School of Computing & Communications | Lancaster University

Machine Learning Applications to Throughput Analysis

MSc Data Science Dissertation

Kieran Molloy

December 15, 2023

Table of contents

Objectives

1. Can regression models be used to predict throughput?
2. What is the most accurate model to predict throughput?
3. Can time series analysis extract further insight from the regression models?

Main Objective - Regression Modelling

- What is the impact of menu on throughput?
- What is the impact of operations on throughput?
- What are the differences between lines and factories?

Problem Introduction

Target

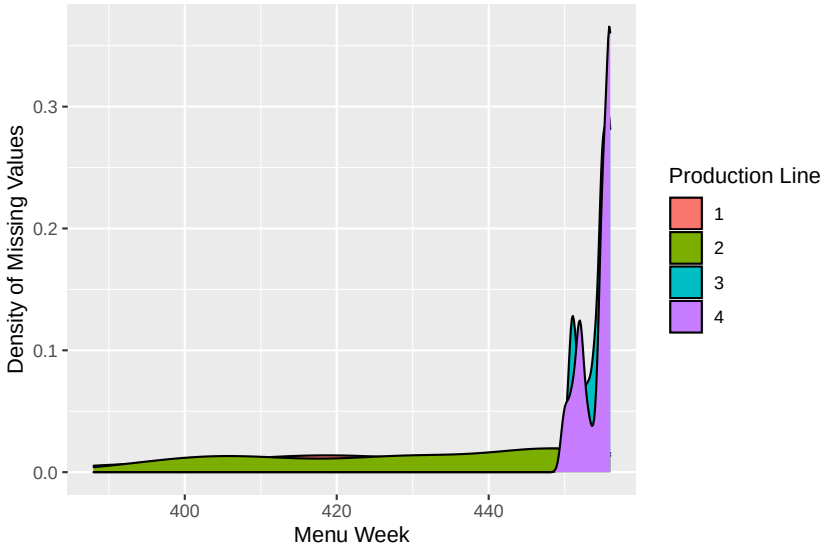
Throughput - Total Boxes Completed per minute run time.

Data is recorded automatically, and summarised by hour.
Otherwise, summarisation procedure is unknown

Useful Notation

1. Factory - 1 and 2
2. Production Lines - 1, 2, 3 and 4
3. Pick Station - Where ingredients are loaded into a box
4. Menu Week - Thursday -> Wednesday Weekly Cycle
5. Shift ... - Various Rotations of Shift Patterns

Missing Values

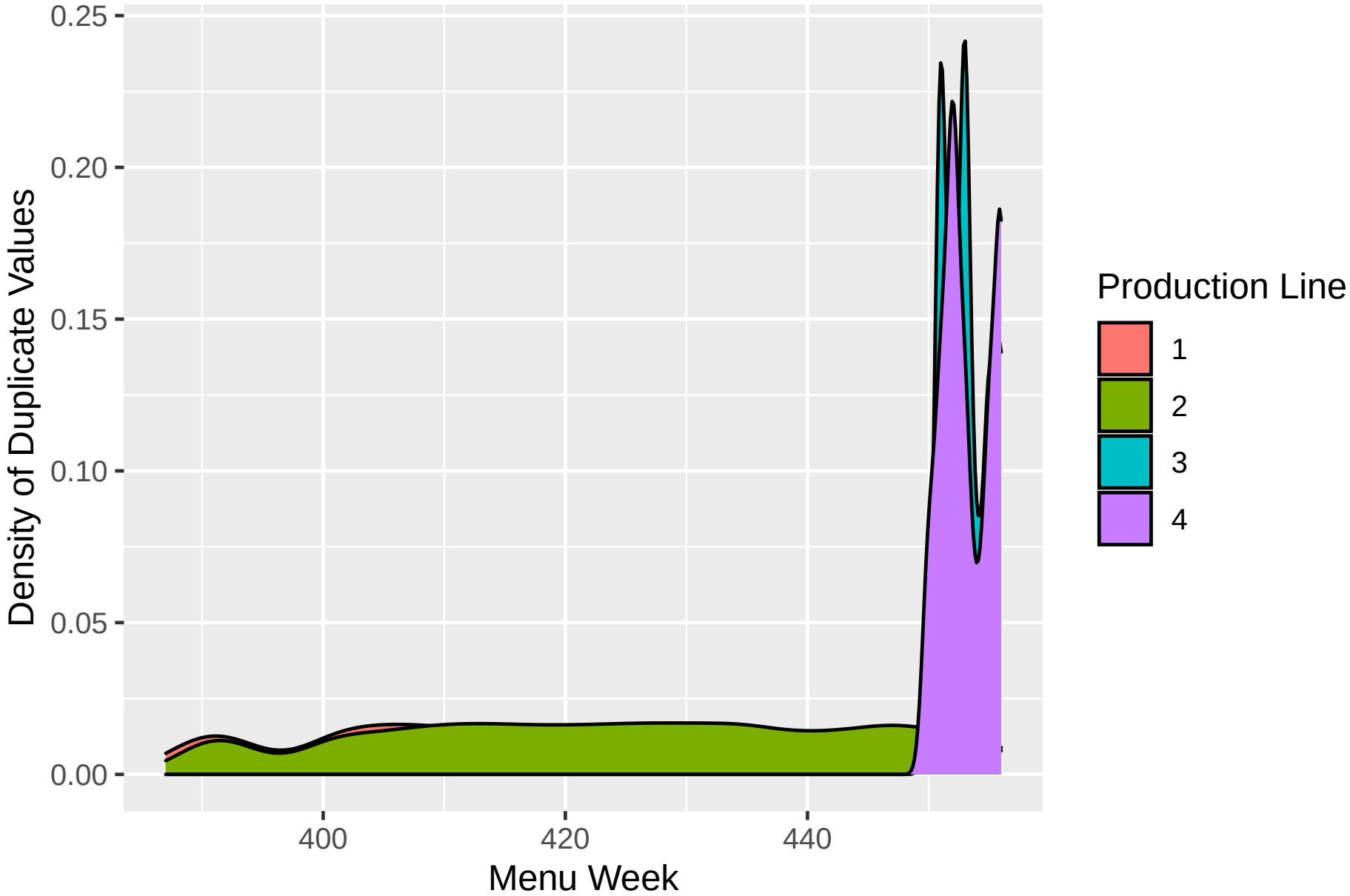


(a) menu week

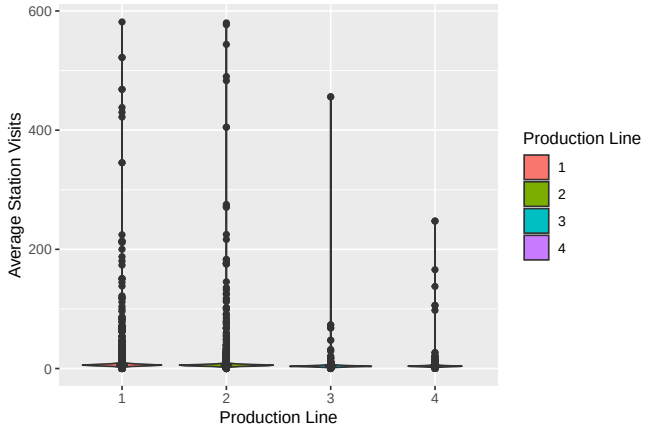
(b) counts of occurrences

Figure: Missing Value Occurrences vs Menu Week

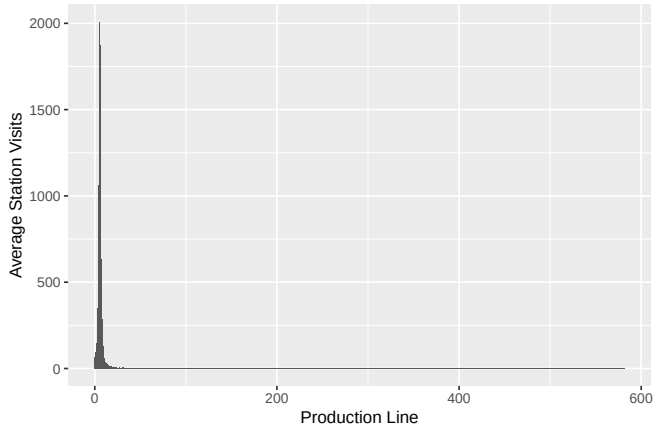
Duplicate Values



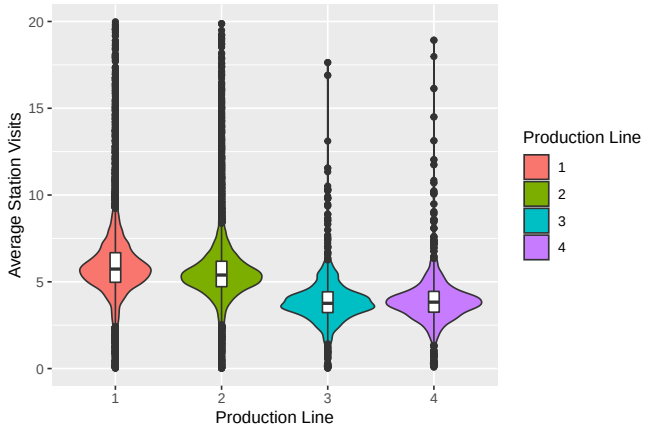
Example Distribution - Avg Station Visits



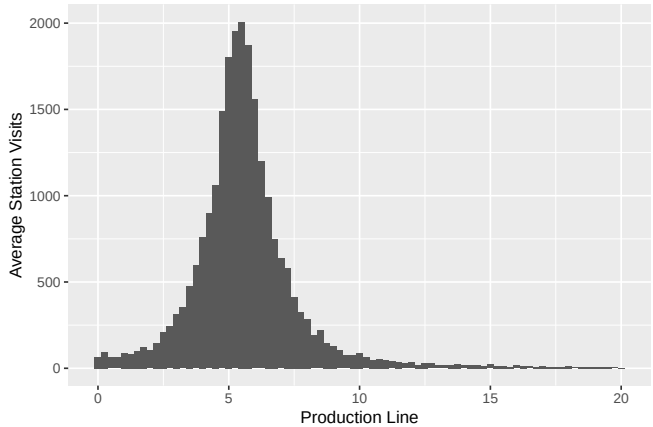
(a) standard violin plot



(b) standard histogram



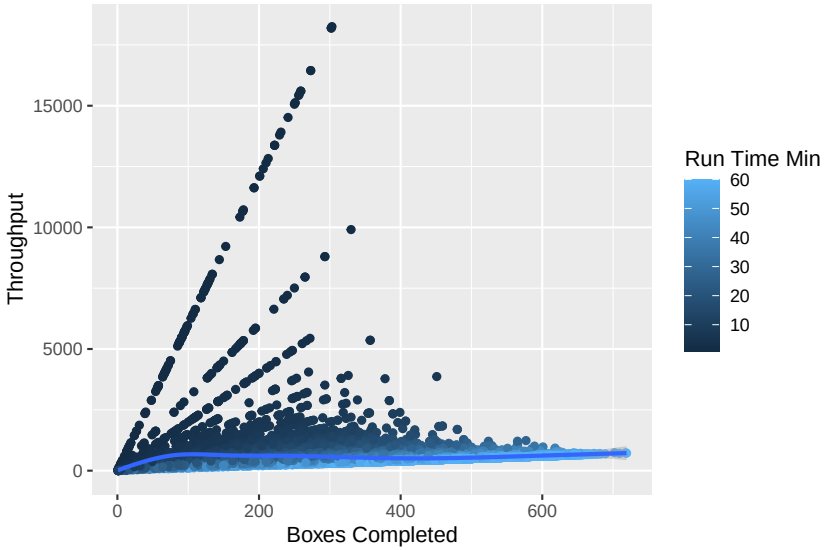
(c) violin plot $x < 20$



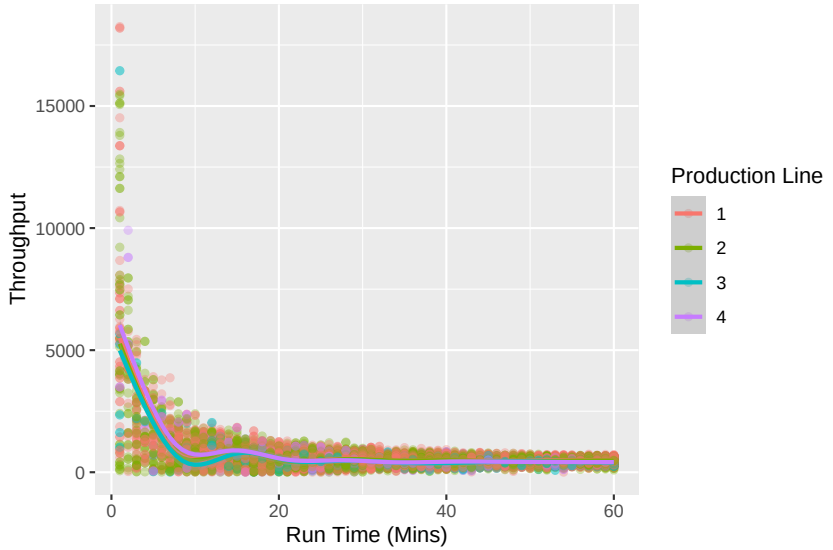
(d) histogram $x < 20$

Figure: Average Station Visits Distribution. Standard and cut $x < 20$

Target Variable



(a) throughput vs boxes completed

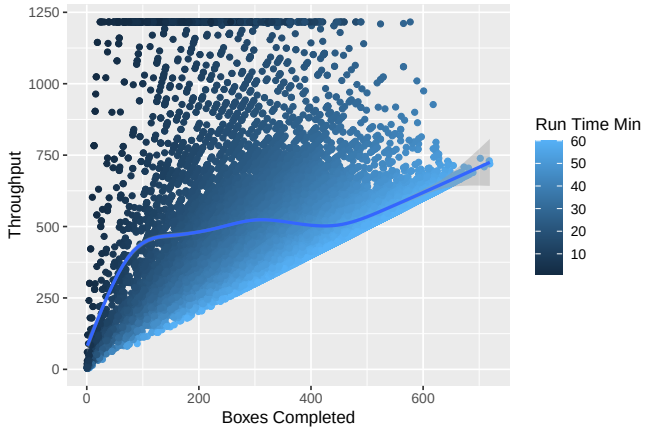


(b) throughput vs run time

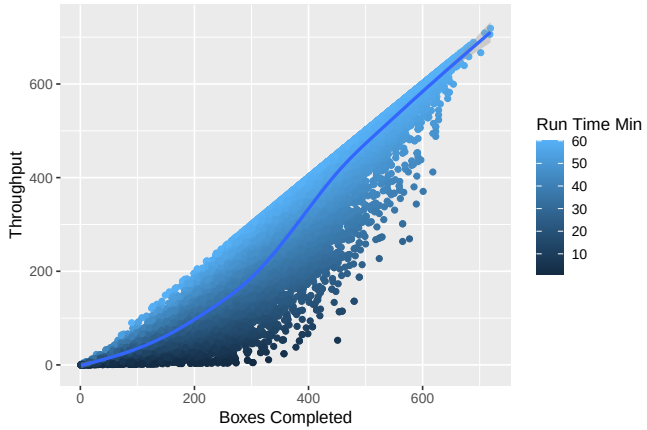
Problematic Points

- How can this relationship be accounted for?

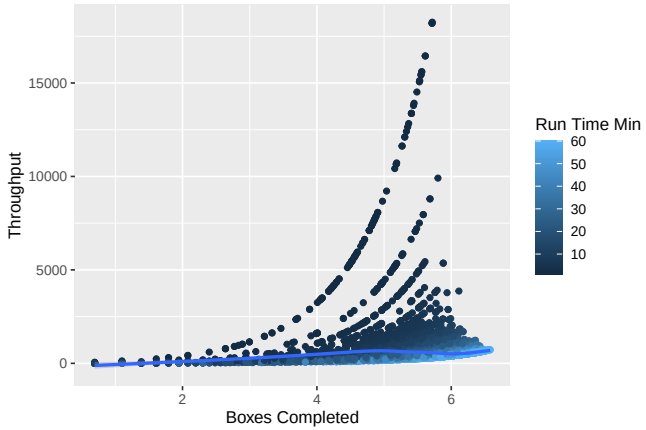
Transformations



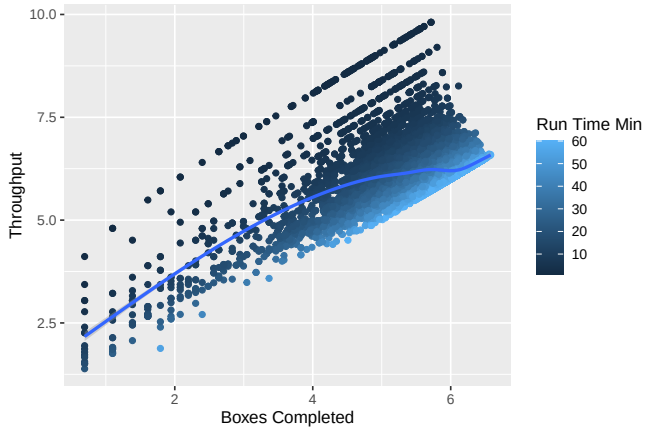
(a) 95% winsorized Standard



(b) Polynomial Transformation



(c) Semi Log Transformation



(d) Natural Log Transformation

Collinearity

Table: Variance Inflation Factor VIF $>$ 10

Variable	VIF	Variable	VIF
Ingredient SKUs	7373.49	Unit Picks per Box	37.11
Add-Ons	6263.64	Pickrate Unit Mins	34.89
Add-Ons Chilled	5151.12	Pickrate SKU Mins	33.63
SKUs Ambient	4128.94	Run Time Mins	28.07
Add-Ons Ambient	1490.15	SKU Release Rate	25.30
SKUs Chilled	941.65	Picked SKU Total	23.35
SKU Picks per Box	58.06	Total Recipe SKUs	20.30
Boxes Released	37.48	Picked Unit Total	15.29

Model Formulation

Default Formulation :

$$\text{throughput} \sim \text{remaining variables}$$

Notably

$$\text{throughput} = \text{run time} * \text{boxes completed}$$

of which both are in the dataset

Reformulation

$$\text{boxes completed} \sim \text{remaining variables} - \text{throughput}$$

Exploratory Takeaways

1. There are some missing values - not introduced totally at random
2. There are significant duplicate values - again, not introduced totally randomly
3. The distribution of variables is questionable
4. There is heteroscedasticity present
5. There is significant collinearity present
6. The target variable has questionable relationships to the dataset

Solution?

Reformulation and Robust Methods

Modelling Considerations

Testing Metrics

MSE, MedAE, R^2, % Deviance, Maximum Error, Tweedie Deviance

Testing Procedure

Splitting the dataset into test, train and validation subsets Using K-Fold Cross Validation with $k = 15$ folds, $n = 15$ repeats

Implementation

Experimentation in R
Data Pipelines in Docker Containers running on AWS written in Python

Robust Experiments

Well Defined Methods

1. M-Estimators
2. Theil-Sen Estimation
3. Huber Loss Function

Less well-defined

1. Least Trimmed Squares
2. Winsorization
3. Median Absolute Error

Conclusion

Experiments showed these were effective in handling the robust nature, but struggled in mapping the non-linear relationship effectively. Decision Tree's proved better, and using robust concepts allowed the benefits of both.

Ordinary Regression Modelling

Table: Summary Results for Standard Testing using repeated kfold $k = 20,\ n = 15$ for multiple metrics and methods

Metric	Score	OLS	SVR	Huber	Theil Sen
MSE	Mean	2620.90	3266.06	33636.47	130589.96
	Std	531.74	234.16	60976.70	249259.13
	Time	00:02.48	15:02.56	00:38.93	05:21.92
MedAE	Mean	30.23	38.38	21.05	11.16
	Std	1.48	2.12	1.08	0.52
	Time	00:01.80	10:54.36	00:28.65	04:28.68
R^2	Mean	0.86	0.83	-0.67	-5.62
	Std	0.02	0.01	2.80	12.75
	Time	00:01.76	16:01.83	00:34.55	04:44.68

Hyperparameter Optimisation

Table: Random Forest Hyperparameter Optimisation Search Optimal Parameters

	Default RF	Rand. S. RF	Grid S. RF
Estimators	100	1200	900
Max. Depth	None	171	160
Min. Leaf Samples	1	2	2
Min. Split Samples	2	2	4
Max. Features	*n* features	*n* features	*n* features
Bootstrapped	True	True	True
Search Time	0	01:16:53.36	03:30:31.42

Random Forest Regression Modelling

Table: Random Forest Hyperparameter Optimisation Search Optimal Results

		Default RF	Rand. S. RF	Grid S. RF
MSE	Mean	30.792	30.016	29.817
	Std	15.137	14.562	14.307
	Time	0:12:55	02:46:10	1:48:11
MedAE	Mean	1.179	1.188	1.073
	Std	0.069	0.0724	0.068
	Time	0:12:59	2:24:06	1:47:01
% Deviance		96.17%	96.62%	96.62%
Max Error		3.3298	2.3168	2.3118
Relative %		0%	0.47%	0.47%

Kernel Random Forest

Random Forests can be rewritten as kernel methods, which are more interpretable and easier to analyse

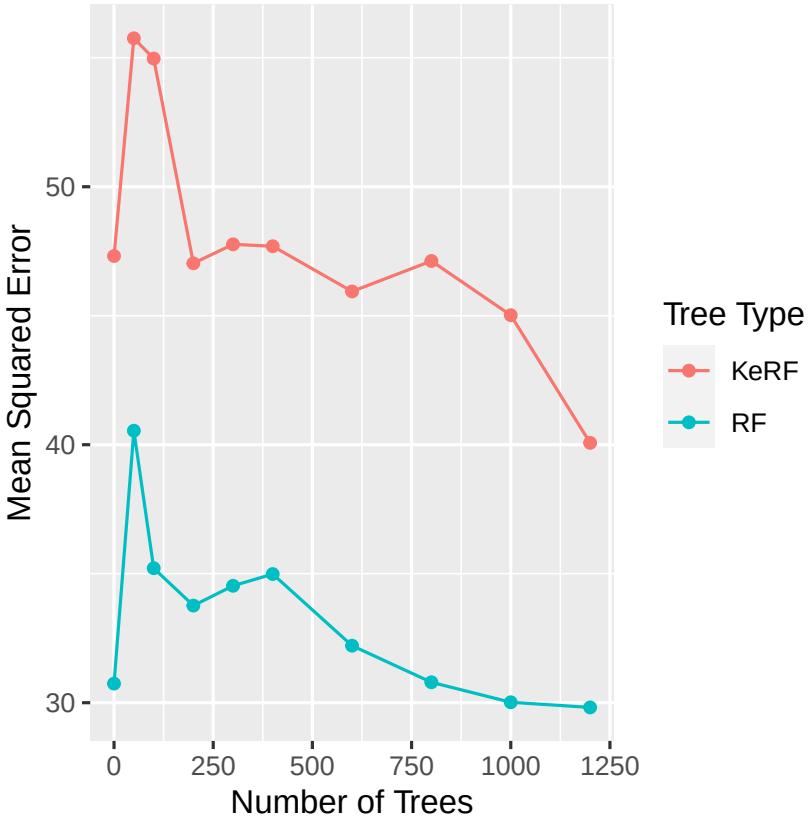
$$m_{M,n}(x, \Theta_1, \ldots, \Theta_M) = \frac{1}{\sum_{j=1}^{M} N_n(x, \Theta_j)} \sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbf{1}_{x_i \in A_n(x, \Theta_j)} \quad (1)$$

where $m_n(x, \Theta_j)$ is the predicted value at point x by the j-th tree, and Θ are independent random variables accounting for the randomness induced by node splitting and random sampling. $A_n(x, \Theta_j)$ is the cell containing x and $N_n(x, \Theta_j) = \sum_{i=1}^{n} \mathbf{1}_{x_i \in A_n(x, \Theta_j)}$. Which finally defines the KeRF as

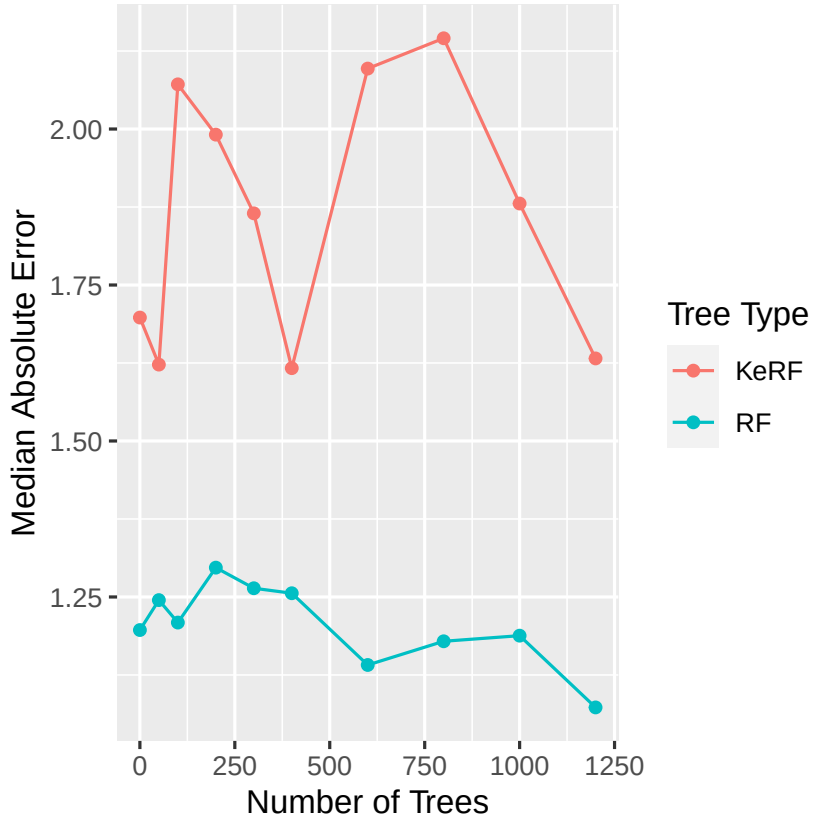
$$m_{M,n}(x, \Theta_1, \ldots, \Theta_M) = \frac{\sum_{i=1}^{n} Y_i K_{M,n}(x, x_i)}{\sum_{\ell-1}^{n} K_{M,n}(x, x_\ell)} \quad (2)$$

with the connection function $K_{M,n}(x, z) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{z \in A_n x, \Theta_j}$

KeRF to Random Forest



(a) MSE



(b) MedAE

Figure: Kernel Random Forest vs Random Forest for Mean Squared Error and Median Absolute Error

Example Tree Nodes

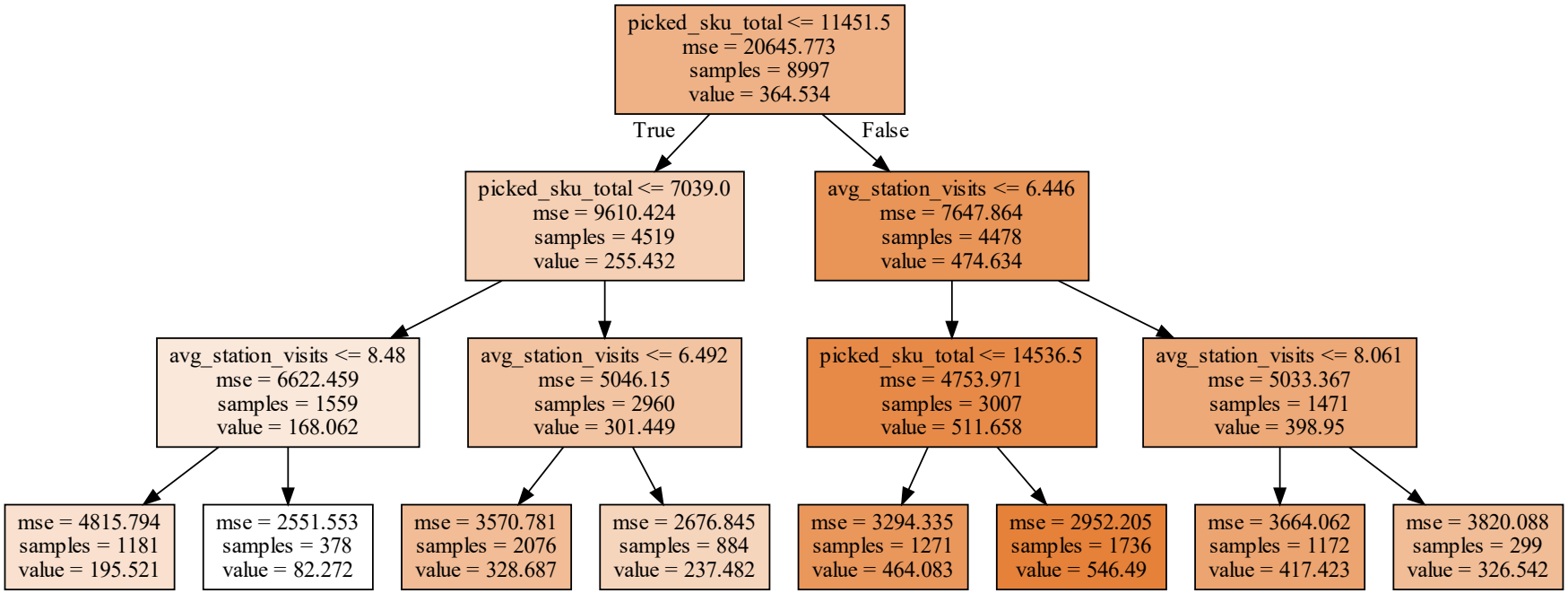


Figure: Random Forest Nodes pruned to $n = 3$ for line 1

Changepoint Definition

Multiple Changepoint Detection using PELT (R.Killick) is defined by

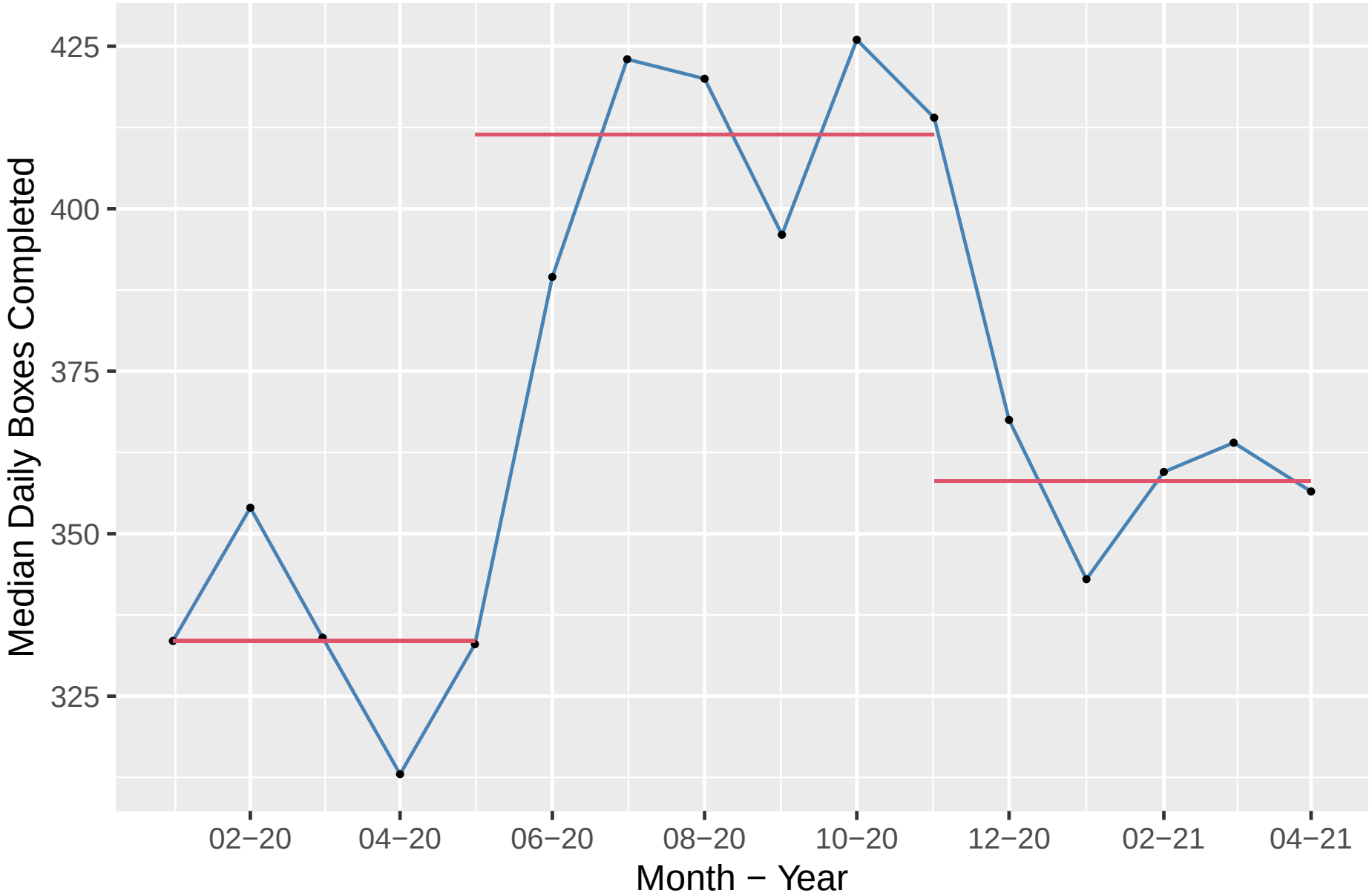
$$F(n) = \min_{\tau_m}\{F(\tau_m) + C(y_{\tau_{m+1}}, \ldots, y_n)\} \tag{3}$$

where C is some cost function. The procedure calculated F_1, then recursively. Each optimal segmentation for F_i is stored up to τ_{m+1}. The penalty function is a CROPS, which is effectively a range of values for which PELT will iteratively test all values. Allowing the production of an elbow plot.

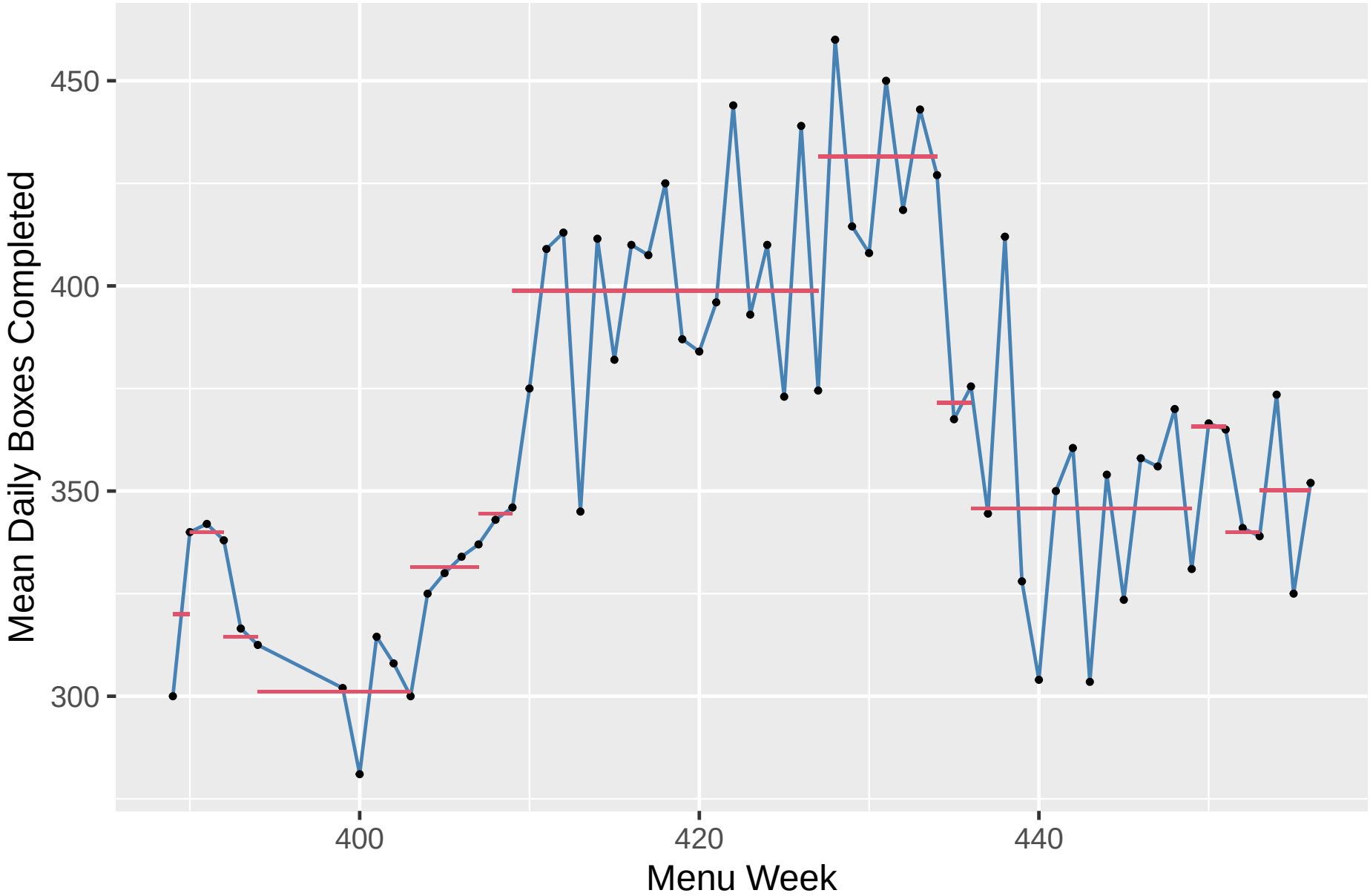
$$\text{C} = [5, 500] \tag{4}$$

The implementation is based upon Killick's R Package 'changepoint'

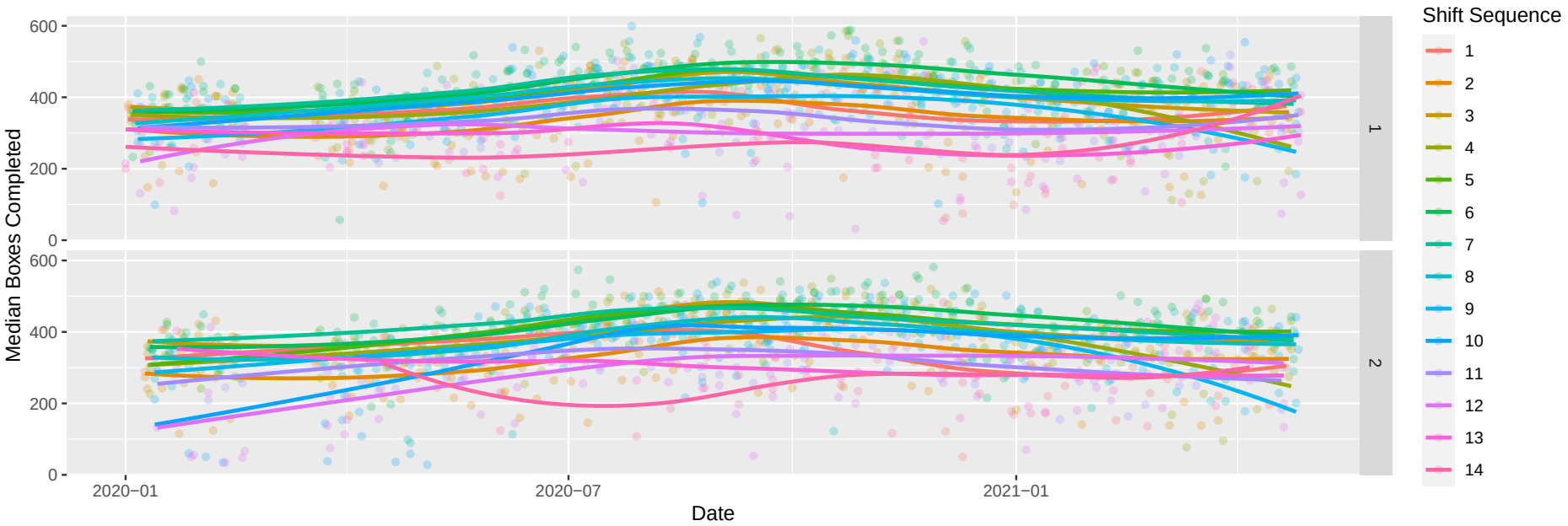
Changepoint - Monthly

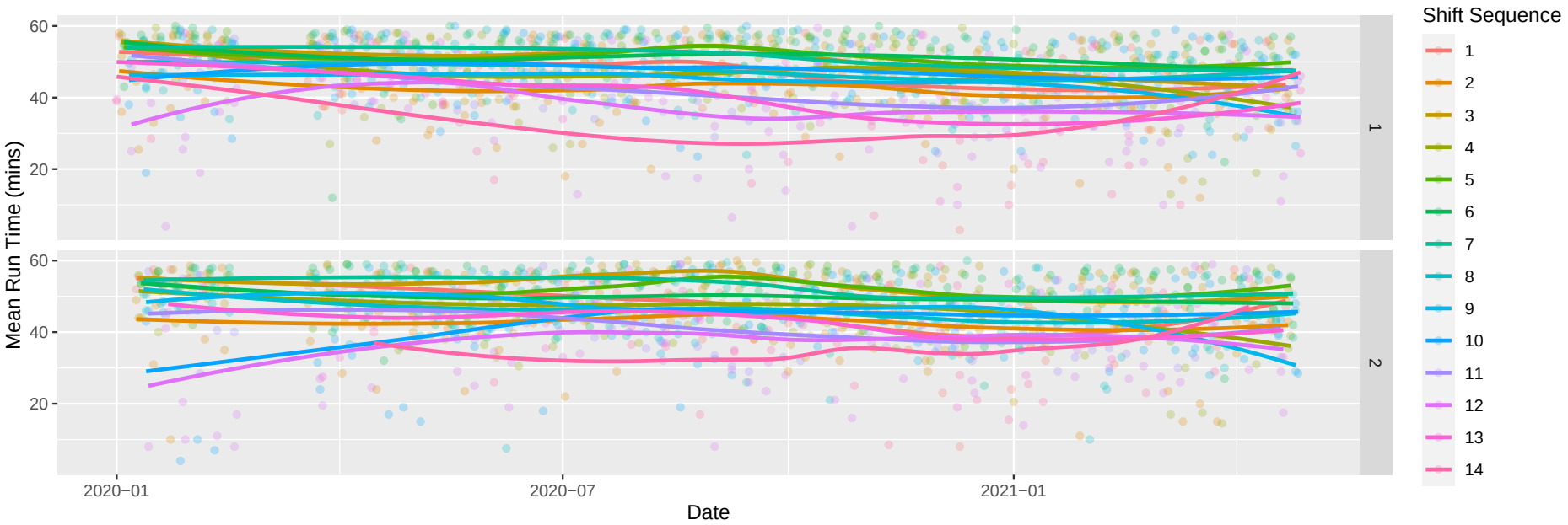


Changepoint - Menu Week-ly



Changepoint - Shift Sequence





Modelling Summary

Model Outcomes

1. Average Vistis $>$ 4.15 $\rightarrow$ Higher Boxes Completed
2. Higher SKUs Picked $\rightarrow$ Higher Boxes Completed
3. Clear Trends within shift type, shift colour and sequence

1. Robust Statistics were vital to make use of dataset
2. Models are Similar for comparable lines. But not factories
3. Solution is (from limited testing) translatable

Aims

1. Can Regression Models be used to predict throughput?
2. What is the most accurate model to predict throughput?
3. Can time series analysis extract further insight from the regression models?

Discussion

Limitations

1. Extreme Levels of Multicollinearity
2. The nature of the data is unknown
3. Validity and Reliability of dataset is questionable

Future Work

1. Use of raw data
2. Markov Chain State modelling
3. Totally-Out-of-Set Testing

Questions?

Thank you for listening

20 Mins for Q&A

References I


Peter J. Huber
Robust Statistics.
Wiley, 2009.


John W. Tukey
A survey of sampling from contaminated distributions
Stanford University Press, 1960.


L. Breiman
Random Forests.
Machine Learning, 2001


E. Scornet
Random Forests and Kernel Methods.
Machine Learning, 2016

References II


R. Killick, P. Fearnhead and I. A. Eckley
Optimal Detection of Changepoints With a Linear Computational Cost.
Journal of the American Statistical Association, 2012

School of Computing & Communications | Lancaster University

Kieran Molloy

Machine Learning Applications to Throughput Analysis

MSc Data Science Dissertation