# 11.3: Final Project

Justin Wisniewski

2022-06-4

## Step 1

### Introduction:

At Waste Management, the adjusted plan metric is one of our main tools to identify opportunities on route. It is compiled of numerous time buckets, such as on property time, disposal time, pre/post trip, travel to/from customer, etc. The adjusted plan is one of the first metrics a route manager will go to for coaching, so it is highly important that the data is correct prior to conversations with a driver. Too many times a data blow will not be corrected, and therefore give a false representation on how a site is performing. The company should be interested in ensuring the data is clean and accurate, as it not only will help drive performance improvement, but also trust between the drivers and management. Furthermore, to show how each time bucket is connected to one another will only help fine tune the adjusted plan in the future.

### Research Questions:

1. What metrics make up the adjusted plan?
2. What transformations/modifications can be made to identify and remove data discrepancies?
3. Which variables are more/less relative to the adjusted plan?
4. Which time buckets are fixed and which are variable?
5. How can we ensure that the customer location and container location are differentiated, as this would affect on property time?
6. Which sites will be a part of the research?
7. What is the leading cause of a driver missing adjusted plan?
8. From a driver specific perspective, should age and/or weight be taken into consideration?
9. Safety being at the forefront of WM, how can this be included/factored into the adjusted plan?
10. Will the variance in clock in time's among the sites need to be controlled?

### Approach

Initially, I would like to start with all sites in the state of Alabama. I believe the first task would be to quickly identify any outliers within the data. The efficiency of the adjusted plan has to have two main metrics to calculate, volume and hours. These would be two metrics to ensure had good data in. It would be inmportant that the container locations were as precise as possible for travel time buckets. Container latitude/longitude values, as well as the drivers confirmation when servicing would be beneficial to improve accuracy. A refined data set free of outliers and discrepancies to help identify trends and actual opportunity being the end goal.

**How the approach addresses (fully or partially) the problem**

> The approach I plan to take will immediately pull all of the garbage data out of what a manager is using to drive results. There is so much data out there, but the discrepancies will muddy up the water, as well as present opportunities that really might not be there. In order for a route manager to have meaningful and effective coaching conversations, it is important that the data is accurate and something he/she can believe in. This approach should shed some light on where actual opportunities are from a driver level perspective vs. the latter being a lot of data that may or may not be accurate.

**Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)**

- Waste Management OPUS Flash Report
  - 99 Columns including all time bucket and metrics from driver punches
- eRouteLogistics
  - Customer export list provides lat / lon as well as ungeocoded containers
- Newton Insights
  - Data is fed directly from OBU used by driver
  - Sites and dates can be filtered how needed

**Required Packages**

- ggplot2
- readxl
- dplyr
- purrr
- Quantpsyc
- ggmap
- lubridate

**Plots and Table Needs**

- Scatterplots to help understand the nature of relationship between variables and the adjusted plan
- Histogram will help in showing what variables make up the time missed in adjusted plan
- Tables with only pertinent data. Volume, hours, as well as the buckets that compile the adjusted plan.
- Density plot to show where the biggest opportunities lie

**Questions for Future Steps**

1. Will I be able to use the ggmap package to assist with ensuring coordinates of container locations are accurate?
2. How will I determine the variables most relative to adjusted plan misses?
3. Could there be a variable/factor not included that would contribute to the variance?
4. What will be the best way to further analyze the different start times for routes?
5. Is there a way to integrate safety metrics?
6. Which plot will be most beneficial in identifying outliers or discrepancies among the data?
7. Can an automated report be sent with identified data blows needing to be corrected?

## Step 2

**How to import and clean data:**

```
setwd("C:/Users/jwiz3/Desktop/Data Statistics/dsc520")
library(readxl)
## Load the `FinalProject/Birmingham.xlsx` to
Birmingham_df <- read_excel("FinalProject/Birmingham.xlsx")
## Load the `FinalProject/Huntsville.xlsx` to
Huntsville_df <- read_excel("FinalProject/Huntsville.xlsx")
## Load the `FinalProject/Moody.xlsx` to
Moody_df <- read_excel("FinalProject/Moody.xlsx")
```

> There are numerous columns within the datasets that will not be important in relation to the adjusted plan. It would be most beneficial to start by selecting only relevant columns. Columns 6, 8, 19, 21, 91, 100, 101, 102, 103. I was able to use is.na to eliminate NA values which was able to eliminate any bad data in the next step.

**What does the final data set look like?:**

```
## Change driver name column to site name
Birmingham_df$Driver <- 'Birmingham'
Huntsville_df$Driver <- 'Huntsville'
Moody_df$Driver <- 'Moody'
## Combine all three data frames
Alabama_df <- do.call("rbind", list(Birmingham_df, Huntsville_df, Moody_df))
## Extract time stamp from pre and post route actual
Alabama_df$preroutetime_component <- format(Alabama_df$`Pre-Route Actual (h:m)`,'%H:%M:%S')
Alabama_df$postroutetime_component <- format(Alabama_df$`Post-Route Actual (h:m)`,'%H:%M:%S')
## Extract time stamp from net idle and non statused time
Alabama_df$idle_component <- format(Alabama_df$`Net Idle`,'%H:%M:%S')
Alabama_df$nonstatus_component <- format(Alabama_df$`Non Statused Time`,'%H:%M:%S')
## Selecting only relevant columns
Alabama_df <- Alabama_df[, c(6,8,19,21,91,100,101,102,103)]
## Drop NA Values
Alabama_df <- Alabama_df[!rowSums((is.na(Alabama_df))),]
# Conversions to datetime object
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

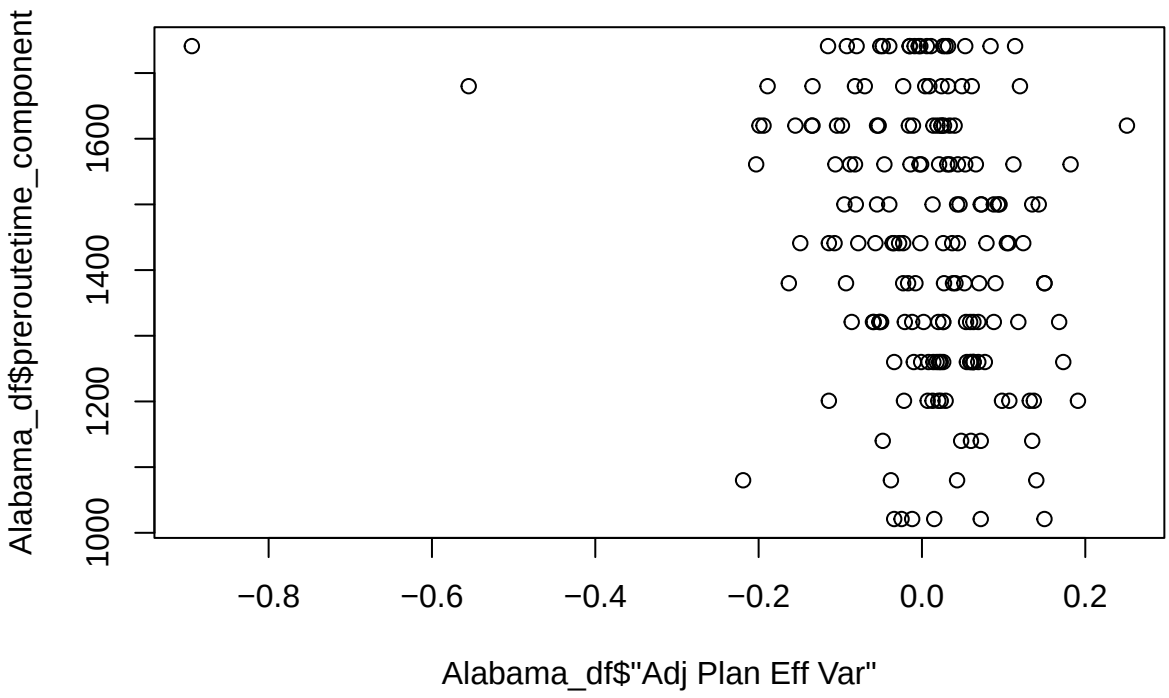
```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(ggplot2)
Alabama_df$preroutetime_component <-  as.duration(hms(Alabama_df$preroutetime_component))
Alabama_df$postroutetime_component <-  as.duration(hms(Alabama_df$postroutetime_component))
```

```r
Alabama_df$idle_component <-  as.duration(hms(Alabama_df$idle_component))
Alabama_df$nonstatus_component <-  as.duration(hms(Alabama_df$nonstatus_component))
## Convert character to factor
Alabama_df$Driver <- as.factor(Alabama_df$Driver)
Alabama_df$MIE <- as.factor(Alabama_df$MIE)
## Convert character to integer
Alabama_df$`Total Actual Units` <- as.integer(Alabama_df$`Total Actual Units`)
## Select relevant data points
Alabama_df = Alabama_df[Alabama_df$'preroutetime_component' > 1000 & Alabama_df$preroutetime_component <
Alabama_df = Alabama_df[Alabama_df$'postroutetime_component' > 600 & Alabama_df$postroutetime_component
summary(Alabama_df)
```

```
##         Driver   MIE     Adj Plan Eff Var    Total Actual Units
##  Birmingham:85   N:122   Min.   :-0.894000   Min.   : 1.000
##  Huntsville:60   Y: 54   1st Qu.:-0.041500   1st Qu.: 5.000
##  Moody     :31           Median : 0.018500   Median : 6.000
##                          Mean   : 0.001585   Mean   : 5.812
##                          3rd Qu.: 0.060250   3rd Qu.: 7.000
##                          Max.   : 0.251000   Max.   :12.000
##     Occ Idle      preroutetime_component
##  Min.   :0.0000   Min.   :1021s (~17.02 minutes)
##  1st Qu.:0.0000   1st Qu.:1260s (~21 minutes)
##  Median :0.0000   Median :1441s (~24.02 minutes)
##  Mean   :0.8636   Mean   :1444.26136363636s (~24.07 minutes)
##  3rd Qu.:1.0000   3rd Qu.:1620s (~27 minutes)
##  Max.   :5.0000   Max.   :1741s (~29.02 minutes)
##  postroutetime_component
##  Min.   :661s (~11.02 minutes)
##  1st Qu.:901s (~15.02 minutes)
##  Median :1201s (~20.02 minutes)
##  Mean   :1176.25568181818s (~19.6 minutes)
##  3rd Qu.:1441s (~24.02 minutes)
##  Max.   :1741s (~29.02 minutes)
##  idle_component
##  Min.   :0s
##  1st Qu.:0s
##  Median :0s
##  Mean   :511.221590909091s (~8.52 minutes)
##  3rd Qu.:735.25s (~12.25 minutes)
##  Max.   :3601s (~1 hours)
##  nonstatus_component
##  Min.   :0s
##  1st Qu.:180s (~3 minutes)
##  Median :540s (~9 minutes)
##  Mean   :1437.69318181818s (~23.96 minutes)
##  3rd Qu.:1395.25s (~23.25 minutes)
##  Max.   :10201s (~2.83 hours)
```

```
plot(Alabama_df$'Adj Plan Eff Var',Alabama_df$preroutetime_component)
```




**Questions for future steps**

I am not completely certain how to change the columns in a (h:m) format, or which format will be the most beneficial. I think it might be easier to try to get it into a decimal format, for example 9.23 hours, or 10.59 hours. I would like to put min and max limits on the actual driver hours, as this will then eliminate any outliers that will affect the end result.

**What information is not self-evident?**

The data does not come with the site name, only driver. It would be beneficial to add a column with the site name for upper management to be able to identify opportunity at a site level. What variable has the biggest impact on the magnitude? The answer to this question will be the highest level view of what management should be focusing on. Magnitude is a metric that shows how much a driver is missing and/or out performing the adjusted plan.

**What are different ways you could look at this data?**

I believe each metric and time bucket tells a different story. For example, it'd be safe to assume that downtime events would have a negative impact on the adjusted plan. However, the adjusted plan does take downtime into consideration, if it logged properly. One way this data can be looked at, is identifying the outliers, and using these to present opportunities to management

on driver tablet usage. Good data in, good data out. The more precise and accurate drivers are on the tablet, the easier it will be to identify opportunties at a driver level. MIE, multiple incident employee is the way the safety metric can be factored into this analysis. Safety, the most important aspect of our business, should be and needs to be included in this data.

**How do you plan to slice and dice the data?**

With the plan to add variable for site, it will be beneficial to splice together the three data frames for the Alabama sites, to ensure an easy to read code and document. I would also like to add a column that will include down in yard and down on route, as these both ultimately are downtime and can be combined for the purposes of this project. When it comes down to over simplifying the end result, I can see the ability to identify at a driver level, the biggest opportunity/relation to the miss on adjusted plan. A driver socrecard in a sense.

**How could you summarize your data to answer key questions?**

Adding all the fixed variable times could be one way to summarize the relationship the other variables have to the adjusted plan. Simply showing each variable alongside the adjusted plan eff var will give management an easy way to see what is most relevant for a site to include in the 30-60-90 plans. Magnitude and frequency are the two most important columns within the datasets. This not only shows us the frequency of a driver missing the route plan, but also the time they are missing it by. Summarizing the data in relation to those two specifically will be most beneficial.

**What types of plots and tables will help you to illustrate the findings to your questions?**

Scatterplots and correlation plots I feel will be the best to illustrate the findings to business questions. The easier it is to idetify what correlates the strongest to the adjusted plan variance, the more efficient are coaching will become.

**Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**

I believe the linear regression models we ran in weeks eight and nine will be the most benefical in answering the research questions for this project. Having the ability to predict the outcome variable based off the predictor will give us an opportunity to get ahead of further adjusted plan misses. Correlation, p-values, r^2 will also be values I want included in the final.

**Questions for future steps.**

My plan is to start from where we did this semester. Building from the basics, to plotting and visualizing the data. I think there will be questions that come along the way with the best way to format and design the visualizations, but I believe this will become trial-and-error for what will portray the best. The machine learning techniques is where I feel I will spend the majority of my time once visualizations and plotting is complete. I will want to get as much data driven from what is already given, in order to have the most evidence behind the "why" of an adjusted plan miss.

## Step 3

### Introduction

At Waste Management, the adjusted plan metric is one of our main resources to identify opportunities on route. The adjusted plan is one of the first metrics a route manager will go to for coaching, so it is highly important that the data is correct prior to conversations with a driver. Too many times a data blow will not be corrected, and therefore give a false representation on how a site is performing. The company should be interested in ensuring the data is clean and accurate, as it not only will help drive performance improvement, but also trust between the drivers and management. Furthermore, to show how each time bucket is connected to one another will only help fine tune the adjusted plan in the future. We will be looking at data from three WM sites in the state of Alabama (Birmingham, Huntsville, and Moody). The primary focus was the adjusted plan eff variance, and which time buckets have the most correlation. Total Actual Units that are hauled has the biggest impact on the efficiency variance. This makes sense, as hauls and hours is how the efficiency number is created. It appears pre-trip has the greatest negative relationship with the adjusted plan. Something to point out would be the strong relationship between idle occurrences and idle time, which also would go hand in hand with each other.

### Problem

The main problem is identifying what time bucket / variable is the biggest opportunity for the WM sites within the state of Alabama. Furthermore, eliminating the outliers that are data discrepancies to ensure that the opportunity is indeed accurate.

### Plan Of Attack

The first task was to import the three sites OPUS data, as well as re-format the columns that kept the (h:m) format preventing certain models from running. Next is eliminating the 90 variables of data within each file that was not necessary for this problem. Lastly, to then integrate all of the sites together to have an overall picture of the opportunity within Alabama. The linear model as well as the correlation below were the first two steps taken to then see what our focus needs to be for 30-60-90 day plans.

### Analysis

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(broom)
library(ggpubr)
library(corrr)
cor(Alabama_df[sapply(Alabama_df,is.numeric)])
```

```
##                          Adj Plan Eff Var Total Actual Units    Occ Idle
## Adj Plan Eff Var               1.00000000         0.40832951 -0.11724605
## Total Actual Units             0.40832951         1.00000000 -0.15095871
## Occ Idle                      -0.11724605        -0.15095871  1.00000000
## preroutetime_component        -0.25815048        -0.06339621  0.03465761
## postroutetime_component       -0.09515105        -0.05576901  0.13464885
## idle_component                -0.11881321        -0.14842120  0.95336290
## nonstatus_component           -0.16791570        -0.11197260  0.01644057
##                          preroutetime_component postroutetime_component
## Adj Plan Eff Var                     -0.25815048             -0.09515105
## Total Actual Units                   -0.06339621             -0.05576901
## Occ Idle                              0.03465761              0.13464885
## preroutetime_component                1.00000000             -0.03003892
## postroutetime_component              -0.03003892              1.00000000
## idle_component                        0.06354329              0.09857180
## nonstatus_component                  -0.06944794              0.05530899
##                          idle_component nonstatus_component
## Adj Plan Eff Var            -0.11881321         -0.16791570
## Total Actual Units          -0.14842120         -0.11197260
## Occ Idle                     0.95336290          0.01644057
## preroutetime_component       0.06354329         -0.06944794
## postroutetime_component      0.09857180          0.05530899
## idle_component               1.00000000          0.03815459
## nonstatus_component          0.03815459          1.00000000
```

```
## Fit a linear model using the `Nonstatus_component` variable as the predictor and `Adj Plan Eff Var` a
nonstatus_component_lm <-  lm(Alabama_df$'Adj Plan Eff Var'~Alabama_df$nonstatus_component,data = Alaba
idle_component_lm <-  lm(Alabama_df$'Adj Plan Eff Var'~Alabama_df$idle_component,data = Alabama_df)
## Fit a linear model using several predictor variables and `Adj Plan Eff Var` as the outcome
AdjPlanMultVar_lm <-  lm(Alabama_df$'Adj Plan Eff Var'~Alabama_df$preroutetime_component+Alabama_df$pos
## View the summary of your model using `summary()`
summary(nonstatus_component_lm)
```

```
##
## Call:
## lm(formula = Alabama_df$"Adj Plan Eff Var" ~ Alabama_df$nonstatus_component,
##     data = Alabama_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89403 -0.04199  0.01263  0.05404  0.23917
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.389e-02  1.018e-02   1.365   0.1741
## Alabama_df$nonstatus_component -8.557e-06  3.808e-06  -2.247   0.0259 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1138 on 174 degrees of freedom
## Multiple R-squared:  0.0282, Adjusted R-squared:  0.02261
## F-statistic: 5.048 on 1 and 174 DF,  p-value: 0.02591
```

```
## View the summary of your new model using `summary()`
summary(AdjPlanMultVar_lm)
```

```
##
## Call:
## lm(formula = Alabama_df$"Adj Plan Eff Var" ~ Alabama_df$preroutetime_component +
##     Alabama_df$postroutetime_component + Alabama_df$idle_component +
##     Alabama_df$nonstatus_component + Alabama_df$MIE, data = Alabama_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85678 -0.04416  0.01729  0.05589  0.25747
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       2.745e-01  6.891e-02   3.984 0.000100 ***
## Alabama_df$preroutetime_component -1.485e-04  4.223e-05  -3.516 0.000563 ***
## Alabama_df$postroutetime_component -3.021e-05  2.591e-05  -1.166 0.245388
## Alabama_df$idle_component         -1.230e-05  1.091e-05  -1.128 0.260931
## Alabama_df$nonstatus_component    -9.268e-06  3.696e-06  -2.508 0.013089 *
## Alabama_df$MIEY                   -1.105e-02  1.843e-02  -0.599 0.549704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1096 on 170 degrees of freedom
## Multiple R-squared:  0.1193, Adjusted R-squared:  0.09337
## F-statistic: 4.605 on 5 and 170 DF,  p-value: 0.0005767
```

```
## Standardized Betas
library('QuantPsyc')
```

```
## Loading required package: boot
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'QuantPsyc'
```

```
## The following object is masked from 'package:base':
##
##     norm
```

```
## Standardized betas for each parameter
lm.beta(nonstatus_component_lm)
```

```
## Alabama_df$nonstatus_component
##                     -0.1679157
```

```
lm.beta(idle_component_lm)
```

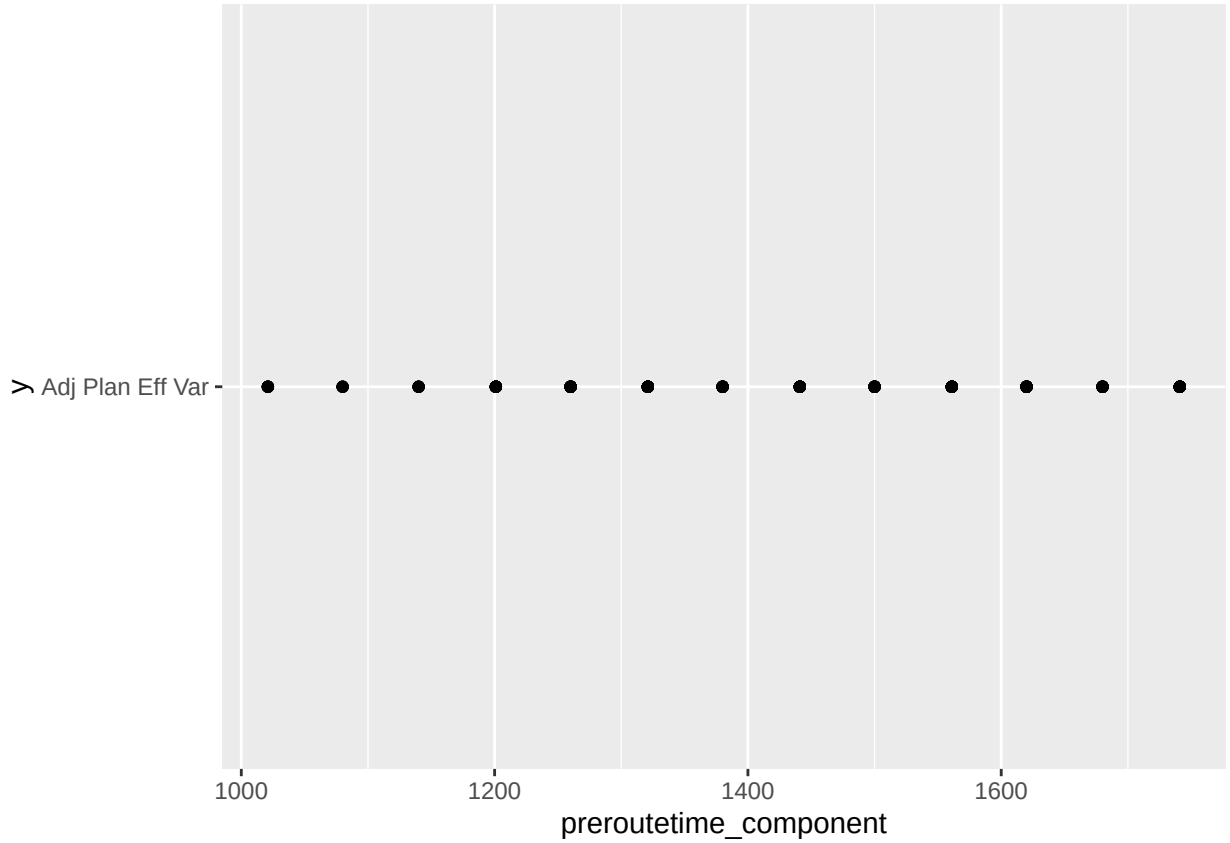
```
## Alabama_df$idle_component
##                -0.1188132
```

```
## Confidence Intervals
confint(AdjPlanMultVar_lm)
```

```
##                                          2.5 %        97.5 %
## (Intercept)                           1.385214e-01  4.105657e-01
## Alabama_df$preroutetime_component    -2.318315e-04 -6.510163e-05
## Alabama_df$postroutetime_component   -8.136038e-05  2.094768e-05
## Alabama_df$idle_component            -3.382918e-05  9.226998e-06
## Alabama_df$nonstatus_component       -1.656344e-05 -1.972459e-06
## Alabama_df$MIEY                      -4.744080e-02  2.534098e-02
```

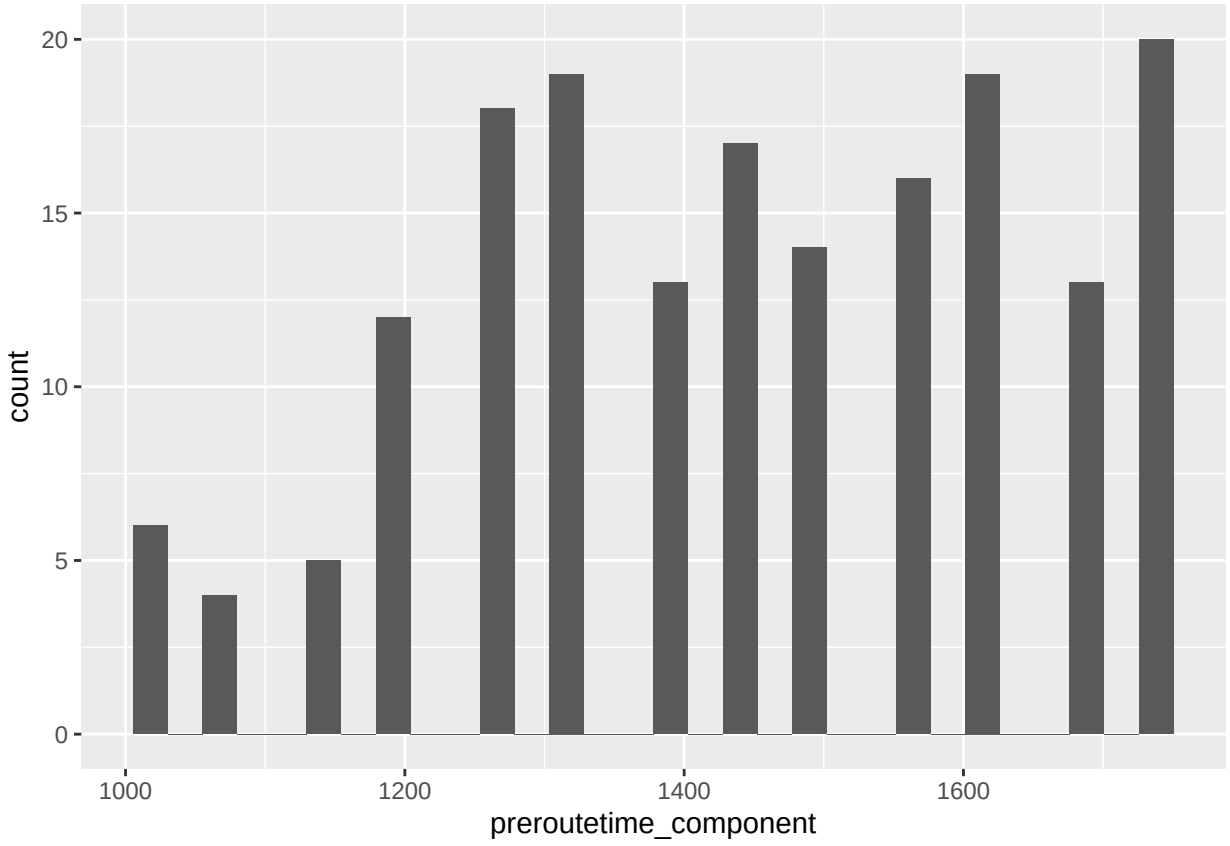
**Plotting**

```
## Using `geom_point() scatterplot for pre route and adj plan
ggplot(Alabama_df, aes(x=preroutetime_component, y='Adj Plan Eff Var')) + geom_point()
```

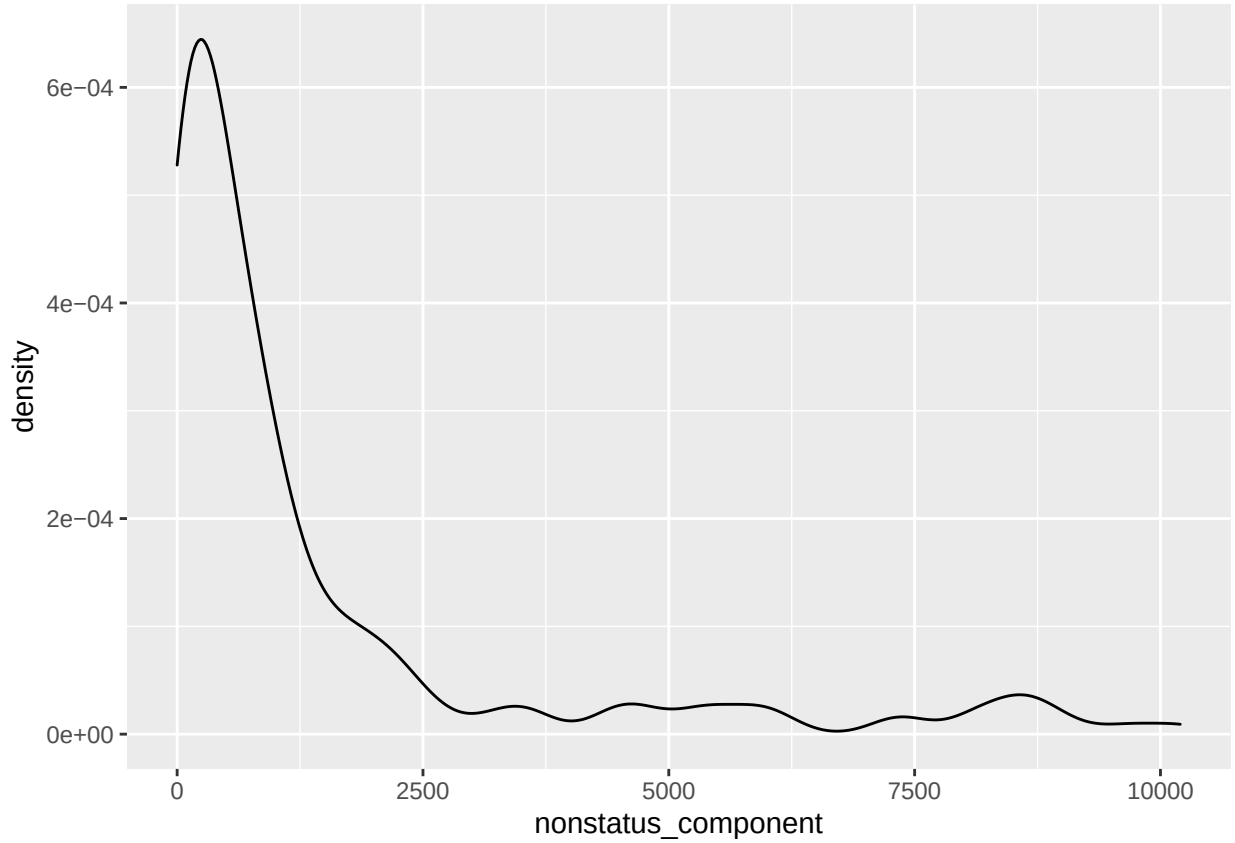



```
## Histogram for pre route
ggplot(Alabama_df, aes(preroutetime_component)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## Density plot for non status
ggplot(Alabama_df, aes(nonstatus_component)) +  geom_density()
```




### Tests

```
library("car")
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
dwt(AdjPlanMultVar_lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.04075549      2.080027   0.654
##  Alternative hypothesis: rho != 0
```

### Implications

The R2 value at the bottom of each summary tells us whether the model is successful in predicting the outcome and if the difference between R2 and adjusted R2 values is small this would indicate that the sample taken is a good representation of the population. First regression model, R2 is 0.02261 so this indicated that nonstatus_component accounted for only 2.26% of the variation in adjusted plan. Multiple regression model, R2 is 0.09337, so this multiple predictor model accounted for 9.34% of the variation in adjusted plan. The inclusion of the new predictors made an impact, but as close as these values are it appears we have a good representation of the WM sites.

### Limitations

From a data standpoint, I believeI had more than I would have even needed for this problem specifically. The biggest concern will still always be bad data. A 24+ hour pre or post trip should stick out, but does't always. I think the biggest limitation is when using the adjusted plan variance, the number is so small that it was affecting the plots and charts trying to portray good info. I think using the actual time the adjusted plan was beat/missed in the (h:m) format would help clean up the look of some of the visualizations.

### Concluding Remarks

I hope this analysis helped paint a better picture of the opportunity in the state of Alabama for WM from an efficiency standpoint, and that to start is simply getting more hauls. Furthermore, proper tablet usage and an increased focus on training for drivers would help from a data perspective.