CLEANING THE DATASET

|  |  |
| --- | --- |
| **Team ID** | **PNT2022TMID42057** |
| **Project Name** | **Car Resale value Prediction** |

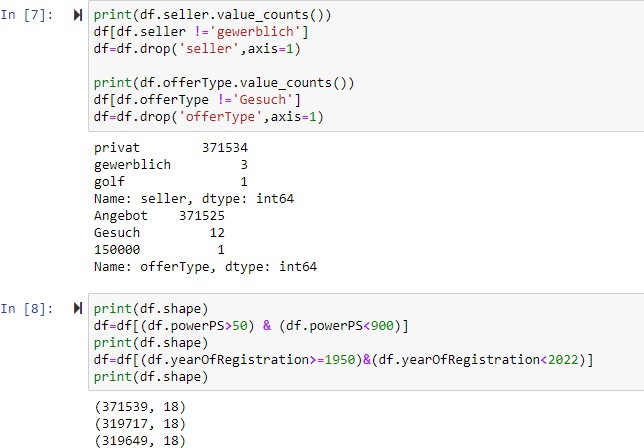
**CLEANING THE DATASET**

print(df.seller.value\_counts()) df[df.seller !='gewerblich'] df=df.drop('seller',axis=1)

print(df.offerType.value\_counts()) df[df.offerType !='Gesuch'] df=df.drop('offerType',axis=1) print(df.shape)

df=df[(df.powerPS>50) & (df.powerPS<900)] print(df.shape)

df=df[(df.yearOfRegistration>=1950)&(df.yearOfRegistration<2022)] print(df.shape)



df.drop(['name','abtest','dateCrawled','nrOfPictures','lastSeen','postalCode','dateCreated'], axis='columns',inplace=True)

new\_df=df.copy()

new\_df=new\_df.drop\_duplicates(['price','vehicleType','yearOfRegistration','gearbox','powerPS','model','kilo meter','monthOfRegistration','fuelType','notRepairedDamage'])

new\_df.gearbox.replace(('manuell','automatik'),('manual','automatic'),inplace=True) new\_df.fuelType.replace(('benzin','andere','elektro'),('petrol','others','electric'),inplace=True)

new\_df.vehicleType.replace(('kleinwagen','cabrio','kombi','andere'),('samll car','convertible','combination','others'),inplace=True)

new\_df.notRepairedDamage.replace(('ja','nein'),('Yes','No'),inplace=True)

new\_df=new\_df[(new\_df.price>=100)&(new\_df.price<=150000)]

new\_df['notRepairedDamage'].fillna(value='not-declared',inplace=True) new\_df['fuelType'].fillna(value='not-declared',inplace=True) new\_df['gearbox'].fillna(value='not-declared',inplace=True) new\_df['vehicleType'].fillna(value='not-declared',inplace=True) new\_df['model'].fillna(value='not-declared',inplace=True)

new\_df.to\_csv("autos\_preprocessed.csv")

