Facebook/01. Data…

Import libraries

FINISHED

```
import org.apache.spark._
import org.apache.spark.SparkContext._
import org.apache.spark.rdd._
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._

val spark = SparkSession.builder().getOrCreate()

import spark.implicits._
```

```
import org.apache.spark._
import org.apache.spark.SparkContext._
import org.apache.spark.rdd._
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._
spark: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@43abb6e0
import spark.implicits._
```

Took 4 sec. Last updated by anonymous at November 09 2017, 8:49:03 PM. (outdated)

Import training set

FINISHED

```
val trainRaw = spark.read
    .option("header","true")
    .option("mode","DROPMALFORMED")
    .csv("/Users/Albert/Data Science/Kaggle
        /Facebook/train.csv")
```

```
trainRaw: org.apache.spark.sql.DataFrame = [row_id: string, x: string ... 4 more fields]
```

Took 0 sec. Last updated by anonymous at November 09 2017, 8:54:33 PM. (outdated)

Count number of rows in train

FINISHED

```
trainRaw.count
```

```
res7: Long = 29118021
```

Took 24 sec. Last updated by anonymous at November 09 2017, 8:54:59 PM. (outdated)

Print schema of train set

FINISHED

```
trainRaw.printSchema()
```

```
root
 |-- row_id: string (nullable = true)
 |-- x: string (nullable = true)
 |-- y: string (nullable = true)
 |-- accuracy: string (nullable = true)
 |-- time: string (nullable = true)
 |-- place_id: string (nullable = true)
```

Took 22 sec. Last updated by anonymous at November 09 2017, 8:54:59 PM. (outdated)

Recast columns to appropriate types

FINISHED

```
val train = trainRaw.select($"row_id".cast(IntegerType).as("row_id")
                        ,   $"x".cast(FloatType).as("x")
                        ,   $"y".cast(FloatType).as("y")
                        ,   $"accuracy".cast(IntegerType).as("accuracy")
                        ,   $"time".cast(IntegerType).as("time")
                        ,   $"place_id".cast(LongType).as("place_id")
                        )
train.printSchema()
```

```
train: org.apache.spark.sql.DataFrame = [row_id: int, x: float ... 4 more fields]
root
 |-- row_id: integer (nullable = true)
 |-- x: float (nullable = true)
 |-- y: float (nullable = true)
 |-- accuracy: integer (nullable = true)
 |-- time: integer (nullable = true)
 |-- place_id: long (nullable = true)
```

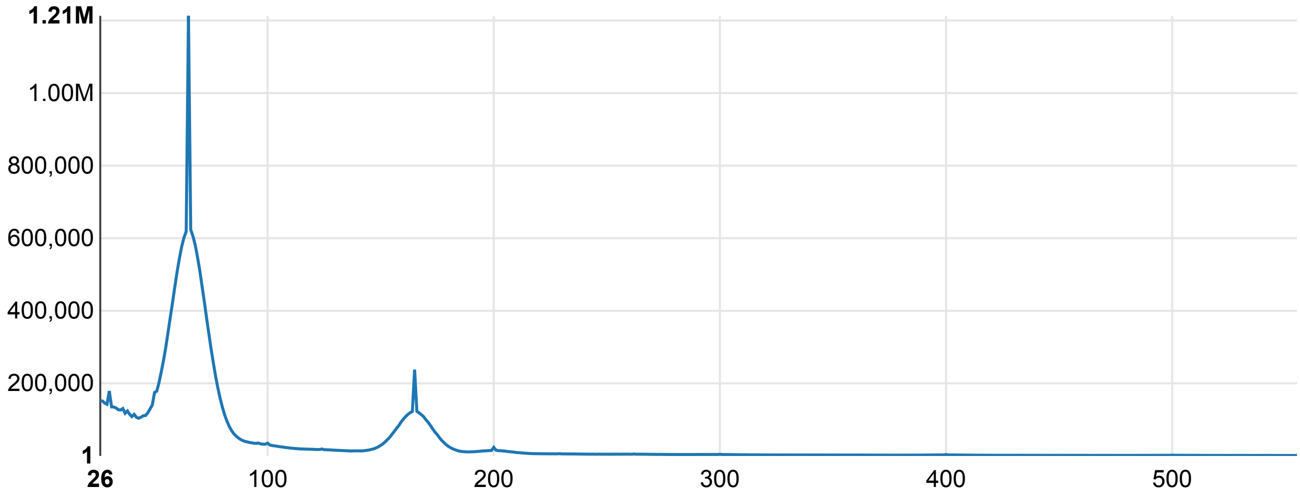
Took 2 sec. Last updated by anonymous at November 09 2017, 8:57:34 PM. (outdated)

Histogram of accuracy

FINISHED

```
z.show(train.groupBy($"accuracy").agg(count("*").as("counts")).sort($"accuracy".desc))
```

settings



```
Results are limited by 1000.
```

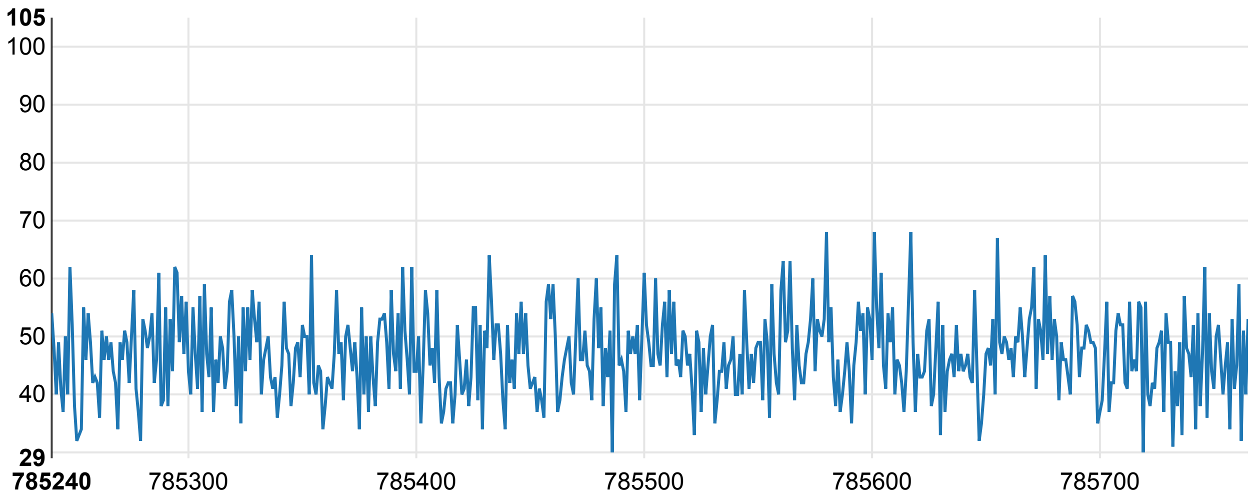
Took 31 sec. Last updated by anonymous at November 09 2017, 9:15:42 PM. (outdated)

Histogram of time

FINISHED

```
z.show(train.groupBy($"time").agg(count("*").as("counts")).sort($"time".desc))
```

settings



```
Results are limited by 1000.
```

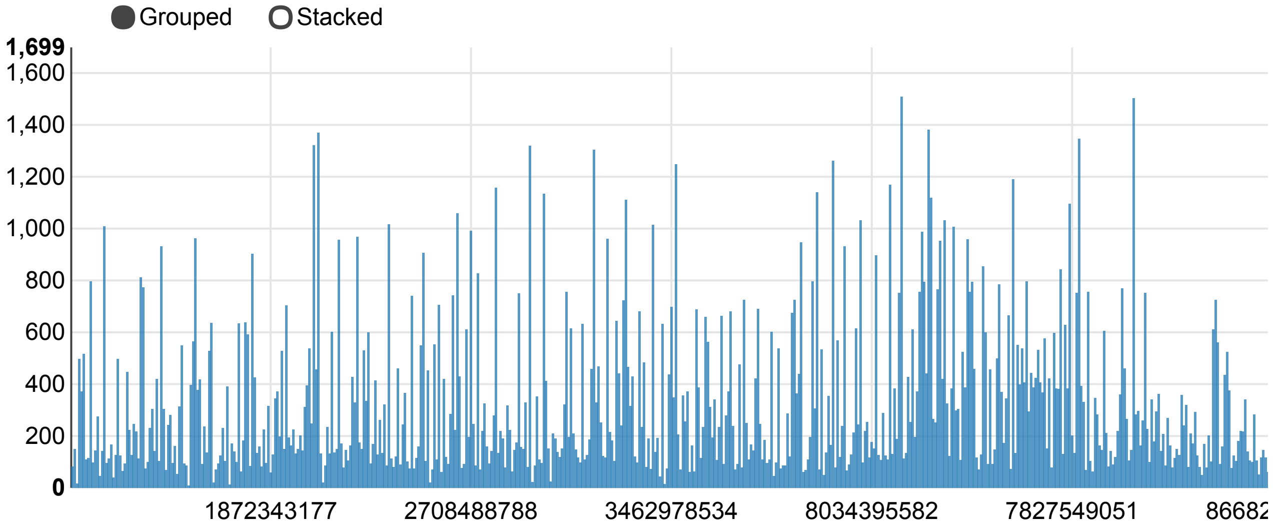
Took 36 sec. Last updated by anonymous at November 09 2017, 9:14:58 PM. (outdated)

Histogram of place_id

FINISHED

```
z.show(train.groupBy($"place_id").agg(count("*").as("counts")))
```

settings ▾



```
Results are limited by 1000.
```

Took 32 sec. Last updated by anonymous at November 09 2017, 9:16:38 PM. (outdated)

READY