STATS 765 Project Milestone 3

Measuring the Impact of the 2021 Residential Tenancies Amendment Act

Chase Robertson - crob873

Goal

The goal of this project is to investigate the impact of the February 2021 Residential Tenancies Amendment Bill by observing changing rental trends on the New Zealand housing market.

Data Source

The data is sourced from the NZ Tenancy Services website. Three CSV files are provided with territorial, regional, and quarterly aggregate bond statistics. In order to translate Location Id's from the quarterly dataset into location names, it is also necessary to retrieve data from the Stats NZ site.

Data Processing

One of the first data processing tasks is to ensure each data column is imported as the correct type, to ensure proper usage later in the analysis.

Some of the ".Bond" and ".Rent" columns of each dataset import as character-type columns, which does not match their numeric contents. I found that removing the thousands-delimiting commas made converting to integer type possible.

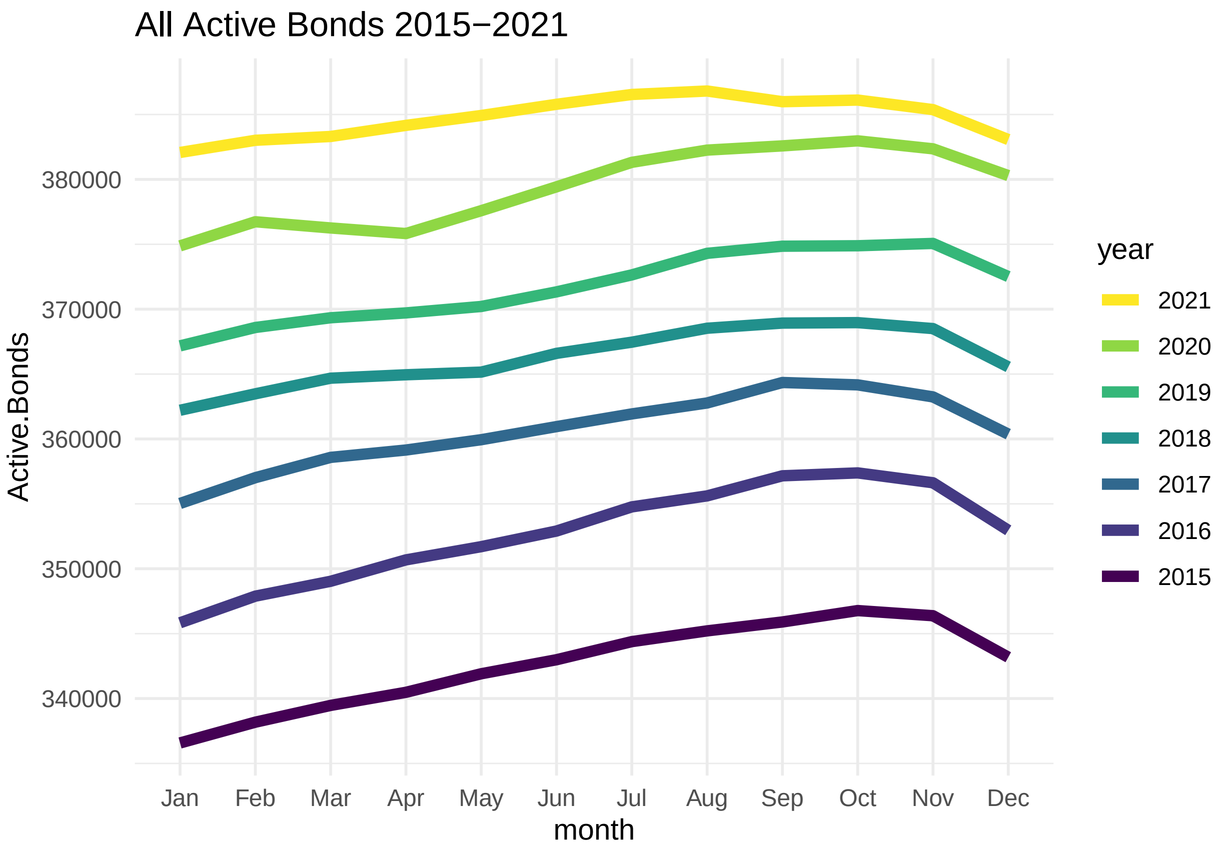
The next data processing step is to convert the `Time Frame` column into more useful `year` and `quarter` columns. The `lubridate` library makes that a simple task.

I also joined the Stats NZ location ID and location name columns with the quarterly report, so that location names could be easily included in future analysis. It was also necessary to manually add ID `-99` to get the location name `"ALL"` to join correctly. All data processing code can be seen in the Appendix.

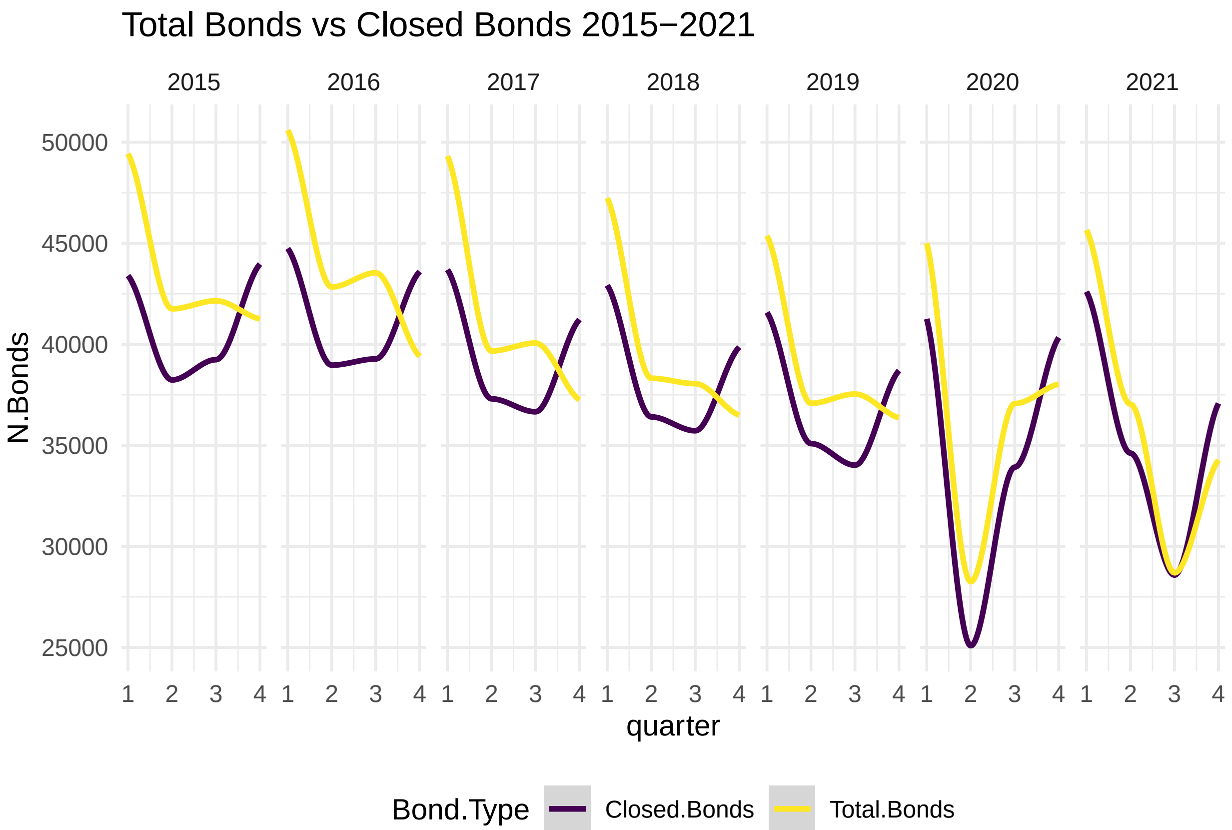
A similar set of steps was followed to import the regional and territorial datasets as well, as can also be seen in the Appendix.

Data Exploration

The first area of the data explored is the active bond numbers before and after Bill implementation, searching for noticeable differences between the monthly data before and after Feb 2021.

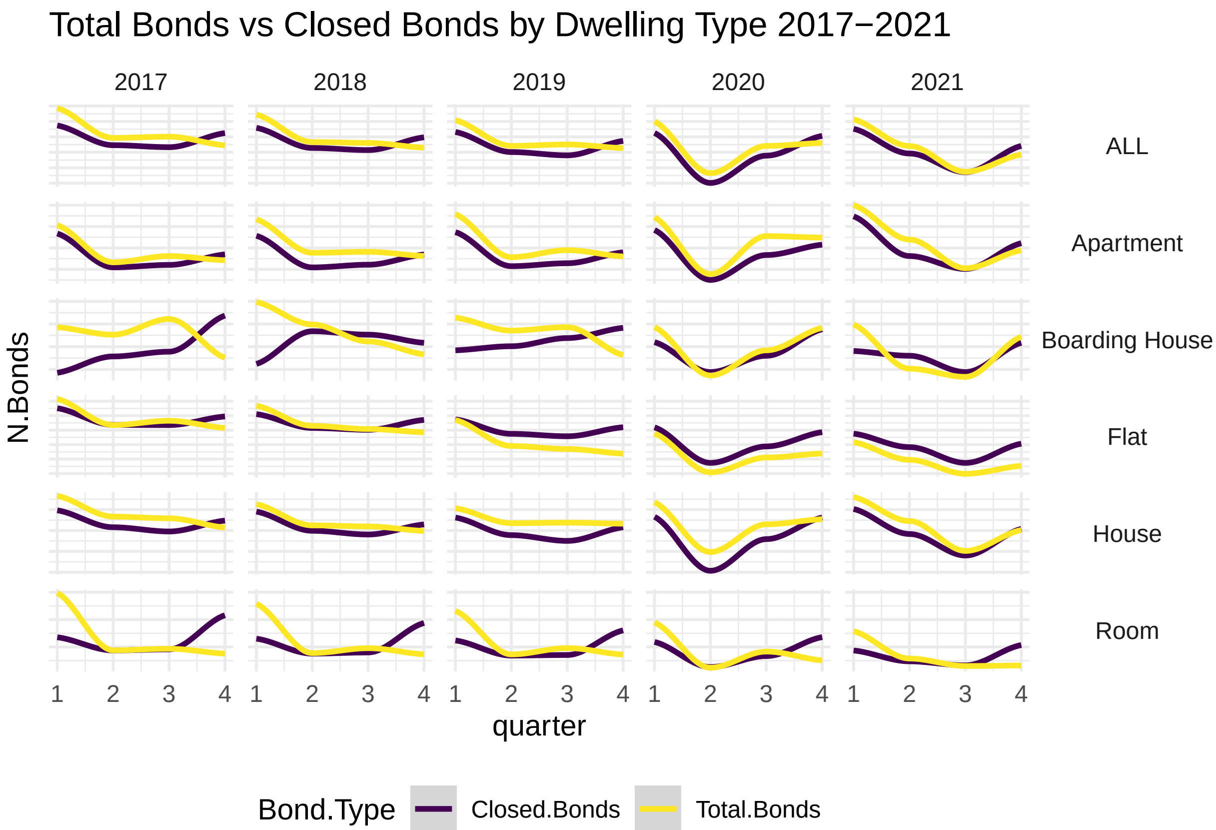


There does seem to be some down turn after Q2 2021, which suggests a slightly delayed effect, or perhaps some external factor. Let's check the number of Total and Closed Bonds as well.

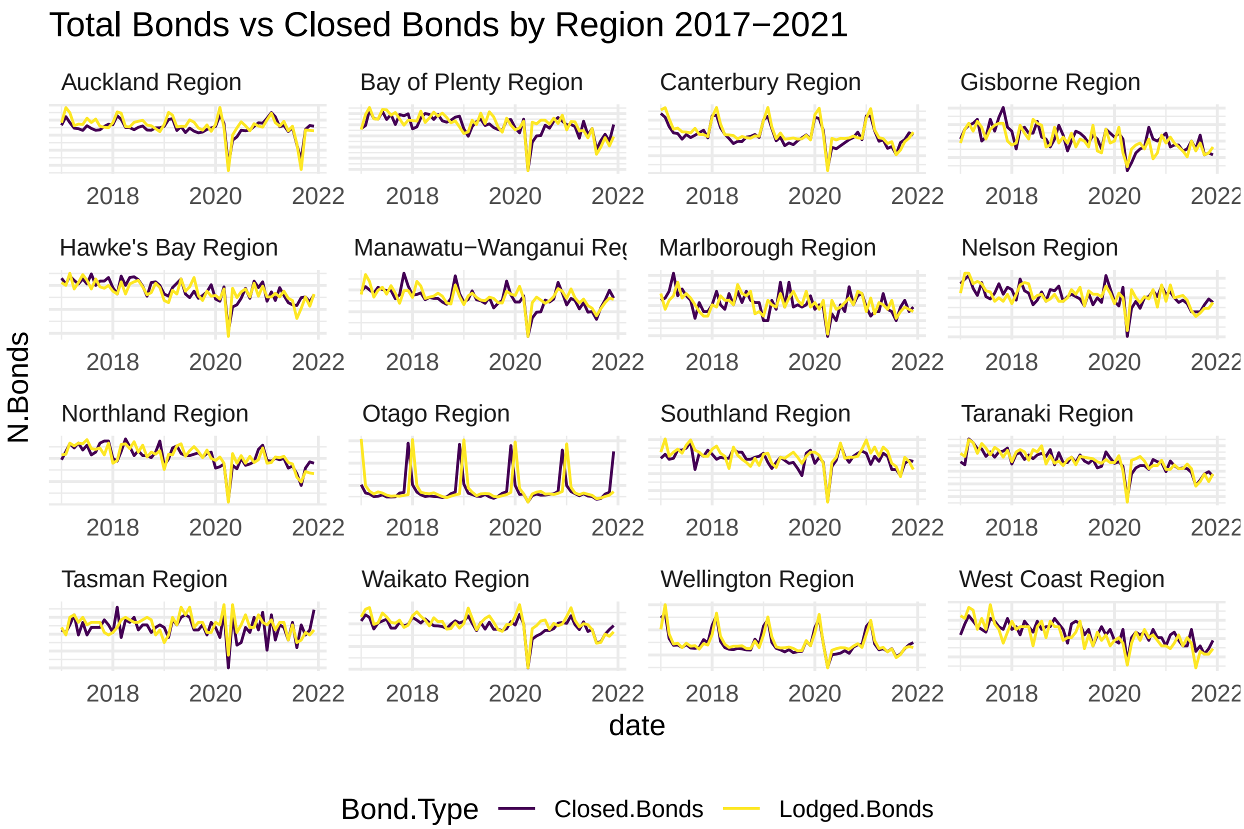


It seems that there may have been some unexpected increase in Closed Bonds toward the end of 2020 / beginning of 2021. Total bonds seem to be on the decline, though the COVID lockdown spikes make any confidence in that assessment difficult.

Let's see if the overall pattern is reflected across all types of dwellings, and narrow the historical scope by a few years.



Again, it seems that COVID lockdowns in early 2020 and late 2021 caused a significant perturbation to bond filings, making any effects from the Bill more difficult to distinguish. However, the monthly regional data may help further expose any varying patterns by location.

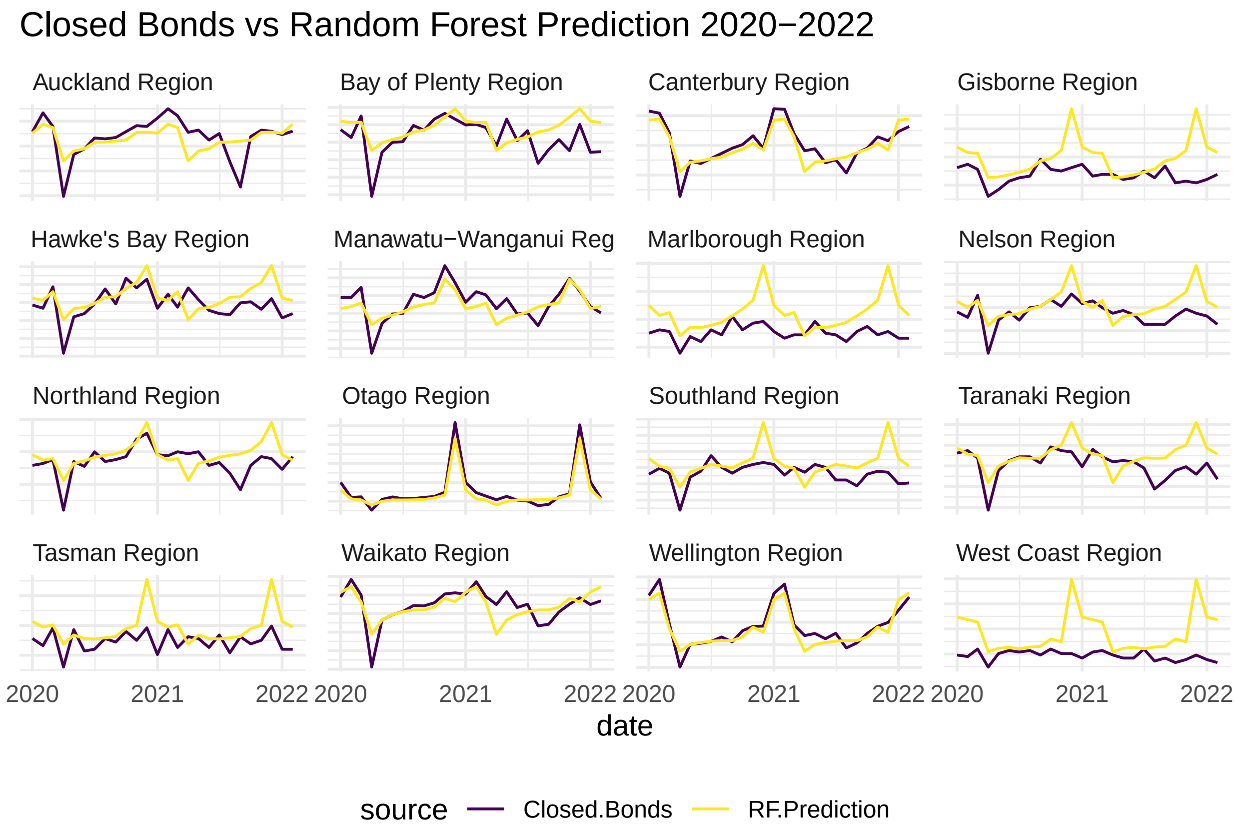


Auckland, Waikato, Wellington, and Canterbury demonstrate the upward trend in closed bonds. Those regions also harbor the largest cities in New Zealand, which begs the question: is this a pattern exclusive to urban environments? Could this be an "urban flight" phenomenon caused by COVID-19, rather than the Bill?

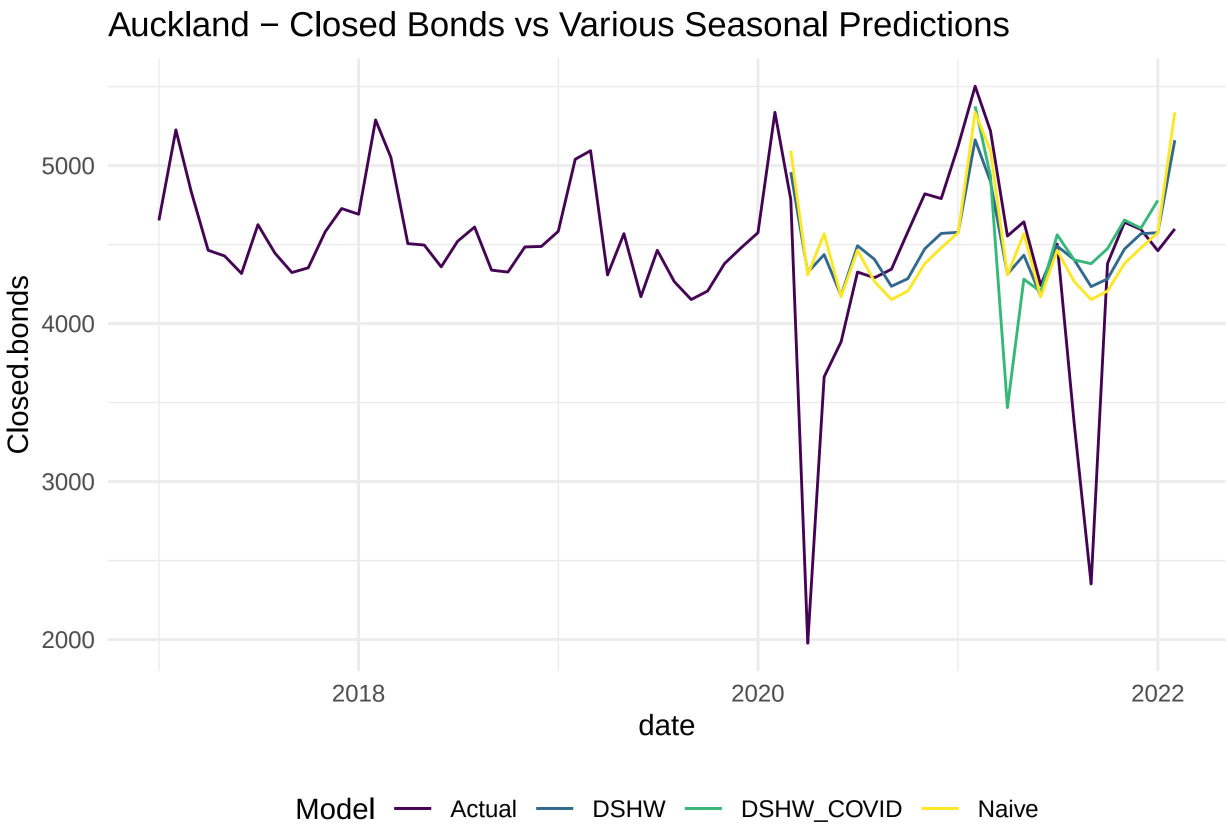
Analytical Plan

Based on the information illuminated in the data exploration, it seems that the effects of the Bill have varied across different dwelling types and locations. One potential path of analysis is constructing a paired t-test to see if there is a statistically significant difference between those bond measures in the time periods before and after the Bill. Perhaps a better way to test that hypothesis is to attempt to infer causality with a Bayesian structural time-series model. By building a Bayesian model from the data before the Bill's implementation, we can attempt to predict the data after the Bill, and compare those predictions with the actual data to infer the Bill's effect. Another way could be to conduct the same analysis with a Random Forest or naive seasonality model.

Below, a Random Forest model trained on all regional monthly data before 2021 (year, month and Location variables), predicts the number of Closed.Bonds in each location. Because it is the same model for all locations, it is not able to capture the varying seasonality of each location individually. Even so, it gives a fair baseline for how predictive models may or may not be able to give an accurate comparison to the data witnessed since before COVID in 2020.

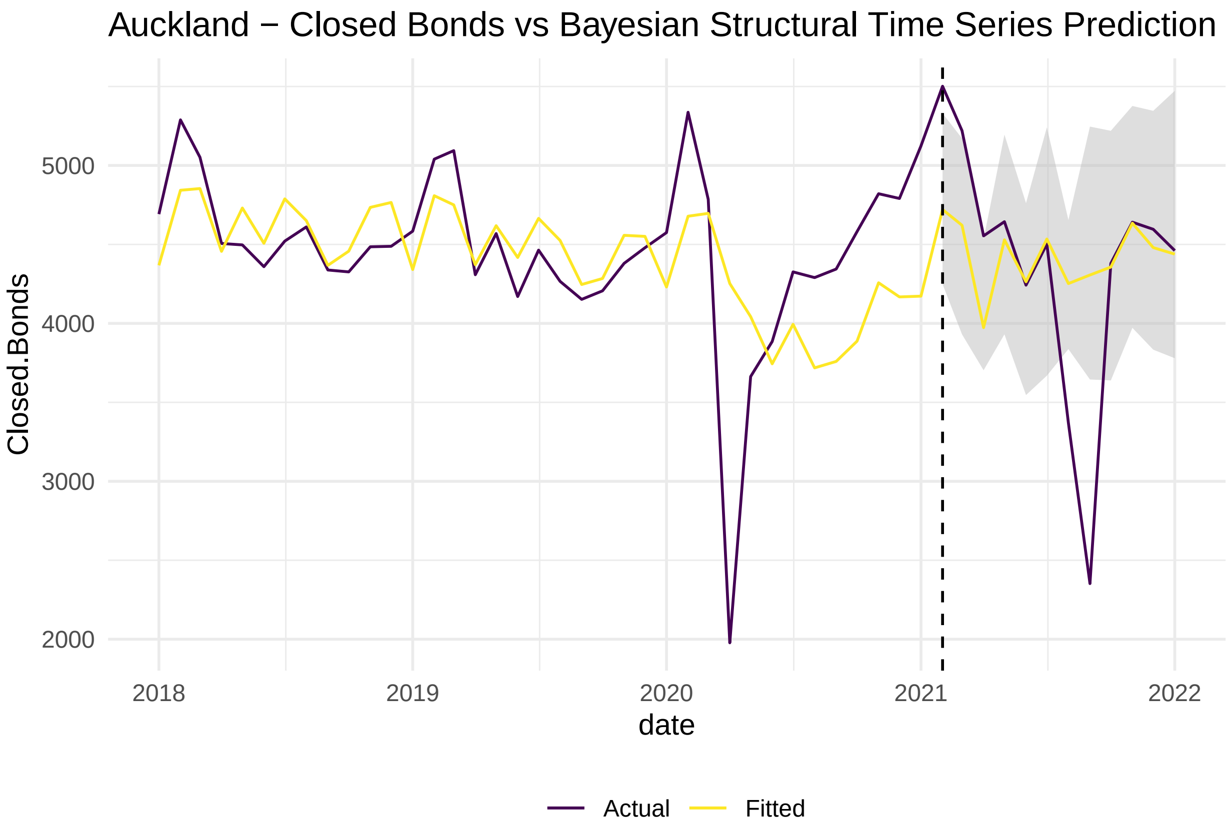


Other models that may give some insight to the effects of the 2021 Bill are naive seasonal and, likely more effective, Double-Seasonal Holt-Winters forecasting. These models are trained on data between the years 2000 and 2020 (or 2021 in the case of `DSHW_COVID`), and their predictions plotted alongside the actual Closed Bonds data since 2020 below. Only Auckland is displayed, as any change in pattern in Auckland would affect more residents than in other regions.



The models which were not trained on 2020 were able to fairly accurately predict the 2021 data, if the late 2021 Omicron lockdown is ignored. This could demonstrate the predictability of 2021's closed bond numbers in Auckland, which suggest minimal effect from the housing Bill.

Perhaps a more informative model is attempted below: a Bayesian Structural Time Series Model.



As illustrated by the gray confidence interval in the above graph, there was only one period after the Bill's implementation where the actual number of closed bonds in Auckland would not have been predicted by this model: during the Omicron lockdown in mid-to-late 2021. This is compelling evidence of the Bill's minimal effect on bond numbers, though further modeling and hypothesis testing will help strengthen this evidence.

Discussion

It is fantastic that NZ Tenancy Services provides this rental data to the public. It is clean and thorough, and provides universal ground truth for discussions about policy effectiveness. In this case, however, the data seems insufficient to prove any claims about the 2021 Bill, because of the incredibly confounding factor of COVID-19. With finding patterns in this data being the goal of this project, it is rather unsatisfactory to look hard and not find any patterns which are not completely overshadowed by lockdown patterns. Unfortunately, that seems to be the hard reality of this data.

Appendix

Appendix includes all code necessary to replicate the report, plus some extra graphs not included in the written analysis.

```
# import quarterly dataset, show names and tabulate types
quarterly <- read.csv("../data/Detailed Quarterly Tenancy.csv")
#names(quarterly)
#table(sapply(quarterly, class))

# remove comma from numeric values and convert columns
library(tidyverse)
q2 <- quarterly %>%
```

```r
    mutate(across(str_subset(names(quarterly),
                             regex(".Bond|(n|e).Rent|.Id")),
                  function(x) {as.integer(gsub(',', '', x))}))

# create year and month columns from Time.Frame
library(lubridate)
q3 <- q2 %>%
    mutate(date = mdy_hm(Time.Frame)) %>%
    mutate(year = year(date)) %>%
    mutate(quarter = quarter(date)) %>%
    mutate(date = paste(year, quarter, sep='-'))

# import and join location names
areas <- read.csv("../data/statsnzstatistical-area-2-2019-centroid-true-CSV/statistical-area-2-2019-cen
locations <- areas %>%
    mutate(ID = SA22019_V1_00) %>%
    mutate(Location = SA22019_V1_00_NAME) %>%
    select(c(ID, Location))
locations <- rbind(locations, list(-99, 'ALL'))
q.df <- left_join(q3, locations, by=c('Location.Id' = 'ID'))
#length(names(q.df))

# import and clean regional dataset
regional <- read.csv("../data/rentalbond-data-regional.csv")
r2 <- regional %>%
    mutate(across(c(Lodged.Bonds, Closed.Bonds),
                  function(x) {as.integer(gsub(',', '', x))}))
r3 <- r2 %>%
    mutate(date = ymd(Time.Frame)) %>%
    mutate(year = year(date)) %>%
    mutate(month = month(date))
r.df <- r3

#import and clean territorial dataset
tla <- read.csv("../data/rentalbond-data-tla.csv")

t2 <- tla %>%
    mutate(across(c(Lodged.Bonds, Active.Bonds),
                  function(x) {as.integer(gsub(',', '', x))}))
t3 <- t2 %>%
    mutate(date = ymd(Time.Frame)) %>%
    mutate(year = year(date)) %>%
    mutate(month = month(date)) #%>%
    #select(-c(Time.Frame, date))
t.df <- t3

# prepare for plotting
library(ggplot2)
theme_set(theme_minimal())
library(viridis)

# not pictured - quarterly Total Bonds (very skewed toward lockdowns)
q.df %>%
```

```
    filter(Location == 'ALL', Dwelling.Type == 'ALL',
           Number.Of.Beds == 'ALL', year > 2014) %>%
    ggplot(aes(x=quarter, y=Total.Bonds, colour=factor(quarter))) +
    facet_grid(~year) + geom_point(size=2) +
    scale_color_viridis(discrete=T) + theme_minimal()

# quarterly Active Bonds
rplot <- r.df %>%
    filter(Location.Id == -99, year > 2014, year < 2022) %>%
    mutate(year = factor(year))

rplot %>%
    ggplot(aes(x=month, y=Active.Bonds, group=year, colour=year)) +
    geom_line(size=2) +
    scale_colour_viridis(discrete=T) +
    guides(colour = guide_legend(reverse = T)) +
    scale_x_discrete('month', limits=month.abb[1:12]) +
    ggtitle("All Active Bonds 2015-2021")

# not pictured - Closed Bonds (skewed by lockdowns)
q.df %>%
    filter(Location == 'ALL', Dwelling.Type == 'ALL',
           Number.Of.Beds == 'ALL', year > 2014) %>%
    ggplot(aes(x=quarter, y=Closed.Bonds, colour=factor(quarter))) +
    facet_grid(~year) + geom_point(size=2) +
    scale_color_viridis(discrete=T) + theme_minimal()

# Total and closed annual bonds per year
q.type = q.df %>%
    filter(Location == 'ALL', Dwelling.Type == 'ALL',
           Number.Of.Beds == 'ALL', year > 2014) %>%
    pivot_longer(c(Total.Bonds, Closed.Bonds),
                 names_to="Bond.Type", values_to="N.Bonds")

q.type %>%
    ggplot(aes(x=quarter, y=N.Bonds, group=Bond.Type, colour=Bond.Type)) +
    facet_grid(~year) +
    geom_smooth() +
    scale_colour_viridis(discrete=T) +
    theme(legend.position = "bottom") +
    ggtitle('Total Bonds vs Closed Bonds 2015-2021')

# total and closed annual bonds per dwelling type
q.building.closed = q.df %>%
    filter(Location == 'ALL', Number.Of.Beds == 'ALL', year > 2016) %>%
    pivot_longer(c(Total.Bonds, Closed.Bonds),
                 names_to="Bond.Type", values_to="N.Bonds")

q.building.closed %>%
    ggplot(aes(x=quarter, y=N.Bonds, group=Bond.Type, colour=Bond.Type)) +
    facet_grid(rows=vars(Dwelling.Type), cols=vars(year), scales='free') +
    geom_smooth() +
    scale_colour_viridis(discrete=T) +
```

```r
    theme(legend.position = "bottom", axis.text.y = element_blank(),
          strip.text.y.right = element_text(angle = 0)) +
    ggtitle('Total Bonds vs Closed Bonds by Dwelling Type 2017-2021')

# lodged and closed bonds by region
r.loc = r.df %>%
    filter(Location.Id > 0, year > 2016, year < 2022) %>%
    pivot_longer(c(Lodged.Bonds, Closed.Bonds),
                 names_to="Bond.Type", values_to="N.Bonds")

r.loc %>%
    ggplot(aes(x=date, y=N.Bonds, colour=Bond.Type)) +
    facet_wrap(~Location, scales='free') +
    geom_line() +
    scale_colour_viridis(discrete=T) +
    theme(strip.text.x = element_text(hjust = 0.1), legend.position = "bottom",
          axis.text.y = element_blank()) +
    ggtitle('Total Bonds vs Closed Bonds by Region 2017-2021')

# plot rent quartiles by region
r.time = r.df %>%
    filter(Location.Id > 0, year > 2014, year < 2022) %>%
    pivot_longer(c(Median.Rent, Upper.Quartile.Rent, Lower.Quartile.Rent),
                 names_to="Type", values_to="Value") %>%
    group_by(year, Type, Location) %>%
    summarise(Annual.Mean = mean(Value))

r.time %>%
    ggplot(aes(x=year, y=Annual.Mean, colour=Type)) +
    facet_wrap(~Location, scales="free") +
    geom_line() +
    scale_colour_viridis(discrete=T) +
    theme_minimal() +
    theme(strip.text.x = element_text(hjust = 0.1),
          legend.position = "bottom") +
    scale_x_discrete(name="year",limits=c(2015,2018,2021))

# train and plot predictions of Random Forest
library(ranger)

fdata <- r.df %>%
  filter(Location.Id > 0)
train <- fdata %>%
  filter(year >= 2000, year < 2021)
test <- fdata %>%
  filter(year > 2019)

forest <- ranger(Closed.Bonds ~ year + month + Location, data=train,
                 mtry=2, importance="impurity")
p <- predict(forest, test)

fplot <- cbind(test, RF.Prediction=p$predictions) %>%
  pivot_longer(c(Closed.Bonds, RF.Prediction), names_to='source', values_to='value')
```

```r
fplot %>%
  ggplot(aes(x=date, y=value, colour=source)) +
  facet_wrap(~Location, scales='free_y') +
  geom_line() +
  scale_colour_viridis(discrete=T) +
  scale_x_date(date_breaks="1 year", date_labels='%Y') +
  theme(strip.text.x = element_text(hjust = 0.1), axis.text.y = element_blank(),
        legend.position = "bottom", axis.title.y=element_blank()) +
  ggtitle('Closed Bonds vs Random Forest Prediction 2020-2022')

# fit and plot naive and double-seasonal Holt-Winters forecasting
library(forecast)
year_start = 2000

timesdf <- r.df %>%
  filter(Location.Id == 2, year >= year_start) %>%
  arrange(year, month) %>%
  dplyr::select(Closed.Bonds)

timesr = ts(timesdf, start=year_start, end=2022, deltat=1/12)

training <- window(timesr, start=c(year_start, 1), end=c(2020, 2))
validation <- window(timesr, start=c(2020, 2))

train_covid <- window(timesr, end=c(2021, 1))
valid_covid <- window(timesr, start=c(2021, 2))

naive = snaive(training, h=length(validation), lambda='auto')
dshw <- dshw(training, period1=4, period2=12, h=length(validation))
dshw_covid <- dshw(train_covid, period1=4, period2=12, h=length(valid_covid))

naive.df <- data.frame(date=zoo::as.Date(time(naive$mean)), Naive=naive$mean)
dshw.df <- data.frame(date=zoo::as.Date(time(dshw$mean)), DSHW=dshw$mean)
dshw_covid.df <- data.frame(date=zoo::as.Date(time(dshw_covid$mean)), DSHW_COVID=dshw_covid$mean)

timesplot <- r.df %>%
  filter(Location.Id == 2, year >= 2017) %>%
  arrange(year, month) %>%
  left_join(naive.df, by=(date='date')) %>%
  left_join(dshw.df, by=(date='date')) %>%
  left_join(dshw_covid.df, by=(date='date')) %>%
  pivot_longer(c(Naive, DSHW, DSHW_COVID), names_to='Model', values_to='Closed.bonds')

timesplot %>%
  ggplot(aes(x=date, y=Closed.bonds, colour=Model)) +
  geom_line(aes(x=date, y=Closed.Bonds, colour='Actual')) +
  geom_line() +
  scale_colour_viridis(discrete=T) +
  theme(legend.position = 'bottom') +
  ggtitle('Auckland - Closed Bonds vs Various Seasonal Predictions')

# fit and plot Bayesian Structural Time Series Model
library(bsts)
```

```r
Y <- train_covid
y <- log10(Y)

ss <- AddLocalLinearTrend(list(), y)
ss <- AddSeasonal(ss, y, nseasons=4)
ss <- AddSeasonal(ss, y, nseasons=12)
bsts.model <- bsts(y, state.specification=ss, niter=500, ping=0)

burn <- SuggestBurn(0.1, bsts.model)

bsts.p <- predict.bsts(bsts.model, horizon=length(valid_covid), burn=burn, quantiles=c(0.025, 0.975))

d2 <- data.frame(c(10^as.numeric(-colMeans(bsts.model$one.step.prediction.errors[-(1:burn),])+y),
                   10^as.numeric(bsts.p$mean)),
                 as.numeric(timesr), as.Date(time(timesr)))
names(d2) <- c("Fitted", "Actual", "Date")

post.interval <- cbind.data.frame(
  10^as.numeric(bsts.p$interval[1,]),
  10^as.numeric(bsts.p$interval[2,]),
  subset(d2, Date>=as.Date('2021-02-01'))$Date)
names(post.interval) <- c("Lower", "Upper", "Date")

### Join intervals to the forecast
d3 <- left_join(d2, post.interval, by="Date")
d3.plot <- subset(d3, year(Date)>2017)

### Plot actual versus predicted with credible intervals for the holdout period
ggplot(data=d3.plot, aes(x=Date)) +
  geom_ribbon(aes(ymin=Lower, ymax=Upper), fill="grey", alpha=0.5) +
  geom_line(aes(y=Actual, colour = "Actual")) +
  geom_line(aes(y=Fitted, colour = "Fitted")) +
  scale_colour_viridis(discrete=T) +
  ylab("Closed.Bonds") + xlab("date") +
  geom_vline(xintercept=as.numeric(as.Date("2021-02-01")), linetype=2) +
  theme(legend.title = element_blank(), legend.position = 'bottom') +
  ggtitle('Auckland - Closed Bonds vs Bayesian Structural Time Series Prediction')
```