Categories

1. [Basic Statistics](https://datascienceplus.com/category/basic-statistics/)

Tags

1. [Data Management](https://datascienceplus.com/tag/data-management/)
2. [R Programming](https://datascienceplus.com/tag/rstats/)

We are following on from our [previously loaded data of NYC bus delays](https://datascienceplus.com/nyc-bus-delays/), and from our simple Cubist regression fit. We know that we need to add in more predictors, and we have access to company level data. A warning to the faint-hearted: this post is essentially all data manipulation and joins. No pretty graphs, no conclusions. I am documenting the slog because the slog is required. We will move on to better things afterwards.

**Preparation of company and route level data**

We read in all of our data and take a look:

#---II\_Breakdown\_and\_Delays

##NR: this actually isn't what I want here. Where is the other data?

in\_csv <- "../data/II\_Bus\_Breakdown\_and\_Delays.csv"

ii\_breakdowns <- read\_csv(in\_csv)

ii\_breakdowns %>% glimpse

*## Observations: 259,637*

*## Variables: 21*

*## $ School\_Year "2015-2016", "2015-2016", "201...*

*## $ Busbreakdown\_ID 1212699, 1212700, 1212701, 121...*

*## $ Run\_Type "Special Ed AM Run", "Special ...*

*## $ Bus\_No "48186", "2518", "235", "2102"...*

*## $ Route\_Number "N758", "L530", "K168", "K216"...*

*## $ Reason "Other", "Mechanical Problem",...*

*## $ Schools\_Serviced "75485", "21854", "18366", "21...*

*## $ Occurred\_On 2015-09-02 06:27:00, 2015-09-...*

*## $ Created\_On 2015-09-02 06:29:00, 2015-09-...*

*## $ Boro "Nassau County", "Brooklyn", "...*

*## $ Bus\_Company\_Name "BORO TRANSIT, INC.", "RELIANT...*

*## $ How\_Long\_Delayed "25 minutes", NA, "30MINS", "2...*

*## $ Number\_Of\_Students\_On\_The\_Bus 0, 0, 0, 1, 0, 0, 0, 9, 0, 2, ...*

*## $ Has\_Contractor\_Notified\_Schools "Yes", "Yes", "Yes", "Yes", "Y...*

*## $ Has\_Contractor\_Notified\_Parents "No", "Yes", "Yes", "Yes", "Ye...*

*## $ Have\_You\_Alerted\_OPT "No", "Yes", "No", "No", "No",...*

*## $ Informed\_On 2015-09-02 06:29:00, 2015-09-...*

*## $ Incident\_Number NA, NA, NA, NA, NA, NA, NA, NA...*

*## $ Last\_Updated\_On 2015-09-02 06:29:16, 2015-09-...*

*## $ Breakdown\_or\_Running\_Late "Running Late", "Breakdown", "...*

*## $ School\_Age\_or\_PreK "School-Age", "School-Age", "S...*

##-------data loading

#---III\_Drivers\_and\_Attendants

in\_csv <- "../data/III\_Drivers\_and\_Attendants.csv"

iii\_drivers <- read\_csv(in\_csv)

iii\_drivers %>% glimpse

*## Observations: 832*

*## Variables: 5*

*## $ School\_Year "2018-2019", "2018-2019", "2018-2019", "2018-...*

*## $ Vendor\_Name "ACADEMY EXPRESS LLC", "E-Z BUS LLC (B2321)",...*

*## $ Type\_of\_Service "Coach", "School Age", "School Age", "School ...*

*## $ Active\_Employees 65, 5, 3, 2, 1, 3, 6, 226, 22, 18, 13, 2, 20,...*

*## $ Job\_Type "Driver", "Attendant", "Driver", "Driver&Atte...*

#---IV\_Routes

in\_csv <- "../data/IV\_Routes.csv"

iv\_routes <- read\_csv(in\_csv)

iv\_routes %>% glimpse

*## Observations: 42,114*

*## Variables: 14*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016"...*

*## $ Route\_Number "C911", "J499", "J500", "J501", "J502...*

*## $ Service\_Type "D2D", "D2D", "D2D", "D2D", "D2D", "D...*

*## $ Vehicle\_TypeDescription "Non-Wheelchair Accessible Alternativ...*

*## $ Route\_Start\_Date "03/21/2016", "09/09/2015", "09/01/20...*

*## $ Vendor\_Code "VN", "RV", "HT", "HT", "HT", "HT", "...*

*## $ Vendor\_Name "VAN TRANS LLC (B2192)", "RELIANT TRA...*

*## $ Vendor\_Affiliation "VAN TRANS LLC (B2192)", "RELIANT TRA...*

*## $ `Garage \_Street\_Address` "670 Hillside Road", "297 NORMAN AVEN...*

*## $ Garage\_City "Pelham Manor", "Brooklyn", "BROOKLYN...*

*## $ Garage\_State "NY", "NY", "NY", "NY", "NY", "NY", "...*

*## $ Garage\_Zip 10803, 11222, 11224, 11224, 11224, 11...*

*## $ XCoordinates 1034574, 1000787, 983439, 983439, 983...*

*## $ YCoordinates 265547.6, 204481.0, 148884.0, 148884....*

#---V\_Routes\_by\_Transportation\_Sites

in\_csv <- "../data/V\_Routes\_by\_Transportation\_Sites.csv"

v\_rbt <- read\_csv(in\_csv)

v\_rbt %>% glimpse

*## Observations: 98,506*

*## Variables: 3*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016", "2015-2016...*

*## $ Route\_Number "J698", "J699", "J700", "J700", "J701", "J701", "...*

*## $ OPT\_Code 75004, 75580, 75003, 75140, 75003, 75140, 75003, ...*

#---VI\_Transportation\_Sites

in\_csv <- "../data/VI\_Transportation\_Sites.csv"

vi\_transport <- read\_csv(in\_csv)

vi\_transport %>% glimpse

*## Observations: 18,220*

*## Variables: 18*

*## $ School\_Year "2016-2017", "2016-2017", "2016-...*

*## $ OPT\_Code "01001", "01569", "01600", "0161...*

*## $ Name "Comprehensive Kids Devel. Scho"...*

*## $ Affiliation "Other Religion", "Public", "Pub...*

*## $ Site\_Type "School", "School", "School", "S...*

*## $ Street\_Address "*[*101 Norfolk Street*](https://www.google.com/maps/search/101+Norfolk+Street?entry=gmail&source=g)*", "525 EAST ...*

*## $ City "Manhattan", "Manhattan", "Manha...*

*## $ State "NY", "NY", "NY", "NY", "NY", "N...*

*## $ Zip 10002, 10002, 10002, 10002, 1000...*

*## $ Longitude -73.98727, -73.97544, -73.99063,...*

*## $ Latitude 40.71879, 40.71876, 40.72249, 40...*

*## $ Door\_To\_Door\_Service "Yes", "Yes", "No", "Yes", "Yes"...*

*## $ Stop\_To\_School\_Service "No", "No", "No", "No", "Yes", "...*

*## $ Common\_Carrier\_Svc\_Metrocards "No", "Yes", "Yes", "Yes", "Yes"...*

*## $ Site\_Reimbursement "No", "No", "No", "No", "No", "N...*

*## $ Mid\_Day\_Service "No", "No", "No", "Yes", "Yes", ...*

*## $ D2D\_Late\_Day\_Programs "No", "No", "No", "No", "No", "N...*

*## $ S2S\_Late\_Day\_Programs "No", "No", "No", "No", "No", "N...*

#---VII\_Vehicles

in\_csv <- "../data/VII\_Vehicles.csv"

vii\_vehicles <- read\_csv(in\_csv)

vii\_vehicles %>% glimpse

*## Observations: 764*

*## Variables: 4*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016",...*

*## $ Vendor\_Name "ACADEMY EXPRESS LLC", "ACME BUS CORP....*

*## $ Vehicle\_TypeDescription "Motor Coach Bus", "Mini-Wagon", "Mini...*

*## $ Active\_Vehicles 22, 71, 6, 62, 1, 9, 71, 6, 63, 47, 67...*

#---VIII\_PreK\_Riders\_by\_Transportation\_Site

in\_csv <- "../data/VIII\_PreK\_Riders\_by\_Transportation\_Site.csv"

viii\_prt <- read\_csv(in\_csv)

viii\_prt %>% glimpse

*## Observations: 1,268*

*## Variables: 5*

*## $ School\_Year "2017-2018", "2017-2018", "2017-2018", "2017-...*

*## $ OPT\_Code "C225", "E011", "E019", "E082", "E148", "E187...*

*## $ Site\_Name "Highbridge Advisory Council", "Crossroads Sc...*

*## $ School\_Name "Marshall England", "Crossroads School for Ch...*

*## $ Number\_of\_Riders 50, 2, 1, 3, 3, 2, 90, 67, 22, 6, 7, 8, 4, 3,...*

#---IX\_PreK\_Vendors\_by\_Transportation\_Site

in\_csv <- "../data/IX\_PreK\_Vendors\_by\_Transportation\_Site.csv"

ix\_pvt <- read\_csv(in\_csv)

ix\_pvt %>% glimpse

*## Observations: 1,833*

*## Variables: 5*

*## $ School\_Year "2017-2018", "2017-2018", "2017-2018", "2017-2018"...*

*## $ OPT\_Code "C053", "C104", "C149", "C153", "C157", "C158", "C...*

*## $ Site\_Name "Bank Street Family Center", "PAL WORLD OF CREATIV...*

*## $ School\_Name "Bank Street Family Center", "KIDS CENTRIC, INC", ...*

*## $ Vendor\_Name "PHILLIPS BUS SERVICE", "L & M BUS CORP.", "L & M ...*

And we begin to engage in some 'exploratory' data cleaning. That is, we aren't quite yet sure what we are going to use, and what we aren't going to use. But we know dirt when we see it.

First we try to clean up iv\_routes a little:

#iv\_routes

check\_similar <- function(df){

df %>%

mutate(

diff= stringdist(Vendor\_Affiliation, str\_trunc(Vendor\_Name,str\_length(Vendor\_Affiliation))),

flag\_company\_same = ifelse(diff <= str\_length(Vendor\_Affiliation)/6, 1L, 0L),

diff = NULL

)

}

iv\_routes %>% check\_similar() %>% filter(flag\_company\_same != 1) %>%

distinct(Vendor\_Name, Vendor\_Affiliation)

*## # A tibble: 12 x 2*

*## Vendor\_Name Vendor\_Affiliation*

*##*

*## 1 EMPIRE STATE BUS CORP. ALLIED TRANSIT CORP.*

*## 2 LOGAN BUS COMPANY INC. LITTLE LINDA BUS CO.,INC.*

*## 3 LORINDA ENT. LTD. LITTLE LINDA BUS CO.,INC.*

*## 4 CONSOLIDATED BUS TRANS. INC. BORO TRANSIT, INC.*

*## 5 LORISSA BUS SERVICE INC. BOBBY`S BUS CO. INC.*

*## 6 GRANDPA`S BUS CO., INC. BOBBY`S BUS CO. INC.*

*## 7 LITTLE RICHIE BUS SERVICE LITTLE LINDA BUS CO.,INC.*

*##* [*8 THIRD AVENUE*](https://www.google.com/maps/search/8+THIRD+AVENUE?entry=gmail&source=g) *TRANSIT JOFAZ TRANSPORTATION INC.*

*## 9 LOGAN TRANSPORTATION SYSTEMS BOBBY`S BUS CO. INC.*

*## 10 CONSOLIDATED BUS TRANSIT, INC. BORO TRANSIT, INC.*

*## 11 LORINDA ENTERPRISES, LTD. LITTLE LINDA BUS CO.,INC.*

*##* [*12 THIRD AVENUE*](https://www.google.com/maps/search/12+THIRD+AVENUE?entry=gmail&source=g) *TRANSIT, INC JOFAZ TRANSPORTATION INC.*

#cleaning that needs to be done prior to welding

iv\_cleaned <- iv\_routes %>%

mutate(Route\_Start\_Date = mdy(Route\_Start\_Date),

Garage\_City = tolower(Garage\_City)

) %>%

dplyr::rename(

"Garage\_Street\_Address" = 'Garage \_Street\_Address',

"Garage\_XCoord" = "XCoordinates",

"Garage\_YCoord" = "YCoordinates"

) %>%

check\_similar()

iv\_cleaned %>% glimpse

*## Observations: 42,114*

*## Variables: 15*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016",...*

*## $ Route\_Number "C911", "J499", "J500", "J501", "J502"...*

*## $ Service\_Type "D2D", "D2D", "D2D", "D2D", "D2D", "D2...*

*## $ Vehicle\_TypeDescription "Non-Wheelchair Accessible Alternative...*

*## $ Route\_Start\_Date 2016-03-21, 2015-09-09, 2015-09-01, 2...*

*## $ Vendor\_Code "VN", "RV", "HT", "HT", "HT", "HT", "H...*

*## $ Vendor\_Name "VAN TRANS LLC (B2192)", "RELIANT TRAN...*

*## $ Vendor\_Affiliation "VAN TRANS LLC (B2192)", "RELIANT TRAN...*

*## $ Garage\_Street\_Address "670 Hillside Road", "297 NORMAN AVENU...*

*## $ Garage\_City "pelham manor", "brooklyn", "brooklyn"...*

*## $ Garage\_State "NY", "NY", "NY", "NY", "NY", "NY", "N...*

*## $ Garage\_Zip 10803, 11222, 11224, 11224, 11224, 112...*

*## $ Garage\_XCoord 1034574, 1000787, 983439, 983439, 9834...*

*## $ Garage\_YCoord 265547.6, 204481.0, 148884.0, 148884.0...*

*## $ flag\_company\_same 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...*

Next we join this to v\_rbt, trying as best we can to figure out how they join, and to account for any observations which can't be joined as cleanly. 'School\_Year' and 'Route\_Number' appear to be our main joining headers, but this join leaves out 'OPT\_Code' for some observations. These observations are not represented in the correct school year. We take a guess, and assign these an OPT\_Code based only on the route, hoping that things haven't changed from year to year.

v\_rbt %>% group\_by(School\_Year,Route\_Number) %>% count %>% filter(n>1)

*## # A tibble: 27,712 x 3*

*## # Groups: School\_Year, Route\_Number [27,712]*

*## School\_Year Route\_Number n*

*##*

*## 1 2015-2016 J510 2*

*## 2 2015-2016 J690 2*

*## 3 2015-2016 J700 2*

*## 4 2015-2016 J701 2*

*## 5 2015-2016 J702 2*

*## 6 2015-2016 J703 2*

*## 7 2015-2016 J704 2*

*## 8 2015-2016 J705 2*

*## 9 2015-2016 J706 2*

*## 10 2015-2016 J707 2*

*## # ... with 27,702 more rows*

# There are 27712 School\_Year, Route\_Number combinations in v\_rbt which appear more than once. Why is that?

#Some year/route combinations have more than one OPT\_Code. Why?

#OPT\_Code represents an organization in a specific location. Each Route/School\_Year combination thus can serve multiple organizations in multiple locations.

#This is going to cause us some problems later on, since we will have bus level data with multiple organizations attached. It will probably require some summary operations.

v\_joined <- iv\_cleaned %>%

left\_join(v\_rbt, by = c("School\_Year", "Route\_Number"))

#62 observations have no OPT\_code after this join. Check if the Route\_Numbers can be matched alone

iv\_cleaned %>% anti\_join(v\_rbt, by = c("School\_Year", "Route\_Number")) %>%

distinct(Route\_Number) #59 routes

*## # A tibble: 59 x 1*

*## Route\_Number*

*##*

*## 1 Z003*

*## 2 T003*

*## 3 T004*

*## 4 Q8227*

*## 5 T008*

*## 6 T009*

*## 7 Z001*

*## 8 Z002*

*## 9 Z004*

*## 10 Z005*

*## # ... with 49 more rows*

#question: are all of these routes represented in the v\_rbt, even if the year is wrong?

iv\_cleaned %>% anti\_join(v\_rbt, by = c("School\_Year", "Route\_Number")) %>%

inner\_join(v\_rbt, by = "Route\_Number") %>% distinct(Route\_Number) #39 routes can be filled by ignoring year. Check.

*## # A tibble: 39 x 1*

*## Route\_Number*

*##*

*## 1 Z003*

*## 2 T004*

*## 3 T008*

*## 4 T009*

*## 5 Z001*

*## 6 Z002*

*## 7 Z004*

*## 8 Z005*

*## 9 X2353*

*## 10 X9353*

*## # ... with 29 more rows*

iv\_cleaned %>% anti\_join(v\_rbt, by = "Route\_Number") %>% distinct(Route\_Number) #Yes, 20 routes left for which we have no OPT information.

*## # A tibble: 20 x 1*

*## Route\_Number*

*##*

*## 1 T003*

*## 2 Q8227*

*## 3 X8304*

*## 4 X8338*

*## 5 X8339*

*## 6 X8406*

*## 7 X9874*

*## 8 X9875*

*## 9 X9881*

*## 10 M8068*

*## 11 M8246*

*## 12 M9068*

*## 13 M9246*

*## 14 X8621*

*## 15 X9621*

*## 16 D002*

*## 17 Q2455*

*## 18 Q9455*

*## 19 A002*

*## 20 I001*

#OK, so I have 39 routes which need to join the OPT\_Code on. But I only want to perform this join on the members of v\_joined which are missing OPT data

v\_sub <- v\_joined %>%

filter(is.na(OPT\_Code)) %>%

mutate(OPT\_Code = NULL) %>%

left\_join(v\_rbt %>% select(Route\_Number,OPT\_Code), by = "Route\_Number")

v\_corrected <- v\_joined %>%

filter(!is.na(OPT\_Code)) %>%

rbind(v\_sub) #and 20 routes remain without an OPT\_Code

v\_corrected %>% glimpse

*## Observations: 98,632*

*## Variables: 16*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016",...*

*## $ Route\_Number "C911", "J499", "J500", "J501", "J502"...*

*## $ Service\_Type "D2D", "D2D", "D2D", "D2D", "D2D", "D2...*

*## $ Vehicle\_TypeDescription "Non-Wheelchair Accessible Alternative...*

*## $ Route\_Start\_Date 2016-03-21, 2015-09-09, 2015-09-01, 2...*

*## $ Vendor\_Code "VN", "RV", "HT", "HT", "HT", "HT", "H...*

*## $ Vendor\_Name "VAN TRANS LLC (B2192)", "RELIANT TRAN...*

*## $ Vendor\_Affiliation "VAN TRANS LLC (B2192)", "RELIANT TRAN...*

*## $ Garage\_Street\_Address "670 Hillside Road", "297 NORMAN AVENU...*

*## $ Garage\_City "pelham manor", "brooklyn", "brooklyn"...*

*## $ Garage\_State "NY", "NY", "NY", "NY", "NY", "NY", "N...*

*## $ Garage\_Zip 10803, 11222, 11224, 11224, 11224, 112...*

*## $ Garage\_XCoord 1034574, 1000787, 983439, 983439, 9834...*

*## $ Garage\_YCoord 265547.6, 204481.0, 148884.0, 148884.0...*

*## $ flag\_company\_same 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...*

*## $ OPT\_Code 75666, 75760, 75006, 75437, 75437, 759...*

Next we work on vi\_transport, including looking into how unique OPT\_Codes really are. We transform the vehicle descriptions into counts of the various kinds of passengers and attributes of each vehicle. This approach is a kind of 'type agnostic' approach, so that a new vehicle type can be predicted based on its properties, rather than just being a new predictor which the model cannot account for.

vi\_cleaned <- vi\_transport %>%

mutate(City = tolower(City),

OPT\_Code = as.integer(OPT\_Code)

) %>%

dplyr::rename(

"location\_city" = "City",

"location\_state" = "State",

"location\_zip" = "Zip",

"location\_longitude" = "Longitude",

"location\_latitude" = "Latitude"

) %>%

filter(!is.na(OPT\_Code))

#OPT\_Codes with non-duplicate lat/long

vi\_code\_and\_first\_location <- vi\_cleaned %>%

group\_by(OPT\_Code, School\_Year) %>%

distinct(location\_longitude,location\_latitude) %>%

count() %>%

filter(n>1) %>%

inner\_join(vi\_cleaned, by = c("OPT\_Code","School\_Year")) %>%

filter(row\_number() == 1) %>%

select(OPT\_Code, School\_Year, location\_longitude, location\_latitude)

mismatches\_location <- vi\_cleaned %>%

inner\_join(vi\_code\_and\_first\_location, by = c("OPT\_Code", "School\_Year")) %>%

select(OPT\_Code, School\_Year, location\_longitude.x, location\_longitude.y, location\_latitude.x, location\_latitude.y) %>%

mutate(

diff\_longitude = location\_longitude.x - location\_longitude.y,

diff\_latitude = location\_latitude.x - location\_latitude.y

) %>%

filter(diff\_longitude > 0 | diff\_latitude > 0)

#So there are 65 non unique OPT\_Codes, it seems, with 85 that produce duplicates.

#When performed with OPT\_Code and School\_Year, it seems there are no matches.

#So I need make my match by OPT\_Code / School\_Year. OPT\_Codes are changing from year to year.

vi\_joined <- v\_corrected %>%

left\_join(vi\_cleaned, by = c("OPT\_Code", "School\_Year"))

vi\_joined %>%

distinct(Vehicle\_TypeDescription)

*## # A tibble: 10 x 1*

*## Vehicle\_TypeDescription*

*##*

*## 1 Non-Wheelchair Accessible Alternative (NWC)*

*## 2 Mini-Wagon*

*## 3 Type A or B Flex Vehicle*

*## 4 Standard Bus (SE)*

*## 5 Standard Bus (GE)*

*## 6 Ambulance Transportation Service*

*## 7 Hydraulic Lift*

*## 8 Type C or D Flex*

*## 9 Ramp-Wagon*

*## 10 Wheelchair Accessible Alternative (WC)*

#Vehicle\_TypeDescription: Number\_of\_Riders, vehicle\_reg\_seats, vehicle\_disabled\_seats, vehicle\_amblatory\_seats, vehicle\_lift, vehicle\_attendant

vehicle\_typedescription <- data.frame(Vehicle\_TypeDescription =

c("Non-Wheelchair Accessible Alternative (NWC)", "Mini-Wagon", "Type A or B Flex Vehicle", "Standard Bus (SE)", "Standard Bus (GE)","Ambulance Transportation Service", "Hydraulic Lift", "Type C or D Flex","Ramp-Wagon","Wheelchair Accessible Alternative (WC)", "Motor Coach Bus"),

vehicle\_max\_riders = c(3,15,6,36,60,1,16,35,8,4, 55),

vehicle\_reg\_seats = c(3,15,6,36,60,0,0,35,0,0, 55),

vehicle\_disabled\_seats = c(0,0,6,0,0,0,8,10,4,2, 0),

vehicle\_ambulatory\_seats = c(0,0,0,0,0,1,8,0,4,2, 0),

vehicle\_lift = as.integer(c(0,0,1,0,0,0,1,1,0,1, 0)),

vehicle\_attendant = as.integer(c(1,1,1,0,0,0,1,1,1,1,0)),

vehicle\_emt = c(0,0,0,0,0,2,0,0,0,0,0),

vehicle\_bathroom = c(0,0,0,0,0,0,0,0,0,0,1)

) %>%

mutate(

Vehicle\_TypeDescription = as.character(Vehicle\_TypeDescription)

) %>%

arrange(vehicle\_max\_riders)

# c("Non-Wheelchair Accessible Alternative (NWC)", 3, 3, 0, 0, 0, 1),

# c("Mini-Wagon", 15, 15, 0, 0, 0, 1),

# c("Type A or B Flex Vehicle", 6, 6, 6, 0, 1, 1),

# c("Standard Bus (SE)", 36, 36, 0, 0, 0, 0),

# c("Standard Bus (GE)", 60, 60, 0, 0, 0, 0),

# c("Ambulance Transportation Service", 1, 0, 0, 1, 0, 1),

# c("Hydraulic Lift", 16, 0, 8, 8, 1, 1),

# c("Type C or D Flex", 35, 35, 10, 0, 1, 1),

# c("Ramp-Wagon", 8, 0, 4, 4, 0, 1),

# c("Wheelchair Accessible Alternative (WC)", 4, 0, 2, 2, 1, 1))

vi\_vehicle\_details <- vi\_joined %>%

left\_join(vehicle\_typedescription, by = "Vehicle\_TypeDescription")

#Worth noting that Number\_of\_Riders from ix\_joined has no analogue in this data set. I have that data only for the PreK data, it seems

#Put the garage count by Vendor\_Code, School\_Year

garage\_count\_abs <- v\_corrected %>%

select(School\_Year, Vendor\_Code, Garage\_XCoord, Garage\_YCoord) %>%

group\_by(School\_Year, Vendor\_Code) %>%

distinct(Garage\_XCoord, Garage\_YCoord) %>%

count() %>%

mutate(

garage\_max = n,

garage\_min = n,

n = NULL

)

vi\_garage\_count <- vi\_vehicle\_details %>%

left\_join(garage\_count\_abs, by = c("School\_Year", "Vendor\_Code")) %>%

mutate(

OPT\_Code = as.character(OPT\_Code)

)

vi\_garage\_count %>% glimpse

*## Observations: 98,632*

*## Variables: 42*

*## $ School\_Year "2015-2016", "2015-2016", "2015-...*

*## $ Route\_Number "C911", "J499", "J500", "J501", ...*

*## $ Service\_Type "D2D", "D2D", "D2D", "D2D", "D2D...*

*## $ Vehicle\_TypeDescription "Non-Wheelchair Accessible Alter...*

*## $ Route\_Start\_Date 2016-03-21, 2015-09-09, 2015-09...*

*## $ Vendor\_Code "VN", "RV", "HT", "HT", "HT", "H...*

*## $ Vendor\_Name "VAN TRANS LLC (B2192)", "RELIAN...*

*## $ Vendor\_Affiliation "VAN TRANS LLC (B2192)", "RELIAN...*

*## $ Garage\_Street\_Address "670 Hillside Road", "297 NORMAN...*

*## $ Garage\_City "pelham manor", "brooklyn", "bro...*

*## $ Garage\_State "NY", "NY", "NY", "NY", "NY", "N...*

*## $ Garage\_Zip 10803, 11222, 11224, 11224, 1122...*

*## $ Garage\_XCoord 1034574, 1000787, 983439, 983439...*

*## $ Garage\_YCoord 265547.6, 204481.0, 148884.0, 14...*

*## $ flag\_company\_same 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...*

*## $ OPT\_Code "75666", "75760", "75006", "7543...*

*## $ Name "CARMEL ACADEMY", "THE SINAI SCH...*

*## $ Affiliation "Yeshiva", "Yeshiva", "Nonsectar...*

*## $ Site\_Type "School", "School", "School", "S...*

*## $ Street\_Address "*[*270 LAKE AVENUE*](https://www.google.com/maps/search/270+LAKE+AVENUE?entry=gmail&source=g)*", "110 SOUTH OR...*

*## $ location\_city "connecticut", "new jersey", "ne...*

*## $ location\_state "CT", "NJ", "NJ", "NJ", "NJ", "N...*

*## $ location\_zip 6830, 7039, 7083, 7666, 7666, 87...*

*## $ location\_longitude -73.63757, -74.36126, -74.27827,...*

*## $ location\_latitude 41.04900, 40.77635, 40.70391, 40...*

*## $ Door\_To\_Door\_Service "Yes", "Yes", "Yes", "Yes", "Yes...*

*## $ Stop\_To\_School\_Service "No", "No", "No", "No", "No", "N...*

*## $ Common\_Carrier\_Svc\_Metrocards "Yes", "No", "No", "No", "No", "...*

*## $ Site\_Reimbursement "No", "No", "No", "No", "No", "N...*

*## $ Mid\_Day\_Service "No", "No", "No", "No", "No", "N...*

*## $ D2D\_Late\_Day\_Programs "No", "No", "No", "No", "No", "N...*

*## $ S2S\_Late\_Day\_Programs "No", "No", "No", "No", "No", "N...*

*## $ vehicle\_max\_riders 3, 15, 3, 3, 3, 3, 3, 15, 3, 3, ...*

*## $ vehicle\_reg\_seats 3, 15, 3, 3, 3, 3, 3, 15, 3, 3, ...*

*## $ vehicle\_disabled\_seats 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...*

*## $ vehicle\_ambulatory\_seats 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...*

*## $ vehicle\_lift 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...*

*## $ vehicle\_attendant 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...*

*## $ vehicle\_emt 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...*

*## $ vehicle\_bathroom 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...*

*## $ garage\_max 3, 4, 2, 2, 2, 2, 2, 4, 2, 2, 2,...*

*## $ garage\_min 3, 4, 2, 2, 2, 2, 2, 4, 2, 2, 2,...*

When I initially worked through this data, I tinkered with the pre-K data, ix\_pvt, quite a lot. But in the end, it turned out that this data set only really existed as a way to join vehicle and staff metadata to the main data set via OPT\_Code. The issue is that the school- and pre-K-level data don't share headers. So pulling, say, the garage information from the school-level data leaves NA values in the pre-K.

The driver and vehicle level data is, in the end, what I am after. The school and pre-K data is really just a way of attaching that metadata to the main data. Here is my cleaning work.

#---drivers and attendants, and vehicles, by vendor name---

#NR: next, I want to join on the driver numbers and vehicle numbers metadata.

#This must be joined by Vendor Name and School Year, and might require fuzzy joining.

iii\_drivers %>% glimpse #832 obs

*## Observations: 832*

*## Variables: 5*

*## $ School\_Year "2018-2019", "2018-2019", "2018-2019", "2018-...*

*## $ Vendor\_Name "ACADEMY EXPRESS LLC", "E-Z BUS LLC (B2321)",...*

*## $ Type\_of\_Service "Coach", "School Age", "School Age", "School ...*

*## $ Active\_Employees 65, 5, 3, 2, 1, 3, 6, 226, 22, 18, 13, 2, 20,...*

*## $ Job\_Type "Driver", "Attendant", "Driver", "Driver&Atte...*

vendors\_unique <- vi\_garage\_count %>%

select(Vendor\_Name) %>%

rbind(ix\_pvt %>% select(Vendor\_Name)) %>%

distinct(Vendor\_Name) %>%

arrange(Vendor\_Name) #87 unique vendors

vendors\_unique %>%

anti\_join(iii\_drivers %>% filter(Type\_of\_Service != "Coach"), by = "Vendor\_Name") %>%

glimpse

*## Observations: 1*

*## Variables: 1*

*## $ Vendor\_Name "ROBIN TRANSPORTATION"*

#only "Robin Transportation" isn't covered by either iii or vii. Check check.

#All vendors are covered by both iii\_ and vii\_ \*except\* ROBIN. Goddamn it, Robin.

#iii\_drivers

#needs turning into tidy data. What do I mean by this?

#We are after company level information.

#Per company and school year, we want the following counts

#number of coach services

#number of school services

#number of pre-k services

#number of ambulance services

#number of attendant services

#number of drivers (count doubles)

#number of attendants (count doubles)

#number of EMTs (count doubles)

#number of paramedics (count doubles)

#service counts

iii\_service\_count <- iii\_drivers %>%

group\_by(Vendor\_Name, School\_Year) %>%

summarise(

drivers\_numServ\_coach = sum(Type\_of\_Service == "Coach"),

drivers\_numServ\_school = sum(Type\_of\_Service == "School Age"),

drivers\_numServ\_prek = sum(Type\_of\_Service == "Pre-K"),

drivers\_numServ\_amb = sum(Type\_of\_Service == "Ambulance"),

drivers\_numServ\_att = sum(Type\_of\_Service == "Attendant")

)

#staff counts

iii\_staff\_count <- iii\_drivers %>%

spread(Job\_Type, Active\_Employees, fill = 0) %>%

group\_by(Vendor\_Name, School\_Year) %>%

summarise(

drivers\_total\_driver = sum(Driver + `Driver&Attendant` + `Driver&EMT`+ `Driver&Paramedic`),

drivers\_total\_attendant = sum(`Driver&Attendant` + Attendant),

drivers\_total\_EMT = sum(`Driver&EMT` + EMT),

drivers\_total\_paramedic = sum(`Driver&Paramedic` + Paramedic)

)

#counts left: staff per service, and each category of staff per service.

#I'm only going to count drivers per service

#drivers per service

iii\_drivers\_per\_service <- iii\_drivers %>%

spread(Job\_Type, Active\_Employees, fill = 0) %>%

mutate(

drivers\_total\_driver = Driver + `Driver&Attendant` + `Driver&EMT`+ `Driver&Paramedic`

) %>%

group\_by(Vendor\_Name, School\_Year, Type\_of\_Service) %>%

summarise(

drivers\_dummy = sum(drivers\_total\_driver)

) %>%

ungroup() %>%

spread(Type\_of\_Service, drivers\_dummy, fill = 0) %>%

dplyr::rename(

"drivers\_num\_servAmb" = "Ambulance",

"drivers\_num\_servAttend" = "Attendant",

"drivers\_num\_servCoach" = "Coach",

"drivers\_num\_servPreK" = "Pre-K",

"drivers\_num\_servSchool" = "School Age"

)

#total staff per service

iii\_total\_staff\_per\_service <- iii\_drivers %>%

group\_by(Vendor\_Name, School\_Year, Type\_of\_Service) %>%

summarise(

drivers\_total\_dummy = sum(Active\_Employees)

) %>%

spread(Type\_of\_Service, drivers\_total\_dummy, fill = 0) %>%

dplyr::rename(

"drivers\_staff\_servAmb" = "Ambulance",

"drivers\_staff\_servAttend" = "Attendant",

"drivers\_staff\_servCoach" = "Coach",

"drivers\_staff\_servPreK" = "Pre-K",

"drivers\_staff\_servSchool" = "School Age"

)

iii\_joined <- iii\_service\_count %>%

left\_join(iii\_staff\_count, by = c("Vendor\_Name", "School\_Year")) %>%

left\_join(iii\_drivers\_per\_service, by = c("Vendor\_Name", "School\_Year")) %>%

left\_join(iii\_total\_staff\_per\_service, by = c("Vendor\_Name", "School\_Year"))

#vii\_vehicles

vii\_vehicles %>% glimpse #764 obs

*## Observations: 764*

*## Variables: 4*

*## $ School\_Year "2015-2016", "2015-2016", "2015-2016",...*

*## $ Vendor\_Name "ACADEMY EXPRESS LLC", "ACME BUS CORP....*

*## $ Vehicle\_TypeDescription "Motor Coach Bus", "Mini-Wagon", "Mini...*

*## $ Active\_Vehicles 22, 71, 6, 62, 1, 9, 71, 6, 63, 47, 67...*

#total vehicles of each kind is actually not relevant.

#total seats of each kind

vii\_vehicle\_counts <- vii\_vehicles %>%

left\_join(vehicle\_typedescription, by = "Vehicle\_TypeDescription") %>%

mutate(

vehicle\_total\_max\_riders = Active\_Vehicles \* vehicle\_max\_riders,

vehicle\_total\_reg\_seats = Active\_Vehicles \* vehicle\_reg\_seats,

vehicle\_total\_disabled\_seats = Active\_Vehicles \* vehicle\_disabled\_seats,

vehicle\_total\_ambulatory\_seats = Active\_Vehicles \* vehicle\_ambulatory\_seats,

vehicle\_total\_with\_lifts = Active\_Vehicles \* vehicle\_lift,

vehicle\_total\_with\_attendants = Active\_Vehicles \* vehicle\_attendant,

vehicle\_total\_with\_emts = Active\_Vehicles \* vehicle\_emt,

vehicle\_total\_with\_bathrooms = Active\_Vehicles \* vehicle\_bathroom

) %>%

mutate(

vehicle\_max\_riders = NULL,

vehicle\_reg\_seats = NULL,

vehicle\_disabled\_seats = NULL,

vehicle\_ambulatory\_seats = NULL,

vehicle\_lift = NULL,

vehicle\_attendant = NULL,

vehicle\_emt = NULL,

vehicle\_bathroom = NULL

) %>%

#spread(Vehicle\_TypeDescription, Active\_Vehicles, fill = 0) %>%

group\_by(Vendor\_Name, School\_Year) %>%

summarise\_if(is.numeric,

sum

)

In my actual workflow, I wrote out all of my intermediate outputs as files that I could keep, and joined them in a separate workbook. In lieu of that process, the code below the fold simply renames some of the data sets that we have developed above. Apologies for any confusion: I promise that my project organization is a little less chaotic than this reporting markdown presentation.

vi\_school <- vi\_garage\_count

iii\_staff <- iii\_joined

vii\_vehicles <- vii\_vehicle\_counts

**Joining the data**

After a lot of messing around, we are at the stage where we can join the company level staff and vehicle data onto the main data set. We also use the times to make variables that tell us how close to either rush hour we are.

in\_csv <- "../output/intermediate/ii\_spread.csv"

ii\_spread <- read\_csv(in\_csv)

#Vendor\_Name, School\_Year for iii\_staff and vii\_vehicles

ix\_module <- ix\_pvt %>%

mutate(Service\_Type = "D2D") %>%

select(Vendor\_Name, School\_Year, OPT\_Code, School\_Name, Service\_Type) %>%

inner\_join(iii\_staff, by = c("Vendor\_Name", "School\_Year")) %>%

inner\_join(vii\_vehicles, by = c("Vendor\_Name", "School\_Year"))

#only Robin transportation can't be joined from ix

#The MONTAUK data needs porting over to 2018-2019 from 2017-2018.

vii\_vehicles <- vii\_vehicles %>%

filter(str\_detect(Vendor\_Name,"MONTAUK STUDENT TRANS, INC.") & School\_Year == "2017-2018") %>%

mutate(School\_Year = "2018-2019") %>%

rbind(vii\_vehicles)

iii\_staff <- iii\_staff %>%

filter(str\_detect(Vendor\_Name,"MONTAUK STUDENT TRANS, INC.") & School\_Year == "2017-2018") %>%

mutate(School\_Year = "2018-2019") %>%

rbind(iii\_staff)

vi\_module <- vi\_school %>%

distinct(Route\_Number, Vendor\_Name, School\_Year, Service\_Type) %>%

inner\_join(iii\_staff, by = c("Vendor\_Name", "School\_Year")) %>%

inner\_join(vii\_vehicles, by = c("Vendor\_Name", "School\_Year"))

#Now I need to look at joining these onto the main data set.

#---main join for basic fit

ii\_ix <- ii\_spread %>%

inner\_join(ix\_module, by = c("Schools\_Serviced" = "OPT\_Code", "School\_Year", "Bus\_Company\_Name" = "Vendor\_Name")) #31 421

ii\_vi <- ii\_spread %>%

inner\_join(vi\_module, by = c("School\_Year", "Route\_Number")) #190 067

#I need to put time occured back in the data set.

#Anonymise company, school, route, and so on. This can be used in another, more detailed join some other time.

ii\_joined <- ii\_vi %>%

select(-Route\_Number, -School\_Year, -Schools\_Serviced, -Bus\_Company\_Name, -Vendor\_Name) %>%

rbind(

ii\_ix %>%

select(-Route\_Number, -School\_Year, -Schools\_Serviced, -Bus\_Company\_Name, -School\_Name)

) %>%

mutate(

service\_type\_d2d = (Service\_Type == "D2D")\*1L,

Service\_Type = NULL

)

tz(ii\_joined$Occurred\_On) <- "America/New\_York" #it's set to UTC

#remove zero variance columns

zero\_var <- function(dat) {

out <- lapply(dat, function(x) length(unique(x)))

keep <- rownames(data.frame(which(out > 1)))

dat %>%

select(keep)

}

ii\_times <- ii\_joined %>%

mutate(

time = strftime(Occurred\_On, format="%H:%M:%S", tz = "America/New\_York"),

time\_am = lubridate::am(hms(time))\*1L,

rush\_within = ((hms(time) >= hms("07:00:00") & hms(time) <= hms("09:00:00")) | (hms(time) >= hms("16:00:00") & hms(time) <= hms("18:00:00")))\*1L,

rush\_between = (hms(time) >= hms("09:00:00") & hms(time) <= hms("16:00:00"))\*1L,

rush\_outside = (hms(time) <= hms("07:00:00") | hms(time) >= hms("18:00:00"))\*1L,

#rush\_time\_from\_peak = min(abs(as.numeric(hms(time) - hms("08:00:00"))), abs(as.numeric(hms(time) - hms("17:00:00"))))/60

rush\_min\_from\_peak = pmin(abs(as.numeric(hms(time) - hms("08:00:00"))),abs(as.numeric(hms(time) - hms("17:00:00"))))/60,

time = NULL,

Occurred\_On = NULL,

time\_delayed = time\_delayed - time\_diff\_report,

time\_diff\_report = NULL

) %>%

zero\_var

ii\_times %>% glimpse

*## Observations: 221,375*

*## Variables: 48*

*## $ Busbreakdown\_ID 1212699, 1212701, 1212703, 121...*

*## $ Has\_Contractor\_Notified\_Schools 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...*

*## $ Has\_Contractor\_Notified\_Parents 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, ...*

*## $ Have\_You\_Alerted\_OPT 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...*

*## $ Number\_Of\_Students\_On\_The\_Bus 0, 0, 1, 0, 9, 0, 2, 3, 3, 4, ...*

*## $ time\_delayed 23, 28, 13, 21, 80, 9, 9, 24, ...*

*## $ reported\_before\_resolved 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...*

*## $ School\_Age 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...*

*## $ Reason\_Accident 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_DelayedbySchool 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_FlatTire 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_HeavyTraffic 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, ...*

*## $ Reason\_LatereturnfromFieldTrip 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_MechanicalProblem 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_ProblemRun 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...*

*## $ Reason\_WeatherConditions 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Reason\_WontStart 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_Bronx 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_Brooklyn 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, ...*

*## $ Boro\_Connecticut 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_Manhattan 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...*

*## $ Boro\_NassauCounty 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, ...*

*## $ Boro\_NewJersey 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_Queens 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, ...*

*## $ Boro\_RocklandCounty 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_StatenIsland 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ Boro\_Westchester 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ drivers\_numServ\_school 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...*

*## $ drivers\_numServ\_prek 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ drivers\_total\_driver 567, 185, 22, 567, 105, 567, 2...*

*## $ drivers\_total\_attendant 383, 196, 14, 383, 131, 383, 3...*

*## $ drivers\_num\_servPreK 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ drivers\_num\_servSchool 567, 185, 22, 567, 105, 567, 2...*

*## $ drivers\_staff\_servPreK 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ drivers\_staff\_servSchool 949, 378, 35, 949, 220, 949, 5...*

*## $ Active\_Vehicles 699, 204, 20, 699, 104, 699, 3...*

*## $ vehicle\_total\_max\_riders 23024, 6810, 681, 23024, 1577,...*

*## $ vehicle\_total\_reg\_seats 20472, 6810, 681, 20472, 1305,...*

*## $ vehicle\_total\_disabled\_seats 1276, 126, 0, 1276, 136, 1276,...*

*## $ vehicle\_total\_ambulatory\_seats 1276, 0, 0, 1276, 136, 1276, 2...*

*## $ vehicle\_total\_with\_lifts 81, 21, 0, 81, 17, 81, 13, 0, ...*

*## $ vehicle\_total\_with\_attendants 258, 21, 3, 258, 104, 258, 321...*

*## $ service\_type\_d2d 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...*

*## $ time\_am 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...*

*## $ rush\_within 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, ...*

*## $ rush\_between 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...*

*## $ rush\_outside 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, ...*

*## $ rush\_min\_from\_peak 93, 75, 65, 65, 41, 20, 15, 5,...*