JIFEM IDS PROJECT

How accurately can we predict the price of a house using 12 variables?

Data Cleaning

```
house_prices_sel <- house_prices %>%
  mutate(TotalBath = BsmtFullBath +
           BsmtHalfBath +
           HalfBath +
           FullBath) %>%
```

Creating a new column 'TotalBath', as the sum of all the bathrooms in each house, to summarise

Id	GarageCars	BedroomAbvGr	TotalBath	BldgType	LotArea	Neighborhood	KitchenQual	PavedDrive	TotalBsmtSF	Electrical	OverallQual	ExterQual	YearBuilt
1	2	3	4	1Fam	8450	CollgCr	3	1	856	SBrkr	7	3	2003
2	2	3	3	1Fam	9600	Veenker	3	1	1262	SBrkr	6	3	1976
3	2	3	4	1Fam	11250	CollgCr	3	1	920	SBrkr	7	3	2001
4	3	3	2	1Fam	9550	Crawfor	2	1	756	SBrkr	7	2	1915
5	3	4	4	1Fam	14260	NoRidge	3	1	1145	SBrkr	8	3	2000
6	2	1	3	1Fam	14115	Mitchel	3	1	796	SBrkr	5	4	1993
7	2	3	3	1Fam	10084	Somerst	2	1	1686	SBrkr	8	3	2004
8	2	3	4	1Fam	10382	NWAmes	2	1	1107	SBrkr	7	2	1973
9	2	2	2	1Fam	6120	OldTown	3	1	952	FuseF	7	3	1931
10	1	2	2	2fmCon	7420	BrkSide	2	1	991	SBrkr	5	2	1939
11	1	3	2	1Fam	11200	Sawyer	3	1	1040	SBrkr	5	3	1965
12	2	4	4	1Fam	11924	Nridght	3	1	1175	SBrkr	9	3	2005
							2	1	912	SBrkr	5	2	1962
							2	1	1494	SBrkr	7	3	2006
							2	1	1253	SBrkr	6	2	1960
							3	1	832	FuseA	7	3	1929
							3	1	1004	SBrkr	6	3	1970
							3	1	0	SBrkr	4	3	1967
						W	3	1	1114	SBrkr	5	3	2004
							3	1	1029	SBrkr	5	3	1958

We hoped applying these changes would sufficiently clean the data, making it easier to handle and more accurate when it came to predicting Sale Price

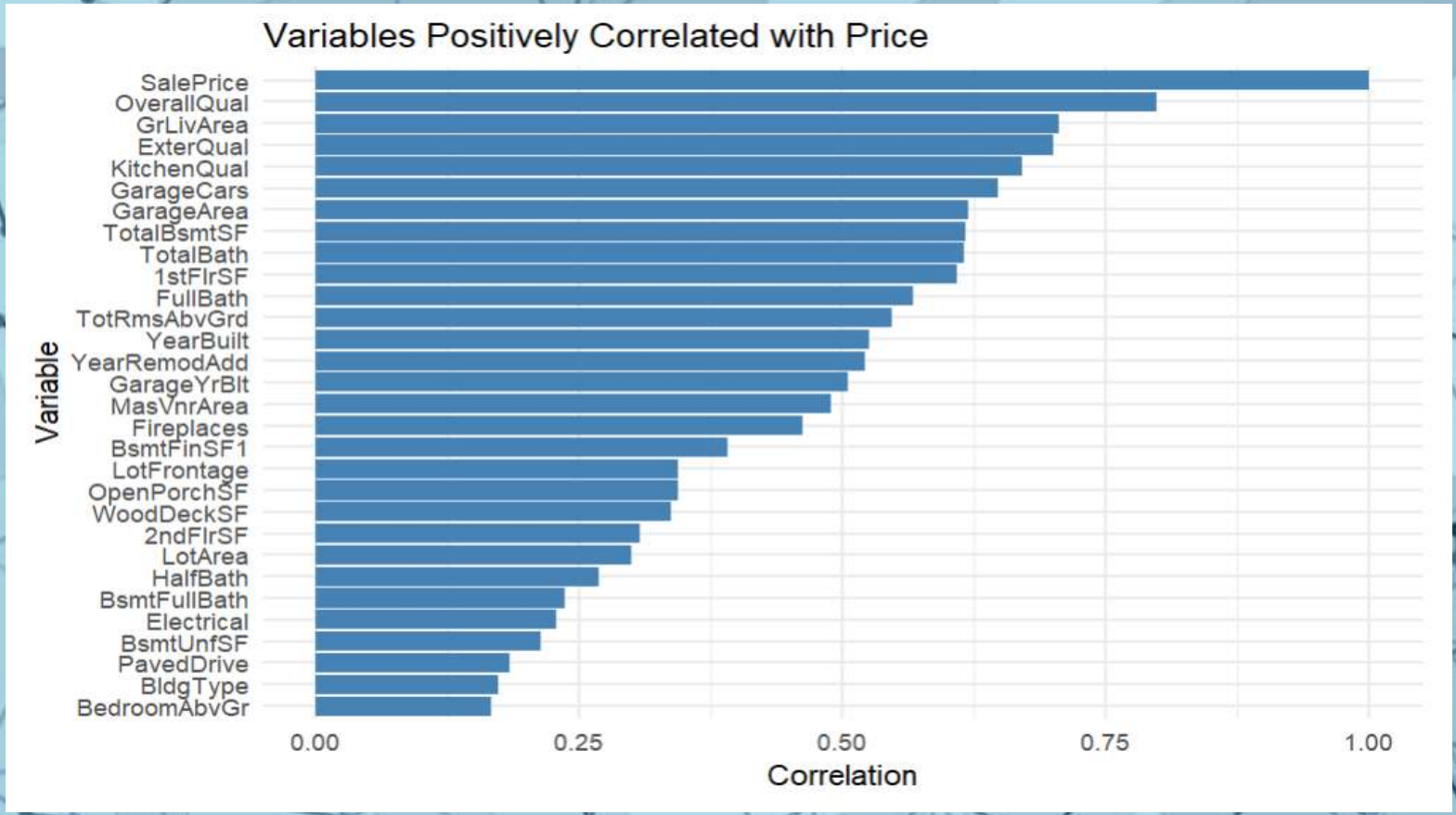
```
qual_map <- c("Ex" = 5,
              "Gd" = 4,
              "TA" = 3,
              "Fa" = 2,
              "Po" = 1)


# Apply mapping
house_prices_sel$KitchenQual <-
qual_map[house_prices_sel$KitchenQual]
house_prices_sel$ExterQual <-
qual_map[house_prices_sel$ExterQual]
house_prices_sel$HeatingQC <-
qual_map[house_prices_sel$HeatingQC]
house_prices_sel$HeatingQC <-
qual_map[house_prices_sel$HeatingQC]
```

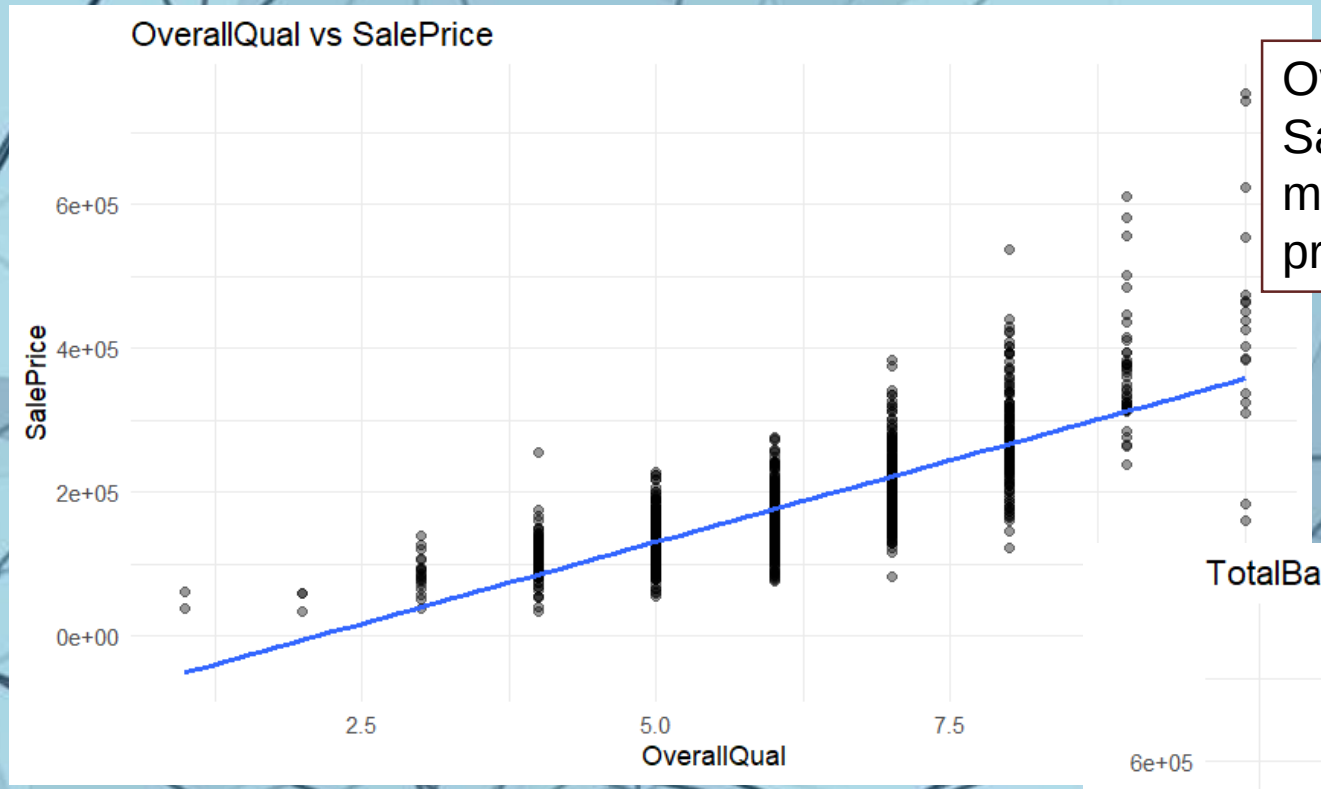
- Changing categorical values to numerical values to help create models and understanding the effect of these explanatory variables.
- We changed all the variables describing the 'quality' of aspects of the house, with Excellent = 5 and Poor =1.

Whether or not a house had a paved drive proved to be a highly predictive variable, we changed it from being a categorical variable of the form 'yes', 'no' to being binary, where 1 is yes and 0 is no.

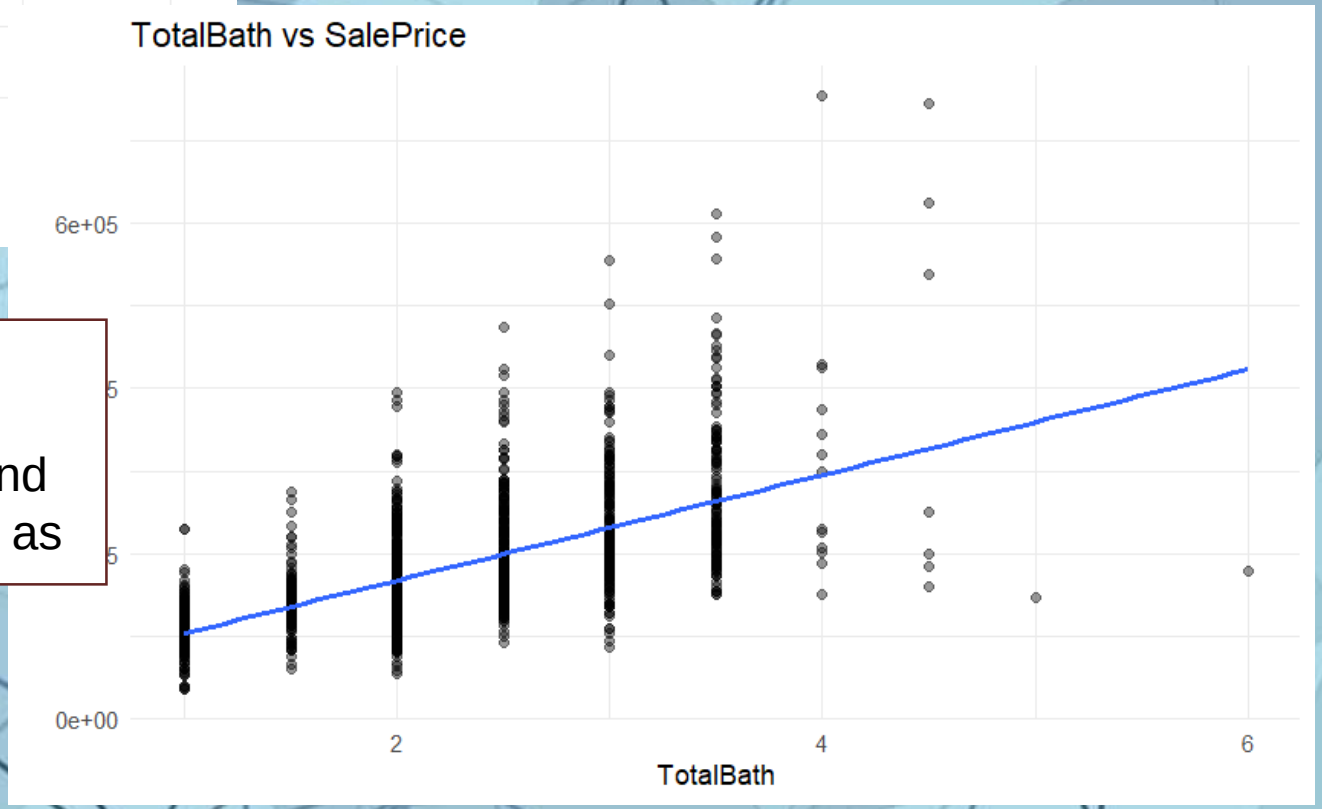
Selecting variables



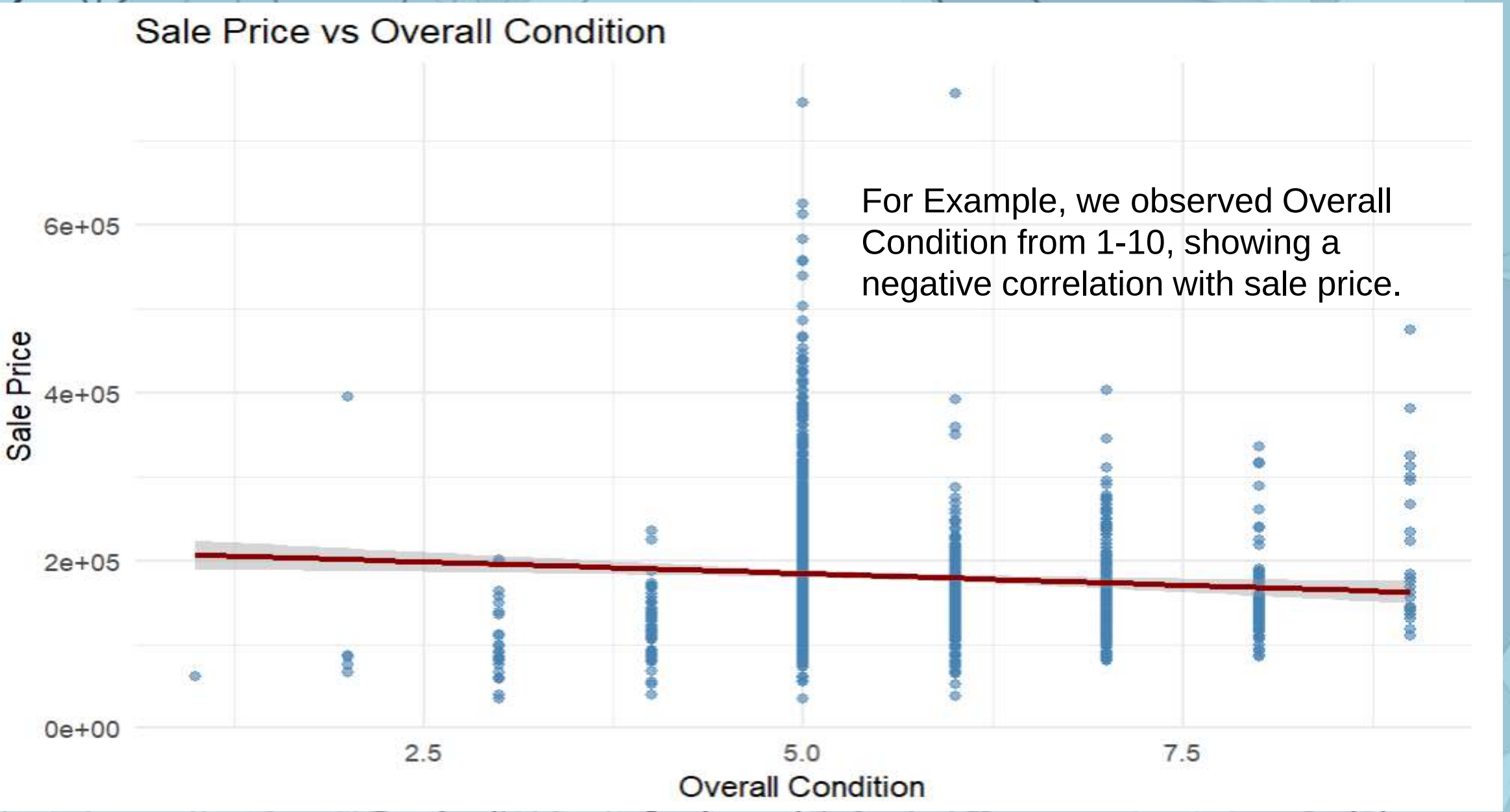
- We originally selected our predictive variables with a bit of guesswork.
- Eventually, we made this correlation bar chart to select the most predictive Variables to use.
- Some of the results from this surprised us.



Overall Quality shows a strong positive correlation with Sale Price. Higher quality homes consistently sell for more, making OverallQual one of the most powerful predictors in the model.



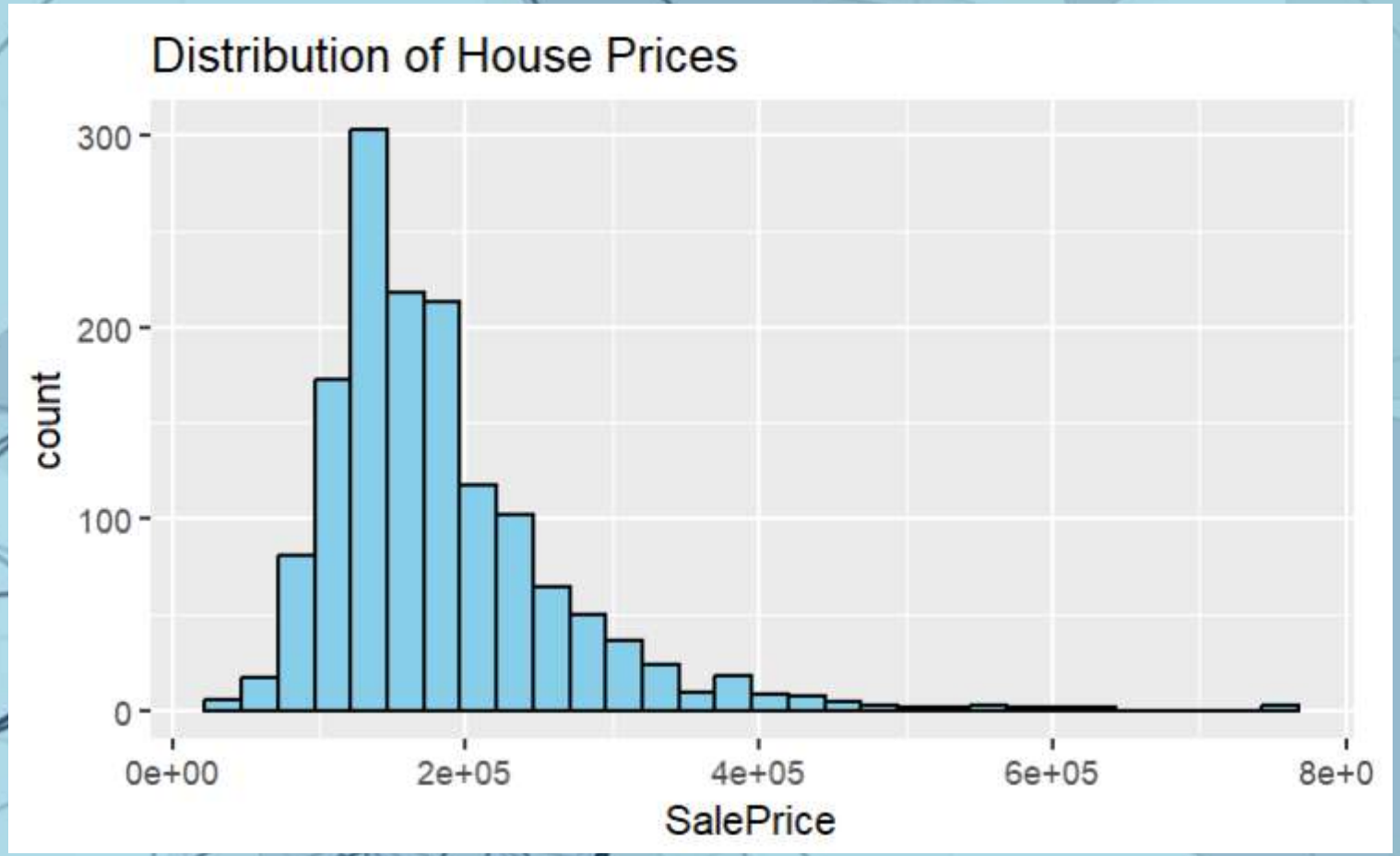
This graph shows that Total Bath has a strong correlation to sale price hence why it is a good predictor to use in our model. The values at 2.5 and 3.5 are due to HalfBath (toilets no bath) counting as 0.5 bathrooms.



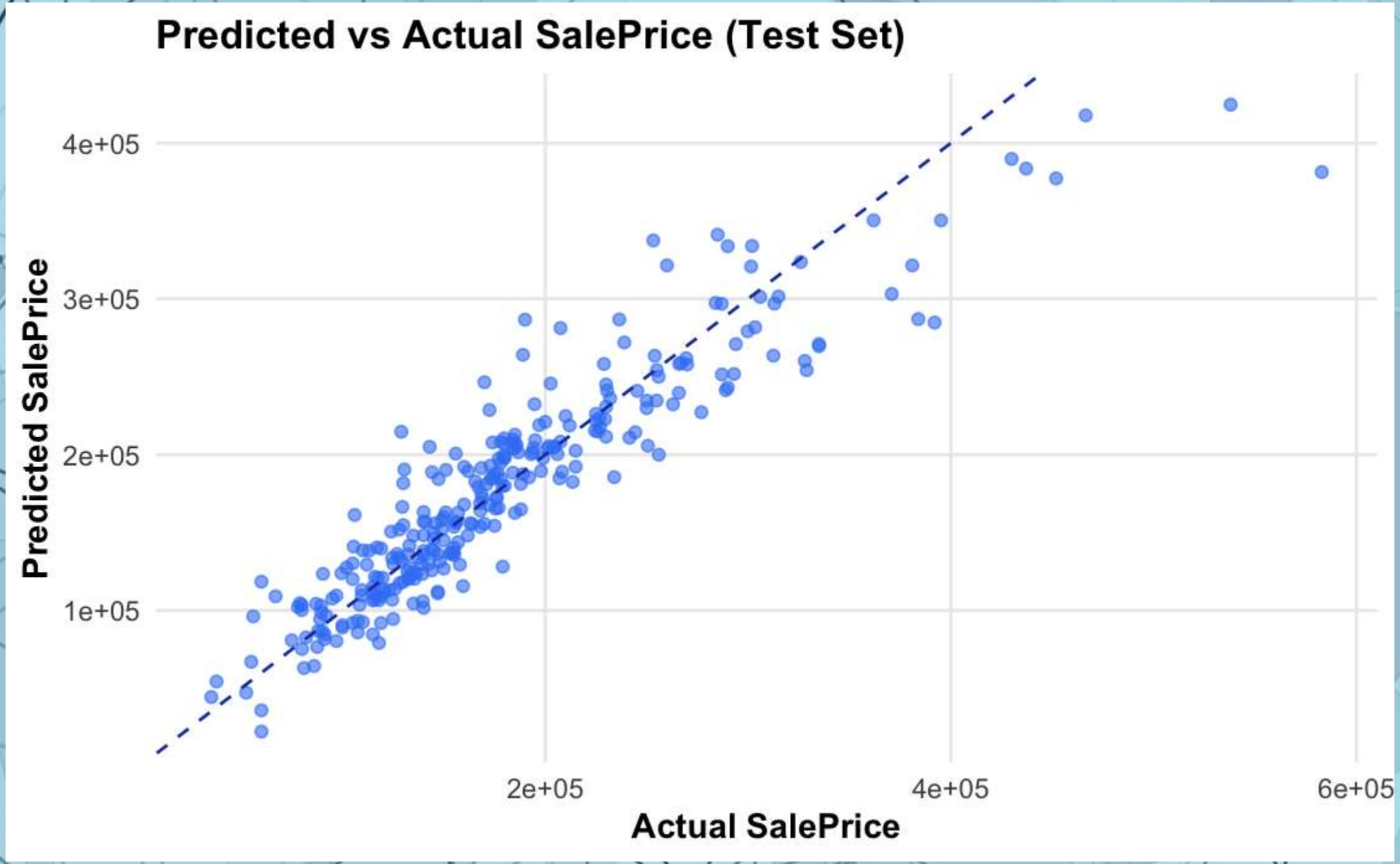
For Example, we observed Overall Condition from 1-10, showing a negative correlation with sale price.

Sale Price

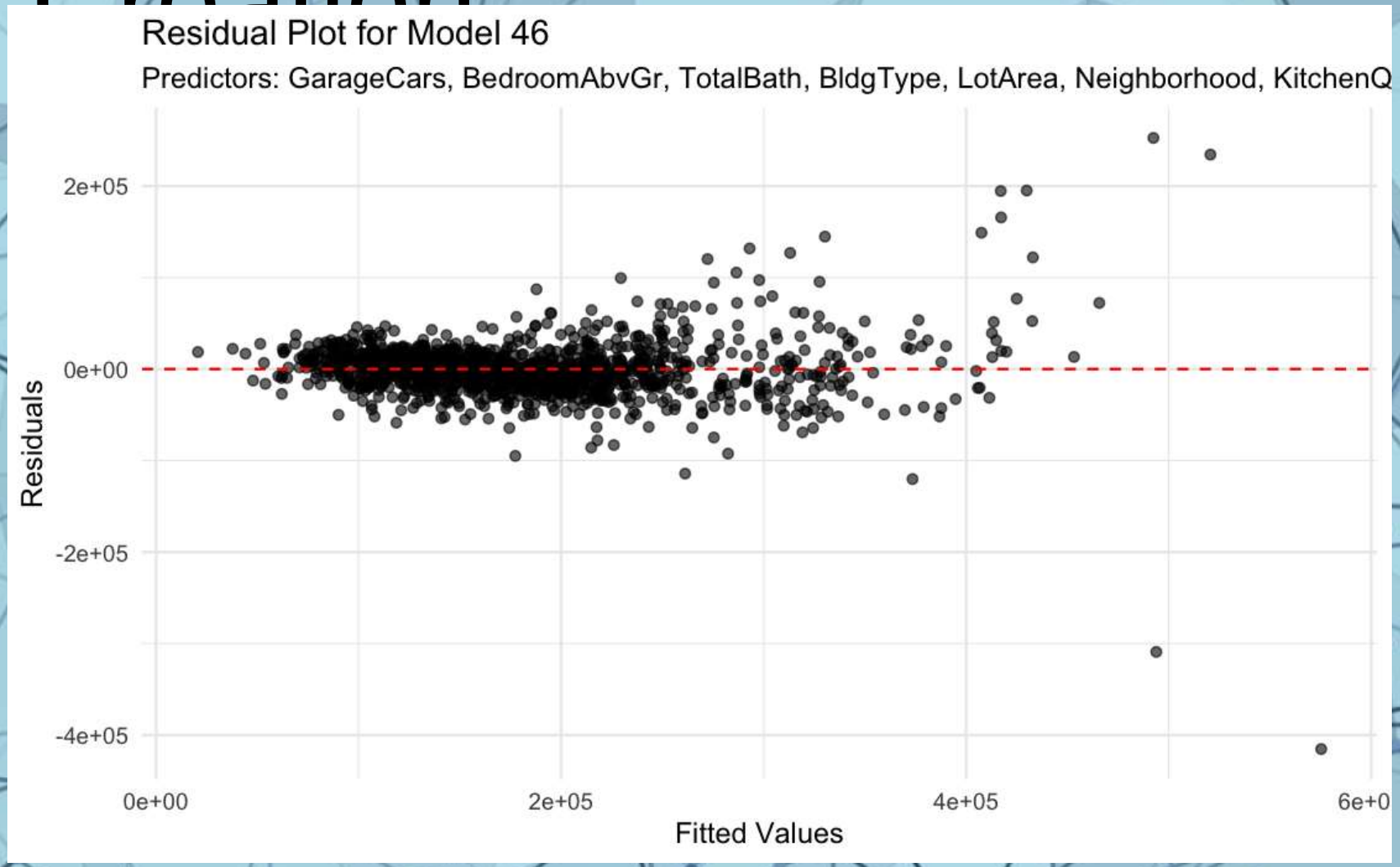
- Range and IQR are both large, confirming sale price would be a suitable variable to try and predict.
- We can also observe this in the bar chart, as we can see the data is spread evenly with enough grouping to suggest it isn't random
- Significant right skew could cause inaccuracies for more expensive



mean	median	IQR	range
180921.2	163000	84025	720100

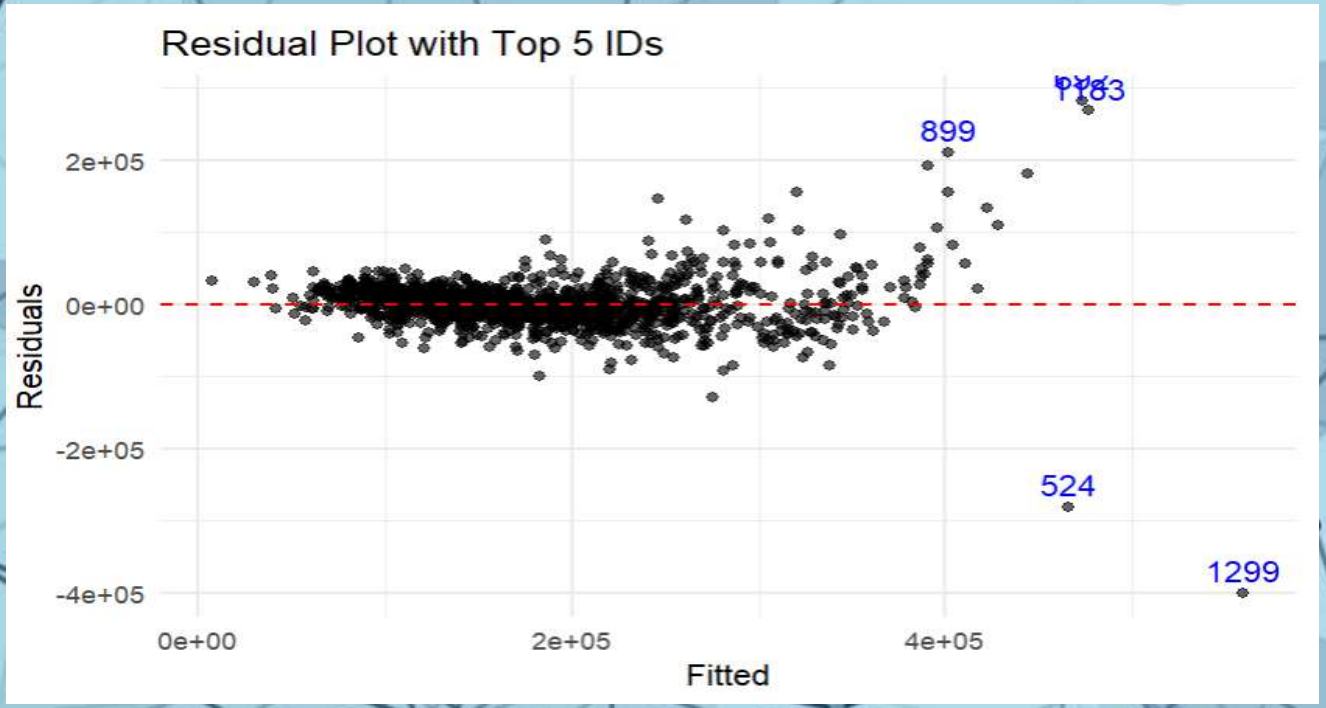

Predicted vs Actual SalePrice (Test Set)
Predicted SalePrice
4e+05
3e+05
2e+05
1e+05
2e+05
4e+05
6e+05
Actual SalePrice

Model Creation



- We used a linear regression model as it's output is far more suitable for the data type, than, for example, a logistic regression model. Which gives binary outputs
- After modelling the house price with various combinations of predictive variables, we settled on these 12 to make the model with the highest R^2 value (0.849).
- We can see that the majority of residual values all lie close to zero, besides some anomalies which we can look at in further detail

Limitations



The house with ID 1299 clearly has the greatest residual value. This value is negative implying the Sale Price was a lot lower than expected. When identifying particularly strong predictive values:

BldgType	TotRmsAbvGrd
1Fam	12

YearBuilt
2008

TotalBsmtSF	OverallQual
6110	10

All but one of our predictive values would suggest a far greater sale price for this house. However, Heating QC = 1, as the scale for Heating QC only runs from 1-5, which is vague, so having a Heating QC of one could imply the house has little to no heating, a massive flaw and something which would affect the Sale Price, yet isn't suitably taken into account by our data, which in turn affects the model.

HeatingQC
1

Ethical considerations

- Despite the accuracy of our model, it fails to take into account things like historical valuation patterns and zoning policies which are both things that affect real housing markets.
- Linear models make complex interactions more simple, this means relying on predictions in the absence of the correct context may lead to misrepresented property values.
- The dataset we used focuses on a specific area and period of time. The patterns here may reinforce local biases and will make the model less accurate for other markets.
- Misuse of this model such as using these house price predictions for financial decisions without proper oversight may lead to things like contribution to inequities and unfair valuation of houses.
- Our data cleaning choices may influence the results we get, if we aren't clear about these assumptions it could lead to over-interpreting the model's predictions.

Thank You for Listening!