香港浸會大學
HONG KONG BAPTIST UNIVERSITY

Course Code: ISEM3035
Course Name: Introduction to Business Data Analytics and Visualization (Section 1)
Lecturer: Dr LI Mengxiang
Group Project: Data analytics on a mortgage dataset

Group:
14223570 WONG Cheung Tsin Alex
14220520 FONG Sun Yui
15223361 TSUI Pui Ying

Table of Content

Executive Summary

This is a report analysing a dataset of the mortgage in the USA in 2017. The objective of the report is too explore insights from the variables that provide significant business value to a financial institute. We assume that the insight would help the company to develop its mortgage business into a more in-depth area in the US. Under this circumstance the team decide to put focus on ***how to allocate the business resources by states in order to optimise the cost effectiveness*** for the company's future possible expansion in USA.

For some general exploration, we would first generate some random hypothesis to see if there are any significant relationships between location fields as well as statistical fields. The aim of this stage is to have an idea on which factors has more influence on the mortgage and second mortgage value, and to come up with some characteristics of some districts.

The software we are going to use includes R studio and Tableau. For R, we would mainly deploy linear regression to examine and predict the relationship between target variables and the mortgage. For Tableau, we would be using it for visualising the distribution of mortgage and second mortgage respective to the top 5 states (as our focus is to explore which states has more promotional value).

Last but not least, after the analysis result is interpreted, correlation between the selected variables would be shown, and we would suggest which districts have the greatest potential to be invested and give out some advice on the focus on states which the mortgage company may consider to put more business efforts into.

The Idea

We want to analyse the data set out of marketing and promotional aspects, by giving suggestions on how to strategically invest in according to the ranking to the value of coefficients. The ultimate goal is ***to find an optimal location for future business development***.

As for the preliminary stage, we would seek for a general picture by plotting all relationships between the geographical data and the statistical data (plot()). By discovering the variables which might have statistically significant correlations between one another, we would further break some of the variables down, in order to drill down into the insight. By that means, we decided to choose the top five states (according to their annual growth rate) to conduct a break down analysis against other statistical variables.

The description of data set

This data set is retrieved from 2011-2015 ACS 5-year U.S. Census Reports and 2016 U.S. Gazetteer Files. From a 60-months predictive table, it is covering all geographical areas' information. Since using the largest sample data set, this research is higher in accuracy and thus is most reliable examining the traits of very small population geographies.

Generally speaking, the dataset consists location field datas with respect to statical field datas.There are 1090 misused or non-effective datas in the set, they are screened out from our original dataset. For the location field, it consists of states, county, city and place, latitude and longitude geographic information, which is valuable for precise and area-targeted investigation. However, most of the nominal datas are just continuous and do not make some investigation value, whereas some of the fields (i.e. ALand, Awater, lat, lng) have such a huge variance in surface area which could not be compared at all.

Meanwhile, statistical field could be categorized into monthly mortgage and owner costs, monthly owner costs, gross rent, household and family income. However, most of those datas do not construct linear relationships or are too logical to predict, so they are not qualify for the investigation or do not carry much business value. Amongst those fields, monthly owner costs and mortgage median would be our main investigation variables. Both of these variables could provide business insights with comparison to the cities' population growth and GDP growth rate.

Having subtracted and narrowed down the variables, we still found the dataset is too big to generalise some interesting insights, therefore, we are going to subsequently break them into small pieces for more precise analysis, with the assistance of some external data and data visualisation.

The description of the targeting variables

An investigation into 6 variables in the dataset to mortgage would be conducted, for five cities with most GDP growth rate. We have illuminated the unpredictable and descriptive data, (e.g. city, place, post code, type, lat, lng), and conducted a correlation test to identify those variables which have a possible linear relationship with the targeted variables. Below is the list of those variables:

1. The median Monthly Owner Costs of a specified geographic location (hc_median)
2. The median Monthly Owner Costs of a specified geographic location (hi_median)
3. The Marriage rate of a specified geographic location (Married)
4. The Divorced rate of a specified geographic location (Divorced)
5. The median Male Age of a specific geographic location (Male_age_median)
6. Percentage of males with at least high school degree (hs_degree)

External data source respective to the growth rate of the states are quoted. They include GDP per capita and population growth. One of which would be selected as a determiner of the ranking compared to those 6 variables.

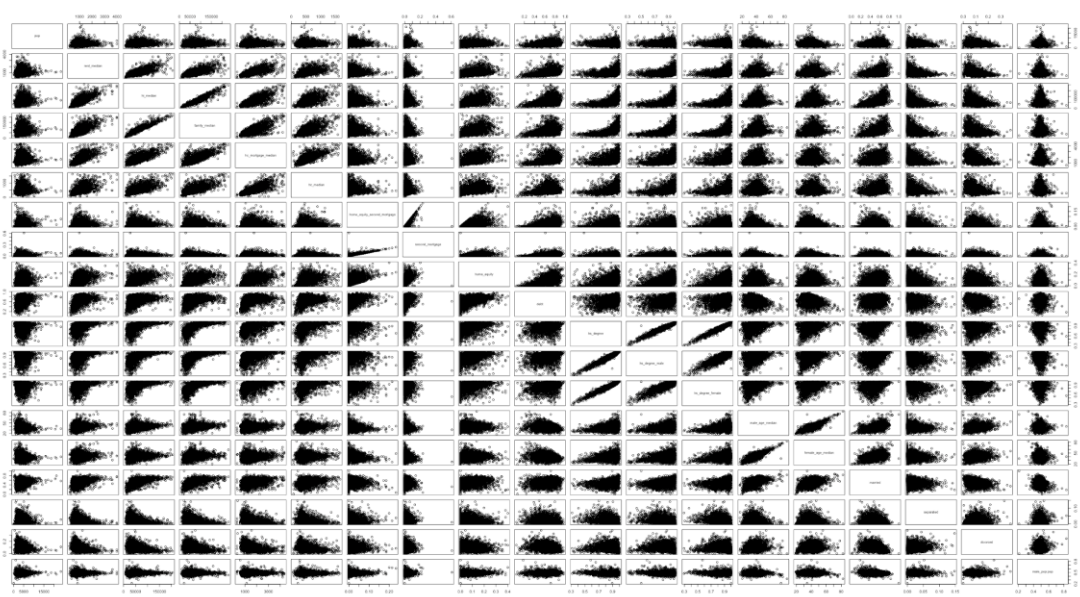
The description of analysis method

We would adopt the plot function in studio R to understand the relationship between data. After that, we will use out domain knowledge to select the factors related to the mortgage median and adopt the linear regression analysis in R studio to understand the relationship between data and mortgage median. To elaborate, linear regression represents means of supervised learning, in which least squares would be used to fit this data set analysis. By linear regression we study the relationship between the mortgage median and the selected variables, and also examine whether the variables relationship are strong or weak against mortgage median. We observed the whole model by the values of coefficient, standard error and the p-value, which represents the correlations, relative percentage of error of the model, and also the statistical significance respectively.

Finally, we will visualise the mortgage data for understanding the data by state (using Tableau Visualisation). This is a simple plotting of datas we selected in order to help building better business decision for the executive level.

The analysis procedure

We have obtained a general picture after running the plot function in studio R. Below is the graph showing the result:



Within these 5 sub-set, including state “Washington”, “Colorado”, “Nevada”, “Arizona” and “Utah”, we run linear regression to find the significant of the valuable factor selected. How the valuable factors are selected? By using best subset method, we use the reference provided by the

R algorithm and modified with the domain knowledge in the mortgage industry. “Hc_median”, “hi_median”, “married”, “male_age_median”, “hs_degree” and “divorced” are selected for all 5 states. However, there are some states may not follow this pattern. We would investigate them one by one. After running the regression model, we would sort the variable factors by the correlation with P-value considered. We will then drill down to investigate the top 3 factors for each state, with plotting the graph to visualize the correlation.

The analysis results

By looking into the p-value, we will only look at to the factor that have a value closed to zero, indicated with 3 “star” mark in R summary function. “hc_median” and “hi_median” are factors which are significant in all 5 state.

For Washington state, “married”, “hs_degree” and “divorced” are significant in p-value.
For Colorado state, “married”, “divorced” and “male_age_median” are significant in p-value.
For Nevada, there are no other significant factor except “hc_median” and “hi_median”.
For Arizona and Utah, only “married” is the significant variable.

Generally speaking, For Washington, Colorado and Nevada, all the variables selected have the p-value under 0.05. However, for Nevada, only “hc_median” and “hi_median” have the p-value under 0.05. For Arizona, “hs_degree” has a p-value higher than 0.05. We will then exclude the factor higher than p-value of 0.05 and start ranking the factor by the correlation.

As “divorced”, “married” and “hs_degree” are counted in decimal place format. We may divide the correlation result by 100 to get a picture of how 1 percent change could affect the result. (Please refer to Appendix I for details of R analysis result.)

The business implications and potential business values

Business implications of Regression analysis

When choosing a location for new branch or for business development. A firm may consider different factors such as the education level, median age of district, to decide the best location to open a store or branch.

In this study we found that there are 6 variables (hc_median, hi_median, married, divorced, male_age_median, hs_degree) are correlated to the mortgage median value. We believe the mortgage median is directly related to the business of a financial institution, since the higher the mortgage median, the higher amount loan will be landed. By considering the factor correlated with the mortgage median, the company may consider the “Right” factors when opening a branch.

Below are the table of Coefficients of factor correlated with the mortgage median. We suggest the firm to consider the coefficient of factor with the demographic data visualised in the Tableau (Please refer to link in Appendix III) to consider which stated should they open a new branch.

State	Factors	Coefficient
Washington	hc_median	1.23300
	hi_median	0.01058
	married	-884.20000
	hs_degree	472.60000
	divorced	-813.70000
Colorado	hc_median	1.01900
	hi_median	0.00924
	married	-844.90000
	divorced	-1594.00000
	male_age_median	14.11000
Nevada	hc_median	1.21500
	hi_median	0.00935
Arizona	hc_median	1.44200
	hi_median	0.00996
	married	-334.60000
Utah	hc_median	0.89530
	hi_median	0.01072
	married	-582.00000

One common factor we observed in 5 states is that the coefficients of "hc_median" are closer to 1 ,and the values of "hi_median" have a tendency closed to 0.01, which means the correlation of these two factor with the mortgage median is very small.

In Washington, we noted that the "married" and "divorced" has a negative correlation with the mortgage median, and "hs_degree" has a positive correlation with the mortgage median, where the correlation of "hs_degree" is 472.6, which is the highest among 5 states, a 1% increase in "hs_degree" . Therefore the firm should choose a location where the "married" and "divorced" is low and "hs_degree" is high.

In Colorado , we noted that the "married" and "divorced" has a negative correlation with the mortgage median where the coefficient of divorced is the lowest among 5 state(-1594), which means a 1% increase in "married" will cause a $1594 decrease in the mortgage median. and "hc_median", "male_age_median" and "hi_median" has a positive correlation with the mortgage median. Therefore the firm should choose a location where "married" and "divorced" is low and the "male_age_median" is high.

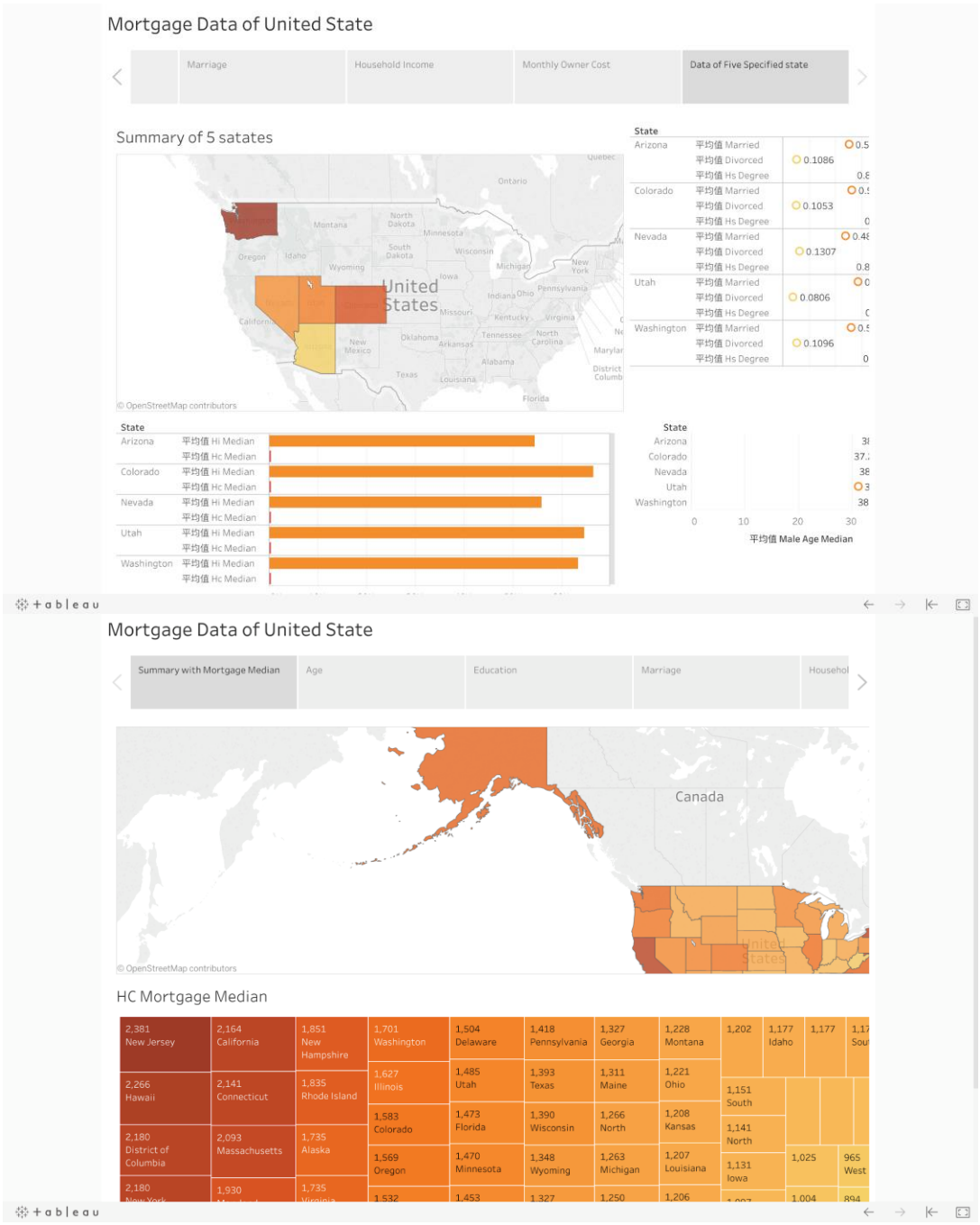
In Arizona, the correlation of "married" is -334, which means a 1 % increase in "married" will cause a $334 decrease in the mortgage median. Therefore the firm should choose a location where the "married" is low.

In Utah, the correlation of "married" is -582, which means a 1% increase in "married" will cause a $582 decrease in the mortgage median. Therefore the firm should choose a location where the "married" rate is low.

Data Visualisation

First, we have shown the Household cost, Mortgage Median, Age, Education, Marriage, Household Income, an Monthly Owner cost data to show a general picture of data across the united states by states.

After that, we have summarize the data of the 5 specific state to create a dashboard to demonstrate the average of 6 variables by state (Including hc_median, hi_median, married, divorced, male_age_median, hs_degree) and we have shown the data by descending order in circle view, bar chart and in a heat Map for easier comprehension. (Please refer to Appendix III) By reading the visualised data, users may understand the ranking for different factors affecting the mortgage median by state , thus, select the best state for business development.



(Sample: Mortgage Data of United State)

Conclusion

Outcome

In this project, we started by filtering out the unrelated factors to eliminate the noise in data, and plotted a scatter diagram to understand the correlation between different factors. Then we conducted a regression analysis on the data by state to find out the factors which are correlated with the mortgage to help the firm understand which factor should it focus on to select the optimal location for business development. Also, we have visualized the data by business illustration tool “Tableau” to have a general picture of the related factors and we filtered the 5

states with highest GDP growth rate which we believe is the indicator of the states with highest potential economics growth.

There are different datas and factors we can include into the study. However, we found that most of the factors are unrelated or we considered have no business value, also there are a lot of missing data in the raw data, therefore we have excluded those data in in this study. We found that the marriage rate is the only common factor among 5 states which has a negative correlation with the mortgage median.

Limitation

However, there are certain limitations we noticed in this study which we think that it might have higher business values. First, there are insufficient amount of datas to study the correlation of the average house size and the mortgage median. Second, we could not study the correlation between different cities, since the data only show a average value per a specified area. Third, we do not have any individual data of the area and could not perform a valuable analysis on the individual data for suggestion on the mortgage data. Therefore, we suggest the company to provide another supporting data set to fill up the missing gap we mentioned, in order to perfect the analysis and the insights we provide.

Citation

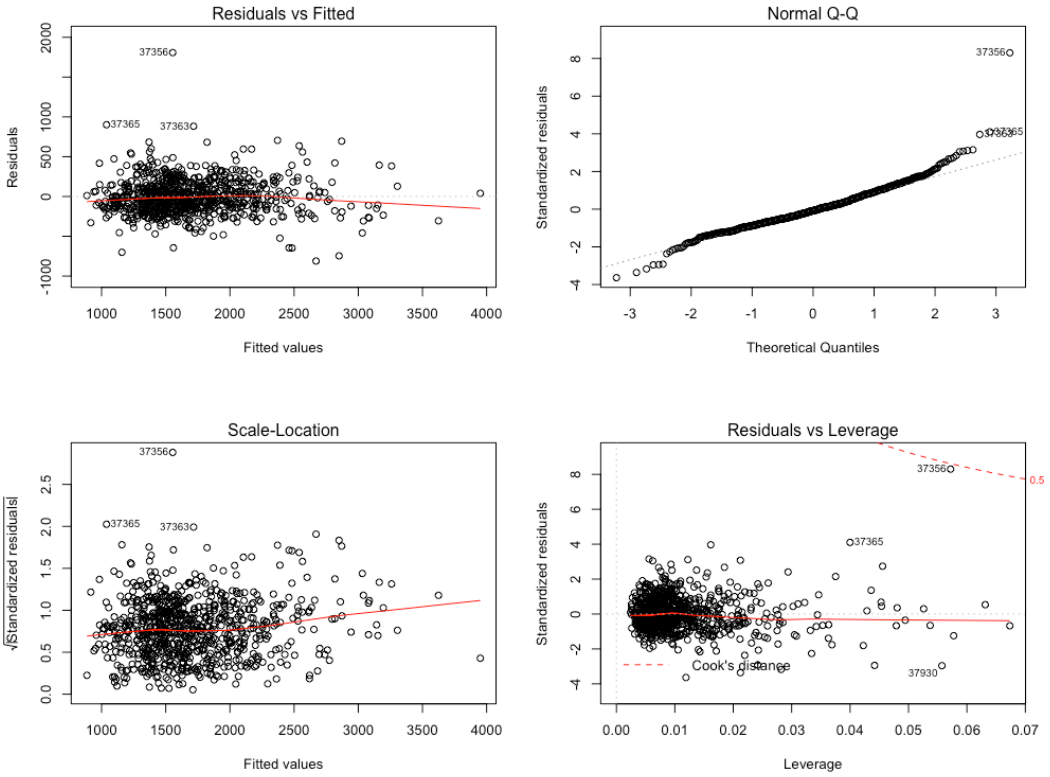
1. States With the Top GDP Growth. (n.d.). Retrieved from http://www.governing.com/topics/finance/states-top-real-gdp-growth-2017.html
2. Glodan Oak Research, mortgage data in the USA in 2017

Appendix

I. Summary of Regression Analysis by State

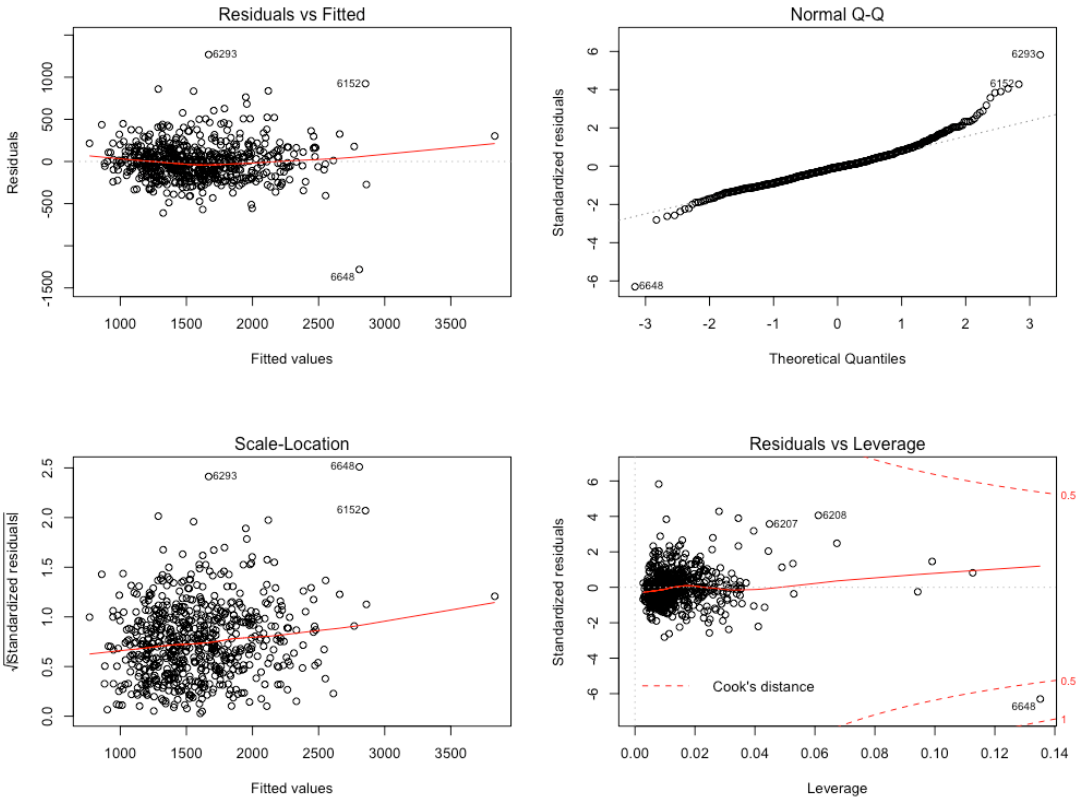
Summary of Regression Analysis on Washington

```
> summary(lm_Washington)

Call:
lm(formula = hc_mortgage_median ~ hc_median + hi_median + married +
    separated + male_age_median + female_age_median + hs_degree +
    married + separated + divorced, data = Washington)

Residuals:
    Min      1Q  Median      3Q     Max
-810.89 -140.80  -18.28  125.62 1807.85

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.406e+02  9.803e+01   3.474 0.000540 ***
hc_median         1.253e+00  6.860e-02  18.260  < 2e-16 ***
hi_median         1.058e-02  5.833e-04  18.141  < 2e-16 ***
married          -8.842e+02  1.046e+02  -8.450  < 2e-16 ***
separated         5.923e+02  5.064e+02   1.170 0.242544
male_age_median   3.575e+00  2.302e+00   1.553 0.120767
female_age_median 4.210e-01  2.105e+00   0.200 0.841515
hs_degree         4.726e+02  1.118e+02   4.229 2.63e-05 ***
divorced         -8.137e+02  2.278e+02  -3.572 0.000375 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 224.3 on 788 degrees of freedom
Multiple R-squared:  0.7946,    Adjusted R-squared:  0.7925
F-statistic:   381 on 8 and 788 DF,  p-value: < 2.2e-16
```



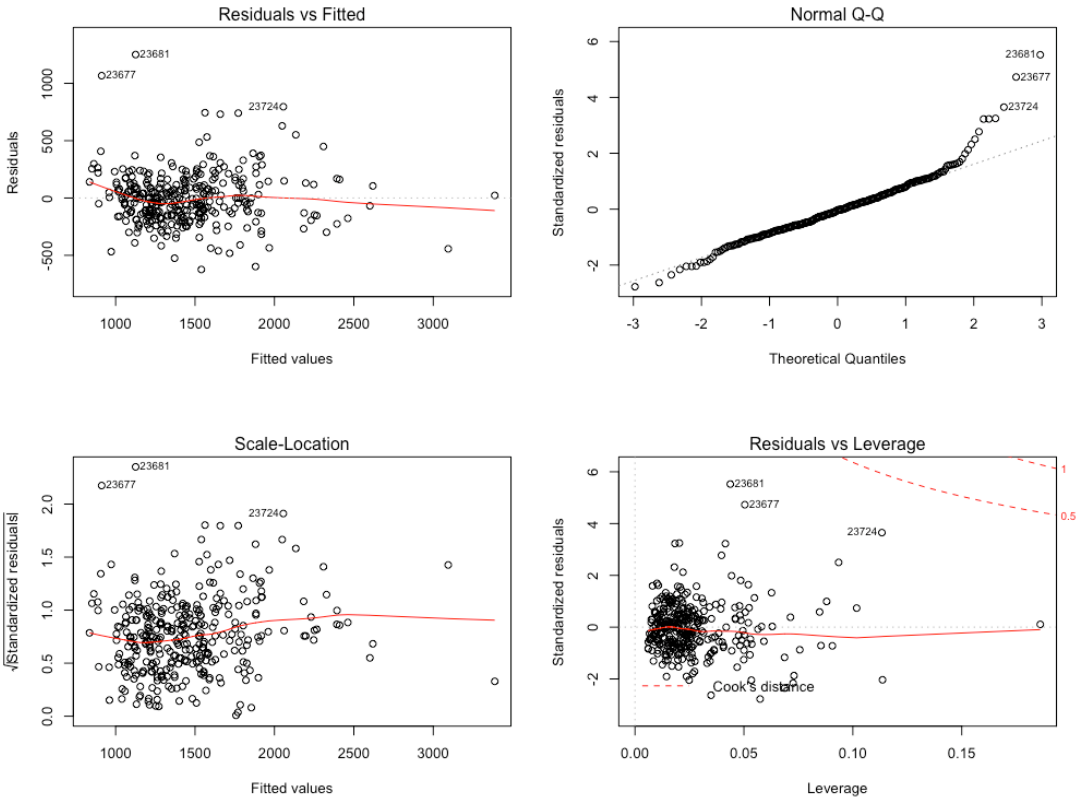
Summary of Regression Analysis on Colorado

```
> summary(lm_Colorado)

Call:
lm(formula = hc_mortgage_median ~ hc_median + hi_median + married +
    separated + male_age_median + female_age_median + hs_degree +
    married + separated + divorced, data = Colorado)

Residuals:
     Min       1Q   Median       3Q      Max
-1281.68  -132.44   -11.24   105.41  1268.66

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.345e+02  1.107e+02   4.829 1.72e-06 ***
hc_median          1.019e+00  7.901e-02  12.897  < 2e-16 ***
hi_median          9.236e-03  5.468e-04  16.890  < 2e-16 ***
married           -8.449e+02  9.913e+01  -8.523  < 2e-16 ***
separated         -3.918e+02  6.201e+02  -0.632   0.5278
male_age_median    1.411e+01  2.482e+00   5.684 2.00e-08 ***
female_age_median -5.270e+00  2.109e+00  -2.499   0.0127 *
hs_degree          3.148e+02  1.221e+02   2.577   0.0102 *
divorced          -1.594e+03  2.188e+02  -7.286 9.51e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 218.7 on 634 degrees of freedom
Multiple R-squared:  0.7472,    Adjusted R-squared:  0.744
F-statistic: 234.3 on 8 and 634 DF,  p-value: < 2.2e-16
```



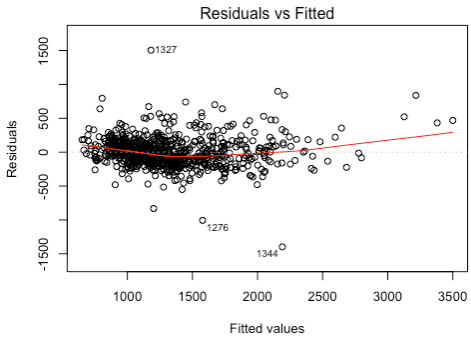
Summary of Regression Analysis on Nevada

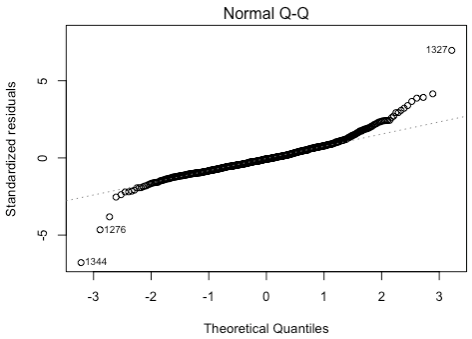
```
> summary(lm_Nevada)

Call:
lm(formula = hc_mortgage_median ~ hc_median + hi_median + married +
    separated + male_age_median + female_age_median + hs_degree +
    married + separated + divorced, data = Nevada)

Residuals:
    Min      1Q  Median      3Q     Max
-623.68 -141.81   -6.33  115.36 1251.10

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.074e+02  1.216e+02   1.706    0.089 .
hc_median          1.215e+00  9.628e-02  12.620   <2e-16 ***
hi_median          9.349e-03  9.877e-04   9.465   <2e-16 ***
married           -1.324e+02  1.767e+02  -0.750    0.454
separated          1.074e+03  7.272e+02   1.477    0.141
male_age_median    2.634e+00  3.138e+00   0.839    0.402
female_age_median  1.188e-02  2.906e+00   0.004    0.997
hs_degree          1.268e+02  1.449e+02   0.875    0.382
divorced           1.590e+02  3.108e+02   0.512    0.609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 231.5 on 334 degrees of freedom
Multiple R-squared:  0.7031,    Adjusted R-squared:  0.696
F-statistic: 98.86 on 8 and 334 DF,  p-value: < 2.2e-16
```

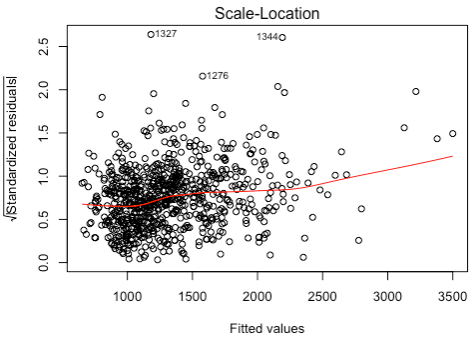


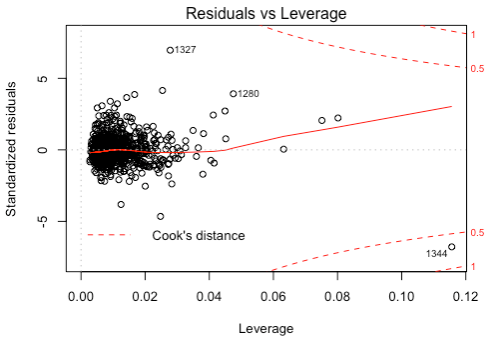
Summary of Regression Analysis on Arizona

```
> summary(lm_Arizona)

Call:
lm(formula = hc_mortgage_median ~ hc_median + hi_median + married +
    separated + male_age_median + female_age_median + hs_degree +
    married + separated + divorced, data = Arizona)

Residuals:
     Min       1Q   Median       3Q      Max
-1397.71  -123.40   -16.22   108.06  1505.11

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.357e+02  6.800e+01   3.466 0.000557 ***
hc_median          1.442e+00  7.878e-02  18.305  < 2e-16 ***
hi_median          9.963e-03  5.725e-04  17.402  < 2e-16 ***
married           -3.346e+02  9.478e+01  -3.531 0.000440 ***
separated         -3.239e+02  4.892e+02  -0.662 0.508127
male_age_median    3.432e+00  1.834e+00   1.871 0.061695 .
female_age_median -5.903e-01  1.808e+00  -0.327 0.744077
hs_degree          8.848e+01  8.898e+01   0.994 0.320383
divorced          -3.651e+02  1.933e+02  -1.889 0.059337 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 219.1 on 764 degrees of freedom
Multiple R-squared:  0.7756,    Adjusted R-squared:  0.7732
F-statistic:   330 on 8 and 764 DF,  p-value: < 2.2e-16
```



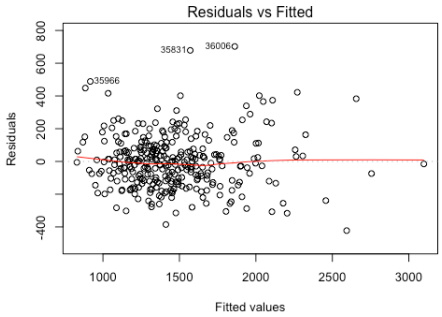


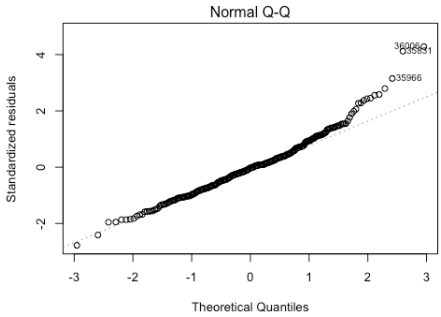


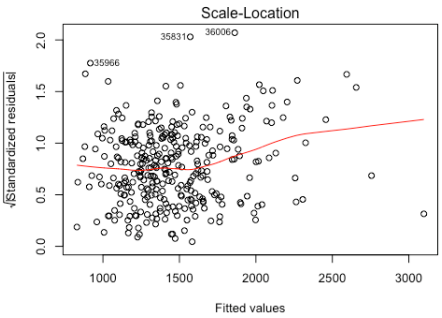


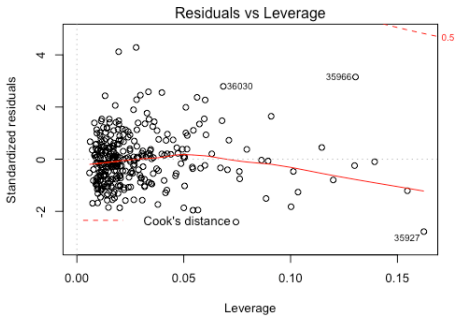
Summary of Regression Analysis on Utah

```
> summary(lm_Utah)

Call:
lm(formula = hc_mortgage_median ~ hc_median + hi_median + married +
    separated + male_age_median + female_age_median + hs_degree +
    married + separated + divorced, data = Utah)

Residuals:
    Min      1Q  Median      3Q     Max
-422.53 -111.55   -3.82   80.89  702.29

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.306e+02  1.425e+02   2.321  0.02095 *
hc_median          8.953e-01  1.135e-01   7.887 5.35e-14 ***
hi_median          1.072e-02  6.743e-04  15.902  < 2e-16 ***
married           -5.082e+02  1.082e+02  -4.695 4.01e-06 ***
separated         -6.345e+02  6.418e+02  -0.989  0.32358
male_age_median    5.253e+00  2.932e+00   1.792  0.07412 .
female_age_median -1.203e+00  2.755e+00  -0.437  0.66261
hs_degree          3.393e+02  1.517e+02   2.237  0.02601 *
divorced          -7.458e+02  2.777e+02  -2.686  0.00763 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 166.2 on 310 degrees of freedom
Multiple R-squared:  0.8106,    Adjusted R-squared:  0.8057
F-statistic: 165.8 on 8 and 310 DF,  p-value: < 2.2e-16
```









II. Data Visualisation of Statical data by State

https://public.tableau.com/views/DA_8/MortgageDataofUnitedState?:embed=y&:display_count=yes&publish=yes