[redacted]

The Never-ending Roller Coaster of Gas Prices

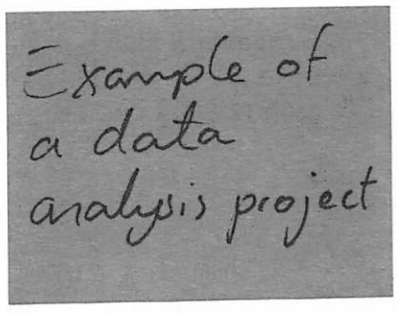


Figure 3: Seasonality

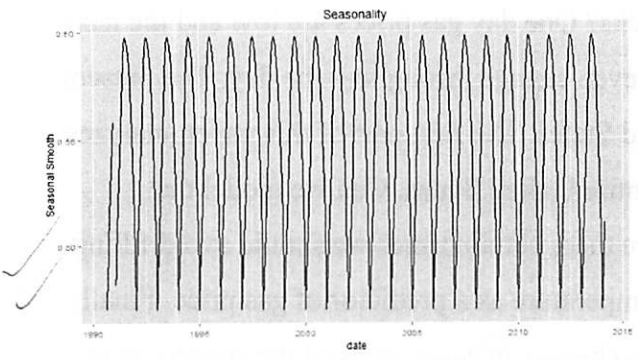
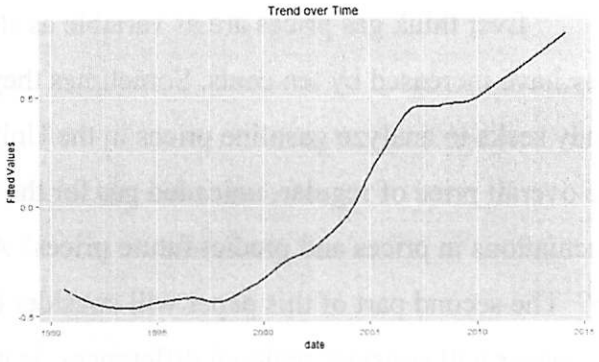


Figure 4: Trend



Recall that the data is collected every Monday. The naïve assumption is that we have weekly data and can use seasonality with a period of 52. However, further analysis shows that there are 52-53 Mondays each year depending upon the year. My initial guess is that the years with 53 Mondays follows the same pattern as leap years. A little internet research shows that it's a bit more complicated. There will be 53 Mondays when a non-leap year starts on a Monday or when a leap year starts on a Sunday or a Monday. The Gregorian calendar cycle repeats every 400 years. In those 400 years, there are 303 non-leap years and 97 leap years .There are 43 non-leap years that start on a Monday and 13+15=28 leap years that start on either a Sunday or Monday for a total of 43+28=71 years with 53 Mondays (Wikipedia). Hence the probability of a year having 53 Mondays is approximately 0.1775. Therefore a seasonal period of 52.1775 should account for the difference in the number of data points per year.

To force the price data to be stationary, I take the first difference to get rid of the trend and then difference using a lag of 52.1775 to remove the seasonality. Figure 5 below shows the residuals. The residuals do not look quite stationary after differencing out the trend and seasonality. However, there are more factors than just trend and seasonality that influence gas prices. The huge drop in 2009 gas prices seen in figure 1 is actually due to international budget crises. This study does not consider how the price of oil changes over time which could have a huge impact on gas prices. Other considerations include politics, the recession, and the oil spill.

The Never-ending Roller Coaster of Gas Prices

Ever think gas prices are as variable as stocks? One day gas prices are low and the next they have increased by ten cents. Sometimes they even change throughout the day. The present study seeks to analyze gasoline prices in the United States. The first part of this paper analyzes the overall price of regular, unleaded gas for the entire United States. Can we model the fluctuations in prices and predict future prices? Are there certain times we should avoid filling up? The second part of this paper will consider temperature as a predictor of gas price. Finally, this paper will consider regional differences. Is gas cheaper in some areas of the country than others? What is the trade-off between gas price and weather?

Part I: Analysis of United States Data

The United States Energy Information Administration (EIA) keeps records of gasoline prices per gallon for various cities and regions across the country. Every Monday, prices are collected from a sample of 900 retail outlets. Note that the prices listed are only for Monday, not a weekly average. Data is available back to August 1990 for the United States. For the purposes of this study, only regular, unleaded gas is considered.

Figure 1: Raw Data

Figure 2: Log Transform of Original Data

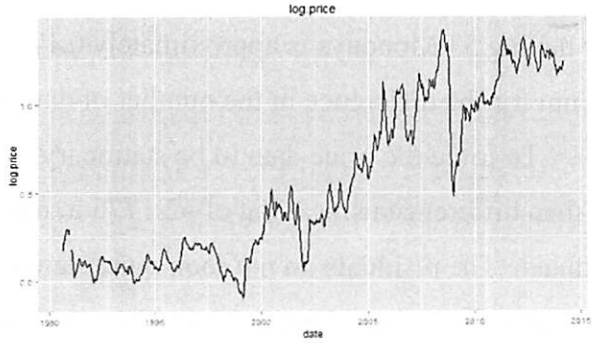


Figure 1 on the left is the raw price data for the United States. We see a clear upward trend with a large dip around 2009. Taking the log transform of the data reduces the variability. Fitting a smooth using week as an explanatory variable we see that prices follow a yearly pattern (figure 3). We can remove the seasonality and then model the trend. Figure 4 confirms that gas prices are increasing over time in addition to fluctuating seasonally. Removing the seasonality and the trend we are left with a stationary time series with which we can fit with a SARIMA model.

what does the seasonality look like oh duh next page

Figure 3: Seasonality

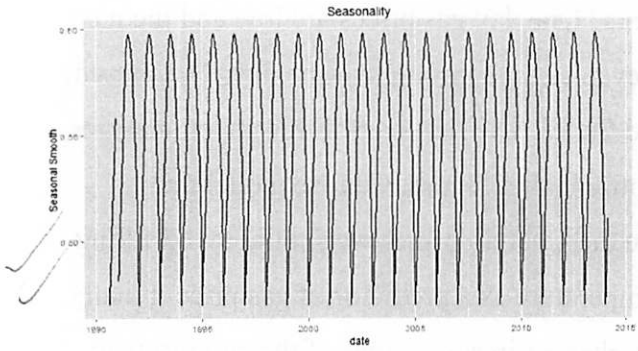
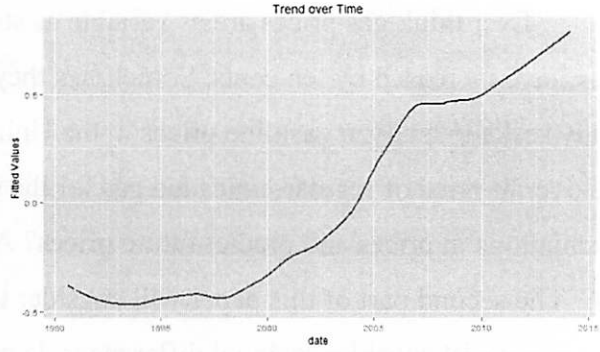


Figure 4: Trend



Recall that the data is collected every Monday. The naïve assumption is that we have weekly data and can use seasonality with a period of 52. However, further analysis shows that there are 52-53 Mondays each year depending upon the year. My initial guess is that the years with 53 Mondays follows the same pattern as leap years. A little internet research shows that it's a bit more complicated. There will be 53 Mondays when a non-leap year starts on a Monday or when a leap year starts on a Sunday or a Monday. The Gregorian calendar cycle repeats every 400 years. In those 400 years, there are 303 non-leap years and 97 leap years .There are 43 non-leap years that start on a Monday and 13+15=28 leap years that start on either a Sunday or Monday for a total of 43+28=71 years with 53 Mondays (Wikipedia). Hence the probability of a year having 53 Mondays is approximately 0.1775. Therefore a seasonal period of 52.1775 should account for the difference in the number of data points per year.

To force the price data to be stationary, I take the first difference to get rid of the trend and then difference using a lag of 52.1775 to remove the seasonality. Figure 5 below shows the residuals. The residuals do not look quite stationary after differencing out the trend and seasonality. However, there are more factors than just trend and seasonality that influence gas prices. The huge drop in 2009 gas prices seen in figure 1 is actually due to international budget crises. This study does not consider how the price of oil changes over time which could have a huge impact on gas prices. Other considerations include politics, the recession, and the oil spill.

Figure 5: Residuals

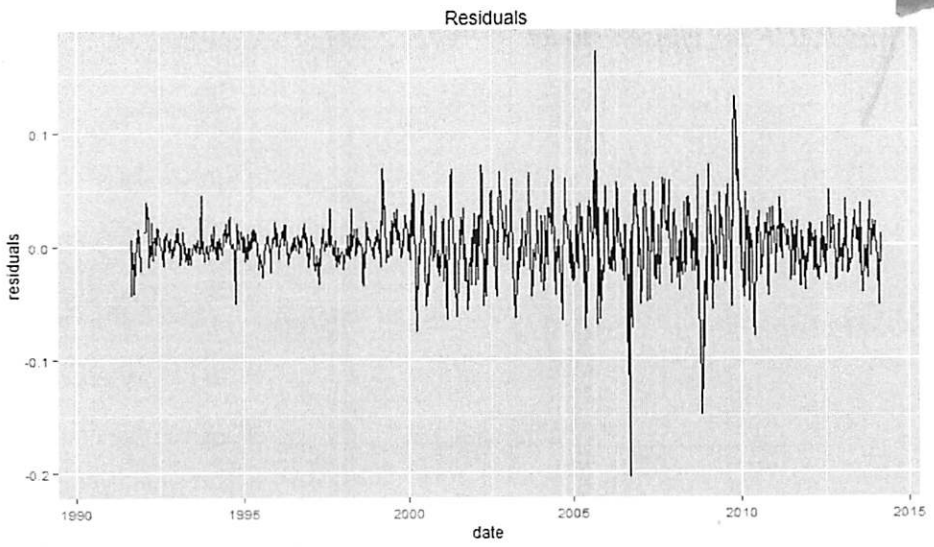
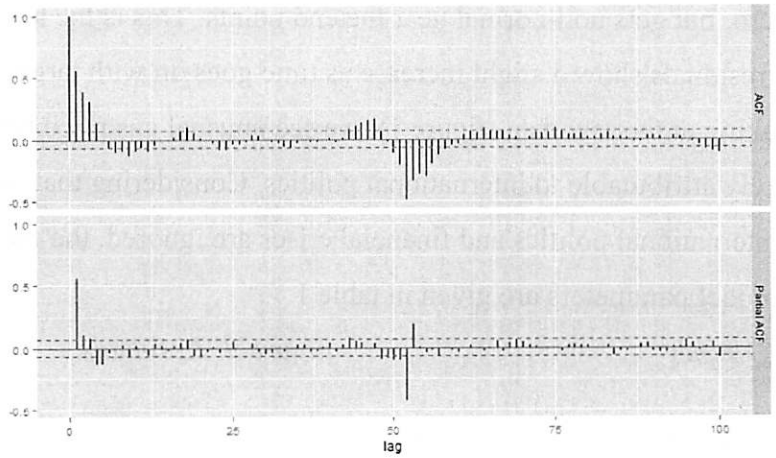


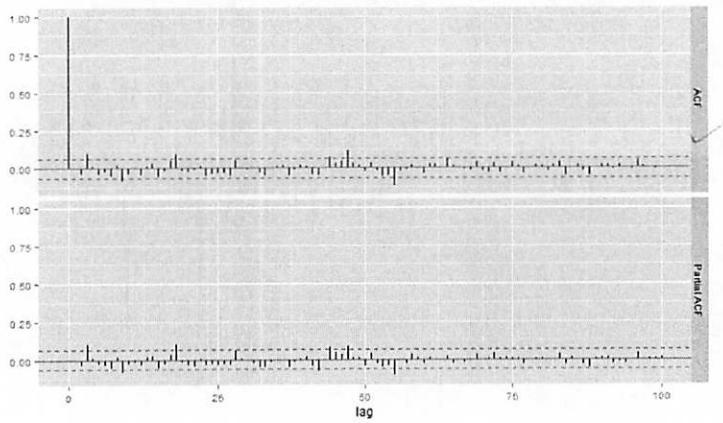
Figure 6: ACF and PACF for residuals



Above are the autocorrelation and partial autocorrelation plots for the residuals of the log transformed price data. The autocorrelation function is decreasing, while the partial autocorrelation plot cuts off after lag 1 and lags 52 and 53. The correlation at lag 53 is a bit worrisome, but it is probably is a result of the way I chose the seasonal lag. The correlation is much less than when I naively used a lag of 52. This suggests an AR part and AR seasonal part. My initial guess is SARIMA (2,1,0)x(1,1,0)_52.1775. This gave a low AIC value, but the correlation plots show correlations beyond lag 1. Considering other models, the best fit in terms of correlation functions is SARIMA (2,1,0)x(1,1,1)_52. Figure 7 shows the autocorrelation and

partial autocorrelation functions for the chosen model. The correlations quickly go to zero with some negligible correlations around lag 53.

Figure 7: ACF and PACF for chosen model



The other model diagnostics are the normal QQ plot of the residuals (figure 8) and the plot of residuals against time (figure 9) are shown below. Now, the normal QQ plot looks good for quantiles near zero, but gets non-normal near the end points. This is the best normal QQ plot given the data. The residuals show a slight increase as time goes on with larger residuals around 2006 and 2009. Looking at the raw data (figure 1), we see unusual gas price behavior around those years most likely attributable to international politics. Considering that other factors such as the price of oil, international politics and financial crises are ignored, the model is considered good enough. The model parameters are given in table 1.

Figure 8 Normal QQ Plot

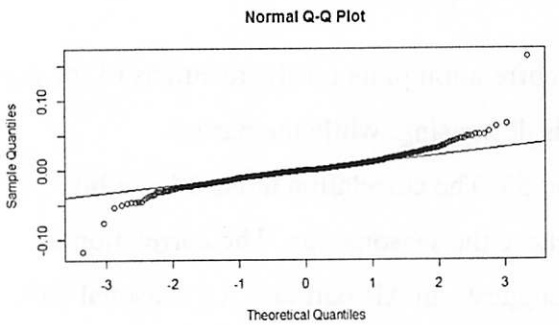


Figure 9: Residuals

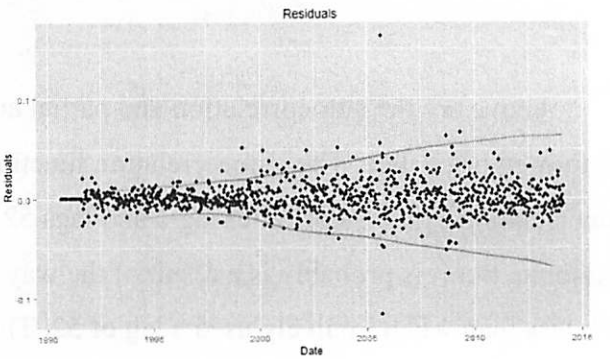


Table 1: Model Parameters

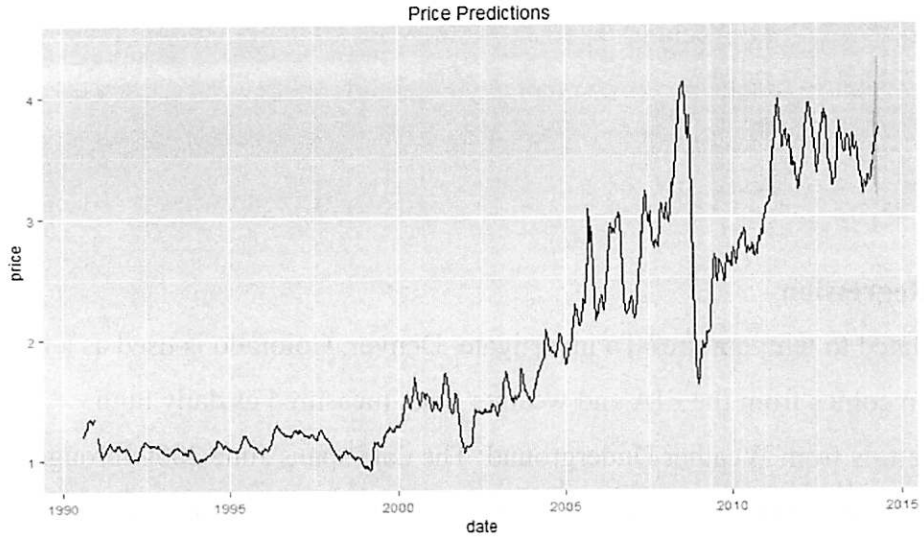
	ar1	ar2	sar1	sma1
Value	0.4876	0.0780	-0.0097	-0.9702
SE	0.0293	0.0294	0.0339	0.0689

Given our model, we can predict future gas prices. Table 2 gives the predicted gas prices for the next 8 time points corresponding to the next two months. Now, the data I used ends on February 24, so we can check the predicted values against the actual values. For March 3, 2014 the EIA reported a national average for unleaded regular gas of $3.479. The model does pretty well! Figure 10 plots the predicted values with 95% confidence bands. We see that the model provides a convincing continuation of the observed data.

Table 1: Predicted Gas Prices

Date	Price	SE
2014-03-03	3.4896	1.0157
2014-03-10	3.5326	1.0284
2014-03-17	3.5664	1.0405
2014-03-24	3.6147	1.0518
2014-04-07	3.6476	1.0622
2014-04-14	3.6768	1.0718
2014-04-21	3.6953	1.0806
2014-04-28	3.7218	1.0889

Figure 10: Predictions



The above analysis takes a simplistic approach to fitting a model and predicting future observations. As noted, a seasonal lag of 52.1775 is used to account for there being 53 Mondays in some years. A better way to take out the seasonality is to model the seasonality using a sinusoidal model, remove the seasonality, then fit a model and back-transform to get predictions

on the original scale. It would also be nice to have daily data instead just Monday as we know that gas prices rise and fall throughout the week.

What about some practical interpretations of the data? Figure 11 shows the average price of gas per month. Gas prices are indeed higher in the summer. We see an increase in gas prices in the beginning of the year and then prices decrease from September to the end of the year. For the average weekly price data (Figure 12) we see peaks at weeks 22 and 36 corresponding to Memorial Day and Labor Day weekends respectfully. The peak around week 9 corresponds to the end of February (don't think anything is special about that, but gas is cheaper). Gas prices rise near the end of the year and the beginning of the year corresponding to Christmas and New Years travel. The take away message is that we can use date to predict the gas price to a certain extent. Due to the variability of gas prices with respect to politics and financial matters, I would be skeptical in making claims much past two months out.

Figure 11: Average Price by Month

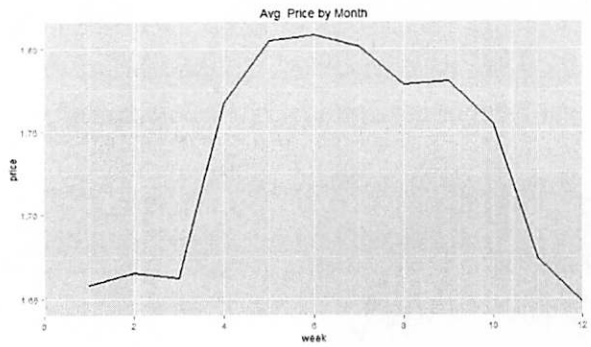
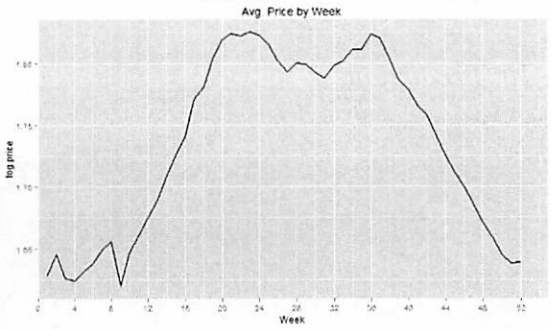


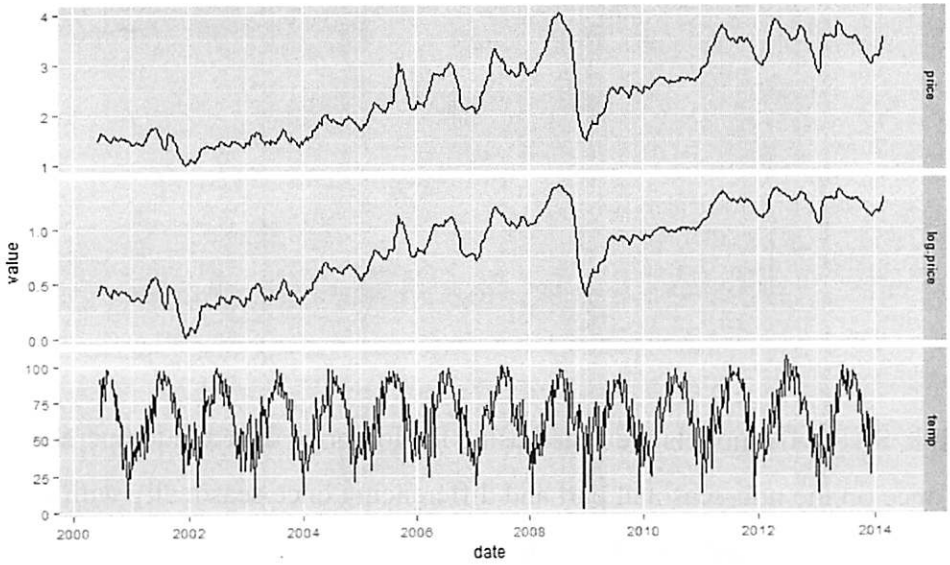
Figure 12: Average Price by Week



Part II: Temperature Regression

Are gas prices related to temperature? To investigate, Denver, Colorado is used as a test case. Gas price data again comes from the EIA and weather data, measured as daily high temperatures in Fahrenheit, is from Weather Underground. The data spans June 2000 through December 2013. The pattern of Denver gas prices over time is similar to the national pattern. Further examination of the Denver weather data shows that Denver has four distinct seasons in terms of temperature. Figure 13 displays the variables of interest. The price data is log transformed and will be compared to the weather data. The question of interest is whether we can use temperature to predict gas prices.

Figure 13: Denver (top: price, middle: log transform of price, bottom: temperature)



Before proceeding to fit a model, both series must be transformed into stationary time series. To remove the trend in the price data, we will take the first difference. To remove the seasonality, the same convention as above is used differencing once with a lag of 52.1775. A simple scatterplot (figure 14) suggests that there might be a relationship between temperature and gas price, but this relationship is not particularly strong. Fitting a model with uncorrelated errors, temperature is a significant predictor of gas prices (p=0.024). However, this naively assumes that the errors are uncorrelated. Upon further examination, we find that the residuals are correlated (figure 15). This suggests that we need both an AR and seasonal AR component.

Figure 14: Gas Prices v. Temperature (stationary series)

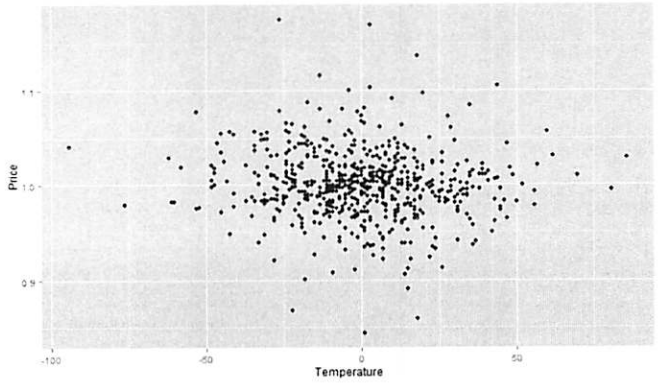
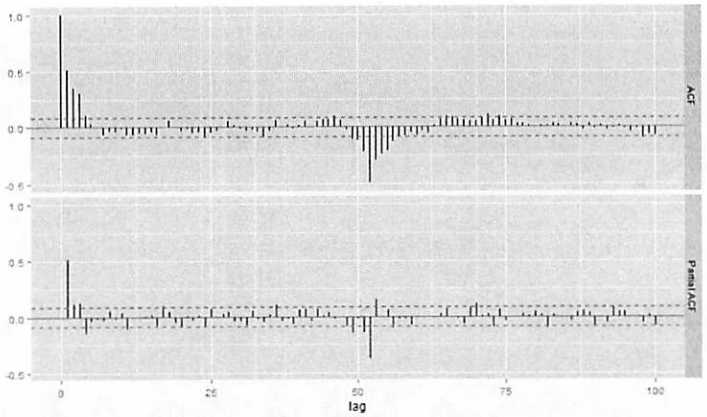
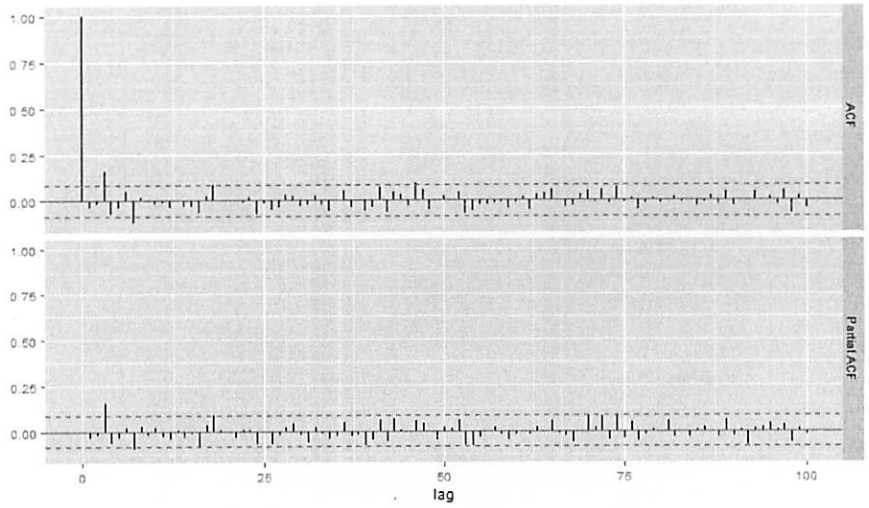


Figure 15: ACF and PACF assuming no correlation structure



Multiple SARIMA models are tested using the log price data and temperature data differencing once on the non-seasonal part and differencing once seasonally with a lag of 52.1774. The best model in terms of AIC values (we want smallest) and autocorrelation and partial autocorrelation plots is SARIMA (2, 1, 0) x (1, 1, 1)_52.1775. We again run into the same problems as with the USA data. The autocorrelation (partial) plots quickly go to zero (figure 16), but there is a slight correlation at lag 3 regardless of the model. This does not make any practical sense and due to the low correlation, it will be ignored. The normal QQ plot does not look great around the edges. It is particularly difficult to fit this model as most models tested either did not converge or resulted in an error. The residuals are very similar to figure 9 with higher residuals around 2009. Overall, this is the best model considering we are ignoring external factors affecting gas prices.

Figure 16: ACF and PACF for SARIMA (2, 1, 0) x (1, 1, 1)_52.1775



Accounting for the correlation structure in the errors, there is suggestive but not definitive evidence that temperature is to gas prices ($p=0.08$ from Wald Test using temperature as regression parameter assuming SARIMA $(2, 1, 0) \times (1, 1, 1)\_52.1775$ errors). We estimate that a one degree increase in maximum temperature is associated with a one cent increase in the median gas price. A more practical interpretation is that both temperature and gas prices are related to the time of year, but not necessarily each other. Note that these results only hold for Denver, CO. Other areas of the country might show a different relationship. Nonetheless, I find the above analysis is interesting.

Part III: Region Analysis

When I first moved to Mississippi, I was shocked by the low gas prices. I always assumed that gas prices were fairly consistent across the county, but Southern prices were considerably lower than in Oregon. For instance, last summer I rarely paid more than $3.05 a gallon in Mississippi. The tradeoff was that Mississippi is a horrible place to live in the summer due to heat and unrelenting humidity. The last part of this paper compares gas prices around the country. Where can you find cheap gas? And would you actually want to live there?

Below is a map of the areas I considered (they also represent places I would consider living in). The cities and/or areas selected are: Seattle, San Francisco, Denver, Minnesota, Chicago, Gulf Coast and Massachusetts. Data comes from the EIA and spans June 2003 to February 2014 and again we are looking at unleaded, regular gasoline. All seven areas show an overall pattern similar to the whole United States. There is a big drop in 2009, but overall gas prices have been increasing over the time period specified.

Figure 17: Map

In appendix II are color figures comparing the time series for the locations. From the data, I found the location of the highest gas prices per each day and created a ranking of the locations based on which place consistently had the highest price and so forth. The final order is below. We see that San Francisco has the highest price and the Gulf Coast has the lowest. Table 2 shows the yearly averages for the last couple of years. Consistently San Francisco and Seattle have the highest gas prices and Denver and the Gulf Coast have the lowest gas prices.

1) San Francisco, avg. price $3.1165
2) Seattle, avg. price $2.9428
3) Chicago, avg. price $2.9278
4) Massachusetts, avg. price $2.7715
5) Minnesota, avg. price $2.7129
6) Denver, avg. price $2.7007
7) Gulf Coast, avg. price $2.6510

Table 2: Yearly Averages

	2013	2012	2011
San Francisco	3.9012	4.0399	3.8318
Seattle	3.6835	3.8510	3.7244
Chicago	3.7526	3.8509	3.7409
Massachusetts	3.5370	3.6411	3.5277
Minnesota	3.4451	3.5205	3.5130
Denver	3.4085	3.4761	3.3780
Gulf Coast	3.2994	3.4155	3.3634

Based on the above observation analysis, the Gulf Coast and Denver are the best places to live if I anticipate driving a lot. That is good, since at least along the Gulf, everything is very far apart so everyone has to drive (public transportation is a foregin concept). On that note, San Francisco and Seattle have good public transportation systems, so there are alternatives to driving. Chicago, Massachusetts and Minnesota fall in the middle. The gas prices there are comparable to Oregon, but those places are very cold in the winter. Overall, I have to conclude that deciding on a place to live based on gas prices is probably not a good idea. However, it is interesting to see how gas prices vary across the country.

References & Data Sources

Denver, Colorado Forecast: Weather Underground (Maximum Temperature). The Weather Underground Inc. Archieved from original 5 June 2000 to 30 December 2013. Retrieved 28 Feb. 2014.

Doomsday Rule. In Wikipedia. Retrieved 10 March 2014.

Weekly Retail Gasoline & Diesel Prices. United States Energy Information Administration. Retrieved from: www.eia.gov

Appendix I

```
<load necessary packages and functions>
```

Part 1: USA

```
#load in USA gas price data
usa <- read.csv('USA.csv',header=F,skip=3,,nrow=1228,col.names=c('date','price'))

usa$date <- parse_date_time(usa$date,"mdy")
usa$month <- month(usa$date)
usa$year <- year(usa$date)

#gets number of weeks per year
wks.per.yr <- ddply(usa, "year", summarise,num.wks = sum(as.numeric(date>0)))
usa$week <-
c(1:20,rep(1:52,5),1:53,rep(1:52,4),1:53,rep(1:52,5),1:53,rep(1:52,4),1:53,1:52,1:8)

#--------------------log.price------------------#
#plot raw data
qplot(date,price,data=usa,geom='line',main='Entire United States')

#take log of data
usa$log.price <- log(usa$price)
qplot(date,log.price,data=usa,geom='line',main='log.price')

#look at seasonality
lo.fit <- loess(log.price~week,data=usa,na.action=na.exclude)
usa$seas.smooth <- fitted(lo.fit)
qplot(usa$date,usa$seas.smooth,geom='line',xlab='date',ylab='Seasonal
Smooth',main='Seasonality')


weekly <- ddply(usa,'week',summarise,avg.log.price=mean(log.price,na.rm=T))
qplot(week[1:52],avg.log.price,data=weekly,geom='line',main='Avg. log.price by
Week',xlab='Week',ylab='log.price')+
  coord_cartesian(xlim=c(0, 55)) +
  scale_x_continuous(breaks=seq(0, 55, 4))
qplot(weekly$week[1:52],exp(weekly$avg.log.price[1:52]),geom='line',main='Avg. Price
by Week',xlab='Week',ylab='log.price')+
  coord_cartesian(xlim=c(0, 55)) +
  scale_x_continuous(breaks=seq(0, 55, 4))

#look at monthly pattern
monthly <- ddply(usa, "month", summarise,avg.log.price = mean(log.price, na.rm =
TRUE))
qplot(month,avg.log.price,data=monthly,geom='line',main='Avg. log.price by
Month',xlab='week',ylab='log.price')+
  coord_cartesian(xlim=c(0, 12)) +
  scale_x_continuous(breaks=seq(0, 12, 2))
qplot(month,exp(avg.log.price),data=monthly,geom='line',main='Avg. Price by
Month',xlab='week',ylab='price')+
  coord_cartesian(xlim=c(0, 12)) +
  scale_x_continuous(breaks=seq(0, 12, 2))

#look at yearly pattern
yearly <- ddply(usa, "year", summarise, avg.log.price = mean(log.price, na.rm = TRUE))
yearly$transform <- exp(yearly$avg.log.price)
qplot(year,avg.log.price,data=yearly,geom='line',main='Avg. log.price by Year')

#subtract off seasonal pattern
usa$deseas <- usa$log.price-usa$seas.smooth
```

```
qplot(date,deseas,data=usa,geom='line',main='Deseasonalized')

#fit the trend
lo.fit.trend <- loess(deseas~as.numeric(date),data=usa,na.action=na.exclude,span=0.4)
usa$trend <- fitted(lo.fit.trend)
qplot(date,trend,data=usa,geom='line',ylab='Fitted Values',main='Trend over Time')

#subtract off trend
usa$resid <- usa$deseas-usa$trend

qplot(date,deseas,data=usa,geom='line')
qplot(date,resid,data=usa,geom='line')

#-------------------------------------------------------------#
source(url("http://stat565.cwick.co.nz/code/get_acf.R"))

usa$diff <- c(NA,diff(usa$log.price,lag=1))
usa$seas <- c(rep(NA,52),diff(usa$diff,lag=52.1775))
qplot(date,seas,data=usa,geom='line',main='Residuals',ylab='residuals')

examine_corr(usa$seas,lag.max=100,na.action=na.exclude)

fit1 <- arima(usa$log.price,order=c(2,1,0),seasonal=list(order=c(0,1,0),period=52.1775))
fit2 <- arima(usa$log.price,order=c(2,1,0),seasonal=list(order=c(1,1,0),period=52.1775))
fit3 <- arima(usa$log.price,order=c(2,1,0),seasonal=list(order=c(1,1,1),period=52.1775))
fit4 <- arima(usa$log.price,order=c(2,1,0),seasonal=list(order=c(1,0,1),period=52.1775))
fit5 <- arima(usa$log.price,order=c(0,1,0),seasonal=list(order=c(2,0,1),period=52.1775))
fit6 <- arima(usa$log.price,order=c(1,1,0),seasonal=list(order=c(2,0,1),period=52.1775))
fit7 <- arima(usa$log.price,order=c(1,1,0),seasonal=list(order=c(2,1,0),period=52.1775))
fit8 <- arima(usa$log.price,order=c(1,1,0),seasonal=list(order=c(1,1,0),period=52.1775))
fit9 <- arima(usa$log.price,order=c(2,1,0),seasonal=list(order=c(2,1,0),period=52.1775))
fit10 <- arima(usa$log.price,order=c(3,1,0),seasonal=list(order=c(1,1,0),period=52.1775))

examine_corr(fit3$resid,na.action=na.exclude,lag.max=100)#better

qqnorm(fit3$residuals)
qqline(fit3$residuals)

qplot(usa$date,fit3$resid,xlab='Date',ylab='Residuals',main='Residuals')

#-----------------------PREDICT-------------------------------#

pred <- as.data.frame(predict(fit3,n.ahead=8))
pred$transform <- exp(pred$pred)
pred$se.transform <- exp(pred$se)
pred$date <-
ymd(20140303,20140310,20140317,20140324,20140331,20140407,20140414,20140421)
qplot(date, price, data = usa, geom = "line",main='Price Predictions') +
  geom_ribbon(aes(ymin = exp(pred- 2*se), ymax = exp(pred + 2*se), y = NULL), data =
pred, alpha = 0.2) +
  geom_line(aes(y = exp(pred)), data = pred)
```

Part II: Regression with Temperature

```
#load in denver gas price data
denver <- read.csv('denver.csv',header=F,skip=3,,nrow=718,col.names=c('date','price'))
denver$date <- parse_date_time(denver$date,"mdy")
qplot(date,price,data=denver,geom='line',main='Denver Price Data')
denver$log.price <- log(denver$price)

#load in denver weather data
denver.temp <- read.csv('denver_temp.csv',header=F, skip=1,col.names=c('date','temp'))
denver.temp$date <- ymd(denver.temp$date)
```

```
#join gas data and weather data using gas data as reference
data <- join(denver,denver.temp,by='date',type='left')

qplot(date, value, data = melt(data, id.vars = "date"), geom = "line") +
  facet_grid(variable ~ ., scale = "free")

#examine data
data$date <- ymd(data$date)
data$month <- month(data$date)

qplot(date,temp,data=data,main='Denver Temperature',geom='line')

monthly.price <- ddply(data,'month',summarise,avg.price=mean(price,na.rm=TRUE))
qplot(month,avg.price,data=monthly.price,,geom='line',main='Denver Avg. Price per
Month')+
  coord_cartesian(xlim=c(0, 12)) +
  scale_x_continuous(breaks=seq(0, 12, 2))

monthly <- ddply(data, "month", summarise,avg.temp = mean(temp, na.rm = TRUE))
qplot(month,avg.temp,data=monthly,geom='line',main='Avg. Temp by
Month',xlab='month',ylab='temp')+
  coord_cartesian(xlim=c(0, 12)) +
  scale_x_continuous(breaks=seq(0, 12, 2))

#create stationary time series
data$temp.diff <- c(NA,diff(data$temp,lag=1))
data$temp.diff2 <- c(rep(NA,52),diff(data$temp.diff,lag=52.1775))
qplot(date,temp.diff2,data=data,geom='line',xlab='temp',main='Stationary Temperature')

data$price.diff <- c(NA,diff(data$log.price,lag=1))
data$price.diff2 <- c(rep(NA,52),diff(data$price.diff,lag=52.1775))
qplot(date,price.diff2,data=data,geom='line',xlab='price',main='Stationary Price')

qplot(temp,price,data=data,main='Price v. Temp')
qplot(temp.diff2,exp(price.diff2),data=data,'Stationary Time Series Comparison',
      xlab='Temperature',ylab='Price')

mod <- lm(price.diff2~temp.diff2,data=data)
examine_corr(residuals(mod),lag.max=100,na.action=na.exclude)

#This suggests that we have both seasonal and non-seasonal AR parts

fit1 <- with(data,arima(log.price,xreg=temp,order=c(1,1,0),
                        seasonal=list(order=c(2,1,0),period=52.1775)))#-2717

fit2 <- with(data,arima(log.price,xreg=temp,order=c(1,1,0),
                        seasonal=list(order=c(1,1,1),period=52.1775)))#-2821

fit3 <- with(data,arima(log.price,xreg=temp,order=c(1,1,0),
                        seasonal=list(order=c(1,1,0),period=52.1775)))#-2656

fit4 <- with(data,arima(log.price,xreg=temp,order=c(2,1,0),
                        seasonal=list(order=c(1,1,1),period=52.1775)))#-2823!!!!!

examine_corr(residuals(fit4),lag.max=100,na.action=na.exclude)

#I choose fit4 (although it doesn't look great)
qqnorm(residuals(fit4))
qqline(residuals(fit4))

qplot(data$date,fit4$resid,xlab='Date',ylab='Residuals',main='Residuals')

est <- exp(coef(fit4)["temp"])
```

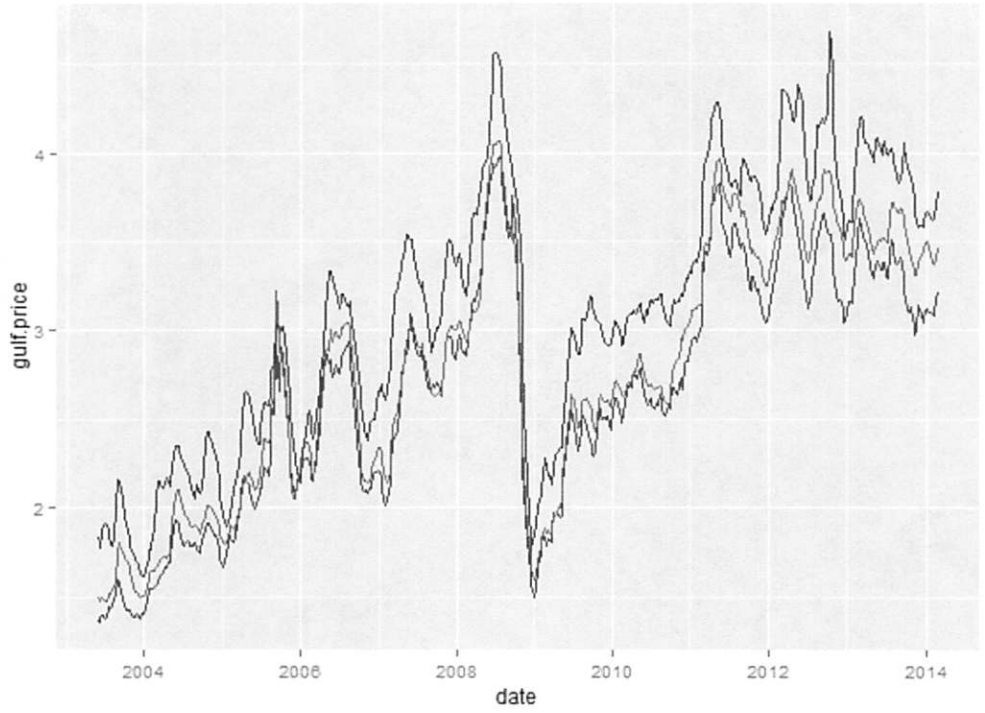
```
2*(1 - pnorm(abs( fit4$coef["temp"] / sqrt(diag(fit4$var.coef)["temp"]) )))  #p=0.1260
```

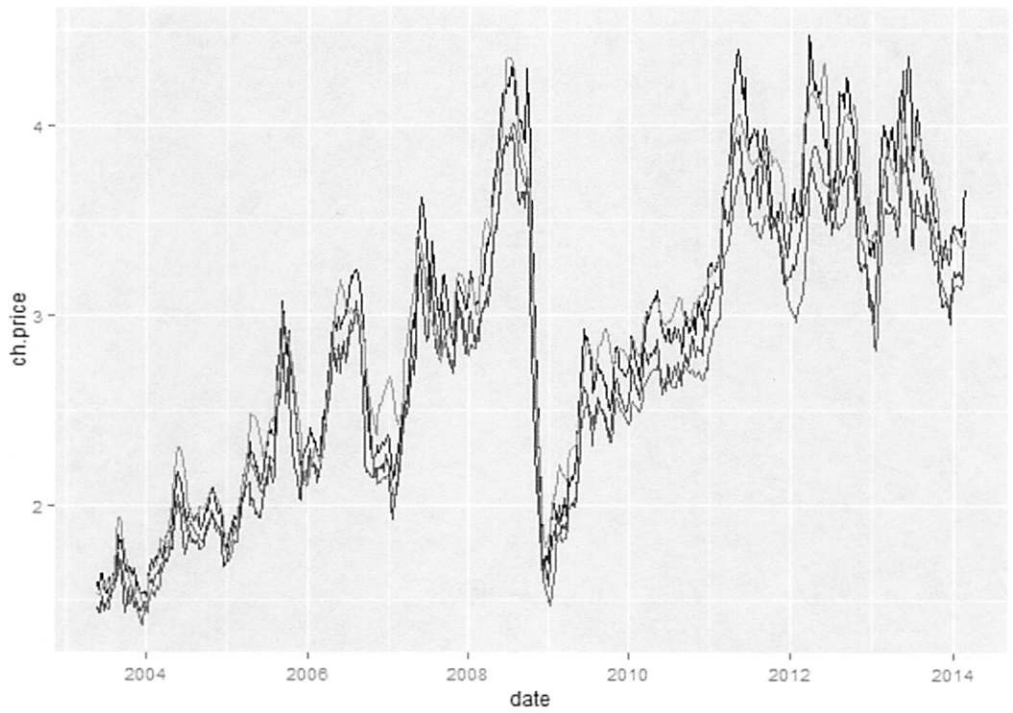
Part III: Regional Comparison

```
gulf <- read.csv('Gulf Coast.csv', skip=3,nrow=1138, header=F,
      col.names=c('date','gulf.price'))
mass <- read.csv('Massachusetts.csv', skip=3, nrow=562, header=F,
      col.names=c('date','ma.price'))
seattle <- read.csv('SEATTLE/Seattle.csv', skip=3, nrow=562, header=F,
      col.names=c('date','sea.price'))
denver <- read.csv('DENVER/Denver.csv', skip=3, nrow=718, header=F,
      col.names=c('date','den.price'))
minnesota <- read.csv('Minnesota.csv', skip=3, nrow=718, header=F,
      col.names=c('date','mn.price'))
sanfran <- read.csv('Sanfran.csv', skip=3, header=F, nrow=718,
      col.names=c('date','sf.price'))
chicago <- read.csv('Chicago.csv', skip=3, header=F, nrow=718,
      col.names=c('date','ch.price'))

gulf$date <- parse_date_time(gulf$date,"mdy")
gulf$year <- year(gulf$date)
mass$date <- parse_date_time(mass$date,"mdy")
mass$year <- year(mass$date)
seattle$date <- parse_date_time(seattle$date,"mdy")
seattle$year <- year(seattle$date)
denver$date <- parse_date_time(denver$date,"mdy")
denver$year <- year(denver$date)
minnesota$date <- parse_date_time(minnesota$date,"mdy")
minnesota$year <- year(minnesota$date)
sanfran$date <- parse_date_time(sanfran$date,"mdy")
sanfran$year <- year(sanfran$date)
chicago$date <- parse_date_time(chicago$date,"mdy")
chicago$year <- year(chicago$date)

price1 <- join(mass,gulf,by='date',type='left')
price2 <- join(price1,seattle,by='date',type='left')
price3 <- join(price2,denver,by='date',type='left')
price4 <- join(price3,minnesota,by='date',type='left')
price5 <- join(price4,sanfran,by='date',type='left')
price <- join(price5,chicago,by='date',type='left')

rm(price1,price2,price3,price4,price5)

gulf.yr <- ddply(gulf, "year", summarise, avg.price = mean(gulf.price, na.rm = TRUE))
mass.yr <- ddply(mass, "year", summarise, avg.price = mean(ma.price, na.rm = TRUE))
sea.yr  <- ddply(seattle,"year", summarise, avg.price = mean(sea.price, na.rm = TRUE))
den.yr  <- ddply(denver, "year", summarise, avg.price = mean(den.price, na.rm = TRUE))
mn.yr   <- ddply(minnesota,"year",summarise, avg.price = mean(mn.price, na.rm = TRUE))
sf.yr   <- ddply(sanfran, "year", summarise, avg.price = mean(sf.price, na.rm = TRUE))
ch.yr   <- ddply(chicago, "year", summarise, avg.price = mean(ch.price, na.rm = TRUE))

a <- data.matrix(price)
rownames(a) <- NULL
colnames(a) <- NULL
a <- a[,-1]
head(a)
#----- order 1-MA, 2-GULF, 3-SEA, 4-DEN, 5-MN, 6-SF, 7-CH -----------#
matrix <- t(a)
head(matrix)
cost <- rep(NA,562)
cheap <- rep(NA,562)
for (i in 1:562){
  cost[i] <- which.max(matrix[,i])
```

```
  cheap[i] <- which.min(matrix[,i])
}

table(cost)
table(cheap)

#------- Order 1-MA, 2-SEA, 3-DEN, 4-MN, 5-CH ----------------------#
mat2 <- matrix[-6,] #remove the highest Sanfran
mat2 <- mat2[-2,]   #remove the lowest Gulf
cost2 <- rep(NA,562)
cheap2 <- rep(NA,562)
for (i in 1:562){
  cost2[i] <- which.max(mat2[,i])
  cheap2[i] <- which.min(mat2[,i])
}

table(cost2)
table(cheap2)

#----------- order 1-MA, 2-MN, 3-CH ------------------------------#
mat3 <- mat2[-3,]  #remove the next highest Seattle
mat3 <- mat3[-2,]  #remove the next lowest Denver
cost3 <- rep(NA,562)
cheap3 <- rep(NA,562)
for (i in 1:562){
  cost3[i] <- which.max(mat3[,i])
  cheap3[i] <- which.min(mat3[,i])
}
table(cost3)
table(cheap3)

#FINAL ORDER of CHEAP to EXPENSIVE: GULF DEN MN MASS CH SEA SF

#plots

ggplot()+
  geom_line(data=price,aes(date,y=gulf.price,color='Gulf Coast'))+
  geom_line(data=price,aes(date,y=ma.price,color='Massachusetts'))+
  geom_line(data=price,aes(date,y=sf.price,color='San Francisco'))+
  scale_colour_manual("",
                      breaks = c("Gulf Coast", "Massachusetts","San Francisco" ),
                      values = c("blue", "red", "darkgreen"))


ggplot()+
  geom_line(data=price,aes(date,y=ch.price,color="Chicago"))+
  geom_line(data=price,aes(date,y=den.price,color="Denver"))+
  geom_line(data=price,aes(date,y=mn.price,color="Minnesota"))+
  geom_line(data=price,aes(date,y=sea.price,color="Seattle"))+
  scale_colour_manual("",
                      breaks = c("Chicago", "Denver","Minnesota","Seattle"),
                      values = c("black", "tan4", "purple4","darkorange1"))
```

Appendix II





cross correlation might be interesting to look at.