Rutgers University

Multivariate Project

Driving a 2010 **Nissan Versa**

Author:
Immanuel Williams

Contents

December 21, 2013

0.1 Introduction

The purpose of this project is to utilize various statistical tools (primarily multivariate statistical methods) to analyze data regarding the gas I put in my 2010 Nissan Versa. The research questions for this project are as follows:

I) Is there a significant difference between gas stations?

II) Which gas station is the best for my car?

  1) By the amount of miles:

    1) In general (controlling for type of driving)

    2) On the highway vs. In the city

  2) Which gas station is cheaper?

    1) In general (controlling for region)

    2) Middlesex County (NJ) vs. Central Maryland

I will consider various methods that may answer questions that are not stated above but will be discussed in the conclusion of this paper. The following section will thoroughly describe each variable in the data set and discuss the assumptions used for this project; the next section will show all the descriptive statistics and plots for the data. The subsequent sections answer each question stated above. This paper concludes with a general remark about the gas pumped into my Nissan Versa.

0.2 Description of Data and Assumptions

0.2.1 Description of Data

Each time I pumped my gas I recorded the following variables:

1. Per Gallon: How much each gallon of gas cost.
2. Gallons: The number of gallons pumped into the car
3. Total: How much the gas cost in total (Per Gallon x Gallons)
4. Begin Miles: The starting mileage point when gas was pumped into car
5. End Miles: The next time I pumped gas
6. # of Miles: the number of miles between fills
7. Date: The day the gas was filled

8. # of days: days between when gas was pumped
9. Description: Where I was driving (majority: driving in MD or NJ / driving to MD or NJ)
10. Travel Destination: Dummy variable for Description
    (a) A: Driving to NJ
    (b) B: Driving in NJ
    (c) C: Driving to MD
    (d) D: Driving in MD
    (e) E: Driving to DE
    (f) F: Driving to Canada (from NJ)
    (g) G: Driving to NJ (from Canada)
    (h) H: Driving to VA (from NJ)
    (i) I: Driving to NC (from NJ)
    (j) J: Driving to VA (from NC)
    (k) K: Driving to MD (from VA)
11. Travel: Highway (1) vs City coded (0)
12. Destination: majority Jersey (1) vs MD coded (0)
13. Gas Station: the gas station where the car was filled
14. Gas Type: Dummy variable for Gas station
    (a) A: Shell
    (b) B: Exxon
    (c) C: Mobil
    (d) D: BP
    (e) E: Sunoco
    (f) F: Royal Farms
    (g) G: WAWA
    (h) H: Random

0.2.2 Assumptions

Each time I filled my car with gas, I considered it to be an independent event of the other times I filled my car. Although this is not true, for some statistical test to be done this assumption needs to be made. A big problem with this assumption is the idea of gas mixing; it is uninterpretable to determine which gas had what effect on the car. I tried to remedy this situation by waiting till the gas tank was practically empty, which is very dangerous.

Some of the variables and factors are discussed in these analyzes and some are ignored, this is done to better interpret the results and for potential research questions in the future. The number of gallons variable is used to verify the Nissan claims of miles per gallon even though there is colinearity between number of gallons and total cost. This may not be true due to the variance in gas station prices. Gas stations that I went to only once were ignored such as Royal Farms, WaWa and Delta. The times I went to North Carolina and Canada, the variable for location was ignored but all the other variables were included. I tried to make all my analysis about driving from/to/in New Jersey and Maryland. Number of Days is hard to interpret because I would fill up in Maryland or New Jersey and spend a couple days in the other state. This is hard to interpret but the analysis will be done so that reasonable results can be drawn. Also, since three receipts were lost before information could be recorded, that data is unavailable for this analysis.

0.3 Descriptive Statistics

	Destination						
		N	Mean	SD	Median	Min	Max
Gallon	Maryland	30	9.99	1.93	10.4555	5.4313	12.023
	New Jersey	38	10.01	1.46	10.4555	5.21	11.663
Number of days	Maryland	29	4.17	2.25	4	1	9
	New Jersey	38	5.58	3.44	5.5	0	16
Number of miles	Maryland	30	284.77	52.70	294	143	358
	New Jersey	36	263.42	47.98	263.5	118	367
Cost per Gallon	Maryland	30	3.43	0.17	3.39	3.059	3.72
	New Jersey	38	3.49	0.19	3.469	3.139	3.89

	Travel Destination						
		N	Mean	SD	Median	Min	Max
Gallon	Driving to NJ	17	9.96	1.67	10.44	6.17	11.66
	Driving in NJ	20	9.99	1.32	10.46	5.21	11.23
	Driving to MD	16	9.66	1.66	10.29	5.76	11.57
	Driving in MD	14	10.36	2.20	11.37	5.43	12.02
Number of days	Driving to NJ	17	3.59	2.43	3	0	9
	Driving in NJ	20	7.35	3.30	7	2	16
	Driving to MD	16	4.56	2.58	5	1	9
	Driving in MD	13	3.69	1.75	3	2	8
Number of miles	Driving to NJ	15	295.87	45.48	305	181	367
	Driving in NJ	20	238.85	36.00	248	118	280
	Driving to MD	16	302.38	28.03	310	240	334
	Driving in MD	14	264.64	66.87	278	143	358
Cost per Gallon	Driving to NJ	17	3.50	0.20	3.49	3.18	3.86
	Driving in NJ	20	3.47	0.20	3.43	3.14	3.89
	Driving to MD	16	3.43	0.16	3.39	3.06	3.70
	Driving in MD	14	3.43	0.17	3.39	3.13	3.72

Type of Gas							
		N	Mean	SD	Median	Min	Max
Gallon	Shell	13	9.92	1.86	10.44	5.76	12.02
	Exxon	24	10.09	1.66	10.71	5.21	11.57
	Mobil	5	10.88	0.94	11.07	9.59	11.81
	BP	19	9.67	1.71	10.07	6.17	11.69
	Sunoco	9	10.46	0.56	10.37	9.80	11.75
Number of miles	Shell	13	289.85	33.66	292	234	367
	Exxon	24	260.67	59.03	253.5	118	334
	Mobil	5	243.20	64.88	250	181	336
	BP	17	264.18	55.54	261	178	358
	Sunoco	9	282.44	34.69	278	220	332
Number of days	Shell	13	5	1.83	6	2	8
	Exxon	24	5.21	3.55	5	1	16
	Mobil	5	3	0.71	3	2	4
	BP	18	3.56	3.07	3	0	12
	Sunoco	9	6.22	3.53	7	1	12
Cost per Gallon	Shell	13	3.54	0.20	3.60	3.23	3.86
	Exxon	24	3.53	0.20	3.56	3.139	3.959
	Mobil	5	3.46	0.22	3.38	3.319	3.84
	BP	19	3.42	0.18	3.39	2.99	3.71
	Sunoco	9	3.37	0.23	3.39	3.059	3.72

City/Highway							
		N	Mean	SD	Median	Min	Max
Gallon	City	34	10.14	1.72	10.64	5.21	12.02
	Highway	42	9.65	1.69	10.29	5.76	11.66
Number of days	City	33	5.91	3.30	5.00	2	16
	Highway	42	3.93	2.77	3.50	0	12
Number of miles	City	34	249.47	51.71	251	118	358
	Highway	40	287.80	44.44	297	178	367
Cost per gallon	City	34	3.45	0.19	3.41	3.13	3.89
	Highway	42	3.47	0.21	3.48	2.99	3.96

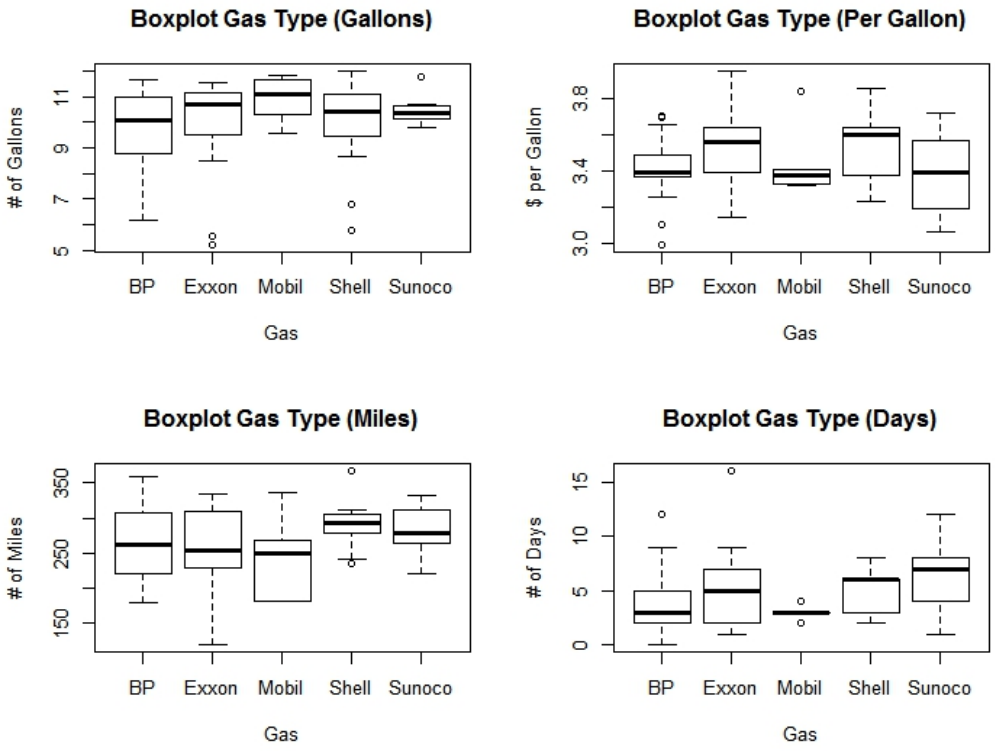

Figure 1: Figure 1. Boxplot by Gas Type

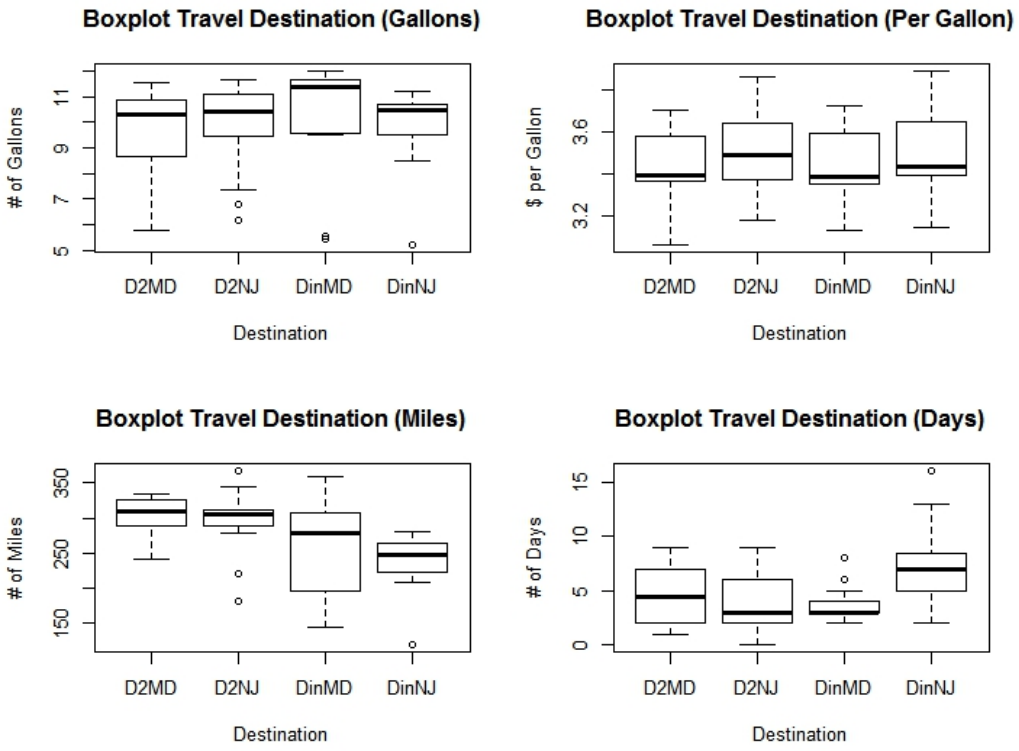

Figure 2: Figure 2. Boxplot by Travel Destination

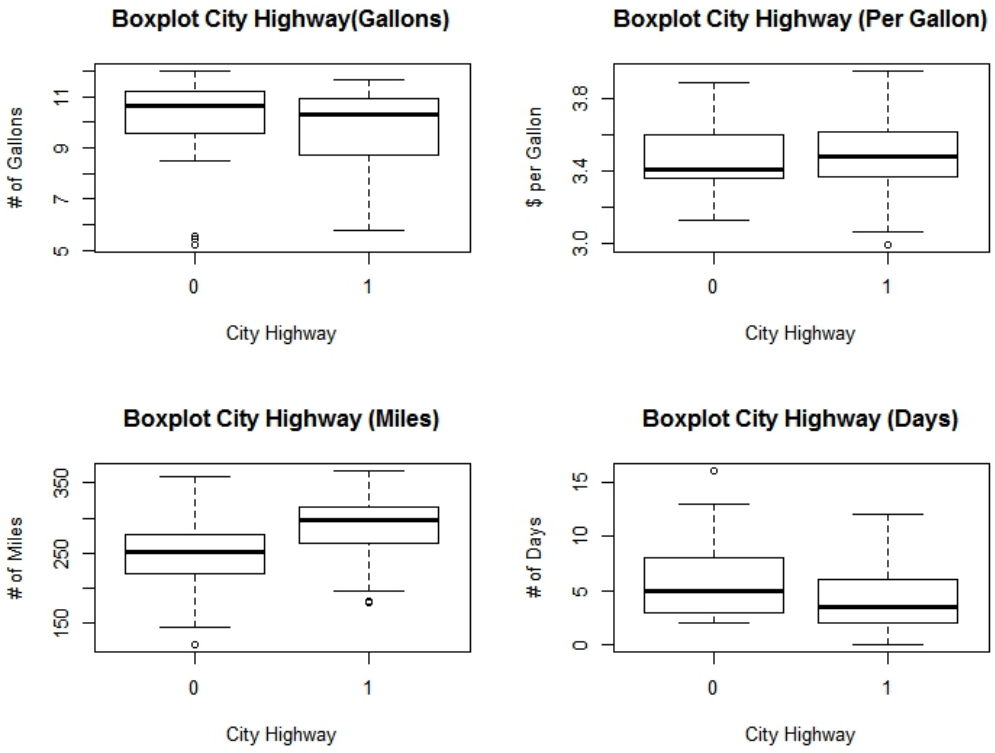

Figure 3: Figure 3. Boxplot by City Highway

0.4 Analysis

0.4.1 Is there a significant difference between gas stations?

There are many ways to answer the research questions for this paper. I will perform analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) to determine if there is a difference between the gas stations.The MANOVA method is used to determine if the gas station differ with respect to, the amount of gallons I put in my car, cost per gallon, the number of miles and the number of days.

```
## Method One: MANOVA
# Y1: Gallons, PerGallon, numofMiles, numofdays
Y1=cbind(cardata[,1],cardata[,2],cardata[,3],cardata[,4])
fit1=manova(Y1~factorD)
summary.manova(fit1)

          Df  Pillai  approx F num Df den Df Pr(>F)
factorD    7  0.76323 2.1896    28    260  0.0007886 ***
Residuals 65
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

This summary shows that there is a significant difference between gas stations with respect to number of gallons, number of days, number of miles and cost per gallon. This may not be conclusive, if we consider ANOVA we may get different results which will show if the gas stations with respect to each response variable differ. The ANOVA method is shown below:

```
## Method Two: Four seperate ANOVAs
fit1aa=aov(cardata[,1]~factorD) ## The # of Gallon
summary(fit1aa)
            Df Sum Sq Mean Sq F value Pr(>F)
factorD      7  48.41   6.916   2.744 0.0143 *
Residuals   68 171.37   2.520
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
### INT: There is a difference between
### gas stations with respect to # of gallons

fit1a=aov(cardata[,2]~factorD) ## Cost Per Gallon
```

```
summary(fit1a)
            Df Sum Sq Mean Sq F value Pr(>F)
factorD      7 0.3817 0.05453   1.408  0.217
Residuals   68 2.6345 0.03874
### INT: No difference between gas
### stations with respect to cost per gallon

fit1b=aov(cardata[,3]~factorD) ## Number of Miles
summary(fit1b)
            Df Sum Sq Mean Sq F value Pr(>F)
factorD      7  22194    3171   1.231  0.299
Residuals   66 170059    2577
2 observations deleted due to missingness
### INT: No difference between gas stations
### with respect to number of miles

fit1c=aov(cardata[,4]~factorD) ## Number of days
summary(fit1c)
            Df Sum Sq Mean Sq F value Pr(>F)
factorD      7    138  19.720    2.21 0.0442 *
Residuals   67    598   8.925
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
1 observation deleted due to missingness
### INT: Slight difference between
### gas stations with respect to number of days
```

ANOVA models shows that gas stations have a significant difference with respect to the number of days and the number of gallons pumped into the car, but does not have a significant difference with respect to the number of miles and cost per gallon. The difference between number gallons between gas stations is not meaningful to this data set because the number of gallons is not independent of each pump and the violation of the assumption above is need to make sense of these findings. The ANOVA results are not conclusive because it is controlling for region and type of driving which may have an effect on gas prices and the number of miles. Type of driving is synonymous to Travel in the description section meaning city vs. highway driving. In the next section we will not control for the region and type of driving to determine the best gas station, if the difference is notable, we can conclude that there is a significant difference, not controlling for region.

0.4.2 Which gas station is the best?

There are many characteristics which make up a good gas station such as cost, the effect of gas on the number of miles and how many days the gas lasts in the car. These factors are very subjective because everybody drives in a different way and has different preferences on what they consider to be good qualities in gas. In this paper we only consider the number of miles and cost per gallon. Coincidently, these are the same variable that did not show significant differences with respect to the ANOVA. We compare gas stations by region with respect to cost of per gallon and compare gas station by city/highway with respect to number of miles.

Which gas station is cheaper?

When controlling for region we use a regression model to determine which gas stations is cheap with respect to the cost of gas.

```
> ##       a) In general (controling for region)
> model3=lm(carbox[,2]~factorDd)
> summary(model3)

Call:
lm(formula = carbox[, 2] ~ factorDd)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4270 -0.1365 -0.0170  0.1030  0.4328

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.41700    0.04559  74.946   <2e-16 ***
factorDdExxon   0.10925    0.06103   1.790   0.0781 .
factorDdMobil   0.03860    0.09989   0.386   0.7004
factorDdShell   0.11869    0.07153   1.659   0.1019
factorDdSunoco -0.04967    0.08042  -0.618   0.5390
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1987 on 65 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.09857,Adjusted R-squared:  0.0431
F-statistic: 1.777 on 4 and 65 DF,  p-value: 0.1442
```

Although all the variables are not significant, some conclusions can still be drawn from this regression analysis. When controlling for region the order of the cheapest gas station cost of gas is:

1. Sunoco
2. BP
3. Mobil
4. Exxon
5. Shell

When not controlling for region we use a regression model to determine which gas stations is cheapest in Maryland and New Jersey with respect to the cost of gas.

```
> ##  b) New Jersey vs Maryland
> model4=lm(carbox[,2]~factorCc*factorDd)
> summary(model4)

Call:
lm(formula = carbox[, 2] ~ factorCc * factorDd)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39836 -0.08511  0.00387  0.12911  0.35264

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.50450    0.06952  50.410   <2e-16 ***
factorCc1                 -0.05725    0.09197  -0.623   0.5363
factorDdExxon             -0.04361    0.08975  -0.486   0.6291
factorDdMobil             -0.13483    0.12041  -1.120   0.2680
factorDdShell             -0.10730    0.10312  -1.041   0.3029
factorDdSunoco            -0.06975    0.10992  -0.635   0.5285
factorCc1:factorDdExxon    0.13372    0.11727   1.140   0.2594
factorCc1:factorDdMobil    0.01658    0.21708   0.076   0.9394
factorCc1:factorDdShell    0.28230    0.13373   2.111   0.0396 *
factorCc1:factorDdSunoco -0.14050    0.15152  -0.927   0.3581
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1703 on 52 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.2704,Adjusted R-squared:  0.1441
F-statistic: 2.141 on 9 and 52 DF,  p-value: 0.04212
```

Not controlling for region we find that the order for cheap gas stations in New Jersey and Maryland is:

1. Sunoco (NJ)
2. Mobil (MD)
3. Shell (MD)
4. Sunoco (MD)
5. BP (NJ)
6. Exxon (MD)
7. BP (MD)
8. Mobil (NJ)
9. Exxon (NJ)
10. Shell (NJ)

Overall Sunoco has the cheapest gas in NJ and the third cheapest in MD. Gas was typically more expensive in NJ than in MD. These differences are not significant so we cannot be conclusive about the results.

By the amount of miles:

When controlling for the type of driving used a regression model to determine if there is a difference between driving on the highway vs. in the city regarding the different gas stations with respect to the amount of miles obtained.

```
##   1) By the  amount of miles:
##       a) In general (controling for type of driving)
model1=lm(carbox[,3]~factorDd)
summary(model1)
Call:
```

```
lm(formula = carbox[, 3] ~ factorDd)
Residuals:
     Min       1Q   Median       3Q      Max
-142.667  -28.551   -1.011   36.750   93.824

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      264.18      12.58  20.995   <2e-16 ***
factorDdExxon     -3.51      16.45  -0.213    0.832
factorDdMobil    -20.98      26.39  -0.795    0.430
factorDdShell     25.67      19.11   1.343    0.184
factorDdSunoco    18.27      21.39   0.854    0.396
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 51.88 on 63 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.06948,Adjusted R-squared:  0.0104
F-statistic: 1.176 on 4 and 63 DF,  p-value: 0.33
```

Though all the variables are not significant, some conclusions can still be drawn from this regression analysis. The order below represents the gas station that provides the most mileage:

1. Shell
2. Sunoco
3. BP
4. Exxon
5. Mobil

When not controlling for the type of driving I used a regression model to determine if there is a difference between driving on the highway vs. in the city regarding the different gas stations with respect to the amount of miles obtained.

```
 ##       b) on the highway vs in the city?
model2=lm(carbox[,3]~factorBb*factorDd)
summary(model2)
```

```
Call:
lm(formula = carbox[, 3] ~ factorBb * factorDd)

Residuals:
     Min       1Q   Median       3Q      Max
-100.308  -19.170    2.091   20.670   93.000

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                272.70      14.06  19.401  < 2e-16 ***
factorBb1                  -20.70      21.90  -0.945  0.34857
factorDdExxon              -54.39      18.70  -2.909  0.00513 **
factorDdMobil              -16.70      29.26  -0.571  0.57037
factorDdShell               11.80      34.43   0.343  0.73304
factorDdSunoco             -2.20       26.30  -0.084  0.93361
factorBb1:factorDdExxon    113.12      28.48   3.971  0.00020 ***
factorBb1:factorDdMobil    -11.30      46.11  -0.245  0.80727
factorBb1:factorDdShell     27.02      40.59   0.666  0.50824
factorBb1:factorDdSunoco    42.20      37.00   1.141  0.25872
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 44.45 on 58 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.3712,Adjusted R-squared:  0.2736
F-statistic: 3.804 on 9 and 58 DF,  p-value: 0.0008037
```

The order of gas stations based on driving on the highway is:

1. Exxon
2. Sunoco
3. Shell
4. Mobil
5. BP

The order of gas stations based on driving in the city is:

1. Shell
2. BP

3. Sunoco
4. Mobil
5. Exxon

This regression model does not have significant difference, which shows that the mileage may not have a difference when you control for type of driving. However, when you do not control for this there is a significant difference between certain gas stations. After this analysis, my suggestion is to used Shell or Sunoco for my 2010 Nissan Versa.

0.5 Conclusion

Based on the results there are many conclusions that can be drawn about driving my Nissan Versa. I should use Shell or Sunoco gas in my car for many reasons even though Shell is expensive. Based on the data it seems as though I get more mileage using these gas stations. The Sunoco gas station is pretty much the cheapest gas station. Though the regression model did not show this, the descriptive data and boxplots showed that the number of days were greatest with Shell and Sunoco.

0.6 Further Research

There is a substantive amount of research that could be done from here. I could use various time series models to predict the outcome of gas based on the data. I could use more variables such as temperature, cost of living in each state, road conditions in both state and seasonal information. Once I collect various types of variables I would like to perform a principal component analysis to determine which variables has an effect on my efficacy of the gas I put in the car. I could also propose a method that remedies the idea of gas mixing.