Analysis of

Sairam Reddy Reddipalli

CONTENTS

I. Executive Summary

Prelude

Car Allowance Rebate System or Cash for Clunkers is an US government designed program intended to incentivize US citizens towards purchasing the new more fuel efficient vehicles by turning in their gas guzzlers or vehicles which consumed more gasoline. Incentive ranged from 3500 USD to 4500 USD. CARS formed a platform between US citizens and car dealers with citizens on the rebated side. The obtained Gas Guzzlers are scrapped so that the engines are un-usable. The program was started on July 1st 2009 and almost 700,000 transactions have been carried out till October 2009.

Challenge

The challenge is to visualize the findings about the density of transactions in various states, purchasing pattern and behavioral patterns of buyers towards various metrics like Mileage, Cost etc. across the United States.

II. Data Description

The data set consists of each successful final transaction processed or committed.
Total Features of the Data Set: 42
Total Transactions of the Data Set: 677,238
Number of States Involved in the Program: 55
Data Set Analyzed: Final Paid Transaction Database text file (via ftp)

Assumptions and Changes:

Since quite a few details were missing in the data description and time constraints, I had to make a few assumptions:
New Feature: 'mileage_diff', difference between trades in vehicle mileage and new vehicle mileage is added to each transaction.
New Feature: 'customer_cost', difference between new vehicle MSRP and incentive received is added to each transaction.
All NA values to the fields that are to be analyzed are either eliminated or made to 0.
NA values across features 'trade_in_mileage', 'new_vehicle_car_mileage', 'new_car_category' are eliminated.
All 'Unlisted' values in column 'new_vehicle_category' are treated as a single different type.

III. Research Questions

Question 1:

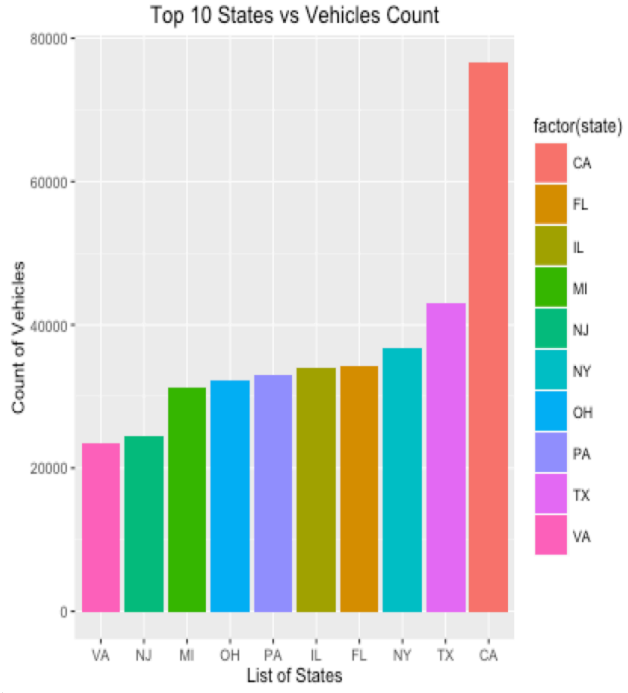
Define one or more metrics that can be used to measure the success of the program. Identify the 10 most successful states and the 10 least successful states based on your Metric(s), and show the performance in these 20 states.

Purpose:

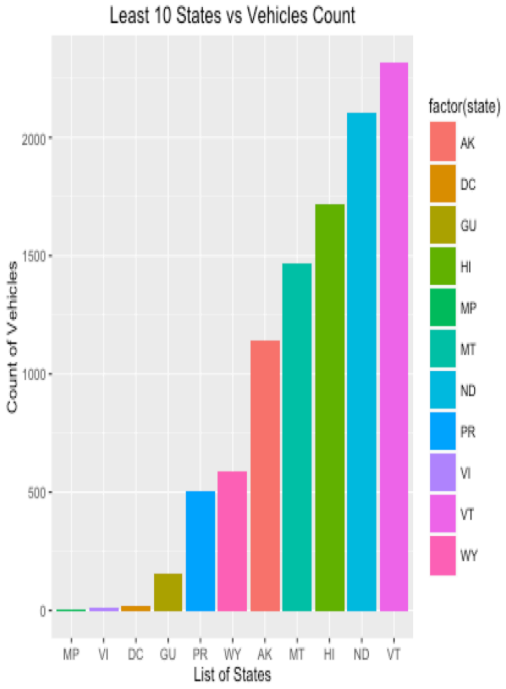
Number of Transactions: Number of successful transactions happened in each state will reveal the success of the campaign in each state on relative basis. Higher the transactions specify higher the gas-guzzlers dumped, sky rocketing success of the program.

Visualization 1:

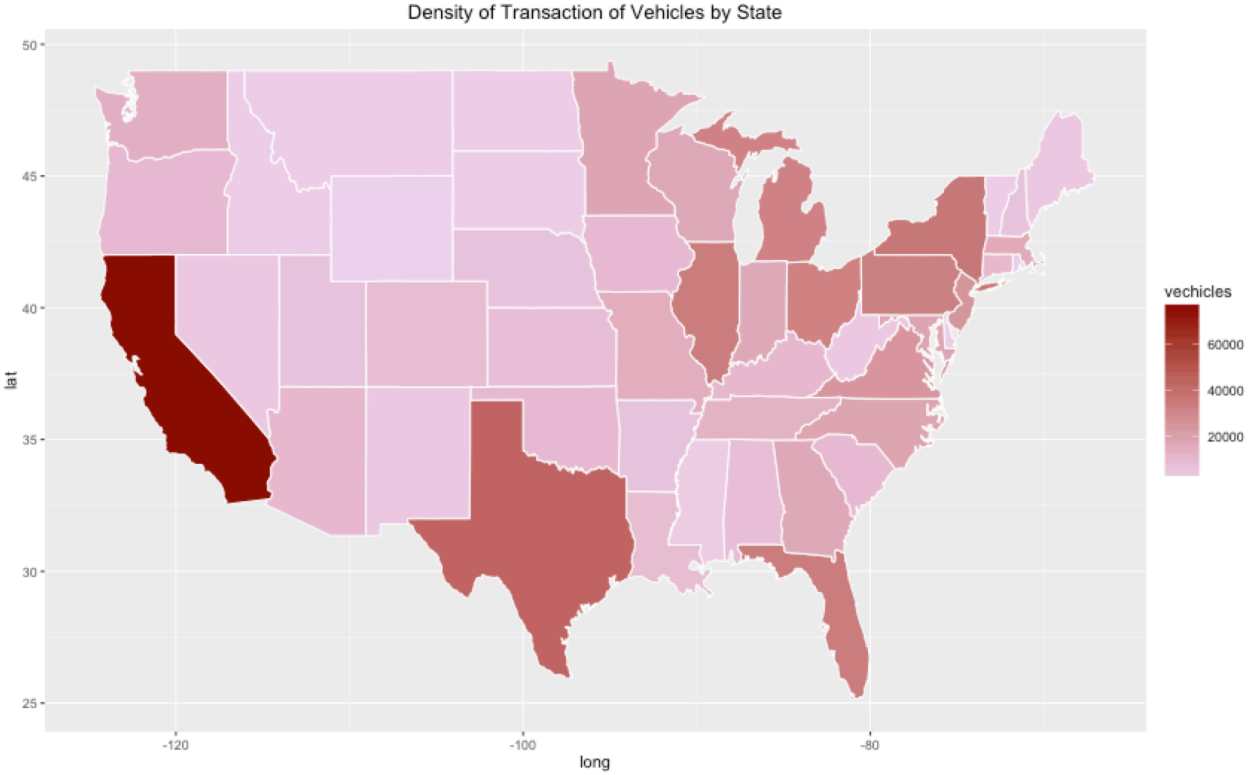
Visualization '1a' represents 10 states with highest transactions. We observe California has highest transactions by more than 30,000 to the next nearest state Texas. So, Californians utilized this program more than any other state citizens.



Visualization '1a'



Visualization '1b'



Visualization '1c'

Visualization '1b' represents 10 states with least transactions. As expected smaller states like DC, MP, VI have very less transactions while amazingly huge states like AK, WY haven't utilized the program. This might have to do with low density of population and low awareness.

Visualization '1c' represents all the states with darkness of the color indicating the number of transactions. Bigger states like California and Texas are amongst those with high transactions.

Limitations of Visualization 1: We don't know density of population of cars or people in the state, so it is not possible to compare success of program in smaller states like Maryland to California based on sheer number of transactions

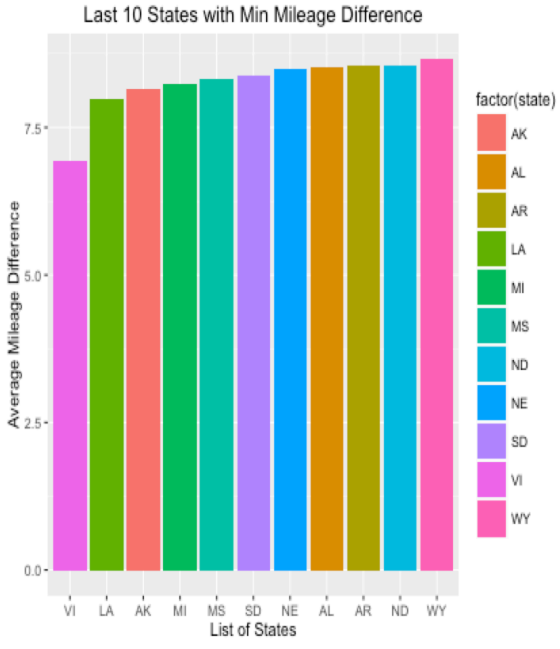
Average Mileage Difference in each State: It is the difference between mileage of new vehicle purchased and mileage of trade in vehicle. Greater the difference, greater is the success of the campaign.

Visualization 2:

Visualization '2a' represents 10 states with mean maximum mileage difference (mileage of trade in – new vehicle). We observe California has highest mean mileage difference by more than 10.2, followed by Oregon, Utah and Arizona. So, Californians utilized this program more than any other state citizens by giving away less fuel-efficient vehicles and purchasing more fuel-efficient vehicles.

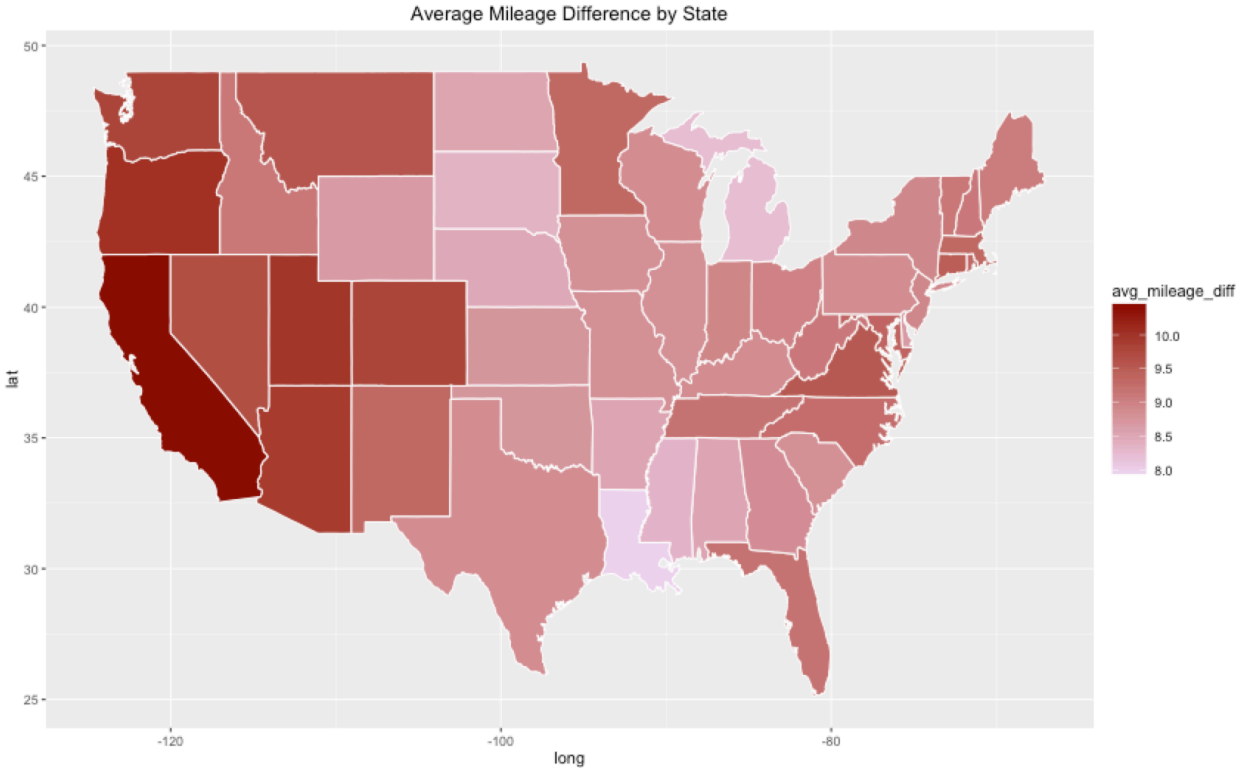
Visualization '2b' represents 10 states with mean minimum mileage difference (mileage of trade in – new vehicle). We observe Virgin Islands has least mean mileage difference by less than 8.5, followed by Louisiana, Arkansas and Michigan. So, Virgin Islands having very less transactions, Louisiana, Arkansas, Michigan utilized this program least effectively than any other state citizens.

Visualization '2a'



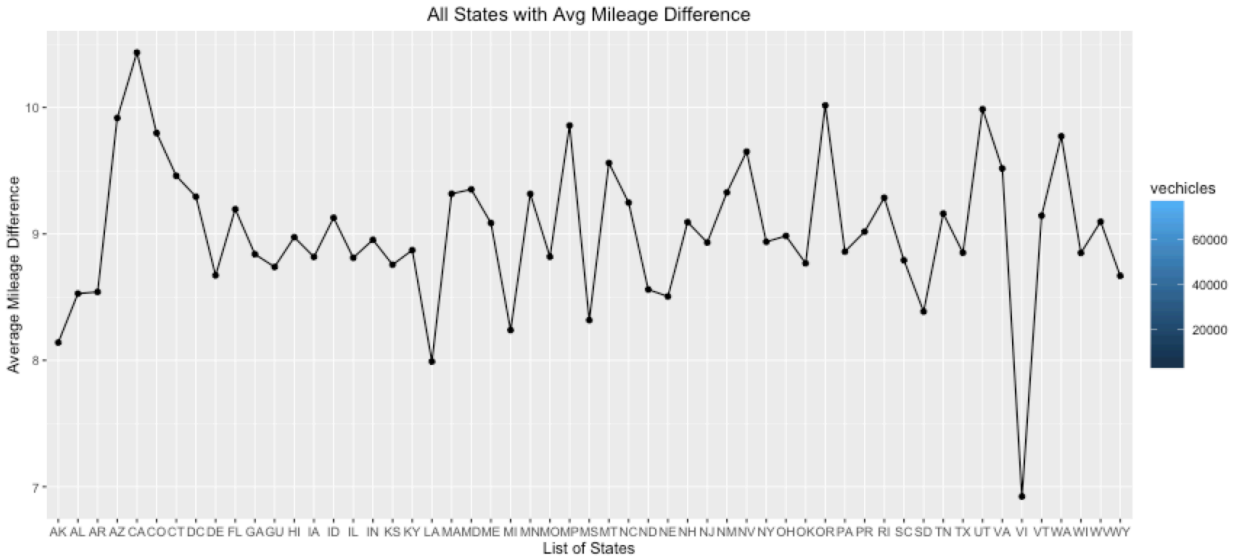
Visualization '2b'

Visualization '2c' represents all the states with darkness of the color indicating the maximum mean mileage difference. We see west coast people along with Utah, Arizona have bought more fuel-efficient vehicles.



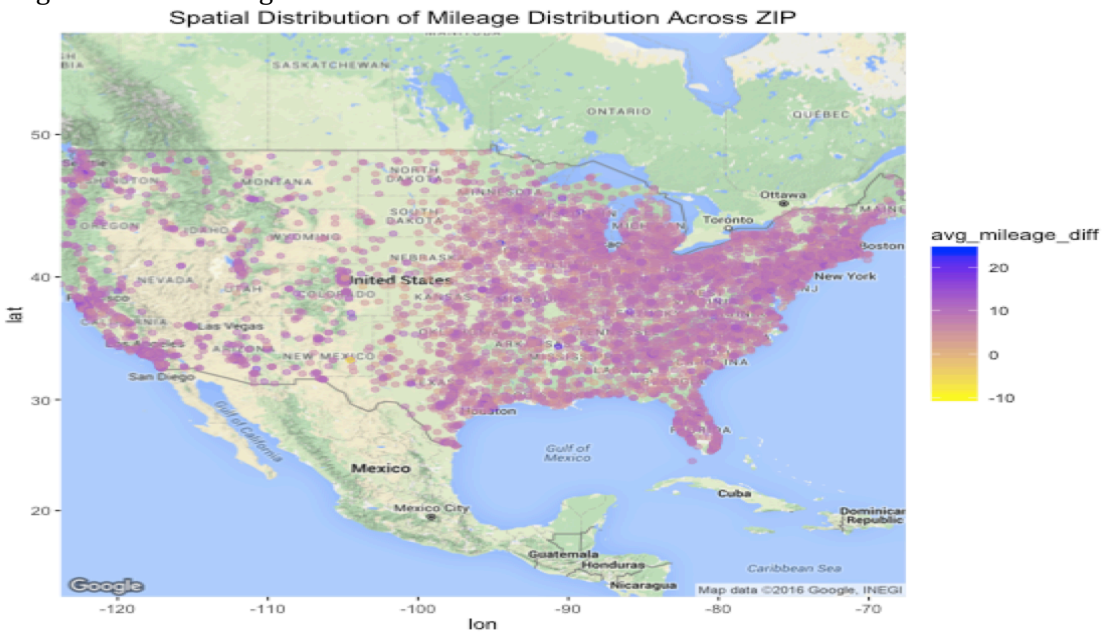
Visualization '2c'

Visualization '2d' indicates average mileage difference across all the states with California being highest closely followed by Utah, Oregon while, Virgin Islands, Louisiana, Michigan being lowest.



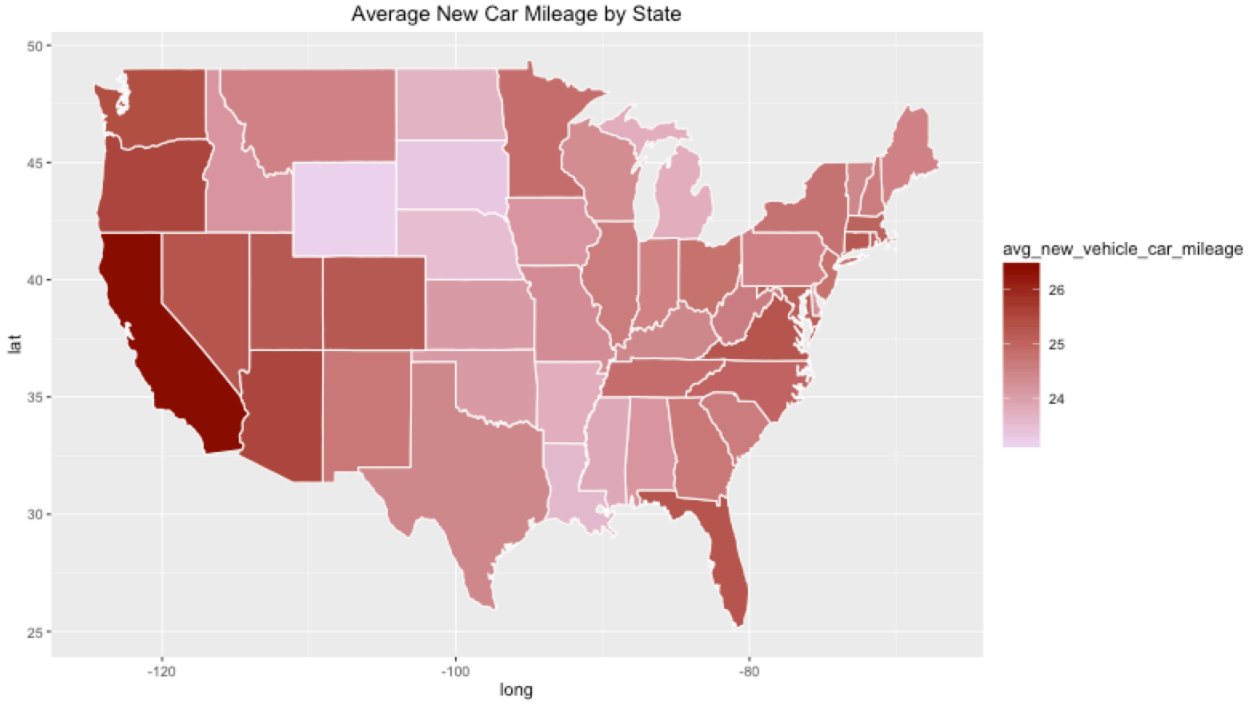
Visualization '2d'

Visualization '2e' indicates spatial distribution of vehicle sales by county with darkness indicating mileage difference being the maximum.



Visualization '2e'

Visualization '2f' indicates new vehicle mileage with darkness indicating average mileage for new car being the maximum. California, Utah, Oregon, Washington all bought higher mileage vehicles (mostly West Coast States).

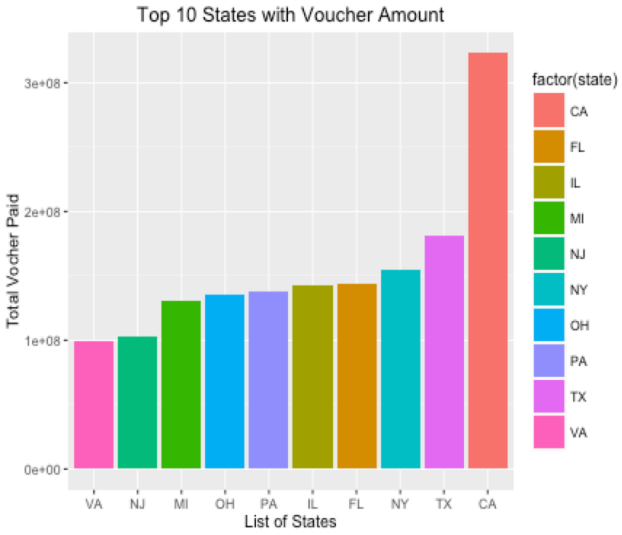


Visualization '2f'

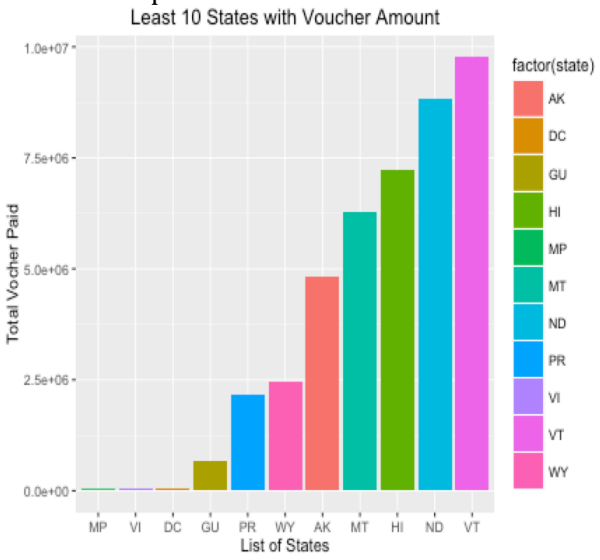
Sum of Invoice Amount: It is sum of Incentive paid by the US Government to each state citizen. The greater the amount paid, more the vehicle transaction and greater the impact of the program. It heavily relies on number of transactions and density of cars in the state.

Visualization 3:

Visualization '3a' represents 10 states with mean maximum total incentive received from Government. As expected we observe California has highest total incentive received by more than 300 million USD, followed by Texas, NY and Florida. So, Californians utilized this program more than any other state citizens by receiving incentive amount to purchase more fuel-efficient vehicles.



Visualization '3a'



Visualization '3b'

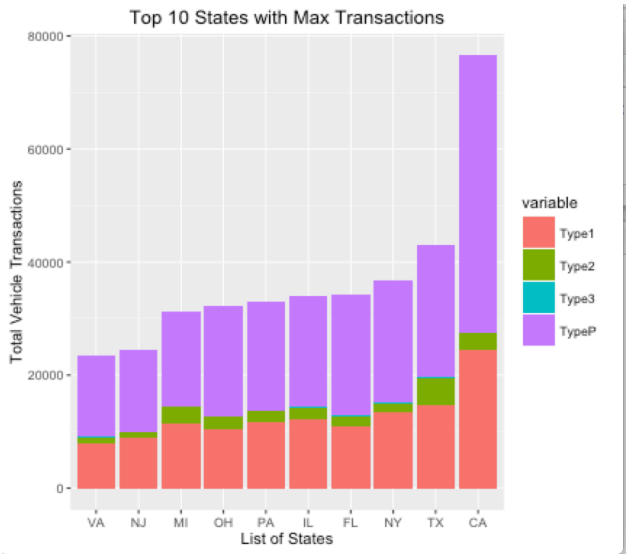
Visualization '3b' represents 10 states with mean minimum total incentive received from Government. As expected we observe smaller states like MP, Virgin Islands, DC have lowest total incentive received, followed by Guam, Puerto Rico and Wyoming. Amazingly, despite the size of Wyoming, this program was least efficient by receiving very less incentive amount to purchase more fuel-efficient vehicles.

New Vehicle Category: Analyzing the new vehicle category we could conclude the success of the program. Greater the percentage of passenger vehicle purchased greater the mileage and hence indicating the success of the program. Greater the percentage of truck vehicle purchased lesser the mileage and hence indicating the failure of the program.

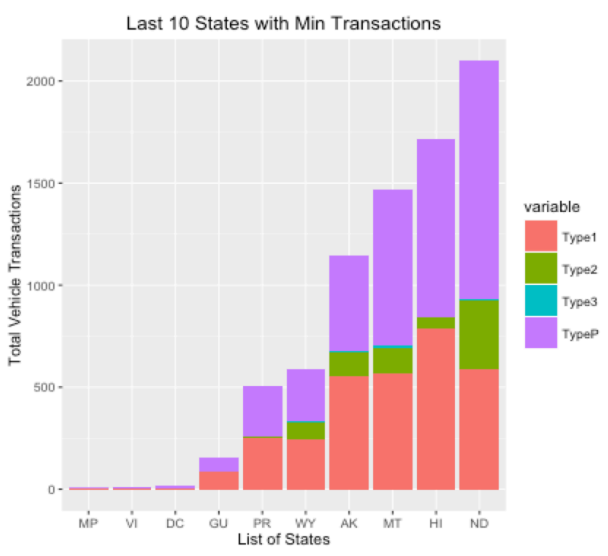
Visualization 4:

Visualization '4a' represents 10 states with maximum total transactions, with color indicating the category of vehicle purchased.

Visualization '4b' represents 10 states with minimum total transactions, with color indicating the category of vehicle purchased. We observe North Dakota citizens buying more proportion of type 2 vehicles than anyone.



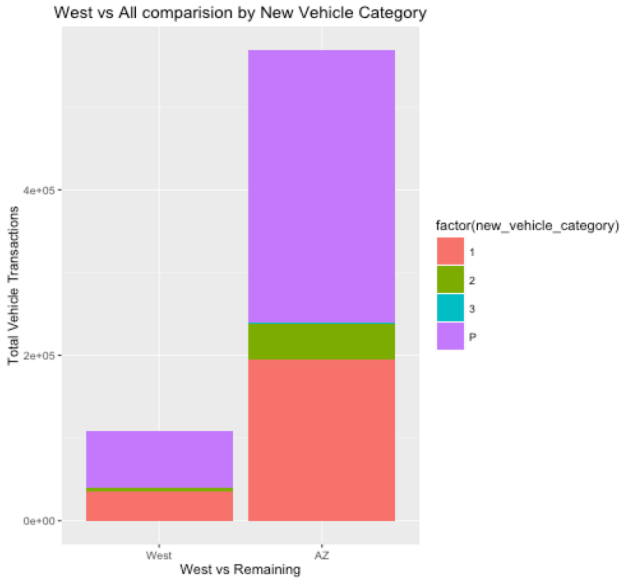
Visualization '4a'



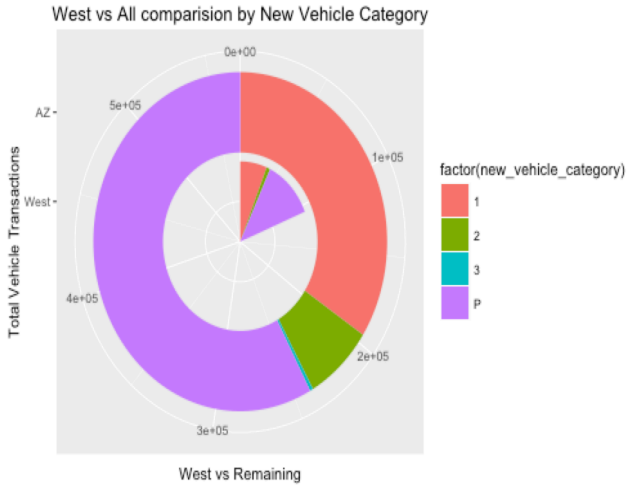
Visualization '4b'

Visualization 5:
'West' indicates west coast states of CA, OR, WA, HI, AL.
'**AZ**' indicates all the other states excluding 'West'.

Visualization '5a' indicates distribution of categories of new vehicles purchased by people of West Coast and rest of the country.
Visualization '5b' shows the pie chart, with inner pie indicating distribution of categories of new vehicles purchased on West Coast and outer pie representing rest of the states.



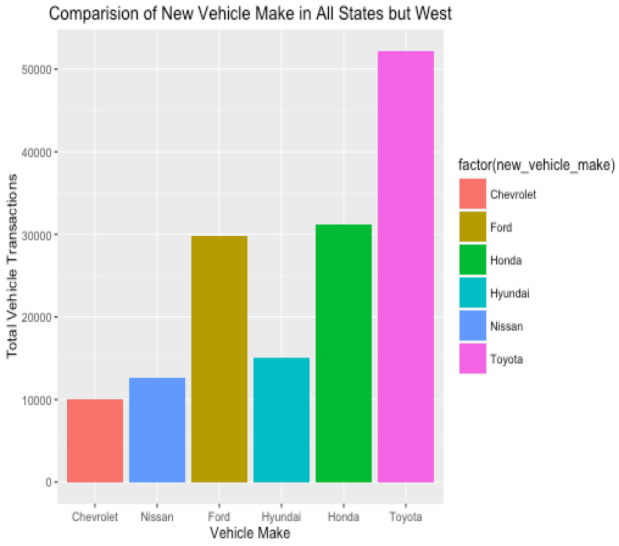
Visualization '5a'



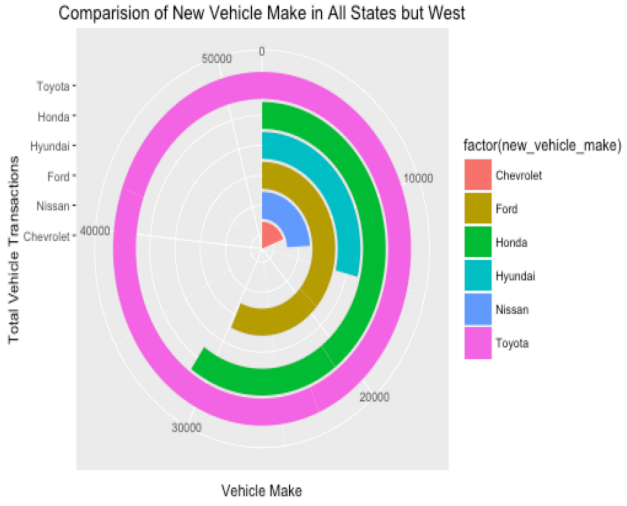
Visualization '5b'

Visualization '5c' indicates distribution of make of new vehicles purchased by people of all states except "West". Toyota is highest seller of its vehicles followed by Honda, Ford, Hyundai by order.

Visualization '5d' indicates the above in pie chart.



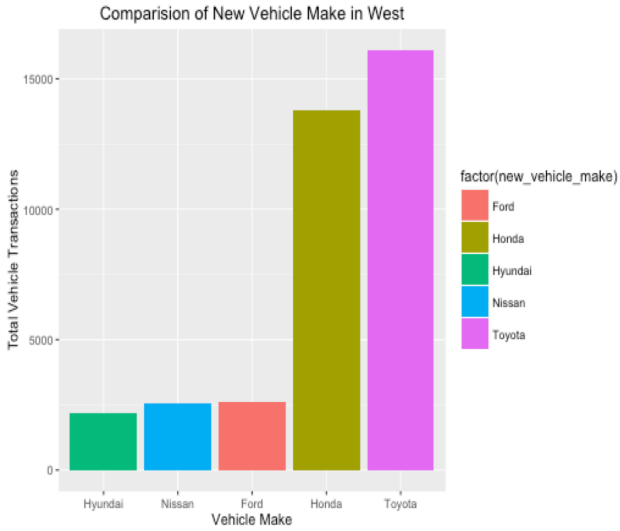
Visualization '5c'



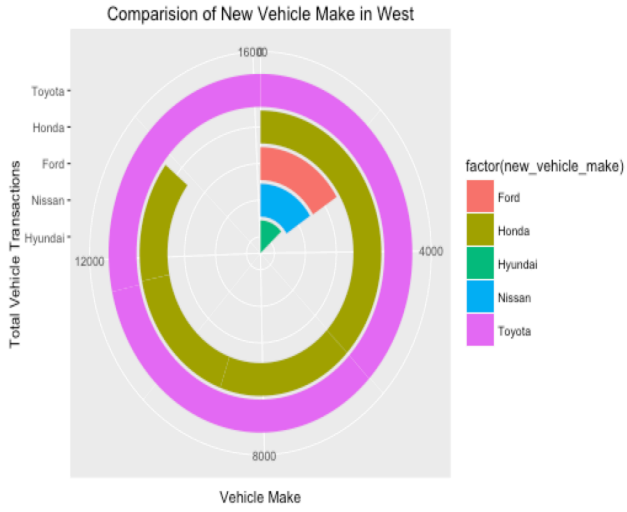
Visualization '5d'

Visualization '5e' indicates distribution of make of new vehicles purchased by people of states "West". Toyota is highest seller of its vehicles; interestingly Toyota is closely competed by Honda with Ford, Nissan and Toyota sharing very small pie on West Coast. West Coast people bought more percentage of Honda vehicles than east coast people.

Visualization '5f' indicates the above in pie chart.



Visualization '5e'



Visualization '5f'

Question 2:
Did West Coast consumers purchase more fuel-efficient cars than consumers in other regions? Please support your answer with analysis and visualizations.

Purpose:
List of West Coast States Assumed: CA, OR, WA, HI, AL.
Mean of Mileage of vehicles purchased by West Coast States: 26.05115 = m1
Mean of Mileage of vehicles purchased by all but West Coast States: 24.64772 = m2
Now there appears to be difference if not significant between means mileages of new vehicles purchased in those two regions. Let's back this up with **two-sample t test**.

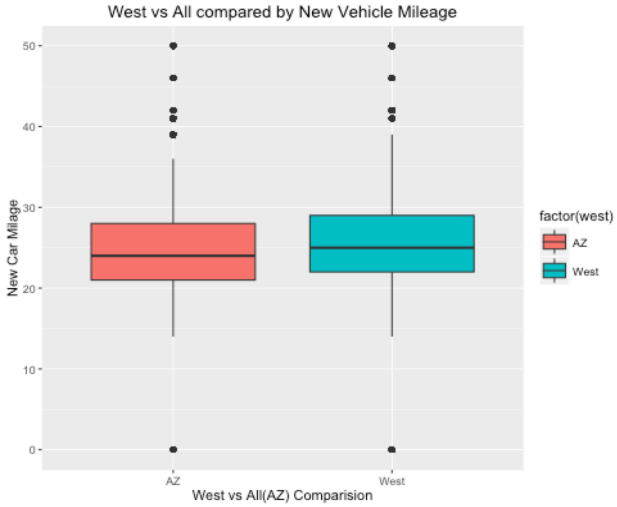
Hypothesis Testing:
H0: Means of mileages of both the regions are same = m1 –m2 = 0
H1: Means of mileages of both the regions are not same = m1 –m2 != 0
Confidence Interval at 95% level = (1.3603,1.4465)
p-value: < 2.2e-16
Since p-value is less than 0.05 **we can reject the null hypothesis**.
There is significant evidence that West coast people bought more fuel-efficient vehicles.

Evidence 1 : t - test

```
        Welch Two Sample t-test

data:  only_west and not_west
t = 63.787, df = 137550, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.360301 1.446547
sample estimates:
mean of x mean of y
 26.05115  24.64772
```

Evidence 2: Boxplot between West coast states and Remaining (defined as 'AZ').



We can clearly see boxplot of West is positioned higher than 'AZ' (all) indicating west coast consumers purchasing more fuel-efficient cars.

Question 3:
From the data, can you find any behavioral patterns that help us understand how consumers buy new vehicles? Can the tendencies you found (if any) be applied to the general population? Support your answer with models / visualizations / analyses, as appropriate.

Purpose:
Now let us build a model to predict what category of vehicle the customers purchase. I chose to go ahead with 'randomForest' because of its capability to handle the factors exceptionally well with out having to create dummy variables. 'RandomForest' are also well known for their sampling and dealing with issues of boot strapping. But 'randomForest' only allows 53 variables per feature as a factor. So, I have decided to merge CA, OR and WA into one factor "West".
Features Utilized:

Dependent Variable:
new_vehicle_category: Dependent Variable, factor with levels 'P', '1', '2', '3'.

Independent Variables:

Independent Variable	Type	Comments
state	Factor	CA, OR, WA merged to 'West'
Sale_date	Numeric	Converted date to numeric value to capture trend
Sales-type	Factor	Factor with two levels 'LEASED' or 'PURCHASED'
Invoice_amount	Integer	Incentive by government
Trade_in_vehicle_category	Factor	Category of Trade-in vehicle
Trade_in_mileage	Integer	Mileage of trade-in vehicle
Trade_in_odometer_reading	Integer	Odometer reading of new vehicle
Mileage_diff	Integer	Mileage of trade in – mileage of new
Age_diff	integer	Age of old vehicle

Training Data Set: 70% of the original data set separated randomly.
Test Data Set: 30% of the original data set separated randomly.

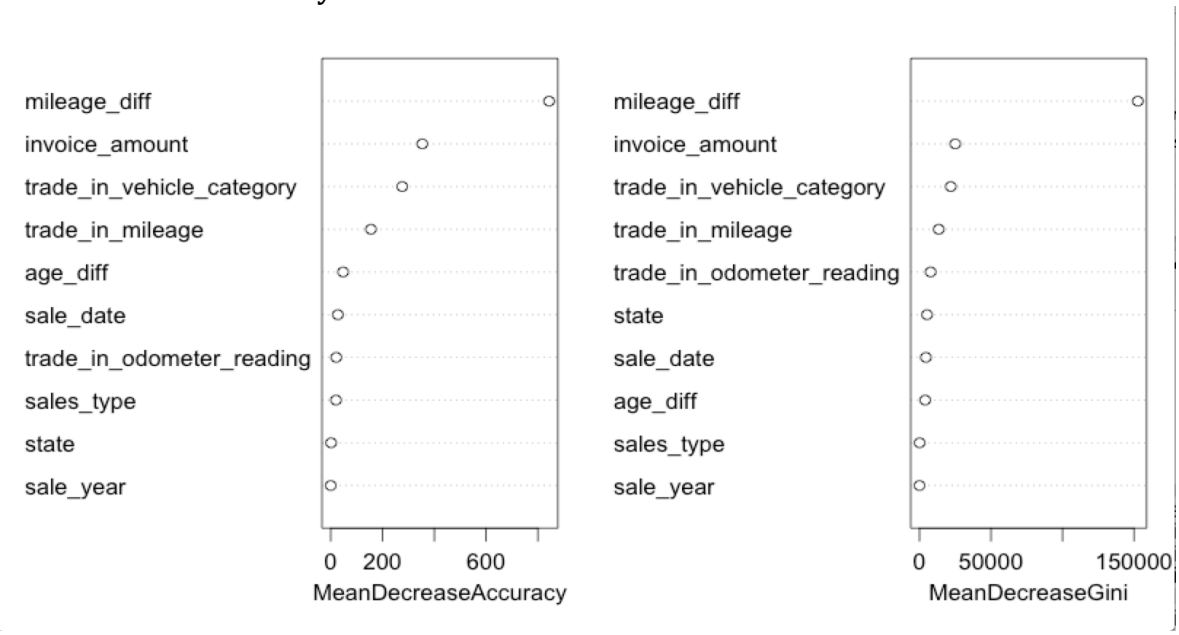
Number of Trees (ntree): 1000
Variables selected at once (mtry): 3 (square root of total variables)
Training set accuracy: **95.16%**
Test set Accuracy: **95.01%** (Indicating Random Forest's very less over-fitting)

Training Set Evidence:
Accuracy and miss-classification Error rates of training set:

```
Call:
 randomForest(formula = new_vehicle_category ~ ., data = train,      ntree = 1000, mtry = 3, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 4.84%
Confusion matrix:
       1     2    3      P class.error
1 144224   267    0  16628  0.10486038
2    865 32311   13    122  0.03002011
3      5   319 1277      5  0.20485679
P   4720    11    0 273299  0.01701615
```

Mean Decrease Accuracy and Mean Decrease Gini:



Mean Decrease Accuracy: It tells how worse the model behaves without the each variable, so a high decrease in accuracy would be expected for very predictive variables. People are actually concerned about mileage difference, which is why it is most important variable followed by invoice_amount, trade_in_vehicle_category.

Mean Decrease Gini: Gini measures how pure the nodes are at the end of the tree. High value indicates variable is very important. Here mileage_diff is sigle most important variable.

Test Set Evidence:

```
          Reference
Prediction      1      2      3      P
         1  61780    369      3   1986
         2    104  13789    133      5
         3      0      5    536      0
         P   7462     62      0 116938

Overall Statistics

               Accuracy : 0.9501
                 95% CI : (0.9492, 0.9511)
    No Information Rate : 0.5854
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9057
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: P
Sensitivity            0.8909  0.96935 0.797619   0.9833
Specificity            0.9824  0.99872 0.999975   0.9107
Pos Pred Value         0.9632  0.98275 0.990758   0.9395
Neg Pred Value         0.9456  0.99769 0.999329   0.9747
Prevalence             0.3413  0.07001 0.003308   0.5854
Detection Rate         0.3041  0.06787 0.002638   0.5756
Detection Prevalence   0.3157  0.06906 0.002663   0.6126
Balanced Accuracy      0.9366  0.98403 0.898797   0.9470
```

Question 4:
The program was declared "wildly successful" by the government. Is there sufficient data from NHTSA to support that conclusion? If not, what additional data will you need in order to determine if the government was right? Explain why this additional data is needed and what you would use it for. (You are not expected to actually go and find that extra data.) If you don't need additional data, was the government right -- and why?

Purpose:
Why the data is not sufficient?

1. During October 2014 average new vehicle fuel economy hits 24.1 miles per gallon, which means fuel economy is on rise irrespective of this program. (25.1 is our dataset's average)
2. One important aspect to consider is during the program in 2009, did aggregate fuel efficiency improve during the period?
3. Calculating the aggregate fuel economy before the program inaugurated and after the program is completed would give great insights into interpretation of the success of program.
4. Average age of trade-in cars is about 20 years, meaning many customers might be in line to replace their cars irrespective of incentive program, in that case government just gave away incentive voucher for nothing.
5. Only way to interpret this is by surveying list of customers who received the coupons if they wanted to change the cars irrespective of the incentive scheme.

What data is required?

1. Survey data to identify which customers were willing to buy new car irrespective of program. Greater such category of customers worse the program.
2. Data before the program was initiated and data after the program has completed, would help analyze the situation.
3. Distribution of vehicles owned across all the states in United States, most people living in Manhattan might not be very interested in the program.
4. Predicted average fuel efficiency by the end of the program vs fuel efficiency after the program is concluded would provide us great insights.

Nevertheless based on the above facts it just is too difficult to analyze if the program is successful with the above data. On base terms gas-guzzlers are replaced with more fuel-efficient cars with the program.

VII. Appendix

Tools Used: R
Packages Used: dplyr, ggplot2, caret, randomForest, data.table, map, reshape2, zipcode, ggmap.

R code (Could not include knit because of time taken by randomForest)**:**

```
library(dplyr)
library(reshape2)
library(ggplot2)
library(data.table)
library(maps)
library(zipcode)
library(ggmap)
library(randomForest)
library(caret)
setwd("~/Downloads/MISC/Edmunds")
cars <- data.table(read.csv('CARS.csv'))
cars[,c(2,3,4,5,6,10,11,12,13,17,18,21,28,29,30,31,32,33) :=NULL]
str(cars)
# cars <- merge(cars, states, by = 'state')
# cars <- subset(cars, new_vehicle_car_mileage != 0)
# cars <- subset(cars, trade_in_mileage != 0)
# cars <- subset(cars, new_vehicle_MSRP > invoice_amount)
# cars <- subset(cars, trade_in_mileage != new_vehicle_car_mileage)
cars$invoice_date <- as.Date(cars$invoice_date, format = "%d-%b-%y")
cars$sale_date <- as.Date(cars$sale_date, "%d-%b-%y")
cars$mileage_diff <- cars$trade_in_mileage - cars$new_vehicle_car_mileage
cars$mileage_diff <- -cars$mileage_diff
cars$customer_cost <- cars$new_vehicle_MSRP - cars$invoice_amount
# cars$amount_per_mile <- cars$customer_cost/cars$mileage_diff
state_summary <- summarise(group_by(cars, state),
                           sum_invoice_amount = sum(invoice_amount),
                           avg_invoice_amount = mean(invoice_amount),
                           avg_trade_in_mileage = mean(trade_in_mileage),
                           avg_trade_in_odometer_reading = mean(trade_in_odometer_reading),
                           avg_new_vehicle_car_mileage = mean(new_vehicle_car_mileage),
                           avg_new_vehicle_MSRP = mean(new_vehicle_MSRP),
                           avg_mileage_diff = mean(mileage_diff),
                           avg_customer_cost = mean(customer_cost))
head(state_summary)
nrow(state_summary)
states <- data.table(table(cars$state))
colnames(states)[1:2] <- c('state', 'vechicles')
state_summary <- merge(state_summary, states, by = 'state')
state_summary <- (state_summary[order(vechicles, decreasing = TRUE),])
ggplot(state_summary[1:10,], aes(reorder(factor(state), vechicles),
                                 vechicles, fill = factor(state))) + geom_bar(stat = "identity") + xlab('List of States')
+ ylab('Count of Vehicles') + ggtitle('Top 10 States vs Vehicles Count')

ggplot(state_summary[45:55,], aes(reorder(factor(state), vechicles),
```

```
                    vechicles, fill = factor(state))) + geom_bar(stat = "identity") + xlab('List of States')
+ ylab('Count of Vehicles') + ggtitle('Least 10 States vs Vehicles Count')

# Maps
data(state)
state_map <- data.frame("state" = state.abb, "Longitude" = state.center$x,
                "Latitude" = state.center$y, "region" = state.name)
state_map$region <- tolower(state_map$region)
map_summary <- state_summary[,c(1,6,8,10),with=FALSE]
state_map <- merge(map_summary, state_map, by = 'state')
state_map <- na.omit(state_map)
raw <- map_data("state")
head(raw)
state_map <- merge(state_map, raw, by = 'region')
p <- ggplot()
p <- p + geom_polygon(data=state_map, aes(x=long, y=lat, group = group, fill=vechicles),
            colour="white") + scale_fill_continuous(low = "thistle2", high = "darkred",
guide="colorbar")
p <- p + ggtitle('Density of Transaction of Vehicles by State')

p2 <- ggplot()
p2 <- p2 + geom_polygon(data=state_map, aes(x=long, y=lat, group = group, fill=avg_mileage_diff),
            colour="white") + scale_fill_continuous(low = "thistle2", high = "darkred",
guide="colorbar")
p2 <- p2 + ggtitle('Average Mileage Difference by State')

p3 <- ggplot()
p3 <- p3 + geom_polygon(data=state_map, aes(x=long, y=lat, group = group,
fill=avg_new_vehicle_car_mileage),
            colour="white") + scale_fill_continuous(low = "thistle2", high = "darkred",
guide="colorbar")
p3 <- p3 + ggtitle('Average New Car Mileage by State')

# Part 2 Invoice Amount
invoice_summary <- (state_summary[order(sum_invoice_amount, decreasing = TRUE),])
ggplot(invoice_summary[1:10,], aes(reorder(factor(state), sum_invoice_amount),
                    sum_invoice_amount, fill = factor(state))) + geom_bar(stat = "identity") +
xlab('List of States') + ylab('Total Vocher Paid') + ggtitle('Top 10 States with Voucher Amount')
ggplot(invoice_summary[45:55,], aes(reorder(factor(state), sum_invoice_amount),
                    sum_invoice_amount, fill = factor(state))) + geom_bar(stat = "identity") +
xlab('List of States') + ylab('Total Vocher Paid') + ggtitle('Least 10 States with Voucher Amount')


# Part 3 avg_mileage_diff
mileage_summary <- data.frame(state_summary[order(avg_mileage_diff, decreasing = TRUE),])
ggplot(mileage_summary[1:10,], aes(reorder(factor(state), avg_mileage_diff),
                    avg_mileage_diff, fill = factor(state))) + geom_bar(stat = "identity") +
geom_bar(stat = "identity") + xlab('List of States') + ylab('Average Mileage Difference') +
ggtitle('Top 10 States with Max Mileage Difference')
ggplot(mileage_summary[45:55,], aes(reorder(factor(state), avg_mileage_diff),
```

```
                    avg_mileage_diff, fill = factor(state))) + geom_bar(stat = "identity") +
geom_bar(stat = "identity") + xlab('List of States') + ylab('Average Mileage Difference') +
ggtitle('Last 10 States with Min Mileage Difference')


# Part 4 involves total transactions, type of cars
category_summary <- summarise(group_by(cars, new_vehicle_category, state),
                    sum_invoice_amount = sum(invoice_amount),
                    avg_invoice_amount = mean(invoice_amount),
                    avg_trade_in_mileage = mean(trade_in_mileage),
                    avg_trade_in_odometer_reading = mean(trade_in_odometer_reading),
                    avg_new_vehicle_car_mileage = mean(new_vehicle_car_mileage),
                    avg_new_vehicle_MSRP = mean(new_vehicle_MSRP),
                    avg_mileage_diff = mean(mileage_diff),
                    avg_customer_cost = mean(customer_cost))

count <- data.table(table(cars$new_vehicle_category, cars$state))
colnames(count)[1:3] <- c('new_vehicle_category', 'state', 'count')
category_summary <- merge(category_summary, count, by = c('new_vehicle_category', 'state'))
category_summary <- (category_summary[order(state),])
category_summary <- dcast(data = category_summary, state ~ new_vehicle_category, value.var =
"count", fun.aggregate = sum)
# category_summary <- transform(category_summary, rowSums(category_summary[,-1]))
colnames(category_summary)[2:5] <- c('Type1', 'Type2', 'Type3', 'TypeP')
category_summary <- melt(data = category_summary)
category_summary <- merge(category_summary, states, by = c('state'))
category_summary <- (category_summary[order(vechicles, decreasing = TRUE),])
ggplot(category_summary[1:40], aes(reorder(factor(state), value), value, fill = variable)) +
geom_bar(stat = "identity") + xlab('List of States') + ylab('Total Vehicle Transactions') + ggtitle('Top
10 States with Max Transactions')
ggplot(category_summary[181:220], aes(reorder(factor(state), value), value, fill = variable)) +
geom_bar(stat = "identity") + xlab('List of States') + ylab('Total Vehicle Transactions') +
ggtitle('Last 10 States with Min Transactions')


# Extra Credit Time Series and Spatial
ggplot(state_summary, aes(x = factor(state), y = avg_mileage_diff, fill = vechicles, group = 1) ) +
geom_point() + geom_line() + xlab('List of States') + ylab('Average Mileage Difference') + ggtitle('All
States with Avg Mileage Difference')
county_summary <- summarise(group_by(cars, ZIP), sum_invoice_amount = sum(invoice_amount),
                    avg_invoice_amount = mean(invoice_amount),
                    avg_trade_in_mileage = mean(trade_in_mileage),
                    avg_trade_in_odometer_reading = mean(trade_in_odometer_reading),
                    avg_new_vehicle_car_mileage = mean(new_vehicle_car_mileage),
                    avg_new_vehicle_MSRP = mean(new_vehicle_MSRP),
                    avg_mileage_diff = mean(mileage_diff),
                    avg_customer_cost = mean(customer_cost))
head(county_summary)
data(zipcode)
```

```
county_summary$ZIP <- clean.zipcodes(county_summary$ZIP)
county_summary <- merge(county_summary, zipcode, by.x ='ZIP', by.y = 'zip')
maps <- get_map(location = 'united states', zoom = 4, maptype = "terrain", source = 'google', color = 'color')
ggmap(maps) + geom_point(
  aes(x=longitude, y=latitude, show_guide = TRUE, color = avg_mileage_diff),
  data = county_summary, alpha = .5, na.rm = T) +
  scale_color_gradient(low="yellow", high ="blue") + ggtitle('Spatial Distribution of Mileage Distribution Across ZIP')

# month <- strftime(cars$sale_date, "%m")
# year <- strftime(cars$sale_date, "%Y")
# df <- data.frame(month, year, cars$new_vehicle_MSRP)
# df.agg <- aggregate(df$cars.new_vehicle_MSRP ~ month + year, df, FUN = sum)
# df.agg$date <- as.POSIXct(paste(df.agg$year, df.agg$month, "01", sep = "-"))
# head(df.agg)



# Question 2
cars$west <- cars$state
# cars$state[cars$state == c('CA', 'OR', 'WA', 'HI', 'AL')] <-'West'
# cars$west<-recode(cars$state,"c('CA','OR','WA', 'HI', 'AL')='West'")
# cars$west[cars$west != 'West'] <- 'Other'
levels(cars$west) <- c(levels(cars$west), "West")
levels(cars$west)[levels(cars$west) %in%  c("CA","WA","OR","HI","AL")] <- "West"
levels(cars$west)[(levels(cars$west) != "West")] <- "AZ"
coast_category <- summarise(group_by(cars, west, new_vehicle_category),
                            sum_invoice_amount = sum(invoice_amount),
                            avg_invoice_amount = mean(invoice_amount),
                            avg_trade_in_mileage = mean(trade_in_mileage),
                            avg_new_vehicle_car_mileage = mean(new_vehicle_car_mileage),
                            avg_mileage_diff = mean(mileage_diff))

only_west <- subset(cars,west == 'West')
not_west <- subset(cars, west == 'AZ')
only_west <- data.frame(only_west$new_vehicle_car_mileage)
not_west <- data.frame(not_west$new_vehicle_car_mileage)
t.test(only_west,not_west)
count <- data.table(table(cars$new_vehicle_category, cars$west))
colnames(count)[1:3] <- c('new_vehicle_category', 'west', 'count')
coast_category <- merge(coast_category, count, by = c('new_vehicle_category', 'west'))
plot1 <- ggplot(coast_category, aes(reorder(factor(west), count), (count), fill = factor(new_vehicle_category))) + geom_bar(stat = "identity")
plot1 + xlab('West vs Remaining') + ylab('Total Vehicle Transactions') + ggtitle('West vs All comparision by New Vehicle Category')
plot2 <- plot1 + coord_polar(theta = "y")
plot2 + ylab('West vs Remaining') + xlab('Total Vehicle Transactions') + ggtitle('West vs All comparision by New Vehicle Category')
```

```
# Pie Charts
car <- data.frame(table(cars$new_vehicle_model, cars$west))
colnames(car)[1:3] <- c('new_vehicle_model', 'west', 'count')
car_summary <- summarise(group_by(cars, west, new_vehicle_model, new_vehicle_make), sum_invoice_amount = sum(invoice_amount), avg_invoice_amount = mean(invoice_amount), avg_trade_in_mileage = mean(trade_in_mileage), avg_trade_in_odometer_reading = mean(trade_in_odometer_reading), avg_new_vehicle_car_mileage = mean(new_vehicle_car_mileage), avg_new_vehicle_MSRP = mean(new_vehicle_MSRP), avg_mileage_diff = mean(mileage_diff), avg_customer_cost = mean(customer_cost))
car_summary <- merge(car_summary, car, by = c('new_vehicle_model','west'))
car_summary <- (car_summary[order(count, decreasing = TRUE),])
car_summary_west <- subset(car_summary, west == 'West')
car_summary_notwest <- subset(car_summary, west == 'AZ')
plot3 <- ggplot(car_summary_notwest[1:10], aes(reorder(factor(new_vehicle_make), count),
                                (count), fill = factor(new_vehicle_make))) + geom_bar(stat = "identity")
plot3 + xlab('Vehicle Make') + ylab('Total Vehicle Transactions') + ggtitle('Comparision of New Vehicle Make in All States but West')
plot4 <- plot3 + coord_polar("y")
plot4 + ylab('Vehicle Make') + xlab('Total Vehicle Transactions') + ggtitle('Comparision of New Vehicle Make in All States but West')
plot5 <- ggplot(car_summary_west[1:10], aes(reorder(factor(new_vehicle_make), count), (count), fill = factor(new_vehicle_make))) + geom_bar(stat = "identity")
plot5 + xlab('Vehicle Make') + ylab('Total Vehicle Transactions') + ggtitle('Comparision of New Vehicle Make in West')
plot6 <- plot5 + coord_polar(theta = "y")
plot6 + ylab('Vehicle Make') + xlab('Total Vehicle Transactions') + ggtitle('Comparision of New Vehicle Make in West')


# Question 3: Modeling Behaviour
cars$sale_year <- format(cars$sale_date, "%Y")
cars$sale_year <- as.character(cars$sale_year)
cars$sale_year <- as.numeric(cars$sale_year)
cars$age_diff <- cars$sale_year - cars$trade_in_year
cars$sale_date <- as.numeric(cars$sale_date)
cars[,c(1,2,4,5,7,11,12,13,14,17,19,20,21,22,23,24,26,27) := NULL]
str(cars)
levels(cars$state) <- c(levels(cars$state), "West")
levels(cars$state)[levels(cars$state) %in%  c("CA","WA","OR")] <- "West"

# K fold cross validation R
x <- sample(nrow(cars),0.7*nrow(cars),replace = FALSE)
train <- cars[x,]
test <- cars[-x,]
train <- na.omit(train)
test <- na.omit(test)
cars_rf <-randomForest(new_vehicle_category~.,data=train,
                ntree=1000, mtry=3, importance = TRUE)
```

```
cars_rf
varImpPlot(cars_rf)
pred <-predict(cars_rf, test[,-8, with = FALSE])
pred <- data.frame(pred)
ref <- data.frame(test[,8,with=FALSE])
confusionMatrix(pred$pred, ref$new_vehicle_category)
```