



## Zadání diplomové práce

<b>Název:</b>	STK portál
<b>Student:</b>	Bc. Daniel Brotz
<b>Vedoucí:</b>	Mgr. Adam Szabó
<b>Studijní program:</b>	Informatika
<b>Obor / specializace:</b>	Znalostní inženýrství
<b>Katedra:</b>	Katedra aplikované matematiky
<b>Platnost zadání:</b>	do konce letního semestru 2023/2024

### Pokyny pro vypracování

Cílem této práce je vytvořit portál pomáhající lidem vybrat si vozidlo. Na portále bude možné dohledat proběhlé prohlídky zadaného vozidla, závady, které se na STK objevily, a součástí budou také detailní statistiky s možností porovnání různých typů vozidel.

Proveďte rešerši stávajících řešení (např. <https://www.cebia.cz>, <https://vindecoder.eu/cz/>, <https://www.carvertical.com/cz>), seznamte se s poskytnutými daty, proveďte rešerši vhodných metod pro hlubší analýzu dat z STK. Na základě rešerše budou pak vybrány a implementovány konkrétní metody, pokud je třeba, navrhněte nutné transformace. Implementujte model, který na základě typu auta, jeho stáří a historie bude predikovat závady, které se na něm mohou v blízké době vyskytnout. Navrhněte vhodné metriky a analýzy a též formu prezentace výsledků na webovém portále. Ten následně implementujte a otestujte.

Vstupem budou XML data za období 2018-2022, které poskytne vedoucí práce.





**FAKULTA  
INFORMAČNÍCH  
TECHNOLOGIÍ  
ČVUT V PRAZE**

Diplomová práce

**STK portál**

*Bc. Daniel Brotz*

Katedra aplikované matematiky  
Vedoucí práce: Mgr. Adam Szabó

9. května 2024



---

## **Poděkování**

Chtěl bych poděkovat v prvé řadě Mgr. Adamovi Szabó za motivující vedení při vývoji této práce a stejně tak Mgr. Martinu Marešovi za nesčetné hodiny konzultací, které mi oba věnovali. Rád bych poděkoval také své rodině a přátelům jak za uživatelské testování a cenné podněty k práci, tak za jejich podporu po celou dobu mého studia.



---

## Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisu. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisu, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 9. května 2024

.....

České vysoké učení technické v Praze  
Fakulta informačních technologií

© 2024 Daniel Brotz. Všechna práva vyhrazena.

*Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí a nad rámec oprávnění uvedených v Prohlášení na předchozí straně, je nezbytný souhlas autora.*

### Odkaz na tuto práci

Brotz, Daniel. *STK portál*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2024.

---

# Abstrakt

Práce se zabývá vytěžováním znalostí z dat o kontrolách vozidel na stanicích technické kontroly (STK) a registru vozidel České republiky. V úvodu je provedena rešerše komerčních služeb pro kontrolu stavu ojetých vozů Cebia, Vindecoder a Carvertical. Analyzováno je také několik portálů provozovaných většinou Ministerstvem dopravy ČR, které zdarma poskytují různá data o vozidlech. Na základě rešerše jsou navrženy a implementovány metody, které z uvedených datových zdrojů získávají statistiky např. o spolehlivosti vozidel různých značek či anomálních prohlídkách na STK. Použity jsou také metody strojového učení pro výstupy jako je predikce nájezdu či závad na konkrétních vozech. Výsledky jsou prezentovány ve webové aplikaci *STK portál*, která je zpřístupňuje široké veřejnosti a nabízí tak bohatý přehled o vozovém parku ČR i bezplatnou alternativu ke komerčním službám kontroly ojetých vozidel.

**Klíčová slova** STK, registr vozidel, technické kontroly, webová aplikace, datová analýza, detekce anomalií, Cebia, Vindecoder, Carvertical, Datová kostka, Kontrola tachometru

---

# Abstract

This thesis focuses on extracting knowledge from records on vehicle inspections at technical inspection stations (STK) and the vehicle registry of the Czech Republic. In the introduction, a research of commercial used vehicle health check services such as Cebia, Vindecoder and Carvertical is conducted. Several web portals, mostly operated by the Ministry of Transport of the Czech Republic, which provide various vehicle data for free are also analyzed. Based on the research, methods are proposed and implemented to extract statistics from the mentioned data sources, e.g., on the reliability of vehicles of different brands or anomalous inspections at the STK. Machine learning methods are also used for outputs such as predicting mileage or defects on specific vehicles. The results are presented in the *STK portal* web application, which makes them available to the general public, offering a comprehensive overview of the Czech vehicle fleet as well as a free alternative to commercial vehicle history checking services.

**Keywords** vehicle technical inspections, vehicle registry, web application, data analysis, machine learning, anomaly detection, Cebia, Vindecoder, Carvertical, Datová kostka, Kontrola tachometru

---

# Obsah

<b>Úvod</b>	<b>1</b>
<b>1 Cíle práce</b>	<b>3</b>
<b>I Rešerše a návrh řešení</b>	<b>5</b>
<b>2 Dosavadní situace</b>	<b>7</b>
2.1 Standard VIN . . . . .	7
2.2 Technické kontroly vozidel v ČR . . . . .	8
2.3 Původní STK portál . . . . .	9
2.3.1 Datové zdroje . . . . .	9
2.3.2 Použité metody . . . . .	10
2.3.3 Webová prezentace . . . . .	11
2.4 Služby pro spotřebitele . . . . .	13
2.4.1 Cebia . . . . .	13
2.4.2 Vindecoder . . . . .	15
2.4.3 Carvertical . . . . .	17
2.5 Datové portály . . . . .	19
2.5.1 Datová kostka . . . . .	19
2.5.2 Kontrola tachometru . . . . .	20
2.5.3 Svaz dovozců automobilů . . . . .	21
2.6 Výsledky rešerše . . . . .	22
<b>3 Datové zdroje</b>	<b>25</b>
3.1 Prohlídky na STK . . . . .	25
3.2 Registr vozidel . . . . .	29
3.3 Seznam stanic . . . . .	33
3.4 Číselník závad . . . . .	33

<b>4 Metody zpracování dat</b>	<b>35</b>
4.1 Klasifikace a regrese . . . . .	35
4.2 Shlukování . . . . .	37
4.3 Detekce odlehlých hodnot . . . . .	38
4.4 Modelování časových řad . . . . .	39
<b>5 Návrh řešení</b>	<b>41</b>
5.1 Webový portál a cíle analýzy . . . . .	41
5.1.1 Úvodní stránka . . . . .	41
5.1.2 Stanice technické kontroly . . . . .	41
5.1.3 Vozidla . . . . .	44
5.1.4 Srovnávač vozidel . . . . .	45
5.1.5 Ostatní požadavky . . . . .	46
5.2 Architektonický návrh . . . . .	46
5.2.1 Datový modul . . . . .	47
5.2.2 Databáze a API . . . . .	47
5.2.3 Webová aplikace . . . . .	48
<b>II Implementace</b>	<b>49</b>
<b>6 Datová analýza a návrh modelů</b>	<b>51</b>
6.1 Předzpracování dat . . . . .	51
6.1.1 Prohlídky na STK . . . . .	51
6.1.2 Registr vozidel . . . . .	52
6.1.3 Seznam stanic . . . . .	53
6.1.4 Číselník závad . . . . .	53
6.2 Predikce závad . . . . .	54
6.3 Predikce vytíženosti stanic . . . . .	55
6.4 Predikce nájezdu vozidel . . . . .	58
6.5 Detekce anomálních kontrol . . . . .	59
6.6 Ostatní výsledky . . . . .	60
<b>7 Výsledky analýzy</b>	<b>63</b>
7.1 Predikce závad . . . . .	63
7.2 Predikce vytíženosti stanic . . . . .	64
7.3 Predikce nájezdu vozidel . . . . .	67
7.4 Detekce anomálních kontrol . . . . .	67
7.5 Shrnutí výsledků . . . . .	68
<b>8 Softwarový projekt</b>	<b>71</b>
8.1 Infrastruktura . . . . .	71
8.2 Pipeline . . . . .	72
8.3 Webová aplikace . . . . .	73

<b>9 Závěr</b>	<b>77</b>
<b>Literatura</b>	<b>79</b>
<b>A Seznam použitých zkratek</b>	<b>85</b>
<b>B Obsah digitální přílohy</b>	<b>87</b>



---

# Seznam obrázků

2.1	Výřez z detailu stanice na původním STK portálu. . . . .	12
2.2	Výsledky kontrol podle kraje na původním STK portálu. . . . .	12
2.3	Výřez prověření vozidla na portálu Cebia. . . . .	14
2.4	Historie nájezdu vozidla na portálu Cebia. . . . .	15
2.5	Výřez náhledu před zakoupením reportu od Vindecoder. . . . .	16
2.6	Úvod analýzy od Carvertical. . . . .	17
2.7	Historie nájezdu vozidla v analýze Carvertical. . . . .	18
2.8	Výřez zobrazení dat na portálu Datová kostka. . . . .	20
2.9	Detail prohlídky na webu Kontrola tachometru. . . . .	21
2.10	Graf vývoje registrací nových vozidel podle SDA. . . . .	22
3.1	Počet VIN podle četnosti jejich výskytu v kontrolách. . . . .	27
3.2	Počet VIN podle délky kódu. . . . .	28
3.3	Vybrané varianty zápisu modelu Škoda Octavia, červen 2021. . . . .	28
5.1	Návrh úvodní stránky webu. . . . .	42
5.2	Návrh stránky stanic a jejího detailu. . . . .	43
5.3	Návrh stránky vozidel a detailu konkretního vozidla. . . . .	45
6.1	Vývoj počtu prohlídek na stanici 3102. . . . .	56
6.2	Detail vývoje počtu prohlídek na stanici 3102. . . . .	56
6.3	Autokorelační grafy četnosti prohlídek na stanici 3102. . . . .	57
6.4	Analýza reziduí predikce SARIMA na stanici 3102. . . . .	58
6.5	Příklad výsledku shlukování pro analýzu flotil. . . . .	61
7.1	Predikce závad v sekci detailu vozidla. . . . .	64
7.2	Predikce vytíženosti stanice 3102 různými modely. . . . .	65
7.3	Vizualizace predikce vytíženosti stanice na webu. . . . .	66
7.4	Predikce nájezdu v sekci detailu vozidla. . . . .	67
7.5	Histogram podílu všech anomálních prohlídek na celkovém počtu prohlídek. . . . .	68

7.6	Histogram počtů opakování prohlídky s úspěchem na jiné stanici. . .	68
8.1	Diagram struktury Docker compose projektu. . . . .	71
8.2	Diagram struktury pipeline modulu. . . . .	72
8.3	Diagram struktury podmodulu pro import dat v pipeline. . . . .	73
8.4	Výřez sekce o stanicích na webu. . . . .	74
8.5	Detail analýzy podílů barev nově registrovaných vozidel na webu. .	75
8.6	Graf stavu vozidel podle data registrace na webu. . . . .	75

---

# **Seznam tabulek**

4.1	Matice záměn.	37
6.1	Rozsahy vyzkoušených parametrů při tréninku modelu predikce závad.	55
6.2	Rozsahy vyzkoušených parametrů při tréninku modelu predikce nájezdu.	59
7.1	Evaluace predikce závad.	63
7.2	Evaluace predikce vytíženosti stanic.	66
7.3	Seznam výsledků analýzy.	70



---

# Úvod

Průměrný věk vozidla v České republice v posledních letech trvale stoupá, vozový park se časem obměnuje a s tím i preference spotřebitelů [1]. Každý, kdo uvažuje o nákupu ať už nového či ojetého vozidla, ale hledá především spolehlivý vůz, který by zároveň odpovídal jeho osobním požadavkům. V reakci na tuto poptávku po informacích o vozidlech proto vzniklo množství webových služeb, které nabízí různě detailní a většinou zpoplatněná data o historii vozidla, pojistných událostech, nájezdu i o technických parametrech. Cílem této diplomové práce je proto nabídnout témtoto službám alternativu, která by využila otevřená data a zpřístupnila tak veřejnosti zdarma co největší množství informací jak o vozovém parku obecně, tak o konkrétních vozidlech.

Diplomová práce navazuje opět pod záštitou laboratoře OpenDataLab na práci Aleksandry Parkhomenko *Portál výsledků analýzy dat a dalších informací o STK* [2]. Výstupem této práce byl webový portál zabývající se analýzou prohlídek na stanicích technické kontroly (STK) v roce 2018. Nově vyvinutý nástupce tohoto portálu umožní periodicky doplňovat data z STK a zároveň zahrnout i další datové sady za účelem provedení výrazně rozsáhlejší analýzy českého vozového parku. Přidanými datovými sadami budou anonymizovaný výtah z registru vozidel, informace odvozené ze seznamu STK a číselník závad odhalovaných na technických kontrolách.

Ministerstvo dopravy (dále také MDČR) poskytuje prostřednictvím několika webových portálů velké množství otevřených dat a na základě zákona 106/1999 Sb., o svobodném přístupu k informacím, lze obdržet ještě větší objem. Využití všech těchto informací je ale podmíněno nutností data různými cestami získat, začítit je do strukturované a konzistentní podoby a připravit k dalšímu zpracování. Teprve následně lze navrhnout konkrétní metody a vytvořit vizualizace, které odhalí skrytý potenciál dat. Využití začíná získáním statistik o vývoji popularity jednotlivých značek a modelů v čase, pohledem na elektrifikaci vozového parku nebo podílu importovaných ojetých vozů na celkovém objemu nově registrovaných vozidel. Zapojením modelů strojového

## ÚVOD

---

učení lze predikovat například výskyt poruch či změnu nájezdu na konkrétních vozidlech a spojením všech těchto výsledků můžeme srovnávat jak konkrétní vozidla, tak celé modelové řady navzájem a poskytnout tak cenné informace spotřebitelům.

Diplomová práce sleduje tento postup a v úvodní části proto mapuje několik stávajících, často komerčních portálů zaměřených zejména na spotřebitele. Následující kapitoly detailně popisují získané datové sady a jejich rozličné vady pramenící jednak poměrně nevyhnutelně z podstaty jejich vzniku, jednak z okolností jejich uchovávání v prostředí státní správy. Na základě dostupných dat jsou poté navrženy výsledky, které je jejich analýzou možné vyvodit, a jsou popsány konkrétní výpočetní metody. Navazuje návrh webového portálu pro ucelenou prezentaci výsledků jakož i infrastruktury pro zpracování a uchovávání dat.

Druhá část práce se věnuje implementaci navrženého systému. Zabývá se nejprve předzpracováním dat, aplikacemi popsaných metod a evaluací získaných výsledků. Nakonec pak popisuje okolnosti vývoje webového portálu a sestavení všech softwarových součástí do celku připraveného k nasazení.

# KAPITOLA **1**

---

## Cíle práce

Zastřešujícím cílem práce je tvorba webového portálu, který nabídne návštěvníkům jednak informace pro podporu rozhodování při koupi vozidla, jednak detailní statistiky o stavu českého vozového parku. K dosažení tohoto cíle je zapotřebí seznámit se s aktuální situací a dostupnými daty, na základě této rešerše navrhnout cíle analýzy a ty poté na webu vizualizovat.

Úvodním cílem je proto uvedení problematiky technických kontrol vozidel v ČR a analýza několika webových komerčních služeb i bezplatných datových portálů souvisejících s tímto tématem.

Druhý cíl spočívá v seznámení s datovými zdroji a jejich předzpracováním. To sestává z transformací, které očistí data od zjevně chybných záznamů, doplní ve vhodných případech chybějící údaje a uloží výsledek do databáze. Tento podsystém pro příjem surových dat musí podporovat jejich postupné periodické doplňování, např. při aktualizaci registru vozidel či získání nových záznamů kontrol na STK.

Následuje datová analýza, která se skládá z cílů se dvěma odlišnými úrovněmi složitosti. Naplnění jednoduších cílů analýzy spočívá ve výpočtech základních statistik a v agregacích či transformacích tabulek s daty. Příkladem je analýza vývoje popularity různých značek automobilů v čase podle počtu nově registrovaných vozů dané značky v každém roce. Složitější cíle vyžadují aplikaci metod strojového učení či rozhodování na základě empiricky odvozených pravidel – patří sem

- predikce závad na konkrétním vozidle;
- predikce nájezdu jednotlivých vozidel;
- detekce anomálních prohlídek na STK;
- predikce vytíženosti stanic.

## **1. CÍLE PRÁCE**

---

Poslední cíl, tj. vizualizace je naplněn implementací webového portálu, který musí být dobře přístupný i z mobilních zařízení pro možnost okamžité kontroly vozidla. Obsah portálu se bude odvíjet od výsledků rešerše a dostupných dat, mezi jeho základní sekce budou patřit

- statistiky týkající se vozového parku obecně;
- statistiky o prohlídkách vozidel na STK;
- detaily konkrétních vozidel včetně historie jejich prohlídek, závad a predikcí;
- detaily jednotlivých stanic a analýza chování na nich;
- srovnání konkrétních vozidel i celých modelových řad.

Na základě otevřených dat tak vznikne ucelený webový portál orientovaný na spotřebitele i zájemce z řad široké veřejnosti, jenž poskytne jak grafickou prezentaci vystihující stav českého vozového parku, tak i detailní pohled na každé vozidlo.

# Část I

## Rešerše a návrh řešení



# KAPITOLA **2**

---

## **Dosavadní situace**

Tato kapitola obsahuje úvod do problematiky STK a analýzy vozidel. Začíná základní terminologií a popisem užívaných standardů. Následně se věnuje původnímu STK portálu, na nějž tato práce navazuje. Nakonec analyzuje vybrané současné služby pro spotřebitele zabývající se ověřováním ojetých vozů a také portály poskytující otevřená data.

### **2.1 Standard VIN**

**VIN** (vehicle identification number) je unikátní identifikátor motorového vozidla definovaný normou ISO 3779. Jedná se o alfanumerický kód o 17 znacích rozdelený na 3 části, které identifikují výrobce, popis vozidla a sériové číslo. Aby se předešlo záměně znaků při zhoršení čitelnosti kódu (bývá vyražen do kovových částí vozidla, které např. mohou zkrodotovat), VIN nesmí obsahovat písmena O, I a Q. Popis jednotlivých částí, jak jsou uvedeny níže, je definován prováděcím nařízením Evropské komise 2021/535 [3].

**WMI** (world manufacturer identifier) je tvořen třemi znaky identifikujícími výrobce. Kód přiděluje stát, ve kterém se nachází hlavní místo podnikání výrobce. Pro výrobce produkující méně než 500 vozidel ročně je posledním znakem vždy 9.

**VDS** (vehicle description section) označuje pěti znaky obecné vlastnosti vozidla. Typicky se jedná o model, motorizaci, typ karoserie či výbavu [4].

**VIS** (vehicle indicator section) tvoří osm znaků, z nichž poslední čtyři jsou vždy číslice. Tato sekce slouží ve spojení s ostatními částmi VIN k identifikaci konkrétního vozidla. Nevyužitá místa jsou vyplňena číslicí 0.

## 2. DOSAVADNÍ SITUACE

---

### 2.2 Technické kontroly vozidel v ČR

Podle § 40 zákona č. 56/2001 Sb., o podmínkách provozu vozidel na pozemních komunikacích musí osobní automobily absolvovat kontrolu technického stavu každé 2 roky, s výjimkou nových vozů, které absolvují první kontrolu po 4 letech od registrace [5]. Pro nákladní automobily a autobusy platí povinnost kontroly každý rok, naopak u motocyklů stačí prohlídka vždy po 4 letech. Tímto opatřením je zajištěna bezpečnost provozovaných vozidel jak pro jejich posádku, tak pro okolí. Prohlídky se provádějí na stanicích technické kontroly (STK), které musí pro svůj provoz splnit požadavky určené § 16 vyhlášky č. 211/2018 Sb. o technických prohlídkách vozidel.

Prohlídek je několik typů podle okolností jejich provádění a rozsahu, za nejdůležitější pro účely této práce lze považovat následující:

- pravidelná technická prohlídka;
- opakování technická prohlídka – do 30 po předchozí prohlídce, kde byly zjištěny vážné nebo nebezpečné závady;
- technická prohlídka silničních vozidel určených k přepravě nebezpečných věcí;
- technická prohlídka před schválením technické způsobilosti vozidla;
- technická prohlídka před registrací;
- technická prohlídka na žádost zákazníka;
- technická silniční kontrola – prováděná mimo STK pomocí mobilní kontrolní jednotky.

Kontrolní úkony určuje příloha vyhlášky č. 211/2018 Sb. a úkony se dělí do následujících kategorií [6]:

- identifikace vozidla;
- brzdové zařízení;
- řízení;
- výhledy;
- svítily, světlomety, odrazky a elektrické zařízení;
- nápravy, kola, pneumatiky a zavěšení náprav;
- podvozek a části připevněné k podvozku;
- jiné vybavení (pásy, zámky, bezpečnost, elektronické systémy);

- obtěžovaní okolí;
- další prohlídky vozidel k dopravě osob kategorie M2 a M3.

Zjištěné závady se dělí do tří kategorií dle závažnosti, kategorie definuje § 49 zákona 56/2001 Sb. [5] následovně:

- **lehké (A)**: nemají významný vliv na provozní vlastnosti vozidla, bezpečnost provozu na pozemních komunikacích ani životní prostředí;
- **vážné (B)**: ovlivňují provozní vlastnosti vozidla, jsou způsobilá ohrozit provoz na pozemních komunikacích, můžou nepříznivě působit na životní prostředí nebo spočívají ve vážném nedostatku v identifikaci vozidla;
- **nebezpečné (C)**: bezprostředně ohrožují bezpečnost jízdy silničního vozidla, provoz na pozemních komunikacích nebo životní prostředí.

Následující odstavce dále stanovují, že lehké závady nebrání v provozu vozidla, ale provozovatel je povinen je odstranit. V případě zjištění vážných závad je nutné je odstranit a do 30 dnů absolvovat opakovanou kontrolu. Nebezpečná závada pak znamená povinnost vozidlo odtáhnout, závadu odstranit a následně vozidlo opět přistavit k opakované prohlídce – vozidlo se tedy nesmí pohybovat po vlastní ose na veřejných komunikacích. Je-li zjištěna závada kategorie B, resp. C, vozidlo je „částečně způsobilé“, resp. „nezpůsobilé“. Nejsou-li zjištěny žádné závady či jen závady kategorie A, výsledek se označuje „způsobilé“.

Součástí technické kontroly je také zaznamenání stavu počítače kilometrů (tj. nájezdu) vozidla a fotografická dokumentace vozidla ze všech stran, včetně detailních fotografií VIN. Hlavním obsahem je kontrola stavu vozidla, všechny zjištěné závady se zapisují do protokolu o kontrole. Všechny tyto informace jsou pak shromažďovány v databázi, k níž kontrolní technici přistupují přes internet. [5]

## 2.3 Původní STK portál

Jak bylo zmíněno v úvodu, tato práce navazuje na STK portál Aleksandry Parkhomenko, která se ve své bakalářské práci [2] věnovala analýze dat o kontrolách vozidel na STK za rok 2018. Následující popis vychází jednak z textu této práce, jednak přímo z webových stránek, kde je původní portál dostupný, tj. <https://stk-test.opendatalab.cz>.

### 2.3.1 Datové zdroje

Klíčovou datovou sadou je seznam všech kontrol vozidel, které v roce 2018 proběhly na stanicích technické kontroly v ČR [7]. Jedná se o otevřená data

## 2. DOSAVADNÍ SITUACE

---

dostupná v Národním katalogu otevřených dat [8], která nejsou chráněna autorským právem a mohou tedy být volně využívána.

Data jsou dostupná jako jediný soubor ve formátu XML obsahující téměř 4 miliony záznamů. Záznam obsahuje informace o stanici (její číselný kód), datu a čase kontroly, výsledku technické kontroly i měření emisí, dále některé údaje o stavu vozidla (nájezd a počty zjištěných závady dle závažnosti) a základní parametry vozidla jako jeho kategorie, tovární značku a model či VIN.

Podpůrnou datovou sadu tvoří číselník stanic technické kontroly dostupný jako tabulka ve formátu Microsoft Excel na webu Ministerstva dopravy ČR [9]. Tento číselník řazený podle krajů obsahuje detaily ke každé stanici definované jejím kódem a umožňuje je tak spojit s datovou sadou o kontrolách. Mimo název stanice a strukturované adresy obsahuje kontaktní údaje a rozsah oprávnění (které kategorie vozidel může stanice kontrolovat).

Pro potřeby datové analýzy je doplněn ještě jeden zdroj, kterým jsou statistiky pro výpočet kapacit kontrolních linek na stanicích, rovněž publikovaný MDČR [10].

### 2.3.2 Použité metody

Autorčina práce s daty spočívá jednak v základní exploratorní analýze, kdy jsou představuje jednotlivé sloupce datové sady o kontrolách a prezentuje statistiky jako počet chybějících hodnot, průměrné hodnoty či vzájemné korelace.

Následuje analýza shluků za použití algoritmu K-means, kdy je nutné data předpřipravit jednak odebráním některých sloupců, jednak konverzí hodnot jiných sloupců do číselné podoby. Autorka poté data normalizuje a redukuje dimenzionalitu výsledku pomocí PCA, výsledkem je rozdělení do tří shluků lišících se především značkami zahrnutých vozidel, na základě čehož prohlašuje data za nevhodná pro shlukovou analýzu.

Na základě pravidel pocházejících z expertní znalosti problematiky je provedena analýza podezřelého chování na STK. Jelikož data obsahují i údaj o přesném čase provedení prohlídky, je možné vytěžit z webu data o provozní době jednotlivých stanic a porovnat je s časovou značkou prohlídky. Závěrem tohoto šetření je, že některé automaticky vytěžené údaje o provozní době byly nepravdivé a musely být ručně opraveny. Většina zbylých kontrol proběhla do 60 minut kolem pracovní doby a téměř všechny zkoumané stanice nějakou takovou kontrolu provedly.

Pro vybrané stanice autorka zjišťuje počet jejich kontrolních linek a kombinací těchto informací s odhady pracnosti kontrol zveřejněnými MDČR odhaluje anomálně krátké trvání kontrol na jedné ze stanic. Jelikož počet kontrolních linek není dostupný např. v rámci číselníku STK, není tato analýza proveditelná automatizovaně pro všechny stanice.

Za anomální jsou považována také výrazně častější setkání vozidel stejné značky na kontrole, než lze odhadnout z celkových počtů vozidel jednotlivých značek v datasetu. Autorka vypočítává odhad počtu setkání vozidel každých

dvou značek a po vyfiltrování pouze osobních automobilů dochází ke zjištění, že na některých stanicích se určité dvojice značek potkávají i dvakrát častěji, než bylo odhadnuto.

Završení datové analýzy je tvořeno algoritickou detekcí anomalií za použití shlukovací metody DBSCAN. Tato metoda, která počet shluků určuje automaticky a nikoliv na základě přímého parametru, odděluje vozidla zejména na základě jejich výsledků technické a emisní kontroly, při druhé konfiguraci parametrů pak data rozděluje podle typu prohlídky (pravidelná, opakována apod.) na časté a řídce se objevující typy.

#### 2.3.3 Webová prezentace

Webový portál je dostupný na adrese <https://stk-test.opendatalab.cz>. Úvodní stránka obsahuje odkaz na datovou sadu v NKOD. Ústředním prvkem je zde mapa krajů ČR, která je ale pouhým obrázkem a po kliknutí na libovolný kraj nenásleduje žádná navigační akce ani zobrazení kontextových detailů, jak bylo lze očekávat.

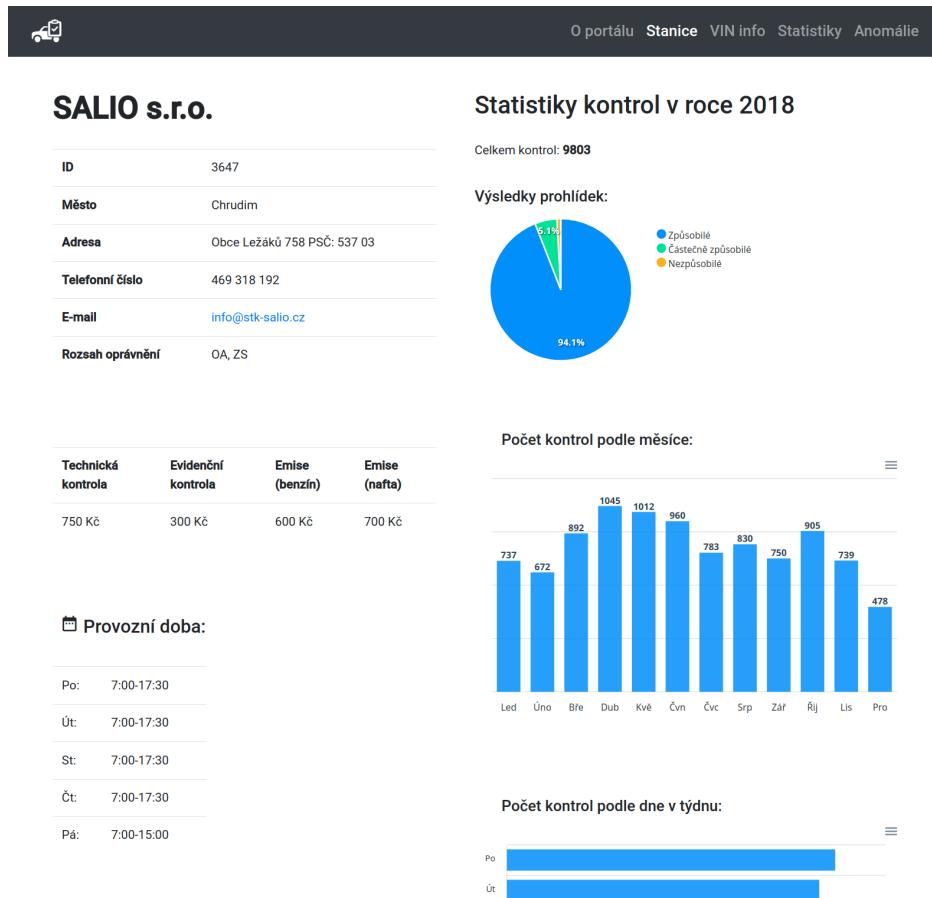
Portál je dále přehledně dělen na několik obrazovek podle funkcionality. První funkcí je filtr/vyhledávač spojený se seznamem STK, odkud lze přejít na detailní stránku každé stanice (obrázek 2.1). Tam má uživatel k dispozici jednak údaje z číselníku, dále informace o tehdejší ceně prohlídek podle typu a provozní době, které bohužel nejsou aktualizované. Následuje koláčový graf podílu výsledků kontroly (způsobilé, částečně způsobilé či nezpůsobilé vozidlo). K dispozici jsou dále sloupcové grafy udávající počty provedených kontrol podle měsíce, dne v týdnu apod. Koláčový graf je využitý rovněž pro zobrazení nejčastěji kontrolovaných značek, bohužel z něj kvůli příliš velkému počtu malých výsečí není příliš dobře patrný poměr mezi četnostmi jednotlivých značek. Zajímavé jsou bezesporu počty anomálních prohlídek podle typu anomálie. Pro lepší zasazení těchto čísel do kontextu by se zde ale hodil např. histogram počtů anomalií napříč všemi stanicemi – závažnost je indikována pouze oranžovým výstražným trojúhelníkem nebo červeným kolečkem a z těchto symbolů není jasné, jak byla závažnost určena.

Druhá sekce, která jako jediná není součástí bakalářské práce a byla doplněna později, umožňuje zobrazit surové záznamy o kontrolách na základě zadání alespoň deseti znaků VIN kódu vozidla. Ze záznamu o kontrole lze např. zjistit její výsledek nebo přejít na detail stanice, kde kontrola proběhla.

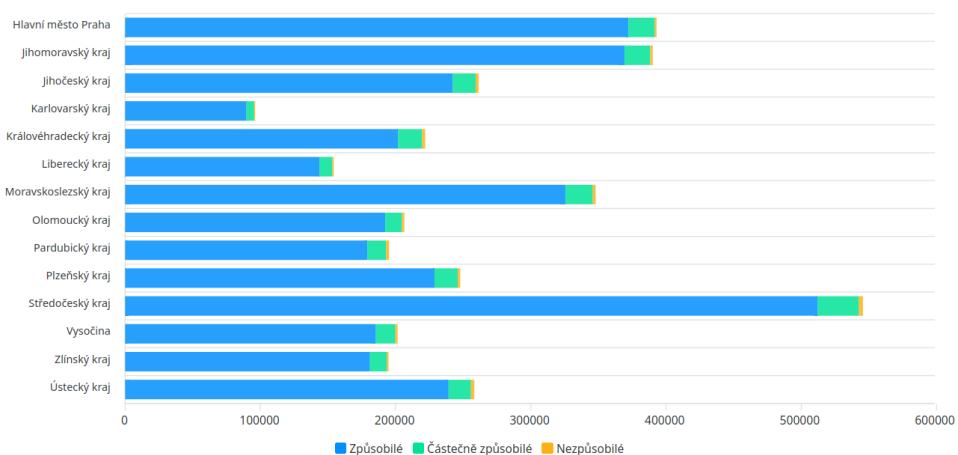
Následují celkové statistiky kontrol. Uživatel opět formou koláčového grafu dostává informaci o podílu výsledků všech prohlídek jak pro technickou kontrolu, tak pro měření emisí. Zajímavá je vizualizace průměrných cen pravidelných a evidenčních kontrol podle krajů. Stejným způsobem jsou pak rozděleny i počty kontrol a vidíme tak například (obrázek 2.2), že ve Středočeském kraji probíhá obecně větší počet prohlídek než kdekoli jinde, ale z grafu bohužel není dobře patrný rozdíl (či jeho neexistence) v poměrné úspěšnosti prohlídek napříč krají.

## 2. DOSAVADNÍ SITUACE

Obrázek 2.1: Výřez z detailu stanice na původním STK portálu.



Obrázek 2.2: Výsledky kontrol podle kraje na původním STK portálu.



Poslední sekce portálu se zabývá výsledky detekce anomalií. K dispozici je histogram počtu prohlídek konaných mimo pracovní dobu podle celé hodiny ve dni a odkazy na stanice, které se na nich v největší míře podílely. Pro čtyři stanice je zde porovnaný odhad pracnosti provedených prohlídek s celkovou dotací pracovních minut na základě otevřací doby; dvě z nich vykazují anomální chování. Zbytek sekce se zabývá bližší analýzou kontrol na těchto dvou anomálních stanicích, dále analýzou chybně zadávaných dat a seznamem nejčastějšími a anomálními souběhy značek vozidel na kontrolách.

## 2.4 Služby pro spotřebitele

Tato podkapitola se zaměří na několik spotřebitelsky orientovaných webů, jež poskytují informace zejména pro potenciální kupce ojetých osobních automobilů. Většina z nich umožňuje zobrazit základní analýzu bezplatně. Za poplatek obvykle v rámci stokorun pak zpřístupňuje rozšířené informace. Jelikož portál, který je výstupem této práce, má poskytovat informace zdarma, následující text se bude věnovat zejména bezplatně dostupným informacím z analyzovaných webů, aby bylo možné funkcionality porovnat.

### 2.4.1 Cebia

Cebia je česká společnost s více než 30letou historií v oblasti prověřování ojetých vozidel. Hlavní službou je poskytování informací o vozidlech na základě jejich VIN kódu prostřednictvím webových stránek <https://www.cebia.cz>. Společnost mimo to nabízí např. mobilní aplikaci pro evidenci stavu vozidla či fyzické služby jako homologaci vozidel nebo satelitní zabezpečení. Informace o vozidle lze vyhledávat na základě VIN kódu a jejich rozsah se může u každého vozidla lišit, protože data pocházejí z různých státních registrů, autoservisů či pojíšťoven z mnoha zemí. Následující popis je proto založen na ukázce prověření, kterou lze zobrazit na <https://cz.cebia.com/payed/detail/example>, náhled poskytuje obrázek 2.3. [11]

Analýza je uvedena kartou se základními údaji o vozidle, kam patří značka a model, typ karoserie, základní vlastnosti pohonné jednotky (palivo, výkon, objem) a data výroby a první registrace. Zbytek webové stránky je rozdělen do sekcí, na něž lze přeskočit pomocí tabulky s obsahem po levé straně. Součástí tohoto obsahu bohužel nejsou nadpisy sdružující jednotlivé sekce do tří kapitol – historie vozidla, kontrola závazků a stav vozidla.

Přehled historie nabízí detailní časovou osu záznamů nájezdu, servisních úkonů a poškození, technických i emisních kontrol a dalších událostí. Pro každou prohlídku na STK je k dispozici protokol, kde lze najít údaje velmi podobné otevřené datové sadě STK 2018, která byla popsána v předchozí kapitole. Navíc jsou zde ale vyjmenovány konkrétní nalezené závady, je uvedena registrační značka vozidla a součástí je i detailní protokol měření emisí. Tato

## 2. DOSAVADNÍ SITUACE

Obrázek 2.3: Výřez prověření vozidla na portálu Cebia.

Datum	Stav tachometru	Země	Popis
2008	prosinec	0 km	Vozidlo vyrabeno
2009	leden	Česko	Přihlášení vozidla v ČR
	leden	Česko	Vystavení technického průkazu
	duben	Česko	Poškození
	květen	Česko	Servisní záznam
	červenec	Česko	Servisní záznam
	červenec	Česko	Poškození
	srpen	Česko	Poškození

data jsou bohužel dostupná pouze jako soubory v PDF, nikoliv uspořádaná přímo na webové stránce např. do interaktivní tabulky.

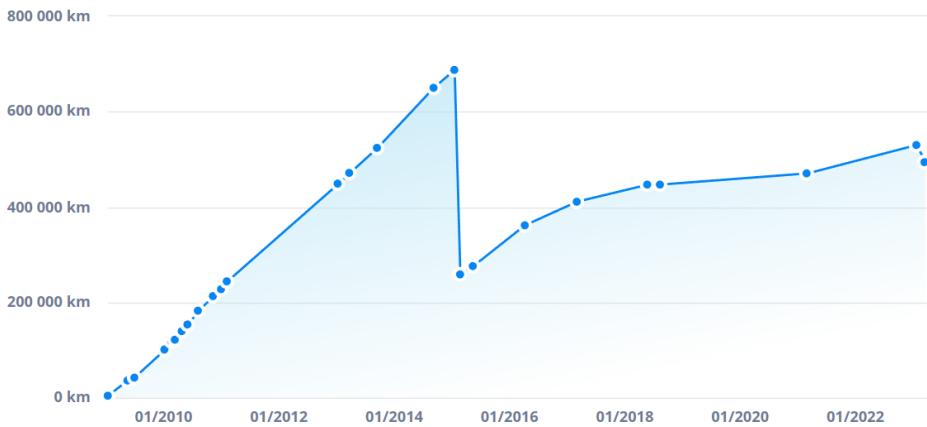
Díky údajům o změnách stavu čítače najetých kilometrů je dostupný graf vývoje nájezdu (obr. 2.4). Tato data pochází z mnoha zdrojů – STK, prohlídek v autorizovaných servisech atd., takže hustota datových bodů může být poměrně velká. Množství dat usnadňuje odhalení známého podvodu „stočení tachometru,“ o kterém Cebia tvrdí, že se vyskytuje u více než třetiny ojetých vozů [12, 13].

Zajímavými sekci jsou záznamy o pojistných událostech. Pro jednotlivé události portál nabízí informaci o zemi, kde vznikla, o výši škody a někdy také o přibližném místě poškození. Navazující část zabývající se servisními úkony obsahuje záznamy, z nichž se uživatel dozví stav tachometru a výčet materiálu použitého k opravě. Materiál není nijak segmentován např. podle funkčního celku, jehož se týká (motor, převodovka, chlazení apod.), údaje se zdají být poskytované v podobě, jak byly zadány, tj. často s chybějící diakritikou nebo nekonzistentním formátováním.

Kontrola závazků sestává z výčtu databází odcizených vozidel a informací, jestli je v některé z nich vozidlo evidováno. Následuje informace o aktuálně běžícím leasingu či jiné formě financování, které by vozidlo mohlo podléhat.

## 2.4. Služby pro spotřebitele

Obrázek 2.4: Historie nájezdu vozidla na portálu Cebia.



Sekce je zakončena kontrolou, jestli vozidlo nebylo provozováno jako taxi.

V poslední části webu nalezneme odhad tržní ceny vozidla, dále svolávací akce, které mohl výrobce vyhlásit, a nakonec technické informace o vozidle. Ty sestávají ze seznamu výbavy a z výpisu z registru vozidel, který je uveden formou jednoduché tabulky, jejíž obsah je patrně opět zobrazením surových dat. To lze odvodit např. z hodnoty spotřeby paliva 0 litrů na 100 km pro vzorové vozidlo, jehož analýza je zde popisována.

Ověření vozidla prostřednictvím portálu Cebia tedy nabízí mnoho detailů týkajících se stavu vozidla, důležitým prvkem jsou data o pojistných událostech a servisních úkonech. Pro spotřebitele mohou být rovněž důležité informace o závazcích, které by mohly kupujícího nepříjemně překvapit. V oblasti technických údajů a záznamů z STK je možné nalézt nevyužitý potenciál dat. Ta mohla být např. automatizovaně vytěžena z PDF protokolů a zahrnuta na webu interaktivní formou. Zároveň by bylo možné zkombinovat data o závadách a evaluoват či dokonce predikovat celkový technický stav vozidla. Celkově se ale jedná o kvalitní analýzu, jen je třeba mít na paměti, že ne všechna popsaná data mohou být dostupná pro každé vozidlo, které si uživatel chce nechat prověřit.

### 2.4.2 Vindecoder

Vindecoder je česká společnost nabízející prověření vozidla na základě VIN kódu. Nabízí podobné informace jako Cebia s výjimkou dat od pojišťoven a servisů. Oproti Cebii také není orientovaná pouze na fyzické osoby hledající radu, ale poskytuje také API pro firmy, které potřebují zjednodušit správu svého vozového parku. [14]

Součástí webových stránek služby <https://vindecoder.eu> bohužel není žádný příklad, jak vypadá zakoupená analýza – návštěvník si může pouze

## 2. DOSAVADNÍ SITUACE

prohlédnout vzorové odezvy jednotlivých API. Lze ale zadat libovolný VIN a web zobrazí dostupnost konkrétních položek, jak je vidět na obrázku 2.5. Následující shrnutí proto neobsahuje popis vizuální reprezentace dat.

Obrázek 2.5: Výřez náhledu před zakoupením reportu od Vindecoder.

The screenshot shows a search interface for a car. At the top, it says: "Shromáždili jsme následující záznamy o Vašem vozidle. Pro detaily si kupte report." Below this, there are three main sections:

- Prověření v databázích odcizených vozidel**: 6 databáze. This section lists countries checked against databases of stolen vehicles:
  - Česká republika (✓)
  - Rumunsko (✓)
  - Slovinsko (✓)
  - Maďarsko (✓)
  - Slovensko (✓)
  - Vincario Stolen DB (✓)
- Autovraky**: 1 databáze. This section lists the country checked against the AutoWreck database:
  - Česká republika (✓)
- Základní identifikace vozidla**: 8 záznamů. This section lists basic vehicle identification details with checkmarks:
  - Značka (✓)
  - Model (✓)
  - Modelový rok výroby (✓)
  - Kategorie (✓)
  - Karoserie (✓)
  - Řada (✓)
  - Pohon (✓)
  - Specifikace vozidla (✓)

Podobně jako Cebia nabízí Vindecoder hledání vozidla v různých středo- a východoevropských databázích kradených vozidel. Navíc umí vůz najít v registru autovraků v rámci ČR. Další spotřebitelsky zajímavou informací je odhad tržní hodnoty vozidla, který zahrnuje vývoj ceny vozidla v čase a rozložení nabídek na mapě včetně cenové mapy. Portál dokáže zobrazit také historii nájezdu či informace o výrobci, datu výroby a registraci.

Zbytek analýzy se týká technických parametrů vozu. Kromě základních informací odpovídajících obsahu technického průkazu jako je značka, model či kategorie vozidla je k dispozici detailní specifikace motoru. Ta kromě paliva, výkonu a objemu zahrnuje také kód motoru, druh převodovky včetně počtu rychlostních stupňů a dokonce emisní standard<sup>1</sup>. Specifikace pokračuje dalšími detaily jako je počet válců, objem oleje nebo spotřeba. Získat lze též

<sup>1</sup> Jedná se o tzv. Euro standard, který byl zaveden v roce 1992 a od té doby vyšlo celkem 6 revizí. Standard omezuje maximální povolené emise oxidu dusíku, oxidu uhelnatého a dalších emisí, které od Euro 7 mají zahrnovat i emise z otěru brzdových destiček a pneumatik. [15, 16]

## 2.4. Služby pro spotřebitele

rozměry vozidla, maximální rychlosť či zatížení, počty kol, dveří či sedadel a zkontovalovat barvu vozu.

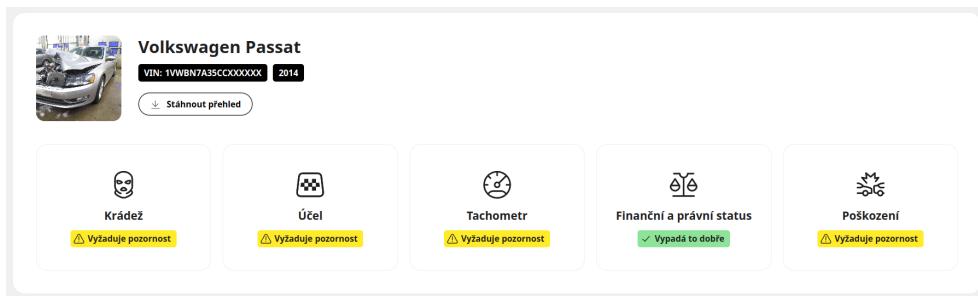
V rámci business-to-business nabídky Vindecoder poskytuje čtyři různé API. První z nich, *VIN Decoder*, obsahuje technické specifikace shodné s odpovídající sekci spotřebitelské analýzy popsáne výše. Druhá, *Vehicle Market Value* za zabývá odhadem tržní hodnoty vozu v závislosti na nájezdu a regionu. Třetí API *Stolen Check* dodává výpis přítomnosti VIN v databázích kradených vozidel. Poslední možností je *OEM VIN Lookup*, sestávající jednak opět z technických parametrů a navíc výpisu jednotlivých prvků výbavy. Nabídka Vindecoderu je tedy proti Cebii spíše chudší (chybí servisní a pojišťovací informace a data z STK), ale vzhledem k dostupnosti API může cílit nejen na spotřebitele. Služba tak dostává svému jménu, protože hlavním cílem se zdá být skutečně pouze dekódování VIN.

### 2.4.3 Carvertical

Carvertical je litevská společnost, jež od roku 2019 buduje globální databázi historie vozidel. Pokrývá většinu evropských států, USA a Austrálii. Z hlediska cílové skupiny se podobá spíše Cebii, mří také na spotřebitele kupující ojeté osobní automobily. Pro Českou republiku lze navíc vozidla kontrolovat buď na základě VIN, nebo čísla registrační značky. Webové stránky <https://www.carvertical.com> nabízí vzorovou analýzu, na níž je založen popis služby. [17]

Report je uveden přehledným zobrazením klíčových oblastí zkoumání a jejich stavem (obr. 2.6). Uživatel tak na první pohled vidí, jaké problémy by se vozidlo mohly týkat a kliknutím může přejít na příslušnou sekci.

Obrázek 2.6: Úvod analýzy od Carvertical.



Analýza je uvedena chronologicky seřazenou fotogalerií vozidla. Fotografie mohou zachycovat automobil např. po havárii a odhalit tak rozsah škod. Následuje ověření využívání vozu v taxislužbě a přítomnost záznamu v evidenci kradených vozidel. Rozsah těchto detailů je poměrně velký, Carvertical kontroluje databáze vozů užívaných v taxislužbách a autoškolách, policejních vozidel či autopůjčoven. Odcizení se kontroluje pro vzorový report v 17 převážně evropských zemích.

## 2. DOSAVADNÍ SITUACE

Přehled historie nájezdu na obrázku 2.7 kombinuje jak data o daném vozidle, tak statistiku průměrného nájezdu podobných modelů. Nechybí vizuálně výrazné upozornění v případě, že mohlo dojít ke stočení tachometru. Chybí však důkladnější popisy obou os grafu, stejně tak zobrazení konkrétní hodnoty datového bodu např. při najetí myší.

Obrázek 2.7: Historie nájezdu vozidla v analýze Carvertical.



Navazující sekce kombinuje data o finančním stavu vozidla, tj. zdali podléhá leasingu či jiné formě financování, spolu s výpisem technických prohlídek a evidencí zlikvidovaných vozidel. Pro technické kontroly zde bohužel chybí jakékoli detaily kromě země, data a úspěšnosti. Vzhledem k litevským kořenům společnosti a faktu, že vzor obsahuje jeden takto nekonkrétní záznam o prohlídce v Litvě, pravděpodobně rozsah poskytnutých dat ani pro ČR nebude širší, avšak tuto domněnku nelze s jistotou potvrdit.

Uživatelsky přívětivě je vizualizována historie poškození vozidla. Web nabízí 3D model vzorového vozu a při zvolení pojistné události zvýrazňuje oblast na vozidle, kde k poškození došlo. Pro každé poškození nechybí odhad nákladů na opravu ani země a datum jeho vzniku. Žádné další detaily jako např. protokoly o provedených opravách však dostupné nejsou.

Předposlední sekce zobrazuje zemi původu, rok výroby a základní technické specifikace. Následuje velmi detailní výčet výbavy vozu. Na samotném konci reportu jsou předchozí data zobrazena v časové ose. Z hlediska uživatelské přívětivosti a vizuálního zpracování se jedná ze tří popisovaných služeb o tu nejpropracovanější. Oproti Cebii neposkytuje zdaleka takové množství

detailních informací, nicméně vše dostupné je zobrazeno přehledně. Přílišnou jednoduchost lze shledat jen u grafu historie nájezdu, který zbytečně skrývá konkrétní hodnoty.

## 2.5 Datové portály

Kromě komerčních služeb existují i různé webové portály, kde může uživatel získat informace bezplatně. Pochopitelně zde není dostupný stejný charakter dat jako u placených služeb, ale pro účely dalšího statistického zpracování mohou být takto získatelná data velmi užitečná. Několik takových zdrojů analyzuje tato podkapitola.

### 2.5.1 Datová kostka

MDČR spustilo v roce 2021 projekt Datová kostka [18], který umožňuje veřejnosti prostřednictvím webových stránek <https://www.dataovozidlech.cz> přistupovat k vybraným údajům z registru vozidel. Registr vozidel je kvůli obsahu osobních údajů neveřejný [19], proto Datová kostka nabízí omezená data – chybí např. informace o majiteli či přidělených registračních značkách.

Portál poskytuje detailní možnosti filtrování dat. Tato data mohou být následně zobrazena přímo ve webovém prohlížeči, nebo lze zažádat o jejich výdej. V takovém případě je nutné zadat jméno a e-mailovou adresu, na kterou je obvykle v horizontu hodin doručen odkaz ke stažení souboru ve formátu CSV.

Filtrování je v případě zobrazení dat v prohlížeči i výdeje do CSV téměř tožné; při výdeji dat lze však zvolit několik dodatečných omezení, např. podle typu vlastníka (fyzická, právnická či neztotožněná osoba) či typu registrační značky (standardní, vývozní, na přání, apod.). Filtrování podle značek a modelů vozidel je vždy omezeno na 150 nejčastějších hodnot, zřejmě z důvodu anonymizace. Méně časté hodnoty jsou pak ve výdeji dat nahrazeny textem JINÉ.

Shodné možnosti filtrování se týkají těchto technických parametrů vozidla:

- druh vozidla řádu prvního (osobní, nákladní, ...) a druhého (kombi, sedan, ...) a kategorie vozidla (M1, N1, ...);
- tovární značka, obchodní označení (model) a výrobce;
- motor (výrobce, palivo, interval objemu a výkonu);
- interval emisí (obecných, městských a mimo město);
- karoserie (výrobce, druh a barva);
- interval počtu míst k sezení, stání a lžížek,

## 2. DOSAVADNÍ SITUACE

---

- interval rozměrů a hmotností;
- přítomnost spojovacího zařízení;
- intervalu nejvyšší rychlosti, spotřeby a poměru výkonu k hmotnosti;
- řazení převodovky (automatická, manuální) pro vozidla evidovaná do roku 2019.

Náhled zobrazení dat je na obrázku 2.8. Portál umožňuje mimo výpis všech vozidel také zobrazit jejich počet podle kraje, okresu či obce. Lze tak například jednoduše zjistit, kolik je aktuálně provozovaných vozů Škoda Octavia na dieslový pohon v každém kraji. Portál tedy až na mírně nepraktický způsob výdeje dat nabízí užitečné informace v uživatelsky přívětivém prostředí.

Obrázek 2.8: Výřez zobrazení dat na portálu Datová kostka.

Status vozidla	Druh vozidla	Druh vozidla 2.ř. TP	Kategorie	Tovární značka	Obchodní označení	Výrobce vozidla	Palivo	Výrobce motoru
PROVOZOVANÉ	NÁKLADNÍ AUTOMOBIL	JINÉ	N1	VW	GOLF	VOLKSWAGEN	Nafta	VOLKSWAGEN
PROVOZOVANÉ	NÁKLADNÍ AUTOMOBIL	JINÉ	N1	VW	GOLF	VOLKSWAGEN	Nafta	VOLKSWAGEN
PROVOZOVANÉ	NÁKLADNÍ AUTOMOBIL	JINÉ	N1	VW	GOLF	VOLKSWAGEN	Nafta	VOLKSWAGEN
PROVOZOVANÉ	NÁKLADNÍ AUTOMOBIL	JINÉ	N1	VW	GOLF	VOLKSWAGEN	Nafta	VOLKSWAGEN
PROVOZOVANÉ	NÁKLADNÍ AUTOMOBIL	JINÉ	N1	VW	GOLF	VOLKSWAGEN	Nafta	VOLKSWAGEN

### 2.5.2 Kontrola tachometru

Dalším projektem MDČR je web <https://www.kontrolatachometru.cz> určený primárně pro získávání výpisů technických prohlídek, jejichž součástí je zápis stavu najetých kilometrů dle počítáče ve vozidle. Tento portál zpočátku nabízel pouze výpis najetých kilometrů v čase, později byl ale rozšířen o zobrazení některých detailů o technických kontrolách. [20]

Po zadání VIN a opsání kontrolního kódu z obrázku za účelem ochrany před automatickým sběrem dat se zobrazí jednoduchá tabulka s řádkem pro každou kontrolu na STK. Záznam obsahuje datum, typ prohlídky (technická kontrola, měření emisí), číslo protokolu, druh prohlídky a nájezd. Pro každý záznam lze získat detaily prezentované na obrázku 2.9. Funkcionalita webu je velmi jednoduchá, ale splňuje svůj účel a poskytuje občanům bezplatně základní informace, které mohou pomoci před koupí ojetého vozu.

## 2.5. Datové portály

Obrázek 2.9: Detail prohlídky na webu Kontrola tachometru.

Detail prohlídky CZ-[REDACTED]									
Datum prohlídky	11.01.2023								
Prohlídka	STK								
Druh prohlídky	Pravidelná								
Stav km	410 127								
Poznámka									
Číslo TP (dokladu)	[REDACTED]								
Kategorie vozidla	M1								
Druh vozidla	OSOBNÍ AUTOMOBIL								
Tovární značka	AUDI								
Obch. označení (typ)	A6 AVANT								
Datum první registrace	[REDACTED]								
Typ motoru	CLAA								
Druh technické způsobilosti	způsobilé								
Seznam závad	<table><tr><td>1.1.11.3.1</td><td>Mírná povrchová koroze/oxidace brzdového potrubí.</td></tr><tr><td>5.1.1.5.1</td><td>Netěsnost nápravy.</td></tr><tr><td>5.3.2.2.1</td><td>Tlumič pěrování je poškozený nebo netěsný avšak tato závada nemá vliv na provozní vlastnosti vozidla.</td></tr><tr><td>6.2.1.1.1</td><td>Povrchová koroze kabiny, karoserie nebo nástavby vozidla.</td></tr></table>	1.1.11.3.1	Mírná povrchová koroze/oxidace brzdového potrubí.	5.1.1.5.1	Netěsnost nápravy.	5.3.2.2.1	Tlumič pěrování je poškozený nebo netěsný avšak tato závada nemá vliv na provozní vlastnosti vozidla.	6.2.1.1.1	Povrchová koroze kabiny, karoserie nebo nástavby vozidla.
1.1.11.3.1	Mírná povrchová koroze/oxidace brzdového potrubí.								
5.1.1.5.1	Netěsnost nápravy.								
5.3.2.2.1	Tlumič pěrování je poškozený nebo netěsný avšak tato závada nemá vliv na provozní vlastnosti vozidla.								
6.2.1.1.1	Povrchová koroze kabiny, karoserie nebo nástavby vozidla.								
<a href="#">Zavřít</a>									

### 2.5.3 Svaz dovozců automobilů

Svaz dovozců automobilů (SDA) sdružuje přes 30 firem věnujících se dovozu automobilů do ČR. Ve spolupráci s MDČR zpracovává každý měsíc statistiky, které jsou dostupné na webových stránkách svazu <https://portal.sda-cia.cz>. [21]

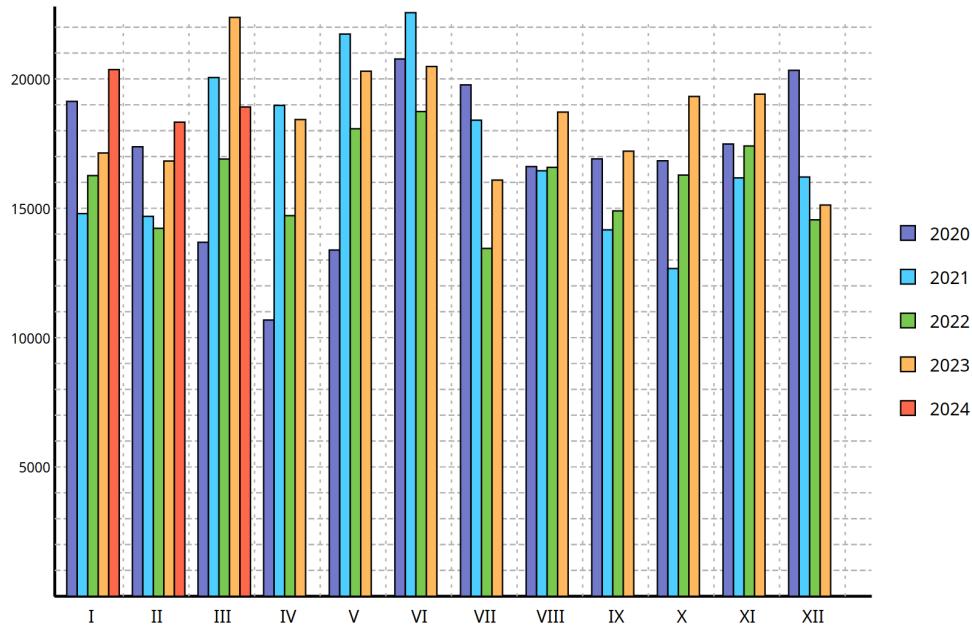
První oblastí statistik jsou registrace vozidel v ČR. Souhrnná statistika umožňuje zobrazit počet registrovaných nových i ojetých vozidel za každý měsíc počínaje rokem 2004. Počty jsou navíc rozděleny podle kategorií vozidel (osobní, lehká užitková a další) a tabulka obsahuje také podíl dané kategorie na registracích každého typu (nové či ojeté). Registrace obou typů lze rovněž zobrazit v separátních tabulkách, kde řádky tvoří roky a sloupce měsíce. Tabulka je doplněna nepříliš přehledným sloupcovým grafem (obr. 2.10), jenž vizualizuje vývoj posledních pěti let.

Další oblastí jsou vyřazená vozidla. Stejně jako v předchozím případě lze zvolit měsíc, případně celý rok, pro nějž se následně zobrazí počty a podíly vyvezených a zrušených (tedy fyzicky zlikvidovaných) vozidel podle kategorie.

## 2. DOSAVADNÍ SITUACE

---

Obrázek 2.10: Graf vývoje registrací nových vozidel podle SDA.



Neméně zajímavá je statistika stavu vozového parku. Pro každý rok je v tabulce zobrazen počet, meziroční přírůstek a průměrné stáří vozidel opět podle kategorie. Portál data nabízí navíc ve formátech Microsoft Excel či PDF a jedná se tak o dobře dostupný zdroj základních informací o objemu vozidel v českém vozovém parku.

## 2.6 Výsledky rešerše

Předchozí kapitoly popsaly jednak komerční služby pro spotřebitele, jednak portály poskytující informace v surovějším formátu bezúplatně. Hlavní předností komerčních služeb je z hlediska spotřebitele přístup k datům z pojišťoven či STK, což umožňuje zobrazit důkladnou analýzu a odhalit podvodné jednání prodejců ojetých vozů. Informace obou služeb, které poskytují vzor analýzy (Cebia a Carvertical), jsou přehledně strukturované. Cebia upřednostňuje množství informací, Carvertical naopak maximální jednoduchost prezentace.

Mezi datovými portály ale stojí web Kontrola tachometru, který poskytuje také informace o historii nájezdu a zároveň alespoň nějaký odhad stavu vozidla díky evidenci poruch zjištěných na STK. Spojil-li by si spotřebitel tato data s výpisem vozidla z Datové kostky, získal by dobrou představu o tom, zdali se jej prodejce ojetého vozu nesnaží oklamat. Z hlediska uživatelské přívětivosti je ale takový postup krajně nepraktický a nelze očekávat, že běžný spotřebitel by měl ochotu tímto způsobem informace získávat.

## 2.6. Výsledky rešerše

---

Z popisovaných portálů pouze Datová kostka a SDA nabízí nějakou formu statistických přehledů o vozovém parku v ČR. V obou případech se však uživatel, který nechce data stahovat a lokálně pomocí dalších výpočetních metod analyzovat, musí spokojit s pouhými sumami či případně průměry podle několika málo kritérií.

Kromě webů, které byly představeny výše, existují ještě další, jež obsahují nějakou formu vyhledávače a přehledu stanic technické kontroly. Jedná se např. o <https://www.stanice-technicke-kontroly.cz> [22] nebo <http://www.seznam-stk.cz> [23], jimiž se již zabývala A. Parkhomenko ve své bakalářské práci. Jejím závěrem bylo, že uvedené portály kromě seznamu a výčtu několika detailů ke každé stanici neobsahují žádná další data jako třeba analýzu proběhlých kontrol. [2]

Z výše uvedeného proto vyplývá, že chybí služba, která by občanům poskytla zajímavější analýzy vozového parku a kontrol na STK jednoduše čitelnou formou. Hledá-li spotřebitel informace o ojetém vozidle, musí se navíc v dosavadní situaci buď spokojit s bezplatnou avšak jednoduchou kontrolou tachometru, nebo získat velmi podrobná data prostřednictvím některé z placených služeb, aniž by měl na výběr nějakou formu střední cesty mezi těmito dvěma volbami.



# KAPITOLA **3**

---

## Datové zdroje

Smyslem kapitoly je detailně popsat datové zdroje, jež práce využívá. Součástí je kromě popisu formátu datových souborů též exploratorní analýza a z ní plynoucí výčet nutných transformací a začištění, která musí být provedena před další analýzou.

### 3.1 Prohlídky na STK

Stěžejní datovou sadou jsou pro tuto práci bezpochyby záznamy o všech proběhlých technických kontrolách na STK. Jedná se o data poskytovaná MDČR pravidelně za každý uběhlý měsíc. Jejich časové rozpětí začíná rokem 2018 a pro účely této práce je omezeno do konce roku 2022, přestože vzniklý software bude schopen zpracovat i novější data.

Částí těchto dat (rokem 2018) a jejím detailním popisem se zabývá bakalářská práce A. Parkhomenko [2, str. 11–12 a 25–33], jak bylo shrnuto v kapitole 2.3.1. Novější data se v některých ohledech liší, avšak rozdílů není mnoho. Následující popis se proto zaměří primárně na novější data a odchylky pro rok 2018 budou shrnuty ke konci kapitoly.

Data jsou členěna do XML souborů po měsících. Následuje seznam atributů všech záznamů a jejich popis:

- **CisP** – číslo protokolu kontroly;
- **VIN** – unikátní identifikátor vozidla;
- **DrTP** – druh kontroly jako např. pravidelná, evidenční apod.;
- **DatKont** – datum (pro rok 2018 i čas) provedení kontroly;
- **Km** – stav počítáče najetých kilometrů při prohlídce;
- **CisTP** – číslo technického průkazu vozidla;

### 3. DATOVÉ ZDROJE

---

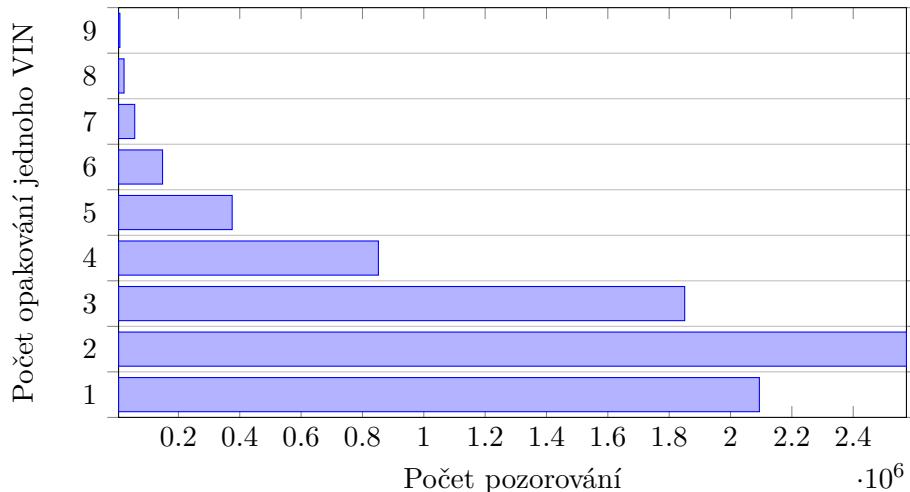
- Ct – kategorie vozidla (M1, N1 apod.);
- DrVoz – druh vozidla, např. osobní automobil či motocykl
- TZn – tovární značka vozidla (název výrobce);
- ObchOznTyp – obchodní označení (model) vozidla;
- RokVyr – rok výroby vozidla, hodnoty chybí;
- DatPrvReg – datum první registrace vozidla (i mimo ČR);
- TypMot – alfanumerický kód typu motoru
- STK – čtyřciferný kód stanice, na které kontrola probíhala;
- Vysl – výsledek kontroly, viz kapitolu 2.2;
- Zav – seznam nalezených závad, resp. jejich číselných kódů;

Detailní výsledky exploratorní analýzy jsou dostupné v digitální příloze v adresáři `exploratory_analysis/inspections`. Exploratorní analýza byla provedena pomocí Python nástroje `ydata-profiling` [24] pro data sloučená za jednotlivé roky kvůli paměťové náročnosti. Tento nástroj generuje pro numerické atributy histogramy a korelační diagramy, pro kategorické atributy výčty nejčastějších hodnot. Počítá také podíl chybějících hodnot a základní statistiky (průměr, počet unikátních hodnot apod.). Mimo to bylo ručně vyčteno několik dalších analýz pomocí Python knihovny `pandas` [25]. Vybrané výsledky jsou diskutovány níže.

Jak je vidět z dříve popsaných komerčních služeb, pro uživatele je velmi důležité znát historii zkoumaného vozidla. Prohlídky na STK jsou pro tento účel vhodným zdrojem. Je proto dobrou zprávou, že většina VIN (odpovídajících jednotlivým vozidlům) se v celé datové sadě (v plném rozsahu 2018–2022) alespoň jednou opakuje, viz graf 3.1. Nalezneme zde nicméně přes dva miliony VIN, které mají pouze jeden výskyt. To může být způsobené vozidly, která zanikla do roku 2020 a mají tedy pouze jednu prohlídku v našem rozsahu (pokud nebyly opakovány), nebo vozidly, která jsou naopak nová a měla v tomto intervalu pouze svou první, evidenční prohlídku. Další možností je též vývoz vozidla z ČR, kdy další technická kontrola již spadá pod cílový stát vývozu.

VIN je důležitým atributem také proto, že je obsažen i v následující datové sadě, tj. registru vozidel a může proto sloužit k propojení obou datových sad. Aktuální standard, tedy 17písmenný alfanumerický kód, ale nebyl v platnosti vždy a je proto nutné zpracovat odchylky. Pokud např. vidíme 16písmennou hodnotu, může se jednat o chybu zadavatele (vozidla s nestandardním VIN existují, takže nelze očekávat, že by systém na STK zadání kratšího kódu neumožnil). Obsahuje-li atribut VIN ale třeba jen 3třípísmennou hodnotu, pouze manuální kontrolou v kombinaci s kompletním registrem vozidel bychom mohli

Obrázek 3.1: Počet VIN podle četnosti jejich výskytu v kontrolách.



ověřit jeho validitu. Navíc bychom také pravděpodobně při spojení datasetu kontrol na registr vozidel dosáhli stavu, kdy k jednomu vozidlu v registru jsou přiřazeny prohlídky několika různých vozů. Z grafu 3.2 plyne, že nestandardních VIN je v datech nesrovnatelně méně než těch standardních. Proto je lze zcela odstranit, aniž bychom přišli o to nejcennější, tj. historii moderních vozů, které často mění majitele a data o nich jsou tedy uživatelsky zajímavá.

Textové atributy kontroly jsou rovněž zajímavým předmětem zkoumání, protože obsahují mnoho různých způsobů zápisu zřejmě stejného údaje. Pro ilustraci tohoto faktu je na grafu 3.3 zobrazen počet výskytů několika vybraných variant zápisu modely Škoda Octavia z června 2021. Variant je celkem 37, z toho 28 se jich vyskytuje méně než desetkrát. Aby bylo možné data smysluplně seskupovat a na skupinách např. podle modelu provádět výpočty, tyto varianty musí být sdruženy do jedné. Přímočarým způsobem je odebrat veškeré dodatky následující po „základním“ názvu modelu, tedy např. veškeré varianty v grafu 3.3 přejmenovat na „Octavia“. Jediný problém může nastat ve chvíli, kdy dodatek v názvu je oficiálním jménem modelu (např. VW Golf vs. Golf Variant).

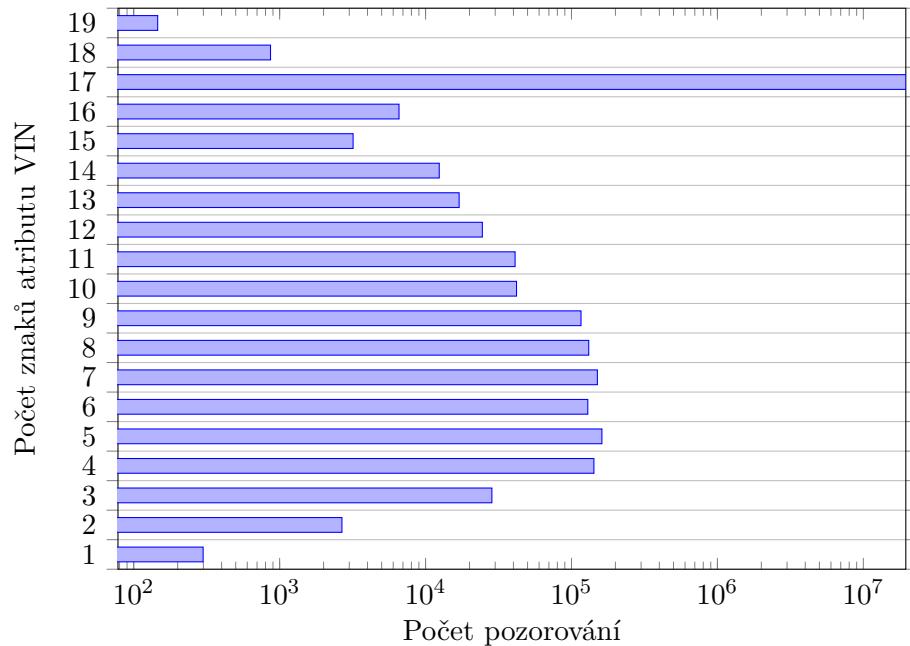
Atribut STK obsahuje čtyřciferný kód stanice, pomocí něhož lze kontroly spojit s dalším z následujících datasetů, seznamem stanic. Vyskytují se zde ale i kódy, které v seznamu stanic neexistují. Takové záznamy nelze jednoduše opravit, a proto mohou být odebrány.

Datum kontroly a první registrace se vyskytuje v několika formátech. Do března 2019 se jedná o ISO formát včetně časové značky. Zbytek dat obsahuje pouze datum, navíc v roce 2019 je od dubna do července datum uvedeno ve formátu **MM/dd/YYYY**. Později je pak datum v českém formátu **dd.MM.YYYY**.

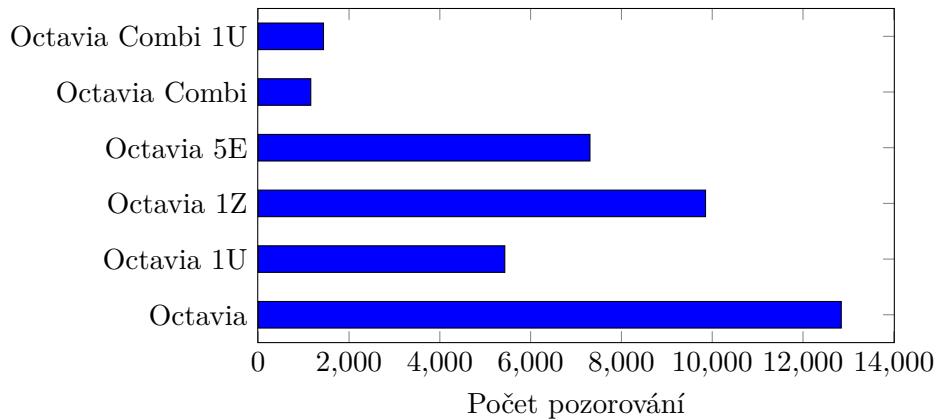
### 3. DATOVÉ ZDROJE

---

Obrázek 3.2: Počet VIN podle délky kódu.



Obrázek 3.3: Vybrané varianty zápisu modelu Škoda Octavia, červen 2021.



Zároveň se vyskytují některé anomální hodnoty, např. 1. leden 1753, což je minimální hodnota typu datetime pro MS SQL server [26]. Zde je proto nutné chybné hodnoty odstranit.

Extrémní hodnoty se vyskytují také v hodnotě nájezdu vozidla. Zde je poměrně obtížné určit, jaká hodnota je chybná a jaká jen anomální. Je tedy

nutné určit vhodnou hranici. Zároveň se např. za celý rok 2020 v atributu vyskytuje 8,2 % nulových hodnot.

Některé atributy vzhledem k cílům práce neposkytují žádný užitek a mohou proto být zahrozeny. Vzhledem k dostupnosti registru vozidel popsaného níže jsou také některé atributy duplikované. V takovém případě lze upřednostnit registr vozidel, protože tak odpadne nutnost rozhodování, kterou z potenciálně různých hodnot z několika prohlídek jednoho vozidla využít. Zahrozeno proto může být číslo protokolu, číslo technického průkazu (TP), kategorie a druh vozidla, rok výroby a datum první registrace a typ motoru.

Datová sada pro rok 2018 se liší několika způsoby. Zaprvé se jedná o jediný XML soubor, kterému navíc chybí mateřský XML uzel a preambule (záznamy jsou za sebou jako jednotlivé uzly **record**, což odporuje stromové struktuře XML). Data dále neobsahují informace o zjištěných konkrétních závadách, pouze sumy závad podle kategorií. Chybí také čísla protokolů a TP.

## 3.2 Registr vozidel

Pomocí Datové kostky lze samoobslužným způsobem získat anonymizovaný export registru vozidel. Pro účely této práce je použita verze exportu z 24. února 2023. Při žádosti o výdej dat bylo filtrování zcela vynecháno a byly zvoleny všechny dostupné sloupce.

Výstupem výdeje dat je CSV soubor o velikosti cca 7,1 GB v kódování Windows-1250 se svislou čarou jako oddělovačem. Soubor obsahuje zhruba 7 milionů záznamů. Exploratorní analýza byla provedena opět pomocí nástroje **ydata-profiling**, jednotlivé sloupce či skupiny sloupců jsou popsány níže.

**Rok výroby** chybí v 30,1 % případů, zejména pro vozidla od roku 2010, pročež se na tuto hodnotu nelze spolehnout např. pro využití v predikčních modelech a může proto být zahrozena. Přítomny jsou extrémní hodnoty jako 0 či 29 012 015.

**Stav** nabývá čtyř různých hodnot a je vždy vyplněn. Hodnoty PROVOZOVANÉ, ZÁNIK, VÝVOZ a VYŘAZENO Z PROVOZU jsou samopopisné.

**1. registrace** je uvedena v ISO formátu s časem vždy rovným 0:00:00. Označuje datum první registrace vozidla i mimo ČR, např. pokud bylo vozidlo importováno jako ojeté ze zahraničí. Hodnota chybí v méně než tisícině případů. Podezřele často (v jednom procentu případů) se jedná o 1. leden 1990.

**1. registrace ČR** má stejný formát jako 1. registrace, může se od ní lišit právě v případě importu ojetého vozu. Hodnota je vždy vyplněna. Podobně jako výše je 0,5 % dat rovno 1. lednu 1990.

### **3. DATOVÉ ZDROJE**

---

**ZTP** je zkratkou pro základní technický popis schváleného typu vozidla. Tento popis slouží jako podklad k vyplnění technického průkazu [27]. Jelikož základní technické popisy nejsou součástí datové základny této práce a hodnotu tak nelze s ničím spojit, sloupec může být zahozen.

**ES** je prázdný sloupec, který proto může být zahozen.

**Druh** nabývá hodnot jako OSOBNÍ AUTOMOBIL, MOTOCYKL apod. Chybí pouze v 0,3 % případů a lze pomocí něj filtrovat vozidla. K dispozici je celkem 117 různých hodnot, např. mnoho typů přívěsů pro nákladní vozidla či traktory.

**Druh 2. ř.** upřesňuje předchozí atribut. Označuje např. tvar karoserie (hodnoty jako KOMBI nebo HATCHBACK), druh nákladního vozidla (SKŘÍŇOVÝ, TAHAČ NÁVĚSŮ) apod. V souboru se nachází 137 různých hodnot, avšak pouze necelých 60 % záznamů je vyplněných. Zajímavostí je, že pro osobní vozidla dominuje typ karoserie kombi – počet hatchbacků a sedanů je zhruba o třetinu menší.

**Kategorie** označuje kategorie vozidla ve smyslu § 4 vyhlášky 153/2023 o schvalování technické způsobilosti vozidel a technických podmínek provozu vozidel na pozemních komunikacích [28]. Nejčastější kategorií je M1, tj. motorová vozidla pro přepravu osob s nejvýše osmi místy pro pasažéry bez míst pro stání, která má přes 12 milionů výskytů.

**Tovární značka** nabývá hodnot jako Škoda, VW či Ford. Toto pole je vyplněno pouze pokud se jedná o jednu ze 150 nejčastějších továrních značek, zřejmě kvůli anonymizaci vydaných dat. Z tohoto důvodu chybí čtvrtina hodnot úplně a 2,5 % obsahuje text JINÉ. V záznamech o kontrolách na STK jsou ale značky vyplněny téměř vždy a spojením obou datových sad podle VIN tedy lze mnoho hodnot do registru doplnit.

**Varianta název** je pole s nejasně určeným významem. Často zejména pro novější vozidla obsahuje kód motoru. Pro starší vozidla zde ale bývá třeba dodatek k názvu modelu – např. některé Škody 120 L mají právě ono L uvedeno v tomto sloupci. Jelikož se tak na jeho význam nelze spolehnout, měl by být sloupec zahozen.

**Verze název** někdy obsahuje jako prefix hodnotu předchozího sloupce, jindy jsou zde zcela jiné alfanumerické kódy, nebo hodnota chybí úplně. Atribut proto rovněž nemůže být hromadně využit.

**Obchodní označení** obsahuje hodnoty jako FABIA, OCTAVIA apod. V dalším textu tedy bude označované také jednoduše jako model vozidla. Obdobně jako u obchodního označení je vyplněno pouze 150 nejčastějších hodnot, ostatní jsou uvedeny jako JINÉ (19,5 %) nebo zcela chybí – dohromady tak hodnota není k dispozici téměř u poloviny záznamů. Naštěstí je však lze také mnohdy doplnit z kontrol na STK. Tak jako u kontrol na STK jsou zde často modely uvedeny s různými dodatky. Navíc sloupec obsahuje zjevně nesmyslné hodnoty, např. kombinace značky a modelu jako „Audi Favorit“ (kde měl být model podle ostatních údajů zřejmě „RS6“), dále „Škoda Golf“ nebo „VW Cิตigo“. Tyto anomálie jsou přítomny běžně v jednotkách, maximálně desítkách případů.

**Tvar karoserie** je dalším poměrně nespolehlivým sloupcem. Někdy kopíruje Druh 2. řádu, jindy obsahuje text NEUVEDENO anebo chybí. Je proto vhodné jej také zanedbat.

**Typ** opět obsahuje nepříliš vypovídající data. Pro některé záznamy se zřejmě jedná o první dvě části VIN, kdy pak skutečný sloupec pro VIN obsahuje pouze jeho zbylou část. Často se jedná o dodatek k modelu vozidla. Sloupec proto může být zahozen.

**VIN** obsahuje identifikační číslo vozidla. Sloupec je velmi důležitý pro propojení registru vozidel s kontrolami na STK. Vzhledem k orientaci výsledného webu na spotřebitele a zájemce o statistiky lze registr vozidel omezit pouze na osobní automobily, nákladní automobily a motocykly. Ze 14 milionů záznamů pak zbude po odstranění duplikátů a jiných než 17znakových hodnot 78 %, odebraná jsou zejména starší vozidla, která nemohou být spojena s daty o kontrolách a pro analýzu tak nejsou zajímavá.

**Ostatní sloupce** popisují technické parametry vozidla. Z hlediska zobrazení na webu se jedná o údaje, které musí být poskytnuty uživateli ve většině případů v surové podobě, protože jejich validace by vyžadovala buď získání technických specifikací od výrobců, anebo ruční práci na milionech záznamů (předchozí sloupce určené k zahození by mohly být také zahrnuty ve webové prezentaci, ale jejich popis by mnohdy mohl být pro uživatele matoucí). Následuje proto pouze jejich souhrn s případnými poznámkami.

- **Výrobce vozidla** – obvykle shodný s tovární značkou, uvedeno 150 nejčastějších, jinak JINÉ;
- **Motor/Výrobce** – opět uvedeno 150 nejčastějších, jinak JINÉ;
- **Motor/Max. výkon** – hodnota ve W;
- **Motor/Min otáčky** – otáčky za minutu, při nichž platí max. výkon;

### 3. DATOVÉ ZDROJE

---

- **Motor/Zdvihový objem** – hodnota v cm<sup>3</sup>;
- **Palivo** – jednotlivá paliva oddělená +, pokud vozidlo má více pohonů;
- **Karoserie/číslo**;
- **Karoserie/Výrobce** – nekonzistentně vyplňené, chybí nebo obsahuje JINÉ;
- **Míst celkem** – nekonzistentně formátované, hodnoty jako např. 5 nebo 2+3;
- **Míst k sezení** – obdobně jako předchozí sloupec;
- **Míst k stání** – pro vozidla bez míst k stání buď 0 nebo chybí;
- **Lůžek** – počet lůžek;
- **Barva** – jedna z deseti barev, neobsahuje podrobnou definici barvy od výrobce;
- **Max. zatížení střechy** – hodnota v kg;
- **Objem cisterny** – hodnota v l;
- **Rozvor, Rozchod 1–4, Délka, Šířka, Výška, Rozměry ložné plochy** – sloupce s hodnotami v mm;
- **Přípustné a povolené hmotnosti** – sloupce s hodnotami v kg;
- **Spojovací zařízení (SZ)** – název třídy SZ, nejčastěji TŘÍDA A50-X;
- **Přípustné a povolené hmotnosti SZ** – sloupce s hodnotami v kg;
- **Nápravy druh** – zřejmě značí, které nápravy jsou hnané;
- **Nápravy počet** – celkový počet náprav;
- **Pneumatiky a ráfky** – sloupce se specifikacemi povolených rozměrů pneumatik a ráfků;
- **Max. rychlosť** – uvedena v km/h, obsahuje zjevně nesprávné hodnoty jako 0;
- **Spotřeby ve městě a mimo město** – hodnoty v l/100 km;
- **Převodovka** – MANUÁLNÍ nebo AUTOMATICKÁ, chybí pro vozidla od roku 2019;
- **Brzdy** – sloupce s pravdivostní hodnotou přítomnosti provozních, parkovacích nouzových a odlehčovacích brzd;

- **Hluk** – hodnoty v dB pro hluk stojícího a jedoucího vozidla;
- **Emise CO<sub>2</sub>** – hodnota v g/km pro emise ve městě a mimo město;
- **Prohlídka status** – PLATNÁ, NEPLATNÁ nebo ØSTATNÍ, vztahuje se k technické kontrole;

## 3.3 Seznam stanic

MDČR poskytuje na svých webových stránkách ke stažení kromě mnoha statistik také seznam stanic technické kontroly [9]. Jedná se o soubor ve formátu MS Excel a v této práci je použita verze platná k 15. 3. 2023. Jednotlivé záznamy jsou seskupeny podle kraje, kde se stanice nachází. Ke každé stanici jsou dostupné následující údaje:

- čtyřciferný kód stanice rozdelený uprostřed tečkou, ale jinak odpovídající hodnotám z datové sady kontrol;
- seznam typů vozidel, jež má stanice oprávnění kontrolovat, např OA (osobní automobil), NA (nákladní automobil) apod.;
- umístění stanice rozdelené do sloupců na PSČ, město, ulici s č.p., ORP, okres a kraj;
- obchodní firma (název) provozovatele;
- kontakty, tj. telefonní číslo či čísla oddělená čárkou, e-maily oddělené středníkem;

Vzhledem k malé velikosti souboru je formátování údajů poměrně uniformní a data tedy mohou být dobře strojově zpracována. Po odebrání tečky z čísla stanice tak lze celý záznam spojit s datovou sadou kontrol na STK a třídit kontroly geograficky.

## 3.4 Číselník závod

Jelikož pro kontroly na STK jsou od roku 2019 známy konkrétní kódy závod, jež byly na každém vozidle nalezeny, bylo by pro uživatele pohodlnější, aby viděl také název závady a její závažnost podobně, jako je to např. na webu Kontroly tachometru. Číselník závod ale není k dispozici v žádném formátu připraveném pro přímočaré strojové zpracování. Na webu <https://www.zakonyprolidi.cz> jsou ale dostupná úplná znění zákonů včetně příloh, mimo jiné také vyhláška č. 211/2018 Sb. o technických prohlídkách vozidel [6], která ve své příloze č. 1 zahrnuje seznam kontrolních úkonů. Jejich čísla odpovídají právě kódům závodů, tudíž lze toto znění ve formátu HTML vytěžit a z tabulky požadované údaje získat.

### **3. DATOVÉ ZDROJE**

---

Tabulka je strukturovaná podle úrovní hierarchie kontrolních úkonů. Jejich kódy jsou ve formátu čísel oddělených tečkami, kde každá tečka znamená zanoření v hierarchii. Na první úrovni jsou okruhy uvedené v kapitole 2.2, tj. závady týkající se brzd, světlometů, řízení apod. Úroveň zanoření každé jednotlivé závady není konzistentní, pomyslný strom má tedy listy v proměnlivé hloubce. Pro každou závadu lze z tabulky získat kód závady, textový popis obvykle o délce jedné věty a stupeň závažnosti.

Vyhľáška prochází časem různými změnami, a proto se i mění celkový počet závad. Z předchozích znění však lze potvrdit, že jednotlivé závady mají napříč novelizacemi stejné kódy, přidávány jsou bezkonfliktně se zpětnou kompatibilitou. Při spojování tohoto datasetu s kontrolami na STK se tedy může stát, že popis závady bude chybět, ale pokud k dispozici bude, neměl by být chybný.

# KAPITOLA **4**

---

## **Metody zpracování dat**

Tato kapitola popisuje několik různých metod datové analýzy, jež jsou dále v práci evaluovány pro těžbu znalostí z dostupných dat. Metody jsou členěny podle druhu výstupu, jež poskytují; pro každý typ výstupu je čtenář uveden do problematiky, jsou nastíněny možné přístupy a pro vybrané metody následuje detailní popis.

### **4.1 Klasifikace a regrese**

Klasifikace je v kontextu strojového učení problém zařazování datových bodů do předem určených tříd. Spadá mezi techniky supervizovaného učení, které využívá trénovací data s již známými třídami. Příklady běžně používaných metod zahrnují:

- lineární klasifikátory, které dělí datové body na základě lineární kombinace jejich příznaků;
- logistickou regresi poskytující pravděpodobnost příslušnosti datového bodu do třídy;
- Support Vector Machines založené na maximalizaci „mezery“ mezi daty rozdělenými nadrovinou;
- rozhodovací stromy, které rekurzivně rozdělují množinu dat podle vybraného příznaku a lokálně optimalizované hranice. [29]

Regresní analýza zkoumá funkční vztahy mezi proměnnými, díky čemuž nachází využití ve velké škále odvětví. Proměnné jsou typicky rozděleny na vysvětlovanou a vysvětlující. Předmětem zkoumání je pak vztah, který by mezi nimi našel zákonitou souvislost – chceme přitom, aby vysvětlující proměnné byly navzájem co nejméně závislé a naopak co nejvíce souvisely s vysvětlovanou proměnnou. [30, s. 1–2]

Datové sady, kterými se tato práce zabývá, obsahují kromě číselných hodnot také velké množství kategorických (ve zdroji textových) hodnot. Řešením, které umožňuje následnou aplikaci např. lineární regrese nebo rozhodovacího stromu, je konverze kategorických hodnot na číselné. Běžně se tak děje pomocí tzv. *dummy variables*, tedy indikátorových proměnných, které nabývají pouze hodnot 0 nebo 1, aby označily příslušnost vzorku k jedné ze tříd. [30, s. 129]

Problém s tímto přístupem ale v kontextu dostupných dat tkví v množství kategorií. Jen různých modelů a značek jsou v datech přítomny desítky tisíc, což by výsledný model činilo neúměrně velkým. Jako alternativa se proto nabízí CatBoost. Jedná se o model využívající gradient boosting nad rozhodovacími stromy, který byl od začátku navrhován pro modelování kategorických dat. [31, s. 1]

Jednou z alternativ pro konverzi kategorických hodnot na číselné je nahrazení kategorií průměrnou hodnotou vysvětlované proměnné pro vzorky v dané kategorii. Máme-li vektor vysvětlovaných hodnot  $\mathbf{Y}$ , matici dat  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$  obsahující vektory  $\mathbf{X}_i = (x_{i,1}, \dots, x_{i,m})$ , pak výskyt kategorické hodnoty  $x_{i,j}$  nahradíme výrazem

$$\frac{\sum_{k=1}^n \mathbf{1}_{x_{k,j}=x_{i,j}} \cdot \mathbf{Y}_k}{\sum_{k=1}^n \mathbf{1}_{x_{k,j}=x_{i,j}}}, \quad (4.1)$$

kde  $\mathbf{1}_{a=b}$  se rovná 1 pokud  $a = b$ , jinak 0. Tímto způsobem jsou ale vysvětlovaná data propagována do jisté míry mezi vysvětlující proměnné, a výsledkem proto může být přeúčení modelu. Přeúčení lze pak řešit tak, že průměr vypočteme z oddělené části trénovacích dat a pouze druhou část využijeme pro samotné trénování. [31, s. 2]

CatBoost popsaný postup vylepšuje tak, že využívá celá data, díky čemuž nepřichází o vzorky pro trénování. Pomocí náhodné permutace  $\sigma = (\sigma_1, \dots, \sigma_n)$  jsou data promíchána a kategorická hodnota pro každý vzorek je nahrazena průměrem vysvětlované hodnoty pro vzorky ve stejné kategorii umístěné před nahrazovaným. Hodnotu  $x_{\sigma_p}$  tak nahrazujeme

$$\frac{\sum_{k=1}^{p-1} \mathbf{1}_{x_{\sigma_k,j}=x_{\sigma_p,j}} \cdot \mathbf{Y}_{\sigma_k} + a \cdot P}{\sum_{k=1}^{p-1} \mathbf{1}_{x_{k,j}=x_{i,j}}}, \quad (4.2)$$

kde parametry  $a > 0$  a  $P$  redukují šum vznikající u kategorií s nízkou frekvencí výskytu. Hodnot parametrů se používá několik a výsledkem je více nových příznaků, které dohromady vedou ke zlepšení celkové kvality modelu. [31, s. 2]

Pro zefektivnění výpočtu CatBoost používá *oblivious trees*, tedy rozhodovací stromy, ve kterých je na jedné hladině ve všech uzlech využito stejně dělicí kritérium. Takto vzniklé stromy jsou vyvážené a méně náchylné k přeúčení. Další výhodou je zrychlení predikce. List lze reprezentovat jako binární vektor konstantní délky a jeho index při predikci pak získat binárními operacemi. [31, s. 4]

Tabulka 4.1: Matice záměn.

Celkem P + N	Predikované jako P	Predikované jako N
Skutečné P	TP (true positive)	FN (false negative)
Skutečné N	FP (false positive)	TN (true negative)

Jednotlivé rozhodovací stromy jsou spojeny technikou *gradient boosting*, jež obvykle lineárně kombinuje mnoho nepříliš kvalitních stromů tak, že sloučený výsledek dosahuje mnohem lepší přesnosti. Opakováně je konstruován jednoduchý rozhodovací strom, který přesností jen mírně předčí náhodnou predikci. Každý následující strom má přitom za úkol predikovat zbylou informaci, kterou předchozí stromy nevysvětlily. Výsledkem je tedy postupná minimalizace hladovým způsobem za pomocí gradientního sestupu. [32]

Pro evaluaci kvality klasifikace lze využít běžné míry *precision* a *recall*. V případě klasifikace do dvou tříd označíme jednu ze tříd jako pozitivní s počtem výskytů  $P$ , druhou jako negativní s  $N$  výskyty. Podle toho, jak model vzorky klasifikoval, je označíme podle tzv. matice záměn vyobrazené v tabulce 4.1. Precision je pak poměr skutečně pozitivních a všech pozitivně predikovaných, tj.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall nebo také senzitivita je poměr predikovaných pozitivních ke skutečně pozitivním:

$$\text{TPR} = \frac{\text{TP}}{\text{P}}$$

a udává tedy, kolik pozitivních vzorků model odhalil. Alternativně lze obě hodnoty zkombinovat pomocí  $F_1$  skóre, které je jejich harmonickým průměrem a definuje se následovně:

$$F_1 = \frac{2 \times \text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}}$$

[33, s. 364–369]

## 4.2 Shlukování

Shlukování spočívá v zařazování datových bodů do několika skupin (shluků) tak, aby body v rámci jednoho shluku byly co nejpodobnější a napříč dvěma shluky co nejrozdílnější. Podobnost a rozdílnost je přitom založena na vhodné metodě měření vzdálenosti dvou bodů. Shlukovací metody se dělí do následujících skupin podle principu jejich funkce:

- metody dělící data do vzájemně disjunktních shluků, kdy datum náleží do shluku, pokud je dostatečně blízko jeho středu;

- hierarchické metody, kdy je soubor dat opakovaně dělen, dokud nevzniknou skupiny s požadovanými vlastnostmi;
- metody založené na hustotě dat, kdy datový bod patří do shluku, pokud má kolem sebe dostatečný počet sousedů;
- metody používající vícerozměrnou mřížku, podle které dělí prostor dat; [33, s. 443]

Metoda DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je založená na hledání regionů s vysokou hustotou dat. Na rozdíl od prvních dvou typů shlukovacích metod tedy dokáže identifikovat i shluky s jinými než kulovými tvary (nezáleží na vzdálenosti příslušných bodů od nějakého středu). [33, s. 471]

DBSCAN pracuje na základě dvou parametrů. Parametr  $\epsilon > 0$  určuje vzdálenost ohraničující „sousedství“ každého bodu. Pokud je v jeho sousedství nalezeno alespoň  $MinPts$  dalších bodů, pak je takový bod považován za *jádro*. Algoritmus postupně navštěvuje jednotlivé body. Pokud je bod jádrový, stane se zakladatelem nového shluku. Do tohoto shluku jsou následně přidáváni jeho sousedi. Jestliže jsou sousedi zároveň jádry, přidají se i jejich sousedi a postup se opakuje. Rozšiřování shluku končí, když přidaný bod není jádrem. Zbylé body, které nejsou jádry a zároveň nemají za souseda jiný jádrový bod, se označí jako šum. [33, s. 472]

### 4.3 Detekce odlehlých hodnot

Identifikace odlehlých či anomálních hodnot je využívaná jednak při začítování dat před jejich analýzou, jednak pro detekci anomálií jako cílových objektů zájmu. Techniky lze rozdělit na tradiční a pokročilé. Mezi tradiční patří metody založené na vzdálenosti bodů (kam spadá i zmíněné shlukování) a na projekci dat do jiného prostoru, kde lze anomálie dobře oddělit. Postupem času ale vznikají i další metody, které se snaží adaptovat mimo jiné na rostoucí počet či dimenzionalitu dat. [34]

Jednou z takových metod je využití neuronové sítě typu autoencoder. Autoencoder je neuronová síť, jejímž cílem je na výstupu získat stejná data jako na vstupu, zatímco se ve skryté vrstvě nachází „úzké hrdlo“, tj. menší počet neuronů než na vstupu a výstupu. Skrytá vrstva tedy vytváří reprezentaci (embedding) vstupních dat v prostoru nižší dimenze. Na rozdíl od jiných metod pro redukci dimenzionality jako je PCA, která je diskutována níže, umožňuje autoencoder zachytit i nelineární vztahy v datech a potenciálně tak vytvořit lepsí embedding. [35]

Detekce anomalií na základě redukce dimenzionality pak pracuje tak, že převede data do prostoru nižší dimenze a nazpátek. To v případě autoencoderu znamená jednoduše vložení dat na vstup, dopředný běh neuronové sítě a získání výstupu. Předpokládá se, že autoencoder najde při učení takový embedding, který optimálně transformuje většinu dat, tj. neanomální datové body. Proto jsou pak za anomálie označena ta data, která se na výstupu výrazně liší od původní vstupní reprezentace (mají vysokou rekonstrukční chybu). [35]

Problémem takto získaných anomalií je však jejich vysvětlitelnost uživateli. Pokud z výstupu není jasné, jak byl získán, jeho důvěryhodnost může utrpět. Pomocí techniky SHAP (Shapley Additive Explanations) je ale možné odhalit propojení datových příznaků s vysokou rekonstrukční chybou a těch, které chybu zavinily – příznaky (či jen několik s největší chybou) jsou ohodnoceny podle toho, do jaké míry se podílí na výsledku.

Metoda pracuje s autoencoderem způsob black-box analýzy, tj. zpracovává pouze vstup a výstup bez znalosti konkrétních detailů architektury. Vliv konkrétního příznaku je izolován do lineárního vysvětlovacího modelu pomocí tak, že se použije již natrénovaný vysvětlovaný model a ostatní příznaky jsou nahrazeny náhodným vzorkováním z blízkých datových bodů. [36]

## 4.4 Modelování časových řad

Modelování časových řad se zabývá popisem sady určitých měření v čase za účelem nalezení vzorců nebo předpovědi budoucího vývoje. Triviálně lze předpovídat na základě střední hodnoty dosavadního měření, podle poslední známé hodnoty či proložením přímky mezi prvním a posledním bodem měření. Ze složitějších metod jsou běžné např. modely exponenciálního vyhlazování, které hledá střední cestu mezi předpovědí poslední a střední hodnoty. Existují též rozšíření, která dokážou pracovat s trendem (postupnou systematickou změnou střední hodnoty) či sezónností (periodickým opakováním určitého vzorce). Dalším populárním modelem, který dokáže podchytit jak sezónnost, tak trend, je SARIMA (Seasonal Autoregressive Integrated Moving Average) model. Jedná se o složení modelu klouzavých průměru (MA) a autoregresního (AR) modelu spolu s integrační složkou. [37]

Autoregresní model predikuje následující pozorování jako lineární kombinaci několika předchozích. Počet předchozích měření využitých k predikci (lagů) se označuje jako řád modelu, označujeme AR( $p$ ). Model řádu  $p$  tedy pro pozorování  $y_1$  až  $y_t$ ,  $t \geq p$ , parametry  $\phi_i$ , chybu  $\varepsilon_t \sim \mathcal{N}(0, 1)$  a případnou konstantu  $c$  lze zapsat

$$y_t = c + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (4.3)$$

Řád lze identifikovat z chování grafu parciální autokorelační funkce (PACF), který by měl obsahovat odpovídající počet významných lagů a poté strmě klesnout. [37]

#### 4. METODY ZPRACOVÁNÍ DAT

---

Model klouzavých průměrů využívá k vyjádření predikce lineární kombinaci chyb předchozích odchylek  $\varepsilon_t \sim \mathcal{N}(0, 1)$ . Obdobně jako v předchozím případě lze model MA( $q$ ) vyjádřit jako

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \quad (4.4)$$

Autokorelační (ACF) a parciální autokorelační funkce v případě MA modelu vykazují opačné chování, tj. řád odhadneme jako počet významných lagů v ACF, významnost lagů v PACF by měla klesat pozvolna. [37]

Složením obou modelů vzniká ARMA( $p, q$ ). Aby bylo možné modelovat časové řady vykazující trend, lze je navíc differencovat (diferencovaná řada  $y'_i$  se získá jako  $y_i - y_{i-1}$  pro prvky původní řady  $y_i$ , kde  $i \geq 1$ ), čímž vzniká model ARIMA( $p, d, q$ ), kde  $d$  značí počet diferenciací. Nakonec sezónnost lze podchytit přidáním sezónní složky posunuté o počet pozorování odpovídajících periodě opakování vzorce v datech, tedy např. o 12 pozorování při měsíční vzorkovací frekvenci a roční periodě. Výsledkem je kýžený model SARIMA( $p, d, q)(P, D, Q)s$  o sezóně (periodě)  $s$  a ARIMA parametrech napříč sezónou  $P, D, Q$ . [37]

# KAPITOLA **5**

---

## Návrh řešení

Tato kapitola se zabývá v první části vytyčením konkrétních cílů pro datovou analýzu a rozvržením jejich výsledků na webovém portálu spolu s dalšími funkcionalitami. Druhá část popisuje návrh architektury softwarového systému včetně zvolených technologií.

### 5.1 Webový portál a cíle analýzy

Webová aplikace by měla být rychlá, přehledná pro uživatele a měla by být snadno použitelná i na mobilních zařízeních. Web je členěn na několik sekcí sdružujících jednotlivé oblasti analýzy. Sekce jsou mezi sebou provázané odkazy, aby mezi nimi uživatel mohl jednoduše přecházet a získávat tak kontext k obsahu, který zrovna sleduje.

#### 5.1.1 Úvodní stránka

Úvodní stránka slouží jako rozcestník, který uživateli představuje celý web. Obsahuje vyhledávač, pomocí něhož znalý uživatel rychle přejde na detail konkrétní stanice či vozidlo. Pro ty návštěvníky, kteří jsou na webu poprvé, shrnuje jednotlivé sekce, jejich popis a vybrané grafy, které uživatele zaujmou. Součástí je též statistika množství dostupných dat, zobrazená jako počty dostupných záznamů a jejich časové rozpětí. Konceptuální návrh úvodní stránky bez vybraných analýz je na obrázku 5.1.

#### 5.1.2 Stanice technické kontroly

Sekce o stanicích poskytuje jednak místo pro analýzy týkající se všech kontrol na STK, jednak pro rozcestník k detailům každé stanice. Funkcionalitu rozcestníku obstarává vyhledávač a interaktivní mapa. Vyhledávač poskytuje možnost hledání stanice podle adresy, názvu firmy nebo kontaktních údajů.

## 5. NÁVRH ŘEŠENÍ

Obrázek 5.1: Návrh úvodní stránky webu.



Uživatel je po stisknutí vyhledávacího tlačítka přesměrován na stránku s výsledky, odkud může přejít na detail stanice, nebo je informován, že nebyly nalezeny žádné výsledky. Mapa zobrazuje stanice jako body zájmu. Je možné ji přibližovat a nalézt tak např. všechny stanice v místě pobytu uživatele. Po zvolení stanice se otevře dialog se základními informacemi a možností přejít na její detail.

Souhrnné statistiky o stanicích jsou nastíněny v obsahu, za nímž následuje zobrazení jednotlivých grafů. Každý graf umožňuje přejít na stránku s detailním popisem, který uvede uživatele do problematiky a vysvětlí, jak vizualizaci interpretovat. Dostupné jsou následující analýzy.

- Průměrné výsledky kontrol
  - Počet kontrol podle výsledku: celkový počet prohlídek v každém roce a rozpad podle výsledku.
  - Poměrný výsledek populárních značek: základní přehled o spolehlivosti vozidel podle značky.
  - Poměrný výsledek populárních modelů: srovnání spolehlivosti pro několik nejpopulárnějších modelů.
- Závady
  - Nejčastější závady podle kategorie: kategorie závad (brzdy, podvozek, řízení apod.) seřazené podle četnosti jejich výskytu.
  - Nejčastější konkrétní závady: prvních  $n$  nejčastějších závad na všech vozidlech obecně.
  - Nejčastější důvody neúspěšné kontroly podle značek: pro  $n$  nejpopulárnějších značek nejčastější závady, které byly příčinou neúspěšného výsledky kontroly.

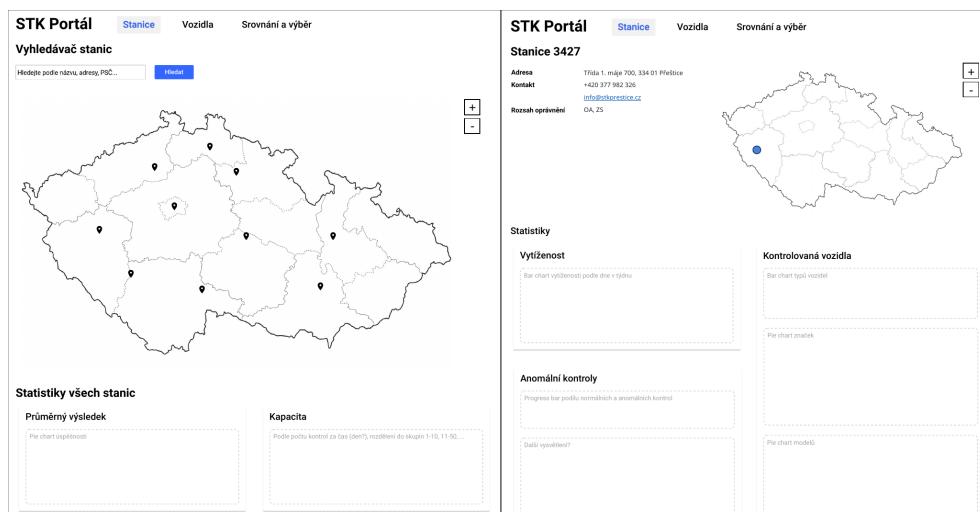
## 5.1. Webový portál a cíle analýzy

- Průměrný počet závad podle závažnosti v rámci jednotlivých krajů.
- Anomální kontroly
  - Podíl všech anomálních kontrol: srovnání počtu výskytů anomalií podle stanice, aby měl uživatel představu, kolik anomalií se vyskytuje běžně a jaké počty detekovaných anomalií jsou již podezřele velké.
  - Anomální kontroly podle typu anomaly: obdobně zobrazené rozdělení počtu anomalií pro jednotlivé typy anomaly.

Portál by měl sloužit mimo jiné jako místo k nalezení vhodné stanice pro absolvování technické kontroly. Základem podstránky s detailem stanice jsou proto informace vycházející ze seznamu stanic obohacené o zobrazení stanice na mapě. U stanice nechybí kontaktní údaje a jaké typy vozidel kontroluje.

Důležitým prvkem je predikce vytíženosti stanice, která by měla uživateli umožnit výběr optimálního dne k návštěvě. Základní statistikou, která se tyká každé stanice, je také výčet nejčastěji kontrolovaných vozidel podle značky a modelu – ilustruje totiž jejich regionální popularitu. Součástí analýzy je rovněž detekce anomálních kontrol. Anomalita může být různých druhů, pro každý z nich je zobrazen počet výskytů na dané stanici spolu se srovnáním počtů výskytů napříč všemi stanicemi. Základní grafický návrh spojující obecnou stránku stanic a její detail je zobrazen na obrázku 5.2.

Obrázek 5.2: Návrh stránky stanic a jejího detailu.



## 5. NÁVRH ŘEŠENÍ

---

### 5.1.3 Vozidla

Druhou sekcí webové aplikace je analýza vozového parku. Rozvržení kopíruje sekci o stanicích pro jednoduchou orientaci uživatele. Pro přístup k detailům o konkrétních vozidlech se v úvodu stránky nachází vyhledávač VIN kódů, který po zadání kódu přesměruje uživatele na stránku s detailem popsanou níže. Analýza založená zejména na registru vozidel se týká následujících okruhů.

- Stáří a nájezd vozidel
  - Průměrný věk osobních automobilů (OA): vývoj průměrného stáří všech osobních automobilů v čase.
  - Průměrný věk OA podle typu pohonu: předchozí graf rozšíří o vývoj věků podle typů pohonů sdružených do několika kategorií (např. všechny druhy plynu dohromady) pro jednoduchost.
  - Průměrný nájezd kontrolovaných OA.
  - Průměrný nájezd kontrolovaných OA podle kraje.
- Značky a modely
  - Popularita značek:  $n$  nejpopulárnějších značek seřazených podle počtu nově registrovaných vozidel za rok.
  - Popularita modelů: totéž pro konkrétní modely.
- Alternativní pohony
  - Typ pohonu nově registrovaných vozidel: vývoj počtu nově registrovaných vozidel za rok rozdělený podle typu jejich pohonu.
  - Elektrifikace nově registrovaných vozidel: detail předchozího grafu omezený na typy pohonu obsahující složku „elektropohon“.
  - Celkový nájezd podle typu pohonu.
- Import ojetých vozidel
  - Průměrné stáří ojetých vozidel při importu.
  - Poměr nových a importovaných ojetých OA.
- Zajímavosti
  - Podíl barev nově registrovaných vozidel: poměrné zastoupení konkrétních barev laku na celkovém objemu zaregistrovaných vozidel podle roku registrace.
  - Status vozidel podle data registrace: poměrné zastoupení aktuálního stavu vozidel podle roku registrace – zdali jsou v provozu, vyvezené, zaniklé apod.

## 5.1. Webový portál a cíle analýzy

Detail vozidla jako podstránka této sekce poskytuje konkrétní informace složené jednak z registru vozidel, jednak z historie kontrol. V úvodu jsou vyjmenovány všechny dostupné sloupce z registru vozidel – tvoří tak sekci o technických parametrech vozu. Následuje tabulka známých kontrol daného vozidla na STK. Tato tabulka umožňuje řazení kontrol podle data, detail stanice, kde kontrola probíhala a seznam nalezených závad včetně jejich popisu a závažnosti. Druhou část stránky tvoří výstupy analýzy. Patří mezi ně graf historie a predikce nájezdu, predikce závad, které by se na vozidle mohly na příští prohlídce objevit a nakonec informace týkající se participace tohoto vozidla na anomálních kontrolách. Grafický návrh je znázorněn na obrázku 5.3.

Obrázek 5.3: Návrh stránky vozidel a detailu konkretního vozidla.

The screenshot displays two side-by-side views of a web portal for vehicle management. The left view shows a search interface for vehicles, featuring fields for VIN and model, and a search button. Below this is a section titled 'Statistiky vozového parku ČR' (Statistics of the Czech vehicle fleet) with a 'Proměny v čase' (Changes over time) chart. The right view shows a detailed view for a specific vehicle (VIN), including sections for 'Detail vozidla <VINxxxxxx>', 'Seznam kontrol' (List of inspections), 'Statistiky' (Statistics), and 'Nájezd' (Journey). The 'Seznam kontrol' table lists inspection results for three dates in August 2018, showing various status categories like 'Způsobilé' (Approved) and 'Nezaplacené' (Unpaid).

### 5.1.4 Srovnávač vozidel

Jelikož webový portál má usnadnit spotřebitelům výběr vozidla, součástí je také srovnávač konceptuálně podobný např. srovnávač produktů v interneto-vém obchodě. Srovnávač se dělí na dva typy. První umožňuje porovnat dvě konkrétní vozidla vyhledaná podle VIN. Obsahem srovnání je tabulka porovnávající technické informace z registru vozidel a také veškerý obsah stránek detailů obou vozidel přizpůsobený svým grafickým rozvržením ke srovnávání.

Druhá varianta srovnávače se zabývá srovnáním modelů vozidel. Ke zvolení obou modelů jsou k dispozici textová vyhledávací pole, která uživateli nabízí automatické doplnění rozepsaného názvu značky či modelu. Portál následně zobrazuje srovnání průměrného nájezdu obou vozidel podle jejich stáří. Pro porovnání dostupných motorizací je uživateli dále nabídnuta detailní tabulka vyjmenovávající jednotlivé motorizace obsažené v registru vozidel. Tabulka umožňuje filtrace podle typu pohonu a řazení dle objemů a výkonů motorů.

## 5. NÁVRH ŘEŠENÍ

---

Posledním prvkem srovnání je graf vývoje průměrného počtu závad na kontrolách obou vozidel podle jejich věku.

### 5.1.5 Ostatní požadavky

Vzhledem k dostupnosti číselníku závad jeho strukturované zobrazení tvoří samostatnou sekci. Její součástí je jednak stručný popis datového zdroje a významu závažností závad, jednak samotný výčet závad spolu s obsahem, pomocí něhož lze přeskočit na libovolnou kategorii. Web nabízí také doplňkovou sekci o portálu, kde jsou popsány všechny datové zdroje.

## 5.2 Architektonický návrh

Softwarový celek, který zahrnuje veškeré operace od zpracování dat po jejich webovou prezentaci, je rozdělen do těchto čtyř hlavních částí:

- datový modul pro předzpracování a provedení výpočtů;
- databáze datových zdrojů a výsledků analýzy;
- API pro zpřístupnění databáze strojově a z webu;
- webová aplikace přistupující k API.

Software je kontejnerizován pomocí technologie Docker, která umožňuje zabalit jednotlivé moduly do samostatných kontejnerů, jež obsahují veškeré potřebné závislosti. Na rozdíl od nasazení pomocí virtuálního stroje má Docker menší požadavky na režii, zatímco poskytuje potřebnou míru izolace od hostujícího serveru. Jedná se proto o vhodný způsob distribuce softwaru složeného z více funkčních celků. [38]

Každá ze shora vyjmenovaných částí odpovídá jednomu kontejneru. Datový modul a API jsou propojeny virtuální sítí pro přístup k databázi. API je pak stejně jako server webové aplikace dostupné na otevřeném portu, kam lze přistoupit z internetu. Uživatel tak není limitován na prohlížení výsledků analýzy a zdrojových dat pouze prostřednictvím webu, ale může použít API přímo a zpracovávat odpovědi strojově.

Software a jeho konfigurace včetně orchestrace pomocí Docker compose je spravována v systému pro správu verzí kódu Git [39]. Vzhledem k tomu, že software je určen široké veřejnosti a zpracovává otevřená data, je Git repozitář dostupný pod open-source licencí také na portálu GitHub<sup>2</sup>.

---

<sup>2</sup><https://github.com/opendatalabcz/STK-portal>

### 5.2.1 Datový modul

Zpracování je implementováno v prostředí Jupyter Notebook [40] s Python backendem. Pro ukládání datových sad v paměti a základní transformace se používá knihovna Pandas [25], která umožňuje rychlé operace s datovými tabulkami včetně exportu a načítání pro persistenci. Tento přístup dovoluje rychlý vývoj jednotlivých analýz nezávisle na sobě s okamžitou odevzrou.

V produkčním nasazení je ale spouštění Jupyter Notebooků krajně nevhodné, protože vyžaduje ruční spouštění jednotlivých analýz a čekání na jejich dokončení. Obsah notebooků je proto následně rozložen do několika Python modulů, které jsou postupně spouštěny hlavním skriptem. Takto vzniklá pipeline umožní v budoucnosti případnou paralelizaci některých úkonů, což může přinést zrychlení běhu. Když se tedy např. každý měsíc doplní nová data o prohlídkách na STK nebo se aktualizuje export registru vozidel, postačí spustit Docker kontejner a veškeré operace v pipeline proběhnou automaticky.

Pipeline je složená z následujících modulů:

- **ingestion**: import a předzpracování datových zdrojů s uložením do databáze;
- **analysis**: datová analýza seskupená podle datových zdrojů;
  - **inspections**: výpočet statistik a aplikace modelů strojového učení pro analýzy dostupné na webu v sekci o stanicích;
  - **vehicles**: totéž pro sekci o vozidlech a srovnávač.

### 5.2.2 Databáze a API

Pro ukládání dat a výsledků analýzy slouží relační databáze PostgreSQL [41]. Jedná se o projekt s více než třicetiletou historií a širokou škálou funkcí, který poskytuje nástroje k zabezpečení integrity dat a rychlý přístup k nim.

O zpřístupnění dat v databázi se stará PostgREST [42], což je REST API server, který automaticky z PostgreSQL relací tvorí API endpointy. Pro projekty jako je tento jeho výhoda spočívá v tom, že vývojář nemusí ručně programovat API server, jehož jedinými funkcemi by bylo přenosilání dat z databáze se zajištěním zabezpečení, stránkování a podobných běžných možností typických pro REST.

Oba produkty poskytují oficiální image Docker kontejnerů [43, 44], což z nich činí komponenty s jednoduchým způsobem nasazení a začlenění do projektu.

### 5.2.3 Webová aplikace

K vývoji webové aplikace byl zvolen framework Next.js [45] založený na knihovně React [46]. Tato vyspělá abstrakce nad Reactem pro vývojáře automatizuje klíčové kroky jako tvorbu projektu, komplikaci a konfiguraci. Součástí jsou subsystémy zajíšťující routing, načítání dat a optimalizaci obrázků a fontů, které tvůrce použije. Implementace navigace mezi jednotlivými podstránkami využívá App router, který na základě adresářové struktury generuje jednotlivé stránky a podporuje některé novější funkce Next.js.

Pro jednoduchou implementaci konzistentního uživatelského rozhraní je doplněna knihovna AntDesign [47], která nabízí všechny typické používané prvky jako navigační lištu, tabulkou s podporou filtrování a řazení, indikátory načítání apod. Aby web mohl zobrazovat interaktivní grafy, používá React wrapper nad knihovnou Chart.js [48]. Pomocí ní je možné tvořit dynamické vizualizace, které dovolují např. skrýt jednotlivé datové sady či zobrazit přizpůsobené popisky datových bodů při najetí myši. Druhá zmíněná funkce dovoluje mimo jiné správně formátovat číselné hodnoty a jednotky podle českých zvyklostí či zobrazit vhodnou formu plurálu pro malá celá čísla. Pro zobrazení interaktivní mapy je zvolena knihovna Leaflet [49], pro kterou je rovněž dostupný wrapper umožňující pohodlné využití v React frameworku.

# Část II

## Implementace



# Datová analýza a návrh modelů

Tato kapitola popisuje implementaci předzpracování dat a aplikaci navržených metod datové analýzy. Podkapitoly se zabývají jednotlivými cíli analýzy, poslední podkapitola shrnuje jednodušší cíle, které spočívají ve výpočtu základních statistik a transformací. Veškeré operace, ke kterým se vztahují uváděné údaje o časové náročnosti běhu algoritmů, byly prováděny na počítači s procesorem AMD Ryzen 5600H, grafickým akcelerátorem Nvidia RTX 3060 a zhruba 28 GB dostupné operační paměti, pokud není uvedeno jinak.

## 6.1 Předzpracování dat

### 6.1.1 Prohlídky na STK

Prvním krokem je konverze dat z XML formátu do Pandas `DataFrame` objektu. Pro tento účel je navržen jednoduchý XML parser založený na ElementTree balíčku [50], jenž je součástí distribuce Pythonu. Parser prochází jednotlivé záznamy, tagy typu `record`, a přidává jejich atributy do Python seznamů. Z těchto seznamů je následně vytvořen Pandas `DataFrame` jako výstup. Typová inference je ponechaná na automaticce Pandas, protože hodnoty jsou následně zpracovávány, což zahrnuje i konverzi na optimální typ.

Čištění dat spočívá v odebrávání příliš poškozených záznamů, případně v úpravě konkrétních atributů. Prvním problémem se ukazuje být více způsobů reprezentace chybějící hodnoty. Ručně lze v datech nalézt tři možnosti, a sice `---`, `nan` a prázdný řetězec; všechny jsou pak nahrazeny za speciální hodnotu `np.nan`, kterou Pandas označuje chybějící záznam.

Datová sada je následně vyfiltrována pouze na prohlídky vozidel typu osobní automobil, nákladní automobil a motocykl. Toto omezení je zvoleno proto, aby výsledky datové analýzy byly lépe interpretovatelné a aby se zmírnil počet outlierů, kterými by velmi pravděpodobně různé traktory (zpravidla velmi staré) či speciální pracovní vozidla byly – vzhledem k zaměření výsledného webu se nejedná o ztrátu cenných dat.

## 6. DATOVÁ ANALÝZA A NÁVRH MODELŮ

---

Odstraněny jsou také všechny záznamy, v nichž chybí zásadní atributy (číslo stanice, datum a typ prohlídky, její výsledek či nájezd vozidla). Naopak nevýznamné jsou např. informace o typu motoru kontrolovaného vozidla, takže jejich absence nevadí. Dále se odeberou záznamy s jiným než 17znakovým VIN a pomocí srovnání s číselníkem závad také záznamy s kódem STK, který v číselníku neexistuje. Složitější kontrola VIN bohužel není možná, protože pro velkou část přítomných VIN kontrolní číslice není povinnou součástí.

Následují transformace jednotlivých atributů. Typ prohlídky je přeložen na jednoduchou kratší hodnotu do angličtiny, datum prohlídky převedeno z různých řetězců na standardní datový typ `datetime` a prohlídky s datem mimo měsíc, jehož soubor byl zpracováván, se zahazují jako chybné.

Ve spojení s číselníkem závad jsou pro každou prohlídku získány počty prohlídek podle závažnosti a na základě kódů prohlídek samotných se přidají počty prohlídek podle kategorie první úrovně (např. brzdy, osvětlení, apod.). Dalším přidaným sloupcem se stává stáří vozidla v čase prohlídky, získané jako rozdíl data prohlídky a první registrace.

Nejsložitější úprava spočívá v řešení problému popsánoho v kapitole 3.1 o datové sadě, tj. různých způsobech pojmenování modelu vozidla, kdy se občas vyskytuje model navíc s nějakým druhotným označením např. generace či výbavy, takže hodnota je chybně víceslovňá. Zvolen je přístup založený na frekvenci výskytu jednotlivých hodnot. Nejprve se připraví počty výskytů jednoslovňých a dvouslovňých hodnot. Dvouslovňům hodnotám s méně než 25 výskytů se pak druhé slovo odebere, pokud první slovo samo má alespoň 50 výskytů. Stejný postup se opakuje pro trojslovňá označení s konverzí na dvouslovňá, přičemž konstanty jsou určeny manuálním zkoumáním výsledků tak, aby bylo opraveno co nejvíce chyb, ale neporušily se správná jména vzácně se vyskytujících vozidel.

Posledním krokem je odebrání duplicitních záznamů. Vzniká tak tabulka o 17 933 514 záznamech způsobilá k dalšímu zpracování. Z důvodu paměťové náročnosti přitom musí být předzpracování prováděno po jednotlivých měsících – pro načtení celé sady totiž nestačí ani 24 GB operační paměti. Začištěná data jsou následně ještě jednou podrobena automatické exploratorní analýze pomocí `ydata-profiling`, jejíž výsledky jsou rovněž součástí digitální přílohy.

### 6.1.2 Registr vozidel

Import registru vozidel je na jedné straně díky CSV formátu poměrně přímočarý, na druhé straně však opět kvůli paměťové náročnosti musí být data zpracovávána po částech. Jako vhodná hodnota zabírající pouze nižší jednotky GB paměti se ukazuje zpracování po milionu řádků.

Zpracování principiálně kopíruje postup u datasetu prohlídek na STK. Chybějící hodnoty jsou v registru vozidel navíc označeny textem `NEUVEDENO` a `JINÉ`. Datum lze díky ISO formátu načíst automaticky a stačí tak pouze odfiltrovat extrémní hodnoty. Složitější zpracování se týká pouze počtu míst

pro pasažéry a typu pohonu – ten je převeden na datový typ seznamu (vozidlo může mít např. elektropohon a navíc spalovat benzín), aby se s ním lépe pracovalo při analýze.

Protože registr vozidel má vyplněná jména značek a modelů pouze pro 150 nejčastějších hodnot, chybějící data jsou doplněna z datasetu kontrol, kde je dostupné vše. Poté jsou začištěny víceslovné názvy modelů stejným způsobem jako u kontrol s jediným rozdílem, a sice že počty výskytů jsou získány z celého registru najednou. Získání počtů výskytů se provádí ještě před dávkovým zpracováním, díky načítání jediného sloupce se neprojevuje problém s nedostatkem paměti.

Problematickou fází předzpracování se ukazuje být nutnost odebrání odlehých hodnot v mnoha numerických sloupcích (výkon a objem motoru, maximální rychlosť apod.). Pro každý sloupec je na základě exploratorní analýzy nalezena vhodná mezní hodnota a veškeré výskytu ji překračující nahrazeny np. `nan`. Výsledkem je tabulka obsahující 11 121 895 záznamů.

### 6.1.3 Seznam stanic

Pandas nabízí pohodlné načtení dat ze souborů MS Excel, díky čemuž je možné tabulku jednoduše převést na `DataFrame`. Jedinými úpravami jsou rozdelení sloupců obsahujících více hodnot (typy kontrolovaných vozidel, e-maily a telefonní čísla) na Pythonové seznamy. Pro jednoduchou implementaci vyhledávání stanic podle těchto polí musejí však původní verze těchto sloupců zůstat – díky tomu lze využít full text search funkci zabudovanou v PostgREST API popisované níže.

Aby mohly být stanice zobrazeny v interaktivní mapě, je zapotřebí získat jejich zeměpisné souřadnice. K tomu slouží služba Geolokátor<sup>3</sup>, jež poskytuje API pro získávání nejen souřadnic, ale také různých statistických a demografických údajů na základě vložené adresy. Pomocí tohoto nástroje lze doplnit souřadnice pro 91 % stanic.

### 6.1.4 Číselník závad

Zpracování číselníku závad je poměrně obtížné, jelikož číselník ve formátu HTML tabulky není uzpůsoben pro strojové čtení. Tabulka proto musí být nejprve ručně vykopírována ze zdrojového kódu na webu Zákony pro lidi a uložena jako XML soubor. Ten se následně zpracuje již zmíněným XML parserem obohaceným o sadu pravidel, která se stará o přeskakování nadpisů v tabulce, aby ve výsledku zůstaly pouze trojice kód závady, popis a závažnost.

---

<sup>3</sup><https://geolokator.profinit.cz>

## 6.2 Predikce závad

Klíčovým cílem výše popisovaných spotřebitelských portálů je bezesporu varování kupujícího před vozidly, která jsou ve špatném technickém stavu a jejich koupě tedy nemusí být ve výsledku výhodná. Je proto nasnadě takovou funkciionalitu přinést i v rámci této práce. Přirozenými zdrojovými daty jsou přitom kontroly na STK, které obsahují výčty zjištěných závad. Závady zjištěné na STK sice nemusí plně odpovídat realitě a nepodchytí např. pojistné události, kde také došlo k poškození vozidla, ale přesto mohou přinést užitečnou představu o stavu vozidla.

Aby bylo možné využít supervizované učení, cílem je pro každou kategorii závad predikovat pravděpodobnost, že vozidlo na příští technické kontrole bude mít alespoň jednu závadu v dané kategorii. Predikce konkrétního počtu závad by byla také proveditelná, ale bylo by pak nutné řešit formát výsledků (není optimální uživateli zobrazit, že vozidlo bude mít 1,5 závady). Pravděpodobnost navíc umožňuje srozumitelné grafické zobrazení a lze ji dobře porovnávat napříč vozidly díky jasnemu rozsahu možných hodnot.

Pro predikci je zvolen klasifikační model CatBoost, který dovoluje predikovat jak přímo třídu datového bodu (zda v dané kategorii bude či nebude závada), tak v případě dvou kategorií i pravděpodobnost. Jelikož kategorií závad je v použitém číselníku deset, ale pro poslední tři je k dispozici velmi málo kontrol, které by tyto závady obsahovaly, natrénováno je 7 modelů, kdy každý předpovídá pravděpodobnost závady v jedné kategorii. Všechny modely přitom vychází ze stejných dat.

Dataset kontrol na STK se nejprve omezí pouze na vozidla, která mají přítomny alespoň dva záznamy. Z tohoto seznamu jsou následně vytvořeny trénovací body přidáním následujících příznaků:

- vybrané příznaky z registru vozidel (informace o motoru, pohonu, modelu a značce) na základě spojení podle VIN;
- čtyři kategorické příznaky nabývající hodnoty 0/1, které indikují přítomnost benzinového, naftového, elektrického a plynového pohonu v kombinaci – jednotlivé druhy plynu se takto sjednotí a z hlediska prostoru příznaků se tímto vytvoří souvislost mezi vozidly, která mají neprázdný průnik pohonů;
- vysvětlované proměnné jako nové příznaky indikující pravdivostní hodnotu nenulovosti počtu závad v jednotlivých kategoriích na následující prohlídce – trénovací bod tak může být vytvořen pouze z prohlídky, která pro dané vozidlo není v datasetu poslední, tj. je znám příští stav vozu.

Objem a výkon motoru, který u některých záznamů chybí, je doplněn jako průměr dané značky a modelu.

Nalezení optimálního nastavení CatBoost modelu probíhá ručním prohledáváním prostoru hyperparametrů v rozsahu uvedeném v tabulce 6.1 (poslední dva kroky při zmenšování learning rate jsou velikosti 0,01). Optimální bylo použít automatické prohledávání na větším počtu bodů v prostoru parametrů, což znemožňuje velká výpočetní a tudíž i časová náročnost trénování v kombinaci s dostupnými hardwareovými zdroji – na uvedené sestavě trvá trénink jedné sady modelů cca 4 hodiny. K evaluaci je použita predikce kategorie (nikoliv pravděpodobnosti příslušnosti do kategorie 1) a ztrátová funkce **Logloss** aplikovaná na testovací dataset, který vznikl oddělením 5 % zdrojových dat.

Tabulka 6.1: Rozsahy vyzkoušených parametrů při tréninku modelu predikce závad.

Parametr	Minimální hodnota	Maximální hodnota	Velikost kroku
Počet iterací	500	1000	100
Learning rate	0.03	0.4	0.05
Hloubka stromu	4	8	1

Při zmenšování learning rate postupně klesá hodnota precision a stoupá recall, po snížení pod 0,05 se ale obě metriky začínají zlepšovat. Jako nejlepší volba se ukazuje 1000 iterací při learning rate 0,03 a hloubce stromu 8. Kvůli paměťové náročnosti musí být trénování prováděno po částech, kdy každý další model vychází z parametrů předcházejícího a výsledný model je součtem všech dílcích.

### 6.3 Predikce vytíženosti stanic

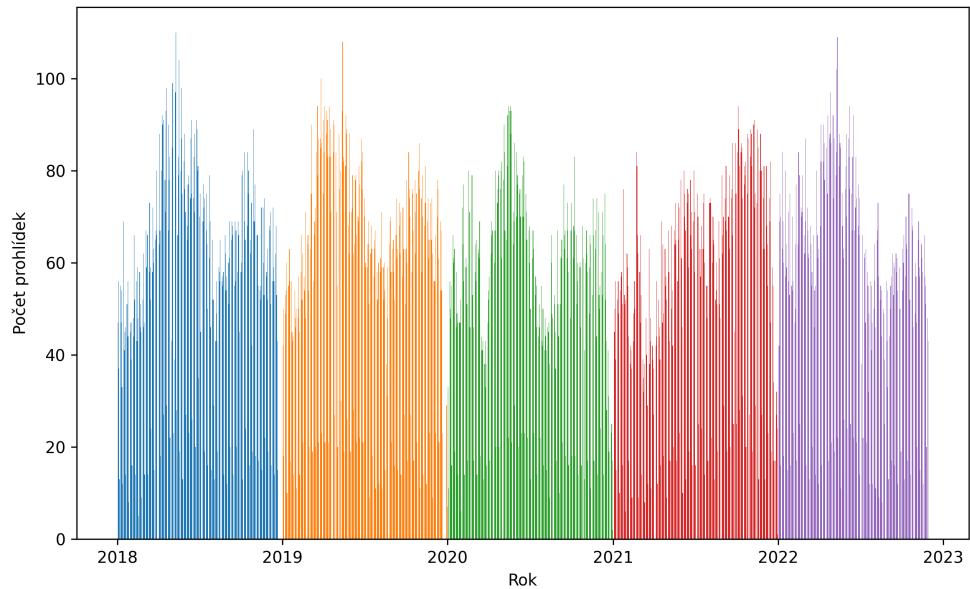
Predikce vytíženosti umožňuje uživateli získat představu o vytíženosti stanice a zvolit vhodný den k její návštěvě. Založená je na časové řadě počtu provedených prohlídek v každém dni, kterou lze z datasetu o kontrolách jednoduše vypočítat. Protože data o prohlídkách jsou poskytována nejméně s měsíčním zpožděním, predikce by měla být rozumně přesná alespoň na dva měsíce do budoucnosti, aby byla vždy dostupná aspoň měsíční předpověď.

Seznámení s daty a evaluace je provedena s využitím historie pražské stanice č. 3102, která kontroluje osobní automobily. Na obrázku 6.1 je zobrazen vývoj počtu prohlídek v celém známém období (do listopadu 2022). Časová řada nevykazuje trend. Ukazuje se sezónnost na bázi pololetí či případně celého roku – v létě počet prohlídek klesá, na jaře a na podzim je vyšší. Od druhé poloviny roku 2020 do první poloviny 2021 lze spatřit poněkud menší maxima než v jiných letech, interpretace dočasnými změnami ve společnosti v období pandemie Covid-19 je zde nasnadě. Při bližším pohledu (obr. 6.2) je také jasné patrná týdenní sezónnost, pátky mají obvykle zhruba třetinový počet prohlídek a víkendy nulový.

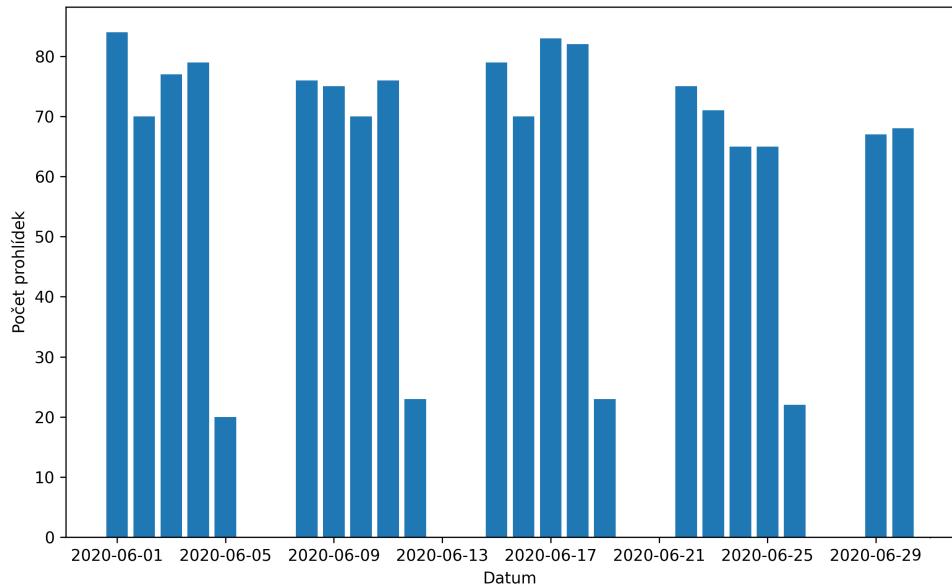
## 6. DATOVÁ ANALÝZA A NÁVRH MODELŮ

---

Obrázek 6.1: Vývoj počtu prohlídek na stanici 3102.



Obrázek 6.2: Detail vývoje počtu prohlídek na stanici 3102.



Pro predikci jsou zvoleny k porovnání tři přístupy.

- První spočívá v triviálním výpočtu průměrného počtu prohlídek pro každý den v týdnu.
- Druhý vypočítává průměrnou návštěvnost v „tomtéž“ dni napříč roky.

### 6.3. Predikce vytíženosti stanic

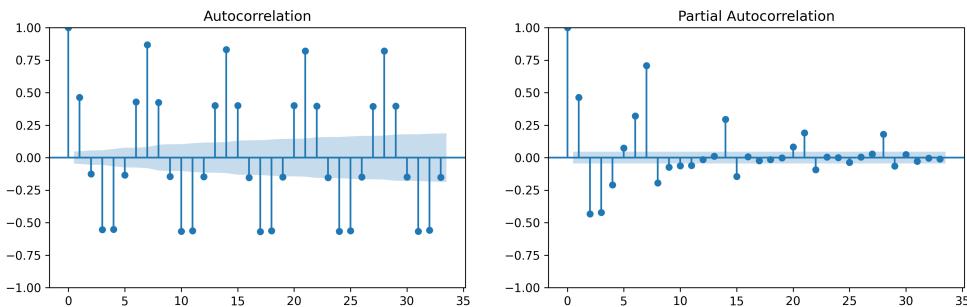
Vezmou se vždy n-té dny v týdnu pro m-tý týden každého roku a těchto pět hodnot (jedna za každý rok) se zprůměruje.

- Třetí metoda využívá týdenní SARIMA( $p, d, q$ )7 model.

Vzájemné srovnání kvality predikcí je provedeno na pomocí MSE (mean squared error), dává smysl penalizovat větší odchylky nelineárně více. Trénovací data tvoří roky 2018–2021, testovací jsou leden až březen 2022.

Pro identifikaci charakteristiky procesu před trénováním SARIMA jsou nejprve analyzovány autokorelační grafy, jejich náhled poskytuje obrázek 6.3. Z grafů (parciální) autokorelace procesu je jasné patrná perioda 7. Vzhledem k velmi pomalu klesající každé 7. hodnotě v grafu ACF mohl být vhodný AR(2) proces s periodou 7, PACF klesá poměrně rychle, i když dokonalá charakteristika AR procesu toto není. Lze tedy očekávat, že proces by mohl být smíšený. Pomocí ADF testu se následně potvrzuje (slabá) stacionarita.

Obrázek 6.3: Autokorelační grafy četnosti prohlídek na stanici 3102.

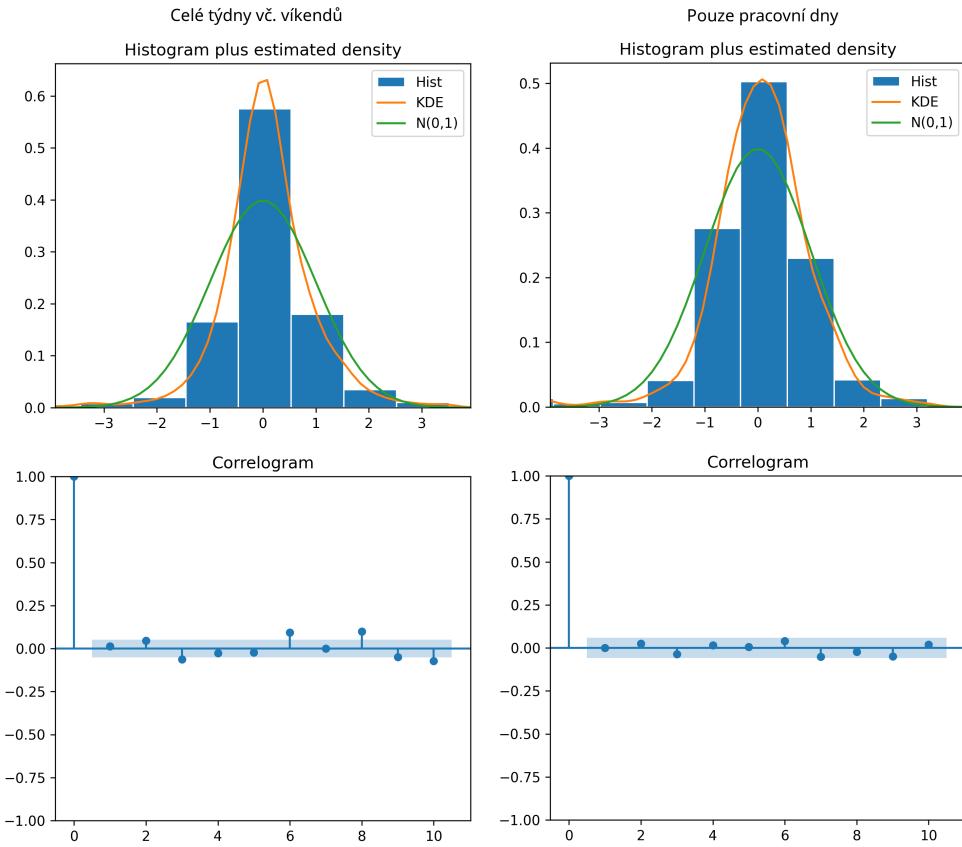


Za použití AIC (Akaike Information Criterion) jsou nalezeny optimální parametry SARIMA modelu, pro všechny hodnoty je použitý rozsah 0–2. Statistické vlastnosti modelu se ukazují být poměrně dobré, zvolený model SARIMA(1,0,2)(2,0,1)7 vysvětluje rezidua s výjimkou poněkud větší korelace v sedmém lagu, tedy v rámci sezóny. Natrénován je proto druhý model na datech pouze z pracovních dní, jehož autokorelace reziduí je již akceptovatelná, stejně jako jádrový odhad, jak je vidět ve srovnání obou modelů na obrázku 6.4.

## 6. DATOVÁ ANALÝZA A NÁVRH MODELŮ

---

Obrázek 6.4: Analýza reziduí predikce SARIMA na stanici 3102.



## 6.4 Predikce nájezdu vozidel

Analýza nájezdu vozidla je standardní součástí výše popisovaných webových služeb, predikce proto může být zajímavým rozšířením. Implementace probíhá podobným způsobem jako u predikce závad včetně předzpracování dat, vstup CatBoost regresního modelu je zkonstruován stejně. Liší se pouze vysvětlovaná proměnná, kterou tvoří přírůstek nájezdu na následující prohlídce. Obdobně jako u predikce závad tedy pro vozidlo, které má v databázi např. tři známé prohlídky, vznikají dva trénovací body, protože pro poslední prohlídku není nájezd na příští prohlídce známý – jeho přírůstek je předmětem predikce.

Problémem regresního modelu je, že může predikovat i záporné hodnoty, v případě nájezdu by jej mohl predikovat v budoucnosti snížený, což není možné (předpověď nemá za cíl odhalit hrozbu přetočení tachometru). Jednoduchým řešením je proto predikovat přírůstek a zápornou predikci zahodit. Pro tyto případy pak predikce není dostupná. Aby se model navíc vyhnul modelům a značkám, které jsou příliš vzácné a byly by proto špatně predikovatelné, model je aplikován pouze na vozidla, která se mají k dispozici alespoň

1000 prohlídek Tato hodnota je odhadnuta z kvality predikce závad na velmi vzácných závadách – jedná se rověž o CatBoost model.

Nalezení optimálních parametrů probíhá stejným způsobem jako u predikce závad. Vyzkoušené parametry uvádí tabulka 6.2, kdy se na základě předchozí zkušenosti learning rate zkouší jemněji v nižších hodnotách, které se ukázaly výhodnější. Optimální hodnotu tvoří learning rate 0,03 při hloubce stromu 12 a 1000 iteracích. Obecně ale platí, že rozdíl v kvalitě predikce není příliš velký, nejlepší model je pouze o 4 % lepší než nejhorší.

Tabulka 6.2: Rozsahy vyzkoušených parametrů při tréninku modelu predikce nájezdu.

Parametr	Minimální hodnota	Maximální hodnota	Velikost kroku
Počet iterací	500	1000	100
Learning rate	0.03	0.1	0.01
Hloubka stromu	8	12	2

## 6.5 Detekce anomálních kontrol

Parkhomenko se ve své bakalářské práci zabývá detekcí anomálií pomocí shlukovacího algoritmu DBSCAN [2, s. 44], proto jsou v této práci otestovány alternativní přístupy.

První z nich využívá principu detekce anomálií na základě rekonstrukční chyby v autoencoderu. Jako vstupní data se používají předzpracované záznamy o kontrolách na STK. Jednoduchá třívrstevná architektura se skrytu vrstvou o čtvrtinové šířce vůči vstupní a výstupní vrstvě dosahuje rozumné průměrné chyby, ale interpretace rekonstrukční chyby pomocí DeepExplaineru zvýrazňuje pouze vztah mezi najetými kilometry a stářím vozu.

Po standardizaci dat a odebrání dummy variables značící typ kontroly se chyba i kvalita její interpretace značně zlepšuje. Ruční zkoumání dat ale ukázuje, že jsou jako anomální často oddělované prohlídky, kde nebyla zjištěna žádná závada. Dalším častým vysvětlením anomaly je neobvyklý vztah mezi stářím vozu a počtem najetých kilometrů – často se jedná o vozy, které mají nulový nájezd, což poukazuje na chybu v datech.

Trénink autoencoderu probíhá v prostředí Google Colab pomocí GPU akcelerátoru Tesla T4, kdy trénování 100 epoch pro jeden měsíc záznamů o kontrolách OA trvá zhruba 30 minut. Protože získané anomálie nejsou příliš zajímavé a výpočetní náročnost je vysoká, tento přístup se ukazuje jako nevhodný.

Druhý přístup proto detekuje anomálie na základě ručně definovaných pravidel odvozených z exploratorní analýzy dat. Pravidla jsou vytvořena pro následující typy anomálií.

## 6. DATOVÁ ANALÝZA A NÁVRH MODELŮ

---

První anomálií je prohlídka v nadměrně vytíženém dni, kdy proběhlo výrazně více kontrol než je průměrem, může indikovat zvýšené nároky na personál a potenciální vliv na průběh prohlídky. Za výjimečně frekventovaný den se považuje takový, kdy počet provedených prohlídek je o dvě standardní odchylky vyšší než průměr daného měsíce. Tato hranice je zvolena tak, aby bylo označeno pouze malé procento dní a tempo práce muselo tedy na stanici být nadstandardní.

Druhý typ anomaly se týká situací, kdy vozidlo na STK neprošlo a následně opakovalo kontrolu na jiné stanici, kde již prošlo. Většinou totiž platí, že prohlídka je opakována na stejně stanici a výjimka je tudíž anomálií.

Třetí anomalita označuje kontroly, při kterých bylo zjištěno alespoň o 5 lehkých závad méně než na předchozí pravidelné kontrole. Zaměřuje se pouze na lehké závady, protože většina z nich se týká koroze či mírného opotřebení některých dílů na podvozku. Tyto závady by neměly být opravovány tak důsledně jako závažné závady, kvůli kterým vozidlo neprojde kontrolou, proto je anomální, když se jejich počet takto výrazně sníží.

## 6.6 Ostatní výsledky

Ostatní statistiky vyjmenované v návrhu webového portálu spočívají v jednodušších datových transformacích. Jejich výpočet je implementován pomocí SQL operací tvořících materialized views s výsledky. Pokud vyjadřovací schopnosti jazyka SQL nejsou dostačující, data se nejprve načtou z databáze do Pandas `DataFrame`. Python skript tato data následně zpracovává a zapisuje zpět do nové tabulky v databázi.

Příkladem je analýza vývoje podílu jednotlivých typů pohonu na nově registrovaných vozidlech v jednotlivých letech. Ve zdrojové tabulce je typ pohonu uveden jako seznam textových hodnot. Ten je v Python skriptu rozdělen na jednotlivé typy, které jsou následně sloučeny do těchto kategorií:

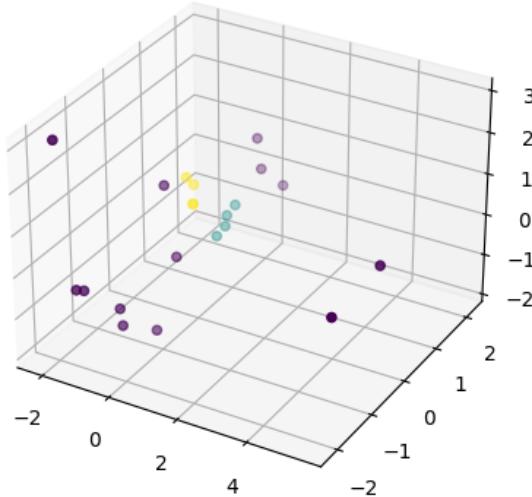
- **Benzin:** záznam obsahuje pouze hodnotu `Benzin`;
- **Nafta:** záznam obsahuje buď `Nafta` nebo `BIO Nafta`;
- **Elektropohon:** záznam je jakoukoli kombinací pohonů obsahující `Elektropohon`, např. tedy také (`Benzin, Elektropohon`);
- **Plyn:** záznam je kombinací obsahující kterýkoli plyn, např. `LPG`, (`Benzin, LNG`) nebo `Vodík`;
- **Ostatní:** např. samostatná hodnota `Etanol 85%`.

Výsledkem je tabulka obsahující řádek pro každý rok s počty výskytů jednotlivých kategorií ve sloupcích.

Poslední analýzou, která pro nemožnost ověření kvality výsledků není součástí výsledného portálu, je analýza firemních flotil. Tato analýza je založená na předpokladu, že vozidla, která náleží do jedné flotily, se vyskytují na STK v společně, v krátkém časovém intervalu. Jedním z možných přístupů ke shlukování je vytvoření datových sad obsahujících prohlídky z jedné stanice v určitém časovém intervalu, např. dvou týdnů. Datovým bodům je následně redukována dimenzionalita pomocí PCA (Principal Component Analysis). Tato metoda data transformuje podle nové lineární báze, v níž jsou zobrazena tak, aby byl maximalizován rozptyl podél jednotlivých os [51]. Druhým možným způsobem redukce dimenzionality je autoencoder, kdy se jako výstup využije jeho vnitřní vrstva, v níž vzniká embedding do prostoru nižší dimenze. Na výsledku je aplikován shlukovací algoritmus DBSCAN.

Příklad jeho výstupu na obrázku 6.5 při redukci do tří dimenzí ukazuje, jak jsou data rozdělena na dva shluky (žluté a modré body) a osamocené fialové body, jež nepatří do žádného shluku. Pro důkladnou evaluaci tohoto přístupu by ale bylo nutné získat velké množství testovacích dat ve formě seznamů VIN kódů různých flotil. Několik takto získaných seznamů ale odhaluje fakt, že flotily jsou často tvořeny i různě starými vozy různých značek a předpoklad o společném absolvování kontrol nemusí platit.

Obrázek 6.5: Příklad výsledku shlukování pro analýzu flotil.





# KAPITOLA 7

## Výsledky analýzy

Tato kapitola shrnuje výsledky hlavních aplikací nástrojů datové analýzy. Ostatní výsledky získané jednoduššími metodami datových transformací jsou předmětem přílohy, která obsahuje screenshoty vybraných grafů z webového portálu.

### 7.1 Predikce závad

Jak je vidět z výsledků evaluace modelu v tabulce 7.1, predikce třídy je vzhledem k relativní jednoduchosti trénovacích dat, kdy se k předpovědi používá pouze bezprostředně předcházející prohlídka, velmi kvalitní.

Tabulka 7.1: Evaluace predikce závad.

Kategorie závad	Precision	Recall	F <sub>1</sub>
0. identifikace vozidla	0,7768	0,7302	0,7528
1. brzdové zařízení	0,8463	0,8063	0,8258
2. řízení	0,7824	0,7472	0,7644
3. výhledy	0,7838	0,7408	0,7617
4. svítily, světlomety, ...	0,8149	0,7903	0,8024
5. nápravy, kola, ...	0,7235	0,8578	0,7849
6. podvozek	0,8847	0,8423	0,8630

Jelikož uživatel uvidí přímo pravděpodobnosti příslušnosti ke všem třídám, rozhodující pro něj bude srovnání jednotlivých pravděpodobností, aby se mohl zaměřit na určité kategorie závad. Z hlediska uživatele také je užitečné, když model predikuje také na základě obecných parametrů vozidla a zobrazí tak uživateli jakýsi průměr závadovosti daného modelu a motorizace obohacený o informaci o stáří a nájezdu konkrétního vozu.

Tyto předpoklady jsou manuálně ověřeny na příkladech několika různě starých vozidel různých značek. Při pohledu na historii kontrol každého z nich a se

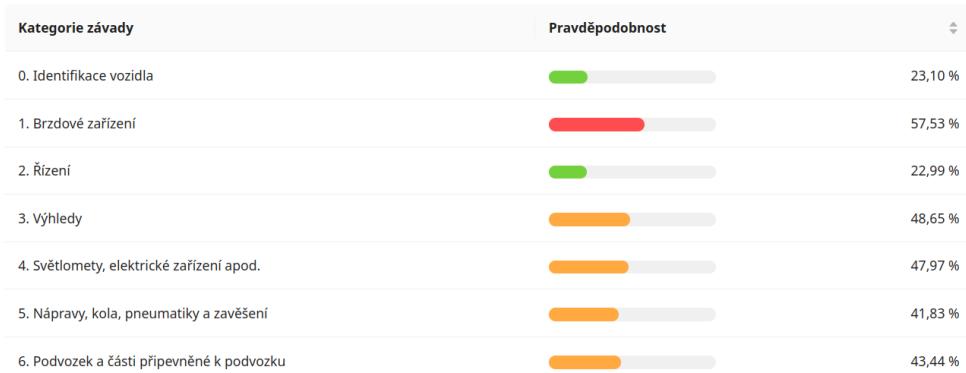
## 7. VÝSLEDKY ANALÝZY

---

znalostí jejich skutečného technického stavu se jeví smysluplné a odpovídající realitě; vizualizace jednoho z nich je vyobrazena na screenshotu 7.1.

Pro využití v produkci jsou modely natrénovány na všech dostupných datech s předpokladem, že se predikce zahrnutím všech dat ještě zpřesní. Parametry získaných sedmi modelů se ukládají pro opakované použití na nových datech, aby nebylo nutné pokaždé opakovat trénink – ten je spolu s případnou úpravou hyperparametrů možné zopakovat, když přibude větší množství dat.

Obrázek 7.1: Predikce závad v sekci detailu vozidla.



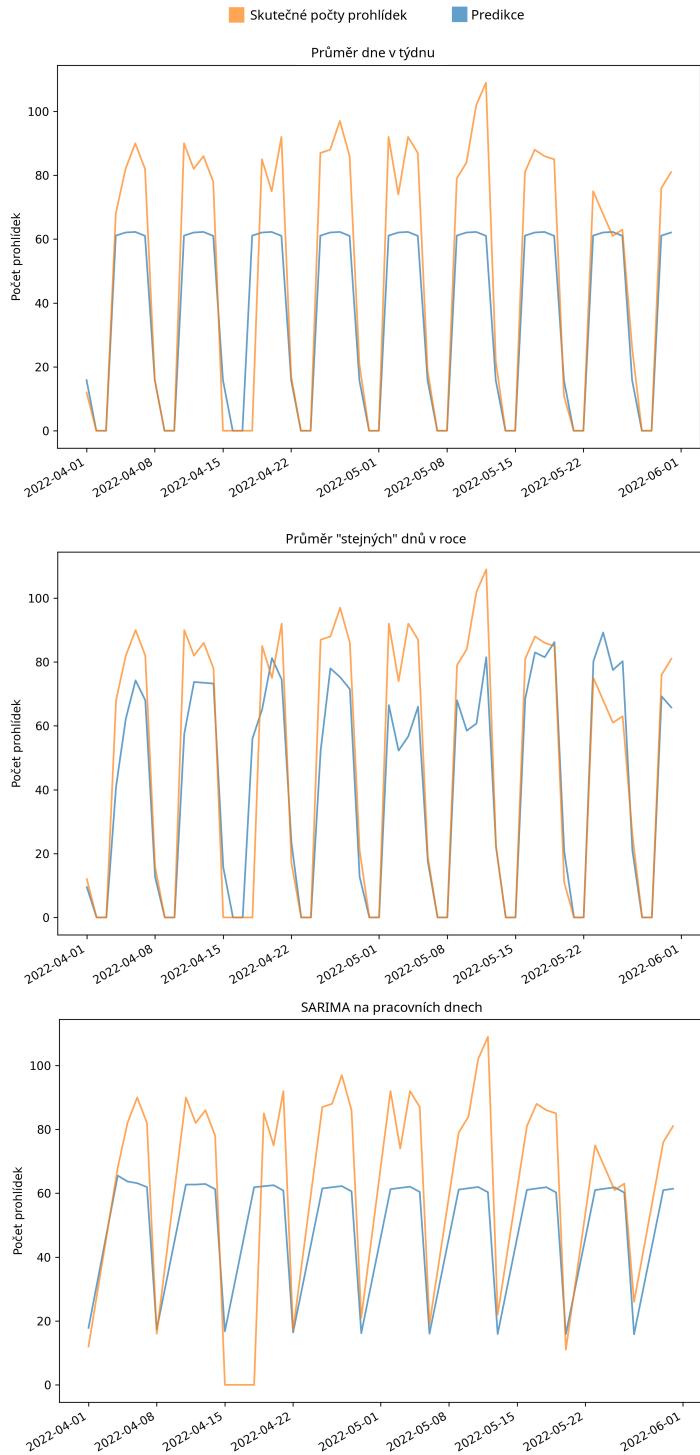
## 7.2 Predikce vytíženosti stanic

Všechny tři modely jsou evaluovány na dvou testovacích intervalech, protože data vykazují roční sezónnost, a bylo by proto neprůkazné snažit se ověřit modely pouze na jedné části roku. Porovnání predikce jednotlivých modelů pro duben–květen 2022 je vidět na obrázku 7.2.

Triviální metoda průměru dne v týdnu se zdá poměrně uvěřitelná. Týdenní sezónnost je podchycena dobře – páteční zkrácená otevírací doba je jasné vidět a víkendy jsou správně predikovány s nulovou hodnotou. Počty prohlídek v pracovních dnech jsou ale v zobrazeném dubnu a květnu systematicky podhodnoceny. Toto lze přičítat sezónnímu charakteru vývoje celé časové řady.

## 7.2. Predikce vytíženosti stanic

Obrázek 7.2: Predikce vytíženosti stanice 3102 různými modely.



## 7. VÝSLEDKY ANALÝZY

---

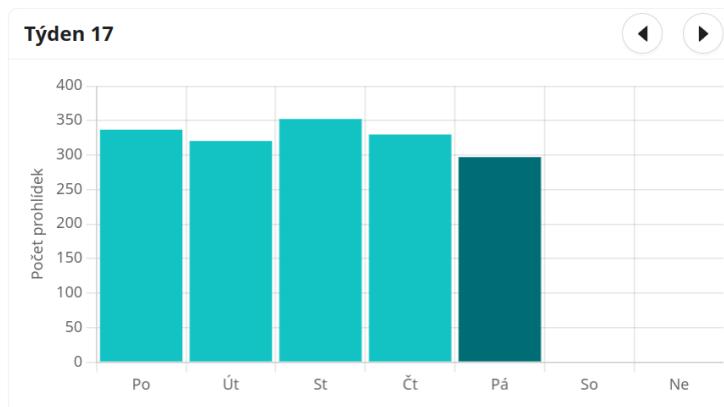
Metoda průměrování stejných dnů ve stejných týdnech napříč roky („metoda stejných dnů“) produkuje vizuálně „zajímavější“ výsledky. Pátky předpovídá opět dobře stejně jako víkendy. Počty prohlídek nejsou zatíženy systematickou chybou jako v předchozím případě, protože do predikce konkrétního dne se nezapočítají dny v jiném ročním období.

Třetí přístup, tj. SARIMA, vypadá velmi podobně jako metoda triviálního průměru, predikce je ale systematicky podhodnocená. Ani jedna metoda se samozřejmě nedokáže vypořádat se svátky (na obrázku 7.2 je znát Velký pátek 15. 4. 2022), jejichž den v týdnu se mění.

Konkrétní výsledky evaluace predikce pro dva intervaly jsou uvedeny v tabulce 7.2. V periodě leden–únor si nejlépe vede nejjednodušší přístup s MSE 82,2569. Duben až květen 2022 ale zdaleka nejlépe předpovídá metoda stejných dnů; ostatní mají vzájemně srovnatelnou kvalitu.

Z těchto výsledků lze předpokládat jednak větší robustnost metody stejných dnů, jednak může být její výstup využity také jako statistika průměrné návštěvnosti v průběhu roku. Proto je nakonec tento přístup zvolen jako nejlepší, jelikož uživateli nabízí nejvíce informací rozumné kvality. Jak je vidět na screenshotu 7.3, uživatel může procházet všechny týdny v roce, pro něž vidí průměrnou vytíženost. Graf se přitom zobrazí při načtení stránky na aktuálním týdnu se zvýrazněným dnem v týdnu pro snadnou orientaci.

Obrázek 7.3: Vizualizace predikce vytíženosti stanice na webu.



Tabulka 7.2: Evaluace predikce vytíženosti stanic.

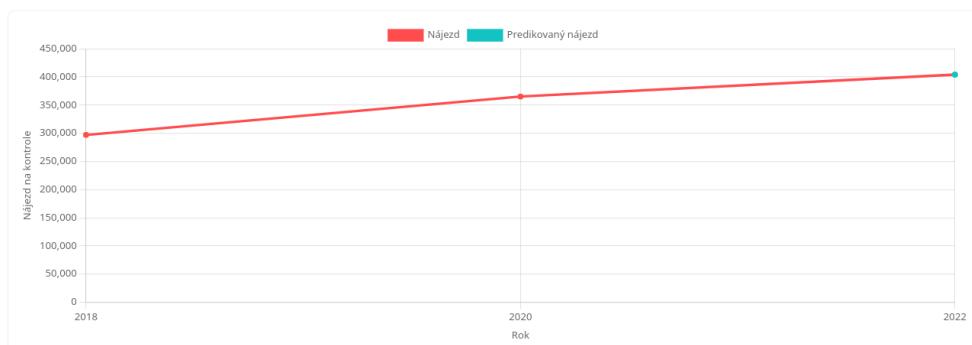
Model	MSE za leden–únor. 2022	MSE za duben–květen. 2022
Průměr dne v týdnu	85,2569	378,5891
Průměr téhož dne napříč lety	243,5802	275,2807
SARIMA	208,3789	378,1977

### 7.3 Predikce nájezdu vozidel

Získaný model je hodnocen na základě znalosti nájezdu několika vybraných vozidel běžných značek. Z pohledu na historické nájezdy vypadá predikce smysluplně, vizualizace historie s predikcí dohromady je přehledná (viz screenshot 7.4). Bohužel v případě, kdy došlo v polovině známé historie ke změně majitele, se změnil charakter užívání automobilu a predikovaný nájezd je proto podhodnocený téměř o polovinu.

Na základě testování pomocí testovacích dat jako 5 % celkového objemu podává predikce najetých kilometrů chybu MSE o hodnotě 15 077,965. Tato hodnota není příliš uspokojivá, ale lze očekávat, že s postupným doplňováním informací o historii se bude zlepšovat – méně často zastoupené značky vozidel jsou totiž v kvalitě predikce znevýhodněné.

Obrázek 7.4: Predikce nájezdu v sekci detailu vozidla.



### 7.4 Detekce anomálních kontrol

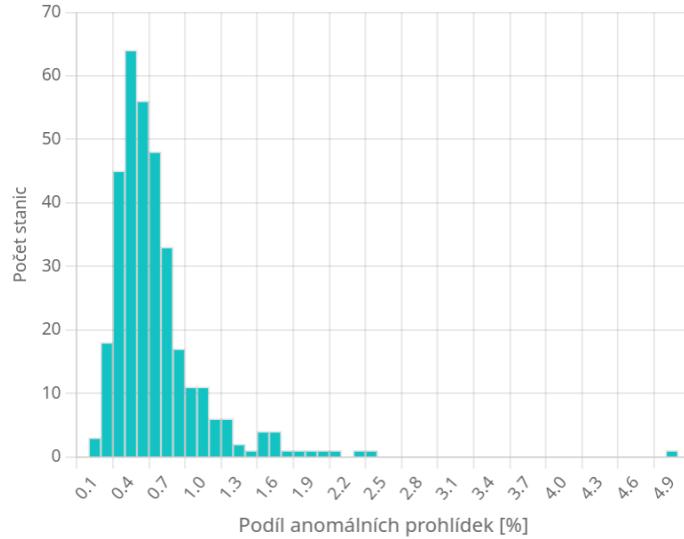
Zobrazení výsledků detekce anomalií je na webu realizováno pomocí histogramu. Ze screenshotu 7.5 je patrné, že rozdelení stanic podle podílu zjištěných anomalií na všech provedených prohlídkách se podobná normálnímu rozdělení s prodlouženým pravým chvostem. Je proto na uživateli, aby na základě histogramu interpretoval pozici konkrétní stanice a určil, zda je její chování podezřelé. Bylo by sice možné stanovit hranici a oddělit např. 90. a vyšší percentil jako podezřelé stanice, ale takové označení by bylo zcela arbitrární.

Jednotlivé typy anomalií jsou zobrazeny v histogramech podle absolutního počtu výskytů anomálie (např. screenshot 7.6). Kombinací pohledu na absolutní počty konkrétních anomalií a podíl všech anomalií na celkovém počtu prohlídek je tak možné získat přehled o podezřelém chování na stanici.

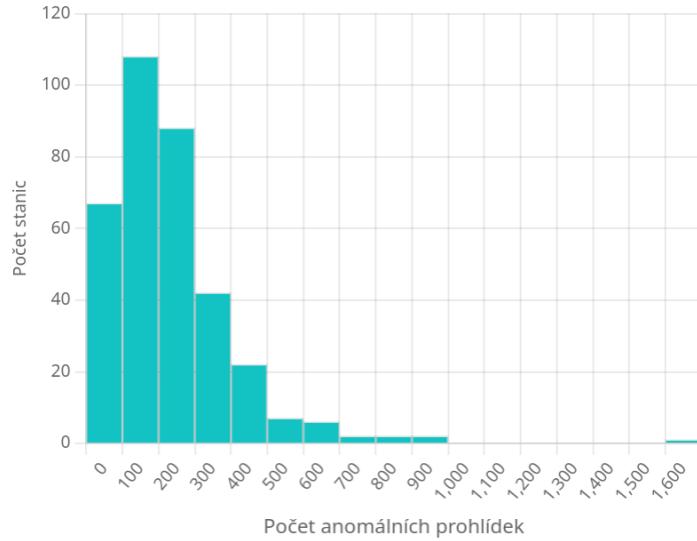
## 7. VÝSLEDKY ANALÝZY

---

Obrázek 7.5: Histogram podílu všech anomálních prohlídek na celkovém počtu prohlídek.



Obrázek 7.6: Histogram počtů opakování prohlídky s úspěchem na jiné stanici.



## 7.5 Shrnutí výsledků

Kromě výše uvedených výsledků je k dispozici velké množství dalších statistik, které se zabývají stavem českého vozového parku z pohledu registru vozidel i kontrol na STK. Jejich seznam spolu se zevrubným popisem uvádí tabulka 7.3.

## 7.5. Shrnutí výsledků

---

Název výsledku	Popis
<b>Průměrné výsledky kontrol</b>	
Počet kontrol podle výsledku	Celkový počet kontrol provedených v každém roce rozdělený podle výsledku kontroly.
Poměrný výsledek populárních značek	Percentuální podíl výsledků všech kontrol dané značky pro každý rok.
Poměrný výsledek populárních modelů	Percentuální podíl výsledků všech kontrol daného modelu pro každý rok.
<b>Závady</b>	
Nejčastější závady podle kategorie	Počet všech zjištěných závad na kontrolách pro každý rok rozdělený do kategorií (brzdy, vozek, nápravy apod.).
Nejčastější konkrétní závady	Nejčastěji zjištěné závady na všech kontrolách pro každý rok.
Nejčastější důvody neúspěšné kontroly	Pro nejpopulárnější značky (podle počtu kontrol pro každý rok) tři nejčastější závady závažnosti B nebo C, které se objevily na kontrolách, kde bylo vozidlo shledáno částečně způsobilým nebo nezpůsobilým.
Průměrný počet závad podle závažnosti	Průměrný počet závad dané závažnosti zjištěný na každé kontrole podle kraje, kde sídlí stanice provádějící kontrolu.
<b>Anomální kontroly</b>	
Podíl všech anomálních kontrol	Histogram rozdělení podílů anomálních prohlídek napříč stanicemi.
Prohlídky v nadmerně vytížených dnech	Histogram počtů anomálních prohlídek na jednotlivých stanicích.
Prohlídky s mizejícími závadami	Histogram anomalií jako výše.
Prohlídky s úspěšným opakováním na jiné stanici	Histogram anomalií jako výše.
<b>Stáří a nájezd vozidel</b>	
Průměrný věk osobních automobilů (OA)	Odhad průměrného věku osobních automobilů provozovaných pro každý rok.
Průměrný věk OA podle typu pohonu	Odhad průměrného věku osobních automobilů provozovaných pro každý rok podle jejich typu pohonu.
Průměrný nájezd kontrolovaných OA	Jak průměrně ojetá auta jsou auta při návštěvě STK v rámci celé ČR. Nejde ale o průměrný nájezd všech provozovaných vozidel, protože vozy do 4 let stáří se na STK prakticky nevyskytují.

## 7. VÝSLEDKY ANALÝZY

---

Průměrný nájezd kontrolovaných OA podle kraje	Jak průměrně ojetá auta jsou auta při návštěvě STK v každém kraji, s omezením jako výše.
<b>Značky a modely</b>	
Popularita značek	Zobrazení nejčastěji registrovaných značek s počty registrací pro každý rok a graf vývoje počtu registrací vybraných značek.
Popularita modelů	Obdobná vizualizace pro nejčastější modely.
<b>Alternativní pohony</b>	
Typ pohonu nově registrovaných vozidel	Celkový počet vozidel nově registrovaných v ČR v každém roce, rozdelený podle typu pohonu.
Elektrifikace nově registrovaných vozidel	Celkový počet vozidel nově registrovaných v ČR v každém roce, rozdelený podle jednotlivých typů elektrifikace, tj. čistě elektrická vozidla či hybridy s naftovým či benzinovým agregátem.
Celkový nájezd podle typu pohonu	Součet stavů najetých kilometrů všech vozidel, která byla pro každý rok na technické kontrole, rozdelený podle typu pohonu vozidel.
<b>Import ojetých vozidel</b>	
Průměrné stáří ojetin při importu	Za importovaný je považován automobil, jehož rozdíl v obecné první registraci a první registraci v ČR je alespoň 365 dní. Hodnota v grafu udává průměrný věk všech takovýchto osobních automobilů, které byly v daném roce poprvé registrované v ČR.
Poměr nových a importovaných ojetých OA	Vývoj poměru nových a importovaných ojetých vozidel na nových registracích v každém roce s využitím stejné metodiky pro určení importu jako výše.
<b>Zajímavosti</b>	
Podíl barev nově registrovaných vozidel	Vývoj podílu jednotlivých barev karoserie na registracích v každém roce.
Status vozidel podle data registrace	Podíl stavů vozidel podle jejich data první registrace v ČR platný k datu získání exportu registru vozidel.

Tabulka 7.3: Seznam výsledků analýzy.

# KAPITOLA 8

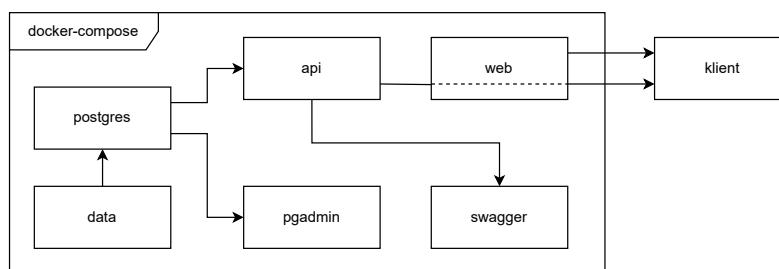
## Softwarový projekt

Tato kapitola popisuje implementaci softwarového celku, který řídí zpracování, uchování a prezentaci dat. Představuje podstatné vlastnosti jednotlivých součástí a řešení problémů, které při se implementaci objevily.

### 8.1 Infrastruktura

Infrastruktura složená jako Docker compose projekt odpovídá diagramu na obrázku 8.1, kde šipky zobrazují směr toku dat. Základem je kontejner `postgres`, do něhož ukládá zdrojová data a výsledky výpočtu datová pipeline v kontejneru `data`. Služba `web` obsahuje Next.js server pro poskytování webového frontendu klientům. Z klienta jsou pak posílány API požadavky zpět na server, který funguje jako proxy (přerušovaná čára) pro PostgREST běžící v kontejneru `api`. Pro inspekci dat administrátorem přímo z databáze, resp. z API, slouží služby PgAdmin [52], resp. SwaggerUI [53] ve stejnojmenných kontejnerech.

Obrázek 8.1: Diagram struktury Docker compose projektu.



Databázi PostgreSQL je nutné inicializovat tak, aby obsahovala role umožňující zabezpečený přístup PostgREST serveru k datům. Toho je docíleno pomocí skriptu, jež je využitý kontejnerem automaticky při jeho prvním spuštění, když je databáze prázdná. Do skriptu jsou pomocí Docker compose předány

environment variables s hesly, jež je třeba nastavit. Všechny citlivé údaje a konfigurace např. portů, kde mají být služby dostupné, se předávají stejným způsobem i do samotného Docker compose souboru. Takto je docíleno bezpečného a snadno konfigurovatelného nasazení na server.

## 8.2 Pipeline

Datové operace popsané výše jsou implementovány jako Python modul tvořící pipeline, jejíž strukturu popisuje obrázek 8.2 (šipky naznačují pořadí volání modulů). Tato pipeline je navržena pro opakovaná spouštění, mezi nimiž jsou doplněna data (přidány kontroly na STK a aktualizovány ostatní datové zdroje).

Obrázek 8.2: Diagram struktury pipeline modulu.

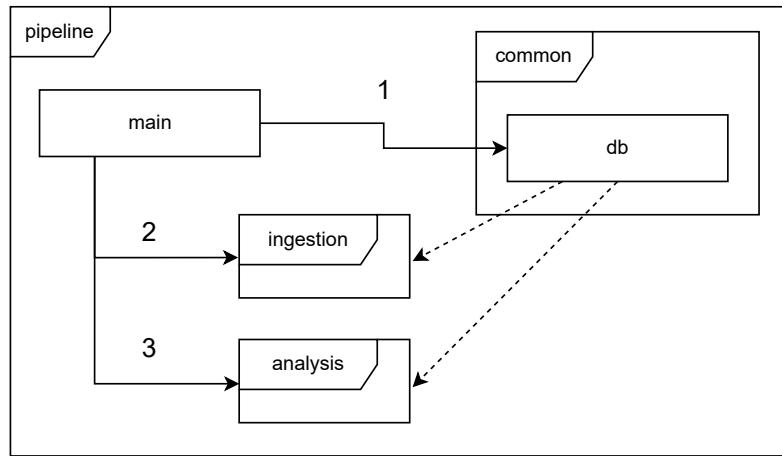
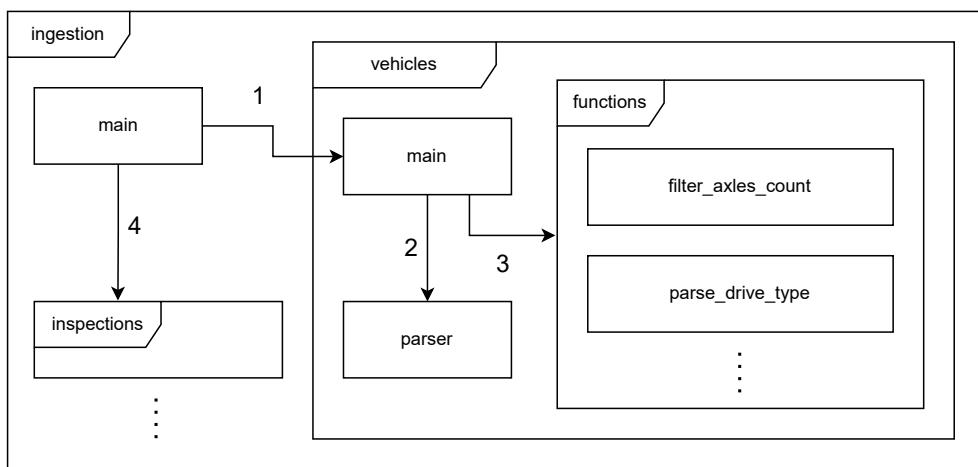


Image s modulem je sestaven pomocí jednoduchého `Dockerfile` založeného na oficiálním Python 3.11 image [54]. Na tomto základu jsou doinstalovány knihovny pro přístup k databázi a další Python závislosti. Nakonec se do image kopíruje samotný kód. Aby uživatel při spuštění kontejneru vždy viděl živý výstup programu, je také třeba vypnout buffering textového výstupu v Pythonu nastavením speciální environment variable. Pokud není nastavena, výstup se někdy i na několik minut zastaví a zdá se, že proces selhal.

Pipeline je rozdělena na fázi datového importu a analýzy. Import implementovaný v podmodulu `ingestion` spočívá v načtení souborů s datovými zdroji, předzpracování a začištění dat a jejich uložení do databáze. Soubory se načítají z disku hostujícího počítače, program dostává jejich umístění stejně jako parametry pro připojení k databázi prostřednictvím environment variables. To umožňuje spouštět pipeline i lokálně mimo Docker pro účely testování. Obě fáze pipeline využívají třídu v podmodulu `common`, která se stará o připojení k databázi a poskytuje metody pro přístup k ní.

Podmodul `ingestion` je členěn podle datových zdrojů, o jejichž import se stará; struktura je naznačena na obrázku 8.3. Skript `parser` načítá data do Pandas `DataFrame` a následně jsou spouštěny jednotlivé transformace definované v podmodulu `functions`. Každá operace akceptuje jako argument `DataFrame` a vrací jeho upravenou kopii. Některé funkce mají dodatečné argumenty, které jsou potřeba např. pro předání globálních statistik v případě, kdy transformace probíhají po částech kvůli paměťové náročnosti.

Obrázek 8.3: Diagram struktury podmodulu pro import dat v pipeline.



Registr vozidel stejně jako číselníky stanic a závad se v databázi vždy přeší novým zdrojovým souborem – v případě číselníků se jedná o velmi malé množství dat, takže není třeba import více optimalizovat. Do aktualizovaného registru vozidel typicky přibudou nová vozidla, takže by na první pohled mohlo stačit pouze doplnit tento rozdíl. Jelikož se ale u starších vozidel mohl změnit jejich stav (např. pokud vůz zanikl), registr je třeba importovat také celý znova. Inkrementální import pouze chybějících záznamů je možný jen v případě datové sady kontrol na STK.

Podmodul `analysis` se nakonec stará o zpracování importovaných dat. Aby byly výsledky zapsané do databáze čitelné prostřednictvím API, je nutné ke každé zapisované tabulce nastavit příslušná přístupová práva. Díky tomu lze v budoucnu doplnit další výpočet nově požadovaného výstupu pouhým přidáním jednoho Python souboru a není třeba udržovat definici schématu na odděleném místě.

## 8.3 Webová aplikace

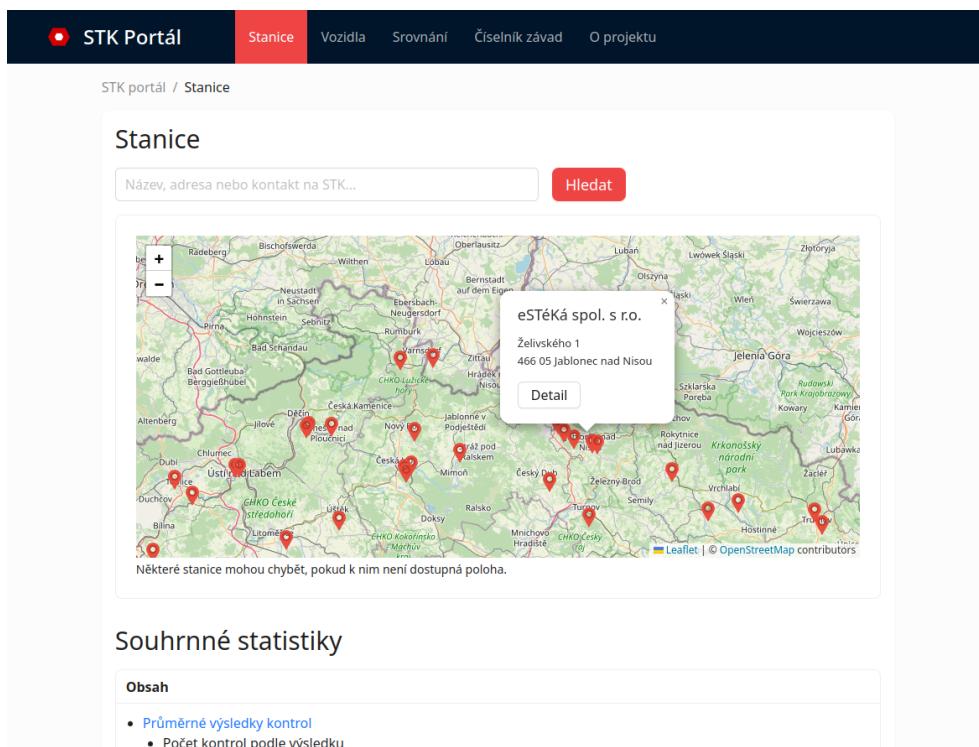
Kontejnerizace webové aplikace je zajištěna pomocí oficiálního vzoru dostupného ve veřejném repozitáři Next.js [55]. Do vzorového `Dockerfile` je doplněno předání argumentu udávajícího adresu API serveru, aby aplikační server

## 8. SOFTWAROVÝ PROJEKT

věděl, kam přesměrovávat API požadavky. Adresu API proto nelze změnit pouhou úpravou environment variables v Docker compose konfiguraci – image kontejneru je při změně nutné znovu sestavit.

Webová aplikace se skládá ze stránek se shodnou strukturou sestávající z navigační lišty, navigačního pásu zobrazujícího aktuální pozici ve stromu stránek, samotného obsahu a patičky. Pro efektivní vývoj je tato kostra definována jako sada React komponent, které se využívají v několika šablonách pro každou ze sekcí portálu. Náhled této struktury je vidět na obrázku 8.4, který ukazuje začátek sekce o stanicích. Podobným způsobem je strukturována sekce o vozidlech. Domovská stránka pak obsahuje krátký textový úvod a několik vybraných grafů, které nastiňují rozsah dostupných dat.

Obrázek 8.4: Výřez sekce o stanicích na webu.

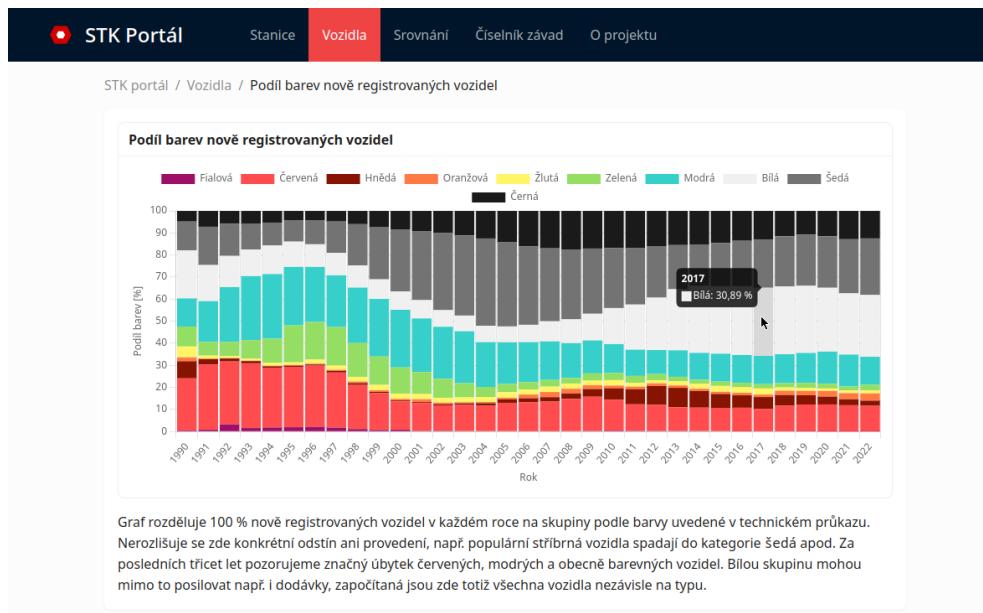


Majoritu stránek tvoří detailní zobrazení jednotlivých analýz s jejich popisem. Jednou z takových je např. stránka o podílu barev nově registrovaných vozidel, jejíž náhled je k dispozici na obrázku 8.5. Všechny grafy dovolují zobrazit přesnou hodnotu datového bodu při najetí myší a skrýt libovolnou datovou sadu kliknutím na její výčet v legendě. Zásadním prvkem je popis grafu, který uživateli vysvětluje způsob čtení grafu a upozorňuje před možnými úskalími v interpretaci prezentovaných informací.

Velmi důležitým prvkem v prezentaci informací vizuální formou je barva. Dobře zvolená barevná paleta působí na příjemce dojmem důvěryhodnosti

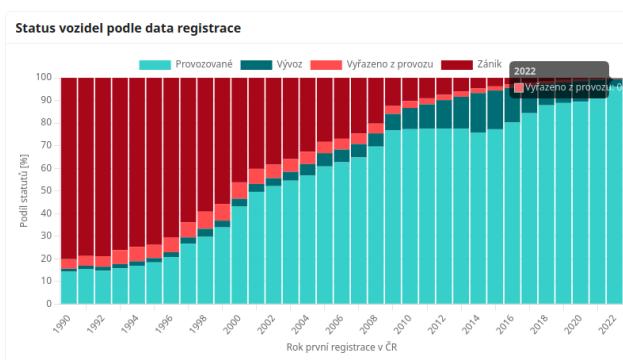
### 8.3. Webová aplikace

Obrázek 8.5: Detail analýzy podílů barev nově registrovaných vozidel na webu.



a usnadňuje čtení poskytovaných dat mimo jiné tím, že na sebe např. přílišnou křiklavostí nepřitahuje zbytečnou pozornost a nechává především mluvit data samotná. Pro celý web je jako primární barva zvolena červená, která významově koresponduje s barvou kontrolních nálepek vylepovaných na registrační značky při technických kontrolách. Jako doplňková barva se k ní přidává azurová. Jelikož v mnoha grafech je nutné zobrazit i čtyři různé skupiny dat (např. jako na obrázku 8.6), k těmto dvěma základním barvám se přidávají ještě jejich utlumené varianty s nižší světlostí. Celkový dojem je tedy harmonický, obě základní barvy k sobě dobře pasují a jejich utlumené varianty se při čtení grafů snadno odlišují, aniž by narušovaly jednotný dojem z celé prezentace. [56]

Obrázek 8.6: Graf stavu vozidel podle data registrace na webu.



## **8. SOFTWAROVÝ PROJEKT**

---

Veškeré vizualizace a doprovodné texty k nim jsou otestovány na svou přístupnost laickému uživateli na několika dobrovolnících z odlišných oborů. Testování je provedeno jednak pozorováním pohybu neznámého uživatele na webu, jednak řízeným rozhovorem s účelem ověřit především správnou interpretaci grafu na základě jeho popisu. Prohlížení webu je přitom otestováno jak na stolním počítači, tak na několika mobilních telefonech, aby byl naplněn požadavek responzivity. Tím je docíleno uživatelské přívětivosti webového portálu a jednoznačnosti poskytovaných informací.

# KAPITOLA 9

---

## Závěr

Hlavním výstupem práce je webový portál poskytující široké veřejnosti nový pohled na data poskytovaná Ministerstvem dopravy ČR. Uživatelsky přívětivý web dostupný na adrese <https://stk.opendatalab.cz> se přizpůsobí počítačové i mobilní obrazovce a umožňuje tak zkonto rolovat si historii a parametry ojetého vozidla kdekoliv v terénu. Data jsou přitom dostupná nejen ve formě grafického rozhraní, ale také pomocí strojově čitelné API.

Práce přináší náhled na aktuální situaci na trhu služeb pro kontrolu ojetých vozů a zjišťuje slabiny v prezentaci poskytovaných informací. Jedná se např. o problémy s chybějícím strukturováním dat, která jsou dostupná pouze ve formě PDF dokumentů vyžadujících ruční analýzu, či o nevhodné vizualizace, které skrývají přesné údaje. Na základě těchto nedostatků je navržen portál, který data zobrazuje s důrazem na detail a správnou interpretaci uživatelem prostřednictvím doplňujících vysvětlivek. Jeho součástí jsou kromě výpisů informací o konkrétních vozidlech také detailní statistiky o kontrolách vozidel na STK a vozovém parku ČR obecně.

Přínosem práce je rovněž podrobná analýza dat poskytovaných Ministerstvem dopravy a implementace systému pro jejich začištění a konverzi do strukturované a dále zpracovatelné reprezentace. Na základě těchto dat jsou navrženy metody analýzy a výsledkem je softwarový celek, který různými metodami získává více než 30 různých výsledků, které jsou následně prezentovány na webu. Mezi analýzy využívající metody strojového učení patří například predikce závad, které by se v budoucnu na konkrétním voze mohly vyskytnout. Podobné výstupy pak doplňují webový výpis informací o každém vozidle a STK.

Podstatnou část webové prezentace tvoří statistiky srovnávající časový vývoj veličin popisujících vozový park. Průměrné stáří osobních automobilů se například ukazuje být pro rok 2022 jen cca 13,8 roku na rozdíl od běžně uváděných 15,9. Tento výsledek může být zapříčiněn technikou odhadu množiny aktuálně provozovaných vozidel, který bylo nutné provést, protože datum ukončení provozu není z dat dostupné. Relativní nízkost průměrného věku ale

## 9. ZÁVĚR

---

může být také způsobena tím, že v použitém registru vozidel nejsou zahrnutý veterány a některá další vozidla, která by průměr značně vychýlila směrem nahoru.

Jiným naopak nepřekvapivým výsledkem je, že průměrný nájezd vozidla v Praze je nižší než ve zbytku Česka, a to dokonce téměř o 12 % než v případě Moravskoslezského kraje, jež je v tomto žebříčku druhý. Zajímavý je také celkový nájezd všech naftových vozidel kontrolovaných na STK v roce 2022, který je se svými 343 biliony kilometry dvakrát vyšší než nájezd vozidel na benzin. Jednoduchým vysvětlením by přitom mohlo být, že nákladní vozidla by využívala v mnohem vyšší míře naftový pohon a jejich kumulativní nájezd by byl zřejmě proti těm osobním značně významnější. S menším nájezdem pražských vozidel souvisí také jejich, oproti některým krajům i trojnásobně, menší průměrný počet zjištěných závad na prohlídkách. Oba tyto výsledky korespondují s vyšší ekonomickou úrovni centra České republiky.

Portál nabízí rovněž srovnání jednotlivých značek a modelů vozidel. Anekdotální dojmy některých řidičů tak nabývají exaktnějších rysů, když se například ukazuje, že vozy značky Opel neuspějí při technické kontrole o 40 % častěji než zástupci nejpopulárnější Škody. Za zmínku stojí též nejčastější důvody neúspěšné kontroly. Zatímco u Škody se jedná o problémy s účinností parkovací brzdy, majitelé vozů značky Volkswagen nejčastěji pohoří na neoprávněných úpravách, které nejsou správně zapsány v technické dokumentaci vozu. Závadou s nejvyšší četností obecně se ukazuje být povrchová koroze různých součástí vozidla, v roce 2022 se objevila u téměř 1,2 milionu kontrol. Tyto výsledky přinášejí jasně kvantifikovaný pohled na spolehlivost různých značek a mohou tak posloužit spotřebitelům při jejich nákupním rozhodnutí.

Rozsahem dostupných informací při kontrole vozu se webový portál nevyrovná komerčním nabídkám zejména kvůli absenci dat o pojistných událostech, ale spojuje zdarma několik různých zdrojů, které by si jinak každý zájemce musel složitě obstarat. V kombinaci s datovou analýzou na základě strojového učení se ale rozsah informací rozšiřuje a *STK portál* se tak stává smysluplnou alternativou k aktuálně dostupným službám.

---

## Literatura

- [1] Svaz dovozců automobilů: Přehled stavu vozového parku. [online], 2023, [cit. 2024-05-07]. Dostupné z: <https://portal.sda-cia.cz/stat.php?v#rok=2023&mesic=12&kat=stav&vyb=&upr=&obd=m&jine=false&lang=CZ&str=vpp>
- [2] Parkhomenko, A.: *Portál Výsledků Analýzy Dat a Dalších Informací o STK*. Bakalářská práce, České vysoké učení technické v Praze. Fakulta informačních technologií, Praha, 2020. Dostupné z: <http://hdl.handle.net/10467/88156>
- [3] Commission Implementing Regulation (EU) 2021/535 of 31 March 2021 Laying down Rules for the Application of Regulation (EU) 2019/2144 of the European Parliament and of the Council as Regards Uniform Procedures and Technical Specifications for the Type-Approval of Vehicles, and of Systems, Components and Separate Technical Units Intended for Such Vehicles, as Regards Their General Construction Characteristics and Safety (Text with EEA Relevance), [online]. [cit. 2024-04-09]. Dostupné z: [http://data.europa.eu/eli/reg\\_impl/2021/535/oj/eng](http://data.europa.eu/eli/reg_impl/2021/535/oj/eng)
- [4] autoDNA: VIN číslo, [online]. [cit. 2024-04-09]. Dostupné z: <https://www.autodna.cz/vin-cislo>
- [5] ČESKO: Zákon č. 56/2001 Sb., o podmínkách provozu vozidel na pozemních komunikacích - znění od 1. 4. 2024, [online]. [cit. 2024-04-08]. Dostupné z: <https://www.zakonyprolidi.cz/cs/2001-56>
- [6] ČESKO: Vyhláška č. 211/2018 Sb., o technických prohlídkách vozidel - znění od 1. 1. 2024, [online]. [cit. 2024-04-08]. Dostupné z: <https://www.zakonyprolidi.cz/cs/2018-211>
- [7] ČESKÝ ROZHLAS: STK 2018. [online], [cit. 2024-04-05]. Dostupné z: <https://data.gov.cz/datov%C3%A1-sada?iri=https%>

## LITERATURA

---

- 3A%2Fdata.gov.cz%2Fzdroj%2Fdatov%C3%A9-sady%2Fhttps---un7pp4qfr5.execute-api.eu-west-1.amazonaws.com-prod-package\_show-id-stk\_md\_2018
- [8] Ministerstvo vnitra ČR: Portál o Datech České Republiky. [online], [cit. 2024-04-04]. Dostupné z: <https://data.gov.cz/>
  - [9] Ministerstvo dopravy ČR: STK - Seznam STK podle krajů. [online], [cit. 2024-04-05]. Dostupné z: <https://www.mdcr.cz/Dokumenty/Silnicni-doprava/STK/STK-Seznam-STK-dle-kraju?returl=/Dokumenty/Silnicni-doprava/STK>
  - [10] Ministerstvo dopravy ČR: Statistiky pro výpočty kapacit. [online], [cit. 2024-04-05]. Dostupné z: <https://www.mdcr.cz/Statistiky/Silnicni-doprava/STK/Statistiky-pro-vypocty-kapacit>
  - [11] Cebia.cz. [online], [cit. 2024-04-06]. Dostupné z: <https://www.cebia.cz>
  - [12] Novinky, Cebia: Cebia prověřuje i auta ze zahraničí. Kde bere informace?, [online]. 2021, [cit. 2024-04-07]. Dostupné z: <https://www.novinky.cz/clanek/komerjni-clanky-cebia-proveruje-i-auta-ze-zahranici-kde-bere-informace-40374201>
  - [13] Šidlák, M.: Pět let bezzubého zákona. Fabia stočená o 760 tisíc kilometrů, bez postihu, [online]. 2023, [cit. 2024-04-07]. Dostupné z: [https://www.idnes.cz/auto/zpravodajstvi/staceni-tachometru-cebia-pajer.A230919\\_111104\\_automoto\\_fdv](https://www.idnes.cz/auto/zpravodajstvi/staceni-tachometru-cebia-pajer.A230919_111104_automoto_fdv)
  - [14] Vindecoder.eu. [online], [cit. 2024-04-08]. Dostupné z: <https://www.vindecoder.eu>
  - [15] Ramírez Pérez, S. M.: Euro 7 Motor Vehicle Emission Standards, [online]. Listopad 2023, [cit. 2024-04-08]. Dostupné z: [https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/754573/EPRS\\_ATA\(2023\)754573\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/754573/EPRS_ATA(2023)754573_EN.pdf)
  - [16] Halleux, V.: Development of Euro 7 Emission Standards for Cars, Vans, Lorries and Buses | Legislative Train Schedule, [online]. 2024, [cit. 2024-04-08]. Dostupné z: <https://www.europarl.europa.eu/legislative-train/theme-a-european-green-deal/file-post-euro6vi-emission-standards>
  - [17] Carvertical.com. [online], 2024, [cit. 2024-04-08]. Dostupné z: <https://www.carvertical.com>
  - [18] Ministerstvo dopravy ČR: Datová kostka. [online], [cit. 2024-04-09]. Dostupné z: <https://www.dataovozidlech.cz>

- [19] Ministerstvo dopravy ČR: Data o vozidlech – Poskytnuté informace. [online], 2023, [cit. 2024-04-09]. Dostupné z: [https://www.mdcr.cz/Ministerstvo/Zadost-o-poskytnuti-informace-\(1\)/Poskytnute-informace/Data-o-vozidlech](https://www.mdcr.cz/Ministerstvo/Zadost-o-poskytnuti-informace-(1)/Poskytnute-informace/Data-o-vozidlech)
- [20] Ministerstvo dopravy ČR: Kontrola tachometru. [online], [cit. 2024-04-09]. Dostupné z: <https://www.kontrolatachometru.cz>
- [21] Svaz dovozců automobilů. [online], 2024, [cit. 2024-04-10]. Dostupné z: <https://portal.sda-cia.cz>
- [22] Bartos015: STK Stanice technické kontroly a emise. [online], [cit. 2024-04-10]. Dostupné z: <https://www.stanice-technicke-kontroly.cz>
- [23] Agentura Kryštof s.r.o.: Seznam-STK. [online], [cit. 2024-04-10]. Dostupné z: <http://www.seznam-stk.cz>
- [24] YData contributors: ydata-profiling. [software], 2024, [cit. 2024-04-14]. Dostupné z: <https://github.com/ydataai/ydata-profiling>
- [25] The Pandas Development Team: Pandas. [software], 2024, [cit. 2024-04-14]. Dostupné z: <https://pandas.pydata.org>
- [26] Microsoft Learn contributors: datetime (Transact-SQL). [online], 2022, [cit. 2024-04-14]. Dostupné z: <https://learn.microsoft.com/en-us/sql/t-sql/data-types/datetime-transact-sql?view=sql-server-ver16>
- [27] Ministerstvo dopravy ČR: Příloha 18 metodiky 1/2012-150-METO/1. [online], [cit. 2024-04-14]. Dostupné z: [https://www.mdcr.cz/getattachment/Dokumenty/Silnicni-doprava/Schvalovani-vozidel/Metodiky/1\\_2012150METO\\_2.pdf.aspx?lang=cs-CZ](https://www.mdcr.cz/getattachment/Dokumenty/Silnicni-doprava/Schvalovani-vozidel/Metodiky/1_2012150METO_2.pdf.aspx?lang=cs-CZ)
- [28] ČESKO: Vyhláška č. 153/2023 Sb., o schvalování technické způsobilosti vozidel a technických podmínek provozu vozidel na pozemních komunikacích, [online]. 2023, [cit. 2024-04-15]. Dostupné z: <https://www.zakonyprolidi.cz/cs/2023-153>
- [29] Osisanwo, F.; Akinsola, J.; Awodele, O.; aj.: Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, ročník 48, č. 3, 2017: s. 128–138.
- [30] Chatterjee, S.; Hadi, A.: *Regression Analysis by Example*. Wiley Series in Probability and Statistics, Wiley, 2015, ISBN 9781119122739. Dostupné z: <https://books.google.cz/books?id=zyjWBgAAQBAJ>
- [31] Dorogush, A. V.; Ershov, V.; Gulin, A.: CatBoost: gradient boosting with categorical features support. 2018, 1810.11363. Dostupné z: <http://arxiv.org/abs/1810.11363>

## LITERATURA

---

- [32] Prokhorenkova, L.; Gusev, G.; Vorobev, A.; aj.: CatBoost: unbiased boosting with categorical features. [online], 2017, 1706.09516. Dostupné z: <https://arxiv.org/abs/1706.09516>
- [33] Han, J.; Kamber, M.; Pei, J.: *Data Mining: Concepts and Techniques*. Elsevier Science & Technology, ISBN 978-0-12-381480-7. Dostupné z: <http://ebookcentral.proquest.com/lib/techlib-ebooks/detail.action?docID=729031>
- [34] Boukerche, A.; Zheng, L.; Alfandi, O.: Outlier Detection: Methods, Models, and Classification. *ACM Comput. Surv.*, ročník 53, č. 3, jun 2020, ISSN 0360-0300, doi:10.1145/3381028.
- [35] Chen, Z.; Yeo, C. K.; Lee, B. S.; aj.: Autoencoder-based network anomaly detection. In *2018 Wireless Telecommunications Symposium (WTS)*, 2018, s. 1–5, doi:10.1109/WTS.2018.8363930.
- [36] Antwarg, L.; Shapira, B.; Rokach, L.: Explaining Anomalies Detected by Autoencoders Using SHAP. *CoRR*, ročník abs/1903.02407, 2019, 1903.02407. Dostupné z: <http://arxiv.org/abs/1903.02407>
- [37] Hyndman, R. J.; Athanasopoulos, G.: *Forecasting: principles and practice*. Melbourne, Australia: OTexts, třetí vydání, 2021, [cit. 2024-04-19]. Dostupné z: <https://otexts.com/fpp3/>
- [38] Matthias, K.; Kane, S. P.: *Docker - Up and Running*. Sebastopol, CA: O'Reilly Media, 2015, ISBN 9781491917572.
- [39] Git community: Git. [software], 2024, [cit. 2024-04-22]. Dostupné z: <https://git-scm.com>
- [40] Jupyter community: Project Jupyter. [software], 2024, [cit. 2024-04-20]. Dostupné z: <https://jupyter.org>
- [41] The PostgreSQL Global Development Group: PostgreSQL. [software], 2024, [cit. 2024-04-22]. Dostupné z: <https://www.postgresql.org>
- [42] Joe Nelson, S. C.: PostgREST Documentation. [online], 2024, [cit. 2024-04-22]. Dostupné z: <https://postgrest.org/en/stable/>
- [43] Docker Hub: Postgres – Official Image. [software], 2024, [cit. 2024-04-22]. Dostupné z: [https://hub.docker.com/\\_/postgres](https://hub.docker.com/_/postgres)
- [44] Docker Hub: postgrest/postgrest – Docker Image. [software], 2024, [cit. 2024-04-22]. Dostupné z: <https://hub.docker.com/r/postgrest/postgrest/>
- [45] Vercel: Next.js Documentation. [online], 2024, [cit. 2024-04-22]. Dostupné z: <https://nextjs.org/docs>

- [46] Meta Open Source: React. [software], 2024, [cit. 2024-04-24]. Dostupné z: <https://react.dev>
- [47] Ant Group, Ant Design Community: Ant Design. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://ant.design>
- [48] Chart.js contributors: Chart.js. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://www.chartjs.org>
- [49] Agafonkin, V.: Leaflet. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://leafletjs.com>
- [50] Xml.Etree.ElementTree — The ElementTree XML API. [online], 2024, [cit. 2024-04-20]. Dostupné z: <https://docs.python.org/3/library/xml.etree.elementtree.html>
- [51] Shlens, J.: A Tutorial on Principal Component Analysis. doi: 10.48550/arXiv.1404.1100, 1404.1100. Dostupné z: <http://arxiv.org/abs/1404.1100>
- [52] pgAdmin community: pgAdmin. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://www.pgadmin.org>
- [53] SmartBear Software: Swagger UI. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://swagger.io/tools/swagger-ui/>
- [54] Python Docker Official Image. [software], 2024, [cit. 2024-05-07]. Dostupné z: [https://hub.docker.com/\\_/python/](https://hub.docker.com/_/python/)
- [55] Vercel: Next.js with Docker. [software], 2024, [cit. 2024-05-07]. Dostupné z: <https://github.com/vercel/next.js/tree/canary/examples/with-docker>
- [56] Muth, L. C.: How to pick more beautiful colors for your data visualizations. [online], 2024, [cit. 2024-05-07]. Dostupné z: <https://blog.datawrapper.de/beautifulcolors/>



## **Seznam použitých zkratek**

**ACF** Auto-correlation Function

**ADF** Augmented Dickey-Fuller test

**AIC** Akaike Information Criterion

**API** Application Programming Interface

**ARIMA** Autoregressive Integrated Moving Average

**ARMA** Autoregressive Moving Average

**CSV** Comma Separated Values

**DBSCAN** Density-Based Spatial Clustering of Applications with Noise

**ISO** International Organization for Standardization

**MDČR** Ministerstvo dopravy České republiky

**MSE** Mean Squared Error

**NKOD** Národní katalog otevřených dat

**OA** osobní automobil

**OEM** Original Equipment Maker

**ORP** obec s rozšířenou působností

**PACF** Partial Auto-correlation Function

**PCA** Principal Component Analysis

**PDF** Portable Document Format

## A. SEZNAM POUŽITÝCH ZKRATEK

---

**SARIMA** Seasonal Autoregressive Integrated Moving Average

**SHAP** Shapley Additive Explanations

**STK** stanice technické kontroly

**TP** technický průkaz

**WMI** World Manufacturer Identifier

**XML** Extensible Markup Language

**VIN** Vehicle Identification Number

**VDS** Vehicle Description Section

**VIS** Vehicle Indicator Section

**ZTP** základní technický popis

# PŘÍLOHA **B**

---

## Obsah digitální přílohy

```
readme.txt.....stručný popis obsahu přílohy
exploratory_analysis.....výsledky exploratorní analýzy
    inspections.....kontroly na STK
    vehicles.....registrování vozidel
jupyter.....Jupyter notebooky s přípravou analýzy
    clustering.....analýza flotil
    inspections_outlier_detection.....detekce anomálních kontrol
    stations_weekly_occupancy.....predikce vytíženosti stanic
    vehicles_defect_prediction.....predikce závad na vozidlech
    vehicles_mileage_prediction.....predikce nájezdu
    vehicles_statistics.....statistiky vozového parku
src.....zdrojové kódy implementace
    data.....datová pipeline
    web.....webová aplikace
text.....zdrojový kód textu práce ve formátu LATEX
```