Feature Engineering

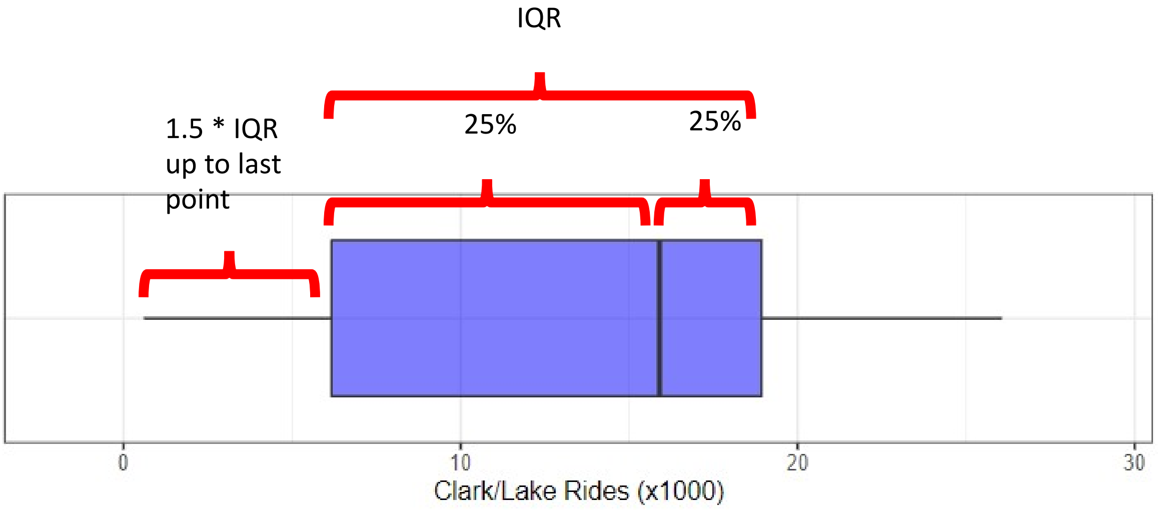
Chapter 4

Stephen Kimel

Chicago Train Ridership

- “Two common mistakes of misunderstanding demand could be made. At one extreme, having too few cars on a line to meet weekday demand would delay riders from reaching their destination and would lead to overcrowding and tension. At the other extreme, having too many cars on the weekend would be inefficient leading to higher operational costs and lower profitability. Good forecasts of demand would help the CTA to get closer to optimally meeting demand.”
- “Our illustration will narrow to predicting daily ridership at the Clark/Lake stop.”

Boxplot

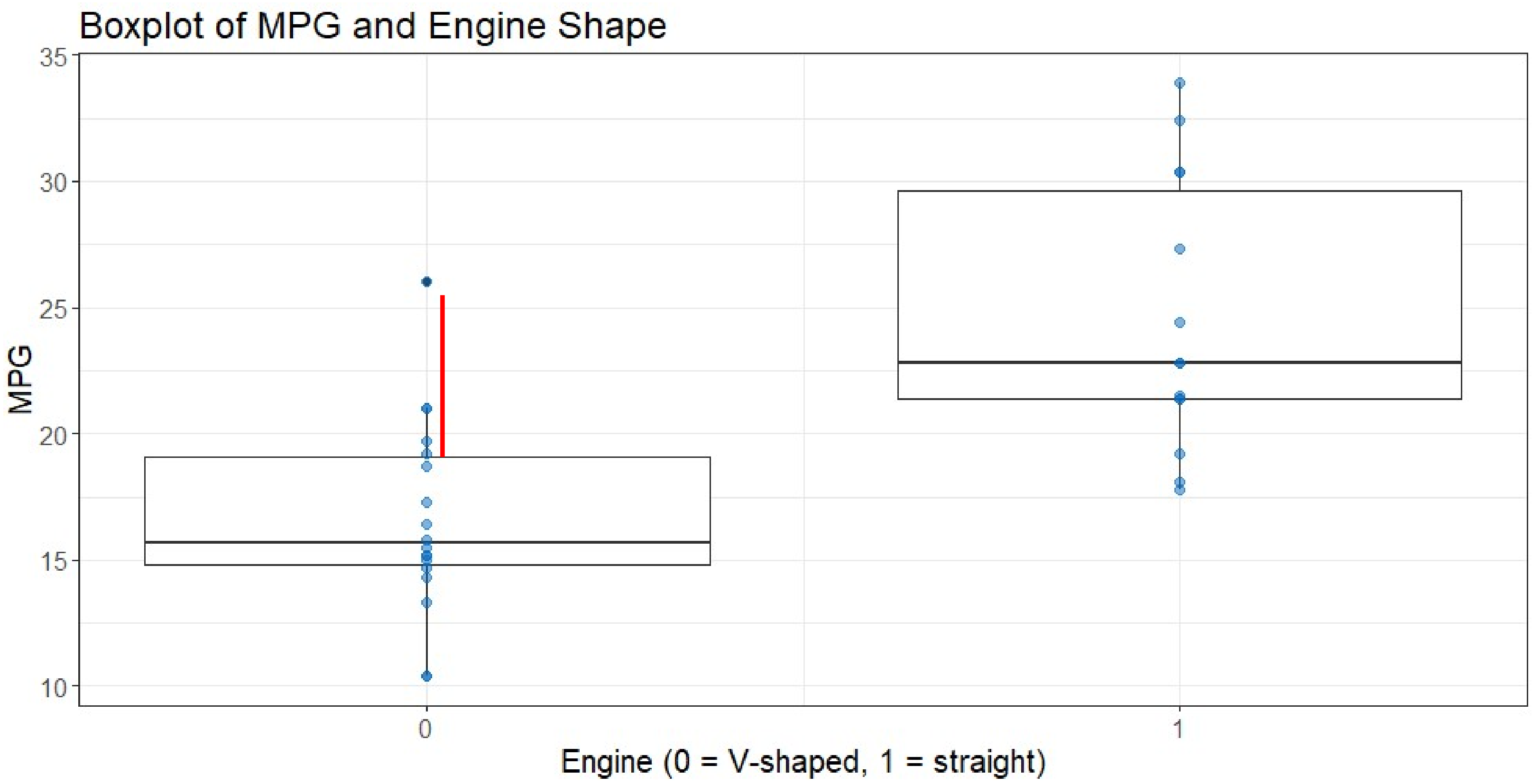


IQR

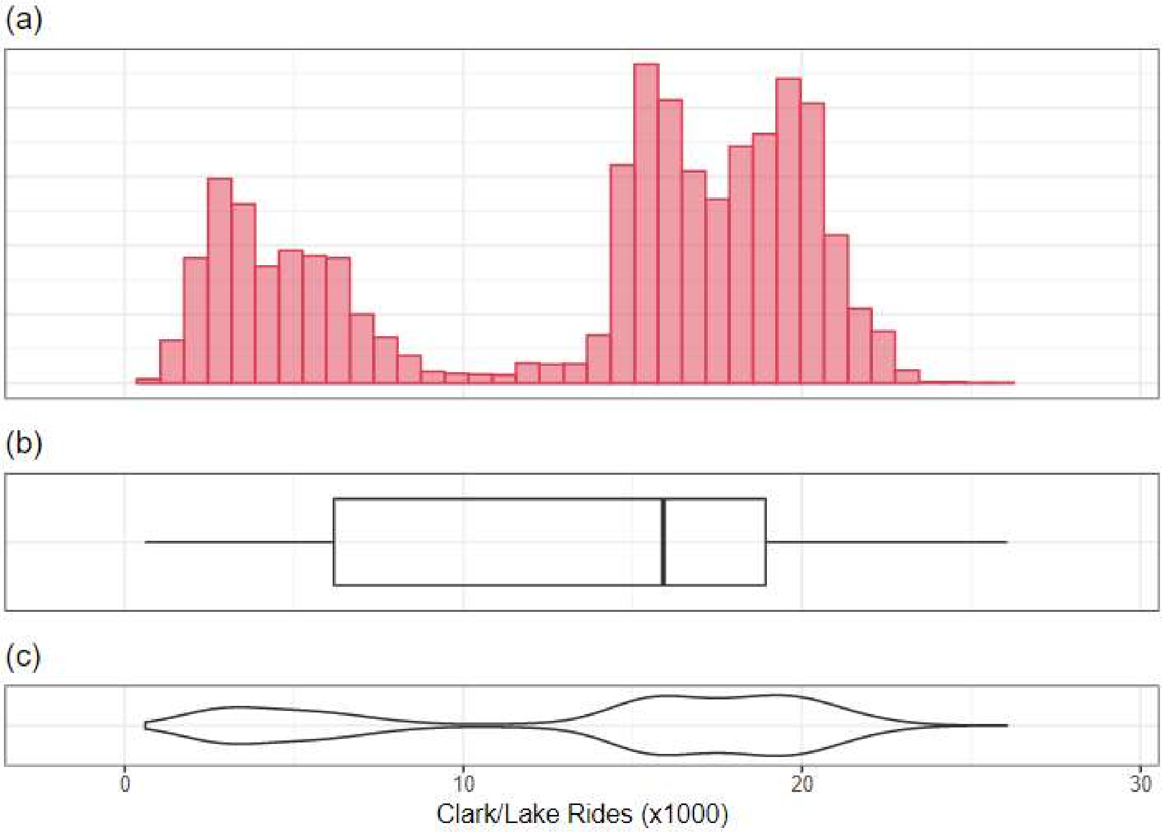
25%

25%

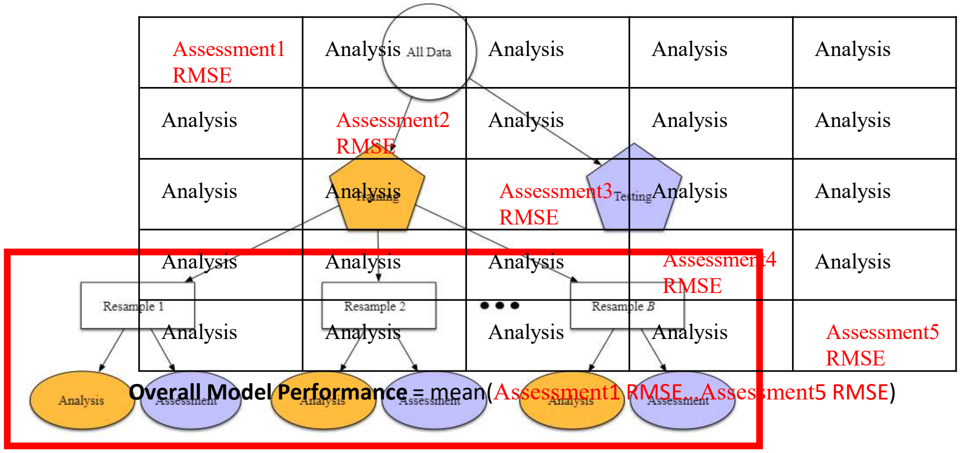
1.5 * IQR up to last point


Boxplot of MPG and Engine Shape
MPG
35
30
25
20
15
10
0
1
Engine (0 = V-shaped, 1 = straight)

Histogram vs. Bloxplot vs. Violinplot



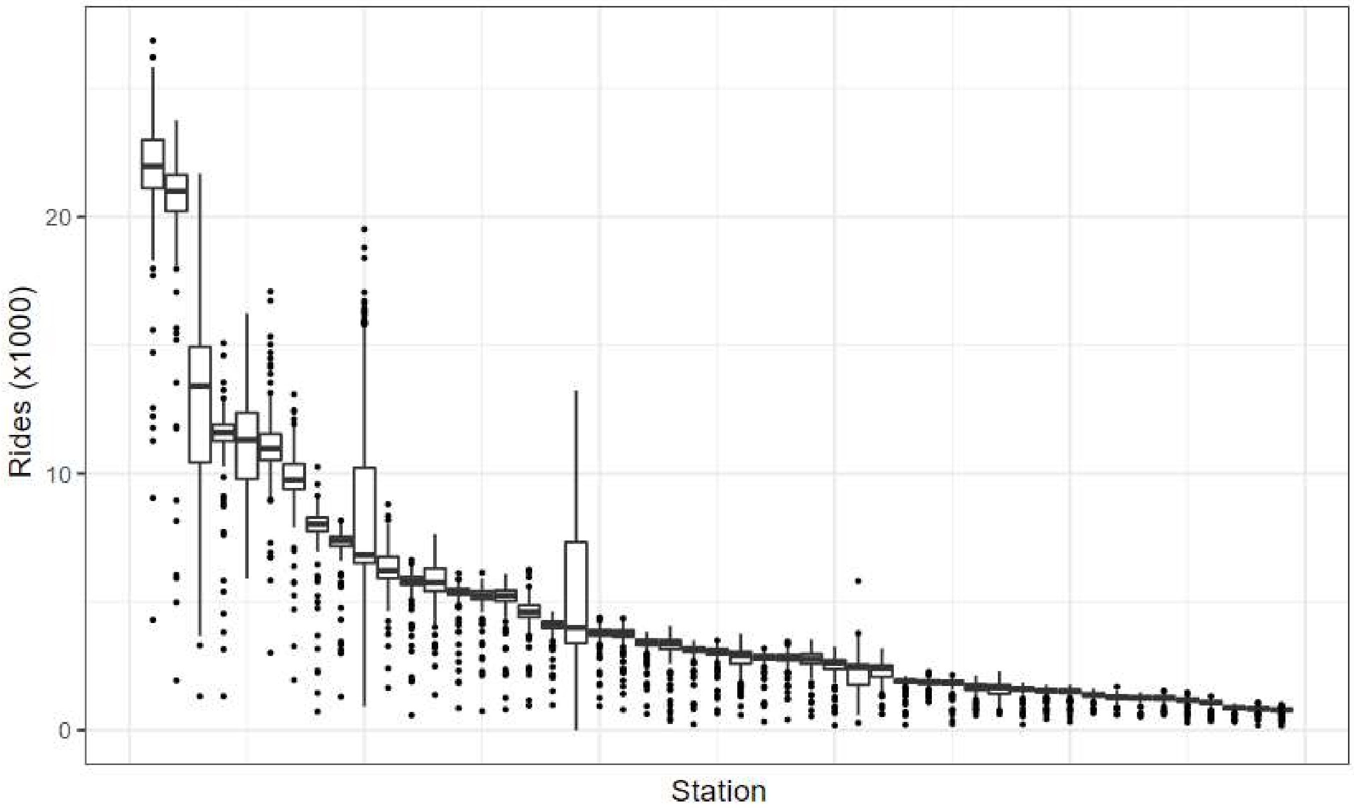
- “Given the range of the daily ridership numbers, there was some question as to whether the **outcome** should be modeled in **the natural units or on the log scale**. On one hand, the **natural units makes interpretation of the results easier** since the RMSE would be in terms of riders. However, if the **outcome were transformed** prior to modeling, it would ensure that ***negative*** ridership **could not be predicted**. The **bimodal nature** of these data, as well as distributions of ridership for each year that have **a longer tail on the right** made this decision difficult. In the end, a **handful of models were fit both ways** to make the determination. The **models** computed in the **natural units** appeared to have slightly **better performance** and, for this reason, all models were analyzed in the natural units.”



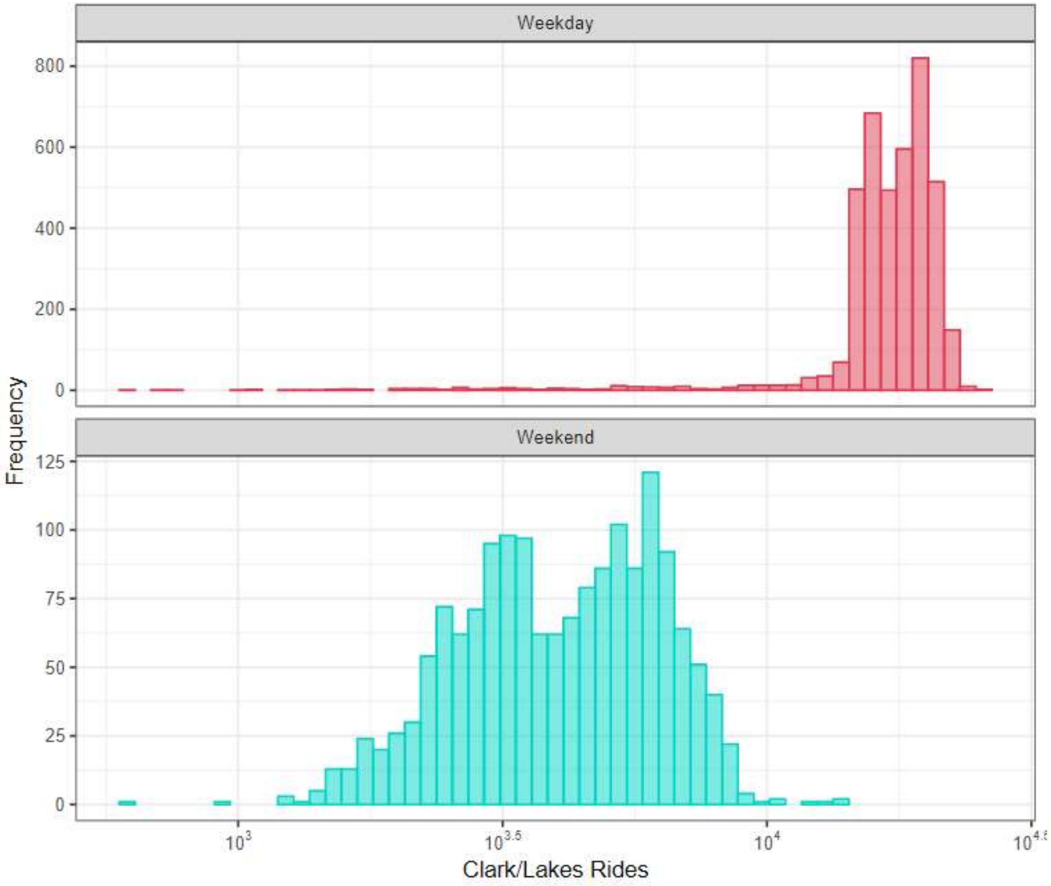
Assessment1 RMSE	Analysis	Analysis	Analysis	Analysis
Analysis	Assessment2 RMSE	Analysis	Analysis	Analysis
Analysis	Analysis	Assessment3 RMSE	Analysis	Analysis
Analysis	Analysis	Analysis	Assessment4 RMSE	Analysis
Analysis	Analysis	Analysis	Analysis	Assessment5 RMSE

Overall Model Performance = mean(Assessment1 RMSE…Assessment5 RMSE)

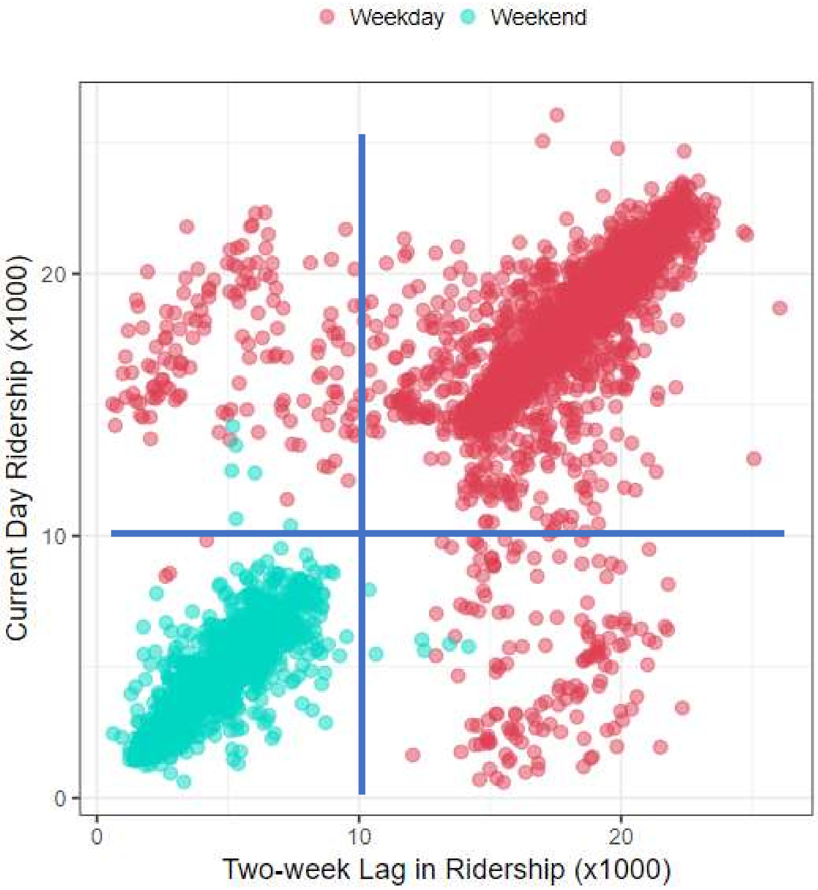
Observing Predictor Variables with Ridership



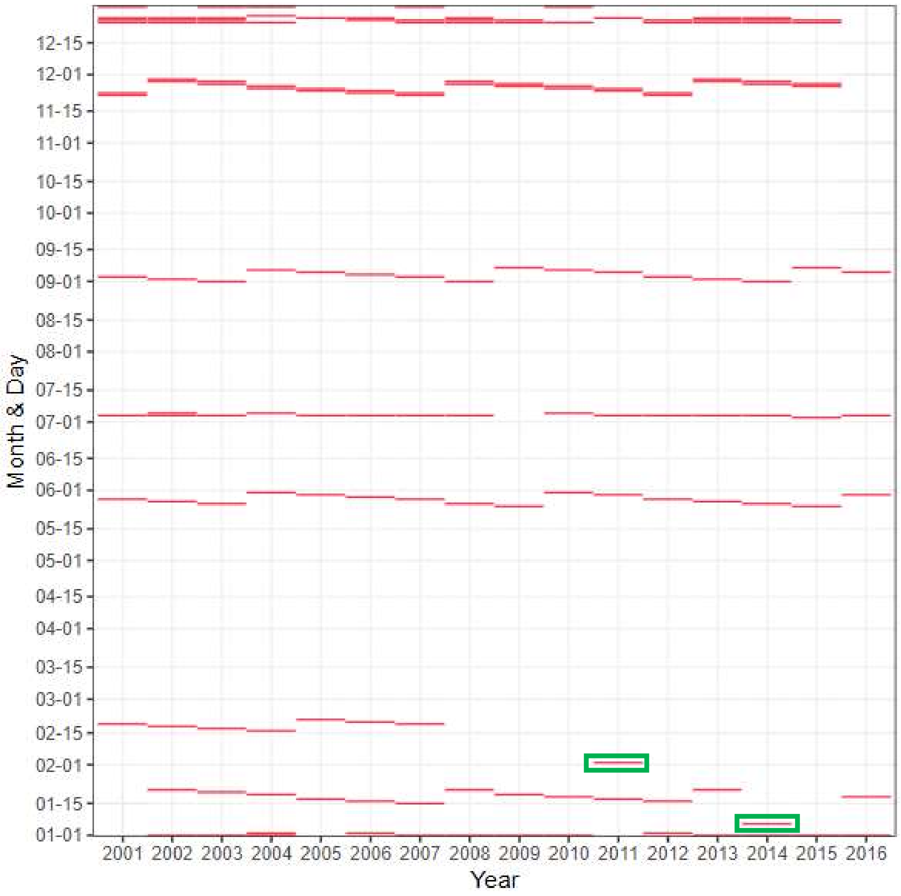
Faceting


Weekday
800
600
400
200
0
Weekend
125
100
75
50
25
0
Frequency
10^3
$10^{3.5}$
10^4
$10^{4.5}$
Clark/Lakes Rides

Scatterplots

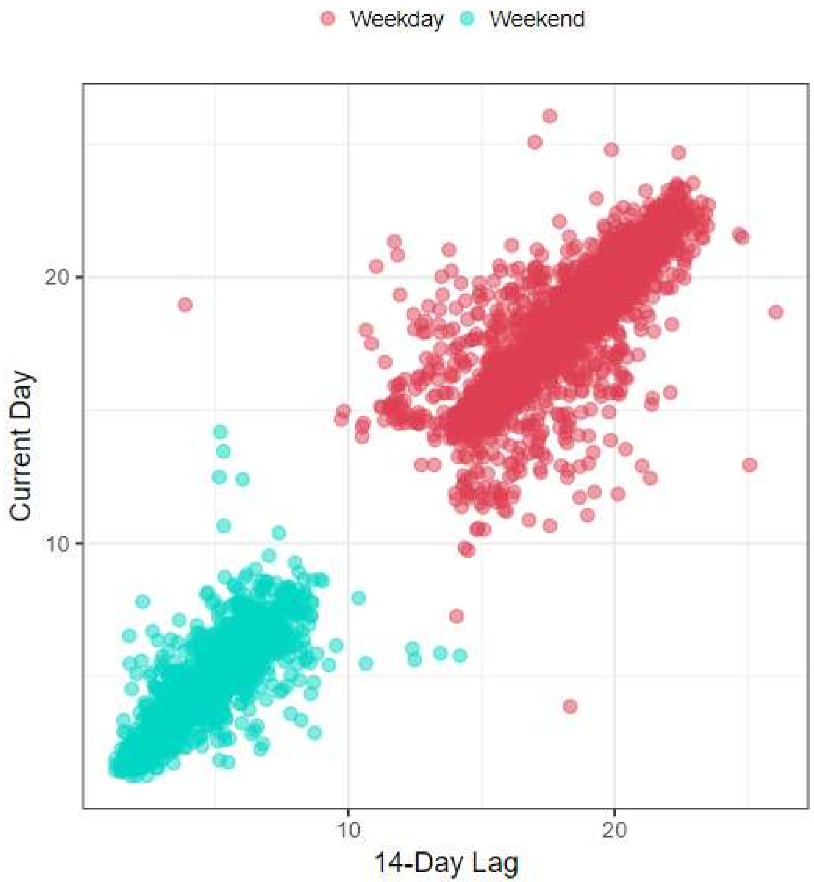


Heatmap

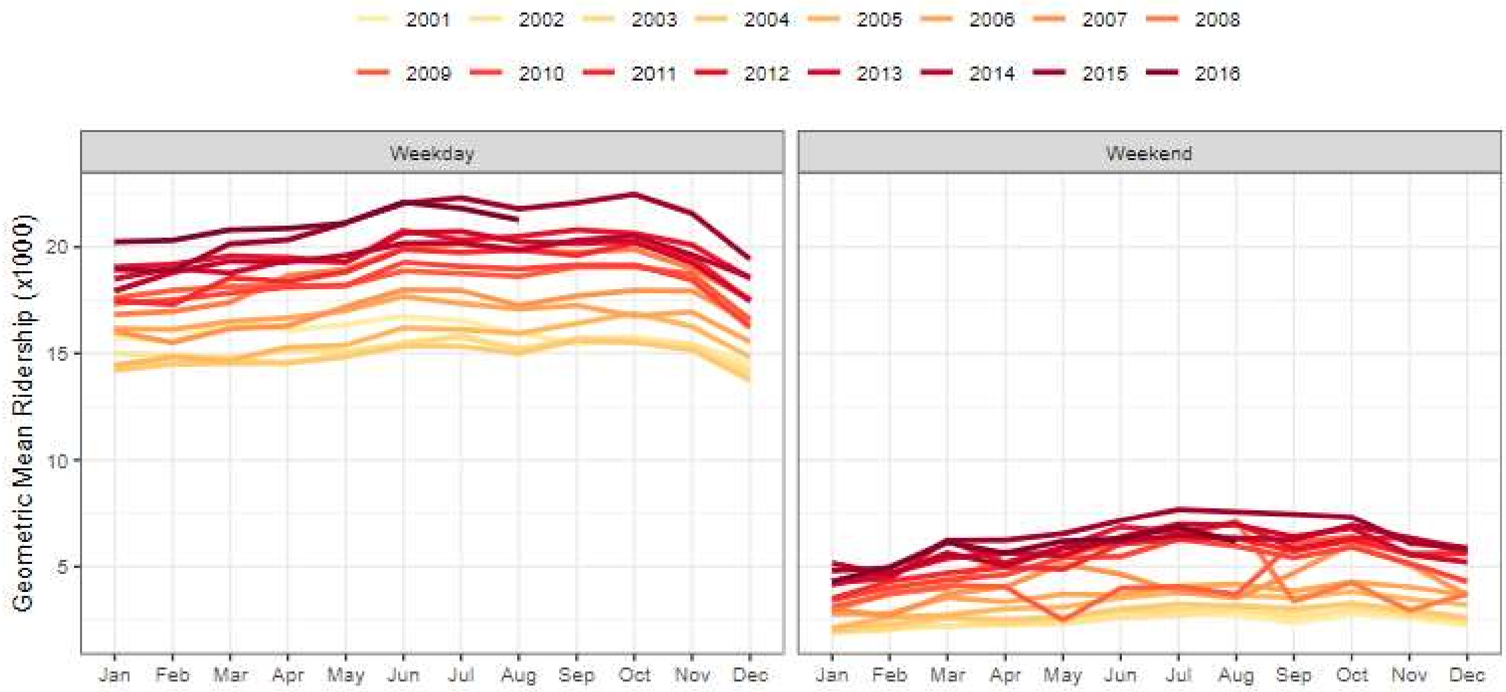


Month & Day
12-15
12-01
11-15
11-01
10-15
10-01
09-15
09-01
08-15
08-01
07-15
07-01
06-15
06-01
05-15
05-01
04-15
04-01
03-15
03-01
02-15
02-01
01-15
01-01
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year

After Some Feature Engineering



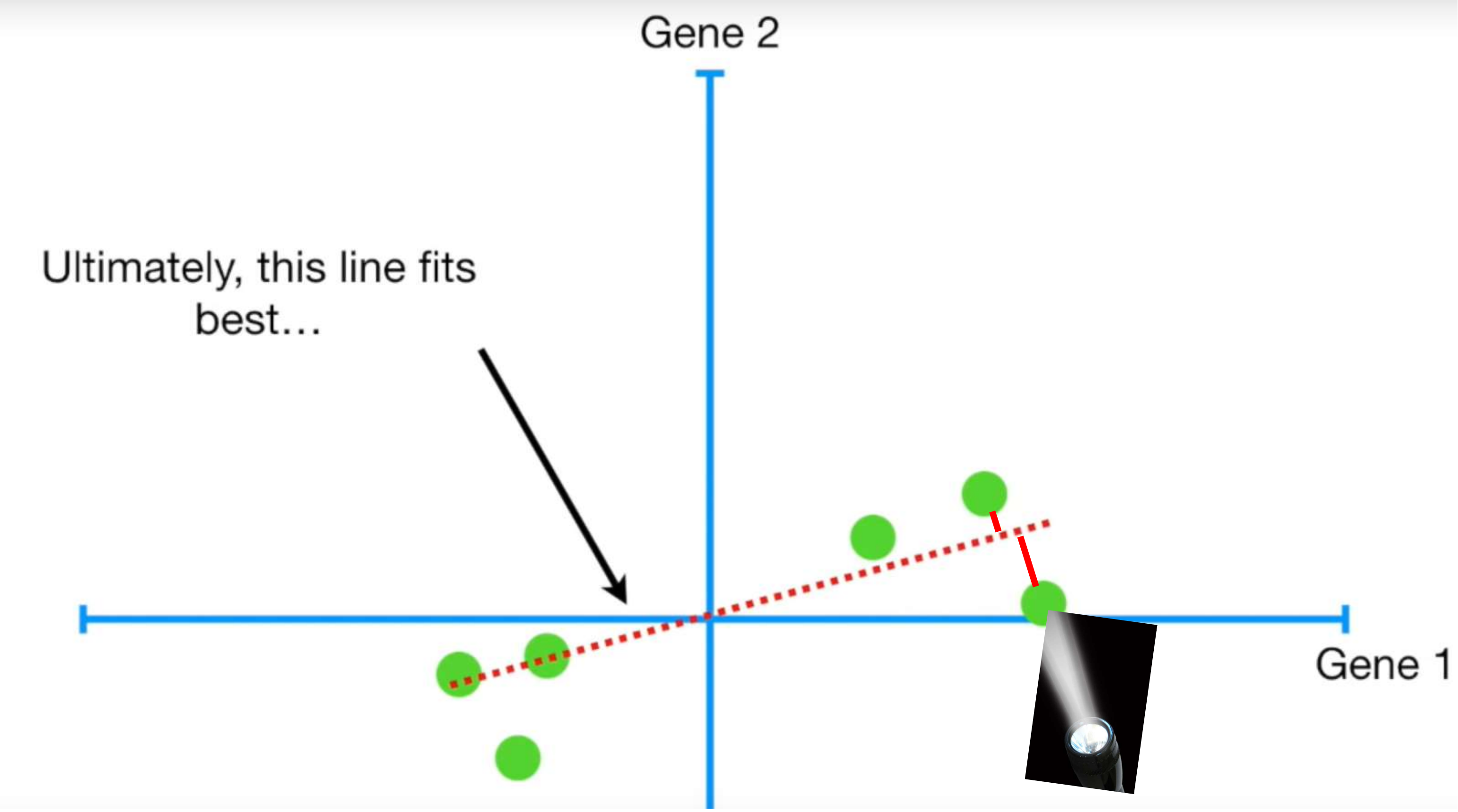
Line Charts


2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Weekday
Weekend
Geometric Mean Ridership (x1000)
20
15
10
5
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

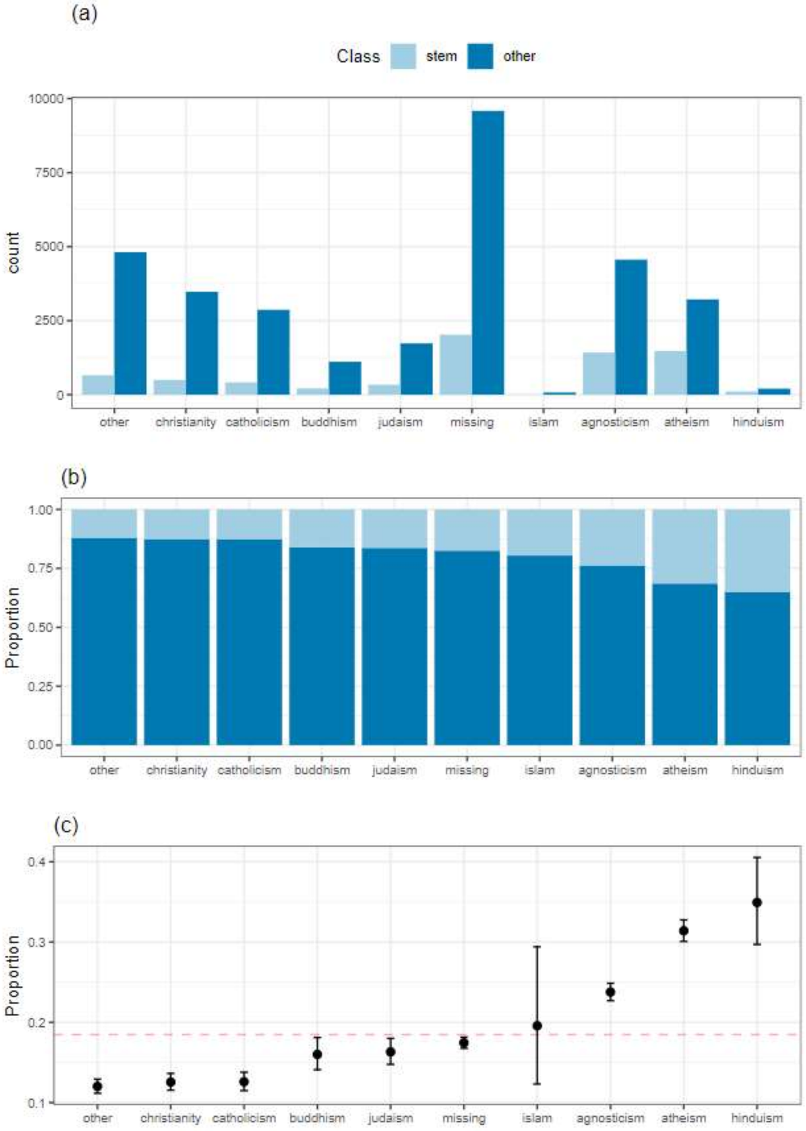
PCA

https://www.youtube.com/watch?v=FgakZw6K1QQ&vl=en

Goal: Retain the most information while reducing the number of dimensions



Are Proportions Different?

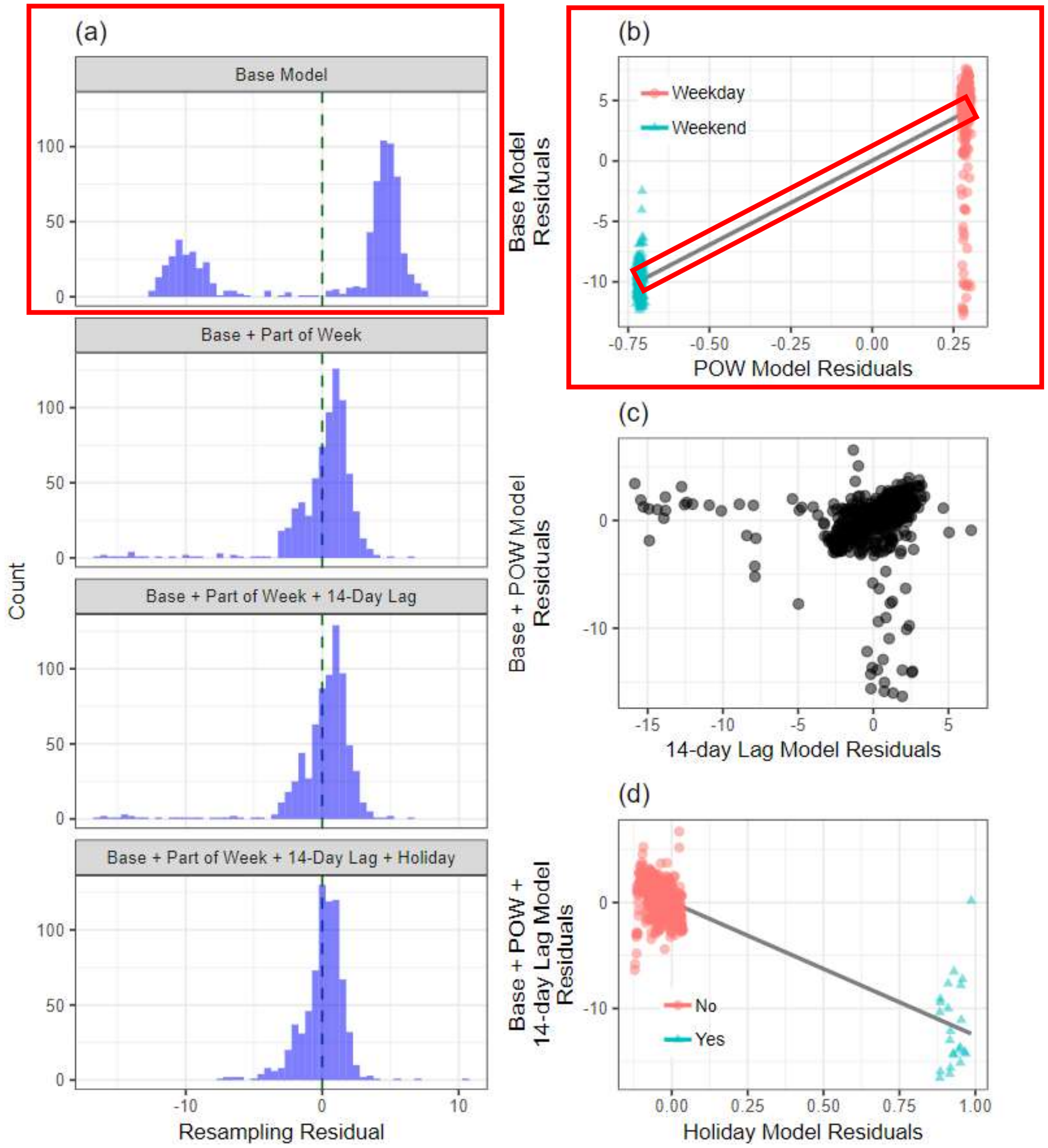


Post-Modeling Charts

Partial Regression Plots

Base Model is:

ridership = month + year + week



weekday = month + year + week