# Olist

**Technical Candidate: Ian Yu**

# Agenda

◆ **Overview**

◆ **Data Wrangling**

◆ **EDA**

◆ **Post-Sales Analysis**

◆ **Future Development**

# My Data
# Philosophy

Context-Centric

Scalable Solution

Post-Sales Analysis

# My
# Deliverables

Module data.py  and Public S3

EDA Notebook

GitHub

manipulates Olist datasets. Default

bucket = "ianyu-public-data"):

cket

urce('s3')
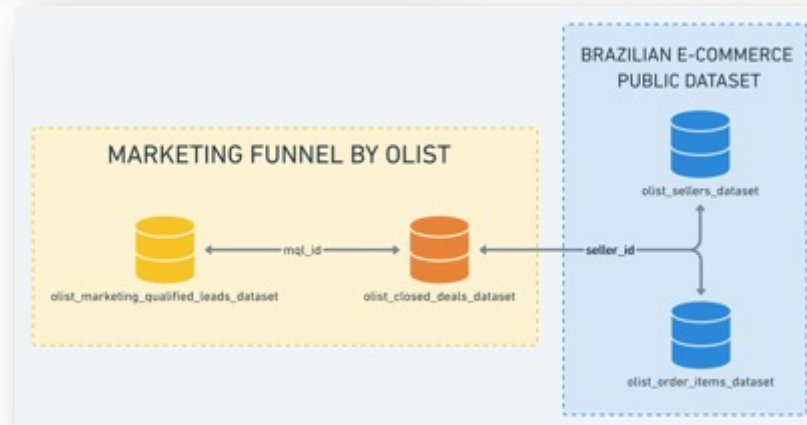
"oli olist"):

lable datase

# Olist
# Overview

Marketplace of Marketplaces

Customer-Led Growth

Post-Sales Analysis

# Olist
## Funnel

Only **1** Conversion Stage

Missing Much Information

SDR Qualifies Leads

SR Closes Deals

Reach

Landing Page

Sign-Up

Consultancy

- Makret Qualified Leads

- Closed Deals

# Data Wrangling

# Much Data is Missing

Features to Remember:

Lead Behaviour Profile is Internal Data

Declared Monthly Revenue

| | |
|---|---|
| mql_id | 0.000000 |
| seller_id | 0.000000 |
| sdr_id | 0.000000 |
| sr_id | 0.000000 |
| won_date | 0.000000 |
| business_segment | 0.118765 |
| lead_type | 0.712589 |
| lead_behaviour_profile | 21.021378 |
| has_company | 92.517815 |
| has_gtin | 92.399050 |
| average_stock | 92.161520 |
| business_type | 1.187648 |
| declared_product_catalog_size | 91.805226 |
| declared_monthly_revenue | 0.000000 |
| dtype: float64 | |

**Top:** Table Missing Values (%)

**Right:** Declared Monthly Revenue Count

| | |
|---|---|
| 0.0 | 94.655582 |
| 100000.0 | 0.593824 |
| 20000.0 | 0.356295 |
| 25000.0 | 0.356295 |
| 10000.0 | 0.356295 |
| 30000.0 | 0.356295 |
| 120000.0 | 0.237530 |
| 5000.0 | 0.237530 |
| 250000.0 | 0.237530 |
| 50000.0 | 0.237530 |
| 300000.0 | 0.237530 |
| 15000.0 | 0.237530 |
| 60000.0 | 0.237530 |
| 1000.0 | 0.118765 |
| 8000.0 | 0.118765 |
| 4000.0 | 0.118765 |
| 6.0 | 0.118765 |
| 180000.0 | 0.118765 |
| 50000000.0 | 0.118765 |
| 8000000.0 | 0.118765 |
| 200000.0 | 0.118765 |
| 210000.0 | 0.118765 |
| 150000.0 | 0.118765 |
| 130000.0 | 0.118765 |
| 500000.0 | 0.118765 |
| 6000.0 | 0.118765 |
| 40000.0 | 0.118765 |

Name: declared_monthly_rev

# Lead
# Generation

Search accounts for 48.52% of Lead Generation

Successful SEO or Word of Mouth

Unclear Why Social Was Successful!

| origin | leads generated |
|---|---|
| organic_search | 28.70% |
| paid_search | 19.82% |
| social | 16.87% |
| other | 16.36% |
| direct_traffic | 6.23% |
| email | 6.16% |
| referral | 3.55% |
| display | 1.47% |
| other_publicities | 0.81% |

# Lead Behaviour
# Profile

Total Missing Values: **20%**

Can't Do Predictive Modeling

High Bias If Used

3 SDRs Keep Missing Data



Cats Behaviour Dominates the Closed Deals
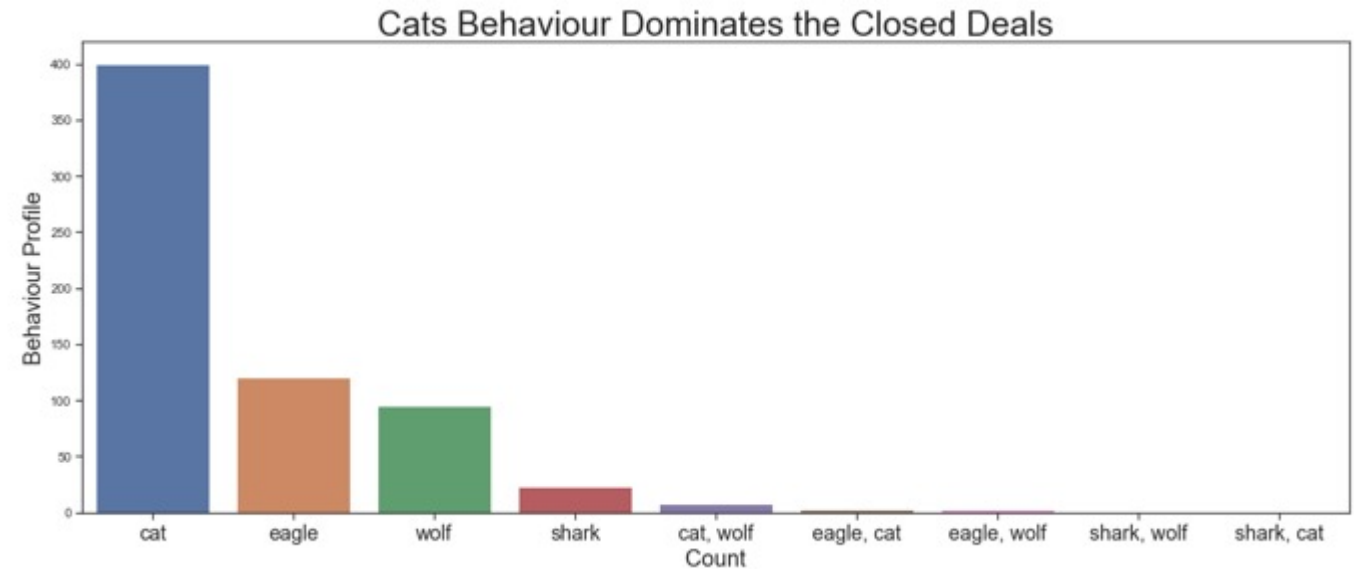
| sdr_id | total closed | total missing profile | percentage |
|---|---|---|---|
| 56bf83c4bb35763a51c2baab501b4c67 | 74 | 34 | 45.95% |
| fdb16d3cbbeb5798f2f66c4096be026d | 34 | 27 | 79.41% |
| de63de0d10a6012430098db33c679b0b' | 53 | 41 | 77.36% |

**Top:** Lead Behaviour Profile count

**Bottom:** SDRs who keep fail at indicating Behaviour Profile

# So What?

Dive Deeper Into Social Activities

Why the 3 SDRs Are Not Able to Indicate Lead Behaviour Profile

Classify Declared Monthly Revenue In Broad Ranges
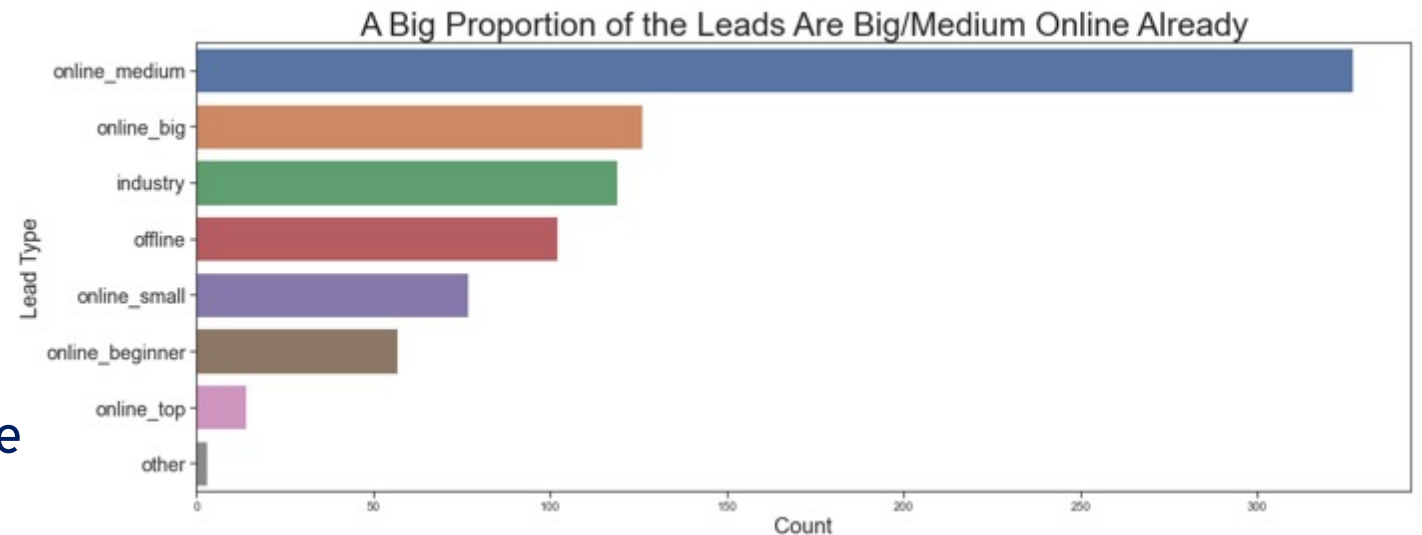
# Exploratory Data Analysis

# Lead
## Type

Big/Medium Online: 55%

They Should Know Their Revenue
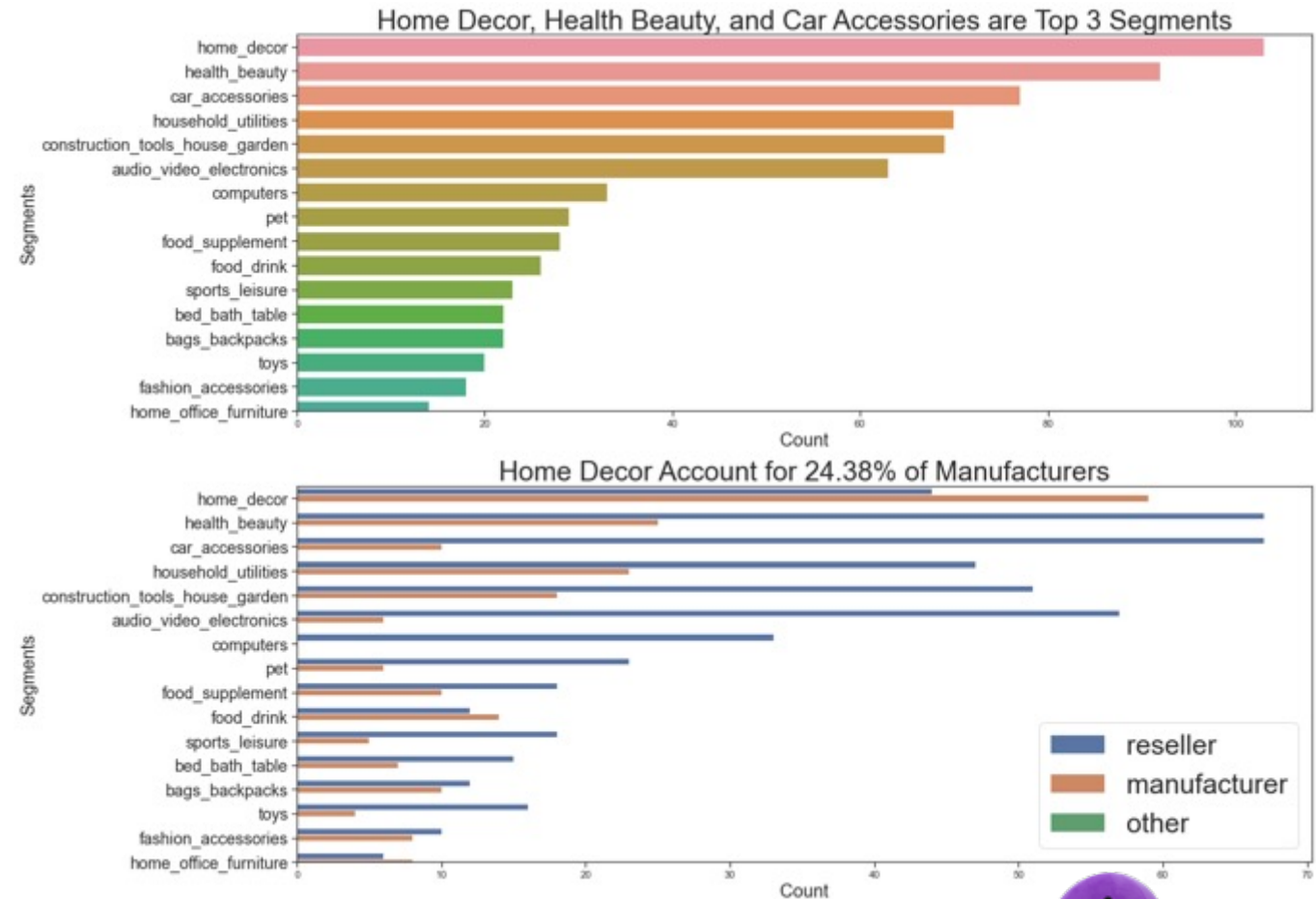
Why Declared Monthly Revenue Has 94% of Missing Values?



A Big Proportion of the Leads Are Big/Medium Online Already

# Business
# Segment

30% Manufacturers

70% Resellers

¼ of Manufacturers are Home Décor



Home Decor, Health Beauty, and Car Accessories are Top 3 Segments



Home Decor Account for 24.38% of Manufacturers

# Lead-to-Closed Origin

Search: 48% >>> 55%

Social Dropped in Proportion: 7%

Could Be a Costly Origin!

```
Marketing Qualified Leads Origin   Closed Deals Media Origin

organic_search        28.7000      organic_search        32.121212
paid_search           19.8250      paid_search           23.393939
social                16.8750      other                 23.272727
other                 16.3625      social                 9.090909
direct_traffic         6.2375      direct_traffic         6.545455
email                  6.1625      referral               2.909091
referral               3.5500      email                  1.696970
display                1.4750      display                0.606061
other_publicities      0.8125      other_publicities      0.363636
Name: origin, dtype: float64      Name: origin, dtype: float64
```

```
Percentage Difference

direct_traffic        0.307955
display              -0.868939
email                -4.465530
organic_search        3.421212
other                 6.910227
other_publicities    -0.448864
paid_search           3.568939
referral             -0.640909
social               -7.784091
Name: origin, dtype: float64
```
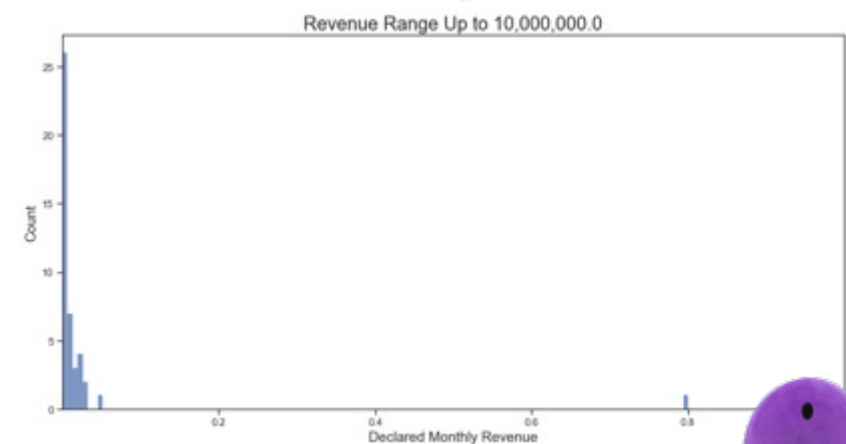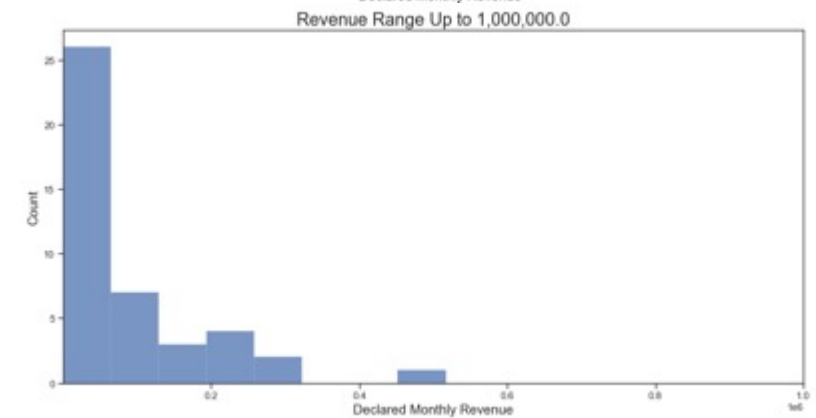
# Declared Monthly Revenue **Ranges**

45 Samples out of 842

Most Are Under 60,000R

| Ranges | Class |
| --- | --- |
| 0-50,000R | Minor |
| 50,000-300,000R | Small |
| 300,000R+ | Medium |



Revenue Range Up to 100,000.0



Revenue Range Up to 1,000,000.0



Revenue Range Up to 10,000,000.0

# Sales
# Duration

MQL-Closed: ~10%

1/3 of the Deals Closed on the 7th Day

Momentum Forever Below 1 on 23rd day



Half of the Deals Were Closed In 14 Days

- Deals Closed by nth Day
- Rate of Increase for n-1 Day

Number of Days Between First Contact Day to Deal Closed

# Sales
# Duration

Top 5 Average: 22 Days

Overall Average: 105 Days

Super Sales: Number 6

| sr_id | declared_monthly_revenue | lead_to_close |
|---|---|---|
| 9e4d1098a3b0f5da39b0bc48f9876645 | 0.000000 | 17.000000 |
| fbf4aef3f6915dc0c3c97d6812522f6a | 0.000000 | 21.966102 |
| c638112b43f1d1b86dcabb0da720c901 | 0.000000 | 22.727273 |
| 060c0a26f19f4d66b42e0d8796688490 | 0.000000 | 25.218750 |
| 6565aa9ce3178a5caf6171827af3a9ba | 0.000000 | 26.013699 |
| 4ef15afb4b2723d8f3d81e51ec7afefe | 383053.435115 | 30.183206 |
| 85fc447d336637ba1df43e793199fbc8 | 1562.500000 | 33.703125 |
| 9ae085775a198122c5586fa830ff7f2b | 0.000000 | 41.489796 |
| 2695de1affa7750089c0455f8ce27021 | 1754.385965 | 44.596491 |
| d3d1e91a157ea7f90548eef82f1955e3 | 5062.500000 | 47.625000 |
| 495d4e95a8cf8bbf8b432b612a2aa328 | 7000.000000 | 49.850000 |
| 56bf83c4bb35763a51c2baab501b4c67 | 12500.000000 | 55.125000 |
| de63de0d10a6012430098db33c679b0b | 154326.923077 | 62.884615 |
| a8387c01a09e99ce014107505b92388c | 24038.692308 | 96.653846 |
| 068066e24f0c643eb1d089c7dd20cd73 | 7037.037037 | 122.333333 |
| 34d40cdaf94010a1d05b0d6212f9e909 | 19000.000000 | 172.000000 |
| b90f87164b5f8c2cfa5c8572834dbe3f | 0.000000 | 175.000000 |
| 4b339f9567d060bcea4f5136b9f5949e | 31250.000000 | 194.500000 |
| 9d12ef1a7eca3ec58c545c678af7869c | 43333.333333 | 214.333333 |
| 9749123c950bf8363ace42cb1c2d0815 | 103571.428571 | 234.000000 |
| 0a0fb2b07d841f84fb6714e35c723075 | 6000.000000 | 306.000000 |
| 6aa3b86a83d784b05f0e37e26b20860d | 8000.000000 | 321.000000 |

# So **What?**

Explore SDRs Are Not Recording Declared Monthly Revenue

Home Décor Manufacturers May Present Opportunities

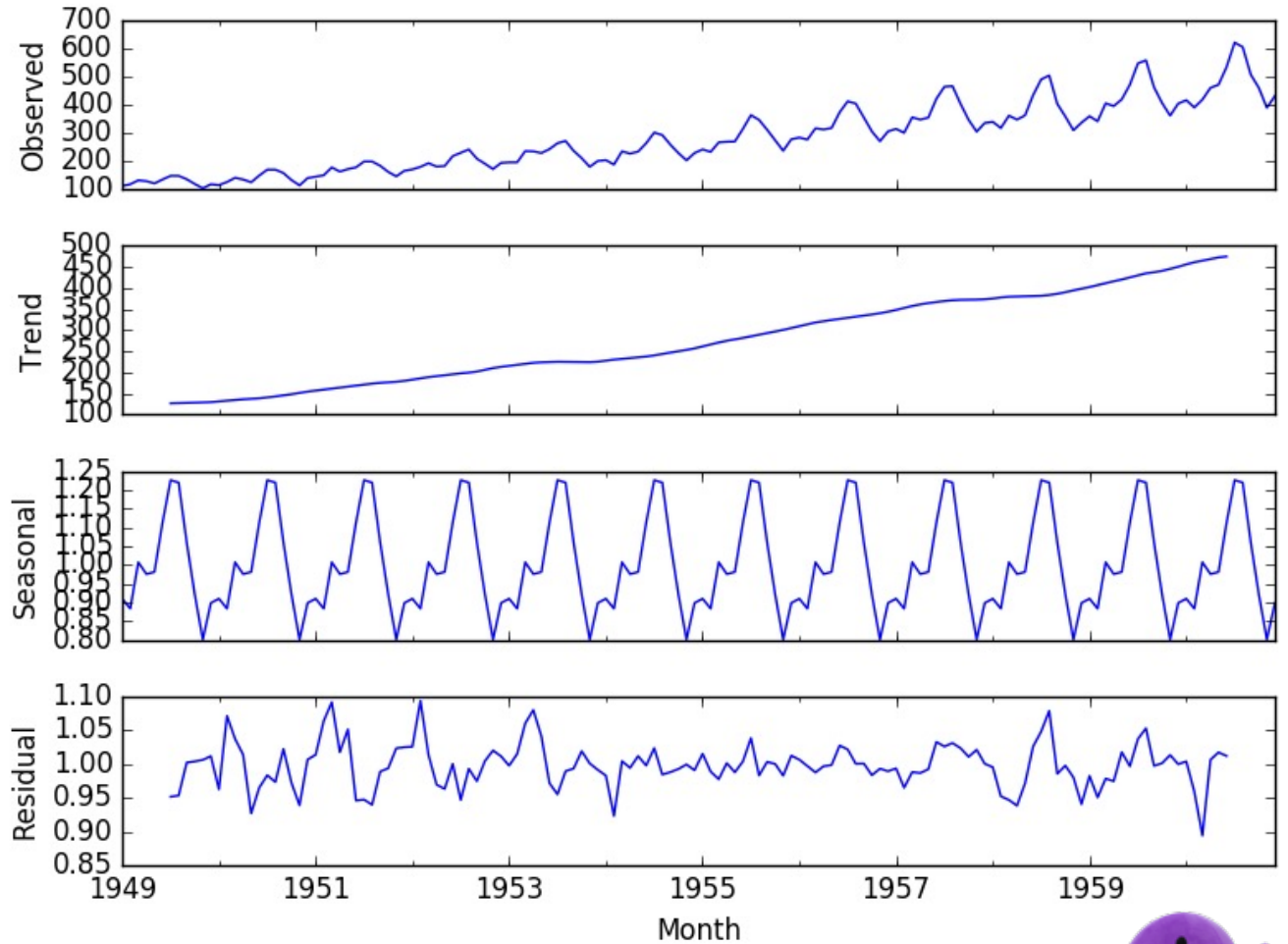Need Data on Lost Deals

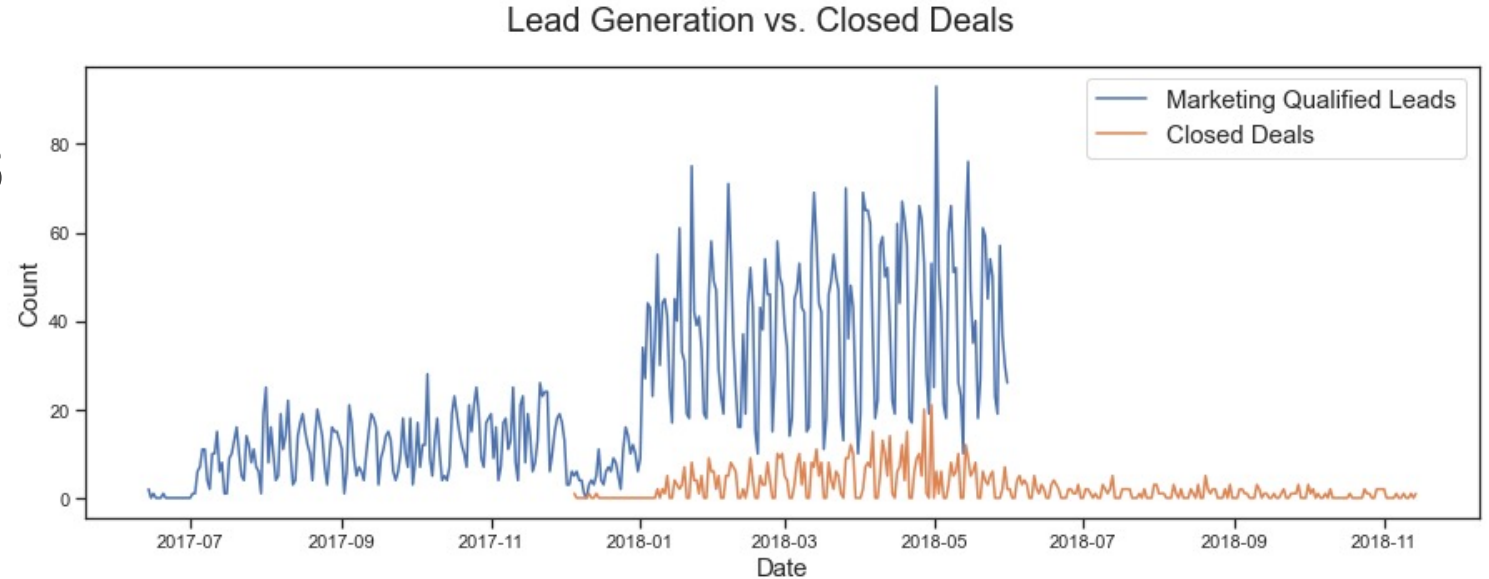# Time Series Analysis

# What Is
## TSA?

Trend

Seasonality

Noise/Residual

# Is Closed Deals
## Stationary?

Lead Generation vs. Closed Deals



Lead Generation is Not Stationary

Closed Deals Became Stationary by Mid June

```
From May
ADF Statistic: -1.684110
p-value: 0.439318

From June
ADF Statistic: -2.834552
p-value: 0.053505

From July
ADF Statistic: -3.511988
p-value: 0.007684
```
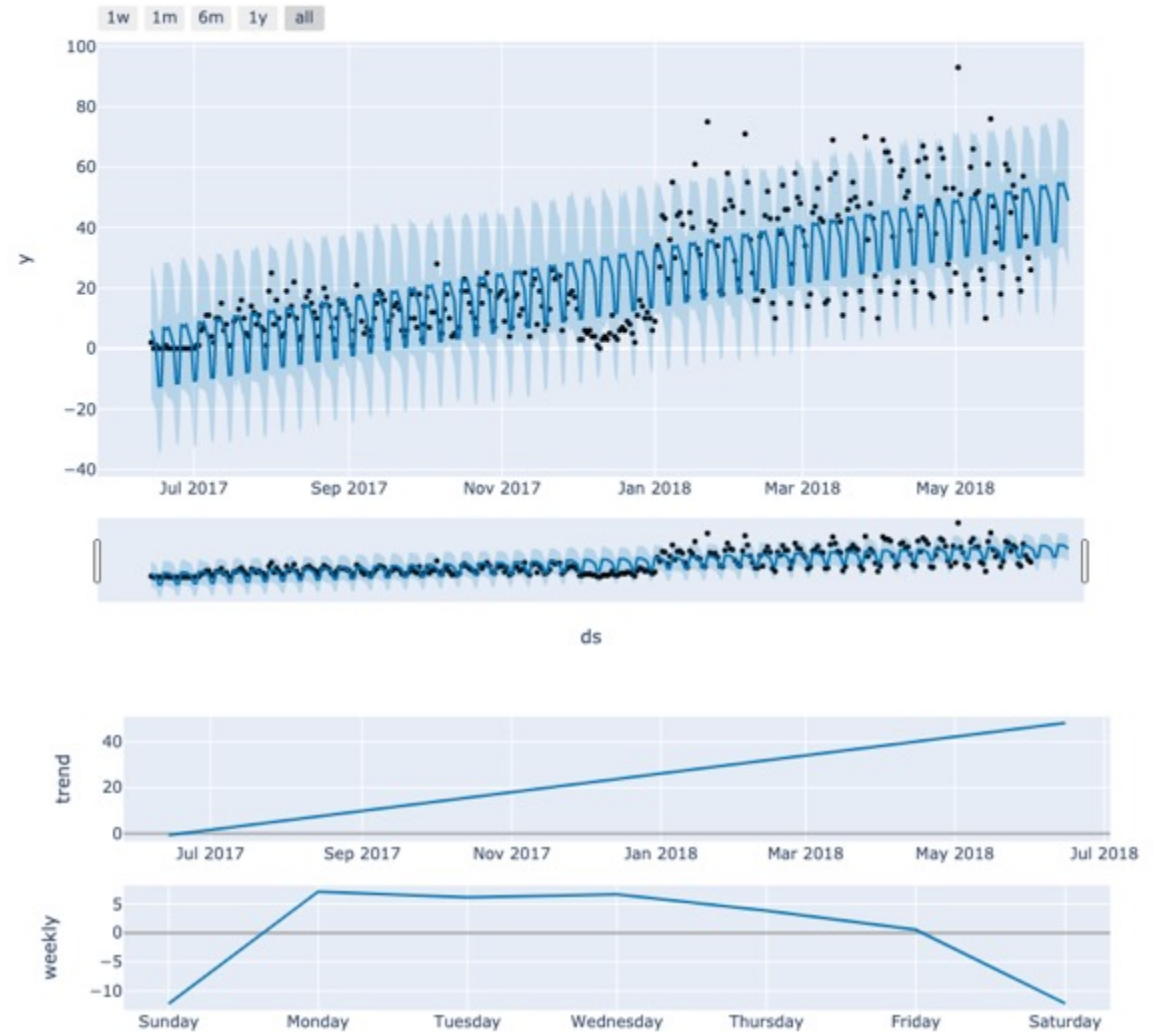
# FBProphet
## MQL

General Uptrend

Monday Generates Leads
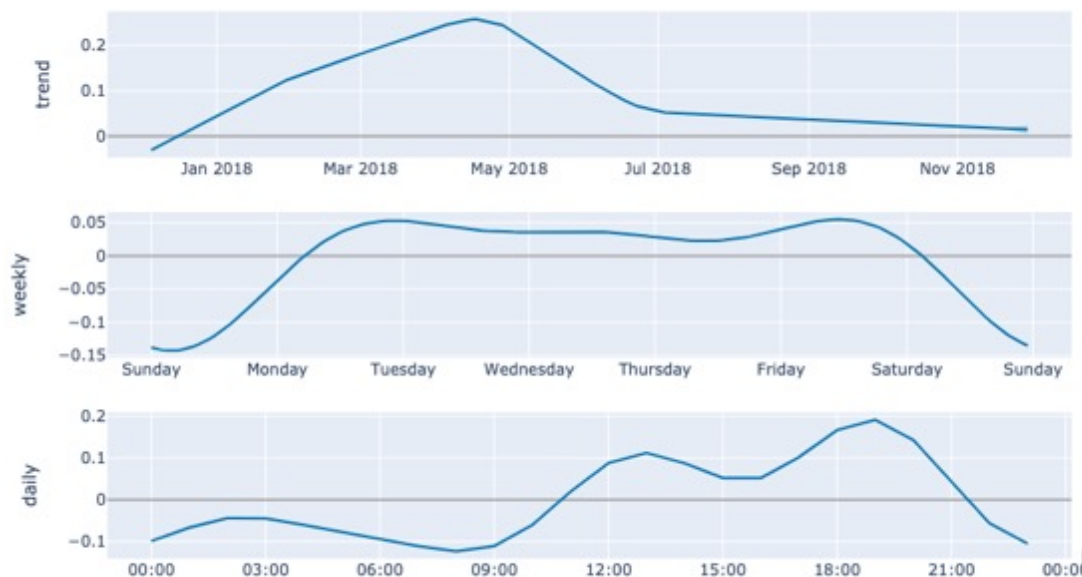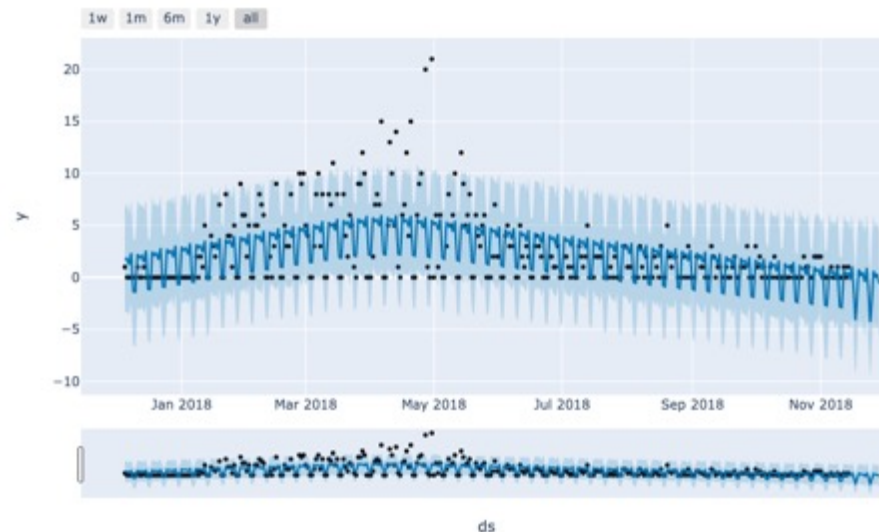
Can't Determine Yearly Seasonality

# FBProphet
## Closed Deals



Humbling Trend

Start and End of Week Close Deals

Lunch / Dinner Time

# So **What?**

Need At Least 2 Years Worth of Data

Q: Is Monday Generating Sales Through Social?

Explore Validity of Closing on Meal-Time

# Post-Sales Analysis

# Post-Sales Analysis

Single Seller Analysis

Flexible to Individual or Holistic

Tools to Follow-up with Customers

```python
##############################################################
# Post-Sales Analysis: Aims to be Developed Into a Post-Sales Class #
##############################################################


def single_seller(ps, feature='price', seller_id = None, resample = '1D', function = 'cumsum'):
    '''
    A custom function that helps examine a seller's performance over time
    '''
    if seller_id == None:
        seller = random.choice([x for x in ps['seller_id']])
    else:
        seller = seller_id


    sum_ = pd.DataFrame(ps[ps['seller_id']==seller][feature].resample(resample).sum())

    if function == 'daily_sum':
        return sum_

    elif function == 'daily_mean':
        return sum_.mean()

    elif function == 'cumsum':
        return sum_.cumsum()

    elif function == 'total_growth':
        start = float(sum_.cumsum().iloc[0])
        end = float(sum_.cumsum().iloc[-1])
        return ((end - start) / start)

    else:
        raise Exception("Invalid function")
```
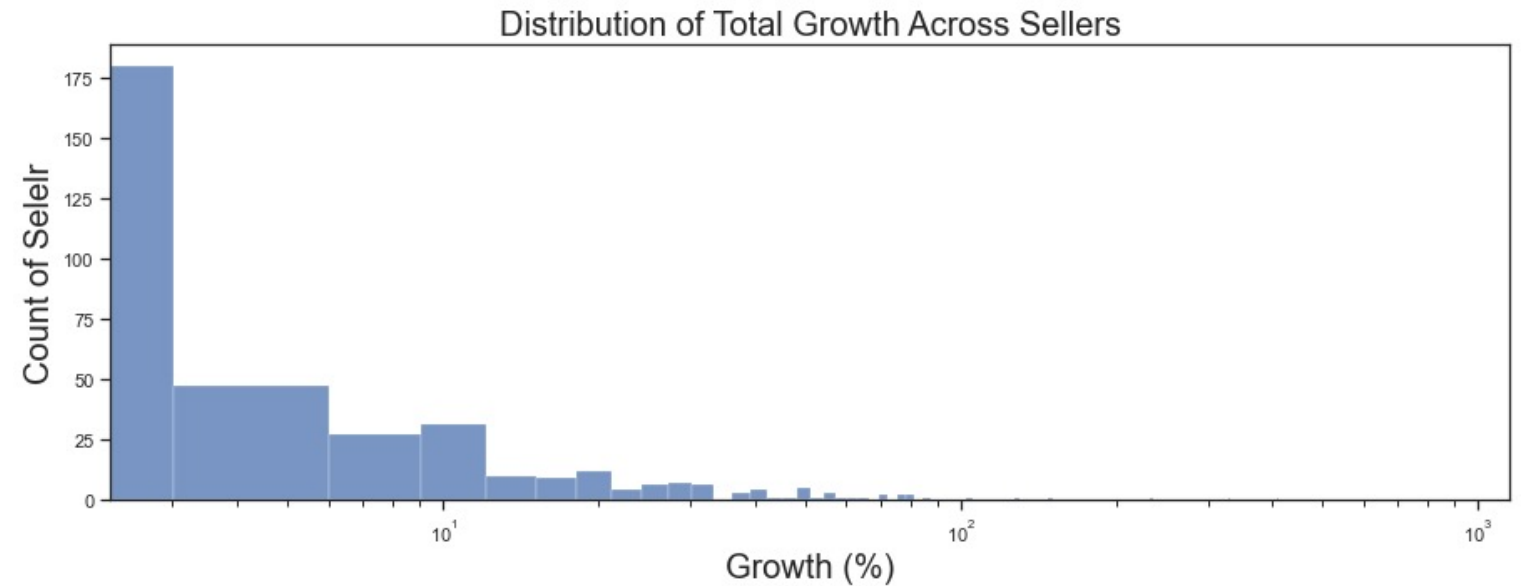
# Aggregate Analysis

In the given 216 Days

70% of the Sellers

Never Reached 10% Total Growth



Distribution of Total Growth Across Sellers

# Future Development

# Monitor **Metrics**

| Metrics | Why? |
|---------|------|
| MQL-Closed Conversion | Will it drop? (Recall TSA) |
| Average Closing Duration | Sales Representative Efficiency |
| Missing Lead Behaviour Profile | SDR Efficiency |
| Converting Proportion from Social | Potential Marketing Cost Dead Weight |

# Other **Development**

Develop data.py

Create a Knowledge Base & Wiki

More Statistical Analysis and Time Series Techniques

Machine Learning (Future)