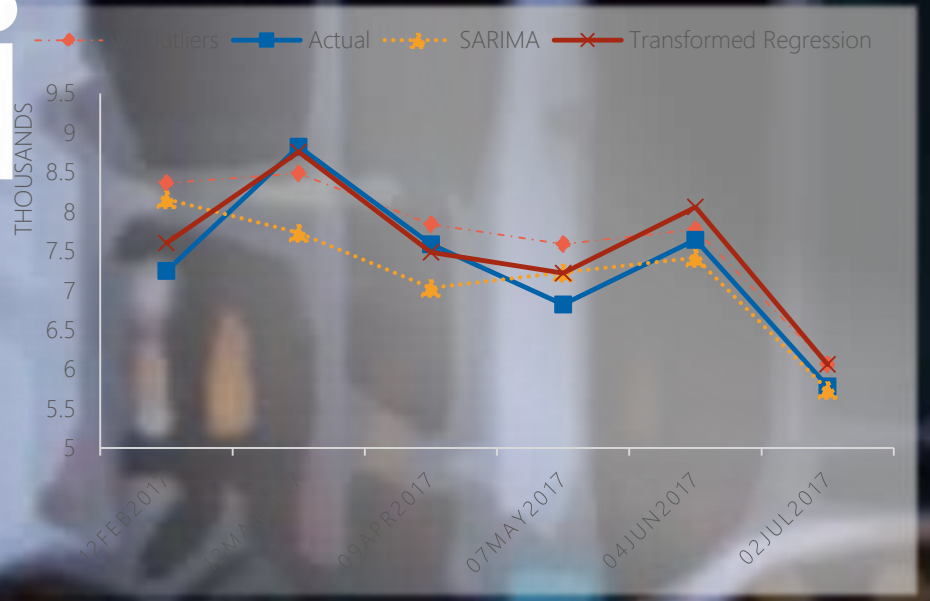
Chicago Taxi

Nathaniel Green



Texas A&M Masters of Data Analytics Capstone Project

Business Context

Business Context

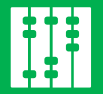
Disrupted Model
Business Questions:
Revenue & Demand

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

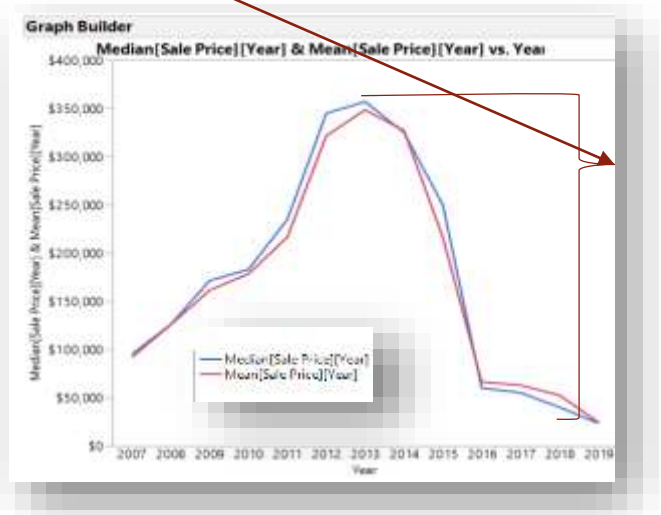
Industry Questions: Level of Disruption Sunsetting on Taxi License Business

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 fi
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/medallionowners/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.

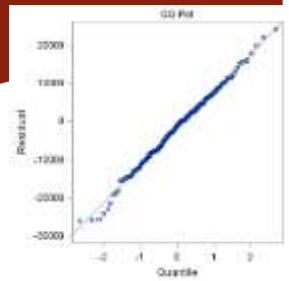
Values context:
Daily Revenue: Mean = $862k : Median = $889k
Daily Demand: Mean = 67.4k trips : Median = 69.2k trips

Taxi Data:
Executive Summary – Findings

Executive Summary: Finding

Process Keynotes

Independent Variables Used to Predict Demand and Revenue



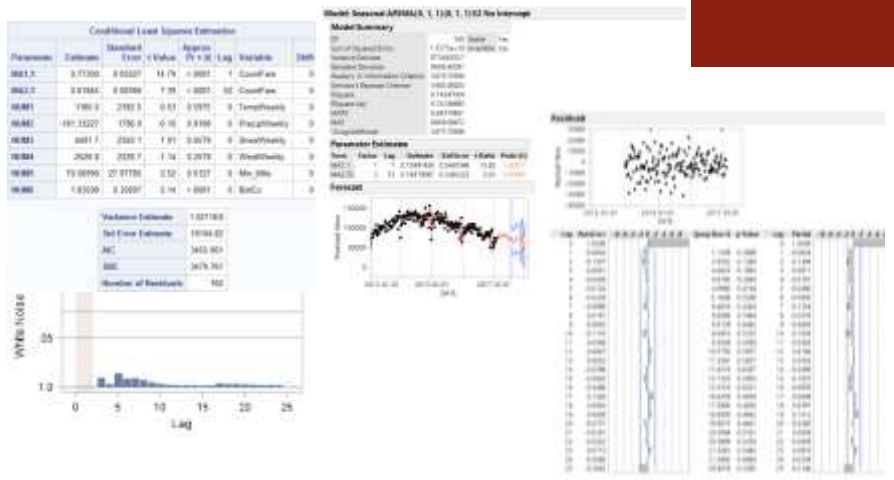
Weekly SARIMA (0,1,1)(0,1,1) 52

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

Companies and transfers were most useful coefficients for predictions

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

Weather coefficients were significant in Transformed Regression

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

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

Data Collection

Chicago Data Portal

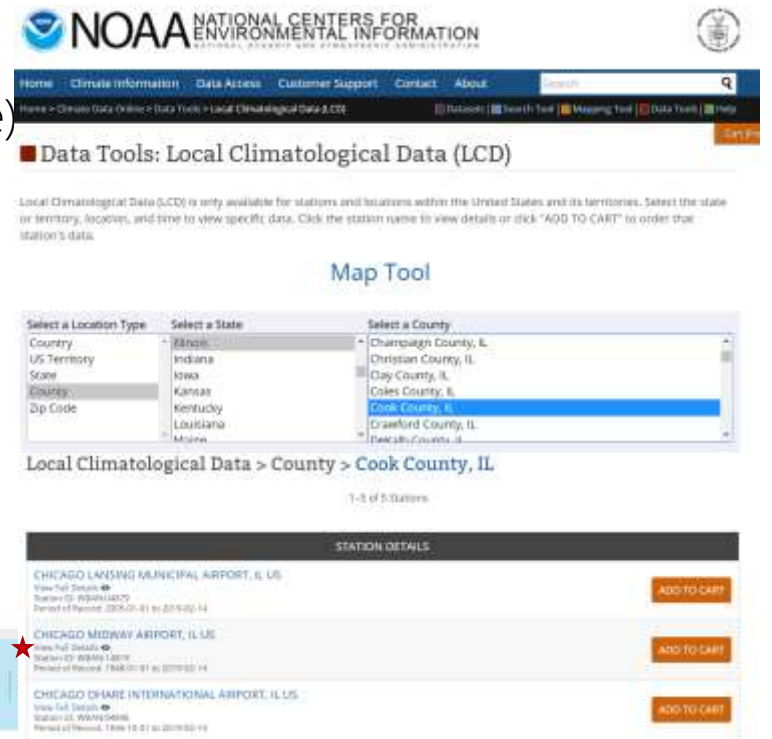
NOAA Weather Data

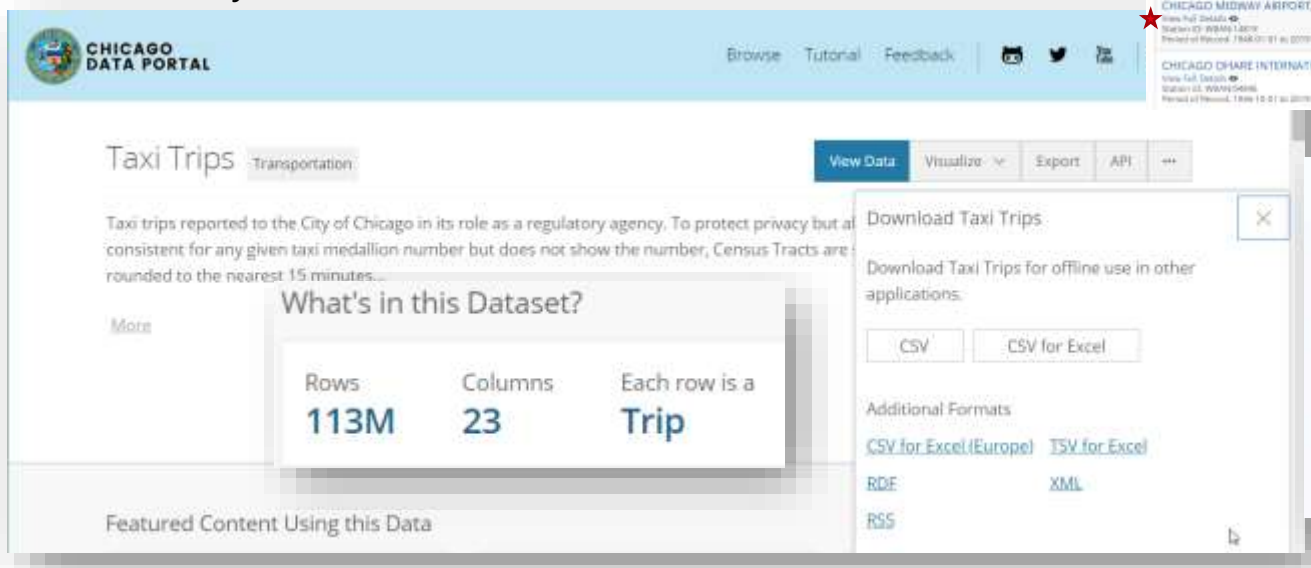
Chicago Data Portal (https://data.cityofchicago.org/)

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)

Weather Data (https://www.ncdc.noaa.gov/cdo-web/datasets/)

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





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

Data Collection

Data and Fare Summaries

Characteristics 100% of data | Fare characteristics using 1% random sample

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

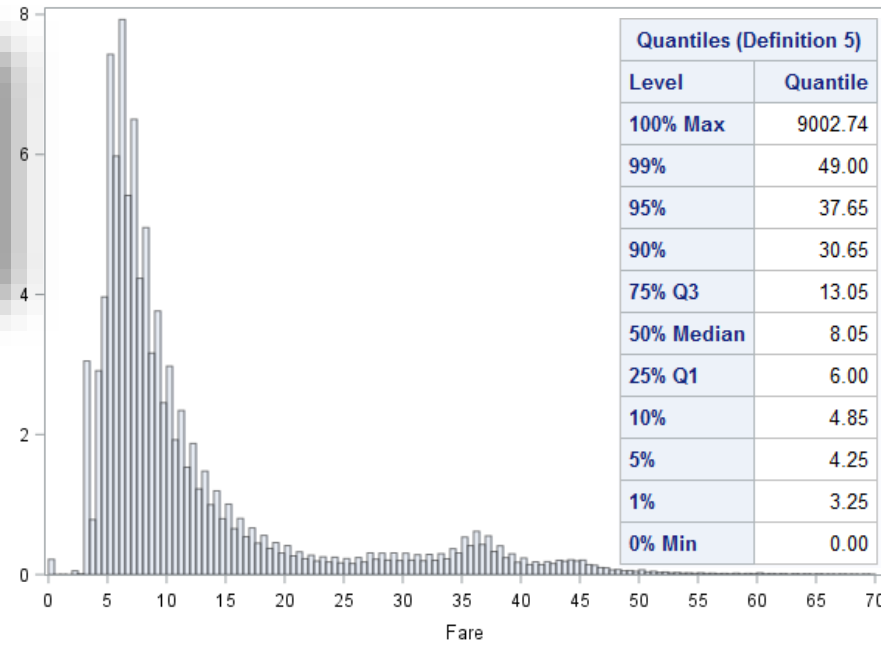
Missing as a percentage of All Data:

	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%

Moments

N	1128585	Sum Weights	1128585
Mean	12.732341	Sum Observations	14369529
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



Quantiles (Definition 5)

Level	Quantile
100% Max	9002.74
99%	49.00
95%	37.65
90%	30.65
75% Q3	13.05
50% Median	8.05
25% Q1	6.00
10%	4.85
5%	4.25
1%	3.25
0% Min	0.00

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

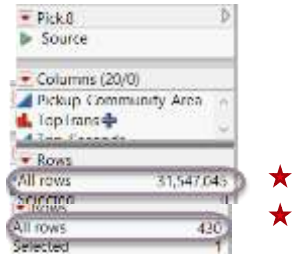
Fares Summarized as Revenue

Top 12 Demand Areas

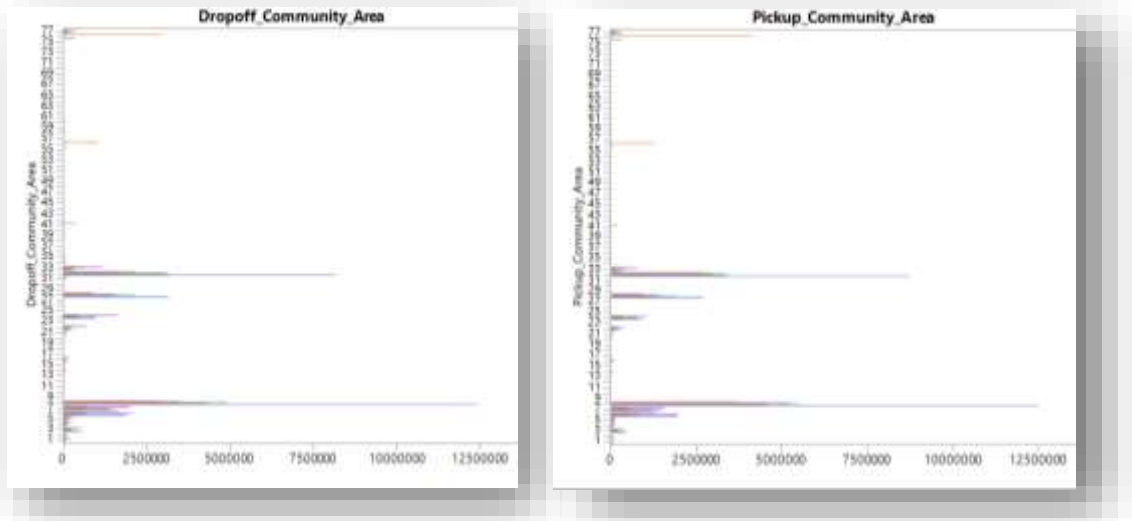
Pickup_Community_Area	% of Total	N
8	33.41%	32,104,338
32	22.35%	21,477,142
28	8.83%	8,481,004
6	7.09%	6,815,397
✈ 76	6.09%	5,853,830
7	5.70%	5,477,208
24	3.72%	3,575,711
33	2.29%	2,204,222
3	1.71%	1,643,982
✈ 56	1.69%	1,619,459
22	1.23%	1,184,590
77	1.12%	1,073,453

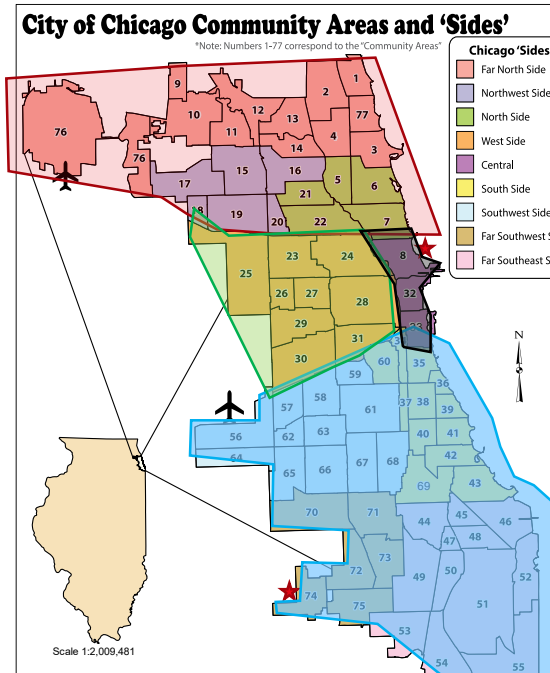
Dropoff_Community_Area	% of Total	N
8	30.08%	28,201,515
32	19.71%	18,480,324
28	9.88%	9,267,949
6	7.49%	7,021,590
7	6.69%	6,275,255
24	4.83%	4,532,021
✈ 76	3.78%	3,540,993
33	3.02%	2,831,048
3	2.04%	1,912,080
22	1.87%	1,757,540
✈ 56	1.31%	1,228,219
77	1.28%	1,196,975

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

Data Cleaning

Taxi Data

Chose to remove and adjust missing and outliers.

1. Removed Observations with Missing Community Areas (Total)
    1. Both to and from for better accountability of Chicago roundtrip fares.
        1. Transfer areas useful as model predictors.
    2. Trends for Demand and Revenue Across 4.5 years for these 19.75M observations
        1. Trailing 12 months drops off significantly
    3. 50% of other outliers included missing community areas
        1. 300% increase in outliers between missing community population compared to total population

2. Removed Observations with Missing and Outliers in Other Variables
    1. Missing Fares, Fares = 0, and Fares >= 500 observations = 260k
        1. 148k observations without Communities (57%)
    2. Trip miles < 90 observations = 69k
        1. 13k observations without Communities (20%)
    3. Trip Seconds < 20k (5hours) observations = 42k
        1. 16k observations without Communities (38%)

3. Now What? Adjustments to Post-Processed Data:
    1. Add 10% to Demand for most area model outputs
        1. Over time the percentages between the total data set and cleaned data set have narrowed:
        2. 22.3% greater demand in city in year 0-1, compared to 11.2% for trailing year, and 10.3% for last 90 days
    2. Characteristics of available community area
        1. Pickup skewed towards airports – suggest airports add 20% pickup demand

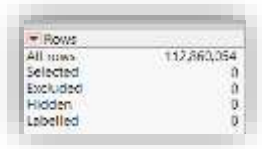
Removed taxi data

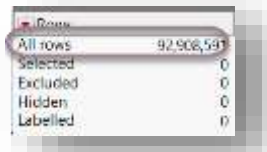
- Fare between 1 and 499
- Trip_Miles < 90
- Dropoff_Community_Area > 0
- Pickup_Community_Area > 0
- Trip_Seconds < 20000
- Sort by Time and Data

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

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%





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)

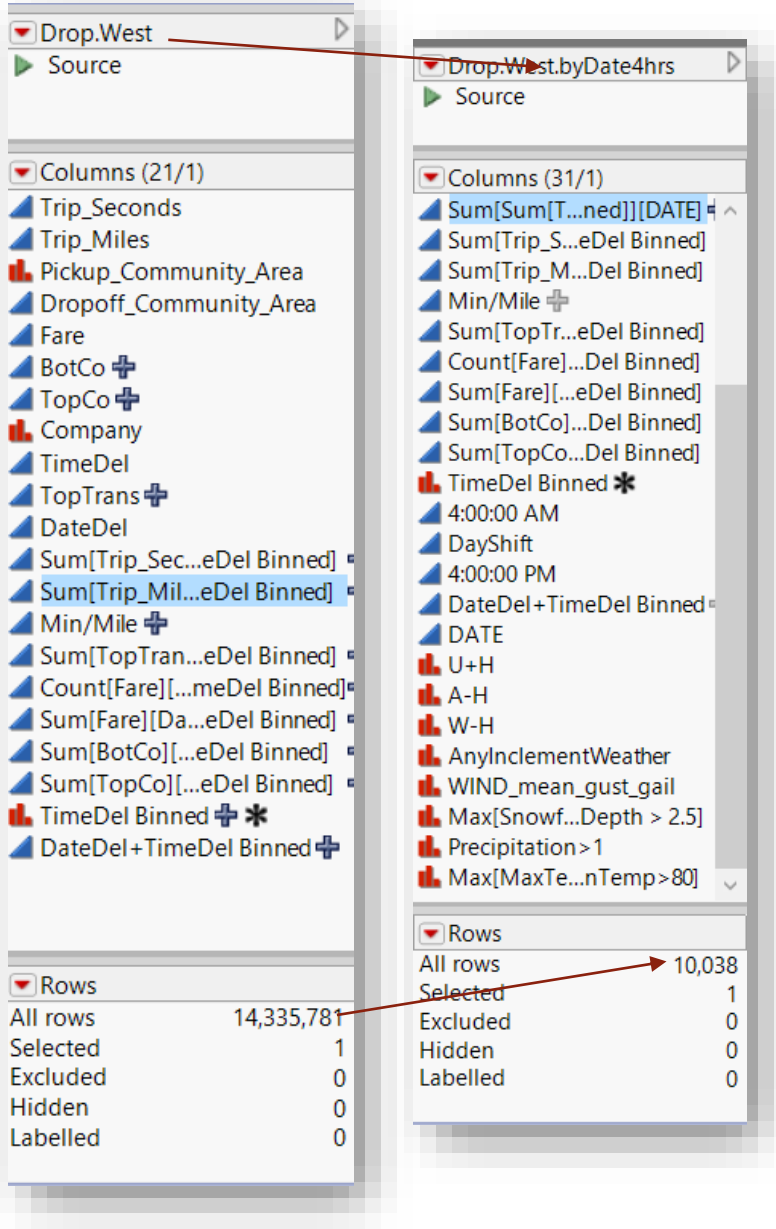
Created time-based bins and variables:

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



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

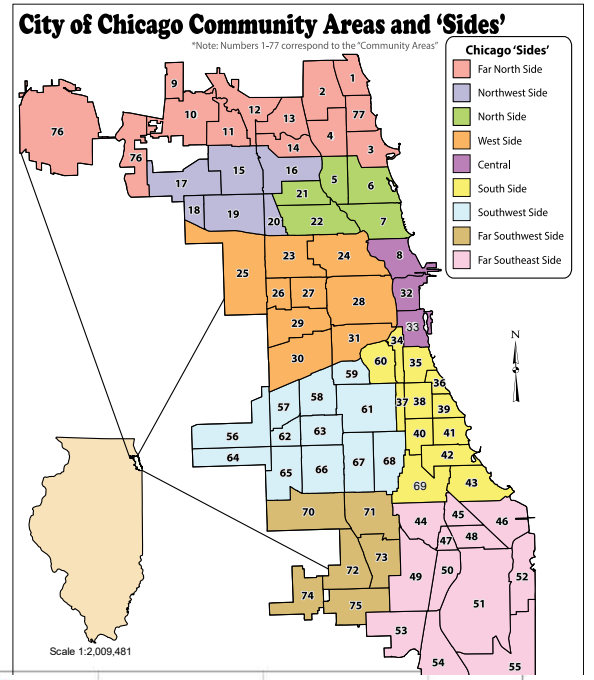
Model and professional summaries of project

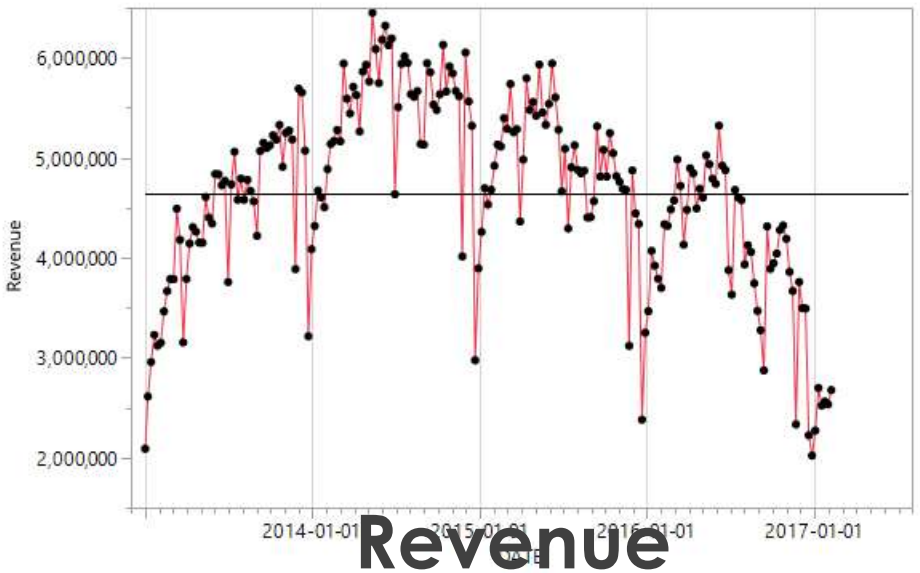
Model Roadmap

SARIMA Time-Series

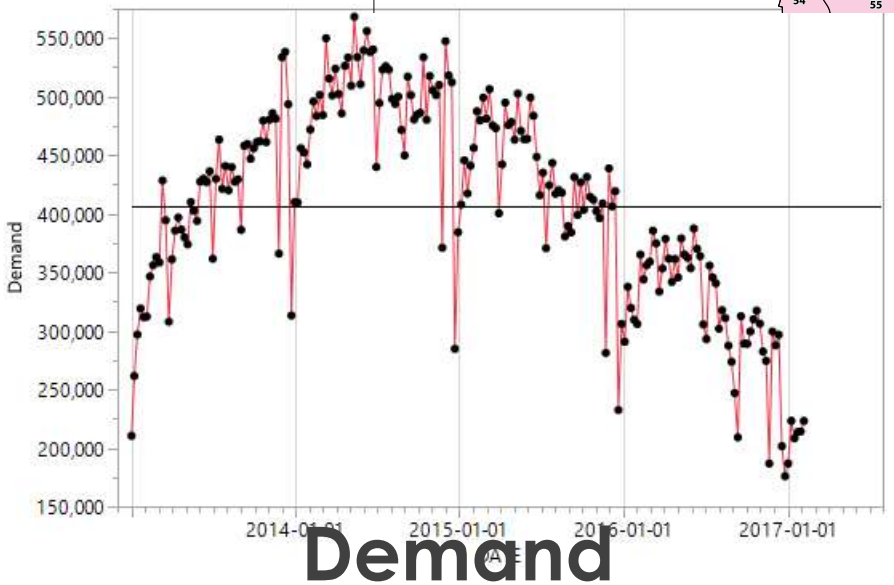
Model outlines:
Inputs / Validity Tests
Effects on Targets (responses, outliers, etc)
Model Takeaways

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)





Revenue



Demand

Chicago Total Demand Weekly

City-wide Demand Weekly

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

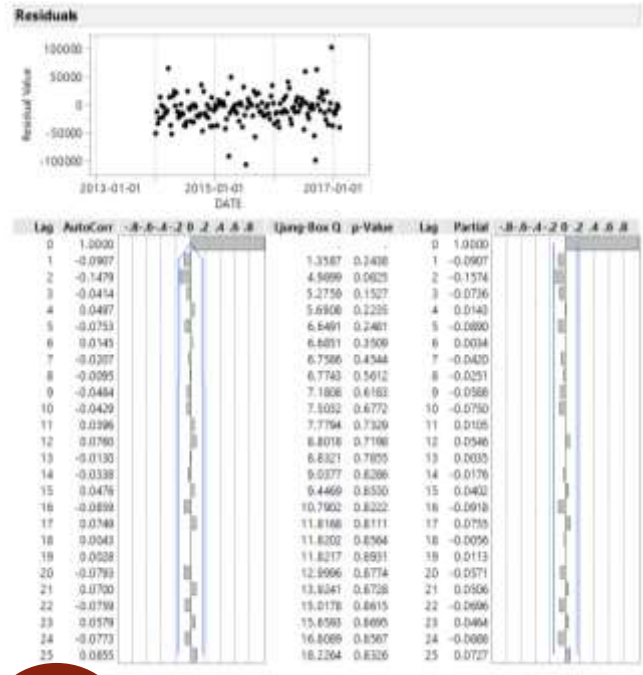
Compared to:

SAS Multi-Determinant Transformed Regression w/ Outliers

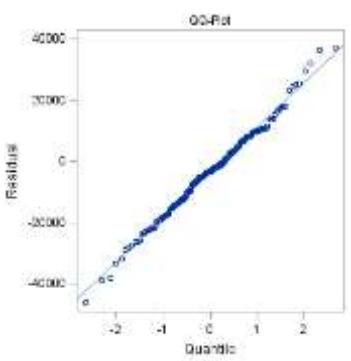
Demand (24 period)	SARIMA	Transformed Regression
Avg Absolute Variance %	10.04%	**9.85%**
Avg Abs Var	24,348.6	**23,717.1**
Avg % Var	-4.86%	**-4.50%**
Avg Variance	(12,265.6)	**(11,499.5)**
SBC	3788.66	**3697.423**

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

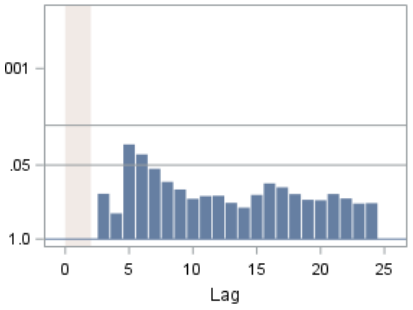


Transformed Regression

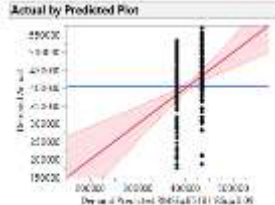


Maximum Likelihood Estimation

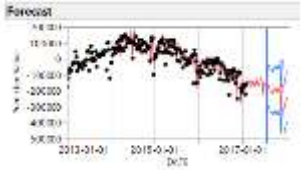
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.57551	0.07070	8.14	< .0001	1	Demand	0
MA2,1	0.43415	0.12314	3.53	0.0004	52	Demand	0
NUM1	1372.9	3396.7	0.40	0.6861	0	TempWeekly	0
NUM2	267.02454	2879.6	0.09	0.9261	0	PrecipWeekly	0
NUM3	10804.0	3959.4	2.73	0.0064	0	SnowWeekly	0
NUM4	-441.87358	3411.8	-0.13	0.8970	0	WindWeekly	0
NUM5	-87167.4	12311.0	-7.08	< .0001	0	AO133	0
NUM18	40110.4	14051.8	2.85	0.0043	0	AO204	0
NUM19	-41282.0	12297.0	-3.36	0.0008	0	AO95	0
I20	33926.6	12993.0	2.61	0.0090	0	LS45	0



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



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



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

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

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**

City-wide Revenue Weekly

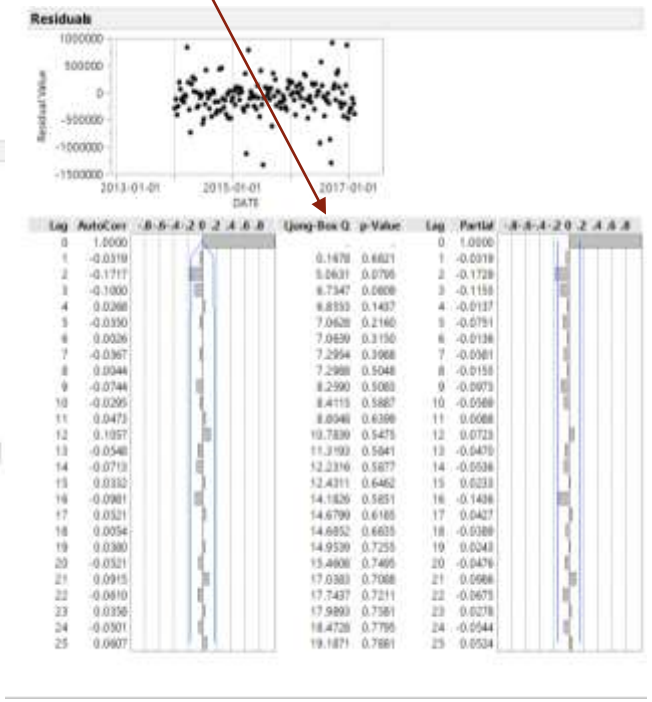
SARIMA (0,1,1)(0,1,1)52 Time-Series

Compared to

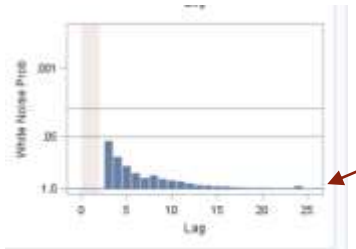
SAS Multi-Determinant Transformed Regression w/ Outliers

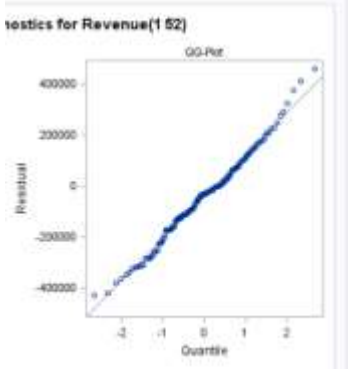
White Noise Hypothesis is Validated w./ autocorrelated & partial residuals.

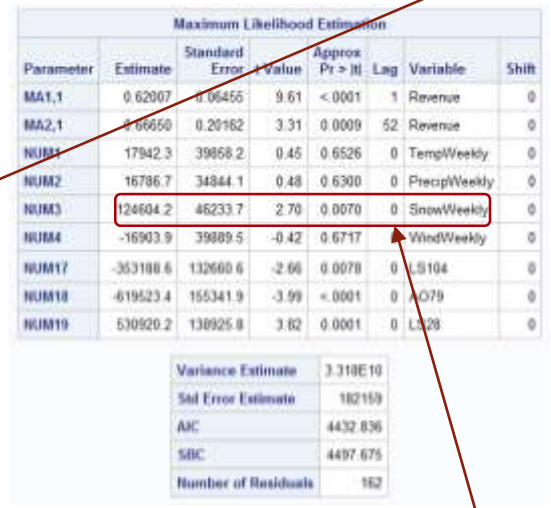
SARIMA

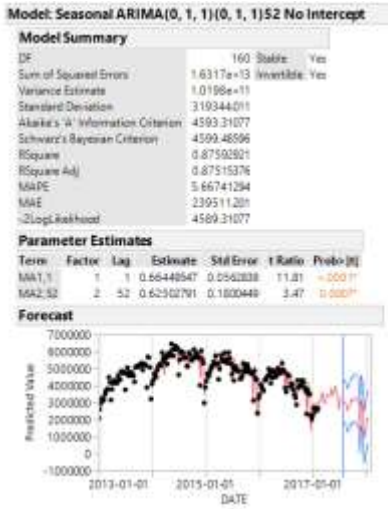


Transformed Regression









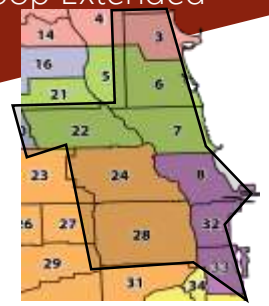
Snow statistically significant effects of + $125k / wk
Outliers and Determinants improved performance model 25% in t+1 – t+6 (t+1 – t+24 shown).

Loop Demand

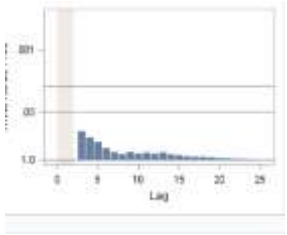
Loop Weekly Demand Drop-off, Comparing:

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

Loop
Loop Extended

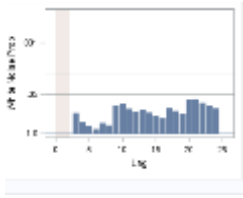


Demand Loop	SARIMA	SARIMA w/ Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	11.23%	**11.09%**	14.04%
Avg Abs Var	16,725.4	**15,952.9**	19,165.6
Avg % Var	-7.50%	**-5.21%**	14.04%
Avg Variance	(11,444.3)	**(7,964.9)**	19,165.6
SBC	3617.97	**3546.48**	3363.49



29% improved predictions in Multi-determinant Regression models

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**
	3485.88	3441.89	**3437.1**



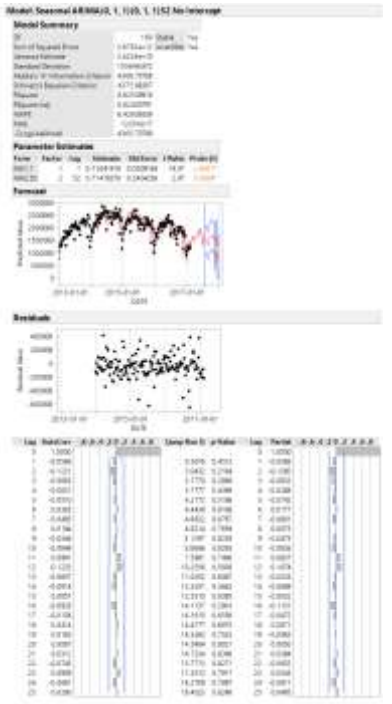
White Noise Hypothesis is Validated in models using auto and partial residuals

Transformed Regression

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

Top company statistically signif. effect + .385 per unit

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



Areas 8 – Drop-off Tabulation

Company	N
Taxi Affiliation Services	5942894
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%

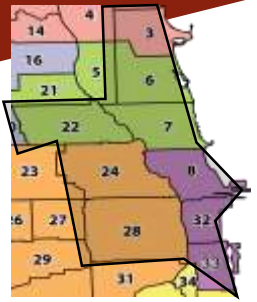
45% & 21% improved predictions in Transformed Regression models

Loop Revenue

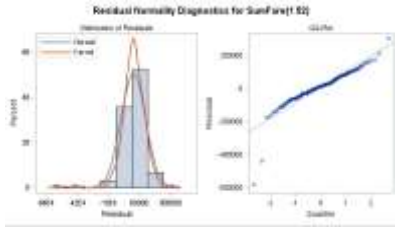
Loop Weekly Revenue Drop-off, Comparing:

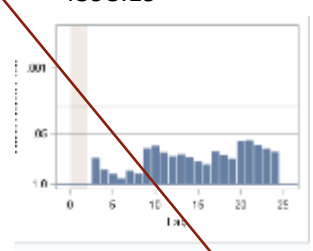
SARIMA (0,1,1)(0,1,1)52:
SARIMA w/ Outliers:
Transformed Multi-Determinant

Loop
Loop Extended

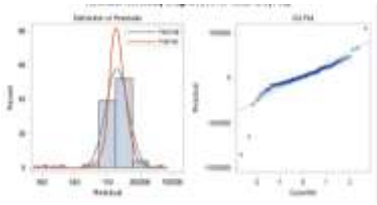


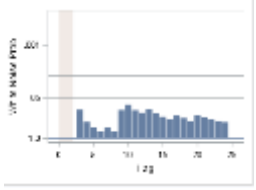
Revenue Loop	SARIMA	SARIMA Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	14.58%	13.49%	**8.04%**
Avg Abs Var	245,154.3	222,529.4	**118,773.4**
Avg % Var	-10.74%	-8.86%	**-1.33%**
Avg Variance	(186,145.4)	(152,532.2)	**(18,787.8)**
SBC	4375.88	4398.19	**4214.09**





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**



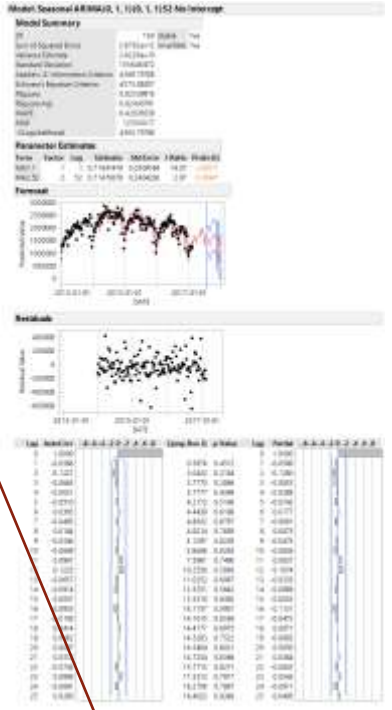


Transformed Regression

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

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

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

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

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

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

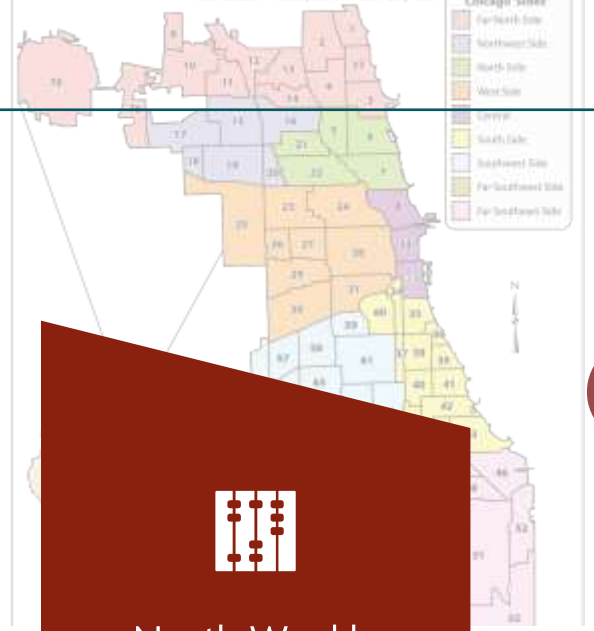
$3812/wk, $2572/wk

Determinants:
Temp
Percip
Snow
Wind
TopCo

66% improved predictions of Demand Transformed Regression models

60% improved predictions of Revenue Transformed Regression models

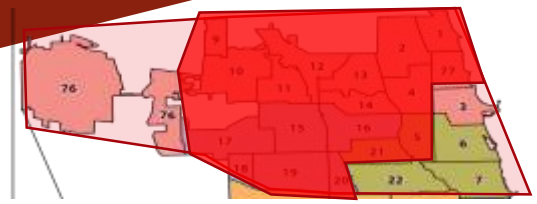
North Revenue



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	14.32%	13.80%	**4.77%**
			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%**
			91,031.3
			-6.09%
			(63,281.5)
SBC	4283.19	4197.42	**2291.96**

38% improved predictions in Transformed Regression models

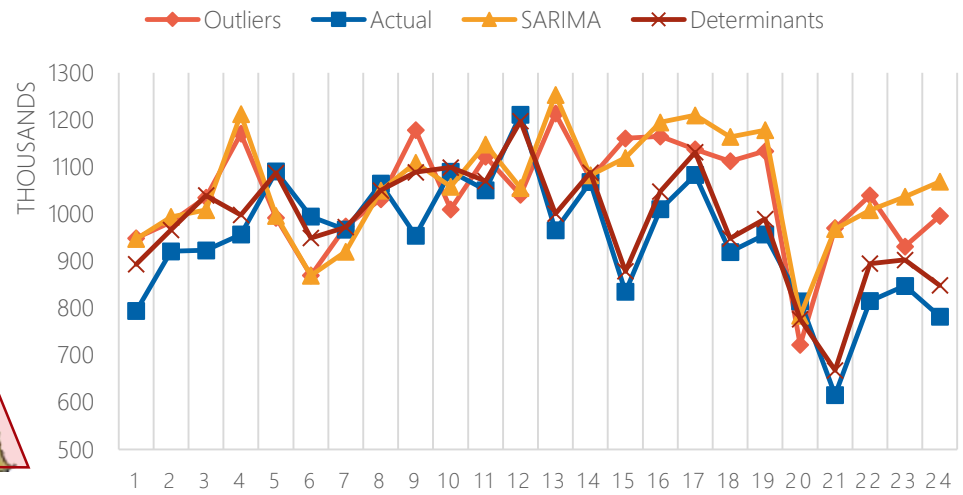
Pickup T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.80%	9.93%	**8.13%**
			7,598.2
			-7.68%
			(7,160.6)
SBC	3599.84	3521.59	**3366.41**

17% improved predictions in Transformed Regression models

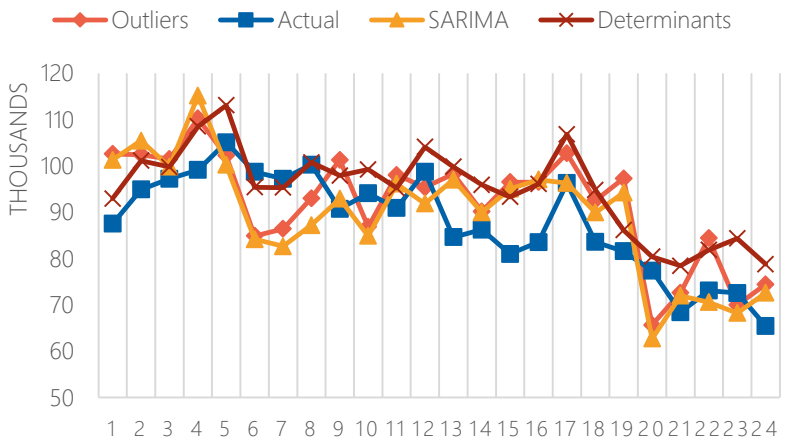
Dropoff T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.35%	8.15%	**7.23%**
			15,159.7
			-4.60%
			(9,381.8)
SBC	3744.09	3564.07	**3575.553**

23% improved predictions in Transformed Regression models

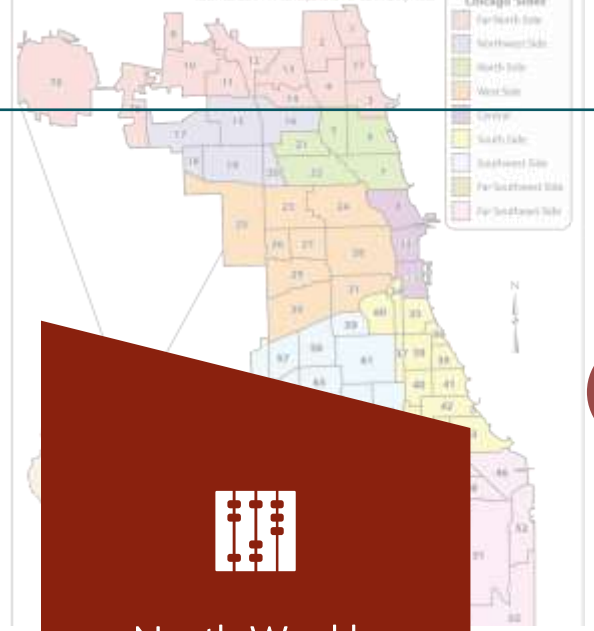
24-week validation: Pickup Revenue T1&2



Pickup Revenue T2



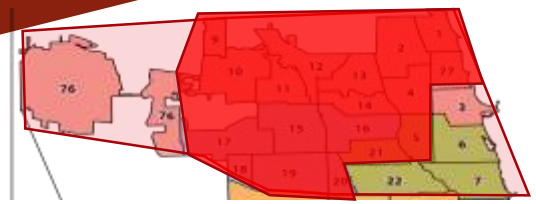
North Demand



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%**
			818.4
			-0.86%
			(368.2)
SBC	3396.13	3350.333	**2834.705**

85% improved predictions in Transformed Regression models

Dropoff T1&2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	11.55%	10.76%	**2.14%**
		4	**1,064.9**
		%	**1.37%**
		4.2)	**654.7**
SBC	3402.97	3342.82	**2877.408**

82% improved predictions in Transformed Regression models

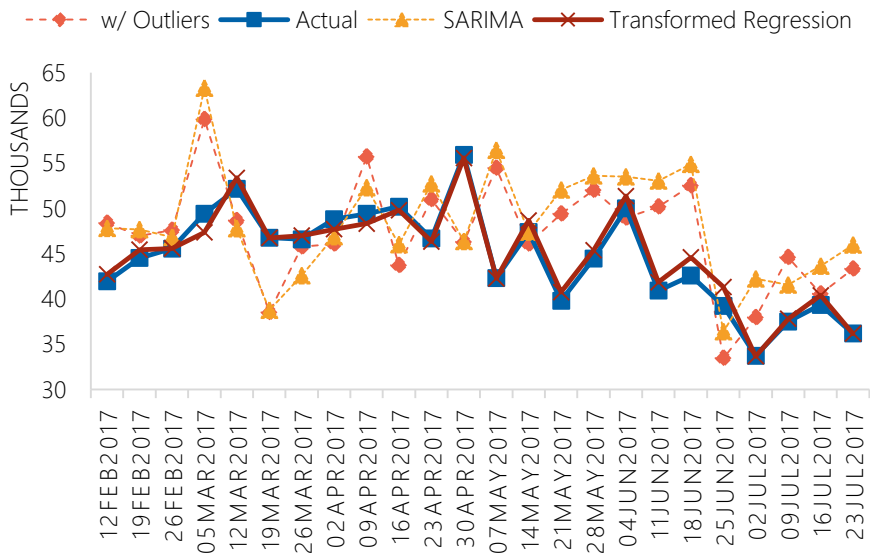
Pickup T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	10.67%	8.06%	**3.41%**
			244.0
			-1.68%
			(108.6)
SBC	2818.99	2753.69	**2462.349**

68% improved predictions in Transformed Regression models

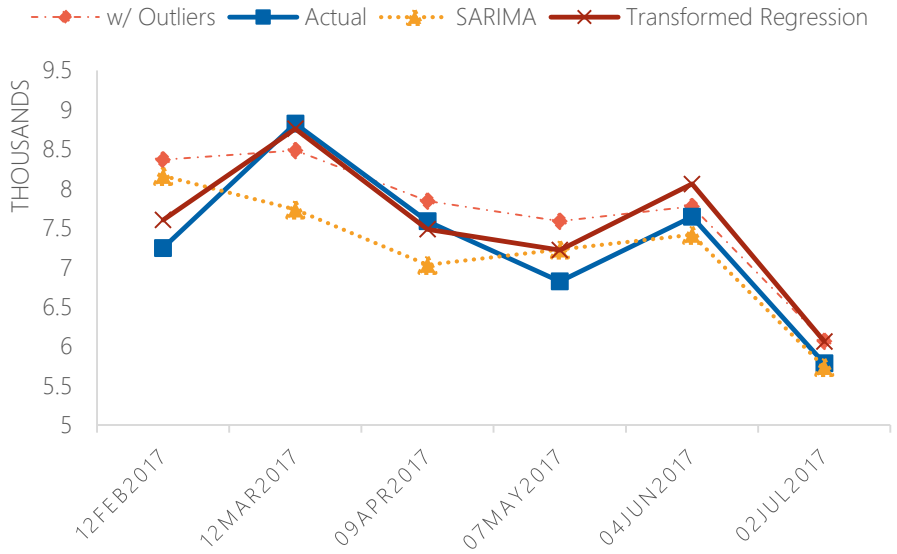
Dropoff T2 Areas	SARIMA	w/ Outliers	Multi-determinant
Avg Abs % Variance	9.50%	8.63%	**4.02%**
			491.8
		%	1.95%
			249.1
SBC	2895.04	2827.914	**2291.96**

58% improved predictions in Transformed Regression models

24-week validation: Pickup Demand T1&2



Pickup Demand T2 (MO.)



Chicago Total Daily

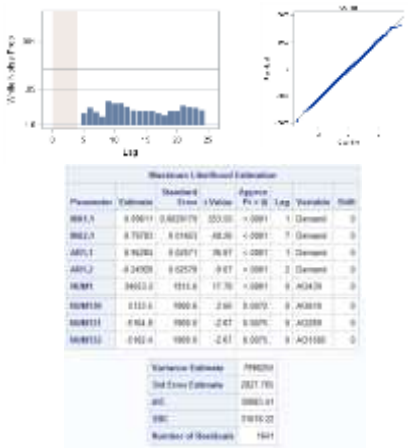
City-wide Daily

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

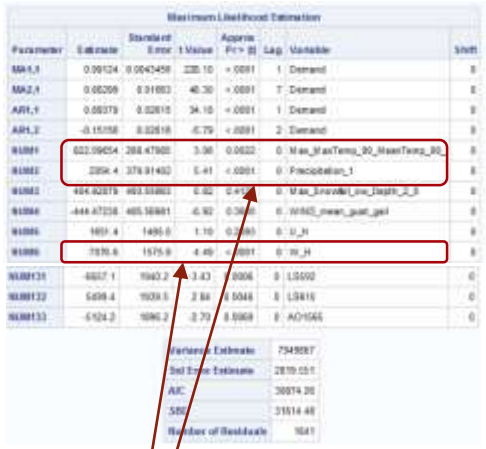
Demand (CHICAGO)	SARIMA (2,1,1)(0,1,1)7	SARIMA Outliers	Transformed Regression Multi-Determinant
Avg Abs % Variance	7.31%	26.39%	3.75%
Avg Abs Var	2,293.2	5,745.4	1,024.3
Avg % Var	-7.15%	26.39%	-2.29%
Avg Variance	(2,258.4)	5,745.4	(622.4)
SBC	32,679.91	31,618.22	31,614.48

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**

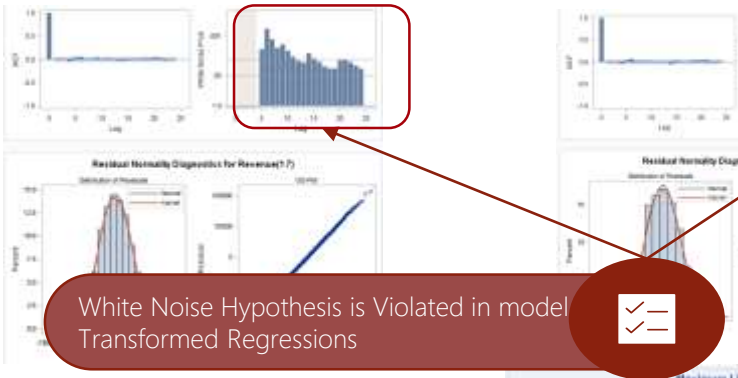
Outliers



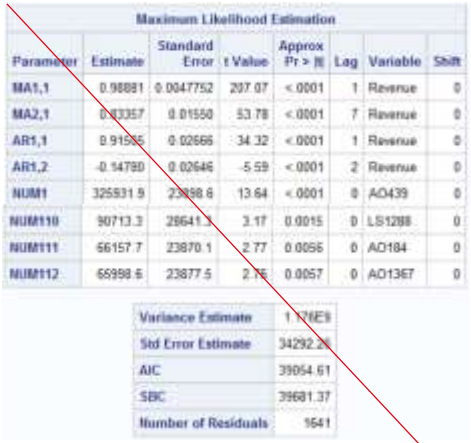
Transformed Regression



White Noise Hypothesis is Validated in DEMAND models using residuals



White Noise Hypothesis is Violated in model Transformed Regressions

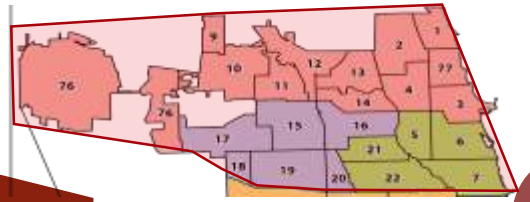


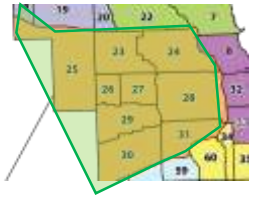
Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.98881	0.0047752	207.07	< .0001	1	Revenue	0
MA2,1	0.03357	0.01550	53.78	< .0001	7	Revenue	0
AR1,1	0.91595	0.02666	34.32	< .0001	1	Revenue	0
AR1,2	-0.14780	0.02646	-5.59	< .0001	2	Revenue	0
NUM1	325931.9	23898.6	13.64	< .0001	0	AO439	0
NUM110	90713.3	28641.3	3.17	0.0015	0	LS1288	0
NUM111	66157.7	23870.1	2.77	0.0056	0	AO184	0
NUM112	65998.6	23877.5	2.76	0.0057	0	AO1367	0

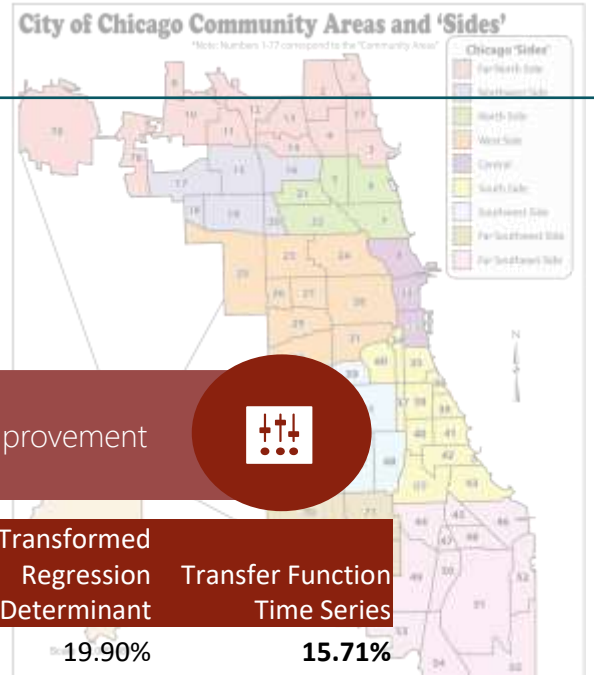
Variance Estimate	1.176E9
Std Error Estimate	34292.2
AIC	39054.61
SBC	39681.37
Number of Residuals	1641

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

Areas
Northern & Western Daily







Pickup Demand and Revenue

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

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

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