Comparing Overnight Occupancy in Homeless Shelters During 2014 Oil downturn and Covid-19 Pandemic

DATA 602

October 18, 2020

Jordan Keelan / Ali Raza / Abrie Le Roux

Abstract

The purpose of this report is to apply specific statistical methods to help understand overnight occupancy in homeless shelters across the province of Alberta, from 2013 to 2022. Given the fact that Alberta's economy is closely tied to the oil and gas industry, we expected an increase in homeless shelter occupancy following both the Fall 2014 crash in oil prices and the 2020 crash in oil prices due to the COVID-19 pandemic. Furthermore, due to the various business closures and quarantine rules imposed by the Government of Alberta, we expected the COVID-19 pandemic to have a large effect on homeless shelter occupancy. This brings us to our two topics of investigation:

Topic of Investigation #1 – Which economic downturn, brought on by the crash in oil prices and/or COVID-19 pandemic had a larger effect on homeless shelter occupancy?

- Initially, the plan was to compare the mean monthly total occupancy in all shelters in Alberta during a predetermined period, both post 2014 and post 2020 downturns.
- The analysis showed a significantly higher shelter occupancy rate in 2014. This was an unexpected result, but a line plot did show a decreasing trend in shelter occupancy from 2013 to 2019, a sharper drop in total occupancy in 2020 and a sharp increase in the winter of 2021/2022. The decision was made to use linear regression to model the decreasing trend in shelter occupancy (on a dataset ending in 2019, before COVID) and using that to estimate the expected Q2 2022 value (Q2 2022 is the last full quarter in the dataset), had the COVID-19 pandemic not occurred. From this, we can better understand the effect the pandemic had on shelter occupancy.

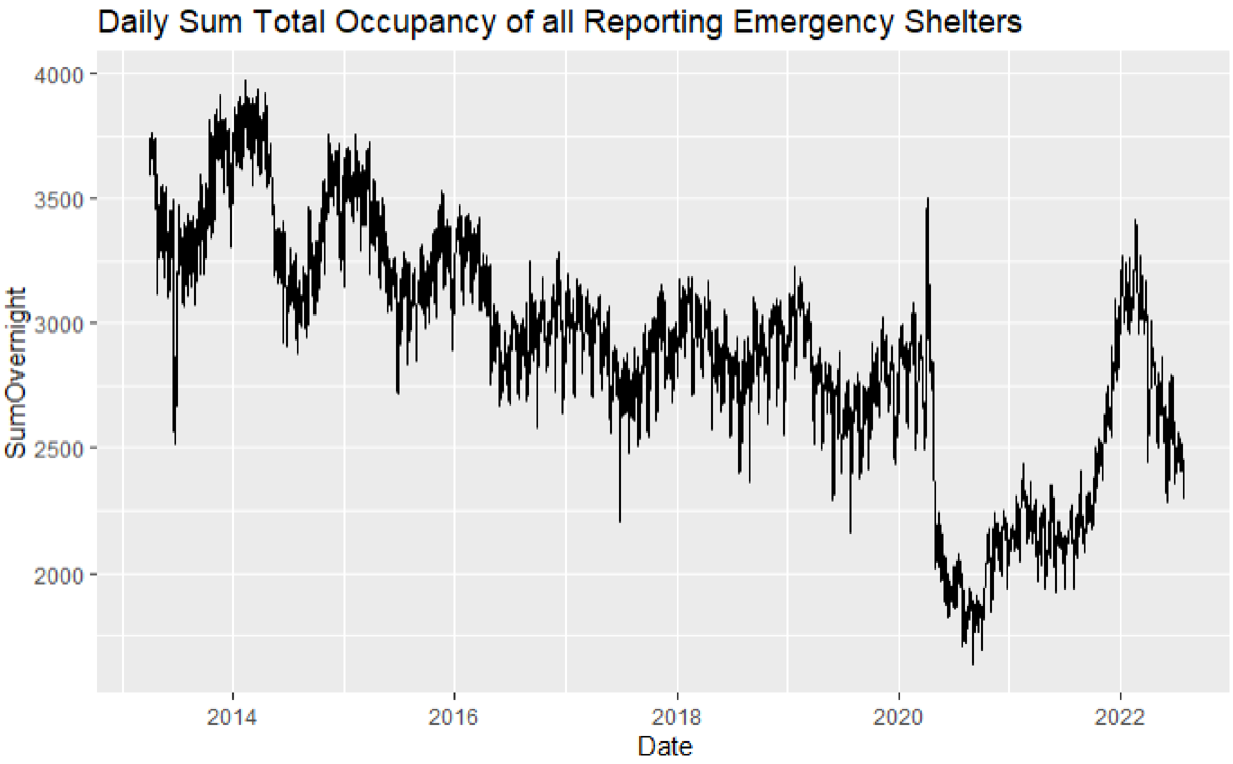
Topic of Investigation #2 – Which economic downturn resulted in a higher women's-only shelter occupancy, as a proportion of the total monthly occupancy?

- *"The 'shadow' pandemic"* (Sawhney) is an article on the Government of Alberta Website that discusses the increased domestic violence as a result of government mandated quarantining. The most obvious reasoning being that couples and families were forced to spend more time together and indoors. In addition, the problem was compounded by joblessness that resulted in the abused often not being able to afford to move out of an abusive home.
- This investigation was performed through a bootstrap analysis of the difference in proportion of women's shelters vs total shelter occupancy both after the 2014 oil crash and after the 2020 oil crash. In addition a permutation test was used to test the hypothesis.

The Dataset

Our dataset is "*Emergency Shelters Daily Occupancy AB - Emergency Shelters Daily Occupancy AB - 2013-22*" (Alberta Government, 2022). The data was accessed via The Government of Alberta Open Data platform and published by the Community and Social Services sector. It is provided with an *Open Government License – Alberta*, which gives us permission to use this dataset in any medium or form. Our dataset contains aggregated data from the years 2013-2022 which ensures us that it is not antiquated and useful for the timelines we wish to analyze.

The dataset consists of daily occupancy data (both daytime only and overnight) from 16 municipalities in Alberta over several different shelter types. These shelter types include adult emergency, women emergency, and intoxicated persons shelters. One of the shelter types that is included in the dataset are COVID-19 Isolation Sites. We removed this shelter type from the dataset as this shelter type is only for the purposes of quarantine if positive for COVID-19 or for close contacts of positive cases. In addition, we only considered overnight shelter stays in this analysis, as those are a better indication of true homelessness.



<u>Figure 1: Daily Sum of Total Occupancy for All Reporting Shelters from Q2 2013 to Q2 2022.</u>

As you can see above, the total daily shelter overnight occupancy in the province of Alberta shows a steady, negative decline from the start of the dataset, until the first quarter of 2020. At that point, shelter occupancy appears to sharply drop, before picking up again in 2021 and showing a sharp increase in 2022. This showed us that a statistical inference analysis on total occupancy after the 2014 downturn and after the 2020 downturn would not fully explain the topic of investigation due to various factors:

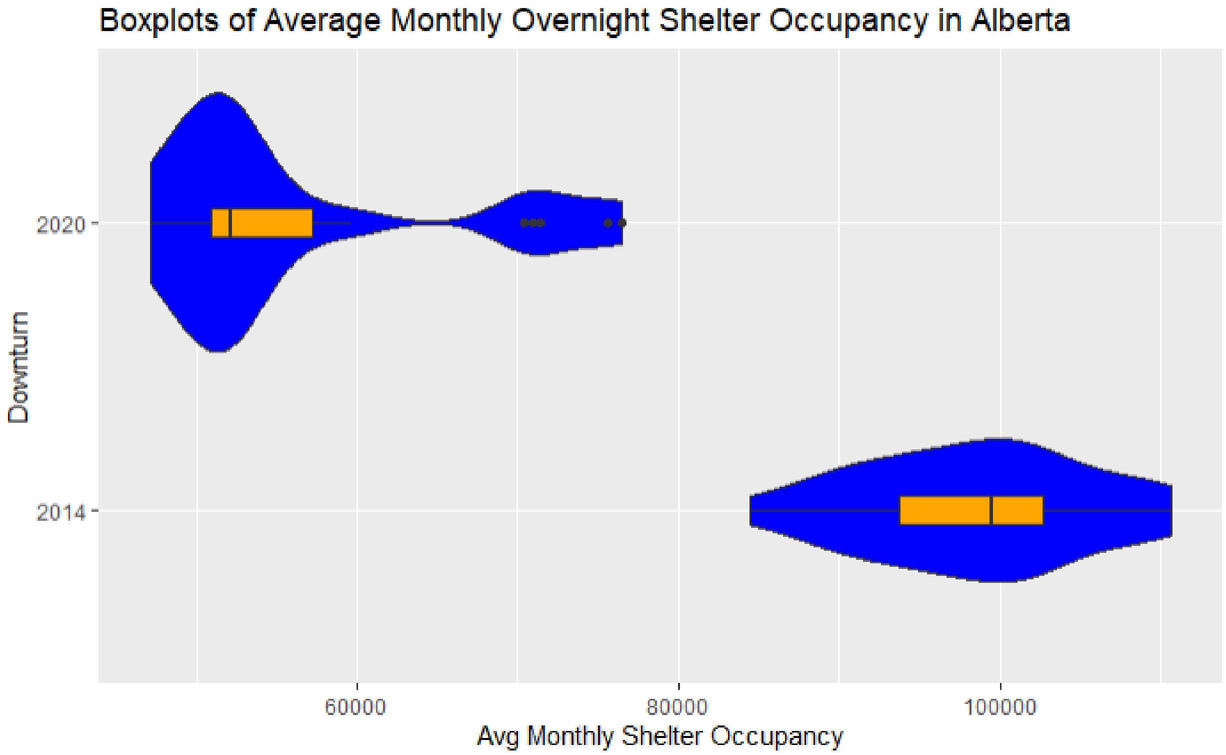
1. The sharp drop in shelter occupancy in 2020-2021 is inferred to not be due to less people in need of shelter, but rather due to a lag associated with the shelters' abilities to house people while a highly contagious virus is spreading. As was the case with various "essential" businesses or institutions, it took several months for shelters to adapt to the situation and transform their systems to allow for indoor social distancing.
2. Fear of infection could have deterred certain people from visiting shelters during the early phases of the pandemic.
3. The development and approval of multiple vaccines in late 2020 helped the above-mentioned institutions adapt even faster to living with the virus, possibly explaining the sharp rise in the winter of 2021/2022.

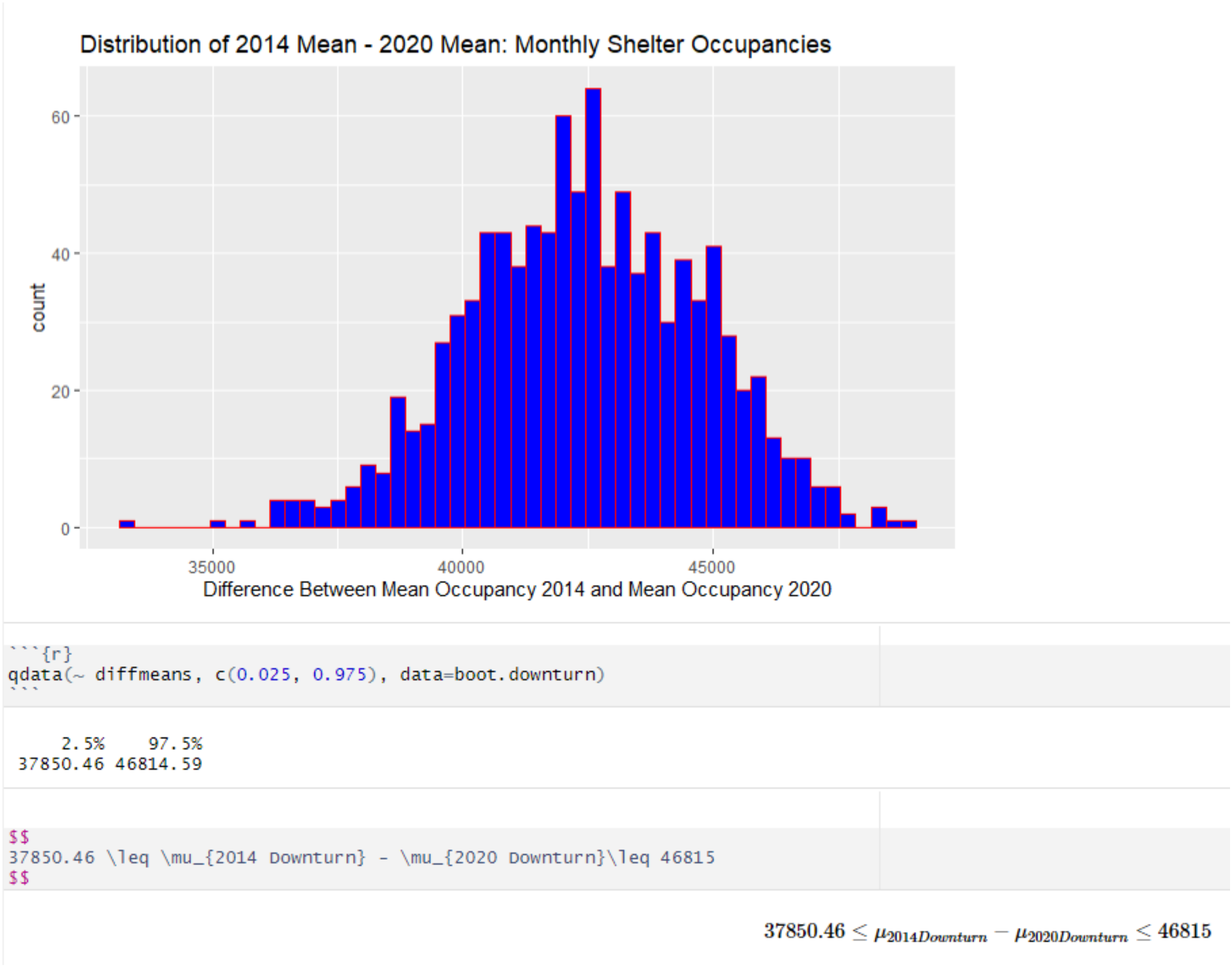
Topic of Investigation #1

The initial work conducted on this topic was to understand the difference between occupancy after both the 2014 and 2020 downturns. As the dataset contains occupancy from April 2013 until the end of July 2022, we only have 28 full months of data after the start of the oil crash in March of 2020. Therefore, we modified the parent data into two datasets of 24 months each. This allowed for equal representation of winter months in the data, which is often associated with increased occupancy. So the datasets were divided as follows:

Set 1: October 2014 to September 2016 (inclusive); further referred to as 2014 Downturn

Set 2: April 2020 to March 2022 (Inclusive); further referred to as 2020 Downturn or COVID-19




Distribution of 2014 Mean - 2020 Mean: Monthly Shelter Occupancies
count
60
40
20
0
35000
40000
45000
Difference Between Mean Occupancy 2014 and Mean Occupancy 2020

~~~
```{r}
qdata(~ diffmeans, c(0.025, 0.975), data=boot.downturn)
```
~~~

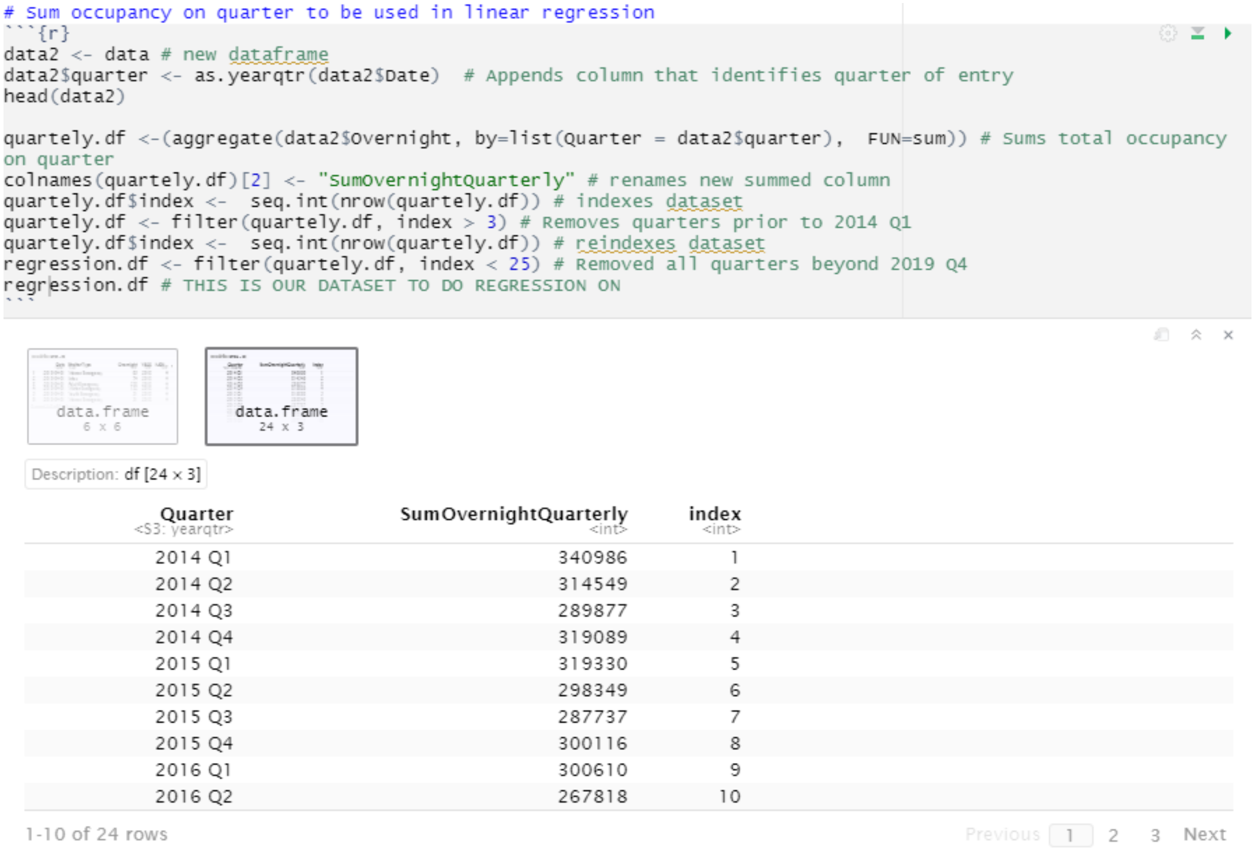
```
     2.5%    97.5%
 37850.46 46814.59
```

```
$$
37850.46 \leq \mu_{2014 Downturn} - \mu_{2020 Downturn}\leq 46815
$$
```

$$37850.46 \leq \mu_{2014 Downturn} - \mu_{2020 Downturn}\leq 46815$$

The visualizations above show how the 2014 shelter occupancy greatly exceeded the 2020 shelter occupancy. This confirms what we inferred from our initial visualizations.

To allow for linear regression analysis on time-based data, we first had to divide our dataset into quarters. This is starting with Q1 2014 and ending with Q4 2019. While the dataset starts in Q2 2013, we wanted to include an equal number of summer and winter quarters to allow for the best possible curve fit. As expected, occupancy increases in the winter months. The code is below:



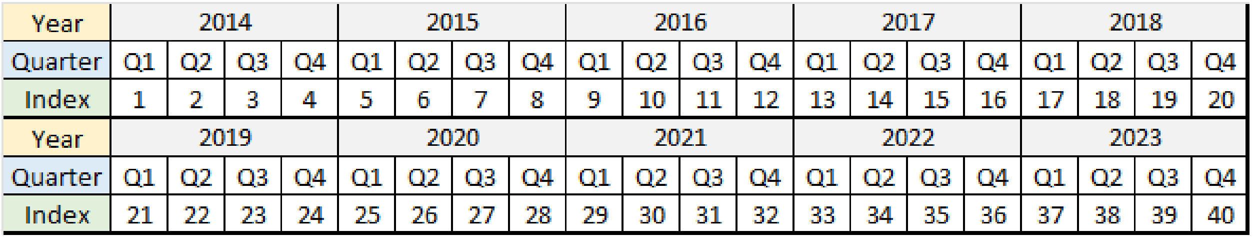
~~~
# Sum occupancy on quarter to be used in linear regression
```{r}
data2 <- data # new dataframe
data2$quarter <- as.yearqtr(data2$Date)  # Appends column that identifies quarter of entry
head(data2)

quartely.df <-(aggregate(data2$Overnight, by=list(Quarter = data2$quarter),  FUN=sum)) # Sums total occupancy
on quarter
colnames(quartely.df)[2] <- "SumOvernightQuarterly" # renames new summed column
quartely.df$index <-  seq.int(nrow(quartely.df)) # indexes dataset
quartely.df <- filter(quartely.df, index > 3) # Removes quarters prior to 2014 Q1
quartely.df$index <-  seq.int(nrow(quartely.df)) # reindexes dataset
regression.df <- filter(quartely.df, index < 25) # Removed all quarters beyond 2019 Q4
regression.df # THIS IS OUR DATASET TO DO REGRESSION ON
```
~~~

Description: df [24 × 3]

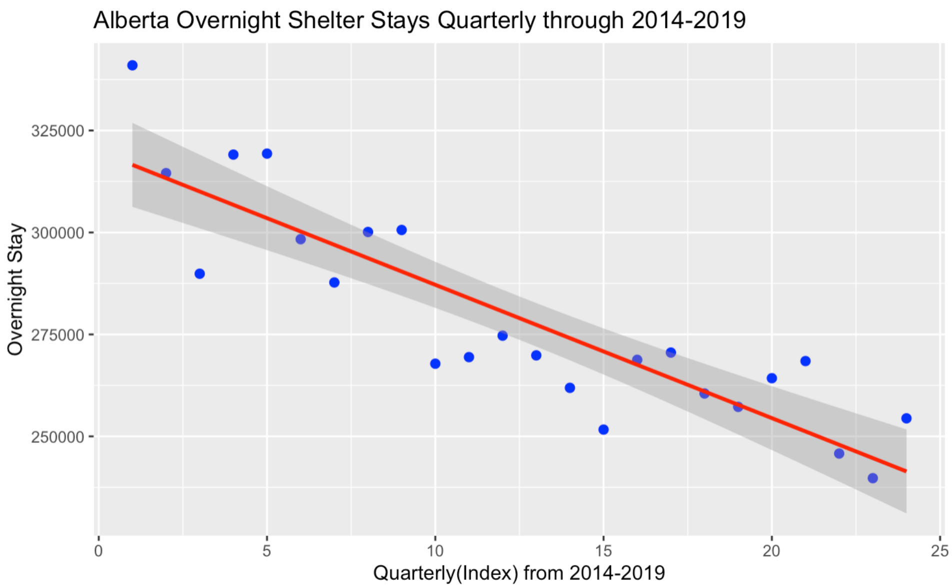
```
     Quarter        SumOvernightQuarterly    index
<S3: yearqtr>                       <int>    <int>
     2014 Q1                       340986        1
     2014 Q2                       314549        2
     2014 Q3                       289877        3
     2014 Q4                       319089        4
     2015 Q1                       319330        5
     2015 Q2                       298349        6
     2015 Q3                       287737        7
     2015 Q4                       300116        8
     2016 Q1                       300610        9
     2016 Q2                       267818       10
1-10 of 24 rows                       Previous  1  2  3  Next
```

Our regression analysis with respect to our model is visualized below where we wish to investigate the correlation of our data points through the summed quarterly overnight stays through the years 2014-2019. Plotting the data yields the following:

Note we have given our quarters an index in order to model them in decimal numbers–rather than date–on the x-axis. That index dictionary is shown below:



Year	2014				2015				2016				2017				2018			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Year	2019				2020				2021				2022				2023			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Index	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40



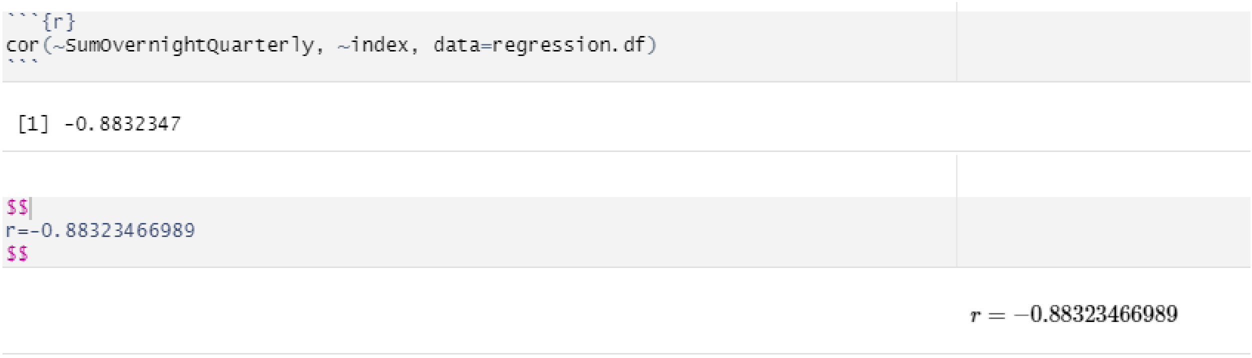
We will attempt to express the response variable of our overnight quarterly summed occupancy in our shelter types as a linear function of the predictor variable to the Quarters (Index) throughout the years we wish to investigate from our data set above.

Therefore the linear regression model we wish to estimate is …

$$\hat{R}_{SumOvernightQuarterly,i} = \beta_0 + \beta_1 * \hat{R}_{Quarter,i} + e_i$$

As you can see, there appears to be a negative linear correlation, meaning average shelter occupancy is decreasing over the time period specified.

Using the cor() function, we find a r-value of -0.8832, indicating a strong negative correlation.



~~~
```{r}
cor(~SumOvernightQuarterly, ~index, data=regression.df)
```
~~~

```
 [1] -0.8832347
```

```
$$
r=-0.88323466989
$$
```

$$r = -0.88323466989$$

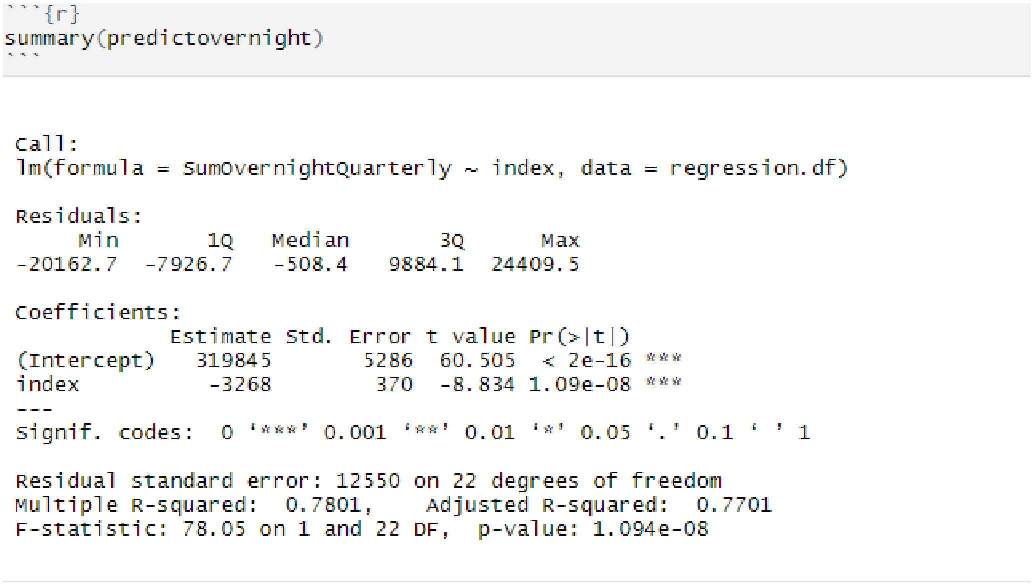
In our next step to estimate our linear regression model we will find our estimates for the model.

Estimating The Model:

From the use of the lm function via R we get our estimated linear regression model to be…

$$\hat{R}_{SumOvernightQuarterly,i} = 319844.866 - 3268.403 * \hat{R}_{Quarter,i} \qquad \text{(Note: There is no } e_i \text{ term on the estimate of the model)}$$

We can interpret our estimation model above by saying when the quarter increases by 1 unit, then the overnight quarterly occupancy rate will decrease by an average of -3268 people per quarter. When the quarter is 0 (start of the dataset) the homeless shelter on average has an occupancy in that quarter is 319844..



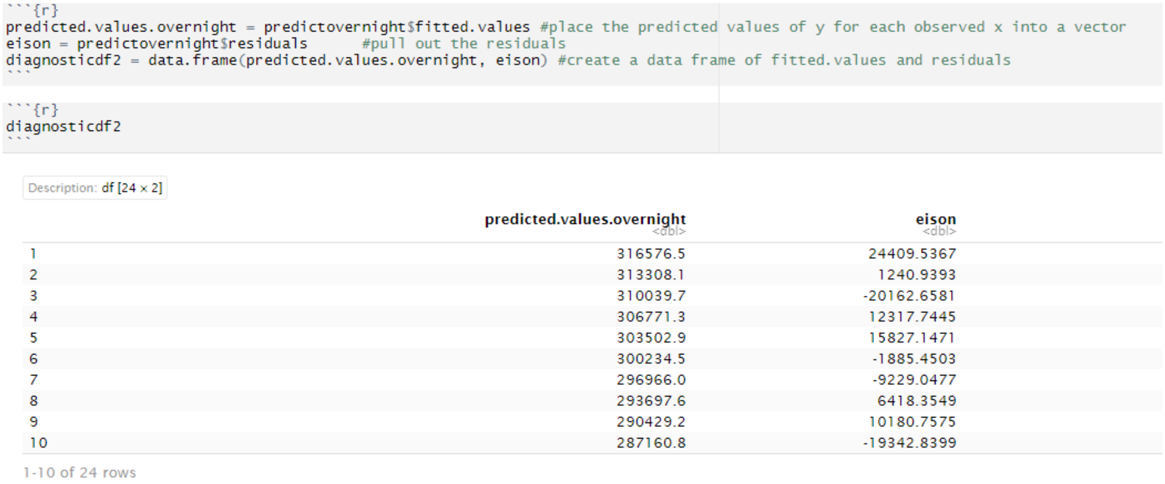
~~~
```{r}
summary(predictovernight)
```
~~~

```
Call:
lm(formula = SumOvernightQuarterly ~ index, data = regression.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-20162.7  -7926.7   -508.4   9884.1  24409.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   319845       5286  60.505  < 2e-16 ***
index          -3268        370  -8.834 1.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12550 on 22 degrees of freedom
Multiple R-squared:  0.7801,	Adjusted R-squared:  0.7701 
F-statistic: 78.05 on 1 and 22 DF,  p-value: 1.094e-08
```

$$r^2 = 0.77010819$$

In addition, the calculated R-squared value is equal to 0.7701, which tells us that approximately 77% of the variation of overnight shelter occupancy is explained by its negative linear relationship with/dependency on the quarter of the year.

Condition Checking:



~~~
```{r}
predicted.values.overnight = predictovernight$fitted.values #place the predicted values of y for each observed x into a vector
eison = predictovernight$residuals      #pull out the residuals
diagnosticdf2 = data.frame(predicted.values.overnight, eison) #create a data frame of fitted.values and residuals
```
~~~

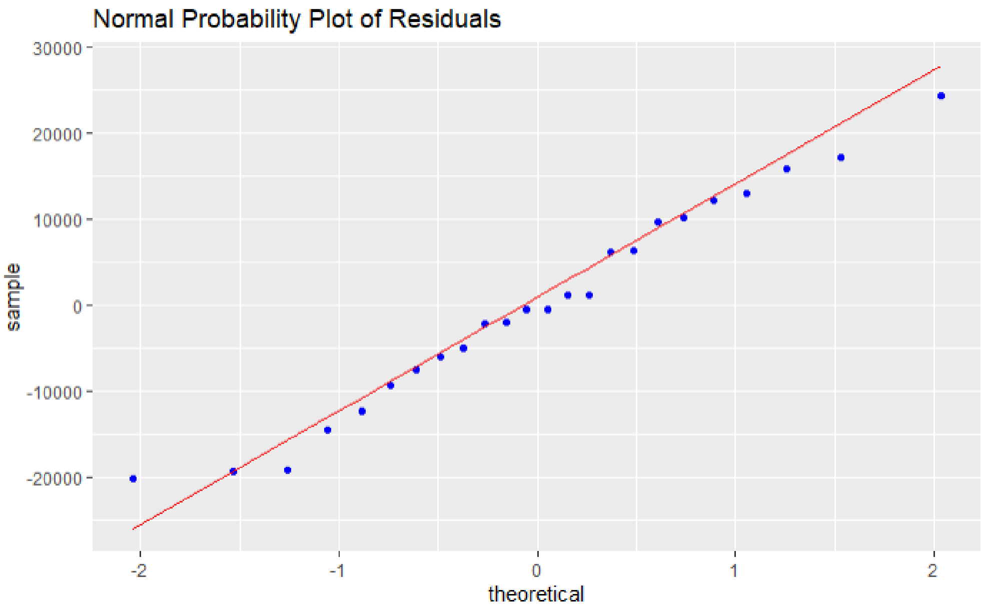
~~~
```{r}
diagnosticdf2
```
~~~

Description: df [24 × 2]

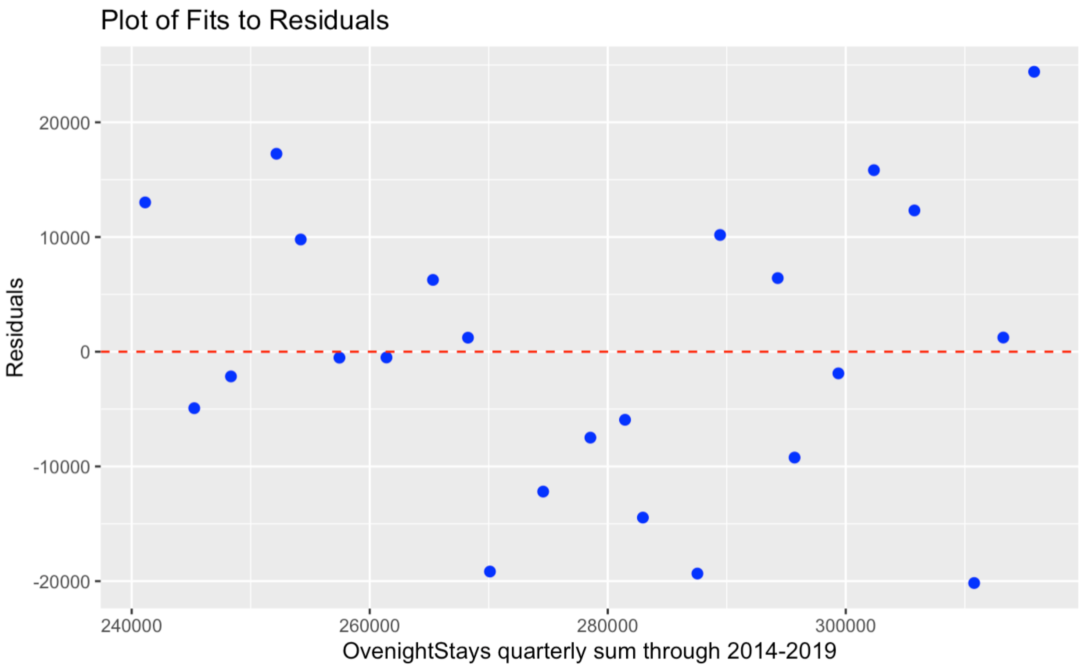
	predicted.values.overnight <dbl>	eison <dbl>
1	316576.5	24409.5367
2	313308.1	1240.9393
3	310039.7	-20162.6581
4	306771.3	12317.7445
5	303502.9	15827.1471
6	300234.5	-1885.4503
7	296966.0	-9229.0477
8	293697.6	6418.3549
9	290429.2	10180.7575
10	287160.8	-19342.8399

1-10 of 24 rows

Below we are visualizing the normality of our residuals



Based on our visual above we can infer that the residuals of our model show normality.

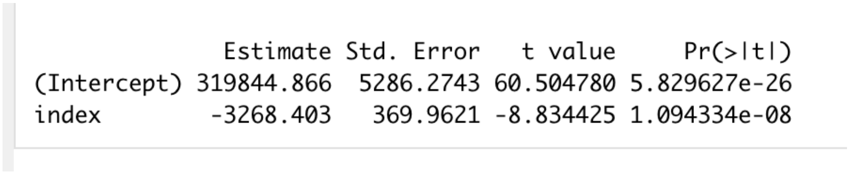


From the visual above we can see that our data is homoscedastic and checks both the conditions in order to perform our linear model.

For further reference we summed our standard of errors and got the value to be 0.00000000001000444, a relatively small number which is another good indication that our residuals are normal.

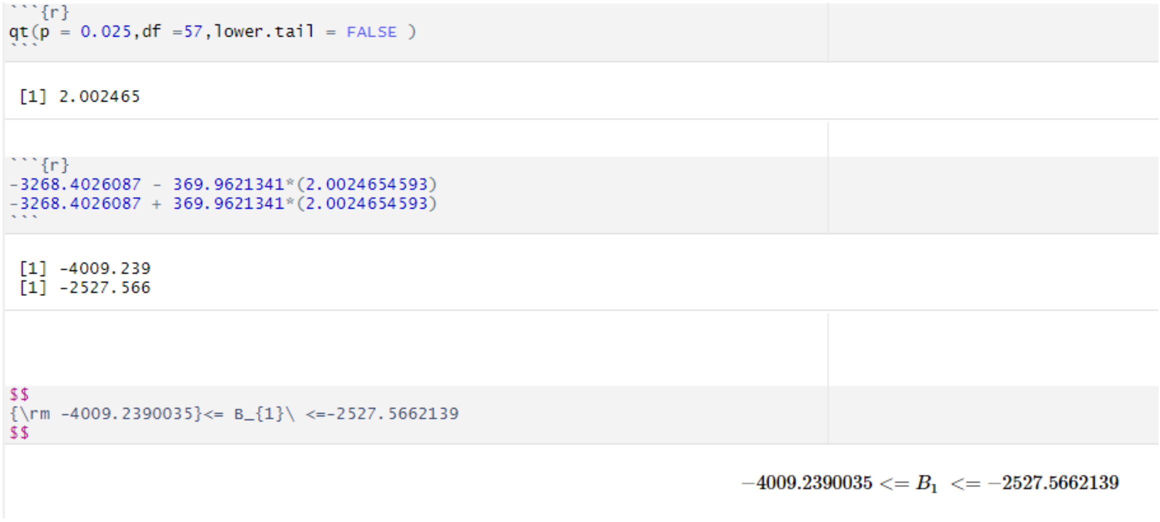
Hypothesis Development to check the negative linearity of our model is valid.

$$\mathrm{H}_0 : \beta_1 = (\leq)0 \qquad \mathrm{H}_A : \beta_1 < 0$$



```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 319844.866  5286.2743 60.504780 5.829627e-26
index        -3268.403   369.9621 -8.834425 1.094334e-08
```

And this yields a t-value of -8.83 and a subsequent P-value of 0.00000000547167. Due to the small size of this P-value, we can reject the null hypothesis as there is a 0.00000000547167 probability of observing stronger evidence against the null hypothesis. Therefore, we can conclude that there is a negative linear correlation between shelter occupancy and indexed months within the dataset. We can also find a 95% confidence interval on the value of β_1:

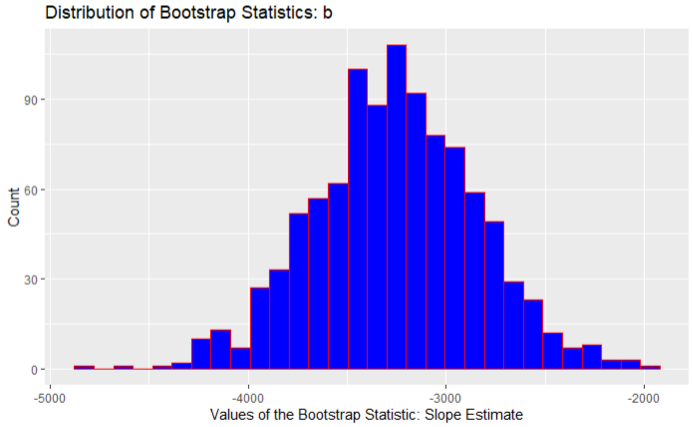
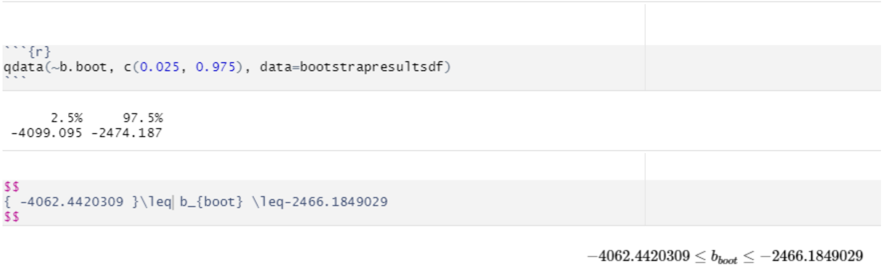


~~~
```{r}
qt(p = 0.025,df =57,lower.tail = FALSE )
~~~

 [1] 2.002465

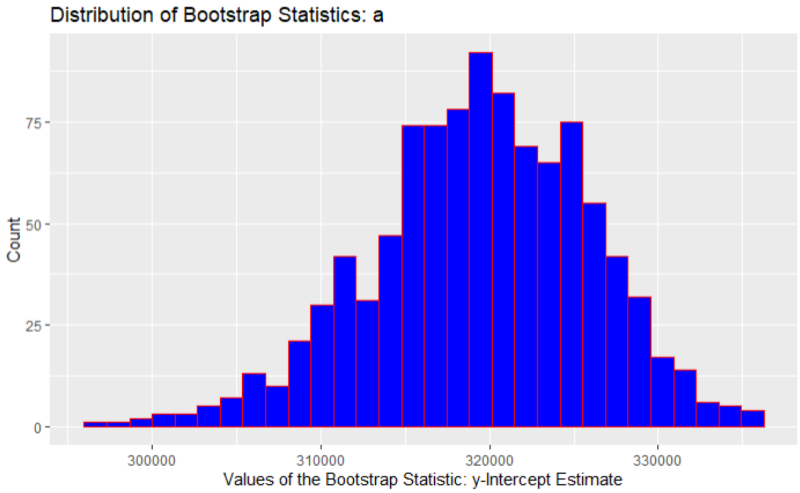
```{r}
-3268.4026087 - 369.9621341*(2.0024654593)
-3268.4026087 + 369.9621341*(2.0024654593)
```

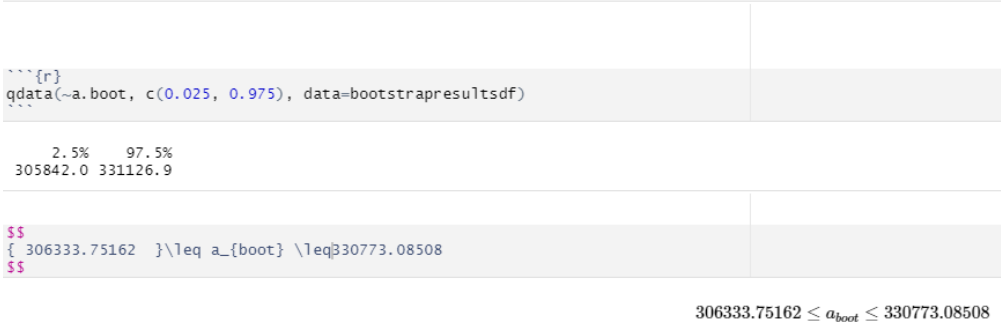
 [1] -4009.239
 [1] -2527.566

$$
{\rm -4009.2390035}<= B_{1}\ <=-2527.5662139
$$
```

$$-4009.2390035 <= B_1\ <= -2527.5662139$$

Bootstrapping the value of β_1 yields a confidence interval shown below:


Distribution of Bootstrap Statistics: b
90
60
30
0
Count
-5000
-4000
-3000
-2000
Values of the Bootstrap Statistic: Slope Estimate



```
```{r}
qdata(~b.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```

      2.5%   97.5%
 -4099.095 -2474.187

$$
{ -4062.4420309 }\leq b_{boot} \leq-2466.1849029
$$
```

$$-4062.4420309 \leq b_{boot} \leq -2466.1849029$$

And bootstrapping the value of A yields the confidence interval below:
```





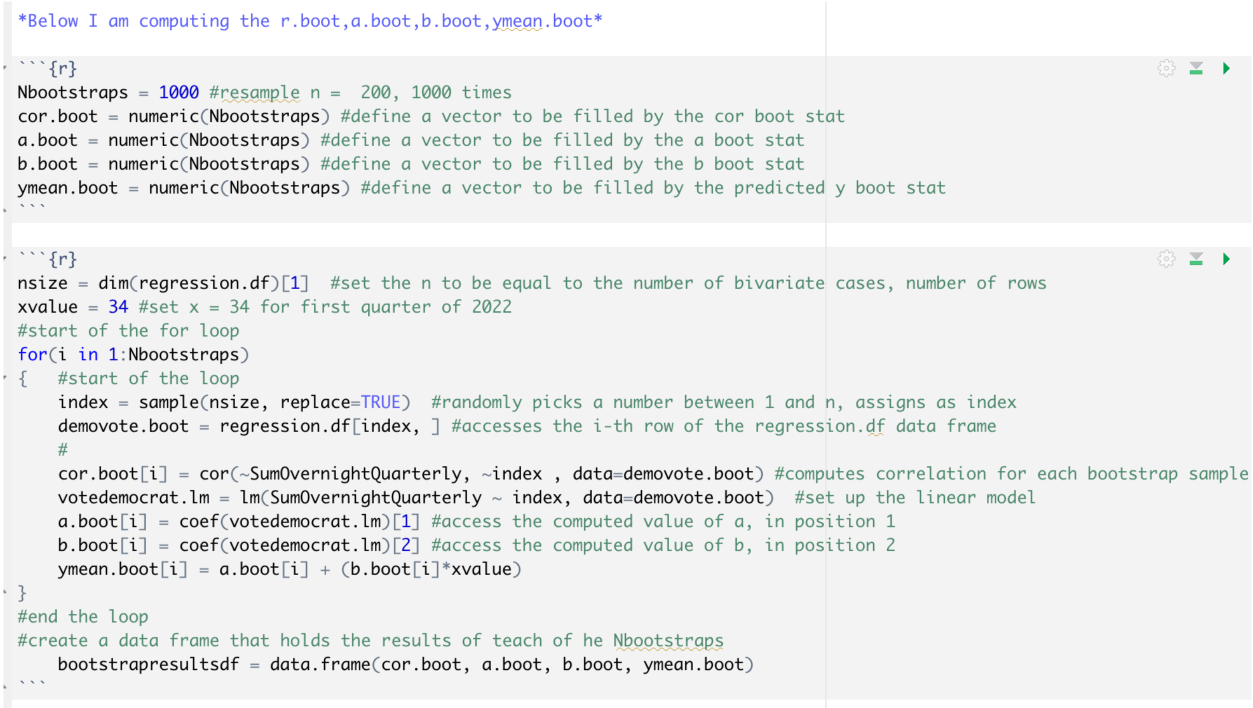
~~~
```{r}
qdata(~a.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```
~~~

```
    2.5%    97.5%
 305842.0 331126.9
```

```
$$
{ 306333.75162  }\leq a_{boot} \leq330773.08508
$$
```

$$306333.75162 \leq a_{boot} \leq 330773.08508$$

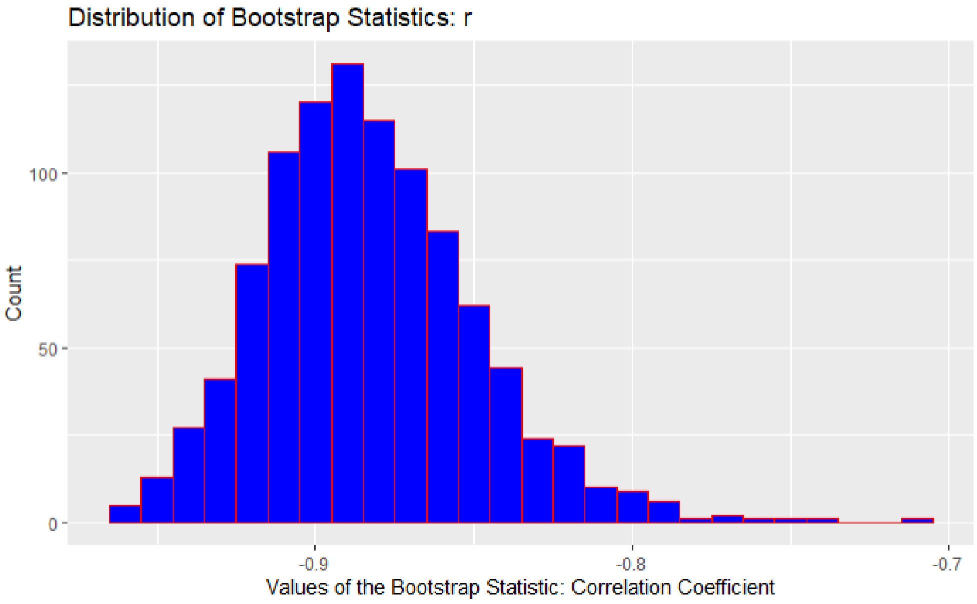
We can also bootstrap the r-value:

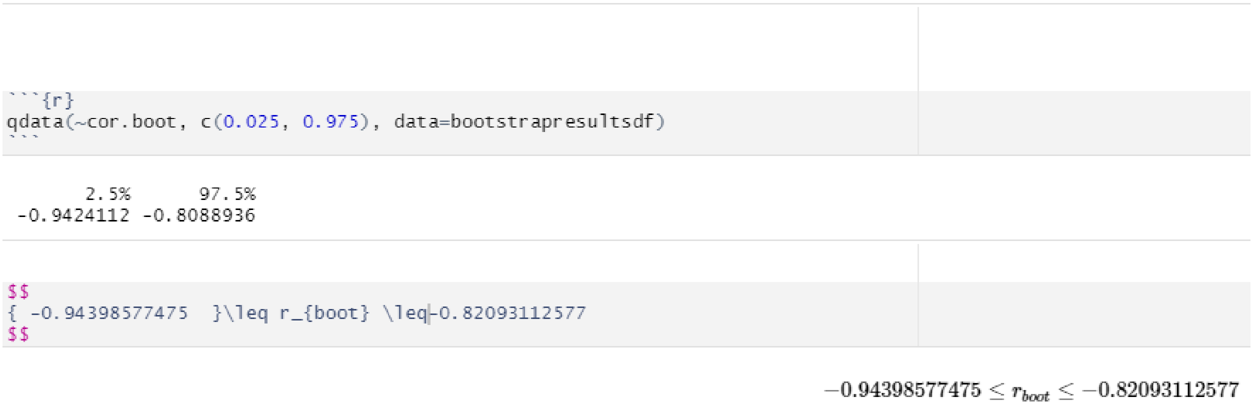


Below I am computing the r.boot,a.boot,b.boot,ymean.boot

~~~
```{r}
Nbootstraps = 1000 #resample n =  200, 1000 times
cor.boot = numeric(Nbootstraps) #define a vector to be filled by the cor boot stat
a.boot = numeric(Nbootstraps) #define a vector to be filled by the a boot stat
b.boot = numeric(Nbootstraps) #define a vector to be filled by the b boot stat
ymean.boot = numeric(Nbootstraps) #define a vector to be filled by the predicted y boot stat
```
~~~

~~~
```{r}
nsize = dim(regression.df)[1]  #set the n to be equal to the number of bivariate cases, number of rows
xvalue = 34 #set x = 34 for first quarter of 2022
#start of the for loop
for(i in 1:Nbootstraps)
{   #start of the loop
    index = sample(nsize, replace=TRUE)  #randomly picks a number between 1 and n, assigns as index
    demovote.boot = regression.df[index, ] #accesses the i-th row of the regression.df data frame
    #
    cor.boot[i] = cor(~SumOvernightQuarterly, ~index , data=demovote.boot) #computes correlation for each bootstrap sample
    votedemocrat.lm = lm(SumOvernightQuarterly ~ index, data=demovote.boot)  #set up the linear model
    a.boot[i] = coef(votedemocrat.lm)[1] #access the computed value of a, in position 1
    b.boot[i] = coef(votedemocrat.lm)[2] #access the computed value of b, in position 2
    ymean.boot[i] = a.boot[i] + (b.boot[i]*xvalue)
}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
    bootstrapresultsdf = data.frame(cor.boot, a.boot, b.boot, ymean.boot)
```
~~~





~~~
```{r}
qdata(~cor.boot, c(0.025, 0.975), data=bootstrapresultsdf)
```
~~~

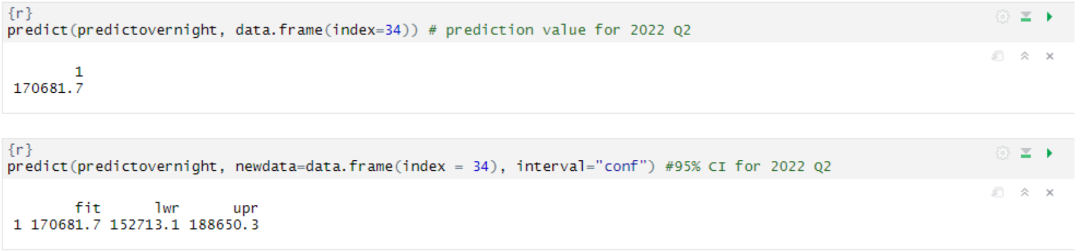
```
      2.5%      97.5%
 -0.9424112 -0.8088936
```

```
$$
{ -0.94398577475  }\leq r_{boot} \leq-0.82093112577
$$
```

$$-0.94398577475 \leq r_{boot} \leq -0.82093112577$$

Do Current Emergency Shelters Fit the Pre-2020 Linear Regression?

Building on our linear model we attempt to determine if emergency shelter occupancy levels have deviated from our pre-2020 negative linear regression model. We have chosen 2022 Q2 (X_{INDEX} = 34) where the sum emergency shelter occupancy was 194,373 to test our model. Referring to Figure 1, this seemed like a good quarter to choose, as there is a clear break in trend in 2020 Q1 that seems to return to previous trend around 2022 Q2. Should we derive a 95% confidence interval for that quarter that contains this occupancy rate, we should conclude that emergency shelter occupancy rates have returned to the pre-covid trend.



```
{r}
predict(predictovernight, data.frame(index=34)) # prediction value for 2022 Q2
```

```
       1
170681.7
```

```
{r}
predict(predictovernight, newdata=data.frame(index = 34), interval="conf") #95% CI for 2022 Q2
```

```
       fit      lwr      upr
1 170681.7 152713.1 188650.3
```

$$95\%\ CI:\ 152713.1 > X_{2022Q2} > \ 188650.3$$

The above results show that the model defines a Q4 2022 occupancy value approximately between 152,713.1 and 188,650.3, with 95% confidence. This does not contain our measured value of 194,373.

Therefore, we can conclude that in 2022 Q2, shelter occupancy deviates from our modeled trend from 2014-2019.

Topic of Investigation #2

Our data set contains a column that classifies each reporting shelter by shelter type. These shelter types include: Women Emergency, Intox, Adult Emergency, Winter Emergency, Youth Emergency, Short Term Supportive, Family Emergency, Long Term Supportive, COVID19 Expanded Shelter, COVID 19 Isolation Site, COVID19 Social Distancing Measures. It would be an interesting exploration of the data to understand the differences in proportions of these different sub-groups. It was a common topic of public discourse through the pandemic, that due to more people being confined indoors, rates of domestic violence were higher than in previous years (Sawhney). Our next investigation tests the hypothesis: Was the proportion of emergency shelter occupants that stayed in *Women Emergency* shelter types larger during the COVID19 pandemic than it was during the years following the economic downturn of 2014?

$$H_0 : p_{womenShelter2020} - p_{womenShelter2014} = 0 \qquad H_A : p_{womenShelter2020} - p_{womenShelter2014} > 0$$

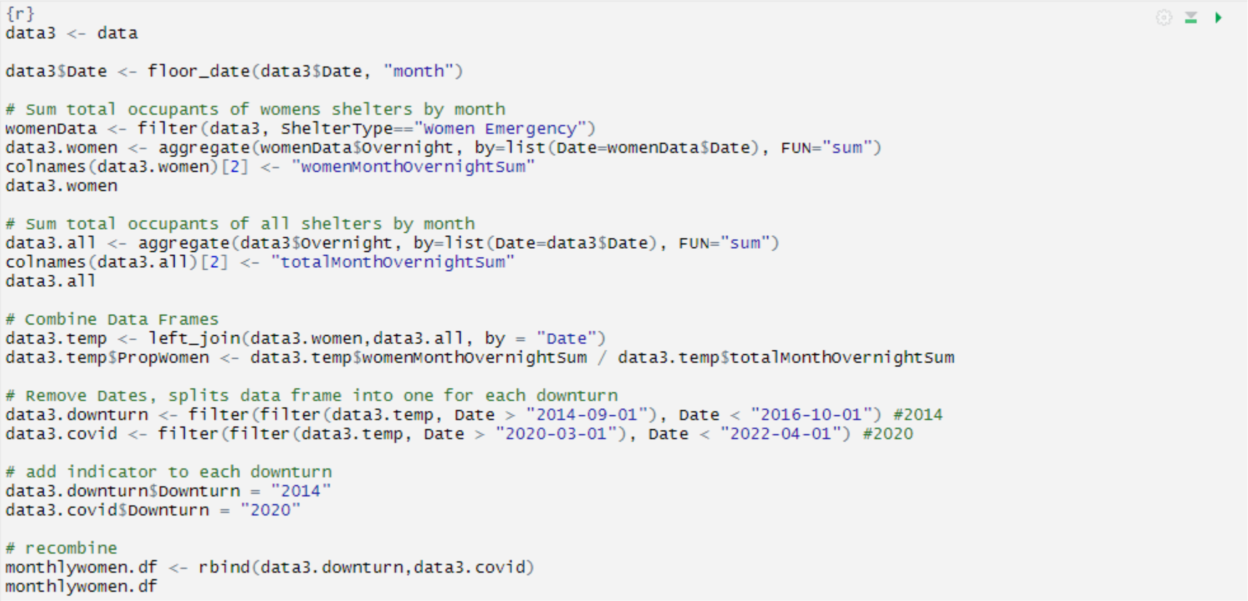
We can see from

Preparing Data

Starting with our original dataset we will derive a table containing the monthly summed totals of all *Women Emergency* shelter types, the monthly summed totals of occupants of all emergency shelters, a calculated field of the monthly proportion of occupants from *Women Emergency* shelter types, and an identity column for each downturn.

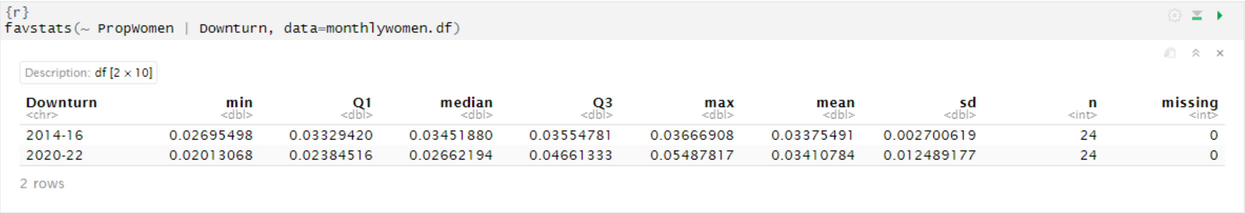
To prepare the data, first the total occupancy for *Women Emergency* shelter types are summed for each month with an aggregate function. The same is then done for all shelter types and these data frames are combined.

Our initial sample period for the 2014 downturn will begin October of 2014 and continue to September of 2016, inclusively. Our sample period for COVID19 will begin April of 2020 and continue to February of 2022, inclusively. So next, months outside of these are removed by splitting the data set into two data frames, coding each with its identity, and recombining them into our working data frame, *monthlywomen.df.* Seen below:



```r
{r}
data3 <- data

data3$Date <- floor_date(data3$Date, "month")

# Sum total occupants of womens shelters by month
womenData <- filter(data3, ShelterType=="Women Emergency")
data3.women <- aggregate(womenData$Overnight, by=list(Date=womenData$Date), FUN="sum")
colnames(data3.women)[2] <- "womenMonthOvernightSum"
data3.women

# Sum total occupants of all shelters by month
data3.all <- aggregate(data3$Overnight, by=list(Date=data3$Date), FUN="sum")
colnames(data3.all)[2] <- "totalMonthOvernightSum"
data3.all

# Combine Data Frames
data3.temp <- left_join(data3.women,data3.all, by = "Date")
data3.temp$PropWomen <- data3.temp$womenMonthOvernightSum / data3.temp$totalMonthOvernightSum

# Remove Dates, splits data frame into one for each downturn
data3.downturn <- filter(filter(data3.temp, Date > "2014-09-01"), Date < "2016-10-01") #2014
data3.covid <- filter(filter(data3.temp, Date > "2020-03-01"), Date < "2022-04-01") #2020

# add indicator to each downturn
data3.downturn$Downturn = "2014"
data3.covid$Downturn = "2020"

# recombine
monthlywomen.df <- rbind(data3.downturn,data3.covid)
monthlywomen.df
```

Stats on the proportion of total emergency shelter occupants that are from women emergency for each time-period:

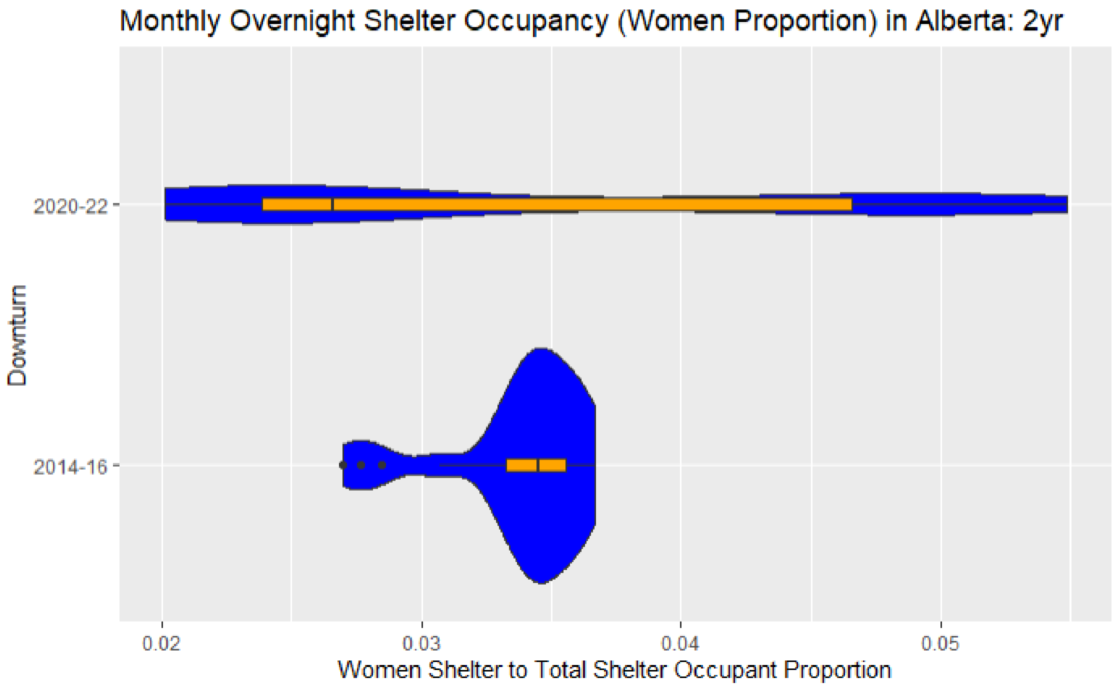


```r
{r}
favstats(~ PropWomen | Downturn, data=monthlywomen.df)
```

Description: df [2 x 10]

Downturn	min	Q1	median	Q3	max	mean	sd	n	missing
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
2014-16	0.02695498	0.03329420	0.03451880	0.03554781	0.03666908	0.03375491	0.002700619	24	0
2020-22	0.02013068	0.02384516	0.02662194	0.04661333	0.05487817	0.03410784	0.012489177	24	0

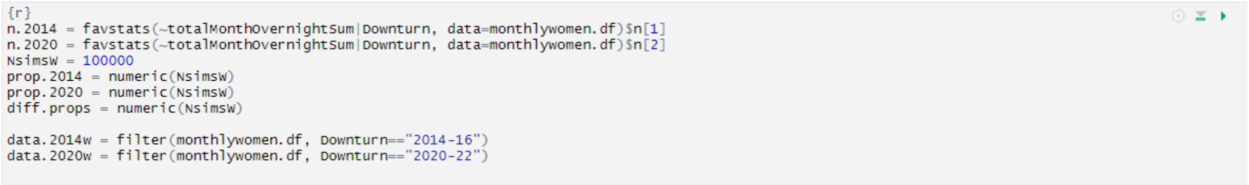
2 rows



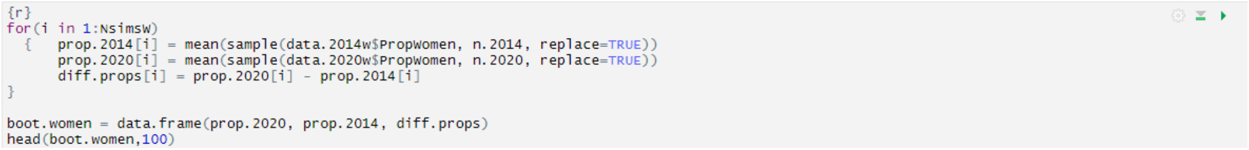
Looking at the boxplots for the two sample proportions above, we see a significantly larger IQR for 2020-22. Variance was significantly less in 2014-16. 2020-22 has a very distinct right-skew, and can be interpreted as not normally distributed. So to statistically analyze this data, we will require non-parametric methods to test our hypothesis.

24 Month Bootstrap Confidence Interval

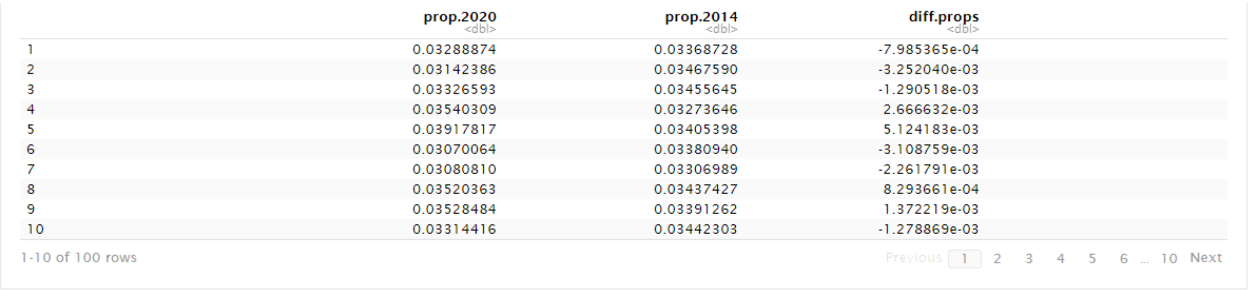
Here we derive our bootstrap distribution for $p_{womanShelter2020}$ - $p_{womanShelter2014}$. We have chosen to run 100,000 iterations. This is likely unnecessarily high, but we see no harm.



```
{r}
n.2014 = favstats(~totalMonthOvernightSum|Downturn, data=monthlywomen.df)$n[1]
n.2020 = favstats(~totalMonthOvernightSum|Downturn, data=monthlywomen.df)$n[2]
NsimsW = 100000
prop.2014 = numeric(NsimsW)
prop.2020 = numeric(NsimsW)
diff.props = numeric(NsimsW)

data.2014w = filter(monthlywomen.df, Downturn=="2014-16")
data.2020w = filter(monthlywomen.df, Downturn=="2020-22")
```



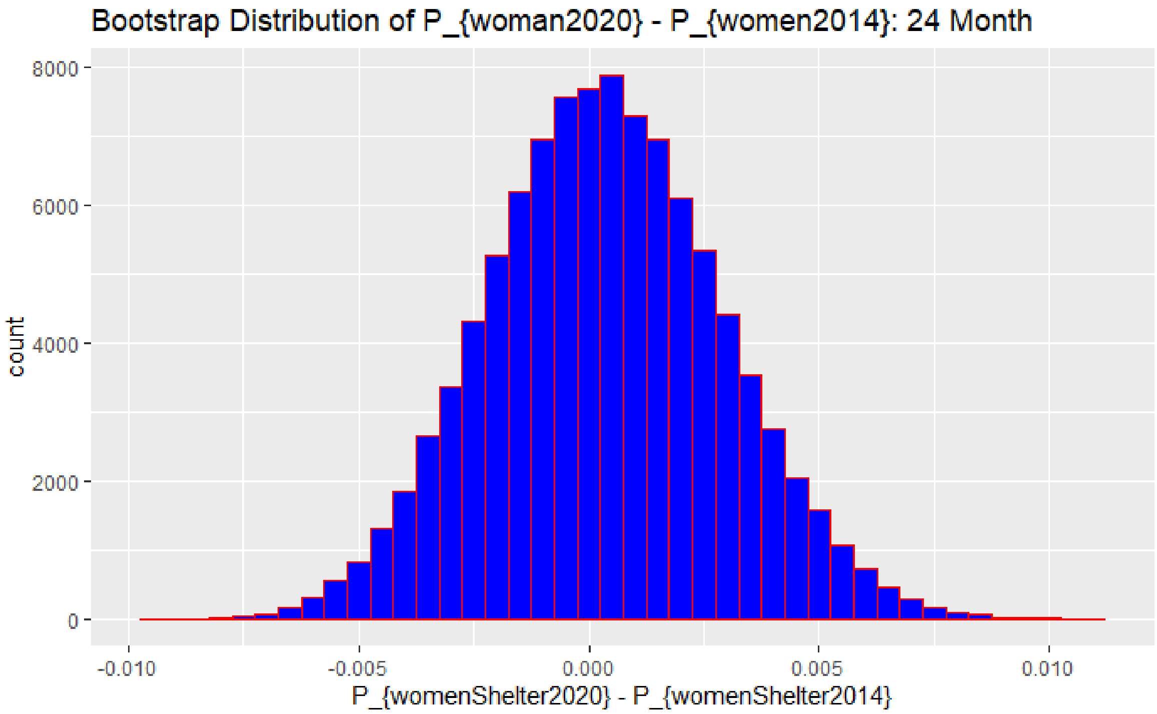
```
{r}
for(i in 1:NsimsW)
  {   prop.2014[i] = mean(sample(data.2014w$PropWomen, n.2014, replace=TRUE))
      prop.2020[i] = mean(sample(data.2020w$PropWomen, n.2020, replace=TRUE))
      diff.props[i] = prop.2020[i] - prop.2014[i]
}

boot.women = data.frame(prop.2020, prop.2014, diff.props)
head(boot.women,100)
```



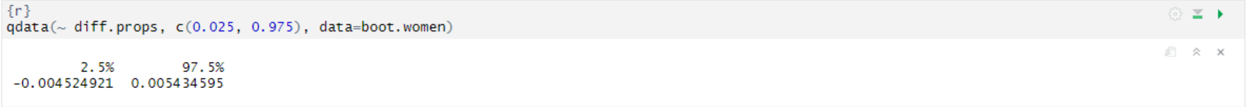
	prop.2020 <dbl>	prop.2014 <dbl>	diff.props <dbl>
1	0.03288874	0.03368728	-7.985365e-04
2	0.03142386	0.03467590	-3.252040e-03
3	0.03326593	0.03455645	-1.290518e-03
4	0.03540309	0.03273646	2.666632e-03
5	0.03917817	0.03405398	5.124183e-03
6	0.03070064	0.03380940	-3.108759e-03
7	0.03080810	0.03306989	-2.261791e-03
8	0.03520363	0.03437427	8.293661e-04
9	0.03528484	0.03391262	1.372219e-03
10	0.03314416	0.03442303	-1.278869e-03

1-10 of 100 rows

With our bootstrap distribution calculated, we can visualize it with ggplot:



Above, we see that our bootstrap distribution clearly straddles zero, implying no difference in proportions. Below, we calculate our 95% confidence interval:



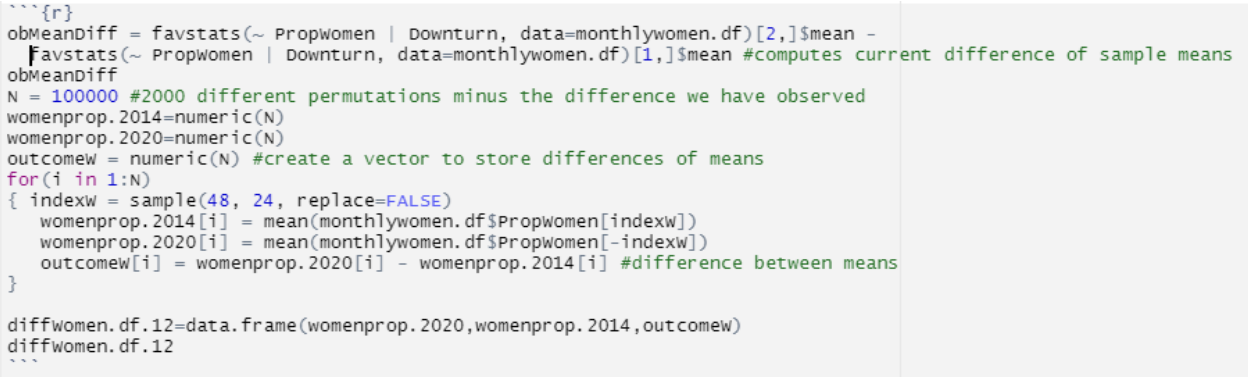
```
{r}
qdata(~ diff.props, c(0.025, 0.975), data=boot.women)

        2.5%        97.5%
-0.004524921  0.005434595
```

$$95\%CI : -0.00452 < p_{womanShelter2020} - p_{womanShelter2014} < 0.0055$$

Since we see that the 95% confidence interval consists of 0 in the interval we can interpret that there is no difference between the proportion of women shelters in 2020 to women shelters in 2014 in the province of Alberta.

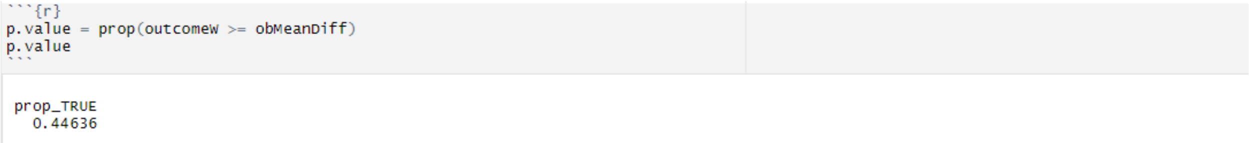
24 Month Permutation Test To Investigate The Women Proportion in Shelters

As this is not the result we expected, we hope to better understand the difference of proportion with a permutation test. Below we conduct a permutation test with 100,000 iterations, this should

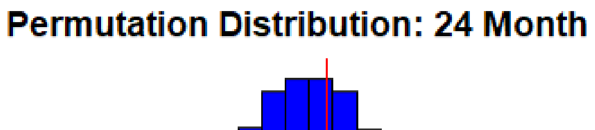


~~~
```{r}
obMeanDiff = favstats(~ PropWomen | Downturn, data=monthlywomen.df)[2,]$mean -
  favstats(~ PropWomen | Downturn, data=monthlywomen.df)[1,]$mean #computes current difference of sample means
obMeanDiff
N = 100000 #2000 different permutations minus the difference we have observed
womenprop.2014=numeric(N)
womenprop.2020=numeric(N)
outcomeW = numeric(N) #create a vector to store differences of means
for(i in 1:N)
{ indexW = sample(48, 24, replace=FALSE)
   womenprop.2014[i] = mean(monthlywomen.df$PropWomen[indexW])
   womenprop.2020[i] = mean(monthlywomen.df$PropWomen[-indexW])
   outcomeW[i] = womenprop.2020[i] - womenprop.2014[i] #difference between means
}

diffWomen.df.12=data.frame(womenprop.2020,womenprop.2014,outcomeW)
diffWomen.df.12
```
~~~

provide very consistent output and precise computed values.



~~~
```{r}
p.value = prop(outcomeW >= obMeanDiff)
p.value
```

prop_TRUE
  0.44636
~~~



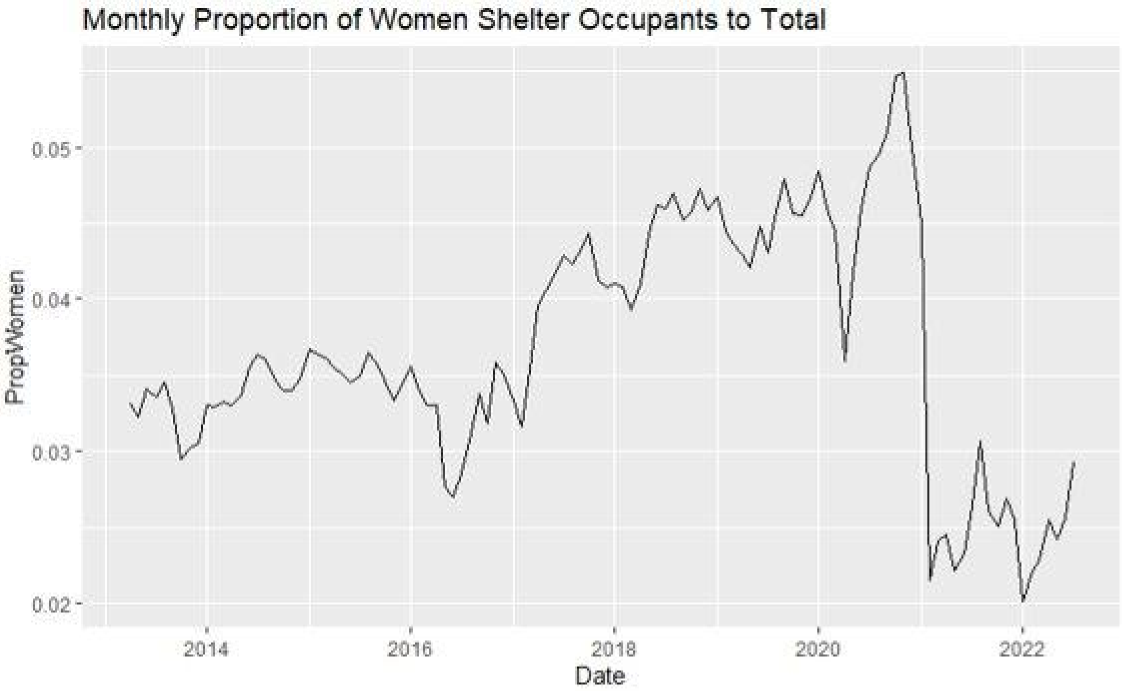
Looking at our permutation distribution and observed difference (red line), it's obvious that there is no statistical difference here. With a p-value of 0.4463, our observed difference is close to the expected value of the distribution. This evidence again supports our conclusion to fail to reject the null hypothesis.

24 Month Conclusion

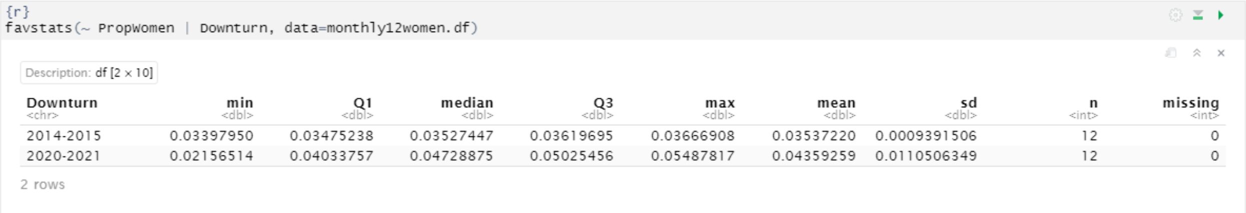
In conclusion both non-parametric tests suggest with 95% confidence that the Women Shelter proportion of total emergency shelter occupants was not greater during the first 24 months of the 2020 COVID-19 pandemic and downturn than it was in the 2014 downturn.

Taking a closer look at our data

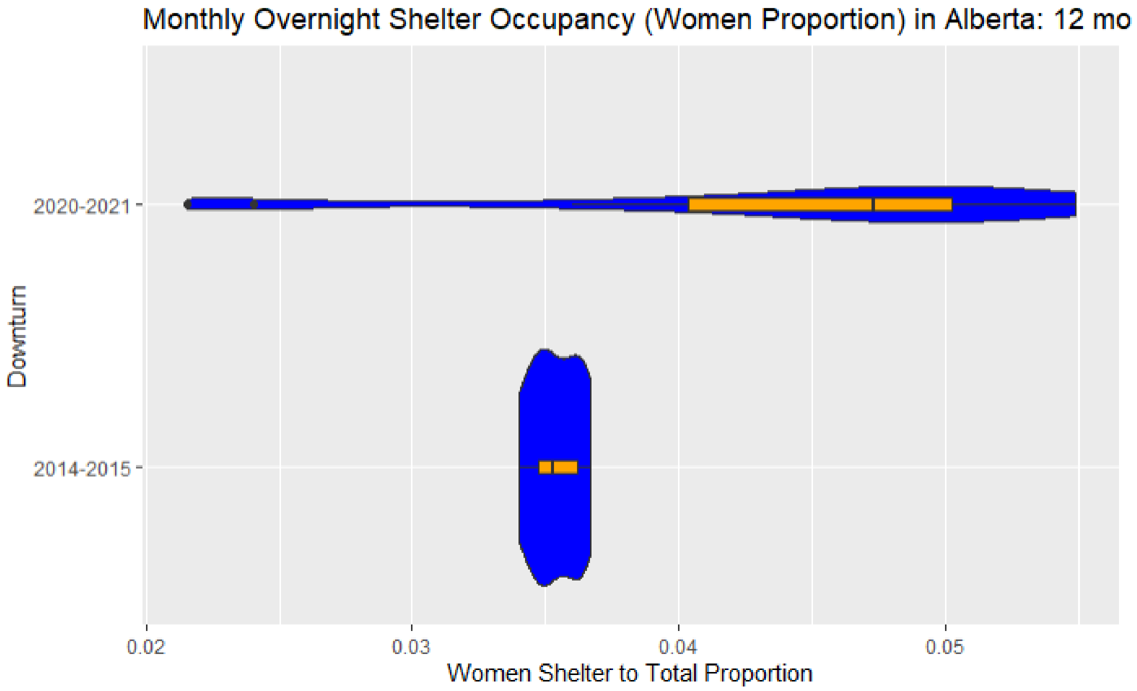
Looking at the plot below, showing the monthly proportion of women's shelter occupants over time, we see a substantiation increase in this proportion during the first year of the pandemic.


Monthly Proportion of Women Shelter Occupants to Total

If we change our two sample sets to 12 month periods after each oil crash, would we see a different outcome? The new data frames are October 2014 to September 2015 and April 2020 to March 2022. The hypothesis stays the same, let's investigate:

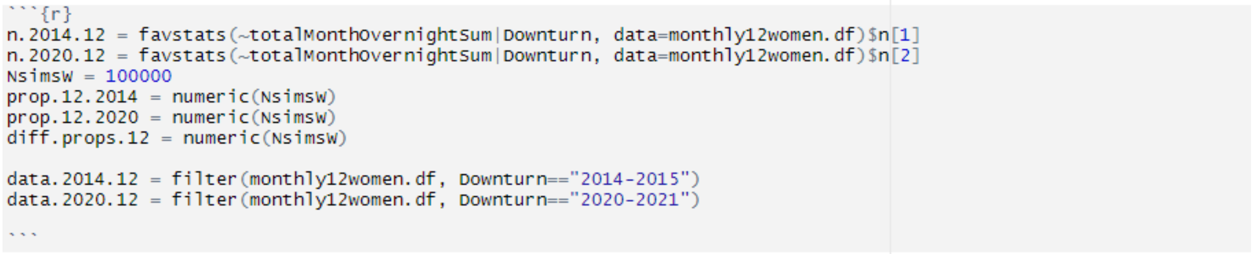

```
{r}
favstats(~ PropWomen | Downturn, data=monthly12women.df)
```

```
Description: df [2 × 10]

Downturn         min         Q1     median         Q3        max       mean           sd    n  missing
<chr>          <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>        <dbl> <int>   <int>
2014-2015 0.03397950 0.03475238 0.03527447 0.03619695 0.03666908 0.03537220 0.0009391506   12        0
2020-2021 0.02156514 0.04033757 0.04728875 0.05025456 0.05487817 0.04359259 0.0110506349   12        0

2 rows
```

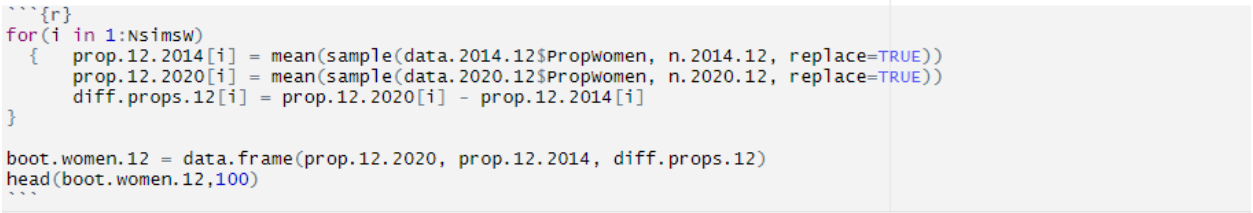


In these shortened time periods, we see a significant reduction in the IQR from 2020-21. As well, note that the median value for 2020-21 is now greater than the median of 2014-15. 2020-21 has also swapped to a left-skewed distribution. So again–as we must use non-parametric analysis–we will conduct a bootstrap confidence interval and permutation test.

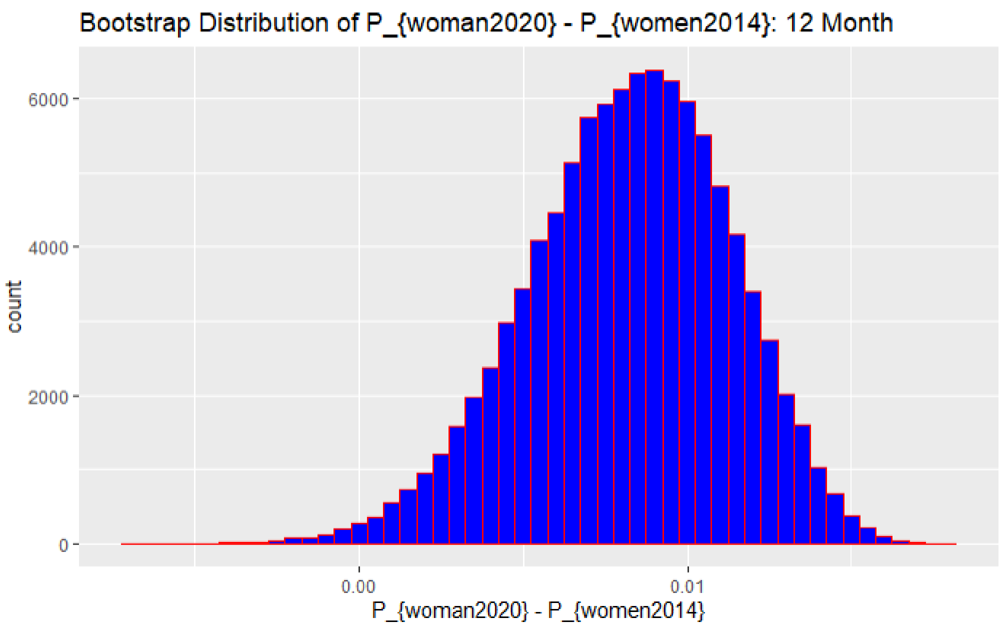
12 Month Bootstrap Confidence Interval



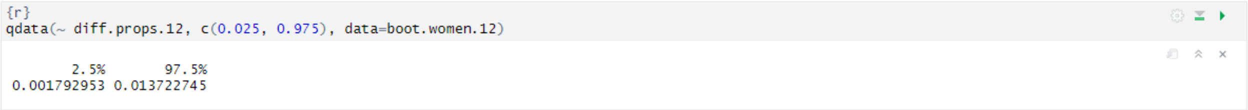
~~~
```{r}
n.2014.12 = favstats(~totalMonthOvernightSum|Downturn, data=monthly12women.df)$n[1]
n.2020.12 = favstats(~totalMonthOvernightSum|Downturn, data=monthly12women.df)$n[2]
NsimsW = 100000
prop.12.2014 = numeric(NsimsW)
prop.12.2020 = numeric(NsimsW)
diff.props.12 = numeric(NsimsW)

data.2014.12 = filter(monthly12women.df, Downturn=="2014-2015")
data.2020.12 = filter(monthly12women.df, Downturn=="2020-2021")

```
~~~



~~~
```{r}
for(i in 1:NsimsW)
  {   prop.12.2014[i] = mean(sample(data.2014.12$PropWomen, n.2014.12, replace=TRUE))
      prop.12.2020[i] = mean(sample(data.2020.12$PropWomen, n.2020.12, replace=TRUE))
      diff.props.12[i] = prop.12.2020[i] - prop.12.2014[i]
}

boot.women.12 = data.frame(prop.12.2020, prop.12.2014, diff.props.12)
head(boot.women.12,100)
```
~~~



Above, the bootstrap distribution this time does cross zero, but only on its left tail. Below, we calculate our 95% confidence interval:

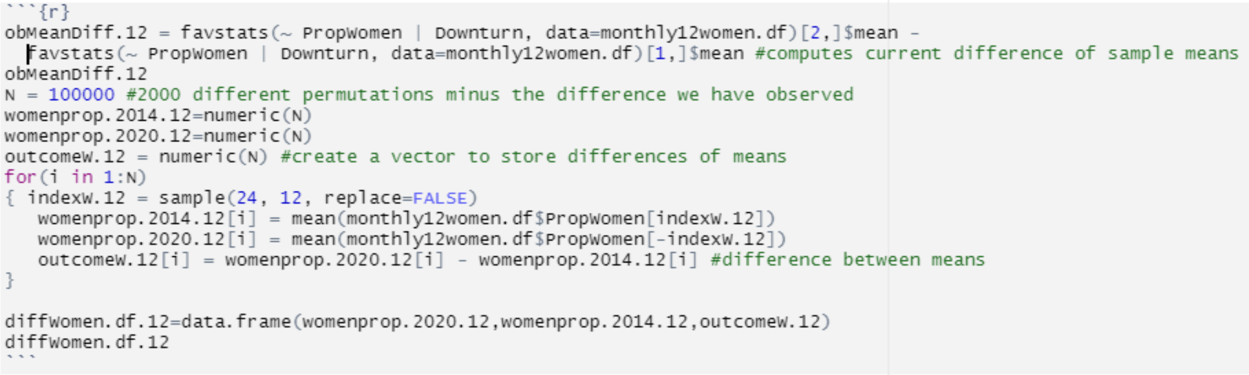


```
{r}
qdata(~ diff.props.12, c(0.025, 0.975), data=boot.women.12)

        2.5%        97.5%
 0.001792953 0.013722745
```

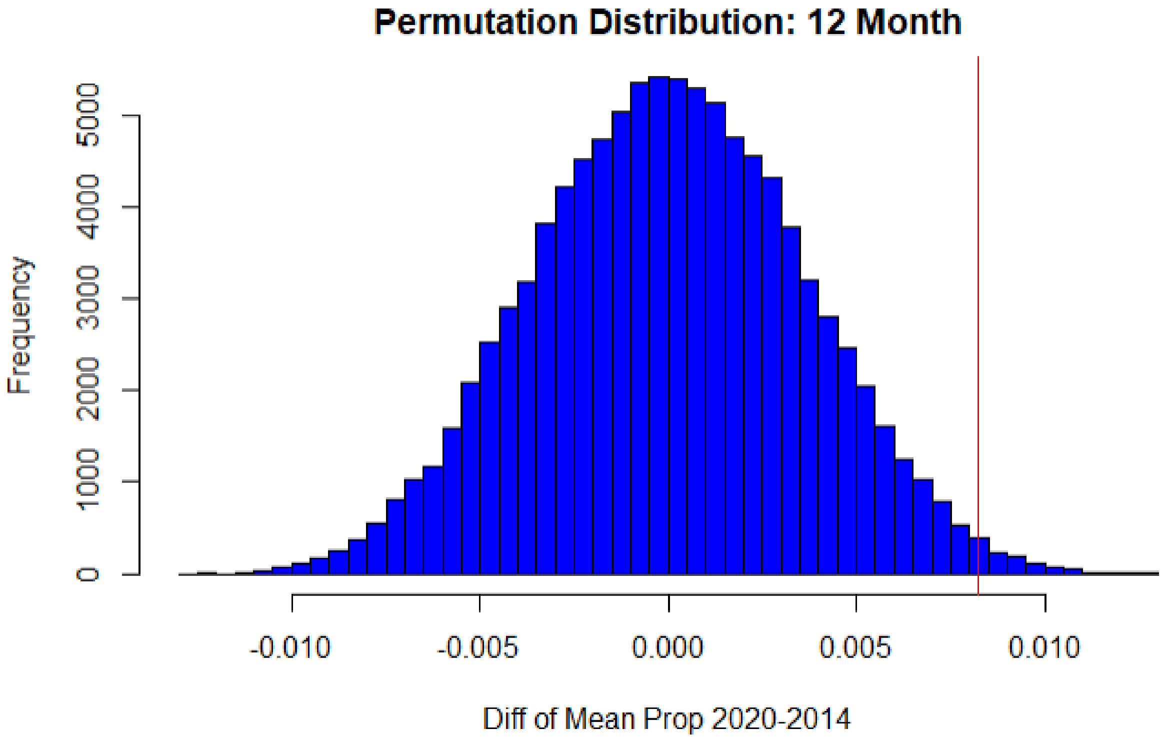
Our 95% bootstrap confidence interval for difference of proportion is entirely positive. This evidence supports the rejection of the null hypothesis when our time sample is reduced to 12 months.

12 Month Permutation Test

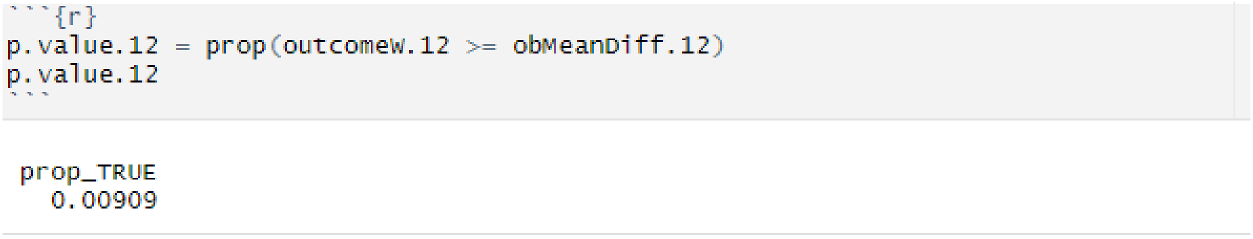
Below we conduct out 12 month permutation test to confirm our rejection of the null hypothesis:



~~~
```{r}
obMeanDiff.12 = favstats(~ PropWomen | Downturn, data=monthly12women.df)[2,]$mean -
  favstats(~ PropWomen | Downturn, data=monthly12women.df)[1,]$mean #computes current difference of sample means
obMeanDiff.12
N = 100000 #2000 different permutations minus the difference we have observed
womenprop.2014.12=numeric(N)
womenprop.2020.12=numeric(N)
outcomeW.12 = numeric(N) #create a vector to store differences of means
for(i in 1:N)
{ indexW.12 = sample(24, 12, replace=FALSE)
   womenprop.2014.12[i] = mean(monthly12women.df$PropWomen[indexW.12])
   womenprop.2020.12[i] = mean(monthly12women.df$PropWomen[-indexW.12])
   outcomeW.12[i] = womenprop.2020.12[i] - womenprop.2014.12[i] #difference between means
}

diffWomen.df.12=data.frame(womenprop.2020.12,womenprop.2014.12,outcomeW.12)
diffWomen.df.12
```
~~~



Above we see our permutation distribution is much different than in our 24 month analysis. Our observed difference of proportion means is far to the right tail. Below we calculate our p-value:



~~~
```{r}
p.value.12 = prop(outcomeW.12 >= obMeanDiff.12)
p.value.12
```
~~~

```
 prop_TRUE
   0.00909
```

With a p-value of 0.0091 we confirm our bootstrap confidence interval outcome to reject the null hypothesis.

12 Month Conclusions

Both of our non-parametric tests agree with 95% confidence that when we reduce our time sample to 12 months following each downturn we do have statistical evidence to reject the null hypothesis. We conclude that for the 12 months following each downturn, the 2020 COVID-19 pandemic did have a larger proportion of Women Shelter occupants than the 2014 downturn.

Conclusions

Which economic downturn, brought on by the crash in oil prices and/or COVID-19 pandemic had a larger effect on homeless shelter occupancy?

Our analysis showed that comparing two different 24 month periods after the two oil downturns showed that the 2014-2016 period yielded much higher total shelter occupancy in Alberta. However, after a closer look at the data, the COVID-19 homelessness data did not deviate from the linear trend that had been followed by the overall shelter occupancy from 2014-2019. From performing our linear regression we found there to be a strong negative linear correlation with an r value of -0.88323. We then estimated our linear model to the Q2 2022 (Index=34) . We found our estimated overnight stay to be 208719, which falls in your 95% bootstrap confidence interval for our estimation of Q2 2022 which was approximately between 191389 to 226049. This indicates that the COVID-19 pandemic and associated crash in oil prices did not have a statistically significant effect on homeless shelter occupancy. In addition, the data also shows a negative trend in homeless shelter occupancy after the 2014 oil crash. (this was the linear model).

Further investigation is required to investigate why, aside from the COVID-19 'disruption', the total shelter occupancy has decreased steadily through two major oil crashes. One potential study would be to perform a similar analysis on other Canadian provinces, with less economic dependence on the oil industry.

Which economic downturn resulted in a higher women's-only shelter occupancy, as a proportion of the total monthly occupancy?

Our analysis showed that with 24-month sample periods after each oil crash (2014-2016 and 2020-2022), there was no statistical difference between the proportions of womens-only shelter occupancy ('women's proportion') between the two periods. However, a visualization of this proportion over time indicated that there was a substantial increase in this 'women's proportion' in the first year of the pandemic, followed by a sharp drop. Thus, shortening the two sample periods to 12 months resulted in the 2020-2021 data having a larger 'women's proportion'. This can potentially be explained by the fact that the majority of the quarantining and business closures were in the first year of the pandemic, resulting in the most time spent confined at home.

Further investigation could look at occupancies in other shelter types as indicators of the 'Shadow Pandemic'. These shelter types include youth shelters and to a lesser extent, family shelters.

References

“11.5 Symmetric and skewed data | Statistics.” *Siyavula*, https://ng.siyavula.com/read/maths/grade-11/statistics/11-statistics-05. Accessed 18 October 2022.

Government of Alberta, 2019. “Emergency Shelters Daily Occupancy AB - Emergency Shelters Daily Occupancy AB - 2013-22.” *Open Government Program*, 8 February 2019, https://open.alberta.ca/dataset/funded-emergency-shelters-daily-occupancy-ab/resource/b7080b66-25ea-4c30-ac47-02b64353637f. Accessed 18 October 2022.

Government of Alberta, 2022. “Oil Prices.” *Alberta Economic Dashboard*, http://economicdashboard.alberta.ca/oilprice. Accessed 18 October 2022.

Makowichuk, Darren. “Two years of COVID-19: A timeline of the pandemic in Alberta.” *Calgary Herald*, 15 March 2022, https://calgaryherald.com/news/local-news/two-years-of-covid-19-a-timeline-of-the-pandemic-in-alberta. Accessed 18 October 2022.

Sawhney, Rajan. “The 'shadow' pandemic | Alberta.ca.” *Government of Alberta*, 20 April 2021, https://www.alberta.ca/article-the-shadow-pandemic.aspx. Accessed 17 October 2022.

Appendix:

.Rmd file appended below: