Exploratory Data Analysis of Bike Rental System in Washington, D.C.

Ana Laguna Pradas

Synopsis

In this project, we performed some exploratory analysis on the data set from the kaggle competition, which contains historical data of bike sharing system in Washington, D.C. from the beginning of 2011 to the end of 2012. In this competition, the goal is to combine past rental patterns with historical weather data in order to forecast bike rental demand. To do so, we built several models including Poisson regression, Quasi-Poisson regression, Linear regression (with log transformation of the response variables) and Negative Binomial regression, however none of the models fit the data very well. Therefore we could only explore the relationship between the bike rental and all the explanatory variables.

Background

Following is a short background introduction from the kaggle competition website:

"*Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able to rent a bike from one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.*

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city."

Understanding the Data Set

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, `weather conditions`, `precipitation`, `day of the week`, `season`, `hour of the day`, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in the Capital Bikeshare website. The data set used in this project was aggregated on daily basis and then added the corresponding weather and seasonal information. Weather information are extracted from the i-weather website.

Count of rented bikes is also correlated to some events in the town, which are easily traceable via search engines. For instance, a query like “2012-10-30 Washington D.C.” on Google returns results related to Hurricane Sandy. Therefore the data can be used for identification and validation of anomaly or event detection algorithms as well.

The original data set contains 16 variables in total:

- instant: record index
- dteday: date
- season: seasons (1: spring, 2: summer, 3: fall, 4: winter)

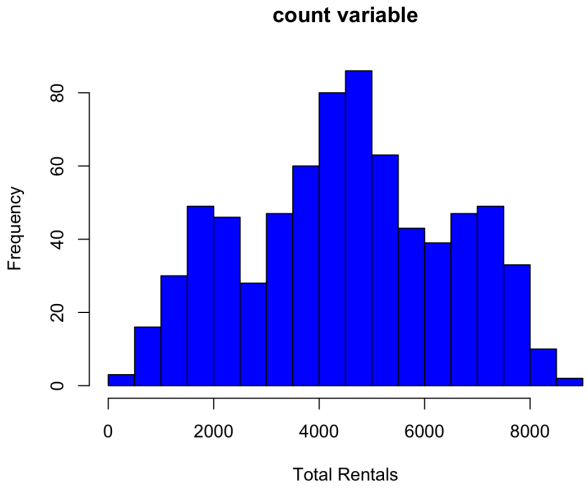
- yr: year (0: 2011, 1: 2012)
- mnth: month (1 to 12)
- holiday: whether day is holiday or not (extracted from the goverment website)
- weekday: day of the week
- workingday: if the day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit:
    - 1: Clear, Few clouds, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- temp: Normalized temperature in Celsius. The values are divided by 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided by 50 (max)
- hum: Normalized humidity. The values are divided by 100 (max)
- windspeed: Normalized wind speed. The values are divided by 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

There are some variables in the data set that are not in our scope of work. For this reason, we removed the irrelevant variables and
made `season`, `holiday`, `workingday` and `weather` as factor variables. The cleared data set looks like as it follows:

```
      dteday season holiday workingday weathersit      temp    atemp 1
 2011-01-01      1       0          0          2 14.110847 18.18125 2
2011-01-02      1       0          0          2 14.902598 17.68695 3 2
011-01-03       1       0          1          1  8.050924  9.47025 4 20
11-01-04        1       0          1          1  8.200000 10.60610 5 201
1-01-05         1       0          1          1  9.305237 11.46350 6 2011
-01-06          1       0          1          1  8.378268 11.66045       h
um windspeed  cnt 1 80.5833 10.749882  985 2 69.6087 16.652113  801 3
43.7273 16.636703 1349 4 59.0435 10.739832 1562 5 43.6957 12.522300 16
00 6 51.8261  6.000868 1606
```

Exploratory Analysis with Plots

First of all, we want to observe how the response variable `cnt` is distributed.



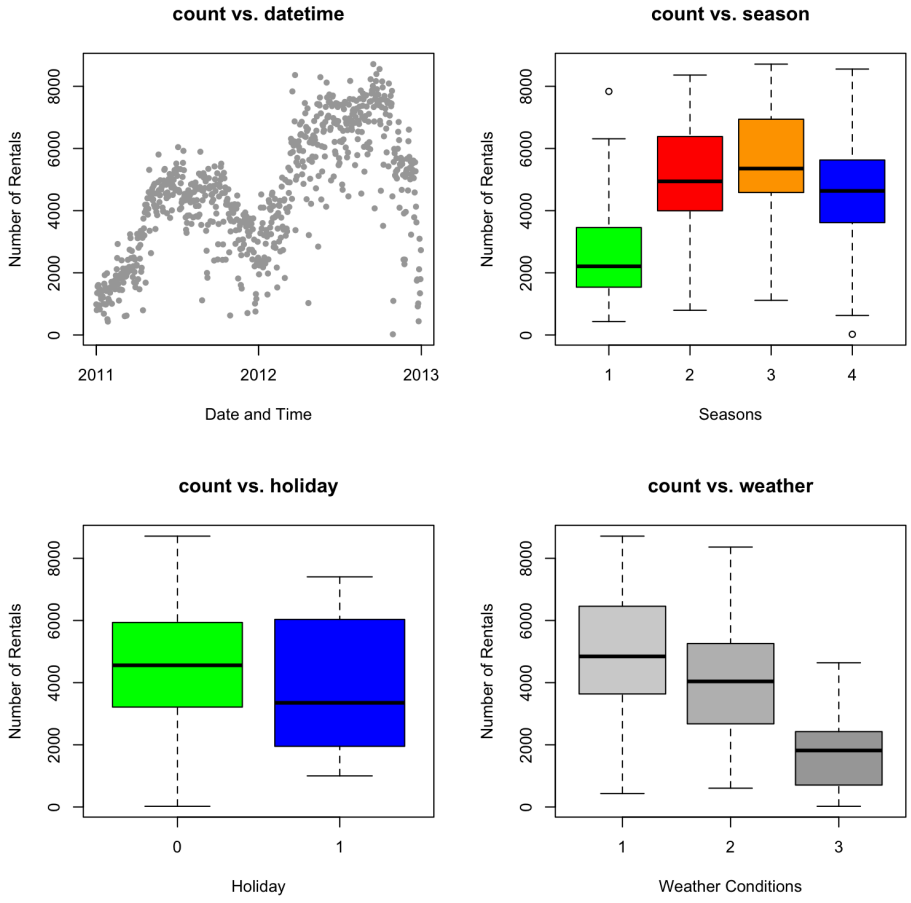
From the histogram above, it seems that the number of total rented bikes follow a nearly normal distribution. Many studies have revealed that Poisson distribution can be used to analyze count data. The mean and variance of Poisson distribution are the same, and when the mean is getting larger, Poisson distribution approximates a normal distribution.

```
        mean          var        ratio     4504.3488 3752788.2083
833.1478
```

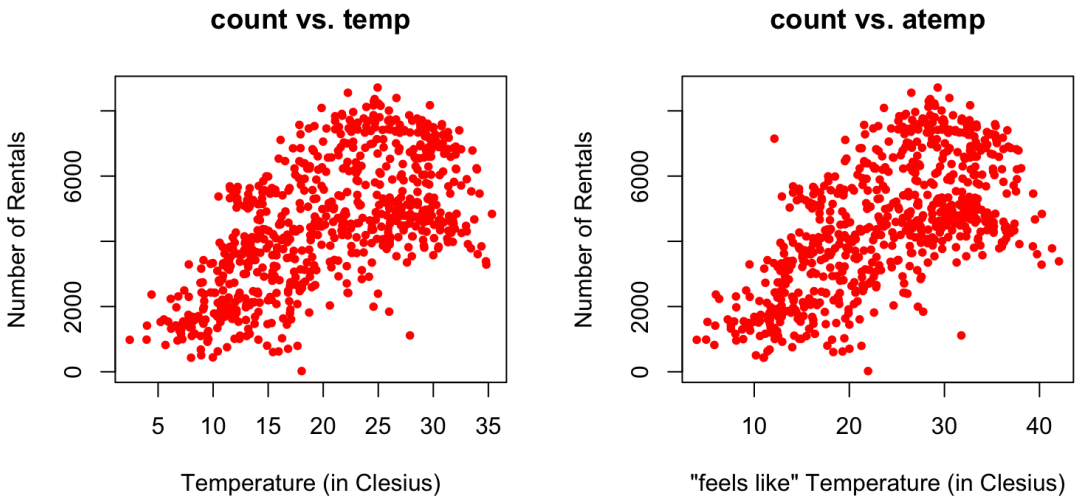
We can see there is a huge difference between mean and variance of the `cnt` variable. Clearly, it doesn't follow a standard Poisson distribution. For count data, if variance is larger than mean, it means the data has over dispersion. There are several ways to deal with it. We will discuss it later in the regression analysis part.

Next, we looked at the relationship between the response variable and each explanatory variable. We selected some plots with obvious patterns as shown below.The first plot shows the relationship between `cnt` variable and date. We can see that the overall trend increased during the two-year time span. And within each year, there are huge amount of bike rentals during summer and fall seasons.

The second plot shows the relationship between `cnt` variable and season, which confirms the conclusion we got above. The average numbers of bike rentals are the highest during summer and fall.

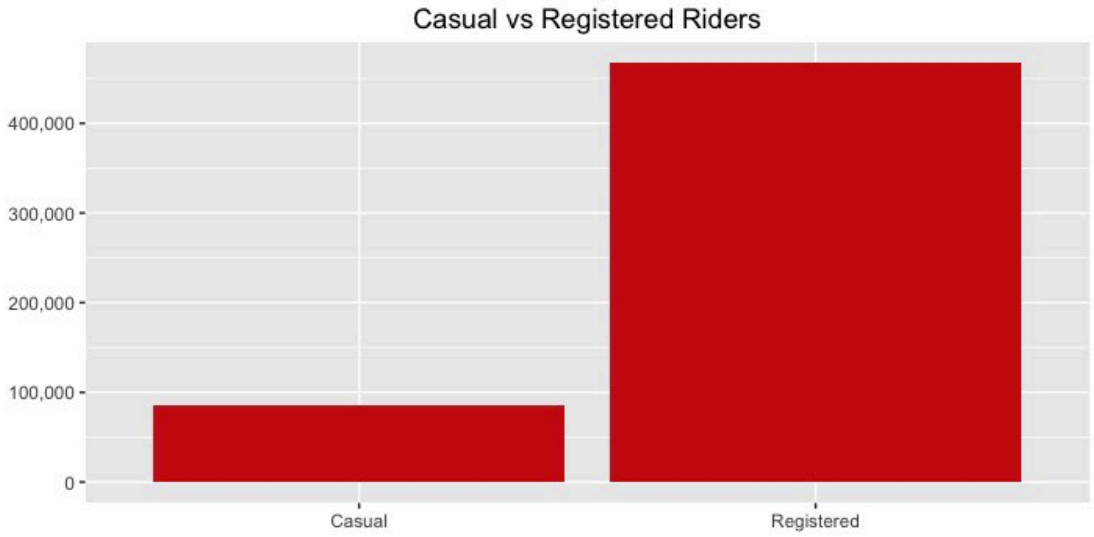


The third plot shows the relationship between `cnt` variable and holiday. We can see that the average number of bike rentals on non-holiday is higher than holiday, but has more variability as well.



The forth plot shows the relationship between `cnt` variable and weather. There is a clearly decreasing trend of bike rentals when weather conditions grow worse.These two plots show the relationship between `cnt` variable and temperature. There seems to be a linear relationship between them, which means more people will rent bikes when it gets warmer. However, the data seems to be scattered with a lot of variability, so the linear relationship might be weak if there is any.

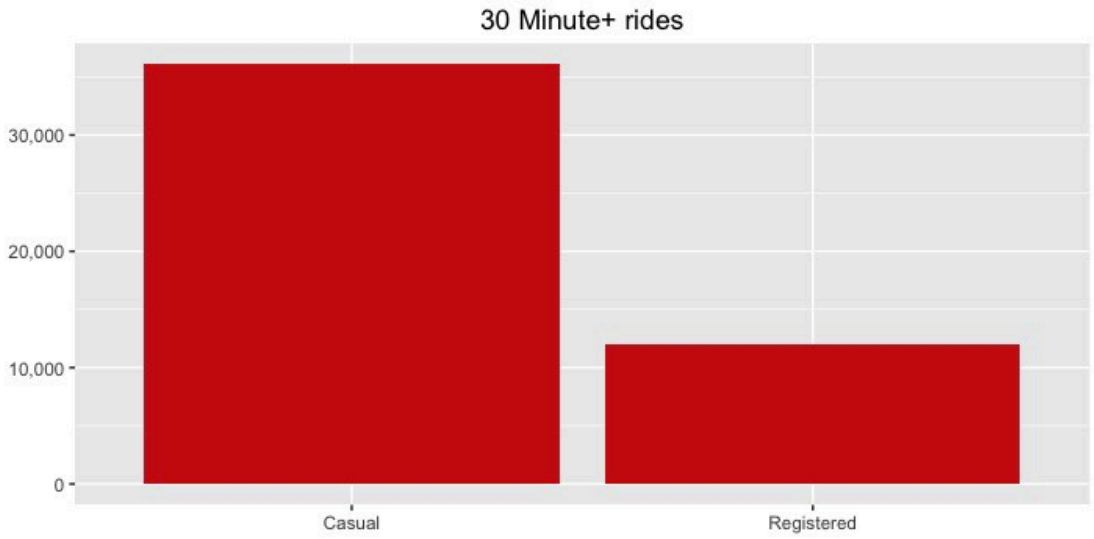
The next thing I wanted to see was the breakout of Registered users Vs. Casual users.



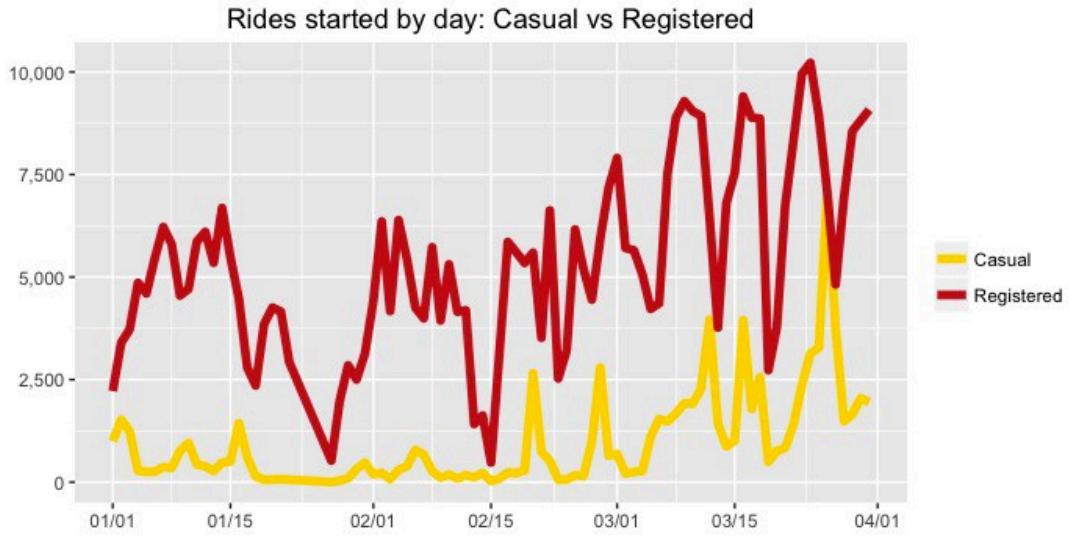
Looking at the breakdown should not surprise anyone. Registered users hopped on a bike 467,432 times compared to casual users at 84,967. Together, those riders logged some impressive minutes!

(Estimated due to rounding conversion from millisecond to second)
9,145,670 Minutes
152,428 Hours
6,351 Days
And over 17 years!

The next breakdown I wanted to see was by ride time. While using a bike, the first 30 minutes are free for both registered and casual bikers. Registered users pay a small fee ($1.50) for usage over 30 minutes while casual users pay a some-what larger fee ($2) which start to scale up in increments of 30 minute timeframes. It is a common practice for bikers to check-in a bike within the 30 minute span and immediately check one out again.

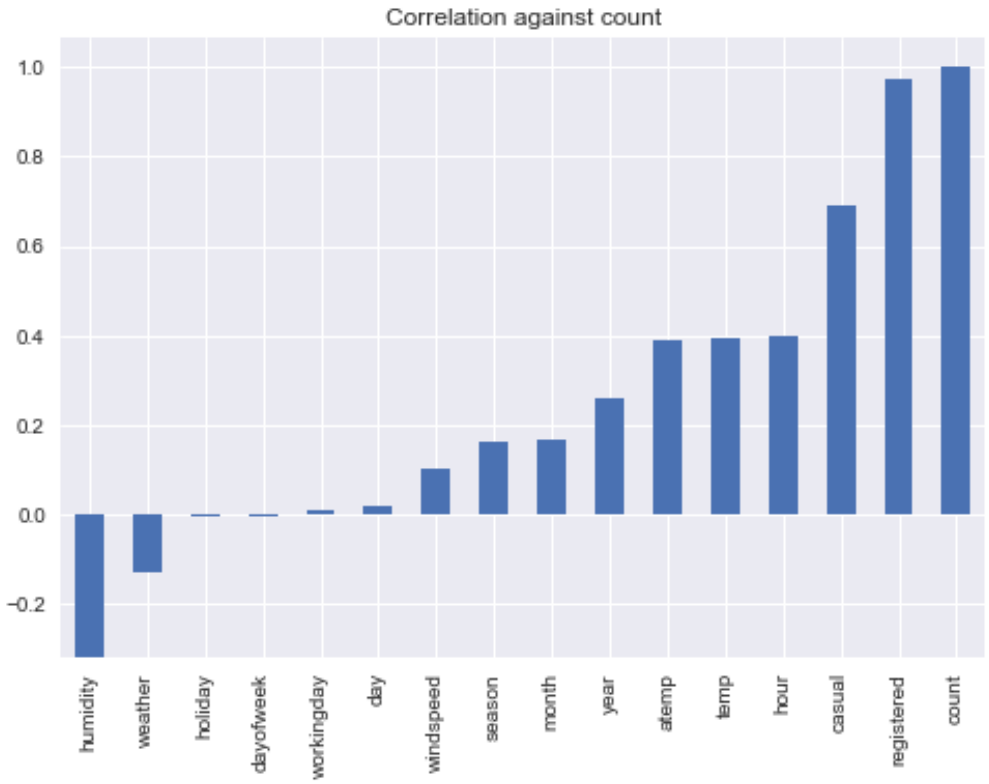


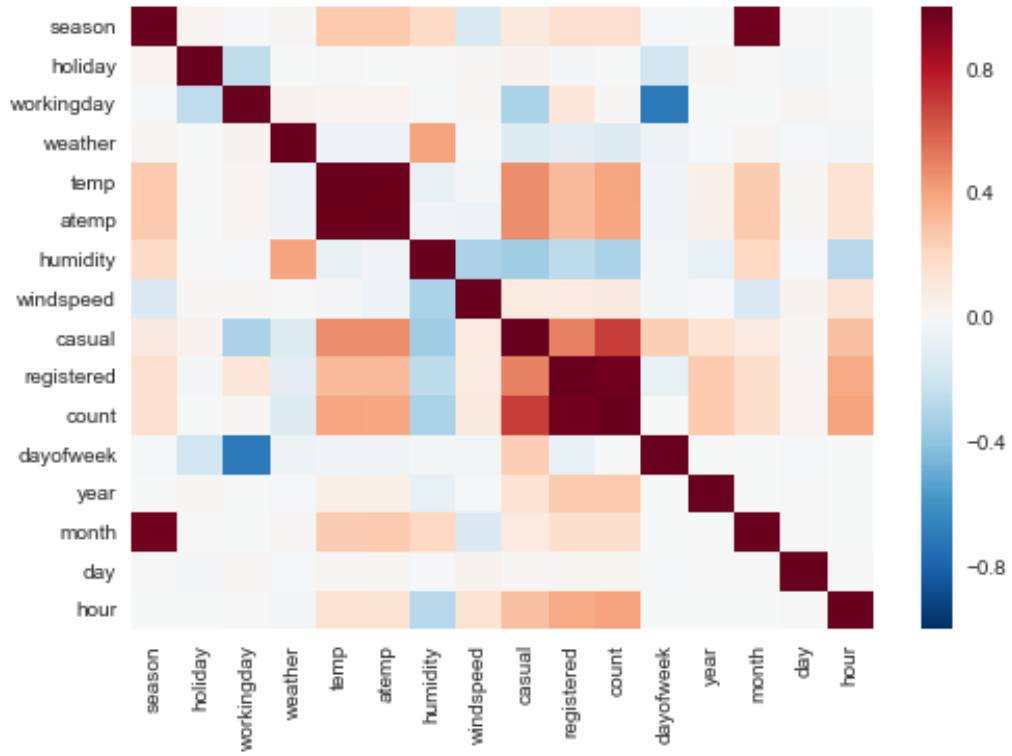
The next graph concerning users I want to look at is the number of trips started based off of a riders member type by day.



This graph shows a major drop late January (Snowzilla) and Valentines Day (another snow day we had, I think?) to essentially zero. What is interesting here and also expected, is the boom in casual rides around the peek bloom for the Cherry Blossoms (March 18th—23rd).

The last graphs show us the correlation among variables:





Some important aggregations of our dataset:

Top Starting Stations

Rank	Starting Station	Trips
1	Columbus Circle / Union Station	13,120
2	Massachusetts Ave & Dupont Circle NW	9,560
3	Lincoln Memorial	9,388
4	Jefferson Dr & 14th St SW	8,138
5	Thomas Circle	7,479
6	15th & P St NW	7,401
7	14th & V St NW	6,568
8	New Hampshire Ave & T St NW	6,491
9	Eastern Market Metro / Pennsylvania Ave & 7th St SE	5,649
10	17th & Corcoran St NW	5,514

Causal Riders: Top Starting and Ending Destinations

Rank	Starting to Ending Station	Count
1	Jefferson Dr & 14th St SW to Jefferson Dr & 14th St SW	962
2	Lincoln Memorial to Jefferson Dr & 14th St SW	899
3	Jefferson Dr & 14th St SW to Lincoln Memorial	894
4	Lincoln Memorial to Jefferson Memorial	886
5	Lincoln Memorial to Lincoln Memorial	792
6	Smithsonian / Jefferson Dr & 12th St SW to Smithsonian / Jefferson Dr & 12th St SW	544
7	Smithsonian / Jefferson Dr & 12th St SW to Lincoln Memorial	506
8	Ohio Dr & West Basin Dr SW / MLK & FDR Memorials to Ohio Dr & West Basin Dr SW / MLK & FDR Memorials	396
9	Jefferson Memorial to Lincoln Memorial	393
10	Lincoln Memorial to Smithsonian / Jefferson Dr & 12th St SW	384

Registered Riders: Top Starting and Ending Destinations

Rank	Starting to Ending Station	Count
1	Columbus Circle / Union Station to 8th & F St NE	977
2	8th & F St NE to Columbus Circle / Union Station	790
3	Adams Mill & Columbia Rd NW to Calvert St & Woodley Pl NW	708
4	Lincoln Park / 13th & East Capitol St NE to Eastern Market Metro / Pennsylvania Ave & 7th St SE	655
5	New Hampshire Ave & T St NW to Massachusetts Ave & Dupont Circle NW	612
6	Eastern Market Metro / Pennsylvania Ave & 7th St SE to Lincoln Park / 13th & East Capitol St NE	573
7	Columbus Circle / Union Station to 11th & H St NE	547
8	11th & H St NE to Columbus Circle / Union Station	540
9	Columbus Circle / Union Station to 6th & H St NE	520
10	D St & Maryland Ave NE to Columbus Circle / Union Station	508