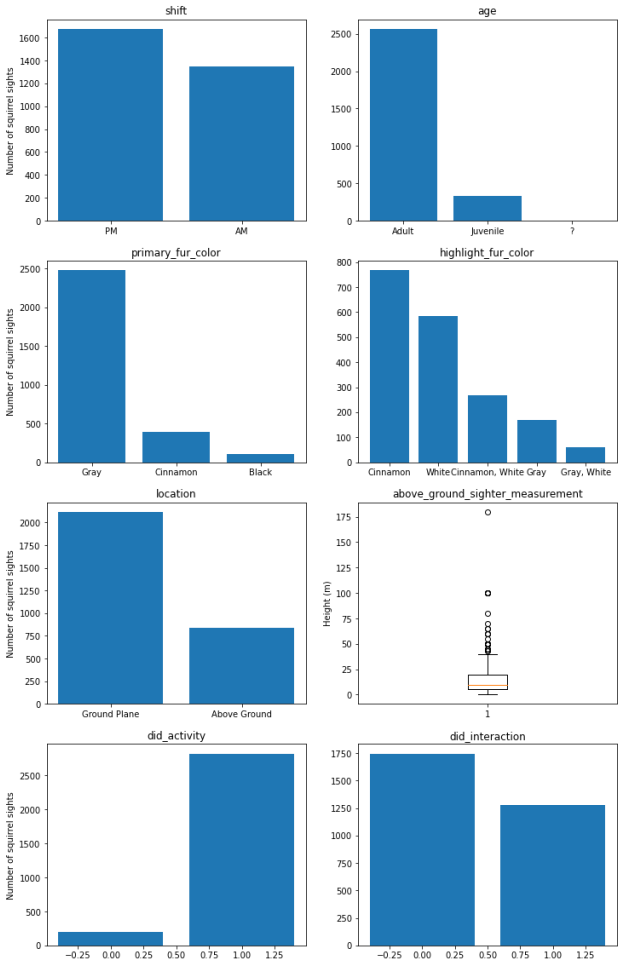
Rudolf Nosek
2DAV

Report projektu z predmetu Manažment dát
Téma: Veveričky v central parku, NYC

Ciele projektu boli predovšetkým získať čo najviac informácii z datasetu o nájdených veveričkách, tieto informácie vizualizovať, využiť inteligentným spôsobom na predikciu, výsledky vyvesiť na webovú stránku a vytvoriť webovú aplikáciu na predikciu atribútov veveričky. Dáta sme čerpali z github repozitára, ktorý viete nájsť tu: https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-10-29.

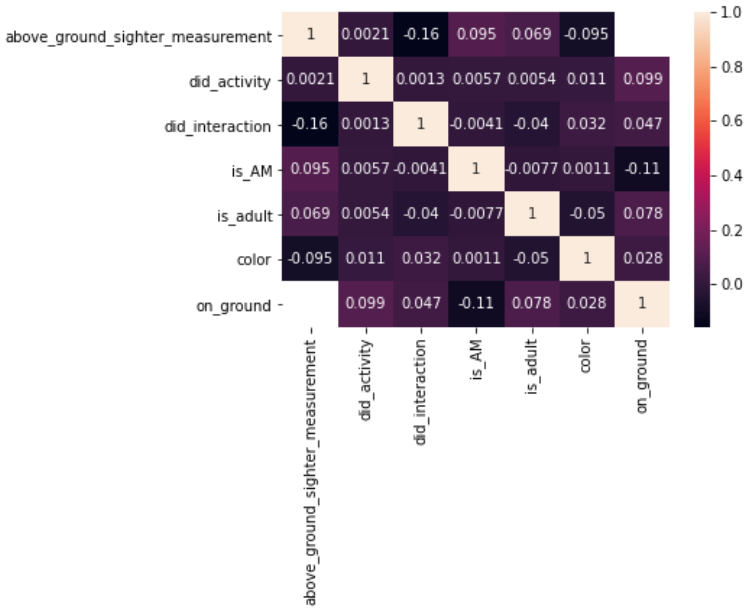
Po spracovaní dát sme mali k dispozícii 36 atribútov o 3023 nálezov veveričiek. V prvom rade sme sa snažili počet atribútov zredukovať, podľa počtu vyplnených hodnôt alebo podľa relevantnosti k naším cieľom. V prvom rade sme zistili, že dáta máme z časového interval od 6.10.2018 až 20.10.2018. Keďže jeden z atribútov bolo id veveričky, zistili sme že, v rámci tohto časového úseku bol len 5 veveričiek zahliadnutých viackrát. Ako je vidieť z obrázku č.1, väčšina veveričiek bolo vidieť po obede, iba 11% veveričiek je dospievajúcich. Ďalšie informácie zo spomínaných grafov sú: veveričky sú väčšinou nájdene na zemi, ak už sú veveričky v nejakej výške, tak interkvartálne rozpätie je 5 až 20 metrov, viac ako 90% veveričiek robila nejakú aktivitu. Napriek tomu väčšina si človeka nevšímala.



Obrázok 1

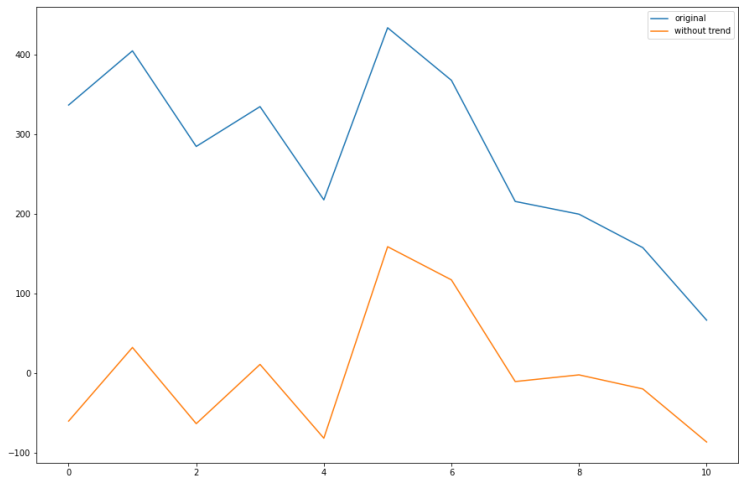
Rudolf Nosek
2DAV

Ďalším krokom bolo nájdenie korelácii atribútov, ktoré sa dali vyjadriť numericky. Jediná premenná, ktorá je otázne prekonvertovaná do numerickej premennej je primárna farba srsti: rozhodli sme sa že svetlejšia farba predstavuje vyššie číslo, to jest čierna 0, šedá 1, škoricová 2. Podlá korelačnej matice z obrázku 2 vieme vytiahnuť dve signifikantnejšie korelácie. Prvá je záporná korelácie medzi výškou veveričky nad zemou a interakcie. To by sme očakávali keďže ťažko môže veverička interagovať s človekom keď je vysoko na strome. Druha je zaujímavejšia. To či veverička bola na zemi a to či veverička bola sparená pred obedom sú tiež záporne korelované. To znamená že veveričky po obede sú častejšie na zemi.



Obrázok 2

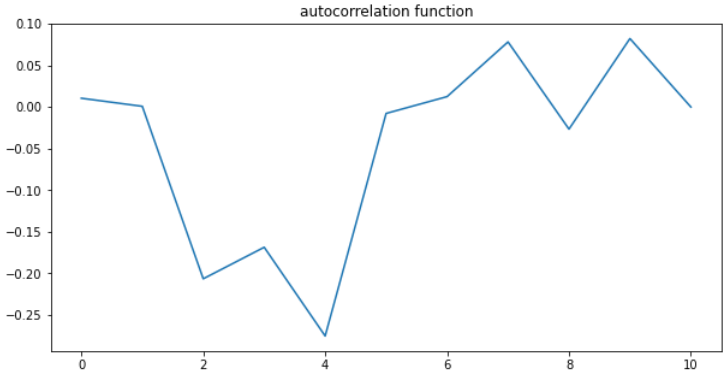
Analýza časového radu počtov nájdených veveričiek je ďalšou našou úlohou. Na obrázku 3 môžete vidieť tento časový rad a ten istý bez zisteného lineárneho trendu cez lineárnu regresiu.



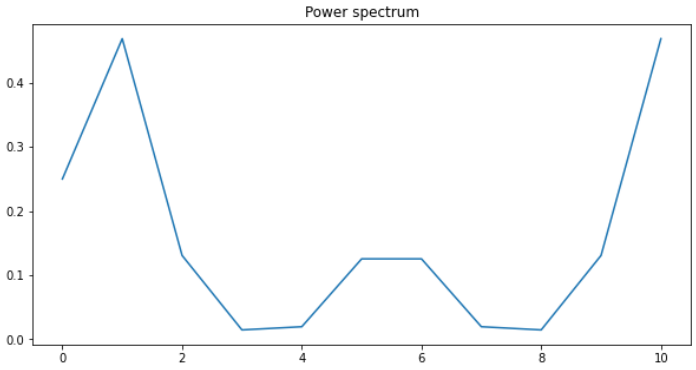
Obrázok 3

Aby sme vedeli predikovať časový rad do budúcnosti budeme potrebovať príslušnú metódu/model. V tomto projekte požijeme tzv. AR model (auto regresný model). Model používa p bodov z minulosti na

vypočítanie ďalšieho bodu v jednoduchej p-dimenzionálnej lineárnej funkcie. Zistenie parametra p sa da cez analýzu tzv. auto korelačnej funkcie(obrázok 4) a jej spektrálneho rozkladu(obrázok 5). Pokúsili sme sa o to ale kvôli nedostatku bodov bol spektrálny rozklad malo informatívny.



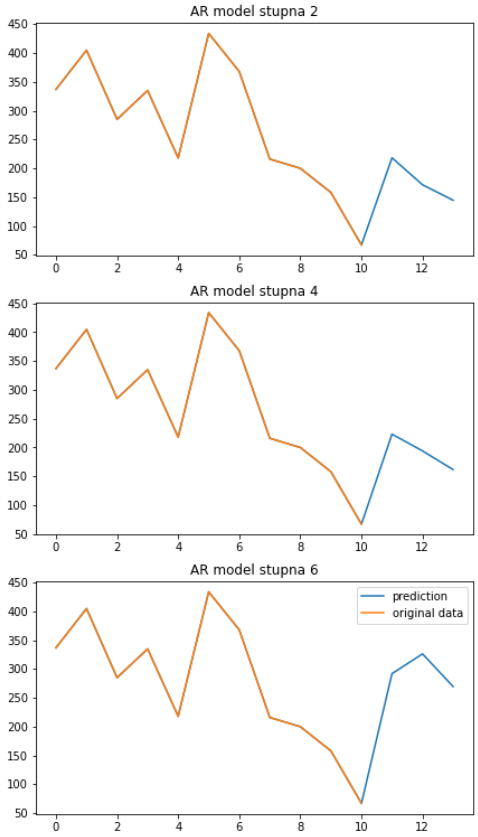
Obrázok 4



Obrázok 5

Preto sme vyskúšali niekoľko AR modelov a na obrázku 6 môžete vidieť akú predikciu by určili.

Rudolf Nosek
2DAV



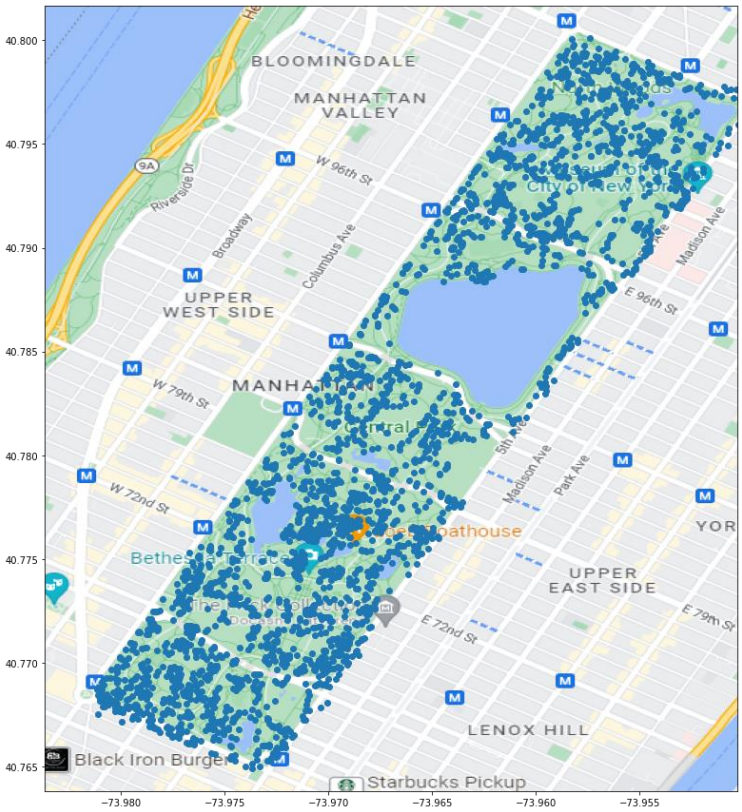
Obrázok 6

Takéto časove rady (avšak už bez snahu o predikciu) sme skonštruovali z iných atribútov. Na obrázku 7 je vidieť že, najväčšia odchýlka z booleovských premenných ma premenná určujúca či bola veverička nájdená pred ci po obede. Je možný lineárny trend v pocte veveričiek na zemi.

Rudolf Nosek
2DAV

Obrázok 7

Predtým ako ukončime analýzu sme ešte vizualizovali výskyty veveričiek na mape (obrázok 7).



Obrázok 8

Rudolf Nosek
2DAV

Posledná časť projektu bolo vytvorené stránky pre naše grafy a tvorba webovej aplikácie pre kvalifikátor veveričiek. Keďže nemáme veľa dát rozhodli sme sa použiť pravdepodobnostný model, konkrétne tzv. naivný baysovský kvalifikátor. Vstup pre model sú atribúty nájdene v korelačnej matici, okrem farby, ktorú sme sa rozhodli predikovať. Model vracia rozdelenie pravdepodobnosti na základe vstupu. Po úspešnom otestovaní modelu a funkčnej stránke sme pokladali projekt za hotoví.

V rámci projektu boli niektoré úlohy náročnejšie: vizualizácia mapy, predikcia časového radu a tvorba kvalifikátora. Jednoduchšie časti projektu bolo spracovanie dát, keďže sme ich mali vo veľmi peknej forme. Spätne by sme odporučili skúsiť hľadať iný ako lineárny trend (polynomialny), využitie zložitejšieho modelu pre predikciu časového radu ako ARMA alebo nechať užívateľa si vybrať predikovanú premennú, možno použiť nejakú vlastnú baysovsku sieť či úplne iný model.