eda

January 10, 2021

1 EDA

2 Überblick der Daten

Wir lesen die CSV-Daten ein und verschaffen uns einen Überblick über die Anzahl Objekte und Features.

[4]:
```
42175937
```

[5]:
```
(153627, 69)
```

[6]:

	Id	AreaLiving	AreaProperty	BuiltYear	FloorNumber	ForestDensityL	\
0	7135329	140.0	501.0	2016	NaN	0.418964	
1	7170979	143.0	277.0	2004	NaN	0.033259	
2	7172246	160.0	712.0	1945	NaN	0.000000	
3	7172252	351.0	496.0	2016	NaN	0.037575	
4	7172733	400.0	1800.0	1975	NaN	0.095162	

	ForestDensityM	ForestDensityS	GroupNameDe	HouseObject	…	\
0	0.555985	0.730714	Haus	True	…	
1	0.074061	0.076468	Haus	True	…	
2	0.000000	0.000000	Haus	True	…	
3	0.000000	0.000000	Haus	True	…	
4	0.097193	0.153314	Haus	True	…	

	gde_private_apartments	gde_social_help_quota	gde_tax	gde_workers_sector1	\
0	1358.0	3.660512	8.73	17.0	
1	3476.0	3.634717	6.13	0.0	
2	2806.0	2.512344	9.79	167.0	
3	131.0	1.734104	9.15	12.0	
4	1181.0	1.056052	2.97	0.0	

	gde_workers_sector2	gde_workers_sector3	gde_workers_total	\
0	162.0	358.0	537.0	
1	2250.0	2787.0	5041.0	
2	1694.0	1138.0	2999.0	
3	10.0	17.0	39.0	
4	27.0	701.0	732.0	

```
   location_has_street  location_is_complete  PurchasePrice
0                    0                     0       745000.0
1                    1                     1       780000.0
2                    0                     0       570000.0
3                    0                     0       920000.0
4                    0                     0      3950000.0

[5 rows x 69 columns]
```

Das CSV beinhaltet 153'627 Spalten (Objekte) und 69 Kolonnen (Features). Mittels `head()` betrachten wir uns die ersten 5 Einträge. Da die Standardeinstellung des Notebooks die 69 Features mittig schneidet, geben wir die alle Features separat aus.

[7]:
```
Index(['Id', 'AreaLiving', 'AreaProperty', 'BuiltYear', 'FloorNumber',
       'ForestDensityL', 'ForestDensityM', 'ForestDensityS', 'GroupNameDe',
       'HouseObject', 'LastUpdate', 'Latitude', 'Locality', 'Longitude',
       'Name', 'NoisePollutionRailwayL', 'NoisePollutionRailwayM',
       'NoisePollutionRailwayS', 'NoisePollutionRoadL', 'NoisePollutionRoadM',
       'NoisePollutionRoadS', 'PopulationDensityL', 'PopulationDensityM',
       'PopulationDensityS', 'RealEstateTypeId', 'Renovationyear',
       'RiversAndLakesL', 'RiversAndLakesM', 'RiversAndLakesS', 'Rooms',
       'SourceId', 'StateShort', 'StreetAndNr', 'TravelTimeMiv',
       'WorkplaceDensityL', 'WorkplaceDensityM', 'WorkplaceDensityS', 'Zip',
       'distanceToTrainStation', 'gde_area_agriculture_percentage',
       'gde_area_forest_percentage', 'gde_area_nonproductive_percentage',
       'gde_area_settlement_percentage', 'gde_average_house_hold',
       'gde_empty_apartments', 'gde_foreigners_percentage',
       'gde_new_homes_per_1000', 'gde_politics_bdp', 'gde_politics_cvp',
       'gde_politics_evp', 'gde_politics_fdp', 'gde_politics_glp',
       'gde_politics_gps', 'gde_politics_pda', 'gde_politics_rights',
       'gde_politics_sp', 'gde_politics_svp', 'gde_pop_per_km2',
       'gde_population', 'gde_private_apartments', 'gde_social_help_quota',
       'gde_tax', 'gde_workers_sector1', 'gde_workers_sector2',
       'gde_workers_sector3', 'gde_workers_total', 'location_has_street',
       'location_is_complete', 'PurchasePrice'],
      dtype='object')
```

3 Fehlende Werte

Wir werfen einen Blick auf die `NAs`, also die fehlenden Werte. Folgende Tabelle listet diese je Feature auf.

[9]:
```
                        feature  number_na
0                   FloorNumber      87695
1                Renovationyear     138326
2                   StreetAndNr      46435
```

3	TravelTimeMiv	2
4	gde_area_agriculture_percentage	2
5	gde_area_forest_percentage	2
6	gde_area_nonproductive_percentage	2
7	gde_area_settlement_percentage	2
8	gde_average_house_hold	2
9	gde_empty_apartments	2
10	gde_foreigners_percentage	2
11	gde_new_homes_per_1000	2
12	gde_politics_bdp	32803
13	gde_politics_cvp	5058
14	gde_politics_evp	29595
15	gde_politics_fdp	2920
16	gde_politics_glp	18442
17	gde_politics_gps	11113
18	gde_politics_pda	91628
19	gde_politics_rights	30535
20	gde_politics_sp	2046
21	gde_politics_svp	893
22	gde_pop_per_km2	2
23	gde_population	2
24	gde_private_apartments	2
25	gde_social_help_quota	2
26	gde_tax	2
27	gde_workers_sector1	2
28	gde_workers_sector2	2
29	gde_workers_sector3	2
30	gde_workers_total	2

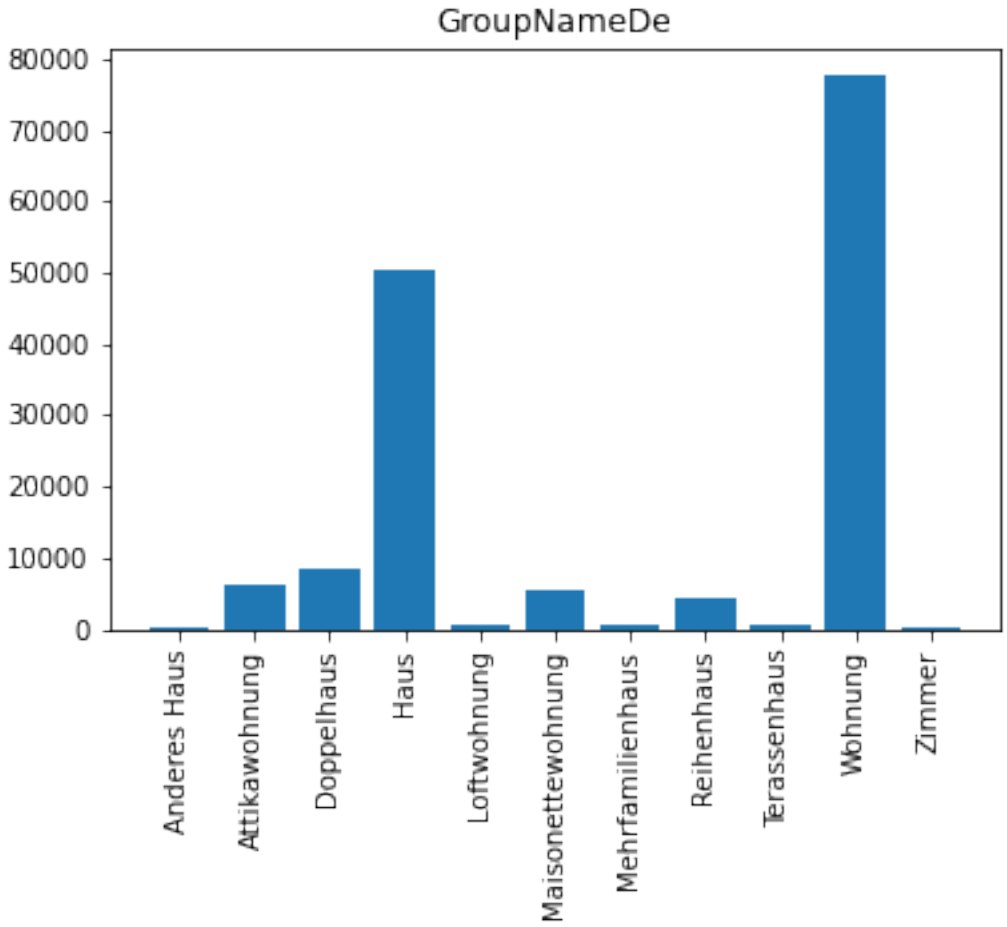
Wir sehen, dass von den insgesamt 69 Features Werte bei 30 Features fehlen. Bei mehr als der Hälfte dieser, fehlen sehr wenige, sodass diese vernachlässigbar sind. Diese werden für die weitere Analyse gelöscht.

Bei weiteren Featuers fehlt hingegen ein beachtlicher Teil an entsprechenden Angaben, beispielsweise: - Renovationyear: 138'326 - FloorNumber: 87'695

Speziell beim Renovationsjahr dürfen die fehlenden Werte aber nicht als nicht vorhandene Information gedeutet werden. Vielmehr wurde das Haus womöglich schlicht (noch?) nicht renoviert.

Eine ähnliche Aussage kann auch für die `FloorNumber` gemacht werden: ist das Objekt ein ganzes Haus, so scheint die Angabe einer Etage nicht nötig resp. sinnvoll.

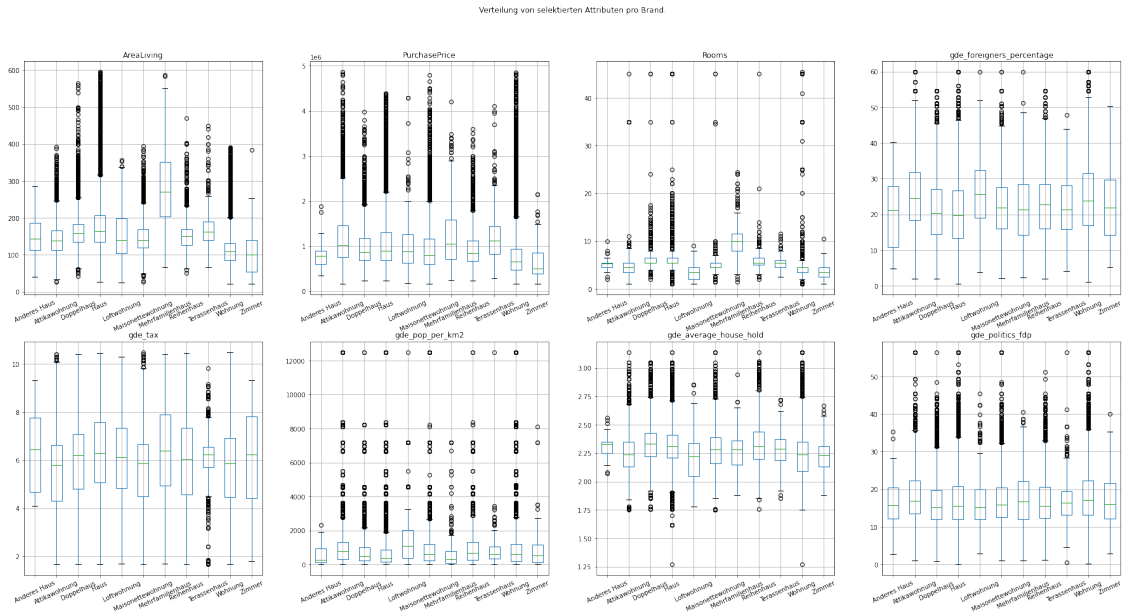
4 Grobüberblick von numerischen Attributen



[10]:
```
count      153627
unique         11
top       Wohnung
freq        77499
Name: GroupNameDe, dtype: object
```

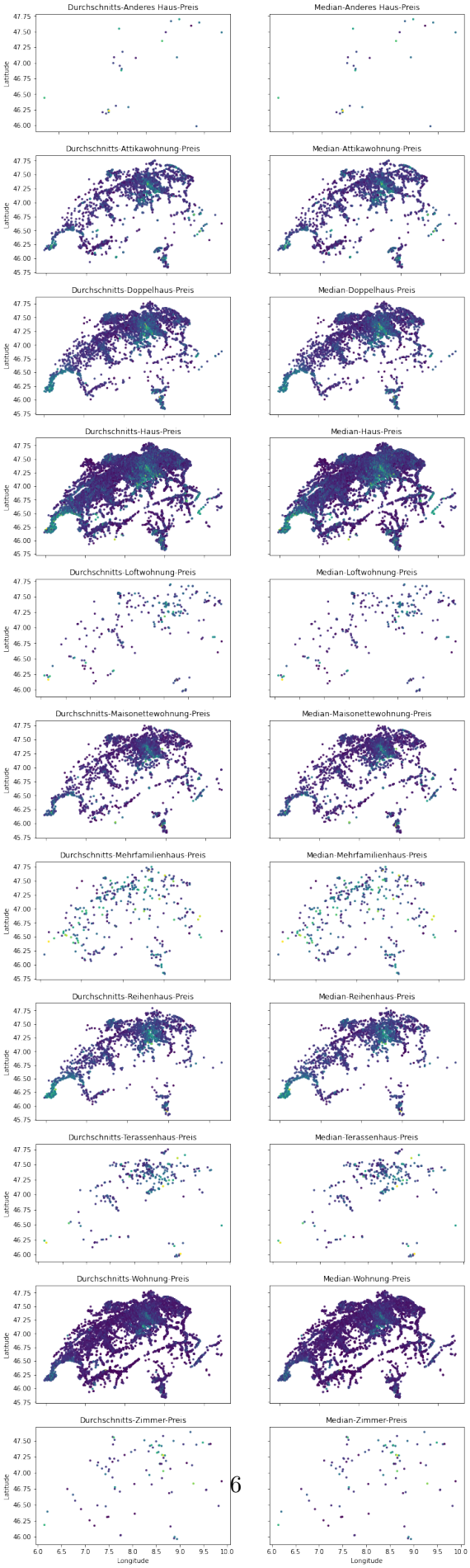
```
/Users/ericwinter/opt/miniconda3/envs/py38/lib/python3.8/site-
packages/pandas/plotting/_matplotlib/tools.py:201: UserWarning: When passing
multiple axes, layout keyword is ignored
  warnings.warn(
/Users/ericwinter/opt/miniconda3/envs/py38/lib/python3.8/site-
packages/pandas/plotting/_matplotlib/boxplot.py:385: UserWarning: When passing
multiple axes, sharex and sharey are ignored. These settings must be specified
when creating axes
  ax = boxplot(
```

```
[11]: array([<AxesSubplot:title={'center':'AreaLiving'}>,
             <AxesSubplot:title={'center':'PurchasePrice'}>,
             <AxesSubplot:title={'center':'Rooms'}>,
             <AxesSubplot:title={'center':'gde_foreigners_percentage'}>,
             <AxesSubplot:title={'center':'gde_tax'}>,
             <AxesSubplot:title={'center':'gde_pop_per_km2'}>,
             <AxesSubplot:title={'center':'gde_average_house_hold'}>,
             <AxesSubplot:title={'center':'gde_politics_fdp'}>], dtype=object)
```



Im obigen Plot, der einige Features dem `GroupNameDe`, also dem "Haustyp", gegenüberstellt, lässt sich ein Trend erkennen, der aber im Allgemeinen ähnlich ist – will heissen: die Boxplots verlaufen je Häusertyp (grob vereinfacht) ähnlich; und dort, wo sie abweichen, erscheinen sie plausibel. Z. B. scheint klar, dass ein Mehrfamilienhaus mehr Räume hat, als eine Loftwohnung.

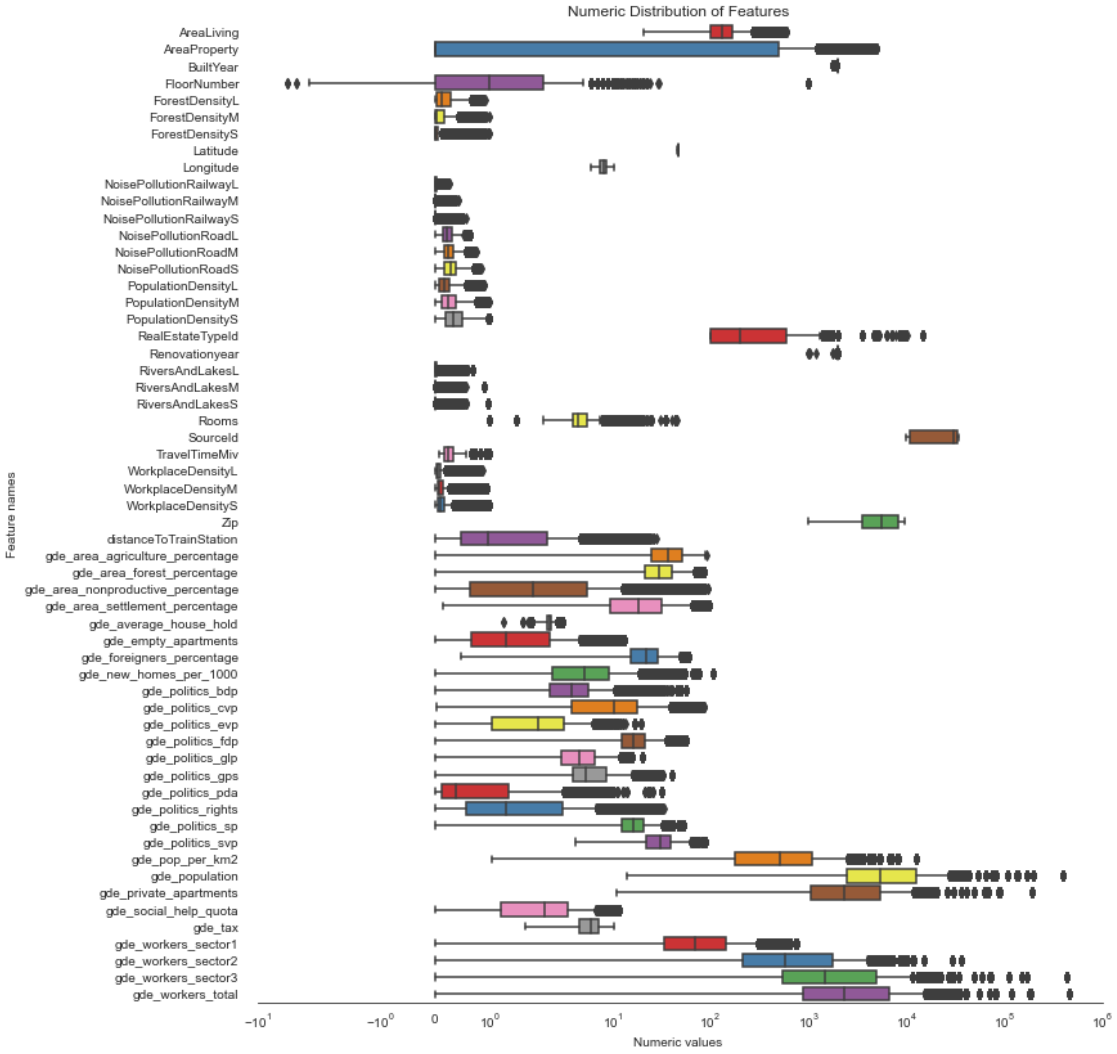
Im folgenden Plot untersuchen wir nun, wie die Lage einer Immobilie Einfluss auf ihren Preis hat. Wir nehmen dazu den Durchschnittspreis pro `GroupNameDe`.


Durchschnitts-Anderes Haus-Preis
Median-Anderes Haus-Preis
Durchschnitts-Attikawohnung-Preis
Median-Attikawohnung-Preis
Durchschnitts-Doppelhaus-Preis
Median-Doppelhaus-Preis
Durchschnitts-Haus-Preis
Median-Haus-Preis
Durchschnitts-Loftwohnung-Preis
Median-Loftwohnung-Preis
Durchschnitts-Maisonettewohnung-Preis
Median-Maisonettewohnung-Preis
Durchschnitts-Mehrfamilienhaus-Preis
Median-Mehrfamilienhaus-Preis
Durchschnitts-Reihenhaus-Preis
Median-Reihenhaus-Preis
Durchschnitts-Terassenhaus-Preis
Median-Terassenhaus-Preis
Durchschnitts-Wohnung-Preis
Median-Wohnung-Preis
Durchschnitts-Zimmer-Preis
Median-Zimmer-Preis
Latitude
Longitude

Deutlich lässt sich erkennen, dass die Lage einer Immobilie Einfluss auf ihren Preis hat. Die – für ihre hohen Preise bekannten – Regionen am/um den Zürichsee, Genfersee, Luganersee und die Umgebung um St. Moritz "leuchten" auf. Die hellen Farben symbolisieren einen hohen, die dunklen Farben einen tiefen Immobilienpreis.

Bei der Prediction wird der Standort eine wichtige Rolle einnehmen; entweder durch die Longitude und Latitude oder durch die Postleitzahl.

Interessant ist auch die Verteilung aller numerischen Werte auf der Symlog-Skala. Dies ist eine Skala, welche – ausser man ist nahe bei 0 – der Log-Skala entspricht. Um 0 herum ist diese linear, womit die Explosion des Logs auf "+/- unendlich" verhindert wird.



Viele Features sind schief-verteilt. Mit Hilfe einer Box Cox Transformation lässt sich diese Eigenschaft abschwächen. Somit könnte unser Modell stabiler werden.

Die “Skewness” ist das Verhältnis vom Mittelwert zur Standardabweichung: desto höher, desto “skewter”.

```
There are 20 numerical features with Skew > 0.5 :
```

[14]:
```
RiversAndLakesS           9.870534
RealEstateTypeId          6.261766
Rooms                     5.363506
NoisePollutionRailwayS    4.905221
RiversAndLakesM           4.858933
RiversAndLakesL           3.624726
NoisePollutionRailwayM    3.609659
WorkplaceDensityL         3.284066
distanceToTrainStation    3.208185
AreaProperty              3.008660
WorkplaceDensityM         2.661331
NoisePollutionRailwayL    2.628042
ForestDensityS            2.562313
WorkplaceDensityS         2.399048
ForestDensityM            1.647688
AreaLiving                1.607749
ForestDensityL            1.007906
PopulationDensityL        1.002755
PopulationDensityM        0.647705
NoisePollutionRoadL       0.540166
dtype: float64
```

5 Wie korrelieren die numerischen Werte mit dem Preis?

Wir berechnen die Korrelation der einzelnen Features mit dem Verkaufspreis mit Hilfe des Spearman-Koeffizienten. Dieser betrachtet die Kovarianz mit den Standardabweichungen von den zu vergleichenden Attributen und vergleicht deren Ränge. Die Ränge der Werte sind die nach der Reihenfolge der nach der Grösse seiner

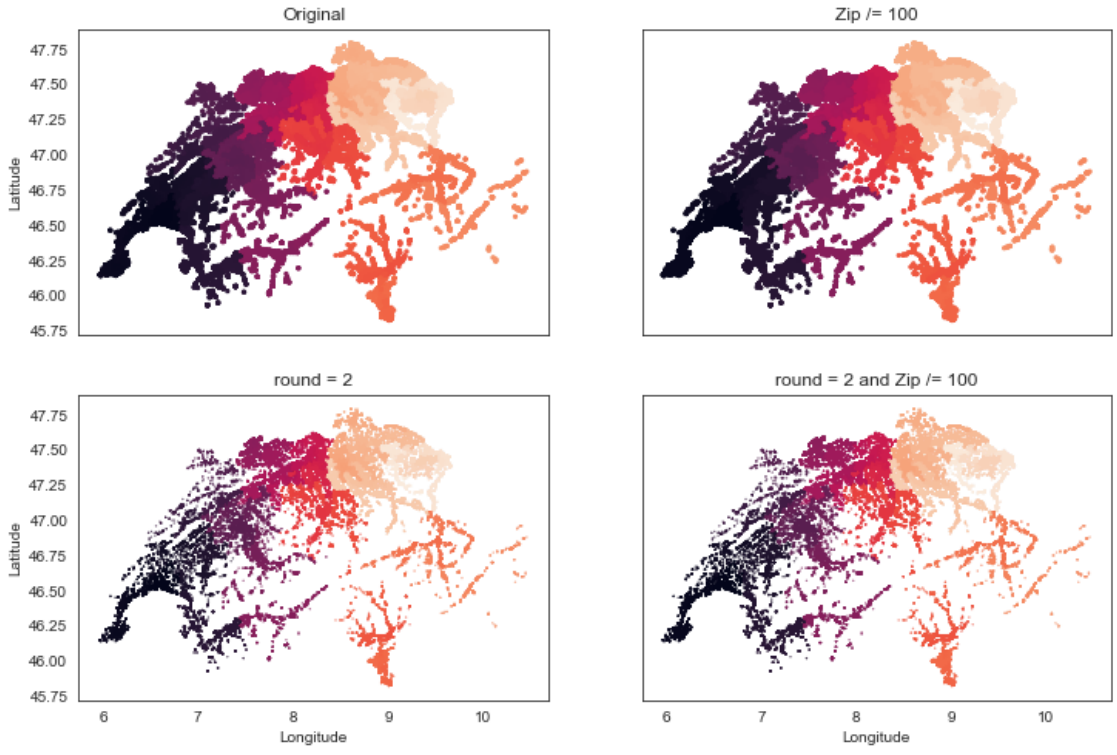
Der Spearman-Koeffizient wird auch als Rangkorrelationskoeffizient genannt, weil er die Korrelation nicht zwischen den Datenpunkten selbst, sondern zwischen ihren Rängen berechnet.

[15]:
```
AreaLiving                          0.677319
AreaProperty                        0.302468
BuiltYear                           0.156482
HouseObject                         0.260701
PopulationDensityL                  0.118138
RealEstateTypeId                    0.247618
Rooms                               0.508444
TravelTimeMiv                      -0.301922
WorkplaceDensityL                   0.131320
WorkplaceDensityM                   0.101484
gde_area_nonproductive_percentage  -0.135720
```

```
gde_area_settlement_percentage       0.219687
gde_empty_apartments                -0.186640
gde_foreigners_percentage            0.132322
gde_politics_bdp                    -0.203626
gde_politics_fdp                     0.212381
gde_politics_gps                     0.144317
gde_politics_svp                    -0.115015
gde_pop_per_km2                      0.231552
gde_population                       0.150899
gde_private_apartments               0.146832
gde_tax                             -0.176127
gde_workers_sector3                  0.143093
gde_workers_total                    0.118875
PurchasePrice                        1.000000
Name: PurchasePrice, dtype: float64
```

Korrelationen können nur zwischen Zahlen und nicht zwischen Zeichen (Strings) berechnet werden.

Bei der Postleitzahl darf nicht vorschnell ein Zusammenhang zwischen ihrer Höhe und dem Verkaufspreis konstruiert werden. Die Höhe einer Postleitzahl hat keine besondere Bedeutung, vielmehr hat sie kategorialen Charakter. Da die Postleitzahl viele Werte annehmen kann, müssen wir diese Werte one-hot-encoden. Das führte aber zu (zu) vielen "Einzelkolonnen", weswegen wir hier vereinfachen: wir kürzen die Postleitzahl auf 2 oder 3 Stellen.

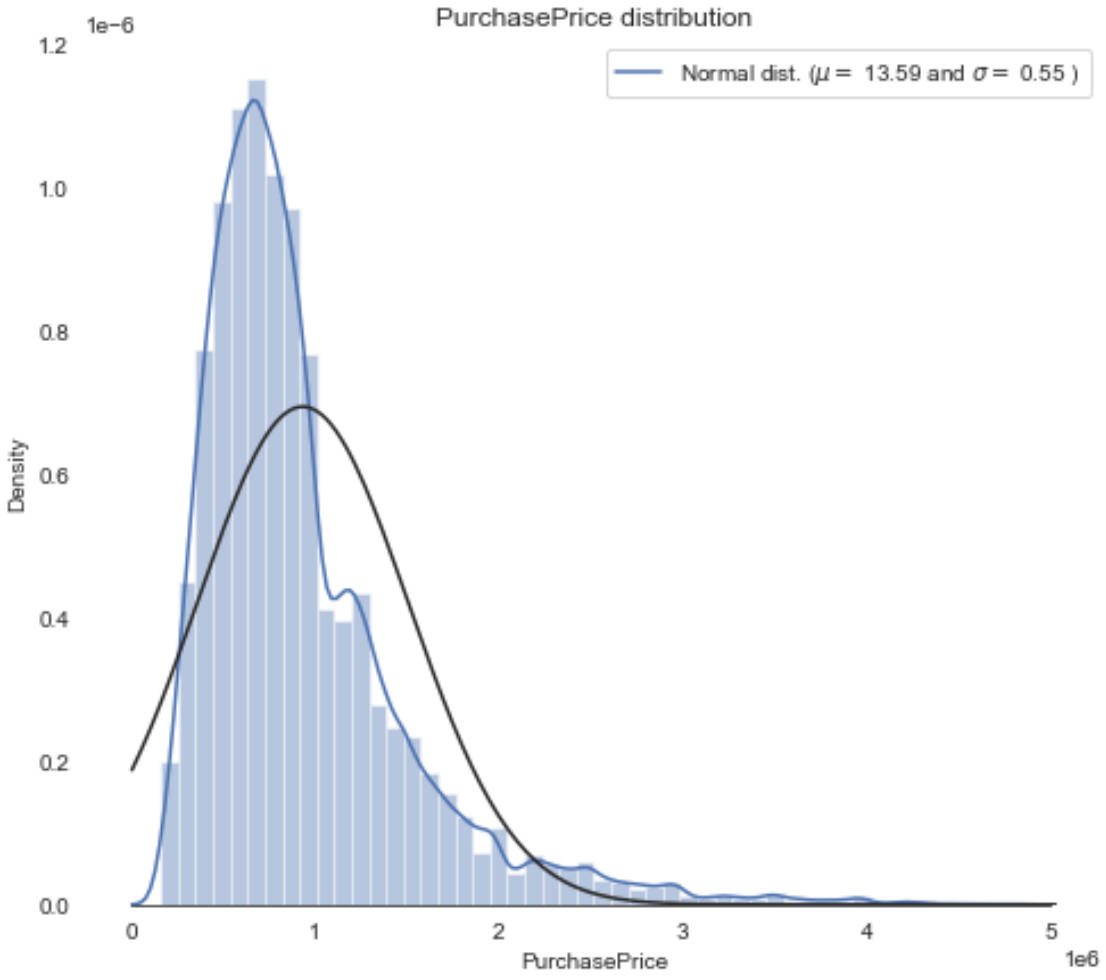


Mit den Postleitzahlen können die einzelnen "Orte" sehr genau lokalisiert resp. zugeordnet werden. Auch beim Kürzen und Runden der Postleitzahl ist die Lokalisation nach wie vor möglich. Wir beschränken uns somit auf die ersten beiden Ziffern der Postleitzahl und verzichten auf die exakte Position mittels Longitude und Latitude.

6 Verteilung des Kaufpreises

```
/Users/ericwinter/opt/miniconda3/envs/py38/lib/python3.8/site-
packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a
deprecated function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar flexibility)
or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
mu = 13.59 and sigma = 0.55
```



Der Verkaufspreis ist nicht normalverteilt. Wir nehmen somit den `log` und stellen seine Verteilung erneut dar.

```
/Users/ericwinter/opt/miniconda3/envs/py38/lib/python3.8/site-
packages/seaborn/distributions.py:2557: FutureWarning: `distplot` is a
deprecated function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar flexibility)
or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

```
mu = 13.59 and sigma = 0.55
```