STA302 Final Group Project

Professor: Dr. Antonio Herrera Martin

Final Group 12

Binhe Jia	***1007651710***
William Kwan	***1007158898***
Xiaoxu (Rita) Liu	***1007854779***

Introduction

What is the most significant characteristic of a house that impacts its sale price?

Research Question
The most obvious characteristics that are known to determine a house's price are the total size in square feet of the house, the condition, the neighbourhood, the type of house (single, detached, condominium, etc), number of levels, number of rooms or bathrooms, and the age. We will use the house's sale price in dollars as the response variable and the initial predictors will be the characteristics mentioned. Comparing the magnitude and sign of each predictor coefficient given by a multiple linear regression model will help us to objectively determine which predictor or factor is the most significant.

Motivation
Over the past years, the demand for houses has surged, greatly outpacing the increase in housing supply (Day, 2018). In the US, the median sale price of a house has shot up from $317,100 to $442,600 in just a two year period from 2020 to 2022 (St.Louis Fed). While economic factors definitely play a role, understanding how the intrinsic characteristics determine housing prices will help buyers estimate prices and choose affordable homes that are aligned with the key advantages they value. A predictive model can also help real estate investors to decide market values of houses.

Related Literature
Several papers have discussed and modelled the significance of a property's characteristics in relation to its sale price. Focusing on differentiating crimes into violent and property crimes, Cigdem-Bayram's paper indicates that an increase in the per capita crime rate will lower property prices (Cigdem‐Bayram, 2019). Peng and Chen's article points out that the residential buildings quality and environmental living quality demonstrated statistically significant effects on house prices (Peng & Chen, 2016). Ebru's paper with similar variables as our data, supposes that having larger garage(s) will lead to a higher sale price (Ebru, 2011). Many studies however, lack specific and extensive housing attributes, such as the basement size, the number of bathrooms, and house condition, etc. This report will provide additional analysis towards such specific attributes and spur further discussion towards predicting house prices.

Methods
The data was compiled by Dean De Cook, uploaded on Kaggle for educational purposes. It is an overview of housing sales in Ames, Iowa in the US from 2006 to 2010, with 80 variables to describe the characteristics of a house sold.

Variable Selection
For each variable, we check the values to determine if it should be included in our initial model. Variables such as PoolQC, with a lot of NaN or 0 values, or variables that are within a domain (GarageCars, GarageFinish, etc are replaced in favour of GarageArea) are removed. Variables with a lack of variability such as Street and LandContour were removed.
Once we had our initial model, we first did a hypothesis test (t-test) on each predictor, $H_0: \beta_i = 0$ against $H_1: \beta_i \neq 0$. Predictors that cannot reject the null hypothesis (p-value > 0.05) suggest that it is irrelevant in determining the response variable. We also combine this

with ANOVA test on the entire model, H_0: $\beta_1 = \ldots = \beta_{p-1} = 0$ against H_A: at least some $\beta_i \neq 0$ to determine what variables to eliminate.

After, we formed reduced models and tested it against the initial model using a partial F-test. Rejecting H_0 would mean that at least one of the predictors removed $\beta_1, \ldots, \beta_k$ were significant. We also performed a t-test on the reduced model to confirm that the predictors we were using were statistically significant.

When we arrive at a final model, we use VIF to check for multicollinearity between the variables. If we see a VIF > 5, there would be dominant multicollinearity issues between variables. A VIF of around 1 would indicate small but not significant multicollinearity.

Model Validation

Once we have selected our potential models from the previous section, we first check conditional mean response by plotting the response values vs the fitted values. If the response values are following a linear fashion, then we can also use a pairs plot within our predictors to check for linearity. Any violation in linearity, i.e. following a curve instead of a line would deem the model flawed for further analysis.

Next, we can plot residual vs fitted, residual vs predictor and QQ plots to check for linearity, uncorrelated errors, and constant variance. If we see a pattern in any of the residual vs predictor plots or the residual vs fitted plot, it is a good sign that linearity is broken.

Moreover an increase or decrease in the spread of residuals, will mean a violation of constant variances of the errors. Since we assume normality of errors for hypothesis tests, we can check this condition using a QQ plot. If we see big deviations from the QQ-line, our model might be breaking the normality condition.

Finally, if we see any clustering in any of the residual plots, this can be attributed to a violation of uncorrelated covariances of the errors. This would most likely be a result of a poor selection of the predictor variables. Thus, if this occurs, we would have to find a different model with different predictors.

Model Violations and Diagnostics

To diagnose these issues, we will employ a variety of statistical techniques. If constant variance is violated, we can make a variance stabilising transformation on our response such as the square root or natural log. We can also use other transformations to deal with non-linearity in the predictors and response, such as a power transformation or a Box-Cox method. However, the Box-Cox method will give us a more complicated transformation, we will rely on it to determine a simpler one. If none of these are able to prevent the issues, then we should consider a different model with different predictors.

Results

Description of Data

Table 1a,b: Summary of variables in final model. For categorical variables, information about counts, categories, and proportion are displayed. For numerical variables, the min, quantiles, median, mean and max are displayed.

Summary of Variables			
	Neighbourhood Crime	**BldgType**	**BsmtQual**

Values And Counts	**Extremely High**: 288 (24.39%) **High**: 80 (0.07%) **Slightly Above Average**: 32 (0.03%) **Slightly Below Average:** 324 (27.43%) **Low**: 169 (14.31%) **Extremely Low**: 288 (24.39%)	**TownHouse**: 117 (9.90%) **Duplex**: 20 (1.70%) **1Fm**: 1024 (86.7%) **2FmCon**: 20 (1.70%)	**Typical/Average**: 544 (46.10%) **Excellent**: 112 (9.48%) **Fair**: 32 (2.71%) **Good**: 493 (41.74%)
	HouseStyle	**Functional**	
Values	**1Story**: 642 (54.36%) **1.5Story**: 142 (12.02%) **2Story**: 382 (32.35%) **2.5Story**: 15 (1.27%)	**Typical (No Defects)**: 1104 (93.48%) **Major Deductions (1 or 2)**: 14 (1.19%) **Minor Deductions (1 or 2)**: 51 (4.32%) **Moderate Deductions**: 11 (0.09%) **Severe Deductions**: 1 (0.0008%)	

	Lot Area	**Overall Cond**	**Year Built**	**Total BsmtSF**	**GrLiv Area**	**Garage Area**	**Sale Price**	**Total Baths**
Min	1300	2.000	1872	4.654	334	160.0	37900	1.000
Median	9675	5.000	1971	1004.0	1484	471.0	162000	2.000
Mean	11019	5.591	1970	1076.7	1538	468.7	182116	2.467
Max	215245	9.000	2010	6110.0	5642	1418.0	755000	6.000
Std.Dev	1.05e4	1.13	3.07e1	4.50e2	5.35e2	2.17e2	8.20e4	9.39e-1

Density Plots for numerical variables and histograms for categorical variables are shown in **App. Fig. 2**

Analysis Process

At first, our preliminary model consisted of 15 predictors (Appendix Table 1), giving us a R^2 of 0.8044. We performed the model reduction techniques mentioned in the methods section. A t-test on the initial model and ANOVA suggested keeping statistically significant variables that rejected the null hypothesis ($p < 0.05$). We tested that a reduced model with only 12 predictors (Table 1) was able to give a R^2 of 0.8037, so we chose this reduced model. A partial F-test confirmed our decision (p=0.74), suggesting that the variables we removed were not statistically significant. Finally, we performed StepAIC, using the initial model to start and allowing it to reduce variables. It was able to agree with our 12 variable reduced model. Using this reduced model, we performed a VIF check. Some of our variables had a VIF around 5, suggesting that there is multicollinearity. Density plots for response and most numerical variables also showed a normal distribution (Appendix Figure 2).

Goodness of Final Model and Diagnostics

First, we assessed all assumptions discussed in Methods. We saw that the response vs fitted plot was able to provide us with an appropriate linear conditional mean response as the response values were following the fitted line (Figure 1a). Furthermore, we didn't notice a pattern in any of the pairs plots (Appendix Figure 3).

Moving on to the residuals vs predictors plots, we saw that LotArea, TotalBsmtSF, GrLivArea, and GarageArea residuals (Figure 3) were increasingly more spread out as the predictor increased. This was also observed in the residuals vs fitted plot (Figure 2a). This would correspond to a slight violation of constant variance known as heteroskedasticity. Aside from that, we don't see any clustering or sequencing patterns in any of the residual

plots, so there is no correlation in the errors. In the QQ-plot (Figure 2b), there is a slight deviation from the QQ-line towards the end.

To fix these issues, we performed transformations on the numerical variables with heteroskedasticity. We first used BoxCox on the response variable, which suggested a (lambda) of near 0, meaning we should use a logarithm on the response. This was able to stabilise the variance on the response variable. We then also used the logarithm transformation on the aforementioned numerical variables, which was able to stabilise the variance towards the larger predictor values of the plot. Lastly, we did not see any issues with the box plots on the categorical variables (Figure 4).

Fig 1. Response vs Fitted (Prior to transformation left, Post Transformation right). Pairs plot in Appendix 4,

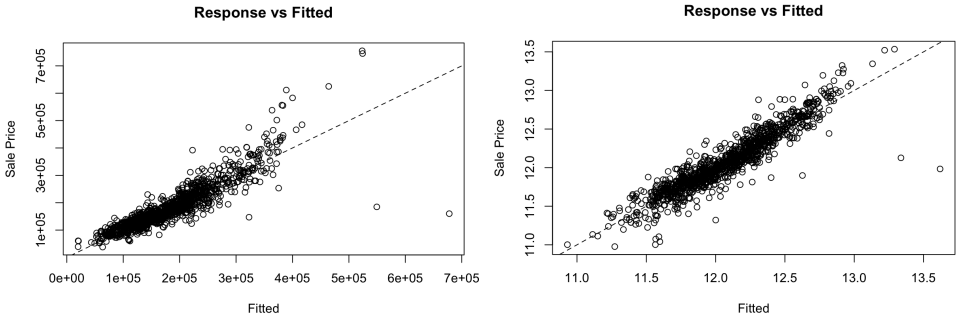


Fig 2. Residual vs Fitted for Response (Prior to transformation), QQ plot (Prior to transformation)

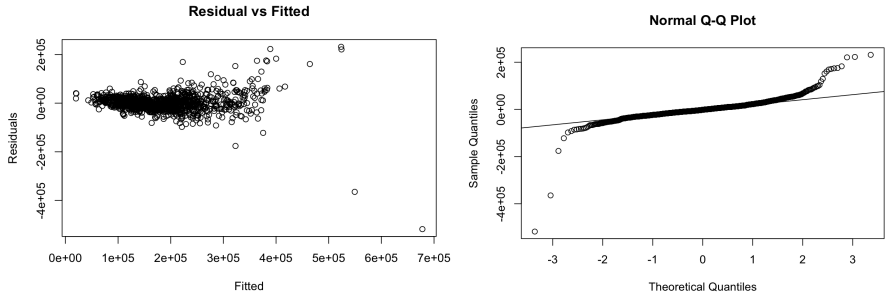


Fig 3. Residual vs Fitted for numerical variables (Prior to transformation)

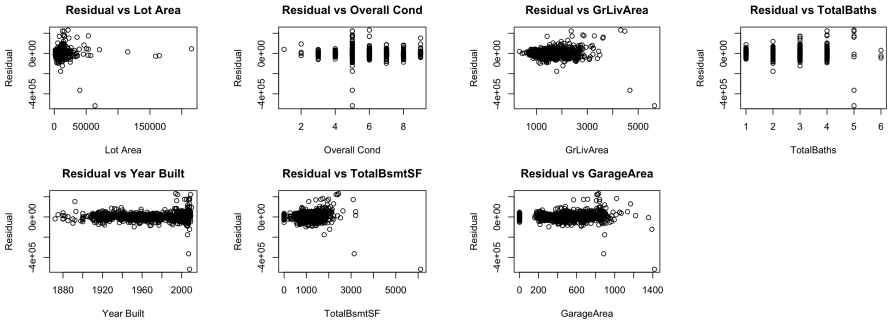
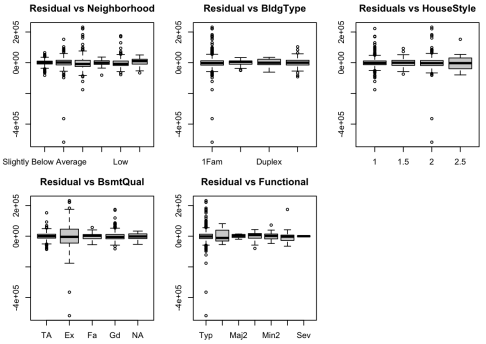


Fig 4. Box plots for categorical residual vs fitted



Research Findings (Final Model and MLR Summary)

$$\ln(SalePrice) = \beta_0 + \beta_1 \ln(LotArea) + \beta_2 I(Neighborhood\_ExtermelyHigh) + \beta_3 I(Neighborhood\_High) + \beta_4 I(Neighborhood\_SlightlyAbove) + \beta_5 I(Neighborhood\_Low) + \beta_6 I(Neighborhood\_ExtermelyLow) + \beta_7 I(BldgType\_2Fm) + \beta_8 I(BldgType\_Duplex) + \beta_9 I(BldgType\_Twnhs) + \beta_{10} I(HouseStyle\_1.5) + \beta_{11} I(HouseStyle\_2) + \beta_{12} I(HouseStyle\_2.5) + \beta_{13} OverallCond + \beta_{14} YearBuilt + \beta_{15} I(BsmtQual\_Ex) + \beta_{16} I(BsmtQual\_Fa) + \beta_{17} I(BsmtQual\_Gd) + \beta_{18} \ln(TotalBsmtSF) + \beta_{17} \ln(GrLivArea) + \beta_{18} I(Functional\_Maj1) + \beta_{19} I(Functional\_Maj2) + \beta_{20} I(Functional\_Min1) + \beta_{21} I(Functional\_Min2) + \beta_{22} I(Functional\_Mod) + \beta_{23} I(Functional\_Sev) + \beta_{24} \ln(GarageArea) + \beta_{25} TotalBaths$$

Table 2: Summary of fitted model, values for coefficients, standard errors, t-statistics and p-values. Binary Dummy variables encoding was used for the categorical variables (one-hot).

Variable	Estimates	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.625751	0.646390	-0.968	0.333213	
ln(LotArea)	0.079915	0.012225	6.537	9.39e-11	***
Neighborhood_ExtremelyLow	0.058779	0.013306	4.417	1.09e-05	***
Neighborhood_High	-0.057593	0.018793	-3.065	0.002230	**
BldgType_2fmCon	-0.100814	0. 034153	-2.952	0.003223	**
BldgTypeDuplex	-0.275128	0. 034286	-8.025	2.49e-15	***
HouseStyle_1.5	-0.053271	0.018965	-2.809	0.005056	**
HouseStyle_2	-0.072498	0.019832	-3.656	0.000268	***
OverallCond	0.067265	0.004475	15.030	< 2e-16	***
YearBuilt	0.002947	0.000312	9.444	< 2e-16	***
BsmtQual_Ex	0.220161	0.022074	9.974	< 2e-16	***
BsmtQual_Gd	0.058253	0.013932	4.181	3.12e-05	***
ln(TotalBsmtSF)	0.138873	0.025381	5.472	5.47e-08	***
ln(GrLivArea)	0.569201	0.030554	18.629	< 2e-16	***
FunctionalMaj1	-0.141052	0.047703	-2.957	0.003171	**
FunctionalMaj2	-0.219219	0.075039	-2.921	0.003552	**
FunctionalMin1	-0.096973	0. 029399	-3.299	0.001002	**
FunctionalMin2	-0.124131	0.030287	-4.099	4.45-05	***
FunctionalMod	-0.139819	0.044857	-3.117	0.001872	**
FunctionalSev	-0.575418	0.145871	-3.945	8.47e-05	***
ln(GarageArea)	0. 090222	0.015061	5.991	2.79-09	***
TotalBaths	0.035779	0. 007399	4.835	1.51e-06	***
HouseStyle_2.5	-0.061923	0.045376	-1.365	0.172627	

BsmtQual_Fa	-0.012778	0.028369	-0.450	0.652483	
BldgTypeTwnhs	0.020535	0.019180	1.071	0.284558	
Neighborhood_Low	0.021994	0.014129	1.557	0. 119828	
Neighborhood_SlightlyAbove Average	-0.053067	0. 027643	-1.920	0.055138	
Neighborhood_ExtremelyHigh	-0.000961	0.015543	-0.062	0.950709	

Discussion

Final Model Interpretation and Importance

Neighborhood_ExtremelyLow, *Neighborhood_Low*, *BldgType_Twnhs, HouseStyle_2.5* and *BsmtQual_FA* were unable to reject the null hypothesis. As expected, variables such as *Functional* (with any reported defects), *Neighborhood_High* (crime rate) and *BldgType* (Duplex or 2Family) had significant negative correlation with the *SalePrice*. Other variables either had a positive correlation as expected or a slight negative correlation.

As a reference point for categorical variables, we used a 1 story, 1 Family house in a Slightly Below Average Neighborhood in crime rates, with a typical quality basement, and typical (normal) functionality. An increase in the families living in the house (in *BldgType*), the neighbourhood's crime rate and the house's defects all contribute negatively to the sale price. Our model is able to achieve a high adjusted R-squared value of 0.8561 and residual standard error of 0.1449, indicating the model is able to predict the data with high accuracy.

Our conclusion is that the transformed ground living area in square feet is the most statistically significant factor that positively impacts a house's sale price among the variables. A larger and better quality basement (*BsmtQual_Ex, TotalBsmtSF)* also had a significant correlation. This predictive model agreed with Jafari's conclusion that the square footage of the unit is most significant (Jafari&Akhavian, 2019). This is an obvious conclusion, as it is generally accepted that a larger house equates to a bigger sale price. Moreover, Jafari also concluded that the location of the house was significant, which we also agreed, evident by negative correlation of the neighbourhood variable in relationship to higher crime rates.

Limitations

The collinearity of the model had VIF values around 5, indicating that there are some collinearities within our model. However, this is to be expected because housing characteristics are interrelated, such as LotArea and GrLivArea.

To identify problematic observations, we used techniques such as finding leverage, outliers and influential data points. For leverage, we had a cutoff of $2\frac{13}{1181}$, for outliers, $[-\ 4, 4]$, and for influential points, a 50th percentile of F(13, 1168) for D_i and $2\sqrt{\frac{13}{1181}}$ for $DFFITS_i$. Contextually, we deemed removing these observations unnecessary as they were crucial for the completeness of the dataset and they only had a minor but non-significant influence in our regression model. Our study only sampled from one city (Ames, Iowa) and might not be completely accurate for other US cities. However, national-wide data sets generally lack more detailed characteristics of houses. Furthermore, our model can be more accurate in regards to predicting based on location with more information about the neighbourhood, such as proximity hot spots and regional characteristics (surrounding environment conditions).

Citations:

1. Librarysearch.library.utoronto.ca. (n.d.). https://librarysearch.library.utoronto.ca/discovery/fulldisplay?docid=alma991106181058206196&context=L&vid=01UTORONTO_INST%3AUTORONTO&lang=en&search_scope=UTL_AND_CI&adaptor=Local+Search+Engine&tab=Everything&query=any%2Ccontains%2Cwhy%2CAND&mode=advanced&pfilter=lang%2Cexact%2Ceng%2CAND&offset=0

   Ryan-Collins, J. (2019). *Why can't you afford a home?* Polity Press.

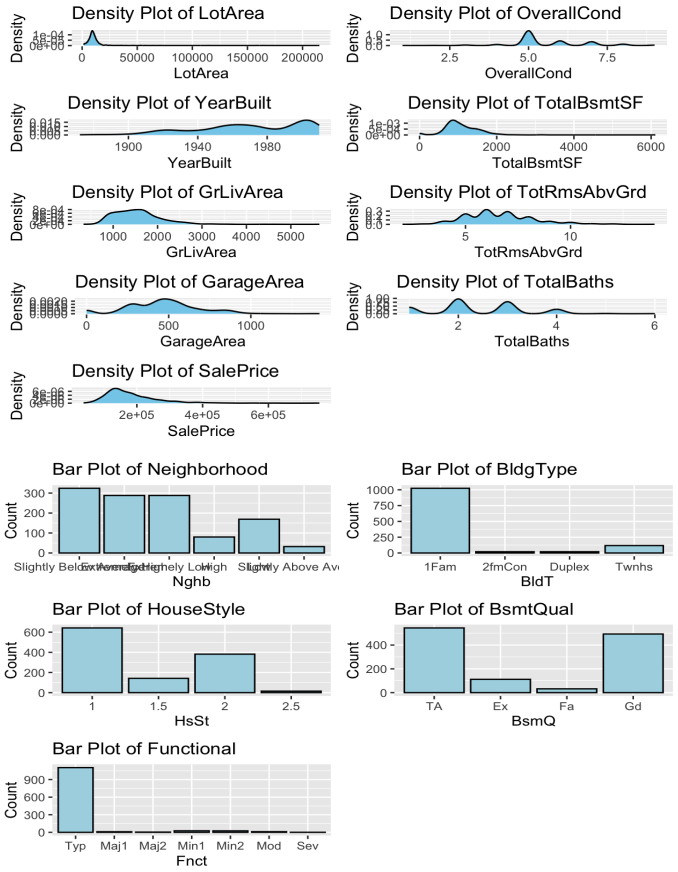
2. *House prices - advanced regression techniques*. Kaggle. (n.d.). https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data
3. Areavibes. (n.d.). Ames, IA area guide. https://www.areavibes.com/ames-ia/
4. Khatiwada, U. (2024, March 19). *Exploring Toronto's rental : Kijiji Data Analysis*. Kaggle. https://www.kaggle.com/datasets/umeshkhatiwada/toronto-kijiji-rental-data/code
5. Cigdem‐Bayram, M., & Prentice, D. (2019). How Do Crime Rates Affect Property Prices? *The Economic Record*, *95*(S1), 30–38. https://doi.org/10.1111/1475-4932.12455
6. Peng, T.-C., & Chen, C.-F. (2016). The effect of quality determinants on house prices of eight capital cities in Australia: A dynamic panel analysis. *International Journal of Housing Markets and Analysis*, *9*(3), 355–375. https://doi.org/10.1108/IJHMA-06-2015-0028
7. *Federal Reserve Bank of St. Louis*. Saint Louis Fed Eagle. (n.d.). https://www.stlouisfed.org/
8. Dettling, L. J., & Kearney, M. S. (2011). *House Prices and Birth Rates: The Impact of the Real Estate Market on the Decision to Have a Baby*. National Bureau of Economic Research.
9. Day, C. (2018). Australia's Growth in Households and House Prices. *Australian Economic Review*, *51*(4), 502–511. https://doi.org/10.1111/1467-8462.12280
10. Ebru, Ç., & Eban, A. (2011). Determinants of house prices in Istanbul: a quantile regression approach. *Quality & Quantity*, *45*(2), 305–317. https://doi.org/10.1007/s11135-009-9296-x
11. Jafari, A., & Akhavian, R. (2019, June 18). *Driving forces for the US residential housing price: A predictive analysis*. Built Environment Project and Asset Management. https://www.emerald.com/insight/content/doi/10.1108/BEPAM-07-2018-0100/full/html

Appendix

Appendix Table 1: Information about categorical and numerical variables

Variables	
MSZoning (removed in final model)	Represented by high density (condos), medium density (apartments), and low density zones
LotArea	Total Lot Size in Square Feet
Neighbourhood	The neighbourhood variable corresponds to the crime rate of each. It ranges from extremely high (highest crime rate in Ames) to extremely low (lowest crime rate in Ames).
BldgType	Type of dwelling. 1Fam corresponds to a single family detached house. 2FamCon corresponds to two family conversion, originally a single family house. Townhouse and Duplex are self-explanatory.
HouseStyle	Number of stories
OverallCond	Rates the overall condition of the house provided by the dataset from 1 to 10 (1=very poor, 10=excellent)
YearBuilt	Original finished construction date
ExterCond (removed in final model)	Present condition of the material on the exterior from Po (poor) to Ex (excellent)
TotalBsmtSF	Total Square Feet of Basement Area. Note that all houses in our dataset had a basement.
GrLivArea	Above ground living area in square feet
TotRmsAbvGrd (removed in final model)	Number of rooms (not bath) above ground
Functional	Home Functionality (assume typical unless deductions). Syntax: Min1 = one minor deduction. Maj2 = two major deductions. Mod = moderate deductions. Sev = severely damaged
GarageArea	Total square feet of garage area
TotalBaths	Number of bathrooms
SalePrice	Sale price of the house in US dollars

Appendix Figure 2: Density plot for numerical variables. Histogram for categorical variables



Appendix Figure 3: Pairwise scatterplot