Part 1

I. Identify the question

Purchasing family housing is one of the major expenses that almost all families need to face. How to buy a cost-effective house has become a hot topic in society. There are many factors that affect housing prices, such as location, environment, daylighting, house material and so on. Therefore, we will research and analyze "what factors are directly and highly related to housing prices" as an exploratory question.

II. The source of data

Our data source is selected from the kaggle website. In the "house price" competition, the competition provides 3 data tables, namely ‘train’, ‘test’ and ‘samples’. We choose the ‘train’ data table as our data source for this project.

III. Sampling strategy

We use the random function in excel to randomly number 1460 data, and then select 50 data with largest random numbers as the database of this project.

Part 2

Through the full description of data, we can have a brief idea about how the target variable was distributed and better understand the data structure. From the data description, the cheapest sale price is $34,900, the most expensive house price is $370,878 and the range is $335,978. In the sample data, we can tell from the mean, median and mode that the data was positively skewed since the mean is greater than median. And the skewness is 0.9083. According to the standard deviation, we can infer that the data was widely distributed instead of narrowed. By using Sturges’ formula, we divided the target variable into seven segments and each of the class width is 50,854.9. By counting the frequency, we know that 42% of house price in the sample dataset was from $136,609 to $187,464. We also observed a few outliers from the boxplot which are 2 houses priced over $330,000. The percentile also allows us to know the data ranking. 25% of the house sale price is under $135,475 and 75% of house sale price is under $208,550.

Part 3.

To apply adequate hypothesis tests for the features that take impacts on the pre-owned houses' selling price, a test of the price itself compared with the one reported in the market needs to be considered initially in order to evaluate whether the sale price is changeable or stable regardless of the changes. To meet this purpose, an average sale price of a pre-owned house constructed in 2005 is found as $222,000 to be a reference announced by Fed Prime Rate (2021), and the parameter of interest in this case is the mean and we plan to find out whether the mean sale price upon our sample data is lower than this national average price with a confidential level as 95% (as mostly house included in our sample is built before 2005 and it assumes that these is a increasing trend existing in the house's sale market). Hence, the hypotheses are:

$$H_0: \mu = 222000$$

$$H_1: \mu < 222000$$

μ : the average sale price calculated based on our sample data. Then, a t-test statistic is implemented with the formula structured as below to calculate the t-value and p-value.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{171954 - 222000}{\frac{67756.74}{\sqrt{50}}} = -5.22$$

$$p - value = 1.79394E - 06 < 0.05$$

Therefore, with a 5% level of significance, the null hypothesis should be rejected and it can be concluded that the average sales price is lower than the national average sale price of the pre-owned house established in 2005. Furthermore, an equivalent approach to p-value for conducting hypothesis tests can also be utilized by comparing the critical value directly to the t-value, and it generates the same result as what we found above with the fact t-value is less than the critical value which is -1.68. In addition, the 99% confidence interval for average house sale price is [$146,273.57, $197,634.43] which also tests the hypothesis and indicates that the average house selling price is different from the sales price of houses built in the subsequent year. The detailed calculation process can be found in the Excel attached.

After analyzing the sale price as a single variable, we combine it with quantitative and nominal variables such as the year period and the kitchen quality which we will use to apply the further hypothesis tests. Firstly, it is concerned about the impacts taken by the year period of selling, and the focus of this test is to find the gap between the sale prices of houses pre-owned for less than 50 years and those owned for longer than 50 years before the sales. Based on this aim, the data has been divided into two samples and a F-test is drafted to find the mean and variance, and then the hypothesis is taken out as below with 95% confidence level, where g represents the price gap.

$$H_0: g = 30000$$

$$H_1: g > 30000$$

To simplify the procedure, we use Excel's data analysis tool to help us to gain the result table, which shows that the p-value is less than 0.05 that the null hypothesis should be rejected.

As for the analysis practiced on the sale price and the kitchen quality, we assume that there is a belief that at least 15% sales of pre-owned houses follow the rule that the house tends to have an excellent kitchen quality if its value is greater than \$200,000 for sale. To test this assumption, we have 22 total samples with a certain selling price and the number of those owning an excellent kitchen is 4, and a hypothesis test for a population proportion will be launched at the 5% level of significance. Moreover, as long as $n \times p \geq 5$ and $n \times (1-p) \geq 5$, it regards that the resulting distribution of proportions is approximately normal.

$$H_0: p = 0.15$$

$$H_1: p > 0.15$$

$$\hat{p} \sim N\left(p, \frac{\sqrt{p(1-p)}}{n}\right), z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

As p-value equals 0.024 that is less than 0.05, this is sufficient evidence to conclude, at the 5% significant level, that at least 15% sales of pre-owned houses follow the rule that the house will have an excellent kitchen if its price is higher than \$200,000.

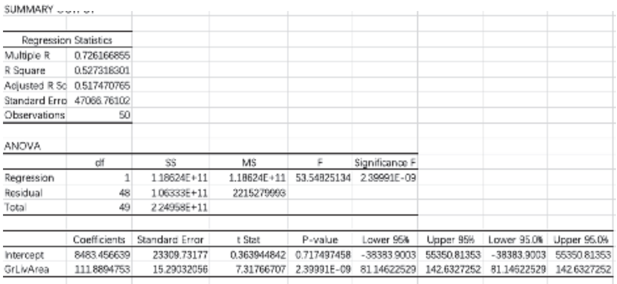
Part 4

Based on common sense, we believe that there is a certain relationship between living area (GrLivArea) and house sale prices (SalePrice). So, what kind of relationship exists between them, we can explore through regression analysis. Regression analysis refers to a statistical analysis method that determines the quantitative relationship between two or more variables. According to textbook, Regression analysis enables us to estimate the strength and direction of relations between variables. So, we can know the strength of the impact of living area (GrLivArea) on house sale prices (SalePrice) and their relationship through regression analysis. In the regression analysis, we use the living area (GrLivArea) as the independent variable and the house sale prices (SalePrice) as the dependent variable. We sample fifty living area data for regression analysis. We used two software to do the regression analysis. The first one is excel. Through the data analysis of excel, we can get the basic parameters. We can get the coefficients of GrLivAREAd of the regression line as 111.889, intercept as 8483.457, and R square as 0.527.the relationship between living area (GrLivArea) and house sale prices (SalePrice) as the following:

SalePrice=8483.457 + 111.889*GrLivAREA

In addition, the regression line graph is made through python, and we can clearly see the causality between the living area (GrLivArea) and house sale prices (SalePrice) through the graph.

Therefore, through regression analysis, we can draw a conclusion: the larger the house area, the higher the price. And the living area (GrLivArea) has a very large impact on the price (111.889).



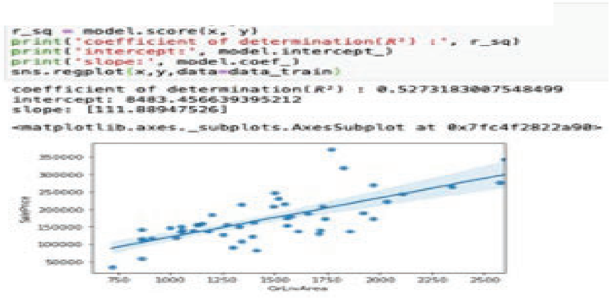
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726166855
R Square	0.527318301
Adjusted R Sc	0.517470765
Standard Erro	47066.76102
Observations	50

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1.18624E+11	1.18624E+11	53.54825134	2.39991E-09
Residual	48	1.06333E+11	2215279993		
Total	49	2.24958E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	8483.456639	23309.73177	0.363944842	0.717497458	-38383.9003	55350.81353	-38383.9003	55350.81353
GrLivArea	111.8894753	15.29032056	7.31766707	2.39991E-09	81.14622529	142.6327252	81.14622529	142.6327252



```
r_sq = model.score(x, y)
print('coefficient of determination(R²) :', r_sq)
print('intercept:', model.intercept_)
print('slope:', model.coef_)
sns.regplot(x,y,data=data_train)
```

```
coefficient of determination(R²) : 0.5273183007548499
intercept: 8483.456639395212
slope: [111.88947526]

<matplotlib.axes._subplots.AxesSubplot at 0x7fc4f2822a90>
```

Reference:

Fed Prime Rate (2021). “*Average & Median Sale Price for A Previously Occupied (Used) Home in The United States*”. [Online]. Available at: http://www.fedprimerate.com/preowned_used-home_sales_price_history.htm