test Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

test

default

FINISHED

```
%pyspark
### Import Libraries
import pandas as pd
from pyspark.sql.types import StringType
from pyspark import SQLContext
import numpy as np  # arrays and numerical processing
from sklearn.cluster import KMeans  # cluster analysis by partitioning
from sklearn.metrics import silhouette_score as silhouette_score
from __future__ import division, print_function
import plotly.plotly as py
import plotly.graph_objs as go
from datetime import datetime
from pyspark.sql import functions as F
from pyspark.sql.functions import col,udf, unix_timestamp
from pyspark.sql.types import DateType
from pyspark.sql.functions import from_unixtime
from pyspark.sql.functions import col, unix_timestamp, round
```

Took 1 sec. Last updated by anonymous at March 25 2017, 12:55:41 PM.

FINISHED

```
%pyspark
road_df1 = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load
/*.csv")
road_df2 = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load
/*.csv", schema=custschema)
road_df3 = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load
/*.csv")
```

Took 10 sec. Last updated by anonymous at March 25 2017, 12:55:58 PM.

FINISHED

```
%pyspark
road_df.printSchema()
```

```
root
 |-- status: string (nullable = true)
 |-- avgMeasuredTime: string (nullable = true)
 |-- avgSpeed: string (nullable = true)
 |-- extID: string (nullable = true)
 |-- medianMeasuredTime: string (nullable = true)
 |-- TIMESTAMP: string (nullable = true)
 |-- vehicleCount: string (nullable = true)
 |-- _id: string (nullable = true)
 |-- REPORT_ID: string (nullable = true)
```

Took 1 sec. Last updated by anonymous at March 24 2017, 3:32:59 PM.

test

test Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

test

default

FINISHED

```
%pyspark
from pyspark.sql.types import *
custschema = StructType([
StructField("status", StringType(), True),
StructField("avgMeasuredTime", StringType(), True),
StructField("avgSpeed", StringType(), True),
StructField("extID", StringType(), True),
StructField("medianMeasuredTime", StringType(), True),
StructField("TIMESTAMP", StringType(), True),
StructField("vehicleCount", StringType(), True),
StructField("_id", StringType(), True),
StructField("REPORT_ID", StringType(), True)
])
```

Took 0 sec. Last updated by anonymous at March 24 2017, 3:53:52 PM.

FINISHED

```
%pyspark
```

Took 0 sec. Last updated by anonymous at March 24 2017, 4:03:12 PM.

FINISHED

```
%pyspark
road_columns = list(road_df.columns)

meta_df = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load(
```

Took 0 sec. Last updated by anonymous at March 24 2017, 4:07:10 PM.

FINISHED

```
%pyspark
road_df1 = road_df1.unionAll(road_df2)
road_df1.count()
```

```
19276612
```

Took 17 sec. Last updated by anonymous at March 24 2017, 4:05:45 PM.

FINISHED

```
%pyspark
road_df1 = road_df1.unionAll(road_df3)
road_df1.count()
```

```
33551334
```

Took 27 sec. Last updated by anonymous at March 24 2017, 4:06:28 PM.

FINISHED

```
%pyspark

meta_df = meta_df.drop('extID')
meta_df = meta_df.drop('_id')
metadata_columns = list(meta_df.columns)
merged_df = road_df1.join(meta_df, (road_df1.REPORT_ID == meta_df.REPORT_ID) ).drop(meta_df.RI
```

test

Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

test

default

```
 |-- status: string (nullable = true)
 |-- avgMeasuredTime: string (nullable = true)
 |-- avgSpeed: string (nullable = true)
 |-- extID: string (nullable = true)
 |-- medianMeasuredTime: string (nullable = true)
 |-- TIMESTAMP: string (nullable = true)
 |-- vehicleCount: string (nullable = true)
 |-- _id: string (nullable = true)
 |-- REPORT_ID: string (nullable = true)
 |-- POINT_1_STREET: string (nullable = true)
 |-- DURATION_IN_SEC: string (nullable = true)
 |-- POINT_1_NAME: string (nullable = true)
 |-- POINT_1_CITY: string (nullable = true)
 |-- POINT_2_NAME: string (nullable = true)
 |-- POINT_2_LNG: string (nullable = true)
 |-- POINT_2_STREET: string (nullable = true)
 |-- NDT_IN_KMH: string (nullable = true)
```

Took 1 sec. Last updated by anonymous at March 24 2017, 4:07:30 PM.

FINISHED

```
%pyspark

road_merged_df = merged_df.select([c for c in merged_df.columns if c not in{'DURATION_IN_SEC'
road_merged_df.printSchema()

#type( rd_df)
```

```
root
 |-- status: string (nullable = true)
 |-- avgMeasuredTime: string (nullable = true)
 |-- avgSpeed: string (nullable = true)
 |-- extID: string (nullable = true)
 |-- medianMeasuredTime: string (nullable = true)
 |-- TIMESTAMP: string (nullable = true)
 |-- vehicleCount: string (nullable = true)
 |-- _id: string (nullable = true)
 |-- REPORT_ID: string (nullable = true)
 |-- POINT_1_STREET: string (nullable = true)
 |-- POINT_1_NAME: string (nullable = true)
 |-- POINT_1_CITY: string (nullable = true)
 |-- POINT_2_LNG: string (nullable = true)
 |-- POINT_2_STREET: string (nullable = true)
 |-- NDT_IN_KMH: string (nullable = true)
 |-- POINT_2_POSTAL_CODE: string (nullable = true)
 |-- POINT_2_COUNTRY: string (nullable = true)
```

Took 0 sec. Last updated by anonymous at March 24 2017, 4:08:01 PM.

FINISHED

```
%pyspark
rd_df = road_merged_df.select([c for c in road_merged_df.columns if c not in
{'DURATION_IN_SEC','DISTANCE_IN_METERS','POINT_2_NAME'}])
```

test Intitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

Took 0 sec. Last updated by anonymous at March 24 2017, 4:08:08 PM.

test

FINISHED default

```
%pyspark
format = "yyyy-MM-dd'T'HH:mm:ss"
rd_df = rd_df.select('avgMeasuredTime','avgSpeed','TIMESTAMP','vehicleCount','REPORT_ID','POI
,'POINT_1_LNG',  from_unixtime(unix_timestamp('TIMESTAMP',format)).cast("timestamp").alias('d
rd_data_selected = rd_df.select([c for c in rd_df.columns if c not in
{'TIMESTAMP','REPORT_ID','avgMeasuredTime','vehicleCount'}])
rd_data_sel = rd_data_selected.select('avgSpeed','POINT_2_LNG','POINT_1_LAT','POINT_2_LAT','P
```

Took 1 sec. Last updated by anonymous at March 24 2017, 6:58:59 PM.

FINISHED

```
%pyspark
#newdf = elevDF.select(year(elevDF.date).alias('dt_year'), month(elevDF.date).alias('dt_month
dayofyear(elevDF.date).alias('dt_dayofy'), hour(elevDF.date).alias('dt_hour'), minute(elevDF.
).alias('dt_week_no'), unix_timestamp(elevDF.date).alias('dt_int'))
```

Took 0 sec. Last updated by anonymous at March 24 2017, 6:58:39 PM.

FINISHED

```
%pyspark
rd_data_sel.show()
```

```
+--------+-----------------+-----------------+-----------------+------------------+----+
|avgSpeed|      POINT_2_LNG|      POINT_1_LAT|      POINT_2_LAT|       POINT_1_LNG|hour|
+--------+-----------------+-----------------+-----------------+------------------+----+
|      93|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      90|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      84|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      88|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      90|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      88|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      87|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      84|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      88|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      90|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      80|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      79|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      85|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      87|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
|      88|10.10702759027481|56.21399075104399|56.21740644105506|10.145073305557275|   2|
```

Took 4 sec. Last updated by anonymous at March 24 2017, 6:56:35 PM.

FINISHED

```
%pyspark
#rd_data_selected.show()
temp = rd_data_selected.select('avgSpeed','date',F. weekofyear('date').alias('weekofyear') )
```

test

test

Zeppelin

```
%pyspark
temp1 = temp.groupBy('weekofyear').agg(mean('avgSpeed').alias("mean"))
```

FINISHED

default

Took 0 sec. Last updated by anonymous at March 24 2017, 7:05:41 PM.

```
%pyspark
temp1.repartition(1).write.csv("/Users/jyothi/Downloads/timecsv2")
```

FINISHED

Took 37 min 22 sec. Last updated by anonymous at March 24 2017, 7:43:17 PM.

```
%pyspark
temp1.count()
```

FINISHED

34

Took 3 min 4 sec. Last updated by anonymous at March 25 2017, 11:22:08 AM.

```
%pyspark
#type(rd_data_sel)
from pyspark.sql.functions import col, unix_timestamp, round
from pyspark.sql.functions import hour, mean
rd_data_set = rd_data_sel.groupBy('hour','POINT_1_LAT' , 'POINT_1_LNG','POINT_2_LAT','POINT_2_
```

FINISHED

Took 0 sec. Last updated by anonymous at March 24 2017, 4:08:46 PM.

```
%pyspark
rd_data_set =rd_data_set.withColumn("mean", round(rd_data_set.mean, 3))
rd_data_set =rd_data_set.withColumn("POINT_1_LNG", round(rd_data_set.POINT_1_LNG, 9))
rd_data_set =rd_data_set.withColumn("POINT_2_LNG", round(rd_data_set.POINT_2_LNG, 9))
rd_data_set =rd_data_set.withColumn("POINT_1_LAT", round(rd_data_set.POINT_1_LAT, 9))
rd_data_set =rd_data_set.withColumn("POINT_2_LAT", round(rd_data_set.POINT_1_LAT, 9))
rd_data_set.registerTempTable("latdata")
rd_data_set.show()
```

FINISHED

```
+----+------------+------------+------------+------------+------+
|hour| POINT_1_LAT| POINT_1_LNG| POINT_2_LAT| POINT_2_LNG|  mean|
+----+------------+------------+------------+------------+------+
|   2|56.230013293|10.293740482|56.230013293|10.277436491|72.638|
|   4|56.139627711|10.177235595|56.139627711|10.154338146|68.362|
|   5|56.153135888|10.139060373|56.153135888|10.157440225|38.712|
|   8|56.159418709|10.187763761|56.159418709|10.182659856|15.359|
|   1|56.155914111|10.176133542|56.155914111|10.182616306|38.772|
|  20|56.141823496|10.204370308|56.141823496|10.198852845|40.106|
|   8|56.141823496|10.204370308|56.141823496|10.198852845|33.839|
|  12|56.210806442|10.269670627|56.210806442|10.267762216|50.721|
|   6|56.170278565|10.186652602|56.170278565|10.194526469|15.809|
|  12|56.163896974|10.186162247|56.163896974|10.194526469|34.162|
|   6|56.178710604|10.184485324|56.178710604|10.182183959|29.174|
|  22|56.154610138|10.157440225|56.154610138|10.139025504|46.526|
```

test

Zeppelin

Took 2 min 26 sec. Last updated by anonymous at March 24 2017, 4:11:15 PM.

test

default

FINISHED

```
%pyspark
type(rd_data_set)
```

```
<class 'pyspark.sql.dataframe.DataFrame'>
```

Took 0 sec. Last updated by anonymous at March 23 2017, 7:39:40 AM.

FINISHED

```
%pyspark
rd_data_set.count()
```

```
10632
```

Took 1 min 57 sec. Last updated by anonymous at March 24 2017, 4:13:19 PM.

FINISHED

```
%pyspark
rd_data_set.repartition(1).write.csv("/Users/jyothi/Downloads/myc2.csv")
```

Took 2 min 17 sec. Last updated by anonymous at March 24 2017, 4:41:41 PM.

FINISHED

```
%pyspark

final_df= sqlContext.read.format("com.databricks.spark.csv").option("header", "true").load(
```

Took 1 sec. Last updated by anonymous at March 24 2017, 1:12:25 AM.

FINISHED

```
%pyspark

final_df.count()
```

```
16032
```

Took 2 sec. Last updated by anonymous at March 24 2017, 1:12:44 AM.

FINISHED

```
%pyspark
final_df.repartition(1).write.csv("/Users/jyothi/Downloads/final.csv")
```

Took 2 sec. Last updated by anonymous at March 24 2017, 1:13:34 AM.

FINISHED

```
%r
library(readr)
library(dplyr)
library(data.table)
library(sqldf)
library(ggplot2)
library(ggmap)
library(maps)
data1 <- read_csv("/Users/jyothi/Downloads/myc1.csv/road.csv")
```

test Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled |Untitled

Took 12 sec. Last updated by anonymous at March 24 2017, 5:12:39 PM.

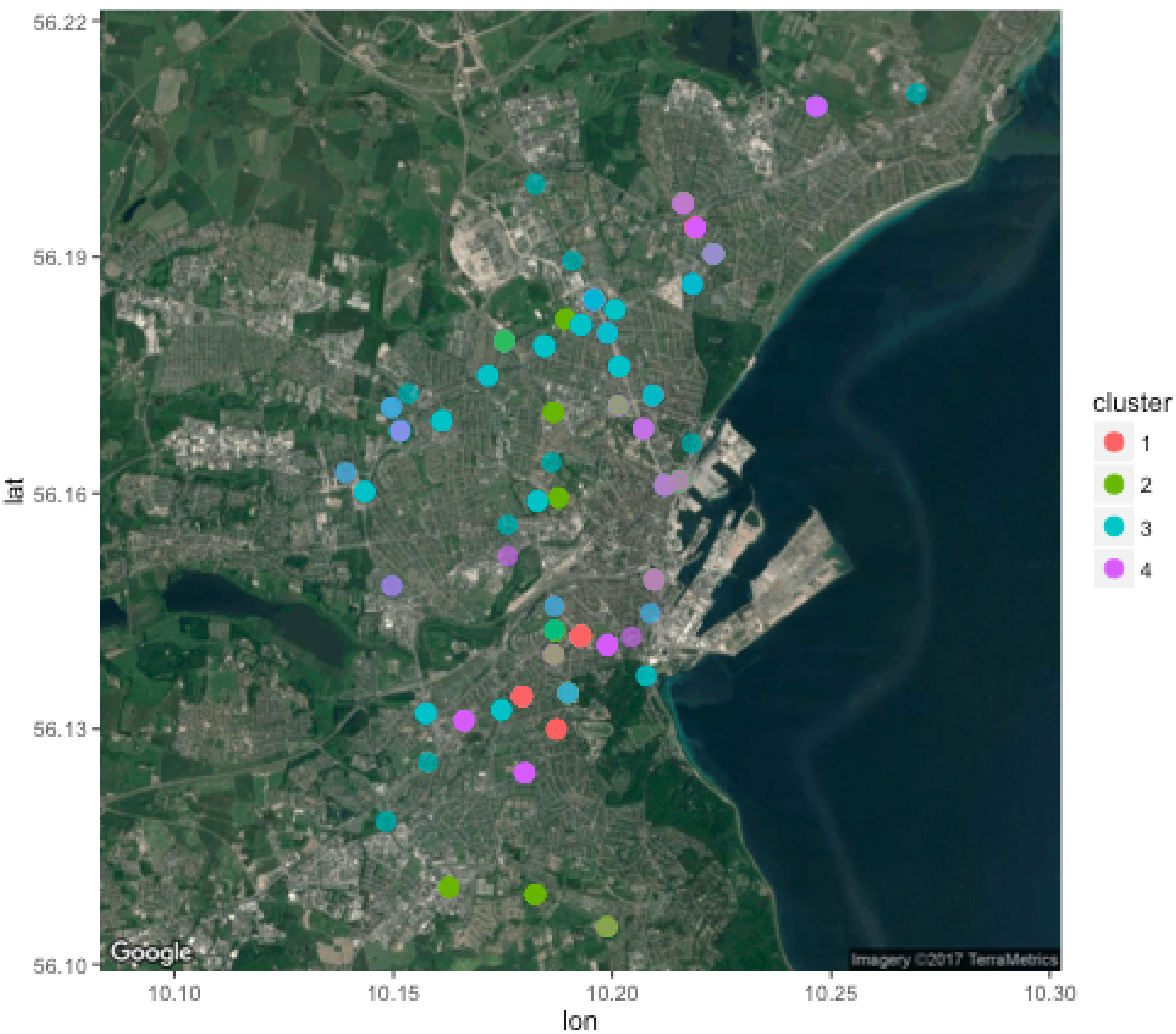
Zeppelin

test

FINISHED

default

```
%r
tbl1<-subset(data1, data1$hr_group == 1 & data1$mean < 30)
#tbl1<-subset(tbl, data1$mean > 50)
result4<-kmeans(tbl1,4)
tbl1$cluster <- result4$cluster
#View(tbl1)
tbl1$cluster <- as.factor(tbl1$cluster)
mapgilbert <- get_map(location = c(lon = mean(tbl1$long1), lat = mean(tbl1$lat1)), zoom = 12,
# plotting the map with some points on it
ggmap(mapgilbert) + geom_point(data = tbl1, aes(x = as.numeric(long1), y = as.numeric(lat1),c
= 20) + guides(fill=FALSE, alpha=FALSE, size=FALSE)
```



Took 5 sec. Last updated by anonymous at March 24 2017, 5:13:30 PM.

test Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

Zeppelin

FINISHED

test

default

```
%r
tbl2<-subset(data1, data1$hr_group == 2 & data1$mean < 30)
result4<-kmeans(tbl2,4)
tbl2$cluster <- result4$cluster
#View(tbl2)
tbl2$cluster <- as.factor(tbl2$cluster)
mapgilbert <- get_map(location = c(lon = mean(tbl2$long1), lat = mean(tbl2$lat1)), zoom = 11,
# plotting the map with some points on it
ggmap(mapgilbert) + geom_point(data = tbl2, aes(x = as.numeric(long1), y = as.numeric(lat1),c
= 20) + guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

Took 6 sec. Last updated by anonymous at March 24 2017, 5:14:06 PM.

FINISHED

test

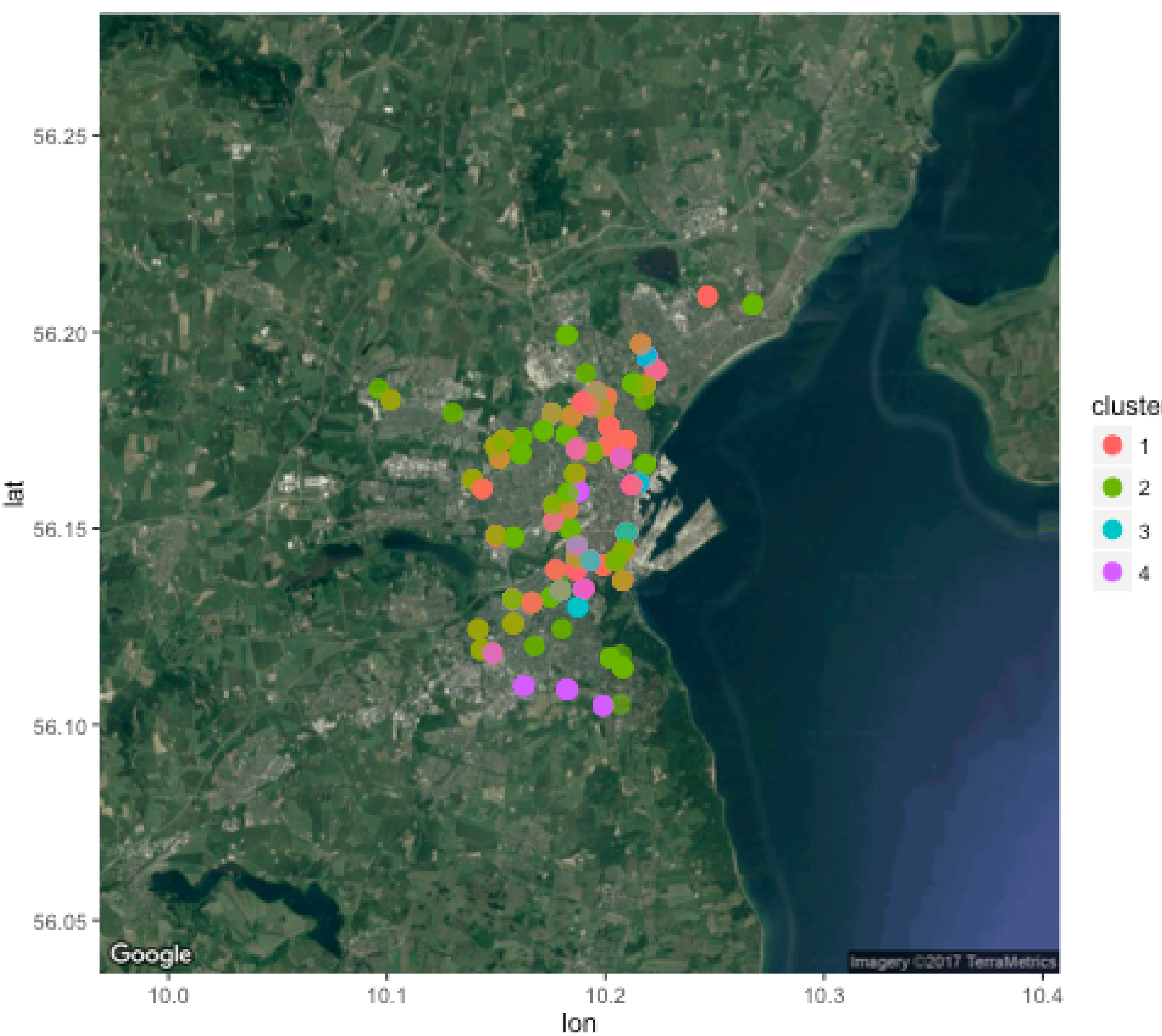
```
%r
```

Zeppelin

test

default

```
result4<-kmeans(tbl3,4)
tbl3$cluster <- result4$cluster
#View(tbl3)
tbl3$cluster <- as.factor(tbl3$cluster)
mapgilbert <- get_map(location = c(lon = mean(tbl3$long1), lat = mean(tbl3$lat1)), zoom = 11,
# plotting the map with some points on it
ggmap(mapgilbert) + geom_point(data = tbl3, aes(x = as.numeric(long1), y = as.numeric(lat1),c
shape = 20) + guides(fill=FALSE, alpha=FALSE, size=FALSE)
```



Took 5 sec. Last updated by anonymous at March 24 2017, 5:14:52 PM.

FINISHED

```
%r
tbl4<-subset(data1, data1$hr_group == 4 & data1$mean < 30)
result4<-kmeans(tbl4,4)
tbl4$cluster <- result4$cluster
```

test

Zeppelin

Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

```
tbl4$cluster<-as.factor(tbl4$cluster)
mapgilbert <- get_map(location = c(lon = mean(tbl4$long1), lat = mean(tbl4$lat1)), zoom = 12,
# plotting the map with some points on it
ggmap(mapgilbert) + geom_point(data = tbl4, aes(x = as.numeric(long1), y = as.numeric(lat1), c
= 20)  + guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

default

```
Error in check_for_XQuartz(): X11 library is missing: install XQuartz from xquart
z.macosforge.org
```

Took 5 sec. Last updated by anonymous at March 24 2017, 5:15:14 PM.

FINISHED

```
%r
tbl5<-subset(data1, data1$hr_group == 5 & data1$mean < 30)
result4<-kmeans(tbl5,4)
```

test

test Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled Untitled

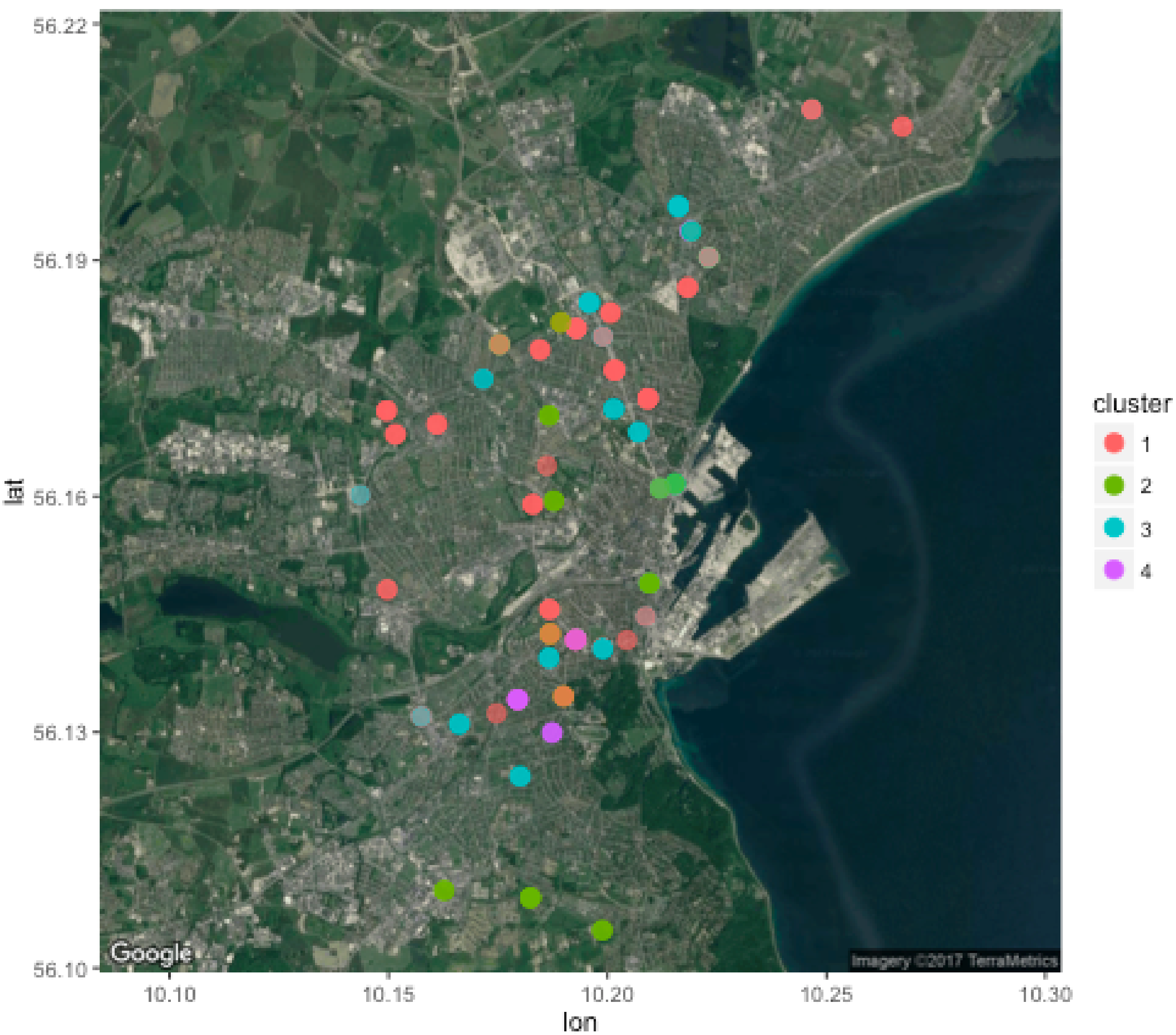
Zeppelin

test

default

```
...w(tbl5)
tbl5$cluster <- as.factor(tbl5$cluster)
mapgilbert <- get_map(location = c(lon = mean(tbl5$long1), lat = mean(tbl5$lat1)), zoom = 12,
# plotting the map with some points on it
ggmap(mapgilbert) + geom_point(data = tbl5, aes(x = as.numeric(long1), y = as.numeric(lat1)
shape = 20) + guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

```
Error in check_for_XQuartz(): X11 library is missing: install XQuartz from xquart
z.macosforge.org
```



Took 5 sec. Last updated by anonymous at March 24 2017, 5:15:35 PM.

Zeppelin

READY

test

default