KING COUNTY HOUSE SALES ANALYSIS

CONTENTS

INTRODUCTION

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

This information 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

The project seeks to optimize the design and marketing of new properties, identify investment opportunities, and make data-driven decisions about the development and sale of properties.

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?

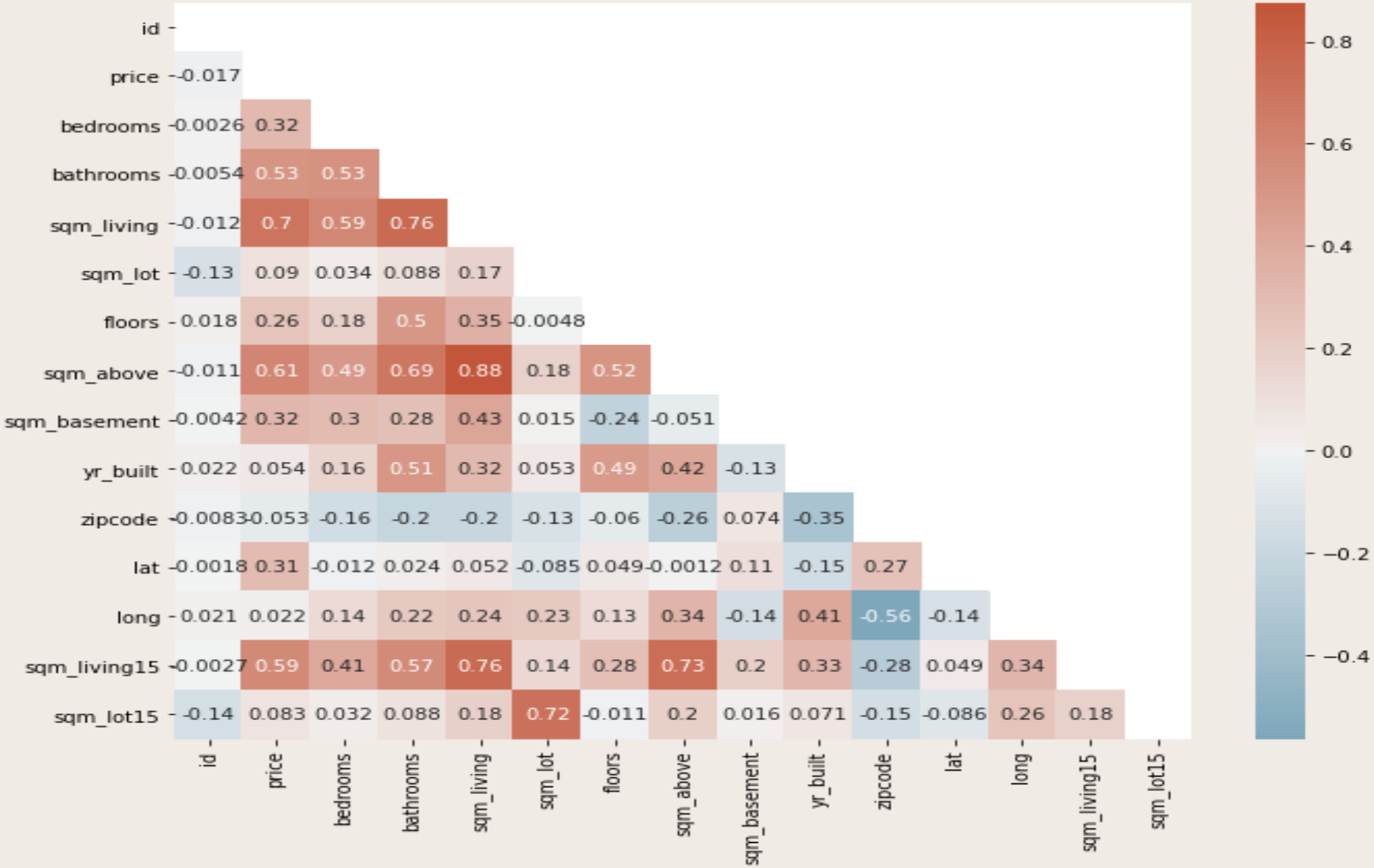
THE DATA

DATASETS USED

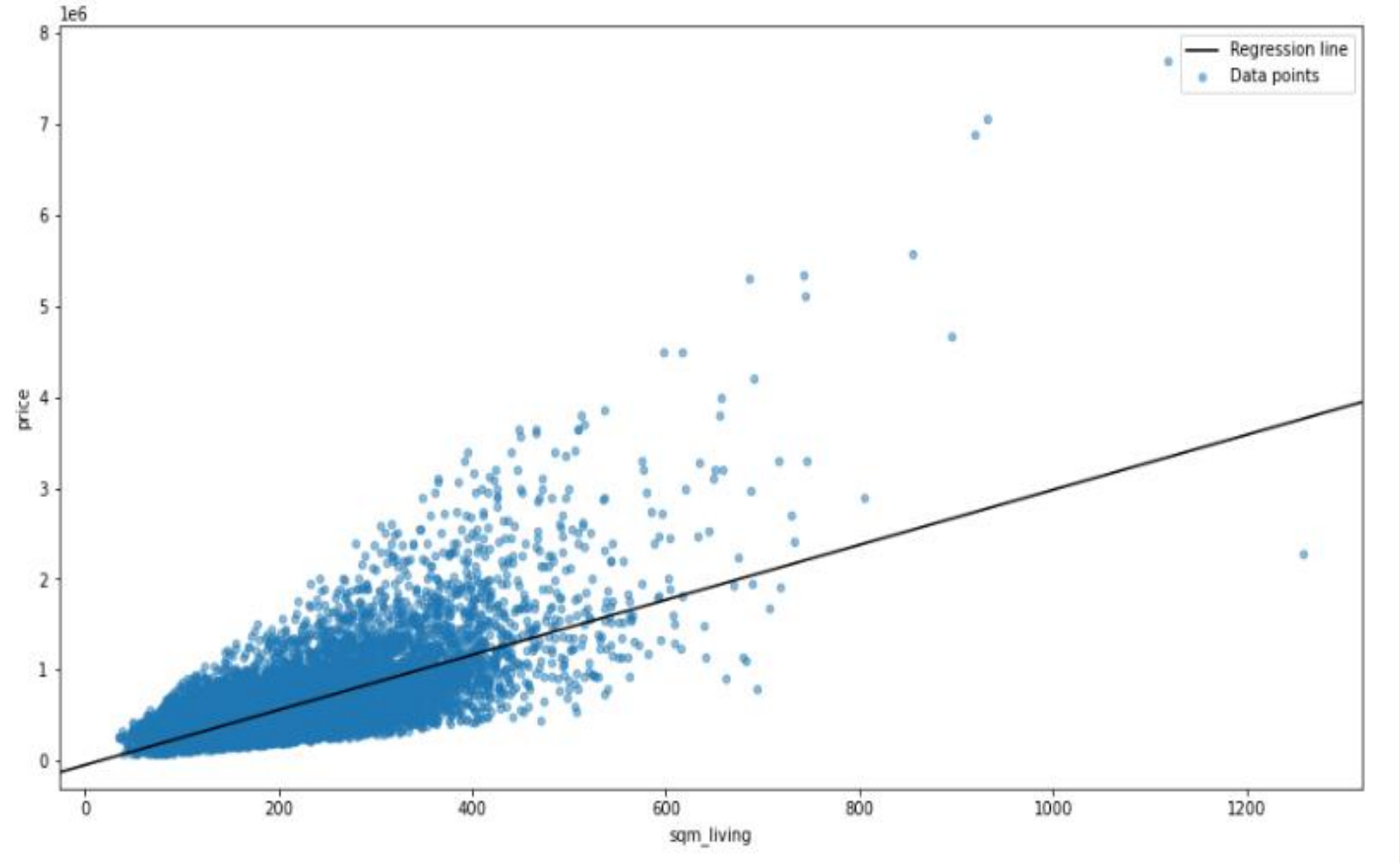
This project uses the King County House Sales dataset. It includes information about:

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 zip-code, the latitude and longitude co-ordinates and sqft_living15(the size of 15 properties around it)
4. **Additional features** such as: the view and Whether the house is on a waterfront

CORRELATION MATRIX



BASELINE MODEL



Results…

The Mean Absolute Error (MAE) for the baseline model showed that the model is off by about $173,829.544 in the prediction.

The model was able to capture only 49.2% variance in price.

To improve this, a multiple linear regression model was built.

ITERATED MODEL

	coefficient	p-value
const	6.408900e+06	0.000000e+00
bedrooms	-4.494520e+04	9.673014e-96
bathrooms	4.910154e+04	4.109894e-44
sqm_living	1.861823e+03	0.000000e+00
sqm_lot	-2.635470e+00	3.991863e-11
floors	2.666171e+04	1.453689e-12
condition	1.792535e+04	6.555596e-13
grade	1.236595e+05	0.000000e+00
sqm_basement	1.573565e+01	7.440282e-01
yr_built	-3.673171e+03	0.000000e+00
view_AVERAGE	5.542823e+04	8.668392e-14
view_EXCELLENT	4.909560e+05	0.000000e+00
view_FAIR	1.118657e+05	7.354961e-20
view_GOOD	1.269046e+05	7.327627e-36

From the results obtained, these were coefficients and 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%** variance in 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.

ITERATED MODEL 2

In this model, the Zip-code of the homes were included into the analysis.

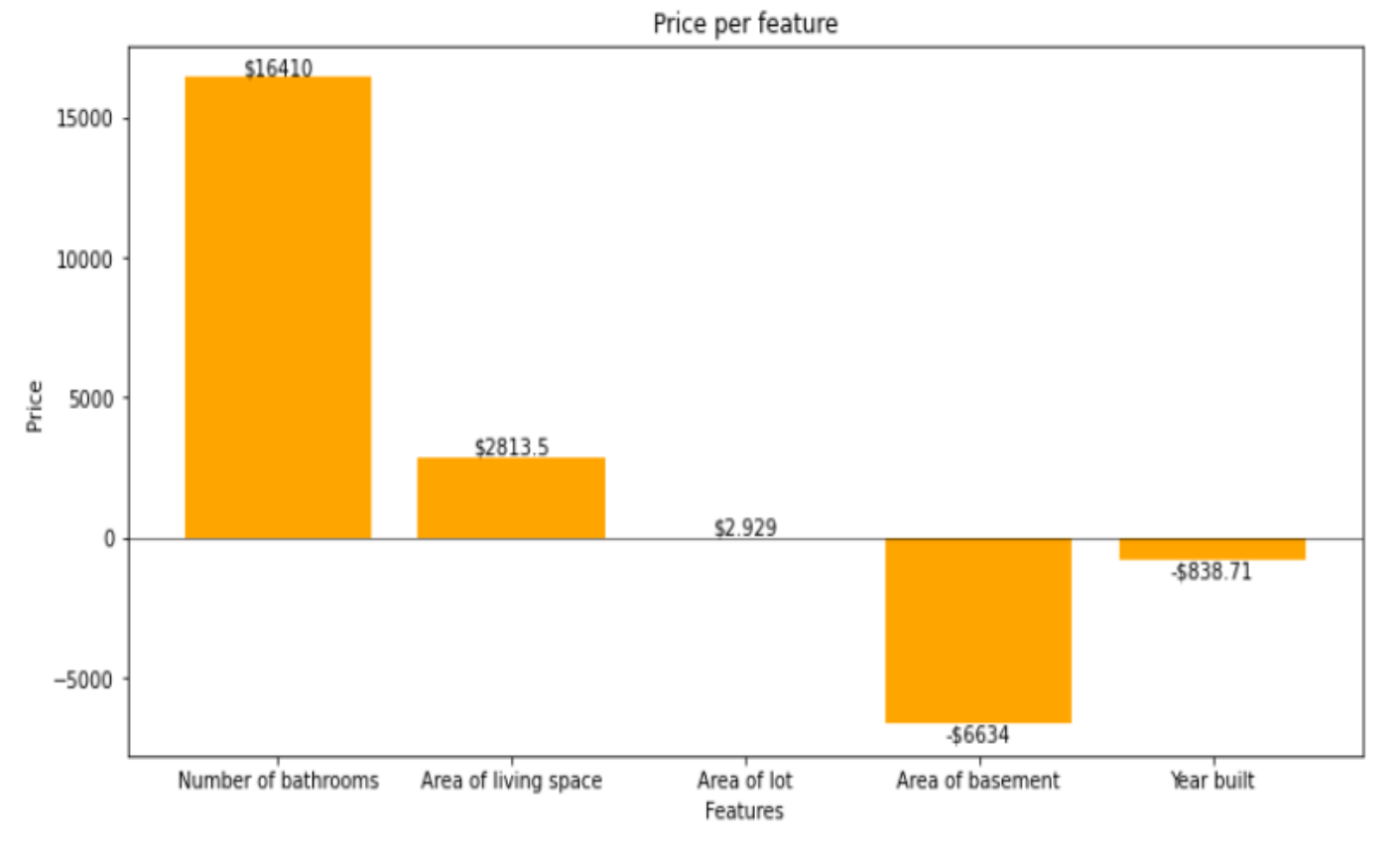
From the results obtained, the model prediction has improved and is only off by about **$ 109,000** from the previous **$140,694**.

The model was able to capture **73.6%** of variance in price which is an **improvement to the previous model.**

Overall, this **model was the best performing model,** therefore, it was selected for further prediction.

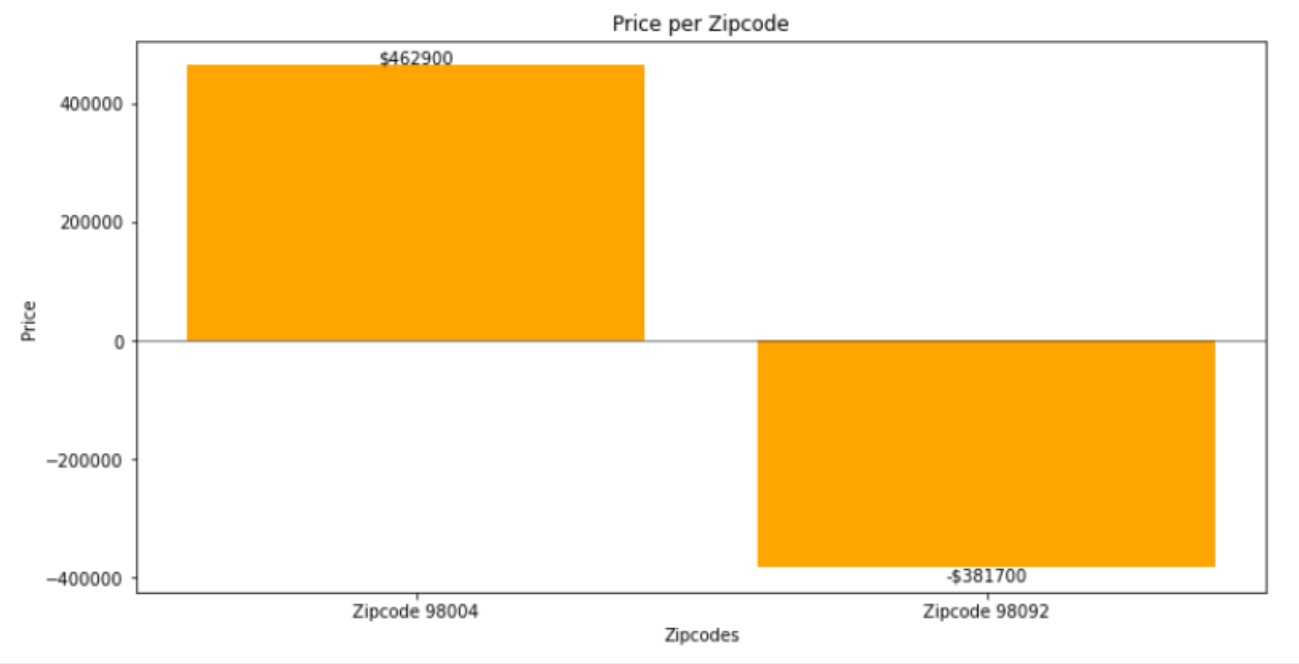
	coef	std err	t	P>|t|
const	1.774e+06	1.17e+05	15.118	0.000
bathrooms	1.641e+04	2929.362	5.603	0.000
sqm_living	2813.4898	26.608	105.739	0.000
sqm_lot	2.9292	0.367	7.977	0.000
sqm_basement	-663.3922	38.149	-17.389	0.000
yr_built	-838.7064	60.961	-13.758	0.000
zipcode_98001	-3.546e+05	1.27e+04	-27.905	0.000
zipcode_98002	-3.401e+05	1.55e+04	-21.951	0.000
zipcode_98003	-3.461e+05	1.38e+04	-25.164	0.000
zipcode_98004	4.629e+05	1.33e+04	34.779	0.000
zipcode_98005	-2.094e+04	1.66e+04	-1.262	0.207

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.

A plot showing variation of price according to zip-codes, relative to zip-code 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

Multiple Linear Model 2 was chosen as the final model as it explained about 74 % of the variance in price, about 10% more than Multiple Linear Model 1.

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.
2. An increase in the area of the living space by 1 square meter had the second highest associated increase in price.
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

- A larger size(square meters) of the living space.
- Increase the number of bathrooms
- A view also adds to the value of the house
- Building houses in the zip-code '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.
- Market research such as: population would improve the recommendation

Thank You.
Any questions?

MAUREEN KITANG'A

SAMUEL KYALO

PRISCILLA KAMIRI

JIMCOLLINS WAMAE

STEVE GITHINJI

LEO KARIUKI