KING COUNTY HOUSE SALES PREDICTION

20th April 2023

INTRODUCTION

Real estate developers are interested in identifying factors that influence the sale price of homes in King County, as well as using models to predict the sale price of homes based on these factors.

The information about these factors can be used to optimize the design and marketing of new properties, identify investment opportunities, and make data-driven decisions about the development and sale of properties.

PROJECT OVERVIEW

This project seeks to provide insights into the available data on house sales in the area and provide recommendations to a real estate developer.
The project will leverage a wide range of data points including the size of the houses, their quality, their age, and additional features such as the views and whether they have a waterfront, in order to establish the relationships between these features and the sale price of those houses.

THE REAL ESTATE BUSINESS

PROBLEM QUESTIONS

The project seeks to answer the following questions:

- ❖ Which house features have the highest influence on the price?
- ❖ How does the size of the property influence the sale price of homes in King County?
- ❖ How does the house neighborhood affect the prices?
- ❖ How accurately can we predict the sale price of homes in King County based on the available features?

THE DATA

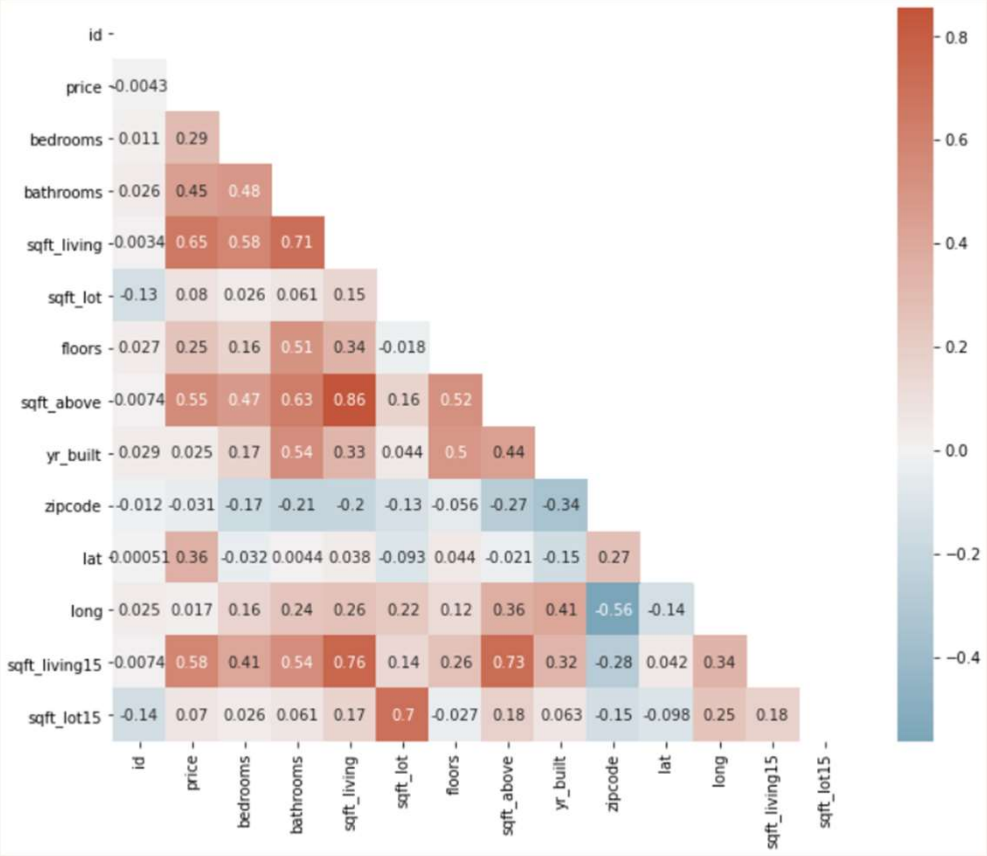
DATASETS USED

This project uses the King County House Sales dataset. It includes information about:

1. The **size of the homes** which is described by the area in square feet of: The lot, the living area, the basement, the area above ground and the number of bedrooms and bathrooms.
2. The **quality of the homes** which is described by the features: condition, grade and the year built.
3. The **neighborhood** around which the property stands which is described by: the zipcode, the latitude and longitude co-ordinates and the size of 15 properties around it.
4. **Additional features** such as: the view and Whether the house is on a waterfront

DATA ANALYSIS

A correlation matrix



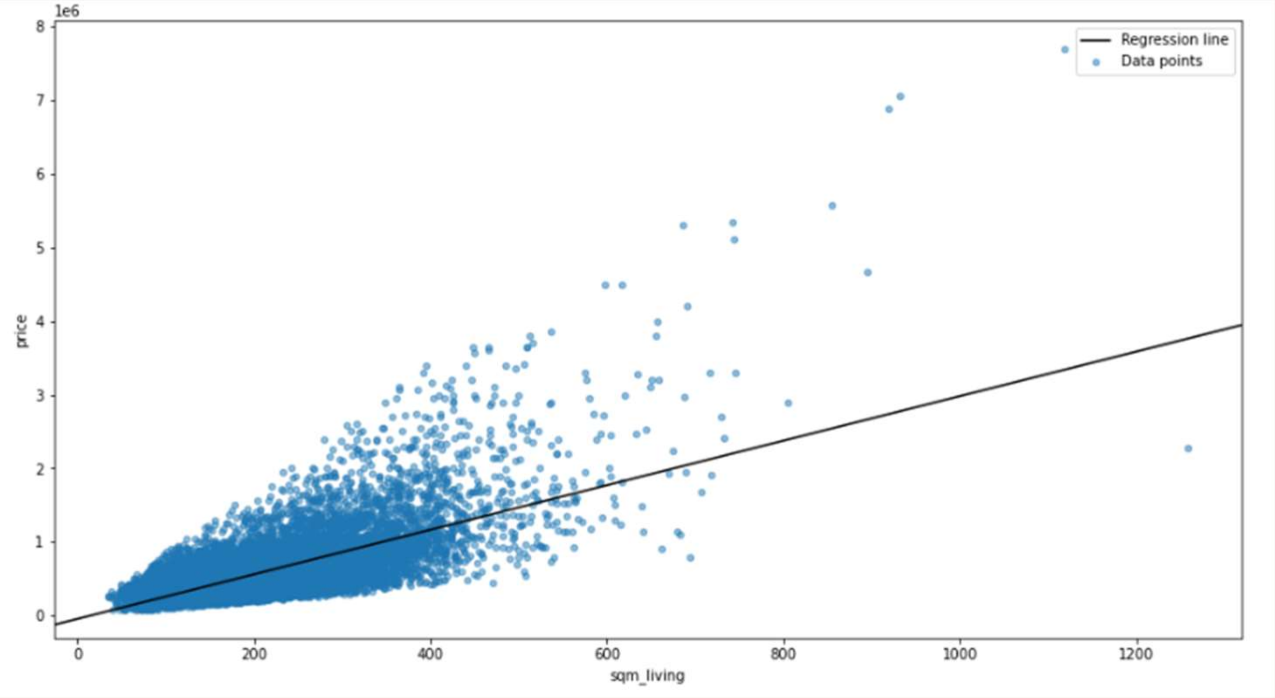
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	yr_built	zipcode	lat	long	sqft_living15	sqft_lot15
id														
price	-0.0043													
bedrooms	0.011	0.29												
bathrooms	0.026	0.45	0.48											
sqft_living	0.0034	0.65	0.58	0.71										
sqft_lot	-0.13	0.08	0.026	0.061	0.15									
floors	0.027	0.25	0.16	0.51	0.34	-0.018								
sqft_above	0.0074	0.55	0.47	0.63	0.86	0.16	0.52							
yr_built	0.029	0.025	0.17	0.54	0.33	0.044	0.5	0.44						
zipcode	-0.012	-0.031	-0.17	-0.21	-0.2	-0.13	-0.056	-0.27	-0.34					
lat	0.00051	0.36	-0.032	0.0044	0.038	-0.093	0.044	-0.021	-0.15	0.27				
long	0.025	0.017	0.16	0.24	0.26	0.22	0.12	0.36	0.41	-0.56	-0.14			
sqft_living15	0.0074	0.58	0.41	0.54	0.76	0.14	0.26	0.73	0.32	-0.28	0.042	0.34		
sqft_lot15	-0.14	0.07	0.026	0.061	0.17	0.7	-0.027	0.18	0.063	-0.15	-0.098	0.25	0.18	

Analysis of the correlation matrix:

- Before modeling, a correlation matrix was built.
- Correlation is a measure of the linear relationship between two variables.
- The higher the value, the stronger the relationship.
- The most strongly correlated feature with the target column price is the square footage of the living space with a correlation of 0.65.
- This suggests that houses with a larger living area are likely to have higher prices than those with smaller living spaces.

MODELING

1. Baseline Model

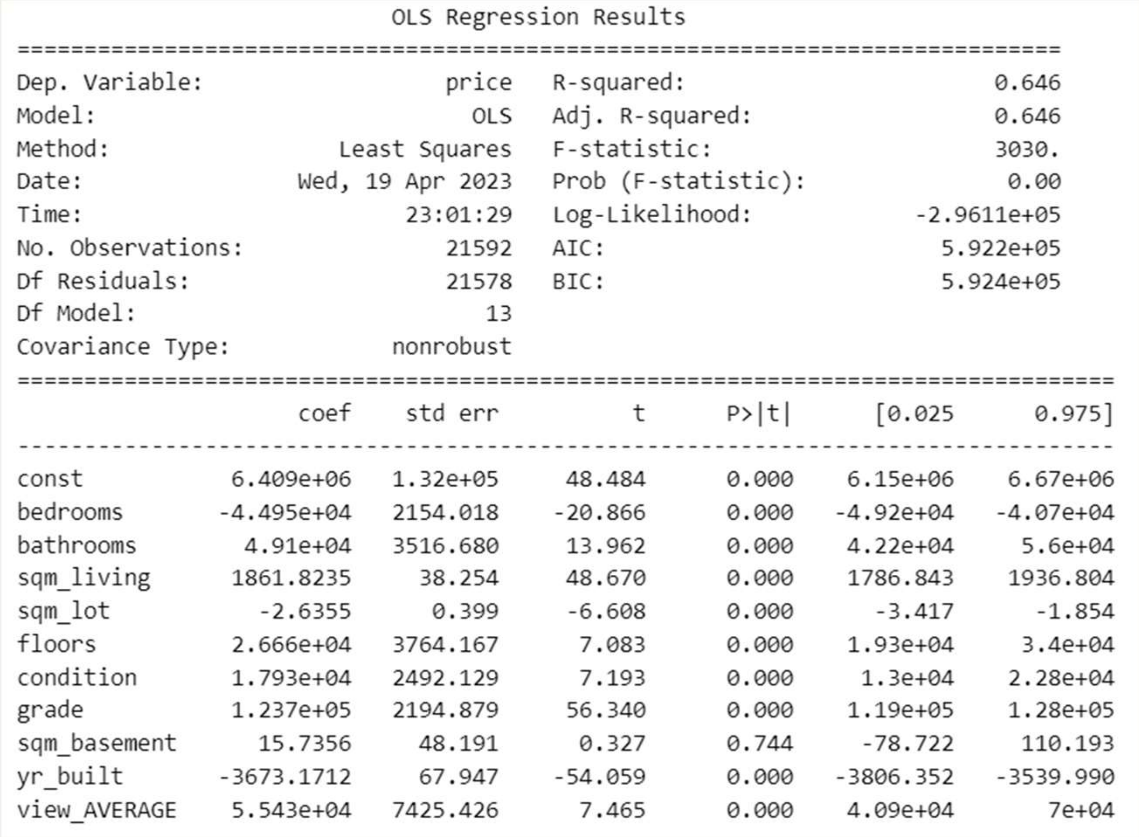


Analysis of the model:

- The first approach we took to modelling was building a baseline model.
- The model is off by about $174,000.
- The model was able to capture only 49.2% of the variance in price.
- To improve this, a multiple linear regression model was built.

MODELING

2. ITERATED MODEL



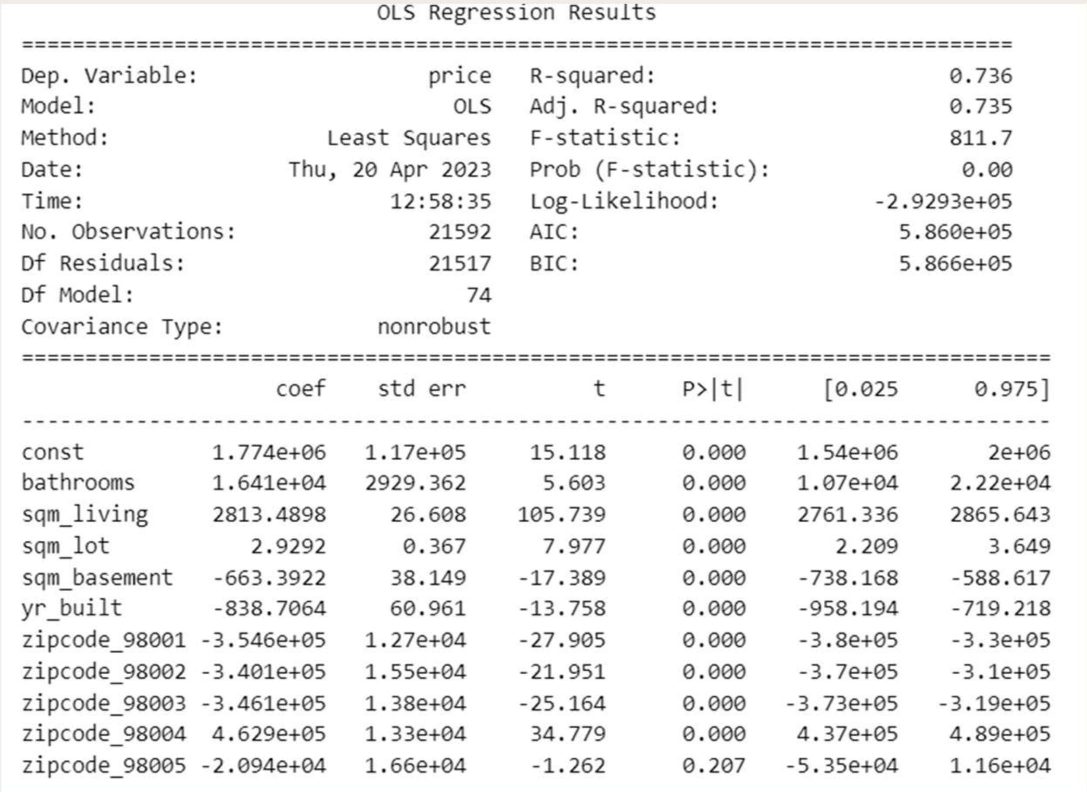
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.646
Model:                            OLS   Adj. R-squared:                  0.646
Method:                 Least Squares   F-statistic:                     3030.
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:01:29   Log-Likelihood:            -2.9611e+05
No. Observations:               21592   AIC:                         5.922e+05
Df Residuals:                   21578   BIC:                         5.924e+05
Df Model:                          13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const           6.409e+06   1.32e+05     48.484      0.000    6.15e+06    6.67e+06
bedrooms       -4.495e+04   2154.018    -20.866      0.000   -4.92e+04   -4.07e+04
bathrooms        4.91e+04   3516.680     13.962      0.000    4.22e+04     5.6e+04
sqm_living      1861.8235     38.254     48.670      0.000    1786.843    1936.804
sqm_lot           -2.6355      0.399     -6.608      0.000      -3.417      -1.854
floors          2.666e+04   3764.167      7.083      0.000    1.93e+04     3.4e+04
condition       1.793e+04   2492.129      7.193      0.000     1.3e+04    2.28e+04
grade           1.237e+05   2194.879     56.340      0.000    1.19e+05    1.28e+05
sqm_basement      15.7356     48.191      0.327      0.744     -78.722     110.193
yr_built       -3673.1712     67.947    -54.059      0.000   -3806.352   -3539.990
view_AVERAGE    5.543e+04   7425.426      7.465      0.000    4.09e+04       7e+04
```

Analysis of the model:

- The **goal of the multiple linear regression** is to analyze the relationship between the price and two or more of the other features.
- From the results obtained, the model is off by about **$140,694** in its prediction, which is an improvement to the baseline model.
- The model was able to capture **64.5%** of the variance of the price, which is an **improvement to the previous model.**
- Overall, this **model performed better than the baseline model**, however, in order to better improve the proportion of price that can be explained by the house features, another multiple linear regression was performed, where more features were added to the model.

MODELING

2. Final Model



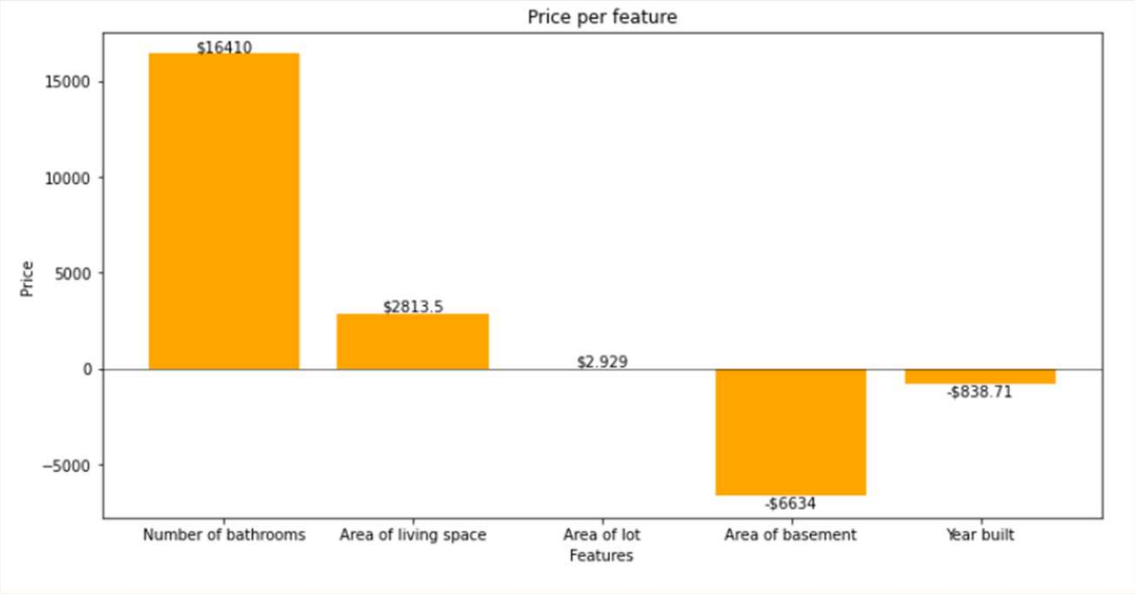
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.736
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     811.7
Date:                Thu, 20 Apr 2023   Prob (F-statistic):               0.00
Time:                        12:58:35   Log-Likelihood:            -2.9293e+05
No. Observations:               21592   AIC:                         5.860e+05
Df Residuals:                   21517   BIC:                         5.866e+05
Df Model:                          74
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          1.774e+06   1.17e+05     15.118      0.000    1.54e+06       2e+06
bathrooms      1.641e+04   2929.362      5.603      0.000    1.07e+04    2.22e+04
sqm_living     2813.4898     26.608    105.739      0.000    2761.336    2865.643
sqm_lot           2.9292      0.367      7.977      0.000       2.209       3.649
sqm_basement   -663.3922     38.149    -17.389      0.000    -738.168    -588.617
yr_built       -838.7064     60.961    -13.758      0.000    -958.194    -719.218
zipcode_98001 -3.546e+05   1.27e+04    -27.905      0.000    -3.8e+05    -3.3e+05
zipcode_98002 -3.401e+05   1.55e+04    -21.951      0.000    -3.7e+05    -3.1e+05
zipcode_98003 -3.461e+05   1.38e+04    -25.164      0.000   -3.73e+05   -3.19e+05
zipcode_98004  4.629e+05   1.33e+04     34.779      0.000    4.37e+05    4.89e+05
zipcode_98005 -2.094e+04   1.66e+04     -1.262      0.207   -5.35e+04    1.16e+04
```

Analysis of the model:

- In this model, the zipcode of the homes were included into the analysis.
- From the results obtained, the model prediction has improved and is only off by about $ 109,000 from the previous $140,694.
- The model was able to capture 73.6% of the variance in price, which is an improvement to the previous model.
- Overall, this model was the best performing, therefore, it was selected for prediction.

RESULTS

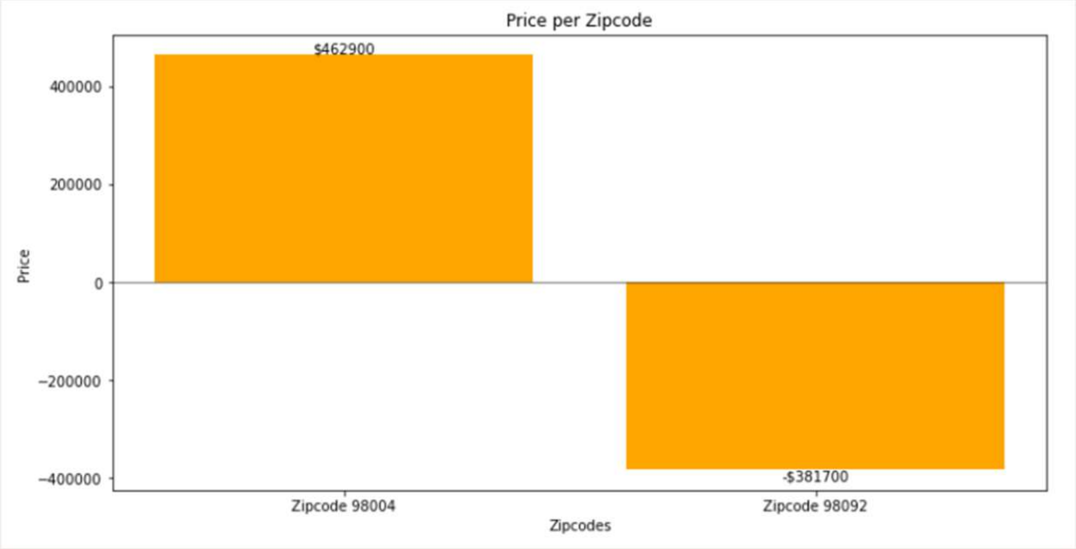
1. A plot showing the most significant and least significant features



- A one-unit increase in the number of bedrooms is associated with an increase of $ 16,410 in home price.
- An increase of one square meter of living space is associated with an increase of $ 2,813.50 in home price.
- An increase of one square meter of the lot size is associated with an increase of $ 2.9292 in home price.
- An increase of one square meter of the basement is associated with a decrease of $ 838.71 in home price.
- An increase in the year the home was built is associated with a decrease of $ 1070.3272 in home price.

RESULTS

1. A plot showing variation of price according to zipcodes, relative to zipcode 98103



- Compared to zipcode_98103, zipcode_98004 has the highest increase of $462,900 in home price.
- Compared to zipcode_98103, zipcode_98092 has the highest decrease of $381,700 in home price.

CONCLUSION

The final model was chosen as it explained about 74 % of the variance in price, about 10% more than the iterated model.

It also had a lower Mean Absolute Error, by about $ 32,000. From this model:

1. The bathroom is associated with bringing the highest increase in sale price at $16,410 for each bedroom added.
2. An increase in the area of the living space by 1 square meter had the second highest associated increase in price at $2,813.50 for each square meter added
3. Compared to zipcode_98103, zipcode_98004 has the highest increase of $462,900 in home price.

In the final model, prediction of the house prices is off by about $109,000.

RECOMMENDATIONS

When building new houses, The Real Estate Developer should prioritize:

1. Increasing the number of bathrooms.
2. The size(square meters) of the living space.
3. Building houses in the postal area of `zipcode_98004`

CHALLENGES & NEXT STEPS

- The study had drawbacks in that it had many missing values.
- A further study may be required with a larger dataset for better insights

NEXT STEPS

- To develop a dashboard to display insights for easier access of information.
- Add more features in the model that may result to sale’s profit.

Thank You.
Any questions?

1. JIMCOLLINS WAMAE
2. LEO KARIUKI
3. MAUREEN KITANGA
4. PRISCILA KAMIRI
5. SAMUEL KYALO
6. STEVE GITHINJI