Zkušební okruhy ke státní závěrečné zkoušce KMA/SZZ2 Statistika a zpracování dat

Obecné informace

Státní závěrečná zkouška z předmětu KMA/SZZ2 *Statistika a zpracování dat* má dvě části:

- přípravná (praktická) část
  - Student za účasti alespoň dvou členů zkušební komise realizuje náhodný výběr jednoho okruhu z příslušného souboru zkušebních okruhů, viz typové úlohy níže.
  - Student má k dispozici 8 hodin na řešení zadané úlohy (rozdělených na promýšlení/ analýzu, návrh, vývoj, implementaci a tvorbu dokumentace/ reportu).
  - Termín přípravné části je stanoven 3 až 10 dní před závěrečnou ústní částí.
  - Funkční řešení musí být implementováno prostřednictvím technologií poskytnutých katedrou matematiky. Používání vlastních technologií (hardware, software) je však při řešení přípustné.
  - Během řešení lze používat libovolné dostupné materiály a zdroje, nicméně je přísně zakázáno využívat pomoci jiných osob nebo s nimi jakkoliv spolupracovat.
  - Výsledek přípravné části zhodnotí před konáním závěrečné ústní části vybraní odborníci. Zkušební komisi podají souhrnné hodnocení a připraví pro ni otázky. Student nebude s hodnocením ani otázkami seznámen.
- závěrečná ústní část (ve stejný den jako ústí zkouška KMA/SZZ1 *Matematika s aplikacemi*): 30 minut (prezentace zadání a řešení, odpovědi na otázky zkušební komise bez přípravy)

Zadání úloh

Jde o typové úlohy, které naznačují rozsah dovedností, které by měly být při řešení aplikovány. Skutečné zadání může být jiné.

Okruh A – zpracování dat

Cílem úlohy je zpracovat a analyzovat data z on-line obchodu, která obsahují informace o objednávkách, produktech a zákaznících. Úkolem je připravit data (pokročilé manipulace s daty), provést explorační analýzu a identifikovat klíčové vztahy. Výstupem bude **komentovaný report** obsahující zpracování a analýzu vstupních dat.

1) **Načtení a příprava dat**: Načtěte data z CSV souborů `objednavky.csv`, `produkty.csv` a `zakaznici.csv` do prostředí R/Pythonu. Každý z těchto souborů obsahuje různé typy dat, včetně číselných, faktorových, textových (řetězce) a logických. Zkontrolujte kvalitu dat a proveďte úpravy, jako je odstranění chybějících hodnot a duplicit, převedení proměnných do vhodných datových typů.

2) **Manipulace s daty**: Slučte tabulky `objednavky`, `produkty` a `zakaznici` podle relevantních klíčů. Získejte přehled o počtu objednávek, prodejích a zákaznících podle kategorií produktů, regionů a časových období pomocí operací filtrování, seskupování a sumarizace.

3) **Práce s daty v dlouhém a širokém formátu**: Převeďte data do dlouhého formátu (například jestliže na vstupu budou data v širokém formátu o měsíčních prodejích produktů ve sloupcích pro každý měsíc, pak může být užitečné je převést do tří sloupců, tj. dlouhého formátu, tak, že v každém řádku bude trojice produkt–měsíc–prodeje), proveďte potřebné analýzy a následně je převeďte zpět do širokého formátu.

4) **Práce s textem a regulárními výrazy**: Každý zákazník má v databázi uloženou emailovou adresu. Vaším úkolem je rozdělit tuto emailovou adresu na dvě části: před zavináčem a za ním a tyto informace uložit do dvou nových sloupců `email_name` a `email_domain`. Dále, některé z produktových kategorií mají v názvu zeměpisné označení (např. „Káva – Kolumbie“). Vytvořte nový sloupec `product_country`, který bude obsahovat zeměpisné označení extrahované z názvu produktu. Pokud produkt žádné zeměpisné označení nemá, sloupec by měl obsahovat hodnotu NA.

5) **Explorační analýza dat a vizualizace**: Identifikujte a vizualizujte klíčové vztahy v datech. Konkrétně vytvořte soubor souvisejících grafů (grafy sdílející stejné osy a jsou rozdělené na základě jedné nebo více kategorií) pro zobrazení rozdílů v prodejnosti produktů podle kategorie a regionu. Tj. chceme vidět množství prodaných produktů pro každou kategorii a pro každý region v jedné vizualizaci. To nám umožní rychle porovnat, jak se prodejnost liší mezi různými regiony a kategoriemi produktů.

6) **Komentovaný report**: Nakonec všechny výsledky představte ve formě komentovaného reportu (např. v Quarto nebo Jupyter Notebook). Tento report by měl obsahovat:

   - **Technická část**: Zde by měl být jasný a strukturovaný postup řešení počínaje představením dat a datového modelu. Jednotlivé kroky řešení by měly být řádně komentovány na úrovni zdrojového kódu.
   - **Interpretační část**: Tato část by měla obsahovat prezentaci a interpretaci klíčových zjištění a výsledků (s využitím tabulek a grafů).

Celkově by měl být report strukturovaný, srozumitelný a obsahovat všechny klíčové aspekty řešení.

Okruh B – využití statistických metod pro zpracování dat z dotazníkových šetření

Cílem úlohy je analyzovat data z rozsáhlého dotazníkového šetření (např. ve formátu CSV nebo XLSX) a identifikovat významné vztahy mezi různými demografickými skupinami a odpověďmi na otázky. Úkolem je provést explorační analýzu dat, formulovat statistické hypotézy a testovat je pomocí inferenčních statistických metod. Výstupem bude **komentovaný report** obsahující statistickou analýzu vstupních dat.

1) **Načtení a příprava dat**: Načtěte data do prostředí R/Pythonu, zkontrolujte kvalitu dat a proveďte potřebné úpravy, jako je například odstranění chybějících hodnot a duplicit nebo převedení proměnných do vhodných datových typů.

2) **Explorační analýza dat**: Prozkoumejte vztahy mezi jednotlivými proměnnými a demografickými faktory (věk, pohlaví, vzdělání, velikost obce atd.) a identifikujte klíčové vztahy a zájmové skupiny. Použijte metody popisné statistiky a vizualizační techniky, jako jsou histogramy, boxploty, scatterploty a heatmapy.

3) **Formulace statistických hypotéz**: Na základě zjištěných vztahů a skupin navrhněte statistické hypotézy, které budou testovány pomocí inferenčních statistických metod.

4) **Testování statistických hypotéz**: Použijte metody inferenční statistiky, jako jsou intervaly spolehlivosti, regresní analýza, analýza rozptylu (ANOVA), nebo neparametrické testy, pro testování navržených hypotéz a určení statistické významnosti zjištěných vztahů. U statisticky významných výsledků posuďte i věcnou významnost (effect size). Vždy ověřte předpoklady používaných metod.

5) **Komentovaný report**: Prezentujte výsledky statistické analýzy a vizualizace ve formě komentovaného reportu (Quarto, Jupyter Notebook aj.), který obsahuje výstupy z jednotlivých kroků, interpretaci výsledků a závěry.

Okruh C – vytvoření interaktivního dashboardu

Cílem úlohy je vytvoření **interaktivního dashboardu** znázorňujícího vybrané ekonomické ukazatele podniku. Preferovaným programem pro zpracování úlohy je Power BI, v případě zájmu je možné použít i jiný software, např. Tableau, Qlik.

1) Načtěte data z CSV souborů do Power BI (např. obchodní data o prodejích apod.). Každý z těchto souborů obsahuje data různých typů. Zkontrolujte kvalitu dat a proveďte úpravy, jako je odstranění chybějících hodnot a duplicit, data převeďte do vhodných datových typů.

2) Jednotlivé datové soubory propojte do datového modelu.

3) Vytvořte hierarchii `Continent` – `Country`.

4) V dimenzi `Country` nastavte kategorii dat na State/Province (Země/Region) a v dimenzi `Continent` nastavte kategorii dat na Continent (Kontinent), aby bylo možné data vizualizovat na mapě.

5) Vytvořte hierarchii `Year` – `MonthNumber`.

6) Pomocí funkcí vytvořte další dimenze, konkrétně finanční ukazatele, které vyplývají z ostatních ukazatelů v datasetu:

   - `GrossSales` jako `UnitsSold` × `SalePrice`,
   - `SalesAfterDiscount` jako `GrossSale` − `Discounts`,
   - `COGS` (Cost of Goods Sold) jako `UnitsSold` × `ManufacturingPrice`,
   - `Profit` jako `SalesAfterDiscount` − `COGS`.

7) Vytvořte několikastránkovou sestavu, na jejíchž dashboardech bude k dispozici:

   - přehledová tabulka všech dostupných informací,

- celková hodnota `COGS` a `Profit`,
- srovnání finančních výsledků (viz bod 6) po jednotlivých segmentech,
- srovnání finančních výsledků po jednotlivých produktech,
- srovnání finančních výsledků po jednotlivých státech,
- srovnání prodaného množství (`UnitsSold`) na úrovni produktů při různých typech slevových akcí (`DiscountBand`),
- srovnání celkového zisku (`Profit`) na úrovni produktů při různých typech slevových akcí (`DiscountBand`),
- vhodný graf TOP 5 produktů po jednotlivých segmentech z pohledu celkového zisku (umožněte uživateli volbu segmentu),
- srovnání finančních ukazatelů za rok 2014 a 2013 (absolutní i procentuální nárůst dle segmentu/produktu/státu),
- zisk jednotlivých států vizualizovaný na mapě.

Předpokládá se, že stránky sestavy budou vhodně pojmenované, grafy a tabulky budou mít vhodné názvy a popisky os, v případě potřeby budou jednotlivé stránky doplněny stručnými komentáři k použitým datům/tabulkám/grafům.

8) Umožněte uživateli v rámci jednotlivých stránek filtrovat minimálně dle segmentu, státu, produktu (např. pomocí průřezu).

9) Vhodně nastavte interakce mezi jednotlivými tabulkami/grafy.

10) Interpretujte získané výsledky (zpracujte jednostránkové manažerské shrnutí).

Okruh D – analýza časových řad

Cílem úlohy je zpracovat a analyzovat časové řady, a to typicky vybraných makroekonomických ukazatelů, jako jsou údaje o ekonomické výkonnosti (např. HDP), vývoji cenové hladiny (např. CPI, inflace) či statistiky zaměstnanosti (např. různé indikátory zaměstnanosti a nezaměstnanosti). Poskytnutá data mohou pocházet z různých časových období, být prezentována v různých časových škálách (měsíčně, ročně atp.) a mohou se týkat různých území (např. různé úrovně NUTS). Data mohou být také dále strukturována různými způsoby, například v případě statistik zaměstnanosti mohou být rozdělena podle věku a pohlaví. Výstupem bude **komentovaný report** obsahující zpracování a analýzu vstupních dat s důrazem na vizualizaci a interpretaci výsledků.

1) **Načtení a příprava dat**: Načtěte data do prostředí R či Python. Zkontrolujte kvalitu dat a proveďte potřebné úpravy, jako je převedení dat do správného formátu a časového měřítka, odstranění nebo nahrazení chybějících hodnot, odstranění duplicit, odlehlých hodnot apod.

2) **Explorační analýza dat a vizualizace**: Prozkoumejte vztahy mezi jednotlivými proměnnými a faktory. Identifikujte a vizualizujte klíčové vztahy v datech. Využijte metody popisné statistiky (např. různé typy průměrů, absolutní a relativní míry dynamiky, diference vyšších řádů, míry variability), testování stacionarity, metody korelační analýzy

a vizualizační techniky, jako jsou například spojnicové grafy časových řad, histogramy, boxploty, scatterploty či korelogramy. Volbu metod přizpůsobte charakteru dat a omezením metod.

3) **Dekompozice časových řad**: Proveďte dekompozici vybraných časových řad. Vysvětlete, proč volíte aditivní, nebo multiplikativní model. K odhadu trendové složky použijte metodu klouzavých průměrů.

4) **Modelování časových řad a predikce**: Na vybrané časové řady aplikujte vybrané regresní modely, modely exponenciálního vyrovnávání a ARIMA modely. Postupujte vždy tak, že data v časové řadě rozdělíte na trénovací a testovací, trénovací data využijete k nastavení modelů a testovací data k ověření výkonnosti modelů a jejich schopnosti předpovídat budoucí hodnoty. Pro každý model vypočítejte metriky, jako jsou například RMSE a MAPE, a to jak na trénovacích, tak na testovacích datech. Porovnejte tyto metriky napříč různými modely. Modely použijte k předpovědím budoucích hodnot zvolených časových řad. Pro budoucí hodnoty určete intervaly spolehlivosti.

5) **Komentovaný report**: Prezentujte výsledky zpracování a analýz vstupních časových řad ve formě komentovaného reportu (např. Quarto nebo Jupyter Notebook). Zaměřte se na přehlednost výstupů (tabulky, grafy) a jejich interpretaci. Komentujte postup řešení. Proveďte reflexi a diskusi nad celým procesem a výsledky. Diskutujte o možných omezeních použitých metod a dat.

Okruh E – použití metod lineárního programování

Cílem úlohy je aplikace metod lineárního programování na úlohy z praxe použití SW knihoven (PULP, CVXOPT). Výstupem bude **Jupyter Notebook** obsahující matematickou formulaci modelu, implementovaný funkční model, diskuse řešení a grafické výstupy.

1) Na základě slovního popisu a dodaných dat sestavte matematický model ve tvaru lineárního programování (varianty distribučního problému, směšovacího problému, toky sítí apod.), např.:

> *Uvažujte dvoustupňový dopravní problém s pěti producenty, třemi mezisklady a čtyřmi odběrateli, kde kapacity a požadavky všech účastníků jsou dány tabulkou. Uvažujte v modelu i platbu za skladování za kus, která je pro každý z meziskladů jiná (rovněž zadáno tabulkou). Ceny za přepravu mezi účastníky po jednotlivých trasách, včetně maximálních a minimálních kapacit tras, jsou zadány taktéž. Sestavte matematický model úlohy a nalezněte jeho optimální řešení.*

2) Implementujte model za použití SW knihoven a nalezněte jeho optimální řešení (případně množinu optimálních řešení).

3) Ověřte citlivost řešení na změnu podmínek, např. za pomoci duální formulace daného problému určíme duální proměnné (stínové ceny), nebo budeme opakovaně měnit požadované parametry modelu, např.:

- *Ve výše uvedeném modelu ověřte jeho chování při změně nákladů na skladování v prvním meziskladu na intervalu* $\langle \min, \max \rangle$*, ostatní parametry modelu ponechte stejné. Zaměřte se na celkové náklady a chování přepravních cest.*
- *Ověřte vliv maximálních kapacit vybraných přepravních cest na původní model. Uvažujte změnu maximální kapacity v diskrétním intervalu mezi polovinou a dvojnásobkem původní hodnoty.*

4) Diskuse řešení