

## Walmart Recruiting: Trip Type Classification

Wiki page. All important documents related to things like- **domain** of this problem, **important links**, **how to run code**, etc. go here.

Stuff to read before starting (and for future reference)

- Data Description and Download [TODO](#).
- [Forums!!](#)

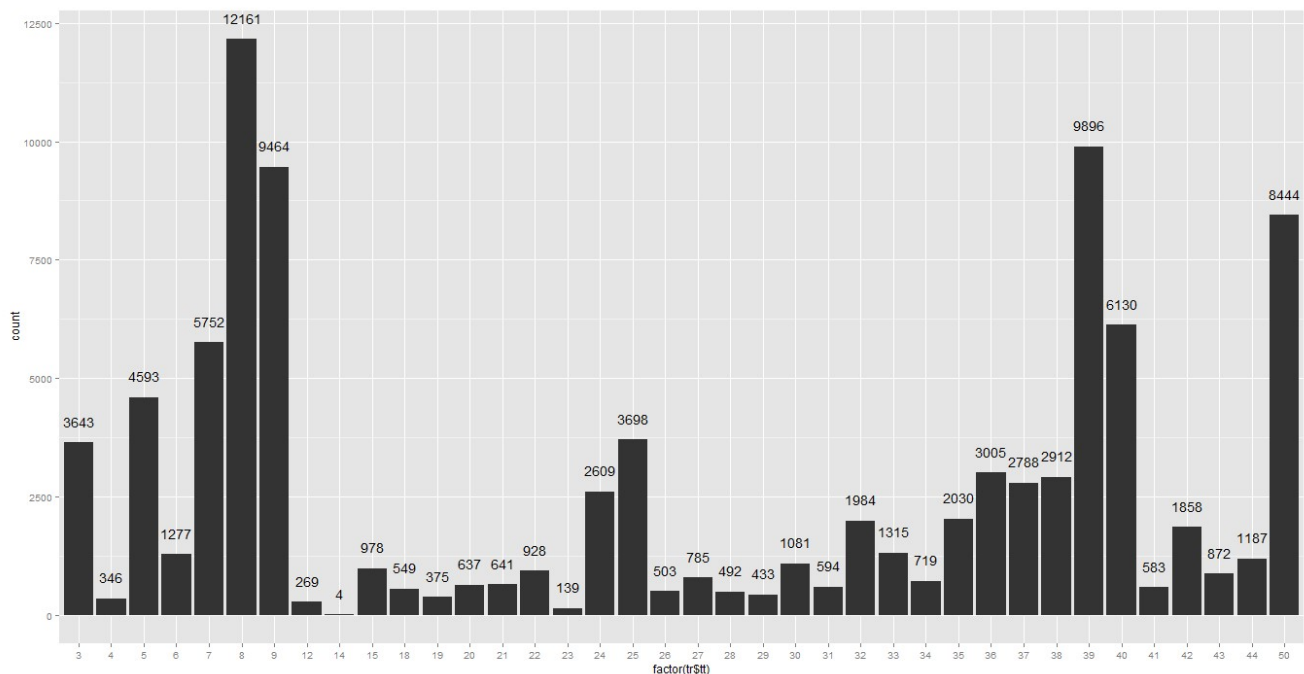
[All About Data-TODO!!](#) <- Put all data related details in this page

- train.csv(~32.9 MB):
  - Columns- 7
  - Rows- 647054
  - Unique VisitNumber Rows- 95674
- test.csv(~31.4 MB):
  - Columns- 7
  - Rows- 653646
  - Unique VisitNumber Rows- 95674

### Details about features

- TripType[int] - a categorical id (type of shopping trip). Ground truth. TripType\_999 is an "other" category.  
Total unique - 38 [3, 4, 5, 6, 7, 8, 9, 12, 14, 15, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 999]

```
tr<- train[,.(tt=unique(TripType),records=N),VisitNumber] ggplot() + aes(factor(tr$tt))+ geom_histogram(binwidth=1) + stat_bin(binwidth=1, geom="text", aes(label=..count..), vjust=-1.5)
```



- VisitNumber[int] - an id corresponding to a single trip by a single customer  
Total unique - 95674(both)
- Weekday[String-ENUM] - the weekday of the trip

Total unique - 7 ["Friday, Monday, Saturday, Sunday, Thursday, Tuesday, Wednesday"]

- Upc[long][NA] - the UPC number of the product purchased(min-834, max-978970666419)  
Total unique - 97715(train), 98148(test)
- ScanCount[int] - the number of the given item that was purchased. A negative value indicates a product return.  
Total unique - 39(train)[-12-to-71], 49(test)[-21-to-55]  
Only in train - 5, Only in test - 15
- DepartmentDescription[String-ENUM][NA=NULL] - a high-level description of the item's department  
Total unique - 69(train) ["HEALTH AND BEAUTY AIDS"],  
68(test) ["1-HR PHOTO, ACCESSORIES, AUTOMOTIVE, BAKERY, BATH AND SHOWER, BEAUTY, BEDDING, BOOKS AND MAGAZINES, BOYS WEAR, BRAS & SHAPEWEAR, CAMERAS AND SUPPLIES, CANDY, TOBACCO, COOKIES, CELEBRATION, COMM BREAD, CONCEPT STORES, COOK AND DINE, DAIRY, DSD GROCERY, ELECTRONICS, FABRICS AND CRAFTS, FINANCIAL SERVICES, FROZEN FOODS, FURNITURE, GIRLS WEAR, 4-6X AND 7-14, GROCERY DRY GOODS, HARDWARE, HOME DECOR, HOME MANAGEMENT, HORTICULTURE AND ACCESS, HOUSEHOLD CHEMICALS/SUPP, HOUSEHOLD PAPER GOODS, IMPULSE MERCHANDISE, INFANT APPAREL, INFANT CONSUMABLE HARDLINES, JEWELRY AND SUNGLASSES, LADIES SOCKS, LADIESWEAR, LARGE HOUSEHOLD GOODS, LAWN AND GARDEN, LIQUOR,WINE,BEER, MEAT - FRESH & FROZEN, MEDIA AND GAMING, MENS WEAR, MENSWEAR, NULL, OFFICE SUPPLIES, OPTICAL - FRAMES, OPTICAL - LENSES, OTHER DEPARTMENTS, PAINT AND ACCESSORIES, PERSONAL CARE, PETS AND SUPPLIES, PHARMACY OTC, PHARMACY RX, PLAYERS AND ELECTRONICS, PLUS AND MATERNITY, PRE PACKED DELI, PRODUCE, SEAFOOD, SEASONAL, SERVICE DELI, SHEER HOSIERY, SHOES, SLEEPWEAR/FOUNDATIONS, SPORTING GOODS, SWIMWEAR/OUTERWEAR, TOYS, WIRELESS"]
- FinelineNumber[int][NA=empty\_space] - a more refined category for each of the products, created by Walmart  
Total unique - 5196(train)[0-to-9998], 5204(test)[0-to-9999]  
Only in train - 150(length(trainFLN[!(trainFLN %in% testFLN)])), Only in test - 158(length(testFLN[!(testFLN %in% trainFLN)]))

## R - important tutorials

- <https://github.com/dmlc/xgboost/blob/master/R-package/vignettes/discoverYourData.Rmd>

## Software Download

- [Python 3.5.0](#)- make sure to download 64bit(file name:python-3.5.0-amd64-webinstall)
  - [Python for Windows Guide](#)
- [R-Windows Binaries](#)
- [R-Studio Download](#)
- [Tortoise Git](#)
- [Git Binaries](#)

Updated 2015-12-13