Predicting House Prices in King County

Part One

Task A – Stating the Question

Predicting the sale price of properties is an important and challenging problem. It is also big business with companies such as Zoopla and RightMove offering online valuations of houses using machine learning techniques. In the US, a similar company called Zillow has launched a $1 million competition, called the Zillow Prize, to improve their house price estimates (Zillow Promotions, 2017).

This report provides details of our attempt to predict house prices using various regression methods. We will make use of data of house prices in King County, Washington State, USA for sales in 2014 and 2015.

Before posing our main question we pose two interesting descriptive questions:

1. What is the relationship between the number of floors in a house and the area of the land it sits on?
2. Have house sizes become bigger or smaller over time?

Our main, predictive question is:

> 3. *"Is it possible to predict the sale price of a house from information about that house such as the size, number of bedrooms, condition etc?"*

Our questions have the following properties, indicative of a high-quality data science question:

1. They are interesting
2. They are specific
3. They are plausible
4. They are answerable

Question 1 is interesting because it is not obvious whether people prefer fewer floors or larger gardens. When the area of the plot is larger there is the option to have a large house build on a single floor and more spread but this comes at the expense of a potentially larger garden. Question 1 examines this trade-off. It is also specific because we can provide a definite answer about whether a clear relationship exists and what that relationship is. It is plausible that the number of floors in a house may be related to the area of the plot because we generally see that where land area is sparse buildings are taller – e.g. in city centres. It is reasonable, therefore, to suggest that the same relationship may apply to houses. It is certainly answerable from our data which provides details on both the number of floors and the area of the plot.

Question 2 is interesting because it asks the non-trivial question of whether houses built more recently are, in general, larger or smaller than houses built decades ago. This could have an impact on social studies as it may indicate something about how people's demands and expectations have changed. It may also be of interest to economists considering the question of cost-of-living who may include factors such as house prices but may not account for changing sizes of properties. This phenomenon is becoming recognised in food products and has been dubbed "shrinkflation" (Ochirova 2017).

The question is also specific because both of the features are continuous and so we can determine the correlation between them and determine whether there is a trend. It is certainly plausible that house sizes have changed over time because we might expect house sizes to reflect various factors that have changed over time such as family sizes and average incomes.

Finally, the question is answerable within the time-frame available because we have a large collection of data ready which we can use to provide an analysis.

The main predictive question is interesting because, as mentioned, the problem of predicting house prices is certainly not trivial and an accurate solution is extremely valuable. If companies are willing to pay $1 million for an improvement in their algorithms, then it is evident that the problem is both challenging and important.

It is also specific because it is possible to be precise about what is being predicted – house prices – and the accuracy of predictions can be measured with various standard metrics.

The question is plausible because it is reasonable to suppose that the main factors in the price of a house are the physical condition of that house, including its size, number of rooms, condition etc. Furthermore, we have evidence of its plausibility in that many companies provide exactly this service already.

Finally, the question is answerable to the extent that we will be able to generate a prediction of house prices and measure the accuracy of those predictions. House prices datasets are not difficult to obtain and can be reasonably small and simple which will make it possible to analyse within the time frame available. However, we note that extremely accurate predictions require large amounts of data and very complex algorithms as evidenced by the need for global competitions.

For the purposes of this report we will use a dataset of house prices from King County which is an area in the US State of Washington that includes Seattle. The dataset was uploaded to the Kaggle website (Kaggle.com, 2017) by the user harlfoxem. Unfortunately, the user has not indicated the source of the data but we note that King County has an open data platform where the data may have originated (King County, 2017).

The dataset itself consists of 21,613 examples with 19 features. The features are:

Feature Name	Type	Comment
Bedrooms	Integer	
Bathrooms	Real	Values are decimal and calculated as: Full bathrooms (ensuite) = 1 Half bathrooms (separate)= 0.5 Powder rooms (only toilet and sink) = 0.25 (Dahlin, 2016)
Sqft_living	Integer	Total area of the house
Sqft_lot	Integer	Total area of the plot

Floors	Real	
Waterfront	Boolean	Does the property have a waterfront view
View	Integer	How many times the property has been viewed
Condition	Categorical	The condition of the property
Grade	Categorical	Grade according to the King County grading system
Sqft_above	Integer	Total area of the house above ground
Sqft_basement	Integer	Total area of any basement
Yr_built	Year	
Yr_renovated	Year	
Zipcode	Categorical	
Lat	Real	Latitude
Long	Real	Longitude
Sqft_living15	Integer	If there have been any renovations this value may be different to sqft_living
Sqft_lot15	Integer	If there have been any renovations this value may be different to sqft_living

From the inclusion of the “Grade” feature which is based on a system particular to King County, it would seem likely that the data originated from an official source. On the other hand, the existence of the “View” feature which relates to the number of viewings the property received suggests that the data originated with a real estate agent or company.

In either case, there is no reason to call into question the general accuracy of the data because values such as prices, sizes etc do not appear random and there would appear to be no reason to invent values. Nevertheless, some caution is appropriate before applying models trained on this data to more general cases.

Task B – Data Exploration

The aim of this section is to identify any potential issues with the features and decide on how to handle those issues. The primary concern is with outliers and unexpected distributions, though we might potentially encounter issues such as unexpected scales.

Price

The first part of the data we examine is the Price, which is the target in our study. Figure 1 shows the histogram of prices.

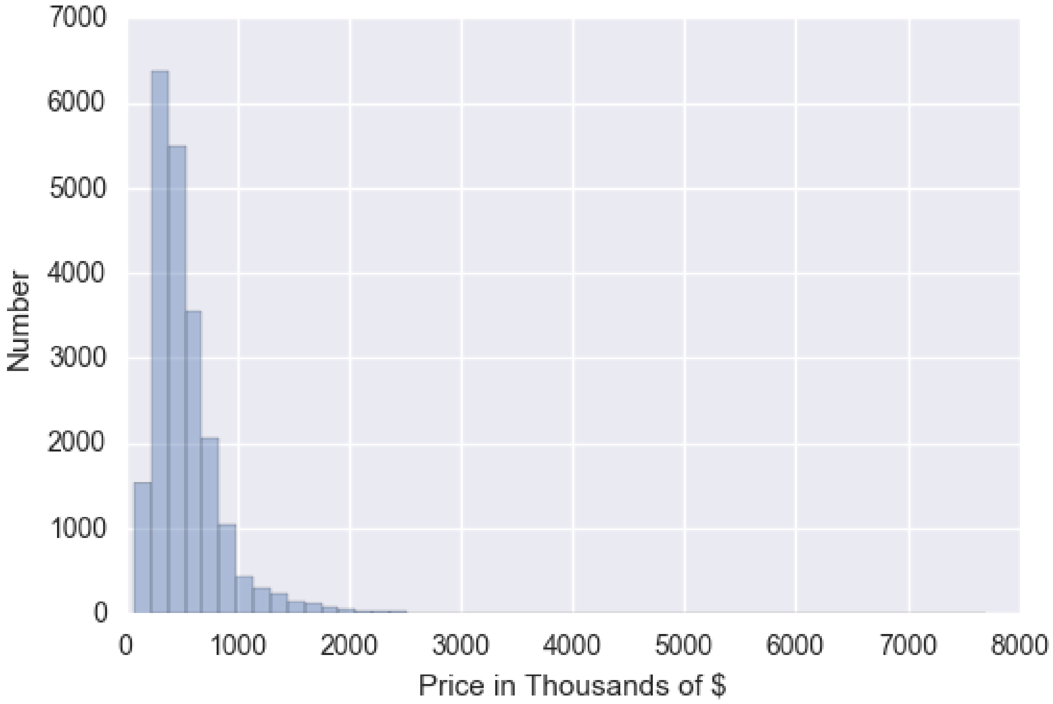

Figure 1: Histogram of House Prices

The histogram shows a normal distribution with a long tail, which is exactly what we would expect for data of this kind. The most expensive house cost $7.7 million and there were a total of 1,465 houses sold for more than $1 million. The median house price was $450,000.

Bedrooms

In the bedrooms data, there is an anomaly. The median number of bedrooms is just 3 but there is one entry with 33 bedrooms. Although such houses do exist, the price of this house is only $640,000, it has only 1.75 bathrooms and is on a single floor. This suggests that the property does not actually have 33 bedrooms at all. Most likely, the house has 3 bedrooms and the 33 is a data entry error. However, without knowing this for sure, the safer option is to exclude that entry from future analysis.

Bathrooms

As mentioned in the previous section, the values for bathroom are unusual because it is possible to have half- and quarter-bathrooms. Figure 2 shows the histogram of values for the number of bathrooms.

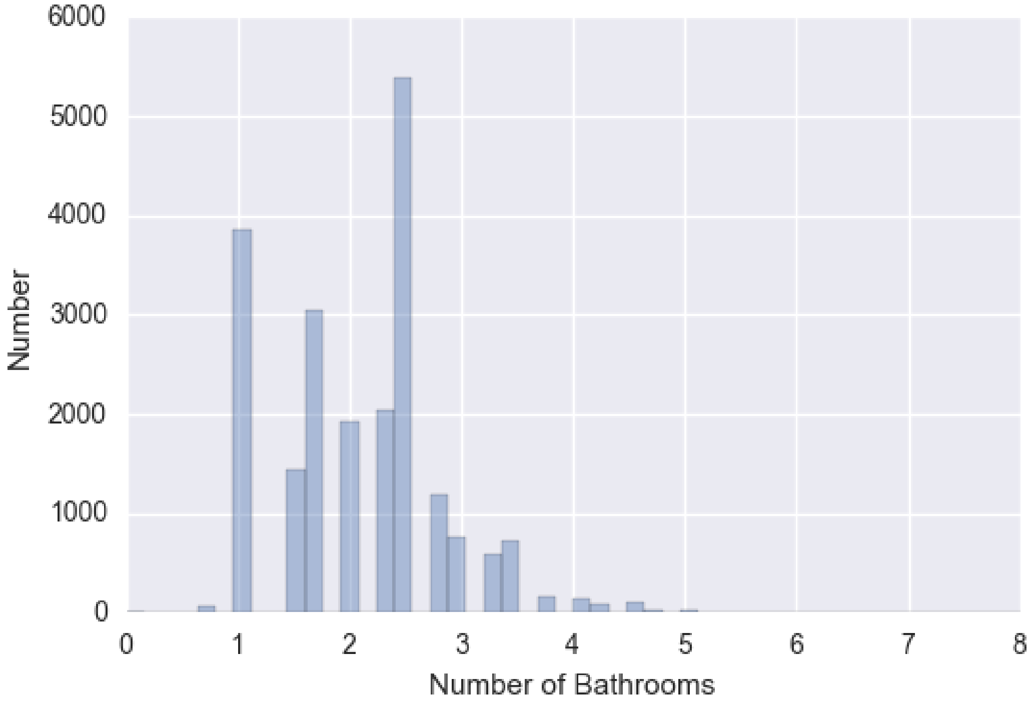

Figure 2: Histogram of the number of bathrooms

The median number of bathrooms is 2.25 and the mean is 2.1 which is consistent with a more or less normal distribution as seen in Figure 2. In this case we have a distribution that is close to normal because the tail is not very long.

Sqft_living and sqft_lot

The `sqft_living` feature gives the total area of the living space of the house in square feet. The median is 1,910 square feet and the largest property has an area of 13,540 square feet. Figure 3 below shows the histogram of values which is a normal distribution with a long tail, as we might expect.

The feature of `sqft_lot` gives the total area of the plot which includes the garden and grounds of a property. Unlike the total area of the living space, the area of the plot is extremely heavily skewed. A histogram of the values is shown in Figure 4, but the histogram is curtailed at 50,000 square feet because of the very long tail.

The median value is 7,619 square feet and the mean is 15,107.4 square feet.

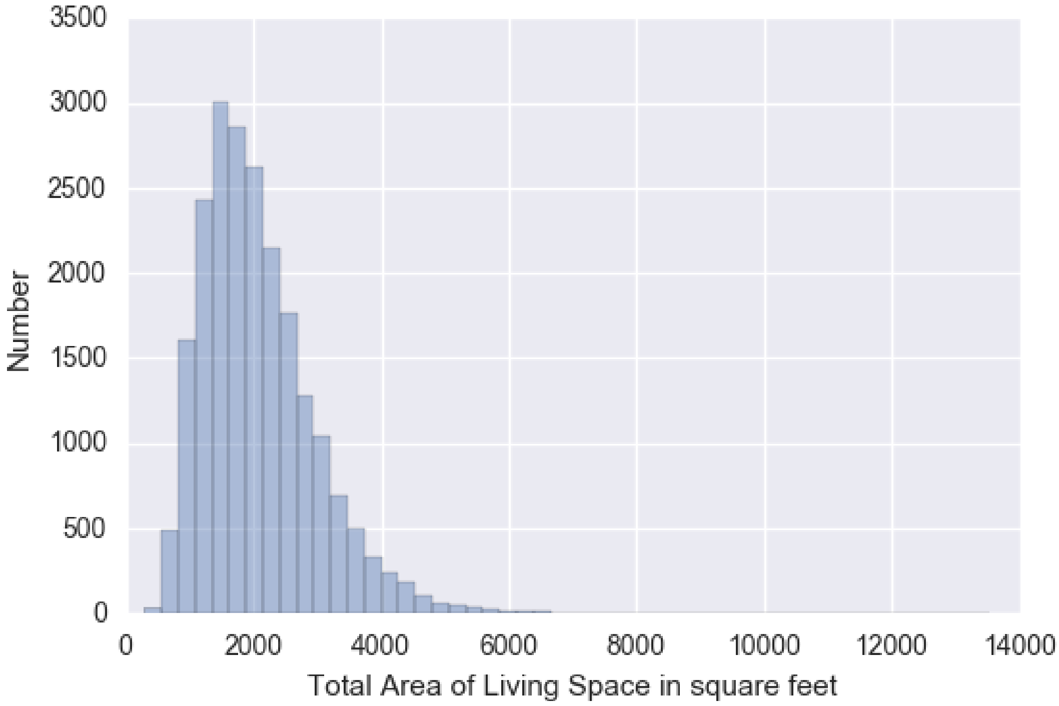

Figure 3: Histogram of values for sqft_living feature

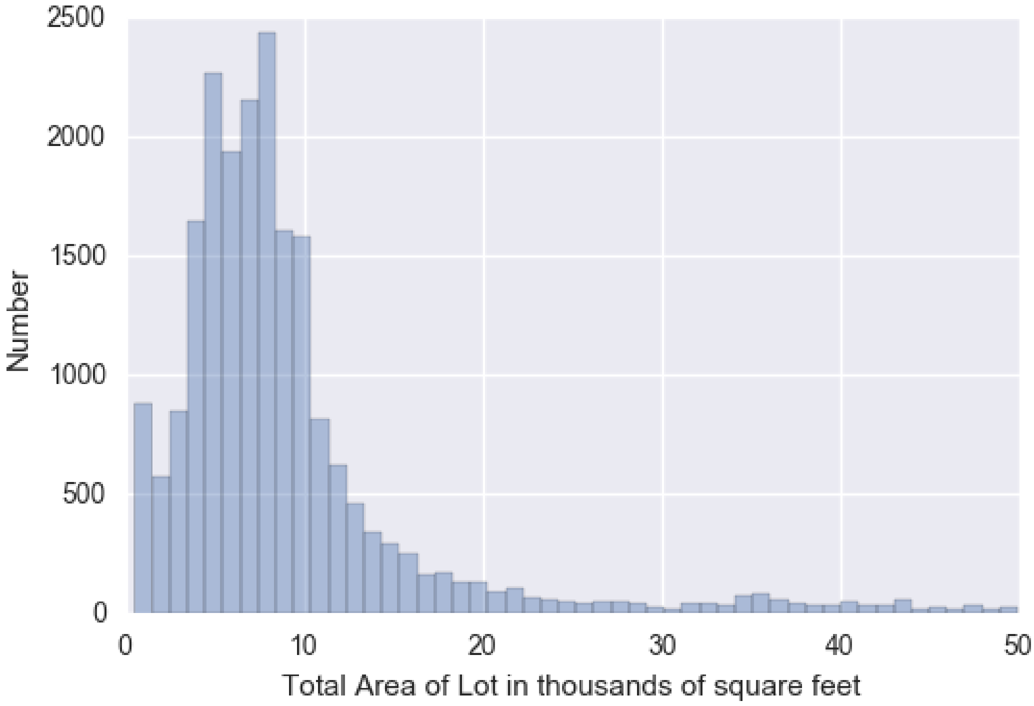

Figure 4: Histogram of values for sqft_lot feature

Floors

The number of floors in properties can be half values (e.g. 1.5, 2.5 etc) which is most likely because of Mezzanine floors. The median value is 1.5 floors. 10,679 of the properties are on a single floor, 8,241 on two floors and 1,910 on 1 and a half floors. The number with different values (up to a maximum of 3.5) are much smaller.

Waterfront, Condition and Grade

These features are all categorical and can be discussed together. The `waterfront` feature is 0 if the property does not have a waterfront view and 1 if it does. In the dataset we have 21,449 houses do not have a waterfront view and only 163 do. This suggests that this feature will not be very helpful in the general case.

The features condition and grade are assigned by the King County authorities and their definitions can be found on the King County website[1]. Condition is “Building Condition” and is “relative to age and grade”. It has values between 1 and 5, where 1 indicates a worn-out property and 5 indicates a very good condition. Grade is “Building Grade” which “represents the construction quality of improvements. Grades run from grade 1 to 13.”

The table below shows the number of properties with different conditions

Condition Rating	Number with Rating
1	30
2	172
3	14,031
4	5,679
5	1,700

The distribution for the grade is a normal distribution, as we would expect, with the median value being 7 which is defined by King County as “Average”.

View

It is unclear what this feature indicates because one source suggests it is whether the property has been viewed and another suggests it is a measure of the quality of the view from the property. It has values 0 to 4 but the vast majority (90.2%) of houses have a value of 0. Therefore, since the meaning of the feature is unclear and it has little differentiating power, it will be ignored in future analysis and prediction.

Sqft_basement and sqft_above

These two features are related in that sqft_above is simply sqft_living less sqft_basement. Therefore, if sqft_living is included in the set of features used for prediction and sqft_basement is used as well, there is no value in including sqft_above because it is a redundant feature.

A histogram of values for sqft_basement is shown in Figure 5 below. As the distribution shows, the majority of properties have no basement and indeed the median value is 0. Of those properties with a basement, the median size is 700 square feet and the mean is 742.4 which suggests a normal distribution with a slight skew owing to a long tail. The largest value is 4,820 square feet in a property with a total living area of 9,640 square feet.

Overall, since 60.7% of properties do not have a basement, this may not turn out to a very useful feature.

Year Built

This feature gives the year the property was built in. There are no missing values and all properties were built between 1900 and 2015. As the histogram in Figure 6 shows, the age of the houses sold

[1] http://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r

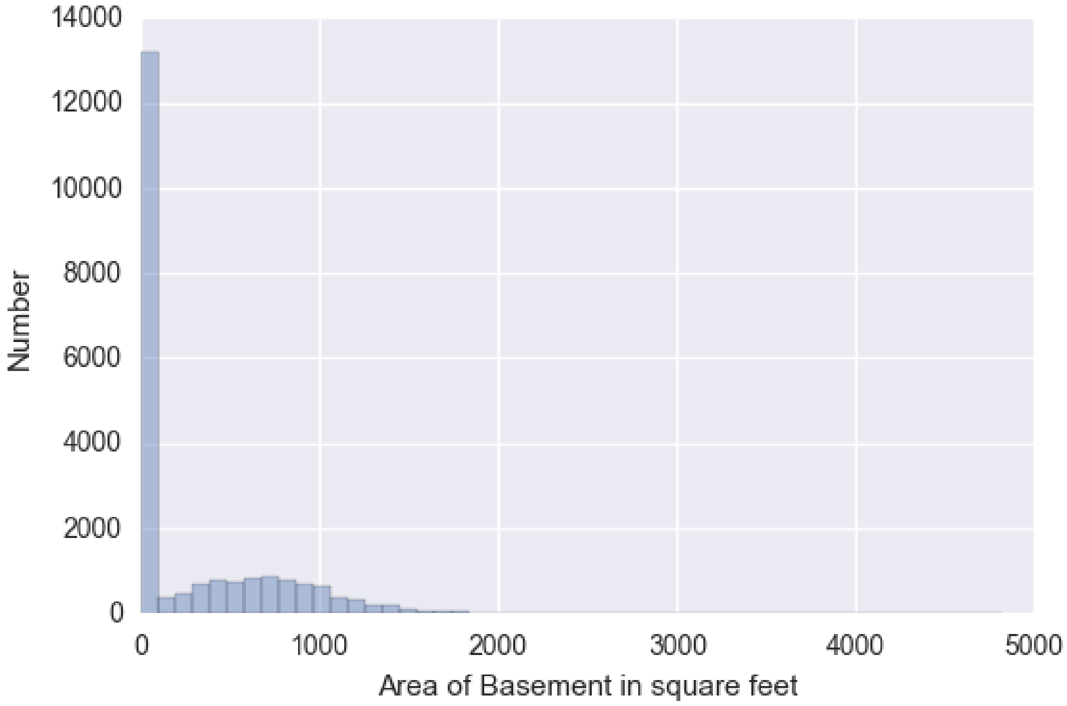

Figure 5: Histogram of sqft_basement feature values

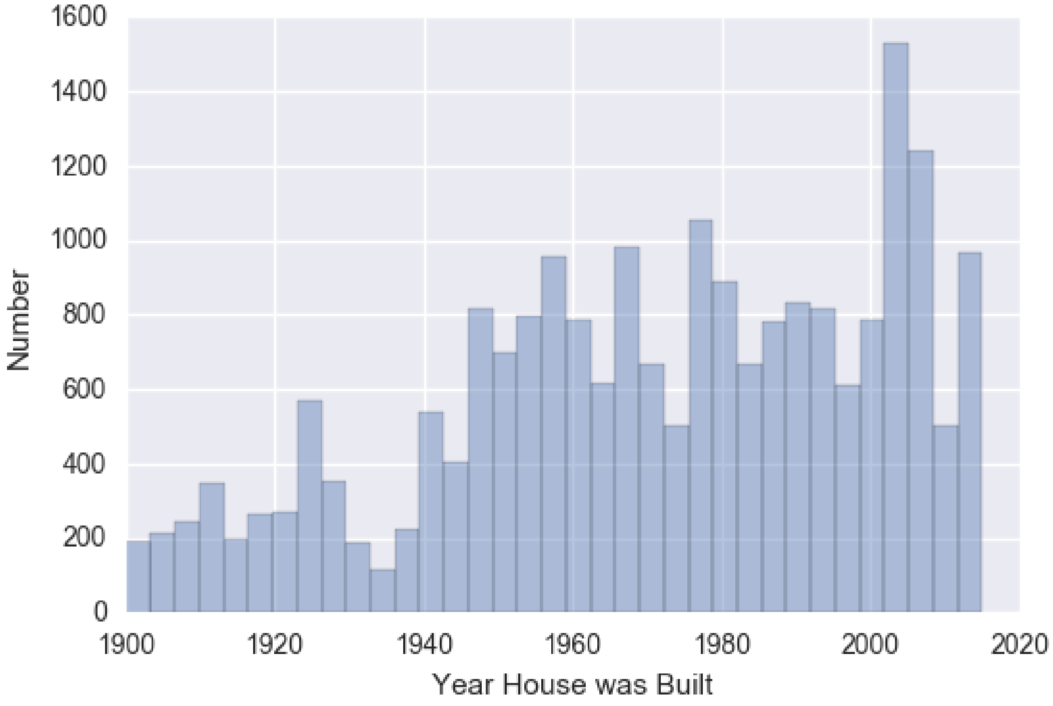

Figure 6: Histogram of years properties sold were born in

does not follow a simple pattern of any kind. However, it likely follows the distribution of general house building, hence we see a significant drop in the late 1930s and a spike in the early 2000s.

Year Renovated

This feature records the most recent year of renovation to the property. It is not clear how much renovation is required for this value to be updated nor who recorded the information. 95.8% of all properties have a 0 for this feature indicating that they have never been renovated. This is extremely unlikely and suggests that the information was not collected in the vast majority of cases. Given this unreliability and the lack of clarity concerning its meaning, the feature should be ignored in future analysis.

Zipcode

This feature is categorical, indicating the area the property sold is in. In total there are 70 unique zipcodes in the data with between 50 and 601 properties sold in each zipcode.

Longitude and Latitude

These features provide precise coordinates for the property sold. Whilst the area of a property is a key determinant in its value, this information is already recorded in the zipcode. The precision of longitude and latitude is perhaps too much for the simple algorithms that will be applied in this report. These features will therefore be ignored for further analysis.

Sqft_living15 and sqft_lot15

These features record values for the size of the living are and plot for properties in 2015, which may be different to the values in 2014 (the values in sqft_living and sqft_lot) because of renovations, new building work and land purchases or sales.

The median values for the changes for living area and lot area are 260 square feet and 520 square feet respectively. This compares to medians of 1,910 and 7,619 for the original values. In other words, the typical change is 13.6% for living area and 6.8% for the lot.

Moreover, only 11.9% of properties had the same values for living area in 2015 as in 2014 and 24.9% had the same values for lot area. Without further information this calls into question the accuracy of the values in these features because it seems very unlikely that 90% of all houses sold undergo significant building work in the year before or after their sale. It is even less credible that more than three quarters of sales are immediately preceded or followed by a change in the lot size.

Therefore, owing to the questionable values of the features these two features will be ignored in any further analysis.

Task C – Data Preparation

Before moving on to predictive analysis, it is necessary to prepare the data by removing outliers, unnecessary or unreliable features and normalising the data values.

From the preceding discussion, there are a number of features that will be excluded from future analysis. These are: View, sqft_above, sqft_basement, Yr_renovated, long, lat, sqft_living15 and sqft_lot15. These features will therefore be removed in a copy of the dataset to be used for further analysis. We also observed one erroneous examples where the number of bedrooms was listed as 33. This example will also be removed.

Further preparation is needed in two ways. Firstly, because we will initially be applying linear regression, non-ordinal, categorical features must be transformed into a series of Boolean indicator variables. Of the remaining features, this only applies to the zipcode. Unfortunately, there are 70 unique zipcodes in the data which would mean that we would end up with more than 70 features to be used for learning which is excessive at this point.

Therefore, for the initial analysis we will also ignore the zipcode feature. An alternative option would be to partition the data based on the zipcode and apply predictive analysis to subsets of the data where all examples have the same zipcode. However, this would mean reducing the number of examples for more than 20,000 to at most 600.

Finally, there is the issue of normalisation arising from the different range of values in each feature. When using multivariate linear regression, the model will sum the contributions from various features and features with higher values will inevitably count for more when perhaps there is no reason for them to do so. The straightforward solution is to normalise the ranges of all features into the range (0,1).

Part Two

Task A – Model Building I

In this section we will answer our two descriptive questions:

1. What is the relationship between the number of floors in a house and the area of the land it sits on?
2. Are houses with a waterfront view significantly more expensive than other houses? If we control for all other factors does this change?

Question One

This questions deals with the relationship between two variables of different types. Specifically, the number of floors is an ordinal variable – meaning that it is categorical but with an order – whereas the area of the land is a continuous variable. This rules out a number of correlation measures such as Pearson's Correlation Coefficient.

To get a feel for the answer we can plot a boxplot of the distributions of the plot sizes, grouped according to the number of floors. This is shown in Figure 7 below.

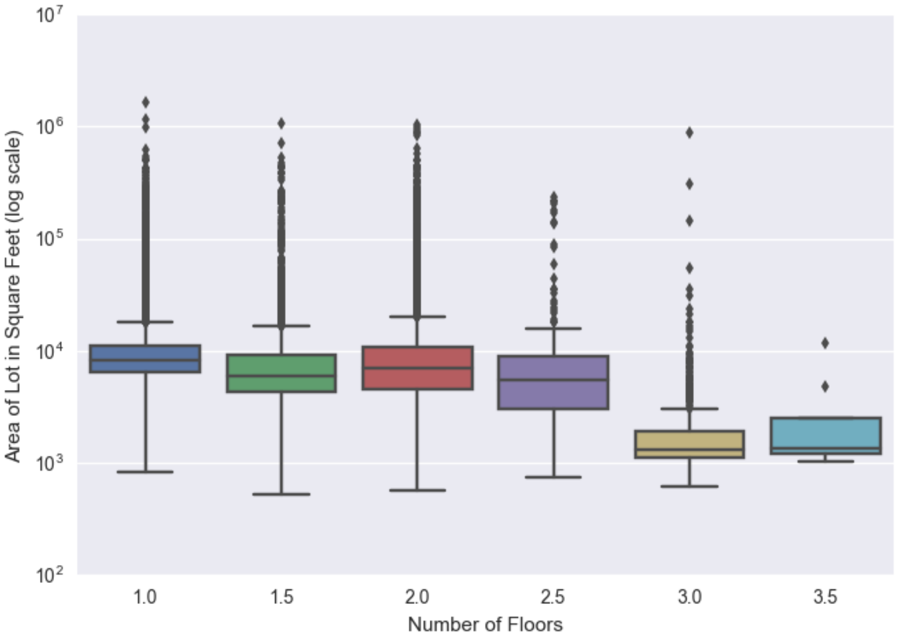

Figure 7: The distribution of the area of lots grouped according to the number of floors

The central, horizontal line in each box indicates the median for that group. The results suggest a relationship between the two such the area of the lot tends to be higher when there are fewer floors. This relationship is less pronounced for houses with fewer than three floors but houses with three or three and a half floor have a significantly lower median lot area than other houses.

One metric that can be applied to measure the relationship between a continuous variable and an ordinal one is Spearman's Rank Correlation Coefficient. This metric does not measure the correlation between the values of the variables but between their ranks once ordered. The metric returned is known as Spearman's ρ (as compared to Pearson's r).

For the case of number of floors and area of lot, Spearman's ρ is -0.23 ($p < 0.001$). This confirms what we can see in the boxplot, namely that there is a negative relationship between the two variables. However, Spearman's ρ varies in the range (-1,1) so a value of -0.23 indicates only a weak relationship.

A final avenue of exploration is to apply hypothesis testing to determine whether the distribution of lot sizes is statistically significantly different depending on the number of floors. That is, we can ask whether there is a statistically significant difference between the distribution of lot sizes for houses with one floor and houses with two floors (for example). The most common type of this kind of test would be Student's t-test but that is only applicable when the distributions are normal, and we know that the distribution of lot sizes is heavily skewed. We must therefore use the non-parametric Mann-Whitney U-test.

Using this test, we find that there is a statistically significant difference between the distributions of the area of the lot depending on whether there is one, one and a half, two or three floors. That is, pairwise Mann-Whitney tests on these groups returned a p-value much less than 0.001 indicating that we can reject the null hypothesis that the groups were drawn from the same distribution.

Question Two

This questions asks us to examine the relationship between house sizes over time. The first step towards answering this question is to produce a scatter plot of the house sizes against year of construction, which is shown in Figure 8 on the next page.

The results show that there has been an upward trend in the size of houses over the period 1900 to 2015. The cyan line is the line of best fit for the data. We can use Pearson's Correlation Coefficient to evaluate the strength of the relationship. The Pearson r value is 0.05 ($p < 0.001$) which indicates a statistically significant relationship but a weak one.

However, arguably Pearson's Correlation is inappropriate in this case for two main reasons. Firstly, the year built may be considered an ordinal rather than a continuous variable. Secondly, the relationship may not be linear as there may have been changes over time. We therefore also applied Spearman's Rank Correlation Coefficient. This, though, confirmed the weakness of the relationship, giving a ρ value of -0.04 indicating a very weak relationship (and in fact a negative relationship).

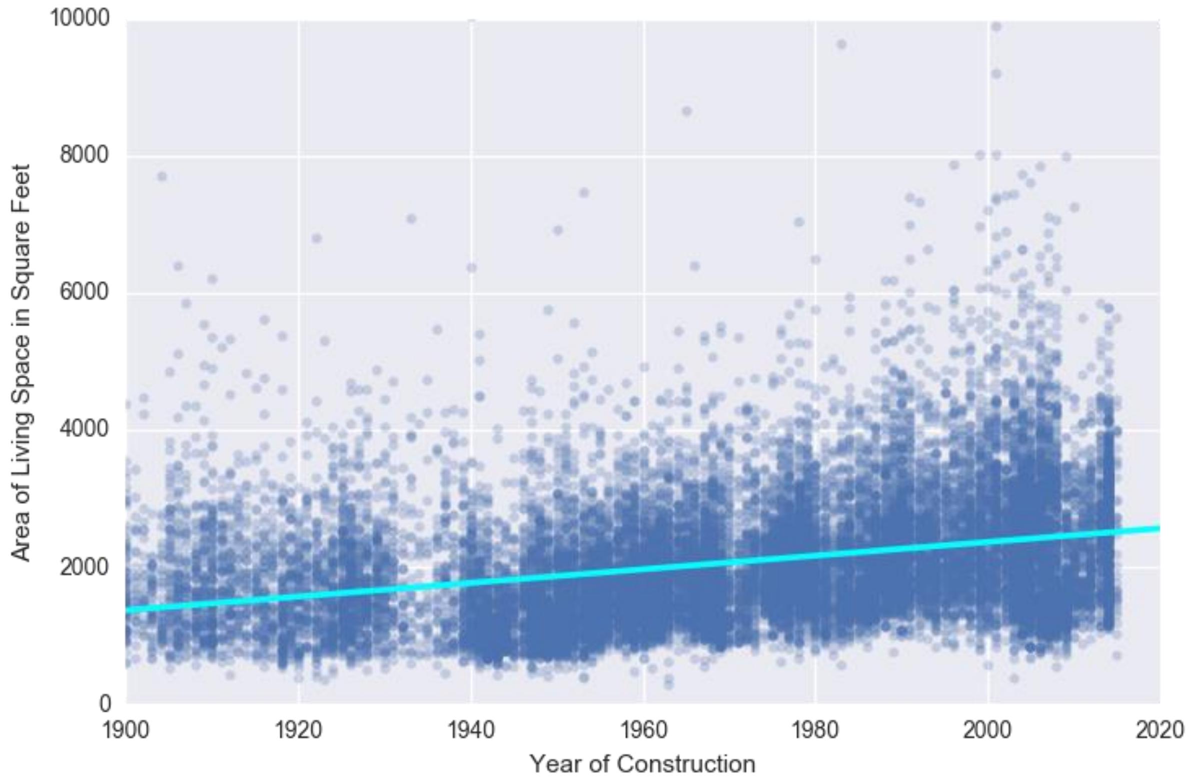

Figure 8: The area of houses against the year they were built in

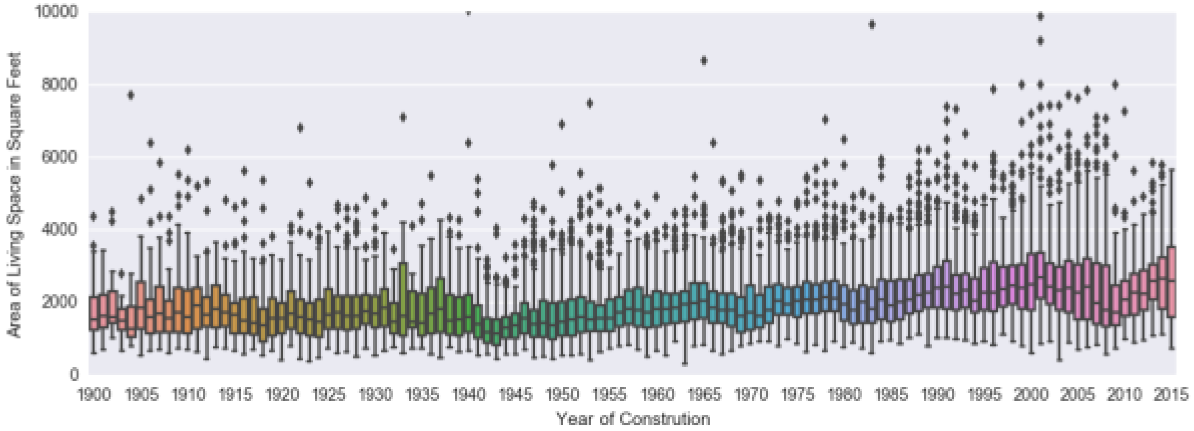

Figure 9: Boxplot of house sizes over time

To conclude the discussion, Figure 9 above shows the boxplot of the house sizes by year of construction. Looking at the medians, it appears that in the early part of the period (1900 to 1940) there was not much change in the typical house size. In the post-war period (1945 to 2000) we see a sustained upward trend in the median house size which perhaps roughly corresponds to the sustained period of economic growth. This trend is reversed between 2000 and 2010 when the economy suffered two significant crashes (the dot-com bubble bursting and the credit crunch). Since 2010, however, there has been a significant upward trend once more.

This mixed picture explains the lack of correlation as found by Pearson's Correlation Coefficient and suggests that a more complex time-series analysis is required. However, such an analysis is beyond the scope of this report and must be left for a future exercise.

Task B – M Model Building II

Predictive Analysis – Univariate Linear Regression

For the first attempt at predicting house prices from the data, we will use univariate linear regression. This is the simplest form of regression which uses a single feature in a straight line model. Although we do not expect to achieve accurate results with such a simple model it will provide some baseline metrics to compare performance to. Further, there remains the possibility (however small) of univariate linear regression giving acceptable accuracy in which case it would be preferred.

For all predictive analysis in this report, k-fold cross validation is used to reduce the risk and impact of overfitting. Specifically, the data is randomly split into ten folds and each fold is used once as the testing set and nine times as part of the training set.

The metric we use is the Root Mean Square Error (RMSE) which gives a figure in $ for the average difference between the predicted values and the true values. We also put this into context by providing the normalised RMSE which is the RMSE divided by the true mean of house prices and is expressed as a percentage.

The results are presented in the table below:

Feature	RMSE	NRMSE
bedrooms	$347,331.6	64.3%
bathrooms	$311,955.5	57.8%
sqft_living	$261,356.3	48.4%
sqft_lot	$364,982.8	67.6%
floors	$354,092.8	65.6%
waterfront	$353,286.2	65.4%
condition	$366,060.9	67.8%
grade	$272,901.0	50.5%
yr_built	$366,352.6	67.8%

The results show that univariate linear regression does not work all that well, with the best result (sqft_living) still being on average almost 50% incorrect.

Multivariate Linear Regression

To improve the results, we create a model which combines all the features into a single model.

This method leads to a RMSE of $218,281.8 with a normalised RMSE of 40.4% which is a significant improvement over the univariate linear regression models.

When applying multivariate linear regression, the model produces coefficients for each feature which indicates the weight given to each feature and whether the learned relationship is positive (the price increases as the feature value increases) or negative. The table below shows the weights for each feature in the multivariate model:

Feature	Coefficienct
bedrooms	-42,731.7
bathrooms	37,713.7
sqft_living	167,430.4
sqft_lot	-9,697.8
floors	12,099.4
waterfront	60,918.3
condition	12,521.9
grade	149,117.9
yr_built	-112,311.8

We can see that the most important features are the size of the living area, the grade and the year of construction, with older houses being worth less than newer builds. It is also interesting that the learned model has produced a negative relationship between the number of bedrooms and the price but a positive one between the number of bathrooms and the price. This seems counter-intuitive and is probably the result of an unexpected interaction between features.

Polynomial, Multivariate Regression

Taking the above approach further, we apply polynomial multivariate regression, with powers up to three. The polynomial features are the base features, the base features squared and cubed and also the base features all multiplied in pairs and in triples. This takes the number of features used from 9 to 220.

The result is that the RMSE is $207,060.1 with a normalised RMSE of 38.3%. Therefore, we see an improvement but not a significant one.

Ridge Regression

One of the dangers of polynomial regression is that it can lead to overfitting. Using cross-validation can help to notice when this is happening but does not prevent it. Typically, when a model is overfit, the coefficients are very large. In the polynomial multivariate regression model, the average coefficient (in absolute terms) is 13,559,365 which is extremely large and therefore indicates significant overfitting.

Ridge Regression is a technique that can help reduce overfitting by applying a “punishment” to terms with large coefficients. This is a form of regularisation and can lead to more accurate results.

Using Ridge Regression brings down the average absolute coefficient to 116,761 which is two orders of magnitude smaller. The RMSE is $205,430.7 and the normalised RMSE is 38.0%. It is therefore evident that although the coefficients were reduced, this has not led to a significant increase in accuracy.

Nearest Neighbour Regression

The final method we consider is nearest neighbour regression (knn) which is a form of lazy learning. Unlike the other methods we have used, knn does not build a model from the features to the price. Rather, it uses the features of the example being predicted to identify the k examples in the training set that are most similar. It then averages the prices of those k houses and uses that as the predicted price.

For this report, we used k=5 and applied a weighted average by distance. The RMSE was $204,267.66 and the normalised RMSE was 37.8%.

Task C – Interpret

We have applied numerous machine learning algorithms to the problem of predicting house prices in King County. The table below summarises the best results:

Method	RMSE	NRMSE
Univariate Linear Regression	$261,356.3	48.4%
Multivariate Linear Regression	$218,281.8	40.4%
Polynomial Linear Regression	$207,060.1	38.3%
Ridge Regression	$205,430.7	38.0%
Nearest Neighbour Regression	$204,267.7	37.8%

The results show that predicting house prices is an extremely challenging problem. Our best result still contained an average error of over $200,000.

Part of the problem is that there is a small number of features, exacerbated by our decisions to remove some we considered unreliable or difficult to use. We note, however, that when using nearest neighbour regression we can include many of the difficult to handle features such as zipcode, latitude and longitude. Using 14 features (i.e. only excluding view and yr_renovated) with knn gives a RMSE of $158,452 and a normalised RMSE of 29.3% which is a significant improvement. This suggests that using a larger set of features is critical to achieving high accuracy in this problem.

We are also limited to using relatively simple algorithms. It is likely that more complex algorithms, such as artificial neural networks, may provide better results.

Conclusion

In this report, we attempted to predict the price of houses in King County. We found that the best we could achieve was a RMSE of a little more than $150,000 which we do not consider to be very accurate. Predicting house prices is an extremely complex and challenging problems because houses vary widely and house prices are not only based on the physical properties of a house but also on the emotional, social and financial position of the parties involved.

Our results indicate that in order to provide accurate predictions of house prices, a very large number of features must be used and that they most likely need to be combined with a powerful, complex and non-linear model.

References

Dahlin, E. (2016). The Rebuilding and Restoration of America: Get What You Want Not What You're Given. Rosedog Books, p.340.

Kaggle.com. (2017). House Sales in King County, USA | Kaggle in Class. [online] Available at: https://inclass.kaggle.com/harlfoxem/housesalesprediction [Accessed 17 Oct. 2017].

King County. (2017). King County | Open Data | King County | Open Data. [online] Available at: https://data.kingcounty.gov/ [Accessed 17 Oct. 2017].

Ochirova, N. (2017). The impact of Shrinkflation on CPIH, UK: January 2012 to June 2017. Office for National Statistics.

Zillow Promotions. (2017). *Zillow Prize*. [online] Available at: https://www.zillow.com/promo/zillow-prize/ [Accessed 17 Oct. 2017].