Library Computer Usage Analysis

Computer Science Industry Advisory Board Meeting - November 27, 2017

Brown Biggers, Nickolas Lloyd, Michael Ellis, Patricia Tanzer

Background

- The University of North Carolina at Greensboro University Libraries host over 300 public machines.
- These machines have numerous configurations:
  - Desktop vs. Laptop
  - Single- vs. Dual-monitors
  - Adjacent window vs. no window
- These machines have been allocated based upon perceived need.
- Machines report logon data in a format that timestamps a change of state:
  - In-use
  - Available
  - Restarted
  - Shutdown
- Machines do not report duration of state.

Questions to answer

- What characteristics define the popularity of a machine, as measured by its usage?
  - Dual monitors
  - Proximity to external window
- Do external phenomena (e.g. weather, location, etc.) affect a machine's usage?
  - Hourly rainfall
  - Hourly temperature
  - Location within library (Reading room, CITI lab, etc.)
- Do patterns of usage correlate with campus calendars?

Data Sets

- Library machine logon data
  - datestamp, machine, state
- Machine attributes
  - adjacentWindow, dualMonitor, is245, etc.
- Weather data
  - NOAA weather station data for PTI Airport
    - Temperature
    - Precipitation
- Library gate counts
  - At-open and at-close gate counts for two main library entrances.
- Semester schedule
  - Fall/Spring breaks, exam weeks, reading day, commencement, holidays

Methodologies

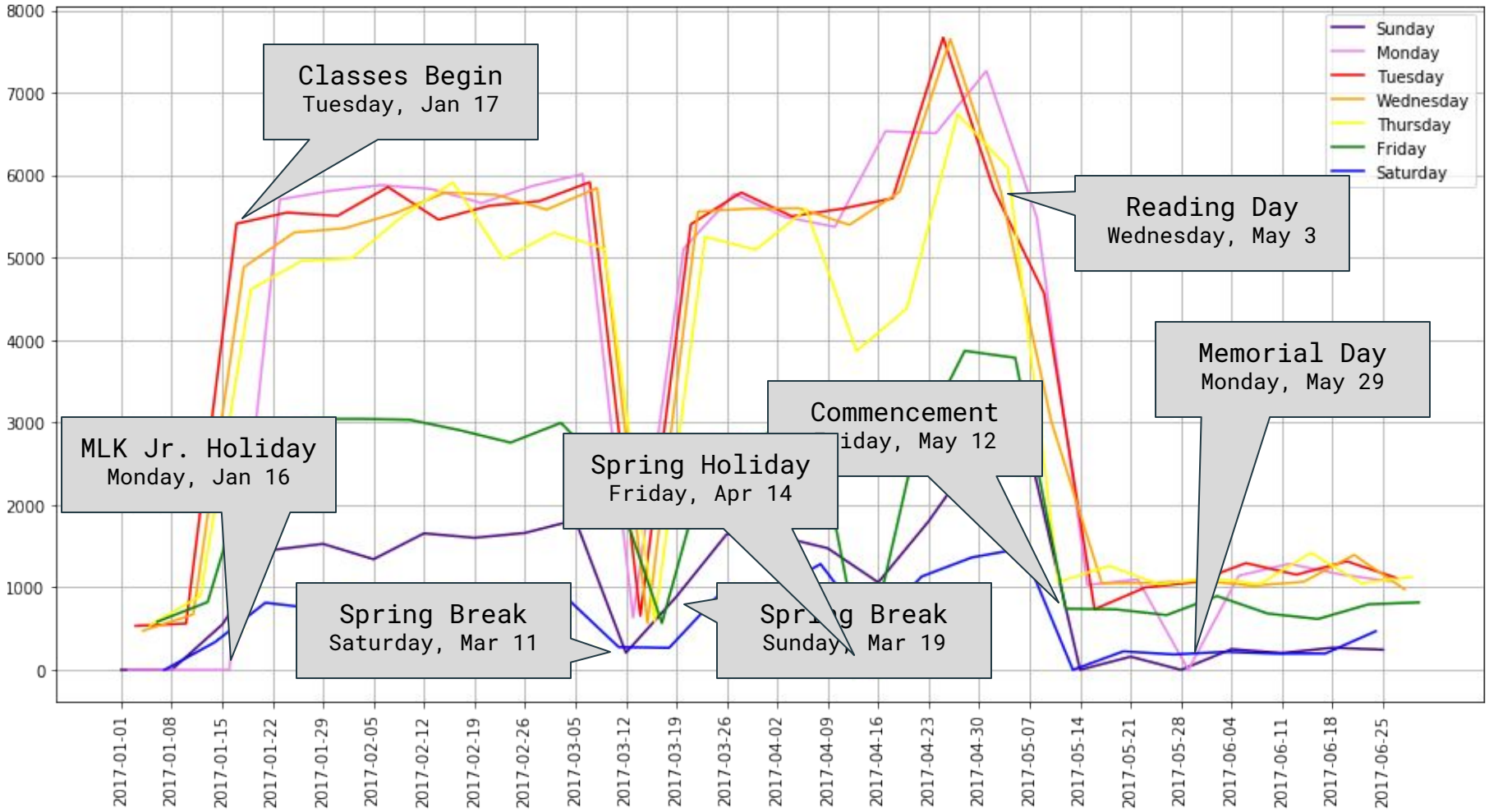
- Transforming Data:
  - Usage of machines to per-hour statistics
- Correlation (machine usage):
  - Gate count
  - Temperature
  - Precipitation
  - Location
  - Attributes
- Hypothesis Testing:
  - Investigate p-value of above correlations
- Visualization of Data
- Conclusions

Decoding the Weather

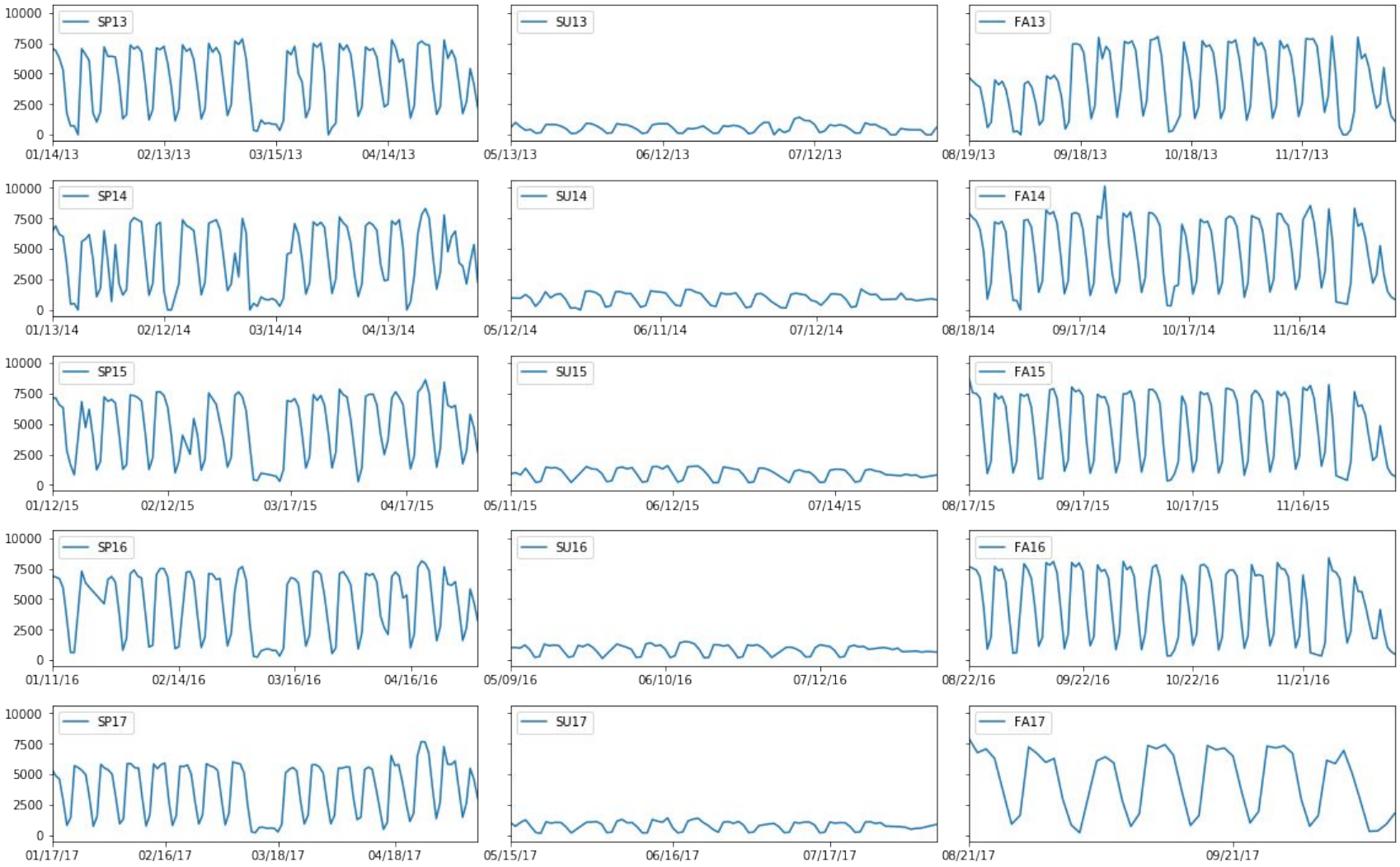
```
('METAR: ', 'METAR KEWR 111851Z
VRB03G19KT 2SM R04R/3000VP6000FT
TSRA BR FEW015 BKN040CB BKN065
OVC200 22/22 A2987 RMK AO2 PK WND
29028/1817 WSHFT
1812 TSB05RAB22 SLP114 FRQ
LTGICCCCG TS OHD AND NW-N-E MOV
NE P0013 T02270215')
```

station: KEWR
type: routine report, cycle 19 (automatic report)
time: Wed Oct 11 18:51:00 2017
temperature: 22.7 C dew point: 21.5 C
wind: variable at 3 knots, gusting to 19 knots
wind: WNW at 28 knots at 18:17
visibility: 2 miles
visual range: on runway 04R, from 3000 to greater than 6000 meters
pressure: 1011.5 mb
precipitation: 0.13in
weather: thunderstorm with rain; mist
sky: -a few clouds at 1500 feet
-broken cumulonimbus at 4000 feet
-broken clouds at 6500 feet
-overcast at 20000 feet
remarks: - Automated station (type 2)
- peak wind 28kt from 290 degrees at 18:17
- wind shift at 18:12
- frequent lightning (intracloud,cloud-to-cloud,cloud-to-ground)
- thunderstorm overhead and NW-N-E moving NE

What about these gate counts?

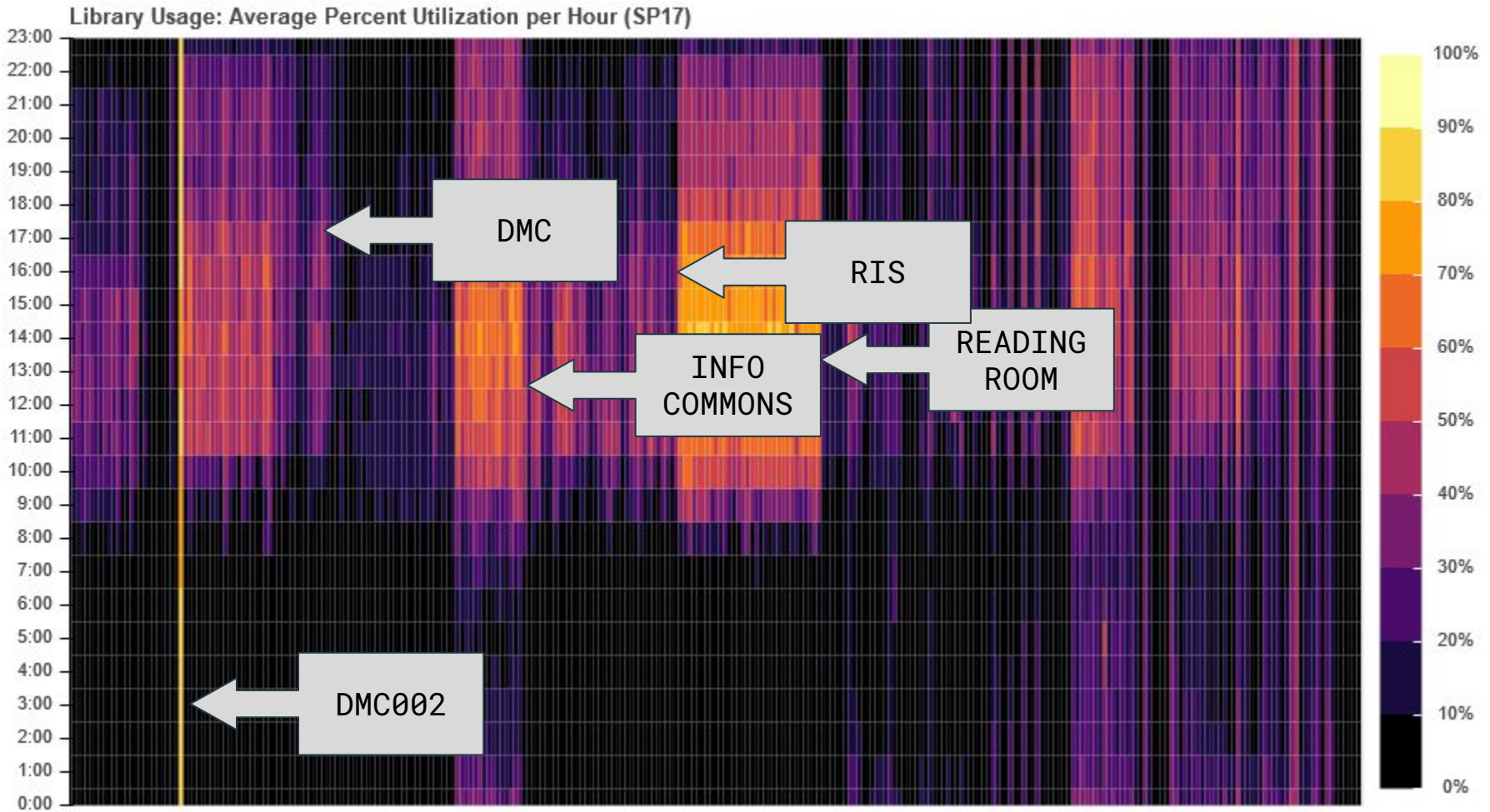


Spring 2017 Gate Count per day, grouped by day-of-week

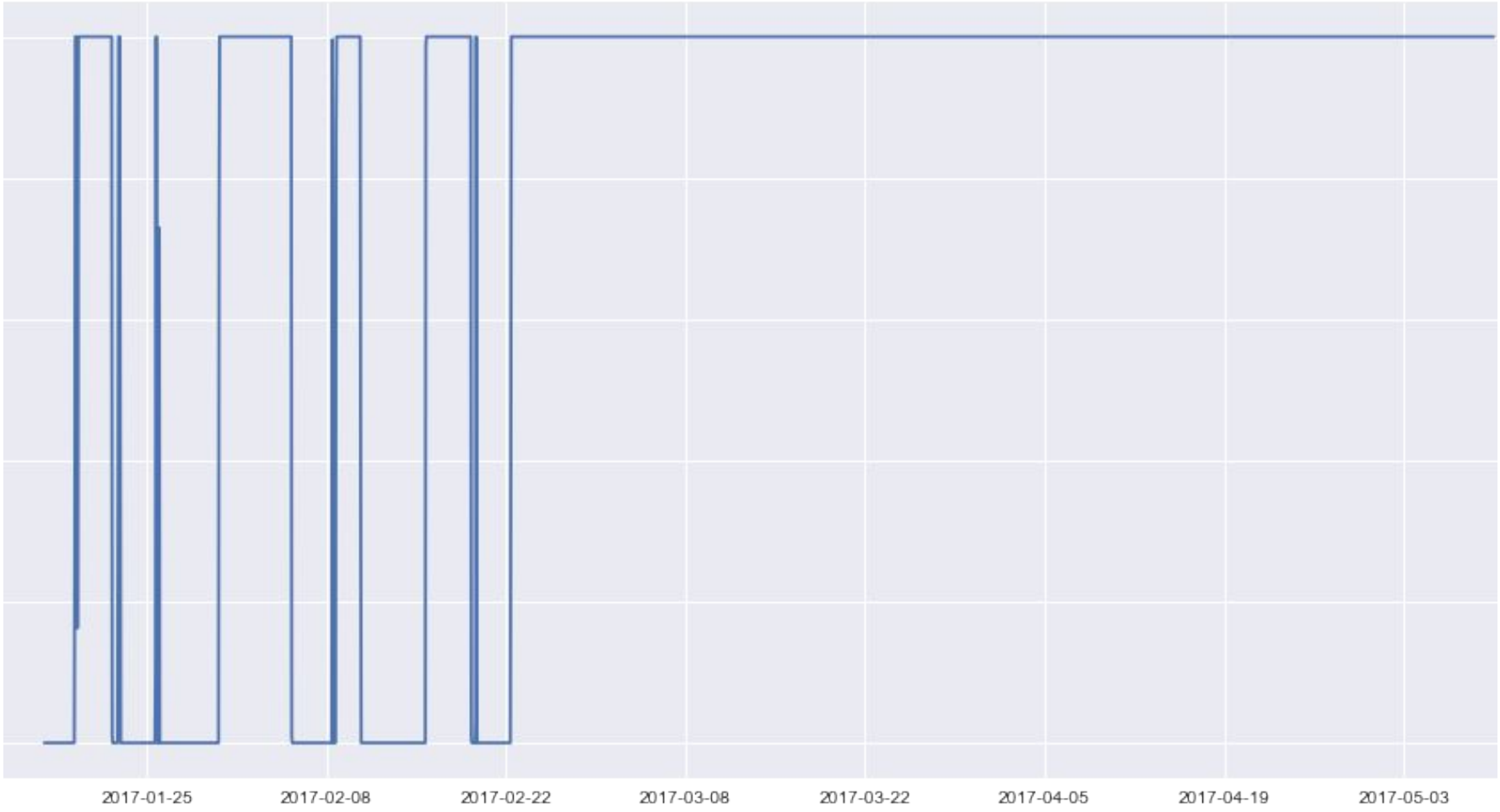


Gate Count per day, separated by semester.

Now let’s look at computer usage….



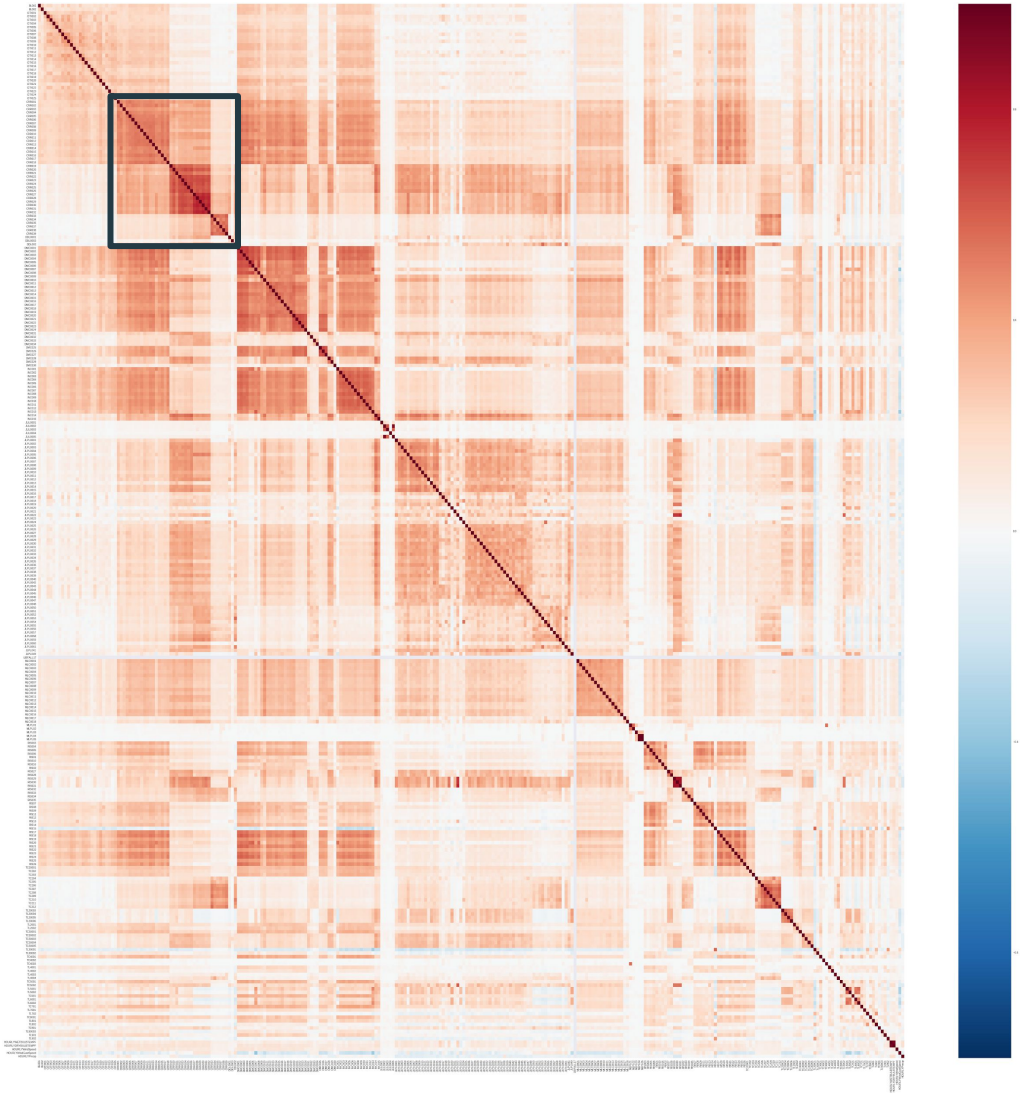
Usage per hour (Spring 17), sorted by region



DMC002 - Spring 2017

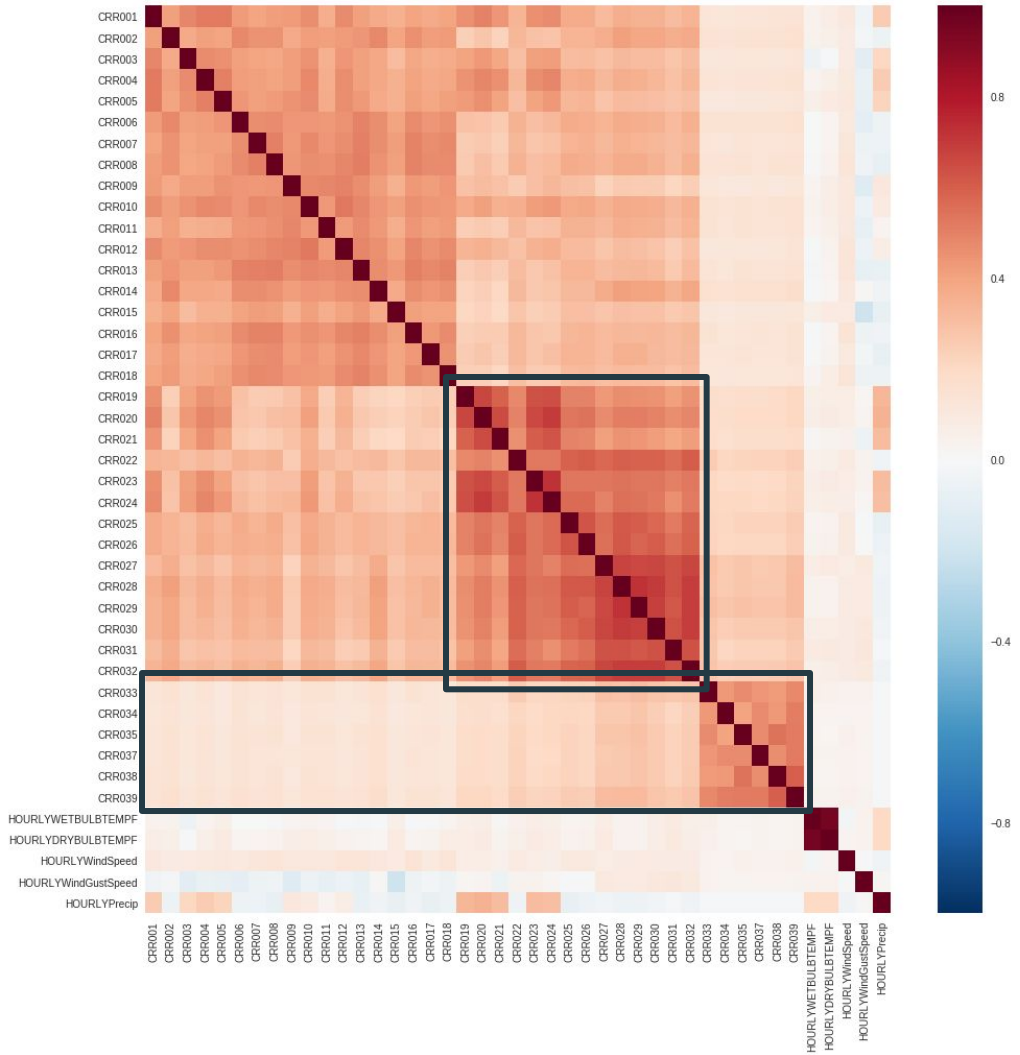
Correlation Heatmap:

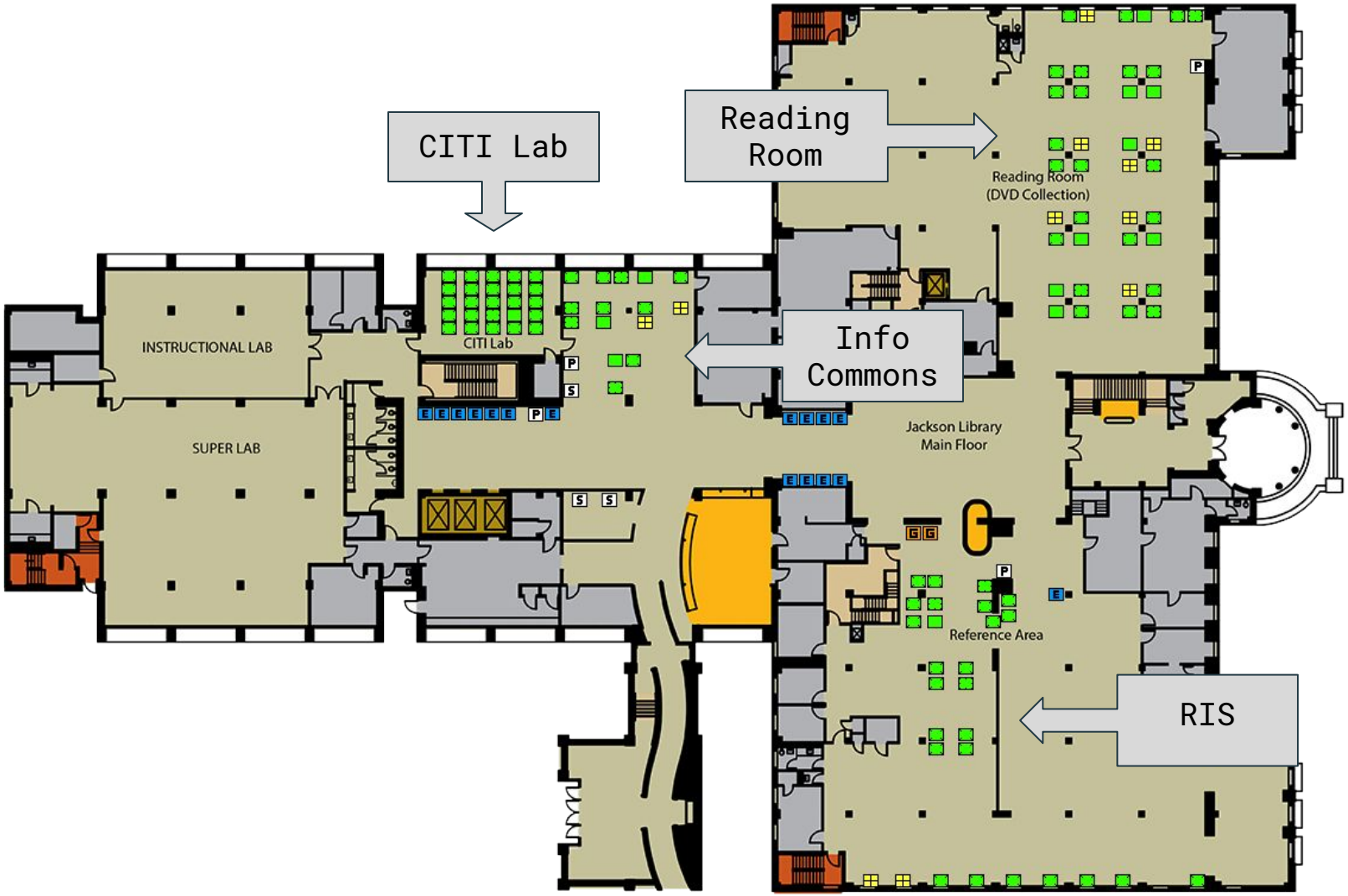
- Hourly usage vs. hourly temperature and precipitation.
- Correlation is weak between specific machines and weather conditions
- Correlation is stronger within physical regions.



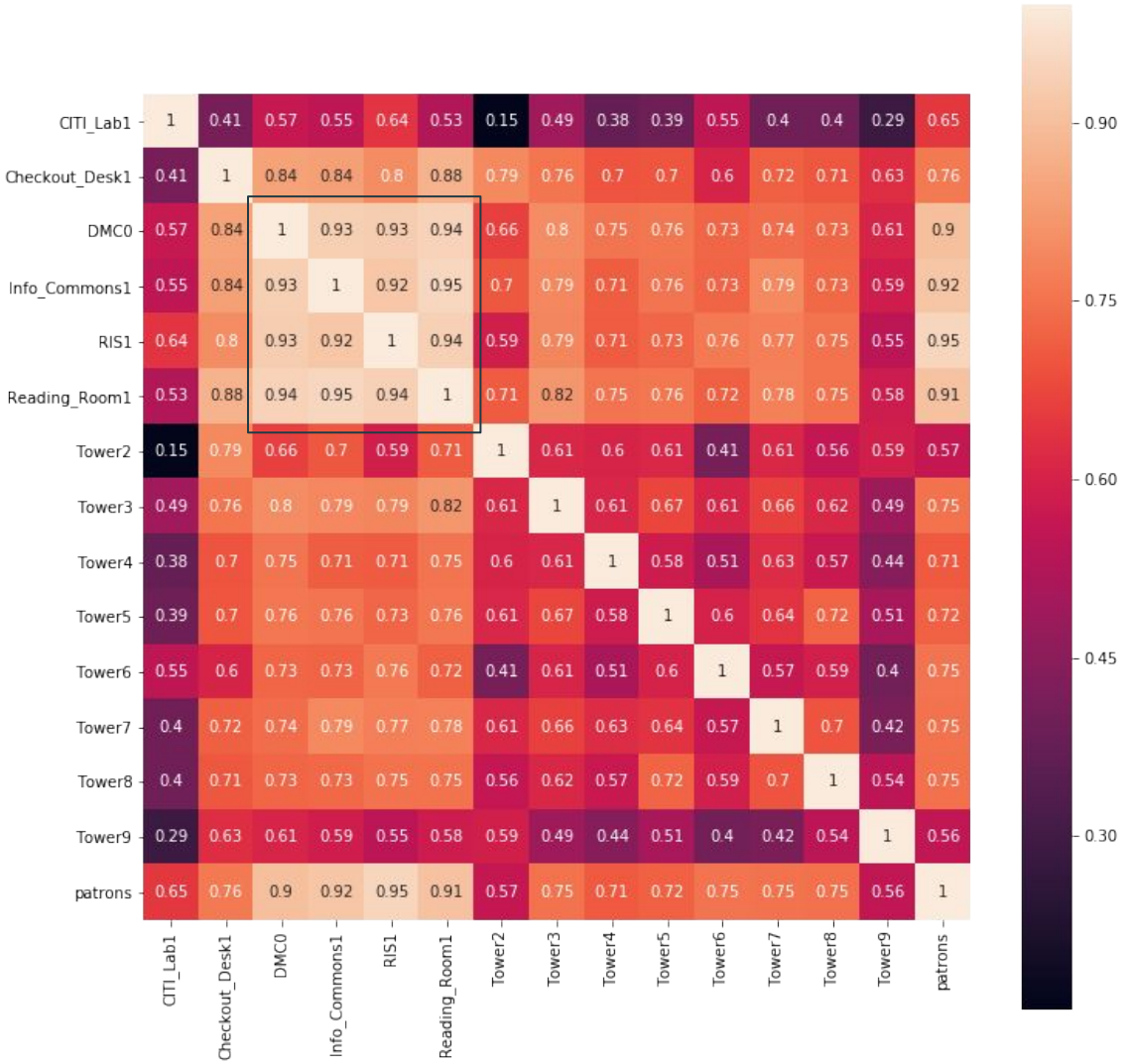
Correlation Heatmap Closeup:

- Usage of machines in a physical region is more strongly correlated to other computers' utilization within that region, than with environmental changes.



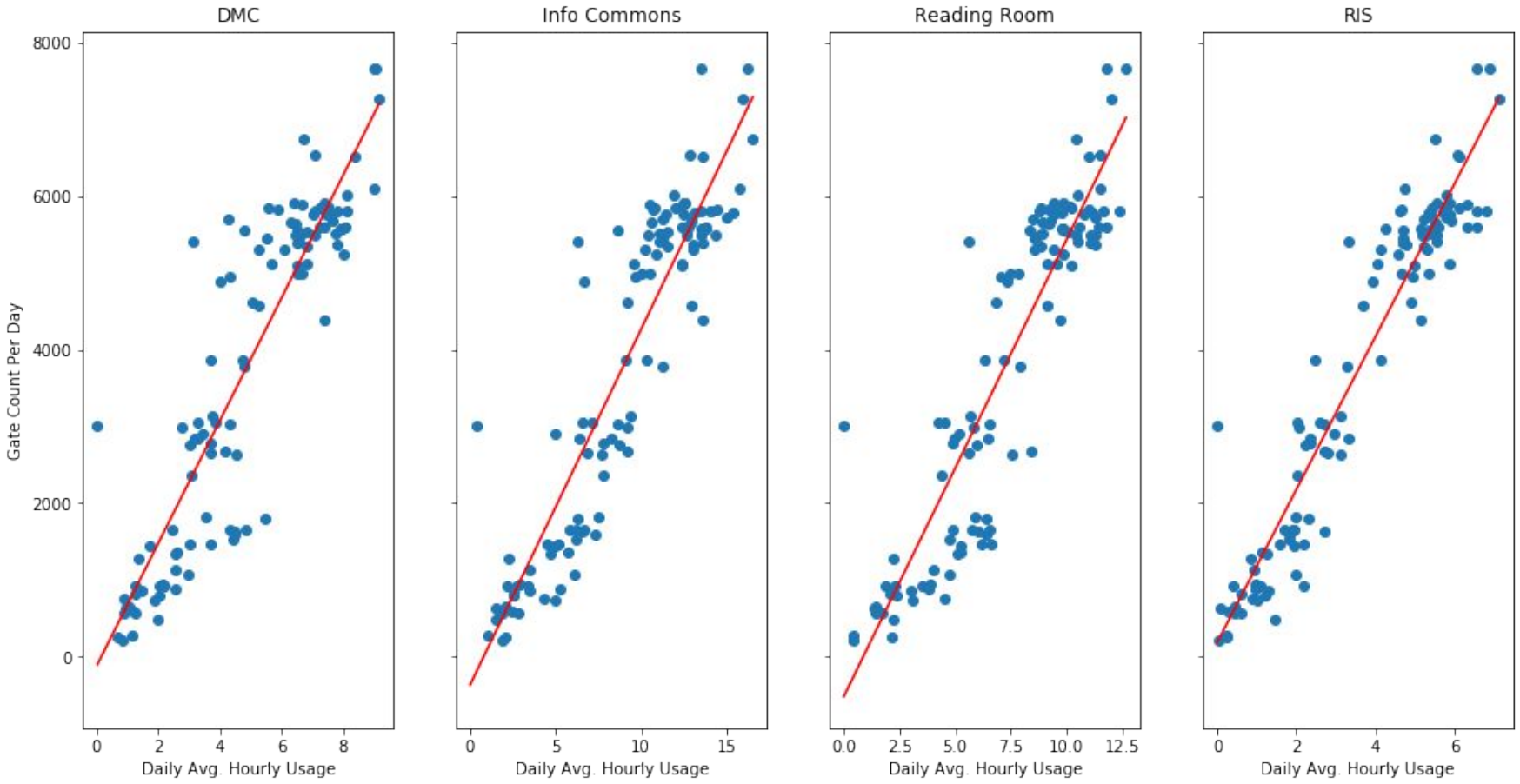


Jackson Library First Floor Regions



Regional daily average usage vs. daily gate count

Linear Regression of four regions (correlation >= .9)



Let’s see why these are so popular…

Criteria of machines:

- Popular regions: Reading Room, Info Commons, RIS, DMC
- Characteristics:
  - First floor (except DMC)
  - Mostly dual monitors
  - Plenty of natural light (windows)
- Compare with DMC
  - Not first floor
  - Some windows
- Compare with CITI Lab
  - First floor
  - Single monitors
  - Windows

Hypothesis: monitor count & location

- With a 95% confidence (p-value <= 0.05):
- **Null Hypothesis 1:** *There is no difference between the use of computers with dual monitors and those with only one. (Spring 2016 - present)*
  - **False!** There is at least a 97% correlation (**p-value 0.03**) in all four semesters between increased hours of computer use and the increased number of monitors.
- **Null Hypothesis 2**: *There is no difference between the use of computers near windows and those that are not.*
  - **True!** There was not enough evidence to reject this hypothesis.

Predictive Analysis

- Given a set of attributes for a new desktop machine (such as floor, number of monitors, availability hours, proximity to windows) **we can predict whether the machine will have high or low use.**
- **High use defined as >23% use per hour, on average**. This is the top quartile of our test data.
- Our algorithm can predict with about **73% accuracy** whether a machine with certain characteristics will be high or low use.

Summary

Take Away Points:

- Computers with dual monitors are preferable to those without.
- People tend to go to anywhere on the first floor of Jackson before the Tower
  - The CITI lab appears to be an exception - possibly because:
    - it is often reserved as a classroom
    - it is mostly single monitor machines
  - The DMC is also an exception - not on first floor
    - Almost entirely dual monitors

Other conditions:

- Proximity to a window does not appear to affect computer usage.
- Neither temperature nor precipitation influenced computer usage or gate counts

How we did it

- Github for sharing content and collaborating on code
- Data Munging and Wrangling through Pandas & NumPy
- Data Visualization via Seaborn, Matplotlib, and Bokeh
- Code editing in Jupyter Notebook (iPython)
- Machine Learning: Decision Tree and Random Forest models, using Sklearn
- https://github.com/UNCG-CSE/Library-Computer-Usage-Analysis/tree/final

143 lines (130 sloc) | 6.82 KB

```python
import pandas as pd
import numpy as np

from pandas.tseries.offsets import *
import functools
import itertools


def _concat(lists):
    #return functools.reduce(lambda x, y: x+y, lists, [])
    return itertools.chain(*lists)

# Compute the intervals during which each computer was in use.
#
# @df must be a pandas DataFrame with columns ['machineName', 'datestamp',
# 'state'].
#
# Returns a list of (machineName,[(timestamp,timerange)]) corresponding to
# sessions of machine usage.
def toUsageIntervals(df):
    # Bin records by `machineName'.
    names = df['machineName'].unique()
    grouped = df.sort_values(by='datestamp').groupby(by='machineName')
    binned = [(name, group.loc[:,'datestamp':'state']) for (name, group) in grouped]

    # This function helps f() construct a mask to collapse runs of in-use or
    # not-in-use records into a single row.
    def g(x, xs):
        if x == False:
            return [True] + xs[1:]
        else:
            return [True] + ([False] * (len(xs) - 1))

    def f(frame):
        # Mark which records correspond to 'in-use'.
        inUse = (frame['state'] == 'in-use').values
        grouped = itertools.groupby(inUse)
        # Filter down to those records which indicate 'in-use', or which occur
        # directly after such a record.
        mask = list(_concat([g(x,list(xs)) for (x,xs) in grouped]))
```