Ideas for final slides:

Part 1: objective and assumptions - Same as EDA Presentation ( mention recommendations/ suggestions for company too)

Part 2: Cluster Analysis, Summary Statistics for selected 50 stations - Overview of dataset, the clustering process and final dataset we are using (choose random forest because we are doing multiple clustering)

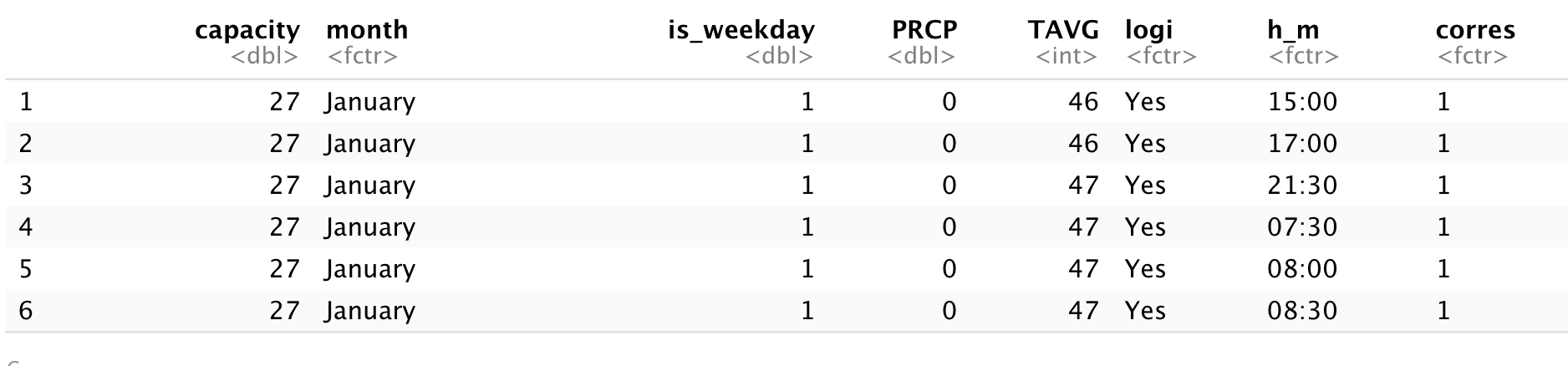
Part 3: Model building: random forest model

A non-technical summary of the methodology used to create a prediction model

RF: 1 ppt - Introduce random forest, the reason why we are using random forest

A summary of the accuracy of the prediction model - Plot, confusion matrix, the estimated accuracy

A plan to maintain the model going forward:

Based on test: 

The current dataset used for training is 2019’s data.

Decide time intervals to re-train the model (1 month) to update the model based on recent trends (2020 and 2021 models will act differently due to COVID):

* Window of 1 year (keep adding 1 month recent data and deleting old 1 month), keep retraining the model every 1 month
* [August 2020-August 2021] -> [Sep 2020 - Sep 2021] -> so on..
* Ensure that all relevant variables are tracked consistently (e.g. 2020 March stops tracking the bike\_id per trip)

-

Actionable recommendations describing how the model can be used to lower

operational costs or increase revenue:

Model to predict availability at a station: identify stations that have low/high/not enough availabilities around this area

We want to fill the needs for each cluster, so each cluster will have the different needs while minimizing cost for each stations’ reshuffling ( reduce unnecessary reshuffling, and increase necessary reshuffling)

Next steps for a phase 2 project : Expand our models in cluster 1 to other stations, and maybe other clusters to get a more general idea of reshuffling needs

A technical appendix that explains the statistical details of the methodology used

in the prediction model