Visualization

Data Science: Jordan Boyd-Graber

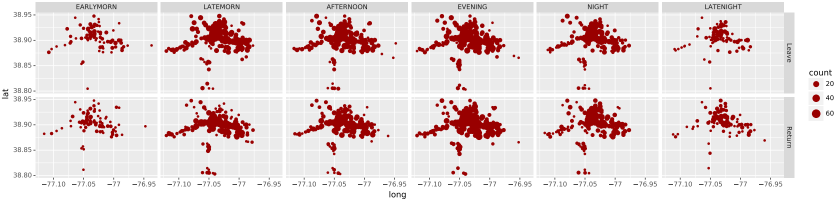
University of Maryland

FEBRUARY 20, 2018

Download Data

- In ds-hw data directory
- **Two** csv files
- Already cleaned

Replicate This Plot



```
reorder_categories
geom_point
facet_grid
```

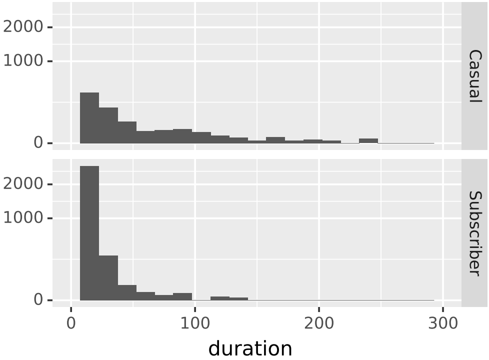
Stations

```
stations = pd.read_csv("stations.csv")

levels = [’EARLYMORN’, ’LATEMORN’, ’AFTERNOON’, ’EVENING’, ’NIGHT’, ’LATENIGHT’]
stations[’time’] = stations[’time’].astype(’category’)
stations[’time’] = stations[’time’].cat.reorder_categories(levels)

p = ggplot(stations)
p += geom_point(stations, aes(x=’long’, y=’lat’, size = ’count’),
                             color="#990000")
p += facet_grid("type ~ time")
p.save("stations.pdf", scale=0.6, height=4, width=18)
```

Replicate This Plot



```
geom_histogram binwidth
scale_y_sqrt
xlim
facet_grid
```

Rides

```
def duration(time):
    fields = [int(x[:-1]) for x in time.split()]
    return fields[0] * 60 + fields[1] + fields[2]/60.

rides[’duration’] = rides.apply(lambda row:
                                 duration(row[’duration’]),
                                 axis=1)

p = ggplot(rides)
p += geom_histogram(aes(x=’duration’), binwidth=15)
p += scale_y_sqrt()
p += xlim(0, 300)
p += ylab("")
p += facet_grid("subscription ~ .")
p.save("duration.pdf", scale=0.6, height=3, width=4)
```