Part 3

In [ ]:

```r
setwd("/home/leoKraushaar/Documents/School/Year 3/Semester 2/STAT 413/Project/protests/")
set.seed(42)
```

Libraries

In [ ]:

```r
library(MASS)
library(dplyr)
```

Clean Data

In [ ]:

```r
newMonth <- function(x) {
    if (x %in% c("December", "January", "February")) {
        return("Winter")
    } else if (x %in% c("March", "April", "May")) {
        return("Spring")
    } else if (x %in% c("June", "July", "August")) {
        return("Summer")
    } else {
        return("Fall")
    }
}
```

In [ ]:

```r
new_retail <- read.csv("data/clean/new_retail.csv")[, -1]

new_retail$season <- as.factor(sapply(new_retail$month, newMonth))
new_retail$month <- NULL
colnames(new_retail)[1] <- "prov"

new_retail$year <- as.numeric(new_retail$year)
new_retail$prov <- as.factor(new_retail$prov)

head(new_retail)
```

A data.frame: 6 × 4

	prov	retail	year	season
	<fct>	<int>	<dbl>	<fct>
1	Alberta	6726992	2017	Winter
2	British Columbia	7277591	2017	Winter
3	Manitoba	1749096	2017	Winter
4	New Brunswick	1049815	2017	Winter
5	Newfoundland and Labrador	800919	2017	Winter
6	Northwest Territories	65317	2017	Winter

In [ ]:

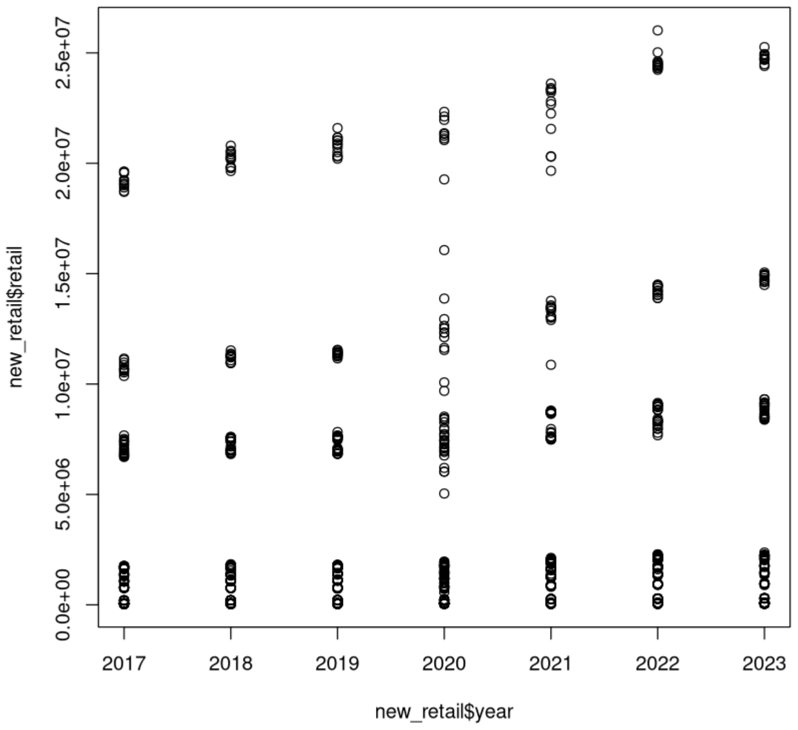
```r
data_2023 <- data[as.character(data$year) == "2023", ]

total_protests <- aggregate(protests ~ prov, data=data_2023, sum)
total_protests
```

A data.frame: 13 × 2

prov <fct>	protests <int>
Alberta	139
British Columbia	284
Manitoba	118
New Brunswick	61
Newfoundland and Labrador	61
Northwest Territories	6
Nova Scotia	85
Nunavut	11
Ontario	627
Prince Edward Island	29
Quebec	270
Saskatchewan	56
Yukon	18

```
In [ ]: plot(new_retail$year, new_retail$retail, type="p")
```



```
In [ ]: retail_predictor <- step(lm(retail ~ ., data=new_retail))
        summary(retail_predictor)
```

```
Start:  AIC=29552.41
retail ~ prov + year + season

         Df  Sum of Sq        RSS   AIC
<none>                 5.9960e+14 29552
- season  3 8.0735e+12 6.0767e+14 29561
- year    1 1.8588e+14 7.8548e+14 29845
- prov   12 4.2392e+16 4.2991e+16 34194
```

```
Call:
lm(formula = retail ~ prov + year + season, data = new_retail)

Residuals:
     Min       1Q   Median       3Q      Max 
-7709160  -300011   -45066   329186  3847004 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   -409202370   22826475 -17.927  < 2e-16 ***
provBritish Columbia              731375     115239   6.347 3.24e-10 ***
provManitoba                    -5471763     115239 -47.482  < 2e-16 ***
provNew Brunswick               -6192468     115239 -53.736  < 2e-16 ***
provNewfoundland and Labrador   -6572306     115239 -57.032  < 2e-16 ***
provNorthwest Territories       -7335838     115239 -63.657  < 2e-16 ***
provNova Scotia                 -5891235     115239 -51.122  < 2e-16 ***
provNunavut                     -7363982     115239 -63.902  < 2e-16 ***
provOntario                     14308586     115239 124.164  < 2e-16 ***
provPrince Edward Island        -7168491     115239 -62.205  < 2e-16 ***
provQuebec                       5010660     115239  43.480  < 2e-16 ***
provSaskatchewan                -5572432     115239 -48.355  < 2e-16 ***
provYukon                       -7327016     115239 -63.581  < 2e-16 ***
year                              206291      11300  18.256  < 2e-16 ***
seasonSpring                     -232696      63923  -3.640 0.000285 ***
seasonSummer                      -55290      63923  -0.865 0.387264    
seasonWinter                     -100333      63923  -1.570 0.116806    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 746800 on 1075 degrees of freedom
Multiple R-squared:  0.9861,	Adjusted R-squared:  0.9859 
F-statistic:  4772 on 16 and 1075 DF,  p-value: < 2.2e-16
```

In [ ]:

```
data <- read.csv("data/merged_data.csv")[, -1]
data$food <- NULL
data$manufac <- NULL
```

In [ ]:

```
head(data)
```

A data.frame: 6 × 8

	year	month	GEO	pop	protests	retail	oil	power
	<int>	<chr>	<chr>	<int>	<int>	<dbl>	<int>	<int>
1	2022	April	Alberta	4480956	17	7989056	3983	6069621
2	2022	April	British Columbia	5310164	42	8959229	77433	5240902
3	2022	April	Manitoba	1405197	2	2083495	6290	2168371
4	2022	April	New Brunswick	801778	5	1340707	1818	1171958
5	2022	April	Newfoundland and Labrador	529249	2	920444	77160	686123
6	2022	April	Northwest Territories	44828	0	76390	0	58889

In [ ]:

```
summary(model)
```

```
Call:
glm.nb(formula = protests ~ prov + retail + season, data = data, 
    init.theta = 8.30561596, link = log)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    4.994e+00  1.256e+00   3.975 7.03e-05 ***
provBritish Columbia           8.155e-01  1.631e-01   5.000 5.73e-07 ***
provManitoba                  -1.915e+00  9.211e-01  -2.079 0.037605 *  
provNew Brunswick             -2.624e+00  1.044e+00  -2.513 0.011965 *  
provNewfoundland and Labrador -3.070e+00  1.113e+00  -2.757 0.005833 ** 
provNorthwest Territories     -5.358e+00  1.269e+00  -4.224 2.40e-05 ***
provNova Scotia               -2.413e+00  9.935e-01  -2.429 0.015133 *  
provNunavut                   -4.985e+00  1.263e+00  -3.947 7.92e-05 ***
provOntario                    5.597e+00  2.449e+00   2.285 0.022314 *  
provPrince Edward Island      -4.113e+00  1.216e+00  -3.382 0.000719 ***
provQuebec                     2.190e+00  9.316e-01   2.351 0.018711 *  
provSaskatchewan              -2.582e+00  9.418e-01  -2.741 0.006116 ** 
provYukon                     -4.138e+00  1.245e+00  -3.325 0.000884 ***
retail                        -2.605e-07  1.496e-07  -1.741 0.081671 .  
seasonSpring                  -6.502e-02  8.264e-02  -0.787 0.431441    
seasonSummer                  -5.522e-01  8.629e-02  -6.399 1.57e-10 ***
seasonWinter                  -2.287e-01  8.873e-02  -2.578 0.009946 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(8.3056) family taken to be 1)

    Null deviance: 2091.2  on 298  degrees of freedom
Residual deviance:  349.2  on 282  degrees of freedom
AIC: 1585

Number of Fisher Scoring iterations: 1


              Theta:  8.31 
          Std. Err.:  1.46 

 2 x log-likelihood:  -1548.997 
```

In [ ]:

```
data$month <- sapply(data$month, newMonth)
colnames(data)[2] <- "season"
```

In [ ]:

```
standardize <- function(x, mu, std) {
    return((x-mu)/std)
}

# data$pop <- sapply(data$pop, function(x) standardize(x, mean(data$pop), sd(data$pop)))
```

In [ ]:

```
colnames(data)[colnames(data) == "GEO"] <- "prov"

data$prov  <- as.factor(data$prov)
data$season <- as.factor(data$season)
data$year <- as.factor(data$year)
```

In [ ]:

```
total_protests <- data[as.character(data$year) == "2023", ] %>% group_by(prov) %>% summarise(total = sum(protests))
total_protests %>% group_by(prov) %>% summarise(mean = mean(total))
```

A tibble: 13 × 2

prov	mean
<fct>	<dbl>
Alberta	139
British Columbia	284
Manitoba	118
New Brunswick	61
Newfoundland and Labrador	61
Northwest Territories	6
Nova Scotia	85
Nunavut	11
Ontario	627
Prince Edward Island	29
Quebec	270
Saskatchewan	56
Yukon	18

```
In [ ]: # data$retail <- NULL
        head(data)
```

A data.frame: 6 × 8

	year	season	prov	pop	protests	retail	oil	power
	<fct>	<fct>	<fct>	<int>	<int>	<dbl>	<int>	<int>
1	2022	Spring	Alberta	4480956	17	7989056	3983	6069621
2	2022	Spring	British Columbia	5310164	42	8959229	77433	5240902
3	2022	Spring	Manitoba	1405197	2	2083495	6290	2168371
4	2022	Spring	New Brunswick	801778	5	1340707	1818	1171958
5	2022	Spring	Newfoundland and Labrador	529249	2	920444	77160	686123
6	2022	Spring	Northwest Territories	44828	0	76390	0	58889

Build Model

```
In [ ]: # std_data <- data
        # std_data$retail <- sapply(std_data$retail, function(x) {standardize(x, mean(std_data$retail), sd(std_data$retail)


        # model <- glm.nb(protests ~ prov + retail + season, data=std_data)

        model <- glm.nb(protests ~ prov + retail + season, data=data)
        summary(model)
```

```
Call:
glm.nb(formula = protests ~ prov + retail + season, data = data, 
    init.theta = 8.30561596, link = log)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                    4.994e+00  1.256e+00   3.975 7.03e-05 ***
provBritish Columbia           8.155e-01  1.631e-01   5.000 5.73e-07 ***
provManitoba                  -1.915e+00  9.211e-01  -2.079 0.037605 *  
provNew Brunswick             -2.624e+00  1.044e+00  -2.513 0.011965 *  
provNewfoundland and Labrador -3.070e+00  1.113e+00  -2.757 0.005833 ** 
provNorthwest Territories     -5.358e+00  1.269e+00  -4.224 2.40e-05 ***
provNova Scotia               -2.413e+00  9.935e-01  -2.429 0.015133 *  
provNunavut                   -4.985e+00  1.263e+00  -3.947 7.92e-05 ***
provOntario                    5.597e+00  2.449e+00   2.285 0.022314 *  
provPrince Edward Island      -4.113e+00  1.216e+00  -3.382 0.000719 ***
provQuebec                     2.190e+00  9.316e-01   2.351 0.018711 *  
provSaskatchewan              -2.582e+00  9.418e-01  -2.741 0.006116 ** 
provYukon                     -4.138e+00  1.245e+00  -3.325 0.000884 ***
retail                        -2.605e-07  1.496e-07  -1.741 0.081671 .  
seasonSpring                  -6.502e-02  8.264e-02  -0.787 0.431441    
seasonSummer                  -5.522e-01  8.629e-02  -6.399 1.57e-10 ***
seasonWinter                  -2.287e-01  8.873e-02  -2.578 0.009946 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(8.3056) family taken to be 1)

    Null deviance: 2091.2  on 298  degrees of freedom
Residual deviance:  349.2  on 282  degrees of freedom
AIC: 1585

Number of Fisher Scoring iterations: 1


              Theta:  8.31 
          Std. Err.:  1.46 

 2 x log-likelihood:  -1548.997 
```

Perform Monte Carlo Simulation

```
In [ ]: data <- data[c("prov", "season", "retail", "year")]

        dim(data[as.character(data$year) == "2023", ][, c(-3)])
```

143 · 3

Create New Data

1. High Retail Sales

```
In [ ]: num_iterations <- 10000
```

```
results <- c()

for (i in 1:num_iterations) {

    # Get constant values
    blank_data <- data[as.character(data$year) == "2023", ][, c(-3)]

    blank_data$year <- 2030
    # Predict retail uniformly from interval
    blank_data <- cbind(blank_data, predict.lm(retail_predictor, newdata=blank_data, interval = "prediction"))
    blank_data <- as.data.frame(blank_data)
    pred_retails <- runif(n=nrow(blank_data), min=blank_data$lwr, max=blank_data$upr)

    blank_data$retail <- pred_retails
    blank_data[, c("fit", "lwr", "upr")] <- NULL
    blank_data$year <- NULL

    # Predict protests
    blank_data$protests <- predict.glm(model, newdata=blank_data, type="response")
    # Round off to nearest integer
    blank_data$protests <- round(blank_data$protests)

    rownames(blank_data) <- NULL
    results <- rbind(results, blank_data)
}
```

In [ ]:

```
dim(data)
```

In [ ]:

```
# write.csv(results2, "data/montecarlo/2030.csv")
```