Predict Future Sales

2020-11-17

Contents

Data Descriptions **1**

Sales Train . . . 2

Items . . . 2

Data Preparation **3**

Analysis **4**

Total Sales by shop . . . 4

Items sold by shop . . . 5

Popular Items by shop . . . 6

Total item Category by shop . . . 7

Popular Item Category by shop . . . 8

Highest Sales Grossing Product Category . . . 9

List of top 3 product category sales and top 3 item sales by Shop ID . . . 9

Total Sales by day and month . . . 11

Total Sales per Year . . . 13

Item Sold per Weekdays . . . 14

Prediction **14**

Arima Model . . . 14

Data Visualization . . . 14

Stationarity test . . . 15

Autocorrelation . . . 16

Partial Autocorrelation . . . 17

Forecasting . . . 19

Exponential Smoothing . . . 20

Data Descriptions

File_Names	Description
sales_train.csv	the training set. Daily historical data from January 2013 to October 2015.
test.csv	the test set. You need to forecast the sales for these shops and products for November 2015.
sample_submission.csv	a sample submission file in the correct format.
items.csv	supplemental information about the items/products.
item_categories.csv	supplemental information about the items categories.
shops.csv	supplemental information about the shops.

Sales Train

Feature	Description	Data Type
date	sale date	date in format dd/mm/yyyy
date_block_num	a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33	integer
shop_id	unique identifier of a shop	integer
item_id	unique identifier of a product	integer
item_price	current price of an item	numeric
item_cnt_day	number of products sold. You are predicting a monthly amount of this measure	numeric

```
## 'data.frame':    2935849 obs. of  6 variables:
##  $ date          : chr  "02.01.2013" "03.01.2013" "05.01.2013" "06.01.2013" ...
##  $ date_block_num: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ shop_id       : int  59 25 25 25 25 25 25 25 25 25 ...
##  $ item_id       : int  22154 2552 2552 2554 2555 2564 2565 2572 2572 2573 ...
##  $ item_price    : num  999 899 899 1709 1099 ...
##  $ item_cnt_day  : num  1 1 -1 1 1 1 1 1 1 3 ...
```

Items

Feature	Description	Data Type
item_name	name of item	character
item_id	unique identifier of a product	integer
item_category_id	unique identifier of item category	integer

```
## [1] "Data Structure"
```

```
## 'data.frame':    22170 obs. of  3 variables:
##  $ item_name       : chr  "!              (   .)         D" "!ABBYY FineReader 12 Professional Editio
##  $ item_id         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ item_category_id: int  40 76 40 40 40 40 40 40 40 40 ...
```

Data Preparation

```
## [1] "Data Structure"
```

```
## 'data.frame':    2935849 obs. of  7 variables:
##  $ date            : chr  "02.01.2013" "03.01.2013" "05.01.2013" "06.01.2013" ...
##  $ date_block_num  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ shop_id         : int  59 25 25 25 25 25 25 25 25 25 ...
##  $ item_id         : int  22154 2552 2552 2554 2555 2564 2565 2572 2572 2573 ...
##  $ item_price      : num  999 899 899 1709 1099 ...
##  $ item_cnt_day    : num  1 1 -1 1 1 1 1 1 1 3 ...
##  $ item_category_id: int  37 58 58 58 56 59 56 55 55 55 ...
```

The 'date' column should be a 'date' type and need to be convert. Also, it may be useful to separate the date into year, month, day and weekdays.

```
## [1] "Data Structure"
```

```
## 'data.frame':    2935849 obs. of  12 variables:
##  $ date            : Date, format: "2013-01-02" "2013-01-03" ...
##  $ date_block_num  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ shop_id         : int  59 25 25 25 25 25 25 25 25 25 ...
##  $ item_id         : int  22154 2552 2552 2554 2555 2564 2565 2572 2572 2573 ...
##  $ item_price      : num  999 899 899 1709 1099 ...
##  $ item_cnt_day    : num  1 1 -1 1 1 1 1 1 1 3 ...
##  $ item_category_id: int  37 58 58 58 56 59 56 55 55 55 ...
##  $ year            : Factor w/ 3 levels "2013","2014",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ month           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day             : int  2 3 5 6 15 10 2 4 11 3 ...
##  $ weekday         : Factor w/ 7 levels "Friday","Monday",..: 7 5 3 4 6 5 7 1 1 5 ...
##  $ weeknumber      : chr  "01" "01" "01" "01" ...
```

```
## [1] "Data Summary"
```

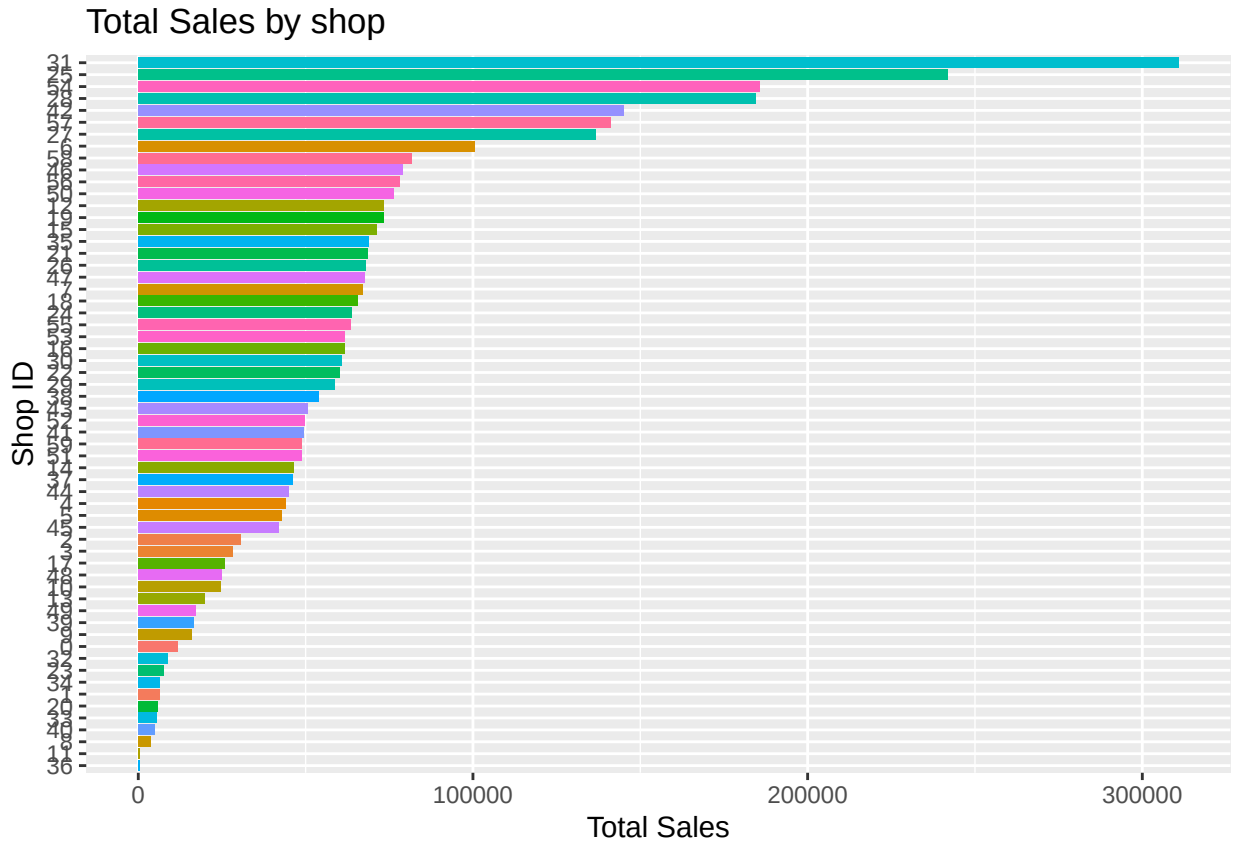
```
##       date            date_block_num     shop_id          item_id
##  Min.   :2013-01-01   Min.   : 0.00   31     : 235636   20949  :  31340
##  1st Qu.:2013-08-01   1st Qu.: 7.00   25     : 186104   5822   :   9408
##  Median :2014-03-04   Median :14.00   54     : 143480   17717  :   9067
##  Mean   :2014-04-03   Mean   :14.57   28     : 142234   2808   :   7479
##  3rd Qu.:2014-12-05   3rd Qu.:23.00   57     : 117428   4181   :   6853
##  Max.   :2015-10-31   Max.   :33.00   42     : 109253   7856   :   6602
##                                       (Other):2001714   (Other):2865100
##    item_price        item_cnt_day      item_category_id    year
##  Min.   :    -1.0   Min.   : -22.000   40     : 564652   2013:1267562
##  1st Qu.:   249.0   1st Qu.:   1.000   30     : 351591   2014:1055861
##  Median :   399.0   Median :   1.000   55     : 339585   2015: 612426
##  Mean   :   890.9   Mean   :   1.243   19     : 208219
##  3rd Qu.:   999.0   3rd Qu.:   1.000   37     : 192674
##  Max.   :307980.0   Max.   :2169.000   23     : 146789
##                                        (Other):1132339
##      month             day             weekday        weeknumber
##  Min.   : 1.000   Min.   : 1.00   Friday   :439298   Length:2935849
```

```
##  1st Qu.: 3.000   1st Qu.: 8.00   Monday   :337074   Class :character
##  Median : 6.000   Median :16.00   Saturday :590359   Mode  :character
##  Mean   : 6.248   Mean   :15.85   Sunday   :503104
##  3rd Qu.: 9.000   3rd Qu.:24.00   Thursday :367280
##  Max.   :12.000   Max.   :31.00   Tuesday  :345772
##                                   Wednesday:352962
##  item_cnt_month
##  Min.   : -22.000
##  1st Qu.:   1.000
##  Median :   2.000
##  Mean   :   7.401
##  3rd Qu.:   5.000
##  Max.   :2253.000
##
```

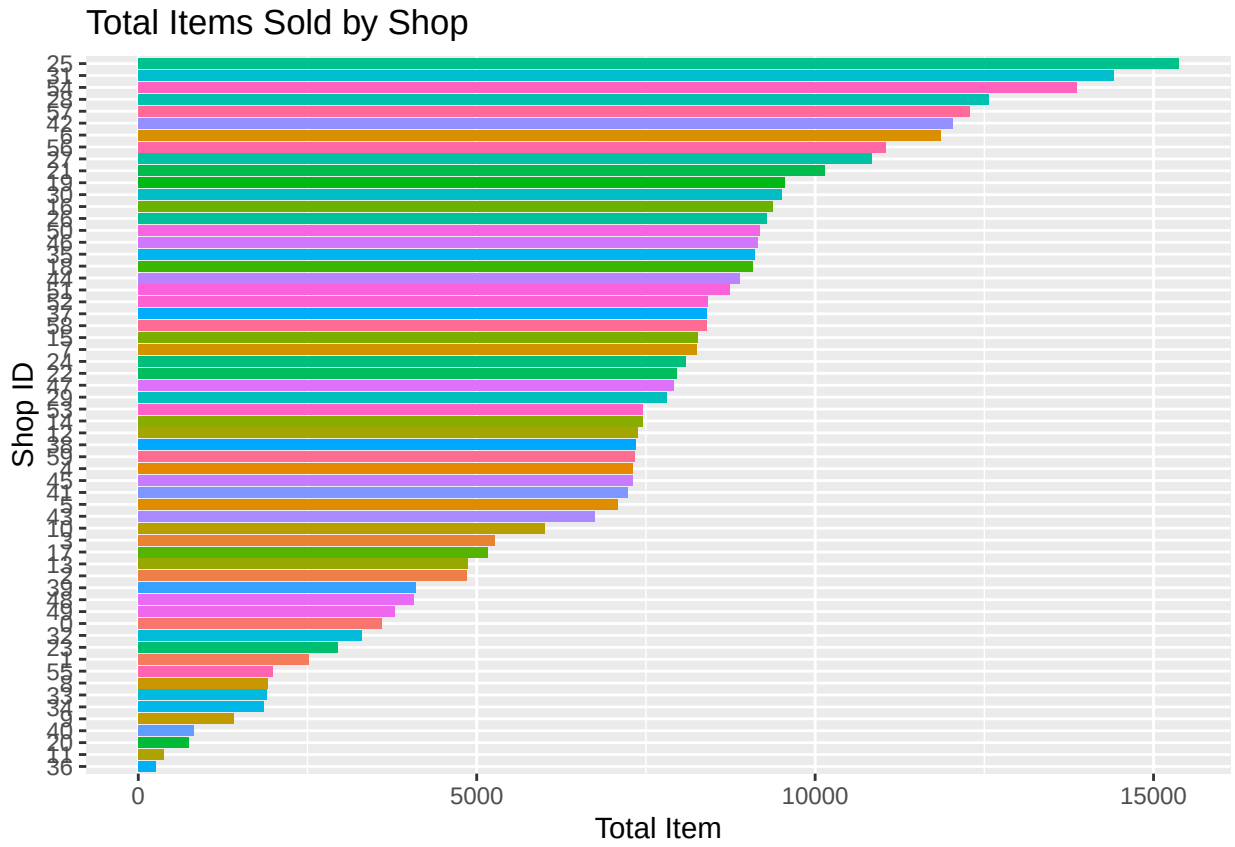
As we can see there are no missing values.

Analysis

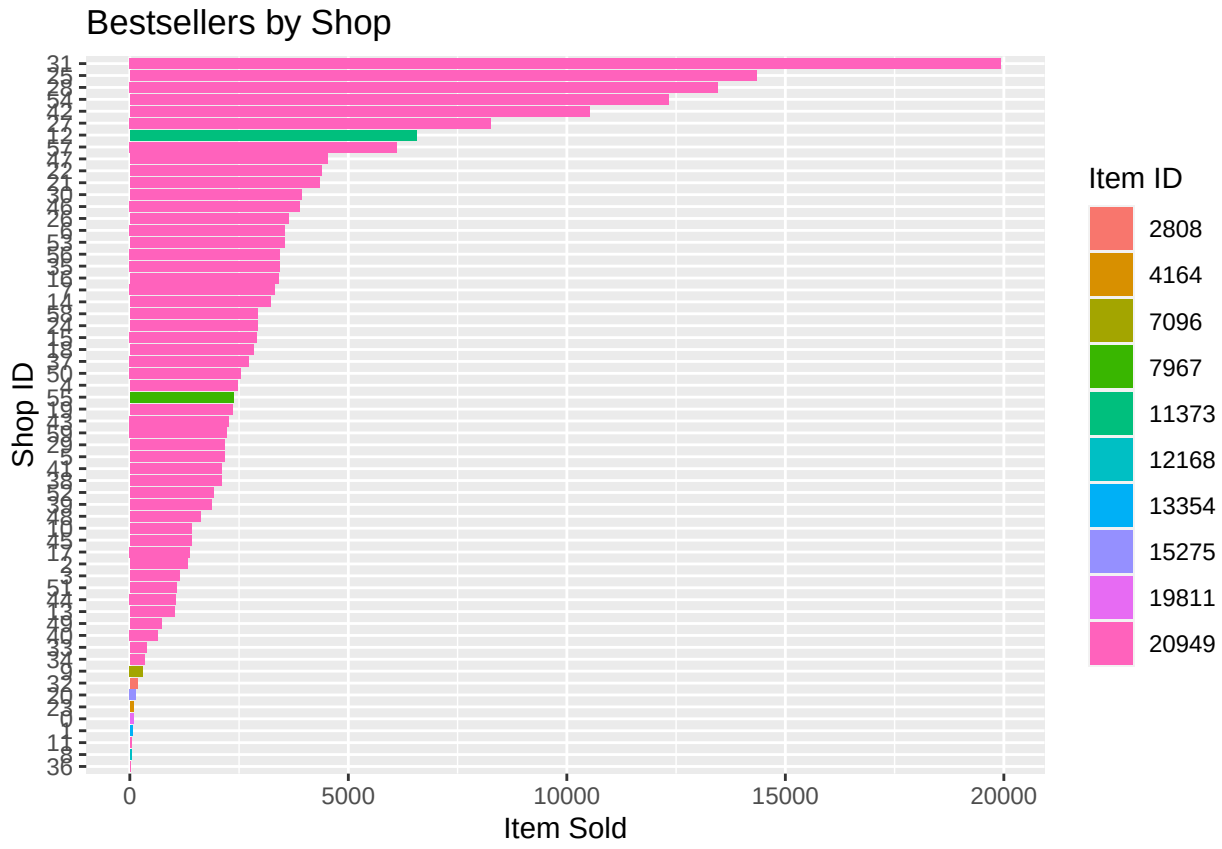
Total Sales by shop



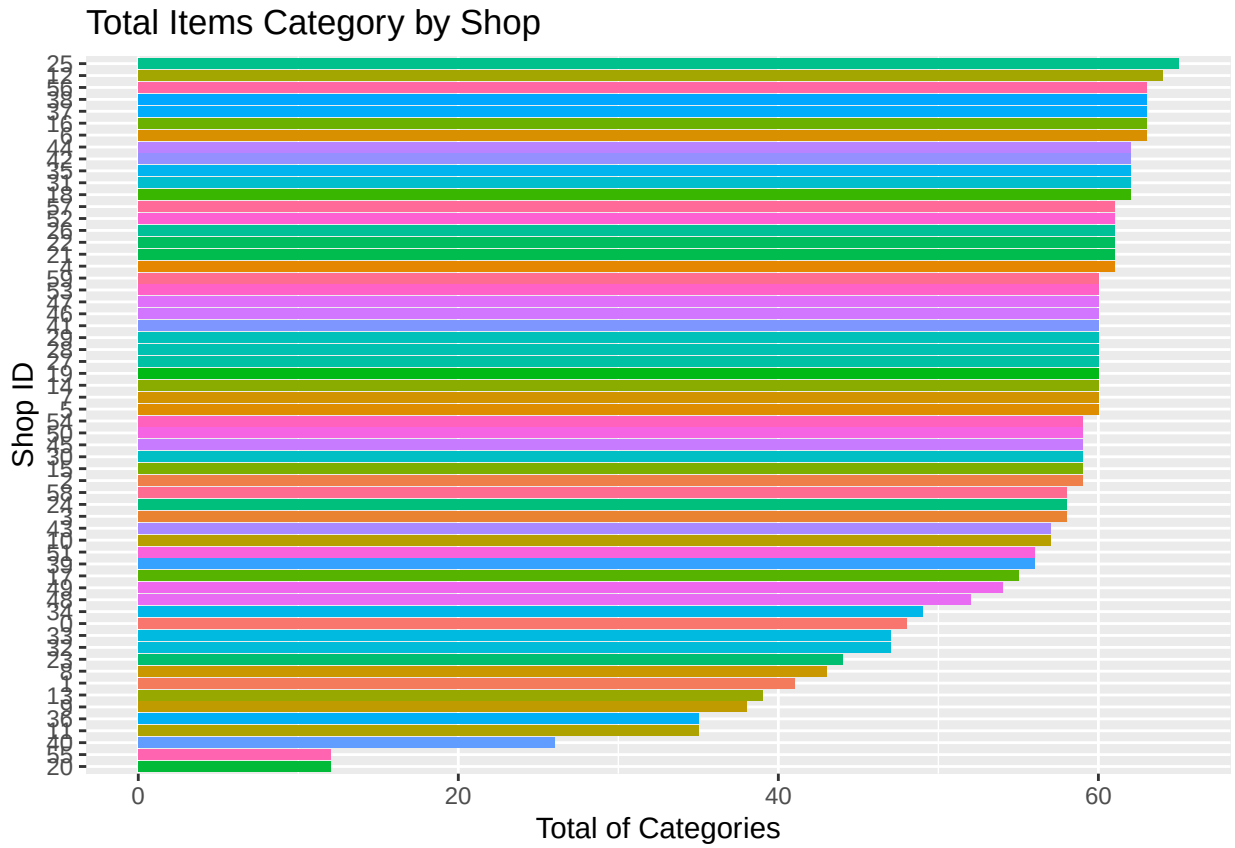
Items sold by shop



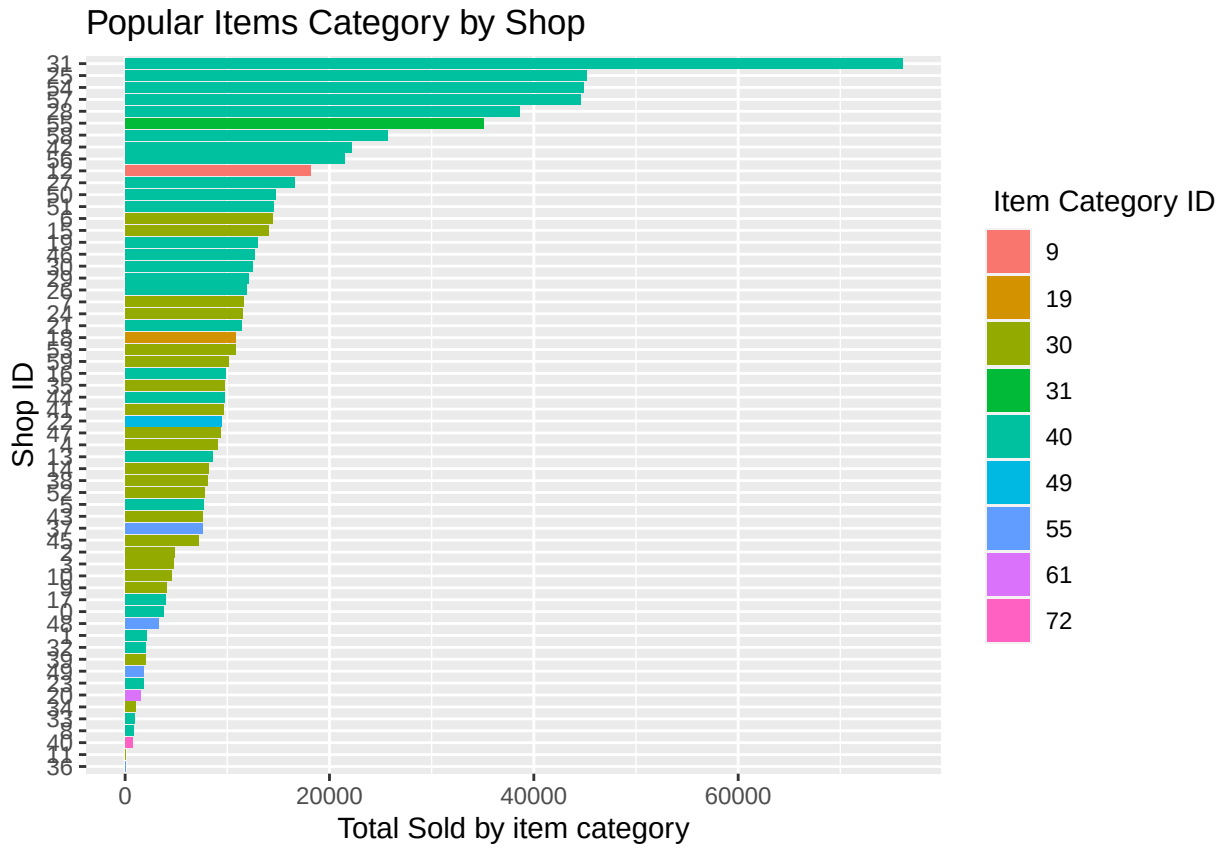
Popular Items by shop



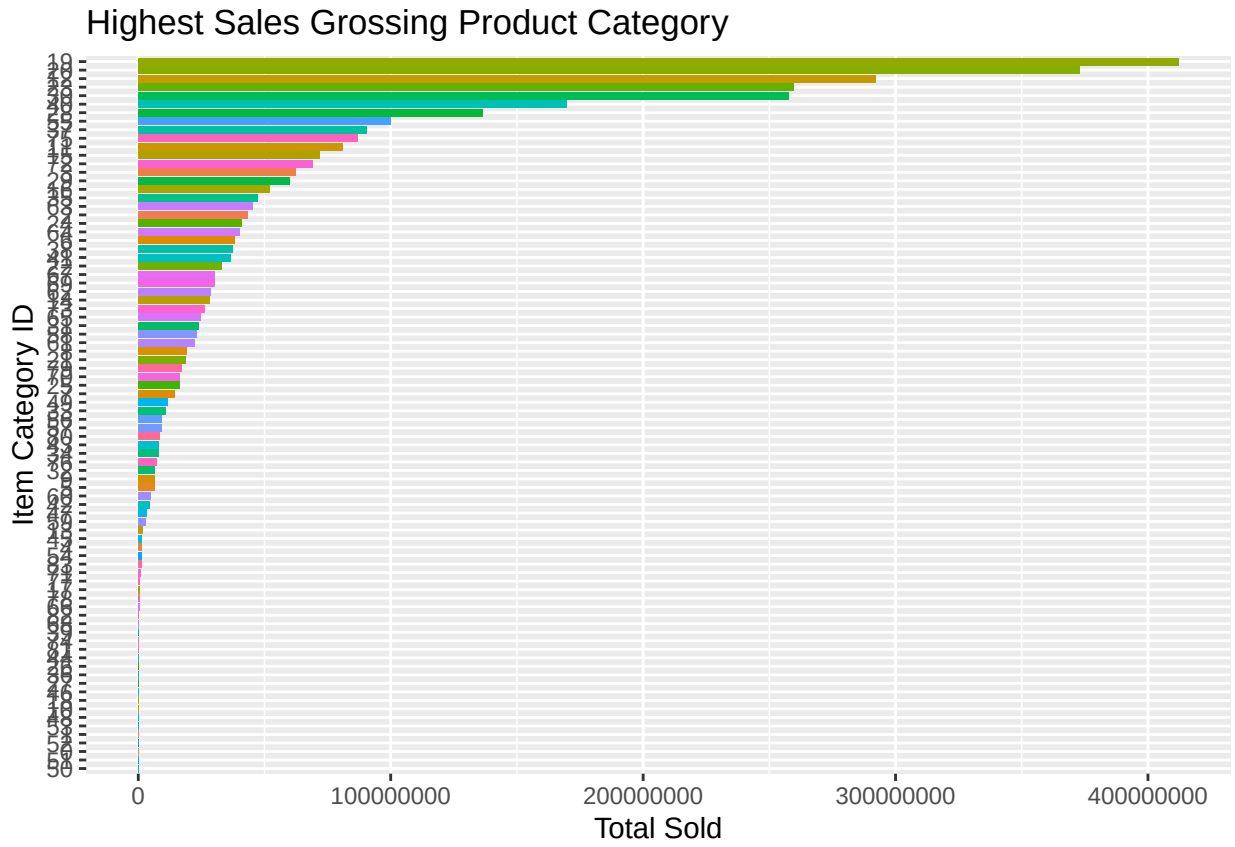
Total item Category by shop



Popular Item Category by shop



Highest Sales Grossing Product Category

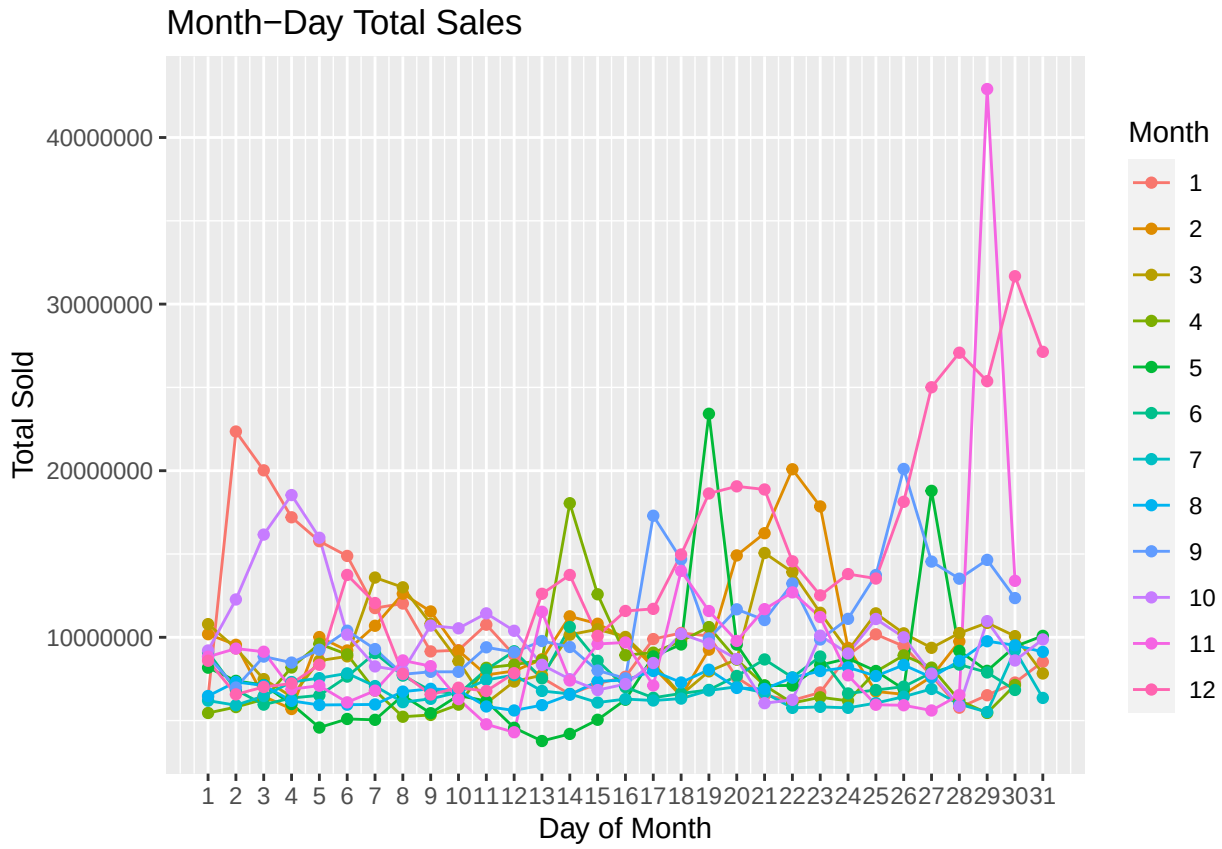


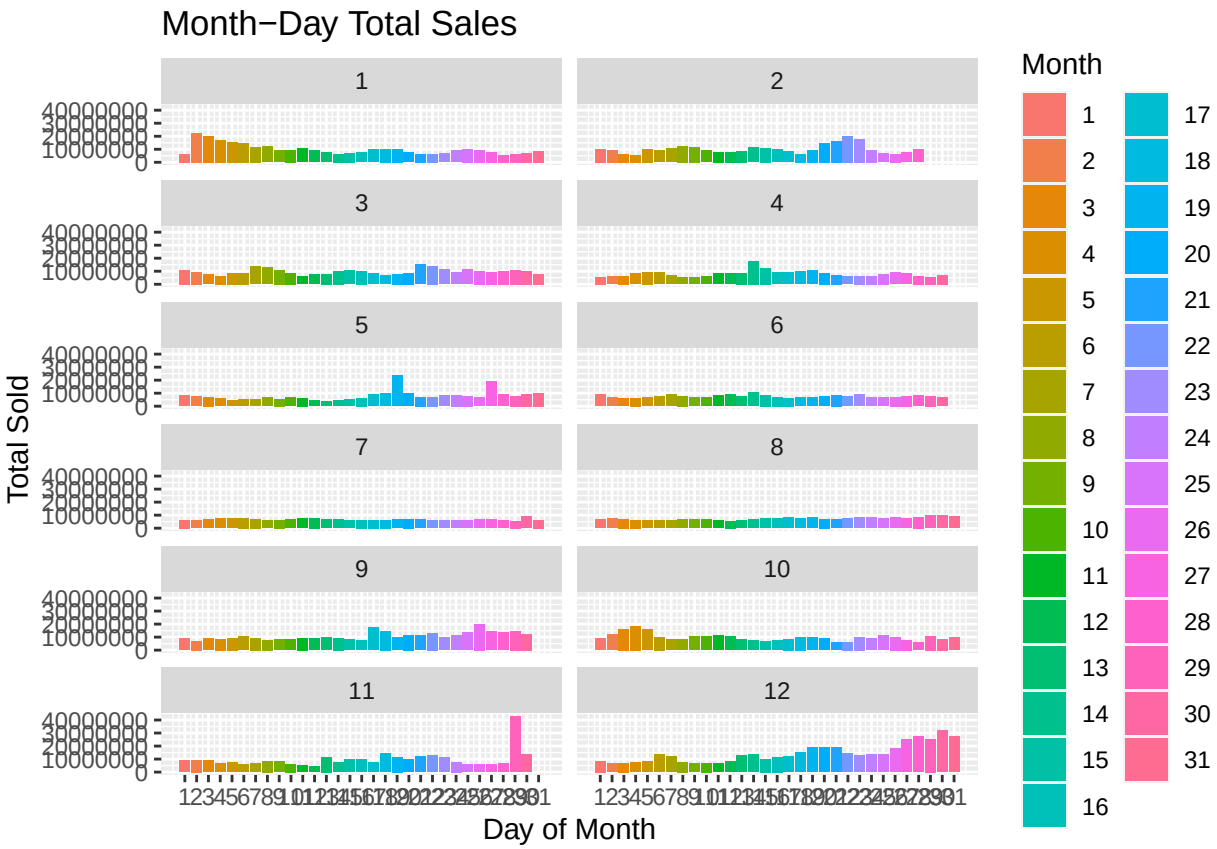
List of top 3 product category sales and top 3 item sales by Shop ID

shop_id	Category_ID	Item_ID
0	40, 30, 55	19811, 14752, 4163
1	40, 30, 55	13354, 13351, 14447
10	30, 40, 55	20949, 4181, 5822
11	30, 40, 55	20949, 4181, 7087, 16825
12	9, 30, 49	11373, 11370, 11369
13	40, 55, 30	20949, 13345, 13351
14	30, 55, 40	20949, 2808, 17717
15	30, 40, 55	20949, 4181, 5822
16	40, 30, 55	20949, 4181, 7856
17	40, 55, 19	20949, 3732, 6675
18	19, 30, 55	20949, 3732, 5822
19	40, 30, 55	20949, 5822, 3732
2	30, 19, 23	20949, 3732, 17717
20	61, 63, 72	15275, 13246, 9396
21	40, 55, 30	20949, 17717, 16832
22	49, 30, 40	20949, 482, 4181
23	40, 30, 55	4164, 2445, 1830, 6738
24	30, 40, 55	20949, 4181, 4178
25	40, 55, 30	20949, 2808, 3732

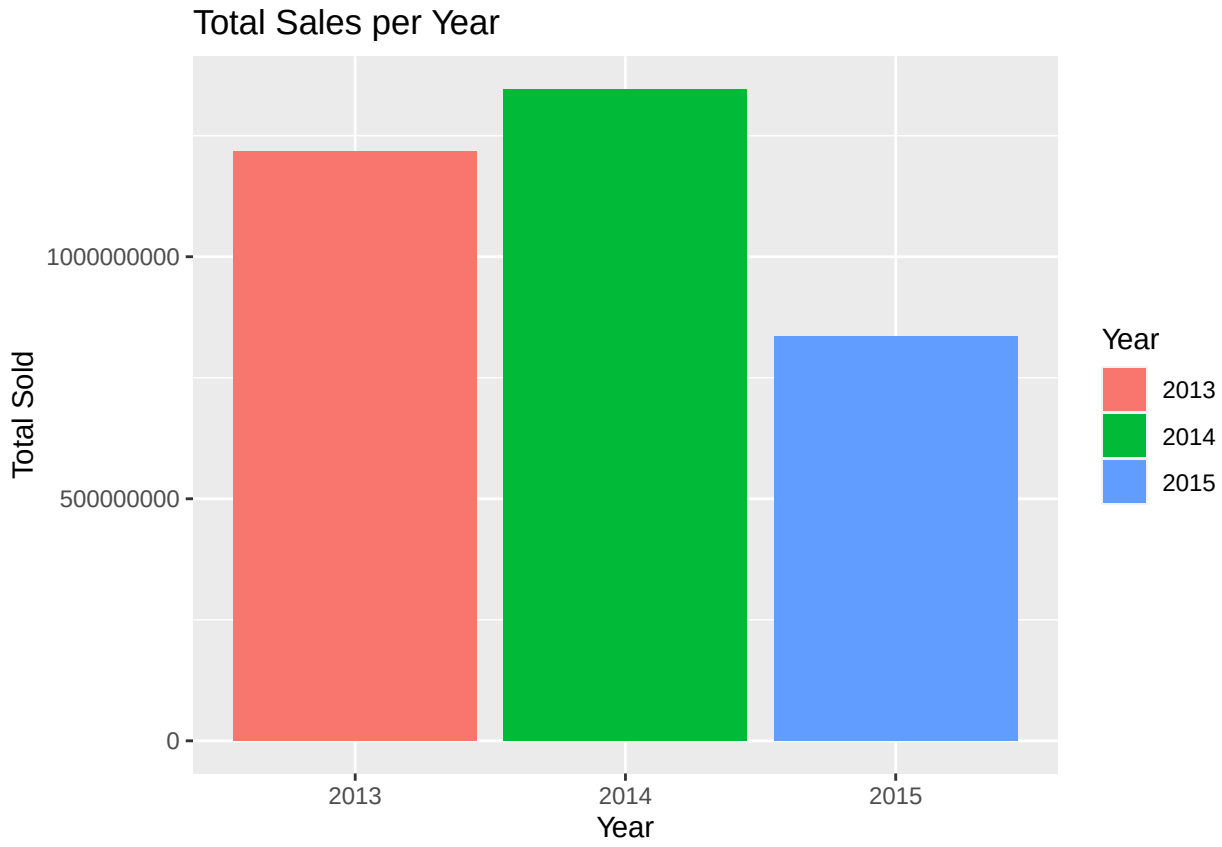
shop_id	Category_ID	Item_ID
26	40, 30, 55	20949, 2808, 5822
27	40, 30, 55	20949, 17717, 8057
28	40, 55, 30	20949, 3732, 5822
29	40, 30, 55	20949, 2808, 5822
3	30, 40, 55	20949, 17717, 3732
30	40, 30, 55	20949, 3732, 2808
31	40, 30, 55	20949, 5822, 17717
32	40, 30, 55	2808, 6738, 12168
33	40, 55, 30	20949, 17717, 5822
34	30, 20, 19	20949, 3731, 17717
35	30, 40, 55	20949, 2808, 17717
36	55, 20, 30	20949, 10201, 1583, 2423
37	55, 30, 40	20949, 3731, 2808
38	30, 19, 55	20949, 17717, 3732
39	30, 71, 55	20949, 3731, 17717
4	30, 40, 55	20949, 17717, 7856
40	72, 63, 71	20949, 20609, 20608
41	30, 40, 19	20949, 2808, 3331
42	40, 30, 55	20949, 17717, 5822
43	30, 19, 40	20949, 3732, 5822
44	40, 55, 30	20949, 13345, 13354
45	30, 40, 55	20949, 2808, 2814
46	40, 30, 55	20949, 5822, 2808
47	30, 19, 40	20949, 3732, 17717
48	55, 30, 23	20949, 17717, 6503
49	55, 30, 40	20949, 17717, 6503
5	40, 30, 55	20949, 17717, 3732
50	40, 30, 55	20949, 3732, 2808
51	40, 55, 30	20949, 13354, 13344
52	30, 40, 55	20949, 3732, 17717
53	30, 55, 40	20949, 3734, 17717
54	40, 55, 30	20949, 3732, 2808
55	31, 54, 34	7967, 492, 9249
56	40, 30, 55	20949, 5822, 13351
57	40, 30, 55	20949, 4178, 4870
58	40, 30, 23	20949, 4870, 3077, 12134
59	30, 40, 55	20949, 17717, 4181
6	30, 40, 55	20949, 17717, 3732
7	30, 40, 55	20949, 17717, 3732
8	40, 30, 55	12168, 3432, 2808
9	30, 61, 70	7096, 6457, 19436

Total Sales by day and month

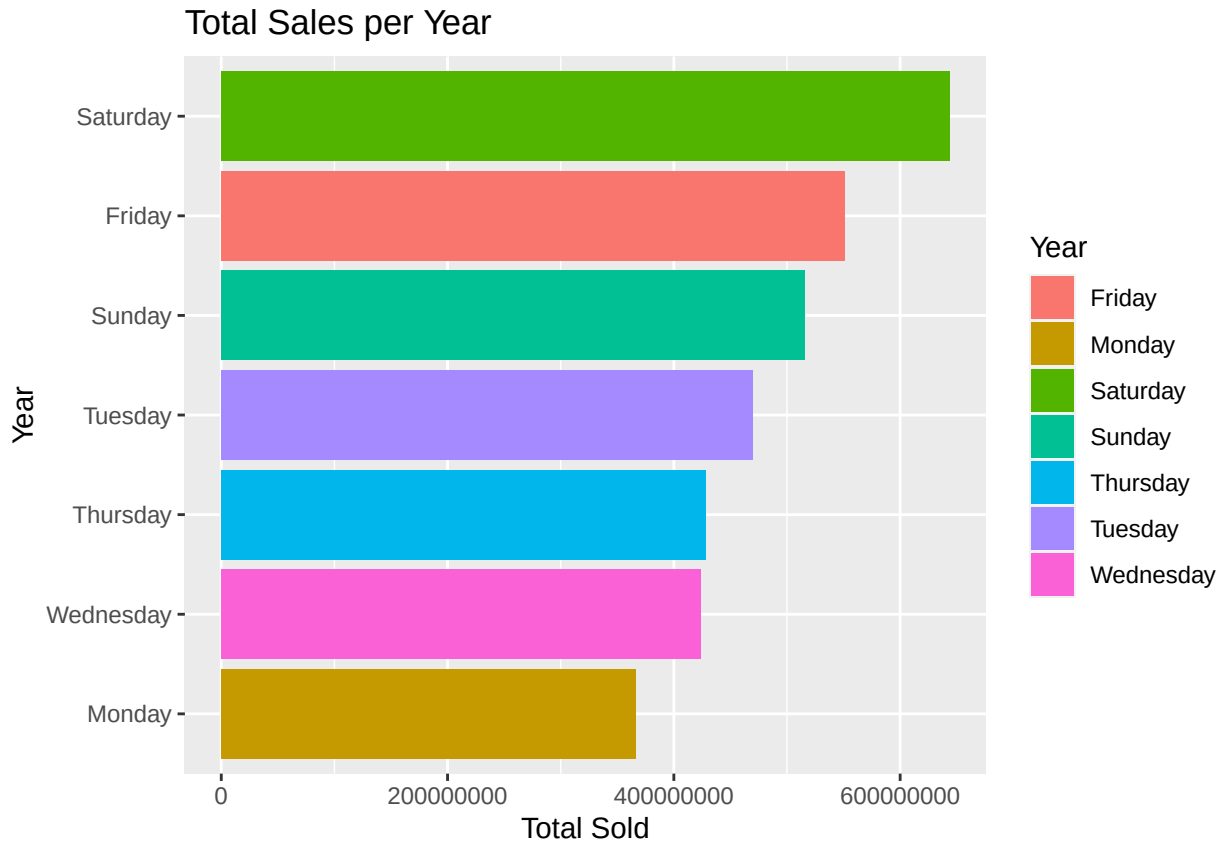



Month−Day Total Sales
1
2
3
4
5
6
7
8
9
10
11
12
Total Sold
Day of Month
Month
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

Total Sales per Year



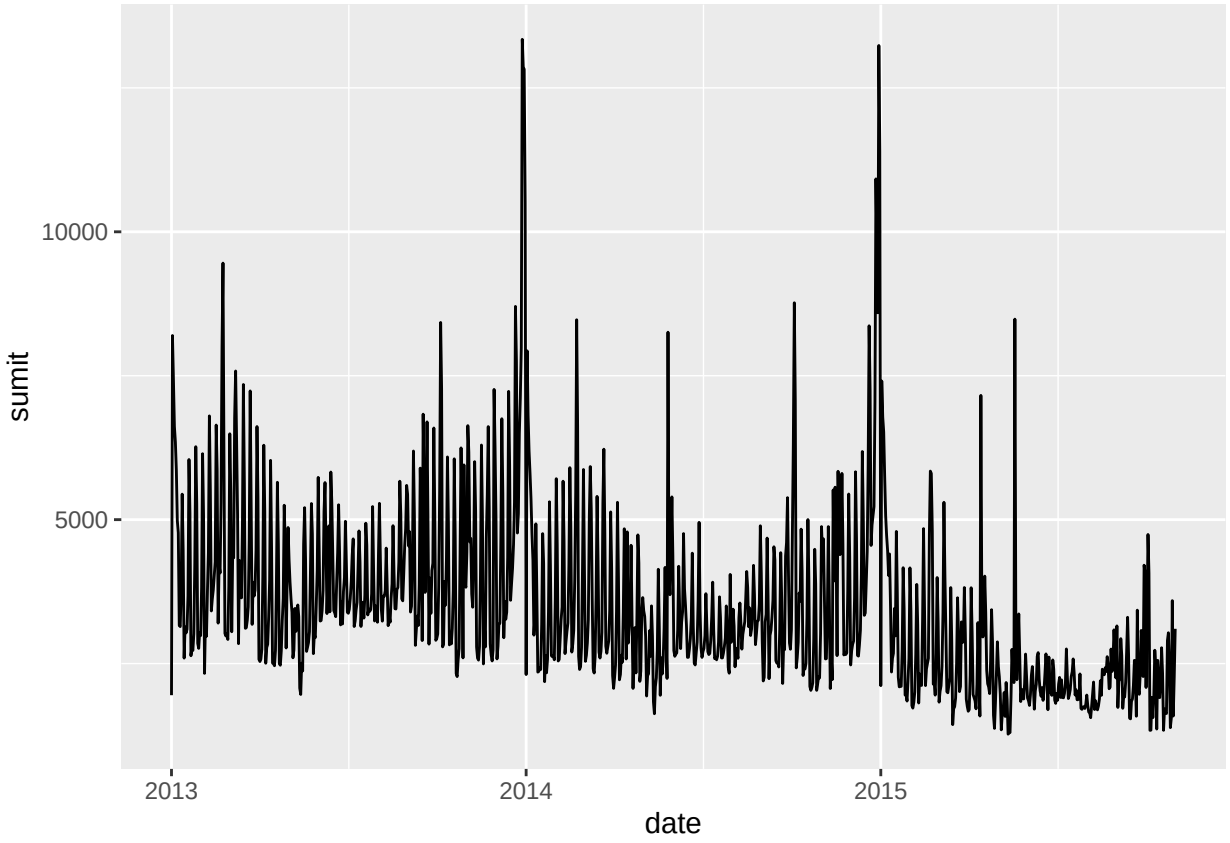
Item Sold per Weekdays



Prediction

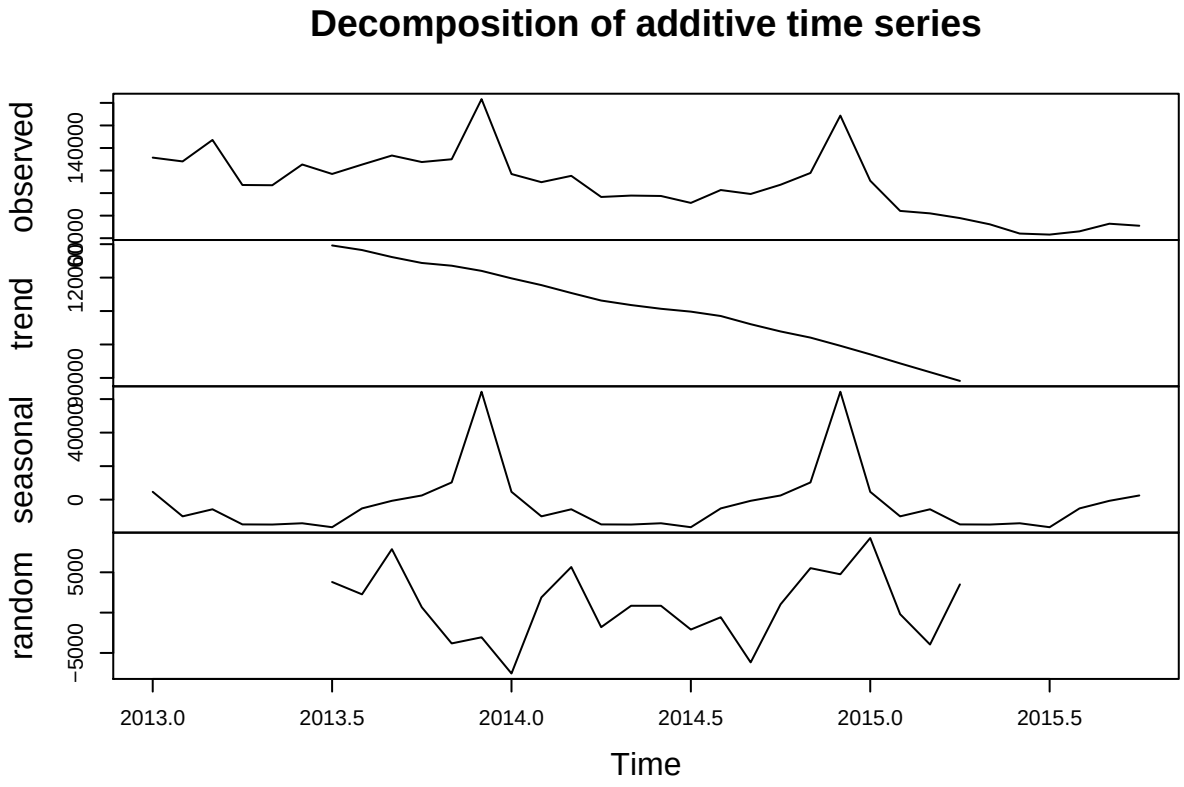
Arima Model

Data Visualization



Stationarity test The Augmented Dickey–Fuller (ADF) t-statistic test: small p-values suggest the data is stationary and doesn't need to be differenced stationarity. High p-value (>0.05) shows that the data is non stationary.

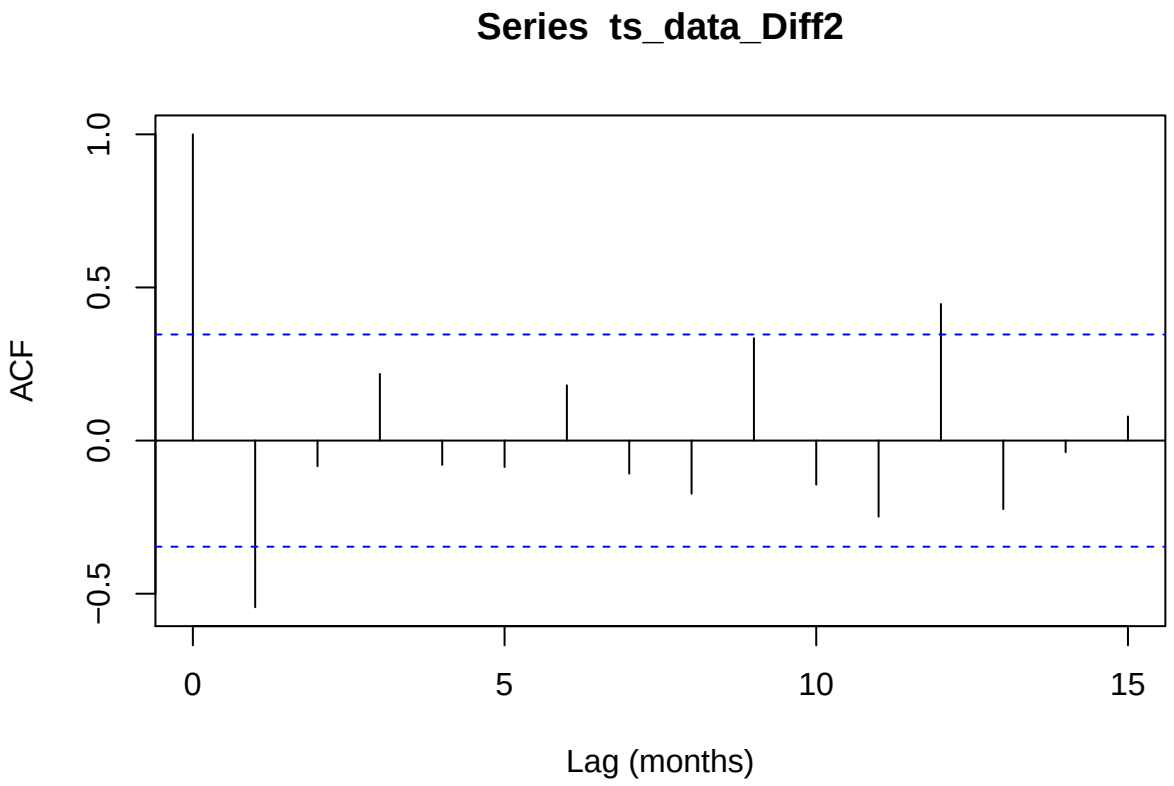
```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -2.743, Lag order = 3, p-value = 0.2851
## alternative hypothesis: stationary
```



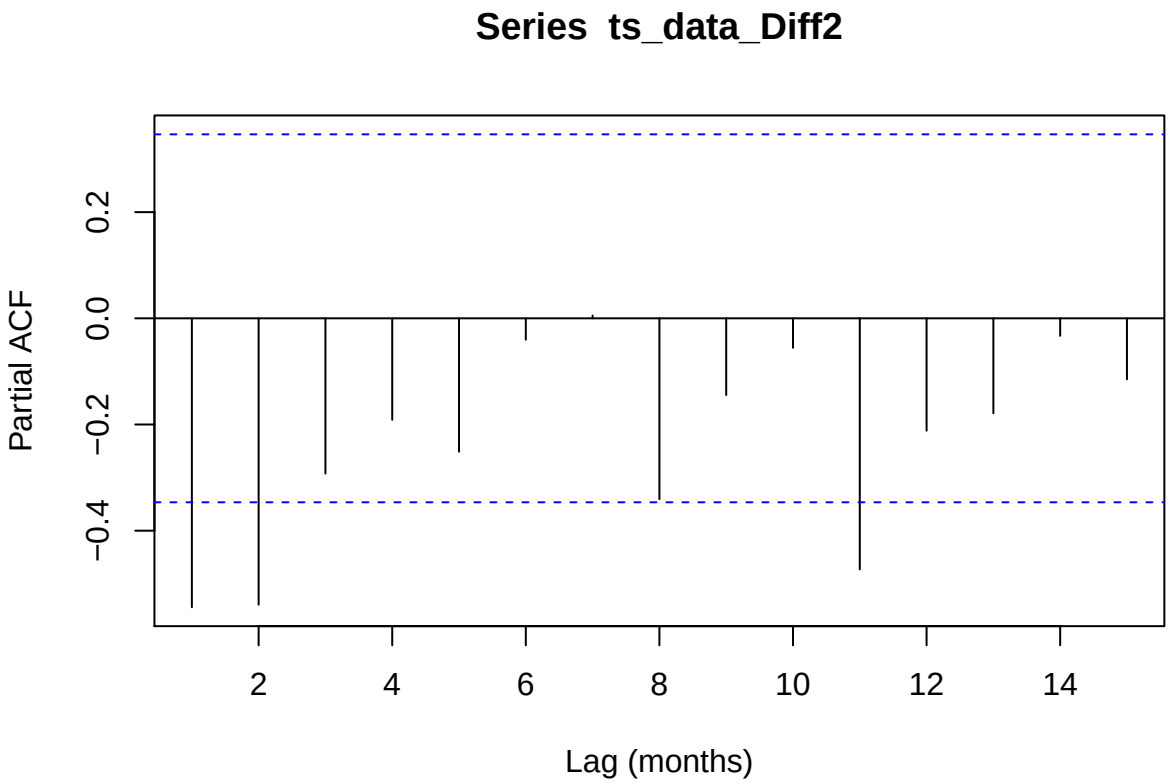

```
## [1] "total_sales"
```

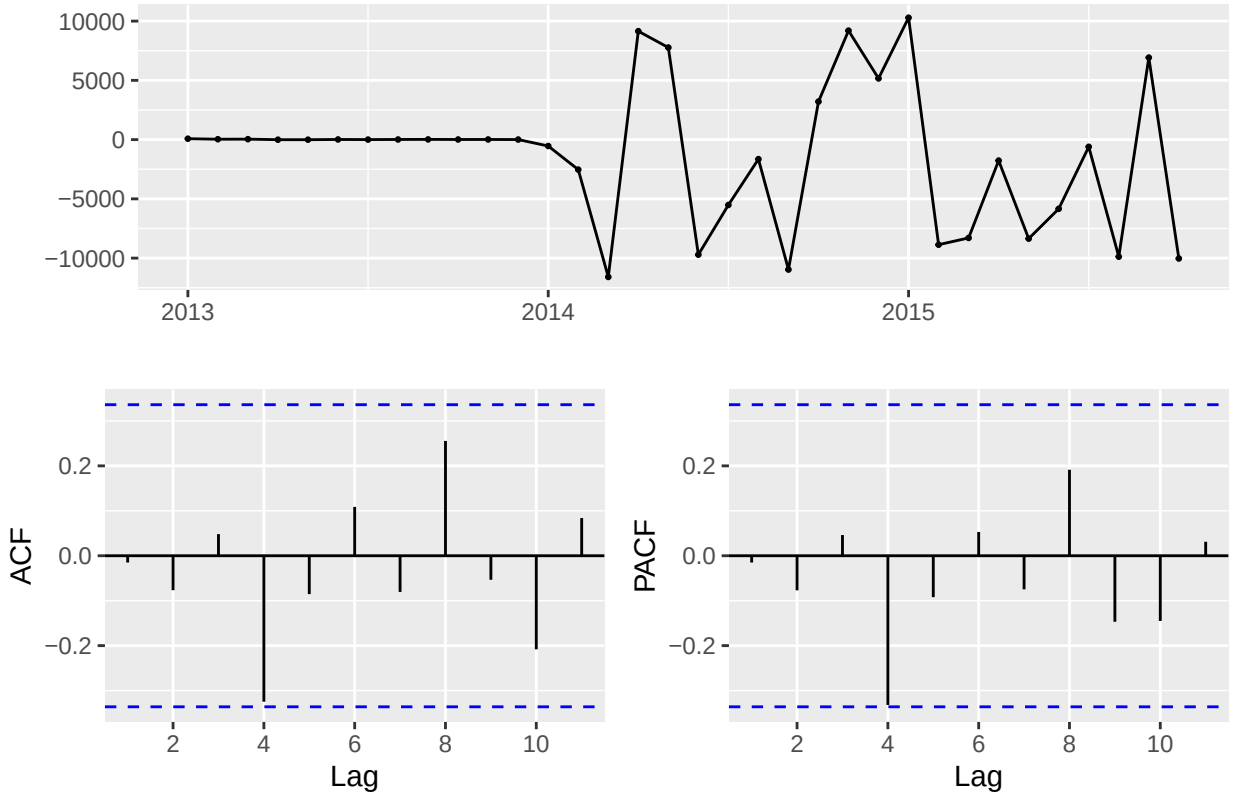
```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data_Diff2
## Dickey-Fuller = -4.2109, Lag order = 3, p-value = 0.01416
## alternative hypothesis: stationary
```

Autocorrelation

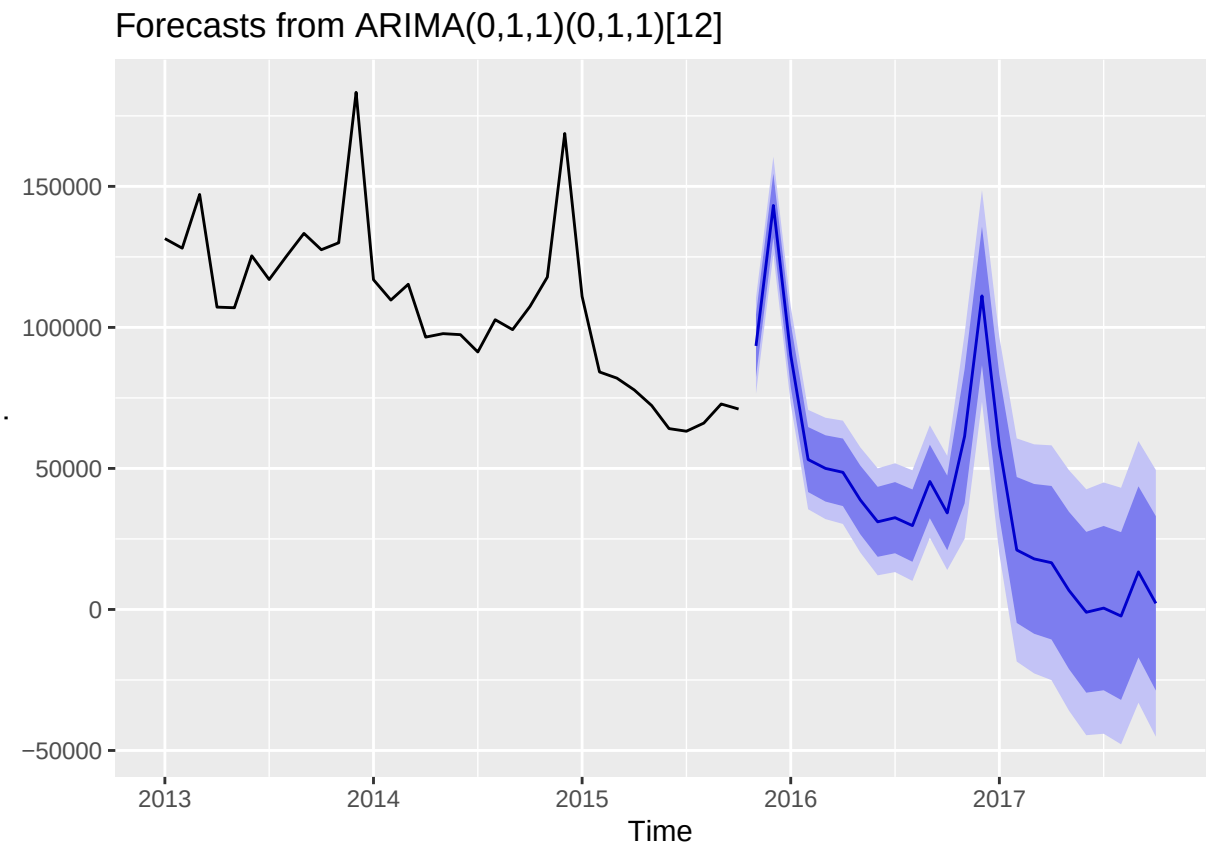


Partial Autocorrelation


Series ts_data_Diff2
Partial ACF
0.2
0.0
−0.2
−0.4
2
4
6
8
10
12
14
Lag (months)



Forecasting



Exponential Smoothing

Soon.