Στατιστικά και Οπτικοποίηση Δεδομένων στις Τροχιές Λεωφορείων της Αθήνας (ΟΑΣΑ)

Απαλλακτική Εργασία Εξαμήνου: Γεωγραφικά και Πληροφοριακά Συστήματα 2021-2022

ΓΕΩΡΓΙΟΣ ΜΠΑΛΤΖΑΚΗΣ
ΑΜ: Π18105
Email: george-baltzakis@hotmail.gr

Περιεχόμενα

Εισαγωγή

- Στην παρούσα εργασία υλοποιήθηκε μία πρώτη στατιστική ανάλυση σε τροχιές λεωφορείων από το σύστημα μέσων μαζικής μεταφοράς της Αθήνας (ΟΑΣΑ Α.Ε.).
- Η συγκεκριμένη βάση δεδομένων αποτελείται από τις στάσεις των λεωφορείων, τις διαδρομές τους και τα γεωγραφικά αποτυπώματα τους.
- Αρχικά εφαρμόζεται η εισαγωγή των τριών πινάκων, έπειτα ακολουθεί μία πρώτη «ματιά» στα υπάρχοντα δεδομένα, ο καθαρισμός και η μορφοποίηση των δεδομένων και τέλος τα στατιστικά στοιχεία που προέκυψαν, συνοδευόμενα απ' τα διαγράμματα τους.
- Για λόφους ευκολίας, δεν συμπεριλήφθηκαν όλες οι εγγραφές του πίνακα με τις τροχιές των λεωφορείων, παρά μόνο ένα διάστημα δύο εβδομάδων.

Προ απαιτούμενα

- Στην συγκεκριμένη εργασία χρησιμοποιήθηκε η γλώσσα προγραμματισμού *Python 3.7* με τη βοήθεια ορισμένων βιβλιοθηκών της.
- Πιο συγκεκριμένα, χρησιμοποιήθηκαν:
  - *Python 3.7.11*
  - *Pandas 1.0.3*
  - *NumPy 1.18.1*
  - *GeoPandas 0.7.0*
  - *MatPlotLib 3.5.1*
  - *GeoPy 2.2.0*
  - *Bokeh 2.0.0*
  - *ST_Visions*

Τα Δεδομένα που χρησιμοποιήθηκαν

bus_trajectories
routeid
vehicleid
datetime
lat
lon

bus_routes
shape_id
shape_pt_lat
shape_pt_lon
shape_pt_sequence
shape_dist_traveled

bus_stops
stop_id
stop_code
stop_name
stop_lat
stop_lon

Η βάση δεδομένων είναι διαθέσιμη εδώ

Φόρτωση των Δεδομένων

- Το πρώτο βήμα είναι η φόρτωση των δεδομένων στο περιβάλλον της *Python*.
- Έπειτα εισάγουμε τους πίνακες και διαγράφουμε τις διπλότυπες εγγραφές.
- Ρίχνουμε μια πρώτη «ματιά» στα δεδομένα.

Φόρτωση των Δεδομένων 2

```
ROUTES table
<class 'pandas.core.frame.DataFrame'>
Int64Index: 54551 entries, 0 to 54550
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   shape_id             54551 non-null  int64
 1   shape_pt_lat         54551 non-null  float64
 2   shape_pt_lon         54551 non-null  float64
 3   shape_pt_sequence    54551 non-null  int64
 4   shape_dist_traveled  52987 non-null  float64
dtypes: float64(3), int64(2)
memory usage: 2.5 MB
```

[4]:

	shape_id	shape_pt_lat	shape_pt_lon	shape_pt_sequence	shape_dist_traveled
0	3185	37.949377	23.637303	182	42985.0
1	3185	37.949221	23.636149	183	43087.0
2	3185	37.948560	23.635824	184	43166.0
3	3185	37.946659	23.635988	185	43378.0
4	3185	37.945976	23.635657	186	43459.0

```
STOPS table
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7989 entries, 0 to 7988
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   stop_id    7989 non-null   int64
 1   stop_code  7989 non-null   object
 2   stop_name  7989 non-null   object
 3   stop_lat   7989 non-null   float64
 4   stop_lon   7989 non-null   float64
dtypes: float64(2), int64(1), object(2)
memory usage: 374.5+ KB
```

[5]:

	stop_id	stop_code	stop_name	stop_lat	stop_lon
0	60605	060605	ΚΑΤΕΧΑΚΗ	37.999062	23.770535
1	60607	060607	ΓΗΡΟΚΟΜΕΙΟ	37.995774	23.767887
2	60608	060608	ΕΡΥΘΡΟΣ ΣΤΑΥΡΟΣ	37.992499	23.766326
3	60610	060610	ΖΕΡΒΑ	37.989746	23.764364
4	60611	060611	ΑΜΠΕΛΟΚΗΠΟΙ	37.987389	23.761900

```
TRAJECTORIES table
<class 'pandas.core.frame.DataFrame'>
Int64Index: 30818606 entries, 0 to 32012259
Data columns (total 5 columns):
 #   Column     Dtype
---  ------     -----
 0   routeid    int64
 1   vehicleid  int64
 2   datetime   object
 3   lat        float64
 4   lon        float64
dtypes: float64(2), int64(2), object(1)
memory usage: 1.4+ GB
```

[6]:

	routeid	vehicleid	datetime	lat	lon
0	3376	65118	Mar 16 2020 08:55:07:000AM	37.839379	23.871029
1	3376	65152	Mar 16 2020 08:55:00:000AM	37.914188	23.900087
2	2024	7029	Mar 16 2020 08:54:53:000AM	38.005803	23.749358
3	2024	7056	Mar 16 2020 08:55:09:000AM	37.977853	23.732996
4	2024	7024	Mar 16 2020 08:45:21:000AM	38.002349	23.741560

Προετοιμασία των Δεδομένων

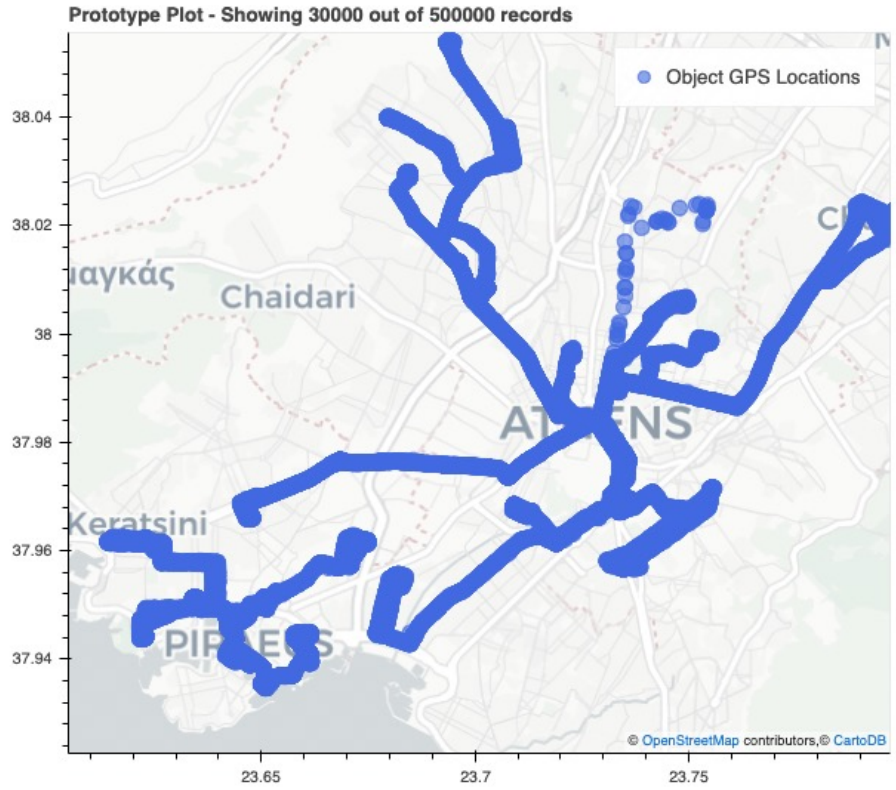
- Σε αυτό το βήμα γίνεται προετοιμασία και μορφοποίηση των δεδομένων για τη καλύτερη μελλοντική τους διαχείριση.
- Αρχικά γίνεται **ταξινόμηση των δυο πρώτων πινάκων** (ROUTES, STOPS).
- Όσον αφορά τον πίνακα TRAJECTORIES, πρέπει να **μετασχηματιστεί η στήλη DATETIME από απλό κείμενο σε *datetime* τύπο**. Επίσης παρατηρείται ότι τα microseconds είναι σε όλες τις εγγραφές μηδέν, επομένως μπορούμε να τα διαγράψουμε.
- Τώρα λοιπόν που η στήλη DATETIME είναι τύπου *datetime*, μπορούμε να **επιλέξουμε το διάστημα των δύο εβδομάδων και να ταξινομήσουμε τον πίνακα βάσει τον αριθμό της διαδρομής, τον αριθμό του λεωφορείου και το χρόνο .**

Προετοιμασία των Δεδομένων 2

- Το επόμενο βήμα είναι η **αναζήτηση των κενών τιμών (null values) στους πίνακες**.
- Είναι αισθητή η ύπαρξη null τιμών, οι οποίες εντοπίζονται στον πίνακα ROUTES, στη στήλη SHAPE_DIST_TRAVELED, στις θέσεις *null_values_index[5] (βλ. oasa_documentation.pdf)*.
- Για να αντιμετωπίσουμε αυτό το πρόβλημα, **γίνεται αντικατάσταση των τιμών αυτών με την απόσταση από το σημείο[κ-1] και το σημείο[κ],** αν το χαρακτηριστικό SHAPE_PT_SEQUENCE είναι μεγαλύτερο του 1, *αλλιώς* η απόσταση γίνεται 0, διότι αν SHAPE_PT_SEQUENCE ισούται με 1, τότε το σημείο[κ] είναι η αρχή της διαδρομής.
- Το επόμενο βήμα είναι να **μετατρέψουμε τις στήλες που εκφράζουν γεωγραφική τοποθεσία από δεκαδικούς αριθμούς σε γεωγραφικού τύπου δεδομένα**. Με αυτό τον τρόπο, στον κάθε πίνακα προστίθεται μία στήλη GEOM όπου αποτελείται από το γεωγραφικό σημείο της κάθε εγγραφής.

Προετοιμασία των Δεδομένων 3

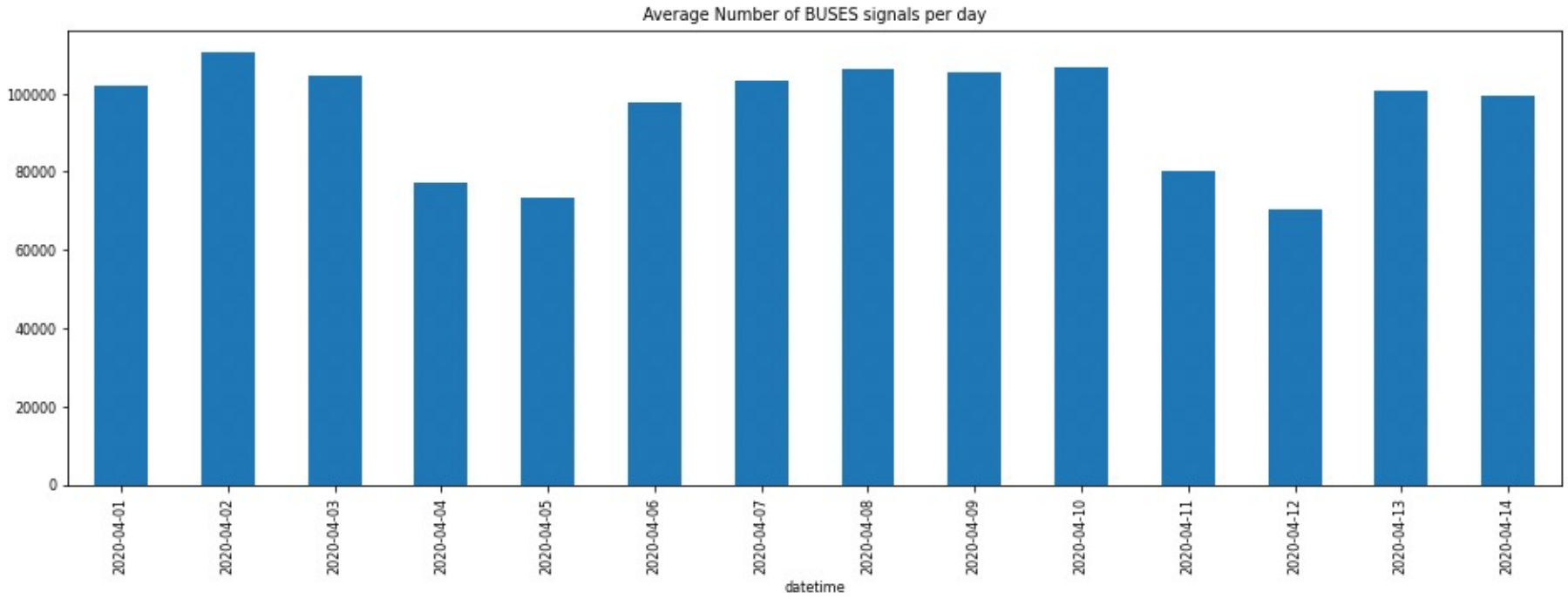
- Τέλος, ρίχνουμε μία γεωγραφική, αυτή τη φορά, «ματιά», στον πίνακα TRAJECTORIES.



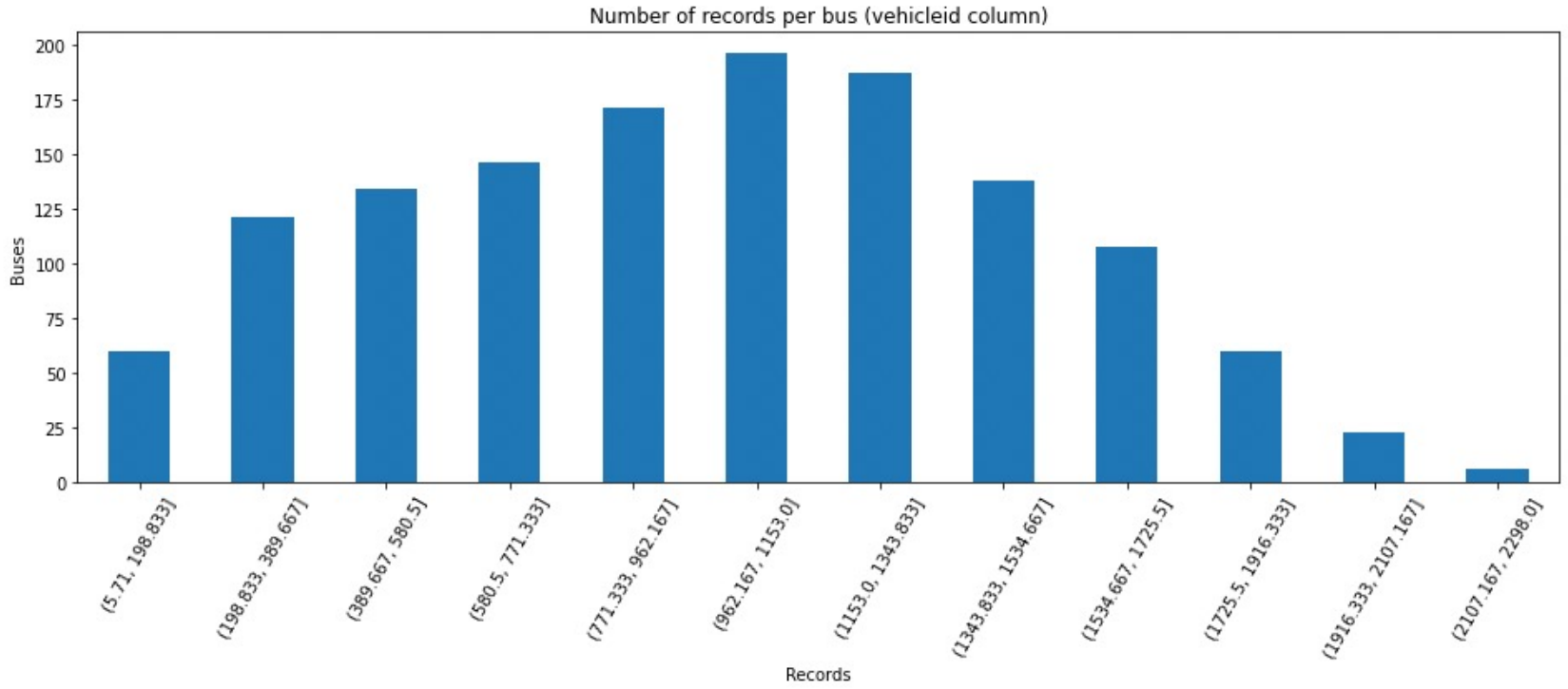
Καθαρισμός των Δεδομένων

- Σε αυτό το βήμα γίνεται καθαρισμός των δεδομένων από τον θόρυβο που περιέχουν τα δεδομένα.
- Αρχικά διαγράφουμε τις διπλότυπες εγγραφές του πίνακα TRAJECTORIES, αυτή τη φορά, βάσει των χαρακτηριστικών DATETIME και VEHICLEID.
- Έπειτα, προσθέτουμε στον πίνακα TRAJECTORIES τα χαρακτηριστικά: VELOCITY, BEARING, ACCELERATION.
- Κατόπιν, παρατηρούμε αν η ταχύτητα ενός λεωφορείου είναι πάνω από 100 χιλ/ώρα. Αν ναι τότε θεωρείται θόρυβος και διαγράφουμε τη συγκεκριμένη εγγραφή.
- Αργότερα, δημιουργούμε ένα χωρικό ευρετήριο R-Tree.

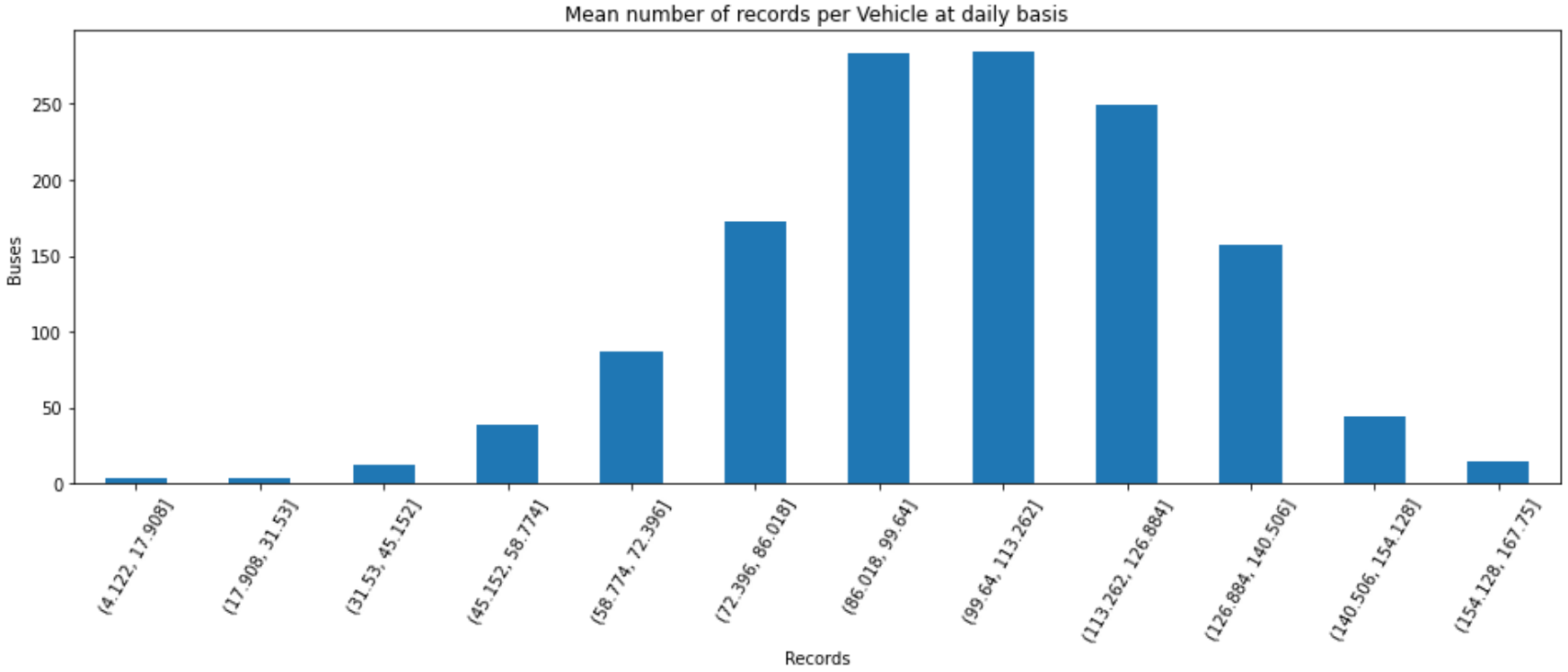
Στατιστικά και Οπτικοποίηση Δεδομένων



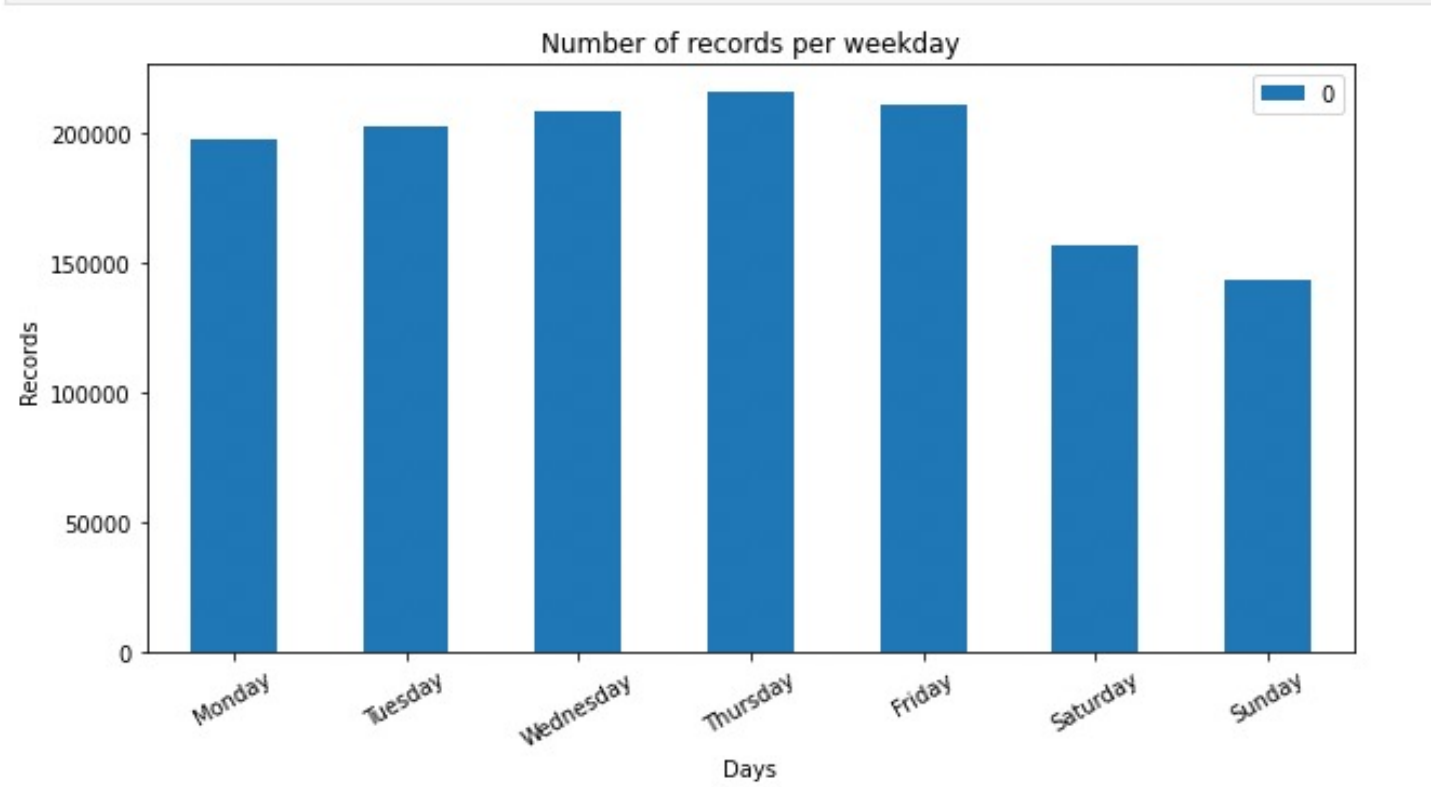
Στατιστικά και Οπτικοποίηση Δεδομένων



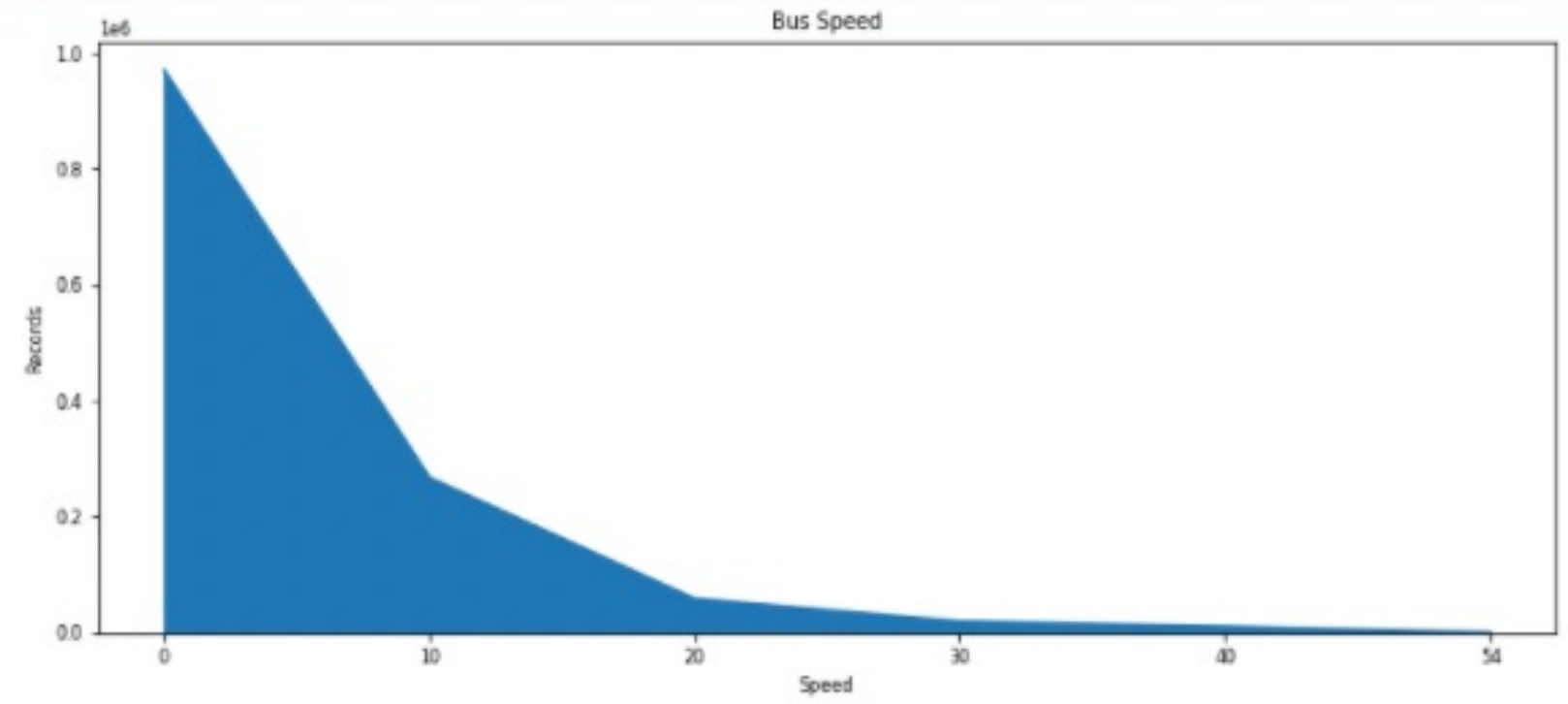
Στατιστικά και Οπτικοποίηση Δεδομένων



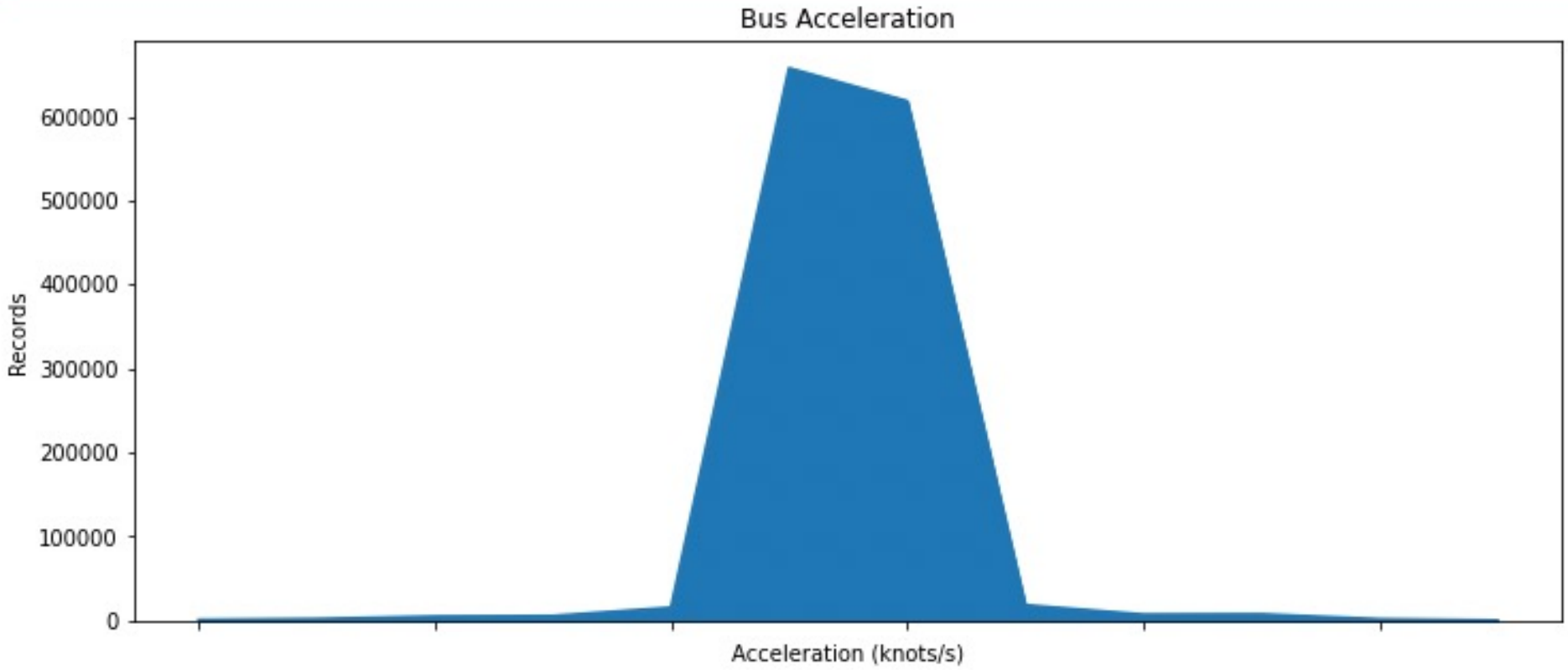
Στατιστικά και Οπτικοποίηση Δεδομένων



Στατιστικά και Οπτικοποίηση Δεδομένων

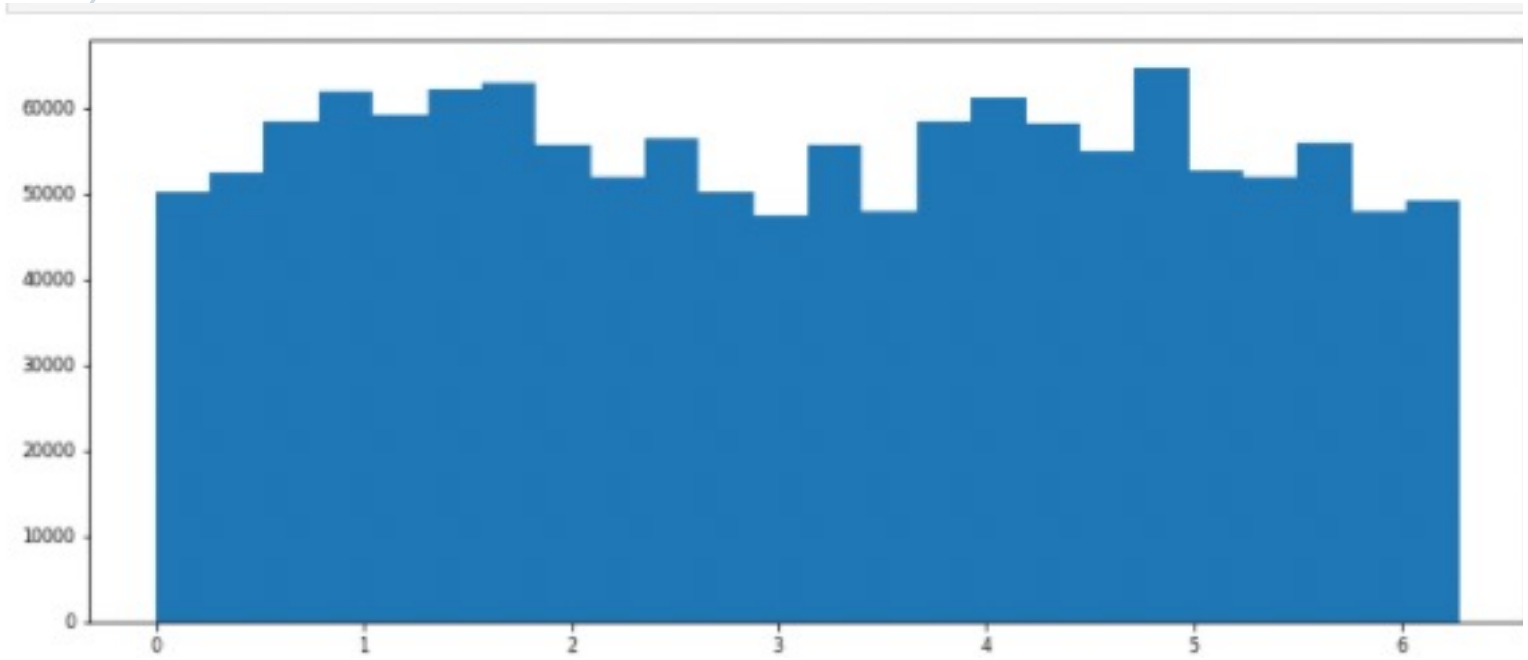


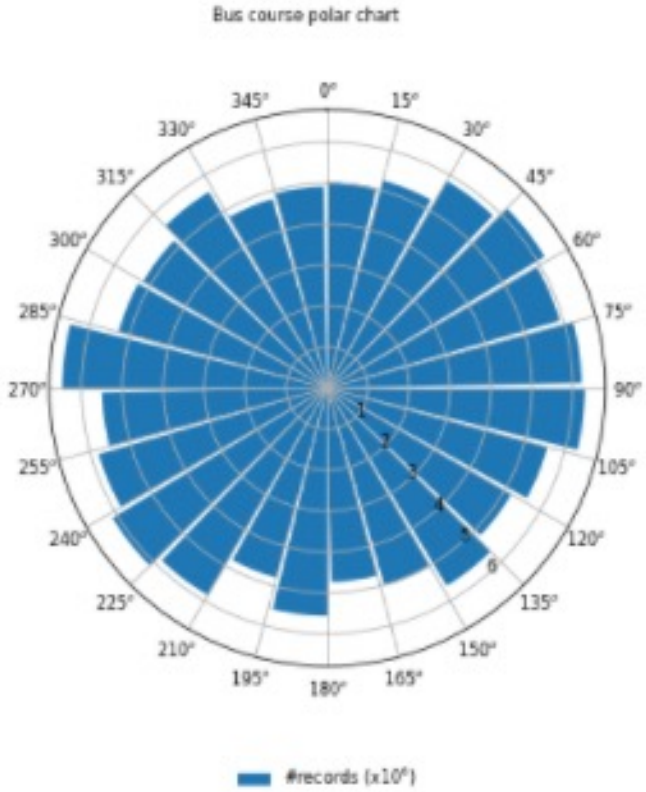
Στατιστικά και Οπτικοποίηση Δεδομένων



Στατιστικά και Οπτικοποίηση Δεδομένων

Bearing Plots





Ευχαριστώ για την προσοχή σας!