Business Forecasting Midterm

OLIST Sales Predictions

Ali, Mamadou S., Mamadou T., & Mackenzie

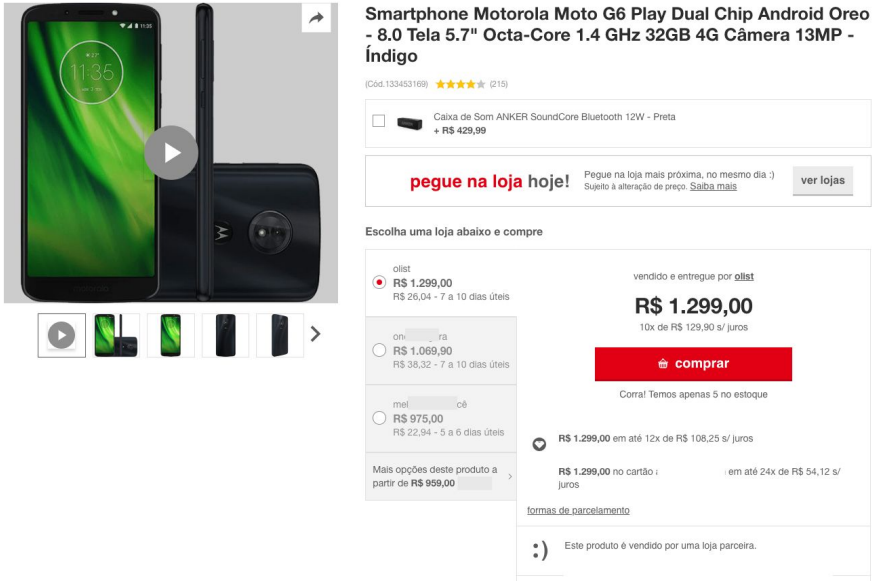
Agenda

1. About the Data & Introduction to the Problem
2. Data Collection & Preparation
3. Exploratory Data Analysis
4. Modeling Approach
5. Modeling Validation & Selection for Each Goal
6. Final Results

About OLIST & the Data

OLIST provides **e-commerce support for small & medium business** looking to win customers. It is a **online marketplace** (much like Amazon) where entrepreneurs can sell their products. The company is located in Brazil.

- 9 data tables
- Over 100K orders from 2016 to 2018
- From Kaggle



Introduction

Three main objectives:

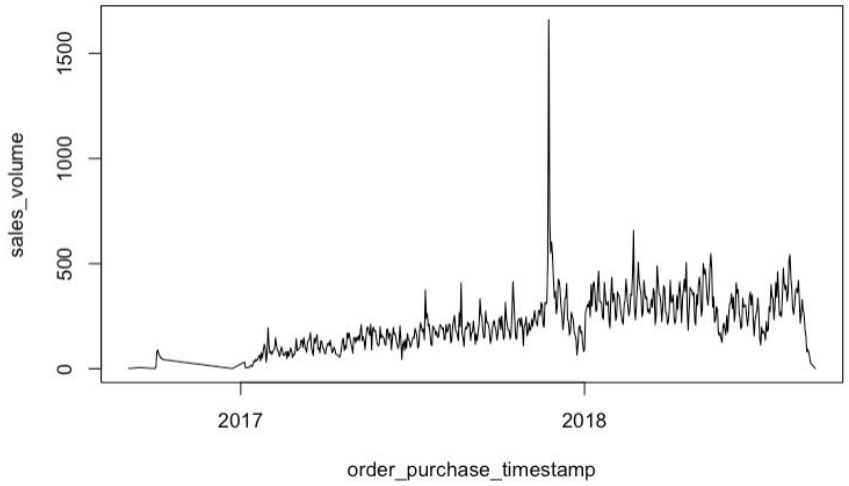
1. provide a **three-month forecast of future sales** for a Brazilian E-commerce startup called OLIST,
2. determine the **three best-selling categories**, and forecast their growth over the next three months, and
3. determine the **fastest-growing category**, and forecast its growth over the next three months

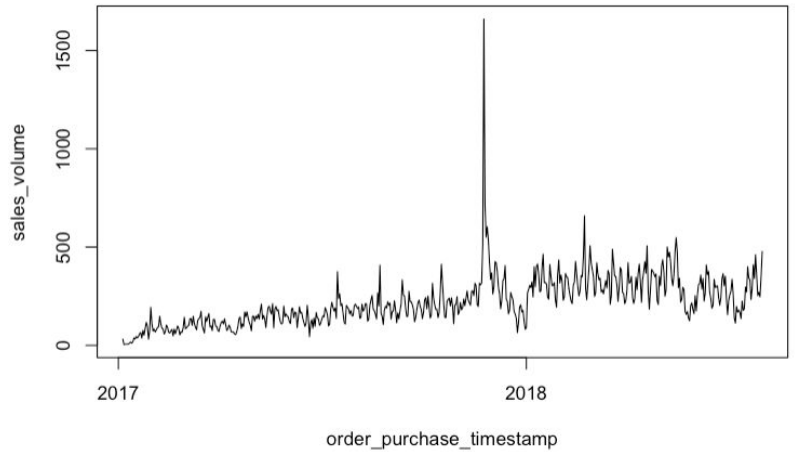
Some Definitions

- Sales → quantity of items being sold
  - Why not revenue?

- Growth → change in quantity of items being sold over a period of time

Growth Rate = (Final Value – Initial Value) / Initial Value





	order_id <chr>	order_purchase_timestamp <date>	product_id <chr>	price <dbl>	total_price <dbl>	qty <int>	product_category_name_english <chr>
105027	63943bddc261676b46f01ca7ac2f7bd8	2018-02-06	f1d4ce8c6dd66c47bbaa8c6781c2a923	174.90	174.90	1	baby
105028	83c1379a015df1e13d02aae0204711ab	2017-08-27	b80910977a37536adeddd63663f916ad	205.99	205.99	1	home_appliances_2
105029	11c177c8e97725db2631073c19f07b62	2018-01-08	d1c427060a0f73f6b889a5c7c61f2ac4	179.99	179.99	1	computers_accessories
105030	11c177c8e97725db2631073c19f07b62	2018-01-08	d1c427060a0f73f6b889a5c7c61f2ac4	179.99	359.98	2	computers_accessories
105031	66dea50a8b16d9b4dee7af250b4be1a5	2018-03-08	006619bbed68b000c8ba3f8725d5409e	68.50	68.50	1	health_beauty

5 rows

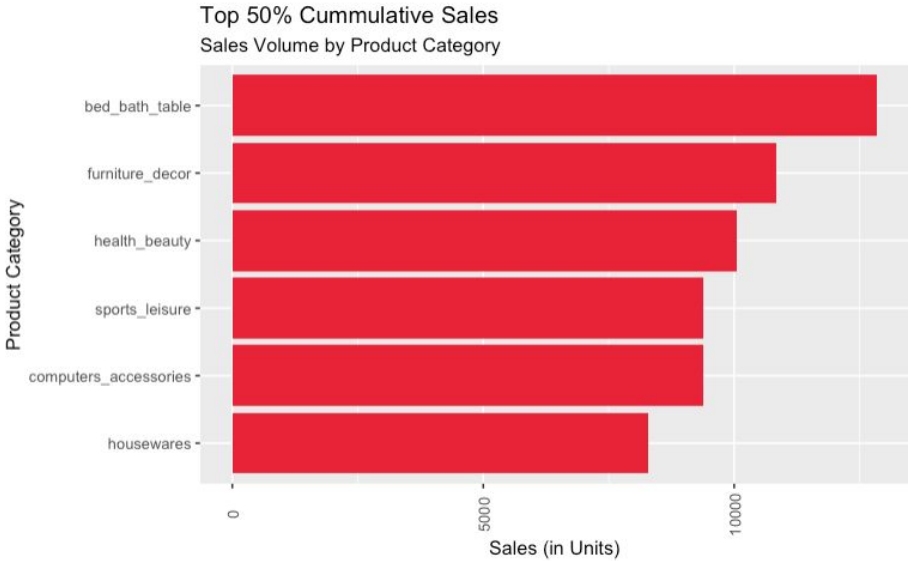
Data Preparation - Dates & Added Variables

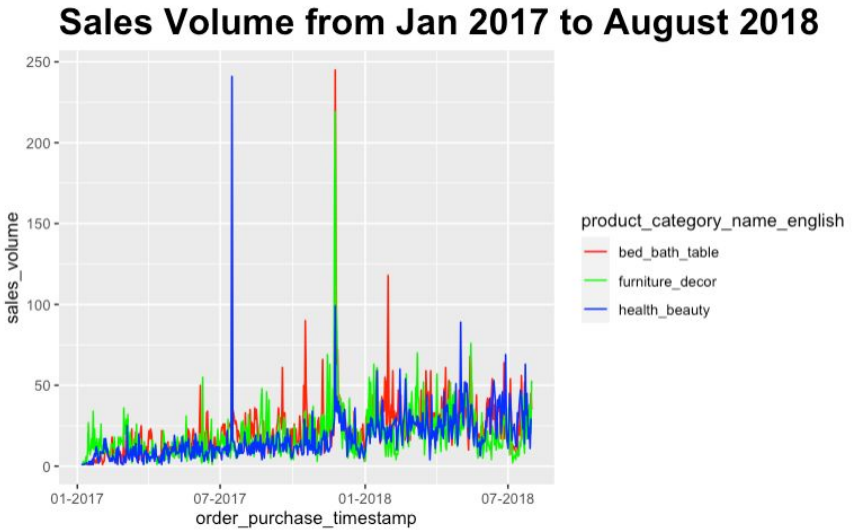
- Cut off dates from 01/01/2017 to 07/31/2018
  - Before 01/01/2017, data not clear or consistent
  - Last date in dataset (09/03/2018) was not accurate for ‘qty’

- Added variables
  - Date → year, month, day of week
  - Weekend → isWeekend
  - Holiday → isBlackFriday

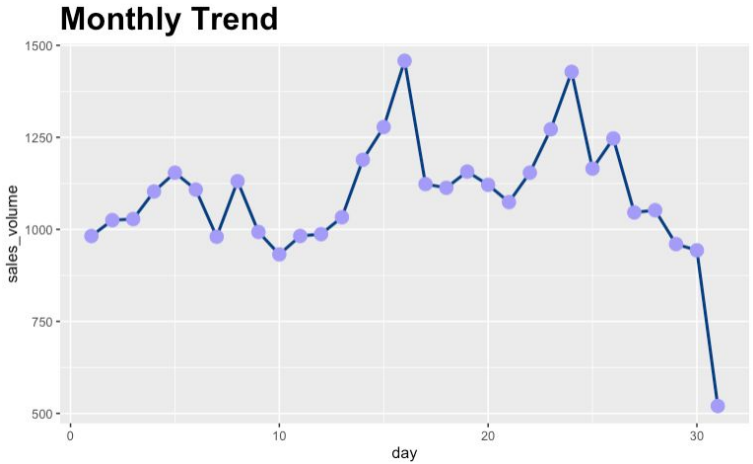
Exploratory Data Analysis

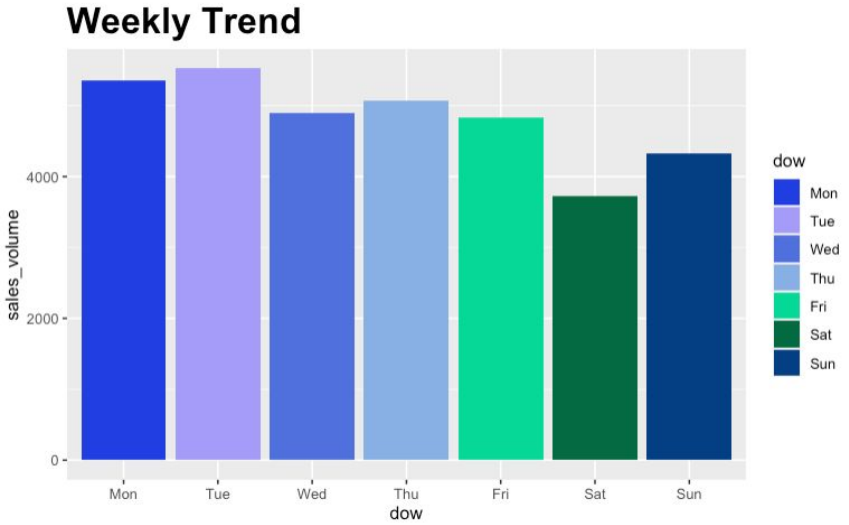
Note: Sales → quantity of items being sold





Time Trends


Monthly Trend
sales_volume
1500
1250
1000
750
500
0
10
20
30
day


Weekly Trend
sales_volume
4000
2000
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
dow
dow
Mon
Tue
Wed
Thu
Fri
Sat
Sun

Modeling Approach

For each goal:

- Created and used **training** and **validation** sets to validate best model simulating the situation in the midterm
  - Training set ranges from 01/01/2017 to 04/31/2018
  - Validation set is last three months of data, ranges from 05/01/2018 to 07/31/2018
- Used **AIC** and **MSE** to select best model
- Predicted `sales_volume` for next three months (08/01/2018 to 10/31/2018), 92 days

For goals 2 & 3:

- Summed predicted `sales_volume` values and used this to calculate growth

Goal 1 - Model Validation & Selection

- Tested OLS, SARIMAX, and Panel models

For Goal 1:

Model	AIC	MSE
TSLM	4320.302	19065.84
SARIMA	5486.460	10501.4
Panel	5345	18563.0008

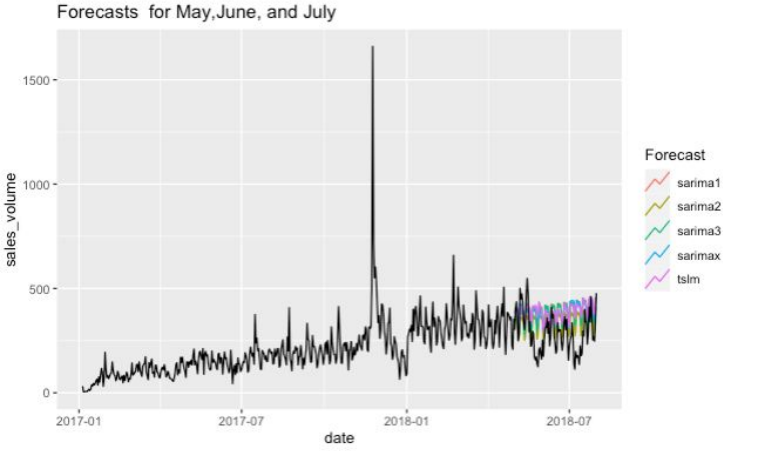
Goal 1: Overall Sales Prediction

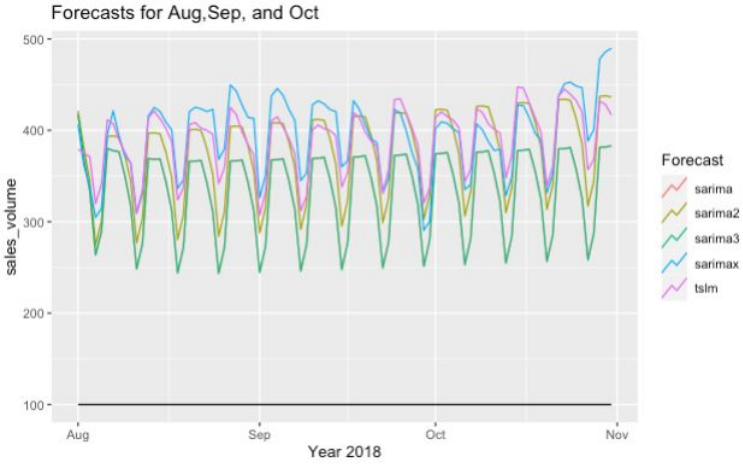
```r
# modeling the Whole Data
alltsb
fit.all <- alltsb[,-c(24,25)]%>%
        model(sarima = ARIMA(sales_volume ~ 0 + pdq(4, 1, 2) + PDQ(0, 1, 1, period = 7)),
              sarima2 = ARIMA(sales_volume ~ 0 + pdq(1, 1, 1) + PDQ(0, 1, 1, period = 7)),
              sarima3 = ARIMA(sales_volume ~ 0 + pdq(4, 1, 2) + PDQ(0, 1, 1, period = 7)),
              sarimax =  ARIMA(sales_volume ~ month + weekday  + day + trend() + fourier(period = "week",1)+
                               fourier(period = "month",3) + fourier(period = "year",5)),
              tslm = TSLM(sales_volume ~ month + trend() + season("week") + fourier(period = "month",3)))
fit.all %>% report()
```

product_category <chr>	.model <chr>	sigma2 <dbl>	log_lik <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>
all	sarima1	5997.477	-2738.230	5486.460	5486.588	5507.255
all	sarima2	6175.068	-2742.496	5492.992	5493.078	5509.629
all	sarima3	5975.707	-2735.761	5487.523	5487.833	5520.796
all	sarimax	5546.951	-2737.832	5547.664	5553.664	5697.995
all	tslm	7534.013	-2816.660	4320.302	4323.394	4428.875

product_category <chr>	.model <chr>	sigma2 <dbl>	log_lik <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>
all	sarima	5616.969	-3247.311	6510.622	6510.881	6545.317
all	sarima2	5756.978	-3254.948	6517.896	6517.967	6535.243
all	sarima3	5616.969	-3247.311	6510.622	6510.881	6545.317
all	sarimax	5316.136	-3253.520	6577.040	6581.733	6729.321
all	tslm	8077.399	-3377.861	5181.619	5184.191	5294.742

5 rows | 1–9 of 18 columns

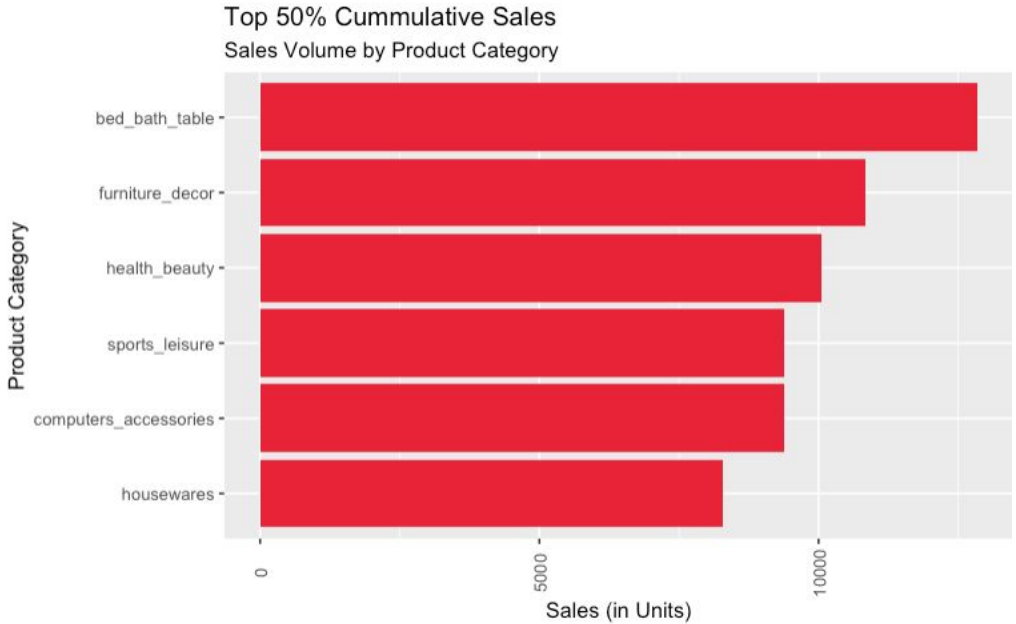




Goal 2: Determination of Best-Selling Categories

Summed the quantity of sales for each category

1. Bed_bath_table
2. Furniture_decor
3. Health_beauty



Goal 2 - Model Validation & Selection

- Reminder: these values correspond to the prediction of `sales_volume` for the next three months, not growth

For Goal 2 (bed_bath_table):

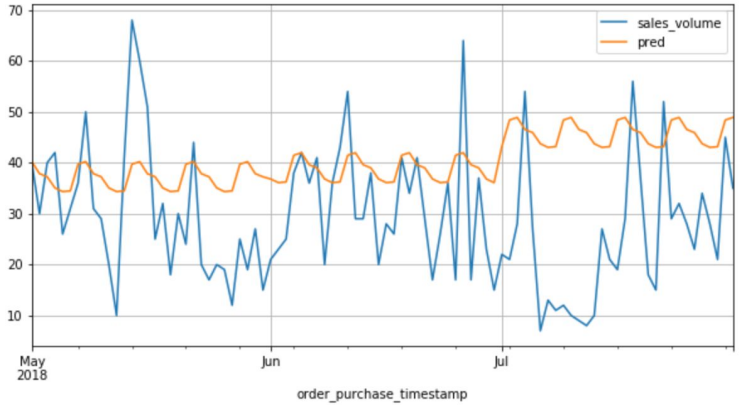
Model	AIC	MSE
OLS	3742	322.7053
SARIMAX	3900.031	182.1917
Panel	3663	319.8653

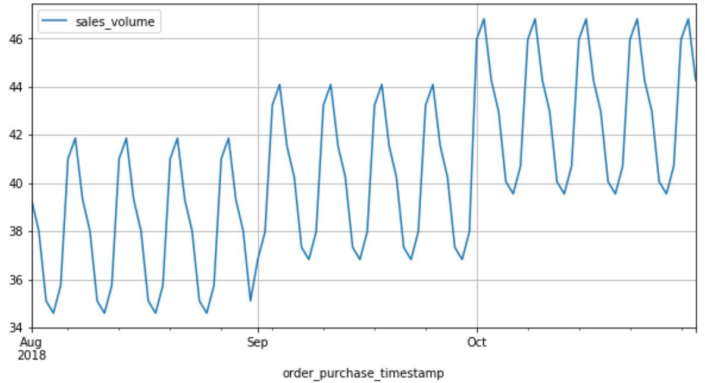
Goal 2 - Bed_bath_table Sales Prediction

```
"sales_volume ~ C(year) + C(month) + C(weekday) + isBlackFriday + isWeekend"
```

	order_purchase_timestamp	sales_volume
0	2018-08-01	39.313277
1	2018-08-02	38.017004
2	2018-08-03	35.103271
3	2018-08-04	34.595708
4	2018-08-05	35.748980
...	...	...
87	2018-10-27	39.549617
88	2018-10-28	40.702890
89	2018-10-29	45.964006
90	2018-10-30	46.819891
91	2018-10-31	44.267187

92 rows × 2 columns





Goal 2 - Model Validation & Selection

- Used AIC & MSE to determine which model performed the best on the validation set

For Goal 2 (furniture_decor):

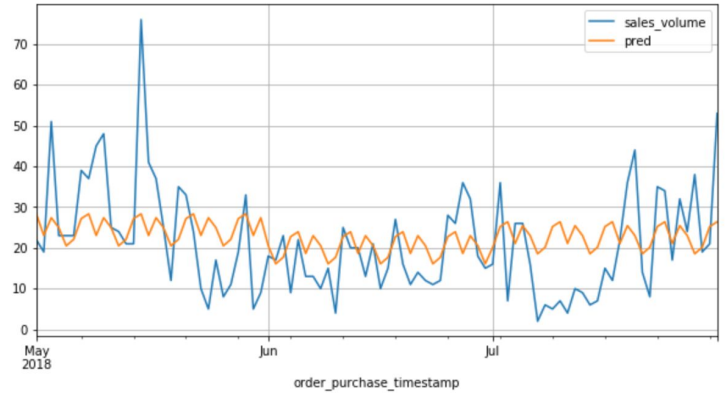
Model	AIC	MSE
OLS	3825	189.5485
SARIMAX	3885.655	192.2151
Panel	3811	187.5343

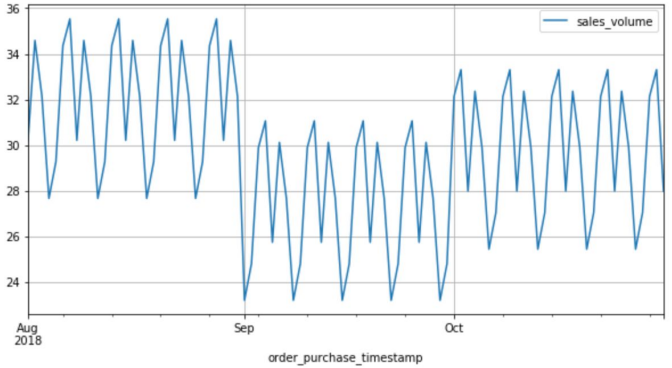
Goal 2 - Furniture_decor Sales Prediction

```
"sales_volume ~ C(year) + C(month) + C(weekday) + isBlackFriday + isWeekend"
```

	order_purchase_timestamp	sales_volume
0	2018-08-01	30.217929
1	2018-08-02	34.586143
2	2018-08-03	32.145882
3	2018-08-04	27.676941
4	2018-08-05	29.279667
...	...	...
87	2018-10-27	25.456469
88	2018-10-28	27.059195
89	2018-10-29	32.138574
90	2018-10-30	33.309538
91	2018-10-31	27.997457

92 rows × 2 columns





Goal 2 - Model Validation & Selection

- Used AIC & MSE to determine which model performed the best on the validation set

For Goal 2 (health_beauty):

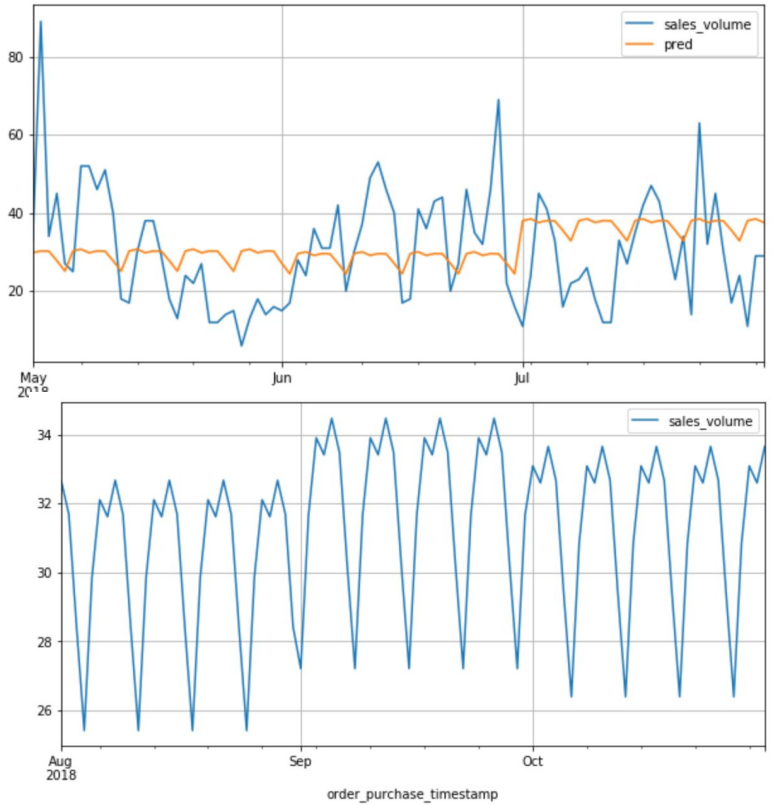
Model	AIC	MSE
OLS	3756	221.7046
SARIMAX	3806.792	222.2342
Panel	3756	221.7046

Goal 2 - Health_beauty Sales Prediction

```
"sales_volume ~ C(year) + C(month) + C(weekday) + isBlackFriday + isWeekend"
```

	order_purchase_timestamp	sales_volume
0	2018-08-01	32.675281
1	2018-08-02	31.692351
2	2018-08-03	28.430882
3	2018-08-04	25.418194
4	2018-08-05	29.871244
...	...	...
87	2018-10-27	26.398398
88	2018-10-28	30.851448
89	2018-10-29	33.090267
90	2018-10-30	32.598665
91	2018-10-31	33.655485

92 rows × 2 columns



Growth for Goal 2:

We defined “growth” as: **Growth Rate = (Final Value – Initial Value) / Initial Value**

1. Bed bath table

product_category_name_english <chr>	sales_volume_2017 <int>	sales_volume_2018 <dbl>	growth_rate <dbl>
bed_bath_table	1998	3713	0.8583584

1 row

2. Furniture_decor

product_category_name_english <chr>	sales_volume_2017 <int>	sales_volume_2018 <dbl>	growth_rate <dbl>
furniture_decor	1613	2737	0.6968382

1 row

3. health_beauty

product_category_name_english <chr>	sales_volume_2017 <int>	sales_volume_2018 <dbl>	growth_rate <dbl>
health_beauty	1189	2871	1.414634

1 row

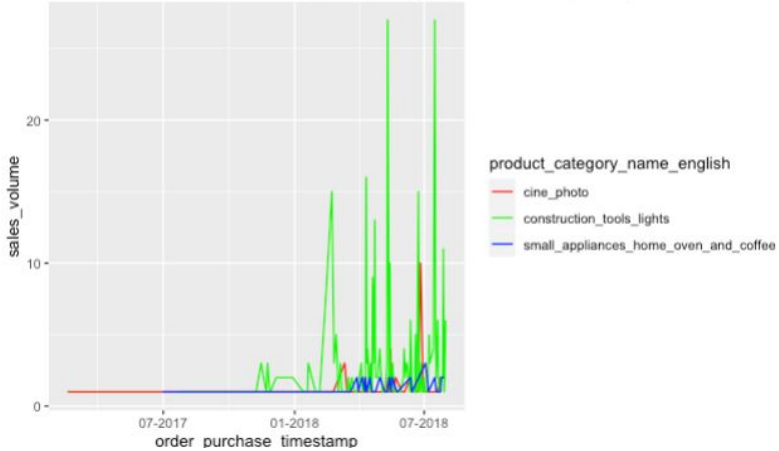
Goal 3: Determination of Fastest-Growing Category

Since this problem very dependent in seasonality and each has different factors that increases the sales during that month. We decided to compare may , june and july sales form 2017 to the same months in 2018 to see what is the growth rate between these periods. Note, we are using the simple growth rate in this case. Growth in sales and not in revenue

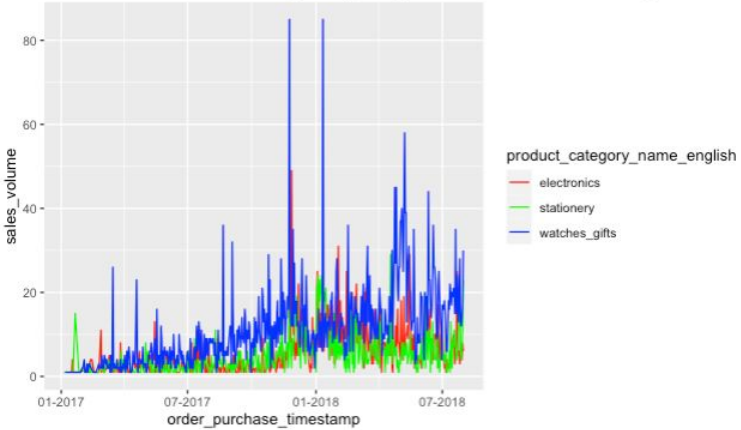
Growth Rate = (Final Value – Initial Value) / Initial Value

product_category_name_english <chr>	sales_volume_2017 <int>	sales_volume_2018 <int>	growth_rate <dbl>
construction_tools_lights	1	206	205.00000000
small_appliances_home_oven_and_coffee	1	28	27.00000000
cine_photo	2	46	22.00000000
drinks	6	127	20.16666667
construction_tools_construction	24	469	18.54166667
industry_commerce_and_business	6	97	15.16666667
books_technical	7	111	14.85714286

Sales Volume for the fastest-growing categories with the highest growth rate betwee



Sales Volume for the fastest-growing categories that has a min of selling 100 items



Goal 3 - Model Validation & Selection

- Used AIC & MSE to determine which model performed the best on the validation set

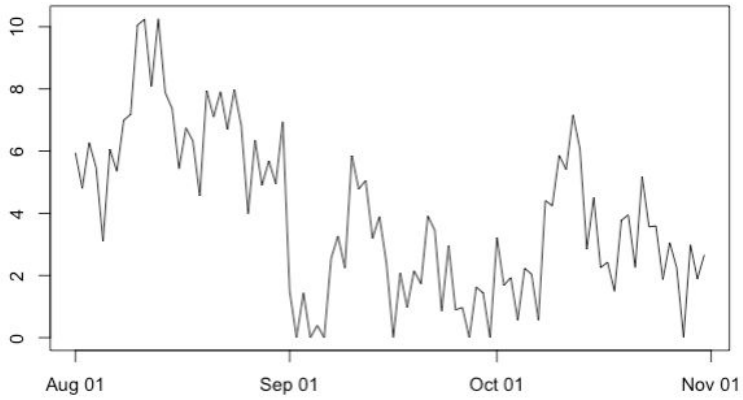
For Goal 3:

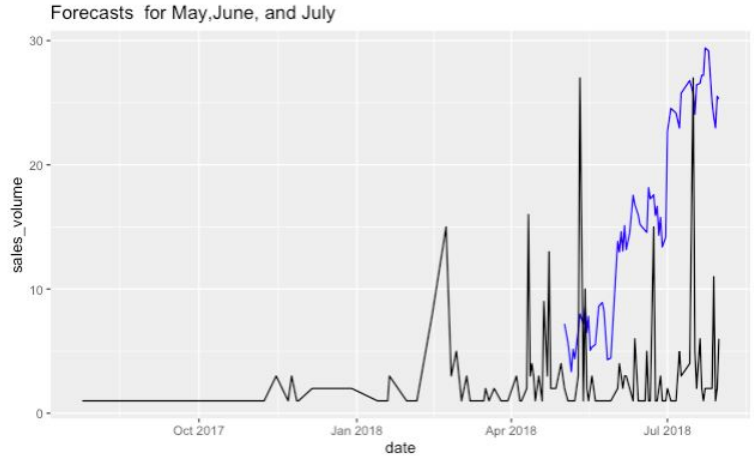
Model	AIC	MSE
TSLM in R	144.4293	#######
SARIMAX	289.796	27.5508
Panel	292.4	28.5889

Goal 3: Construction_tools_lights Sales Prediction

```
# whole_data model

fit.construction.all <- alltsb%>%
        model(
                tslm = TSLM(sales_volume ~ month + trend() + season("week") + fourier(period = "month",3)))
fit.construction.all %>% report()
```





Growth for Goal 3:

We defined “growth” as: **Growth Rate = (Final Value – Initial Value) / Initial Value**

1. construction_tools_lights

product_category_name_english <chr>	sales_volume_2017 <int>	sales_volume_2018 <dbl>	growth_rate <dbl>
construction_tools_lights	5	361.0567	71.21134

Final Thoughts:

- On **Black Friday**, the volume of sales more than **tripled**

  → we recommend that OLIST use this model to plan future inventory and coordinate work schedules

- Addition of holiday calendar
  - Anticipate increased demand