ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ ΣΤΟ ΜΑΘΗΜΑ
ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΙΑ ΤΗΝ ΑΠΟΦΥΓΗ ΕΓΚΛΗΜΑΤΙΚΩΝ ΠΕΡΙΩΧΩΝ ΣΤΟ ΒΑΝΚΟΥΒΕΡ

ΟΜΑΔΑ

Λεονάρδος Φρέρης 2696
Μάριος-Χρυσόστομος Ασκητής 2760

Υπεύθυνος Καθηγητής : Βασιλακόπουλος Μιχαήλ

ΒΟΛΟΣ 2022

ΕΙΣΑΓΩΓΗ

Σκοπός της εργασίας

Το Βανκούβερ, όπως όλες οι μεγάλες πόλεις στο μέγεθός του, είναι ένα πολύ ασφαλές μέρος για να ζήσετε ή να επισκεφθείτε με αρκετά χαμηλό ποσοστό εγκληματικότητας.Ωστόσο, εξακολουθούν να υπάρχουν εγκλήματα και παράνομες δραστηριότητες σε σημαντικό βαθμό. Στην εργασία, θα θέλαμε να αναλύσουμε και να προβλέψουμε εάν ένα άτομο θα αντιμετωπίσει έγκλημα, με δεδομένο πότε και πού. Πότε ως προς την ώρα και την ημέρα και πού ως προς τη γειτονιά που βρίσκεται το άτομο. Ολοκληρώνοντας την εργασία, θα πρέπει να κατανοήσουμε καλύτερα ποιο μέρος της πόλης πρέπει να αποφύγουμε σε μια συγκεκριμένη στιγμή. Εάν η ανάλυση είναι επιτυχής, θα είναι μια τεράστια βοήθεια τόσο για το ευρύ κοινό όσο και για την αστυνομία του Βανκούβερ.

Τεχνικές εξόρυξης που θα χρησιμοποιηθούν

Θα δοκιμάσουμε δύο ταξινομητές (classifiers), το δέντρο αποφάσεων (Desicion Tree) και των Κ κοντινότερων γειτόνων (K-Nearest neighbour), στο σύνολο των δεδομένων μας. Αφού λάβουμε πρωτεύοντα αποτελέσματα, θα επιλέξουμε για το μοντέλο μας τον ταξινομητή με την μεγαλύτερη ακρίβεια.

Ποιοι είναι οι ισχυρισμοί μας;

Προτού προχωρήσουμε στην ανάλυση, χρησιμοποιώντας τις γνώσεις μας για το Βανκούβερ και την κατανόησή μας για το πώς λειτουργεί ένας άνθρωπος, θα κάνουμε τους ακόλουθους ισχυρισμούς :

- Το ποσοστό εγκληματικότητας θα εκτιναχθεί το απόγευμα, όταν οι άνθρωποι φεύγουν από τη δουλειά και το σχολείο.
- Το ποσοστό εγκληματικότητας θα εκτιναχθεί τις ημέρες εκτός λειτουργίας, στις οποίες περιλαμβάνονται τα Σαββατοκύριακα και οι αργίες.
- Θα υπάρχουν γειτονιές με υψηλά συγκεκριμένα ποσοστά εγκληματικότητας, όπως το Downtown Eastside και τα εγκλήματα ιδιοκτησίας, όπως αναφέρθηκε παραπάνω.

Δεδομένα

Πηγή δεδομένων

Τα δεδομένα προέρχονται από τον κατάλογο ανοιχτών δεδομένων του Βανκούβερ. Συλλέγεται από το Αστυνομικό Τμήμα του Βανκούβερ με την πάροδο των ετών και ενημερώνεται μία φορά την εβδομάδα τα πρωινά της Κυριακής. Το συγκεκριμένο σύνολο δεδομένων που θα χρησιμοποιήσουμε στο έργο περιλαμβάνει δεδομένα εγκληματικότητας από το 2003/01/01 έως το 2019/12/31.Επιλέξαμε να μην συμπεριλάβουμε στα δεδομένα τα τελευταία δύο χρόνια και τους τελευταίους μήνες του τρέχουν έτους (2022) για να εξάγουμε πιο ασφαλή συμπεράσματα. Περιλαμβάνεται ο σύνδεσμος προς το σύνολο δεδομένων:
https://geodash.vpd.ca/opendata/
Όπως φαίνεται παρακάτω, τα χαρακτηριστικά του συνόλου δεδομένων περιλαμβάνουν:

1. TYPE: Είδος εγκλήματος
2. YEAR: Έτος κατά το οποίο σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
3. MONTH: Μήνας κατά τον οποίο σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
4. DAY: Ημέρα κατά την οποία σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
5. HOUR: Ώρα κατά την οποία σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
6. MINUTE: Λεπτό κατά το οποίο σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
7. HUNDRED BLOCK: Γενικευμένη τοποθεσία της αναφοράς εγκληματικής δραστηριότητας
8. NEIGHBOURHOODN: Γειτονιά όπου σημειώθηκε η αναφερόμενη εγκληματική δραστηριότητα
9. X: Τιμές συντεταγμένων
10. Y: Τιμές συντεταγμένων

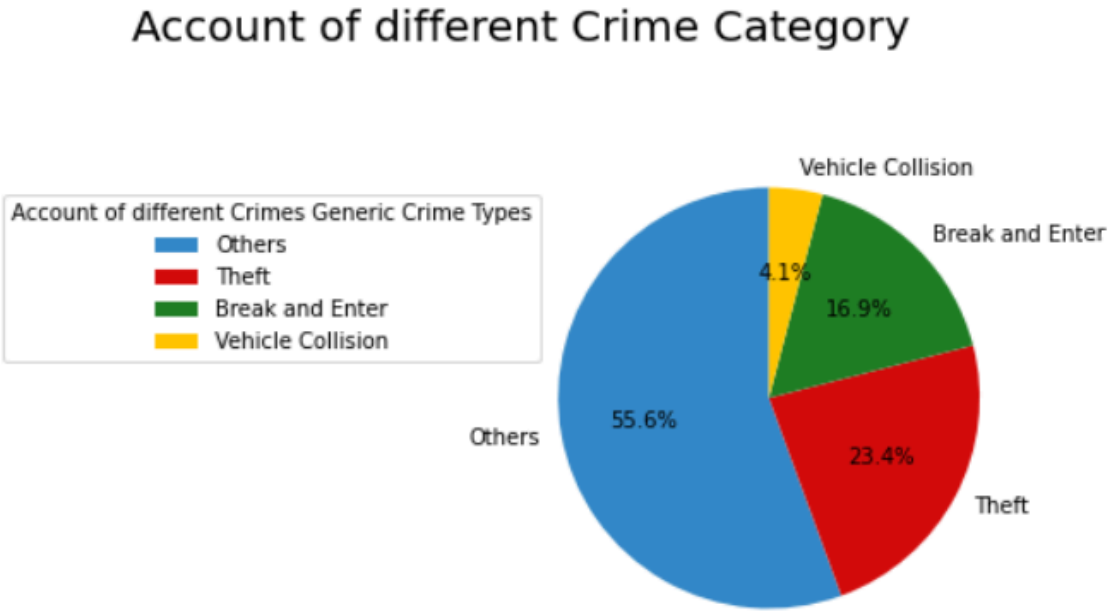
Προεπεξεργασία Δεδομένων

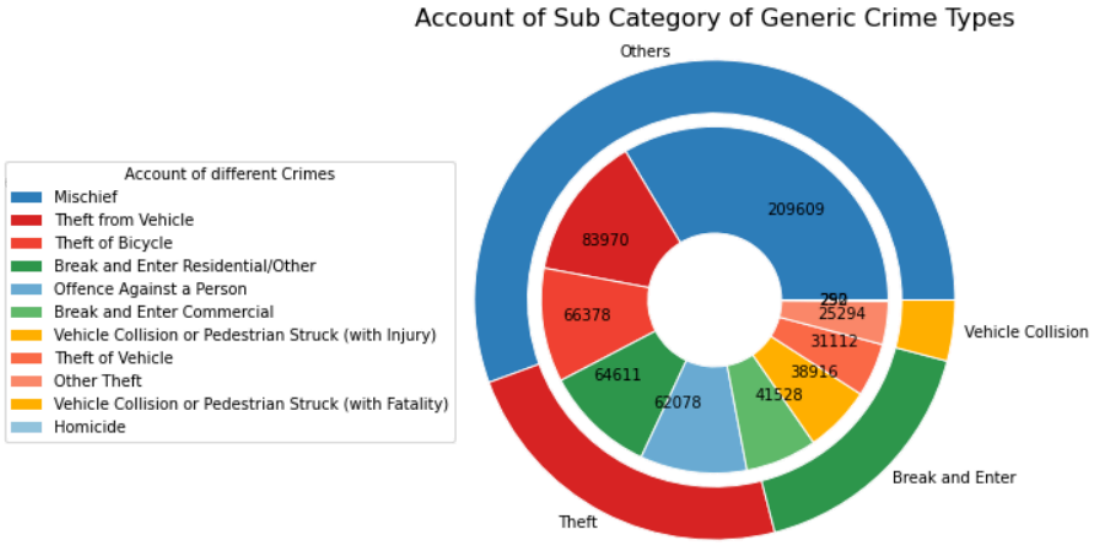
Αρχικά παρατηρήσαμε, πως υπήρχαν μη πλήρης εγγραφές. Δηλαδή εντοπίσαμε στήλες με άγνωστες τιμές. Γ' αυτό παράξαμε κώδικα που μας δείχνει τα ποσοστά ελλιπών τιμών για κάθε στήλη. Εντοπίσαμε ότι μόνο οι στείλες NEIGHBOURHOODN , HUNDRED BLOCK, X και Y έχουν NaN τιμές και πιο συγκεκριμένα πολλές περισσότερες η στήλη NEIGHBOURHOODN απ' τις υπόλοιπες. Ως εκ'τούτου προσπαθήσαμε να συμπηρώσουμε τα κενά της στήλης NEIGHBOURHOODN με την βοήθεια των των συντεταγμένων X, Y . Ειδικότερα ελέγχουμε τις συντεταγμένες X, Y των τιμών NEIGHBOURHOODN που είναι NaN.Αν είναι μηδέν διαγράφεται η εγγραφή από το σύνολο των δεδομένων, ενώ αν είναι διάφορες του μηδενός η τιμή NEIGHBOURHOODN προσεγγίζεται βάσει κοντινών συντεταγμένων του X, Y.

Στην συνέχεια, προσθέσαμε μια νέα στήλη στο dataset μας με όνομα CATEGORY στην οποία θα κατηγοριοποιήσουμε τα εγκλήματα της στήλης TYPE σε πιο γενικές κατηγορίες (Theft, Break and Enter, Vehicle collision, Others). Επίσης δημιουργούμε τις στήλες DATETIME και WEEKDAY, οι οποίες θα περιέχουν την ημερομηνία και την ημέρα της εβδομάδας κάθε εγγραφής αντίστοιχα. Τέλος δημιουργύμε τις στήλες ISHOLIDAY, ISOFFDAY για να ελένχουμε αν είναι ημέρα διακοπών ή αργίας.

Οπτικοποίηση ανάλυσης δεδομένων

Καταρχάς, προσπαθήσαμε να εντοπίσουμε το είδος και το πλήθος των εγκλημάτων. Με την βοήθεια μιας πίτας παρατηρούμε το ποσοστό κάθε γενικής κατηγορίας εγκλήματος που λαμβάνει χώρα στο Βανκούβερ και μέσω δύο δακτυλίων καταγράφουμε το πλήθος των διαφορετικών υποκατηγοριών εγκλημάτων σε σύγκριση με τις γενικές κατηγορίες.

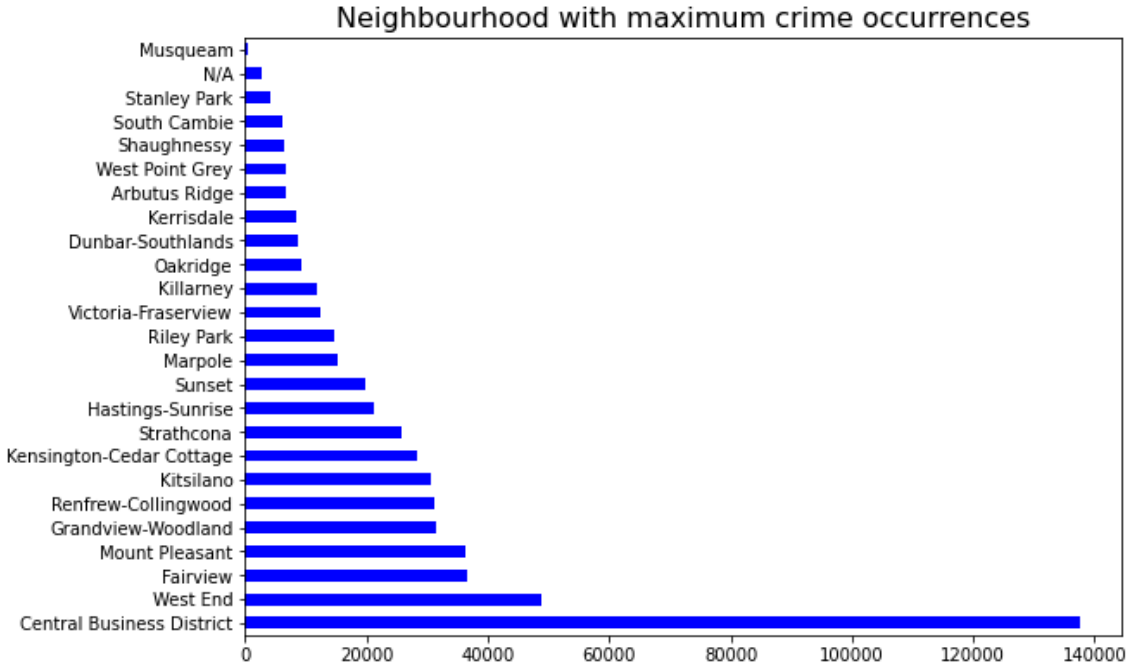




Έτσι παρατηρήσαμε πως η κατηγορία εγκλήματος Others είναι πιο συχνή στην πόλη μιας και το έγκλημα Mischief (κακούργημα) εμφανίζει το μεγαλύτερο πλήθος. Στην συνέχεια, ακολουθούν η κατηγορία Theft με την πιο πολυπληθή υποκατηγορία της να είναι η κλοπή από αυτοκίνητο (Theft from Vehicle), η κατηγορία Break and Enter που εντοπίζεται πιο πολύ στα εμπορικά καταστήματα παρά στις κατοικίες και τέλος η κατηγορία Vehicle Collision με αισθητα μικρό ποσοστό και ιδιαίτερα αυτα που επιφέρουν θάνατο. Είναι σημαντικό να αναφέρουμε πως συναντάμε πολύ σπάνια περιστατικά ανθρωποκτονιών. Άρα μπορούμε να συμπεράνουμε πως τα εγκλήματα δεν είναι μεγάλα στο πλήθος τους και πλήττουν κυρίως τις περιουσίες παρά τις ζωές των κατοίκων της πόλης και συμφωνεί με την αρχική μας εκτίμηση πως το Βανκούβερ αποτελεί μια

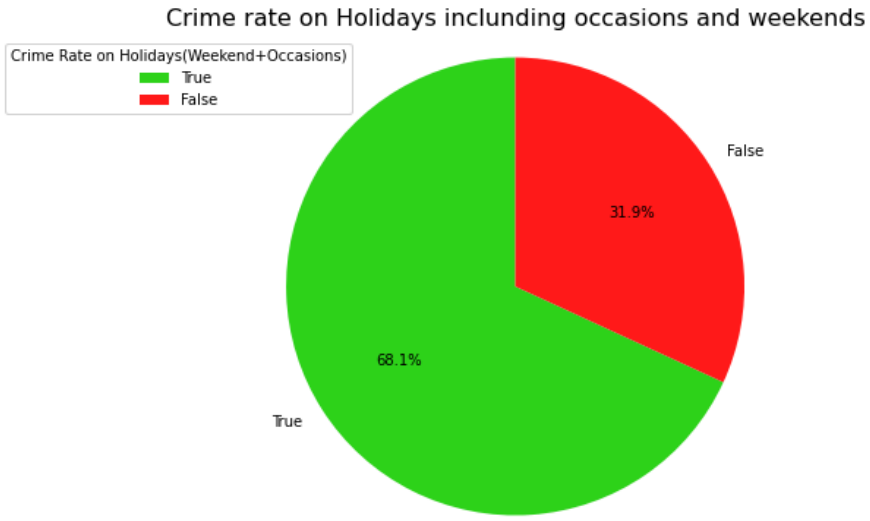
ασφαλή πόλη για το μέγεθος της.

Στην συνέχεια, προχωρήσαμε στην ανάλυση των στηλών NEIGHBOURHOODN και HUNDRED BLOCK για να εντοπίσουμε τις γειτονιές με την περισσότερη εγκληματικότητα.

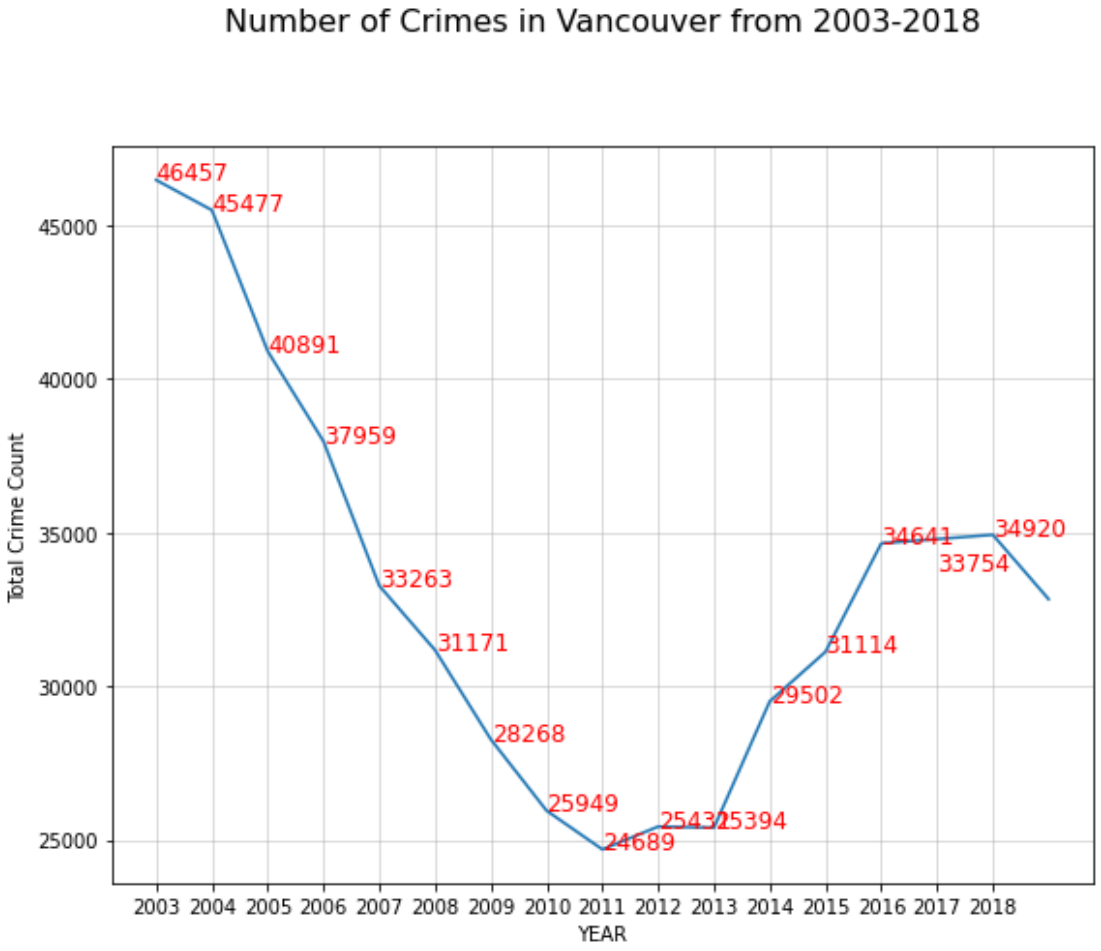


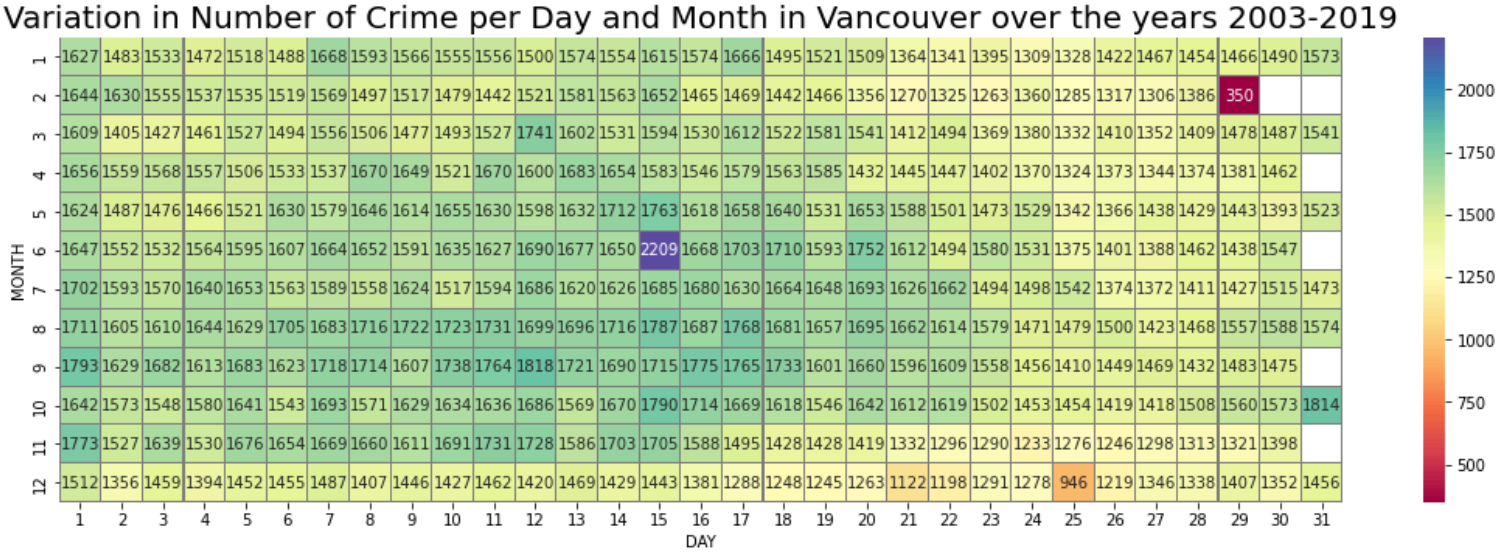
Παρατηρώντας το παραπάνω διάγραμμα βλέπουμε ότι η γειτονιά με την μεγαλύτερη εγκληματικότητα είναι η Central Businness District, με συντριπτική διαφορά από τις υπόλοιπες.

Έπειτα με την βοήθεια της στήλης ISOFFDAY θέλουμε να εξετάσουμε αν υπάρχει μια συσχέτιση μεταξύ της εγκληματικότητας και της στήλης αυτής. Για τον λόγο αυτό δημιουργήσαμε την παρακάτω πίτα.



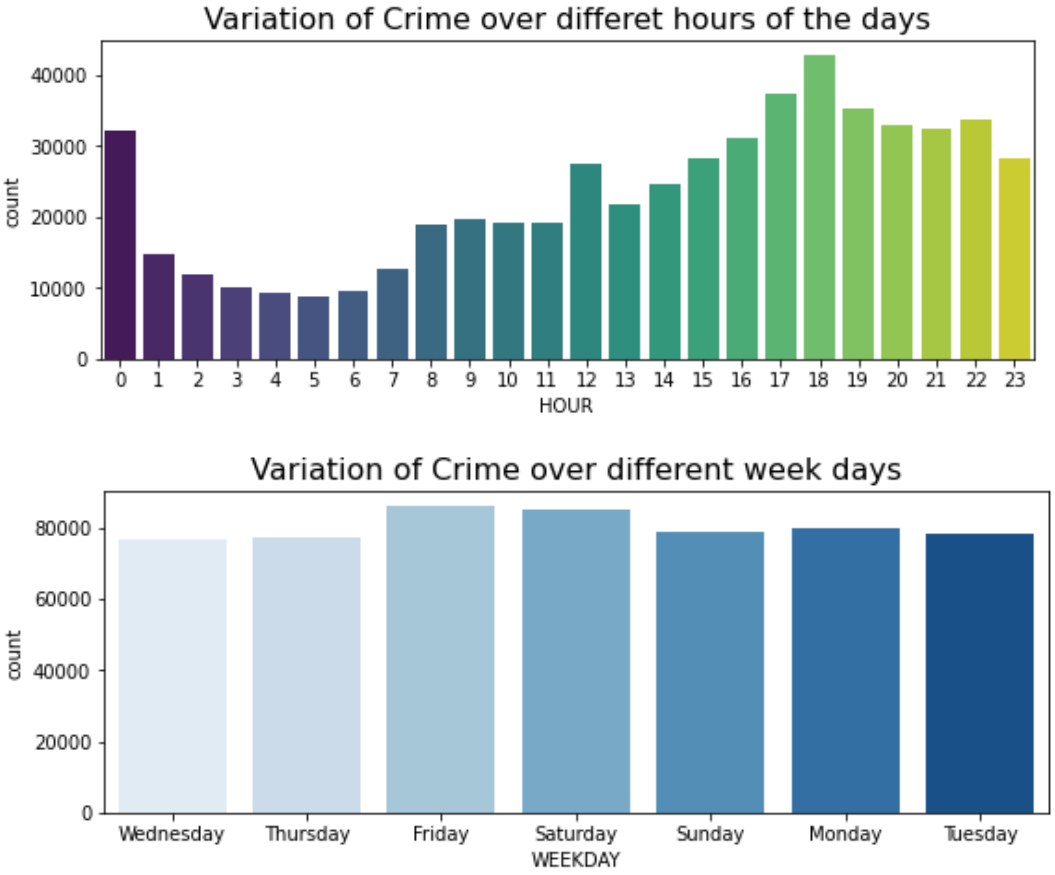
Πράγματι, κατά 70% περίπου αληθεύει ότι έχουμε μεγαλύτερη εγκληματικότητα τις ημέρες που έχουμε διακοπές (αργίες και σαββατοκύριακα).
Ως προς την ανάλυση των στηλών YEAR, MONTH, DAY δημιουργήσαμε ένα lineplot, όπου βλέπουμε τα συνολικά εγκλήματα ανά χρονία και ένα heatmap με τον συνολικό αριθμό εγκλημάτων για κάθε μέρα του χρόνου.





όσον αφορά την εξάρτηση εγκληματικότητας με την στήλη YEAR, δεν μπορούμε να πούμε σιγουριά πως υπάρχει καθώς βλέπουμε πως μέχρι το 2011 έχουμε μία πτωτική τάση και μετά έχουμε μια αύξηση που δεν ξεπερνά αυτήν μεταξύ του 2003-2007. Ενώ για τις ημέρες κάθε χρόνου μπορούμε να παρατηρήσουμε πως έχουμε ελάχιστη εγκληματικότητα τις ημέρες της περιόδου των Χριστουγέννων με αποκορύφωμα στης 25/12. Στις 29/2 βλέπουμε την ελάχιστη τιμή του heatmap αλλά δεν μπορούμε να την συμπεριλάβουμε στα συμπεράσματα μας καθώς εμφαωίζεται μόνο στα δίσεκτα έτη και είναι λογικό να έχουμε τόσο μικρό συνολικό αριθμό εγκλημάτων εκείνη την ημέρα. Παράλληλα βλέπουμε πως η μεγαλύτερη εγκληματικότητα γενικά εμφανίζεται τις μέρες του καλοκαιριού και ιδιαίτερα στην μέση των μηνών.Τέλος μας προξένησε το ενδιαφέρον ο μεγάλος συνολικός αριθμός των εγκλημάτων στης 15/6, οπότε αναζητήσαμε τα γεγονότα που συνέβησαν εκείνες τις μέρες στο Βανκούβερ και ανακαλύψαμε πως το 2011 συνέβη μια δημόσια αναταραχή στο κέντρο της πόλης μετά τον τελικό του Stanley Cup το οποίο επέφερε μεγάλη έξαρση εγκληματικότητας στην πόλη.

Τελειώνοντας την ανάλυση ελέγχουμε πως κυμαίνεται η εγκληματικότητα με το πέρασμα των ωρών και των ημερών της εβδομάδας. Για τον σκοπό αυτό δημιουργήσαμε τα δύο παρακάτω barplots:



Έτσι, για την στήλη HOUR παρατηρούμε ότι τις απογευματινές και βραδινές ώρες μέχρι τα μεσάνυχτα έχουμε έντονη εγκληματικότητα σε σχέση με τις υπόλοιπες ώρες τη μέρας. Συγκεκριμένα, την μέγιστη την εντοπίζουμε στης 18:00. Επιπλέον, για την στήλη WEEKEND παρατηρούμε ότι την μεγαλύτερη εγκληματικότητα την εμφανίζουν η Παρασκευή και το Σάββατο.

Εκπαίδευση μοντέλου

Πρώτα απ' όλα θα πρέπει να επιλέξουμε τα χαρακτηριστικά (στήλες) που θα χρησιμοποιηθούν στο μοντέλο μας. Μέσω της ανάλυσης και οπτικοποίησης των δεδομένων που έγινε προηγουμένως, συμπεράναμε πως η εγκληματικότητα έχει την μεγαλύτερη εξάρτηση με τις στήλες : MONTH, DAY, HOUR, ISOFFDAY και NEIGHBOURHOODN. Επομένως, τα δεδομένα εισαγωγής (dataX) του μοντέλου μας θα αποτελούν οι στήλες που αναφέραμε ενώ τα δεδομένα εξόδου (dataY) θα αποτελεί η στήλη TYPE μιας και θέλουμε να προβλέψουμε το είδος του εγκλήματος. Γ' αυτό το λόγο προηγουμένος προστέσαμε εγραφές NO CRIME γιατί σε αντίθεση περίπτωση το μοντέλο μας θα προέβλεπε πάντα ότι θα συνέβαινε κάποιο έγκλημα.

Προτού προχωρήσουμε στην ανάπτυξη του μοντέλου θα πρέπει να μετατρέψουμε τις στήλες NEIGHBOURHOODN, ISOFFDAY, TYPE σε αριθμούς αφού αποτελούν κατηγορικές τιμές.

Στην συνέχεια, ξεκινάμε να αναπτύσσουμε το μοντέλο μας διαχωρίζοντας τα δεδομένα εισόδου και εξόδου (dataX, dataY) σε δεδομένα εκπαίδευσης και ελέγχου (trainX, trainY, testX, testY). Έπειτα, εφαρμόζουμε τις μεθόδους ταξινόμησης. Επιλέξαμε να χρησιμοποιήσουμε τους ταξινομητές K κοντινότερων γειτόνων και δέντρο απόφασης, όπου αναφέρθηκαν στο μάθημα αλλά και να βγάλουμε μια πιο ασφαλή ακρίβεια στο μοντέλο. Κατά την εφαρμογή των δύο ταξινομητών χρησιμοποιήσαμε τις προεπιλεγμένες παραμέτρους, που μας παρέχει η βιβλιοθήκη μηχανικής μάθησης της python, sklearn και καταφέραμε να πετύχουμε και με τις δύο μεθόδους ακρίβεια της τάξεως του 84%.Είναι σημαντικό να αναφερθεί ότι επιλέξαμε να χρησιμοποιήσουμε 5 γείτονες για την μέθοδο των K κοντινότερων γειτόνων διότι, μετά από αρκετές δοκιμές είδαμε πως σε αυτήν την τιμή K είχαμε αισθητά τη μεγαλύτερη αύξηση στην ακρίβεια, ενώ για μεγαλύτερα K η ακρίβεια συνέχιζε να αυξάνεται ελάχιστα χωρίς να ξεπεράση το 86%.
Για την εφαρμογή του μοντέλου επιλέξαμε το δέντρο απόφασης ως ταξινομητή και παίρνουμε ως είσοδο πότε (ώρα, ημέρα, μήνα, αν είναι αργία) και που (γει-

τονιά) επιθυμεί να βρεθεί ο χρήστης. Ως έξοδο παίρνουμε το μύνημα αν είναι ασφαλής ή όχι η μετάβαση του.

Συμπεράσματα

Ας ρίξουμε μια ματιά στους ισχυρισμούς μας ξανά. Το ποσοστό εγκληματικότητας θα εκτιναχθεί το απόγευμα, όταν οι άνθρωποι φεύγουν από τη δουλειά και το σχολείο αυτό είναι σωστό. Μπορούμε να δούμε ότι ο αριθμός των εγκλημάτων κορυφώνεται γύρω στις 18:00. Αυτή είναι η στιγμή που οι άνθρωποι φεύγουν από τη δουλειά και το σχολείο. Το ποσοστό εγκληματικότητας θα εκτιναχθεί τις ημέρες εκτός λειτουργίας, στις οποίες περιλαμβάνονται τα Σαββατοκύριακα και οι αργίες. Αυτό είναι επίσης εν μέρει σωστό. Πιστεύαμε ότι τα ποσοστά εγκληματικότητας θα πρέπει να είναι υψηλότερα τις ημέρες εκτός λειτουργίας. Αντίθετα, μπορούμε να δούμε ότι ο αριθμός των εγκλημάτων την Παρασκευή και το Σάββατο είναι υψηλότερος από την υπόλοιπη εβδομάδα. Αυτό είναι λογικό επειδή οι περισσότεροι άνθρωποι ετοιμάζονται να απολαύσουν τα Σαββατοκύριακα τους την Παρασκευή αφού φύγουν από τη δουλειά (να θυμάστε ότι ο αριθμός των εγκλημάτων είναι υψηλότερος όταν οι άνθρωποι αρχίζουν να φεύγουν από τη δουλειά και το σχολείο) και ετοιμάζονται να επιστρέψουν στη δουλειά την Κυριακή. που μπορεί να τους εμποδίσει να φύγουν. Είναι ενδιαφέρον να σημειωθεί ότι ενώ η Παρασκευή και το Σάββατο έχουν τους δύο υψηλότερους αριθμούς εγκλημάτων της εβδομάδας, η Κυριακή δεν είναι πολύ πίσω, όντας 3η στο chart. Αυτό αποδεικνύει και τους ισχυρισμούς μας. Επιπλέον, με απλά μαθηματικά και λίγη γνώση της αναλογίας, μπορούμε να δούμε ότι οι ημέρες εκτός λειτουργίας (Σαββατοκύριακα και αργίες) έχουν υψηλότερα ποσοστά εγκληματικότητας, αν και μόνο κατά περίπου 10% σε σύγκριση με άλλες ημέρες. Θα υπάρχουν γειτονιές με υψηλά συγκεκριμένα ποσοστά εγκληματικότητας, όπως το Downtown Eastside και τα εγκλήματα ιδιοκτησίας, όπως αναφέρθηκε παραπάνω. Αυτό δεν αναλύεται σωστά, αλλά με μια γρήγορη επιθεώρηση του συνόλου δεδομένων, αυτό είναι επίσης σωστό. Εκτός από γειτονιές με υψηλά συγκεκριμένα ποσοστά εγκληματικότητας, υπάρχουν και μέρες με υψηλά ποσοστά ειδικής εγκληματικότητας. Είτε πρόκειται για επαναλαμβανόμενη ημέρα είτε για μη επαναλαμβανόμενη ημέρα. Για παράδειγμα, φαίνεται να υπάρχει υψηλό ποσοστό επίθεσης την ημέρα της Πρωτοχρονιάς.