

# Retail Sales Forecasting

## Contents

<b>Introduction</b>	<b>2</b>
<b>Data Wrangling</b>	<b>3</b>
libraries . . . . .	3
Load data . . . . .	3
Training data . . . . .	3
Stores . . . . .	4
Items . . . . .	5
Transactions . . . . .	5
Oil . . . . .	6
Holidays . . . . .	6
Missing values . . . . .	7
Extract time series features . . . . .	8
Holiday Events, convert character features to factors . . . . .	8
Joining item data with train data ————— . . . . .	10
Joining stores and transactions data for analysis . . . . .	10
transactions_stores, convert character features to factors . . . . .	11
<b>Data Storytelling</b>	<b>11</b>
count of perishable items . . . . .	11
top 20 item families . . . . .	12
top item families and their class . . . . .	12
difference between holiday and non holiday sales . . . . .	13
variation of sales by holiday type . . . . .	13
perishable vs non perishable sales . . . . .	14
promotion vs non promotion avg sales - promotion has significant impact on sales. . . . .	14
transactions by city . . . . .	15
transactions by store cluster . . . . .	16
transactions by store type . . . . .	17
transactions by store . . . . .	18
top 20 items by sales . . . . .	19
sales by day of the month . . . . .	20
weekly sales trend . . . . .	21

# Introduction

The training data includes dates, store and item information, whether that item was being promoted, as well as the unit sales. Additional files include supplementary information that may be useful in building your models.

## File Descriptions and Data Field Information

### `train.csv`

- Training data, which includes the target `unit_sales` by date, `store_nbr`, and `item_nbr` and a unique id to label rows.
- The target `unit_sales` can be integer (e.g., a bag of chips) or float (e.g., 1.5 kg of cheese). Negative values of `unit_sales` represent returns of that particular item.
- The `onpromotion` column tells whether that `item_nbr` was on promotion for a specified date and `store_nbr`.
- Approximately 16% of the `onpromotion` values in this file are NaN.
- NOTE: The training data does not include rows for items that had zero `unit_sales` for a store/date combination. There is no information as to whether or not the item was in stock for the store on the date, and teams will need to decide the best way to handle that situation. Also, there are a small number of items seen in the training data that aren't seen in the test data.

### `stores.csv`

- Store metadata, including city, state, type, and cluster.
- `cluster` is a grouping of similar stores.

### `items.csv`

- Item metadata, including family, class, and perishable.
- NOTE: Items marked as perishable have a score weight of 1.25; otherwise, the weight is 1.0.

### `transactions.csv`

- The count of sales transactions for each date, `store_nbr` combination. Only included for the training data timeframe.

### `oil.csv`

- Daily oil price. Includes values during both the train and test data timeframe. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

### `holidays_events.csv`

- Holidays and Events, with metadata
- NOTE: Pay special attention to the `transferred` column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where `type` is `Transfer`. For example, the holiday `Independencia de Guayaquil` was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are `type` `Bridge` are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the `type` `Work Day` which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the `Bridge`.

- Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).
- Additional Notes
- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

## Data Wrangling

### libraries

```
library('ggplot2')
library('dplyr')
library('readr')
library('data.table')
library('tibble')
library('tidyr')
library('stringr')
library('forcats')
library('lubridate')
library('reshape2')
```

### Load data

*training* data is 4.7 GB in size with 126 million rows. 10% of this data is sampled for exploratory analysis.

```
set.seed(32)
train_data <- sample_frac(as.tibble(fread('../data/raw/train.csv')),0.1)
stores <- as.tibble(fread('../data/raw/stores.csv'))
items <- as.tibble(fread('../data/raw/items.csv'))
transactions <- as.tibble(fread('../data/raw/transactions.csv'))
oil <- as.tibble(fread('../data/raw/oil.csv'))
holidays_events <- as.tibble(fread('../data/raw/holidays_events.csv'))
```

### Training data

```
summary(train_data)
```

##	id	date	store_nbr	item_nbr
##	Min. :	3	Length:12549704	Min. : 96995
##	1st Qu.: 31400001	Class :character	1st Qu.:12.00	1st Qu.: 522721
##	Median : 62754887	Mode :character	Median :28.00	Median : 959500
##	Mean : 62761157		Mean :27.47	Mean : 972698

```
## 3rd Qu.: 94146149          3rd Qu.:43.00   3rd Qu.:1353969
## Max.    :125497031        Max.    :54.00   Max.    :2127114
##   unit_sales      onpromotion
## Min.    : -1768.000      Mode :logical
## 1st Qu.:    2.000      FALSE:9603856
## Median :    4.000      TRUE :782413
## Mean    :    8.557      NA's :2163435
## 3rd Qu.:    9.000
## Max.    :20748.000
```

```
glimpse(train_data)
```

```
## Observations: 12,549,704
## Variables: 6
## $ id          <int> 25073733, 118143635, 115930312, 15194656, 73818906...
## $ date        <chr> "2014-06-08", "2017-06-07", "2017-05-17", "2013-12...
## $ store_nbr   <int> 31, 4, 23, 6, 46, 33, 33, 46, 50, 42, 51, 39, 38, ...
## $ item_nbr    <int> 258376, 1963265, 1457411, 1239795, 1113847, 129635...
## $ unit_sales  <dbl> 1, 3, 1, 4, 5, 2, 10, 1, 3, 5, 1, 12, 15, 1, 18, 2...
## $ onpromotion <lgl> FALSE, FALSE, FALSE, NA, FALSE, NA, FALSE, FALSE, ...
```

- There is a unique *id* to label our observations.
- The store numbers are integers (*store\_nbr*) ranging from 1 to 54. Item numbers (*item\_nbr*) are integers.
- *onpromotion* is a logical feature, describing whether the item in question had been assigned a special promotion pricing at the time in the specific store. This feature contains many NA values.
- *unit\_sales* is our target feature. Negative values mean that this particular item was returned (source).

## Stores

```
summary(stores)
```

```
##   store_nbr      city      state      type
## Min.    : 1.00   Length:54   Length:54   Length:54
## 1st Qu.:14.25   Class :character Class :character Class :character
## Median :27.50   Mode  :character Mode  :character Mode  :character
## Mean    :27.50
## 3rd Qu.:40.75
## Max.    :54.00
##   cluster
## Min.    : 1.000
## 1st Qu.: 4.000
## Median : 8.500
## Mean    : 8.481
## 3rd Qu.:13.000
## Max.    :17.000
```

```
glimpse(stores)
```

```
## Observations: 54
## Variables: 5
## $ store_nbr <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ city      <chr> "Quito", "Quito", "Quito", "Quito", "Santo Domingo",...
## $ state     <chr> "Pichincha", "Pichincha", "Pichincha", "Pichincha", ...
## $ type      <chr> "D", "D", "D", "D", "D", "D", "D", "D", "D", "B", "C", "B...
## $ cluster   <int> 13, 13, 8, 9, 4, 13, 8, 8, 6, 15, 6, 15, 15, 7, 15, ...
```

- Stores are identified by their *city* (e.g. “Quito”) and *state* (e.g. “Pichincha”), according to their *store\_nbr* which connects this information to the *train* data. Along with the *type* of the store, these should be encoded as factors.
- *cluster* describes a “grouping of similar stores” (source).

## Items

```
summary(items)
```

##	item_nbr	family	class	perishable
##	Min. : 96995	Length:4100	Min. :1002	Min. :0.0000
##	1st Qu.: 818111	Class :character	1st Qu.:1068	1st Qu.:0.0000
##	Median :1306198	Mode :character	Median :2004	Median :0.0000
##	Mean :1251436		Mean :2170	Mean :0.2405
##	3rd Qu.:1904918		3rd Qu.:2990	3rd Qu.:0.0000
##	Max. :2134244		Max. :7780	Max. :1.0000

```
glimpse(items)
```

```
## Observations: 4,100
## Variables: 4
## $ item_nbr <int> 96995, 99197, 103501, 103520, 103665, 105574, 10557...
## $ family   <chr> "GROCERY I", "GROCERY I", "CLEANING", "GROCERY I", ...
## $ class    <int> 1093, 1067, 3008, 1028, 2712, 1045, 1045, 1045, 104...
## $ perishable <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ...
```

- The *items* are grouped into a broad *family* (e.g. “BREAD/BAKERY”) and an integer *class* column. Once more, these will be factors.
- *perishable*, an identifier whether the item will go bad over time. It is encoded as an integer but would work better as a logical feature, since the only values appear to be “0 vs 1”: perishable (e.g. milk) vs not perishable (e.g. DVDs).
- *item\_nbr* is the key column relating this data set to *train*

## Transactions

```
summary(transactions)
```

```
##      date      store_nbr    transactions
## Length:83488   Min.    : 1.00   Min.    : 5
## Class :character 1st Qu.:13.00   1st Qu.:1046
## Mode  :character Median :27.00   Median :1393
##              Mean  :26.94   Mean  :1695
##              3rd Qu.:40.00   3rd Qu.:2079
##              Max.   :54.00   Max.   :8359
```

```
glimpse(transactions)
```

```
## Observations: 83,488
## Variables: 3
## $ date      <chr> "2013-01-01", "2013-01-02", "2013-01-02", "2013-0...
## $ store_nbr <int> 25, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
## $ transactions <int> 770, 2111, 2358, 3487, 1922, 1903, 2143, 1874, 32...
```

- This data set gives us an additional total number of transactions per *store\_nbr* for a given *date*. This information is only available for the training data.

## Oil

```
summary(oil)
```

```
##      date      dcoilwtico
## Length:1218   Min.    : 26.19
## Class :character 1st Qu.: 46.41
## Mode  :character Median : 53.19
##              Mean  : 67.71
##              3rd Qu.: 95.66
##              Max.   :110.62
##              NA's   :43
```

```
glimpse(oil)
```

```
## Observations: 1,218
## Variables: 2
## $ date      <chr> "2013-01-01", "2013-01-02", "2013-01-03", "2013-01-...
## $ dcoilwtico <dbl> NA, 93.14, 92.97, 93.12, 93.20, 93.21, 93.08, 93.81...
```

## Holidays

```
summary(holidays_events)
```

```
glimpse(holidays_events)
```

- Holidays and special events also come in the shape of a time series with a *date* column.
- There is a *type* of the holiday, a qualifier whether it's regional (*locale*) and in which region it applies (*locale\_\_name*), as well as the name of the holiday in the feature *description*.
- *transferred* is a logical column indicating whether this specific holiday was moved to a different day that year.

```
sum(is.na(train_data))
```

```
sum(is.na(stores))
```

```
sum(is.na(items))
```

```
sum(is.na(transactions))
```

```
sum(is.na(oil))
```

7

```
sum(is.na(holidays_events))
```

```
## [1] 0
```

```
sum(is.na(stores))
```

```
## [1] 0
```

- *train\_data* contains the majority of NAs in the *onpromotion* feature.
- *oil* contains 43 NAs .

## Extract time series features

```
train_data$date <- ymd(train_data$date)
train_data <- train_data %>%
  mutate(year = year(date), month = month(date), day = day(date),
         weekday = wday(date), week_of_year = week(date))
glimpse(train_data)
```

```
## Observations: 12,549,704
## Variables: 11
## $ id          <int> 25073733, 118143635, 115930312, 15194656, 7381890...
## $ date        <date> 2014-06-08, 2017-06-07, 2017-05-17, 2013-12-09, ...
## $ store_nbr   <int> 31, 4, 23, 6, 46, 33, 33, 46, 50, 42, 51, 39, 38,...
## $ item_nbr    <int> 258376, 1963265, 1457411, 1239795, 1113847, 12963...
## $ unit_sales  <dbl> 1, 3, 1, 4, 5, 2, 10, 1, 3, 5, 1, 12, 15, 1, 18, ...
## $ onpromotion <lgl> FALSE, FALSE, FALSE, NA, FALSE, NA, FALSE, FALSE,...
## $ year        <dbl> 2014, 2017, 2017, 2013, 2016, 2013, 2016, 2016, 2...
## $ month       <dbl> 6, 6, 5, 12, 3, 6, 5, 2, 5, 7, 4, 11, 5, 2, 12, 6...
## $ day         <int> 8, 7, 17, 9, 19, 26, 9, 4, 3, 15, 25, 17, 1, 26, ...
## $ weekday     <dbl> 1, 4, 4, 2, 7, 4, 2, 5, 3, 7, 3, 5, 1, 1, 7, 3, 2...
## $ week_of_year <dbl> 23, 23, 20, 49, 12, 26, 19, 5, 18, 28, 17, 46, 18...
```

## Holiday Events, convert character features to factors

```
glimpse(holidays_events)
```

```
## Observations: 350
## Variables: 6
## $ date        <chr> "2012-03-02", "2012-04-01", "2012-04-12", "2012-04...
## $ type        <chr> "Holiday", "Holiday", "Holiday", "Holiday", "Holid...
## $ locale      <chr> "Local", "Regional", "Local", "Local", "Local", "L...
## $ locale_name <chr> "Manta", "Cotopaxi", "Cuenca", "Libertad", "Riobam...
## $ description <chr> "Fundacion de Manta", "Provincializacion de Cotopa...
## $ transferred <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, F...
```



```
str(holidays_events)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':   350 obs. of  6 variables:
## $ date      : chr  "2012-03-02" "2012-04-01" "2012-04-12" "2012-04-14" ...
## $ type      : chr  "Holiday" "Holiday" "Holiday" "Holiday" ...
## $ locale    : chr  "Local" "Regional" "Local" "Local" ...
## $ locale_name: chr  "Manta" "Cotopaxi" "Cuenca" "Libertad" ...
## $ description: chr  "Fundacion de Manta" "Provincializacion de Cotopaxi" "Fundacion de Cuenca" "Can
## $ transferred: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
## - attr(*, ".internal.selfref")=<externalptr>
```

```
holidays_events$date <- ymd(holidays_events$date)
holidays_events <- holidays_events %>%
  mutate(
    type      = as_factor(type),
    locale    = as_factor(locale),
    locale_name = as_factor(locale_name)
  )
```

```
head(holidays_events)
```

```
## # A tibble: 6 x 6
##   date      type    locale locale_name description      transferred
##   <date>    <fct>   <fct>   <fct>         <chr>          <lgl>
## 1 2012-03-02 Holiday Local    Manta         Fundacion de Manta FALSE
## 2 2012-04-01 Holiday Region~ Cotopaxi     Provincializacion de ~ FALSE
## 3 2012-04-12 Holiday Local    Cuenca        Fundacion de Cuenca FALSE
## 4 2012-04-14 Holiday Local    Libertad      Cantonizacion de Libe~ FALSE
## 5 2012-04-21 Holiday Local    Riobamba      Cantonizacion de Riob~ FALSE
## 6 2012-05-12 Holiday Local    Puyo          Cantonizacion del Puyo FALSE
```

```
summary(holidays_events)
```

```
##      date      type      locale      locale_name
## Min.   :2012-03-02  Holiday :221   Local    :152   Ecuador  :174
## 1st Qu.:2013-12-23  Transfer : 12   Regional:  24   Quito    : 13
## Median :2015-06-08  Additional: 51  National:174   Riobamba : 12
## Mean   :2015-04-24  Bridge   :  5                      Guaranda : 12
## 3rd Qu.:2016-07-03  Work Day :  5                      Latacunga: 12
## Max.   :2017-12-26  Event    : 56                      Ambato   : 12
##                                     (Other)  :115
##
## description      transferred
## Length:350      Mode :logical
## Class :character FALSE:338
## Mode  :character TRUE :12
##
##
##
##
```

```
holidays_events %>% group_by(type) %>% count() %>% arrange(desc(n))
```

```
## # A tibble: 6 x 2
## # Groups:   type [6]
##   type      n
##   <fct>    <int>
## 1 Holiday    221
## 2 Event      56
## 3 Additional  51
## 4 Transfer   12
## 5 Bridge      5
## 6 Work Day    5
```

## Joining item data with train data

```
train_data_items_holidays <- train_data %>%
  left_join(items) %>%
  left_join(holidays_events, by = "date")

head(train_data_items_holidays)
```

```
## # A tibble: 6 x 19
##   id date      store_nbr item_nbr unit_sales onpromotion year month
##   <int> <date>      <int>    <int>      <dbl> <lgl>      <dbl> <dbl>
## 1 2.51e7 2014-06-08      31    258376      1 FALSE      2014     6
## 2 1.18e8 2017-06-07       4    1963265     3 FALSE      2017     6
## 3 1.16e8 2017-05-17      23    1457411     1 FALSE      2017     5
## 4 1.52e7 2013-12-09       6    1239795     4 NA        2013    12
## 5 7.38e7 2016-03-19      46    1113847     5 FALSE      2016     3
## 6 7.49e6 2013-06-26      33     129635     2 NA        2013     6
## # ... with 11 more variables: day <int>, weekday <dbl>,
## #   week_of_year <dbl>, family <chr>, class <int>, perishable <int>,
## #   type <fct>, locale <fct>, locale_name <fct>, description <chr>,
## #   transferred <lgl>
```

## Joining stores and transactions data for analysis

```
transactions_stores <- transactions %>%
  left_join(stores)

head(transactions)
```

```
## # A tibble: 6 x 3
##   date      store_nbr transactions
##   <chr>      <int>      <int>
## 1 2013-01-01      25         770
## 2 2013-01-02       1        2111
## 3 2013-01-02       2        2358
## 4 2013-01-02       3        3487
## 5 2013-01-02       4        1922
## 6 2013-01-02       5        1903
```

```
head(transactions_stores)
```

```
## # A tibble: 6 x 7
##   date      store_nbr transactions city      state      type cluster
##   <chr>      <int>      <int> <chr>    <chr>      <chr>    <int>
## 1 2013-01~      25          770 Salinas  Santa Elena D           1
## 2 2013-01~       1          2111 Quito   Pichincha D          13
## 3 2013-01~       2          2358 Quito   Pichincha D          13
## 4 2013-01~       3          3487 Quito   Pichincha D           8
## 5 2013-01~       4          1922 Quito   Pichincha D           9
## 6 2013-01~       5          1903 Santo Do~ Santo Domingo de~ D           4
```

transactions\_stores, convert character features to factors

```
transactions_stores$date <- ymd(transactions$date)
transactions_stores <- transactions_stores %>%
  mutate(
    city      = as_factor(city),
    state     = as_factor(state),
    type      = as_factor(type),
    cluster   = as_factor(cluster)
  )
head(transactions_stores)
```

```
## # A tibble: 6 x 7
##   date      store_nbr transactions city      state      type cluster
##   <date>      <int>      <int> <fct>    <fct>      <fct>    <fct>
## 1 2013-01-01      25          770 Salinas  Santa Elena D           1
## 2 2013-01-02       1          2111 Quito   Pichincha D          13
## 3 2013-01-02       2          2358 Quito   Pichincha D          13
## 4 2013-01-02       3          3487 Quito   Pichincha D           8
## 5 2013-01-02       4          1922 Quito   Pichincha D           9
## 6 2013-01-02       5          1903 Santo Do~ Santo Domingo ~ D           4
```

## Data Storytelling

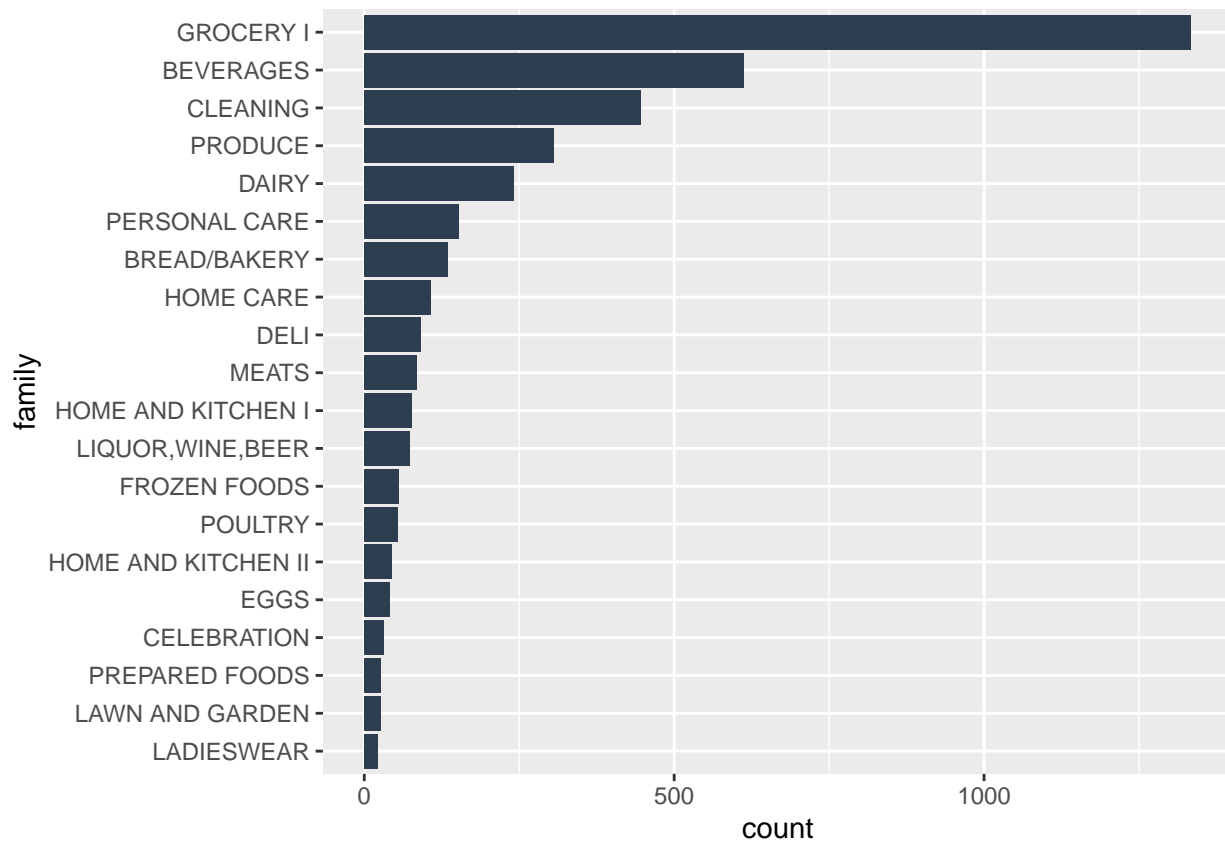
count of perishable items

```
items %>%
  group_by(perishable) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   perishable [2]
##   perishable      n
##   <int> <int>
## 1       0 3114
## 2       1  986
```

## top 20 item families

```
items %>% group_by(family) %>%  
  summarise(count = n()) %>%  
  top_n(20) %>%  
  mutate(family = family %>% as_factor() %>% fct_reorder(count)) %>%  
  ggplot(aes(family, count)) +  
  geom_col(fill = "#2c3e50") +  
  coord_flip()
```



## top item families and their class

```
items %>% group_by(class, family) %>%  
  summarise(count = n()) %>%  
  arrange(desc(count))
```

```
## # A tibble: 337 x 3  
## # Groups:   class [337]  
##   class family    count  
##   <int> <chr>      <int>  
## 1  1016 GROCERY I    133  
## 2  1040 GROCERY I    110
```

```
## 3 1124 BEVERAGES 100
## 4 1034 GROCERY I 98
## 5 1122 BEVERAGES 81
## 6 1072 GROCERY I 70
## 7 1032 GROCERY I 65
## 8 1148 BEVERAGES 62
## 9 1120 BEVERAGES 61
## 10 3020 CLEANING 60
## # ... with 327 more rows
```

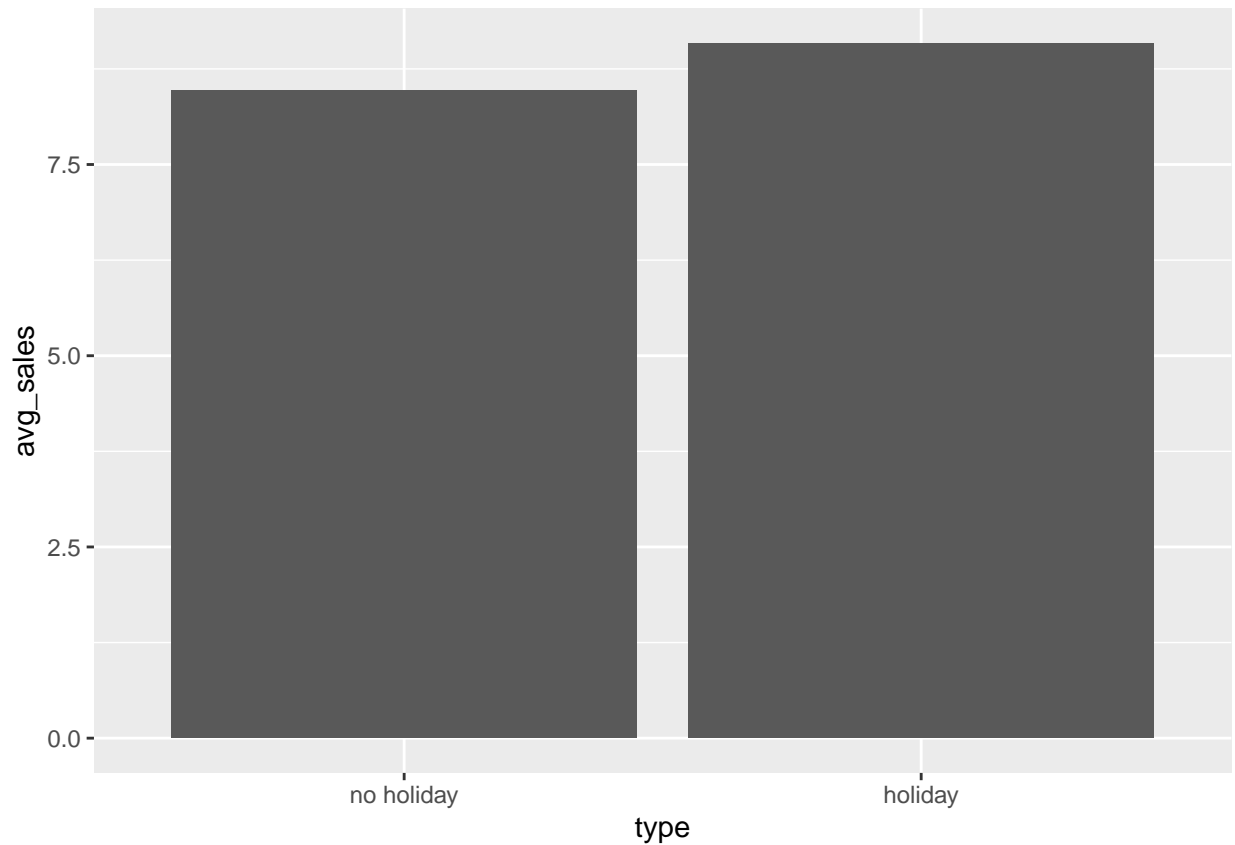
## difference between holiday and non holiday sales

```
train_data_items_holidays <- train_data_items_holidays %>%
  mutate(type = as_factor(ifelse(is.na(locale), "no holiday", "holiday")))
head(train_data_items_holidays)
```

```
## # A tibble: 6 x 19
##       id date      store_nbr item_nbr unit_sales onpromotion year month
##   <int> <date>      <int>    <int>    <dbl> <lgl>      <dbl> <dbl>
## 1 2.51e7 2014-06-08      31    258376      1 FALSE      2014     6
## 2 1.18e8 2017-06-07       4    1963265      3 FALSE      2017     6
## 3 1.16e8 2017-05-17      23    1457411      1 FALSE      2017     5
## 4 1.52e7 2013-12-09       6    1239795      4 NA        2013    12
## 5 7.38e7 2016-03-19      46    1113847      5 FALSE      2016     3
## 6 7.49e6 2013-06-26      33     129635      2 NA        2013     6
## # ... with 11 more variables: day <int>, weekday <dbl>,
## #   week_of_year <dbl>, family <chr>, class <int>, perishable <int>,
## #   type <fct>, locale <fct>, locale_name <fct>, description <chr>,
## #   transferred <lgl>
```

## variation of sales by holiday type

```
train_data_items_holidays %>%
  group_by(type) %>%
  summarise(avg_sales = mean(unit_sales)) %>%
  ggplot() +
  geom_col(aes(x = type, y = avg_sales))
```



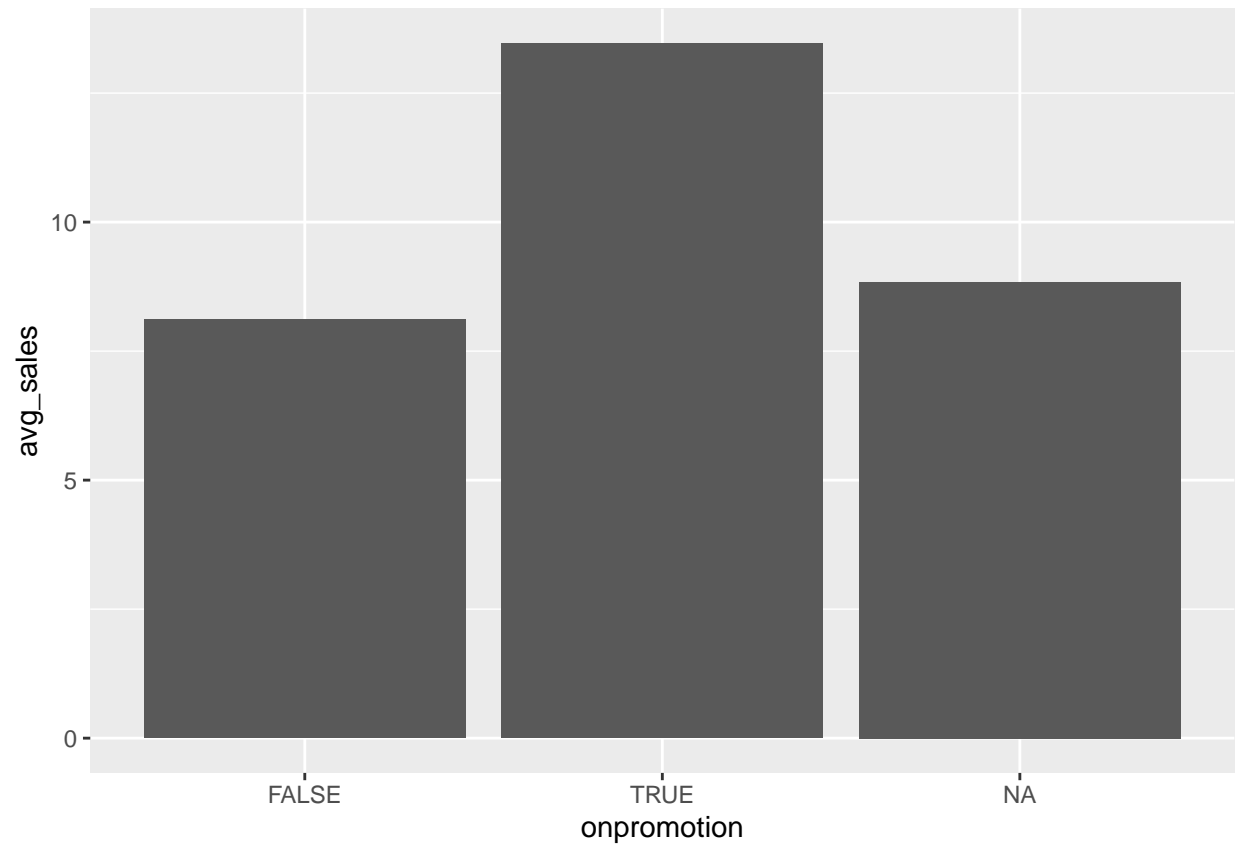
### perishable vs non perishable sales

```
train_data_items_holidays %>%
  group_by(perishable) %>%
  summarise(total_sales = sum(unit_sales))
```

```
## # A tibble: 2 x 2
##   perishable total_sales
##   <int>      <dbl>
## 1     0  74742416.
## 2     1  34963302.
```

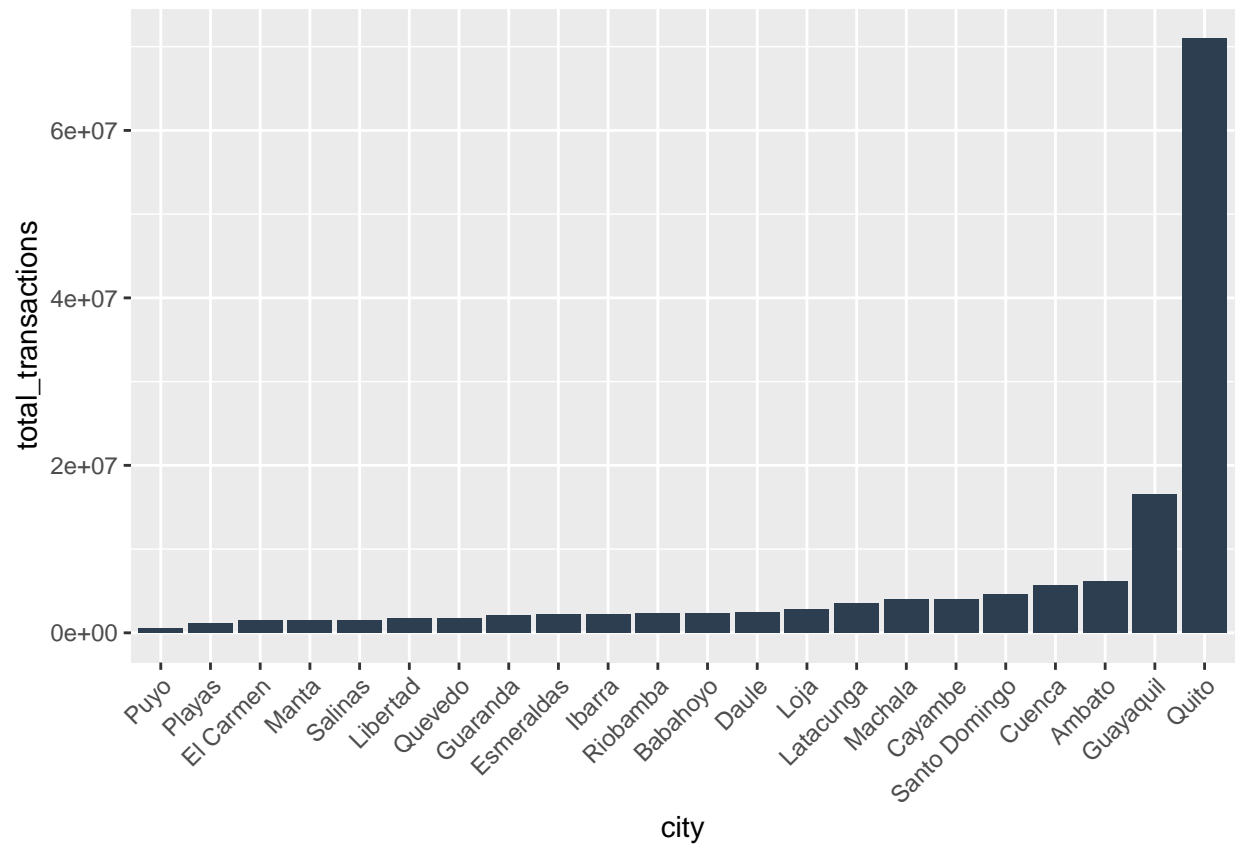
promotion vs non promotion avg sales - promotion has significant impact on sales.

```
train_data_items_holidays %>%
  group_by(onpromotion) %>%
  summarise(avg_sales = mean(unit_sales)) %>%
  ggplot() +
  geom_col(aes(x = onpromotion, y = avg_sales))
```



### transactions by city

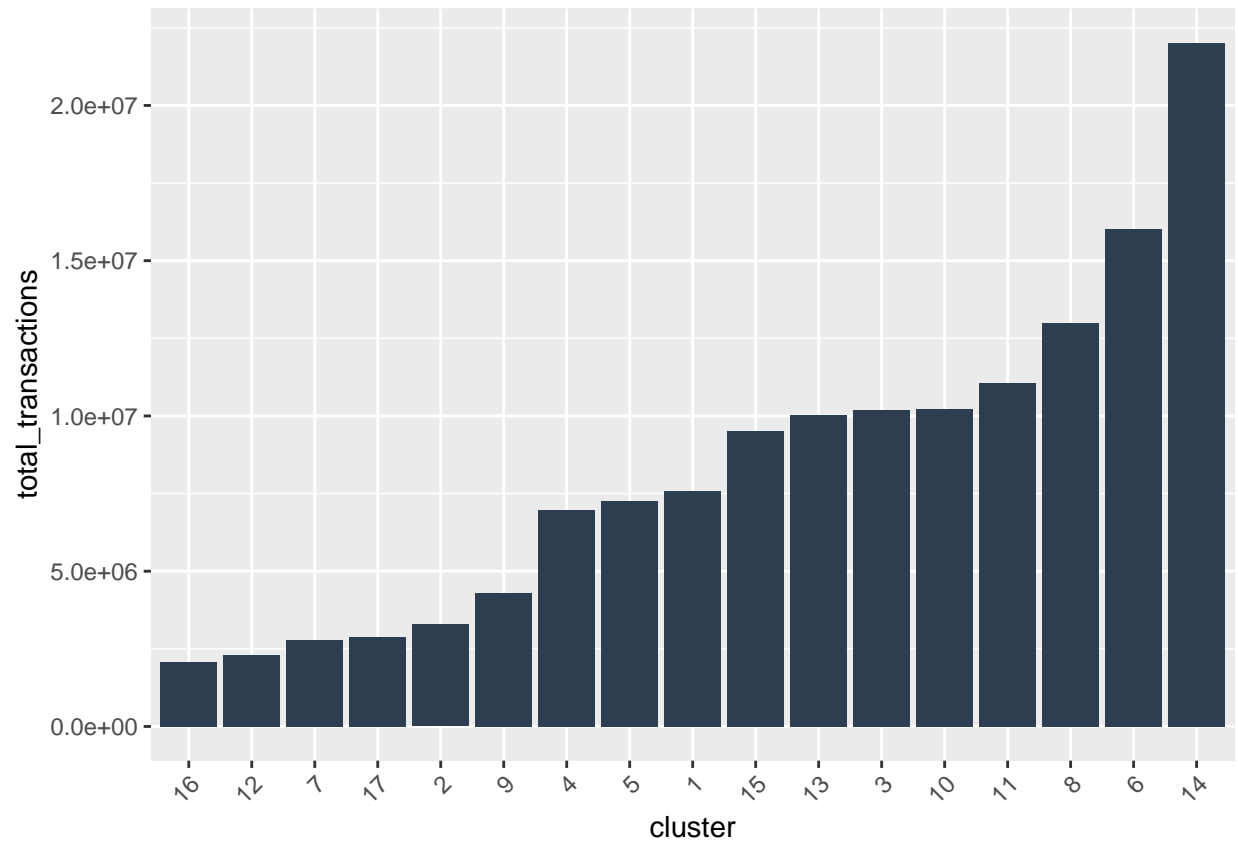
```
transactions_stores %>%  
  group_by(city) %>%  
  summarise(total_transactions = sum(transactions)) %>%  
  mutate(city = city %>% as_factor() %>% fct_reorder(total_transactions)) %>%  
  ggplot(aes(city, total_transactions)) +  
  geom_col(fill = "#2c3e50") +  
  theme(  
    axis.text.x = element_text(angle = 45, hjust = 1)  
  )
```



## transactions by store cluster

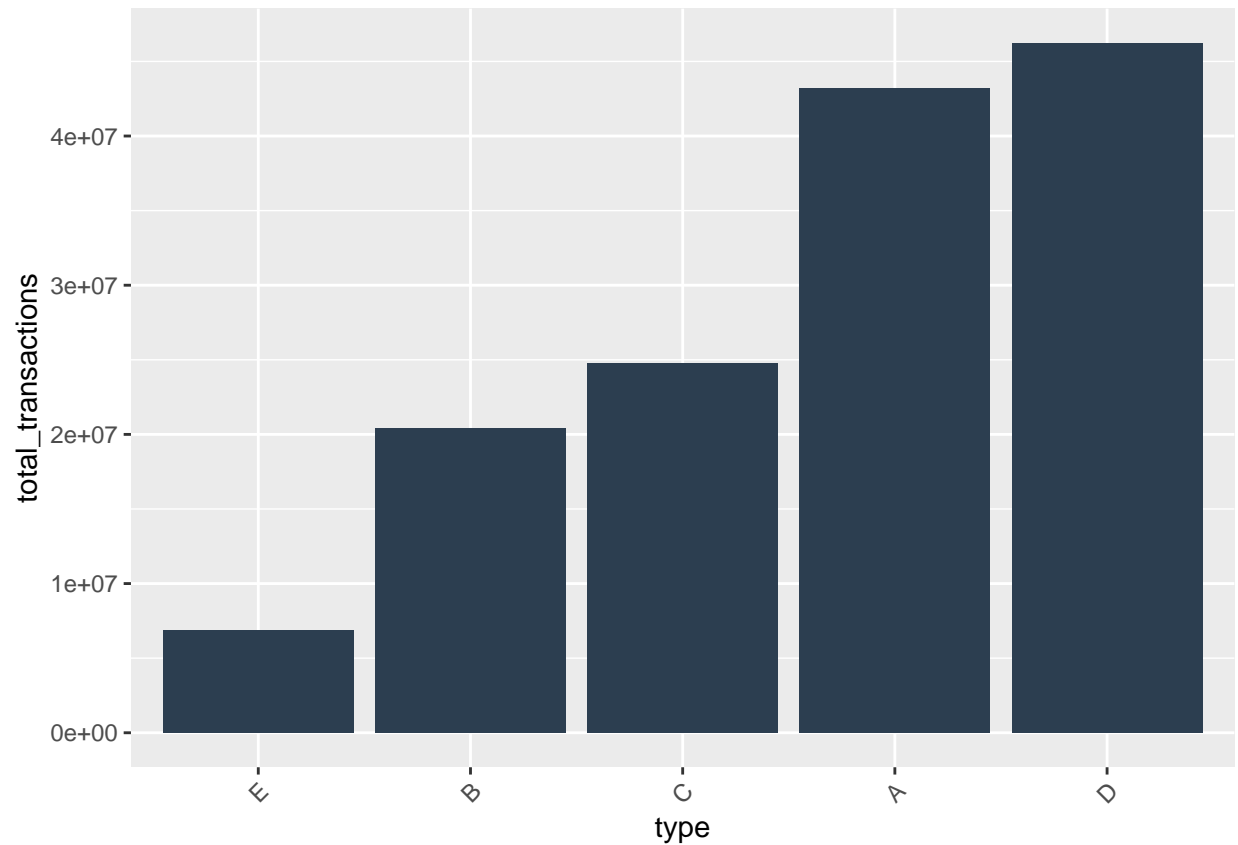
```
transactions_stores %>%
  group_by(cluster) %>%
  summarise(total_transactions = sum(transactions)) %>%
  mutate(cluster = cluster %>% as_factor() %>% fct_reorder(total_transactions)) %>%
  ggplot(aes(cluster, total_transactions)) +
  geom_col(fill = "#2c3e50") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```





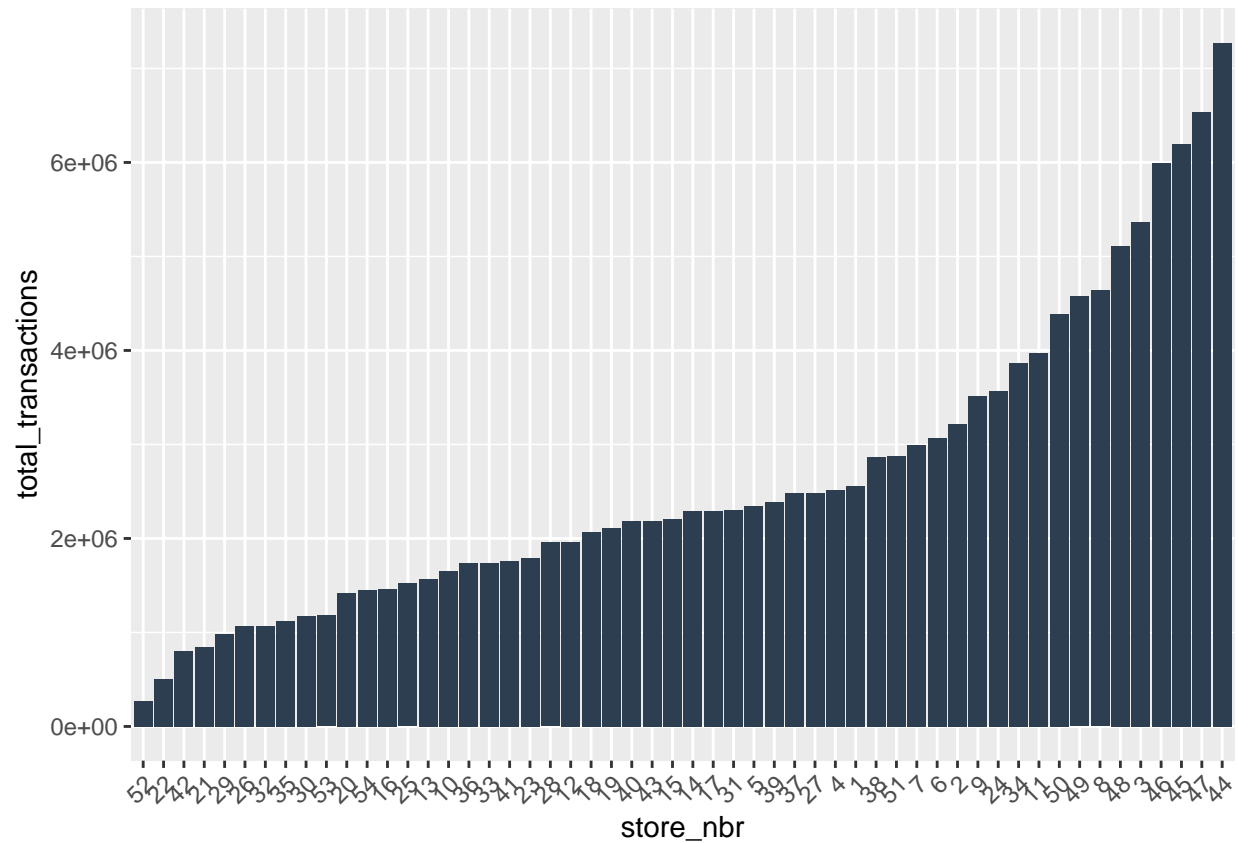
## transactions by store type

```
transactions_stores %>%
  group_by(type)%>%
  summarise(total_transactions = sum(transactions)) %>%
  mutate(type = type %>% as_factor() %>% fct_reorder(total_transactions)) %>%
  ggplot(aes(type, total_transactions)) +
  geom_col(fill = "#2c3e50") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



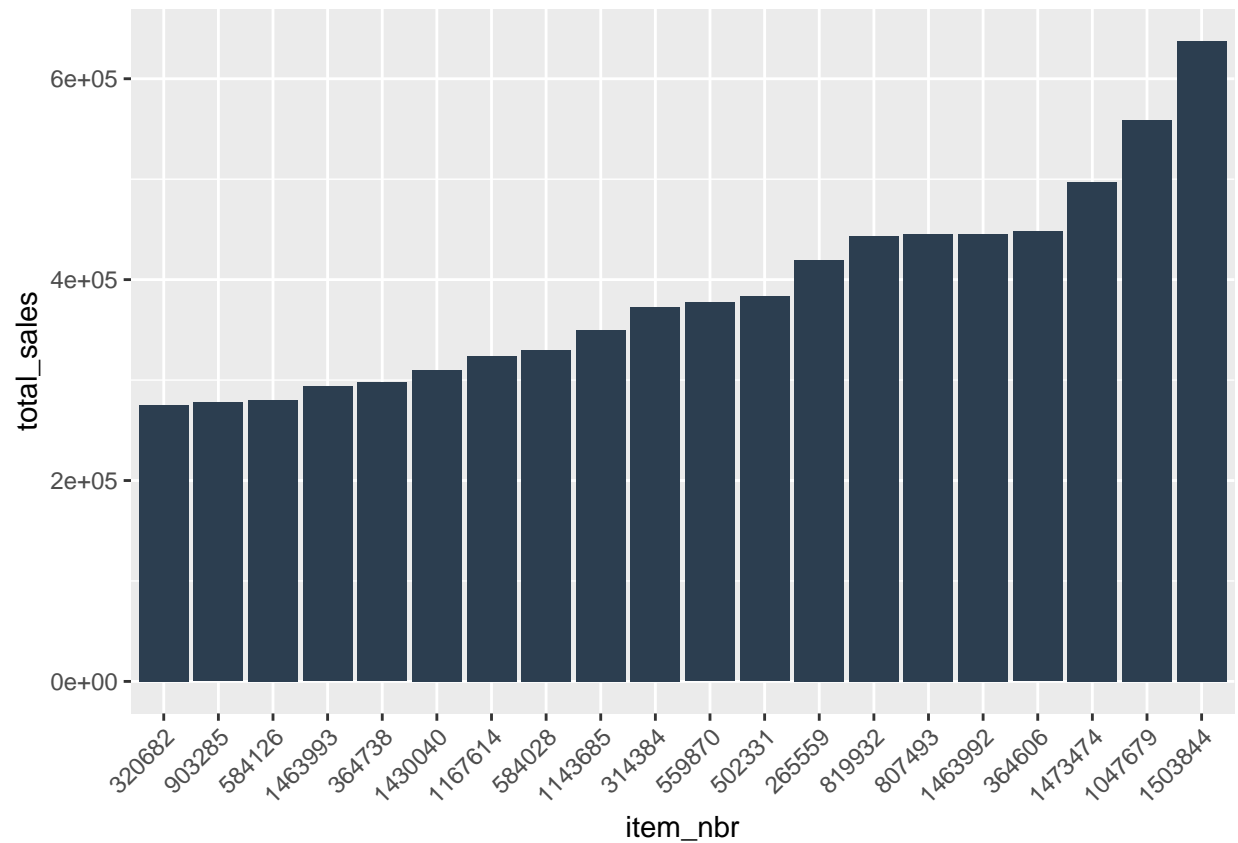
### transactions by store

```
transactions_stores %>%
  group_by(store_nbr)%>%
  summarise(total_transactions = sum(transactions)) %>%
  mutate(store_nbr = store_nbr %>% as_factor() %>% fct_reorder(total_transactions)) %>%
  ggplot(aes(store_nbr, total_transactions)) +
  geom_col(fill = "#2c3e50") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



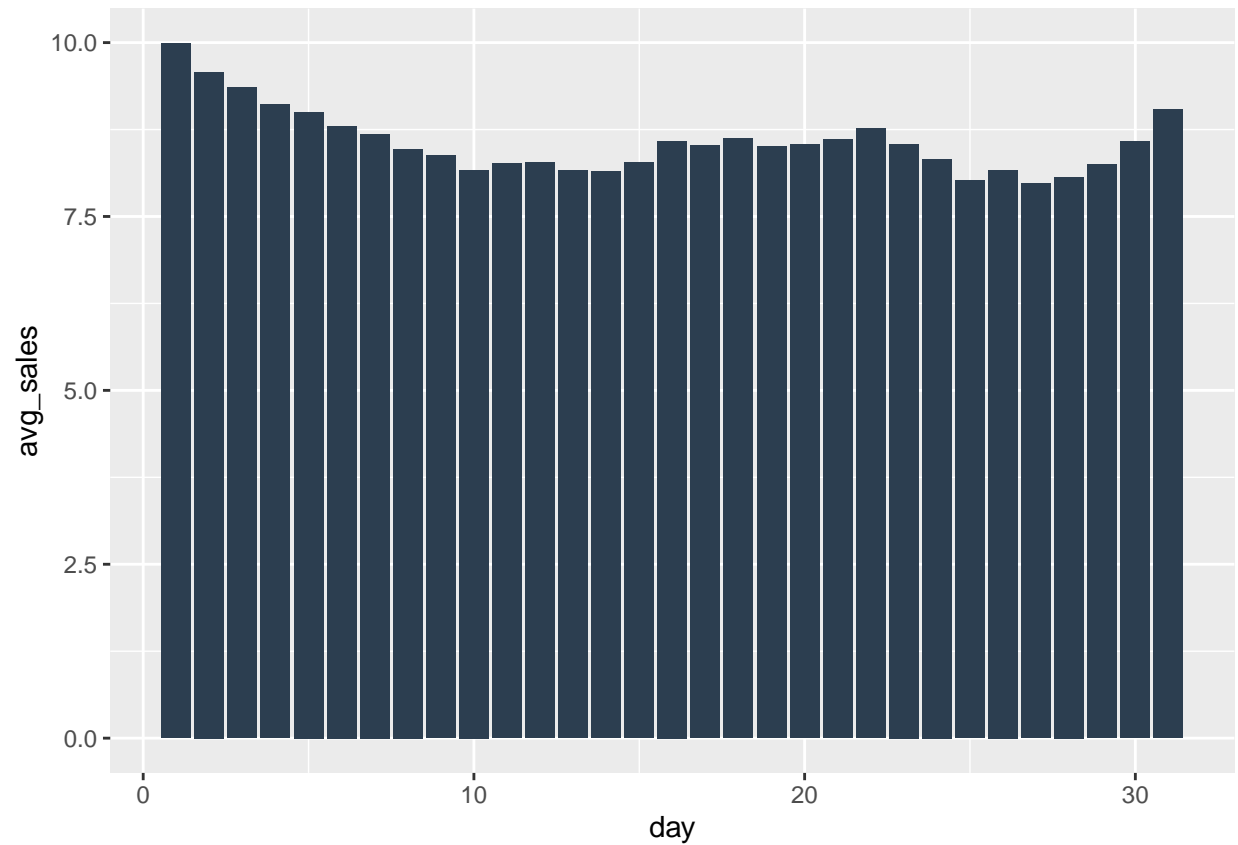
## top 20 items by sales

```
train_data_items_holidays %>%
  group_by(item_nbr)%>%
  summarise(total_sales = sum(unit_sales)) %>%
  arrange(desc(total_sales)) %>%
  top_n(20) %>%
  mutate(item_nbr = item_nbr %>% as_factor() %>% fct_reorder(total_sales)) %>%
  ggplot(aes(item_nbr, total_sales)) +
  geom_col(fill = "#2c3e50") +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



### sales by day of the month

```
train_data_items_holidays %>%
  group_by(day)%>%
  summarise(avg_sales = mean((unit_sales))) %>%
  ggplot() + geom_col(aes(x = day, y = avg_sales),fill = "#2c3e50")
```



### weekly sales trend

```
train_data_items_holidays %>%  
  group_by(weekday)%>%  
  summarise(avg_sales = mean(unit_sales)) %>%  
  ggplot() + geom_col(aes(x = weekday, y = avg_sales), fill = "#2c3e50") +  
  ggtitle("weekly sales") +  
  xlab("weekday") + ylab("avg sales")
```

