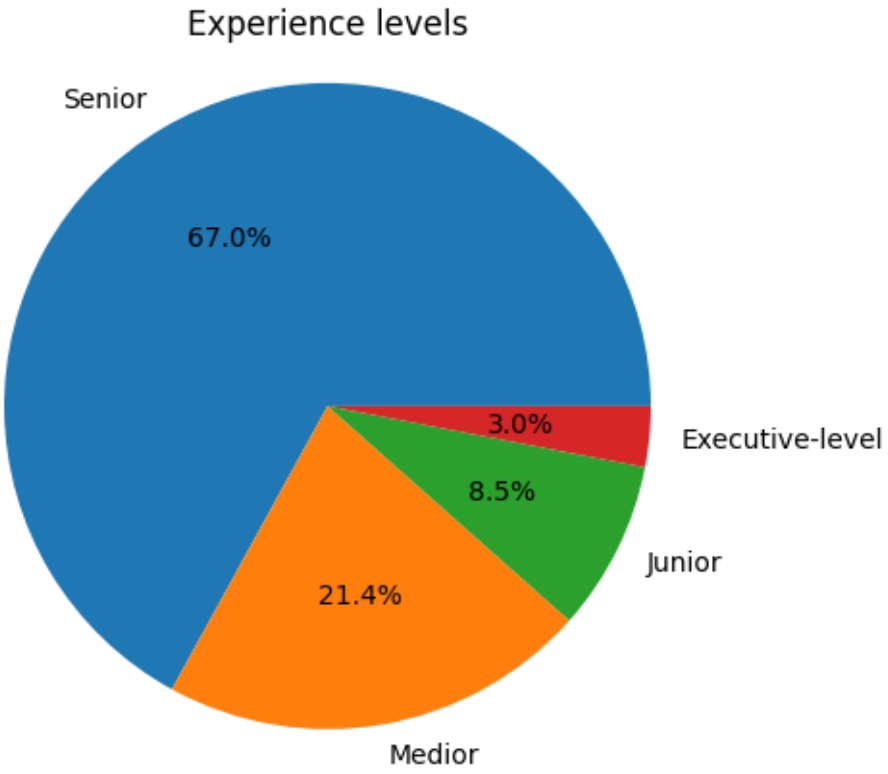
2. Vizualizace dat

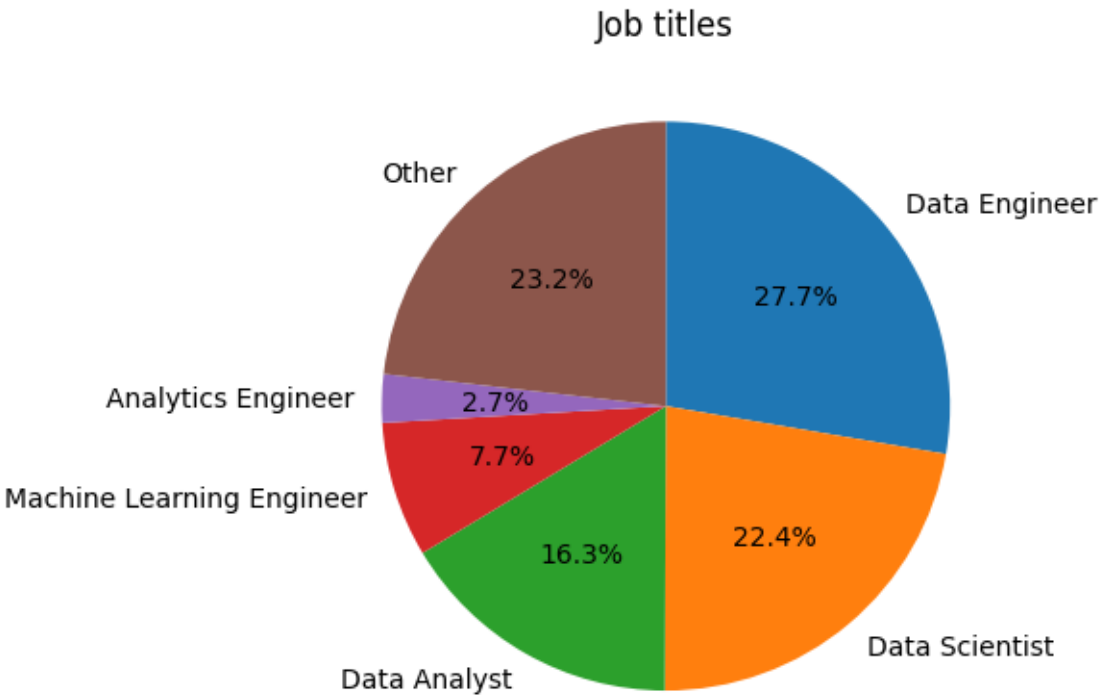
Zadání:
V jednom ze cvičení jste probírali práci s moduly pro vizualizaci dat. Mezi nejznámější moduly patří matplotlib (a jeho nadstavby jako seaborn), pillow, opencv, aj. Vyberte si nějakou zajímavou datovou sadu na webovém portále Kaggle a proveďte datovou analýzu datové sady. Využijte k tomu různé typy grafů a interpretujte je (minimálně alespoň 5 zajímavých grafů). Příklad interpretace: z datové sady pro počasí vyplynulo z liniového grafu, že v létě je vyšší rozptyl mezi minimální a maximální hodnotou teploty. Z jiného grafu vyplývá, že v létě je vyšší průměrná vlhkost vzduchu. Důvodem vyššího rozptylu může být absorpce záření vzduchem, který má v létě vyšší tepelnou kapacitu.
https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023

Řešení:
Vybrala jsem si datovou sadu „Data Science Salaries 2023 “, kterou můžete najít na tomto odkazu: https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023.

Nejdřív jsem si udělala analýzu této datové sady, abych zjistila s jakými daty pracuju.

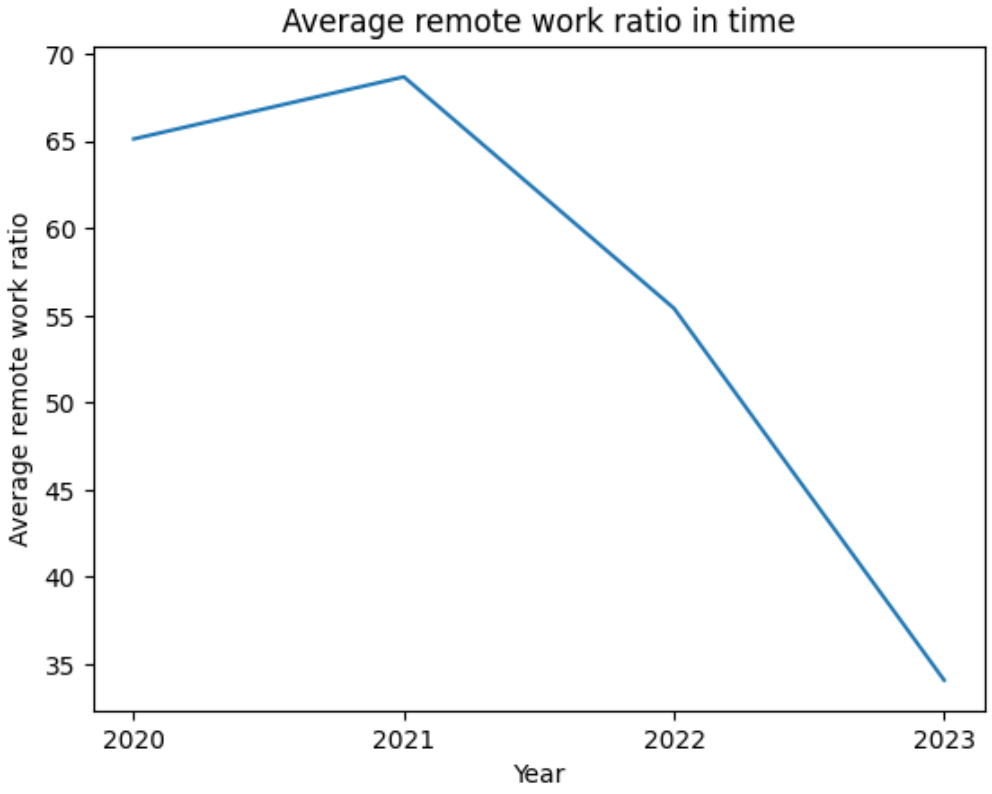


První graf ukazuje úroveň zkušeností pracovníků v odvětví data science. Na grafu vidíme, že většina pracovníků poměrně zkušená a jsou to senioři v oboru. Naopak nejméně lidí ze vzorku má úroveň vedoucího pracovníka.



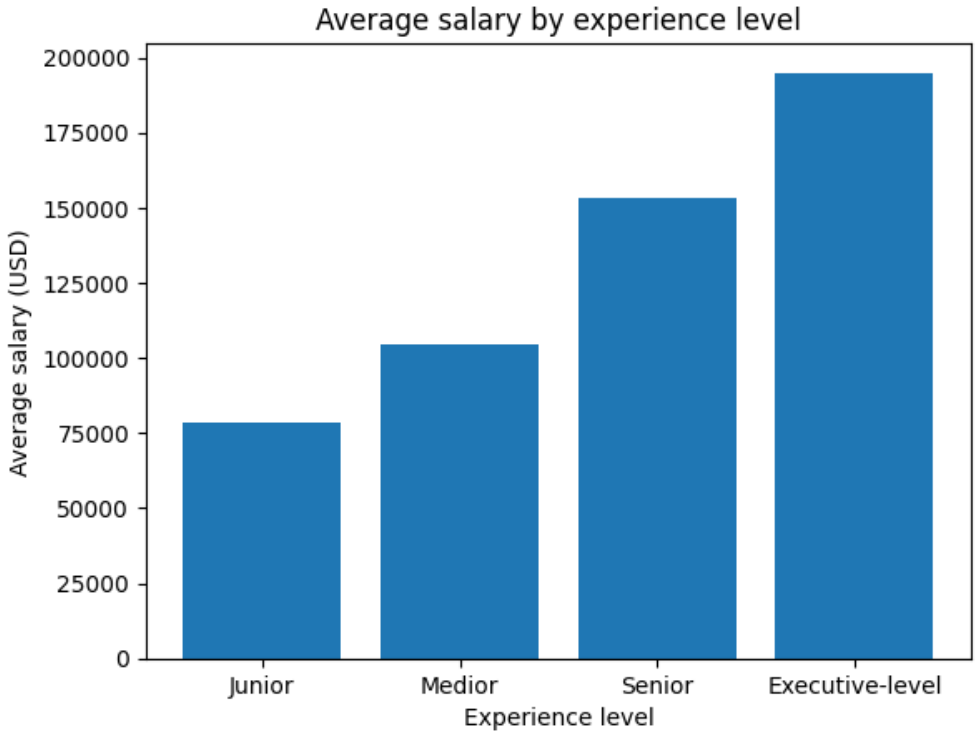
Další koláčový graf zobrazuje na jakých pozicích pracují daní zaměstnanci. Graf ukazuje 5 nejčastějších pracovních pozic v této datové sadě a poté podíl všech ostatních. V grafu nejsou všechny pozice, protože je jich tolik, že by se to do grafu nevešlo.

Potom, co jsem zjistila, jak zhruba datová sada vypadá, tak se podívám na nějaká zajímavější data. Začnu podílem práce z domova v čase.

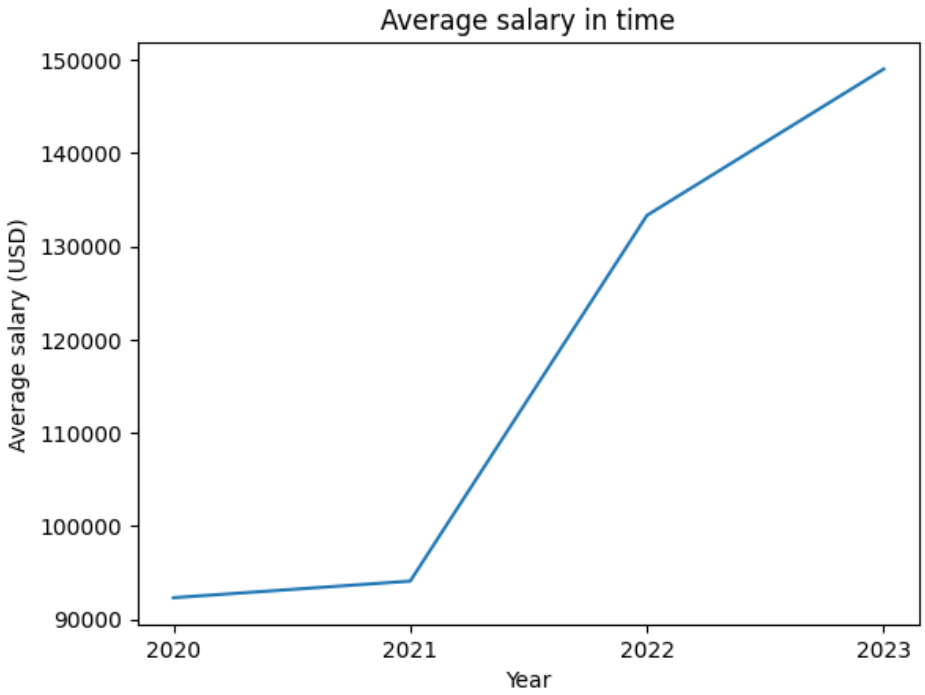


Na tomto grafu je vidět, že největší podíl práce z domova byl v roce 2021 a 2020. Důvodem je nejspíše covid-19. V těchto letech byla nejpřísnější epidemiologická opatření a lidé častěji pracovali z domova. Data science je obor, který nepotřebuje osobní přítomnost pracovníků na pracovišti, a tak je práce z domova častá. Na grafu je i vidět, že potom, co se situace uklidnila, tak se lidi vrátili zpět do kanceláří a podíl práce z domova velmi klesl.

Další grafy se zabývají výškou platu v závislosti na různých aspektech.



Jako první tu mám graf, který zobrazuje výši platu v závislosti na zkušenostech pracovníka. Nejspíš nikoho nepřekvapí, že plat roste úměrně ke zkušenostem zaměstnance.



Na dalším grafu je zobrazený vývoj platu v čase. Mezi lety 2020 a 2021 průměrný plat příliš nevzrostl, ale pak je velký skok mezi rokem 2021 a 2022. Zjištění z tohoto grafu nejsou nijak zvlášť překvapivá a to, že průměrný plat se s každým rokem zvyšuje.

Z posledního grafu je vidět, že práce v oboru data science se nejvíc vyplatí v Severní Americe. Průměrné platy v Severní Americe přesahují hodnotu 140 000 dolarů ročně, což je v přepočtu zhruba 3 milióny Kč. Na druhém místě se nachází Evropa s velkým rozdílem v platu. Průměrná hodnota ročního platu v Evropě je někde okolo 80 000 dolarů. Následuje Austrálie, Afrika, Asie a nakonec Jižní Amerika.