ECON 5P04 Forecast Report
Natural Gas Consumption

Greg DeVilliers (5741251)
Zack Lansfield (5697693)

April 20, 2019

Table of Contents

1 Introduction

The following report utilizes natural gas consumption for the years 2010 to 2019 for the USA with consumption being reported in billions of cubic feet. Data was retrieved from FRED and originally collected by the U.S. Energy Information Administration (EIA).

2 Trend and Seasonality

A time series and seasonal plot were generated for the years 2010 to 2019 and reported in Figures 1 & 2. From the time series plot, a deterministic trend is observed. It's clear that consumption is increasing at a non-stochastic rate; this is further supported by the Augmented Dickey-Fuller Test reporting a value of -5.7808 (p-value < 0.01) and therefore rejecting the null hypothesis of a stochastic trend. This result is intuitive as we expect natural gas consumption to increase with increasing population and decrease with efficiency improvements, both of which are relatively non-stochastic trends in the United States.

The seasonality plot achieves a visualization of monthly (or seasonal) effects within the data. It's seen that there are repeating highs and lows at the same time period across the years and it is known that this can be captured via seasonal dummies. This is intuitive as well, natural gas is large source of energy in the USA with some primary uses being electricity generation and heating. In winter months, much more use can be expected as additional heating is required. In summer months, there would be be moderate use due to electricity consumption for things like air conditioning units. In between these, relatively low use is predicted. The seasonality plot confirms this hypothesis, but it is further investigated via ACF plots in the section that follows.

Figure 1: Time Series Plot

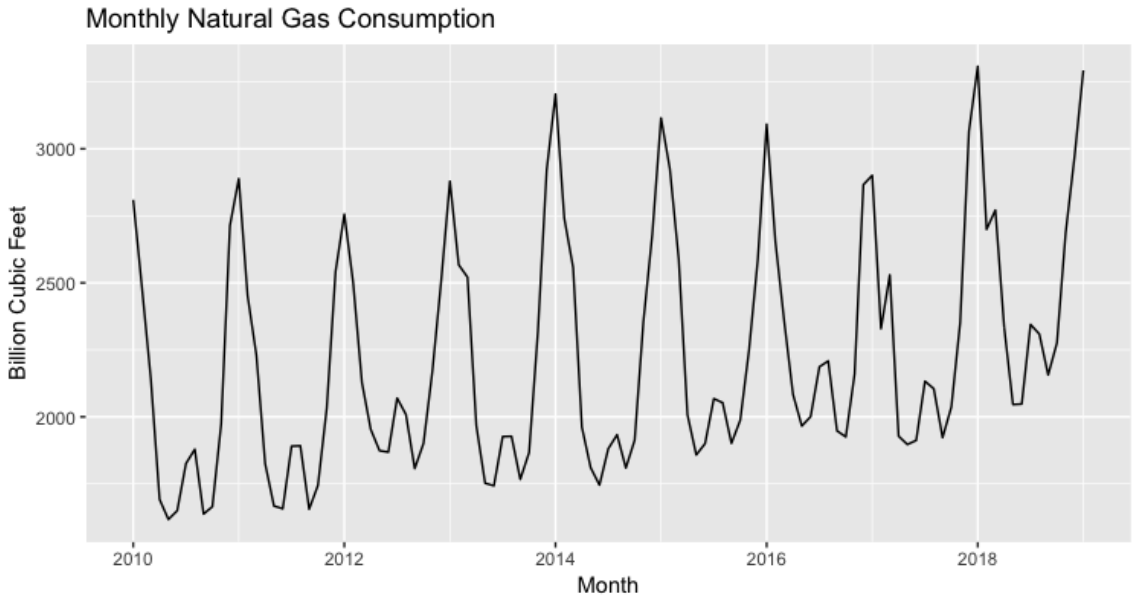
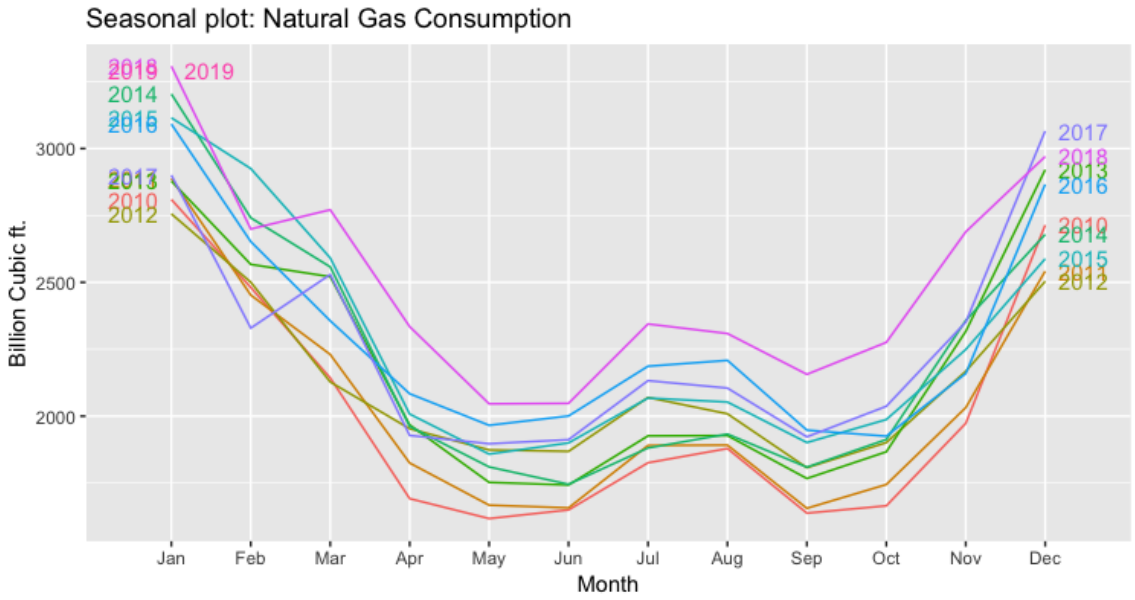


Figure 2: Seasonal Time Series Plot

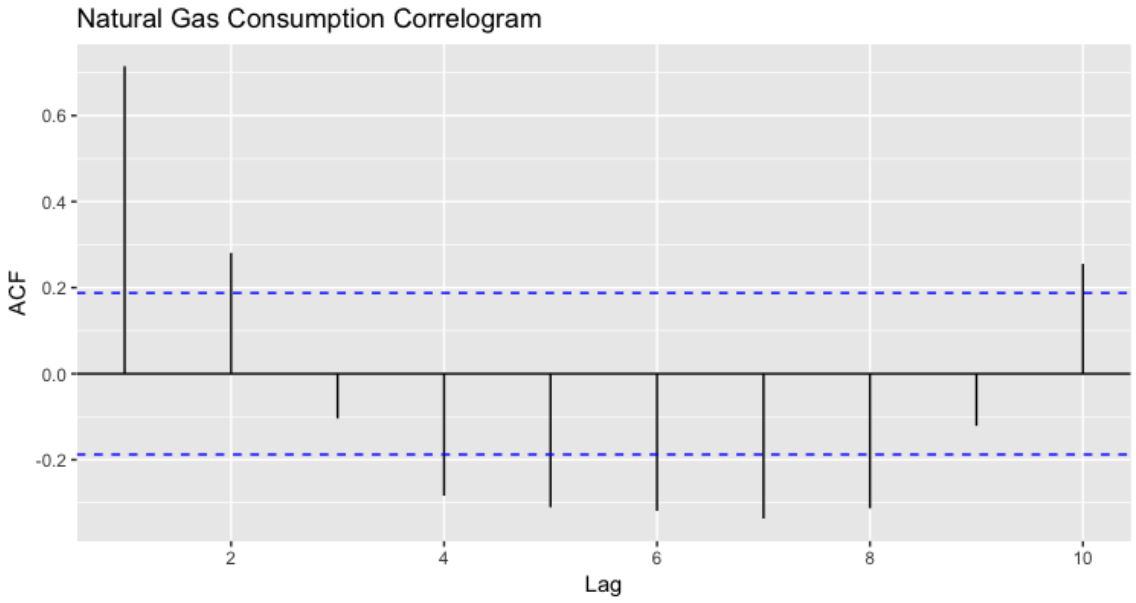


3 Autocorrelation

Autocorrelation values are reported in the table and visualized via correlogram in Figure 3. Due to high autocorrelation values following a seasonal trend as previously discussed, the data is not white noise. The correlogram acts as further evidence of the original hypothesis that each season has some unique effect on natural gas consumption. What the correlogram implies is essentially that months are highly positively correlated with the previous month, and negatively correlated with months 4-8 back. Using the example of January, this would be consumption being positively correlated to (or similar to) the energy in December, but negatively correlated to the consumption in May to September. Again, this would show consumption is similar among winter months, but opposite to consumption among summer months (and vice versa).

Figure 3: Correlogram and Autocorrelation Values

Lag	1	2	3	4	5	6	7	8	9	10
Correlation	0.715	0.281	-0.104	-0.284	-0.310	-0.319	-0.337	-0.313	-0.121	0.255



4 Forecasting Methods

Forecasts were generated using the average, naive, and seasonal naive methods with a training set ranging from 2010 to 2016. Accuracy of the three methods is reported in Table 1 results reported in Tables 2, 3, & 4. A visualization is provided in Figure 4.

The average and naive method provide constant forecasts across all months, not accounting for seasonality. The result of this is a poor, straight line estimate based off the average or last observed value. The data shows clear signs of seasonal trends, and so neither of these methods are ideal. The seasonal naive method provides forecasts based on the last observed value for the month in question. The result is a replication of each month's consumption value substituted for each future month. This does not capture the overall upward trend we see across years, but at least captures the seasonal trend within them.

The measures of accuracy used are all some form of error measurement. It is best to keep these as low as possible. In general, the method with the lowest measures of error should be selected. In Table 1, it can be seen that the naive method performs better than the average method in the training set, but does worse than the average method in the test set. Both of these methods do poorly compared to the seasonal naive method, which reports errors about 50% of the next best in each scenario. This is to be expected as it has already been determined that seasonality exists in the data, and this is the only method of the three used that captures it.

Table 1: Measures of Accuracy

		Average	Naive	Seasonal Naive
RMSE	Training	411.07	287.95	136.53
	Test	473.75	465.41	226.31
MAE	Training	350.49	232.20	105.99
	Test	341.48	422.67	180.41
MAPE	Training	16.173	10.544	4.7688
	Test	12.855	18.882	7.3379
MASE	Training	3.3068	2.1907	1.0000
	Test	3.2217	3.9877	1.7021

Table 2: Average Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2
2017	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2
2018	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2	2140.2
2019	2140.2											

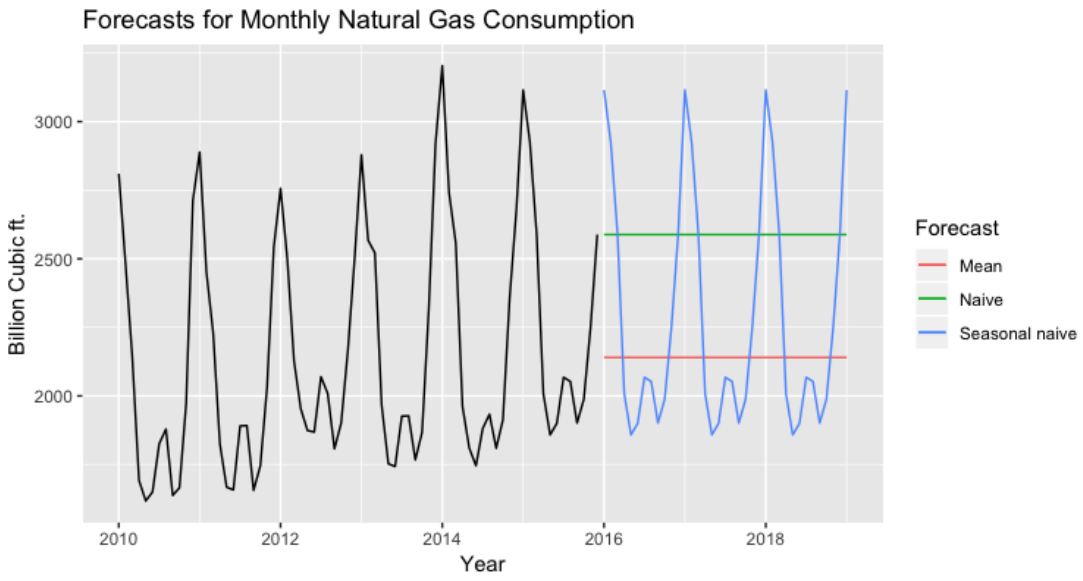
Table 3: Naive Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2
2017	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2
2018	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2	2588.2
2019	2588.2											

Table 4: Seasonal Naive Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	3115.0	2925.2	2591.3	2007.9	1858.1	1899.9	2067.7	2052.7	1901.3	1987.3	2249.1	2588.2
2017	3115.0	2925.2	2591.3	2007.9	1858.1	1899.9	2067.7	2052.7	1901.3	1987.3	2249.1	2588.2
2018	3115.0	2925.2	2591.3	2007.9	1858.1	1899.9	2067.7	2052.7	1901.3	1987.3	2249.1	2588.2
2019	3115.0											

Figure 4: Forecasts



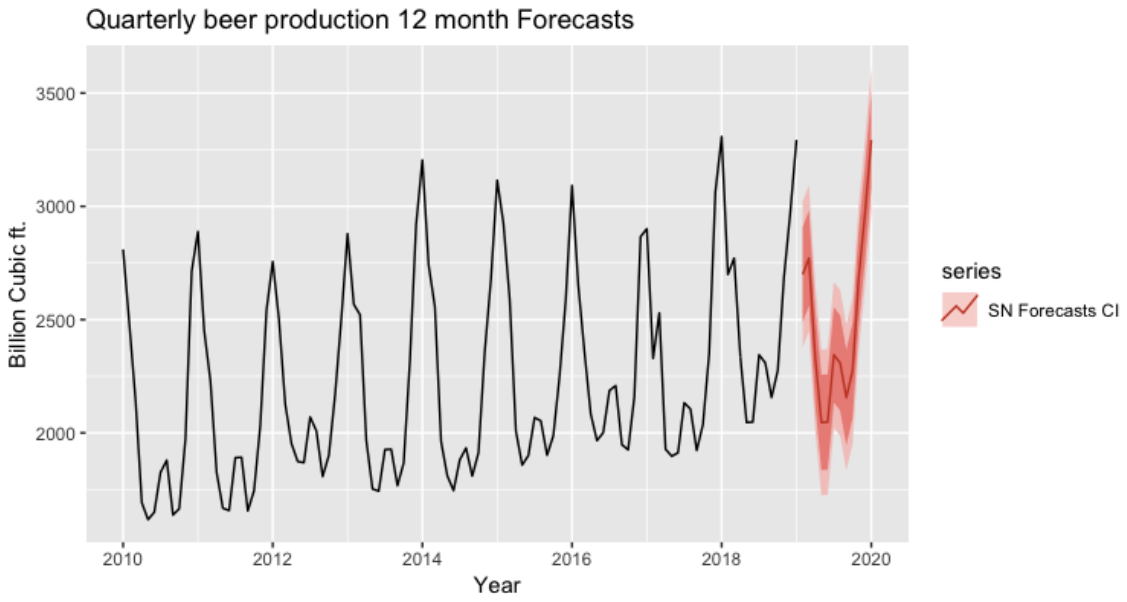
5 Future Forecasts

Based on the measures of accuracy and seasonality present in the data, the seasonal naive forecasting method is recommended. A 12 month forecast is found in Figure 5.

The 80% and 95% confidence intervals are shown by the light and dark red areas around the point forecasts. The method does a good job of capturing seasonality but fails to capture the upward trend. Regardless, this method was proven to be the best of the three methods on the basis of accuracy and it can be expected to capture consumption well outside of the annual upward trend.

Figure 5: Feb 2019 - Jan 2020 Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2019	NA	2699.4	2771.3	2334.6	2046.1	2047.9	2344.5	2308.8	2156.3	2275.8	2688.9	2970.3
2020	3292.8											



6 Regression-based Forecast Analysis

The trend, seasonal dummies, and trend with seasonal dummies regressions are reported in Tables 6, 7, & 8. Residual plots are reported in Figures 6, 7, & 8. Note that the season dummies refer to months of the year and that season1 (January) was dropped to avoid perfect multicollinearity.

The regression summaries reveal that a trend-only regression is insufficient. It reports a low F statistic and has poor goodness of fit (R^2= 0.029). Additionally, the trend on its own is shown to be insignificant via a high p-value (0.152). This is improved upon in the seasonal dummies regression. This regression has a large F statistic (50.9), high R^2 (0.90), and, given their p-values, all variables are significant on their own. Including both the trend and seasonal dummies yields the best results, achieving the highest F statistic (101.6) and R^2 (0.95).

Analyzing the residuals tells a similar story. In the case of trend-only regression, the residuals in Fig. 6 show a pattern of high peaks at the beginning/end of every year with smaller peaks around the middle of the year. This was previously discussed and revealed to be a seasonal effect. This is known to be true due to that pattern disappearing in Fig. 7, the seasonal dummies regression. The residuals of this regression are still problematic as there is a clear upward trend. This is solved in the third iteration utilizing both trend and seasonal dummy variables; Fig. 8 shows no clear pattern and is primarily white noise. Furthermore, this regression shows the least autocorrelation among residuals. Although none of the three regressions have perfectly normal residuals, the third attempt is clearly the best.

7 Coefficient of Variation

The coefficient of variation is, by definition, the ratio of standard deviation to the mean. The values are reported in Table 5. Since lower here is better, the regression with both trend and seasonal dummy variables is the best of the three options. The difference between the trend-only and seasonal dummy regressions is extremely large here, revealing that a model which doesn't consider seasonality fails by a large margin.

Table 5: Coefficients of Variation

	Trend	Season	Trend & Season
CV	173737.87	23552.72	11544.41

Table 6: Trend Only Regression

Coefficient	Estimate	Std. Error	t value	p-value
Intercept	2017.127	97.843	20.616	<2e-16
Trend	3.372	2.329	1.448	0.152
R^2= 0.02907		F= 2.096		

Table 7: Seasonal Dummies Only Regression

Coefficient	Estimate	Std. Error	t value	p-value
Intercept	2942.08	57.19	51.440	<2e-16
season2	-330.80	80.89	-4.090	0.000131
season3	-580.17	80.89	-7.173	1.27e-09
season4	-1040.92	80.89	-12.869	<2e-16
season5	-1178.87	80.89	-14.575	<2e-16
season6	-1181.52	80.89	-14.607	<2e-16
season7	-998.57	80.89	-12.345	<2e-16
season8	-993.30	80.89	-12.280	<2e-16
season9	-1179.10	80.89	-14.577	<2e-16
season10	-1095.85	80.89	-13.548	<2e-16
season11	-759.33	80.89	-9.388	2.21e-13
season12	-284.07	80.89	-3.512	0.000852
R^2= 0.9032		F= 50.9		

Table 8: Trend with Seasonal Dummies Regression

Coefficient	Estimate	Std. Error	t value	p-value
Intercept	2802.158	43.459	64.477	< 2e-16
trend	4.514	0.561	8.046	4.58e-11
season2	-335.314	56.328	-5.953	1.55e-07
season3	-589.194	56.336	-10.458	4.67e-15
season4	-1054.458	56.350	-18.713	< 2e-16
season5	-1196.921	56.370	-21.233	< 2e-16
season6	-1204.085	56.395	-21.351	< 2e-16
season7	-1025.649	56.426	-18.177	< 2e-16
season8	-1024.896	56.462	-18.152	< 2e-16
season9	-1215.210	56.504	-21.507	< 2e-16
season10	-1136.473	56.551	-20.096	< 2e-16
season11	-804.470	56.604	-14.212	< 2e-16
season12	-333.717	56.662	-5.890	1.97e-07
R^2= 0.9538		F= 101.6		

Figure 6: Trend Only Regression Residuals

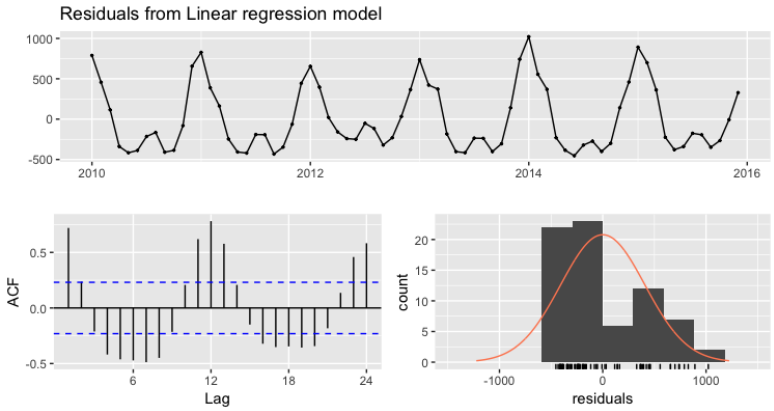


Figure 7: Seasonal Dummies Regression Residuals

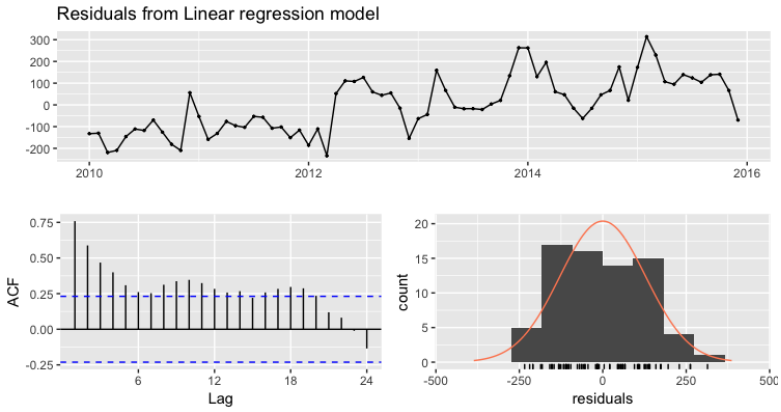
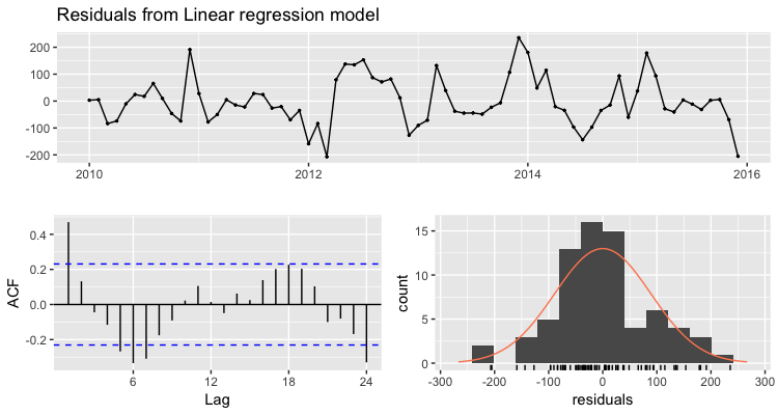


Figure 8: Trend & Seasonal Dummies Regression



8 Test Set Regression Forecasts

The forecasts of the three regression methods over the test set are reported in Figures 9, 10, & 11.

The time series plots provide a good visual explanation as to why a trend-only regression does so poorly. On average, the distance between the data and the fitted line is massive. The seasonal dummy regression does a far better job, but as time goes on the average distance between the fitted and actual lined becomes larger due to the uncaptured upward trend. This is solved by combining both; in this scenario the fitted and actual lines trace each other perfectly at some points with minimal difference elsewhere.

Figure 9: Trend Regression Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	2263.3	2266.7	2270.0	2273.4	2276.8	2280.2	2283.5	2286.9	2290.3	2293.6	2297.0	2300.4
2017	2303.8	2307.1	2310.5	2313.9	2317.2	2320.6	2324.0	2327.4	2330.7	2334.1	2337.5	2340.9
2018	2344.2	2347.6	2351.0	2354.3	2357.7	2361.1	2364.5	2367.8	2371.2	2374.6	2377.9	2381.3
2019	2384.7											

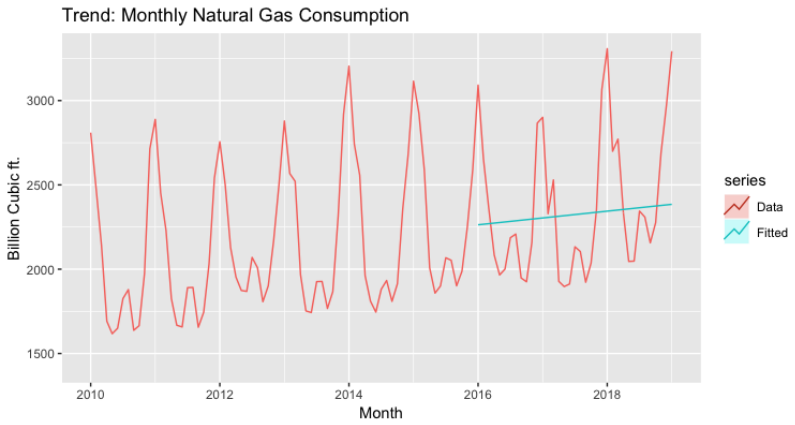


Figure 10: Seasonal Dummies Regression Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	2942.1	2611.3	2361.9	1901.2	1763.2	1760.6	1943.5	1948.8	1763.0	1846.2	2182.8	2658.0
2017	2942.1	2611.3	2361.9	1901.2	1763.2	1760.6	1943.5	1948.8	1763.0	1846.2	2182.8	2658.0
2018	2942.1	2611.3	2361.9	1901.2	1763.2	1760.6	1943.5	1948.8	1763.0	1846.2	2182.8	2658.0
2019	2942.1											

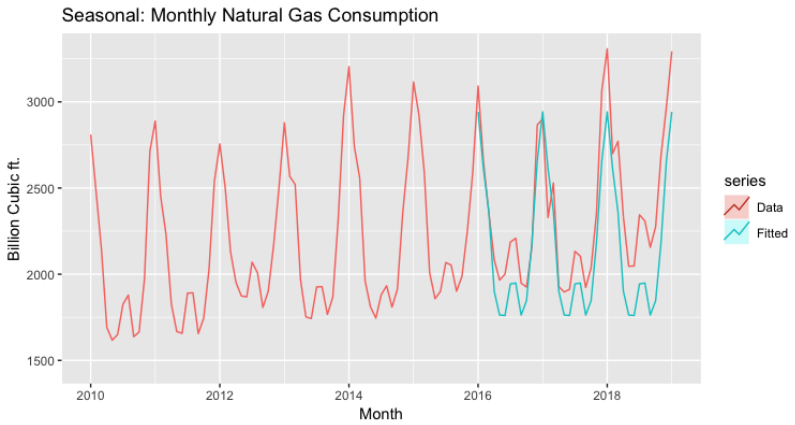
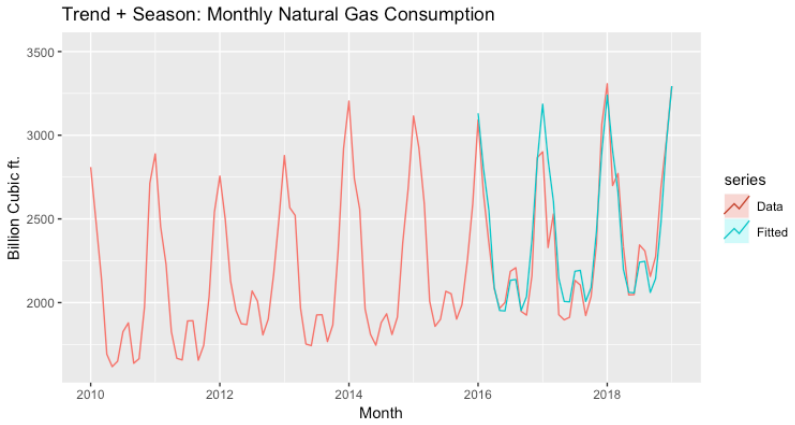


Figure 11: Trend & Seasonal Dummies Regression Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	3131.7	2800.9	2551.5	2090.7	1952.8	1950.1	2133.1	2138.4	1952.6	2035.8	2372.3	2847.6
2017	3185.8	2855	2605.7	2144.9	2007	2004.3	2187.3	2192.5	2006.7	2090	2426.5	2901.8
2018	3240	2909.2	2659.8	2199.1	2061.1	2058.5	2241.4	2246.7	2060.9	2144.1	2480.7	2955.9
2019	3294.2											



9 Regression Forecast Accuracy

Table 9 contains the RMSE and MAPE values for each of the three regression methods used. The results confirm the hypothesis that using both trend and seasonal dummy variables yields the best forecasts. Similar to the coefficient of variation case, the difference between the trend and seasonal dummies regressions is much larger than the difference between the seasonal dummies and trend & seasonal dummies regression. This is additional evidence that seasonality plays a large role for this data set, regardless of accuracy measure or forecasting method. The in-sample and out-of-sample results are the same: the trend with seasonal dummies performed the best, followed by seasonal dummies only, then trend only. Trend with seasonal dummies does better by a fair margin, but the trend only regression does awful job both in and out-of sample.

Table 9: Measures of Accuracy

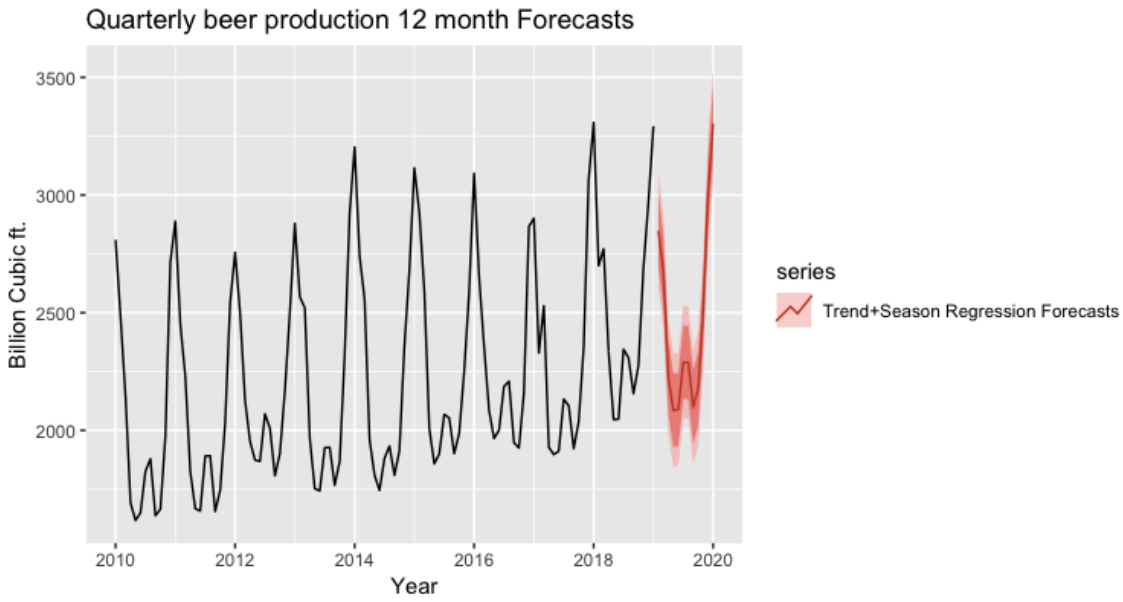
	Trend	Seasonal	Trend & Seasonal
RMSE	405.0496	127.8908	88.3128
MAPE	16.0269	5.011272	3.039428

10 Future Regression Forecasts

The trend with dummies model was chosen due to it showing the best measures of accuracy and highest R^2. The forecasts for the next 12 months not contained in the data set are reported in Figure 12. Over the next year, this model predicts the highest usage in January 2020 (something seen year after year in the original data) with a small peak in June and July. Elsewhere, the model predicts lower rates of natural gas consumption. Overall, there is an expected increase of natural gas consumption across all months with standard, expected seasonality.

Figure 12: Feb 2019 - Jan 2020 Forecasts

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2019	NA	2847.5	2678.6	2225.8	2085.2	2089.2	2289.3	2288.2	2098.1	2177.1	2508.3	3014.4
2020	3303.1											



A R Input

```
install.packages("ggplot2")
library(ggplot2)
install.packages("forecast")
library(forecast)
install.packages("fpp2")
library(fpp2)

ts_natural_gas<- ts(NATURALGAS$NATURALGAS, start=c(2010, 1), end=c(2019, 1),
                    frequency=12)

#Question 1--> Use full data set or just traning set?

#Using autoplot and ggseasonplot on full time period

autoplot(ts_natural_gas) + ggtitle("Monthly Natural Gas Consumption") +
  xlab("Month") + ylab("Billion Cubic Feet")

autoplot(ts_natural_gas) + ggtitle("Monthly Natural Gas Consumption") +
  xlab("Month") + ylab("Billion Cubic Feet")

ggseasonplot(ts_natural_gas, year.labels=TRUE, year.labels.left=TRUE) +
  ylab("Billion Cubic ft.") + ggtitle("Seasonal plot: Natural Gas Consumption")

#Question 2

ggAcf(ts_natural_gas, lag=10) + ggtitle("Natural Gas Consumption Correlogram") +
  xlab("Lag") + ylab("ACF")

gas_auto_corr<- ggAcf(ts_natural_gas, plot = FALSE)
gas_auto_corr_10<- gas_auto_corr[1:10,]

#Question 3
#Obtaining Forecasts
mean_fore<- meanf(training_gas, h=37)
mean_fore_list<- mean_fore$mean
naive_fore<- naive(training_gas, h=37)
naive_fore_list<- naivve_fore$mean
snaive_fore<- snaive(training_gas, h=37)
snaive_fore_list<- snaive_fore$mean

#Plotting Forecasts
training_gas <- window(ts_natural_gas,start=2010,end=c(2015,12))
autoplot(training_gas) + ggtitle("Quarterly beer production") +
```

```
  xlab("Year") + ylab("Megalitres")
# Plot some forecasts
autoplot(training_gas) +
  autolayer(meanf(training_gas, h=37)$mean, series="Mean") +
  autolayer(naive(training_gas, h=37)$mean, series="Naive") +
  autolayer(snaive(training_gas, h=37)$mean, series="Seasonal naive") +
  ggtitle("Forecasts for Monthly Natural Gas Consumption") +
  xlab("Year") + ylab("Billion Cubic ft.") +
  guides(colour=guide_legend(title="Forecast"))

#Measures of Accuarcy
gas_acc<- window(ts_natural_gas, start=2016)
accuracy(mean_fore, gas_acc)
accuracy(naive_fore, gas_acc)
accuracy(snaive_fore, gas_acc)

#Question 4
#Now using the full data set. Make forecasts for the next 12 months using
preferred method
#I chose to use snaive as we saw the 12th lag of each
observations is highlt correlated

snaive_fore_full<- snaive(ts_natural_gas, h=12)

autoplot(ts_natural_gas) + ggtitle("Quarterly beer production 12 month Forecasts") +
  xlab("Year") + ylab("Billion Cubic ft.")
# Plot some forecasts
autoplot(ts_natural_gas) +
  autolayer(snaive(ts_natural_gas, h=12), series="SN Forecasts CI")+
  ggtitle("Quarterly beer production 12 month Forecasts") +
  xlab("Year") + ylab("Billion Cubic ft.")

#Estimate 3 TS regressions: 1. trend only 2. Seasonal Only 3. Both

#Trend only
trend_only_reg<- tslm(training_gas ~ trend)
checkresiduals(trend_only_reg)
#Seasonal Dummies only
seasonal_only_reg<- tslm(training_gas ~ season)
checkresiduals(seasonal_only_reg)
#Tredn and Seasonal Regression
trend_season_reg<- tslm(training_gas ~ trend + season)
checkresiduals(trend_season_reg)

#Question 6. Use CV function to choose best model
```

```
model_check<- rbind(CV(trend_only_reg),CV(seasonal_only_reg),CV(trend_season_reg))

#Question 7: Forecasting over test set and plotting over all data
testing_gas <- window(ts_natural_gas,start=2016,end=c(2019,1))

#Trend only regression
trend_fore<- forecast(trend_only_reg, newdata = testing_gas)
autoplot(ts_natural_gas, series="Data") +
  autolayer(trend_fore, level = FALSE ,PI = TRUE ,series="Fitted") +
  xlab("Month") + ylab("Billion Cubic ft.") +
  ggtitle("Trend: Monthly Natural Gas Consumption")

#Season Only
season_fore<- forecast(seasonal_only_reg, newdata = testing_gas)
autoplot(ts_natural_gas, series="Data") +
  autolayer(season_fore, level = FALSE ,PI = TRUE ,series="Fitted") +
  xlab("Month") + ylab("Billion Cubic ft.") +
  ggtitle("Seasonal: Monthly Natural Gas Consumption")

#Trend and Season
trend_season_fore<- forecast(trend_season_reg, newdata = testing_gas)
autoplot(ts_natural_gas, series="Data") +
  autolayer(trend_season_fore, level = FALSE ,PI = TRUE ,series="Fitted") +
  xlab("Month") + ylab("Billion Cubic ft.") +
  ggtitle("Trend + Season: Monthly Natural Gas Consumption")

#Question 8

q8_validation<- rbind(accuracy(trend_fore, testing_gas),
accuracy(season_fore, testing_gas),
accuracy(trend_season_fore, testing_gas))
q8_validation

#Question 9: Using favorite regression: Make forecasts for 12 months.
Present in graph and table

full_tslm<- tslm(ts_natural_gas ~ trend + season)
full_tslm_fore<- forecast(full_tslm, h=12)
autoplot(ts_natural_gas) + ggtitle("Quarterly beer production 12 month Forecasts")+
  xlab("Year") + ylab("Billion Cubic ft.")
# Plot some forecasts
autoplot(ts_natural_gas) +
  autolayer(forecast(full_tslm, h=12), series="Trend+Season Regression Forecasts")+
  ggtitle("Quarterly beer production 12 month Forecasts") +
  xlab("Year") + ylab("Billion Cubic ft.")
```