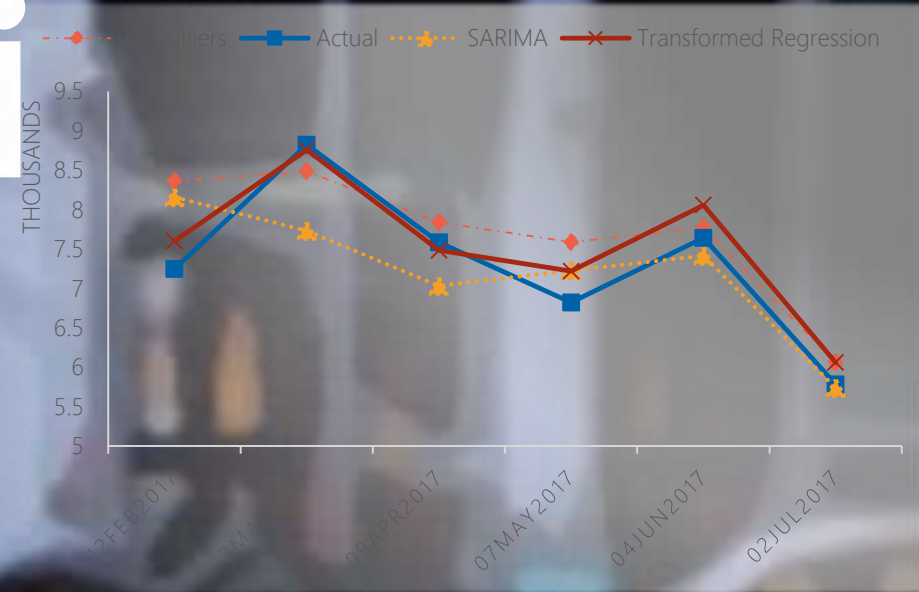


Chicago Taxi

Nathaniel Green



Texas A&M
Masters of Data
Analytics
Capstone Project

Business Context



Business Context

Disrupted Model
Business Questions:
Revenue & Demand



Data Collection

Chicago Data Portal
NOAA Weather
Parse and Aggregate



Data Cleaning

113M observations
2 Dependent
Independent Variables:
5 Time/Day
5 Weather
4 Area
3 Misc.



Model Exploration / Selection

Time-Series SARIMA
SARIMA w/ Outliers
Transformed Regress.
Transfer Function T-S
Dependent Variables:
1) Fares
2) Fare \$'s



Summary

Model and professional
summaries of project

Executive Summary

Seeking answers to:

What Influences Revenue? What Influences Demand?



Demand and
Revenue in Context:
Business & Industry

Business Questions:
Value of Predictions

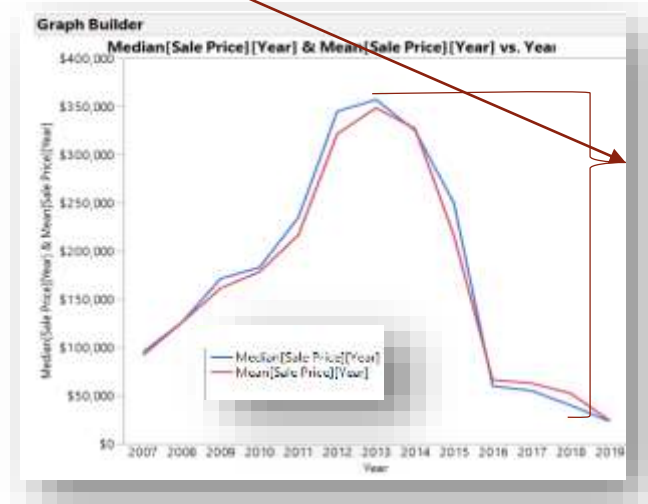
Industry Questions:
Level of Disruption
Sunsetting on Taxi
License Business

Answers usefulness in business:

- Owner / operators
 - License value peaked in 2013, at median valuations of \$350,000
 - Today prices are under reserve auction prices - set around \$30,000
 - Find a strategy in this disrupted industry
- Rideshare Companies
 - Price wars vs price protections

Answers usefulness in public domain:

- City Department Interests
 - Tax revenue
 - Traffic congestion
 - Commutable planning



* NYC Mayor Signs Bill Capping New Ride-Hail Licenses

Associated Press • Tue, Aug 14, 2018

NEW YORK (AP) - New York City, the largest American market for Uber, has become the first U.S. city to regulate the growth of app-based rides. [\(Read More\)](#)

**

With nearly half of Chicago cabs in foreclosure or idled, cabbies' hopes riding on New York-style ride-share limits

*<https://apnews.com/5ac2b84c246441018eb4ce985703e5db>

**<https://www.chicagotribune.com/business/ct-biz-chicago-taxis-ride-share-limits-20180823-story.html>

***<https://www.chicago.gov/content/dam/city/depts/bacp/publicvehicleinfo/medallionnews/medallionsales8222007to1312019.pdf>

Taxi Data: Sliced in Time and Areas



Executive Summary: Data

Framework:

4.5 years
1-1-13 to 7-31-17

9 official 'Sides'
77 areas
2 airports

Time Slices

- Weekly data is useful from a medallion rental basis
 - 4.5 years = 239 weeks
- Daily data is useful for weekly seasonality for operations
- Shifts are 12 hours (rent includes AM or PM rush hour)
- Rush Hour starts at either 4AM or 4PM (binned 4 hours)
 - 4.5 years = 1673 days
- Challenges with sensitivity in averages and aggregated (e.g. weeks with holidays, weekend rush hours)

Area Slices

- Dominant areas = Loop, Near West, and Airports. Subordinate areas = most South and outskirts.
- Commuter trends (in to and out of destination areas – pick up and drop-off)
- Airport fares unique:
 - Added fee to pickup at airport
 - Possible business opportunity – find efficient times against traffic.
- Challenge with traffic impacts. The goal is to make lots of trips efficiently.

Values context:

Daily Revenue: Mean = \$862k : Median = \$889k

Daily Demand: Mean = 67.4k trips : Median = 69.2k trips

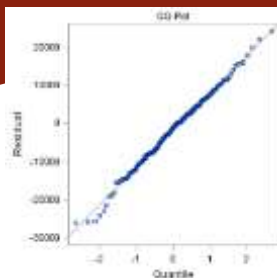
Taxi Data: Executive Summary – Findings



Executive Summary: Finding

Process Keynotes

Independent Variables
Used to Predict
Demand and Revenue



Weekly SARIMA (0,1,1)(0,1,1) 52

Daily SARIMA (2,1,1)(0,1,1) 7



Performance metrics:

- SBC scores
- Variances in holdout term (x24)

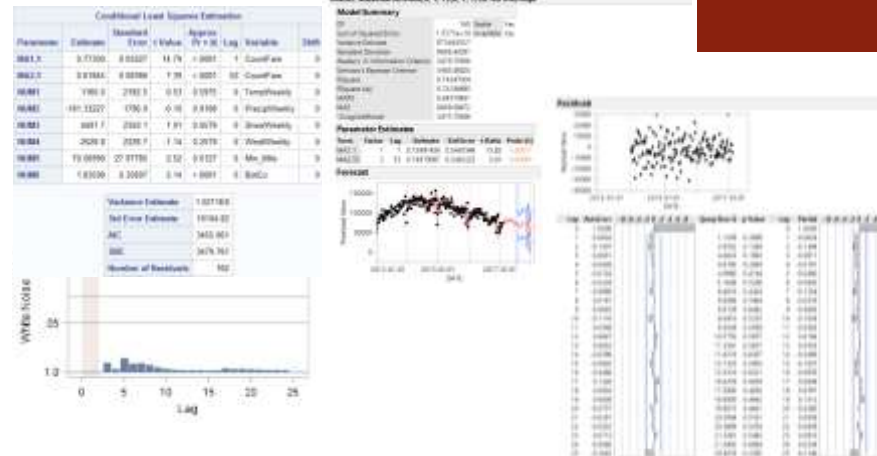
Validation process:

- Autoregressive and Part.-Autoregressive Residuals for white noise hypothesis
- Constant means
- Coefficients valid and parsimonious



Transformed Regression

SARIMA



Time-series: seasonality, autocorrelation, and moving averages fairly consistent, so p and q coefficients were similar

Weather coefficients were significant in Transformed Regression



Companies and transfers were most useful coefficients for predictions

Holidays and weekday (v. Sat / Sun) had statistical significance in Transformed Regression



Trip efficiency used in many models, but rush hours and shifts not effective predictors in transfer-function time-series.



Data Collection



Business Context

Disrupted Model
Business Questions:
Revenue & Demand



Data Collection

Chicago Data Portal
NOAA Weather
Parse and Aggregate



Data Cleaning

113M observations
2 Dependent
Independent Variables:
5 Time/Day
5 Weather
4 Area
4 Misc.



Model Exploration / Selection

Time-Series SARIMA
SARIMA w/ Outliers
Transformed Regress.
Transfer Function T-S
Dependent Variables:
1) Fares
2) Fare \$'s



Summary

Model and professional
summaries of project

Data Procurement

Chicago Data Portal (<https://data.cityofchicago.org/>)

1. Taxi Trips under Transportation Dept
2. Other Chicago data procured (collectively exhaustive)
 1. L-Station Entries (CTA Ridership) – not used
 2. Bus Routes (CTA Ridership) – not used
 3. Divvy Trips (Bike sharing) – not used
 4. Park Event Permits (Park District Source) – not used
 5. Not Available: Rideshare Company Data (UBER / LYFT)



Data Collection

Chicago Data Portal

NOAA Weather Data

Weather Data

(<https://www.ncdc.noaa.gov/cdo-web/datasets/>)

1. National Oceanic and Atmospheric Administration :
1673 days 60 weather-coded columns

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION

Home Climate Information Data Access Customer Support Contact About

Search

Home > Climate Data > Local Climatological Data (LCD)

Data Tools: Local Climatological Data (LCD)

Local Climatological Data (LCD) is only available for stations and locations within the United States and its territories. Select the state or territory, location, and time to view specific data. Click the station name to view details or click "ADD TO CART" to order the station's data.

Map Tool

Select a Location Type: Country, US Territory, State, County, Zip Code

Select a State: Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine

Select a County: Champaign County, IL, Christian County, IL, Clay County, IL, Cook County, IL, Crawford County, IL, DeKalb County, IL

Local Climatological Data > County > Cook County, IL

1-3 of 5 Stations

STATION DETAILS

CHICAGO LANING MUNICIPAL AIRPORT, IL US
View Full Details
Station ID: 95044 LCD
Period of Record: 2009-01-01 to 2019-02-14

CHICAGO MIDWAY AIRPORT, IL US
View Full Details
Station ID: 95044 LCD
Period of Record: 2009-01-01 to 2019-02-14

CHICAGO O'HARE INTERNATIONAL AIRPORT, IL US
View Full Details
Station ID: 95044 LCD
Period of Record: 2009-01-01 to 2019-02-14

CHICAGO DATA PORTAL

Browse Tutorial Feedback

Taxi Trips Transportation

View Data Visualize Export API

Taxi trips reported to the City of Chicago in its role as a regulatory agency. To protect privacy but at consistent for any given taxi medallion number but does not show the number, Census Tracts are rounded to the nearest 15 minutes.

What's in this Dataset?

Rows: 113M Columns: 23 Each row is a Trip

Download Taxi Trips

Download Taxi Trips for offline use in other applications:

CSV CSV for Excel

Additional Formats

CSV for Excel (Europe) TSV for Excel

RDF XML

RSS

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
LATITUDE	4138	-41.8661000	0.1289156	-41.7372000	-41.9950000
LONGITUDE	4138	-87.8555500	0.0780594	-87.9336000	-87.7775000
ELEVATION	4138	195.4000000	6.4007735	189.0000000	201.8000000
DATE	4138	20393.00	597.3409656	19359.00	21427.00
AWND	2066	9.8283107	3.6449425	2.0100000	26.6200000
PGTM	1	1455.00		1455.00	1455.00
PRCP	4113	0.1182227	0.3331031	0	5.1100000
SNOW	4136	0.1233075	0.7114359	0	17.2000000
SNWD	4133	0.5681587	1.9530284	0	18.1000000
TAVG	1925	51.2472727	20.1844444	-8.0000000	87.0000000
TMAX	4138	59.9929918	21.7731109	-2.0000000	97.0000000
TMIN	4138	-43.1522475	20.0437043	-16.0000000	80.0000000

Data Description

Distribution of Fares

Characteristics 100% of data

Fare characteristics using 1% random sample



Data Collection

Data and Fare
Summaries

All Data w/ Relevant Columns

The MEANS Procedure

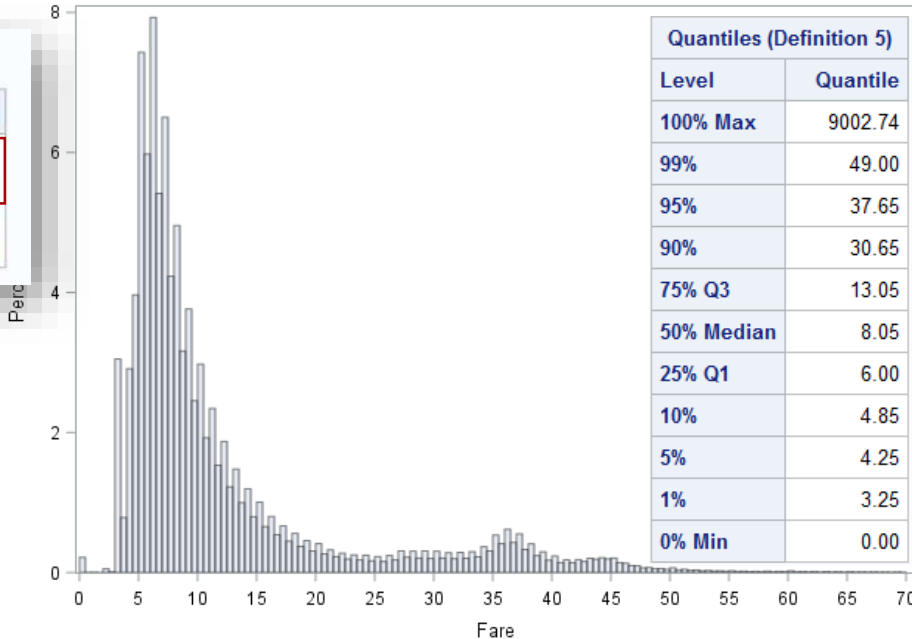
Variable	N	N Miss	Minimum	Maximum	Mean
Trip_End_Timestamp	112844186	15868	-1893369600	1817220600	1737017190
Trip_Seconds	111570347	1289707	0	86399.00	747.0295367
Trip_Miles	112859421	633	0	3460.00	2.7755557
Pickup_Community_Area	96077712	16782342	1.0000000	77.0000000	22.4048165
Dropoff_Community_Area	93764216	19095838	1.0000000	77.0000000	21.2125782
★ Fare	112858978	1076	0	9999.99	12.8008601
DateDel	112844186	15			
TimeDel	112844186	15			

The MEANS Procedure

Variable	N	N Miss
Pickup_Community_Area	96077712	16782342
Dropoff_Community_Area	93764216	19095838
Fare	112858978	1076
Trip_Seconds	111570347	1289707

Moments			
N	1128585	Sum Weights	1128585
Mean	12.732341	Sum Observations	14389529
Std Deviation	45.5765771	Variance	2077.22438
Skewness	148.23712	Kurtosis	25275.4268
Uncorrected SS	2527279948	Corrected SS	2344322204
Coeff Variation	357.959131	Std Error Mean	0.04290171

Distribution of Fare



Missing as a percentage of All Data:

	Fare	Pickup	Dropoff	Ttl Comm.
Missing N	1,076	16,782,342	19,095,838	19,754,821
Missing as % of total	0.00%	14.87%	16.92%	17.50%

Data Description Demand

- Parsed data aggregated into 4 super areas, split into:
 - high-demand areas in and out (except South)



Data Collection

Demand Drilldown

Fares Counted as
Demand

Fares Summarized as
Revenue

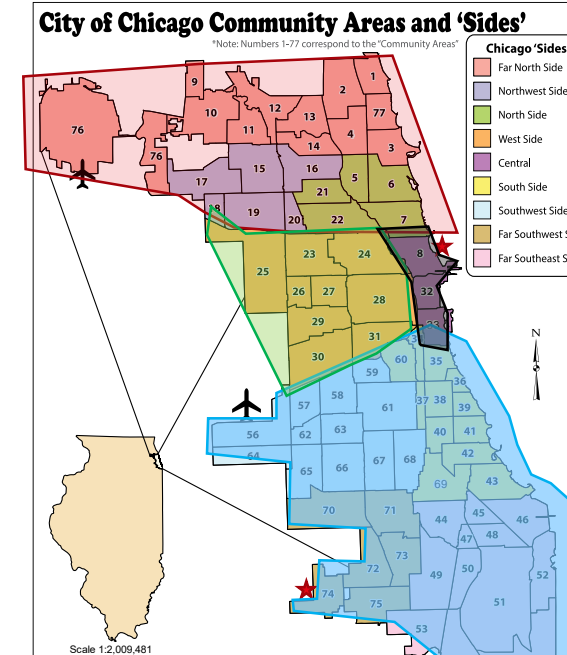
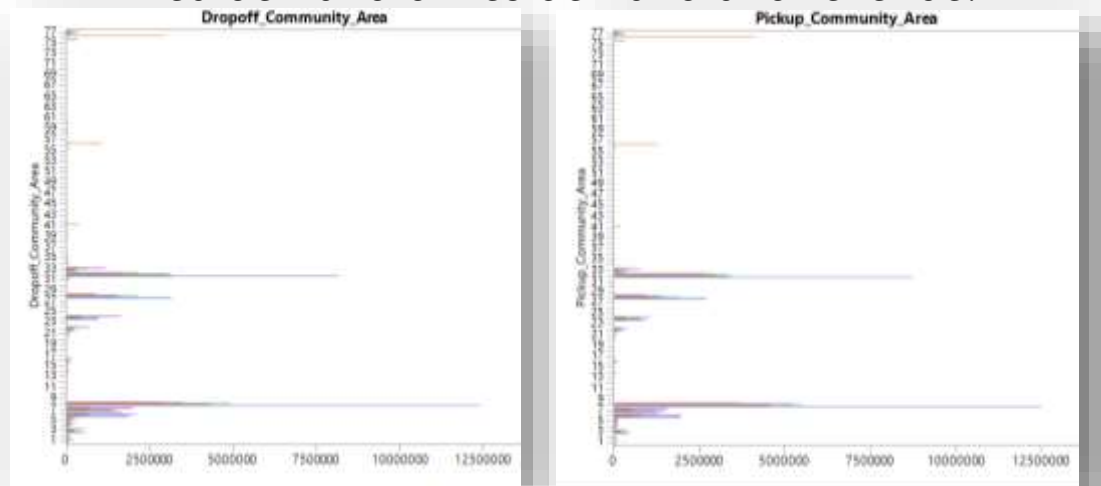
Top 12 Demand Areas

Pickup_Community_Area	% of Total	N	Dropoff_Community_Area	% of Total	N
8	33.41%	32,104,338	8	30.08%	28,201,515
32	22.35%	21,477,142	32	19.71%	18,480,324
28	8.83%	8,481,004	28	9.88%	9,267,949
6	7.09%	6,815,397	6	7.49%	7,021,590
✈️ 76	6.09%	5,853,830	7	6.69%	6,275,255
7	5.70%	5,477,208	24	4.83%	4,532,021
24	3.72%	3,575,711	✈️ 76	3.78%	3,540,993
33	2.29%	2,204,222	33	3.02%	2,831,048
3	1.71%	1,643,982	3	2.04%	1,912,080
✈️ 56	1.69%	1,619,459	22	1.87%	1,757,540
22	1.23%	1,184,590	✈️ 56	1.31%	1,228,219
77	1.12%	1,073,453	77	1.28%	1,196,975

Area Contrast



- Area demand drives demand and revenue.



Data Cleaning



Business Context

Disrupted Model
Business Questions:
Revenue & Demand



Data Collection

Chicago Data Portal
NOAA Weather
Parse and Aggregate



Data Cleaning

93M observations
2 Dependent
Independent Variables:
5 Time/Day
5 Weather
4 Area
4 Misc.



Model Exploration / Selection

Time-Series SARIMA
SARIMA w/ Outliers
Transformed Regress.
Transfer Function T-S
Dependent Variables:
1) Fares
2) Fare \$'s



Summary

Model and professional
summaries of project

Data Cleaning

Outliers and Missing Data

Removed taxi data

Fare between 1 and 499
Trip_Miles < 90
Dropoff_Community_Area > 0
Pickup_Community_Area > 0
Trip_Seconds < 20000
Sort by Time and Data



Data Cleaning

Taxi Data

Chose to remove and
adjust missing and
outliers.

- Removed Observations with Missing Community Areas (Total)
 - Both to and from for better accountability of Chicago roundtrip fares.
 - Transfer areas useful as model predictors.
 - Trends for Demand and Revenue Across 4.5 years for these 19.75M observations
 - Trailing 12 months drops off significantly
 - 50% of other outliers included missing community areas
 - 300% increase in outliers between missing community population compared to total population

2. Removed Observations with Missing and Outliers in Other Variables

- Missing Fares, Fares = 0, and Fares >= 500 observations = 260k
 - 148k observations without Communities (57%)
- Trip miles < 90 observations = 69k
 - 13k observations without Communities (20%)
- Trip Seconds < 20k (5hours) observations = 42k
 - 16k observations without Communities (38%)

3. Now What? Adjustments to Post-Processed Data:

- Add 10% to Demand for most area model outputs
 - Over time the percentages between the total data set and cleaned data set have narrowed:
 - 22.3% greater demand in city in year 0-1, compared to 11.2% for trailing year, and 10.3% for last 90 days
- Characteristics of available community area
 - Pickup skewed towards airports – suggest airports add 20% pickup demand



Periods	Mean % Diff Total and Full Demands
All	17.2%
Year 0-1	22.3%
Last 365	11.2%
Last 120	10.3%
Last 90	10.3%
Last 60	10.2%
Last 24	10.3%

Top 4 Areas Missing Transfers

Pickup_Community_Area	Missing % of Total	N_Missing	Dropoff_Community_Area	Missing % of Total	N_Missing
76	39.36%	1,170,086	8	20.69%	135,002
8	17.10%	508,378	32	14.73%	96,128
32	11.69%	347,499	76	11.56%	75,418
56	5.72%	170,040	28	8.22%	53,622

Rows	112,862,354
All rows	0
Selected	0
Excluded	0
Hidden	0
Labelled	0

Rows	92,908,591
All rows	0
Selected	0
Excluded	0
Hidden	0
Labelled	0

Data Cleaning Variables



Data Cleaning

Weather Variables,

Taxi Dependent &
Independent Variables

Time Bins and Variables

Dependent and Misc. Taxi Independent Variables:

Dependent Variables:

- Demand = Fare count
- Revenue = Fare summary

Independent Variables

- Top Company
- Bottom Company
- Top Transit (to or from)
- Min / Miles Efficiency

Created 5 independent binary variables from weather:

Any inclement weather

Precipitation event > 1mm any
precipitation

Snowfall event > 1MM snow fall . Snow
depth event > 5MM snow on ground

Wind event: Max (5 sec gust, 120 sec
gust, daily average)

Temperature event: Max(average, high /
low)

Created time-based bins and variables:

Grouped by: Week, Date, 4hr bin
x 2, shifts x 2

- Binned Time = 4 hour x 6 bins (1,0)
- Shifts = night 4pm-3:59am (1,0), day 4am –
3:59pm (1,0)
- Rush Hour = AM 4am (1,0), PM 4pm (1,0)

Added three day binary variables

W-H = Weekday, A-H =
Saturday*, U+H = Sundays
and Holidays

* Saturday prior to St Patrick
Day is treated as a Holiday

Drop.West
Source

Columns (21/1)

- Trip_Seconds
- Trip_Miles
- Pickup_Community_Area
- Dropoff_Community_Area
- Fare
- BotCo
- TopCo
- Company
- TimeDel
- TopTrans
- DateDel
- Sum[Trip_Sec...eDel Binned]
- Sum[Trip_Mil...eDel Binned]
- Min/Mile
- Sum[TopTran...eDel Binned]
- Count[Fare][...meDel Binned]
- Sum[Fare][Da...eDel Binned]
- Sum[BotCo][...eDel Binned]
- Sum[TopCo][...eDel Binned]
- TimeDel Binned
- DateDel+TimeDel Binned

Rows

All rows	14,335,781
Selected	1
Excluded	0
Hidden	0
Labelled	0

Drop.West.byDate4hrs
Source

Columns (31/1)

- Sum[Sum[T...ned]][DATE]
- Sum[Trip_S...eDel Binned]
- Sum[Trip_M...Del Binned]
- Min/Mile
- Sum[TopTr...eDel Binned]
- Count[Fare][...Del Binned]
- Sum[Fare][...eDel Binned]
- Sum[BotCo][...Del Binned]
- Sum[TopCo...Del Binned]
- TimeDel Binned
- 4:00:00 AM
- DayShift
- 4:00:00 PM
- DateDel+TimeDel Binned
- DATE
- U+H
- A-H
- W-H
- AnyInclementWeather
- WIND_mean_gust_gail
- Max[Snowf...Depth > 2.5]
- Precipitation > 1
- Max[MaxTe...nTemp > 80]

Rows

All rows	10,038
Selected	1
Excluded	0
Hidden	0
Labelled	0

Modeling Exploration / Selection



Business Context

Disrupted Model
Business Questions:
Revenue & Demand



Data Collection

Chicago Data Portal
NOAA Weather
Parse and Aggregate



Data Cleaning

93M observations
2 Dependent
Independent Variables:
5 Time/Day
5 Weather
4 Area
4 Misc.



Model Exploration / Selection

Time-Series SARIMA
SARIMA w/ Outliers
Transformed Regress.
Transfer Function T-S
Dependent Variables:
1) Fares
2) Fare \$'s



Summary

Model and professional
summaries of project

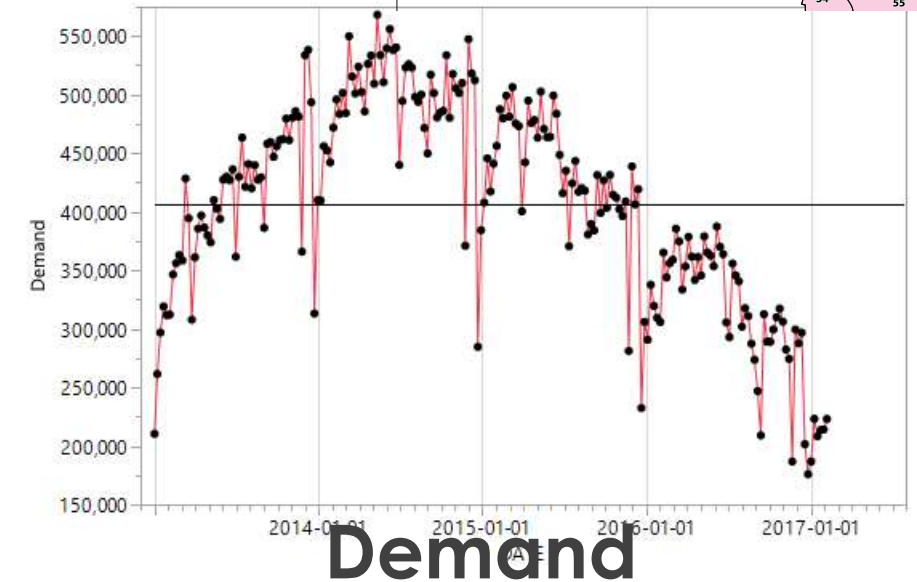
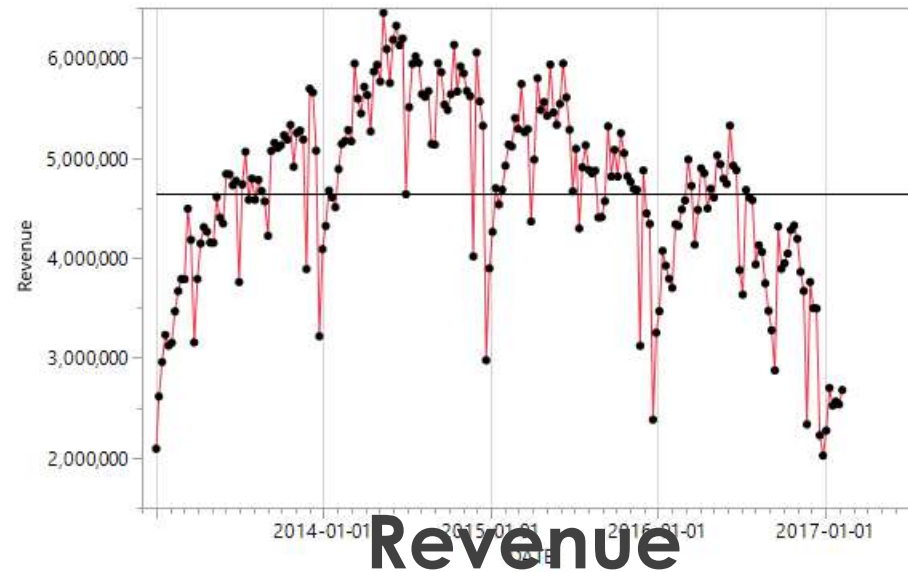
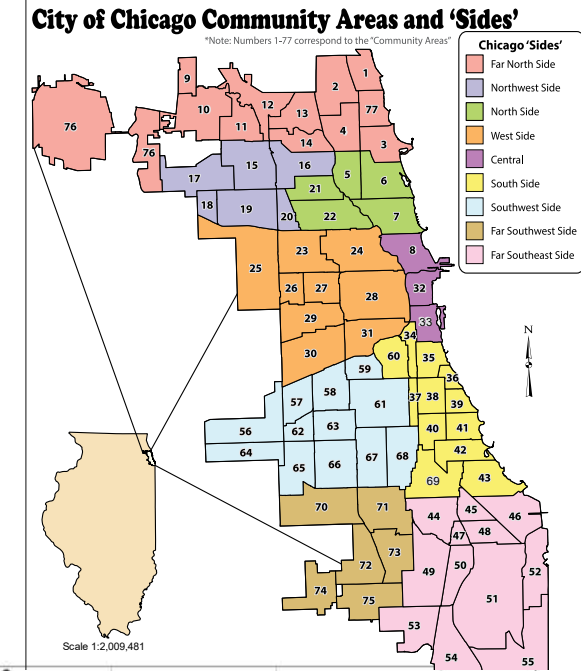
Model Roadmap



SARIMA Time-Series

Model outlines:
Inputs / Validity Tests
Effects on Targets
(responses, outliers, etc)
Model Takeaways

1. Models of Demand and Revenue at entire City Level:
 1. Time-Series Analysis – SARIMA (0,1,1)(0,1,1)₅₂, compared to
 2. Transformed Regression Analysis with residuals from weather with controlled outliers
2. Models of Pickup and Dropoff per Demand and Revenue in Areas: Loop, North and Airports, comparing:
 1. SARIMA Time-Series
 2. SARIMA Time-Series with controlled outliers
 3. Transformed Regression (with multiple determinant independent variables)
3. Model Comparison Advanced
 1. Daily SARIMA and Transformed Regression (including new temporal determinants)
 2. Transfer Function Time-Series (with efficiency, top transfer, or top company inputs)



Chicago Total Demand Weekly



Plotted regression residuals.
SARIMA (0,1,1)(0,1,1)52 with best SBC score.
All Individual Weather

SARIMA (0,1,1)(0,1,1)52 = Best

Demand (24 period)	SARIMA	Transformed Regression
Avg Absolute Variance %	10.04%	9.85%
Avg Abs Var	24,348.6	23,717.1
Avg % Var	-4.86%	-4.50%
Avg Variance	(12,265.6)	(11,499.5)
SBC	3788.66	3697.423

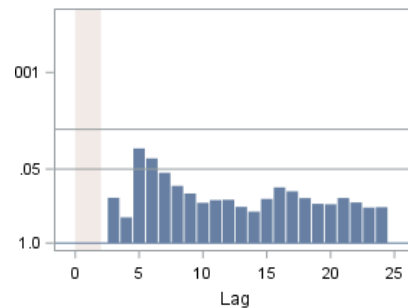
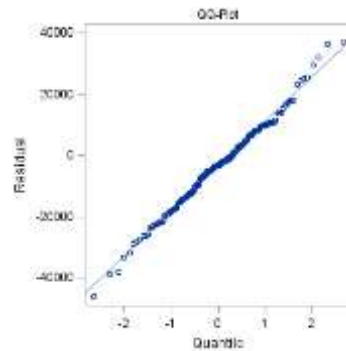


City-wide
Demand
Weekly

SARIMA (0,1,1)(0,1,1)52
Time-Series

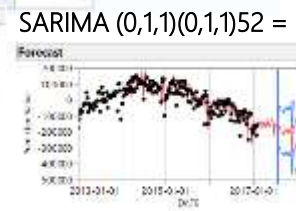
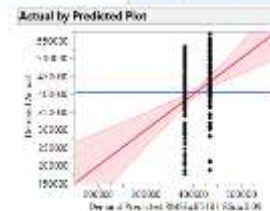
Compared to:
SAS Multi-Determinant
Transformed
Regression w/ Outliers

Transformed Regression

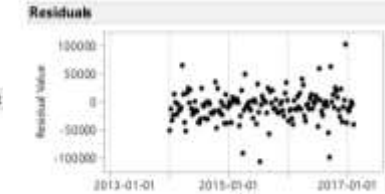
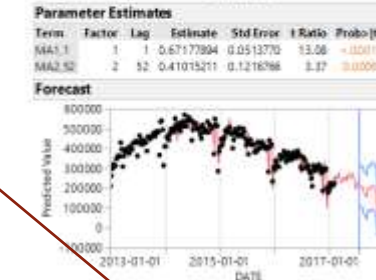


Maximum Likelihood Estimation						
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable
MA1,1	0.57551	0.07070	8.14	<.0001	1	Demand
MA2,1	0.43415	0.12314	3.53	0.0004	52	Demand
NUM1	1372.9	3396.7	0.40	0.6861	0	TempWeekly
NUM2	267.02454	2879.6	0.09	0.9261	0	PrecipWeekly
NUM3	10804.0	3959.4	2.73	0.0064	0	SnowWeekly
NUM4	-441.87358	3411.8	-0.13	0.8970	0	WindWeekly
NUM5	-87167.4	12311.0	-7.08	<.0001	0	AO133
NUM18	40110.4	14051.8	2.85	0.0043	0	AO204
NUM19	-41282.0	12297.0	-3.36	0.0008	0	AO95
NUM20	33926.6	12993.0	2.61	0.0090	0	LS45

Variance Estimate	2.5881E8
Std Error Estimate	16087.63
AIC	3629.496
SBC	3697.423
Number of Residuals	162



Model: Seasonal ARIMA(0, 1, 1)(0, 1, 1)52 No Intercept				
Model Summary				
DF	160	Stable	Yes	
Sum of Squared Errors	1.2004E+11	Invertible	Yes	
Variance Estimate	75025935			
Standard Deviation	27390.8714			
Akaike's AIC Information Criterion	3782.4848			
Schwarz's Bayesian Criterion	3788.8803			
RSquare	0.91326351			
RSquare-Adj	0.91272141			
MAPE	5.37657644			
MAE	20008.5352			
-2LogLikelihood	3778.4848			



Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	1.3187	0.2436	0	1.0000
1	-0.0907	1	0.9907	1	-0.0907
2	-0.1479	4.9699	0.0623	2	-0.1574
3	-0.0414	5.2758	0.1527	3	-0.0736
4	0.0487	5.6908	0.2235	4	0.0143
5	-0.0753	6.5491	0.2481	5	-0.0890
6	0.0143	6.8811	0.3509	6	0.0034
7	-0.0207	6.7586	0.4344	7	-0.0420
8	-0.0095	6.7743	0.5612	8	-0.0251
9	-0.0484	7.1808	0.6183	9	-0.0568
10	-0.0420	7.5052	0.6772	10	-0.0750
11	0.0396	7.7794	0.7329	11	0.0125
12	0.0760	8.8018	0.7108	12	0.0546
13	-0.0130	8.8321	0.7805	13	0.0033
14	-0.0338	9.0377	0.8286	14	-0.0176
15	0.0476	9.4469	0.8350	15	0.0402
16	-0.0899	10.7902	0.8222	16	-0.0918
17	0.0748	11.8186	0.8111	17	0.0755
18	0.0043	11.8202	0.8564	18	-0.0056
19	0.0028	11.8217	0.8931	19	0.0113
20	-0.0793	12.9996	0.8774	20	-0.0571
21	0.0700	13.9241	0.8728	21	0.0566
22	-0.0799	15.0178	0.8615	22	-0.0696
23	0.0578	15.8593	0.8495	23	0.0464
24	-0.0773	16.8089	0.8507	24	-0.0800
25	0.0855	18.2204	0.8326	25	0.0727

2%+ Improved prediction in
Transformed Regression
Model with weather and outliers



Snow statistically signif. – effect
increased trips 10.8k / week



SARIMA (0,1,1)(0,1,1)52 = Best

Chicago Total Revenue Weekly

Revenue (CHICAGO) SARIMA Transformed Regression

Avg Abs % Variance 13.25% 13.14%

Avg Abs Var 449,227.7 454,538.3

Avg % Var -9.98% -11.60%

Avg Variance (348,370.1) (405,625.5)

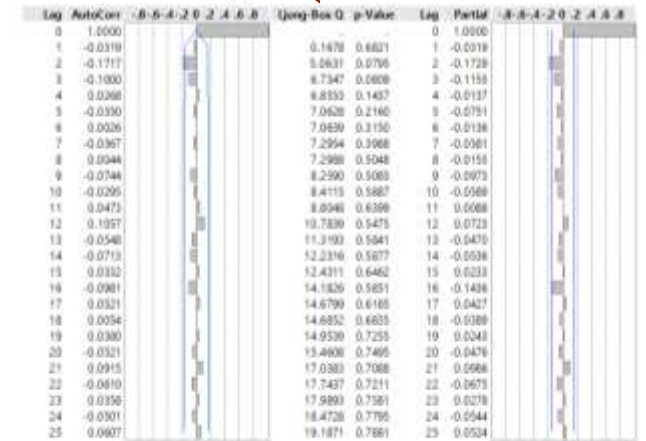
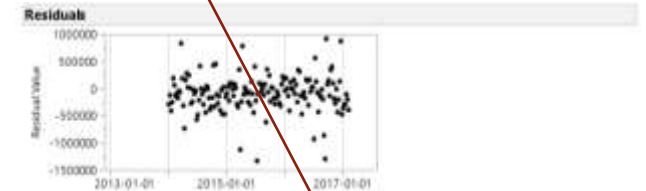
SBC 4599.49 4497.675

Transformed Regression



White Noise Hypothesis is Validated w./ autocorrelated & partial residuals.

SARIMA

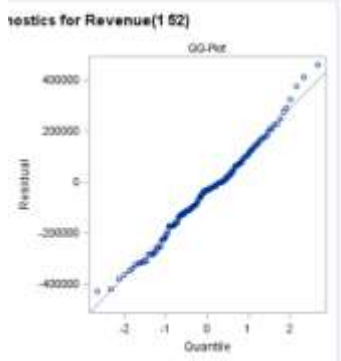
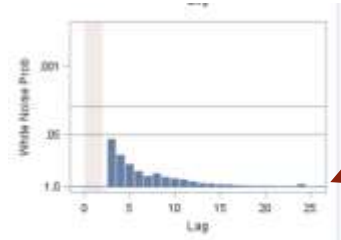
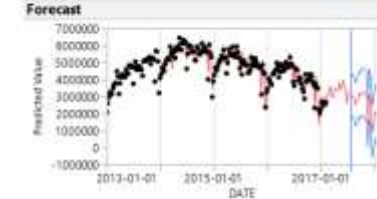


Maximum Likelihood Estimation						
Parameter	Estimate	Standard Error	Value	Approx Pr > t	Lag	Variable
MA1,1	0.62907	0.06455	9.61	<.0001	1	Revenue
MA2,1	0.66650	0.20162	3.31	0.0009	52	Revenue
NUM1	17942.3	39858.2	0.45	0.6526	0	TempWeekly
NUM2	16796.7	34844.1	0.48	0.6300	0	PrecipWeekly
NUM3	124604.2	46233.7	2.70	0.0070	0	SnowWeekly
NUM4	-16903.9	39889.5	-0.42	0.6717	0	WindWeekly
NUM17	-353188.6	132660.6	-2.66	0.0078	0	LS104
NUM18	-619523.4	155341.9	-3.99	<.0001	0	PO79
NUM19	530920.2	138925.8	3.82	0.0001	0	LS28

Variance Estimate	3.318E10
Std Error Estimate	182159
AIC	4432.836
SBC	4497.675
Number of Residuals	162

Model Summary			
DF	160	Statistic	Yes
Sum of Squared Errors	1.6317e+13	Invertible	Yes
Variance Estimate	1.0198e+11		
Standard Deviation	319344.011		
Akaike's AIC Information Criterion	4593.31077		
Schwarz's Bayesian Criterion	4599.48596		
RSquare	0.87502921		
RSquare Adj	0.87515376		
MAPE	5.66741294		
MAE	239511.201		
-2LogLikelihood	4569.31077		

Parameter Estimates					
Term	Factor	Lag	Estimate	Std Error	t Ratio
MA1,1	1	1	0.56448947	0.0502038	11.21
MA2,52	2	52	0.62902791	0.1800449	3.47



City-wide Revenue Weekly

SARIMA (0,1,1)(0,1,1)52 Time-Series

Compared to SAS Multi-Determinant Transformed Regression w/ Outliers

Snow statistically significant effects of + \$125k / wk Outliers and Determinants improved performance model 25% in t+1 – t+6 (t+1 – t+24 shown).

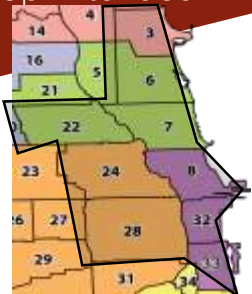


Loop Demand

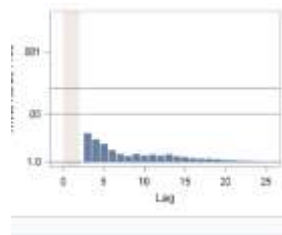


Loop Weekly Demand Drop-off, Comparing: SARIMA (0,1,1)(0,1,1)52: SARIMA w/ Outliers: Transformed Multi-Determinant

Loop Loop Extended



Demand Loop	SARIMA	SARIMA w/ Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	11.23%	11.09%	14.04%
Avg Abs Var	16,725.4	15,952.9	19,165.6
Avg % Var	-7.50%	-5.21%	14.04%
Avg Variance	(11,444.3)	(7,964.9)	19,165.6
SBC	3617.97	3546.48	3363.49



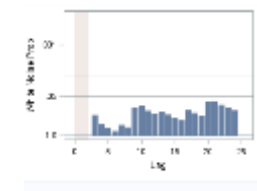
29% improved predictions in Multi-determinant Regression models



Outlier Maximum = 20, alpha=0.005 (3x iterate)



Demand LoopExt	SARIMA	SARIMA w/ Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	13.87%	12.91%	9.79%
Avg Abs Var	10,838.2	9,575.0	6,708.4
Avg % Var	-11.21%	-7.28%	0.48%
Avg Variance	(8,777.9)	(5,580.2)	512.9
SBC	3485.88	3441.89	3437.1



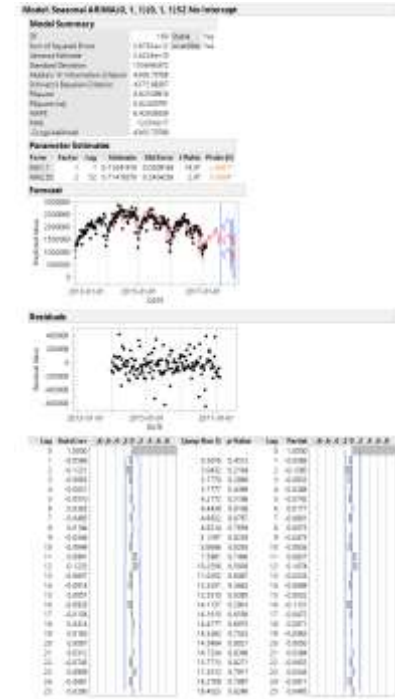
White Noise Hypothesis is Validated in models using auto and partial residuals

Transformed Regression

Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1.1	0.58145	0.05275	10.86	<.0001	1	CountFare	0
MA2.1	0.34499	0.05683	3.60	0.0004	52	CountFare	0
NUM1	-41961.9	10577.3	-5.86	<.0001	0	AD194	0
NUM2	-42300.6	8832.3	-4.79	<.0001	0	AO133	0
NUM11	31608.0	10908.8	2.98	0.0034	0	AO171	0
NUM12	-24497.1	8842.5	-2.77	0.0063	0	AO79	0
NUM13	-26795.9	10514.4	-2.55	0.0118	0	AO184	0

Variance Estimate	1.2991E8
Std Error Estimate	11384.73
AIC	3500.165
SBC	3546.479
Number of Residuals	162

SARIMA (0,1,1)(0,1,1)52 Loop



Areas 8 – Drop-off Tabulation

Company	N
Tax Affiliation Services	3942894
Dispatch Taxi Affiliation	1915945
Blue Ribbon Taxi Association Inc.	1389635
Choice Taxi Association	1142652
Northwest Management LLC	723551
KOAM Taxi Association	303494
Top Cab Affiliation	200765
Chicago Medallion Leasing INC	95891

TopCo	% of Total
0	43.38%
1	56.62%

Top company statistically signif. effect + .385 per unit



45% & 21% improved predictions in Transformed Regression models

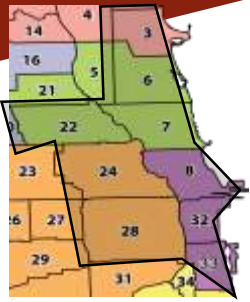


Loop Revenue

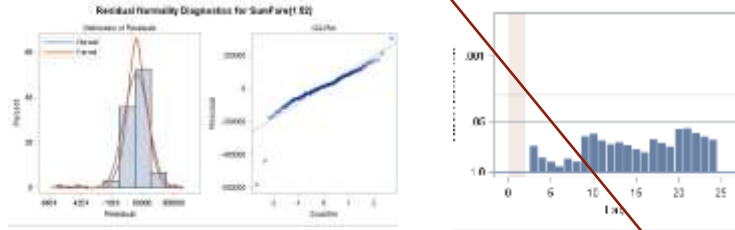


Loop Weekly Revenue Drop-off, Comparing: SARIMA (0,1,1)(0,1,1)52: SARIMA w/ Outliers: Transformed Multi-Determinant

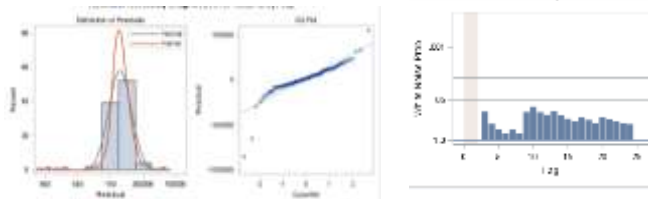
Loop Loop Extended



Revenue Loop	SARIMA	SARIMA Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	14.58%	13.49%	8.04%
Avg Abs Var	245,154.3	222,529.4	118,773.4
Avg % Var	-10.74%	-8.86%	-1.33%
Avg Variance	(186,145.4)	(152,532.2)	(18,787.8)
SBC	4375.88	4398.19	4214.09



Revenue LoopExt	SARIMA	SARIMA Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	9.10%	12.27%	7.18%
Avg Abs Var	207,225.3	288,264.8	157,304.4
Avg % Var	-2.14%	-4.05%	1.78%
Avg Variance	(62,834.2)	(111,489.1)	42,502.9
SBC	4505.77	4404.71	4335.46



Transformed Regression

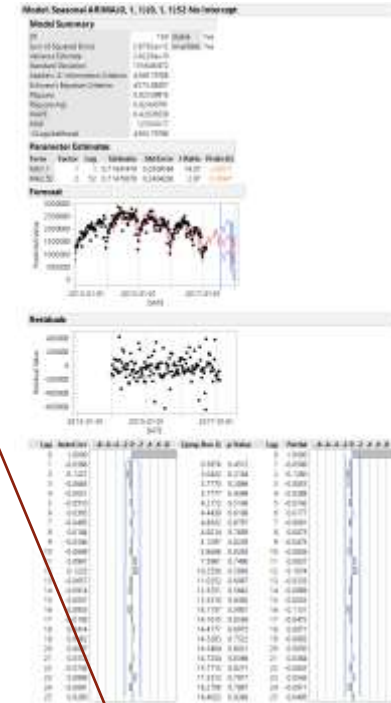
Conditional Least Squares Estimation						
Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag Variable	Shift
MA1,1	0.66174	0.06298	10.51	< .0001	1 SunFare	0
MA2,1	1.00000	0.09997	10.42	< .0001	52 SunFare	0
NUM1	-27303.7	18455.9	-1.48	0.1411	0 TempWeekly	0
NUM2	26218.2	15295.5	-1.84	0.0671	0 PrecipWeekly	1
NUM3	-12935.4	20465.1	-0.63	0.5283	0 SnowWeekly	0
NUM4	-5325.2	21343.2	-0.25	0.8033	0 WindWeekly	0
NUM5	11.16160	0.90652	12.31	< .0001	0 TopCo	0
NUM6	29.81438	6.07702	4.91	< .0001	0 BotCo	0

Precipitation, Co (Top & Bot) are statistically signif.

Conditional Least Squares Estimation						
Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag Variable	Shift
MA1,1	0.58744	0.05919	9.93	< .0001	1 SunFare	0
MA2,1	1.00000	0.10582	9.91	< .0001	52 SunFare	0
NUM1	-25718.4	26207.1	-1.13	0.2586	0 TempWeekly	0
NUM2	-38769.7	21796.9	-1.81	0.1001	0 PrecipWeekly	0
NUM3	-6572.5	29941.3	-0.22	0.8265	0 SnowWeekly	0
NUM4	4590.1	38445.6	0.15	0.8864	0 WindWeekly	0
NUM5	12.38449	0.91398	13.56	< .0001	0 TopCo	0
NUM6	25.65277	5.93410	4.40	< .0001	0 BotCo	0

Companies (Top & Bot) are statistically significant

Loop



Percip only weather determinant w./Stat.Sig.: - \$28k / wk in Loop



Areas

Airport Demand and Revenue



Airports & Loop Weekly Demand

Comparison of SARIMA
to Transformed
Regression using Top
Co, Outliers, and
Weather (icons shown)

Midway, O'Hare

Determinants:

Temp
Percip
Snow
Wind
TopCo


66% improved predictions of Demand
Transformed Regression models



	Airport 1 Dropoff Demand	Airport 1 Transformed Regression	Airport 1 SARIMA	Airport 2 Transformed Regression	Airport 2 SARIMA
Avg Abs % Var		8.03%	24.72%	5.70%	23.21%
Avg Abs Var		211.3	951.5	491.3	2,717.1
Avg % Var		0.06%	-23.20%	3.82%	-19.14%
Avg Variance		7.2	(906.9)	358.2	(2,387.3)
SBC		2021.621	2573.63	2238.55	2895.04

 -148/wk

	Airport 1 Dropoff Revenue	Airport 1 Transformed Regression	Airport 1 SARIMA	Airport 2 Transformed Regression	Airport 2 SARIMA
Avg Abs % Var		8.33%	24.39%	7.67%	23.30%
Avg Abs Var		6,803.6	28,074.1	25,614.9	100,741.8
Avg % Var		1.21%	-22.23%	-2.48%	-19.17%
Avg Variance		1,307.5	(26,048.6)	(6,149.9)	(88,061.5)
SBC		3236.727	3658.96	3636.526	4072.83



 \$2701/wk

	Airport 1 Pickup Demand	Airport 1 Transformed Regression	Airport 1 SARIMA (1,12)(0,1,1)52	Airport 2 Transformed Regression	Airport 2 SARIMA (0,1,2)(0,1,1)52
Avg Abs % Var		7.83%	25.52%	5.45%	13.83%
Avg Abs Var		304.7	1,373.7	835.8	2,392.5
Avg % Var		-4.05%	-25.44%	1.47%	-10.24%
Avg Variance		(154.0)	(1,370.5)	268.4	(1,849.9)
SBC		2039.095	2573.37	2333.075	2931.59

  -36/wk, 35/wk

 -243 / wk

	Airport 1 Pickup Revenue	Airport 1 Transformed Regression	Airport 1 SARIMA	Airport 2 Transformed Regression	Airport 2 SARIMA
Avg Abs % Var		6.08%	22.06%	8.57%	13.45%
Avg Abs Var		7,205.2	35,485.7	49,753.8	87,961.8
Avg % Var		0.11%	-21.73%	1.76%	-8.93%
Avg Variance		527.8	(35,085.4)	13,469.5	(60,493.4)
SBC		3281.752	3675.06	3689.695	4100.166

  \$3812/wk, \$2572/wk

60% improved predictions of Revenue
Transformed Regression models



North Revenue

Pickup T1&2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	14.32%	13.80%	4.77%
Avg Abs Var	45,208.3	-3.72%	(35,624.3)
SBC	4281.79	4203.38	4026.043

67% improved predictions in Transformed Regression models



Dropoff T1&2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	14.34%	13.86%	8.84%
Avg Abs Var	91,031.3	-6.09%	(63,281.5)
SBC	4283.19	4197.42	2291.96

38% improved predictions in Transformed Regression models



Pickup T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.80%	9.93%	8.13%
Avg Abs Var	7,598.2	-7.68%	(7,160.6)
SBC	3599.84	3521.59	3366.41

17% improved predictions in Transformed Regression models

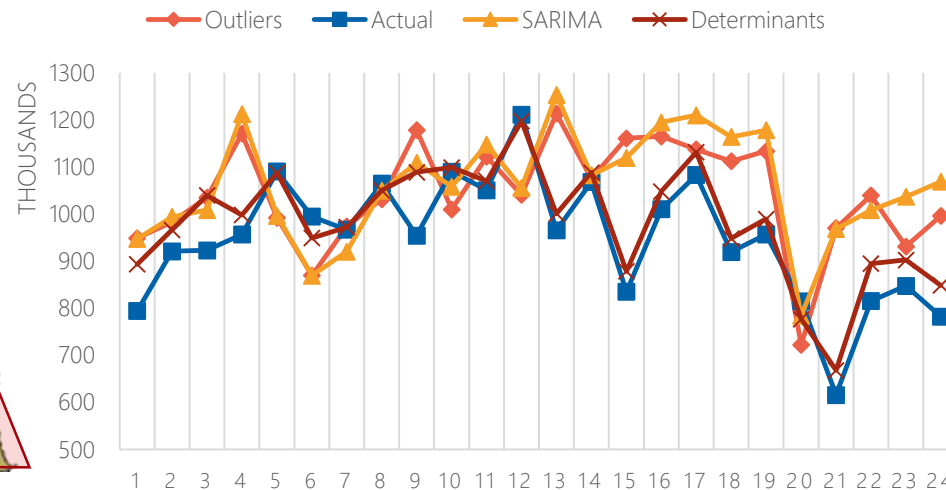


Dropoff T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.35%	8.15%	7.23%
Avg Abs Var	15,159.7	-4.60%	(9,381.8)
SBC	3744.09	3604.07	3575.553

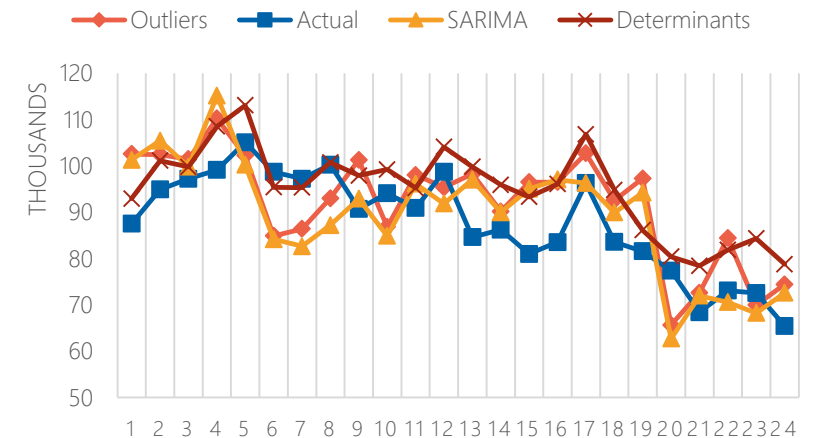
23% improved predictions in Transformed Regression models



24-week validation: Pickup Revenue T1&2



Pickup Revenue T2



North Weekly Pickup & Dropoff

SARIMA (0,1,1)(0,1,1)52:
SARIMA w/ Outliers:
Transformed Multi-Determinant

North, North 2nd tier

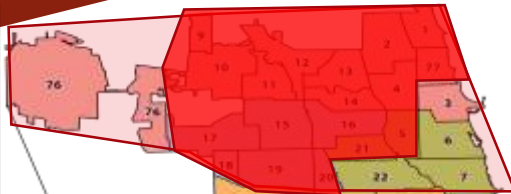
North Demand



North Weekly Pickup & Dropoff

SARIMA (0,1,1)(0,1,1)52:
SARIMA w/ Outliers:
Transformed Multi-Determinant

North, North 2nd tier



Pickup T1&2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	13.34%	12.27%	1.80%
Avg Abs Var	818.4	-0.86%	(368.2)
SBC	3396.13	3350.333	2834.705

85% improved predictions in Transformed Regression models



Pickup T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	10.67%	8.06%	3.41%
Avg Abs Var	244.0	-1.68%	(108.6)
SBC	2818.99	2753.69	2462.349

68% improved predictions in Transformed Regression models



Dropoff T1&2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	11.55%	10.76%	2.14%
Avg Abs Var	1,064.9	1.37%	(654.7)
SBC	3402.97	3342.82	2877.408

82% improved predictions in Transformed Regression models



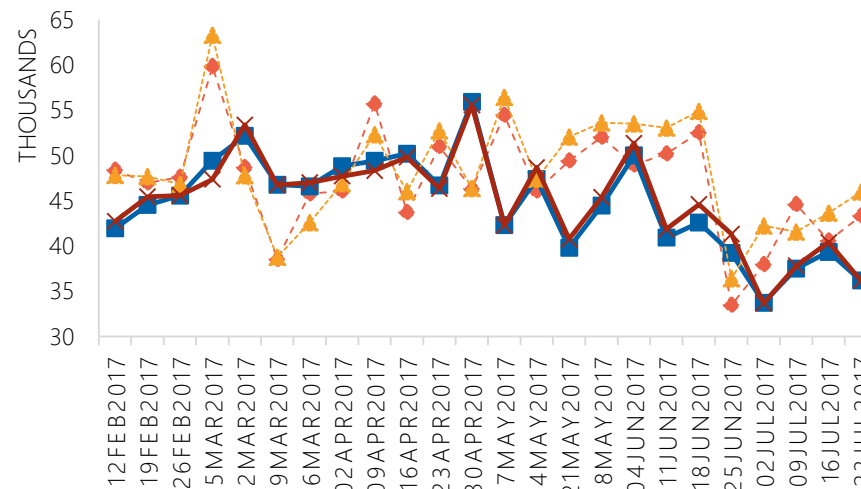
Dropoff T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.50%	8.63%	4.02%
Avg Abs Var	491.8	1.95%	(249.1)
SBC	2895.04	2827.914	2291.96

58% improved predictions in Transformed Regression models



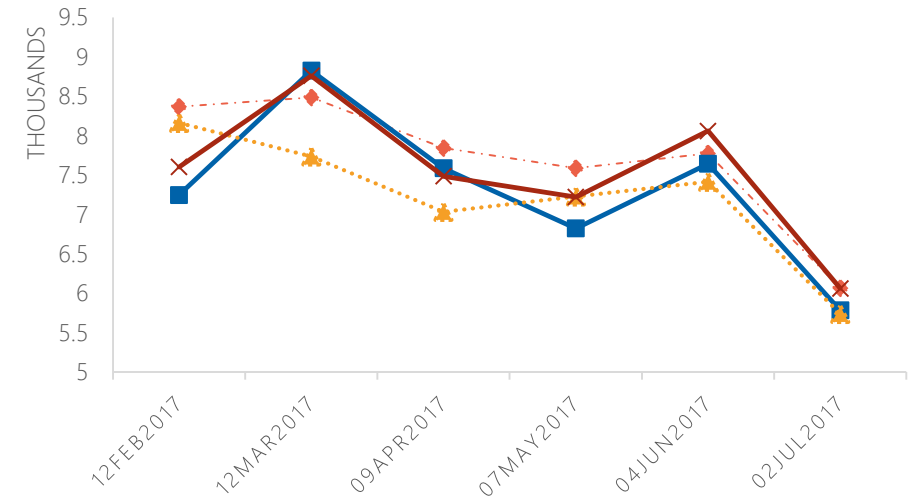
24-week validation: Pickup Demand T1&2

w/ Outliers Actual SARIMA Transformed Regression



Pickup Demand T2 (MO.)

w/ Outliers Actual SARIMA Transformed Regression



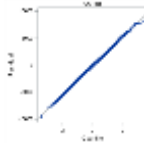
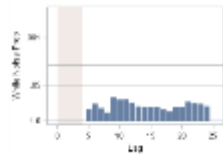
Chicago Total Daily



City-wide
Daily

SARIMA (2,1,1)(0,1,1)7:
SARIMA w/ Outliers:
Transformed Multi-
Determinant

Demand (CHICAGO)	SARIMA (2,1,1)(0,1,1)7	SARIMA Outliers	Transformed Regression Multi- Determinant
Avg Abs % Variance	7.31%	26.39%	3.75%
Avg Abs Var	2,293.2	5,745.4	1,024.3
Avg % Var	-7.15%	26.39%	-2.29%
Avg Variance	(2,258.4)	5,745.4	(622.4)
SBC	32,679.91	31,618.22	31,614.48
Outliers		Transformed Regression	



Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag	Variable	Shift
MA1,1	0.00011	0.0000179	0.00011	< .0001	1	Demand	0
MA2,1	0.00001	0.0000179	0.00001	< .0001	7	Demand	0
AR1,1	0.00001	0.0000179	0.00001	< .0001	1	Demand	0
AR1,2	0.00001	0.0000179	0.00001	< .0001	2	Demand	0
NUM1	0.00001	0.0000179	0.00001	< .0001	0	AC439	0
NUM110	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM111	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM112	0.00001	0.0000179	0.00001	< .0001	0	AC1367	0

White Noise Hypothesis is Validated in
DEMAND models using residuals

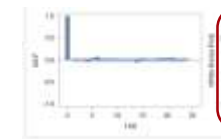
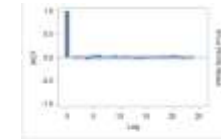


Stat significant effects on Daily Demand:
Weekday +7070 trips, Precipitation + 2054 trips
Temperature + 822 trips



Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag	Variable	Shift
MA1,1	0.00011	0.0000179	0.00011	< .0001	1	Demand	0
MA2,1	0.00001	0.0000179	0.00001	< .0001	7	Demand	0
AR1,1	0.00001	0.0000179	0.00001	< .0001	1	Demand	0
AR1,2	0.00001	0.0000179	0.00001	< .0001	2	Demand	0
NUM1	0.00001	0.0000179	0.00001	< .0001	0	AC439	0
NUM110	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM111	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM112	0.00001	0.0000179	0.00001	< .0001	0	AC1367	0

Revenue (CHICAGO)	SARIMA Outliers	Transformed Regression Multi- Determinant	SARIMA (2,1,1)(0,1,1)7
Avg Abs % Variance	39.20%	6.15%	9.02%
Avg Abs Var	96,530.6	24,064.6	27,606.4
Avg % Var	39.20%	-5.53%	7.88%
Avg Variance	96,530.6	(22,103.8)	23,296.8
SBC	39,710.15	39,710.15	40,540.20



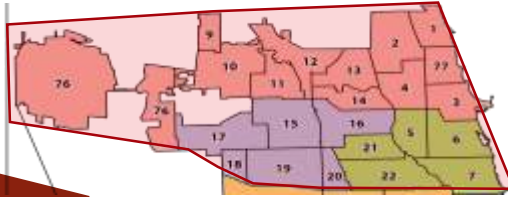
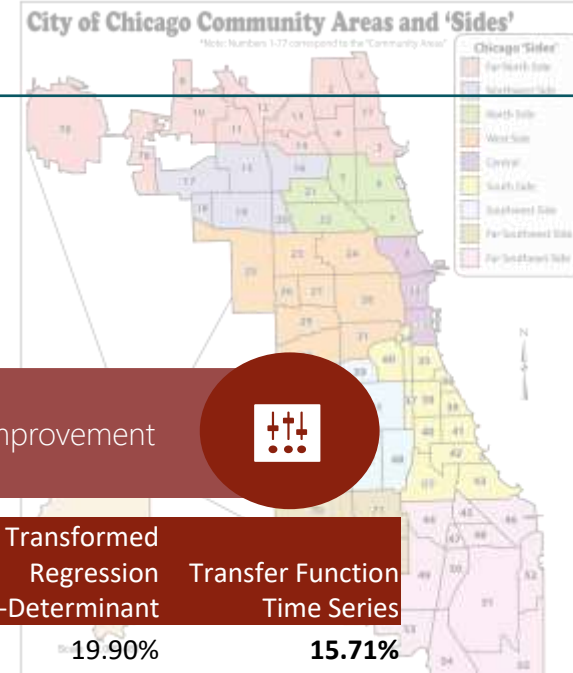
White Noise Hypothesis is Violated in model
Transformed Regressions



Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag	Variable	Shift
MA1,1	0.00011	0.0000179	0.00011	< .0001	1	Revenue	0
MA2,1	0.00001	0.0000179	0.00001	< .0001	7	Revenue	0
AR1,1	0.00001	0.0000179	0.00001	< .0001	1	Revenue	0
AR1,2	0.00001	0.0000179	0.00001	< .0001	2	Revenue	0
NUM1	0.00001	0.0000179	0.00001	< .0001	0	AC439	0
NUM110	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM111	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM112	0.00001	0.0000179	0.00001	< .0001	0	AC1367	0

Parameter	Estimate	Standard Error	t Value	Approx. Pr > t	Lag	Variable	Shift
MA1,1	0.00011	0.0000179	0.00011	< .0001	1	Revenue	0
MA2,1	0.00001	0.0000179	0.00001	< .0001	7	Revenue	0
AR1,1	0.00001	0.0000179	0.00001	< .0001	1	Revenue	0
AR1,2	0.00001	0.0000179	0.00001	< .0001	2	Revenue	0
NUM1	0.00001	0.0000179	0.00001	< .0001	0	AC439	0
NUM110	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM111	0.00001	0.0000179	0.00001	< .0001	0	AC184	0
NUM112	0.00001	0.0000179	0.00001	< .0001	0	AC1367	0

Areas Northern & Western Daily



12.2% improvement

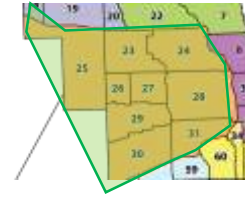


Revenue (North)	SARIMA (3,1,3)(0,1,1)7	Transformed Regression Multi-Determinant	Transfer Function Time Series
Avg Abs % Variance	12.23%	9.17%	8.05%
Avg Abs Var	12,585.3	9,626.4	8,396.2
Avg % Var	10.17%	8.63%	6.82%
Avg Variance	10,243.6	9,043.5	7,037.8
SBC	37655	35718	35634

Transfer Function Inputs for North Daily (trans):
Efficiency Input -64.7% prediction accuracy
Top Company Input -16.4% prediction accuracy



Demand (North)	SARIMA (2,1,1)(0,1,1)7	Transformed Regression Multi-Determinant	Transfer Function Time Series
Avg Abs % Variance	7.07%	1.82%	1.83%
Avg Abs Var	390.5	94.7	97.7
Avg % Var	-2.62%	0.22%	-0.87%
Avg Variance	(156.5)	1.3	(54.3)
SBC	21195	23815	23598



21.1% improvement



Revenue (West)	SARIMA (2,1,1)(0,1,1)7	Transformed Regression Multi-Determinant	Transfer Function Time Series
Avg Abs % Variance	24.92%	19.90%	15.71%
Avg Abs Var	806.3	634.1	575.1
Avg % Var	15.19%	18.05%	3.08%
Avg Variance	472.0	565.3	106.2
SBC	29880.71	27527.36	27580.48

Demand (West)	SARIMA (2,1,1)(0,1,1)7	Transformed Regression Multi-Determinant	Transfer Function Time Series
Avg Abs % Variance	20.77%	10.64%	7.58%
Avg Abs Var	47.9	24.0	19.5
Avg % Var	13.54%	8.90%	3.37%
Avg Variance	29.8	19.4	7.8
SBC	21194.78	17832.51	18303.44

28.7% improvement



Pickup
Demand and
Revenue

SARIMA (2,1,1)(0,1,1)7:
SARIMA w/ Outliers:
Transformed Multi-
Determinant:
Transfer Function w/
Top Company

Summary



Industry / Business Context

Disrupted Model
Business Questions:
Revenue & Demand



Data Collection

Chicago Data Portal
NOAA Weather
Parse and Aggregate



Data Cleaning

113M observations
2 Dependent
Independent Variables:
5 Time/Day
5 Weather
4 Area
4 Misc.



Model Exploration / Selection

Time-Series SARIMA
SARIMA w/ Outliers
Transformed Regress.
Transfer Function T-S
Dependent Variables:
1) Fares
2) Fare \$'s



Summary

Model and professional
summaries of project

Summary – Report Conclusion



Summary

Best Model(s)

Further Studies

Data Insights: BEST MODELS

- BEST MODEL Transfer Function Time Series (Conditionally)
 - Outperformed Transformed Regression Model in West Daily w/ Top Company Input
 - Performed marginally better in North Daily w/ Top Transfer Input
 - Validated models compared to attempted models was a small %

Model Choice

- Transformed Regression Model using Top Company
 - Improved predictions of over 80% in the North (Average Absolution Variance %)
 - Better than SARIMA and SARIMA w/ outliers, except when the model was unstable.
 - Independent Variables with Statistically Significant: top transfer, top company, bottom company, weather, holidays, weekdays have statistical significance under certain conditions. High VIFs with Top Company and Top Transfer resulted in one or other.

Next Steps: Top Company / Top Transfer in Complex Models

- Studio Forecast with MECE data, efficiencies, and more granularly binned datasets for more complex models
- Machine learning to frame complex data analytics into production-level models:
 - Bootstrap variables w/ more collectively exhaustive dataset in Random Forest (repeat Studio models)
 - Bayesian Statistics using Top Company / Transfer in Neural Networks to test impact in sliced temporal spatial data

Summary

Professional Impact



Summary

Challenges:
Big Data – handling,
cleaning, processes

Professional
Development:
Time-series, outputs,
processes

Overcame Major Challenges:

- Parceling large data and pulling through variables requires forethought
- Cleaning data in a 'dark room' requires testing
- Major trends and averages make repeatable models, but validations are required for in stable and reliable predictions. Validation stopped overfitting models a lot.

Professional Development

- Time-series statistics acumen advanced. Tools, approaches and methods in time-series are at levels of Lead Contributor / Manager.
- Gained skills to derive insights from analytics in reportable / digestible formats.
- Ability to think about large data structures and work through hard problems into manageable sizes – with meaningful output for audience.



Thank You