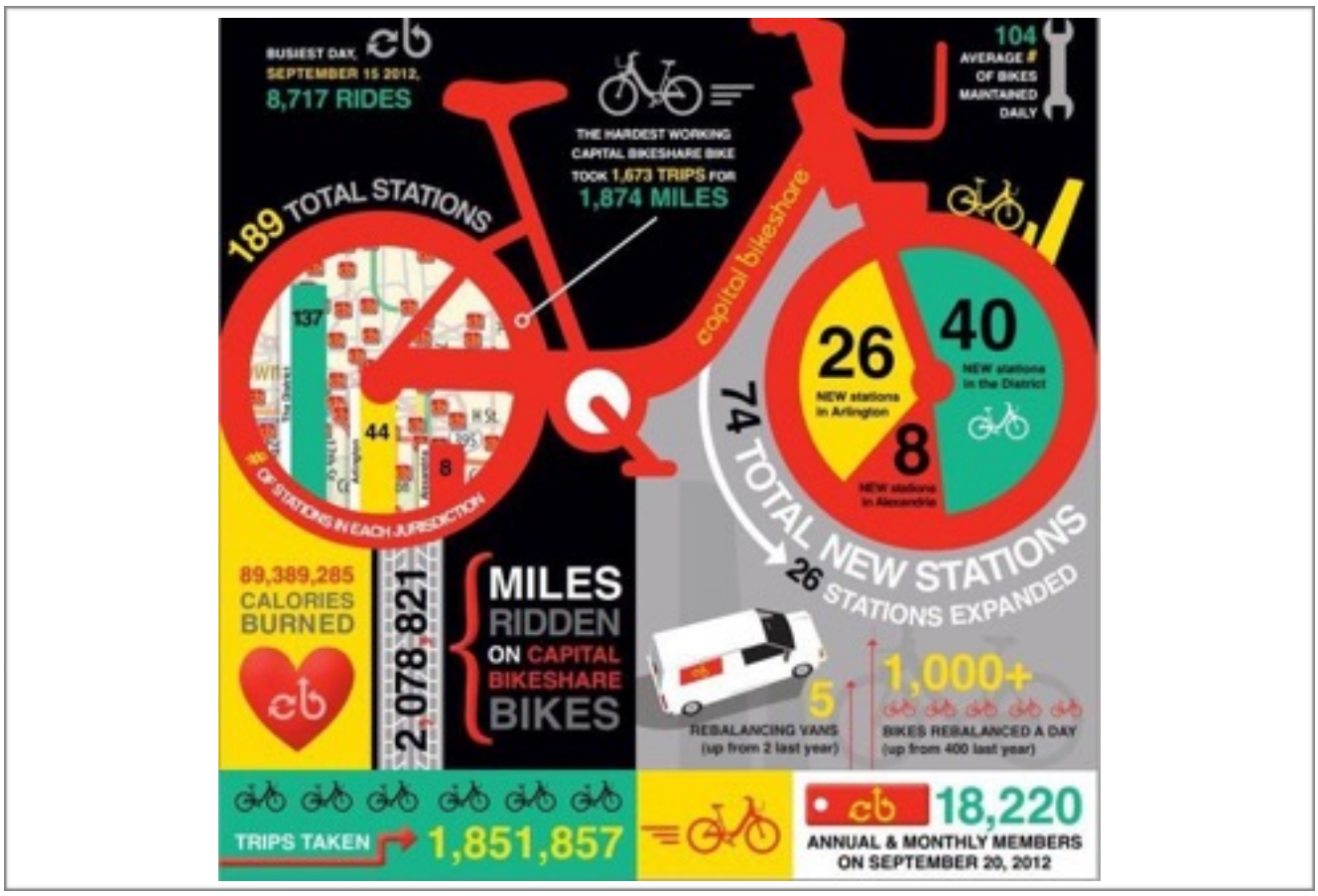
Forecasting Washington D.C.'s



Hourly Bikeshare Demand:

A Kaggle Competition

Bennet Voorhees

[DRAFT: To do on page 5]

The Question

The goal of this Kaggle competition is to forecast hourly bike share demand in the the entire Capital Bikeshare region. Or, in Kaggle's words :

> "In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C."

…it may not look explicit in the description, but Kaggle actually wants folks to forecast demand using information from the past. In other words (in Kaggle's Words:

> "Your model should only use information which was available prior to the time for which it is forecasting."

So there are really two complimentary questions to be answered:

- **Which of the given data determine Capital Bikeshare demand?**
- **What is the hourly demand of rentals given those factors?**

The Data

The dataset provided by Kaggle contains information about time, weather, and demand only. The data are already very tidy; there are also no missing observations for any of the variables. A codebook is provided at the end of the document.

Table 1: "Raw Data" provided by Kaggle

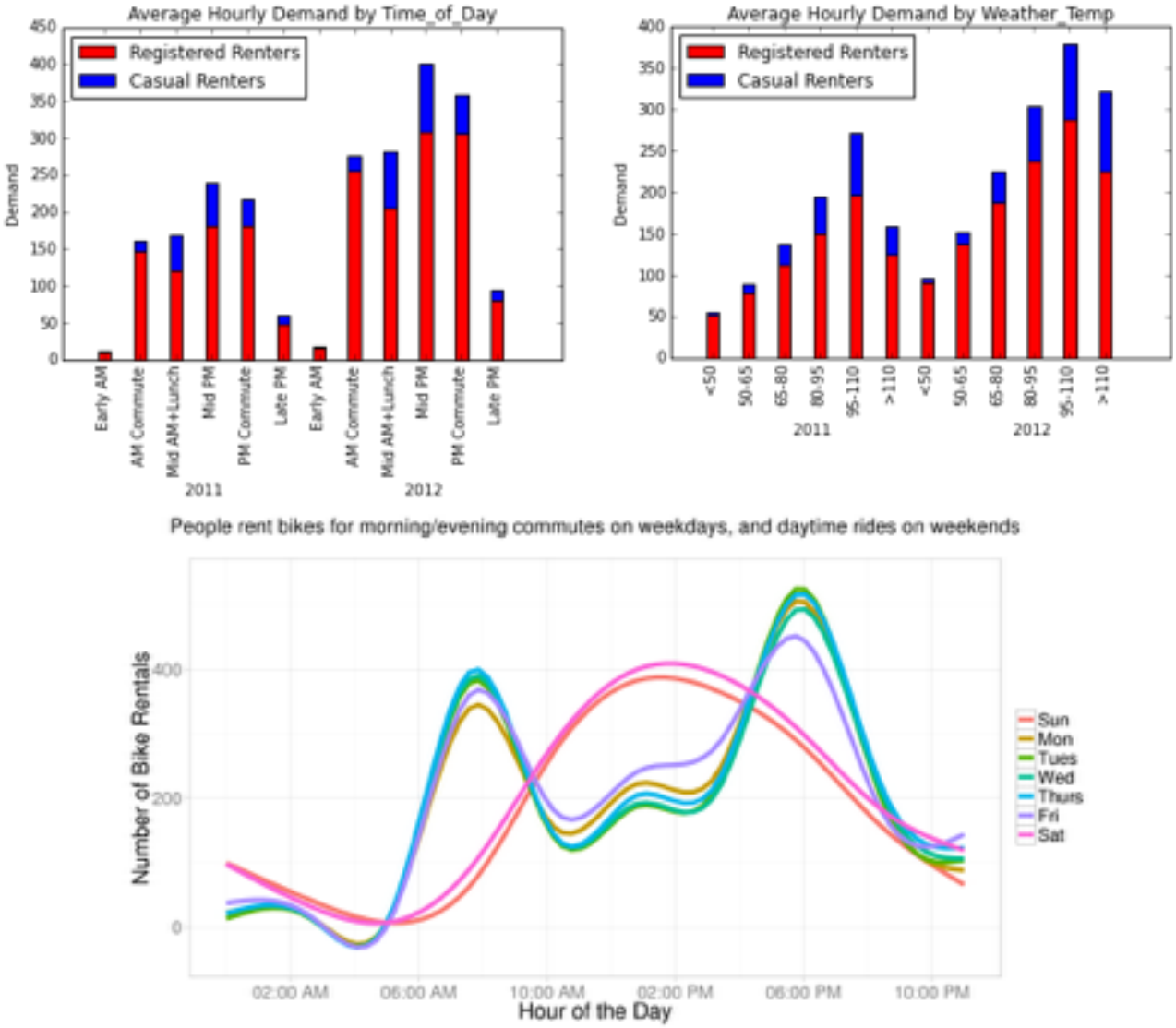
Time	Weather	Demand
Year/Month/Date/Hour	Season	Registered
Holiday?	Conditions	Casual
Workday?	Temperature (2)	
	Windspeed	
	Humidity	

The only processing the data really need is breaking apart the time variable into its individual components (year, month, date, weekday, and hour). Some of these variables will be turned into categorical dummy variables (which I'll discuss below).

Exploratory Analysis

After feature engineering, I've taken a look at all of the variables in the dataset vis-a-vis the two demand variables (casual rider rentals and registered rider rentals). Figure 1 shows a breakdown of—what I think—are the most intuitive relationships.

Figure 1: Demand vis-a-vis Time of Day and Temperature



Note: Bottom center chart was produced by Kaggle.

Model Specification

Since I estimating continuous response variable (demand), I will be using a regression model.

To start, I wanted to see the explanatory power of every variable. Table 2 shows each variable's explanatory power for explaining variation in the two types of demand. Given that, on average, hourly casual demand is 36 and hourly rental demand is 155, not one of these variables alone has good explanatory power.

Table 2: Individual explanitory power of all variables

RMSE (Casual Users)	Variable	RMSE (Registered Users)	Variable
41.23763	temp	138.428462	Hour
41.44142	atemp	140.143321	temp
46.616853	Workday	140.233751	atemp
47.040938	Hour	143.16491	Year
47.049444	humidity	144.513247	humidity
47.518661	Weekday	147.626472	Workday
47.764509	Weekday_5	147.691272	weather
48.075754	Weekday_6	147.752817	weather_3
48.336524	weather	148.114548	Weekday
48.45974	Weekday_2	148.140326	Weekday_6
48.48876	weather_3	148.183837	Season
48.527297	Weekday_1	148.212098	Month
48.528159	Weekday_3	148.317117	Weekday_5
48.665017	weather_2	148.45709	Weekday_3
48.707491	Weekday_4	148.505499	windspeed
48.739319	weather_4	148.536638	Weekday_4
48.740383	holiday	148.541232	Weekday_1
48.752152	Day	148.562039	weather_2
48.775191	windspeed	148.586099	holiday

RMSE (Casual Users)	Variable	RMSE (Registered Users)	Variable
49.114748	Year	148.600002	weather_4
49.522064	Season	148.614343	Weekday_2
49.772434	Month	148.63934	Day

• TO DO:

1. **Make all of the variables into lag variables given the rules of the competition.**
2. **Use the variables from above to start constructing the multivariable model. I have already begun experimenting (with no real methodology). You can see that in 03_Linear_Regression.py in this folder.**
3. **Make several submissions to Kaggle to reduce RMSE as much as possible (Kaggle ranks by a close metric to this —RMSLE).**

CODE BOOK:

from: http://www.kaggle.com/c/bike-sharing-demand/data

datetime - hourly date + timestamp

season:

4. spring
5. summer
6. fall
7. winter

holiday - whether the day is considered a holiday

workingday - whether the day is neither a weekend nor holiday

weather :

1. Clear, Few clouds, Partly cloudy, Partly cloudy
2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

temp - temperature in Celsius

atemp - "feels like" temperature in Celsius

humidity - relative humidity

windspeed - wind speed

casual - number of non-registered user rentals initiated

registered - number of registered user rentals initiated

count - number of total rentals