# Извештај за практично истраживање

|  |
| --- |
| **Истраживачи: име, презиме, број индекса (највише четири особе, ред по особи)** |
| Miloš Sirar IN 3/2020 |

|  |
| --- |
| **Шифра истраживања: кодни назив (једна или две речи)** |
| **Vreme\_Šangaj** |

|  |
| --- |
| **Наслов истраживања: сажето исказана суштина истраживања (до 15 речи)** |
| Eksplorativna analiza podataka o vremenskim prilikama (parametrima) i kvalitetu vazduha u Šangaju u periodu 2014-2021 |

|  |
| --- |
| **Предмет, циљ, методи и очекивани резултати истраживања (бар 100 речи)** |
| **Istraživanje** obuhvata učitavanje podataka o vremenskim prilikama i kvalitetu vazduha u Šangaju za naznačeni period, analizu podataka koji su učitani, predstavljanje podataka pomoću tabela i grafikona, traženje povezanosti između različitih faktora i donošenje zaključaka na osnovu analize.  **Cilj** je da se kroz ovo istraživanje stekne šira slika o oblasti analize podataka i zaključivanja na osnovu dobijenih rezultata. Pored toga bitna je i tematika: razumevanje podataka o vremenu i kvalitetu vazduha i njihov međusobni uticaj.  Što se tiče **metoda**, planirano je da se koristi programski jezik Python (i potrebne biblioteke). Koristiće se metode za obradu podataka i tekstualnih fajlova, kao i metode za kreiranje grafikona i tabela.  **Rezultati istaživanja** bi trebalo da pokažu kako su se vreme i vremenski uslovi menjali tokom perioda i njihovu potencijalnu povezanost sa kvalitetom vazduha. Sve ovo će biti predstavljeno grafički da bi se lakše uočile zavisnosti i dali zaključci. |

|  |
| --- |
| **Опис коришћених скупова података (бар 30 речи по скупу података)** |
| Od tri skupa podataka koja su navedena u ideji za praktično istaživanje korišćena su samo dva:  **Skup podataka 2** – 2503 podatka o kvalitetu vazduha tokom perioda 2014-2021, sadrži 7 kolona, podaci su u csv formatu. Izvor: Kaggle  **Skup podataka 3** – 2587 podataka, još dodatnih informacija koje će po potrebi biti uvrštene u istraživanje, sadrži 25 kolona, podaci su u csv formatu. Izvor: Kaggle  Link do svih skupova podataka: https://www.kaggle.com/datasets/erhankul/shanghai-air-pollution-and-wheather-20142021?select=Shanghai+AQI+and+Wheather+2014-2021.csv |

|  |
| --- |
| **Опис обављених активности и њихове расподеле међу истраживачима (бар 150 речи)** |
| **Prikupljanje, pregled i prebacivanje podataka u prikladan oblik za dalji rad**   * Korišćena su dva skupa podataka koja su naznačena * Oni su učitani pod nazivima podaci\_vreme\_sangaj i podaci\_kvalitet\_vazduha\_sangaj * Prvo je odrađen pregled i čišćenje podataka podaci\_vreme\_sangaj * Provera da li postoje nepostojeće vrednosti * Promenjeni su i prilagođeni nazivi kolona na srpski jezik * Uklonjene su kolone koje nisu potrebne za ovo istraživanje, odnosno one čiji podaci neće biti parametri istraživanja * Promenjeni su tipovi kolona na tipove koji su prikladniji (datum) * Dodate su nove kolone radi lakše dalje manipulacije i radom podacima (godina\_merenja, mesec\_merenja, dan\_merenja) * Zatim je odrađen postupak pregleda i čišćenja podataka i u podaci\_kvalitet\_vazduha\_sangaj * Promenjeni su i prilagođeni nazivi kolona na srpski jezik * Promenjeni su tipovi kolona (datum i svi parametri za kvalitet vazduha u tip float) * Ovde nisu uklonjene nepostojeće vrednosti, već su zamenjene sa vrednošću -1.0, koja kasnije neće imati uticaj na podatke, a ovi podaci mogu da se obrađuju iako ne postoje svi parametri kvaliteta vazduha * Kada su oba skupa podataka dovedena u pogodan oblik za dalji rad, oni se spajaju u jedan skup podataka. Taj skup podataka ima naziv podaci, i on se koristi u nastavku istraživanja kao glavni skup podataka   **Početna istraživanja**   * Na početku se gleda broj podataka za svih 8 godina u periodu od 2014. do 2021. godine * Uočava se da za 2021. godinu ne postoji relevantan broj podataka i on se isključuje iz daljeg toka istraživanja * Celokupno istraživanje biće rađeno za 7 godina, 2014-2020 * Zatim se prelazi na izračunavanje AQI (Air Quality Index) parametra na osnovu parametara o kvalitetu vazduha, tj. podataka o zastupljenosti štetnih čestica i gasova u vazduhu * Čestice koje se gledaju su PM 2.5 i PM 10, dok gasovi koji se gledaju su O3 (ozon), NO2 (azot dioksid), SO2 (sumpor dioksid) i CO (ugljen monoksid) * Za to je potrebno da se izračunaju sub indeksi na osnovu standardizovanih formula, svaki parametar ima svoje formule * Kada se to izračuna, moguće je izračunati i AQI indeks * To se radi tako što se uzme maksimalna vrednost jednog od sub indeksa koji je izračunat * U zavisnosti od vrednosti AQI indeksa se dodeljuje i tekstualni opis kakvo je stanje vazduha, tj. kakav je kvalitet vazduha * Svi podaci o sub indeksima, AQI indeksima i AQI objašnjenje su dodati u skup podataka podaci   **Početak analize podataka**   * Postoji nekoliko različitih analiza koje su odrađene i sada će svaka biti ukratko opisana   Analiza 1 i 2   * Ove dve analize služe za prikaz promene prosečnih maksimalnih temperatura, odnosno prosečnih minimalnih temperatura tokom godina. Ovom analizom sam primetio da se od 2014. godine svake naredne godine povećava prosečna maksimalna (minimalna) temperatura. Prikazano je i povećanja za svaku godinu i na kraju se došlo do toga da je u 2020. godine prosečna maksimalna temperatura povećana u odnosu na 2014. za 1.73 °Cdok se prosečna minimalna temperatura povećala za 2.59 °C.   Analiza 3, 4, 5 i 6   * Ove četiri analize su međusobno povezane. Cilj analize 3 i 4 je da prikažu prosečne maksimalne (minimalne) temperature po godinama za svaki mesec. Tu je primećeno da se u julu mesecu kroz godine smanjuje maksimalna prosečna temperatura, a da se u avgustu i semptembru povećava. Zbog toga su urađene analize 5 i 6, koje prikazuju prosečne maksimalne (minimalne) temperature po mesecima za svaku godinu, kako bi se videla još bolje ta promena za mesece jul, avgust i septembar.   Analiza 7 i 8   * One se nadovezuju na analize 3, 4, 5 i 6 tako što prikazuju prosečne maksimalne temperature za mesece avgust i semptembar, gde se vidi kako se kroz svaku godinu menjaju dnevne temperature u ovim mesecima.   Analiza 9   * Ovde je prikazana ukupna količina padavina mm u svakoj od godina koje se posmatraju u istraživanju. U 2019. i 2020. godini je došlo do povećanja količine padavina.   Analiza 10   * U ovoj analizi je prikazana prosečna vidljivost za sve godine.   Analiza 11   * Ova analiza se nastavlja na analizu 10, odnosno posmatra isti parametar – vidljivost. Ovde se prikazuju oni podaci kod kojih je vidljivost ispod 6. U nastavku ove analize se traži neka povezanost sa oblačnošću, pošto je uviđeno da je oblačnost prilično velika za ove vrednosti. Došlo se do zaključka da je vidljivost u slaboj negativnoj korelaciji sa oblačnošću.   Analiza 12   * Ovde se prikazuje AQI indeks, odnosno njegovo objašnjenje ukupno za ceo period.   Analiza 13   * Prikazuje se AQI indeks, odnosno njegovo objašnjenje za svaku od godina.   Analiza 14   * U ovoj analizi se proverava korelacija između AQI indeksa i različitih vremenskih parametara. Ti vremenski parametri su: maksimalna dnevna temperatura, minimalna dnevna temperatura, padavine mm, vidljivost, brzina vetra km/h, oblačnost, pravac vetra. Na kraju ovog istraživanja se došlo do zaključka da AQI indeks nije u korelaciji ni sa jednim od ovih parametara. Postoji slaba korelacija između nekih parametara za neke godine, ali ukupno za ceo period ili da ima neki kontinuitet ne postoji..   Deskriptivna statistika   * Za kraj je odrađena deskriptivna statistika nad podacima, odnosno nad vrednostima nad kojima to ima smisla. Prikazani su count, mean, std, min, 25%, 50%, 75%, max za podatke.   **Zaključak**  Nakon svih analiza i istraživanja došlo se do nekoliko zaključaka:   * Dolazi do povećanja prosečne maksimalne i minimalne temperature kroz godine * Najveće temperature se pomeraju sa jula na avgust * Septembar ima značajno povećanje temperature * Postoji slaba negativna korelacija između vidljivosti i oblačnosti * Prikazani su AQI indeksi pomoću kojih se došlo do informacija koliko je zagađen vazduh * Pokazano je da ne postoje značajne korelacije između parametara vremenskih prilika i AQI indeksa   Šta bi mogli da budu faktori zagađenja vazduha:   * Sagorevanje ulja * Motorna vozila * Industrijska prašina * Hemijski otpadi koji odlaze u atmosferu   Šta se radi na tome da se ovo poboljša:   * Šangaj postaje zeleniji grad * Veliki broj novih parkova * Smanjenje emisije ugljenika * ... |

|  |
| --- |
| **Кратак опис претходних истраживања која су изведена над коришћеним скуповима података од стране других особа и приказ главних разлика између садашњег и претходних истраживања (бар 70 речи)** |
| Nad ovim skupom podataka je jedan autor radio takođe istraživanje. Koliko sam shvatio, njegovo istraživanje se fokusiralo na aspekt povezanosti AQI parametra i nekih bolesti. Sličnosti između istraživanja jesu to što je u oba rada rađeno računanje AQI indeksa, ali ostatak je generalno različit. Što se tiče nedostajućih vrednosti, ja sam ih brisao, odnosno menjao sa -1.0, dok je autor radio njihovu dopunu sa srednjom vrednošću. Nastavak je skroz različit, jer smo ispitivali podatke za totalno različite domene. Autor je takođe kasnije radio predikcije i pravio je model, što u mom istraživanju nema, jer to nije bio cilj istraživanja. |

|  |
| --- |
| **Анализа испуњења циља и остварења очекиваних резултата (бар 30 речи)** |
| Uspeo sam da uradim obradu i analizu podataka i njihovo predstavljanje putem grafikona, teksta ili tabela. Naučio sam više o vremenskim parametrima, njihovim značenjima i o parametrima koji određuju kvalitet vazduha. Pokušao sam da pronađem neke zavisnosti između različitih parametara, uglavnom se ispostavilo da oni nisu previše međusobno povezani, ali je bilo slučajeva gde se vidi jasna povezanost. Rezultati su bili uglavnom očekivani, jer zagađenost vazduha ipak najviše zavisi od ljudskog faktora, odnosno zagađenja prirode. |

|  |
| --- |
| **Могућности за примену истраживања у пракси (бар 30 речи)** |
| Ovo istraživanje bi bilo primenjivo u praksi ukoliko bi se baš svi parametri detaljno proučili i našla baš svaka zavisnost između svakog pojedinačnog obeležja. Za to je potrebno malo više domenskog znanja. Takođe, podaci na nivou jednog sata, a ne na nivou jednog dana bi sigurno bili još više reprezentativniji i mogle bi se uočiti neke zavisnosti koje se trenutno ne vide. |

|  |
| --- |
| **Идеје за побољшање истраживања (бар 30 речи)** |
| Da bi se poboljšalo istraživanje trebao bi veći skup podataka, tačnije češći skup podataka kao što sam napisao. Verovatno bi bilo bolje da imam i podatke sa više različitih tačaka merenja iz jednog grada, a ne samo sa jedne tačke, pošto je površina grada dosta velika. Veće domensko znanje. Veće praktično znanje iz statistike kako bi se mogle uočiti neke zavisnosti, odnosno analiza i testovi koji bi se trebalo primeniti nad podacima. |