A close up of a sign

Description automatically generated

Univerzitet u Beogradu,

Elektrotehnički fakultet

**Semestralni rad**

Softversko inženjerstvo velikih baza podataka

|  |  |
| --- | --- |
| **Mentor:**  Prof. dr Miroslav Bojović | **Student:**  Nina Grujić 2019/3300 |

Sadržaj

[1. Postavka zadatka 3](#_Toc49084728)

[2. Detaljna analiza podataka 4](#_Toc49084729)

# 1. Postavka zadatka

Cilj zadatka je da se korišćenjem različitih algoritama mašinskog učenja predvidi brzina kojom su napušteni ljubimci širom Malazije usvajani. Podaci korišćeni za obučavanje i testiranje preuzeti su sa sajta [PetFinder.my](https://www.petfinder.my/). Na raspolaganju su sledeće datoteke:

* train.csv – podaci za obučavanje;
* test.csv – podaci za testiranje;
* breed\_labels.csv – sadrži tip ljubimca (Type - 1 za pse, 2 za mačke) i naziv rase (BreedName) pridruženih uz identifikator rase (BreedID);
* color\_labels.csv – sadrži nazive boja (ColorName) pridruženih uz identifikator boje (ColorID);
* state\_labels.csv – sadrži nazive država (StateName) pridruženih uz identifikator države (StateID).

U train.csv i test.csv datotekama na raspolaganju su sledeći podaci:

* **PetID** – jedinstvena heš vrednost za svakog ljubimca ili grupu ljubimaca;
* **Type** – tip ljubimca, 1 za pse, 2 za mačke;
* **Name** – naziv ljubimca;
* **Age** – starost ljubimca izražena u mesecima;
* **Breed1** – primarna rasa ljubimca;
* **Breed2** – sekundarna rasa ljubimca ako je ljubimac rasni mešanac;
* **Gender** – pol ljubimca (1 – muški pol, 2 – ženski pol, 3 – i muški i ženski pol, ako je u pitanju grupa koja se odnosi na ljubimce oba pola);
* **Color1** – primarna boja ljubimca;
* **Color2** – sekundarna boja ljubimca, ako posotji;
* **Color3** – tercijalna boja ljubimca, ako postoji;
* **MaturitySize** – veličina ljubimca (1 – mali, 2 – srednji, 3 – veliki, 4 – veoma veliki, 0 – nije specificirano);
* **FurLength** – dužina krzna (1 – kratko, 2 – srednje, 3 – dugačko, 0 – nije specificirano);
* **Vaccinated** – da li je ljubimac vakcinisan (1 – da, 2 – ne, 3 – ne zna se);
* **Dewormed** – da li je ljubimac očišćen od buva (1 – da, 2 – ne, 3 – ne zna se);
* **Sterilized** – da li je ljubimac sterilisan (1 – da, 2 – ne, 3 – ne zna se);
* **Health** – opšte zdravlje ljubimca (1 – zdrav, 2, manje povrede, 3 veće povrede);
* **Quantity** – Broj ljubimca dat na usvajanje;
* **Fee** – novčana nadoknada za ljubimca;
* **State** – država ljubimca;
* **RescuerID** – jedinstvena heš vrednost osobe/organizacije koja je prijavila ljubimca;
* **VideoAmt** – broj video klipova pridruženih uz svakog ljubimca/grupu ljubimaca;
* **PhotoAmt** – broj slika pridruženih uz svakog ljubimca/grupu ljubimaca;
* **Description** – opis ljubimca.

Vrednost koju treba predvideti je brzina usvajanja ljubimaca. Brzina usvajanja ljubimaca data je u koloni **AdoptionSpeed** skupa za obučavanje i ima sledeće vrednosti:

* **0** – ljubimac je usvojen istog dana kada je i izlistan;
* **1** – ljubimac je usvojen između 1. i 7. dana od momenta kada je izlistan;
* **2** – ljubimac je usvojen između 8. i 30. dana od momenta kada je izlistan;
* **3** – ljubimac je usvojen između 31. i 90. dana od momenta kada je izlistan;
* **4** – nije zabeleženo usvajanje ljubimca nakon 100 dana od momenta kada je ljubimac izlistan.

# 2. Detaljna analiza podataka

Možda najvažniji korak u mašinskom učenju jeste analiza i vizuelizacija podataka. Analiza podataka pomaže u konstrukciji modela odnosno odabiru parametara od značaja kao i transformaciji parametara i kreiranju novih na osnovu već postojećih parametara.

Na raspolaganju je train.csv datoteka koja raspolaže sa **14993** redova. Najpre treba videti da li je baza kompletna, odnosno da li neki podaci nedostaju. Postoji **1257** redova u kojima nedostaje kolona **Name** i **12** redova u kojima nedostaje kolona **Description**.

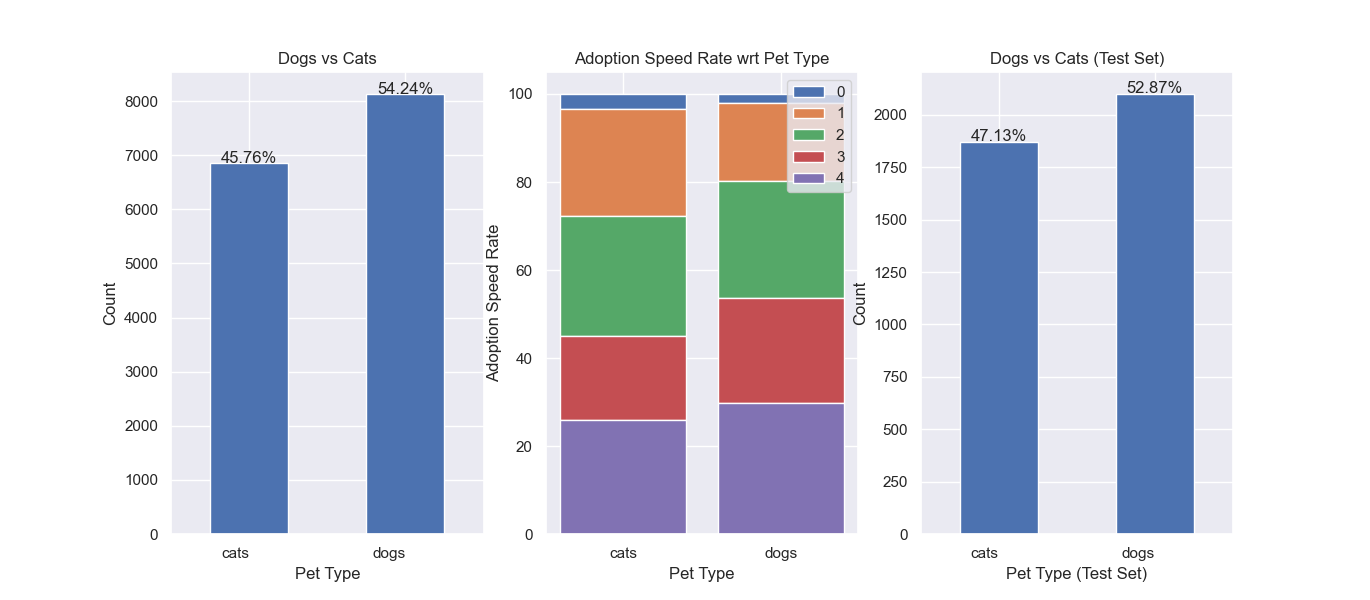
A screenshot of a social media post

Description automatically generated

**Slika 1:** Brzine usvajanja - raspodela

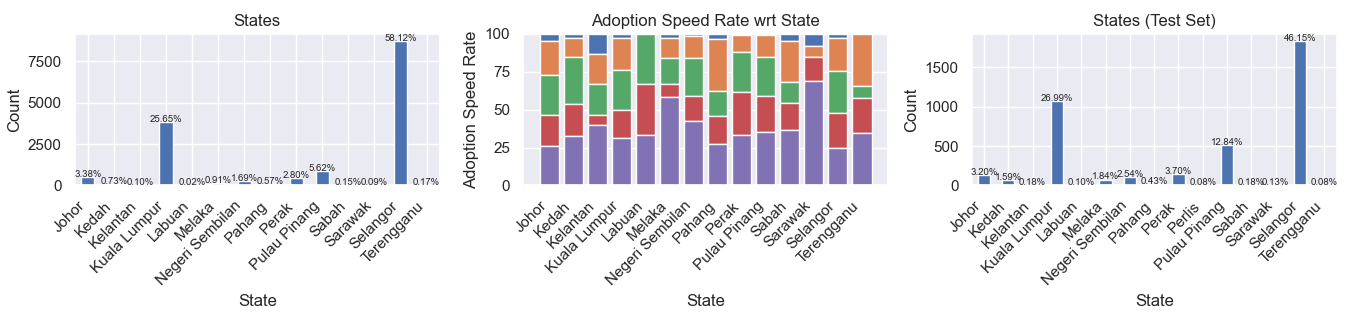
Možemo zaključiti da je znatno više onih životinja koje su usvojene, nego onih koje nisu. Međutim kada je u pitanju raspodela brzine usvajanja po klasama, vidimo da je jako mali procenat životinja koje su usvojene istog dana kada su i izlistane, dok je procenat životinja koje nisu usvojene dominantan.

Kada je u pitanju tip ljubimca, može se zaključiti da je u skupovima podataka i za obučavanje i za testiranje nešto veći broj pasa nego mačaka, a da su mačke brže usvajane od pasa.



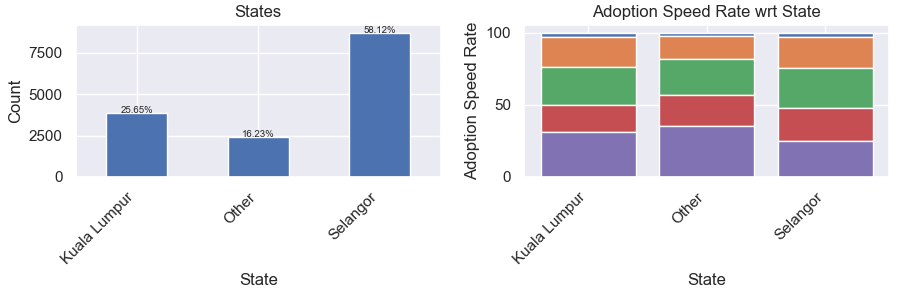
**Slika 2:** Raspodela tipa ljubimca u skupu za obučavanje i testiranje i raspodela brzina usvajanja u odnosu na tip

Kada je u pitanju država u kojoj se ljubimac nalazi, može se primetiti da je najzastupljenija država Selangor, zatim Kuala Lumpur, dok su ostale države prisutne u znatno manjem broju. Kuala Lumpur ima najveći procenat životinja koje pripadaju klasi 0, dok Selangor ima najmanji procenat životinja koje pripadaju klasi 4.



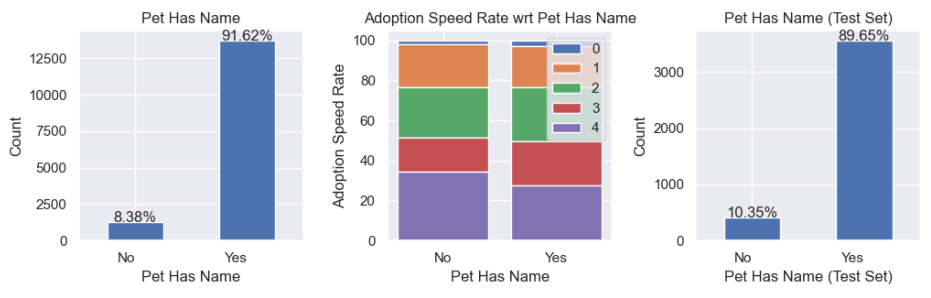
**Slika 3:** Raspodela država u skupu za obučavanje i testiranje i raspodela brzina usvajanja u odnosu na države

S obzirom da su ostale države prisutne u manjem broju, možemo ih podeliti u tri kategorije: Selangor, Kuala Lumpur, i ostale države. Kod tako raspodeljenih država vidimo da je usvajanje kod ostalih država sporije u odnosu na Kuala Lumpur i Selangor, ali da razlike nisu značajne.



**Slika 4:** Raspodela država podeljenih u kategorije u skupu za obučavanje i testiranje i raspodela brzina usvajanja u odnosu na kategoriju države

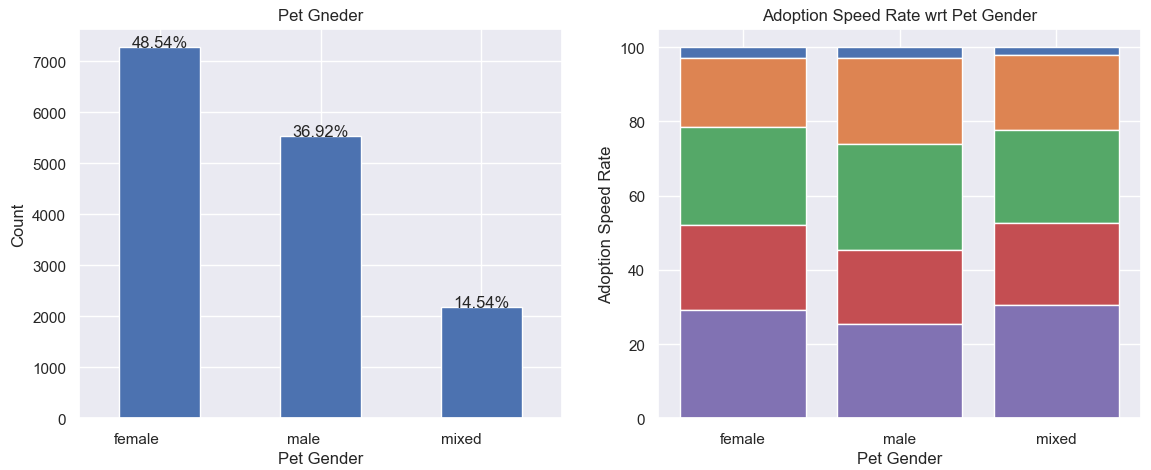
Kada je u pitanju ime životinje, posotji određeni broj ljubimaca kojima ime nije dodeljeno. Zbog toga je interesantno videti da li postojanje imena utiče na brzinu usvajanja:



**Slika 5:** Postojanje imena u skupu za obučavanje i testiranje i raspodela brzina usvajanja u odnosu na postojanje imena

Može se primetiti da je broj ljubimaca koji imaju ime znatno veći od broja ljubimaca koji ime nemaju. Što se tiče brzine usvajanja, može se primetiti da su ljubimci sa imenom više usvajani u odnosu na ljubimce bez imena. Može se kreirati nova odlika **HasName** koja će u modelu biti korišćena umestoodlike **Name**.

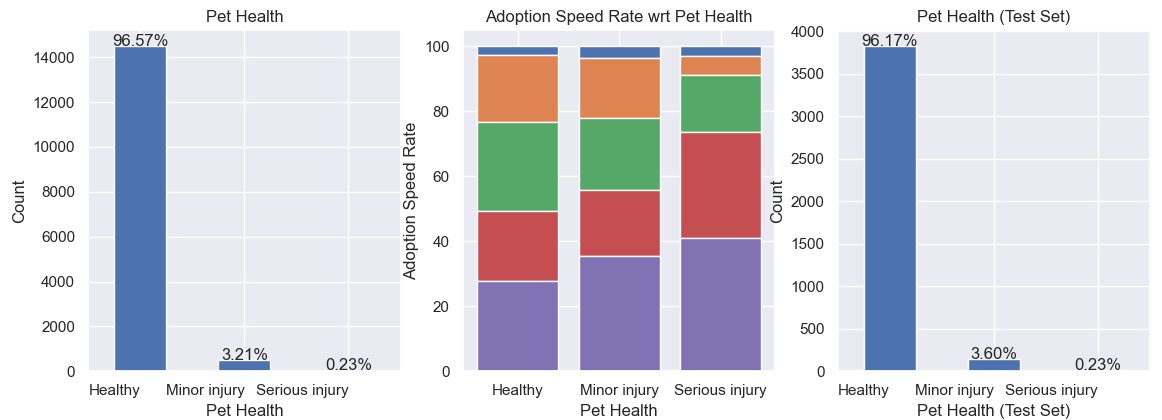
Što se tiče pola ljubimca, najzastupljeniji je ženski pol. Takođe se može ptimetiti da su ljubimci muškog pola, nešto brže usvajani od ženskog i „mešanog“ pola.



**Slika 6:** Zastupljenost pola i raspodela klasa brzina usvajanja u odnosu na pol

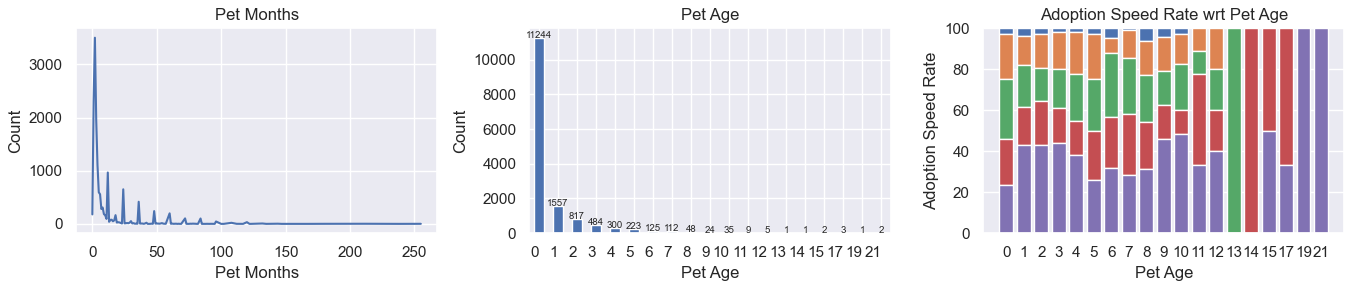
Na sličan način mogu se analizariti parametri koji svedoče da li je ljubimac vakcinisan, očišćen od buva kao i da li je sterilisan. Zanimljivo je da su ljubimci koji nisu vaksinicani, koji nisu očišćeni od buva i koji nisu sterilisani nešto brže usvajani od ljubimaca koji jesu.

Što se tiče opšteg zdravlja ljubimca, ljubimci koji su dobreg opšteg zdravlja su najbrže usvajani, dok su ljubimci sa ozbiljnim povredama najsporije usvajani. Međutim kako je i u skupu za obučavanje i u skupu za testiranje broj ljubimaca sa povredama veoma mali, ovaj parametar se možda može i zanemariti.



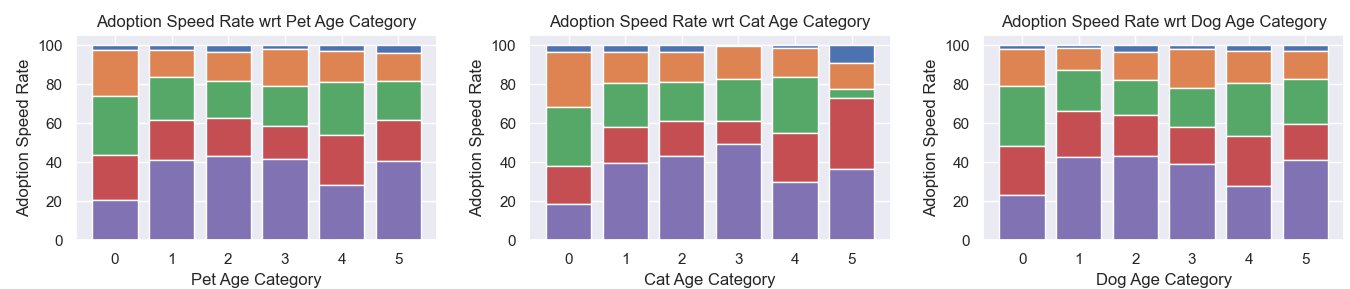
**Slika 7:** Zastupljenost opšteg zdravlja ljubimca i raspodela brzine usvajanja u odnosu na opše zdravlje

Godine deluju kao veoma bitan faktor za brzinu usvajanja. Radi vizuelizacije, najpre su meseci života ljubimca transformisani u godine, a potom u kategorije. Može se primetiti da su ljubimci koji nisu navršili ni godinu dana najzastupljeniji, kao i da su ti ljubimci najbrže usvajani. Brzina usvajanja kod starijih ljubimaca je nasumična, mada takvih ljubimaca ima znatno manje.



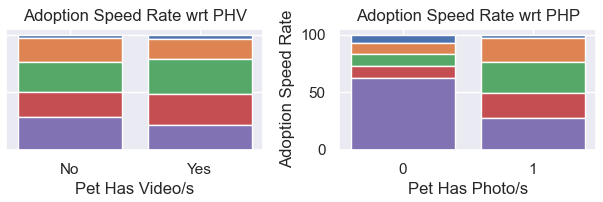
**Slika 8:** Zastupljenost meseca života, godina života i raspodela brzine usvajanja u odnosu na godine života ljubimca

Pošto je na primer nekad teško proceniti da li neki ljubimac ima 4 godine ili 7 godina, pomislila sam kako bi možda bilo značajnije mesece podeliti u određene kategorije (0 – [0,6), 1 – [6, 12), 2 – [12,36), 3- [36, 60), 4 – [60, 96), 5 – [96, +)). Zanimljivo je da su mačke koje imaju 8 i više godina najbrže usvajane ako posmatramo klase 0, 1 i 2.



**Slika 9:** Raspodela brzine usvajanja u odnosu na senioritet svih tipova, mačaka i pasa

Još jedan veoma bitan faktor za brzinu usvajanja ljubimca je broj raspoloživih fotografija i video klipova. Retko koji ljubimac ima priložen video klip, dok se broj fotografija pridruženih uz ljubimce kreće između 0 i 30. Ljubimci koji imaju video, su nešto brže usvajani u odnosu na ljubimce koji nemaju video, dok je broj neusvojenih ljubimaca bez ijedne fotografije znatno veći od ljubimaca koji imaju makar jednu fotografiju.



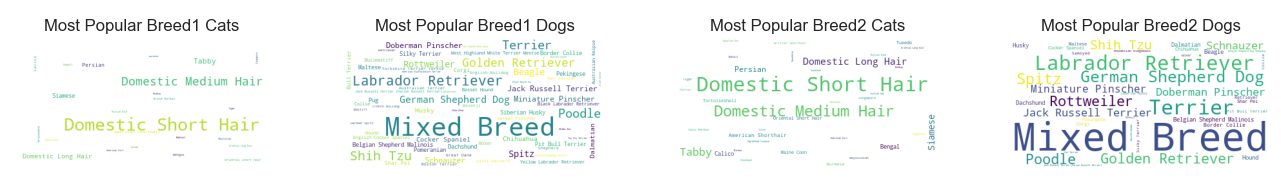
**Slika 10:** Raspodela brzine usvajanja u odnosu na postojanje videa i fotografija

Što se tiče dužine krzna, najzastupljeniji su ljubimci sa kratkim krznom, ali su takvi ljubimci sporije usvajani od ljubimaca sa srednjim i dugačkim krznom. Ljubimaca sa dugačkim krznom ima najmanje ali kod njih je zabeležena veća brzina usvajanja. Što se tiče same veličine ljubimca, najzastupljeniji su srednji ljubimci koji su najsporije usvajani, dok veoma velikih ljubimaca ima znatno najmanje, ali takvi ljubimci su brže usvajani od ostalih.

Boja ljubimca gotovo i da ne utiče na brzinu usvajanja, a na osnovu parametara **Color1**, **Color2** i **Color3** kreiran je novi parametar **Colorfulness** koji svedoči o šarenosti ljubimaca i koji može imati jednu od sledeće tri vrednosti: 1 – ljubimac je jednobojan, 2 – ljubimac ima 2 boje, 3 – ljubimac ima 3 boje. Međutim, analiza je pokazala da ovaj parametar apsolutno ne utiče na brzinu usvajanja, tako da se može i zanemariti.

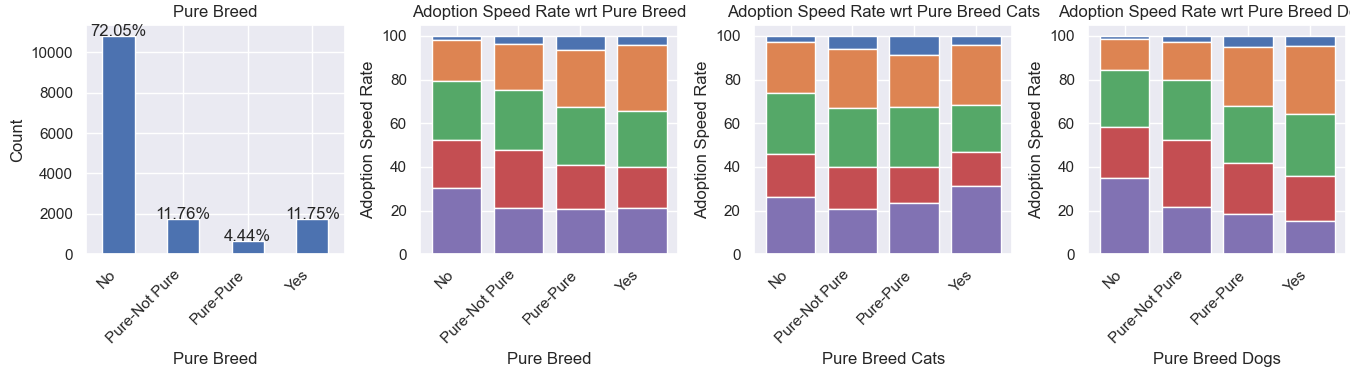
Što se tiče grupe ljubimaca datih na usvajanje, najviše ima individualnih ljubimaca, a jedna grupa može brojati i do 20 ljubimaca, mada su grupe koje broje 6 ili više ljubimaca znatno manje pristunije od ostalih grupa. Brzina usvajanja kod takvih grupa je znatno manja. Između ostalih grupa nisu zabeležene preterane razlike u brzini usvajanja.

Rasa ljubimca je još jedan logičan parametar koji bi mogao uticati na brzinu usvajanja. Na raspolaganju je 307 različitih rasa mačaka i pasa. Da li je neki ljubimac rasan ili nije je interesantan parametar koji vredi razmotriti. Međutim, među postojećim rasama postoje i „rase“ na osnovu kojih se može zaključiti da su pojedini ljubimci mešanci. Takve rase su na primer **Mixed Breed, Domestic Short Hair, Domestic Medium Hair, Domestic Long Hair**. Ove rase su i najzastupljenije.



**Slika 11:** Zastupljenost rasa među ljubimcima

Kreiran je novi parametar **PureBreed**, koji određuje koliko je ljubimac rasan. Može imati sledeće vrednosti: 0 – mešanac, 1 – čistorasni ljubimac, 2 – mešanac dve čiste rase, 3 – mešanac čiste rase i mešanca. Naravno, najzastupljniji su mešanci, dok mešanaca dve čiste rase ima najmanje. Zanimljivo je primetiti da je kod pasa ovaj faktor dosta bitan kod brzine usvajanja, pa se tako može izvući zaključak da su čisto rasni psi najbrže usvajani, dok su mešanci najsporije usvajani. Posmatrajući oba tipa ljubimca, najbrže su usvajani mešanci dve čiste rase.



**Slika 12:** Zastupljenost čistorasnih ljubimaca i mešanaca i raspodela klasa brzine usvajanja u odnosu na mešance i čistorasne ljubimce

Što se tiče osobe/organizacije koja je prijavila ljubimca na usvajanje, postoje i organizacije koje su prijavile i više od 300 ljubimaca. Među prvih 5 osoba/organizacija sa najvećim brojem prijavljenih ljubimaca, osoba/organizacija sa najvećim brojem ljubimaca beleži najmanji procenat neusvojenih ljubimaca. Osobe/organizacije sa brojem prijavljenih ljubimaca su svrstane u jednu kategoriju.

Iako naizgled deluje da je novčana nadoknada dosta bitan faktor koji može uticati na to da li će ljubimac biti usvojen i kojom brzinom, ovde to nije slučaj. Najveći je broj ljubimaca za koje ne postoji nikakva novčana nadoknada, međutim brzina usvajanja kod takvih ljubimaca se gotovo i ne razlikuje od ljubimaca sa novčanom nadoknadom između 210 i 3000, mada je takvih ljubimaca dosta manje.