Druhá část projektu - základní analýza dat

Cílem této částu projektu je provést základní analýzu datasetu Statistika nehodovosti Policie ČR. Data pro jednotlivé roky jsou v CSV souboru stažena a jsou k dispozici na stránce: ehw.fit.vutbr.cz/izv/data.zip. Popis dat naleznete na stránce https://ehw.fit.vutbr.cz/izv. Řešení se skládá ze 5 úkolů, přičemž výstupem každého úkolu bude jedna funkce implementovaná v jazyce Python.

Cíl

Cílem je vytvořit kód, který vizualizuje tři různé závislosti v datech. Kód bude součástí jednoho souboru `analysis.py`, jehož **kostru naleznete v elearningu** v IS VUT. Předpokládá se, že budete **primárně pracovat s knihovnami Pandas a Seaborn** + je dovolené využít všechny knihovny zmiňované během přednášek. Matplotlib používejte pouze pro doladění vizuální podoby. Doporučený postup řešení je pouze nápovědou, můžete samozřejmě problém řešit jiným způsobem, pokud to povede ke stejnému výsledku bez nárustu komplikovanosti.

Odevzdávání a hodnocení

Soubor `analysis.py` odevzdejte do 15. 12. 2023. Hodnotit se bude zejména:

- správnost výsledků
- vizuální zpracování grafů
- kvalita kódu
    - efektivita implementace (nebude hodnocena rychlost, ale bude kontrolováno, zda nějakým způsobem řádově nezvyšujete složitost)
    - korektní práce s Pandas a Seaborn
    - přehlednost kódu
    - dodržení standardů a zvyklostí pro práci s jazykem Python (PEP8)
    - dokumentace kódu

Celkem lze získat až 20 bodů, přičemž k zápočtu je nutné získat z této části minimálně 1 bod.

Úkol 1: Načtení dat (až 4 body)

Signatura funkce, která bude odpovídat tomuto úkolu:

```
def load_data(filename : str) -> pd.DataFrame:
```

Funkcionalita

- Funkce načte data obsažená v ZIP souboru získaného z adresy http://ehw.fit.vutbr.cz/izv/data.zip . Argument `filename` určuje cestu k souboru. Soubor nestahujte z webu.
- Tento soubor obsahuje ZIP soubory pro každý rok, kdy jsou data rozdělená v rámci krajů (čísla krajů a zkratky naleznete v popisu datového souboru).
- Vaším úkolem je postupně projít všechny potřebné soubory a roky a načíst data do jednoho DataFrame.
- Přidejte jeden sloupec *region* obsahující třípísmennou zkratku daného regionu.
- Není dovoleno vytvářet pomocné soubory.
- Data ve sloupcích nemodifikujte (zpracování dat bude součástí funkce `parse_data`).
- Názvy sloupců musí odpovídat názvům specifikovaném v popisu datového souboru (t.j. p1, p36, p37, …) + sloupec *region*.

Tipy: pracujte s třídou zipfile.ZipFile(), při načítání dat přes Pandas berte v úvahu fakt, že data jsou v kódování `cp1250`.

Upozornění: Další skripty budou využívat vaši implementaci načítání datového rámce. Bez této implementace není možné hodnotit další úkoly**.**

Úkol 2: Formátování a čištění dat (až 4 body)

Signatura funkce, která bude odpovídat tomuto úkolu:

```
def parse_data(df : pd.DataFrame,
               verbose : bool = False) -> pd.DataFrame:
```

Funkcionalita

- Funkce dostane na vstup DataFrame získaný voláním funkce `load_data()`.
- Vytvoří se nový DataFrame.
- Funkce vytvoří v DataFrame sloupec *date*, který bude ve formátu pro reprezentaci data (berte v potaz pouze datum, t.j. sloupec *p2a*)
- S výjimkou sloupce region reprezentujte vhodné sloupce pomocí kategorického datového typu. Měli byste se dostat pod velikost v paměti 0.7 GB. Sloupec *region* je vhodné ponechat v původní podobě pro lepší práci s figure-level funkcemi Seaborn.
- Sloupce obsahující čísla s desetinnou čárkou reprezentujte jako float (zejména slouce *d* a *e*).
- Odstraňte duplikované záznamy dle identifikačního čísla (*p1*).
- Při povoleném výpisu (`verbose == True`) spočítejte kompletní (hlubokou) velikost všech sloupců v datovém rámci před a po vaší úpravě a vypište na standardní výstup pomocí funkce print následující 2 řádky
  `orig_size=X MB`
  `new_size=X MB`
  Čísla vypisujte na 1 desetinné místo a počítejte, že 1 MB = 10^6 B.

Tipy: Převod do formátu data provedete funkcí `pd.to_datetime`

Upozornění: Další skripty budou využívat vaši implementaci parsování datového rámce. Bez této implementace není možné hodnotit další úkoly**.**

Úkol 3: Počty nehod podle stavu řidiče (až 4 body)

Signatura funkce, která bude odpovídat tomuto úkolu:

```
def plot_state(df: pd.DataFrame, fig_location: str = None,
               show_figure: bool = False):
```

Funkcionalita

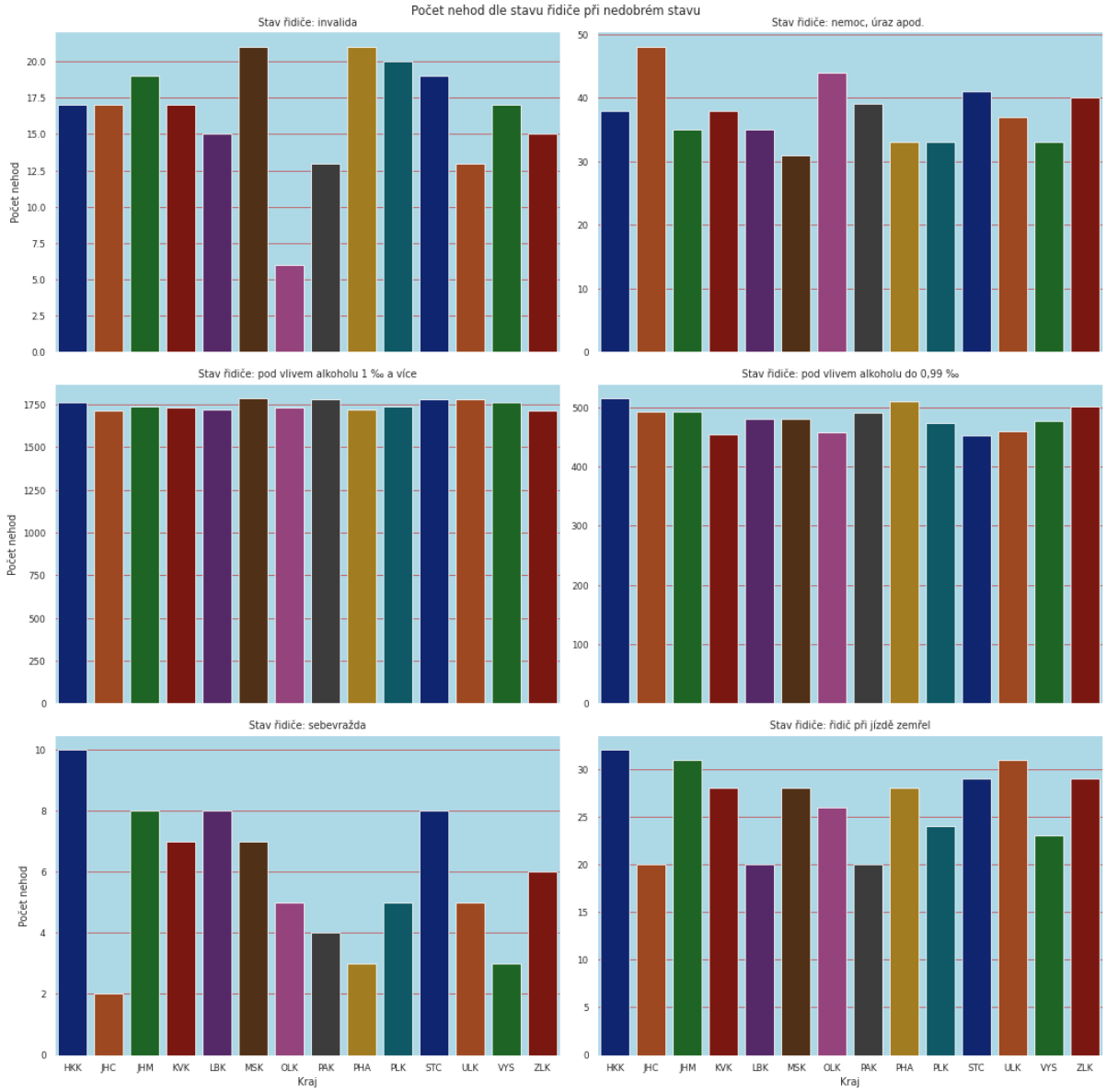
Vytvořte graf počtu nehod v jednotlivých regionech, který uložte do souboru specifikovaného argumentem `fig_location` a případně zobrazte pokud argument `show_figure` je **`True`**. Argument `df` odpovídá DataFrame jenž je výstupem funkce `parse_data`. Graf se bude skládat z 6 podgrafů uspořádaných do mřížky o třech řádcích a dvou sloupcích.

Požadavky:

1) Pracujte se sloupcem *p57*, rozlišujte pouze stavy označené číslem 1 až 9. Je vhodné využít náhrady.
2) Nastavte správně titulky jednotlivých podgrafů.
3) Upravte zobrazení tak, aby se grafy a popisky nepřekrývaly.
4) Graf upravte tak, aby popisky na osách, titulky atd. dávaly smysl. Dle zásad dobré vizualizace (viz přednáška 4) zvolte vhodnou barvu a vhodný styl. U podgrafu nastavte vlastní pozadí.

Tip: nejdříve je vhodné si nahradit celočíselné hodnoty ve sloupci *p57* vhodnými řetězci a vytvořit nějaký *pomocný* sloupec, který potom budete sčítat při agregaci `groupby`. Vhodným figure-level grafem pak můžete tato data vizualizovat.

Příklad výstupu: (použita náhodná data a záměrně zcela nevhodná grafická forma)


Počet nehod dle stavu řidiče při nedobrém stavu
Stav řidiče: invalida
Stav řidiče: nemoc, úraz apod.
Stav řidiče: pod vlivem alkoholu 1 ‰ a více
Stav řidiče: pod vlivem alkoholu do 0,99 ‰
Stav řidiče: sebevražda
Stav řidiče: řidič při jízdě zemřel
Počet nehod
Kraj
HKK JHC JHM KVK LBK MSK OLK PAK PHA PLK STC ULK VYS ZLK

Úkol 4: Alkohol v jednotlivých hodinách (až 4 body)

Signatura funkce, která bude odpovídat tomuto úkolu:

```
def plot_alcohol(df: pd.DataFrame, fig_location: str = None,
                 show_figure: bool = False):
```

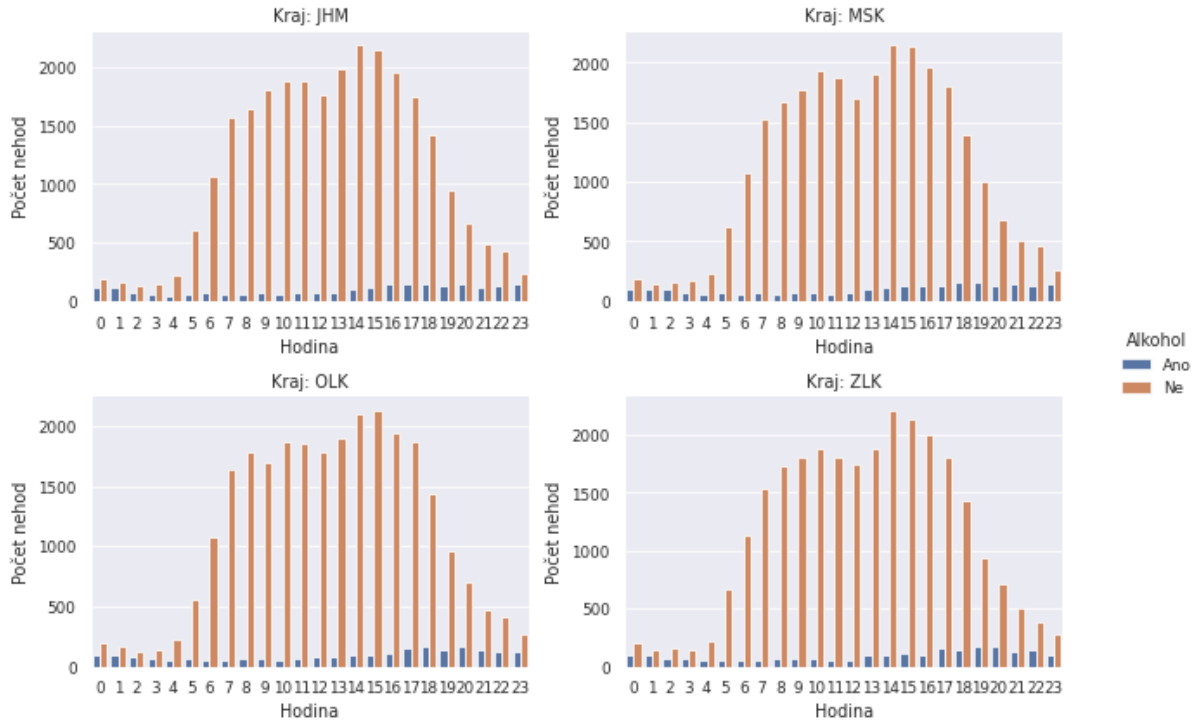
Funkcionalita

Vytvořte graf počtu nehod v jednotlivých regionech, který uložte do souboru specifikovaného argumentem `fig_location` a případně zobrazte pokud `show_figure` je `True`. Argument `df` odpovídá DataFrame jenž je výstupem funkce `parse_data`. Pro čtyři vámi vybrané kraje znázorněte počet nehod pod vlivem (p11 >= 3) a bez vlivu alkoholu (p11 in [1, 2]) v jednotlivých hodinách (p2b). Nehody, kdy čas není znám, nezapočítávejte.

Požadavky a doporučený postup:

1) Přepočtěte sloupec p2b na hodiny (celočíselným dělením) a neznámé hodnoty odstraňte.
2) Určete si vhodným způsobem, zda byl v nehodě přítomný alkohol.
3) Data správně agregujte pomocí `groupby` a vykreslete vhodným figure-level grafem.
4) Upravte zobrazení tak, aby se grafy a popisky nepřekrývaly.
5) Graf upravte tak, aby popisky na osách, titulky atd. dával smysl.

Příklad výstupu: (použita náhodná data)



Úkol 5: Zavinění nehody v čase (až 5 bodů)

Signatura funkce, která bude odpovídat tomuto úkolu:

```
def plot_fault(df: pd.DataFrame, fig_location: str = None,
               show_figure: bool = False):
```

Funkcionalita

Vytvořte graf počtu nehod v jednotlivých regionech, který uložte do souboru specifikovaného argumentem `fig_location` a případně zobrazte pokud `show_figure` je **`True`**. Argument `df` odpovídá DataFrame jenž je výstupem funkce `parse_data`. Pro čtyři vámi vybrané kraje pro různé měsíce vykreslete čárový graf, který bude zobrazovat pro jednotlivé měsíce (osa X- sloupec *date*) počet nehod podle zavinění. Uvažujte pouze zavinění chodcem, zvířetem a řidičem motorového či nemotorového vozidla.

Požadavky a doporučený postup:

1. Vyberte čtyři kraje a vyfiltrujte všechny nehody.
2. Na základě zavinění (p10) určete textovou hodnotu zavinění - využívejte jen hodnoty 1, 2, 3 a 4.
3. Transformujte tabulku tak, aby pro každý den a region byl v každém sloupci odpovídajícím následkům počet nehod (`pivot_table`)
4. Pro každý kraj proveďte podvzorkování na úroveň měsíců a převeďte správně do *stacked formátu*.
5. Vykreslete čárový graf a omezte osu X od 1. 1. 2016 do 1. 1. 2023.
6. Vykreslete patřičné grafy upravené tak, aby popisky na osách, titulky atd. dávaly smysl.

Příklad výstupu: (použita náhodná data)

Poznámky k implementaci

Soubor, který vytvoříte, bude při hodnocení importovaný a budou volány jednotlivé funkce. Mimo tyto funkce, část importů a dokumentační řetězce nepište žádný funkční kód. Blok na konci souboru ohraničený podmínkou

```
if __name__ == "__main__":
    pass
```

naopak můžete upravit libovolně pro testovací účely. Dále můžete přidat další funkce (pokud budete potřebovat), pro názvy těchto funkcí použijte prefix "_".

Stručnou dokumentaci všech částí (souboru a funkcí) uveďte přímo v odevzdaných souborech. Respektuje konvenci pro formátování kódu PEP 257 [PEP 257 -- Docstring Conventions] a PEP 8 [PEP 8 -- Style Guide for Python Code].

Grafy by měly splňovat všechny náležitosti, které u grafu očekáváme, měl by být přehledný a jeho velikost by měla být taková, aby se dal čitelně použít v šířce A4 (t.j. cca 18 cm). Toto omezení není úplně striktní, ale negenerujte grafy, které by byly přes celý monitor.

Grafy v zadání jsou pouze ukázkové. Data byla randomizována a vaše výsledky budou vypadat jinak. Není nutné ani chtěné, aby grafy vizuálně vypadaly stejně - vlastní invence směrem k větší přehlednosti a hezčímu vzhledu se cení! Co není přímo specifikováno v zadání můžete vyřešit podle svého uvážení.

Doporučený postup není nutné dodržet. Důležité je však to, aby výsledky odpovídaly zadání (včetně podvýběru dat a podobně). U argumentů funkcí `fig_location` můžete počítat s tím, že adresář, kam se mají data ukládat, již existuje.

Dotazy a připomínky

Dotazy a připomínky směřujte na fórum v IS VUT případně na mail mrazek@fit.vutbr.cz.