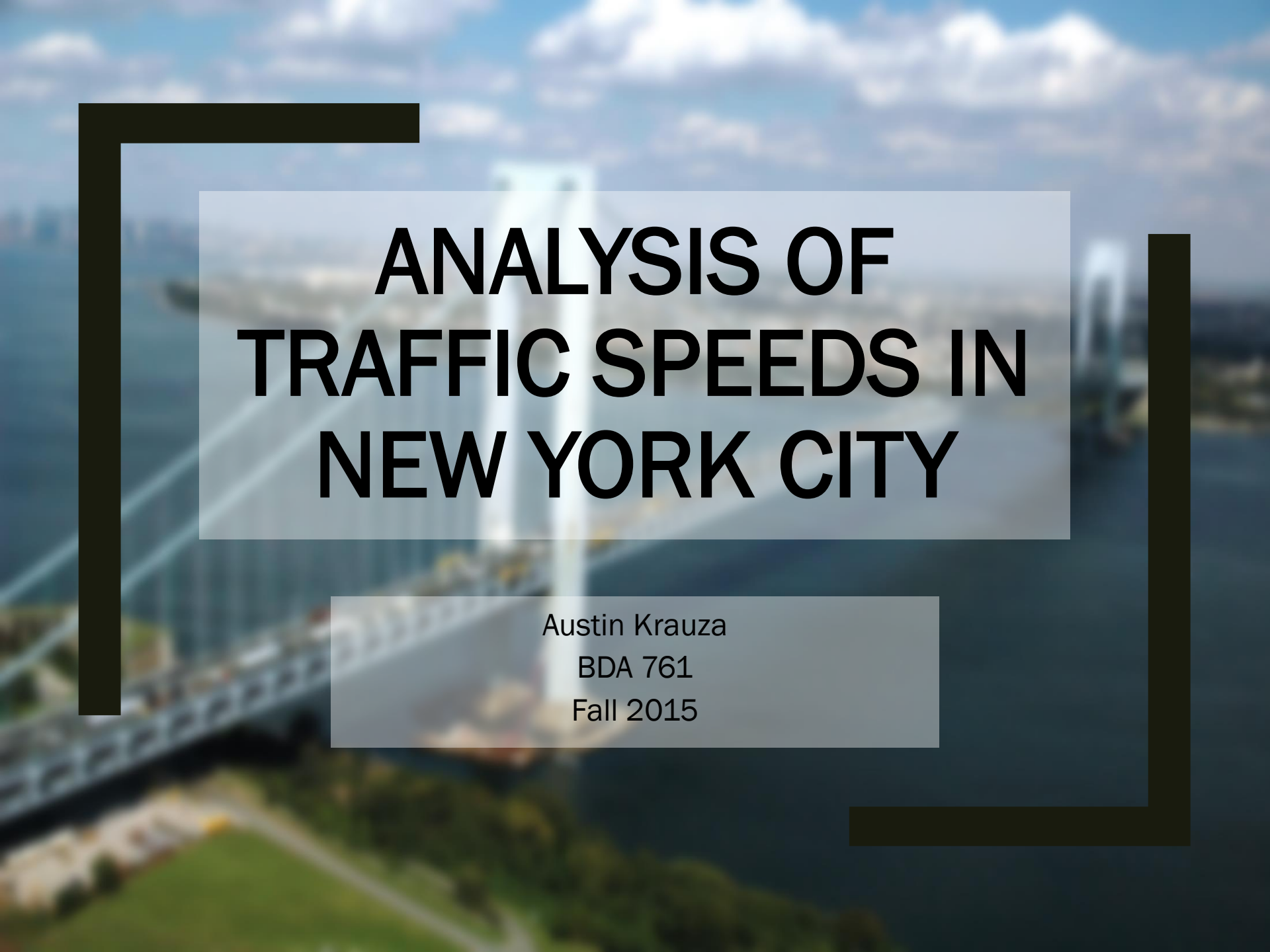


ANALYSIS OF TRAFFIC SPEEDS IN NEW YORK CITY

Austin Krauza

BDA 761

Fall 2015

Problem Statement

- How can Amazon Web Services be used to conduct analysis of large scale data sets?
  - *Data set contains over 80 million records in CSV Format*
- How does the average speed of the Verrazano- Narrows Bridge and the Holland tunnel fluctuate:
  - *Over a 168 Hour Period (One Week)*
  - *Over 11 Months (September 2014- July 2015)*

Software Packages Used

- Microsoft Excel
- SAS (Statistical Analysis System)
- Amazon Web Services
  - *Amazon Elastic Map Reduce (EMR)*
  - *Hive*
  - *Hadoop*
  - *Hue*
  - *Amazon S3 Web Storage*

What is Amazon Web Services?

- Cloud Computing Platform
- Offers various services offsite
- Low cost usage for users
- Provides various platforms
  - *Hadoop*
  - *AWS S3*
  - *MapReduce*

Advantages to using AWS

- Low cost to the user
- Easily scalable
- Provides simple interfaces for novice users
- Allows full customization for advanced users

Information Sources

- Data collected from TRANSCOM scraped using a PHP Script

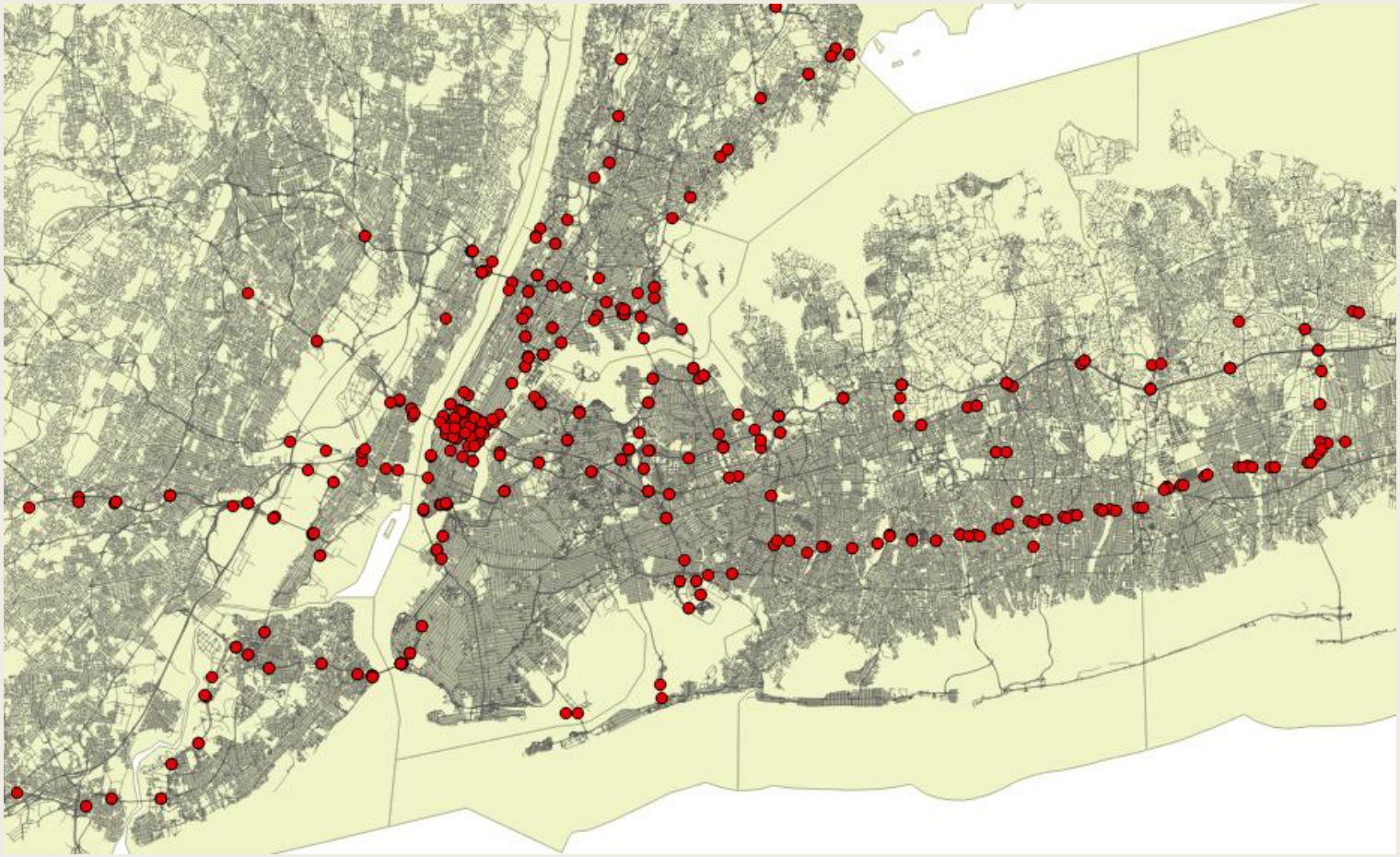
```
// Start the output

        $timeStamp = $searchResultsObj['return']['dataResponse']['linkConditions']['@attributes']['Timestamp'];
        $datetime= explode('T',$timeStamp);
        $date= $datetime[0];
        $time= explode('-',$datetime[1]);
        $time= $time[0];

        $hour= date('H:i', strtotime($time));
        $fullDate= date('Y-m-d H:i:s', strtotime($date.$time));
        $filename= $date.$time;
        file_put_contents("$filename.txt", $datafile);


            foreach ($searchResultsObj['return']['dataResponse']['linkConditions']['linkCondition'] as
            $currentitem):
                $id= $currentitem['ID'];
                $type= $currentitem['DataType'];
                //$speed= round(($currentitem['Speed']/1.609344),0);
                $speed= $currentitem['Speed'];
                $travelTime= $currentitem['TravelTime'];
                $travelTimeFloat= $currentitem['TravelTimeFloat'];
                $query= "INSERT INTO traffic071214 (date, time, stationID, type, speed, travelTime,
                travelTimeFloat) VALUES ('$fullDate', '$hour', '$id', '$type', '$speed', '$travelTime', '
                $travelTimeFloat')";
                $queryres= mysql_query($query);
            endforeach;
```

Sample Data

id	date	time	stationID	type	speed	travelTime	travelTimeFloat
1	11/14/2014 23:50	23:50:00	4616439	Averaged	90	94	94
2	11/14/2014 23:50	23:50:00	4575368	Averaged	106	208	208
3	11/14/2014 23:50	23:50:00	4616246	Averaged	92	76	76
4	11/14/2014 23:50	23:50:00	4616223	Averaged	76	86	86
5	11/14/2014 23:50	23:50:00	4575379	Averaged	92	558	558
6	11/14/2014 23:50	23:50:00	4616352	Averaged	90	135	135
7	11/14/2014 23:50	23:50:00	20484203	Averaged	97	54	54
8	11/14/2014 23:50	23:50:00	4575426	Averaged	114	190	190
9	11/14/2014 23:50	23:50:00	5419028	Averaged	111	12	12
10	11/14/2014 23:50	23:50:00	5361701	Averaged	69	107	107

Sensors on the Staten Island Expressway

Location of Sensors in New York City



Clean-up Using SAS

```
data dec2;
        set dec2;
        year=substr(VAR2,1,4);
        month=substr(VAR2,6,2);
        day=substr(var2,9,2);
        newdate= mdy(month,day,year);
        dow=weekday(newdate);
        hour=substr(var3,1,2);
        minute=substr(var3,4,2);
        how=(((weekday(newdate)-1)*24)+hour);
run;

data dec1;
        set dec1;
        format newdate date9.;
run;

proc summary data=dec2 noprint;
        class newdate;
        output out=o1;
run;
```

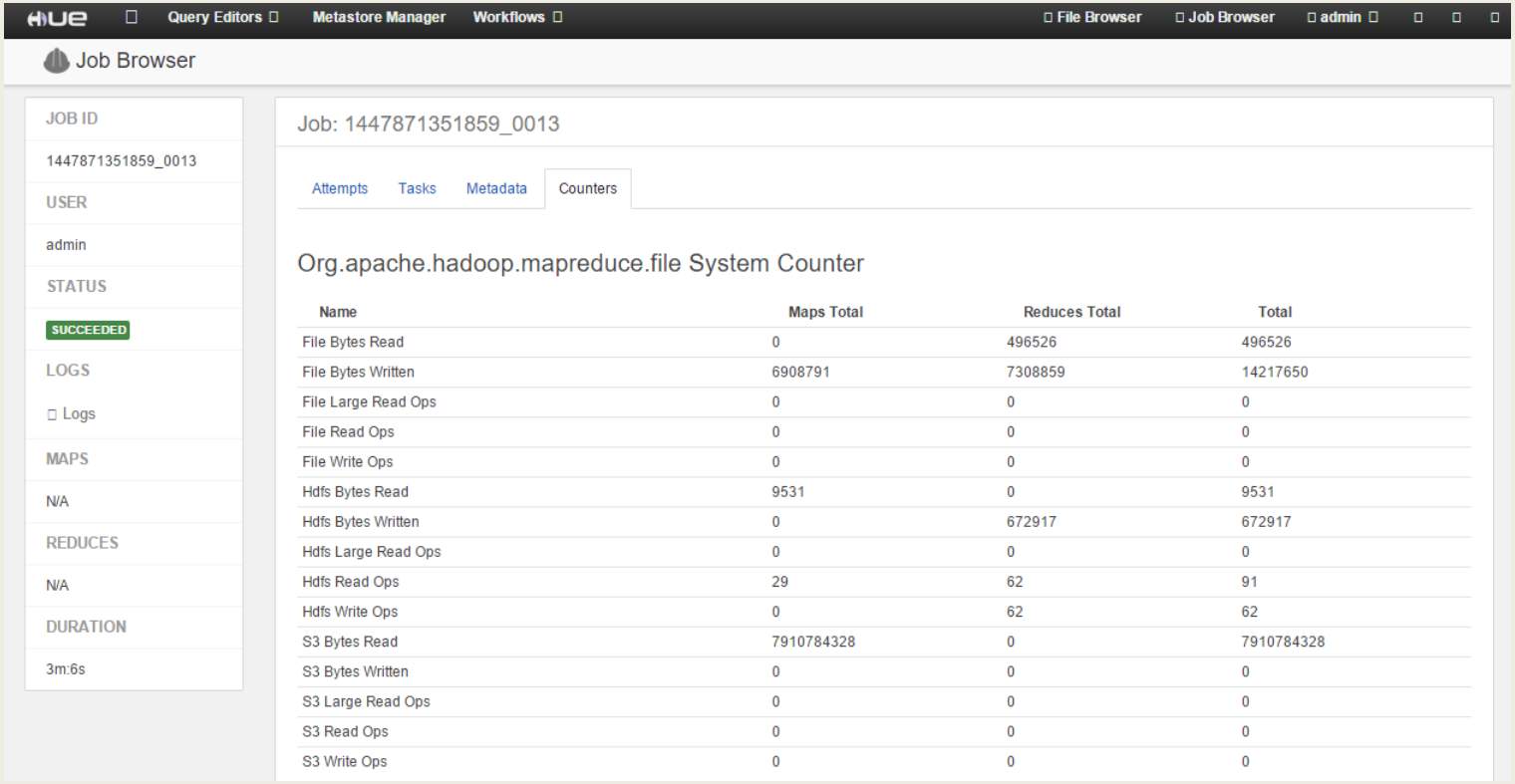
Hive Script: External Table

```
drop table transcomEXT;
CREATE external TABLE `transcomEXT`(
        `id` int,
        `datetime` string,
        `time` string,
        `stationid` int,
        `type` string,
        `speed` int,
        `traveltime` int,
        `traveltimefloat` int,
        `year` smallint,
        `month` int,
        `day` bigint,
        `date` string,
        `dow` int,
        `hour` bigint,
        `minute` bigint,
        `how` int)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://traffic-111715/data/';
```

Hive Query: Analysis

```
select
        avg(speed) as avgSpeed,
        CONCAT(year,'-',month,'-','1') as month1,
        how as HourWeek,
        stationid as station
from transcomext
where stationid in
        (4763652,4763649,4616219,4763655,4763648,
        4616204,4751366,4751367,4456501,4456502)
group by
        stationid,
        how,
        CONCAT(year,'-',month,'-','1');
```

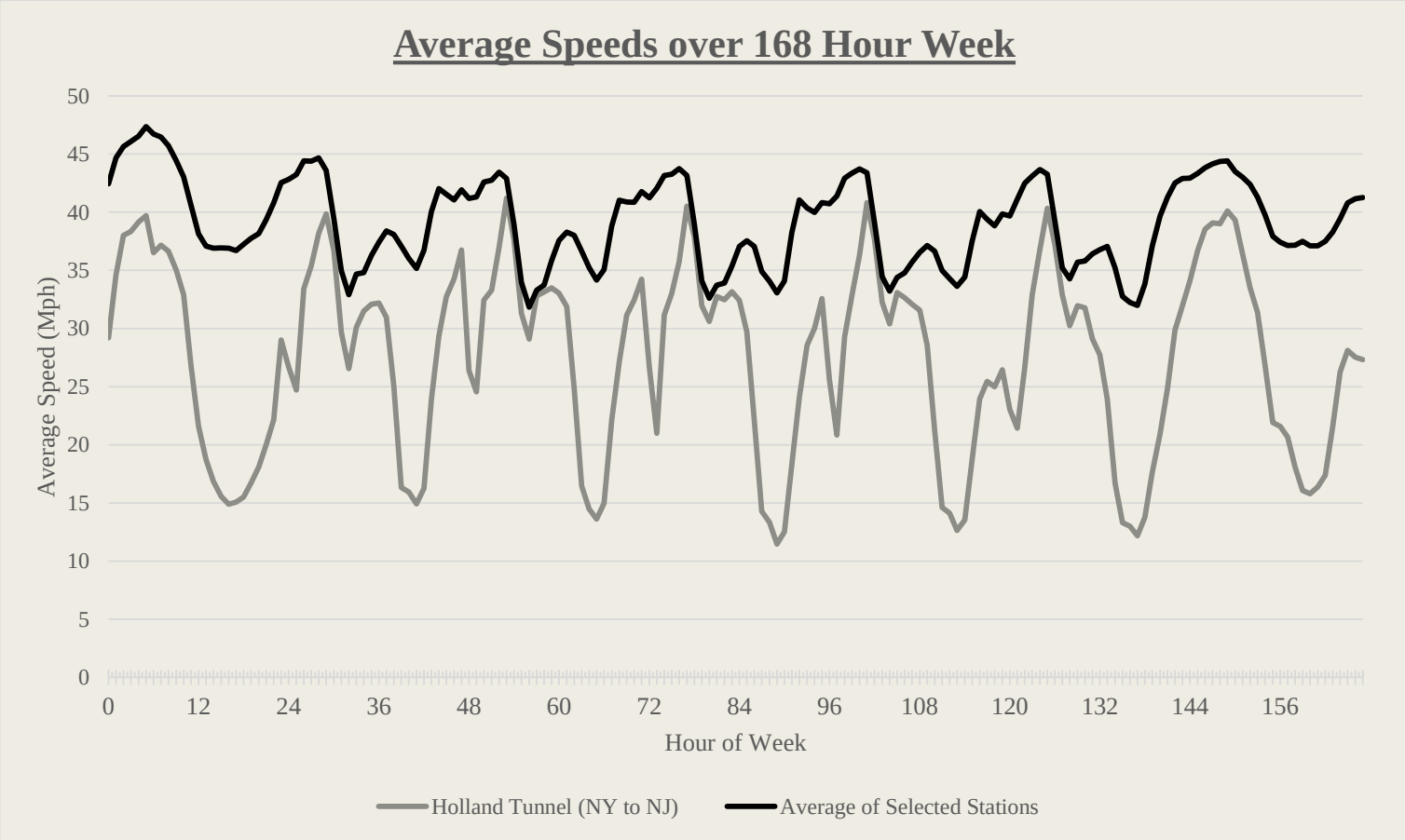
Results of Map Reduce Job



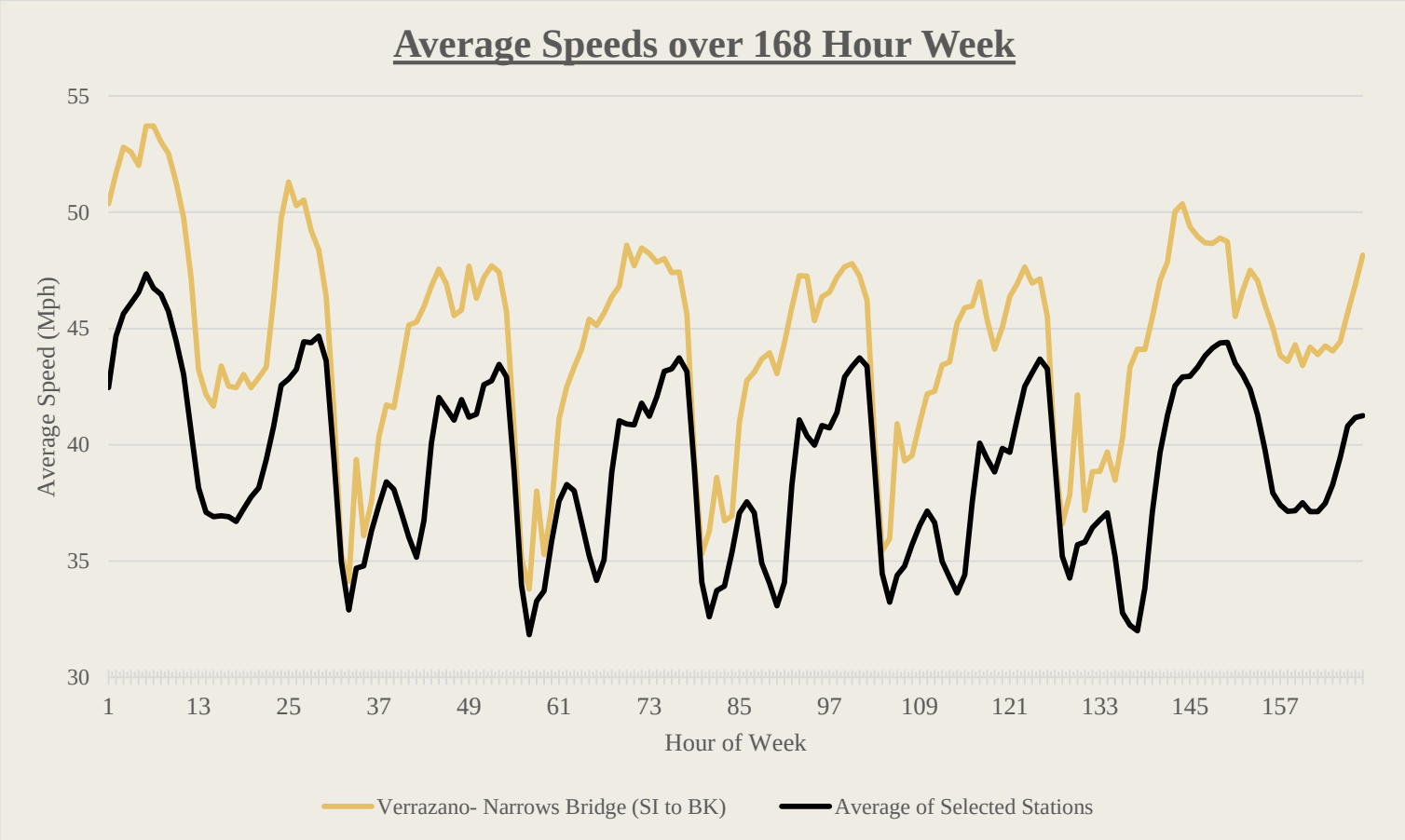
Results of Map Reduce Job

Statistic	Value
Duration	3 minutes 6 seconds
File Written	14.21765 MB
HDFS Written	0.672917 MB
S3 Bytes Read	7910.784328 MB (7.9 GB)
Map Input Records	79904047
Map Functions Completed	29
Reduce Functions Completed	31

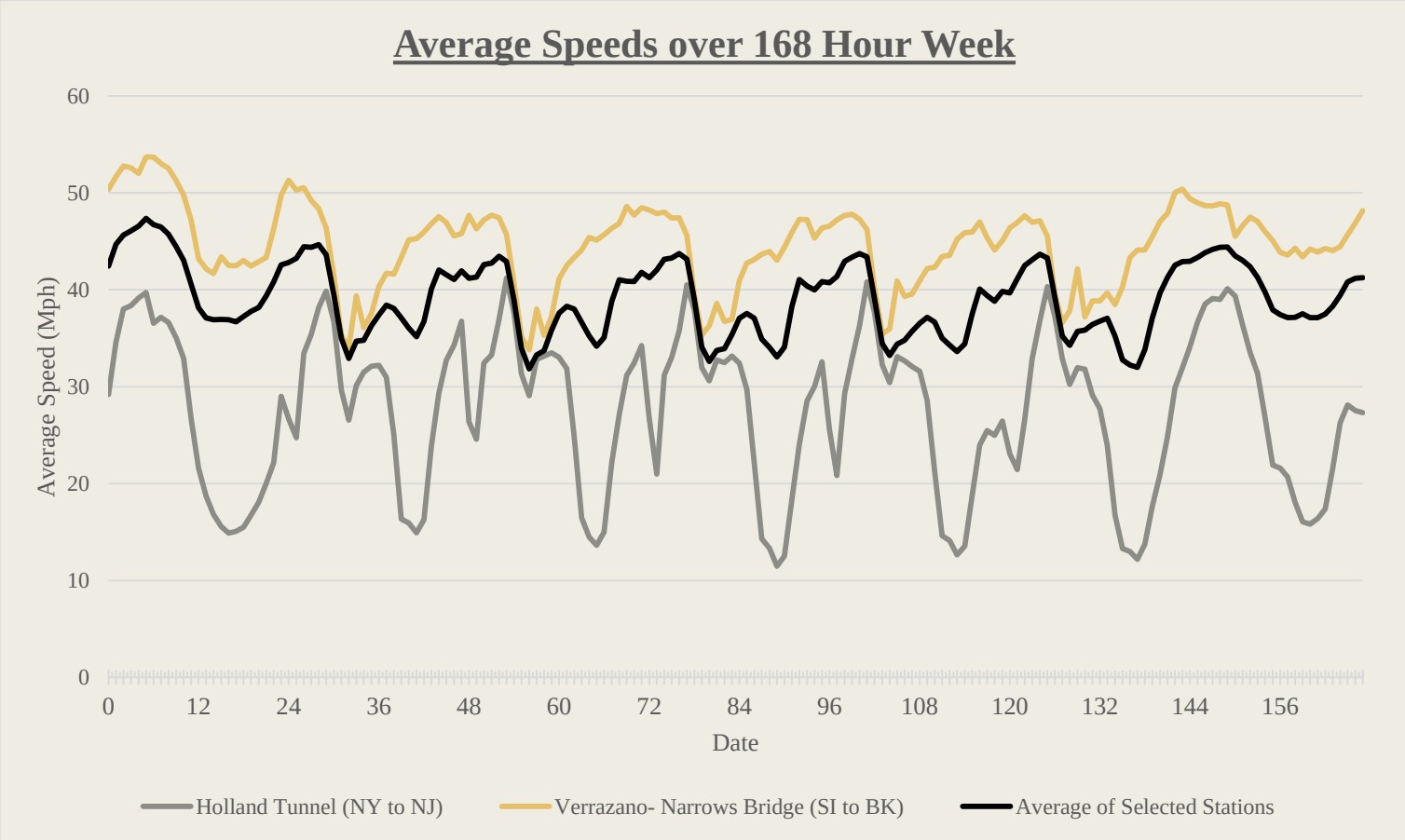
Analysis


Average Speeds over 168 Hour Week
Average Speed (Mph)
50
45
40
35
30
25
20
15
10
5
0
0
12
24
36
48
60
72
84
96
108
120
132
144
156
Hour of Week
Holland Tunnel (NY to NJ)
Average of Selected Stations

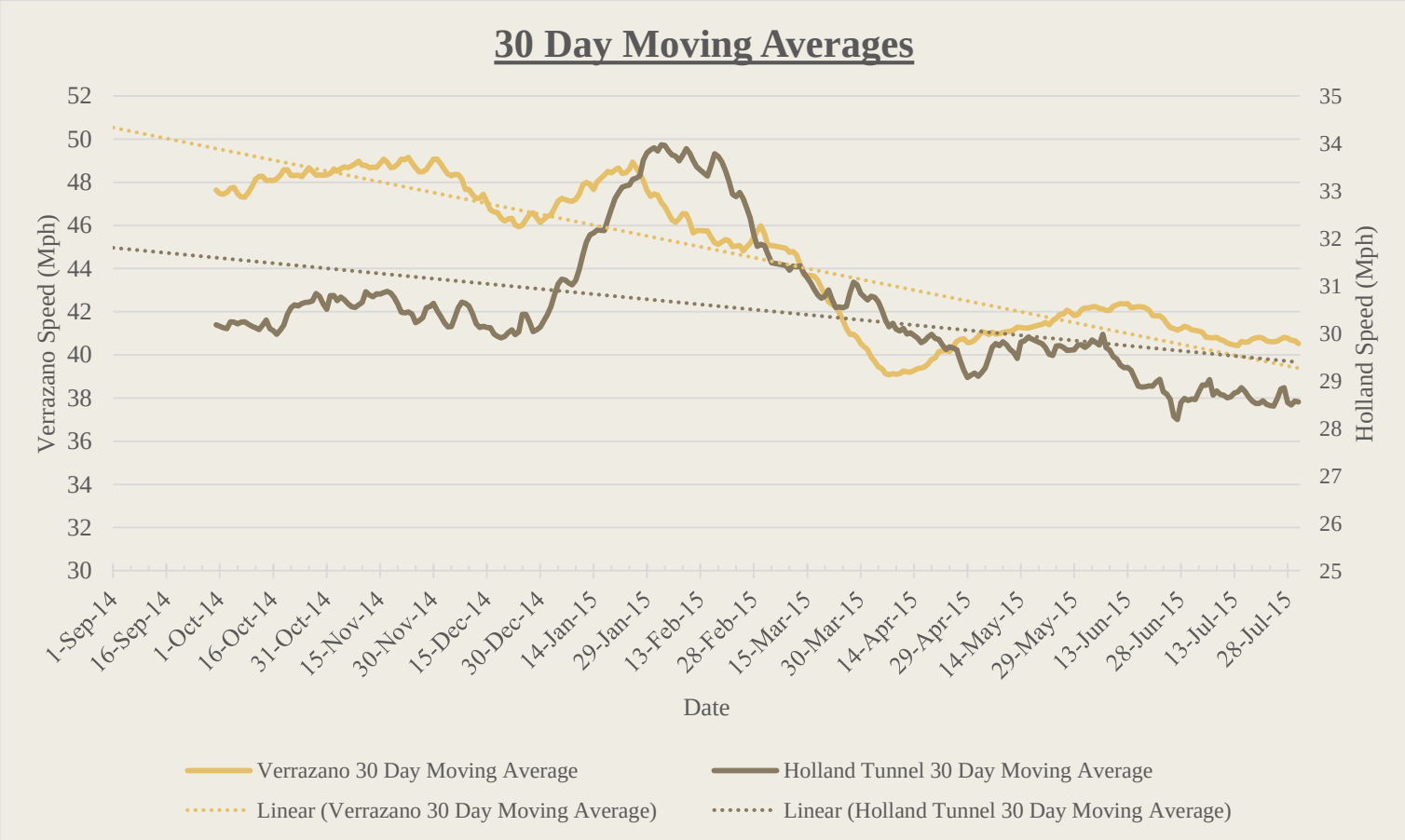
Analysis


Average Speeds over 168 Hour Week
55
50
45
40
35
30
Average Speed (Mph)
1
13
25
37
49
61
73
85
97
109
121
133
145
157
Hour of Week
Verrazano- Narrows Bridge (SI to BK)
Average of Selected Stations

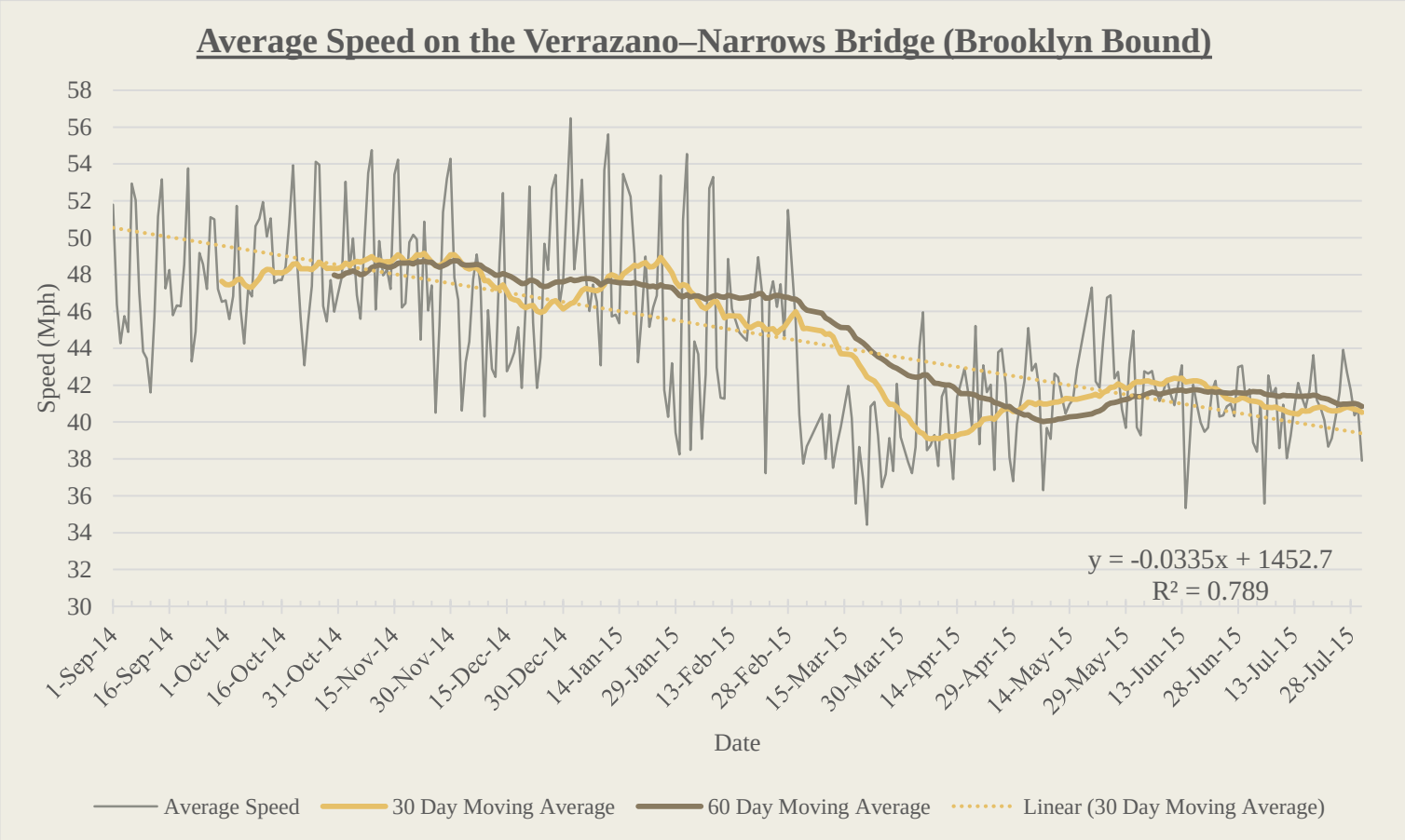
Analysis


Average Speeds over 168 Hour Week
60
50
40
30
20
10
0
Average Speed (Mph)
0
12
24
36
48
60
72
84
96
108
120
132
144
156
Date
Holland Tunnel (NY to NJ)
Verrazano- Narrows Bridge (SI to BK)
Average of Selected Stations

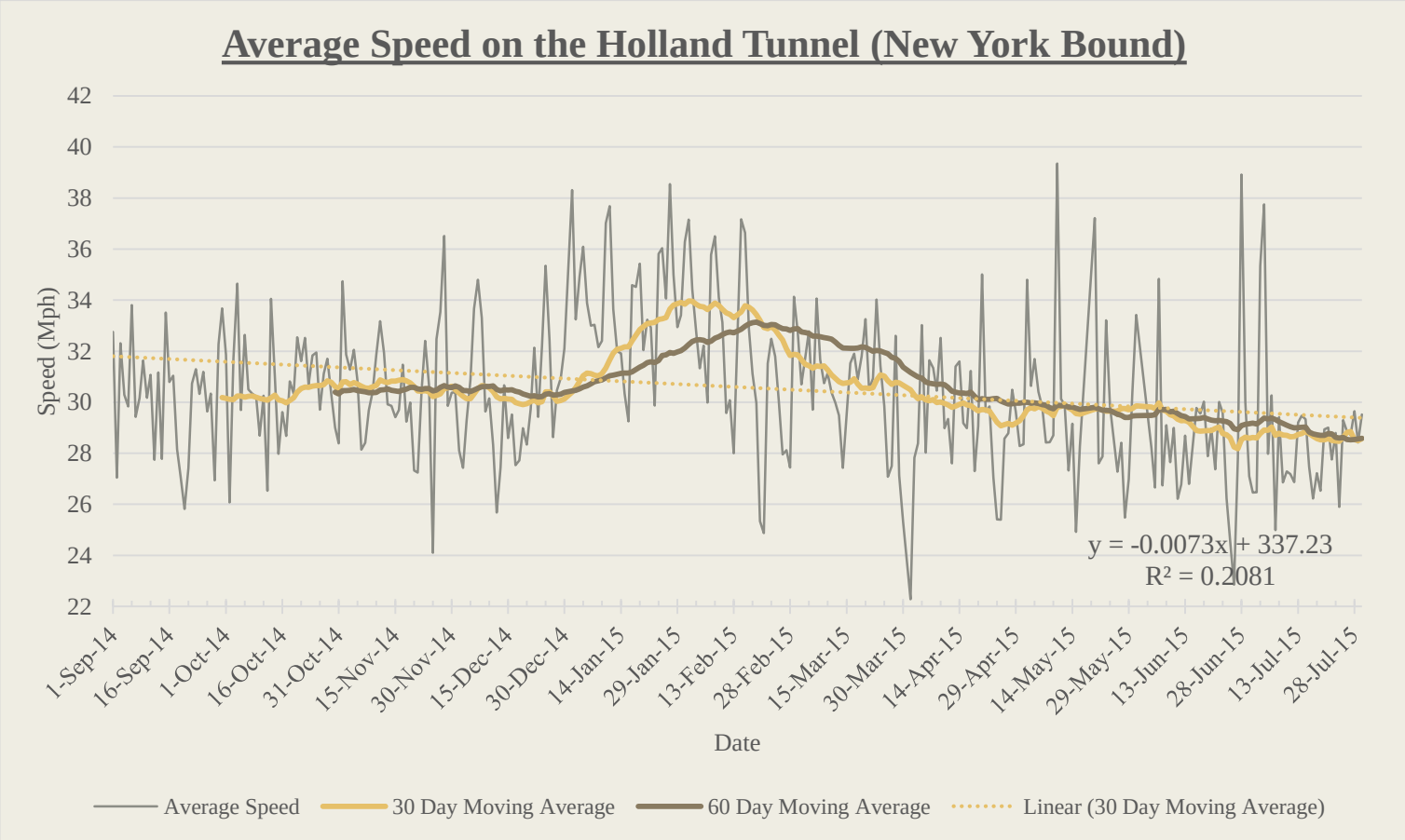
Analysis


30 Day Moving Averages
Verrazano Speed (Mph)
Holland Speed (Mph)
52
50
48
46
44
42
40
38
36
34
32
30
35
34
33
32
31
30
29
28
27
26
25
1-Sep-14
16-Sep-14
1-Oct-14
16-Oct-14
31-Oct-14
15-Nov-14
30-Nov-14
15-Dec-14
30-Dec-14
14-Jan-15
29-Jan-15
13-Feb-15
28-Feb-15
15-Mar-15
30-Mar-15
14-Apr-15
29-Apr-15
14-May-15
29-May-15
13-Jun-15
28-Jun-15
13-Jul-15
28-Jul-15
Date
Verrazano 30 Day Moving Average
Holland Tunnel 30 Day Moving Average
Linear (Verrazano 30 Day Moving Average)
Linear (Holland Tunnel 30 Day Moving Average)

Analysis


Average Speed on the Verrazano–Narrows Bridge (Brooklyn Bound)
Speed (Mph)
58
56
54
52
50
48
46
44
42
40
38
36
34
32
30
y = -0.0335x + 1452.7
R² = 0.789
1-Sep-14
16-Sep-14
1-Oct-14
16-Oct-14
31-Oct-14
15-Nov-14
30-Nov-14
15-Dec-14
30-Dec-14
14-Jan-15
29-Jan-15
13-Feb-15
28-Feb-15
15-Mar-15
30-Mar-15
14-Apr-15
29-Apr-15
14-May-15
29-May-15
13-Jun-15
28-Jun-15
13-Jul-15
28-Jul-15
Date
Average Speed
30 Day Moving Average
60 Day Moving Average
Linear (30 Day Moving Average)

Analysis


Average Speed on the Holland Tunnel (New York Bound)
Speed (Mph)
42
40
38
36
34
32
30
28
26
24
22
y = -0.0073x + 337.23
R² = 0.2081
1-Sep-14
16-Sep-14
1-Oct-14
16-Oct-14
31-Oct-14
15-Nov-14
30-Nov-14
15-Dec-14
30-Dec-14
14-Jan-15
29-Jan-15
13-Feb-15
28-Feb-15
15-Mar-15
30-Mar-15
14-Apr-15
29-Apr-15
14-May-15
29-May-15
13-Jun-15
28-Jun-15
13-Jul-15
28-Jul-15
Date
Average Speed
30 Day Moving Average
60 Day Moving Average
Linear (30 Day Moving Average)

Regression Analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.532820115
R Square	**0.283897275**
Adjusted R Square	0.281436441
Standard Error	2.852563774
Observations	293

ANOVA

	df	*SS*	*MS*	*F*
Regression	1.00E+00	9.39E+02	9.39E+02	1.15E+02
Residual	2.91E+02	2.37E+03	8.14E+00	
Total	2.92E+02	3.31E+03		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	5.85E+00	3.60E+00	1.62E+00	1.06E-01
HOT30Day	1.27E+00	1.18E-01	1.07E+01	6.89E-23

Low Periods: VNZ to Brooklyn

Rank	Speed (MPH)	HOW	Time (EST)
168	33.78938594	56	Tuesday 8am
167	34.12049655	32	Monday 8am
166	35.14218241	55	Tuesday 7am
165	35.27610664	31	Monday 7am
164	35.28588222	58	Tuesday 10am

Low Periods: Holland Tunnel to NY

Rank	Speed (MPH)	HOW	Time (EST)
168	13.75552926	138	Friday 7pm
167	12.171702450	137	Friday 6pm
166	13.52144944	114	Thursday 7pm
165	15.08261256	17	Thursday 6pm
164	15.49752670	18	Thursday 5pm

Conclusions

- How can Amazon Web Services be used to conduct analysis of large scale data sets?
  - *Amazon Web Services is an effective resource to analyze large scale data sets*
  - *Data is stored into the Hadoop File System using Amazon S3 Storage Systems*
  - *Data processed using Map Reduce after pre-processing*
- How does the average speed of the Verrazano- Narrows Bridge and the Holland tunnel fluctuate?
  - *Highs:*
    - VZN to Brooklyn: 2 am
    - HOT to NY: 4 am
  - *Lows:*
    - VZN to Brooklyn: 7 am
    - HOT to NY: 5 pm

Further Research

- Predictive Analysis to:
  - *Determine the speed at a given time*
  - *Determine the best route using real time traffic conditions*