

10.3: Final Project Step 1 & 2

Justin Wisniewski

2022-05-24

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 important 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

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, 37, 38, 41, 44, 59, 62, 69, 81, 84, 92, 93, 95, 97. I would also like to add a column that is a total of the down in yard and down on route times. I believe we can get a better handle of relevant data points by using Total Actual Units. I'm thinking looking at Total Actual Units above 3, as well as eliminating any actual hours of 0 or 14+. This will be a formatting issue that I will need to dig into further for the final data set.

What does the final data set look like?:

```
## Select relevant data points, Total Actual Units > 3
Birmingham_df = Birmingham_df[Birmingham_df$'Total Actual Units'>3, ]
Huntsville_df = Huntsville_df[Huntsville_df$'Total Actual Units'> 3, ]
Moody_df = Moody_df[Moody_df$'Total Actual Units'> 3, ]
## Selecting only relevant columns
Birmingham_df <- Birmingham_df[, c(6,8,19,37,38,41,44,59,62,69,81,84,92,93,95,97)]
Huntsville_df <- Huntsville_df[, c(6,8,19,37,38,41,44,59,62,69,81,84,92,93,95,97)]
Moody_df <- Moody_df[, c(6,8,19,37,38,41,44,59,62,69,81,84,92,93,95,97)]
```

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 opportunities 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 identify 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 beneficial 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.