Carnegie Mellon University

Project Progress Report

April 19th, 2021

Frank Kovacs, Ning Gao, Pragya Jain, Wonil Lee

Agenda

- ❖ Introduction
  - ➢ Team & Client Profile
  - ➢ Project Scope
- ❖ Data & EDA
- ❖ Methods
  - ➢ Differences Filters
  - ➢ Trimmed Weighted Average Filters
- ❖ Results
- ❖ Next Steps
- ❖ Q&A

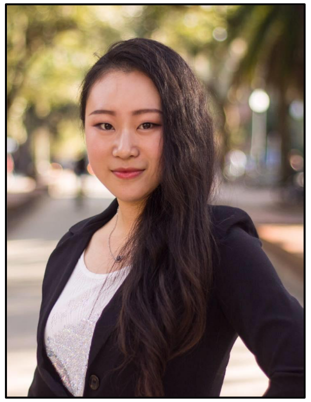
Introduction

Team

Frank Kovacs

- CMU Statistics & Machine Learning ‘19
- Software & Data Research
- Research with Delphi COVIDcast and ISLE

Ning Gao



- Georgia Tech Industrial & Systems Engineering ‘20
- Research with NSF LeapHi Program
- Past work experience in the telecom industry

Pragya Jain

- Past work experience in the insurance industry
- Associate Actuary
- B.E. from NSIT, New Delhi

Wonil Lee

- Past work experience in Consulting (2+ years)
- CMU Tepper & Statistics ‘18
- R, SQL, and Python

NPD Group Overview

- NPD Group is a **Market research company**
- “Raw data assets into insights”
- Specialize in general merchandise and food service
- Market leader
    - **8B+** B2B transactions / yr

Objective & Scope

- "...explore using unsupervised learning methods to help identify common data collection errors to help guide further analyst review."
- **Goals**
  - <u>**Detect anomalies in time series datasets**</u>
  - Identify common data collection errors
  - Facilitate further data analyst review
  - Automate data error flagging processes

 Source: Capstone Handout

Data

NPD Project Dataset Overview - Key Variables

- **Merchant ID and Name**
- **Acquire Type ID**
- **Receipt_count**
- **Sum_total_paid**
- **Item_total**
- **Sum_items_distinct**
- **Sum_item_spend**
- **Panelists**

NPD Project Dataset Overview - Main Datasets

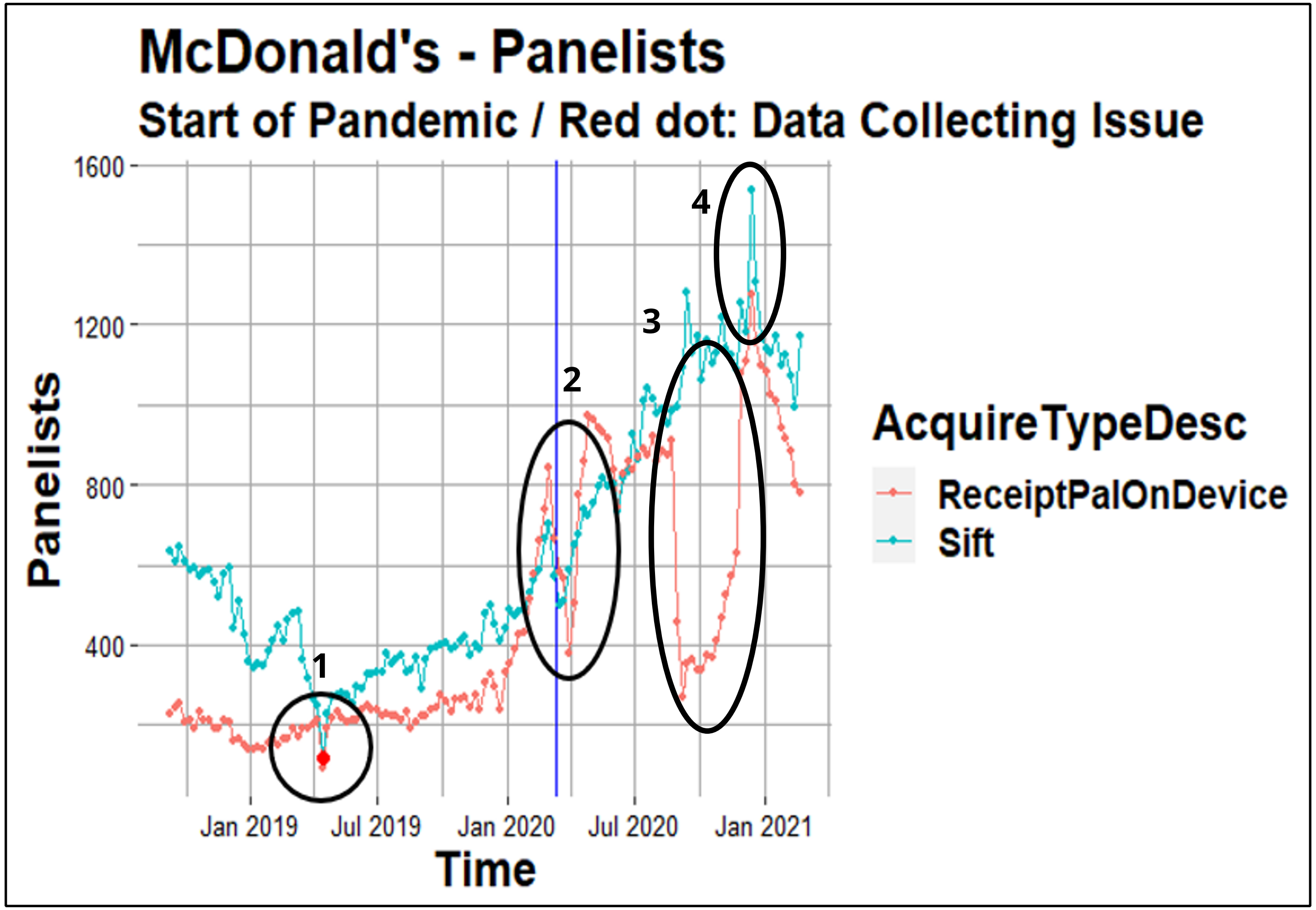
- **Source Data**
  - 516 rows, 8 columns
  - Weekly values of the receipt_count, sum_total_paid, sum_items_distinct, sum_item_spend, panelists by 4 different data source types (iPhone, Android, Sift, and Receipt pal on device)
- **Retailer Data**
  - 983,953 rows, 11 columns
  - Weekly values of the receipt_count, sum_total_paid, sum_items_distinct, sum_item_spend, panelists by individual merchants and by different data sources
- **Issue Data**
  - 31 rows, 5 columns
  - Dataset of when (the Acquired date) and where (merchant name & source type) the data collecting error occurred

EDA

EDA

- **Existing Flags**
  - Issues logged by client in the past 2 years were shared
- **Data Preparation**
  - Data sanity checks
  - Merged ‘Retailer Data’ with ‘Issue Data’
- **EDA Plots**
  - Generated time series visualizations for individual merchants and marked issues logged by client with a ‘Red Dot’
  - Start of Pandemic marked with a vertical line - March 11th, 2020
  - Highlighted potential unmarked anomalies

EDA Example



- Anomalies right after start of COVID-19 not marked
- Dips with smaller amplitude detected but dips with bigger amplitude left unmarked
- Detection of rapid dips over 2-3 weeks but not over 1 months
- Preference for marking dips over peaks

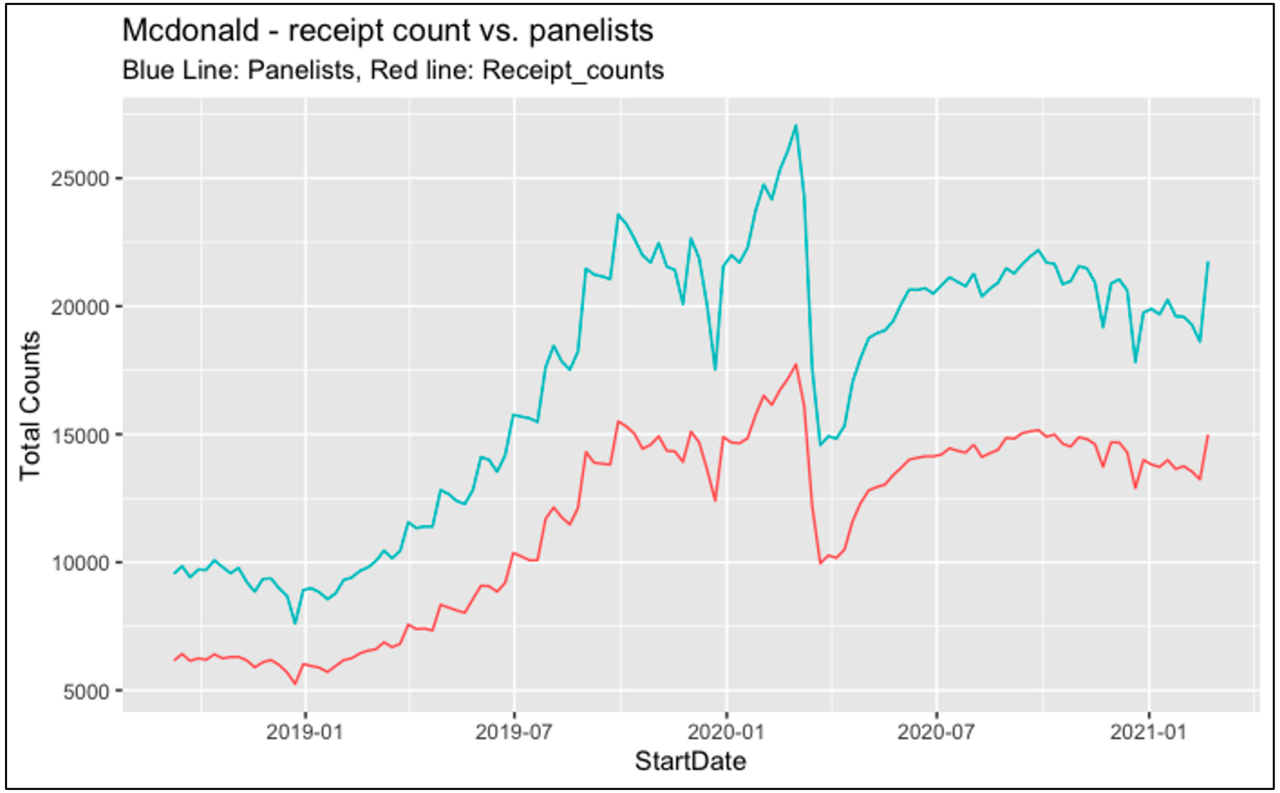
Variable Selection

- Out of 6 response columns in client's data, following pairs of columns were observed to be correlated. The table below displays the correlation values:

Pair of Columns	Correlation
Receipt_count & Panelists	**0.9937866**
Sum_total_paid & Sum_item_spend	**0.9990581**
Item_total & Sum_items_distinct	**0.9991476**

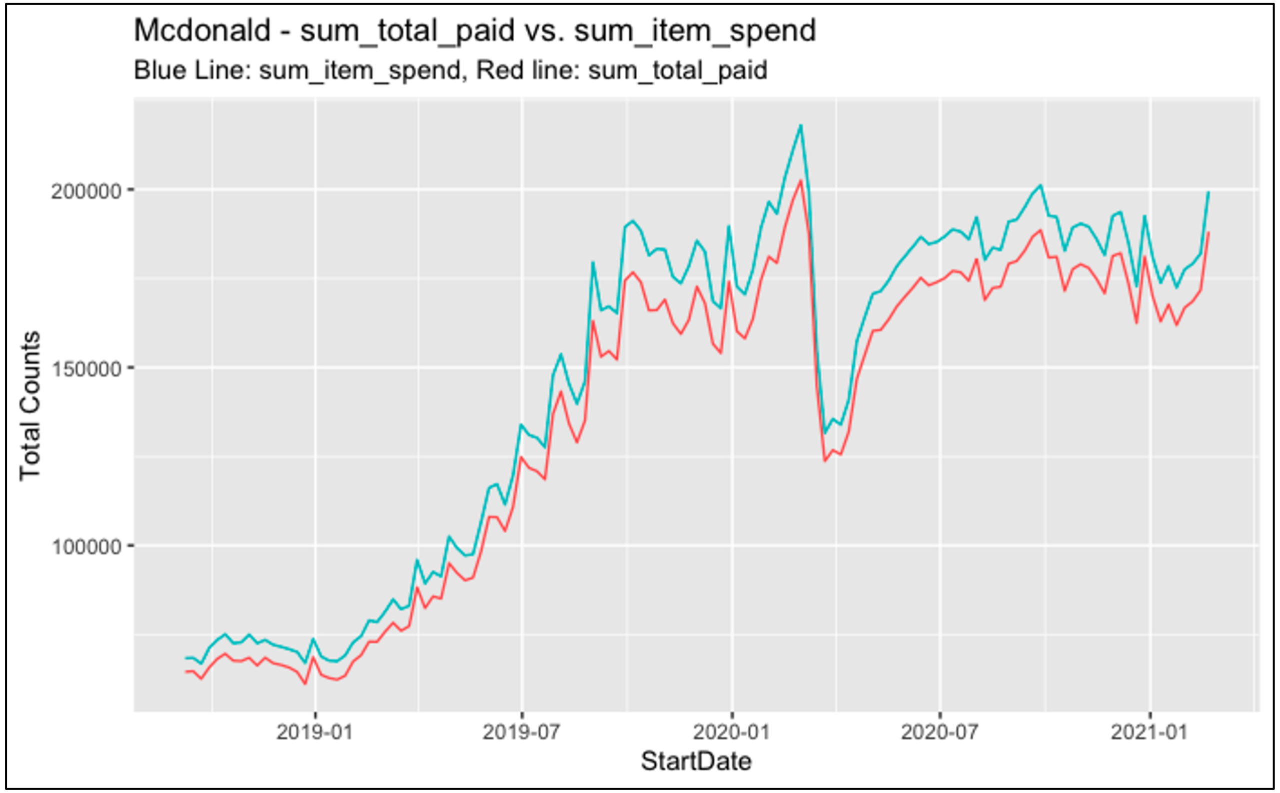
Example

Variable Selection - Receipt Count vs Panelists



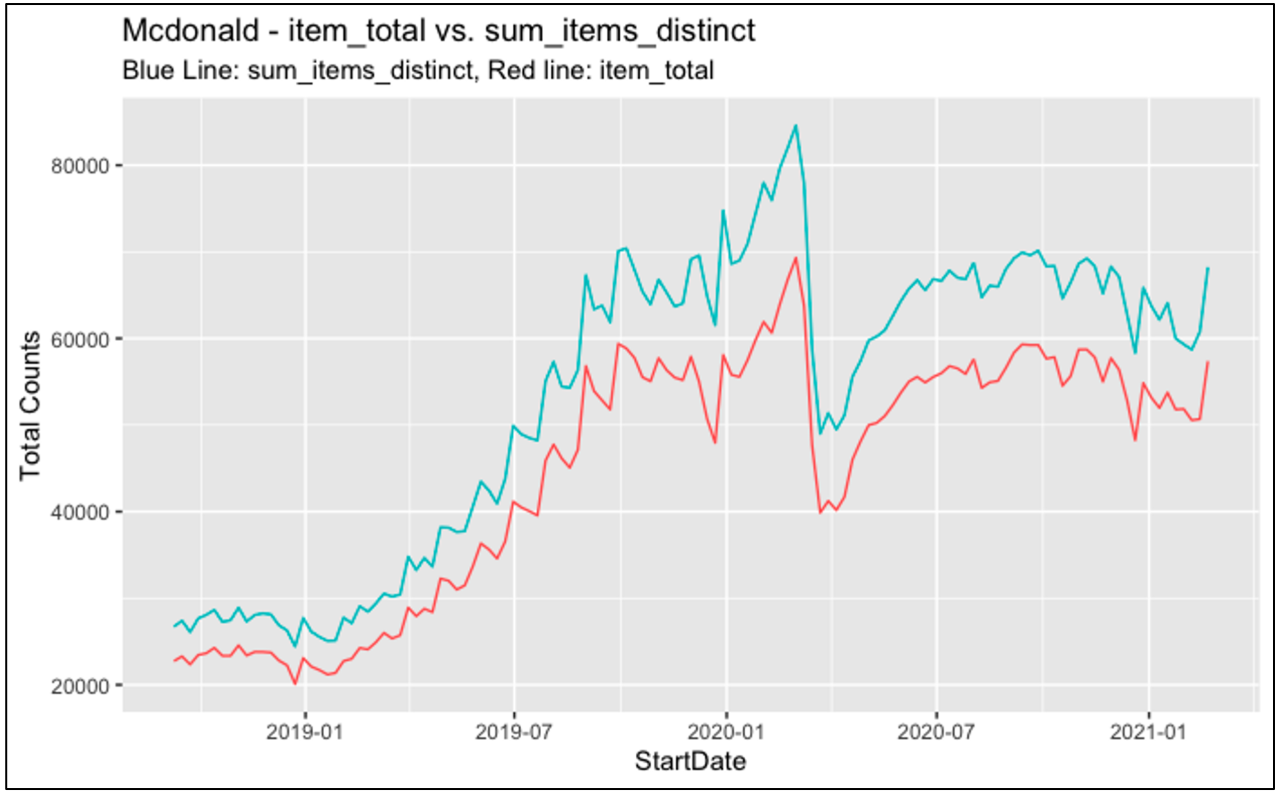
Example

Variable Selection - Sum Total Paid vs Sum Item Spend



Example

Variable Selection - Items Total vs Sum Items Distinct



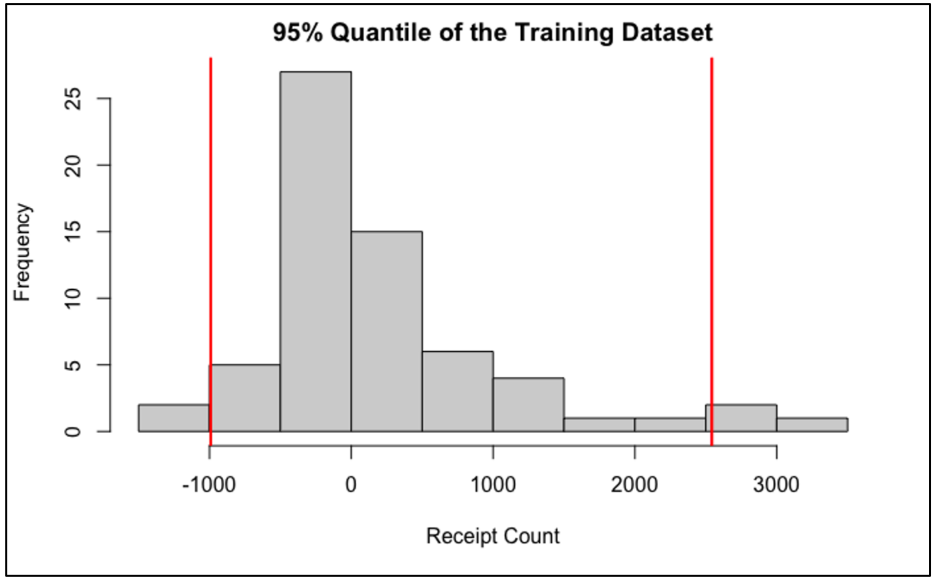
Methods

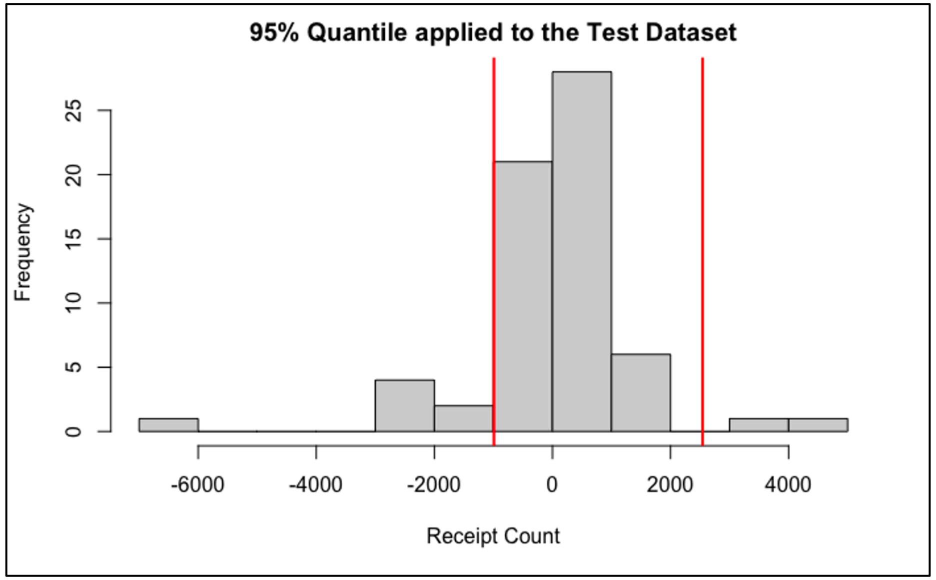
First Difference Filter - Logic

- Calculating multiple quantiles for the first difference analysis
  - First differences are calculated for a selected window size
  - Selected Quantile is calculated for all first differences
  - Outliers outside of calculated quantile are marked as anomaly

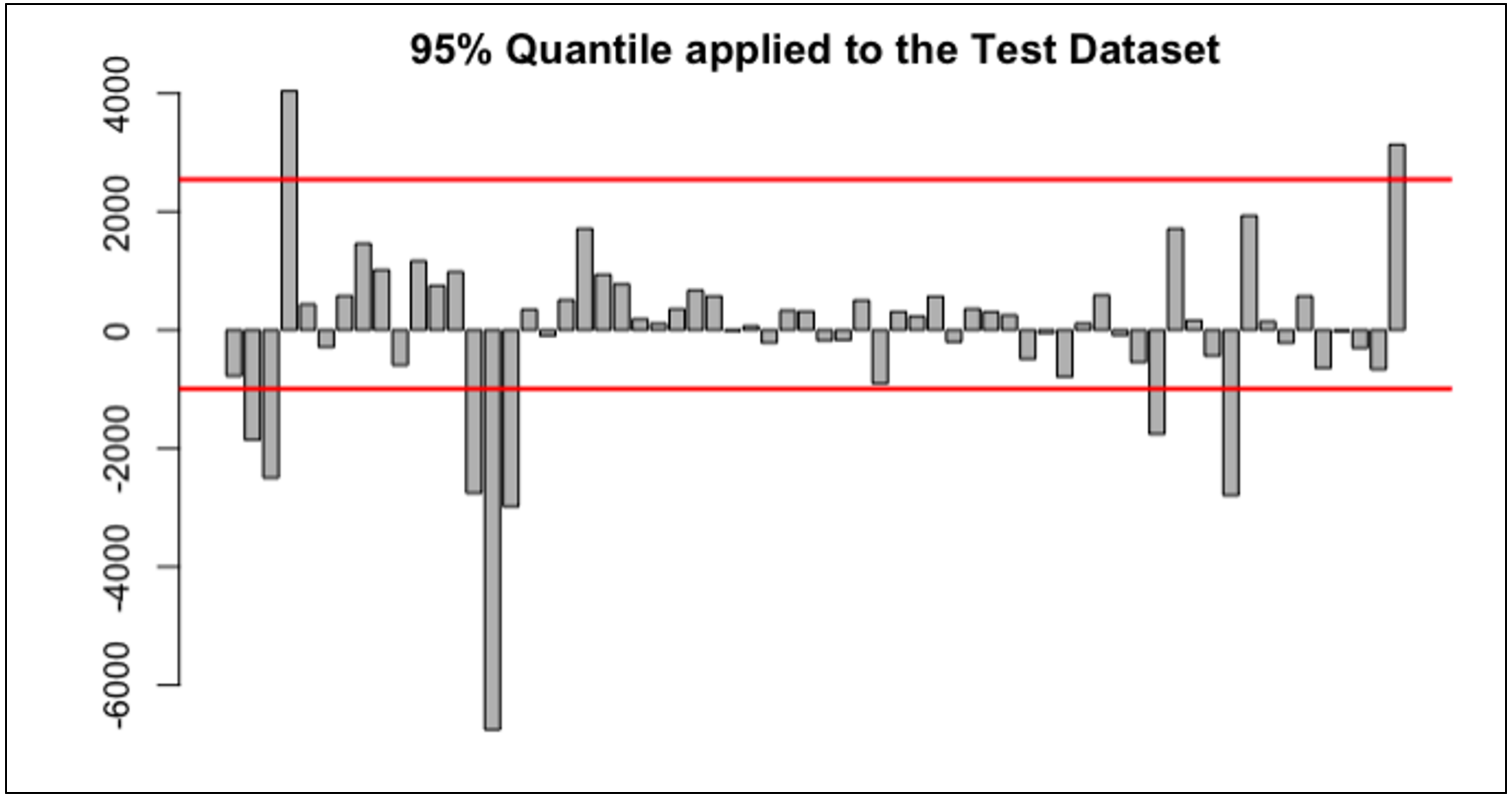
- Shifting Window
  - For each point in the future (say at ‘t+1’), **4** filters will perform anomaly detection
  - Each filter has the same window size ‘n’ but covers different points in the past
    - Filter 1 covers points from ‘t-n’ till ‘t’
    - Filter 2 covers points from ‘t-n-1’ till ‘t-1’ and so on

First Difference Filter - Sanity Check 1/2





First Difference Filter - Sanity Check 2/2

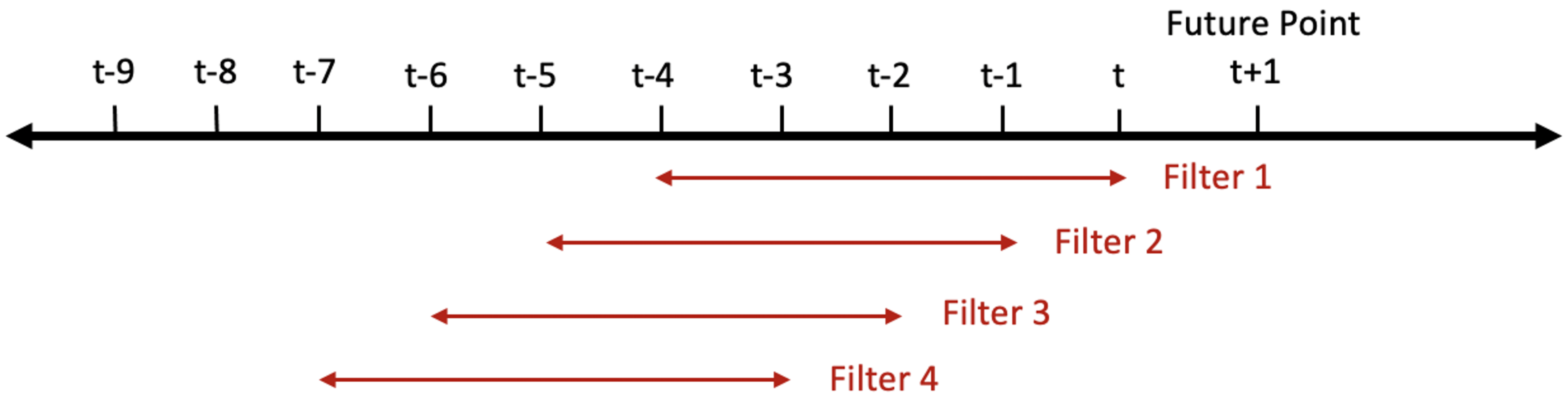


Trimmed Moving Weighted Average Filter - Logic

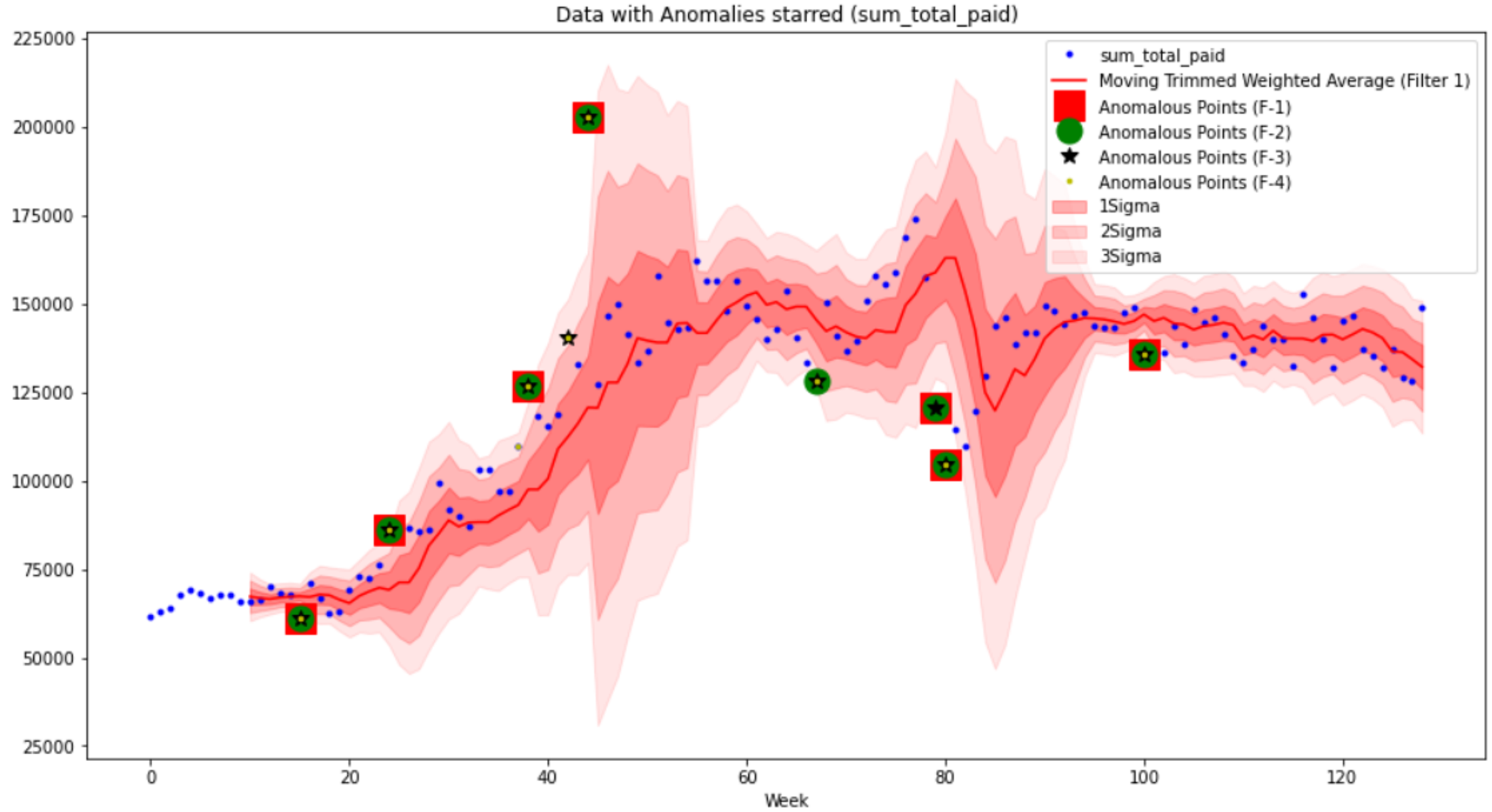
- Calculating robust estimate for moving weighted average
  - Trimming
    - Median is calculated for a selected window size
    - Outliers (max 20% points) are removed based on ranked distance from the median
  - Weighted Average calculation for new set of points
    - Weights increase with increasing time (more weight to recent points)
    - Weights are generated from a half normal distribution

- Shifting Window
  - For each point in the future (say at ‘t+1’), **4** filters perform anomaly detection
  - Each filter has the same window size ‘n’ but covers different points in the past
    - Filter 1 covers points from ‘t-n’ till ‘t’
    - Filter 2 covers points from ‘t-n-1’ till ‘t-1’ and so on

MA Analysis - Filters Illustration



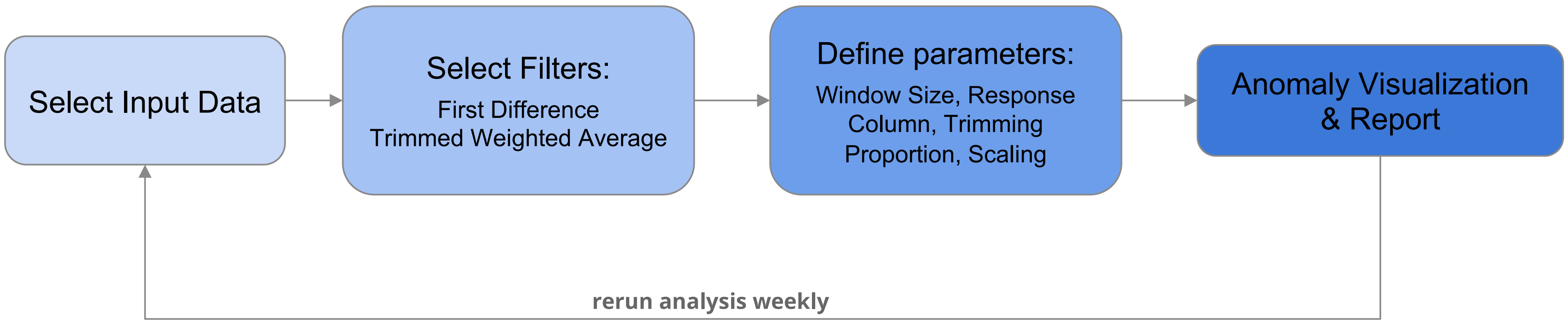
MA Analysis - Drafted Visualization



Result

Result Overview

A python package will be shared with the client. Its broad components are described below:



Next Steps

Next Steps

- Finalizing results for both filtering methods
  - Use robust estimate for standard deviation
  - Clean up plots
- Integration of all developed functions into the user-interface
- Client Feedback
- IDMRAD Paper Documentation and Review

Q&A

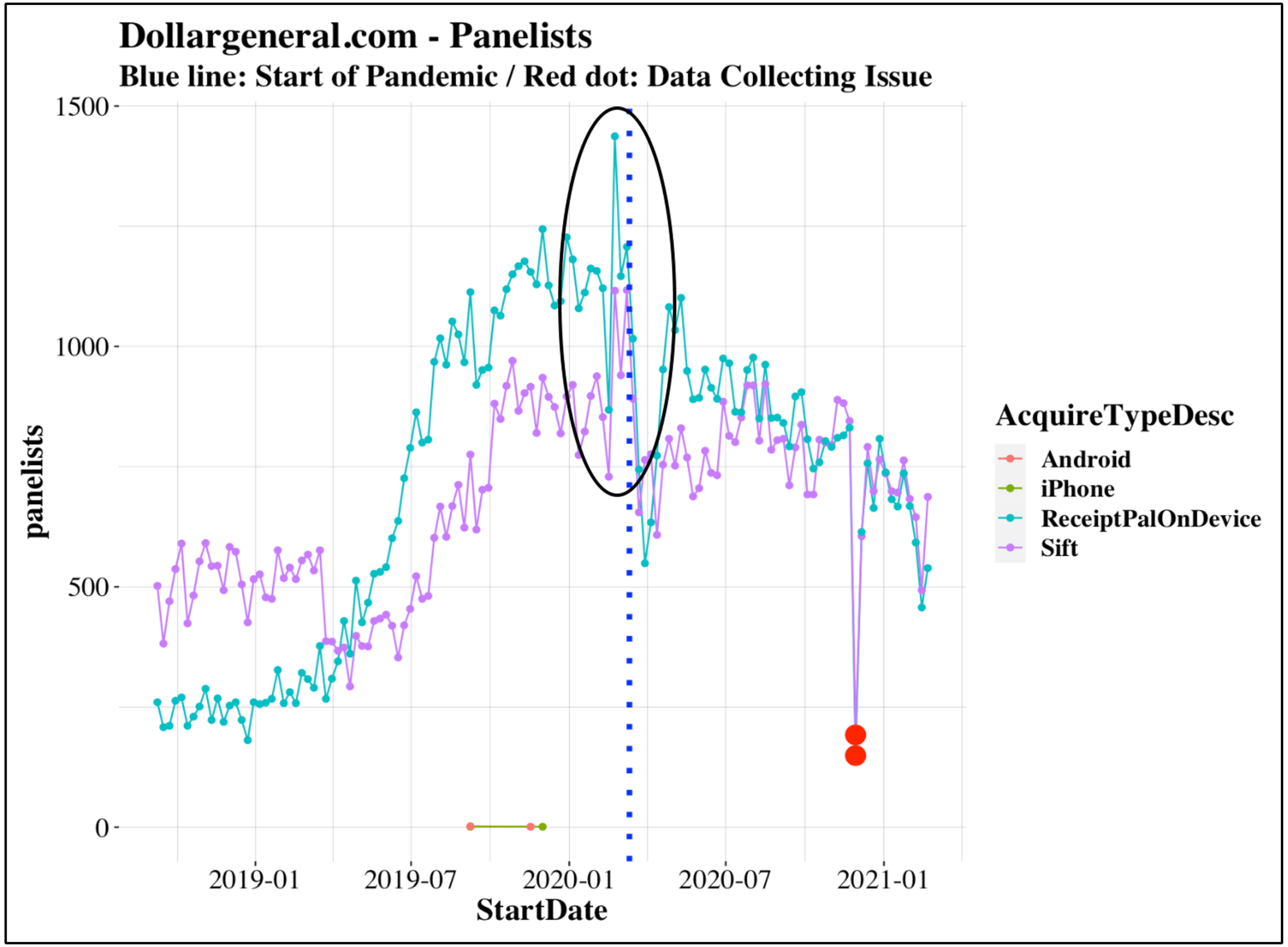
THANK YOU!

Appendix

Issues logged | Visualization Explanation

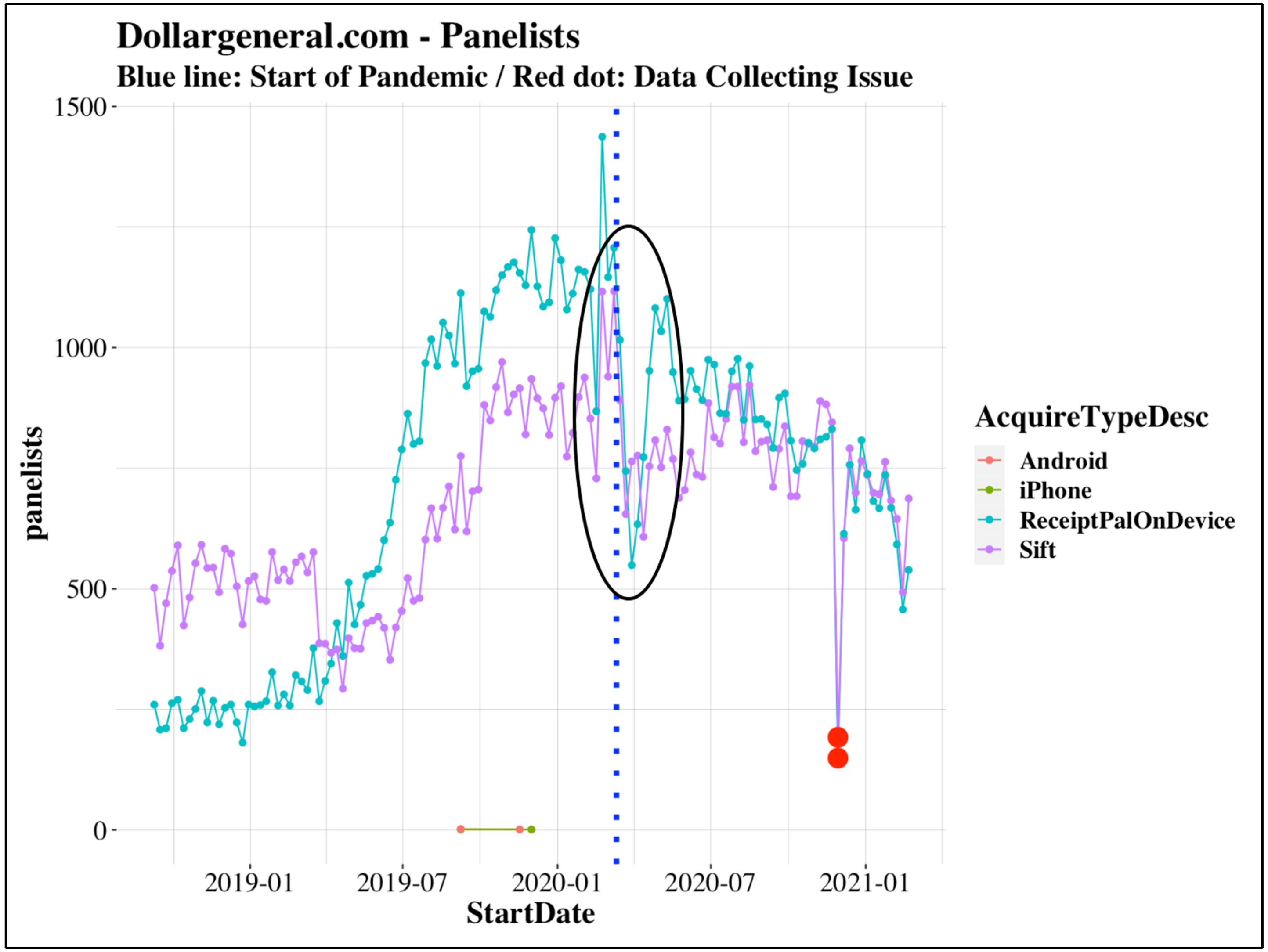
Preference for marking dips as issues

- A peak of significant amplitude has not been marked as an issue
- A dip of similar amplitude has been marked as an issue



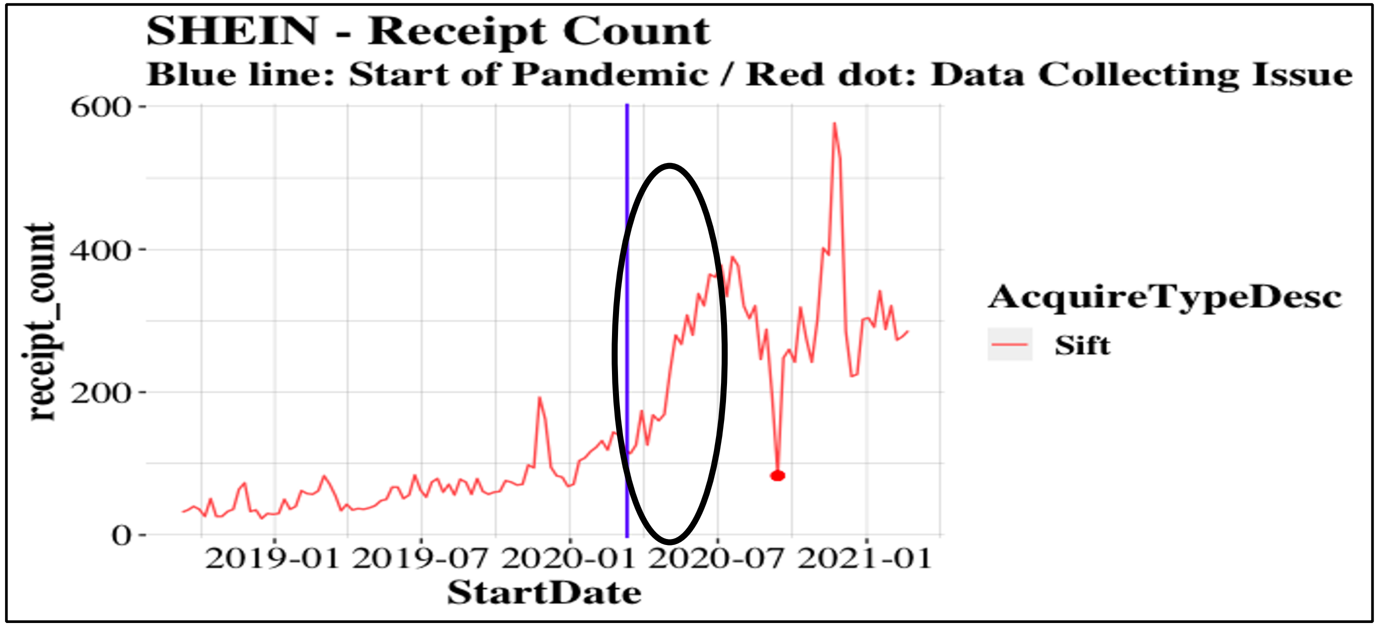
Anomalies right after Pandemic Start not marked

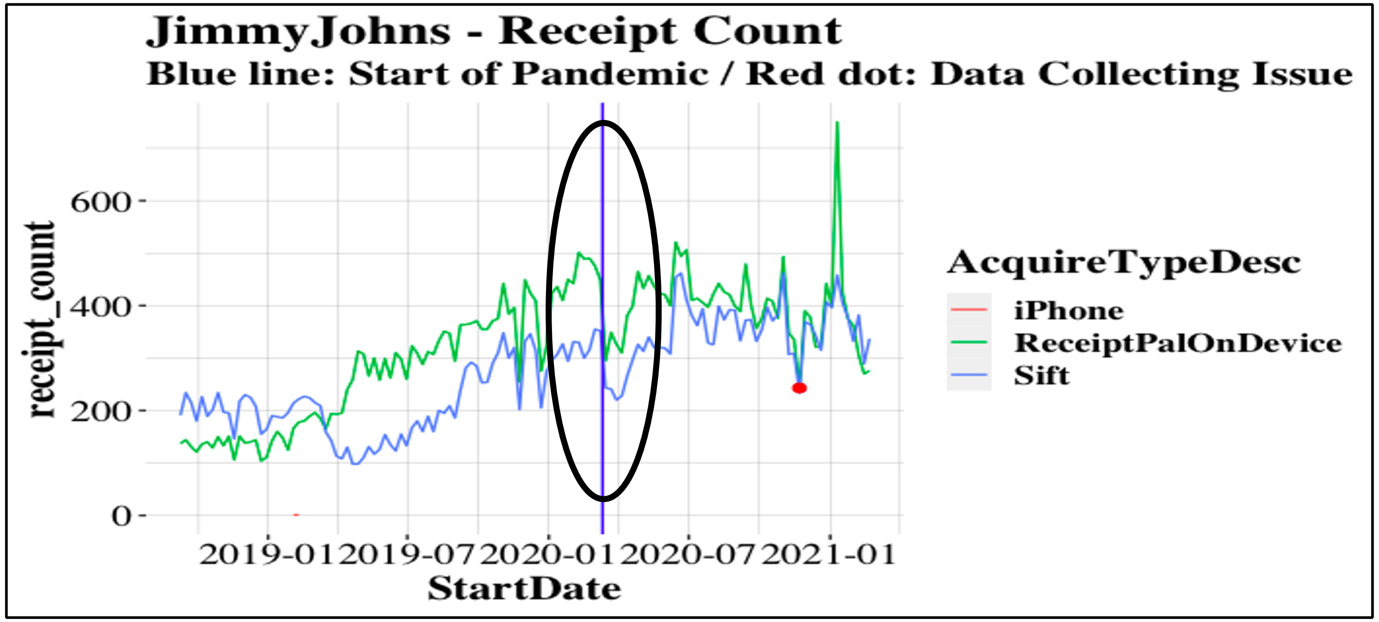
- Detected drop near Christmas 2020
- Did not detect drop near start of Covid-19



Anomalies right after Pandemic Start not marked

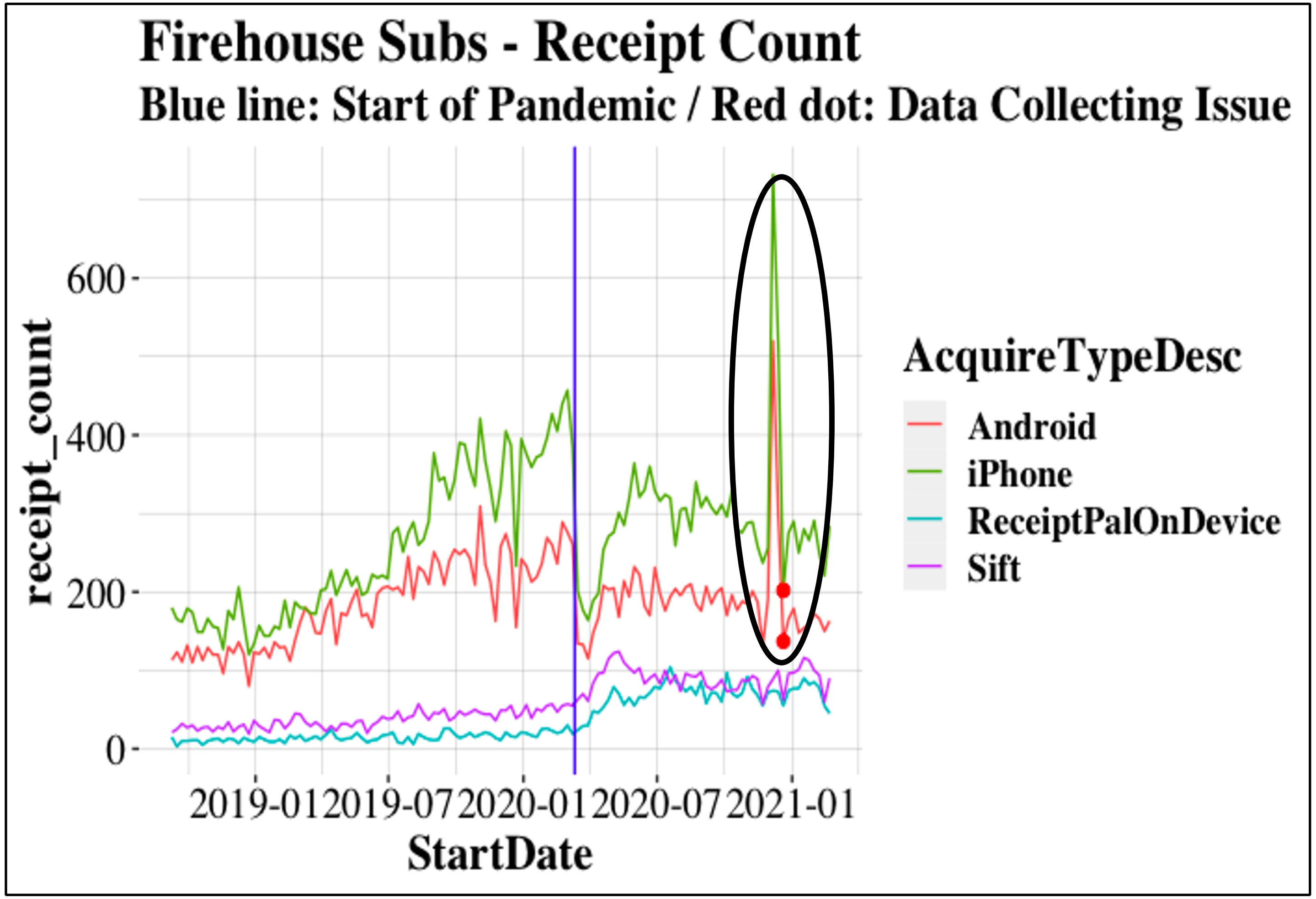
- The anomalies after the Pandemic outbreak is not captured
  - Sudden Decrease
  - Sudden Increase
- We would like to know whether the current algorithm considers impact of COVID19





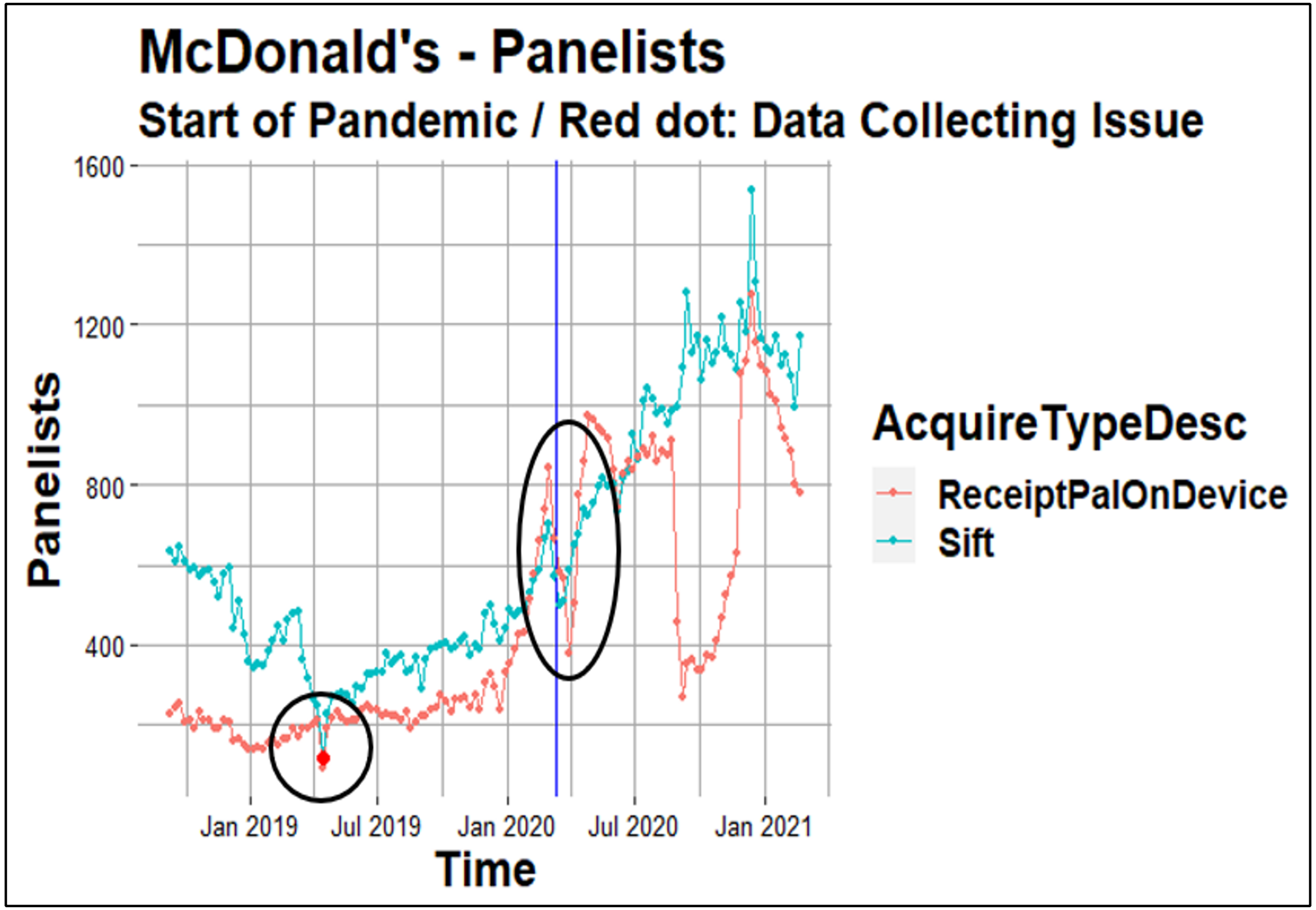
Delayed detection of Sudden Shifts

- Anomalies are detected in delayed manners
  - The error was detected 3 weeks after the first abnormal value



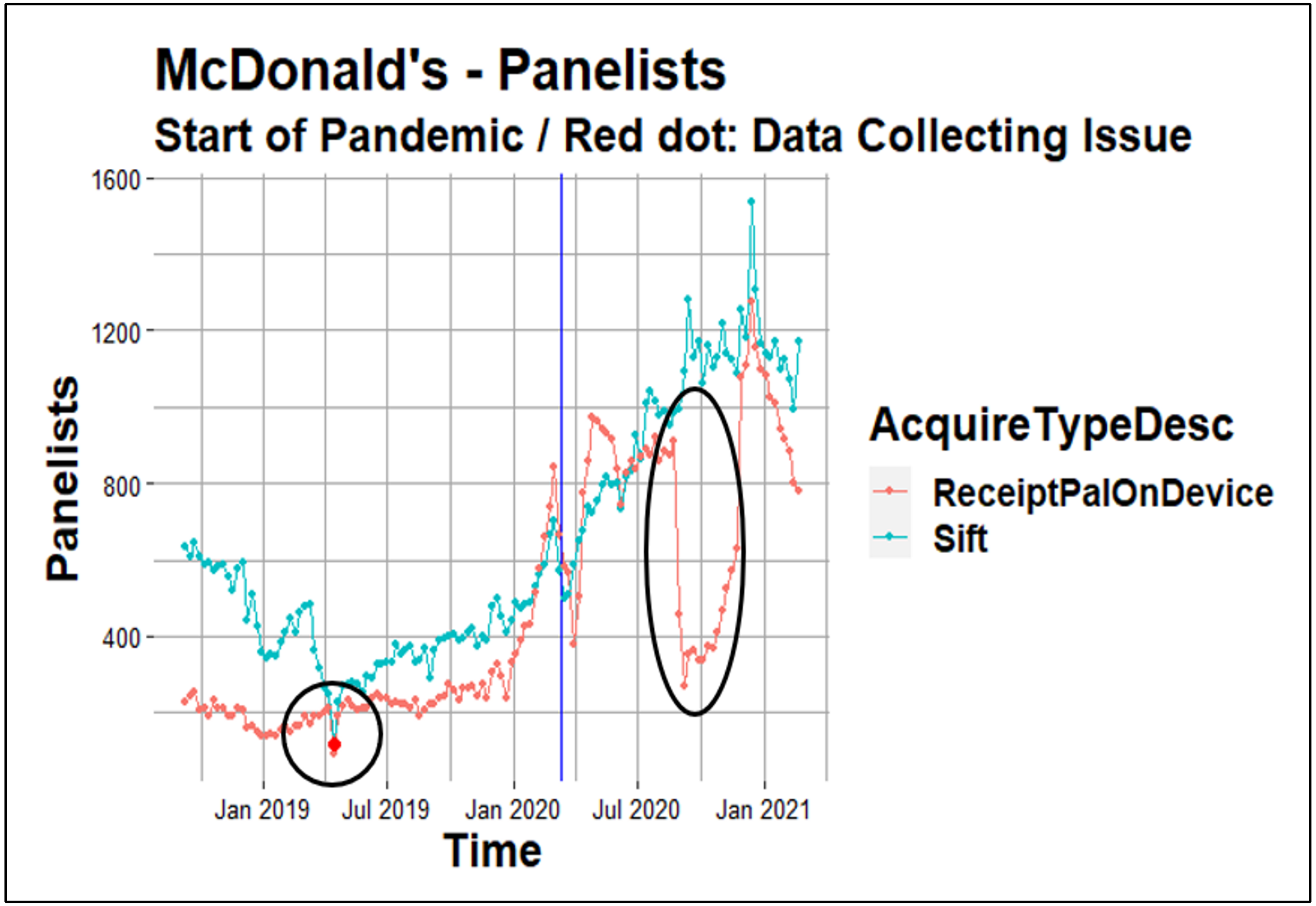
Detection of rapid dips over 2-3 weeks but not over 1 months

- Small, sharp drop of panelists using Sift around April 2019
- Bigger, consistent 1 month drop around pandemic time
- Do we want to detect long-term anormales?



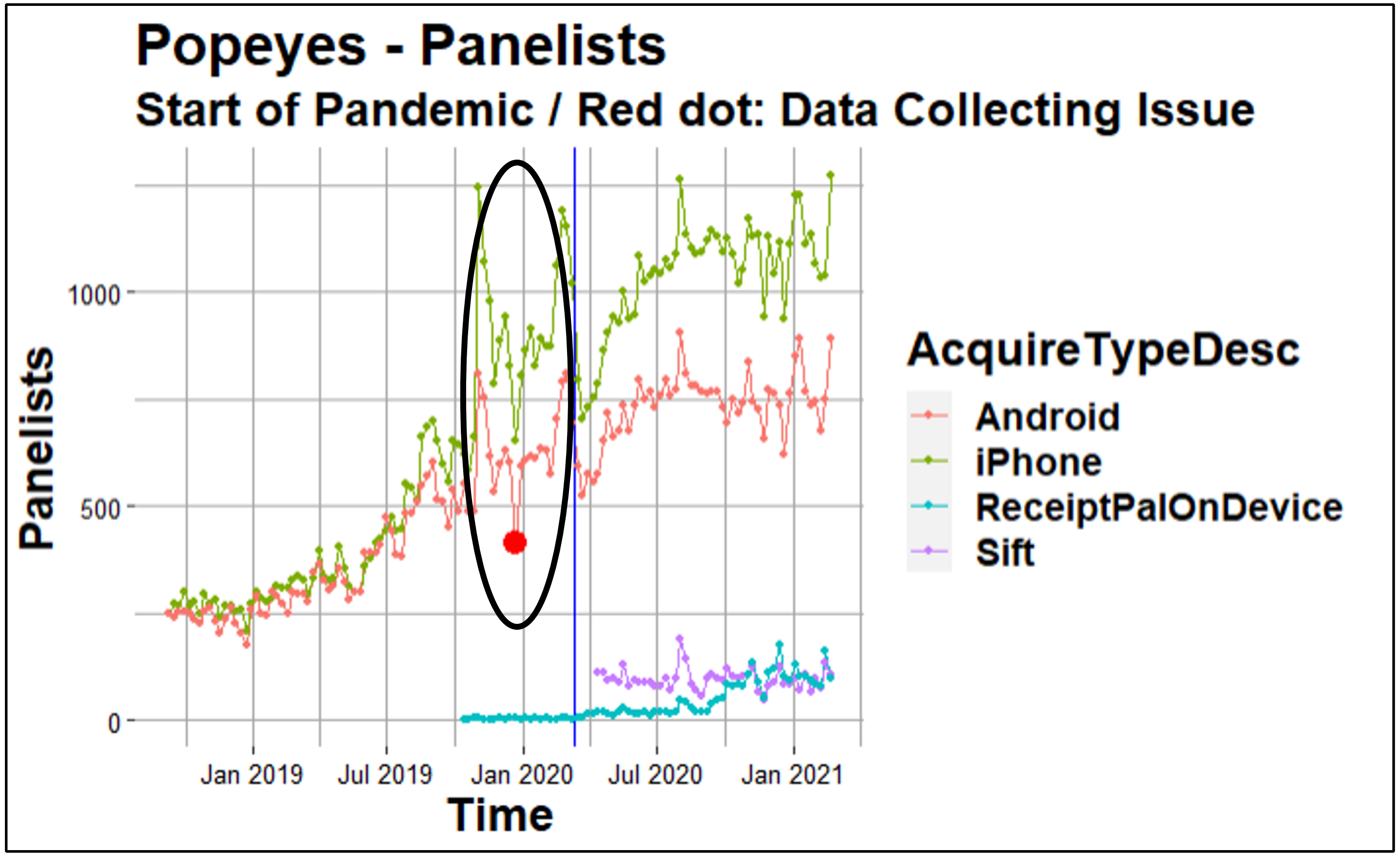
Smaller Amplitude of dips detected but not bigger ones

- Small, sharp drop of panelists using Sift around April 2019
- Huge, sharp drop from September to November 2020
- Drops were similar, but one is detected, the other is not



Dips for all Acquire- Types not marked

- Dip in Android was marked as an issue
- Dip in iPhone for a similar time period and amplitude, but was not marked



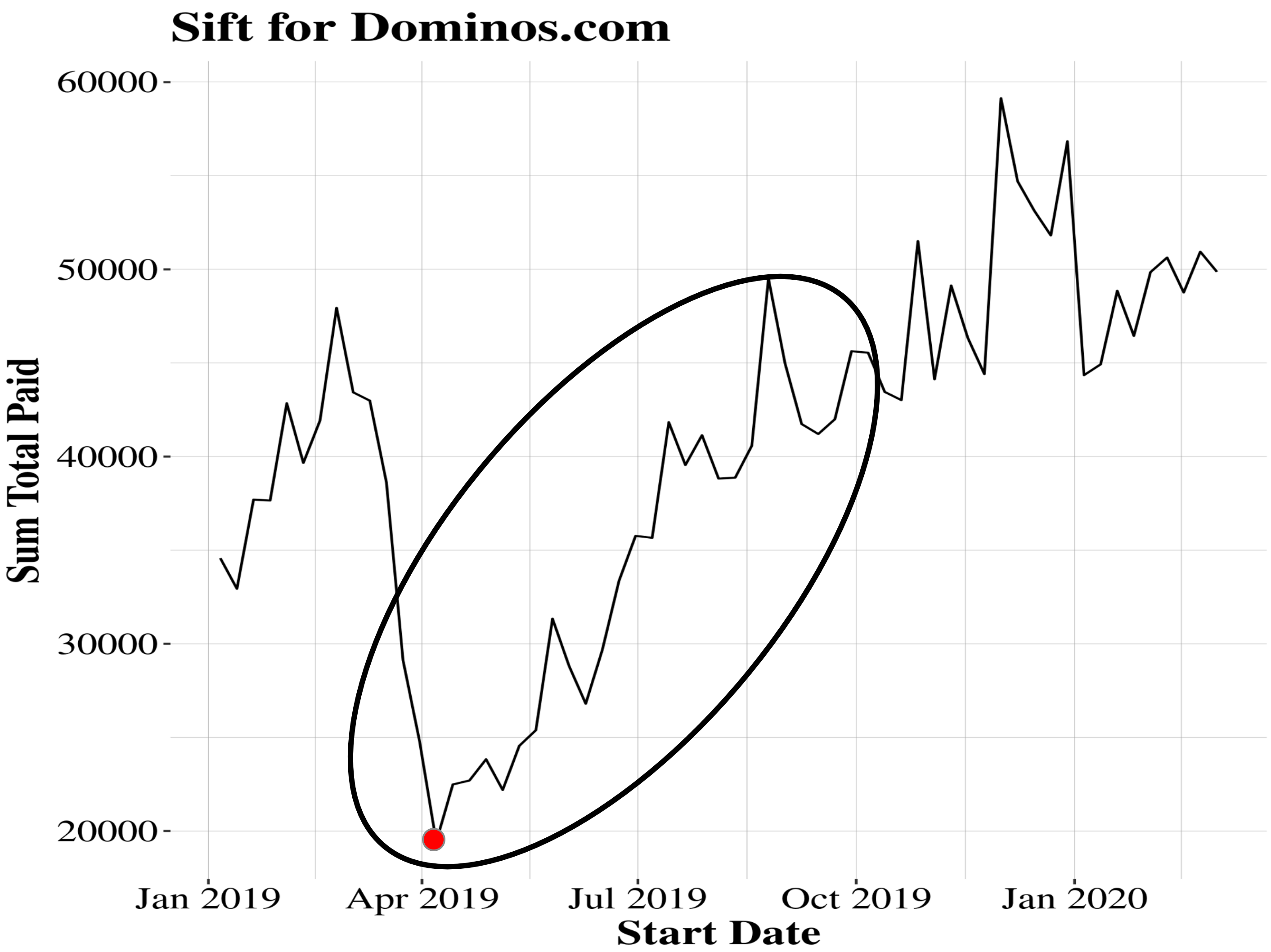
Missing data - Jersey Mike's

- Issues flagged for the dates that did not have data
- No definition given, assume missing data

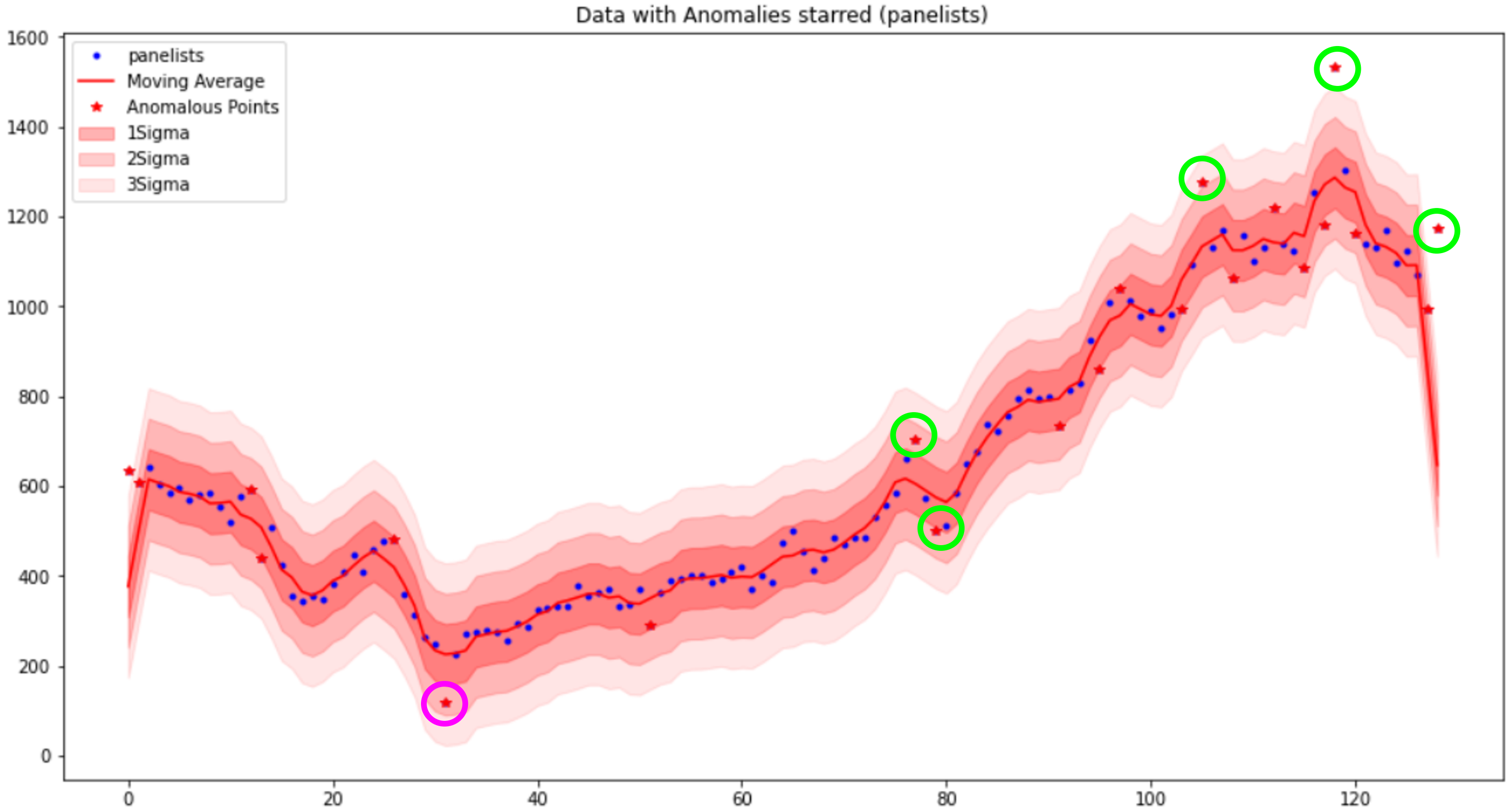
Date	Retailer Data Present?	Issue Found?
7/19/20	Yes	No
7/26/20	No	Yes
8/2/20	No	Yes
8/9/20	No	Yes
8/16/20	Yes	No
8/23/20	Yes	No
8/30/20	Yes	No
9/6/20	No	Yes
9/13/20	No	Yes
9/20/20	No	Yes
9/27/20	No	Yes
10/4/20	Yes	Yes
\vdots	\vdots	\vdots

Should Trends be Flagged?

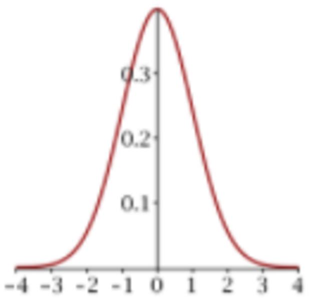
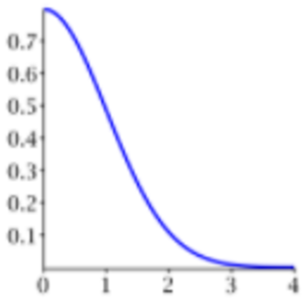
- Many of the graphs had clear trends (slow drift) that were not detected.
- We see a shift starting from April 2019 that increases until approximately November 2019
- Should we detect trends? If so, what kind of trends?



Moving Average Prototype



Half Normal Distribution

	Normal $\mathcal{N}(\mu, \sigma^2)$	**Half-normal** $\mathcal{H}(\sigma)$
		
Support	$x \in \mathbb{R}$	$x \in \mathbb{R}_{\geq 0}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\sqrt{\frac{2}{\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$
Mean	μ	$\sigma\sqrt{\frac{2}{\pi}}$
Variance	σ^2	$\sigma^2\left(1 - \frac{2}{\pi}\right)$