Μελέτη Κοινωνικών Φαινομένων και Επίδραση της Πανδημίας του Κορωνοϊού στη Νέα Υόρκη

Παπαμαύρος Δημήτριος
Επιστήμη Δεδομένων και Μηχανική Μάθηση
ΣΗΜΜΥ,Ε.Μ.Π.
Αθήνα, Ελλάδα
dimitrispapamavros@gmail.com

Ορφανουδάκης Φίλιππος Σκόβελεφ
Επιστήμη Δεδομένων και Μηχανική Μάθηση
ΣΗΜΜΥ,Ε.Μ.Π.
Αθήνα, Ελλάδα
phil.orfa@gmail.com

I. Περίληψη

Στο πλαίσιο αυτής της εργασίας αξιοποιούμε δεδομένα που αφορούν το 311 Service Calls στην πόλη της Νέας Υόρκης για να εξάγουμε πληροφορίες για την επίδραση του κορωνοϊού στην εξέλιξη της κοινωνίας και πολιτείας.

Ο τηλεφωνικός αριθμός 311 αποτελεί μια ειδική γραμμή διαθέσιμη σε πολλές περιοχές του Καναδά και των Η.Π.Α.. Μέσω αυτού του αριθμού οι πολίτες έχουν τη δυνατότητα να καταγγέλλουν στις δημοτικές αρχές της πόλης του διάφορες μικρό-παραβάσεις, όπως η δυνατή μουσική σε κάποιο διαμέρισμα, ή να καταθέτουν αιτήματα σε αυτές όπως η κοπή δέντρων. Τα δεδομένα που αφορούν αυτές τις κλήσεις(για τη Νέα Υόρκη) και περιέχουν πληροφορίες όπως το είδος της κλήσης και το είδος της καταγγελίας αλλά η περιοχή είναι δημόσια και βρίσκονται συγκεντρωμένα στη βάση δεδομένων NYC311 Service Requests from 2010 to Present που μας παρέχεται από το NYC Open Data [1]. Όπως αναφέραμε παραπάνω, δεν θα ασχοληθούμε με βαριές παραβάσεις αλλά με καταγγελίες υποδεέστερης σημασίας, οι οποίες κρύβουν και την κοινωνική και πολιτική συμπεριφορά των κατοίκων μιας περιοχής, πχ "Street Condition".

Έχοντας σαν κύριο γνώμονα την αιτία της αλλαγής της καθημερινότητας, δηλαδή τον κορωνοϊό ακολουθήσαμε κάποια βήματα για να εξερευνήσουμε τις βασικές διαφορές στην ποιότητα ζωής των κατοίκων της Νέας Υόρκης. Μελετάμε ξεχωριστά κάθε "περιοχή" της Νέας Υόρκης από το Manhattan έως το Bronx και συγκρίνουμε πως το διαφορετικό επίπεδο ζωής αντιδράει στις ιδιαίτερες συνθήκες μιας πανδημίας.

Το πρώτο βήμα που ακολουθήσαμε είναι να εξάγουμε ένα dataset για την εποχή πριν την πανδημία και ένα για την εποχή της πανδημίας. Στη συνέχεια παρουσιάζουμε κάποιες βασικές συγκρίσεις τόσο για την φύση όσο και την ποσότητα των καταγγελιών με σκοπό να κατανοήσει ο αναγνώστης την γενική εικόνα των dataset μας.

Έπειτα χωρίζουμε την ανάλυση μας σε 3 μέρη : Clustering Analysis, Association Rules Analysis και Assumptions Check - Text Mining. Σε κάθε ένα από αυτά εμβαθύνουμε πλέον σε συγκεκριμένες "περιοχές" της Νέας Υόρκης και τέλος εξάγουμε τα αντίστοιχα συμπεράσματα μας.

II. Εισαγωγή

Τον Δεκέμβριο του 2019, στην πόλη Ουχάν της Κίνας, ξεκίνησε η εξάπλωση του νέου κορωνοϊού SARS-CoV-2. Αρχικά διαγνώστηκαν σαν περιπτώσεις πνευμονίας και ύστερα στις 9 Ιανουαρίου 2020 ανακοινώθηκε ότι πρόκειται για νέο στέλεχος κορωνοϊού. Ως τις 27 Σεπτεμβρίου 2020 έχουν επιβεβαιωθεί πάνω από 33,000,000 κρούσματα, 1,000,000 θάνατοι και έχουν ανακάμψει πάνω από 24,000,000 άνθρωποι.

Συγκεκριμένα στη πόλη της Νέας Υόρκης διαμένουν 8.4 εκατομμύρια άνθρωποι και διαθέτει 5 προάστια τα οποία χαρακτηρίζονται ως τουριστικοί προορισμοί και περιοχές ποικιλομορφίας ως προς τις εθνικότητες. Συνεπώς θεωρούμε την Νέα Υόρκη ως ένα αντιπροσωπευτικό δείγμα της παγκόσμιας κοινωνίας. Η Νέα Υόρκη εμφάνισε το πρώτο θετικό κρούσμα κορωνοϊού στις 01/03/2020 και από τον μήνα Μάρτιο μπήκε σε καθολικό lockdown το οποίο οδηγήθηκε σε σταδιακή άρση μέχρι τον Ιούλιο του 2020.

Αρκετά σημαντικό είναι να εντοπίσουμε τις βασικές αλλαγές στην καθημερινότητα και την συμπεριφορά των ανθρώπων με την επίδραση τόσο ενός lockdown όσο και τον φόβο που προκαλεί μια πανδημία.

Για τον σκοπό αυτό δημιουργούμε 2 datasets. Το πρώτο αποτελείται από όλες τις πληροφορίες σχετικά με τις καταγγελίες που πραγματοποιήθηκαν την περίοδο 01/03/2019-29/02/2020 και το δεύτερο την περίοδο 01/03/2020-01/03/2021.

Τα dataset μας αποτελούνται από 2263571 και 2649627 καταγγελίες αντιστοίχως και οι πιο σημαντικές πληροφορίες για τις καταγγελίες που θα εκμεταλλευτούμε κατά βάση είναι το είδος της καταγγελίας (για παράδειγμα Ηχορύπανση), η ημερομηνία που πραγματοποιήθηκε η καταγγελία, το τμήμα το οποίο διαχειρίστηκε τη καταγγελία όπως και η διεύθυνση του συμβάντος.

III. Συγγενείς Εργασίες

Έχουν εντοπιστεί ορισμένες εργασίες που χρησιμοποίησαν το συγκεκριμένο dataset, παρόλα αυτά οι περισσότερες από αυτές στοχεύουν είτε στο visualization των δεδομένων είτε στην δημιουργία κάποιου μοντέλου πρόβλεψης καταγγελιών. Καμία από τις εργασίες που έχουμε βρει δεν αναφέρεται στη σύγκριση των 2 εποχών.

Το πρώτο project[2] που μας κίνησε το ενδιαφέρον έχει σαν βασικό σκοπό την δημιουργία μιας σελίδας που να κάνει visualize τις πληροφορίες που εξάγει. Οι πληροφορίες που εξάγει έχουν να κάνουν με στατιστικά στοιχεία σχετικά με τις πιο δημοφιλείς καταγγελίες καθώς και τις ώρες που αυτές πραγματοποιούνται. Στο τέλος του project, χρησιμοποιείται ο αλγόριθμος k-NN ο οποίος βασιζόμενος στη περιοχή προσπαθεί να προβλέψει το είδος του θορύβου (Μουσική, Αυτοκίνητα κ.α.) σε μελλοντικές καταγγελίες.

Το δεύτερο project[3] στο οποίο θα αναφερθούμε έχει σαν κύριο στόχο την χρήση και τη σύγκριση αλγορίθμων Supervised Learning για την κατασκευή μοντέλων πρόβλεψης μελλοντικών παραπόνων. Συγκεκριμένα έχουμε την χρήση k-NN, SVM, Logistic Regression και Decision Trees και την σύγκριση των μεταξύ τους αποτελεσμάτων.

Το τρίτο paper[10] αφορά την αξιοποιήσει των δεδομένων από τις κλήσεις στον αριθμό 311 με σκοπό την εύρεση ενός μοντέλο το οποίο θα μπορεί να χρησιμοποιηθεί για την εύρεση αστικών περιοχών οι οποίες αναμένεται να υποβαθμιστούν ή όχι. Για να γίνει αυτό οι συγγραφείς προτείνουν τη χρήση της λογιστικής παλινδρόμησης.

Το τέταρτο και τελευταίο paper[18] που θα σχολιάσουμε αποτελεί ίσως το κίνητρο για τη θεματολογία της εργασίας. Στο άρθρο αυτό οι συγγραφείς αναλύουν τα δεδομένα του 311 Service Requests για την Νέα Υόρκη, την Βοστώνη και το Σικάγο και μέσω της χρήσης κατάλληλων μεθόδων ανάλυσης και classification καταλήγουν στο συμπέρασμα ότι τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν στη δημιουργία ενός κοινωνικού και οικονομικού προφίλ για κάθε περιοχή. Ένα τέτοιο προφιλ είναι ικανό να περιγράψει την περιοχή ποιοτικά όπως επίσης να φανεί χρήσιμο στη λήψη αποφάσεων σχετικά με την αγοροπωλησία ακινήτων καθώς και την κατασκευή regression μοντέλων που να προβλέπουν οικονομικούς, εκπαιδευτικούς, δημογραφικούς δείκτες.

Οι αναλύσεις που αποφασίσαμε να κάνουμε έχουν εμπνευστεί και από άλλες σχετικές εργασίες που όλες τους όμως δεν ξεφεύγουν από την θεματολογία των 3 πρώτων που παρουσιάστηκαν παραπάνω.

IV. Μεθοδολογίες

A. Βασική Σύγκριση των **2** Εποχών

Στην ενότητα αυτή στόχο έχουμε να εκμεταλλευτούμε την μεγάλη διαφοροποίηση στο επίπεδο ζωής, που παρατηρείται στα προάστια της Νέας Υόρκης. Πιο συγκεκριμένα θα κατασκευάσουμε ένα προφίλ για κάθε ένα από αυτά, ορίζοντας συγκεκριμένα χαρακτηριστικά και έπειτα θα μελετήσουμε συγκεκριμένες αλλαγές στο πλήθος και το φύση των καταγγελιών στην εποχή πριν την πανδημία αλλά και κατά τη διάρκεια της.

Αρχικά η Νέα Υόρκη χωρίζεται στα εξής 5 "διαμερίσματα" : Brooklyn, Manhattan, Bronx, Staten Island, Queens. Ο πληθυσμός κάθε "διαμερίσματος" για την χρονιά 2020[4] καθώς παρατηρούμε ελάχιστες διαφορές για τις χρονιές πριν και κατά την διάρκεια της πανδημίας είναι:

- Brooklyn : 2,648,452
- Manhattan : 1,638,281
- Bronx : 1,446,788
- Staten Island : 487,155
- Queens : 2,330,295

Πυκνότητα πληθυσμού[5] (people per square mile) :

- Brooklyn : 36,732
- Manhattan : 70,826
- Bronx : 35,000
- Staten Island : 8,112
- Queens : 21,000

Κατά Κεφαλήν Εισόδημα[6] (in U.S. Dollars):

- Brooklyn : 34,173
- Manhattan : 76,592
- Bronx : 21,778
- Staten Island : 36,907
- Queens : 31,930

Ποσοστό πληθυσμού που κατέχουν τουλάχιστον πτυχίο Bachelor's[7]:

- Brooklyn : 37%
- Manhattan : 61%
- Bronx : 20%
- Staten Island : 33%
- Queens : 31.5%

Ποσοστό πληθυσμού που είναι κάτω από το όριο της φτώχειας[6] (poverty):

- Brooklyn : 37%
- Manhattan : 14.1%
- Bronx : 26.2%
- Staten Island : 33%
- Queens : 11%

Ποσοστό άνεργου πληθυσμού[8]:

- Brooklyn : 4.5%
- Manhattan : 4.2%
- Bronx : 9.1%
- Staten Island : 4%
- Queens : 4.6%

Ποσοστό ανήλικου πληθυσμού[6]:

- Brooklyn : 22.7%
- Manhattan : 14.3%
- Bronx : 24.6%
- Staten Island : 21.8%

- Queens : 20%

Ποσοστό African-American πληθυσμού[6]:

- Brooklyn : 33.8%
- Manhattan : 17.8%
- Bronx : 43.6%
- Staten Island : 11.6%
- Queens : 20.7%

Τα συμπεράσματα που μπορούν να προκύψουν για την ποιότητα και την φύση της καθημερινότητας του κάθε "διαμερίσματος" δεν αποτελεί γνωστικό μας πεδίο, παρόλα αυτά μπορούμε να θεωρήσουμε το Manhattan ως μια πυκνοκατοικημένη περιοχή που προσφέρει περισσότερες ευκαιρίες για καλύτερη ποιότητα ζωής. Στη συνέχεια μπορούμε να κατατάξουμε το Queens ως ένα πιο αραιοκατοικημένο "διαμέρισμα", που έχει χαμηλό δείκτη φτώχειας και αρκετά υψηλό κατά κεφαλήν εισόδημα. Στη συνέχεια το Staten Island είναι η πιο απομακρυσμένη περιοχή και χαρακτηριστικό της είναι ότι διαθέτει μόνο έναν αυτοκινητόδρομο, συνεπώς καταλαβαίνουμε ότι είναι διαφορετική η φύση της καθημερινότητας των ανθρώπων συγκριτικά με τα υπόλοιπα "διαμερίσματα". Έπειτα το Brooklyn θυμίζει αρκετά στους δείκτες το Queens αλλά ο υψηλός δείκτης φτώχειας υποδεικνύει ένα κατώτερο ποιοτικό επίπεδο ζωής. Τέλος το Bronx ίσως προσφέρει τις λιγότερες ευκαιρίες για ένα καλό επίπεδο ζωής καθώς έχει το χαμηλότερο κατά κεφαλήν εισόδημα ενώ ταυτόχρονα το υψηλότερο ποσοστό ανεργίας. Για καλύτερη εποπτεία των επόμενων συγκρίσεων μας και συμπερασμάτων αποδίδουμε την εξής αξιολόγηση:

- Manhattan ★★★★★
- Queens ★★★★
- Staten Island ★★★
- Brooklyn ★★★
- Bronx ★

Έχοντας κατασκευάσει ένα προφίλ για κάθε "διαμέρισμα", μπορούμε να προχωρήσουμε στην ανάλυση ορισμένων καταγγελιών.

Αρχικά παρουσιάζουμε έναν πίνακα με το πλήθος των καταγγελιών ανά "διαμέρισμα" :

Before Covid-19		
Borough Name	Sum of Complaints	Pie%
Manhattan	461228	20.5%
Quenns	565851	25.2%
Staten Island	115154	5.1%
Brooklyn	706997	31.5%
Bronx	396264	17.6%

During Covid-19		
Borough Name	Sum of Complaints	Pie%
Manhattan	534386	20.3%
Quenns	608525	23.2%
Staten Island	112715	4.3%
Brooklyn	750876	28.6%
Bronx	620551	23.6%

Before Covid-19	
Borough Name	Calls Per Capita
Manhattan	0.28
Quenns	0.24
Staten Island	0.24
Brooklyn	0.27
Bronx	0.27

During Covid-19	
Borough Name	Calls Per Capita
Manhattan	0.33
Quenns	0.26
Staten Island	0.23
Brooklyn	0.28
Bronx	0.43

Παρατηρούμε πολύ μεγάλη αύξηση καταγγελιών στο Bronx, το "διαμέρισμα" που έχουμε κατατάξει ως τελευταίο με βάση την ποιότητα ζωής. Επίσης μια σημαντική αύξηση παρατηρείται στο Manhattan που είναι η πιο πυκνοκατοικημένη περιοχή. Συνεπώς το συμπέρασμα μας με βάση τις καταγγελίες είναι ότι είναι αντιστρόφως ανάλογο της ποιότητας ζωής και ανάλογο της πυκνότητας του πληθυσμού, από εδώ και στο εξής θα έχουμε τον παρακάτω συμβολισμό :

- Ποιότητα Ζωής ↗
- Πυκνότητα Πλυθυσμού ↘

Στη συνέχεια θα μελετήσουμε συγκεκριμένα είδη καταγγελιών που θεωρούμε ότι μπορούν να εκφράσουν κάποιο κοινωνικό φαινόμενο. Για τον σκοπό αυτό δοκιμάσαμε να εφαρμόσουμε Latent Dirichlet Allocation (LDA)[11] και να εξάγουμε κάποια βασικά θέματα (topics), όμως ο όγκος των δεδομένων μας και η έλλειψη υπολογιστικών πόρων δεν μας το επέτρεψε. Συνεπώς μελετήσαμε τις διάφορες κατηγορίες καταγγελιών και καταλήξαμε σε 4 θέματα με την εξής ομαδοποίηση :

- Ποιότητα Ατμόσφαιρας : 'Air Quality'
- Οδική Κυκλοφορία : 'Blocked Driveway', 'Noise - Vehicle','Traffic'
- Κατάσταση Αστέγων : 'Homeless Street Condition', 'Homeless Person Assistance', 'Homeless Encampment'
- Συμπεριφορά Κατοίκων σε Εξωτερικούς Χώρους : 'Urinating in Public','Noise - Street/Sidewalk','Sidewalk Condition','Root/Sewer/Sidewalk Condition', 'Illegal Tree Damage','Noise - Park'

a) Ποιότητα Ατμόσφαιρας και Οδική Κυκλοφορία*:* Ομαδοποιούμε τις 2 αυτές κατηγορίες στην ίδια παράγραφο γιατί χρησιμοποιούμε και τις 2 για την εξαγωγή των συμπερασμάτων μας. Αρχικά θα παρουσιάσουμε τον αντίστοιχο χάρτη διασποράς, ώστε να παρατηρήσουμε την αλλαγή αναλυτικά σε όλη την έκταση της Νέας Υόρκης για τα συμπεράσματα μας θα αναλύσουμε και την κατηγορία "Οδική Κυκλοφορία" καθώς μπορεί να έχει άμεση σχέση με την ποιότητα του αέρα.

Συνεχίζουμε με την Οδική Κυκλοφορία και ομοίως θα παρουσιάσουμε τον χάρτη διασποράς για την συγκεκριμένη κατηγορία.

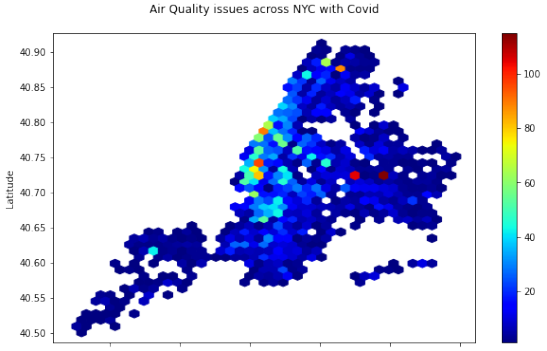

Figure 1. Air Quality Complaints During Covid-19

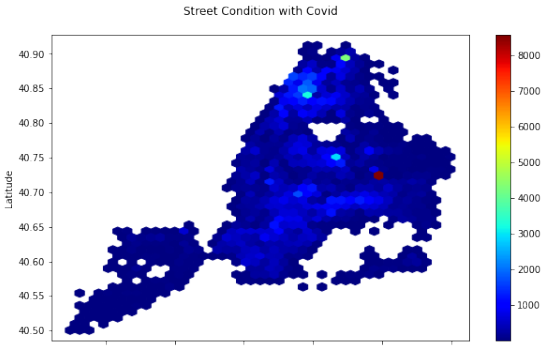

Figure 3. Traffic Complaints During Covid-19

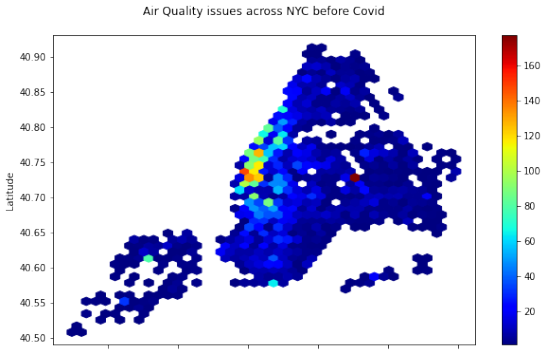

Figure 2. Air Quality Complaints Before Covid-19

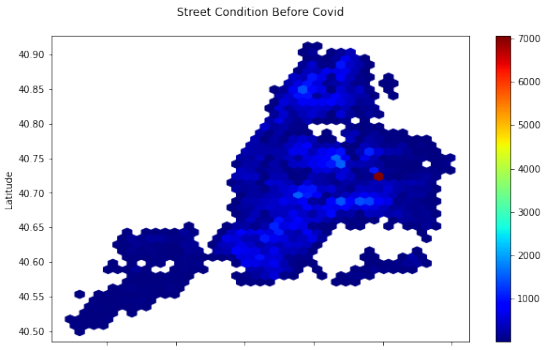

Figure 4. Traffic Complaints Before Covid-19

Η βασική πληροφορία που εξάγουμε από την κλίμακα είναι ότι υπάρχει μείωση των καταγγελιών στην εποχής της πανδημίας, που μπορεί έμμεσα να εξηγεί μια βελτίωση στη κοινωνική συμπεριφορά των κατοίκων. Το Manhattan φαίνεται να διατηρεί την πρώτη θέση και στις 2 εποχές, γεγονός που οφείλεται κυρίως στην πυκνότητα του πληθυσμού.

Air Quality			
Borough Name	Before Covid-19	During Covid-19	Difference
Manhattan	2747	1873	-31,8%
Quenns	1385	1589	+14,7%
Staten Island	311	265	-14,7%
Brooklyn	2285	1888	-17,3%
Bronx	471	721	+53,1%

Από τον παραπάνω πίνακα έχουμε σημαντική μείωση στο Manhattan, στο Brooklyn και στο Staten Island. Από την άλλη υπάρχει αύξηση τόσο στο Queens όσο και στο Bronx , δύο περιοχές που δεν μοιράζονται κάποια χαρακτηριστικά από αυτά που έχουμε αναφέρει παραπάνω. Πριν αποφασίσουμε

Οι αλλαγές που βλέπουμε αντιστοιχούν ακριβώς στις αλλαγές που είδαμε στις "Air Quality" καταγγελίες , δηλαδή αύξηση στο Bronx και στο Queens. Μελετώντας αυτά τα αποτελέσματα, ανακαλύψαμε ότι τα 2 αεροδρόμια της Νέας Υόρκης, το LaGuardia και το JFK βρίσκονται στο Queens, γεγονός που επηρεάζει κατά πολύ την οδική κυκλοφορία. Αυτό φαίνεται και στον παρακάτω πίνακα :

Traffic Complaints			
Borough Name	Before Covid-19	During Covid-19	Difference
Manhattan	14988	22088	+47.37%
Quenns	70757	67356	-4.81%
Staten Island	4868	4567	-6.2%
Brooklyn	62703	60813	-3.01%
Bronx	32848	49750	+51.4%

Συγκεντρωτικά μπορούμε να οδηγηθούμε στις εξής παρατηρήσεις. Πράγματι το Queens λόγω των αεροδρομίων έχει την χειρότερη Οδική Κυκλοφορία. Παρατηρούμε μεγάλη αύξηση στην κυκλοφορία στο Manhattan και στο Bronx παρόλο που είχαμε μεγάλη μείωση των καταγγελιών για Air

Quality στο πρώτο μεγάλη αύξηση στο δεύτερο. Επομένως βγάζοντας πλέον από την εξίσωση το Queens λόγω των ιδιαιτεροτήτων έχουμε ότι για το Air Quality, αναμένουμε :

- Ποιότητα Ζωής ↗
- Πυκνότητα Πλυθυσμού ↘

και για την Οδική Κυκλοφορία δεν καταλήγουμε σε κάποιο συμπέρασμα καθώς ίσως οφείλονται και εξωγενείς παράγοντες που να οδήγησαν σε καταγγελίες, όπως η αλλαγή στην ψυχολογία.

b) Κατάσταση Αστέγων*:* Στην παράγραφο αυτή μπορεί να παρερμηνευθεί ο τρόπος έκφρασης, για αυτό διευκρινίζουμε ότι αύξηση σημαίνει χειροτέρευση της κατάστασης. Ακολουθούν οι αντίστοιχοι χάρτες για να παρατηρήσουμε την κλίμακα και πιθανόν κάποια αλλαγή στη διασπορά.

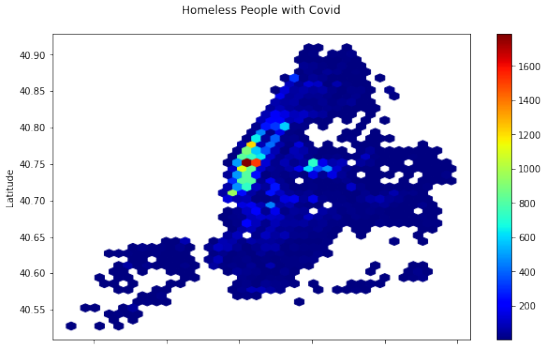

Figure 5. Homeless People During Covid-19

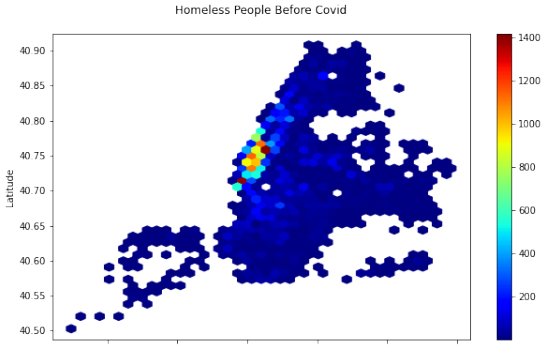

Figure 6. Homeless People Before Covid-19

Παρατηρούμε στο Manhattan τις περισσότερες καταγγελίες για την κατάσταση των αστέγων, παρόλο που στο Bronx το ποσοστό του άνεργου πληθυσμού είναι διπλάσιο. Αυτό μπορεί να οφείλεται στην κοινωνική "ψαλίδα", καθώς σε μια κοινωνία με υψηλό επίπεδο ζωής είναι εύκολα παρατηρήσιμο κάποιο σημάδι χαμηλού επιπέδου. Ακολουθεί ο πίνακας με τις διαφορές.

Homeless People Complaints			
Borough Name	Before Covid-19	During Covid-19	Difference
Manhattan	19598	24145	+23.2%
Quenns	2703	6171	+128.3%
Staten Island	2544	502	+97.63%
Brooklyn	5966	7659	+28.38%
Bronx	1565	2167	+38.46%

Σε αυτή την κατηγορία παρατηρείται το φαινόμενο ότι ανεξαρτήτως κοινωνικής κατάστασης της περιοχής η κατάσταση των άστεγων αυξάνεται. Συνεπώς έχουμε :

- Covid-19 ↗

c) Συμπεριφορά Κατοίκων σε Εξωτερικούς Χώρους*:* Προσθέσαμε αυτή τη κατηγορία καθώς δίνει μια εικόνα για την ψυχολογία των ανθρώπων. Η πίεση μιας πανδημίας μπορεί να οδηγήσει σε πιο ακραίες συμπεριφορές, γεγονός που μπορεί να εξηγήσει και άλλα φαινόμενα , όπως την γενικότερη αύξηση των καταγγελιών λόγω του ότι οι άνθρωποι καταλήγουν πιο ευερέθιστοι. Παρουσιάζουμε τον αντίστοιχο χάρτη διασποράς.

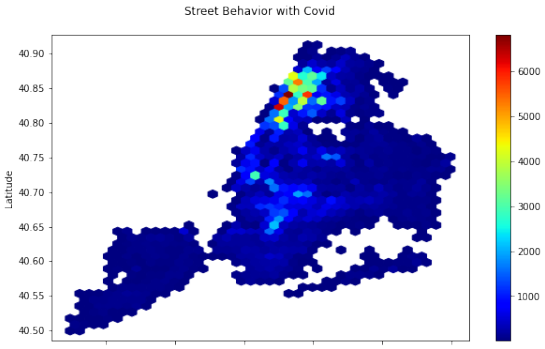

Figure 7. Street Behavior During Covid-19

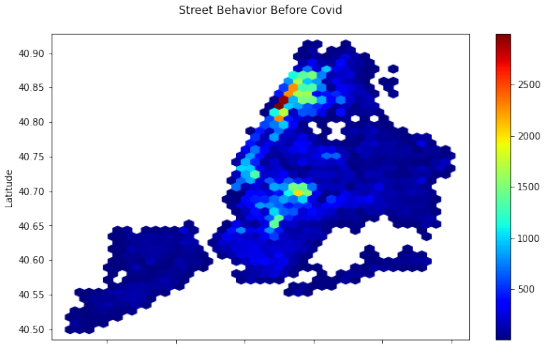

Figure 8. Street Behavior Before Covid-19

Μεγάλο πλήθος καταγγελιών εμφανίζεται στο Bronx, το οποίο βρίσκεται τελευταίο στη κλίμακα που σχηματίσαμε για την ποιότητα ζωής. Επίσης η κλίμακα μας προϊδεάζει ότι θα έχουμε μεγάλες αυξήσεις σε κάθε "διαμέρισμα".

Homeless People Complaints			
Borough Name	Before Covid-19	During Covid-19	Difference
Manhattan	43676	81004	+85.47%
Quenns	27704	36891	+33.16%
Staten Island	6738	6984	+3.7%
Brooklyn	50798	60396	+18.9%
Bronx	32190	68822	+113.8%

Πολύ σημαντικές αυξήσεις παρατηρούνται που υποδηλώνουν την πίεση των κατοίκων και την ευερέθιστη συμπεριφορά. Ίσως πιο σημαντική έρευνα θα αποτελούσε η συμπεριφορά των κατοίκων μέσα στους οικιακούς τους χώρους. Τέλος το συμπέρασμα μας για την συμπεριφορά είναι το εξής :

- Covid-19 ↗

B. **Clustering**

Σκοπός αυτής της ενότητας είναι η παρουσίαση μιας προσπάθειας ομαδοποίησης των γειτονιών της Νέας Υόρκης σύμφωνα με τα πιο συχνά αναφερόμενα προβλήματα στον αριθμό 311 καθώς επίσης και το πως αυτές επηρεάστηκαν λόγω της πανδημίας. Για να πραγματοποιήσουμε clustering στα δεδομένα μας θα πρέπει πρώτα να επέμβουμε σε αυτά με σκοπό να δημιουργήσουμε κατάλληλα χαρακτηριστικά πάνω στα οποία θα βασιστεί η συσταδοποίηση. Αρχικά διορθώνουμε τους καταγεγραμμένους τύπους παραπόνων (π.χ. θέρμανση, παράνομη στάθμευση) καθώς όντας διάνυσμα χαρακτήρων τυπογραφικά λάθη αλλά και σημεία στίξης μπορούν να επηρέασουν την ανάλυση μας. Στη συνέχεια αρκεί να υπολογίσουμε το πλήθος των καταγγελιών ανά τύπο παραπόνου και zip-code, λόγω του μεγάλου πλήθους των δεδομένων και της δυσανάλογης υπολογιστικής δύναμης μας επιλέγουμε μόνο τις καταγγελίες που εμφανίζονται τουλάχιστον 10 φορές. Μέσω αυτής της προεπεξεργασίας καταλήγουμε σε καινούργια σετ δεδομένων, ένα για τη περίοδο πριν την πανδημία και ένα για τη περίοδο αυτής. Κάθε ένα από αυτά διαθέτουν τόσες γραμμές όσες και ο αριθμός των μοναδικών zip-codes στο αρχικό σετ δεδομένων και αντίστοιχα τόσες στήλες όσοι οι μοναδικοί τύποι παραπόνων. Για να συσταδοποίησουμε τα παραπάνω δεδομένα χρησιμοποιήθηκαν dbscan[12], Gaussian Mixtures Models[17], k-Means[16] και Hierarchical Clustering[19] (Divisive και Agglomerative) παρόλο αυτά παρατηρήθηκε ότι οι δύο πρώτες μέθοδοι αδυνατούν, λόγω της φύσης των δεδομένων, να παράξουν επαρκείς ομαδοποιήσεις. Τόσο για τα δεδομένα πριν την πανδημία όσο και κατά τη διάρκεια αυτής καταλήγουμε σε καλά αποτελέσματα χρησιμοποιώντας συσσωρευτική ιεραρχική ομαδοποίηση με τη μέθοδο Ward και κλαδεύοντας το δέντρο έτσι ώστε να καταλήξουμε σε τέσσερις συστάδες καθώς επίσης και με τον αλγόριθμο k-Means χρησιμοποιώντας τέσσερα κέντρα. Ανάμεσα σε αυτές τις δύο μεθόδους επιλέχθηκε ο αλγόριθμος k-Means καθώς παρουσίαζε το μεγαλύτερο average silhouette (0.58 για τα δεδομένα πριν την πανδημία και 0.52 για τα δεδομένα κατά τη διάρκεια της πανδημίας). Στα Σχήματα 9-10 παρουσίαζουμε τα αποτελέσματα αυτής της ομαδοποίησης.

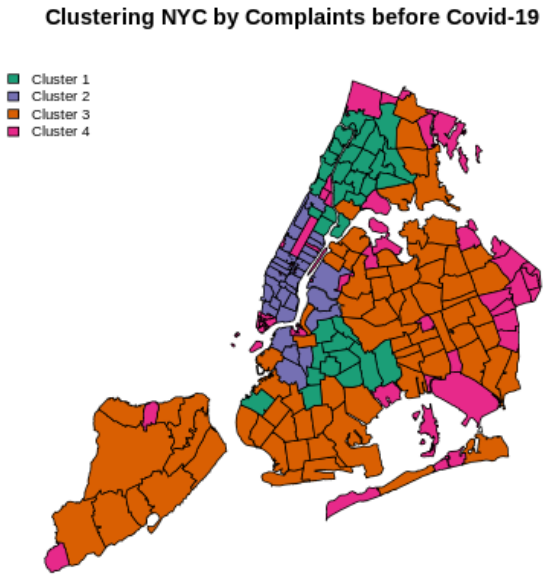

Figure 9. Clustering Zip-Codes by Complaint Type Before Covid

Before Covid-19			
Φτωχές Γειτονιές (Cluster-1)	for hire vehicle complaint	derelict bicycle	homeless person assistance
Εύπορες Γειτονιές (Cluster-2)	flooring/stair	door/window	water leak
Τυπικοί Νεοϋρκέζοι (Cluster-3)	panhandling	ferry complaint	taxi complaint
Προάστια (Cluster-4)	damaged tree	sewer	abandoned vehicle

Table I
Μεταβλητές με μεγάλες τιμές κέντρων ανά CLUSTER πριν τη πανδημία

Παρατηρούμε εύκολα ότι πριν και κατά την διάρκεια της πανδημίας οι ομαδοποιήσεις έχουν μεταβληθεί ελάχιστα κάτι το οποίο μας οδηγεί σε ένα εύλογο συμπέρασμα που δεν είναι άλλο από το ότι οι κοινότητες, όπως αυτές διαμορφώνονται μετά την ομαδοποίηση, παρουσιάζουν παρόμοια προβλήματα τόσο πριν όσο και κατά τη διάρκεια αυτής ανεξάρτητα εαν αυτά τα προβλήματα έχουν αλλάξει. Κοιτώντας τα κέντρα του κάθε cluster μπορούμε να πάρουμε μια εικόνα σχετικά με το ποια είναι τα πιο σημαντικά προβλήματα σε κάθε cluster. Αυτό είναι εφικτό καθώς τα δεδομένα μας αφορούν το πλήθος από κάθε είδος παραπόνων ανά zip-code. Συνεπώς οι μεταβλητές με τις μεγαλύτερες τιμές κέντρων ανά cluster αφορούν και τα πιο συχνά-σημαντικά προβλήματα για κάθε cluster αντίστοιχα. Φυσική απόρροια αυτού είναι η εύρεση μιας φυσικής ερμηνείας των συστάδων λαμβάνοντας πάντα υπόψη και τη χωρική απεικόνιση αυτών στον χάρτη. Όπως παρουσιάζεται και στους Πίνακες I-II χωρίζεται σε τέσσερις συστάδες, τις Εύπορες Γειτονίες, τις Φτωχές Γειτονιές, τα Προάστια και τέλος τους Τυπικούς Νεοϋορκέζους, οι οποίοι καταλαμβάνουν και το μεγαλύτερο μέρος.

Παρατηρούμε ότι οι Εύπορες Περιοχές τις Νέας Υόρκης απασχολούνται τόσο πριν όσο και κατά τη διάρκεια της

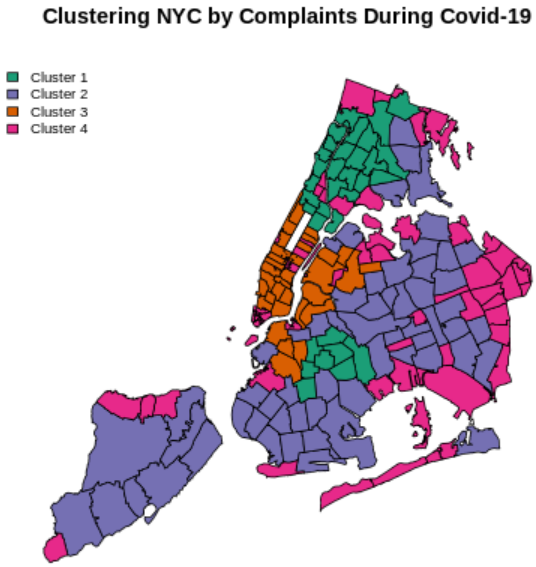

Figure 10. Clustering Zip-Codes by Complaint Type During Covid

During Covid-19			
Φτωχές Γειτονιές (Cluster-1)	outdoor dining	noise	covid-19 non-essential construction
Τυπικοί Νεοϋρκέζοι (Cluster-2)	lifeguard	mosquitoe	x-ray machine or equipment
Εύπορες Γειτονιές (Cluster-3)	water leak	door/window	appliancet
Προάστια (Cluster-4)	damaged tree	overgrown tree/branche	building/use

Table II
Μεταβλητές με μεγάλες τιμές κέντρων ανά CLUSTER στη πανδημία

πανδημίας με θέματα τα οποία αφορούν τη βελτίωση κτιριακών εγκαταστάσεων μιας καθώς τα χρήματα δεν αποτελούν πρόβλημα για αυτούς. Παρόμοια είναι και οι κατάσταση στα Προάστια με τους κάτοικους να μην αλλάζουν ιδιαίτερα τις συνήθειες τους και να χρησιμοποιούν το 311 για να παραπονεθούν για προβλήματα τα οποία αφορούν προβλήματα που αφορούν περιοχές εκτός αστικών κέντρων όπως πεσμένα δέντρα και βουλωμένοι υπόνομοι. Αντίθετα στις Φτωχές Γειτονίες η εικόνα είναι εντελώς διαφορετική. Πριν την πανδημία από τις καταγγελίες των κατοίκων είναι εμφανής μια εικόνα εγκατάλειψης σε αυτές τις περιοχές καθώς οι κάτοικοι τηλεφωνούν στο 311 κυρίως για να ζητήσουν βοήθεια για κάποιο άστεγο συμπολίτη τους είτε να παραπονεθούν για υπηρεσίες εκμίσθωσης αυτοκινήτων και παρατημένα ποδήλατα. Παρόλο αυτά με την έλευση της πανδημίας και την εφαρμογή μέτρων περιορισμού η εικόνα αλλάζει δραστικά. Οι κάτοικοι των συγκεκριμένων περιοχών φαίνεται να είναι οι μόνοι οι οποίοι επηρεάστηκαν έντονα από την πανδημία καθώς η εμφάνιση αυτής αντικατοπτρίζεται και στο είδος των παραπόνων στο 311 σε αντίθεση με άλλες περιοχές. Πιο συγκεκριμένα λόγω της εφαρμογής των μέτρων περιορισμού οι κάτοικοι καταγγέλλουν συμπολίτες οι οποίοι δεν τηρούν τα μέτρα περιορισμού είτε τρώγοντας σε εξωτερικούς χώρους είτε πραγματοποιώντας μη-απαραίτητες οικοδομικές διεργασίες και φυσικά παραπονιούνται για προβλήματα θορύβου τα οποία όμως δεν διευκρινίζεται τι αφορούν. Τέλος οι Τυπικοί Νεοϋρκέζοι όντας στη πλειονότητα τους commuters παραπονιούνται πριν τη πανδημία κυρίως για θέματα τα οποία αφορούν τη καθημερινή προσπάθεια τους να προσεγγίσουν τους χώρους εργασίας τους. Παρόλο κατά τη διάρκεια της πανδημίας μας απέδειξαν ότι μπορούν να παραπονεθούν για μεγάλη θεματολογία ζητημάτων

C. Association Rules

Στην ενότητα αυτή θέλουμε να εντοπίσουμε ποια ζευγάρια καταγγελιών εμφανίζονται συχνά, δηλαδή ποιες καταγγελίες σχετίζονται μεταξύ τους στα διαφορετικά "διαμερίσματα" και να εξάγουμε κάποια συμπεράσματα με την αλλαγή της κοινωνικής συμπεριφοράς των κατοίκων. Η διαδικασία που ακολουθεί πραγματοποιείται για κάθε "διαμέρισμα" ξεχωριστά όπως επίσης και για κάθε περίοδο ξεχωριστά. Ο στόχος είναι να παρατηρήσουμε αν αλλάζει η φύση των συσχετιζόμενων καταγγελιών και αν αυτό μπορεί να αποδοθεί στην πανδημία και στα χαρακτηριστικά κάθε "διαμερίσματος".

Η μεθοδολογία που ακολουθούμε είναι η χρήση του αλγορίθμου a priori[9]. Αρχικά θα πρέπει να πραγματοποιήσουμε κάποια προεπεξεργασία στα δεδομένα μας.

Πρώτα κατασκευάζουμε 10 μεταβλητές (1 για κάθε προάστιο ανα εποχή) και ομαδοποιούμε σύμφωνα με τον zip-code, συνεπώς τα αντίστοιχα basket μας είναι οι καταγγελίες ανά zip-code. Στην συνέχεια πραγματοποιούμε one hot encoding στα δεδομένα μας έτσι ώστε να μην υπολογίσουμε το πλήθος των καταγγελιών ανά basket. Έπειτα τα δεδομένα μας είναι έτοιμα για να τοποθετηθούν σαν είσοδος στον αλγόριθμο. Στο σημείο αυτό να τονίσουμε πως ο αλγόριθμος a priori, όπως και ο fp-growth[14] χρίζουν αρκετά μεγάλη υπολογιστική πολυπλοκότητα, συνεπώς έχοντας ένα αρκετά μεγάλο dataset έπρεπε να πραγματοποιήσουμε κάποιες περικοπές. Η πρώτη ήταν να επιλέξουμε σαν περίοδο μελέτης, τον Μάρτιο καθώς τότε εφαρμόστηκε το lockdown στη Νέα Υόρκη και επίσης να θέσουμε σαν ποσοτικό όριο σε κάθε basket το 10, δηλαδή μόνο αν έχει πραγματοποιηθεί απο ένα zip-code μια καταγγελία πάνω από 10 φορές τη λαμβάνουμε υπόψη. Τέλος ορίζουμε το minimum support και minimum confidence ανάλογα σε κάθε περίπτωση ώστε να μην προκύψει πρόβλημα με την μνήμη του περιβάλλοντος που χρησιμοποιούμε. Προκύπτουν τα εξής αποτελέσματα :

a) **Manhattan***:*

- Πριν την Πανδημία (min support=0.3,min confidence=0.95):
  - 'Illegal Parking', 'Request Large Bulky Item Collection'
  - 'Noise - Residential', 'Street Condition'
  - 'HEAT/HOT WATER', 'Street Condition'

- Κατά τη διάρκεια της Πανδημίας(min support=0.4,min confidence=1):
  - 'Consumer Complaint', 'Illegal Parking', 'Street Condition'
  - 'Consumer Complaint', 'HEAT/HOT WATER'
  - 'Noise - Residential', 'Street Condition'

Παρατηρούμε την εμφάνιση του 'Consumer Complaint' στις καταγγελίες που προυπήρχαν. Μια ένδειξη της πίεσης που αρχίζει να ασκείται στους κατοίκους που αποτυπώνεται ως ευερέθιστη συμπεριφορά. Επίσης λόγω του lockdown είχαμε αύξηση των υπηρεσιών Courier που μπορεί να οδήγησαν σε αυτές τις καταγγελίες.

b) **Queens***:*

- Πριν την Πανδημία (min support=0.5,min confidence=1):
  - 'Blocked Driveway', 'Illegal Parking', 'Noise - Residential'
  - 'Blocked Driveway', 'Illegal Parking', 'Street Condition'
  - 'Illegal Parking', 'Noise - Residential', 'Street Condition'
- Κατά τη διάρκεια της Πανδημίας (min support=0.3,min confidence=0.8):
  - 'Blocked Driveway','Consumer Complaint','Sewer'

Κατά την διάρκεια της πανδημίας συνεχίζονται οι καταγγελίες να σχετίζονται με την κατάσταση στους εξωτερικούς χώρους, αλλά πλεον από τους ίδιους ανθρώπους εμφανίζονται και οι καταγγελίες 'Consumer Complaint' για τους λόγους που αναφέραμε παραπάνω.

c) **Staten Island***:*

- Πριν την Πανδημία (min support=0.85,min confidence=1):
  - 'Illegal Parking', 'Request Large Bulky Item Collection'
  - 'Request Large Bulky Item Collection', 'Street Condition'
  - 'Electronics Waste Appointment', 'Street Condition'
- Κατά τη διάρκεια της Πανδημίας (min support=0.9,min confidence=1):
  - 'Street Light Condition', 'Illegal Parking'
  - 'Street Light Condition', 'Noise - Residential'
  - 'Noise - Residential', 'Street Condition'

Παρατηρούμε ότι κατα τη διάρκεια της πανδημίας συχνό "ζευγάρι" καταγγελίας έχει να κάνει με την μεγαλύτερη ασφάλεια στη μετακίνηση.

d) **Brooklyn***:*

- Πριν την Πανδημία (min support=0.8,min confidence=1):
  - 'Blocked Driveway', 'HEAT/HOT WATER', 'Illegal Parking'
  - 'HEAT/HOT WATER', 'Noise - Residential', 'Street Condition'
  - 'Noise - Residential', 'Request Large Bulky Item Collection', 'Street Condition'
- Κατά τη διάρκεια της Πανδημίας (min support=0.9,min confidence=1):
  - 'Consumer Complaint', 'Illegal Parking', 'Street Condition'
  - 'General Construction/Plumbing', 'Noise - Residential'
  - 'General Construction/Plumbing', 'Street Condition'

Παρατηρείται μια τάση στη διάρκεια της πανδημίας για καταγγελίες που αφορούν τον οικιακό χώρο, γεγονός που υποδεικνύει εφαρμογή του lockdown.

e) **Bronx***:*

- Πριν την Πανδημία (min support=0.75,min confidence=1):
  - 'HEAT/HOT WATER', 'Illegal Parking'
  - 'Illegal Parking', 'Street Condition'
  - 'Blocked Driveway', 'HEAT/HOT WATER', 'Illegal Parking'
- Κατά τη διάρκεια της Πανδημίας (min support=0.85,min confidence=1):
  - 'Consumer Complaint', 'Illegal Parking', 'Street Condition'
  - 'Illegal Parking', 'HEAT/HOT WATER'
  - 'Noise - Residential', 'Consumer Complaint'

Παρατηρείται πάλι το 'Consumer Complaint' να έχει αντικαταστήσει ένα μέρος των καταγγελιών για την κατάσταση της οδικής κυκλοφορίας.

D. Assumptions Check - Text Mining

a) **Assumptions Check***:* Έχοντας στη διάθεση μας την ακριβή ημερομηνία και ώρα της καταχώρησης κάθε καταγγελίας καθώς και πότε επιλύθηκε το πρόβλημα δεν γίνεται να μην μπούμε στο πειρασμό να ελέγξουμε πως επίδρασε η πανδημία στις επιδόσεις των κρατικών φορέων στην επίλυση των ζητημάτων. Επιπροσθέτως μας ενδιαφέρει να ερευνήσουμε αν υπάρχει κάποια "προτίμηση" των αρχών ως προς ποια "διαμερίσματα" θα εξυπηρετηθούν πρώτα.

Για να ερευνήσουμε τη πρώτη υπόθεση θα χρησιμοποιήσουμε Two Sample Independent t-test[13] με άγνωστες και άνισες διακύμανσης για τους δύο πληθυσμούς καθώς έχουμε αρκετά μεγάλα μεγέθη δείγματος. Πραγματοποιώντας τον έλεγχο υπόθεσης καταλήγουμε στο ότι ο μέσος χρόνος ανταπόκρισης στα αιτήματα των πολιτών όχι μόνο δεν είναι ίδιος πριν και κατά τη διάρκεια της πανδημίας (t-test $p.value << 0.001$) αλλά οι έκτακτες καταστάσεις φαίνεται να έχουν επιδράσει θετικά στην ανταπόκριση των αρχών βελτιώνοντας τους χρόνους.

Όσο αφορά τη δεύτερη υπόθεση χρησιμοποιήθηκε η Ανάλυση Διακύμανσης[20] για τον έλεγχο της. Καταλήξαμε στο συμπέρασμα ότι υπάρχουν διαφορές στο μέσο χρόνο επίλυση των προβλημάτων ανά περιοχή πριν και μετά την πανδημία (F-test $p.value << 0.001$). Πιο συγκεκριμένα

παρατηρούμε ότι τα προβλήματα φαίνεται να διευθετούνται γρηγορότερα κατά μέσο όρο στο Bronx(πριν και κατά την διάρκεια της πανδημίας) ενώ αντίθετα στο State Island χρειάζεται περισσότερος χρόνος κατά μέσο όρο για την επίλυση τους. Σε γενικές γραμμές παρατηρούνται παρόμοιες συμπεριφορές ως προς τον χρόνο επίλυσης των προβλημάτων μεταξύ των Queens-Brooklyn, Manhattan-Bronx κατά την πανδημία ενώ πριν από αυτή παρόμοια συμπεριφορά παρατηρείται μεταξύ των Queens-Brooklyn-Manhattan.

b) **Text Mining***:* Σε αυτή την ενότητα θα προσπαθήσουμε να εκμεταλλευτούμε την μεταβλητή Descriptor η οποία περιέχεται στο σετ δεδομένων και αφορά την περιγραφή της καταγγελίας σε μορφή κειμένου. Χρησιμοποιώντας κλασικές τεχνικές επεξεργασίας φυσικής γλώσσας όπως αφαίρεση stopwords, lemmatization και tokenazation έχουμε την δυνατότητα να βρούμε τις πιο σημαντικές λέξεις που εμφανίζονται στο πεδίο της περιγραφής. Στα σχήματα 11-12 παρουσιάζουμε κάποιες από αυτές πριν και κατά τη διάρκεια της πανδημίας. Όσο μεγαλύτερη η λέξη τόσο πιο σημαντική. Παρατηρούμε ότι στην πόλη που δεν κοιμάται ποτέ οι κάτοικοι της παραπονιούνται κυρίως για ζητήματα που αφορούν πάρτι και δυνατή μουσική.

Figure 11. WordCloud for Complaints Description Before Covid-19

Figure 12. WordCloud for Complaints Description During Covid-19

Πριν την πανδημία κυριαρχούν λέξεις οι οποίες σχετίζονται με τη παραλαβή απορριμμάτων και τη συμπεριφορά των ανθρώπων σε εξωτερικούς χώρους ενώ κατά τη διάρκεια της πανδημίας υπάρχει μια μετατόπιση προς λέξεις που σχετίζονται άμεσα με κτιριακές υποδομές.

Επιπροσθέτως έχει ενδιαφέρον ο εντοπισμός διαφορών ανάμεσα στις πέντε περιοχές, όπως αυτές περιγράφηκαν στη πρώτη ενότητα, ως προς τη περιγραφή των παραπόνων. Για να το κάνουμε αυτό υπολογίζουμε το tf idf[15] για κάθε λέξη στο σύνολο των δεδομένων ανά περιοχή τόσο πριν αλλά και κατά τη διάρκεια της πανδημίας. Στη συνέχεια θα παραθέτουμε ορισμένες ενδεικτικές λέξεις με υψηλούς δείκτες για κάθε περιοχή πριν και κατά τη διάρκεια της πανδημίας:

- ***Bronx***
  - Πριν την Πανδημία:
    - ∗ 'fuse'
    - ∗ 'scaffolding'
  - Κατά τη διάρκεια της Πανδημίας
    - ∗ 'feed'
    - ∗ 'insurance', 'communication'
    - ∗ 'roofing'

Επιπλέον και στις δύο περιόδους εμφανίζονται λέξεις που σχετίζονται με τη χρήση ναρκωτικών ουσιών όπως 'niddle', 'syringe'.

- ***Brooklyn***
  - Πριν την Πανδημία:
    - ∗ 'ewaste'
    - ∗ 'appointment'
  - Κατά τη διάρκεια της Πανδημίας
    - ∗ 'establishment'
    - ∗ 'depression', 'communication'
    - ∗ 'roofing'
- ***Manhattan***
  - Πριν την Πανδημία:
    - ∗ 'glasses'
    - ∗ 'pedicab'
    - ∗ jewellery'
  - Κατά τη διάρκεια της Πανδημίας
    - ∗ 'maladjusted'
    - ∗ 'wallet', 'bag'
    - ∗ 'clothing'
- ***Queens***
  - Πριν την Πανδημία:
    - ∗ 'ewaste'
    - ∗ 'appointment'
  - Κατά τη διάρκεια της Πανδημίας
    - ∗ 'depression',
    - ∗ 'wallet', 'bag'
    - ∗ 'phones'
- ***Staten Island***
  - Πριν την Πανδημία:
    - ∗ 'ewaste'
    - ∗ 'appointment'
  - Κατά τη διάρκεια της Πανδημίας
    - ∗ 'fuse', 'receptacles'
    - ∗ 'deck'
    - ∗ 'junction'

V. Συμπεράσματα

Συνοπτικά χρησιμοποιήσαμε το NYC Open Data το οποίο είναι ένα σύνολο δεδομένων που αφορά μικρές καταγγελίες στη Νέα Υόρκη, με έναν διαφορετικό τρόπο απότι εμφανίζεται στη βιβλιογραφία.

Αξιοποιήσαμε την ποικιλομορφία στα "διαμερίσματα" της Νέας Υόρκης και σε συνδυασμό με την πανδημία του Covid-19 προσπαθήσαμε να παρατηρήσουμε αλλαγές σε κοινωνικά φαινόμενα. Θεωρήσαμε ότι οι μικρές καταγγελίες μπορούν να εκφράσουν την κοινωνική κατάσταση μιας περιοχής.

Αρχικά ορίσαμε με βάση την σημασιολογία των καταγγελιών κάποια κοινωνικά φαινόμενα, με σκοπό να τα μελετήσουμε. Παρατηρήσαμε όξυνση αρνητικών φαινομένων όπως η ποιότητα της ατμόσφαιρας σε περιοχές που ήδη πάσχουν από χαμηλό επίπεδο ζωής ή σε περιοχές που έχουν υψηλή πυκνότητα πληθυσμού. Παρατηρήσαμε επίσης όξυνση αρνητικών φαινομένων ανεξαρτήτως χαρακτηριστικών περιοχής, όπως η κατάσταση των αστέγων.

Στη συνέχεια με μεθόδους clustering έγινε η ομαδοποίηση των περιοχών της Νέας Υόρκης με βάση τις καταγγελίες και εντοπίσαμε τα κοινά χαρακτηριστικά αυτών των περιοχών και ποιες καταγγελίες κυριαρχούν.

Έπειτα μελετήσαμε την τάση των κατοίκων στον συνδυασμό των καταγγελιών ανά περιοχή και πως η πανδημία άλλαξε την κατεύθυνση τους. Όπως είδαμε πριν την πανδημία οι κάτοικοι είχαν την τάση να συνδυάζουν την πλειοψηφία των καταγγελιών με παρατηρήσεις εξωτερικού χώρου, ενώ με την πανδημία πλέον το ενδιαφέρον στράφηκε και στην ποιότητα της εξυπηρέτησης που δεχόντουσαν ως καταναλωτές.

Επιπροσθέτως ερευνήσαμε την επιρροή της πανδημίας στις αποδόσεις των κρατικών μηχανισμών στην διευθέτηση των καταγγελιών. Και καταλήξαμε ότι τόσο η πανδημία όσο και το "διαμέρισμα" από το οποίο έχει προέλθει η καταγγελία επιδρούν στατιστικά σημαντικά στο μέσο χρόνο διευθέτηση τους.

Τέλος χρησιμοποιώντας την περιγραφή των παραπόνων προχωρήσαμε σε text mining σε μια προσπάθεια εύρεσης των σημαντικότερων λέξεων που χαρακτηρίζουν τις καταγγελίες των πολιτών και εν γένει τα σημαντικότερα προβλήματα. Μέσα από αυτή τη διαδικασία καταφέραμε να αποτυπώσουμε την αλλαγή στη ψυχοσύνθεση των πολιτών μετά την έλευση της πανδημίας καθώς παρατηρείτε μια αλλαγή στο είδος των λέξεων που χρησιμοποιούνται στη περιγραφή των παραπόνων ανά "διαμέρισμα".

Βιβλιογραφία

[1] https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9. [Online].

[2] https://oikonang.github.io/social_data_visualization/prj/. [Online].

[3] https://schwarzwald-ai.medium.com/predicting-new-york-city-311-service-requests-7a08a6014f25. [Online].

[4] https://data.cityofnewyork.us/City-Government/2020-population/t8c6-3i7b/data. [Online].

[5] https://worldpopulationreview.com/boroughs/staten-island-population. [Online].

[6] https://www.census.gov/quickfacts/fact/table/newyorkcountymanhattanboroughnewyork, bronxcountybronxboroughnewyork, queenscountyqueensboroughnewyork, kingscountybrooklynboroughnewyork, richmondcountystatenislandboroughnewyork, newyorkcitynewyork/HSG010219. [Online].

[7] https://www.towncharts.com/New-York/Education/Manhattan-borough-NY-Education-data.html. [Online].

[8] https://data.cccnewyork.org/data/bar/85/unemployment-rate#85/a/1,15,28,2,47,62/131/62. [Online].

[9] Rakesh Agrawal, Ramakrishnan Srikant, et al. "Fast algorithms for mining association rules". In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. Citeseer. 1994, pp. 487–499.

[10] Jessica Athens et al. "Using 311 data to develop an algorithm to identify urban blight for public health improvement". In: *PLOS ONE* 15.7 (July 2020), pp. 1–11. DOI: 10.1371/journal.pone.0235227. URL: https://doi.org/10.1371/journal.pone.0235227.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[12] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proc. of 2nd International Conference on Knowledge Discovery and*. 1996, pp. 226–231.

[13] William Sealy Gosset. "The Probable Error of a Mean". In: *Biometrika* 6.1 (Mar. 1908). Originally published under the pseudonym "Student"., pp. 1–25. URL: http://dx.doi.org/10.2307/2331554.

[14] Jiawei Han, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation". In: *ACM sigmod record* 29.2 (2000), pp. 1–12.

[15] Karen Spärck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 28.1 (1972). URL: http://www.soi.city.ac.uk/~ser/idfpapers/ksj_orig.pdf.

[16] Stuart P. Lloyd. "Least squares quantization in PCM." In: *IEEE Trans. Inf. Theory* 28.2 (1982), pp. 129–136. URL: http://dblp.uni-trier.de/db/journals/tit/tit28.html#Lloyd82.

[17] G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker, 1988.

[18] Lingjing Wang et al. "Structure of 311 service requests as a signature of urban location". In: *PloS one* 12.10 (2017), e0186314.

[19] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244. URL: http://www.jstor.org/stable/2282967.

[20] F. Yates. “The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes”. In: *Journal of the American Statistical Association* 29.185 (1934), pp. 51–66. DOI: 10.1080/01621459.1934.10502686. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1934.10502686. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1934.10502686.