Analysis of Ames Housing Dataset

Joseph, Supriya, Mei Lian, Eng Soon

Problem Statement

We are a Data Science team at Propnex pte ltd, a real estate company. We have been with tasked with analysing the Ames Housing Dataset and finding the key factors that influence the Sale Price of a house in Ames city.

Through this analysis, we will develop a regression model to predict the sale prices. We will make recommendations to homeowners on improving their property value and set expectations for their property's Sale Price.

Ames Housing Dataset

Individual residential properties sold in Ames, Iowa from 2006 to 2010.

Train Data Set	2,051 rows, 81 columns
Test Data Set	878 rows, 80 columns

Categorical	Nominal Data	23	MS_SubClass, Garage Type
	Ordinal Data	23	Overall Quality, Overall Condition
Numerical	Discrete Data	20	Year Sold, Bedroom
	Continuous Data	15	Ground Living Area, Lot Area

Data cleaning

Handling missing data

- Check for actual values of NaN values with na_filter=false
- Most NaN values are actually NA (Not applicable)

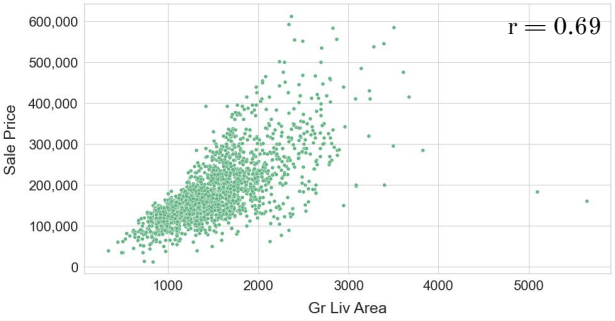
How we imputed the data for some columns

- If values are missing in dataframe, we would try to impute it by enquiring:-

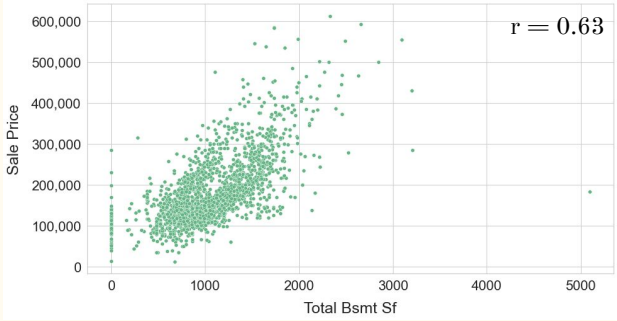
a. Is the amount of data missing significant? (eg. Lot Frontage)
b. What other features would have highest correlation with missing feature
c. If we are unable to identify (b), we can impute with values (mean/median/mode) from the feature with the missing values.
d. If not, we would most probably fill it as NA

Exploratory Data Analysis (Numerical Variables)

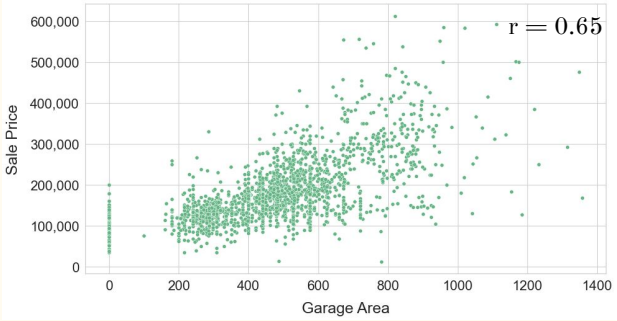
Ground living Area vs Sale Price



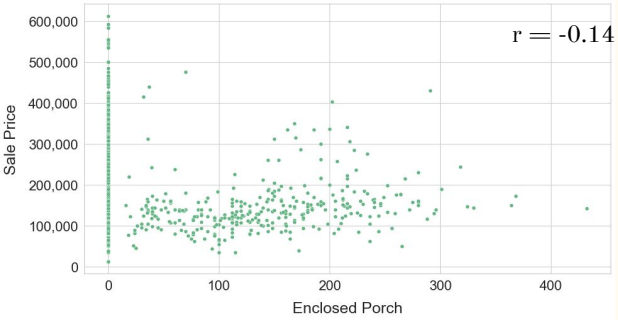
Total Basement Sqft vs Sale Price



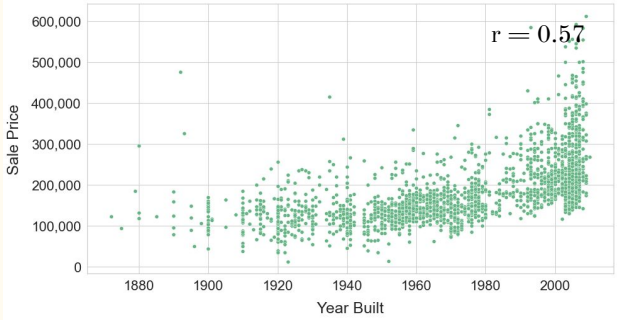
Garage Area vs Sale Price



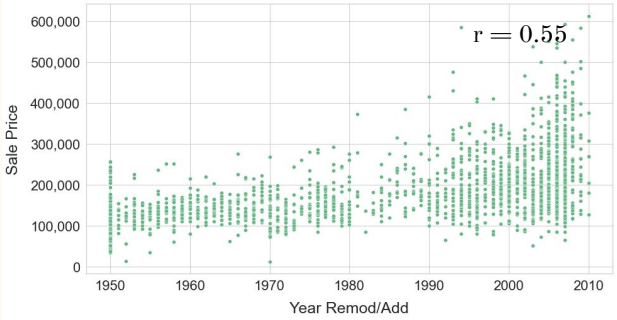
Open Porch Sqft vs Sale Price



Year Built vs Sale Price

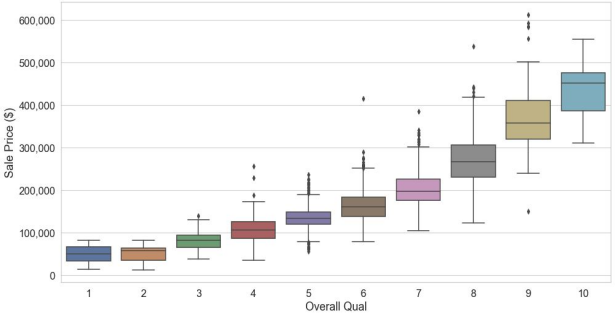


Remodel Year vs Sale Price

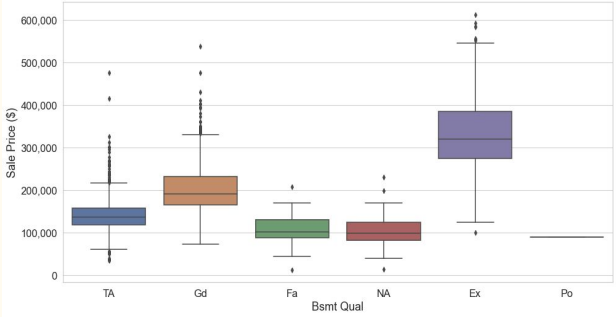


Exploratory Data Analysis (Categorical Variables)

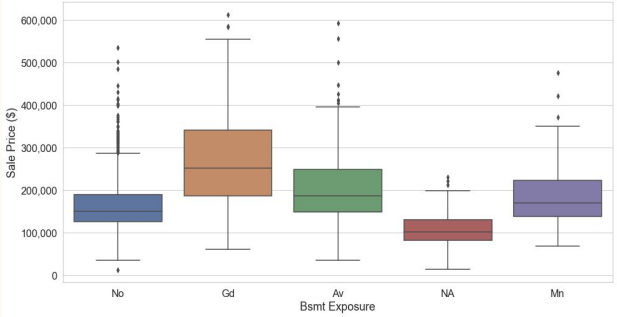
Overall Quality vs Sale Price



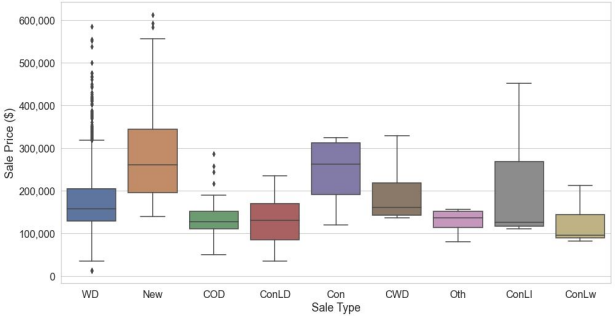
Basement Quality vs Sale Price



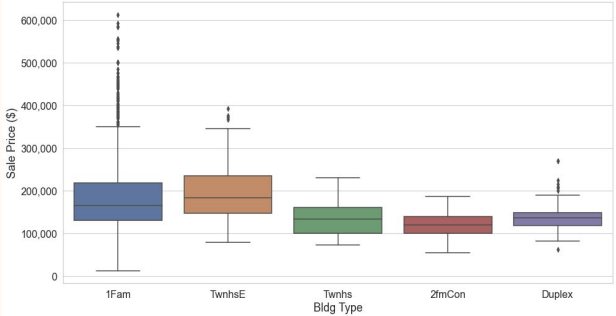
Basement Exposure vs Sale Price



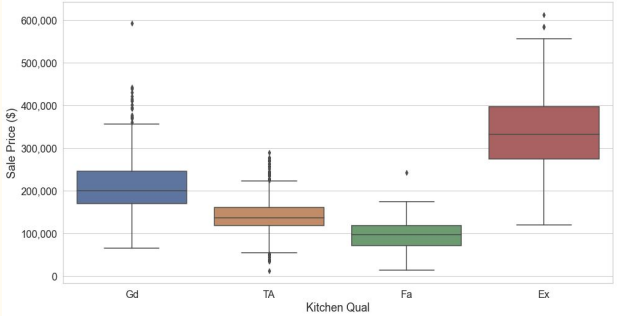
Sale Type vs Sale Price



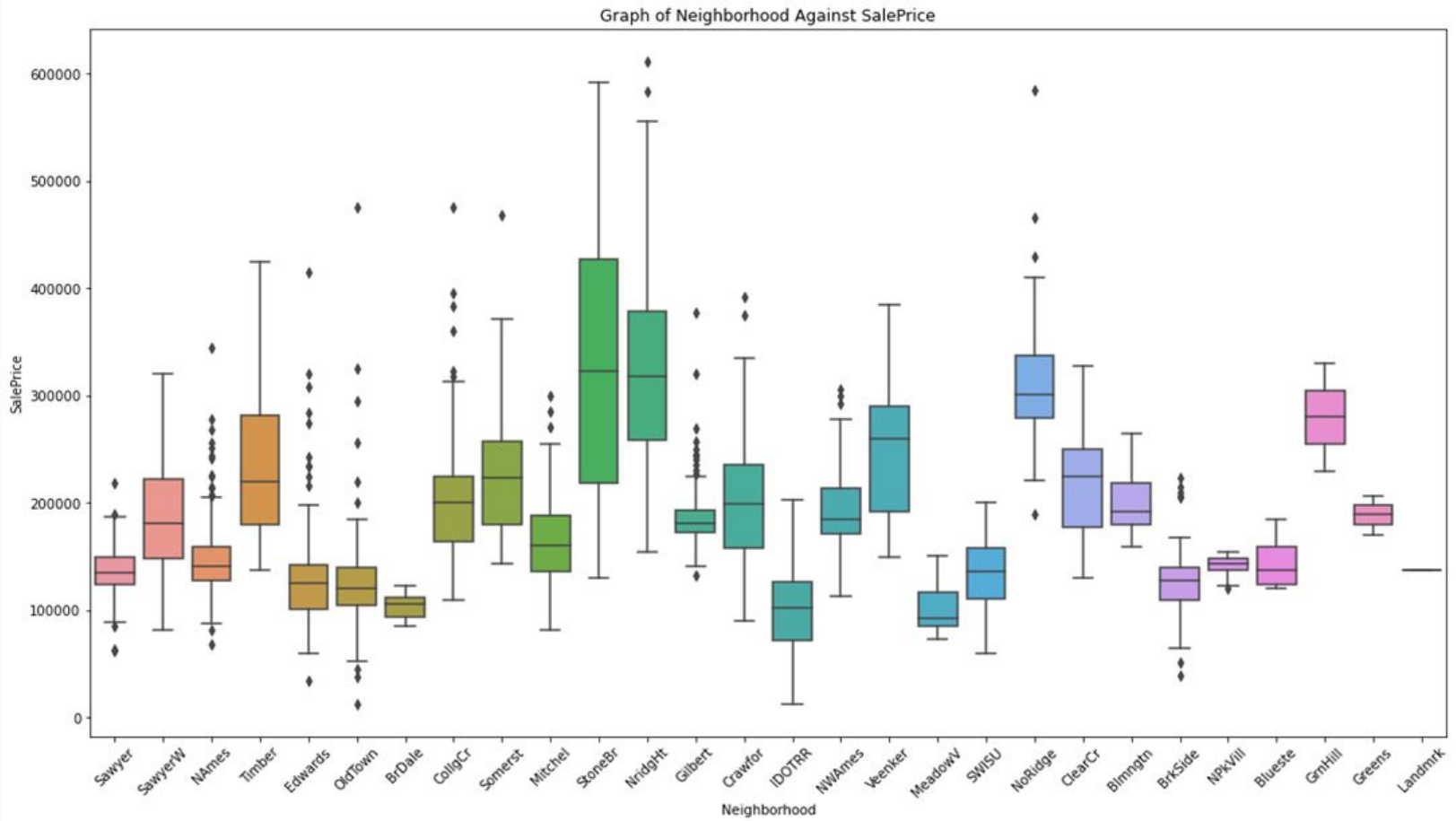
Building Type vs Sale Price



Kitchen Quality vs Sale Price

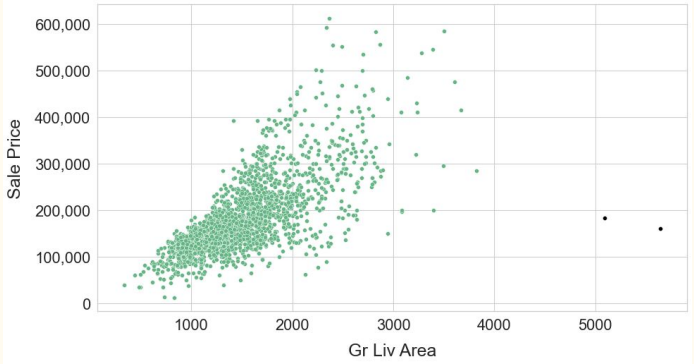


Exploratory Data Analysis (Categorical Variables)

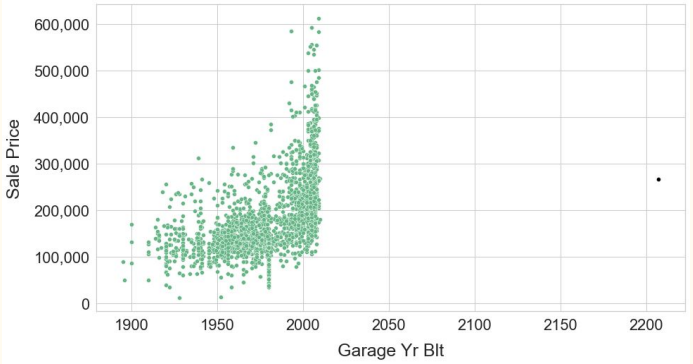


Handling Outliers

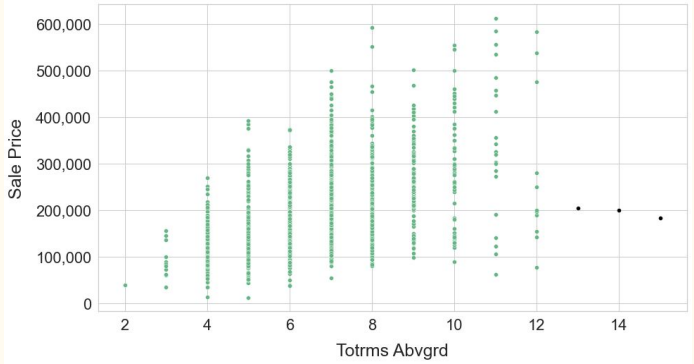
Ground living Area vs Sale Price



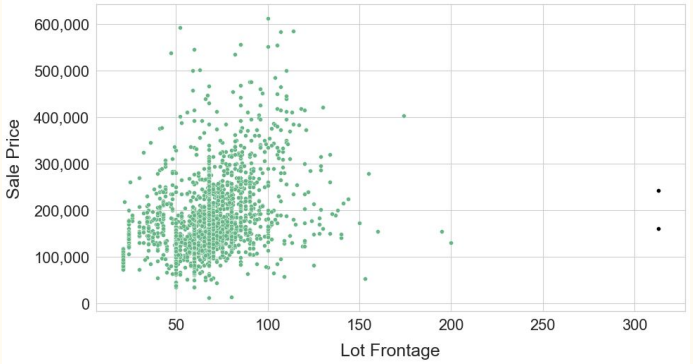
Garage Year Built vs Sale Price



Total Rooms Above ground vs Sale Price



Lot Frontage vs Sale Price



Pre-Processing the Data

- Handling Ordinal Variables
  - Eg: External quality, Basement Condition
    - Before: Excellent, Good, Average/Typical, Fair, Poor
    - After: 5, 4, 3, 2, 1
- One Hot Encoding
  - Eg: Heating, Neighborhood
    - Before: Heating (with sub-categories)
    - After: Heating_Floor, Heating_GasA, Heating_GasW, Heating_Grav, Heating_OthW, Heating_Wall
- Standardisation

Modelling - Lasso regression

R^2 score
Train data: 0.93
Validation data: 0.91

Cross val score: 0.91

Feature Engineering

- Ground Living Area * Overall Quality
- Basement Quality * Basement Exposure
- Garage Cars * Garage Area
- Overall Condition * Remodel date
- Lot Frontage * Lot Area
- Overall Quality * Functionality

Modelling - Lasso & Ridge regression

Lasso

R^2 score
Train data: 0.92
Validation data: 0.90

Cross val score: 0.91

Ridge

R^2 score
Train data: 0.94
Validation data: 0.93

Cross val score: 0.93

Model Coefficients

Variable	Coefficient	Absolute (Coefficient)
gr_liv_area	20,314.41	20,314.41
overall_qual	11,598.11	11,598.11
gr_liv_area*overall_qual	8,678.98	8,678.98
bsmtfin_sf_1	8,284.08	8,284.08
total_bsmt_sf	5,869.06	5,869.06
lot_area	5,822.88	5,822.88
year_built	5,764.04	5,764.04
exter_qual	5,150.10	5,150.10
bsmt_qual*bsmt_exposure	4,986.68	4,986.68
overall_cond	4,557.10	4,557.10
1st_flr_sf	4,293.77	4,293.77
neighborhood_StoneBr	4,282.09	4,282.09
neighborhood_NridgHt	4,087.44	4,087.44
kitchen_qual	4,030.58	4,030.58
mas_vnr_area	3,630.57	3,630.57
year_remod/add	3,445.92	3,445.92
screen_porch	3,385.24	3,385.24
functional	3,356.80	3,356.80
neighborhood_GrnHill	3,275.59	3,275.59
exterior_1st_BrkFace	3,236.05	3,236.05
sale_type_New	3,178.57	3,178.57
roof_style_Mansard	(3,039.32)	3,039.32
garage_cars*garage_area	3,000.79	3,000.79
fireplaces	2,855.81	2,855.81
lot_frontage	2,733.63	2,733.63
garage_type_NA	(2,639.20)	2,639.20
bsmt_qual	2,457.01	2,457.01
ms_zoning_FV	2,380.66	2,380.66
neighborhood_Crawfor	2,291.73	2,291.73
lot_config_CulDSac	2,262.15	2,262.15

Positive Factors

- Ground living area
- Overall quality
- Basement Type 1 finished square feet
- Total square feet of basement area
- Lot area
- Year built
- External quality
- Basement quality with good exposure

Negative Factors

- Roof type - Mansard
- No garage
- Neighborhood College Creek

Conclusions

- Bigger houses fetch better Sale Price
  - High Ground living area
  - High Basement Area
  - High Lot Area
- Better quality houses fetch better Sale Price
  - Overall quality
  - External quality
- Good Basement quality & exposure enhances the Sale Price
- Newer houses positively influence the Sale Price
- Not having a garage lowers the Sale Price

Recommendations to Home Owners

- Home owners can consider undertaking renovation and improving the overall quality of the house by using material of better quality and finish.
  - Avoid roof type Mansard
  - Add a fireplace
  - Improve kitchen quality
  - Exterior covering of the house BrickFace
  - Add a screen porch
- They should focus on specifically improving the external quality using better quality material on the exterior.
- Home owners may set their expectations about their property value according to:
  - Size of their Ground Living Area, Basement, Lot Area
  - Year in which the house was built
  - Basement quality & exposure
  - Neighbourhood (Stone Brook, Northridge Heights, Green Hills have better Sale Price)

Thank You

Q&A

Mansard Roof and disadvantages

1. Costs and installation difficulties. Usually take much longer to install as require a greater degree of labor and more materials.

2. Not very weather resistant. The second, flatter slope on the upper section of the roof does not allow accumulating moisture to drain as quickly as it does on the lower sections.

Homeowners may need to deal with ponding water on the roof. If left alone for too long, these sections can quickly cause leaks and develop mold inside and outside the roofing system, while weighing down on the roof's structure itself.

3. Greater degree of maintenance when it comes to managing this roofing system, especially in stormier areas. Debris can collect at the flatter portion of the roof surface, causing long term problems throughout the unit itself.

Disadvantage of staying in Neighborhood College Creek

https://www.iowastatedaily.com/news/college-creek-concerns-ames-residents/article_fcf2a4ae-2698-5117-a149-09ae533f36dd.html

College Creek concerns Ames residents

By Katie Robb (Iowa State Daily)
Mar 6, 2001

As warmer weather begins to melt the snow and send runoff back to the local waterways, College Creek has once again become a concern for Ames residents."We will be monitoring the creek again in the spring when the snow thaws and the creek is flowing again," said John Dunn, assistant director of the Ames Water Pollution Control Plant.Last fall, the creek showed abnormally high levels of fecal chloroform, which spread waterborne diseases."The banks of the creek in West Ames basically have sewage lining the walls. It's knee deep in some spots," said Eric Abrams, co-founder of the College Creek Action Committee.The sewage contained high levels of fecal chloroform, bacteria that originate in the intestines of mammals. According to the City's College Creek Web site, www.city.ames.ia.us/waterweb/CollegeCreek/college_creek.htm, the bacteria themselves are harmless, but their presence indicates the water has been contaminated with the fecal material of humans or other animals.Local residents were concerned about health risks and formed CCAC with the help of Abrams and Anita Mahr-Lewis.The contamination levels are highest near the start of the creek by the Bentwood Subdivision, said Bob Kindred, assistant city manager for the City of Ames.Dunn said the biggest problem is present in the first few thousand feet of the creek, but normal levels of bacteria are found once it reaches South Dakota Avenue.Abrams, however, estimates the polluted portion of the creek extends as far as two miles, to approximately the location of Ames Middle School, 321 State St.The city gave the residents the option of cleaning out the creek, but residents decided to let it degrade naturally, he said.Although the problem first came to the attention of officials a year ago, many believe it has been an issue for years."The mobile home parks have been there for years and years," Dunn said. "This has probably been a problem for some time."Two restaurants, The Broiler, 6008 Lincoln Way, and Chef's Inn, 6400 Lincoln Way, were cited as sources of the problem. Dunn said three mobile home parks, LPM Homes Country Terrace, 6100 W. Lincoln Way, Hillsdale Mobile Home Park, located on a county road south of Lincoln Way, and Crestview Mobile Home Park, 5615 W. Lincoln Way, were also responsible for the contamination.The restaurants have corrected the problem and Country Terrace was granted an extension until June 1, Dunn said. The three mobile-home parks didn't meet the deadline, but because the homes are out of city limits the counties and the Department of Natural Resources have to address the problem, Kindred said.The other two mobile homes parks will be referred to the Iowa attorney general by DNR, Dunn said. The parks need to have new sewer systems or they will be subject to fines from the state attorney general.The city also hopes to expand its water-quality assessment to other creeks, streams and lakes in the area."We're trying to get volunteers to help with ongoing water monitoring," Kindred said. "Later, this month we're going to bring back a program of proposed testing for all creeks in the area."The program would involve testing local waterways four times a year. The proposal will be presented during the March 27 city council meeting, Kindred said."Lots of people come into contact with the water in the creek," Kindred said. "This is a major problem."