```
library(tidyjson)
```

```
##
## Attaching package: 'tidyjson'

## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggplot2)
library(dplyr)
```

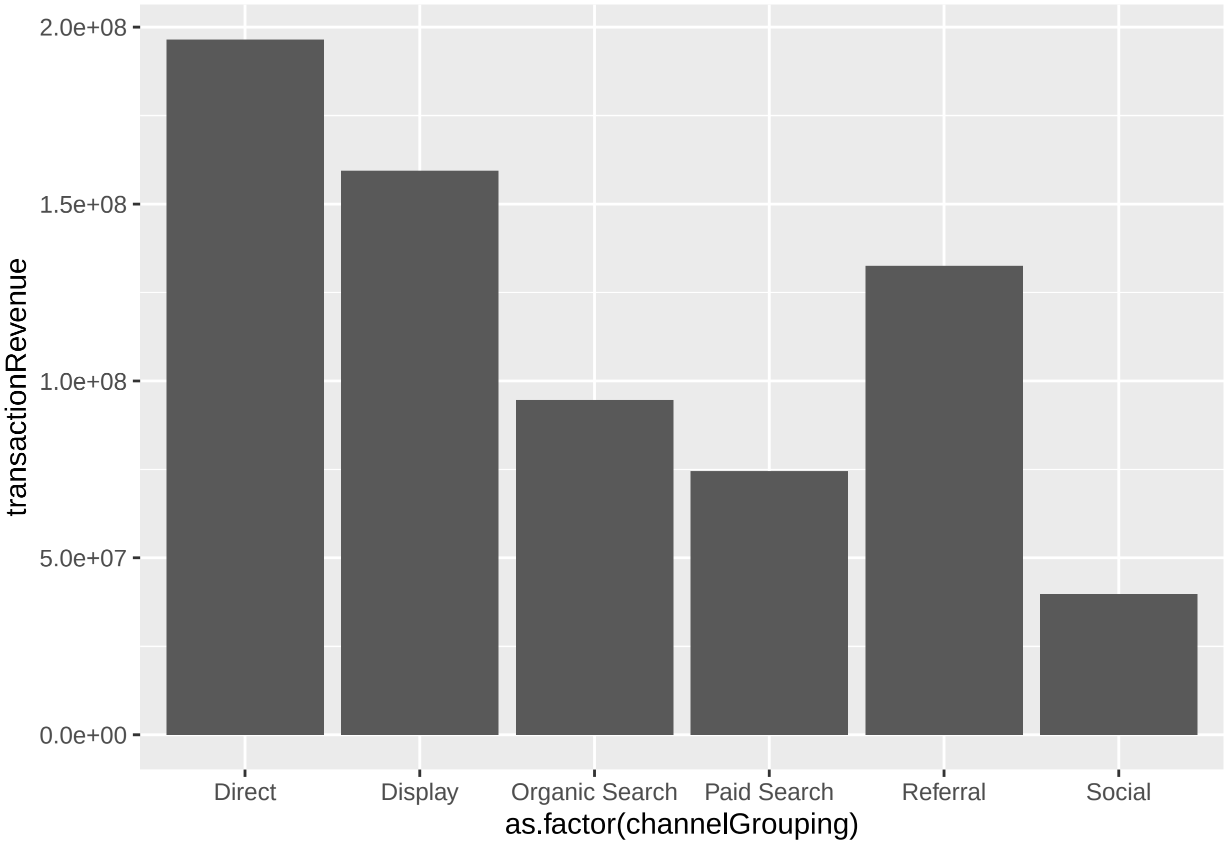
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Load a subset of the data
g_df = read.csv("train.csv", nrows = 100000)
```

```
# Extract the Geographic Data
geo_df = g_df %>% as.tbl_json(json.column="geoNetwork") %>% spread_all %>% select(sessionId, continent,

# Extract the Transaction & Page Visit Data
trans_df = (g_df %>% as.tbl_json(json.column="totals") %>% spread_all ) %>% filter(!is.na(transactionRev

# Combine
total_df = merge(geo_df, trans_df, by="sessionId")

# Cast for convenience
total_df$transactionRevenue = as.numeric(total_df$transactionRevenue)
```

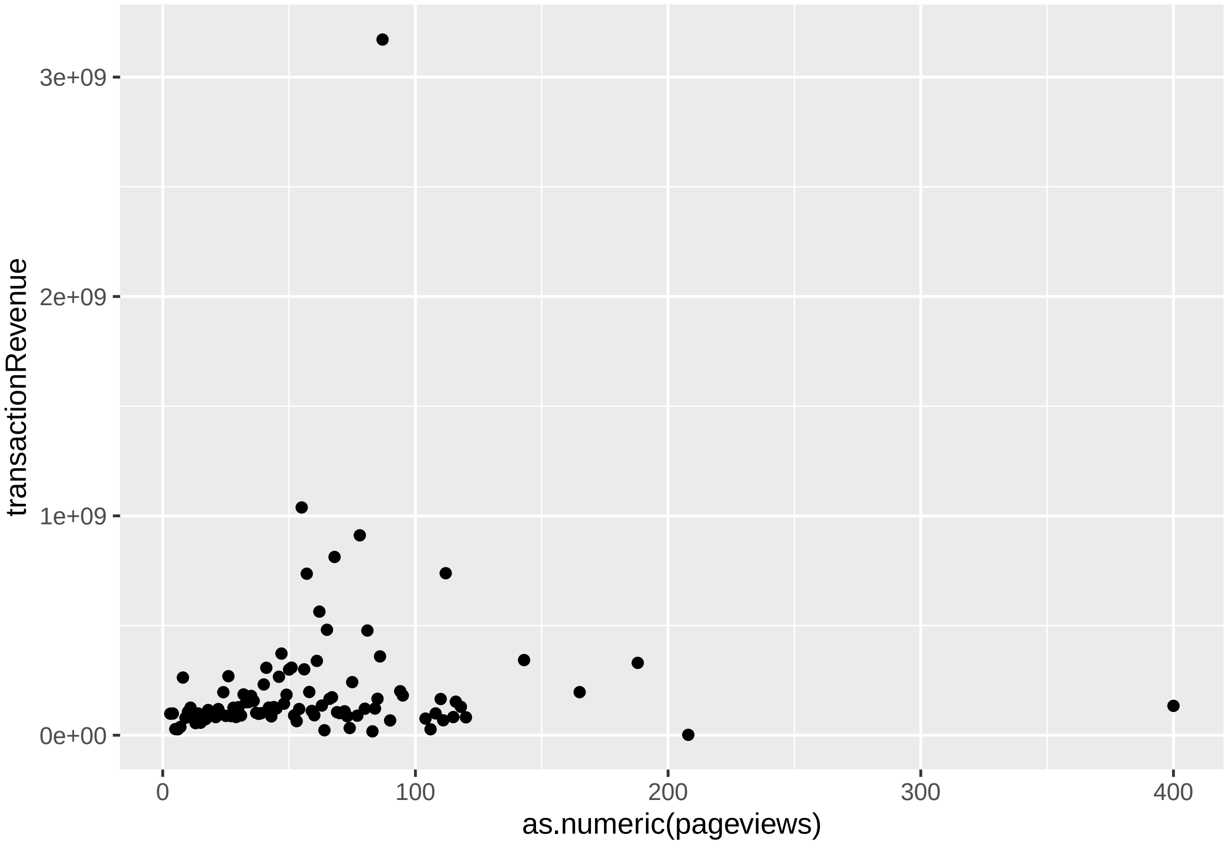
Mean transaction revenue by Channel Grouping

```
ggplot(total_df, aes(as.factor(channelGrouping), transactionRevenue)) + geom_bar(stat = "summary", fun
```



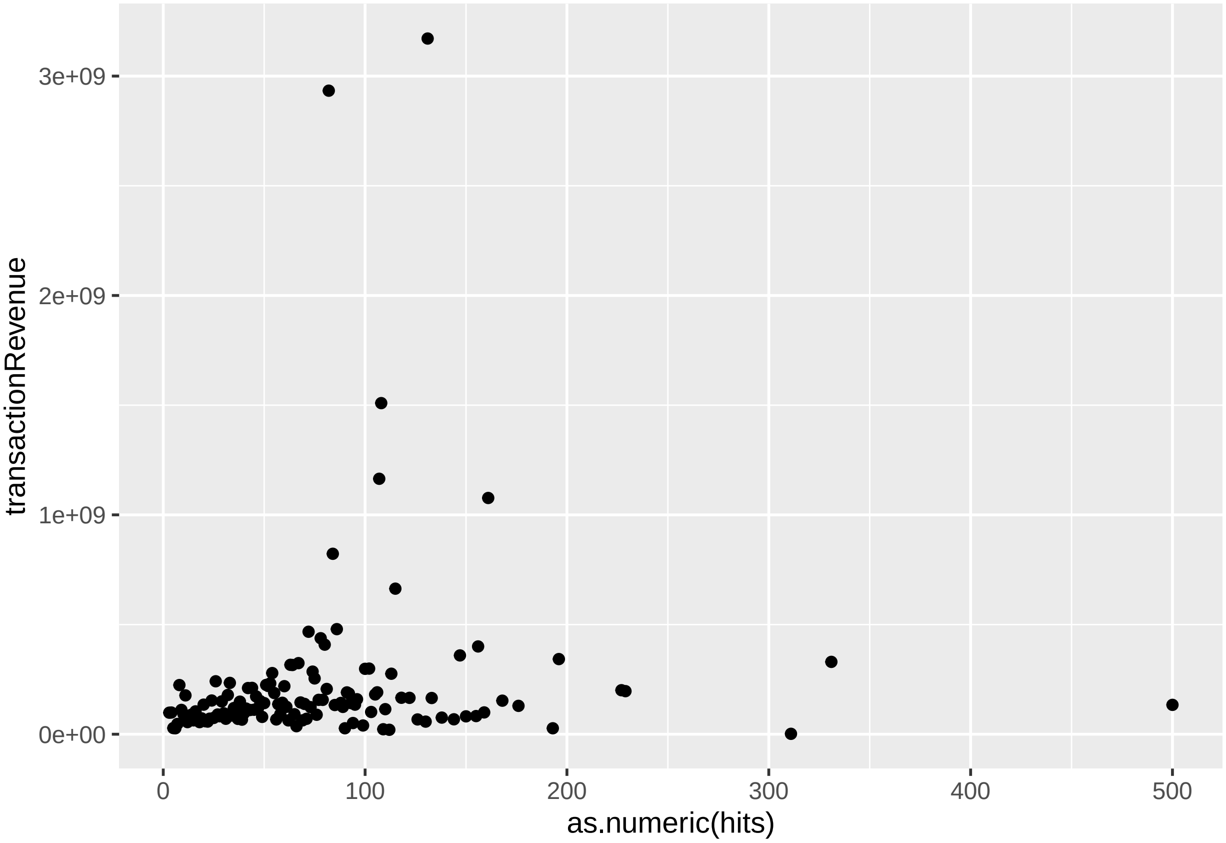
Mean Revenue by Page Views

```
ggplot(total_df, aes(as.numeric(pageviews), transactionRevenue)) + geom_point(stat = "summary", fun = "
```



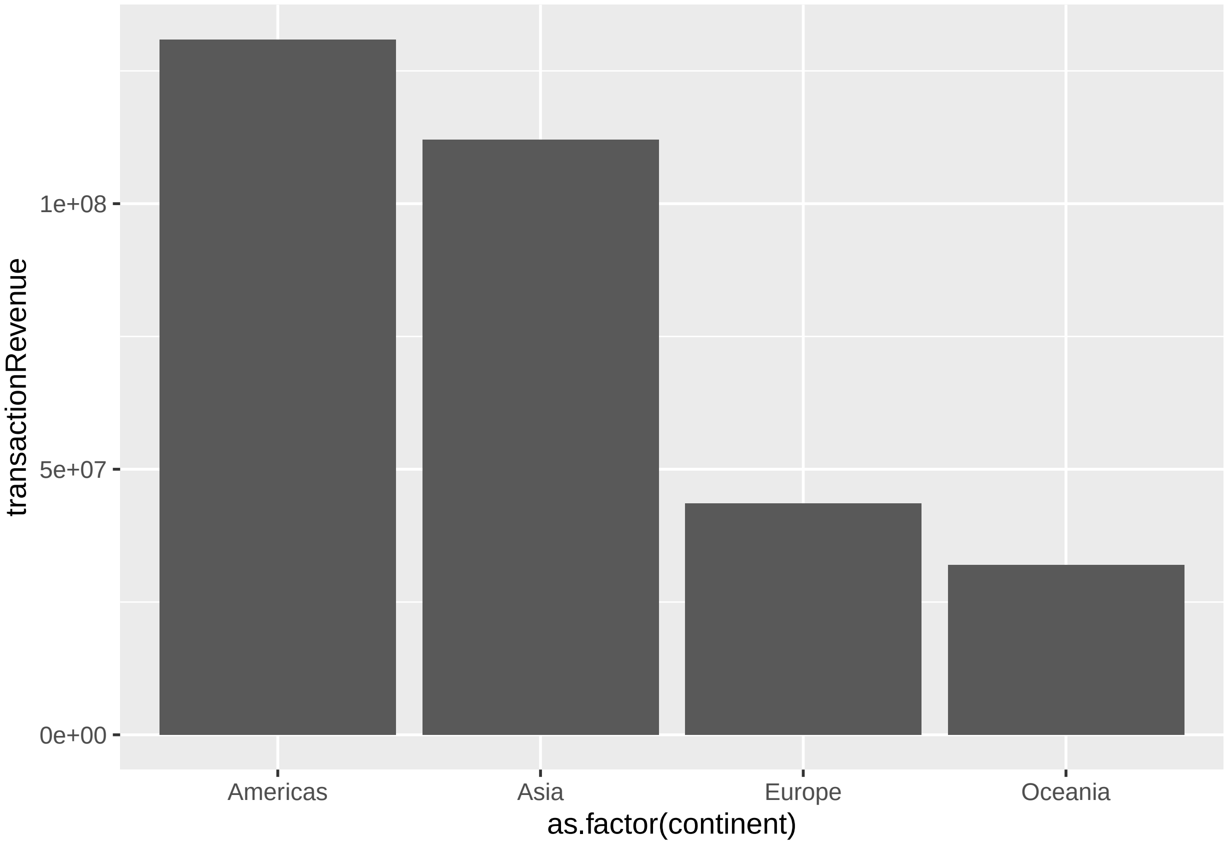
Mean Revenue by Number of Hits

```
ggplot(total_df, aes(as.numeric(hits), transactionRevenue)) + geom_point(stat = "summary", fun = "mean")
```



Mean Revenue by Continent

```
ggplot(total_df, aes(as.factor(continent), transactionRevenue)) + geom_bar(stat = "summary", fun = "mea
```



Interaction Model incorporating the above parameters

```
initial_model = lm(transactionRevenue ~ (as.factor(channelGrouping) + as.numeric(pageviews) + as.numeric
fit_model = step(initial_model, direction="backward", k=2, trace=FALSE) # Fit Using AIC
summary(fit_model)
```

```
##
## Call:
## lm(formula = transactionRevenue ~ as.factor(channelGrouping) +
##     as.numeric(pageviews), data = total_df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -791259991  -95060525  -58711805  -11199076 5307486950
##
## Coefficients:
##                                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                 139361616   21707765   6.420 1.87e-10
## as.factor(channelGrouping)Display           -33709121   69491091  -0.485  0.62769
## as.factor(channelGrouping)Organic Search -105368856   24346537  -4.328 1.61e-05
## as.factor(channelGrouping)Paid Search    -122092897   42490329  -2.873  0.00412
## as.factor(channelGrouping)Referral        -66675528   22357695  -2.982  0.00291
## as.factor(channelGrouping)Social         -137400697   88407471  -1.554  0.12037
## as.numeric(pageviews)                         2131310     383068   5.564 3.16e-08
##
## (Intercept)                              ***
## as.factor(channelGrouping)Display
## as.factor(channelGrouping)Organic Search ***
## as.factor(channelGrouping)Paid Search    **
## as.factor(channelGrouping)Referral       **
## as.factor(channelGrouping)Social
## as.numeric(pageviews)                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 298800000 on 1392 degrees of freedom
## Multiple R-squared:  0.03603,    Adjusted R-squared:  0.03187
## F-statistic: 8.671 on 6 and 1392 DF,  p-value: 2.763e-09
```

Log Model Interaction incorporating the above paramters

```
initial_model = lm(log(transactionRevenue) ~ (as.factor(channelGrouping) + as.numeric(pageviews) + as.nu
fit_model = step(initial_model, direction="backward", k=2, trace=FALSE) # Fit Using AIC
summary(fit_model)
```

```
##
## Call:
## lm(formula = log(transactionRevenue) ~ as.factor(channelGrouping) +
##     as.numeric(pageviews) + as.factor(channelGrouping):as.numeric(pageviews),
##     data = total_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5232 -0.7106 -0.1008  0.6365  4.5321
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                     17.618188
## as.factor(channelGrouping)Display                               -0.625465
## as.factor(channelGrouping)Organic Search                        -0.594048
## as.factor(channelGrouping)Paid Search                           -1.095410
## as.factor(channelGrouping)Referral                               0.046531
## as.factor(channelGrouping)Social                                 0.693260
## as.numeric(pageviews)                                            0.011556
## as.factor(channelGrouping)Display:as.numeric(pageviews)          0.032919
## as.factor(channelGrouping)Organic Search:as.numeric(pageviews)   0.009741
## as.factor(channelGrouping)Paid Search:as.numeric(pageviews)      0.021109
## as.factor(channelGrouping)Referral:as.numeric(pageviews)        -0.001285
## as.factor(channelGrouping)Social:as.numeric(pageviews)          -0.072942
##                                                                 Std. Error
## (Intercept)                                                       0.119276
## as.factor(channelGrouping)Display                                 0.651907
## as.factor(channelGrouping)Organic Search                          0.159900
## as.factor(channelGrouping)Paid Search                             0.350708
## as.factor(channelGrouping)Referral                                0.139204
## as.factor(channelGrouping)Social                                  1.151403
## as.numeric(pageviews)                                             0.003474
## as.factor(channelGrouping)Display:as.numeric(pageviews)           0.023499
## as.factor(channelGrouping)Organic Search:as.numeric(pageviews)    0.004671
## as.factor(channelGrouping)Paid Search:as.numeric(pageviews)       0.011521
## as.factor(channelGrouping)Referral:as.numeric(pageviews)          0.004003
## as.factor(channelGrouping)Social:as.numeric(pageviews)            0.061776
##                                                                 t value Pr(>|t|)
## (Intercept)                                                    147.709  < 2e-16
## as.factor(channelGrouping)Display                               -0.959 0.337505
## as.factor(channelGrouping)Organic Search                        -3.715 0.000211
## as.factor(channelGrouping)Paid Search                           -3.123 0.001825
## as.factor(channelGrouping)Referral                               0.334 0.738232
## as.factor(channelGrouping)Social                                 0.602 0.547206
## as.numeric(pageviews)                                            3.326 0.000904
## as.factor(channelGrouping)Display:as.numeric(pageviews)          1.401 0.161466
## as.factor(channelGrouping)Organic Search:as.numeric(pageviews)   2.085 0.037215
## as.factor(channelGrouping)Paid Search:as.numeric(pageviews)      1.832 0.067121
## as.factor(channelGrouping)Referral:as.numeric(pageviews)        -0.321 0.748227
## as.factor(channelGrouping)Social:as.numeric(pageviews)          -1.181 0.237907
##
## (Intercept)                                                    ***
## as.factor(channelGrouping)Display
## as.factor(channelGrouping)Organic Search                       ***
## as.factor(channelGrouping)Paid Search                          **
## as.factor(channelGrouping)Referral
## as.factor(channelGrouping)Social
## as.numeric(pageviews)                                          ***
## as.factor(channelGrouping)Display:as.numeric(pageviews)
## as.factor(channelGrouping)Organic Search:as.numeric(pageviews) *
## as.factor(channelGrouping)Paid Search:as.numeric(pageviews)    .
## as.factor(channelGrouping)Referral:as.numeric(pageviews)
## as.factor(channelGrouping)Social:as.numeric(pageviews)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.164 on 1387 degrees of freedom
## Multiple R-squared:  0.08659,    Adjusted R-squared:  0.07935
## F-statistic: 11.95 on 11 and 1387 DF,  p-value: < 2.2e-16
```