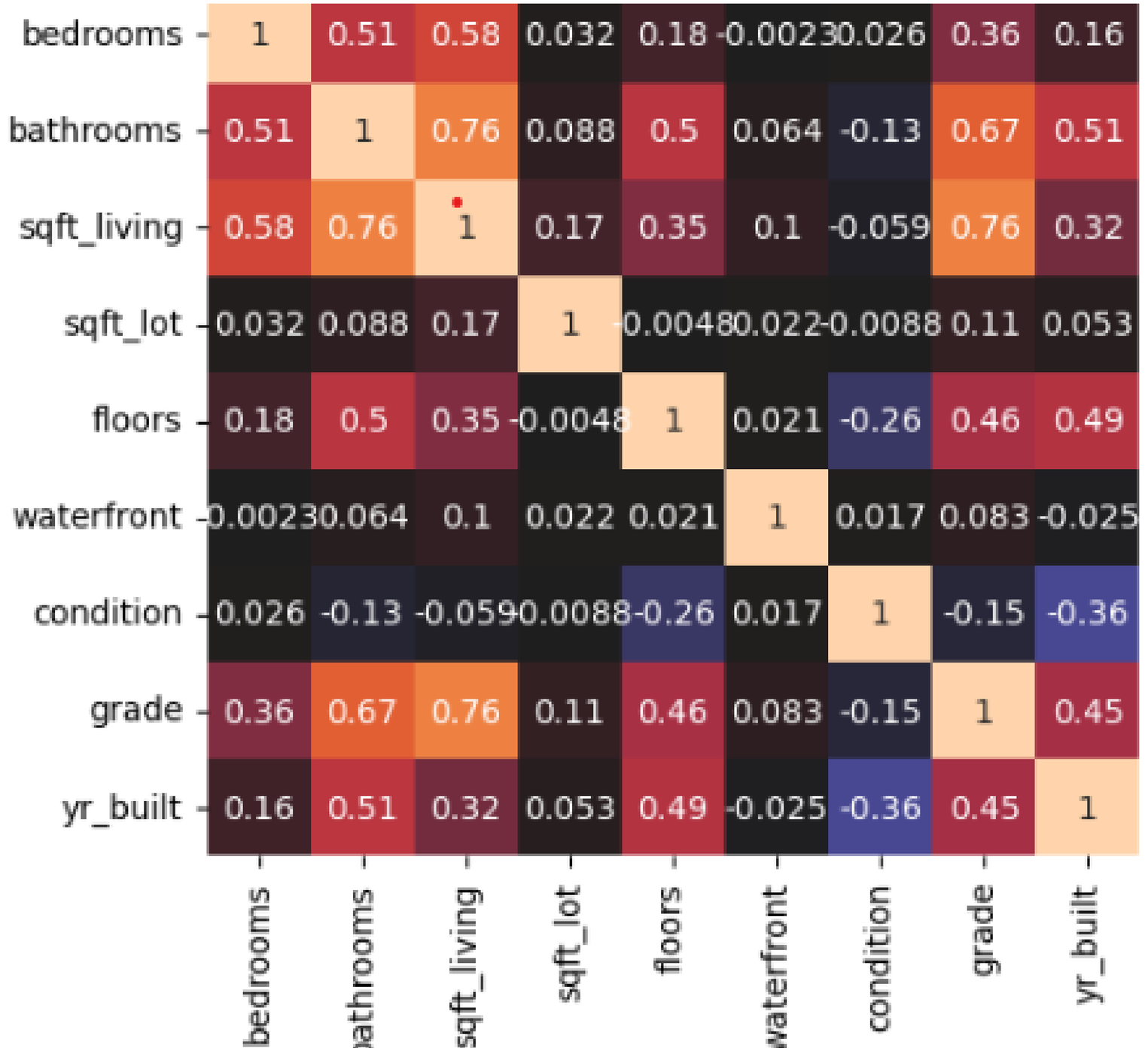
Student name: Bonface Mangoka¶
Student pace: part time
Scheduled project review date/time: 30/03/2023
Instructor name: Samuel Jane
Blog post URL: https://github.com/bonmwash/project_2

Business Problem

In this project, I am to provide expected/estimated home prices to homeowners based on the logistics of their home. This can also give insight on how home renovations might increase the estimated value of their homes, and what type of potential renovations are best.

The Data

This project uses the King County House Sales dataset, which is found in kc_house_data.csv in the data folder in this repo. The description of the column names can be found in column_names.md in the same folder. As with most real-world data sets, the column names are not perfectly described, so i had to do some research to understand the data.



sqft_living and grade = 0.76

sqft_living and bathrooms = 0.76

grade and bathrooms = 0.67

Let's remove grade and bathrooms for this model. We will also use our previous outlier filter, as this seems to be a step in the right direction.

```
regression_final = LinearRegression()
regression_final.fit(X_final, y_final)

y_hat_final = regression_final.predict(X_final)
rmse_final = np.sqrt(mean_squared_error(y_final, y_hat_final))

print(f'Test Root Mean Square Error: {rmse_final}')
```

Test Root Mean Square Error:
0.2617520712050872

Conclusion¶

We have an r-squared value of ~0.73, which is the most accurate model in our analysis.