Data

Each week the Consumer Financial Protection Bureau (CFPB) sends thousands of consumers' complaints about financial products and services to companies for response. Those complaints are published here CFPB after the company responds or after 15 days. Fields include:

date_received, product, sub-product, issue, sub-issue, Consumer complaint narrative, Company public response, company, state, zip code, Consumer consent provided?, submitted via, Date sent to company, Company response to consumer, Timely response?, Consumer disputed?, Complaint ID

```
'data.frame':   1048575 obs. of  18 variables:
 $ date_received               : Factor w/ 3329 levels "1/1/2012","1/1/2014",..: 3215 3163 837 1476 1476 1476 1476 1476 1476 1456 ...
 $ product                     : Factor w/ 18 levels "Bank account or service",..: 8 7 8 7 7 7 7 7 7 8 ...
 $ Sub.product                 : Factor w/ 77 levels "","(CD) Certificate of deposit",..: 35 17 35 17 17 17 17 17 17 42 ...
 $ Issue                       : Factor w/ 163 levels "Account opening, closing, or management",..: 14 76 36 76 76 76 113 113 113 145 ...
 $ Sub.issue                   : Factor w/ 219 levels "","Account information incorrect",..: 59 96 84 96 96 96 192 216 192 194 ...
 $ Consumer.complaint.narrative: Factor w/ 381781 levels "","''' We placed {$1300.00} in suspense on XX/XX/XXXX, because we needed to verify two items
deposited into your ac"| __truncated__,..: 338251 1 276817 1 1 1 1 1 1 1 ...
 $ Company.public.response     : Factor w/ 12 levels "","Company believes complaint caused principally by actions of third party outside the control
or direction of the company",..: 1 12 1 1 1 1 1 1 1 1 ...
 $ Company                     : Factor w/ 4973 levels "(Former)Shapiro, Swertfeger & Hasty, LLP",..: 4598 1692 1491 4597 4597 4597 4597 4597 4597 28
58 ...
 $ State                       : Factor w/ 63 levels "","AA","AE","AK",..: 15 48 37 15 15 15 11 53 54 37 ...
 $ ZIP.code                    : Factor w/ 19402 levels "","007XX","009XX",..: 6081 1314 3769 5818 5740 5855 15059 7231 14080 3981 ...
 $ Tags                        : Factor w/ 4 levels "","Older American",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Consumer.consent.provided.  : Factor w/ 6 levels "","Consent not provided",..: 3 2 3 1 1 1 1 1 1 1 ...
 $ Submitted.via               : Factor w/ 6 levels "Email","Fax",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ Date.sent.to.company        : Factor w/ 3210 levels "1/1/2014","1/1/2015",..: 3103 3069 808 1419 1419 1419 1419 1419 1419 1399 ...
 $ Company.response.to.consumer: Factor w/ 9 levels "","Closed","Closed with explanation",..: 3 5 3 8 8 8 8 8 8 8 ...
 $ Timely.response.            : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Consumer.disputed.          : Factor w/ 3 levels "N/A","No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Complaint.ID                : int  3384392 3379500 3433198 4211591 4211597 4210257 4211588 4211594 4211569 4205355 ...
```

```
                                        Company           State          ZIP.code                                  Tags
TRANSUNION INTERMEDIATE HOLDINGS, INC.:169324   CA      :128676         :125529                                   :924335
EQUIFAX, INC.                         :169062   FL      :124546   33Oxx :  11811   Older American                 : 34349
Experian Information Solutions Inc.   :154843   TX      :105431   33Lxx :  10353   Older American, Servicemember: 10845
BANK OF AMERICA, NATIONAL ASSOCIATION : 28764   NY      : 70421   300xx :   9048   Servicemember                  : 79046
JPMORGAN CHASE & CO.                  : 28006   GA      : 63362   770xx :   8002
CAPITAL ONE FINANCIAL CORPORATION     : 26216   PA      : 42771   606xx :   6415
(Other)                               :472360   (Other):513368   (Other):877417
       Consumer.consent.provided.      Submitted.via   Date.sent.to.company                 Company.response.to.consumer
                     : 41888          Email     :   74   9/8/2017  :  2834   Closed with explanation     :888438
Consent not provided:382273          Fax       : 10597   1/5/2021  :  2328   Closed with non-monetary relief:100526
Consent provided    :431248          Phone     : 50901   9/9/2017  :  2241   Closed with monetary relief    : 37568
Consent withdrawn   :  2302          Postal mail: 24403   12/29/2020:  2166   In progress                    : 14554
N/A                 :153168          Referral  : 60845   12/15/2020:  2141   Untimely response              :  5579
Other               : 37696          Web       :901755   12/17/2020:  2129   Closed                         :   996
                                                         (Other)   :1034736   (Other)                        :   914
Timely.response. Consumer.disputed.  Complaint.ID
No :  18008      N/A:1007577         Min.   :   110
Yes:1030567      No :  35059         1st Qu.:3120924
                 Yes:   5939         Median :3487602
                                     Mean   :3383367
                                     3rd Qu.:3822580
                                     Max.   :4213178
```

Planning

The data pulled captures the consumer complaints between 2011 and 2021 for various financial products such as student loans, mortgages etc., however for this post we will single out a product of interest that we will use for analysis. To start with, let's narrow our data to only 2 variables: **'date_received' and 'product'**

From the str () function output, we saw that eate_recieved was a chr (character) instead of date so we need transform this variable to its appropriate type. As seen most consumers complained regarding mortgages so we will subset our data to focus on this product. We will use the lubridate package for this. We probably would like to perform time series on yearly data () so we will extract Year, Month & Day and then later take monthly sums of the complaints across all the years using dplyr. Finally, transforming the data to a timeseries so we can begin the time series analysis. We will slice the data such that it starts from January 2011 and ends in December 2021 and then perform some exploratory analysis.

Analysis

2011 starts off at a very low dip that steadily spikes until mid-year then shows a series of ups and downs, this pattern is repeated through out the time span. The data seems to have a weak seasonality & without a clear trend. Wondering what might have caused the sharp spike in early 2016 its clear that most consumers filed more complaints compared to the other years. Disassembling the data shows seasonality a bit more clear but still no clear trend. Finally, the seasonal plot shows the seasonal patterns a bit more clearly and makes it easy to spot that 2011 began with a low minimum and early 2016 had a sharp spike and in 2018.

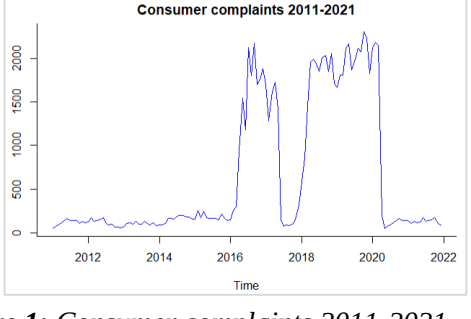

Figure 1: *Consumer complaints 2011-2021*

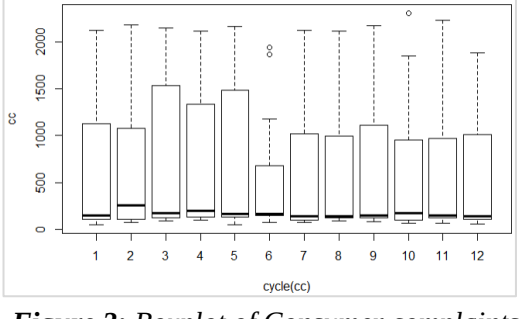

Figure 2: *Boxplot of Consumer complaints 2011-2021*

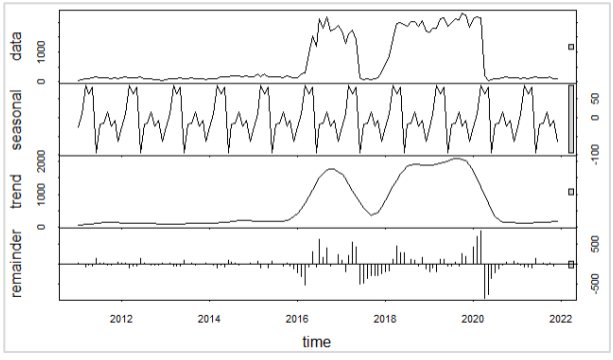

Figure 3: *Periodic plot over time-series*

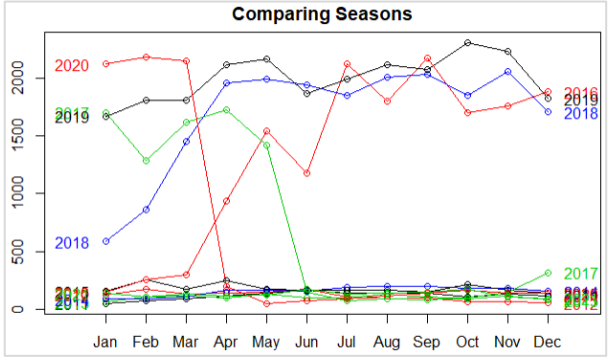

Figure 4: *Comparing seasons with respect to (months)*

Test for Stationarity: One of the common requirements in many time series techniques is that data must be stationary, meaning that the mean, variance and autocorrelation do not change over time. The test outputs a p-value greater than 0.05 therefore the data are not stationary. data: cc Dickey-Fuller = -2.7414, Lag order = 5, p-value = 0.2682 | alternative hypothesis: stationary.

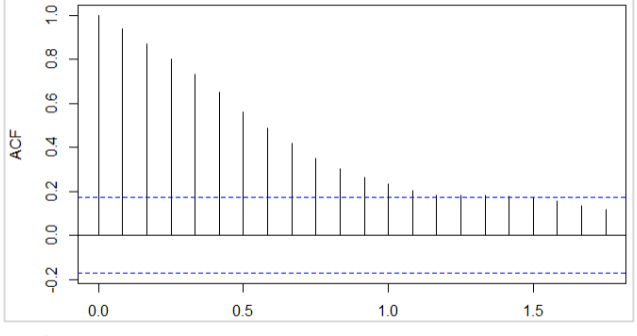

Figure 5: *ACF Series cc*

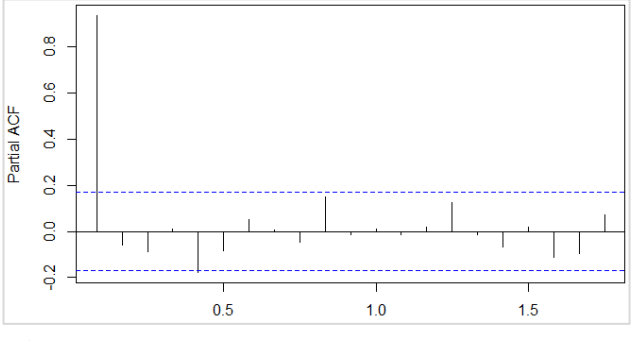

Figure 6: *PACF series cc*

The analysis above shows a lag correlation of 2 for the ACF test and a lag correlation of 1 for the PACF. So, if we had to manually assess which ARIMA models to use, a good point to start would be with p = 2 & q = 1 values or p = 1 and q = 1.

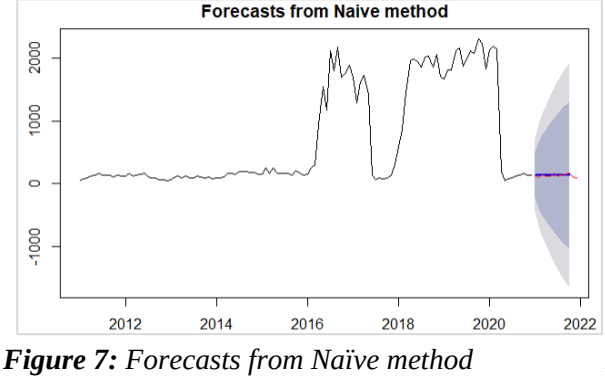

Figure 7: *Forecasts from Naïve method*

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
naive	0.00	20.00	16.20	-2.09	11.91	0.02	0.02	0.79
mod_exp	-6.08	23.76	19.58	-7.99	16.17	0.03	0.04	0.90
mod_arima	-6.08	23.76	19.58	-7.99	16.17	0.03	0.04	0.90
mod_tbats	-8.31	24.43	20.33	-9.76	16.98	0.03	0.04	0.93
mod_bats	-8.31	24.43	20.33	-9.76	16.98	0.03	0.04	0.93
mod_sts	-32.58	48.03	42.06	-29.34	34.87	0.06	-0.48	1.79
mod_stl	-217.58	250.01	217.58	-178.03	178.03	0.33	0.54	8.96
mod_neural	-1235.75	1465.80	1312.02	-940.14	1023.03	2.02	-0.04	54.97

Figure 8: *Model Performance Diagnostics*

We can see that ARIMA models did not perform well on this data. Structural models seem to have done a better job at fitting the data. A naive model which forecasts a flat line is also included in this analysis. We use the Forecast package which is pretty efficient in model fitting. The Structural model seem to have done better than other models on this data based on almost all diagnostic metrics.

Bibliographic Reference:

Themes, M. H. (2021) *Time Series Analysis using R – forecast package* Retrieved from https://www.r-bloggers.com/2014/04/time-series-analysis-using-r-forecast-package/

Mwanza, E. (December 18, 2016) *Consumer Financial Protection Bureau-A Forecast of Consumer Complaints in R* Retrieved from https://stats.idre.ucla.edu/r/faq/how-can-i-generate-bootstrap-statistics-in-r/

(An official website of the United States government, 2021) *Consumer Complaint Database* Retrieved from https://www.consumerfinance.gov/data-research/consumer-complaints/