red", "blue"), lty=1, cex=0.8)
```




```
library(ggplot2)
ggplot(data, aes(x=version, y=pages))  + geom_boxplot(alpha = .7) + geom_jitter(width = .05, alpha = .2)
```




```
library(ggplot2)
ggplot(data, aes(x=portal, y=pages))  + geom_boxplot(alpha = .7) + geom_jitter(width = .05, alpha = .2)
```



Both version and portal show that the change in factor has observably different impact on page visits. We can again see that Consumer portal has a lower mean in the distribution. We can see tha version 1 has more normaly distributed and has a lower mean than version 0.

2.3 Normality check

```r
library(pander)
pander(tapply(data$pages, data$version, shapiro.test))
```

- **0**:

Table 1: Shapiro-Wilk normality test: `X[[i]]`

Test statistic	P value
0.8529	9.927e-22 * * *

- **1**:

Table 2: Shapiro-Wilk normality test: `X[[i]]`

Test statistic	P value
0.8236	1.887e-22 * * *

```r
pander(tapply(data$pages, data$portal, shapiro.test))
```

- **0**:

Table 3: Shapiro-Wilk normality test: `X[[i]]`

Test statistic	P value
0.7953	1.53e-24 * * *

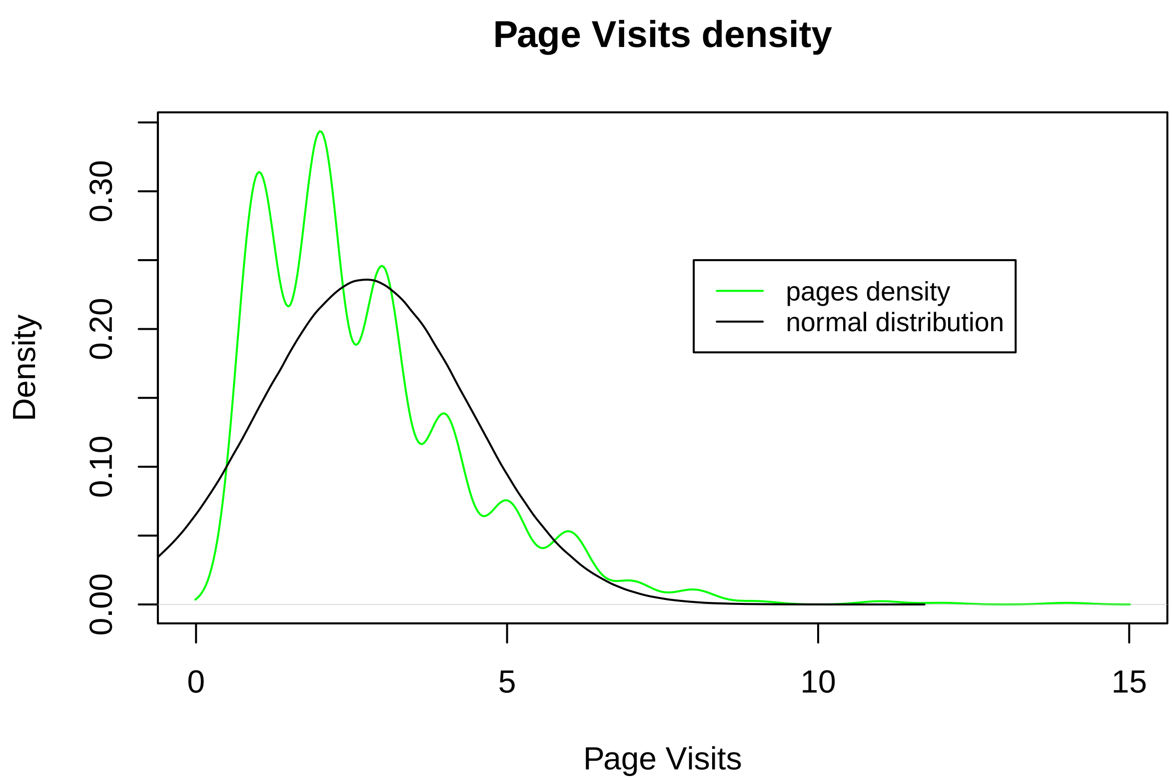
- **1**:

Table 4: Shapiro-Wilk normality test: `X[[i]]`

Test statistic	P value
0.8649	2.059e-20 * * *

Page visits for both version and portal are **NOT** normally distributed therefore we cannot simply make conclusions based on observing the distributions.

```r
plot(density(data$pages), col='green', main="Page Visits density", xlab = "Page Visits")
lines(density(rnorm(10000*length(data$pages), mean = mean(data$pages), sd = sd(data$pages))))
legend(8, 0.25, legend=c("pages density", "normal distribution"), col=c("green", "black"), lty=1, cex=0
```



```r
shapiro.test(data$pages)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$pages
## W = 0.8436, p-value < 2.2e-16
```

Judging by the graph and the Shapiro Wilk test we can see that Page Visits are also **NOT** normally distributed therefore we need to perform some extra analysis.

2.4 Model analysis

Lets build some linear models and see the significanse of the variables. We should first see the Homogeniety of variance across all the conditions first.

```r
# Model analysis
library(car) #Package includes Levene's test
```

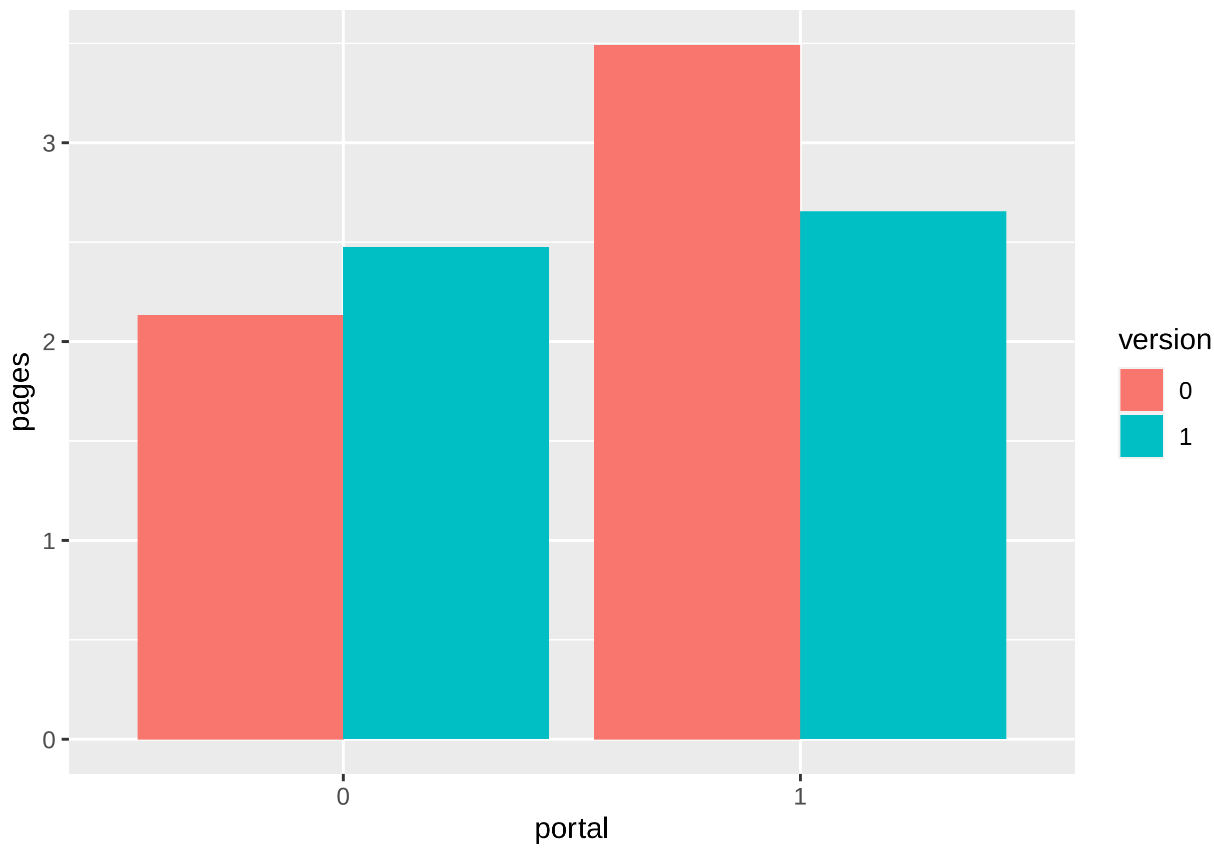
```
## Loading required package: carData
```

```r
leveneTest(data$pages, interaction(data$version, data$portal))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  1.0835 0.3551
##       995
```

We see that the Levene test shows that there is no significant difference between the 4 conditions.

```r
library(ggplot2)
ggplot(data, aes(portal , pages, fill = version)) + stat_summary(fun = mean, geom = "bar", position="do
```




```
s = subset(data, portal == 0 & version == 0)$pages
c(mean(s), sd(s))
```

```
## [1] 2.135458 1.517095
```

```
s = subset(data, portal == 0 & version == 1)$pages
c(mean(s), sd(s))
```

```
## [1] 2.47561 1.54306
```

```
s = subset(data, portal == 1 & version == 0)$pages
c(mean(s), sd(s))
```

```
## [1] 3.492701 1.767104
```

```
s = subset(data, portal == 1 & version == 1)$pages
c(mean(s), sd(s))
```

```
## [1] 2.653509 1.600698
```

The figure shows mean page visits for each of 4 conditions. The most difference is between (portal 1 version 0) and (portal 0 version 0). The mean and standard deviations have been printed after that.

```
bar = ggplot2::aes(data$pages, data$version, fill=data$portal)


model0 = lm(pages ~ 1, data=data)
model1 = lm(pages ~ version, data=data)
model2 = lm(pages ~ portal, data=data)
model12 = lm(pages ~ version + portal, data=data)
```

```
model123 = lm(pages ~ version + portal + version:portal, data=data)

pander(anova(model0, model1), caption="Version as main effect on Page Visits")
```

Table 5: Version as main effect on Page Visits

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
998	2858	NA	NA	NA	NA
997	2838	1	19.9	6.99	0.008324

```
pander(anova(model0, model2), caption="Portal as main effect on Page Visits")
```

Table 6: Portal as main effect on Page Visits

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
998	2858	NA	NA	NA	NA
997	2695	1	162.9	60.28	2.033e-14

```
pander(anova(model123), caption="Effect of Version + Portal + interaction on Page Visits")
```

Table 7: Effect of Version + Portal + interaction on Page Visits

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
version	1	19.9	19.9	7.636	0.005828
portal	1	158.6	158.6	60.86	1.546e-14
version:portal	1	86.47	86.47	33.18	1.118e-08
Residuals	995	2593	2.606	NA	NA

We see that version does significantly affect page visits ($p = 0.008324$), the same goes for portal ($p = 2.033e\text{-}14$). We see that the Interracion effect is also significant ($p=1.118e\text{-}08$) for the Page Visits.

2.5 Simple effect analysis

As the interaction is significant, we carried out a Simple Effect Analysis.

```
# Simple Effect analysis
data$interaction = interaction(data$version, data$portal) # merge 2 factors
levels(data$interaction) # see levels of interaction
```

```
## [1] "0.0" "1.0" "0.1" "1.1"
```

```
# create contrasts to multiply
contrastSimple = c(1,-1,0,0)
contrastComplex = c(0,0,1,-1)
SimpleEff = cbind(contrastSimple, contrastComplex)
contrasts(data$interaction) = SimpleEff
simpleEffModel = lm(pages ~ interaction, data=data)
pander(summary.lm(simpleEffModel))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.689	0.05118	52.54	2.964e-289
interactioncontrastSimple	-0.1701	0.07241	-2.349	0.01904
interactioncontrastComplex	0.4196	0.07235	5.799	8.94e-09
interaction	0.7676	0.1024	7.498	1.428e-13

Table 9: Fitting linear model: pages ~ interaction

Observations	Residual Std. Error	R^2	Adjusted R^2
999	1.614	0.09271	0.08998

We can see here by the Simple effect analysis that the interaction effect is significant. The simple contrast is significant (p. = 0.01904) and so is the complex contrast (p. = 8.94e-09) for the page visits.

2.6 Report section for a scientific publication

portal	version	μ	σ
0	0	2.135458	1.517095
0	1	2.47561	1.54306
1	0	3.492701	1.767104
1	1	2.653509	1.600698

We analyzed the data and found observable impact of version and portal on page visits by their distribution. We found that neither of version, portal or page visits are normally distributed. We observed the means of the 4 means of the groups formed by the different version and portals (4 levels): they are given in the above table.

We conducted a Model analysis and saw that all three: Version ($F(1,997) = 6.99, p. = 0.008324$), Portal ($F(1,997) = 60.28p. = 2.033e - 14$) and Interaction ($F(1,995) = 7.498, p. = 1.428e - 13$), independently have an impact on Page Visits.

As the interaction effect was significant we conducted a Simple Effect analysis. It revealed a significant difference ($t = -2.349, p. = 0.01904$) for simple contrast and a significant difference ($t = 5.799, p. = 8.94e - 09$) for the complex contrast.