Monte Carlo Analysis on Gasoline Price

Group Members:

- Yue Wan (yuewan2)
- Yumian Liu (yumianl2)
- Theodore Andrew (tandrew2)
- Aldo Sanjoto (sanjoto2)
- Changhao Ying (cying4)

1. Introduction:

Stepping into today's modern world, gasoline price is gaining more attention than usual. It is both due to the fact of increasing vehicles on street and also because of the role of gasoline as a direct and evident indicator for our energy resources consumption. "No prices are more visible to the public than gasoline prices", said the U.S. Bureau of Labor. Therefore, because of its explicit features and arising importance, we would like to discover more about the hidden facts lying behind the gasoline price across years. To be more specific, we wanted to make predictions on future gasoline price based on the empirical model from our analysis.

The data we were working on is the monthly price of gasoline from 1976 to 2004, which is from *mathforum*.org. By looking at the data, we started to think whether the monthly distribution of gasoline price is from the same data generating process and whether the price is increasing in a constant rate or by some other arbitrary functions. We first did several permutations test and excluded the month/season factors' influence on price distribution. Continued our work on average price of each year, we needed to simulate more data since this is only a monthly data of 29 years, which means that there are only 348 observations. By doing the Bootstrap Resampling Method, we were able to estimate the average pricing distribution, the ECDF for each year. After we retrieved the ECDF, which also indicates its likelihood distribution of average price for each year, we started our bayesian analysis. The reason why we used bayesian analysis is simple. The average gasoline price for each year is not independent and must rely on the price from that of previous year more or less. In order to validate our model, we generate predictions of average price from our bayesian model for year 1977-2004 and compared them with the existing data. At last, we are able to predict the price of 2005 and even the price of the distant future.

2. Methods:

a. Permutation Test:

After we cleaned the data, we first created visualizations of the average price for each year to check for the general trend of price:

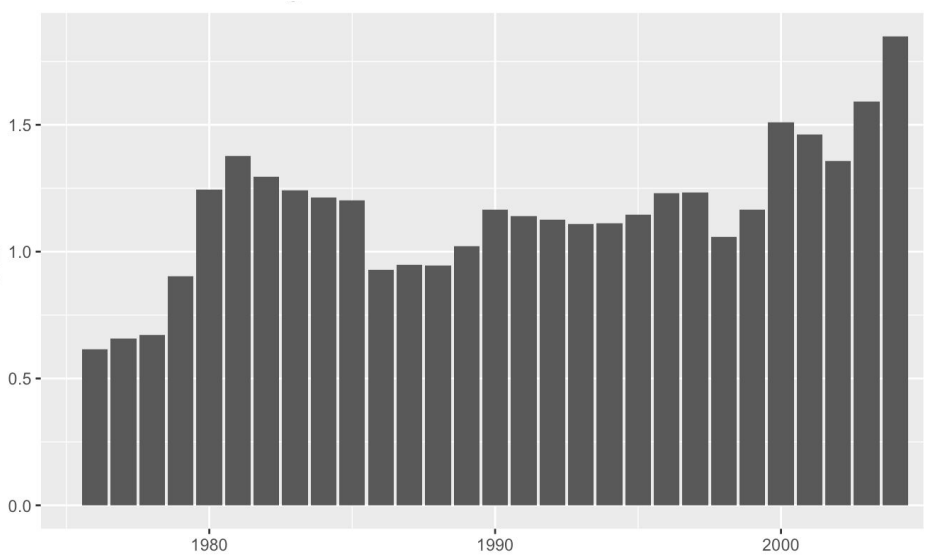

Figure 1: Average gasoline price for year 1976-2004

The average gasoline price is mostly going upward after each year, which is consistent with the economics growth. Abnormal price jump and drop such as the average price for 1979 and 1986 will be discussed later in conclusion section. To explore deeper into the price difference within each month, we subsequently plotted the monthly price difference across years:

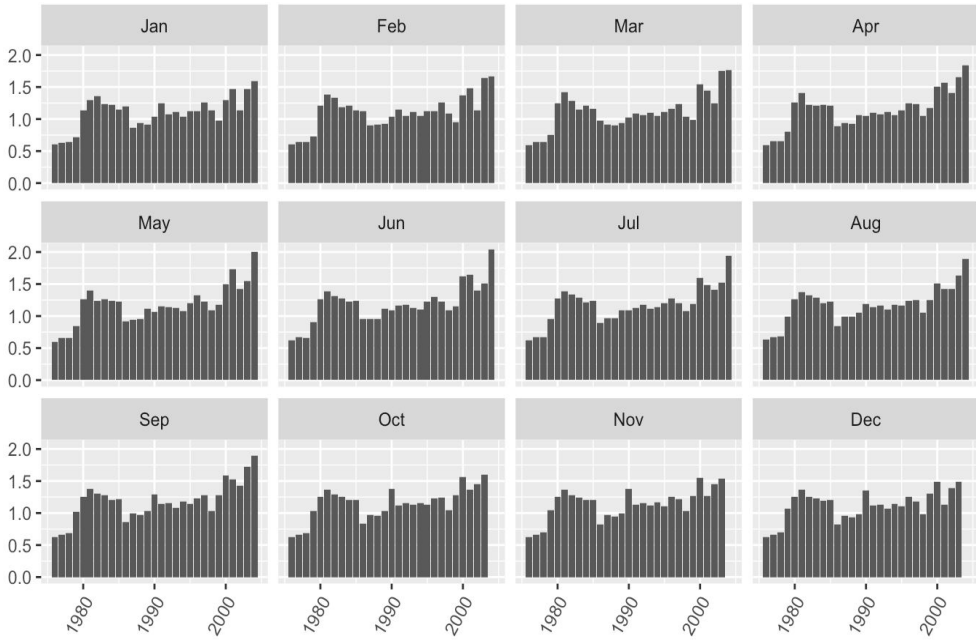

Figure 2: Gasoline price change within each month

From figure 2, we can see that all the price plot have a similar shape, which means that the price difference between month may come from the same distribution. In other words, the price difference between months may not be significant enough to impact the price difference between years. If this is truly the case, then it will be reasonable for us to put the price difference between months aside and focus on analyzing the average price. In order to test our hypothesis, we generated 66 month price data pairs (e.g. {January, February}, {January, March}) and performed the permutation test for each pair. We both did a t-statistic permutation test, which is sensitive to difference in the means of the two distribution, and a Kolmogorov-Smirnov permutation test, which is sensitive to the maximum difference between two cdf. For each test, we generated 10000 permutations of the pair data to get the final p-value for our null hypothesis.

Furthermore, in order to make the test more concrete and practical, we grouped our month data into four seasons to conduct the permutation test again since season can be a stronger indicator for gasoline price regarding to the weather and climate. The result for both permutation tests prove that those price changes were coming from the same distribution (detail in result section). We then continued our analysis by focusing on the average gasoline price of each year.

b. Bootstrap Resampling:

Since our later bayesian analysis will merely focus on the average gasoline price of each year, we need to find the ECDF, or in other words, the likelihood distribution of the average price for each year. Therefore, we did the bootstrap resampling method on the monthly gasoline price with in each year. For each bootstrap resamples, we calculated the mean of each sample and store its value. After 10000 times of resampling, we eventually retrieved 10000 mean price value and formed a likelihood distribution. We then repeated this process for the other 28 years of data and retrieved 29 likelihood distribution of the average gasoline price.

c. Bayesian Analysis:

Based on the bootstrap resampling, we found out that the distribution of price for each year is normal. In order to perform our analysis, we choose the normal conjugate prior distribution. Therefore, to compute the expected posterior mean, which is the predicted price for each year, we used posterior normal mean formula to find it.

$$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$$

Figure 3: Posterior hyperparameter for normal conjugate prior distribution

First of all, we started to compute the predicted price for year 1977. Because we did not have an informative prior mean, we pick the prior normal mean equals to 1. The formula for the

posterior normal mean involves fix sigma value, which is standard deviation. Because there was not really a big variation within the histogram of each year, we decide to set fix sigma equals to 1. For the following years, we set the prior mean equals to the posterior mean that we obtain from previous year. Finally, we achieved our goal to find the predicted average price for 2005.

d. Shiny App:

At last, we created a shiny app to show users the working demo of our project. Panels include the visual EDA of the data and the procedure of analyzing gasoline price.

3. Results:

a. Permutation Test:

The results was not surprising. For the first permutation test which performed on 66 sample pairs, all of the 66 p-values were much larger than 0.05. Also for the second permutation test in which we grouped the month data into season, all the p-values were much larger than 0.05 as well. Therefore, we failed to reject the null hypothesis that all price difference within each month is highly probable to come from the same distribution.

```
1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 0.9183082 0.7119288 1.0000000 0.9895010
1.0000000 1.0000000 0.7306269 0.6293371 0.4625537 0.8774123 0.9013099 1.0000000 0.6139386 0.5177482
0.3672633 1.0000000 1.0000000 0.7192281 0.6229377 0.4426557 1.0000000 0.6958304 0.5929407 0.4271573
0.5914409 0.4978502 0.3365663 0.8801120 0.6642336 0.7695230
```

Figure 4.1: p-values from the 66 data pairs t-statistic permutation test

```
0.9487051 0.7815218 0.9987001 0.7869213 0.7781222 0.7854215 0.3614639 0.3472653 0.5844416 0.5579442
0.9983002 0.9977002 0.7898210 0.5656434 0.3619638 0.5689431 0.2134787 0.2134787 0.3700630 0.3617638
0.7829217 0.9532047 0.5656434 0.3613639 0.5747425 0.3741626 0.5749425 0.3479652 0.5250475 0.7500250
0.9462054 0.7792221 0.9463054 0.9465053 0.7888211 0.7568243 0.9424058 0.8758124 0.9973003 0.9984002
0.9988001 0.9516048 0.9907009 0.9922008 0.8807119 0.9983002 0.9979002 0.9981002 0.9893011 0.9237076
0.8749125 0.9981002 0.9982002 0.9920008 0.9259074 0.9216078 0.9989001 0.9606039 0.9939006 0.8031197
0.9313069 0.7393261 0.7431257 0.9999000 0.9911009 0.9907009
```

Figure 4.2: p-values from the 66 data pairs ks-statistic permutation test

```
1.0000000 1.0000000 0.2929707 0.5870413 0.1395860 0.2983702
```

Figure 4.3: p-values from the t-statistic 4 seasons permutation test

```
0.8131187 0.8578142 0.9323068 0.9943006 0.3207679 0.3856614
```

Figure 4.4: p-values from the ks-statistic 4 seasons permutation test

b. Bootstrap Resampling:

Below is the graph we created after we retrieved the mean of each bootstrap resamples of each year.

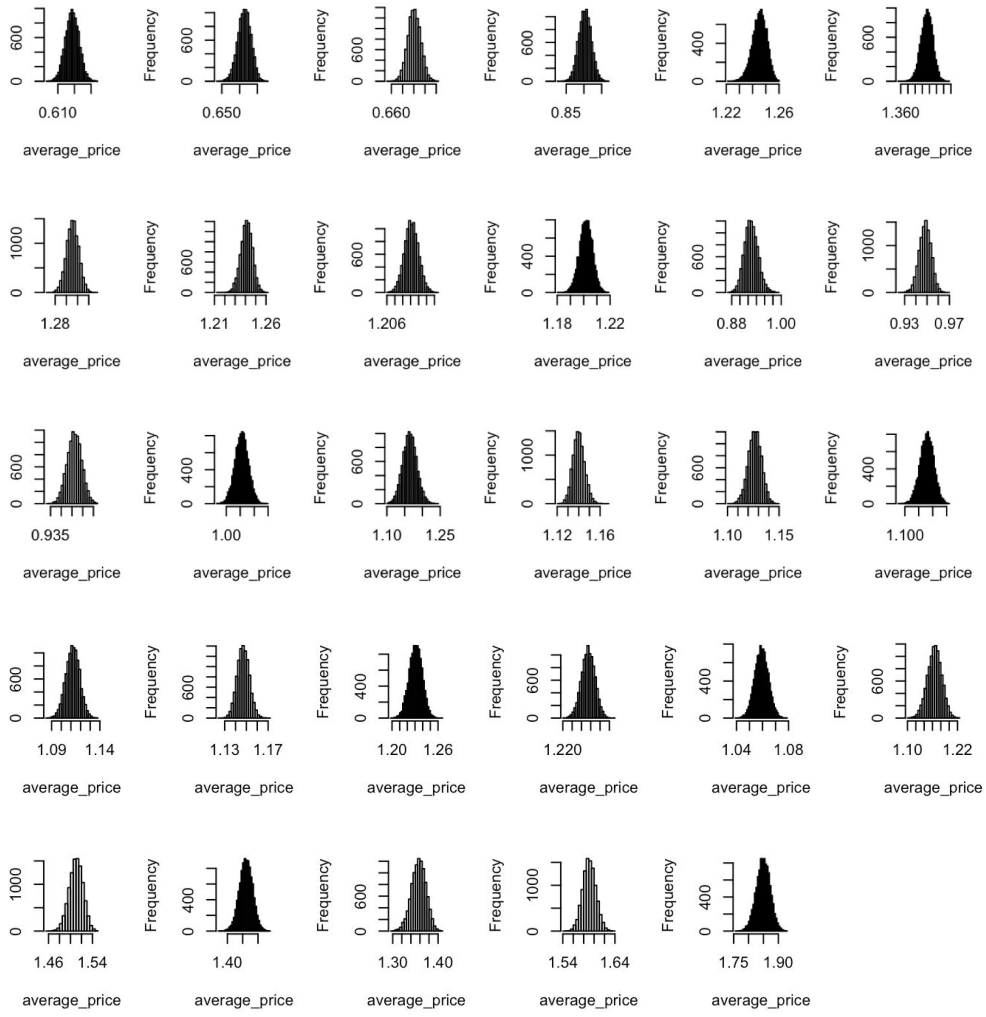

Figure 5: The likelihood distribution of average price of year 1976-2004

It is obvious that all the price likelihood distribution came from the normal distribution with different mean but similar variance. The mean of each likelihood distribution also followed an increasing pattern, which is consistent with what we observed at the very beginning. Most sigma of the likelihood distribution are around 0.003~0.005. For future bayesian analysis, we set the sigma to be fixed for convenience. Disadvantages will be discussed in conclusion section.

c. Bayesian Analysis:

This is the table of predicted price for each year, the price we predicted for 2005 is 1.16 dollar per gallon:

Gasoline Predicted Average Price

Year	Predicted Price
1977	0.62
1978	0.64
1979	0.65
1980	0.71
1981	0.82
1982	0.91
1983	0.97
1984	1.00
1985	1.02
1986	1.04
1987	1.03
1988	1.02
1989	1.02
1990	1.02
1991	1.03
1992	1.04
1993	1.04
1994	1.04
1995	1.05
1996	1.05
1997	1.06
1998	1.07
1999	1.07
2000	1.07
2001	1.09
2002	1.10
2003	1.11
2004	1.13
2005	1.16

Figure 6: Predicted price from Bayesian Analysis

Based on the table, the predicted average price for each year increases. To show clearer relation of predicted average price between each year, we also plot the graph of the predicted price below.

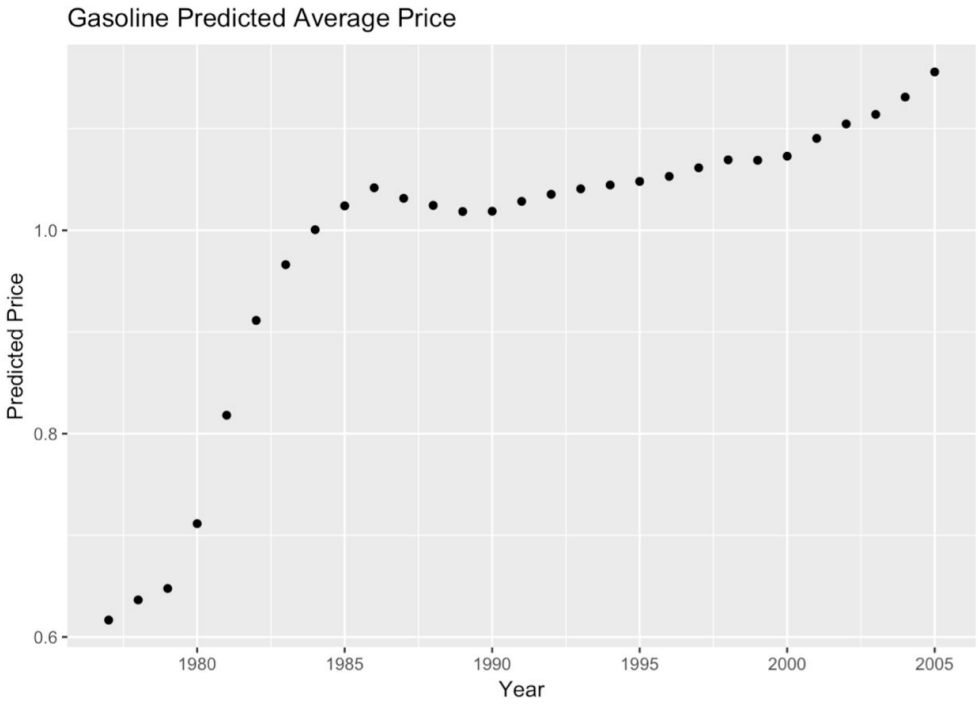

Figure 7: The visualization of our predicted price

4. Conclusion/discussion:

The predicted price from 1977 to 2004 is more of a validation of our analysis. Below is the comparison between the actual average price and the predicted average price we retrieved from our method:

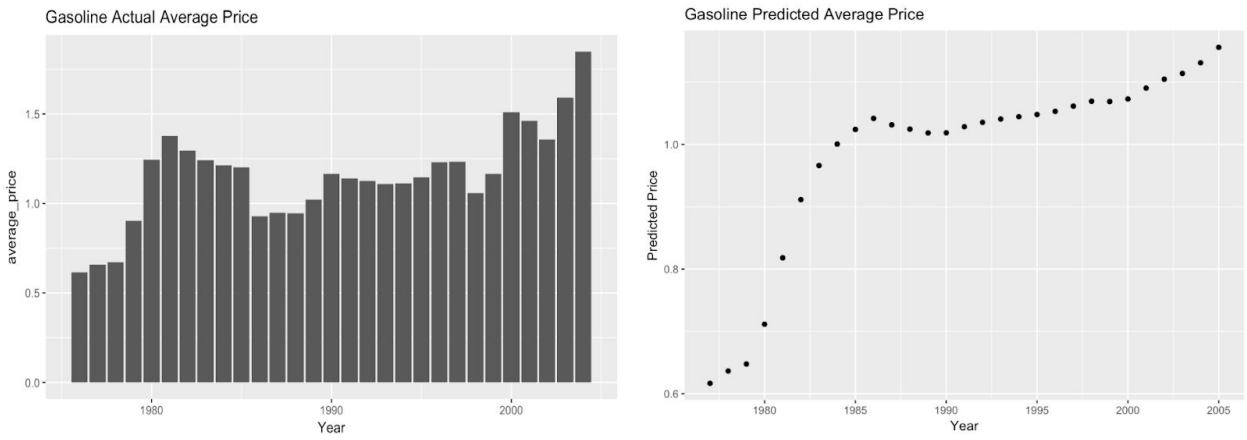

Figure 8: The comparison between actual data and the predicted data

It is obvious to see that our predicted price follow the same pattern compared to the actual data, which to some extent validify our Bayesian Model. Our model is not limited to only predict the price of 2015. In fact, all we need is the price data of current year will we be able predict the gasoline price of next year. In other words, we are able to predict the gasoline price of 2019 based on the current gasoline data in 2018.

Another interesting thing we found from the original data and from our predicted curve is the first price peak at around 1980s followed by a price drop. According to our research, an energy crisis happened in 1979 which explained the dramatic increase in gasoline price (Monthly Energy Review, 2015). Because of this crisis, nations started to put more efforts on oil production and technology innovation. However, another crisis happened in couple years later and it is called the 1980s oil glut. At the opposite of deficit, because of the surplus of the crude oil, the gasoline price collapsed (Annual Energy Review, 2006). Both cases can be counted as abnormality in gasoline market. Fortunately, predicted future price based on the gasoline price of the previous year, our Bayesian Model successfully captured those abnormalities, which further proved its strength.

However, this method still has certain disadvantages. Although our predicted graph shows the same pattern as the actual one, our predicted price is biased downward. This can be caused by some reasons:

At the very beginning, for convenience purposes, we fix the sigma of each likelihood distribution to be 1. However, fixed sigma is not realistic. In fact, when we are calculating the average price from bootstrap resampling, the variance of the bootstrap estimator mean, although

similar, is still different from others. To be more specific, we discovered that the variance of average price tended to increase dramatically after 2000. There might be some hidden variables that are also affecting the gasoline price. The true variation of the price is what our model fail to explain. Because of this, if we want to predict the gasoline price more accurately, we may need something more complex than the normal conjugate prior distribution in our Bayesian Analysis.

Another thing that is worth noticing is that when we are doing a permutation test, because we set our significance level to 0.05 we failed to reject our null hypothesis. According to Figure 4.3, the lowest is 0.1396, which is the sample difference between summer and winter. Raising the significance level higher will impact our decision. In other words, the gasoline price difference between winter and summer can be rather significant and should not be simply ignored. Future study which can bring seasons into our predictive model is needed and also promising.

5. Reference:

"Monthly Energy Review" (PDF), U.S. Energy Information Administration, November 2015

"Annual Energy Review", U.S. Energy Information Administration, 2006

"Beyond the Numbers", U.S. Bureau of Labor Statistics, Volume 2/Number 23, September 2013

6. Appendix:

a. Data source:

http://mathforum.org/workshops/sum96/data.collections/datalibrary/Price_of_Gasoline.XL.xls

b. Code snippets for Permutation Test:

```
perm_test = function(x, y, type){
  B = 10000
  z = c(x, y)
  nu = 1:58
  reps = numeric(B)
  if (type == "t"){
    t0 = t.test(x, y, var.equal = TRUE)$statistic
  } else{
    t0 = ks.test(x, y, exact =FALSE)$statistic
  }
  for (i in seq_len(B)){
    perm = sample(nu, size = 29, replace = FALSE)
    x1 = z[perm]
    y1 = z[-perm]
    if (type == "t"){
      reps[i] = t.test(x1, y1, var.equal = TRUE)$statistic %>% abs()
    } else{
      reps[i] = ks.test(x1, y1, exact = FALSE)$statistic
    }
  }
  p = mean(c(t0, reps) >= t0)
  return(p)
}
```

```
perm_simulation = function(data_list, type, nsample){
  final_results = c()
  for (k in seq_len(nsample)){
    results = rep(0, nsample-k)
    for (i in seq_len(nsample-k)){
      x = data_list[[k]]
      y = data_list[[k+i]]
      results[i] = perm_test(x, y, type)
    }
    final_results = c(final_results, results)
  }
  return(final_results)
}
```

c. Code snippets for Bootstrap Resampling Method:

```
bootstrap = function(x){
  T=mean(x)
  B=10000
  average_price=numeric(B)
  for(b in 1:B){
   xb=sample(x,50,replace=TRUE)
   average_price[b]=mean(xb, na.rm = TRUE)
  }
  return(average_price)
}

a = 1976:2004
normal_mean = numeric(length(a))
result = matrix(0, 10000, 29)
for(i in seq_along(a)){
  key = as.character(a[i])
  result[,i] = bootstrap(gasoline_perm[gasoline_perm$Year == key,]$Price)
}
```

d. Code snippets for Bayesian Analysis:

```
Year = seq(1977, 2005, 1)
post_mean = numeric(length(Year))
post_sigma_2 = numeric(length(Year))
prior_mean = 1
sigma_2_prior = 1
sigma_2_fix = 0.004
n_month = 12

post_mean[1] = (1 / (1 / sigma_2_prior + n_month / sigma_2_fix) ) *
               (prior_mean / sigma_2_prior + gasoline_price$Price[1] / sigma_2_fix)

post_sigma_2[1] = (1 / sigma_2_prior + n_month / sigma_2_fix) ^ -1

for (i in 2:length(post_mean)) {
  post_mean[i] = (1 / (1 / post_sigma_2[i-1] + n_month / sigma_2_fix) ) *
                 (post_mean[i-1] / post_sigma_2[i-1] + gasoline_price$Price[i] / sigma_2_fix)
  post_sigma_2[i] = (1 / post_sigma_2[i-1] + n_month / sigma_2_fix) ^ -1
}

gasol_predicted_avg_price = data.frame(Year = Year, Predict_Price = post_mean)
plot(gasol_predicted_avg_price$Year, gasol_predicted_avg_price$Predict_Price)
```

e. Overview of the Shiny App: