Flexport

Data Analyst Assignment

Ravi Dayabhai
August 30, 2019

Contents

Motivations & Context

Flexport is **growing fast**,...

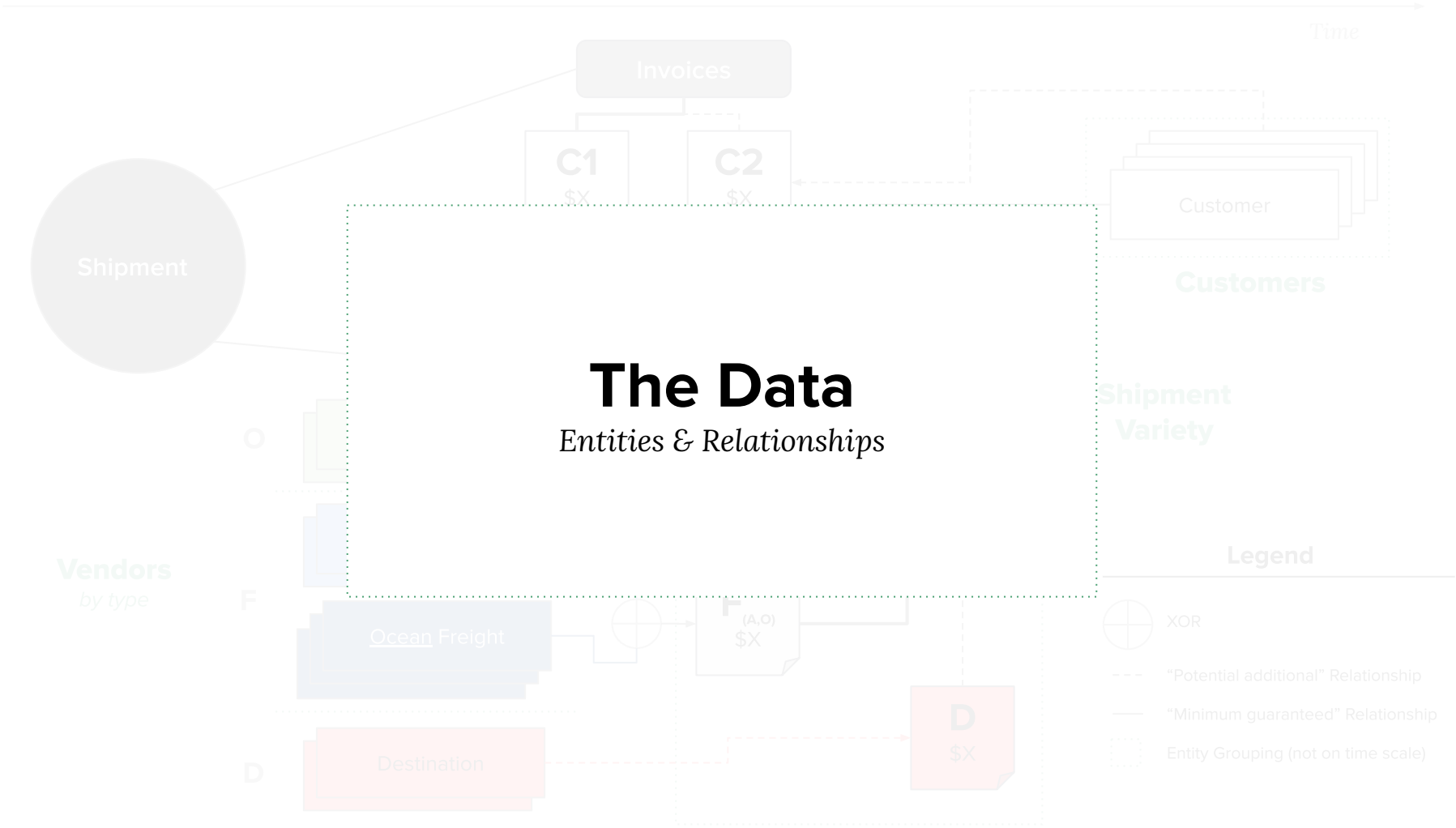
...**providing more services** in new and old product lines...

...generating **increased invoicing** [more clients] and **billing** [by vendors] for shipments.

...and creating need to **understand timeliness of invoices and bills** & get **visibi into our financials**.

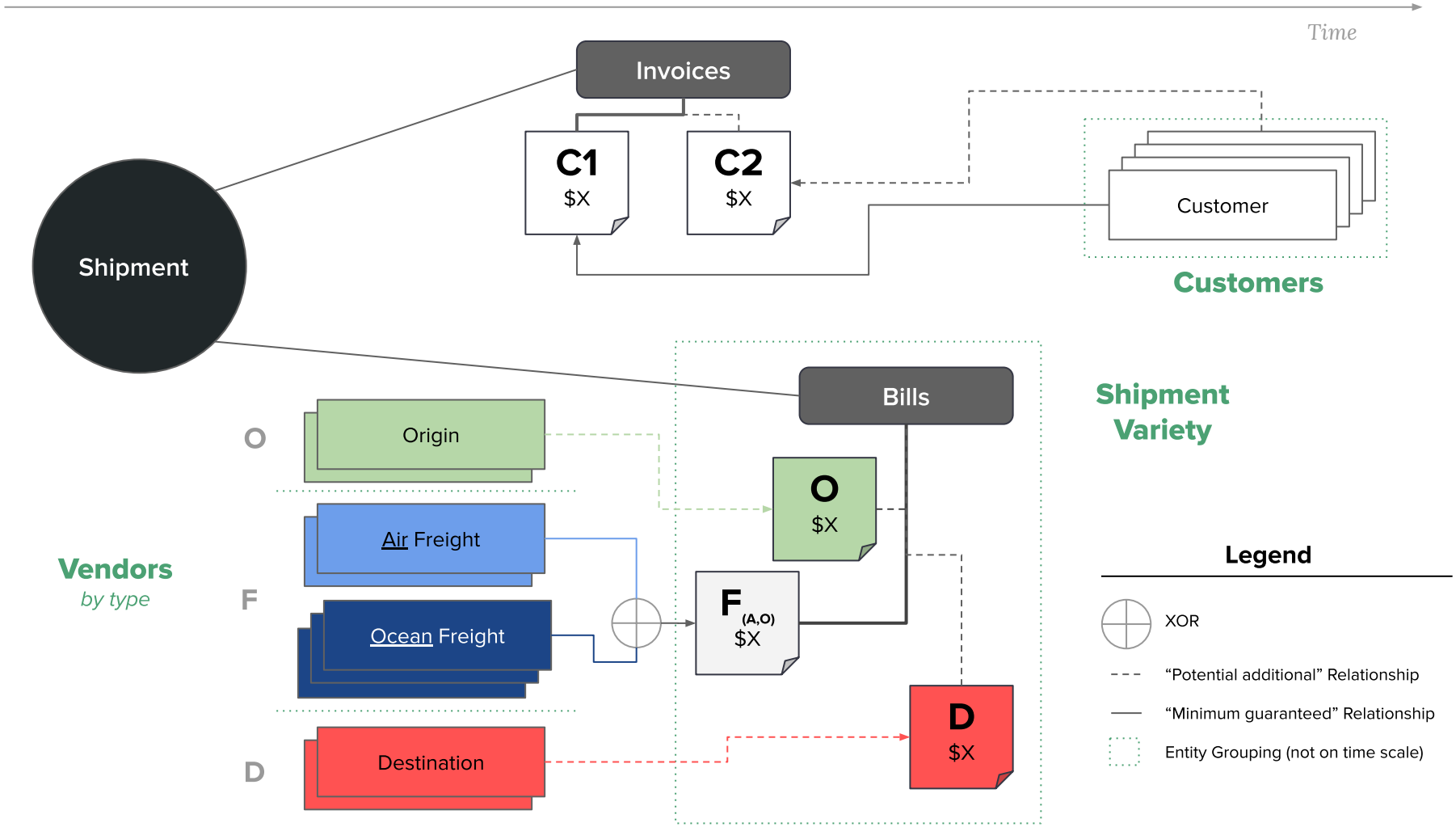
OPS

FIN

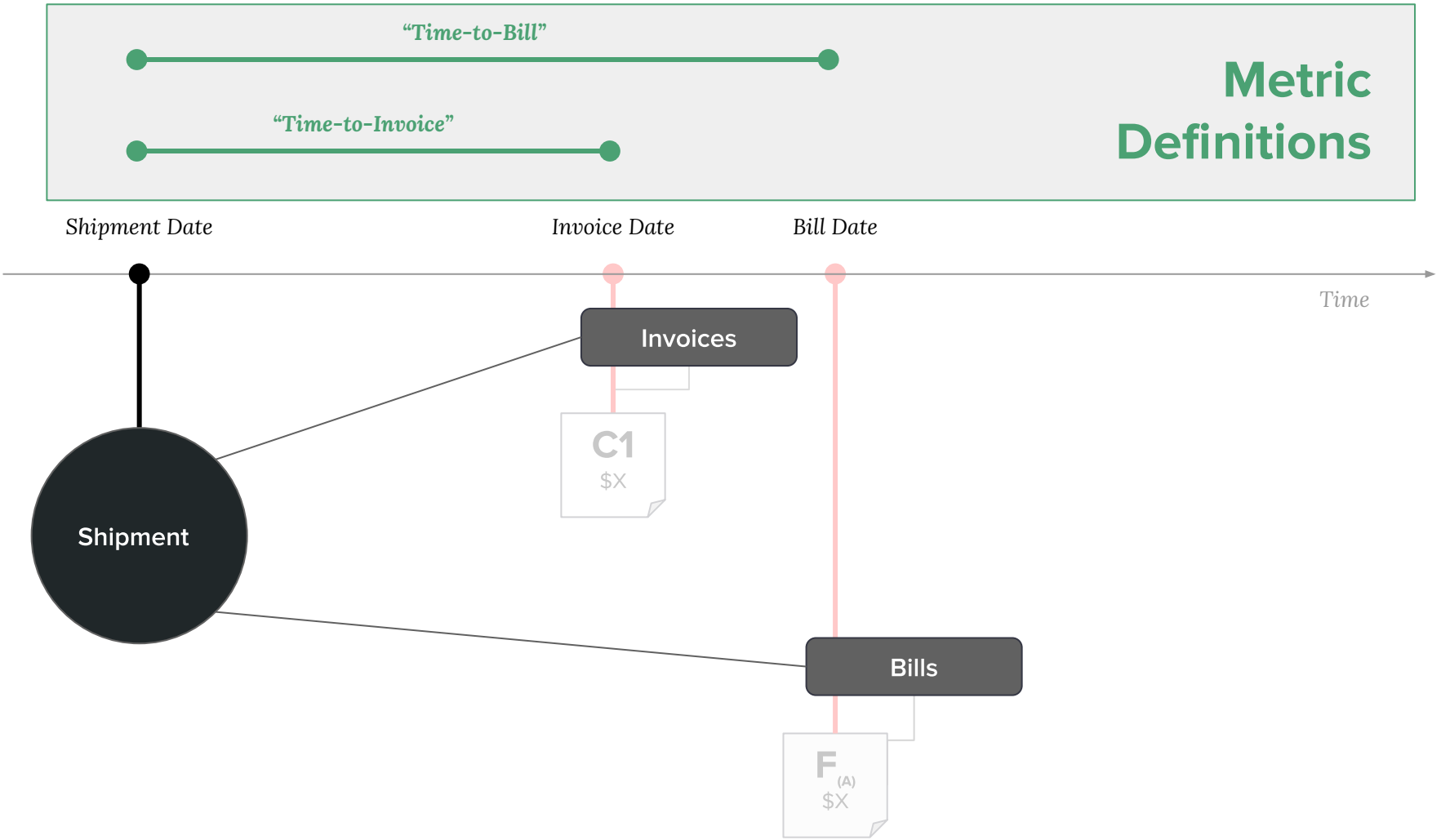


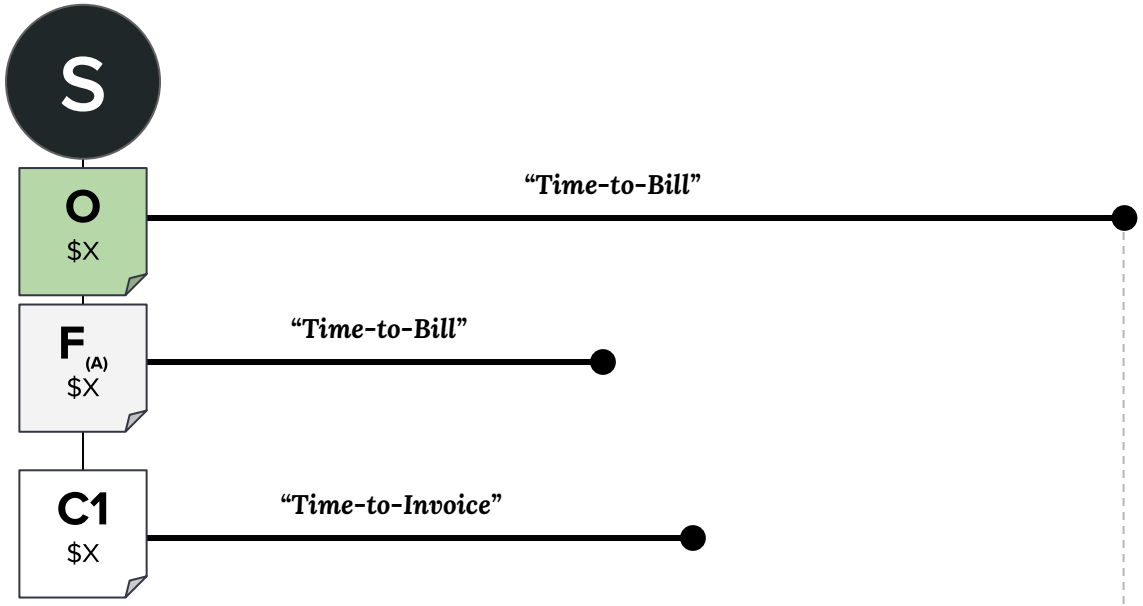
The Data

Entities & Relationships


Time
Invoices
C1
$X
C2
$X
Customer
Customers
Shipment
Bills
Shipment Variety
O
Origin
Air Freight
F
Ocean Freight
Vendors
by type
D
Destination
O
$X
F (A,O)
$X
D
$X
Legend
XOR
"Potential additional" Relationship
"Minimum guaranteed" Relationship
Entity Grouping (not on time scale)

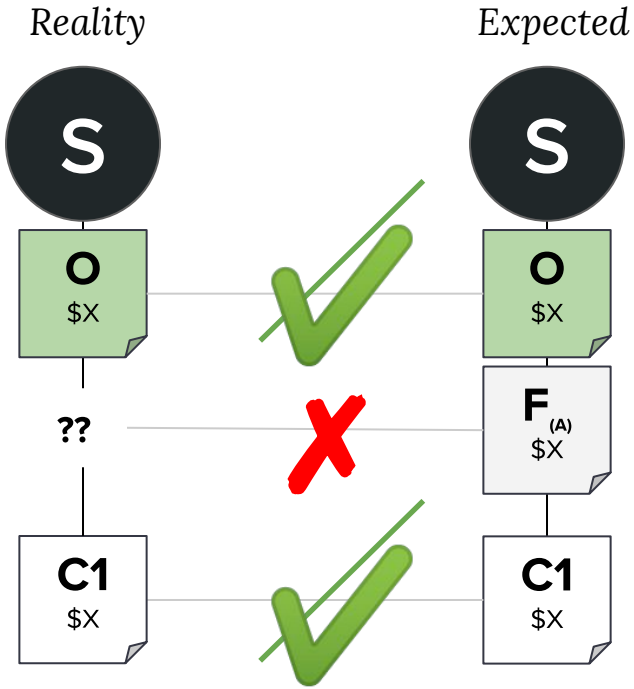
Metric Definitions


"Time-to-Bill"
"Time-to-Invoice"
Metric Definitions
Shipment Date
Invoice Date
Bill Date
Time
Invoices
C1
$X
Shipment
Bills
F
(A)
$X



"Time-to-close" for a particular shipment is the **maximum of all "time-to-bill"(s) and "time-to-invoice"(s)** for bills and invoices associated with said shipment.

Metric Definitions

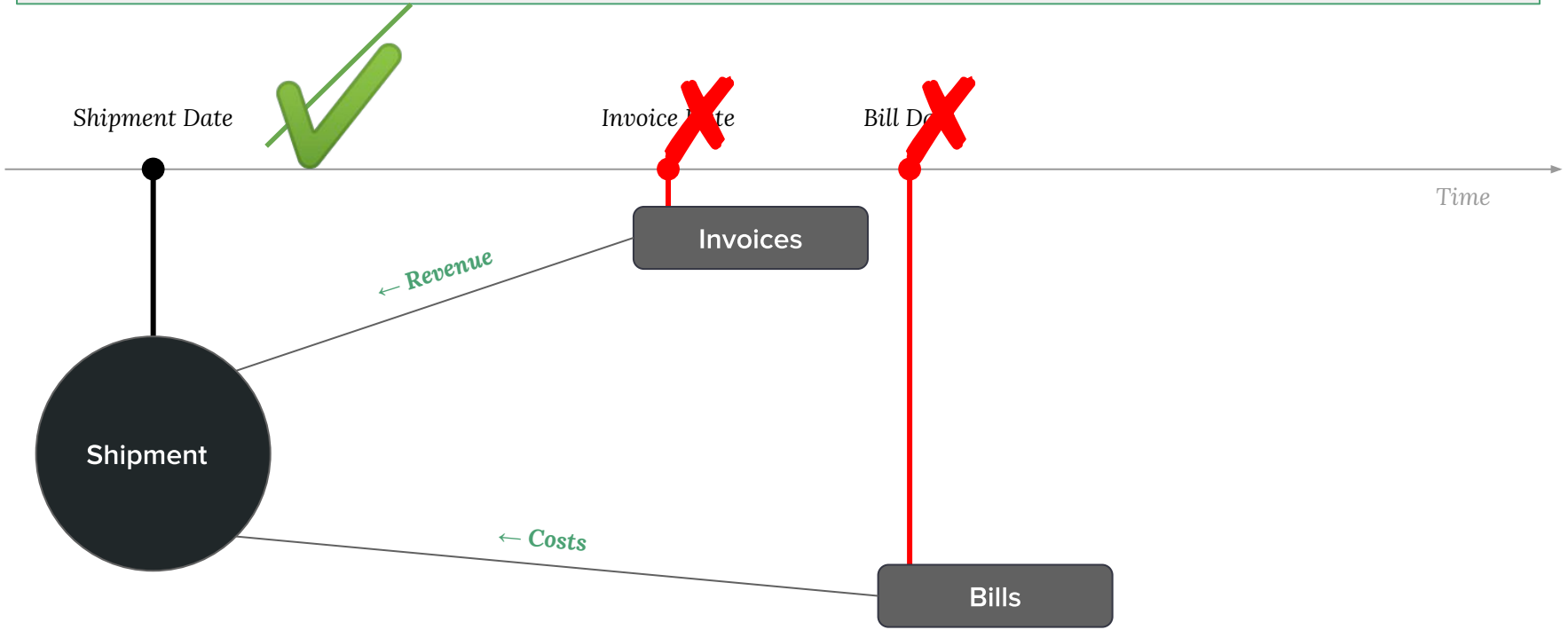


A "missing bill" (or "missing invoice") occurs when **we have information for only a few, but *not all* of the bills** (or invoices, respectively) **associated with a shipment** for services rendered *or* **we have record of a bill or invoice with NULLs.**

Metric Definitions

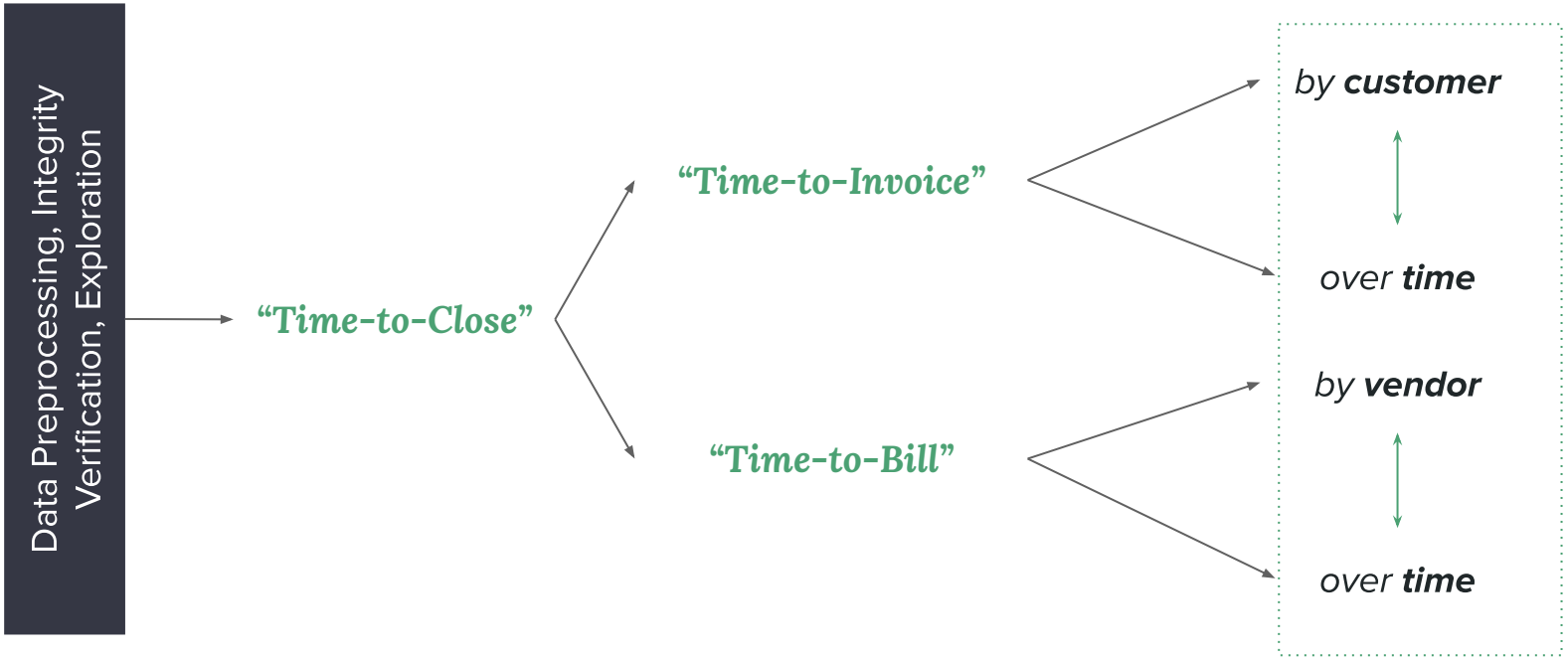
Metric Definitions

"Net revenue" is **revenue less cost**, where **revenue** is the sum of invoiced amounts and **cost** is the sum of billed amounts, **both recognized on the shipment date**, for a given shipment.



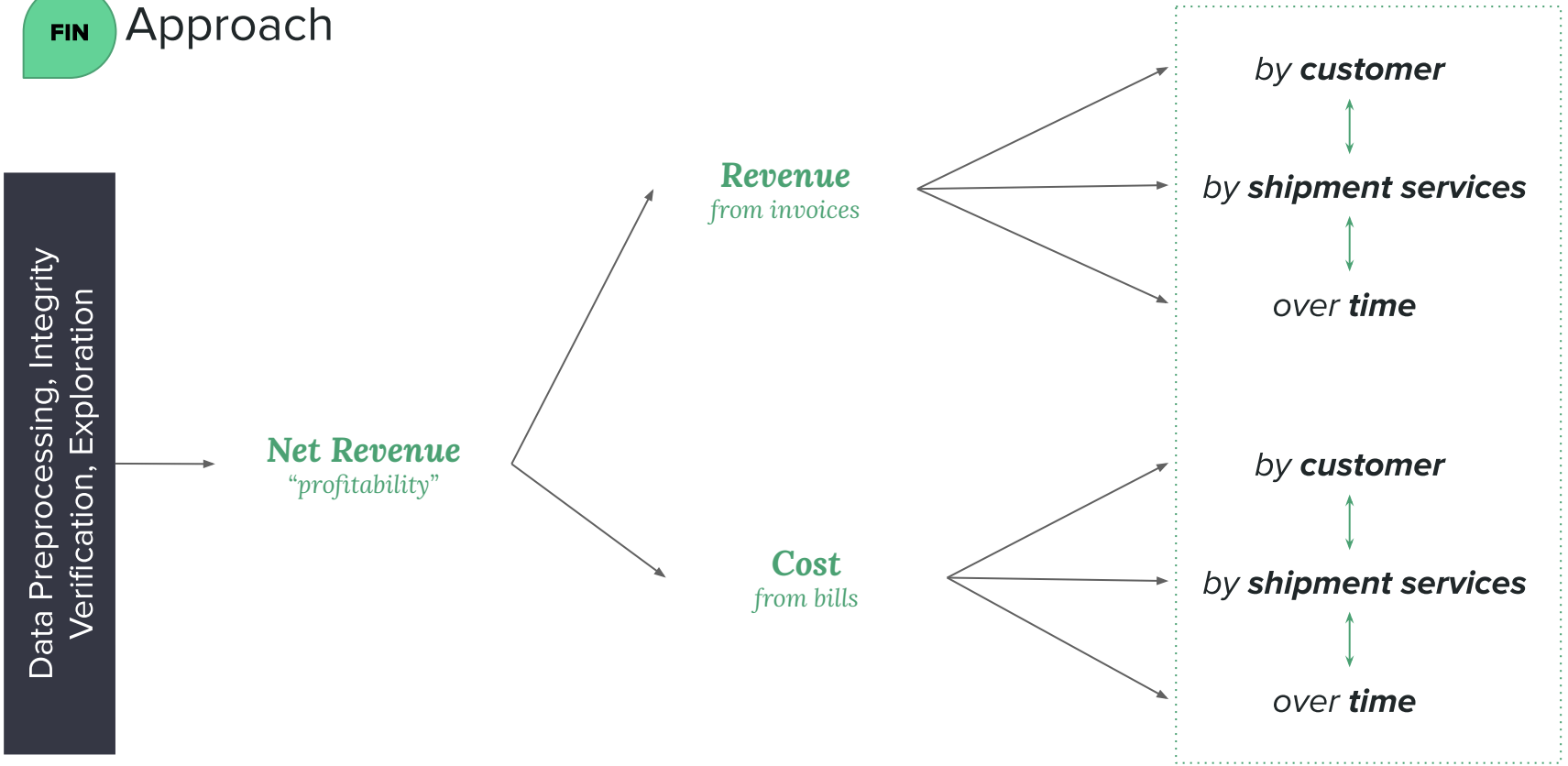
Approach

OPS Approach



Note: This isn't exhaustive. Subsequent summary visualizations attempt to capture multiple relationships due to economy of time!

FIN Approach



Note: This isn't exhaustive. Subsequent summary visualizations attempt to capture multiple relationships due to economy of time!

Analysis

Before diving in, let's inspect the data to get sense of shipments...

100

 # shipments

From January 2016 to May 2017

108

 # invoices

80

 # shipments invoicing 1 client

14

 # shipments invoicing 2 clients

221

 # bills

83

 # origin bills

98

 # freight bills

40

 # destination bills

49

 # air freight bills

49

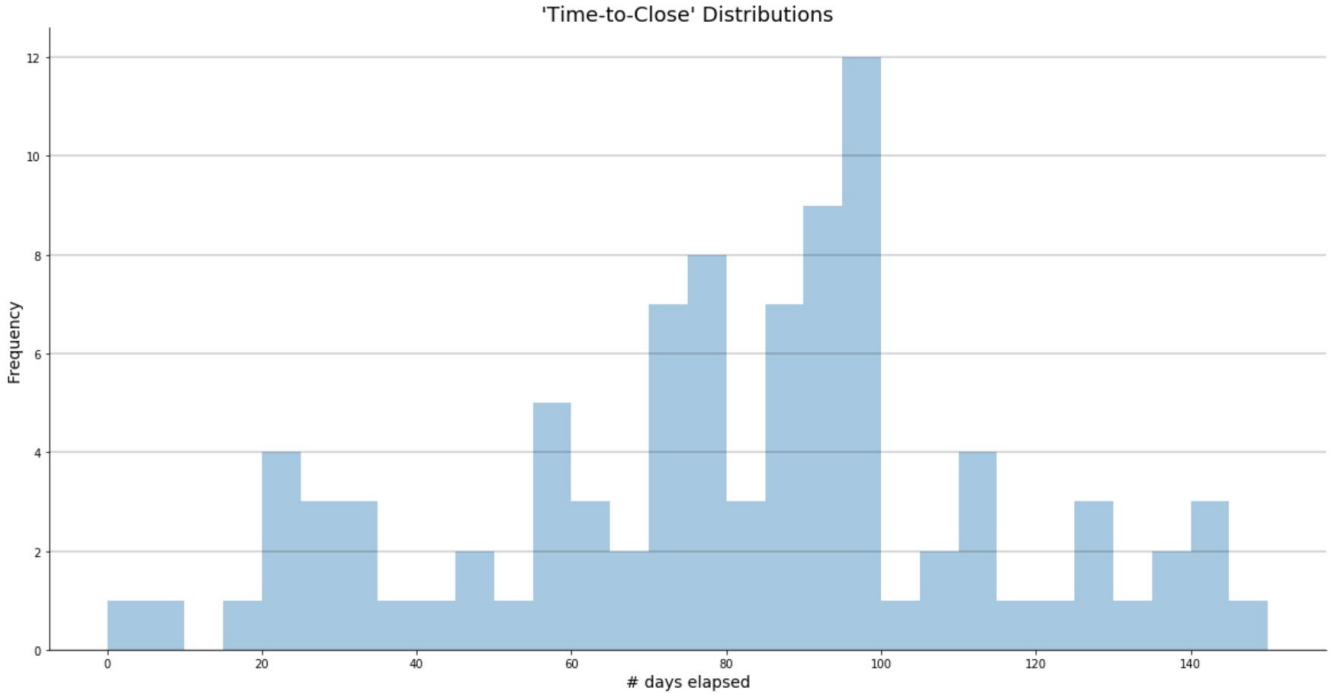
 # ocean freight bills

→ *Every* shipment uses freight services!

...and what data we're missing.

Distribution of “time-to-close” suggests fair amount of variance…

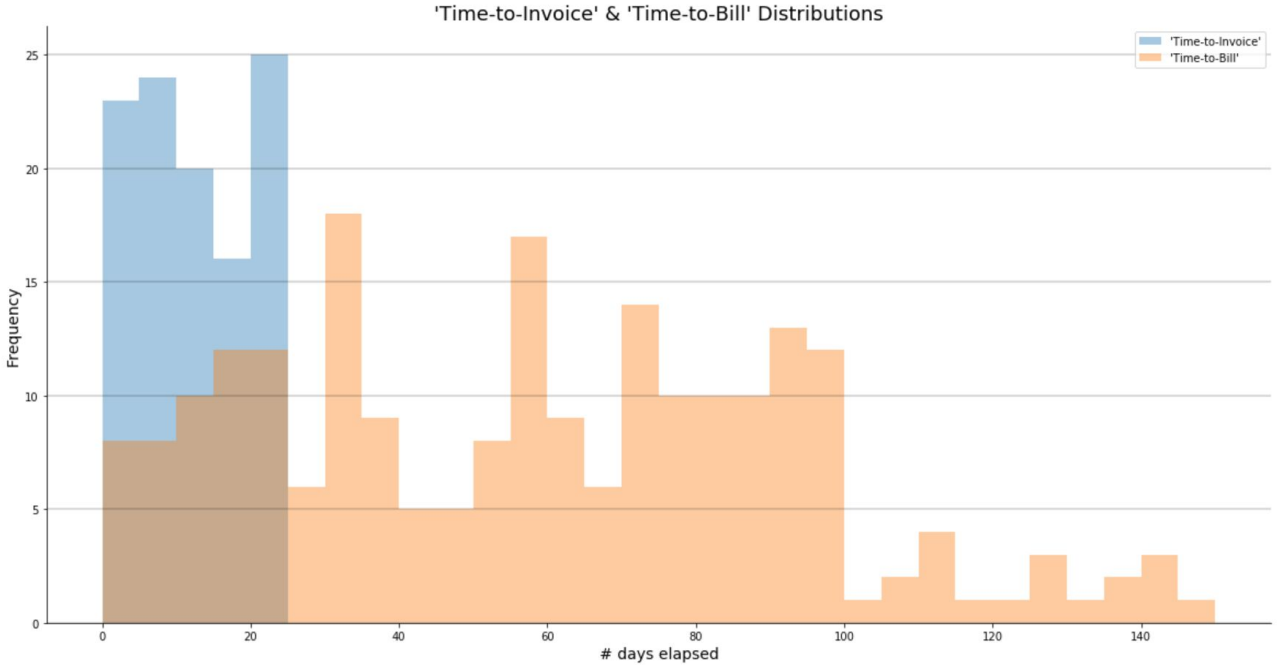
OPS



➔ “Closing” shipments seems to take **quite long**: **c. 80 days** (or almost a full quarter), on average.

...and we see that “time-to-bill” is the leading culprit.

OPS

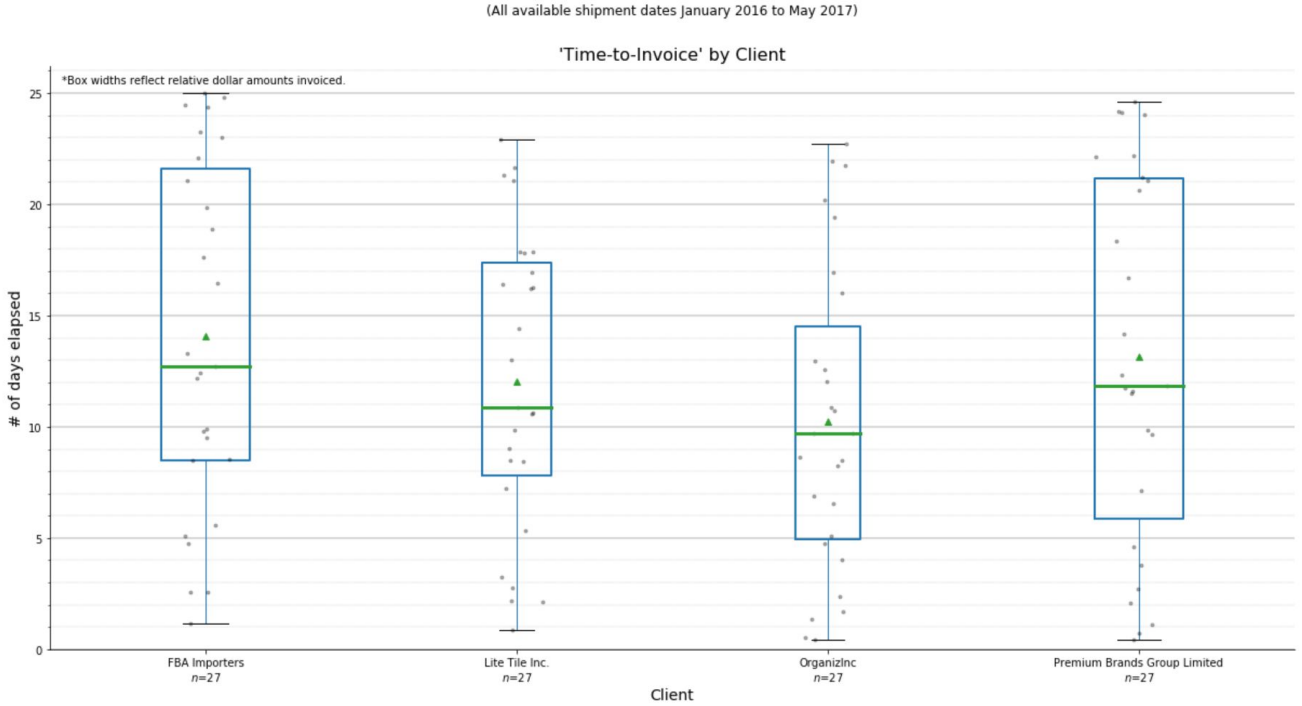


- ➔ Only **c. 10%** of bills are billed before their shipment's corresponding invoice.
- ➔ Only **c. 4%** of shipment “time-to-close” is determined by relatively slower invoicing.

“Time-to-invoice” is roughly the same among our customers…

OPS

FIN

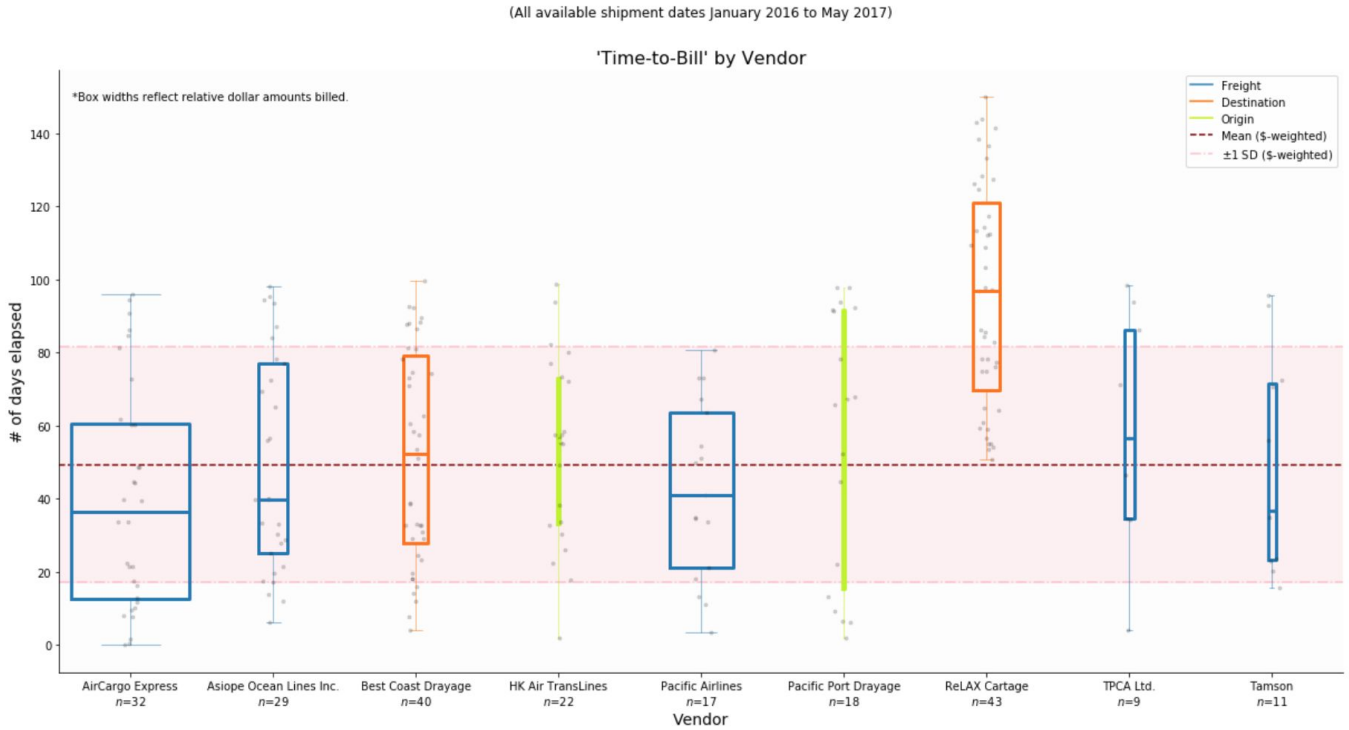


→ **c. 9-13 days** is roughly the **dollar-weighted average “time-to-invoice”** for any client, but this measure is **pretty spread out** (c. 8 days standard deviation) for each client.

...but one vendor in particular (ReLAX Cartage) is very slow to bill!

OPS

FIN

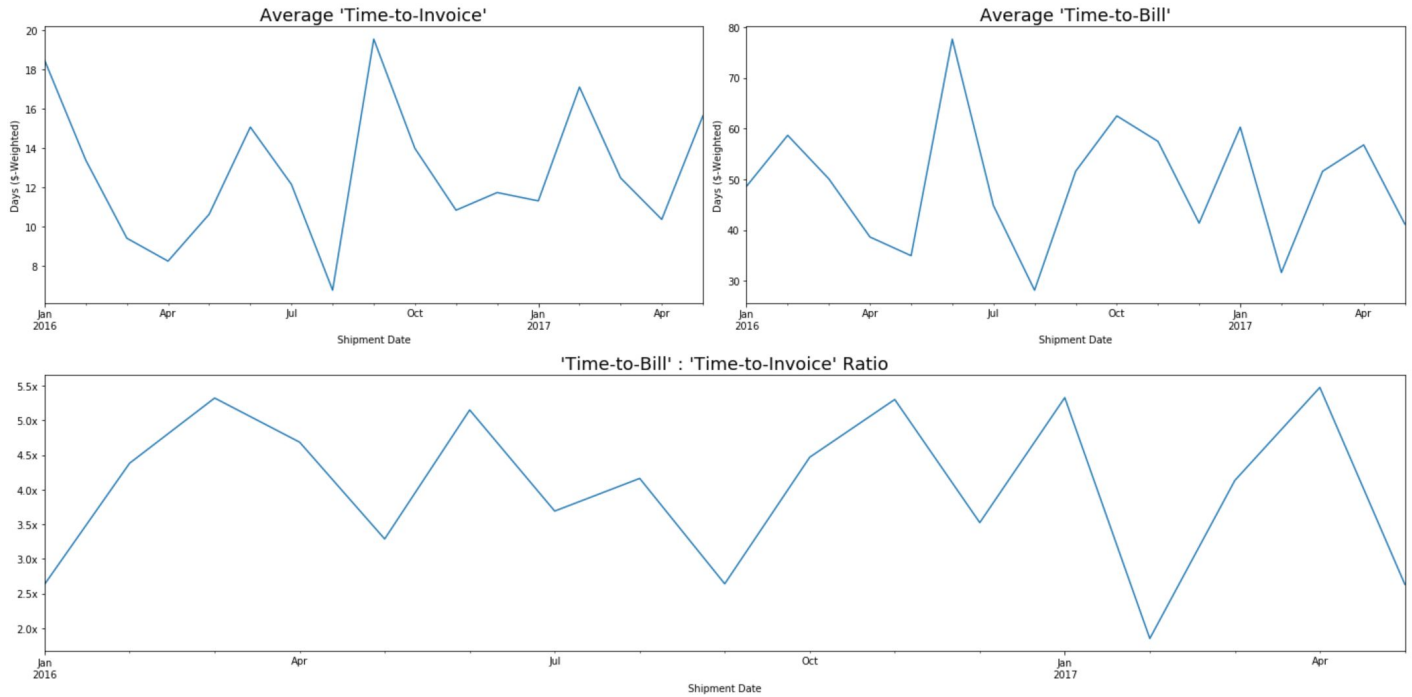


➔ Also, **the more expensive service-providers generally bill quicker**: 49-day dollar-weighted average vs. 58-day simple average "time-to-bill".

Given dearth of data, it's difficult to talk about trends over time...

OPS

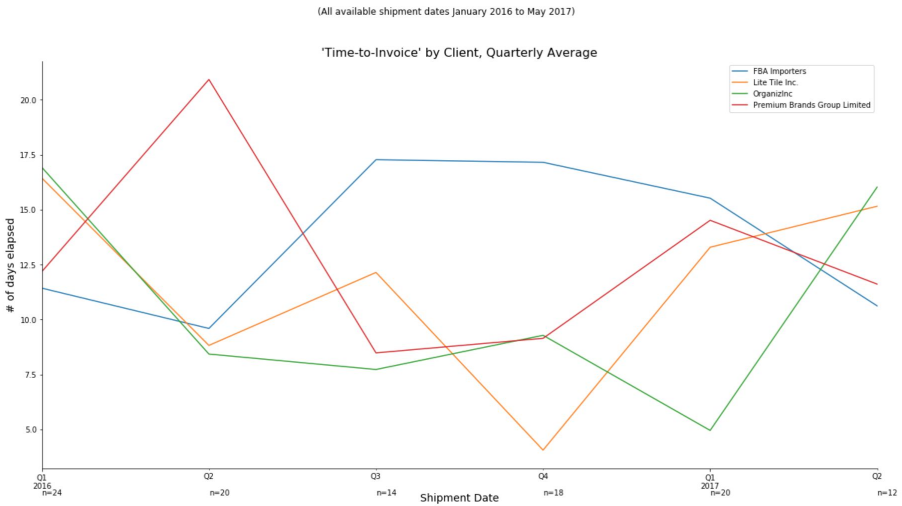
FIN

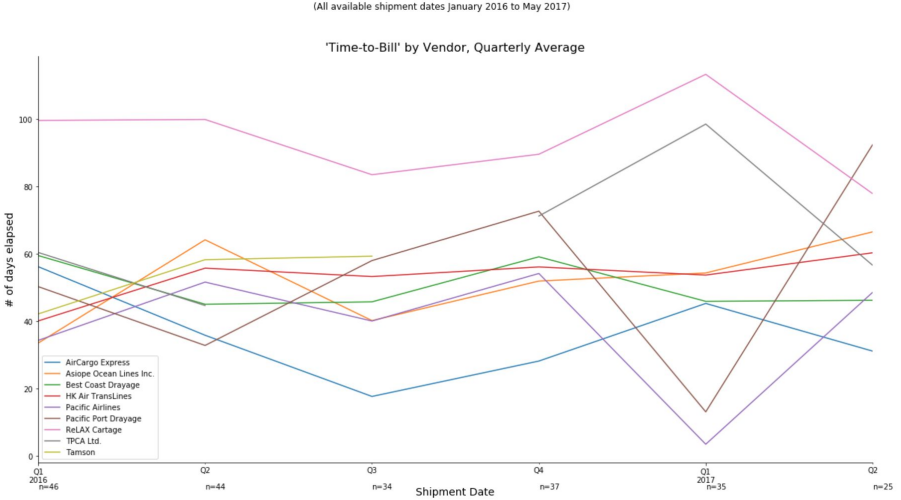


- ➔ Roughly, more activity in Q1, Q4 (slowing thru Q2, Q4) when measured by invoices and bills.
- ➔ Looking at **autocorrelations** for seasonality **didn't suggest further investigation**.

...and vendor and client cuts over time told largely the same story.

OPS

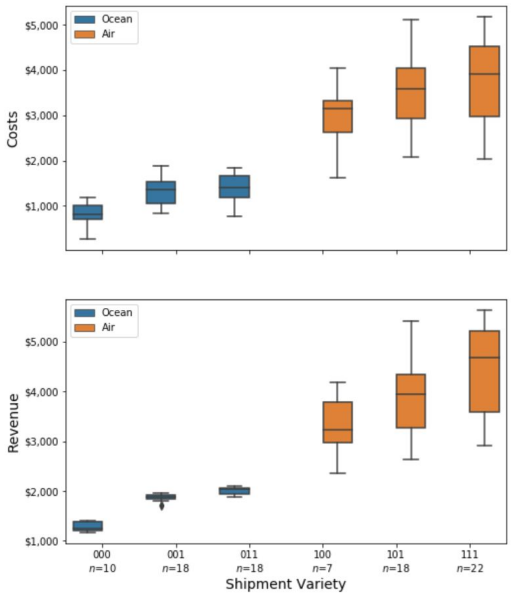


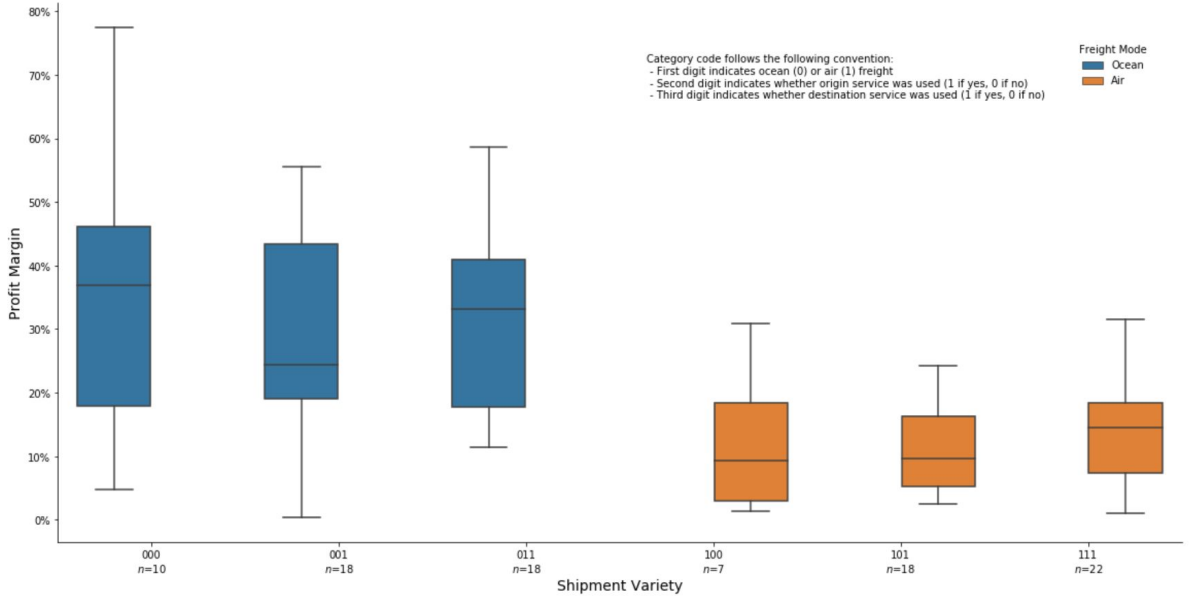


→ **ReLAX Cartage has always been slower** than vendor peers to bill.

Profitability depends on decision to transport via ocean or air...

FIN

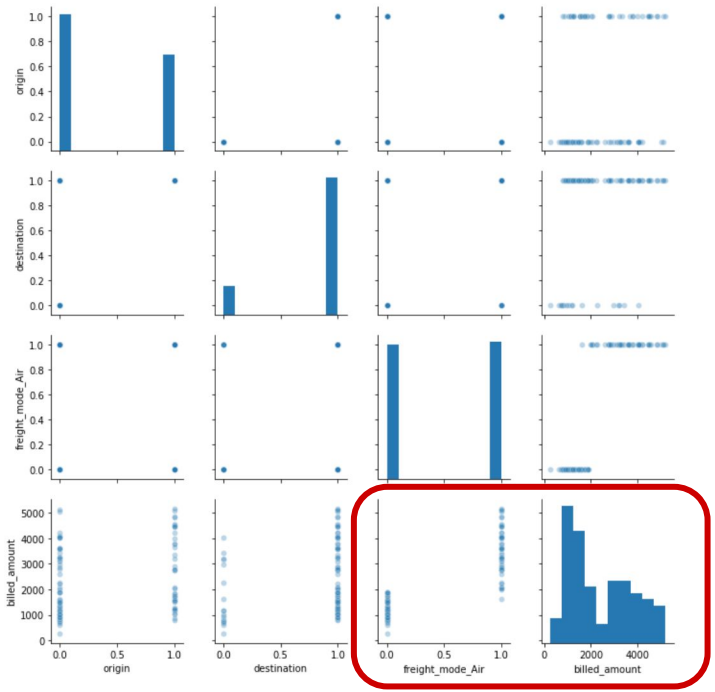




→ The driver for this profitability is **lower cost** of ocean freight, **not higher pricing**.

...and regression *might* be a good candidate method to predict costs...

FIN



OLS Regression Results

Dep. Variable:	billed_amount	**R-squared:**	0.759
Model:	OLS	**Adj. R-squared:**	0.752
Method:	Least Squares	**F-statistic:**	97.81
Date:	Thu, 29 Aug 2019	**Prob (F-statistic):**	1.17e-28
Time:	20:53:18	**Log-Likelihood:**	-787.74
No. Observations:	97	**AIC:**	1543.
Df Residuals:	93	**BIC:**	1554.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	775.7349	173.647	4.467	0.000	430.906	1120.564
x1	208.2174	151.432	1.375	0.172	-92.497	508.931
x2	487.6117	196.247	2.485	0.015	97.905	877.319
x3	2255.4024	138.043	16.338	0.000	1981.275	2529.529

Omnibus:	2.144	**Durbin-Watson:**	2.088
Prob(Omnibus):	0.342	**Jarque-Bera (JB):**	1.538
Skew:	-0.218	**Prob(JB):**	0.464
Kurtosis:	3.436	**Cond. No.**	5.48

→ Coefficients approximate factors: **x1 is origin services** (binary), **x1 is destination services** (binary), and **x3 is air freight** (binary).

...and regression *might* be a good candidate method to predict costs...

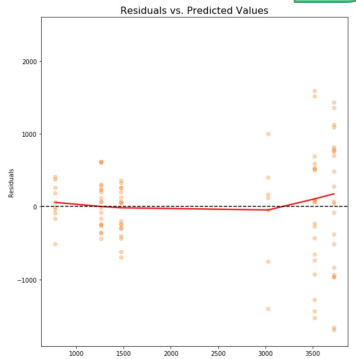
FIN

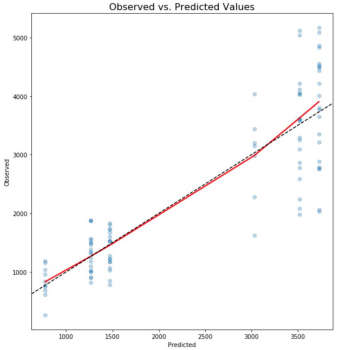
OLS Regression Results

Dep. Variable:	billed_amount	R-squared:	0.759
Model:	OLS	Adj. R-squared:	0.752
Method:	Least Squares	F-statistic:	97.81
Date:	Thu, 29 Aug 2019	Prob (F-statistic):	1.17e-28
Time:	20:53:18	Log-Likelihood:	-767.74
No. Observations:	97	AIC:	1543.
Df Residuals:	93	BIC:	1554.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	775.7349	173.647	4.467	0.000	430.906	1120.564
x1	208.2174	151.432	1.375	0.172	-92.497	508.931
x2	487.6117	196.247	2.485	0.015	97.905	877.319
x3	2255.4024	138.043	16.338	0.000	1981.275	2529.529

Omnibus:	2.144	Durbin-Watson:	2.088
Prob(Omnibus):	0.342	Jarque-Bera (JB):	1.538
Skew:	-0.218	Prob(JB):	0.464
Kurtosis:	3.436	Cond. No.	5.48





➔ I worry about **non-constant variance (heteroskedasticity) for freight type** — our predictions won't be as good for those shipments that use air transportation.

...given evidence of being able to generalize decently well [maybe].

FIN

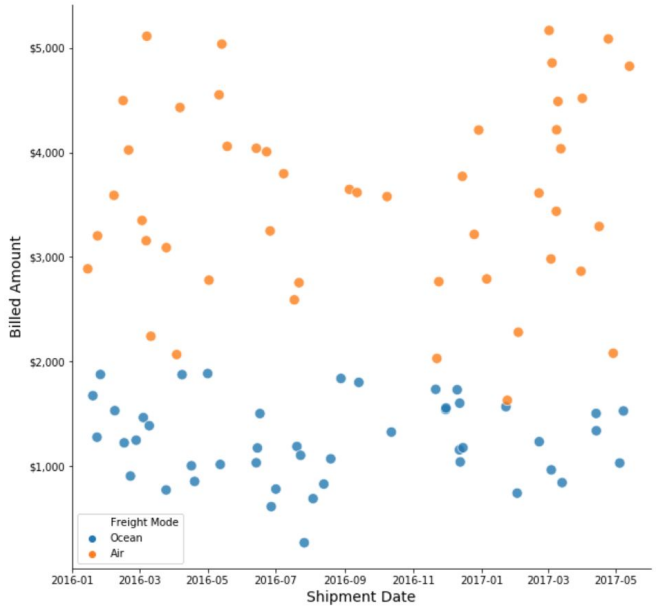
Random Train-Test

```
Training data: 67 records
Testing data: 30 records
Random State: 67
IS R^2: 0.7526 | IS RMSE: 656.3579
OS R^2: 0.7501 | OS RMSE: 701.7332
--------------------------------
Training data: 67 records
Testing data: 30 records
Random State: 41
IS R^2: 0.7567 | IS RMSE: 652.8822
OS R^2: 0.7282 | OS RMSE: 709.6886
--------------------------------
Training data: 67 records
Testing data: 30 records
Random State: 61
IS R^2: 0.7862 | IS RMSE: 607.7329
OS R^2: 0.6904 | OS RMSE: 784.1906
--------------------------------
Training data: 67 records
Testing data: 30 records
Random State: 95
IS R^2: 0.7354 | IS RMSE: 691.3490
OS R^2: 0.8000 | OS RMSE: 605.7226
--------------------------------
Training data: 67 records
Testing data: 30 records
Random State: 12
IS R^2: 0.7413 | IS RMSE: 679.3947
OS R^2: 0.7883 | OS RMSE: 633.9426
--------------------------------
```

vs.

"Back to the Future" Train-Test

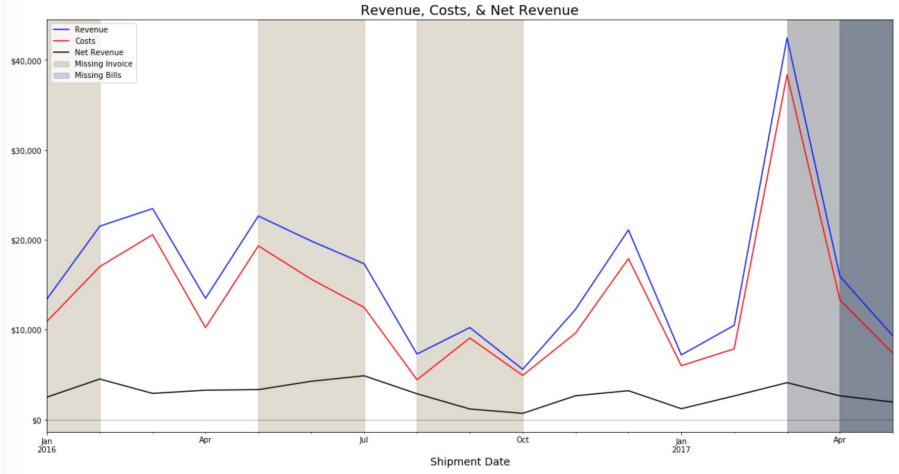
```
Training data: 21 records
Testing data: 76 records
Cut-off date: 2016-03-31 00:00:00
IS R^2: 0.7952 | IS RMSE: 562.6216
OS R^2: 0.7457 | OS RMSE: 694.9761
--------------------------------
Training data: 40 records
Testing data: 57 records
Cut-off date: 2016-06-30 00:00:00
IS R^2: 0.7939 | IS RMSE: 615.8809
OS R^2: 0.7150 | OS RMSE: 718.4413
--------------------------------
Training data: 55 records
Testing data: 42 records
Cut-off date: 2016-09-30 00:00:00
IS R^2: 0.8085 | IS RMSE: 587.6223
OS R^2: 0.6716 | OS RMSE: 770.7303
--------------------------------
Training data: 70 records
Testing data: 27 records
Cut-off date: 2016-12-31 00:00:00
IS R^2: 0.7975 | IS RMSE: 577.0079
OS R^2: 0.6218 | OS RMSE: 890.9002
--------------------------------
Training data: 87 records
Testing data: 10 records
Cut-off date: 2017-03-31 00:00:00
IS R^2: 0.7805 | IS RMSE: 622.4212
OS R^2: 0.5709 | OS RMSE: 976.3895
--------------------------------
```



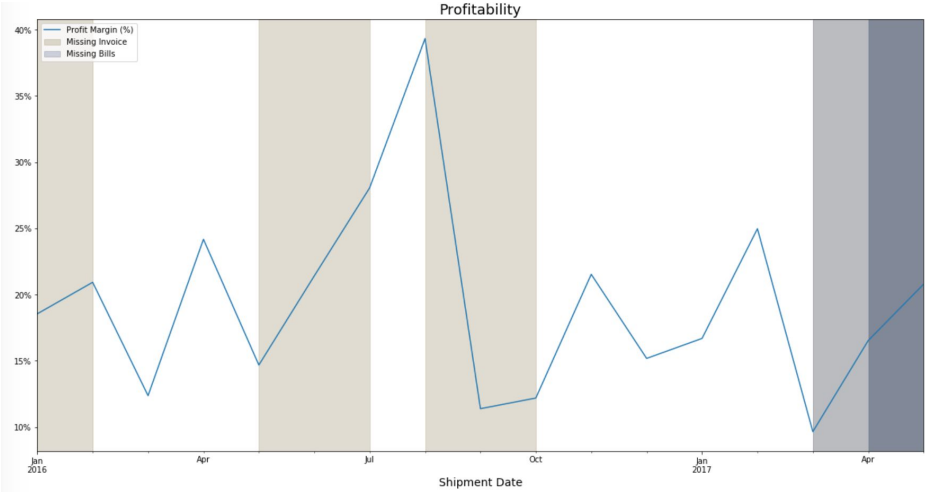
→ If granted more time, **revisiting distribution of shipment varieties over time** or would be a good next step. Also **additional data could better specify the model.**

Finally, it's difficult to pick out profitability trends given missing data...

FIN



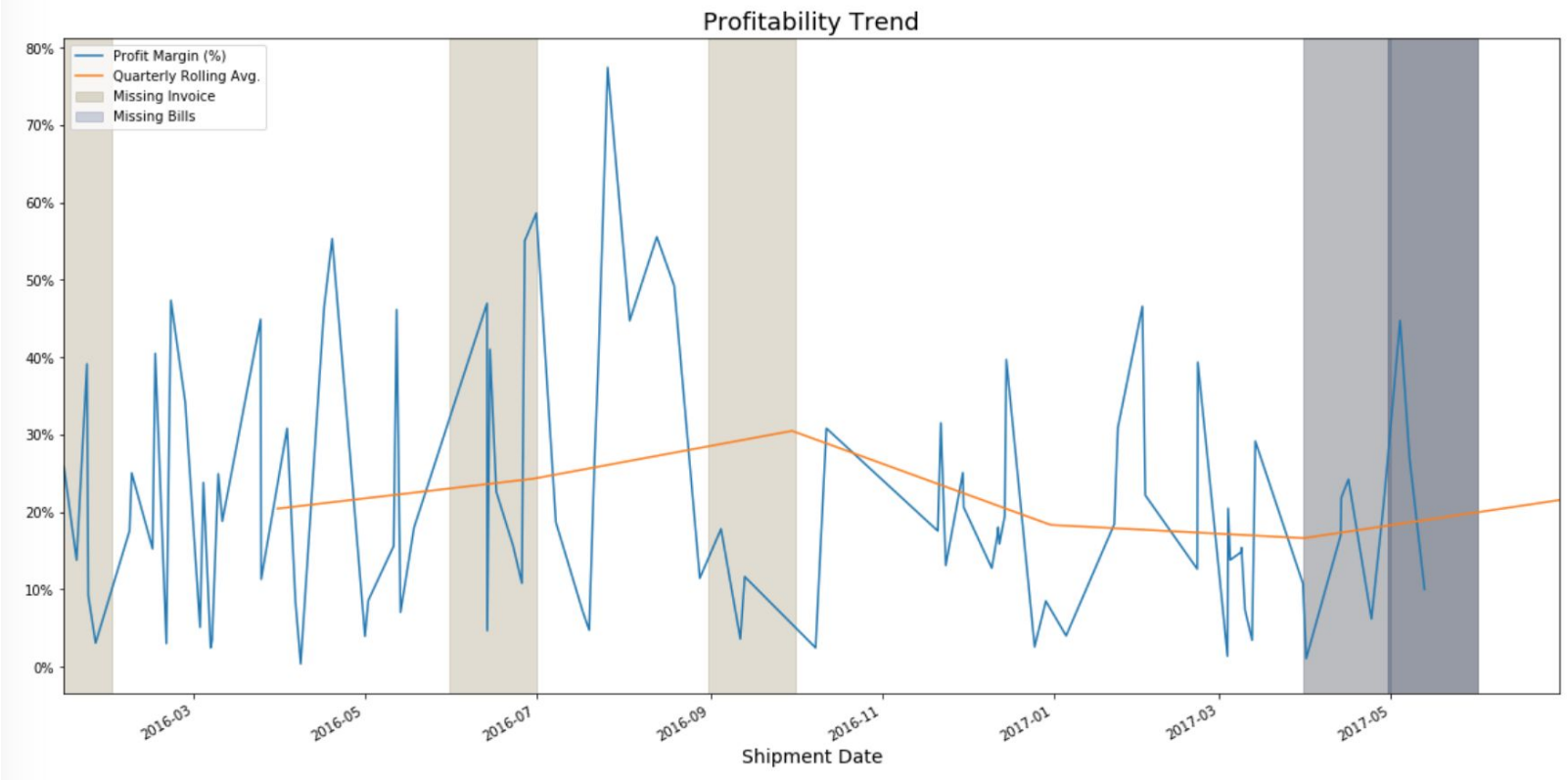
Note: Only shows shipments for which we have complete information.



→ **Missing bills** seem only to **appear in the later months**, and **better clarity** can be achieved **by fixing longer-standing missing invoicing**.

...even on a broader scale.

FIN



Note: Only shows shipments for which we have complete information.

Q&A