Benchmark Model: Simple Linear Regression

Group: Firat Saritas, Simon Staehli

This notebook describes our workflow of an initial try with a Simple Linear Regression and how we improved the simple model by detection/removal of outliers and transformation of the attributes. More attempts regarding to this are available in our notebooks listed in sources.

Data Import

At the beginning we start to import the file and take a quick look whether the data has been imported correctly and what it looks like.

---

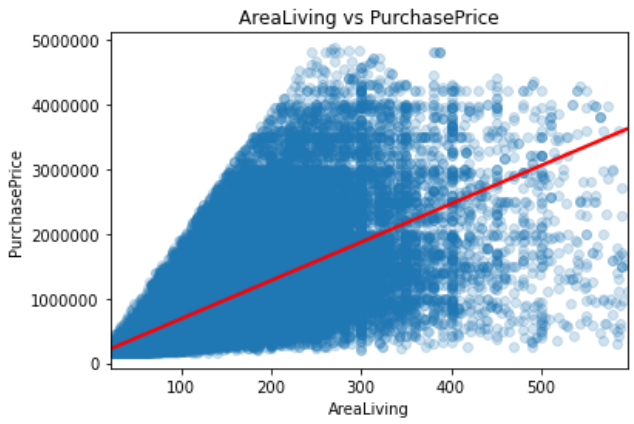
1. Simple Linear Regression (Non-transformed)

This will be our first try to fit a regression to our data and analyze the results of it. Therefore we already have done Exploratory Analysis to detect outliers, missing and unlogical data. The notebooks are listed in the sources. Select the data for the simple linear regression with two features. In this case we have the features:

X = 'AreaLiving' as independent variable

y = 'PurchasePrice' as dependent variable

Let us investigate those two attributes by plotting them in a scatterplot.



The plot shows us the dependency between those two attributes. In red we fitted a linear regression to the data. We can see that there is a certain limit for each size of the Living Area regarding to its PurchasePrice. From 300m2 there is like a break acc. to the Price of the object. For each Price for a 1qm wie have the largest variances in PurchasePrice for the bigger objects with are bigger than 250m2. We think to predict the higher prices will be more problematic than the lower prices.

1.1 Model-Building

Now we build our model. Therefore we perform a train-test-split of our data with the ratio 80/20 and fit our data afterwards.

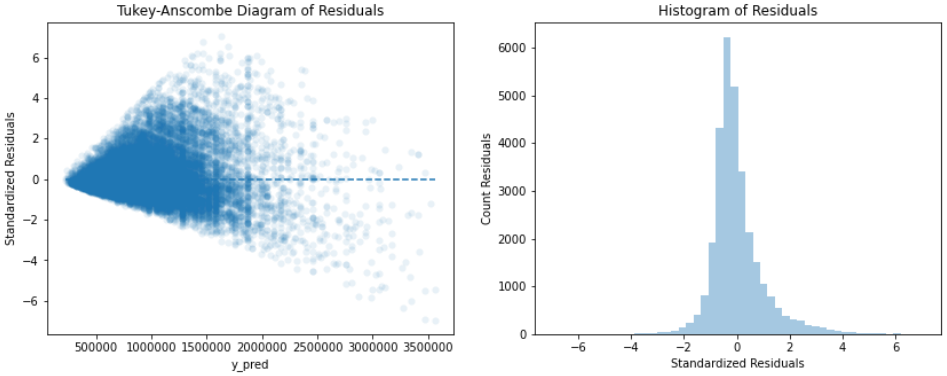
```
R2 Score:  0.3914312193071785
```

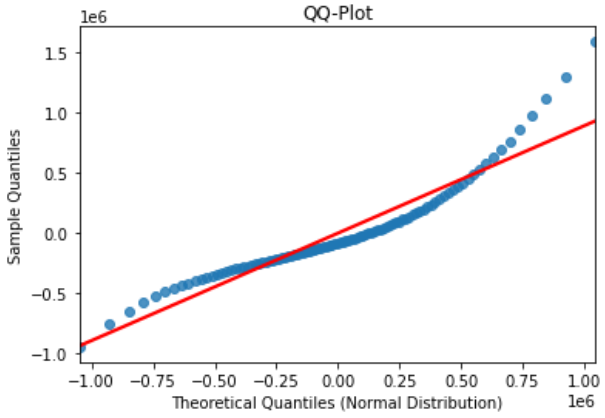
The R2-Score reached indicates a bad fit, but from now it can be used as a comparisson for further work on the model.

1.2 Analysis of Residuals

Plot the Residuals and check the underlying distribution of the them.

```
Mean ABS Error:  309996.07
Mean ABS Percentage Error:  0.3763321471343064
Median ABS Error:  217698.19
```





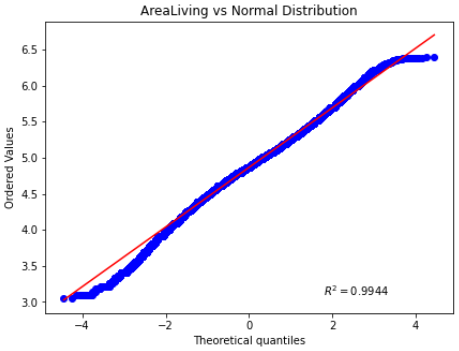
Observations:

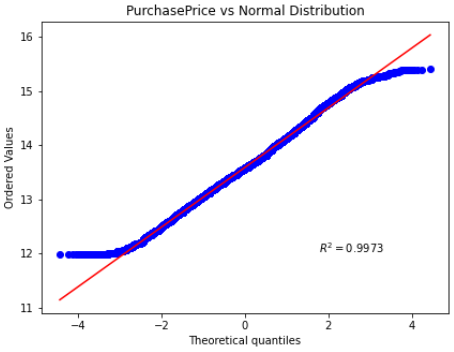
We can see on the Tukey-Anscombe Diagram of Resiudals, that we predict expensive properties too low. We expected this to happen, because of the large variance of the PurchasePrice for bigger objects. We need a transformation which decreases the higher values and increases the lower values. The distribution of the data is not normal. We can see skewness of the right side. To get a closer look to the quantiles we compared it with the normal distribution. Especially in the upper quantiles we have the biggest deviation as well as for lower quantiles from the normal distribution. Our predictions do not match with the normal distribution throughout the price range.

1.3 Transformation of Attributes

Our goal is to achieve a normal distribution of the data with the transformation. Therefore we chose the box-cox transformation to achieve this. We have tried other transformation as well, but this transformation delivered us the best results. All other attempts are visible in other notebooks listed in the sources.

$$\lambda = 0 : \log(X) \quad \text{or} \quad \lambda \neq 0 : \frac{X^\lambda - 1}{\lambda}$$






```
<function __main__.box_cox(lmbda_X, lmbda_y)>
```

Now we test the box-cox transformed input data of Arealiving and PurchasePrice against the theoretical sample quantiles of a normal distribution with same mean and standard deviation as our sample data. The R2-Score indicates the fit fo the regression line on our sample and theoretical quantiles.

AreaLiving:

As we just increase the lambda value for X, we can see an S-shaped curve which does not fit the normal distribution at all. With a transformation of X with lambda 0.1 we get the best results for approximating a Normal distribution of our sample data.

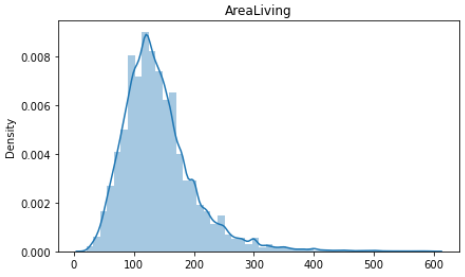
PurchasePrice:

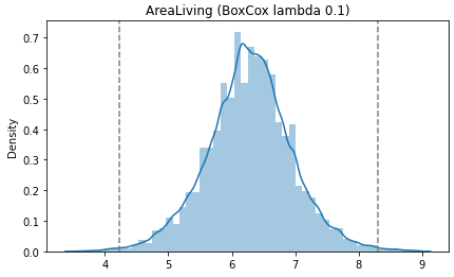
Same here with the "AreaLiving". In our case a lambda factor of 0 which means that a logarithmic transformation makes most sense for this attribute, because it fits best with the theoretical quantiles of a normal distribution.

1.4 Handling of Outliers

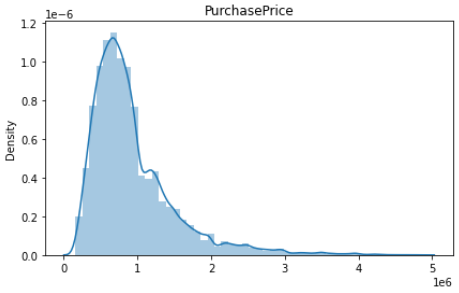
We decided to remove the outliers according to the z-values from the standard normal distributionen. The Z-Score tells us how many standard deviations a data point is away from the Mean. We take 3 times standard deviation as threshold.

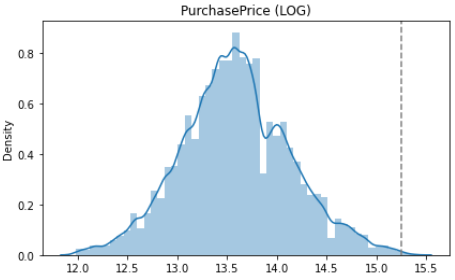
$$z = \frac{x - \overline{x}}{s_{\overline{x}}} = \frac{x - \mu}{\sigma}$$





On the left side we have the untransformed distribution of AreaLiving on the right side the transformed distribution of AreaLiving with lambda 0.1. We can see that we have a bunch of upper outlier as well as lower outliers of the attribute AreaLiving. They are visualized by the grey-dotted line in the right distribution plot.





Above the distribution of the PurchasePrice is visualized. On the left-hand side the unstransformed attribute and on the right the transformed attribute with the logarithm. It's visible that we only have upper outliers in the Log-transformed attribute which is an indicator for right-skewed distribution.

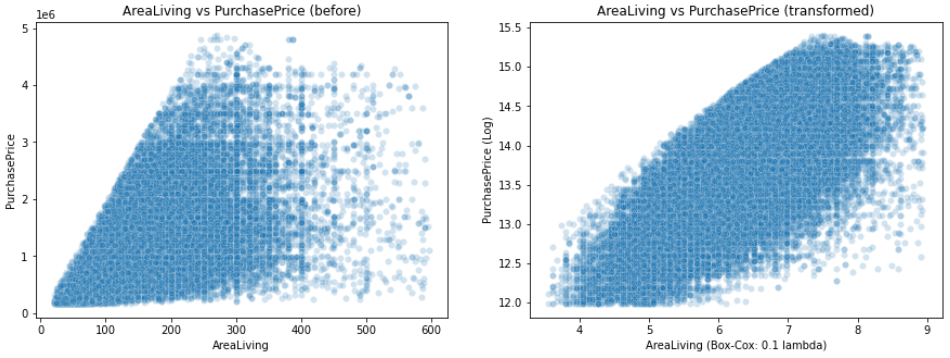
Now we want to remove the outliers, that have a higher score than 3 and a lower score than -3:

---

2. Regression with transformed Attributes

2.1 Check Before and After the Transformation

Now we will have a look on our initial scatter plot with the untransformed attributes and compare it with the scatterplot of the transformed scatterplot.



On the left-hand side we can see both attributes plotted against each other like at the beginning. and on right-hand side the scatterplot with the two transformed attributes. If one would chose which plot to draw a line in, which describes the data best, it would be clearly in the right plot.

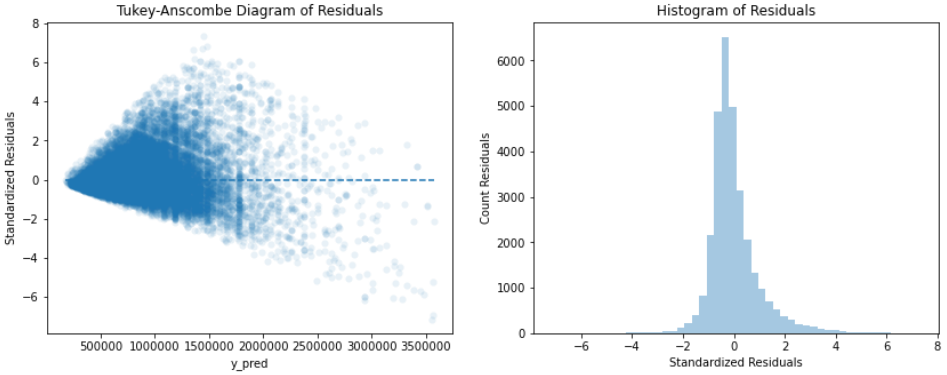
2.2 Train Model

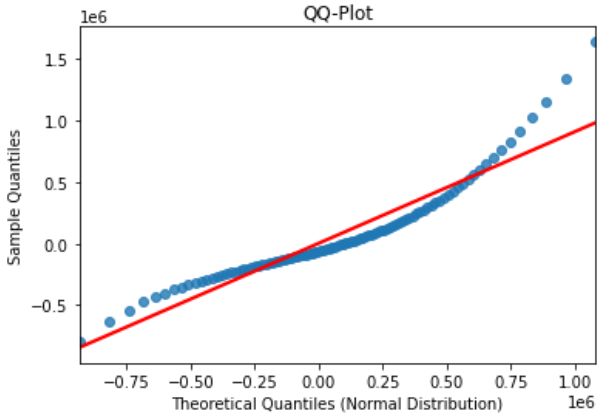
```
R2 Score:  0.48025863655786294
```

```
The first unchanged model had a R2 score of:  39.143 %
The new transformed model has a R2 score of:  48.026 %
Thus we have an increase in the R2 score of:   8.883000000000001 %
```

2.3 Model Metrics

```
Mean ABS Error:  295680.38
Mean ABS Percentage Error:  0.3283729524242097
Median ABS Error:  187457.5
--------------------------------------------------
Residual Plot and Histogram after the Transformation
```





Observations:

As we can see on the Residuals plot on the upper left, we have the same pattern in our data. Of course it is to mention that we removed some data from AreaLiving, which we suspected as outlier, that is why the the pattern in our residuals plot looks now more compressed than before. Also the distribution of the residuals has not approximated to a normal distribution. Although we have had not eliminated the pattern in the residual plot, we were able to increase our model score by approximately 10%.

Conclusion:

In the end we can say that the creation of this model helped us a lot to understand the target variable (PurchasePrice) better and what results are possible to get with just a Simple Regression Model.

---

3. Sources

APA-Generator: https://www.scribbr.ch/zitieren/apa-generator/ (https://www.scribbr.ch/zitieren/apa-generator/)

Notebooks:

Simon Staehli, S. S. (2020). Simon_Test.ipynb [Jupyter Notebook]. Simon Staehli.

Firat Saritas, F. S. (2020). ML_Firat.ipynb [Jupyter Notebook]- Firat Saritas.

Firat Saritas, F. S. & Simon Staehli, S. S. (2020). EDA_Master.ipynb [Jupyter Notebook].

Webpages:

Michael Graber, M. G. (2020). Immobilienpreisrechner. DS-Spaces. https://ds-spaces.technik.fhnw.ch/immobilienrechner/ (https://ds-spaces.technik.fhnw.ch/immobilienrechner/)

Mahbubul Alam, M. A. (2020). Towards Data Science. Towards Data Science. https://towardsdatascience.com/z-score-for-anomaly-detection-d98b0006f510 (https://towardsdatascience.com/z-score-for-anomaly-detection-d98b0006f510)

Scott, D. S. (2004). Box Cox Transformation. Onlinestatbook. http://onlinestatbook.com/2/transformations/box-cox.html (http://onlinestatbook.com/2/transformations/box-cox.html)